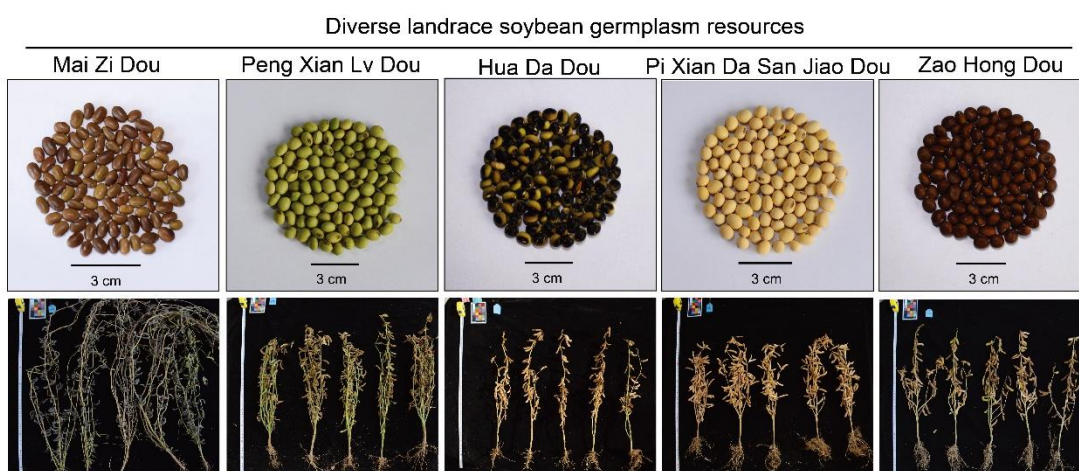
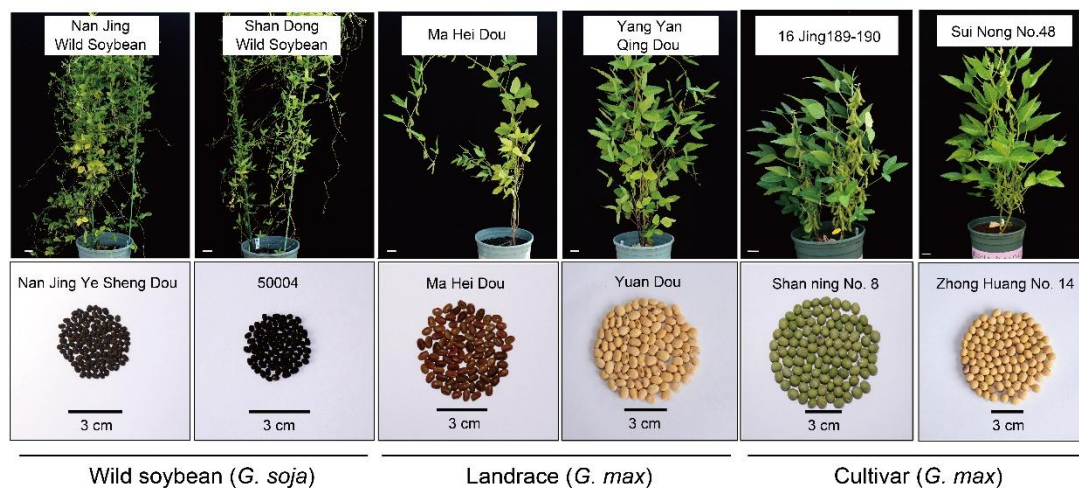
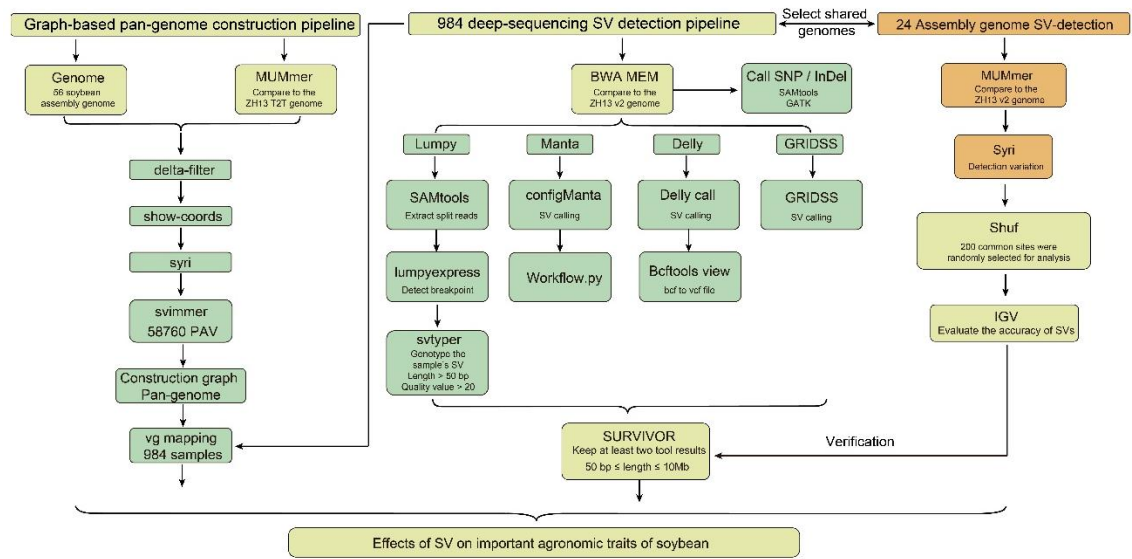


**Analysis of deep-resequencing data of 984 soybean accessions reveals
structural variations underlying agronomic traits**

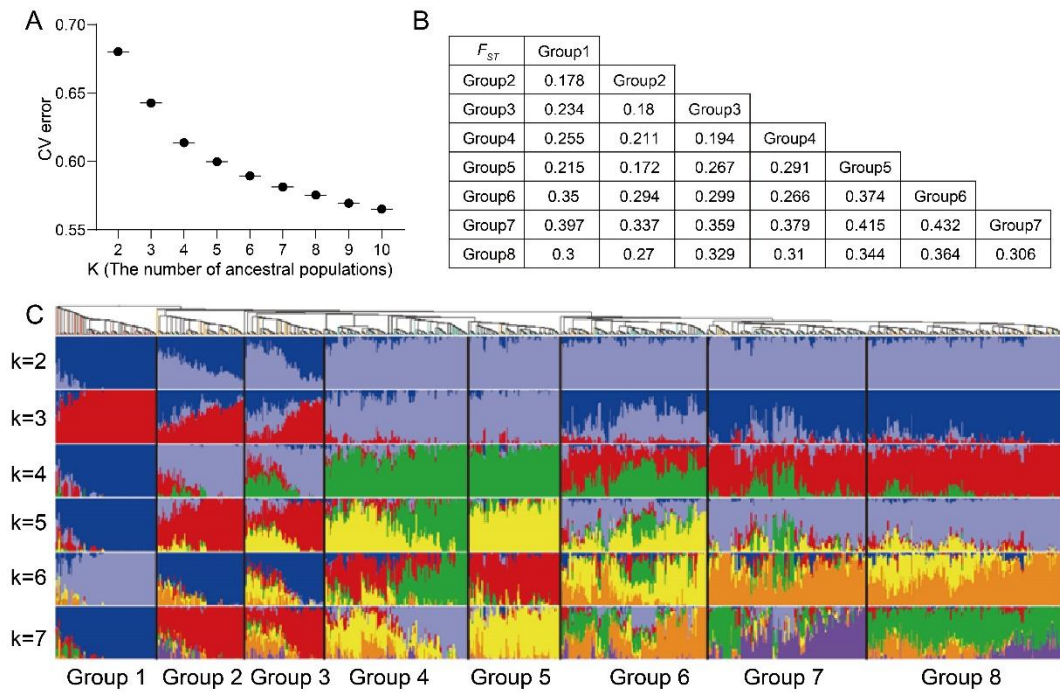
Li et al.



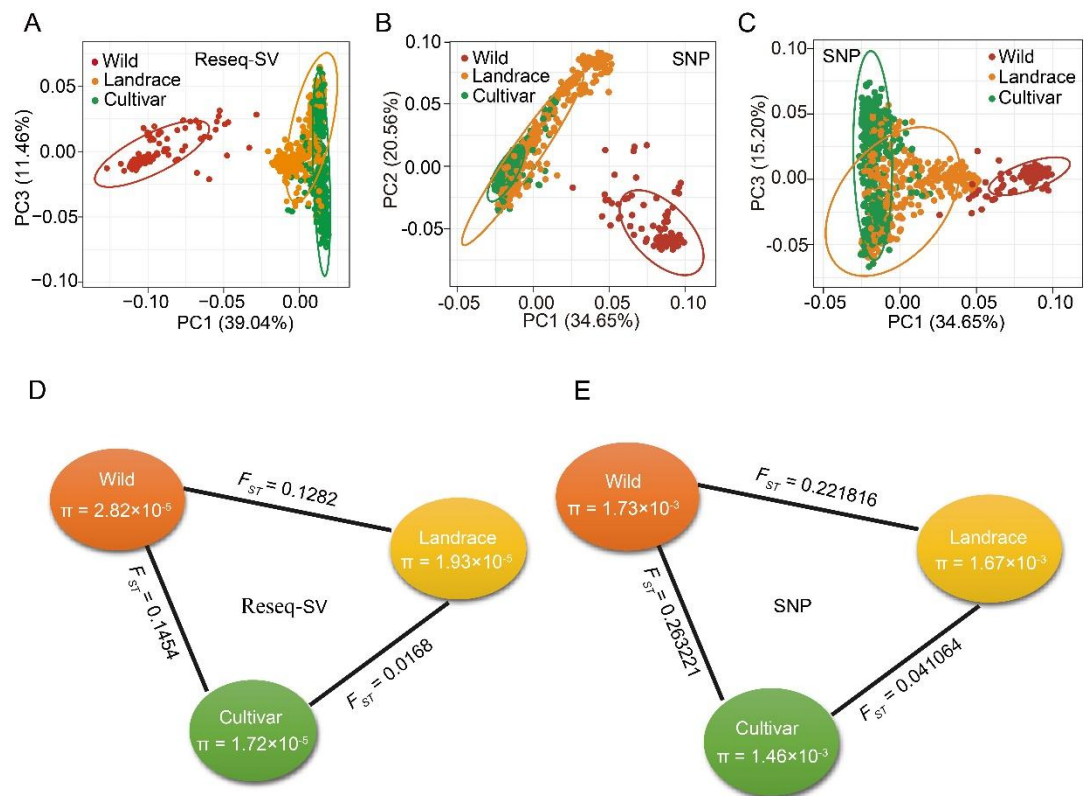
Supplementary Fig. 1. Collection of diverse soybean germplasm resources. All germplasm resources are available at SoyOD (https://bis.zju.edu.cn/soyod/phenome/Germplasm_search).



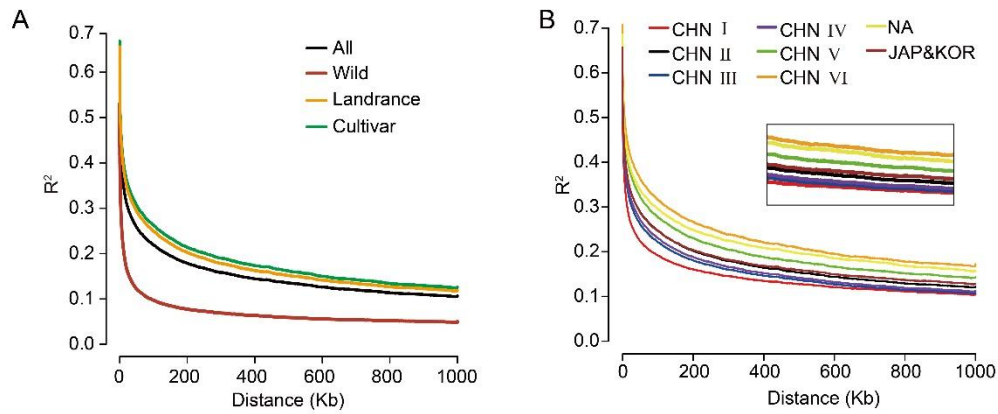
Supplementary Fig. 2. Comprehensive structural variation detection pipeline integrating 984 resequenced accessions and 56 assembled soybean genomes.



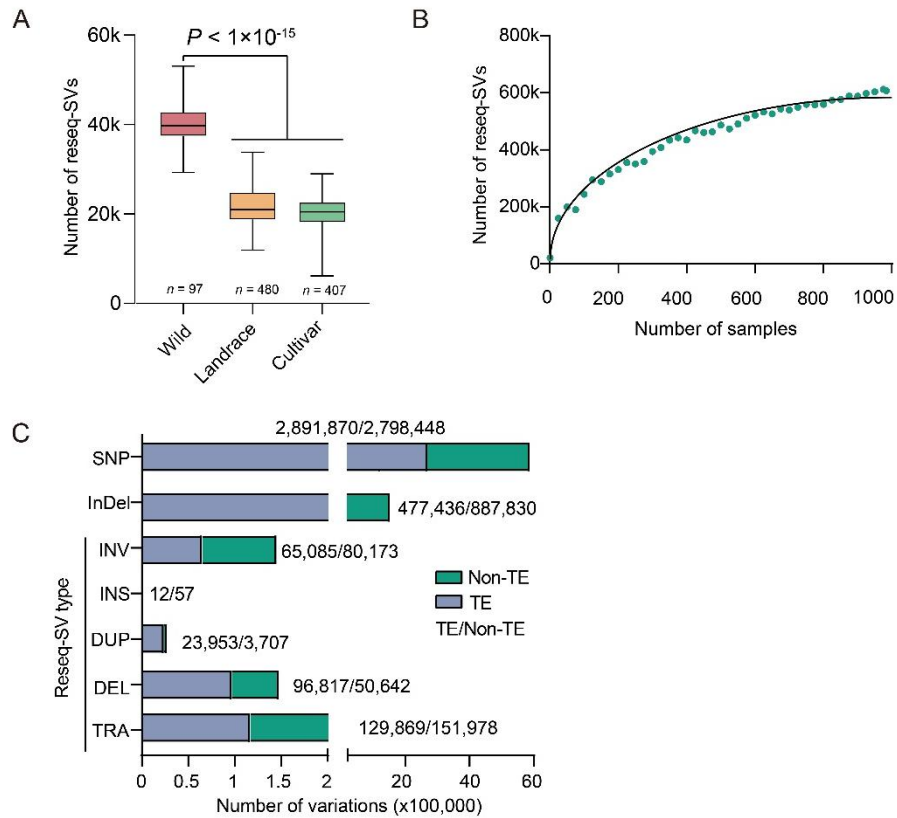
Supplementary Fig. 3. Population structure analysis using admixture. (A) Cross-validation error (CV) for different K values. The optimal number of ancestral populations was determined by the K with the lowest CV error. (B) Pairwise F_{ST} matrix between groups. (C) Ancestral composition of each accession inferred by ADMIXTURE. Each vertical bar represents an individual accession, partitioned into colored segments proportional to the contribution of each ancestral population. The K (the number of ancestral populations) ranged from 2 to 7.



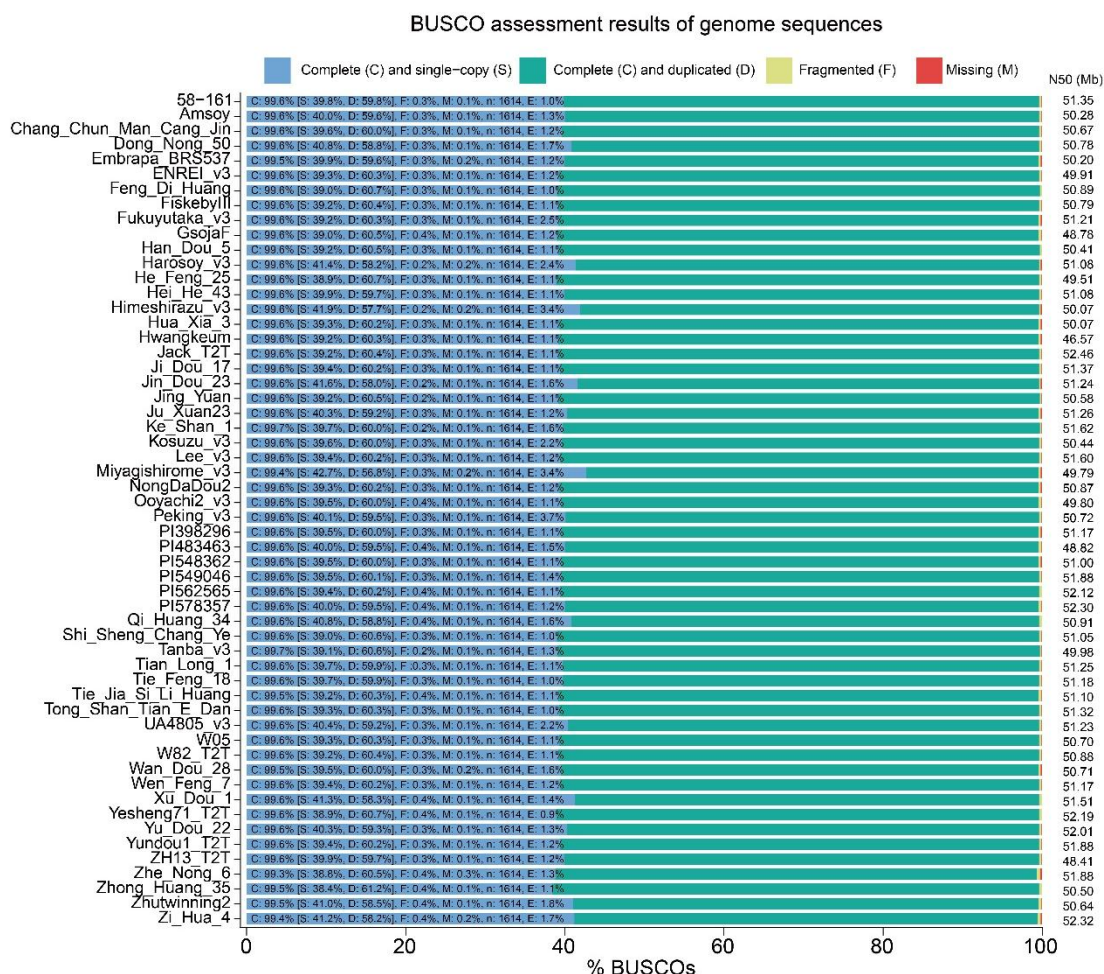
Supplementary Fig. 4. Population structure and nucleotide diversity analysis. (A-C) PCA plots based on SV and SNP data. Different colors represent wild soybeans (red), landraces (orange) and cultivars (green). (D, E) Nucleotide diversity (π) was calculated for wild soybeans, landraces and cultivars using reseq-SV and SNP data, respectively.



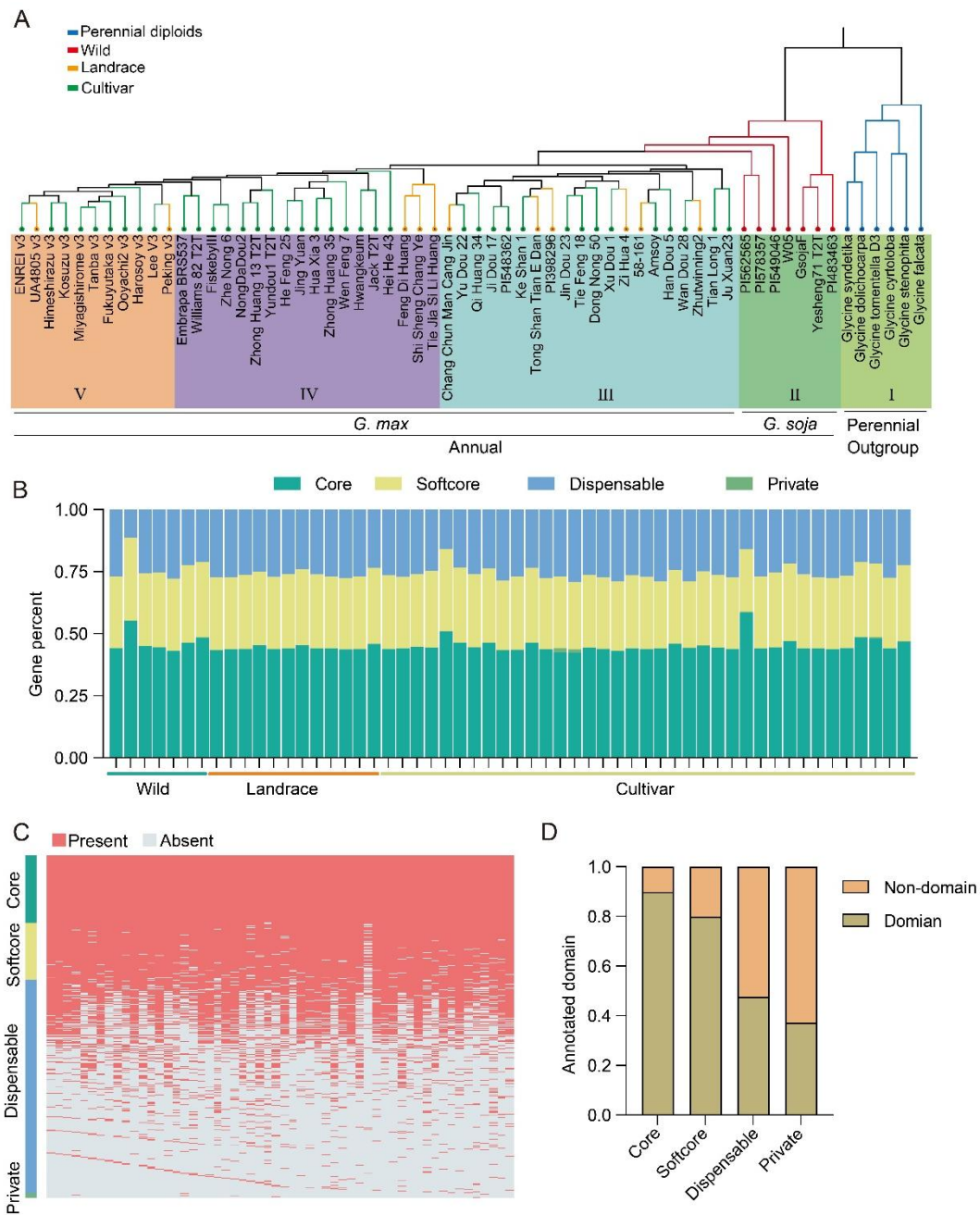
Supplementary Fig. 5. Linkage disequilibrium analysis. Relationship of average LD with genomic distance in wild, cultivar and landrace using SNP data (A). Decay of LD in different ecological regions SNP data (B). The x-axis indicates the physical distance (Kb) between SNP pairs on the same chromosome, and the y-axis represents the squared correlation coefficient (R^2). The decay of LD, as measured by R^2 , of SNPs as a function of physical distances between markers. In Glycine, eight clades are designated as China (I-VI), North America (NA), Japan and Korea (JAP& KOR).



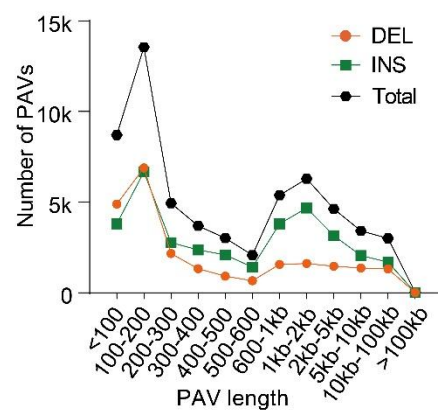
Supplementary Fig. 6. Evaluation and annotation of reseq-SVs based on 984 accessions. (A) Comparison of reseq-SV counts among wild soybeans, landraces, and cultivated varieties. n represents the sample size per group. Statistical significance was assessed using a two-tailed t -test. The box edges represent the upper and lower quartiles, with the center line indicating the median and the whiskers extending to the minimum and maximum values. (B) Simulation showing the increase in the number of reseq-SVs detected as the sample size increases. (C) Overlap between different types of reseq-SVs and transposable elements (TEs).



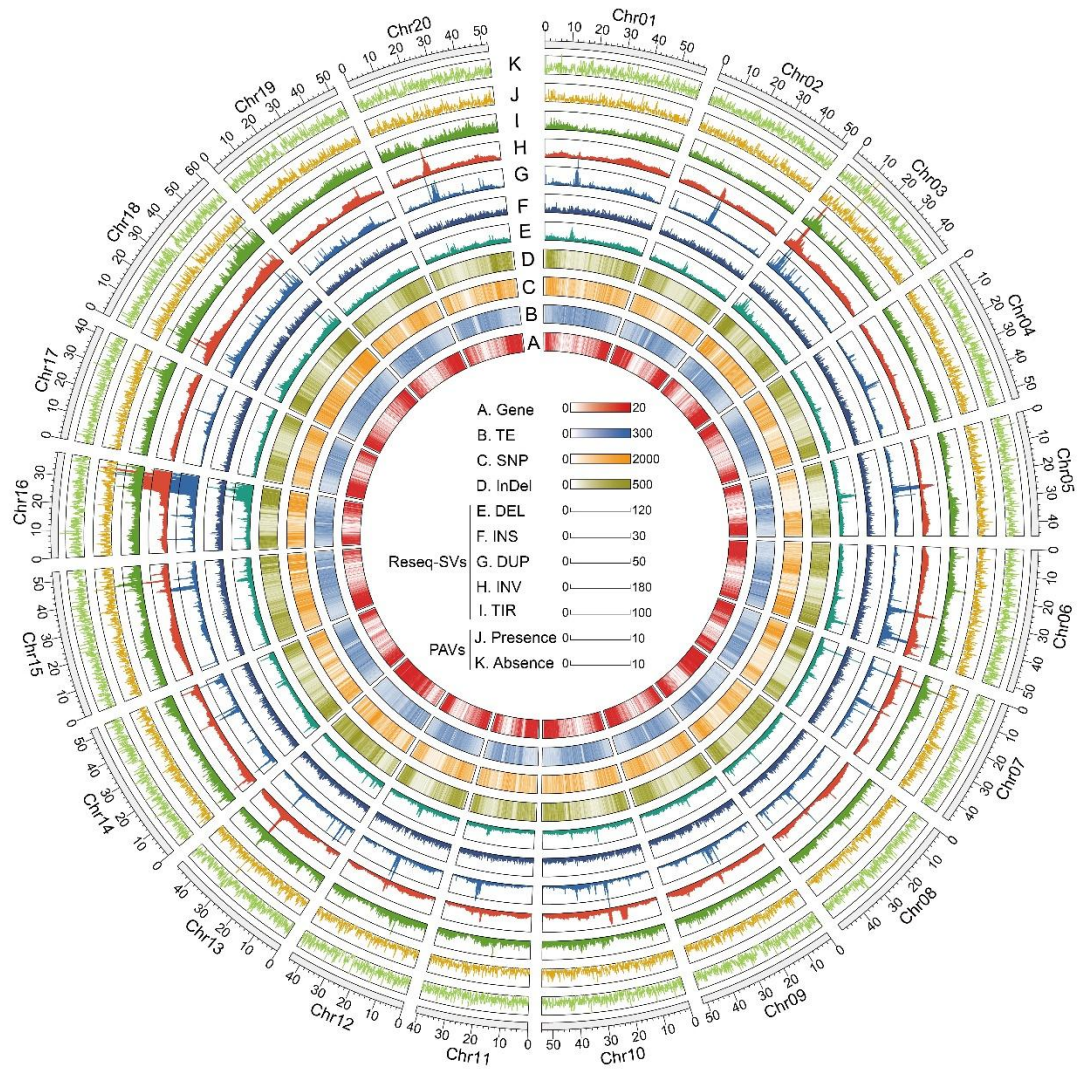
Supplementary Fig. 7. BUSCO evaluation of 56 soybean genome sequences and assembly-stats calculating N50. C: Complete BUSCOs; S: Complete and single-copy BUSCOs; D: Complete and duplicated BUSCOs; F: Fragmented BUSCOs; M: Missing BUSCOs; n: Total BUSCO groups searched. Average N50 = 50.81 Mb. Average single-copy Embryophyta genes = 99.58%.

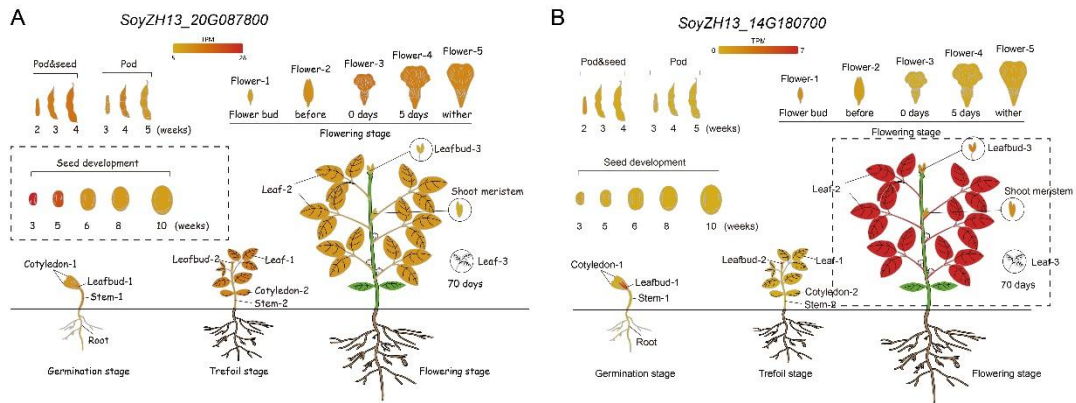


Supplementary Fig. 8. Phylogenetic relationships of Glycine species and PFAM domain distribution across gene families. (A) Phylogenetic relationships analysis of 62 genomes. Six perennial accessions were included as outgroups. Number of classified genes in each genome (B) and presence/absence information of pan gene families of all populations (C). (D) Proportion of genes with PFAM domains in core, soft core, dispensable, and private genomes.

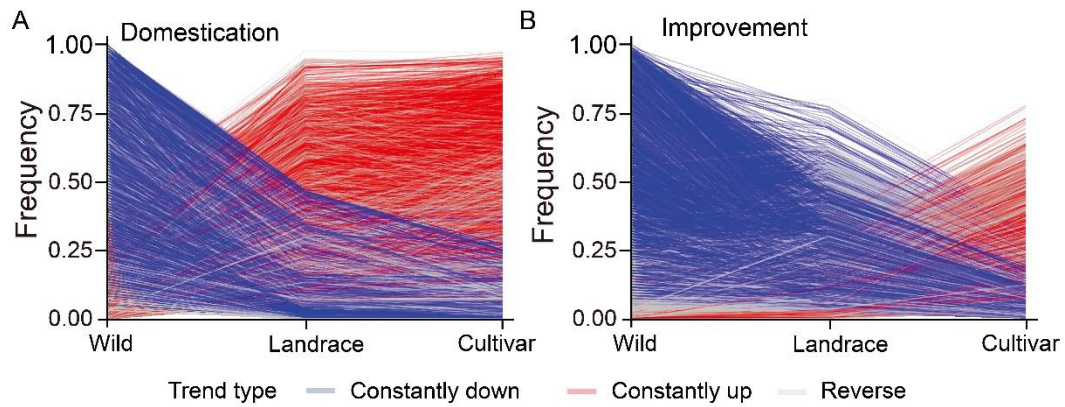


Supplementary Fig. 9. Size distribution of PAVs by types.

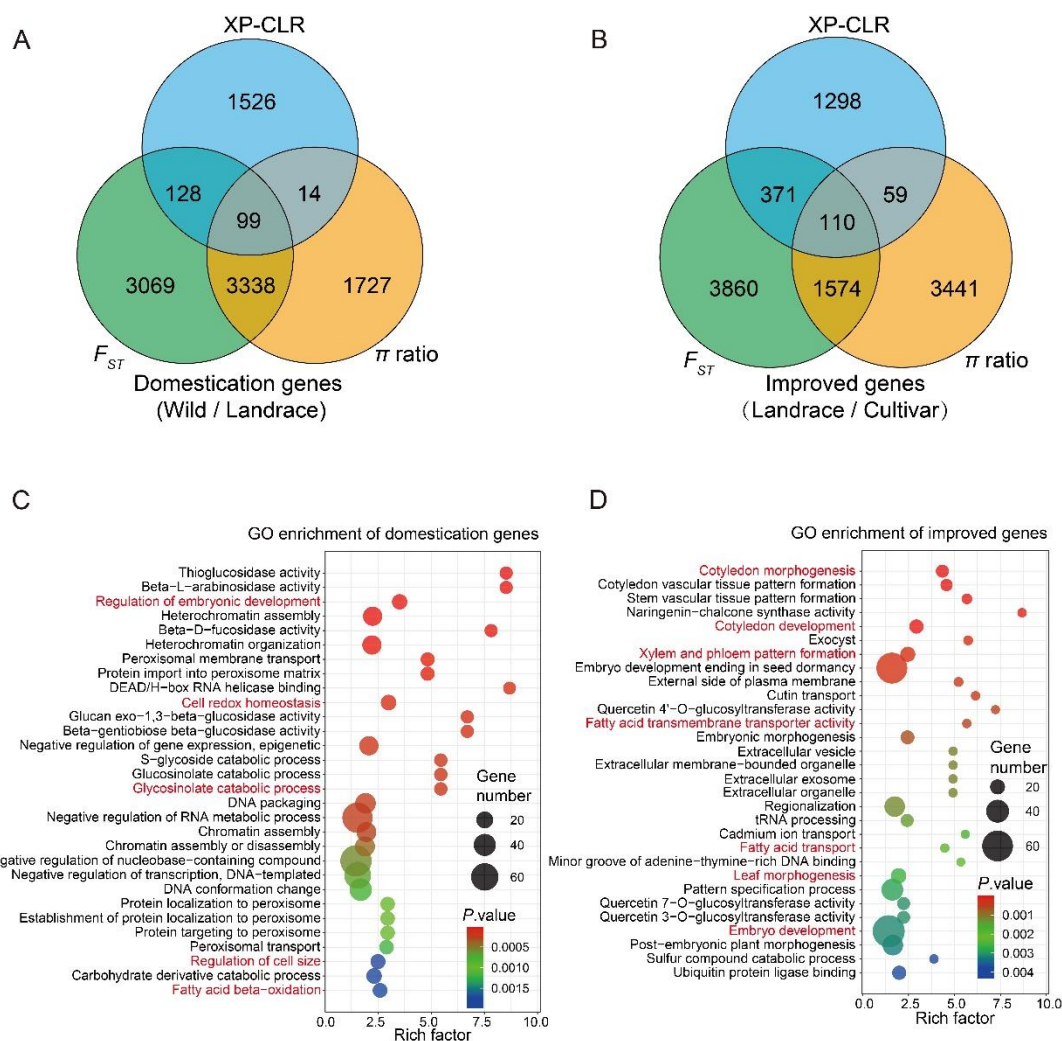




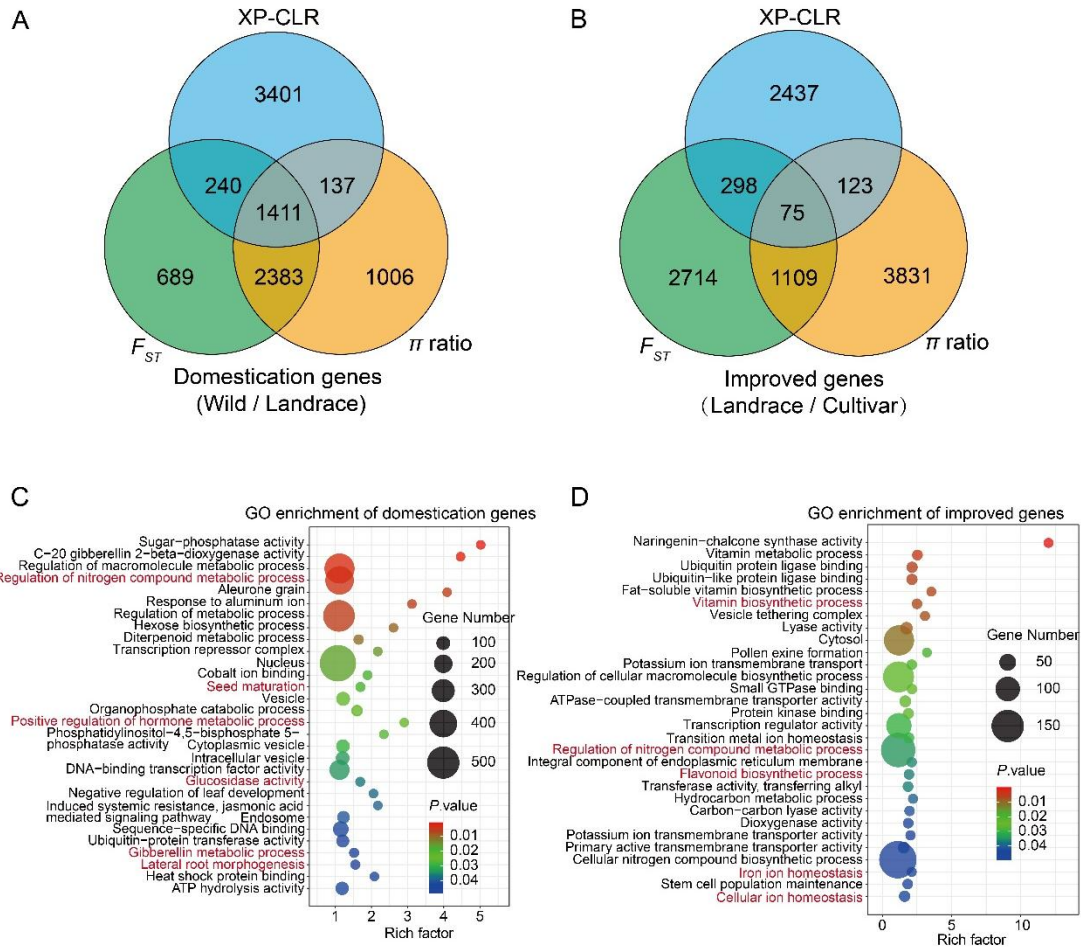
Supplementary Fig. 11. Tissue expression of candidate genes associated with 100-seed weight and leaf shape. (A, B) Tissue expression of the *SoyZH13_20G087800* and *SoyZH13_14G180700* genes. The expression diagrams of different tissues were generated using SoyOD website (<https://bis.zju.edu.cn/soyod/transcriptom/tissue>).



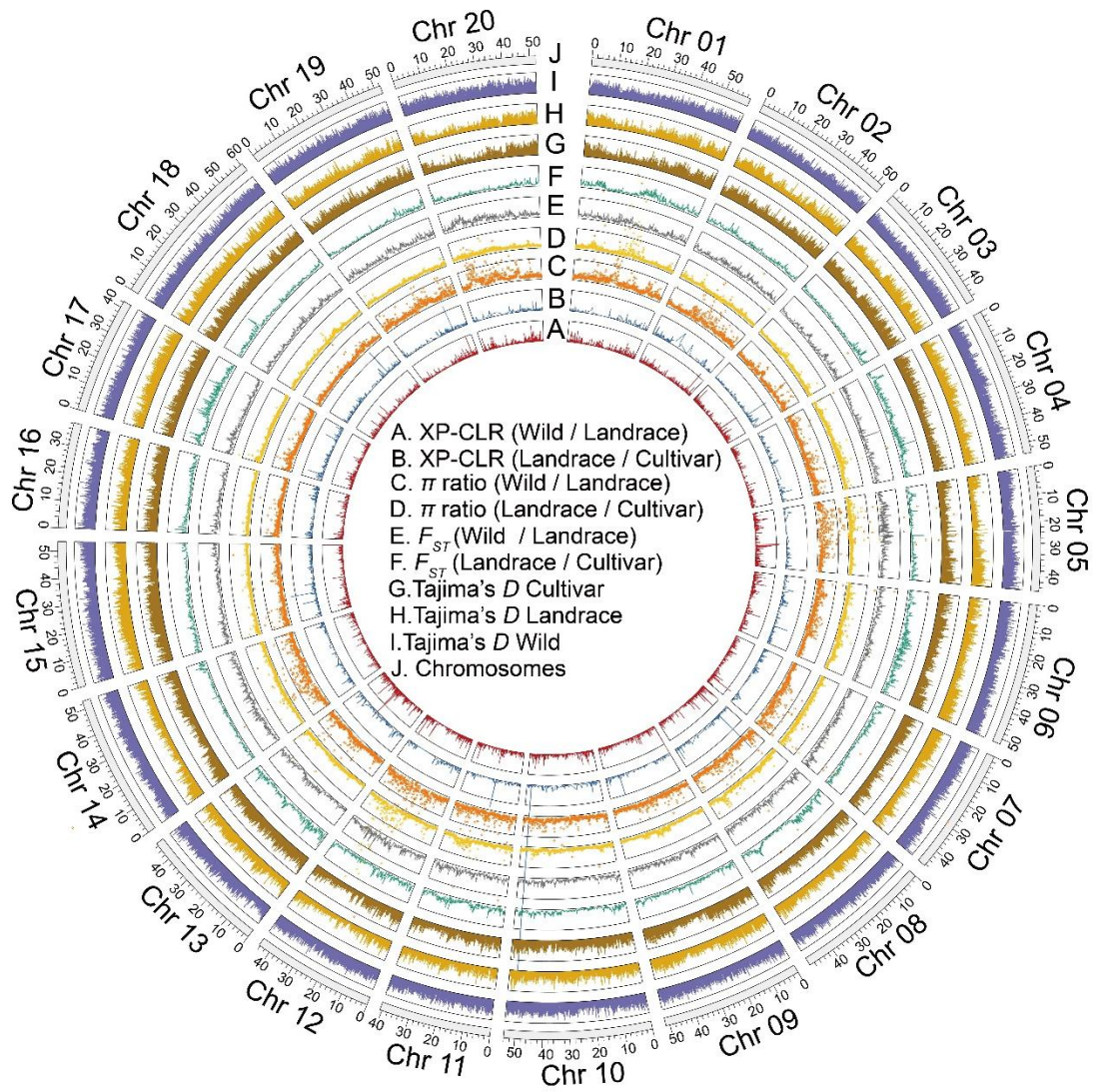
Supplementary Fig. 12. Identification of selection signals during soybean domestication and improvement. (A) Occurrence frequency patterns of putative selected reseq-SVs during domestication. (B) Occurrence frequency patterns of putative selected reseq-SVs during improvement. FDR computed using a two-sided Fisher's exact test.



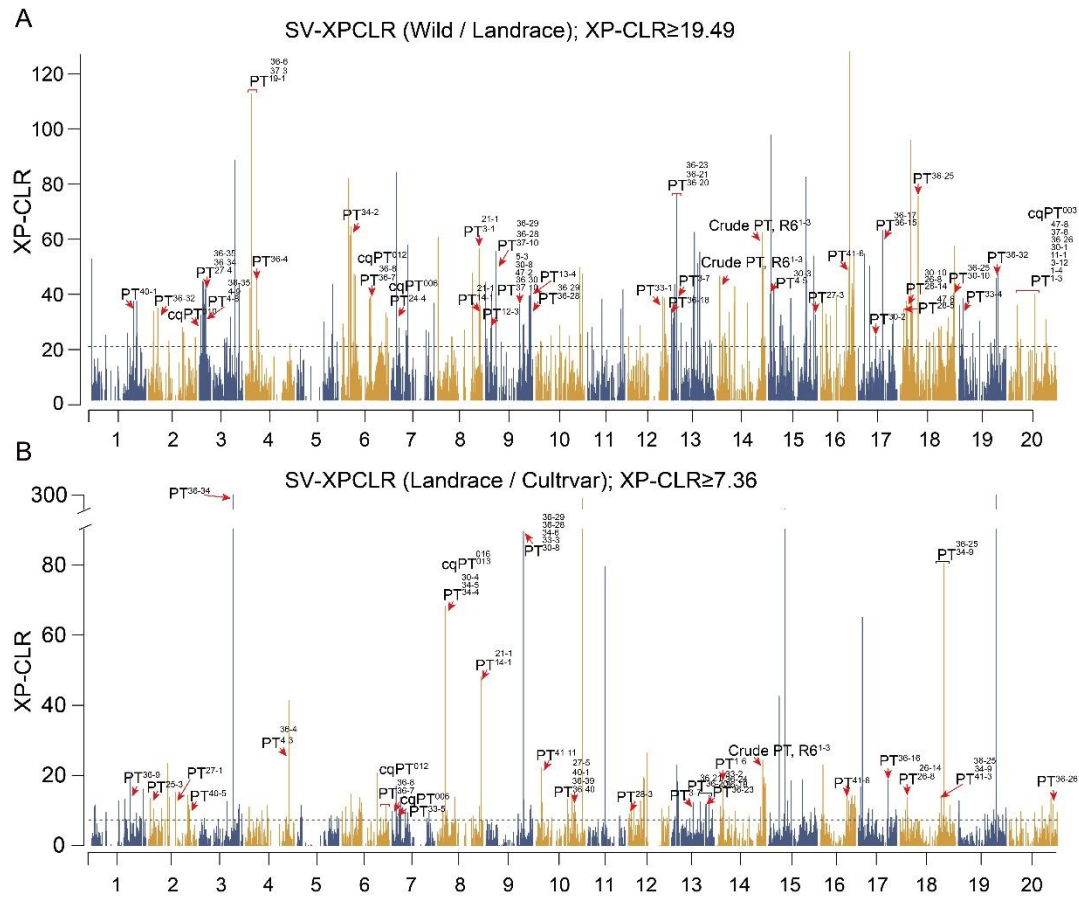
Supplementary Fig. 13. Gene counts and GO enrichment analysis during domestication and improvement using reseq-SV data. (A, B) Venn diagrams showing the number of selected genes during domestication (A) and improvement (B). Population selection was performed using XP-CLR, F_{ST} and π ratio screening. (C, D) GO enrichment analysis of candidate genes during domestication (C) and improvement (D). The size of the circles represents the number of genes, and the color indicates the significance of the GO terms.



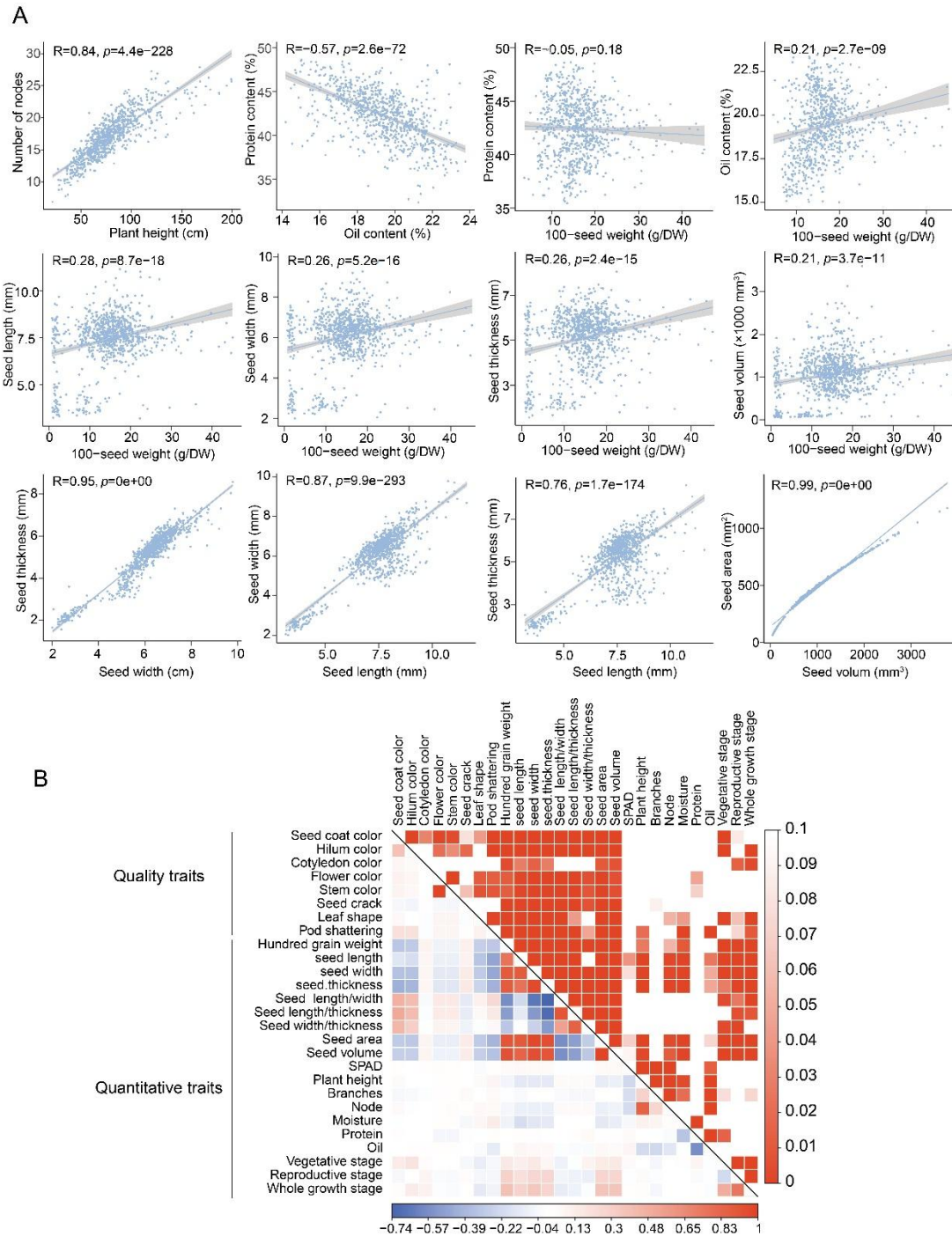
Supplementary Fig. 14. Gene counts and GO enrichment analysis during domestication and improvement using SNP data. (A, B) Venn diagrams showing the number of selected genes during domestication (A) and improvement (B). Population selection was performed using XP-CLR, F_{ST} and π ratio screening. (C, D) GO enrichment analysis of candidate genes during domestication (C) and improvement (D). The size of the circles represents the number of genes, and the color indicates the significance of the GO terms.



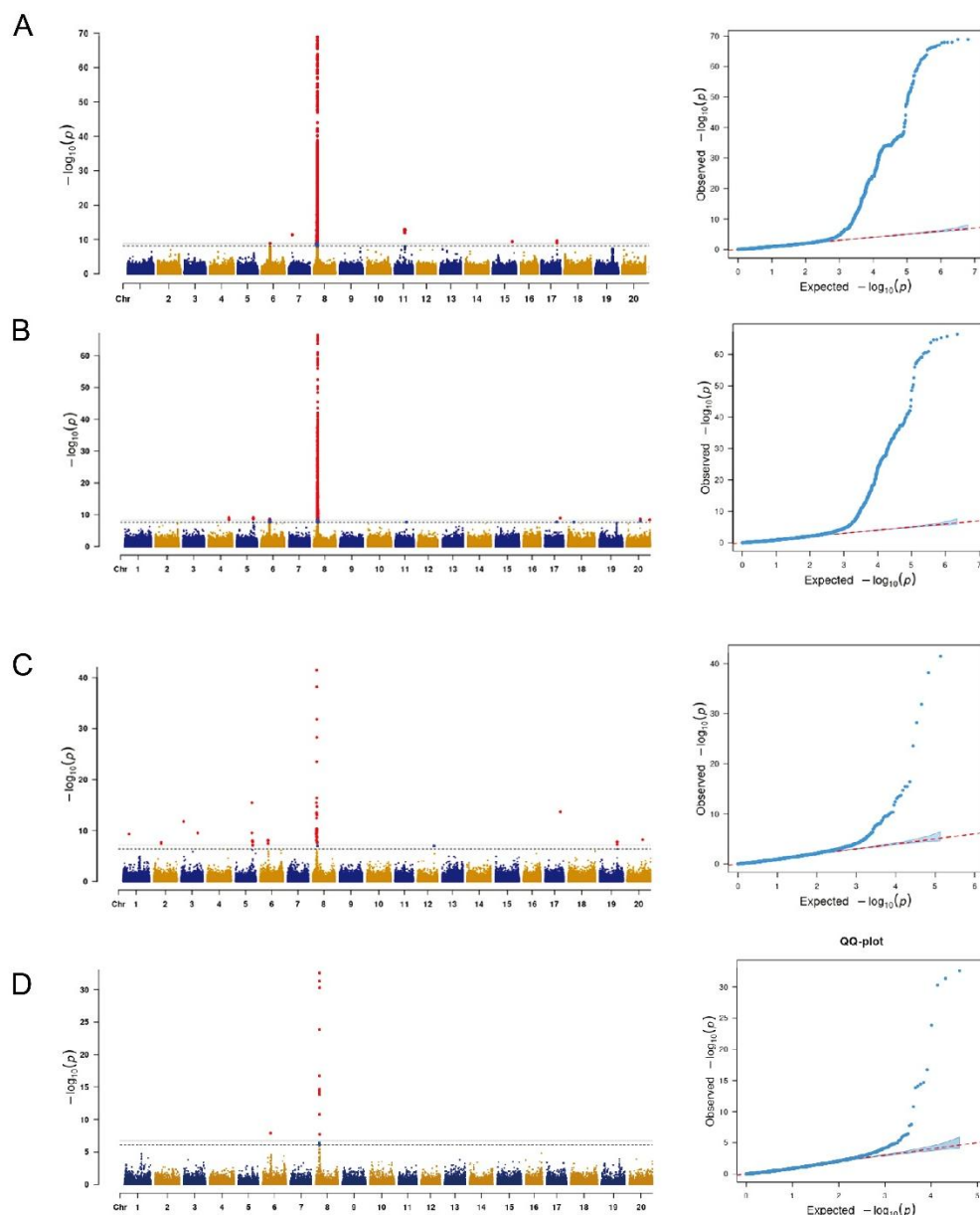
Supplementary Fig. 15. Genome-wide selective sweep maps. Selective sweeps were identified using XP-CLR, Tajima's D , π ratio and F_{ST} based on reseq-SV data.



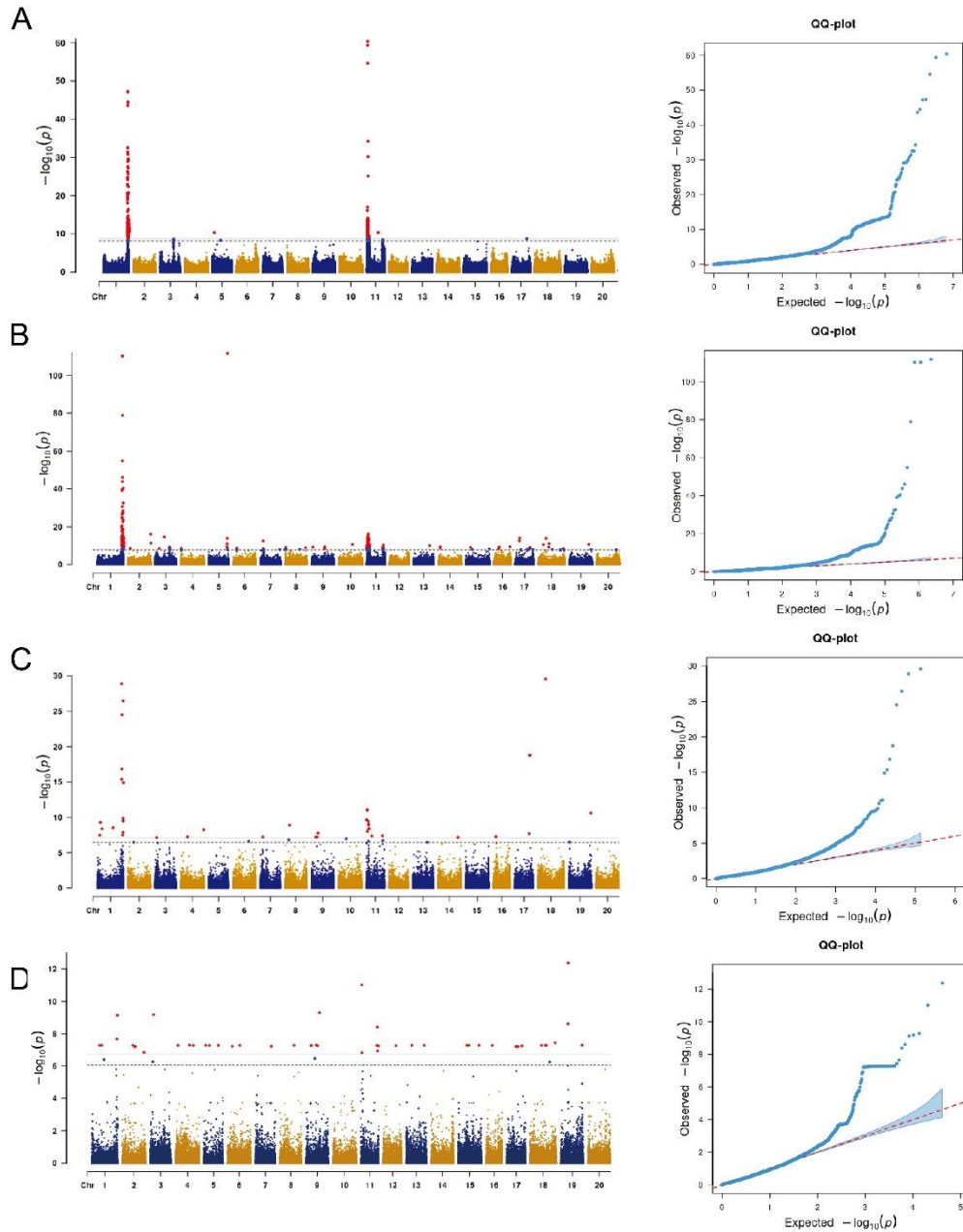
Supplementary Fig. 16. Genomic functional annotation of selective sweeps during domestication and improvement. (A) Genome-wide distribution of selective signals during domestication (from wild to landrace; XP-CLR threshold ≥ 19.49). (B) Genome-wide distribution of selective signals during improvement (from landrace to cultivars; XP-CLR threshold ≥ 7.31). Horizontal dashed lines indicate the selection threshold corresponding to the top 5% of XP-CLR values. Protein-related QTLs (PT) were retrieved from SoyOD (<https://bis.zju.edu.cn/soyod/>) to assess their overlap with genomic intervals under selection identified by reseq-SV analysis.



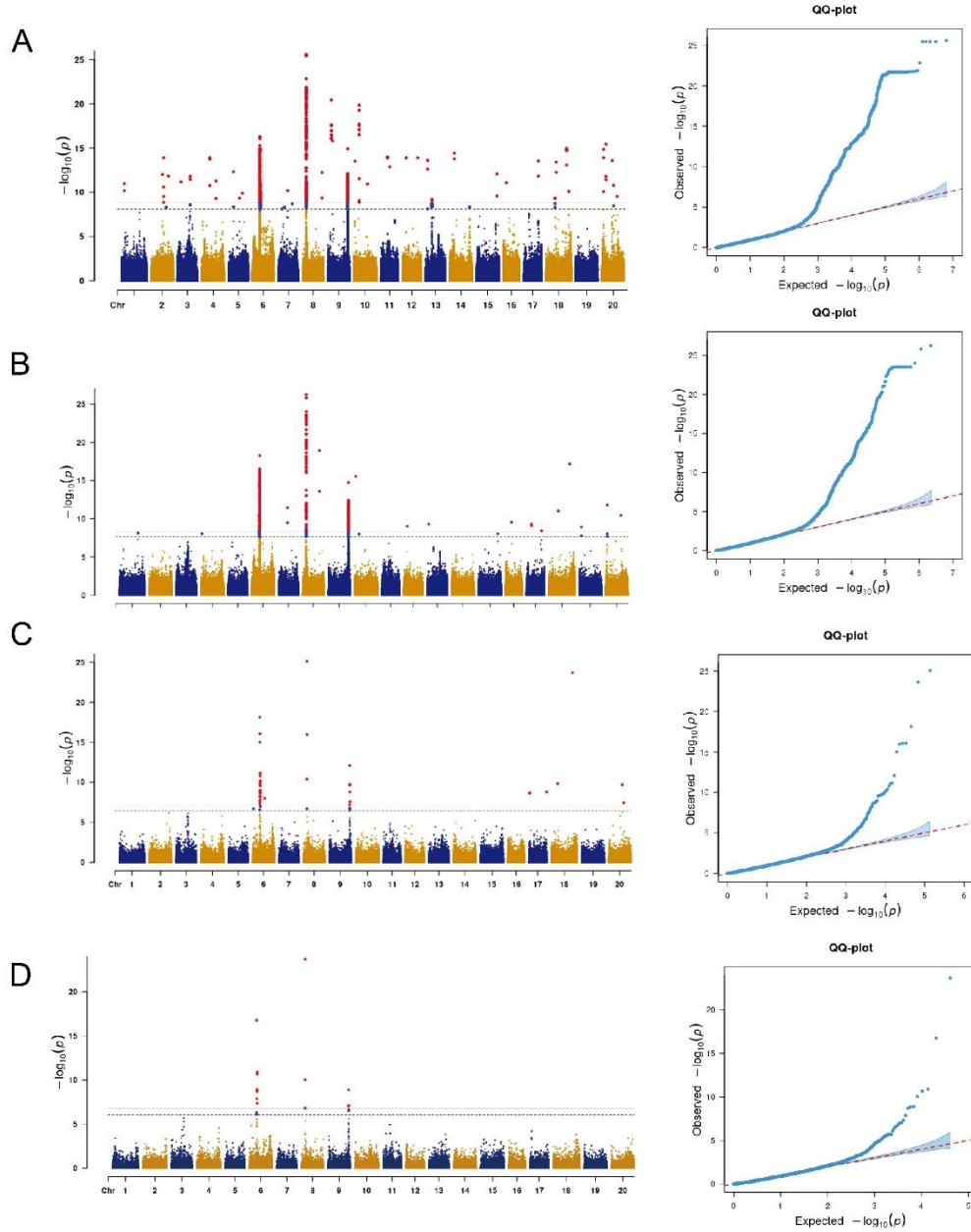
Supplementary Fig. 17. Correlation analysis among different traits. (A) Correlation matrix displaying relationships among traits. (B) The upper triangle shows p values, while the lower triangle shows Pearson correlation coefficients.



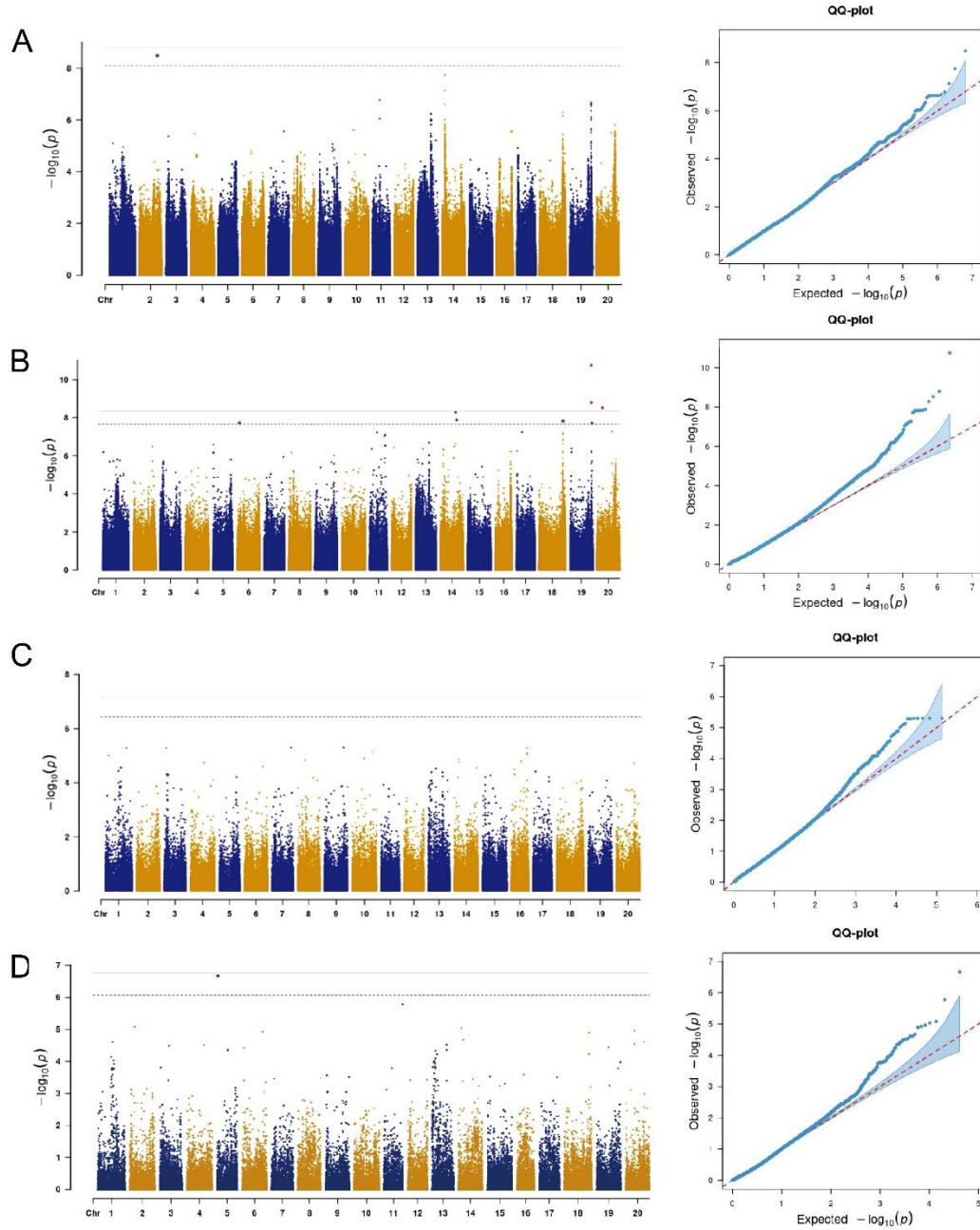
Supplementary Fig. 18. GWAS of seed coat color using EMMAX method. Manhattan and quantile-quantile plots for seed coat color based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



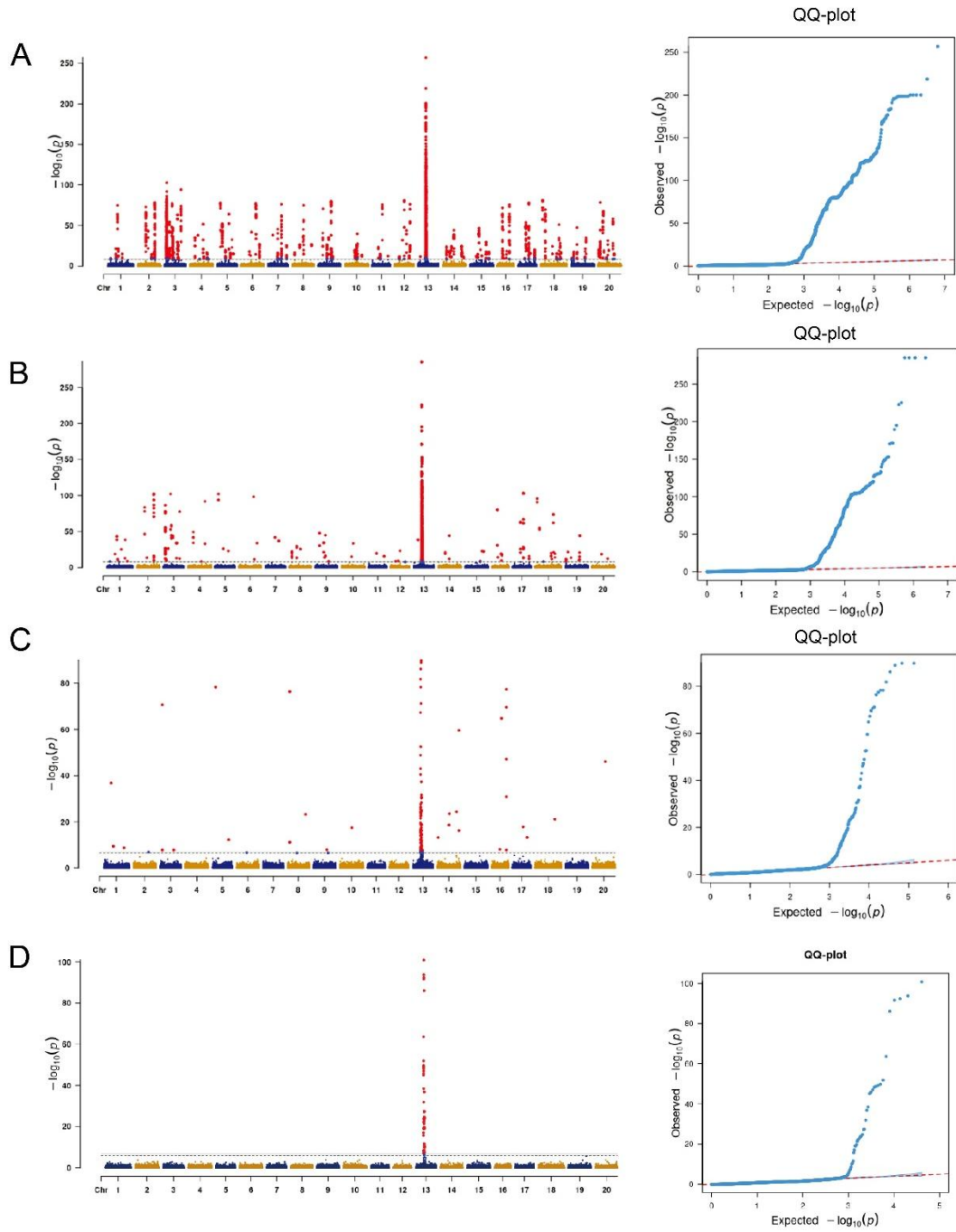
Supplementary Fig. 19. GWAS of cotyledon color using EMMAX method. Manhattan and quantile-quantile plots for cotyledon color based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



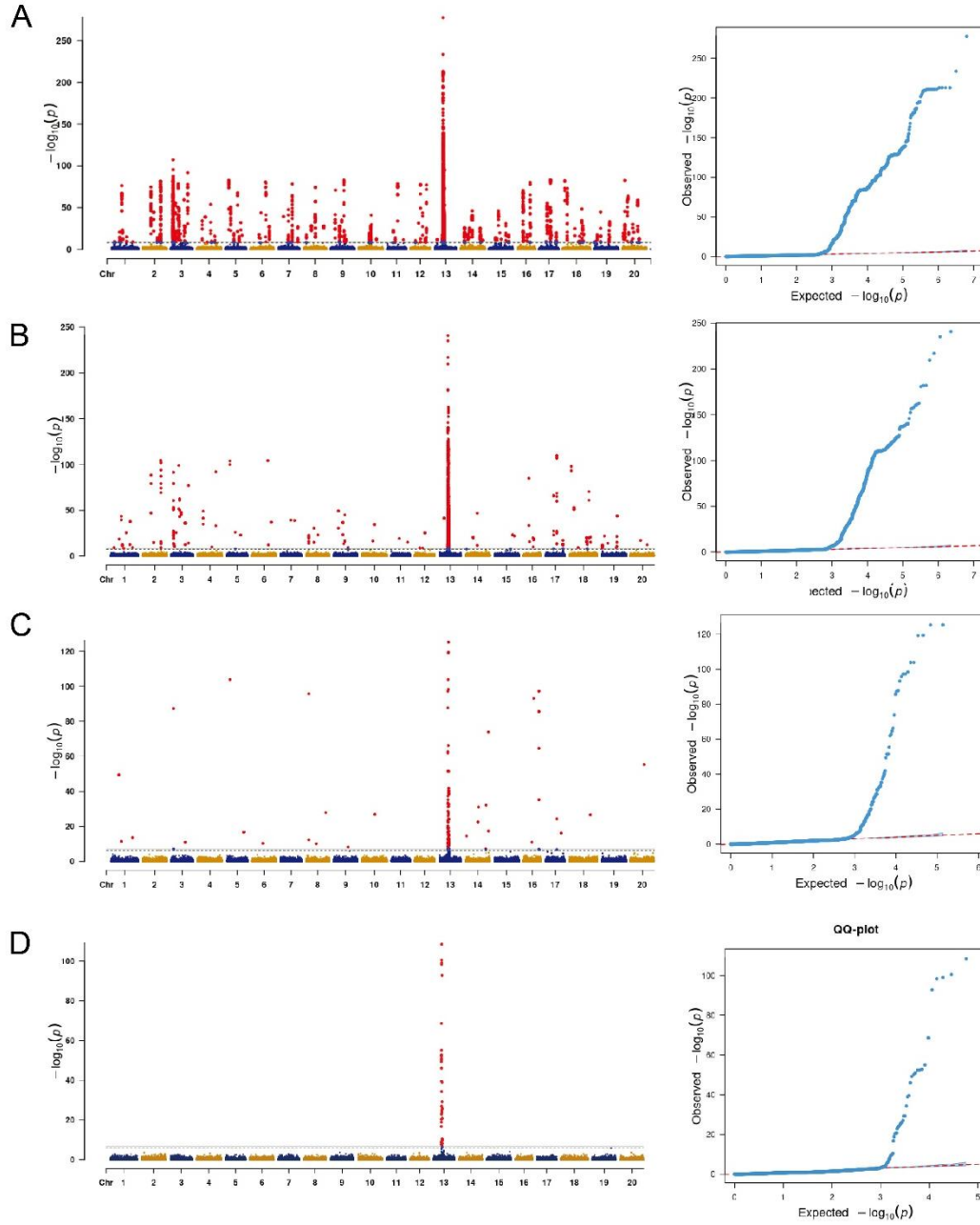
Supplementary Fig. 20. GWAS of hilum color using EMMAX method. Manhattan and quantile-quantile plots for hilum color based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



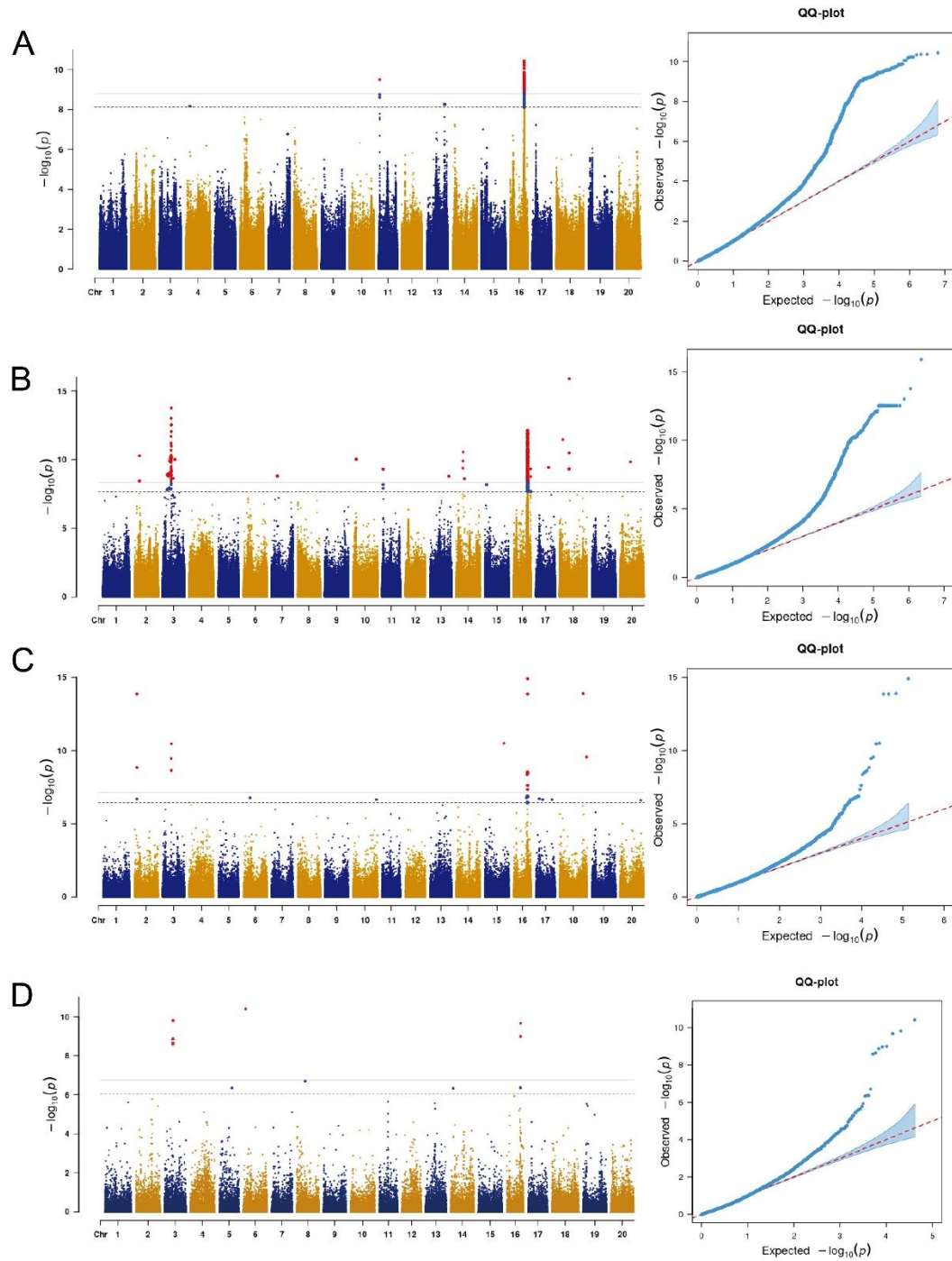
Supplementary Fig. 21. GWAS of seed crack using EMMAX method. Manhattan and quantile-quantile plots for seed crack based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



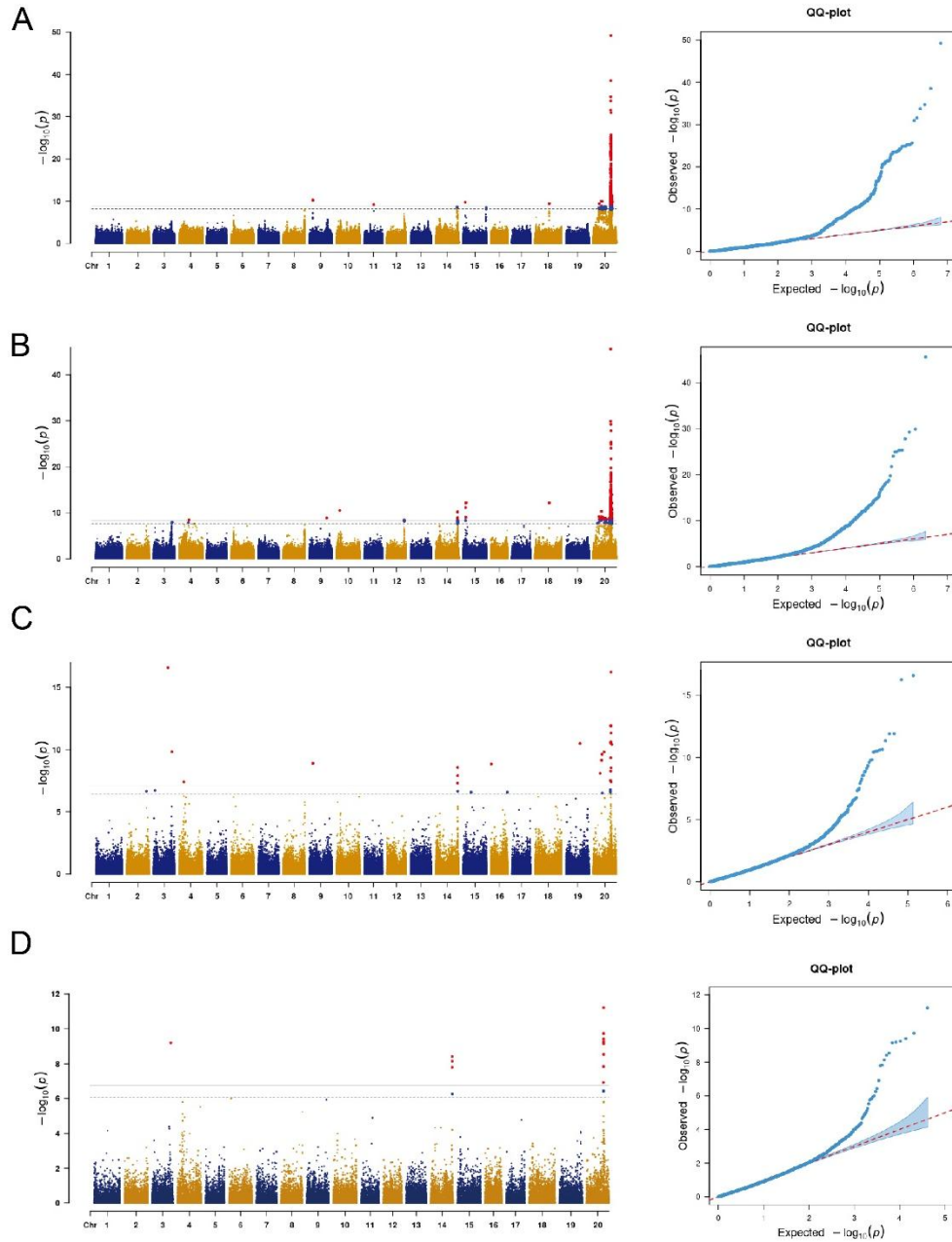
Supplementary Fig. 22. GWAS of flower color using EMMAX method. Manhattan and quantile-quantile plots for flower color based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



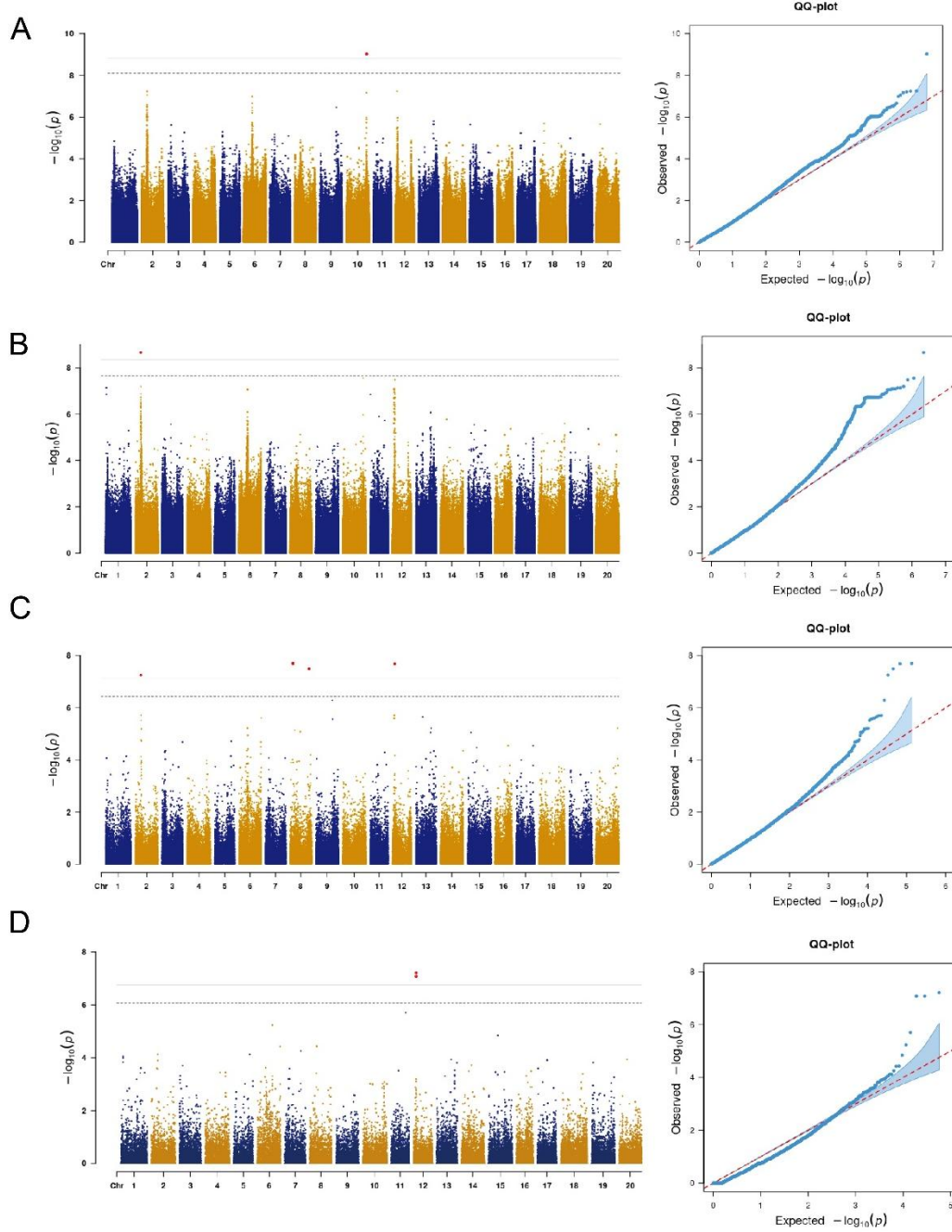
Supplementary Fig. 23. GWAS of hypocotyl color using EMMAX method. Manhattan and quantile-quantile plots for hypocotyl color based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



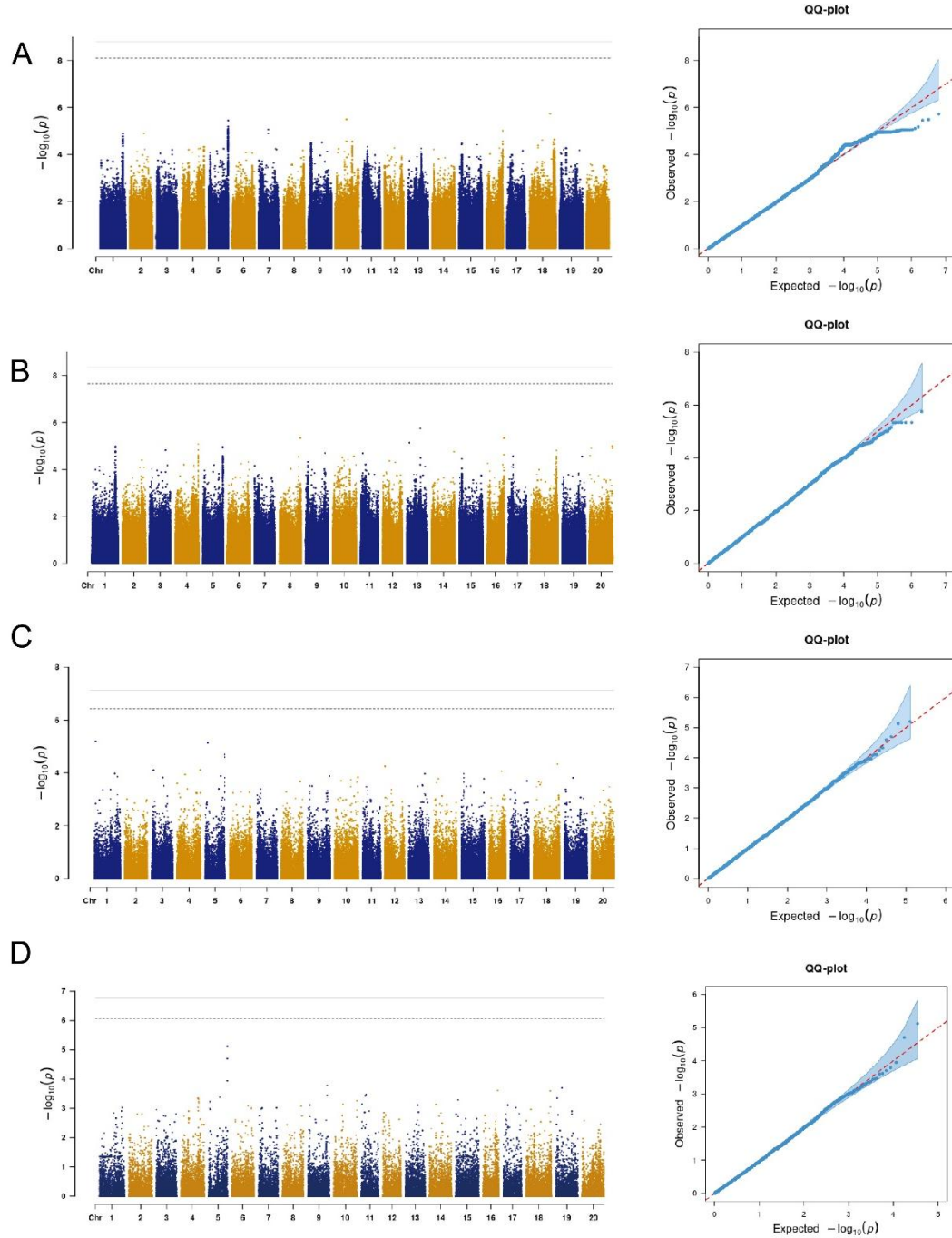
Supplementary Fig. 24. GWAS of pod shattering using EMMAX method. Manhattan and quantile-quantile plots for pod shattering based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



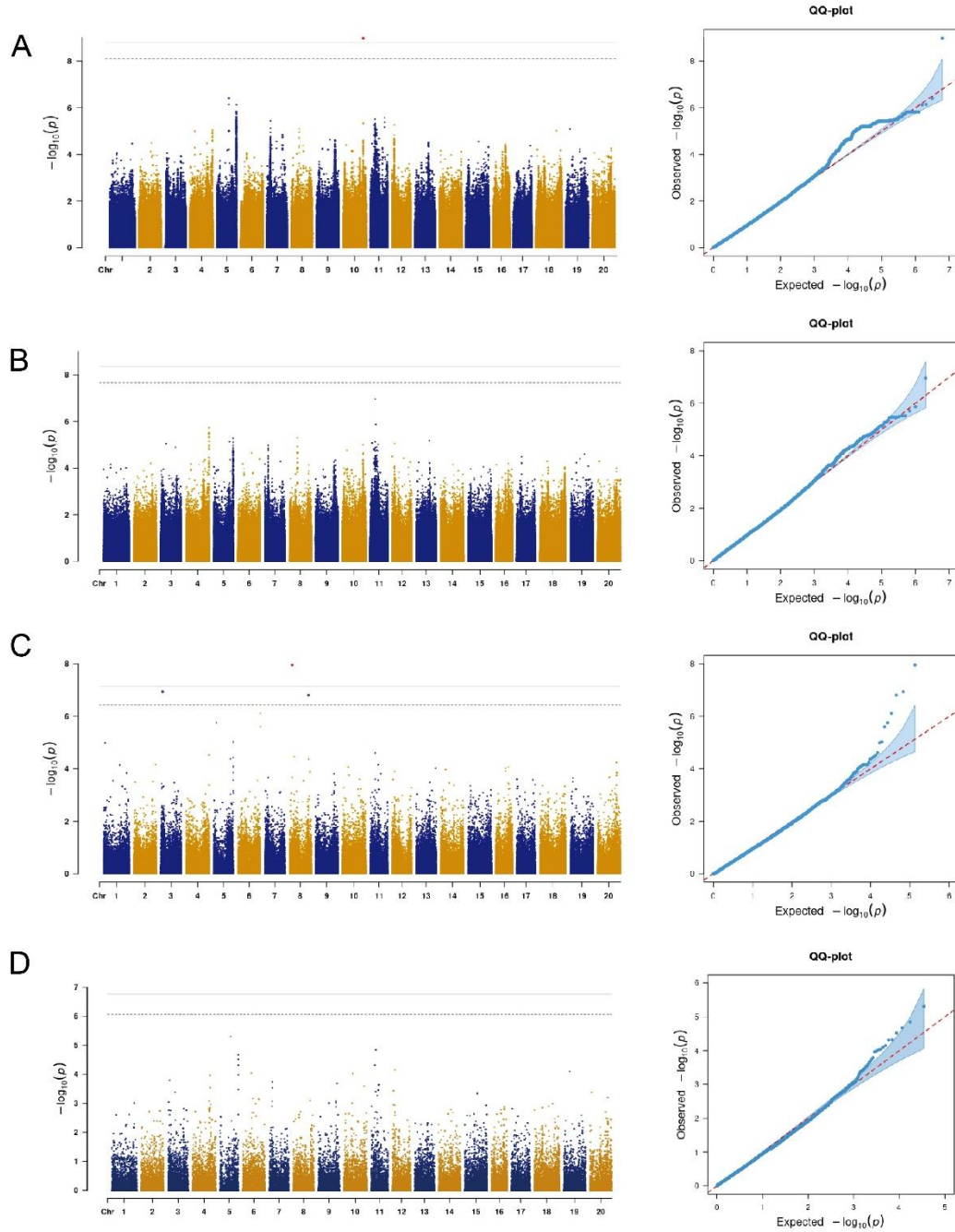
Supplementary Fig. 25. GWAS of leaf shape using EMMAX method. Manhattan and quantile-quantile plots for leaf shape based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



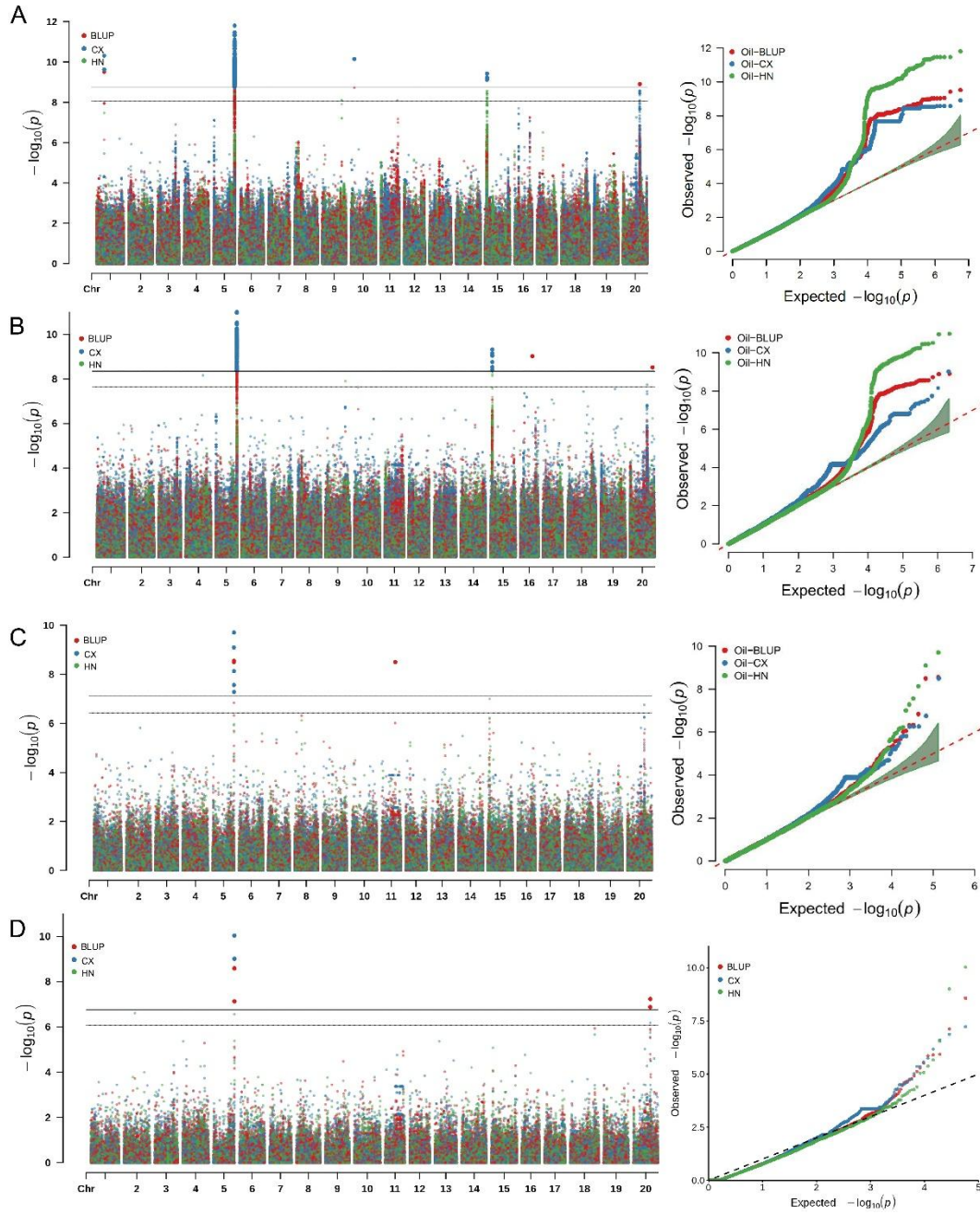
Supplementary Fig. 26. GWAS of vegetative stage using EMMAX method. Manhattan and quantile-quantile plots for vegetative stage based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



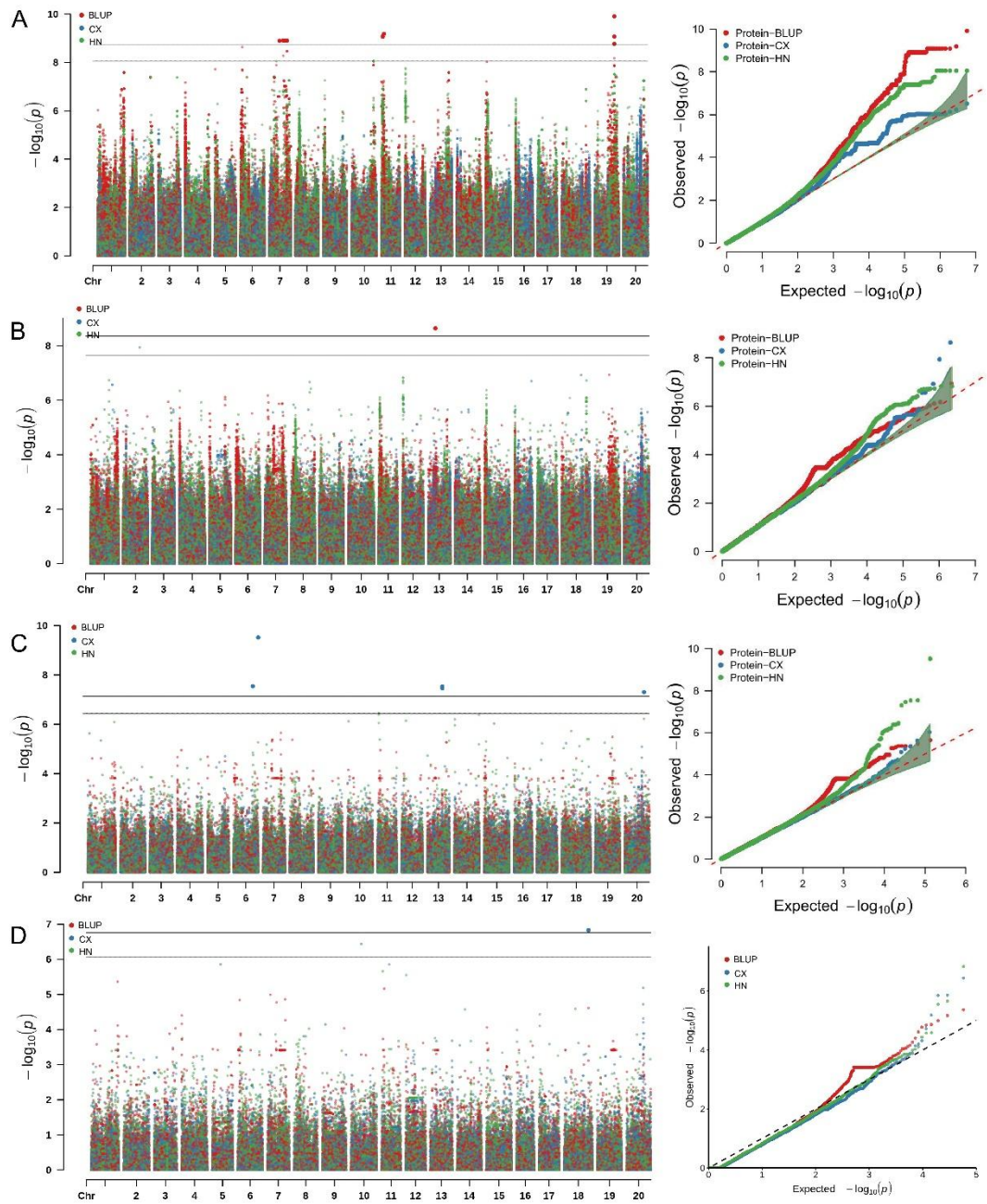
Supplementary Fig. 27. GWAS of reproductive stage using EMMAX method. Manhattan and quantile-quantile plots for reproductive stage based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



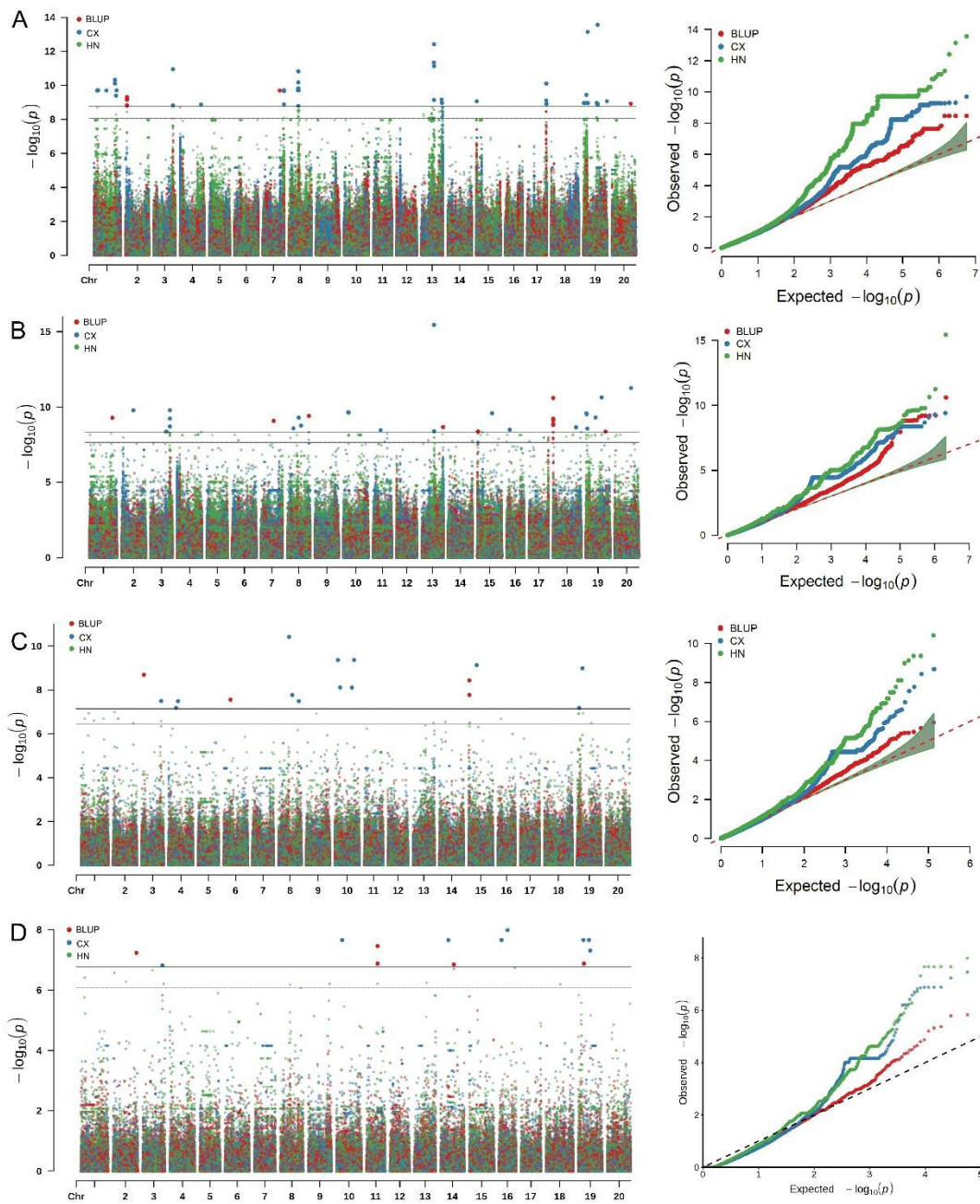
Supplementary Fig. 28. GWAS of whole growth stage using EMMAX method. Manhattan and quantile-quantile plots for whole growth stage based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



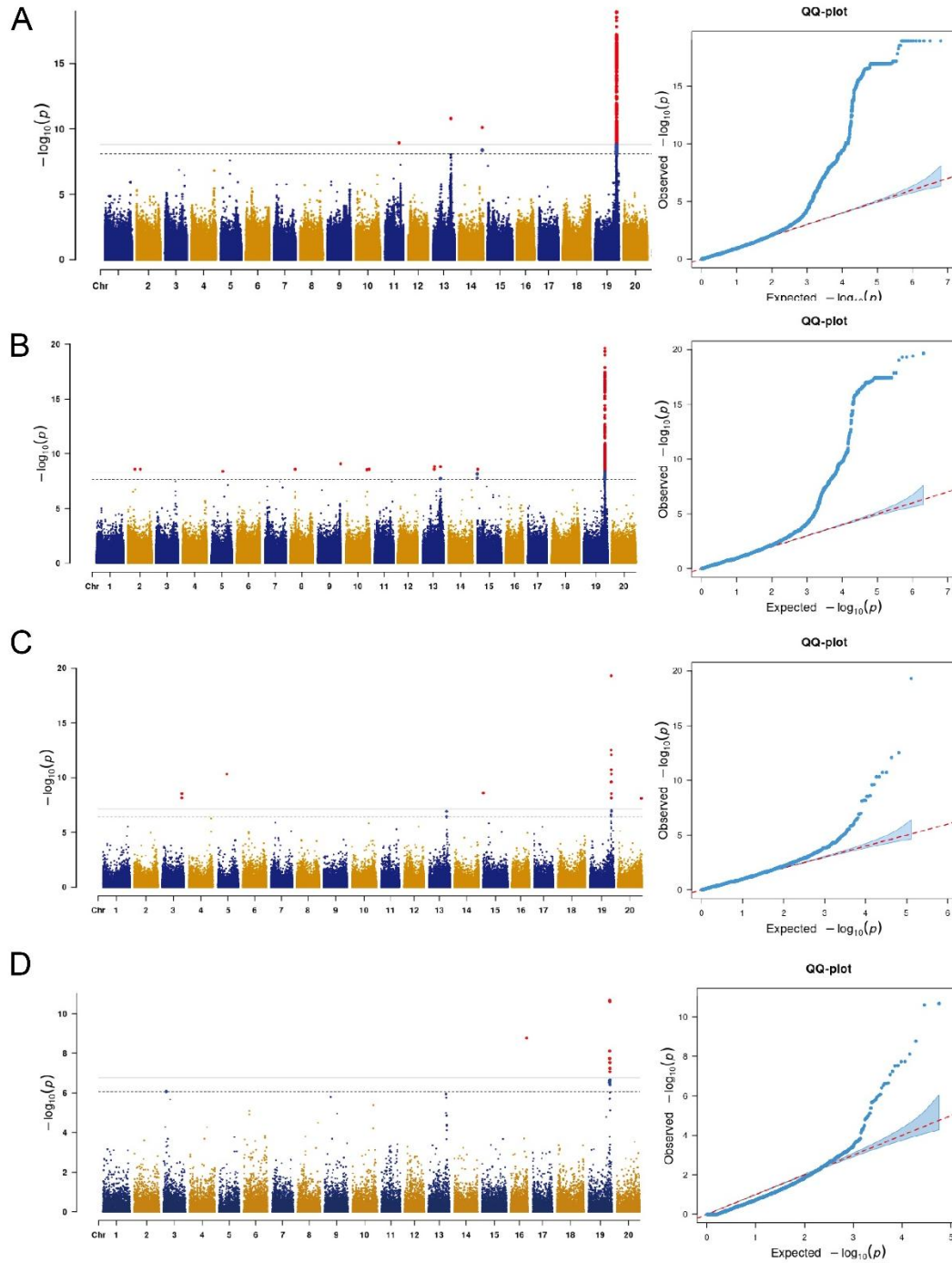
Supplementary Fig. 29. GWAS of oil content using EMMAX method. Manhattan and quantile-quantile plots for oil content based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



Supplementary Fig. 30. GWAS of protein content using EMMAX method. Manhattan and quantile-quantile plots for protein content based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).

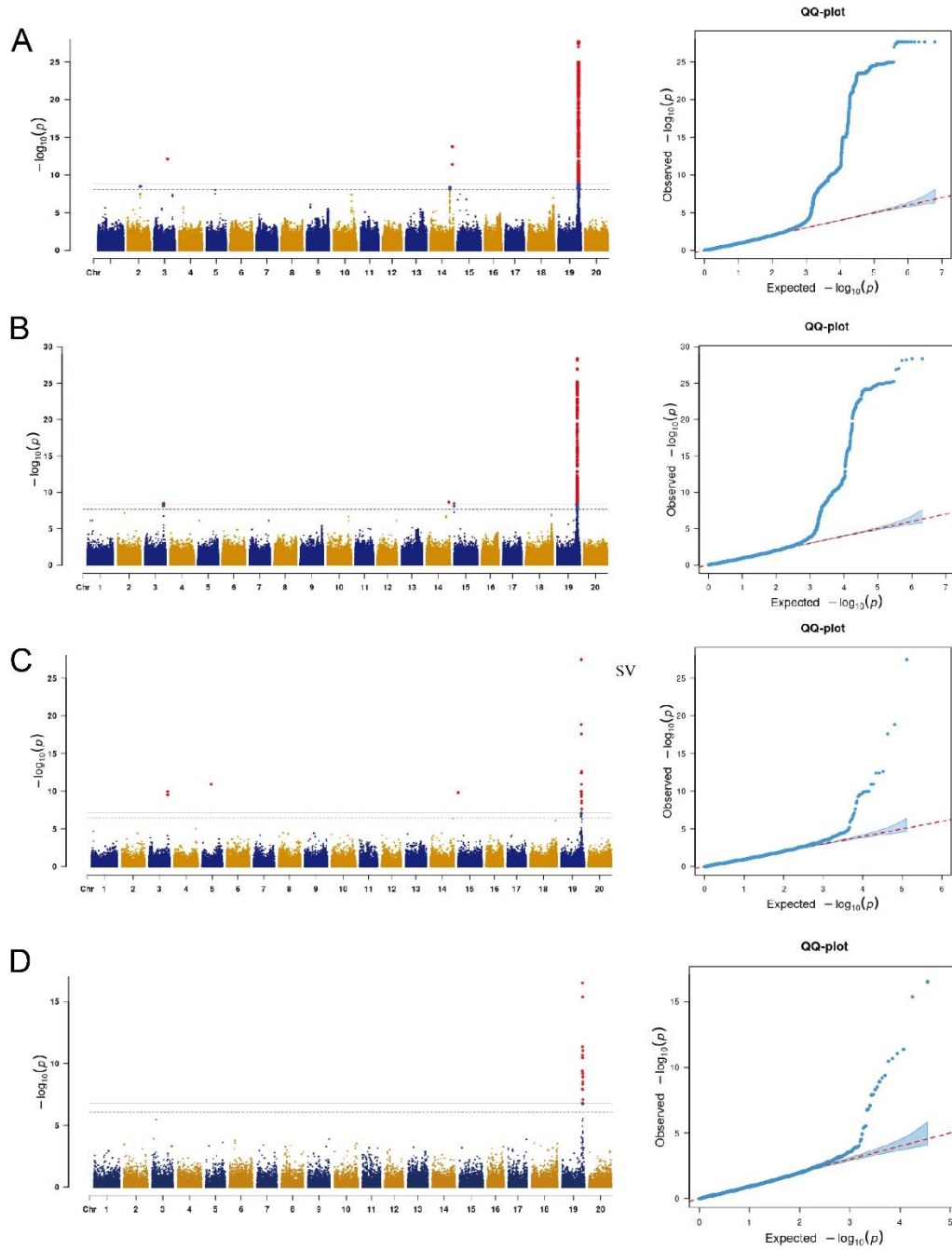


Supplementary Fig. 31. GWAS of seed moisture using EMMAX method. Manhattan and quantile-quantile plots for seed moisture based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



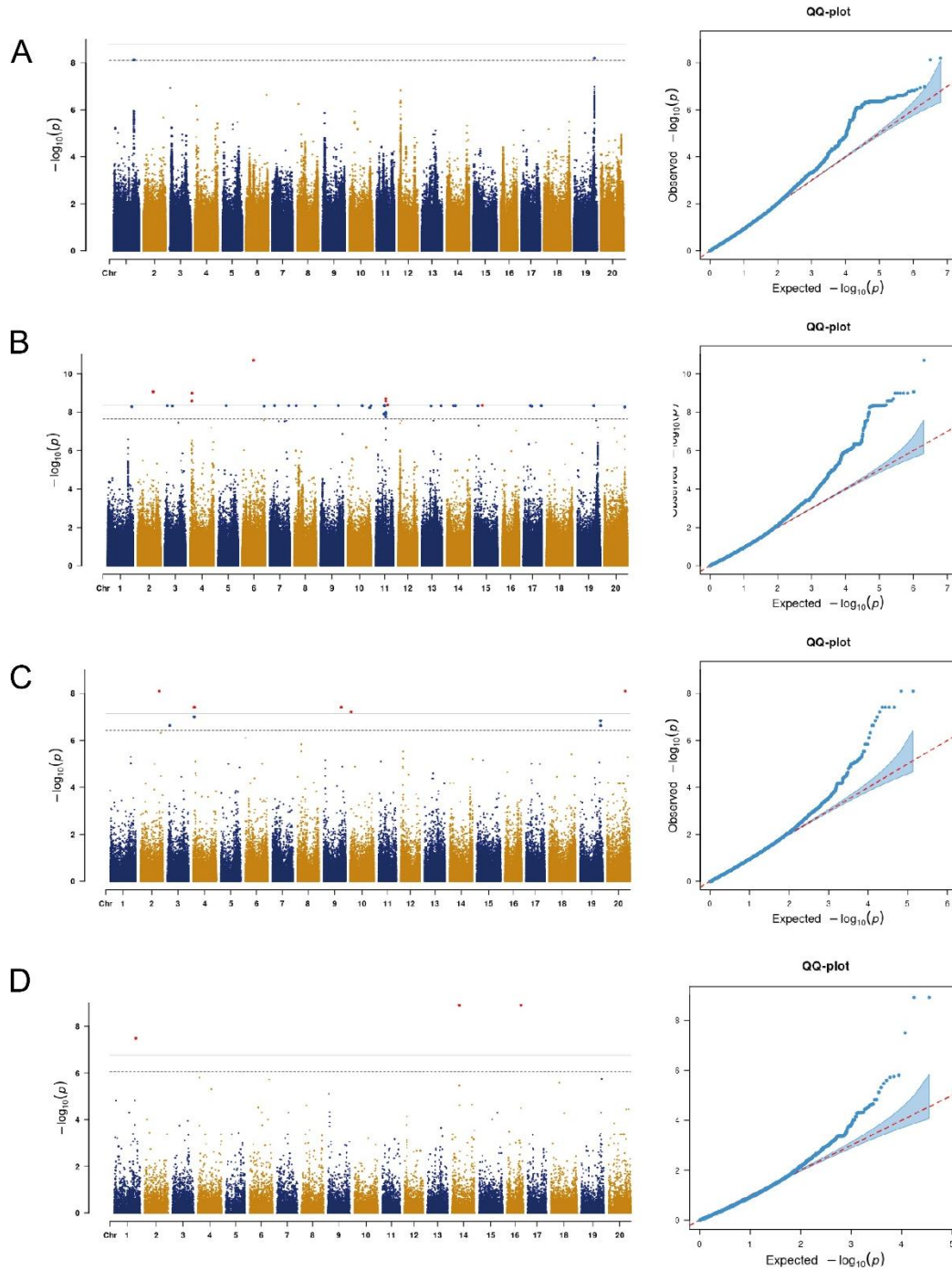
Supplementary Fig. 32. GWAS of plant height using EMMAX method. Manhattan and quantile-quantile plots for plant height based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines,

respectively (where n is the total number of variants).



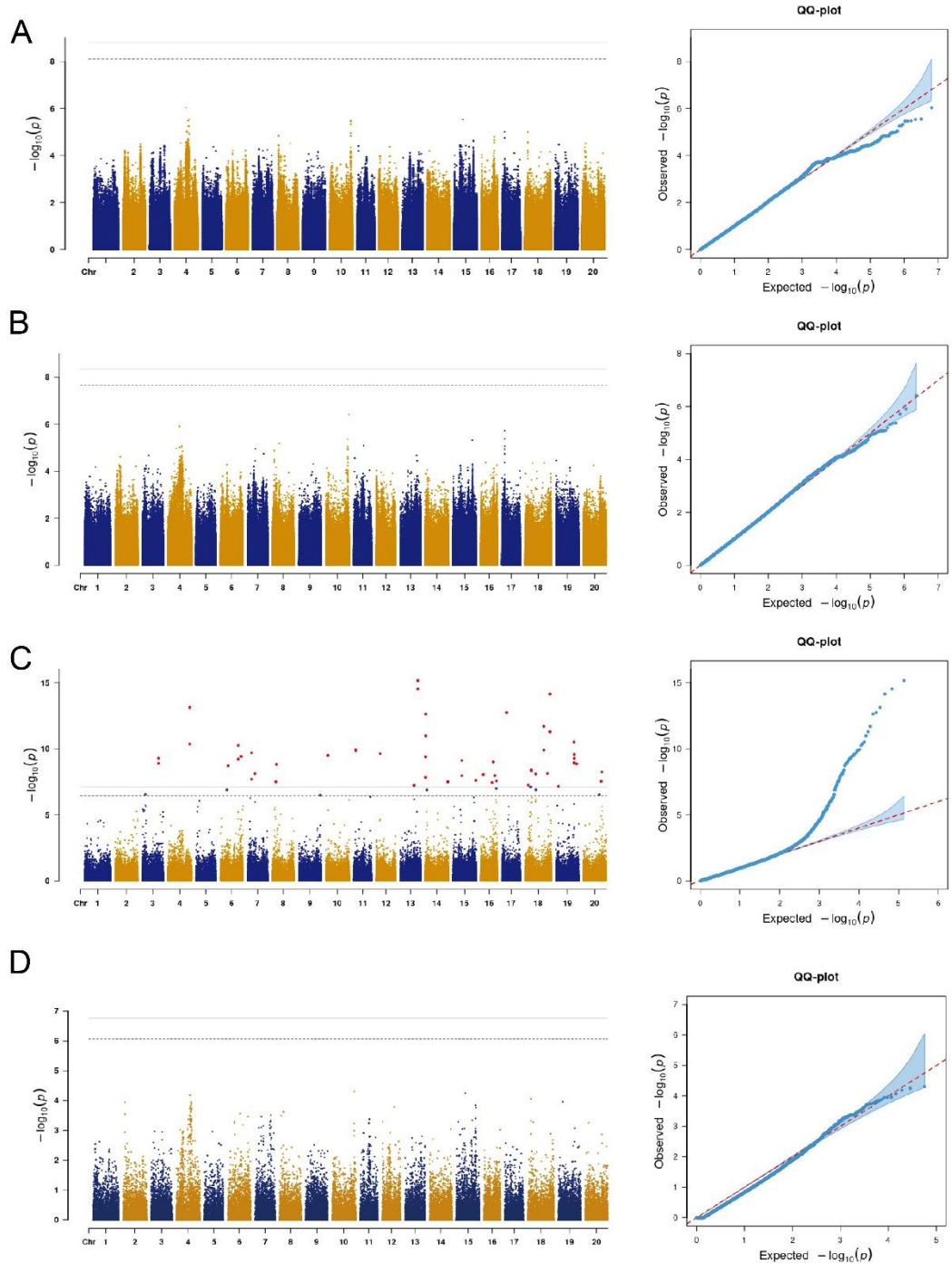
Supplementary Fig. 33. GWAS of number of nodes using EMMAX method.

Manhattan and quantile-quantile plots for number of nodes based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).

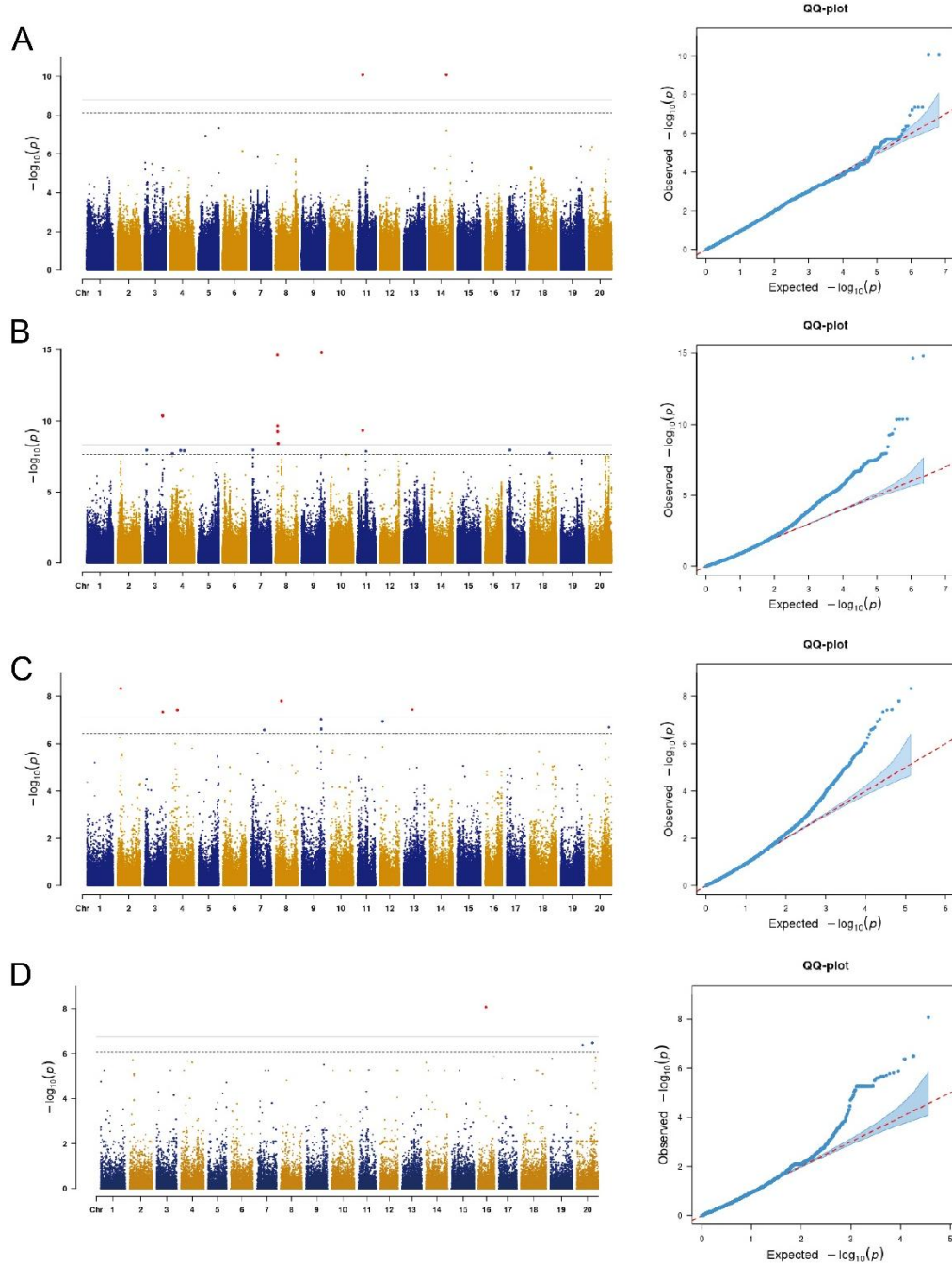


Supplementary Fig. 34. GWAS of branch numbers using EMMAX method.

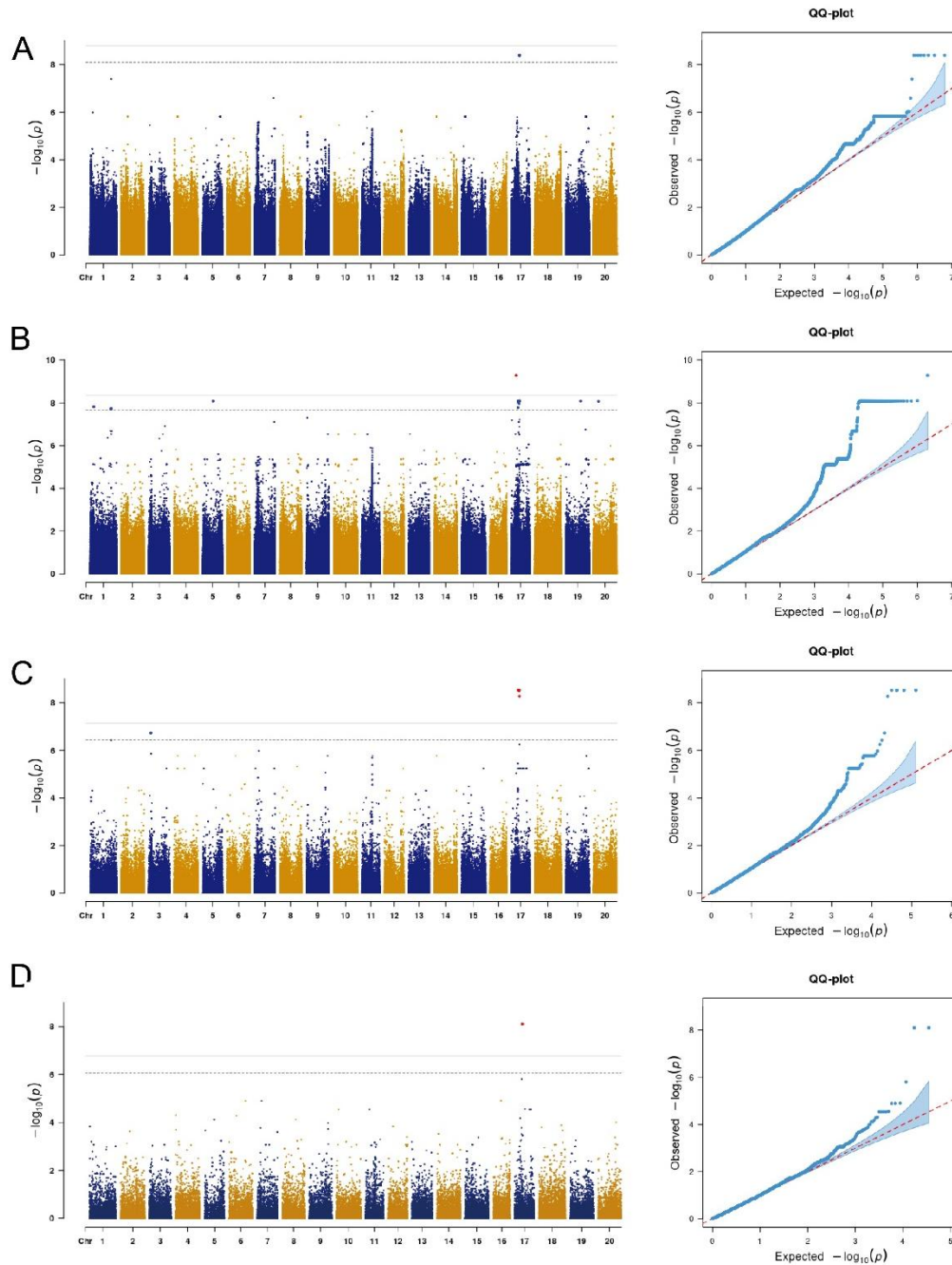
Manhattan and quantile-quantile plots for branch numbers based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



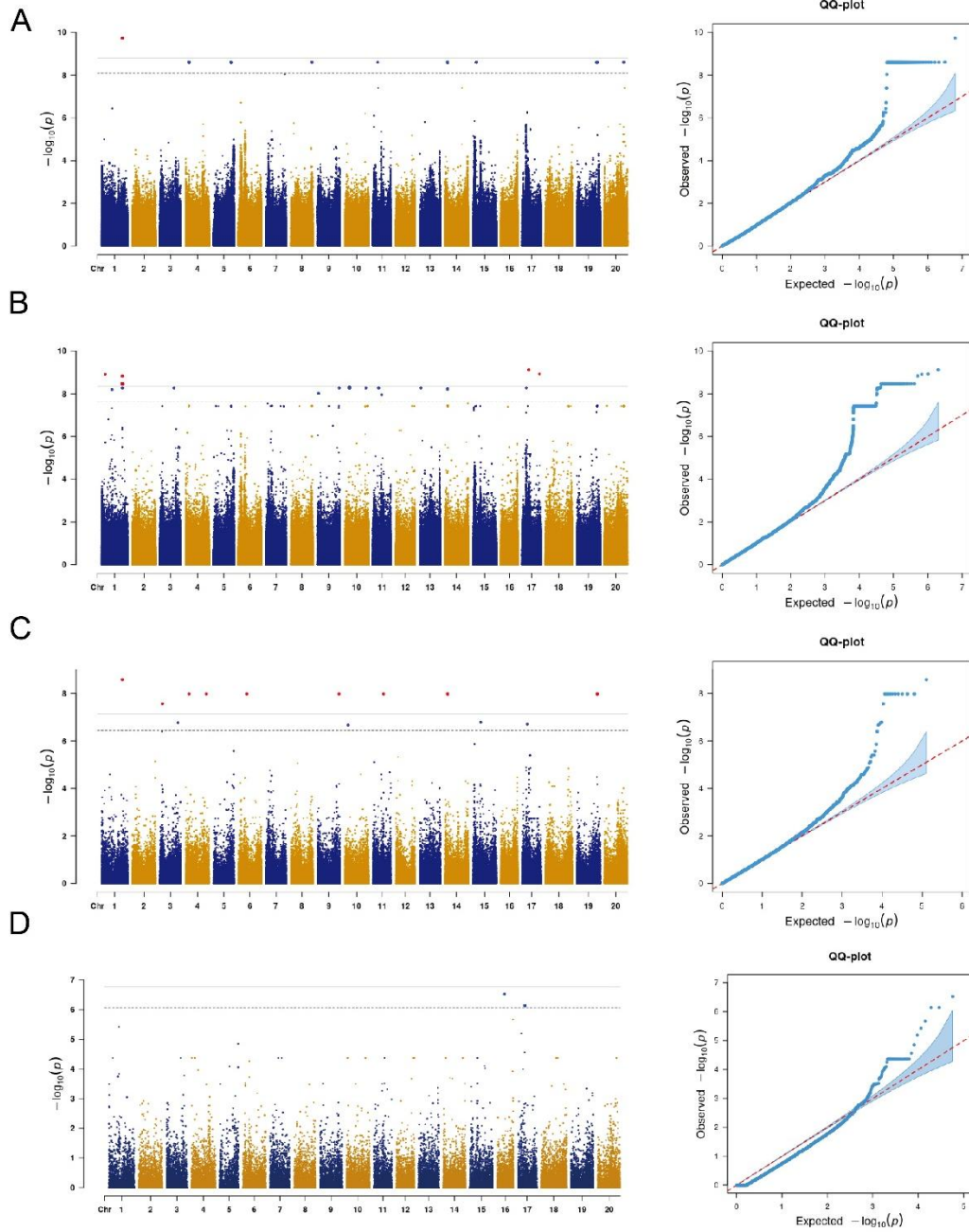
Supplementary Fig. 35. GWAS of SPAD value using EMMAX method. Manhattan and quantile-quantile plots for SPAD value based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



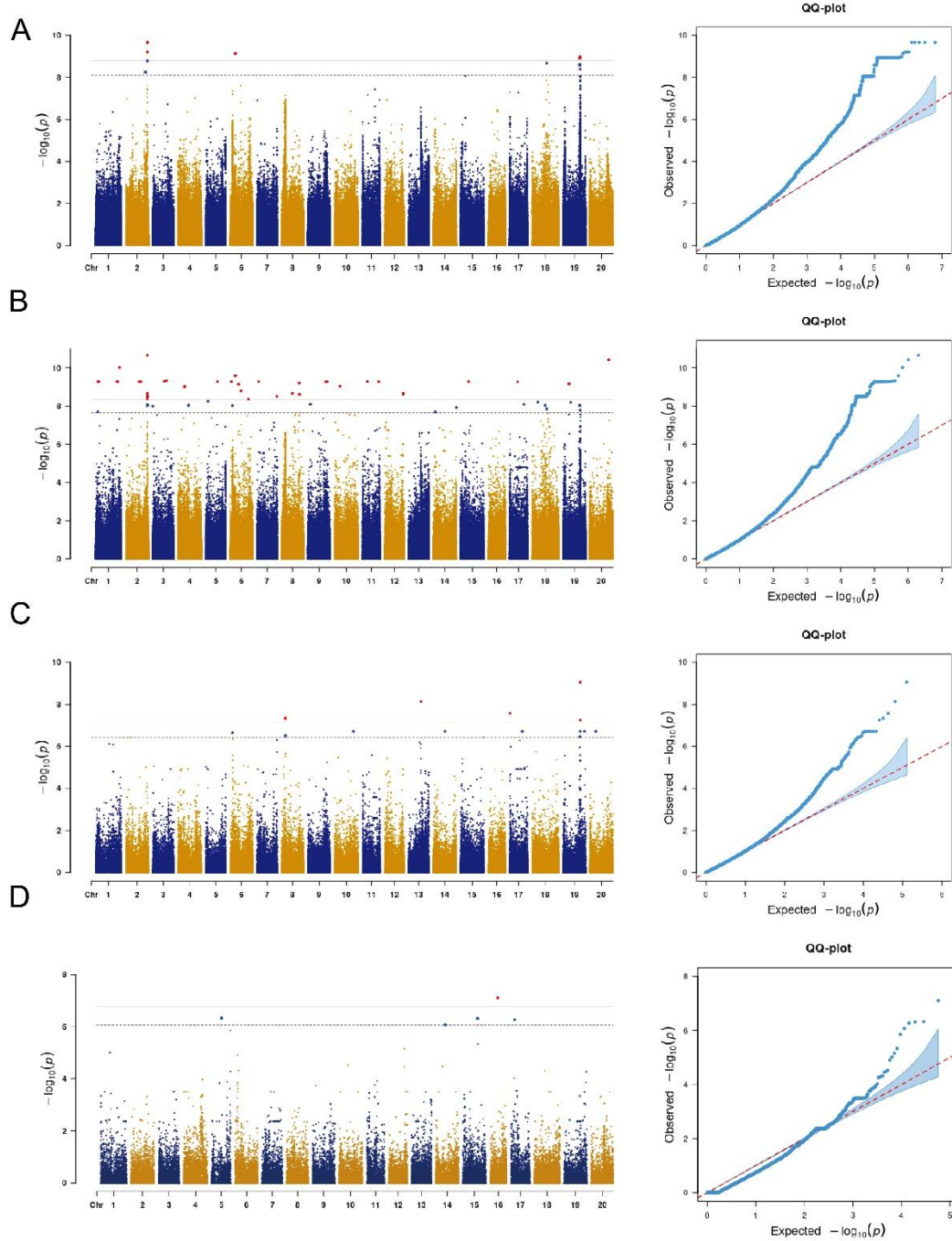
Supplementary Fig. 36. GWAS of 100-seed weight using EMMAX method. Manhattan and quantile-quantile plots for 100-seed weight based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



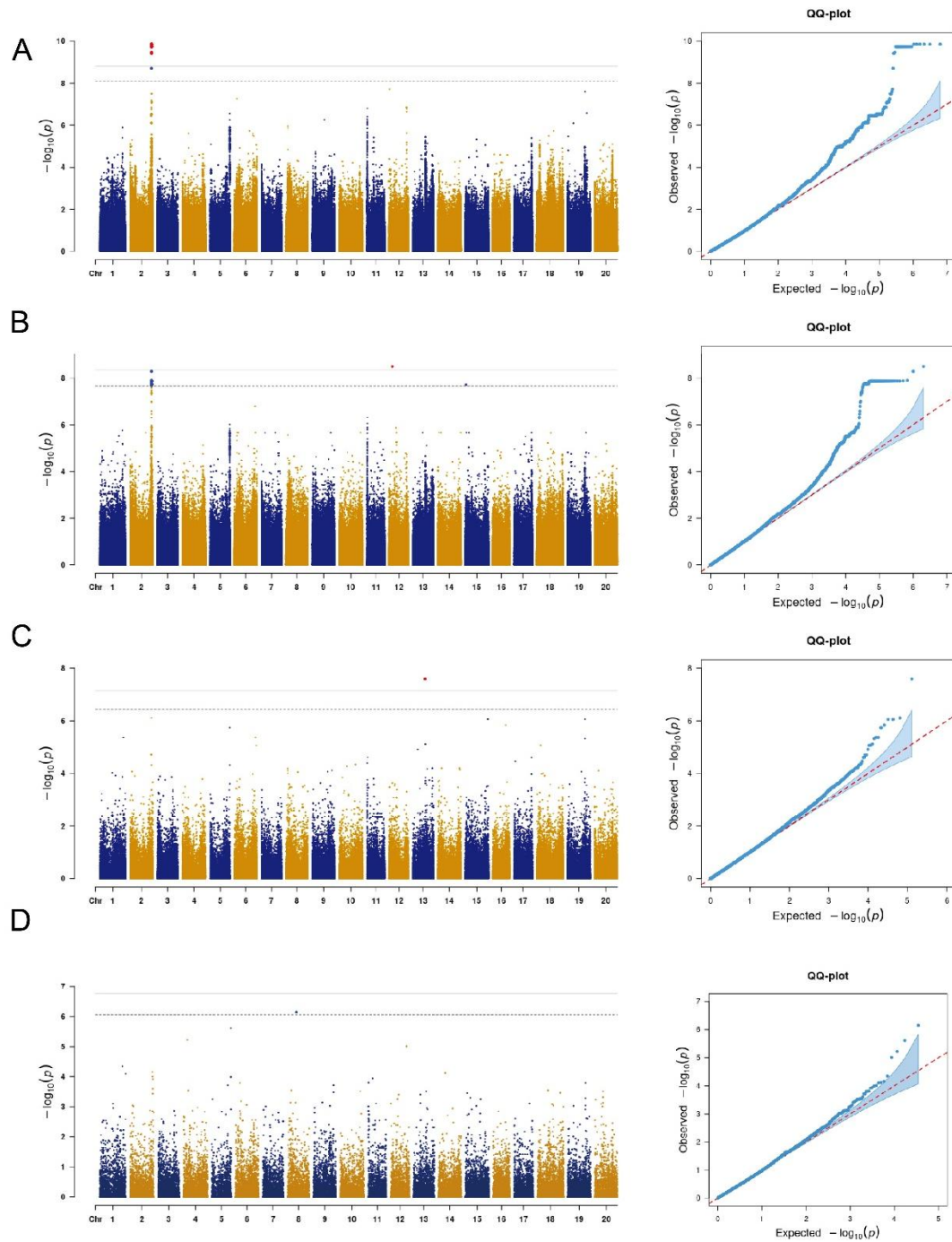
Supplementary Fig. 37. GWAS of seed length using EMMAX method. Manhattan and quantile-quantile plots for seed length based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



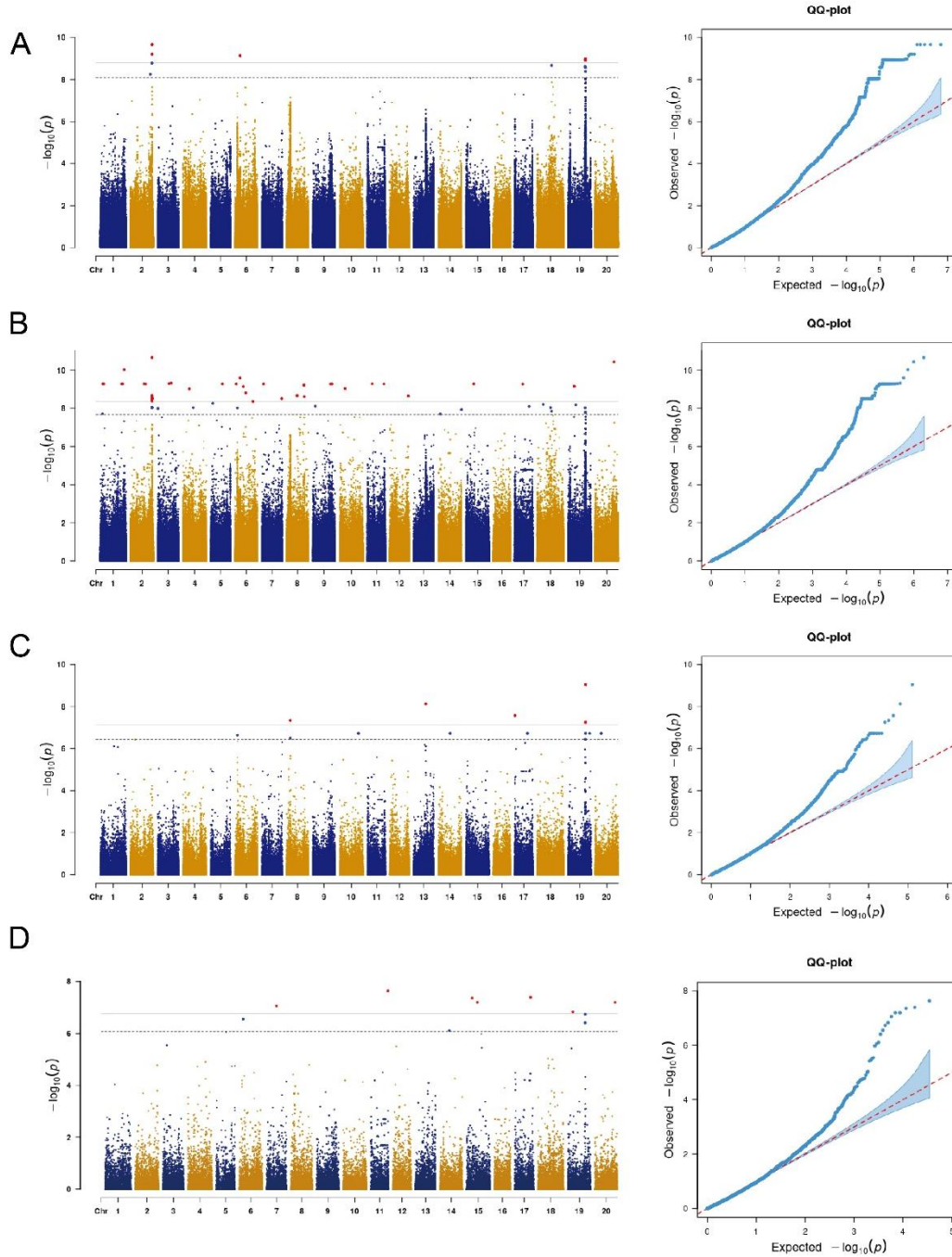
Supplementary Fig. 38. GWAS of seed width using EMMAX method. Manhattan and quantile-quantile plots for seed width based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



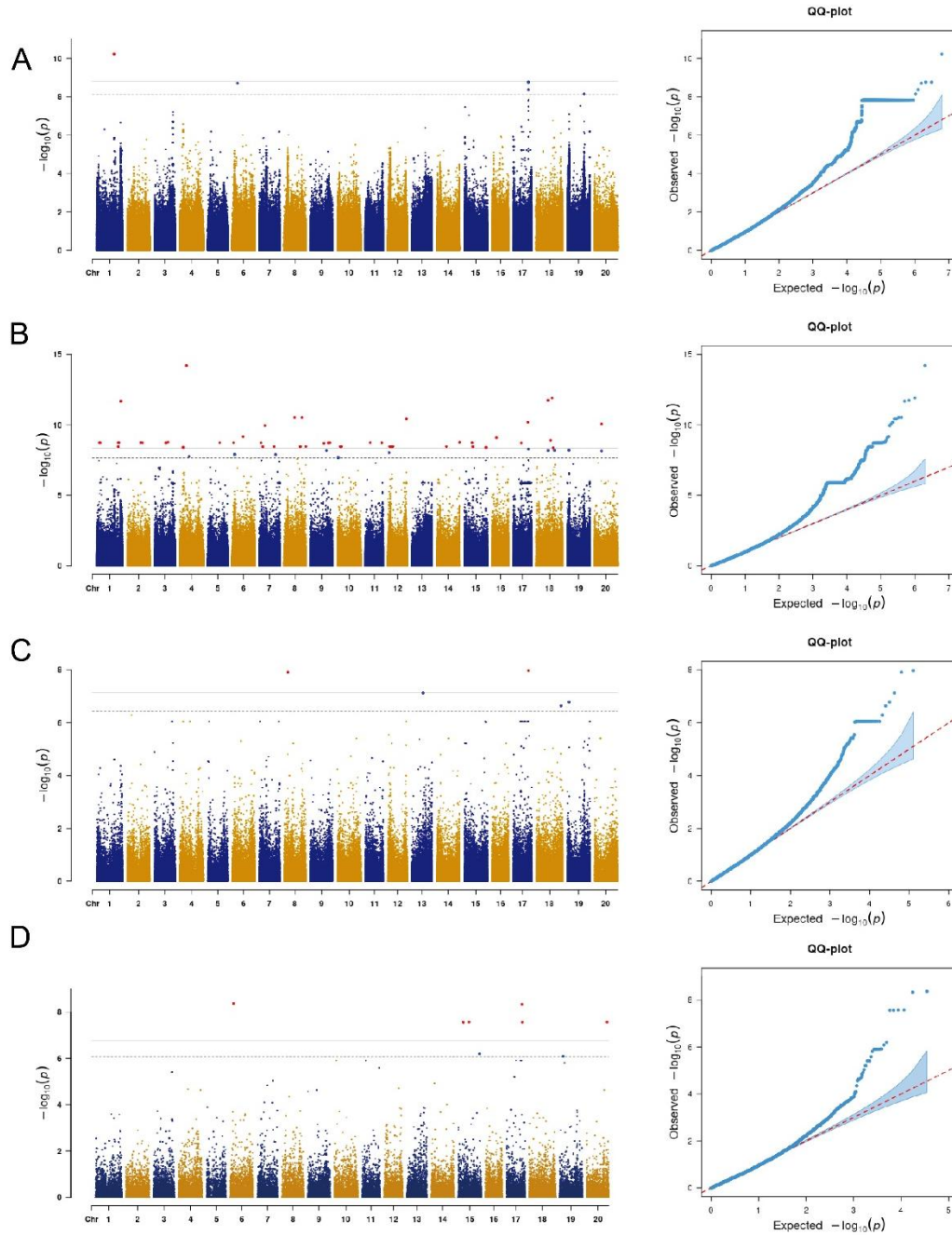
Supplementary Fig. 39. GWAS of seed thickness using EMMAX method. Manhattan and quantile-quantile plots for seed thickness based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



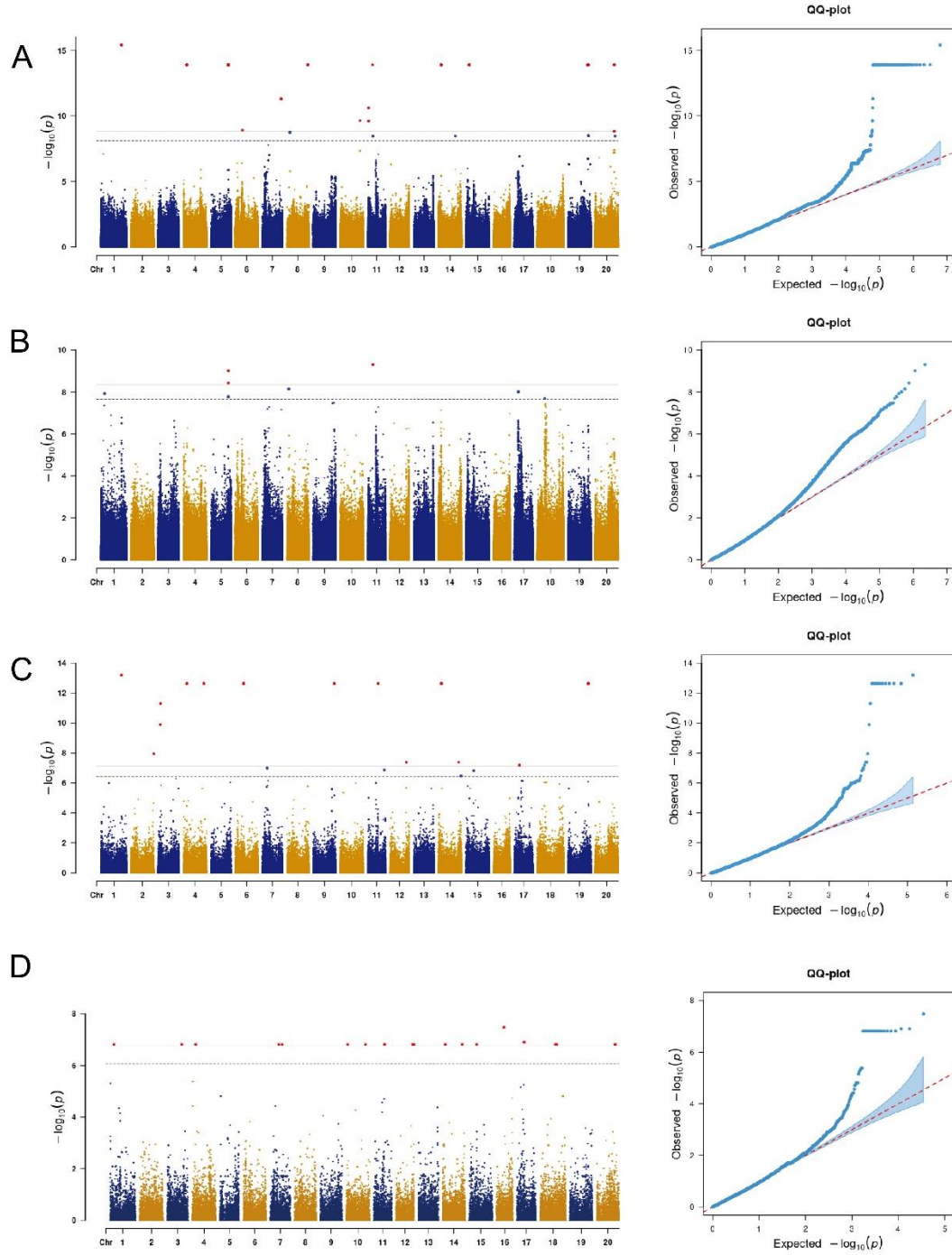
Supplementary Fig. 40. GWAS of seed length/width using EMMAX method. Manhattan and quantile-quantile plots for seed length/width based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



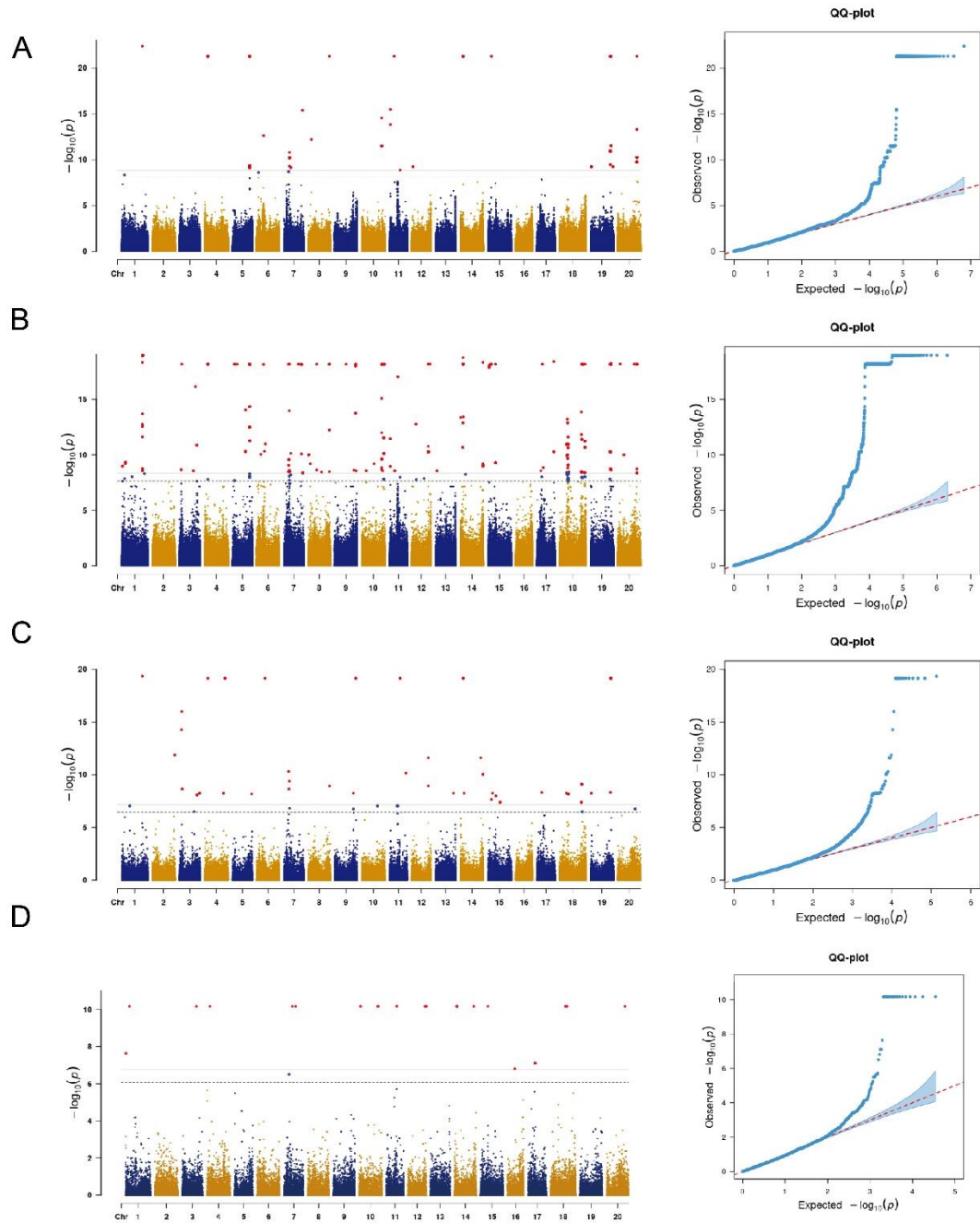
Supplementary Fig. 41. GWAS of seed length/thickness using EMMAX method. Manhattan and quantile-quantile plots for seed length/thickness based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



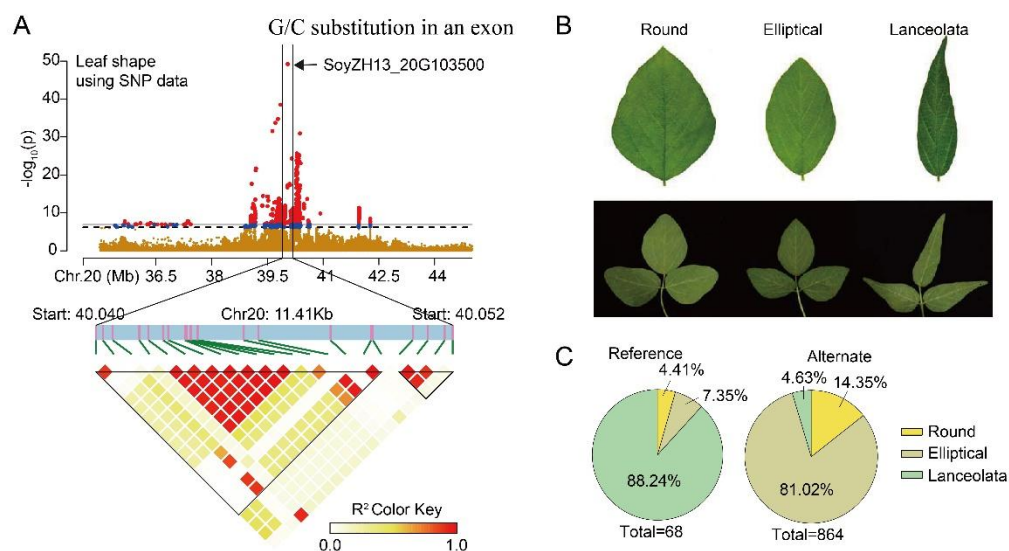
Supplementary Fig. 42. GWAS of seed width/thickness using EMMAX method. Manhattan and quantile-quantile plots for seed width/thickness based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



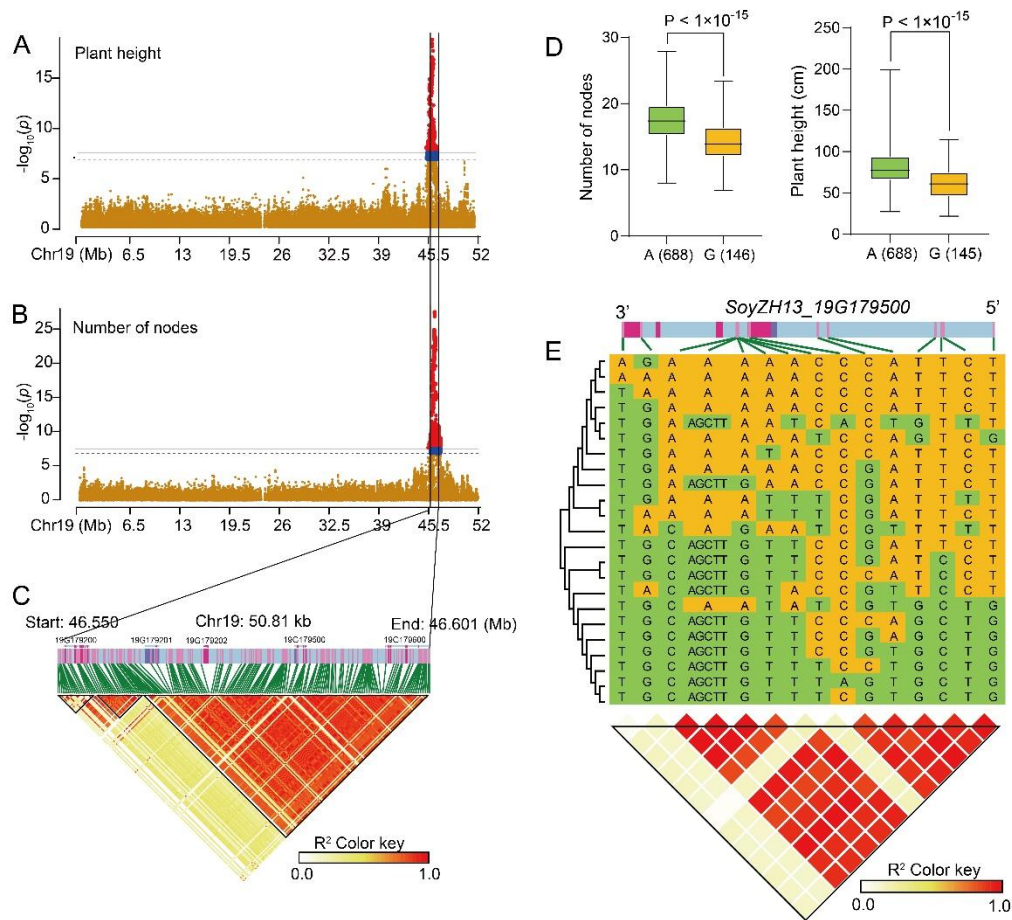
Supplementary Fig. 43. GWAS of seed area using EMMAX method. Manhattan and quantile-quantile plots for seed area based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



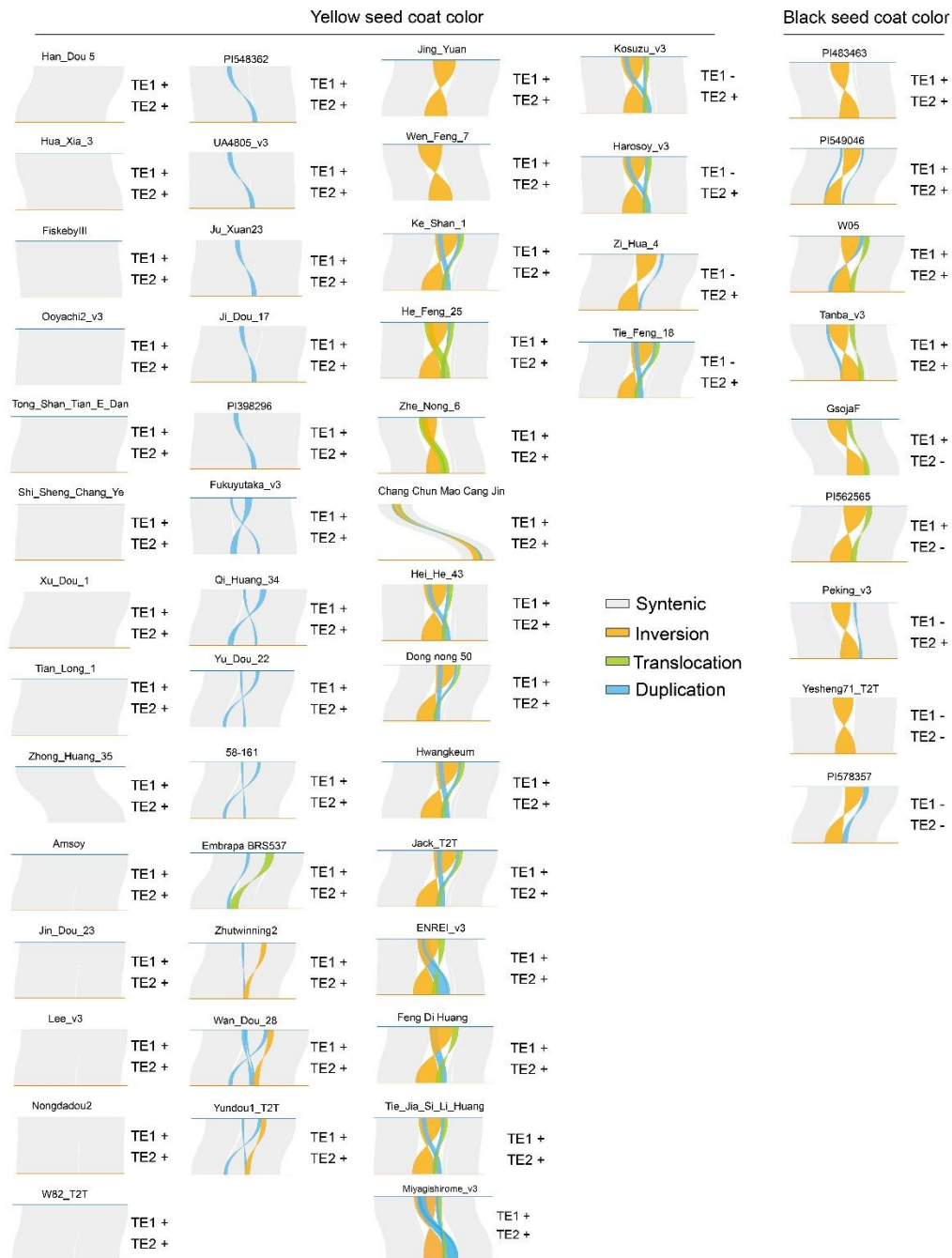
Supplementary Fig. 44. GWAS of seed volume using EMMAX method. Manhattan and quantile-quantile plots for seed volume based on SNPs (A), InDels (B), reseq-SVs (C) and PAVs (D) data. The horizontal dashed lines indicate the significance thresholds. Trait-associated variants above the threshold are highlighted in red. The thresholds were set at $-\log_{10}(0.05/n)$ and $-\log_{10}(0.01/n)$, represented by dashed and solid lines, respectively (where n is the total number of variants).



Supplementary Fig. 45. Identification and functional characterization of a candidate leaf shape locus. (A) A significant locus associated with a JAGGED-like zinc finger protein controls leaf shape. The top panel shows the local Manhattan plot from the SNP-based GWAS of leaf shape, while the bottom panel displays the LD heatmap for a 12-kb region on Chromosome 20. (B) The predominant leaf shapes are round, elliptical, and lanceolate. (C) Genotype distribution of *SoyZH13_20G103500* genes.

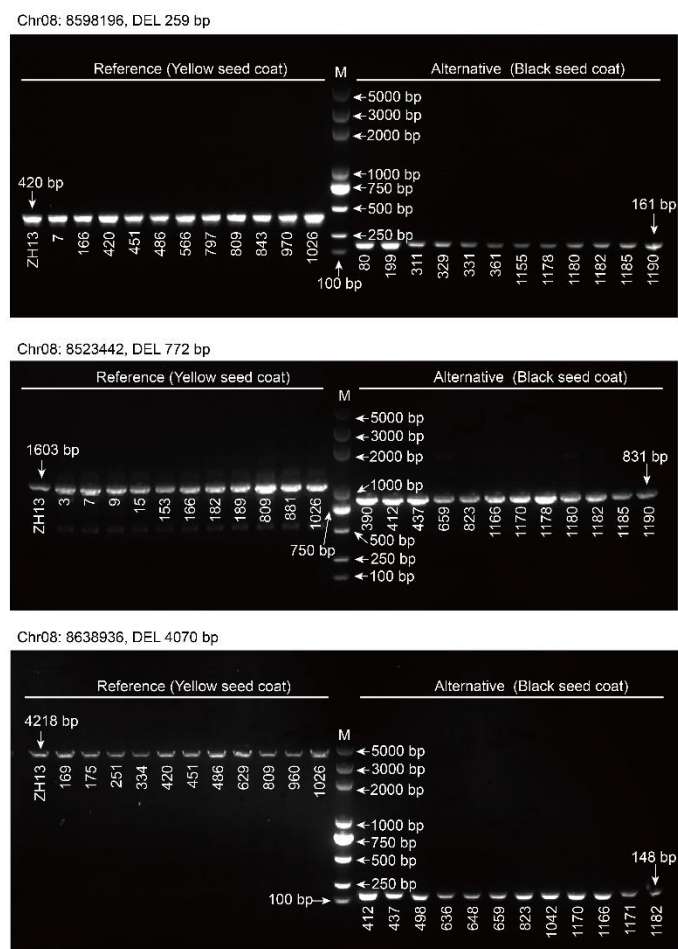


Supplementary Fig. 46. SNP-GWAS and haplotype analysis of plant height and node numbers. (A) GWAS of plant height based on SNP data using the EMMAX method. (B) GWAS of node number based on SNP data using the EMMAX method. (C) LD heatmap of a 51 kb genomic region. (D) Allelic variation of the T to A polymorphism in SoyZH13_19G179500 among 984 soybean accessions and its association with node number and plant height. Numbers in parentheses represent the sample size per group. Statistical significance was assessed using a two-tailed t -test. The box edges represent the upper and lower quartiles, with the center line indicating the median and the whiskers extending to the minimum and maximum values. (E) Haplotype analysis of SoyZH13_19G179500 using SNP and InDel data. The LD heatmap shows strong linkage in red.

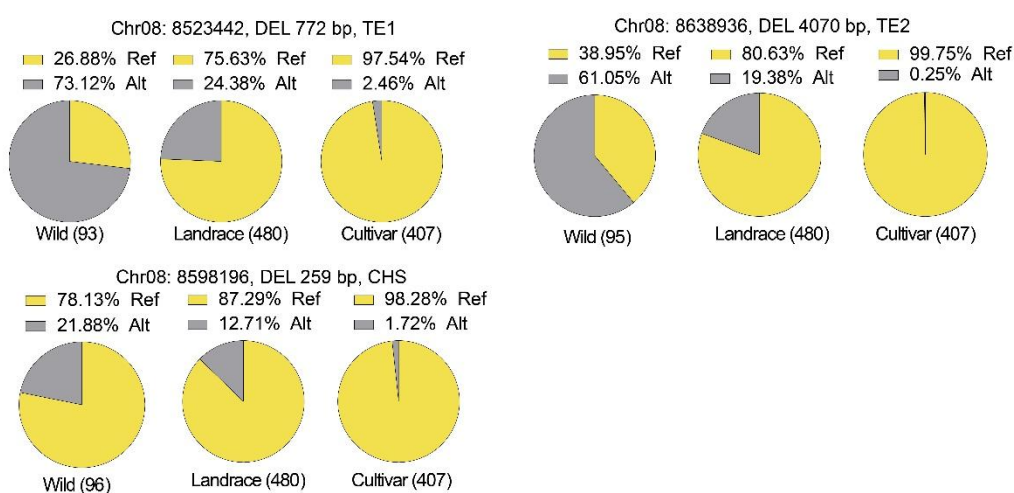


Supplementary Fig. 47. Structural variations in the CHS region revealed by pan-genome analysis. The 8.2–8.7 Mb region was analyzed to identify structural variations. The ‘+’ and ‘-’ denote the presence and absence of TE insertions, respectively.

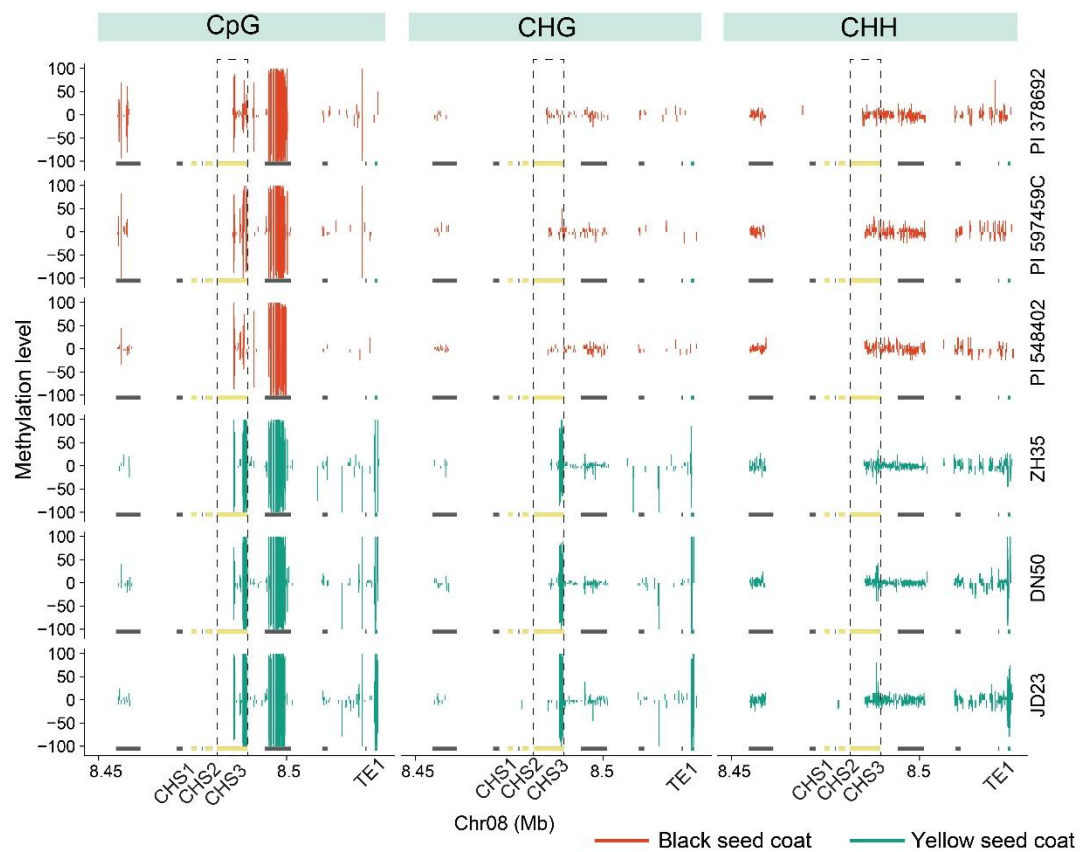
A



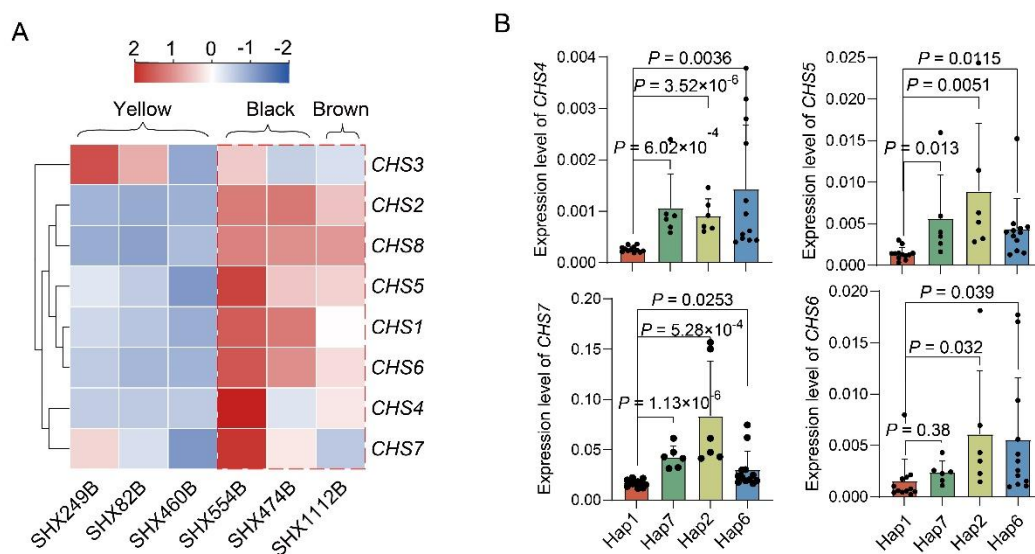
B



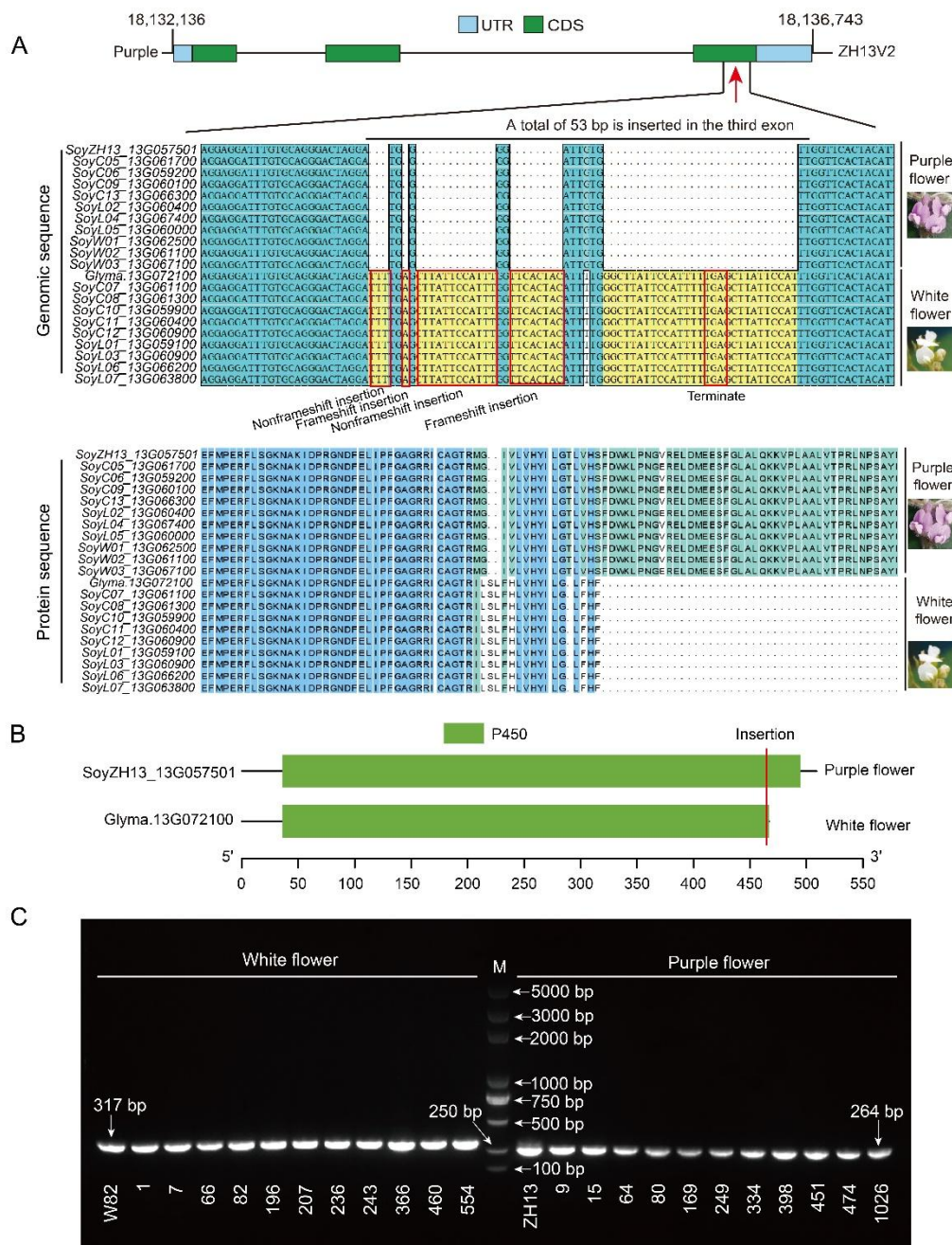
Supplementary Fig. 48. PCR verification and haplotypes of the three identified loci across wild, landrace, and cultivated soybeans. (A) PCR amplification in these sites. Germplasm IDs are provided in Supplementary Table 1. (B) Frequencies of these sites across wild soybeans, landraces, and cultivars.



Supplementary Fig. 49. Methylation profiles (CHG/CHH/CpG) in the Chr8:8.45–8.55 Mb region. The positive and negative values represent the methylation levels of the positive and negative strands.

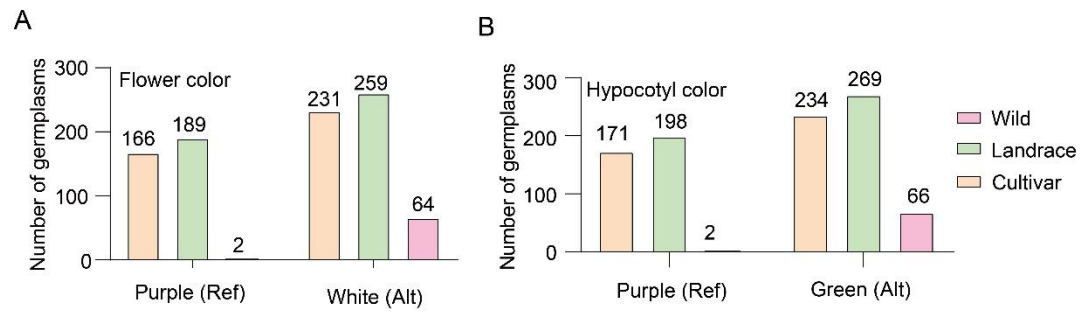


Supplementary Fig. 50. Expression level of the *CHS* gene in different soybean varieties. (A) RNA-seq. (B) RT-PCR. Data are mean $\pm$ SD ($n = 6-12$). Statistical significance was assessed using a two-tailed *t*-test.

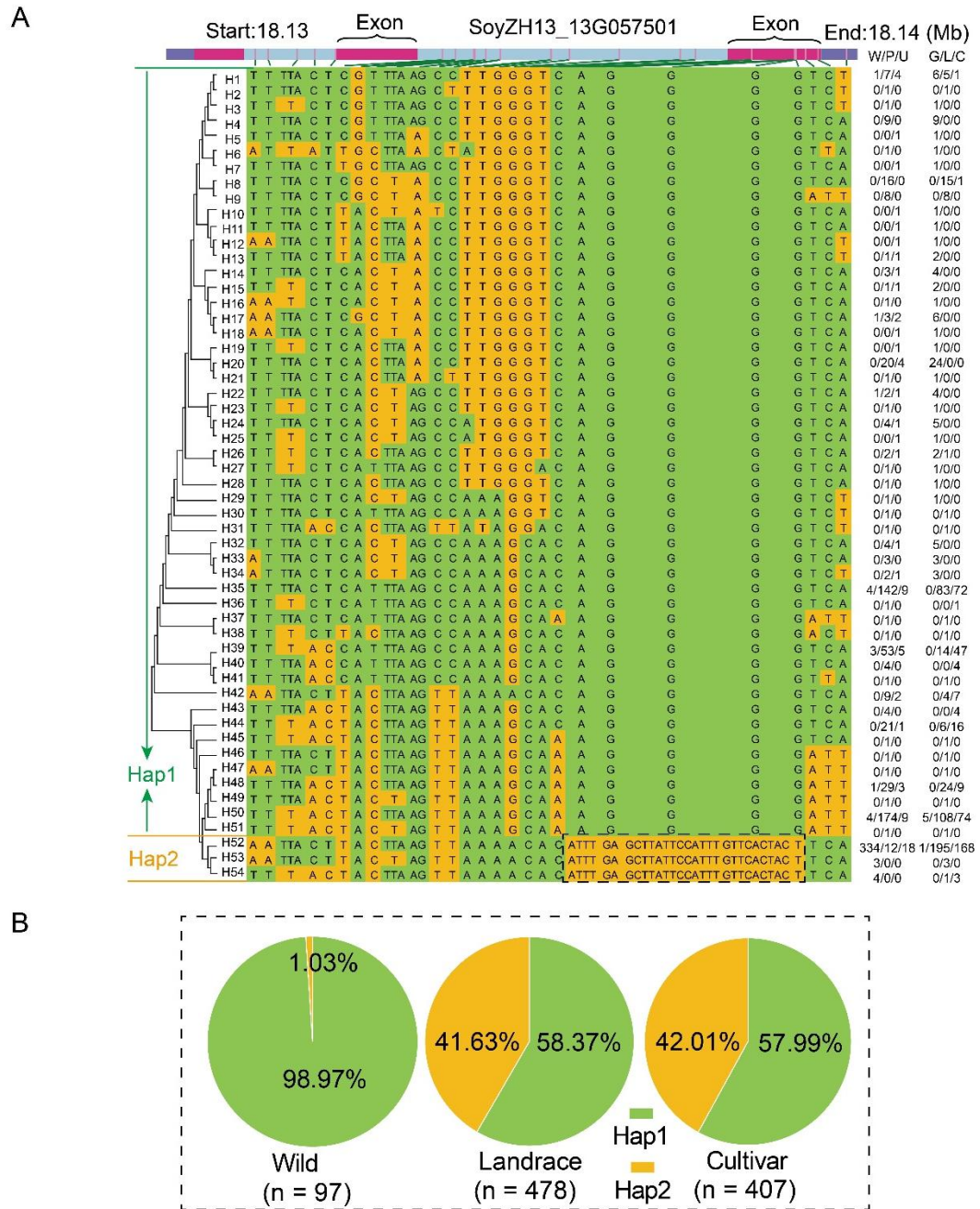


Supplementary Fig. 51. Sequence alignment and PCR verification of *SoyZH13_13G057501* gene across different germplasm accessions. (A) Alignment of genomic and amino acid sequences of *SoyZH13_13G057501* from different germplasm. The top panels show the gene structures in Zhonghuang 13 and Williams 82. The red arrow indicates a sequence insertion that results in premature termination of the amino acid sequence. (B) Domain analysis of white and purple flower variants. (C) PCR amplification of the insertion and corresponding flower colors in different

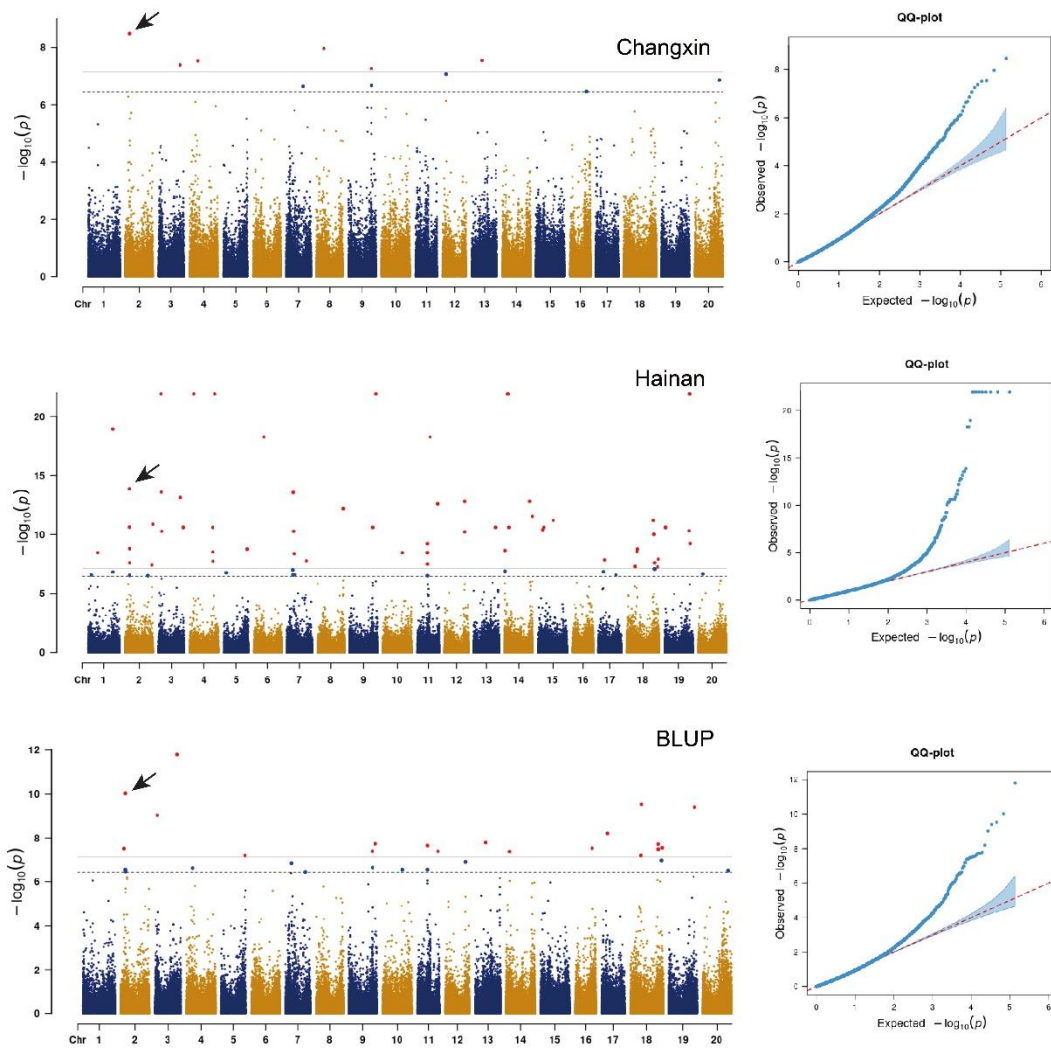
soybean accessions. The germplasm IDs are provided in Supplementary Table 1.



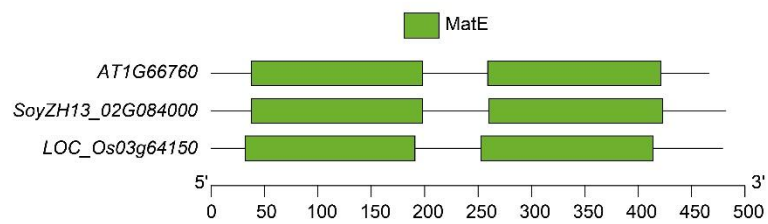
Supplementary Fig. 52. Distribution of flower and hypocotyl color phenotypes in wild, landrace, and cultivated accessions. (A) Flower color. (B) Hypocotyl flower.



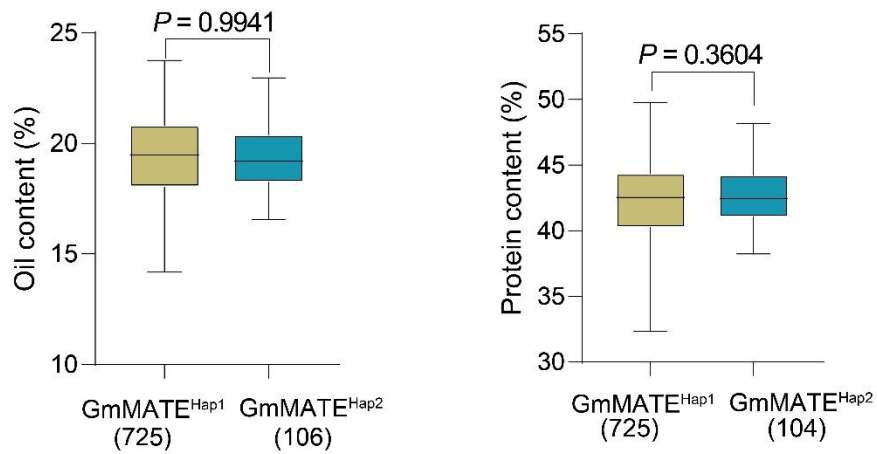
Supplementary Fig. 53. Haplotype analysis of *SoyZH13_13G057501* gene. (A) Haplotypes detected in the genomic region of *SoyZH13_13G057501* based on SNPs and InDels from 984 resequenced accessions. W, White; P, Purple; U, Unknown; G, *G.soja*; L, Landrace; C, Cultivar. (B) Frequency distribution of the two haplotypes of *SoyZH13_13G057501* in wild soybeans, landraces and cultivars.



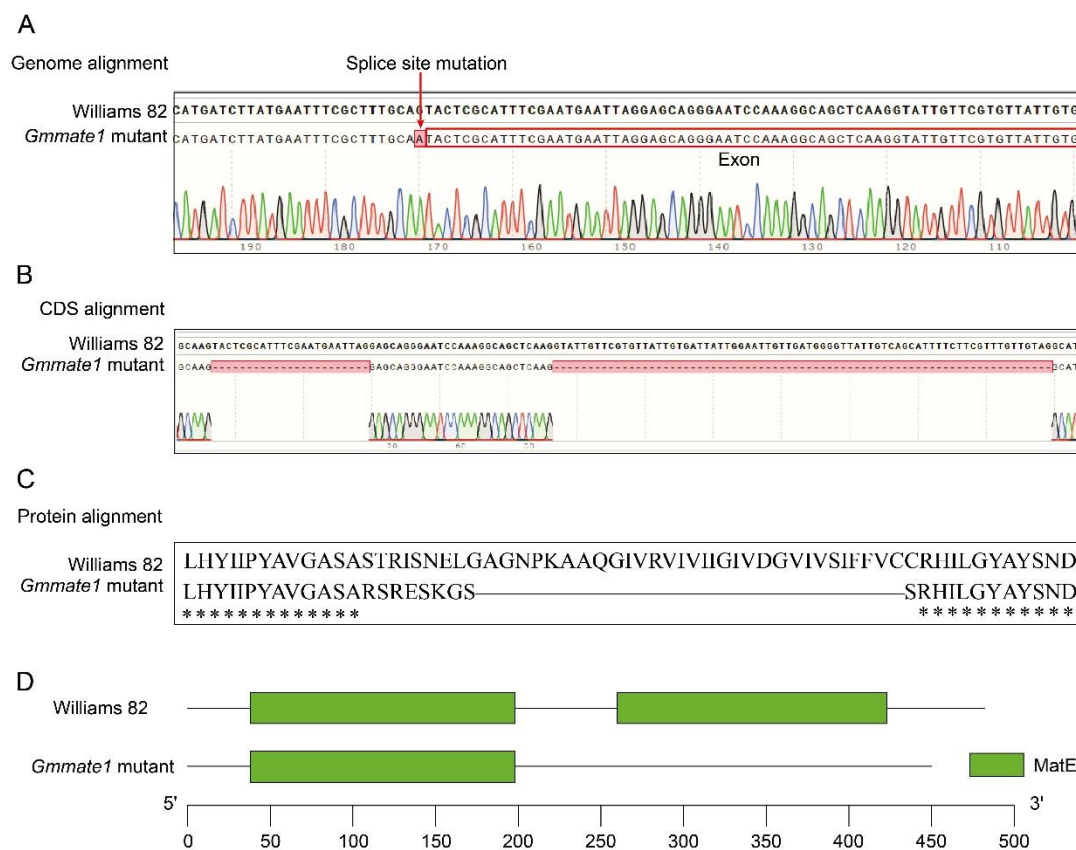
Supplementary Fig. 54. Reseq-SV-based GWAS results for 100-seed weight across multiple years and environments.



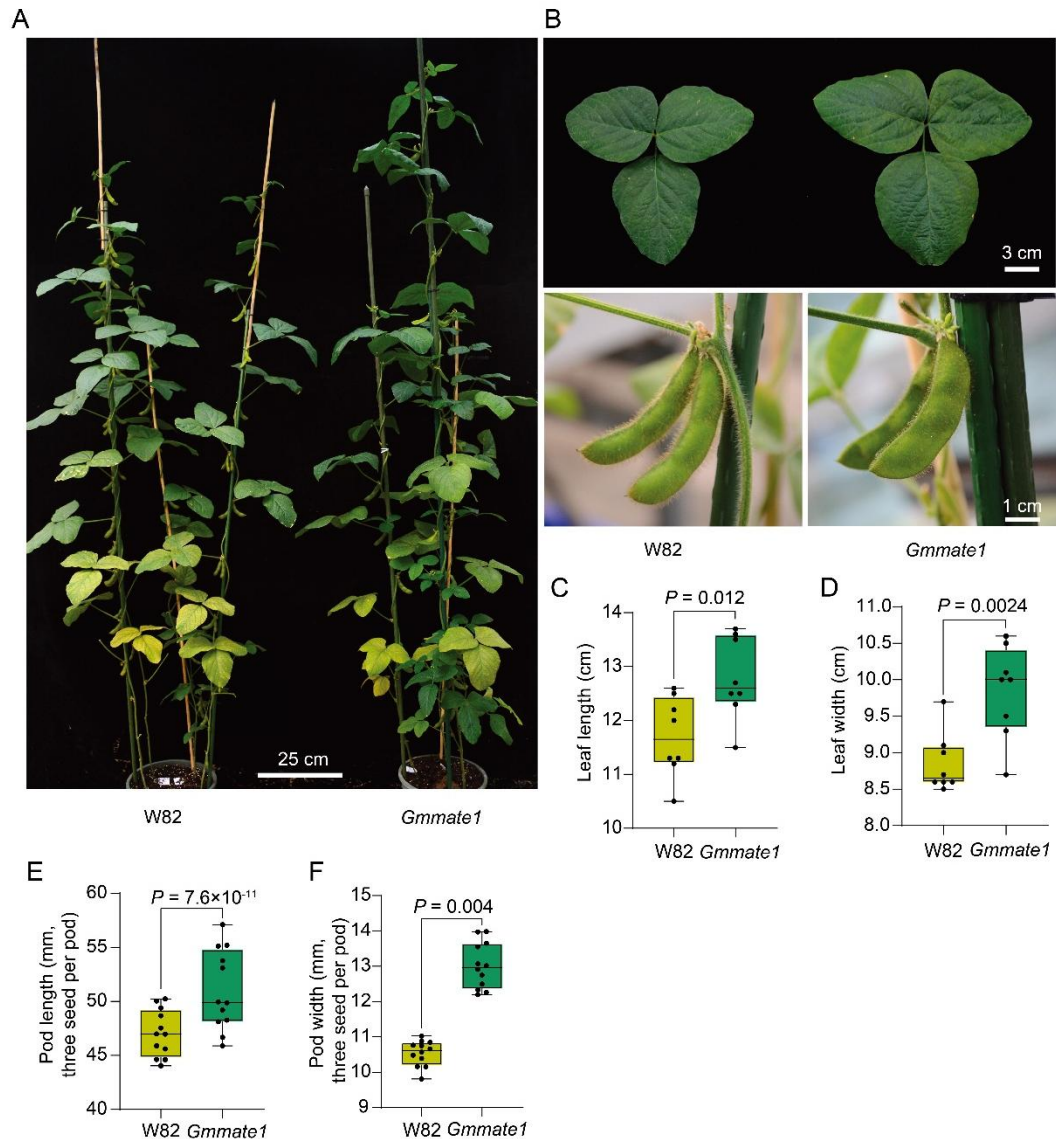
Supplementary Fig. 55. GmMATE1 protein domains. Domain analysis of GmMATE1 with homologs from Arabidopsis and rice.



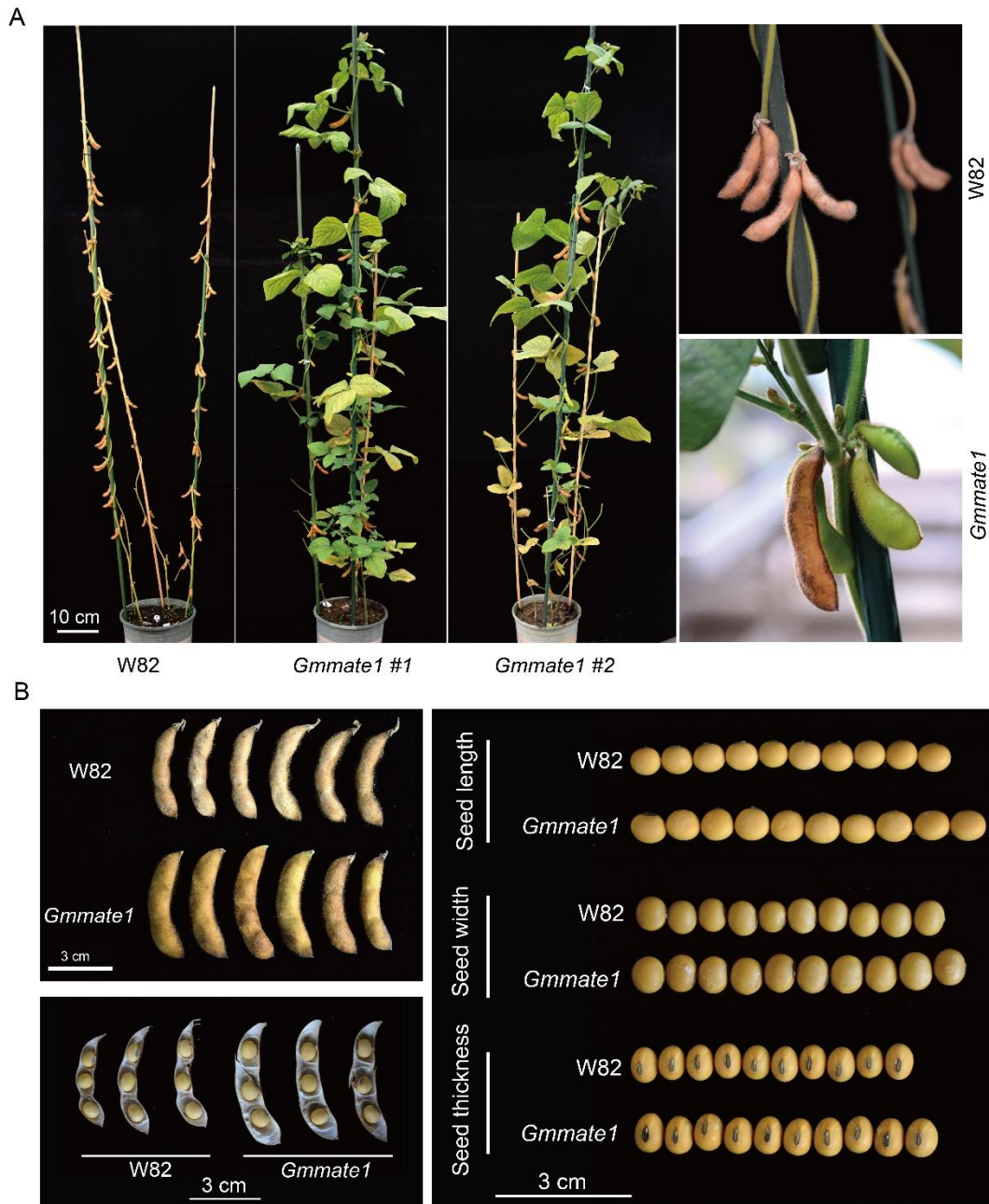
Supplementary Fig. 56. Comparison of oil and protein content between *GmMATE1* haplotypes. Numbers in parentheses represent the sample size per group. The box edges represent the upper and lower quartiles, with the center line indicating the median and the whiskers extending to the minimum and maximum values.



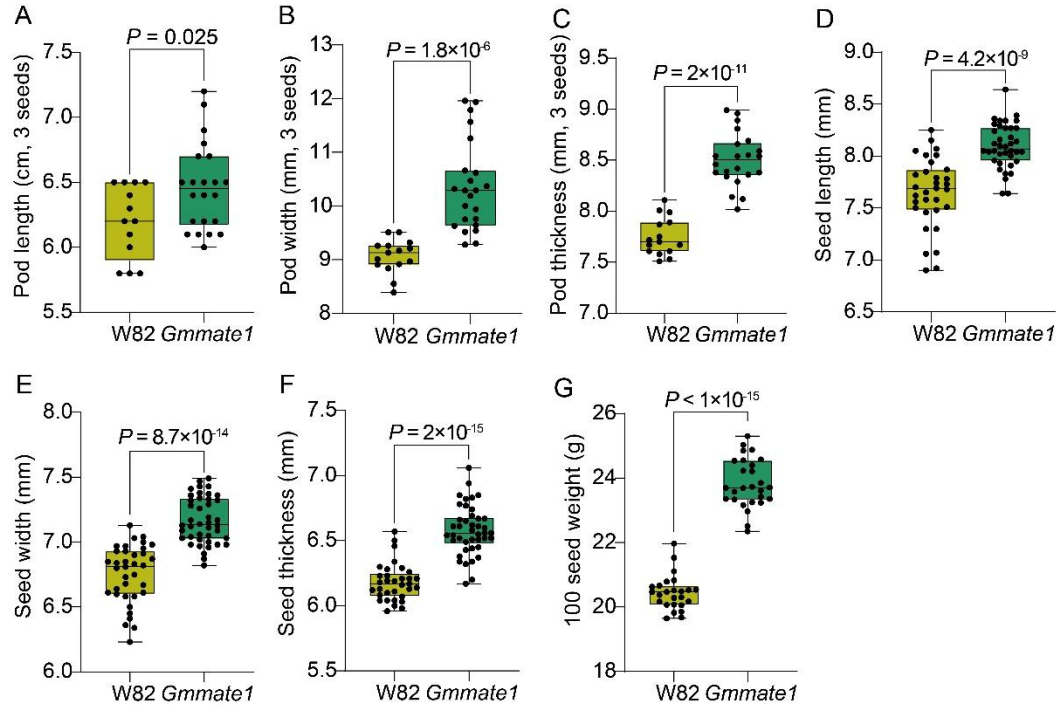
Supplementary Fig. 57. Identification of EMS-induced *Gmmate1* mutants. (A) Genomic sequence alignment. (B) Alignment of coding sequences (CDS). (C) Protein sequence alignment. (D) Comparison of protein domains between the *Gmmate1* mutant and wild-type (Williams 82).



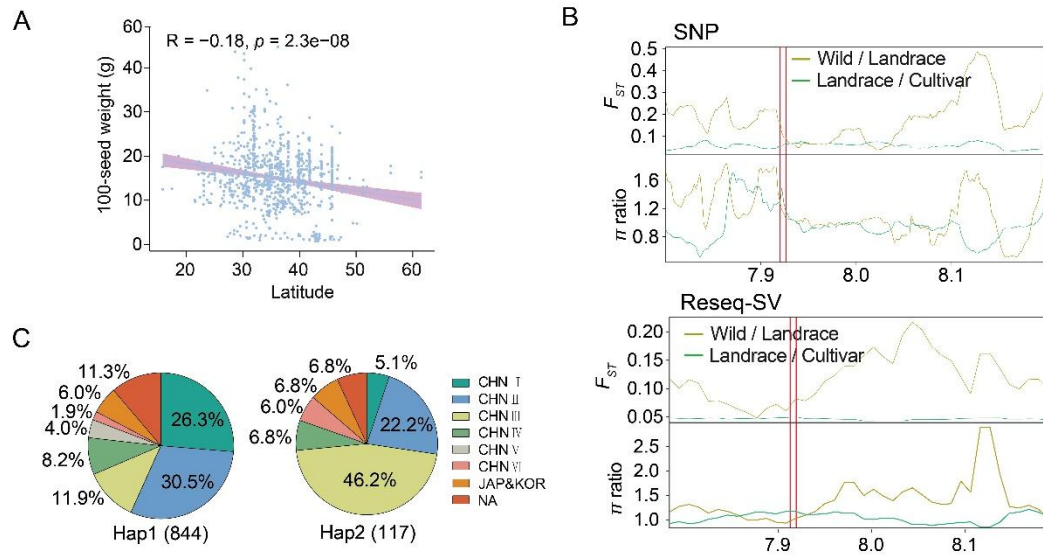
Supplementary Fig. 58. Phenotype of Gmmate1 mutant plants in 7 weeks. (A) Whole-plant phenotype after 7 weeks in a greenhouse. (B) Phenotypes of trifoliate leaves and pods. (C-D) Measurement of leaf length and leaf width ($n = 8$). (E-F) length and width of three-seeded pods ($n = 12$). Statistical significance was assessed using a two-tailed t -test.



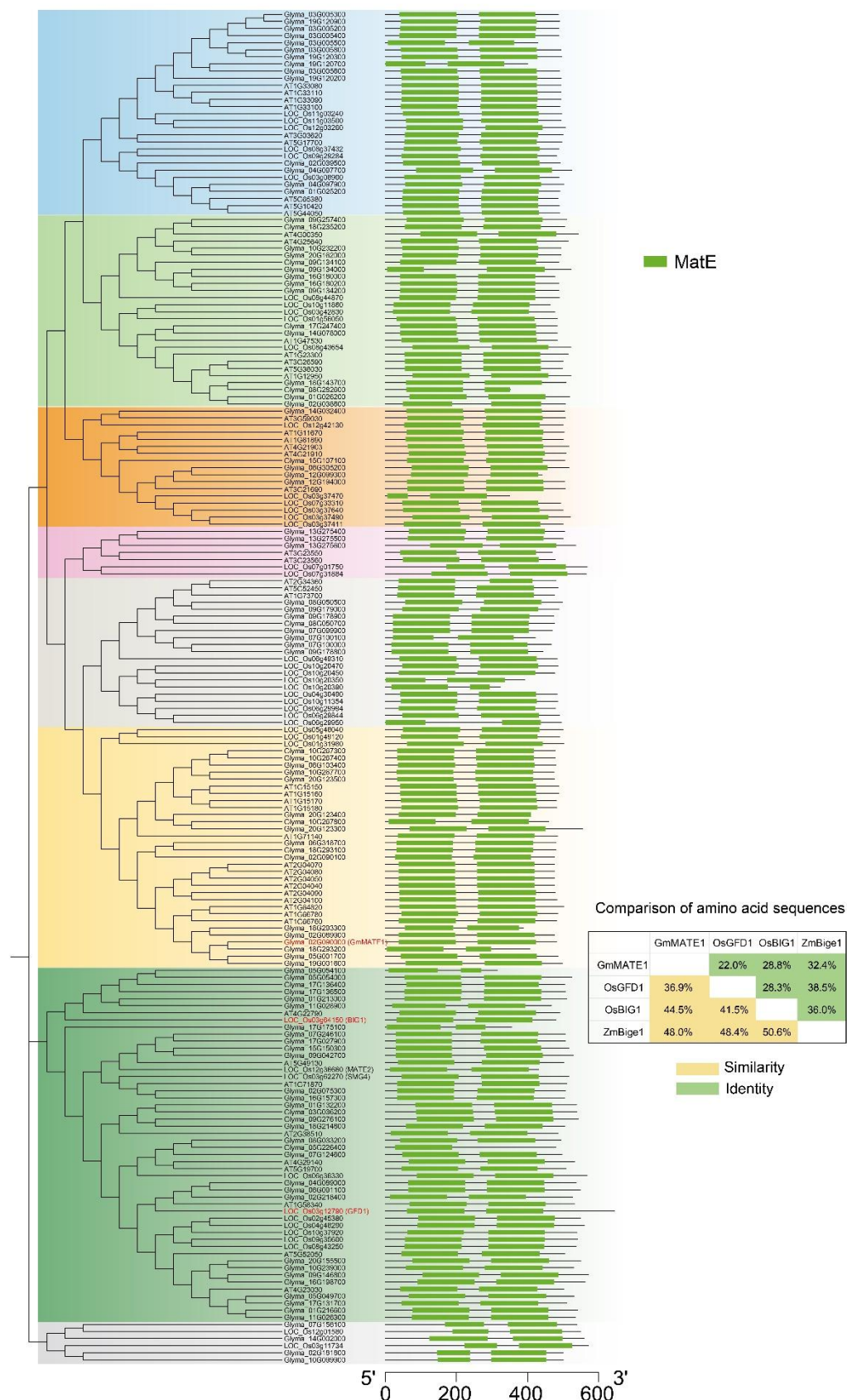
Supplementary Fig. 59. Phenotype of *Gmmate1* mutant plants in 14 weeks. (A) Whole-plant phenotype after 14 weeks in a greenhouse. (B) Pod and seed phenotypes of W82 and the *Gmmate1* mutant.



Supplementary Fig. 60. Pod and seed traits of Gmmate1 mutant plants. (A-C) Length, width and thickness of three-seeded pods ($n = 13-23$). (D-F) Length, width and thickness of seeds ($n = 32-42$). (G) 100-seed weight ($n = 25-27$). Statistical significance was assessed using a two-tailed t -test.

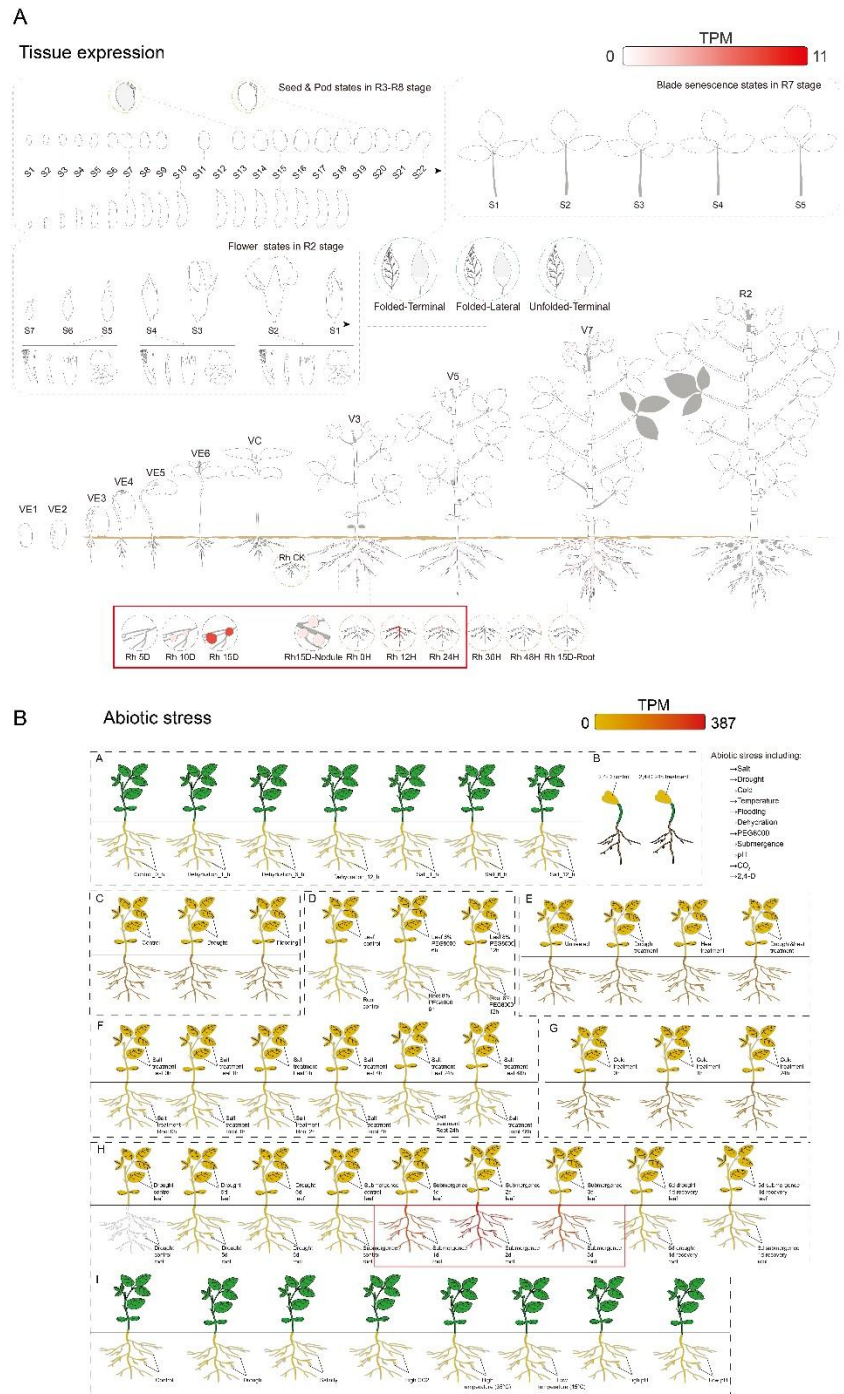


Supplementary Fig. 61. Selection signals and geographical distribution of GmMATE1. (A) Correlation between 100-seed weight and latitude. (B) F_{ST} and π ratio calculated from SNP and reseq-SV data comparing wild, landrace, and cultivated soybean groups. (C) Distribution of distinct GmMATE1 haplotypes across ecological regions.



Supplementary Fig. 62. The *MATE* gene family in soybean, rice, and *Arabidopsis*

thaliana.



Supplementary Fig. 63. Expression analysis of *GmMATE1*. (A) Tissue-specific expression profile of *GmMATE1* in various soybean tissues. (B) Expression levels of *GmMATE1* under different abiotic stress conditions. The expression diagrams of different tissues were generated using SoyOD and SoyOmics website.

Supplementary Table 1. Number of reseq-SVs across length ranges and types.

Region	DEL	DUP	INV	INS	TRA	Total
50-100 bp	47702	3	27667	32	-	75404
100-200 bp	20577	760	49644	17	-	70998
200-300 bp	8584	1474	25463	5	-	35526
300-400 bp	6286	1248	5809	2	-	13345
400-500 bp	5706	1162	3293	1	-	10162
500-600 bp	4390	961	2242	0	-	7593
600 bp-1 KB	9416	2470	3269	0	-	15155
1-2 kb	10103	3070	2752	0	-	15925
2-5 kb	12041	4106	2782	0	-	18929
5-10 kb	8338	2761	2201	0	-	13300
10-100 kb	11190	6621	6726	0	-	24537
100 kb-1 M	2033	1929	6920	0	-	10882
1 M-10 M	1093	1095	6490	0	-	8678
Total	147459	27660	145258	57	281847	602281

Notes: “-” indicates that the reseq-SV length cannot be determined.

Supplementary Table 2. Transcriptome-based expression profiles of the *CHS* gene cluster.

Gene ID	Gene symbol	Description	SHX460B	SHX249B	SHX82B	SHX366B	SHX554B	SHX474B	SHX1112B
			Yellow			Black			Brown
<i>SoyZH13_08G103000</i>	<i>CHS1</i>	chalcone synthase 5	1.41	2.49	2.02	1.69	7.04	6.35	3.53
<i>SoyZH13_08G103101</i>	<i>CHS2</i>	chalcone synthase 3	0.67	0.69	0.63	0.95	1.84	1.86	1.46
<i>SoyZH13_08G103102</i>	<i>CHS3</i>	chalcone synthase							
		CHS4	0.5	1.92	1.42	0.47	1.28	0.75	0.84
<i>SoyZH13_08G103702</i>	<i>TE1</i>	Cytochrome P450							
		71D10	2.09	2.37	1.89	7.87	1.6	3.08	6.56
<i>SoyZH13_08G104300</i>	<i>CHS4</i>	chalcone synthase 2	0.24	0.29	0.24	0.61	4.54	0.71	1.49
<i>SoyZH13_08G104700</i>	<i>CHS5</i>	chalcone synthase 3	0.74	1.67	1.36	1.21	3.81	2.49	2.37
		chalcone synthase							
<i>SoyZH13_08G104701</i>	<i>CHS6</i>	CHS4	0.93	1.42	1.01	1.22	5.35	4.34	3.01
<i>SoyZH13_08G104702</i>	<i>CHS7</i>	chalcone synthase 1	0.56	1.11	0.85	0.84	1.66	1.05	0.76
<i>SoyZH13_08G104703</i>	<i>CHS8</i>	chalcone synthase 3	0.66	0.54	0.46	0.88	1.89	1.8	1.75
		methyletransferase-							
<i>SoyZH13_08G105500</i>	<i>TE2</i>	like protein 13	60.39	88.09	84.38	58.64	37.91	36.88	33.31

Note: Samples were collected at 16–18 days past flowering. Values represent transcripts per million reads (TPM). Seeds were sampled at 14–16 days post flowering.

Supplementary Table 3. Sample information for McrBC-qPCR analysis.

Simple number	ID	Seed coat color	TE1	DEL	TE2
64	SHX64C	Yellow	Ref	Ref	Ref
82	SHX82B	Yellow	Ref	Ref	Ref
166	SHX166B	Yellow	Ref	Ref	Ref
366	SHX366B	Yellow	Ref	Ref	Ref
486	SHX486B	Yellow	Ref	Ref	Ref
809	SHX809B	Yellow	Ref	Ref	Ref
1166	SHX1166B	Black	Alt	Ref	Alt
1171	SHX1171B	Black	Alt	Ref	Alt
1176	SHX1176B	Grey	Alt	Alt	Ref
1172	SHX1197B	Black	Alt	Ref	Alt
1178	SHX1178B	Black	Alt	Alt	Ref
1190	SHX1190B	Black	Alt	Alt	Ref

Note: "Ref" and "Alt" indicate the presence and absence of the insertion, respectively.