JIPB Journal of Integrative Plant Biology

Research Article
https://doi.org/10.1111/jipb.70392

Natural variation in *GmSOP5* regulates seed oil and protein content during soybean domestication[OO]

Yu Tian[1†], Hao Zhang[1†], Zihao Zheng[2], Delin Li[1], Jindong Li[1,3], Yingpu Che[1,3], Xing Li[4], Changjiang Li[4], Huawei Gao[1,3], Long Yan[5], Chaosen Zhao[6], Jiantao Guan[7], Jianan Han[1,7], Qi Wang[1], Shiyu Guo[1], Dong Bai[1], Jochen C. Reif[8], Ying-Hui Li[1,3*] and Li-Juan Qiu[1*]

1. State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
2. Department of Agronomy, Iowa State University, Ames IA 50011-1051, USA
3. National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya 572024, China
4. State Key Laboratory of Plant Environmental Resilience, College of Biological Sciences, China Agricultural University, Beijing 100193, China
5. Institute of Cereal and Oil Crops, Hebei Academy of Agricultural and Forestry Sciences, Hebei 050035, China
6. Jiangxi Province Key Laboratory of Oil Crops Biology, Crops Research Institute of Jiangxi, Nanchang 330200, China
7. Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
8. Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben 06466, Germany
†These authors contributed equally to this work.
*Correspondences: Li-Juan Qiu (qiulijuan@caas.cn, Dr. Qiu is fully responsible for the distribution of the materials associated with this article); Ying-Hui Li (liyinghui@caas.cn)

Yu Tian

Li-Juan Qiu

ABSTRACT

Seed oil content, protein content, and yield are agronomically important, correlated traits that determine the economic value of soybean (*Glycine max*). However, improving seed quality and yield simultaneously is challenging because gains in one breeding target often compromise the other, and the genetic basis of this trade-off is poorly understood. Here, we performed a genome-wide association study of 429 diverse soybean accessions and identified *Seed Oil and Protein 5* (*SOP5*), which encodes a kinesin protein, as a key locus associated with seed oil and protein content. Knockout and overexpression experiments demonstrated that *GmSOP5* positively affects seed oil content and 100-seed weight and negatively influences seed protein content. *GmSOP5* is located in a selective sweep region, and the domestication-related *GmSOP5*^{H1} allele is nearly fixed in cultivated soybean, contributing to increased seed size, weight, and oil content and reduced protein content. Field trials demonstrated that neither loss-of-function *GmSOP5*-edited mutants, which have increased seed protein content, nor *GmSOP5*-overexpression lines, which have increased seed oil content, differed significantly in yield from wild-type plants, because changes in plant architecture were offset by changes in seed weight. Our results shed light on soybean domestication and suggest how pleiotropy can be harnessed in breeding to enhance seed quality without compromising yield.

Keywords: domestication, GWAS, kinesin, seed oil and protein content, soybean

Tian, Y., Zhang, H., Zheng, Z., Li, D., Li, J., Che, Y., Li, X., Li, C., Gao, H., Yan, L., et al. (2026). Natural variation in *GmSOP5* regulates seed oil and protein content during soybean domestication. J. Integr. Plant Biol. **00:** 1–13.

INTRODUCTION

Soybean (*Glycine max* [L.] Merr.) is an economically important legume crop, accounting for approximately 56% of international oil seed production and providing over 25% of the protein consumed by humans and livestock (Lu et al., 2020; Wang et al., 2020). Cultivated soybean was domesticated from its annual wild progenitor *Glycine soja* (Sieb. & Zucc.) approximately 5,000 years ago in China (Li et al., 2023). During the domestication process, cultivated varieties diverged from wild populations across a suite of morphological and physiological traits, collectively termed the "domestication syndrome" (Hammer, 1984; Olsen and Wendel, 2013). Wild soybean generally produces smaller seeds with a high protein content, whereas cultivated soybean produces larger seeds with a high oil content (Patil et al., 2018; Wang et al., 2020). Together, the three interrelated traits of seed size, oil content, and protein content are the key factors that determine seed quality and yield in soybean (Goettel et al., 2022; Hu et al., 2023a). However, breeding for simultaneous improvement of seed quality and yield remains challenging because of the negative correlation between yield and protein content (Patil et al., 2017; Goettel et al., 2022).

Numerous genomic regions associated with seed quality have been identified through genome-wide association studies (GWASs) and quantitative trait locus (QTL) mapping (SoyBase database, https://www.soybase.org/). These studies have revealed that seed oil and protein content are complex quantitative traits governed by multiple genes. To date, a number of key regulators have been cloned and characterized, including *POWR1* (Goettel et al., 2022), *GmSWEET10a/10b* (Wang et al., 2020), *GmST05* (Duan et al., 2022)/*GmMFT* (Cai et al., 2023), *GmSop20* (Zheng et al., 2025), *FA9* (Qi et al., 2023), *PC08* (Liu et al., 2025), *SW14* (Zhang et al., 2025), and *GmSMS6* (Li et al., 2025). Nonetheless, the specific molecular mechanisms and individual genes involved in seed quality formation, particularly those that emerged or were fixed after domestication, remain to be clarified. Understanding how selection has acted upon these interrelated traits during soybean domestication and improvement and determining the genomic basis for trade-offs between seed quality and yield are therefore crucial for the continued improvement of soybean through molecular breeding.

Kinesins are microtubule-based motor proteins that drive essential cellular processes, including microtubule dynamics, morphogenesis, cell wall synthesis, chromosome segregation, and organelle transport (Ali and Yang, 2020). Although the functional diversity of kinesins has been well characterized in animals, much less is known about their roles in plants, and research has focused predominantly on model species such as Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*) (Reddy and Day, 2001; Richardson et al., 2006). For example, Arabidopsis *KAC1* and *KAC2* mediate chloroplast movement (Suetsugu et al., 2010), and rice *OsKCH2* mediates microtubule motility (Tseng et al., 2018). In soybean, *GmMs1* (an ortholog of *TES/NACK2*) has been shown to influence male fertility (Nadeem et al., 2021), but the potential involvement of kinesins in modulating seed quality components has not been explored previously.

In this study, we identified *Seed Oil and Protein 5* (*SOP5*) as a key locus associated with soybean seed quality through a GWAS and identified a mutation associated with significantly lower oil content and higher protein content, a stop-gain single-nucleotide polymorphism (SNP) in the coding region of the kinesin gene Glyma.05G085400 (*GmSOP5*). Haplotype analysis, association analysis, and gene knockout and overexpression experiments demonstrated that *GmSOP5* has pleiotropic effects on seed protein and oil content, seed size, 100-seed weight, and plant architecture. The $GmSOP5^{H1}$ allele, associated with higher seed oil content, greater 100-seed weight, and lower seed protein content, was under strong artificial selection during soybean domestication and has reached near-complete fixation in cultivated soybeans. A few accessions carry the rare high-protein $GmSOP5^{H2}$ allele, which was derived from $GmSOP5^{H1}$ after soybean domestication. Remarkably, *GmSOP5*-edited mutants with higher seed protein content and *GmSOP5*-overexpression lines with higher seed oil content showed no significant differences in yield compared with the corresponding wild-type control plants, as changes in plant architecture were offset by changes in seed weight. This study provides insight into early soybean domestication and identifies a genetic pathway that balances the trade-off between seed quality and yield in soybean and other seed crops.

RESULTS

GWAS reveals *SOP5* as a major locus for seed oil and protein content

To identify key genetic loci associated with seed oil and protein content, we performed a GWAS using 3,808,470 SNPs and 3 years of phenotypic data from 429 diverse soybean accessions (the association panel) (Figure S1; Table S1; Li et al., 2023). Using a mixed linear model, we identified 37 significant loci: 27 associated with oil content, six with protein content, and four with both traits (Figure 1A–D; Table S2). Among these, two previously characterized pleiotropic genes, *ST1* (Li et al., 2022) and *GmST05* (Duan et al., 2022)/*GmMFT* (Cai et al., 2023), were recovered. Notably, a locus on chromosome 5 (chr5:15,709,660; C/T) had the third-highest association with seed oil content and was also significantly associated with protein content (Figure 1A, C). This locus, which we designated *Seed Oil and Protein* 5 (*SOP5*), was located within several QTLs associated with seed oil content and protein content (Table S2), supporting its role in regulating these traits.

We performed linkage disequilibrium (LD) analysis of the 200-kb region centered on the lead SNP at the *SOP5* locus and identified three protein-coding genes (Glyma.05G085300, Glyma.05G085400, and Glyma.05G085500) based on the Williams 82 (Wm82) v2 reference genome (Figure 1E; Table S3). No variations were identified in the exonic regions of two of the genes (Glyma.05G085300 and Glyma.05G085500) within the association panel (Table S3). The significant SNP

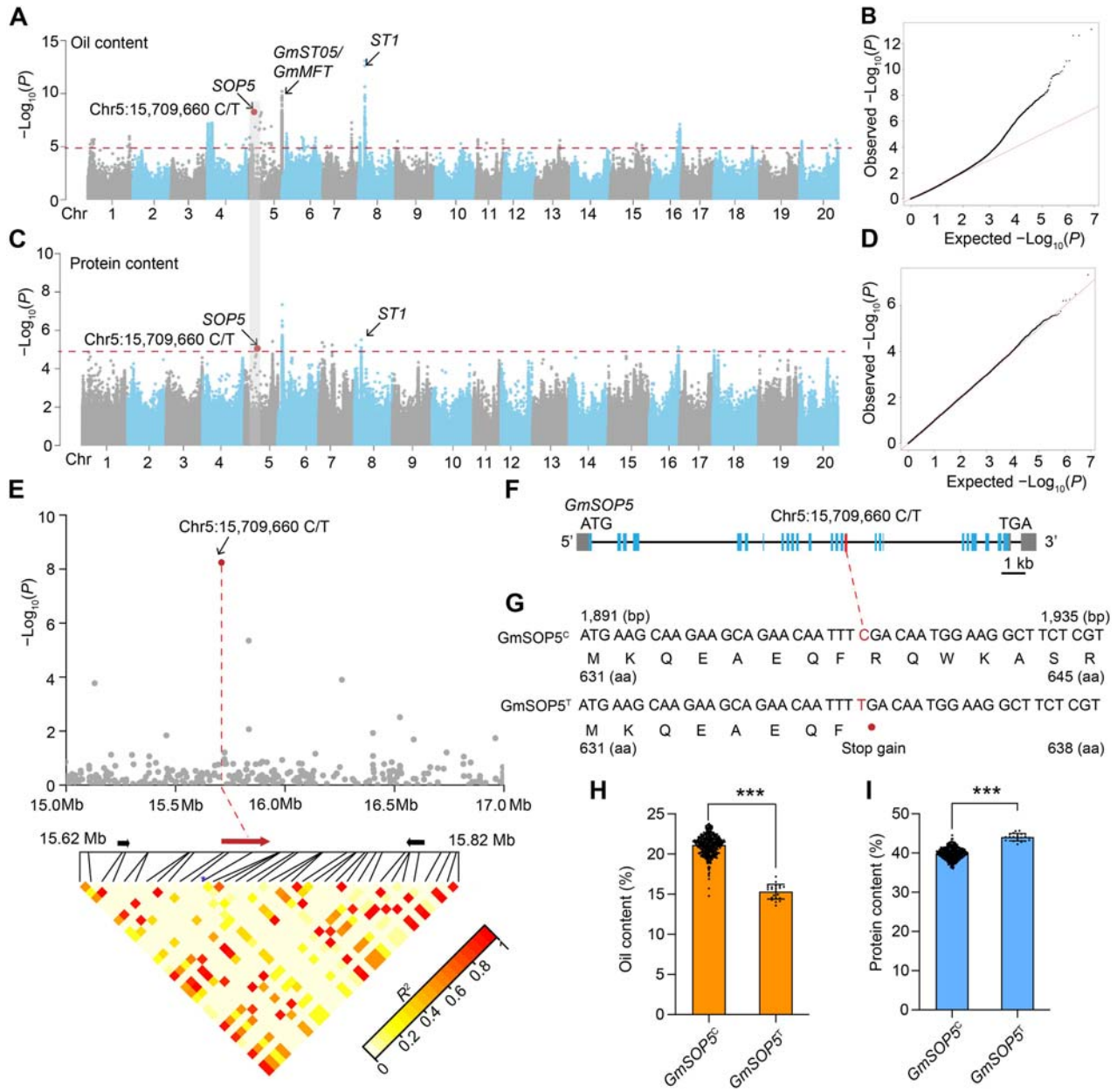

Figure 1. A stop-gain single-nucleotide polymorphism (SNP) mutation in *GmSOP5* is significantly associated with seed oil and protein content in soybean

(A, **C)** Manhattan plots of GWAS results for average seed oil content **(A)** and average seed protein content **(C)** in 429 diverse soybean accessions over 3 years (2018, 2019, and 2020) in Shijiazhuang. The red dot in each graph indicates the most significant SNP (chr5:15,709,660, C/T) at the *SOP5* locus. **(B**, **D)** Quantile–quantile plots for seed oil content **(B)** and protein content **(D)** GWAS based on a mixed linear model. **(E)** Local Manhattan plot for oil content in a 15–17 Mb region on chromosome 5 (top) and linkage disequilibrium plot for the 200 kb interval on chromosome 5 surrounding the peak SNP (chr5:15,709,660 C/T) (bottom). **(F)** Gene structure of *GmSOP5*. The blue and gray rectangles and black lines represent exons, untranslated regions, and introns, respectively. The red vertical line indicates the position of the most significant SNP (chr5:15,709,660 C/T). **(G)** Sequence comparison of $GmSOP5^C$ from Wm82 and $GmSOP5^T$ from XHD. The asterisk indicates the significant SNP mutation in Xiaoheidou ($GmSOP5^T$) that leads to the production of a truncated protein lacking 393 amino acids at the C terminus. **(H**, **I)** The effects of the $GmSOP5^C$ ($n = 405$) and $GmSOP5^T$ ($n = 22$) alleles on seed oil content **(H)** and seed protein content **(I)** in the association panel; data are means ± *SD*. *P* values were obtained using two-tailed Student's *t*-tests (***$P < 0.001$).

(chr5:15709660, C/T) was located within the 16th exon of the second gene, Glyma.05G085400, which encodes a kinesin protein consisting of 1,031 amino acids (Figure 1F). Compared with the $GmSOP5^C$ allele from Wm82, the mutant $GmSOP5^T$ allele contained a stop codon (TGA) instead of an arginine codon (CGA), leading to the loss of 393 amino acids at the C terminus, including part of the predicted coiled-coil domain and the entire tail region (Figures 1G, S2 and S3A). The mutation is located in a region that shows high conservation in both dicot and monocot species (Figure S3B) and clearly alters the predicted 3D protein structure of $GmSOP5^T$ compared with that of $GmSOP5^C$ (Figure S3C). Accessions carrying the $GmSOP5^C$

allele exhibited significantly higher oil content (4.4%) and significantly lower protein content (2.2%) than those with the *GmSOP5*^{T} allele (Figure 1H, I). These observations strongly suggest that Glyma.05G085400 is the causal gene underlying the association of *SOP5* with seed oil and protein content.

GmSOP5 influences seed oil and protein content in soybean

To further explore whether Glyma.05G085400 is the *GmSOP5* gene responsible for differences in seed oil and protein content, we used the CRISPR/Cas9 system to knock out Glyma.05G085400 in the Wm82 background, which harbors the *GmSOP5*^{C} allele. We obtained two independent mutant lines, *GmSOP5*-CR-1 (1-base pair deletion) and *GmSOP5*-CR-2 (5-base pair deletion), which produced truncated proteins of 113 and 119 amino acids, respectively (Figures 2A, S4). Compared with wild-type Wm82, both *GmSOP5*-edited lines had 2.3% lower oil content and 2.1% higher protein content (Figure 2C). In addition, the seed length, width, and thickness of the two *GmSOP5*-edited lines were all significantly smaller, resulting in an average reduction in 100-seed weight of 1.3 g relative to the Wm82 control (Figures 2A, C, S5A).

We also overexpressed the coding sequence of *GmSOP5*^{C} from Wm82 in the landrace Xiaoheidou (XHD), which carries *GmSOP5*^{T}, and generated two independent transgenic overexpression lines (*GmSOP5*^{C}-OE-1 and *GmSOP5*^{C}-OE-2) with increased transcript levels of *GmSOP5* (1.9- and 2.2-fold higher, respectively), as confirmed by reverse transcription quantitative polymerase chain reaction (RT-qPCR) (Figures 2B, S6). Relative to wild-type XHD, the two overexpression lines exhibited 7.1% higher seed oil content but 3.5% lower seed protein content (Figure 2D). The overexpression lines also produced larger seeds, with increased 100-seed weight, as well as increased seed length, width, and thickness (Figures 2D, S5B). These

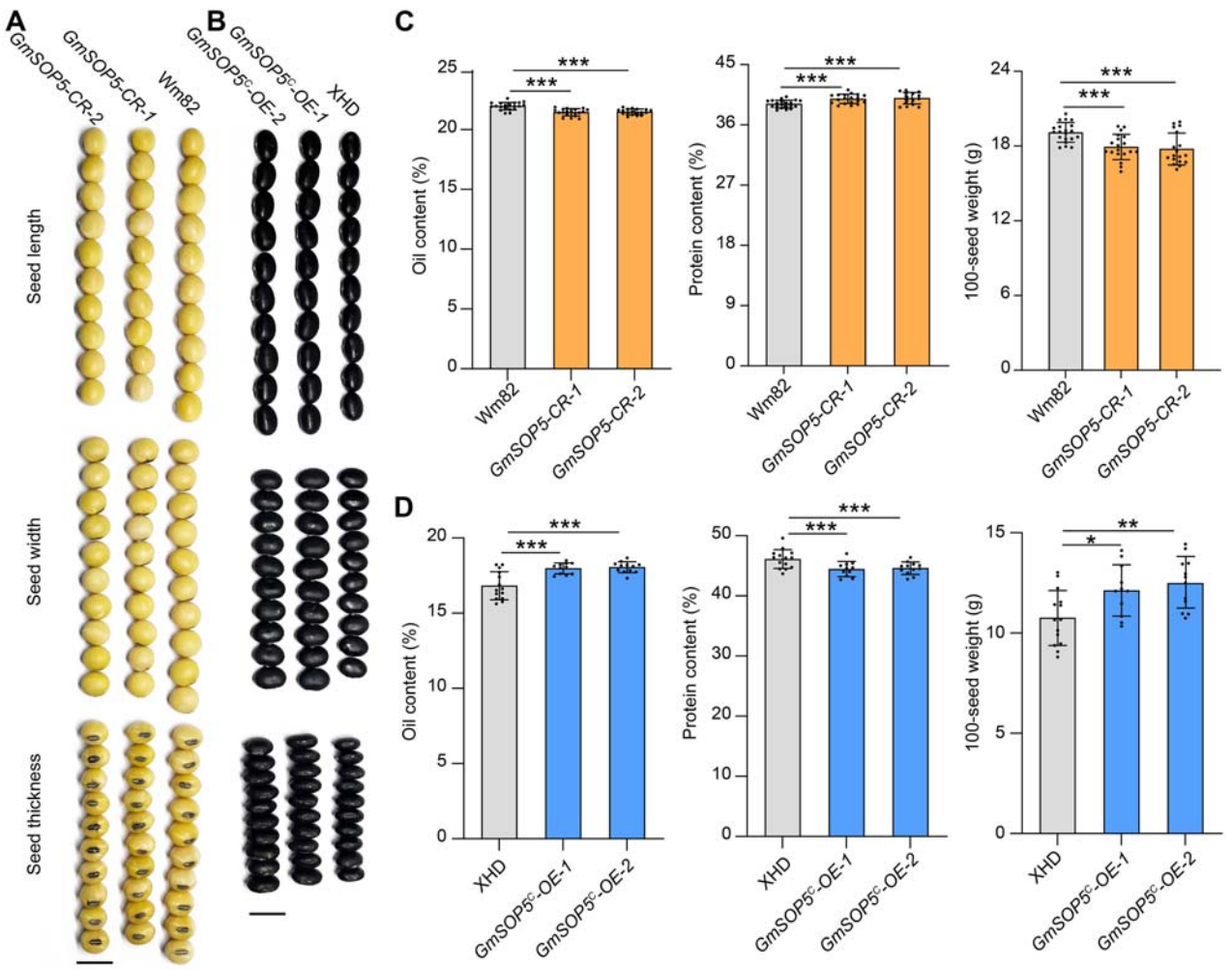

Figure 2. *GmSOP5* positively affects seed oil content and 100-seed weight but negatively influences seed protein content in soybean
(A, B) Seeds of wild-type Wm82 and two *GmSOP5*-edited lines (*GmSOP5*-CR-1 and *GmSOP5*-CR-2) **(A)**, and wild-type XHD and two T3 overexpression lines (*GmSOP5*^{C}-OE-1 and *GmSOP5*^{C}-OE-2) **(B)**. Scale bars, 1 cm. **(C, D)** Seed oil content, seed protein content, and 100-seed weight in mature seeds from wild-type Wm82, and two *SOP5*-edited lines (*GmSOP5*-CR-1 and *GmSOP5*-CR-2) ($n = 25$, 19, and 18, from left to right) **(C)** and wild-type Xiaoheidou (XHD and two T3 overexpression lines (*GmSOP5*^{C}-OE-1 and *GmSOP5*^{C}-OE-2) ($n = 14$, 11, and 12, from left to right) **(D)**. Data are means ± *SD*. *P* values were obtained using two-tailed Student's *t*-tests (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

results confirm that Glyma.05G085400 corresponds to *SOP5*, which promotes increased seed oil content and 100-seed weight while reducing seed protein content in soybean.

GmSOP5 affects the expression of genes related to sugar metabolism and transport

GmSOP5 encodes a kinesin protein homologous to Arabidopsis *AtKINESIN-4A/FRA1* (Figure S7; Zhong et al., 2002; Kong et al., 2015; Augustine, 2020; Ganguly et al., 2020) and is expressed in all tissues examined (i.e., leaf, stem, root, floral bud, and developing seeds) (Figure 3A). Most members of the kinesin-4 subfamily contain putative nuclear localization signals (NLSs) and have nuclear functions in animals, such as chromosome-arm orientation and oscillation, mitosis, and meiosis (Wang and Adler, 1995; Levesque and Compton, 2001; Sheng et al., 2018). However, the subcellular localization and functions of kinesin-4 proteins in plants remain largely unknown. Using the LOCALIZER program (Sperschneider et al., 2017), we established that the full-length GmSOP5 protein ($GmSOP5^C$) contains four putative NLSs (NLS1–NLS4) (Figure S8). Whereas $GmSOP5^C$ was predominantly localized in the nucleus, we observed significantly lower levels of nucleus-localized $GmSOP5^T$ (Figures 3B, S9). We also detected unknown punctate structures in cells expressing $GmSOP5^T$ that were absent in cells expressing $GmSOP5^C$ (Figure 3B). These results suggest that the stop-gain SNP in $GmSOP5^T$ alters or impairs SOP5 function by removing a large region of the protein, thereby changing its subcellular localization.

To further characterize the mechanisms by which *GmSOP5* influences oil content, protein content, and 100-seed weight, we performed RNA sequencing of seed coats and embryos from the Wm82 control and the two *GmSOP5*-edited lines (CR-1 and CR-2) (Table S4). We identified 1,963 differentially expressed genes (DEGs) between the two *GmSOP5*-edited lines and wild-type Wm82: 836 upregulated and 1,127 downregulated (Figure 3C; Table S5). Among these DEGs, 270 were identified only in the embryo, 1,620 only in the seed coat, and 73 in both the embryo and the seed coat (Table S5).

To examine tissue-specific differences associated with *GmSOP5* knockout, we performed independent Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) enrichment analyses of the seed coat

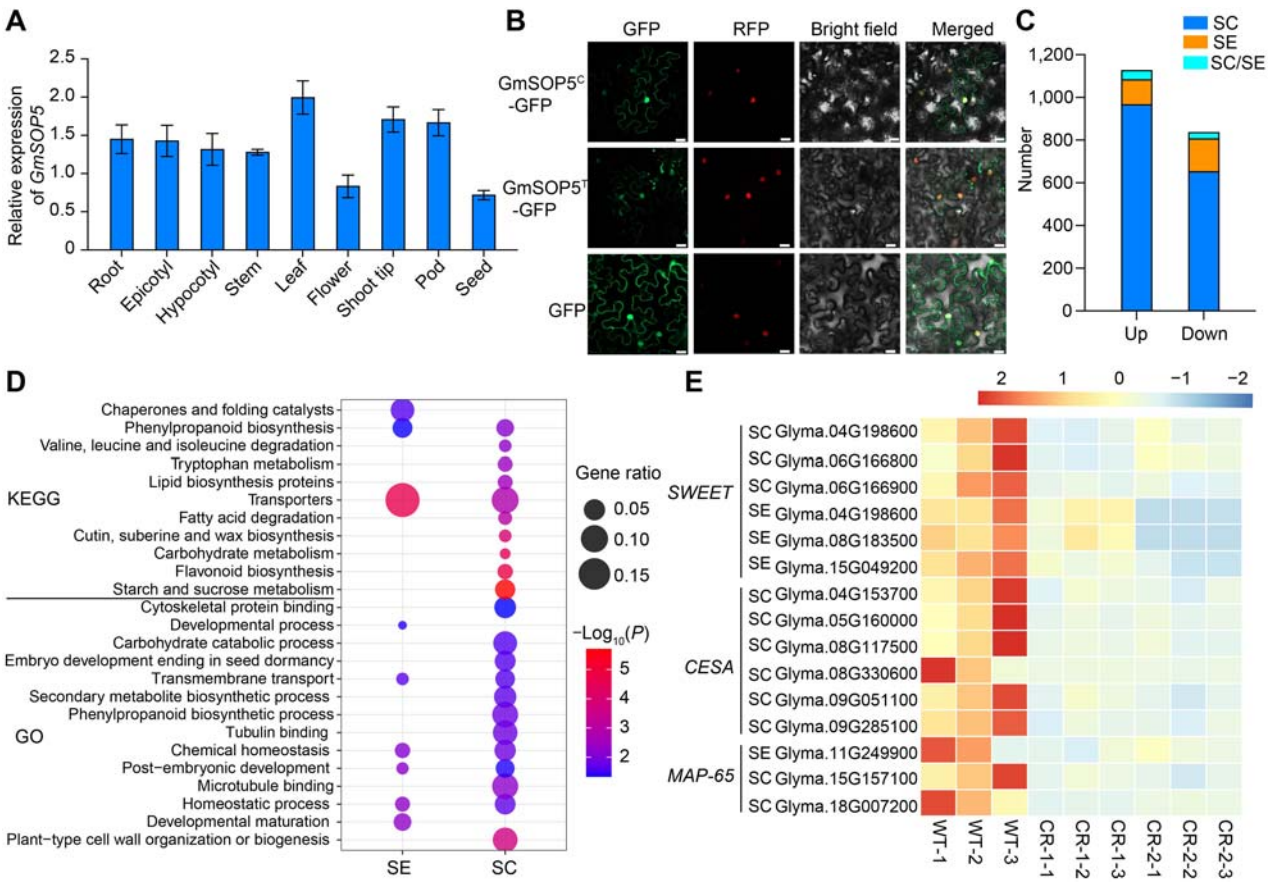

Figure 3. *GmSOP5* affects the expression of genes related to sugar metabolism and transport

(A) *GmSOP5* expression in major tissues of soybean Wm82 as measured by RT-qPCR. *GmActin11* was used as the reference control to calculate relative expression; data are means ± *SD* ($n = 3$). **(B)** Subcellular localization of full-length $GmSOP5^C$-GFP and $GmSOP5^T$-GFP fusion proteins in epidermal cells of *Nicotiana benthamiana*. AHL22-RFP was used as the nuclear marker. Scale bars, 20 μm. **(C)** DEGs between Wm82 and two *GmSOP5*-edited lines (CR1 and CR2) in the seed coat (SC) and seed embryo (SE). **(D)** KEGG pathways and GO terms enriched in the DEGs. **(E)** Heatmaps highlighting key downregulated genes in SC and SE. WT: Wm82.

and embryo DEGs (Figure 3D). The seed coat DEGs were significantly enriched in KEGG pathways and GO terms related to starch and sucrose metabolism, carbohydrate metabolism, fatty acid degradation, cell wall organization (biological process), and microtubule binding (molecular function). The embryo DEGs were significantly enriched in pathways and terms related to chaperones and folding catalysts, anatomical structure maturation (biological process), and developmental maturation (biological process). In addition, pathways and terms related to transporters, phenylpropanoid biosynthesis, postembryonic development (biological process), and homeostatic processes (biological process) were enriched in DEGs shared by the seed coat and embryo. These observations suggest that *GmSOP5* influences seed size and composition through tissue-specific effects on processes such as starch and sucrose metabolism, carbohydrate metabolism, cell wall organization, and transporter activity.

A number of key genes involved in sugar transport (*SWEET12*/Glyma.06G166800, *SWEET13*/Glyma.04G198600/Glyma.06G166900, *SWEET10a*/Glyma.15G049200, and *SWEET10b*/Glyma.08G183500), cellulose synthesis (*CSLE1*/Glyma.08G330600, *CESA4*/Glyma.09G051100/Glyma.15G157100, *CESA7*/Glyma.04G153700, and *CESA8*/Glyma.08G117500/Glyma.05G160000), and microtubule networks (*MAP65-6*/Glyma.09G285100 and *MAP65-8*/Glyma.11G249900/Glyma.18G007200) were significantly downregulated in the embryo and/or seed coat of *GmSOP5*-edited lines relative to Wm82 (Figure 3E). We specifically analyzed the expression of genes involved in oil and protein metabolism (Figure S10), such as those associated with lipid biosynthesis (*DGAT1A*/Glyma.13G106100, *AAE*/Glyma.02G239500, and *GPAT9*/Glyma.05G131100) and fatty acid desaturation (*FAD2-2*/Glyma.10G278000) and those encoding oleosin protein (*OLEO1*/Glyma.20G196600), acyl-ACP thioesterase (*FATA1B*/Glyma.08G349200), and key transcription factors (*ABI5*/Glyma.13G317000, *ABI3b*/Glyma.08G357600, and *WRI1a*/Glyma.15G221600). Among these, only *FAD2-2* (Glyma.10G278000) was significantly downregulated in the embryos of the edited lines relative to Wm82 (Figure S10).

Two genes previously reported to influence seed oil and protein content, *GmSWEET10a*/Glyma.15G049200 and *GmSWEET10b*/Glyma.08G183500 (Wang et al., 2020), showed significantly lower expression in the *GmSOP5*-edited lines than in the Wm82 control (Figure 3E). Previous studies have shown that the expression of these genes is positively associated with seed oil content and seed weight but negatively associated with seed protein content (Wang et al., 2020; Cai et al., 2023). RT-qPCR confirmed that both genes were downregulated in *GmSOP5*-CR-1 and *GmSOP5*-CR-2 and upregulated in $GmSOP5^{C}$-OE-1 and $GmSOP5^{C}$-OE-2 relative to the Wm82 and XHD controls, respectively (Figure S11). These results suggest that *GmSOP5* may influence the expression of genes related to sugar metabolism, sugar transport, and the cytoskeleton, thereby exerting pleiotropic effects on seed oil and protein content as well as on 100-seed weight.

Evolution and selection of *GmSOP5* during the domestication and subsequent geographic spread of soybean

To determine whether *GmSOP5* was under selection during soybean domestication, we compared nucleotide diversity (π, F_{st}, and Tajima's *D*) across the 2-Mb genomic region of *GmSOP5* using previously published high-quality genome sequences of 2,214 *G. soja* and *G. max* accessions, each with well-documented genotype and origin information (Li et al., 2023). The results of π, F_{st}, and Tajima's *D* calculations all indicate the presence of a selective sweep at the *SOP5* locus (Figure 4A–C).

Across the re-sequencing panel, we identified three nonsynonymous variants in the coding region of *GmSOP5*: Two leading to amino acid substitutions and one to a premature stop codon. Based on these variants, *GmSOP5* can be classified into four major haplotypes (H1–H4) (Figure 4D, E; Table S6). Three of the four major haplotypes, $GmSOP5^{H1}$, $GmSOP5^{H3}$, and $GmSOP5^{H4}$, were present in *G. soja* and accounted for 18.3%, 69.5%, and 12.2% of all *G. soja* accessions, respectively (Figure 4F), suggesting that the $GmSOP5^{H3}$ haplotype is the predominant haplotype in *G. soja*. Geographical information for these *G. soja* accessions revealed that those carrying the $GmSOP5^{H3}$ haplotype were spread across East Asia, whereas those with the $GmSOP5^{H1}$ and $GmSOP5^{H4}$ haplotypes occurred mainly in the Huang–Huai region of China (Figure 4G). Notably, the frequency of the $GmSOP5^{H1}$ allele increased from *G. soja* (18.3%) to landraces (97.5%) and improved cultivars (98.9%) (Figure 4F).

We next investigated the associations between the three *GmSOP5* haplotypes present in *G. soja* (H1, H3, and H4) and the seed-related traits of the *G. soja* accessions. The 100-seed weight and oil content were significantly higher in *G. soja* accessions with the $GmSOP5^{H1}$ haplotype than in those with other haplotypes, whereas protein content was significantly lower in *G. soja* accessions with $GmSOP5^{H1}$ (Figure 4H). This result was consistent with the selection signal identified around the *SOP5* locus and suggests that the $GmSOP5^{H1}$ allele was under strong artificial selection during domestication, likely due to its association with larger seeds and/or higher oil content. This allele then spread both north and south and became almost completely fixed during the early domestication and diversification of soybean (Figure 4I).

Intriguingly, we identified a rare haplotype, $GmSOP5^{H2}$ (carrying the $GmSOP5^{T}$ stop-gain SNP) in 27 *G. max* accessions (1.3%), including 22 landraces and five improved cultivars (Table S6). These landraces were mainly collected from northern China and displayed desirable agronomic traits such as increased protein content and more branches and pods per plant without a significant reduction in seed weight per plant, which may explain their local retention (Figures 4I, S12). A median-joining network clearly showed that both the $GmSOP5^{H1}$ and $GmSOP5^{H4}$ haplotypes arose from $GmSOP5^{H3}$ through two different SNP mutations. Consistent with this scenario, two *Glycine tomentella* accessions and most of the *G. soja* accessions carried $GmSOP5^{H3}$, suggesting that $GmSOP5^{H3}$ was

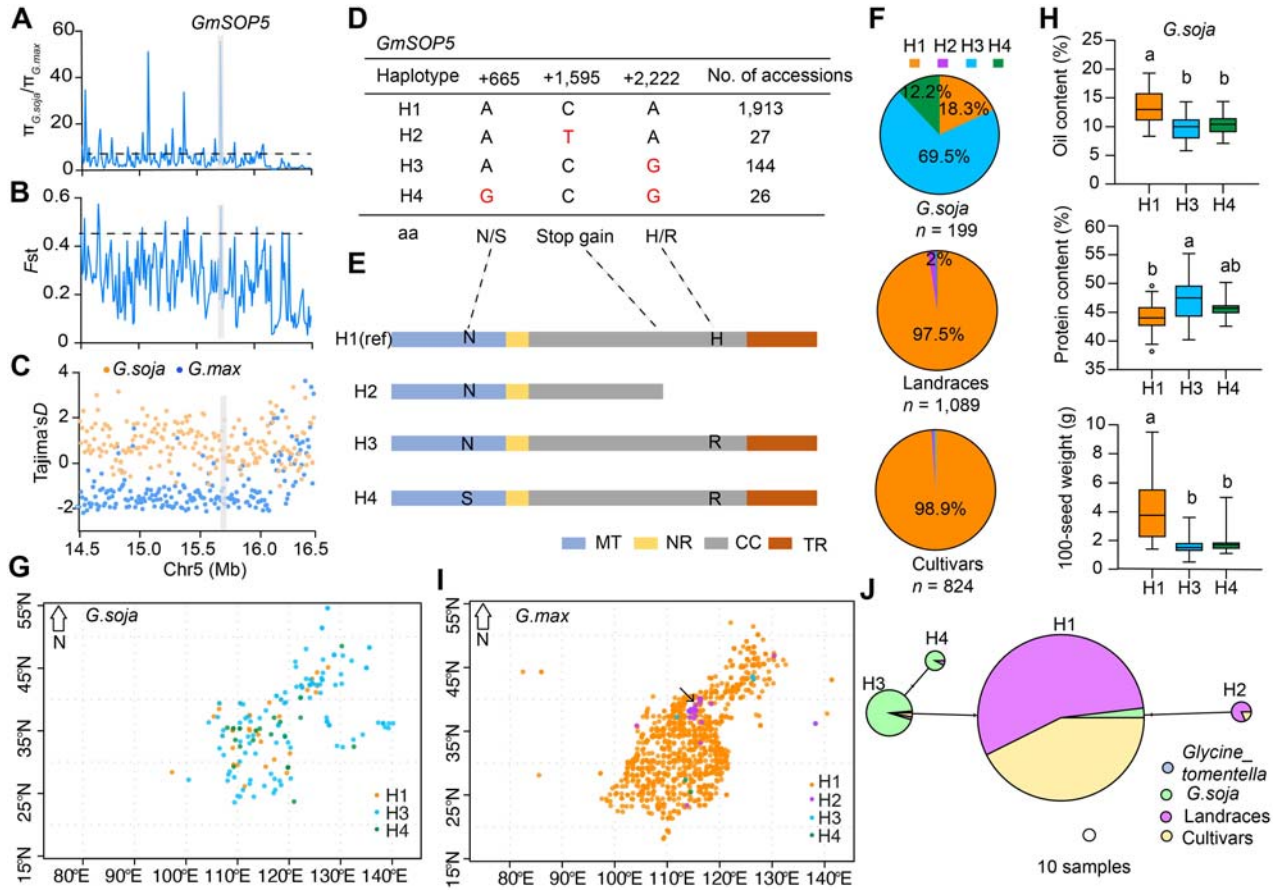

Figure 4. Evolution and selection of *GmSOP5* during soybean domestication and geographic spread

(A–C) Analysis of genetic diversity across the 2-Mb region of *GmSOP5*. Values of π **(A)**, F_{st} **(B)**, and Tajima's *D* **(C)** between *Glycine soja* and *G. max* across the 2-Mb region of *GmSOP5* are shown. The dashed horizontal lines in **(A)** and **(B)** represent the genome-wide threshold for selection signals, defined as the top 10% of scores across the entire genome. The gray rectangle in each graph highlights the position of *GmSOP5*. **(D)** Haplotype analysis of *GmSOP5*. The large-effect SNPs (those that cause a change in amino acid residues) shown here were derived from a previously reported diversity panel (Li et al., 2023). The ATG start codon is defined as position +1 bp. Red letters indicate differences from H1. **(E)** Structures of the proteins encoded by four *GmSOP5* haplotypes. CC, coiled-coil domains; MT, motor domain; NR, neck region; TR, tail region. The amino acid abbreviations before and after the slashes are the reference and alternative amino acids, respectively. **(F)** Frequency distribution of the four haplotypes in *G. soja* accessions, landraces, and cultivars. **(G)** Geographic distribution of different *GmSOP5* haplotypes in 189 *G. soja* accessions collected from East Asia, the region of origin of soybean. **(H)** Effects of three *GmSOP5* haplotypes (H1, H3, and H4) on seed oil content, seed protein content, and 100-seed weight in *G. soja* accessions ($n = 32$, 79, and 13, from left to right); data are means $\pm$ *SD*. Box edges represent the interquartile range, and the black line within the box indicates the median. Bars with the same lowercase letter are not significantly different. **(I)** Geographic distribution of different *GmSOP5* haplotypes in 1,680 *G. max* accessions collected from East Asia, the region of origin for soybean. **(J)** Median-joining network showing the relatedness among the four haplotypes. Each pie represents a haplotype; the size of the pie indicates the number of accessions with that haplotype, and its colored sectors indicate the different types of accessions.

the ancestral haplotype (Figure 4J). The novel *GmSOP5*^{H2} allele was derived from *GmSOP5*^{H1} by a stop-gain SNP mutation after soybean domestication (Figure 4J).

GmSOP5 can balance seed number per plant and 100-seed weight, thus maintaining yield in soybean

Increasing the protein content of soybean seeds is challenging because protein content tends to be negatively correlated with yield (Florencia et al., 2016; Kim et al., 2016; Guo et al., 2022; Liu et al., 2023). To examine the yield performance of *GmSOP5* gene-edited and overexpression lines, we performed field tests at three sites with widely different latitudes (Sanya, Shijiazhuang, and Beijing). The two *GmSOP5*-edited lines (*GmSOP5*-CR-1 and *GmSOP5*-CR-2) produced significantly more branches (~1.1 to 4.7 more), pods (~4.0 to 17.9 more), and seeds (~8.2 to 52.7 more) per plant but had a lower 100-seed weight (~0.5 to 2.0 g less) than the Wm82 control, with no significant differences in plant height (Figures 5A–E, S13A). There were also no significant differences in seed weight per plant or plot yield, as the increased seed number per plant in the edited lines offset their reduced 100-seed weight (Figure 5F, G).

By contrast, the two *GmSOP5*-overexpression lines (*GmSOP5*-OE-1 and *GmSOP5*-OE-2) showed significant increases in seed oil content and 100-seed weight compared with the XHD control but had ~2.7 to 2.8 fewer branches and ~41.7 to 56.6 fewer pods per plant, resulting in ~57.5 to 82.1 fewer seeds per plant (Figure S14). Again, there were no significant

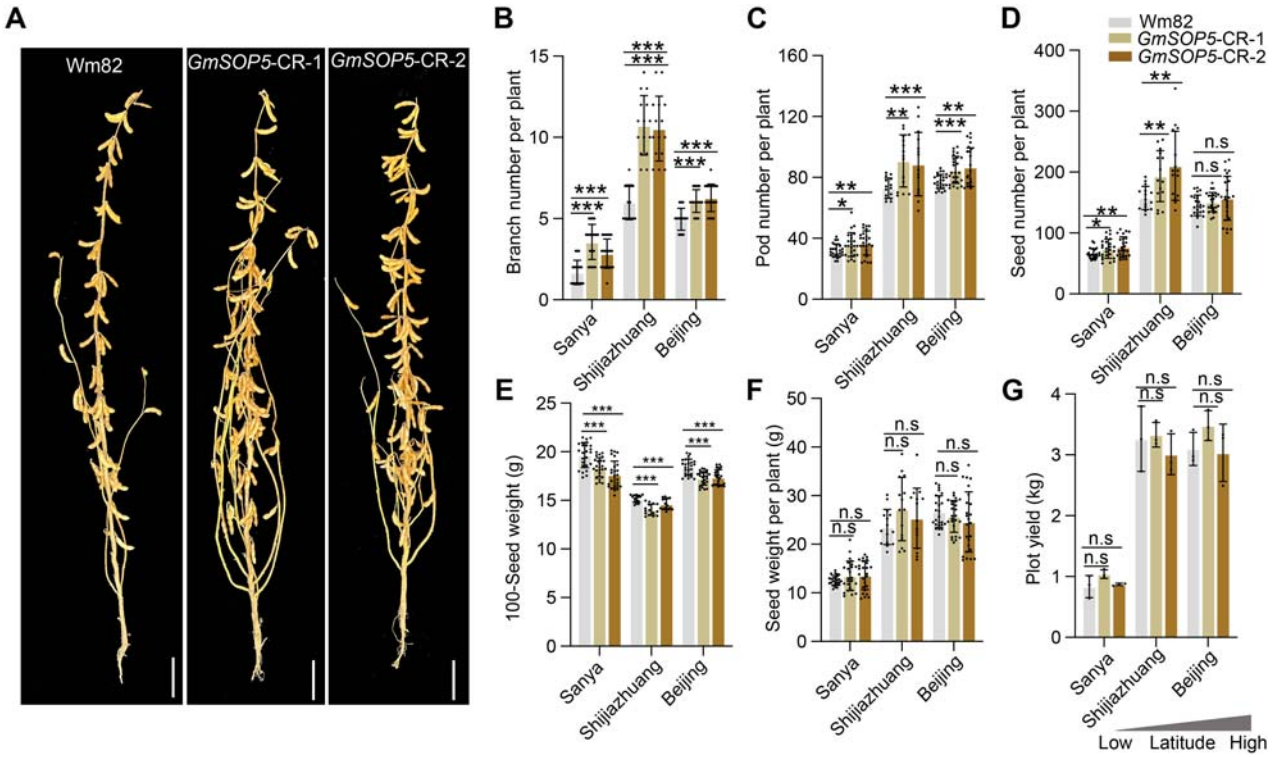

Figure 5. *GmSOP5* gene-edited lines have a reduced 100-seed weight but an increased seed number per plant, thus maintaining yield under field conditions

(A) A Wm82 control plant and two *GmSOP5*-edited lines (*GmSOP5*-CR-1 and *GmSOP5*-CR-2). Scale bars, 10 cm. **(B–G)** Phenotypic differences between the Wm82 control, *GmSOP5*-CR-1, and *GmSOP5*-CR-2 grown in three locations (Sanya in 2022, Shijiazhuang in 2023, and Beijing in 2023): branch number per plant **(B)**, pod number per plant **(C)**, seed number per plant **(D)**, 100-seed weight **(E)**, seed weight per plant **(F)**, and plot yield **(G)**. Sanya (Wm82, *GmSOP5*-CR-1, *GmSOP5*-CR-2; $n = 25$ each), Shijiazhuang ($n = 15$ each), Beijing ($n = 25$ each). The plot sizes were 2 m^2 for Sanya and 6 m^2 for Shijiazhuang and Beijing; each genotype had three replicate plots per site. Data are means ± *SD*. *P* values were obtained using two-tailed Student's *t*-tests (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

differences in plant height or seed weight per plant between XHD and the overexpression lines (Figures S13B, S14). Together, these results demonstrate that *GmSOP5* has pleiotropic effects on seed traits and plant architecture, suggesting that it could serve as a valuable target for enhancing seed quality while maintaining yield.

DISCUSSION

Kinesins are microtubule-based motor proteins that participate in diverse cellular processes, including microtubule dynamics and morphogenesis, cell wall synthesis, and organelle transport (Ali and Yang, 2020). Although the kinesin superfamily is well characterized in animals, less is known about the roles of kinesins in plants, and most studies have focused on Arabidopsis and rice (Reddy and Day, 2001; Richardson et al., 2006). Although *GmMs1* was recently identified as a regulator of male fertility in soybean (Nadeem et al., 2021), the broader physiological roles of kinesins have not been characterized in this major crop. Our study reveals that *GmSOP5*, an ortholog of *AtFRA1* (Zhong et al., 2002) and *OsGDD1/BC12* (Zhang et al., 2010; Li et al., 2011), encodes a kinesin-4 superfamily protein with broad pleiotropic effects. Expression of *GmSOP5* influences multiple seed and plant architectural traits, helping to balance the trade-off between oil/protein content and yield in soybean. These results expand our understanding of kinesin-4 function and underscore its potential for use in multi-trait crop improvement.

Comparative sequence analysis revealed that although the motor and coiled-coil domains of GmSOP5 are highly conserved, its tail region exhibits substantial divergence (Table S7), a hallmark of cargo specificity in kinesin proteins (Zhu and Dixit, 2012). We identified four putative NLSs in the full-length $GmSOP5^{H1}$ protein. By contrast, the truncated $GmSOP5^{H2}$ variant lacks two C-terminal NLSs owing to a premature stop codon, a loss that alters its subcellular localization and likely affects its function. Our RNA-seq analysis revealed that *GmSOP5* predominantly influences pathways related to sugar metabolism, transport, and cytoskeletal organization but revealed few direct changes in the expression of classical oil or protein biosynthesis genes. We interpret this observation as follows: *GmSOP5* encodes a kinesin motor protein, which is not expected to directly regulate the enzymatic steps of fatty acid or storage protein synthesis.

Instead, its effect on seed composition likely occurs through upstream control of carbon partitioning and nutrient transport, which in turn alters the balance of oil and protein allocation. Notably, knockout or overexpression of *GmSOP5* altered the expression of *GmSWEET10a* and *GmSWEET10b*, two established regulators of soybean seed oil and protein content (Figure S11). However, yeast one-hybrid assays demonstrated that GmSOP5 does not directly bind to the promoters of *GmSWEET10a* or *GmSWEET10b*, thus ruling out direct transcriptional regulation (Figure S15). We hypothesize that GmSOP5 influences *GmSWEET10a/10b* expression indirectly, possibly through effects on transport, cellular sugar status, microtubule-mediated trafficking of transcription factors, or nuclear positioning. The specific downstream pathways that connect *GmSOP5* to changes in seed composition and plant architecture remain an intriguing question for future mechanistic studies.

The negative correlation between seed protein content and yield remains a major bottleneck in soybean breeding (Patil et al., 2017; Guo et al., 2022; Hu et al., 2023b). Traditional high-protein alleles, such as *POWR1* (Goettel et al., 2022) and *GmST05* (Duan et al., 2022), often carry a "yield drag". In the current study, *GmSOP5* balanced seed quality and yield. In *GmSOP5*-edited lines, the reduced 100-seed weight was offset by increased branch and pod numbers, maintaining seed weight per plant (Figures 5, S16), whereas in overexpression lines, the increased seed size was offset by reduced branch and pod numbers. Although *GmSOP5* transcript levels were only moderately increased (1.9- and 2.2-fold) in the two overexpression lines, their consistent and substantial phenotypic changes—including higher oil content, lower protein content, and larger seeds—support the biological relevance of this moderate upregulation. In addition, the introduced full-length $GmSOP5^{C}$ allele likely complemented the defective endogenous $GmSOP5^{T}$ allele in the landrace XHD. The observed trade-off between seed number and seed size could be consistent with altered sugar transport and resource allocation: smaller seeds allow for a greater seed set, whereas larger seeds necessitate reduced sink strength elsewhere. These observations suggest a potential strategy for improving oil or protein content while maintaining yield stability in soybean breeding programs.

We also identified a rare high-protein haplotype, $GmSOP5^{H2}$, which was present in only a few landraces and improved cultivars (Figure 4I; Table S6). This allele provides a valuable genetic resource for molecular marker-assisted selection and biotechnological intervention, offering a rare opportunity to enhance soybean protein levels without compromising the productivity required for global food security.

Domesticated crops are characterized by a set of traits referred to as the domestication syndrome (Doebley et al., 2006). In soybean, these traits include increased seed size, fewer branches, loss of seed dormancy and dispersal ability, and changes in photoperiod sensitivity, flowering, and maturation time (Liu et al., 2007; Lu et al., 2020, 2022). Although genes such as *SHAT1-5* (pod shattering) (Dong et al., 2014) and *GmHs1-1* (seed hardness) (Sun et al., 2015) have been identified in soybean, few genes are known to have contributed simultaneously to multiple facets of the domestication syndrome during early selection. We demonstrated that *GmSOP5* is located within a selective sweep (Figure 4A–C) and that the $GmSOP5^{H1}$ haplotype underwent intense artificial selection during early domestication, reaching near fixation in cultivated soybean. This selection likely facilitated the simultaneous transition toward increased seed size, higher oil content, and fewer branches (Figure S12). These results provide critical insights into early soybean evolution, illustrating how selection on a single pleiotropic gene can contribute to multiple key domestication traits.

MATERIALS AND METHODS

Plant materials, growth conditions, and phenotyping

The 429 soybean accessions (the association panel) used for the GWAS were planted and grown under natural conditions during the summer season (June to October) at the experimental station of the Hebei Academy of Agricultural and Forestry Sciences, Shijiazhuang (37.8° N, 114.5° E) in 2018, 2019, and 2020. Each accession was planted in a 1.8 × 0.8-m plot with two rows and a spacing of 10 cm between plants in each row. The CRISPR/Cas9 knockout lines, *GmSOP5*-overexpression transgenic lines, and WT (Wm82 and XHD) were planted in a randomized complete block design with three replications per environment. The length of each row was 3 m, with 0.10 m between plants in each row and 0.50 m between rows. The plot area for yield measurement was 2 m^2 in Sanya (18.2° N, 109.5° E) and 6 m^2 in Shijiazhuang and Beijing (40.1° N, 116.7° E). The knockout plants used for RNA sequencing and expression analysis were grown under long-day conditions (16 h light/8 h dark) in a plant growth chamber (RDN-1000C, Ningbo Yanghui Instrument Co.) with a light intensity of 500 lux.

Oil and protein contents of uniform dry seeds collected from the association panel, the CRISPR/Cas9 knockout lines, the overexpression lines, and the wild-type controls were measured using a multifunctional near-infrared system (MPA; Bruker Optics, Ettlingen, Germany). Each sample was measured three times to generate three sets of spectral data, and the average value was used as the final phenotype. For 100-seed weight, at least three replicates were examined. Seed length, width, and thickness were measured using three replicates of at least 10 representative dry seeds each from the CRISPR/Cas9 knockout mutants, transgenic plants, and their wild-type controls, and the average value of the three replicates was used as the final phenotype.

GWAS

We identified 3,808,470 high-quality SNPs (minor allele frequency [MAF] > 0.05, missing rate < 0.2, and heterozygosity < 0.1) relative to the Wm82.a2.v1 soybean reference

genome (Schmutz et al., 2010) (https://phytozome-next.jgi.doe.gov) in the association panel. These accessions were a subset of 2,214 previously re-sequenced soybean accessions (Li et al., 2023) and were used to perform GWAS for average seed oil content and average seed protein content across 3 years (2018, 2019, and 2020). The population structure (Q) was evaluated using Admixture software (Alexander et al., 2009). Kinship matrices were calculated using the "centered_IBS" method in Tassel v5 (Bradbury et al., 2007). GWAS was performed using a mixed linear model implemented in Tassel v5. Significantly associated SNPs were defined using a threshold of 1 divided by the effective SNP number, which was determined in PLINK v1.90 with the settings—indep-pairwise 100 kb 50 0.2, resulting in a cutoff of $P = 1.15E{-}05$.

Population genetics analysis

We analyzed genetic diversity using 4,467,028 high-quality SNPs (MAF > 0.05, missing rate < 0.2, and heterozygosity < 0.1) from our previously re-sequenced panel of 2,214 wild soybean accessions, landraces, and improved cultivars. To investigate pairwise genomic differentiation between *G. soja* and *G. max*, we calculated F_{st} and π values in 10-kb sliding windows using VCFtools (0.1.16) (Danecek et al., 2011). The diversity ratio ($\pi_{G.soja}/\pi_{G.max}$) was calculated as described previously (Wang et al., 2018). For each method, the top 10% of genomic sequences were defined as selective sweeps.

Phylogeny and analysis of protein structure

Protein orthologs of GmSOP5 were identified at the National Center for Biotechnology Information using BlastP (https://blast.ncbi.nlm.nih.gov). Orthologs with high similarity to GmSOP5 from representative species of each genus were downloaded. A phylogenetic tree was constructed using the neighbor-joining method in MEGA v7 (Kumar et al., 2016) with 1,000 bootstrap replicates. Coiled-coil domains in GmSOP5^C were predicted using the Paircoil2 program (McDonnell et al., 2006), and the cutoff score for defining the coiled-coil as positive was 0.5. Protein structures were predicted using I-TASSER (Yang et al., 2015).

DNA and RNA extraction, PCR, and RT-qPCR

Genomic DNA was isolated from fresh trifoliate leaves using a Plant Genomic DNA Kit (#DP305, Tiangen). Messenger RNA was extracted using the TRIzol reagent (#15,596,018, Invitrogen) according to the manufacturer's instructions. Genomic DNA was removed and first-strand cDNA was synthesized using the FastKing RT Kit (with gDNase) (#KR116, Tiangen). KOD One PCR Master Mix (#KMM-201, Toyobo) was used to amplify DNA fragments. RT-qPCR was performed using TB Green Premix Ex Taq (Ti RNaseH Plus) (#RR420A, Takara) on a CFX96 Real-Time System (Bio-Rad) according to the manufacturer's instructions. Gene expression was normalized to that of *GmActin11* (Glyma.18G290800). Three biological replicates and three technical replicates were performed for each gene. All primers used are listed in Table S8.

RNA sequencing

Three replicate samples of seed coats and embryos were collected from developing seeds of wild-type Wm82 and two *GmSOP5*-edited lines at 25 d post-flowering for RNA sequencing. RNA sequencing libraries were constructed using the NEBNext Ultra RNA Library Prep Kit and sequenced on the Illumina NovaSeq. 6000 platform to obtain 150-bp paired-end reads. Bioinformatic analyses were performed as described previously (Li et al., 2024). In brief, Trimmomatic (v0.36) (Bolger et al., 2014) was used to remove low-quality bases or reads, and high-quality clean reads were mapped to the soybean Wm82 v2 reference genome (Schmutz et al., 2010) using STAR (v2.7.1a) (Dobin et al., 2013). Read counts were obtained using featureCounts (Liao et al., 2014). The R package DESeq2 (Love et al., 2014) was used to identify DEGs (FDR-adjusted P value < 0.05, $|\text{Log}_2(\text{fold change})| \geq 1$). Gene expression values were normalized as transcripts per kilobase million (TPM). Enriched KEGG pathways and GO terms in the DEG set were identified using TBtools (Chen et al., 2020).

Subcellular localization

To observe the subcellular localization of GmSOP5 proteins, the coding sequences of *GmSOP5*^{C} from Wm82 and *GmSOP5*^{T} from XHD were cloned into the pUC2300-35S-eGFP vector with the green fluorescent protein (GFP) fused at the C terminus. The resulting constructs were transformed into young leaves of *Nicotiana benthamiana* by *Agrobacterium* infiltration. Leaves were imaged 48 h after infiltration using an LSM 980 laser confocal scanning microscope (Zeiss Microsystems, Jena, Germany). GmSOP5^C and GmSOP5^T protein levels were detected by immunoblotting with an anti-GFP antibody (#AE011, ABclonal), and an anti-actin antibody (#AC009, ABclonal) was used as a protein-loading control.

Vector construction and plant transformation

The *GmSOP5* gene sequence (Glyma.05G085400, Wm82.a2.v1) was downloaded from the Phytozome database. To generate the *CaMV35S*-driven *GmSOP5* construct, the full-length coding sequence of *GmSOP5* was amplified from Wm82 and inserted into the binary vector p3301-FLAG under the control of the *CaMV35S* promoter. The resulting construct was introduced into *Agrobacterium tumefaciens* strain EHA105 and then transferred into the soybean landrace XHD by *Agrobacterium*-mediated cotyledonary node transformation, as described previously (Han et al., 2019). The construct for CRISPR/Cas9-mediated knockout of *GmSOP5* contained two single guide RNAs (sgRNAs) targeting the second exon and was designed using the CRISPRdirect online tool (http://crispr.dbcls.jp/) (Naito et al., 2015). The sgRNAs were cloned into the pHSE401 plasmid, which was then introduced into *A. tumefaciens* strain EHA105 for transformation into the soybean cultivar Wm82. The gene-specific primers used to amplify DNA fragments corresponding to the target sites are listed in Table S8.

Data availability statement

The genome sequencing data previously reported and RNA-Seq sequencing data used in this study are available at NCBI Sequence Read Archive under accession PRJNA681974, PRJNA1069343, and PRJNA1061047, respectively.

ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China (2021YFD1201600), the Agricultural Science and Technology Innovation Program (ASTIP) of the Chinese Academy of Agricultural Sciences, and the earmarked fund for CARS (CARS-04-PS01).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

L.J.Q. and Y.H.L. conceived the project and designed the research. Y.T., H.Z., D.L.L., X.L., C.J.L., J.T.G., and J.N.H. performed the research and analyzed the data. Y.T., J.D.L., Y.P.C., H.W.G., C.S.Z., L.Y., Q.W., S.Y.G., and D.B. collected the phenotypic data. Y.T., H.Z., Z.H.Z., J.C.R., Y.H.L., and L.J.Q. jointly wrote the manuscript. All authors read and approved the final manuscript.

Edited by: Qingxin Song, Nanjing Agricultural University, China

Received Mar. 16, 2026; **Accepted** Aug. 24, 2026

OO: OnlineOpen

REFERENCES

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

Ali, I., and Yang, W.C. (2020). The functions of kinesin and kinesin-related proteins in eukaryotes. Cell Adh. Migr. **14:** 139–152.

Augustine, R.C. (2020). RA1 kinesin prevents cell wall deposition from going off the rails. Plant Cell **32:** 2455–2456.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30:** 2114–2120.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633–2635.

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.

Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y., and Xia, R. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. Mol. Plant **13:** 1194–1202.

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.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics **29:** 15–21.

Doebley, J.F., Gaut, B.S., and Smith, B.D. (2006). The molecular genetics of crop domestication. Cell **127:** 1309–1321.

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.

Duan, Z., Zhang, M., Zhang, Z., Liang, S., Fan, L., Yang, X., Yuan, Y., Pan, Y., Zhou, G., Liu, S., et al. (2022). Natural allelic variation of *GmST05* controlling seed size and quality in soybean. Plant Biotechnol. J. **20:** 1807–1818.

Florencia, P., Lucas, B., and Rotundo, J.L. (2016). Variation in seed protein concentration and seed size affects soybean crop growth and development. Crop Sci. **56:** 3196–3208.

Ganguly, A., Zhu, C., Chen, W., and Dixit, R. (2020). FRA1 kinesin modulates the lateral stability of cortical microtubules through cellulose synthase-microtubule uncoupling proteins. Plant Cell **32:** 2508–2524.

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.

Guo, B., Sun, L., Jiang, S., Ren, H., Sun, R., Wei, Z., Hong, H., Luan, X., Wang, J., Wang, X., et al. (2022). Soybean genetic resources contributing to sustainable protein production. Theor. Appl. Genet. **135:** 4095–4121.

Hammer, K. (1984). Das domestikationssyndrom. Kulturpflanze **32:** 11–34.

Han, J., Guo, B., Guo, Y., Zhang, B., Wang, X., and Qiu, L.J. (2019). Creation of early flowering germplasm of soybean by CRISPR/Cas9 technology. Front. Plant Sci. **10:** 1446.

Hu, Y., Liu, Y., Wei, J.J., Zhang, W.K., Chen, S.Y., and Zhang, J.S. (2023a). Regulation of seed traits in soybean. aBIOTECH **4:** 372–385.

Hu, Y., Liu, Y., Tao, J.J., Lu, L., Jiang, Z.H., Wei, J.J., Wu, C.M., Yin, C.C., Li, W., Bi, Y.D., et al. (2023b). GmJAZ3 interacts with GmRR18a and GmMYC2a to regulate seed traits in soybean. J. Integr. Plant Biol. **65:** 1983–2000.

Kim, M., Schultz, S., Nelson, R.L., and Diers, B.W. (2016). Identification and fine mapping of a soybean seed protein QTL from PI 407788A on Chromosome 15. Crop Sci. **56:** 219–255.

Kong, Z., Ioki, M., Braybrook, S., Li, S., Ye, Z.H., Julie Lee, Y.R., Hotta, T., Chang, A., Tian, J., Wang, G., et al. (2015). Kinesin-4 functions in vesicular transport on cortical microtubules and regulates cell wall mechanics during cell elongation in plants. Mol. Plant **8:** 1011–1023.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. **33:** 1870–1874.

Levesque, A.A., and Compton, D.A. (2001). The chromokinesin Kid is necessary for chromosome arm orientation and oscillation, but not congression, on mitotic spindles. J. Cell Biol. **154:** 1135–1146.

Li, B., Yang, C., Yong, B., Wang, Y., Zhu, W., Gu, Y., An, Z., Yu, H., Chen, M., and He, C. (2025). A 14-3-3 modulator of seed weight and quality for unlocking the yield potential of soybean. Nat. Commun. **16:** 10547.

Li, D., Wang, Q., Tian, Y., Lyv, X., Zhang, H., Hong, H., Gao, H., Li, Y.F., Zhao, C., Wang, J., et al. (2024). TWAS facilitates gene-scale trait genetic dissection through gene expression, structural variations, and alternative splicing in soybean. Plant Commun. **5:** 101010.

Li, J., Jiang, J., Qian, Q., Xu, Y., Zhang, C., Xiao, J., Du, C., Luo, W., Zou, G., Chen, M., et al. (2011). Mutation of rice BC12/GDD1, which encodes a kinesin-like protein that binds to a GA biosynthesis gene

promoter, leads to dwarfism with impaired cell elongation. Plant Cell **23:** 628–640.

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.

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.

Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30:** 923–930.

Liu, B., Fujita, T., Yan, Z.H., Sakamoto, S., Xu, D., and Abe, J. (2007). QTL mapping of domestication-related traits in soybean (*Glycine max*). Ann. Bot. **100:** 1027–1038.

Liu, S., Liu, Z., Hou, X., and Li, X. (2023). Genetic mapping and functional genomics of soybean seed protein. Mol. Breed. **43:** 29.

Liu, S., Zeng, P., Ban, D., Zhang, C., Kong, F., Li, X., and Hou, X. (2025). A natural allele of *PC08* lost in domestication contributes to soybean seed storage protein accumulation. Proc. Natl. Acad. Sci. U.S.A. **122:** e2508709122.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq. 2. Genome Biol. **15:** 550.

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.

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.

McDonnell, A.V., Jiang, T., Keating, A.E., and Berger, B. (2006). Paircoil2: Improved prediction of coiled coils from sequence. Bioinformatics **22:** 356–358.

Nadeem, M., Chen, A., Hong, H., Li, D., Li, J., Zhao, D., Wang, W., Wang, X., and Qiu, L. (2021). *GmMs1* encodes a kinesin-like protein essential for male fertility in soybean (*Glycine max* L.). J. Integr. Plant Biol. **63:** 1054–1064.

Naito, Y., Hino, K., Bono, H., and Ui-Tei, K. (2015). CRISPRdirect: Software for designing CRISPR/Cas guide RNA with reduced off-target sites. Bioinformatics **31:** 1120–1123.

Olsen, K.M., and Wendel, J.F. (2013). A bountiful harvest: Genomic insights into crop domestication phenotypes. Annu. Rev. Plant Biol. **64:** 47–70.

Patil, G., Mian, R., Vuong, T., Pantalone, V., Song, Q., Chen, P., Shannon, G.J., Carter, T.C., and Nguyen, H.T. (2017). Molecular mapping and genomics of soybean seed protein: A review and perspective for the future. Theor. Appl. Genet. **130:** 1975–1991.

Patil, G., Vuong, T.D., Kale, S., Valliyodan, B., Deshmukh, R., Zhu, C., Wu, X., Bai, Y., Yungbluth, D., Lu, F., et al. (2018). Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping. Plant Biotechnol. J. **16:** 1939–1953.

Qi, Z., Guo, C., Li, H., Qiu, H., Li, H., Jong, C., Yu, G., Zhang, Y., Hu, L., Wu, X., et al. (2023). Natural variation in *Fatty Acid 9* is a determinant of fatty acid and protein content. Plant Biotechnol. J. **22:** 759–773.

Reddy, A.S., and Day, I.S. (2001). Kinesins in the Arabidopsis genome: A comparative analysis among eukaryotes. BMC Genomics **2:** 2.

Richardson, D.N., Simmons, M.P., and Reddy, A.S. (2006). Comprehensive comparative analysis of kinesins in photosynthetic eukaryotes. BMC Genomics **7:** 18.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463:** 178–183.

Sheng, L., Hao, S.L., Yang, W.X., and Sun, Y. (2018). The multiple functions of kinesin-4 family motor protein KIF4 and its clinical potential. Gene **678:** 90–99.

Sperschneider, J., Catanzariti, A.M., DeBoer, K., Petre, B., Gardiner, D.M., Singh, K.B., Dodds, P.N., and Taylor, J.M. (2017). LOCALIZER: Subcellular localization prediction of both plant and effector proteins in the plant cell. Sci. Rep. **7:** 44598.

Suetsugu, N., Yamada, N., Kagawa, T., Yonekura, H., Uyeda, T.Q., Kadota, A., and Wada, M. (2010). Two kinesin-like proteins mediate actin-based chloroplast movement in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U.S.A. **107:** 8860–8865.

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.

Tseng, K.F., Wang, P., Lee, Y.J., Bowen, J., Gicking, A.M., Guo, L., Liu, B., and Qiu, W. (2018). The preprophase band-associated kinesin-14 *OsKCH2* is a processive minus-end-directed microtubule motor. Nat. Commun. **9:** 1067.

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.

Wang, S.Z., and Adler, R. (1995). Chromokinesin: A DNA-binding, kinesin-like nuclear protein. J. Cell Biol. **128:** 761–768.

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.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: Protein structure and function prediction. Nat. Methods **12:** 7–8.

Zhang, C., Li, W., Tan, C., Huang, M., Wu, H., Liu, S., Liu, H., Li, X., Miao, Y., Liu, B., et al. (2025). Natural allelic variation in *SW14* determines seed weight and quality in soybean. Nat. Commun. **16:** 8070.

Zheng, H., Feng, X., Wang, L., Shao, W., Guo, S., Zhao, D., Li, J., Yan, L., Miao, L., Sun, B., et al. (2025). *GmSop20* functions as a key coordinator of the oil-to-protein ratio in soybean seeds. Adv. Sci. **12:** e05181.

Zhang, M., Zhang, B., Qian, Q., Yu, Y., Li, R., Zhang, J., Liu, X., Zeng, D., Li, J., and Zhou, Y. (2010). Brittle Culm 12, a dual-targeting kinesin-4 protein, controls cell-cycle progression and wall properties in rice. Plant J. **63:** 312–328.

Zhong, R., Burk, D.H., Morrison, W.H., and Ye, Z.H. (2002). A kinesin-like protein is essential for oriented deposition of cellulose microfibrils and cell wall strength. Plant Cell **14:** 3101–3117.

Zhu, C., and Dixit, R. (2012). Functions of the Arabidopsis kinesin superfamily of microtubule-based motor proteins. Protoplasma **249:** 887–899.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article: http://onlinelibrary.wiley.com/doi/10.1111/jipb.70392/suppinfo

Figure S1. Correlation analyses of seed oil content and protein content in 3 years (2018, 2019, and 2020) in Shijiazhuang

Figure S2. Verification of the stop-gain SNP (chr05:15,709,660 C/T) between Wm82 and XHD by Sanger sequencing
Figure S3. Sequence analysis of GmSOP5
Figure S4. Sanger sequencing for *GmSOP5*-edited lines obtained by CRISPR-Cas9 gene editing
Figure S5. Statistical analysis of seed length, width, and thickness of wild-type Wm82 and two *GmSOP5*-edited lines (*GmSOP5*-CR-1 and *GmSOP5*-CR-2) ($n = 20$, 19, and 18, from left to right), and wild-type XHD and two $GmSOP5^{C}$ overexpression lines ($GmSOP5^{C}$-OE-1 and $GmSOP5^{C}$-OE-2) ($n = 14$, 11, and 12, from left to right)
Figure S6. Identification of *GmSOP5* overexpression lines
Figure S7. Neighbor-joining phylogenetic tree of GmSOP5 and its homologs from dicot and monocot species
Figure S8. Predicted nuclear localization signals (NLS) on the full-length $GmSOP5^{C}$ protein
Figure S9. Expression and relative nuclear localization of $GmSOP5^{C}$-GFP and $GmSOP5^{T}$-GFP in transiently transformed *N. benthamiana*
Figure S10. Differential expression of known oil- and protein-related genes in the seed coat and embryo
Figure S11. Validation of the expression levels of *GmSWEET10a* and *GmSWEET10b* by RT-qPCR
Figure S12. The comparison of seed quality and yield-related traits in lines carrying $GmSOP5^{H1}$ or $GmSOP5^{H2}$ haplotypes among landraces from Hebei and Beijing
Figure S13. Comparison of plant height between wild-type Wm82 and two *GmSOP5*-edited lines (*GmSOP5*-CR-1 and *GmSOP5*-CR-2) and between wild-type XHD and two $GmSOP5^{C}$ overexpression lines ($GmSOP5^{C}$-OE-1 and $GmSOP5^{C}$-OE-2)
Figure S14. The comparison of yield-related traits between wild-type XHD and two *GmSOP5* overexpression lines ($GmSOP5^{C}$-OE-1 and $GmSOP5^{C}$-OE-2)
Figure S15. Yeast one-hybrid assay demonstrating that GmSOP5 does not bind directly to the promoter of *GmSWEET10a* or *GmSWEET10b*
Figure S16. A proposed model illustrating the role of *GmSOP5* in balancing seed quality and yield during soybean breeding
Table S1. Information on 429 diverse soybean accessions for association analysis
Table S2. Summary of QTL significantly associated with soybean seed oil and protein content
Table S3. Information on candidate genes in the *SOP5* locus
Table S4. Descriptive statistics of RNA sequencing for wild-type Wm82 and two *GmSOP5*-edited lines (*GmSOP5*-CR-1 and *GmSOP5*-CR-2)
Table S5. List of DEGs identified by RNA sequencing between wild-type Wm82 and two GmSOP5-edited lines (*GmSOP5*-CR-1 and *GmSOP5*-CR-2)
Table S6. The haplotype and geographic information of soybean accessions in this study
Table S7. The sequence alignment of $GmSOP5^{H1}$ and other well-known Kinesin4A subfamily members, classified into motor domain, middle region, and tail region
Table S8. Primers used in the study