

Genomic dissection of divergent selection reveals a roadmap for reciprocal improvement of Chinese and American soybean^{oo}

Chengzhi Jiao^{1,2†} , Delin Li^{1,3†} , Hao Zhang^{1†} , Zhongren Zhang^{4†} , Xueqing Wang^{1†} , Long Yan⁵ , Tianqing Zheng¹ , Chaosen Zhao⁶ , Hongmei Qiu⁷ , Guangnan Xing⁸ , Guoyu Hu⁹ , Like Liu¹⁰ , Qiang Wang¹¹ , Wen Huang¹² , Xing Wang¹³ , Yumin Wang⁷ , Long Miao¹⁴ , YanFei Li¹ , Xin Jing⁴ , Xiaojian Fang² , Yu Tian¹ , Ying-Hui Li^{1*} , Xiaobo Wang^{1,14*}  and Li-Juan Qiu^{1*} 

1. Key Laboratory of Grain Crop Genetic Resources Evaluation and Utilization (MARA), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

2. State Key Laboratory of Wheat Improvement, Shandong Laboratory of Advanced Agricultural Sciences, Peking University Institute of Advanced Agricultural Sciences, Weifang 261325, China

3. Xianghu Laboratory, Hangzhou 311231, China

4. Smartgenomics Technology Institute, Qingdao 266111, China

5. Institute of Cereal and Oil Crops, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang 050031, China

6. Crops Research, Institute of Jiangxi Academy of Agricultural Sciences, Nanchang 330200, China

7. Soybean Research Institute, Jilin Academy of Agricultural Sciences, Changchun 130033, China

8. Soybean Research Institute, Nanjing Agricultural University, Nanjing 210095, China

9. Crop Institute of Anhui Academy of Agricultural Sciences, Hefei 230031, China

10. School of Life Sciences, Liaocheng University, Liaocheng 252059, China

11. Institute of Crop Resources, Heilongjiang Academy of Agricultural Sciences, Harbin 150086, China

12. Tong Hua Academy of Agricultural Sciences, Meihoukou 135007, China

13. Xuzhou Institute of Agricultural Sciences in XuHuai Region of Jiangsu, Xuzhou 221131, China

14. School of Agronomy, Anhui Agricultural University, Hefei 230036, China

[†]These authors contributed equally to the work.

*Correspondences: Li-Juan Qiu (qiujuan@caas.cn), Dr. Qiu is fully responsible for the distribution of all materials associated with this article; Ying-Hui Li (liyingshui@caas.cn); Xiaobo Wang (wangxiaobo@ahau.edu.cn)



Chengzhi Jiao



Li-Juan Qiu

ABSTRACT

Soybean (*Glycine max* [L.] Merr.) was domesticated in China over five millennia ago; yet, its global expansion has resulted in two distinct, independently evolved breeding pools in China and the Americas. While regional breeding has optimized specific traits, a comprehensive roadmap for the reciprocal

exchange of these “breeding-specific” beneficial alleles is lacking. Here, we performed a large-scale genomic dissection of 2,377 soybean accessions, integrating 8.4 million high-quality SNPs with multi-environment phenotyping across 11 locations. Through genome-wide association studies (GWAS), we identified 261 significant loci (223 novel) governing six key agronomic traits, revealing that Chinese and American germplasms have accumulated distinct sets of beneficial alleles of seed size/protein in China and oil/yield in the Americas. To harness this untapped potential, we developed the Soybean Targeted Improvement Navigation System (STINS), a framework integrating population genomics, GWAS, and genomic selection (GS). We demonstrate the potential of targeted beneficial allele introgression between populations via gene modification for rapid trait improvement. Proof-of-concept functional evaluation through CRISPR/Cas9-mediated knockout of *GmST05* in an American cultivar and overexpression of a resistant

***GmSNAP18* haplotype in a Chinese cultivar provided experimental support for STINS-predicted targets for protein content and SCN resistance improvement. Our study provides a navigable genomic resource for “knowledge-driven design breeding”, offering a blueprint for the global exchange of elite alleles to ensure future food security.**

Keywords: beneficial alleles, genomic selection, GWAS, soybean, STINS

Jiao, C., Li, D., Zhang, H., Zhang, Z., Wang, X., Yan, L., Zheng, T., Zhao, C., Qui, H., Xing, G., et al. (2026). Genomic dissection of divergent selection reveals a roadmap for reciprocal improvement of Chinese and American soybean. *J. Integr. Plant Biol.* **00**: 1–15.

INTRODUCTION

Soybean is a cornerstone of global agriculture, providing a primary source of vegetable oil and protein for both human nutrition and animal feed. Originating from domestication in central China over 5,000 years ago (Jeong et al., 2019; Li et al., 2023), soybean evolved into a diverse array of landraces. A small fraction of this genetic diversity was subsequently introduced to Europe and North America in the 18th century, and later to South America (Qiu et al., 2011). This geographical expansion, which established Brazil, the United States, and Argentina as leading producers (Qiu et al., 2011), demanded profound genetic adaptations to disparate agro-ecological zones. As a photoperiod-sensitive short-day plant, its successful cultivation in higher-latitude regions was enabled by natural and human selection of mutations in key flowering time genes, such as the *E1* family, *Tof11*, and *Tof12/GmPRR3b* (Tsubokura et al., 2014; Lu et al., 2017; Li et al., 2020). Conversely, adaptations like the long juvenile trait, conferred by mutations in *J* and *Tof16*, facilitated its expansion into tropical regions (Dong et al., 2021). Parallel to these adaptive changes, breeders have consistently targeted genes governing plant architecture, such as *Dt1* and *Dt2*, to enhance yield potential (Liu et al., 2010; Liang et al., 2022).

Distinct regional market demands and pathogen pressures have further sculpted divergent breeding goals, driving genomic divergence between major production regions. Chinese breeding programs have traditionally emphasized high protein content for direct food uses, such as tofu and soymilk, while American programs have prioritized a balance between oil content and yield for industrial and feed markets (Grieshop and Fahey, 2001; Xu et al., 2022). This divergence extends to biotic stress resistance, with soybean mosaic virus (SMV) being a paramount threat in China (Lin et al., 2022) and soybean cyst nematode (SCN) inflicting massive economic losses in the Americas (Allen et al., 2017). Consequently, Chinese and American breeders have likely selected for different sets of beneficial alleles from the shared genetic foundation of East Asian landraces, potentially leading to significant genomic and phenotypic differentiation.

Recently, the dissemination of key known genes was well described with 8,105 soybean accessions for adaptation and selection across ecological (Zhu et al., 2025), a comprehensive, genome-wide understanding of the genetic divergence underlying this adaptation and breeder selection remains incomplete. Key questions persist: How is the

genomic structure of modern cultivars partitioned between these two major breeding germplasms, and what is the extent of genetic diversity retained after the founding event and selection breeding? What are the specific genetic loci responsible for their distinct agronomic trait profiles, and how many valuable alleles remain unexploited? To what extent have unique sets of beneficial alleles been preferentially accumulated in each population, and can they be utilized for reciprocal improvement? A comprehensive framework capable of answering these questions would therefore integrate population genomics, GWAS, and predictive breeding to systematically evaluate this untapped potential for reciprocal improvement.

To address these questions, we analyzed a worldwide collection of 2,377 soybean accessions, including wild soybeans, landraces, and improved cultivars from China and the Americas, utilizing over 8.4 million high-quality SNPs derived from whole-genome resequencing data (Zhou et al., 2015; Li et al., 2020, 2023; Liu et al., 2020). We integrated high-throughput genotyping with multi-environment phenotyping for six key agronomic traits. Our study (i) elucidates the population genetic structure and quantifies the genetic diversity and relationship between Chinese and American germplasm; (ii) identifies trait-associated loci through GWAS, revealing a reservoir of novel, high-value loci beyond previously known genes; (iii) quantifies the differential accumulation of beneficial alleles for divergent breeding targets in Chinese and American improved cultivars, defining breeding-specific allele sets; and (iv) validates the functional potential of these alleles through genomic selection models and a proof-of-concept gene-modified strategy, establishing a framework for targeted genetic improvement. By dissecting the genomic basis of this divergence, our work provides a valuable resource and a navigable system for the strategic utilization of alien beneficial alleles to accelerate future, knowledge-driven soybean breeding.

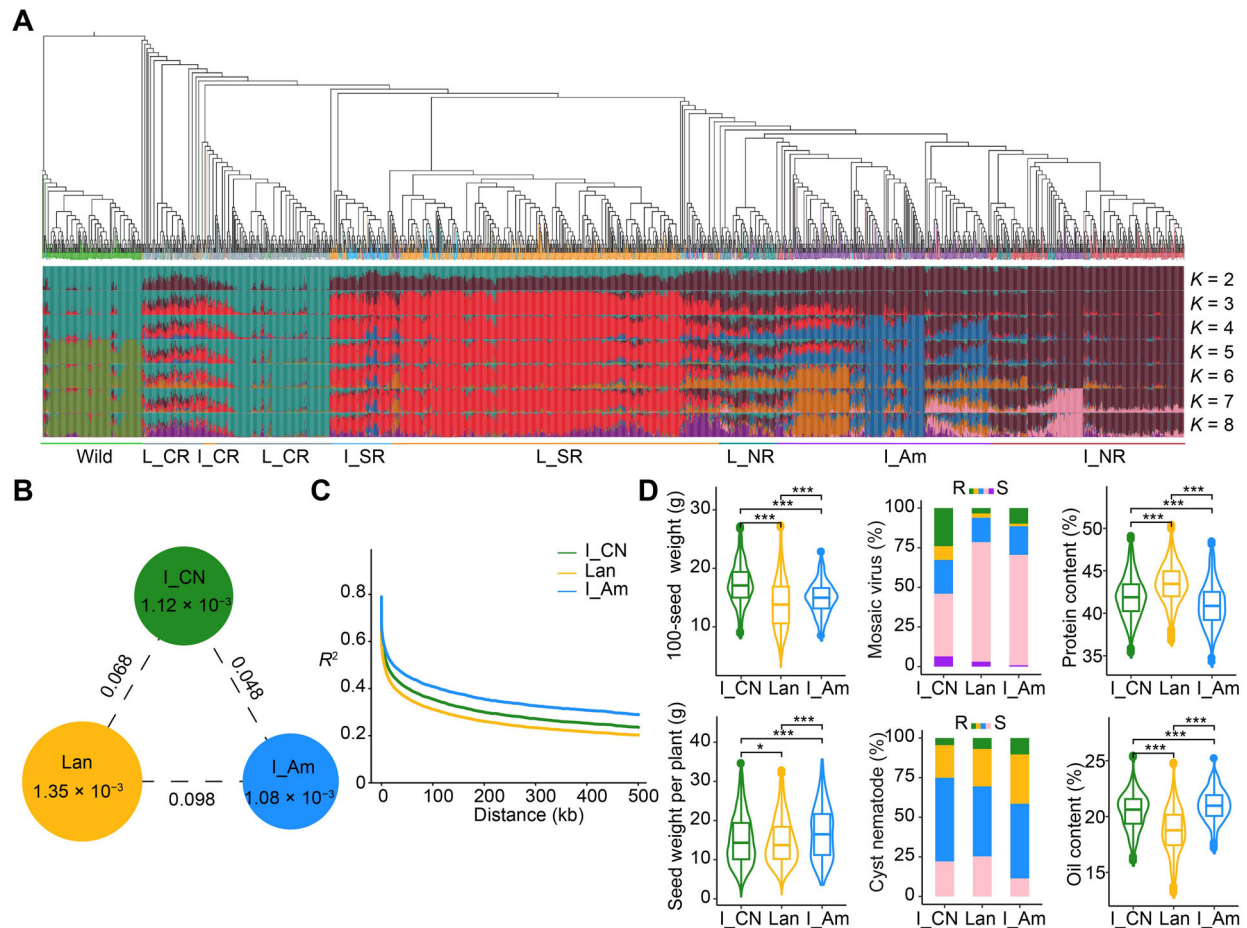
RESULTS

Population genetic structure and diversity of soybean from China and the Americas

To investigate the genomic divergence between Chinese and American soybean cultivars in a broader geographic context, we analyzed 2,377 soybean accessions, comprising wild accessions, landraces, and improved cultivars from diverse

geographic regions (Dataset S1). Specifically, the panel consists of 218 wild accessions, 1,101 landraces, and 1,058 improved cultivars. Among wild accessions, 179 originate from China, 19 from Japan, nine from Russia, 10 from Republic of Korea, and one from the United States. The landraces comprise 1,087 accessions from China, one from Democratic People's Republic of Korea, and 13 from Republic of Korea. The improved cultivars include 598 accessions from China, 432 from the United States, and 28 from Canada. All accessions had publicly available whole-genome resequencing data (Zhou et al., 2015; Li et al., 2020, 2023; Liu et al., 2020). Raw resequencing data were aligned to the Williams 82 v4 reference genome (Schmutz et al., 2010), yielding 61,254,543 single nucleotide polymorphisms (SNPs). After stringent filtering, 8,485,501 high-quality SNPs were retained for downstream analyses (Figure S1; Table S1). Based on these SNPs, we systematically analyzed the population structure and

genetic diversity of soybeans (Figure 1A). At $K=2$, all accessions were clearly separated into wild and cultivated soybeans. At $K=8$, cultivated soybeans were subdivided into seven sub-populations: landraces from the Huang-Huai region of China (L_CR) and improved cultivars from the same region (I_CR); landraces from South China (L_SR) and improved cultivars from South China (I_SR); landraces from North China (L_NR) and improved cultivars from North China (I_NR); and improved cultivars from the United States and other American countries (I_Am) (Figure 1A). At $K=2-4$, L_CR showed genetic similarity to wild soybeans, supporting the hypothesis of soybean domestication in the Huang-Huai region (Li et al., 2023). From $K=5-8$, landraces and improved cultivars from the same region clustered together, indicating independent breeding efforts within each ecological zone. I_Am clustered closely with L_NR and shared a proportion of genetic components,



suggesting that American improved cultivars likely originated from the introduction and subsequent improvement of North Chinese landraces (Figure 1A).

To efficiently compare breeding processes between China and the Americas, we merged the three Chinese landrace groups (L_NR, L_CR, L_SR) into a single pre-improvement group (Lan). Similarly, the three Chinese improved cultivar groups (I_NR, I_CR, I_SR) were combined into a post-improvement Chinese group (I_CN), while I_Am was retained as the American improved group. Genetic diversity analysis revealed that Lan had the highest nucleotide diversity (1.35×10^{-3}), followed by I_CN (1.12×10^{-3}), and I_Am (1.08×10^{-3}) (Figure 1B). The fixation index value (F_{ST}) between the Lan and I_Am groups was 0.098, followed by Lan and I_CN (0.068), and I_CN and I_Am (0.048) (Figure 1B). Linkage disequilibrium (LD) decay showed a consistent trend (Figure 1C).

We further compared agronomic traits among the three groups. Both Chinese and American improved cultivars showed significant increases in 100-seed weight compared to Lan (average value, I_CN: 17.19, I_Am: 14.90, Lan: 13.94), with a greater increase in Chinese cultivars, indicating that seed weight is a common breeding target, but with stronger emphasis in China. For resistance to mosaic virus, both improved groups exhibited enhanced resistance (I_CN: 2.97, I_Am: 3.50, Lan: 3.72), though the improvement was more pronounced in Chinese cultivars. In contrast, protein content decreased significantly in both improved groups relative to Lan (I_CN: 41.92, I_Am: 40.98, Lan: 43.41), but the reduction was less in Chinese cultivars, possibly due to the use of soybeans for feed in China. For seed weight per plant (I_CN: 15.09, I_Am: 16.81, Lan: 14.66), resistance to SCN (I_CN: 2.94, I_Am: 2.60, Lan: 2.88), and oil content (I_CN: 20.43, I_Am: 20.94, Lan: 18.72), American cultivars showed stronger selection gains. These results indicate that Chinese breeding has prioritized seed weight, SMV resistance, and protein content, whereas American breeding has focused more on seed weight per plant, SCN resistance, and oil content (Figure 1D).

GWAS identifies candidate genes and loci for breeding traits of soybean

To comprehensively identify functional genes and quantitative trait loci (QTL) associated with key breeding traits, we collected phenotypic data for six traits from 1,910 cultivated soybean accessions (landraces and improved cultivars) evaluated across 11 locations ranging from 28°41' to 45°45' N (Nanchang, Wuhang, Hefei, Nanjing, Mengcheng, Xuzhou, Liaocheng, Shijiazhuang, Tonghua, Gongzhuling, Harbin) (Figure S2). Traits included yield-related traits (100-seed weight, seed weight per plant), quality traits (seed oil and protein content), and resistance traits (SCN and SMV). Best linear unbiased predictors (BLUPs) values were calculated for each trait with multiple-year and environment data. Phenotypic distributions and cross-environment consistencies confirmed high data quality suitable for GWAS.

Using 8,485,501 high-quality SNPs, we performed GWAS with the FarmCPU model, setting a significance threshold of $-\log_{10}(P) = 8$ (Bonferroni correction). Candidate loci were filtered to retain only those detected in BLUP values and in at least two individual environments. Within LD blocks, only the peak SNP with the strongest signal was retained. In total, 261 peak SNPs associated with breeding traits were identified, corresponding to 3,626 candidate genes, including well-known cloned genes such as *GmSTO5/GmMFT*, *GmSNAP18*, *GmGA3ox1*, *GmPRR3b*, *GmSWEET39*, *POWR1*, and *Dt1* (Figures 2A, S3–S8; Dataset S2; Liu et al., 2010, 2017; Li et al., 2020; Miao et al., 2020; Duan et al., 2022; Goettel et al., 2022; Hu et al., 2022; Cai et al., 2023a). The distribution of these candidate loci varied significantly among Lan, I_CN, and I_Am (Figure 2B). Among them, 38 were previously known, while 223 were novel (Dataset S3). Phenotypic variation explained (PVE) by known loci was generally higher than that of novel loci; however, there are still 100 novel loci that showed PVE values exceeding those of some known loci, which are regarded as the most promising candidates for future soybean molecular breeding (Figure 2C). Novel loci with PVE above the lower quartile of known loci were most frequently associated with 100-seed weight and protein content (Figure 2D), highlighting their potential for future breeding.

To characterize the associated loci, we performed detailed genetic analyses for represented loci of two yield-related and two quality-related key traits. For 100-seed weight, three novel loci were further investigated: Gm06-5,660,496, Gm06-13,756,943, and Gm12-5,506,964. The beneficial haplotypes of Gm06-5,660,496 and Gm06-13,756,943 were more frequent in Chinese than American cultivars and were selected across all Chinese regions but not in the Americas (Figure S9). Gm06-5,660,496 was strongly associated with Glyma.06G073800, which was an unknown functional gene, whereas Gm06-13,756,943 was associated with Glyma.06G165800 (encoding BIC1A_XENLA protein) (Table S2). Gm12-5,506,964 was the known causal SNP of the flowering gene *GmPRR3b* ($-\log_{10}(P) = 8.68$) via changes in the stop codon (Figure S10A, B). Its beneficial large-seed haplotype (Hap1: TT) was slightly more frequent in Chinese than American cultivars (Figure S10C, D), with the highest frequency in North China (99.53%), lowest in Central China (53.23%), and similar frequencies in South China (95.52%) (Figure S10E). For protein content, two novel loci (Gm10-45,121,943 and Gm19-46,531,530) showed higher beneficial high-protein haplotype frequencies in Chinese cultivars, with selection in both China and the Americas (Figure S11). Gm10-45,121,943 was strongly associated with Glyma.10G218500, which contained a WD domain. Gm19-46,531,530 was associated with Glyma.19G203400, which was glutamate carboxypeptidase (Table S2). For seed weight per plant, two loci (Gm06-49,320,584 and Gm10-45,600,754) had higher beneficial high-yield haplotype frequencies in American cultivars (Figure S12). Gm06-49,320,584 was strongly associated with Glyma.06G309300, which was a

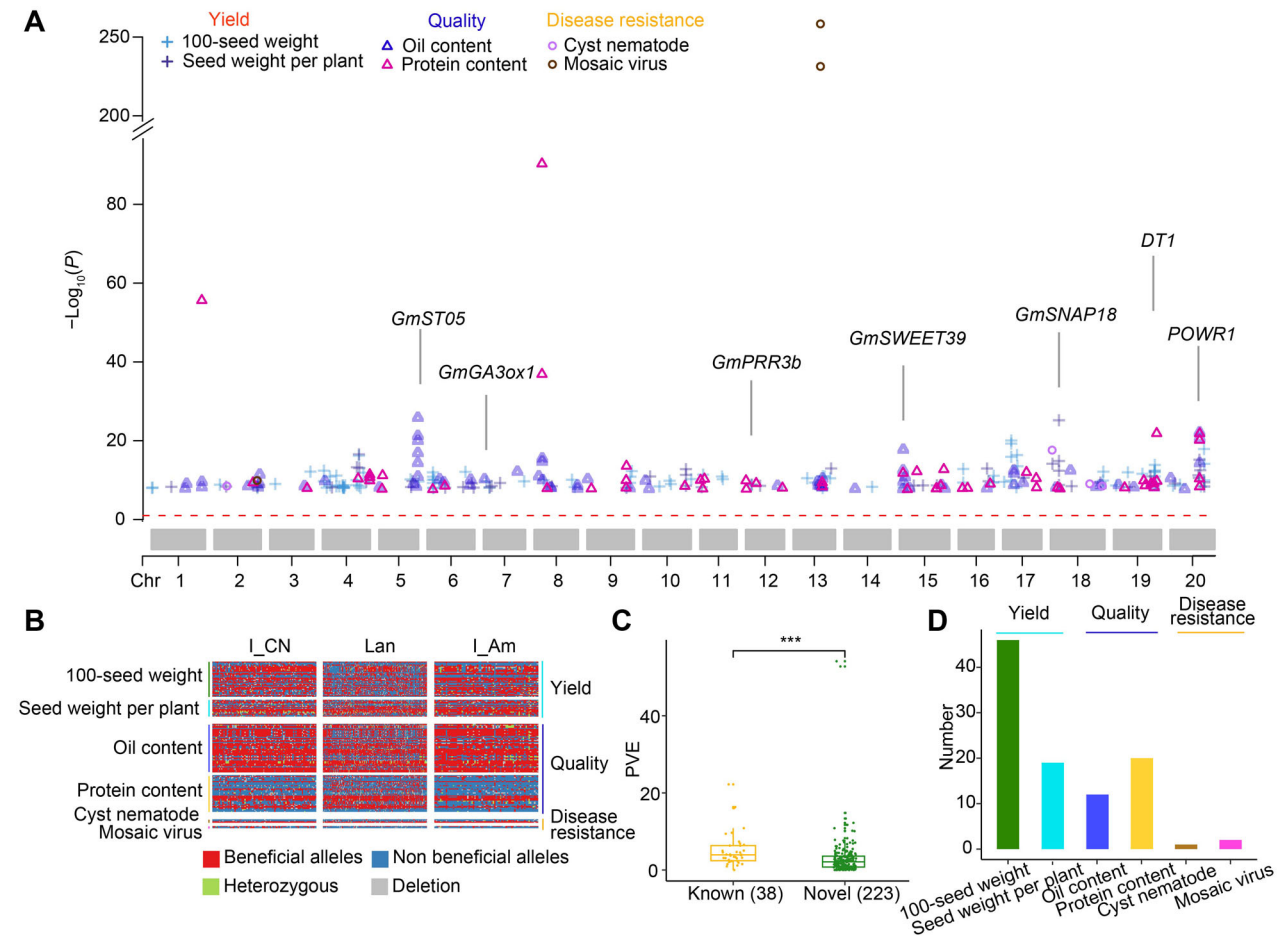


Figure 2. Novel agronomy trait-associated loci and their potential benefits in breeding

(A) Combined Manhattan plot of six agronomy traits across the 20 chromosomes. **(B)** Allele types are indicated for each locus (per horizontal line), with beneficial alleles in red and other alleles in blue. Each individual's allele type is depicted vertically. **(C)** Differences in the proportion of variation explained (PVE) between known genes and novel loci, along with the corresponding P -value from a One-tailed t -test, are shown (* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$). **(D)** Number of novel alleles with a PVE higher than the lower quartile value of known genes is displayed for four traits.

choline kinase. Gm10-45,600,754 was associated with Glyma.10G224700, which was an unknown functional gene (Table S2). For oil content, two loci (Gm08-46,684,397 and Gm08-17,840,625) also showed Americas-specific high-oil haplotype frequencies (Figure S13). The two loci were associated with Glyma.08G361500 and Glyma.08G218500, which were callose synthase and ATPase, respectively (Table S2).

Distribution characteristics of beneficial alleles in Chinese and American soybeans

To compare the difference of genome selection resulted from the independently Chinese and American soybean breeding, we focused on six traits with divergent breeding priorities: 100-seed weight, seed weight per plant, seed protein content, seed oil content, SCN resistance, and SMV resistance. Beneficial alleles for six traits were identified based on phenotypic effects and breeding directions (Figure S14; Dataset S4). Additive effects were significant in Lan, I_CN, and I_Am, with trait values improving as beneficial alleles accumulated, though correlation coefficients

varied (Figure 3A). This indicates that these loci have been consistently selected during breeding.

For the three traits prioritized in Chinese breeding (high 100-seed weight, SMV resistance, high protein content), I_CN accumulated more beneficial alleles than I_Am. Conversely, for the three US-prioritized traits (high seed weight per plant, SCN resistance, high oil content), I_Am had more beneficial alleles (Figure 3B). Based on the distribution of favorable allele frequencies, we used a percentile-based criterion and defined the top 20% of loci with the largest frequency differences as region-specific beneficial alleles. Loci with higher frequency differences in I_CN relative to I_Am were defined as Chinese-specific, and vice versa. This yielded 60 Chinese-specific beneficial alleles (25 for 100-seed weight, four for SMV resistance, 31 for protein content) and 55 Americas-specific beneficial alleles (20 for seed weight per plant, three for SCN resistance, 32 for oil content) (Figure 3C). These specific alleles were enriched in genomic regions differentiating I_CN and I_Am, with some overlapping known functional genes, such as *E3*, *GmFATA1B*, and so on

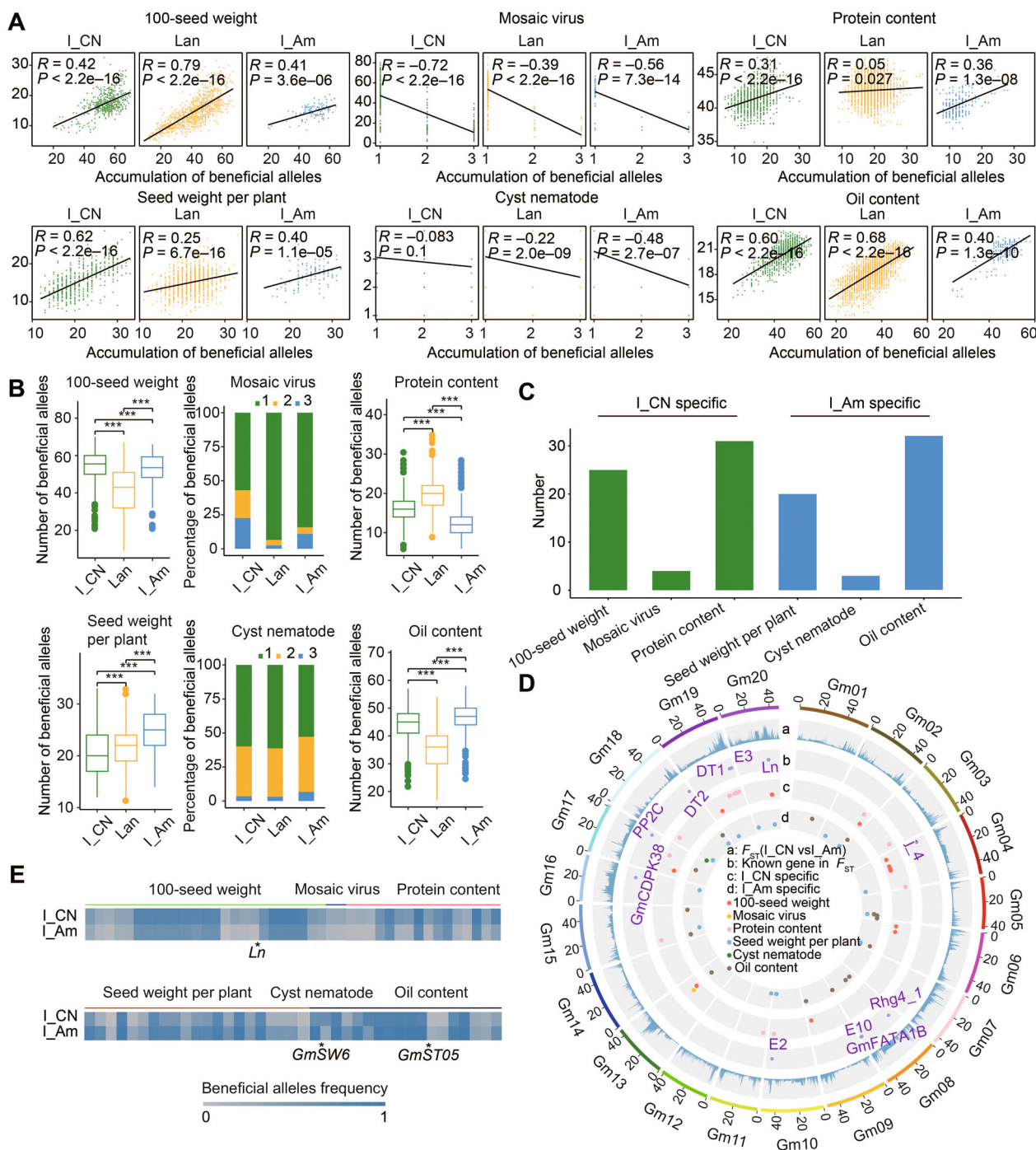


Figure 3. Differential alleles at trait-associated loci preferred by Chinese and American breeders

(A) Relationship between the number of beneficial alleles and phenotype values. The correlation and corresponding correlation test are indicated for each sub-population. Each dot represents one soybean accession in a certain environment. (B) Difference in the accumulation of beneficial alleles among sub-populations. One-tailed t -test significance: * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. (C) Number of specific beneficial alleles in I_CN and I_Am. (D) Genetic improvement differences between Chinese and American improved soybeans: (a) F_{ST} between I_CN and I_Am sub-populations, (b) Known genes located within the top 10% F_{ST} regions, (c) I_CN-specific loci. The frequency of beneficial alleles was calculated separately for I_CN and I_Am sub-populations. For each locus, the difference in frequency (I_CN-I_Am) was determined, with the top 20% quantile defined as CN-specific loci. Analogously, the top 20% quantile of the frequency difference (I_Am-I_CN) was defined as I_Am-specific loci (d). (E) The frequency of specific beneficial alleles in I_CN and I_Am.

(Xu et al., 2013; Cai et al., 2023b; Li et al., 2024). Most of the specific alleles are newly discovered and are more distributed in the highly divergent genome regions of soybeans between China and the Americas (Figure 3D). The frequency

distribution differences of specific alleles selected for breeding in I_CN and I_Am. These include several known functional genes associated with key agronomic traits. For instance, Ln (Jeong et al., 2012), which is associated with

leaflet shape and number of seeds per pod, has a beneficial allele genotype designated as C. The frequency of this genotype was 0.37 in the I_CN and 0.05 in the I_Am. Similarly, *GmST05/GmMFT* (Duan et al., 2022; Cai et al., 2023a), an oil content-associated gene, has higher-oil genotype A. The frequency of this genotype is 0.86 in the I_CN and 0.95 in the I_Am (Figure 3E). These specific beneficial alleles provide important breeding modification loci for the targeted improvement of soybeans.

Genome selection model for soybean improvement through simulated genetic modification

We evaluated the potential contribution of identified beneficial alleles to breeding using genome selection (GS). Four GS algorithms, including the linear model rrBLUP and three nonlinear machine-learning algorithms (SVR, RF, and LGB), were tested for accuracy and stability. In each of 100 iterations, 80% of the panel was randomly selected as the training population to predict the remaining 20% using seven allele sets: randomly selected, common I_CN and I_Am beneficial alleles, I_CN-specific, I_Am-specific, all I_CN beneficial alleles, all I_Am beneficial alleles, and a combined set. All four algorithms showed the same overall pattern as rrBLUP: For 100-seed weight and protein content, I_CN-specific alleles significantly improved prediction accuracy compared to I_Am-specific alleles. For grain weight per plant

Divergent selection in Chinese and American soybean

and oil content, I_Am-specific alleles performed better. Using all beneficial alleles from I_CN or I_Am improved accuracy for respective target traits, and the combined set achieved the highest accuracy, validating the effectiveness of the detected alleles (Figures 4A, S15). To further verify the robustness of these observations, we performed GS analyses separately for the Chinese and American subpopulations using the same four algorithms. Consistent conclusions were obtained across subpopulations, supporting the reliability of our findings.

We simulated the improvement of Chinese soybeans using Americas-specific beneficial alleles for grain weight per plant and oil content. Using the rrBLUP algorithm, genomic estimated breeding values (GEBVs) for I_CN were significantly lower than for I_Am. Substituting I_CN alleles with Americas-specific beneficial alleles progressively increased GEBVs, reaching I_Am levels after replacing 9 loci for seed weight per plant and 10 loci for oil content (Figure 4B). Similarly, using Chinese-specific alleles to improve American soybeans for 100-seed weight and protein content required replacing 9 and 4 loci, respectively, to reach I_CN levels (Figure 4B). Further substitutions continued to enhance performance. These results demonstrate the potential of targeted allele introgression via gene editing or genetic modification for rapid trait improvement.

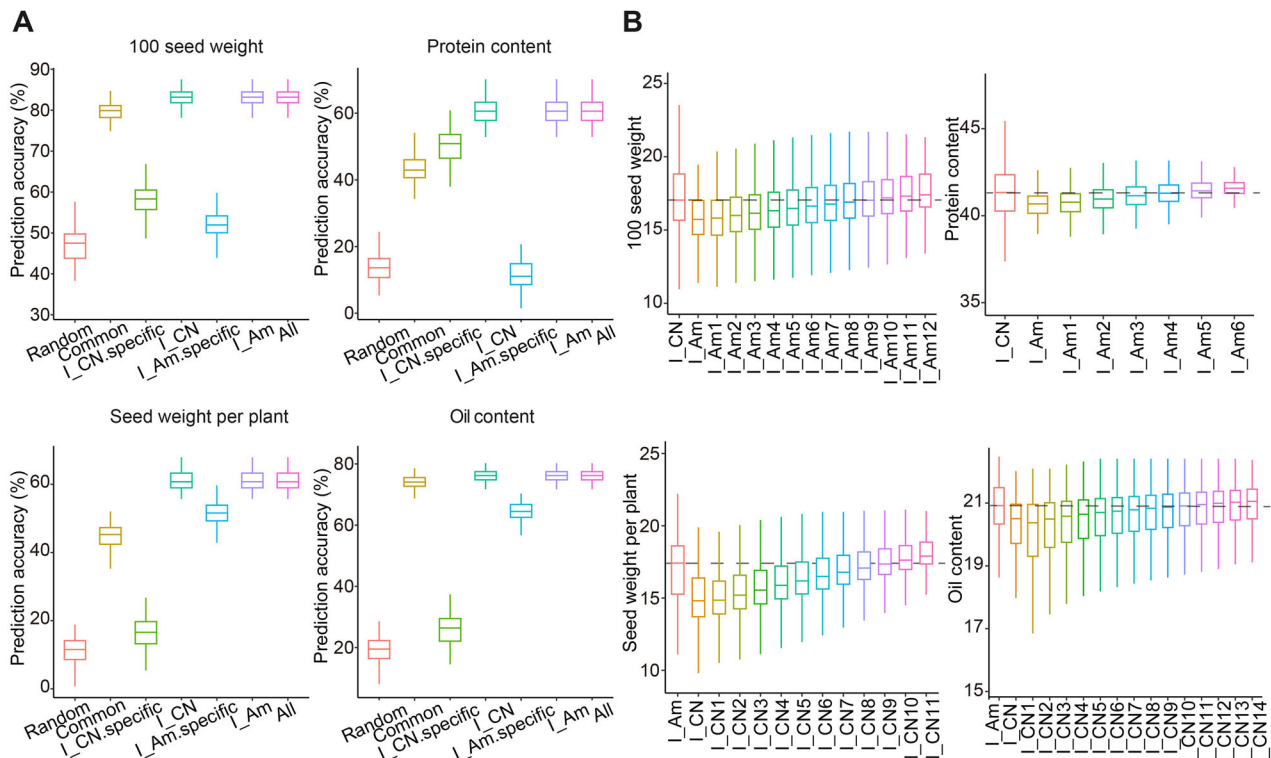


Figure 4. Construction of GS prediction model and gene-editing-enhanced prediction system based on beneficial alleles

(A) Genomic prediction accuracies for 100-seed weight, protein content, seed weight per plant, and oil content using different databases. (B) Based on I_CN beneficial alleles for 100-seed weight and Protein content, genomic estimated breeding values (GEBVs) for I_Am. Based on I_Am beneficial alleles for seed weight per plant and oil content, genomic estimated breeding values (GEBVs) for I_CN.

Soybean targeted improvement navigation system and its breeding application

Here, we present a Soybean Targeted Improvement Navigation System (STINS, <https://sfgb.rmbreeding.cn>) that integrates genomic selection and gene modification technologies to enable precision trait improvement. The data layer contains 2,377 soybean accessions, six key phenotypes, genome-wide genotypes, GWAS-associated loci, and China-/Americas-specific beneficial alleles across 256 functional loci. The algorithm layer includes genomic selection models (rrBLUP, SVR, RF, and LGB) for phenotype prediction and simulation of allelic substitution effects. The application layer describes user inputs and outputs, including target germplasm selection, trait selection, retrieval of favorable/non-favorable alleles, prioritization of editable loci, and predicted breeding values after simulated modification. Initially, we defined trait prioritization profiles for Chinese and American breeding programs, with the former emphasizing 100-seed weight, SMV resistance, and protein content, and the latter prioritizing seed weight per plant, SCN resistance, and oil content. A comprehensive database of breeding-beneficial alleles was established, incorporating both China- and Americas-specific beneficial alleles across 256 functional loci for six key traits. Central to the system is a genomic selection model trained on Chinese and American soybean populations, which not only predicts phenotypic values from genotype data but also simulates the outcomes of gene modification on target traits by: (i) selecting the sample that needs to be improved; (ii) selecting the phenotype that needs to be improved, which will output all the favorable allele and non-favorable allele information corresponding to this sample's phenotype;

(iii) selecting the loci that need to be edited; and (iv) predicting the breeding value of the simulated sample using genomic selection (Figures 5, S16).

To validate the practicality and accuracy of this navigation system, and to test whether its predicted elite alleles could confer measurable improvements across diverse genetic backgrounds, we selected two representative candidate genes with strong genetic effects and clear agronomic relevance (*GmSNAP18* for nematode resistance and *GmST05* for protein content) for simulation experiments and breeding verification (Liu et al., 2017; Duan et al., 2022; Cai et al., 2023a). We identified a lead SNP (A/G) in the promoter region of the *GmSNAP18* gene via GWAS (Figure 6A). The *GmSNAP18* gene, a core gene associated with resistance to soybean cyst nematode, is located at the *rhg1* site and encodes the α -soluble NSF attachment protein (Liu et al., 2017). The beneficial allele at this locus is *GmSNAP18*^{GG} (Figure 6B), and the frequency of the beneficial haplotype in American soybeans is significantly higher than that in Chinese soybeans (Figure 6C). The STINS system was employed to perform genetic simulation on the Chinese soybean variety Huachun6, which has a non-beneficial allele. It was predicted that the SCN race 3 female index (SCN3-FI) would significantly decrease after genetic modification, and this gene can serve as a target for resistance improvement of Chinese varieties (Figure 6D). Through overexpression of the disease-resistant haplotype, the resistance of the Chinese soybean variety Huachun6 to soybean cyst nematode was significantly enhanced in two independent overexpression lines (Figure 6E). We also identified a protein content-associated locus featuring an intronic lead SNP (A/C) in *GmST05/GmMFT* (Figure 6F), a known protein content gene, which we hypothesized as the causal gene for this locus

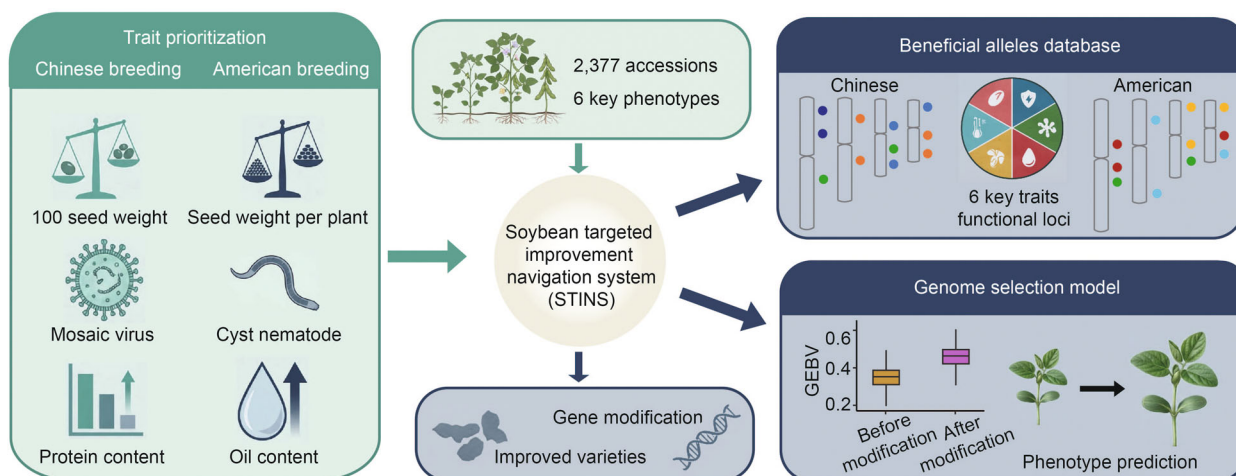


Figure 5. Construction of a soybean targeted improvement navigation system (STINS)

Traits prioritized differently during soybean improvement in China and Americas. Chinese breeders favored 100-seed weight, resistance to mosaic virus, and protein content, while American breeders favored seed weight per plant, resistance to soybean cyst nematode, and oil content. A breeding beneficial allele database of soybean was established, encompassing China- and Americas-specific beneficial alleles. Construct a genome selection model that predict the phenotypic values after gene modification. Based on the optimal gene modification target identified through the breeding navigation system, targeted genetic improvements were implemented using gene editing or genetic modification to develop a variety exhibiting enhanced target traits.

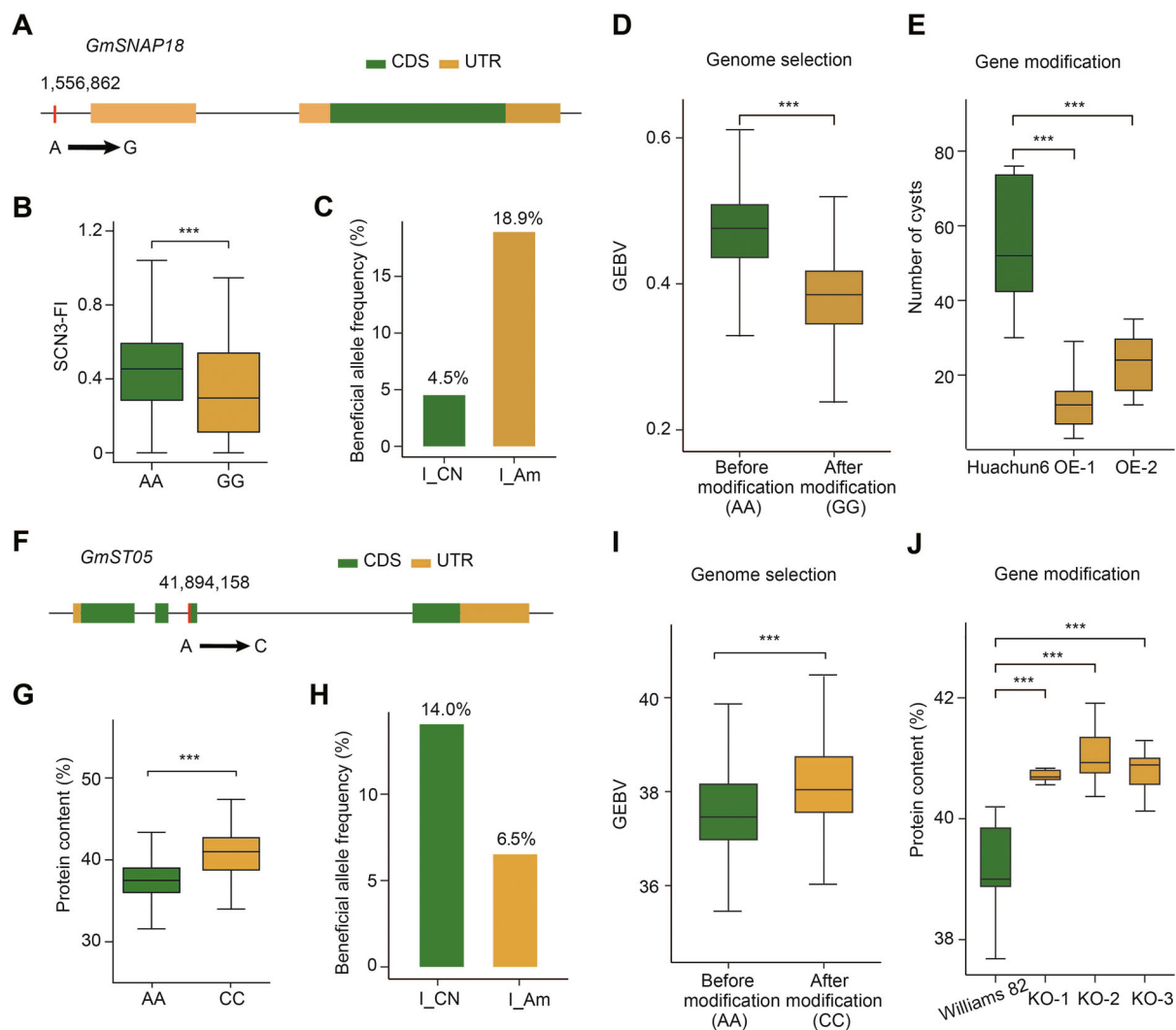


Figure 6. The breeding improvement application of STINS

(A) The gene structure and the target for breeding modification of *GmSNAP18*. (B) The impact of the two haplotypes of *GmSNAP18*. The GG genotype is the breeding-beneficial allele. (C) The gene frequencies of the beneficial allele of *GmSNAP18* in China and the Americas. (D) The Chinese soybean variety Huachun6 contains the non-beneficial allele AA of *GmSNAP18*. By simulating the predicted breeding value after gene modification through the STINS, the SCN3 incidence index of Huachun6 decreased significantly. (E) Comparison of the SCN incidence index of Huachun6 before and after the overexpression experiment. (F) The gene structure and the target for breeding modification of *GmST05*. (G) The impact of the two haplotypes of *GmST05*. The CC genotype is the breeding-beneficial allele. (H) The gene frequencies of the beneficial allele of *GmST05* in China and the Americas. (I) The American soybean variety Williams 82 contains the non-beneficial allele AA of *GmST05*. By simulating the predicted breeding value after gene modification through the STINS, the protein content of Williams 82 increased significantly. (J) Comparison of protein content of Williams 82 before and after gene modification experiment.

(Duan et al., 2022; Cai et al., 2023a). The beneficial *GmST05*^C allele was associated with higher protein content but occurred at low frequency in I_Am (Figure 6G). In contrast, I_CN has a higher frequency of the beneficial haplotype as a result of selective breeding (Figure 6H). The navigation system identified *GmST05* as an optimal target and predicted a significant protein content improvement for American soybean following gene editing (Figure 6I). The American cultivar Williams 82 possesses a non-beneficial haplotype *GmST05*^A. Given that the GWAS-identified intronic locus in *GmST05* is linked with coding-region variants, we employed CRISPR/Cas9 to target its first exon in Williams 82, generating a frameshift mutation that resulted in premature translation termination (Zhang et al., 2026). Consequently, knockout of this unbeneficial gene led to

significantly elevated protein contents in the three deletion mutant lines, validating both the predictive power and practical utility of our system (Figure 6J). Together, the *GmST05* editing experiment and the auxiliary functional evaluation of *GmSNAP18* provide proof-of-concept experimental support for the utility of STINS in prioritizing candidate targets for soybean improvement.

DISCUSSION

The intensive breeding efforts in China and the Americas, operating under distinct agro-ecological and market pressures, have created two of the world's most important

soybean gene pools. Our study, by integrating high-density genomic data from 2,377 accessions with multi-environment phenotyping for six key traits, provides a genome-wide dissection of the genetic divergence between these two germplasms. We move beyond descriptive population structure to explicitly link genomic variation to divergent trait selection, identify the functional loci and alleles underpinning these differences, and demonstrate a practical framework for their utilization in future breeding. While the phenotypic divergence toward higher protein in Chinese soybeans and higher oil in American soybeans has long been appreciated, our work provides a genome-wide characterization of the beneficial alleles driving this divergence, identifies numerous novel trait-associated loci, establishes a simulation-guided reciprocal improvement framework to deploy these alleles for soybean breeding, and offers a roadmap for targeted crop improvement.

Genomic footprints of divergent breeding histories and objectives

Our population structure analysis confirms that Chinese and American improved cultivars, despite sharing a common ancestral origin in Chinese landraces, have formed distinct genomic clusters (Figure 1A). The clear separation of I_CN and I_Am subpopulations at $K = 8$ underscores the powerful effect of decades of independent breeding. The genetic proximity of I_Am to the North Chinese landraces (L_NR) is consistent with historical records of germplasm introduction (Qiu et al., 2011) and suggests that the initial genetic base for American breeding was relatively narrow.

The observed reduction in genetic diversity from landraces (Lan) to improved cultivars in both China (I_CN) and the Americas (I_Am), with I_Am exhibiting the lowest diversity (Figure 1B, C), is a classic signature of selection and genetic bottleneck (Fu, 2015). This indicates that while both breeding programs have selected on a subset of the landrace diversity, the American bottleneck may have been more severe, potentially due to its narrower founding population and intensive selection for adaptation to the Midwest's highly productive but relatively uniform environments (Hyten et al., 2006). In contrast, the higher retained diversity in Chinese improved cultivars can be attributed to multiple factors: (i) A broader parental diversity, with breeding programs actively utilizing a wide spectrum of landrace and exotic germplasm; (ii) China's highly complex and variable geography and climate—spanning from the cool, high-latitude north to subtropical, low-latitude south—which imposed selective pressures for a broad genetic base to accommodate widely divergent photoperiods and local climates in this photoperiod-sensitive short-day plant (Zhou et al., 2015); (iii) diversified breeding objectives in China, which simultaneously aim to meet the demands for high oil content for crushing as well as superior protein quality and functionality for diverse traditional soyfood processing. This multipurpose breeding strategy has necessitated and subsequently preserved a wider range of genetic variation.

The phenotypic divergence we observed is a direct reflection of these genomic differences and divergent breeding goals. The stronger emphasis on 100-seed weight in China aligns with the culinary preference for large-seeded soybeans for products like tofu and soymilk (Grieshop and Fahey, 2001; Xu et al., 2022). Conversely, the Americas' focus on seed weight per plant and oil content is driven by the demand for vegetable oil and animal feed (Wilcox, 2001). The differential improvement in resistance to SMV and SCN perfectly mirrors the geographic prevalence of these pathogens (Allen et al., 2017; Lin et al., 2022), demonstrating how pathogen pressure directly shapes the allelic landscape of major resistance genes. In summary, the genomic and phenotypic profiles of Chinese and American soybean germplasms have been uniquely shaped by a combination of local environments, market demands, and stresses.

The genetic architecture of key agronomic traits and the reservoir of unexploited alleles

Our GWAS successfully identified a large set of 261 loci trait-associated with six agronomic traits, validating known genes and uncovering a wealth of novel candidates (Figure 2A). The identification of loci harboring known functional genes, such as *GmPRR3b* for adaptation (Li et al., 2020) and *Dt1* for plant architecture (Liu et al., 2010), confirms the robustness of our association panel and methodology. More importantly, the discovery of 223 novel loci, many of which have PVE values comparable to known loci (Figure 2C, D), significantly expands the genetic toolkit available to breeders. The novel loci with high PVE suggest that there remain substantial genetic gains to be made by mining this “reservoir” of unexploited variation. Our 223 newly detected loci provide rich candidate resources for soybean functional genomics. Although these loci are supported by population-scale evidence, further experimental validation via gene editing or transgenics will be prioritized in our future research to uncover their molecular functions.

Our result implies pleiotropic effects on bottom pod height, section number, and 100-seed weight, a phenomenon increasingly recognized in plants (Dong et al., 2022). The differential haplotype frequency of this locus between I_CN and I_Am provides a clear genetic explanation for part of the observed phenotypic difference in seed weight. Similarly, the discovery of novel loci for protein and oil content with biased haplotype frequencies and expression patterns offers concrete targets for understanding and manipulating these complex quality traits (Warrington et al., 2015).

Breeding-driven allelic selection and its implications for cross-germplasm improvement

Our findings, which pinpoint specific beneficial alleles that were preferentially accumulated under divergent breeding goals in Chinese and American germplasms, are strongly supported by a recent, independent genomic atlas of 8,105 soybeans. That study also underscores how regional selection pressures have sculpted the soybean genome during

its global dissemination (Zhu et al., 2025). Our work moves beyond descriptive genomics to define the very unit of selection—the “region-specific beneficial allele”—and quantitatively demonstrate its potential for cross-germplasm improvement. A central finding of our study is the systematic divergence in beneficial allele accumulation for traits under divergent selection (Figure 3A, B). The concept of “region-specific beneficial alleles” (Figure 3C) quantitatively captures the genomic signature of decades of targeted selection. The fact that these specific alleles are often fixed or nearly fixed in one germplasm while being rare in the other (Figure 3E) highlights a significant opportunity. It suggests that each germplasm possesses unique, valuable alleles that the other lacks, creating a substantial potential for reciprocal genetic improvement.

Our genomic selection simulations provide powerful empirical support for this notion (Figure 4A). The significant boost in prediction accuracy for Americas-priority traits (seed weight per plant, oil content) when using Americas-specific alleles, and vice-versa for Chinese-priority traits, validates the functional value of these allele sets. Most importantly, the “breeding by editing” simulation demonstrates that targeted introgression of a relatively small number of these alleles can theoretically bridge the phenotypic gap between I_CN and I_Am (Figure 4B). These findings enable us to formulate an actionable, simulation-guided breeding strategy for soybean improvement.

For instance, the low frequency of beneficial alleles for seed weight per plant and oil content in I_CN represents a major “allelic debt” that limits the yield and quality potential of Chinese cultivars. Conversely, the deficiency of certain 100-seed weight alleles in I_Am could leave American germplasm vulnerable if market demands shift. Our identified allele sets provide a precise shopping list for addressing these weaknesses. The traditional approach of wide cross-breeding is often hampered by linkage drag (Huang et al., 2023). However, modern technologies like gene editing and overexpression now allow for the precise introduction of these beneficial alleles without disrupting otherwise elite genetic backgrounds (Bortesi and Fischer, 2015), dramatically accelerating the breeding cycle. This was demonstrated by editing the *GmST05* to enhance protein content in an American elite cultivar.

A paradigm shift toward knowledge-driven design breeding

The collective findings of this study argue for a paradigm shift from phenotype-based, iterative breeding toward a knowledge-driven design approach. The Soybean Targeted Improvement Navigation System (STINS) we present, which integrates genomic selection models with a database of breeding-beneficial alleles and proof-of-concept functional evaluation of prioritized targets, is a concrete step in this direction. For *GmSNAP18*, overexpression of the resistant haplotype was used as an initial functional evaluation following the previously established complementation strategy

Divergent selection in Chinese and American soybean

for this gene (Liu et al., 2017). This approach allowed us to assess the resistance-enhancing potential of the STINS-prioritized haplotype in a common genetic background, although direct validation of the predicted nucleotide substitution will require endogenous allele replacement. This integrated resource empowers breeders to make informed decisions on parental selection and to design optimal haplotypes for specific environments and markets (Hammer et al., 2020).

Future work should focus on the functional characterization of the high-value novel loci identified here, particularly those with large effects. Furthermore, exploring epistatic interactions among these beneficial alleles will be crucial for designing optimal genetic combinations rather than simply stacking individual loci (Gu et al., 2023). As the climate changes and market demands evolve, the ability to rapidly and precisely reconfigure elite cultivars using a global pool of beneficial alleles, as demonstrated here between Chinese and American germplasm, will be paramount for ensuring global food and feed security.

MATERIALS AND METHODS

Planting and phenotyping

1,910 soybean landraces and cultivars of the previously studied panel (Li et al., 2023) were phenotyped for agronomy and resistance traits. The agronomy traits were phenotyped at 11 different locations, and for 2–3 years at each location from 2017 to 2019. Each soybean accession was planted in a 1.8 × 0.8 m plot with two rows and a spacing of 10 cm between seedlings in each row. SCN and SMV resistance were both assessed in Harbin (45°45' N, 126°41' E) in 2017 and 2018.

Six traits were phenotyped: 100-seed weight, seed weight per plant, seed oil content, protein content, SCN resistance, and SMV resistance. Yield-related traits (100-seed weight, seed weight per plant) and quality traits (oil and protein content) were evaluated across 10 geographic locations in China from 2017 to 2019, with data collected over 2–3 years per location. The environments included Harbin (2017–2019), Gongzhuling (2018–2019), Tonghua (2017–2019), Shijiazhuang (2018–2019), Liaocheng (2017–2019), Xuzhou (2018–2019), Nanjing (2018–2019), Hefei (2018–2019), Wuhan (2018–2019), Mengcheng (2017–2019) and Nanchang (2017–2019). Phenotyping for these traits followed the methods described by Qiu et al. (2006). The two biotic stress resistance traits were assessed in Harbin in 2017 and 2018. Resistance to SCN (race 3) was evaluated using the method of Li et al. (2016), and resistance to SMV (strain SC3) was scored using a previously established metric.

Trait BLUP values estimation

The BLUP values were estimated across multiple years using the mixed linear model with the lme4 package in R (v3.2.2). The formula was as follows: $Y = \mu + \text{Line} + \text{Loc} + (\text{Line} \times \text{Loc}) + \text{Rep}(\text{Loc}) + \varepsilon$. Where Y represents the

phenotype, Line and Loc denote the respective random effects of the genotype and the environment, Line $\times$ Loc is the interaction effect between the genotype and environment, Rep (Loc) is the random effect of replication nested in the environment, and ϵ represents the random residual error effect. The BLUP values were used to analyze the phenotypic distribution in Lan, I_CN, and I_Am populations. The significance of differences for phenotypic data was tested using a *t*-test at the significance levels of 0.05 or 0.01.

Genotyping

The raw sequencing of pair-end reads was trimmed to remove low-quality reads as below: (i) reads with $\geq 10\%$ unidentified nucleotides (N); (ii) reads with > 10 nt aligned to the adaptor, with $\leq 10\%$ mismatches allowed; (iii) reads with $> 50\%$ bases having Phred quality < 5 ; and (iv) putative PCR duplicates generated via PCR amplification in the library construction process.

The remaining high-quality reads were mapped to the soybean reference genome (Williams 82 assembly V4.0) using Burrows-Wheeler Aligner software (mem -t 4 -k 32 -M) (Li and Durbin, 2009). In order to reduce mismatches generated by PCR amplification before sequencing, duplicated reads were removed using SAMtools (v0.1.1) (Li and Handsaker, Wysoker, et al., 2009). After alignment, the genomic variants were identified with the HaplotypeCaller module and the GVCF model by Genome Analysis Toolkit (GATK) software (McKenna et al., 2010). All of the GVCF files were then merged. A raw population genotype file with the SNPs was created in the HaplotypeCaller module, which was filtered with the following parameters: minor allele frequency (MAF) $\geq 1\%$, with miss rate $\leq 20\%$.

Phylogenetic tree and population structure

248,019 representative SNPs were used for phylogenetic tree construction and population structure analysis. An individual-based neighbor-joining (NJ) tree was constructed based on the p-distance using the software TreeBest (v1.9.2) (Vilella et al., 2009) with 1,000 bootstrap replications. The population genetic structure was performed with ADMIXTURE (v1.23) (Alexander et al., 2009), with the preset number of assumed genetic clusters *K* and 10,000 iterations for each run. Principal components analysis (PCA) was conducted to evaluate genetic structure using PLINK software (v 1.9) (Purcell et al., 2007).

Selective sweep analysis

To identify genome-wide selective sweeps, we scanned the genome in 10 kb sliding scales with a step size of 5 kb and calculated the genome-wide distribution of fixation index (F_{ST}) and the genetic diversity (π) by vcfTools implementing the Weir and Cockerham method, respectively. The regions with the diversity ratio values and F_{ST} scores in the top 5% were considered to have been subjected to selective sweeps, respectively.

Linkage disequilibrium analysis

To estimate and compare the pattern of linkage disequilibrium (LD) for I_CN, Lan, and I_Am accessions, the squared correlation coefficient (r^2) between pairwise SNPs was computed using the software LDdecay. Parameters in the program were

set as: “-MaxDist 500 -MAF 0.05 -Miss 0.2.” The average r^2 value was calculated for pairwise markers in a 500 kb window and averaged across the whole genome.

Genome-wide association study (GWAS)

For each single GWAS, the SNPs were filtered with a MAF of $> 5\%$ among the studied individuals. The GWAS was conducted using the fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al., 2016) with population structure controlled by the first four components from SNP PCA results. The resulting *P*-values were adjusted with Bonferroni correction at a level of $\alpha = 0.05$ to define trait-associated SNPs.

For the estimation of PVE by SNPs, the GREML method (REstricted Maximum Likelihood) in GCTA software (version 1.24.2) (Yang et al., 2011) was used to estimate the ratio of genetic variance to phenotypic variance of 12 traits using genome-wide markers.

Definition of beneficial alleles

All peak SNPs with $-\log_{10}(P) \geq 8.0$ for BLUP-GWAS or $-\log_{10}(P) \geq 8.0$ in at least 2 years were used to identify beneficial alleles. Firstly, we calculated the effect of each allele on phenotype in soybean cultivars and removed the false-positive loci (*t*-test, $P > 0.05$). Secondly, we defined breeding beneficial alleles at the associated loci based on at least one of six phenotypic characteristics, that are, positive effect to increase 100-seed weight, seed weight per plant, protein content or oil content, improve resistance to mosaic virus or resistance to soybean cyst nematode. For each locus, the difference in frequency (I_CN - I_Am) was determined, with the top 20% quantile defined as CN-specific loci. Analogously, the top 20% quantile of the frequency difference (I_Am - I_CN) was defined as I_Am-specific loci.

Genome selection and STINS

The beneficial alleles and the traits data were used for genomic prediction with four models, including rrBLUP (Endelman, 2011), SVR, RF, and LGB. rrBLUP is a classical linear model, assuming marker effects follow a normal distribution and primarily captures additive genetic effects. SVR, RF, and LGB are nonlinear machine-learning algorithms that can capture non-additive genetic effects such as epistasis. 80% of the panel was randomly selected as a training population to predict the genomic estimated phenotype value of the remaining 20% on seven allele datasets: the randomly selected common I_CN and I_Am beneficial allele, I_CN specific beneficial alleles, I_Am specific beneficial alleles, I_CN beneficial alleles, I_Am beneficial alleles, and the combined beneficial alleles. This was conducted using 100 iterations of cross-validation, and prediction accuracy was based on Pearson correlation between the Genomic estimated breeding values (GEBVs) and the observed phenotype values. GEBVs were predicted by rrBLUP, SVR, RF, and LGB, respectively.

For each target trait, we first selected region-specific beneficial loci that exceeded a predefined effect-size threshold based on PVE (PVE $> 2\%$ for most traits; for oil

content, a lower threshold of 0.3% was used because the individual effect sizes of oil-related loci were generally smaller). We then focused only on loci at which the target cultivar carried the non-beneficial allele. These candidate loci were ranked from large to small effect and were progressively substituted with the counterpart population-specific beneficial alleles. After each cumulative round of substitution, the phenotype was re-predicted using the rrBLUP model. Thus, the simulation did not replace all region-specific alleles indiscriminately; it replaced only non-beneficial loci in the target genetic background, in a stepwise manner ordered by predicted contribution.

We demonstrated the practical utility of STINS using 16NF1005_1006 as a case study. 16NF1005_1006 was from I_CN and showed less resistance to soybean cyst nematode compared to I_Am. Our goal was to enhance its resistance. We applied the STINS to achieve this. First, we selected 16NF1005_1006 as a sample and soybean cyst nematode (SCN3-FI) as the phenotype that needed to be improved, and all the favorable and non-favorable allele information was output. Second, we selected loci (non-favorable alleles) that needed to be edited. Finally, the breeding value was predicted using our pre-trained genomic selection models. The results show that the SCN3-FI resistance value of 16NF1005_1006 decreased from 0.5381 to 0.485, a reduction of 9.87%, indicating that the SCN3-FI resistance of 16NF1005_1006 significantly improved.

Plasmid construction and plant transformation

To generate CRISPR/Cas9-engineered *st05* mutants, a single sgRNA was designed using the CRISPR-P online platform (<http://crispr.hzau.edu.cn>). The sgRNA fragment was cloned into the *Bsa*I site of the pBSN401 plasmid (Xing et al., 2014). A CRISPR/Cas9 expression vector was then constructed, in which the Cas9 coding sequence was driven by the 35S promoter, while sgRNA was controlled by the Arabidopsis U6 promoter. The construct was introduced into *Agrobacterium tumefaciens* strain EHA105 and transformed into the soybean cultivar Williams 82 via the cotyledon-node method. Homozygous mutants were obtained by self-pollination of candidate lines and confirmed by Sanger sequencing. The PCR primers are: forward, 5'-GCAACCACCAGAACTCTC-3'; reverse, 5'-ACCAATGTATCCACTCGCGC-3'. To generate *GmSNAP18* overexpression (OE) soybean lines, the full-length coding sequence (CDS) of the resistant haplotype of *GmSNAP18* (Bayless et al., 2016) was synthesized and cloned into the plant binary vector pCAMBIA3301 under the control of the 35S promoter. The resulting overexpression construct contained the bar gene as a selectable marker, conferring resistance to phosphinothricin. All recombinant plasmids were verified by Sanger sequencing prior to transformation. The confirmed constructs were introduced into *Agrobacterium tumefaciens* strain EHA105, and soybean transformation was performed in the variety Huachun6 using the cotyledon-node method. Positive T₀ transgenic plants were self-pollinated to obtain homozygous T₂/T₃ overexpression lines. The transcript levels of *GmSNAP18* in

Divergent selection in Chinese and American soybean

independent OE lines were quantified by RT-qPCR, and lines exhibiting stable and significantly elevated expression were selected for subsequent analyses. The qPCR primers are: forward, 5'-ATGTCGATGCTGCACATTGC-3'; reverse, 5'-GCAGCAAACCTGGGCAACTTT-3'.

Data availability statement

The raw genome re-sequencing data of the worldwide derived soybean panel were previously deposited: NCBI Sequence Read Archive (SRA) under accession numbers PRJNA552939 (Li et al., 2020) and PRJNA681974 (Qin et al., 2023); and Genome Sequence Archive (GSA) of the National Genomics Data Center (NGDC), BIG Data Center, under accession number PRJCA002030 (Liu et al., 2020).

ACKNOWLEDGEMENTS

This work was supported by the Beijing Rural Revitalization Agricultural Science and Technology Project (NY2401030000), the National Key Research and Development Program of China (2021YFD1201600), the earmarked fund for CARS (CARS-04-PS01), the Agricultural Science and Technology Innovation Program (ASTIP), and Bio-breeding Laboratory of Anhui Province (2025SWYZ0412).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

L.Q., along with Y.-H.L., X.B.W., and C.J., designed and coordinated the project. L.Y., C.Z., H.Q., G.X., G.H., L.L., Q.W., W.H., X.W., Y.W., L.M., Y.F.L., and Y.T. collected the phenotyping data. C.J., D.L., Z.Z., and X.J. contributed to the data analyses. C.J., X.F., and T.Z. designed and deployed a soybean targeted improvement navigation system. H.Z. and X.Q.W. contributed to transgenic and gene editing experiments. C.J. and D.L. wrote the draft manuscript. L.Q., X.B.W., and Y.-H.L. revised and completed the manuscript. All authors read and approved the final manuscript.

Edited by: Qingxin Song, Institute of Genetics and Developmental Biology, CAS, China

Received Jun. 11, 2026; **Accepted** Sep. 20, 2026

OO: OnlineOpen

REFERENCES

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**: 1655–1664.

- Allen, T.W., Bradley, C.A., Sisson, A.J., Byamukama, E., Chilvers, M.I., Coker, C.M., Collins, A.A., Damicone, J.P., Dorrance, A.E., Dufault, N.S., et al. (2017). Soybean yield loss estimates due to diseases in the United States and Ontario, Canada, from 2010 to 2014. *Plant Health Prog.* **18**: 19–27.
- Bayless, A.M., Smith, J.M., Song, J., McMinn, P.H., Teillet, A., August, B.K., and Bent, A.F. (2016). Disease resistance through impairment of α -SNAP-NSF interaction and vesicular trafficking by soybean *Rhg1*. *Proc. Natl. Acad. Sci. U.S.A.* **113**: E7375–E7382.
- Bortesi, L., and Fischer, R. (2015). The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol. Adv.* **33**: 41–52.
- Cai, Z., Xian, P., Cheng, Y., Yang, Y., Zhang, Y., He, Z., Xiong, C., Guo, Z., Chen, Z., Jiang, H., et al. (2023b). Natural variation of *GmFATA1B* regulates seed oil content and composition in soybean. *J. Integr. Plant Biol.* **65**: 2368–2379.
- Cai, Z., Xian, P., Cheng, Y., Zhong, Y., Yang, Y., Zhou, Q., Lian, T., Ma, Q., Nian, H., and Ge, L. (2023a). MOTHER-OF-FT-AND-TFL1 regulates the seed oil and protein content in soybean. *New Phytol.* **239**: 905–919.
- Dong, L., Cheng, Q., Fang, C., Kong, L., Yang, H., Hou, Z., Li, Y., Nan, H., Zhang, Y., Chen, Q., et al. (2022). Parallel selection of distinct *Tof5* alleles drove the adaptation of cultivated and wild soybean to high latitudes. *Mol. Plant* **15**: 308–321.
- Dong, L., Fang, C., Cheng, Q., Su, T., Kou, K., Kong, L., Zhang, C., Li, H., Hou, Z., Zhang, Y., et al. (2021). Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. *Nat. Commun.* **12**: 5445.
- Duan, Z., Zhang, M., Zhang, Z., Liang, S., Fan, L., Yang, X., Yuan, Y., Pan, Y., Zhou, G., Liu, S., et al. (2022). Natural allelic variation of *GmST05* controlling seed size and quality in soybean. *Plant Biotechnol. J.* **20**: 1807–1818.
- Endelman, J.B. (2011). Ridge regression and other kernels for genomic selection with R Package rrBLUP. *Plant Genome* **4**: 250–255.
- Fu, Y. (2015). Understanding crop genetic diversity under modern plant breeding. *Theor. Appl. Genet.* **128**: 2131–2142.
- Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D., Song, Q., Pantalone, V.R., Song, B., Yu, D., et al. (2022). *POWR1* is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nat. Commun.* **13**: 3051.
- Grieshop, C., and Fahey, G. (2001). Comparison of quality characteristics of soybeans from Brazil, China, and the United States. *J. Agric. Food Chem.* **49**: 2669–2673.
- Gu, Z., Gong, J., Zhu, Z., Li, Z., Feng, Q., Wang, C., Zhao, Y., Zhan, Q., Zhou, C., Wang, A., et al. (2023). Structure and function of rice hybrid genomes reveal genetic basis and optimal performance of heterosis. *Nat. Genet.* **55**: 1745–1756.
- Hammer, G.L., McLean, G., van Oosterom, E., Chapman, S., Zheng, B., Wu, A., Doherty, A., and Jordan, D. (2020). Designing crops for adaptation to the drought and high-temperature risks anticipated in future climates. *Crop Sci.* **60**: 605–621.
- Hu, D., Li, X., Yang, Z., Liu, S., Hao, D., Chao, M., Zhang, J., Yang, H., Su, X., Jiang, M., et al. (2022). Downregulation of a gibberellin 3 β -hydroxylase enhances photosynthesis and increases seed yield in soybean. *New Phytol.* **235**: 502–517.
- Huang, K., Jahani, M., Gouzy, J., Legendre, A., Carrere, S., Lázaro-Guevara, J.M., González Segovia, E.G., Todesco, M., Mayonade, B., Rodde, N., et al. (2023). The genomics of linkage drag in inbred lines of sunflower. *Proc. Natl. Acad. Sci. U.S.A.* **120**: e2205783119.
- Hyten, D.L., Song, Q., Zhu, Y., Choi, I.Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., and Cregan, P.B. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. U.S.A.* **103**: 16666–16671.
- Jeong, N., Suh, S.J., Kim, M.H., Lee, S., Moon, J.K., Kim, H.S., and Jeong, S.C. (2012). *Ln* is a key regulator of leaflet shape and number of seeds per pod in soybean. *Plant Cell* **24**: 4807–4818.
- Jeong, S.C., Moon, J.K., Park, S.K., Kim, M.S., Lee, K., Lee, S.R., Jeong, N., Choi, M.S., Kim, N., Kang, S.T., et al. (2019). Genetic diversity patterns and domestication origin of soybean. *Theor. Appl. Genet.* **132**: 1179–1193.
- Li, C., Li, Y., Li, Y., Lu, H., Hong, H., Tian, Y., Li, H., Zhao, T., Zhou, X., Liu, J., et al. (2020). A domestication-associated gene *GmPRR3b* regulates the circadian clock and flowering time in soybean. *Mol. Plant* **13**: 745–759.
- Li, D., Wang, Q., Tian, Y., Lyv, X., Zhang, H., Hong, H., Gao, H., Li, Y., Zhao, C., Wang, J., et al. (2024). TWAS facilitates gene-scale trait genetic dissection through gene expression, structural variations, and alternative splicing in soybean. *Plant Commun.* **5**: 101010.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- Li, Y., Qin, C., Wang, L., Jiao, C., Hong, H., Tian, Y., Li, Y., Xing, G., Wang, J., Gu, Y., et al. (2023). Genome-wide signatures of the geographic expansion and breeding of soybean. *Sci. China Life Sci.* **66**: 350–365.
- Li, Y., Shi, X., Li, H., Reif, J.C., Wang, J., Liu, Z., He, S., Yu, B., and Qiu, L. (2016). Dissecting the genetic basis of resistance to soybean cyst nematode combining linkage and association mapping. *Plant Genome* **9**: 1–11.
- Liang, Q., Chen, L., Yang, X., Yang, H., Liu, S., Kou, K., Fan, L., Zhang, Z., Duan, Z., Yuan, Y., et al. (2022). Natural variation of *Dt2* determines branching in soybean. *Nat. Commun.* **13**: 6429.
- Lin, F., Chhapekar, S.S., Vieira, C.C., Da Silva, M.P., Rojas, A., Lee, D., Liu, N., Pardo, E.M., Lee, Y.C., Dong, Z., et al. (2022). Breeding for disease resistance in soybean: A global perspective. *Theor. Appl. Genet.* **135**: 3773–3872.
- Liu, B., Watanabe, S., Uchiyama, T., Kong, F., Kanazawa, A., Xia, Z., Nagamatsu, A., Arai, M., Yamada, T., Kitamura, K., et al. (2010). The soybean stem growth habit gene *Dt1* is an ortholog of Arabidopsis TERMINAL FLOWER1. *Plant Physiol.* **153**: 198–210.
- Liu, S., Kandoth, P.K., Lakhssassi, N., Kang, J., Colantonio, V., Heinz, R., Yeckel, G., Zhou, Z., Bekal, S., Dapprich, J., et al. (2017). The soybean *GmSNAP18* gene underlies two types of resistance to soybean cyst nematode. *Nat. Commun.* **8**: 14822.
- Liu, X., Huang, M., Fan, B., Buckler, E.S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **12**: e1005767.
- Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G., Zhang, H., Liu, Z., Shi, M., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell* **182**: 162–176.
- Lu, S., Zhao, X., Hu, Y., Liu, S., Nan, H., Li, X., Fang, C., Cao, D., Shi, X., Kong, L., et al. (2017). Natural variation at the soybean *J* locus improves adaptation to the tropics and enhances yield. *Nat. Genet.* **49**: 773–779.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**: 1297–1303.
- Miao, L., Yang, S., Zhang, K., He, J., Wu, C., Ren, Y., Gai, J., and Li, Y. (2020). Natural variation and selection in *GmSWEET39* affect soybean seed oil content. *New Phytol.* **225**: 1651–1666.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**: 559–575.

- Qin, C., Li, Y., Li, D., Zhang, X., Kong, L., Zhou, Y., Lyu, X., Ji, R., Wei, X., Cheng, Q., et al. (2023). *PH13* improves soybean shade traits and enhances yield for high-density planting at high latitudes. *Nat. Commun.* **14**: 6813.
- Qiu, L., Chang, R., Liu, Z., Guan, R., and Li, Y. (2006). *Descriptors and data standard for soybean (Glycine spp.)*. Beijing (in Chinese): China Agriculture Press.
- Qiu, L., Chen, P., Liu, Z., Li, Y., Guan, R., Wang, L., and Chang, R. (2011). The worldwide utilization of the Chinese soybean germplasm collection. *Plant Genet. Resour.* **9**: 109–122.
- Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178–183.
- Tsubokura, Y., Watanabe, S., Xia, Z., Kanamori, H., Yamagata, H., Kaga, A., Katayose, Y., Abe, J., Ishimoto, M., and Harada, K. (2014). Natural variation in the genes responsible for maturity loci E1, E2, E3 and E4 in soybean. *Ann. Bot.* **113**: 429–441.
- Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**: 327–335.
- Warrington, C.V., Abdel-Haleem, H., Hyten, D.L., Cregan, P.B., Orf, J.H., Killam, A.S., Bajjalieh, N., Li, Z., and Boerma, H.R. (2015). QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. *Theor. Appl. Genet.* **128**: 839–850.
- Wilcox, J.R. (2001). Sixty years of improvement in publicly developed elite soybean lines. *Crop Sci.* **41**: 1711–1716.
- Xing, H., Dong, L., Wang, Z., Zhang, H., Han, C., Liu, B., Wang, X., and Chen, Q. (2014). A CRISPR/Cas9 toolkit for multiplex genome editing in plants. *BMC Plant Biol.* **14**: 327.
- Xu, C., Wu, T., Yuan, S., Sun, S., Han, T., Song, W., and Wu, C. (2022). Can soybean cultivars with larger seed size produce more protein, lipids, and seed yield? A meta-analysis. *Foods* **11**: 4059.
- Xu, M., Xu, Z., Liu, B., Kong, F., Tsubokura, Y., Watanabe, S., Xia, Z., Harada, K., Kanazawa, A., Yamada, T., et al. (2013). Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. *BMC Plant Biol.* **13**: 91.
- Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**: 76–82.
- Zhang, H., Tian, Y., Zhao, C., Shan, Z., Yang, L., Gu, Y., Gao, H., Liu, Z., Li, D., Ge, T., et al. (2026). An underutilized genetic component in the regulation of soybean seed weight. *Nat. Commun.* **17**: 6639.

- Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**: 408–414.
- Zhu, Z., Wang, Y., Liu, S., Wang, S., Li, J., Fang, C., Liu, Y., Yang, X., Tian, D., Song, S., et al. (2025). Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean. *Cell* **188**: 6519–6535.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article: <http://onlinelibrary.wiley.com/doi/10.1111/jipb.70421/supinfo>

Dataset S1. Summary of the geographic origin and re-sequencing information of soybean accessions

Dataset S2. Summary of SNP hits in the GWAS analysis

Dataset S3. Summary of known and novel SNPs with PVE

Dataset S4. Summary of favorable alleles

Figure S1. The whole-genome SNP density distribution

Figure S2. Location of phenotypic data used in GWAS

Figure S3. GWAS Manhattan plots and quantile-quantile (Q-Q) plots for 100-seed weight

Figure S4. GWAS Manhattan plots and Q-Q plots for cyst nematode

Figure S5. GWAS Manhattan plots and Q-Q plots for mosaic virus

Figure S6. GWAS Manhattan plots and Q-Q plots for oil content

Figure S7. GWAS Manhattan plots and Q-Q plots for protein content

Figure S8. GWAS Manhattan plots and Q-Q plots for seed weight per plant

Figure S9. Identification of candidate locus associated with 100-seed weight

Figure S10. Identification of candidate locus associated with 100-seed weight

Figure S11. Identification of candidate locus associated with protein content

Figure S12. Identification of candidate locus associated with seed weight per plant

Figure S13. Identification of candidate locus associated with oil content

Figure S14. The distribution of beneficial alleles on chromosomes of six agronomic traits

Figure S15. Genomic prediction accuracies for 100-seed weight, protein content, seed weight per plant, and oil content using different databases and models

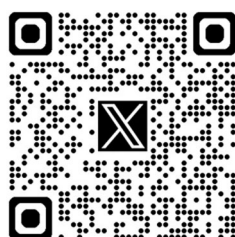
Figure S16. A schematic workflow of the STINS breeding module for targeted trait improvement

Table S1. Whole-genome SNP information

Table S2. Summary of functional annotations of candidate genes in GWAS



Scan the QR code to view
JIPB on WeChat
(WeChat: [jipb1952](#))



Scan the QR code to view
JIPB on X
(X: [@JIPBio](#))



Scan the QR code to view
JIPB on Bluesky
(Bluesky: [jipb.bsky.social](#))