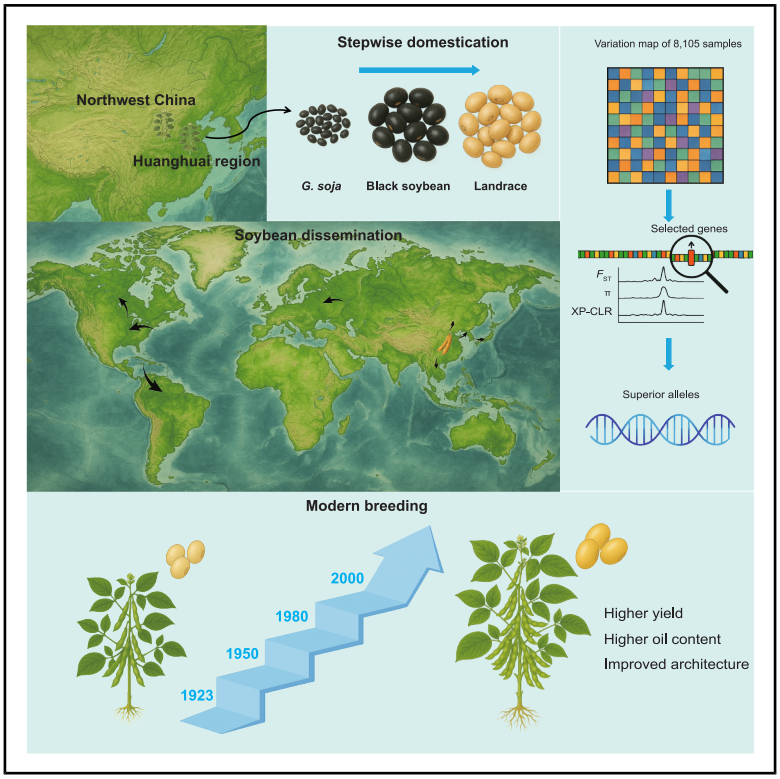
Cell

Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean

Graphical abstract



Authors

Zhou Zhu, Yalin Wang, Shulin Liu, ..., Dongmei Tian, Shuhui Song, Zhixi Tian

Correspondence

tianzhixi@yzwlab.cn

In brief

A genome-wide analysis of 8,105 soybean accessions reveals its stepwise domestication process, in which black soybean acts as an evolutionary intermediate, and identifies region- and era-specific selected genes that have shaped its global adaptation and breeding priorities.

Highlights

- Black soybean is an evolutionary intermediate during soybean stepwise domestication
- Huanghuai region and northwest China are two soybean domestication centers
- Spatiotemporal selection of genes during soybean dissemination and improvement
- Established online quantitative trait nucleotide library and variation map database

Zhu et al., 2025, Cell *188*, 1–17
November 13, 2025
https://doi.org/10.1016/j.cell.2025.09.007

CellPress

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

Cell

CellPress

Article

Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean

Zhou Zhu,[1,8] Yalin Wang,[1,8] Shulin Liu,[1] Shoudong Wang,[2] Juxu Li,[1,3] Chao Fang,[1] Yucheng Liu,[4] Xiaoyue Yang,[5] Dongmei Tian,[6] Shuhui Song,[6,7] and Zhixi Tian[1,7,9,*]

[1]Yazhouwan National Laboratory, Sanya 572025, China
[2]State Key Laboratory of Black Soils Conservation and Utilization, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China
[3]College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China
[4]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[5]State Key Laboratory of Tree Genetics and Breeding, Co-Innovation Center for Sustainable Forestry in Southern China, College of Ecology and Environment, Nanjing Forestry University, Nanjing 210037, China
[6]National Genomics Data Center, China National Centre for Bioinformation & Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[7]University of Chinese Academy of Sciences, Beijing 100039, China
[8]These authors contributed equally
[9]Lead contact
*Correspondence: tianzhixi@yzwlab.cn
https://doi.org/10.1016/j.cell.2025.09.007

SUMMARY

After millennia of domestication, dissemination, and improvement, soybean has evolved into a globally significant leguminous crop. Addressing how soybeans adapt to diverse planting environments and breeding objectives will facilitate future breeding advancements. Here, we systematically investigated the genes under selection of 8,105 soybean accessions underlying domestication, dissemination, and improvement. The analyses revealed that black soybeans serve as a critical domestication intermediate, and soybean domestication traits were selected in a stepwise manner. Comparisons across accessions from diverse geographical areas and historical eras identified numerous selected genes that have contributed to trait enhancement and environmental adaptation during the global dissemination and unveiled a temporal shift of breeding priorities in soybean improvement in China. To highlight the allele utilization among soybean varieties, we constructed a variation map and quantitative trait nucleotide (QTN) library. Our findings provide valuable insights and serve as a critical resource for understanding soybean domestication and informing breeding strategies.

INTRODUCTION

Soybean (*Glycine max* (L.) Merr.) was first domesticated from its wild progenitor (*Glycine soja* Siebold & Zucc.) in China approximately 5,000 to 6,000 years ago and subsequently spread to Eastern and Southeast Asia by the mid-fifteenth century, reaching Europe in 1740, North America in 1765, and Central and South America in the early twentieth century. The expansion of soybean planting areas has greatly increased total soybean production. Nowadays, it has become a globally significant leguminous crop that serves as a primary source of protein and oil for both human consumption and animal feed.[1,2]

Food and Agriculture Organization (FAO) statistical data reveal that global soybean production has increased approximately 13-fold over the past 6 decades. This growth mirrors the escalating demand for soybeans. In reality, the substantial increase in total soybean production can primarily be attributed to the expansion of cultivated areas. In view of the escalating demands from an expanding global population, it is imperative to significantly enhance soybean production.[3] This necessity is further compounded by the challenges posed by climate change, pests, and diseases. Therefore, there is a substantial need to develop soybean cultivars with enhanced yield and oil and protein content, as well as improved stress resilience.[4–6]

The domestication and dissemination entailed rigorous selection for agronomically desirable traits. For instance, in the domestication process, the traits of a more upright plant architecture, stronger stems, reduced pod dehiscence, larger seed size, and increased oil content in seeds were strongly selected,[7–10] whereas, in the dissemination process, the traits of adaptation to diverse environmental conditions, including variations in photoperiod, temperature, disease, and

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

pest pressures, were significantly selected.[11–17] In the past several decades, continuous genetic improvement significantly increased the yield and seed oil content in developed varieties.[18–22]

The selections in these processes have driven dynamic changes in the soybean genome.[23,24] Understanding the genetic changes underlying soybean domestication, dissemination, and improvement; identifying key genes involved in agronomic trait selection; and elucidating their genetic diversity within natural populations are critical for the comprehensive utilization of soybean genetic resources. This knowledge is essential to accelerate breeding programs focused on enhancing yield, stress tolerance, and adaptability to diverse environmental conditions.[25–27]

Over the past decade, a number of studies have examined the genetic diversity and genomic introgression among wild soybeans, landraces, and modern cultivars.[28–31] These investigations have offered valuable insights into the processes of selection and genomic introgression that have occurred during soybean domestication and improvement. However, our understanding of the genetic evolution of the extensively developed genetic resources remains limited. For instance, the genetic variations of the key genes conferring desirable agronomic traits during domestication and global dissemination are not fully characterized, and even the soybean domestication origin and detailed process are still unclear and controversial.[32–34] Furthermore, modern cultivars are primarily high-yield, high-oil varieties with yellow seed coats, and the significance of certain distinctive landraces, such as black soybeans, in the domestication process is largely unexplored.

Here, we comprehensively investigated the evolutionary trajectory of 8,105 soybean accessions, encompassing wild relatives, landraces, and improved cultivars, to elucidate the processes of soybean domestication, dissemination, and improvement. We found that black soybeans are an important intermediate in the history of soybean domestication and revealed the stepwise selection of traits and genes. We also elucidated the allelic diversity of the selected genes in the soybean dissemination and improvement. Moreover, we constructed a variation map comprising 8,105 soybean accessions and established a soybean quantitative trait nucleotide (QTN) library and an online genetic variation database (https://ngdc.cncb.ac.cn/soyomics/breedingtips) for selected genes. Our findings shed light on the intricate interplay of domestication events, breeding strategies, and environmental pressures in shaping current soybean diversity, offering avenues for the identification of functional genes and genetic improvement.

RESULTS

Genetic variations of 8,105 soybean accessions

To investigate the complex processes of soybean domestication and selective breeding, we compiled resequencing data of a total of 8,105 soybean accessions from multiple published studies (see STAR Methods). These accessions consisted of 1,334 wild soybeans, 1,045 landraces, 5,716 improved cultivars, and 10 *G. gracilis* (Table S1; Figure S1A) that spanned a broad geographical range, encompassing China, the Korean Peninsula, Southeast Asia, India, Japan, various European countries, Russia, the United States, Brazil, and Canada (Figure 1A).

We mapped the raw sequencing data (~68.19 Tb of paired-end reads) against the soybean "Zhonghuang 13" (ZH13_v2.0) reference genome[35,36] and identified 48,563,254 single-nucleotide polymorphisms (SNPs) and 10,820,675 insertions/deletions (indels). As expected, the genetic variations are mainly located in the intergenic regions and the intronic regions (Figure S1B). Consistent with previous studies,[7,29,37–40] the genetic diversity of wild soybeans was found to be higher compared with both landraces and improved cultivars (Figure S1C), while the extent of linkage disequilibrium (LD), as indicated by r^2, was lower in wild soybeans relative to both landraces and improved cultivars (Figure S1D).

Genomic architecture of black soybeans uncovers potential origins of soybean domestication

Population structure and phylogenetic analyses demonstrated that all wild lines clustered into a single group; meanwhile, the classification of landraces and improved cultivars exhibited a pronounced correlation with their geographical distribution (Figures 1B and 1C), which was in accordance with previous reports.[23,24,41–43] Notably, our admixture analysis and phylogenetic analysis revealed that certain cultivated soybeans possess distinct genomic structures that differ significantly from those of other cultivated soybeans (Figures 1B, 1C, and S1E). Upon reviewing their sample information, most of these materials were classified as black soybeans, which are characterized by their distinctive black seed coats. In fact, most studies of black soybeans have thus far been limited to their nutritional and compositional properties,[44] with few studies having been conducted at the population-genetic level.

Further investigation of the phylogenetic tree revealed that these black soybeans were evolutionarily more closely related to wild soybeans (Figure 1B). It has been widely acknowledged that the domestication process of soybeans involved a transition from black seed coats to yellow seed coats.[7,29,37,45–50] Given the distinctive genomic features of black soybeans, we proposed that they may serve as a significant intermediate in the domestication process of soybeans. We performed F_{ST} analysis and discovered that the F_{ST} differentiation index between black soybeans and wild soybeans (F_{ST} = 0.220245) was smaller than that between wild soybeans and cultivated soybeans (F_{ST} = 0.276497), and F_{ST} between cultivars and landraces was minimal (F_{ST} = 0.0101152).

Archaeological and historical evidence strongly suggested that cultivated soybeans were domesticated from wild soybeans in China approximately 5,000–6,000 years ago.[32,39] However, the location of the domestication center remains a subject of debate, with some scholars proposing northern China, the Yellow River basin, the Huang-Huai plain in central China, or the region between the Yellow River and the Huai River.[29,32,39] We investigated the geographical distribution of the black soybeans and determined that they could be divided into two subgroups in the phylogenetic tree: clade 1 predominantly distributed in the Huanghuai region (Huang-Huai-Hai Plain, encompassing the Yellow River, Huai River, and Hai River basins) and clade 2 primarily distributed in northwest China (region of northwestern

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

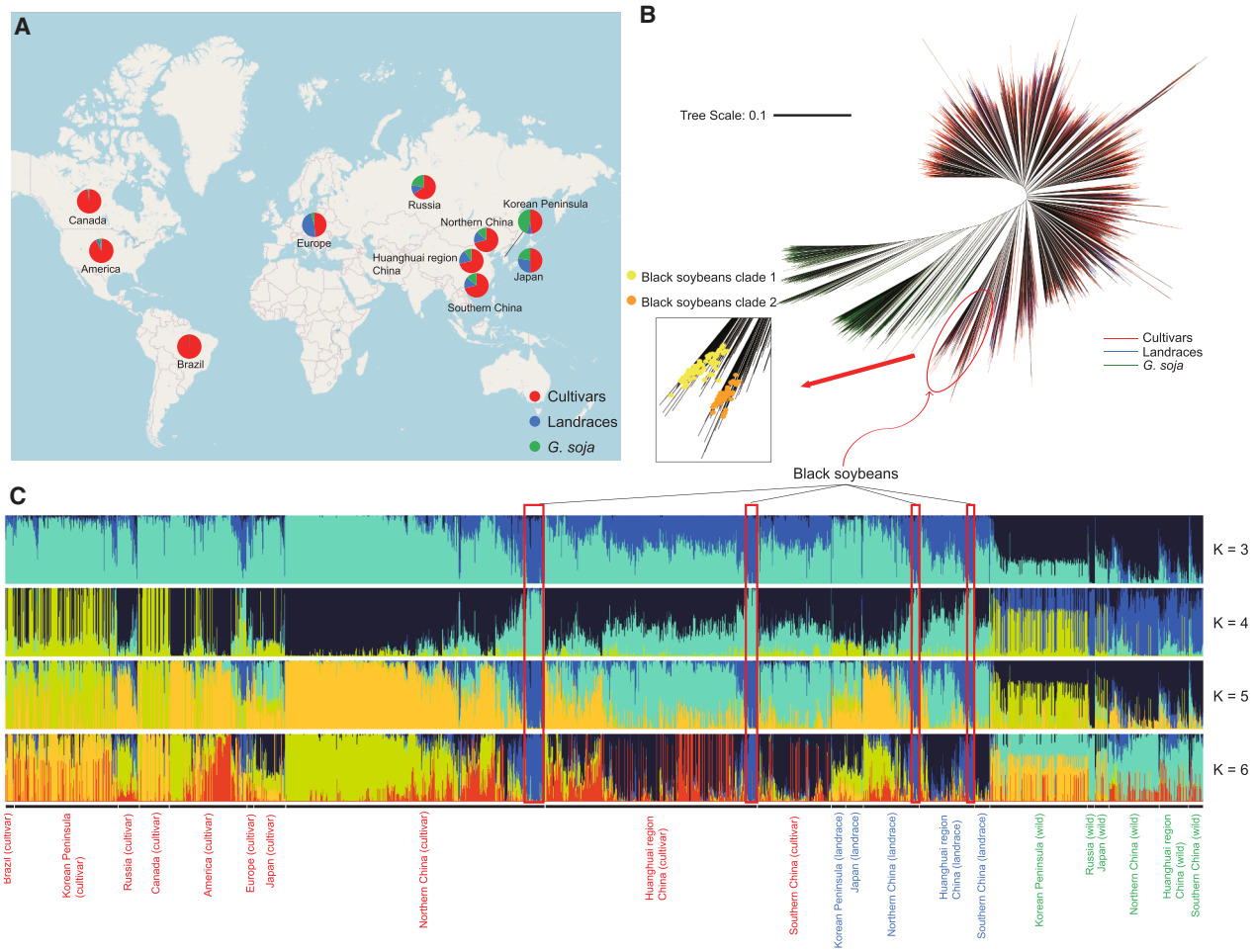


Figure 1. Population structure of 8,105 soybean accessions

(A) Geographical distribution of 8,105 soybean accessions and the proportion of different types of accessions in each region.

(B) Neighbor-joining phylogenetic tree of 8,105 soybean accessions based on whole-genome SNPs. Box in the lower left corner highlights the position of black soybeans on the tree, and branch lengths are proportional to genetic distance (scale bar, 0.1 substitutions per site).

(C) Admixture analysis of 8,105 soybean accessions.

See also Figure S1 and Table S1.

China, including provinces such as Shanxi, Shaanxi, Gansu, and Inner Mongolia) (Figure S1F), indicating the existence of at least two soybean domestication centers: one in the northwest China and another in the Huanghuai region. We subsequently performed TreeMix analysis to model the gene flow among wild soybeans, black soybeans, landraces, and cultivars from northern China and the Huanghuai region. The analysis indicated that gene flow occurred among wild soybeans, black soybeans, cultivars, and landraces (Figure 2A). This finding was consistent with previous reports that multiple introgression and admixture events took place during soybean domestication.[28,39] The complex patterns of gene flow and genetic drift among wild soybeans, black soybeans, and landraces were further supported by *D* statistics (Figure S1G) and f_3 statistics (Table S2). These results supported the hypothesis that black soybeans serve as an intermediate in the domestication process with possible gene flow.

Progressive selection of agronomic traits during the soybean domestication

Previously, to identify potential selective sweeps during soybean domestication, genomic comparisons were made between wild soybeans and landraces or cultivars.[34,37] Given that black soybeans serve as an intermediate in the domestication process, comparisons between wild soybeans and black soybeans, as well as between black soybeans and landraces, will provide additional evidence and insights into the selection dynamics of soybean domestication. Therefore, we conducted a genome-wide scan to identify selective sweeps with multiple methods (see STAR Methods), including F_{ST}, cross-population composite likelihood ratio test (XP-CLR), π ratio, and raised accuracy in sweep detection (RAiSD).[37,51–53] A total of 135 high-quality selective sweeps were detected between wild soybeans and black soybeans, while 486 selective sweeps were identified between black soybeans and landraces (Figures S2A and S2B).

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

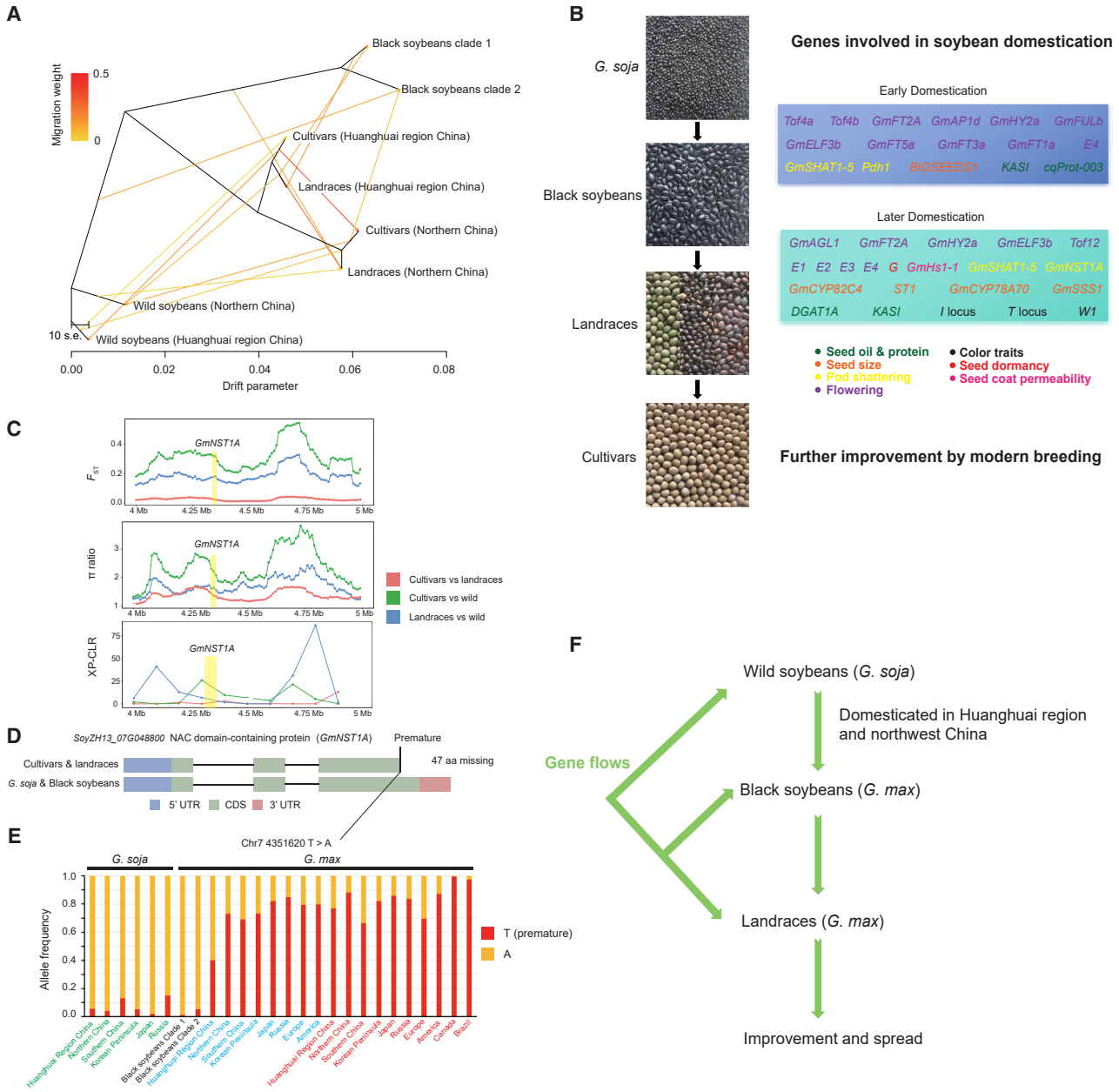


Figure 2. Progressive selection of agronomic traits during the soybean domestication

(A) Gene flow analysis of cultivars, landraces, black soybeans, and wild soybeans from northern China and the Huanghuai region. Colored arrows denote inferred gene flow. The scale bar at bottom left ("10 SE") represents ten standard errors for the drift parameter.

(B) Genes under selection in the wild soybeans to black soybeans stage and the black soybeans to landraces stage.

(C) Analysis of F_{ST}, π ratio, and XP-CLR surrounding the *GmNST1A* gene region.

(D) Gene structure of two haplotypes of the *GmNST1A*.

(E) Allele frequency of the single-nucleotide variant introducing a premature stop codon in *GmNST1A* across accessions.

(F) Proposed model for soybean domestication. Possible gene flows are marked to reflect the complex introgression events during soybean domestication.

See also Figures S2 and S3 and Table S2.

The analysis enabled us to elucidate that the selection of the domestication traits was progressive. For instance, previously reported genes associated with different domestication traits had been identified.[4,7,54] We found that *Tof4a*, *Tof4b*,[55] *GmFT3a*,[56] *AP1d*,[57] *FT1a*,[58] *GmFULb*,[59] *GmFT5a*,[56,60] *E4*,[61] *BIGSEEDS1*,[62] *KASI*,[63] *cqProt-003*,[64,65] *GmSHAT1-5*,[66] and *Pdh1*[8,67] controlling flowering time, seed size, oil and protein content, and pod shattering, respectively, were selected during the stage from wild soybeans to black soybeans (Figure 2B), suggesting that these genes and the related traits had undergone strong selection at

the earlier stage from wild soybeans to the black soybeans, whereas *E1*,[68,69] *E2*, *GmAGL1*,[70,71] *E3*, *E4*, *Tof12*,[72,73] *GmELF3b*, *GmCYP78A70*,[74,75] *GmCYP82C4*, *ST1*,[76] *GmSSS1*,[77] *G*,[78] *GmHs1-1*,[79,80] *GmSHAT1-5*, *KASI*, *DGAT1A*, *I* locus,[45,81] *T* locus,[45] and *W1* controlling flowering time, seed size, seed dormancy, seed coat permeability, oil and protein content, and color traits, respectively, were selected during the stage from black soybeans to landraces (Figure 2B), suggesting that these genes and the related traits were selected during the later domestication stage from black soybeans to the landraces. A previous study suggested that the domestication of seed coat color occurred later than that of seed morphology and quality.[76] Our results of the domestication of color genes were concentrated in the stage from black soybeans to landraces strongly confirmed the hypothesis. Notably, we also found a selection of *E4*, *GmABAS1*,[82] *GmHY2a*,[83] *GmFT2A*, *GmELF3b*, *GmSHAT1-5*, and *KASI* in both procedures, indicating that these genes underwent a continuous selection during the domestication process.

Progressive selection was further demonstrated by changes in the haplotype frequencies of key genes at different stages of domestication. For example, *G* and *GmSHAT1-5* were both reported to be key domestication genes controlling dormancy[78] and pod shattering,[66,84] respectively. We found that the black soybeans captured a similar haplotype to that of the wild soybeans at *G* but significantly different at *GmSHAT1-5* (Figure S2C). The frequency changes at the casual variant of the *G*, and the phylogenetic tree also indicated that the domestication of the *G* into a short-dormancy allele suitable for modern cultivation occurred at the later domestication stage (Figures S2C–S2E).

Moreover, the analysis enabled us to identify potential genes that may play important roles in the domestication (Table S2). For instance, as the homolog to *GmSHAT1-5*, *GmNST1A* (*SoyZH13_07G048800*) has been hypothesized to have a similar function to regulate pod shattering in soybean[8,66]; however, this hypothesis lacks direct genetic evidence. We found that *GmNST1A* also experienced significant selection in the landraces and cultivars (Figure 2C). Black soybeans mostly harbored the same haplotype as wild soybeans (full length, defined as Hap2), whereas landraces and cultivars tended to carry a single-nucleotide mutation leading to a premature stop codon (truncated, defined as Hap1), resulting in a deletion of 47 amino acids (Figures 2D, 2E, and S3A), suggesting pod shattering was continuously refined during soybean domestication, with *GmNST1A* being selected at a later stage than its homolog *GmSHAT1-5*. In addition, subcellular localization in tobacco leaves and *Arabidopsis* protoplasts demonstrated that the truncation of the GmNST1A protein did not affect its localization: both GmNST1A haplotype proteins localize to the nucleus (Figures S3B and S3C). To identify the GmNST1A target genes, we queried the SoyTFBase transcription factor binding prediction platform (www.soytfbase.cn), which indicated a potential interaction between GmNST1A and the *GmSHAT1-5* promoter. Given that a previous study had directly linked *GmSHAT1-5* expression level to pod-shattering resistance in soybean,[66] we were interested in whether the two GmNST1A haplotypes differ in their ability to regulate *GmSHAT1-5* transcription. Yeast one-hybrid assay confirmed that both GmNST1A^{Hap1} and GmNST1A^{Hap2} bind to the *GmSHAT1-5* promoter (Figure S3D). Electrophoretic mobility shift assays (EMSAs) further demonstrated that GmNST1A^{Hap1} has a higher binding affinity for the *GmSHAT1-5* promoter than GmNST1A^{Hap2} (Figure S3E). Finally, dual-luciferase (dual-LUC) reporter assays in *Arabidopsis* protoplasts revealed that the Hap1 variant exhibited a significantly stronger activation of the *GmSHAT1-5* promoter compared with Hap2 (Figures S3F and S3G). These results indicated that natural allelic variation in *GmNST1A* modulates its transcriptional activation strength on *GmSHAT1-5*, suggesting a potential molecular mechanism by which the *GmNST1A* gene contributes to pod shattering during domestication.

Taken together, we proposed a domestication model (Figure 2F) based on current discovery in which wild soybeans were initially domesticated into black soybeans, with early selection focusing on seed size, oil content, protein content, pod shattering, and flowering time. Subsequent domestication phases involved further selection for pod shattering, flowering time, oil and protein content, and seed size, while also targeting additional traits including seed coat color, seed coat permeability, and seed dormancy. In addition, we also detected gene flow among landraces, black soybeans, and wild soybeans (Figures 2A and S1G), indicating that complex introgression events occurred during long-term domestication. This process ultimately led to the development of modern cultivars. It is important to emphasize that while many traits were continuously improved throughout domestication, the genes targeted by artificial selection varied across different stages.

Selection in the soybean dissemination

The global dissemination of soybeans from China has followed a relatively well-documented expansion route.[52,85] We applied the same methods (see STAR Methods) to systematically identify selective sweeps and selected genes in the genomes of cultivars from different countries. Our findings revealed that numerous reported genes that associated with soybean agronomic traits, such as flowering time, stress resistance, and yield and grain quality, were differentially selected along the dissemination route (Figure 3A; Table S3). For instance, given that soybeans are an environmentally sensitive species, the selection of photoperiod and rhythm-related genes is directly associated with the latitude distribution of soybean cultivation. *E2*, a homolog of the *Arabidopsis GIGANTEA*, has been identified as a major gene controlling flowering time in soybeans.[61] Our genome-wide association study (GWAS) for latitude of 5,716 cultivars further identified *E2* as the major-effect quantitative trait locus (QTL) (Figure 3B), indicating it plays a critical role in determining the latitudinal distribution of soybean cultivation. Haplotype analysis demonstrated that *E2* exhibits selection in both low- and high-latitude areas, with Hap2 predominating in low latitudes and Hap1 predominating in high latitudes (Figure 3C; Table S3). Other genes, such as *Tof12*,[72] *FT2b*,[56] *E1*, *Tof8*,[86] and *FT2a*/*E9*,[60] had also undergone significant artificial selection in different latitude regions. However, distinct haplotypes of these genes were preferred in high- and low-latitude areas (Table S3). Notably, we found *Dt1*, a core-effect gene regulating soybean plant architecture and growth habit,[87,88] also exhibited differential selection across various countries (Figure S4B). Accessions in China exhibited a high

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

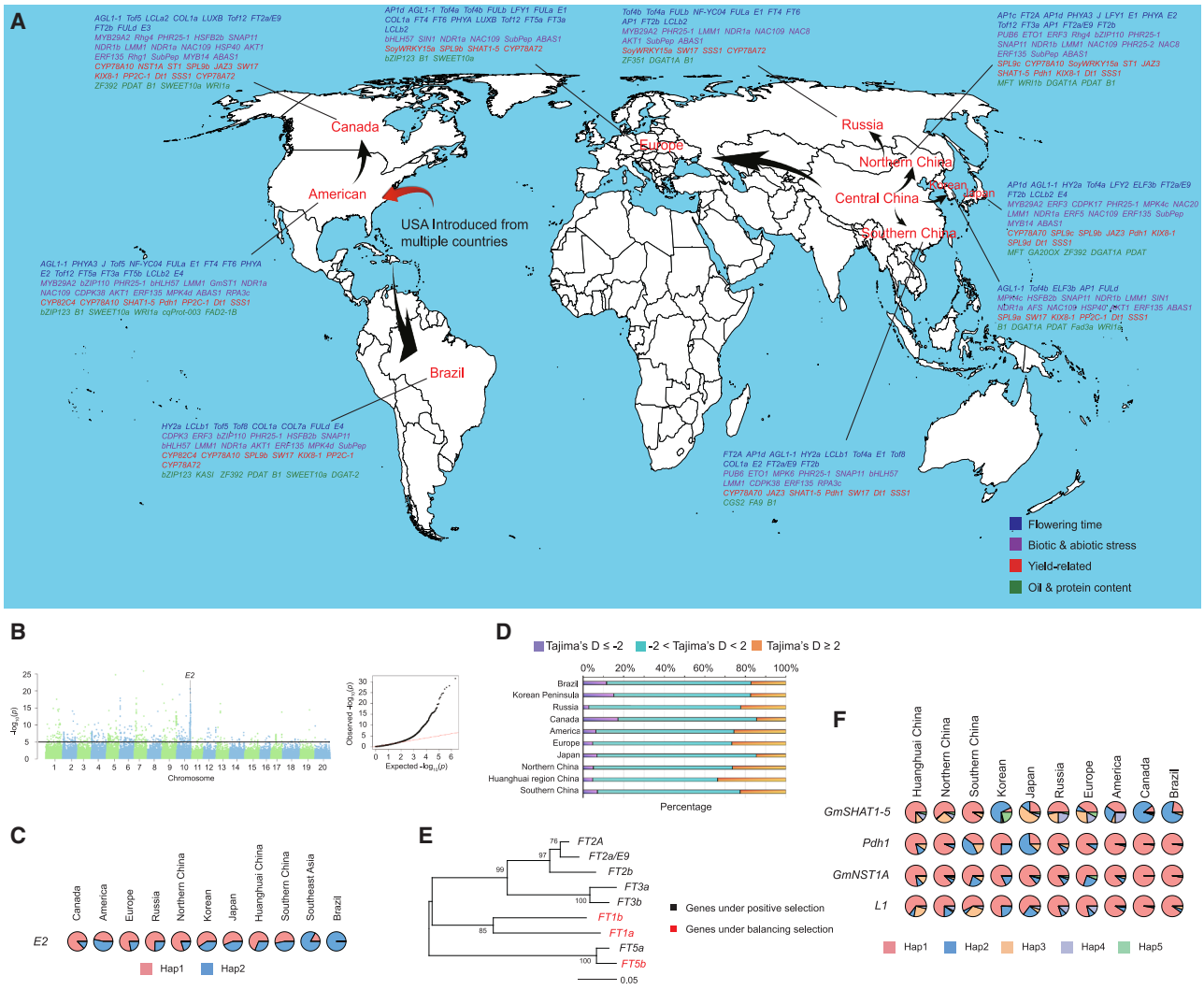


Figure 3. Selection in the soybean dissemination

(A) Genes under selection across different regions during soybean dissemination.

(B) GWAS results for the latitude of 5,716 cultivated soybeans. The horizontal line in the figure represents the threshold of 1×10^{-5}.

(C) Haplotype distribution of *E2* in cultivars from different regions.

(D) Distribution of Tajima's D values in cultivars from different countries.

(E) The neighbor-joining phylogenetic tree of nine *FT* family genes. Scale bar, 0.05 amino acid substitutions per site.

(F) Haplotype distribution of pod-shattering-related genes in cultivated soybeans from different countries.

See also Figure S4 and Table S3.

degree of allelic diversity, whereas countries with higher yields (e.g., America and Brazil) predominantly selected for Hap1, which is associated with indeterminate growth habit. This may imply that the selection for the indeterminate growth habit could potentially contribute to enhanced soybean yields.

It has been reported that balancing selection also played important roles in facilitating plant adaptation.[89–91] Here, we also screened the genes under balancing selection across the cultivars and landraces from different countries using Tajima's D statistics (Figures 3D and S4A). We found that numerous genes undergo balancing selection apart from directional selection during soybean dissemination, and notably, the genes controlling the same trait experienced differential selection. For instance, the results demonstrated that several genes function as key regulators in the soybean flowering network,[92,93] including *FT2a/E9*, *FT2A*, *FT2b*, *FT3a*, *FT3b*, *FT5a*, and *GmELF3b*, appear to be under positive selection, whereas *FT1a*, *FT1b*, *FT5b*, and *GmELF4a* exhibited signatures of balancing selection (Figures 3A, 3E, and S4A). Previous studies have demonstrated that, despite belonging to the same gene family, different *FT* members exhibit divergent functions in regulating flowering and maturation in soybean.[94] For example, *GmFT2a/5a* function as flowering activators, whereas *GmFT1a* acts as a flowering inhibitor.[95] These findings may suggest that

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

the advantageous alleles of the flowering activators appeared to be rapidly fixed through directional selection in specific geographic or environmental contexts, thereby ensuring the basic stability and efficiency of flowering time and circadian rhythms. By contrast, the flowering inhibitor under balancing selection enables soybean to finely tune gene expression in response to varying photoperiods, temperatures, and climatic conditions, allowing for flexible adjustment of flowering time and enhancing regional adaptability.

We also observed that certain pod-shattering-related genes, such as *Pdh1* and *GmSHAT1-5*, exhibited high levels of haplotype polymorphism in some countries (Figure 3F; Table S3), despite pod dehiscence having undergone strong improvement in modern cultivars. This finding is consistent with the observation from a recent study.[42]

Selection during the breeding along different eras in China

Yield, protein content, and oil content are primary objectives in soybean breeding and improvement programs. We first investigated the variations of the hundred-seed weight, protein content, and oil content across cultivars from different eras by categorizing them into four groups: before 1960, 1960–1980, 1980–2000, and after 2000. Our observations revealed that with the advancement of breeding over time, there was a slight increase in protein content accompanied by a decrease in oil content during the period from 1960 to 1980, while the yield remained relatively stable during this period. After which, the hundred-seed weight and oil content showed continuous increases, whereas the protein content significantly decreased (Figure 4A). The character changes indicated a shift in soybean breeding priorities, initially focusing on high-protein varieties, followed by a subsequent emphasis on developing high-yielding and high-oil soybean varieties in later stages. The shift might also be partially driven by the socioeconomic change, as China's rapid economic growth and changing dietary demands shift from high-protein cultivars primarily used in traditional soybean products (e.g., tofu and soy milk) toward high-yield and high-oil cultivars better suited for vegetable oil production and animal feed.[96]

To elucidate the genes involved in the improvement of these traits, we performed a comprehensive comparative analysis of genetic selection across four era groups, identifying numerous selective sweeps throughout the breeding history (Figure 4B). Our results indicated that the selection pressure on genes regulating oil and protein content has decreased over the breeding eras, while the selection for genes associated with yield has increased in later stages, which is consistent with the shift in soybean breeding priorities (Figure 4A). Moreover, during the period from 1980 to 2000, the selection of high-yield genes was more closely associated with genes controlling seed size, such as *GmCYP78A10*, *GmCYP78A57*, *GmCYP82C4*, and *ST1*. By contrast, after 2000, the focus of high yield gradually shifted to genes controlling plant architecture apart from seed size, like *SPL9c*.[97,98] This shift may indicate a change in breeding strategies.

It has been reported that genetic introgression from wild soybeans plays a crucial role in augmenting genetic diversity and improving agronomic traits in cultivated varieties.[28] We investigated the introgression of wild soybean genomic regions in the cultivars from different eras using the ABBA-BABA test.[99] Among these introgressed regions, we identified genes related to stress resistance (e.g., *GmNDR1b* and *GmMYB14*) and flowering (e.g., *GmFT2a/E9*, *GmHY2a*, *GmELF3b*, and *GmSOC1a*) (Figure 4B). A significant constraint on China's soybean production is the limited arable land available for soybeans. Premium land typically prioritized allocated to staple crops such as rice and maize, leaving much of the soybean production to be carried out on marginal salt-alkaline soils.[100] Therefore, improving soybean adaptability to such adverse conditions—particularly through enhanced tolerance to salinity, alkalinity, drought, and diseases, as well as adaptability to varying photoperiods—is of critical importance.

Breeding elite cultivars with concurrent high-yield, high-oil, and high-protein content is an optimal objective for soybean breeding programs. However, oil and protein content typically display a significant negative correlation (Figure 4C), which can be attributed to the pleiotropic and opposing effects exerted by associated functional genes in the protein and oil synthesis pathway.[18,22,101] Indeed, a compilation of the GWAS signals revealed numerous overlapping loci influencing both protein and oil content (Figure 4D). However, certain cultivars exhibit both high-oil and high-protein content simultaneously (e.g., QiHuang 34 and ZhongHuang 301), suggesting the potential to successfully breed elite cultivars with dual high-value traits.

Here, we combined genomic selection (GS) and GWAS (see STAR Methods) using the total protein and oil content (Figures 4E, S5A, and S5B) as a phenotype instead of oil content and protein content, respectively (Figures S5C–S5F). Surprisingly, we discovered a significant signal on chromosome 13 (Figure 4E; Table S3). Of the GWAS significant regions, we identified *Glyma.13G041300* as one promising candidate gene. *Glyma.13G041300* is annotated as *GmSWEET30a*. Previous studies revealed that the sugars will eventually be exported transporter (SWEET) family proteins in soybean, like GmSWEET10a and GmSWEET10b, function as bidirectional sugar transporters and are highly associated with seed oil content and protein content by influencing sugar allocation from seed coat to embryo.[22,102] We found that *GmSWEET30a* also highly expresses in the soybean seed coat (Figure S5G), which was similar to *GmSWEET10a* and *GmSWEET10b*. We hypothesized that *GmSWEET30a* may also affect soybean total protein and oil content in analogous pathway. The natural allelic variations of the *GmSWEET30a* gene resided within its promoter region (Table S3). Further investigation revealed that accessions carrying Hap2 exhibited higher total protein and oil content, whereas those carrying Hap1 showed lower levels of these components (Figure 4F). Using a dual-LUC reporter system, we assessed the transcriptional activity of each promoter haplotype and found that Hap2 induced significantly greater activation compared with Hap1 (Figure 4G). These findings suggest that the expression level of *GmSWEET30a* is likely positively correlated with total protein and oil content. We then generated double mutants of *GmSWEET30a* and its homologous gene (Figure S5H), *Glyma.14G160100* (*GmSWEET30b*), in the Williams 82 background using CRISPR-Cas9. Compared with the wild type, the homozygous knockout line exhibited a significant reduction in total protein and oil content (Figure 4H). Notably,

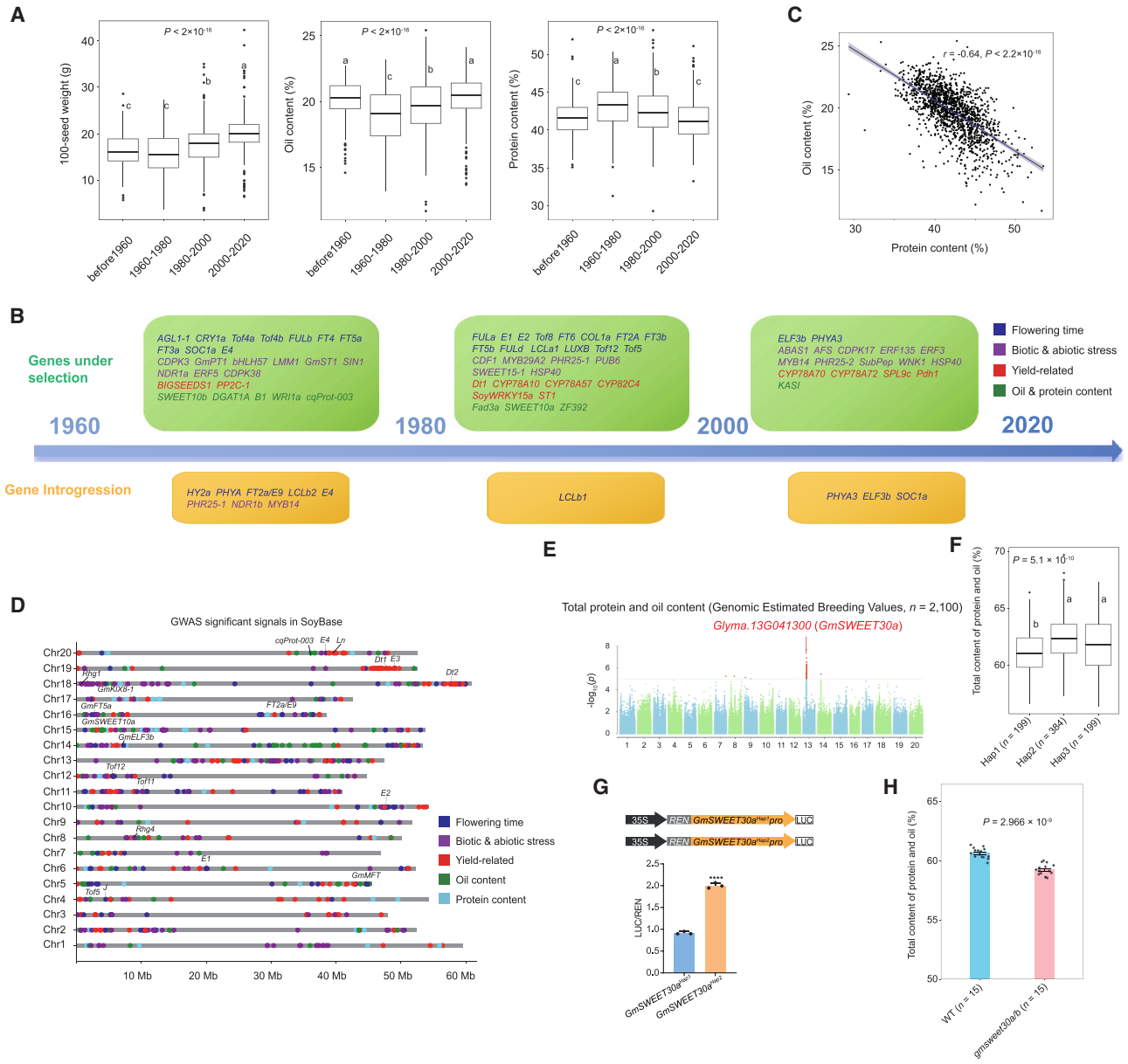


Figure 4. Selection during the breeding along different eras in China

(A) Comparison of 100-seed weight, protein content, and oil content of soybean accessions from different eras. The soybean accessions are classified into four groups according to collected information: before 1960 (n = 157), 1960–1980 (n = 275), 1980–2000 (n = 479), and 2000–2020 (n = 608). Statistical tests were performed using the Kruskal-Wallis test and Dunn's post hoc test in R.

(B) Genes under selection (green box) and genes involved in introgression (orange box) of soybean accessions from different eras.

(C) Pearson correlation between protein and oil content.

(D) GWAS significant signals sourced from the SoyBase website.

(E) GWAS results for total protein and oil content. The horizontal line in the figure represents the threshold of 1×10^{-5}.

(F) Comparison of phenotypic differences among haplotypes of *GmSWEET30a*. Statistical test was performed by the Kruskal-Wallis test and Dunn's post hoc test in R.

(G) Comparison of promoter activity between two *GmSWEET30a* haplotypes using a dual-LUC assay. Bars represent mean ± SD of three independent biological replicates. (****p < 0.0001, Student's t test).

(H) Total seed protein and oil content (%) in wild-type (WT, blue) and *gmsweet30a/b* double-mutant (pink) plants. Data are presented as mean ± SD (n = 15). Statistical test was conducted by a two-tailed Student's t test.

See also Figure S5 and Table S3.

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

GmSWEET30a and *GmSWEET30b* share high homology, and the edited form of *GmSWEET30b* lacks only two amino acids without causing a frameshift mutation (Figure S5I). Our transgenic results indicate that GmSWEET30a and its homolog play an important role in determining soybean seed quality.

Genetic variation database for selected genes

We constructed a genetic database with an emphasis on the variation of these selected genes and QTLs (https://ngdc.cncb.ac.cn/soyomics/breedingtips). Detailed investigation of the allele distribution and combination of selected genes will provide insight into the path for future breeding. We then examined the haplotype distribution of key selected genes in 223 Chinese core backbone parental varieties from different production regions, which represent valuable genetic resources for soybean breeding (Figure 5A; Table S4). The investigation revealed that certain genes had undergone significant selective pressures, leading to the fixation of their advantageous alleles in the majority of core backbone parental varieties, such as *GmCYP78A10*, *GmCYP78A70*, *BIGSEEDS1*, *SoyWRKY15a*, and *GmSWEET10a* (Figure 5A). However, we also identified certain genes that have not been uniformly selected in these core parental varieties (Figure 5A). For instance, *GmSSS1*, a gene that encodes a SPINDLY (SPY, an O-linked N-acetylglucosamine transferase-like protein) homolog reported to affect soybean seed size[77]; *GmZF392*, a gene that encodes a tandem CCCH-type zinc finger protein that was reported to affect soybean oil synthesis[103]; and *cqProt-003*, which is a major gene influencing soybean protein and oil content that encodes a CONSTANS, CONSTANS-like, and TOC1 (CCT) domain-containing protein.[64,65]

Moreover, we found that the emphasis on high-yield, disease-resistant, and high-oil content in modern soybean breeding had led to reduced genetic polymorphism in these 223 backbone parental varieties (Figure 5B). In addition, strong artificial selection may result in the loss of superior alleles at certain genes. For instance, the *I* locus, a major determinant of seed coat color (Figure 5C), is in close physical proximity to *Rhg4* (Figure 5D), a major locus conferring resistance to soybean cyst nematode disease.[104,105] The selection for yellow seed coat color during the soybean domestication and improvement resulted in a loss of *Rhg4* superior haplotype (Hap4) in cultivated soybeans due to genetic drag (Figures 5D–5F; Table S2). This observation is consistent with the generally superior resistance exhibited by black soybeans compared with cultivars with yellow seed coats.

We also constructed a QTN library (Table S4) of soybean key genes from multiple published literature or databases, which was similar to the sophisticated work in rice.[106] We calculated the allelic frequency of 92 important QTNs across a diverse panel of soybean accessions (Figure S6). For several well-characterized domestication genes, such as *GmSWEET10a*,[22] *Tof11*, *Tof12*,[72] and *G*. We examined the allele frequency distributions of their causal variants and observed that the frequency spectrum in black soybeans closely parallels that of wild soybeans (Figure S6; Table S4). These results further support our conclusion that black soybeans represent an intermediate stage in domestication.

Interestingly, we also found a strong selective pressure acted on certain nodulation-related genes during domestication, such as *VAMP721a*, *VAMP721d*,[107] *GmNFR5α*/*GmRj5*, and *GmNFR5β*/*GmRj6*,[108–110] characterized by reduced polymorphism and convergent haplotypes (Figure 6A). Based on the functional characterization of these genes, we speculated that domestication and artificial selection have also affected soybean nodulation, nitrogen fixation, and symbiosis. GmRj5/GmNFR5α has been reported to interact with the type III effector nodulation outer protein L (NopL) and *glycine max* remorin 1a (GmREM1a) to promote symbiosis in soybean.[109] In wild soybeans, *GmRj5* exists in four major haplotypes, each distinguished by non-synonymous and synonymous substitutions (Table S4). During domestication, $GmRj5^{Hap1}$ and $GmRj5^{Hap2}$ were positively selected, whereas $GmRj5^{Hap3}$ and $GmRj5^{Hap4}$ were eliminated (Figure 6A). We then assessed the interaction strength of all four GmRj5 haplotypes with NopL and GmREM1a using membrane yeast two-hybrid assay (Y2H), bimolecular fluorescence complementation (BiFC), and co-immunoprecipitation (coIP) assay (Figures 6B–6E). $GmRj5^{Hap1}$ and $GmRj5^{Hap2}$ exhibited significantly stronger protein-protein interactions with NopL and GmREM1a than $GmRj5^{Hap3}$ and $GmRj5^{Hap4}$, suggesting that the latter were purged due to their diminished ability to participate in the symbiosis signaling cascade. These results highlight that enhanced symbiotic efficiency remains a critical objective in contemporary soybean breeding.

DISCUSSION

Because of its distinct and excellent seed-quality profile, soybeans have served as the most significant source of plant oil and protein.[5] Given the rising population and continuous improvement in living standards, there is a pressing need to substantially enhance soybean production; however, this endeavor presents significantly greater challenges. Long-term artificial selection during domestication, dissemination, and improvement has developed diverse soybean germplasms, which provide a valuable resource for investigating genetic variability, enabling potential solutions to current and future agricultural challenges.[54]

Through a comprehensive investigation of the evolutionary trajectory of 8,105 soybean accessions, we determined that black soybeans serve as an intermediary in soybean domestication. Based on Chinese archaeological records and genomic evidence, we proposed that there are two centers of soybean domestication in China: the Huanghuai region and northwest China. We also compared selection differences between the two domestication centers. Overall, the genomic divergence of black soybeans from the two centers was modest (F_{ST} = 0.0896879). The differential genes identified were primarily related to seed size, flowering, and stress resistance (Figures S2F and S2G), likely reflecting environmental and climatic differences in early domestication across distinct regions.

The primary difference between black and yellow seed coats in soybean lies in the abundance of certain secondary metabolites (e.g., phenols, flavonoids, and anthocyanins).[111] Interestingly, a previous study using an F_3 population derived from cultivated and wild soybeans found that seed coat color may also be associated with seed dormancy and overwintering ability,[112] indicating that seed coat pigmentation or seed coat structure

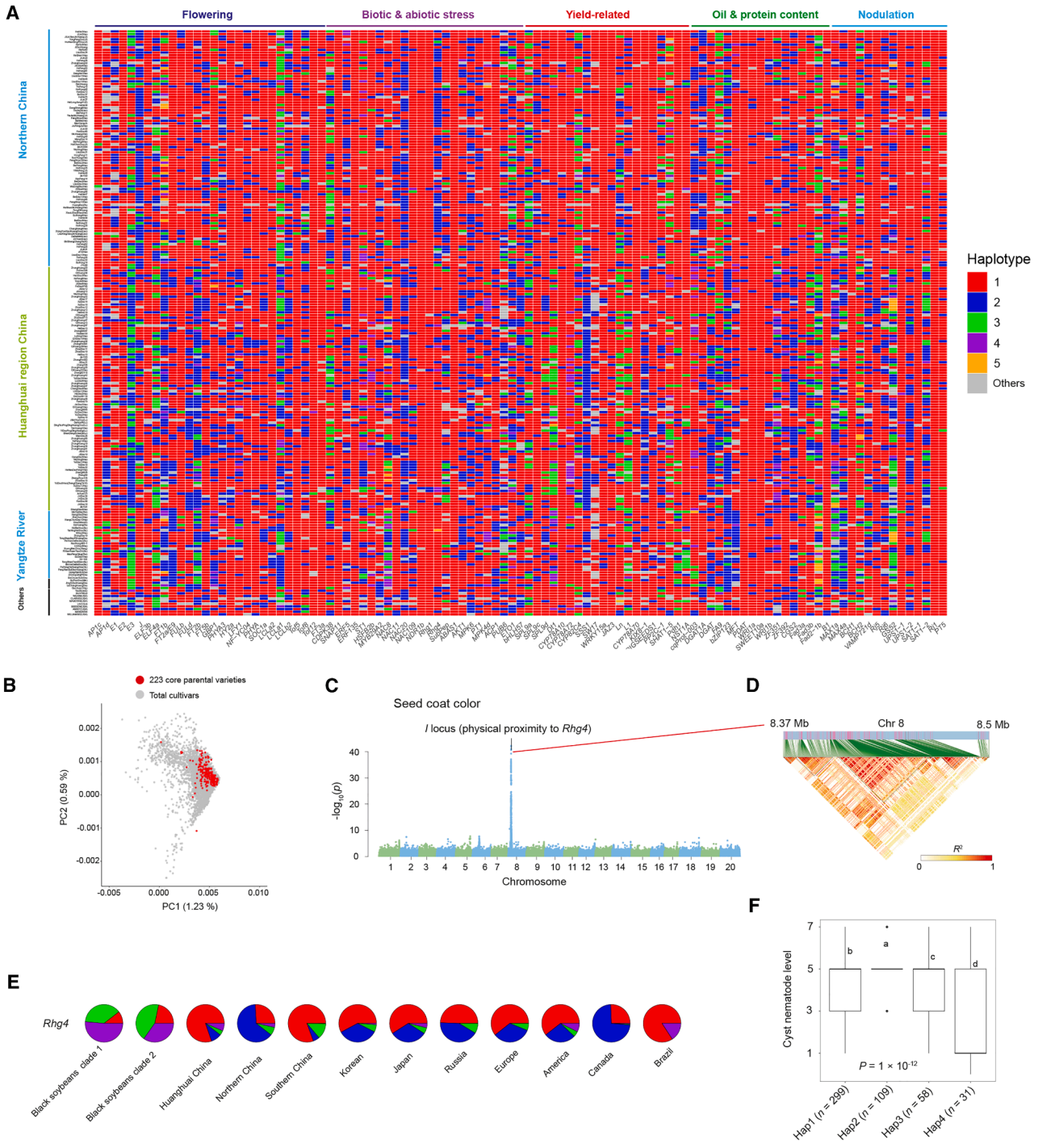


Figure 5. Analysis for selected genes in modern cultivars

(A) Haplotype distribution of selected genes in 223 core soybean parental varieties from different production regions.

(B) Principal-component analysis of 223 core soybean parental varieties.

(C) GWAS results for seed coat color.

(D) LD block within the 8.37–8.50 Mb region of chromosome 8.

(E) Haplotype distribution of *Rhg4* among black soybeans and cultivars from different countries. (Hap1, red; Hap2, blue; Hap3, green; Hap4, purple).

(F) Comparison of the phenotypes of different haplotypes of *Rhg4*. Lower cyst nematode levels represent stronger cyst nematode resistance. A statistical test was performed by the Kruskal-Wallis test and Dunn's post hoc test in R.

See also Figure S6 and Table S4.

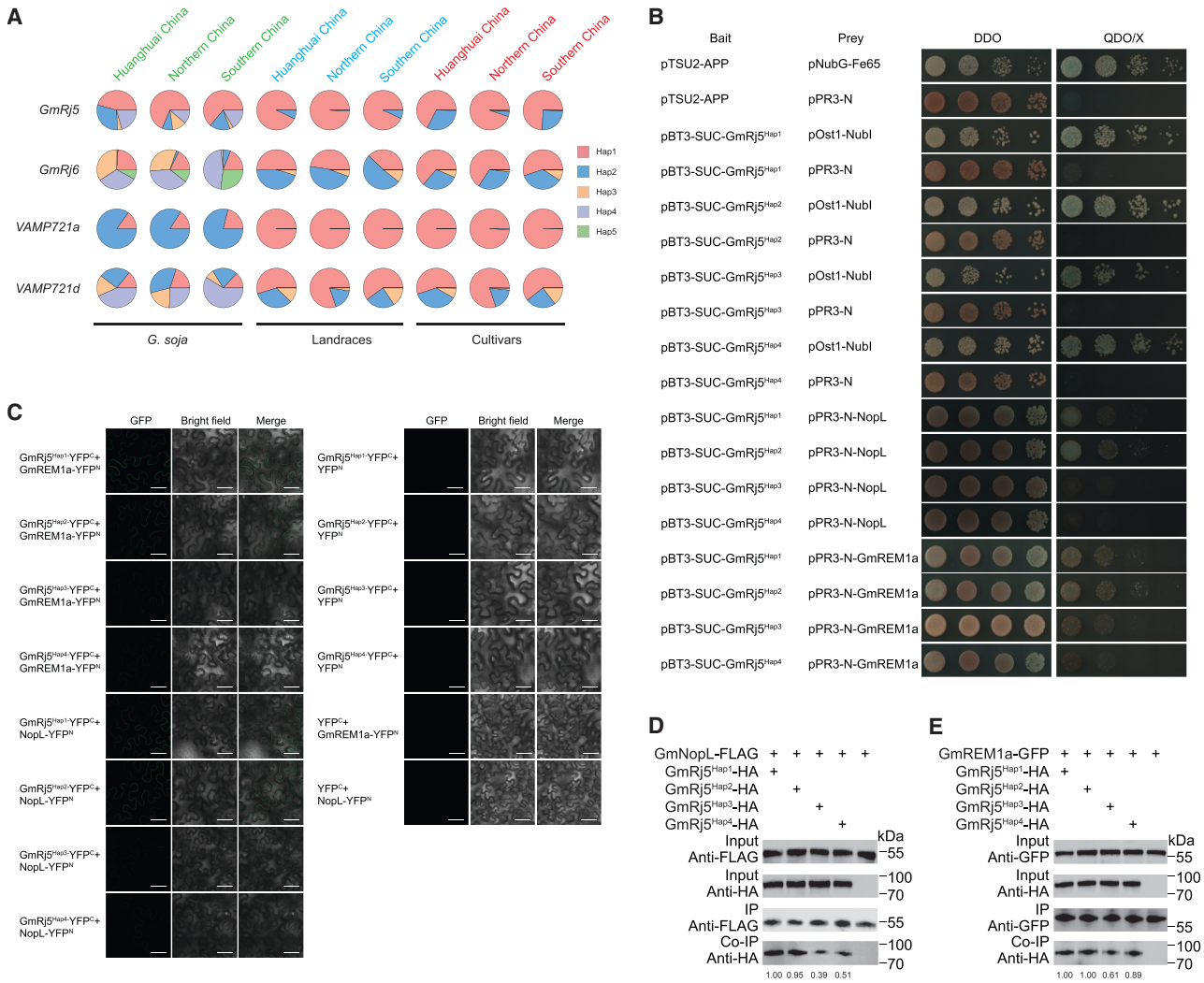


Figure 6. Interactions between different GmRj5/GmNFR5α protein haplotypes and NopL, GmREM1a

(A) Haplotype distribution of four nodulation-related genes in wild soybeans, landraces, and modern cultivars.

(B) Membrane Y2H of interactions between four GmRj5 haplotypes and NopL and GmREM1a proteins.

(C) BiFC assay of four GmRj5 haplotypes with NopL and GmREM1a in tobacco leaves. Scale bar, 50 μm.

(D and E) Interaction of NopL and GmREM1a with different GmRj5 haplotypes in coIP assays. Relative interaction intensities are shown below.

See also Table S4.

may serve additional functions. Black soybeans appeared to harbor more cryptic genomic variations compared with yellow varieties, apart from the obvious difference in seed coat color. As an intermediate product of domestication, black soybeans may retain a greater number of favorable alleles from their wild relatives. For instance, during the outbreak of soybean cyst nematode in the United States in the last century, the introduction of the Chinese landrace "Beijing small black soybean" provided a crucial source of resistance genes, which was also supported by our analysis of the *Rhg4* gene. In the future, elite alleles present in black soybean germplasms could be mined to improve modern cultivars, either through direct allele introgression or by conventional hybrid breeding to enhance agronomic performance in elite cultivars.

In our analysis of modern cultivars from various countries, we found that although pod shattering is a characteristic domestication trait, some modern cultivars in certain regions still carry alleles that predispose them to become shattering. One possibility is that regional variations in agricultural practices, environmental conditions, and selection pressures have resulted in the retention of distinct alleles. Additionally, multiple domestication events or continuous gene flows from wild soybeans could contribute to maintaining high genetic diversity in these genes.

Members of the SWEET family encode bidirectional sugar transporters. *GmSWEET10a* is a well-established domestication locus that affects seed size, protein content, and oil content. A recent study applied an AlphaFold-guided approach to engineer GmSWEET10a and its homolog in order to improve soybean

seed quality.[113] In our GWAS analysis of seed quality, we identified another SWEET family member, *GmSWEET30a*. In the future, leveraging transgenic approaches and artificial intelligence-driven design to engineer multiple SWEET family proteins may be a critical strategy for enhancing seed protein and oil content in soybean.

To date, most soybean domestication studies have focused on traits related to yield and quality, such as growth habit, dormancy, seed size, protein content, oil content, flowering, and pod shattering.[7] Few studies have investigated how nodulation and symbiosis-related phenotypes have changed during soybean domestication. Unlike other crops such as rice and maize, this characteristic is specific to soybean, making research in this area particularly compelling. In our selection scans and haplotype analyses, we identified several nodulation-related genes under strong selection. The domestication of soybean has preferentially retained GmRj5 haplotypes with strong interactions with the type III effector NopL and GmREM1a, highlighting selection for symbiotic efficiency. Prior studies have also shown that *VAMP721a* is important for pectin dynamics and bacterial release in soybean nodules.[107] *VAMP721a* encodes a vesicle-associated membrane protein. In wild soybeans, this gene is predominantly present as Hap2, whereas in landraces and cultivated soybeans, it is almost fixed as Hap1 (Table S4). The two haplotypes differ by a non-synonymous mutation. Future molecular investigations of *VAMP721a* and other nodulation-related genes will enhance our understanding of how domestication has influenced nodulation and symbiosis.

Limitations of the study

Despite incorporating a substantial amount of resequencing data, the majority of accessions in our dataset were sourced from East Asia, with a notably smaller representation from other major soybean-exporting countries, including the United States, Brazil, and Argentina. Furthermore, phenotypic information was frequently incomplete, thereby constraining our analyses to a certain extent. While we successfully identified numerous selection intervals and candidate genes, functional validation through transgenic approaches or molecular assays proved to be inadequate. Future research endeavors will prioritize experimentally validating these candidate genes and investigating their potential applications within breeding programs.

RESOURCE AVAILABILITY

Lead contact

Further information or requests for reagents and resources should be addressed to the lead contact, Zhixi Tian (tianzhixi@yzwlab.cn).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- The data and software referenced in the key resources table are publicly available as indicated.
- The analysis code and scripts have been uploaded to GitHub for public access (https://github.com/sibs-zz/ScricptsForSoybean/).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was funded by the Yazhouwan National Laboratory project (2310ZX01), the National Natural Science Foundation of China (grant no. 32388201), the Xplorer Prize, and the Taishan Scholars Program to Z.T. We acknowledged Professor Guodong Wang (Institute of Genetics and Developmental Biology, Chinese Academy of Sciences) for assistance with soybean seed-quality assessments. We also acknowledged Professor Dawei Xin (Northeast Agricultural University, Harbin, China) for providing the NopL-, GmREM1a-, and GmNFR5-related vectors.

AUTHOR CONTRIBUTIONS

Z.T. designed the experiments and managed the project. Z.Z., C.F., X.Y., and Y.L. performed the data analyses and visualization. Y.W. performed molecular validation experiments. J.L. measured the seed-quality traits. S.W. and S.L. provided the transgenic soybean lines. S.S. constructed the online database. Z.Z. and Z.T. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Soybean accessions and resequencing data
  - Variants calling and annotation
  - Genomic structure analysis
  - Genome-wide association analysis
  - Selection signal detection
  - Analysis of gene haplotypes
  - Analysis of gene flow
  - f_3 statistics calculation
  - Genomic selection
  - Construction of transgenic materials and measurement of soybean quality traits
  - Plasmid construction and plant transformation
  - Subcellular localization of proteins
  - Yeast one-hybrid assay
  - Dual membrane system Y2H assay
  - Electrophoretic mobility shift assay
  - Transcriptional activity assay
  - Co-immunoprecipitation assay
  - Bimolecular fluorescence complementation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.09.007.

Received: April 18, 2025
Revised: July 25, 2025
Accepted: September 5, 2025

REFERENCES

1. Rizzo, G., and Baroni, L. (2018). Soy, soy foods and their role in vegetarian diets. Nutrients *10*, 43. https://doi.org/10.3390/nu10010043.

2. Singer, W.M., Lee, Y.C., Shea, Z., Vieira, C.C., Lee, D., Li, X., Cunicelli, M., Kadam, S.S., Khan, M.A.W., Shannon, G., et al. (2023). Soybean genetics, genomics, and breeding for improving nutritional value and reducing antinutritional traits in food and feed. Plant Genome *16*, e20415. https://doi.org/10.1002/tpg2.20415.

3. Ray, D.K., Mueller, N.D., West, P.C., and Foley, J.A. (2013). Yield trends are insufficient to double global crop production by 2050. PLoS One *8*, e66428. https://doi.org/10.1371/journal.pone.0066428.

4. Zhang, M., Liu, S., Wang, Z., Yuan, Y., Zhang, Z., Liang, Q., Yang, X., Duan, Z., Liu, Y., Kong, F., et al. (2022). Progress in soybean functional genomics over the past decade. Plant Biotechnol. J. *20*, 256–282. https://doi.org/10.1111/pbi.13682.

5. Song, H., Taylor, D.C., and Zhang, M. (2023). Bioengineering of soybean oil and its impact on agronomic traits. Int. J. Mol. Sci. *24*, 2256. https://doi.org/10.3390/ijms24032256.

6. Liu, S., Zhang, M., Feng, F., and Tian, Z. (2020). Toward a "green revolution" for soybean. Mol. Plant *13*, 688–697. https://doi.org/10.1016/j.molp.2020.03.002.

7. Lu, S., Fang, C., Abe, J., Kong, F., and Liu, B. (2022). Current overview on the genetic basis of key genes involved in soybean domestication. aBIOTECH *3*, 126–139. https://doi.org/10.1007/s42994-022-00074-5.

8. Zhang, J., and Singh, A.K. (2020). Genetic control and geo-climate adaptation of pod dehiscence provide novel insights into soybean domestication. G3 (Bethesda) *10*, 545–554. https://doi.org/10.1534/g3.119.400876.

9. Lyu, X., Li, Y.H., Li, Y., Li, D., Han, C., Hong, H., Tian, Y., Han, L., Liu, B., and Qiu, L.J. (2023). The domestication-associated *L1* gene encodes a eucomic acid synthase pleiotropically modulating pod pigmentation and shattering in soybean. Mol. Plant *16*, 1178–1191. https://doi.org/10.1016/j.molp.2023.06.003.

10. Zhou, L., Luo, L., Zuo, J.F., Yang, L., Zhang, L., Guang, X., Niu, Y., Jian, J., Geng, Q.C., Liang, L., et al. (2016). Identification and validation of candidate genes associated with domesticated and improved traits in soybean. Plant Genome *9*, plantgenome2015.09.0090. https://doi.org/10.3835/plantgenome2015.09.0090.

11. 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. https://doi.org/10.1016/j.molp.2021.10.004.

12. Gong, Z. (2020). Flowering phenology as a core domestication trait in soybean. J. Integr. Plant Biol. *62*, 546–549. https://doi.org/10.1111/jipb.12934.

13. Li, C., Li, Y.H., 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. https://doi.org/10.1016/j.molp.2020.01.014.

14. Li, J., Li, Y., Agyenim-Boateng, K.G., Shaibu, A.S., Liu, Y., Feng, Y., Qi, J., Li, B., Zhang, S., and Sun, J. (2024). Natural variation of domestication-related genes contributed to latitudinal expansion and adaptation in soybean. BMC Plant Biol. *24*, 651. https://doi.org/10.1186/s12870-024-05382-0.

15. Wang, H., Li, X., Su, F., Liu, H., Hu, D., Huang, F., Yu, D., and Wang, H. (2022). Soybean *CALCIUM-DEPENDENT PROTEIN KINASE17* positively regulates plant resistance to common cutworm (*Spodoptera litura* Fabricius). Int. J. Mol. Sci. *23*, 15696. https://doi.org/10.3390/ijms232415696.

16. Zhang, G., Chen, M., Li, L., Xu, Z., Chen, X., Guo, J., and Ma, Y. (2009). Overexpression of the soybean *GmERF3* gene, an AP2/ERF type transcription factor for increased tolerances to salt, drought, and diseases in transgenic tobacco. J. Exp. Bot. *60*, 3781–3796. https://doi.org/10.1093/jxb/erp214.

17. Zhao, X., Jing, Y., Luo, Z., Gao, S., Teng, W., Zhan, Y., Qiu, L., Zheng, H., Li, W., and Han, Y. (2021). *GmST1*, which encodes a sulfotransferase, confers resistance to soybean mosaic virus strains G2 and G3. Plant Cell Environ. *44*, 2777–2792. https://doi.org/10.1111/pce.14066.

18. Cai, Z., Xian, P., Cheng, Y., Zhong, Y., Yang, Y., Zhou, Q., Lian, T., Ma, Q., Nian, H., and Ge, L. (2023). MOTHER-OF-FT-AND-TFL1 regulates the seed oil and protein content in soybean. New Phytol. *239*, 905–919. https://doi.org/10.1111/nph.18792.

19. Cao, P., Zhao, Y., Wu, F., Xin, D., Liu, C., Wu, X., Lv, J., Chen, Q., and Qi, Z. (2022). Multi-omics techniques for soybean molecular breeding. Int. J. Mol. Sci. *23*, 4994. https://doi.org/10.3390/ijms23094994.

20. Zhang, D., Zhang, H., Hu, Z., Chu, S., Yu, K., Lv, L., Yang, Y., Zhang, X., Chen, X., Kan, G., et al. (2019). Artificial selection on *GmOLEO1* contributes to the increase in seed oil during soybean domestication. PLoS Genet. *15*, e1008267. https://doi.org/10.1371/journal.pgen.1008267.

21. 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. https://doi.org/10.1111/nph.16250.

22. Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y.C., Liu, Z., Frommer, W.B., Ma, J.F., Chen, L.Q., et al. (2020). Simultaneous changes in seed size, oil content and protein content driven by selection of *SWEET* homologues during soybean domestication. Natl. Sci. Rev. *7*, 1776–1786. https://doi.org/10.1093/nsr/nwaa110.

23. Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., and Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome *8*, eplantgenome2015.04.0024. https://doi.org/10.3835/plantgenome2015.04.0024.

24. Bayer, P.E., Valliyodan, B., Hu, H., Marsh, J.I., Yuan, Y., Vuong, T.D., Patil, G., Song, Q., Batley, J., Varshney, R.K., et al. (2022). Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. Plant Genome *15*, e20109. https://doi.org/10.1002/tpg2.20109.

25. Fudge, J.B. (2022). Flowering time: soybean adapts to the tropics. Curr. Biol. *32*, R360–R362. https://doi.org/10.1016/j.cub.2022.03.030.

26. 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. https://doi.org/10.1038/s41467-021-25800-3.

27. Wang, Y., Yuan, L., Su, T., Wang, Q., Gao, Y., Zhang, S., Jia, Q., Yu, G., Fu, Y., Cheng, Q., et al. (2020). Light- and temperature-entrainable circadian clock in soybean development. Plant Cell Environ. *43*, 637–648. https://doi.org/10.1111/pce.13678.

28. Wang, X., Chen, L., and Ma, J. (2019). Genomic introgression through interspecific hybridization counteracts genetic bottleneck during soybean domestication. Genome Biol. *20*, 22. https://doi.org/10.1186/s13059-019-1631-5.

29. Li, Y.H., 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. https://doi.org/10.1007/s11427-022-2158-7.

30. Kim, M.S., Lozano, R., Kim, J.H., Bae, D.N., Kim, S.T., Park, J.H., Choi, M.S., Kim, J., Ok, H.C., Park, S.K., et al. (2021). The patterns of deleterious mutations during the domestication of soybean. Nat. Commun. *12*, 97. https://doi.org/10.1038/s41467-020-20337-3.

31. Yang, C., Yan, J., Jiang, S., Li, X., Min, H., Wang, X., and Hao, D. (2022). Resequencing 250 soybean accessions: new insights into genes associated with agronomic traits and genetic networks. Genomics Proteomics Bioinformatics *20*, 29–41. https://doi.org/10.1016/j.gpb.2021.02.009.

32. 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. https://doi.org/10.1007/s00122-018-3271-7.

33. Lee, G.A., Crawford, G.W., Liu, L., Sasaki, Y., and Chen, X. (2011). Archaeological soybean (*Glycine max*) in East Asia: does size matter? PLoS One *6*, e26720. https://doi.org/10.1371/journal.pone.0026720.

34. Chung, W.H., Jeong, N., Kim, J., Lee, W.K., Lee, Y.G., Lee, S.H., Yoon, W., Kim, J.H., Choi, I.Y., Choi, H.K., et al. (2014). Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA Res. *21*, 153–167. https://doi.org/10.1093/dnares/dst047.

35. Shen, Y., Liu, J., Geng, H., Zhang, J., Liu, Y., Zhang, H., Xing, S., Du, J., Ma, S., and Tian, Z. (2018). De novo assembly of a Chinese soybean genome. Sci. China Life Sci. *61*, 871–884. https://doi.org/10.1007/s11427-018-9360-0.

36. Shen, Y., Du, H., Liu, Y., Ni, L., Wang, Z., Liang, C., and Tian, Z. (2019). Update soybean Zhonghuang 13 genome to a golden reference. Sci. China Life Sci. *62*, 1257–1260. https://doi.org/10.1007/s11427-019-9822-2.

37. 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. https://doi.org/10.1038/nbt.3096.

38. Han, Y., Zhao, X., Liu, D., Li, Y., Lightfoot, D.A., Yang, Z., Zhao, L., Zhou, G., Wang, Z., Huang, L., et al. (2016). Domestication footprints anchor genomic regions of agronomic importance in soybeans. New Phytol. *209*, 871–884. https://doi.org/10.1111/nph.13626.

39. Sedivy, E.J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol. *214*, 539–553. https://doi.org/10.1111/nph.14418.

40. Kofsky, J., Zhang, H., and Song, B.H. (2018). The untapped genetic reservoir: the past, current, and future applications of the wild soybean (*Glycine soja*). Front. Plant Sci. *9*, 949. https://doi.org/10.3389/fpls.2018.00949.

41. Song, Q., Hyten, D.L., Jia, G., Quigley, C.V., Fickus, E.W., Nelson, R.L., and Cregan, P.B. (2015). Fingerprinting soybean germplasm and its utility in genomic research. G3 (Bethesda) *5*, 1999–2006. https://doi.org/10.1534/g3.115.019000.

42. Yano, R., Li, F., Hiraga, S., Takeshima, R., Kobayashi, M., Toda, K., Umehara, Y., Kajiya-Kanegae, H., Iwata, H., Kaga, A., et al. (2025). The genomic landscape of gene-level structural variations in Japanese and global soybean *Glycine max* cultivars. Nat. Genet. *57*, 973–985. https://doi.org/10.1038/s41588-025-02113-5.

43. Du, H., Fang, C., Li, Y., Kong, F., and Liu, B. (2023). Understandings and future challenges in soybean functional genomics and molecular breeding. J. Integr. Plant Biol. *65*, 468–495. https://doi.org/10.1111/jipb.13433.

44. Li, S., Chen, J., Hao, X., Ji, X., Zhu, Y., Chen, X., and Yao, Y. (2024). A systematic review of black soybean (*Glycine max* (L.) Merr.): Nutritional composition, bioactive compounds, health benefits, and processing to application. Food Front. *5*, 1188–1211. https://doi.org/10.1002/fft2.376.

45. Yuan, B., Yuan, C., Wang, Y., Liu, X., Qi, G., Wang, Y., Dong, L., Zhao, H., Li, Y., and Dong, Y. (2022). Identification of genetic loci conferring seed coat color based on a high-density map in soybean. Front. Plant Sci. *13*, 968618. https://doi.org/10.3389/fpls.2022.968618.

46. Yang, K., Jeong, N., Moon, J.K., Lee, Y.H., Lee, S.H., Kim, H.M., Hwang, C.H., Back, K., Palmer, R.G., and Jeong, S.C. (2010). Genetic analysis of genes controlling natural variation of seed coat and flower colors in soybean. J. Hered. *101*, 757–768. https://doi.org/10.1093/jhered/esq078.

47. Senda, M., Jumonji, A., Yumoto, S., Ishikawa, R., Harada, T., Niizeki, M., and Akada, S. (2002). Analysis of the duplicated *CHS1* gene related to the suppression of the seed coat pigmentation in yellow soybeans. Theor. Appl. Genet. *104*, 1086–1091. https://doi.org/10.1007/s00122-001-0801-4.

48. Kasai, A., Kasai, K., Yumoto, S., and Senda, M. (2007). Structural features of *GmIRCHS*, candidate of the *I* gene inhibiting seed coat pigmentation in soybean: implications for inducing endogenous RNA silencing of chalcone synthase genes. Plant Mol. Biol. *64*, 467–479. https://doi.org/10.1007/s11103-007-9169-4.

49. Owen, F.V. (1928). Inheritance studies in soybeans. III. Seed-coat color and summary of all other mendelian characters thus far reported. Genetics *13*, 50–79. https://doi.org/10.1093/genetics/13.1.50.

50. Woodworth, C.M. (1921). Inheritance of cotyledon, seed-coat, hilum and pubescence colors in soy-beans. Genetics *6*, 487–553. https://doi.org/10.1093/genetics/6.6.487.

51. Huang, X., Kurata, N., Wei, X., Wang, Z.X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. Nature *490*, 497–501. https://doi.org/10.1038/nature11532.

52. Cho, Y., Kim, J.Y., Kim, S.K., Kim, S.Y., Kim, N., Lee, J., and Park, J.L. (2024). Whole-genome sequencing analysis of soybean diversity across different countries and selection signature of Korean soybean accession. G3 (Bethesda) *14*, jkae118. https://doi.org/10.1093/g3journal/jkae118.

53. Alachiotis, N., and Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. Commun. Biol. *1*, 79. https://doi.org/10.1038/s42003-018-0085-8.

54. Tian, Z., Nepomuceno, A.L., Song, Q., Stupar, R.M., Liu, B., Kong, F., Ma, J., Lee, S.H., and Jackson, S.A. (2025). Soybean2035: A decadal vision for soybean functional genomics and breeding. Mol. Plant *18*, 245–271. https://doi.org/10.1016/j.molp.2025.01.004.

55. Fang, C., Sun, Z., Li, S., Su, T., Wang, L., Dong, L., Li, H., Li, L., Kong, L., Yang, Z., et al. (2024). Subfunctionalisation and self-repression of duplicated *E1* homologues finetunes soybean flowering and adaptation. Nat. Commun. *15*, 6184. https://doi.org/10.1038/s41467-024-50623-3.

56. Wu, F., Sedivy, E.J., Price, W.B., Haider, W., and Hanzawa, Y. (2017). Evolutionary trajectories of duplicated *FT* homologues and their roles in soybean domestication. Plant J. *90*, 941–953. https://doi.org/10.1111/tpj.13521.

57. Chen, L., Nan, H., Kong, L., Yue, L., Yang, H., Zhao, Q., Fang, C., Li, H., Cheng, Q., Lu, S., et al. (2020). Soybean *AP1* homologs control flowering time and plant height. J. Integr. Plant Biol. *62*, 1868–1879. https://doi.org/10.1111/jipb.12988.

58. Wang, L., Sun, S., Wu, T., Liu, L., Sun, X., Cai, Y., Li, J., Jia, H., Yuan, S., Chen, L., et al. (2020). Natural variation and CRISPR/Cas9-mediated mutation in *GmPRR37* affect photoperiodic flowering and contribute to regional adaptation of soybean. Plant Biotechnol. J. *18*, 1869–1881. https://doi.org/10.1111/pbi.13346.

59. Escamilla, D.M., Dietz, N., Bilyeu, K., Hudson, K., and Rainey, K.M. (2024). Genome-wide association study reveals *GmFulb* as candidate gene for maturity time and reproductive length in soybeans (*Glycine max*). PLoS One *19*, e0294123. https://doi.org/10.1371/journal.pone.0294123.

60. Nan, H., Cao, D., Zhang, D., Li, Y., Lu, S., Tang, L., Yuan, X., Liu, B., and Kong, F. (2014). GmFT2a and GmFT5a redundantly and differentially regulate flowering through interaction with and upregulation of the bZIP transcription factor GmFDL19 in soybean. PLoS One *9*, e97669. https://doi.org/10.1371/journal.pone.0097669.

61. Jiang, B., Nan, H., Gao, Y., Tang, L., Yue, Y., Lu, S., Ma, L., Cao, D., Sun, S., Wang, J., et al. (2014). Allelic combinations of soybean maturity Loci *E1*, *E2*, *E3* and *E4* result in diversity of maturity and adaptation to different latitudes. PLoS One *9*, e106042. https://doi.org/10.1371/journal.pone.0106042.

62. Ge, L., Yu, J., Wang, H., Luth, D., Bai, G., Wang, K., and Chen, R. (2016). Increasing seed size and quality by manipulating *BIG SEEDS1* in legume species. Proc. Natl. Acad. Sci. USA *113*, 12414–12419. https://doi.org/10.1073/pnas.1611763113.

63. Dobbels, A.A., Michno, J.M., Campbell, B.W., Virdi, K.S., Stec, A.O., Muehlbauer, G.J., Naeve, S.L., and Stupar, R.M. (2017). An induced chromosomal translocation in soybean disrupts a *KASI* ortholog and is

associated with a high-sucrose and low-oil seed phenotype. G3 (Bethesda) *7*, 1215–1223. https://doi.org/10.1534/g3.116.038596.

64. Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D., Song, Q., Pantalone, V.R., Song, B.H., Yu, D., et al. (2022). *POWR1* is a domestication gene pleiotropically regulating seed quality and yield in soybean. Nat. Commun. *13*, 3051. https://doi.org/10.1038/s41467-022-30314-7.

65. Marsh, J.I., Hu, H., Petereit, J., Bayer, P.E., Valliyodan, B., Batley, J., Nguyen, H.T., and Edwards, D. (2022). Haplotype mapping uncovers unexplored variation in wild and domesticated soybean at the major protein locus *cqProt-003*. Theor. Appl. Genet. *135*, 1443–1455. https://doi.org/10.1007/s00122-022-04045-8.

66. Dong, Y., Yang, X., Liu, J., Wang, B.H., Liu, B.L., and Wang, Y.Z. (2014). Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nat. Commun. *5*, 3352. https://doi.org/10.1038/ncomms4352.

67. Hu, D., Kan, G., Hu, W., Li, Y., Hao, D., Li, X., Yang, H., Yang, Z., He, X., Huang, F., et al. (2019). Identification of loci and candidate genes responsible for pod dehiscence in soybean via genome-wide association analysis across multiple environments. Front. Plant Sci. *10*, 811. https://doi.org/10.3389/fpls.2019.00811.

68. Gao, Y., Zhang, Y., Ma, C., Chen, Y., Liu, C., Wang, Y., Wang, S., and Chen, X. (2024). Editing the nuclear localization signals of *E1* and *E1Lb* enables the production of tropical soybean in temperate growing regions. Plant Biotechnol. J. *22*, 2145–2156. https://doi.org/10.1111/pbi.14335.

69. Wan, Z., Liu, Y., Guo, D., Fan, R., Liu, Y., Xu, K., Zhu, J., Quan, L., Lu, W., Bai, X., et al. (2022). CRISPR/Cas9-mediated targeted mutation of the *E1* decreases photoperiod sensitivity, alters stem growth habits, and decreases branch number in soybean. Front. Plant Sci. *13*, 1066820. https://doi.org/10.3389/fpls.2022.1066820.

70. Zeng, X., Liu, H., Du, H., Wang, S., Yang, W., Chi, Y., Wang, J., Huang, F., and Yu, D. (2018). Soybean MADS-box gene *GmAGL1* promotes flowering via the photoperiod pathway. BMC Genomics *19*, 51. https://doi.org/10.1186/s12864-017-4402-2.

71. Chi, Y., Wang, T., Xu, G., Yang, H., Zeng, X., Shen, Y., Yu, D., and Huang, F. (2017). *GmAGL1*, a MADS-Box gene from soybean, is involved in floral organ identity and fruit dehiscence. Front. Plant Sci. *8*, 175. https://doi.org/10.3389/fpls.2017.00175.

72. Lu, S., Dong, L., Fang, C., Liu, S., Kong, L., Cheng, Q., Chen, L., Su, T., Nan, H., Zhang, D., et al. (2020). Stepwise selection on homeologous *PRR* genes controlling flowering and maturity during soybean domestication. Nat. Genet. *52*, 428–436. https://doi.org/10.1038/s41588-020-0604-7.

73. Wang, L., Niu, F., Wang, J., Zhang, H., Zhang, D., and Hu, Z. (2024). Genome-wide association studies prioritize genes controlling seed size and reproductive period length in soybean. Plants (Basel) *13*, 615. https://doi.org/10.3390/plants13050615.

74. Dai, A.H., Yang, S.X., Zhou, H.K., Tang, K.Q., Li, G., Leng, J.T., Yu, H., Zhang, Y.H., Gao, J.S., Yang, X., et al. (2018). Evolution and expression divergence of the *CYP78A* subfamily genes in soybean. Genes (Basel) *9*, 611. https://doi.org/10.3390/genes9120611.

75. Zhao, B., Dai, A., Wei, H., Yang, S., Wang, B., Jiang, N., and Feng, X. (2016). Arabidopsis KLU homologue *GmCYP78A72* regulates seed size in soybean. Plant Mol. Biol. *90*, 33–47. https://doi.org/10.1007/s11103-015-0392-0.

76. Li, J., Zhang, Y., Ma, R., Huang, W., Hou, J., Fang, C., Wang, L., Yuan, Z., Sun, Q., Dong, X., et al. (2022). Identification of *ST1* reveals a selection involving hitchhiking of seed morphology and oil content during soybean domestication. Plant Biotechnol. J. *20*, 1110–1121. https://doi.org/10.1111/pbi.13791.

77. Zhu, W., Yang, C., Yong, B., Wang, Y., Li, B., Gu, Y., Wei, S., An, Z., Sun, W., Qiu, L., et al. (2022). An enhancing effect attributed to a nonsynonymous mutation in *SOYBEAN SEED SIZE 1*, a *SPINDLY*-like gene, is exploited in soybean domestication and improvement. New Phytol. *236*, 1375–1392. https://doi.org/10.1111/nph.18461.

78. Wang, M., Li, W., Fang, C., Xu, F., Liu, Y., Wang, Z., Yang, R., Zhang, M., Liu, S., Lu, S., et al. (2018). Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat. Genet. *50*, 1435–1441. https://doi.org/10.1038/s41588-018-0229-2.

79. Sun, L., Miao, Z., Cai, C., Zhang, D., Zhao, M., Wu, Y., Zhang, X., Swarm, S.A., Zhou, L., Zhang, Z.J., et al. (2015). *GmHs1-1*, encoding a calcineurin-like protein, controls hard-seededness in soybean. Nat. Genet. *47*, 939–943. https://doi.org/10.1038/ng.3339.

80. Yan, H., Tian, D., Zhang, Q., Wen, J., Wang, Z.Y., and Chai, M. (2024). *GmHs1-1* and *GmqHS1* simultaneously contribute to the domestication of soybean hard-seededness. Plants (Basel) *13*, 2061. https://doi.org/10.3390/plants13152061.

81. Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., et al. (2020). Pan-genome of wild and cultivated soybeans. Cell *182*, 162–176.e13. https://doi.org/10.1016/j.cell.2020.05.023.

82. Ku, Y.S., Ni, M., Muñoz, N.B., Xiao, Z., Lo, A.W.Y., Chen, P., Li, M.W., Cheung, M.Y., Xie, M., and Lam, H.M. (2020). ABAS1 from soybean is a 1R-subtype MYB transcriptional repressor that enhances ABA sensitivity. J. Exp. Bot. *71*, 2970–2981. https://doi.org/10.1093/jxb/eraa081.

83. Zhang, Z., Yang, S., Wang, Q., Yu, H., Zhao, B., Wu, T., Tang, K., Ma, J., Yang, X., and Feng, X. (2022). Soybean *GmHY2a* encodes a phytochromobilin synthase that regulates internode length and flowering time. J. Exp. Bot. *73*, 6646–6662. https://doi.org/10.1093/jxb/erac318.

84. Li, S., Wang, W., Sun, L., Zhu, H., Hou, R., Zhang, H., Tang, X., Clark, C. B., Swarm, S.A., Nelson, R.L., et al. (2024). Artificial selection of mutations in two nearby genes gave rise to shattering resistance in soybean. Nat. Commun. *15*, 7588. https://doi.org/10.1038/s41467-024-52044-8.

85. Zhou, X., Wang, D., Mao, Y., Zhou, Y., Zhao, L., Zhang, C., Liu, Y., and Chen, J. (2022). The Organ size and morphological change during the domestication process of soybean. Front. Plant Sci. *13*, 913238. https://doi.org/10.3389/fpls.2022.913238.

86. Li, H., Du, H., He, M., Wang, J., Wang, F., Yuan, W., Huang, Z., Cheng, Q., Gou, C., Chen, Z., et al. (2023). Natural variation of *FKF1* controls flowering and adaptation during soybean domestication and improvement. New Phytol. *238*, 1671–1684. https://doi.org/10.1111/nph.18826.

87. Tian, Z., Wang, X., Lee, R., Li, Y., Specht, J.E., Nelson, R.L., McClean, P. E., Qiu, L., and Ma, J. (2010). Artificial selection for determinate growth habit in soybean. Proc. Natl. Acad. Sci. USA *107*, 8563–8568. https://doi.org/10.1073/pnas.1000088107.

88. 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. https://doi.org/10.1104/pp.109.150607.

89. Delph, L.F., and Kelly, J.K. (2014). On the importance of balancing selection in plants. New Phytol. *201*, 45–56. https://doi.org/10.1111/nph.12441.

90. Eckshtain-Levi, N., Weisberg, A.J., and Vinatzer, B.A. (2018). The population genetic test Tajima's D identifies genes encoding pathogen-associated molecular patterns and other virulence-related genes in *Ralstonia solanacearum*. Mol. Plant Pathol. *19*, 2187–2192. https://doi.org/10.1111/mpp.12688.

91. Zhao, H., Sun, S., Ding, Y., Wang, Y., Yue, X., Du, X., Wei, Q., Fan, G., Sun, H., Lou, Y., et al. (2021). Analysis of 427 genomes reveals moso bamboo population structure and genetic basis of property traits. Nat. Commun. *12*, 5466. https://doi.org/10.1038/s41467-021-25795-x.

92. Lin, X., Liu, B., Weller, J.L., Abe, J., and Kong, F. (2021). Molecular mechanisms for the photoperiodic regulation of flowering in soybean. J. Integr. Plant Biol. *63*, 981–994. https://doi.org/10.1111/jipb.13021.

93. Luo, X., Yin, M., and He, Y. (2021). Molecular genetic understanding of photoperiodic regulation of flowering time in *Arabidopsis* and soybean. Int. J. Mol. Sci. *23*, 466. https://doi.org/10.3390/ijms23010466.

94. Lee, S.H., Choi, C.W., Park, K.M., Jung, W.H., Chun, H.J., Baek, D., Cho, H.M., Jin, B.J., Park, M.S., No, D.H., et al. (2021). Diversification in functions and expressions of soybean *FLOWERING LOCUS T* genes fine-tunes seasonal flowering. Front. Plant Sci. *12*, 613675. https://doi.org/10.3389/fpls.2021.613675.

95. Liu, W., Jiang, B., Ma, L., Zhang, S., Zhai, H., Xu, X., Hou, W., Xia, Z., Wu, C., Sun, S., et al. (2018). Functional diversification of *Flowering Locus T* homologs in soybean: *GmFT1a* and *GmFT2a*/*5a* have opposite roles in controlling flowering and maturation. New Phytol. *217*, 1335–1345. https://doi.org/10.1111/nph.14884.

96. Jamet, J.-P., and Chaumet, J.-M. (2016). Soybean in China: adaptating to the liberalization. OCL *23*, D604. https://doi.org/10.1051/ocl/2016044.

97. Bao, A., Chen, H., Chen, L., Chen, S., Hao, Q., Guo, W., Qiu, D., Shan, Z., Yang, Z., Yuan, S., et al. (2019). CRISPR/Cas9-mediated targeted mutagenesis of *GmSPL9* genes alters plant architecture in soybean. BMC Plant Biol. *19*, 131. https://doi.org/10.1186/s12870-019-1746-6.

98. Zhao, D., Zheng, H., Li, J., Wan, M., Shu, K., Wang, W., Hu, X., Hu, Y., Qiu, L., and Wang, X. (2024). Natural variation in the promoter of *GmSPL9d* affects branch number in soybean. Int. J. Mol. Sci. *25*, 5991. https://doi.org/10.3390/ijms25115991.

99. Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol. Biol. Evol. *32*, 244–257. https://doi.org/10.1093/molbev/msu269.

100. Cai, X., Jia, B., Sun, M., and Sun, X. (2022). Insights into the regulation of wild soybean tolerance to salt-alkaline stress. Front. Plant Sci. *13*, 1002302. https://doi.org/10.3389/fpls.2022.1002302.

101. Jo, L., Pelletier, J.M., Goldberg, R.B., and Harada, J.J. (2024). Genome-wide profiling of soybean WRINKLED1 transcription factor binding sites provides insight into seed storage lipid biosynthesis. Proc. Natl. Acad. Sci. USA *121*, e2415224121. https://doi.org/10.1073/pnas.2415224121.

102. Zhang, H., Goettel, W., Song, Q., Jiang, H., Hu, Z., Wang, M.L., and An, Y. C. (2020). Selection of *GmSWEET39* for oil and protein improvement in soybean. PLoS Genet. *16*, e1009114. https://doi.org/10.1371/journal.pgen.1009114.

103. Lu, L., Wei, W., Li, Q.T., Bian, X.H., Lu, X., Hu, Y., Cheng, T., Wang, Z.Y., Jin, M., Tao, J.J., et al. (2021). A transcriptional regulatory module controls lipid accumulation in soybean. New Phytol. *231*, 661–678. https://doi.org/10.1111/nph.17401.

104. Kwon, K.M., Masonbrink, R.E., Maier, T.R., Gardner, M.N., Severin, A.J., Baum, T.J., and Mitchum, M.G. (2024). Comparative transcriptomic analysis of soybean cyst nematode inbred populations non-adapted or adapted on soybean *rhg1-a*/*Rhg4*-Mediated resistance. Phytopathology *114*, 2341–2350. https://doi.org/10.1094/PHYTO-03-24-0095-R.

105. Liu, S., Kandoth, P.K., Warren, S.D., Yeckel, G., Heinz, R., Alden, J., Yang, C., Jamai, A., El-Mellouki, T., Juvale, P.S., et al. (2012). A soybean cyst nematode resistance gene points to a new mechanism of plant resistance to pathogens. Nature *492*, 256–260. https://doi.org/10.1038/nature11651.

106. Wei, X., Qiu, J., Yong, K., Fan, J., Zhang, Q., Hua, H., Liu, J., Wang, Q., Olsen, K.M., Han, B., et al. (2021). A quantitative genomics map of rice provides genetic insights and guides breeding. Nat. Genet. *53*, 243–253. https://doi.org/10.1038/s41588-020-00769-9.

107. Gavrin, A., Chiasson, D., Ovchinnikova, E., Kaiser, B.N., Bisseling, T., and Fedorova, E.E. (2016). VAMP721a and VAMP721d are important for pectin dynamics and release of bacteria in soybean nodules. New Phytol. *210*, 1011–1021. https://doi.org/10.1111/nph.13837.

108. Indrasumunar, A., Searle, I., Lin, M.H., Kereszt, A., Men, A., Carroll, B.J., and Gresshoff, P.M. (2011). Nodulation factor receptor kinase 1alpha controls nodule organ number in soybean (*Glycine max* L. Merr). Plant J. *65*, 39–50. https://doi.org/10.1111/j.1365-313X.2010.04398.x.

109. Ma, C., Wang, J., Gao, Y., Dong, X., Feng, H., Yang, M., Yu, Y., Liu, C., Wu, X., Qi, Z., et al. (2024). The type III effector NopL interacts with GmREM1a and GmNFR5 to promote symbiosis in soybean. Nat. Commun. *15*, 5852. https://doi.org/10.1038/s41467-024-50228-w.

110. Indrasumunar, A., Kereszt, A., Searle, I., Miyagi, M., Li, D., Nguyen, C.D. T., Men, A., Carroll, B.J., and Gresshoff, P.M. (2010). Inactivation of duplicated nod factor receptor 5 (NFR5) genes in recessive loss-of-function non-nodulation mutants of allotetraploid soybean (*Glycine max* L. Merr.). Plant Cell Physiol. *51*, 201–214. https://doi.org/10.1093/pcp/pcp178.

111. Žilić, S., Akıllıoğlu, H.G., Serpen, A., Perić, V., and Gökmen, V. (2013). Comparisons of phenolic compounds, isoflavones, antioxidant capacity and oxidative enzymes in yellow and black soybeans seed coat and dehulled bean. Eur. Food Res. Technol. *237*, 409–418. https://doi.org/10.1007/s00217-013-2005-y.

112. Zhang, L., Jia, R., Liu, L., Shen, W., Fang, Z., Zhou, B., and Liu, B. (2023). Seed coat colour and structure are related to the seed dormancy and overwintering ability of crop-to-wild hybrid soybean. AoB Plants *15*, plad081. https://doi.org/10.1093/aobpla/plad081.

113. Wang, J., Zhang, L., Wang, S., Wang, X., Li, S., Gong, P., Bai, M., Paul, A., Tvedt, N., Ren, H., et al. (2025). AlphaFold-guided bespoke gene editing enhances field-grown soybean oil contents. Adv. Sci. (Weinh) *12*, e2500290. https://doi.org/10.1002/advs.202500290.

114. Kajiya-Kanegae, H., Nagasaki, H., Kaga, A., Hirano, K., Ogiso-Tanaka, E., Matsuoka, M., Ishimori, M., Ishimoto, M., Hashiguchi, M., Tanaka, H., et al. (2021). Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. DNA Res. *28*, dsaa032. https://doi.org/10.1093/dnares/dsaa032.

115. Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., Zhang, M., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. *18*, 161. https://doi.org/10.1186/s13059-017-1289-9.

116. Torkamaneh, D., and Belzile, F. (2015). Scanning and filling: ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing Data. PLoS One *10*, e0131533. https://doi.org/10.1371/journal.pone.0131533.

117. Valliyodan, B., Dan, Q., Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C., Li, Y., Joshi, T., Song, L., et al. (2016). Landscape of genomic diversity and trait discovery in soybean. Sci. Rep. *6*, 23598. https://doi.org/10.1038/srep23598.

118. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

119. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

120. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, 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. https://doi.org/10.1101/gr.107524.110.

121. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164. https://doi.org/10.1093/nar/gkq603.

122. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

123. Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics *32*, 1749–1751. https://doi.org/10.1093/bioinformatics/btw044.

124. 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. https://doi.org/10.1086/519795.

125. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.

126. 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. https://doi.org/10.1016/j.ajhg.2010.11.011.

127. Yin, L., Zhang, H., Tang, Z., Yin, D., Fu, Y., Yuan, X., Li, X., Liu, X., and Zhao, S. (2023). HIBLUP: an integration of statistical models on the BLUP framework for efficient genetic evaluation using big genomic data. Nucleic Acids Res. *51*, 3501–3512. https://doi.org/10.1093/nar/gkad074.

128. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296. https://doi.org/10.1093/nar/gkab301.

129. Francis, R.M. (2017). pophelper: an R package and web app to analyse and visualize population structure. Mol. Ecol. Resour. *17*, 27–32. https://doi.org/10.1111/1755-0998.12509.

130. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. *8*, e1002967. https://doi.org/10.1371/journal.pgen.1002967.

131. Fitak, R.R. (2021). OptM: estimating the optimal number of migration edges on population trees using Treemix. Biol. Methods Protoc. *6*, bpab017. https://doi.org/10.1093/biomethods/bpab017.

132. Liu, Y., Zhang, Y., Liu, X., Shen, Y., Tian, D., Yang, X., Liu, S., Ni, L., Zhang, Z., Song, S., et al. (2023). SoyOmics: A deeply integrated database on soybean multi-omics. Mol. Plant *16*, 794–797. https://doi.org/10.1016/j.molp.2023.03.011.

133. Yang, Z., Luo, C., Pei, X., Wang, S., Huang, Y., Li, J., Liu, B., Kong, F., Yang, Q.Y., and Fang, C. (2024). SoyMD: a platform combining multi-omics data with various tools for soybean research and breeding. Nucleic Acids Res. *52*, D1639–D1650. https://doi.org/10.1093/nar/gkad786.

134. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

135. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354. https://doi.org/10.1038/ng.548.

136. Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., Liu, Y., Feng, J., Chen, H., He, Y., et al. (2023). TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. Mol. Plant *16*, 1733–1742. https://doi.org/10.1016/j.molp.2023.09.010.

137. Pook, T., Mayer, M., Geibel, J., Weigend, S., Cavero, D., Schoen, C.C., and Simianer, H. (2020). Improving imputation quality in BEAGLE for crop and livestock data. G3 (Bethesda) *10*, 177–188. https://doi.org/10.1534/g3.119.400798.

138. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

139. Vatsiou, A.I., Bazin, E., and Gaggiotti, O.E. (2016). Detection of selective sweeps in structured populations: a comparison of recent methods. Mol. Ecol. *25*, 89–103. https://doi.org/10.1111/mec.13360.

140. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

141. Zhang, C., Dong, S.S., Xu, J.Y., He, W.M., and Yang, T.L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics *35*, 1786–1788. https://doi.org/10.1093/bioinformatics/bty875.

142. Wang, S., Yokosho, K., Guo, R., Whelan, J., Ruan, Y.L., Ma, J.F., and Shou, H. (2019). The soybean sugar transporter GmSWEET15 mediates sucrose export from endosperm to early embryo. Plant Physiol. *180*, 2133–2141. https://doi.org/10.1104/pp.19.00641.

143. Liang, S., Duan, Z., He, X., Yang, X., Yuan, Y., Liang, Q., Pan, Y., Zhou, G., Zhang, M., Liu, S., et al. (2024). Natural variation in *GmSW17* controls seed size in soybean. Nat. Commun. *15*, 7417. https://doi.org/10.1038/s41467-024-51798-5.

144. Wang, Y., Cheng, J., Guo, Y., Li, Z., Yang, S., Wang, Y., and Gong, Z. (2024). Phosphorylation of ZmAL14 by ZmSnRK2.2 regulates drought resistance through derepressing *ZmROP8* expression. J. Integr. Plant Biol. *66*, 1334–1350. https://doi.org/10.1111/jipb.13677.

145. 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. https://doi.org/10.1038/s41467-022-34153-4.

146. Liu, S., Fan, L., Liu, Z., Yang, X., Zhang, Z., Duan, Z., Liang, Q., Imran, M., Zhang, M., and Tian, Z. (2020). A Pd1-Ps-P1 feedback loop controls pubescence density in soybean. Mol. Plant *13*, 1768–1783. https://doi.org/10.1016/j.molp.2020.10.004.

147. Kong, L., Cheng, J., Zhu, Y., Ding, Y., Meng, J., Chen, Z., Xie, Q., Guo, Y., Li, J., Yang, S., et al. (2015). Degradation of the ABA co-receptor ABI1 by PUB12/13 U-box E3 ligases. Nat. Commun. *6*, 8630. https://doi.org/10.1038/ncomms9630.

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse monoclonal anti-GFP	TransGene	Cat#HT801; RRID: AB_2922385
Mouse monoclonal anti-HA	MBL	Cat#M180-7; RRID: AB_11124961
Mouse monoclonal anti-FLAG	MBL	Cat#M185-7; RRID: AB_2687989
Chemicals, peptides, and recombinant proteins		
FLAG magnetic beads	MBL	Cat#M185-10R
GFP magnetic beads	Lablead	Cat#PGM025
Ni NTA Beads 6FF	Smart-lifesciences	Cat#SA005100
Cocktail	Roche	Cat#04693132001
DO Supplement-Leu/-Trp	Coolaber	Cat#PM2220
DO Supplement -Ade/-His/-Leu/-Trp	Coolaber	Cat#PM2110
DO Supplement -Trp/-Ura	Coolaber	Cat#PM2260
X-GAL	Coolaber	Cat#CX11921
X-α-GAL	Yeasen	Cat#10903ES60
PEG 3350	Sigma-Aldrich	Cat#P4338
Triton X-100	Sigma-Aldrich	Cat#T8787
D-(+)Galactose	Coolaber	Cat#CG5481
PMSF	Coolaber	Cat#SL1079
Bacterial and virus strains		
strain EGY48	Home-made	N/A
strain NMY51	Home-made	N/A
BL21(DE3) competent cell	Sangon	Cat#B528414
Critical commercial assays		
KOD FX Neo	TOYOBO	Cat#KFX-201
2×Taq Plus MasterMix (Dye)	CWBIO	Cat#CW2849L
pEASY-Uni Seamless Cloning and Assembly Kit	TransGen	Cat#CU101
Luciferase assay reagent	Promega	Cat#E1483
Recombinant DNA		
Plasmid:*AD-GmNST1A^{Hap1}*	This paper	N/A
Plasmid:*AD-GmNST1A^{Hap2}*	This paper	N/A
Plasmid:*GmSHAT1-5-pLacZ*	This paper	N/A
Plasmid:*GmRj5-pBT3-SUC*	This paper	N/A
Plasmid:*NopL-pPR3-N*	This paper	N/A
Plasmid:*GmREM1a-pPR3-N*	This paper	N/A
Plasmid:*GmSWEET30a-LUC*	This paper	N/A
Plasmid:*GmSHAT1-5-LUC*	This paper	N/A
Plasmid:*GmNST1A^{Hap1}-GFP*	This paper	N/A
Plasmid:*GmNST1A^{Hap2}-GFP*	This paper	N/A
Plasmid:*GmRj5$^{Hap1\text{-}Hap4}$-YFP^{C}*	This paper	N/A
Plasmid:*NopL-YFP^{N}*	This paper	N/A
Plasmid:*GmREM1a-YFP^{N}*	This paper	N/A
Plasmid:*His-GmNST1A^{Hap1}*	This paper	N/A
Plasmid:*His-GmNST1A^{Hap2}*	This paper	N/A
Plasmid:*GmRj5$^{Hap1\text{-}Hap4}$-HA*	This paper	N/A

(*Continued on next page*)

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Plasmid:*GmREM1a-GFP*	This paper	N/A
Plasmid:*NopL-FLAG*	This paper	N/A
Oligonucleotides		
Primers listed in Table S5	BioSune Biotechnology	N/A
Deposited Data		
Soybean DNA sequencing data	Lu et al.[72]	NGDC BioProject: PRJCA001691
Soybean DNA sequencing data	Liu et al.[81]	NGDC BioProject: PRJCA002030
Soybean DNA sequencing data	Yang et al.[31]	NGDC BioProject: PRJCA002554
Soybean DNA sequencing data	Kajiya-Kanegae et al.[114]	BioProject: PRJDB7250
Soybean DNA sequencing data	Kajiya-Kanegae et al.[114]	BioProject: PRJDB7386
Soybean DNA sequencing data	Kajiya-Kanegae et al.[114]	BioProject: PRJDB7786
Soybean DNA sequencing data	N/A	BioProject: PRJEB1942
Soybean DNA sequencing data	N/A	BioProject: PRJEB31453
Soybean DNA sequencing data	N/A	BioProject: PRJNA175477
Soybean DNA sequencing data	N/A	BioProject: PRJNA227063
Soybean DNA sequencing data	N/A	BioProject: PRJNA243933
Soybean DNA sequencing data	N/A	BioProject: PRJNA248222
Soybean DNA sequencing data	Fang et al.[115]	BioProject: PRJNA257011
Soybean DNA sequencing data	N/A	BioProject: PRJNA274295
Soybean DNA sequencing data	Torkamaneh and Belzile[116]	BioProject: PRJNA287266
Soybean DNA sequencing data	Valliyodan et al.[117]	BioProject: PRJNA289660
Soybean DNA sequencing data	N/A	BioProject: PRJNA291452
Soybean DNA sequencing data	N/A	BioProject: PRJNA294227
Soybean DNA sequencing data	N/A	BioProject: PRJNA295763
Soybean DNA sequencing data	N/A	BioProject: PRJNA356132
Soybean DNA sequencing data	N/A	BioProject: PRJNA383915
Soybean DNA sequencing data	N/A	BioProject: PRJNA384190
Soybean DNA sequencing data	Fang et al.[115]	BioProject: PRJNA394629
Soybean DNA sequencing data	N/A	BioProject: PRJNA449253
Soybean DNA sequencing data	N/A	BioProject: PRJNA484078
Soybean DNA sequencing data	N/A	BioProject: PRJNA552939
Soybean DNA sequencing data	Kim et al.[30]	BioProject: PRJNA555366
Soybean DNA sequencing data	N/A	BioProject: PRJNA597660
Soybean DNA sequencing data	N/A	BioProject: PRJNA639876
Soybean DNA sequencing data	N/A	BioProject: PRJNA681974
Soybean DNA sequencing data	N/A	BioProject: PRJNA743225
Software and Algorithms		
BWA 0.7.18-r1243-dirty	Li and Durbin[118]	https://sourceforge.net/projects/bio-bwa/
SAMtools 1.20	Li et al.[119]	https://sourceforge.net/projects/samtools/
GATK v4.2	McKenna et al.[120]	https://github.com/broadinstitute/gatk/
ANNOVAR	Wang et al.[121]	https://annovar.openbioinformatics.org/en/latest/
VCFtools 0.1.16	Danecek et al.[122]	https://vcftools.github.io/man_latest.html
BCFtools 1.20	Narasimhan et al.[123]	https://github.com/samtools/BCFtools
Plink v1.90b6.21	Purcell et al.[124]	www.cog-genomics.org/plink/1.9/
Admixture v1.3.0	Alexander et al.[125]	www.genetics.ucla.edu/software/admixture
GCTA v1.94.1	Yang et al.[126]	https://yanglab.westlake.edu.cn/software/gcta/
HIBLUP v1.5.0	Yin et al.[127]	https://www.hiblup.com/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R 4.2.0	N/A	https://www.R-project.org/
PHYLIP 3.697	N/A	https://phylipweb.github.io/phylip/
iTOL	Letunic and Bork[128]	https://itol.embl.de/
Pophelper v2.3.1	Francis[129]	https://www.royfrancis.com/pophelper/
Treemix v1.12	Pickrell and Pritchard[130]	https://bitbucket.org/nygcresearch/treemix/
OptM v 0.1.8	Fitak[131]	https://cran.r-project.org/web/packages/OptM/
MEGAX	N/A	https://www.megasoftware.net/
SoyOmics	Liu et al.[132]	https://ngdc.cncb.ac.cn/soyomics/
SoyMD	Yang et al.[133]	https://yanglab.hzau.edu.cn/SoyMD/
BEDtools v2.31.1	Quinlan and Hall[134]	http://bedtools.readthedocs.io/
RAiSD v2.9	Alachiotis and Pavlidis[53]	https://github.com/alachins/raisd
Genomics_general	N/A	https://github.com/simonhmartin/genomics_general
EMMAX beta-07Mar2010	Kang et al.[135]	http://csg.sph.umich.edu/kang/emmax/
TBtools v2.210	Chen et al.[136]	https://github.com/CJ-Chen/TBtools-II
BEAGLE v5.4	Pook et al.[137]	https://faculty.washington.edu/browning/beagle/
SeqKit v2.8.2	Shen et al.[138]	https://github.com/shenwei356/seqkit
Xpclr	Vatsiou et al.[139]	https://github.com/hardingnj/xpclr

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

In this study, the phenotypic data were sourced from the Chinese national soybean evaluation records and databases, while the resequencing data were obtained from multiple published soybean studies. Consequently, detailed descriptions of soybean cultivation, sampling, and resequencing procedures are omitted. For complete information on data sources, please refer to the key resources table.

METHOD DETAILS

Soybean accessions and resequencing data

In this study, we retrieved high-quality paired-end soybean resequencing data from public databases, including the NCBI (National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/) and the National Genomics Data Center (NGDC, https://ngdc.cncb.ac.cn/). The datasets encompassed the following projects: PRJCA001691, PRJCA002030, PRJCA002554, PRJDB7250, PRJDB7386, PRJDB7786, PRJEB1942, PRJEB31453, PRJNA175477, PRJNA227063, PRJNA243933, PRJNA248222, PRJNA257011, PRJNA274295, PRJNA287266, PRJNA289660, PRJNA291452, PRJNA294227, PRJNA295763, PRJNA356132, PRJNA383915, PRJNA384190, PRJNA394629, PRJNA449253, PRJNA484078, PRJNA552939, PRJNA555366, PRJNA597660, PRJNA639876, PRJNA681974, and PRJNA743225. In total, the high-quality paired-end sequencing data for 8,105 soybean samples were download for subsequent analysis. These samples comprised 1,334 wild soybeans, 5,716 cultivated soybeans, 1,045 landraces, and 10 *Glycine gracilis* that originated from the United States, Brazil, China, Canada, Japan and the Korean Peninsula, Russia, Europe, Vietnam, Indonesia, and other countries or regions. The Chinese materials were further classified into several groups based on their geographic distribution: Northern China, the Huanghuai region, and Southern China. Phenotypic data of the soybean materials were collected through literature reviews, germplasm records of soybean varieties, national certification documents, and online database queries (https://ngdc.cncb.ac.cn/soyomics/).[132] The distribution of soybean accessions and the visualization of the world map were performed by R packages sf, rnaturalearth and leaflet.

Variants calling and annotation

We assessed the quality of the downloaded resequencing data using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For data containing adapter sequences or low-quality reads, we performed filtering using Trimmomatic.[140] The parameters were set as follows: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 to remove adapter sequences; MINLEN:36 to discard reads shorter than 36 bp; SLIDINGWINDOW:4:15 to trim low-quality bases using a 4-bp sliding window with an average quality threshold of 15; and LEADING:5 and TRAILING:3 to remove low-quality bases from the 5' and 3' ends, respectively. Other parameters were set to default values.

We used the version 2.0 of the high-quality reference genome of 'Zhonghuang 13' (ZH13_v2.0)[35,36] for sequence alignment. The resequencing data were aligned to the reference genome using BWA-MEM.[118] SAM files were converted to BAM format using

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

SAMtools,[119] followed by sorting with the sort command. Duplicate sequences in the alignment results were marked using the MarkDuplicates tool in Genome Analysis Toolkit (GATK) v4.2[120] with default parameters, allowing for recognition in subsequent variant detection. Variant calling for each individual was performed using the HaplotypeCaller module in GATK, based on the whole-genome alignment of the resequencing data.

We conducted stringent quality control on the obtained soybean genomic variant dataset. SNP and InDel sites were separately extracted using the SelectVariants command in GATK. Filtering parameters for InDels were set as QD < 2.0, FS > 200.0, SOR > 10.0, InbreedingCoeff < −0.8, and ReadPosRankSum < −20.0. For SNPs, filtering parameters were QD < 2.0, FS > 60.0, SOR > 5.5, MQ < 40.0, MQRankSum < −12.5, and ReadPosRankSum < −8.0. The identified variants were annotated using ANNOVAR.[121]

Genomic structure analysis

We performed principal component analysis (PCA) on 8,105 soybean samples using GCTA.[126] PCA was conducted using the parameters –grm to specify the constructed GRM file and –pca 4 to set the number of principal components to four. The output included files ending with eigenval and eigenvec, which were visualized using R.

Ancestral component inference of the collected soybean samples was performed using ADMIXTURE.[125] Bed format files obtained from PCA analysis were analyzed for K = 2-7. The cross-validation (CV) error for each K value was examined, with the minimum CV error indicating the optimal number of clusters. Visualization was conducted using the R package pophelper.[129] A phylogenetic tree was constructed using the neighbor-joining method implemented in the PHYLIP software package and subsequently visualized through the iTOL online platform.[128] Linkage disequilibrium decay analysis was performed using PopLDdecay.[141] The maximum likelihood tree for soybean populations was constructed by the software TreeMix[131] and visualized with R script. In order to determine the number of migration edges in TreeMix analysis, we performed 5 replicates for each m value and ultimately selected the optimal value according to the R package OptM.

Genome-wide association analysis

Genome-wide association studies (GWAS) were conducted using the MLMA function of GCTA software,[126] based on a mixed linear model. Due to the lack of phenotypic data for wild soybeans and population structure, GWAS was performed only on cultivated soybeans and landraces. In the analysis, the first three principal components were used as covariates, and a preliminary significance threshold of 1×10^{-5} was set. Candidate gene annotation referred to the SoyOmics (https://ngdc.cncb.ac.cn/soyomics/) and SoyMD (https://yanglab.hzau.edu.cn/SoyMD/) databases.

Selection signal detection

Selection signals were detected using multiple methods. Based on the identified high-quality SNPs, we calculated nucleotide diversity (π) for wild soybeans, cultivated soybeans, landraces, and improved varieties using VCFtools[122] with a 100-kb sliding window and a 10-kb step size. The population fixation index (F_{ST}) was calculated between wild and cultivated soybeans, between landraces and improved cultivars, among cultivars from different countries, and among varieties from different eras with a 100-kb sliding window and a 10-kb step size. Tajima's D was calculated using VCFtools with a 100-kb sliding window. XP-CLR values across the soybean genome were computed using the Python version of package 'xpclr'[139] with parameters –ld 0.95, –maxsnps 1000, –size 100000, and –step 100000. The genome-wide μ statistic was calculated for cultivated soybeans and landraces using package 'RAiSD',[53] and selective sweeps were identified based on these values. Python scripts were used to transform the calculation results of F_{ST}, XP-CLR, μ statistics, π ratio and then Z-scores were computed. A threshold of Z-score $\geq$ 1.96 was set to identify genomic regions of candidate selective sweeps, and a custom script was used to extract selected genes from the ZH13_v2 annotation files. To avoid fragmentation of candidate intervals, we merged overlapping windows using the merge function in bedtools. The high-quality candidate selective sweeps were determined by support from at least two statistics.

Analysis of gene haplotypes

We utilized a python script and the pysam library for haplotype analysis from VCF files, aimed at identifying and counting the types and numbers of different haplotypes in our dataset. Haplotypes were defined based on coding-region variants and known QTNs. The positional information for all genes was queried from ZH13_v2 annotation. Missing genotypes were imputed using BEAGLE software (version 5.4).[137]

Analysis of gene flow

D-statistics for different combinations among wild soybeans, black soybeans, and landraces were computed using the ABBABABAwindows.py script (https://github.com/simonhmartin/genomics_general). The outgroup was set to the perennial *Glycine tomentella*. To detect the introgression between wild and cultivated soybeans from different eras, f_d statistic was computed based on a tree form ([(P1, P2), P3, O]), where P1 was fixed as the American cultivated soybeans and the perennial *Glycine tomentella* was fixed as the outgroup. P2 was set as cultivated soybeans from different eras, and P3 was set as wild soybeans accessions in China. The f_d statistic was computed in 100 kb non-overlapping windows with the same python script. The output results with $D > 0$ and $0 <= f_d <= 1$

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

were retained. The valid f_d values were converted into Z-scores, and regions with Z-scores >= 1.96 were considered as candidate introgression regions.

f_3 statistics calculation

To test for evidence of admixture, we computed the three-population statistics using the threepop program from the TreeMix package. Briefly, genotype counts for different populations (wild soybeans, black soybeans, and landraces) were converted into TreeMix input format (allele counts per population per SNP) and compressed with gzip. We specified a block size of 2000 SNPs for jackknife estimation of the standard error. By convention, a significantly negative f_3 value from f_3 (C; A, B) indicates that population C derives a measurable fraction of its ancestry from a mixture of populations A and B. Note that a positive f_3 (with a large positive Z-score) indicates that the two-way admixture model is not strongly supported under these parameters. We applied this procedure to all combinations of wild soybeans, black soybeans clade 1/2, and landraces. All f_3 values and Z-scores were reported in Table S2.

Genomic selection

We integrated Genomic selection (GS) and GWAS to analyze total content of oil and protein. We employed the genomic best linear unbiased prediction (GBLUP), considering both additive and dominance effects, and incorporated principal component analysis (PCA) results as covariates to predict the Genomic Estimated Breeding Value (GEBV). GEBV was calculated by HIBLUP software (https://www.hiblup.com).[127] We consider the single trait linear mixed model as followed:

$$y = Xb + Rr + \sum_{i=1}^{k} Z_i u_i + e$$

y is an n-dimensional vector of phenotypic records. X is an $n \times n_b$ design matrix for fixed effects and covariates. b is an n_b-dimensional vector of corresponding fixed effects. R is an $n \times n_r$ design matrix for environmental random effects. r is an n_r-dimensional vector of environmental random effects, assumed to follow a normal distribution $N(0, I\sigma_r^2)$, where I is the identity matrix and σ_r^2 is the environmental variance component. k is the total number of genetic random effects considered in the model. Z_i is an $n \times n_i$ design matrix for the i-th genetic random effect. u_i is an n_i-dimensional vector of genetic effects corresponding to the i-th genetic random effect, assumed to follow a multivariate normal distribution $N(0, K_i\sigma_i^2)$. K_i is the genetic relationship matrix for the i-th genetic effect, derived from genomic information. σ_i^2 is the genetic variance component associated with the i-th genetic effect. e is an n-dimensional vector of residual errors, assumed to follow $N(0, I\sigma_e^2)$, where σ_e^2 is the residual variance component. Both additive and dominant genetic effects were included to comprehensively capture the genetic architecture influencing the trait. The additive genetic relationship matrix (K_A) and the dominant genetic relationship matrix (K_D) were constructed based on genome-wide SNP markers. The first three principal components (PC1-PC3) were included as covariates in the fixed effects matrix X to adjust for population stratification and relatedness among individuals. We used five-fold cross-validation to evaluate the phenotypic prediction accuracy of different traits. Accuracy was defined as the Pearson correlation coefficient between GEBVs and observed phenotypic values.

Construction of transgenic materials and measurement of soybean quality traits

The CRISPR/CAS9 construct was based on the pBlu-gRNA vector and CAS9 MDC123 (Addgene plasmids 59188 and 59184) following the method previously described.[142] The 20-bp target sequence (5'-AGTGATACAAACCAAGAGTG-3') was designed in the fifth exon of the *GmSWEET30a* (*Glyma.13G041300*) coding sequence. The synthesized target sequence was inserted into pBlu-gRNA (Addgene plasmids 59188) at the BbsI site. After digested with EcoRI, the generating gRNA cassette was cloned into CAS9 MDC123 (Addgene plasmids 59184). The resulting construct was transformed into cultivated soybean [*Glycine max* (L.) Merr.] ecotype Williams 82 via *Agrobacterium tumefaciens* strain EHA105. For field experiments, soybean plants were cultivated in Changchun, Jilin Province, China. Mature seed oil and protein contents were measured in a Antaris II target spectrometer (Antaris II, ThermoFisher, USA).

Plasmid construction and plant transformation

To create His-GmNST1A^{Hap1} and His-GmNST1A^{Hap2} recombinant proteins, their coding sequence (CDS) fragments were cloned into the pET28a vector. To generate *GmNST1A^{Hap1}-GFP* and *GmNST1A^{Hap2}-GFP* constructs, the CDSs were cloned into the pUC19-35S-GFP vectors. To construct the plasmid *GmSHAT1-5pro:LUC* for Dual-LUC transcriptional activity assay, the 3 kb *GmSHAT1-5* promoter sequence was amplified and cloned into pGreenII0800-LUC vector. To obtain constructs for promoter activity analysis, a 3123 bp promoter sequence of *GmSWEET30a^{Hap1}* or *GmSWEET30a^{Hap2}* were cloned into pGreenII0800-LUC vector.

The full-length CDSs of *GmRj5$^{Hap1\text{-}Hap4}$* were introduced into the pBT3-SUC vector as bait vectors. To construct the prey vectors, we cloned full-length CDSs of *GmREM1a* and *NopL* into the pPR3-N vector. The coding regions of *GmNST1A^{Hap1}* and *GmNST1A^{Hap2}* were cloned into pB42AD vector. A 319 bp promoter sequence of *GmSHAT1-5* was cloned into pLacZi vector.

To generate the *GmRj5$^{Hap1\text{-}Hap4}$-HA* and *GmREM1a-GFP* or *NopL-FLAG* constructs, we cloned CDS of each gene into the pUC19-35S-HA, pUC19-35S-GFP vector, and pUC19-35S-FLAG vector. The CDSs of *GmRj5$^{Hap1\text{-}Hap4}$* were cloned into the pSPYCE (M) vector to generate *GmRj5$^{Hap1\text{-}Hap4}$-YFP^{C}*. *GmREM1a* and *NopL* were cloned into the pSPYNE173 to generate *GmREM1a-YFP^{N}* or *NopL-YFP^{N}*.

Please cite this article in press as: Zhu et al., Genomic atlas of 8,105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.007

Subcellular localization of proteins

The constructs of GmNST1A^{Hap1}-GFP and GmNST1A^{Hap2}-GFP were co-expressed with RFP in *Arabidopsis* protoplasts and incubated in darkness for 16 h as described previously.[143] Transformed protoplasts were observed using a fluorescence microscope (ZEISS980, Carl Zeiss). GFP fluorescence was excited at 488 nm, while RFP fluorescence was excited at 561 nm.

The GmNST1A^{Hap1}-GFP and GmNST1A^{Hap2}-GFP were co-expressed with H2B-mCherry in *Nicotiana benthamiana*, and fluorescence signals were detected using ZEISS980 at 48 h after transformation. H2B-mCherry was used as control for nuclear localization.

Yeast one-hybrid assay

The constructs were co-transformed into yeast strain EGY48, with the empty vector pB42AD and placZ serving as negative controls. The yeast cells were grown on synthetic drop-out/-Trp/-Ura (-TU) medium or synthetic drop-out/-Trp/-Ura with X-gal (-TU+X-gal) medium for 3 days.

Dual membrane system Y2H assay

Different combinations of the constructs were co-transformed into the yeast strain NMY51. Yeast cells were spread on synthetic dropout media without Leu and Trp (DDO) or without Leu, Trp, His, and Ade (QDO) with X-α-gal. The positive control pTSU2-APP and pNubG-Fe65 were co-transformed into strain NMY51. For functional validation, pBT3-SUC-GmRj5$^{Hap1\text{-}Hap4}$ and pBT3-N (bait plasmid) or pOst1-NubI (prey plasmid) were co-transformed into strain NMY51. Plates were cultivated at 30°C for 3 days.

Electrophoretic mobility shift assay

Electrophoretic mobility shift assay (EMSA) was conducted as previously described.[144] The constructs His-GmNST1A^{Hap1} and His-GmNST1A^{Hap2} were transformed into *Escherichia coli* BL21 cells. The recombinant proteins were induced by 0.2 mM of Isopropyl-beta-D-thiogalactopyranoside (IPTG) and incubated overnight at 22°C. The proteins were purified on Ni Sepharose and eluted with elution buffer (50 mM Tris-HCl, pH 7.2 and 250 mM imidazole). Oligonucleotide probes were synthesized and labeled with biotin at the 5′ end (Sangon). The proteins were incubated with 2 nM biotin-labeled probe, cold probe at 27°C for 20 min using a LightShift Chemiluminescent EMSA Kit (Thermo Fisher). The reaction products were separated by 6.5% native polyacrylamide gel electrophoresis and transferred onto a nitrocellulose membrane. The binding signals were carried out following the manufacturer's instructions. The primers were listed in Table S5.

Transcriptional activity assay

The Dual-Luciferase transcriptional activity assay in *Arabidopsis* protoplasts was performed as described previously.[145] *Arabidopsis* protoplasts were isolated from 3-week-old seedlings. The effector constructs GmNST1A^{Hap1}-GFP or GmNST1A^{Hap2}-GFP was co-transformed with *GmSHAT1-5pro:LUC* reporter construct into *Arabidopsis* protoplasts via the polyethylene glycol (PEG)-mediated protocol. After incubation in darkness for 16 hours, total proteins were extracted from the protoplasts with a Dual-Luciferase Reporter Assay Kit (Promega). The LUC/REN ratio was used to calculate *GmSHAT1-5* promoter activity. 35S:REN was used as an internal control.

The promoter activity analysis was performed using *Arabidopsis* protoplasts as previously described.[146] The *GmSWEET30a^{Hap1}pro:LUC* and *GmSWEET30a^{Hap2}pro:LUC* plasmids were used for transient transformation into *Arabidopsis* protoplasts for 16 hours. Total proteins were extracted from the samples using Dual-Luciferase Reporter Assay Kit (Promega). The LUC/REN ratio was used to calculate *GmSWEET30a* promoter activity. 35S:REN was used as an internal control. The primers were listed in Table S5.

Co-immunoprecipitation assay

To determine the interaction of GmRj5$^{Hap1\text{-}Hap4}$-HA and GmREM1a-GFP or NopL-FLAG, the Co-immunoprecipitation (Co-IP) assay in *Arabidopsis* leaf protoplasts was performed as described in a previous study.[147] GmRj5$^{Hap1\text{-}Hap4}$-HA and GmREM1a-GFP or NopL-FLAG were transformed into protoplasts via the PEG-mediated protocol. Total protein was extracted from protoplasts after incubation for 16 h using IP buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.2% Triton X-100, 1 mM PMSF, 5 mM DTT, and protease inhibitor cocktail) and immunoprecipitated by GFP magnetic beads (MBL) or FLAG magnetic beads (Lablead) for 3 hours at 4°C. The beads were washed five times with PBS buffer (10 mM Na_2HPO_4, 2 mM KH_2PO_4, 150 mM NaCl, 2.7 mM KCl). The immunoprecipitated proteins were boiled in SDS loading buffer. The proteins were separated on 8%-15% SDS polyacrylamide gel electrophoresis, subjected to immunoblot analysis and detected with anti-HA antibodies (MBL), anti-GFP antibodies (TransGen) or anti-FLAG antibodies (MBL). The band intensity was quantified by the Image J software.

Bimolecular fluorescence complementation

The BiFC vectors GmRj5$^{Hap1\text{-}Hap4}$-YFP^{C} and GmREM1a-YFP^{N} or NopL-YFP^{N} were transformed into *Agrobacterium tumefaciens* GV3101 and then injected into *Nicotiana benthamiana* leaves. After 48 h, the GFP signal detected and imaged using ZEISS 980 (Carl Zeiss).

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed using R (4.2.0). The statistical tests used for each experiment were listed in figure legends. When comparing phenotypes, if there were two groups, we used two-tailed wilcox.test or two-tailed t.test in R. When there were more than two groups, we used kruskal.test or ANOVA in R, and the post hoc test was performed by dunnTest function in FSA package or TukeyHSD in multcomp package.

Supplemental figures

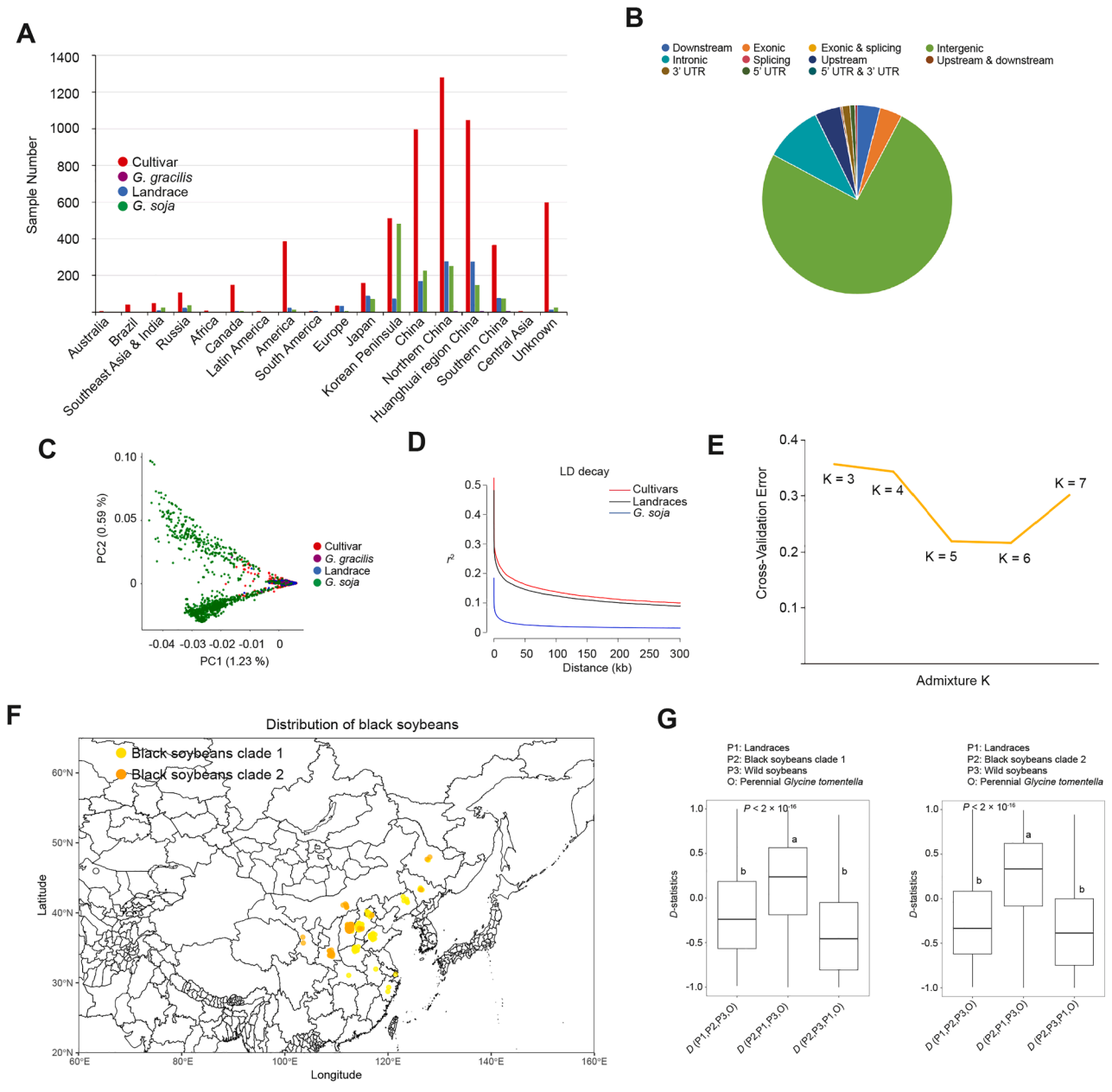


Figure S1. Statistics and genomic variations of 8,105 soybean accessions, related to Figure 1

(A) Numbers of soybean accessions collected from each country.

(B) Annotation of 48,563,254 SNPs.

(C) Principal-component analysis of 8,105 soybean accessions.

(D) LD decay of cultivars, landraces, and wild soybeans. Using $r^2 = 0.1$ as the LD threshold, LD decay distances are approximately 0.2 kb in wild soybeans, 203.4 kb in landraces, and 296 kb in cultivars.

(E) Cross-validation error of admixture analysis from K = 3 to K = 7.

(F) Geographical distribution of two black soybean clades in China.

(G) *D* statistics among wild soybeans, black soybeans, and landraces, calculated using the ABBA-BABA test. Statistical tests were performed by the Kruskal-Wallis test and Dunn's post hoc test in R.

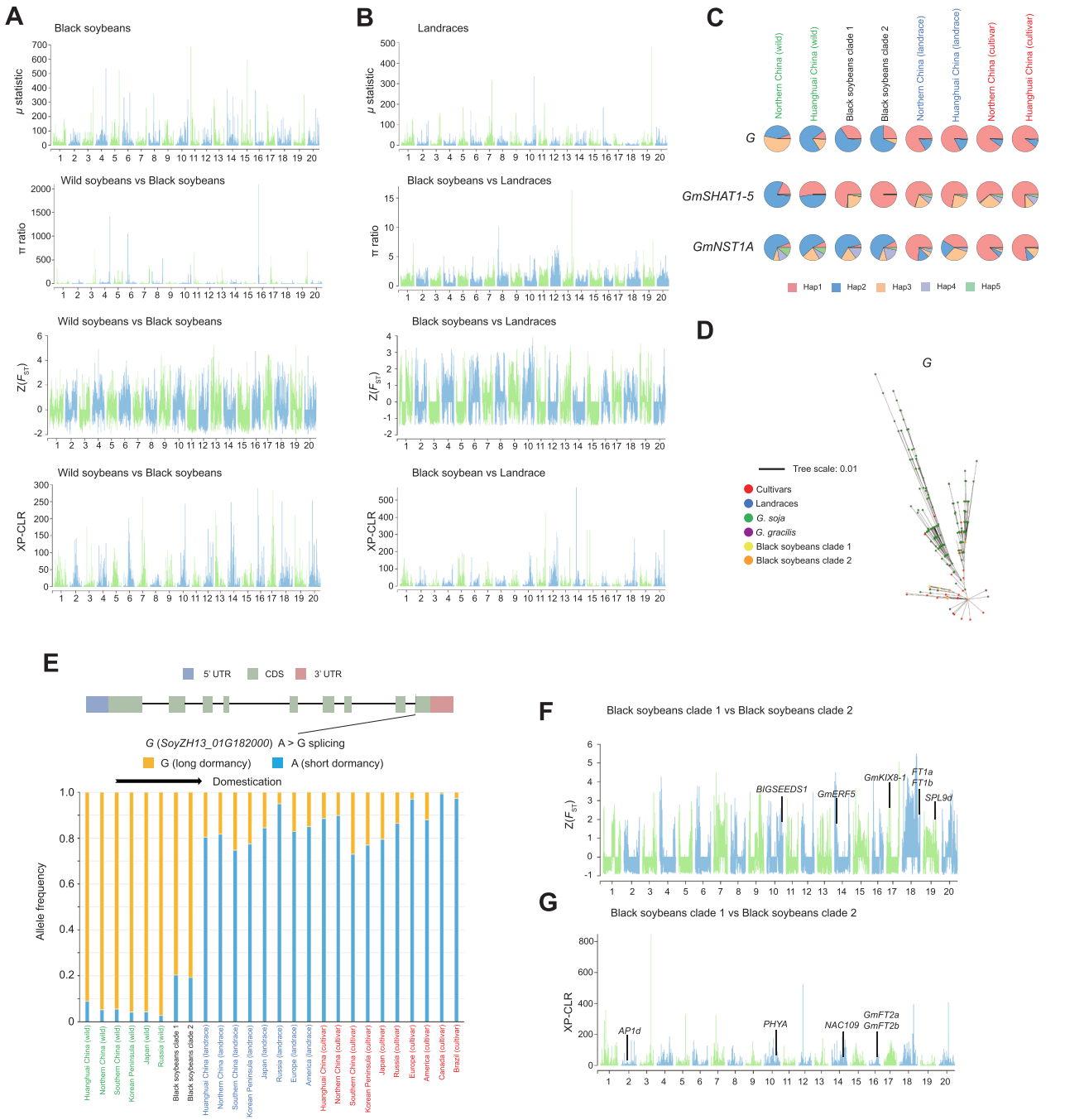


Figure S2. Genes under selection during the soybean domestication, related to Figure 2

(A) Selective sweeps identified in wild soybeans to black soybeans stage, detected by RAiSD, π ratio, F_{ST}, and XP-CLR.

(B) Selective sweeps identified in black soybeans to landrace stage, detected by RAiSD, π ratio, F_{ST}, and XP-CLR.

(C) Haplotype distribution of *G*, *GmSHAT1-5*, and *GmNST1A* across wild soybeans, landraces, cultivars, and black soybeans.

(D) Phylogenetic tree of the *G* gene, with different types of varieties distinguished by different colors. The tree is constructed based on the total SNPs and indels within the gene region, and branch lengths are proportional to genetic distance (scale bar, 0.01 substitutions per site).

(E) Gene structure and functional allele frequency of the *G* gene in different soybean accessions. The *G* gene is not significantly selected during the wild soybeans to black soybeans stage but is selected for the short-dormancy allele in the subsequent black soybeans to landraces stage.

(F) F_{ST} analysis of two clades of black soybeans.

(G) XP-CLR analysis of two clades of black soybeans.

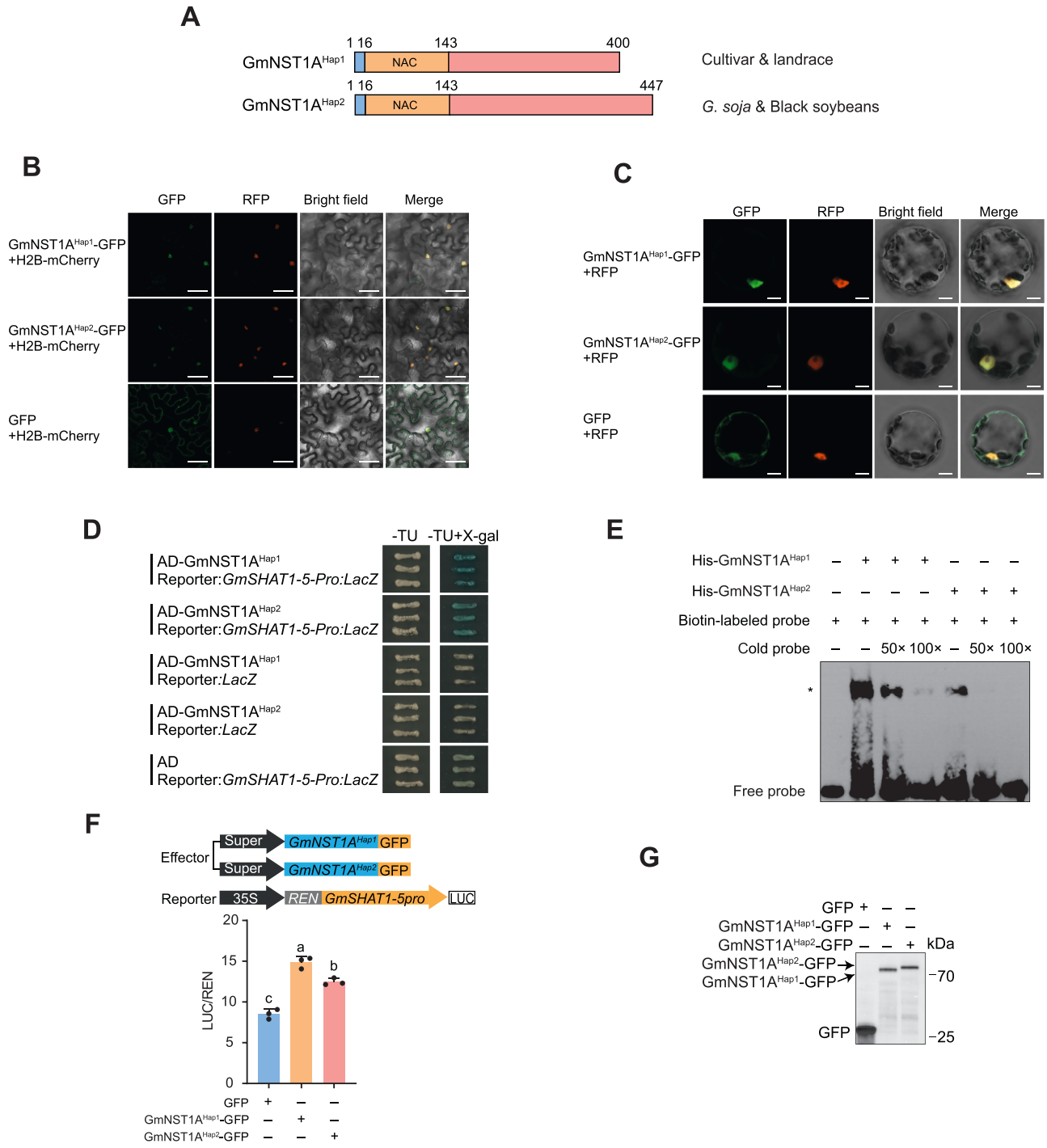


Figure S3. Functional analysis of two *GmNST1A* haplotypes, related to Figure 2

(A) Structure of the two GmNST1A protein variants, with the conserved NAC transcription factor domain highlighted in orange.

(B) Subcellular localization of the two GmNST1A protein variants in tobacco leaves. Fluorescent signals are visualized by confocal microscope. Scale bar, 50 μm.

(C) Subcellular localization of the two GmNST1A protein variants in *Arabidopsis* protoplasts. Fluorescent signals are visualized by confocal microscope. Scale bar, 10 μm.

(D) Interaction of GmNST1A isoforms with the *GmSHAT1-5* promoter in yeast one-hybrid assay. The transfected yeast cells were grown on synthetic drop-out/−Trp/−Ura (−TU) medium (left) and on the same medium supplemented with X-gal (−TU + X-gal; right).

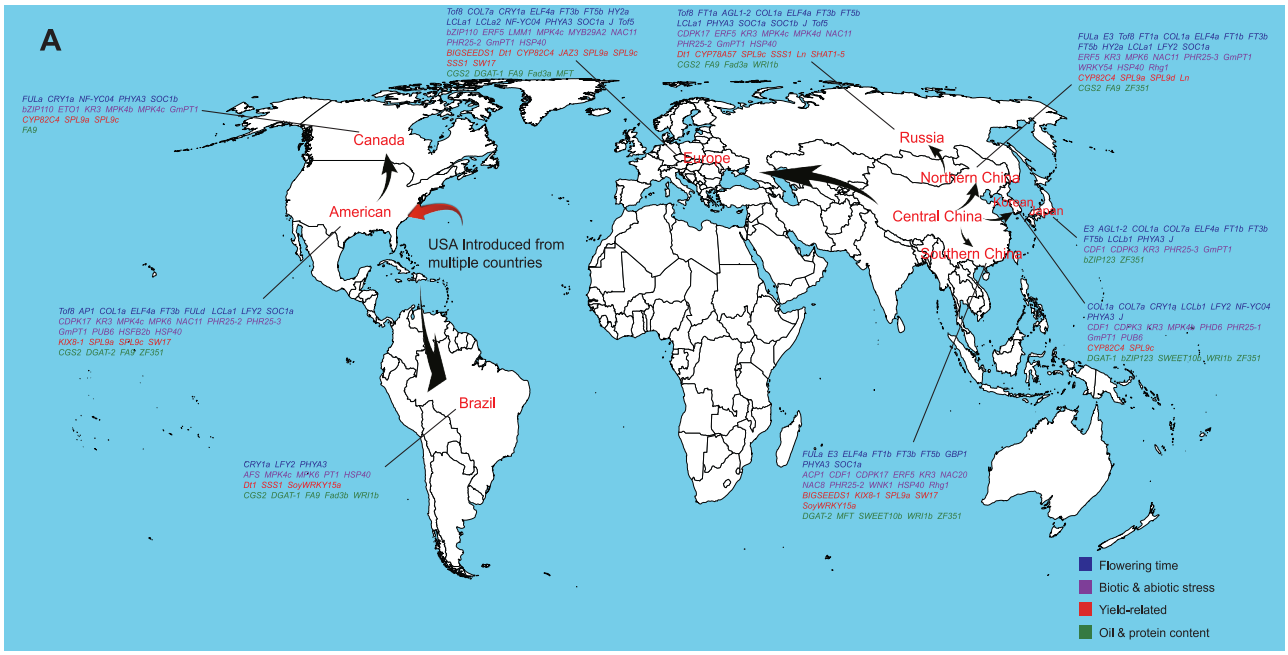
(E) EMSA analysis of the binding of two GmNST1A haplotypes to the *GmSHAT1-5* promoter fragment. A biotin-labeled *GmSHAT1-5* promoter probe is incubated with recombinant His-GmNST1A^{Hap1} or His-GmNST1A^{Hap2} proteins. Lane 1 shows a free probe without protein. Lanes 2–4 contain His-GmNST1A^{Hap1} with no

(legend continued on next page)

competitor (lane 2) and with 50× (lane 3) or 100× (lane 4) molar excess of unlabeled ("cold") probe. Lanes 5–7 contain His-GmNST1A^{Hap2} under the same competitor conditions. Asterisk indicates a GmNST1A-bound probe. The position of the free (unbound) probe is indicated at the bottom.

(F) Dual-LUC assay showing differential activation of the *GmSHAT1-5* promoter by two GmNST1A haplotypes. Top: schematic of the effector and reporter constructs. Effector plasmids express either free GFP or GmNST1A^{Hap1}-GFP or GmNST1A^{Hap2}-GFP fusions under the Super promoter. The reporter contains a Renilla luciferase (REN) gene driven by the constitutive 35S promoter for normalization and a firefly luciferase (LUC) gene under control of the *GmSHAT1-5* promoter. Bottom: quantification of promoter activation in transiently transformed *Arabidopsis* protoplasts. *GmSHAT1-5pro:LUC* and effector-free *GFP*, *Super:GmNST1A^{Hap1}-GFP*, or *Super:GmNST1A^{Hap2}-GFP* were co-transformed into *Arabidopsis* protoplasts. Bars represent mean ± SD from three independent experiments (one-way ANOVA and Tukey's test).

(G) Western blot validation of expression of the two GmNST1A-GFP fusion proteins in the dual-LUC assay. Molecular weight markers (in kDa) are shown on the right.



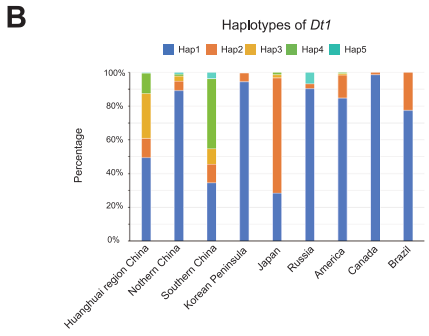


Figure S4. Genes under balancing selection during soybean dissemination, related to Figure 3

(A) Balancing selection was identified by Tajima's D (≥2). Known genes are classified into four categories: flowering time (blue), biotic and abiotic stress (purple), yield-related (red), and oil and protein content (green).

(B) Haplotype distribution of *Dt1* in soybean cultivars. Haplotypes are defined based on six non-synonymous mutations in the gene coding sequence.

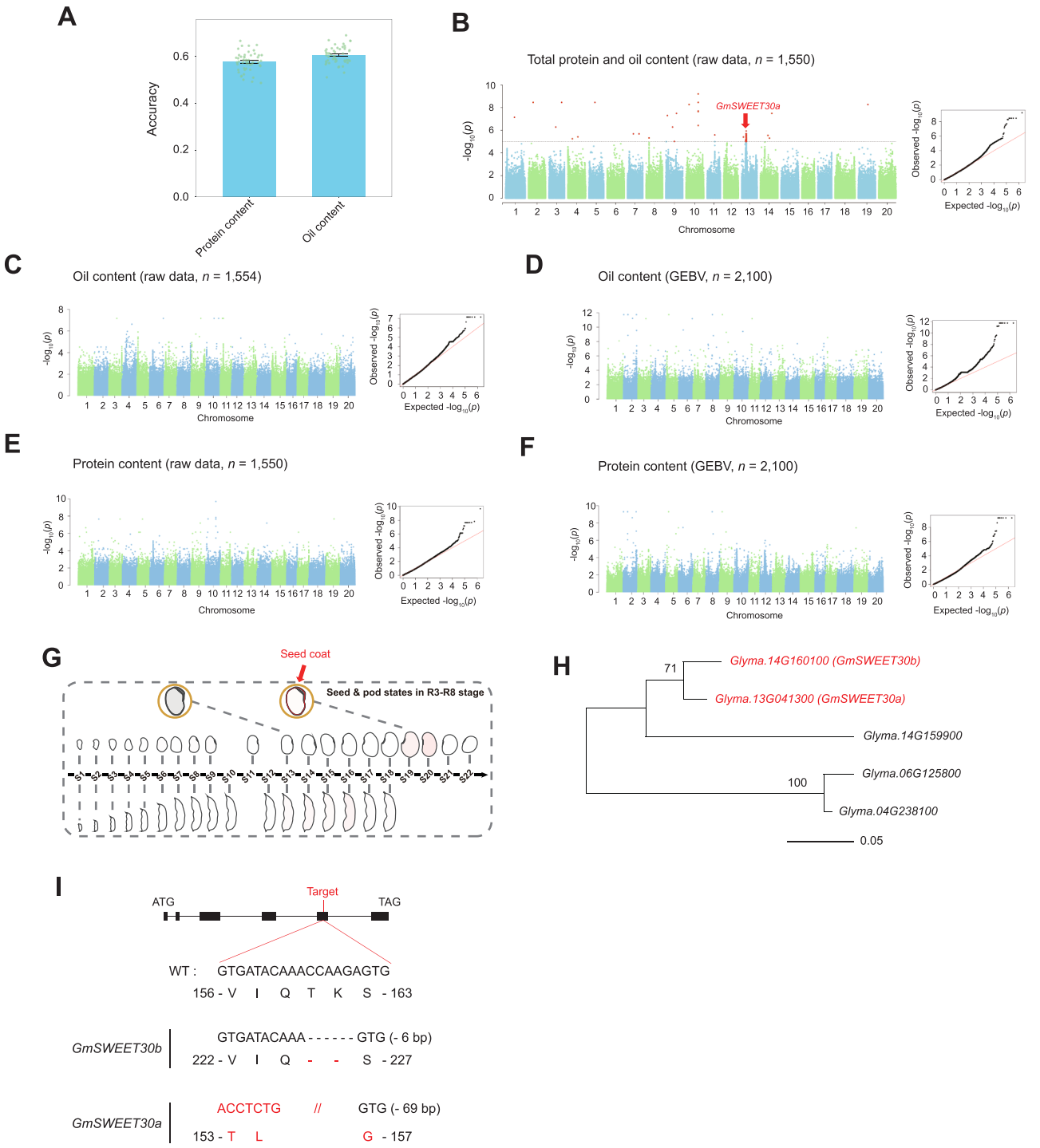


Figure S5. GWAS analyses of soybean oil and protein content, related to Figure 4

(A) Prediction accuracy of genomic estimated breeding values (GEBVs) for protein and oil content calculated by the genomic best linear unbiased prediction (GBLUP) method. Accuracy is assessed by 5-fold cross-validation with 10 replicates per trait. Data are presented as mean ± SD (n = 50). Accuracy is defined as the Pearson correlation coefficient (r) between GEBVs and observed phenotypic values.

(B) GWAS results for total protein and oil content (raw data) in 1,550 soybean samples.

(C) GWAS results for oil content (raw data) in 1,554 soybean samples.

(D) GWAS results for GEBVs (oil content) in 2,100 soybean samples.

(E) GWAS results for protein content (raw data) in 1,550 soybean samples.

(F) GWAS results for GEBVs (protein content) in 2,100 soybean samples.

(legend continued on next page)

(G) Expression pattern of the *GmSWEET30a* across soybean tissues. *GmSWEET30a* is highly expressed in the seed coat (transcripts per million = 298.684, colored in red) during the R3 to R8 developmental stages, with data sourced from the SoyOmics database.
(H) Maximum-likelihood phylogenetic tree of *GmSWEET30a* and four other highly homologous genes. The tree is constructed in molecular evolutionary genetics analysis, version X (MEGA-X), using amino acid sequences, with 1,000 bootstrap replicates. Scale bar, 0.05 amino acid substitutions per site.
(I) Schematic of the *GmSWEET30a* coding region (black boxes, exons; thin lines, introns) showing the CRISPR-Cas9 target site (red line). Below: DNA and predicted protein sequence alignments for the WT and two independent mutant alleles. The WT sequence spanning codons 156–163 (GTG ATA CAA ACC AAG AGT) encodes V-I-Q-T-K-S. For *GmSWEET30b*, a 6-bp deletion (dashes) removes "CCAAGA," resulting in loss of T and K (red dashes). For *GmSWEET30a*, a large fragment deletion causes a frameshift mutation. Amino acid positions are indicated at the ends of each sequence.

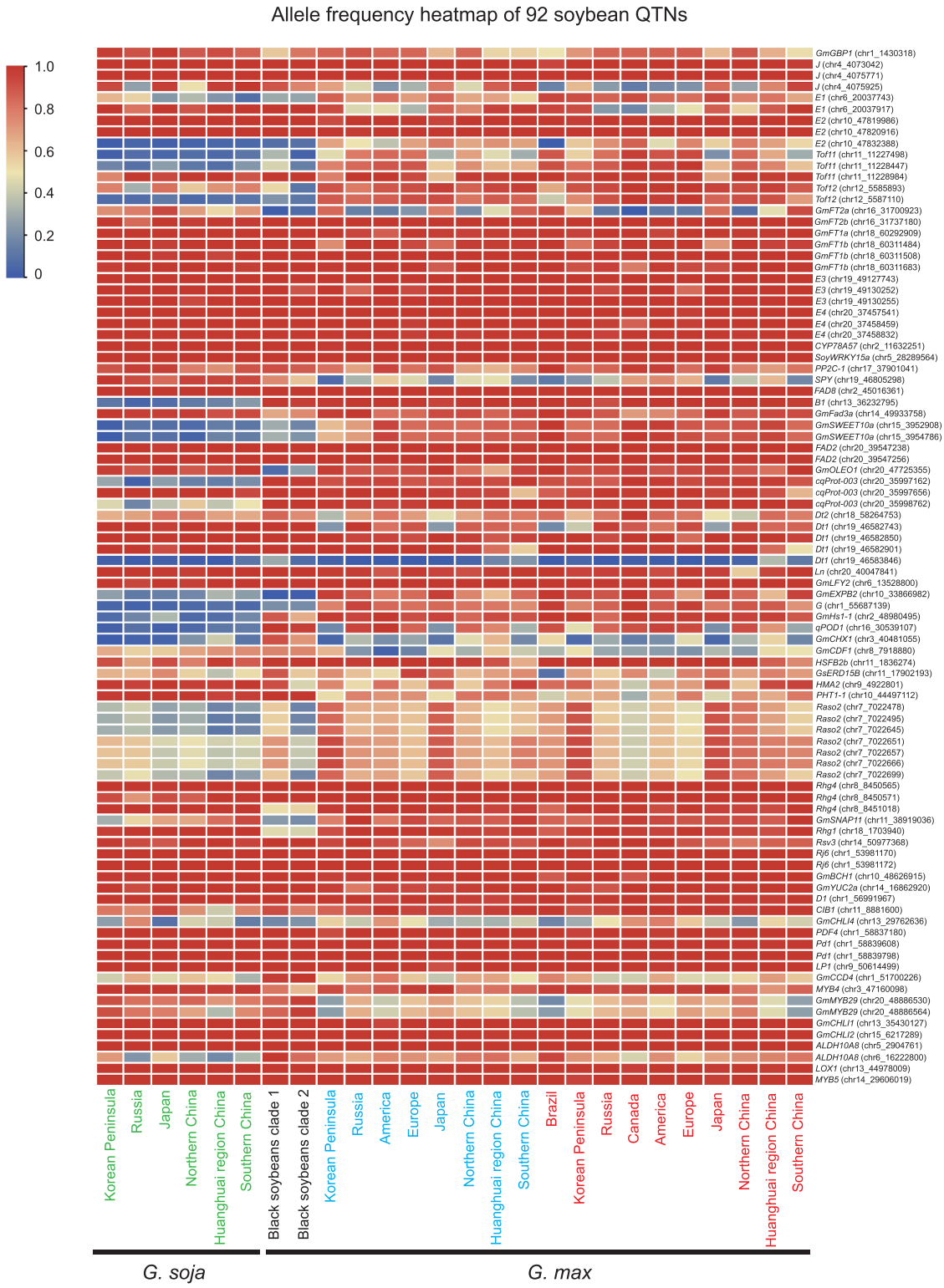


Figure S6. Soybean QTN library for selected genes, related to Figure 5 and Table S4

Allele frequencies of 92 functional variants in different soybean subpopulations: wild soybeans (green), black soybeans (black), landraces (blue), and cultivars (red). The legend on the left reflects the relationship between allele frequency and color. Frequencies are calculated based on the reference allele at each QTN site.