

Time Series Canopy Phenotyping Enables the Identification of Genetic Variants Controlling Dynamic Phenotypes in Soybean

Delin Li^{1†}, Dong Bai^{1†}, Yu Tian^{1†}, Ying-hui Li^{1*}, Chaoseng Zhao², Qi Wang^{1,3}, Shiyu Guo^{1,3}, Yongzhe Gu¹, Xiaoyan Luan⁴, Ruizhen Wang², Jinliang Yang⁵, Malcolm J. Hawkesford⁶, James C. Schnable^{5*}, Xiuliang Jin^{1*}, Li-juan Qiu^{1*}

SUPPORTING INFORMATION

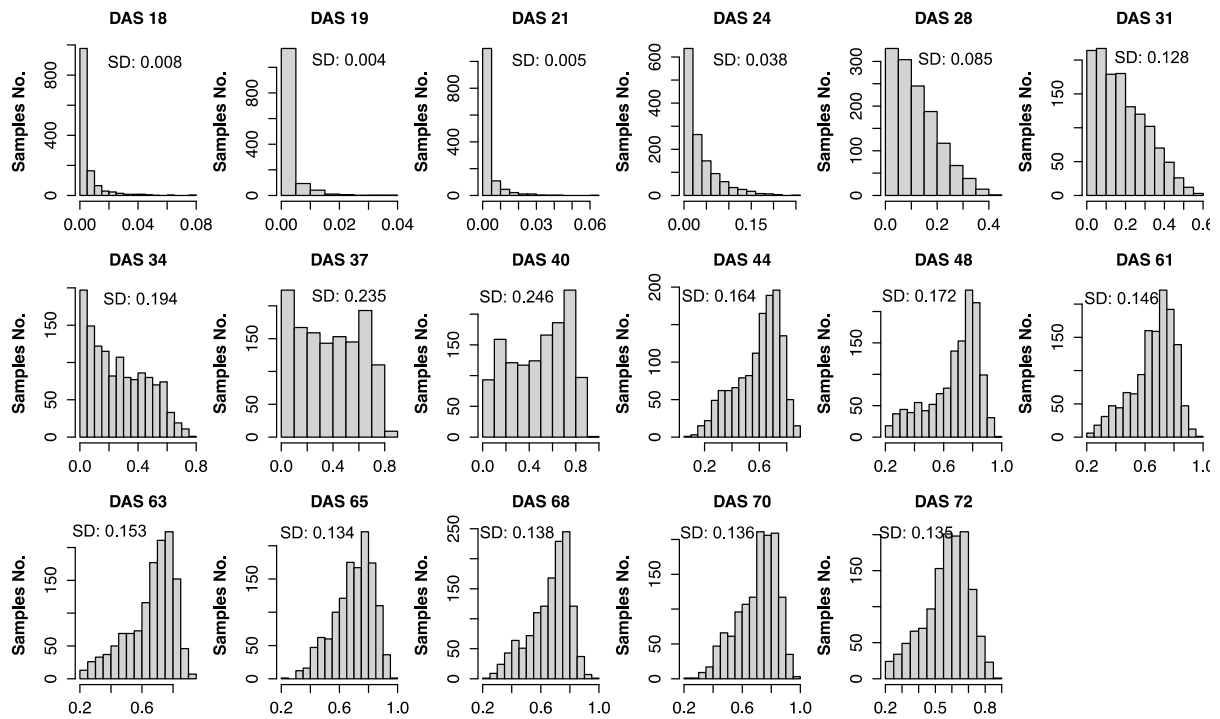


Figure S1. Distribution of canopy coverage for the entire population for each of the 17 time points in Nanchang

DAS, days after sowing. The x-axis shows the canopy coverage, and the y-axis shows the number of soybean accessions. *SD*, standard deviation for canopy coverage.

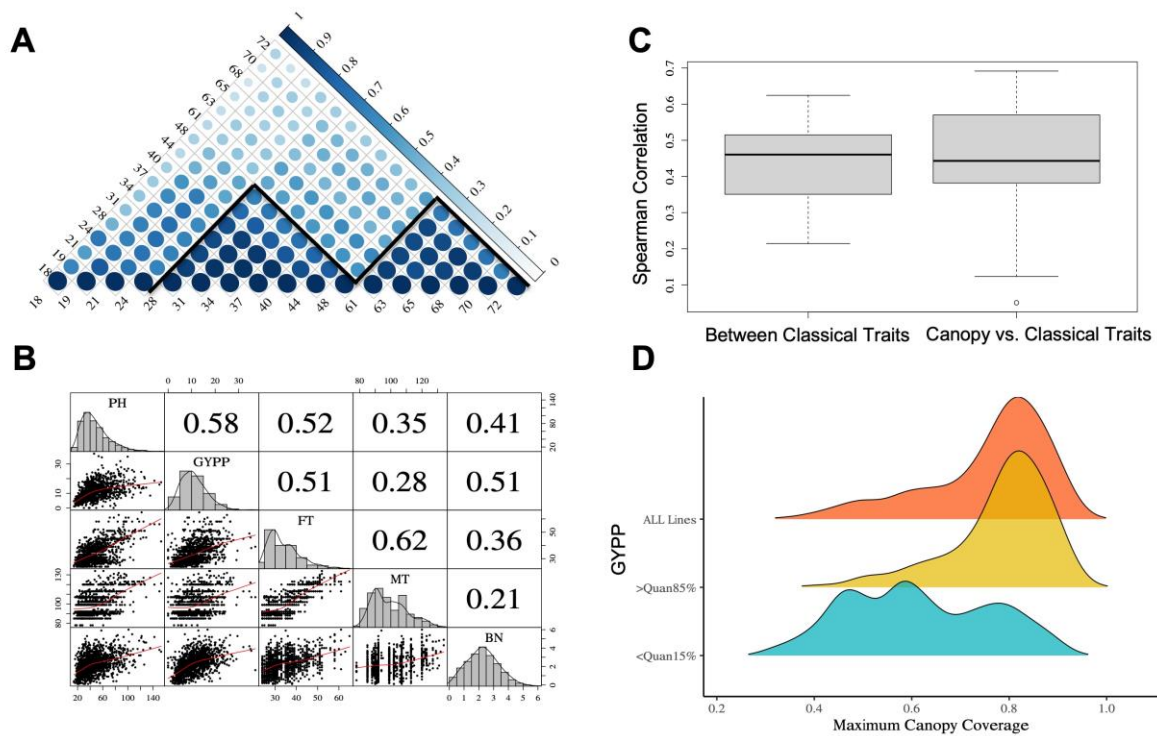


Figure S2. Relationship between canopy coverage and classically scored developmental phenotypes

(A) Pairwise Spearman correlation coefficient between canopy coverage values observed for the same plots at the 17 different time points in Nanchang. Higher correlation is indicated by darker blue shades (see color bar) and larger circles. Row/column labels are the number of days after sowing for individual time points. Thick black lines indicate two groups of time points (first cluster, 28–48 DAS; second cluster, 61–72 DAS) that are distinct from each other but internally highly correlated. **(B)** Distribution and pairwise correlations between the five classical developmental traits assessed in this study (plant height [PH], flowering time [FT], maturity time [MT], branch number [BN], and grain yield per plant [GYPP]). Values given in the top right are the Spearman correlation coefficients for each pairwise comparison. Scatterplots of the plot-by-plot relationships between the same traits and nonlinear fitted lines (in red) are shown in the bottom left triangle. **(C)** Distribution of Spearman's correlation coefficients observed in pairwise comparisons either between classical traits or between classical traits and canopy coverage traits. **(D)** Distribution of maximum canopy coverage observed for individual plots across all time

points considering either all accessions (orange), the top 15% of accessions for yield (yellow), or the bottom 15% of accessions for yield (cyan).

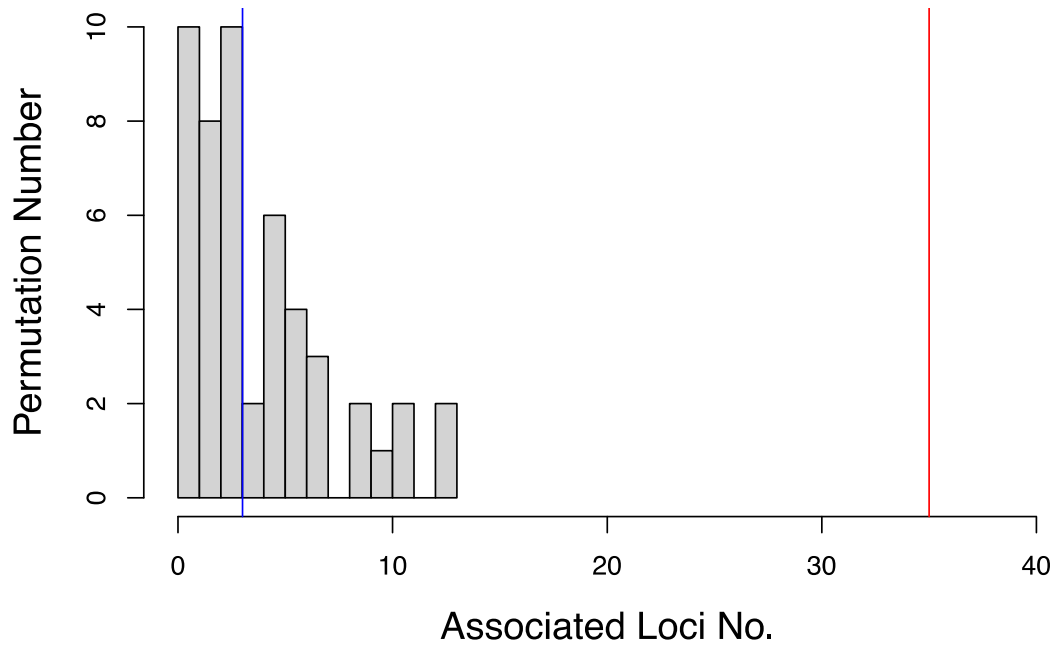


Figure S3. Relationship between the distribution of the number of apparently significant loci identified in GWASs using permuted data and the true number of loci identified when analyzing non-permuted data

The histogram displays the results of GWASs for 13 canopy coverage time points using 50 permutations of genotype/phenotype relationships. The median number of associated loci identified from permuted data (three loci) is indicated by the blue line. The true number of loci identified with the non-permuted data (35) is indicated by the red line.

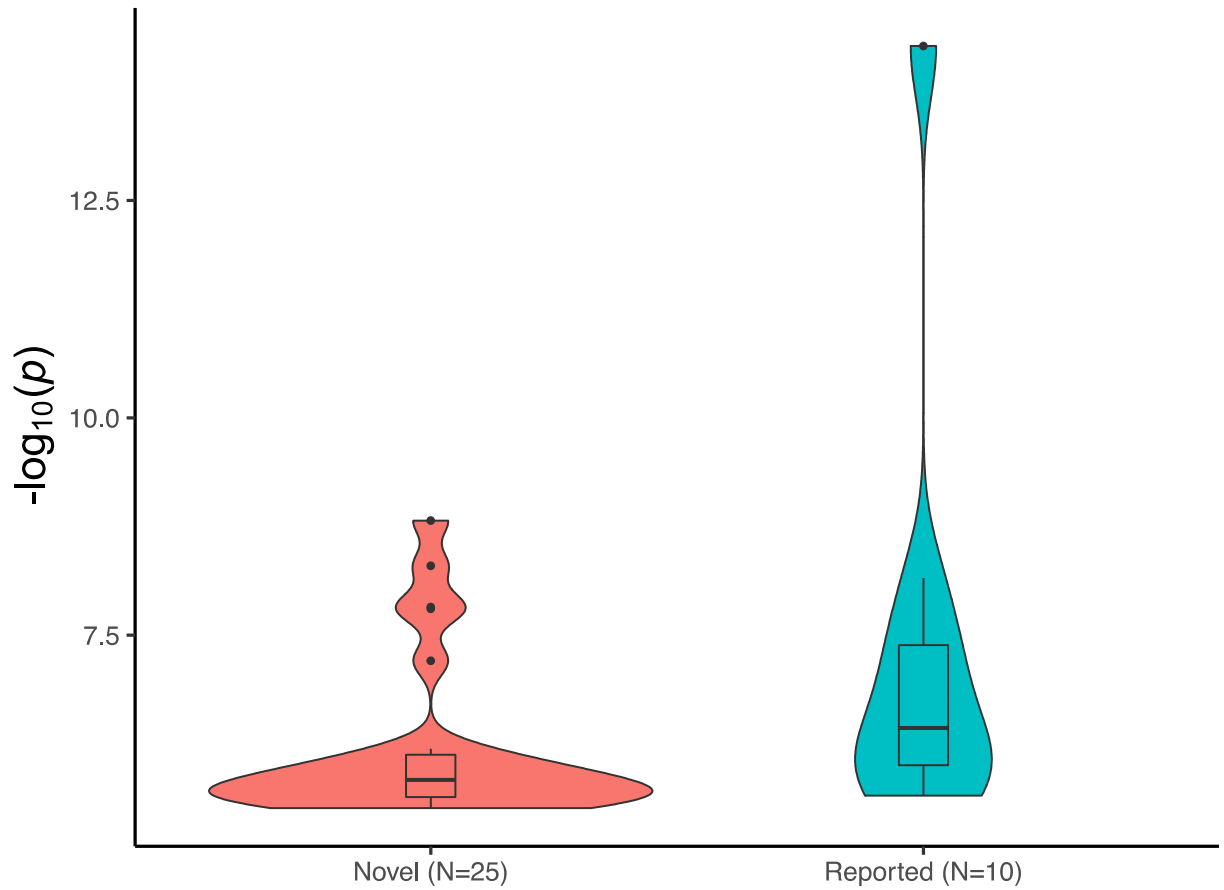


Figure S4. Distribution of p -values for statistically significant GWAS signals detected from either novel (Novel) or previously characterized loci (Reported)

For each unique genomic interval (GWAS peak), the single most significant SNP identified in analysis of canopy coverage from any of the 13 time points is shown. The 10 GWAS signals that overlapped with loci previously reported to influence plant height or flowering time are classified as “Reported” and with the remaining 25 signals are classified as “Novel”.

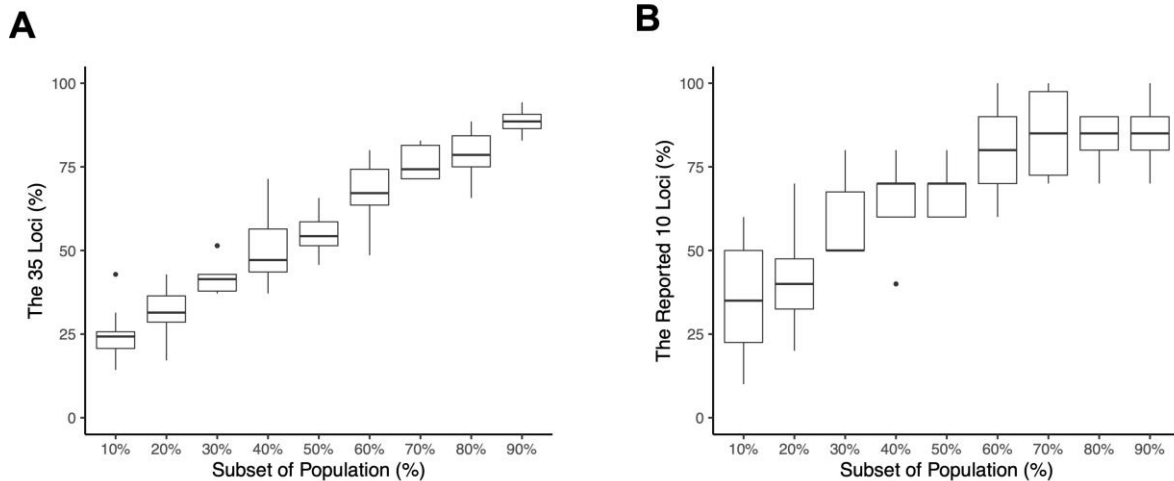


Figure S5. Proportion of loci identified when conducting GWASs using different sized subsets of the complete soybean population

Random subsets consisting of 10%, 20%, and up to 90% of the complete set of 1,303 soybean accessions analyzed in this study were selected and used to conduct GWASs. The subsampling and reanalysis were performed 10 times. **(A)** Average number and distribution of the final set of 35 loci identified in this study that were also identified in GWAS analysis using different sized subsets of the complete population. **(B)** Average number and distribution of the 10 loci corresponding to known genes or QTLs that were previously identified in GWAS analysis using different sized subsets of the complete population.

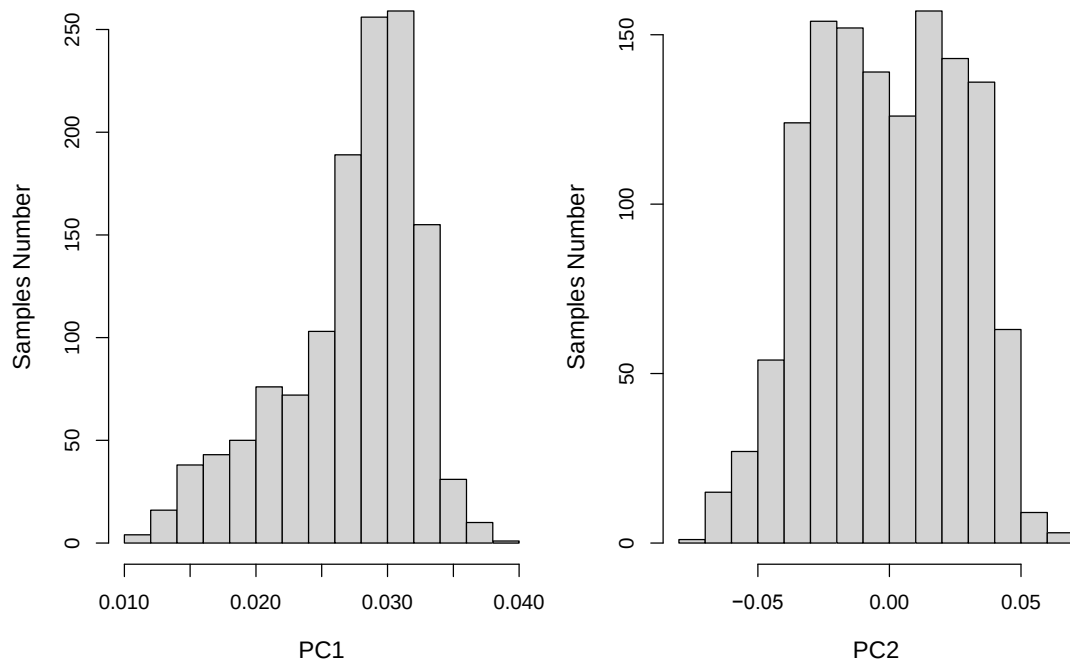


Figure S6. Distribution of the variation in the first two principal components identified when conducting PCA-based dimensionality reduction of observed canopy coverage values at 17 time points in Nanchang

The y-axis (Sample Number) is the number of soybean accessions.

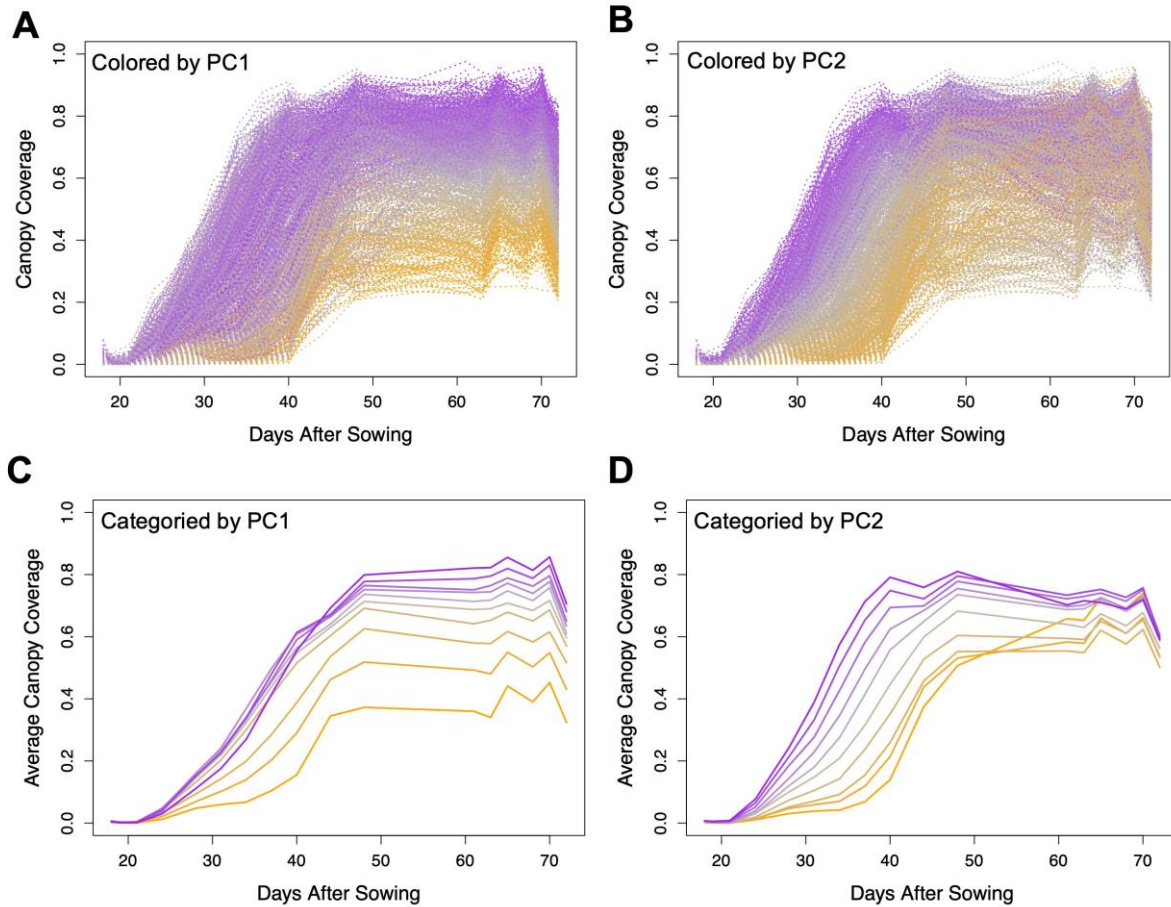


Figure S7. Canopy coverage and its first two principal components in the 1,303 soybean accessions

(A,B) Each dashed line represents one soybean accession planted in Nanchang. The gradient color (purple-grey-orange) indicates the first principal component (PC1) (A) or PC2 (B) values, in decreasing order. (C,D) Distribution of canopy coverage values for 10 groups of accession based on PC1 (C) or PC2 (D) values, separated based on quantiles ([0 to 10%], [10 to 20%], .. [90% to 100%]) for the 1,303 soybean accessions. Average values were calculated for each group and at each time point. The gradient color (purple-gray-orange) represents the 10 quantile groups in decreasing order.

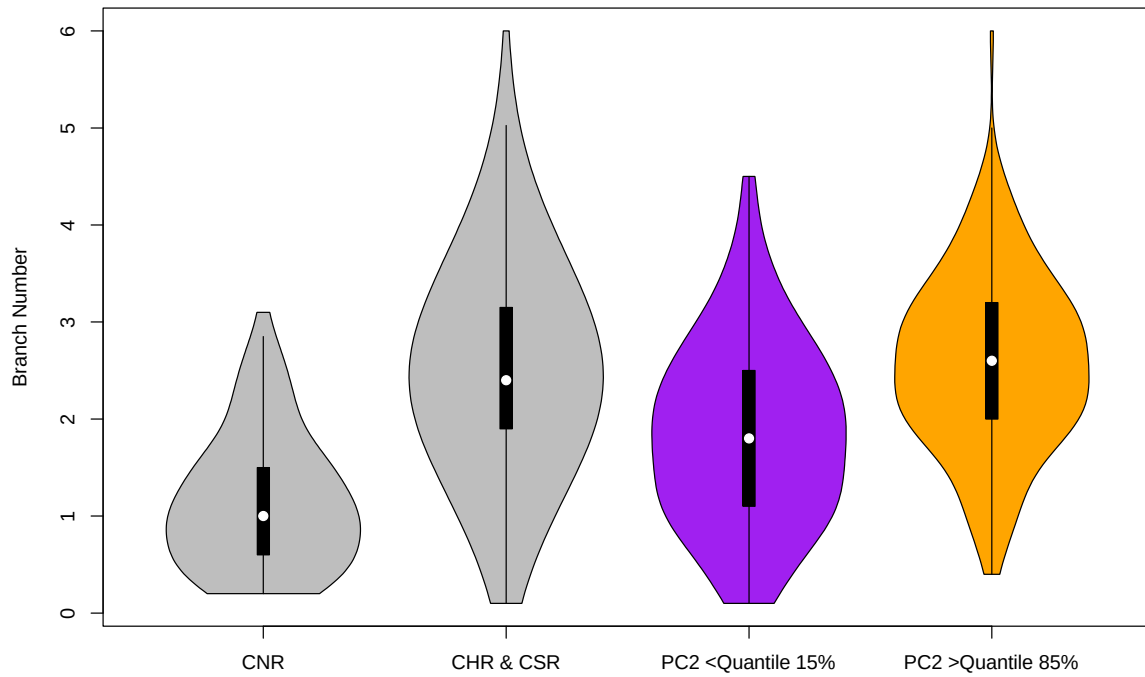


Figure S8. Median branch number and distribution of observed branch numbers for genotypes classified based on subpopulation origin or PC2 score

White points indicate the median branch number for each category. The branch number and PC2 score were obtained from the same plots grown in Nanchang in 2020.

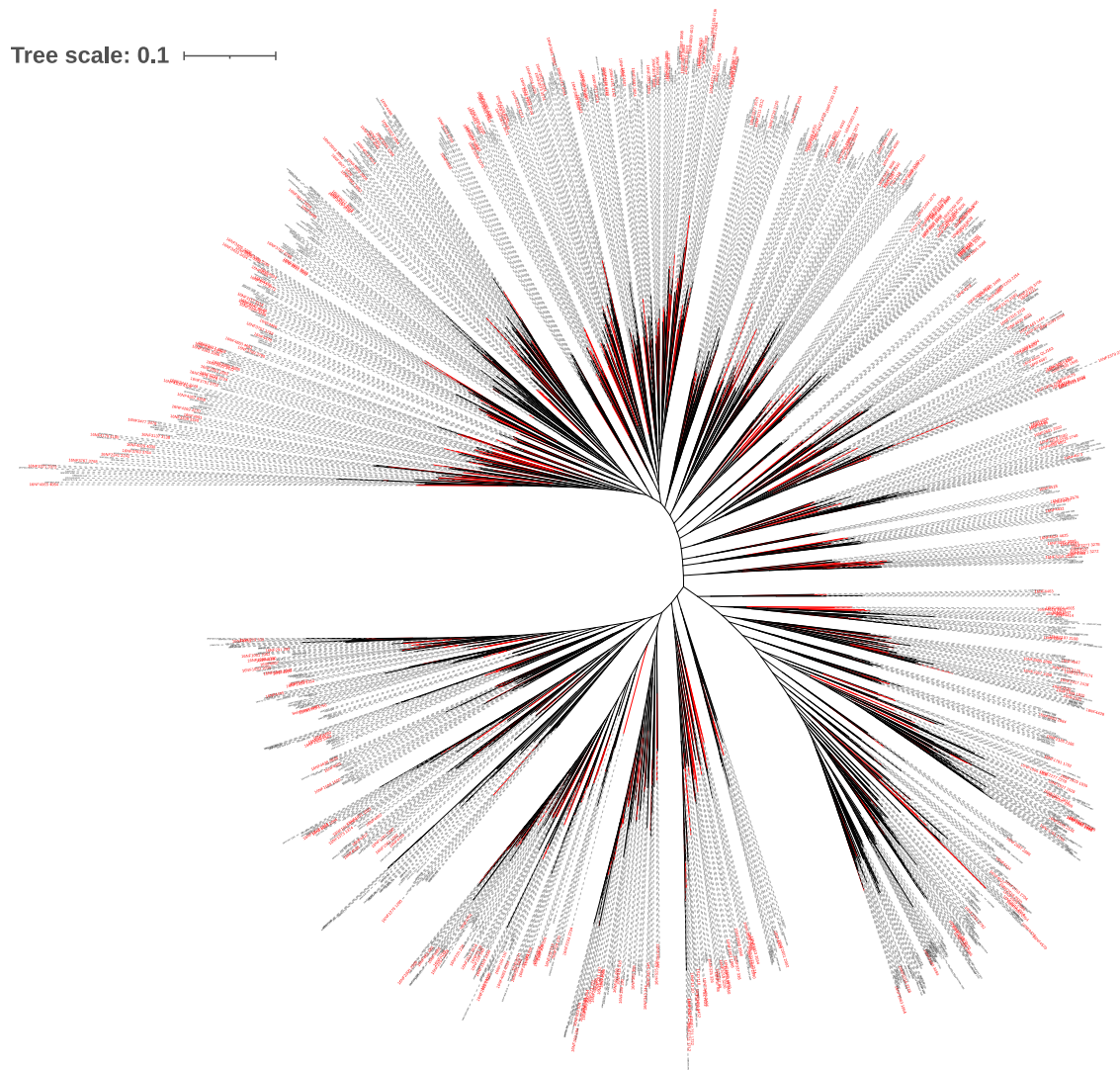


Figure S9. Unrooted phylogenetic tree of the 1,303 soybean accessions

The tree was constructed using data from 1,303 soybean accessions investigated in Nanchang. Phenotyping conducted in Sanya is highlighted by red branches and larger font size.

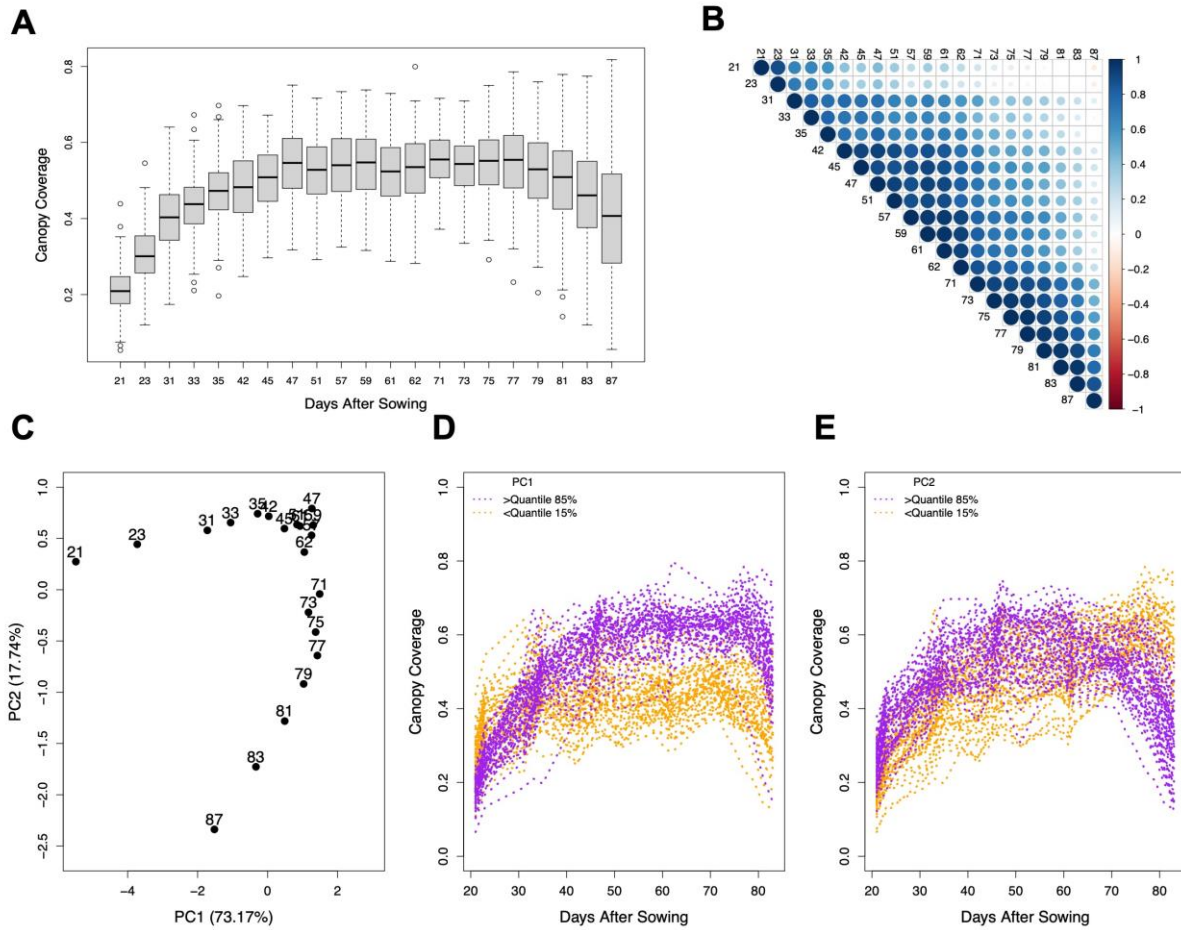


Figure S10. Time-series canopy coverage collected in Sanya

(A) Canopy coverage (CC) of 397 soybean accessions collected over 21 time points. (B) Pairwise Spearman's correlation for CC at each of the 21 time points. (C) Principal component analysis (PCA) performed between the time-series CC. The first two principal components (PCs) are plotted for the 21 time points. (D, E) CC for the top 15% (purple) or bottom 15% (orange) of PC1 values (D) and PC2 values (E) across 21 time points collected in Sanya.

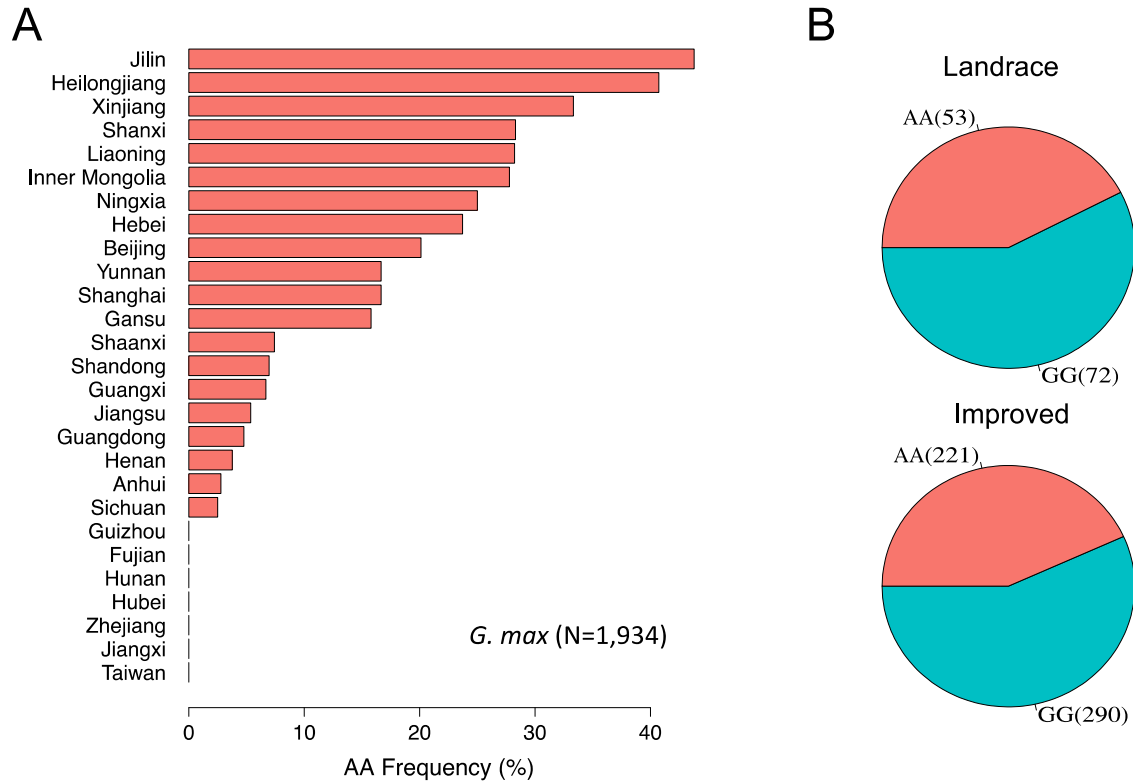


Figure S11. Distribution of the two alleles of the peak SNP within the *C-1-1* locus across China and between landraces and improved varieties in northeast China

(A) Genotype frequency of the leading SNP at locus *C-1-1*. The leading SNP (Wm82 v2 1-292611) was converted to its corresponding ZH13 v2 physical position (1-882783). There were 1,934 *G. max* cultivars with Chinese province information (Liu et al., 2020) that used the ZH13 v2 genome as reference. The frequency of the AA genotype was calculated for each province. The provinces with fewer than five accessions are not listed.

(B) Genotype pie chart for landrace and improved cultivars for the top two provinces in **(A)**.

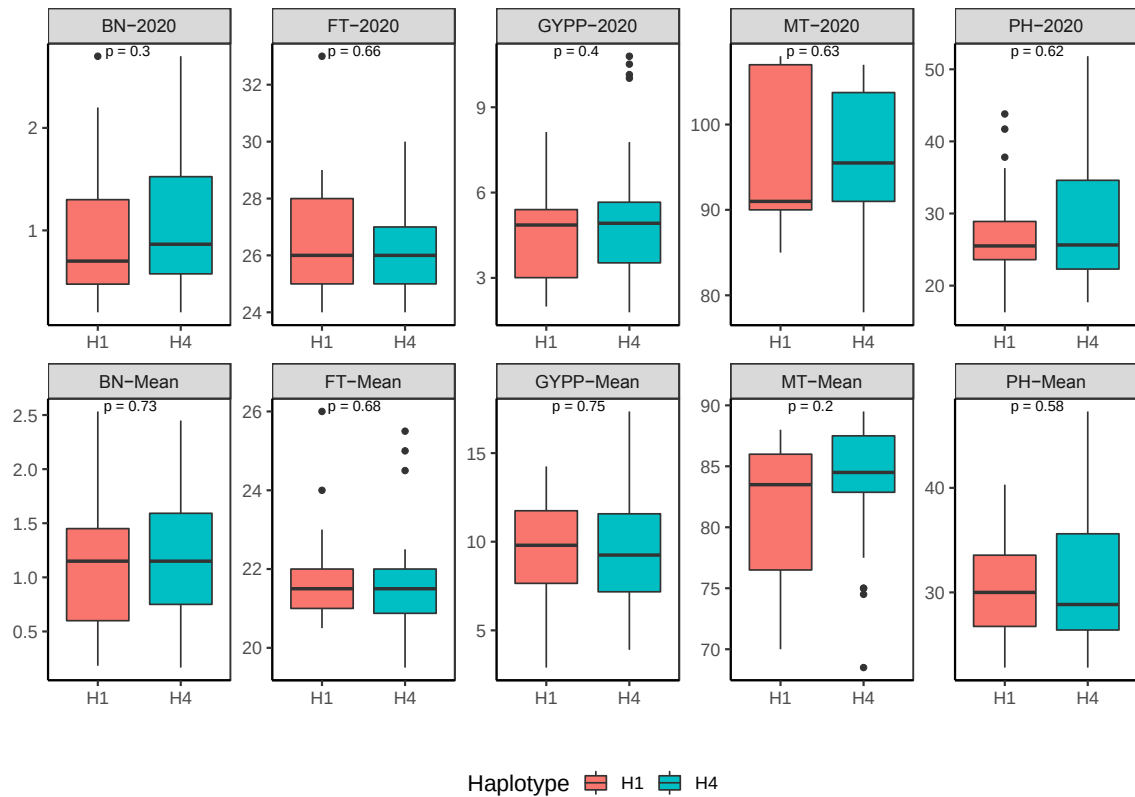


Figure S12. Comparison of observed trait values for improved soybean cultivars from the CNR subpopulation carrying different haplotypes at the C-1-1 locus when grown and phenotyped in Nanchang

Top row: five traits phenotyped in 2020 in Nanchang; bottom row: five traits phenotyped in Nanchang in both 2017 and 2018; the observed values were averaged across the two years. FT and MT in 2020 were reported as days from sowing day. FT-Mean and MT-Mean from 2017/2018 were reported as days from germination. The units of GYPP and PH are grams and centimeters, respectively. The *p*-values shown were calculated using a one-tailed *t*-test comparing values between groups of individuals carrying one of the two haplotypes.

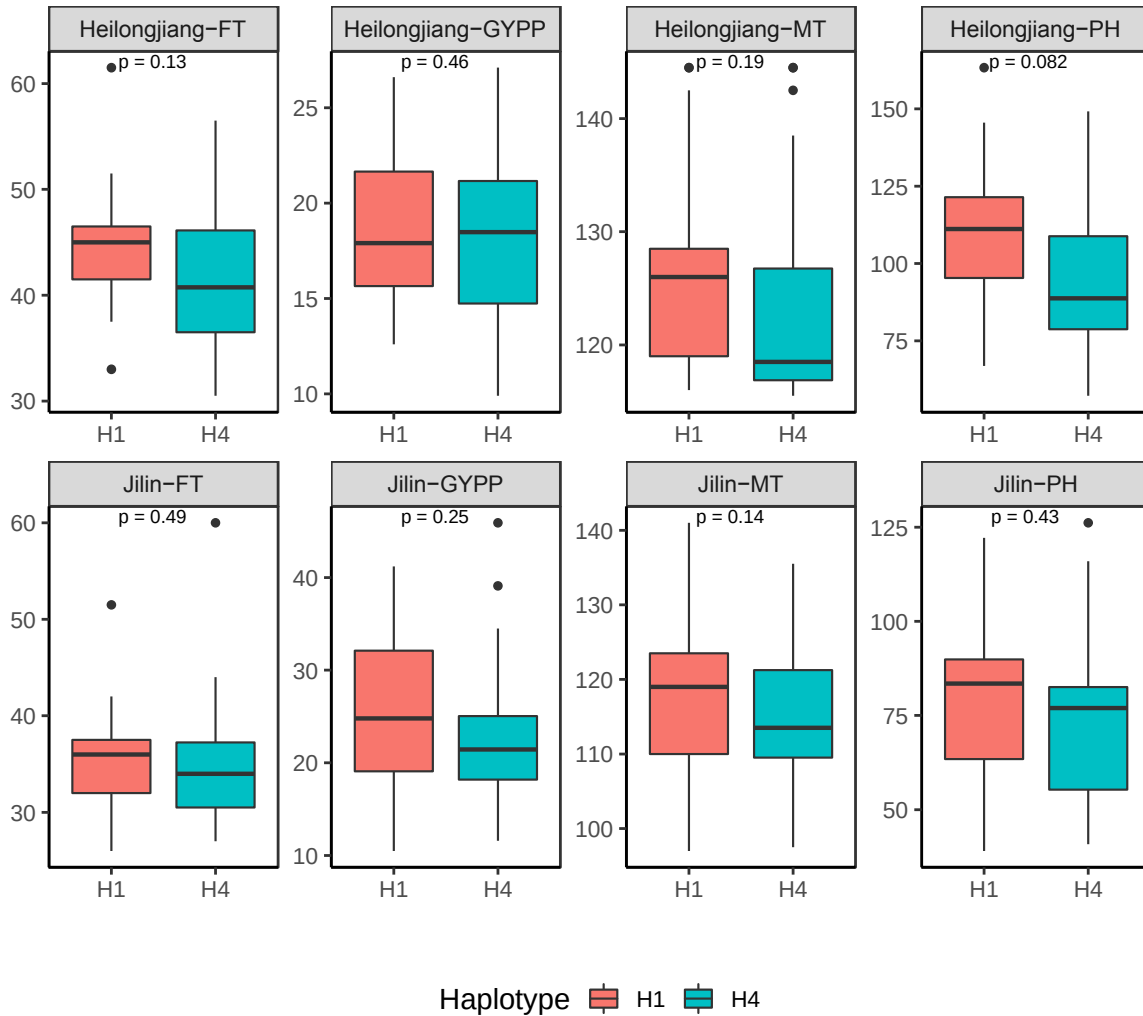


Figure S13. Comparison of manually scored plant traits for improved cultivars of the CNR subpopulation carrying different haplotypes of the *C-1-1* locus evaluated in Heilongjiang and Jilin

Heilongjiang and Jilin are in the Northeastern province of China and show local adaptation by soybean accessions from the CNR subpopulation. The phenotypic values of each trait are averages from 2018 and 2019. The p -values were obtained by one-tailed t -tests between phenotypes from the two haplotype groups.

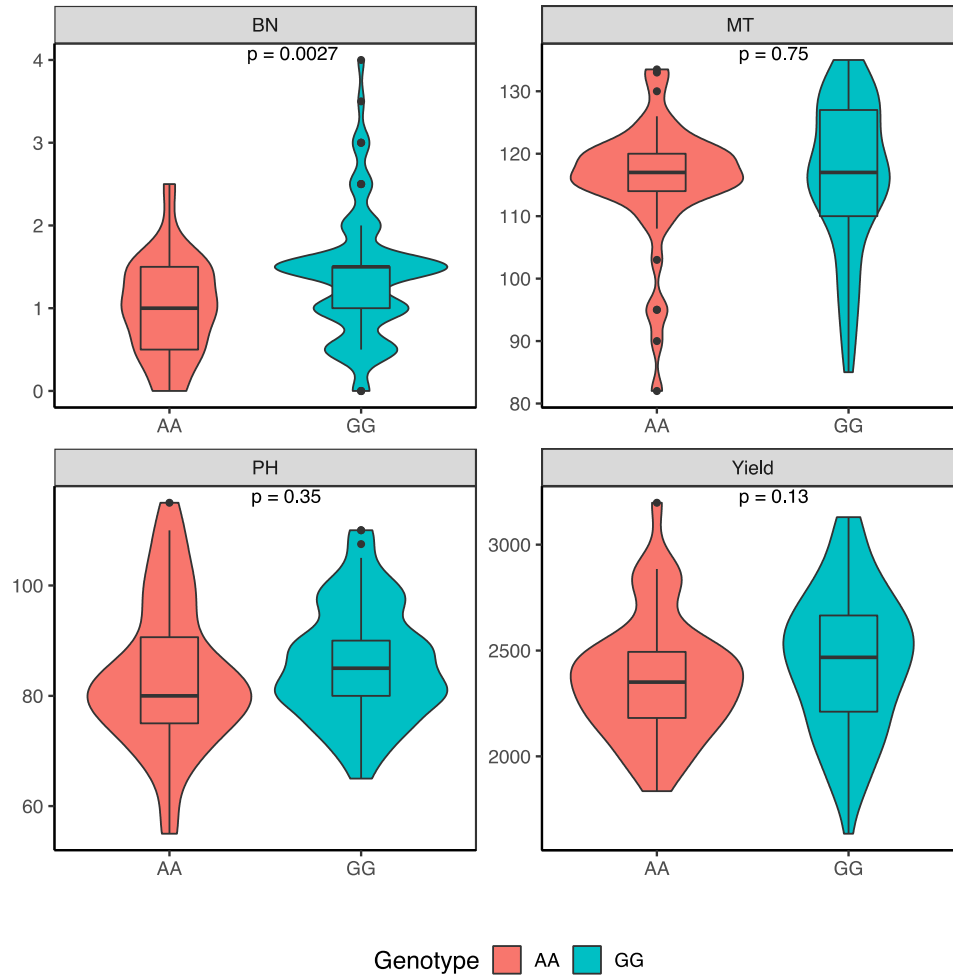


Figure S14. Distribution of traits as a function of genotype for the *C-1-1* locus in improved cultivars developed in Heilongjiang and Jilin

The 128 improved cultivars were developed in Heilongjiang and Jilin. Data for four traits were extracted from their 2-year regional trials records. MT is recorded as days from germination. Yield is reported in kg/hm², and PH is reported in cm. The *p*-values were obtained by one-tailed *t*-tests between two genotype groups.

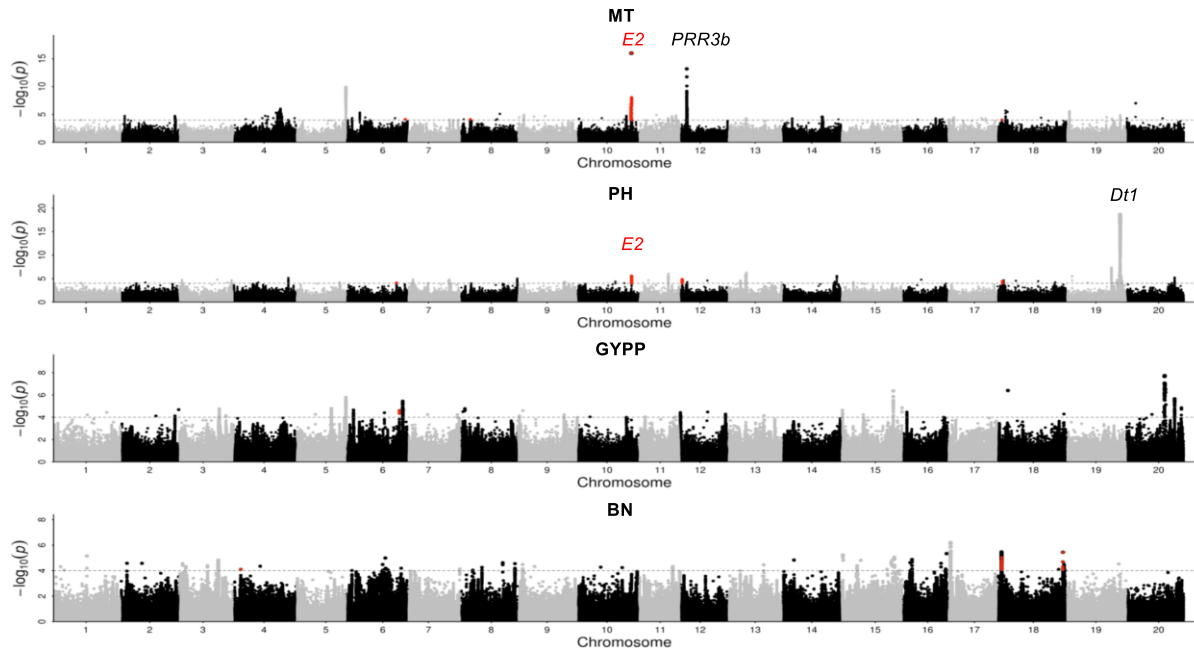


Figure S15. GWAS results conducted using manually scored classical developmental phenotypes in soybean

Manhattan plots for maturity time (MT), plant height (PH), grain yield per plant (GYPP), and branch number (BN) using the average value from 2017 and 2018 data collected in Nanchang and derived from the same population with canopy coverage data. The gray dashed lines indicate the cutoff ($-\log_{10}(p)$, 4) for the identification of the 35 loci in the canopy coverage analysis. Red circles mark SNPs that are both significant in these analyses and were separately identified as significant in the GWAS analysis conducted using canopy coverage data. The known gene *E2* was identified by MT and canopy coverage and is labeled in red. The known genes *Dt1* and *PRR3b* were identified using manually collected trait data but not when using canopy coverage data and are labeled in black.

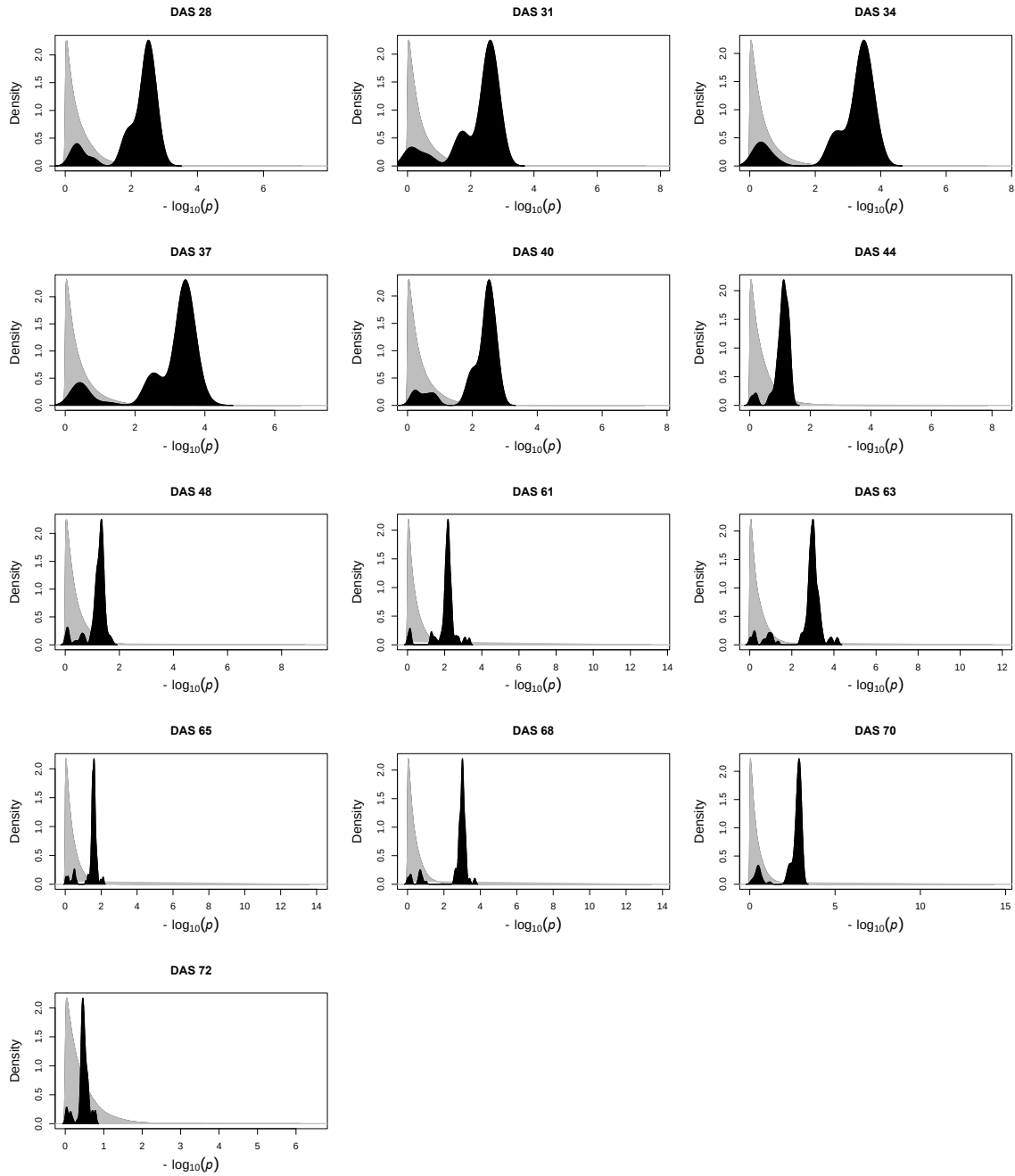


Figure S16. Distribution of p -values observed for SNPs in the C-8-9 interval (black) relative to the overall genome-wide distribution of p -values (gray)

Density plots of the $-\log_{10}(p)$ of SNPs located within the C-8-9 interval (Chr8: 11,601,216 to 11,623,966 bp) (black) and genome wide (gray). The p -values were extracted from the GWAS of canopy coverage for each time point. DAS, days after sowing.

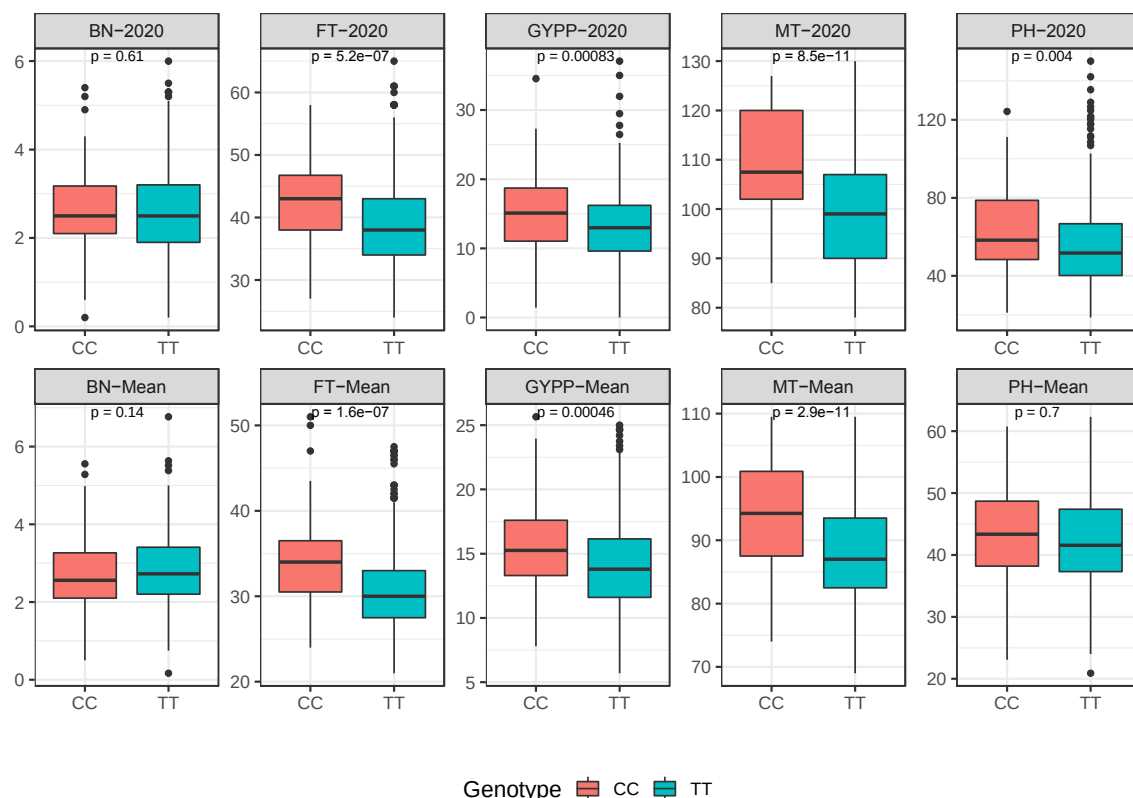


Figure S17. Traits in Nanchang between CSR landraces differing in their genotype at locus C-8-9

Top row: five traits phenotyped in 2020 in Nanchang; bottom row: average values from 2017 and 2018 in Nanchang. The FT and MT of 2020 are shown in days from sowing day to be comparable to canopy coverage, or from germination day in 2017 and 2018. GYPP is reported in g; PH is reported in cm. The p -values were obtained by one-tailed t -tests between two groups.

Table S1. Material information and traits collected in Nanchang in 2020

Table S2. Flight logs

Table S3. List of the 35 canopy coverage–associated loci and leading p -values from each GWAS in this study

Table S4. List of the 10 out of 35 canopy coverage–associated loci were previously reported to be linked or associated with flowering time or plant height

Table S5. Average canopy coverage collected in Sanya

Table S6. Summary of the three genes and polymorphisms within locus *C-1-1*

Table S7. Genotype summary of locus *C-1-1* for *G. soja* with Chinese province information

Table S8. Average phenotyping data for traits collected in Nanchang in 2017 and 2018

Table S9. Average phenotyping data of traits collected in Heilongjiang and Jilin provinces 2018 and 2019

Table S10. Two-year regional trials records of 128 improved cultivars developed in Heilongjiang and Jilin provinces

Table S11. Genotype summary of locus *C-8-9* for *G. soja* with Chinese province information