Single-cell resolution transcriptome atlases of soybean root organs reveal new regulatory programs controlling the nodulation process

Sergio Alan Cervantes-Pérez  
University of Nebraska-Lincoln

Prince Zogli  
GALY.Co

Sandra Thibivilliers  
University of Nebraska-Lincoln

Sutton Tennant  
University of Nebraska-Lincoln

Md Sabbir Hossain  
University of Nebraska-Lincoln

Hengping Xu  
University of Nebraska-Lincoln

Ian Meyer  
University of Nebraska-Lincoln

Akash Nooka  
University of Nebraska-Lincoln

Sai Subhash Mahamkali Venkata Subramanyam  
University of Nebraska-Lincoln

Pengchong Ma  
University of Nebraska-Lincoln  https://orcid.org/0009-0003-2385-8077

Qiuming Yao  
University of Nebraska Lincoln

Michael Naldrett  
University of Nebraska-Lincoln

Benjamin Smith  
University of California, Berkeley

Samik Bhattacharya  
Resolve Biosciences

Jasper Kläver  
Resolve Biosciences

Marc Libault (✉ marc.libault@unl.edu)  
University of Nebraska-Lincoln  https://orcid.org/0000-0001-7419-9129

Article

Keywords: Glycine max, root, nodule, single-cell RNA-seq, GmFWL/CNR family, membrane microdomains
Abstract

The soybean root system is complex. In addition to being composed of various cell types, the soybean root system is also composed of the nodule, an organ in which mutualistic symbiosis with the nitrogen-fixing bacterium *Bradyrhizobium diazoefficiens* occurs. Notably, the soybean root nodule is characterized by a central infection zone where the atmospheric nitrogen is fixed and assimilated by the symbiont, resulting from the close cooperation between the plant cell and the bacteria. To date, the cellular complexity of the soybean root and nodule has not been characterized. Applying single nucleus RNA-seq technology, we characterized the transcriptomic signature of the soybean root and nodule cell types and revealed the existence of different sub-populations of *B. diazoefficiens*-infected cells in the soybean nodule including those actively involved or not in nitrogen fixation. The mining of the single cell-resolution atlas also helped reevaluate the role of known and identified new genes controlling the nodulation process. For instance, we functionally characterized the role of a new microdomain-associated GmFWL3 protein in the soybean nodule. Our study reveals the unique cellular complexity of the soybean nodule and helps redefine the concept of cell types when focusing on the infection zone of the nodule.

Introduction

Legumes are engaged in a mutualistic symbiotic interaction with Rhizobia, nitrogen-fixing soil bacteria (e.g., *Bradyrhizobium diazoefficiens* for soybean)\(^1\). This biological process named nodulation has economic and ecologic impacts on agriculture because it helps mitigate the applications of nitrogen fertilizers, reduces environmental pollution, and supports sustainable agricultural practices. For most legume species (e.g., *Medicago truncatula*, *Lotus japonicus*, *Glycine max*, *Phaseolus vulgaris*, *Pisum sativum*), the nodulation process is initiated by the infection of the root hair cells by Rhizobia. Concomitant to this infection, the root outer cortical cells are engaged in de novo cell divisions leading to the formation of the nodule primordia. Ultimately, a new root organ, the nodule, emerged where differentiated bacteria named bacteroids fix and assimilate the atmospheric nitrogen for the plant\(^2,3\). However, based on its determination, nodule organogenesis and the cellular organization of the nodule differ\(^4,5\). Mature indeterminate nodules (e.g., *M. truncatula* and *P. sativum*) can be divided into four biologically and microscopically distinct zones. In zone I, at the tip of the nodule, a permanent nodule meristem subsists. In zone II, rhizobia infect the plant cells. Zone III is the nitrogen fixation zone where the bacteroids fix and assimilate for the plant the atmospheric nitrogen. In zone IV the nodule cells senesce. Unlike indeterminate nodules, mature determinate nodules (e.g., *G. max*, *L. japonicus*, and *P. vulgaris*) are not organized in visually distinct zones associated with different stages of interaction between the plant cells and rhizobia. As a result, all the plant cells colonized by rhizobia are located in the center of the nodule.

During the past 20 years, many -omics studies conducted on different legumes led to the identification of many genes controlling the nodulation process including those involved in the symbiosis between the infected cells of the nodule and the bacteroids\(^6-13\). Among them, the emergence of high-throughput sequencing technologies led to the establishment of several legume transcriptomic atlases that notably revealed differences in the transcriptomic profile of the root and the nodule and identified hundreds of nodule-specific genes\(^14-17\). In *M. truncatula*, the transcriptome of each of the 4 zones of the nodule was established using laser microdissection\(^18\). This technology has recently been superseded by emerging single-cell and single-nucleus transcriptomic technologies (sc and sNucRNA-seq)\(^19-26\). For instance, single-cell RNA-seq technology was recently used to capture the transcriptomic profile of the indeterminate Medicago nodule cells and confirmed zone-specific transcriptomics programs of the infected cells of the nodule\(^27\). In determinate nodules such as the soybean nodule, our understanding of the cellular composition and associated transcriptomics programs of the infected cells remain vague. Here, we applied single-nucleus RNA-seq technology to establish the transcriptomic signature of the soybean root and nodule cell types. Our analysis revealed, for the first time, the cellular complexity of the infection zone of the soybean nodule and the differential regulation of genes between different sub-populations of
infected cells. Upon mining the soybean single-cell root and nodule atlases, we functionally characterized the role of GmFWL3, a new microdomain-associated protein controlling the infection rate of soybean cells by B. diazoefficiens. Our study reveals the unique cellular complexity of the soybean nodule allowing a deeper understanding of the molecular processes governing nodulation.

Single-cell resolution transcriptome atlas of the soybean root

To establish the transcriptomic profile of each cell type composing the soybean root, we applied sNucRNA-seq technology on three independent soybean root replicates (Fig.S1 to S4). The single-nucleus root transcriptome atlas is composed of 14,639 high-quality nuclei, captures means of 1,949 unique molecular identifiers (UMIs) and 1,363 expressed genes per nucleus, and covers the expression of 75.8% of the coding soybean transcriptome (42,391 out of the 55,897 predicted soybean protein-coding genes; Table.S1). This atlas nicely overlaps but also better covers the transcriptome of the soybean root when compared to previously reported bulked transcriptomes that identified 39,709 and 36,354 expressed genes in the root and root tip, respectively\textsuperscript{15} (Fig.S5). Upon applying the Uniform Manifold Approximation and Projection (UMAP) dimensional reduction technique, the root nuclei were distributed in 16 distinct cell clusters according to their transcriptomic profiles (Fig. 1a). Except the root hair cell cluster 3 which is characterized by higher numbers of UMIs and expressed genes, the remaining 15 clusters have similar transcriptional activities (Fig. 1b). Looking at the activity of the 55,897 protein-coding genes across the 16 root clusters, we identified 14,088 ubiquitously expressed genes (i.e., genes found expressed in each cluster). Among them, 2,753 are constitutively expressed across the 16 clusters (i.e., less than a 4-fold change in activity between the clusters where the gene is the highest and least expressed) (Table.S1). We also identified 424 cell-type marker genes (fold-change > 20 between the two clusters where the gene is the most expressed; a minimum expression of 0.1/10,000 UMIs, expressed in at least 20% of the nuclei composing the cluster where the gene was found specifically expressed) (Table.S1).

To functionally annotate these clusters, we first applied Molecular Cartography\textsuperscript{™} (MC) technology developed by Resolve Biosciences on a cross-section of the soybean root to analyze the transcriptional activity of 52 soybean cluster-specific genes in the context of the cellular organization of the soybean root. The transcriptional pattern of these genes led to the annotation of the epidermal, cortical, endodermal, pericycle, cambial, xylem, and phloem clusters (Fig. 1c and 1e, Fig.S6). To accurately annotate the root hair cell cluster, which is a difficult cell type to assess from our MC experiments, due to its unique morphology and peripheral localization in the root, we also looked at the activity of past and newly functionally validated root hair-specific genes (Fig. 1d, Fig.S7). To annotate the soybean “root cap” and “stem cell niche” clusters, cell types not represented on the root cross-section by MC technology, we analyzed the activity of soybean genes orthologous to Medicago truncatula and Arabidopsis thaliana marker genes (Fig. 1d, Fig.S7, Table.S2). Using the same strategy, we confirmed the identity of clusters 10 and 11 as the “endodermal” clusters and refined the identity of cluster 13 as the “pericycle” cluster. Taken together, we functionally annotated the 16 soybean root clusters. Looking at the distribution of the 424 single-cell type marker genes across the 16 annotated soybean root clusters, we found that they were restricted to 9 clusters (i.e., the root hair cluster #3, root cap #4, dividing cells # 6, cortex #7, endodermis #10 and 11, xylem #14, phloem fiber #15, and phloem #16; Table.S1), likely a reflection of the biological specialization of these cell types.

Single-cell resolution transcriptome atlas of the soybean nodule

The single-cell transcriptome atlas of 28 days post-infection soybean nodules was generated from two independent replicates (Fig.S8 to S10) and is composed of 7,830 nuclei with the means of 1,058 UMIs and 647 expressed genes per cell, with a total of 37,119 expressed genes (covering ~ 66.4% of the soybean protein-coding genes) (Table.S1). Similarly to the root, the sNucRNA-seq datasets from the soybean nodules overlap well with a previously published soybean nodule transcriptome\textsuperscript{15} (Fig.S11). The UMAP of the nodule nuclei revealed 11 different cell clusters named A to K.
(Fig. 2a). Among them, clusters F and G are characterized by significantly higher numbers of UMIs and expressed genes per nucleus compared to the other clusters, suggesting the higher transcriptomic activities of the cells composing these two clusters (Fig. 2b). Using the same parameters as described above, we found 950 ubiquitously expressed genes including only 16 constitutively expressed genes across the 11 clusters and 250 cluster-marker genes. Among the 16 constitutively expressed genes of the nodule, 8 were also identified as constitutively expressed in the soybean root system (i.e., Glyma.17G073300, Glyma.19G248000, Glyma.09G156600, Glyma.07G091800, Glyma.18G222400, Glyma.12G056500, Glyma.13G212500, and Glyma.08G262900). The very limited number of constitutively expressed genes across the nodule clusters suggests the high transcriptional diversity existing between nodule cell types (Table. S1).

To annotate these 11 clusters, we applied MC technology on a cross-section of a mature nodule. Upon analysis of the spatial activity of 28 soybean cell-type-specific marker genes (Fig. 2c; Fig. S12), we annotated clusters A and B as the “inner/outer cortical cell” cluster (Fig. 2e; blue), cluster C as the “vascular endodermis” cluster (Fig. 2f, light pink), cluster D as the “sclereid layer” (Fig. 2e, pink), and cluster E as the “vascular bundle” cluster (Fig. 2f, orange). Upon applying MC technology using probes designed against nine *B. diazoefficiens* genes, we identified both the infected cells (Fig. 2g; Fig. S13 and uninfected cells of the nodule (Fig. 2g, white arrow). Our MC experiments also revealed the spatial activity of two soybean genes expressed in the uninfected cells (Glyma.06G235500, Glyma.06G002000; Fig. 2h, green arrows) and four soybean genes expressed in the “infected cells” of the nodule (Glyma.17G195900, Glyma.01G164600, Glyma.15G210100, and Glyma.05G216000; Fig. 2h; red; Fig. S14). The transcriptional activity of Glyma.06G235500, Glyma.06G002000 in clusters J and K in conjunction with the activity of Glyma.17G195900, Glyma.01G164600, Glyma.15G210100 in clusters F and G, and Glyma.05G216000 in cluster H (Fig. 2c) led to the annotation of clusters J and K as the “uninfected cells” and clusters F, G, and H as the “infected cells” of the nodule (Fig. S15).

To annotate cluster I, we analyzed the biological function of the genes specifically expressed in this cluster. They include genes encoding a metallothionein28 (Glyma.18G180800), two CAP proteins, proteins associated with soybean nodulation-related traits29 (i.e., cysteine-rich secretory proteins antigen 5 and Pathogenesis-related 1 proteins; Glyma.15G062300 and Glyma.13G252600), a γ-thionin antimicrobial protein30 (Glyma.16G100400), and two TCTPs, proteins hampering the program cell death31 (Glyma.09G044200 and Glyma.15G148900) were specifically or preferentially expressed in cluster I, and to some extent, in cluster H (Fig. 2d, Fig. S16). These transcriptional patterns suggest that the cells composing cluster I are under high stress; potentially senescing. Taken together, we functionally annotated 10 out of 11 cell clusters of the nodule (Fig. 2a) including three cell clusters associated with the infected cells of the soybean nodule (i.e., clusters F, G, and H), and provide information regarding the physiological status of cluster I. The 250 nodule single-cell type marker genes were distributed across the inner/outer cortex cluster A, vascular endodermis cluster C, sclereid layer cluster D, vascular bundle cluster E, the “infection zone” cluster G, and the “stressed cells” cluster I (Table. S1).

When comparing the list of root and nodule cell-type marker genes, 42 root and nodule marker genes were repetitively expressed in the root xylem cluster #14 and the nodule sclereid layer cluster D (Table. S1). Looking at the predicted function of these 42 genes, we found an over-representation of genes involved in the biosynthesis and organization of the plant cell wall, a specific biochemical characteristic shared between the root xylem and the nodule sclereid cells. Overall, the absence of shared marker genes between the root and nodule cell types reflects the unique specialization of each cell type composing these two root organs. Conducting a similar comparative transcriptomic analysis of the genes identified as root and nodule cell-type-specific using Molecular Cartography™, we confirmed that root xylem-specific genes are also expressed in the sclereid layer of the nodule. Besides, we found that root phloem-, pericycle-, endodermal-, and epidermal-specific genes are specifically expressed in the infection zone, vascular bundle, vascular endodermis, and in the inner/outer cortical cells of the nodule, respectively. This result suggests that the plant reallocates the use of root cell-type-specific genes to fulfill unique function on the different cell-types composing the nodule.
The infection zone of the soybean nodule is composed of different populations of rhizobia-infected cells

Our single-cell resolution transcriptomic analysis helps reveal the differences and similarities existing in the transcriptional profile of the different cell types composing the root and nodule organs. Upon conducting a multidimensional scaling plot (MDS) analysis, we found that most of the cell clusters of the nodule are transcriptionally different from the root cell clusters (Fig. 3a), confirming the unique functions of the different cell types composing the nodule organ. Among them, the infected clusters F and G as well as the cells of the sclereid layer (cluster D) have unique transcriptional profiles. The uninfected cells of the nodule (clusters J and K), another subpopulation of infected cells (cluster H), and the cells composing the nodule “stress” cluster I are also characterized by specific transcriptional profiles. This MDS analysis also supports the hypothesis that the infected cells composing clusters F and G are transcriptionally different from the infected cells composing cluster H. To further reveal these differences, we conducted a comparative transcriptomic analysis between clusters F, G, and H (Fig. 3b). We only identified 253 differentially expressed genes (DEGs) between clusters F and G (Fig. 3b; Table.S3). Among them, we identified several up-regulated genes in cluster G controlling the catabolism and metabolism of cyclic beta-glucans, bacterial carbohydrates acting as suppressors of the host defense, regulators of osmotic potential, and in the interaction with host membranes\(^32\,33\) (Table.S4). When comparing the activity of soybean genes between clusters F&G vs. H, we identified 3,731 up-regulated genes in clusters F&G (Table.S3) associated with “RNA processing”, “vesicle-mediated transport”, and “chromatin modification” functions (Table.S4) including genes orthologous to \emph{MtCCS52A}, a gene controlling the endoreduplication rate of infected cells during rhizobia differentiation (Fig. 3c; Table.S5). In cluster H, we identified 331 up-regulated genes (Table.S3) associated with the “nodulation” process and the “biosynthesis of organonitrogen compounds” (Table.S4). This result suggests that cluster H cells are involved in an active nitrogen fixation process (Fig. 3b). This observation was further confirmed based on the transcriptional activity of four soybean leghemoglobin genes in cluster H (Fig. 3d; Table.S5). Therefore, our data suggest that at least two biologically distinct populations of infected cells compose the infected zone of the soybean mature nodule: infected but non-nitrogen-fixing cells composing clusters F and G, and the infected nitrogen-fixing cells of cluster H. To confirm the differential biological functions of the cells composing the clusters F, G, and H, we analyzed the activity of soybean genes orthologous to legume genes controlling various aspects of legume-rhizobia symbiosis. Among them, we found that clusters F and G are characterized by the preferential expression of genes controlling host-range restriction and the plant defense system as previously defined\(^1\) (Fig. 3e and f; Table S5). Interestingly, the genes controlling bacterial maturation were most expressed in cluster G and, to a lesser extent, in cluster H (Fig. 3g; Table.S5). These results support previously published spatial metabolomics observations that revealed the biochemical heterogeneity in the infection zone of the nodule\(^1\,2\,3\)4.

We hypothesize that the two sub-populations composing the soybean nodule are physiologically and biologically similar to those composing the zones II and III of the \emph{M. truncatula} nodule, respectively. To verify this hypothesis, we reanalyzed published Medicago nodule single-cell RNA-seq datasets (see method section for details) and annotated clusters 4 and clusters 6&7 as the infection zone (i.e., zone II) and the nitrogen-fixation zones (zone III) of the Medicago nodule based on the expression of selected marker genes, respectively\(^27\) (Fig. 3h; Fig.S17; Table.S6). Cluster 5 seems to be composed of cells in a transitional stage between cluster 4 and clusters 6&7 (Fig.S17). The transcriptional activity of Medicago \emph{CCS52A}, leghemoglobin genes, as well as the genes characterized as regulators of the host-range, defense, and bacterial maturation responses\(^1\), indicate that \emph{MtCCS52A} is most broadly and highly expressed in cluster 4 and, at a lower level, in cluster 5 (Fig. 3i; Table.S5). Oppositely, the leghemoglobin genes are most expressed in nitrogen-fixing clusters 6 and 7 (Fig. 3j; Table.S5). We also observed the preferential activity of the host range restriction and nodule defense genes in cluster 5, and the expression of the genes controlling bacterial maturation are preferentially expressed in cluster 6 and, to a lower level, in clusters 5 and 7 (Fig. 3k to m; Table.S5). These results suggest that the Medicago clusters 4 and 5 are biologically and physiologically similar to soybean clusters F and G while the Medicago clusters 6 and 7 share
similarities with the soybean cluster H. The soybean clusters F&G and cluster H sub-populations share similar biological functions compared to those previously described in the infection zone II (clusters 4–5) and the nitrogen fixation zone III of the Medicago nodule (clusters 6 and 7), respectively. Our single-cell resolution transcriptomic analysis redefined our current knowledge of the functional diversity in the infected cells in the soybean nodule.

Our Molecular Cartography™ datasets highlighted three preferentially expressed genes in the infected cells of the nodule (Glyma.17G195900, Glyma.15G210100, and Glyma.05G216000; Fig.S14). Our nodule sNucRNA-seq datasets led to a slightly different conclusion because they reveal the preferential abundance of Glyma.17G195900 and Glyma.15G210100 transcripts in the nuclei composing the cluster F&G the preferential accumulation of Glyma.05G216000 transcripts in the nuclei composing the cluster H, respectively (Fig. 2c). Taking advantage of the high-resolution of Molecular Cartography™ technology, we identified two different populations of B. diazoefficiens-infected soybean cells according to the density of the Glyma.17G195900 transcripts in the nucleus of the cells (Fig. 2h, Fig. 4c and f). These same cells showed the same differential nuclear vs. cytosolic abundance of Glyma.15G210100 transcripts (Fig. 4d and g) but did not display this trend when analyzing the distribution of the Glyma.05G216000 transcripts (Fig. 4e and h). Based on this observation, we hypothesize that the cells composing the clusters F&G are characterized by the nuclear retention of Glyma.17G195900 and Glyma.15G210100 transcripts leading to the unique transcriptomic signature in these nuclei as highlighted by sNucRNA-seq technology. Interestingly, these cells are not retaining Glyma.05G216000 transcripts in their nucleus suggesting that the infected cells of the soybean nodule have the capability to regulate the transport of selected transcripts through the nuclear membrane, likely a post-transcriptional mechanism to fine tune protein translation.

GmFWL3, similarly to GmFWL1, controls the infection of the soybean nodule cells by B. diazoefficiens

One member of the soybean plasma membrane microdomain-associated FWL family, GmFWL1, has been reported as a major regulator of the infection of the nodule cells by B. diazoefficiens and chromatin accessibility15 and interacts with microdomain-associated proteins35. Our single-cell RNA-seq approach allowed us to revisit the expression pattern of GmFWL1 in mature soybean nodules; showing its preferential expression in the rhizobia-infected clusters F, G, and H but not in the root clusters (Fig. 5a and b). Considering the over-representation of GO terms related to “RNA processing” and “chromatin condensation” in the pool of cluster F&G DEGs (Fig. 3c), our results further support the role of GmFWL1 as a critical regulator of the infection of the nodule cells by rhizobia potentially through changes in chromatin condensation15.

GmFWL3 is the only other member of the FWL family that shares a pattern of expression similar to GmFWL1 [i.e., GmFWL3 is not expressed in the root (Fig. 5a) but is preferentially expressed in the infected cells of the nodules (Fig. 5b and c; clusters F, G, and H)]. We hypothesize that GmFWL3 encodes for another microdomain-associated protein that regulates the infection of the soybean cells by rhizobia. This hypothesis is supported by previous studies revealing the localization of GmFWL1 and GmFWL3 proteins in the symbiosome membrane10,35. To verify this hypothesis, we deleted the conserved PLAC8 domain of GmFWL3 (Fig.S18) by expressing two guide RNAs (i.e., GmFWL3T1 and T2) and using CRISPR-Cas-mediated genome-editing technology. Upon editing, the number of nodules on the GFP-positive CAS9/GmFWL3T1-T2 transgenic roots significantly decreased compared to GFP-positive Cas9/pAH595 control transgenic roots (Fig. 5d and e; Student T-test, p-value = 0.009). At the cellular level, upon staining of bacteroids with the green fluorescent SYTO13 dye, we observed a significant decrease in the intensity of the fluorescent signal in the CAS9/GmFWL3T1-T2 transgenic roots compared to control roots (Fig. 5f and g; Student T-test, p-value = 0.009). We conclude that the microbial infection of the nodule cells was impaired upon mutagenesis of GmFWL3.

GmFWL3 and GmFWL1 belong to large membrane microdomain protein complexes
To support the role of *GmFWL3* as a microdomain-associated protein, we analyzed the GmFWL3 cellular and sub-cellular localization by expressing N- and C-terminal translational fusions between the GFP and GmFWL3 in tobacco leaf cells. Both the N- and C-terminal GFP-GmFWL3 chimeric proteins were found punctually localized in the plasma membrane of the cells (Fig. 6a to c, grey arrows; Fig.S19). Plasmolysis assays (Fig.S20a to i) and protoplast isolation experiments (Fig.S20j to l) provided further evidence of the punctuate localization of GmFWL3-GFP at the plasma membrane. Given the similar punctuate localizations for GmFWL1, GmFWL3 is likely a plasma membrane microdomain-associated protein. However, opposite to GmFWL1 which is exclusively localized in the plasma membrane, GmFWL3-GFP is also associated with the nucleus and the nuclear membrane (Fig. 6a, white arrows). We validated this nuclear membrane localization by co-localizing in tobacco epidermal leaf cells GmFWL3-GFP with CFP-AtSUN1, a nuclear envelope marker (Fig.S21) and in soybean root epidermal cells and root hairs for GmFWL3 (Fig.S22) using epiuorescent confocal microscopy. To assess the subcellular localization of GmFWL3 in the infected cells of the soybean nodule, cell types where *GmFWL3* is most expressed, we applied high-resolution Transmission Electron Microscopy (TEM) combined with immunogold labeling on transgenic soybean nodules expressing N-terminal and C-terminal myc-tagged GmFWL3 proteins under the control of the pFMV promoter. We observed GmFWL3 localized in the plasma, symbiosome, vacuolar, and vesicular membranes, and in the nucleus of the infected cells of the nodule (Fig. 6d, white arrows, Fig.S23).

To support the role of GmFWL3 as a microdomain-associated protein, we conducted co-immunoprecipitation assays on 30-day-old GFP-positive transgenic soybean nodules expressing N- and C-terminal myc-tagged GmFWL3 proteins to identify GmFWL3 protein partners (Fig.S24). Across three independent biological replicates, we identified a total of 321 proteins co-immunoprecipitated with the N-terminal or the C-terminal GmFWL3-myc tagged proteins in at least two replicates, but not against the myc tag alone (Fig. 6e, Table.S7). Among them, 5 FWL/PLAC8 proteins including GmFWL1, 6 vacuolar ATPases, 9 SPFH-domain microdomain associated proteins (i.e., Prohibitin/Flotollin/Remorin), 12 aquaporins, and 21 proteins with functions related to vesicle trafficking interacted with GmFWL3 (Fig. 6e, Table.S7). Looking at the transcriptional activity of the 321 genes encoding these proteins, 10 were preferentially expressed in the clusters F, G, and H (fold-change ≥ 4 between the expression levels in the most highly expressed F, G, and H cluster vs. the most highly expressed cluster in the remaining nodule clusters and the root clusters; Fig. 6f). These genes encode three FWL/PLAC8 proteins (including GmFWL1 and GmFWL3), three remorins/flotillins/prohibitins, well-characterized microdomain-associated proteins, one CASP-like, one sulfate transporter, one vesicle-associated membrane protein, and one receptor-like kinase 1. Among these 10 genes, nine are co-expressed in clusters F and G, and, to a lesser extent, in cluster H (Fig. 6f, red characters; Fig.S25 and 26). The co-expression of these genes in the infected cells of the nodule and the interaction of their proteins with GmFWL3 suggest the formation of a quaternary protein structure localized in the microdomain fraction of biological membrane to control the symbiosis between the soybean nodule cells and bacteroids. GmFWL1 and 3 proteins play a central and likely redundant role in forming this protein complex as revealed by the integration of the list of GmFWL3 and GmFWL1 co-immunoprecipitated proteins. When comparing the proteins co-immunoprecipitated by the two FWL proteins, we identified that 63 proteins interacting with GmFWL3 also interact with GmFWL1 (Table.S7). These included 7 SPFH-domain microdomain-associated proteins, 4 vacuolar ATPases, 15 proton-ATPase/GTPases, 4 aquaporins, 4 receptor kinases, 4 integral membrane proteins, 3 proteins with transport activity, and 3 proteins associated with vesicle trafficking. Interestingly among these shared binding partners is GmFLWT2/4 (Glyma.06G065600), the soybean ortholog to the *Medicago truncatula* MtFLWT2 and MtFLWT4. This result shows that GmFWL1 and GmFL3 proteins cooperate to support a network of microdomain-associated proteins to control the infection of soybean nodule cells by *B. diazoefficiens*.

**Discussion**

The nodule is a root organ specialized in the fixation and assimilation of atmospheric nitrogen. This biological process is the product of the symbiotic interaction between nitrogen-fixing rhizobia and the plant. Microscopic observations
revealed the differential cellular organization according to the determination of the nodule. Indeterminate nodules (e.g., *M. truncatula*) are organized in different zones that reflect differences in the developmental stages of the plant cells and their relationship with their symbiont. These zones include an active meristematic zone at the tip of the indeterminate nodule (zone I), the infection zone by the bacteria (zone II), the nitrogen fixation zone (zone III), and the plant cell senescence zone (zone IV). Recent molecular studies revealed the differential transcriptomic and epigenomic profiles existing between the different zones composing the indeterminate *M. truncatula* nodule. In determinate nodules (e.g., soybean), such zonation does not exist leading to the assumption that the soybean infection zone is cellularly and molecularly homogeneous. This assumption has been recently contradicted through the spatial profiling of over 100 metabolites using matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry imaging (MALDI-FTICR-MSI). Our single cell-resolution transcriptomic analysis confirmed this cellular heterogeneity in the infection zone of the nodule. In addition to distinguishing the uninfected (clusters J and K) and infected cells of the infection zone of the nodule (clusters F, G, and H) based on their transcriptomic profiles, we also observed different transcriptomic signatures between the cells composing the rhizobia-infected clusters F&G and H including the strong induction of the expression of the soybean leghemoglobin genes in cluster H. This result suggests that only the cells composing cluster H are actively fixing the atmospheric dinitrogen and allows us to reconsider the definition of a plant cell type not only based on differences existing in the morphology, physiology, or relative localization in an organ but also based on its molecular attribute. From a functional perspective, our data raise the possibility that the plant is capable of allocating a limited number of rhizobia-infected cells to an active nitrogen fixation process.

The transcriptional differences existing between clusters F&G and cluster H can also be very subtle. The high resolution and sensitivity of Molecular Cartography™ technology combined with the analysis of the sNucRNA-seq datasets revealed the differential nuclear and cytosolic distribution of the transcripts encoded by the Glyma.17G195900 and Glyma.15G210100 genes. More specifically, we observed two populations of infected cells according to the preferential retention of transcripts in the plant nucleus. Because our single-cell transcriptomic analysis is based on the use of isolated nuclei, we conclude that the cells composing the clusters F&H are cells characterized by the nuclear retention of the transcripts encoded by the Glyma.17G195900, Glyma.01G164600, and Glyma.15G210100 transcripts. Oppositely, the cells characterized by the absence of Glyma.17G195900, Glyma.01G164600, and Glyma.15G210100 transcript in the nuclei belong to cluster H. Based on these results, we hypothesize that plant cells use the nuclear retention of transcripts as an additional regulatory mechanism to control protein translation. This mechanism is independent of the level of expression of the gene and helps to fine tune the regulation of protein biosynthesis. In the context of the nodulation process, we hypothesize that the differential compartmentation of these transcripts between the cells composing the clusters F&G and cluster H is associated with the unique physiology of these two groups of rhizobia-infected cells.

The root and nodule single nucleus transcriptome atlases also help refine the identification of genes controlling biological processes. For instance, microdomain-associated proteins are central in controlling the nodulation process (i.e., flotillins, remorins, and FWL protein-coding genes have been previously reported to control legume nodulation). Here, we identified a new microdomain-associated protein-coding gene, *GmFWL3*, preferentially expressed in the infected cells of the nodule (i.e., clusters F, G, and H). While the profiles of expression of the *GmFWL1* and *FWL3* genes are similar at the single cell level (Fig.5c and Fig.S25), the subcellular localization of the GmFWL3 protein in the infected cells of the nodule is broader than previously reported for GmFWL1. Specifically, besides being localized in the plasma membrane of the infected cells of the nodule, we found GmFWL3 in the membrane of the symbiosome. This observation associated with the significant decrease in the rate of bacterial infection of the nodule cells upon knock-out of GmFWL3 suggests that the microdomain fraction of the symbiosis membrane plays a critical role in the communication between the plant cells and the bacteroids. The identification of other microdomain-associated proteins interacting with GmFWL3, including GmFWL1, and their co-expression in the same cell type of the nodule, support their interaction and the formation of a protein network in the microdomain fraction of the biological membranes of the rhizobia-infected cells.
Our findings highlight the future impact of single-cell genomics to dissect complex biological processes such as refining the biological concept of “cell type” by including high-resolution molecular attributes in this definition. Besides revealing the spatial organization of plant cells in an organ, we also demonstrate the potential of single-cell genomics in exploring the sub-cellular compartmentation of molecules. Here, the differential localization of selected transcripts between cell types is an additional level of control of protein translation, independent of the level of transcription of the genes. Revealing the regulatory mechanisms controlling the differential compartmentation of transcripts between cells is important to understand in order to maximize the impact of precision genetic engineering in plant biology.

**Methods**

**Bacterial culture**

*Escherichia coli, Agrobacterium rhizogenes* (K599), and *Agrobacterium tumefaciens* (GV3101) strains were grown in LB medium supplemented with appropriate antibiotics at 37°C and 30°C, respectively. *Bradyrhizobium diazoefficiens* USDA110 was grown in HM medium supplemented with 50 µg/ml chloramphenicol. Rhizobia cultures were grown for 3 d and pelleted at 4000 x g for 10 minutes, followed by washing and diluting to an OD₆₀₀ₙ₉ m of 0.1 in the nutritive NPNS solution for inoculation.

**Plant material**

Soybean (*Glycine max Williams 82*) seedlings were sterilized as previously described. The seeds were then placed on agar B&D medium in absence of nitrogen, and placed in a growth chamber. The root tips of six days-old plants were collected and processed to generate the sNucRNA-seq libraries. To isolate mature nodules, three days-old seedlings were germinated on B&D minus nitrogen agar media and inoculated with *Bradyrhizobium diazoefficiens* USDA 110 suspension (OD₆₀₀ₙ₉ = 0.1). After 72 hours of incubation in the dark, the seedlings were transferred in vermiculite:perlite mix (3:1) and grown in the growth chamber (16 hours light/ 8 hours dark) at 20–26°C. Mature nodules were collected 32 days after bacterial inoculation.

**Nuclei isolation, sNucRNA-seq library preparation, and sequencing**

For nuclei isolation, roots, and nodules were chopped and passed through a 30 and 40 µm cell strainers as previously described. The filtered nuclei were purified by cell sorting using FACS Aria II™ 603 cell sorter (BD Biosciences). An average of 80,000 to 100,000 nuclei were collected for each sample. The three sNucRNA-seq libraries of root and two for nodule were constructed following the Chromium™ Single Cell 3’ Library & Gel Bead Kit v3.1 protocol (10x Genomics). The sequencing of single-indexed, paired-end libraries was performed on an Illumina™ NovaSeq 6000 platform according to the 10x Genomics recommendations.

**Cloning, molecular constructs**

*Conventional cloning procedures:* All constructs were generated using either classical restriction enzyme-ligation or Gateway cloning strategies ([www.lifetechnologies.com](http://www.lifetechnologies.com)). For Gateway cloning, *GmFWL3* cDNA and *FWL3-GFP* were cloned into pDONR-Zeo by BP clonase reaction. Constructs were sequenced to confirm the integrity of the genes cloned. The *GmFWL3*-pDONR plasmid was used along with the pMDC43 and pMDC83 destination vectors in an LR reaction to generate p35S::GFP-FWL3 and p35S::FWL3-GFP translational fusion constructs. The p35S::CFP-AtSUN1 was previously described.
To generate myc-tag chimeric proteins, AttR1-CmR-ccdb-AttR2-10xMyc, and 10xMyc-AttR1-CmR-ccdB-AttR2 cassettes were amplified from pGWB20 and pGWB21\(^5\), respectively, and cloned in place of the HA tag in the CGT3304 plasmid using BamHI and Eco53kI restriction sites. After DNA sequence confirmation, the resulting promoter FMV::AttR1-CmR-ccdb-AttR2-10xmyc-tnos and pFMV::10xmyc-AttR1-CmR-ccdB-AttR2-tnos cassettes were excised by SbfI restriction enzyme digestion and ligated into AKK1467 at the SbfI restriction site, thereby creating AKK1467B-10myc-GW and AKK1467B-GW-10myc Gateway-compatible destination plasmids (See Fig.S27). The GmFWL3::DoNR-Zeo plasmid was used in an LR reaction to clone GmFWL3 cDNA into the modified AKK1467B-10myc-GW and AKK1467B-GW-10myc Gateway destination plasmids, generating pFMV::10myc-GmFWL3 and pFMV::GmFWL3-10myc constructs. The pFMV::10myc-FWL3, pFMV::FWL3-10myc, p35S::FWL3-GFP, p35S::GFP-FWL3 and their respective controls, pFMV::10Myc, and p35S::GFP, were transformed into Agrobacterium rhizogenes (strain K599) for hairy root transformation.

**CRISPR-Cas9 design and screening for mutation by band-shift and sequencing:** The guide RNAs used to knock out selected genes using CRISPR-Cas9 technology were designed using the guide RNA designer website\(^6\). Two GmFWL3 target sequences (referred to hereafter as GmFWL3-T1 and T2) were independently cloned into the Esp3I and BsaI sites of pAH595 guide RNA entry vector under the control of the AtU6 and At7SL promoters, respectively, using the Golden Gate method\(^7\). Upon cloning, the pAH595-GmFWL3-T1-T2 donor plasmid was created. Subsequently, the pAH595-GmFWL3-T1-T2 entry vector was used with pNJB184-CAS9 entry vector in a two-fragment multi-site Gateway LR clonase reaction. The cassettes were recombined into the pUB-GW-GFP binary vector\(^8\) which carries a GFP selectable marker for screening transgenic FWL3-CRISPR-Cas9 hairy roots\(^9\). The empty pAH595 donor plasmid combined with pNJB184-CAS9 in an LR reaction was used as a control. The resulting pLjUB::Cas9-pU6/At7SL::GmFWL3-T1-T2 binary plasmid (hereafter CAS9/GmFWL3-T1-T2) was transformed into Agrobacterium rhizogenes (K599) for hairy root transformation and nodule phenotyping.

To characterize the nature of the mutations induced by CRISPR/Cas9 in the FWL3 gene, the genomic DNA of transgenic soybean roots expressing the GFP was extracted. Using the Polymerase Chain Reaction (PCR) assay, the regions spanning the target sites were amplified using the GFWL3-5utr forward (5′-CCAAGTCCAATAACTATGCTTGAG-3′) and reverse primers (5′-TCAACGGCTCATGCCC-3′), then sequenced for analysis.

**Plant transformation and confocal laser scanning microscopy**

**Tobacco leaf infiltration and protoplast isolation:** Nicotiana benthamiana leaves were co-infiltrated with A. tumefaciens (GV3101) expressing the virus RNA-silencing suppressor protein HC-Pro to enhance the expression of the transgene, and the following constructs: p35S::GFP-GmFWL3, p35S::GmFWL3-GFP, and p35S::GFP. Three days after infiltration, the tobacco leaf cells were imaged using a Nikon A1 confocal microscope. To produce transgenic tobacco leaf protoplasts, infiltrated epidermal leaf cells were incubated in MKM medium [(9% mannitol, 0.037%KCl, 0.2M MOPS, and pH 6.0) supplemented with 0.05% driselase, 0.02% macerozyme R10, and 0.1% onozuka R10 (all Sigma, http://www.sigmaaldrich.com/)] for 3 h in dark conditions and at room temperature. The epidermal tobacco protoplasts expressing the translational fusions of GFP-FWL3 and FWL3-GFP as well as GFP control were imaged using a Nikon A1 confocal microscope. Plasma membrane staining was conducted by infiltrating 5µM of FM4-64 (SynaptoRed™ C2, Biotium # 70020) prior to microscopy.

**Soybean hairy root transformation:** Eleven-day-old soybean Williams 82 plants were used for hairy root transformation\(^2\). Three days after transformation, the shoot explants on rockwool cubes were watered with nutritive NPNS solution and allowed to grow for an additional 10 d. Then, the plants were transferred to autoclaved vermiculite perlite (i.e., mixed at 3:1), and grown for an additional 14 days to allow plants to develop roots. Plants were then inoculated with USDA110.
Thirty-das-old transgenic nodules, characterized by the expression of the GFP reporter, were collected under a Nikon SMZ25 epifluorescence stereoscope.

**Co-immunoprecipitation assay**

Co-immunoprecipitation experiments were performed as previously described using transgenic nodules isolated from GFP-positive transgenic roots (see above). Total protein extracts used for co-immunoprecipitation assays were obtained by grinding transgenic nodules in protein extraction buffer (50 mM Tris-MES pH 7.5, 300 mM sucrose, 150 mM NaCl, 10 mM potassium acetate, 5 mM EDTA, Sigma plant protease inhibitor cocktail and 1% Triton X-100). After a 30 min incubation in the protein extraction buffer, the proteins were filtered through a 40 µm filter (Fisherbrand #22-363-547) and then centrifuged (15000 x g, 10 minutes, 4°C). Prior to the co-immunoprecipitation assay, western blot assays were conducted to detect the tagged proteins using anti-Myc-HRP antibodies (Fisher R951-25) diluted 1:1000 in TBS-0.05% Tween20/2% skimmed milk. Co-immunoprecipitated proteins were isolated by applying total protein extracts to 50 µl of anti-myc Tag MicroBeads (Miltenyi Biotech, #130-091-284) on ice for 30 min and separated through a µColumn (Miltenyi Biotech # 130-042-701) in the magnetic field of the µMACS Separator system (Miltenyi Biotech,#s 130-042-602 and 130-042-303) according to the manufacturer’s protocol. Samples were eluted with SDS-PAGE sample buffer for further analysis as described below. Three replicates each for both N and C-terminal c-myc fusions to GmFWL3 were compared to three replicates of c-myc alone controls.

**Mass spectrometry and identification of GmFWL3-binding protein partners**

Co-immunoprecipitated proteins were denatured at 95°C for 5 min before they were loaded on a 10% Bolt Bis-Tris Plus gel. After brief electrophoresis to concentrate the proteins at the top of the gel, the proteins were fixed and then stained with colloidal Coomassie blue (Sigma). The areas of gel containing the co-immunoprecipitated proteins were excised and subjected to reduction and alkylation with DTT and iodoacetamide, respectively, and then washed with ammonium bicarbonate/acetonitrile to remove stain and SDS.

Trypsin digestion was carried out overnight at 37°C. Peptides were extracted from the gel pieces, dried down, and resuspended in 0.1% trifluoroacetic acid. Samples were desalted using an Oasis HLB µElution solid-phase extraction plate (Waters Corp, Milford, MA). Eluates were dried down and resuspended in 2.5% acetonitrile, and 0.1% formic acid. Peptides were then run by nanoLCMS/MS using a 2h gradient on a 0.075 mm x250 mm CSH C18 column (Waters Corp, Milford, MA) feeding into a Q-Exactive HF mass spectrometer.

All MS/MS samples were analyzed using Mascot (Matrix Science, London, UK; version 2.6.1). Mascot was set up to search the *Glycine max* protein database (Wm82.a4.v1l; 92226 records) for tryptic peptides. Mascot was searched with a fragment ion mass tolerance of 0.060 Da and a parent ion tolerance of 10.0 PPM. Deamidated asparagine and glutamine, oxidized methionine, and carbamidomethylated cysteine were specified in Mascot as variable modifications. Scaffold (version 4.8.9, Proteome Software Inc., Portland, OR) was used to validate MS/MS-based peptide and protein identifications. Peptide identifications were accepted with a probability of 80% or greater and < 1% FDR by the Peptide Prophet algorithm with Scaffold delta-mass correction. Protein identifications were accepted with a probability greater than 99.0%, < 1% FDR and at least 2 peptides per protein.

**Immunogold Labeling-Transmission Electron Microscopy (Imunogold-TEM)**

Thirty-day-old transgenic nodules expressing the pFMV::myc-GmFWL3, pFWL1::GmFWL3-myc, and pFMV::myc transgenes were isolated based on the observation of the GFP under a Nikon SMZ25 fluorescence stereoscope. The
nODULES WERE IMMEDIATELY FIXED IN 4% GLUTARALDEHYDE, 1% PARAFORMALDEHYDE, AND 0.2 M SODIUM CACODYLATE BUFFER (pH 7.2) FOR 1 H AT ROOM TEMPERATURE. FIXED NODULES WERE CUT IN HALF AND STORED AT 4°C OVERNIGHT IN THE BUFFER. FOR THE IMMUNOGOLD LABELING, SAMPLES WERE EMBEDDED IN LR-WHITE RESIN AND THEN SECTIONED TO GENERATE 100 NM THICK CROSS-SECTIONS. THE SECTIONS WERE COLLECTED ONTO NICKEL GRIDS, BLOCKED IN 1X PBS-TWEEN-BSA [0.05% (V/V) TWEEN 20 AND 3% BSA] FOR 30 MINUTES, AND RINSED IN PBS-0.05% TWEEN (PBST). THE GRIDS WERE THEN LABELED WITH ANTI-MYC ANTIBODIES (R950-25, THERMOFOISHER) DILUTED 1:50 IN PBS-0.05% TWEEN-1% BSA FOR 1 HOUR. AFTER THREE WASHES IN PBST, THE LABELED SAMPLES WERE INCUBATED FOR 1 HOUR WITH THE SECONDARY ANTIBODY CONJUGATED WITH 10 NM COLLOIDAL GOLD (A-31561, THERMOFOISHER) AND DILUTED 1:100 IN PBS-0.05% TWEEN-1% BSA. THE SECTIONS WERE RINSED 3 TIMES IN PBS-0.05% TWEEN 20 THEN ONCE IN DEIONIZED WATER. ELECTRON MICROSCOPE IMAGES WERE COLLECTED WITH A HITACHI H7500 TEM OPERATED AT 80 KV. WHEN NECESSARY, SECTIONS WERE STAINED USING URANYL ACETATE AND LEAD CITRATE. FOR EACH TRANSGENE, THREE INDEPENDENT BIOLOGICAL REPLICATES WERE PROCESSED AND OBSERVED UNDER THE TRANSMISSION ELECTRON MICROSCOPE.

**SNucRNA-seq DATA PRE-PROCESSING, INTEGRATION, AND CLUSTERING**

Each SNucRNA-seq library was processed individually using the 10x Genomics Cell Ranger software v6.1.1.0 for the demultiplexing and for the alignment against the soybean reference genome from Ensembl Plants database (i.e., Glycine_max_v2.1.52; http://ftp.ensemblgenomes.org/pub/plants/release-52/fasta/glycine_max/). The background contamination was subtracted after the alignment of the reads using using SoupX\(^\text{12}\), and doubles were filtered out using the DoubletDetection prediction method\(^\text{13}\). Finally, we applied a minimum threshold of 500 UMIs to remove the nuclei with lower transcriptional content. Upon normalization, the integration anchors were defined for the combined set of 3 SNucRNA-seq datasets of the root and 2 SNucRNA-seq datasets for the nodule using the tool Seurat V4\(^\text{14}\). The dimensional reduction for the complexity of the data was performed using the UMAP method with the first 40 principal components and selecting the top 2000 variable genes for the clustering by using the FindClusters method from Seurat V4. We used the Seurat object for the soybean nodule UMAP to generate expression distribution plots of the set of genes using the function RidgePlot from Seurat V4. The DotPlot function of Seurat V4 was used to generate the DotPlot expression figures.

**CELL TYPE ANNOTATION BY USING MOLECULAR CARTOGRAPHY™**

To annotate the soybean root and nodule cell clusters, we identified the top specific-expressed genes for each cluster using FindAllMarkers function in Seurat V4. Probes were designed by Resolve Biosciences (Monheim am Rhein, Germany) and hybridized against 10 μm thick cross-sections of fixed root and nodule and embedded in paraffin. Upon hybridization, the cross sections were stained with DAPI and calcofluor white, and microscopic observations were conducted to reveal the position of the nuclei and cells, respectively. The microscopic images with transcript locations were analyzed using the “Molecular Cartography™” plugin for ImageJ analysis software provided by Resolve Biosciences.

**CELL TYPE ANNOTATION BY USING ORTHOLOG GENE MARKERS**

To support the annotation of soybean root cell clusters, we identified soybean orthologous genes to previously functionally validated cell-type specific gene markers in the root of *Arabidopsis thaliana*\(^\text{15}\) and *Medicago truncatula*\(^\text{16}\) from the literature (Table.S2). Gene orthology was based on the identification of syntenic regions by CoGe (genomeevolution.org/coge) between the *Glycine max* and *Arabidopsis thaliana*, and *Glycine max* and *Medicago truncatula* genomes.

**UMAP VISUALIZATION**
For visualization, all the sNucRNA-seq libraries for the root, nodule, and the integration of both were combined using the Cell Ranger `aggr` function from 10X Genomics and we used Loupe software from 10X Genomics to visualize the integrations.

**Differential Gene Expression analysis and Gene Ontology analysis**

To identify differentially expressed genes (DEGs) between clusters, we applied the DEsingle software\(^{17}\), a Zero-inflated negative binomial distribution method\(^{18}\), on the raw read counts using a p-value < 0.05 and Fold Change greater than 1.5. Gene ontology enrichment analyses were performed on the DGEs using the Plant Transcriptional Regulatory web page and applying a threshold p-value ≤ 0.01 (http://plantregmap.gao-lab.org/go.php).

**Comparison of soybean sNucRNA-seq and bulk RNA-seq**

To evaluate the depth and sensitivity of the soybean root and nodule single nuclei transcriptome atlases, we compared our pseudo-bulk sNucRNA-seq datasets with previously published root and nodule bulk RNA-seq datasets\(^ {19}\). Using the database of the legume information system (LIS), we extracted bulk expression datasets (2022/11/14; https://data.legumeinfo.org/Glycine/max/expression/Wm82.gnm2.ann1.expr.Wm82.Libault_Farmer_2010/; identifiers #SRR037385, SRR037386, and SRR037387 of the nodule, root tip, and entire root, respectively) before comparing the number of expressed genes between bulk and pseudo-bulk RNA-seq libraries.

**Multi-Dimensional Scaling analysis of the soybean root and nodule transcriptomes**

To evaluate the level of similarity across the single-nuclei transcriptomes of the soybean root (16 cell clusters) and nodule (11 cell clusters), we conducted a Multidimensional scaling (MDS) analysis using the library `ggfortify` in R (version 4.2.2). Specifically, a classical multidimensional scaling was performed to calculate a distance matrix between the different objects.

**Reanalysis of the *M. truncatula* nodule single-cell RNA-seq datasets**

Upon mining the *Medicago truncatula* nodule single-cell RNA-seq datasets (i.e., SAMC899255, and SAMC899256 from the National Genomics Data Center), we individually processed both datasets using the 10X Genomics Cell Ranger v6.1.1.0 pipeline to map the sequencing reads against the *M. truncatula* reference genome (https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/). Upon removing the “MIX” cluster 0 as described previously\(^ {20}\), we applied the same analytical methodology as described above. For visualization purposes, we generated expression Ridge plots on selected Medicago and soybean genes using the RidgePlot function of Seurat V4.

**Analysis of Nuclear Retention of Infected Cell Transcripts**

Cross sections of soybean nodule were stained with both DAPI and Calcoflour White to reveal nuclei and cell walls, respectively. Ten micron thick high-resolution Z-stack images of stained soybean nodule cross-sections were generated through confocal microscopy using a Plan-APOCHROMAT 50× / 1.2 auto-correction and auto-immersion (water) with Magnification changer 0.5×. Each individually tiles were scanned 296µm x 296µm, in pixel 2144x2144, stitched and used for image analysis. Image analysis was conducted using the following ImageJ macros: https://github.com/Llamero/Libault_Transcript_Localization. First, gridlines left over from the tile scans were removed by linear interpolation across the gap between tiles using the ImageJ macro "Fill grid lines." Count masks of the nuclei were
generated from 16-bit z-stack DAPI stained images using the ImageJ macro “Threshold Stack” using default parameters and setting thresholds for particle analysis filters for mean fluorescence, area of fluorescence, standard deviation of fluorescence, and circularity of fluorescent bodies. Thresholds were set for each filter to optimally define infection zone nuclei. Ideal parameters to identify cell wall boundaries were identified by watershedding Gaussian blurred 16-bit Z-stack Calcofluor White images using the ImageJ macro “Segment Cells”. (Default values were used for “Segment Cells” macro; including Gaussian blur value = 20, Minimum/maximum watershed tolerance = 100/1000) Watershed tolerance values were set to minimize excessive segmentation of cells, and ideal parameters were saved to be used for final analysis macro. Transcript coordinate data for all genes analyzed was provided by Resolve Biosciences for each nodule cross section tested using Molecular Cartography. Genes used in this study (17G195900, 15G210100, 05G216000) were selected as they are specifically expressed in nodule infection zone cells identified to be infected by *Bradyrhizobium bacteria* based on presence of the bacterial transcript BAC47034 in analyzed cell populations. Using the nuclear mask, defined cell segmentation parameters, and transcript coordinate data, the ImageJ macro “Distance Histogram By Cell” performed a Euclidian distance transform to distinguish nuclear and cytosolic localized transcripts for infected cells of the infection zone. Cells identified as having no nuclei from our generated datasets were removed from the analysis and generation of scatter plots.

**Declarations**

**Data availability**

The sNucRNA-seq data generated in this study have been deposited with the National Center for Biotechnology Information (NCBI) bioproject number PRJNA938968 and the GEO number GSE226149. The webpage supporting access to the data from this manuscript is http://soybeancellatlas.org.

**Acknowledgments**

We thank Dr. David Evans (Oxford Brookes University, UK), and Dr. Tsuyoshi Nakagawa (Shimane University, Matsue, Japan) for sharing the p35S: CFP-AtSUN1 and pGWB20/21 plasmids, respectively. We also thank Ms. Rosa Angelica Bedolla-Gaxiola for the artistic work (Fig.1a and.2a). We thank Dirk Anderson, Jules Russ, and Dr. You Zhou of the Flow Cytometry, and the Microscopy Core Research Facilities at the University of Nebraska-Lincoln Center for Biotechnology for helping with nuclei sorting, confocal microscopy, and transmission electron microscopy. This research was funded by the Nebraska Soybean Board, the NSF awards #2127485 and #1854326, and the Nebraska Research Initiative.

**Author contribution**

SACP, PZ, and ML designed the experiments; STh generated the single-cell RNA-seq libraries; PZ, MSH, HX, IM, and AN performed the molecular cloning, Co-IP, and confocal observations, supervised the biochemical and TEM experiments; MJN carried out the mass spectrometry identification of the GmFWL3 co-immunoprecipitated proteins, JK, SB, and STe conducted the molecular cartography™ experiments, QY and PM created the web-based tools for data visualization, SACP, PZ, STh, STe, SSMVS, and ML analyzed the results, SACP, PZ, and ML wrote the manuscript.

**References**


**Figures**
Establishment of the soybean root single nucleus transcriptome atlas. **a.** Uniform Manifold Approximation and Projection (UMAP) plot of 14,369 soybean root nuclei according to their transcriptomic profiles. These nuclei were clustered in 16 different groups. **b.** Distribution of the number of Unique Molecular Identifiers (UMI) and expressed genes per root cluster (Tukey’s test with P-value<0.05 reported to highlight differences between clusters). **c.** Dotplot representation of the expression of 52 soybean root cell-type-specific marker genes validated using the Molecular Cartography technology (Fig.S6). **d.** Dotplot representation of the expression of root cell-type specific marker genes identified based on previous functional genomic studies and based on their orthologous relationships with root cell-type-specific *Arabidopsis thaliana* and *Medicago truncatula* marker genes (Fig.S7). For these two dotplot figures, the percentage of nuclei expressing the gene of interest (circle size) and the mean expression (circle color) of the genes are shown. **e.** Integrated analysis of the expression of several soybean root marker genes using Molecular Cartography technology on a soybean root cross-section. Left panel: Detection of transcripts from epidermal (blue) and cortical (purple) markers genes. The epidermal marker genes are GLYMA-19G255500, CYP93A1, GLYMA-06G259400, G4DT, GLYMA-09G099900, GLYMA-10G070200, GLYMA-01G156200, GLYMA-07G130800, GLYMA-20G061300, GLYMA-02G149100, GLYMA-04G010600, and GLYMA-17G133100. The cortical marker genes are GLYMA-06G235500, GLYMA-11G221200, GLYMA-09G216800, and GLYMA-15G169100. Central panel: Detection of transcripts from endodermal (light pink color; see arrows in the magnified...
The endodermal marker genes are GLYMA-14G218700, and GLYMA-16G106800. The pericycle marker genes are GLYMA-02G003700, GLYMA-05G023700, GLYMA-11G078300, GLYMA-08G125800, and GLYMA-09G127000. Right panel: Detection of transcripts from xylem (red) and phloem (brown) markers genes. The xylem marker genes are GLYMA-15G245800, GLYMA-04G063800, GLYMA-06G065000, GLYMA-13G334500, GLYMA-15G040000, GLYMA-18G197400, and GLYMA-15G179500. The phloem marker genes are GLYMA-07G006500, GLYMA-05G216000, GLYMA-15G274200, GLYMA-11G243100, and GLYMA-12G154300. ED, endodermis; PF, phloem fiber; RH, root hair cells; RC, root cap cells; SCN, Stem Cell Niche.

Figure 2

Establishment of the soybean nodule single nucleus transcriptome atlas. a. UMAP plot of 7,830 soybean nodule nuclei according to their transcriptomic profiles. These nuclei were clustered in 11 different groups (clusters A to K). b. Distribution of the number of Unique Molecular Identifiers (UMI) and expressed genes per nodule cluster (Tukey’s test with
P-value < 0.05 reported to highlight differences between clusters. **c.** Dotplot representation of the expression of 34 soybean nodule cell-type-specific marker genes validated using the Molecular Cartography technology (Fig. S12 and S14).

**d.** Dotplot representation of the expression of 6 nodule cluster I-specific marker genes. The dot size of c and d represents the percentage of cells in which each gene is expressed. For these two dotplot figures, the percentage of nuclei expressing the gene of interest (circle size) and the mean expression (circle color) of the genes are shown. **e to h.** Integrated analysis of the expression of several soybean nodule marker genes using Molecular Cartography technology on a soybean nodule cross-section. **e.** Detection of transcripts from the inner/outer cortical cells (blue) and from the sclereid layer (pink). The inner/outer cortical marker genes are GLYMA-16G039800, GLYMA-19G255500, CYP93A1, GLYMA-06G259400, G4DT, GLYMA-01G156200, GLYMA-20G061300, and GLYMA-03G079500; The sclereid layer marker genes are GLYMA-04G063800, GLYMA-06G065000, GLYMA-13G334500, and GLYMA-15G040000. **f.** Detection of transcripts from the vascular endodermis (light pink) and vascular bundle (orange). The vascular endodermis marker genes are GLYMA-04G218700, GLYMA-14G227200, GLYMA-16G106800, and GLYMA-20G151700. The vascular bundle marker genes are GLYMA-06G256000, GLYMA-10G139200, GLYMA-02G003700, GLYMA-07G231500, GLYMA-18G062100, GLYMA-08G125800, GLYMA-11G078300, GLYMA-09G127000, GLYMA-13G334500, and GLYMA-15G040000. **g.** Detection of *B. diazoefficiens* transcripts in the infection zone of the nodule (yellow color) from 9 different genes: BAC45727, BAC46169, BAC47034, BAC48395, BAC51072, BAC51722, BAC52602, BAC52793, and BAC52805. **h.** Detection of plant transcripts in *B. diazoefficiens* infected (red) and uninfected cells (green; green arrows). The infected cell marker genes are GLYMA-17G195900, GLYMA-01G164600, GLYMA-15G2100100, and GLYMA-05G216000. The uninfected marker genes are GLYMA-06G235500, and GLYMA-06G002000.
Figure 3

Identification of distinct populations of *B. diazoefficiens*-infected cells based on their transcriptional profiles. 

**a.** Comparative cell Multi-Dimensional Scaling (MDS) plot of the 16 root and 11 nodule clusters. **b.** Identification of differentially expressed genes (DEGs) between the *B. diazoefficiens*-infected cell cluster of the soybean nodule (i.e. clusters F, G, and H). For each population of DEGs, we highlight the top enriched “gene ontology” categories. **c to g.** Ridge plot distributions of the expression of the soybean CCS52A (c.) and leghemoglobin genes (d.), as well as genes controlling host-range restriction (e.), nodule defense (f.), and bacterial maturation (g.) as defined by Roy et al., 2020 (Table.S4). **h.** Enhanced UMAP plot of 4,368 Medicago nodule nuclei according to their transcriptomic profiles. The single cell RNA-seq datasets were mined from Ye et al., 2022. These nuclei were clustered in 8 different groups. **i to m.** Ridge plot distributions of the expression of the Medicago CCS52A(i.) and leghemoglobin genes (j.), and the genes controlling host-range restriction (k.), nodule defense (l.), and bacterial maturation (m.) as defined by Roy et al., 2020 (Table.S4).
**Figure 4**

**Differential distribution of plant gene transcripts between the nuclear and cytosolic compartments of the *B. diazoefficiens*-infected cells of the soybean nodule.**

**a.** DAPI staining of the infection zone of the soybean nodule to label plant nuclei (white arrows).

**b.** Detection of *B. diazoefficiens* BAC47034 transcripts using Molecular Cartography technology to reveal the set of infected cells in the infection zone of the soybean nodule.

**c to e.** Detection of Glyma.17G195900 (c), Glyma.15G210100 (d), and Glyma.05G216000 (e) transcripts using Molecular Cartography technology that revealed the differential distribution of transcripts between the nucleus and the cytosol of the *B. diazoefficiens*-infected cells of the nodule.

**f to h.** Scatter plots showing the differential distribution of the Glyma.17G195900, Glyma.15G210100, and Glyma.05G216000 transcripts in the nucleus of the infected cells of the soybean nodule. For each scatter plot, the x- and y-axes show the percentage of nuclear transcript vs. total transcripts for two genes in each of the 775 *B. diazoefficiens*-infected cells analyzed: Glyma.15G210100 (x-axis) vs. Glyma.17G195900 (y-axis) (f); Glyma.15G210100 (x-axis) vs. Glyma.05G216000 (y-axis) (g); and Glyma.17G195900 (x-axis) vs. Glyma.05G216000 (y-axis) (h).
For each scatter plot and each populations of cells, the linear trendlines and its equation are indicated.

Figure 5

**Functional characterization of the GmFWL3 gene, a new microdomains-associated protein-encoding gene.** Expression patterns of the GmFWL genes expressed in at least one of the sixteen root(a) and eleven nodule clusters (b). c. Preferential expression of the GmFWL3 gene in the 11 soybean nodule clusters. GmFWL3 is most expressed in the clusters F, G, and H, the infected cells clusters of the nodule. The color scale reflect the level of activity of the genes and is linear. d. Mutagenesis of GmFWL3 using CRISPR-Cas9 technology. Representative transgenic soybean roots transformed with CAS9/pAH595 control (pUB-CAS9-pAH595) and CAS9/GmFWL3-T1-T2 transgenes. The GFP signal was used as a reporter of the transgenic roots (white arrow). e. Average number of nodules growing on 150 pUB-CAS9-pAH595 and 201 pUB-CAS9/GmFWL3-T1-T2 GFP-positive transgenic roots (student t-test: p-value = 0.009). f. Representative SYTO13 staining of transgenic soybean roots transformed with CAS9/pAH595 control (pUB-CAS9-pAH595) and CAS9/GmFWL3-T1-T2 transgenes. This staining reveal the density of rhizobial bacteria in the infected cells of the soybean nodules. g. Quantification of the infection rate in with CAS9/pAH595 control and CAS9/GmFWL3-T1-T2 transgenic nodule cells upon SYTO13 staining of the bacterial and nuclear DNAs (E) (AnovaSingle Factor Test a=0.05, *P<1e-100, n≈350).
Subcellular localization of GmFWL3 in tobacco leaf using confocal microscopy (a-c) and in soybean nodules using transmission electron micrography after immunogold labeling (d). Optical cross sections of tobacco leaf cells transiently expressing p35S::GmFWL3-GFP (a) and counterstained with membrane dye SynaptoRed (FM-64, b) exhibit a punctate plasma membrane localization of GFP-GmFWL3 fusion proteins in tobacco leaf cells. The co-localization of GFP-GmFWL3 signal with membrane dye FM64 confirms its membrane localization (c). Scale Bar: 20 µm. Representative images of gold particle distribution in the plasma membrane after immunogold labeling against the c-myc-tagged GmFWL3 chimeric protein (d). In addition to its plasma membrane localization (d), clusters of gold particles were detected in the nucleus, nuclear envelope, symbiosomal membranes, vacuolar membrane, plasma membrane, and in vesicular and other membrane bound organelles (Fig.S23). This result suggests that the myc-GmFWL3 protein is localized in these biological membranes. White arrow heads point to clusters of the gold particles. CW, cell wall; PM, plasma membrane. Scale bar: 400nm. Pie chart of the distribution of putative GmFWL3 soybean proteins interacting partners according to their biological function (e). Dotplot representation of the expression of 10 soybean genes interacting with GmFWL3 including GmFWL3 itself. Among them, nine are preferentially expressed in clusters F, G, and H (Fig.S25 and S26). The percentage of nuclei expressing the gene of interest (circle size) and the mean expression (circle color) of the genes are shown.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.