Plant Communications
Research article

CellPress
Partner Journal

# Spatiotemporal transcriptomic and metabolomic landscapes of wild soybean seed development reveal regulatory mechanisms of nutrient accumulation

Peiyan Liu[1,2], Mingyang Li[1,2], Ping Ma[1], Hao Yan[1], Chunyan Liu[1], Zhenbang Hu[1], Mingliang Yang[1], Qingshan Chen[1,*] and Ying Zhao[1,*]

[1]National Key Laboratory of Smart Farm Technology and System, Key Laboratory of Soybean Biology, Chinese Education Ministry, College of Agriculture, Northeast Agricultural University, Harbin 150000, China

[2]These authors contributed equally to this article.

***Correspondence: Qingshan Chen (qshchen@126.com), Ying Zhao (tianshi198937@126.com)**

https://doi.org/10.1016/j.xplc.2025.101580

## ABSTRACT

**Seed development is a pivotal stage of the soybean life cycle, directly determining yield and nutritional quality related to oil and protein contents. However, the spatiotemporal mechanisms underlying cell differentiation and nutrient accumulation during seed growth remain to be resolved, especially in wild soybean (*Glycine soja*), which harbors rich genetic diversity for quality traits. Here, spatial transcriptomics and metabolomics were combined to dissect the dynamics of cell differentiation and nutrient accumulation in *G. soja* seeds at the mid-maturity stage. Differential expression analysis revealed distinct patterns of accumulation in adaxial versus abaxial parenchyma cells of the embryo: abaxial cells are enriched in protein metabolism pathways, whereas adaxial cells are focused on lipid metabolism pathways, consistent with previous reports on spatial nutrient accumulation in *G. soja* seeds. Pseudotemporal trajectory analyses supported a sequential pattern of transcriptional regulation underlying these differences. Analysis of cell–cell communication provided insight into the interactions that may mediate cell-type-specific differences among seed cells. Key genetic regulators and differentially abundant metabolites were identified through the integration of spatial transcriptomics and metabolomics, and *GsMAPK23-4* was identified as a core candidate gene linked to nutrient metabolism in the cotyledon. Functional validation confirmed that *GsMAPK23-4* modulates seed quality: knockout mutants had significantly higher levels of amino acids and proteins. These findings reveal cellular characteristics and differentiation processes in *G. soja* seeds at the mid-maturity stage, providing a molecular basis for understanding this phase and generating targets to improve soybean yield and quality.**

**Key words:** spatial transcriptomics, spatial metabolomics, seed development, soybean, crop wild relative

**Liu P., Li M., Ma P., Yan H., Liu C., Hu Z., Yang M., Chen Q., and Zhao Y.** (2026). Spatiotemporal transcriptomic and metabolomic landscapes of wild soybean seed development reveal regulatory mechanisms of nutrient accumulation. Plant Comm. **7**, 101580.

## INTRODUCTION

Cultivated soybean (*Glycine max*) was domesticated from its annual wild relative (*Glycine soja*) in China approximately 5000 years ago (Hymowitz, 1970; Li et al., 2008). Soybean is an extremely economically valuable legume crop that is cultivated throughout the globe, and in 2023 alone, its worldwide harvest was 398.2 million tons (Volkova and Smolyaninova, 2024). As soybean seeds are a rich source of edible oil (20% on average) and protein (40% on average), they are a leading source of nutrients and energy for both humans and animals, and they are frequently used in the preparation of various industrial products. However, cultivated soybean has undergone multiple genetic bottlenecks during domestication and improvement, resulting in decreased genetic diversity (Kim et al., 2021). By contrast, wild soybean has retained higher genetic diversity, together with enhanced genetic quality, broad ecological adaptability (Zhang et al., 2014), and multi-layered resistance to

 

both biotic and abiotic stressors (Kim et al., 2023). The collection and conservation of wild soybean relatives, together with the dissection of their genomic features, have the potential to improve soybean breeding and promote sustainable agriculture and food production. Studies on soybean seed development can aid in the development of synthetic or breeding-based approaches to enhancing seed yield and quality through incorporation of additional nutrients (Egamberdieva et al., 2021), reduction in allergen content (Wiederstein et al., 2023), or increases in yield in terms of seed size or number (Duan et al., 2022).

Seed development in flowering plants is complex, consisting of temporally distinct morphogenesis and maturation phases that are initiated by a double-fertilization event (Jo et al., 2019; Wang et al., 2022). In soybean seeds, this event results in the differentiation of three functionally, anatomically, and genetically distinct tissues: the embryo, endosperm, and seed coat (Lin et al., 2017). During seed morphogenesis, the anatomical specification of the embryo and endosperm into functional domains takes place first. Proliferation of the endosperm causes it to occupy the majority of the embryo sac after fertilization, and it is thought to provide nourishment for embryo development (Le et al., 2007). The cellular components of the endosperm are gradually absorbed by the embryo, such that mature seeds contain only a vestigial layer of endosperm cells, termed the aleurone. The embryo undergoes differentiation into two predominant regions: the cotyledon, a storage organ that provides nutrients to the seedling during germination, and the axis containing the shoot and meristems, which ultimately give rise to the next seedling (Lin et al., 2017). After the initiation of embryonic and endosperm development, the division and differentiation of ovule-integument cells give rise to a seed-coat layer that protects the underlying embryo and endosperm.

During the maturation phase, which largely occurs after morphogenesis with only limited overlap, cellular proliferation and differentiation are interrupted (Pelletier et al., 2017). The major active processes at this stage include extensive biosynthesis and accumulation of storage compounds, pronounced increases in embryo cell size associated with cell expansion, acquisition of desiccation tolerance, and establishment of dormancy (Jo et al., 2019). These processes have major agronomic implications for seed crops such as soybean. During seed maturation, cotyledon cells undergo endoreduplication that is thought to support the production of highly abundant seed-storage compounds (Lin et al., 2017). When seed development is complete, the embryo and other structures enter a state of developmental arrest, desiccation, and metabolic quiescence, in which they generally remain until germination (Jo et al., 2019). The appropriate spatiotemporal coordination of these processes is essential for proper seed growth and development. However, the mechanisms necessary for the functional differentiation of individual seed regions and the coordination of these processes during seed development are not well understood.

Insight into the genetic networks active within seeds has been obtained through approaches including RNA sequencing (Machado et al., 2020), microarray analyses (Almeida-Silva et al., 2021), and laser-capture microdissection (Schnabel et al., 2023). Patterns of gene expression can reflect both spatial differences among various cell regions and temporal differences across various stages of development. Pronounced gene expression reprogramming takes place during the early-maturity stage, which entails a rapid increase in the nutrient content within the cotyledon. During this stage, genes involved in the accumulation of seed-storage reserves, including those that encode storage proteins and fatty acid biosynthesis enzymes, are activated. Combinations of laser-capture microdissection and GeneChips or microarrays have been used to explore gene expression in different subregions of soybean seeds and across different phases of development. Recent advances in high-throughput single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics now enable the direct probing of cellular heterogeneity and intercellular communication by identifying markers that can associate cellular localization with cell type for detailed transcriptomic investigation of tissues (Wong et al., 2018; Yu et al., 2023; Sang and Kong, 2024). In addition to spatial transcriptomics, spatial metabolomics has been used to identify metabolites in specific tissue subregions and to investigate their spatiotemporal dynamics during development (Nakabayashi et al., 2021). Notably, recent studies have integrated spatial transcriptomics with spatial metabolomics to address biological questions, demonstrating the synergistic potential of these two omics approaches (Rao et al., 2021; Ge et al., 2025).

We used an integrated multi-omics approach to investigate wild soybean seeds, asking how cellular heterogeneity underpins the spatiotemporal regulation of oil and protein accumulation. We integrated high-throughput spatial transcriptomics, spatial metabolomics, and scRNA-seq data to establish a high-resolution, spatially resolved overview of gene expression at the mid-maturity stage of seeds from wild soybean *G. soja* at a crucial point in nutrient accumulation. Cell types were classified, and predicted cell-type annotations were used to support spatially resolved transcriptomic analyses, followed by the reconstruction of differentiation lineages and the identification of putative regulatory factors involved in shaping the accumulation of storage compounds. This combined spatially resolved transcriptomic and metabolomic approach provides a robust foundation for hypothesis generation and subsequent testing in studies aimed at characterizing patterns of spatiotemporal gene expression and accumulation of key metabolites involved in wild soybean seed development and nutrient accumulation at the cellular and tissue levels.

## RESULTS

### High-quality spatial transcriptomics data from wild soybean seeds

Soybean seed development is broadly separated into the cotyledonary, early-maturity, mid-maturity, late-maturity, and dry-seed stages. Here, spatial transcriptomics data were generated using wild-type *G. soja* seeds collected at the mid-maturity stage, the critical period for accumulation of compounds associated with seed quality, including the major products protein and oil (Figure 1A). Seed oils and storage proteins accumulate progressively during seed development (Santos-Mendoza et al., 2008). Transmission electron microscopy (TEM) of oil bodies during seed development confirmed the accumulation

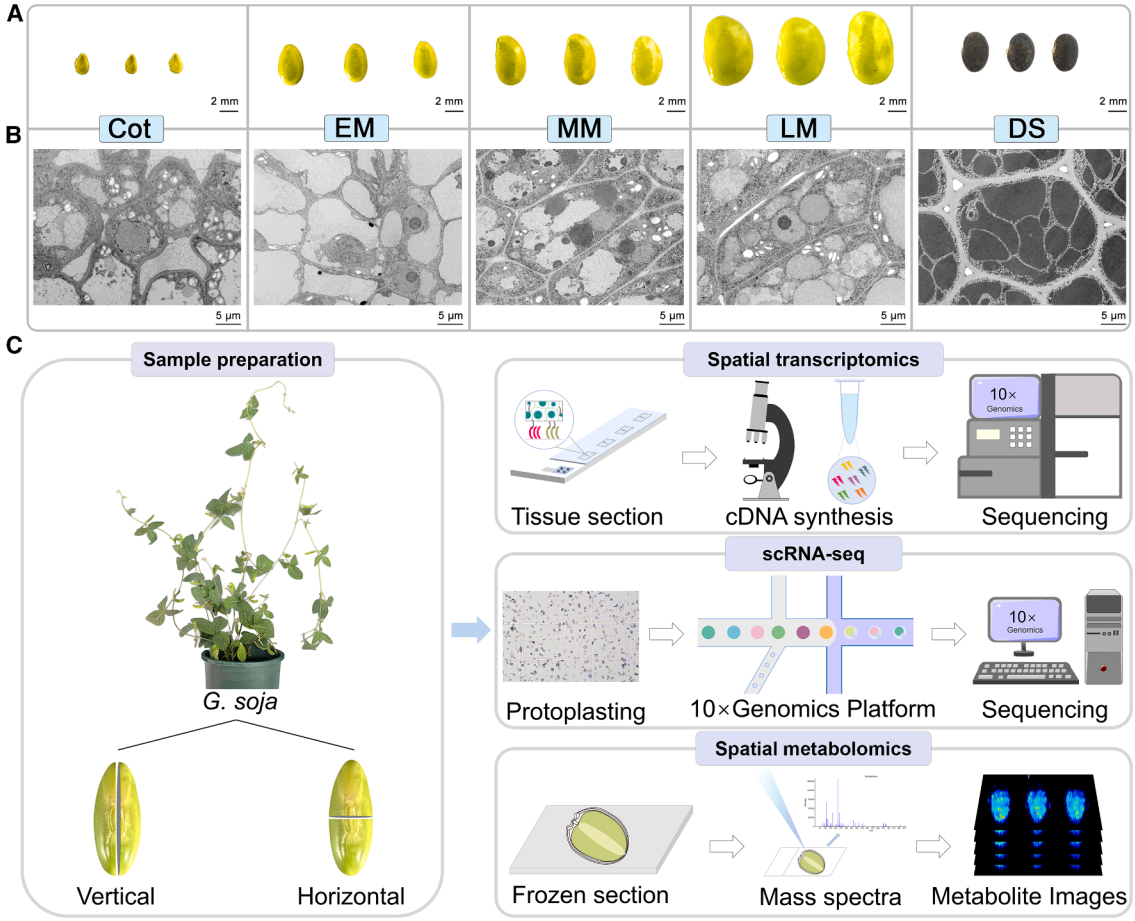


**Figure 1. *G. soja* seed development and overview of analyses performed in this study.**

**(A)** Morphological changes in *G. soja* seeds at different stages of seed development: the cotyledon primordium formation stage (Cot), early-maturity stage (EM), mid-maturity stage (MM), late-maturity stage (LM), and dry-seed stage (DS). Scale bars: 2 mm.

**(B)** Transmission electron microscopy images of *G. soja* seed development. Scale bars: 5 μm.

**(C)** Summary of data-acquisition workflows for spatial transcriptomics, scRNA-seq, and spatial metabolomics analyses. After paraffin embedding and sectioning, section-staining cell wall analyses were performed. *In situ* RNA capture was performed using sections loaded on spatial transcriptomics chips, followed by cDNA amplification, library construction, and sequencing. Seeds were also used for spatial metabolomics.

of stored lipids. During the cotyledonary stage, a single large vacuole occupied the majority of each cell (Figure 1B). At the early-maturity stage, cells began showing obvious protein-storage vacuoles, with a limited number of these electron-dense vacuoles evident on the periphery of individual cells. During the mid-maturity stage, active synthesis of oil bodies and protein-storage vacuoles was evident. By the late-maturity stage, oil-body synthesis was nearly complete. At the dry-seed stage, the seeds were mature, and their cells contained high levels of protein-storage vacuoles and oil bodies.

Isolated seeds at the mid-maturity stage were frozen, sectioned in two planes, and used for analyses. Spatial transcriptomics analyses were performed for spots 55 μm in diameter on two sections, one each in the horizontal and vertical planes. Seed cross-sections were prepared to 10-μm thickness, corresponding to one cell layer with the thickness of a single cell. The permeabilization of these tissue sections was optimized using *G. soja*–specific conditions (Supplemental Figures 1 and 2). Multiple selected sections were then stained with hematoxylin and eosin before analysis with the 10× Genomics Visium platform (Figure 1C). Spatial transcriptomics sequencing yielded 1961 and 1219 valid spots for the horizontal and vertical sections, respectively, with 36 890 and 38 225 corresponding valid genes (Supplemental Table 1). Distinct quality-control results were observed for these sections (Supplemental Figure 3). For scRNA-seq analysis of wild soybean, high-purity, viable protoplasts were isolated from seeds, and the 10× Genomics scRNA-seq platform was used for droplet-based scRNA-seq analyses (Figure 1C). Data were successfully generated for 29 515 cells, with a mean of 55 364 sequencing reads per cell and a median of 2924 expressed genes (Supplemental Table 2). Genes induced during the process of protoplast preparation were eliminated during quality control (Supplemental Table 3). After quality control, data from 20 672 cells were retained for subsequent analysis (Supplemental Figure 4C–4G).

In brief, we generated high-quality transcriptomic data from two wild soybean seed sections using spatial transcriptomics and from 20 672 soybean seed cells by scRNA-seq at the

mid-maturity stage, enabling the generation of a spatiotemporal transcriptomic map and characterization of dynamic gene-expression patterns in developing wild soybean seeds.

### Spatial clustering and cell-type classification of *G. soja* seeds

The spatial and histological characteristics of the seed samples were used to separate them into anatomically distinct regions: embryo parenchyma (Emb parenchyma), embryo axis (Emb axis), embryo epidermis (Emb epidermis), seed-coat epidermis (SC epidermis), seed-coat hilum (SC hilum), seed-coat vasculature and parenchyma (SC vasculature & parenchyma), and endosperm (Figure 2A). The cell-type composition of these samples was analyzed using Seurat to perform unsupervised clustering analyses based only on gene-expression data. Optimized resolution parameters were used to ensure correct clustering of the cell types, and the results were visualized using t-distributed stochastic neighbor embedding (t-SNE) plots (Becht et al., 2019). The analysis revealed 9 cell clusters in the horizontal section and 10 cell clusters in the vertical section (Figure 2B). After mapping these clusters onto the seed sections using the corresponding spatial data, the sections were used as a reference for cell-clustering analyses. Next, the spatial transcriptomic data were annotated as different cell types associated with different anatomical regions and sample planes on the basis of their spatial and histological features.

In the horizontal tissue section, the identified cells included H–Emb parenchyma-adaxial cells (cluster H_1), H–Emb parenchyma-abaxial1 cells (cluster H_2), H–Emb parenchyma-abaxial2 cells (cluster H_3), H–endosperm cells (cluster H_5), H–SC epidermis cells (cluster H_6), H–SC hilum cells (cluster H_7), H–SC vasculature & parenchyma1 cells (cluster H_4), H–SC vasculature & parenchyma2 cells (cluster H_8), and H–Emb epidermis cells (cluster H_9). The vertical tissue sections included V–Emb parenchyma1 cells (cluster V_1), V–Emb parenchyma2 cells (cluster V_2), V–Emb parenchyma3 cells (cluster V_3), V–Emb axis-hypocotyl cells (cluster V_4), V–Emb axis-plumule cells (cluster V_7), V–Emb epidermis cells (cluster V_5), V–SC hilum cells (cluster V_6), V–SC epidermis cells (cluster V_8), V–SC vasculature & parenchyma cells (cluster V_9), and V–endosperm cells (cluster V_10) (Figure 2C and Supplemental Table 4). The spatial transcriptomics data were then analyzed to identify potential marker genes that showed cell-type-specific expression patterns (Supplemental Table 5). The expression of tissue-specific marker genes was scored in accordance with previously published spatial transcriptomics identification of cultivated soybean in different cell clusters using the "AddModuleScore" function in the Seurat package to confirm the presence of the correct cell types in the different clusters; reasonable consistency was observed between the identified cell types and their spatial arrangements in the prepared sections (Zhang et al., 2025) (Figure 2D). RNA *in situ* hybridization analyses of representative marker genes confirmed the reliability of the *G. soja* cell populations and the identification of marker genes (Supplemental Figure 5).

### Cell-specific functions of Emb parenchyma cell types

Soybean embryos originate from apical cells produced by the division of fertilized eggs, with the cotyledons serving as storage organs for reserves that will nourish the germinated seedling (Danzer et al., 2015). To explore the similarities and differences between abaxial and adaxial Emb parenchyma cells, we analyzed differential gene expression using the spatial transcriptomics datasets (Figure 3A). This resulted in the identification of a total of 5979 differentially expressed genes (DEGs) (Supplemental Table 6). We performed Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of the abaxial and adaxial Emb parenchyma cell populations to assess associations of the DEGs with biological pathways in different cell types, confirming that the enriched pathways and terms were related to the cell types of interest (Figure 3B). For instance, most abaxial and adaxial Emb parenchyma cells showed enrichment in the same biological processes, although abaxial Emb parenchyma cells were also enriched in pathways involved in glutathione metabolism, lysine degradation, cysteine and methionine metabolism, lysine biosynthesis, and cyanoamino acid metabolism, all of which are associated with storage proteins (Supplemental Table 7). By contrast, the adaxial Emb parenchyma cells were also enriched in pathways associated with α-linolenic acid metabolism, glycerolipid metabolism, glycerophospholipid metabolism, linoleic acid metabolism, and biosynthesis of unsaturated fatty acids, all of which are related to lipid storage (Supplemental Table 8). Previous studies have shown that, during the mid-maturity stage, starch levels are two-fold lower and storage-protein deposits are larger in abaxial than adaxial tissues, whereas lipid deposition occurs predominantly in adaxial regions (Borisjuk et al., 2005).

To explore the dynamics of gene expression underlying cellular specialization in *G. soja* Emb parenchyma, we performed a subclustering analysis of Emb parenchyma cells from both the horizontal and vertical sections (Figure 3C and 3E). This subclustering yielded five subpopulations in horizontal Emb parenchyma cells (H-Emb parenchyma_1 to H-Emb parenchyma_5) and six subpopulations in vertical Emb parenchyma cells (V-Emb parenchyma_1 to V-Emb parenchyma_6). To further dissect the functional heterogeneity within these subpopulations, we performed gene set variation analysis (GSVA) to evaluate patterns of pathway activity across the subclusters (Figure 3D and 3F).

For the horizontal Emb parenchyma subclusters (Figure 3C), H-Emb parenchyma_3, which spatially corresponded to the adaxial region, showed prominent enrichment in pathways related to lipid metabolism and biosynthesis, including Carotenoid biosynthesis (ko00906), Sesquiterpenoid and triterpenoid biosynthesis (ko00909), and Glycerophospholipid metabolism (ko00564) (Supplemental Table 9). These are all core pathways that support lipid synthesis and accumulation, consistent with our earlier finding that adaxial regions have a tendency for lipid storage. By contrast, the remaining subpopulations (H-Emb parenchyma_1, H-Emb parenchyma_2, H-Emb parenchyma_4, and H-Emb parenchyma_5), which mapped to abaxial-like domains, exhibited heightened activity in pathways linked to protein metabolism, including Protein export (ko03060) and Protein processing in endoplasmic reticulum (ko04141). These pathways are critical for protein maturation and transport and align with the larger storage-protein deposits observed in abaxial tissues. For the vertical Emb parenchyma subclusters (Figure 3E), we analyzed pathway

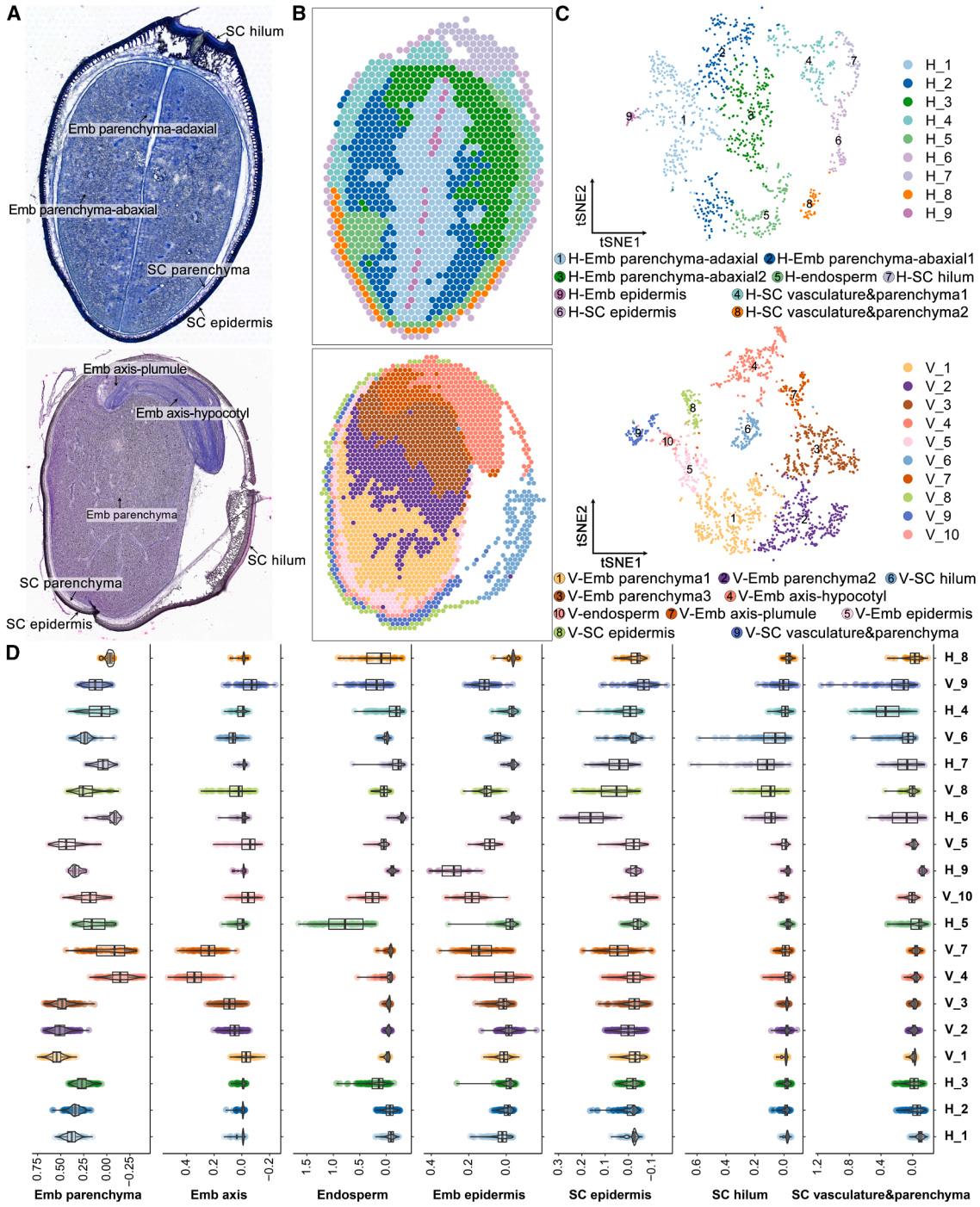


**Figure 2. Spatial transcriptomic analysis of cellular organization in *G. soja* seeds.**

**(A)** Toluidine blue–stained paraffin-embedded *G. soja* seeds were sectioned horizontally (top) and vertically (bottom).

**(B)** Cell clusters colored as in **(C)** mapped onto stained tissue sections.

**(C)** t-SNE plots showing spatial transcriptomics–based cell clusters represented in different colors.

**(D)** Scoring of different types of cell marker genes in each cluster.

activities by spatial location: V-Emb parenchyma_1 (upper region) showed heightened activity in protein metabolism-related pathways, including Ribosome (ko03010), RNA transport (ko03013), and Biosynthesis of amino acids (ko01230). These pathway enrichments confirm its bias toward protein-related functions. The middle-region subclusters (V-Emb parenchyma_2, V-Emb parenchyma_5, and V-Emb parenchyma_6) exhibited a transitional profile: V-Emb parenchyma_2 and V-Emb parenchyma_6 were enriched in lipid pathways, including Glycerolipid metabolism (ko00561) and Biosynthesis of unsaturated fatty acids (ko01040), whereas V-Emb parenchyma_5 was enriched in aspects of protein processing,

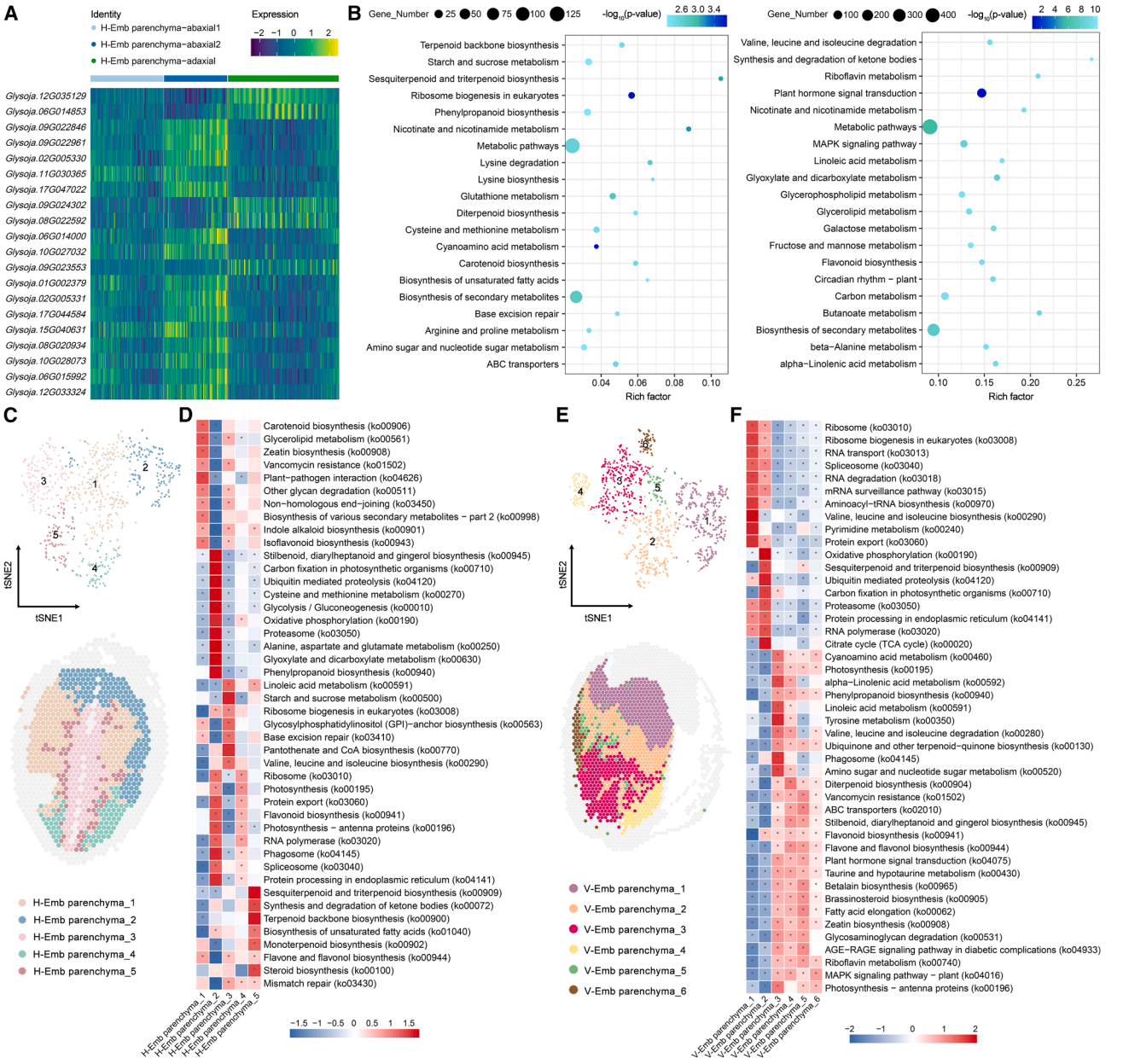


**Figure 3. Reconstruction of the cellular organization of seeds on the basis of spatial transcriptomics.**

**(A)** Heatmap showing the expression levels of differentially expressed genes (DEGs) in the Emb parenchyma-abaxial and Emb parenchyma-adaxial cell types. The color gradient indicates the level of gene expression.

**(B)** KEGG enrichment analysis of DEGs between cotyledon cell populations. Left: upregulated DEGs enriched in the Emb parenchyma-abaxial cells. Right: downregulated DEGs enriched in the Emb parenchyma-adaxial cells.

**(C)** Top: t-SNE plots showing Emb parenchyma-adaxial cell subclusters. Bottom: subclusters mapped to their spatial locations.

**(D)** Heatmap of GSVA scores for Emb parenchyma-adaxial cell subclusters. Asterisks (*) indicate gene sets with $P < 0.05$ as determined by limma analysis.

**(E)** Top: t-SNE plots showing Emb parenchyma-abaxial cell subclusters. Bottom: subclusters mapped to their spatial locations.

**(F)** Heatmap of GSVA scores for Emb parenchyma-adaxial cell subclusters. *$P < 0.05$.

including Protein processing in endoplasmic reticulum (ko04141) and Aminoacyl-tRNA biosynthesis (ko00970). The lower-region subclusters (V-Emb parenchyma_3 and V-Emb parenchyma_4) displayed prominent enrichment in lipid biosynthesis pathways, including Linoleic acid metabolism (ko00591), Glycerophospholipid metabolism (ko00564), and Sphingolipid metabolism (ko00600) (Supplemental Table 10), confirming their tendency toward lipid storage.

Collectively, the Emb parenchyma cells in *G. soja* exhibit sophisticated spatial differentiation in function: in the horizontal Emb parenchyma cells, adaxial Emb parenchyma cells mainly undertake functions related to lipid synthesis and accumulation, whereas abaxial Emb parenchyma cells focus primarily on functions associated with protein synthesis and storage; the functional distribution of vertical Emb parenchyma cells forms a clear gradient, in which cells in the upper regions are dominated

by protein-related metabolic functions, cells in the lower regions are dominated by lipid-related metabolic functions, and cells in the middle regions integrate both of these metabolic processes. This spatial partitioning ensures the efficient allocation of metabolic resources, supporting the accumulation of nutrients in developing wild soybean seeds.

## Pseudotime analysis reconstructs a continuum of functional states in Emb parenchyma cells

To clarify the transcriptional dynamics that accompany the functional specialization of Emb parenchyma cells, we performed RNA velocity and pseudotime trajectory analyses. We used scVelo, a likelihood-based dynamic model, to estimate the latent time of individual cells (La Manno et al., 2018); RNA velocity analysis predicted a predominant direction of transcriptional change originating from clusters H-Emb parenchyma_2 and V-Emb parenchyma_2 (Figure 4A and 4B). This prediction was corroborated by CytoTRACE analysis, which indicated that cells within the H-Emb parenchyma_2 and V-Emb parenchyma_2 clusters exhibited transcriptional signatures of a less differentiated state, suggesting that they represent a putative early state in the continuum of Emb parenchyma cell differentiation.

We next constructed pseudotime trajectories to model the progression of functional states. For horizontal Emb parenchyma cells, the pseudotemporal trajectory exhibited a directional trend in gene expression, starting from a state enriched in abaxial Emb parenchyma genes and progressing toward a state enriched in adaxial Emb parenchyma genes (Figure 4C). This pseudotemporal ordering of transcriptional states directly reflects the known functional difference between abaxial and adaxial Emb parenchyma cells, which have been shown to undergo endoreduplication asynchronously (Lin et al., 2017). This pseudotemporal progression aligns with the spatial gradient of function we observed and was associated with dynamic changes in gene expression patterns (Supplemental Figure 6B). The pseudotime trajectory contained one branchpoint, dividing the cells into three distinct states (Supplemental Figure 6C). We identified four gene expression modules along this trajectory (Figure 4D and Supplemental Table 11). Genes in module 3, enriched at the trajectory origin, were significantly associated with the Gene Ontology (GO) terms cytoplasmic translation (GO:0002181), translational elongation (GO:0006414), and ribosomal small subunit assembly (GO:0042274) (Figure 4E), consistent with the pronounced protein synthesis and storage activities of abaxial cells. By contrast, genes in module 1, enriched at the trajectory endpoint, were associated with terms like modification-dependent protein catabolic process (GO:0019941), cellular protein modification process (GO:0036211), and lignin biosynthetic process (GO:0009809), suggesting a shift toward lipid metabolism and tissue maturation in adaxial cells.

The visualized transcriptional profiles of the vertical Emb parenchyma cells were then mapped along pseudotime trajectories (Figure 4F). A comparative analysis revealed that the density distribution of the cell states differed among the various groups, which may reflect dynamic changes during the acquisition of distinct functional states by vertical Emb parenchyma cells (Supplemental Figure 6E and 6F). To evaluate the processes involved in the functional diversification of vertical Emb parenchyma cells, we constructed a heatmap to illustrate the expression patterns of the DEGs, which were divided into four modules (Figure 4G and Supplemental Table 12). As expected, the genes in module 1 were significantly enriched in GO terms associated with seed maturation (GO:0010431) and lipid transport (GO:0006869), both of which are closely related to nutrient storage. The genes in module 4 were significantly enriched in the GO terms protein maturation (GO:0051604) and chitin catabolic process (GO:0006032), which are closely related to protein synthesis in soybean Emb parenchyma cells (Figure 4H).

Together, the pseudotime trajectories, discrete gene modules, and GO enrichment analyses demonstrate that Emb parenchyma cells exist in a continuum of transcriptional states that correspond to spatiotemporally distinct functional programs. This transcriptional gradient aligns with their spatial functional gradients and provides a foundation for deciphering the mechanisms underlying their functional specialization. Notably, many transcription factors also displayed branch-dependent expression patterns across the pseudotemporal continuum. These transcription factors are important regulatory factors that may be involved in regulating the establishment of distinct Emb parenchyma cell states (Supplemental Figure 6G and 6H and Supplemental Tables 13 and 14).

## Characterization of the regulation of wild soybean seed development by analysis of cell–cell communication

We next used the integrated spatial data from the horizontal and vertical sections to investigate the cellular and molecular dynamics involved in seed development. Unsupervised clustering analysis identified similar cell compositions in both sections (Figure 5B). The identified cell types were used to construct horizontal and vertical structural diagrams of soybean seed sections (Figure 5A). Integration and clustering of the same cell types were seen in both the horizontal and vertical sections of soybean seeds. To explore the cell-specific functions of different seed cell types, we performed GO enrichment analysis (Figure 5C). Cotyledon cells were enriched in processes related to nutrient accumulation during seed development, including seed maturation and protein folding (Tulpan et al., 2015; Yamasaki et al., 2017). Axis cells were enriched in processes and cellular compartments associated with growth, development, and energy conversion, including the cytosol, translation initiation factor activity, and helicase activity (Zhang and Fernie, 2018; Fatahi et al., 2024). Seed-coat cells were enriched in activities linked to energy production and nutrient transport, including GTPase activity, the tricarboxylic acid cycle, and structural molecule activity (Zeng et al., 2017). GO terms related to "GTP binding," "cytosol," and "protein folding" were enriched in most of the cell types. This may indicate that they provide global support for key processes such as cell differentiation during seed development by facilitating various activities, including signal coordination and substance synthesis.

We next investigated potential cell–cell communications between the various cell populations during the mid-maturity stage.

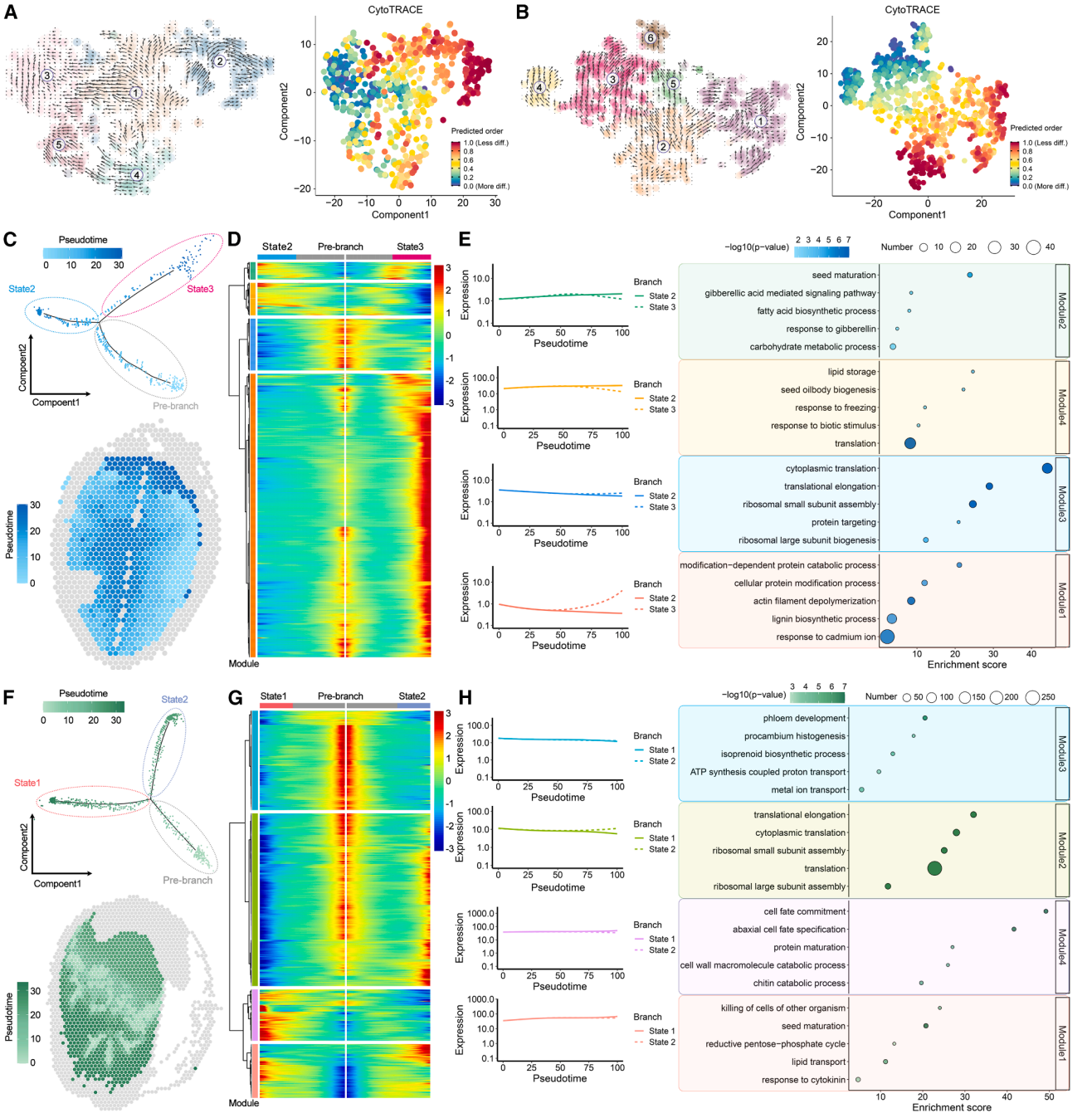


**Figure 4. Simulation of the pseudotemporal trajectory of Emb parenchyma cells and analysis of gene expression patterns.**

**(A)** RNA velocity (left) and CytoTRACE (right) visualization for horizontal cotyledon cells as streams in the t-SNE space.

**(B)** RNA velocity (left) and CytoTRACE (right) visualization for vertical cotyledon cells as streams in the t-SNE space.

**(C)** Pseudotime analysis of horizontal cotyledon cellular differentiation, with dots representing individual cells.

**(D)** Heatmaps of gene expression and representative gene clustering in the pseudotime analyses.

**(E)** Gene expression trends with cell branching based on differentiation fate outcomes. Significantly enriched GO terms associated with specific clusters are shown on the right.

**(F)** Pseudotime analysis of vertical cotyledon cellular differentiation, with dots representing individual cells.

**(G)** Heatmaps of gene expression and representative gene clustering in the pseudotime analyses.

**(H)** Gene expression trends with cell branching based on differentiation fate outcomes. Significantly enriched GO terms associated with specific clusters are shown on the right.

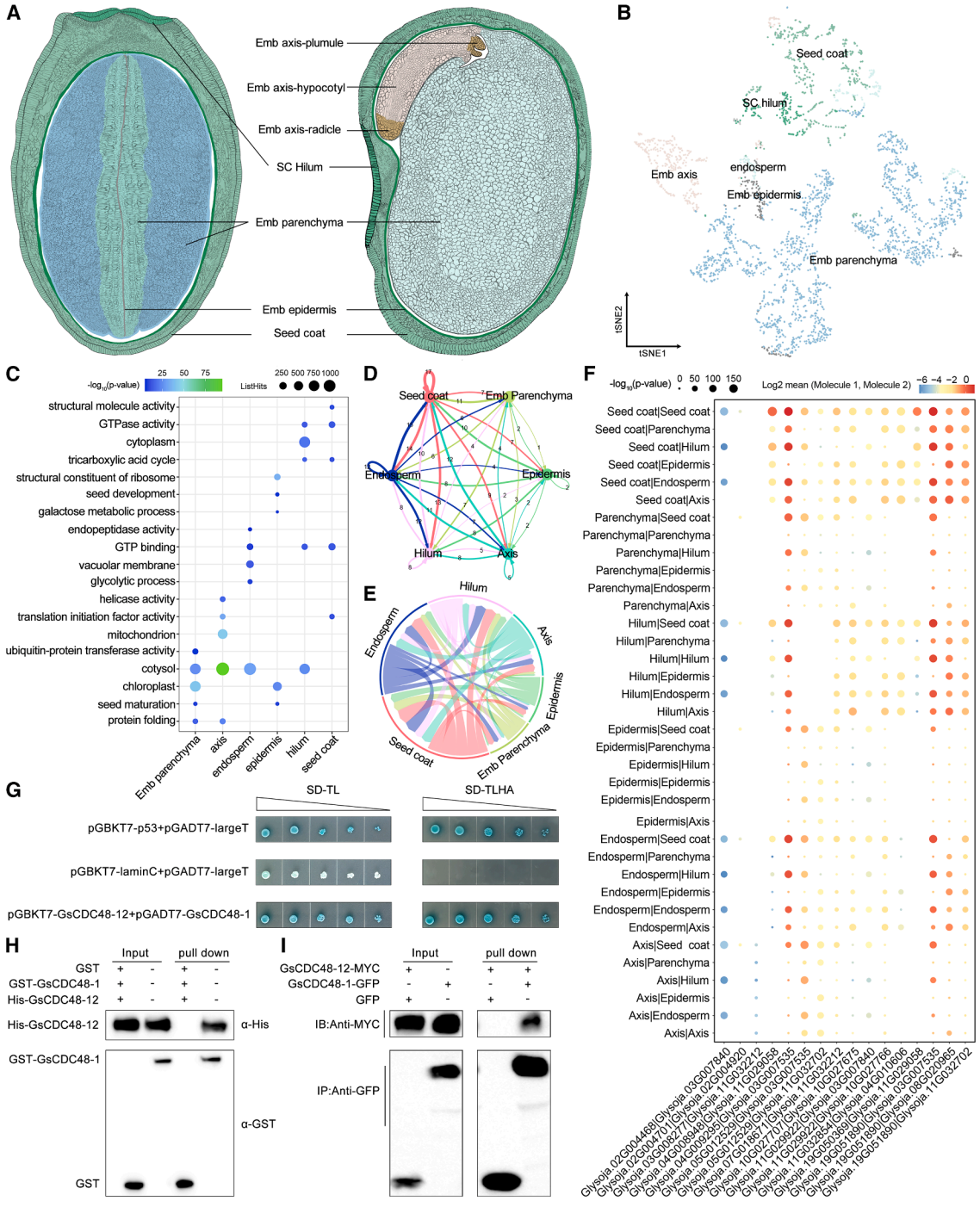


**Figure 5. Seed regulation based on ligand–receptor analysis.**

**(A)** *G. soja* seed anatomy, with different colors representing different cell types.

**(B)** t-SNE plots highlighting the cell types detected in *G. soja* seeds. Dots correspond to individual cells, colored and labeled as in **(A)**.

**(C)** GO enrichment analysis of genes specifically expressed in individual cell types. Dot size represents the number of genes, and dot color denotes the adjusted *P* value in the related pathway.

**(D)** Network plots showing ligand–receptor interaction events in seed cells. Cell–cell communication is indicated by connecting lines. Line thickness and node size indicate the number of ligand–receptor interactions.

**(E)** Chord diagram of cell–cell communication between pairs of cell types. Line width indicates the number of significant ligand–receptor pairs.

**(F)** Top ligand–receptor pairs with $p < 0.05$ show different ligand-receptor interaction strengths. Columns are scaled by maximum ligand–receptor expression.

**(G)** Yeast two-hybrid assay confirming the protein–protein interaction between GsCDC48-12 and GsCDC48-1. Serial dilutions and growth on selective medium demonstrate the strength of the interaction.

**(H)** GST pull-down assay showing the physical interaction between GsCDC48-12 and GsCDC48-1.

**(I)** Co-IP assay showing the interaction between GsCDC48-12 and GsCDC48-1.

Ligand–receptor enumeration was performed to analyze the interactions between each pair of cell types (Figure 5D). The high number of interactions in seed coats suggested active communication with other cell types. Notably, in terms of both ligands and receptors, the seed coats ranked highest in number of interactions, implying an auto- or intercellular manner of regulation. These results imply that seed-coat cells act as regulatory hubs in cell–cell communication networks. Unexpectedly, the number of interactions shown by endosperm cells was relatively high, despite the lower number of these cells. We next constructed a cell–cell interaction map by correlating ligands to their corresponding receptors in seed cells (Figure 5E). Analyses of the networks suggested that the seed-coat cells may act as potential central hubs in cell–cell communication networks, given their high frequency of predicted ligand–receptor interactions, potentially regulating key processes during mid-maturity seed development. We then focused our analysis on the top 10 ligand–receptor pairs, ranked according to the average scoring method, as these interactions are likely to play important roles in cell–cell communication (Figure 5F). Notably, a number of ligand–receptor pairs were detected in most cell–cell pairs, including Glysoja.19G051890 (*GsCDC48-12*)–Glysoja.03G007535 (*GsCDC48-1*), Glysoja.04G009295 (14-3-3-LIKE PROTEIN GF14 OMEGA)–Glysoja.03G007535 (*GsCDC48-1*), and Glysoja.19G051890 (*GsCDC48-12*)–Glysoja.08G020965 (ribosome biogenesis ATPase). Among the ligand–receptor pairs, the GsCDC48 protein–GsCDC48 protein interaction was the most frequently observed.

We next focused on confirming the interaction between GsCDC48-12 and GsCDC48-1, two CDC48 family members that exhibited the highest expression levels and ligand–receptor interaction scores in our dataset. Yeast cells co-expressing GsCDC48-12 with GsCDC48-1 exhibited robust growth and distinct blue coloration on SD/−Leu/−Trp/−His/−Ade/X-α-Gal medium, indicating effective interaction, unlike the negative controls (Figure 5G). Consistent with this finding, co-immunoprecipitation (co-IP) and GST pull-down assays also provided evidence for the physical association between GsCDC48-12 and GsCDC48-1, reinforcing the specificity of this interaction (Figures 5H and 5I). These comprehensive *in vitro* and *in vivo* biochemical assays systematically demonstrate that GsCDC48-12 interacts with and modulates GsCDC48-1. Given that CDC48 complexes are known to participate in cell division, plant growth, and various cellular processes (Bae et al., 2009; Inès et al., 2024; Nicolas-Francès et al., 2024), our data suggest a model in which CDC48-mediated intercellular communication may play a role in coordinating seed development. Notably, the physical interaction between GsCDC48-12 and GsCDC48-1 and the high ligand–receptor interaction frequency of seed-coat cells provide a foundation for investigating intercellular communication mechanisms during mid-maturity seed development, with their specific regulatory roles in tissue differentiation and nutrient accumulation requiring further functional verification.

### Correlations between spatial transcriptomics and spatial metabolomics data

An airflow-assisted desorption electrospray ionization–mass spectrometry imaging (AFADESI–MSI) approach was used to map the spatial distribution of metabolites detected in *G. soja* seeds (Figure 6A). Representative overall average mass spectra in negative-ionization mode are presented in Supplemental Figure 7. We identified 503 metabolites, the largest proportions of which were lipids (13.3%), amino acids and derivatives (11.5%), and carbohydrates and derivatives (11.5%) (Supplemental Table 15). The six most enriched metabolic pathways were associated with 67 lipids and 58 amino acids and their derivative metabolites (Figure 6B and Supplemental Table 16), suggesting that they are critical to soybean quality traits. We identified numerous abundant metabolites during the mid-maturity stage of seed development, such as L-glutamine, L-proline, and (R)-2-methylmalate, all of which were highly accumulated in cotyledon tissues. We next performed integrated analyses based on the joint point-to-point spatial transcriptomics and spatial metabolomics datasets. The workflow included three major steps: (1) alignment of the metabolite and transcriptome data, (2) image transformation, and (3) dimensionality reduction by weighted nearest-neighbor integration and uniform manifold approximation and projection (Supplemental Figure 8) (Becht et al., 2019; Ravi et al., 2022). Genes related to these metabolites were also identified (Supplemental Table 17). To dissect regulatory links between transcripts and metabolites, we prioritized the 36 metabolites associated with the top six enriched pathways mentioned above, analyzed pairwise correlations between these metabolites and highly variable genes (Figure 6C), and visualized the spatial distribution of these metabolites (Figure 6D).

The 20 most highly variable genes were subjected to detailed correlation analysis with the 36 metabolites, and a correlation heatmap was generated (Figure 7A). On the basis of the functions of strongly correlated metabolites in this heatmap, we selected *Glysoja.16G043314* and its associated metabolite L-proline. L-proline has a number of functions, including the maintenance of redox homeostasis under stressful conditions, control of permeability, and promotion of protein synthesis and structural stability (Szabados and Savouré, 2010; Bach and Takagi, 2013; Lee et al., 2013; Meena et al., 2019). *Glysoja.16G043314* encodes the MAPK23-4 mitogen-activated protein kinase (MAPK), which plays important roles in many aspects of plant growth, development, and environmental stress responses (Wu and Wang, 2024). However, it has not previously been associated with protein accumulation in soybean seeds.

### Functional verification of *GsMAPK23-4*

To characterize the potential role of *GsMAPK23-4*, its coding sequence was fused in-frame with green fluorescent protein (GFP) (hereafter GsMAPK23-4–GFP). The resulting construct and a free-GFP positive-control construct were transiently expressed in *Arabidopsis* mesophyll protoplasts, and GsMAPK23-4–GFP localized to the cytosol (Figure 7B).

To confirm the identification and function of *GsMAPK23-4*, we generated stable transgenic lines in which GsMAPK23-4–GFP was overexpressed from the ubiquitous and strong CaMV *35S* promoter, and the *GsMAPK23-4* ortholog was knocked out by CRISPR–Cas9 gene editing in the DN50 *G. max* cultivar (Figure 7C–7F and Supplemental Figures 9, 10A, and 10B). The mutant alleles (designated *gsmapk23-4-1* to *gsmapk23-4-6*) encoded truncated proteins that varied in length from 97 to 213

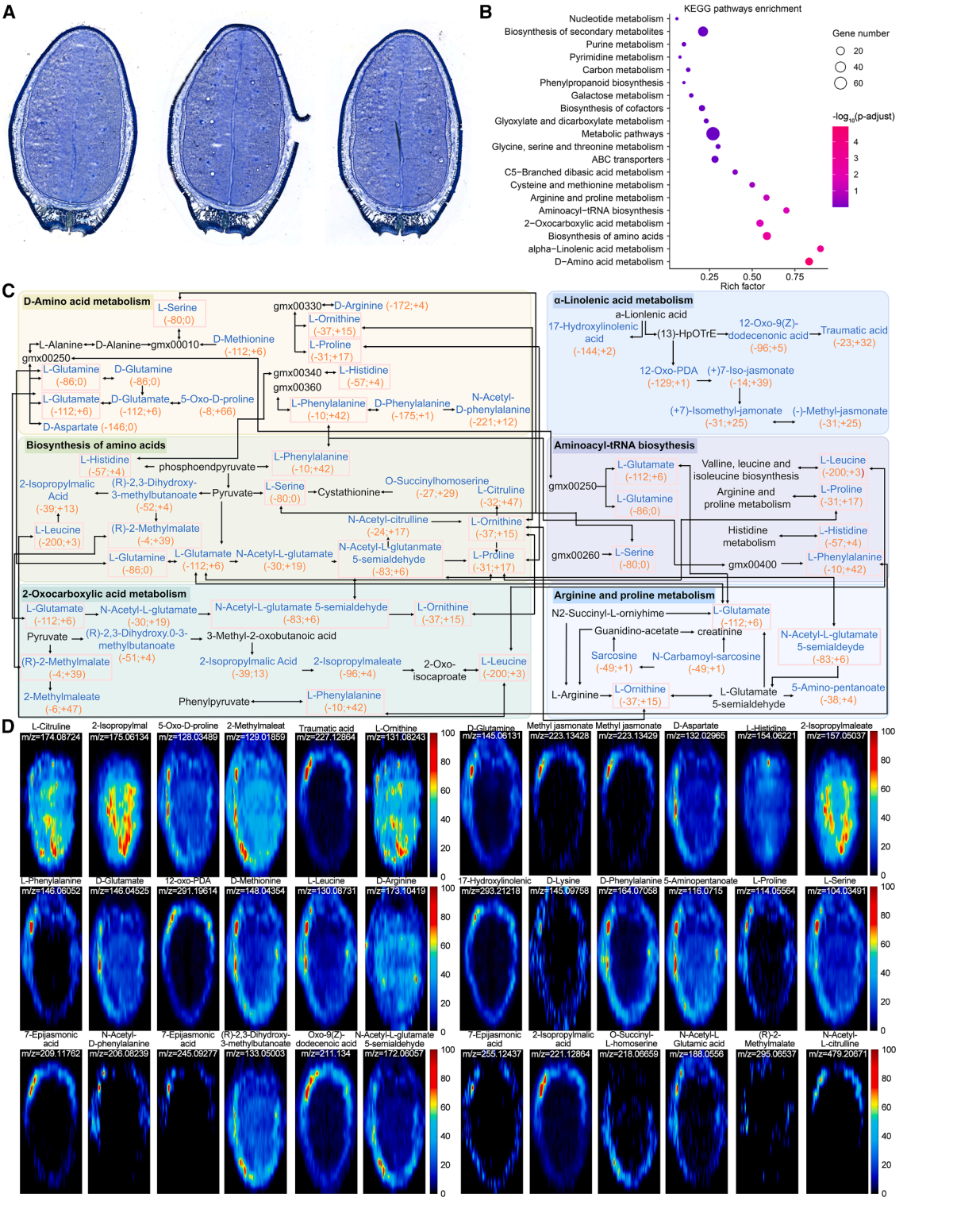


**Figure 6. Integration of spatial transcriptomics and spatial metabolomics analyses.**

**(A)** Hematoxylin and eosin staining of *G. soja* seeds.

**(B)** KEGG enrichment of metabolites identified in *G. soja* seeds.

**(C)** Metabolite identification and associated genes in the enriched metabolic pathways. Transformations or metabolic processes are marked with arrows. +, positive correlation; −, negative correlation.

**(D)** Spatial distributions of representative metabolites as mapped by AFADESI–MSI. Relative abundance is shown from blue (low) to red (high).

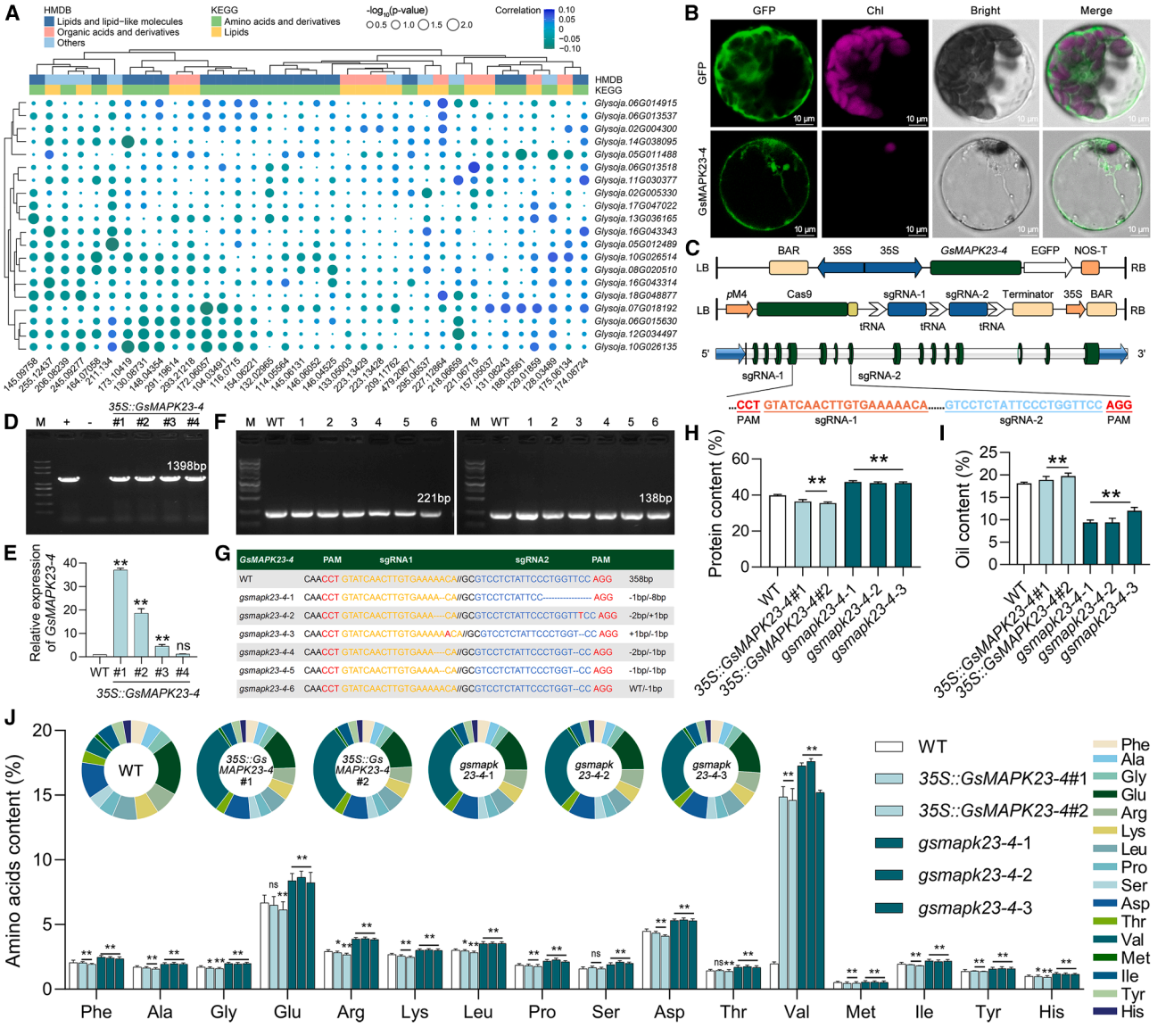


**Figure 7. Phenotypic analyses of seed oil and protein levels in mutant and overexpression lines.**

**(A)** Heatmap of the expression of genes and metabolites.

**(B)** Confocal microscopy images showing the subcellular localization of GsMAPK23-4–GFP upon transient expression in *Arabidopsis* protoplasts. "GFP" denotes the free-GFP control. Scale bars: 10 μm.

**(C)** Top: schematic of the pSOY1-*GsMAPK23-4* vector. Bottom: *GsMAPK23-4* gene structure and single-guide RNA (sgRNA) target sites used for CRISPR-Cas9 gene editing.

**(D)** PCR detection of *GsMAPK23-4*-overexpressing transgenic lines. M, Trans2K Plus DNA marker; +, positive control of the plasmid used for transformation; −, negative control.

**(E)** *GsMAPK23-4* expression in the transgenic overexpression lines as determined by RT–qPCR.

**(F)** PCR analysis of the CRISPR-edited lines (*gsmapk23-4*-1 to *gsmapk23-4*-6). Left: sgRNA 1. Right: sgRNA 2. M, DNA marker; WT, wild-type DN50.

**(G)** Sanger sequencing confirmation of gene-editing outcomes in the mutant lines after targeting *GsMAPK23-4 with* sgRNA 1 and sgRNA 2. Insertions and deletions are indicated with red letters and blue dashes, respectively. Sequence alterations relative to the wild-type sequence for individual genes are annotated at right.

**(H and I)** Protein **(H)** and oil **(I)** contents of seeds from the wild type (WT), transgenic *GsMAPK23-4*-overexpressing lines 1 and 2 (*35S::GsMAPK23-4*#1 and *35S::GsMAPK23-4*#2), and three mutant alleles (*gsmapk23-4*-1 through *gsmapk23-4*-3).

**(J)** Amino acid composition of seeds from wild-type DN50 (WT), transgenic *GsMAPK23-4*-overexpressing lines 1 and 2 (*35S::GsMAPK23-4*#1, *35S::GsMAPK23-4*#2) and three mutant alleles (*gsmapk23-4*-1 through *gsmapk23-4*-3).

Values in **(E) and (H)–(J)** are presented as the mean ± SD ($n$ = 3). ns, not significant, $P > 0.05$; $^{*}P < 0.05$ and $^{**}P < 0.01$; one-way ANOVA followed by Dunnett's test.

amino acids, compared with 451 residues for the wild-type protein (Figure 7G). Two independent transgenic overexpression lines and three knockout lines were selected for functional investigation. Protein levels were significantly higher in mature seeds of the knockout lines and significantly lower in those of the *GsMAPK23-4*-overexpressing lines compared with the wild type (Figure 7H and Supplemental Figure 10C). Conversely, oil levels were lower in the three knockout lines and higher in the overexpression lines compared with the wild type (Figure 7I and Supplemental Figure 10D). Overexpression and knockout had no obvious effects on key agronomic traits or overall developmental phenotype (Supplemental Figure 11).

We next analyzed levels of 16 hydrolyzed amino acids in seeds from these lines. Relative to wild-type DN50, the *gsmapk23-4-1 to gsmapk23-4-3 mutants showed* a statistically significant increase in amino acid content, whereas the contents of most protein-associated amino acids were significantly reduced in the *GsMAPK23-4*-overexpressing seeds (Figure 7J). Together, these data highlight a major role for *GsMAPK23-4* in protein and oil accumulation within soybean seeds.

### Spatial annotation of single-cell clusters in developing *G. soja* seeds

Principal-component analysis (PCA) and graph-based unsupervised clustering without known markers led to the identification of 14 distinct cell clusters. The clusters were visualized with 2D t-SNE plots, with all clusters seen in two biological replicates (Supplemental Figures 4A, 4B, 12A, and 12B). In general, compartments with the greatest degree of overlap in terms of gene expression are expected to be more closely associated with one another and may share functional characteristics (Qin et al., 2016). Hierarchical clustering of Pearson correlation coefficients revealed clear correlations between compartments from closely adjacent anatomical tissues (Supplemental Figure 12C).

Using the spatial transcriptomics sequencing data from seed sections and the scRNA-seq data from 20 672 *G. soja* cells at the mid-maturity stage, we developed a spatiotemporal transcriptomic map illustrating gene expression patterns in developing soybean seeds. The overall correlation coefficient between the scRNA-seq and the spatial transcriptomics datasets was close to 0.65, likely owing to effects of the enzymatic digestion process necessary for protoplast preparation (Supplemental Figure 4H). As there are relatively few marker genes in *G. soja* seeds, the successful assignment of several cell clusters was not possible. To overcome this issue, robust cell-type decomposition (RCTD) was used to monitor spatial patterns of gene expression. Cell-type annotations were performed for all cell clusters in the spatial sections, with a darker red color denoting a higher proportion of the corresponding cell type (Supplemental Figure 12D). Expression levels of known marker genes were also analyzed in the scRNA-seq dataset to facilitate the association of specific clusters with corresponding cell types (Supplemental Figure 12E). For instance, KTi3 and KTi1 (Glysoja.08G022501 and Glysoja.01G000984), which were uniquely expressed in embryos, were expressed at high levels in the cotyledon region (Perez-Grau and Goldberg, 1989), while β-conglycinin Gy4 (Glysoja.10G025927) is also reportedly expressed at the mid-maturity stage and was highly expressed in the cotyledon region (Harada et al., 1989). Moreover, flavonoid 3′-hydroxylase, which corresponds to the T locus and is responsible for controlling pubescence and seed coat color, was highly expressed in SC_AL2 (Toda et al., 2012). Specific expression of the GPI-anchored lipid transport protein LTPG1 (Glysoja.13G037184) was also observed in SC_AL. Consistent expression of these marker genes was also noted in different spatial sections. This integrated spatial transcriptomics and scRNA-seq analysis facilitated the successful identification of cell types in *G. soja* seeds and the preparation of a comprehensive seed-cell molecular map.

## DISCUSSION

Our study provides a cell-level perspective on seed development in wild soybean, revealing how spatial organization and transcriptional trajectories contribute to oil and protein accumulation. Wild soybean, the ancestor of cultivated soybean, has been proposed as a useful genetic source for increasing these components in soybean crops (Du et al., 2023). Despite thorough descriptions of the organization and structures of seed cells in cultivated soybean, there has been no comprehensive analysis of how cells differentiate in wild soybean seeds. Here, we constructed and compared the developmental trajectories of cells during development of *G. soja* seeds to generate a cellular molecular map. The findings provide new insight into the processes involved in cellular differentiation during soybean seed development.

Fully characterizing the development of soybean seeds requires a thorough understanding of the changes in gene expression that take place in specific cell populations during this process. Here, we used spatial transcriptomics to generate cellular maps and molecular networks associated with the differentiation trajectories of cells in *G. soja* seeds. Horizontal and vertical sections were used for spatial transcriptomic sequencing with the 10× Visium platform, facilitating analysis of the spatial patterns of gene expression in seeds while also clarifying the cellular composition of seed tissues.

Previous soybean spatial transcriptomics data, together with the spatial transcriptomics data obtained in this study, were used to annotate the different cell types. The spatial location of each cell type was successfully determined, enabling us to characterize the cell composition of wild soybean seeds. RNA *in situ* hybridization was used to verify marker genes specific to the different clusters, and the results were consistent with those of the comprehensive bioinformatics analysis (Supplemental Figure 5). In summary, this study successfully classified the various cell types found in soybean seeds, providing a data resource that will contribute to understanding the regulatory network of seed development at the cellular level.

Horizontal Emb parenchyma cells display distinct functional specializations: abaxial cells are enriched in pathways for protein storage, whereas adaxial cells prioritize lipid biosynthesis. This spatial divergence in gene expression aligns with the metabolic gradients observed in mid-maturity seeds in previous studies, in which abaxial tissues accumulated more storage proteins and adaxial regions more lipids. The functional specialization of Emb parenchyma cells follows distinct spatiotemporal

trajectories governed by modular gene expression, providing a molecular basis for their functional diversification in nutrient accumulation. Together, our pseudotime analyses define a continuum of transcriptional states in Emb parenchyma cells that underlies their spatial gradient in function. Although our study and prior work (Lin et al., 2017) have established clear molecular and physiological distinctions between adaxial and abaxial parenchyma cells, the precise developmental lineage relationships between these cell types, as highlighted in legume seed biology (Le et al., 2007), remain an open question. Our data do not directly address cellular origins; rather, they provide a high-resolution map of the transcriptional programs associated with functional specialization. The transcription factors we identified as varying across this pseudotemporal continuum represent candidate regulators for the establishment and maintenance of these distinct cell states. Future studies using lineage-tracing techniques will be essential to determine how these transcriptional states are linked to the embryonic development of parenchyma cell lineages. In particular, a pseudotime analysis of cells in the Emb parenchyma led to further refinement of Emb parenchyma cell types and investigation of their functional roles. The results provide insight into the gene expression patterns of individual cell types involved in various cellular and molecular processes, including seed growth and development, energy storage, and nutrient accumulation. These spatial transcriptomic profiles lay a foundation for more detailed exploration of metabolite–transcript regulatory networks and hypothesis-led investigations of the functions of metabolites, tissues, and cell types in seed development.

Spatial transcriptomic analyses and predictions of ligand–receptor interactions suggested that seed-coat cells may act as central regulatory hubs in intercellular communication (with the highest inferred interaction frequency) and revealed that endosperm cells show unexpectedly high signaling activity despite their low abundance. These observations provide testable hypotheses for tissue regulatory roles and lay the groundwork for validating communication mechanisms during the mid-maturity stage of seed development. The CDC48 protein complex emerged as a key mediator, and the physical interaction between GsCDC48-12 and GsCDC48-1 was experimentally validated by yeast two-hybrid, GST pull-down, and co-IP assays. Notably, although CDC48 is known for intracellular roles in protein degradation and cell-cycle regulation (Bae et al., 2009; Nicolas-Francès et al., 2024), our identification of its intercellular interaction suggests that it may act as a communication module linking the seed coat to other tissues. This hypothesis, supported by the role of the seed coat as a communication hub, provides a foundation for investigating how intercellular communication coordinates tissue differentiation and nutrient accumulation in seeds. Overall, existing evidence supports the idea that CDC48 may be involved in coordinating these seed developmental processes, but its regulatory role *in planta,* including whether it directly mediates seed coat–embryo signaling, requires further functional verification.

Seed development proceeds through three main stages: (1) embryo growth, cellular division, and morphogenesis; (2) seed maturation and the accumulation of reserve compounds; and (3) desiccation and the onset of dormancy (Weber et al., 2005). Soybean seeds contain large amounts of lipids, starch, and amino acids, which serve as a source of nutrition for the seeds. Amino acids and lipids play roles in the regulation of soybean seed development and the accumulation of nutrients (Gupta et al., 2017). The seed yields and oil contents of cultivated soybean varieties tend to be higher than those of ancestral wild varieties, but their protein contents tend to be lower (Sedivy et al., 2017). Here, we performed transcriptomics and metabolomics analyses of wild soybean seeds at the mid-maturity stage using an AFADESI–MSI approach to gain comprehensive insights into the key metabolites that are most abundant during seed development. Metabolic pathway enrichment analyses indicated that lipids, amino acids, and their derivatives were present at high levels in soybean seeds at mid-maturity. Further identification of these metabolites at a higher spatial resolution and identification of genes that may govern the course of seed development will help to generate a new understanding of metabolism at the tissue and even the single-cell level, enabling improvements in the nutritional quality of soybean seeds.

Soybean is a nutrient-rich crop that contains all the amino acids essential for the human diet (Kudełka et al., 2021). Further improvement of cultivated soybean varieties in terms of protein and amino acid contents is an important goal in soybean breeding, and the abundant genetic diversity of wild soybean offers significant possibilities for such improvement. Analyses of correlations between spatial transcriptomic and spatial metabolomic data identified *GsMAPK23-4* (Glysoja.16G043314), which was associated with L-proline content in *G. soja* seeds. As conserved protein kinases expressed across eukaryotic species, MAPKs are key regulators of plant development and environmental stress responses (Cristina et al., 2010; Manna et al., 2023). *NaMAPK4* has a positive effect on vegetative growth but negatively affects reproductive growth (Hettenhausen et al., 2011). MAPKs have been shown to regulate gene expression in *Arabidopsis* following exposure to abiotic or biotic stressors (Bigeard and Hirt, 2018). To characterize the function of *GsMAPK23-4*, we generated overexpression and knockout lines. Compared with wild-type DN50, *gsmapk23-4* mutant plants had a higher protein content, as well as increased contents of proline, threonine, and other amino acids, but their oil content was lower. *GsMAPK23-4*-overexpressing plants showed the opposite characteristics. The phenotypes of the mutants and overexpression plants thus suggest that *GsMAPK23-4* is a negative regulator of soybean protein content, providing a reference for future efforts to enhance the nutritional content of soybean seeds. The gene and associated metabolites identified here may be linked to the regulation of the developmental processes active in these seeds and may influence the quality of different soybean cultivars. The data provided establish a strong phenotypic and spatial association between *GsMAPK23-4* and proline-mediated protein accumulation, but whether this reflects a causal mechanism remains to be determined.

Because relatively little is known about marker genes for soybean cell types, the spatial transcriptomics data were deconvoluted for the scRNA-seq clusters. This may have led to bias in cell-type identification, as excessively high resolution can introduce artificially divided subpopulations, whereas insufficient resolution may obscure true cellular heterogeneity. We therefore used the RCTD method. Numerous previous studies have also

explored single-cell sequencing identification using spatial transcriptomics results, and this has become a pivotal approach for single-cell annotation (Xu et al., 2023; Yao et al., 2024). Here, although the resolution was fuzzy for the definition of cell types in some boundary areas, we were able to effectively integrate single-cell data and spatial expression patterns overall. However, theoretically, the number of single-cell cell types should be greater than the number of spatial transcriptomics cell types, but it is unrealistic to expect complete annotation of all cell populations of single cells using spatial localization. Owing to the lack of specific marker genes for these seed cell types, we will undertake in-depth analysis using single-cell sequencing data in our next study.

In summary, we created a cellular multi-omics map by combining spatial transcriptomics data with spatial metabolomics data and identified highly enriched metabolites that were differentially abundant during the mid-maturity stage of seed development in wild soybeans. These findings provide a foundation for future analyses of soybean seed development with single-cell transcriptomic and metabolomic resolution. By mapping cell-type-specific gene expression and metabolic networks, we revealed the spatial organization of nutrient accumulation, identified key ligand–receptor communication hubs in the seed coat, and delineated the developmental trajectories of cotyledon cells. Importantly, we functionally validated *GsMAPK23-4* as a novel negative regulator of protein and oil content. This work provides a comprehensive multi-omics resource and critical genetic insights for future efforts aimed at improving seed composition in soybean.

## METHODS

### Plant growth conditions

Wild-type *G. soja* ZYD00006 and transgenic and wild-type *G. max* Dongnong no. 50 (DN50) were cultivated at 23°C ± 2°C under a 16-h light/8-h dark photoperiod with 1000 μmol·m$^{-2}$ s$^{-1}$ irradiance and 60%–80% relative humidity. *G. soja* seed coats were scored with a knife (without damaging the hilum) before germination. Wild soybean seeds were collected during the mid-maturity stage (Soybean Ontology: SOY: 0001293).

### Protoplast isolation

After collection, seeds were soaked in 20 mM MES (pH 5.7) with 3% (w/v) Cellulase R10, 1% (w/v) Macerozyme R10, 1% (w/v) Snailase, 1% (w/v) cutinase, 0.4 M mannitol, and 20 mM KCl to isolate protoplasts. The samples (two wild soybean seeds at the mid-maturity stage) were then soaked in the enzymolytic solution and vacuum infiltrated for 10 min to promote entry of the solution into the tissue samples. After dissociation, the cells were passed through a cell sieve to remove large cellular aggregates and obtain a single-cell suspension.

### scRNA-seq

The freshly prepared single-cell suspension was adjusted to 700–1200 cells/μl according to the operation manual of the 10× Genomics Chromium Next GEM Single Cell 3′ Reagent Kit v3.1 (no. 1000268) for library construction. The constructed library was sequenced using the Illumina NovaSeq 6000 platform to obtain 150-bp paired-end reads.

### scRNA-seq data analyses

Cell barcodes were demultiplexed and reads were mapped to the *G. soja* W05 genome using the Cell Ranger pipeline (v5.0.0, 10× Genomics). The algorithms in this package generated expression matrices for all cells and all genes. Seurat (version 3.1.1) (Hao et al., 2021) was used for subsequent quality control, retaining high-quality cells with > 200 genes, > 1000 unique molecular identifiers (UMIs), and a $\log_{10}$GenesPerUMI value > 0.7, and doublets were removed from further analyses with DoubletFinder. To obtain normalized gene expression data, library size normalization was performed using the NormalizeData function. The global-scaling normalization method LogNormalize was used to normalize the gene expression measurements for each cell in terms of total expression, with multiplication by a scaling factor (10 000 by default) and log transformation of the results.

The top 2000 most variable genes were determined using the Seurat function FindVariableGenes (mean.function = FastExpMean, dispersion.function = FastLog VMR). PCA was performed to reduce dimensionality with the RunPCA function. Graph-based clustering was performed with the FindClusters function to cluster cells according to their gene expression profiles. We visualized the clusters on a 2D map produced with t-SNE (Van der Maaten and Hinton, 2008). The FindAllMarkers function (test.use = presto) was used to identify the marker genes for each cluster.

Protoplast preparation has been shown to trigger protoplasting-responsive genes that are not expressed during normal plant growth (Xu et al., 2021). To avoid interference of these genes in subsequent analyses, protoplasting-responsive genes from the model species *Arabidopsis* (Denyer et al., 2019) were used to identify wild soybean homologs from the soybean genome using Blastn, retaining genes with an E value $\leq 1 \times 10^{-5}$ and percentage identity > 70%. Thus, we obtained a list of relevant genes in wild soybean (Supplemental Table 3) that were excluded from further analyses.

### Spatial transcriptomics sequencing

Seeds collected during the mid-maturity stage were embedded directly in OCT (Sakura, 4583) before cryosectioning. Tissue sections were mounted on spatially barcoded microarray slides using a Visium Spatial Gene Expression Slide & Reagents Kit (10× Genomics, 1000184). RNA was extracted, and qualified samples with RIN > 7.0 were selected for hematoxylin and eosin staining. Tissues were cleared for 12 min using an optimized protocol to enable release of the maximum amount of RNA, which was then bound to spatially barcoded spots. The captured mRNA was used for cDNA synthesis and library preparation for the 10× Visium platform according to the Visium Spatial Gene Expression User Guide (CG000239 rev. A, 10× Genomics).

### Analysis of spatial transcriptomic data

The FASTQ files were processed and aligned to the *G. soja* W05 reference genome using Space Ranger software (version 2.0.1) from 10× Genomics, with UMI counts summarized for each barcode. To distinguish tissue-overlaying spots from the background, tissue-overlaying spots were identified from the images. The filtered UMI count matrix was then analyzed using the Seurat (version 4.3.0) (Stuart et al., 2019) R package.

Sctransform (Hafemeister and Satija, 2019) was used to normalize data and identify the top 3000 most variable genes. PCA was performed to reduce dimensionality on the log-transformed gene–barcode matrices of top variable genes. Graph-based clustering was performed to cluster cells according to their gene expression profiles with the FindClusters function. Cells were visualized using a 2D t-SNE (nonlinear dimensionality reduction) algorithm. The FindAllMarkers function (test.use = presto) was used to identify the marker genes for each cluster. FindAllMarkers was implemented in Seurat (v4.1.3) with the settings "min.pct = 0.25" and "logfc.threshold = 0.58" to identify genes enriched in the clusters.

The AddModuleScore function in the Seurat R package was used to define the scores of the marker-genes gene set. All analyzed features were binned on the basis of averaged expression, and the control features were randomly selected from each bin.

DEGs were identified using the function FindMarkers (test.use = presto) with $P < 0.05$ and |avg_log2FoldChange| > 1.0 as the thresholds for significant differential expression. GO enrichment and KEGG pathway enrichment analyses of DEGs were performed using R (v4.0.3) on the basis of the hypergeometric distribution. The criteria for screening GO and KEGG terms were as follows: the number of genes enriched in terms was greater than 2, and the enrichment significance was less than 0.05. GSVA was performed using the GSVA package (v1.34.0, Barcelona, Spain) (Hänzelmann et al., 2013) and the KEGG database (Kyoto, Japan) (Kanehisa et al., 2017).

### Deconvolution

The RCTD (v1.1.0) (Cable et al., 2022) method for cell-type deconvolution was used with default parameters, with a minimum of 1 cell per cell type and 1 UMI per spot, setting the run.RCTD function doublet_mode to FALSE, thereby enabling the inference of cell-type composition for each locus.

### Cell–cell communication analysis

The PlantPhoneDB package (version 1.0.0) (Xu et al., 2022) was used to predict ligand–receptor interactions. First, the normalized expression matrix was provided as input, then the corresponding plant ligand–receptor database was selected, and finally the LRscore function was used to calculate the scores of the ligand–receptor interactions.

### CytoTRACE analysis

CytoTRACE (Gulati et al., 2020) (version 0.3.3) was used to predict differentiation states from the spatial transcriptomics data. The raw counts were loaded, and the CytoTRACE function was then used to determine the recommended time order of cells with the enableFast = TRUE parameter. Visualization was performed using the plotCytoTRACE function.

### RNA velocity analyses

Recounting of spliced and unspliced mRNA with the Python script velocyto.py (La Manno et al., 2018) (https://github.com/velocyto-team/velocyto.py) was performed on the Cell Ranger output folder to facilitate RNA velocity analyses. Values for each gene in each cell were established and translated to RNA velocity vectors in low-dimensional space using the velocyto.R package (v0.6), projecting the velocity vectors onto t-SNE plots generated with Seurat.

### Pseudotime analyses

The developmental pseudotime was determined using the Monocle2 package (version 2.9.0) (Trapnell et al., 2014). Raw counts were first converted from a Seurat object into a CellDataSet object with the importCDS function in Monocle. The differentialGeneTest function of the Monocle2 package was used to select ordering genes ($q < 0.01$) that were likely to be informative in the ordering of cells along the pseudotime trajectory. Dimensional reduction clustering analysis was performed with the reduceDimension function, followed by trajectory inference with default parameters of the orderCells function. Gene expression was plotted with the plot_genes_in_pseudotime function to track changes over pseudotime.

### RNA *in situ* hybridization

Mid-maturity-stage seeds were fixed for 24 h in Formalin-Acetic Acid-Alcohol (FAA) solution at 4°C, then dehydrated in an ethanol gradient and embedded in paraffin. The tissues were then sliced with a microtome to obtain 3- to 5-μm-thick sections. Digoxigenin (DIG)-labeled specific primers were designed for marker genes (Supplemental Table 17). After pretreatment, probes were hybridized to the sections at 60°C for 48 h, and anti-DIG–horseradish peroxidase (HRP) was then applied to detect the associated signal.

### Preparation of spatial metabolomics sections

OCT was applied to the bottom of an embedding block, samples were placed in the center of the block, more OCT was added, and any air bubbles were removed from the tissue sections with tweezers. Samples were then sectioned at −20°C with a cryomicrotome. After preliminary sample trimming, sections were cut at a 30-μm thickness, exchanging the cutting blade for one of 10-μm thickness when approaching the target area. Sections were placed on a Superfrost Plus positive-charge anti-slip slide and flattened with a paintbrush, and hematoxylin and eosin staining was performed for one section, followed by storage at −80°C for subsequent imaging.

### AFADESI–MSI analyses

MSI analyses of *G. soja* were performed with an AFADESI–MSI platform coupled with a Q-Orbitrap mass spectrometer (Q Exactive, Thermo Scientific, Germany) using a spray voltage of 7.0 kV and an extraction gas flow rate of 45 $l \cdot min^{-1}$. The spray solvent consisted of an 8:2 mixture of acetonitrile and water at a flow rate of 5 $\mu l \cdot min^{-1}$, with 0.7 MPa nitrogen as the spray gas. Tissue sections were mounted on glass slides, fixed with a 3D electrical moving stage (Beijing Optical Instrument Factory, China), and scanned using a sprayer at a constant 0.2 $mm \cdot s^{-1}$ rate in the *x* direction, with a 0.2-mm vertical step separation in the *y* direction.

AFADESI–MSI analyses were performed through a full MS scan in negative-ion mode with the following MS parameters: capillary temperature 350°C, S-lens voltage 55 V, maximum injection time 200 ms, automatic gain control target $3 \times 10^6$, resolution 70 000, and *m/z* range 70–1000 Da. Xcalibur 3.0 Qual Browser (Thermo Fisher Scientific) was used to convert raw MS data files to .cdf format. Images were analyzed with the custom MassImager 1.0 program.

### Integration of spatial transcriptomics and metabolomics data

Spatially resolved transcriptomic and MSI data were aligned through feature-based alignment and dynamic transformation (Ravi et al., 2022). Manual alignment was achieved with the ManualAlignImages function in Semla Shiny (Larsson et al., 2023), using this same approach for processing RNA data and MSI data points with hematoxylin and eosin-stained images. Sections that did not align, particularly at the edges or in areas that contained artifacts, were excluded from further analyses. MSI data were subjected to PCA to exclude any data outside of the tissue sections. The nearest pairs in the spatial transcriptomics and MSI datasets were identified using the R package dbscan with k-NN (k = 6), filtering out any pairs with a distance greater than 45 pixels. For individual MSI points, only the closest RNA neighbor was retained to ensure one-to-one mapping. The aligned data were stored using newly generated Seurat objects (Hao et al., 2024).

### Correlations between spatial transcriptomics and metabolomics data

Pearson correlation coefficients were calculated for metabolites and associated mRNA levels among samples in the same group to quantify spatial transcriptional–metabolic regulatory relationships, with values closer to 1 denoting greater concordance between mRNA and metabolite levels. The results were assembled in a heatmap to evaluate correlations between DEGs and differentially abundant metabolites in a given spatial area. Given the point-to-point correspondence between spatial transcriptomics and spatial metabolomics datasets and their weighted matrices, integrated data-based correlation analyses should confer greater accuracy.

### Correlation heatmap

Genes with high spatial variability were identified using the FindVariableFeatures function of the Seurat R package (https://satijalab.org/seurat/reference/findvariablefeatures). The top 2000 variable genes were selected on the basis of the screen performed above, and we selected the 36 metabolites that were significantly enriched in pathways.

Associations between mRNA levels and metabolites among samples were analyzed using Pearson correlation coefficients, with values closer to 1 indicating greater concordance between mRNAs and metabolites in these samples.

### Transmission electron microscopy (TEM) analysis of cotyledon cells in *G. soja*

Seeds were sampled at the mid-maturity stage to examine the ultrastructure of the cotyledon cells. The cotyledons were cut into 1-$mm^3$ blocks and then prepared for TEM sectioning as described previously (Mori et al., 2004). The ultra-thin sections were evaluated and imaged using a transmission electron microscope (HITACHI HT7800, Hitachi Co., Ltd., Japan) at 120 kV.

### Subcellular-localization analysis

A GsMAPK23-4–GFP fusion protein was constructed by inserting the *GsMAPK23-4* coding region into the GFP-tagged pBI121 vector using gene-specific primers (Supplemental Table 18). After Polyethylene Glycol (PEG)–calcium-mediated transformation for 15 h, GsMAPK23-4–GFP plasmids were transiently expressed in *Arabidopsis* mesophyll protoplasts (Yoo et al., 2007). An empty GFP vector was used as a positive control (Chen et al., 2022a, 2022b).

Subcellular localization was determined by confocal laser-scanning microscopy (Carl Zeiss LSM 710, Germany). GFP imaging was performed at an excitation wavelength of 488 nm and an emission wavelength of 507 nm.

### Construction of expression vectors and plant transformation

The pSOY1-*GsMAPK23-4* overexpression vector was generated by cloning the *GsMAPK23-4* coding sequence from *G. max* DN50 cDNA (which is identical to that of *GsMAPK23-4*) into the pSOY1 vector containing the CaMV *35S* promoter. Colony PCR was used to verify successful construction of the overexpression vector (Supplemental Figure 10A).

For CRISPR-Cas9 gene editing, the CRISPR-P 2.0 tool was used to design two sgRNAs targeting *GsMAPK23-4* (Liu et al., 2017), which were then subcloned into the pGES401 plasmid; the resulting construct was introduced into *Agrobacterium tumefaciens* EHA105. *G. max* DN50 served as the cloning target and background for genetic transformation.

### Yeast two-hybrid (Y2H) assay

The coding sequence of *GsCDC48-12* was inserted into the pGBKT7 vector using the EcoRI and BamHI restriction sites to create pGBKT7-*GsCDC48-12*. Concurrently, the coding sequence of *GsCDC48-1* was inserted into the pGADT7 vector. Different combinations of recombinant and empty vectors were co-transformed into the Y2H Gold yeast strain. Transformed yeast cells were plated on X-α-Gal-supplemented SD/−Leu/−Trp for selection and interaction verification. Interaction assays were refined on quadruple-dropout medium lacking leucine, tryptophan, histidine, and adenine (SD/−Leu/−Trp/−His/−Ade) supplemented with X-α-Gal.

### GST pull-down assay

The coding sequences of *GsCDC48-1* and *GsCDC48-12* were inserted into the pGEX-6P-1 and pET-28a plasmids fused with GST and His tags, respectively. The recombinant plasmids were introduced into *Escherichia coli* Rosetta (DE3) strains. Protein expression was induced by addition of 0.5 mM Isopropyl β-D-1-Thiogalactopyranoside (IPTG), followed by incubation at 16°C for 16 h. Cells were lysed in lysis buffer (1× PBS, 1 mM PMSF) and sonicated, and supernatants were collected by centrifugation at 12 000 rpm for 10 min. GST-tagged GsCDC48-1 was isolated using GST magnetic beads (Solarbio) and eluted with GST elution buffer (150 mM NaCl, 50 mM Tris–HCl [pH 8.0], and 25 mM reduced glutathione). After three washes, the GST-tagged protein-bound beads were mixed with supernatants containing His-tagged GsCDC48-12. After 4 h of incubation, the beads were washed with GST washing buffer (150 mM NaCl, 50 mM Tris–HCl [pH 7.5], and 2 mM EDTA) to remove nonspecifically bound proteins. Bound proteins were eluted with GST elution buffer. Interactions were analyzed by immunoblotting using anti-His (1:5000, M20001, Abmart) and anti-GST (1:5000, 30901ES50, Yeasen) antibodies.

### Co-IP assay

The coding sequences of *GsCDC48-12* and *GsCDC48-1* were cloned into the pCAMBIA1302-4×Myc-CMV-3×FLAG vector (Myc-tagged) and the pCAMBIA1302 plant expression vector (GFP-tagged), respectively. The recombinant constructs were introduced into *A. tumefaciens* strain GV3101 and subsequently infiltrated into 1-month-old tobacco leaves. After a 48-h incubation, proteins were isolated from the infiltrated leaves using IP Lysis Buffer (Thermo) supplemented with 1× Protease Inhibitor Cocktail (Roche, 04693132001). The extracts were then incubated with protein A/G agarose beads (Yeasen) in conjunction with an anti-GFP antibody (Roche) and incubated at 4°C for approximately 16 h to pull down the protein complexes. The immunoprecipitated proteins were visualized by immunoblotting with anti-GFP (Roche) and anti-Myc (Abclonal) antibodies (1:5000 each). HRP goat anti-mouse IgG (Abclonal) and HRP goat anti-rabbit IgG (Proteintech) secondary antibodies were used.

### RT–qPCR

Total RNA was extracted with the TRIzol reagent, DNA was removed by RNase-free DNase treatment, and first-strand cDNA was synthesized from the RNA templates. *ACTIN11* served as a reference gene for normalization, and the $2^{-\Delta\Delta Ct}$ method was used to calculate relative gene expression. All analyses were performed in triplicate.

### Phenotypic analysis of seeds

Protein, oil, and amino acid contents of seeds were measured using the PerkinElmer DA7250 NIR (near-infrared) grain analyzer (Perten Instruments, Sweden).

### Statistical analysis

Statistical analysis was performed using one-way analysis of variance (ANOVA) followed by Dunnett's test in GraphPad Prism version 8.0 (GraphPad, USA). $P < 0.05$ was considered statistically significant. Values for individual samples represent the mean of three replicates.

### Data availability

The scRNA-seq and spatial transcriptomics data have been deposited at the NCBI Gene Expression Omnibus under accession numbers GEO: GSE281215 and GSE281070. The spatial metabolomics data can be accessed at https://metaspace2020.eu/project/liu-2024-1. The complete cellular multi-omics map generated in this study is available at a dedicated online repository (URL: http://www.soycellmarker.com/).

## FUNDING

This study was financially supported by the Natural Science Foundation of Heilongjiang Province of China (ZL2024C007) and the National Natural Science Foundation of China (U23A20192, 31971899, 32272093, and 32272072).

## ACKNOWLEDGMENTS

No conflict of interest is declared.

## AUTHOR CONTRIBUTIONS

Y.Z. and Q.C. conceived and designed the project. P.L. and M.L. performed the experiments. C.L., Z.H., and M.Y. contributed to project discussions. P.L., P.M., and H.Y. analyzed the data. P.L. wrote the

manuscript draft. Y.Z. and M.L. revised the manuscript. All authors have read and approved the final manuscript.

## SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

Received: January 4, 2025
Revised: February 13, 2025
Accepted: October 27, 2025
Published: October 29, 2025

## REFERENCES

**Almeida-Silva, F., Moharana, K.C., and Venancio, T.M.** (2021). The state of the art in soybean transcriptomics resources and gene coexpression networks. silico Plants **3**:diab005. https://doi.org/10.1093/insilicoplants/diab005.

**Bach, T.M.H., and Takagi, H.** (2013). Properties, metabolisms, and applications of (L)-proline analogues. Appl. Microbiol. Biotechnol. **97**:6623–6634. https://doi.org/10.1007/s00253-013-5022-7.

**Bae, H., Choi, S.M., Yang, S.W., Pai, H.S., and Kim, W.T.** (2009). Suppression of the ER-localized AAA ATPase NgCDC48 inhibits tobacco growth and development. Mol. Cells **28**:57–65. https://doi.org/10.1007/s10059-009-0101-4.

**Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W.** (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. **37**:38–44. https://doi.org/10.1038/nbt.4314.

**Bigeard, J., and Hirt, H.** (2018). Nuclear Signaling of Plant MAPKs. Front. Plant Sci. **9**:469. https://doi.org/10.3389/fpls.2018.00469.

**Borisjuk, L., Nguyen, T.H., Neuberger, T., Rutten, T., Tschiersch, H., Claus, B., Feussner, I., Webb, A.G., Jakob, P., Weber, H., et al.** (2005). Gradients of lipid storage, photosynthesis and plastid differentiation in developing soybean seeds. New Phytol. **167**:761–776. https://doi.org/10.1111/j.1469-8137.2005.01474.x.

**Chen, J., Zhang, P., Zhao, Y., Zhao, J., Wu, X., Zhang, R., Cha, R., Yao, Q., and Gao, Y.** (2022). Nitroreductase-instructed supramolecular assemblies for microbiome regulation to enhance colorectal cancer treatments. Sci. Adv. **8**:eadd2789. https://doi.org/10.1126/sciadv.add2789.

**Cable, D.M., Murray, E., Zou, L.S., et al.** (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. Nat. Biotechnol. **40**:517–526. https://doi.org/10.1038/s41587-021-00830-w.

**Chen, X., Wang, Y., Wang, J.N., Cao, Q.C., Sun, R.X., Zhu, H.J., Zhang, Y.R., Ji, J.D., and Liu, Q.H.** (2022). m(6)A modification of circSPECC1 suppresses RPE oxidative damage and maintains retinal homeostasis. Cell Rep. **41**:111671. https://doi.org/10.1016/j.celrep.2022.111671.

**Cristina, M., Petersen, M., and Mundy, J.** (2010). Mitogen-Activated Protein Kinase Signaling in Plants. Annu. Rev. Plant Biol. **61**:621–649. https://doi.org/10.3390/ijms23052744.

**Danzer, J., Mellott, E., Bui, A.Q., Le, B.H., Martin, P., Hashimoto, M., Perez-Lesher, J., Chen, M., Pelletier, J.M., Somers, D.A., et al.** (2015). Down-Regulating the Expression of 53 Soybean Transcription Factor Genes Uncovers a Role for SPEECHLESS in Initiating Stomatal Cell Lineages during Embryo Development. Plant Physiol. **168**:1025–1035. https://doi.org/10.1104/pp.15.00432.

**Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., and Timmermans, M.C.P.** (2019). Spatiotemporal Developmental Trajectories in the Arabidopsis Root Revealed Using High-Throughput Single-Cell RNA Sequencing. Dev. Cell **48**:840–852.e5. https://doi.org/10.1016/j.devcel.2019.02.022.

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

**Duan, Z., Zhang, M., Zhang, Z., Liang, S., Fan, L., Yang, X., Yuan, Y., Pan, Y., Zhou, G., Liu, S., et al.** (2022). Natural allelic variation of GmST05 controlling seed size and quality in soybean. Plant Biotechnol. J. **20**:1807–1818. https://doi.org/10.1111/pbi.13865.

**Egamberdieva, D., Ma, H., Reckling, M., Omari, R.A., Wirth, S., and Bellingrath-Kimura, S.D.** (2021). Interactive effects of biochar, nitrogen, and phosphorous on the symbiotic performance, growth, and nutrient uptake of soybean (Glycine max L.). Agron **12**:27. https://doi.org/10.3390/agronomy12010027.

**Fatahi, B., Sorkheh, K., and Sofo, A.** (2024). Deciphering the possible role of RNA-helicase genes mechanism in response to abiotic stresses in rapeseed (Brassica napus L.). BMC Plant Biol. **24**:206. https://doi.org/10.1186/s12870-024-04893-0.

**Ge, X., Yu, X., Liu, Z., Yuan, J., Qin, A., Wang, Y., Chen, Y., Qin, W., Liu, Y., Liu, X., et al.** (2025). Spatiotemporal transcriptome and metabolome landscapes of cotton somatic embryos. Nat. Commun. **16**:859. https://doi.org/10.1038/s41467-025-55870-6.

**Gulati, G.S., Sikandar, S.S., Wesche, D.J., Manjunath, A., Bharadwaj, A., Berger, M.J., Ilagan, F., Kuo, A.H., Hsieh, R.W., Cai, S., et al.** (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. Science **367**:405–411. https://doi.org/10.1126/science.aax0249.

**Gupta, M., Bhaskar, P.B., Sriram, S., and Wang, P.H.** (2017). Integration of omics approaches to understand oil/protein content during seed development in oilseed crops. Plant Cell Rep. **36**:637–652. https://doi.org/10.1007/s00299-016-2064-1.

**Hafemeister, C., and Satija, R.** (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. **20**:296. https://doi.org/10.1186/s13059-019-1874-1.

**Hänzelmann, S., Castelo, R., and Guinney, J.** (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinf. **14**:7. https://doi.org/10.1186/1471-2105-14-7.

**Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., and Satija, R.** (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat. Biotechnol. **42**:293–304. https://doi.org/10.1038/s41587-023-01767-y.

**Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al.** (2021). Integrated analysis of multimodal single-cell data. Cell **184**:3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

**Harada, J.J., Barker, S.J., and Goldberg, R.B.** (1989). Soybean beta-conglycinin genes are clustered in several DNA regions and are regulated by transcriptional and posttranscriptional processes. Plant Cell **1**:415–425. https://doi.org/10.1105/tpc.1.4.415.

**Hettenhausen, C., Baldwin, I.T., and Wu, J.** (2012). Silencing MPK4 in Nicotiana attenuata Enhances Photosynthesis and Seed Production But Compromises Abscisic Acid-Induced Stomatal Closure and Guard Cell-Mediated Resistance to Pseudomonas syringae pv tomato DC3000. Plant Physiol. **158**:759–776. https://doi.org/10.1104/pp.111.190074.

**Hymowitz, T.** (1970). On the domestication of the soybean. Econ. Bot. **24**:408–421. https://doi.org/10.1007/BF02860745.

**Inès, D., Courty, P.E., Wendehenne, D., and Rosnoblet, C.** (2024). CDC48 in plants and its emerging function in plant immunity. Trends Plant Sci. **29**:786–798. https://doi.org/10.1016/j.tplants.2023.12.013.

**Jo, L., Pelletier, J.M., and Harada, J.J.** (2019). Central role of the LEAFY COTYLEDON1 transcription factor in seed development. J. Integr. Plant Biol. **61**:564–580. https://doi.org/10.1111/jipb.12806.

**Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.** (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. **45**:D353–D361. https://doi.org/10.1093/nar/gkw1092.

**Kim, D.Y., Heo, J.H., Pack, I.S., Park, J.-H., Um, M.S., Kim, H.J., Park, K.W., Nam, K.-H., Oh, S.D., Kim, J.K., et al.** (2021). Natural hybridization between transgenic and wild soybean genotypes. Plant Biotechnol. Rep. **15**:299–308. https://doi.org/10.1007/s11816-021-00685-2.

**Kim, W.J., Kang, B.H., Moon, C.Y., Kang, S., Shin, S., Chowdhury, S., Jeong, S.-C., Choi, M.-S., Park, S.-K., Moon, J.-K., et al.** (2023). Genome-wide association study for agronomic traits in wild soybean (Glycine soja). Agron **13**:739. https://doi.org/10.3390/agronomy13030739.

**Kudełka, W., Kowalska, M., and Popis, M.** (2021). Quality of soybean products in terms of essential amino acids composition. Molecules **26**:5071. https://doi.org/10.3390/molecules26165071.

**La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al.** (2018). RNA velocity of single cells. Nature **560**:494–498. https://doi.org/10.1038/s41586-018-0414-6.

**Larsson, L., Franzén, L., Ståhl, P.L., and Lundeberg, J.** (2023). Semla: a versatile toolkit for spatially resolved transcriptomics analysis and visualization. Bioinformatics **39**:btad626. https://doi.org/10.1093/bioinformatics/btad626.

**Le, B.H., Wagmaister, J.A., Kawashima, T., Bui, A.Q., Harada, J.J., and Goldberg, R.B.** (2007). Using Genomics to Study Legume Seed Development. Plant Physiol. **144**:562–574. https://doi.org/10.1104/pp.107.100362.

**Lee, B.-R., Muneer, S., Park, S.-H., Zhang, Q., and Kim, T.-H.** (2013). Ammonium-induced proline and sucrose accumulation, and their significance in antioxidative activity and osmotic adjustment. Acta Physiol. Plant. **35**:2655–2664. https://doi.org/10.1007/s11738-013-1297-7.

**Li, Y., Guan, R., Liu, Z., Ma, Y., Wang, L., Li, L., Lin, F., Luan, W., Chen, P., Yan, Z., et al.** (2008). Genetic structure and diversity of cultivated soybean (Glycine max (L.) Merr.) landraces in China. Theor. Appl. Genet. **117**:857–871. https://doi.org/10.1007/s00122-008-0825-0.

**Lin, J.-Y., Le, B.H., Chen, M., Henry, K.F., Hur, J., Hsieh, T.-F., Chen, P.-Y., Pelletier, J.M., Pellegrini, M., Fischer, R.L., et al.** (2017). Similarity between soybean and Arabidopsis seed methylomes and loss of non-CG methylation does not affect seed development. Proc. Natl. Acad. Sci. USA **114**:E9730–E9739. https://doi.org/10.1073/pnas.1716758114.

**Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K., and Chen, L.L.** (2017). CRISPR-P 2.0: An Improved CRISPR-Cas9 Tool for Genome Editing in Plants. Mol. Plant **10**:530–532. https://doi.org/10.1016/j.molp.2017.01.003.

**Machado, F.B., Moharana, K.C., Almeida-Silva, F., Gazara, R.K., Pedrosa-Silva, F., Coelho, F.S., Grativol, C., and Venancio, T.M.** (2020). Systematic analysis of 1298 RNA-Seq samples and construction of a comprehensive soybean (Glycine max) expression atlas. Plant J. **103**:1894–1909. https://doi.org/10.1111/tpj.14850.

**Manna, M., Rengasamy, B., and Sinha, A.K.** (2023). Revisiting the role of MAPK signalling pathway in plants and its manipulation for crop improvement. Plant Cell Environ. **46**:2277–2295. https://doi.org/10.1111/pce.14606.

**Meena, M., Divyanshu, K., Kumar, S., Swapnil, P., Zehra, A., Shukla, V., Yadav, M., and Upadhyay, R.S.** (2019). Regulation of L-proline biosynthesis, signal transduction, transport, accumulation and its vital role in plants during variable environmental conditions. Heliyon **5**:e02952. https://doi.org/10.1016/j.heliyon.2019.e02952.

**Mori, T., Maruyama, N., Nishizawa, K., Higasa, T., Yagasaki, K., Ishimoto, M., and Utsumi, S.** (2004). The composition of newly synthesized proteins in the endoplasmic reticulum determines the transport pathways of soybean seed storage proteins. Plant J. **40**:238–249. https://doi.org/10.1111/j.1365-313X.2004.02204.x.

**Nakabayashi, R., Hashimoto, K., Mori, T., Toyooka, K., Sudo, H., and Saito, K.** (2021). Spatial metabolomics using imaging mass spectrometry to identify the localization of asparaptine A in Asparagus officinalis. Plant Biotechnol. **38**:311–315. https://doi.org/10.5511/plantbiotechnology.21.0504b.

**Nicolas-Francès, V., Besson-Bard, A., Meschini, S., Klinguer, A., Bonnotte, A., Héloir, M.C., Citerne, S., Inès, D., Hichami, S., Wendehenne, D., and Rosnoblet, C.** (2024). CDC48 regulates immunity pathway in tobacco plants. Plant Physiol. Biochem. **211**:108714. https://doi.org/10.1016/j.plaphy.2024.108714.

**Pelletier, J.M., Kwong, R.W., Park, S., Le, B.H., Baden, R., Cagliari, A., Hashimoto, M., Munoz, M.D., Fischer, R.L., Goldberg, R.B., and Harada, J.J.** (2017). LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. Proc. Natl. Acad. Sci. USA **114**:E6710–E6719. https://doi.org/10.1073/pnas.1707957114.

**Perez-Grau, L., and Goldberg, R.B.** (1989). Soybean Seed Protein Genes Are Regulated Spatially during Embryogenesis. Plant Cell **1**:1095–1109. https://doi.org/10.1105/tpc.1.11.1095.

**Qin, Y., Pan, J., Cai, M., Yao, L., and Ji, Z.** (2016). Pattern Genes Suggest Functional Connectivity of Organs. Sci. Rep. **6**:26501. https://doi.org/10.1038/srep26501.

**Rao, A., Barkley, D., França, G.S., and Yanai, I.** (2021). Exploring tissue architecture using spatial transcriptomics. Nature **596**:211–220. https://doi.org/10.1038/s41586-021-03634-9.

**Ravi, V.M., Will, P., Kueckelhaus, J., Sun, N., Joseph, K., Salié, H., Vollmer, L., Kuliesiute, U., von Ehr, J., Benotmane, J.K., et al.** (2022). Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. Cancer Cell **40**:639–655.e13. https://doi.org/10.1016/j.ccell.2022.05.009.

**Sang, Q., and Kong, F.** (2024). Applications for single-cell and spatial transcriptomics in plant research. New Crops **1**:100025. https://doi.org/10.1016/j.ncrops.2024.100025.

**Santos-Mendoza, M., Dubreucq, B., Baud, S., Parcy, F., Caboche, M., and Lepiniec, L.** (2008). Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. Plant J. **54**:608–620. https://doi.org/10.1111/j.1365-313X.2008.03461.x.

**Schnabel, E., Thomas, J., El-Hawaz, R., Gao, Y., Poehlman, W.L., Chavan, S., Pasha, A., Esteban, E., Provart, N., Feltus, F.A., et al.** (2023). Laser Capture Microdissection Transcriptome Reveals Spatiotemporal Tissue Gene Expression Patterns of Medicago truncatula Roots Responding to Rhizobia. Mol. Plant Microbe Interact. **36**:805–820. https://doi.org/10.1094/MPMI-03-23-0029-R.

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

**Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R.** (2019). Comprehensive Integration of Single-Cell Data. Cell **177**:1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

**Szabados, L., and Savouré, A.** (2010). Proline: a multifunctional amino acid. Trends Plant Sci. **15**:89–97. https://doi.org/10.1016/j.tplants.2009.11.009.

**Toda, K., Kuroiwa, H., Senthil, K., et al.** (2012). The soybean F3'H protein is localized to the tonoplast in the seed coat hilum. Planta **236**:79–89. https://doi.org/10.1007/s00425-012-1590-5.

**Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L.** (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. **32**:381–386. https://doi.org/10.1038/nbt.2859.

**Tulpan, D., Leger, S., Tchagang, A., and Pan, Y.** (2015). Enrichment of Triticum aestivum gene annotations using ortholog cliques and gene ontologies in other plants. BMC Genom. **16**:299. https://doi.org/10.1186/s12864-015-1496-2.

**Van der Maaten, L., and Hinton, G.** (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. **9**:2579–2605.

**Volkova, E., and Smolyaninova, N.** (2024). Analysis of world trends in soybean production. BIO Web Conf. **141**:01026. https://doi.org/10.1051/bioconf/202414101026.

**Wang, W., Xiong, H., Sun, K., Zhang, B., and Sun, M.X.** (2022). New insights into cell-cell communications during seed development in flowering plants. J. Integr. Plant Biol. **64**:215–229. https://doi.org/10.1111/jipb.13170.

**Weber, H., Borisjuk, L., and Wobus, U.** (2005). Molecular physiology of legume seed development. Annu. Rev. Plant Biol. **56**:253–279. https://doi.org/10.1146/annurev.arplant.56.032604.144201.

**Wiederstein, M., Baumgartner, S., and Lauter, K.** (2023). Soybean (Glycine max) allergens— a review on an outstanding plant food with allergenic potential. ACS Food Sci. Technol. **3**:363–378. https://doi.org/10.1021/acsfoodscitech.2c00380.

**Wong, K., Navarro, J.F., Bergenstråhle, L., Ståhl, P.L., and Lundeberg, J.** (2018). ST Spot Detector: a web-based application for automatic spot and tissue detection for spatial Transcriptomics image datasets. Bioinformatics **34**:1966–1968. https://doi.org/10.1093/bioinformatics/bty030.

**Wu, G., and Wang, W.** (2024). Recent advances in understanding the role of two mitogen-activated protein kinase cascades in plant immunity. J. Exp. Bot. **75**:2256–2265. https://doi.org/10.1093/jxb/erae020.

**Xu, C., Ma, D., Ding, Q., Zhou, Y., and Zheng, H.L.** (2022). PlantPhoneDB: A manually curated pan-plant database of ligand-receptor pairs infers cell-cell communication. Plant Biotechnol. J. **20**:2123–2134. https://doi.org/10.1111/pbi.13893.

**Xu, X., Crow, M., Rice, B.R., Li, F., Harris, B., Liu, L., Demesa-Arevalo, E., Lu, Z., Wang, L., Fox, N., et al.** (2021). Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery. Dev. Cell **56**:557–568.e6. https://doi.org/10.1016/j.devcel.2020.12.015.

**Xu, Y., Zhang, T., Zhou, Q., Hu, M., Qi, Y., Xue, Y., Nie, Y., Wang, L., Bao, Z., and Shi, W.** (2023). A single-cell transcriptome atlas profiles early organogenesis in human embryos. Nat. Cell Biol. **25**:604–615. https://doi.org/10.1038/s41556-023-01108-w.

**Yamasaki, Y., Gao, F., Jordan, M.C., and Ayele, B.T.** (2017). Seed maturation associated transcriptional programs and regulatory networks underlying genotypic difference in seed dormancy and size/weight in wheat (Triticum aestivum L.). BMC Plant Biol. **17**:154. https://doi.org/10.1186/s12870-017-1104-5.

**Yao, J., Chu, Q., Guo, X., Shao, W., Shang, N., Luo, K., Li, X., Chen, H., Cheng, Q., Mo, F., et al.** (2024). Spatiotemporal transcriptomic landscape of rice embryonic cells during seed germination. Dev. Cell **59**:2320–2332.e5. https://doi.org/10.1016/j.devcel.2024.05.016.

**Yoo, S.D., Cho, Y.H., and Sheen, J.** (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nat. Protoc. **2**:1565–1572. https://doi.org/10.1038/nprot.2007.199.

**Yu, X., Liu, Z., and Sun, X.** (2023). Single-cell and spatial multi-omics in the plant sciences: Technical advances, applications, and perspectives. Plant Commun. **4**:100508. https://doi.org/10.1016/j.xplc.2022.100508.

**Zeng, F.S., Menardo, F., Xue, M.F., Zhang, X.J., Gong, S.J., Yang, L.J., Shi, W.Q., and Yu, D.Z.** (2017). Transcriptome Analyses Shed New Insights into Primary Metabolism and Regulation of Blumeria graminis f. sp. tritici during Conidiation. Front. Plant Sci. **8**:1146. https://doi.org/10.3389/fpls.2017.01146.

**Zhang, X., Luo, Z., Marand, A.P., Yan, H., Jang, H., Bang, S., Mendieta, J.P., Minow, M.A.A., and Schmitz, R.J.** (2025). A spatially resolved multi-omic single-cell atlas of soybean development. Cell **188**:550–567.e19. https://doi.org/10.1016/j.cell.2024.10.050.

**Zhang, Y., and Fernie, A.R.** (2018). On the role of the tricarboxylic acid cycle in plant productivity. J. Integr. Plant Biol. **60**:1199–1216. https://doi.org/10.1111/jipb.12690.

**Zhang, Y., Schernthaner, J., Labbé, N., Hefford, M.A., Zhao, J., and Simmonds, D.H.** (2014). Improved protein quality in transgenic soybean expressing a de novo synthetic protein, MB-16. Transgenic Res. **23**:455–467. https://doi.org/10.1007/s11248-013-9777-5.