Structural variations contribute to genetic diversity and heterosis in rice

Zhiwu Dan[1,2], Yunping Chen[1,2], Wenchao Huang[1]

[1]State Key Laboratory of Hybrid Rice, Key Laboratory for Research and Utilization of Heterosis in Indica Rice, the Ministry of Agriculture, The Yangtze River Valley Hybrid Rice Collaboration & Innovation Center, College of Life Sciences, Wuhan University, Wuhan, China.
[2]These authors contributed equally: Zhiwu Dan, Yunping Chen.

e-mail: wenchaoh@whu.edu.cn

Abstract

The contributions of structural variants (SVs) to rice heterosis, which has significantly boosted crop yield, remain largely unexplored. In this study, we assembled 12 high-quality rice genomes and revealed that SVs are pivotal in the evolution and variation of *indica* and *japonica*. Our findings indicate that SVs trigger gene expression changes which explain the variance in heterosis of 17 traits, highlighting the importance of SVs in harnessing rice heterosis.

Introduction

Heterosis refers to the phenomenon where heterozygous hybrids outperform their parents[1], significantly enhancing crop yield and helping to address global food shortages[2]. Despite the success cultivation of hybrid rice[3], the genetic mechanisms underlying rice heterosis remain obscure. Notably, the roles of structural variants, which require high-quality genome assemblies, have not been comprehensively studied in relation to rice heterosis. Here, we assembled 12 rice genomes and underscored the impact of structural variants on heterosis of 17 agronomic traits.

Results and discussion

We selected 12 *indica* and *japonica* accessions based on previously reported genome sequencing results[4] (Fig. 1a and Supplementary Fig. S1). These inbred lines (ILs) are phenotypically diverse varieties sourced from China, India, and Italy (Fig. 1b). The 12 ILs were crossed using a complete diallel design, and F_1 hybrids resulting from inter-subspecies exhibited significantly decreased seed setting rate and increased plant height compared to those from intra-subspecies (Fig. 1c and Supplementary Fig. S2), which is consistent with traditional hybrid breeding observations.

To perform genome assembly, we obtained genomic data for the 12 ILs, achieving an average sequencing depth at 30.46× (Supplementary Fig. S3). The average contig N50, scaffold N50, and total assembled genome sizes for the 12 ILs were 28.74 Mb, 31.75 Mb, and 396.36 Mb, respectively, with an average anchor ratio of 98.18%. The average quality value of bases (63.06), re-mapping rate of HiFi reads (99.51%), Benchmarking Universal Single-Copy Orthologs (BUSCO) (98.61%), and

long terminal repeats (LTRs) assembly index (22.13) all indicated high-quality genome assemblies (Supplementary Fig. S4). Compared to five other reference genomes, the newly assembled genomes for YB and Balilla were significantly different, highlighting the contributions of these selected accessions to current rice genomic resources (Supplementary Fig. S5).

A pan genome for the 12 ILs, which comprised 47,490 gene families and a total of 489,713 genes, was assembled. The gene number increased with the addition of genomes, reaching a plateau at $n = 10$, indicating the representativeness of the selected ILs (Fig. 1d). Unlike the trend observed in pan genes, the number of core genes decreased as the number increased, with the lowest point reached at $n = 12$, underscoring the specificity of each IL (Fig. 1e). Furthermore, the gene annotation ratios decreased from core to softcore to dispensable genes, demonstrating the conservation of core genes amidst diversification (Fig. 1f).

Genomic variants—especially structural variants (SVs: presence/absence variants, copy number variants, translocations, and inversions)—were compared between YB and the other 11 ILs. Consistent with the patterns observed for single nucleotide polymorphisms (SNPs) and insertions and deletions (InDels) (Fig. 1g), both the number and size of SVs gradually increased from *indica* to *japonica* (Fig. 1h-i). Two large inversions were detected in chromosome 6 (inversion_6: 5.83 Mb) and 8 (inversion_8: 5.90 Mb) (Fig. 1j), which were between the six *indica* and six *japonica* or exclusive to YB and JR2 (Supplementary Fig. S6-S10), respectively, indicating that inversions contribute to the evolutionary divergence of *indica* and *japonica*.

Unlike the pan-genome, the number of pan-SVs did not plateau even at $n = 11$, suggesting the necessity of including more accessions for a comprehensive exploration of SVs (Fig. 1k). Different from the core, softcore, and private SVs (Fig. 1l), the number of dispensable SVs increased from *indica* to *japonica* (Fig. 1m). The lengths of different SV types varied (Fig. 1n), with three large absence variants (AVs) (0.73-1.16 Mb) detected (Fig. 1o). Furthermore, DNA transposons and LTRs were the predominant types in repeat sequences (Fig. 1p).

We performed paired genomic comparisons and identified both unique and common genomic variants for each pair of ILs (Fig. 2a and Supplementary Fig. S11). The unique genomic variants were closely interconnected and approximately half of the unique genomic variants—57.45% of SNPs, 50.42% of InDels, and 50.56% of SVs—were located in intergenic regions (Fig. 2b), emphasizing the significance of these regions in the differentiation of ILs. In contrast to the other SV types, the majority of unique inversions (77.94%) were found in exonic regions, suggesting a link between inversions and gene expression levels. Furthermore, the genetic relatedness of the 12 ILs based on unique SVs, particularly translocations, showed slight variations when assessed through unique SNPs and InDels, demonstrating the function roles of SVs in shaping genomic differences between ILs (Fig. 2c and Supplementary Fig. S12).

Correlation analysis between the number of genomic variants and heterosis of the 17 traits was performed and the correlation patterns of the two groups of variants exhibited discrepancies (Fig. 2d). Importantly, the unique variants had more

significant correlations with heterosis compared to common ones (Fig. 2e). Though the number of unique SVs accounted for only 1.99% and 10.24% of those of SNPs and InDels, respectively, they demonstrated comparable degrees of correlation with heterosis (Supplementary Fig. S13). Among the five types of SVs, the correlation patterns of unique inversions—though fewer in number yet larger in size than unique translocations—indicated that SV size might influence heterosis. We observed that an increase in unique SVs corresponded with a slight rise in heterosis for seed setting rate in intra-subspecific F_1 hybrids, while a dramatic decrease in *indica-japonica* F_1 hybrids (Fig. 2f). The inter-subspecific hybrids possessed significantly more SVs than their intra-subspecific counterparts (Fig. 2g and Supplementary Fig. S14). Furthermore, SVs exhibited stronger correlations with heterosis for tiller number in inter-subspecific F_1 hybrids, emphasizing their roles in heterosis among inter-subspecies.

To further investigate the impact of SVs on gene expression changes, we conducted RNA-sequencing analysis for the 12 ILs (Fig. 2h and Supplementary Fig. S15). Drawing from previously reported inheritance patterns of metabolite levels in rice F_1 hybrids, we first proved that transcriptomic additive and partially dominant effects played significant roles in heterosis (Fig. 2i). We next focused on genes in the two inversions as previously shown (Fig. 2j). Consistent with the presence of inversion_6, the 12 ILs were accurately classified into two groups based on gene expression levels (Fig. 2k-l). Unexpectedly, genes in inversion_6 with additive and partially dominant effects were predictive of heterosis (Supplementary Fig. S16). Based on genes in inversion_8, JR2 was found to be closer to *indica* than to *japonica*. Though only a limited number of genes differed between the *indica* and *japonica*, the genes in inversion_8 well explained the variance in heterosis. Moreover, most genes in AV_6 were expressed and effectively divided the 12 ILs into two groups (Fig. 2m-n). The genes in AV_6 also demonstrated comparable predictability for heterosis with those of differentially expressed genes among ILs and inversions (Fig. 2o). These results confirmed the impact of SVs on gene expression levels, which in turn explained the variance in heterosis.

The models such as dominance (complementation of loci)[5-8], over-dominance (heterozygosity of loci)[9-11], and epistasis (interaction of loci)[12] have been proposed to elucidate the mechanisms underlying heterosis. In this study, our findings indicate that both inversions and AVs are linked to changes in gene expression levels, and their additive and partially dominant effects well explain the variance in heterosis. Consequently, we propose that the over-dominance model may serve as the genetic basis for rice heterosis, with presence/absence variants representing extreme forms of heterozygosity. Furthermore, the SVs distinguishing *indica* from *japonica* could be precisely engineered to harness the strong inter-subspecific heterosis.

Methods

Plant materials

Twelve rice inbred lines (ILs) were selected from accessions reported in previous studies[4,13-15]. Among these, six ILs—Yuetai B (YB), Chenghui9348 (9311K), 610234,

C418, Wuyunjing8, and Balilla—are well-established cultivars known for high grain yields or have been instrumental in breeding elite inbred or hybrid rice (https://www.ricedata.cn). Based on clustering results from metabolite profiles generated using GC- or LC-MS platforms[13,14], low-density SSR markers[15], and high-density SNPs or InDels[4], these 12 inbred lines are representatives of *indica* and *japonica* accessions. Plants of the selected inbred lines for genome sequencing were grown under long-day conditions, maintaining a 16-hour light/8-hour dark photoperiod at 30 °C in a plant culture room. The seedlings were sown in 96-well PCR plates (without bottoms) and placed in a plastic basin filled with Yoshida nutrient solution (pH = 5.5). After 20 days of growth, leaves from each genotype were harvested for experimentation. For full-length RNA sequencing of YB and Balilla, the plants were divided into four tissue parts: root, middle leaf (the youngest leaf), leaf without middle leaf, and stem.

Additionally, seedling of the 12 rice inbred lines were placed in an intelligent breeding room for RNA sequencing. The environmental conditions were set to 28 °C with a 16-hour light/8-hour dark photoperiod and controlled humidity. Each inbred line occupied half of a PCR plate and was maintained in a single basin containing four liters of nutrient solution, which was changed every three days. Fifteen-day-old seedings (without root) were harvested and frozen in liquid nitrogen. The middle five plants from each row were pooled to create one biological replicate, with three biological replicates established for each genotype.

The 12 inbred lines were crossed using a complete diallel design. As previously detailed[4], a total of 17 agronomic traits were recorded for both the inbred lines and their corresponding F_1 hybrids. These traits included grain yield per plant (YPP), grain number per panicle (GNP), seed setting rate (SSR), thousand-grain weight (TGW), plant height/tiller number (at seedling, elongation, flowering and maturation stages) (SSPH, ESPH, FSPH, MSPH and SSTN, ESTN, FSTN, MSTN), heading date (HD), tiller number per plant (TPP), panicle length (PL), and the number of primary and secondary branches per panicle (PBN and SBN). Better-parent heterosis was calculated using the equation: $BPH = (F_1 - P_{high})/P_{high}$, where P_{high} represents the value of the high parent. For the heading date, parents with shorter heading dates were designated as P_{high}, and the calculated values were subsequently multiplied by -1.

Genomic sequencing

SMRTbell libraries (~17 kb) for each rice inbred line were constructed using the SMRTbell Express Template Prep Kit 3.0 and sequenced on the Pacbio Revio platform (Pacific Biosciences) at Frasergen Bioinformatics (Wuhan, China). The SMRTbell libraries were size-selected using the BluePippin system (Sage Science) and purified with a 1× concentration of SMRTbell cleanup beads. Library quality was assessed using Femto pulse and Qubit 3.0 (Life Technologies), resulting in an average of 12.31 Gb of HiFi reads (approximately 30.46× coverage of the genome) obtained for each inbred line.

Hi-C libraries for YB and Balilla were prepared and sequenced using the BGI PE150 platform, generating a total of 51.81 Gb and 49.27 Gb of reads (approximately 127.801× coverage of the genome) for the two samples, respectively. The raw data

were filtered using Trimmomatic (v0.40)[16] with parameters LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:15. The quality of the clean data was evaluated with default parameters using FastQC[17].

RNA-sequencing

Full-length RNA sequencing technology was employed to aid in the annotation of genes in YB and Balilla. Total RNA was extracted from four tissues of the seedling samples, and equal amounts of RNA were pooled before constructing SMRTbell libraries. Full-length cDNA was synthesized using the SMARTer PCR cDNA Synthesis Kit (Clonetech) with the primer sequences "AAGCAGTGGTATCAACGCAGAGTACATGGGG" and "AAGCAGTGGTATCAACGCAGAGTAC". The products were purified and size-filtered (>1 kb) prior to sequencing. The PacBio Sequel II platform at Frasergen Bioinformatics was used for sequence analysis, with raw data (an average of 36.14 Gb) filtered and preprocessed using SMRTlink (v10.1, --minLength=50). Iso-seq analysis was performed to obtain full-length and non-concatemer (FLNC) sequences. These FLNC sequences were mapped to reference genomes using GMAP (V19.06.10)[18], retaining those with the highest PID to determine gene structure.

To investigate the impact of structural variants on gene expression changes, RNA sequencing was also conducted to seedlings of the 12 rice inbred lines (with three biological replicates per genotype). Total RNA was extracted for each sample to construct a DNA library. The Illumina Hiseq PE150 platform at Frasergen Bioinformatics was selected to generate 150 bp paired-end reads, with more than 6 Gb of raw data per sample. The raw data were filtered using SOAPnuke (v2.1.0, -lowQual=20, -nRate=0.005, -qualRate=0.5)[19]. Clean reads were aligned to the YB genome using Hisat2 (v2.2.1, --no-unal)[20]. Transcripts were assembled using StringTie[21], and clean reads were mapped to reference transcripts using bowtie2 (v2.3.5, --sensitive -k 50 --no-mixed --no-discordant)[22]. The read count of each transcript was calculated using RSEM (v1.3.3)[23].

Genome assembly

Draft genomes of the 12 rice inbred lines were assembled using HiFi reads with parameter -l3 via HiFiasm (v0.19.5)[24]. De novo genome assemblies of two representative ILs, YB and Balilla, were performed with the aid of Hi-C data, yielding an average sequencing depth at 127.80×. The clean Hi-C reads for YB and Balilla were aligned to contigs using Juicer (v1.6, -SP5M)[25]. Contact matrices were generated with 3D-DNA (v180922, -r1)[26], and corrections were applied using JuiceBox (v1.11.08). The genomes at the chromosome level of the remained 10 ILs were assembled based on one of these two reference genomes using Ragtag (v2.1.0, scaffold). We assessed the accuracy, identity, and completeness of the 12 assemblies. The completeness of the assemblies was assessed using Benchmarking Universal Single-Copy Orthologue (v3.0.2, embryophyta_odb10) analysis[27]. The long terminal repeats (LTR) assembly index (LAI) was calculated for each chromosome using LTR tools[28], including LTR_retriever (v2.9.0, -infinder ltrfinder.scn -inharvest ltrharvest.scn), LTR_FINDER (v1.0.7, -D 15000 -L 7000 -C -M 0.7), and GenomeTools (v1.6.2, ltrharvest -similar 70 -vic 10 -seed 20 -seqids yes -maxlenltr

7000 -maxtsd 6). HiFi reads were re-mapped to the assemblies with parameters "-ax map-hifi" using Minimap2 (v2.21)[29]. Additionally, quality value of HiFi reads was scored using a k-mer-based method with the Merqury pipeline (v1.3)[30].

Genome annotation

Repetitive sequences (including tandem repeats and transposable elements), protein-coding genes, and non-coding RNAs were annotated for the genomes of YB and Balilla. Tandem repeats were identified using the Tandem Repeats Finder (TRF, v4.09.1, 2 7 7 80 10 50 2000 -d -h)[31]. Transposable elements were annotated through an integrated approach combining do novo and homology-based (Repbase, www.girinst.org/repbase[32]) strategies. The de novo repeat library was constructed using LTR_FINDER (v1.0.7, -w 2 -C)[33] and RepeatModeler (v2.0.1)[34]. Transposable elements were detected using RepeatMasker (v4.1.2)[35] with parameters "-nolow -no_is -norna -parallel 2", searching against both the Repbase[32] and de novo repeat library. Additionally, RepeatProteinMask (1.36)[35] with parameters "-engine ncbi -noLowSimple -pvalue 0.0001" was utilized to identify transposable elements in Repbase[32]. The percentages of repetitive sequences in the genomes of YB and Balilla were 57.35% and 55.44%, respectively, with transposable elements (TEs) comprising 55.09% and 53.10% of their respective genomes. Among these TEs, DNA transposons and LTRs were the predominant types, accounting for 24.26-25.67% and 24.9-28.64% of the genomes, respectively.

To annotate gene structure and function, we integrated three strategies: homology-based (using five genomes), de novo, and transcriptome-based (full-length RNA sequencing) predictions. In the homology-based approach, protein-coding sequences from the genomes of 9311[36], Minghui63[37], Zhenshan97[37], R498[38], and Nipponbare[39] were compared to the YB and Balilla genomes using tblastn (v2.11.0+, -evalue 1e-05)[40]. Gene structures of the aligned sequences were predicted using Exonerate (v2.4.0, --model protein2genome)[41]. The lengths and GC content of genes and CDS, as well as the lengths and numbers of exon and intron per genes, were compared between the seven genomes. De novo annotation was conducted using Augustus (v3.4.0, --noInFrameStop=true --strand=both)[42], GlimmerHMM (v3.0.4)[43], and GeneMark (v4.65)[44]. The annotated gene sets were integrated to non-redundant sets using MAKER (v3.01.03)[45], and further polished by incorporating full-length RNA sequencing results using PASA (v2.4.1)[46]. RNA-sequencing results were aligned using HiSat2 (v.2.2.1)[20], and transcripts were assembled with StringTie (v.2.1.7)[47]. Additionally, transcripts were de novo assembled using Trinity (v2.8.5)[48]. This resulted in YB containing 44,332 genes and Balilla containing 43,706 genes. Gene functional analysis was conducted with Diamond BLASTP (v2.0.7, -c 1 -f 5 -e 1e-5 --very-sensitive -k 10)[49] against five databases: Non-Redundant (NR, ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), Kyoto Encyclopedia of Genes and Genomes (KEGG)[50], Gene Ontology (GO)[51], TrEMBL[52], and Swiss-Prot[52]. Protein domains were annotated using InterProScan (v5.50-84.0, --applications Pfam)[53] based on the InterPro database[54]. We revealed that 94.4% of the genes in YB and 95.35% in Balilla have assigned functions.

Non-coding RNAs, including rRNA, tRNA, snRNA, and miRNA, were also

annotated for the two genomes. tRNA was identified using tRNAscan-SE (v2.0.9)[55], while rRNA was detected using RNAmmer (v1.2, -S euk -multi -m lsu,ssu,tsu)[56]. miRNA and snRNA were annotated with Infernal (v1.1.4)[57] using Rfam (v14.6)[58]. miRNA and rRNA were present in higher percentages within the genomes. The completeness of the genomic annotations was further assessed using BUSCO (v5.2.2, embryophyta_odb10)[27]. Here, 98.5% and 98.6% of the gene sets in the two genomes contained complete BUSCOs in embryophyta.

For the annotation of the remaining ten genomes, Liftoff[59] was employed. The YB genome was chosen as the reference genome for five *indica* accessions: R465, Mianhui725, 9311K, 610234, and Qianlijing. The Balilla genome served as the reference for five *japonica* accessions: R4115, C418, 110080, Wuyunjing8, and JR2.

Pan-genome analysis

Protein sequences of the 12 inbred lines were initially aligned using BLASTp (-evalue 1e-5)[60]. The BLAST results were processed with OrthoFinder (v2.3.12)[61] to classify gene families. Gene families shared among all 12 inbred lines were defined as core gene families, while those identified in more than 90% of inbred lines were termed softcore gene families. Gene families specific to a single inbred line were classified as private gene families, and the remaining gene families were considered dispensable. Genes in the four types of gene families were further analyzed in the Pfam[62], GO[51], and KEGG[50] databases to determine their conservation.

Identification of genomic variants

Genomic variants, including single nucleotide polymorphisms (SNPs), insertions and deletions (InDels) (2-50 bp), and structural variants (SVs), were detected in the 12 rice inbred lines, with the genome of YB selected as the reference genome and the remained 11 as query genomes. SNPs and InDels were detected using MUMmer4 (Nucmer: -c 40 -l 90, Delta-filter: -1, Show-coords: -THrd, Show-SNPs: -ClrTH)[63]. SVs, including presence-absence variants (PAVs), inversions, translocations, and copy number variants (CNVs), were detected with the assistance of SyRI (v1.6.3)[64]. The distribution of SNPs, InDels, and SVs was analyzed across various genomic regions or functions, including: 1. Upstream (1 kb regions before the gene start codon); 2. Downstream (1 kb regions after the gene stop codon); 3. Intergenic regions; 4. UTR3 (3' untranslated regions); 5. UTR5 (5' untranslated regions); 6. Splicing junction regions (< 10 bp); 7. Intronic regions; 8. Exonic regions; 9. Synonymous or nonsynonymous mutants; 10. Frameshift or nonframeshift insertions/deletions; 11. Stopgain (gain of stop codons); 12. Stoploss (loss of stop codons). Variants simultaneously located in two regions, such as UTR5 and UTR3, were showed as Region 1;Region 2. The functions of SNPs, InDels, and SVs was annotated with ANNOVAR[65]. Genomic variants between each pair of the 11 inbred lines were classified into unique and common variants, representing different and shared variants, respectively. On average, unique genomic variants comprised 2,018,539 SNPs, 391,298 InDels, and 40,082 SVs (17,690 AVs, 11,975 PVs, 6,327 CNVs, 149 inversions, and 3,941 translocations). For correlations with heterosis, no notable change in correlation patterns was observed between stop-gain and stop-loss variants in unique SNPs or InDels, nor between synonymous and nonsynonymous variants in

unique SNPs, or between frameshift and nonframeshift insertions/deletions in unique InDels, all of which emphasized the importance of variations between the two parents in heterosis.

Pan-SV analysis

A pan-SV analysis was conducted for the SVs of the 11 query genomes using Vg (v1.43)[66]. For SVs excluding PVs, an SV was assigned when the ratio of overlapping to total locations in two inbred lines exceeded 50%. For PVs, a PV was designated when the distance between PV locations in two inbred lines was less than 5 bp and the length ratio between the shorter and longer samples was greater than 50%. Circles illustrating genomic variants on chromosomes were generated using Circos[67]. Genetic relatedness among the 12 inbred lines was visualized using Cytoscape[68]. Core, softcore, dispensable, and private SVs were assigned using the same criteria as pan-genes. Among the total SVs, 878 identified in all 11 ILs were classified as core SVs, while 1,640 detected in 10 ILs were categorized as softcore SVs. Additionally, 66,655 SVs identified in 2-9 ILs were classified as dispensable SVs, and 51,413 identified in only one IL were considered private SVs.

Construction of phylogenetic trees

A phylogenetic tree for 223 rice inbred lines was constructed using genotypic data based on 33 pairs of single sequence repeat markers from a previous study[4]. Phylogenetic trees based on the number of differential SNPs, InDels, and CNVs among the 12 rice inbred lines were built using the neighbour-joining method with MEGA (v10.2.6)[69]. The number of SNPs, InDels, and CNVs of the inbred lines were retrieved from a previous report[4].

Normalization and differential analysis of transcriptomics data

Read counts of 53,564 genes in each experimental sample were used for data normalization and differential analyses through MetaboAnalyst 5.0[70]. The interquantile range was selected to filter genes, with a threshold set at 40%. Sum, none, and auto-scaling were chosen for sample normalization, data transformation, and data scaling, respectively. Differential analysis among the 12 inbred lines was implemented using the ANOVA module in MetaboAnalyst 5.0[70], with a raw *P*-value cutoff set at 0.001 for post-hoc analysis. Principal component analysis (PCA) and heatmap results for the 12 inbred lines were obtained through corresponding modules on the website. PERMANOVA was conducted in PCA and the permutation number was 999. Euclidean for distance measure and ward for clustering method were set in heatmaps. Differential genes were selected from the original read count table and averaged across the three biological replicates. The averaged data were further filtered and normalized for each genotype with the above parameters. Due to the limited number of genes in inversions and absence variants, no filtering was applied to these genes. The genes in three large AVs, including AV_6, AV_8, and AV_11, were analyzed. Despite AV_8 and AV_11 being larger in size, the genes in these AVs were rarely expressed in the seedling samples. Only genes in AV_6 was further analyzed.

Partial least square (PLS) analysis

PLS analysis was conducted to investigate the relationship between various inheritance patterns at the transcriptomic level and better-parent heterosis of the 17

agronomic traits using SPSS 20.0. The maximum number of latent factors was set to 50. The inheritance patterns were calculated based on normalized gene expression levels using equations outlines in a previously study[4]. The genetic effects included: 1. Additive effect: average levels of the female (P_f) and male (P_m) parents; 2. Partially dominant effects (PD): partially dominant _0.3 (PD_0.3) = $P_m+0.3\times(P_f-P_m)$ and partially dominant _0.7 (PD_0.7) = $P_m+0.7\times(P_f-P_m)$; 3. Dominant effects: transcript levels of the female (dominant_female) and male (dominant_ male) parents.

Statistical analysis

Boxplots were created using BoxPlotR[71]. Analysis of variance (ANOVA) with the least significant difference for post hoc comparisons was performed using SPSS 20.0. This analysis compared the following parameters: lengths of genes, lengths of coding sequences, lengths of exons, lengths of introns, GC content of genes, GC content of coding sequences, number of exons, and number of introns across seven genomes (including two newly assembled genomes, YB and Balilla, along with five representative rice genomes). Additionally, we assessed the number of genomic variants and the degrees of better-parent heterosis for 17 traits in F_1 hybrids from both intra- and inter-subspecies. The adjusted R^2 values of partial least square models for heterosis across traits were also compared. An independent samples *t*-test (weighted) was conducted to evaluate the number of correlations between unique and common genomic variants and heterosis. SPSS was also employed to conduct Spearman correlation to investigate the relationships between different types of unique genomic variants and between genomic variants and heterosis.

Data availability

Raw reads of PacBio SMRT sequencing and RNA-sequencing and genome assemblies of the 12 rice inbred lines have been deposited in the NCBI database under BioProject accession PRJNA1178436. Source data are available with this paper.

Code availability

All software used in the study is publicly available on the internet.

Acknowledgements

We thank C. Zhu for funding of our rice research. This study was supported by the National Natural Science Foundation of China (31801439 to Z.D., 32472185 to W.H., and 32101667 to Y.C.), the China Postdoctoral Science Foundation (2019M660186 to Z.D., and 2022T150500 to Z.D., and 2023T160497 to Y.C.), the Key Research and Development Program of Hubei Province (2022BFE003 to W.H.), the National Key R&D Program of China (2017YFD0100400 to W.H.) and the Hubei Agriculture Science and Technology Innovation Center Program to W.H..

Author contributions

Z.D. conceived and designed the project. Z.D. and Y.C. prepared rice seedling samples, conducted genomic, transcriptomic, phenotypic, and statistical analyses, and interpretated the results. W.H. collected the rice accessions. Z.D. and W.H.

constructed the diallel cross population and recorded the phenotypes. Z.D., Y.C., and W.H. wrote the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Shull, G.H. What is "heterosis"? *Genetics* **33**, 439-446 (1948).
2. Hochholdinger, F. & Yu, P. Molecular concepts to explain heterosis in crops. *Trends Plant Sci.*, 10.1016/j.tplants.2024.07.018 (2024).
3. He, Q. *et al.* Hybrid rice. *Engineering* **6**, 967-973 (2020).
4. Dan, Z. *et al.* Additive and partially dominant effects from genomic variation contribute to rice heterosis. *bioRxiv*, 10.1101/2024.07.16.603817 (2024).
5. Bruce, A.B. The Mendelian theory of heredity and the augmentation of vigor. *Science* **32**, 627-628 (1910).
6. Jones, D.F. Dominance of linked factors as a means of accounting for heterosis. *Proc Natl Acad Sci U S A* **3**, 310-312 (1917).
7. Wang, B. *et al.* De novo genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. *Nat. Genet.* **55**, 312-323 (2023).
8. Gu, Z. *et al.* Structure and function of rice hybrid genomes reveal genetic basis and optimal performance of heterosis. *Nat. Genet.* **55**, 1745-1756 (2023).
9. Krieger, U., Lippman, Z.B. & Zamir, D. The flowering gene *SINGLE FLOWER TRUSS* drives heterosis for yield in tomato. *Nat. Genet.* **42**, 459-463 (2010).
10. Semel, Y. *et al.* Overdominant quantitative trait loci for yield and fitness in tomato. *Proc. Natl. Acad. Sci. USA* **103**, 12981-12986 (2006).
11. East, E.M. Heterosis. *Genetics* **21**, 375-397 (1936).
12. Jiang, Y., Schmidt, R.H., Zhao, Y. & Reif, J.C. A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nat. Genet.* **49**, 1741-1746 (2017).
13. Dan, Z. *et al.* Metabolic prediction of important agronomic traits in hybrid rice (*Oryza sativa* L.). *Sci. Rep.* **6**, 21732 (2016).
14. Dan, Z., Chen, Y., Zhao, W., Wang, Q. & Huang, W. Metabolome-based prediction of yield heterosis contributes to the breeding of elite rice. *Life Sci. Alliance* **3**, e201900551 (2020).
15. Dan, Z. *et al.* Balance between a higher degree of heterosis and increased reproductive isolation: a strategic design for breeding inter-subspecific hybrid rice. *PLoS ONE* **9**, e93122 (2014).
16. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).
17. Wingett, S.W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* **7**, 1338 (2018).
18. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-75 (2005).
19. Chen, Y. *et al.* SOAPnuke: a MapReduce acceleration-supported software for integrated

quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1-6 (2018).
20. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907-915 (2019).
21. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290-5 (2015).
22. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 7 (2010).
23. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
24. Cheng, H. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332-1335 (2022).
25. Durand, N.C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95-8 (2016).
26. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).
27. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-2 (2015).
28. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
29. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
30. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
31. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).
32. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
33. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265-8 (2007).
34. Flynn, J.M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**, 9451-9457 (2020).
35. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, 4 10 1-4 10 14 (2009).
36. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79-92 (2002).
37. Zhang, J. *et al.* Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. USA* **113**, E5163-E5171 (2016).
38. Du, H. *et al.* Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* **8**, 15324 (2017).

39. Shang, L. *et al.* A complete assembly of the rice Nipponbare reference genome. *Mol. Plant* **16**, 1232-1236 (2023).
40. Gertz, E.M., Yu, Y.K., Agarwala, R., Schaffer, A.A. & Altschul, S.F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 41 (2006).
41. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
42. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435-9 (2006).
43. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-9 (2004).
44. Borodovsky, M. & Mcininch, J. GENMARK: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17**, 123-133 (1993).
45. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
46. Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654-66 (2003).
47. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
48. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652 (2011).
49. Buchfink, B., Reuter, K. & Drost, H.G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366-368 (2021).
50. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199-205 (2014).
51. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29 (2000).
52. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-70 (2003).
53. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-40 (2014).
54. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, D213-21 (2015).
55. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955-964 (1997).
56. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100-8 (2007).
57. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-5 (2013).
58. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121-4 (2005).
59. Shumate, A. & Salzberg, S.L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639-1643 (2021).

60. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
61. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
62. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412-D419 (2021).
63. Marcais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
64. Goel, M., Sun, H., Jiao, W.B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
65. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
66. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875-879 (2018).
67. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-45 (2009).
68. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482-517 (2019).
69. Stecher, G., Tamura, K. & Kumar, S. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol* **37**, 1237-1239 (2020).
70. Pang, Z. *et al.* Using MetaboAnalyst 5.0 for LC-HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. *Nat. Protoc.* (2022).
71. Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. BoxPlotR: a web tool for generation of box plots. *Nat. Methods* **11**, 121-2 (2014).

Figures and figure legends

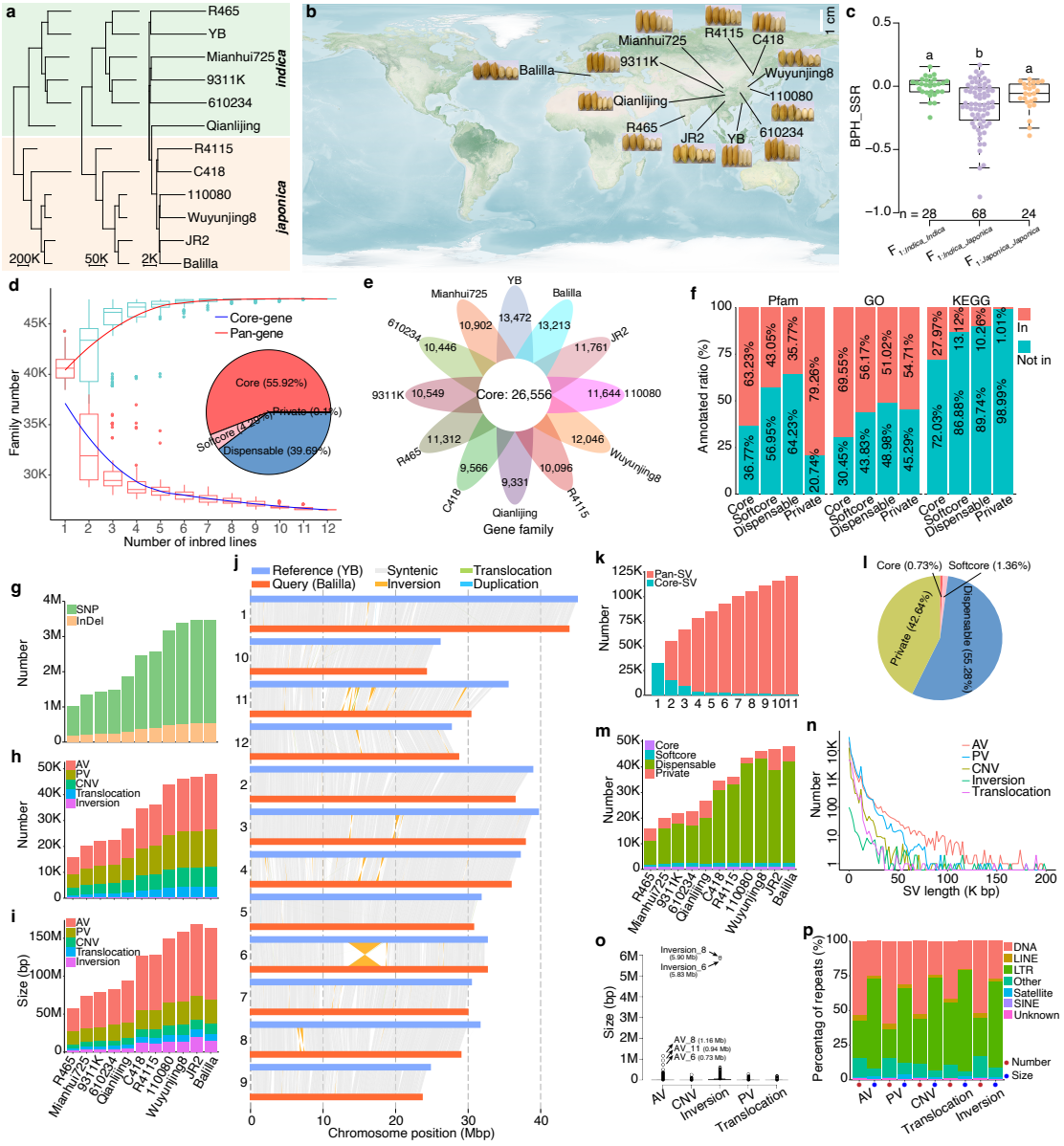

Fig. 1 Pan-genome and pan-structural variants of 12 rice inbred lines. a, Neighbour-joining trees of 12 rice ILs based on SNPs, InDels, and CNVs from left to right. **b**, Geographic origins of 12 ILs. **c**, Better-parent heterosis for seed-setting rate (BPH_SSR) of both intra- and inter-subspecific F_1 hybrids. **d**, Variation of gene families in the pan- and core-genome. **e**, Core and non-core gene families. **f**, Gene annotation ratios cross three databases. **g-i**, Number and size of genomic variants in 11 ILs. **j**, Genomic profiles of YB and Balilla. **k**, Variation of SVs in the pan- and core-SVs. **l**, Percentages of four types of SVs. **m**, Number of four types of SVs in 11 ILs. **n-o**, Number and size of SVs. **p**, Repeat types for all SVs.

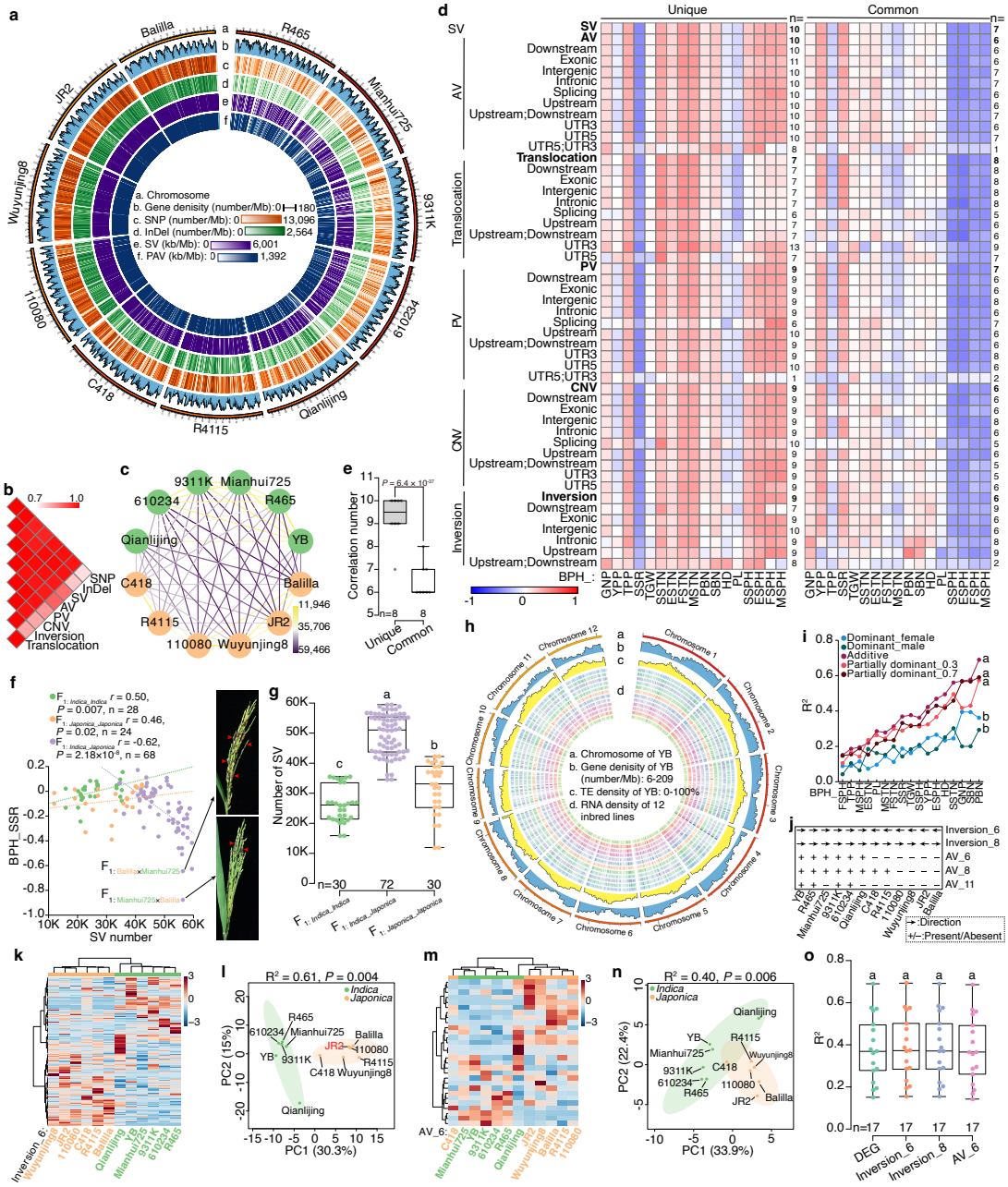

Fig. 2 Impact of SVs on gene expression and heterosis. a, Genomic comparisons between YB and the other 11 ILs. **b**, Heatmap for the correlations between different types of unique genomic variants. **c**, Genetic relatedness of 12 ILs based on unique SVs. **d**, Correlation patterns between SVs and heterosis of 17 traits. **e**, Correlation number of unique and common variants. *P*-value indicates independent *t*-test (weighted). **f**, Relationship between SV number and BPH_SSR in both intra- and inter-subspecific F_1 hybrids. Mature seeds with black and green hulls are full and empty seeds, respectively. *P*-values indicate Spearman correlation. **g**, SV number of intra- and inter-subspecific F_1 hybrids. **h**, Transcriptomic profiles of 12 ILs. **i**, Changes in adjusted R^2 for heterosis of 17 traits. **j**, Genomic structure of two inversions and three AVs. **k-n**, Heatmaps and PCA plots of 12 ILs based on expression levels of genes in Inversion_6 and AV_6. *P*-values indicate PERMANOVA. **o**, Comparisons of adjusted R^2 for predicting heterosis based on

differentially expressed genes (DEGs) of 12 ILs, Inversion_6, Inversion_8, and AV_6. ANOVA with LSD in post-hoc test (P-value < 0.05) was performed in **i** and **o**.