Supplementary Information for

Divergent selection and genetic introgression shape the genome landscape of heterosis in hybrid rice

Zechuan Lin (School of Advanced Agriculture Sciences and School of Life Sciences, State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China)
Peng Qin (Department of rice breeding, Hunan Yahua Seed Scientific Research Institute, Changsha 410119, China and State Key Laboratory of Hybrid Rice; Key Laboratory of Southern Rice Innovation & Improvement, Ministry of Agriculture and Rural Affairs, Changsha 410119, China)
Xuanwen Zhang (Department of rice breeding, Hunan Yahua Seed Scientific Research Institute, Changsha 410119, China and State Key Laboratory of Hybrid Rice; Key Laboratory of Southern Rice Innovation & Improvement, Ministry of Agriculture and Rural Affairs, Changsha 410119, China)
Chenjian Fu (Department of rice breeding, Hunan Yahua Seed Scientific Research Institute, Changsha 410119, China and State Key Laboratory of Hybrid Rice; Key Laboratory of Southern Rice Innovation & Improvement, Ministry of Agriculture and Rural Affairs, Changsha 410119, China)
Hanchao Deng (Department of molecular marker, Shenzhen Institute of Molecular Crop Design, Shenzhen 518107, China; Department of molecular marker, henzhen Agricultural Science and Technology Promotion Center, Shenzhen 518055, China)
Xingxue Fu (Department of rice breeding, Hunan Yahua Seed Scientific Research Institute, Changsha 410119, China and State Key Laboratory of Hybrid Rice; Key Laboratory of Southern Rice Innovation & Improvement, Ministry of Agriculture and Rural Affairs, Changsha 410119, China)
Zhen Huang (Department of rice breeding, Hunan Yahua Seed Scientific Research Institute, Changsha 410119, China and State Key Laboratory of Hybrid Rice; Key Laboratory of Southern Rice Innovation & Improvement, Ministry of Agriculture and Rural Affairs, Changsha 410119, China)
Shuqin Jiang (Department of Crop genomics and Bioinformatics, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China)
Chen Li (Rice Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China)
Xiaoyan Tang (Department of molecular marker, Shenzhen Institute of Molecular Crop Design; Guangdong Provincial Key Laboratory of Biotechnology for Plant Development, College of Life Sciences, South China Normal University, Guangzhou 510631, China)
Xiangfeng Wang (Department of Crop genomics and Bioinformatics, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China)
Guangming He (School of Advanced Agriculture Sciences and School of Life Sciences, State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China)
Yuanzhu Yang (Department of rice breeding, Hunan Yahua Seed Scientific Research Institute, Changsha 410119, China and State Key Laboratory of Hybrid Rice; Key Laboratory of Southern Rice Innovation & Improvement, Ministry of Agriculture and Rural Affairs, Changsha 410119, China)
Hang He (School of Advanced Agriculture Sciences and School of Life Sciences, State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China)

www.pnas.org/cgi/doi/10.1073/pnas.1919086117

Xing Wang Deng (School of Advanced Agriculture Sciences and School of Life Sciences, State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China)

Xing Wang Deng
Email: deng@pku.edu.cn
Hang He
Email: hehang@pku.edu.cn
Yuanzhu Yang
Email: yzhuyah@163.com

This PDF file includes:

Supplementary text
Figures S1 to S9
Tables S1 to S2
Legends for Datasets S1 to S10
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S10

Supplementary Information Text

Supplementary Materials and Methods

Hybrid population construction, genotyping and phenotyping. About one-hundred restorer lines that were improved from widely used backbone restorer lines or cultivars by breeders with over fifteen years of efforts and ~100 male sterile lines that were frequently applied to the breeding of super hybrid rice were selected as candidate parents in our study. In our experimental trial in the year 2014 (denoted as Pop I), we crossed 105 restorer lines (paternal) to 78 male sterile lines (maternal) to produce 8,000 hybrids. The male parents of both populations were comprised of high-quality conventional rice (like Jinnongsimiao, Youjingzhan and Hejingzhan) and improved lines which developed from the yield, grain quality or disease and pest resistance improvement of commercial hybrid parents. The female parents of Pop I were comprise of 6 commercially frequently used three-line male sterility lines (like 9311A and 398A) and 13 present-day commercially frequently used two-line male sterility lines (like Y58S, 638S and Guangzhan63-4S) and 24 relative improved lines, while the female parents of Pop II were 8 commercially frequently used two-line male sterility lines and 26 relative improved lines. None of male parents was employed during the construction of both hybrid populations, but 8 commercially frequently used two-line male sterility lines were simultaneously employed at both populations (Table S1). We performed phenotyping for ten important agronomical traits towards ~1,000 hybrids in Pop I at Changsha, China at the year of 2014 (denote as 2014CS), ~1,000 hybrids in Pop II at Changsha, China at the year of 2015 (denote as 2015CS) and ~700 hybrids in Pop II at Hefei, China at the year of 2015 (denote as 2015HF). All 171 male and 105 female parents were genotyped using a 50k SNP chip (includes ~50,000 SNPs) to obtain high-quality genotype data for each line. Polygenetic UPGMA trees that demonstrated the relationship among restorer or male sterile lines were constructed with MEGA6 (1), and lines close to any other line were manually removed, leaving 66 paternal and 43 maternal lines (Table S1). For the final step, 1,000 hybrid offspring were selected under the criteria that every paternal line should have more than twelve hybrids, whereas every maternal line has more than twenty hybrids. Pop I was cultivated in Changsha, China, in the summer of 2014 (denoted as 2014CS). About 74 lines that were unable to heading because of sensitive to photoperiod or with missing phenotype were removed, leaving 926 hybrids for the following analysis. A similar hybrid design and selection processes was employed in the experimental trial of the year 2015 to select another 1,000 hybrids encompassing 65 paternal and 32 maternal lines (denoted as Pop II, Table S1). Pop II was cultivated in both Changsha and Hefei, China (denoted as 2015CS and 2015HF, respectively), in the summer of 2015. About 78 lines at the PopIIin 2015CS that were unable to heading or with missing phenotype were removed, leaving 922 hybrids for the following analysis. About 437 hybrids at the Pop II in 2015HF were removed due to unable to heading, with missing phenotype or affected by diseases, leaving 563 hybrids for the following analysis. Sixteen plants in the middle of each block were selected to phenotype the heading date, grain yield per plant, plant height, panicle number per plant, straw weight and panicle weight. Biomass per plant was defined as the sum of the panicle weight and straw weight per tiller, and the harvest index was defined as panicle weight divided by biomass per tiller.

QTL mapping. The hybrid parental lines were genotyped with 50k chip designed in our previous study (2). SNPs failed to match the following criteria were removed: (1). SNPs with missing rate lower than 30%; (2). No more than 5% of heterozygous genotype; (3). SNPs without polymorphism among lines. After filtering, 34,788 of SNPs were retained on the following analysis except selective sweep analysis where we also included non-polymorphism SNPs. The missing genotypes of all parents were imputed with Begale (Version 3.0) (3) by using another 1,571 genotyped individuals on our 50k chip platform. Hybrid genotypes were obtained by combining the haploid genome of their corresponding parents. However, the heterozygous loci of the parents may produce ambiguous genotypes in hybrids. Therefore, we first classified these loci as missing and then imputed the genotypes of all hybrids again to fill these missing loci. After imputation, the missed genotyping rate among hybrids was limited to 5%.

The QTLs were mapped with all SNPs (MAF > 0.05) with additive, dominant/recessive and over-dominant liner mixed model. For additive model, we used the origin genotype matrix of all markers. For dominant and over-dominant model, we recoded the markers into dummy genotypes: (1) For dominant/recessive model, we recoded Dd into dd (recessive model) or DD (dominant model) while keep the DD and dd; (2) For over-dominant model, we recode dd into DD and Dd into dd, which suggests heterozygous genotype compared against the homozygous. The dummy markers were also fitted with MAF > 0.05. The kinship matrix were respectively calculated with origin (for additive model) and dummy genotype matrix (for dominant/recessive and over-dominant models), and genome-wide association study for all models were performed with EMMAX with default parameters (4). The genome-wide significant threshold (FDR < 0.05) was determined by 300 times of permutation, which finally draws a p-value cutoff of –logP = 5 that all models at all traits could pass the FDR cutoff.

Population structure analysis. We selected the 50k SNP loci of the ~5,000 reported resequenced lines (46,572 of SNPs) as their genotypes, and removed wild rice lines, leaving 4,214 landraces and improved varieties which cover all Asia cultivated rice sub-populations (5-9). These varieties were integrated with the genotype of hybrid parents, and population structure was conducted with ADMIXTURE tool (10) with default parameters. The number of ancestry populations was determined by five-fold cross-validation.

To test the relationship between the dosage of introgressed genome and heterosis of hybrids, we first employed generalized linear model (GLM) to remove year and location effects for each trait with all hybrids in the study, and investigated the correlation between the dosage of introgressed genome and the trait values after regression. Furthermore, to test the relationship between the dosage of introgressed genome and general combining ability (GCA) of male/female parents, we calculated the

GCA for every parental line based on the trait values of hybrids described above, and then test the correlation between the GCA and the dosage of introgressed genome.

Identification of introgressed regions. To detect introgressed genome regions, we resequenced 36 core female and 79 male lines with a depth of ~3.0 × (Table S1), controlled the reads quality with Cutadapt (11), mapped the reads to Nipponbare reference genome (IRGSP 1.0) with BWA men (default parameters) (12), and called variants with GATK (v3.0) best practice pipeline (13). We finally obtained 1.39 M of variants (missing rate < 70%), with 81.64% of which were in common with the variants detected from the 4,214 resequenced landraces and improved varieties. The missing genotypes were then imputed with Beagle 4 with default parameters (3). We removed the parents of three-line hybrids, integrated the leaving hybrid parents with 4,214 resequenced landraces and improved varieties, and employed their common variants to analyze the introgressed genome regions at hybrid parents with four-taxon f_d statistic which calculates the excessively shared derived variants between two taxa (14). For screening introgressed regions from *aus* or other *indica* sub-populations (Ind I and Ind III) to male or female parents (Ind II), we selected *japonica* landraces as the outgroup. For screening introgressed regions from japonica sub-populations (*Tropical japonica* and *Temperate japonica*) to male or female parents (Ind II), we selected *aus* as the outgroup. We selected male as the control group during the detection of female parents and vice versa. We calculated the f_d statistic with a window size of 25 Kb and a step of 10 Kb, removed windows with less than three informative SNPs or with meaningless result ($f_d > 1$, $f_d < 0$ or with Patterson's D statistic < 0). The cutoff to define significant introgressed regions were informed from the population structure analysis results: for male or female parental population, for each sub-population introgression event (eg. Ind I to female, TroJ to male etc.), we removed male/female parental individuals which have less than 0.001% of genome from that sub-population, and selected the average proportion of genome that from the sub-population (eg. 5%) of the leaving individuals as the estimated proportion of genome that introgressed from the sub-population, then genome-wide windows (including informative and non-informative windows) with strongest 5% of f_d value were considered as the introgressed regions. The additive genetic relatedness matrix among individuals was calculated by EMMAX tool with SNPs from the introgressed and non-introgressed regions, respectively. Then G-BLUP was used to partition the heritability of traits into components explained by SNPs from the introgressed and non-introgressed regions (15, 16).

Construction of trees of heterotic loci and analysis of divergent selection. To test whether all QTLs involved in genetic introgression from other populations, we constructed the maximum likelihood tree for all QTLs with their 25 Kb flanking variants of the resequenced hybrid parents, landraces and improved varieties by using FasTree tool (v2.1.9) (17). QTLs with less than 10 variants at their 25 Kb flanking regions were skipped during the analysis. The trees were annotated with population information and their topology was manually analyzed to detect potential introgression events (other sub-population to Ind II).

To uncover potential divergent selection at the heterotic loci, we traced the major derived sub-populations of heterotic alleles by analyzing the QTL tree topology. As we observed large proportion of loci involved in Ind II/*japonica* (one allele from Ind II while the other allele from *japonica*) or Ind II/Ind I (one allele from Ind II while the other allele from Ind I), we respectively investigated the *Fst* value (18) between *indica* and *japonica*, and Ind II and Ind I, with a window size of 100 Kb and a step of 10 Kb. The highest *Fst* value of windows at heterotic loci 2.5 Kb flanking regions were considered as that of the heterotic loci. To generate a negative control, we conducted 1000 times of random resampling towards genome-wide SNPs, with every time sampled heterotic loci equal amount of SNPs, and investigated their *Fst* as described above. The distribution of *Fst* of the resampled SNPs was then compared to that of heterotic loci to test whether they were enriched in the regions with higher *Fst* value. The variants allele frequency difference (AFD) between *indica* and *japonica*, and Ind II and Ind I, was also respectively investigated, and AFD of variants on the 25 Kb flanking regions of heterotic loci was compared against the genome-wide background to test whether they were potentially involved in divergent selection.

To test whether major heterotic gene, *Ghd8*/*DTH8* (19, 20), *Gn1a* (21), *IPA1* (22) and *RPL1* (23), were involved in divergent selection at their derived sub-populations, we employed the genic and 5 Kb flanking variants (25 Kb for *IPA1* gene, since we did not observe any genic variants) of the genes at the resequenced hybrid parents, landraces, improved varieties and wild rice to construct their haplotype network with minimum spanning method implemented in POPART (24). Haplotypes with a frequency that < 2 were removed before the network construction. The haplotypes were classified into categories on the basis of the number of variants among them.

Selected sweep analysis. The reported resequenced landraces and improved varieties were employed as reference population, and our male/female parent was compared against them with an inter-population composition likelihood approach (XP-CLR) (25), to detect potential selective sweeps. As all hybrid parents were genotyped with our 50k chip, but only ~50% of them were resequenced, considering the sample size problem, we only used the 50k chip SNPs during the detection of selective sweeps. Genetic distances between adjacent SNPs were incorporated according to their physical distances in an ultra-high-density genetic map reported in a previous study. We set the XP-CLR parameters following a previous study with modifications: a step of 500 bp and a sliding window size of 0.5cM. After obtaining the results, the chromosomes were divided into 5 Kb bins, and XP-CLR value of every 5 Kb bin was obtained by averaging over all origin 500 bp windows in the bin. Bins located in 0.1 cM around the centromere of each chromosome were removed. The bins with strongest 5% of XP-CLR signals were selected as candidate selective sweeps in the study, and selective sweeps with distance less than 10 Kb were merged.

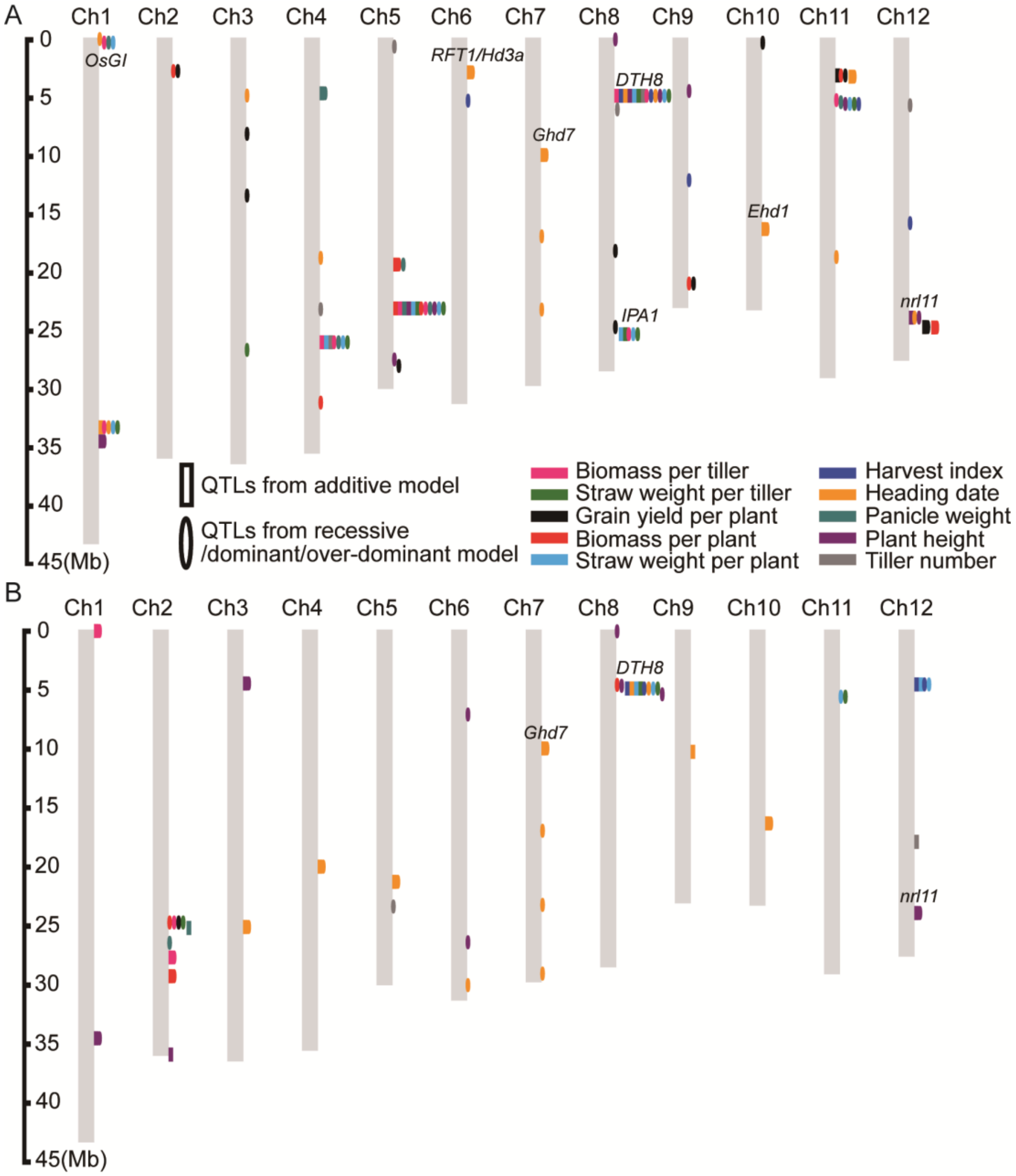

Fig. S1. Chromosomal distribution of potential heterotic loci from 2015CS (A) and 2015HF (B) trials. Ellipses indicate heterotic loci detected using the dominant/overdominant model, while rectangles indicate those detected using the additive model. Different colors represent different traits.

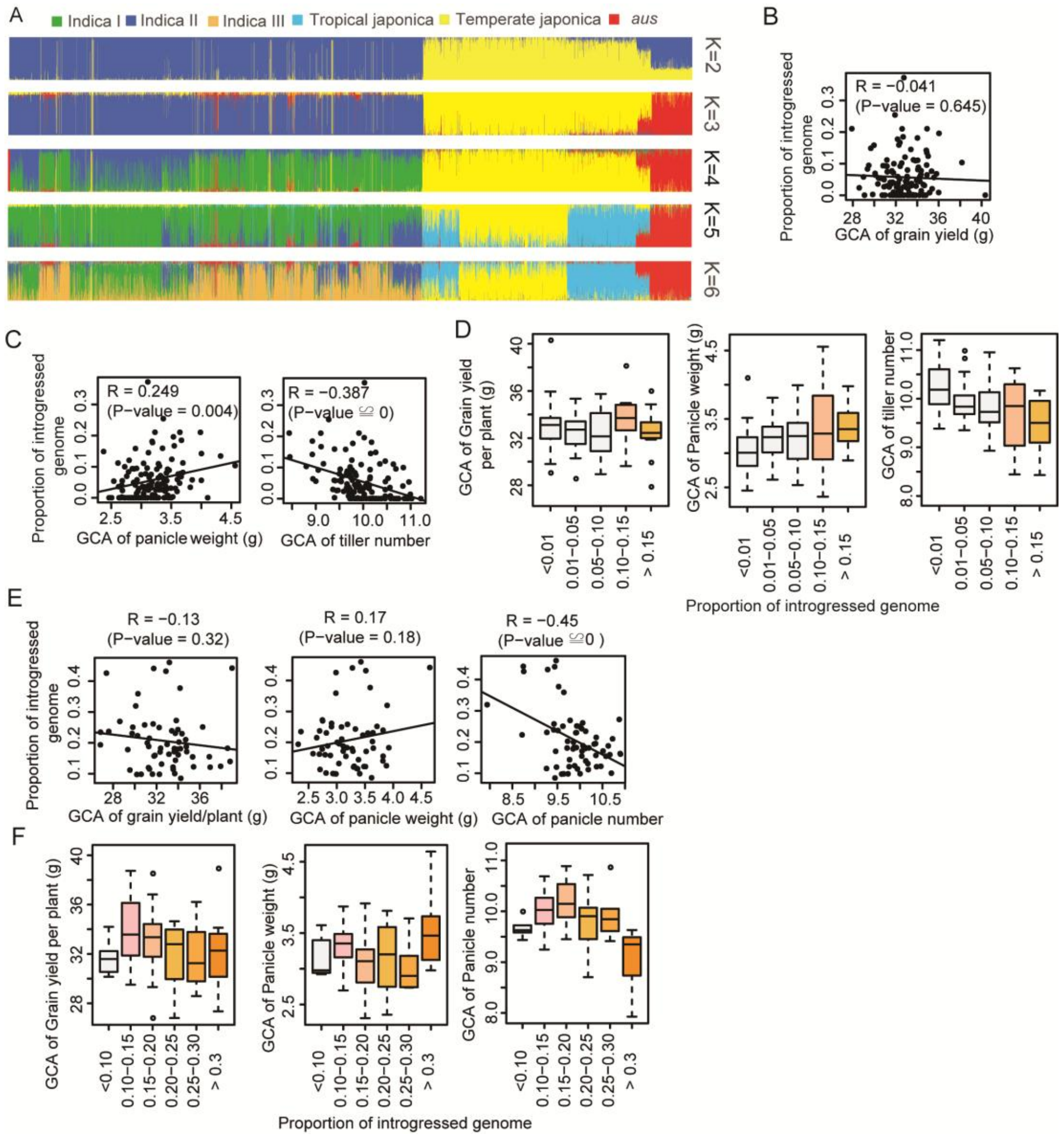

Fig. S2. Exogenous genome introgression improved the heterotic level of male parents. (A) Population structure for ~5,000 landrace strains and improved varieties. The reasonable number of ancestor populations was determined using five-fold cross validation. Only SNPs on the 50k chip were used to conduct the population structure analysis. (B-C) Correlation between the degree of genome introgression and the general combining ability (GCA) of grain yield per plant (left), panicle weight (middle), and tiller number (right) in male parents. All male parents of both hybrid populations were used in the analysis. The year, location, and population effects were regressed out using a general linear model before calculating GCA for each trait. (D) Comparison of the differences in GCA of the three yield traits among male parents with different levels of exogenous genome introgression. The bar in the middle of each box plot indicates the 50th quantile of the GCA value for the group. (E) Correlation between degree of genome introgression and the GCA of grain yield per plant (left), panicle weight (middle), and tiller number (right) for female parents. All female parents of both hybrid populations were used in the analysis. The year, location, and population effects were regressed out using a general linear model before calculating GCA for each trait. (F) Comparison of differences in the GCAs of yield traits among female parents with different levels of exogenous genome introgression. Hybrids were grouped by their levels of exogenous genome introgression and the GCAs of the three yield traits for

each group were investigated and compared. The bar in the middle of each box plot indicates the 50th quantile of the GCA value for the group.

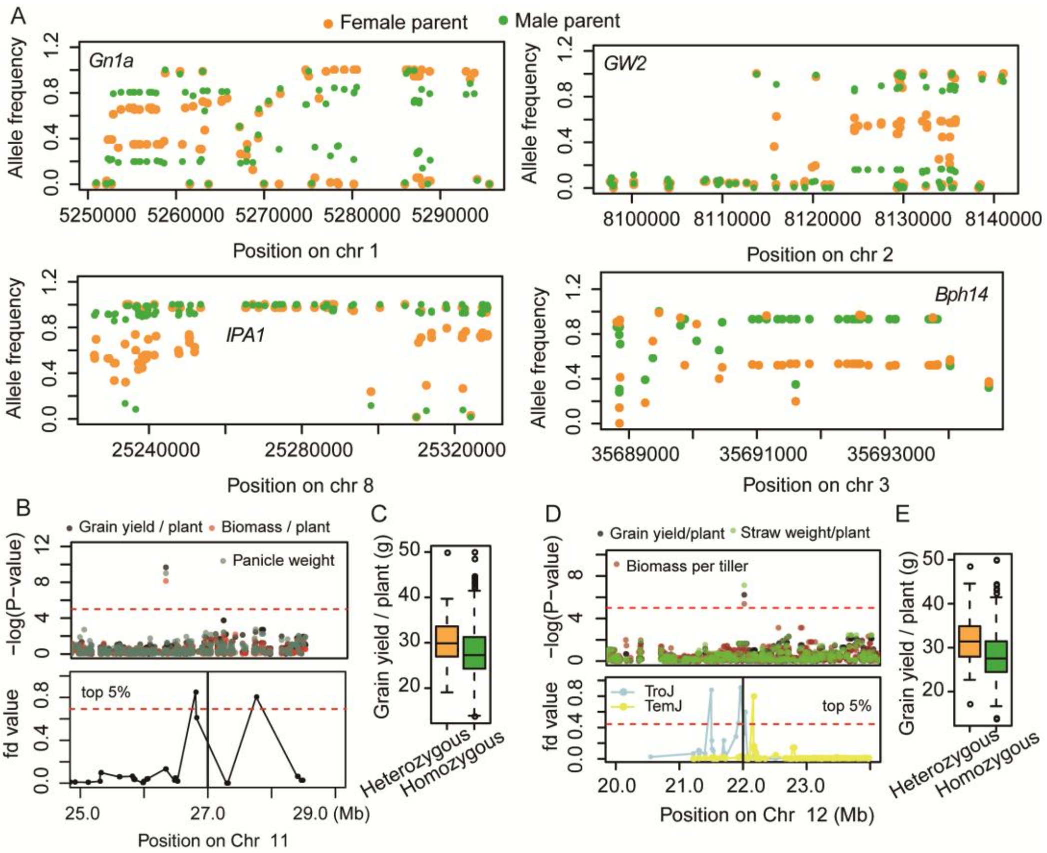

Fig. S3. The effects of parental genetic introgression differences on important genes and heterotic loci. (A) Allele frequency differences between male and female parents for important grain yield (*Gn1a*, *GW2*, *IPA1*) and biotic stress response genes (*Bph14*) located in introgressed regions. (B-C) Heterotic locus on chromosome 11 involved in genetic introgression from *Ind I* to female parents (B, bottom). This locus had a strong heterotic effect on grain yield per plant in the 2014CS trial. (B-C) Heterotic locus on chromosome 12 involved in genetic introgression from *japonica* to female parents (B, bottom). This locus had a strong heterotic effect on grain yield per plant in the 2014CS trial.

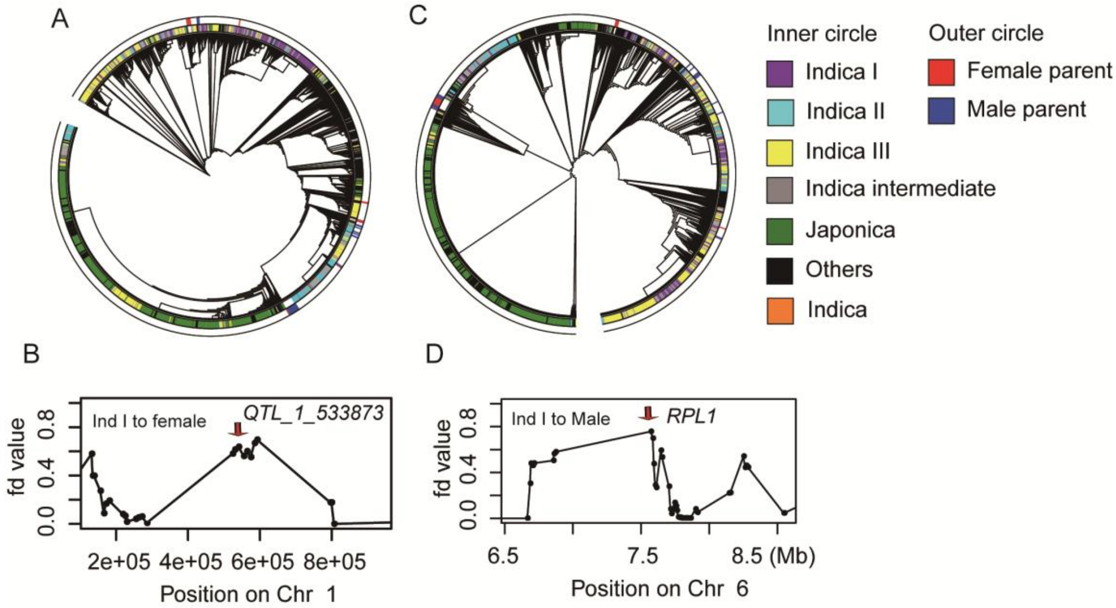

Fig. S4. Polygenetic trees of heterotic loci constructed to detect potential introgression events at these regions. (A) Polygenetic tree of a heterotic locus on chromosome 1 indicates that the locus was involved in genetic introgression from *Ind I* to female parents. Hybrid parents are annotated in the outer circle and landraces and improved varieties are in the inner circle. Different colors represent different sub-populations. (B) The f_d statistic for this locus shows an introgression signal from *Ind I* to the female parents. (C) Polygenetic tree of a heterotic locus that overlapped with the *RPL1* gene indicates that the locus was involved in genetic introgression from Ind I to female parents. Hybrid parents are annotated in the outer circle and landraces and improved varieties are in the inner circle. Different colors represent different sub-populations. (D) The f_d statistic for this locus shows an introgression signal from *Ind I* to male parents.

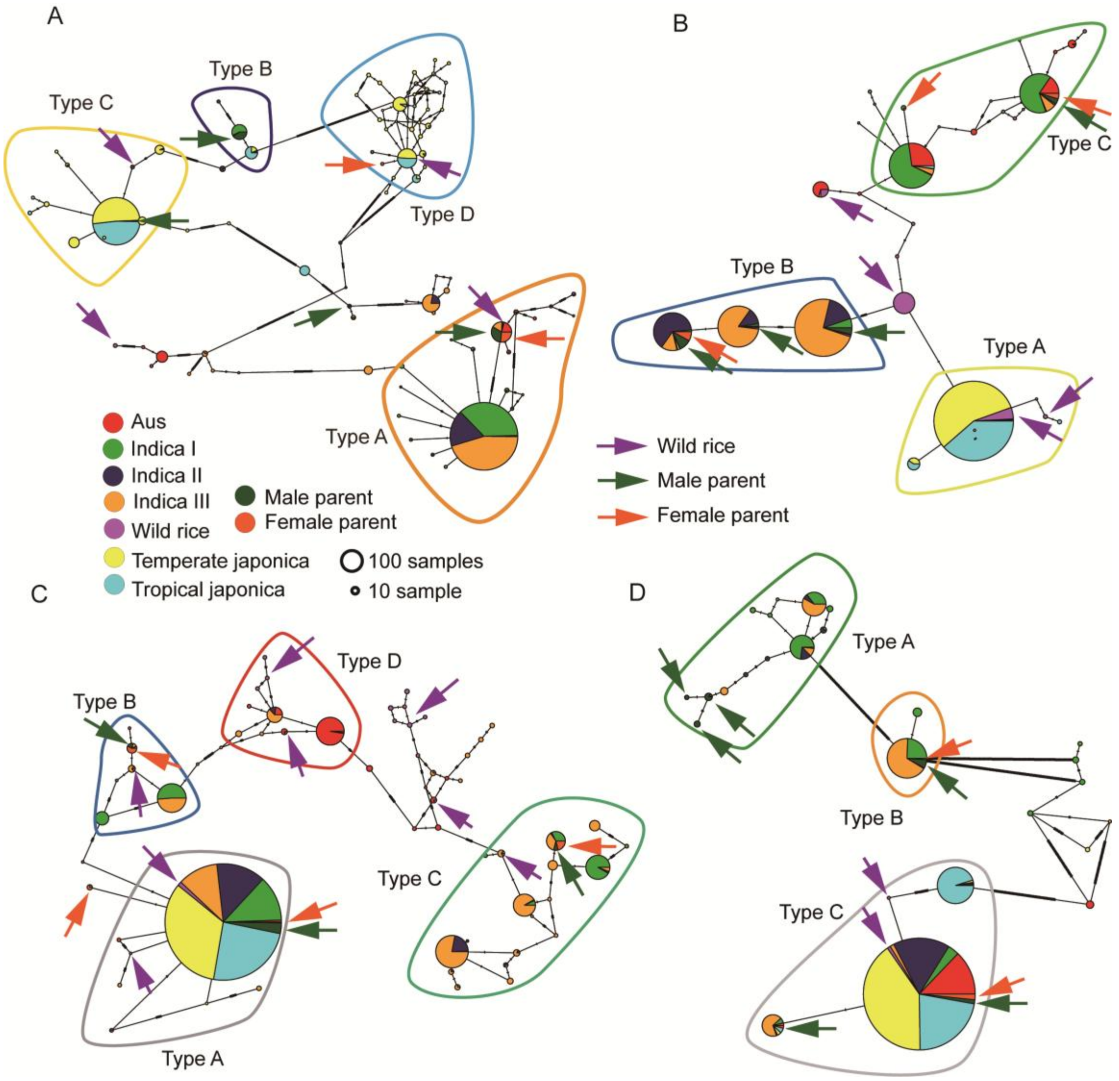

Fig. S5. Haplotype networks for genes *Ghd8/DTH8* (A), *Gn1a* (B), *IPA1* (C), and *RPL1* (D). The haplotye networks were constructed from genetic variants and their 5Kb-flanking regions. Haplotypes with a frequency < 2 were removed before constructing the networks. The networks were constructed using the minimum spanning method. Small bars on the edges of the haplotypes represent the number of variants among them. Haplotypes were classified based on the number of variants. Haplotypes of female parents, male parents and wild rice were marked with orange, darkgreen and purple arrows, respectively.

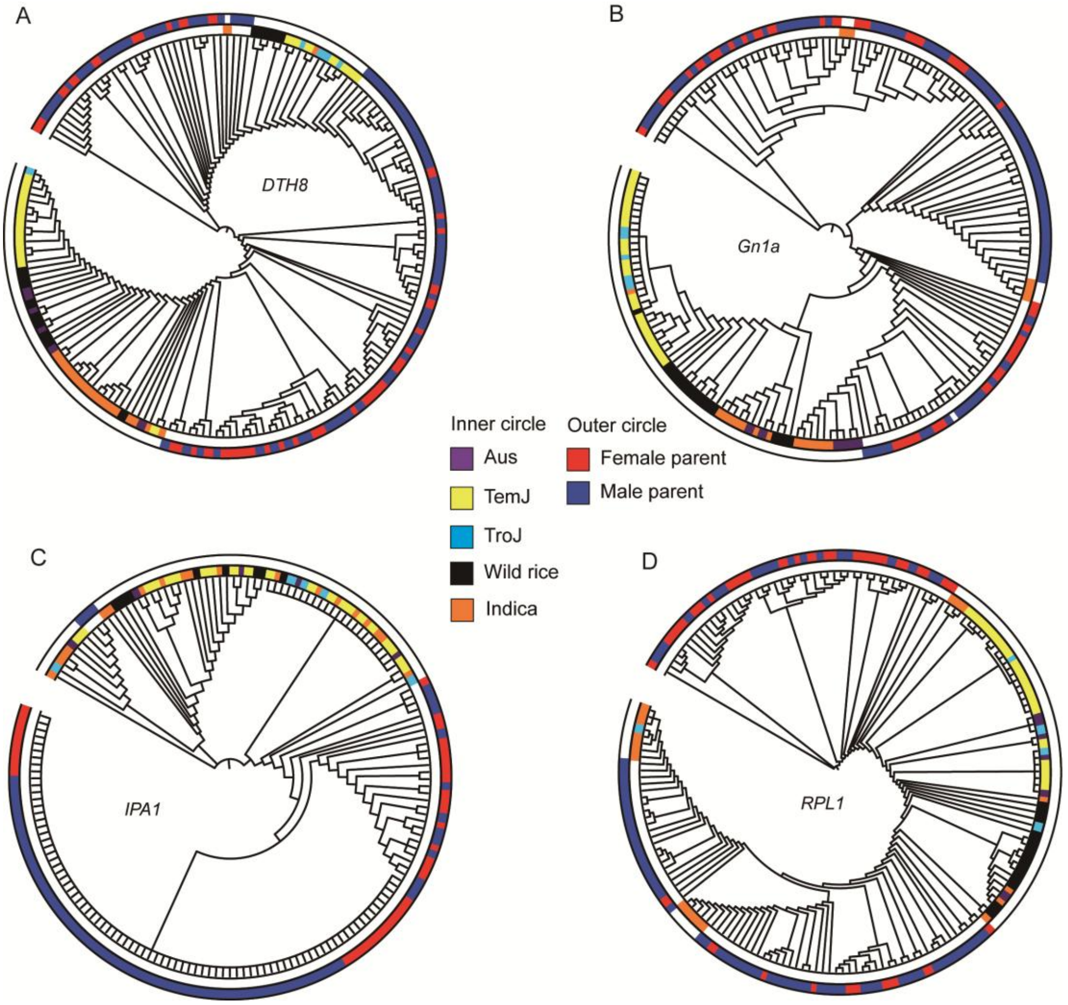

Fig. S6. Gene trees construct with hybrid parents and 66 representative accessions of O. *sativa* and O. *rufipogon* reported in previous study further support the alleles of heterotic genes *DTH8* (A), *Gn1a* (B), *IPA1* (C) and *RPL1* (D) exist in wild rice, divergent among rice sub-populations and were introgressed to rice hybrid parents. The genotypes of 66 representative accessions of O. *sativa* and O. *rufipogon* were called by BLAST to Nipponbare reference genome (IRGSP 1.0), and integrated with the genotypes of hybrid parents to construct the maximum likelihood tree. The result of all genes except *IPA1* is according with that of haplotype network, possibly because the 66 O. *sativa* and O. *rufipogon* did not include the derived landraces of hybrid parents at the locus.

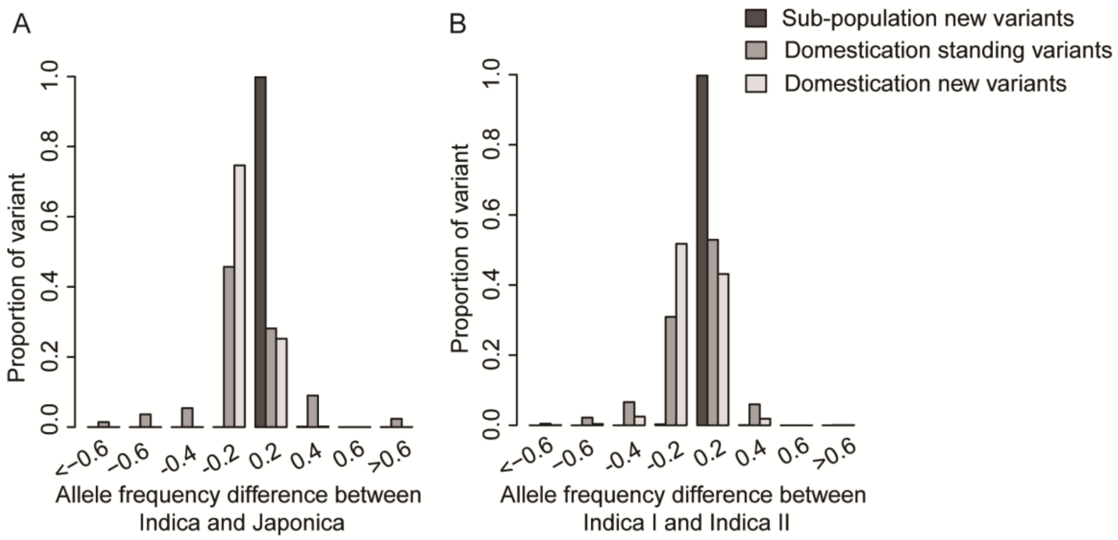

Fig. S7. Domestication of standing variants contributed to divergence at heterotic loci. (A) The AFDs between *indica* and *japonica* for the variants in the 25Kb flanking regions of the *indica*-*japonica*-originating heterotic loci. (B) The AFDs between *Ind I* and *Ind II* for the variants in the 25Kb flanking regions of the *Ind I*-*Ind II*-originating heterotic loci. Standing variants had higher AFDs than the other types of variants, suggesting that they contributed to divergence at the heterotic loci. We analyzed the origin of heterotic loci alleles based on the profiles of the polygenetic trees constructed from the 25Kb flanking variants of the loci. Heterotic loci containing alleles from *indica* and *japonica* were classified as *indica*-*japonica* origin. Loci with alleles from *Ind I* and *Ind II* were classified as Ind I-Ind II origin.

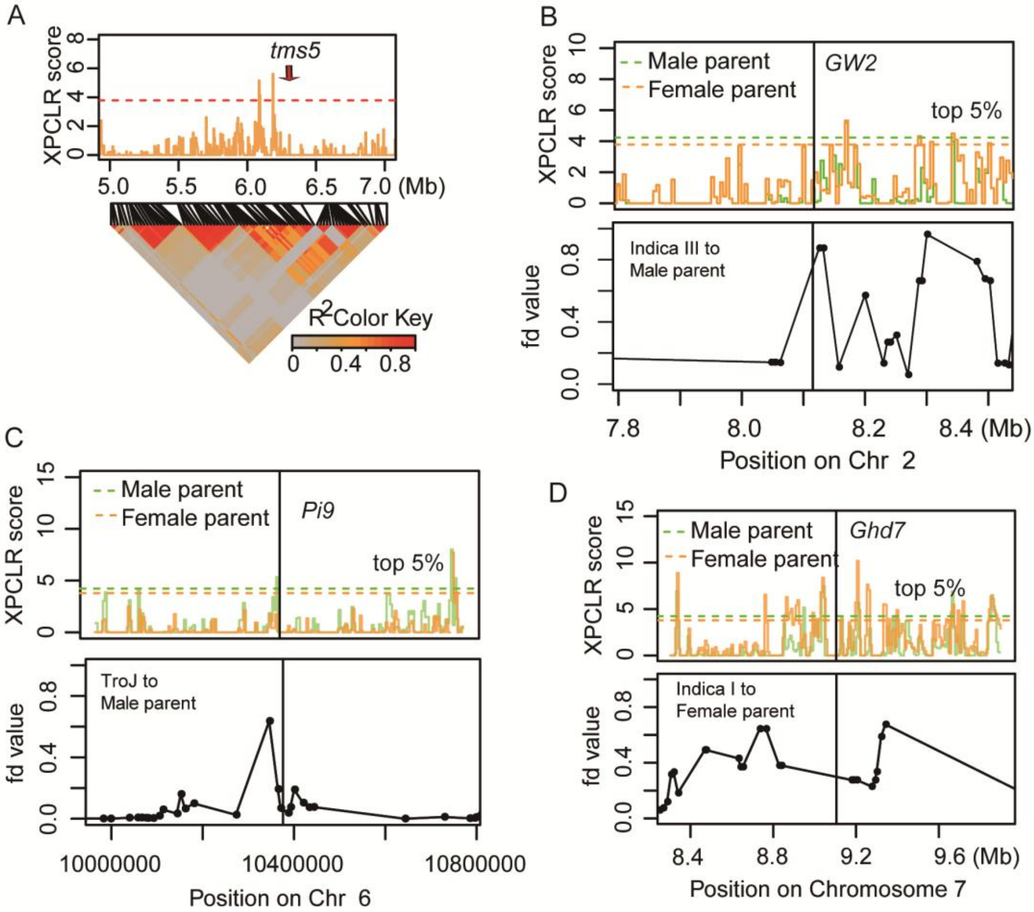

Fig. S8. Introgressed regions encompassing important genes were selected in parents of hybrids during breeding. (A) Selective signals at the *tms5* gene, a male-sterile gene widely used to develop hybrid female parents. (B) The *GW2* gene was involved in genetic introgression from *Ind III* to male parents and was selected in female parents during breeding. (C) The *pi9* gene was involved in genetic introgression from *TroJ* to male parents and was selected in male parents during breeding. (D) The *Ghd7* gene was involved in genetic introgression from *Ind I* to female parents and was selected in both male and female parents during breeding.

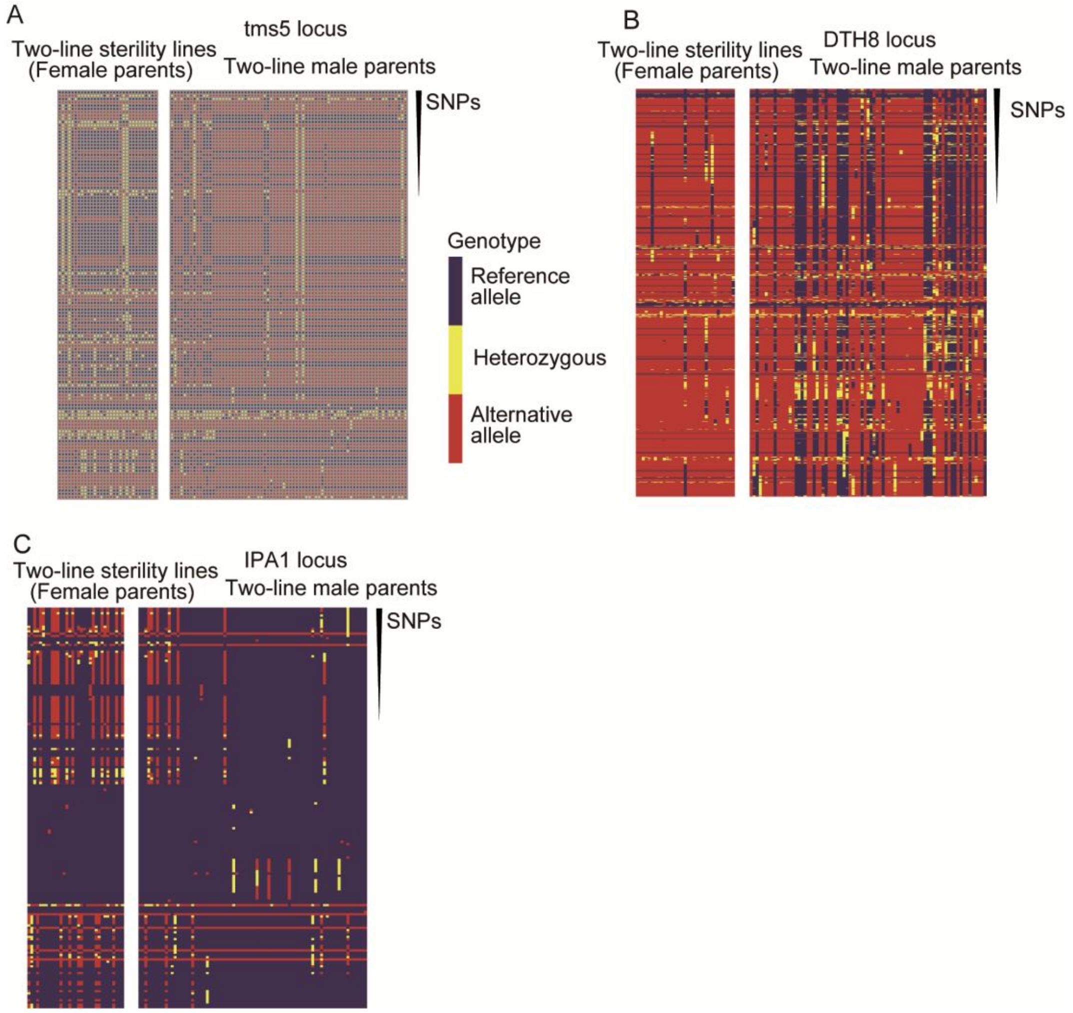

Fig. S9. Frtility-restorer relationship between male and female parents and selection towards dominant effect may explain the genome difference between male and female parents. (A). Genotypic difference between two-line male and female parents at *tms5* locus. Genic and 20 Kb flanking variants were employed to generate the heatmap. (B). Genotypic difference between two-line male and female parents at *DTH8* locus. Genic and 500 Kb flanking variants were employed to generate the heatmap. (C). Genotypic difference between two-line male and female parents at *IPA1* locus. Genic and 500 Kb flanking variants were employed to generate the heatmap.

Table S1. Correlation between dosage of introgressed genome and traits at male parent.

Trait[a]	Indica I	Indica III	*Aus*	Tropical japonica	Temperate japonica
BPP	0.03	0.24**	0.11	0.19*	0.00
BPT	0.09	0.19*	0.29**	0.38**	0.13
GYPP	-0.11	0.22*	-0.08	0.00	-0.13
PH	-0.01	0.27**	0.21*	0.16	-0.11
PW	0.07	0.18*	0.22*	0.30**	0.10
SWPP	0.07	0.16	0.08	0.29**	0.08
SWPT	0.11	0.20*	0.33**	0.42**	0.14
TN	-0.24**	-0.05	-0.39**	-0.35**	-0.22*

[a] BPP: Biomass per plant；BPT: Biomass per tiller；GYPP: Grain yield per plant；PH: Plant height；PW: Panicle weight；SWPP: Straw weight per plant；SWPT: Straw weight per tiller；TN: Tiller number. * indicates significant level at p-value = 0.05, ** at significant level of p-value = 0.01

Table S2. Correlation between dosage of introgressed genome and traits at female parent.

Trait	Indica I	Indica III	Aus	Tropical japonica	Temperate japonica
BPP	0.05	-0.18	-0.07	-0.24	-0.08
BPT	0.31*	0.14	-0.16	-0.25*	-0.21
GYPP	-0.03	-0.13	-0.16	-0.04	-0.14
PH	0.13	-0.28*	-0.23	-0.41**	-0.06
PW	0.30*	-0.12	-0.13	-0.15	-0.22
SWPP	0.03	-0.22	-0.18	-0.29*	-0.11
SWPT	0.29*	-0.16	-0.16	-0.32*	-0.18
PN	-0.49**	-0.05	-0.02	0.10	0.14

[a]BPP: Biomass per plant；BPT: Biomass per tiller；GYPP: Grain yield per plant；PH: Plant height；PW: Panicle weight；SWPP: Straw weight per plant；SWPT: Straw weight per tiller；TN: Tiller number. * indicates significant level at p-value = 0.05, ** at p-value = 0.01

Dataset S1 (separate file). List of hybrid male and female parents.

Dataset S2 (separate file). List of potential heterotic loci detected by GWAS among three experimental trials.

Dataset S3 (separate file). Introgressed regions of male parent.

Dataset S4 (separate file). Introgressed regions of female parent.

Dataset S5 (separate file). List of heterotic loci locate in introgressed regions.

Dataset S6 (separate file). The origin of male and female allele of heterotic loci based on polygenetic tree profiling.

Dataset S7 (separate file). Selective sweep on male parental genome.

Dataset S8 (separate file). Selective sweep on female parental genome.

Dataset S9 (separate file). List of QTLs located in the parental selective sweeps.

Dataset S10 (separate file). Original genotype and phenotype dataset used in the study.

References

1. Tamura K, Stecher G, Peterson D, Filipski A, & Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution* 30(12):2725-2729.
2. Chen H, *et al.* (2014) A high-density SNP genotyping array for rice biology and molecular breeding. *Molecular plant* 7(3):541-553.
3. Browning BL & Browning SR (2016) Genotype imputation with millions of reference samples. *The American Journal of Human Genetics* 98(1):116-126.
4. Kang HM, *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42(4):348-354.
5. Huang X, *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics* 42(11):961.
6. Huang X, *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490(7421):497.
7. Huang X, *et al.* (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature genetics* 44(1):32.
8. Xie W, *et al.* (2015) Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proceedings of the National Academy of Sciences* 112(39):E5411-E5419.
9. Wang W, *et al.* (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557(7703):43.
10. Alexander DH, Novembre J, & Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome research* 19(9):1655-1664.
11. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17(1):10-12.
12. Li H, *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
13. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20(9):1297-1303.
14. Martin SH, Davey JW, & Jiggins CD (2014) Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular biology and evolution* 32(1):244-257.
15. VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. *J Dairy Sci* 91(11):4414-4423.
16. de los Campos G, Gianola D, Rosa GJM, Weigel KA, & Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res* 92(4):295-308.
17. Price MN, Dehal PS, & Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* 5(3):e9490.
18. Weir BS & Cockerham CC (1984) Estimating *F-Statistics* for the Analysis of Population-Structure. *Evolution* 38(6):1358-1370.
19. Wei XJ, *et al.* (2010) DTH8 Suppresses Flowering in Rice, Influencing Plant Height and Yield Potential Simultaneously. *Plant Physiol* 153(4):1747-1758.
20. Yan WH, *et al.* (2011) A major QTL, *Ghd8*, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. *Molecular plant* 4(2):319-330.

21. Ashikari M, *et al.* (2005) Cytokinin oxidase regulates rice grain production. *Science* 309(5735):741-745.

22. Jiao Y, *et al.* (2010) Regulation of *OsSPL14* by *OsmiR156* defines ideal plant architecture in rice. *Nature genetics* 42(6):541.

23. Zhang C-, Yuan WY, & Zhang QF (2012) *RPL1*, a gene involved in epigenetic processes regulates phenotypic plasticity in rice. *Molecular plant* 5(2):482-493.

24. Leigh JW & Bryant D (2015) POPART: full-feature software for haplotype network construction. *Methods Ecol Evol* 6(9):1110-1116.

25. Chen H, Patterson N, & Reich D (2010) Population differentiation as a test for selective sweeps. *Genome research* 20(3):393-402.