nature genetics

Genome-wide association studies of 14 agronomic traits in rice landraces

Xuehui Huang[1,2,10], Xinghua Wei[3,10], Tao Sang[4,10], Qiang Zhao[1,2,10], Qi Feng[1,10], Yan Zhao[1], Canyang Li[1], Chuanrang Zhu[1], Tingting Lu[1], Zhiwu Zhang[5], Meng Li[5,6], Danlin Fan[1], Yunli Guo[1], Ahong Wang[1], Lu Wang[1], Liuwei Deng[1], Wenjun Li[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Tao Huang[1], Taoying Zhou[1], Yufeng Jing[1], Wei Li[1], Zhang Lin[1], Edward S Buckler[5,7], Qian Qian[3], Qi-Fa Zhang[8], Jiayang Li[9] & Bin Han[1,2]

Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. The loci identified through GWAS explained ~36% of the phenotypic variance, on average. The peak signals at six loci were tied closely to previously identified genes. This study provides a fundamental resource for rice genetics research and breeding, and demonstrates that an approach integrating second-generation genome sequencing and GWAS can be used as a powerful complementary strategy to classical biparental cross-mapping for dissecting complex traits in rice.

Rice (*Oryza sativa* L.) is a staple food for more than half of the world population. Rice landraces have evolved from their wild progenitor under natural and human selection, leading to the maintenance of high genetic diversity[1,2]. These cultivated varieties also have a high capacity to tolerate biotic and abiotic stress, resulting in highly stable yields and an intermediate yield under a low-input agricultural system. Identifying the genetic basis of these diverse varieties will provide important insights for breeding elite varieties for sustainable agriculture.

GWAS have emerged as a powerful approach for identifying genes underlying complex diseases at an unprecedented rate[3–6]. However, despite their promise, GWAS have largely not been applied to the dissection of complex traits in crop plants[7–9]. This is due mainly to the lack of effective genotyping techniques for plants and the limited resources for developing high-density haplotype maps like those seen in other well-developed systems, such as the human genome HapMap project[3,4]. Rice is an ideal candidate system for the application of GWAS because it is self-fertilizing and has a high-quality reference genome sequence[10] and phenotyping resources. Such a system should permit the identification of high-quality haplotypes necessary to accurately associate molecular markers with phenotypes.

Here we have genotyped rice landraces through direct resequencing of their genomes by adopting sequencing-by-synthesis technology, which represents a step forward from the oligonucleotide array technology widely used for GWAS[11–13]. More than 500 diverse rice landraces, representing a large collection of rice accessions, were sequenced at approximately onefold genome coverage. The resulting data set captures more of the common sequence variation in cultivated rice than any other data set to date. Using a highly accurate imputation method, we constructed a high-density rice haplotype map and performed GWAS for 14 agronomic traits to identify a substantial number of loci potentially important for rice production and improvement. Some loci were mapped at close to gene resolution, indicating that GWAS of rice landraces could provide an effective approach for gene identification.

RESULTS

Genome sequencing and SNP identification

From a collection of ~50,000 rice accessions originating in China, we have undertaken an effort to build a large sample of morphologically, genetically and geographically diverse landraces for genetic studies. In this study, a total of 517 landraces were selected and comprehensively phenotyped (see Online Methods). We genotyped

[1]National Center for Gene Research, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. [3]State Key Laboratory of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou, China. [4]Department of Plant Biology, Michigan State University, East Lansing, Michigan, USA. [5]Institute for Genomic Diversity, Cornell University, Ithaca, New York, USA. [6]National Center for Soybean Improvement, State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Agriculture, Nanjing Agricultural University, Nanjing, China. [7]US Department of Agriculture–Agricultural Research Service, Ithaca, New York, USA. [8]National Key Laboratory of Crop Genetic Improvement, National Center for Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan, China. [9]National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [10]These authors contributed equally to this work. Correspondence should be addressed to B.H. (bhan@ncgr.ac.cn).

Received 10 May; accepted 27 September; published online 24 October 2010; doi:10.1038/ng.695

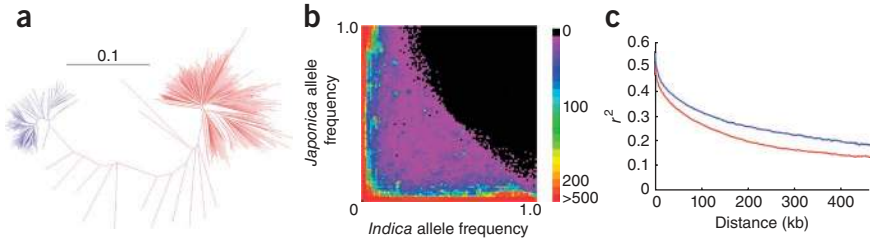

Figure 1 Divergence and geographic origins of 517 rice landraces. (**a**) Neighbor-joining tree constructed from simple matching distance of all SNPs. Red, *indica*; blue, *japonica*; purple, intermediate. (**b**) Comparison of allele frequencies between *indica* and *japonica*. For each SNP, we identified the minor allele across all landraces and then calculated the frequency of this allele in *indica* and *japonica*. Color index indicates the number of SNPs with each set of allele frequencies. (**c**) Genome-wide average LD decay estimated from 373 *indica* (red) and 131 *japonica* (blue) landraces.

these landraces with approximate onefold-coverage genome sequencing using a barcoded multiplex sequencing approach[14] on the Illumina Genome Analyzer II (**Supplementary Table 1** and **Supplementary Fig. 1**). Three additional cultivars with accurate genome sequences were also sequenced as internal controls for evaluating sequence accuracy (**Supplementary Note**). More than 2.7 billion 73-bp paired-end reads were generated. In total, all sequences used for SNP calling comprised ~508-fold coverage of the rice genome.

Sequence reads were aligned to the rice reference genome for SNP identification. We used the alignment of reads to build consensus genome sequences for each rice accession, with a series of filtering criteria that eliminated sequencing and mapping errors (see Online Methods and **Supplementary Note** for details). The resulting consensus sequence of each rice accession covered 27.4% of the reference genome on average (ranging from 12.0% to 46.7%). Comparisons of the consensus sequence against bacterial artificial chromosome (BAC) sequences and high-coverage Illumina data showed that the sequence specificity reached 99.9% (**Supplementary Table 2**). The SNP calling procedure was then based on discrepancies between the consensus sequence and the reference genome. After exclusion of singleton SNPs, the SNP calling error rate was reduced to 2.7% (**Supplementary Fig. 2** and **Supplementary Table 3**). A total of 3,625,200 nonredundant SNPs were identified, resulting in an average of 9.32 SNPs per kb, with 87.9% of the SNPs located within 0.2 kb of the nearest SNP (**Supplementary Fig. 3a**). About 78% of all SNPs were found in intergenic regions; of the remaining SNPs, the largest number were in introns of annotated genes, followed by coding regions and untranslated regions of annotated genes (**Supplementary Fig. 3b**). The chromosomal distribution of the SNPs is shown in **Supplementary Figure 3c**. Despite the high density of our SNP map, however, the recall rate (the rate at which all actual SNPs are recalled) was 20.1%. This was probably due to uneven sampling of short reads from low-coverage sequencing and the complexity and repetitiveness of the rice genome.

To gain insights into potential functional effects of the detected SNPs, we further analyzed the SNPs in coding regions. A total of 167,514 SNPs were found in the coding regions of 25,409 annotated genes with transcript support (RAP2 database). We also found 3,625 large-effect SNPs (SNPs representing mutations predicted to cause large effects). **Supplementary Table 4** lists the types of predicted effects of annotated SNPs. Among the annotated genes, 107 genes were over-represented for large-effect changes, which may indicate that these are incorrect gene annotations or pseudogenes (**Supplementary Table 5**). Moreover, we observed that 11 gene families showed significantly higher ratios of nonsynonymous to synonymous changes ($P < 0.01$), which may reflect positive or relaxed selection (**Supplementary Fig. 4**). These include genes encoding NB-LRR proteins, which are known to be involved in disease resistance.

Population structure and geographic differentiation

The phylogenetic relationships of the 517 selected Chinese rice landraces were determined using the genetic distances calculated

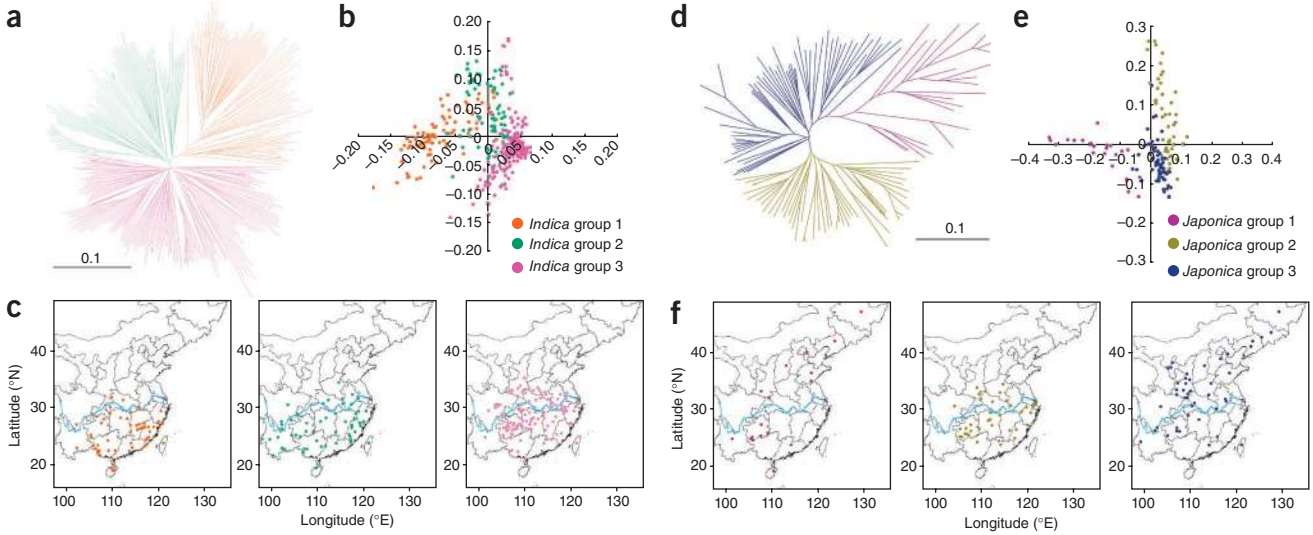

Figure 2 Population structures of Chinese landraces of both subspecies. (**a**) Neighbor-joining tree of 373 *indica* landraces. The three *indica* subgroups identified from the tree are color-coded in **a**–**c**. (**b**) PCA plots of the first two components of 373 *indica* landraces. (**c**) Geographic origins of landraces of each *indica* subgroup. (**d**) Neighbor-joining tree of 131 *japonica* landraces. The three *japonica* subgroups identified from the tree are color-coded in **d**–**f**. (**e**) PCA plots of the first two components of 131 *japonica* landraces. (**f**) Geographic origins of landraces of each *japonica* subgroup.

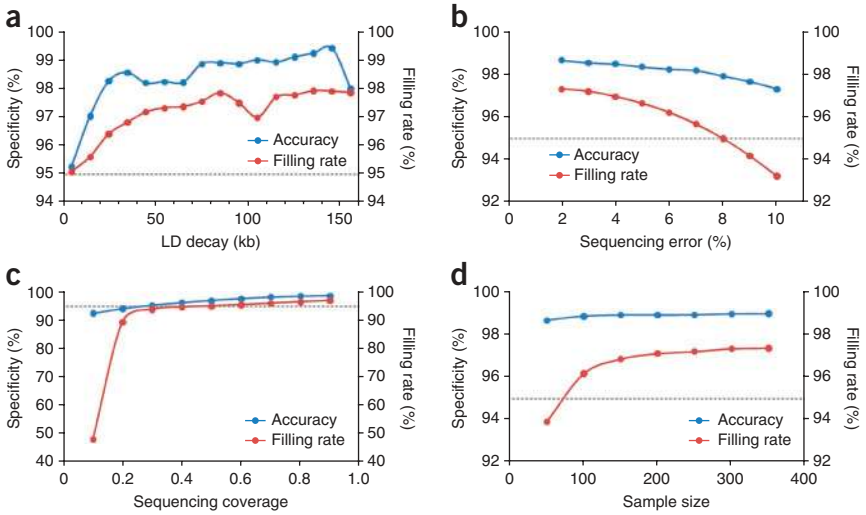

Figure 3 Influence of populational and experimental factors on the performance of the KNN-based imputation method. Performance of the imputation was evaluated by specificity and filling rate. The specificity of the genotype data set after imputation of missing genotypes was assessed against BAC sequences and high-coverage Illumina data. Filling rate was defined as the non–missing data rate of the genotype data set after imputation. Gray horizontal dashed lines indicate 95%, highlighting different scales used on different panels. (**a**) Genomic regions with LD decay range varying from <10 kb to >150 kb in 10-kb intervals. (**b**) Sequencing error rates. Various higher error rates were introduced for the simulation. (**c**) Sequencing coverage. Sequences were removed to simulate data sets derived from lower sequencing coverage. (**d**) Population size. Individuals were randomly removed to create a series of smaller populations for simulation.

from the SNPs (**Fig. 1**). The resulting neighbor-joining tree showed two divergent groups belonging to the two subspecies of cultivated rice, *Oryza sativa* ssp. *indica* and ssp. *japonica*. On the basis of the neighbor-joining tree, we were able to identify 373 typical *indica* and 131 typical *japonica* landraces (**Fig. 1a**). The geographic distribution of *japonica* landraces extends further north than that of *indica* (**Supplementary Fig. 5**). There are 13 intermediate landraces, which may have resulted from occasional historical hybrids between *indica* and *japonica* that experienced partial reproductive isolation.

From the SNP data, sequence diversity (π) was estimated at 0.0024 for all sampled landraces, and 0.0016 and 0.0006 for *indica* and *japonica*, respectively. These estimates suggest that the overall genetic variation of the landraces we studied represents at least 80% of the world's rice cultivars, and the *indica* landraces have much higher genetic diversity than the *japonica* landraces[15,16]. The population-differentiation statistic (F_{ST}) between the *indica* and *japonica* landraces was estimated at 0.55, indicating a very strong population differentiation. After screening all SNPs that were highly differentiated in frequency between *indica* and *japonica*, we found a total of 367,081 (~10%) SNPs that were nearly fixed (with an allele frequency >0.95 in one subspecies and <0.05 in the other) and a total of 127,729 (~3.5%) SNPs that were completely fixed (**Fig. 1b**). These subspecies-specific signatures may reflect, as well as affect, the strong *indica-japonica* differentiation. We observed that the subset of complete-differentiation SNPs had a smaller proportion of coding-region SNPs ($P < 0.0001$) and a

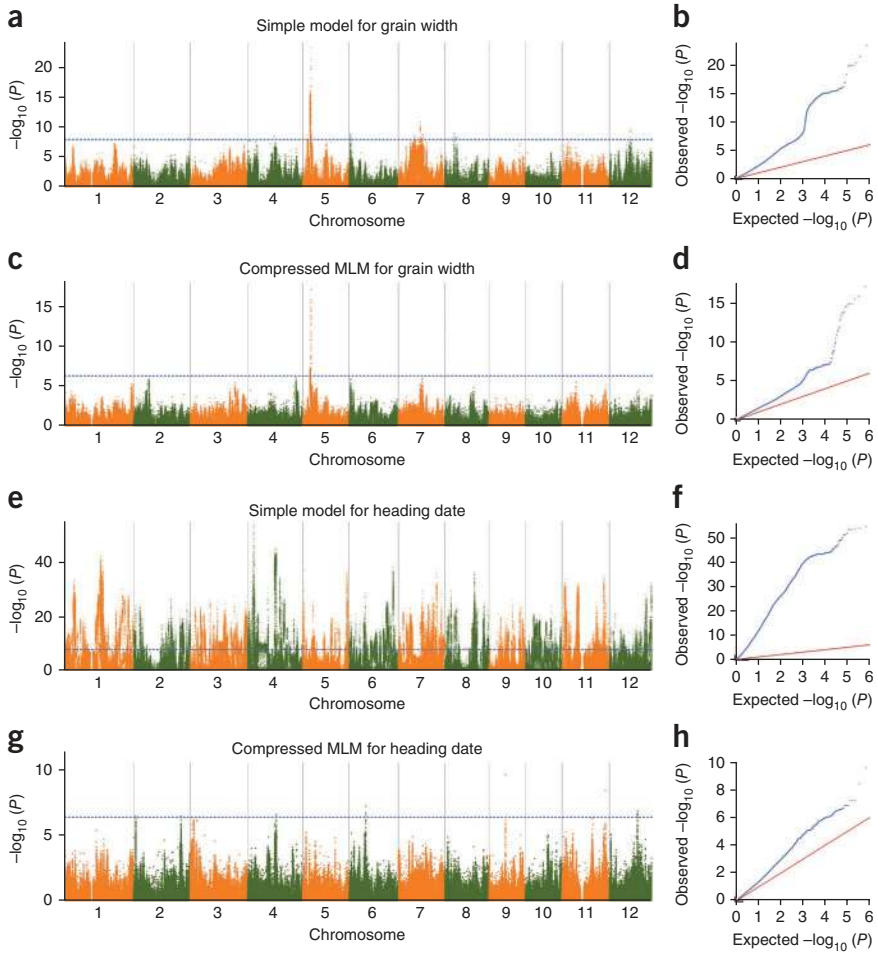

Figure 4 Genome-wide association studies of grain width and heading date. (**a**) Manhattan plots of the simple model for grain width. Negative $\log_{10}$-transformed *P* values from a genome-wide scan are plotted against position on each of 12 chromosomes. Blue horizontal dashed line indicates the genome-wide significance threshold. (**b**) Quantile-quantile plot of the simple model for grain width. (**c**) Manhattan plots of compressed MLM for grain width, as in **a**. (**d**) Quantile-quantile plot of compressed MLM for grain width. (**e**) Manhattan plots of the simple model for heading date, as in **a**. (**f**) Quantile-quantile plot of the simple model for heading date. (**g**) Manhattan plots of compressed MLM for heading date, as in **a**. (**h**) Quantile-quantile plot of compressed MLM for heading date.

lower nonsynonymous-to-synonymous ratio ($P < 0.0001$) than did the set of all SNPs detected in this study. Furthermore, across the whole genome we identified 53 genes that contained large-effect complete-differentiation SNPs; among these might be genes involved in the differentiation of the two subspecies (**Supplementary Table 6**).

We then investigated the population structure within subspecies. According to the neighbor-joining tree as well as the principal-component analysis (PCA)[17], both *indica* and *japonica* had three subgroups, designated 1, 2 and 3 (**Fig. 2**). It has previously been suggested that the photoperiod and temperature clines along latitudes may have been the primary factors driving differentiation of cultivated rice in China[1]. We tested the difference in latitude distribution and found that *indica* group 3 was significantly more northern than *indica* group 1 ($P < 0.0001$) or *indica* group 2 ($P < 0.05$) (**Fig. 2c**). A similar pattern was observed in *japonica*, whose group 3 was significantly more northern than the other two ($P < 0.0001$) (**Fig. 2f**). The measure of population differentiation, F_{ST}, was estimated at 0.17 among the three subgroups of *indica*, suggesting a moderate level of differentiation within *indica*. The genetic differentiation within *japonica* was slightly less ($F_{ST} = 0.14$) but still higher than that between different human populations ($F_{ST} = 0.12$)[3]. The fine-scale maps for the sequence diversity π and the population differentiation F_{ST} of the two subspecies showed great variation along chromosomes (**Supplementary Fig. 6** and **Supplementary Fig. 7**). We observed that some regions had a high F_{ST}, including a total length of 2.1 Mb in *japonica* and 0.6 Mb in *indica* with an $F_{ST} > 0.5$, indicating that they contain loci that may be involved in the geographic adaptation.

Table 1 Genome-wide significant association signals of agronomic traits using the compressed MLM

Trait	Chromosome	Position (IRGSP 4)	Major allele	Minor allele	Minor allele freq.	*P* value (compressed MLM)[a]	Known loci[b]
Tiller number	4	3,760,194	A	T	0.20	3.2×10^{-7}	
	9	23,332,559	A	G	0.34	1.5×10^{-7}	
	10	15,239,407	T	A	0.10	4.1×10^{-7}	
Grain width	5	4,907,158	C	G	0.21	2.7×10^{-9}	
	5	5,341,575	G	A	0.17	7.2×10^{-18}	*qSW5* (ref. 29)
Grain length	3	17,371,398	G	C	0.06	1.3×10^{-10}	*GS3* (ref. 30)
	3	17,637,475	C	A	0.08	2.7×10^{-11}	
	3	23,349,781	A	C	0.13	3.3×10^{-7}	
	5	5,343,949	A	G	0.20	1.7×10^{-7}	
	11	3,072,370	C	T	0.11	3.8×10^{-7}	
Spikelet number	7	18,005,615	C	T	0.44	7.1×10^{-8}	
	10	5,976,140	C	T	0.06	1.3×10^{-7}	
Gelatinization temperature	6	6,726,252	C	T	0.20	7.1×10^{-9}	*ALK* (ref. 26)
Amylose content	6	1,770,929	T	C	0.14	5.0×10^{-26}	*Waxy* (refs. 27,28)
	6	6,189,558	A	T	0.11	3.0×10^{-8}	
	6	6,709,537	C	T	0.19	7.4×10^{-12}	
Apiculus color	6	5,335,519	A	G	0.33	5.6×10^{-27}	*OsC1* (ref. 23)
	6	7,671,184	T	C	0.32	9.4×10^{-9}	
Pericarp color	2	27,066,598	A	G	0.24	2.2×10^{-9}	
	7	6,123,504	A	G	0.34	2.1×10^{-52}	*Rc* (ref. 24)
	8	12,483,076	T	G	0.21	1.3×10^{-11}	
Hull color	6	10,378,142	T	C	0.06	3.8×10^{-7}	
	9	7,366,211	T	C	0.20	3.3×10^{-13}	*Ibf* (ref. 25)[c]
Heading date	2	1,439,288	G	A	0.42	3.9×10^{-7}	
	2	30,818,552	G	C	0.07	3.8×10^{-7}	
	4	18,773,995	A	T	0.25	3.0×10^{-7}	
	6	11,083,237	G	A	0.05	6.6×10^{-8}	
	9	10,738,885	C	A	0.06	2.8×10^{-10}	
	11	28,247,391	C	T	0.12	4.2×10^{-9}	
	12	18,324,888	G	A	0.06	1.4×10^{-7}	
Drought tolerance	1	5,536,395	G	T	0.11	4.1×10^{-7}	
	5	2,275,357	A	C	0.06	2.5×10^{-8}	
	6	28,243,628	C	T	0.09	3.4×10^{-9}	
	11	21,161,361	G	C	0.08	8.5×10^{-12}	
Degree of seed shattering	2	25,025,325	C	T	0.16	4.7×10^{-8}	
	5	948,266	T	C	0.38	2.5×10^{-7}	
	10	2,319,249	T	G	0.06	2.2×10^{-7}	

[a]*P* values of the association signals from the simple model are listed in **Supplementary Table 7**. [b]Details of the known loci are provided in **Figure 4** and the **Supplementary Note**. [c]The causal gene has not yet been identified and confirmed.

Whole-genome patterns of linkage disequilibrium

We then analyzed LD for *indica* and *japonica* landraces using the SNP data. The LD decay rate was measured as the chromosomal distance at which the average pairwise correlation coefficient (r^2) dropped to half its maximum value. Genome-wide LD decay rates of *indica* and *japonica* were estimated at ~123 kb and ~167 kb, where the r^2 drops to 0.25 and 0.28, respectively (**Fig. 1c**). This is in agreement with the previous estimation that cultivated rice has a long-range LD from close to 100 kb to over 200 kb[13,18], which might be a result of self-fertilization coupled with a relatively small effective population size.

We further examined whole-genome patterns of LD in the two subspecies. LD varied widely across the genomes of both *indica* and *japonica* (**Supplementary Fig. 8**), which would presumably lead to differential resolutions of association mapping at different genomic regions. It is noteworthy that the LD decay rates of *indica* and *japonica* were only weakly correlated across the genome (Spearman correlation coefficient is 0.01). This is markedly different from what has been observed for human, where both local and global patterns of LD vary little among different human populations[3]. The differences between *indica* and *japonica* rice may have accumulated from a relatively long history of partial reproductive isolation of these self-fertilized subspecies.

Constructing a high-density haplotype map of the rice genome

Onefold genome sequencing of more than 500 landraces allowed identification of a large number of SNPs with high accuracy. However, the genotype data set contained numerous missing genotype calls, making it insufficient for GWAS. Data-imputation methods have not been developed to deal specifically with low-coverage genome sequencing data. Of the available imputation models, the *k*-nearest neighbor algorithm (KNN) seemed effective for handling a relatively large number of missing genotypes without a reference haplotype map[19,20]. We adopted the KNN algorithm to explore local haplotype

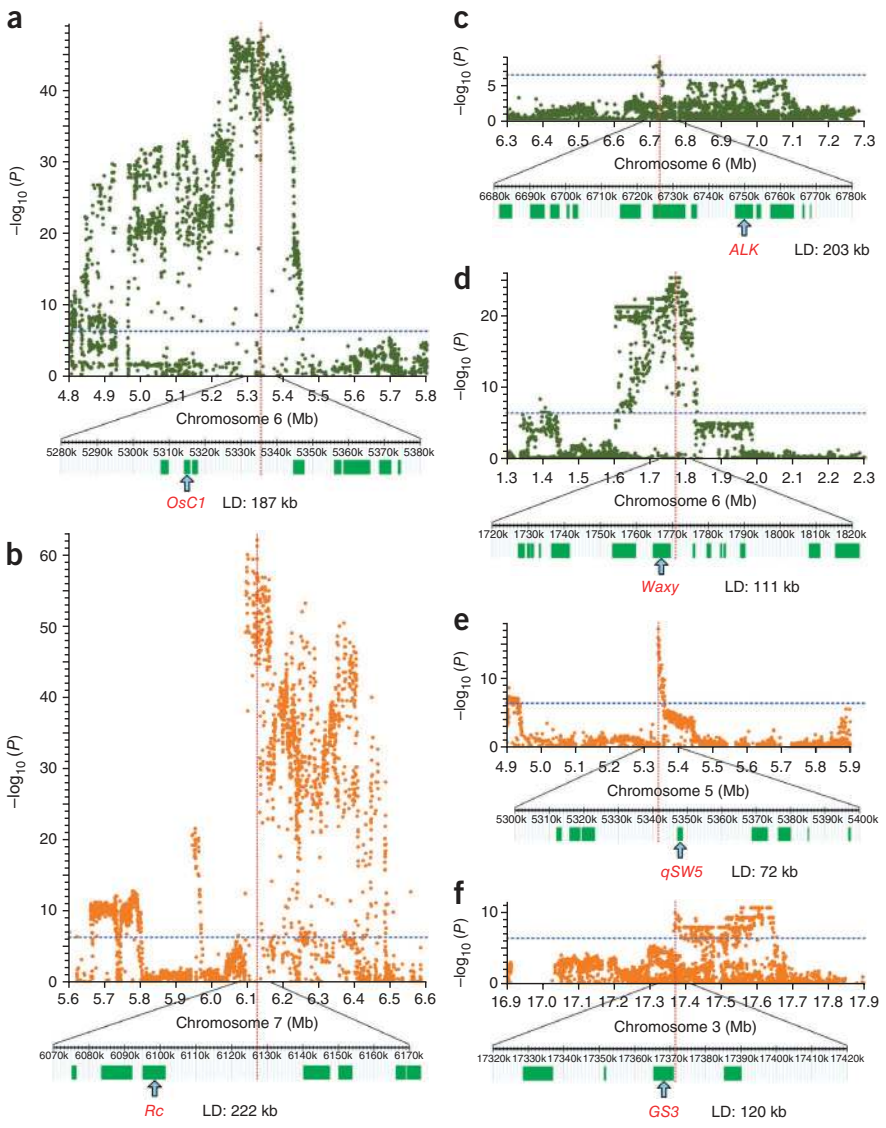

Figure 5 Regions of the genome showing strong association signals near previously identified genes. Top of each panel shows a 0.5-Mb region on each side of the peak SNP (SNP with the lowest *P* value), whose position is indicated by a vertical red line. Negative $\log_{10}$-transformed *P* values from the compressed MLM are plotted on the vertical axis; axis scales are slightly different across panels. Blue horizontal dashed lines indicate the genome-wide significance threshold. Bottom of each panel shows a 50-kb region on each side of the peak SNP, with annotated genes indicated by green boxes. Previously identified genes controlling the traits are labeled. Local LDs of the chromosomal regions containing peak SNPs are given. (**a**) Apiculus color. (**b**) Pericarp color. (**c**) Gelatinization temperature. (**d**) Amylose content. (**e**) Grain width. (**f**) Grain length.

similarity and further developed an algorithm that provided sufficient imputation accuracy and efficiency by optimizing a set of genomic and populational parameters (Online Methods and **Supplementary Note**). The improved algorithm was then used to impute the missing calls of the genotype data set from onefold-coverage genome sequencing (**Supplementary Fig. 9**). The imputation of the genotypes of all 517 landraces reduced missing genotypes from 61.7% to 2.9%, with an accuracy above 98% (**Supplementary Table 3**). It would require more than 20-fold sequencing coverage of the rice genome to yield such a low missing-data rate with a slightly higher accuracy (**Supplementary Fig. 10**). Therefore, our approach, combining second-generation sequencing technology with an effective imputation procedure, permits the quick construction of a high-density haplotype map at a markedly lower cost than microarray-based genotyping.

We then examined the influence of various biological and experimental factors on the performance of the data imputation (**Fig. 3**). Notably, this method performed well even when LD decayed within 10 kb; with this LD decay, the missing-data rate was below 5%, with an accuracy above 95%. This suggests that our imputation method for low-coverage genome sequencing data is also applicable to other genomes with short-range LD.

Genome-wide association studies for 14 agronomic traits

The high-density haplotype map enabled genome-wide association mapping in rice. The strong population structure, along with a slow LD decay rate, makes GWAS in this species not straightforward. To evaluate the performance of GWAS, we carried out GWAS on 14 agronomic traits, which can be divided into five categories: morphological characteristics (tiller number and leaf angle), yield components (grain width, grain length, grain weight and spikelet number), grain quality (gelatinization temperature and amylose content), coloration (apiculus color, pericarp color and hull color) and physiological features (heading date, drought tolerance and degree of seed shattering) (**Supplementary Fig. 11**).

Given the strong population differentiation between the two subspecies of cultivated rice, we did not look for associations across both subspecies. We conducted GWAS for 373 *indica* lines. The sequencing-based genotype data set contained an average density of ~1.7 common SNPs per kb in *indica* (with a minor allele frequency of >0.05). Both the simple model and the compressed mixed linear model (MLM)[21,22] were used to identify association signals. The compressed MLM approach, which took genome-wide patterns of genetic relatedness into account, greatly reduced false positives, as shown in quantile-quantile plots (**Fig. 4** and **Supplementary Figs. 12–23**). A total of 37 association signals were identified with $P < 5 \times 10^{-7}$ from the compressed MLM (**Table 1**). We also identified strong association signals with $P < 10^{-8}$ from the simple model, discarding all but the top five most significant signals for each trait if there was an excess of strong associations (**Supplementary Table 7**). In total, we identified 80 associations for the 14 agronomic traits. The Manhattan plots for both models of all the traits are shown in **Supplementary Figures 12–23**, and detailed information about all significant associations is summarized in **Supplementary Table 8**.

Association signals for six traits were located close to known genes that have been identified previously using mutants or studies of recombinant populations[23–30] (**Fig. 5** and **Supplementary Note**). Although the association resolution varies among loci, mostly owing

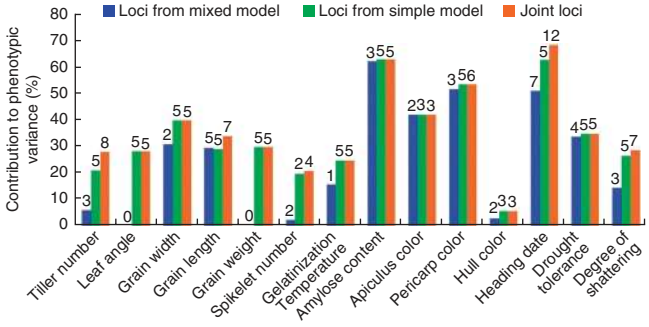

Figure 6 Contributions of identified loci to phenotypic variance of each of 14 agronomic traits. Numbers of loci used to assign contributions to phenotypic variance are indicated at ends of bars. Loci from the compressed MLM are listed in **Table 1**, and loci from the simple model are listed in **Supplementary Table 7**. Joint loci from both models, with redundancy excluded, are listed in **Supplementary Table 8**.

to local LD, the resolutions were all less than 26 kb (within about 1–3 genes). Notably, the peak signals of the GWAS loci often appeared near (but not within) the known genes.

We then screened the causal polymorphisms of three known genes by direct PCR amplification and sequencing, and found that all of them showed a slightly weaker association than peak signals nearby (**Supplementary Table 9**). These results were consistent with similar findings in *Arabidopsis thaliana*[8] and may result from multiple causal polymorphisms of a gene coupled with complex population structure.

Together, the data show that the degree to which population stratification confounds associations varies markedly across traits (**Fig. 4**). An extreme example was observed for heading date (flowering time), a trait that is strongly affected by population structure and controlled by numerous small-effect loci[8,31]. We found that heading date strongly correlated with both population structure and geographic distribution ($R^2 = 0.5$ with the first principle component and $R^2 = 0.3$ with the latitude for the *indica* landraces). Hence, the simple model yielded overwhelming association peaks across the genome (**Fig. 4e**,**f**). Among the peaks, modest association signals were observed around three known genes controlling heading date[32], but these signals did not stand out on the whole-genome scale (**Supplementary Fig. 24**). Although the compressed MLM approach reduced the number of false positives (**Fig. 4g**,**h**), there was too much structure for it to yield substantial statistical power; essentially, there were no statistical solutions that could detect the quantitative trait loci (QTL) affecting structure by GWAS.

We further inspected the genetic architecture of the 14 agronomic traits. Peak SNPs at the identified loci explained ~36% of the phenotypic variance, on average (from 6% to 68% for different traits; **Fig. 6**), which is much higher than for SNPs in GWAS of human[5,6]. Six of the traits had one or two strong peaks of association with relatively large effects; these were traits for colors, grain quality and grain width (**Fig. 4a**–**d**). We observed that most of the major loci controlling these six traits had causal genes identified previously. Of the six known genes mentioned above, five underlie these traits, and these five show the strongest associations (**Fig. 5** and **Supplementary Note**). For other traits, our results suggest that multiple loci with relatively small effects contribute to the phenotypic variance. The new loci identified here are attractive candidates for follow-up studies that could further our understanding of the genetic architecture of these traits.

DISCUSSION

These studies demonstrate that GWAS of rice landraces can be used for genetic mapping of multiple traits simultaneously at a fine resolution. Furthermore, direct resequencing of rice landraces provides a wealth of sequence polymorphisms and high association resolution in GWAS, despite modest rates of LD decay in rice.

Direct resequencing can also enable the detection of structural variation, which will greatly facilitate follow-up studies to determine functional variation. Future studies could identify structural variation from low-coverage genome sequencing data partly by combining information across landraces whose haplotypes are similar. However, for the comprehensive identification of structural-variation events, it will be more effective to deep sequence and assemble a small number of landraces with maximal genetic diversity. Such an approach will soon be feasible, as second-generation sequencing technology continues to improve in terms of both read length and paired-end insert size.

More information will be gained through GWAS of rice landraces as additional phenotypes are evaluated, especially in different environments, and as a larger number of broadly representative landraces are sampled. Several follow-up steps could be taken to pinpoint candidate genes via application of rice functional-genomics approaches[33]. Moreover, for the clinal adaptive traits (for example, flowering time), association mapping will require biparental populations from specific crosses. Constructing collaborative recombinant-mapping populations selected from the sequenced landraces may help to control for population structure, as well as identifying alleles with small effects or low frequency in the population[7,31,34,35]. Joint mapping with this association panel and multiple biparental crosses is likely to be extremely powerful.

In this study, we chose to conduct GWAS for only the *indica* landrace population because its larger sample size and higher genetic diversity provided sufficient power for association analysis. The smaller population size and low genetic diversity from the *japonica* samples within China would limit the power of GWAS. A worldwide effort to collect rice accessions for whole-genome resequencing and comprehensive phenotyping is under way, and associations from this broader sampling can be investigated in the future. For studies aiming to improve map resolution and new-allele identification through continuous population expansion, genome sequencing is an effective genotyping approach for GWAS because it allows new SNPs to be added and imputation efficiency to be improved even at lower sequence coverage. This study therefore lays the foundation for a long-term collective effort to discover valuable genes and alleles from the world germplasm collection for cultivar improvement.

URLs. Annotation of rice SNPs, http://www.ncgr.ac.cn/RiceHapMap/Download; Rice Haplotype Map Project database, http://www.ncgr.ac.cn/RiceHapMap; RAP2 database, http://rapdb.dna.affrc.go.jp/archive/build4.html; EBI European Nucleotide Archive, ftp://ftp.era.ebi.ac.uk/; SEG-Map pipeline, http://www.ncgr.ac.cn/software/SEG/; IRGSP 4.0, http://rgp.dna.affrc.go.jp/IRGSP/Build4/build4.html; Ssaha2 version 2.3, http://www.sanger.ac.uk/Software/analysis/SSAHA2/.

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 numbers ERP000106 for 517 rice landraces, ERP000235 for *indica* cv. Guangluai-4 and ERP000236 for *japonica* cv. Nongken-58.

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

ACKNOWLEDGMENTS

We thank the China National Rice Research Institute for providing the landrace samples, R.A. Wing for critical reading of the manuscript, P. Hu for helping assay rice grain quality and Z. Ning for assistance with sequence alignment. This work was supported by the Chinese Academy of Sciences (KSCX2-YW-N-024), China's Ministry of Science and Technology (2006AA10A102) and Ministry of Agriculture (2008ZX08009-002) and the National Natural Science Foundation of China (30821004) to B.H.

AUTHOR CONTRIBUTIONS

B.H. conceived the project and its components. J.L., Q.-F.Z., T.S. and B.H. contributed to the original concept of the project. Q.F., D.F., Y.G., L.D., Wenjun Li, Y.L. and Q.W. performed the genome sequencing. X.H., Q.Z., Y.Z., C.Z., T.L., K.L. and T.H. performed GWAS and data analysis. Y.Z., Q.Z., C.Z. and X.H. developed the imputation program for data analyses. X.H., Y.Z. and T.S. performed statistical simulations. Z.Z., M.L., Y.Z. and E.S.B. performed GWAS using the compressed mixed linear model. X.W., C.L., A.W., L.W., T.Z., Y.J., Wei Li, Z.L. and Q.Q. collected samples and performed the phenotyping. Q.Z., T.L., Y.Z. and X.H. prepared figures and tables. X.H., T.S. and B.H. analyzed all the data 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://npg.nature.com/reprintsandpermissions/.

1. Zong, Y. *et al.* Fire and flood management of coastal swamp enabled first rice paddy cultivation in east China. *Nature* **449**, 459–462 (2007).
2. Zhang, D. *et al.* Genetic structure and differentiation of *Oryza sativa* L. in China revealed by microsatellites. *Theor. Appl. Genet.* **119**, 1105–1117 (2009).
3. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
4. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
5. 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).
6. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
7. Nordborg, M. & Weigel, D. Next-generation genetics in plants. *Nature* **456**, 720–723 (2008).
8. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
9. Gore, M.A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
10. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
11. Weigel, D. & Mott, R. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.* **10**, 107 (2009).
12. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis* thaliana. *Science* **317**, 338–342 (2007).
13. McNally, K.L. *et al.* Genome-wide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. USA* **106**, 12273–12278 (2009).
14. Huang, X. *et al.* High-throughput genotyping by whole-genome resequencing. *Genome Res.* **19**, 1068–1076 (2009).
15. Caicedo, A.L. *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756 (2007).
16. Zhu, Q. *et al.* Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol.* **24**, 875–888 (2007).
17. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
18. Mather, K.A. *et al.* The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* **177**, 2223–2232 (2007).
19. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
20. Roberts, A. *et al.* Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* **23**, i401–i407 (2007).
21. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
22. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
23. Saitoh, K. *et al.* Allelic diversification at the C (OsC1) locus of wild and cultivated rice: nucleotide changes associated with phenotypes. *Genetics* **168**, 997–1007 (2004).
24. 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).
25. Cui, J. *et al.* Characterization and fine mapping of the *ibf* mutant in rice. *J. Integr. Plant Biol.* **49**, 678–685 (2007).
26. Gao, Z. *et al.* Map-based cloning of the ALK gene, which controls the gelatinization temperature of rice. *Sci. China C Life Sci.* **46**, 661–668 (2003).
27. Wang, Z.Y. *et al.* The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene. *Plant J.* **7**, 613–622 (1995).
28. 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).
29. Shomura, A. *et al.* Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* **40**, 1023–1028 (2008).
30. 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).
31. Buckler, E.S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009).
32. Kim, S.L. *et al.* OsMADS51 is a short-day flowering promoter that functions upstream of Ehd1, OsMADS14, and Hd3a. *Plant Physiol.* **145**, 1484–1494 (2007).
33. Zhang, Q., Li, J.Y., Xue, Y.B., Han, B. & Deng, X.W. Rice 2020: a call for an international coordinated effort in rice functional genomics. *Mol. Plant* **1**, 715–719 (2008).
34. Myles, S. *et al.* Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21**, 2194–2202 (2009).
35. McMullen, M.D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).

ONLINE METHODS

Sampling. We sampled Chinese rice landraces from a collection of ~50,000 rice accessions preserved at the China National Rice Research Institute in Hangzhou, Zhejiang Province. From the germplasm database records of phenotypic variation and geographic origins, we generated a data matrix and conducted a cluster analysis. On the basis of the resulting cluster tree, we sampled accessions to represent the entire range of phenotypic diversity and geographic distribution of the Chinese rice landraces. Depending on availability, 20–30 seeds of each accession were germinated and the seedlings were planted in the experimental field at the China National Rice Research Institute in Hangzhou for phenotypic evaluation.

DNA isolation and genome sequencing. Total genomic DNA was extracted from leaf tissues using the DNeasy Plant Mini Kit (Qiagen). For each landrace, a single individual was used for genome sequencing on the Illumina Genome Analyzer II. Library construction and sample indexing was done as described[14]. Indexed libraries of five landraces were mixed with an equal molar concentration and loaded on 2% agarose gels. Fragments of 300–400 bp were recovered and purified, and then enriched by nine cycles of PCR. The library was loaded into one lane of the Illumina Genome Analyzer II for 2 × 76 bp paired-end sequencing. Image analysis and base calling were done using the Illumina Genome Analyzer processing pipeline (v1.4). PERL scripts in the SEG-Map pipeline were applied to sort raw sequences on the basis of the 5′ indexes. 73-mer reads were obtained after the three-base indexes were trimmed.

Sequence alignment and genotype calling. The 73-bp paired-end reads were mapped to the rice reference genome (IRGSP 4.0) using the software Ssaha2 version 2.3. Aligned reads were picked up with a cutoff of minimum 96% identity over 92% consecutive nucleotides of a read. Only uniquely aligned reads (reads mapped to unique locations in the reference genome) were retained. These reads were used to call the single–base pair genotypes of the consensus sequences across the whole genome by the Ssaha Pileup package (version 0.5). The low-quality bases (base-quality *Q* score in Phred scale <25) were removed, and those called sites with conflicting genotypes among different reads were further excluded. Additionally, we required that the overall depth in each site be <15 to avoid mapping to regions with copy-number variation. Next, the single–base pair genotypes of 520 rice accessions were integrated together for SNP identification. The detailed procedure is provided in the **Supplementary Note**. The consensus sequences of each line at the SNP sites were further retrieved for genotype calling. Four sets of sequencing data, which included BAC-based Sanger sequencing data[10,36] and high-coverage resequencing data of both *indica* and *japonica*, were used to assess genotyping accuracy (**Supplementary Note**).

Phylogenetic and population genetic analyses. Neighbor-joining trees and principal-component analysis plots were used to infer population structure of the rice landraces. A pairwise distance matrix derived from the simple matching distance for all SNP sites was calculated to construct unweighted neighbor-joining trees using the software PHYLIP version 3.66 (ref. 37). Principal-component analysis was done using the software EIGENSTRAT[17]. To minimize the contribution from regions of extensive strong LD, if a pair of SNPs within the 50-kb region had r^2 greater than 0.8, we removed one of them. The first two principal components were plotted against each other for the *indica* population and the *japonica* population, respectively. LD was calculated using the software Haploview with default settings[38]. Pairwise r^2 was calculated for all SNPs in a 500-kb window and averaged across the whole genome. Sequence diversity (π) was calculated in a 100-kb window as the average number of pairwise difference per site for all pairs of total sampled landraces, all pairs of *indica* landraces or all pairs of *japonica* landraces[39]. The population-differentiation statistics (F_{ST}) were computed as described[40], using a 100-kb window, between the *indica* and *japonica* landraces, among the three subgroups of *indica* and among the three subgroups of *japonica*.

Missing genotype imputation. A data-imputation method based on a KNN algorithm was developed for inferring a large number of missing genotypes generated from low-coverage genome sequencing (**Supplementary Fig. 25**). The imputation is performed in a chromosomal region defined by a given number of SNPs—that is, in a window size of *w* SNPs. The window size is allowed to vary according to the size of chromosomal regions in which LD is reasonably strong. The window then slides along a chromosome at a step size of one SNP until the missing data are inferred for the entire chromosome. The detailed algorithm is provided in the **Supplementary Note**.

SNP sites with too much missing data should be excluded for use in imputation. To ensure imputation quality, SNPs with more than 80% missing data and SNPs with minor allele frequency less than 5% were excluded in this study. This method can be more widely applied when haplotype phasing procedure is incorporated to impute heterozygous genotypes.

The specificity of the genotype data set before and after imputation of missing genotypes was assessed using four sets of sequencing data (**Supplementary Note**). The missing-data rate of the genotype data set was calculated as the average proportion of missing calls of the SNP sites. A detailed list of these assessments is provided in **Supplementary Table 3**.

Genome-wide association analysis. Association analyses were conducted using the simple model and the compressed MLM. The genotype data set for *indica* were generated after imputation of missing genotypes, with a total of 671,355 common SNP sites (minor allele frequency > 0.05 in 373 *indica* lines).

For the simple model analysis, we used the following equation:

$$y = X\alpha + e.$$

For the compressed MLM analysis, we used the equation[21,22]

$$y = X\alpha + P\beta + K\mu + e.$$

In these equations, y represents phenotype, X represents genotype, P is the PCA matrix instead of the Q matrix and K is the relative kinship matrix. $X\alpha$ and $P\beta$ represent fixed effects, and $K\mu$ and e represent random effects. The top five principal components were used to build up the P matrix for population-structure correction. The matrix of simple matching coefficients was used to build up the K matrix, and this step was followed by compression[22]. The analyses were performed using PROC MIXED in SAS (SAS Institute).

Phenotyping. For each landrace, five randomly chosen plants were evaluated and their mean was calculated. Tiller number was evaluated when grains fully ripened. On the main tiller, flag leaf angle was measured.

Grain length and width were measured at the maximal values for each grain using an electronic digital caliper. Grain weight was initially obtained by weighing a total of 200 grains, then converting it to 1,000-grain weight, a scale commonly used for yield evaluation. The total number of spikelets produced per panicle was counted manually.

Amylose content was determined according as described[41]. Milled rice flour (50 mg ± 0.5 mg) was digested with 0.5 ml of 95% (vol/vol) ethanol and 4.5 ml of 1 N NaOH overnight, mixed with 0.2 ml 0.2% (wt/vol) I_2 in 2% (wt/vol) KI solution and diluted with 0.1 ml 1 N acetic acid to 10 ml. The amylose-iodine color was measured at 608 nm using a spectrophotometer (Bausch and Lomb Spectronic 20). Gelatinization temperature was determined by an alkali digestion test[42]. The degree of alkali spreading was measured in 1.7% (wt/vol) KOH solution for 23 h in a 30 °C oven.

Heading date was recorded as the number of days from sowing to the time when inflorescences had emerged above the flag leaf sheath for more than half of the individuals of a landrace. The degree of drought tolerance was scored on the basis of the ratio of the burliness rate of the rice landraces in the dry field to that in the wet field. The degree of seed shattering was scored on a scale of 1–3 (easy, medium and hard) when grains fully ripened.

Software and data release. The SNP data set can be found at the Rice Haplotype Map Project database (http://www.ncgr.ac.cn/RiceHapMap). The program for missing-data imputation, implemented in C, can be freely downloaded from the database website.

36. Huang, X. *et al.* Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physiol.* **148**, 25–40 (2008).

doi:10.1038/ng.695

37. Felsenstein, J. PHYLIP: phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
38. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
39. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
40. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196 (2005).
41. Juliano, B. *Rice Chemistry and Technology* 443–513 (American Association of Cereal Chemists, Saint Paul, Minnesota, USA, 1985).
42. Little, R.R., Hilder, G.B. & Dawson, E.H. Differential effect of dilute alkali on 25 varieties of milled white rice. *Cereal Chem.* **35**, 111–126 (1958).