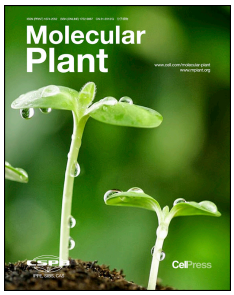
Journal Pre-proof



From DNA to RNA to protein: pQTL mapping and Proteome-Wide Association Study reveal protein-level drivers of soybean traits

Delin Li, Zhan Li, Shiyu Guo, Qi Wang, Xinkang Feng, Hao Zhang, James C. Schnable, Ying-Hui Li, Li-Juan Qiu

PII: S1674-2052(26)00311-4

DOI: https://doi.org/10.1016/j.molp.2026.09.011

Reference: MOLP 2220

To appear in: *Molecular Plant*

Received Date: 9 February 2026

Revised Date: 23 June 2026

Accepted Date: 19 September 2026

Please cite this article as: Li D., Li Z., Guo S., Wang Q., Feng X., Zhang H., Schnable J.C., Li Y.-H., and Qiu L.-J. (2026). From DNA to RNA to protein: pQTL mapping and Proteome-Wide Association Study reveal protein-level drivers of soybean traits. Molecular Plant doi: https://doi.org/10.1016/j.molp.2026.09.011.

From DNA to RNA to protein: pQTL mapping and Proteome-Wide Association Study reveal protein-level drivers of soybean traits

Delin Li[1,2,*,†], Zhan Li[1,†], Shiyu Guo[2], Qi Wang[2,3], Xinkang Feng[2], Hao Zhang[2], James C. Schnable[4], Ying-Hui Li[2,*], Li-Juan Qiu[2,*]

1. Xianghu Laboratory, Hangzhou 311231, China
2. 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
3. Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
4. Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, Nebraska 68583, USA

† These authors contributed equally to this work.

* Correspondence: Ying-Hui Li (liyinghui@caas.cn); Delin Li (lidelin@xhlab.ac.cn); Li-Juan Qiu (qiulijuan@caas.cn)

Running title: Soybean pQTL map and PWAS

Dear Editor,

Proteins are the ultimate functional agents that translate genetic code into phenotype. Protein abundance is known to be dynamic across tissues and development within a given accession (Yang *et al*., 2025) and provides information different from and complementary to mRNA abundance (Walley *et al*., 2016). In humans, protein quantitative trait loci (pQTL) mapping has revealed widespread genetic regulation of protein abundance across populations. In plants, however, this regulatory layer remains largely uncharted (Zhou *et al*., 2021), and whether proteome-wide association study (PWAS) has the potential to recover the "missing heritability" unaccounted for by genomic or transcriptomic variation is unknown. This absence leaves a critical layer of post-transcriptional regulation unresolved in the genotype-to-phenotype map. Population-scale proteomics now makes pQTL mapping and PWAS feasible in crops. Here, we integrate population-scale protein variation with our prior genomic (Li *et al*., 2023), transcriptomic (Li *et al*., 2024), and phenotypic data of various traits to generate a comprehensive pQTL map and demonstrate the capacity of PWAS to directly link protein abundance to agronomic traits in soybean (*Glycine max* [L.] Merr.). Our study provides an integrated view of the chain from DNA to RNA to protein to trait and establishes protein abundance as a direct, heritable mediator of agronomic traits.

We assembled a diverse soybean panel consisting of 111 landraces and 89 improved cultivars (**Supplemental Table 1**), derived from a previously established panel with available genomic and transcriptomic data (**Supplemental Figure 1**). This mini-core collection, representing the Chinese National Soybean GeneBank, captures broad phenotypic and genotypic diversity and encompasses four major subpopulations (**Figure 1A**). Protein abundance was quantified from seedling shoot tissue at the V2 stage, the same tissue source used for the matched transcriptomic dataset. Using a pan-protein reference database (**Supplemental Table 2**) and Astral data-independent acquisition mass spectrometry, we quantified 19,700 proteins, encoded by 17,341 genes, across the entire panel. The pan-proteome approach substantially expanded protein discovery: 12,790 proteins (64.9%) were quantified directly from the Wm82 v2.1 annotation; an additional 6,051 proteins (30.7%) were allelic variants that showed sequence similarity to Wm82 proteins and were quantified from other genomes; and 859 proteins (4.4%) were absent from Wm82, demonstrating the critical value of a population-aware database for capturing true proteomic diversity. Unlike genomic PCA, full-proteome principal components failed to recapitulate subpopulation structure,

whereas UMAP on pQTL proteins or the top 4,000 most variable proteins recovered the four subpopulations (**Supplemental Figure 2**), suggesting proteome variation is regulated beyond genome-wide genetic structure.

To map the genetic basis of proteome variation, we performed pQTL mapping (**Figure 1B**) using 4,295,983 high-quality SNPs (minor allele frequency > 5%) and the BLINK model, which identifies a single representative lead SNP per associated locus. Across the 19,700 proteins quantified, we identified 3,190 pQTL for 2,550 proteins (12.9%) encoded by 2,280 genes, including 52 genes entirely absent from Wm82 (**Figure 1C-D, Supplemental Table 3**). Among the 3,190 total pQTL**,** 1,921 (60.2%) were classified as *cis*-pQTL (lead SNP within 100 kb of the gene). As in eQTL studies, *trans*-pQTL (lead SNP beyond 100 kb of the gene) showed significantly lower phenotypic variance explained (PVE) than *cis*-pQTL (one-tailed Student's t-test, $p = 6.93 \times 10^{-155}$, **Supplemental Figure 3**), highlighting the greater statistical challenge in detecting *trans* associations. The lead SNP was frequently located within the coding regions of the corresponding gene itself (404 out of 1,921, 21.0%; **Figure 1E**). Collectively, this high-resolution pQTL map provides a genome-wide view of the genetic regulation underlying heritable proteome diversity in soybean.

Direct comparison with eQTL (Li *et al.*, 2024) from the same developmental stage revealed 444 proteins (encoded by 415 genes) with (*cis*- or *trans*-) pQTL but without any eQTL, further supporting the existence of post-transcriptional regulation. Specifically, we identified 634 genes with *cis*-pQTL but not with *cis*-eQTL (**Figure 1F**). Notably, the PVE for these protein-only *cis*-pQTL was significantly higher than for those colocalizing with *cis*-eQTL (one-tailed Student's t-test, $p = 0.03$; **Supplemental Figure 3**). Furthermore, the correlation between gene expression and protein abundance was significantly lower (one-tailed Student's t-test, $p = 1.43 \times 10^{-32}$; **Supplemental Figure 4**) for genes with protein-only *cis*-pQTL than for genes with *cis*-pQTL and *cis*-eQTL. Collectively, the higher PVE and lower expression–protein correlation suggest that the protein-only *cis*-pQTL likely reflect genetic regulatory associations rather than statistical false-positives. The 634 genes were significantly enriched for core post-transcriptional and protein-homeostasis processes, including translation (e.g., tRNA aminoacylation, elongation), protein folding and modification, and proteolysis (**Supplemental Figure 5**).

To validate these findings, we selected HSP90 (gene id: *Glyma.02G124500*), a protein-folding factor with significant abundance variation and a strong *cis*-pQTL (2-12458433) but not *cis*-eQTL at the transcript level for molecular validation (**Supplemental Figure 6, Supplemental Table 4**). Based on this *cis*-pQTL, three accessions were randomly selected for each genotype (**Figure 1G**). qRT-PCR analysis verified the level of mRNA expression for *HSP90* was comparable between the groups (**Figure 1H, Supplemental Table 5**), indicating the difference in protein accumulation likely reflected post-transcriptional regulation. Analysis of sequence data identified two non-synonymous SNPs distinguishing the two groups, resulting in the amino acid substitutions Ala478Ser and Ala765Thr (**Figure 1I**). Those two SNPs have LD r^2 of 0.98 (p.Ala478Ser) and 0.43 (p.Ala765Thr) with the *cis*-pQTL, and haplotype analysis of those two SNPs further showed that p.Ala478Ser was the primary haplotype determinant of HSP90 abundance (**Supplemental Figure 7**). We cloned *HSP90* from KA6 (carrying the high protein abundance allele) and HA3 (carrying the low protein abundance allele) and used a protoplast transient expression system, combined with the protein synthesis inhibitor cycloheximide (CHX) to quantify the stability of the proteins produced by the two alleles. Before treatment with CHX, protein levels were similar for both alleles but at 12 hours post-treatment $HSP90^{HA3}$ abundance was significantly lower than that of $HSP90^{KA6}$ (**Figure 1J**), indicating that the allelic variation indeed impacts protein stability. Given that both serine and threonine can act as phosphorylation sites that alter protein stability (Lee *et al.*, 2023), one possible explanation is that the observed amino acid substitutions affect HSP90 stability through altered phosphorylation. Together with the protein-only pQTL genes, this case shows that natural variation in protein abundance can arise directly from protein-stability differences, independent of expression.

Two principal approaches can be used for PWAS: (1) integrating pQTL with results of genome-wide association study (GWAS) from independent studies — an inferred approach widely used in human genetics, which requires the reference panel tissue to match the trait of interest (Wingo *et al.*, 2021); or (2) directly associating measured protein abundance variation with trait variation in the same population — a direct approach. Owing to the availability of inbred lines, the latter approach is uniquely feasible in plants and, as demonstrated for transcriptome-wide association study (TWAS), is less affected by linkage disequilibrium (LD) than GWAS (Li *et al.*, 2021). We applied this direct PWAS approach to soybean pubescence color, conducting GWAS (**Figure 1K**)

and TWAS (**Supplemental Figure 8**) in parallel for comparison. While GWAS identified a broad associated region spanning ~1.1 Mb (**Figure 1L**) that included the known causal locus *T* (Toda *et al.*, 2002) alongside 54 other annotated genes, both PWAS and TWAS specifically pinpointed the T protein/gene itself with the most significant *p* value. In this example, PWAS delivered precise protein-level associations with substantially fewer LD-driven associations than GWAS.

The qualitative pubescence color trait illustrates a consistent flow of genetic information, where variation in DNA is mirrored in both the transcriptome and proteome (**Figure 1M**, **Supplemental Table 6**). However, depending on the biological role of the causal gene, the nature of the mutation, or the genetic architecture of the trait, associations are not always consistent across all three levels. While TWAS is known to be less affected by LD and robust to expression tissue source for quantitative traits, the performance of direct PWAS in complex trait dissection remains largely unexplored in soybean. To evaluate this, we extended our analysis to two polygenic traits, soybean cyst nematode (SCN) resistance and flowering time (FT), conducting GWAS, TWAS, and PWAS in parallel. For SCN resistance (**Supplemental Figure 9**), GWAS identified four loci harboring known genes; TWAS recovered three of these, while PWAS independently identified two. In the case of FT, GWAS detected two known loci and TWAS identified a different set of four known genes (**Supplemental Figure 10**). PWAS identified associations between the abundance of 27 proteins and FT, including two proteins absent from the Wm82 reference genome. Although none of the PWAS-identified proteins correspond to known FT genes in soybean, two of these proteins are homologs of *Arabidopsis* genes with documented roles in regulating flowering (**Supplemental Table 7**). Notably, 11 of the 27 FT-associated proteins had detectable pQTL, and 9 of these were *cis*-pQTL, suggesting that these PWAS signals are less likely to represent purely downstream consequences, though functional validation is still needed for the novel proteins identified by PWAS. Critically, each omics layer—DNA, RNA, and protein—captured a unique subset of associations for those two traits. PWAS thus complements TWAS (pQTL without eQTL) and GWAS (proteins with only *trans*-pQTL). Our study establishes pQTL mapping and PWAS as transformative approaches for plant genetics, offering a direct, protein-centric view of how genetic variation shapes phenotype.

PWAS associations were detected using seedling shoot (V2 stage), a tissue not directly linked to pubescence color or SCN resistance, indicating robustness of PWAS to tissue source. The *cis*-

pQTL of the three known causal genes identified by PWAS (*T*, *NSF07*, and *Rhg1*) colocalized with the corresponding GWAS loci (**Supplemental Table 8**), supporting GWAS–pQTL integration as in inferred PWAS. The proteomic data remain limited by a lower pQTL detection rate (12.9% of quantified proteins) than eQTL (70.2% of expressed genes; Li et al., 2024) and lower PVE from protein abundance than from the lead SNPs for *T*, *NSF07*, and *Rhg1* (**Supplemental Figure 11**), likely reflecting technical and replication constraints (**Supplemental Note**).

Data and code availability

The raw mass spectrometry data for the 200 diverse soybean accessions have been deposited in the OMIX database (China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences) under BioProject PRJCA064953, with accession numbers OMIX017222 and OMIX017228. The protein abundance matrix data are available on figshare at https://doi.org/10.6084/m9.figshare.30997609. All code, example data, and instructions for PWAS are publicly accessible on GitHub (https://github.com/DelinLi/PWAS_Soy).

Funding

This work was supported by the National Key Research and Development Program of China (2021YFD1201601, 2023YFD1202100), National Natural Science Foundation of China (32201759, U22A20473), Zhejiang Provincial Natural Science Foundation (LZYQ25C130001), the earmarked fund for CARS (CARS-04-PS01), and the Agricultural Science and Technology Innovation Program (ASTIP).

Acknowledgments

We would like to thank Dr. Zhiyong Liu (Chinese Academy of Sciences) and Dr. Qiuhong Wu (Xianghu Laboratory) for fruitful discussions.

Declaration of interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Deepseek to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Author contributions

Y.-H.L., D.L., and L.-J.Q. designed the research. Y.-H.L. and L.-J.Q. supervised the project. S.G., Q.W., X.F., and H.Z. collected data. D.L. and Z.L. performed data analyses and conducted wet lab experiments. D.L., Z.L., J.C.S., and Y.-H.L. wrote the manuscript. All authors participated in project discussions and approved the final manuscript.

Supplemental information

Supplemental information is available at *Molecular Plant* Online.

Reference:

Lee, J.M., Hammarén, H.M., Savitski, M.M., and Baek, S.H. (2023). Control of protein stability by post-translational modifications. Nature Communications **14**:201.

Li, D., Liu, Q., and Schnable, P.S. (2021). TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. Plant Physiology **186**:1800-1811.

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

Li, Y.-H., Qin, C., Wang, L., Jiao, C., Hong, H., Tian, Y., Li, Y., Xing, G., Wang, J., Gu, Y., *et al.* (2023). Genome-wide signatures of the geographic expansion and breeding of soybean. Science China Life Sciences **66**:350-365.

Toda, K., Yang, D., Yamanaka, N., Watanabe, S., Harada, K., and Takahashi, R. (2002). A single-base deletion in soybean flavonoid 3'-hydroxylase gene is associated with gray pubescence color. Plant Molecular Biology **50**:187-196.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., *et al.* (2016). Integration of omic networks in a developmental atlas of maize. Science **353**:814-818.

Wingo, A.P., Liu, Y., Gerasimov, E.S., Gockley, J., Logsdon, B.A., Duong, D.M., Dammer, E.B., Robins, C., Beach, T.G., Reiman, E.M., et al. (2021). Integrating human brain proteomes

with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. Nature Genetics **53**:143-146.

Yang, Q., Hou, Z.-Y., Li, L., Wang, L., Li, S.-T., Li, Y., Zhang, X., Huang, H., Ke, Y., Ma, X., *et al.* (2025). Landscape and m^6A post-transcriptional regulation of soybean proteome. Cell Genomics **5**:100926.

Zhou, Q., Fu, Z., Liu, H., Wang, J., Guo, Z., Zhang, X., Tian, R., Liu, Y., Qu, J., and Li, W. (2021). Mining novel kernel size-related genes by pQTL mapping and multi-omics integrative analysis in developing maize kernels. Plant Biotechnology Journal **19**:1489.

Figure 1. **Soybean proteome diversity provides genome-wide protein regulation and protein-level trait association**

(A) Phenotypic and genetic overview of the diversity panel. Seed phenotypic variation is arranged to form the Chinese character '豆' (soybean). Below is an unrooted phylogenetic tree based on genome-wide SNP data, with colors indicating subpopulations (NR, northern China; CR, central China; SR, southern China; Am, American; Mix, admixed) previously classified (Li *et al.*, 2023). **(B)** Schematic of pQTL mapping. pQTL represent genomic loci where genetic variation is associated with changes in protein abundance across a population. **(C)** Genome-wide distribution of pQTL. Each point represents a pQTL associated with a specific protein (*cis*-pQTL in red, *trans*-pQTL in blue). **(D)** Summary of *cis*- and *trans*-pQTL. **(E)** Distribution of distances between the lead SNP of *cis*-pQTL and the corresponding gene. Red bars indicate SNPs located within the gene body of the regulated gene. **(F)** Overlap of genes between pQTL and eQTL. Venn diagram compares protein- and transcript-level associations (eQTL data from Li *et al.*, 2024). **(G–J)** Molecular validation of a *cis*-pQTL for HSP90. **(G)** HSP90 protein abundance in six representative accessions grouped by genotype at the lead *cis*-pQTL SNP. *p* value from a Student's t-test between genotypes (light blue vs. gray). **(H)** *HSP90* transcript levels (mean ± SD, n = 3 replicates) in the same six accessions, analyzed by qRT-PCR using Actin as a reference gene. Lowercase letters denote statistically distinct groups (Tukey's test, $p < 0.05$). The *p* values for the comparison between *cis*-pQTL genotype groups are from Student's t-test. **(I)** Protein structure of HSP90 indicating amino acid substitutions distinguishing the two groups. **(J)** Time course of HSP90 protein levels following cycloheximide (CHX, 10 μM) treatment in protoplasts expressing the KA6 or HA3 allelic variant. A representative blot of three independent replicates is shown, with relative HSP90 protein abundance (quantified by ImageJ) indicated below. Significance symbols are from a one-tailed Student's t-test (HA3 < KA6) across the three replicates: n.s., not significant; $*p < 0.05$, $**p < 0.01$. **(K–M)** Integrative GWAS and PWAS for pubescence color. **(K)** Comparison of GWAS (top; each point is a SNP) and PWAS (bottom; each point is a protein) for pubescence color in the same population. The known causal gene *T* is highlighted. **(L)** Zoomed-in view of the 2-Mb region centered on *T*, showing GWAS (top) and PWAS (bottom) results. **(M)** Multi-omics view of the pubescence color phenotype across the population, showing genotype distribution at the GWAS lead SNP, *T* transcript levels, and T protein abundance, illustrating the flow of genetic information from DNA to phenotype.