doi:10.1038/nature11532

Table of Contents

- Supplementary Table 6. Specificities and missing data proportions of the genotype dataset of *O. rufipogon* population.
- Supplementary Table 7. The list of 1,083 cultivated rice accessions (*O.sativa*) sampled in the collection.
- Supplementary Table 8. The list of 15 accessions of outgroup species sampled in the collection.
- Supplementary Table 9. Genome-wide detection and functional annotation of selective sweep regions in the full population.
- Supplementary Table 10. Genome-wide detection and functional annotation of selective sweep regions in *indica* population.
- Supplementary Table 11. Genome-wide detection and functional annotation of selective sweep regions in *japonica* population.
- Supplementary Table 12. Large-scale genetic mapping of fifteen domestication traits.
- Supplementary Table 13. Co-localization of QTLs for domestication traits.
- Supplementary Table 14. Summary of whole-genome *de novo* assembly of a wild rice accession (*O. rufipogon* W1943).
- Supplementary Table 15. The list of functional variants in coding region with strong alteration in allele frequency between cultivated and wild rice.
- Supplementary Table 16. The list of sequence variants in the promoter region with strong alteration in allele frequency between cultivated and wild rice.
- Supplementary Table 17. The list of functional variants that are nearly fixed in *indica* panel.
- Supplementary Table 18. The list of functional variants that are nearly fixed in *japonica* panel.

Supplementary Sections

Section S1. Phylogenetic relationship reconstruction of AA-genome complex.

A total of 15 rice accessions were selected and sequenced to construct the phylogenetic tree of AA-genome species complex in *Oryza*. The genus *Oryza* contains four sections, including the AA-genome complex. The AA-genome complex contains two species of cultivated rice and five species of wild rice, all of which are diploids with twelve chromosomes. The two domesticated species are Asian cultivated rice (*O. sativa*) and African cultivated rice (*O. glaberrima*). *O. sativa* harbors many agronomic traits of high yield and has a broad geographic distribution across the world. The species contains five genetically distinct ecotype groups. The five groups are *indica*, *aus*, *temperate japonica*, *tropical japonica* and *aromatic*. We selected five representative lines of *O. sativa* from each ecotype group. The five rice accessions are Nipponbare (*temperate japonica*), GP640 (*tropical japonica*), GP294 (*aromatic*), Guangluai-4 (*indica*) and GP110 (*aus*), respectively. Four diverse accessions of common wild rice *O. rufipogon* were selected, which were W1943, W0610, W3105 and W0600, respectively. African cultivated rice *O. glaberrima* is relatively stress tolerant, but still harbored many undesirable phenotypes (for example, shattering). The species *O. glaberrima*, coupled with its close wild relative *O. bathii*, is mainly grown in West Africa. We selected two accessions, W3104 (*O. glaberrima*) and W3106 (*O. bathii*) for deep sequencing. The other species sampled here included *O. longistaminata* (W3101 and W3102) and *O. meridionalis* (W2113 and W2115). The species *O. glumaepatula*, which was mainly located in central and south America, was not included in the sample.

The phylogenetic relationship of these species was determined using genetic distances calculated from whole-genome SNPs. The resulting neighbor-joining tree shows that *O. rufipogon* is the immediate progenitor of *O. sativa*, and that the two species grown in West Africa, *O. glaberrima* and *O. barthii*, are the nearest outgroups for *O. rufipogon* and *O. sativa* (Supplementary Fig. 2). The genomes of two other AA genome species of wild rice, *O. longistaminata* and *O. meridionalis*, contain many

more sequence differences relative to *O. sativa* (approximately one difference per 50 bp), indicating earlier divergence in the history (Supplementary Table 1).

Section S2. Parameter optimization and permutation tests for screening selective sweeps.

We performed whole-genome screening of selective sweeps for *indica*, *japonica* and the full population, respectively. The accessions of both *O. sativa* and *O. rufipogon* were used to screen selection signals in the rice genome. The levels of sequence diversity were calculated in *O. sativa* population (π_c) and *O. rufipogon* population (π_w), respectively, in each 100-kb window, which were then plotted against each other. The plotting showed that the two has a high positive correlation (R^2=0.4). In the procedure for calling selective sweeps, all genomic regions, with the selection signals (π_w/π_c) higher than 3.0, were left as candidates firstly. The genomic regions of low diversity, where both π_c and π_w was lower than 0.001, were further removed. Moreover, those continuous 100-kb regions with high values of π_w/π_c were merged into a single selected locus. The detection threshold, 3.0 (π_w/π_c), was chosen based on permutation tests and the well-characterized domestication loci.

1) It was found that <5% of genomic regions have the ratio of π_w/π_c higher than the cutoff.

2) As a negative control, the value "π_c/π_w" was calculated to detect the loss of diversity regions in *O. rufipogon* population (using *O. sativa* population as the contrast). We found that the highest value is 1.6, thus no positive signals detected.

3) We further reshuffled the species information in the combined population of *O. sativa* and *O. rufipogon*, and then performed selective sweep detection with the same parameters. We performed ten whole-genome screenings using this approach, and found that there were no positive signals passing the empirical threshold.

4) Most well-characterized domestication loci have high selection signals, passing the threshold we set.

Hence, all of the above indicated that the empirical threshold can yield selection signals with very low false discovery rate.

To screen selection signals in *indica* subspecies, both *indica* accessions (including *aus* rice) and wild rice accessions in *Or*-I clade were used. The π_{Ind}–$\pi_{Or\text{-}I}$ plotting showed that the two has a high positive correlation (R^2=0.5). We chose the same threshold (as that used in full population) in detecting selective sweeps in *indica*. The *japonica* accessions (including *temperate japonica*, *tropical japonica* and *aromatic* rice) and wild rice accessions in *Or*-IIIa sub-clade (the sub-clade of *Or*-III that close to *japonica*, mainly located in Southern China) were used to screen selection signals in *japonica* subspecies. We chose the threshold of 14 ($\pi_{Or\text{-}III}/\pi_{Jap}$) in detecting selective sweeps in *japonica*.

Moreover, we also used the method Cross Population Extended Haplotype Homozogysity (XP-EHH) in the analysis of our rice data. The method XP-EHH, which is based on cross-population comparisons to discover alleles that have swept to near-fixation within a population, have been used in the detection of selective sweeps across the human genome and the *Arabidopsis* genome (Sabeti et al., 2007; Horton et al., 2012). We calculated the XP-EHH score, and reveals 34 and 29 sweeps in *indica* and *japonica* rice (Supplementary Fig. 20). Focusing on the eight loci with well-characterized domestication genes, we found that only two of them (*sh4* and *qSW5*) showed selection signals (with the standardized XP-EHH score of > 5), the performance of which showed inferior to the results from simple diversity ratios. It is likely to be due to that cultivated rice is a highly selfing species and a domestication species with long LD, low effective recombination rate and strong population structure (Caicedo et al., 2007), which is quite different the situation in human and *Arabidopsis* populations. The results from XP-EHH were not used in the followed analyses.

Section S3. Phylogenetic and demographic analysis on the origin of Asian cultivated rice.

We employed phylogeographic and phylogenetic approaches to identify the detailed geographic regions where cultivated rice arose, based on the extensive genomic and geographic information of the large sample. The three clades, *Or*-I, *Or*-II and *Or*-III,

were determined according to the neighbor-joining tree as well as principal-component analysis (Fig. 1a and Supplementary Fig. 3), which was constructed from 5 million SNPs in 446 wild rice accessions. The phylogenetic circle of Fig. 2a was constructed from 8 million SNPs in 1,529 rice accessions. According to the phylogenetic tree of 1,529 lines using whole-genome SNP data, the *Or*-III group of *O. rufipogon* can be divided into two major clades, one close to *japonica* rice (namely *Or*-IIIa) and the other close to *Or*-II group relatively (namely *Or*-IIIb) (Fig. 2a). The former was mainly located within Southern China, and the latter was mostly distributed in South Asia (Fig. 1b). We found that, different from that in *O. sativa* (*indica* and *japonica* are very distinct in the phylogenetic tree), the genetic structure in *O. rufipogon* is relatively continuous and grouping of *O. rufipogon* is not absolute. Hence, it should be also appropriate to group *O. rufipogon* into four clades -- *Or*-I, *Or*-II, *Or*-IIIa and *Or*-IIIb.

From whole-genome data, a small number of accessions in the sampling were observed to be the intermediate type between cultivated and wild rice, which included some accessions of *indica* and *Or*-I (see Fig. 2a). The rice accessions of intermediate type, which would be called proto-*indica* (Fuller et al., 2010), may reflect the track that the domesticated *japonica* rice was crossed to local wild rice in India-Indonesia regions.

Using SNP data on 55 domestication loci, the resulting phylogenetic tree showed that the nearest clade for *O. sativa* was 62 wild rice accessions in *Or*-IIIa group (Fig. 4a). The plotting of their geographic locations of the wild rice accessions of the nearest clade shows that the accessions mainly come from Southern China, with the closest relatives located in Guangxi province (Fig. 4b). The result was constant with the observation that the loss-of-function allele of *GS3* (the allele fixed in *japonica*) was from wild rice accessions of Guangxi and Guangdong provinces.

In our sampling, there were a total of 99 wild rice accessions with detailed sampling locations in China (Supplementary Table 2). Another 17 lines that were sampled in China but had no detailed location information were excluded in the following analysis. We calculated the average simple SNP matching coefficients from

all the cultivated rice on the 55 domestication loci and on the *Bh4* locus specially, for wild rice accessions of six provinces. The six provinces, in the order of genetic distance to cultivated rice, were Guangxi province, Guangdong province, Jiangxi province, Hunan province and Yunnan province, respectively (Fig. 4d). Of note, Guangxi is the province of Southern China along its border with Southeast Asia (Fig. 4b).

We also compared genetic distance of wild rice in different geographic regions to *japonica* and *indica* separately around the domestication loci. The results were nearly identical to that from comparing them collectively -- the wild accessions from Southern China are the closest with both *japonica* and *indica* in the domestication loci. Comparison of the distances to *japonica* and *indica* separately were also carried out at the whole genome scale. As expected, the pattern is indeed quite different from that in the domestication loci. The results showed that, at the whole genome scale, the wild accessions from Southern China are the closest with *japonica* while the wild accessions from Southeast Asia (especially Laos, Burma and Thailand) are the closest with *indica* (Supplementary Fig. 23).

Section S4. Selection signals of eight known domestication genes.

Up to now, tens of genes have now been cloned in rice responsible for many agronomic traits. Genes that were cloned based on natural variation (not mutants) in rice can be classified into different groups, with different evolutionary features.

1) For most cloned QTLs, their alleles are responsible for natural intra-species variation. The allelic variations have already existed in the gene pools of wild rice. Due to a high level of genetic differentiation between *indica* and *japonica*, many genes of this type are also related to subspecies divergence (for example, *GS3*, *DPL2* and *Ghd7*).

2) For a few cloned QTLs, their minor alleles were existed in natural population at a low frequency. The minor alleles, however, have large effects on corresponding traits of agronomical importance [for example, *GW2* (Song et al., 2007) and *OsSPL14* (Jiao et al., 2010; Miura et al., 2010)]. Some of them had already been utilized in

modern breeding, which were introduced into more rice varieties during modern breeding (for example, *sd-1*, Sasaki et al., 2002).

3) For a few cloned QTLs underlying typical domestication traits, the favored alleles for this trait under cultivated conditions have been significantly distinct from those under natural conditions of wild rice (for example, *sh4* and *Prog1*).

From the view of evolutionary studies, only genes of the last type can be considered to be domestication-related genes. Up to now, the well characterized domestication-related genes may include *sh4*, *qSH1*, *Prog1*, *Bh4*, *qSW5*, *Rc*, *OsC1* and *Waxy*. Among these genes, *sh4* and *qSH1* are mainly responsible for seed shattering. The *Prog1* gene confers erect plant architecture in cultivated rice. The genes *qSW5* and *Waxy* control grain width and grain quality in rice, respectively. The other three genes *Rc*, *OsC1* and *Bh4* are related to selection for appearances, underlying pericarp color, stigma color and hull color, respectively.

Among them, the five genes *sh4*, *Prog1*, *qSW5*, *OsC1* and *Bh4* were among selective sweeps identified in the full population. The phylogenetic trees around the five domestication loci were constructed based SNPs around 40kb flanking the genes. The other three genes *qSH1*, *Waxy* and *Rc* were among selective sweeps identified in *japonica* population. These gene alleles and their corresponding traits have different patterns of frequency alternation between cultivated and wild rice. The two genes, *sh4* and *Prog1*, are considered to be typical domestication genes because nearly all cultivated rice harbors non-shattering and erect phenotypes. In the wild rice, we observed that only a small number of wild rice accessions also have such phenotypes (however, nearly all accessions of outgroup species have the erect phenotype where the gene is absent in their genomes). As for pericarp color, we observed that many *indica* landraces show red pericarp. The coloration on hull are probably an important indicator of the appearance between cultivated and wild rice, since most *O. rufipogon* accessions show black hull and most *O. sativa* accessions show straw-white hull.

Section S5. Computational simulations of five demographic scenarios of the rice domestication history.

The origin of cultivated rice and its domestication history have long been debated. During past decades, several models were proposed for rice domestication, which were raised from a mass of genetic, phylogenetic and archaeological studies. Below we summarized the most common hypothesizes for rice domestication (Supplementary Fig. 24).

Based on the phenotypic similarity between cultivated rice and the annually occurring form of *O. rufipogon* (also called as *O. nivara* in other studies), the hypothesis of a simple single origin of cultivated rice was proposed (**H1**). This model suggested that the cultivated rice was domesticated from the annual form first, and the two subspecies *indica* and *japonica* were regarded to be formed after domestication (Chang 1976; Khush 1997; Sharma et al., 2000).

However, with the map of rice genome sequence available (Yu et al., 2002; International Rice Genome Sequencing Project, 2005), the comparative genomics analysis estimated that the divergence between *indica* and *japonica* is ~0.4 million years ago, which considerably predates the time of rice domestication (~10, 000 years ago) (Ma et al., 2004; Han and Xue, 2003). Moreover, phylogenetic and population studies using a number of molecular markers in the rice genome also revealed ancient genomic differentiation between rice cultivars (Cheng et al., 2003; Londo et al., 2006; Zhu et al., 2005; also supported by our data). Hence, the hypothesis of independent domestication origins of cultivated rice was proposed (**H2**). This model suggested that *indica* and *japonica* were domesticated independently, from different clades of *O. rufipogon* (Londo et al., 2006).

If *indica* and *japonica* were independently domesticated, we would expect the occurrence of different domestication loci and completely different causal mutations between the two subspecies. However, the key domestication genes (for example, *sh4*, *Prog1* and *Bh4*) in rice genome identified during recent few years show the patterns inconsistent with the independent domestication hypothesis. Nearly all cultivars shared the same domestication alleles with the same causal mutations in these domestication genes, which indicated that they have originated only once, not independently in *indica* and *japonica*. Our data also showed that many loci with

strong signals of selection were nearly identical in *indica* and *japonica* where the divergence between *indica* and *japonica* was extremely low. Under the circumstances, several alternative models were further proposed (Sang and Ge, 2007; He et al., 2011).

One of these hypotheses considers two origins of cultivated rice followed with bidirectional shares of domestication alleles (**H3**). In this model, the two subspecies *indica* and *japonica* were domesticated respectively, and subsequent frequent crosses among the proto-*indica* and proto-*japonica* rice resulted in the shares and fixation of the key domestication alleles. The other two hypotheses are **H4** -- *indica* rice was first domesticated and *japonica* rice was subsequently developed from crosses between *indica* rice and another clade of *O. rufipogon*; and **H5** -- *japonica* rice was first domesticated and *indica* rice was subsequently developed from crosses between *japonica* rice and another clade of *O. rufipogon*.

In this study, the phylogenetic and haplotype analyses using the data of 1,529 rice genomes showed that both subspecies (*indica* and *japonica*) were close to the clade of *O. rifipogon* in Southern China (called *Or*-IIIa in this study) around the domestication loci, while *indica* and *japonica* show close to *Or*-I (a clade of *O. rifipogon* in Southeast Asia and South Asia) and *Or*-IIIa, respectively, in other regions of the rice genome. Hence, the model of **H5** is regarded to be the most likely one.

Computational simulations have been successfully used in the inference of the demographic history in the human population genetics (Hoban et al., 2012). This approach was adopted here to test which model represents the true evolutionary history of rice domestication. We used the forward-in-time simulation program SFS_CODE to simulate the five demographic scenarios (Hernandez, 2008). The SFS_CODE software package was downloaded from the website http://sfscode.sourceforge.net/SFS_CODE/. To simulate the two clades of *O. rufipogon* (*Or*-I and *Or*-III), we divided an ancestral population into two equal sized populations followed with independent evolvements (the command-line argument: -TS 0 0 1). To simulate the domestication loci of cultivated rice that was generated directly from domestication, we allowed the population to be created from one clade

of *O. rufipogon* by the domestication model in the SFS_CODE (the argument "-TD 7 1 3 0.1 10"). To simulate the loci of a subspecies of cultivated rice that was generated from introgression, we allowed the population to be created by the admixture model (the argument "-TJ 15 2 20 2 0 3 0.2 0.8"). For each population, a total of 100 individual lines were simulated. For each individual line, the simulation generated 500 genomic loci (one locus for each simulation replication), and each genomic locus contained a 100-kb sequence.

We calculated the genetic distance between the subspecies of *O. sativa* and each clade of *O. rufipogon* for comparison for both simulated datasets and the real data. It is found that there were significant differences between the real data and the simulated ones in all the scenarios except for the model **H5** ($P < 0.01$, rank sum test), thus further supporting our conclusion.

Section S6. Genetic mapping of domestication traits using BILs and CSSLs.

The BILs and CSSLs were generated from a cross between *O. sativa* (ssp. *indica* cv. Guangluai-4) and *O. rufipogon* (*Or*-IIIa accession W1943) followed with backcrossing to *O. sativa* Guangluai-4. The two parents, *O. rufipogon* W1943 (*Or*-IIIa) and *O. sativa* Guangluai-4 (*indica*), are typical wild rice and cultivated rice, respectively. Both of them have genome sequences, BAC library and full-length cDNA sequences. The domestication gene *Bh4* was cloned using this cross between the two parents (Zhu et al., 2011). Here the recombination population between W1943 and Guangluai-4, including a total of 271 lines, was genotyped by whole-genome sequencing, which generated an ultra-dense recombination bin map. In the analysis procedure, all chromosomes of the 271 lines were aligned and compared for the minimal of 100-kb intervals. Adjacent 100-kb intervals with the same genotype across the entire population were merged into a single recombination bin. A total of 1,793 recombination bins were obtained for the 271 lines. The average physical interval for the recombination bins was 200 kb, ranging from 100 kb to 4.8 Mb. The length of the recombination bin was mainly dependent on local recombination frequency, which has the direct effect on the mapping resolution. According to the high-density

genotype dataset, the genotypes of Guangluai-4/Guangluai-4 (GG), W1943/W1943 (WW) and Guangluai-4/W1943 (GW) take a proportion of 80.2%, 19.2% and 0.6% on average in the population, respectively. The allele frequencies of GG, WW and GW have a relatively even distribution across the rice genome. In the population of 61 CSSLs, the total size of the substituted segments was 1,040 Mb, ~2.6 times that of the rice genome.

Among 58 QTLs identified, there were ten known causal genes -- *Ehd2*, *Ghd7*, *Hd8*, *Bh4 OsC1*, *Prog1*, *qSW5*, *Rc*, *sd-1* and *sh4*, responsible for eight agronomic traits. According to the mapping information of these known genes, we found that the causal genes were always located with the recombination bin with peak LOD value or just the adjacent bin, implying ~200-kb mapping resolution. Among the known genes, the only exception was the QTL for *Ghd7*, where the recombinant frequencies were dramatically low. The situation for mapping *Ghd7* in this population was quite similar with the previous report, where rare recombinants can be found in this local genomic region even within a large population (Xue et al., 2008).

The sites of recombination breakpoints were mapped within 3kb that depended on the high genotyping density, while the bin size was about 200kb on average that depended on recombination rates. We found that the binning less than 100 kb cannot further improved the mapping resolution, probably due to the limited recombinants, the population size and the QTL mapping algorithm (CIM method in the software WinQTLcart). However, with the fine recombinant breakpoints, we can accurately determine the interval of the QTL, which would help the follow-up fine-mappings.

Through inspecting the genetic architecture of the 15 traits, we found that these QTLs accounted for ~51% of the phenotypic variance on average (from 15% to 94% for different traits). Those QTLs with evidence of selection tend to have larger effects on phenotype variation than those without selection (an average of 16.0% for the former versus 9.7% for the latter). Interestingly, multiple QTLs underlying different domestication traits were co-localized within the same selective sweep regions in chromosome 7, which was consistent with the previous finding (Li et al., 2006).

Section S7. Genome sequence and analysis of common wild rice (*O. rufipogon*) W1943.

The reference genome sequence of *japonica* Nipponbare is a powerful resource for functional genomics studies in cultivated rice. The *O.rufipogon* accession W1943, as one parent of the recombination population above, harbors nearly all representative phenotypic traits of wild rice. We have generated both full-length cDNA sequences and BAC clone library for the accession during the past few years. The BACs of the wild rice accession W1943 that we have sequenced has a total length of 656,258 bp. The sequences spanned from 22.7 Mb to 23.4 Mb in rice chromosome 4 (IRGSP 4 release). In this study we further sequenced the accession with 100-fold genome coverage. The final assembly of wild rice W1943 has a total length of 406 Mb. We aligned 2,045 full-length cDNAs of wild rice W1943 (Lu et al., 2008) to the final assembly of W1943 genome, and found that ~98% of the cDNAs (~92% of the bases) had nearly perfect matches in the assembly, indicating that the whole-genome assembly has relatively high genome coverage. Of 2,045 full-length cDNA cloned from W1943 leaves, 2,006 cDNAs (98.1%) have nearly perfect matches in the assembled genome sequences, which indicated the assembly had a high genome coverage. We also aligned 35,187 KOME full-length cDNAs (cloned from Nipponbare) with the Nipponbare genome (rice reference sequence) using the same aligning parameters. Of them, 33,726 cDNAS (95.9%) have their matches onto the Nipponbare genome sequences, which was consistent with a previous estimate that the reference genome sequence covered ~95% of the rice genome (International Rice Genome Sequencing Project, 2005).

The assembled sequence of W1943 was used to perform genome annotation. A total of 65,086 coding genes were generated from *de novo* prediction of the software Fgenesh, with 339 amino acids per predicted protein. Both DNA sequences and protein sequences were aligned against the collection of rice repetitive sequences (Repbase release 16.0). We found that 26,714 predicted genes were transposable element (TE) related. With TE-related genes excluded, we yielded a total of 38,372 non-TE-related gene models of wild rice. The protein sequences were aligned with the

protein sequences of Nipponbare (MSU annotation 7.0), and 33,491 had their corresponding orthologs in Nipponbare, over half of which were changed for at least one amino acid. The left 4,881 coding genes were probably affected by genic indels, large-effect SNPs and other disruptions. Based on the alignment of predicted protein sequences of wild rice W1943 against that of *japonica* cultivar Nipponbare, we found the two representatives of cultivated and wild rice had an average identity of ~99.0% in proteome level. Through clues from 2,045 W1943 full-length cDNAs, gene models generated based on W1943 genome sequences have a better performance in the prediction of gene structure of wild rice.

The final assembly of W1943 was anchored onto rice reference genome sequence to determine the physical locations of the contigs and identify the sequence variants between W1943 and Nipponbare. Of 191,450 contigs, 162,172 contigs (84.7%) were mapped onto 12 rice chromosomes, covering ~321 Mb of the reference genome with an average identity of 97.2%. The left contigs were mostly repetitive sequences, or from physical gaps or large deletions in Nipponbare genome. We also mapped contigs of *indica* 93-11 assembly onto the rice reference genome using the same parameters. Of 50,233 *indica* 93-11 contigs, 32,002 (63.7%) can be mapped onto 12 rice chromosomes, covering ~288 Mb of the reference genome.

Through contig anchor and local sequence alignment, the assembled sequences enabled the detection of nearly all the sequence variants. Only those indels of large size (> 10kb) cannot be identified directly. Sequencing library of long insert size will be constructed, sequenced and used to generate scaffolds and detect larger indels in wild rice (including mated-pair reads and BAC-end sequences).

Section S8. Data release and download links.

Accession codes: Raw sequences have been deposited in the EBI European Nucleotide Archive with accession number ERP001143 for 461 wild rice accessions (including 446 *O. rufipogon* accessions and 15 wild rice accessions of outgroup species), with accession number ERP000729 for the worldwide rice germplasms and with accession number ERP000106 for the Chinese landraces.

The genotype dataset of 1,529 lines of *O. rufipogon* and *O. sativa* can be found at the Rice Haplotype Map Project database (http://www.ncgr.ac.cn/RiceHap3/Genotype). The imputed genotype dataset of 446 *O. rufipogon* accessions for GWAS in wild rice collection is available at http://www.ncgr.ac.cn/RiceHap3/GWAS. The details of the 8 million SNPs, including their position and their ancestral state, are available at http://www.ncgr.ac.cn/RiceHap3/Variant. The *de novo* assembly sequences, BAC sequences, and BLAST searching of *O. rufipogon* W1943 are available at http://www.ncgr.ac.cn/RiceHap3/Assembly.

Supplementary References

1. Han, B. & Xue, Y. Genome-wide intraspecific DNA-sequence variations in rice. *Curr. Opin. Plant Biol.* **6**, 134-138 (2003).
2. Hoban, S., Bertorelle G. & Gaggiotti O. Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.* **13**, 110-122 (2012).
3. Horton, M.W. *et al.* Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212-216 (2012).
4. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793-800 (2005).
5. Jiao, Y. *et al.* Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nat. Genet.* **42**, 541-544 (2010).
6. Miura, K. *et al.* OsSPL14 promotes panicle branching and higher grain productivity in rice. *Nat. Genet.* **42**, 545-549 (2010).
7. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-918 (2007).
8. Sasaki, A. *et al.* A mutant gibberellin-synthesis gene in rice. *Nature* **416**, 701-702 (2002).

9. Sharma, S.D., Tripathy, S. & Biswal, J. Origin of *O. sativa* and its ecotypes. In: Nanda JS, ed. Rice Breeding and Genetics: Research Priorities and Challenges. Science Publications, Enfield. NH. pp. 349–369 (2000).

10. Song, X.J. *et al.* A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* **39**, 623-630 (2007).

11. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. *ssp. indica*). *Science* **296**, 79–92 (2002).

12. 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).

Supplementary Table 1 Sampling and sequence differences detected in phylogenetic relationship reconstruction of AA-genome complex.

Species	Groups	Sampled accessions	Sequence differences (per kb)*
O. sativa ssp indica	*Indica*	Guangluai-4	4.6
O. sativa ssp indica	*Aus*	GP110	5.3
O. sativa ssp japonica	*Temperate japonica*	Nipponbare	0.3**
O. sativa ssp japonica	*Tropical japonica*	GP640	1.9
O. sativa ssp japonica	*Aromatic*	GP294	2.0
O. rufipogon-1	*Or*-I group	W0610	7.9
O. rufipogon-2	*Or*-I group	W3105	6.6
O. rufipogon-3	*Or*-III group	W1943	5.0
O. rufipogon-4	*Or*-II group	W0600	7.9
O. glaberrima	NA	W3104	8.9
O. barthii	NA	W3106	9.0
O. longistaminata-1	NA	W3101	18.9
O. longistaminata-2	NA	W3102	18.0
O. meridionalis-1	NA	W2113	19.0
O. meridionalis-2	NA	W2115	19.0

*Single nucleotide differences between the sampled accession and the reference Nipponbare genome. **The differences (0.3‰) detected between Nipponbare sequenced in this study and the reference Nipponbare genome sequence implied a low rate of sequencing and genotype calling error.

Supplementary Table 3 The fraction of heterozygous genotype calls in five wild rice genomes with high sequencing coverage.

Accession ID	Sequencing Coverage	Heterozygous calls	Heterozygosity proportion*
W1849	9.77	52,152	0.7%
W1943	100.00	105,535	1.3%
W0600	27.30	432,720	5.4%
W3105	40.00	14,486	0.2%
W3106	42.58	9,081	0.1%

*We also selected the wild rice accessions with sequencing coverage > 9, and investigated the heterozygosity based on the overlapped reads that were aligned onto the reference sequence. We calculated the proportion of heterozygosity genotypes at 7,970,358 polymorphic sites.

Supplementary Table 4 The allele frequency spectrum of SNPs segregating in both *O. rufipogon* and *O. sativa*.

MAF in *Or*[*]	Total number	Subset in *Os*[**]	< 0.1 in *Os*	0.1-0.2 in *Os*	0.2-0.3 in *Os*	0.3-0.4 in *Os*	0.4-0.5 in *Os*
<0.1	759,224	494,145	62%	13%	8%	8%	10%
0.1-0.2	731,104	624,289	45%	14%	11%	12%	18%
0.2-0.3	416,440	381,429	34%	16%	13%	14%	23%
0.3-0.4	309,375	291,997	30%	16%	13%	15%	27%
0.4-0.5	256,799	246,540	26%	15%	13%	15%	31%

[*]Minor allele frequency of SNPs segregating in the population of 446 *O. rufipogon* accessions.

[**]The subset of SNPs that also segregate in the population of 1,083 *O. sativa* accessions.

Supplementary Table 5 Genome-wide detection of highly differentiated loci in *O. rufipogon* population.

Chromosome	Interval (Mb, IRGSP 4)	F*st*	Known gene or loci
1	13.1-13.4	0.32	
1	15.9-16.3	0.39	
1	23.1-23.3	0.33	
1	28.6-28.7	0.31	
1	34.3-34.4	0.31	
1	41.2-41.3	0.3	
2	8-8.3	0.31	
2	9.3-9.4	0.36	
2	11.9-12	0.32	
2	12.5-12.8	0.41	
2	15.3-16.2	0.35	
2	17.5-17.7	0.4	
2	18.9-19	0.38	
2	28.3-28.4	0.32	
2	32.2-33.1	0.35	
2	33.6-33.7	0.33	
3	1-1.3	0.33	*OsSOC1*
3	2.7-2.8	0.32	
3	5.1-5.2	0.31	
3	6.2-6.3	0.32	
3	8.9-9.1	0.35	
3	11.8-11.9	0.32	
3	29.1-29.3	0.36	
4	9.8-9.9	0.42	
4	10.2-10.3	0.33	
4	15.6-15.8	0.4	
4	25-25.1	0.3	
4	27.9-28	0.31	
5	15.3-16.2	0.33	
5	21.6-21.7	0.31	
5	26.1-26.3	0.32	
5	29.1-29.2	0.32	
6	4.2-4.4	0.36	*DPL2*
6	8-8.1	0.35	
6	9.7-9.8	0.35	
6	12.3-12.4	0.3	
6	14.2-14.7	0.66	
6	19.4-19.5	0.31	
6	23.4-23.5	0.35	

6	31.5-31.6	0.32	
7	8.1-8.2	0.3	
7	9.1-9.2	0.32	*Ghd7*
7	13.9-14	0.32	
7	14.8-14.9	0.33	
7	15.5-15.6	0.33	
7	18.2-18.3	0.35	
7	19.8-19.9	0.31	
7	25.7-25.8	0.31	
8	13-13.3	0.37	
8	23.3-23.4	0.31	
9	0.8-1.2	0.31	
9	18-18.1	0.3	
10	0.5-1	0.32	
10	10.7-10.8	0.32	
10	14.5-14.6	0.35	
10	15.7-15.8	0.31	
10	20.7-21	0.42	
11	8.1-8.2	0.36	
11	9.9-10	0.38	
11	23.4-23.5	0.32	
12	5.9-6	0.32	*qCTS12*[*]
12	9-9.1	0.32	
12	13.3-14	0.42	
12	14.4-14.5	0.33	
12	16.9-17.1	0.36	
12	19.1-19.2	0.3	
12	20.3-20.6	0.44	
12	20.8-20.9	0.35	

[*]The causal gene has not yet been cloned.

Supplementary Table 6 Specificities and missing data proportions of the genotype dataset of *O. rufipogon* population before and after missing genotypes are inferred.

Type	Before imputing	After imputing
The subset of 446 *O. rufipogon* accessions (2,472,942 SNPs with MAF > 0.05)		
Specificity		
W1943 Illumina data (>20x)	98.7%	97.6%
W0600 Illumina data (>20x)	98.8%	96.8%
W1943 BAC sequences	99.6%	98.7%
Sensitivity	45.8%	-
Missing data rate	50.7%	7.1%
The full set of 1,529 rice accessions (1,614,114 SNPs with MAF > 0.02)*		
Specificity		
Guangluai-4 20x Illumina data	99.7%	99.3%
Guangluai-4 BAC sequences	99.5%	99.3%
Nongken-58 14x Illumina data	99.9%	99.7%
Nipponbare reference genome	99.9%	99.3%
W1943 20x Illumina data	99.2%	98.2%
W0600 20x Illumina data	99.4%	98.2%
W1943 BAC sequences	99.7%	99.4%
Missing data rate	56.4%	3.1%

*The imputation of the full genotype dataset was performed for the GWAS in the full population of *O. sativa–O. rufipogon* complex and will be used in the future study, which was not used in this study. In all population genetic and phylogenetic analyses, only raw genotypes of all the SNPs (before imputing) were used.

Supplementary Table 8 The list of 15 accessions of outgroup species sampled in the collection.

Accession ID	Species	Original producing area
W3104	*O. glaberrima*	Liberia
W3106	*O. barthii*	Burkina Faso
W3101	*O. longistaminata*	Mali
W3102	*O. longistaminata*	Benin
W0636	*O. meridionalis*	Tavoy, Burma
W1239	*O. meridionalis*	Sakor River, Neth. New Guinea
W2111	*O. meridionalis*	Queensland, Australia
W1235	*O. meridionalis*	Taram, Neth. New Guinea
W2117	*O. meridionalis*	13.07 N, 142.07 E, Queensland, Australia
W2075	*O. meridionalis*	15.51 N, 128.44 E, Australia
W1982	*O. meridionalis*	Bogor and Jakarta, Indonesia
W2115	*O. meridionalis*	Queensland, Australia
W0599	*O. meridionalis*	Padang Kemunting, Malaya
W2113	*O. meridionalis*	Queensland, Australia
W2110	*O. meridionalis*	Queensland, Australia

Supplementary Table 9 Genome-wide detection and functional annotation of selective sweep regions in the full population.

Chr	Interval (Mb, IRGSP 4)	Peak (Mb)	π_w/π_c	Associated traits	Known genes (Position)
1	1.2-1.4	1.2	5.3		
1	7.9-8.8	8.3	8.0	Grain width/Stigma exsertion	
1	22.4-23.3	22.7	6.1		
1	28.2-28.3	28.2	3.9		
2	1.1-2.0	1.9	8.0		
2	3.5-3.8	3.5	6.1		
2	18.4-18.5	18.4	5.2		
2	23.8-23.9	23.8	3.6	Stigma exsertion	
3	0.0-1.1	0.0	4.5		
3	12.7-13.2	12.9	4.0		
3	25.5-27.1	26.9	9.9	Panicle length	
3	29.3-29.9	29.8	5.4	Grain number	
3	33.2-33.3	33.2	4.2		
3	34.4-34.5	34.4	3.4		
3	35.8-35.9	35.8	3.7		
4	12.8-13.0	12.8	4.2		
4	15.2-17.5	16.8	10.3	Germination rate	
4	18.3-18.4	18.3	3.4	Awn length	
4	21.1-21.3	21.1	3.9		
4	21.7-21.8	21.7	5.4		
4	23.3-23.8	23.3	6.4	Hull color	*Bh4* (23.3M)
4	25.8-27.5	27.4	9.4		
4	34.3-34.7	34.5	3.6	Shattering	*sh4* (34.6M)
5	4.1-4.2	4.1	3.1		
5	4.8-5.4	4.8	3.9	Grain width	*qSW5* (5.3M)
5	9.1-13.5	10.9	5.6		Centromere
5	29.7-29.9	29.8	5.7		
6	3.7-4.0	3.9	3.9	Stigma exsertion	
6	5-5.6	5.5	4.3	Stigma color	*OsC1* and *S5* (5.3M)
6	22.7-22.8	22.7	3.1		
6	28.5-28.7	28.5	11.6	Grain weight/length	
7	2.4-3.0	2.9	5.9	Tiller angle	*Prog1* (2.8M)
7	3.7-4.1	3.7	7.8		
7	4.9-5.2	5	4.6	Hull color/Stigma exsertion	
7	6.1-6.3	6.2	3.3		
7	6.5-6.6	6.5	3.5		

7	18.2-18.3	18.2	7.2	
7	28.8-29.0	28.9	4.8	Grain length
8	0.9-1.1	0.9	3.1	Tiller angle
8	2.5-2.7	2.5	4.8	
8	23.8-24.9	23.9	13.9	Stigma exsertion/awn length
8	25.3-25.4	25.3	3.1	
8	25.8-25.9	25.8	3.6	
9	8.8-9.0	8.8	4.2	
9	14.5-14.7	14.5	7.0	
9	21.4-21.5	21.4	5.1	
11	8.9-9.0	8.9	3.2	
11	9.9-10.1	9.9	4.8	
12	1.2-1.3	1.2	3.1	
12	2.6-2.9	2.6	5.9	
12	4.3-4.4	4.3	3.3	
12	5.2-5.3	5.2	4.4	
12	20.3-20.6	20.4	5.5	
12	21.4-21.5	21.4	3.1	
12	25.5-25.9	25.6	4.8	

Supplementary Table 10 Genome-wide detection and functional annotation of selective sweep regions in *indica* population.

Chr	Interval (Mb, IRGSP 4)	Peak (Mb)	$\pi_{or\text{-}1}/\pi_{Ind}$	Associated traits	Known genes (Position)
1	1.2-1.3	1.2	3.7		
1	7.4-7.6	7.4	3.7		
1	7.9-8.7	8.0	5.8	Grain width/Stigma exsertion	
1	9.7-9.8	9.7	3.4		
1	22.7-22.8	22.7	4.6		
1	27.9-28.3	27.9	4.2		
1	31.4-31.5	31.4	3.1		
1	37.6-37.7	37.6	3.3		
1	39.7-39.8	39.7	3.4		
2	1.6-2	1.6	7.2		
2	3.5-3.8	3.5	7.2		
2	6.3-6.4	6.3	3.2		
2	11.6-17	16.3	5.5		
2	23.8-23.9	23.8	3.5	Stigma exsertion	
3	4-4.2	4.0	4.5		
3	10.8-11	10.8	4.1		
3	11.6-11.7	11.6	1.6		
3	13.1-13.2	13.1	4.1		
3	25.9-27	26.3	10.7	Panicle length	
3	29.4-29.9	29.8	3.4	Grain number	
3	33.2-33.3	33.2	3.0		
4	15.3-17.5	16.8	9.0	Germination rate	
4	21.7-21.8	21.7	3.8		
4	23.3-23.5	23.4	4.9	Hull color	*Bh4* (23.3M)
4	25.8-27.5	25.8	8.7		
4	34.3-34.7	34.3	5.7	Shattering	*sh4* (34.6M)
5	2.5-2.9	2.8	4.2		
5	3.6-3.7	3.6	3.3		
5	5-5.4	5.1	4.8	Grain width	*qSW5* (5.3M)
5	9.2-13.3	10.9	3.6		Centromere
5	17.3-18.1	17.9	3.4		
5	19.9-21.1	20.7	3.3		
5	28.3-28.4	28.3	4.0		
5	29.7-29.9	29.8	4.2		
6	5.5-5.6	5.5	5.4	Stigma color	*OsC1* and *S5* (5.3M)
6	9.4-9.6	9.4	4.2		
6	21.6-23.1	22.9	3.8		

6	28.5-28.6	28.5	7.3	Grain weigth/length	
6	32.0-32.1	32.0	4.1		
7	2.4-3.2	2.9	3.9	Tiller angle	*Prog1* (2.8M)
7	3.7-4.2	4.0	8.2		
7	4.9-5.1	5.0	3.7	Hull color/Stigma exsertion	
7	28.9-29.1	28.9	4.1	Grain length	
7	29.6-30	29.8	5.8	Awn length/heading date	
8	2.5-2.7	2.5	4.0		
8	15.8-15.9	15.8	3.2		
8	16.3-16.4	16.3	3.2		
8	21.1-22.1	21.6	4.5	Grain width	
8	23.7-25.0	23.9	9.9	Stigma exsertion/awn length	
8	25.9-26.0	25.9	3.2		
9	13.8-14.6	14.5	5.0		
9	21.4-21.5	21.4	3.5		
9	23.5-23.6	23.5	3.5		
10	18.1-18.3	18.2	3.7		
10	19.1-19.2	19.1	3.1		
10	21.7-21.8	21.7	3.2	Grain width	
10	23.5-23.6	23.5	3.0		
11	1.3-1.4	1.3	3.8		
12	2.6-2.9	2.6	4.4		
12	25-25.8	25.2	9.4	Grain weight	

Supplementary Table 11 Genome-wide detection and functional annotation of selective sweep regions in *japonica* population.

Chr	Interval (Mb, IRGSP 4)	Peak (Mb)	$\pi_{or\text{-}III}/\pi_{Jap}$	Associated traits	Known genes (Position)
1	21.8-23.8	21.9	56.0		
1	32.7-32.8	32.7	14.6		
1	37.8-38.3	38.1	14.0	Shattering	*qSH1* (38.2M)
1	39.5-39.6	39.5	14.3		
1	41.9-42	41.9	15.5		
1	42.9-43	42.9	14.5		
2	0.9-2	1.4	32.1		
2	5.5-5.6	5.5	19.1		
2	6.2-6.3	6.2	14.8		
2	10.7-11.7	11.0	31.8		
2	12.7-12.8	12.7	14.0		
2	22.3-22.4	22.3	14.7		
2	28.0-28.5	28.0	17.3		
2	31.6-33.0	31.6	33.5	Grain length	
3	0.2-0.5	0.3	18.0		
3	1.7-1.8	1.7	14.9		
3	9.7-10.1	9.7	24.9		
3	12.7-12.8	12.7	14.7		
3	21.7-21.8	21.7	14.6		
3	23.4-25.7	23.6	21.1		
3	29.4-35.6	34.4	24.5	Grain number	
4	21.7-21.8	21.7	17.2		
4	23.3-26.3	24.4	30.1	Hull color	*Bh4* (23.3M)
4	27.5-29	28.0	24.2		
4	34.6-34.9	34.7	20.3	Shattering	*sh4* (34.6M)
4	35.4-35.9	35.8	21.6		
5	8.9-9.4	9.3	17.8		
5	11.0-11.1	11.0	18.7		
5	13.0-13.1	13.0	14.4		
5	14.4-14.6	14.5	16.5		
5	16.6-17	16.7	26.8		
5	18.3-18.8	18.5	23.2		
5	20.7-20.8	20.7	14.9		
5	22.8-22.9	22.8	14.2		
5	25-25.2	25.1	17.7		
5	25.7-25.9	25.7	16.7		
6	1.7-1.8	1.7	15.1		*Waxy* (1.7M)
6	7.7-7.8	7.7	17.8		
6	28.4-28.6	28.5	20.2	Grain length/grain	

				weight	
7	0.4-0.8	0.7	20.4		
7	3.3-3.4	3.3	15.5		
7	5.8-6.1	5.8	17.6		*Rc* (6.1M)
7	26.0-26.6	26.3	18.0	Tiller angle/panicle length/grain number/grain length	
8	0.6-0.7	0.6	15.0	Tiller angle	
8	2.1-2.2	2.1	17.4		
8	23.9-24	23.9	21.8	Stigma exsertion/awn length	
9	12.4-13.5	12.4	30.4		
9	15.4-15.6	15.5	18.2		
9	16.1-17.1	16.7	27.7	Grain weight	
9	17.9-18	17.9	17.5		
9	18.9-19	18.9	19.2		
9	20.3-23.6	20.9	29.0	Grain number	
10	1.7-1.8	1.7	14.6		
10	14.2-14.6	14.2	21.0		
10	22.4-23	22.4	25.6		
11	0.8-0.9	0.8	21.5		
11	2.4-3.1	2.4	21.7		
11	15.8-15.9	15.8	16.1		
12	0.0-0.2	0.0	16.4		
12	2.1-3.6	2.5	20.5		
12	26.8-26.9	26.8	15.9		
12	27.6-27.7	27.6	16.2		

Supplementary Table 12 Large-scale genetic mapping of fifteen domestication traits identifies 58 QTLs in an *O. sativa* × *O. rufipogon* population through sequencing-based genotyping.

Traits	Chr	Peak bin	Position (Mb)	R^2	Known gene	Selection signal	Analysis panel
Tiller angle	7	bin_1054	2.7	10.9%	*Prog1*	5.9	Full population
	7	bin_1141	25.6	7.1%		18.0	*Japonica* panel
	8	bin_1181	0.7	4.6%		3.1	Full population
	3	bin_0589	36.9	4.5%			
Stigma exsertion	6	bin_0923	4.5	8.2%		3.9	Full population
	2	bin_0351	24.0	6.6%		3.6	Full population
	8	bin_1278	23.5	5.2%		13.9	Full population
Stigma color	6	bin_0929	5.4	72.2%	*OsC1*	4.3	Full population
Shattering	4	bin_0730	34.6	40.2%	*sh4*	3.6	Full population
Pericarp color	7	bin_1072	5.9	93.5%	*Rc*	17.6	*Japonica* panel
Panicle length	7	bin_1144	26.0	14.1%		18.0	*Japonica* panel
	1	bin_0178	39.2	8.1%			
	3	bin_0522	26.3	5.2%		9.9	Full population
Hull color	4	bin_0679	23.1	57.4%	*Bh4*	6.4	Full population
	7	bin_1068	4.9	2.8%		4.6	Full population
	6	bin_0913	3.3	2.1%			
Grain width	5	bin_0773	5.2	18.3%	*qSW5*	3.9	Full population
	5	bin_0780	6.4	17.5%			
	7	bin_1145	26.1	13.8%		18.0	*Japonica* panel
	10	bin_1515	21.6	10.4%		3.2	*Indica* panel
	1	bin_0054	8.1	6.2%		8.0	Full population
	1	bin_0031	5.1	5.2%			

	6	bin_0971	13.1	2.9%			
	8	bin_1269	21.2	2.8%		4.5	*Indica* panel
Grain weight	6	bin_1026	28.4	11.6%		11.6	Full population
	12	bin_1775	25.1	5.4%		9.4	*Indica* panel
	9	bin_1372	16.0	4.7%		27.7	*Japonica* panel
Grain number	7	bin_1155	27.4	14.9%			
	3	bin_0545	29.6	7.1%		5.4	Full population
	7	bin_1144	26.0	6.5%		18.0	*Japonica* panel
	9	bin_1403	20.4	6.1%		29.0	*Japonica* panel
	7	bin_1050	1.5	5.7%			
	7	bin_1094	15.4	5.5%			
	8	bin_1216	6.1	4.3%			
Grain length	6	bin_1026	28.4	20.5%		11.6	Full population
	7	bin_1143	25.9	12.8%		18.0	*Japonica* panel
	7	bin_1153	27.0	11.9%			
	7	bin_1165	28.8	9.8%		4.8	Full population
	2	bin_0376	32.4	7.4%		33.5	*Japonica* panel
	6	bin_0939	6.6	5.8%			
	1	bin_0174	36.9	4.4%			
Germination rate	9	bin_1397	19.8	10.1%			
	4	bin_0645	16.8	4.6%		10.3	Full population
Awn length	4	bin_0673	22.1	16.8%			
	4	bin_0660	19.7	16.7%			
	8	bin_1281	24.2	13.6%		13.9	Full population
	4	bin_0650	18.5	9.8%		3.4	Full population
	4	bin_0685	24.0	5.7%			
	7	bin_1175	30.0	5.3%		5.8	*Indica* panel
	10	bin_1497	18.7	3.2%			
Heading date	7	bin_1088	10.0	22.3%	*Ghd7*		

	7	bin_1175	30.0	20.3%	*Hd2*	5.8	*Indica* panel
	8	bin_1206	4.1	8.3%	*Hd8*		
	10	bin_1477	15.3	5.3%	*Ehd2*		
Plant height	1	bin_0182	40.2	44.5%	*sd-1*		
	7	bin_1158	27.8	7.7%			
	8	bin_1225	9.0	3.1%			
	7	bin_1094	15.4	3.1%			

Supplementary Table 13 Co-localization of QTLs for domestication traits.

Trait	Chromosome	Position (Mb)	Effect size	Known gene
Plant height	1	40.2	44.50%	*sd-1*
Panicle length	1	39.2	8.10%	
Grain weight	6	28.4	11.60%	
Grain length	6	28.4	20.50%	
Plant height	7	15.4	3.10%	
Grain number	7	15.4	5.50%	
Grain length	7	25.9	12.80%	
Panicle length	7	26	14.10%	
Grain number	7	26	6.50%	
Grain width	7	26.1	13.80%	
Grain length	7	27	11.90%	
Grain number	7	27.4	14.90%	
Heading date	7	30	20.30%	
Awn length	7	30	5.30%	
Stigma exsertion	8	23.5	5.20%	
Awn length	8	24.2	13.60%	

Supplementary Table 14 Summary of whole-genome *de novo* assembly of a wild rice accession (*O. rufipogon* W1943)

Item	Statistic
Raw data for whole-genome assembly	
Raw data used in assembly (Gb)	40
Average coverage (x)	100
Contigs from whole-genome assembly	
Total length (Mb)	406.1
Total number	191,450
N50 length (bp)	15,593
Largest (bp)	154,320
Average length (bp)	2,121
Predicted genes from assembly	
Non TE-related genes	38,372
TE-related genes	26,714
Sequence variants detected*	
SNPs	2,621,077
Small indels (≤20bp)	619,132
Structural variants (>20bp)	140,075
Variants with effects on genes	
Non-sysnonymous SNPs	128,010
SNPs that introduced stop codons	3,580
SNPs that disrupt stop codons	1,127
SNPs that disrupt initial codons	597
SNPs that disrupt splice sites	1,309
Frameshift indels	10,386
Non-frameshift indels	32,237

* Between *O. rufipogon* W1943 and *O. sativa japonica* Nipponbare