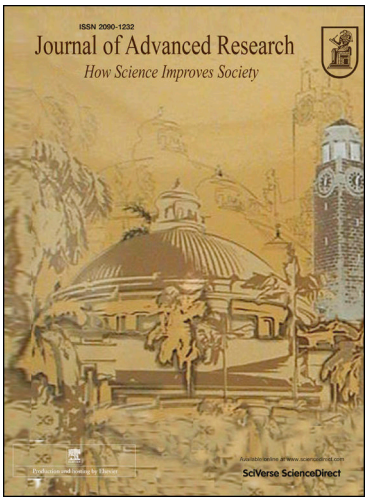
Journal Pre-proofs



Natural variation in *GmSW6* regulates seed weight and quality in soybean

Hao Zhang, Tianli Ge, Shiyu Guo, Guoyu Hu, Like Liu, Xiaoyu Li, Shuang Li, Xiaobo Wang, Xinkang Feng, Haiyang Zheng, Xueqing Wang, Ying-hui Li, Hongning Tong, Li-juan Qiu

PII: S2090-1232(26)00532-1
DOI: https://doi.org/10.1016/j.jare.2026.07.014
Reference: JARE 3071

To appear in: *Journal of Advanced Research*

Received Date: 3 April 2026
Revised Date: 9 June 2026
Accepted Date: 2 July 2026

Please cite this article as: Zhang, H., Ge, T., Guo, S., Hu, G., Liu, L., Li, X., Li, S., Wang, X., Feng, X., Zheng, H., Wang, X., Li, Y-h., Tong, H., Qiu, L-j., Natural variation in *GmSW6* regulates seed weight and quality in soybean, *Journal of Advanced Research* (2026), doi: https://doi.org/10.1016/j.jare.2026.07.014

Natural Variation in *GmSW6* Regulates Seed Weight and Quality in Soybean

Hao Zhang[1,5], Tianli Ge[1,5], Shiyu Guo[1,5], Guoyu Hu[2], Like Liu[3], Xiaoyu Li[1], Shuang Li[1], Xiaobo Wang[4], Xinkang Feng[1], Haiyang Zheng[1], Xueqing Wang[1], Ying-hui Li[1,*], Hongning Tong[1,*], Li-juan Qiu[1,*]

[1]State Key Laboratory of Crop Gene Resources and Breeding / The National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI) / Key Laboratory of Grain Crop Genetic Resources Evaluation and Utilization (MARA) / Bio-breeding Laboratory of Anhui Province, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China.

[2]Crop Institute of Anhui Academy of Agricultural Sciences / Anhui Key Lab of Crops Quality Improving, Hefei 230031, China.

[3]School of Life Sciences, Liaocheng University, Liaocheng 252059, China.

[4]School of Agronomy, Anhui Agricultural University, Bio-breeding Laboratory of Anhui Province, Hefei, 230036, China.

[5]H.Z. T.G. and S.G. contributed equally to this work.

*Correspondence be addressed to: Ying-hui Li (liyinghui@caas.cn), Hongning Tong (tonghongning@caas.cn), Li-juan Qiu (qiulijuan@caas.cn)

Data availability

The data supporting the findings of this study are available within the article and its supplementary data, or from the corresponding authors upon reasonable request.

Acknowledgments

This work was supported by Bio-breeding Laboratory of Anhui Province (2025SWYZ0412), National Key R&D Program of China (2021YFD1201600), National Natural Science Foundation of China (32201756), the Agricultural Science and Technology Innovation Program (ASTIP) of the Chinese Academy of Agricultural Sciences, and the earmarked fund for CARS (CARS-04-PS01).

Author contributions

Y.L., H.T. and L.Q. conceived the study. H.Z., T.G. and S.G. performed the experiments and conducted data analysis. G.H., L.L., X.L., S.L., X.W., X.F., H.Z., and

X.W. provided assistances. H.Z., T.G., S.G. analyzed the data. H.Z., T.G., S.G., Y.L., H.T. and L.Q. wrote the manuscript with the input from all the others. Y.L., H.T. and L.Q. co-supervised the study.

Natural Variation in *GmSW6* Regulates Seed Weight and Quality in Soybean

Abstract

Introduction: Seed weight and nutritional composition (protein and oil content) are critical agronomic traits that collectively determine the yield and quality in soybean. However, the genetic architecture and regulatory mechanisms governing these traits remain poorly understood.

Objectives: This study aimed to identify the key genes and molecular mechanisms governing 100-seed weight and nutritional quality in soybean, providing a theoretical framework for the dual-enhancement of seed weight and protein content.

Methods: A genome-wide association study (GWAS) was conducted using 1,702 diverse soybean cultivars to identify candidate loci associated with seed weight. Functional characterization was conducted through CRISPR-mediated knockout and overexpression analyses. Population genomic analyses were further performed to elucidate the evolutionary history and selection signals of the candidate gene.

Results: We identified *GmSW6* (*Seed Weight 6*), encoding a 2-oxoglutarate Fe(II)-dependent dioxygenase (2OGD), as a master regulator of 100-seed weight. Knockout of *GmSW6* markedly enhanced seed weight and protein content while simultaneously reducing oil content. Mechanistically, we demonstrated that the transcription factor GmSW13 directly activates *GmSW6* expression. This GmSW13-GmSW6 module, in turn, upregulates *GmOLEO1* to coordinately modulate both seed weight and quality. Population genomic analysis revealed that the elite allele, $GmSW6^{G}$, is significantly associated with increased seed weight and has undergone intense positive selection during soybean domestication and modern improvement.

Conclusion: Our findings elucidate a hierarchical genetic pathway governing soybean seed development and provide potent molecular targets for simultaneous improvement of seed weight and protein content in future 'designer' varieties.

Keywords: GWAS, *GmSW6*, *GmSW13*, seed weight, quality, domestication

Introduction

Soybean (*Glycine max* L. Merr.) is a globally indispensable crop, serving as a primary source of high-quality protein and vegetable oil for both human nutrition and animal feed. Given the burgeoning global population, developing high-yield and high-quality cultivars hasve become a critical priority for global food security. Among various yield components, seed weight is a pivotal agronomic attribute; it directly determines total yield and profoundly influences the accumulation of protein and oil. Historically, seed weight has been a primary target of selection during soybean domestication. However, the complex phenotypic correlations among weight, protein, and oil suggest that these traits are governed by an intricate and often antagonistic genetic network [1, 2].

While over 800 quantitative trait loci (QTLs) associated with these traits have been documented in SoyBase (http://www.soybase.org/), only a small fraction of causal genes have been functionally characterized via forward genetics. These identified genes exhibit broad pleiotropy and divergent regulatory paradigms, which can be categorized into three distinct patterns based on their impact on seed traits. First, weight-specific regulators, such as *PP2C-1* (*Phosphatase 2C-1*) [3], *GmGA3ox1* (*Gibberellin 3 Hydroxylase 1*) [4], *GmKIX8-1* (*KIX-domain containing Protein 8-1*) [5], *GmSSS1* (*Soybean Seed Size 1*) [6], *SW16.1* [7], *hsw* [8] and *GmSW17*/*GmSW17.1* [9, 10], exert specific control over seed weight without documented impacts on oil or protein content. In contrast, a second class of regulators, including *PC08* (*Protein Contributor 08*) [11], *GmST05/GmMFT* (*Seed Thickness 05*) [12, 13], *POWR1* (*Protein, Oil, Weight, Regulator 1*) [14], *GmSWEET10a/b* (*Sugars Will Eventually be Exported Transporter 10a/10b*) [15] and *SW14* [16], orchestrate a trade-off pattern where increased seed weight and oil content are positively coupled, yet both are inversely correlated with protein accumulation. Conversely, genes such as *GmOLEO1* (*Oleosin 1*) [17], *GmJAZ3* (*Jamonatezim Domain 3*) [18] and *GmSMS6* (*Small Seed 6*) [19] exhibit an opposite pleiotropic pattern, where enhanced seed weight is accompanied by increased protein content at the physiological expense of oil accumulation. Given the scarcity of genes in this third category, identifying novel regulators that can synergistically enhance both seed weight and protein is critical for overcoming the long-standing trade-off between yield and quality.

The known regulators of soybean seed traits are implicated in diverse biological pathways, ranging from the ubiquitin-proteasome system and phytohormone signaling to transcriptional regulation and enzyme-mediated metabolism [20-22]. Among these, the 2-oxoglutarate Fe(II)-dependent dioxygenase (2OGD) superfamily is of particular interest, as it plays critical roles in both seed development and the biosynthesis of secondary metabolites [23-25]. Our current understanding of 2OGD genes in seed development stems largely from their involvement in GA metabolism, which modulates overall plant growth. Recent studies have demonstrated that 2OGD subfamily members, such as *GA20OX* and *GmGA3ox1*, enhance seed weight, with *GA20OX* additionally promoting oil content [4, 26]. However, the precise molecular mechanisms underlying these functions remain elusive, and non-GA-related 2OGD genes associated with soybean seed weight and quality have rarely been reported.

To address these gaps, we aimed to identify novel 2OGD genes involved in seed development, decipher their molecular mechanisms governing seed weight and quality. To this end~~In this study~~, we identified a novel 2OGD gene, *GmSW6*, which acts as a pleiotropic regulator of soybean seed weight, protein, and oil content. We found that the transcription factor GmSW13 directly promotes *GmSW6* expression, which in turn coordinately enhances the transcription of *GmOLEO1* to modulate both seed weight and quality. Furthermore, we show that the elite allele, *GmSW6*^{G}, has been positively selected during soybean domestication and genetic improvement. These findings provide novel insights into the hierarchical molecular networks governing seed trait determination and offer a valuable genetic resource for the targeted molecular design of high-quality soybean cultivars.

Results

Identification of *GmSW6* as a key regulator of seed weight in soybean

To pinpoint the genes involved in 100-seed weight, we phenotyped a diverse panel of 1,702 cultivated soybean accessions across two eco-geographically distinct locations, Mengcheng (31.88°N, 117.17°E) and Liaocheng city (36.46°N, 115.99°E), over two consecutive years (2018-2019). We observed extensive phenotypic variation in 100-seed weight, ranging from 1.38 g to 41.47 g. Pearson correlation coefficients (r = 0.73–0.86) revealed strong and consistent correlations across different environments (Figure S1 and Table S1). And the broad-sense heritability (h^2) was estimated at 0.92, indicating that the observed diversity is predominantly driven by genetic factors (Table S2).

A genome-wide association study (GWAS) ~~were~~ was subsequently performed using the random model circulating probability unification (FarmCPU) model and 6.58 million high quality SNPs (MAF ≥ 0.05). This analysis uncovered 11 loci significantly associated with 100-seed weight ($-\log_{10} P \geq 8$). Notably, a stable genetic locus on chromosome 6, designated *SW6*, was consistently detected across all environments (Figure S2 and Table S3). Local linkage disequilibrium (LD) decay analysis narrowed this locus down to a 115-kb interval centered on the lead SNP (Chr06:5535154), encompassing nine predicted open reading frames (ORFs) based on the Williams 82 (v4) reference genome (Figure 1a and Table S3). Sequence analysis revealed that six candidate genes (I–VI) possessed nucleotide variations in their promoters, exons, or 3' UTRs (Table S4). However, transcriptomic profiling revealed that only gene V (*Glyma.06G072400*) was significantly upregulated in heavy-seeded accessions compared to light-seeded ones, while others were either negligibly expressed or non-differentially expressed in seeds (Figure 1b, c and Table S5). Further tissue-specific analysis showed that *Glyma.06G072400* is specifically expressed in developing seeds, peaking at 55 days after flowering (55 DAF) (Figure S3a). Taken together, *Glyma.06G072400* was identified as the causal gene underlying the *SW6* locus, and named as *GmSW6*.

Natural variation of *GmSW6* governs ~~with~~ seed weight diversity

Sequence analysis of the 3.1-kb genomic region of *GmSW6* across the 1,702 accessions identified six polymorphic sites defining four major haplotypes (Figure 1d). Accessions harboring H1 ($GmSW6^{G}$) exhibited a significantly higher 100-seed weight than those with H2-H4. Given that H2–H4 share a 10-bp insertion and show no significant phenotypic divergence, they were collectively designated as the $GmSW6^{Ins}$ allele (Figure 1d and Figure S3b). This phenotypic effect remained robust across geographically distinct environments, where $GmSW6^{G}$ consistently outperformed $GmSW6^{Ins}$ (Figure 1e), suggesting that the 10-bp variation is the likely causative variant responsible for the observed transition in seed weight.

GmSW6 encodes a 2-oxoglutarate-dependent dioxygenase (2OGD) containing highly conserved DIOX-N and 2OG-FeII_Oxy domains (Figure S4). Protein sequence

alignment revealed that the catalytic triad (consisting of H-D-H) [27], essential for iron-binding, is conserved across all examined species. The 10-bp insertion in *GmSW6^{Ins}* induces a frameshift mutation that truncates the conserved catalytic triad (H-D-H), specifically causing the loss of a critical histidine (H) residue essential for iron-binding ligand, thereby disrupting the 2OG-FeII_Oxy catalytic site (Figure 1f). 3D protein modeling further confirmed that while the overall scaffold remains intact, the C-terminal structural rearrangement in *GmSW6^{Ins}* likely abolishes its dioxygenase activity (Figure S5a).

Evolutionary analysis indicated that the 10-bp deletion (GmSW6^{G}) is a derived variant unique to *G. max*, as the 10-bp sequence is ancestral and conserved across other legumes (Figure S4). Subcellular localization in *Nicotiana benthamiana* demonstrated that both *GmSW6^{G}* and *GmSW6^{Ins}* localize to the nucleus and plasma membrane, indicating the insertion does not influence protein trafficking (Figure S5b). To facilitate the utilization of this unique variant in breeding, we developed a KASP marker based on the 10-bp polymorphism, which successfully distinguished the two alleles with 100% accuracy in a validation subset, providing an efficient tool for molecular marker-assisted selection of seed weight in soybean (Figure S6).

Knockout *GmSW6* increases seed size via cell expansion

To functionally validate the role of *GmSW6* in regulating 100-seed weight, we employed clustered, regularly interspaced, short palindromic repeat (CRISPR)/CRISPR-associated 9 (Cas9) to generate targeted mutations in the soybean cultivar Williams82 (W82), which carries the *GmSW6^{G}* allele (Figure 2a). We successfully obtained two homozygous mutant lines, *sw6-1* (-1 bp) and *sw6-2* (-2 bp), both of which harbor frameshift mutations (Figure 2a and Figure S7a-c). Phenotypic analysis revealed that knockout of *GmSW6* significantly enhanced seed size and weight. Specifically, seed length and width increased by 6.5% and 5.9%, respectively, leading to a substantial rise in 100-seed weight of 8.7% in *sw6-1* and 11.9% in *sw6-2* (Figure 2b-f). However, this increase in individual seed mass was offset by a slight reduction in the number of pods and seeds per plant. Since other agronomic traits, such as plant height, remained unaffected, the total seed yield per plant showed no significant increase (Figure 2g, h and Figure S7d~~, e~~-f).

To determine the cellular basis for the increased seed size, we performed resin sectioning and scanning electron microscopy (SEM) analyses. The results demonstrated that cell size was significantly increased in the mutants compared to W82, while the total cell number remained comparable between the two groups (Figure 2i-l). These findings collectively demonstrate that the enlargement of seeds in *sw6* mutants is primarily attributed to enhanced cell expansion rather than an increase in cell proliferation.

***GmSW6* regulates seed development independently of its homologs**

The soybean genome contains two GmSW6 homologs, Glyma.04G070500 and

Glyma.04G070600, which shared 61.73% and 50.16% sequence ~~identify~~ identity with GmSW6, respectively (Figure ~~S7f~~S7g). Transcriptomic analysis using Phytozome data revealed distinct expression patterns among these three paralogs. Notably, *GmSW6* exhibited significantly higher expression levels in developing seeds compared to its two homologs, suggesting it is the predominant functional member in this specific tissue (Figure ~~S7g~~S7h). To further investigate whether these homologs provide functional redundancy, we quantified their expression levels in the *sw6* mutant lines. No significant alterations in the transcript levels of either *Glyma.04G070500* or *Glyma.04G070600* were detected in response to the *GmSW6* mutation (Figure ~~S7h~~S7i). This observed lack of transcriptional compensation indicates that *GmSW6* operates independently in the regulation of seed development. Consequently, the phenotypic changes observed in the *sw6* lines are primarily attributable to the loss of *GmSW6* function itself rather than any indirect effects from its homologs.

Elevated *GmSW6* expression reduces seed weight but enhances overall yield

RNA-seq analysis and RT-qPCR revealed that $GmSW6^{Ins}$ transcripts accumulate to significantly higher levels in seeds compared to the $GmSW6^{G}$ isoform (Figure 3a, S8a and Table S6). To determine whether this differential expression is controlled at the transcriptional level, we analyzed the activity of ~2.0 kb promoter sequences from both alleles using a dual-luciferase assay. Interestingly, no significant difference in promoter activity was observed, suggesting that promoter variation is not the primary cause of the divergent expression levels (Figure S8b). Consistent with the transcript levels, transient expression in *Nicotiana benthamiana* leaves showed that $GmSW6^{Ins}$ protein accumulation was markedly higher than that of $GmSW6^{G}$ (Figure 3b, c). The results indicate that the 10-bp insertion in the second exon significantly enhances the stability of *GmSW6* transcripts and proteins, whereas $GmSW6^{G}$ acts as a weak allele due to low transcript and protein persistence.

Consistent with the observation that the knockout of *GmSW6* leads to reduced transcript levels and significantly larger seeds (Figure 2, Figure ~~S7h~~S7i), these results further suggest that *GmSW6* levels are inversely correlated with seed size. To evaluate the impact of elevated *GmSW6* levels on seed traits, we generated overexpressed (OE) transgenic lines of the $GmSW6^{Ins}$ allele under the CaMV 35S promoter in the W82 background (Figure 3d, e). Two representative lines, $GmSW6^{Ins}$-OE1 and $GmSW6^{Ins}$-OE2, were selected for detailed phenotypic analysis (Figure 3f). Overexpression of $GmSW6^{Ins}$ resulted in significantly smaller and lighter seeds, with reductions in both seed length and width (Figure 3g-i). Specifically, the 100-seed weight decreased by 9.1% (18.3 ± 1.0 g) and 8.2% (18.4 ± 1.1 g) in the OE lines compared to W82 (20.0 ± 1.2 g) (Figure 3j). Remarkably, despite the reduction in individual seed weight, the $GmSW6^{Ins}$-OE lines exhibited increased plant height, pod number, and total seed number per plant (Figure 3k-n). These compensatory increases led to a ~7.6% increase in total seed yield per plant relative to the control (Figure 3o). Collectively, these findings demonstrate that *GmSW6* acts as a pivotal regulator of yield architecture, where its overexpression promotes a favorable shift toward increased seed numbers to achieve a net gain in total

yield.

GmSW13 acts upstream of *GmSW6* to regulate seed weight and quality

To identify the upstream regulators of *GmSW6*, we performed a yeast one-hybrid (Y1H) screen using a cDNA library against *GmSW6* promoter region. Among the 16 positive clones identified, four were identical to *qFT13-3* [28], which encodes a pseudo-response regulator (PRR) transcription factor, hereafter designated as *GmSW13* (Table S7). In parallel, promoter sequence analysis of *GmSW6* revealed several cis-regulatory elements, including one G-box motif and two ABRE motifs. These elements contain the core ACGT sequence, which is recognized binding site for PRR-type transcription factors [29] (Table S8).

To validate the physical and functional interaction between *GmSW13* and *GmSW6* promoter, we conducted a systematic biochemical analysis. First, in vivo binding was established using yeast one-hybrid (Y1H) assays, which confirmed that GmSW13 specifically interacts with a *GmSW6* promoter segment (S1) containing the G-box and ABRE motifs (Figure 4a). To further prove this interaction is direct, we performed electrophoretic mobility shift assays (EMSA) for in vitro validation. The results demonstrated that the GST-GmSW13 fusion protein specifically bound to ACGT-containing DNA probes (P1 and P2), whereas no binding was observed for the GST-only control, indicating a high-affinity physical interaction (Figure 4b-d). Finally, to determine the functional consequence of this binding, we performed transient expression assays in *Nicotiana benthamiana* leaves. These assays revealed that GmSW13 significantly enhanced the transcriptional activity of the *GmSW6* promoter (Figure 4e, f). Collectively, these results demonstrate that *GmSW13* acts as a direct upstream activator that modulates *GmSW6* expression through sequence-specific promoter binding.

To further validate this regulatory relationship in soybean, we generated *sw13* mutants (*sw13-1* and *sw13-2*) using the CRISPR/Cas9 system in the W82 background (Figure 4g, Figure S9). Consistent with *GmSW13* functioning as a positive regulator, the expression of *GmSW6* was significantly attenuated in the *sw13* mutants (Figure 4h). Phenotypic characterization further supported the functional role of *GmSW13* in seed development. Compared to W82, *sw13* mutants exhibited marked increases in seed length (2.6-3.6%) and width (2.4-3.0%), which translated into a substantial 9.9-12.3% gain in 100-seed weight (Figure 4j-l), suggesting that *GmSW13* act as a negative regulator of seed size and weight. Collectively, these findings demonstrate that *GmSW13* negatively regulates seed size and weight by directly activating the transcription of *GmSW6*.

Identification of downstream targets of the *GmSW13–GmSW6* module

To identify the downstream effectors of the GmSW13-GmSW6 regulatory module, we performed comparative the transcriptomic profiling via RNA-seq on development seeds of *sw13*, *sw6* and W82. In the *sw13* mutants, we identified 320 differentially expressed

genes (DEGs), while 808 DEGs were identified in *sw6* mutants (Figure S10 and Table S9). Given that *GmSW13* positively regulates *GmSW6* expression, we specifically focused on the overlapping DEGs that exhibited consistent expression patterns in both mutants compared to W82. A total of 84 common DEGs were identified (Figure 5a and Table S10). Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis revealed that these shared DEGs were predominantly involved in sugar metabolic processes, including the starch and sucrose metabolism pathways (Figure 5b). Consistent with these transcriptomic findings, biochemical quantification showed that glucose, fructose, and sucrose levels were significantly lower in *sw13* and *sw6* seeds, whereas these sugars were significantly elevated in $GmSW6^{Ins}$-OE seeds (Figure 5c-e).

Notably, the identified common DEGs included *GmOLEO1*, a gene previously reported to increase seed size and weight while modifying oil and protein content (Table S10) [17]. In alignment with our transcriptomic data, *GmOLEO1* expression was significantly down-regulated in both *sw6* and *sw13* mutants but markedly up-regulated in $GmSW6^{Ins}$-OE plants (Figure 5f). Given the regulatory influence of the *GmSW13-GmSW6* module on *GmOLEO1*, we further investigated its impact on seed quality. *sw13* mutants exhibited a 1.04%-1.22% increase in protein content accompanied by a 0.47%-0.55% decrease in oil content relative to the W82 (Figure 5g, h). Similarly, *sw6* mutants displayed significantly higher protein content and lower oil content compared to the control (Figure 5g h). Collectively, these results demonstrate that *GmSW13-GmSW6* orchestrated a downstream regulatory network, including *GmOLEO1* and sugar metabolism pathway, to coordinately regulate both soybean seed weight and quality.

GmSW13-GmSW6 haplotype combinations enhance soybean seed weight

To investigate whether *GmSW6* was a target of selection during soybean domestication, we analyzed nucleotide diversity (π) and population differentiation (*F*st) using a panel of 2,214 re-sequenced accessions [30]. The significantly reduced genetic diversity in cultivars relative to wild soybeans provides strong evidence of a selective sweep at the *GmSW6* locus (Figure 6a). Notably, the frequency of the favorable $GmSW6^{G}$ allele exhibited a continuously increasing pattern from 5% in wild soybean to 59% in landraces and further to 91% in improved cultivars, a trend that mirrors the historical increase in 100-seed weight (Figure 6b).

Further analysis of the upstream regulator *GmSW13* identified two primary haplotypes, $GmSW13^{H1}$ and $GmSW13^{H2}$, with $GmSW13^{H1}$ associated with significantly higher seed weight (Figure S11a, b). Similar to $GmSW6^{G}$, the frequency of $GmSW13^{H1}$ rose from 4% in wild soybeans to 68% during breeding (Figure S11c). In contrast, while *GmOLEO1* haplotypes also influenced seed weight, their frequency did not show significant directional changes, suggesting *GmOLEO1* was not a primary selection target (Figure S12).

By examining haplotype combinations, we found that accessions harboring the $GmSW13^{H1}$/$GmSW6^{G}$ elite pair exhibited significantly superior seed weight compared

to all other combinations (Figure 6c). Notably, this optimal haplotype pair is present in only 37% of landrace and 75% of improved cultivars, highlighting a substantial underutilized potential for targeted molecular breeding (Figure 6b). Furthermore, ~~Reciprocal background comparisons further substantiated a strong additive genetic effect:~~ *GmSW13*^{H1} consistently outperformed *GmSW13*^{H2} regardless of the *GmSW6* backgrounds, and vice-versa (Figure 6c). These findings indicate that *GmSW13* and *GmSW6* function as a coordinated regulatory~~independent yet synergistic~~ modules~~,~~ where their co-selection of their elite alleles can ~~lead to~~ improve~~maximized~~ seed weight ~~gains~~ in soybean breeding ~~improvement~~ programs.

Geographic analysis revealed that the *GmSW6*^{G} allele predominates in both northern (NR) and southern regions (SR), whereas *GmSW6*^{Ins} is more prevalent in the central region (CR) of China. Crucially, while the elite *GmSW13*^{H1}/*GmSW6*^{G} combination dominates the NR, it remains significantly underrepresented in other areas, appearing in only 13% of CR and 36% of SR accessions (Figure 6d). This geographical disparity suggests that the yield benefits conferred by this specific haplotype module have yet to be fully exploited in CR and SR. These results identify a significant opportunity for regional yield enhancement through the targeted selection and strategic introduction of this elite genetic module into local breeding programs.

Potential for utilization in breeding

To further explore the potential application of *GmSW6* in molecular breeding, we selected a commercially important soybean cultivar Zhonghuang13 (ZH13) widely cultivated in China, as the recipient. We introduced an overexpression vector containing the coding region of *GmSW6*^{Ins} into ZH13 (Figure 6e, f). Phenotypic analysis revealed that the overexpression plants exhibited a significant decrease in seed length, seed width, 100-seed weight, but increase in oil content and seed weight per plant compared to the ZH13 (Figure 6g-i and Figure S13). These findings demonstrate that the *GmSW6* holds potential for application in soybean breeding programs aimed at enhancing oil content and yield across diverse cultivars.

Discussion

Identifying the genetic and molecular mechanisms governing seed yield and quality is fundamental to guiding crop breeding and accelerating variety improvement. While several genes controlling seed weight have been identified, the underlying regulatory networks remain largely elusive, and efficient strategies to simultaneously enhance yield and quality are still lacking. In this study, we identified *GmSW6* as a 2OGD protein that negatively regulates seed weight and protein content, while positively modulating oil content. Although 2OGD enzymes are known regulators of multiple plant processes, their specific involvement in seed weight control has remained poorly defined. Current knowledge is primarily restricted to gibberellin-related family members, such as *GA20OX* and *GmGA3ox1* [4, 26]. To address this gap, our GWAS and functional analyses identified *GmSW6* as a seed weight-associated gene encoding

a 2OGD that phylogenetically diverges from the gibberellin subfamily (Figure S7g). This finding reveals a previously unrecognized role for non-gibberellin 2OGDs in seed weight regulation, expanding our understanding of the metabolic enzymes involved in crop development.

Our investigation elucidated a regulatory pathway through which *GmSW6* modulates seed traits. We demonstrated that GmSW13 binds to the *GmSW6* promoter to activate its transcription, which in turn upregulates *GmOLEO1* expression to influence seed size and composition (Figure 5i). However, the precise biochemical link between GmSW6 and GmOLEO1 requires further characterization. If GmSW6 indeed functions as a canonical 2OGD enzyme, it would most likely participate in metabolic processes and regulate *GmOLEO1* expression indirectly through metabolic or signaling molecules that remain to be identified. However, we cannot completely exclude alternative modes of action. There are precedents in which proteins originally characterized as metabolic enzymes have acquired additional regulatory functions during evolution [8]. Therefore, it remains possible that *GmSW6* possesses functions beyond its predicted enzymatic activity.

Guided by this regulatory framework, it is notable that ~~Notably,~~ while GmSW13 is a homolog of Arabidopsis PRR7, a known negative regulator of flowering [31], its specific role in seed development was previously unclear. Our findings showed that *GmSW13* participates in seed regulation through this specific pathway, providing novel insights into the pleiotropic roles of PRR family proteins. Furthermore, the overexpression of *GmOLEO1* produced pleiotropic agronomic effects: it increased plant height and pod number while reducing 100-seed weight, ultimately resulting in a significant net increase in seed yield per plant [17]. This phenotypic pattern, also observed in *GmSW6*-overexpressing lines (Figure 3), indicates that the GmSW13-GmSW6-GmOLEO1 module may coordinately regulate both individual seed traits and overall plant architecture to enhance yield performance (Figure 5i). Nevertheless, the detailed molecular and genetic mechanisms operating within this regulatory network warrant further investigation.

In many crop species, seed yield and protein content often display a strong negative correlation, presenting a significant challenge for breeders. This trade-off is particularly pronounced in soybean. While gene like *GmSWEET10a/10b* [15], *GmST05/GmMFT* [12, 13] and *SW14* [16] have been linked to these traits, their manipulation typically improves one at the expense of the other. Notably, recent studies have demonstrated that loss of *GmSMS6* function can simultaneously enhance both yield and protein content, offering a promising exception to this typical trade-off [19]. Notably, our results showed that *sw6* mutants exhibited significantly increased protein content without a concomitant reduction in total yield compared to the W82 control (Figure 2h, 5g). These findings suggest that the *GmSW6*-mediated pathway may offer a unique mechanism for overcoming the traditional yield-protein trade-off.

Seed weight has been a primary target during soybean domestication, yet our

analysis reveals that not all favorable alleles have been fully utilized. While both *GmSW6*^{G} and *GmSW13*^{H1} haplotypes (associated with increased seed weight) underwent significant expansion during domestication, the favorable allele *GmOLEO1*^{H1} did not exhibit clear selection signatures (Figure S11, S12). Moreover, the superior haplotype combination *GmSW6*^{G}/*GmSW13*^{H1}, which confers the highest seed weight, remains underutilized in many cultivated soybeans (Figure 6c). These findings underscore the considerable breeding potential of deploying this specific haplotype pair. More broadly, our results suggest that valuable genetic components often remain only partially exploited during crop improvement. Unlocking such "hidden" genetic variation represents a significant opportunity for the future of soybean molecular design and global food security.

Methods

Plant materials and growth conditions

The primary genetic resource consisted of 2,214 accessions from the Chinese National Soybean GeneBank, encompassing a broad spectrum of genetic diversity [30, 32-34]. From this collection, 1,702 cultivars with available 100-seed weight phenotypic data were selected for GWAS. Field trials were conducted at two locations: Mengcheng, Anhui province (31.83°N, 117.25°E) and Liaocheng, Shandong province (36.46°N, 115.99°E), across two consecutive growing seasons (2018 and 2019) using two biological replicates per site.

For functional validation, the soybean cultivar Williams 82 (W82) and zhonghuang13 (ZH13) served as reference genotypes for genetic transformation. Field-based evaluations of W82-derived transgenic lines, including *sw6* and *GmSW6*^{Ins}-OE overexpression lines, were grown in Sanya (18.23°N, 109.9°E) from November 2024 to March 2025. Subsequently, *sw13* mutants and ZH13-derived overexpression lines were evaluated in Beijing (39.91°N, 116.40°E) from June to October 2025. A replicated trial with 18 replicates was used to assess yield. Each plot consisted of 2-meter rows with 0.5-meter spacing, totaling a plot area of 0.1 square meters.

GWAS for 100-seed weight

Association mapping was performed using 6.45 million high-quality SNPs with a minor allele frequency (MAF) > 0.05[30]. FarmCPU was utilized to conduct the association analyses [35]. The significance threshold was set at $P = 1\times10^{-8}$ (0.05 / total SNPs) based on Bonferroni correction. Linkage disequilibrium (LD) was calculated using PLINK software [36].

DNA isolation and detection of *GmSW6* haplotype

Genomic DNA was extracted from young leaves using the cetyltrimethylammonium bromide (CTAB) method [37]. Genetic variation at the *GmSW6* locus, extending from the 2-kb promoter region to the 3'-UTR, was analyzed using existing resequencing data

[30]. To distinguish the 10-bp deletion associated with 100-seed weight, KASP-PCR primers were designed. Each assay utilized two allele-specific forward primers (labeled with FAM and HEX fluorophores) and one common reverse primer. Primer sequences are detailed in Table S11.

Plasmid construction and plant transformation

The CRISPR/Cas9 targets for the first exons of *GmSW6* and *GmSW13* were designed via CRISPR-P (http://crispr.hzau.edu.cn) and cloned into the JRH0645 vector [38]. For overexpression, the CDS of *GmSW6^{Ins}* was amplified from seeds of the Jurongbianqingdou variety at 55 DAF and cloned into the 0641-FLAG vector under the control of the CaMV 35S promoter [39]. Constructs were introduced into *Agrobacterium tumefaciens* strain EHA105 and transformed into W82 using the cotyledon-node method [40]. The specific primers used are listed in Table S11.

RNA extraction and RT-qPCR

Total RNA was extracted using EasyPure® Universal Plant Total RNA Kit (TransGen, China). cDNA synthesis was performed with HiScript IV All-in-One Ultra RT SuperMix (Vazyme, China). RT-qPCR was conducted on a QuantStudio™ 1 Real-Time PCR Instrument using SupRealQ Ultra Hunter SYBR qPCR Master Mix (Vazyme, China). The soybean *Actin11* gene (*Glyma.18G290800*) served as an internal reference. The primers used are listed in Table S11.

Protein stucture prediction and sequence alignment

Three-dimensional structures of GmSW6^{G} and GmSW6^{Ins} proteins were predicted using AlphaFold and visualized with PyMOL. Sequence alignments were performed using MEGA11, with visualization through Jalview and the iTOL online tool (https://itol.embl.de/).

RNA-seq analysis

RNA was extracted from seeds of W82, *sw6*, and *sw13* at 55 DAF (three biological replicates). Sequencing was performed by Novogene. Gene expression was quantified as FPKM (fragments per kilobase per million mapped reads). Differentially expressed genes (DEGs) were identified using criteria of $|\log_2(\text{fold change})| \geq 1$ and $P < 0.05$. KEGG enrichment was analyzed via KOBAS 3.0 (http://bioinfo.org/kobas/genelist/).

Microscopy analysis

Seeds at the 55 DAF were collected and dehydrated in a 1:1 mixture of xylene and 70% ethanol for 1 h at 55 °C. After dehydration, samples were embedded in paraffin, longitudinally sectioned, and stained with alkaline toluidine blue O solution for observation under light microscopy. Sections were subsequently rinsed, dehydrated, cleared with xylene, and mounted with neutral gum for long-term preservation.

For scanning electron microscopy (SEM), the inner surface of the cotyledon from 55 DAF seeds was examined, focusing on the central region for cellular analysis. Cell number and size were quantified using ImageJ software.

Subcellular localization

The CDSs of *GmSW6*^{G} and *GmSW6*^{Ins} were cloned into the pCAMBIA2300-35s-eGFP vector [41] at the *Kpn*I and *Xba*I sites using the In-Fusion system. Red fluorescent protein (RFP) was used as a nucleus membrane marker. Constructs were transformed into EHA105, and then infiltrated into *N. benthamiana* leaves for transient expression. The subcellular localization of the proteins was visualized using a confocal laser scanning microscope (Zeiss LSM980). The primers are provided in Table S11.

Transient expression assay of promoter activity

About 2-kb promoter sequences of *GmSW6*^{G} and *GmSW6*^{Ins} were amplified from genomic DNA of W82 and Jurongbianqingdou, respectively, and cloned into the *Sal*I and *Noc*I sites of the pGreenII0800-LUC vector [42] to generate reporter constructs. The GmSW13-GFP fusion protein serves as the effector, with empty vector as the negative control. Luminescence was measured using the Dual-Luciferase Reporter Assay System (Promega), with the *REN* gene serving as an internal control. The primers used are listed in Table S11.

Y1H assay

The *GmSW13* CDS was cloned into the pGADT7 (AD) vector and the *GmSW6* promoter segment (S1) of was cloned into the pAbAi vector (see Table S11 for primers and cloning sites). Both constructs were co-transformed into the Y1H Gold yeast strain; the empty vector pGADT7 served as a negative control. Binding was detected on SD/-Leu-Ura medium supplemented with Aureobasidin A (AbA).

EMSA assay

The CDS of *GmSW13* was cloned into the pGEX4T-1 vector (see Table S11 for primers and cloning sites) and the recombinant protein GST-GmSW13 was expressed in *Escherichia coli* BL21 and subsequently purified. Biotin-labeled oligonucleotide probes were used to detect DNA-protein complexes following the manufacturer's protocol (Beyotime, GS009).

Genetic diversity analysis

SNP date were filtered to exclude those with > 10% missing data or a minor allele frequency (MAF) < 0.01. The soybean germplasm was classified into three populations: *Glycine soja*, landraces, and cultivars. *F*st values were calculated using VCFtools with a sliding window (5 kb to 2 kb). Nucleotide diversity (π) was evaluated within the same windows to assess genomic divergence.

References

[1] Y. Hu, Y. Liu, J. Wei, W. Zhang, S. Chen, J. Zhang, Regulation of seed traits in soybean, aBIOTECH 4 (2023) 372-385. doi:10.1007/s42994-023-00122-8.

[2] J. Chung, H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, P.B. Cregan, et al. The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci. 43 (2003) 1053-1067. doi:10.2135/cropsci2003.1053.

[3] X. Lu, Q. Xiong, T. Cheng, Q.T. Li, X.L. Liu, Y.D. Bi, et al. A *PP2C-1* allele underlying a quantitative trait locus enhances soybean 100-seed weight. Mol. Plant 10 (2017) 670-684. doi:10.1016/j.molp.2017.03.006.

[4] D. Hu, X. Li, Z. Yang, S. Liu, D. Hao, M. Chao, et al. Downregulation of a gibberellin 3β-hydroxylase enhances photosynthesis and increases seed yield in soybean. New Phytol. 235 (2022) 502-517. doi:10.1111/nph.18153.

[5] C. Nguyen, K. Paddock, Z. Zhang, M.G. Stacey, GmKIX8-1 regulates organ size in soybean and is the causative gene for the major seed weight QTL *qSw17-1*. New Phytol. 229 (2021) 920-934. doi:10.1111/nph.16928.

[6] W. Zhu, C. Yang, B. Yong, Y. Wang, B. Li, Y. Gu, et al. 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 (2022) 1375-1392. doi:10.1111/nph.18461.

[7] X. Chen, C. Liu, P. Guo, X. Hao, Y. Pan, K. Zhang, et al. Differential *SW16.1* allelic effects and genetic backgrounds contributed to increased seed weight after soybean domestication. J. Integr. Plant Biol. 65 (2023) 1734-1752. doi:10.1111/jipb.13480.

[8] S. Wei, B. Yong, H. Jiang, Z. An, Y. Wang, B. Li, et al. A loss-of-function mutant allele of a glycosyl hydrolase gene has been co-opted for seed weight control during soybean domestication. J. Integr. Plant Biol. 65 (2023) 2469-2489. doi:10.1111/jipb.13559.

[9] H. Zhang, L. Yang, S. Guo, Y. Tian, C. Yang, C. Zhao, et al. A natural allelic variant of *GmSW17.1* confers high 100-seed weight in soybean. Crop J. 12 6 (2024) 1709-1717. doi:10.1016/j.cj.2024.10.004.

[10] S. Liang, Z. Duan, X. He, X. Yang, Y. Yuan, Q. Liang, et al. Natural variation in *GmSW17* controls seed size in soybean. Nat. Commun. 15 (2024) 7417. doi:10.1038/s41467-024-51798-5.

[11] S. Liu, P. Zeng, D. Ban, C. Zhang, F. Kong , X. Li, et al. A natural allele of *PC08* lost in domestication contributes to soybean seed storage protein

accumulation. P. Natl. Acad. Sci. USA 122 (2025) e2508709122. doi:10.1073/pnas.2508709122.

[12] Z. Duan, M. Zhang, Z. Zhang, S. Liang, L. Fan, X. Yang, et al. Natural allelic variation of *GmST05* controlling seed size and quality in soybean. Plant Biotechnol. J. 20 (2022) 1807-1818. doi:10.1111/pbi.13865.

[13] Cai, Z. Xian, P. Cheng, Y. Y. Zhong, Y. Yang, Q. Zhou, et al. MOTHER-OF-FT-AND-TFL1 regulates the seed oil and protein content in soybean. New Phytol. 239 (2023) 905-919. doi:10.1111/nph.18792.

[14] Goettel, W. Zhang, H. Li, Y. Z. Qiao, H. Jiang, D. Hou, et al. *POWR1* is a domestication gene pleiotropically regulating seed quality and yield in soybean. Nat. Commun. 13 (2022) 3051. doi:10.1038/s41467-022-30314-7.

[15] S. Wang, S. Liu, J. Wang, K. Yokosho, B. Zhou, Y.C. Yu, et al. Simultaneous changes in seed size, oil content and protein content driven by selection of *SWEET* homologues during soybean domestication. Natl. Sci. Rev. 7 (2020) 1776-1786. doi:10.1093/nsr/nwaa110.

[16] C. Zhang, W. Li, C. Tan, M. Huang, H. Wu, S. Liu, et al. Natural allelic variation in *SW14* determines seed weight and quality in soybean. Nat. Commun. 16 (2025) 8070. doi:10.1038/s41467-025-63582-0.

[17] D. Zhang, H. Zhang, Z. Hu, S. Chu, K. Yu, L. Lv, et al. Artificial selection on *GmOLEO1* contributes to the increase in seed oil during soybean domestication. PLoS Genet. 15 (2019) e1008267. doi:10.1371/journal.pgen.1008267.

[18] Y. Hu, Y. Liu, J. Tao, L. Lu, Z.H. Jiang, J.J. Wei, et al. GmJAZ3 interacts with GmRR18a and GmMYC2a to regulate seed traits in soybean. J. Integr. Plant Biol. 65 (2023):1983-2000. doi:10.1111/jipb.13494.

[19] B. Li, C. Yang, B. Yong, Y. Wang, W. Zhu, Y. Gu, et al. A 14-3-3 modulator of seed weight and quality for unlocking the yield potential of soybean. Nat. Commun. 16 (2025) 10547. doi:10.1038/s41467-025-65598-y.

[20] P. Bailey, B. Ortmann, A. Martinelli, J.W. Houghton, A.S.H. Costa, S.P. Burr, et al. ABHD11 maintains 2-oxoglutarate metabolism by preserving functional lipoylation of the 2-oxoglutarate dehydrogenase complex. Nat. Commun. 11 (2020) 4046. doi:10.1038/s41467-020-17862-6.

[21] T. Lange, N. Atiq, M. Pimenta Lange, GAS2 encodes a 2-oxoglutarate dependent dioxygenase involved in ABA catabolism. Nat. Commun. 14 (2023): 7602. doi:10.1038/s41467-023-43187-1.

[22] T. Hu, Y. Wang, Q. Wang, N. Dang, L. Wang, C. Liu, et al. The tomato 2-oxoglutarate-dependent dioxygenase gene *SlF3HL* is critical for chilling stress

tolerance. Hortic. Res. 6 (2019): 45. doi:10.1038/s41438-019-0127-5.

[23] C. Zhang, L. Gao, J. Sun, J. Jia, Z. Ren, Haplotype variation of Green Revolution gene *Rht-D1* during wheat domestication and improvement. J. Integr. Plant Biol. 56 (2014) 774-780. doi:10.1111/jipb.12197.

[24] J. Peng, D. Richards, N. Hartley, G.P. Murphy, K.M. Devos, J.E. Flintham, et al. 'Green revolution'genes encode mutant gibberellin response modulators. Nature 400 (1999) 256-261. doi:10.1038/22307.

[25] W. Spielmeyer, M. Ellis, P. Chandler, Semidwarf (sd-1),"green revolution" rice, contains a defective gibberellin 20-oxidase gene. P. Natl. Acad. Sci. 99 (2002) 9043-9048. doi:10.1073/pnas.132266399.

[26] X. Lu, Q. Li, Q. Xiong, W. Li, Y.D. Bi, Y.C. Lai, et al. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. Plant J. 86 (2016) 530-544. doi:10.1111/tpj.13181.

[27] J. Dunwell, A. Purvis, S. Khuri, Cupins: the most functionally diverse protein superfamily? Phytochemistry 65 (2004) 7-17. doi:10.1016/j.phytochem.2003.08.016.

[28] Y. Li, L. Zhang, J. Wang, X. Wang, S. Guo, Z. Xu, et al. Flowering time regulator *qFT13‐3* involved in soybean adaptation to high latitudes. Plant Biotechnol. J. 22 (2023) 1164-1176. doi.10.1111/pbi.14254.

[29] T. Liu, J. Carlsson, T. Takeuchi, L. Newton, and E.M. Farre, et al. Direct regulation of abiotic responses by the *Arabidopsis* circadian clock component PRR7. Plant J. 76 (2013) 101-114. doi.10.1111/tpj.12276.

[30] Y. Li, C. Qing, L. Wang, C. Jiao, H. Hong, Y. Tian, et al. Genome-wide signatures of geographic expansion and breeding process in soybean. Sci. China Life Sci. 66 (2023) 350-365. doi:10.1007/s11427-022-2158-7.

[31] Y. Yamamoto, E. Sato, T. Shimizu, N. Nakamich, S. Sato, T. Kato, et al. Comparative genetic studies on the *APRR5* and *APRR7* genes belonging to the APRR1/TOC1 quintet implicated in circadian rhythm, control of flowering time, and early photomorphogenesis. Plant Cell Physiol. 44 (2003) 1119-1130. doi:10.1093/pcp/pcg148.

[32] Y. Guo, Y. Li, H. Hong, L.J. Qiu, Establishment of the integrated applied core collection and its comparison with mini core collection in soybean (*Glycine max*). Crop J. 2 (2014): 38-45. doi:10.1016/j.cj.2013.11.001.

[33] T. Zheng, Y. Li, Y. Li, S. Zhang, C. Wang, F. Zhang, et al. SoyFGB v2. 0: a unique access to variations of Chinese Soybean Gene Bank (CNSGB)

germplasm. Sci. Bull. 67 (2022) 1716-1719. doi:10.1016/j.scib.2022.08.001.

[34] L. Wang, Y. Guan, R. Guan, Y. Li, Y. Ma, Z. Dong, et al. Establishment of Chinese soybean *Glycine max* core collections with agronomic traits and SSR markers. Euphytica 151 (2006) 215-223. doi:10.1007/s10681-006-9142-3.

[35] X. Liu, M. Huang, B. Fan, E.S. Buckler, Z. Zhang, Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet. 12 (2016) e1005767. doi:10.1371/journal.pgen.1005767.

[36] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Human Genet. 81 (2007) 559-575. doi:10.1086/519795.

[37] J. Doyle, DNA protocols for plants-CTAB total DNA isolation. Molecular techniques in taxonomy (1991).

[38] Y. Feng, S. Zhang, J. Li, R. Pei, L. Tian, J. Qi, et al. Dual-function C2H2-type zinc-finger transcription factor GmZFP7 contributes to isoflavone accumulation in soybean. New Phytol. 237 (2022) 1794-1809. doi:10.1111/nph.18610.

[39] Lyu, X. Cheng, Q. Qin, C. Y. Li, X. Xu, R. Ji, et al. GmCRY1s modulate gibberellin metabolism to regulate soybean shade avoidance in response to reduced blue light. Mol. Plant 14 (2021) 298-314. doi:10.1016/j.molp.2020.11.016.

[40] M. Paz, J. Martinez, A. Kalvig, T. Fonger, K. Wang, Improved cotyledonary node method using an alternative explant derived from mature seed for efficient Agrobacterium-mediated soybean transformation. Plant Cell Rep. 25 (2006) 206-213. doi:10.1007/s00299-005-0048-7.

[41] B. Hu, Z. Jiang, W. Wang, Y. Qiu, Z. Zhang, Y. Liu, et al. Nitrate-NRT1.1B-SPX4 cascade integrates nitrogen and phosphorus signalling networks in plants. Nat. Plants 5 (2019) 401-413. doi:10.1038/s41477-019-0384-1.

[42] S. Li, Z. Sun, Q. Sang, C. Qin, L. Kong, X. Huang, et al. Soybean reduced internode 1 determines internode length and improves grain yield at dense planting. Nat. Commun. 14 (2023) 7939. doi:10.1038/s41467-023-42991-z.

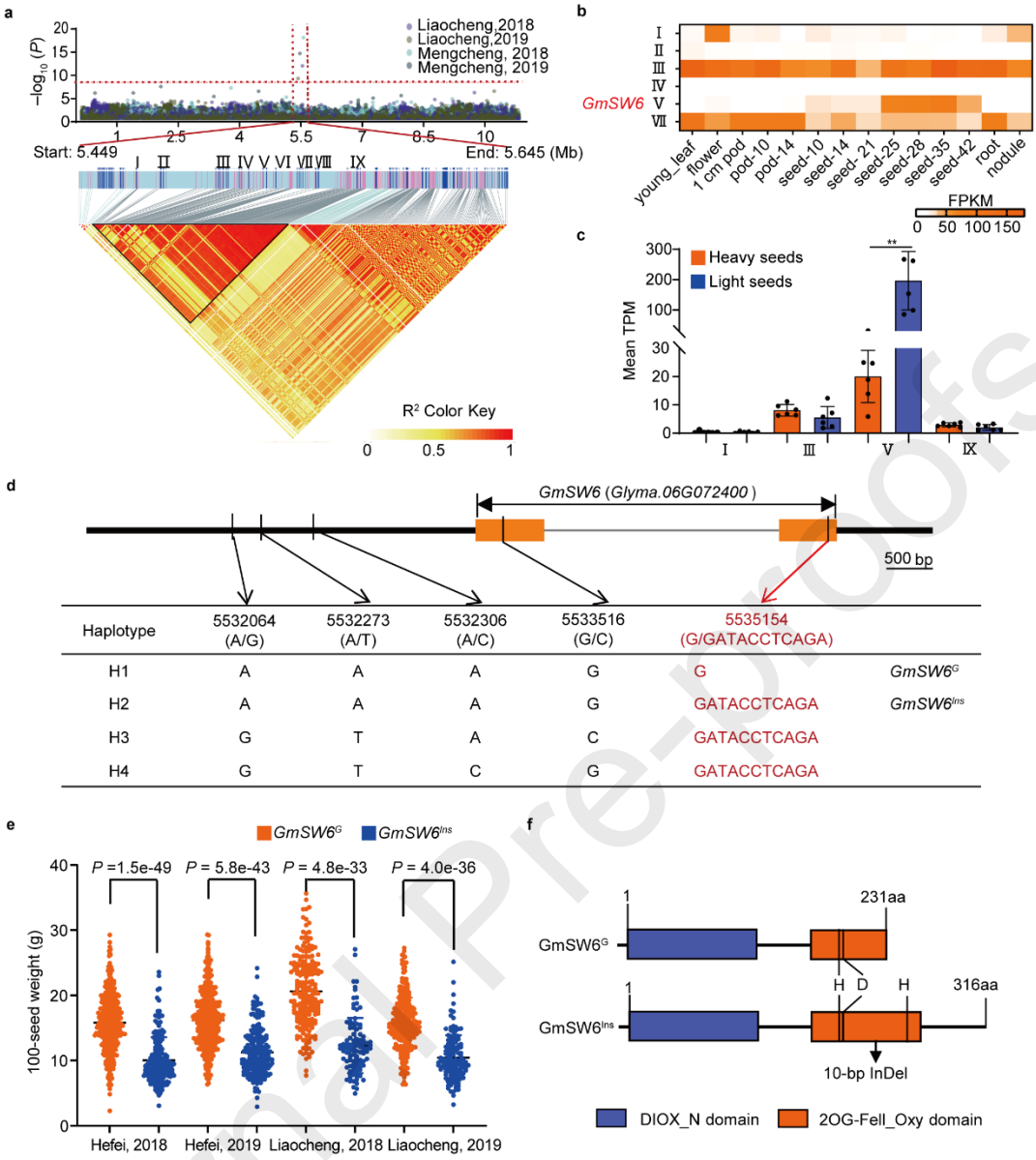


Figure 1 Identification of *GmSW6* as a candidate gene for 100-seed weight in soybean.

(a) Manhattan plot for 100-seed weight. The dashed line indicates the genome-wide significance threshold ($P = 1 \times 10^{-8}$, Bonferroni correction) determined by (top). LD heatmap (pairwise R^2) of the genomic region surrounding the lead SNP at the *SW6* locus (bottom). The physical positions of nine predicted genes (labeled I to IX) within the pick region are indicated. **(b)** Expression heatmap of candidate genes within the *SW6* region based on SoyBase RNA-seq datasets. The color gradient from white to red represents transcript abundance (FPKM, fragments per kilobase per million mapped reads). **(c)** Expression levels of four candidate genes in seeds at 55 DAF from six heavy- and six light-seeded varieties. Transcript abundance is shown as TPM (transcripts per million). **(d)** Haplotypes analysis of *GmSW6*. **(e)** 100-seed weight between accessions carrying the $GmSW6^{G}$ and $GmSW6^{Ins}$ (n = 422, 189, 447, 211, 207, 103, 316 and 137 accessions). **(f)** Protein structure of GmSW6 with and without the 10-bp insertion. The N-terminal DIOX-N domain and the C-terminal

20G-Fell-Oxy domain are highlighted in blue and orange, respectively. Statistical significance was determined by two-tailed Student's *t*-test.

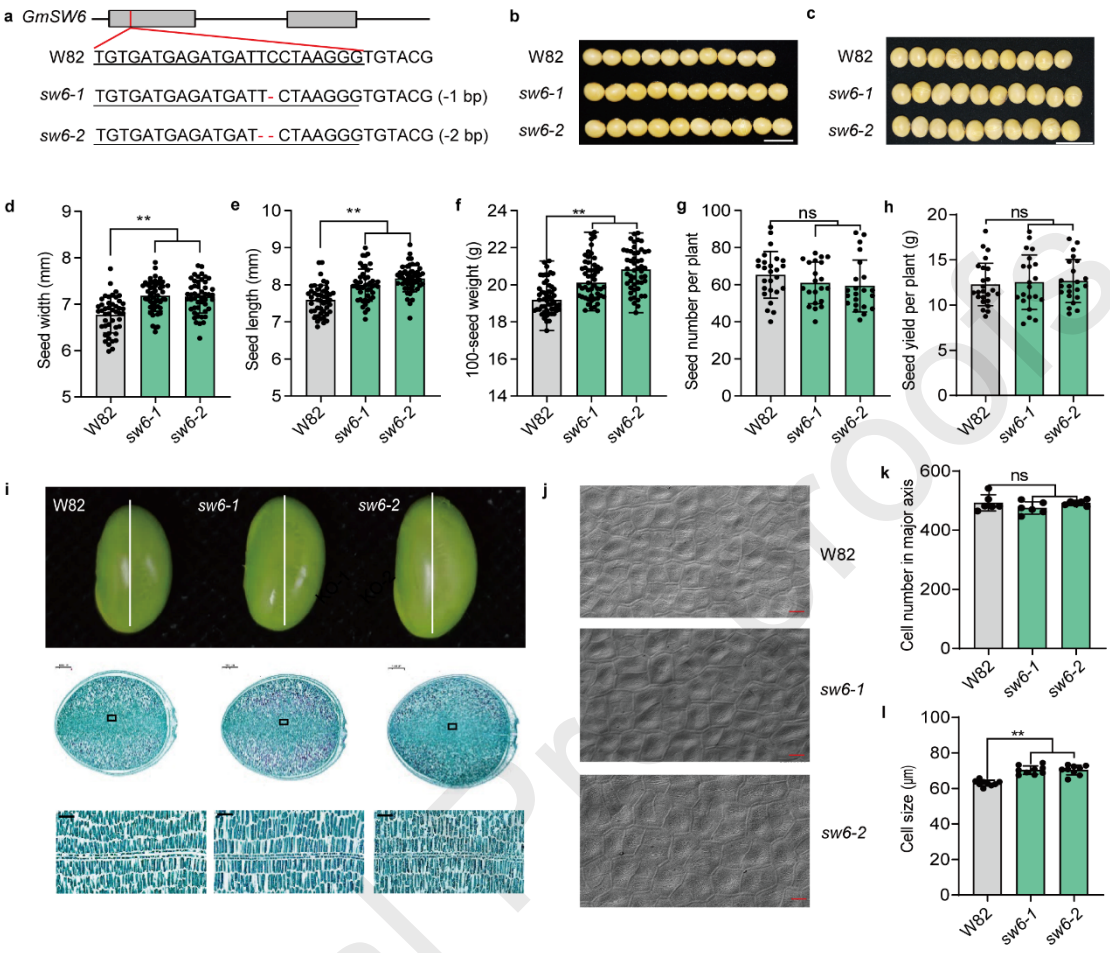


Figure 2 Functional validation of *GmSW6* demonstrates its negative regulation of 100-seed weight through cell expansion.

(a) CRISPR/Cas9-mediated knockout of *GmSW6* in the soybean Williams82 (W82) background. The sgRNA target sequence is underlined and nucleotide deletions are marked by red dashes. **(b, c)** Representative morphology of seed length (**b**) and seed width (**c**) in W82 and *sw6* mutant lines. Scale bar, 1 cm. **(d-f)** Seed width (d), seed length (e) and 100-seed weight (f) between W82 and *sw6* mutant lines (n = 46, 46, 50). **(g, h)** Seed number per plant (g) and seed yield per plant (h) between W82 and *sw6* mutant lines (n = 26, 21, 22). (**i**) Histological analysis of paraffin-sectioned seeds at 55 DAF from W82 and *sw6* mutant. Scale bar, 200 μm. (**j**) Scanning electron micrographs (SEM) showing the ventral surface of cotyledons in 55 DAF seeds of W82 and *sw6* mutants. Scale bar, 10 μm. (**k**) Cell number along the major axis in seeds of W82 and *sw6* mutant (n = 6). (**l**) Cell size in seeds of W82 and *sw6* mutant (n = 10). Data show means ± SD **, $P < 0.01$, ns, $P > 0.05$, Student's *t*-test.

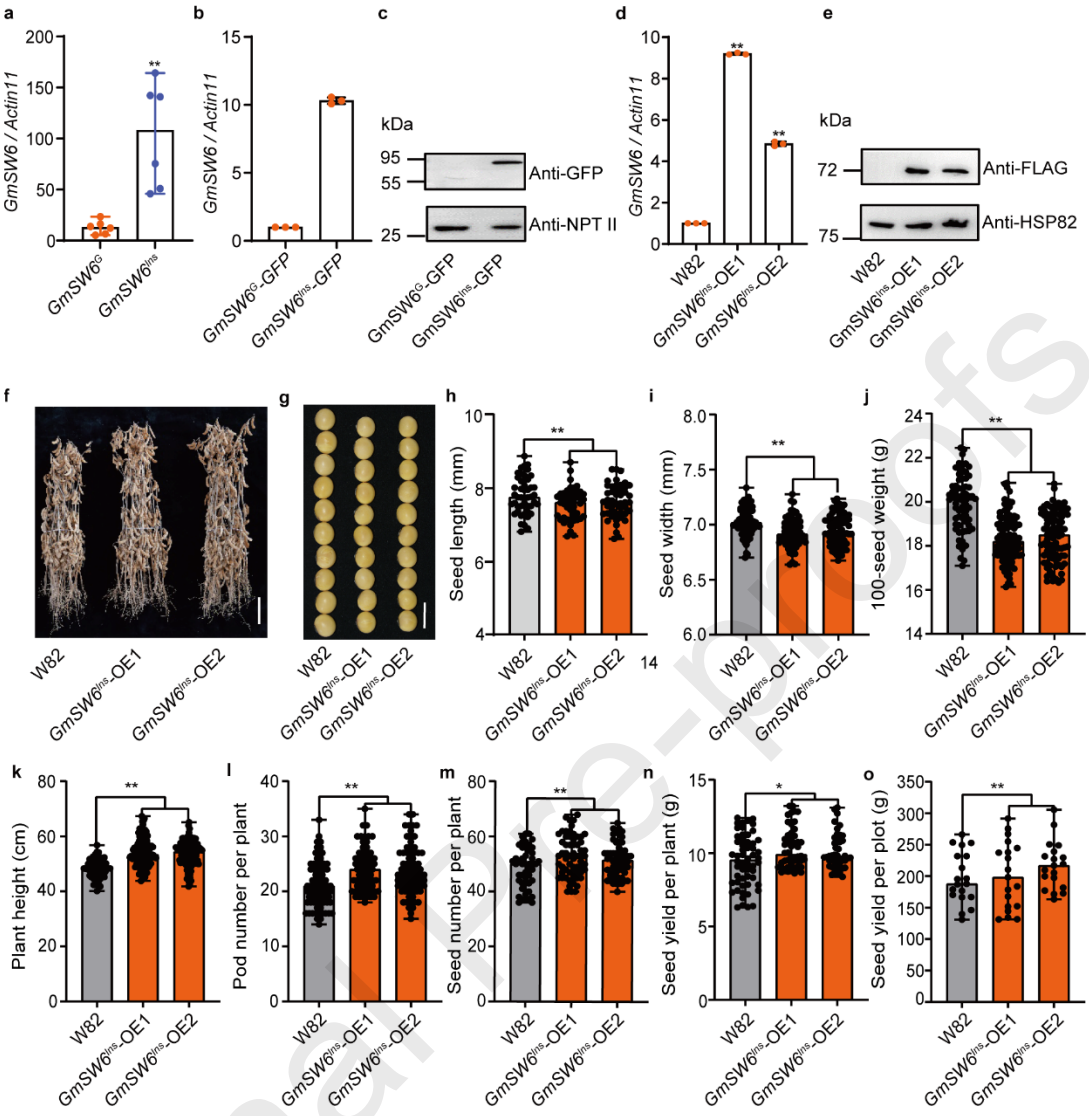


Figure 3 Genetic confirmation of *GmSW6^{Ins}*.

(a) Expression levels of *GmSW6* determined by RT-qPCR in seeds at 55 DAF from accessions carrying the *GmSW6^{G}* and *GmSW6^{Ins}* alleles ($n = 6$). **(b, c)** Ectopic expression analysis in *N. benthamiana* leaves. Transcript level **(b)** and protein level **(c)** of *GmSW6^{G}* and *GmSW6^{Ins}* were detected by RT-qPCR ($n = 3$) and immunoblotting, respectively. NPT-II was used as an internal control. **(d, e)** Molecular validation of overexpression (OE) lines. GmSW6 mRNA levels (d) and protein accumulation **(e)** in 55 DAF seeds of W82 and *GmSW6^{Ins}*-OE lines were analyzed via RT-qPCR ($n = 3$) and immunoblotting. HSP82 was used as an internal control. **(f, g)** Morphological characterization. Representative images of mature plants **(f)** and seed length **(g)** from W82 and *GmSW6^{Ins}*-OE lines. Scale bars, 10 cm in **(f)** and 1 cm in **(g)**. Evaluation of agronomic and yield-related traits. Comparative analysis of seed length (**h**, $n = 47$), seed width (**i**, $n = 69, 72, 72$), 100-seed weight (**j**, $n = 67, 96, 91$), plant height (**k**, $n = 75, 120, 138$), pod number (**l**, $n = 76, 64, 75$), seed number per plant (**m**, $n = 48, 62, 58$), seed yield per plant (**n**, $n = 56, 48, 46$), and seed yield per plot (**o**, $n = 19$) between W82 and *GmSW6^{Ins}*-OE lines. Data show means ± SD. * $P < 0.05$, ** $P < 0.01$, Student's *t*-test.

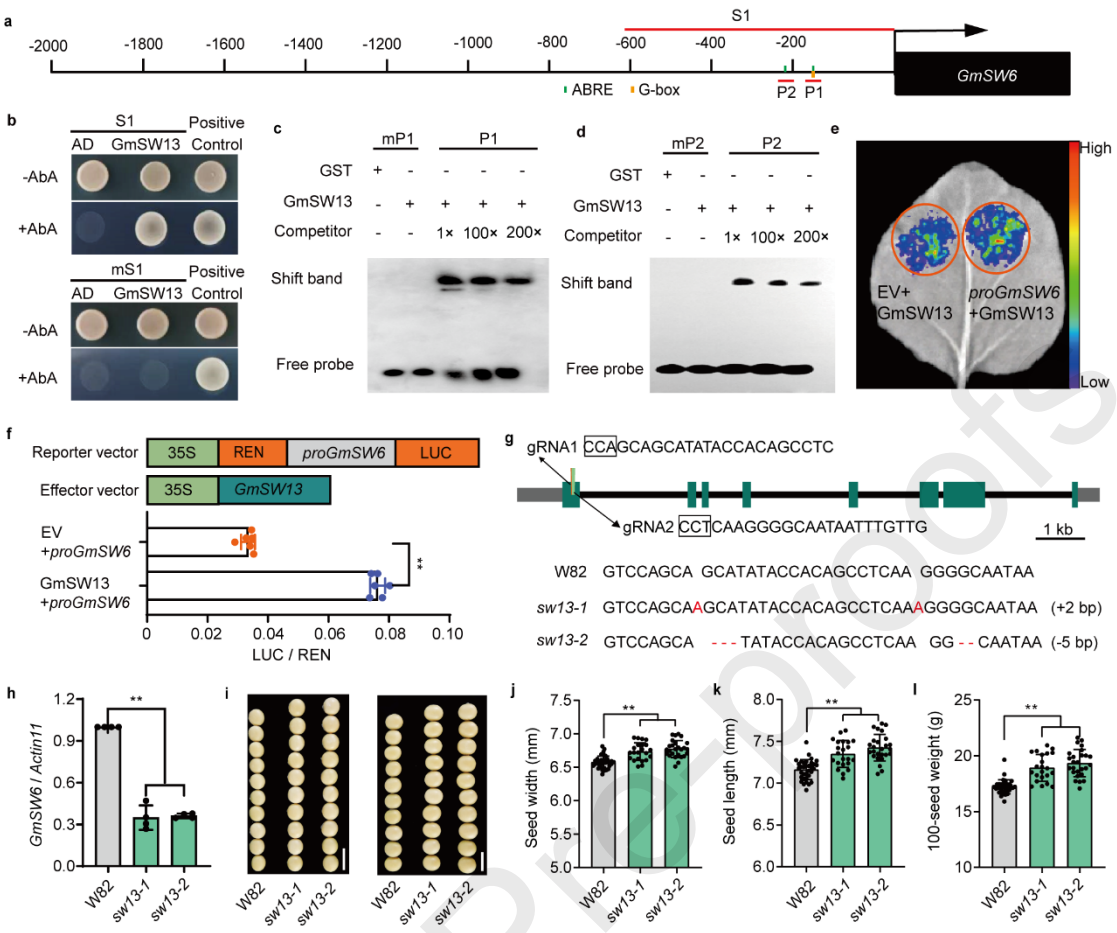


Figure 4 GmSW13 directly activates *GmSW6* expression to regulate soybean seed weight.

(a) Schematic diagram of the *GmSW6* promoter region showing the distribution of G-box (orange) and ABRE (green) *cis*-regulatory motifs. (**b**) Y1H assay demonstrating the binding of GmSW13 to the *GmSW6* promoter segment (S1). The interaction was abolished when the core ACGT sequence of the G-box and ABRE motifs was mutated to AAAA (mS1). (**c**, **d**) EMSA confirming the direct binding of GST-GmSW13 to the *GmSW6* promoter. Specific DNA-protein complexes were formed with biotin-labeled Probe 1 (P1) **(c)** and Probe 2 (P2) **(d)**. The binding was outcompeted by unlabeled cold probes and abolished by mutated probes (mP1 and mP2; ACGT mutated AAAA). (**e**, **f**) Transient Dual-LUC reporter assay in *N. benthamiana* leaves. The schematic of the constructs **(e)** and the relative luciferase activity **(f)** demonstrate that GmSW13 significantly transactivates the *GmSW6* promoter (n = 6). EV, empty vector. **(g)** Targeted mutagenesis of *GmSW13* using CRISPR/Cas9. The genomic structure of the *GmSW13* locus and the two sgRNA target sites are shown. Red letters and dashes indicate nucleotide insertions or deletions in the two representative mutant alleles. **(h)** RT-qPCR analysis of *GmSW6* expression in seeds at 55 DAF from W82 and *sw13* mutants (n = 4). **(i)** Seed length (left) and seed width (right) of the wild type Williams 82 (W82) and *sw13* lines. Scale bar, 1 cm. (**j-l)** seed width (**j**), Seed length (**k**) and 100-seed weight (**l**) for W82 and *sw13* lines (n = 34, 22 and 26). Data show means ± SD. ** $P < 0.01$, Student's t-test.

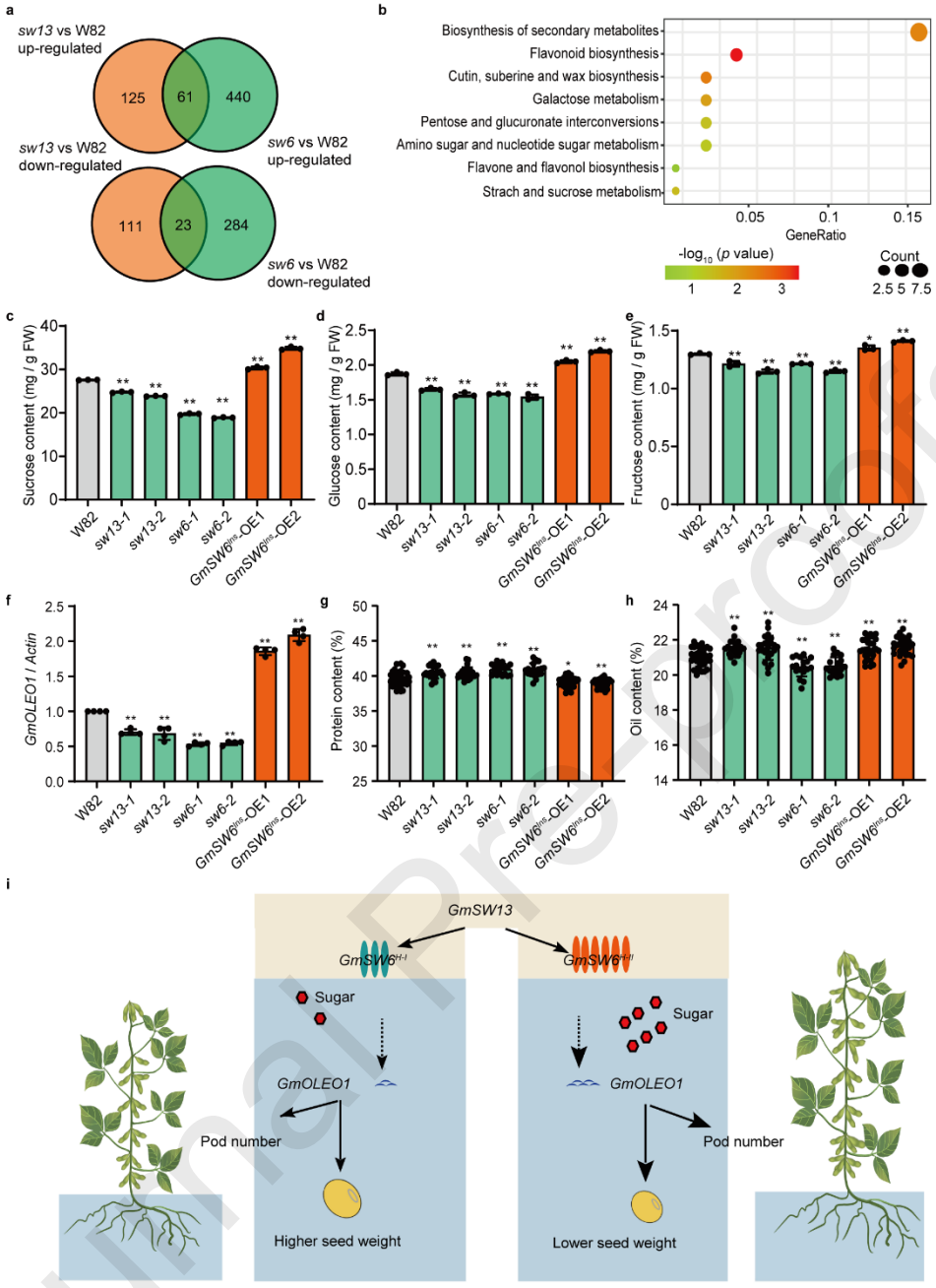


Figure 5 Isolation of GmSW13-GmSW6 downstream genes.

(a) Venn diagram illustrating the overlap of differentially expressed genes (DEGs) that are co-up-regulated or co-down-regulated in both *sw13* and *sw6* mutant seeds compared to W82. KEGG enrichment analysis identifying the primary metabolic and signaling pathways associated with the overlapping DEGs between *sw13* and *sw6*. (**c-e**) Sucrose content (**c**), glucose content (**d**) and fructose content (**e**) in 55 DAF seed between Wm82, *sw13*, *sw6*, and *GmSW6*^{Ins}-OE lines (n = 3). (**f**) Relative expression levels of *GmOLEO1* in 55 DAF seeds of Wm82, *sw13*, *sw6*, and *GmSW6*^{Ins}-OE lines (n = 4). (**g, h**) Protein content (**g**) and oil content (**h**) in Wm82, *sw13*, *sw6*, and *GmSW6*^{Ins}-OE lines (n = 34, 24, 24, 18, 18, 27 and 27). Data show means ± SD. * $P < 0.05$, ** $P < 0.01$, Student's t-test. (**i**) Proposed working model of the GmSW13–GmSW6 regulatory cascade. The

model illustrates how this hierarchical module coordinates 100-seed weight and quality by modulating *GmOLEO1*.

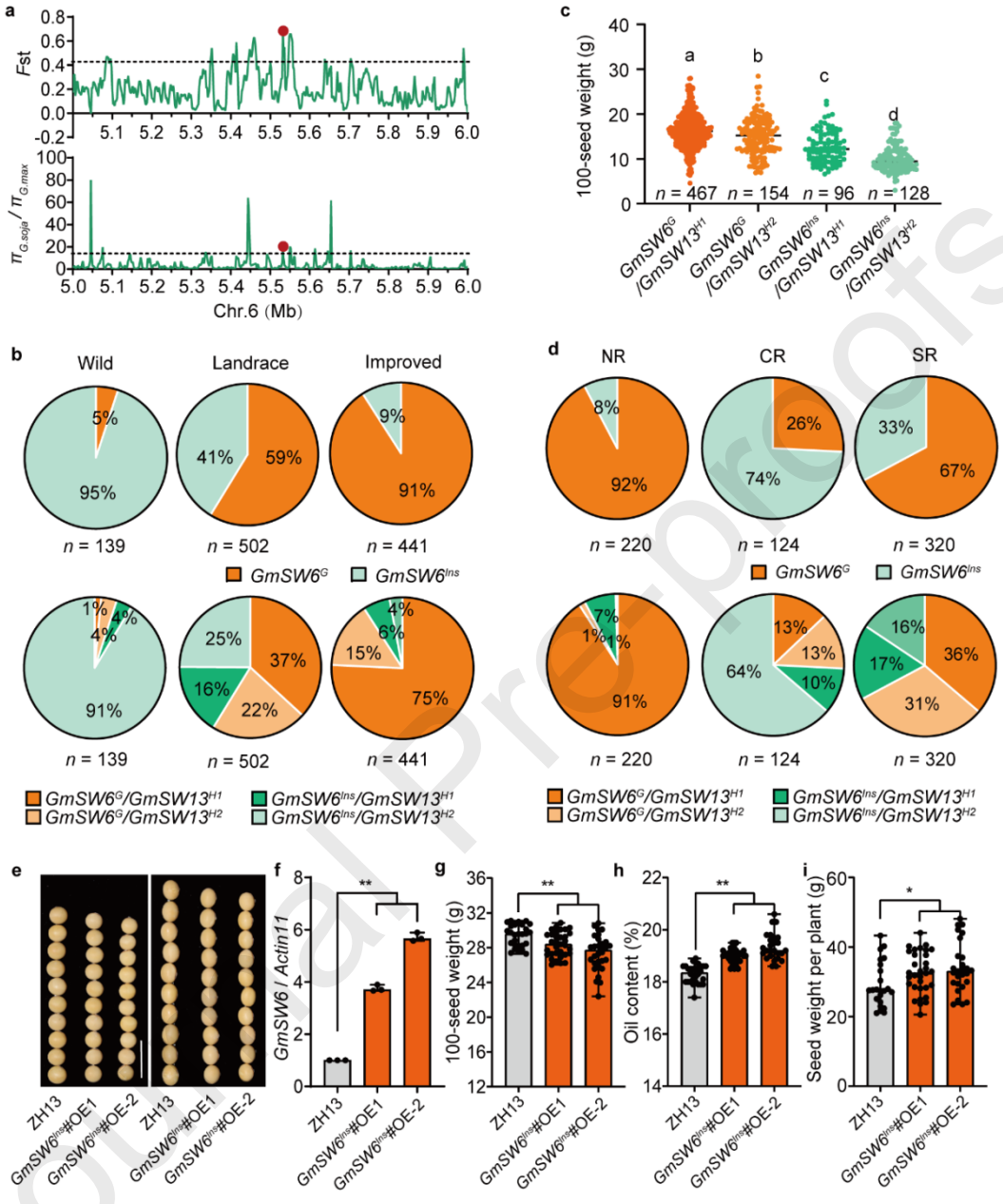


Figure 6 Evolutionary selection and breeding application of the GmSW13–GmSW6 module.

(a) Selection of *GmSW6* during soybean domestication. *F*st and π values between *G. soja* and cultivated soybean (*G. max*) are shown across a 1-Mb genomic region flanking the *GmSW6* locus. **(b)** Allelic frequency shifts of the two GmSW6 alleles and their combinations across wild soybean, landraces, and improved cultivars, illustrating the evolutionary trajectory of the locus. (**c**) Synergistic effects of haplotype combinations on 100-seed weight. **(d)** Distribution patterns of haplotype combinations in various cultivars. Different lowercase letters indicate significant differences ($P < 0.05$) as determined by a two-sided one-way ANOVA. (**e**) Morphological validation in Zhonghuang 13 (ZH13). Representative images of seed width (left) and seed length (right) in

ZH13 and two independent *GmSW6^{Ins}*#OE lines. Scale bar, 1 cm. (**f**) Expression levels of *GmSW6* in 55 DAF seeds of ZH13 and *GmSW6^{Ins}*#OE lines ($n = 3$). (**g-i)** 100-seed weight (**g**), oil content (**h**) and seed weight per plant (**i**) for ZH13 and *GmSW6^{Ins}*#OE lines (n = 34, 22 and 26). Data show means ± SD. * $P < 0.05$, ** $P < 0.01$, Student's *t*-test.

Supplementary materials

Figure S1 Correlation analysis of 100-seed weight across four environments.

Figure S2 Characterization and GWAS for 100-seed weight in a soybean diversity panel.

Figure S3 Tissue-specific expression and haplotypes analysis of *GmSW6*.

Figure S4 Multiple sequence alignment of GmSW6 and its orthologs.

Figure S5 Protein structure and subcellular localization of GmSW6.

Figure S6 KASP genotyping validated a 10-bp deletion specific to the *GmSW6* haplotypes.

Figure S7 Generation of *GmSW6* knockout lines using CRISPR/Cas9.

Figure S8 Haplotype-specific expression and promoter activity of *GmSW6*.

Figure S9 Sequence comparison of GmSW13 in W82 and *sw13* mutants.

Figure S10 Identification of differentially expressed genes (DEGs) by RNA-seq.

Figure S11 Haplotypes analysis of *GmSW13.*

Figure S12 Haplotypes analysis of *GmOLEO1*.

Figure S13 Phenotypic characterization of *GmSW6^{Ins}*#OE lines.

Table S1 Information on 1,702 soybean accessions used for GWAS.

Table S2 Descriptive statistics, ANOVA results and h2 of the 100-seed weight of natural population.

Table S3 11 QTLs associated with 100-seed weight in landraces and improved cultivars measured at Mengcheng and Liaocheng locations in 2018 and 2019.

Table S4 Various located in the predicted gene regions of the *SW6*.

Table S5 Transcriptional profiling of candidate genes across 12 soybean germplasms via RNA-seq analysis.

Table S6 RNA-Seq analysis reveals haplotype-specific expression patterns of *GmSW6*.

Table S7 Potential upstream regulatory genes of GmSW6 screened by yeast one-hybrid.

Table S8 Prediction of cis-regulatory elements in the *GmSW6* promoter.

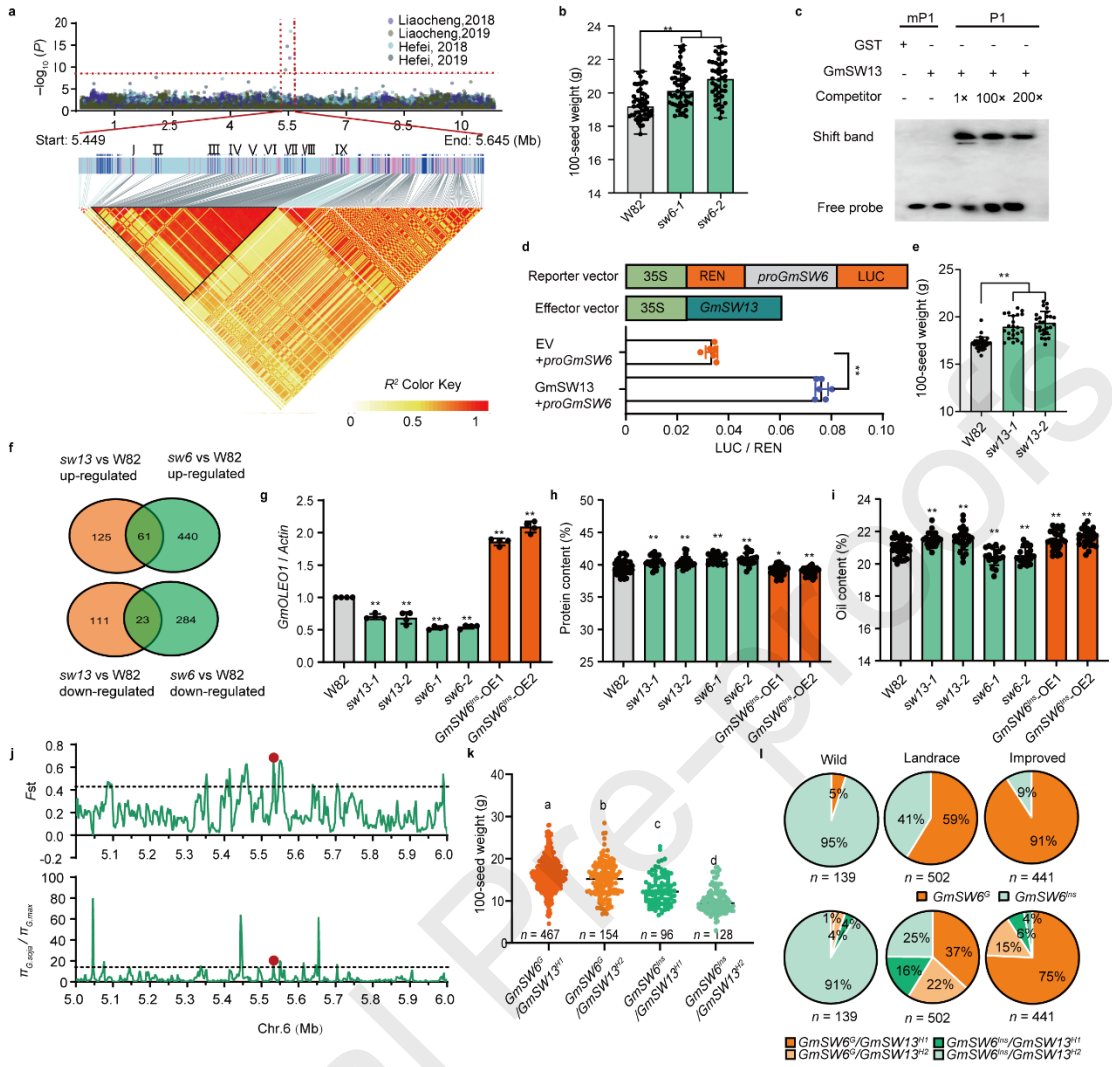
Table S9 DEGs annotation and corresponding homologous gene in Arabidopsis.

Table S10 The gene overlap of the up-regulation of *sw13* and *sw6* or the down-regulation of *sw13* and *sw6*.

Table S11 The information of primers used in this study.

Research Highlights

1. **Gene Discovery:** Identified a novel seed weight gene, *GmSW6*, through genome-wide association studies (GWAS) using a diverse panel of 1,702 soybean accessions.
2. **Functional Variation:** Pinpointed the functional variation (G/Ins) in *GmSW6* and demonstrated that the beneficial allele (*GmSW6*^{G}) was increasingly selected for during domestication and subsequent cultivar improvement.
3. **Novel Pathway:** Uncovered the GmSW13-GmSW6 regulatory module, which represents the first known seed weight and quality regulatory pathway involving non-gibberellin 2-oxoglutarate-dependent dioxygenase proteins in soybean.
4. **Practical Application:** Demonstrated that pyramiding the beneficial alleles of both *GmSW13* and *GmSW6* can significantly enhance seed weight.



Compliance with Ethics Requirements

Ethical approval was not required for this study.