Title page

Structural variations contribute to subspeciation 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 and Yunping Chen.

e-mail: wenchaoh@whu.edu.cn

Abstract

The exploitation of heterosis in rice, namely hybrid rice, has significantly improved global food production. However, the molecular mechanisms underlying rice heterosis, especially in heterosis from inter-subspecies, remain poorly understood. Here, we assemble high-quality genomes for 12 representative *indica* and *japonica* rice. Comparative genomic analyses reveal abundant genetic variations and highlight pivotal roles of structural variants (SVs) in rice subspeciation. The numbers of SVs between parental lines are significantly correlated with heterosis across 17 agronomic traits, with distinct correlation patterns for inter- and intra-subspecific F_1 hybrids. We detect SVs in *S5-ORF5* and functionally validated their roles in heterosis for seed setting rate, underscoring the importance of SVs in breeding inter-subspecific hybrid rice. The SVs, especially inversions that are mostly located in exonic regions, drive gene expression changes that account for heterosis variation. Our findings demonstrate the contributions of SVs to subspeciation and gene expression variations for harnessing rice heterosis.

Keywords: Structural variant, Heterosis, Rice, Subspeciation, Gene expression

Introduction

Heterosis refers to the phenomenon where heterozygous F_1 hybrids exhibit superior performance compared to their parental lines[1], significantly enhancing crop yield and contributing to global food security[2]. Though several well-known genetic models—including dominance (complementation of loci)[3-6], over-dominance (heterozygosity of loci)[7-9], and epistasis (interaction of loci)[10, 11]—have been proposed to explain heterosis, the molecular mechanisms underlying crop heterosis remain largely unknown. Currently, breeding elite hybrid crops relies on labor-intensive and time-consuming field practices. To meet rising food demands driven by population growth and environmental changes, it is essential to further elucidate the mechanisms of crop heterosis and improve the efficiency of hybrid breeding.

The successful breeding and widespread cultivation of hybrid rice have dramatically increased grain yield in China and other Asia countries[12]. To study the molecular mechanisms of yield heterosis in rice, researchers have constructed diverse hybrid populations and performed extensive sequencing and phenotyping analyses. While the contributions of small genomic variants, such as single nucleotide polymorphisms (SNPs) and insertions and deletions (InDels), have been well studied[6, 13-16], the roles of larger structural variants (SVs, > 50 bp) in heterosis, especially the strong heterosis observed in inter-subspecific F_1 hybrids, remains poorly understood. A major limitation in studying SVs in rice heterosis has been the lack of high-quality genome assemblies of parental inbred lines for accurate variant detection.

Recent advances in sequencing technologies, including PacBio, BioNano, and Hi-

67 C platforms, have enabled the assembly of high-quality genomes. Over the past five
68 years, the completeness and accuracy of genome assemblies have significantly
69 improved, leading to the release of multiple rice genomes (including both *indica* and
70 *japonica* subspecies)[17-20]. Given that SVs have been shown to play important roles in
71 tomato improvement and maize heterosis[5, 21], they are likely also critical for rice
72 heterosis.

73 In this study, we selected 12 representative *indica* and *japonica* inbred lines and
74 assembled their genomes for comparative analysis. We identified genomic variants,
75 particularly SVs, and assessed their functional implications in rice sub-speciation.
76 Using these inbred lines as parents, we constructed a diallel cross population and
77 collected phenotypic data for 17 field traits. In combination with gene expression data,
78 we further investigated the contributions of SVs to rice heterosis.

79 Results

80 Selection of 12 representative rice inbred lines

81 To investigate the molecular mechanisms underlying rice heterosis, we selected 12
82 representative *indica* and *japonica* accessions from 223 inbred lines (ILs) based on
83 previous genotyping data[22] (Supplementary Fig. S1). Meanwhile, these 12 ILs were
84 clustered into distinct *indica* and *japonica* groups (Fig. 1a), with different types of
85 genomic variants analyzed from the reported Illumina sequencing data[22]. Among the
86 selected ILs, six—Yuetai B (YB), Chenghui9348 (9311K), 610234, C418, Wuyunjing8,
87 and Balilla—are well-established cultivars known for high grain yields or their
88 widespread use in breeding elite inbred or hybrid rice (https://www.ricedata.cn). These

ILs represent phenotypically and geographically diverse varieties originating from China, India, and Italy (Fig. 1b).

Using a complete diallel cross design, we generated F_1 hybrids from the 12 ILs. We observed that F_1 hybrids derived from inter-subspecies exhibited significantly reduced seed setting rate and increased plant height compared to those of intra-subspecific hybrids (Fig. 1c and Supplementary Fig. S2), which is consistent with traditional hybrid breeding observations.

Genome assembly and annotation of 12 inbred lines

To generate genome assemblies, we obtained genomic data using single-molecule real-time (SMRT) technology on the PacBio Revio platform for the 12 ILs, achieving an average sequencing depth at 30.46× (Supplementary Tables S1-S2). De novo genome assembly for two representative ILs, YB (*indica*) and Balilla (*japonica*), were performed with the assistance of Hi-C data, yielding an average sequencing depth at 127.80× (Supplementary Fig. S3 and Supplementary Table S3). The genomes at the chromosome level of the remaining ten ILs were assembled using YB and Balilla as reference genomes for *indica* and *japonica* subgroups, respectively. The average contig N50, scaffold N50, and total assembled genome sizes for final genome assemblies of the 12 ILs were 28.74 Mb, 31.75 Mb, and 396.36 Mb, respectively, with an average anchor ratio of 98.18%. The high-quality and completeness of these genome assemblies were further confirmed by metrics such as 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)

111 (Supplementary Fig. S4).

112 We annotated repetitive sequences (tandem repeats and transposable elements),
113 protein-coding genes, and non-coding RNAs for the genomes of YB and Balilla
114 (Supplementary Table S4). Repetitive sequences constituted 57.35% (YB) and 55.44%
115 (Balilla) of their genomes, with transposable elements (TEs) comprising 55.09% and
116 53.10%, respectively. Among these TEs, DNA transposons and LTRs were the
117 predominant types, accounting for 24.26-25.67% and 24.9-28.64% of the genomes,
118 respectively. Three strategies, including homology-based (using five previously
119 reported genomes), *de novo*, and transcriptome-based (full-length RNA sequencing)
120 predictions, were integrated for the annotation of protein-coding genes. This resulted in
121 YB containing 44,332 genes and Balilla containing 43,706 genes. We revealed that 94.4%
122 of the annotated genes in YB and 95.35% in Balilla have assigned functions in one of
123 the following six databases: InterPro, Gene Ontology (GO), Kyoto Encyclopedia of
124 Genes and Genomes (KEGG), Swiss-Prot, TrEMBL, and Non-Redundant. Compared
125 to the five established genome assemblies, including Nipponbare, 9311, Minghui63,
126 Zhenshan97, and R498, these newly assembled genomes were significantly different in
127 the lengths of genes and coding sequences, highlighting the value of these new
128 assemblies to current rice genomic resources (Supplementary Fig. S5). Moreover,
129 miRNA and rRNA were the two predominant types of non-coding RNAs in the genome
130 assemblies of both YB and Balilla, which accounted for 0.26-0.28% and 0.34-0.96% of
131 the genomes, respectively.

132 **Validation of genome assemblies using *Phr1***

133 To further validate accuracy of the 12 newly assembled genomes, we analyzed *Phr1*
134 (Fig. 1d), a gene underlies the phenol reaction phenotype of *indica* and *japonica*[23].
135 Phylogenetic result of *Phr1* sequences from the 12 ILs, along with 9311 and
136 Nipponbare, confirmed their expected clustering patterns of subspecies. Compared to
137 9311, *Phr1* contained 24 SNPs and six InDels. Among these InDels, there existed three
138 insertions (1-2, 6, and 7 or 14 bp) and three deletions (2, 18, and 29 bp), with only the
139 18 and 29 bp deletions located in exons. The result of phenol reaction experiment
140 revealed that hulls of five ILs lacking the 18/29 bp deletions turned black within one
141 day, while the remaining nine remained yellow after seven days (Fig. 1e). These results
142 validated the assembly accuracy of these new genomes and implicated importance of
143 the 18/29 bp deletions in outcomes of subspecies-specific phenol reaction.

144 **Pan-genome assembly of 12 inbred lines**

145 A pan genome of the 12 ILs was constructed, comprising 47,490 gene families (Fig. 1f).
146 Among the whole gene families, 26,556 (with 345,960 genes) were identified in all 12
147 ILs and defined as core-gene families (55.92% of gene families and 70.70% of genes),
148 2,037 (with 24,040 genes) identified in 11 ILs were defined as softcore gene families
149 (4.29% of gene families and 4.91% of genes), 18,850 (with 118,436 genes) identified
150 in 2-10 ILs were defined as dispensable families (39.69% of gene families and 24.20%
151 of genes), and 47 (with 892 genes) identified in only one IL were defined as private
152 gene families (0.10% of gene families and 0.18% of genes). The numbers of gene
153 family increased with the addition of genomes, reaching a plateau at $n = 10$, indicating
154 the representativeness of the selected ILs (Fig. 1f). Unlike the trend observed in pan

155 genome, the numbers of core gene family decreased as the number increased, with the
156 lowest point reached at n = 12, underscoring the specificity of each IL (Fig. 1f). For the
157 numbers of non-core gene families among these ILs, YB and Qianlijing had the highest
158 and lowest amount, respectively, showing functional variations between ILs (Fig. 1g).
159 Furthermore, we analyzed the proportion of genes with annotated functions in GO,
160 KEGG, and protein domain annotation (Pfam). The result indicated that the annotation
161 ratios decreased from core to softcore to dispensable genes (Pfam: 63.23%, 43.05%,
162 and 35.77%; GO: 69.55%, 56.17%, and 51.02%; and KEGG: 27.97%, 13.12%, and
163 10.26%), demonstrating the conservation of core genes amidst diversification (Fig. 1h).

164 **Pan-structural variants in the 12 inbred lines**

165 Genomic variants—especially SVs (presence/absence variants, copy number variants,
166 translocations, and inversions)—were identified through comparing YB and the other
167 11 ILs. Consistent with the patterns observed for SNPs and InDels (Fig. 2a), both the
168 number and size (about 57-168 Mb) of SVs gradually increased from *indica* to *japonica*
169 (Fig. 2b-c). And the *japonica* ILs all harbored more SVs than the *indica*. Notably, two
170 large inversions were detected on chromosome 6 (inversion_6: 5.83 Mb) and 8
171 (inversion_8: 5.90 Mb) (Fig. 2d), which were between the six *indica* and six *japonica*
172 or exclusive to JR2 (Supplementary Fig. S6-S10), respectively, indicating that
173 inversions contribute to the evolutionary divergence of *indica* and *japonica*.

174 Unlike the changing trend of pan-genome, the numbers of pan-SVs did not plateau
175 even at n = 11, underscoring the necessity of including more accessions for a
176 comprehensive exploration of SVs (Fig. 2e). Core, softcore, dispensable, and private

177 SVs were assigned using the same criteria as the pan-genome. Among the total SVs,
178 878 (0.73%) identified in 11 ILs were classified as core SVs, while 1,640 (1.36%)
179 detected in 10 ILs were categorized as softcore SVs (Fig. 2f). Additionally, 66,655
180 (55.28%) SVs identified in 2-9 ILs were classified as dispensable SVs, and 51,413
181 (42.64%) identified in only one IL were considered private SVs. Different from the
182 trends of core, softcore, and private SVs, the numbers of dispensable SVs increased
183 from *indica* to *japonica* (Fig. 2g), suggesting the roles of these dispensable SVs played
184 in the subspeciation of rice. The lengths of different SV types varied (Fig. 2h), with
185 three large absence variants (AVs) (0.73-1.16 Mb) detected (Fig. 2i). Since repetitive
186 sequences accounted for more than 50% of the genomes, we analyzed the distributions
187 of SVs and found that these SVs were predominantly detected in DNA transposons and
188 LTRs (Fig. 2j), suggesting regulatory functions of SVs in non-coding regions of the
189 genomes.

190 **Structural variants and heterosis of 17 agronomic traits**

191 We performed paired genomic comparisons and identified both unique and common
192 genomic variants for each pair of ILs, representing different and shared variants,
193 respectively (Fig. 3a, Supplementary Fig. S11 and Supplementary Table S5). On
194 average, unique genomic variants comprised 2,018,539 SNPs, 391,298 InDels, and
195 40,082 SVs (17,690 AVs, 11,975 PVs, 6,327 CNVs, 3,941 translocations, and 149
196 inversions). These unique variants were highly interconnected (Fig. 3b), with
197 approximately half located in intergenic regions (57.45% of SNPs, 50.42% of InDels,
198 and 50.56% of SVs), emphasizing the significance of these regions in the differentiation

199 of ILs. In contrast to other SV types, the majority (77.94%) of unique inversions
200 occurred in exonic regions, suggesting a link between inversions and changes in gene
201 expression levels. Furthermore, genetic relatedness of the 12 ILs assessed using unique
202 SVs, particularly translocations, showed slight variations compared to analyses based
203 on unique SNPs or InDels, demonstrating the distinct functional contributions of
204 different variant types to genomic divergence (Fig. 3c and Supplementary Fig. S12).

205 Correlation analysis between the numbers of genomic variants between the
206 parental lines and heterosis of the 17 traits revealed divergent patterns for unique versus
207 common variants (Fig. 3d). Importantly, the unique variants exhibited stronger
208 associations with heterosis than common variants (Fig. 3e). Meanwhile, no notable
209 change in correlation patterns was observed between stop-gain and stop-loss variants
210 (in SNPs and InDels), synonymous and nonsynonymous SNPs, or frameshift and
211 nonframeshift InDels. These findings emphasized the importance of genomic variations
212 between the two parents in determining heterosis.

213 Though unique SVs accounted for only 1.99% and 10.24% of the total SNPs and
214 InDels, respectively, their correlations with heterosis were comparable. Among the five
215 types of SVs, inversions—though fewer in number than translocations—were larger in
216 size and showed stronger associations with heterosis, suggesting a size-dependent
217 influence. Different from correlation patterns of heterosis for yield and other yield-
218 related traits, the number of unique variants showed negative correlations with heterosis
219 for seed setting rate. Notably, we observed that an increase in unique SVs corresponded
220 with a slight rise in heterosis for seed setting rate of intra-subspecific F_1 hybrids, but a

221 dramatic decrease in *indica-japonica* F_1 hybrids (Fig. 3f and Supplementary Fig. S13).
222 The inter-subspecific hybrids possessed significantly more unique SVs in their parents
223 than their intra-subspecific counterparts (Fig. 3g and Supplementary Fig. S13),
224 emphasizing the roles of these SVs in heterosis of inter-subspecies. The unique SVs
225 between the parental lines also exhibited stronger correlations with heterosis for tiller
226 number in inter-subspecific F_1 hybrids. However, the numbers of unique SVs displayed
227 weak or no correlations with yield heterosis in inter-subspecific F_1 hybrids, likely due
228 to reproductive isolation-driven reductions in seed setting rate.

229 **Structural variants of *S5-ORF5* contribute to heterosis**

230 Given the established role of the *S5* loci in embryo-sac fertility of inter-subspecific F_1
231 hybrids[24, 25], we performed both phylogenetic and genotyping analyses for sequences
232 of three genes (*S5-ORF3*, *S5-ORF4*, and *S5-ORF5*) in the 12 ILs (Fig. 4a).
233 Unexpectedly, two ILs—YB and Qianlijing—clustered together and were identified as
234 wide compatibility varieties (WCVs). With gene sequence of Balilla as the reference,
235 we detected eight SNPs, three deletions (1, 10, and 42 bp), and one AV (94 bp) in *S5-*
236 *ORF5* (Fig. 4a), which acts as the "killer" gene. The 94 bp AV was exclusive to the two
237 WCVs, potentially rescuing embryo-sac viability in their subspecific F_1 hybrids.

238 To investigate contributions of the 94 bp AV to heterosis, we first compared seed
239 setting rate between intra- and inter-subspecific F_1 hybrids lacking the WCV parentage.
240 The seed setting rate of *indica-indica* and *japonica-japonica* F_1 hybrids were similar,
241 both significantly higher than those of *indica-japonica* F_1 hybrids (Fig. 4b). The
242 heterosis for seed setting rate was also lower in inter-subspecific F_1 hybrids (Fig. 4c).

243 However, when YB or Qianlijing was used as a parent, the *indica-japonica* F_1 hybrids
244 achieved seed setting rate as high as the *indica-indica* F_1 hybrids (Fig. 4d-g).
245 Importantly, heterosis for seed setting rate were comparable between the inter- and
246 intra-subspecific F_1 hybrids, confirming the roles of AV in heterosis.

247 **Effects of structural variants on gene expression changes**

248 To further explore the impacts of SVs on gene expression changes, we performed RNA-
249 sequencing and differential expression analyses on the 12 ILs (Fig. 5a and
250 Supplementary Fig. S14). Drawing from previously reported inheritance patterns of
251 metabolite levels in rice F_1 hybrids[22], we demonstrated that transcriptomic additive and
252 partially dominant effects played significant roles in heterosis across the 17 traits,
253 reinforcing contributions of the two main effects to heterosis across different genetic
254 levels (Fig. 5b and Supplementary Fig. S14).

255 We next focused on genes within the two large inversions: inversion_6 (304 genes)
256 and inversion_8 (331 genes) (Fig. 5c). Consistent with the presence pattern of
257 inversion_6, differentially expressed genes (DEGs) were identified and these genes
258 accurately classified the 12 ILs into two groups (Fig. 5d-e and Supplementary Fig. S15).
259 Unexpectedly, genes in inversion_6 with additive/partially dominant effects were
260 predictive of heterosis (Supplementary Fig. S15). Based on genes in inversion_8, JR2
261 was found to be closer to *indica* than to *japonica* (Supplementary Fig. S15). Though
262 only a limited number of genes differed between the *indica* and *japonica*, the genes in
263 inversion_8 well explained the variance in heterosis (Supplementary Fig. S15). Among
264 the three large AVs, the genes in AV_8 (22 genes) and AV_11 (22 genes) were seldomly

265 expressed in seedling samples of the 12 ILs. For AV_6 (34 genes), most genes were
266 expressed and effectively divided these ILs into two groups (Fig. 5f-g). The genes in
267 AV_6 also demonstrated comparable predictability for heterosis with those of all DEGs
268 of 12 ILs and the two inversions (Fig. 5h). These results confirmed the direct impact of
269 SVs on changes in gene expression levels, which in turn explained the variance in
270 heterosis.

271 **Discussion**

272 In this study, we assembled 12 high-quality rice genomes and demonstrated the
273 contributions of genomic variants to rice sub-speciation. Trough comparative genomic
274 analysis, we found that the numbers of genomic variants between the parental lines
275 significantly correlate with heterosis across 17 agronomic traits. A detailed examination
276 of the *S5* loci validates the functional roles of SVs in heterosis for seed setting rate are.
277 We propose that the SVs distinguishing *indica* from *japonica* might be precisely
278 engineered to harness the strong inter-subspecific heterosis. Furthermore, genomic SVs,
279 especially the inversions, drive changes in gene expression levels and the differentially
280 expressed genes in parental lines well explain the variance in heterosis.

281 Though the precise modulation of genomic SVs remains challenging in crops, we
282 believe that the advances in genome editing technologies, such as DualPE[26], will
283 provide concrete evidence for the function roles of specific SVs in crop heterosis.
284 Before the establishment of precise editing methods for SVs in crops, including maize
285 and rice, a population genetic strategy could validate the effects of specific SVs on
286 heterosis of agronomic traits. Firstly, F_1 hybrids carrying targeted SVs in their parents

287 can be used to generate recombinant inbred lines. These recombinant inbred lines—
288 typically requiring at least five generations of self-crossing to achieve phenotypic
289 uniformity—can then be crossed with a male-sterile line to produce F_1 hybrids for
290 heterosis quantification. Genotyping of the recombinant inbred lines and the sterile line
291 can be implemented using standard PCR assays targeting the SVs. Finally, by
292 integrating genotypes of test-cross F_1 hybrid with heterosis measurements, the
293 functional impact of SVs can be assessed.

294 The establishment of causal links between specific SVs and heterosis is another
295 critical issue in the functional analyses of SVs. Gene expression data collected from
296 tissues like young panicles of the above prepared recombinant inbred lines can be
297 adopted to perform correlation analysis or expression GWAS with the assistance of
298 genome sequencing results. Genes associated with heterosis can be identified and SV-
299 associated genes may be further filtered through combining comparative genomic result
300 of the two parents. Since SVs often encompass multiple genes, a single SV may drive
301 gene expression changes across different pathways, thereby regulating heterosis of
302 multiple traits. Thus, it is important to record comprehensive phenotypic data in
303 heterosis analysis.

304 Based on these findings of SVs, we propose that the over-dominance model may
305 serve as the genetic basis for rice heterosis. The PAVs and CNVs between the parents
306 can be regarded as extreme forms of heterozygosity. The translocations can be
307 understood as heterozygosity of location. And the inversions are heterozygosity of
308 direction.

309 # Methods

310 ## Plant materials

311 Twelve rice inbred lines (ILs) were selected from accessions reported in previous
312 studies[16, 22, 27, 28]. Based on clustering results from metabolite profiles generated using
313 GC- or LC-MS platforms[27, 28], low-density SSR markers[16], and high-density SNPs or
314 InDels[22], these 12 inbred lines are representatives of *indica* and *japonica* accessions.
315 Plants of the selected inbred lines for genome sequencing were grown under long-day
316 conditions, maintaining a 16-hour light/8-hour dark photoperiod at 30 °C in a plant
317 culture room. The seedlings were sown in 96-well PCR plates (without bottoms) and
318 placed in a plastic basin filled with Yoshida nutrient solution (pH = 5.5). After 20 days
319 of growth, leaves from each genotype were harvested for experimentation. For full-
320 length RNA sequencing of YB and Balilla, the plants were divided into four tissue parts:
321 root, middle leaf (the youngest leaf), leaf without middle leaf, and stem.

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

330 The 12 inbred lines were crossed using a complete diallel design. As previously

331 detailed[22], a total of 17 agronomic traits were recorded for both the inbred lines and
332 their corresponding F_1 hybrids. These traits included grain yield per plant (YPP), grain
333 number per panicle (GNP), seed setting rate (SSR), thousand-grain weight (TGW),
334 plant height/tiller number (at seedling, elongation, flowering and maturation stages)
335 (SSPH, ESPH, FSPH, MSPH and SSTN, ESTN, FSTN, MSTN), heading date (HD),
336 tiller number per plant (TPP), panicle length (PL), and the number of primary and
337 secondary branches per panicle (PBN and SBN). Better-parent heterosis was calculated
338 using the equation: $BPH = (F_1 - P_{high})/P_{high}$, where P_{high} represents the value of the high
339 parent. For the heading date, parents with shorter heading dates were designated as P_{high},
340 and the calculated values were subsequently multiplied by -1.

341 **Genomic sequencing**

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

350 Hi-C libraries for YB and Balilla were prepared and sequenced using the BGI
351 PE150 platform, generating a total of 51.81 Gb and 49.27 Gb of reads (approximately
352 127.80× coverage of the genome) for the two samples, respectively. The raw data were

filtered using Trimmomatic (v0.40)[29] with parameters LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:15. The quality of the clean data was evaluated with default parameters using FastQC[30].

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)[31], 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

375 of raw data per sample. The raw data were filtered using SOAPnuke (v2.1.0, -
376 lowQual=20, -nRate=0.005, -qualRate=0.5)[32]. Clean reads were aligned to the YB
377 genome using Hisat2 (v2.2.1, --no-unal)[33]. Transcripts were assembled using
378 StringTie[34], and clean reads were mapped to reference transcripts using bowtie2 (v2.3.5,
379 --sensitive -k 50 --no-mixed --no-discordant)[35]. The read count of each transcript was
380 calculated using RSEM (v1.3.3)[36].

381 **Genome assembly**

382 Draft genomes of the 12 rice inbred lines were assembled using HiFi reads with
383 parameter -l3 via HiFiasm (v0.19.5)[37]. The clean Hi-C reads for YB and Balilla were
384 aligned to contigs using Juicer (v1.6)[38]. Contact matrices were generated with 3D-DNA
385 (v180922, -r0)[39], and corrections were applied using JuiceBox (v1.11.08). The genomes
386 of the remained ten ILs were assembled based on the genomes of YB and Balilla using
387 Ragtag (v2.1.0, scaffold)[40]. We assessed the accuracy, identity, and completeness of the
388 12 assemblies. The completeness of the assemblies was assessed using Benchmarking
389 Universal Single-Copy Orthologue (v3.0.2, embryophyta_odb10) analysis[41]. The long
390 terminal repeats (LTR) assembly index (LAI) was calculated for each chromosome
391 using LTR tools[42], including LTR_retriever (v2.9.0, -infinder ltrfinder.scn -inharvest
392 ltrharvest.scn), LTR_FINDER (v1.0.7, -D 15000 -L 7000 -C -M 0.7), and GenomeTools
393 (v1.6.2, ltrharvest -similar 70 -vic 10 -seed 20 -seqids yes -maxlenltr 7000 -maxtsd 6).
394 HiFi reads were re-mapped to the assemblies with parameters "-ax map-hifi" using
395 Minimap2 (v2.21)[43]. Additionally, consensus quality values and completeness of the
396 assemblies were scored using a k-mer-based method with the Merqury pipeline (v1.3)[44].

397 **Genome annotation**

398 Repetitive sequences (including tandem repeats and transposable elements), protein-
399 coding genes, and non-coding RNAs were annotated for the genome seeemblies of YB
400 and Balilla. Tandem repeats were identified using the Tandem Repeats Finder (TRF,
401 v4.09.1, 2 7 7 80 10 50 2000 -d -h)[45]. Transposable elements were annotated through
402 an integrated approach combining do novo and homology-based (Repbase,
403 www.girinst.org/repbase[46]) strategies. The de novo repeat library was constructed using
404 LTR_FINDER (v1.0.7, -w 2 -C)[47] and RepeatModeler (v2.0.1)[48]. Transposable
405 elements were detected using RepeatMasker (v4.1.2)[49] with parameters "-nolow -no_is
406 -norna -parallel 2", searching against both the Repbase[46] and de novo repeat library.
407 Additionally, RepeatProteinMask (1.36)[49] with parameters "-engine ncbi -
408 noLowSimple -pvalue 0.0001" was utilized to identify transposable elements in
409 Repbase[46].

410 To annotate gene structure and function, we integrated three strategies: homology-
411 based (using five genomes), de novo, and transcriptome-based (full-length RNA
412 sequencing) predictions. In the homology-based approach, protein-coding sequences
413 from the genomes of 9311[50], Minghui63[51], Zhenshan97[51], R498[52], and Nipponbare[19]
414 were compared to the YB and Balilla genomes using tblastn (v2.11.0+, -evalue 1e-05)[53].
415 Gene structures of the aligned sequences were predicted using Exonerate (v2.4.0, --
416 model protein2genome)[54]. De novo annotation was conducted using Augustus (v3.4.0,
417 --noInFrameStop=true --strand=both)[55], GlimmerHMM (v3.0.4)[56], and GeneMark
418 (v4.65)[57]. The annotated gene sets were integrated to non-redundant sets using MAKER

419 (v3.01.03)[58], and further polished by incorporating full-length RNA sequencing results
420 using PASA (v2.4.1)[59]. RNA-sequencing results were aligned using HiSat2 (v.2.2.1)[33],
421 and transcripts were assembled with StringTie (v.2.1.7)[60]. Additionally, transcripts were
422 de novo assembled using Trinity (v2.8.5)[61]. Gene functional analysis was conducted
423 with Diamond (v2.0.7, -c 1 -f 5 -e 1e-5 --very-sensitive -k 10)[62] against five databases:
424 Non-Redundant (NR, ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), Kyoto
425 Encyclopedia of Genes and Genomes (KEGG)[63], Gene Ontology (GO)[64], TrEMBL[65],
426 and Swiss-Prot[65]. Protein domains were annotated using InterProScan (v5.50-84.0, --
427 applications Pfam)[66] based on the InterPro database[67].

428 Non-coding RNAs, including rRNA, tRNA, snRNA, and miRNA, were also
429 annotated for the two genomes. tRNA was identified using tRNAscan-SE (v2.0.9)[68],
430 while rRNA was detected using RNAmmer (v1.2, -S euk -multi -m lsu,ssu,tsu)[69].
431 miRNA and snRNA were annotated with Infernal (v1.1.4)[70] using Rfam (v14.6)[71].
432 miRNA and rRNA were present in higher percentages within the genomes. The
433 completeness of the genomic annotations was further assessed using BUSCO (v5.2.2,
434 embryophyta_odb10)[41]. Here, 98.5% and 98.6% of the gene sets in the two genomes
435 contained complete BUSCOs in embryophyta.

436 For the annotation of the remaining ten genomes, Liftoff [72] was employed. The YB
437 genome was chosen as the reference genome for five *indica* accessions: R465,
438 Mianhui725, 9311K, 610234, and Qianlijing. The Balilla genome served as the
439 reference for five *japonica* accessions: R4115, C418, 110080, Wuyunjing8, and JR2.

440 **Pan-genome analysis**

441 Protein coding genes of the 12 inbred lines were initially aligned using BLASTp (-
442 evalue 1e-5)[73]. The BLAST results were processed with OrthoFinder (v2.3.12)[74] to
443 classify gene families. Gene families shared among all 12 inbred lines were defined as
444 core gene families, while those identified in more than 90% of inbred lines were termed
445 softcore gene families. Gene families specific to a single inbred line were classified as
446 private gene families, and the remaining gene families were considered dispensable.
447 Genes in the four types of gene families were further analyzed in the Pfam[75], GO[64], and
448 KEGG[63] databases to determine their conservation.

449 **Identification of genomic variants**

450 Genomic variants, including single nucleotide polymorphisms (SNPs), insertions and
451 deletions (InDels) (2-50 bp), and structural variants (SVs), were detected in the 12 rice
452 inbred lines, with the genome of YB selected as the reference genome and the remained
453 11 as query genomes. SNPs and InDels were detected using MUMmer4 (Nucmer: -c
454 40 -l 90, Delta-filter: -1, Show-coords: -THrd, Show-SNPs: -ClrTH)[76]. SVs, including
455 presence-absence variants (PAVs), inversions, translocations, and copy number variants
456 (CNVs), were detected with the assistance of SyRI (v1.4)[77]. The distribution of SNPs,
457 InDels, and SVs was analyzed across various genomic regions or functions, including:
458 1. Upstream (1 kb regions before the gene start codon); 2. Downstream (1 kb regions
459 after the gene stop codon); 3. Intergenic regions; 4. UTR3 (3' untranslated regions); 5.
460 UTR5 (5' untranslated regions); 6. Splicing junction regions (< 10 bp); 7. Intronic
461 regions; 8. Exonic regions; 9. Synonymous or nonsynonymous mutants; 10. Frameshift
462 or nonframeshift insertions/deletions; 11. Stopgain (gain of stop codons); 12. Stoploss

463 (loss of stop codons). Variants simultaneously located in two regions, such as UTR5
464 and UTR3, were showed as Region 1;Region 2. The functions of SNPs, InDels, and
465 SVs was annotated with ANNOVAR[78].

466 **Pan-SV analysis**

467 A pan-SV analysis was conducted for SVs of the 11 query genomes and the results were
468 further used for constructing graph-based genome using Vg (v1.43)[79]. For SVs
469 excluding PVs, an SV was assigned when the ratio of overlapping to total locations in
470 two inbred lines exceeded 50%. For PVs, a PV was designated when the distance
471 between PV locations in two inbred lines was less than 5 bp and the length ratio between
472 the shorter and longer samples was greater than 50%. Circles illustrating genomic
473 variants on chromosomes were generated using Circos[80]. Genetic relatedness among
474 the 12 inbred lines was visualized using Cytoscape[81].

475 **Construction of phylogenetic trees**

476 A phylogenetic tree for 223 rice inbred lines was constructed using genotypic data based
477 on 33 pairs of single sequence repeat markers from a previous study[22]. Phylogenetic
478 trees based on the number of differential SNPs, InDels, and CNVs among the 12 rice
479 inbred lines were built using the neighbour-joining method with MEGA (v10.2.6)[82].
480 The number of SNPs, InDels, and CNVs of the inbred lines were retrieved from a
481 previous report[22].

482 For phylogenetic analysis of *Phr1* and the *S5* loci, gene sequences from reference
483 inbred lines were obtained from the GenBank database. The *Phr1* sequences of
484 Nipponbare and 9311 corresponded to accession numbers DQ532376 and DQ532391,

485 respectively[23]. For *S5-ORF3-5*[24, 25], the Balilla sequences were retrieved under
486 accession numbers JX138498, JX138502, EU889294. Gene sequences from the 12
487 newly assembled inbred lines were extracted and aligned using Jalview (version
488 2.11.4.1) with default clustal parameters[83]. Phylogenetic trees were constructed with
489 the neighbour-joining method. Genotypes of inbred lines were assigned based on
490 previous reports[23-25]. Gene structures were visualized using the exon-intron graphic
491 maker (http://wormweb.org/exonintron).

492 **Phenol reaction experiment**

493 Seeds from 14 inbred lines were harvested at maturity and stored at -20 ºC. For each
494 inbred line, three seeds were chosen and treated with a 2% (volume/volume) phenol
495 solution at 37 ºC. Changes in hull colour of the seeds were recorded after 1-, 2-, 3-, 5-,
496 and 7 days of treatment.

497 **Normalization and differential analysis of transcriptomics data**

498 Read counts of 53,564 genes in each experimental sample were used for data
499 normalization and differential analyses through MetaboAnalyst 5.0[84]. The interquantile
500 range was selected to filter genes, with a threshold set at 40%. Sum, none, and auto-
501 scaling were chosen for sample normalization, data transformation, and data scaling,
502 respectively. Differential analysis among the 12 inbred lines was implemented using
503 the ANOVA module in MetaboAnalyst 5.0[84], with a raw *P*-value cutoff set at 0.001 for
504 post-hoc analysis. Principal component analysis (PCA) and heatmap results for the 12
505 inbred lines were obtained through corresponding modules on the website.
506 PERMANOVA was conducted in PCA and the permutation number was 999. Euclidean

507 for distance measure and ward for clustering method were set in heatmaps. Differential
508 genes were selected from the original read count table and averaged across the three
509 biological replicates. The averaged data were further filtered and normalized for each
510 genotype with the above parameters. Due to the limited number of genes in inversions
511 and absence variants, no filtering was applied to these genes. The genes in three large
512 AVs, including AV_6, AV_8, and AV_11, were analyzed. Despite AV_8 and AV_11
513 being larger in size, the genes in these AVs were rarely expressed in the seedling
514 samples. Only genes in AV_6 was further analyzed.

515 **Partial least square (PLS) analysis**

516 PLS analysis was conducted to investigate the relationship between various inheritance
517 patterns at the transcriptomic level and better-parent heterosis of the 17 agronomic traits
518 using SPSS 20.0. The maximum number of latent factors was set to 50. The inheritance
519 patterns were calculated based on normalized gene expression levels using equations
520 outlines in a previously study[22]. The genetic effects included: 1. Additive effect: average
521 levels of the female (P_f) and male (P_m) parents; 2. Partially dominant effects (PD):
522 partially dominant _0.3 (PD_0.3) = $P_m+0.3\times(P_f-P_m)$ and partially dominant _0.7
523 (PD_0.7) = $P_m+0.7\times(P_f-P_m)$; 3. Dominant effects: transcript levels of the female
524 (dominant_female) and male (dominant_ male) parents.

525 **Statistical analysis**

526 Boxplots were created using BoxPlotR[85]. Analysis of variance (ANOVA) with the least
527 significant difference for post hoc comparisons was performed using SPSS 20.0. This
528 analysis compared the following parameters: lengths of genes, lengths of coding

529 sequences, lengths of exons, lengths of introns, GC content of genes, GC content of
530 coding sequences, number of exons, and number of introns across seven genomes
531 (including two newly assembled genomes, YB and Balilla, along with five
532 representative rice genomes). Additionally, we assessed the number of genomic variants
533 and the degrees of better-parent heterosis for 17 traits in F_1 hybrids from both intra- and
534 inter-subspecies. The adjusted R^2 values of partial least square models for heterosis
535 across traits were also compared. An independent samples *t*-test (weighted) was
536 conducted to evaluate the number of correlations between unique and common genomic
537 variants and heterosis. SPSS was also employed to conduct Spearman correlation to
538 investigate the relationships between different types of unique genomic variants and
539 between genomic variants and heterosis.

540 Data and code availability

541 Raw reads of PacBio SMRT sequencing and RNA-sequencing of the 12 rice inbred
542 lines have been deposited in the NCBI database under BioProject accession
543 PRJNA1178436. Genome assemblies of the 12 rice inbred lines are available from
544 Figshare repository with DOI: 10.6084/m9.figshare.28397960. Source data are
545 available with this paper.

546 References

547 1. Shull, G.H. What is "heterosis"? *Genetics* **33**, 439-446 (1948).
548 2. Hochholdinger, F. & Yu, P. Molecular concepts to explain heterosis in crops. *Trends Plant Sci.*,
549 10.1016/j.tplants.2024.1007.1018 (2024).
550 3. Bruce, A.B. The Mendelian theory of heredity and the augmentation of vigor. *Science* **32**, 627-628
551 (1910).
552 4. Jones, D.F. Dominance of linked factors as a means of accounting for heterosis. *Proc Natl Acad Sci*
553 *U S A* **3**, 310-312 (1917).
554 5. 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).
6. 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).
7. 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).
8. Semel, Y. et al. Overdominant quantitative trait loci for yield and fitness in tomato. *Proc. Natl. Acad. Sci. USA* **103**, 12981-12986 (2006).
9. East, E.M. Heterosis. *Genetics* **21**, 375-397 (1936).
10. 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).
11. Powers, L. An expansion of Jones's theory for the explanation of heterosis. *Am. Nat.* **78**, 275-280 (1944).
12. He, Q. et al. Hybrid rice. *Engineering* **6**, 967-973 (2020).
13. Huang, X. et al. Genomic architecture of heterosis for yield traits in rice. *Nature* **537**, 629-633 (2016).
14. Lv, Q. et al. Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. *Nat. Commun.* **11**, 4778 (2020).
15. Zhou, G. et al. Genetic composition of yield heterosis in an elite rice hybrid. *Proc Natl Acad Sci U S A* **109**, 15847-15852 (2012).
16. 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).
17. Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542-3558 e3516 (2021).
18. Shang, L. et al. A super pan-genomic landscape of rice. *Cell Res.* **32**, 878-896 (2022).
19. Shang, L. et al. A complete assembly of the rice Nipponbare reference genome. *Mol. Plant* **16**, 1232-1236 (2023).
20. Sun, Z. et al. Dissecting the genetic basis of heterosis in elite super-hybrid rice. *Plant Physiol.* **192**, 307-325 (2023).
21. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145-161 (2020).
22. Dan, Z. et al. Additive and partially dominant effects from genomic variation contribute to rice heterosis. *bioRxiv*, 10.1101/2024.1107.1116.603817 (2024).
23. Yu, Y. et al. Independent losses of function in a polyphenol oxidase in rice: differentiation in grain discoloration between subspecies and the role of positive selection under domestication. *Plant Cell* **20**, 2946-2959 (2008).
24. Chen, J. et al. A triallelic system of S5 is a major regulator of the reproductive barrier and compatibility of indica-japonica hybrids in rice. *Proc Natl Acad Sci U S A* **105**, 11436-11441 (2008).
25. Yang, J. et al. A killer-protector system regulates both hybrid sterility and segregation distortion in rice. *Science* **337**, 1336-1340 (2012).
26. Zhao, Y. et al. Precise deletion, replacement and inversion of large DNA fragments in plants using dual prime editing. *Nat. Plants* **11**, 191-205 (2025).
27. Dan, Z. et al. Metabolic prediction of important agronomic traits in hybrid rice (*Oryza sativa* L.). *Sci. Rep.* **6**, 21732 (2016).

28. 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).
29. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
30. Wingett, S.W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* **7**, 1338 (2018).
31. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005).
32. 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).
33. 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).
34. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290-295 (2015).
35. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 17 (2010).
36. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
37. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332-1335 (2022).
38. Durand, N.C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95-98 (2016).
39. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).
40. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
41. 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**, 3210-3212 (2015).
42. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
43. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
44. 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).
45. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).
46. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
47. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265-268 (2007).
48. 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).
49. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic

643 sequences. *Curr Protoc Bioinformatics* **Chapter 4**, 4 10 11-14 10 14 (2009).
644 50. Yu, J. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79-92
645 (2002).
646 51. Zhang, J. et al. Extensive sequence divergence between the reference genomes of two elite *indica*
647 rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. USA* **113**, E5163-E5171 (2016).
648 52. Du, H. et al. Sequencing and de novo assembly of a near complete indica rice genome. *Nat.*
649 *Commun.* **8**, 15324 (2017).
650 53. Gertz, E.M., Yu, Y.K., Agarwala, R., Schaffer, A.A. & Altschul, S.F. Composition-based statistics
651 and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 41
652 (2006).
653 54. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison.
654 *BMC Bioinformatics* **6**, 31 (2005).
655 55. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**,
656 W435-439 (2006).
657 56. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab
658 initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-2879 (2004).
659 57. Borodovsky, M. & Mcininch, J. GENMARK: Parallel gene recognition for both DNA strands.
660 *Comput. Chem.* **17**, 123-133 (1993).
661 58. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool
662 for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
663 59. Haas, B.J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment
664 assemblies. *Nucleic Acids Res.* **31**, 5654-5666 (2003).
665 60. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2.
666 *Genome Biol.* **20**, 278 (2019).
667 61. Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference
668 genome. *Nat. Biotechnol.* **29**, 644-652 (2011).
669 62. Buchfink, B., Reuter, K. & Drost, H.G. Sensitive protein alignments at tree-of-life scale using
670 DIAMOND. *Nat. Methods* **18**, 366-368 (2021).
671 63. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG.
672 *Nucleic Acids Res.* **42**, D199-205 (2014).
673 64. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29
674 (2000).
675 65. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in
676 2003. *Nucleic Acids Res.* **31**, 365-370 (2003).
677 66. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**,
678 1236-1240 (2014).
679 67. Mitchell, A. et al. The InterPro protein families database: the classification resource after 15 years.
680 *Nucleic Acids Res.* **43**, D213-221 (2015).
681 68. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes
682 in genomic sequence. *Nucleic Acids Res.* **25**, 955-964 (1997).
683 69. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic*
684 *Acids Res.* **35**, 3100-3108 (2007).
685 70. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*
686 **29**, 2933-2935 (2013).

71. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121-124 (2005).
72. Shumate, A. & Salzberg, S.L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639-1643 (2021).
73. 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).
74. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
75. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412-D419 (2021).
76. Marcais, G. et al. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
77. 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).
78. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
79. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875-879 (2018).
80. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-1645 (2009).
81. 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).
82. Stecher, G., Tamura, K. & Kumar, S. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol* **37**, 1237-1239 (2020).
83. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).
84. 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).
85. Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. BoxPlotR: a web tool for generation of box plots. *Nat. Methods* **11**, 121-122 (2014).

Acknowledgements

We thank C. Zhu for funding of our rice research. We thank J. Bao, L. Rong, and S. Zheng from the Frasergen Bioinformatics for their help in genomics and transcriptomics analyses. 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.

Additional information

Supplementary Fig. S1 Neighbour-joining tree for 223 rice accessions.

Supplementary Fig. S2 Heterosis of 16 agronomic traits for F_1 hybrids from intra- or inter-subspecies.

Supplementary Fig. S3 Contact matrices and collinearity diagrams for genomes of YB and Balilla.

Supplementary Fig. S4 Long terminal repeat assembly index (LAI) for 12 chromosomes of 12 rice accessions.

Supplementary Fig. S5 Comparison of two newly assembled genomes with five representative rice genomes.

Supplementary Fig. S6 Genomic comparisons between Mianhui725, 610234, and YB.

Supplementary Fig. S7 Genomic comparisons between 9311K, R465, and YB.

Supplementary Fig. S8 Genomic comparisons between Qianlijing, C418, and YB.

Supplementary Fig. S9 Genomic comparisons between R4115, Wuyunjing8, and YB.

Supplementary Fig. S10 Genomic comparisons between 110080, JR2, and YB.

Supplementary Fig. S11 Genomic variants displayed on the genome of YB.

Supplementary Fig. S12 Genetic relatedness of the 12 rice inbred lines based on genomic variants.

Supplementary Fig. S13 Number of genomic variants and correlation between structural variants and heterosis of intra- and inter-subspecific F_1 hybrids.

Supplementary Fig. S14 Transcriptomic additive and partially dominant effects contribute to heterosis of 17 traits.

Supplementary Fig. S15 Expression levels of genes in two inversions and three absence variants and their relationship with heterosis.

Supplementary Table S1 Genome assembly profile of 12 rice inbred lines.

Supplementary Table S2 Consensus quality values and completeness of the 12 assemblies.

Supplementary Table S3 Hi-C profile of two rice genomes.

Supplementary Table S4 Annotation of two rice genomes.

Supplementary Table S5 Genomic variants of each pair of inbred lines and heterosis of their F_1 hybrids.

Figure and figure legends

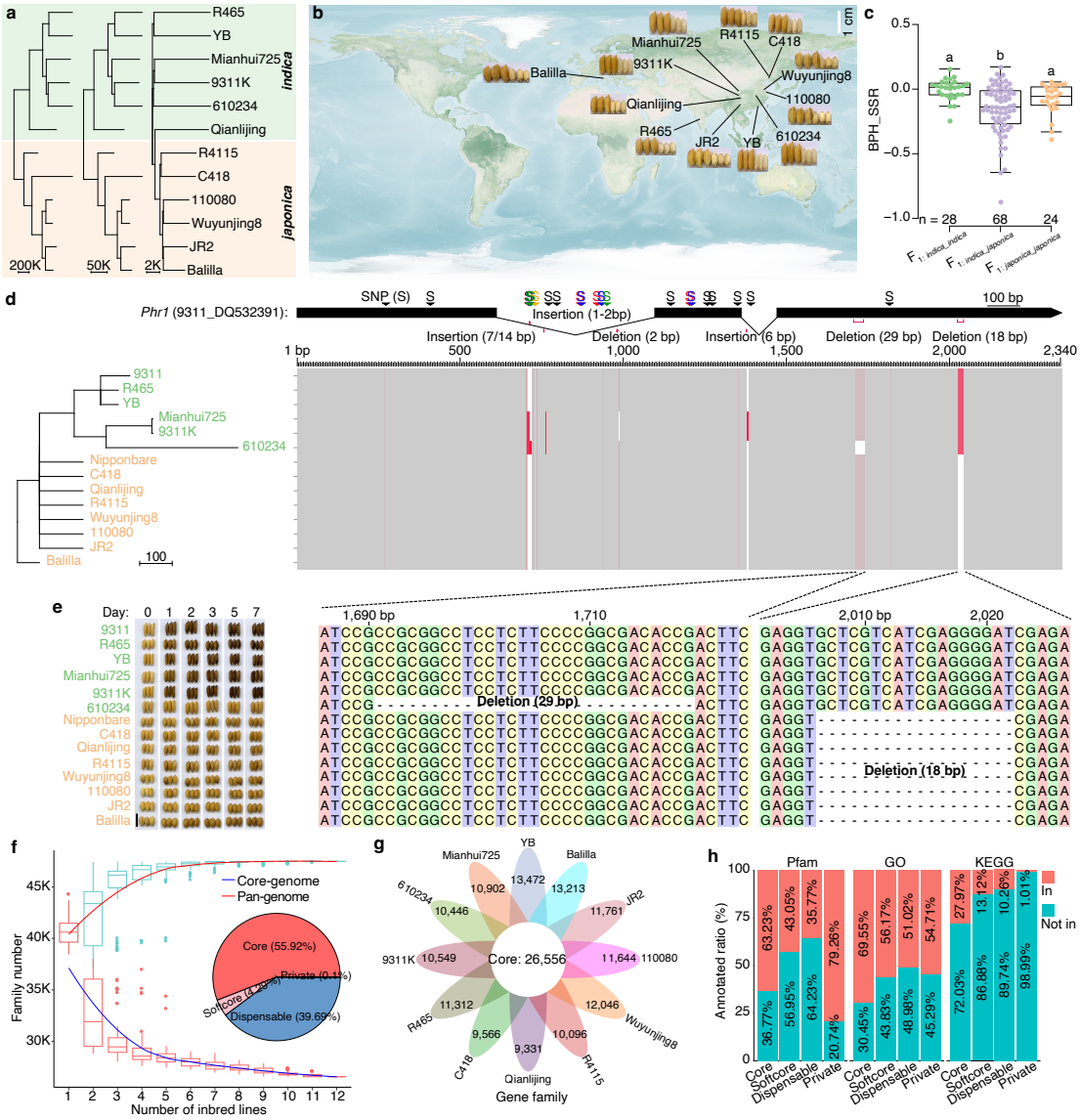

Fig. 1 Genome assemblies and pan-genome of 12 rice inbred lines. a, Neighbour-joining trees of 12 representative rice inbred line (ILs) based on genomic SNPs (left side), InDels (middle), and CNVs (right side) analyzed from previously reported Illumina sequencing results. **b,** Geographic origins of the 12 ILs. **c,** Better-parent heterosis for seed setting rate (BPH_SSR) of intra- and inter-subspecific F_1 hybrids. ANOVA with LSD in post-hoc test (P-value < 0.05) was performed. The numbers of F_1 hybrids are indicated. **d,** Gene structures and neighbour-joining tree of the 12 ILs based on gene sequences of *Phr1*. Nipponbare (*japonica*) and 9311 (*indica*) served as reference accessions. Red regions denote genomic variants, and dashed lines represent deletions. **e,** Phenol reaction phenotypes of 14 ILs. The bar represents 1 cm. **f,** Variation of gene family sizes in the pan- and core-genome. **g,** Core and non-core gene families across the 12 ILs. **h,** Gene annotation ratios cross three databases. Three databases, including Pfam, GO, and KEGG, were selected for annotation analysis.

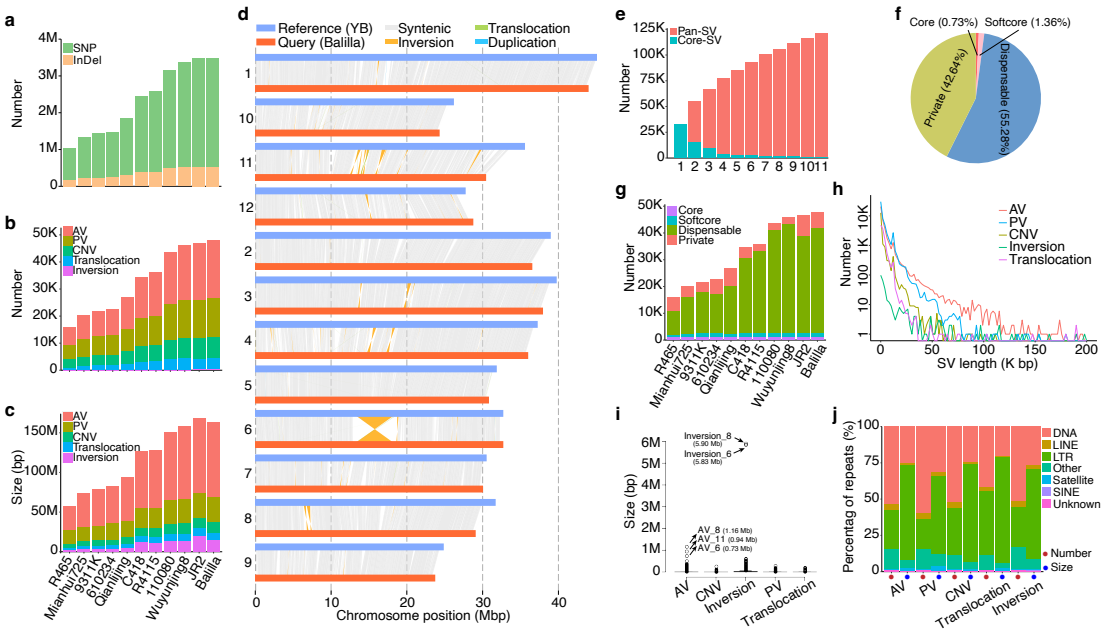

Fig. 2 Genomic variants of 12 ILs. a-c, Number and size distribution of genomic variants in 11 ILs with genome assembly of YB as the reference. **d,** Comparison of genomic profiles between YB and Balilla. **e,** Variation of SVs in the pan- and core-SVs with varied number of ILs. **f,** Percentages of core, softcore, private, and dispensable SVs. **g,** Number of core, softcore, private, and dispensable SVs in 11 ILs. **h-i,** Number and size of SVs. **j,** Percentages of different types of repeats in SVs. The number and size are analyzed.

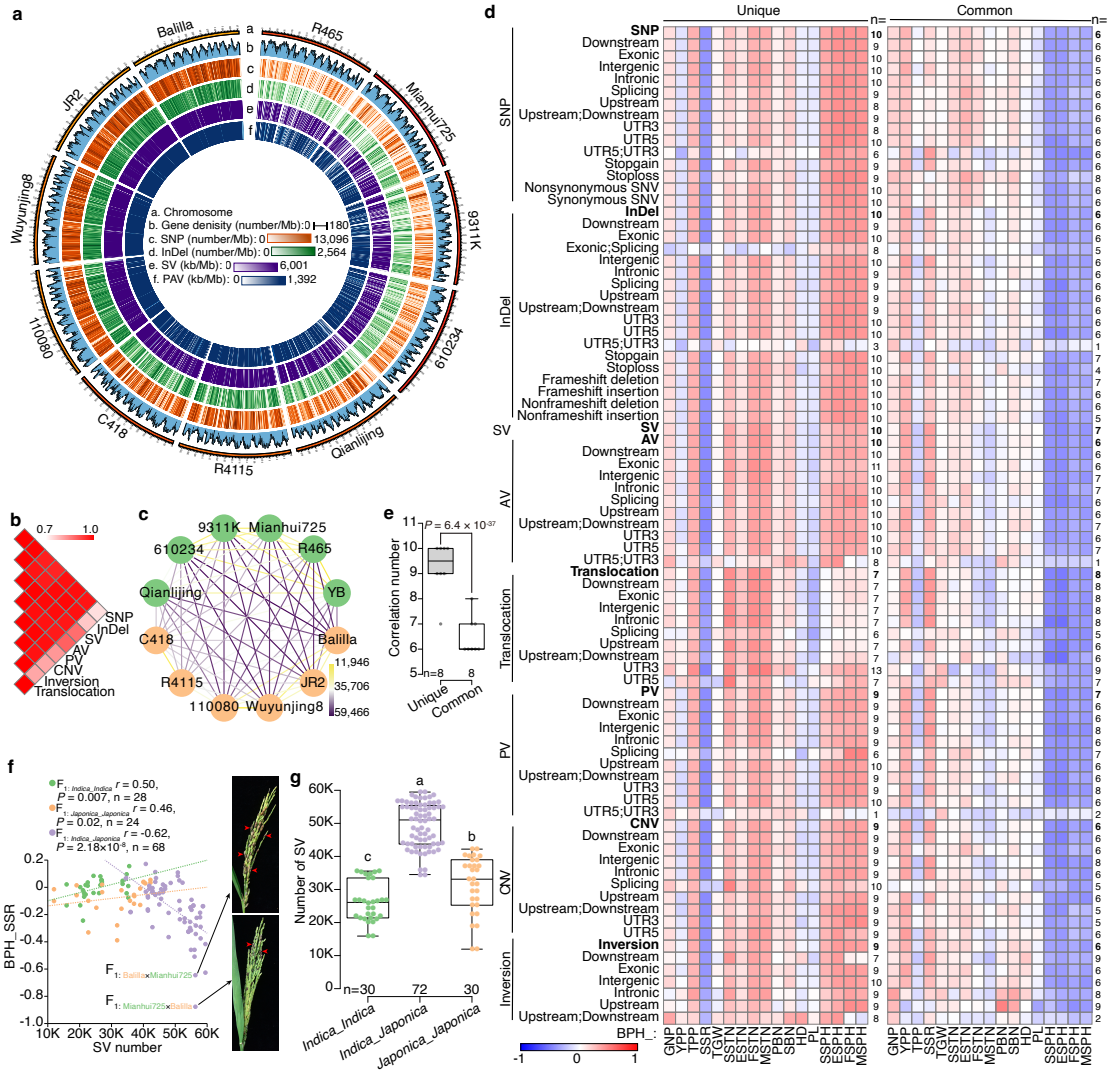

Fig. 3 Connections between genomic variants and heterosis of 17 agronomic traits. **a,** Genomic comparisons of YB against the other 11 ILs. **b,** Heatmap for the correlations between different types of unique genomic variants. **c,** Genetic relatedness of 12 ILs based on unique SVs. **d,** Correlation patterns between genomic variants and heterosis of 17 traits. The genomic variants of each type were divided into different parts according to their locations or functions. Spearman correlation analysis was performed with better-parent heterosis (BPH) of the following traits: grain number per panicle (GNP), tiller number per plant (TPP), seed setting rate (SSR), thousand-grain weight (TGW), seedling stage tiller number (SSTN), elongation stage tiller number (ESTN), flowering stage tiller number (FSTN), maturation stage tiller number (MSTN), yield per plant (YPP), seedling stage plant height (SSPH), elongation stage plant height (ESPH), flowering stage plant height (FSPH), maturation stage plant height (MSPH), panicle length (PL), primary branch number (PBN), secondary branch number (SBN), and heading date (HD). The number besides each row indicates the number of significant correlations (P-value < 0.05). **e,** Comparison of correlation frequencies between unique and common variants. P-value indicates independent t-test (weighted). **f,** Correlations between the number of unique SVs and heterosis for seed setting rate in 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,** The number of SVs in intra- and inter-subspecific F_1 hybrids. ANOVA with LSD in post-hoc test (*P*-value < 0.05) was performed.

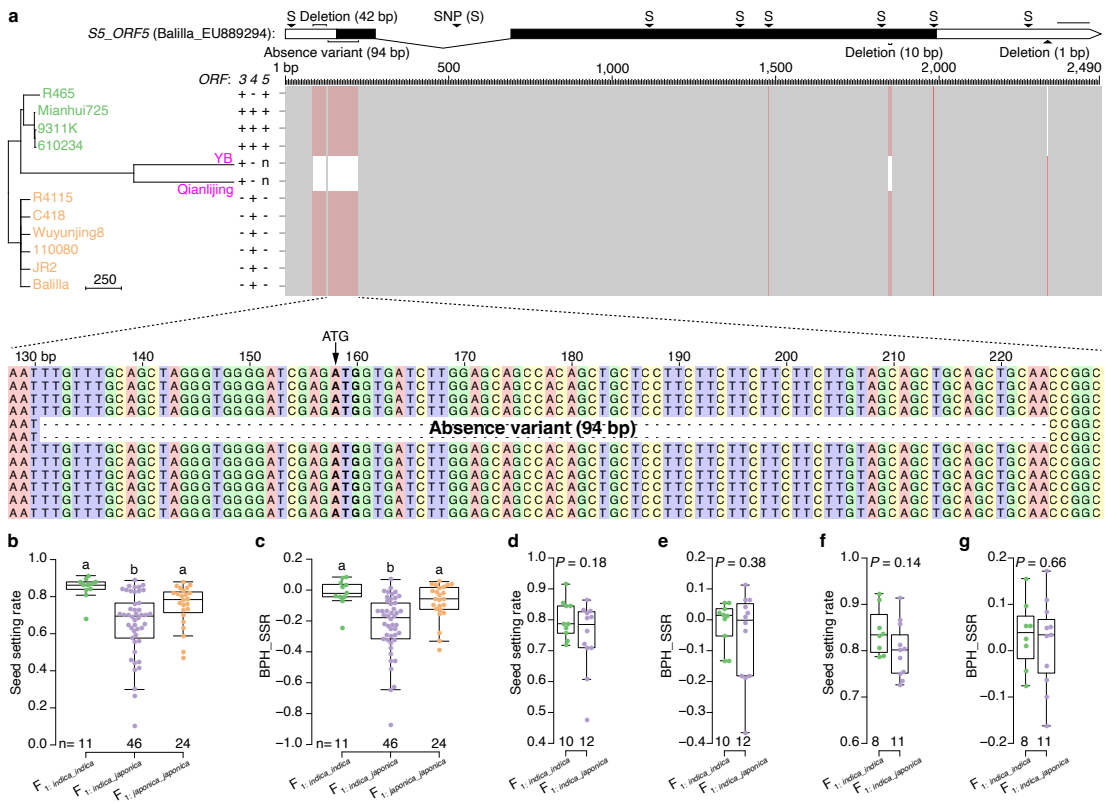

Fig. 4 Contributions of SVs in *S5* to heterosis for seed setting rate. a, Neighbour-joining tree based on *S5* and gene structure of *S5-ORF5* in 12 ILs. Genotypes of *S5-ORF3*, *S5-ORF4*, *S5-ORF5* for 12 ILs were assigned with previously reported rules. The "n" represents wide compatibility varieties. The sequence of Balilla serves as the reference and sequence variants are indicated in red. **b-c,** Comparisons of seed setting rate and heterosis for seed setting rate in intra- and inter-subspecific F_1 hybrids. Two wide compatibility varieties, namely YB and Qianlijing, were excluded for analysis. ANOVA with LSD in post-hoc test (P-value < 0.05) was performed. **d-e,** Comparisons of seed setting rate and heterosis for seed setting rate between intra- and inter-subspecific F_1 hybrids with YB as a parent. **f-g,** Comparisons of seed setting rate and heterosis for seed setting rate between intra- and inter-subspecific F_1 hybrids with Qianlijing as a parent. P values in **d-g** indicate independent sample *t*-test.

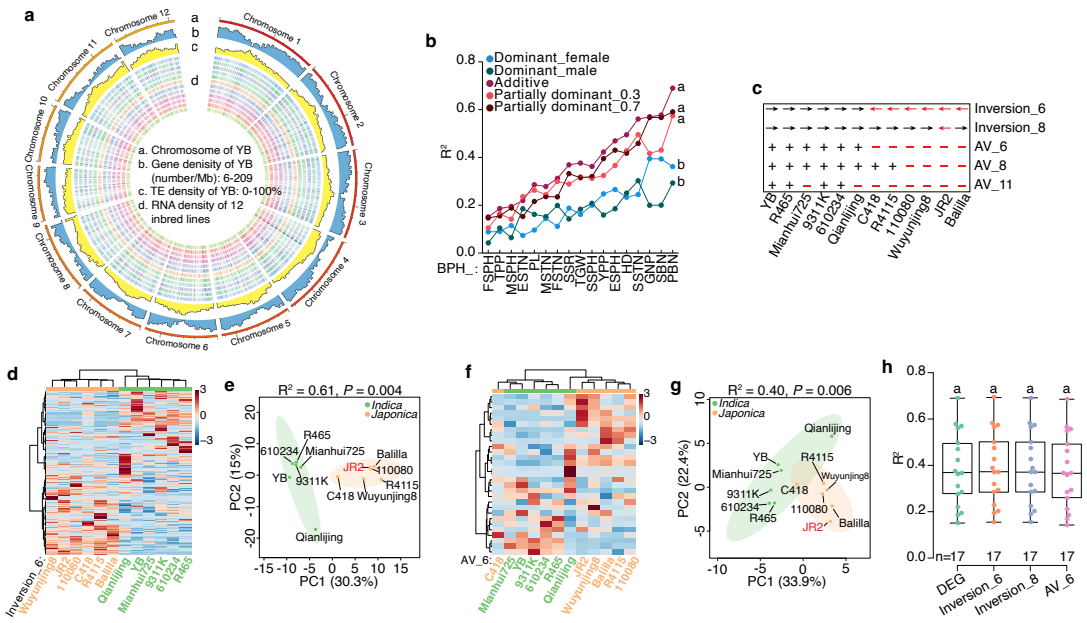

Fig. 5 Relationships among SVs, gene expression, and heterosis. a, Transcriptomic profiles of 12 ILs. Gene expression data were obtained from 15-day-old seedlings. **b,** Changes in R^2 (adjusted) for heterosis of 17 traits. Different inheritance patterns, including dominant (female and male), additive, and partially dominant (0.3 and 0.7) effects, of gene expression levels were used as predictive variables to perform partial least square regression analysis on better-parent heterosis of 17 agronomic traits. The best regression model (with highest adjusted R^2) was selected as the predictive model for heterosis of each trait. Analysis of variance was performed to compare the differences. **c,** Genomic structures of two inversions and three AVs in 12 ILs. Genotypes of YB is as the reference. **d-g,** Heatmaps and PCA plots of 12 ILs based on expression levels of genes within Inversion_6 and AV_6. *P*-values indicate PERMANOVA. **h,** Comparisons of R^2 (adjusted) for predicting heterosis of 17 traits with differentially expressed genes (DEGs) of 12 ILs, genes in Inversion_6, genes in Inversion_8, and genes in AV_6 as predictive variables. ANOVA with LSD in post-hoc test (P-value < 0.05) was performed in **b** and **h**.