Nature Communications

https://doi.org/10.1038/s41467-026-72438-0

Article in Press

An underutilized genetic component in the regulation of soybean seed weight

Received: 23 August 2024

Accepted: 16 April 2026

Cite this article as: Zhang, H., Tian, Y., Zhao, C. *et al.* An underutilized genetic component in the regulation of soybean seed weight. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-72438-0

Hao Zhang, Yu Tian, Chaosen Zhao, Zhihui Shan, Lei Yang, Yongzhe Gu, Huawei Gao, Zhangxiong Liu, Delin Li, Tianli Ge, Wenchao Yin, Zihao Zheng, Ying–hui Li, Hongning Tong & Li–juan Qiu

An underutilized genetic component in the regulation of soybean seed weight

Hao Zhang[1§], Yu Tian[1§], Chaosen Zhao[2], Zhihui Shan[3], Lei Yang[1], Yongzhe Gu[1], Huawei Gao[1,4], Zhangxiong Liu[1], Delin Li[1], Tianli Ge[1], Wenchao Yin[1], Zihao Zheng[5], Ying–hui Li[1,4,*], Hongning Tong[1,4,*] & 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), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[2]Crops Research Institute of Jiangxi Academy of Agricultural Sciences, Nanchang, Jiangxi, 330200, China.

[3]Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, Hubei, 430062, China.

[4]National Nanfan Research Institute, Chinese Academy of Agricultural Sciences, Sanya, Hainan, 572024, China.

[5]Department of Agronomy, Iowa State University, Ames, IA 50011–1051, USA.

[§]These authors contributed equally to this work.

*Corresponding authors: Ying–hui Li (liyinghui@caas.cn), Hongning Tong (tonghongning@caas.cn), and Li–juan Qiu (qiulijuan@caas.cn)

Abstract

Seed weight is a key determinant of soybean yield; however, its underlying genetic regulation remains poorly understood. Here, besides the previously reported positive seed weight regulator GmST05/GmSW5, which encodes a protein homologue of Arabidopsis Mother of TFL1 and FT (MFT), we identify *GmSW19* that exert negative effect on seed weight. *GmSW19* belongs to the basic leucine zipper (bZIP) transcription factor family; it encoded bZIP transcription factor can represses *GmSW5* expression. We further discover that the glycogen synthase kinase 3 (GSK3)-like kinase GmSK21 physically interacts with and phosphorylates GmSW19. A single nucleotide polymorphism (A to C), resulting in a T to P substitution in GmSW19, which alters its phosphorylation by GmSK21 and consequently its protein abundance. Functional analyses reveal that the $GmSW19^{A}$ variant represses *GmSW5* and seed weight stronger than that of $GmSW19^{C}$. Population analyses show that the heavy–seed allele $GmSW19^{C}$ has not been fully utilized in modern soybean breeding. These findings elucidate the genetic components and their possible interaction in determining seed weight and highlight the potential for enhancing yield in soybean.

Introduction

As one of the most successfully domesticated legumes, soybean (*Glycine max*) provides 70.7% of the protein meal and 28.6% of the vegetable oil consumed globally (http://soystats.com), underscoring its central role in food and nutritional security. Enhancing soybean yield is therefore essential, particularly for improving quality of life in developing regions. Seed weight—largely determined by seed size and volume—is a major contributor to overall seed yield. In soybean, seed weight typically has a negative correlation with oil content but a positive correlation with protein levels[1,2]. Natural germplasm collections exhibit extensive and continuous variations in seed weight, reflecting their quantitative nature and complex genetic basis[3,4]. Seed weight is controlled by hundreds of quantitative trait loci (QTLs for seed weight, qSW), yet only a few causal genes underlying these loci have been cloned in recent years, including *GmST05* (*seed thickness 05*; underlying *qSW10–1*)[2,5], *GmSWEET10a* (a sugar transporter gene responsible for *qSW15*)[6], and *GmGA3ox1* (*gibberellin 3β–hydroxylase 1*; underlying *qSW7–1*)[7]. GmST05, the soybean orthologue of the Arabidopsis (*Arabidopsis thaliana*) MFT (mother of TFL1 and FT), positively regulates seed weight, possibly by modulating the expression of *GmSWEET10a*[2]. GmSWEET10a and its paralogue GmSWEET10b, both of which are members of the SWEET sugar transporter family, appear to promote seed weight through effects on sugar allocation between the seed coat and the embryo[6]. GmGA30ox1 increases seed weight but has an opposing effect on seed number[7]. Despite these advances, most causal genes controlling seed weight remain unidentified, and the regulatory networks shaping seed–size variation in soybean are still far from fully understood.

In the model plants Arabidopsis and rice (*Oryza sativa*), many seed– and grain–size genes have been cloned, enabling the construction of several conserved regulatory pathways, including phytohormone signaling, ubiquitination, G–protein signaling, and mitogen–activated protein kinase pathways[8]. Among the phytohormones implicated, brassinosteroids (BRs) appear to play a particularly dominant role in controlling seed size. BRs act primarily through destabilizing a subclade of GSK3/SHAGGY–like kinases, exemplified by BR INSENSITIVE 2 (BIN2) in Arabidopsis and GSK2 in rice[9,10]. These kinases phosphorylate a broad spectrum of substrates, many of which are transcriptional

regulators, to modulate diverse BR responses, including seed development. BR signaling extensively cross−talks with other hormone pathways such as abscisic acid (ABA). For example, BIN2 interacts with and phosphorylates ABA INSENSITIVE5 (ABI5), thereby contributing to BR−ABA crosstalk[9]. As a key transcription factor in ABA signaling, ABI5 regulates multiple ABA−dependent processes, including seed germination and development. ABI5 represses *SHB1* (*SHORT HYPOCOTYL UNDER BLUE1*) transcription to influence seed size and forms a regulatory loop with MFT to modulate seed germination[11,12]. In soybean, several seed–weight genes, such as *GmKIX8–1*[13] and four *GmCYP78A* family genes[14–16], have been identified based on their homology to their Arabidopsis counterparts. In addition, *protein phosphatase 2C* (*PP2C*) has been shown to regulate soybean seed weight by modulating BR signaling[17]. These findings indicate that core seed−size regulatory pathways may be conserved across species. However, despite these advances, the genetic pathways and regulatory networks governing seed weight in soybean remain largely unresolved.

In this work, using a genome–wide association study (GWAS), we identify *GmSW19*, a bZIP transcription factor that negatively regulates seed weight by repressing the expression of the positive regulator *GmST05*/*GmSW5*[2]. Moreover, we demonstrate that GmSK21 interacts with and phosphorylates GmSW19 to stabilize the protein, which in turn targets *GmSW5* to repress its transcription. Importantly, we reveal functional natural variation in *GmSW19* located at one of its phosphorylation sites, leading to differential protein stability and consequently alter regulation of *GmSW5* expression and seed weight. We reveal that the heavy–seed allele of *GmSW19*^{C} remains underutilized in soybean breeding. These findings highlight the substantial potential of *GmSW19* for increasing soybean yield through targeted breeding.

Results

Identification of soybean seed weight loci by GWAS across three environments

We performed a GWAS for 100–seed weight using a panel of 1,501 cultivated soybean accessions, including 938 landraces and 563 improved cultivars, selected from the Chinese primary core collection and the 100–seed weight–based applied core collection (Fig. 1a,

Supplementary Fig. 1 and Supplementary Data 1)[18–20]. This panel captures the broad genetic diversity preserved in the Chinese National GeneBank and has been previously resequenced[21]. Substantial variation in 100–seed weight was observed across the three environments, Sanya (18.2°N, 109.9°E), Wuhan (30.5°N, 114.3°E), and Nanchang (28.3°N, 116.1°E), ranging from 3.0−50.2 g (Fig. 1b). Seed weight measurements were strongly positively correlated among the three environments, with pairwise correlation coefficients (r) ranging from 0.81−0.87 ($P < 0.001$) (Fig. 1b). The broad–sense heritability for 100–seed weight across environments was estimated at 0.86 (Supplementary Table 1), indicating that genetic factors predominantly govern seed–weight variation, while individual accessions display moderate phenotypic plasticity in response to environmental conditions. These phenotypic data provide a robust foundation for identifying candidate genes associated with 100–seed weight in soybean.

We analyzed 6.45 million SNPs with minor allele frequency (MAF) ≥ 0.05 across the 1,501 soybean genomes in our diversity panel[21]. A GWAS for 100–seed weight was conducted in each environment using a linear mixed model in FarmCPU, accounting for population structure and kinship[22]. We consistently identified 21 trait–associated SNPs (TASs) exceeding $-\log_{10}(P \text{ value}) = 8$ across all environments, defining 21 stable genomic regions linked to 100–seed weight (Supplementary Data 2). Approximately two–thirds of these regions overlapped with previously reported QTLs for 100–seed weight (Supplementary Data 2). For example, the SNP at Chr15:3,867,098 falls within the intervals of *qSW6–1*[23] and *qSW33–2*[24] and is adjacent to the known seed–weight gene *GmSWEET10a*[6]. These findings confirm the robustness of our GWAS and pinpoint key genomic regions for candidate genes discovery.

Seed–specific expression of *GmSW19* decreases seed weight

We focused on *qSW19* on chromosome 19 because the lead SNP (Chr19:45120203) showed the strongest association with 100–seed weight in Sanya and strong signals in Wuhan and Nanchang (Fig. 1a). Linkage disequilibrium (LD) analysis revealed that the candidate region was ~82.6 kb, spanning 45.11–45.20 Mb (Fig. 1c). This region contains 11 open reading frames, eight of which harbor nucleotide differences in exons or

promoters, while three have intronic variants (Supplementary Data 2). Haplotype analysis revealed that five SNPs across four genes (*Glyma.19G193600*, *Glyma.19G193900*, *Glyma.19G194300* and *Glyma.19G194500*) were significantly associated with 100–seed weight (Supplementary Fig. 2 and Supplementary Table 2). *Glyma.19G194300* (*Dt1*), an orthologue of the *TFL1* gene in Arabidopsis, was excluded because *Dt1* and its mutants exhibit no significant difference in 100–seed weight under short–day conditions[25,26]. Among the remaining candidates, *Glyma.19G194500* was the only gene with seed–specific expression, peaking at the R7 stage (when the first pod on the main stem shows mature color)[27], whereas the other genes were expressed at minimal levels or were not expressed (Fig. 1d, e). This gene encodes a protein orthologous to Arabidopsis ABI5, a known negative regulator of seed size[11]. These observations indicate that *Glyma.19G194500* is the most likely causal gene underlying *qSW19*. Given the genes is discovered from seed weight QTL cloning, we designated it *GmSW19* for clarity and consistence of the other genes reported in this study.

To identify potential functional polymorphisms, we examined the genetic variation of *GmSW19* across 1,501 soybean germplasms. Although six polymorphisms were detected in the promoter region, expression analysis revealed comparable *GmSW19* levels at the R7 stage in accessions with contrasting seed weights (Supplementary Fig. 3 and Supplementary Table 3), suggesting that promoter variation is unlikely to underlie the functional differences. In the coding region, a single nonsynonymous SNP at Chr19:45,204,441 (A/C) in the first exon was identified, resulting in a threonine–to–proline substitution at position 175 (T175P) (Fig. 1f). Compared with those carrying the A allele ($GmSW19^{A}$), accessions carrying the Chr19:45,204,441–C allele ($GmSW19^{C}$) had significantly heavier seeds across all three environments ($P < 0.001$, Fig. 1g). These findings suggest that this SNP is likely responsible for the functional variation detected by the GWAS.

Functional validation of the role of *GmSW19* in seed weight regulation

To verify the function of *GmSW19*, we generated knockout mutants in the wild–type (WT) cultivar Williams 82 using CRISPR/Cas9 targeting a sequence near the 5’ end of the coding

region. Three homozygous mutants, *sw19–1*, *sw19–2*, and *sw19–3*, carried frameshift mutations caused by a 37–bp deletion, a 1–bp insertion, or a 5–bp deletion, respectively, and were therefore considered null alleles (Fig. 2a). Compared with the WT plants, all the mutants presented consistently larger, longer, and heavier seeds, with increases of ~4.4–6.7%, ~6.0–7.1%, and 9.1–12.2%, respectively (Fig. 2b–e). Seed weight per plant was significantly enhanced by ~5.5–11.7%, primarily due to larger seed cells rather than changes in cell number or density, as indicated by cytological analysis (Fig. 2f–i and Supplementary Fig. 4a-c). Importantly, no significant differences in plant height, branch number, seed number, or pod number were detected between mutants and WT (Supplementary Fig. 4d–g). Consequently, compared with that of the WT, seed yield per plot increased by ~9.5–14.1% (Fig. 2j). These results demonstrate that *GmSW19* functions as a negative regulator of soybean seed size and weight and represents a promising target for yield improvement.

The soybean genome contains two *GmSW19* homologues, *Glyma.13G153200* and *Glyma.10G071700*, whose proteins share 73.73% and 75.45% sequence similarity with GmSW19, respectively (Supplementary Fig. 5a, b). To rule out potential off–target effects, we sequenced the corresponding target regions in both genes across all *sw19* mutants and detected no mutations (Supplementary Fig. 5c). These results confirm that the observed phenotypic changes are attributable to loss of *GmSW19* function.

GmSW19^{A} and GmSW19^{C} differ in their ability to suppress seed weight

To determine whether the identified SNP affects *GmSW19* function, we generated transgenic Williams 82 plants overexpressing each allele under the cauliflower mosaic virus (CaMV) 35S promoter (Fig. 3a). For each construct, two independent lines (*GmSW19^{A}*–OE1/OE2 and *GmSW19^{C}*–OE1/OE2) with comparable transcript levels were selected for analysis (Fig. 3b). Compared with those of the WT plants, the seed sizes and weights of all transgenic plants were reduced, but the seeds of *GmSW19^{A}*–OE lines were consistently smaller than those of the *GmSW19^{C}*–OE lines, whereas the plant height and pod number remained unchanged (Fig. 3c, d and Supplementary Fig. 6). These results

indicate that *GmSW19* negatively regulates seed weight, with $GmSW19^{A}$ having a stronger repressive effect than that of $GmSW19^{C}$.

To explore the impact of the T175P substitution, we compared the subcellular localization of the two protein variants by transiently expressing GFP fusions in *Nicotiana benthamiana* leaves. Both variants localized to the nucleus, as shown by confocal microscopy (Supplementary Fig. 7). The protein structures predicted using ColabFold v1.5.3 (AlphaFold2 with MMseqs2) revealed no spatial differences (Supplementary Fig. 8). However, immunoblot analysis using an anti–FLAG antibody revealed that compared with $GmSW19^{C}$–FLAG, $GmSW19^{A}$–FLAG accumulate at higher levels, despite the detection of similar transcript levels across the four transgenic lines (Fig. 3e), suggesting that the T175P substitution affects protein stability.

GmSK21 interacts with GmSW19

To identify protein partners of $GmSW19^{A}$, we performed a yeast two–hybrid (Y2H) screen using a soybean cDNA library and identified 61 unique clones corresponding to 20 genes (Supplementary Table 4). Among these genes, Glyma.04G063600, a subclade II GSK3–like kinase implicated in BR signaling[28], stood out due to its homology to BIN2/AtSK21 and rice GSK2, which regulate seed size in various species[10,29,30]. We designated this gene *GmSK21*. Subcellular localization revealed that $GmSW19^{A}$ and GmSK21 were colocalized in the nucleus of *N. benthamiana* cells (Supplementary Fig. 9a). Direct interaction was confirmed by point–to–point Y2H, bimolecular fluorescence complementation (BiFC), and luciferase complementation imaging (LCI) assays (Supplementary Fig. 9b–d).

To assess its role in determining seed weight, we generated *GmSK21* knockout mutants in Williams 82. However, no significant changes in seed size were observed, likely due to functional redundancy with homologues sharing >84% sequence similarity, as reported in rice and Arabidopsis (Supplementary Fig. 10, 11a, d–g)[18,31]. We further generated double mutants of *GmSK21* and its closest homolog *GmSK21a* (*Glyma.06G064800*) and found that *sk21 sk21a* mutants presented a significant increase in seed weight (Fig. 3f and Supplementary Fig. 10, 11b, c). Moreover, we generated overexpression lines to investigate

the function of GmSK21. Previous studies have indicated that overexpressing wild–type BIN2 or GSK2 often results in no phenotype due to protein instability, whereas gain–of–function variants (e.g., bin2–D or aGSK2) are stabilized and produce clear effects[32,33]. Accordingly, we overexpressed either wild–type *GmSK21* or a gain–of–function mutant (*GmSK21*^{G787A}, corresponding to rice *GSK2–2*[10]) under the 35S promoter in the soybean cultivars Huachun6 and Williams 82. Whereas wild–type *GmSK21* overexpression caused no phenotypic change, *GmSK21*^{G787A}–OE plants displayed significantly reduced 100–seed weight (Fig. 3g, h and Supplementary Fig. 11h–m). Taken together, these results suggest that GmSK21 likely functions redundantly with its homologues, such as GmSK21a, to inhibit seed weight.

GmSK21 phosphorylates at T175 to stabilize GmSW19

Given that BIN2 phosphorylates ABI5, we hypothesized that GmSK21 could phosphorylate GmSW19, with the T175>P substitution potentially influencing phosphorylation and protein stability. To test this hypothesis, we expressed and purified GST–GmSK21, GmSW19^{A}–His, and GmSW19^{C}–His and performed *in vitro* kinase assays. Using an antiphosphothreonine (anti–pT) antibody, we found that GST–GmSK21 phosphorylated both GmSW19 variants, but the threonine phosphorylation signal was significantly stronger on GmSW19^{A}–His than on GmSW19^{C}–His (Fig. 3i).

GmSW19 contains 14 threonine residues, including T175 (T41, T47, T86, T95, T97, T121, T147, T155, T160, T174, T175, T255, T325 and T367) (Supplementary Fig. 12). To confirm the T175 was the major phosphorylation site, we mutated all the other threonine residues to proline, generating GmSW19^{A}–T175 (only T175 was retained) and GmSW19^{C}–P175 (no threonine). Kinase assays revealed that GmSW19^{A}–T175 retained strong phosphorylation, whereas GmSW19^{C}–P175 showed no detectable threonine phosphorylation (Fig. 3i), indicating that T175 is the primary phosphorylation site.

To determine whether T175 affects GmSW19–GmSK21 interaction, we performed Y2H and co–immunoprecipitation (co–IP) assays. Compared with GmSW19^{C}, GmSW19^{A} exhibited stronger binding to GmSK21 (Fig. 3j). Co-IP experiments revealed that co-

expression of GmSK21-GFP with GmSW19^{A}-FLAG or GmSW19^{C}-FLAG revealed much higher levels of GmSW19^{A}-FLAG pulled down compared to GmSW19^{C}–FLAG, whereas negative controls using GFP alone yielded no signal (Fig. 3k). Finally, real–time quantitative analysis (RT–qPCR) revealed that the loss of *GmSW19* in *sw19* mutants did not significantly alter *GmSK21* transcription (Supplementary Fig. 13), suggesting that the T175P variation primarily affects GmSW19 stability through differential phosphorylation and binding affinity rather than transcriptional regulation.

GmSW19 directly suppresses *GmSW5* transcription to regulate seed weight

To identify downstream targets of GmSW19, we performed RNA–sequencing (RNA–seq) on R7–stage seeds from WT and *sw19* mutants. Across the two alleles, we identified 643 overlapping differentially expressed genes (DEGs) with $|\log_2(\text{fold change})| \geq 1.5$ and $P < 1e^{-3}$, including 360 upregulated and 283 downregulated genes (Supplementary Fig. 14a and Supplementary Data 3). Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis revealed enrichment in fatty acid biosynthesis and metabolism, carbon metabolism, amino acid biosynthesis, and phytohormone signal transduction (Supplementary Fig. 14b). Among the DEGs, 10 had Arabidopsis homologues known to influence seed size or oil content, including *Glyma.05G244100*, a phosphatidylethanolamine–binding protein orthologous to Arabidopsis MFT[30] (Supplementary Fig. 14c).

A strong association peak (SNP Chr05:41,853,526) was detected in the third intron of *Glyma.05G244100*, which we designated *GmSW5* (Fig. 1a). Haplotype analysis identified four haplotypes (H1–H4) based on six polymorphisms, among which H4 was associated with lower 100–seed weight. Accordingly, we grouped the haplotypes into GmSW5$^{H–I}$ (H1–H3) and GmSW5$^{H–II}$ (H4) (Supplementary Fig. 15a, b). Overall, the transcriptome data revealed that compared with GmSW5^{H-II}, GmSW5$^{H–I}$ expressed higher levels of *GmSW5* (Supplementary Fig. 15c and Supplementary Data 4). These observations are consistent with previous reports that *GmST05*/*GmSW5* is a positive regulator of seed weight[2,40].

Given that *GmSW5* expression increased in *sw19* mutants, we hypothesized that GmSW19 directly represses *GmSW5*. Sequence analysis revealed four putative GmSW19/ABI5–binding elements (ACGTs) in the *GmSW5* promoter (Fig. 4a and Supplementary Fig. 16a)[30]. To test binding, we designed four EMSA probes (P1–P4), each containing one ACGT motif. $GmSW19^{A}$ exhibited minimal binding to P1 and P2, as shown by limited competition with cold probes, but bound strongly to P3 and P4, which was efficiently competed by excess cold probes (Fig. 4b, c and Supplementary Fig. 16b, c). Site–directed mutagenesis of ACGT to AAAA abolished binding (Fig. 4b, c). Yeast one–hybrid assays further confirmed that GmSW19 interacts with promoter segments S1 and S2 (corresponding to P3 and P4), and that mutation of the ACGT elements (mS1 and mS2) abolished these interactions (Fig. 4d, e). To determine which elements mediate the transcriptional regulation of *GmSW5* by GmSW19, we performed dual–luciferase reporter assays. Specifically, we tested the wild–type *GmSW5* promoter (*proGmSW5*) as well as three mutant versions in which the P3 and P4 elements were mutated individually (*proGmSW5–3m* and *proGmSW5–4m*) or simultaneously (*proGmSW5–3m4m*). Our results show that GmSW19 significantly represses the activity of the wild–type *GmSW5* promoter. This repressive effect is markedly reduced when either P3 or P4 is mutated and is almost completely abolished when both elements are mutated simultaneously (Supplementary Fig. 16d, e). These results demonstrate that both P3 and P4 cis–elements are required for GmSW19–mediated repression of *GmSW5*, thereby establishing a functional link between the DNA–binding activity of GmSW19 and its transcriptional regulatory effect on *GmSW5*. Consistent with these molecular results, RT–qPCR revealed that *GmSW5* expression was significantly upregulated in *sw19* mutants but downregulated in both $GmSW19^{A}$–OE and $GmSW19^{C}$–OE lines, with $GmSW19^{C}$ exerting a weaker repression (Fig. 4f). Dual–luciferase expression transient assays revealed that both alleles suppressed *GmSW5* promoter activity and that the coexpression of GmSK21 further enhanced repression (Fig. 4g). These results demonstrate that GmSW19 suppresses *GmSW5* to inhibit seed development, with GmSK21 potentially reinforcing this effect.

To determine the genetic relationship, we generated single and double CRISPR/Cas9 mutants of *GmSW19* and *GmSW5*, using only frameshift alleles likely representing

knockouts (Supplementary Fig. 17a, 18a). In contrast to *sw19*, *sw5* single mutants displayed significantly reduced seed weight (Supplementary Fig. 17b–e). Interestingly, the *sw19 sw5* double mutant presented a seed weight similar to that of the WT (Supplementary Fig. 18b–e). Quantitatively, *sw19* increased seed weight by 9.2% in the WT background but only 6.6% in the *sw5* background, indicating that the effect of *GmSW19* is at least partially dependent on *GmSW5* (Fig. 4h), which is consistent with their molecular regulatory relationship.

$GmSW19^C$ is underutilized likely due to close linkage to the *Dt1*/*dt1* locus

Seed weight increased substantially from wild soybean to landraces during domestication and then to cultivars through modern breeding. To investigate the distribution of *GmSW19* alleles, we analyzed 1,719 soybean accessions, including 218 annual wild soybeans (*G. soja*) and the 1,501 cultivated accessions used for GWAS (938 landraces and 563 improved cultivars). Surprisingly, the frequency of the heavy–seed allele $GmSW19^C$ decreased sharply from 59% in wild soybeans to 17% in landraces but then modestly increased to 22% in improved cultivars (Fig. 5a). This pattern suggests that artificial selection favors the light–seed allele $GmSW19^A$ in landraces and cultivars, whereas modern breeding may be beginning to exploit $GmSW19^C$ for increased seed weight.

We hypothesized that the decreased presence of the $GmSW19^C$ allele in landraces is likely due to its close linkage (~16.4 kb) to *Dt1*, a well–known gene controlling growth habit[34]. Wild soybeans carrying *Dt1* are indeterminate, whereas cultivated soybeans can be either indeterminate (*Dt1*) or determinate (*dt1*). Domestication of *Dt1* involved four nonsynonymous SNPs forming four *dt1* haplotypes. When 100–seed weight was compared across *dt1* haplotypes, compared with *dt1–H4*, *dt1–H1*, *H2*, and *H3* showed reduced seed weight (Supplementary Fig. 2g), which is consistent with recent findings that *Dt1–H4* (*dt1–H4*) does not repress seed weight under short–day conditions[26].

Combining the *Dt1* SNPs with the *GmSW19* C/A SNP, we identified six haplotypes across all 1,719 accessions: H1 ($GmSW19^A$ *Dt1*), H2 ($GmSW19^A$ *dt1–H1*), H3 ($GmSW19^A$ *dt1–H2*), H4 ($GmSW19^A$ *dt1–H3*), H5 ($GmSW19^C$ *dt1–H4*), and H6 ($GmSW19^C$ *Dt1*) (Fig. 5b).

H1 and H6 were present in wild soybeans, whereas H1–H5 appeared in landraces and improved cultivars. We predict that H1–H4 originated from wild H1, whereas H5 likely originated from wild H6. Because determinacy was strongly selected during domestication, the selection of *dt1* haplotypes predominated over that of *GmSW19*, explaining the low frequency of $GmSW19^{C}$. Indeed, the frequency of *dt1–H4* (~20% in landraces, ~28% in improved cultivars) closely matched that of $GmSW19^{C}$ (17% and 22%, respectively), whereas the high frequency of other *Dt1*/*dt1* haplotypes explained the predominance of $GmSW19^{A}$. Similar proportional increases from landraces to improved cultivars were observed for *dt1–H4* and $GmSW19^{C}$ (Fig. 5a and Supplementary Fig. 19a).

Geographical analysis further supported this hypothesis. In central China (CR), 58% of the cultivars carried H1, only 4% carried H5 ($GmSW19^{C}$ *dt1–H4*), and the remainder carried H2–H4. In northern China (NR), H2–H4 were absent, leaving H1 and H5, with the frequency of H5 ($GmSW19^{C}$ *dt1–H4*) increasing to 5%. In southern China (SR), the H5 frequency increased to 19%, substantially higher than that in CR or NR, although still relatively low, mirroring the distribution of *dt1–H4* (Fig. 5c and Supplementary Fig. 19b). Interestingly, in NR, where indeterminate growth is favored for photoperiod adaptation, the presence and even increased proportion of H5 ($GmSW19^{C}$ *dt1–H4*) suggest that the heavy–seed trait of $GmSW19^{C}$ is valuable enough that farmers may tolerate minor photoperiod maladaptation to benefit from larger seeds. Taken together, these analyses indicate that $GmSW19^{C}$ has been underutilized during domestication and breeding due to its tight linkage to *dt1*, but is beginning to be exploited for improving seed weight in soybean breeding.

Potential of the *GmSW19–GmSW5* combination for enhancing soybean seed weight

We next analyzed *GmSW5* alleles across 1,719 soybean accessions. Two major haplotypes, $GmSW5^{H–I}$ and $GmSW5^{H–II}$, corresponded to heavy– and light–seed alleles, respectively. The frequency of $GmSW5^{H–I}$ decreased slightly from wild soybeans (61%) to landraces (51%) but then markedly increased in improved cultivars (81%) (Fig. 5d), indicating that modern breeders have strongly selected this allele to increase seed weight, whereas ancient farmers did not. Geographically, $GmSW5^{H–I}$ accounted for 14% the cultivars in the central

region (CR), 100% of those in the northern region (NR), and 50% of those in the southern region (SR) (Fig. 5e), highlighting its widespread adoption, particularly in NR where the genotype has been fixed. This allele could still be further exploited in the SR for seed weight improvement.

Given that *GmSW19* and *GmSW5* function collaboratively at molecular level, we examined whether their haplotype combinations, including *Gmdt1*, influence seed weight. Compared with the other haplotype combinations, accessions carrying the functional combination *H5/GmSW5*$^{H–I}$ (*GmSW19*^{C} *dt1–H4 GmSW5*$^{H–I}$) presented the heaviest seeds (~20.3 g per 100 seeds) (Fig. 5b). In the H1–H5 backgrounds, *GmSW5*$^{H–II}$ was consistently associated with lower seed weight than *GmSW5*$^{H–I}$. These results indicate that the alleles of all three genes contribute significantly to natural variation in soybean seed weight and that pyramiding the beneficial alleles can substantially increase seed weight (Fig. 5b). Despite its potential, the *H5/GmSW5*$^{H–I}$ combination remains underutilized: it represents only 14% of landraces and 27% of improved cultivars (Fig. 5d). Geographically, it is absent in CR, accounting for only 5% and 14% of NR and SR accessions, respectively (Fig. 5e). Collectively, these findings highlight the substantial potential of the *H5/GmSW5*$^{H–I}$ (*GmSW19*^{C} *dt1–H4 GmSW5*$^{H–I}$) combination for further increasing soybean seed weight in breeding programs.

Discussion

Seed weight is a key agronomic trait that affects crop yield and has been extensively studied in many crop species. Although multiple regulatory factors and genetic pathways have been characterized, the molecular mechanisms controlling seed weight in soybean remain largely unexplored. In this study, we identified GmSW19 as a negative regulator of seed weight and revealed a potential interaction that controls this trait, involving its upstream GSK3–like kinases and a downstream target, thereby advancing our understanding of seed weight regulation in soybean. We further pinpointed the causal variation in *GmSW19* that determines its functional divergence and demonstrated that the superior allele, *GmSW19*^{C}, remains largely underutilized due to its tightly linkage with the *Dt1* locus (Fig. 5b). These findings suggest that the causal variant of *GmSW19* represents an attractive target for

precise genome editing, offering a promising strategy for the rapid improvement of seed size in soybean.

Interestingly, discoveries report here resemble the BIN2–ABI5–MFT pathway revealed in Arabidopsis, which also regulates seed germination[9,12]. Because BIN2 is a core inhibitor of BR signaling and ABI5 is a key transcription factor in ABA signaling, this pathway should also participate in diverse biological processes. Thus, the possible interaction among GmSK21/GmSK21a, GmSW19, and GmSW5 may represent a convergent molecular framework that has been broadly conserved across plant species, offering important insights for studying seed development in other crops. For example, in rice, GSK2 functions analogously to BIN2 and GmSK21/GmSK21a[10]; therefore, it is conceivable that GSK2 may target an ABI5 homologue to modulate grain size, a hypothesis worthy of further investigation. Moreover, previous studies have reported that *Dt1* influences seed size, and recent findings have shown that RCN2, a TFL1 homologue in rice, participates in BR responses through the regulation of grain number[35]. Even more intriguingly, TFL1 has been shown to interact with ABI5[36], and in this study, *Dt1* was found to be tightly linked to *GmSW19*. These observations point to a complex yet potentially convergent regulatory network that may be widely conserved in plants.

Seed weight generally positively correlates with seed size and is a primary determinant of yield in soybean; however, the underlying mechanisms remain unclear. The genomic region surrounding *GmSW19* (45.18–46.23 Mb) harbors QTLs associated with seed weight, seed width, seed length and yield[37–40], providing a potential explanation for the association of these traits. We demonstrated that *GmSW19* carries two haplotypes distinguished by an A523–to–C SNP, resulting in a T175–to–P amino acid substitution that alters its phosphorylation pattern. Consistent with the conserved function of SK21 orthologues, AtSK21/BIN2 interacts with and stabilizes ABI5[9]; in soybean, T175 is a critical phosphorylation site that increases GmSW19^{A} protein abundance, thereby repressing *GmSW5* expression (Fig. 4i). This interaction may further influence *GmSWEET10a* transcription, modulating seed size, seed weight and yield[2]. These insights

offer opportunities to select favorable alleles for enhancing specific traits or trait combinations.

As one of the most intensively selected yield traits, 100–seed weight has undergone strong selection during domestication and modern breeding[1–3,41–43]. Although many favorable alleles of known seed weight genes are nearly fixed in modern cultivars[1,2,6,44,45], the *GmSW19*^{C} allele remains rare, accounting for only 17% of landraces and 22% of improved cultivars, representing a valuable genetic resource for high–yield soybean breeding. The underutilization of *GmSW19*^{C} is largely due to its tight genetic linkage to the *Dt1/dt1* locus, which constrains the introgression of *GmSW19*^{C} into genotypes lacking the *dt1–H4* haplotype (and vice versa for *GmSW19*^{A}) (Fig. 5b). Interestingly, in Arabidopsis, TFL1 (the Dt1 orthologue) negatively regulates seed size by stabilizing ABI5, the GmSW19 orthologue[36]. Considering that AtSK21/BIN2 also modulates ABI5 stability, it is plausible that Dt1 and GmSK21/GmSK21a cooperate to regulate GmSW19 protein abundance and, consequently, seed weight in soybean.

Consistent with recent reports, most soybean accessions from NR harbor the favorable *GmST05/GmSW5* haplotype[2,5]. Unlike *GmSW19*, *GmSW5* has been widely utilized in modern breeding. Our geographical distribution analyses indicate that the *H5/GmSW5*$^{H–I}$ combination remains underutilized in NR and CR and is far from fully exploited in SR (Fig. 5e). Given that combinations of the *GmSW19*, *Dt1*, and *GmSW5* alleles influence seed weight and quality, the results of this study provide valuable insights for rational seed design and the targeted improvement of soybean yield and composition. Given that the A/C variation in GmSW19 is specific to soybean but that the potential interaction between GmSK21 and GmSW19 is likely to be highly conserved across plant species, this variation presents a promising candidate target for improving seed size in a broad range of crops.

Methods

Plant materials and growth conditions

Previously published high–quality genome sequences from 2,214 soybean accessions were utilized in this study[18–20]. Among these, 1,501 cultivars with available phenotypic data were

selected for a GWAS, which revealed significant genotype–phenotype correlations. Detailed information for all accessions is provided in Supplementary Data 1.

In this study, the soybean cultivars Williams 82 (WT) and Huachun6 (H6) were used as controls for generating the transgenic lines. To evaluate 100–seed weight, the wild–type (WT and H6), mutant lines (*sw19*, *sw5*, *sk21*, *sk21a*, *sk21 sk21a* and *sw19 sw5*), and overexpression lines (*GmSW19*, GmSK21^{G787A}–OE and *GmSK21*) were grown in the field in Beijing (40.1°N, 116.7°E) from June to October. To assess yield, a field trial with eighteen replicates was conducted in Beijing using plots arranged in 3–meter rows with 0.5–meter spacing, totaling an area of 0.15 square meters.

GWAS analysis

From the resequencing data of 2,214 soybean accessions, 8.79 million SNPs were imputed using Beagle. After filtering based on a minor allele frequency (MAF) > 0.05 and a missing rate < 0.2 in the 1,501 phenotyped accessions, 6.25 million high–quality SNPs were retained for the GWAS of 100–seed weight[21]. The GWAS was performed using the fixed and random model Circulating Probability Unification (FarmCPU)[22], with a Bonferroni–corrected significance threshold of $P < 1 \times 10^{-8}$ (0.05 / SNP number). Linkage disequilibrium (LD) was calculated using PLINK1.9 (version 1.90) software[46]. Genes located in 100–kb genomic windows on both sides of each peak SNP were considered the primary candidate regions, and the candidate region was plotted using LDheatmap[47].

Expression patterns, RNA extraction, and expression analysis

To identify candidate genes highly expressed in seeds, RNA–seq data from various soybean tissues are obtained from SoyBase (http://www.soybase.org). Gene expression levels were represented as a heatmap depicting fragments per kilobase of transcript per million mapped reads (FPKM), which was constructed using GraphPad Prism 7. To examine the tissue–specific expression of *GmSW19*, leaves, flowers, and seeds of Williams 82 were sampled at different developmental stages. Leaves were sampled at the V3 stage (three fully developed trifoliate leaf nodes), flowers were collected during the full–bloom period, and seeds were harvested at the R4 stage (beginning seed), R5 stage (beginning

seed filling), R6 stage (green seed filling the pod cavity), R7 stage (the first pod on the main stem shows mature pod color), and R8 stage (the color of 95% of pods reached that of maturity) [27]. To compare the relative expression levels of *GmSW19*, *GmSW5* and *GmSK21*, seeds from wild–type, mutants and overexpression lines were collected at the R7 stages. Total RNA was extracted from the above samples using an RNAprep Pure Plant Kit (Tiangen, DP432) and then reverse transcribed using a cDNA synthesis kit (Tiangen, AE311). qPCR was performed with SYBR Green PCR Master Mix (Takara, China). The *Actin11* gene was used as the internal control. All the assays included at least three biological replicates. The primer sequences are described in Supplementary Data 5.

Vector construction and transformation

To construct the *GmSW19^{A}–FLAG*, *GmSW19^{C}–FLAG*, *GmSK21* and *GmSK21^{G787A}–FLAG* plant transformation plasmids, the coding sequences (CDSs) of *GmSW19^{A}*, *GmSK21* and its point–mutated version *GmSK21^{G787A}* from Williams 82, as well as *GmSW19^{C}* from ZP03–5373, were individually cloned and inserted into the vector 0641–FLAG[48] between the *Spe*I and *Xho*I sites using the Gateway system following the manufacturer's instructions (Invitrogen). To generate the CRISPR/Cas9–engineered *sw19*, *sw5*, *sk21* and *sw19*/*sw5* mutants, single guide RNAs (sgRNAs) were designed using the CRISPR–P website (http://crispr.hzau.edu.cn). The sgRNA fragment subsequently was cloned and inserted into the pBSN401 plasmid at the *Bsa*I restriction site[49]. A CRISPR/Cas9 expression vector with the *Cas9* CDS driven by the 35S promoter and each customized sgRNA driven from the Arabidopsis *U6* promoter[49] was constructed. All the constructs were introduced into *Agrobacterium tumefaciens* strain EHA105 and then transformed into Williams 82 using the cotyledon–node method[50], except *GmSK21*, which was introduced into Huachun 6 using the same method. Homozygous mutants were produced after self–pollination of candidate mutants and verified by Sanger sequencing. The relevant primer sequences are described in Supplementary Data 5.

Yeast one–hybrid analysis

A yeast one–hybrid assay was conducted to examine the binding of GmSW19 to the *GmSW5* promoter, following an established protocol[11]. The S1 and S2 promoter fragments

and their respective mutant versions (mS1 and mS2) were inserted into the *Xho*I/*Kpn*I sites of the pAbAi vector and transformed into the Y1HGold strain. The full–length coding sequence of *GmSW19* was cloned and inserted into the pGADT7 vector to generate an activation domain (AD) fusion, which was subsequently transformed into yeast strains containing each promoter–reporter construct (S1–AbAi, S2–AbAi, mS1–AbAi, mS2–AbAi). The p53–AbAi and empty pGADT7 vectors were included as positive and negative controls, respectively. Protein–DNA interactions were assessed by growth on SD/–Leu medium supplemented with 500 ng/mL Aureobasidin A (AbA) for 3 days at 30°C. All sequences of primer used in this study are provided in Supplementary Data 5.

Protein structure prediction

The 3D structures of GmSW19^{A} and GmSW19^{C} were predicted using the ColabFold v1.5.3 webserver with AlphaFold2 using MMseqs2 with the default settings (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb).

RNA–seq analysis

Total RNA was extracted from the seeds of Williams82 (WT), *sw19–1*, and *sw19–2* at the R7 stage from three independent biological samples. RNA–seq was performed by Novogene Company (Beijing, China) using an Illumina NovaSeq 6000 (Illumina, USA) to produce paired–end libraries. Gene expression levels were calculated by FPKM values, and differentially expressed genes (DEGs) whose $|\log_2(\text{fold change})| \geq 1.5$ and whose $P <$ 1e–3 were identified. KEGG pathway enrichment of the DEGs was analyzed using the KOBAS 3.0 database (http://bioinfo.org/kobas/genelist/).

Protein expression and EMSA

The CDSs of *CmSW19^{A}* and *CmSW19^{C}* were subsequently cloned and inserted into the pET32a vector for His–tag fusion. Next, recombinant GmSW19^{A}–His and GmSW19^{C}–His were expressed in *Escherichia coli* Rosetta (DE3) and purified with Ni–NTA magnetic beads (Beyotime, P2226). Before EMSA, GmSW19^{A}–His and GmSW19^{C}–His were normalized to the same concentration. The probes were labelled with biotin at the 5’ ends.

Unlabeled mutant probes were added as cold competitors. DNA gel shift assays were performed according to the manufacturer's instructions for the Chemiluminescent EMSA kit (Beyotime, GS009). The probe sequences are listed in Supplementary Data 5.

In vitro kinase assays

In vitro kinase assays were performed according to a previous study, with minor modifications[10]. Briefly, recombinant His–tagged proteins (His, GmSW19^{A}, GmSW19^{C} or their mutated versions GmSW19$^{\triangle A}$–T175 and GmSW19$^{\triangle C}$–P175) and GST–GmSK21 fused proteins were purified from *Escherichia coli* Rosetta (DE3). For *in vitro* kinase assays, purified GST–GmSK21 (1 μg) was incubated with GmSW19–His (5 μg) in 50 mL of kinase buffer (25 mM Tris, pH 7.5, 1 mM $CaCl_2$, and 10 mM $MgCl_2$) supplemented with 100 μM ATP and 1 mM DTT. After the samples were incubated at 30°C for 1 h, the reactions were terminated by the addition of 6× loading buffer and boiling for 5 min. The proteins were then separated via electrophoresis on 10% acrylamide gels. Phosphorylation was detected using an anti–phosphothreonine antibody (Sigma, P6623). The primer sequences are listed in Supplementary Data 5.

Subcellular localization

The full–length CDSs of the two *GmSW19* alleles were cloned individually in–frame and upstream of *GFP* into the vector pCAMBIA2300–35s–eGFP[51]. The resulting construct was subsequently transformed into Agrobacterium strain EHA105, which was then infiltrated into *N. benthamiana* leaves. Infiltrated leaf epidermal cells were observed after 48 h of incubation in the dark using a confocal laser scanning microscope (Zeiss LSM980). The primers used for subcellular localization are listed in Supplementary Data 5.

Transient dual–LUC assay

To generate the *proGmSW5:LUC* reporter construct, a 2–kb *GmSW5* promoter from Williams 82 was cloned and inserted into the pGreen0800–LUC vector[52], which contains a 35S–driven *Renilla luciferase* (*REN*) gene as an internal control. The *35S:GmSW19^{A}* and *35S:GmSW19^{C}* constructs were used as effectors. The indicated combinations of effectors and reporter constructs were co–transformed into *N. benthamiana* leaves via

Agrobacterium–mediated transient expression. The firefly luciferase (LUC) and *Renilla* luciferase (REN) activities were measured using a Dual–Luciferase Reporter Assay System Kit (Promega, E1910). The primers used are listed in Supplementary Data 5.

Y2H assay

Full–length CDSs of *GmSW19*^{A} and *GmSW19*^{C} were individually cloned and inserted into the prey vector pGADT7. The full–length CDS of *GmSK21* was cloned and inserted into the bait vector pGBKT7. All constructs were transformed into the yeast AH109 strain using the Yeastmaker Yeast Transformation System 2 (Clontech). Yeast colonies were selected on synthetic defined (SD) medium lacking Leu and Trp (–L–W) and were then transferred to SD medium lacking Leu, Trp, His, and Ade (–L–W–H–A). To determine the interaction strength, dilutions of saturated yeast cultures (10^{-1} and 10^{-2}) were spotted onto selection plates following incubation at 30°C. The primers used for Y2H assays are listed in Supplementary Data 5.

BiFC assay

The CDSs of *GmSW19* and *GmSK21* were subsequently cloned and inserted into pSPYCE (MR) and pSPYNE (R)173[53] (digested with *BamH*I) to obtain the *GmSW19–cYFP* and *GmSK21–nYFP* constructs for *Agrobacterium*–mediated infiltration of *N. benthamiana* leaves. After infiltration, the plants were grown in the dark for 2 days, after which the BiFC fluorescence signals were observed using a confocal laser scanning microscope (Zeiss LSM980). The primer sequences are described in Supplementary Data 5.

LCI assay

The *GmSW19–cLUC* and *GmSK21–nLUC*[54] constructs were generated as shown in Supplementary Data 5. *Agrobacteria* harboring different constructs were co–infiltrated in appropriate combinations into *N. benthamiana* leaves, and LUC activity of the infiltrated leaves was analyzed for LUC activity using a Tanon–5200 imaging system at 48 h after infiltration. Photographs were taken 5 min after exposure to 1 mM D–luciferin sodium salt substrate.

Co–IP assay

The CDSs of *GmSW19* and *GmSK21* were subsequently cloned and inserted into the vectors 0641[48] and pCAMBIA2300[51] to generate *GmSW19–FLAG* and *GmSK21–GFP*, respectively. These constructs were coinfiltrated into *N. benthamiana* leaves. Total proteins were extracted with immunoprecipitation (IP) buffer containing 50 mM Tris–HCl pH 8.0, 0.5 M sucrose, 1 mM $MgCl_2$, 10 mM EDTA, 5 mM dithiothreitol, and 1× protease inhibitor cocktail, followed by incubation with anti–GFP beads (Sigma, M8823) for 2 h at 4°C. The beads were then washed five times with 1× PBS (phosphate buffered saline). Total proteins and immunoprecipitates were analyzed by immunoblotting using either an anti–FLAG (MBL, M185–7) or anti–GFP antibody (MBL, 598–7). The primers used are listed in Supplementary Data 5.

Data availability

The raw sequence data[21] reported elsewhere and the RNA-Seq sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive under accession PRJNA681974 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA681974/] and PRJNA1449932 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1449932/], respectively. Gene sequences are available at the Phytozome database (https://phytozome-next.jgi.doe.gov/info/Gmax_Wm82_a2_v1). Source data are provided with this paper.

References

1. Li, J. *et al.* Identification of *ST1* reveals a selection involving hitchhiking of seed morphology and oil content during soybean domestication. *Plant Biotechnol. J.* **20,** 1110–1121 (2022).
2. Duan, Z. *et al.* Natural allelic variation of *GmST05* controlling seed size and quality in soybean. *Plant Biotechnol. J.* **20,** 1807–1818 (2022).
3. Smith, T. J., & Camper Jr, H. M. Effects of seed size on soybean performance. *Agrono. J.* **67**, 681–684 (1975).
4. Lu, X. *et al.* The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. *Plant J.* **86**, 530–44 (2016).
5. Cai, Z. *et al*. MOTHER–OF–FT–AND–TFL1 regulates the seed oil and protein

content in soybean. *New Phytol.* **239**, 905–919 (2023).

6. Wang, S. *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**, 1776–1786 (2020).
7. Hu, D. *et al.* Downregulation of a gibberellin 3β–hydroxylase enhances photosynthesis and increases seed yield in soybean. *New Phytol.* **235,** 502–517 (2022).
8. Li, N., Xu, R. & Li, Y. Molecular networks of seed size control in plants. *Annu. Rev. Plant Biol.* **70**, 435–463 (2019).
9. Hu, Y. & Yu, D. BRASSINOSTEROID INSENSITIVE2 interacts with ABSCISIC ACID INSENSITIVE5 to mediate the antagonism of brassinosteroids to abscisic acid during seed germination in Arabidopsis. *Plant Cell* **26**, 4394–4408 (2014).
10. Tong, H. *et al.* DWARF AND LOW–TILLERING acts as a direct downstream target of a GSK3/SHAGGY–like kinase to mediate brassinosteroid responses in rice. *Plant Cell* **24**, 2562–2577 (2012).
11. Cheng, Z. *et al*. Abscisic acid regulates early seed development in Arabidopsis by ABI5–mediated transcription of *SHORT HYPOCOTYL UNDER BLUE1*. *Plant Cell* **26**, 1053–1068 (2014).
12. Xi, W., Liu, C., Hou, X. & Yu, H. *MOTHER OF FT AND TFL1* regulates seed germination hrough a negative feedback loop modulating ABA signaling in Arabidopsis. *Plant Cell* **22**, 1733–48 (2010).
13. Nguyen, C. X., Paddock, K. J., Zhang, Z. & Stacey, M. G. *GmKIX8–1* regulates organ size in soybean and is the causative gene for the major seed weight QTL *qSw17–1*. *New Phytol.* **229**, 920–934 (2021).
14. Du, J. *et al.* Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. *J. Exp. Bot.* **68**, 1955–1972 (2017).
15. Zhao, B. *et al.* Arabidopsis *KLU* homologue *GmCYP78A72* regulates seed size in soybean. *Plant Mol. Biol.* **90**, 33–47 (2016).
16. Wang, X. *et al.* Evolution and association analysis of *GmCYP78A10* gene with seed size/weight and pod number in soybean. *Mol. Biol. Rep.* **42**, 489–96 (2015).
17. Lu, X. *et al.* A *PP2C–1* allele underlying a quantitative trait locus enhances soybean 100–seed weight. *Mol. Plant* **10**, 670–684 (2017).

18. Wang, L. *et al.* Establishment of Chinese soybean *Glycine max* core collections with agronomic traits and SSR markers. *Euphytica* **151**, 215–223 (2006).
19. Zheng, T. *et al.* SoyFGB v2. 0: a unique access to variations of Chinese Soybean Gene Bank (CNSGB) germplasm. *Sci. Bull.* **17,** 1716–1719 (2021).
20. Guo, Y., Li, Y., Hong, H. & Qiu, L. J. Establishment of the integrated applied core collection and its comparison with mini core collection in soybean (*Glycine max*). *Crop J.* **2**, 38–45 (2014).
21. Li, Y. H. *et al.* Genome–wide signatures of geographic expansion and breeding process in soybean. *Sci. China Life Sci.* **65,** 1–16 (2022).
22. Liu, X., Huang, M., Fan, B., Buckler, E.S. & Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome–wide association studies. *PLoS Genet.* **12**, e1005767 (2016).
23. Orf, J. H., Chase, K., Jarvik, T., Mansur, L. M., Cregan, P. B., Adler, F. R. & Lark, K. G. Genetics of soybean agronomic traits: I. comparison of three related recombinant inbred. *Crop Sci.* **39**, 1642–1651 (1999).
24. Moongkanna, J., Nakasathien, S., Novitzky, W. P., Kwanyuen, P., Sinchaisri, P. & Srinives, P. SSR markers linking to seed traits and total oil content in soybean. *J. Agr. Sci.* **44**, 233–241 (2011).
25. Ye, H., Li, L. Guo, H. & Yin, Y. MYBL2 is a substrate of GSK3–like kinase BIN2 and acts as a corepressor of BES1 in brassinosteroid signaling pathway in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **109**, 20142–20147 (2012).
26. Li, X. *et al*. Dt1 inhibits SWEET–mediated sucrose transport to regulate photoperiod–dependent seed weight in soybean. *Mol. Plant* **17**, 496–508 (2024).
27. Fehr, W. R. & Caviness, C. E. Stages of soybean development special report. Cooperative Extension Service, Agriculture and Home Economic Experiment Station Iowa State University, Ames, Iowa. 1–11 (1977).
28. He, C. *et al*. GSK3–mediated stress signaling inhibits legume–rhizobium symbiosis by phosphorylating GmNSP1 in soybean. *Mol. Plant* **14**, 488–502 (2021).
29. Cheng, X. *et al.* A single amino acid substitution in STKc_GSK3 Kinase conferring semispherical grains and its implications for the origin of *Triticum sphaerococcum*. *Plant Cell* **32**, 923–934 (2020).

30. Huang, H. Y. *et al.* BR signal influences Arabidopsis ovule and seed number through regulating related genes expression by BZR1. *Mol. Plant* **6**, 456–469 (2013).
31. Li, J. & Nam, K. H. Regulation of brassinosteroid signaling by a GSK3/SHAGGY–like kinase. *Science* **295**, 1299–1301 (2002).
32. Islam, F., Khan, M. S. S., Guo, X., Wang, D. & Qi, G. Tissue–specific inhibition of brassinosteroids regulates panicle branching and grain yield in rice. *Sci. China Life Sci.* **68**, 1199–1201 (2025).
33. Li, J., Nam, K. H., Vafeados, D. & Chory, J. *BIN2*, a new brassinosteroid–insensitive locus in Arabidopsis. *Plant Physiol.* **127**, 14–22 (2001).
34. Liu, G. *et al.* Geographical distribution of *GmTfl1* alleles in Chinese soybean varieties. *Crop J.* **3**, 371–378 (2015).
35. Liu, Q. *et al.* Precise control of chromatin loop extrusion enhances sustainable green revolution yield in rice. *Nat. Genet.* **57**, 2798–2807 (2025).
36. Zhang, B., Li, C., Li, Y. & Yu, H. Mobile TERMINAL FLOWER1 determines seed size in Arabidopsis. *Nat. Plants* **6**, 1146–1157 (2020).
37. Specht, J. E. *et al*. Soybean response to water: a QTL analysis of drought tolerance. *Crop Sci*. **41**, 493–509 (2001).
38. Wang, J. *et al.* Identification of quantitative trait loci for oil content in soybean seed. *Crop Sci.* **55**, 23–34 (2015).
39. Salas, P., Oyarzo–Llaipen, J. C., Wang, D., Chase, K. & Mansur, L. Genetic map of seed shape in three populations of recombinant inbred lines of soybean (*Glycine max* L. Merr.). *Theor. Appl. Genet.* **113**, 1459–1466 (2006).
40. Mian, M. A. R. *et al*. Molecular markers associated with seed weight in two soybean populations. *Theor. Appl. Genet.* **93**, 1011–1016 (1996).
41. Kato, S. *et al.* A major and stable QTL associated with seed weight in soybean across multiple environments and genetic backgrounds. *Theor. Appl. Genet.* **127**, 1365–1374 (2014).
42. Zhang, J., Song, Q., Cregan, P. B. & Jiang, G. L. Genome–wide association study, genomic prediction and marker–assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* **129**, 117–130 (2016).
43. Zhang, M. *et al.* Progress in soybean functional genomics over the past decade. *Plant*

Biotechnol. J. **20,** 256–282 (2021).

44. Gu, Y. *et al.* Differential expression of a WRKY gene between wild and cultivated soybeans correlates to seed size. *J. Exp. Bot.* **68**, 2717–2729 (2017).
45. Goettel, W. *et al. POWR1* is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nat. Commun.* **13**, 1–11 (2022).
46. Purcell, S. *et al.* PLINK: a tool set for whole–genome association and population–based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
47. Shin, J. H., Blay, S., McNeney, B., & Graham, J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* **16**, 1–9 (2006).
48. Lyu, X. *et al.* GmCRY1s modulate gibberellin metabolism to regulate soybean shade avoidance in response to reduced blue light. *Mol. Plant* **14**, 298–314 (2021).
49. Xing, H. *et al.* A CRISPR/Cas9 toolkit for multiplex genome editing in plants. *BMC plant biology.* **14**, 1–12 (2014).
50. Paz, M. M., Martinez, J. C., Kalvig, A. B., Fonger, T. M. & Wang, K. Improved cotyledonary node method using an alternative explant derived from mature seed for efficient Agrobacterium–mediated soybean transformation. *Plant Cell Rep* **25**, 206–213 (2006).
51. Hu, B. *et al.* Nitrate–NRT1.1B–SPX4 cascade integrates nitrogen and phosphorus signalling networks in plants. *Nat. Plants* **5**, 401–413 (2019).
52. Hellens, R. P. *et al.* Transient expression vectors for functional genomics, quantification of promoter activity and RNA silencing in plants. *Plant Methods* **1**, 1–14 (2005).
53. Yang, Q. *et al.* A stripe rust effector Pst18363 targets and stabilises TaNUDX23 that promotes stripe rust disease. *New Phytol.* **225**, 880–895 (2020).
54. Chen, H. *et al.* Firefly luciferase complementation imaging assay for protein–protein interactions in plants. *Plant Physiol.* **146**, 368–376 (2008).

Acknowledgments

This work was supported by the National Key R&D Program of China (2021YFD1201601), the National Natural Science Foundation of China (32201756,

32425042), the National Science Foundation for Postdoctoral Scientists of China (2021M703554), the earmarked fund for CARS (CARS-04-PS01), the Agricultural Science and Technology Innovation Program (ASTIP) of the Chinese Academy of Agricultural Sciences and the Agricultural Science and Technology Innovation Program (CAAS-CSNCB-202301).

Author contributions

Y.L., H.T. and L.Q. conceived the study. H.Z. and Y.T. conducted the experiments. C.Z., Z.S., L.Y., Y.G., H.G., Z.L., T.G., W.Y. and Z.Z. provided assistance. H.Z., Y.T., D.L., Y.L., H.T. and L.Q. performed the data analysis. H.Z., Y.T., Y.L., H.T. and L.Q. wrote the manuscript with input from all the others. Y.L., H.T. and L.Q. co-supervised the study.

Competing interests

The authors declare that they have no competing interests.

Figure legends

Fig. 1. GWAS reveals that *qSW19* is associated with 100–seed weight in soybean. (a) Manhattan plot for 100–seed weight in 1,501 accessions grown in Sanya, Nanchang, and Wuhan (2017). Dashed line: genome–wide significance (MAF > 0.05). Red arrows: significant SNPs on Chr19 (*qSW19*) and Chr5 (*qSW5*); black arrow: Chr15 locus (likely *GmSWEET10a*). (b) Correlations of 100–seed weight across the three environments. ***P < 0.001 (two–sided *t* tests). (c) Linkage disequilibrium (LD) plot for Chr19:44.80–45.40 Mb. Black triangle: LD block. Green solid lines above: QTL intervals from SoyBase; QTLs for seed size, weight and yield. (d) Heatmap of candidate gene expression within *qSW19* (FPKM, white–red) from SoyBase RNA–seq data. (e) RT–qPCR of *GmSW19* in various tissues (n = 3 biological replicates). (f) Nonsynonymous mutation at Chr19:45204441 (GmSW19A: T; GmSW19C: P). (g) 100–seed weight for $GmSW19^{A}$ and $GmSW19^{C}$ accessions across three environments. The number of accessions (n) for each group is indicated within the plots. Data are presented as the mean ± SD. Two–sided *t* tests was used to generate the P values. Source data are provided as a Source Data file.

Fig. 2. *GmSW19* negatively regulates 100–seed weight. (a) Genomic structure of *GmSW19* and *sw19* mutants. sgRNA target site (underlined) and CRISPR/Cas9–induced deletions (red dashes). (b) Seed width (left) and length (right) of WT (Williams 82) and *sw19* lines. Scale bars, 1 cm. (c–e) Seed length (c), width (d), and 100–seed weight (e) in WT and *sw19* (n = 49, 50, 48 and 50 biologically independent plants, respectively). (f) Seed weight per plant (n = 16 biologically independent plants). (g) Paraffin sections of R7 seeds from WT and *sw19*. Scale bar, 50 μm. (h) Average cotyledon cell area (n = 20 biological replicates). (i) Total cell number along the longitudinal axis (n = 6 seeds). (j) Seed weight per plot (n = 18 biologically independent plants). Data are presented as the mean ± SD. In all the box plots, the interquartile range is indicated between the first and third quartiles, with the central horizontal line denoting the median; the whiskers extend to the minimum and maximum values. Two–sided *t* tests used for significance. ns, P > 0.05; *P < 0.05; **P < 0.01 and ***P < 0.001. Source data are provided as a Source Data file.

Fig. 3. GmSK21 phosphorylates GmSW19 at the variation position to affect seed weight. (a) Seed length of WT (Williams 82), $GmSW19^{A}$–OE, and $GmSW19^{C}$–OE plants.

Scale bar, 1 cm. (b) Relative *GmSW19* expression in WT and OE at the R7 seed stages (n = 4 biological replicates). ***P < 0.001 (two–sided t tests). (c) 100–seed weight in WT, *GmSW19*^{A}–OE, and *GmSW19*^{C}–OE lines (n = 35, 35, 32, 35 and 35 biologically independent plants, respectively). (d) Seed weight per plant in WT, *GmSW19*^{A}–OE, and *GmSW19*^{C}–OE lines (n = 23 biologically independent plants). (e) Immunoblot of GmSW19–FLAG in R7 seeds; HSP82 was used as a loading control (n = 3 biological replicates). **P < 0.01 (two–sided t tests). (f) Seed width of WT, *sk21 sk21a*, and *GmSK21*^{G787A}–OE plants. Scale bar, 1 cm. (g) Relative *GmSK21* expression in R7 seeds (n = 3 biological replicates). **P < 0.01 (two–sided t tests). (h) 100–seed weight in the WT, *sk21 sk21a*, and *GmSK21*^{G787A}–OE (n = 16 biologically independent plants). (i) The GmSK21 phosphorylation site T175 is critical for GmSW19 function; GmSW19^{A}–T175 and GmSW19^{C}–P175 have all other T–sites mutated to Pro. (j) Yeast two–hybrid assay showing the GmSW19–GmSK21 interaction. –L–W: Leu/Trp dropout; –L–W–H–A: Leu/Trp/His/Ade dropout. (k) Co–immunoprecipitation (Co–IP) confirming the GmSW19–GmSK21 interaction in *N. benthamiana*; total proteins were immunoprecipitated with GFP beads and detected with anti–GFP and anti–FLAG antibodies. The experiment is repeated two times with similar results in (i, k). Data are presented as the mean ± SD. In all the box plots, the interquartile range is indicated between the first and third quartiles, with the central horizontal line denoting the median; the whiskers extend to the minimum and maximum values. Different lowercase letters indicate significant differences (P < 0.05), as determined by one-way ANOVA followed by Tukey's honestly significant difference (HSD) test. Source data are provided as a Source Data file.

Fig. 4. GmSW19 represses *GmSW5* to regulate seed weight. (a) Schematic of the four ACGT motif positions, as well as the probes (P1–P4) and segments (S1, S2) for subsequent analyses, indicated by red bars in the promoter of *GmSW5*. (b, c) EMSA revealed that the GmSW19 protein directly binds to the P3 (b) and P4 (c) regions of the *GmSW5* promoter *in vitro*. The mutated forms with ACGT changed to AAAA (mP3 and mP4) were used as negative controls. The GST protein was used as a negative control. (d, e) Yeast one–hybrid results showing GmSW19 binding to S1 (d) and S2 (e) of the *GmSW5* promoter. Transformants were grown on SD/–Leu ± 500 ng·mL^{-1} Aureobasidin A (AbA). AD,

pGADT7 vector. (f) Relative *GmSW5* expression in the WT, *sw19*, *GmSW19*^{A}–OE, and *GmSW19*^{C}–OE lines (n = 4 biological replicates). (g) Transient dual–LUC assay in *N. benthamiana* leaves. Top: LUC activity of *proGmSW5*:LUC coinfiltrated with 35S:GmSW19 ± GmSK21 and empty vector (EV). Middle: schematic of the constructs. Bottom: Statistical data of LUC/REN ratios (n = 4 biological replicates). (h) Statistical data for the 100–seed weight of WT, *sw5*, *sw19*, and *sw19 sw5* (n = 32, 37, 37 and 34 biologically independent plants, respectively). The data in (f–h) are presented as the mean ± SD. The box represents the interquartile range between the first and third quartiles, with the central horizontal line denoting the median; the whiskers extend to the minimum and maximum values. Different lowercase letters denote significant differences at $P < 0.05$, as determined by one-way ANOVA with Tukey's HSD test. (i) A proposed working model illustrating the potential interactions of GmSK21, GmSW19 and GmSW5 in regulating seed weight. GmSK21 interacts with and differentially phosphorylates GmSW19^{A} and GmSW19^{C}, with T175-present only in GmSW19^{A}—serving as the primary phosphorylation site. This modification enhances the stability of GmSW19^{A}, thereby strengthening its repression of *GmSW5* expression and ultimately reducing seed weight. Blue intensity (from light to dark) indicates enhanced protein stability. We used dashed arrows to indicate the possibly interactions given the fact that these interactions are not fully supported by genetic evidences. Source data are provided as a Source Data file.

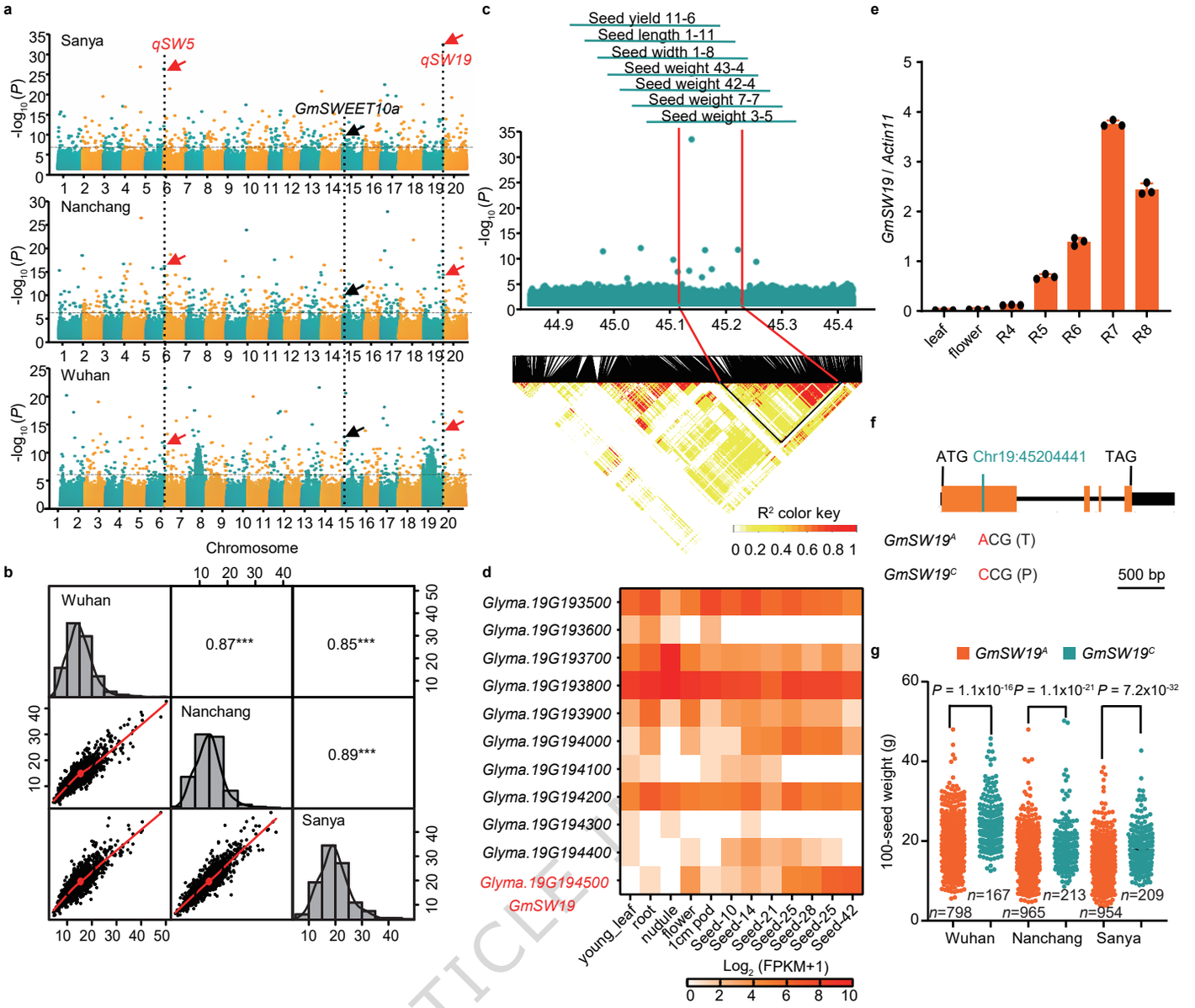
Fig. 5. Utilization of *GmSW19* in soybean breeding. (a) Frequencies of *GmSW19* alleles in wild soybeans, landraces, and improved cultivars (n = 1415, including 188 wild, 767 landraces, and 460 cultivars). (b) Haplotypes of *Dt1* and *GmSW19* across 985 accessions. Six haplotypes (H1–H6) were defined by four *Dt1* SNPs and one *GmSW19* SNP. H1 and H6 are indeterminate; H2–H4 are determinate. In the H1–H5 backgrounds, *GmSW5*$^{H–I}$ confers higher 100–seed weight compared to *GmSW5*$^{H–II}$. (c) Geographic distribution of *GmSW19–Dt1* haplotypes in landraces and cultivars. (d) Frequencies of *GmSW5* alleles (top) and combined *Dt1–GmSW19–GmSW5* genotypes (bottom) in the three germplasm groups. (e) Geographic distribution of *GmSW5* alleles (top) and combined *Dt1–GmSW19–GmSW5* genotypes (bottom) in the three germplasm groups. NR, northern region (including Japan, Korea and the Russian Far East); CR, central region (mid–downstream Yellow

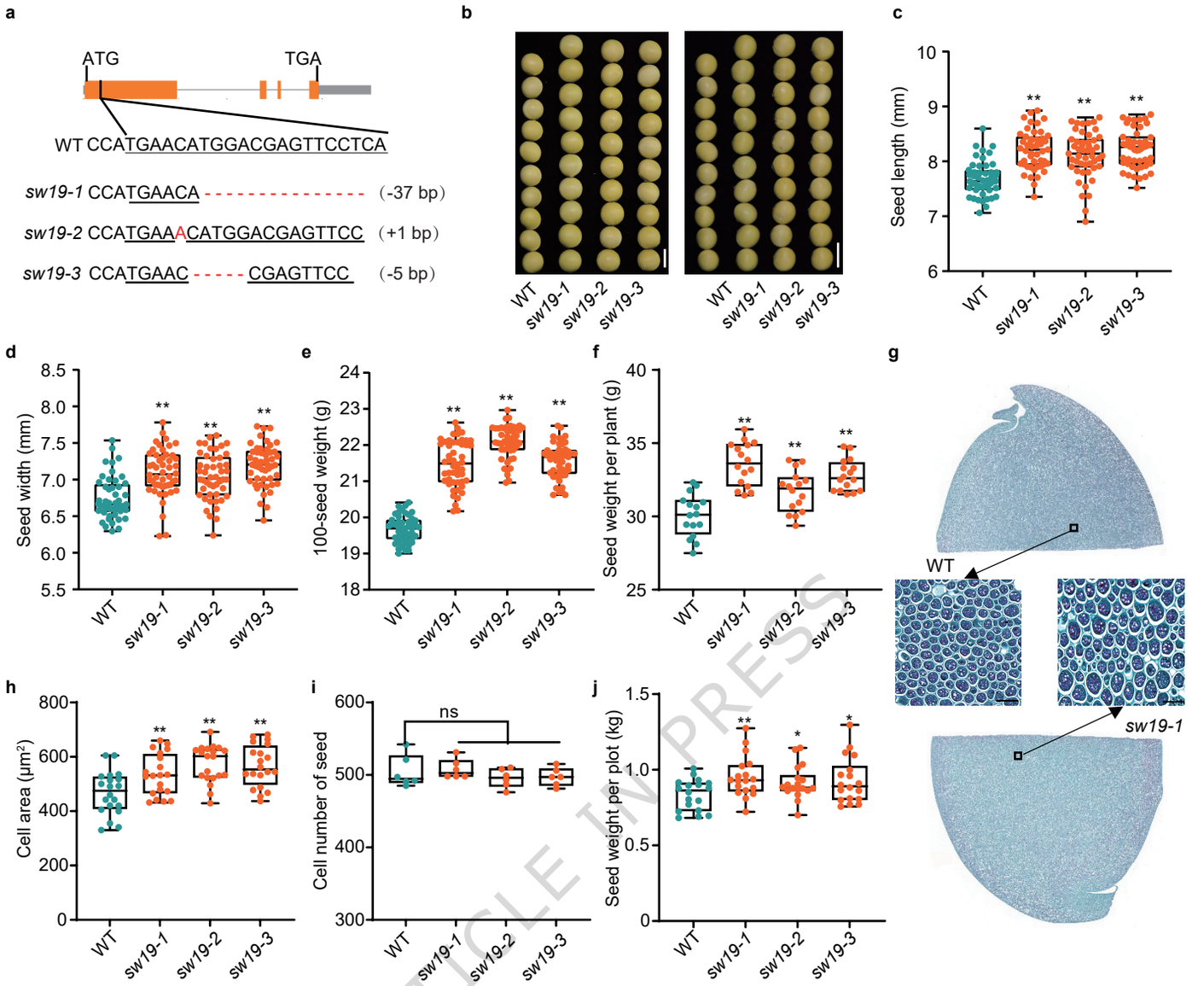
River); SR, southern region. Percentages may not total 100% due to rounding. Source data are provided as a Source Data file.

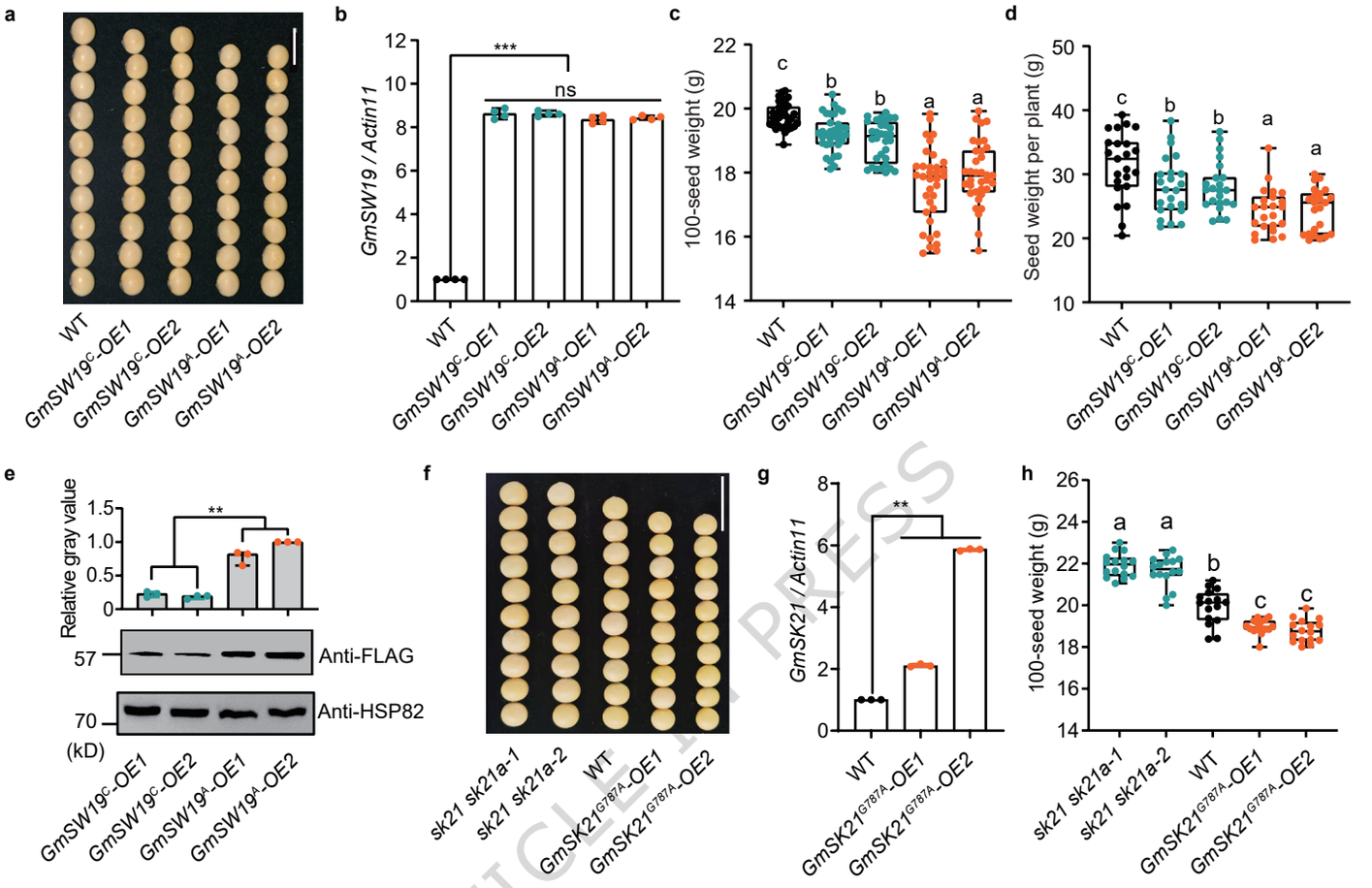
Editor's Summary

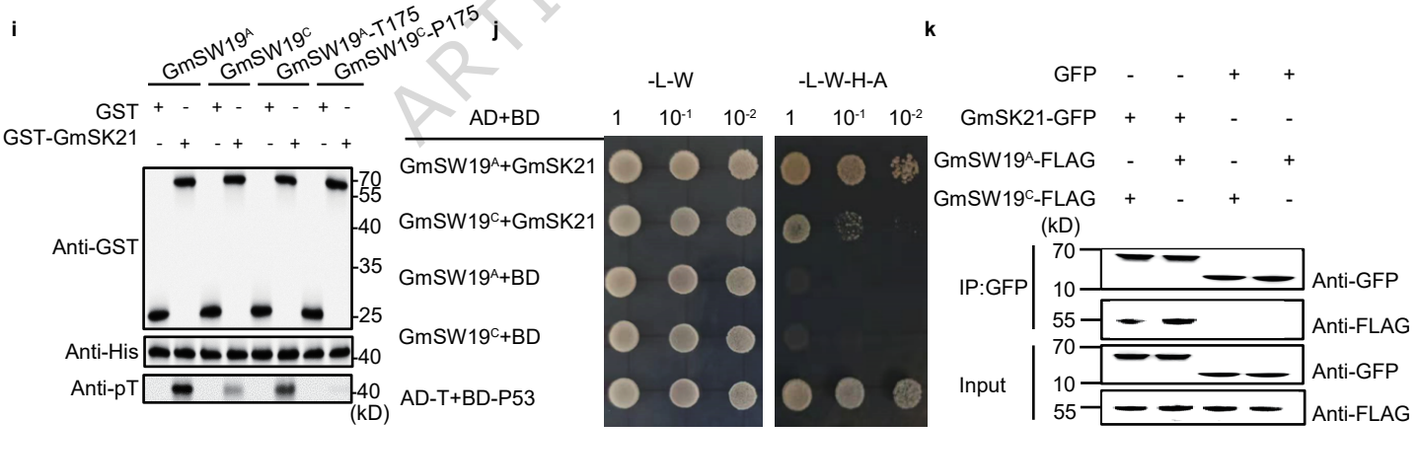
Uncovering the genetic mechanisms that determine seed weight is vital for improving soybean cultivars. Here, the authors identify a bZIP transcription factor, a GSK3-like gene family member, and a MFT ortholog, along with their potential interactions, as key regulators of soybean seed weight.

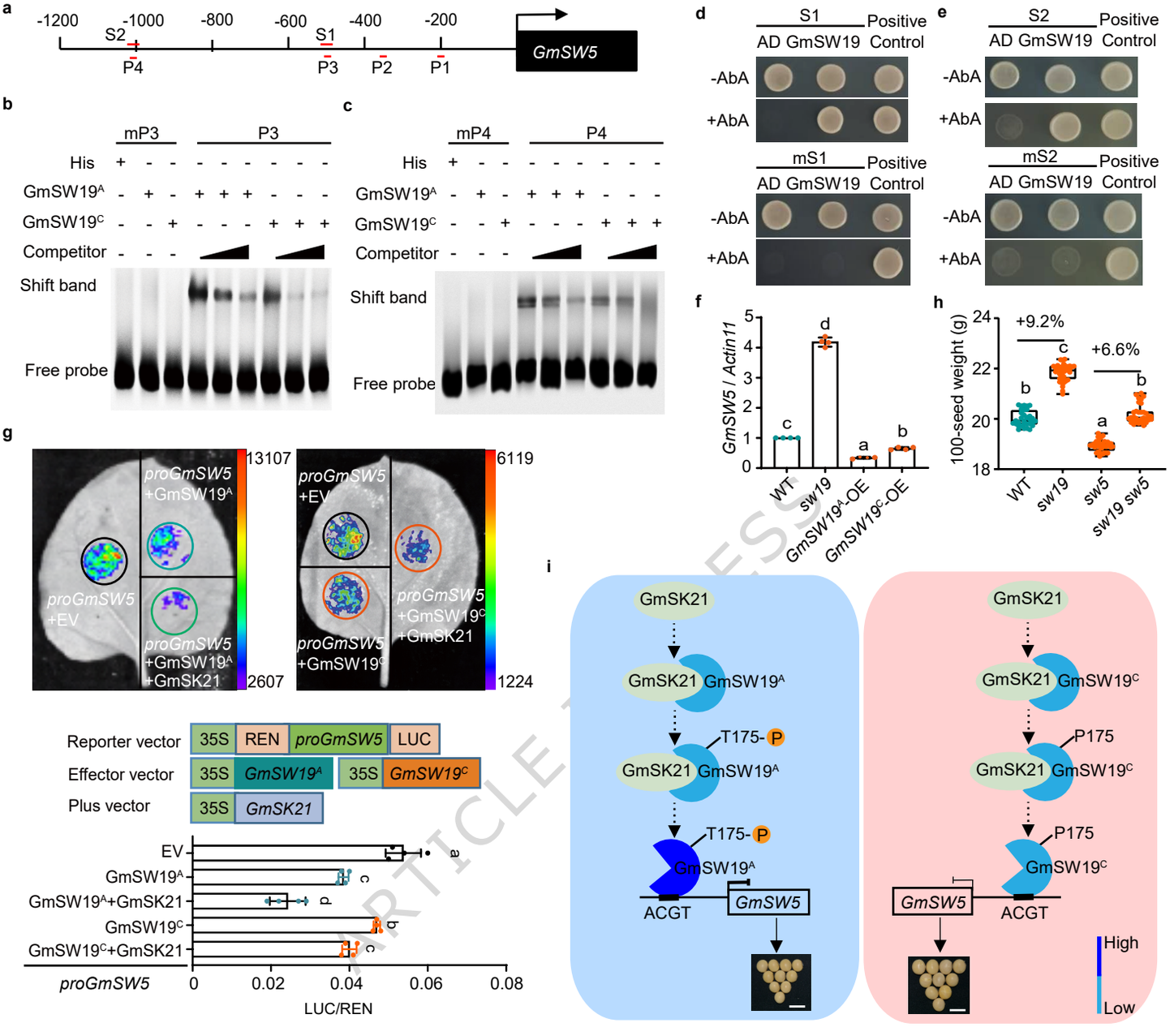
Peer review information: *Nature Communications* thanks Min Ni, Paola Vittorioso and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.










a
-1200
-1000
-800
-600
-400
-200
S2
S1
P4
P3
P2
P1
GmSW5
b
mP3
P3
His
GmSW19A
GmSW19C
Competitor
Shift band
Free probe
c
mP4
P4
d
S1
Positive
AD GmSW19
Control
-AbA
+AbA
mS1
e
S2
mS2
f
GmSW5 / Actin11
WT
sw19
GmSW19A-OE
GmSW19C-OE
h
100-seed weight (g)
+9.2%
+6.6%
sw5
sw19 sw5
g
proGmSW5 +GmSW19A
proGmSW5 +EV
proGmSW5 +GmSW19A +GmSK21
13107
2607
proGmSW5 +GmSW19C
proGmSW5 +GmSW19C +GmSK21
6119
1224
Reporter vector
Effector vector
Plus vector
35S
REN
proGmSW5
LUC
GmSW19A
GmSW19C
GmSK21
EV
GmSW19A+GmSK21
GmSW19C+GmSK21
LUC/REN
i
GmSK21
GmSW19A
T175-P
ACGT
GmSW5
GmSW19C
P175
High
Low