nature plants

Brief Communication

https://doi.org/10.1038/s41477-023-01387-z

Integrated single-nucleus and spatial transcriptomics captures transitional states in soybean nodule maturation

Received: 19 June 2022

Accepted: 3 March 2023

Published online: 13 April 2023

Check for updates

Zhijian Liu[1,8], Xiangying Kong[2,3,4,8], Yanping Long[1,8], Sirui Liu[2,5], Hong Zhang[1], Jinbu Jia[1], Wenhui Cui[2,6], Zunmian Zhang[2,5], Xianwei Song[7], Lijuan Qiu[2], Jixian Zhai[1] ✉ & Zhe Yan[2,3] ✉

Legumes form symbiosis with rhizobium leading to the development of nitrogen-fixing nodules. By integrating single-nucleus and spatial transcriptomics, we established a cell atlas of soybean nodules and roots. In central infected zones of nodules, we found that uninfected cells specialize into functionally distinct subgroups during nodule development, and revealed a transitional subtype of infected cells with enriched nodulation-related genes. Overall, our results provide a single-cell perspective for understanding rhizobium–legume symbiosis.

On compatible host plants, rhizobium infect and form symbiotic organ-nodules in the root, establishing nitrogen-fixing nodules that can convert atmospheric nitrogen into organic ammonia for host plant development. Within this highly heterogeneous tissue, cells of various types have different functions, and some important physiological and transcriptomic programmes are only active in some specific cell types. Despite remarkable advances in the field[1–6], knowledge on specific contributions of different types of cell in nodules as well their relationships during nodule maturation is still limited, especially in determinate nodules.

To reveal cell-type-specific dynamic gene expression during nodule maturation in soybean, we established three single-nucleus libraries on the basis of the 10x Genomics Chromium platform with two different developmental stages of nodules (at 12 days post-infection (dpi) and 21 dpi) (Supplementary Fig. 1), with the corresponding region of roots where the nodules were formed at 21 dpi as control (Fig. 1a). The obtained reads were almost exclusively mapped to the soybean genome, indicating that the captured messenger RNAs were not derived from rhizobia (Supplementary Fig. 2). We obtained a total of 26,712 high-quality single-nucleus transcriptomes in the three libraries covering 39,337 genes, with median number of genes per nucleus of 1,342 and median unique molecular identifiers (UMIs) per nucleus of 1,636 (Supplementary Data 1 and 2). After integration of the three datasets using scVI[7], we obtained 15 cell clusters (Fig. 1b,c) and a series of upregulated genes for each cluster (Extended Data Fig. 1 and Supplementary Data 3). In addition, we also identified differentially expressed genes in each cluster between samples (Supplementary Fig. 3 and Data 4). With known soybean marker genes, homologues of marker genes in other legumes and *Arabidopsis* as well as a public *Arabidopsis* single cell RNA-seq (scRNA-seq) dataset[8], we successfully identified root epidermis (cluster 5), root vascular bundle (cluster 3), nodule vascular bundle (cluster 9), nodule cortex (cluster 1) and infected cells (ICs) in nodule central infected zones (CIZ) (cluster 12) (Extended Data Figs. 2 and 3, and Supplementary Fig. 4 and Data 5). With the benefit of scRNA-seq data from *Arabidopsis* root[8], we also successfully identified some cell subtypes of vascular bundle, including pericycle (subcluster vb-0, vb-5), companion cells (vb-6), phloem-like cells (vb-2) and xylem-like cells (vb-4) (detailed in Supplementary materials and Fig. 5). However,

[1]Institute of Plant and Food Science, Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, China. [2]The National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI), Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, China. [3]Key Laboratory of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China. [4]University of Chinese Academy of Sciences, Beijing, China. [5]School of Agricultural Science and Engineering, Liaocheng University, Liaocheng, China. [6]Key Laboratory of Saline-alkali Vegetation Ecology Restoration (Northeast Forestry University), Ministry of Education, Harbin, China. [7]State Key Laboratory of Plant Genomics and National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [8]These authors contributed equally: Zhijian Liu, Xiangying Kong, Yanping Long. ✉e-mail: zhaijx@sustech.edu.cn; yanzhe@caas.cn

https://doi.org/10.1038/s41477-023-01387-z

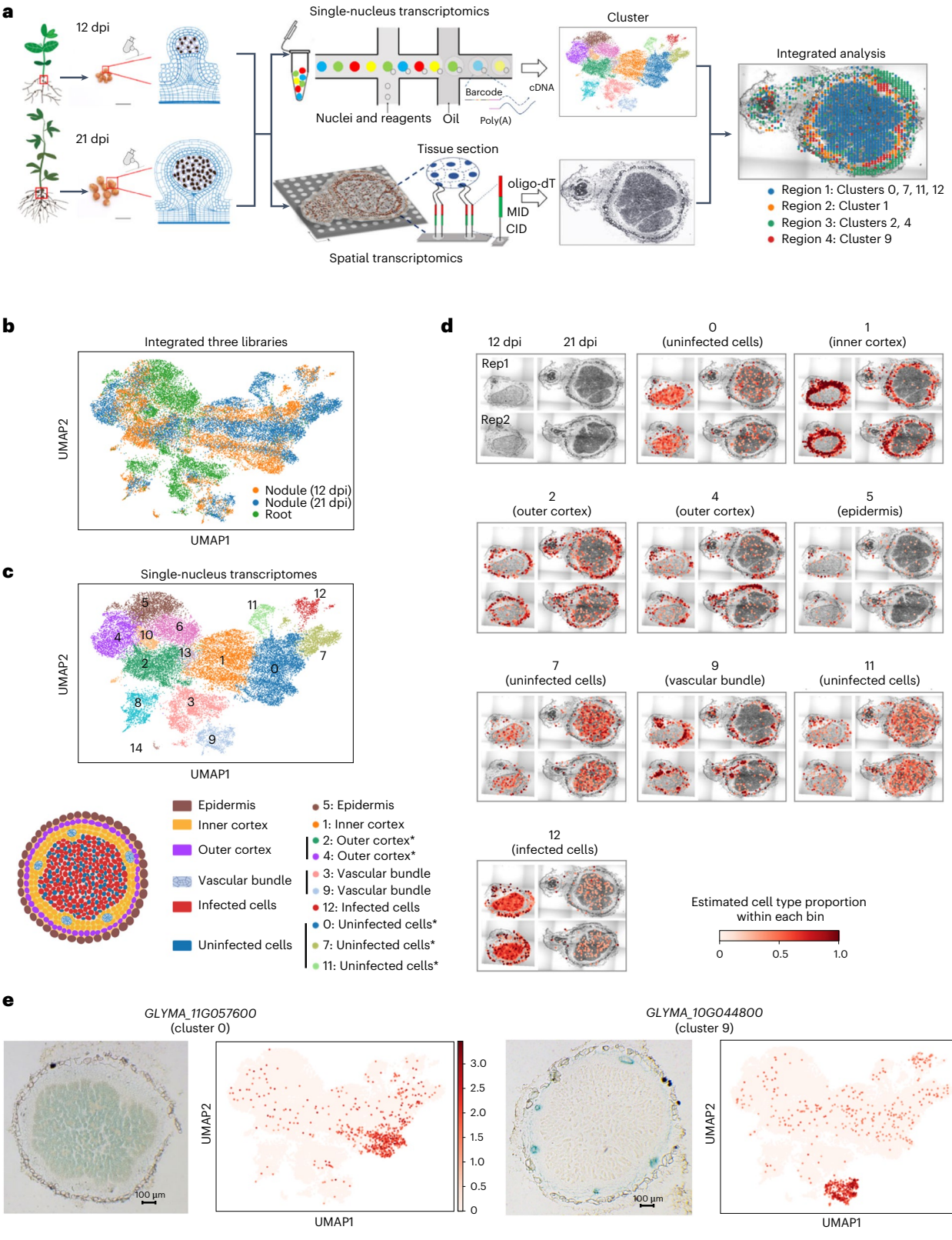


Fig. 1 | Combined single-nucleus and spatial transcriptomes reveal nodule heterogeneity at different developmental stages. a, Schematic diagram of the integration of single-nucleus and spatial transcriptomics analysis. The rightmost panel is integrated analysis plotted on the basis of the results shown in **d**. **b**, Integration of three single-nucleus datasets. **c**, Top: UMAP visualization of 15 identified cell clusters in nodules and roots. '*' indicates that the cluster is annotated by spatial transcriptome. Bottom: cartoon diagram of nodule structure. **d**, Spatial distribution of different cell types in 12- and 21-dpi nodules. Upper left: bright-field image of nodule sections used to prepare the spatial transcriptome. Two replicates were analysed for both 12-dpi and 21-dpi nodules. The other images show the spatial distribution of cell type proportions for each single-nucleus cluster. The colours represent the fraction of single-nucleus transcriptomes of each cluster deconvolved by destVI. **e**, Validation of annotation results by GUS-reporter lines. The left panel of each gene indicates the result of the GUS-reporter line, and the right panel indicates the expression pattern of the gene identified by snRNA-seq. Scale bars, 100 μm. These experiments were repeated in three independent assays and for each section; at least three nodules were analysed, and all showed the same expression pattern.

https://doi.org/10.1038/s41477-023-01387-z

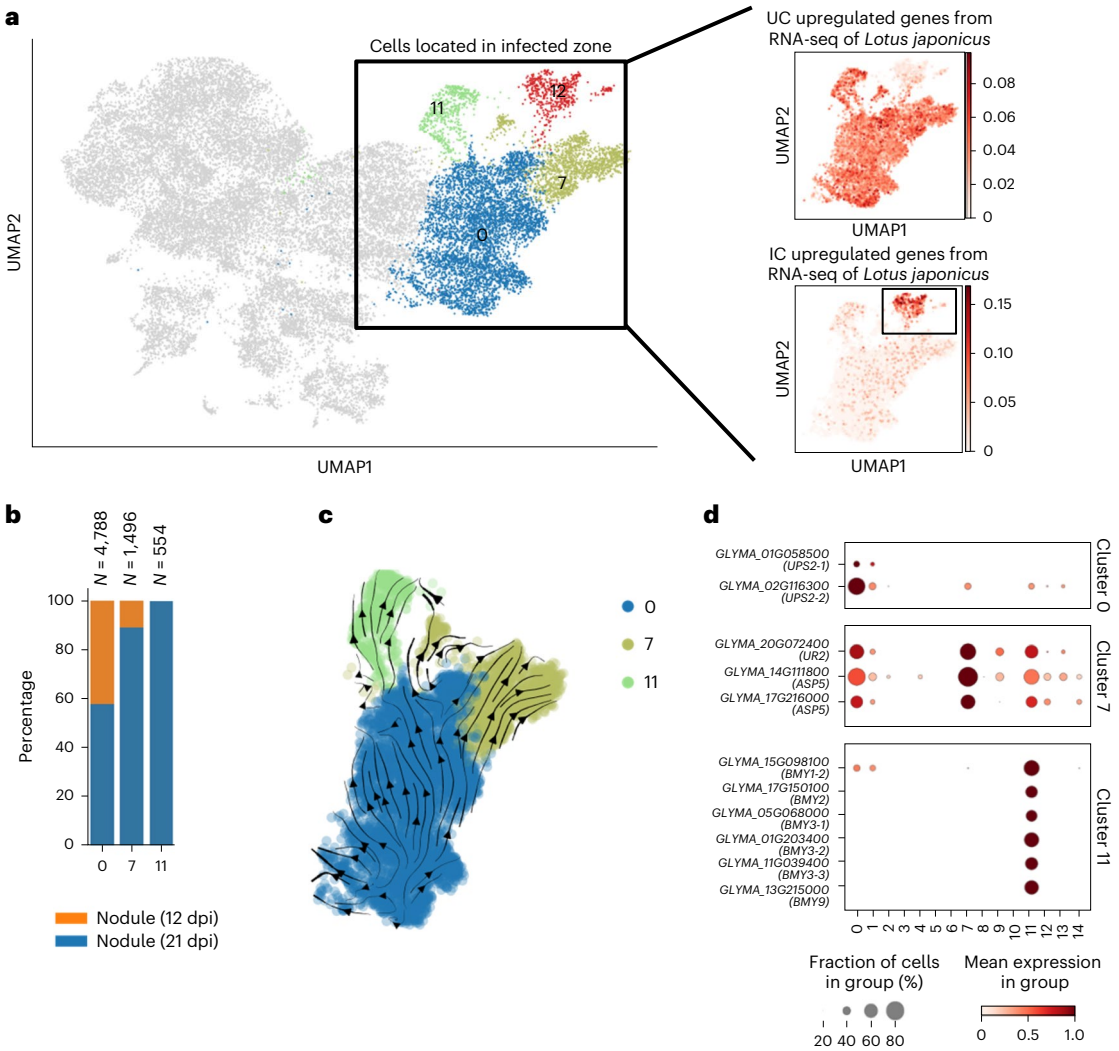


Fig. 2 | Dissection of central infected zone reveals distinct subtypes of nodule cells. a, The gene expression pattern of soybean orthologues of UCs and ICs highly expressed genes identified from *Lotus japonicus* in our soybean single-nucleus datasets. These genes were identified by manual isolation of UCs and ICs and single cell-type transcriptome in ref. 4. Only one-to-one orthologues were used here. Expression levels of gene sets were measured by AUC score. **b**, Bar chart representing the percentage of cells from different samples in each UC cluster. *N* indicates the cell numbers. **c**, Developmental trajectories of UCs inferred using Cellrank and CytoTRACE. Colours represent different IC clusters (0, 7, 11). The arrows refer to the most likely future direction of cell development. The arrow directions of clusters 7 and 11 are almost pointing from cluster 0 to the corresponding clusters, indicating that these cells develop from cluster 0. **d**, Dotplot representing the expression pattern of representative upregulated genes for each UC cluster. The size of the dot indicates the faction of cells expressing the gene, and the colour intensity indicates the expression level.

due to the scarcity of marker genes in soybean nodules, there are still many cell clusters that cannot be successfully assigned, especially those dominated by nodules (Extended Data Fig. 3b). To overcome this problem, we used Stereo-seq[9] to track the spatial expression of genes of the same developmental stage nodules (Fig. 1a and Extended Data Fig. 4). We classified the Stereo-seq data into 6 clusters on the basis of their spatial information and histological features, including CIZ cells (cluster 0), inner cortex (cluster 1), outer cortex (clusters 2, 4), epidermis (cluster 5) and vascular bundle (cluster 3) (Extended Data Fig. 5a), and identified the upregulated genes for each cluster of Stereo-seq data (Supplementary Data 6). With a deconvolution-based approach, we estimated the proportion of single-nucleus transcripts that belonged to each cluster of single-nucleus datasets for each bin, which was merged by several adjacent spots of Stereo-seq data (detailed in Supplementary materials). Therefore, we validated the cluster identities that we detected above and further assigned cluster 0 (in CIZ), 2 (outer cortex), 4 (outer cortex), 7 (in CIZ) and 11 (in CIZ) on the basis of their distribution over space (Fig. 1c,d). We also confirmed the results of the deconvolution-based approach by examining the expression patterns of these upregulated genes of Stereo-seq data in our single-nucleus libraries (Extended Data Fig. 5b and Supplementary Data 6). To validate our final annotation, we performed β-glucuronidase (GUS) staining and RNA in situ hybridization with cell-type-specific genes, and observed corresponding signals in nodules (Figs. 1e and 3c, and Extended Data Figs. 6 and7). In summary, here we successfully classified the major cell types of both root and nodule.

There are four clusters co-localized in the CIZ of nodules: 0, 7, 11 and 12 (Fig. 1d). Recently, uninfected cells (UCs) and infected cells (ICs) were manually isolated from nodules of *Lotus japonocus* and upregulated genes were identified in these two cell types[4]. We therefore investigated expression patterns of soybean orthologues of UC and IC genes from nodules of *Lotus japonocus* in our soybean single-nucleus datasets. Then, we identified cluster 0, 7 and 11 as UCs and assigned cluster 12 as ICs (Fig. 2a, and Supplementary Figs. 6 and 7, and Data 7). In UCs, cluster 0 is shared by nodules at two different developmental stages, while two clusters (7, 11) are almost only found in 21-dpi nodule cells

https://doi.org/10.1038/s41477-023-01387-z

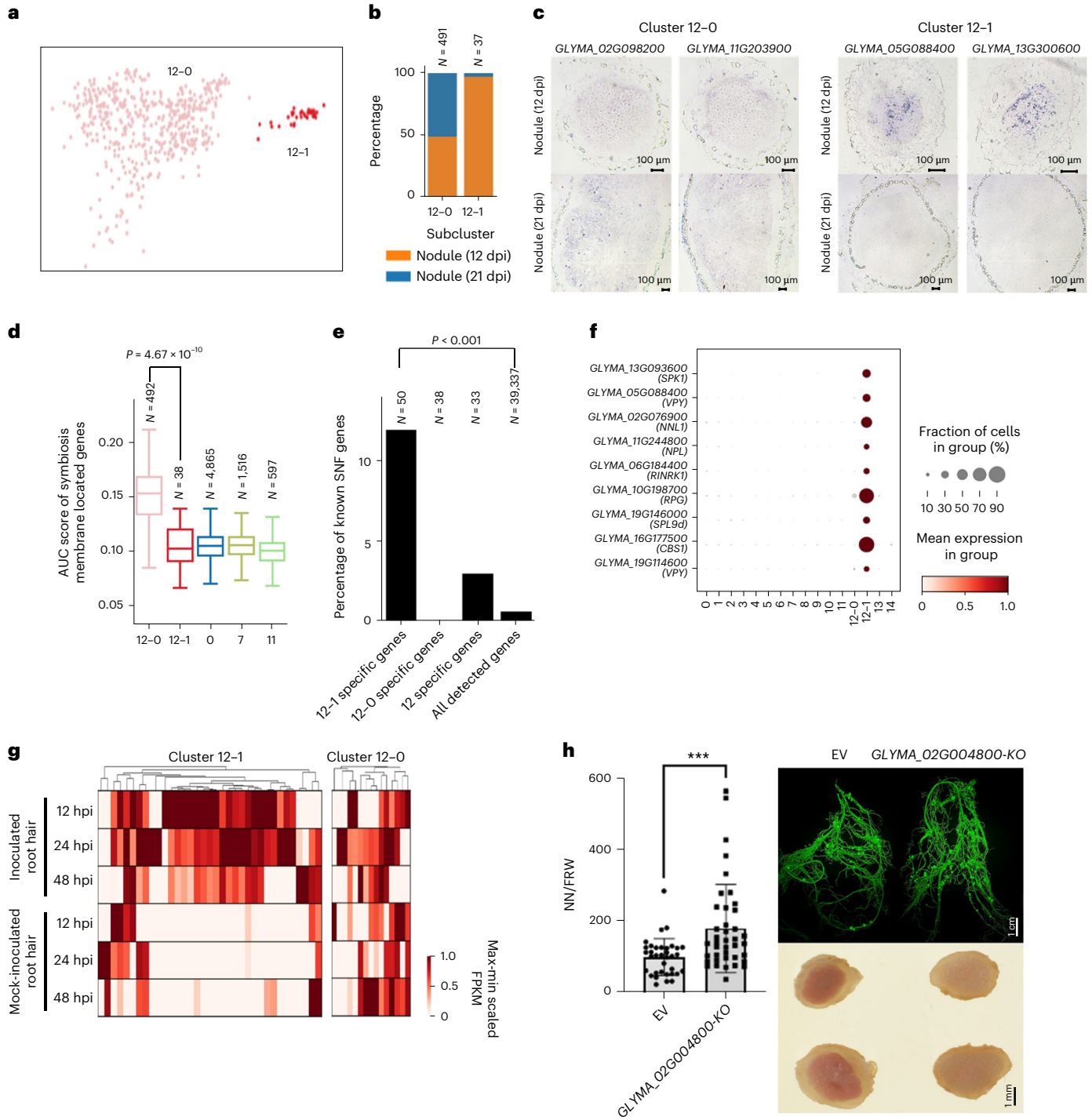


Fig. 3 | A rare subcluster of ICs is essential for nodule maturation and function. **a**, UMAP visualization of identified IC subclusters. **b**, Bar chart representing the percentage of cells from different samples in each sub-cell type. *N* indicates the cell numbers. **c**, Validation of cluster 12-0- and 12-1-specific marker genes by RNA in situ hybridization. These experiments were repeated in three independent assays and for each section; at least three nodules were analysed and all showed the same expression pattern. **d**, AUC score of genes encoding symbiosis membrane protein. *P* values were calculated using two-sided Mood's median test, $P > 0.05$ between cluster 12-1 and the remaining three UC clusters. Cell numbers are the calculated sum of the three libraries: 12-dpi nodules, 21-dpi nodules and roots. Boxplot centre, median; bounds of box, lower quartile (Q_1) and upper quartile (Q_3); minima, $Q_1 - 1.5\ (Q_1\text{-}Q_3)$; maxima, $Q_1 + 1.5\ (Q_1\text{-}Q_3)$. **e**, Percentage of known symbiotic nitrogen fixation genes reported in ref. 1 in different cell-type-specific gene sets. Calculation of the *P* value is detailed in Supplementary Fig. 10. *N* indicates the number of genes. **f**, Dotplot representing the expression pattern of 12-1-specific known symbiotic nitrogen fixation genes. The size of the dot indicates the faction of cells expressing the gene, and the colour intensity indicates the expression level. **g**, Heatmap showing the expression pattern of detected cluster-specific genes for subcluster 12-0 and 12-1 in inoculated and mock-inoculated root hair datasets. FPKM, fragments per kilobase of exon model per million mapped fragments. **h**, Nodulation phenotypes of the cluster 12-1-specific gene *GLYMA_02G004800* knockout line (*GLYMA_02G004800-KO*) after hairy root transformation. Left: nodule number (NN) per g root fresh weight (RFW). $^{***}P < 0.001$ indicate significant differences ($P = 6 \times 10^{-4}$) based on two-sided *t*-test. Error bars represent mean ± s.d. EV, empty vector. *KO*, transgenic knockout line. Top right: fluorescent image showing the expression of GFP which was used as an indicator for transgenic roots. Bottom right: representative transverse sections of nodules from *GLYMA_02G004800* knockout transgenic root and vector control roots (EV). These experiments were repeated in three independent assays and the same results were obtained ($N > 30$).

(Fig. 2b), indicating that cluster 7 and cluster 11 emerge at later stages during nodule maturation. To reveal the differentiation trajectory of these two UC clusters, we performed pseudo-time analysis with all the UC clusters (clusters 0, 7 and 11) using three mainstream trajectory inference algorithms (Fig. 2c and Extended Data Fig. 8) and found that the inferred direction of differentiation was from cluster 0 to cluster 7 and 11, suggesting that these two UC clusters developed from cluster 0 during nodule maturation. In tropical legumes such as soybean, ureides are the primary export forms in root nodules from currently fixed nitrogen. Ureides are reported to be mainly synthesized in UCs and enzymes responsible for ureides biosynthesis present a higher specific activity in the UCs[10,11]. For ureide biogenesis, the uricase and aspartate aminotransferase genes, which are expressed in nodules, are expressed in all three UC clusters and especially upregulated in UC cluster 7 (Fig. 2d), while for ureide transportation, 2 of 3 ureide permease genes are mainly expressed in UC cluster 0 (Fig. 2d). These results reveal a complex compartmentalization in UCs during ureide production and transportation in soybean nodules. Moreover, we found that expression of 6 of 8 beta amylase genes is significantly upregulated in cluster 11 (Fig. 2d), and the pathways associated with polysaccharide catabolic process, starch catabolism, are also activated, indicating that cluster 11 is involved in energy supply for symbiotic nitrogen fixation (Supplementary Fig. 8). Taken together, these results reveal that the UCs can be divided into different functionally specialized sub-cell types and two of them emerge at later stages during nodule development, which can facilitate the exchange of nutrient and energy sources required for symbiosis. We then focused on infected cells, the core sites of symbiotic nitrogen fixation (SNF). Consistent with previous reports, all leghemoglobin genes[12] and some nodulin genes were highly expressed in cluster 12 (Extended Data Fig. 9). We also found that 8 genes encoding sugar transporter and 5 genes belonging to isopropylmalate synthase families are listed in our identified upregulated genes of cluster 12 (Supplementary Data. 3), which is consistent with our gene ontology (GO) enrichment analysis (Supplementary Fig. 9), demonstrating active carbon and nitrogen exchange between soybean and rhizobia in ICs. By reclustering ICs, we found that they could be further divided into two sub-cell types (12-0 with 492 nuclei and 12-1 with 38 nuclei) (Fig. 3a). Subcluster 12-0 is shared by nodules at two different developmental stages, but the small subcluster 12-1 is almost exclusively occupied by the 12-dpi immature nodule (Fig. 3b). Since only a small number of nuclei were clustered to 12-1, to rule out the possibility of overclustering, we verified the expression pattern of specifically expressed genes of these two subclusters using RNA in situ hybridization. These genes are expressed in the CIZ but showed distinct expression pattern, and the 12-1-specific genes were mainly detected in immature nodules, proving that these are two different cell subtypes (Fig. 3c and Extended Data Fig. 7). The expression levels of genes encoding symbiosome membrane protein[13] were much higher in subcluster 12-0 than in 12-1 and other subclusters in UCs (Fig. 3d), indicating a more active movement of solutes between symbionts in subcluster 12-0 in the nodule infection zone. We next checked subcluster 12-1-specific genes and found that nearly 12% of the genes (6/50) are included in known SNF genes previously reported[1] (Extended Data Fig. 10). This proportion is significantly higher than in clusters 12, 12-0 and all detected clusters (Fig. 3e and Supplementary Fig. 10). Three of the remaining 44 genes have been reported as SNF genes in recent years[14,15] (Fig. 3f and Supplementary Data 8). We further found that all these 9 SNF genes, including *SPK1*[16], two homologs of medicago *VPY*[17], *NNL1*[14], *NPL*[18], *RINRK1*[19], *RPG*[20], *SPL9d*[15] and *CBS1*[21], are involved in the formation of infection threads (ITs). ITs are formed in root hair after rhizobium attachment and they assist rhizobium in reaching and finally being released into developing nodules. We analysed the expression of 12-1 cluster genes in soybean root hair in the early stage of rhizobial infection (12, 24 and 48 hpi (hours post-infection)) using public datasets[22] and found that 60% of these genes (21/35) are expressed only after rhizobia inoculation (Fig. 3g). In contrast, of the cluster 12-0-specific genes, only 2 were induced after induction. These results imply that cluster 12-1 could be involved in IT extension and rhizobia release in ICs during nodule maturation, and genes that are specifically expressed in such cells may play a critical role in the interaction between soybean and rhizobium, and the final stages of symbiosis establishment. Then, we explored the function of a subcluster 12-1-specific gene *GLYMA_02G004800* in nodulation (expression pattern is plotted in Extended Data Fig. 10). When the gene was knocked out by CRISPR-Cas9 using hairy root transformation, we found that the nodule number was increased, and the infection zones were white in nearly 50% of transgenic nodules (14/30), providing clues to the importance of cluster 12-1 during nodule maturation (Fig. 3h).

Overall, we provide a cellular atlas by combining single-cell data with spatial transcriptomic data. On the basis of this atlas and experiment results, we identified rare cell subtypes and their important roles for nodule maturation and function. To help the community explore the heterogeneity of different cell types in soybean nodules, we also present a web server (https://zhailab.bio.sustech.edu.cn/single_cell_soybean) to facilitate the use of the datasets generated in this study. In conclusion, we provide a data resource that will contribute to learning the regulatory network of nodule maturation at the single-cell level in the future.

Methods

Plant growth and nodulation

Wild-type soybean (*Glycine max* L. cv Williams 82) seeds were disinfected with chlorine (100 ml NaClO + 4 ml concentrated HCl) and grown on moist sterile filter paper at 22 °C in the dark for 3 d. After germination, the seeds were transferred to pots filled with sterile mixed vermiculite and perlite (2:1, v/v) in a growth chamber with 16 h light/8 h dark cycle and relative humidity of 35% at 28 °C. Nitrogen-free nutrient solution (0.5 mM $MgSO_4$, 0.2 mM $CaCl_2$, 0.15 mM K_2HPO_4, 1 mM K_2SO_4, 0.02 mM $FeCl_3$, 0.5 μM H_3BO_3, 0.1 μM $MnSO_4$, 0.15 μM $ZnSO_4$, 0.04 μM $CuSO_4$, 2.5 pM $NaMoO_4$, 2.5 pM $CoCl_2$ and 2.5 pM $NiSO_4$) was poured twice a week. After the cotyledon spread out, about a week after transferring to pots (about 7 d after germination), the roots were infected with rhizobium strain USDA110 for nodulation.

Hairy root transformation

Agrobacterium rhizogenes-mediated transformation was performed as previously reported[23]. The binary vectors harbouring the gene construct of interest and a green fluorescent protein (GFP) label (Supplementary Data 9) that can be used to identify the transgenic roots were introduced into *Agrobacterium rhizogenes* K599. The primary root was cut off at 1 cm below the cotyledons and excised from 7-day-old soybean (Wm82) seedlings after the real leaves were unfolded. After inoculation with *Agrobacterium*, the infected seedlings were placed into moist sterile vermiculite. The GFP-negative induced hairy roots were removed every 7 d, and the transgenic positive roots were retained until the plants had robust transgenic roots that could sustain plant growth (approximately 2 weeks). Plants were then infected with rhizobia USDA110 for nodulation.

Histochemical analysis of GUS activity

Promoter fragments (~2 kb, upstream of ATG) of marker genes identified by single-nucleus RNA sequencing were amplified (with primers in Supplementary Data 9) and cloned into the GFP-labelled binary vectors to generate the promoter:GUS-GFP constructs pCambia1300-pro-GUS-GFP (Supplementary Data 9). The resulting plasmids were introduced into soybean (Wm82) by *Agrobacterium* K599 as described above. GUS staining was performed as previously described[24,25]. In brief, the fresh nodules were immersed in 5-bromo-4-chloro-3-indolyl-b-D-glucuronic acid (X-Gluc)-containing solution and vacuum infiltrated for 30 min. Histochemical staining for GUS activity was performed at 37 °C for 12 h. After staining, the nodules

were fixed with FAA buffer (5% formaldehyde, 5% acetic acid and 50% ethanol) overnight, washed with ethanol and embedded in paraffin. Then, the nodule was sliced transversely into 25–35 µm sections with a microtome (ZEEDO, HS-3315). GUS activity was observed with a light microscope (BSP-8N) equipped with a camera. For each construct, three individual transgenic lines were generated and at least three nodules from each line were analysed, and all had the same expression pattern.

CRISPR/Cas9-mediated genome editing in soybean hairy roots

CRISPR/Cas9 technology was used to knock out *GLYMA_02G004800*. First, the CDS sequence of *GLYMA_02G004800* was analysed using the online software CRISPR-P (http://crispr.hzau.edu.cn/cgi-bin/CRISPR2/CRISPR), and three guide RNAs (GAGAAAGCAGAGGAGAAAGG, GAGCTCTCTTCCAACCCGGG and TTTGAATGCGAGTGCCACCC) were obtained. The target sequences were cloned into the same single guide RNA (sgRNA) expression cassettes of the pBluescript SK(+)-LjU6-BbsI-gRNA vector (Supplementary Data 9) in series. The resulting construct *GLYMA_02G004800-gRNA* was validated by sequencing. Then the U6 promoter and expression cassettes were cut by *Kpn* I and *Xba* I and ligated to the pCambia1300-GFP-Cas9 vector (Supplementary Data 9). The plasmids were then transformed into *Agrobacterium rhizogenes* strain K599 for hairy root transformation as described above. A preliminary experiment was conducted to test the editing efficiency of the vector. Effective edits occurred in 12 of the 15 GFP-positive independent roots. Three weeks after inoculation with rhizobia, we first sequenced GLYMA_02G004800 from each GFP-positive root, then the number of nodules and fresh root weight on edited roots were analysed.

RNA in situ hybridization

Nodules collected at 12 and 21 dpi were fixed in FAA solution for 24 h at 4 °C and dehydrated in an ethanol series, cleared in a xylene series and embedded in paraffin. Then, 10 µm sections were prepared using a microtome (ZEEDO, HS-3315). For probe labelling, specific 500 bp sequences of target genes were amplified from the complementary DNA (cDNA) of Wm82 with primers (Supplementary Data 9) and cloned to pEASY-Blunt cloning vector (TransGen Biotech, CB111-01). With the resulting vectors as templates, both SP6 and T7 promoter-fused fragments were amplified (with primers in Supplementary Data 9) and purified. Then the digoxigenin-labelled antisense and sense probes were in vitro transcribed with these fragments as templates, respectively using SP6 and T7 RNA polymerase with a DIG RNA labelling kit (Roche, 11175025910). The paraffin on the sections was dissolved with xylene, and the tissue was digested with Proteinase K (Sigma, P2308) and dehydrated with graded alcohol. The prepared probe buffer was added to the tissue sections and incubated for 12 h at 55 °C in a humidified box. After hybridization, the sections were incubated with anti-digoxin antibody (Roche, 11093274910) for 90 min. Then the sections were washed and incubated in a colour reaction solution (Roche, 11681451001) containing blue tetrazolium chloride and 5-bromo-4-chloro-3-indolylphosphate for 36 h in a dark humidified box. Photographs were taken under bright-field illumination using a microscope (BSP-8N). For each probe, at least three nodules were analysed, and all showed the same expression pattern.

Nuclei isolation and 10× single-nucleus RNA-seq library construction

The fresh nodules at 12 and 21 dpi and roots at 21 dpi in the vicinity of mature nodules were collected for single-nucleus RNA-seq. The two time points respectively represent the early and mature stages of soybean nodule development[5,26]. For the 12-dpi nodules, we selected those with a diameter of about 2 mm. For the 21-dpi nodules, we collected representative nodules with a diameter of about 6–8 mm (Supplementary Fig. 1). Nuclei isolation for root and nodules was performed as previously reported[27]. In brief, the root and nodules were chopped in ice-cold 1× Nuclei isolation buffer (NIB, MilliporeSigma, CELLYTPN1) with 1 mM dithiothreitol (Thermo Fisher, R0861), 1× protease inhibitor (Sigma, 4693132001) and 0.4 U µl^{-1} murine RNase inhibitor (Vazyme, R301-03). Then the lysate was filtered with a pre-wet 40 µm strainer and centrifuged at 500 *g* for 5 min at 4 °C. The nuclei pellet was resuspended with 500 µl NIB. For sorting, the nuclei were stained with 4,6-Diamidino-2-phenylindole (DAPI) and loaded into a flow cytometer with a 100 µm nozzle. PBS (1×, 1 ml) with 1% BSA and 0.4 U µl^{-1} murine RNase inhibitor was used as the collection buffer. At least 100,000 nuclei were collected on the basis of the DAPI signal and the nuclear size. The sorted nuclei were pelleted at 4 °C and 500 *g* for 5 min, and then resuspended in 50 µl 1× PBS with 1% BSA and 0.4 U µl^{-1} murine RNase inhibitor. After checking the quality of nuclei and counting under a microscope using the DAPI channel, 20,000–30,000 nuclei were loaded onto the 10x Genomics Chip. Library construction for Illumina sequencing was performed with 10x Chromium Single Cell 3′ Solution v3.1 kit as described previously[28].

Single-nucleus data analysis

Raw reads were mapped to the Glycine_max_v2.1 reference genome[29] by Cell Ranger (v6.0.0) using default parameters but enabling the 'include-introns' option. The matrix was subsequently loaded onto the SCANPY package[30] (v1.8.0) for analysis. For quality control, doublets were removed by ScDblFinder[31] (v1.10.0) with default parameters, except that the expected doublet ratio parameter was determined by the formula $0.01 \times N \div 1{,}000$ (N is the Cell Ranger recovered numbers). Then genes expressed in less than 10 nuclei were discarded; only cells with gene counts between 400 and 4,000, and UMI counts between 600 and 6,000 were kept. To evaluate the effect of quality control parameters, we used stringent (gene counts between 600 and 3,000, UMI counts between 800 and 4,000) and looser criteria (only remove putative doublets) and found that the clustering results were almost unaffected (Supplementary Fig. 11). The matrix was integrated by scVI[7] (v0.16.0) following the manual's tutorial (https://docs.scvi-tools.org/en/0.12.2/user_guide/notebooks/harmonization.html), except that 'n_top_genes' was set to 5,000. We therefore performed leiden algorithm ('resolution' set to 0.3) on nearest-neighbour graph (performed by 'scanpy.pp.neighbors' function, 'n_neighbors' set to 15) built on scVI lower-dimension space for clustering and used the Uniform Manifold Approximation and Projection (UMAP) algorithm (performed by 'scanpy.tl.umap' function, 'resolution' set to 0.3) to visualize the distribution of the data in scVI space. For annotation, we used existing experimentally validated marker genes (Supplementary Data 5) to unveil the identity of each cell cluster and used scANVI (v0.16.0)[32] to validate the annotation results using public single-cell data on *Arabidopsis* roots[8] with the following steps: first, we identified one-to-one orthologues using OrthoFinder[33] (v2.5.4) with default parameters, then combined datasets of the two species on the basis of the orthologue information. Then, annotation transfer was conducted following the manual tutorial (https://docs.scvi-tools.org/en/0.12.2/user_guide/notebooks/harmonization.html), except that the 'n_layer' parameter (which controls the number of encoder and decoder hidden layers of the neural network) of the scANVI model was set to 4. For the reclustering of cluster 12, we used the leiden algorithm and specified the 'restrict_to' parameter in 'scanpy.tl.leiden' ('resolution' set to 0.1).

We identified upregulated genes and specifically expressed genes of each cell cluster using the cellex algorithm[34] (v1.2.2). Upregulated genes are defined as having a specificity score greater than 0.75 and detectable expression in at least 10% of the cells in the corresponding cluster. For cluster-specific expressed genes identified, genes were filtered if they are expressed in less than 20% of cells of the corresponding clusters or more than 1% in the rest of the cells. Then, genes ranked in the top 50 in terms of specificity score were retained. To identify differentially expressed genes (DEGs) of different samples in each cluster, we used the Wald test implemented by diffxpy (v0.7.4). For each sample,

if the number of nuclei in a cluster was below 100, this sample was no longer identified for DEGs in this cluster. After identification, we further filtered DEGs that did not satisfy the following criteria: (1) adjusted *P* value <0.01 and (2) fold change with $\log_2$-transform >4.

In the trajectory inference step, we combined Cellrank[35] (v1.5.1) and CytoTRACE[36] (wrapped in Cellrank) to track the dynamic changes in UCs following the manual's tutorial (https://cellrank.readthedocs.io/en/stable/beyond_rna_velocity.html). First, we extracted all the clusters of UCs and then removed genes that were expressed in no more than 10 cells. Then we computed the first- and second-order moments for each cell ('n_pcs' and 'n_neighbors' set to 30) as recommended by Cellrank's tutorial. By combining the moments and the differentiation scores of each cell calculated by CytoTRACE, we calculated the matrix of directed transition probabilities using Cellrank.

Moreover, we used scVelo[37] (v0.2.4) and monocle3[38] (v1.0.0) to verify the above trajectory inference results. In brief, for scVelo, we first calculated the spliced and unspliced matrices using velocyto, and then calculated the first- and second-order moments on the basis of the matrices (total, spliced and unspliced). Then we calculated full splicing kinetics and estimated the velocities for each gene. Finally, we computed the velocity graph and calculated the RNA-velocity-directed partition-based graph abstraction (PAGA) graph on the basis of the above-mentioned results. For monocle3, we first used the 'preprocess_cds' and 'align_cds' functions in monocle3 package to preprocess the raw count matrix. Then, we used the 'reduce_dimension' function to calculate the principal component analysis (PCA) embedding of our datasets. Finally, we used 'cluster_cell', 'learn_graph' and 'order_cells' functions to get the inferenced trajectory.

In all the above steps, we used the AUCell (v1.18) package[39] to calculate the area under the curve (AUC) score, the clusterProfiler[40] (v4.4) package to perform GO enrichment analysis and the rpy2 (v3.5) package to implement invocation of the R package.

Determining the resolution parameter for clustering

For the leiden algorithm, the 'resolution' parameter has a substantial impact on the number of clusters. To choose an optimal resolution parameter, we checked the results for resolution parameters ranging from 0 to 1.5 in 0.1 intervals. Then we annotated the resulting root dataset by label transferring from the *Arabidopsis* root data and retained only successfully annotated nuclei. The degree of similarity between clustering and annotation results was measured using Adjusted Mutual Information (AMI). We found that the AMI was highest when the resolution was set to 0.2 (Supplementary Fig. 12a). However, at this resolution, most cells in cluster 6 could not be annotated and cluster 6 was merged with clusters 4 and 5 when the unsuccessfully annotated nuclei were added (Supplementary Fig. 12b). When the resolution was set to 0.5, we found that both clusters 0 and 1 were split (Supplementary Fig. 13a), making it difficult to identify specifically expressed genes from the newly generated two subclusters that came from the same cluster (Supplementary Fig. 13b). Taken together, we used a final resolution of 0.3, which also had a high AMI score and gave a reasonable clustering result.

Choice of software tools

For doublet detection, we chose scDblFinder because of its best performance[41,42]. For data integration, benchmark research shows that scVI and scanorama[43] have the best performance[44]. We tested these two methods and found that scVI gave better integration results (Supplementary Fig. 14). Moreover, we also tried another two integration algorithms, Harmony[45] (harmonypy, v0.0.6) and Seurat CCA[46] (v4.1.1), and found that the clustering results from scVI remained clearly bounded on the UMAP plots from the algorithms (Supplementary Fig. 14). Hence, we used the integration results from scVI for downstream analysis.

For trajectory inference, we first chose CytoTRACE because it does not require manual specification of developmental start sites. To validate the results, we added the results of RNA-velocity analysis and Monocle3. All three software support the hypothesis that clusters 7 and 11 develop from cluster 0.

Reclustering and annotation of vascular bundle cells

We extracted the nuclei of clusters 3 and 9 and then concatenated them with the scRNA-seq data of mature *Arabidopsis* vascular bundle cells[8]. In this step, we integrated the obtained data and performed label transfer using scANVI (Supplementary Fig. 5a). We set the 'n_layer' parameter to 5 because distinguishing different cell subtypes requires a higher model complexity. Subsequently, we calculated the euclidean distance of soybean cells in scANVI latent space and constructed the nearest-neighbour graph. Then we clustered these cells using leiden algorithm ('resolution' set to 0.2) and visualized them using UMAP. A total of 7 subclusters were obtained, named vs-0 to vs-6 (Supplementary Fig. 5b).

After clustering, we combined the label transfer results and the specifically expressed genes of subtypes of vascular bundles identified in *Arabidopsis* to annotate our soybean datasets (Supplementary Fig. 5c). These two methods were consistent in the annotation of subcluster vs-0 (xylem pole pericycle), subcluster vs-5 (phloem pole pericycle) and subcluster vs-6 (metaphloem and companion cell). Although most of the cells in subcluster vs-4 were annotated as xylem in label transfer results, xylem-specific expressed genes were only highly expressed in a small number of them, so we defined subcluster vs-4 as xylem-like cells (Supplementary Fig. 5d). Similarly, we defined subcluster vs-2 as phloem-like cells (Supplementary Fig. 5d). Since the two methods were not consistent on subcluster vs-1 and subcluster vs-3, we marked them as unknown subtypes (Supplementary Fig. 5d). Moreover, we identified upregulated expressed genes in different cell subtypes using the above-mentioned criteria (Supplementary Fig. 5e and Data 10).

Identification of ICs and UCs

We used upregulated genes of ICs and UCs of *Lotus japonicus* as previously identified[47] to distinguish UCs and ICs (Supplementary Data 7). Their homologues in soybean were identified by OrthoFinder[33]. We first used all homologues to calculate the AUC score of the two gene sets and identified clusters 0, 7 and 11 as UCs and cluster 12 as ICs (Supplementary Fig. 6). However, given that the soybean genome is highly duplicated[29] and the duplicated genes experience weaker purifying selection[48], we analysed the expression patterns of paralogous genes (2,561 pairs) at single-nucleus level and investigated the correlations of expression. We found a higher correlation and a lower expression difference between paralogous genes relative to others (Supplementary Fig. 15a,b), but only 629 pairs of genes were significantly more correlated than non-homologous genes (Supplementary Fig. 15c,d). Therefore, we used soybean one-to-one orthologues to verify our annotation and obtained consistent results (Fig. 2a, and Supplementary Fig. 7 and Data 7). The annotation results were also confirmed by previously reported IC-specific genes (Extended Data Fig. 2).

Expression difference analysis between paralogous genes

Dene duplication events were identified in ref. 49. Of these, 2,561 pairs of duplication events containing only two genes were kept for subsequent analysis. We first combined data from the same cell cluster to obtain a pseudo-bulk dataset, then used the log-transformed counts per million (CPM) value for calculating the Spearman's rank correlation coefficient and expression difference for each duplication event. After shuffling, we recalculated these two values as control. We used a bootstrap-based method to determine whether the correlation between homologous genes was significantly higher than those of other genes. We divided all detected genes into 50 equal parts according to the average expression of genes in all nuclei. For each gene, we took a randomly selected gene from the part where the paralogue belonged and calculated its correlation coefficient. The above steps were repeated a thousand times. The *P* values were then calculated on the basis of the rank of the original correlation coefficients.

Stereo-seq and data processing

Fresh nodules at 12 and 21 dpi were used for Stereo-seq analysis. Stereo-seq chip preparation and sequencing were performed at the Beijing Genomics Institute (BGI) as previously reported[9]. In brief, the nodules were quickly frozen in liquid nitrogen pre-cooled isopentane and cut into sections at a thickness of 10 mm with a Leica CM1950 cryostat. Tissue sections were adhered to the Stereo-seq chip (generated by BGI, China) surface and fixed with methanol. For histological examination, tissue sections adjacent to it were adhered to glass slides and stained with Fluorescent Brightener 28. Images were acquired with a Motic fluorescence microscope. After washing with 0.1x SSC buffer, the sections on chip were permeabilized using 0.1% pepsin (Sigma, P7000) in 0.01 M HCl buffer. Then the RNAs were released from the permeabilized tissue by washing with 0.1x SSC buffer and captured by the DNA nanoball (DNB). After in situ reverse transcription overnight at 42 °C using SuperScript II (Invitrogen, 18064-014) and tissue removal, cDNA-containing chips were then subjected to Prepare cDNA Release Mix treatment at 55 °C overnight. The released cDNA was purified with 0.8x VAHTSTM DNA Clean beads and amplified with KAPA HiFi Hotstart Ready mix (Roche, KK2602) for 15 cycles. After quantification by Qubit, a total of 20 ng of PCR products were then fragmented with in-house Tn5 transposase. The fragmented DNA were amplified with KAPA HiFi Hotstart Ready mix and Stereo-seq-library primer pairs. PCR products were purified using AMPure XP beads (0.63 and 0.153). After DNB generation, the libraries were finally sequenced on an MGI DNBSEQ-Tx sequencer. The Stereo-seq raw data were preprocessed using SAW (v2.1.0) to generate a spot-gene matrix[9]. In this step, the bin sizes of the 12-dpi nodule and 21-dpi nodule libraries were set to 50 ×50 and 80 ×80, respectively, to obtain the bin-gene matrix. Stained images of sections under a bright-field microscope were overlayed onto the resulting matrix. Then we performed deconvolution on the basis of single-nucleus sequencing data using destVI[50] (v0.16.0). In this step, we removed the clusters obtained from single-nucleus data with less than 500 nuclei, and filtered genes that could not be detected in both the 10x library and the Stereo-seq library. Then we used 'scanpy.pp.highly_variable_genes' to obtain highly variable genes. We set up the CondSCVI model using our single nucleus RNA-seq datasets ('n_layer' set to 4), and then trained the destVI model with Stereo-seq datasets to perform the deconvolution on the basis of the pre-trained CondSCVI model. Using the 'get_proportions' function, the proportion of transcripts from different cell clusters contained in each Stereo-seq bin was estimated.

Before clustering, we used SCTransform[51] (v2) with default parameters to normalize the Stereo-seq datasets. Then, we used mefisto[52] and mofa[53] in muon[54] (v0.1.2) to obtain the embedding of each bin on the basis of the normalized values and spatial information. In this step, 'n_factor' was set to 5 and other parameters were as recommended (https://muon-tutorials.readthedocs.io/en/latest/mefisto/3-MEFISTO-ST.html). We performed the leiden algorithm ('resolution' set to 0.2) on nearest-neighbour graphs ('n_neighbors' set to 15) built on mofa lower-dimension space for clustering and used the UMAP algorithm ('resolution' set to 0.3) to visualize the distribution of the data. For upregulated genes, we used the same criteria as the snRNA-seq analysis, but did not filter low-expression genes due to high sparsity of our Stereo-seq datasets.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data generated in this study are deposited in the China National Center for Bioinformation with accession PRJCA009893. Raw sequencing data are deposited in GSA with accession CRA007122 and processed data are deposited in OMIX with accession OMIX002290. Source data are provided with this paper.

Code availability

The source code to reproduce this project can be accessed at https://github.com/ZhaiLab-SUSTech/soybean_sn_st.

References

1. Roy, S. et al. Celebrating 20 years of genetic discoveries in legume nodulation and symbiotic nitrogen fixation. *Plant Cell* **32**, 15–41 (2020).
2. Roux, B. et al. An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing. *Plant J.* **77**, 817–837 (2014).
3. Roux, B., Rodde, N., Moreau, S., Jardinaud, M.-F. & Gamas, P. in *Plant Transcription Factors* (eds Yamaguchi, N. et al.) 191–224 (Springer, 2018).
4. Wang, L. et al. Single cell-type transcriptome profiling reveals genes that promote nitrogen fixation in the infected and uninfected cells of legume nodules. *Plant Biotechnol. J.* **20**, 616–618 (2022).
5. Fan, W. et al. Rhizobial infection of 4C cells triggers their endoreduplication during symbiotic nodule development in soybean. *New Phytol.* **234**, 1018–1030 (2022).
6. Ye, Q. et al. Differentiation trajectories and biofunctions of symbiotic and un-symbiotic fate cells in root nodules of *Medicago truncatula*. *Mol. Plant* **15**, 1852–1867 (2022).
7. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
8. Shahan, R. et al. A single-cell *Arabidopsis* root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Dev. Cell* **57**, 543–560 (2022).
9. Xia, K. et al. The single-cell stereo-seq reveals region-specific cell subtypes and transcriptome profiling in *Arabidopsis* leaves. *Dev. Cell* **57**, 1299–1310 (2022).
10. Newcomb, E. H. & Tandon, S. R. Uninfected cells of soybean root nodules: ultrastructure suggests key role in ureide production. *Science* **212**, 1394–1396 (1981).
11. Hanks, J. F., Schubert, K. & Tolbert, N. Isolation and characterization of infected and uninfected cells from soybean nodules: role of uninfected cells in ureide synthesis. *Plant Physiol.* **71**, 869–873 (1983).
12. Appleby, C. A. Leghemoglobin and *Rhizobium* respiration. *Annu. Rev. Plant Physiol.* **35**, 443–478 (1984).
13. Luo, Y., Liu, W., Sun, J., Zhang, Z.-R. & Yang, W.-C. Quantitative proteomics reveals key pathways in the symbiotic interface and the likely extracellular property of soybean symbiosome. *J. Genet. Genomics* **50**, 7–19 (2022).
14. Zhang, B. et al. Glycine max NNL1 restricts symbiotic compatibility with widely distributed bradyrhizobia via root hair infection. *Nat. Plants* **7**, 73–86 (2021).
15. Yun, J. et al. The miR156b-GmSPL9d module modulates nodulation by targeting multiple core nodulation genes in soybean. *New Phytol.* **233**, 1881–1899 (2022).
16. Liu, J., Liu, M. X., Qiu, L. P. & Xie, F. SPIKE1 activates the GTPase ROP6 to guide the polarized growth of infection threads in *Lotus japonicus*. *Plant Cell* **32**, 3774–3791 (2020).
17. Murray, J. D. et al. Vapyrin, a gene essential for intracellular progression of arbuscular mycorrhizal symbiosis, is also essential for infection by rhizobia in the nodule symbiosis of *Medicago truncatula*. *Plant J.* **65**, 244–252 (2011).
18. Xie, F. et al. Legume pectate lyase required for root infection by rhizobia. *Proc. Natl Acad. Sci. USA* **109**, 633–638 (2012).
19. Li, X. et al. Atypical receptor kinase RINRK1 required for rhizobial infection but not nodule development in *Lotus japonicus*. *Plant Physiol.* **181**, 804–816 (2019).

20. Arrighi, J.-F. et al. The RPG gene of *Medicago truncatula* controls *Rhizobium*-directed polar growth during infection. *Proc. Natl Acad. Sci. USA* **105**, 9817–9822 (2008).
21. Sinharoy, S. et al. A *Medicago truncatula* cystathionine-β-synthase-like domain-containing protein is required for rhizobial infection and symbiotic nitrogen fixation. *Plant Physiol.* **170**, 2204–2217 (2016).
22. Libault, M. et al. Complete transcriptome of the soybean root hair cell, a single-cell model, and its alteration in response to *Bradyrhizobium japonicum* infection. *Plant Physiol.* **152**, 541–552 (2010).
23. Fan, Y.-l. et al. One-step generation of composite soybean plants with transgenic roots by *Agrobacterium rhizogenes*-mediated transformation. *BMC Plant Biol.* **20**, 208 (2020).
24. Zhao, Y., Wang, T., Zhang, W. & Li, X. SOS3 mediates lateral root development under low salt stress through regulation of auxin redistribution and maxima in *Arabidopsis*. *New Phytol.* **189**, 1122–1134 (2011).
25. Oh, H.-S. et al. The *Bradyrhizobium japonicum* hsfA gene exhibits a unique developmental expression pattern in cowpea nodules. *Mol. Plant Microbe Interact.* **14**, 1286–1292 (2001).
26. Pessi, G. et al. Genome-wide transcript analysis of *Bradyrhizobium japonicum* bacteroids in soybean root nodules. *Mol. Plant Microbe Interact.* **20**, 1353–1363 (2007).
27. Thibivilliers, S., Anderson, D. & Libault, M. Isolation of plant root nuclei for single cell RNA sequencing. *Curr. Protoc. Plant Biol.* **5**, e20120 (2020).
28. Long, Y. et al. FlsnRNA-seq: protoplasting-free full-length single-nucleus RNA profiling in plants. *Genome Biol.* **22**, 66 (2021).
29. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
30. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
31. Germain, P.-L., Lun, A., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research* **10**, 979 (2022).
32. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
33. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
34. Timshel, P. N., Thompson, J. J. & Pers, T. H. Genetic mapping of etiologic brain cell types for obesity. *eLife* **9**, e55851 (2020).
35. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
36. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
37. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
38. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
39. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
40. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
41. Xi, N. M. & Li, J. J. Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. *STAR Protoc.* **2**, 100699 (2021).
42. Xi, N. M. & Li, J. J. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Syst.* **12**, 176–194 (2021).
43. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
44. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
45. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
46. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
47. Wang, L. et al. Single cell-type transcriptome profiling reveals genes that promote nitrogen fixation in the infected and uninfected cells of legume nodules. *Plant Biotechnol. J.* **20**, 616–618 (2022).
48. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3**, research0008.1 (2002).
49. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).
50. Lopez, R. et al. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat. Biotechnol.* **40**, 1360–1369 (2022).
51. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol.* **23**, 27 (2022).
52. Velten, B. et al. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* **19**, 179–186 (2022).
53. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
54. Bredikhin, D., Kats, I. & Stegle, O. Muon: multimodal omics analysis framework. *Genome Biol.* **23**, 42 (2022).

Acknowledgements

This work was supported by the Key Research Program of the Chinese Academy of Sciences, Grant No. ZDRW-ZS-2019-2; Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA28030100 and XDA24010205; the Agricultural Science and Technology Innovation Program; the CAS Project for Young Scientists in Basic Research (YSBR-011) and NSFC General Projects (32272101). The group of J.Z. was supported by a National Key R&D Program of China Grant (2019YFA0903903); an NSFC grant to J.Z. (31871234); the Shenzhen Sci-Tech Fund (KYTDPT20181011104005); the Key Laboratory of Molecular Design for Plant Cell Factory of Guangdong Higher Education Institutes (2019KSYS006); the Stable Support Plan Program of Shenzhen Natural Science Fund Grant (20200925153345004); and the Center for Computational Science and Engineering at Southern University of Science and Technology.

Author contributions

Z.L., X.K., Y.L., S.L., W.C. and Z.Z. performed the experiments. Z.L., Y.L., H.Z. and J.J. analysed the data. Z.Y., J.Z., Z.L., X.K. and Y.L. wrote the manuscript. Z.Y. and J.Z. oversaw the study. L.Q. and X.S. provided conceptual insight.

Competing interests

The authors declare no competing interests.

Additional information

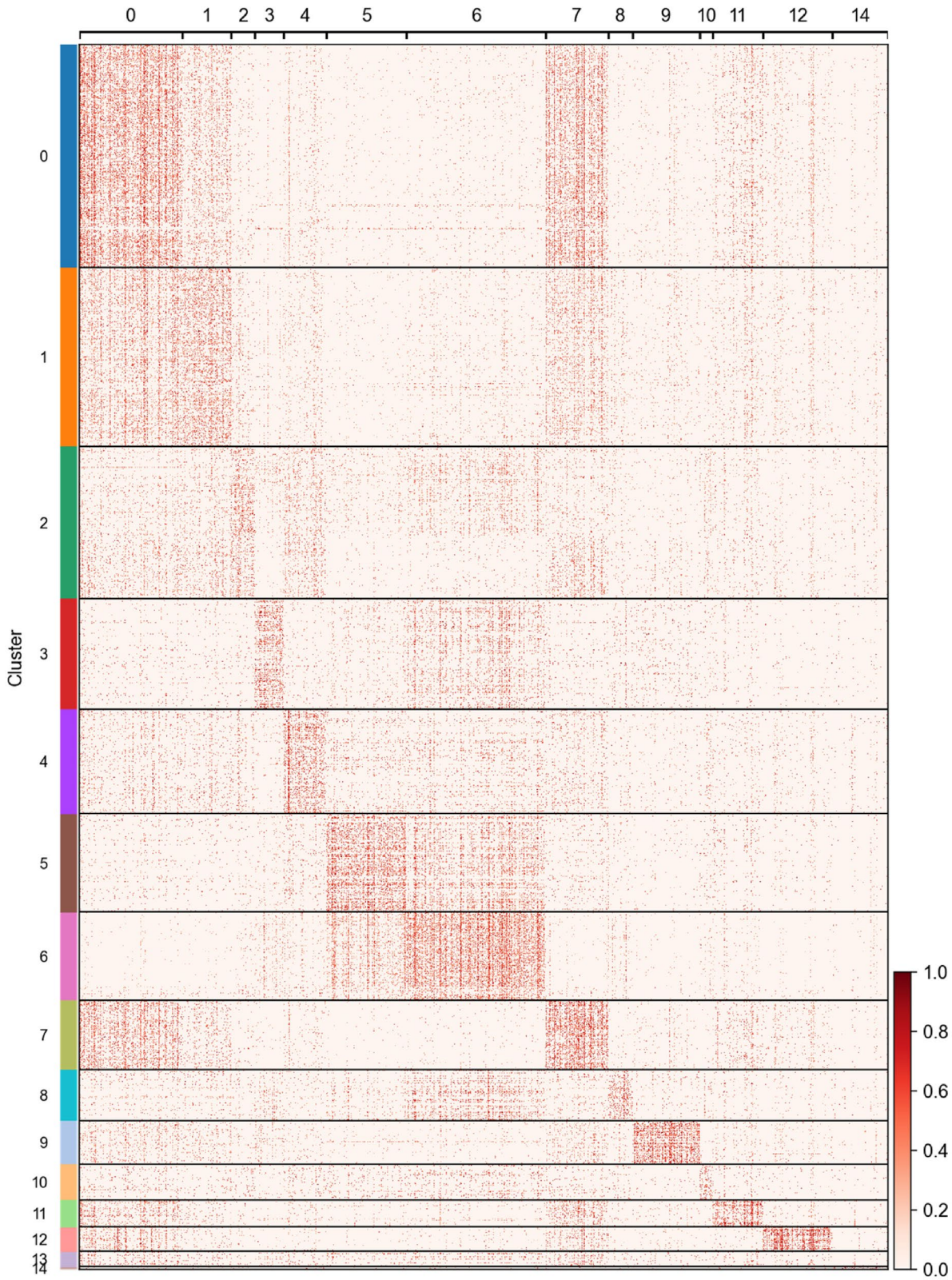
Extended data is available for this paper at https://doi.org/10.1038/s41477-023-01387-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-023-01387-z.

Correspondence and requests for materials should be addressed to Jixian Zhai or Zhe Yan.

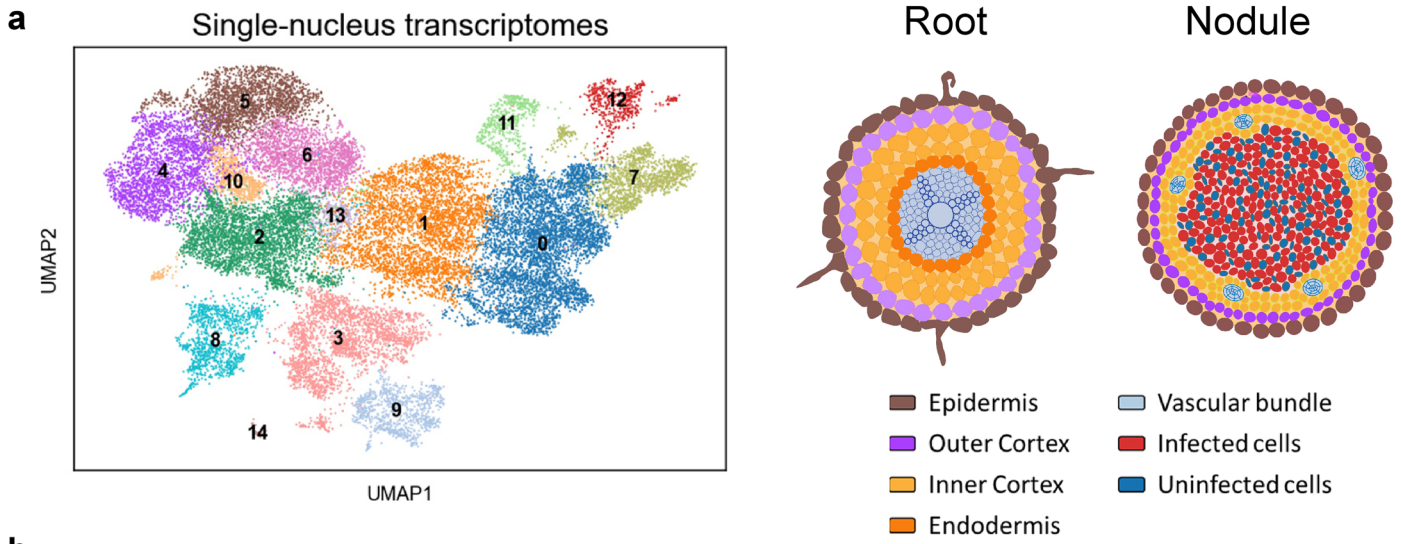
Peer review information *Nature Plants* thanks Hon-Ming Lam and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

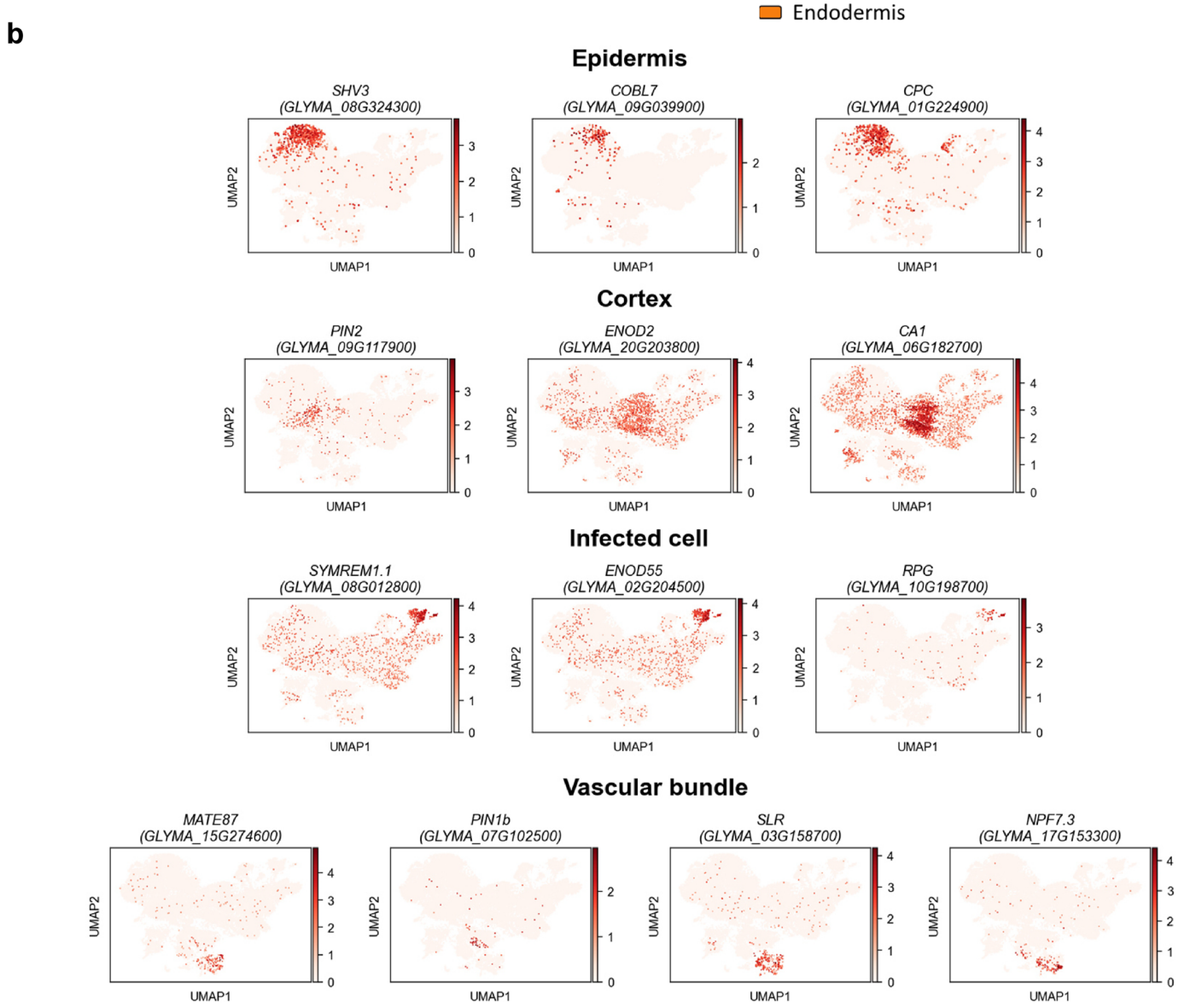
https://doi.org/10.1038/s41477-023-01387-z



Extended Data Fig. 1 | Heatmap representing the expression pattern of up-regulated genes for each cluster.

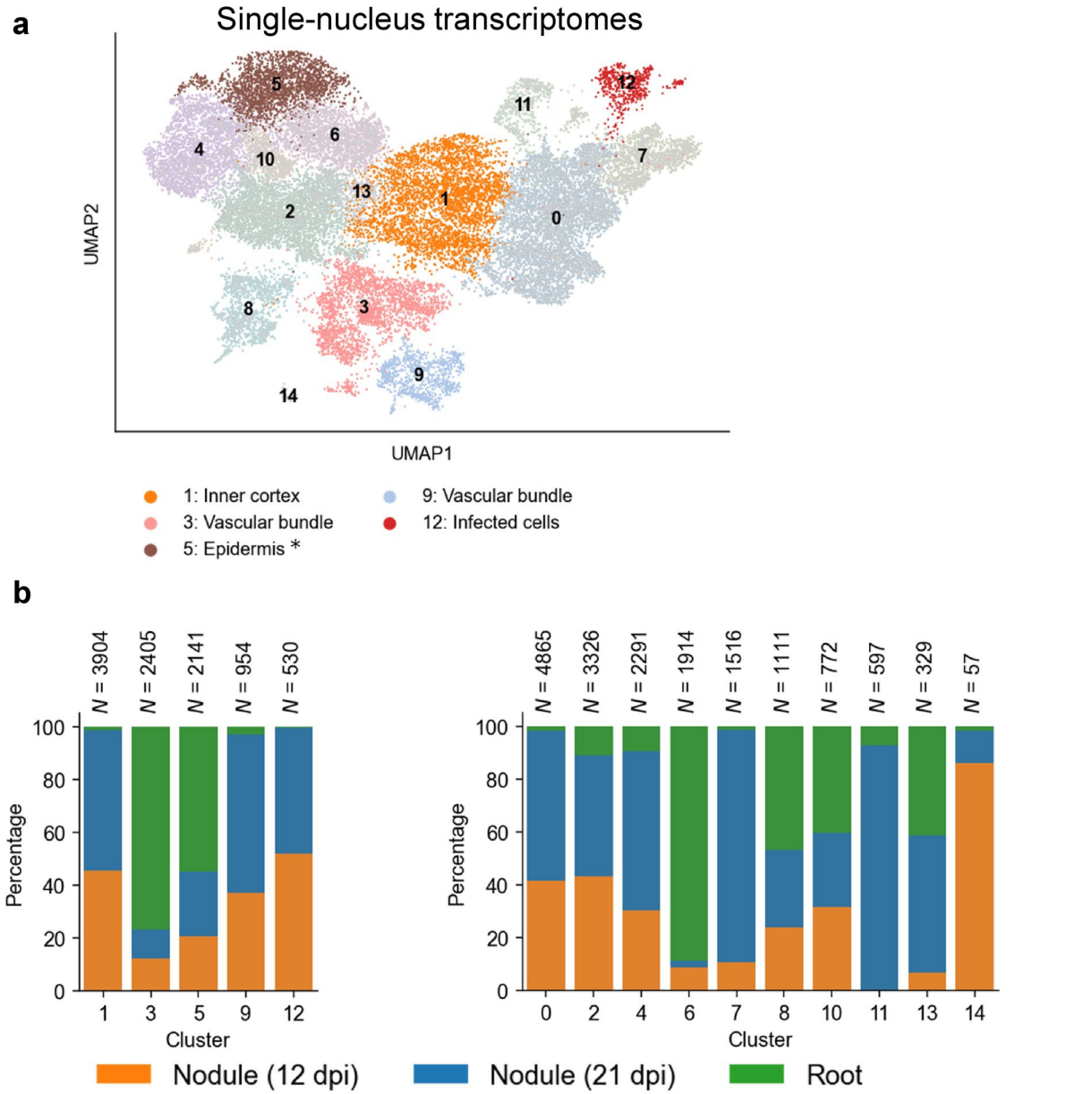
https://doi.org/10.1038/s41477-023-01387-z





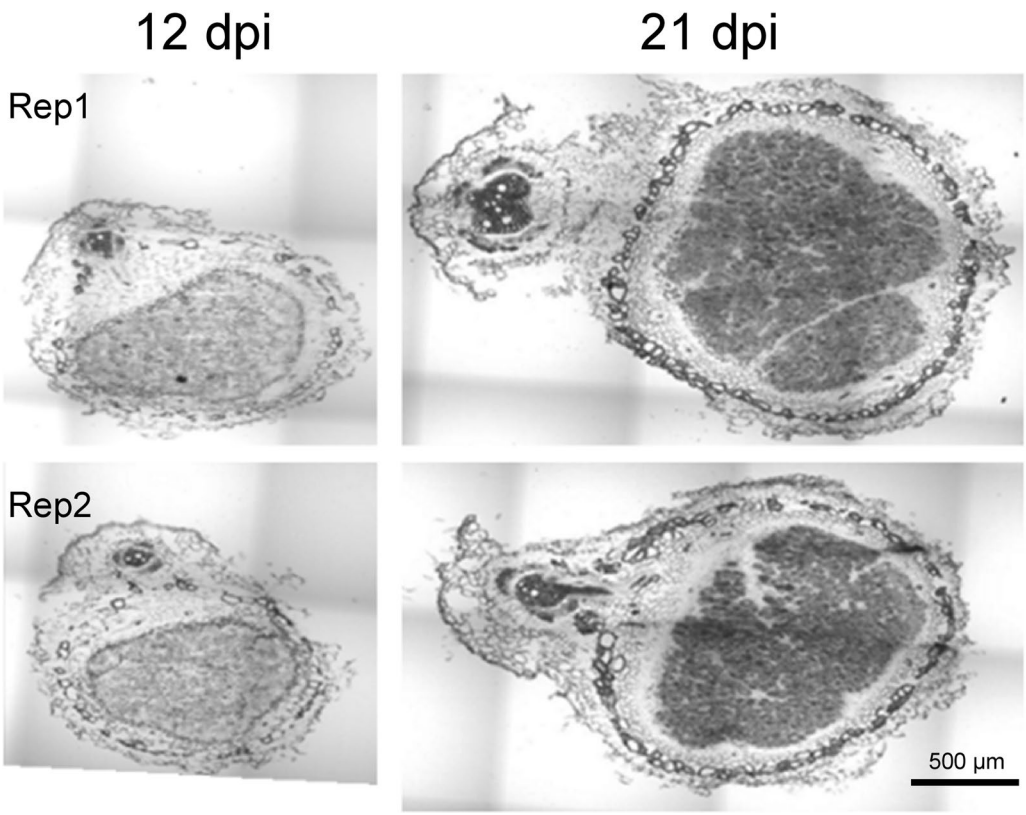
Extended Data Fig. 2 | UMAP visualizations of clustering results (a) and cell-type specific marker genes (b).

https://doi.org/10.1038/s41477-023-01387-z



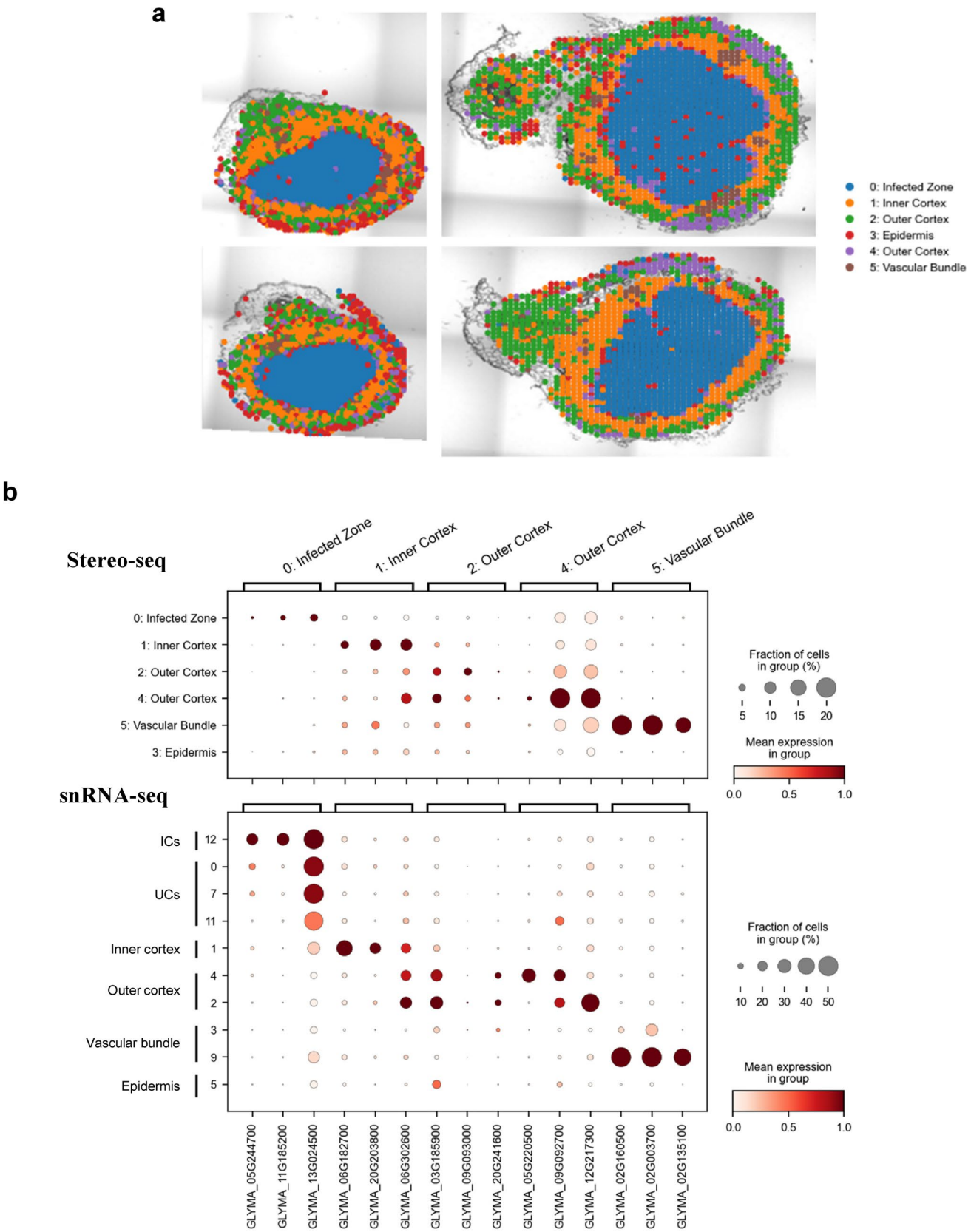
Extended Data Fig. 3 | Annotation results using public resources, including marker genes and scRNA-seq data of Arabidopsis. a. UMAP visualizations of annotation results. Unidentified cluster is masked by grey colour. "*" indicates this cluster is annotated by label transfer method. **b.** Bar chart represents the percentage of different samples in each clusters. Left, successfully identified clusters. Right, un-identified clusters.

https://doi.org/10.1038/s41477-023-01387-z



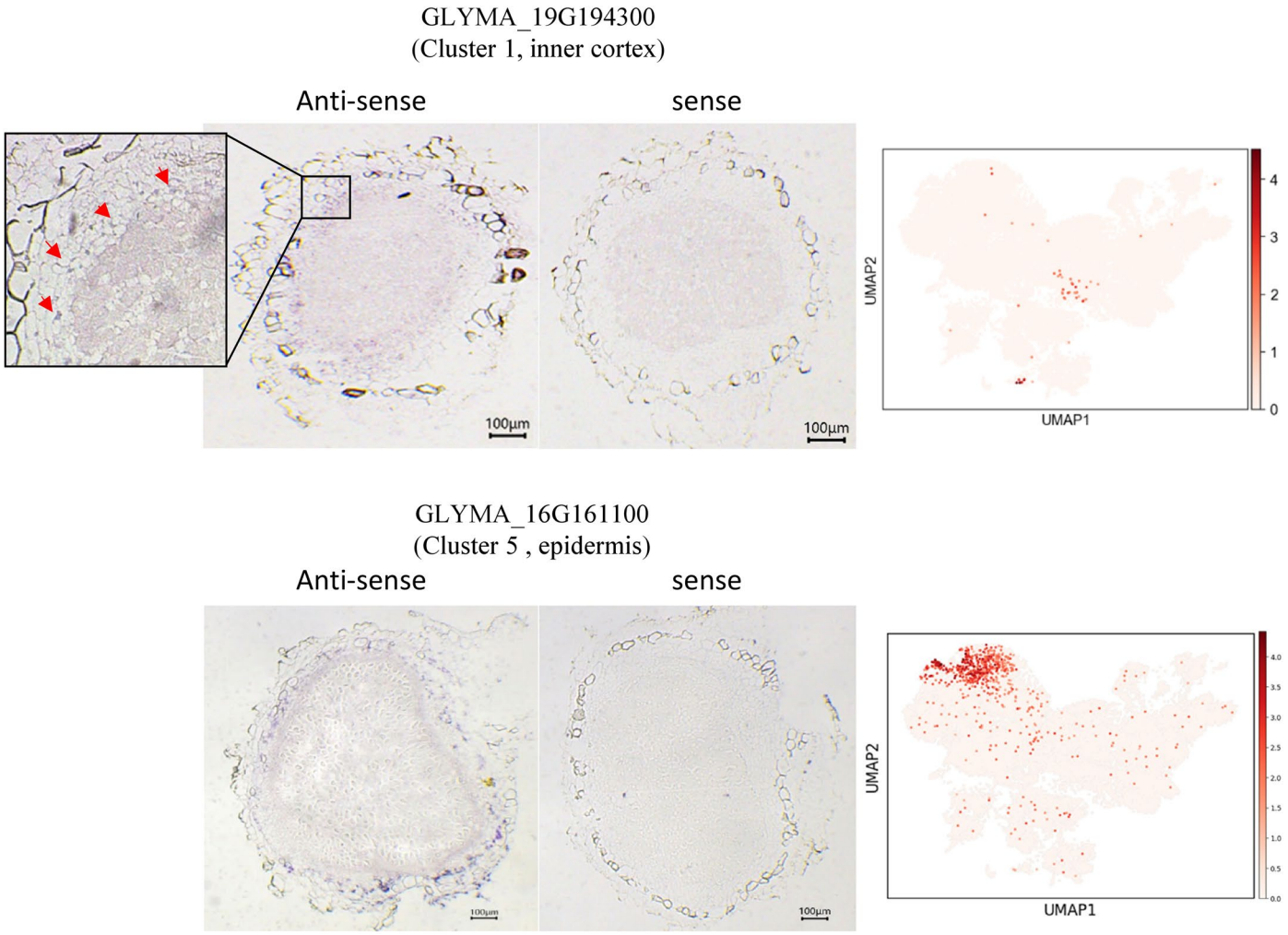
Extended Data Fig. 4 | Bright-field image of soybean nodule sections used to prepare the spatial transcriptome. Two replicates were used as the figure illustrated. Scale bars, 500 μm.

https://doi.org/10.1038/s41477-023-01387-z



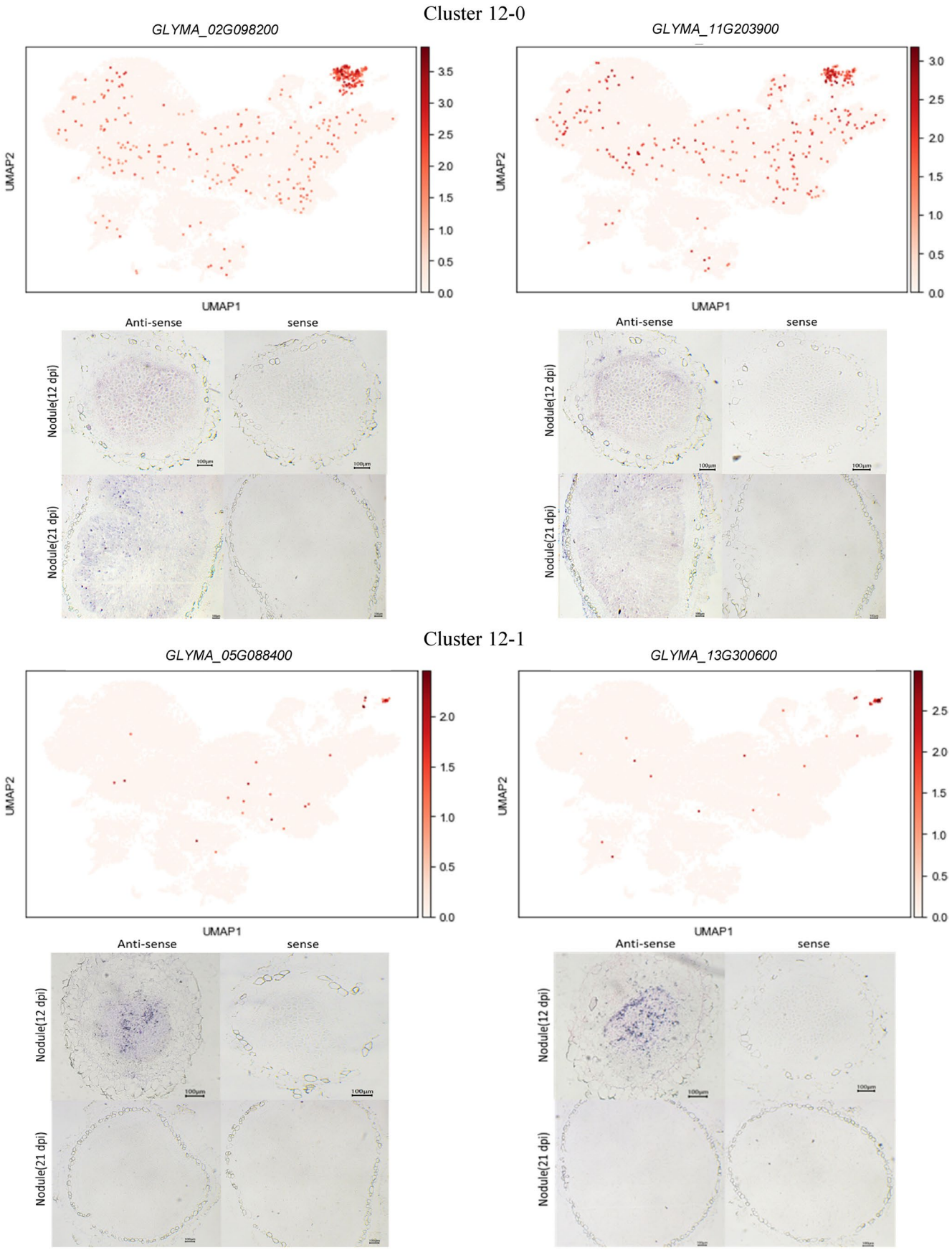
Extended Data Fig. 5 | Using cluster-based method to annotate snRNA-seq datasets. a. Clustering and annotation results of Stereo-seq datasets. **b.** Expression patterns of spatially transcriptome-identified cell-type upregulated genes in Stereo-seq (upper panel) and snRNA-seq (lower panel). We did not identify up-regulated genes in the epidermis from the spatial transcriptomes, so they were not mapped.

https://doi.org/10.1038/s41477-023-01387-z



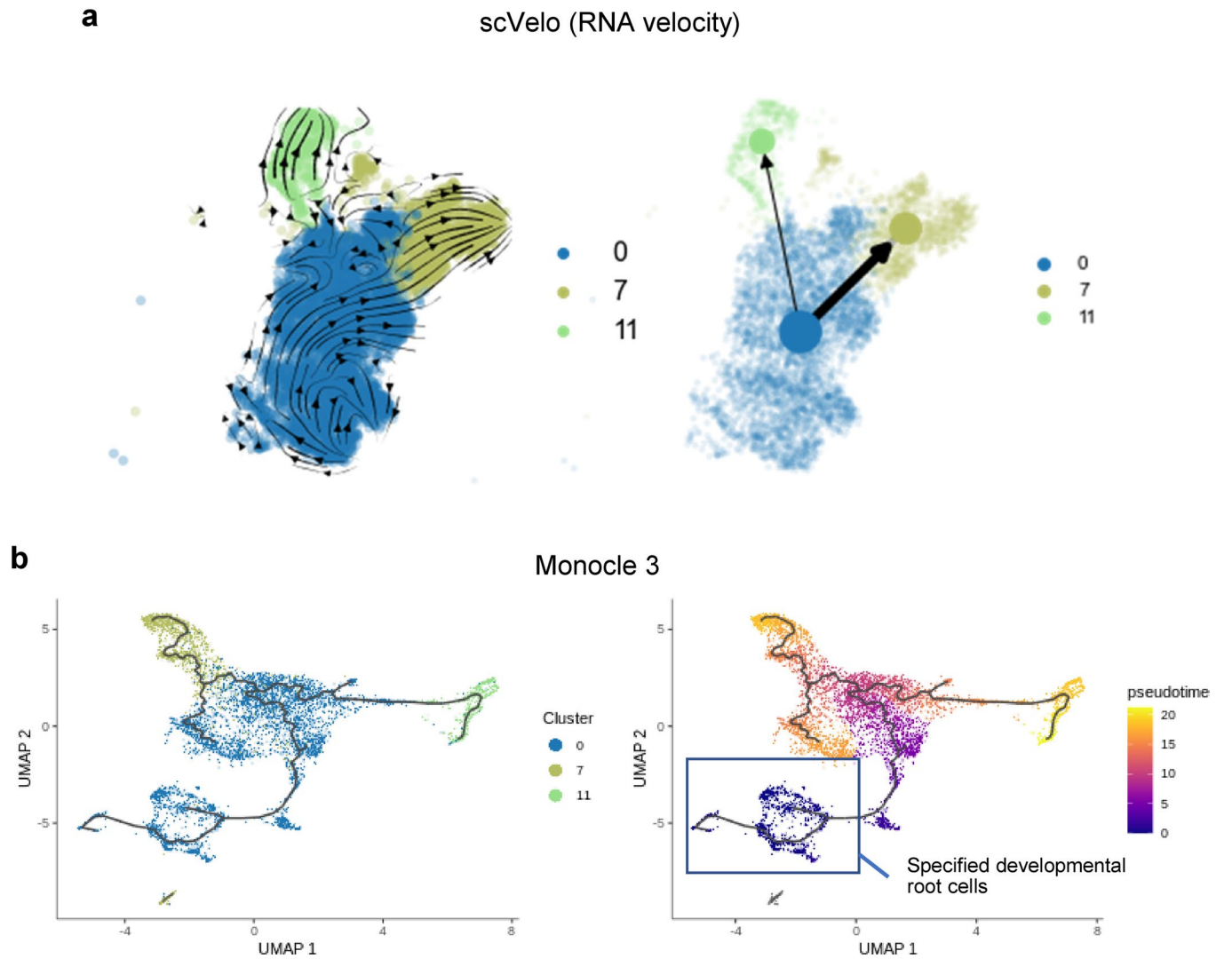
Extended Data Fig. 6 | Validation of cluster-specific marker genes by RNA *in situ* hybridization. These experiments were repeated in three independent assays and for each section, at least three nodules were analysed, and all showed the same expression pattern.

https://doi.org/10.1038/s41477-023-01387-z



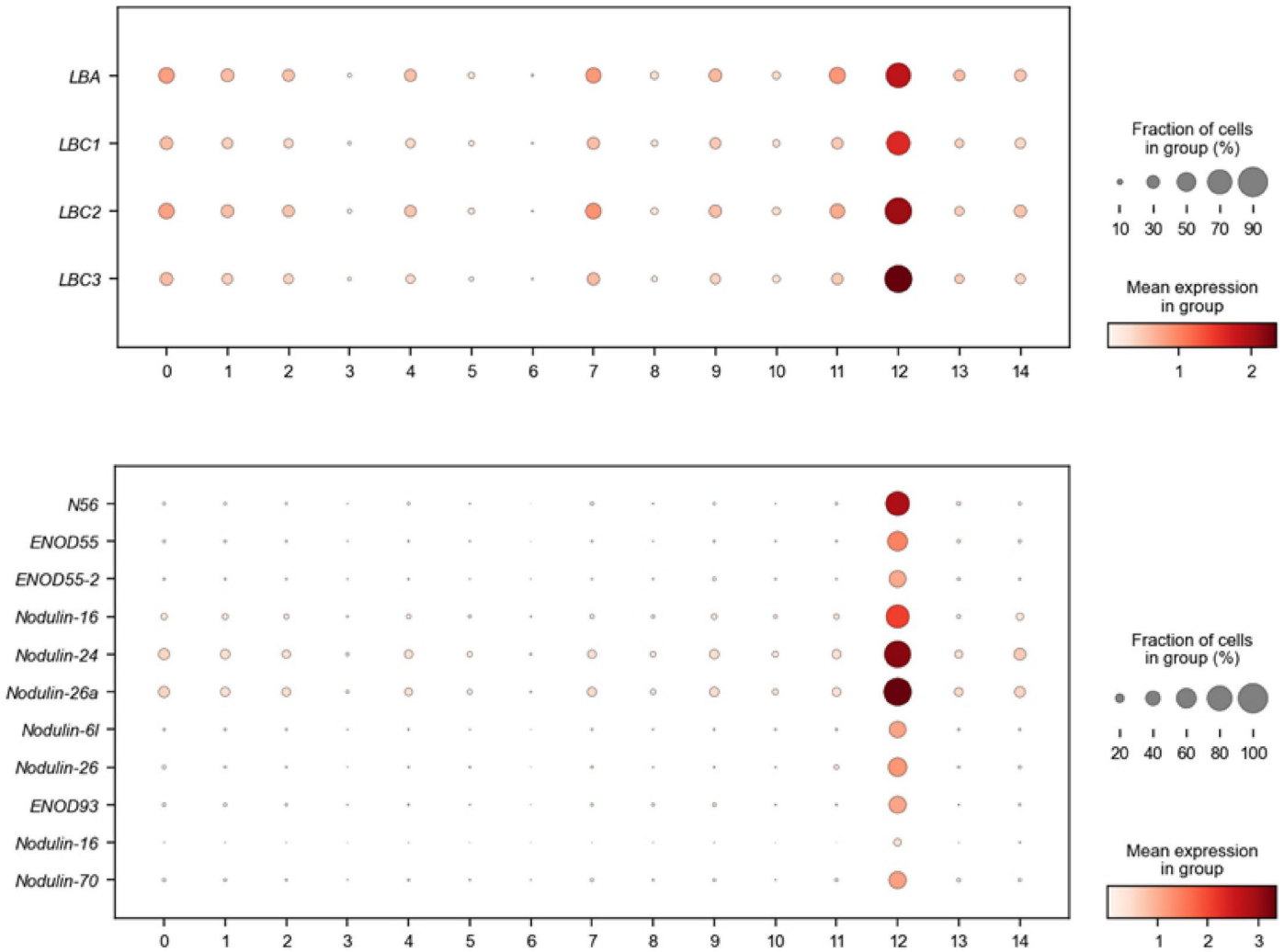
Extended Data Fig. 7 | UMAP visualizations and RNA *in situ* hybridization of expression pattern of 12-0 and 12-1 specific genes that used in Fig. 2g. Scale bars, 100 μm. These experiments were repeated in three independent assays and for each section, at least three nodules were analysed, and all showed the same expression pattern.

https://doi.org/10.1038/s41477-023-01387-z



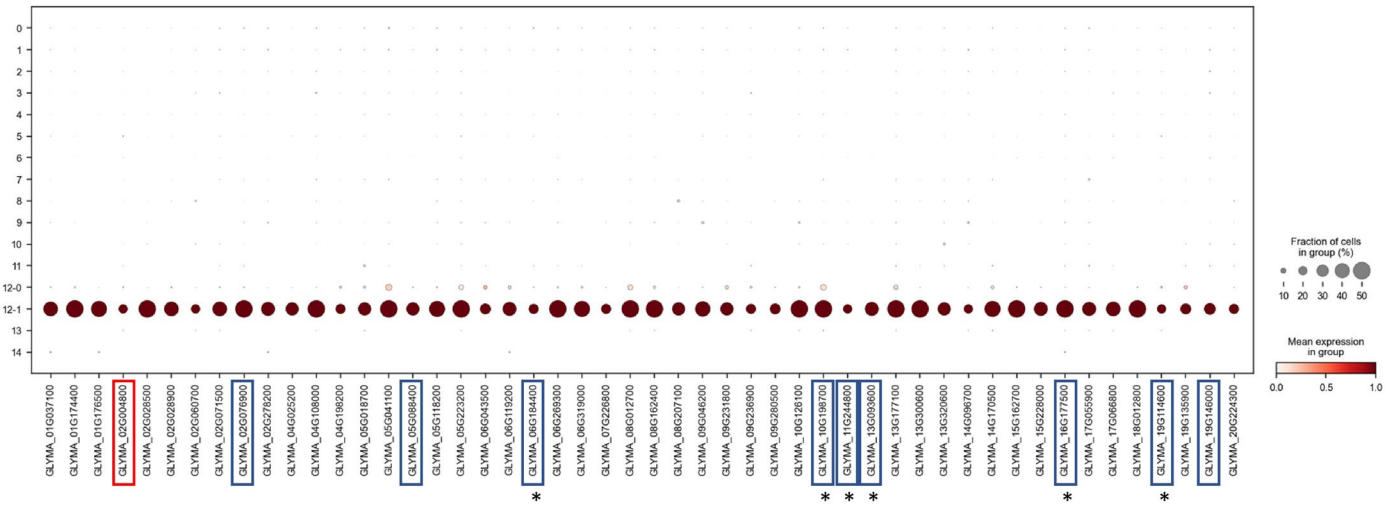
Extended Data Fig. 8 | Developmental trajectory of UCs inferred by scVelo (a) and Monocle 3 (b). a. Left panel, stream plot of RNA velocities on the UMAP embedding. Right panel, partition-based graph abstraction (PAGA) graph with velocity-directed edges. Arrow width indicates the transition probability between different clusters. **b**. Pseudo temporal ordering of nuclei after manually specified developmental root cells.

https://doi.org/10.1038/s41477-023-01387-z



Extended Data Fig. 9 | Expression pattern of four leghemoglobin genes and eleven nodulin genes.

https://doi.org/10.1038/s41477-023-01387-z



Extended Data Fig. 10 | Expression pattern of 12-1 specific genes. Blue box, known SNF genes or homologs of known SNF genes in soybean. Asterisk indicates SNF genes collected by Roy *et al*[1]. Red box, *GLYMA_02G004800*, the example we used to explore the potential function of subcluster 12-1.

nature portfolio

Corresponding author(s): Zhe Yan

Last updated by author(s): Feb 8, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	NA
Data analysis	snRNA-seq data preprocessing: CellRanger (6.0.0), STAR (2.7.8a) stereo-seq data preprocessing: SAW (2.1.0) Analysis: Scanpy(1.8.0), seurat(4.1.1) , scDblFinder(1.10.0), scvi-tools(0.16.0, including scVI, destVI, scANVI), cellrank(1.5.1), cellex(1.2.2), AUCell(1.18), clusterProfiler(4.4), Harmonypy(0.0.6), diffxpy(0.7.4), scVelo(0.2.4), monocle3(1.0.0), muon(0.1.2), scanorama(1.7.1) Visualization: Matplotlib(3.5), plotnine(0.8), seaborn(0.12), patchworklib(0.4.0) Others: Python(3.8), R(4.1.0), rpy2(3.5.1) The source code to reproduce this project can be accessed at https://github.com/ZhaiLab-SUSTech/soybean_sn_st

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data generated in this study were deposited in China National Center for Bioinformation with accession PRJCA009893. Raw sequencing data were deposited in GSA with accession CRA007122 (reviewer link: https://ngdc.cncb.ac.cn/gsa/s/Ki8oBhiE) and processed data were doposited in OMIX with accession OMIX002290 (reviewer link: https://ngdc.cncb.ac.cn/omix/preview/2LlFuSVH)

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample-size was determined based on previous studies and experiments. Sample size of all experiments was sufficient to result in statistical significance and reproducibility.
Data exclusions	Low quality cells were filtered, refer to the supplementary material for details.
Replication	For the spatial transcriptomics, two biological replicates existed for each of the different developmental stages of the nodule. For each GUS construct, three individual transgenic lines were generated and at least three nodules from each line were analyzed, and all had the same expression pattern. For RNA in situ hybridization, at least three nodules were analyzed, and all showed the same expression pattern.
Randomization	Not applicable, as samples were processed identically through standard and in some cases automated procedures (10X genomics, DNA/RNA isolation) that should not have bias outcomes.
Blinding	Not applicable, as the analysis is in genome-wide level and has no obvious bias for specific genes or cells.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging