nature biotechnology

RESOURCE

Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes

Xun Xu[1–3,12], Xin Liu[2,12], Song Ge[4,12], Jeffrey D Jensen[5,12], Fengyi Hu[6,12], Xin Li[1,12], Yang Dong[1,12], Ryan N Gutenkunst[7], Lin Fang[2], Lei Huang[3,4], Jingxiang Li[2], Weiming He[2,8], Guojie Zhang[1,2,4], Xiaoming Zheng[3,4], Fumin Zhang[3], Yingrui Li[2], Chang Yu[2], Karsten Kristiansen[2,9], Xiuqing Zhang[2], Jian Wang[2], Mark Wright[10], Susan McCouch[10], Rasmus Nielsen[1,9,11], Jun Wang[2,9] & Wen Wang[1]

Rice is a staple crop that has undergone substantial phenotypic and physiological changes during domestication. Here we resequenced the genomes of 40 cultivated accessions selected from the major groups of rice and 10 accessions of their wild progenitors (*Oryza rufipogon* and *Oryza nivara*) to >15 × raw data coverage. We investigated genome-wide variation patterns in rice and obtained 6.5 million high-quality single nucleotide polymorphisms (SNPs) after excluding sites with missing data in any accession. Using these population SNP data, we identified thousands of genes with significantly lower diversity in cultivated but not wild rice, which represent candidate regions selected during domestication. Some of these variants are associated with important biological features, whereas others have yet to be functionally characterized. The molecular markers we have identified should be valuable for breeding and for identifying agronomically important genes in rice.

© 2012 Nature America, Inc. All rights reserved.

Asian cultivated rice (*Oryza sativa*) is thought to have been domesticated from divergent populations of Asian wild rice, *O. rufipogon* and *O. nivara*, >10,000 years ago[1,2]. During domestication, rice has undergone significant phenotypic changes in grain size, color, shattering, seed dormancy and tillering. For decades, geneticists have used quantitative trait locus mapping to localize the major causative genes responsible for these traits, yielding a dozen trait-related genes in cultivated rice (for example, *sh4*, *rc* and *prog1*)[3–6]. Additionally, a recent genome-wide association study using genome-wide SNP data for 517 Chinese landraces identified loci that may be associated with 14 agronomic traits[7]. However, quantitative trait locus and gene mapping is labor intensive and time consuming, taking years to construct segregating populations and requiring intensive phenotyping and genotyping. Association mapping is also prone to missing excellent alleles because the favorable alleles tend to be rare and are difficult to detect during regular association analyses[8]. A more recent report tried to identify artificially selected genes[9], but the strategy of pooling many accessions (a strain identified by an International Rice Research Institute (IRRI) accession number) together and using shallow sequencing coverage provided limited variation data for rice. If a comprehensive catalog of genome variation in both cultivated and wild rice were available, it would greatly facilitate the identification of functional variations in elite varieties by comparing genomic variation in an elite variety with data from controls. Dense variation data will also be useful for marker-assisted breeding and gene mapping of rice.

RESULTS

Sequencing and mapping

Cultivated rice is classified into two major subspecies of *O. sativa* (*indica* and *japonica*) and is further subdivided into genetically differentiated groups, including Glaszmann's six groups (I to VI)[10] and Garris *et al.*'s five groups (*indica*, *aus*, *aromatic*, *temperate japonica* and *tropical japonica*)[11]. We selected 40 cultivated rice accessions to represent all of the major groups of Asian cultivated rice (**Supplementary Table 1**), including 11 *tropical japonica* (TRJ), 8 *temperate japonica* (TEJ) and 6 *aromatic* (ARO) that belong to *japonica* rice, and 4 *aus* (AUS) and 9 *indica* (IND) that belong to *indica* rice (**Supplementary Table 1**). In addition, we sampled one accession each from groups III and IV, proposed by Glaszmann[10], which were not included in a previous population study[11]. Among these cultivars, 29 are considered to be landraces and 11 are improved varieties. To strictly control the quality of our sequencing and SNP calling, we also included the Nipponbare strain, which was used to generate the reference rice genome sequence[12]. For wild rice samples,

[1]CAS-Max Planck Junior Research Group on Evolutionary Genomics, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (CAS), Kunming, China. [2]BGI-Shenzhen, Shenzhen, China. [3]Graduate University of Chinese Academy Sciences, Beijing, China. [4]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China. [5]School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [6]Food Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China. [7]Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona, USA. [8]South China University of Technology, Guangdong, China. [9]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [10]Department of Plant Breeding & Genetics, Cornell University, Ithaca, New York, USA. [11]Departments of Integrative Biology and Statistics, University of California, Berkeley, USA. [12]These authors contributed equally to this work. Correspondence should be addressed to W.W. (wwang@mail.kiz.ac.cn) or J.W. (wangj@genomics.org.cn) or R.N. (rasmus_nielsen@berkeley.edu).

Received 3 June; accepted 25 October; published online 11 December 2011; doi:10.1038/nbt.2050

Table 1 Summary of sequencing and variations for cultivated and wild rice

Group	Sample size	Raw data (Gbp)	Raw data depth	Uniquely mapping bases (Gbp)	Mean depth	Mean depth in gene region	SNP (M)	Indel (K)	SV (K)[a]
Total	50	319.2	18.7	182.4	11.8	14.3	6.5	808	94.7
Cultivar total	40[b]	249.8	18.7	163.0	12.2	14.5	4.4	612	62.1
Indica	12	82.5	18.5	47.4	10.6	13.3	3.0	441	38.7
AUS	2	13.4	18.0	7.9	10.6	13.0	0.8	183	17.4
IND	10	69.1	18.6	39.5	10.5	13.4	2.9	414	24.3
Japonica	24	167.3	18.7	115.6	12.9	15.1	2.5	355	28.5
ARO	6	45.6	20.4	29.7	13.3	15.9	1.1	183	15.3
TEJ	8	53.2	17.9	39.5	13.3	15.2	1.1	136	4.3
TRJ	10	68.5	18.4	46.4	12.5	14.6	1.5	208	11.1
Wild total	10	69.4	18.7	39.4	10.6	13.3	5.2	682	40.4
O. rufipogon	5	34.0	18.3	19.2	10.3	12.9	3.5	424	21.1
O. nivara	5	35.4	19.1	20.2	10.9	13.7	3.1	439	21.7

[a]Structural variation (insertion or deletion longer than 100 bp) relative to Nipponbare; the structural variations in the same region in different accessions count as one. [b]Only 36 cultivar accessions can be clearly put in a subgroup, and the other four accessions have admixed genetic backgrounds, as indicated in **Supplementary Table 1**.

five accessions each from *O. rufipogon* and *O. nivara* were collected according to the geographic distribution of wild rice (**Supplementary Table 1** and **Supplementary Fig. 1**).

We sequenced these accessions to >15 × coverage (raw data) using Illumina GA2 instruments. The sequencing quality of these raw reads was generally high (90% with Phred quality score > 27) (**Supplementary Figs. 2** and **3**). We first mapped the short reads back to the IRGSP v4 rice genome[13] by SOAP2 (version 2.20)[14]. The mapping rate in different accessions varied from 79% to 94%, and the final effective mapping depth was >10 × for most accessions (**Table 1** and **Supplementary Table 1**). *Japonica* accessions had the highest mapping rate, whereas the two species of wild rice had relatively low mapping rates. The differences in mapping rates may be due to divergence between the sequenced accessions and the reference. Contamination during sample preparation or sequencing may also cause low mapping rates. However, we found that <10% of unmapped reads can be mapped to sequences from other species (mostly from *Xanthomonas campestris* and *Escherichia albertii*) that might be contamination. Thus, most unmapped reads were either from individual-specific or diverged sequences or from reads with serious sequencing errors.

We assembled unmapped reads for each accession into contigs (**Supplementary Notes**, **Supplementary Table 2** and **Supplementary Data Set 1**), and then used *de novo* gene prediction to annotate a total of 2,031 genes in the contigs, of which 1,552 had homologs in the NCBI nonredundant (nr) database and 1,415 of these had homologs in plants (**Supplementary Table 3**). The average length of these 1,415 genes was substantially shorter than that of the whole genome (957 bp versus 2,300 bp), indicating that many of them are not intact genes and some may be pseudogenes. We chose 17 such 'novel' genes from the wild rice Yuan 3-9 accession for PCR validation (**Supplementary Notes**). Twelve of them could be amplified in Yuan 3-9 but not in Nipponbare, and five could be amplified in both Yuan 3-9 and Nipponbare (**Supplementary Fig. 4**), indicating some of these unmapped genes may be found only in a certain accession but some could represent diverged homologous genes. We were able to detect transcripts for ten of the twelve Yuan 3-9–specific genes using RT-PCR of pooled RNAs extracted from root, shoot and leaf of Yuan 3-9, suggesting that some of these novel genes may be bona fide genes or expressing pseudogenes. We further annotated protein domains in the 1,415 genes by InterProScan[15]. Sixty percent of them could be functionally annotated (**Supplementary Table 3**). Of the 1,415 novel genes that have homologs in plants, 685 were found only in one accession, 1,282 were found in <5 accessions and 319 were found only in wild rice. The most common novel gene, which had an unknown function, was found in 34 accessions.

In addition to novel genes, we also tried to identify genes absent in some accessions. By mapping reads to the reference genome, we found that the genome coverage varied from 84% to 95%. Of the regions that were not covered in at least one accession, 51% were repeat regions (**Supplementary Fig. 5**), and those unmapped regions had substantially higher GC ratios than the whole genome average (75.1% versus 43.5%). In different accessions, there were some genes with unexpected low coverage, whereas in the Nipponbare individual that we sequenced, we did not observe any gene with coverage <10% of the gene length. Thus, we used this criterion to identify 1,327 possible gene loss events, defined as genes with <10% coverage in one or more accessions but >90% coverage in the Nipponbare.

To collect more evidence supporting these gene loss events, we used the paired-end information of reads. Structure variations, including deletions, would result in discordant paired-end reads mapping in the corresponding region and thus can be identified by comparing the insert length of paired-end mapped reads to the experimental insert size[16–18]. We observed that 839 of those 1,327 possible lost genes had such discordant paired-end reads across the low-coverage region (**Supplementary Fig. 6** and **Supplementary Table 4**) and thus were retained in the final gene loss data set. Validation of gene loss events by PCR suggests a low false-positive rate in the final data set. Of nine randomly chosen gene loss events, eight could be validated (**Supplementary Notes**). These lost genes may be responsible for heterosis and thus may be important in breeding programs[19–21]. Forty-nine percent of these genes had no functional or annotation information, and 16% were nonprotein coding genes (**Supplementary Fig. 7**). Overall the lost genes were significantly enriched in the distal regions of chromosomes (Wilcoxon rank sum test, $P = 0.01$, **Supplementary Fig. 8**).

Variation across the rice genome

Using a strict pipeline (**Supplementary Notes**), we identified ~15 million candidate SNPs in all 50 accessions (**Supplementary Data Sets 2** and **3**). To obtain SNPs for population analyses, we excluded SNPs with missing data in any of the 50 accessions, as these will make subsequent inferences unreliable, yielding a final total of 6,496,456 high-quality SNPs (**Table 1**). To our knowledge, this represents the largest high-quality SNP data set obtained in rice. The data may be used to identify important rice genes by serving as molecular markers for designing rice SNP arrays and for breeding. Indeed, by using this data set as a control to represent background genetic variation in rice, we were recently able to identify many tag functional SNPs in elite rice varieties, and our extensive study of one of them has revealed that it functions in the adaptability of upland rice by regulating abscisic acid synthesis (J. Lv, F. Hu, W. Wang *et al.*, unpublished data).

© 2012 Nature America, Inc. All rights reserved.

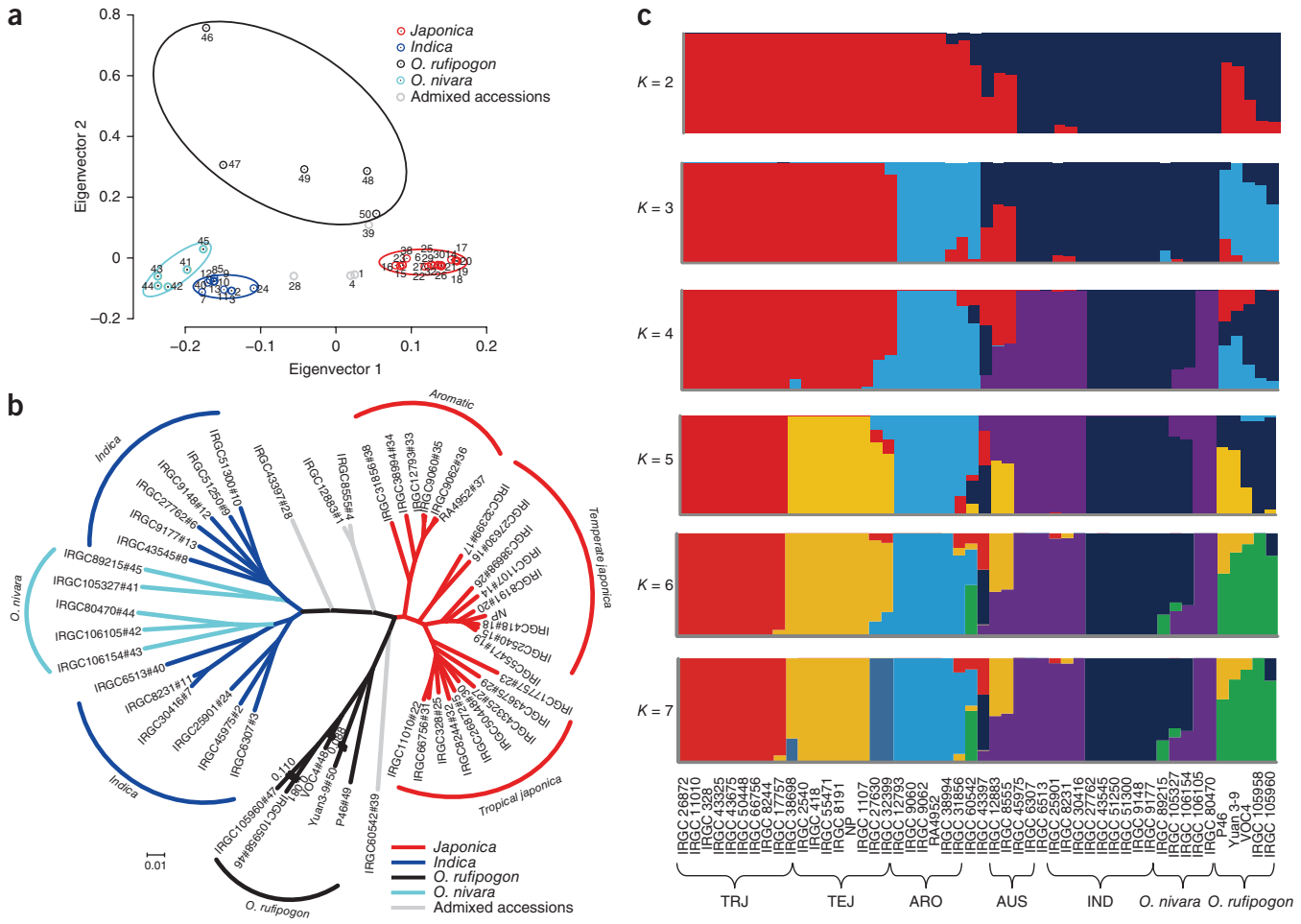

Figure 1 Population structure of Asian rice. (**a**) PCA using all identified SNPs as markers. Most *indica*, *japonica*, *O. nivara* and *O. rufipogon* accessions cluster together, whereas four accessions (IRGC 12883, 8555, 43397 and 60542, marked as 1, 4, 28 and 39) that are located between groups can be explained by admixture and are marked as gray dots. The numbers by each dots are index numbers to the International Rice Research Institute (IRRI) accession numbers in **Supplementary Table 1**. (**b**) Neighbor-joining phylogenetic tree based on all SNPs, with the evolutionary distances measured by *p*-distance with PHYLIP[51]. (**c**) Population structure analysis using FRAPPE[35]. Each color represents one population. Each accession is represented by a vertical bar, and the length of each colored segment in each vertical bar represents the proportion contributed by ancestral populations.

© 2012 Nature America, Inc. All rights reserved.

Of the 6.5 million high-quality SNPs, most (82%) were located in intergenic regions, and only 3.6% were located in coding sequence regions (**Supplementary Table 5**). Among the latter, there were 103,321 synonymous and 132,819 nonsynonymous SNPs. Thus, the ratio of nonsynonymous to synonymous substitutions was 1.29, which is consistent with previous work[22]. This ratio is higher than that of *Arabidopsis* (0.83)[23] but lower than that of soybean (1.61)[24]. We tried to identify gene families with ratios that deviated significantly from the whole genome average (**Supplementary Fig. 9**). Gene families with essential functions (for example, the ubiquitin family and cellulose synthase family) tended to have substantially lower nonsynonymous-to-synonymous substitution ratios, whereas gene families that function in regulatory processes and signal recognition, such as the disease resistance family, had higher ratios.

In addition to SNPs, we also detected 808,000 small insertions and deletions (indels, 1–5 bp) by mapping reads with gaps allowed. We found nearly equal numbers of insertions and deletions. Similar to trends observed for SNPs, rare variants comprised a large proportion of total indels, with ~67% of indels found in <5 accessions (**Supplementary Fig. 10**). Most of the indels were located in intergenic regions, and ~1% (8,232) were located in coding sequences, among which 40% were in-frame, 3-bp indels (**Supplementary Table 6** and **Supplementary Fig. 11**). The 5,161 out-of-frame indels might have generated pseudogenes in different accessions.

Next we applied an assembly-based method to identify larger structural variations. To improve assembly quality, we pooled individuals from the same subgroups, which yielded >80 × raw coverage of the reference rice genome. We identified 94,700 structural variations >100 bp in length (**Table 1**). Based on coverage depth, we also identified 1,676 copy number variations (CNVs) having more copies than the reference genome (**Supplementary Table 7**). Twenty-one percent of the CNVs occurred in more than five individuals. Eight-hundred sixty-five CNVs corresponded to gene regions (**Supplementary Table 8**), of which 338 (39%) were hypothetical or functional unknown genes, and 66 (8%) were nonprotein coding transcripts. Among the genes with

Table 2 Diversity levels of different populations and subpopulations

	Cultivated rice			Wild rice		
Statistic	Combined	*Indica*	*Japonica*	Combined	*O. rufipogon*	*O. nivara*
π per kb	5.4	5.7	3.7	7.7	7.2	6.3
θ_w per kb	6.6	6.8	5.5	11.5	10.6	9.4

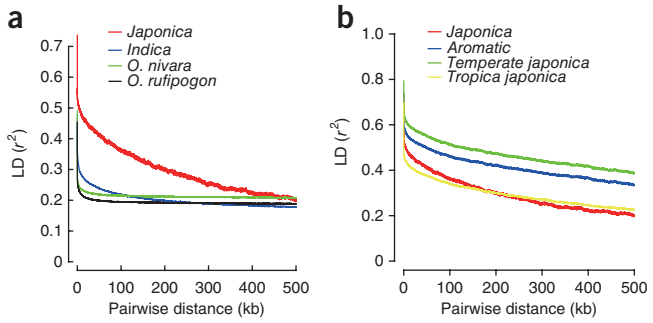

Figure 2 Linkage disequilibrium differences between wild and cultivated rice groups. (**a**) Linkage disequilibrium (LD) decays quickly within 10 kb for *indica*, *O. nivara* and *O. rufipogon*, whereas it extends to 50 kb in *japonica*. (**b**) Different *japonica* subgroups have similar linkage disequilibrium decay patterns, indicating that the overall long linkage disequilibrium in *japonica* is not caused by population substructure. Four accessions with mixed genomic backgrounds are removed from the cultivated population for all analyses in this figure.

functional annotation, many were disease resistance genes (14 are annotated as 'disease resistance protein', four are 'leucine rich repeat containing protein' and three are 'NB-ARC domain containing').

Population structure of cultivated and wild rice

Among the 6.5 million high-quality SNPs, 4,124,470 were found in cultivars. A large proportion (2,953,712; 71.6%) of these SNPs were also found in wild rice accessions, indicating that most genetic variation in cultivated rice is derived from the variation in wild rice. Of the remaining 1,170,758 (28.4%) cultivar-specific SNPs, some might have been false positives owing to the relatively small sample size of wild rice. Most of these cultivar-specific SNPs (720,289, 61.5%) occurred at low frequency (≤4 in the total 72 chromosomes of 36 cultivar accessions, after excluding four admixed accessions (see below)). In contrast, although we only sequenced ten wild accessions, we identified 5,241,380 high-quality SNPs in wild rice. This is consistent with previous studies that showed that wild rice has a much more diverse gene pool than cultivars and thus may contain useful genetic resources for rice improvement[25].

For cultivated and wild rice, we calculated π and θ_w values, two common summary statistics for measuring genetic diversity in a population[26]. The estimated diversity levels (**Table 2**) were roughly twice those found in a previous study[27] of 111 randomly chosen sequence tag sites. However, when we restricted our analysis to regions corresponding to those 111 sites, we obtained diversity levels similar to those found previously[27] (**Supplementary Table 9**). The π values in our *indica* and *japonica* samples were also higher than those observed in 517 Chinese *indica* and *japonica* landraces[7], indicating that worldwide cultivated rice contains more genetic diversity than Chinese landraces. The *japonica* varieties had 49% less diversity than *O. rufipogon*, whereas *indica* had only 9% less diversity than *O. nivara*, although more *japonica* accessions were sampled in our study (**Table 2**). This suggests that *japonica* rice might have undergone a stronger reduction in effective population size than *indica*, possibly owing to a stronger bottleneck during domestication, which has been proposed previously[7,28–30].

To examine genetic population structure and relationships among the major groups of Asian cultivated rice, we conducted principle component analysis (PCA)[31,32] and constructed a neighbor-joining tree[33,34], based on the 6.5 million high-quality SNPs. In the PCA, most of the samples could be divided using the first and second eigenvectors into *indica*, *japonica*, *O. nivara* and *O. rufipogon* groups, with *O. rufipogon* being more dispersed, indicating higher diversity (**Fig. 1a**). In our samples, *indica* was very closely related to *O. nivara*, whereas *japonica* was closer to *O. rufipogon* and farther from *O. nivara*. We used a neighbor-joining tree to cluster accessions based on average genetic distances (**Fig. 1b**). The neighbor-joining tree contained three major groups, corresponding to *O. rufipogon*, *japonica* and *O. nivara+indica,* with a further subdivision of *japonica* into *temperate*, *tropical* and *aromatic* varieties (**Fig. 1b**). The small number of *aus* samples precluded any conclusion with respect to the subdivision of *indica* into *indica* and *aus* varieties. Furthermore, we used the program FRAPPE, which estimates individual ancestry and admixture proportions assuming K populations exist based on a maximum likelihood method[35], to investigate population structure. We analyzed the data by increasing K (the number of populations) from 2 to 7 (**Fig. 1c**). For $K = 2$, we found a division between *O. rufipogon*/*japonica* and *O. nivara*/*indica*. When $K = 4$, we saw a new subgroup (*aromatic*) within the *japonica* group (**Fig. 1c**). When $K = 5$, *tropical japonica* and *temperate japonica* were separated. These results all support the hypothesis that the two cultivated rice subspecies might have been domesticated independently of different populations of wild rice[28–30]—for instance, the population that gave rise to *indica* may have evolved from an ancestral population closely related to our *O. nivara* samples, and the ancestral population that gave rise to *japonica* may have been more closely related to our *O. rufipogon* samples. We also used the indels we identified to construct a phylogenetic tree. The grouping patterns among wild and cultivated rice largely remain the same (data not shown).

To further resolve the evolutionary history of rice, we sequenced ten more *O. rufipogon* and five more *O. nivara* samples at a lower ~3 × coverage (**Supplementary Notes**) to achieve a better geographical resolution of these two wild species (**Supplementary Fig. 1**). We called SNPs for each of these 15 accessions using SOAPsnp (version 1.01)[36] and only retained SNP sites that were covered by at least 6 out of the 15 wild accessions to get informative sites for phylogenetic analyses. By intersecting these SNPs with the set of 6.5 million high-quality SNPs identified in the previous 50 accessions, we obtained 3,668,781 SNPs that could be used to analyze all 40 cultivated and 25 wild rice accessions. PCA, neighbor-joining tree and FRAPPE analyses again showed that *indica* rice is closely related to *O. nivara* (**Supplementary Fig. 12**), indicating a complex evolutionary history between *indica* and *O. nivara*. More interestingly, the tree showed that all *japonica* rices were closer to the five Chinese *O. rufipogon* accessions, especially Dongxiang wild rice from the low Yangtze region, strongly supporting the conjecture that *japonica* might have been independently domesticated from a Chinese *O. rufipogon* population in the Yangzte region[37].

It is noteworthy that a few of the accessions occurred at unexpected positions in both the PCA diagrams and the neighbor-joining trees (**Fig. 1** and **Supplementary Fig. 12**), including International Rice Genebank collection (IRGC) 12883 (no. 1) and IRGC 8555 (no. 4), which were reported to be *aus* (*indica*), and IRGC 43397 (no. 28), which was treated as *tropical japonica* (*japonica*) in previous studies[11], as well as IRGC 60542 (no. 39), which was considered to be in Group IV (*japonica*)[10]. These samples are most likely from accessions with admixed genetic backgrounds among (sub)species and were thus excluded in the following analysis. In addition, previous studies reported that accessions IRGC 26872, 2540 and 32399 were *indica* and that accessions 25901 and 27762 were *japonica*[11], but our whole genome sequence data clustered them into opposite subspecies (**Fig. 1** and **Supplementary Fig. 12**). Our further phenotyping results (data not shown) confirmed that these samples should belong to the cultivar groups that our genomic data suggested, and they were thus reclassified into appropriate cultivar groups in this study. Misidentification of accessions has been reported previously for rice[29].

© 2012 Nature America, Inc. All rights reserved.

To estimate the linkage disequilibrium patterns in different rice groups, we calculated r^2 (ref. 38) between pairs of SNPs using Haploview[39]. Linkage disequilibrium decayed to its half-maximum within <10 kb for *O. rufipogon* and *O. nivara*, 65 kb for *indica* and 200 kb for *japonica* (**Fig. 2a**). This linkage disequilibrium decay pattern again supports the hypothesis of a stronger bottleneck in *japonica* during domestication than that in *indica*. For subpopulations within *japonica*, linkage disequilibrium was also high (**Fig. 2b**), with the half-maximum at ~300 kb, 300 kb and 180 kb for *aromatic*, *temperate japonica* and *tropical japonica*, respectively, which suggests that the high linkage disequilibrium in *japonica* cannot be attributed simply to population substructure. The fine linkage disequilibrium in each rice group will be useful for mapping rice genes.

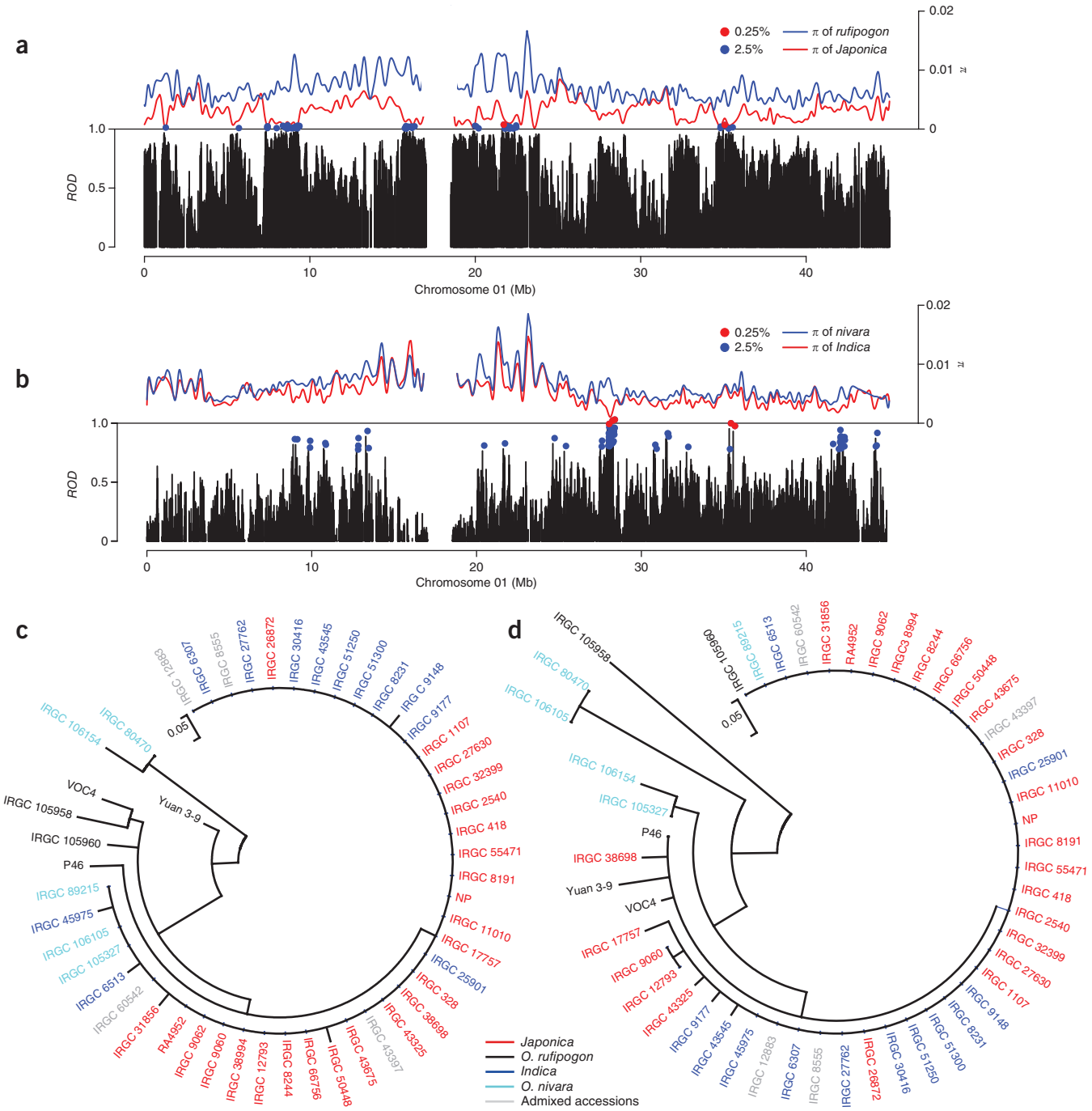

Figure 3 Significant outlier regions (genes) in *ROD* distribution. (**a**) *ROD* for *japonica* relative to *O. rufipogon* across chromosome 1. For other chromosomes see **Supplementary Figure 13**. Regions with a 2.5% significance level of *ROD* are shown in blue; regions with a 0.25% level are shown in red. (**b**) *ROD* for *indica* relative to *O. nivara* across chromosome 1. For other chromosomes see **Supplementary Figure 14**. Regions with a 2.5% significance level of *ROD* are shown in blue; regions with a 0.25% level are shown in red. (**c**,**d**) Gene trees for *prog1* and Os09g0547100, respectively. The tree topologies depart dramatically from the whole genome phylogenetic tree (**Fig. 1b**), and the cultivated rice accessions almost share a single allele, probably because of a selective sweep.

© 2012 Nature America, Inc. All rights reserved.

Regions (genes) affected by artificial selection

Phenotypic traits that were favorably selected by humans to enhance agricultural characteristics usually have low levels of variation and skewed allele frequency spectra[40], parameters that have been successfully used to identify putative artificially selected genes in maize[41–44], cattle[45], silkworms[46] and chickens[47]. Our large SNP data set from both wild and cultivated rice provides an opportunity to identify selected genes by comparing polymorphism levels in cultivated and wild species.

To detect selective sweeps driven by artificial selection, we sought to identify regions with significantly lower levels of polymorphisms in cultivated rice compared with wild rice. We calculated the reduction of diversity ($ROD = 1 - \pi_{cul}/\pi_{wild}$), based on the ratio of diversity in cultivated rice to the diversity in wild rice (π_{cul}/π_{wild}), in nonoverlapping windows of 10 kb along the entire genome for *japonica* relative to *O. rufipogon* and for *indica* relative to *O. nivara*. These calculations are reasonable because *japonica* was clearly domesticated from *O. rufipogon* (**Fig. 1b** and **Supplementary Fig. 12**), and although the evolutionary relationship between *indica* and *O. nivara* might be complex (as discussed above), we had no other choice but to use *O. nivara* as the ancestor of *indica*. We identified regions in the 2.5% or 0.25% right tails of the empirical *ROD* distribution. These regions of lost diversity in cultivars might have experienced ultivar-specific selective sweeps. Quite often, a selective sweep results in a long linkage disequilibrium fragment, and indeed, some candidate 10-kb windows cluster together. An extreme case was observed in *indica* with 2.5% *ROD* cutoff, in which the longest fragment extended to 240 kb. We considered all genes in these regions to be candidate artificially selected genes. When a gene crosses two windows, we counted it as one. It is likely that many of these genes were not themselves subjected to selection but rather have hitchhiked along with the actual gene targeted by artificial selection. To identify actual selected genes, more analyses, including transgenic experiments, are needed.

In *japonica*, we found 739 and 64 10-kb nonoverlapping windows (containing 1,322 and 105 genes), based on 2.5% and 0.25% *ROD* cutoffs, respectively. In *indica*, we found 750 and 75 10-kb nonoverlapping windows, containing 1,265 and 129 genes, respectively (**Fig. 3**, **Supplementary Figs. 13**, **14** and **Supplementary Table 10**). As the genomes of cultivars have been shown to contain long regions of linkage disequilibrium, we also used a 100-kb sliding window to identify candidate regions under selection. We found that all 100-kb windows with significant *ROD* values overlapped with one or more 10-kb regions with significant *ROD* values, whereas only half of the regions identified using 10-kb windows could be identified using 100-kb windows (**Supplementary Notes** and **Supplementary Fig. 15**). This indicates that the 10-kb window scanning was more informative for identifying smaller selected regions, especially for *indica,* which had a lower linkage disequilibrium level as described above.

We further calculated the divergence index F_{ST}[48] using the same nonoverlapping 10-kb window approach, and we found that more than half of the windows in the 2.5% *ROD* tails and >90% of the windows in the 0.25% *ROD* tails were also found in the extreme 2.5% right tails of the F_{ST} distribution (**Supplementary Fig. 16**). About 98% of *indica* and 90% of *japonica* 2.5% *ROD* regions fell in the 5% right tail of F_{ST} distributions.

To identify candidate genes that underwent sweeps in both *japonica* and *indica*, we examined the overlapped regions from both the *japonica* and *indica ROD* distribution 2.5% right tails, yielding 73 genes. These genes could be domestication genes that were swept in both *japonica* and *indica*. Or they could be *indica*-specific or *japonica*-specific selective sweep genes that were possibly independently selected in either subspecies and may be related to morphological and physiological differences between *japonica* and *indica*. Two well-known rice domestication genes, *prog1* (refs. 5,6) and *sh4* (ref. 3), were successfully identified in our putative artificial selection gene set. The region embedding *prog1* showed a very strong selective sweep signal in both cultivar subspecies ($ROD_{indica_nivara} = 0.97$, $ROD_{japonica_rufipogon} = 0.96$, $F_{ST\ indica_nivara} = 0.80$, $F_{ST\ japonica_rufipogon} = 0.91$). The gene tree for *prog1* was also indicative of a selective sweep, with star-like branches in the cultivars but long branches in wild rice (**Fig. 3c**). This gene is responsible for the evolution of erect growth in cultivated rice and is thus a key domestication gene[5,6]. The shattering gene *sh4* (ref. 3) was also selected by humans, although the selection signal was slightly weaker ($ROD_{indica_nivara} = 0.93$, $ROD_{japonica_rufipogon} = 0.89$, $F_{ST\ indica_nivara} = 0.77$, $F_{ST\ japonica_rufipogon} = 0.84$) than *prog1*'s.

To assess possible gene functions targeted by artificial selection in rice domestication or improvement, we used gene family information (from the RAP-DB database of the IRGSP version 4.0, release 2) to annotate these candidate genes. Gene families related to morphology, growth and transcriptional regulation were enriched in the candidate genes (**Supplementary Table 11**), including seven belonging to the auxin-responsive SAUR protein family (Fisher's exact test, $P < 0.01$). This family plays important roles in flowering, plant growth and regulation of plant architecture in a tissue-specific or developmental stage-specific manner[49,50]. This family was also reported to be enriched in maize domestication genes by two previous studies[42,44], suggesting important and general roles of these genes in crop domestication and improvement.

Some of these candidate artificial selection genes have not been functionally annotated yet. One such gene (Os09g0547100), located on chromosome 9, showed a very strong selective sweep signal ($ROD_{indica_nivara} = 0.96$, $ROD_{japonica_rufipogon} = 0.91$, $F_{ST\ indica_nivara} = 0.86$, $F_{ST\ japonica_rufipogon} = 0.84$) and a *prog1*-like gene tree in most cultivars (**Fig. 3d**), except for three *indica* accessions (IRGC 51300, 9148 and 51250) that are clustered with wild rice. Another such gene (Os10g0124100) showed very strong evidence for a selective sweep only in *japonica* ($ROD_{japonica_rufipogon} = 0.96$, $F_{ST\ japonica_rufipogon} = 0.97$; $ROD_{indica_nivara} = 0.21$, $F_{ST\ indica_nivara} = 0.19$). Strikingly, there were 54 fixed SNPs in the coding region of this gene in *japonica* relative to *O. rufipogon*. There is no functional annotation information for this gene yet. All of these functionally uncharacterized or unknown candidate artificial selection genes provide useful guidance for rapidly identifying genes with agronomic significance in rice.

DISCUSSION

In this study, we provide a large genome variation data set for wild and cultivated rice. Millions of SNPs in representative wild and cultivated rice strains provided an unprecedented opportunity to finely resolve the domestication history of cultivated rice. Population structure and phylogenetic analyses not only support the hypothesis that *japonica* and *indica* were independently domesticated, but also further suggest *japonica* was domesticated from the Chinese strain of *O. rufipogon*. We identified thousands of candidate genes that may have been artificially selected during the domestication of one or both of the two cultivated subspecies. The SNPs will be useful as dense markers of genome variation for marker-assisted mapping of important rice traits as well as for rice breeding, and the candidate genes selected during domestication may be agronomically important. The data generated in this study provide a valuable resource for rice improvement.

URLs. Rice reference genome (IRGSP build 4) (IRGSP, 2005) and annotation files were downloaded from RAP-DB (http://rapdb.dna.affrc.go.jp/). IRRI database, http://www.iris.irri.org/germplasm/. SOAP packages, http://soap.genomics.org.cn/.

© 2012 Nature America, Inc. All rights reserved.

METHODS

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

Data access. The sequence data have been deposited into NCBI Short Read Archive under accession number SRA023116. Assembly sequence and novel genes are included in **Supplementary Data Set 1** and can also be downloaded at ftp://rice:ricedownload@public.genomics.org.cn/BGI/rice. The whole genome SNP data set (6.5M SNPs) has been deposited into NCBI dbSNP, with submission number records from ss256302601 to ss26799056. The total SNP data set (15M) can be found in **Supplementary Data Sets 2** and **3** and ftp://rice:ricedownload@public.genomics.org.cn/BGI/rice. The indel and structural variation data set can be found in **Supplementary Data Set 4** and can also be found in ftp://rice:ricedownload@public.genomics.org.cn/BGI/rice.

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

ACKNOWLEDGMENTS

We thank C.-H. Shi (Zhejiang University, China) and X.-H. Wei (China National Rice Research Institute) for assistance in growing rice materials. We are grateful to the International Rice Research Institute (Los Banos, Philippines) for providing most seed samples. This work was supported by the Chinese 973 program (2007CB815700), the National Natural Science Foundation of China (30990242), the Provincial Key Grant of Yunnan Province (2008CC017; 2008GA002), the Shenzhen Municipal Government and the Yantian District local government of Shenzhen, the Ole Rømer grant from the Danish Natural Science Research Council, and a CAS-Max Planck Society Fellowship and the 100 talent program of CAS to W.W., J.W. and S.G. We also acknowledge funding support from the Chinese Ministry of Agriculture (948 program), the Shenzhen Municipal Government of China and grants from Shenzhen Bureau of Science Technology & Information, China (ZYC200903240077A; CXB200903110066A).

AUTHOR CONTRIBUTIONS

W.W., Jun Wang, S.G., X.X., R.N. and F.H. designed the project. X.X., X. Liu, X. Li, J.D.J., M.W., L.F., G.Z., W.H., X. Zheng., Y.L. and R.N. analyzed the data. W.W., X.X., R.N., R.N.G., X. Liu, S.M., K.K. and Jun Wang wrote the manuscript. S.G., F.H., L.H. and F.Z. prepared the samples. X. Zhang, Jian Wang, C.Y., J.L. and Y.D. conducted the experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Kovach, M.J., Sweeney, M.T. & McCouch, S.R. New insights into the history of rice domestication. *Trends Genet.* **23**, 578–587 (2007).
2. Sang, T. & Ge, S. Genetics and phylogenetics of rice domestication. *Curr. Opin. Genet. Dev.* **17**, 533–538 (2007).
3. Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
4. 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).
5. Jin, J. *et al.* Genetic control of rice plant architecture under domestication. *Nat. Genet.* **40**, 1365–1369 (2008).
6. Tan, L. *et al.* Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* **40**, 1360–1364 (2008).
7. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
8. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
9. He, Z. *et al.* Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genet.* **7**, e1002100 (2011).
10. Glaszmann, J.C. Isozymes and classification of Asian rice varieties. *Theor. Appl. Genet.* **74**, 21–30 (1987).
11. Garris, A.J., Tai, T.H., Coburn, J., Kresovich, S. & McCouch, S. Genetic structure and diversity in *Oryza sativa* L. *Genetics* **169**, 1631–1638 (2005).
12. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
13. Goff, S.A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* **296**, 92–100 (2002).
14. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
15. Zdobnov, E.M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
16. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
17. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
18. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
19. Charlesworth, D. & Willis, J.H. The genetics of inbreeding depression. *Nat. Rev. Genet.* **10**, 783–796 (2009).
20. Fu, H. & Dooner, H.K. Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA* **99**, 9573–9578 (2002).
21. Springer, N.M. & Stupar, R.M. Allelic variation and heterosis in maize: how do two halves make more than a whole? *Genome Res.* **17**, 264–275 (2007).
22. McNally, K.L. *et al.* Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. USA* **106**, 12273–12278 (2009).
23. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
24. Lam, H.M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059 (2010).
25. Tanksley, S.D. & McCouch, S.R. Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277**, 1063–1066 (1997).
26. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
27. Caicedo, A.L. *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, e163 (2007).
28. Second, G. Origin of the genic diversity of cultivated rice (*Oryza* spp.): study of the polymorphism scored at 40 isozyme loci. *Jpn. J. Genet.* **57**, 25–57 (1982).
29. Zhu, Q. & Ge, S. Phylogenetic relationships among A-genome species of the genus Oryza revealed by intron sequences of four nuclear genes. *New Phytol.* **167**, 249–265 (2005).
30. Londo, J.P., Chiang, Y.C., Hung, K.H., Chiang, T.Y. & Schaal, B.A. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc. Natl. Acad. Sci. USA* **103**, 9578–9583 (2006).
31. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792 (1978).
32. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
33. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
34. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
35. Tang, H., Peng, J., Wang, P. & Risch, N.J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
36. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
37. Fuller, D.Q. *et al.* The domestication process and domestication rate in rice: spikelet bases from the Lower Yangtze. *Science* **323**, 1607–1610 (2009).
38. Hill, W.G. & Robertson, A. Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics* **38**, 226–231 (1968).
39. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
40. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
41. Tenaillon, M.I., U'Ren, J., Tenaillon, O. & Gaut, B.S. Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* **21**, 1214–1225 (2004).
42. Wright, S.I. *et al.* The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314 (2005).
43. Yamasaki, M. *et al.* A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859–2872 (2005).
44. Yamasaki, M., Wright, S.I. & McMullen, M.D. Genomic screening for artificial selection during domestication and improvement in maize. *Ann. Bot.* **100**, 967–973 (2007).
45. Gibbs, R.A. *et al.* Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**, 528–532 (2009).
46. Xia, Q. *et al.* Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science* **326**, 433–436 (2009).
47. Rubin, C.J. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010).
48. Wright, S. *Evolution and the Genetics of Populations* (The University of Chicago Press, Chicago, 1977).
49. Jain, M. & Khurana, J.P. Transcript profiling reveals diverse roles of auxin-responsive genes during reproductive development and abiotic stress in rice. *FEBS J.* **276**, 3148–3162 (2009).
50. Paponov, I.A. *et al.* The evolution of nuclear auxin signalling. *BMC Evol. Biol.* **9**, 126 (2009).
51. Retief, J.D. Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.* **132**, 243–258 (2000).

© 2012 Nature America, Inc. All rights reserved.

ONLINE METHODS

Sampling. All samples were grown in greenhouses for morphological confirmation.

Library construction and sequencing. Genomic DNA was extracted from fresh or silica gel-dried leaves of a single plant, using the CTAB method as described[52]. For each of the accessions, we only sequenced a single individual. At least 5 μg genomic DNA was used for each accession in constructing sequencing libraries. Paired-end sequencing libraries with insert sizes of ~200 bp or 500 bp were constructed for accessions according to the manufacturer's instructions (Illumina). We used the same workflow as described previously to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking, denaturation and hybridization of the sequencing primers[53]. We sequenced 45 bp or 100 bp at each end and used SolexaPipeline-0.3 to call bases for 45-bp reads and version 1.0 for 100-bp reads from the raw fluorescent images.

Reads mapping. We checked the insert size distribution of each library by Eland in the SolexaPipeline and used the distribution as the input insert size range parameter in a paired-end alignment. We first used SOAPaligner, (SOAP2)[54] to align all short reads to the Nipponbare reference genome. The detailed parameters used are as follows:

"soap2.20 –p 4 –a 1.fq –b 2.fq –D IRGSP_chromosomes_build04.fa.index –o sample.soap -2 sample.single –u unmapped.fa –m 435*1 –x 501*1 –s 35 –l 24 –v 7"

(*1: the insert size was estimated by Eland and given to each library as input).

This process allowed trimming of the reads to map the reads back to the genome (-s 35). But, the actual mapping length of the reads was long (**Supplementary Fig. 3**), with 95% of the reads longer than 50 bp (originally 100 bp). Although we allowed a maximum of 7 mismatches in 100-bp and 3 mismatches in 45-bp reads mapping, the actual missing matches within a single read were less than 2 in 100-bp reads that mapped uniquely to one position in the reference genome (**Supplementary Fig. 17**). We also simulated reads from the reference genome and estimated the mapping rate (**Supplementary Table 12**) by the same mapping process to prove the difficulty in mapping.

Identification of novel genes. We assembled the unmapped reads from each sample into contigs by SOAP*denovo*[55] (default parameters were used and only contigs were constructed, not scaffolds). When identifying novel sequences, we first assembled the unmapped reads separately in each accession, and contigs shorter than 2 kb were excluded. We then used the self-alignment approach to exclude the redundant sequences. In total we identified 5,795 contigs with a total length of 23.2 Mb. We blasted all these candidate "novel" contigs against the reference genome to identify homologs in the Nipponbare genome. We found that 1,403 (24%) contigs indeed have more or less similar homologous sequences in the Nipponbare genome with coverage >30% and identity >80%, indicating these contigs' sequences were very possibly from diverged homologs in the reference genome, which caused mapping difficulty. The remaining 4,392 contigs are either real novel sequences in other accessions or located in nonassembled heterochromatin regions. The GC ratio of these 23.2 Mb sequences is 42.4%, which is comparable to the GC ratio of the genome (43.5%). We conducted *de novo* gene annotation with AUGUSTUS[56] for the 5,795 contigs. After annotation, we excluded the redundant genes that were assembled in different accessions. Only one copy of the genes with more than 90% identity and 90% coverage by BLAT was retained. In total, we annotated 2,031 possible novel genes *de novo*. Then, we used BLASTP to compare the candidate novel genes against the NCBI nr database; 1,552 (76%) of the genes have homologs in the nr database (more than 60% identity and 60% coverage). Of the 1,552 genes, 1,415 (92%) genes have homologs in plants. The other 137 (9%) genes only had homologs in species other than plants, which might have been from contamination. We functionally annotated the 1,415 proteins by InterProScan[15]. The average gene length is substantially shorter than that of the whole genome (957 bp versus 2,300 bp), indicating that many of the annotated genes are not intact. A total of 432 genes are located in the 1,403 contigs with homologs in the Nipponbare reference genome, indicating that they are possibly from diverged homologs in the Nipponbare genome, as pointed out by the reviewer. Other genes may either be from real novel sequences in other accessions or highly diverged subdomain of some genes. We described the homolog information for all the genes in **Supplementary Table 3**. In the 1,415 genes, 60% of the genes can be functionally annotated, but only 24 (1.7% of the total "novel" genes) genes were annotated as transposons or retrotransposons, indicating most of them are genes or part of genes. These may have arisen from the transposition of Helitrons or Pack-MULEs, which commonly create incomplete pseudogenes in plants.

Identification of gene loss events. To identify gene loss events, we first extracted genes with <10% coverage in the gene region in some accessions but >90% coverage in the Nipponbare accession; 1,327 such genes were initially identified as candidates of lost genes. Then, we added paired-end reads to support the gene loss events. As we constructed short-insert-size libraries (200 bp or 500 bp) for sequencing, reads that mapped to the reference genome with obvious longer insert sizes indicated a possible deletion in between. The candidate lost genes with at least one such uniquely mapped split read were finally identified as lost genes. By such a procedure, we identified 839 lost genes in various accessions. We further used PCR to validate the gene loss events. We randomly chose 9 deletions which were longer than 2 kb and used PCR to experimentally validate the deletions. Eight out of nine were clearly shown the deletions are real (**Supplementary Fig. 18**).

SNP calling. Incorrect mapping would have a great effect on the accuracy of variation detection, especially in SNP calling. To get high-quality SNPs, we excluded reads that could be mapped to different genomic positions in the mapping results by SOAP2. Uniquely mapped single-end and paired-end results were used in the SNP calling.

For the 50 accessions that were sequenced to a depth of ~15×, SNPs were called in three steps

1. Likelihoods of genotypes of each individual at every genomic site were calculated by SOAPsnp[36]. In each individual, SNPs were filtered by the quality value given by SOAPsnp, which should be >20, and the base quality at this position should pass the rank-sum test (in SOAPsnp with $P > 0.05$). Then, the sites that were identified as SNPs in at least one individual were identified as possible SNPs. We obtained 15 M raw SNPs at this step.
2. Call SNPs in a population by realSFS, a software that has been applied in a human population SNP calling[57], based on the Bayesian estimation of site frequency at every site. It integrates the likelihoods of genotypes of each individual at each site generated in step 1. Sites with a probability to be variant >0.99, given by realSFS, are further extracted by the following criteria to get the possible SNPs: covered by at least one uniquely mapped read in each of the 50 accessions. Our results show that at a depth above tenfold coverage for a population, the property of real SNP sites (the probability of a variable >0.99) differs significantly from the nonvariable sites as well as the sites that are caused by sequencing and calling errors (**Supplementary Fig. 19**). In this SNP-calling procedure, there were 30,971,130 positions identified by realSFS to be variant (possibility >0.99), and 7,492,068 of them were covered in all 50 accessions by at least one uniquely mapped read. In this procedure, the allele frequency of each position was also estimated by realSFS (**Supplementary Fig. 20**). The population statistics based on SNPs were then estimated using the information on allele frequencies.
3. We obtained the final SNP set by combining the two sets of possible SNPs above (from steps 1 and 2); 6,796,190 were identified as SNPs in both SNP sets. Then, using the genotype information at each SNP position given by SOAPsnp in Step 1, we further filtered the SNPs that deviated from the Hardy-Weinberg principle. Finally, we identified 6,496,456 SNPs. This extremely stringent SNP-calling process guarantees a low false-positive rate in our final SNP data set. We randomly validated 89 selected SNPs in all 50 accessions (**Supplementary Notes** and **Supplementary Table 13**). We also validated the heterozygous SNPs in Nipponbare which were significantly clustered in some regions (**Supplementary Fig. 21**) by randomly picking 62 heterozygous SNPs (**Supplementary Table 14**) by Sanger sequencing (**Supplementary Fig. 22**) (**Supplementary Notes**). These validations confirm the high quality of the SNP data set.

© 2012 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.2050

For the 15 wild accessions that were sequenced at a lower depth (2–3×), we called SNPs for each of the 15 accessions using SOAPsnp[35] and only retained SNP sites that were covered by reads in at least 6 out of the 15 wild accessions to get informative sites for population structure analyses. By imposing these SNPs onto our 6.5-M SNP set identified in the previous 50 accessions, we obtained 3,668,781 SNPs that were further used in the phylogenetic tree, PCA analysis and population structure analyses for all 65 accessions.

Short indel detection. To detect insertions and deletions (shorter than 5 bp), another mapping process with a gap allowed (additional parameter of "-g 5" was used in SOAP2) was performed. Indels (1–5 bp) were called by SOAPindel pipeline (http://soap.genomics.org.cn/) as described in previous studies[46,54]. Indels were identified in each accession and then combined, based on the position and length of the insertions or deletions. We then randomly selected 56 indels to be validated by Sanger sequencing (**Supplementary Notes** and **Supplementary Table 15**).

Structural variation (SV) and copy number variation (CNV). To detect structural variations longer than 10 bp, we applied a process similar to one described previously[58] using assembly. To obtain a better assembly, we combined the sequences from each subgroup. Thus, for each subgroup, the sequences for assembly were more than 50×. Then, the assembly into contigs and scaffolds was done using SOAP*denovo* by default parameters and processes. The assembled scaffold was mapped to the reference genome by BLAT[59] with the –fastmap option. A scaffold was selected as the best aligned one if it covered the longest in length in the region and had the most contig supports. Then, the scaffolds and the 'best alignment' regions of the reference genome were extracted and aligned by LASTZ (http://www.bx.psu.edu/miller_lab/). For those unmapped scaffolds, we further tried to align them against the reference genome using BLASTn[60]. Finally, the structural variations were extracted using all those aligned regions.

For CNV detection, we used the mapping results by mrFAST, which outputs all the possible alignments when a read can be mapped to multiple positions in the genome. Using the mapping results, we calculated the mapping depth of each base of the reference genome in each accession. Then, the nearby bases without significant differences (the total depth distribution is assumed to be in a Poisson distribution) in mapping depth were combined into initial windows. The mean depth of each window was then calculated and compared to other initial windows nearby. Initial windows without significant differences in depth were then combined into larger windows. This process of window merging was done one more time, and the edges and the copy number of each window were decided in this dynamic way. As we detected lost genes earlier, in detecting copy number variations, we only retained regions with more copies than the reference.

Construction of phylogeny. SNPs were used to calculate the genetic distances between different accessions. The *p*-distance between two individuals *i* and *j* is defined to be

$$D_{ij} = \frac{1}{L}\sum_{l=1}^{L} d_{ij}^{(1)}$$

where L is the length of regions where SNPs can be identified, and given the alleles at position l are A/C, $d_{ij}^{(1)}$ was set to 0 if the genotypes of the two individuals were AA and AA; 0.5 if the genotypes of the two individuals were AC and AC; and 1 if the genotypes of the two individuals were AA and AC. Then, a neighbor-joining method was used to construct the phylogenetic tree on the basis of the distance matrix, calculated by the software PHYLIP 3.68 (http://evolution.genetics.washington.edu/phylip.html), and MEGA4 (ref. 34) was used to present the phylogenetic tree.

Population structure inference. We performed a PCA following the procedure as reported[32]. The eigenvector decomposition of the transformed genotype data was performed using the *R* function eigen, and the significance of the eigenvectors was determined with a Tracey-Widom test, implemented in the program twstats, provided by the EIGENSOFT software[32].

We further used the program FRAPPE, which is based on a maximum likelihood method[35], to investigate the population structure.

We ran 10,000 iterations, and the number of clusters (K) was considered from 2 to 7.

Calculation of linkage disequilibrium (LD). To measure LD levels in different populations, we calculated the correlation coefficient (r^2) of alleles after setting

-maxdistance 500 -dprime -minGeno 0.6 -minMAF 0.1 -hwcutoff 0.001

by the software Haploview[39]. Then, it was plotted with R scripts, which drew averaged r^2 against pairwise marker distances.

Estimation of population parameters and detection of putative artificial selection genes.

1. π and θ_w.
   π is defined as the average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population[26]

$$\pi = \sum_{ij} x_i X_j \pi_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{i} x_i x_j \pi_{ij}$$

   where x_i and x_j are the respective frequencies of the *i*th and *j*th sequences, π_{ij} is the number of nucleotide differences per nucleotide site between the *i*th and *j*th sequences and n is the number of sequences in the sample. The summation is taken over all distinct pairs *i,j*, without repetition.
   θ_w is the estimation of the population mutation rate, based on the number of segregating sites[61].
2. F_{ST}.
   F_{ST} is a measure of population differentiation, genetic distance, based on genetic polymorphism data[62], which is defined as

$$F_{ST} = \frac{\Pi_{Between} - \Pi_{Within}}{\Pi_{Between}}$$

   where $\Pi_{Between}$ and Π_{Within} represent the average number of pairwise differences between two individuals sampled from different ($\Pi_{Between}$) or the same (Π_{Within}) population. We specifically analyzed the variations in resistant genes (*R* genes) (**Supplementary Notes**, **Supplementary Table 16** and **Supplementary Fig. 23**).
3. Ratios of nonsynonymous and synonymous polymorphic sites.
   To calculate the synonymous changes relative to nonsynonymous changes at the whole genome level, we defined ratios of nonsynonymous and synonymous polymorphic sites as the sum of π for nonsynonymous sites relative to the sum of π for synonymous sites in one gene. Here we used only the representative transcript for each gene in the rice gene annotation (RAP-DB version 4). Genes with effective lengths (including the exon and intron) shorter than 1,000 bp were removed. If nonsynonymous π or synonymous π in one population (*Japonica*, *Indica*, *O. nivara*, *O. rufipogon*) was 0, the gene was also removed from the gene list.
4. *ROD* values and detection of putative genes under selection.
   Selected regions in cultivars are expected to have a lower diversity compared to the same regions in the wild species. To measure this, we defined the reduction of diversity by *ROD* as

$$ROD = 1 - \frac{\pi_{cul}}{\pi_{wild}}$$

   where π_{cul} and π_{wild} are the values of π for the cultivated and wild varieties, respectively, calculated in 10 kb or 100 kb nonoverlapping windows along the genome.
   By using 10 kb or 100 kb nonoverlapping windows along the genome, we calculated the *ROD* value for each window. The windows with a significantly high *ROD* in the 2.5% or 0.25% right tail of the *ROD* empirical distribution are picked out as candidate selective sweep regions, and genes in these regions are identified as putative genes under selection. Usually, a 10-kb region contains a single gene, but when it contains several genes, we took all of them as candidate artificially selected genes. If a gene crossed two windows, it was only counted once.
5. Analysis of the artificial selection genes.

© 2012 Nature America, Inc. All rights reserved.

The gene family information of the artificial selection genes was obtained from the RAP DB[63]. Then we used χ^2 method to test the enrichment of artificial selection genes in some gene families. If the number of artificial selection genes in one gene family ($n_{family_selected}$) was significantly more than the expected number inferred by considering the number of genes from that gene family (N_{family_total}) in the total gene set (N_{total}) and the total number of genes under selection ($n_{selected}$), then the test of independence between categorizing by gene families and selection was performed using the contingency table below.

	In this gene family	Not in this gene family
Selected	$n_{family_selected}$	$n_{selected} - n_{family_selected}$
Not selected	N_{family_total}	$N_{total} - N_{family_total}$

Then $\chi^2 = \Sigma(O_i\text{-}E_i)^2/E_i$ was used to calculate the χ^2 value, where O_i is the observed value of each cell (as shown in the above table), and E_i is the expected value of each cell calculated by the total number of each column and line. If the test shows significant dependence of whether the one gene is from this gene family with whether this gene is selected, the genes from this family are prone to be selected or not to be selected. And if the number of genes selected in this family is greater than the expectation, we determine this gene family as significantly enriched in selection. The *P*-values of the statistical test were shown in **Supplementary Table 11**.

52. Zhu, Q., Zheng, X., Luo, J., Gaut, B.S. & Ge, S. 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).
53. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
54. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
55. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
56. Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
57. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
58. Li, Y. *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotech.* **29**, 723–730 (2011).
59. Kent, W.J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
60. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
61. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
62. Hudson, R.R., Slatkin, M. & Maddison, W.P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
63. Tanaka, T. *et al.* The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36**, D1028–D1033 (2008).

© 2012 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.2050