Supporting Information

SI Results

SI Result 1. Features of SNPs and genotyping accuracy

For the 6,551,358 high quality SNPs with the minor alleles shared by at least five accessions, approximately 38.0% of SNPs were found with the frequencies of the minor alleles < 0.05, and 34.5% of SNPs were located in genic regions (i.e., exon and intron according to the annotation).

The variety *japonica* cv. Nipponbare with reference genome and two *indica* cv. Zhenshan 97 and Minghui 63, which have relatively high quality sequences available were included in the sample and thus allowed to assess the accuracy of our data processing. The 14-fold Illumina sequences of Zhenshan 97 and Minghui 63 are available in NCBI Sequence Read Archive with accession number SRA012177, and genotypes of consensus sequences were generated using SAMtools (with parameters mpileup -ADSu -C50 -Q20 -q40) and BCFtools. Genotype calls of low-coverage sequencing of Nipponbare, Zhenshan 97 and Minghui 63 before imputing and after imputing were compared with the above-mentioned high quality sequencing data respectively. The percentage of concordant sites gave the estimates of accuracy (**Table S1**). We also genotyped 48 accessions using Illumina Infinium array RiceSNP50 (1). There are 43386 high quality SNP markers in the array and 42759 (98.6%) SNPs were covered by the 6,428,770 well-imputed SNPs. The concordance of genotypes using array hybridization and sequencing can be used to estimate the accuracy of raw genotypes from direct sequencing and after imputation. The results suggested an accuracy of 99.3% (**Table S1**).

SI Result 2. Nucleotide diversity and linkage disequilibrium in different subpopulations

We divided the rice genome into 10 kb segments and estimated the nucleotide diversity (π) in each segment based on the above population classification (**Fig. S2)**. The median diversity of the whole genome was 0.0037 for the whole population, and 0.0018 and 0.00078 for *indica* and *japonica*, respectively. The median diversity levels of *TeJ* and *TrJ* groups (0.00038 and 0.00077) were much lower than *IndI* and *IndII* (0.0013 and 0.0016). The similar sequence diversities of *IndI* and *IndII* indicate that modern breeding process has not altered the overall genetic diversity significantly.

We also calculated Wright's F_{ST} between different subpopulations using SNPs with MAF≥0.05. The population genetic differentiation measured by Wright's F_{ST} between *indica* and *japonica* is high (0.52), similar to previous reports (2). F_{ST} within *indica* and *japonica* are 0.10 and 0.20, respectively.

We further investigated the linkage disequilibrium (LD) based on squared allele-frequency correlations (r^2). There are several studies (3) in LD of rice, and the reported distance of LD decay to 0.2 ranges from < 1 kb to > 1 Mb. To analyze LD decay patterns, SNPs with MAF≥0.05 were randomly selected at density of 3 SNPs per kb and then squared allele-frequency correlations for LD (r^2) of two SNPs within a sliding window of 3 Mb were calculated. The sliding window was moved along each chromosome with step size of 1 Mb. The decay of LD with physical distance was estimated by fitting r^2 to the physical distance between SNPs using lowess. Our results show that the levels of LD in different portions of the rice genome were very heterogeneous. LD decayed rapidly in 11.6% to 41.3% of the genomic regions in different subpopulations, reaching less than 0.2 in only 10 kb (approximate the average gene density of rice) (**Fig. S3**). But in other 4.2% to 33.1% of the regions, LD decay distance was greater than 1 Mb. The median distances of LD decay in *indica* and *japonica* group were 93 kb and 171 kb, respectively. Moreover, *IndII* had far shorter LD decay distance (78 kb) than *IndI* (142 kb), likely due to the efforts of intensive modern breeding. These results also suggested that the resolution of GWAS in rice would be variable for different traits and for different loci associated with the same trait.

SI Result 3. Quantitating the relevance between individual accessions and the allele frequency spectrum of subpopulations using a weighted scoring method

We first calculated the frequency of major allele in each subpopulation at each SNP (denotes as p_{ij}, where i is the SNP index and j is the subpopulation index). For each accession, frequencies of the corresponding alleles in the subpopulation of the SNPs were regarded as scores and averaged:

$$Score_{kj} = \frac{1}{N}\sum_{i=0}^{N}[w_{ik}\ p_{ij} + (1 - w_{ik})(1 - p_{ij})]$$

where w_{ik} is 1 if the allele of accession k in SNP i is the major allele and w_{ik} is 0 if the allele is the minor allele. N is the number of non-missing SNPs in accession k. A greater $Score_{kj}$ means that the contribution of alleles from accession k to population j is bigger, if the accession is really a founder. The distribution of the weighted score for each *indica* accession in different subpopulations is shown in **Figs S5** and **S6**. The accession having the highest score (0.863, Z-score = 2.70, P = 0.007) with the allele frequency spectrum of *IndI-mod* group was still Aijiaonante. And the accession having the highest score (0.792, Z-score = 3.09, P = 0.002) with the allele frequency spectrum of *IndII* group was IR 8.

SI Result 4. Small RNA loci and their targets under selection

Small RNA loci and their targets might also be selected during breeding. We found that osa-miR167j which has targets of *ARF6* and *ARF8* in rice was under selection in *IndI* (Dataset S3). It was shown that miR167, *ARF6* and *ARF8* are involved in auxin-related pathway in developing seeds of rice (4). Interestingly, both this miR167 and another member of the microRNA family (osa-miR167g) were also with high scores of selection in *IndII* (Dataset S4). Although OsmiR156 was not subject to analysis in this study due to its centromere location, a similar miRNA miR529, which shares high homology with miR156 and also targets SBP box genes in plants (5), was significantly selected in both *IndI* and *IndII* (Dataset S3 and S4). We also observed that miR172, which interacts with miR156 to control developmental timing in

Arabidopsis and rice yield trait (6), was selected in *IndII*. Although the selection for *OsSPL14*, a target of OsmiR156 and known to be responsible for rice plant architecture (6), was not significant in both *IndI* and *IndII*, we did find that another two putative targets of both OsmiR156 and OsmiR529, *OsSPL2* and *OsSPL18*, were selected in *IndII* and *IndI* (Dataset S3 and S4), respectively, indicating that they might also play roles in rice breeding.

SI Result 5. GO enrichment analysis for genes under selection

We analyzed whether genes belonging to specific GO categories were more likely to be selected using INRICH (7). The GO classifications of rice genes downloaded from Gramene (http://www.gramene.org) and China Rice Data Center (http://www.ricedata.cn/) were merged. Only the terms in the biological process category with the number of genes between 3 and 150 were used for GO analysis. We first screened 2,106 GO terms using Fisher's exact test and 182 terms with $P < 0.1$ and at least three selected genes in the merged selected regions were examined using INRICH which can perform interval-based enrichment analysis (7). Only the 41,406 non-TE (transposable elements) related genes of rice annotation MSU v6.1 were used as a background gene set. A total of 79 GO gene sets were identified with empirical *P*-value < 0.05 in at least one of three sets of candidate selected regions (*IndI-IndII*, *IndII-IndI* and merged regions) (Dataset S7). In concordance with the results described in the text, we found that genes within the term auxin biosynthetic process (GO:0009851) and indole-containing compound biosynthetic process (GO:0042435) were among the most significant lists, enriched in all three sets of selected regions. Genes involved in hormone biosynthetic process (GO:0042446) and cellular hormone metabolic process (GO:0034754) were enriched in *IndI* and all *indica*, and ones associated with indoleacetic acid biosynthetic process (GO:0009684) and hormone metabolic process (GO:0042445) were enriched in all *indica*. These results suggest probable important roles of genes associated with plant hormones and signal transductions during modern rice breeding. We also found that genes associated with

the term glutamine metabolic process (GO:0006541) were significantly enriched in selected genes of *IndII*, and ones related to glutamine biosynthetic process (GO:0006542) were enriched in the merged regions. It was intriguing to note that genes related to maintenance of DNA methylation (GO:0010216) and DNA methylation (GO:0006306) were very strongly enriched in the merged regions or *IndII*, indicating the likely vital roles of epigenetic modifications of genomic DNA in rice breeding, development and stress responses. Genes involved in circadian rhythm (GO:0007623) were found to be strongly enriched in *IndI*, consistent with the reports that circadian clocks enhance survival, growth vigor, and fitness in plants (8). Ubiquitin-proteasome pathway has been reported to play important roles in regulating rice grain development (9), and interestingly, we also found two genes encoding ubiquitin-conjugating enzyme under the term of negative regulation of developmental process (GO:0051093) were enriched in *IndI*.

SI References

1. Chen H, *et al.* (2014) A High-Density SNP Genotyping Array for Rice Biology and Molecular Breeding. *Mol Plant* 7:541-553
2. Huang X, *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961-967
3. Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. *Mol Biol Evol* 24:875-888
4. Yang JH, Han SJ, Yoon EK, Lee WS (2006) Evidence of an auxin signal pathway, microRNA167-ARF8-GH3, and its response to exogenous auxin in cultured rice cells. *Nucleic Acids Res* 34:1892-1899
5. Cuperus JT, Fahlgren N, Carrington JC (2011) Evolution and functional diversification of MIRNA genes. *Plant Cell* 23:431-442
6. Jiao Y, *et al.* (2010) Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nat Genet* 42:541-544
7. Lee PH, O'Dushlaine C, Thomas B, Purcell SM (2012) INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* 28:1797-1799
8. Dodd AN, *et al.* (2005) Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* 309:630-633
9. Song X-J, Huang W, Shi M, Zhu M-Z, Lin H-X (2007) A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat Genet* 39:623-630
10. Xing YZ, *et al.* (2002) Characterization of the main effects, epistatic effects and their

environmental interactions of QTLs on the genetic basis of yield traits in rice. *Theor Appl Genet* 105:248-257

11. Hua JP, *et al.* (2003) Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc Natl Acad Sci USA* 100:2574-2579

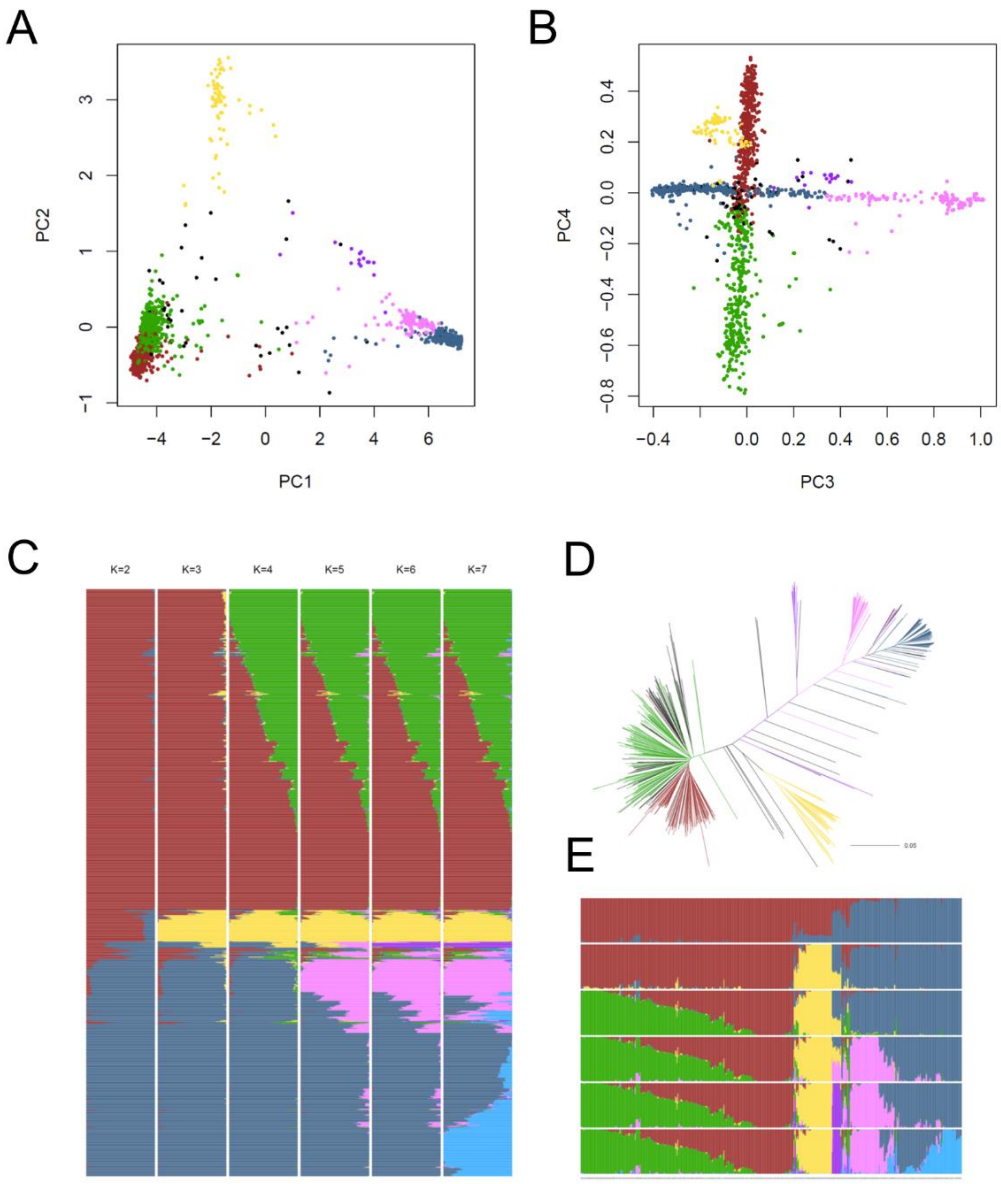


Fig. S1 Population structure of the total of 1479 accessions used in population genetic analysis and 529 accessions sequenced by us. Colors distinguishing different subpopulations are the same as in **Fig. 1**. (**A-B**) Results of the principal components analysis using 188,637 evenly distributed SNPs for 1479 accessions. PC1-4 are the first to fourth principal components. (**C**) The distribution of the estimated subpopulation components for each accession analyzed by ADMIXTURE under different assumptions of ancient clusters K = 2 to 7 for 1479 accessions. (**D**)

Neighbor-joining tree of accessions constructed from matching distance of 188,637 even-distributed SNPs for 529 accessions sequenced by us. (**E**) Same as (**C**) but using only the 529 accessions.

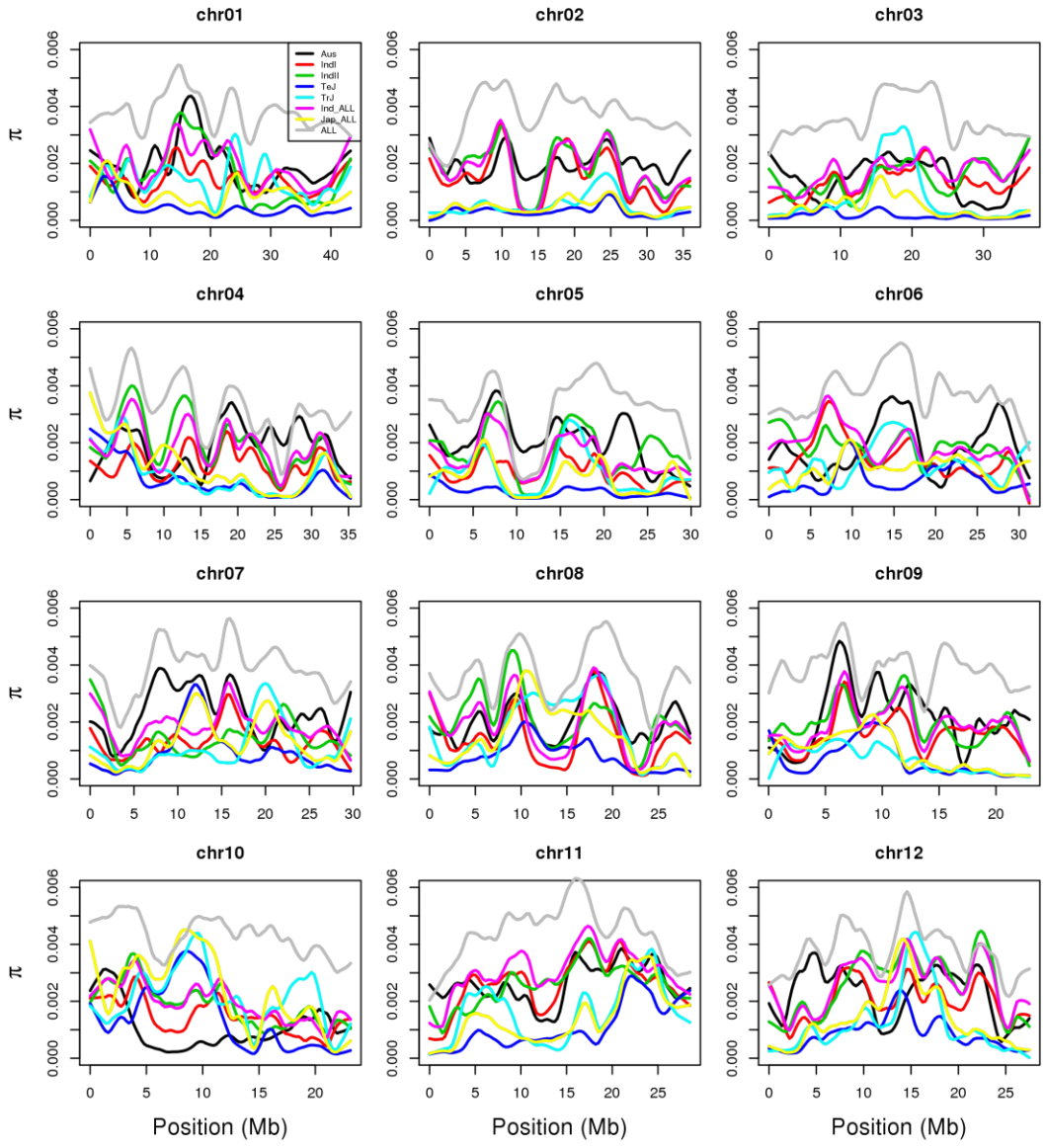


Fig. S2 The distribution of π in different genomic regions and different subpopulations.

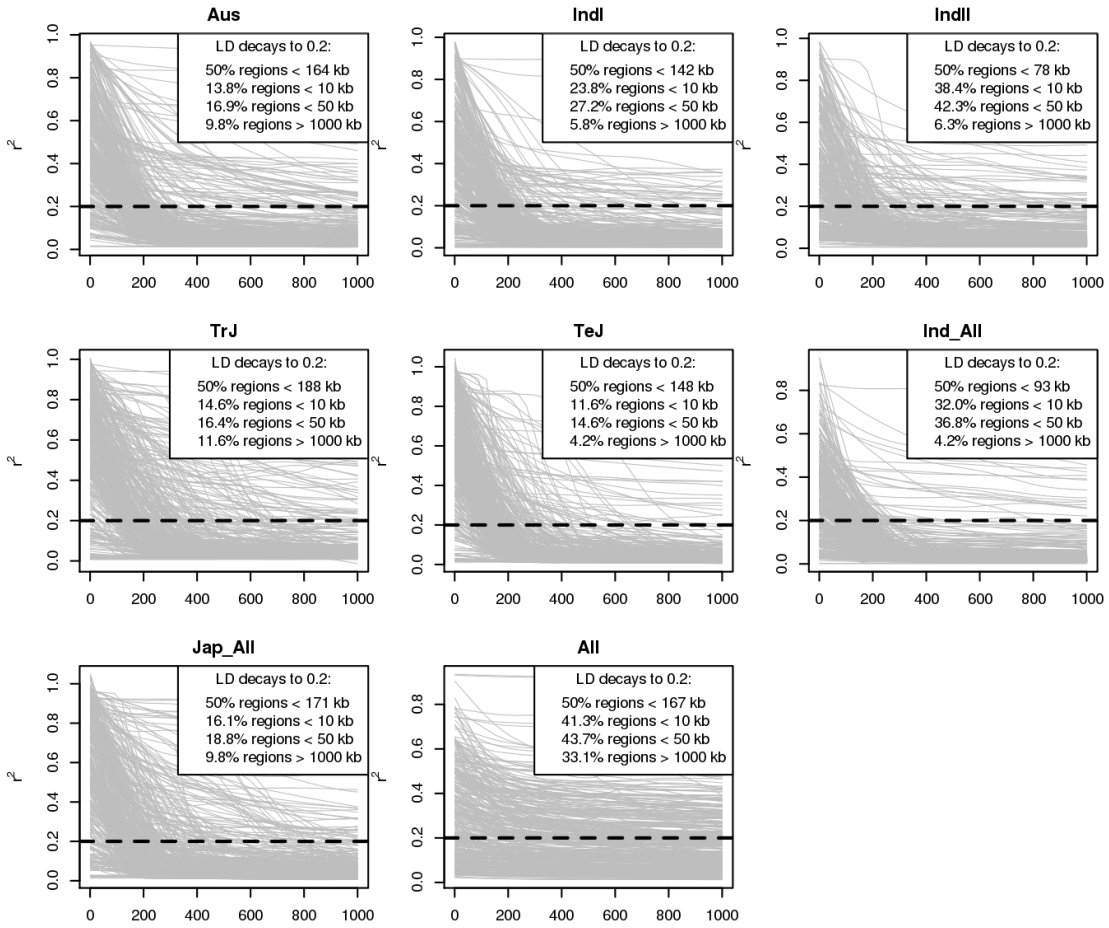


Fig. S3 The distribution of LD decay in different genomic regions and different subpopulations. Each grey line is a LD decay curve in a 3 Mb region.

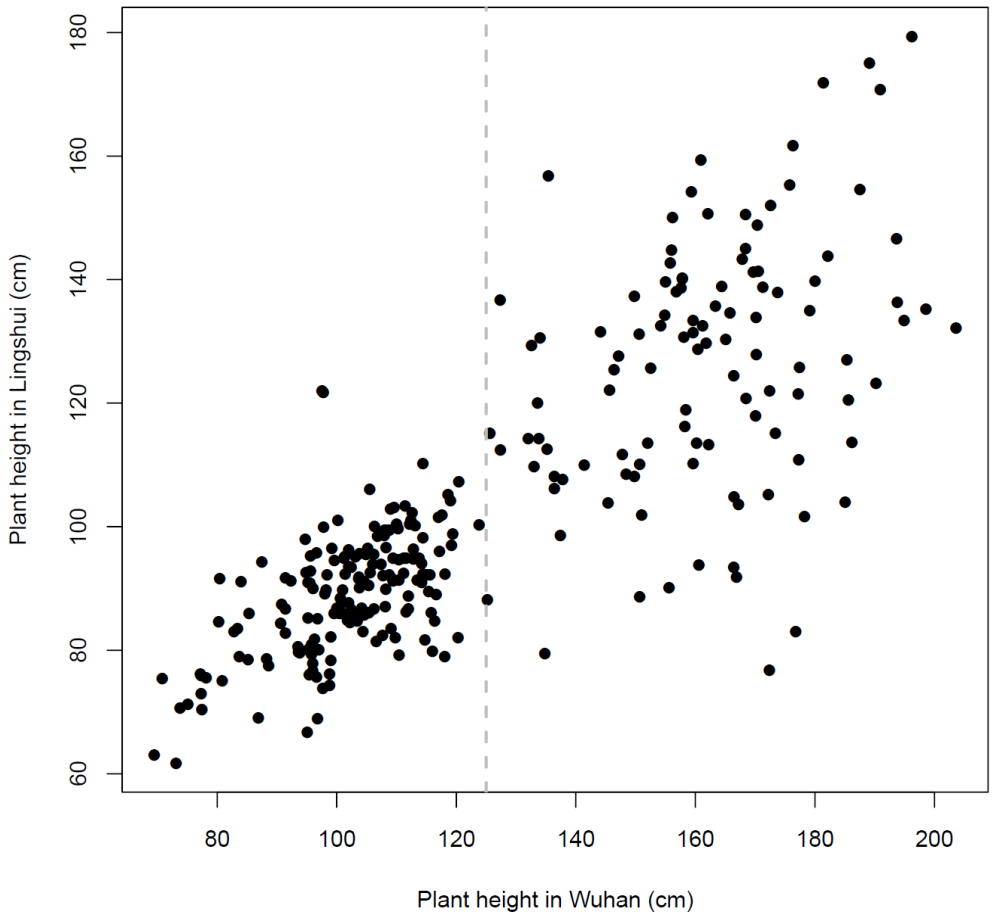


Fig. S4 The distribution of plant height of two field experiments for *indica* accessions. *x*-axis: plant height measured in the summer growing season of 2011 in Wuhan (central China, 30 °28') under a long-day condition (~14h); *y*-axis: plant height measured in Lingshui (southern China, 18 °48') in the spring of 2012 under a relative short-day condition (~11h). Only 282 accessions with available plant height data in both experiments are plotted. The vertical line in gray represents 125 cm.

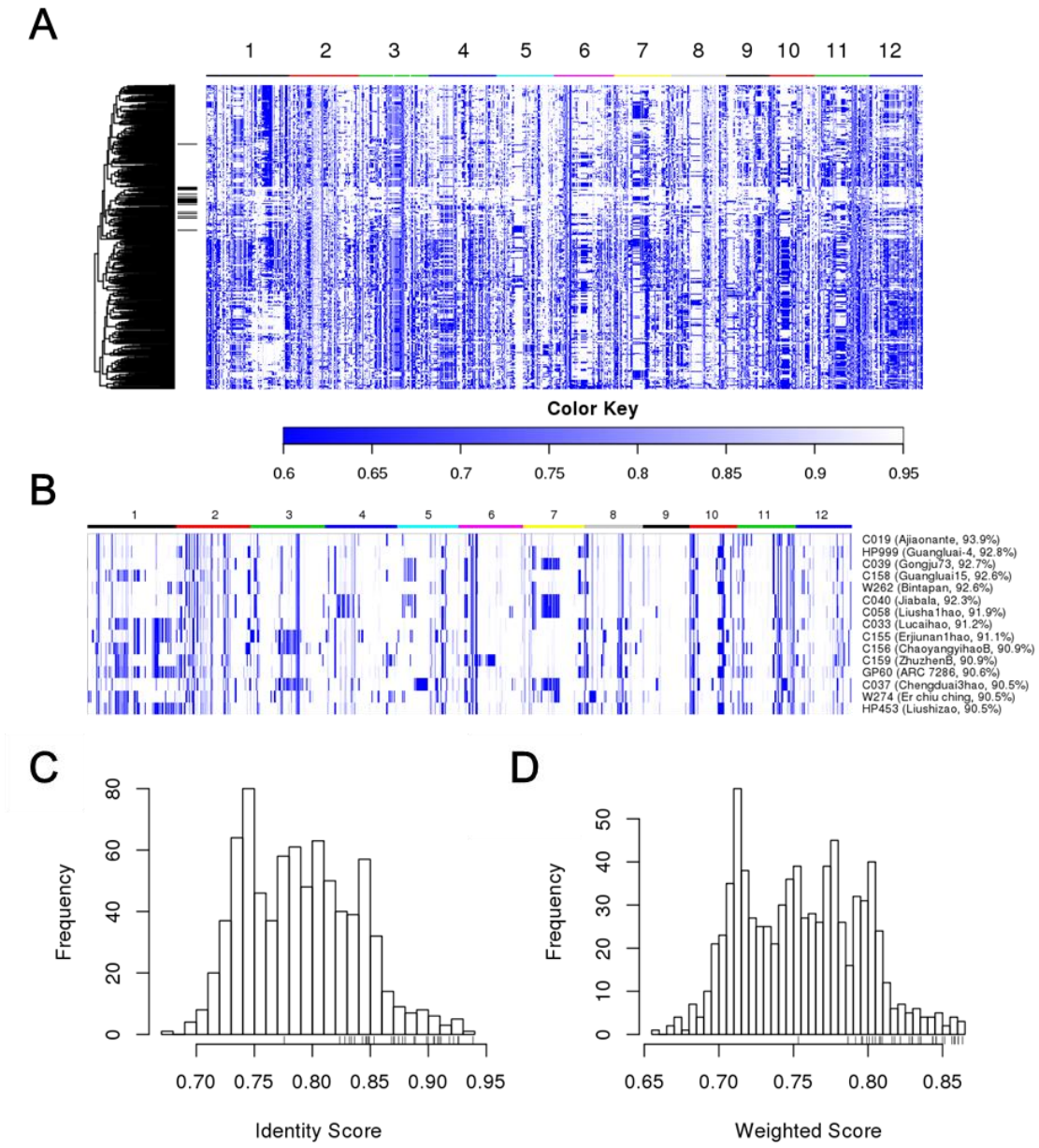


Fig. S5 Similarities between the founder genome of *IndI-mod* and each *indica* accession. (**A**) The genome was divided into 100 kb regions and the proportion of identical genotypes in the region between the founder genome of *IndI-mod* and each accession in *indica* was calculated to draw the heatmap. Larger values (the color close to white) reflect the regions that are more identical to the founder genome. The top horizontal bar indicates rice 12 chromosomes. Each row represents an accession. Rows marked in black are accessions in *IndI-mod*. (**B**) The 15 accessions most similar to the founder genome of *IndI-mod*. The percentage in the parenthesis is the overall proportion of identical genotypes between the founder genome of *IndI-mod* and an

indica accession. (**C**) Distribution of the overall proportion of identical genotypes between the founder genome of *IndI-mod* and each *indica* accession. A rug (near the *x*-axis) in black denotes the marginal distribution of *IndI-mod* accessions. (**D**) Distribution of the overall contribution of each *indica* accession to *IndI-mod* population calculated using the weighted scoring method (SI Result 3).

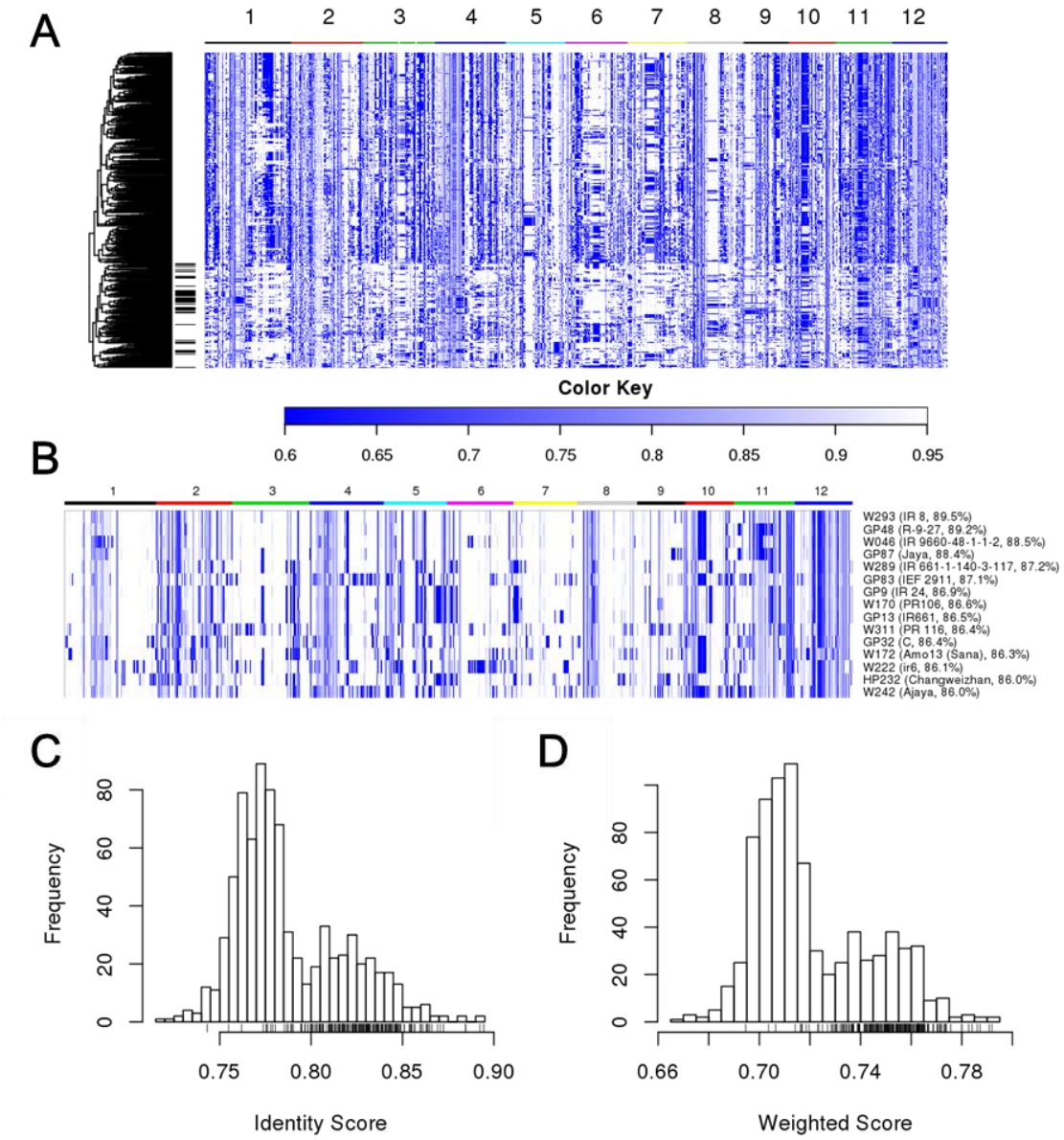


Fig. S6 Similarities between the founder genome of *IndII* and each *indica* accession. (**A**) The genome was divided into 100 kb regions and the proportion of identical genotypes in the region between the founder genome of *IndII* and each accession in *indica* was calculated to draw the heatmap. Larger values (the color close to white) reflect regions to be more identical to the founder genome. The top horizontal bar indicates rice 12 chromosomes. Each row represents an accession. Rows marked in black are accessions released from IRRI. (**B**) The 15 accessions most similar to the founder genome of *IndII*. The percentage in the parenthesis is the overall proportion of identical genotypes between the founder genome of *IndII* and an *indica* accession.

(**C**) Distribution of the overall proportion of identical genotypes between the founder genome of *IndII* and each *indica* accession. A rug (near the *x*-axis) in black denotes the marginal distribution of *IndII* accessions. (**D**) Distribution of the overall contribution of each *indica* accession to *IndII* population calculated using the weighted scoring method (SI Result 3).

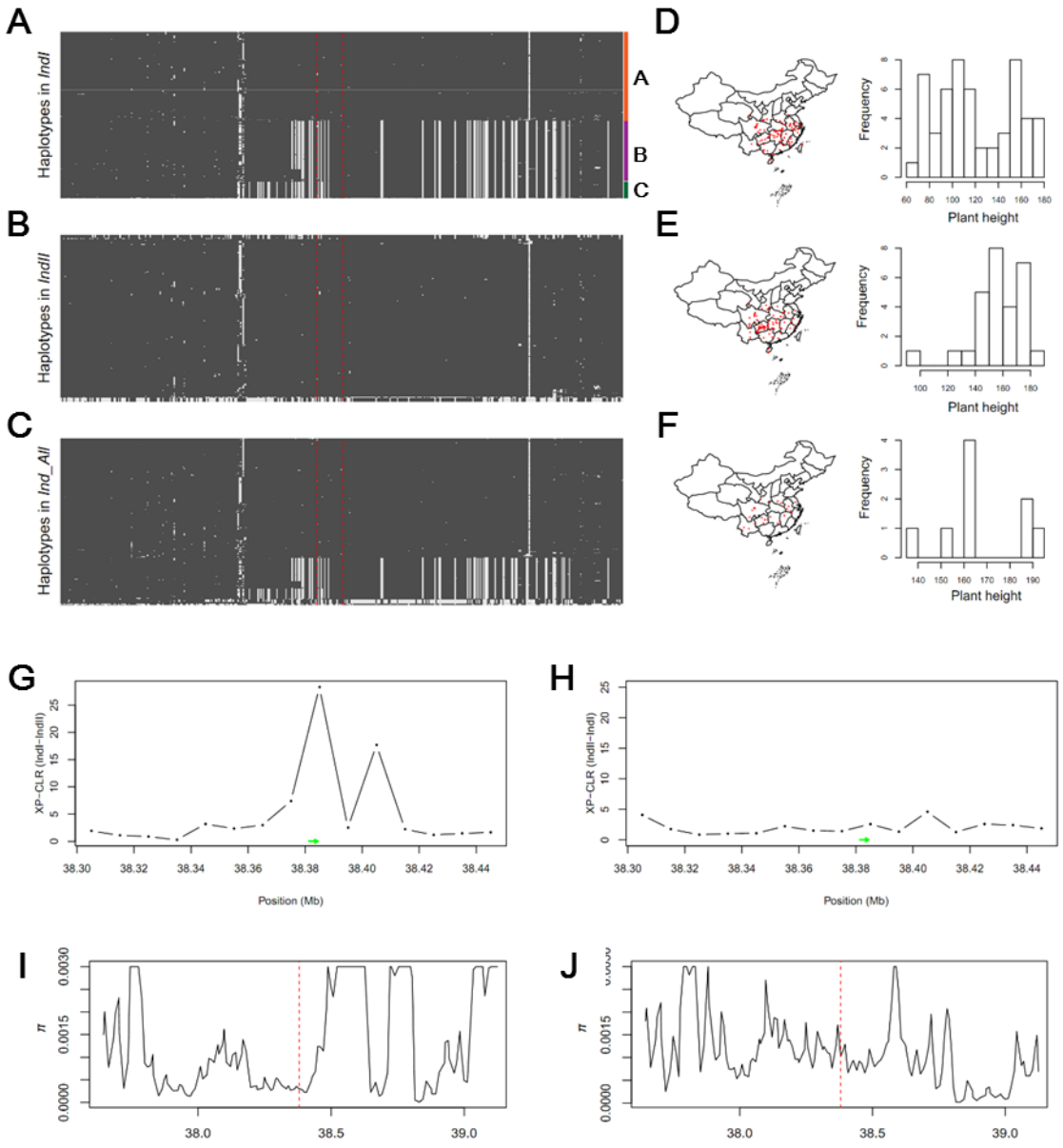


Fig. S7 The haplotypes and diversity in *sd-1* flanking regions and its geographic distribution. (**A-C**) Haplotypes (10 kb regions flanking *sd-1*) of accessions in *IndI*, *IndII* and all *indica*, respectively. The borders of *sd-1* (LOC_Os01g66100) are marked by vertical red lines according to annotations of MSU version 6.1. A total of 448 SNPs were identified in this region. The minor allele in *indica* is denoted in white in panel **A-C** of which each row represents one haplotype. (**D-F**) The geographic distribution and plant height for three main clusters of haplotypes in *IndI*. The corresponding clusters are denoted in (**A**). The haplotype cluster **E** has obvious

preferential geographic distribution in Guizhou Province, southern China, which is probable for adapting to local cultivation conditions. (**G-H**) The XP-CLR score (averaged per 10 kb) around *sd-1* region. The *sd-1* locus is indicated by the green arrow. (**G**) *IndII* as reference population and *IndI* as object population. (**H**) *IndI* as reference population and *IndII* as object population. (**I-J**) The distribution of average diversity (π, averaged per 10 kb with step size of 5 kb) of haplotype cluster e (**I**) and *IndII* (**J**) around *sd-1* region. A diversity value greater than 0.003 was set as 0.003 for presentation.

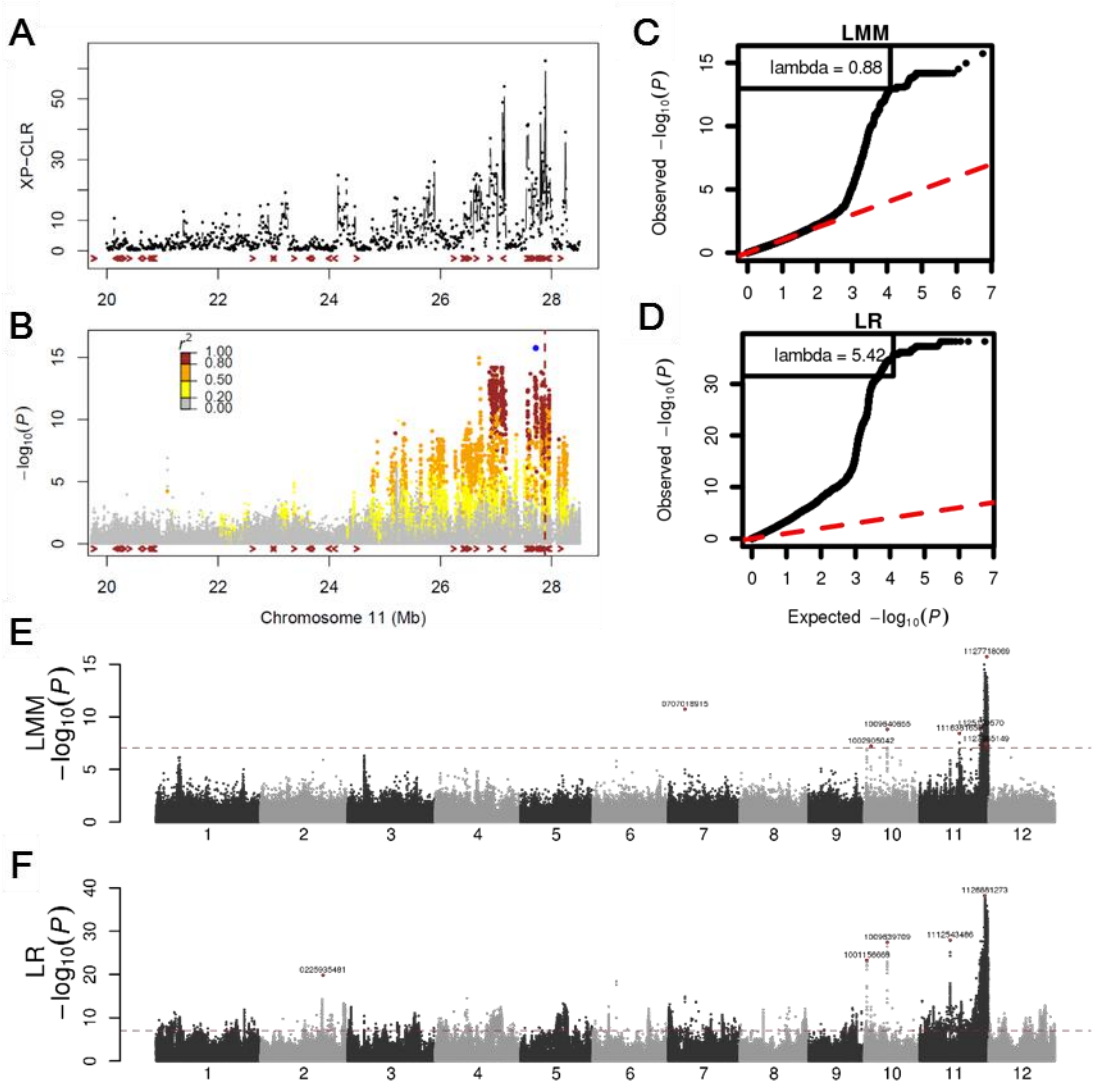


Fig. S8 The XP-CLR near the *Xa4* and *Xa26* region and GWAS results for lesion length of a BB strain PXO341. (**A**) The XP-CLR result near *Xa4* and *Xa26*, using *IndI* as reference population and *IndII* as object population. (**B**) Association result near *Xa26*. The blue point denotes the lead SNP sf1127718069. The colors of the other points represent the linkage disequilibrium for the lead SNP. Each arrow represents a receptor protein kinase gene. The vertical dashed line indicates the position of *Xa26* and *Xa4*. (**C**-**D**) Q-Q plot of the expected null distribution and the observed *p*-value using the mixed model (**c**) and the simple linear regression model (**D**). (**E**-**F**) Genome-wide *p*-values for the mixed model (**E**) and simple linear regression model (**F**). The dashed line in red is the genome-wide significant threshold (8.74×10^{-08}).

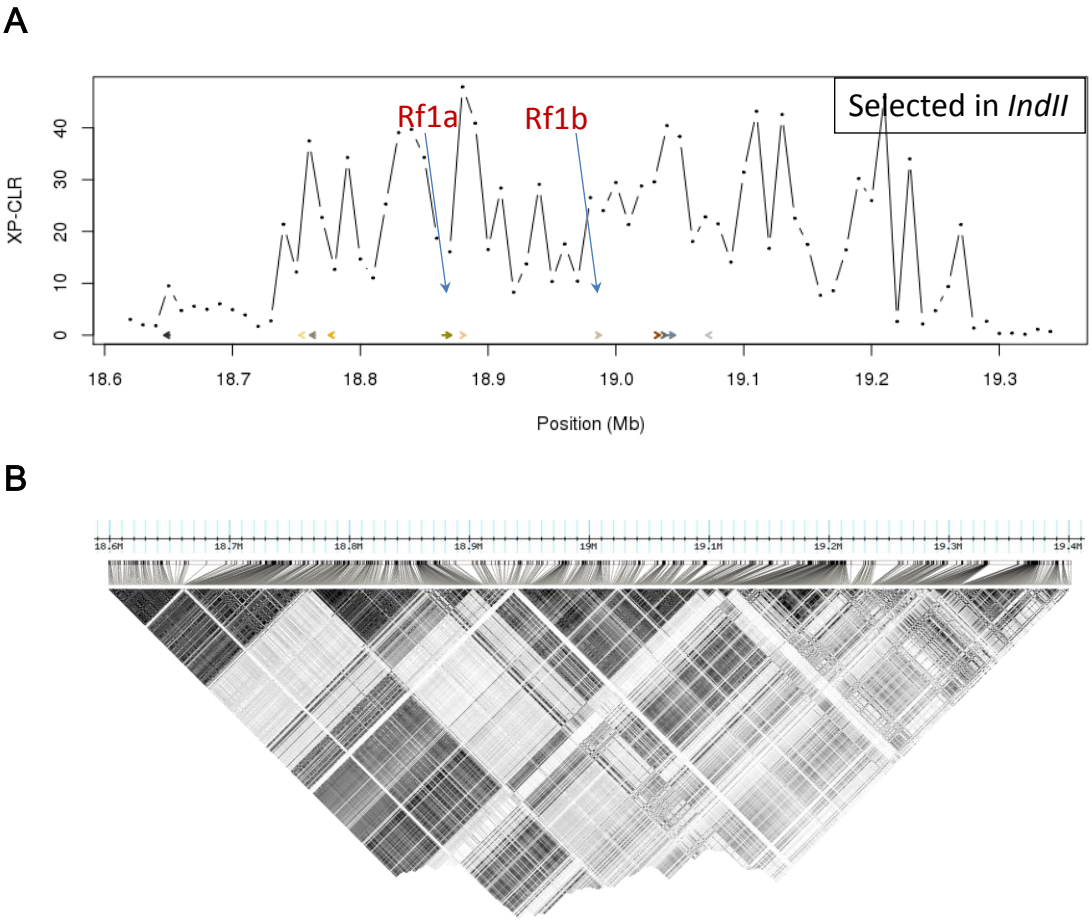


Fig. S9 The XP-CLR and linkage disequilibrium near *Rf1* region. (**A**) The XP-CLR result by using *IndI* as reference population and *IndII* as object population. Each arrow represents a pentatricopeptide repeat domain gene. (**B**) Linkage disequilibrium structure of this region drawn by Haploview based on 1000 randomly sampled SNPs. The grey scale represents the r^2 values (black to white: r^2 = 1 to 0).

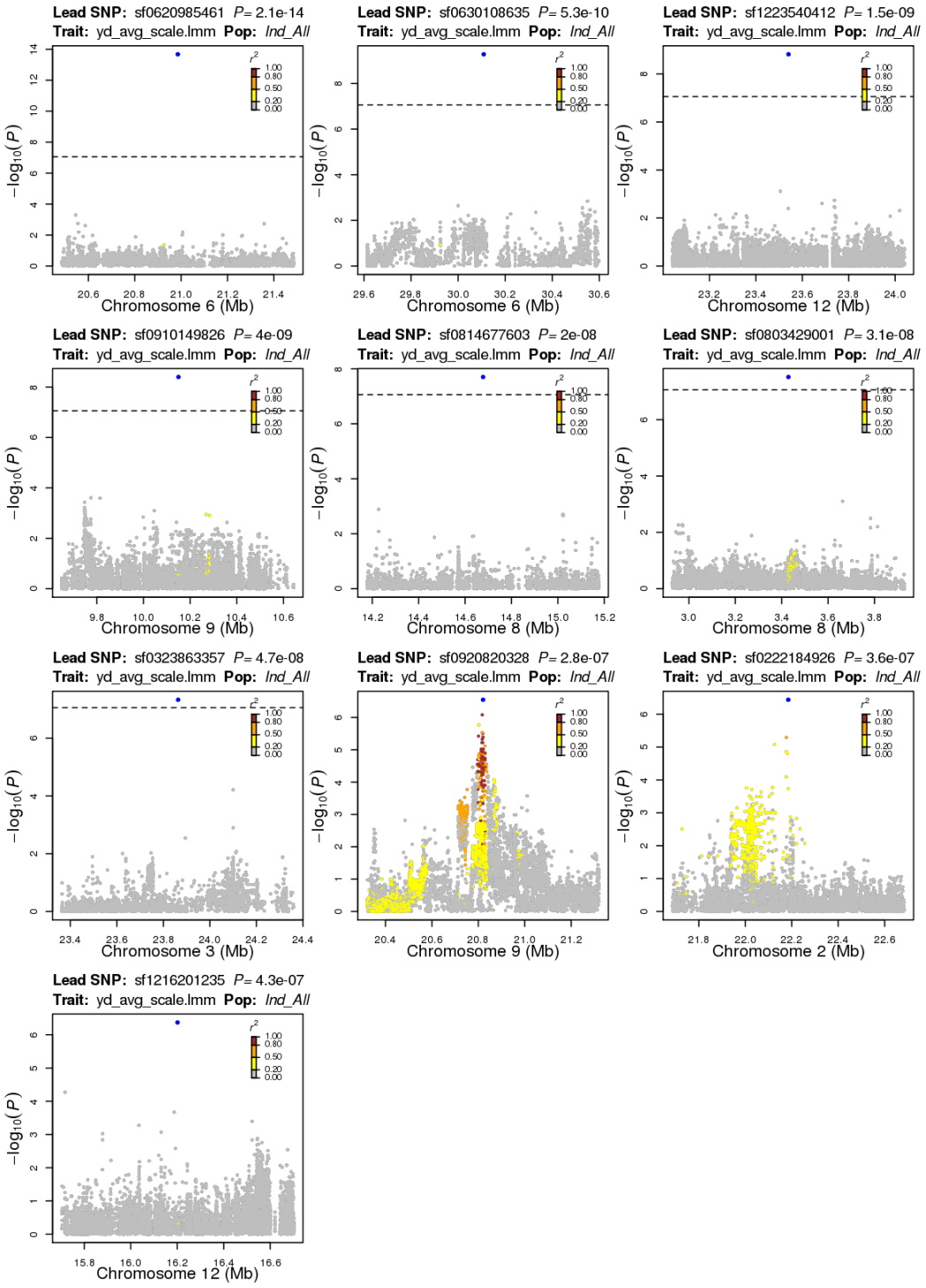


Fig. S10 Association plots of the top ten most significant loci of GWAS for normalized average grain yield using the linear mixed model. Lead SNPs located in the regions are in blue and the colors of the other points represent the degrees of linkage disequilibrium (LD) for the lead SNPs. SNPs in gray are not in LD with the lead SNPs.

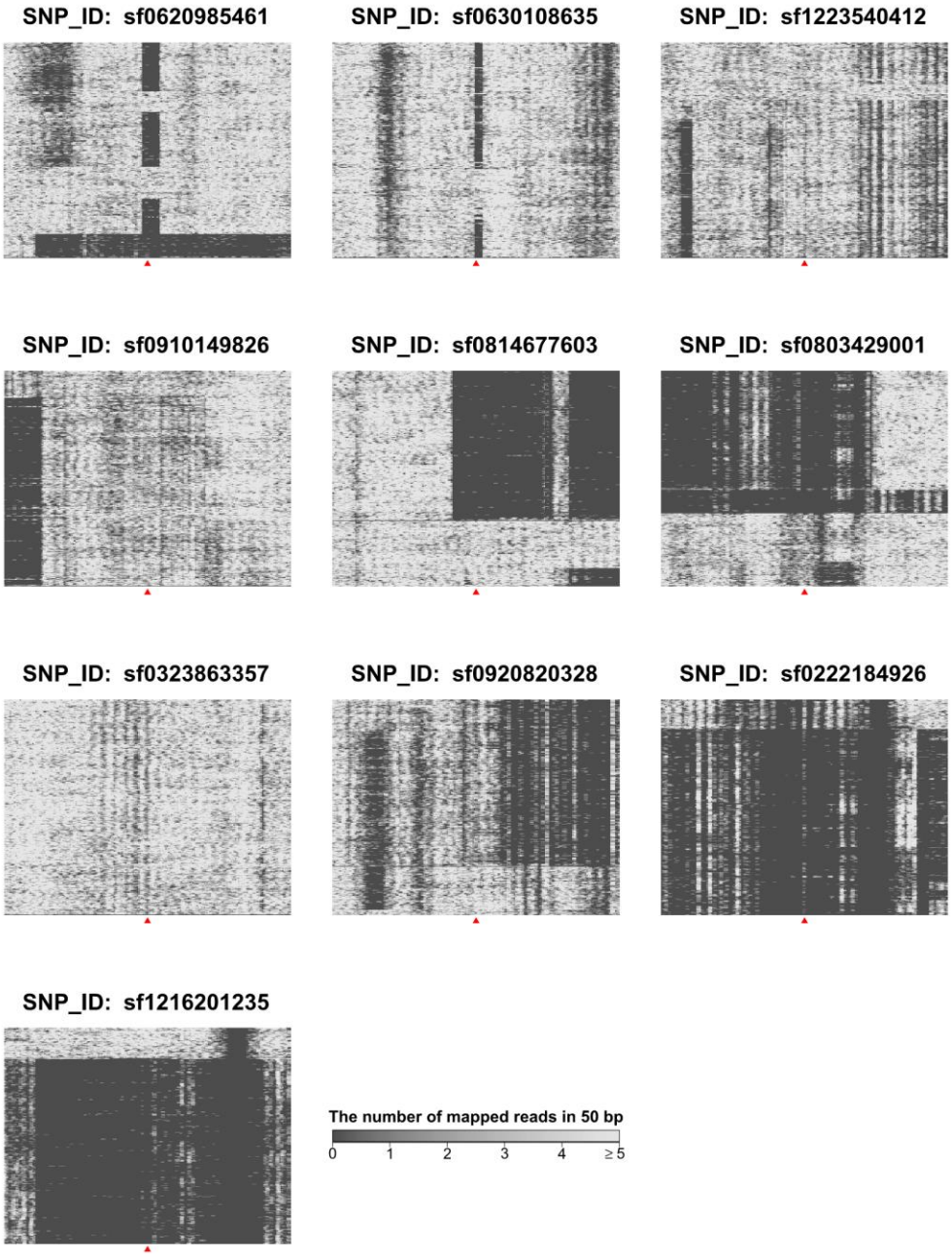
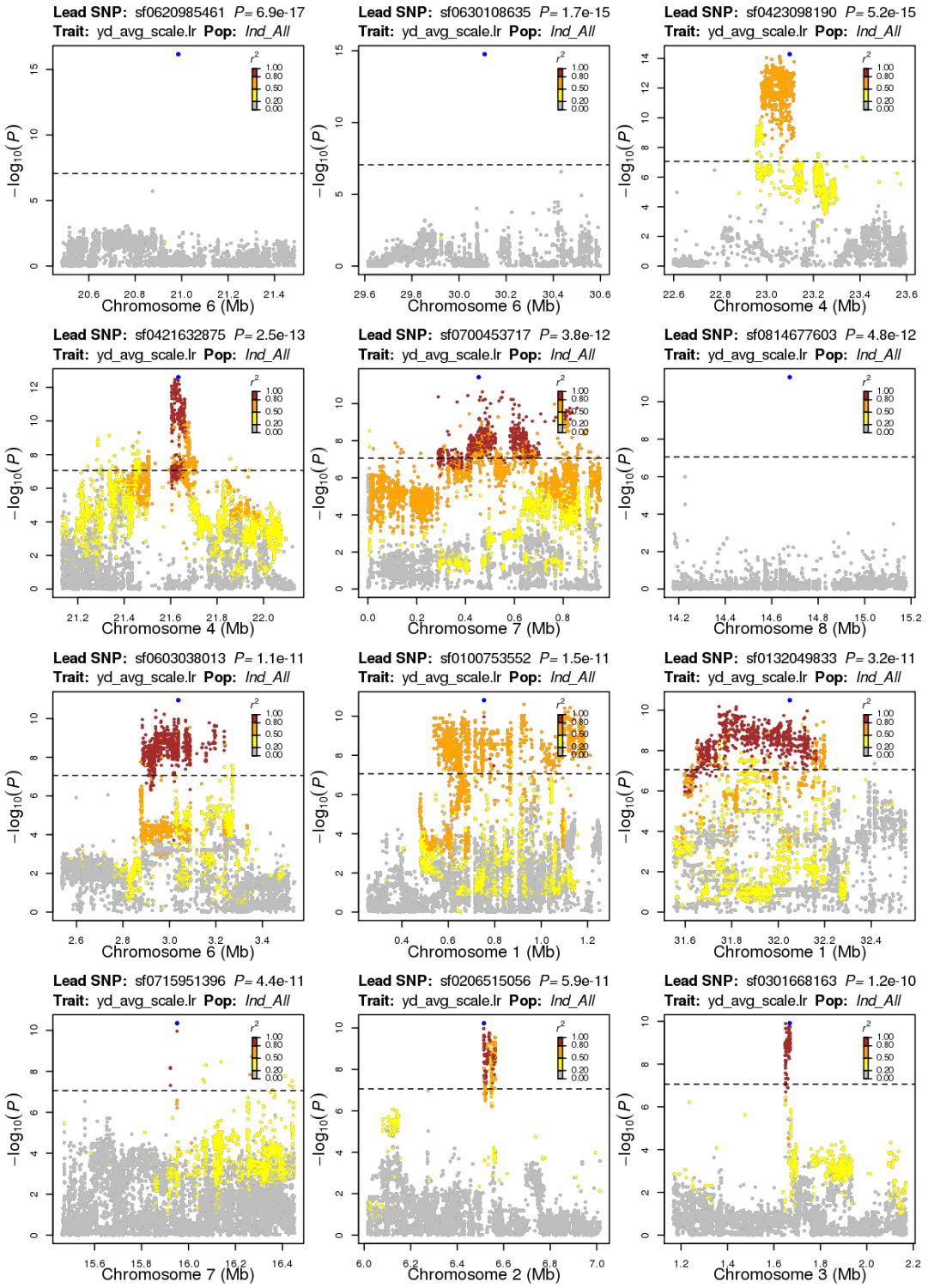


Fig. S11 Sequencing coverage of the top ten most significant loci of GWAS for normalized average rice grain yield using the linear mixed model. For each locus, we counted the number of mapped reads in non-overlapping sliding windows of 50 bp along the flanking 5 kb on each side of the SNP for each of the 295 indica accessions. In a heatmap, each row is an accession and red triangle denotes the location of the lead SNP. The zero-coverage regions are in black. From the distribution of zero-coverage regions, it is clear that some lead SNPs are located in copy number variation (CNV) regions, which may cause imputation errors.


Lead SNP: sf0620985461 P= 6.9e-17
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 6 (Mb)
Lead SNP: sf0630108635 P= 1.7e-15
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 6 (Mb)
Lead SNP: sf0423098190 P= 5.2e-15
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 4 (Mb)
Lead SNP: sf0421632875 P= 2.5e-13
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 4 (Mb)
Lead SNP: sf0700453717 P= 3.8e-12
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 7 (Mb)
Lead SNP: sf0814677603 P= 4.8e-12
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 8 (Mb)
Lead SNP: sf0603038013 P= 1.1e-11
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 6 (Mb)
Lead SNP: sf0100753552 P= 1.5e-11
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 1 (Mb)
Lead SNP: sf0132049833 P= 3.2e-11
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 1 (Mb)
Lead SNP: sf0715951396 P= 4.4e-11
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 7 (Mb)
Lead SNP: sf0206515056 P= 5.9e-11
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 2 (Mb)
Lead SNP: sf0301668163 P= 1.2e-10
Trait: yd_avg_scale.lr Pop: Ind_All
Chromosome 3 (Mb)
−log10(P)
r2
1.00
0.80
0.50
0.20
0.00

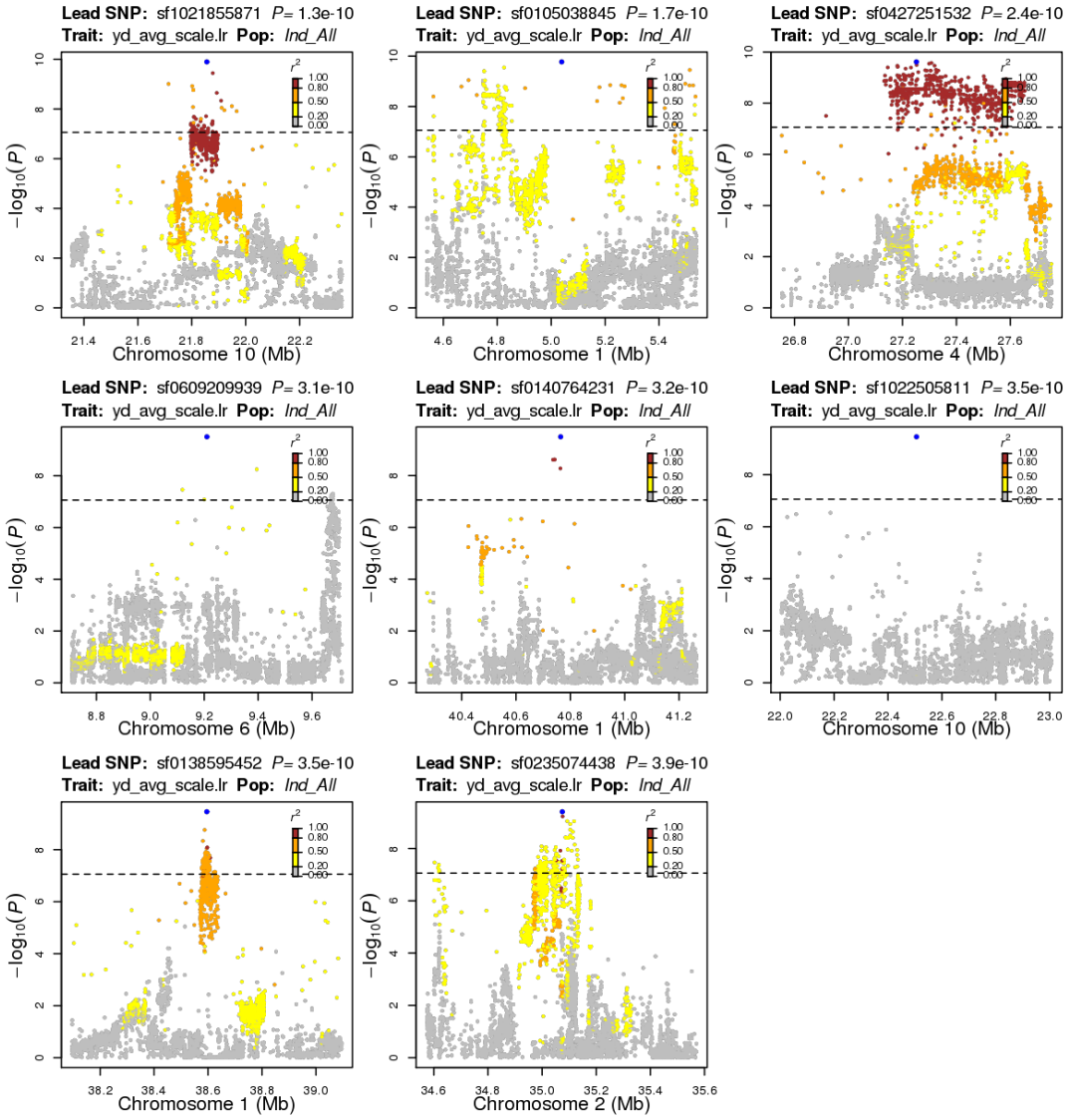


Fig. S12 Association plots of the top twenty most significant loci of GWAS for normalized average rice grain yield using the simple linear regression model. Lead SNPs located in the regions are in blue and the colors of the other points represent the degree of linkage disequilibrium (LD) for the lead SNPs. SNPs in gray are not in LD with the lead SNPs. Six loci (sf0423098190, sf0100753552, sf0132049833, sf 1021855871, sf0105038845 and sf0235074438) are located in the selected regions.

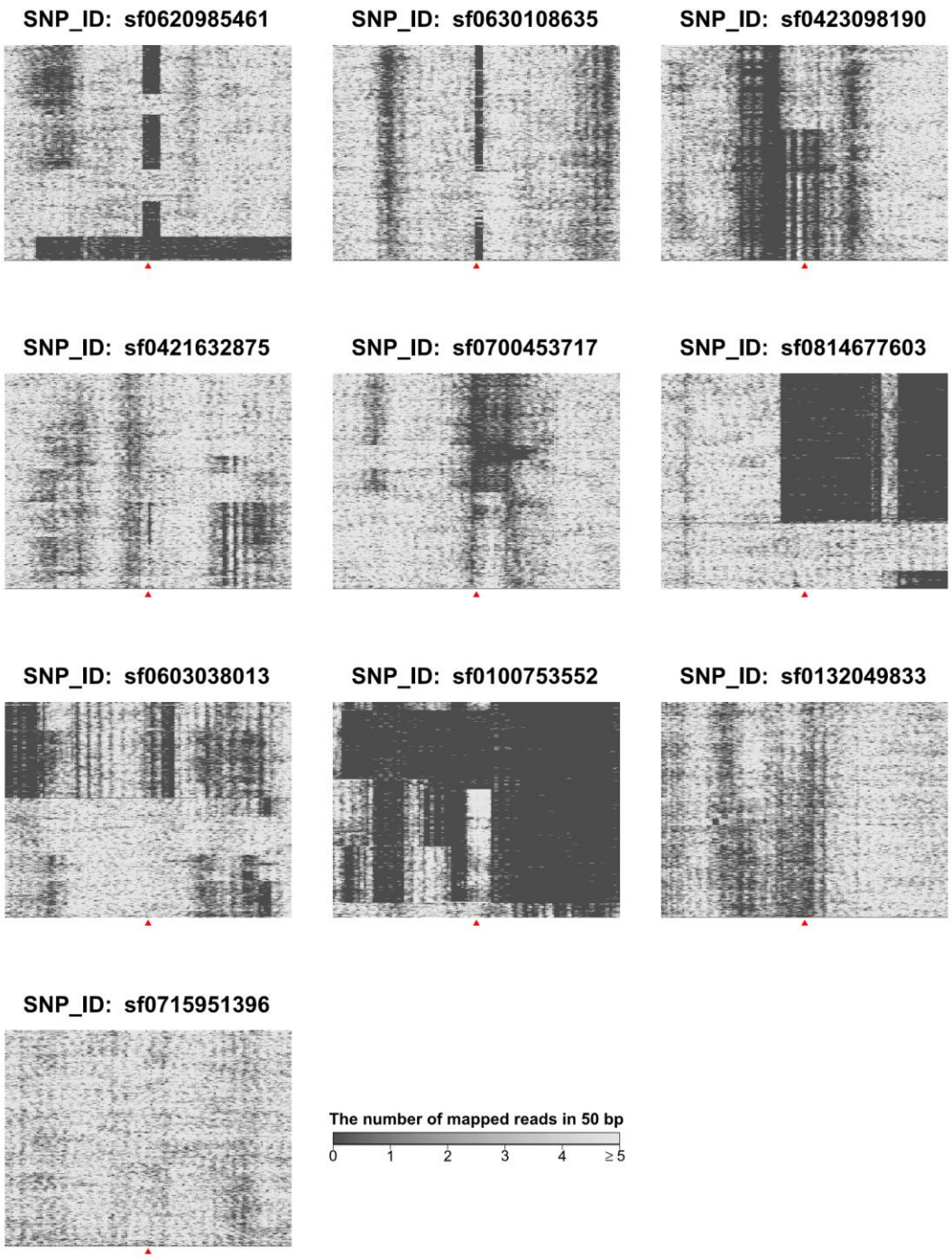


Fig. S13 Sequencing coverage of the top ten most significant loci of GWAS for normalized average rice grain yield using the simple linear regression model. For each loci, we counted the number of mapped reads in non-overlapping sliding windows of 50 bp along the flanking 5 kb on each side of the SNP for each of the 295 indica accessions. In a heatmap, each row is an accession and red triangle denotes the location of the lead SNP. The zero-coverage regions are in black.

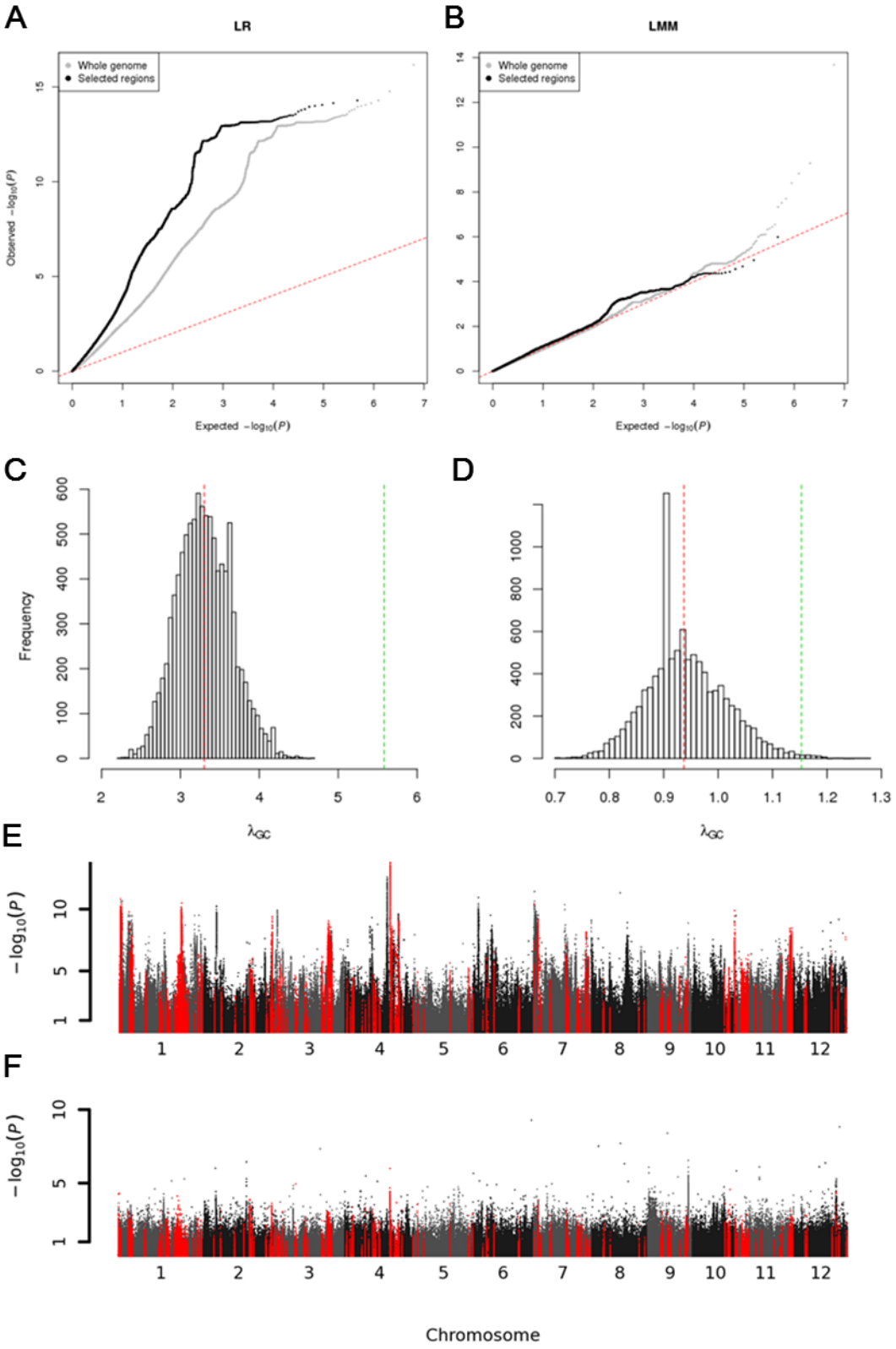


Fig. S14 Selected regions showing enriched effective loci of grain yield. We carried out GWA analysis for normalized average grain yield using different methods. (**A**) QQ plot of the results using simple linear regression model (LR). (**B**) QQ plot of the

results using mixed model (LMM). (**C-D**) Histograms of genomic control factor λ_{GC} of LR (**C**) and LMM (**D**) from 10000 random selected regions with the same width and number as the for selected regions. The vertical red and green lines in **c** indicate the λ_{GC} of LR of whole genome and the selected regions (3.30 and 5.58, respectively). The vertical red and green lines in **D** indicate the λ_{GC} of LMM of whole genome and the selected regions (0.93 and 1.15, respectively). (**E**-**F**) Genome-wide p-values from association analysis of normalized average yield using the simple linear regression model (**E**) and the linear mixed model (**F**). SNPs located in the selected regions are in red.

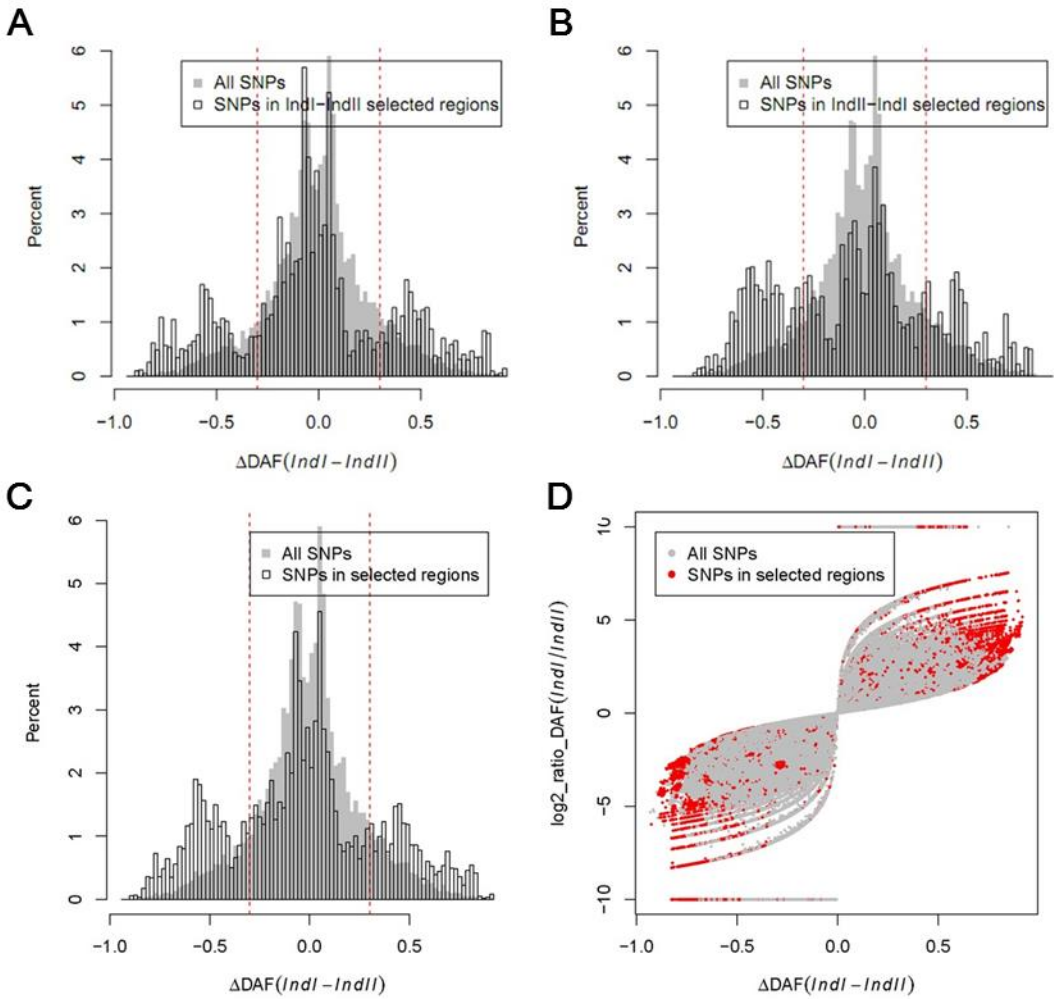


Fig. S15 The distribution of differential derived alleles frequency (ΔDAF) between the two subpopulations. The ΔDAF was calculated by subtracting the derived allele frequency of *IndII* from *IndI*. The two vertical red lines represent the 10th and 90th quantile of ΔDAF of all SNPs, approximately -0.3 and 0.3, respectively. (**A**) SNPs in the *IndI-IndII* selected regions. (**B**) SNPs in the *IndII-IndI* selected regions. (**C**) SNPs in the merged selected regions. (**D**) The *y*-axis is the logarithm (base 2) of ratio of the derived allele frequency of *IndI* and *IndII*.

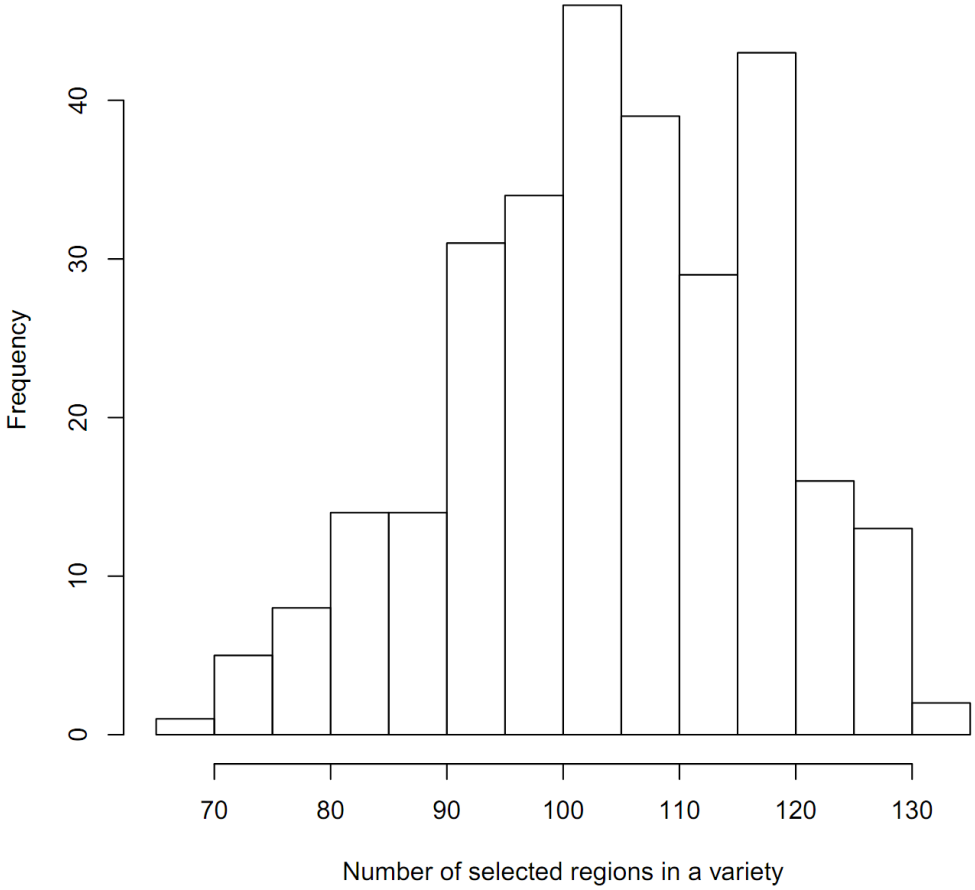


Fig. S16 The distribution of number of regions with derived haplotypes for each accession.

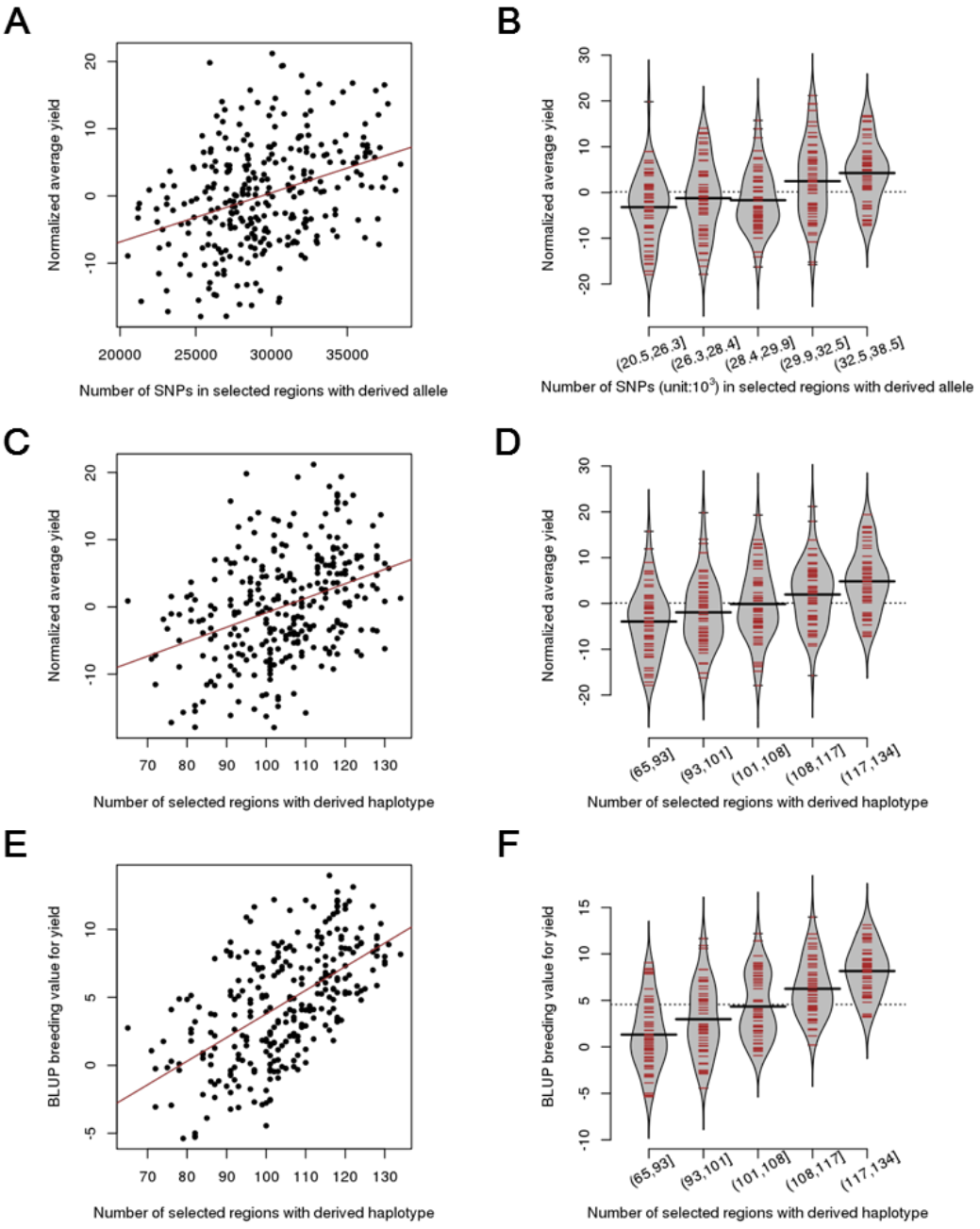


Fig. S17 Correlations between yield and the number of SNPs with derived allele or the number of regions with derived haplotypes for each accession. In drawing the bean plots, the number of SNPs with derived alleles (**B**) or the number of regions with derived haplotypes for each accession (**D,F**) was divided into groups by 20%, 40%, 60% and 80% quantiles. The solid black bar denotes the average for each group.

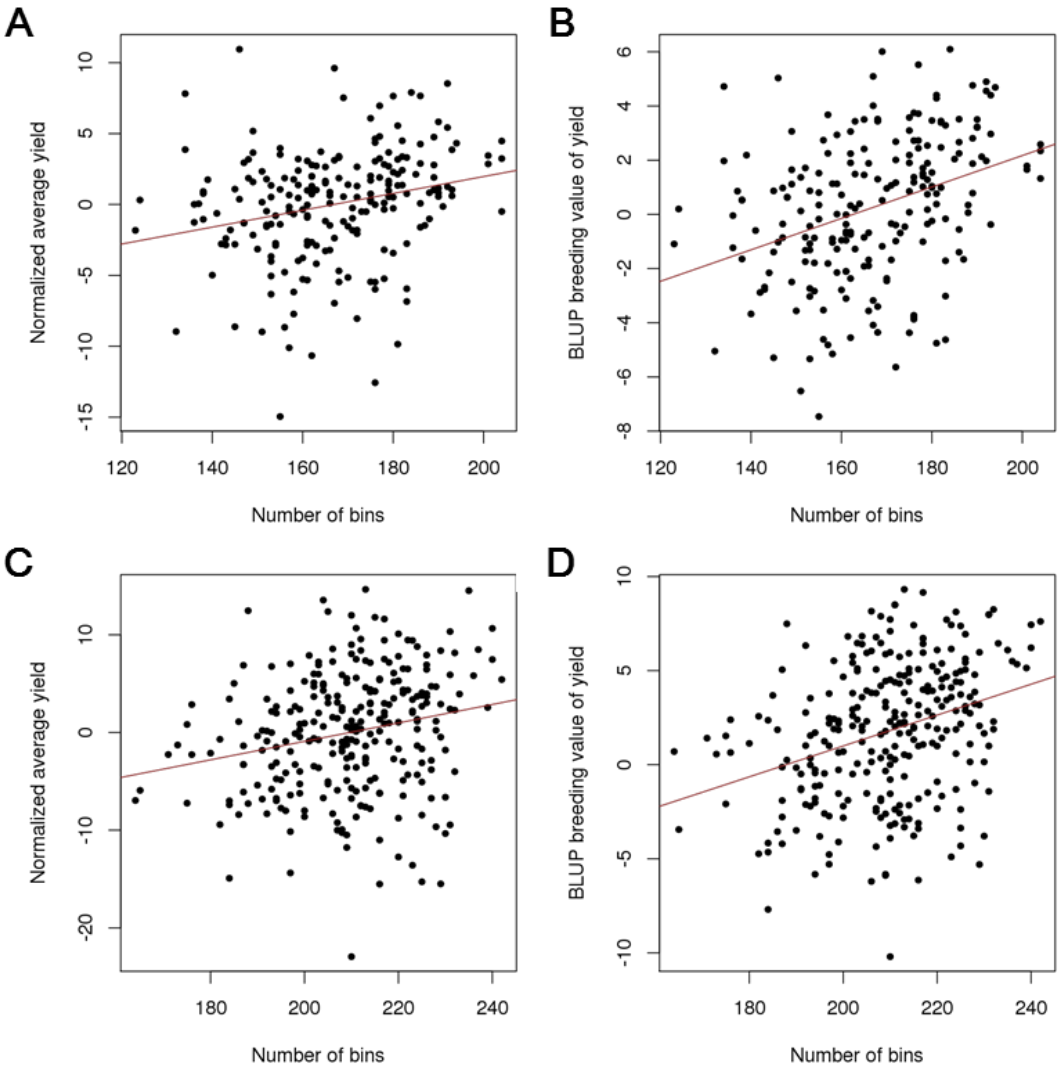


Fig. S18 Correlations between yield, BLUP breeding values of yield and the number of bins with selected genotypes in a RIL and an immortalized F_2 population. (**A**) Normalized average yield of RIL population (N=210, ρ = 0.30, P = 1.0×10^{-5}). (**B**) BLUP breeding values of yield of RIL population (N=210, ρ = 0.40, P = 2.5×10^{-9}). (**C**) Normalized average yield of immortalized F_2 population (N=276, ρ = 0.24, P = 6.8×10^{-5}). (**D**) BLUP breeding values of yield of immortalized F_2 population (N=276, ρ = 0.33, P = 2.2×10^{-8}). For a plant of the immortalized F_2 population, a region, in which one or both of the parental genotypes was under selection, was counted as one selected region. The normalized average yield of RILs and BLUP breeding values were calculated using data of yd97x, yd98x and yd98h only.

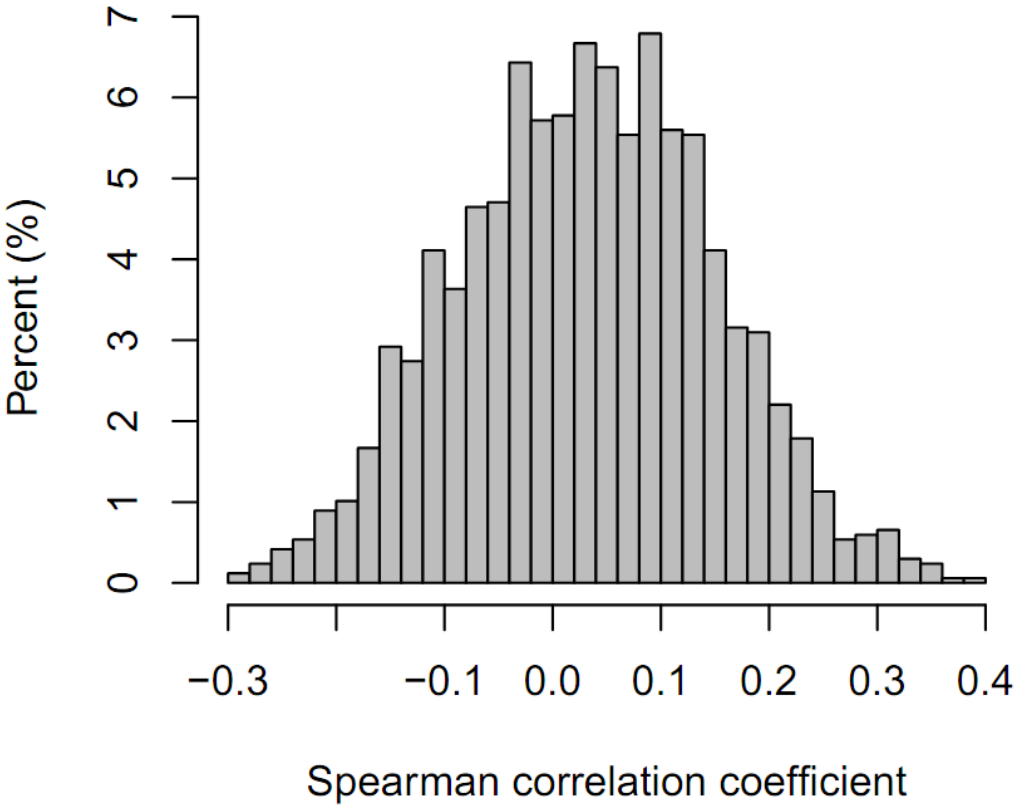


Fig. S19 Distribution of Spearman correlation coefficients between 840 metabolic traits and the number of regions with derived haplotypes for each accession.

Table S1 Missing data rates and accuracy of the complete genotype dataset (6,428,770 SNPs) before and after missing genotypes were inferred. The details of evaluation methods are described in SI Result 1.

Concordance between low-coverage sequencing and pre-existing high coverage sequencing results

Control set	Missing data rate		Accuracy	
	Before imputation	After imputation	Before imputation	After imputation
Nipponbare (HP997)	62.6%	0.18%	99.92%	99.87%
Nipponbare (C146)	12.7%	0.03%	99.94%	99.93%
Minghui 63 (C147)	26.4%	0.36%	99.85%	99.82%
Zhenshan 97 (C145)	26.3%	0.30%	99.92%	99.91%

Concordance between genotyping results of array hybridization and sequencing based on 42759 SNPs

ID	Before imputation			After imputation		
	No. of Concordance	No. of Difference	Accuracy	No. of Concordance	No. of Difference	Accuracy
C002	36096	47	0.999	40765	216	0.995
C003	37070	24	0.999	41385	182	0.996
C005	36855	51	0.999	41505	213	0.995
C006	35121	120	0.997	40584	443	0.989
C025	34681	130	0.996	40723	402	0.990
C026	36751	27	0.999	42213	118	0.997
C027	35055	111	0.997	40862	396	0.990
C028	37893	23	0.999	42154	114	0.997
C029	36971	41	0.999	41805	177	0.996
C037	34992	133	0.996	40670	441	0.989
C038	35183	111	0.997	40607	408	0.990
C040	34886	109	0.997	40722	418	0.990
C041	35093	87	0.998	40764	400	0.990
C042	34557	104	0.997	40785	433	0.989
C043	34737	98	0.997	40845	404	0.990
C048	37636	17	1.000	42085	108	0.997
C050	34500	99	0.997	41079	370	0.991
C052	36930	14	1.000	42075	107	0.997
C054	36828	17	1.000	42018	132	0.997
C056	37195	10	1.000	42248	99	0.998

C060	34266	104	0.997	40775	429	0.990
C063	36595	24	0.999	42157	126	0.997
C064	36754	20	0.999	42051	124	0.997
C065	34054	102	0.997	40570	435	0.989
C067	36327	24	0.999	41734	169	0.996
C074	36182	35	0.999	41585	185	0.996
C091	33896	103	0.997	40811	443	0.989
C094	33592	100	0.997	40539	447	0.989
C096	33733	126	0.996	40426	419	0.990
C101	35957	22	0.999	41737	174	0.996
C103	36364	23	0.999	41719	163	0.996
C106	36821	38	0.999	41782	174	0.996
C110	34568	122	0.996	40572	465	0.989
C113	34759	95	0.997	40674	376	0.991
C114	34291	100	0.997	40618	401	0.990
C120	37714	20	0.999	42262	101	0.998
C123	36634	34	0.999	41582	187	0.996
C124	34098	101	0.997	40599	405	0.990
C129	34101	138	0.996	40505	455	0.989
C131	34655	121	0.997	40587	455	0.989
C134	37211	109	0.997	41698	236	0.994
C135	35296	174	0.995	40747	479	0.988
C137	36808	43	0.999	41739	146	0.997
C140	34618	97	0.997	40528	394	0.990
C143	34911	124	0.996	40659	419	0.990
C144	36774	30	0.999	41785	161	0.996
C148	34376	111	0.997	40780	380	0.991
C149	37072	17	1.000	42195	94	0.998
Total	1711457	3530	0.998	1979310	14023	0.993

Table S2 Correlations between yield and the number of bins with selected genotype in the RIL and immortalized F_2 populations (IMF_2). *N*: sample size; *R* and *P*: Spearman's rank correlation coefficient and *P*-value calculated using the asymptotic *t* approximation by Fisher's Z transform implemented in R function *cor.test*. The yield data of RILs and IMF_2 were from Xing *et al.* (10) and Hua *et al.* (11). Normalized average yield: the yield data from multiple years were centered and averaged when multiple observations for a line were available. The normalized average yield of RILs and BLUP breeding values were calculated using data of yd97x, yd98x and yd98h only. The ridge regression BLUPs of RIL and IMF_2 were calculated using R package rrBLUP with similar procedures of analyzing varieties.

Yield	*N*	*R*	*P*
RILs, yd97x	205	0.17	1.6×10^{-2}
RILs, yd98x	210	0.22	1.1×10^{-3}
RILs, yd98h	209	0.26	1.3×10^{-4}
RILs, yd99h	209	0.13	0.061
RILs, Normalized average	210	0.30	1.0×10^{-5}
RILs, BLUP breeding value	210	0.40	2.5×10^{-9}
IMF_2, yd98	246	0.16	1.2×10^{-2}
IMF_2, yd99	276	0.19	1.6×10^{-3}
IMF_2, Normalized average	276	0.24	6.8×10^{-5}
IMF_2, BLUP breeding value	276	0.33	2.2×10^{-8}

Table S3 Correlations of normalized grain yield and *numDR* with plant compactness and grain-projected area. *N*: sample size; *R* and *P*: Spearman's rank correlation coefficient and *P*-value calculated using the asymptotic *t* approximation. The results showed that although the normalized grain yield was not associated with the two new traits, *numDR* was significantly associated with plant compactness.

Correlations between normalized grain yield and two new traits:

Trait	*N*	*R*	*P*
Plant compactness at the late tillering stage	253	0.11	0.076
Plant compactness at the late booting stage	242	0.041	0.53
Plant compactness at the milk grain stage	165	0.034	0.67
Grain-projected area	287	-0.09	0.13

Correlations of *numDR* with plant compactness and grain-projected area:

Trait	*N*	*R*	*P*
Plant compactness at the late tillering stage	253	0.20	0.012
Plant compactness at the late booting stage	242	0.27	2.6×10^{-5}
Plant compactness at the milk grain stage	165	0.085	0.28
BLUP breeding value of plant compactness at the late tillering stage	295	0.27	2.8×10^{-6}
BLUP breeding value of plant compactness at the late booting stage	295	0.41	3.9×10^{-13}
BLUP breeding value of plant compactness at the milk grain stage	295	0.20	4.6×10^{-4}
Grain-projected area	287	-0.053	0.37
BLUP breeding value of grain-projected area	295	-0.030	0.61