nature plants

Resource

https://doi.org/10.1038/s41477-022-01291-y

# Single-cell RNA sequencing provides a high-resolution roadmap for understanding the multicellular compartmentation of specialized metabolism

Received: 5 December 2021

Accepted: 26 October 2022

Published online: 15 December 2022

Check for updates

Sijie Sun[1,5], Xiaofeng Shen[1,5], Yi Li[1,5], Ying Li[1], Shu Wang[1], Rucan Li[1], Huibo Zhang[1], Guoan Shen[1], Baolin Guo[1], Jianhe Wei[1], Jiang Xu[2], Benoit St-Pierre[3] ✉, Shilin Chen[2,4] ✉ & Chao Sun[1] ✉

Monoterpenoid indole alkaloids (MIAs) are among the most diverse specialized metabolites in plants and are of great pharmaceutical importance. We leveraged single-cell transcriptomics to explore the spatial organization of MIA metabolism in *Catharanthus roseus* leaves, and the transcripts of 20 MIA genes were first localized, updating the model of MIA biosynthesis. The MIA pathway was partitioned into three cell types, consistent with the results from RNA in situ hybridization experiments. Several candidate transporters were predicted to be essential players shuttling MIA intermediates between inter- and intracellular compartments, supplying potential targets to increase the overall yields of desirable MIAs in native plants or heterologous hosts through metabolic engineering and synthetic biology. This work provides not only a universal roadmap for elucidating the spatiotemporal distribution of biological processes at single-cell resolution, but also abundant cellular and genetic resources for further investigation of the higher-order organization of MIA biosynthesis, transport and storage.

Monoterpenoid indole alkaloids (MIAs), with over 3,000 different known structures, are one of the largest and most diverse groups of plant alkaloids; some are used as anticancer, anti-arrhythmic, antihypertensive and antimalarial drugs[1,2]. *Catharanthus roseus* (L.) G. Don is the best-characterized MIA-containing plant and produces more than 130 MIAs, including the anticancer drugs vinblastine and vincristine, and the antihypertensive agents ajmalicine and serpentine[3]. With the ability to produce all major MIA skeletons, *C. roseus* has emerged as the most valuable model system for investigating MIA biosynthesis in plants.

The MIA biosynthesis pathway has been studied extensively in *C. roseus* (Supplementary Fig. 1). The central intermediate, strictosidine, is generated by the coupling of tryptamine (indole moiety) and secologanin (monoterpene moiety) catalysed by the signature enzyme strictosidine synthase[4,5]. Tryptamine is derived from L-tryptophan via decarboxylation, whereas secologanin is synthesized through 11 sequential enzymatic steps from the universal building blocks of all terpenoids: isopentenyl pyrophosphate and dimethylallyl pyrophosphate. Strictosidine is further modified into different MIAs found in nature by a series of tailoring enzymes, including cytochrome P450s,

[1]Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [2]Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China. [3]EA2106 Biomolécules et Biotechnologies Végétales, Université de Tours, Tours, France. [4]Chengdu University of Traditional Chinese Medicine, Chengdu, China. [5]These authors contributed equally: Sijie Sun, Xiaofeng Shen, Yi Li. ✉e-mail: benoit.stpierre@univ-tours.fr; slchen@icmm.ac.cn; csun@implad.ac.cn

alcohol dehydrogenases, acetyltransferases and α/β hydrolases[6–8]. The anticancer drugs vinblastine and vincristine are dimeric indole alkaloids that come from the condensation of catharanthine and vindoline metabolites catalysed by peroxidase in *C. roseus*[9].

Although more than 30 enzymes involved in MIA biosynthesis have been identified in *C. roseus*, their expression patterns at the single-cell level remain largely unknown. Initial studies based on RNA in situ hybridization (RIH) and immunocytochemistry suggested that at least four cell types, namely, internal phloem-associated parenchyma (IPAP) cells, epidermal cells (ECs), laticifer cells (LCs) and idioblast cells (ICs), are involved in the MIA biosynthesis pathway in *C. roseus* leaf tissues[10–12]. Recently, imaging mass spectrometry (MS) and single-cell MS have been used to measure the cell type-specific localization of MIAs in *C. roseus*, providing another way to understand the multicellular compartmentation of MIA biosynthesis in plants[13,14]. These metabolomic studies suggested that most iridoid-type intermediates were localized in the ECs, whereas MIAs, serpentine and vindoline were localized in the ICs.

Here, by optimizing the single-cell RNA sequencing (scRNA-seq) procedure, including protoplasting, gene expression quantification and cell type annotation, we built the first high-resolution single-cell expression atlas of *C. roseus* leaves and provided a successful paradigm in which scRNA-seq technology was applied to dissect the multicellular compartmentation of specialized metabolism in plants at high throughput with transcriptomic analysis at the single-cell level. The expression signature of MIA pathway genes revealed that MIA biosynthesis initializes in IPAP cells; the middle enzymatic steps of the pathway mainly occur in ECs, and the late steps preferentially occur in ICs. The transporters trafficking MIA intermediates between these cells or the subcellular compartments within them were also investigated. A deep understanding of the multicellular compartmentation of MIA biosynthesis will facilitate the development of methods to increase the yields of highly valuable MIAs by engineering native plants or microbial chassis. Moreover, the large-scale leaf single-cell atlas and chromosome-level genome assembly presented here constitute important cellular and genetic resources to further explore the complex MIA pathway architecture in *C. roseus*.

## Results

### Leaf architecture and protoplasting

The anticancer drugs vincristine and vinblastine are synthesized mainly in the leaves of *C. roseus*. Owing to the presence of metabolites that emit distinctive autofluorescence under ultraviolet (UV) light excitation, specialized cells are easy to distinguish from the surrounding parenchyma tissues[15]. The isolated ICs dispersed in the mesophyll emit bluish-white fluorescence, whereas the elongated LCs emit yellow–green fluorescence (Fig. 1a). Both ICs and LCs have been reported to participate in MIA biosynthesis and accumulation. Protoplast preparation is the first key step in scRNA-seq analysis. Therefore, we systematically optimized protoplast isolation protocols and found that the yield and viability of the obtained protoplasts reached a great balance when using a digestion time of 2 h (Supplementary Tables 1 and 2). Unfortunately, we did not detect LCs in the enzymatically digested mixture (Fig. 1b). A possible explanation is that their extremely long tubular shapes hinder their release from the leaf tissue in which they are embedded or that they are easily broken when leaves are cut into thin strips during protoplast preparation. Live and intact cells were distinguished from dead cells and cell debris using trypan blue and fluorescein diacetate staining (Fig. 1c,d). Finally, approximately 65,000 protoplasts with an average viability of 95.67% from three biological replicates were used to generate single-cell transcriptomes.

### Generation of a *C. roseus* leaf cell atlas

A high-quality genome is essential for accurate quantification of a single-cell transcriptome. Here, we assembled a chromosome-level genome of *C. roseus* (BioProject: PRJNA841429) using PacBio long-read sequencing, BioNano optical mapping and high-throughput chromosome conformation capture (Hi-C) technologies (Fig. 2a). We obtained a final assembly of 573.11 Mb anchored onto eight pseudochromosomes with a scaffold N90 of 59.13 Mb. The benchmarking universal single-copy orthologues (BUSCO) evaluation (91.53%) and a mapping rate (96.96%) of Illumina sequencing data revealed high genome completeness. Finally, 37,297 protein-coding genes were predicted in the *C. roseus* genome.

Three single-cell complementary DNA libraries of *C. roseus* leaf protoplasts were subjected to Illumina high-throughput sequencing, and a total of 550.62 Gb of scRNA-seq data were generated (Supplementary Fig. 2 and Supplementary Table 3). Compared with the standard pipeline, Cell Ranger, Alevin reduces the bias caused by multimapping reads and improves the accuracy of gene abundance estimates, especially for highly homologous genes[16] (Extended Data Fig. 1). In addition, the clustering results and downstream analysis were not affected substantially by the use of different quantification pipelines in the current study (Extended Data Figs. 2 and 3). Because the genes involved in specialized metabolism in plants are often present in multiple copies or are members of gene families with relatively high similarity with each other[17,18], we chose Alevin to quantify the scRNA-seq data sets and recovered 40,530 cells with a median of 2,239 genes and a median of 5,931 unique molecular identifiers (UMIs) per cell (Supplementary Table 4). Cells with abnormal gene numbers and UMI counts and high percentages of mitochondrial (>3%) or chloroplast genes (>40%) were filtered out, resulting in 34,392 high-quality cells (Fig. 2b and Supplementary Table 4). By comparing the gene expression profiles of pooled scRNA-seq and leaf bulk RNA-seq, we found that the two data sets were highly correlated ($r = 0.88$), implying that the scRNA-seq data accurately recovered the global transcriptome profile in *C. roseus* leaves. In addition, we examined the protoplasting-induced shift in gene expression in *C. roseus* leaves using a real-time polymerase chain rection (PCR) and found that the expression of most MIA pathway genes was not remarkably affected by protoplasting (Supplementary Fig. 3). Two thousand highly variable genes were subjected to principal component analysis (PCA) for unsupervised clustering to cluster these high-quality cells. Finally, 14 transcriptionally distinct cell clusters were generated, which were represented in two- and three-dimensional uniform manifold approximation and projection (UMAP) plots (Fig. 2c and Supplementary Data 1).

We compiled a list of cell type markers in *C. roseus* using the following strategies to accurately annotate cell clusters: (1) identifying *C. roseus* genes whose localizations have been determined using RIH or immunocytochemistry; and (2) selecting gene markers whose orthologues in other species have been well studied (Supplementary Table 5). According to the expression patterns of these markers, clusters 8, 10, 11 and 12 belonged to the vascular cells (VC), proliferating cells (PCs), IC and IPAP populations, respectively. Clusters 0–5 and 13 were adjacent to each other on the UMAP plot and were all annotated as mesophyll cells (MCs), which indicated high heterogeneity among MCs (Fig. 3a–e, Supplementary Fig. 4 and Extended Data Fig. 4). Clusters 6 and 9 were assigned as ECs, and all guard cell (GC) markers were enriched in clusters 6 and 9, suggesting the existence of GCs in the EC population (Extended Data Fig. 5a). Therefore, the EC population was further dissected into subclusters, and cluster EC_5 was assigned as GC cells with GC markers. In addition, the transcription factor *ARABIDOPSIS THALIANA MERISTEM LAYER 1* (*ATML1*)[19] and wax-associated protein *ECERIFERUM3* (*CER3*)[20] were used to distinguish young and mature ECs. Finally, EC_2 and EC_3 were classified as EC1s (young ECs), whereas EC_0, EC_1 and EC_4 were classified as EC2s (mature ECs) (Fig. 3a and Extended Data Fig. 5b). Owing to the absence of any known marker genes, cluster 7 was named the unknown (UN) population. After annotation, we obtained nine broad cell populations: MC, EC1, EC2, GC, VC, PC, IC, IPAP and UN (Fig. 3f). The identities of most cell populations have been validated by RIH experiments (Fig. 3b–e and Extended Data

https://doi.org/10.1038/s41477-022-01291-y

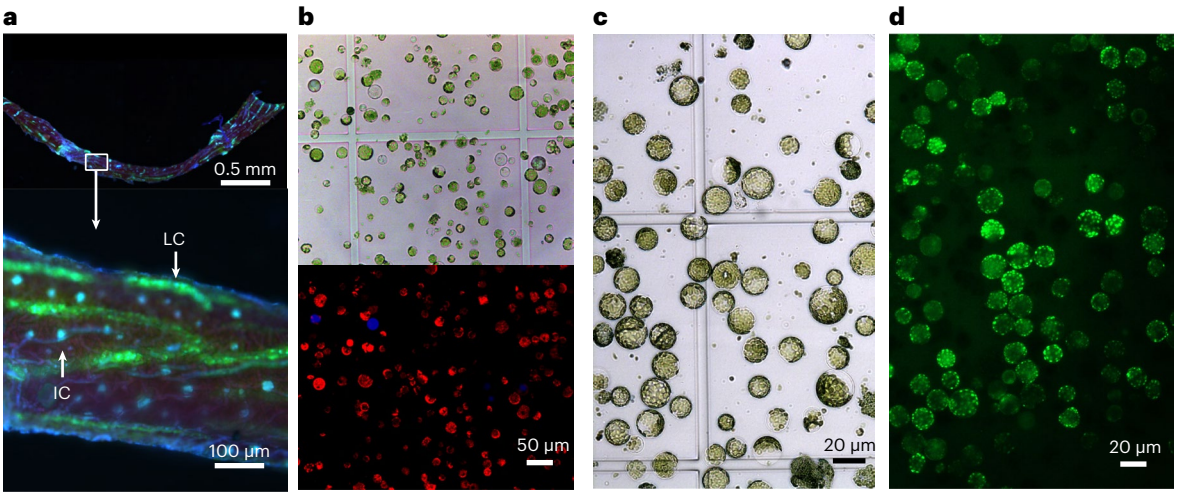



**Fig. 1 | Anatomical features of *C. roseus* leaves and isolation of protoplasts.** **a**, Leaf cross-section under UV excitation showing the distribution of ICs and LCs. **b**, Enzymatically digested mixture under bright light (upper) and UV excitation (lower, 360–380 nm). Chlorophyll exhibits red autofluorescence upon UV excitation. **c**,**d**, Protoplasts stained with trypan blue (**c**, bright field) and fluorescein diacetate (**d**, 488 nm). Live cells cannot be stained by trypan blue and can emit green fluorescence after staining with fluorescein diacetate.

Fig. 4). By detecting differentially expressed genes among distinct cell populations, we also identified a number of cell type-specific marker genes (Supplementary Table 6), which will provide a valuable resource for future functional analysis of developmental processes and specialized metabolism in *C. roseus* leaves.

## Cell type-specific distribution of MIA biosynthesis

We analysed the expression patterns of the characterized genes involved in the MIA biosynthesis pathway at single-cell resolution to explore the spatial distribution of this pathway in *C. roseus* leaves (Supplementary Table 7). In higher plants, isopentenyl pyrophosphate and its isomer dimethylallyl pyrophosphate are common building blocks for terpenoid biosynthesis and are synthesized by the plastidic methyl erythritol phosphate (MEP) or cytosolic mevalonate (MVA) pathway. Using single-cell transcriptomic analysis, we found that all genes in the MEP pathway and those in the first seven steps of the seco-iridoid pathway were expressed at high levels in IPAP cells (Fig. 4a,b), which supported the notion that the terpenoid moiety of MIAs is derived mainly from the MEP pathway[21] and that MIA biosynthesis is initiated in IPAP cells[22,23]. On the other hand, most of the genes in the MVA pathway were preferentially expressed in ECs. The MVA pathway was reported to fuel the biosynthesis of phytosterol, sesquiterpenoids and triterpenoids[21]. This suggests that one or some types of these chemicals were actively produced in ECs. The enzymes in the middle part of the vinblastine biosynthesis pathway from loganic acid methyltransferase to tabersonine synthase and catharanthine synthase (CS), as well as tryptophan decarboxylase (TDC), the last enzyme of the indole biosynthesis pathway, were highly enriched in ECs. Tabersonine is further converted to vindoline by a seven-step biosynthesis pathway (Fig. 4b). Our observations showed that the first three steps mainly occurred in ECs and that the last two steps occurred in ICs, consistent with previous reports based on RIH or immunocytochemistry[10–12]. The localization of the transcripts encoding 2,3-dihydrotabersonine *N*-methyltransferase was analysed for the first time and revealed preferential expression both in ECs and ICs. Heteroyohimbine synthases are not involved in the biosynthesis of vinblastine and vincristine. However, they convert strictosidine aglycones into heteroyohimbine-type MIAs, such as tetrahydroalstonine and ajmalicine. The *C. roseus* genome encodes six heteroyohimbine synthases, and most of them were expressed at high levels in ICs, suggesting that ICs might be an important cell type ensuring the diversity of MIAs in *C. roseus*.

As shown in Fig. 4b, intermediate transport between distinct cell types occurred at least three times during vinblastine biosynthesis. Four families of transporters are involved in the transport of plant alkaloids: the ATP-binding cassette (ABC) protein family, nitrate/peptide family (NPF), multidrug and toxic compound extrusion (MATE) family and purine permease (PUP) family[24]. According to our scRNA-seq analysis, 62 ABC, 27 NPF, 27 MATE and 13 PUP genes were expressed in *C. roseus* leaves (Supplementary Table 8 and Supplementary Fig. 5), among which 15 transporters were preferentially expressed in cell populations involved in MIA biosynthesis, including IPAP cells, EC1s and ICs. Phylogenetic and homology analyses further filtered out several transporters (Supplementary Figs. 6–8) because they were not clustered together in the phylogenetic trees with known transporters trafficking specialized metabolites or showing high similarity with *Arabidopsis* counterparts shuttling phytohormones or nutrient substances. Finally, with the exception of two characterized transporters (CrNPF2.9 and CrTPT2), two ABCs, two MATEs and four PUPs were selected as candidates for transporting MIAs or their intermediates in this study. No IPAP-specific transporter was identified among these candidates, and the efflux of loganic acid across the IPAP plasma membrane remains an enigma. Interestingly, the functionally characterized CrNPF2.6 transporter was proposed to mediate loganic acid uptake into ECs[25], and the *CrNPF2.6* gene showed high expression in both ICs and IPAP cells. Some NPF transporters are passive facilitators that export specialized metabolites along their chemical gradient[24]. Therefore, the role of CrNPF2.6 in IPAP cells should be further investigated. Two EC-specific transporters have been characterized. CrTPT2 from the ABC G family reportedly secreted catharanthine from ECs to the leaf surface[26], whereas CrNPF2.9 is localized in the tonoplast and exports strictosidine into the cytosol[27]. *CrABCG8*, *CrMATE1* and *CrPUP11* were the other three candidates expressed at high levels in the EC1 population, among which the *CrMATE1* gene was clustered together with *TDC* and *STR* in the *C. roseus* genome (Extended Data Fig. 6a). Thus, CrMATE1 was more likely to be a tonoplast-anchored secologanin importer. Because most ABC G transporters related to specialized metabolism were exporters localized at the plasma membrane, CrABCG8 was presumed to be a potential MIA exporter trafficking substrate across the plasma membrane. CrPUP11 and three IC-specific PUP transporters are homologues of *Arabidopsis* PUP3. These AtPUP3-like genes experienced obvious expansion by tandem duplication and formed a large cluster with 14 members in the *C. roseus* genome (Extended Data Fig. 6b). Interestingly,

https://doi.org/10.1038/s41477-022-01291-y

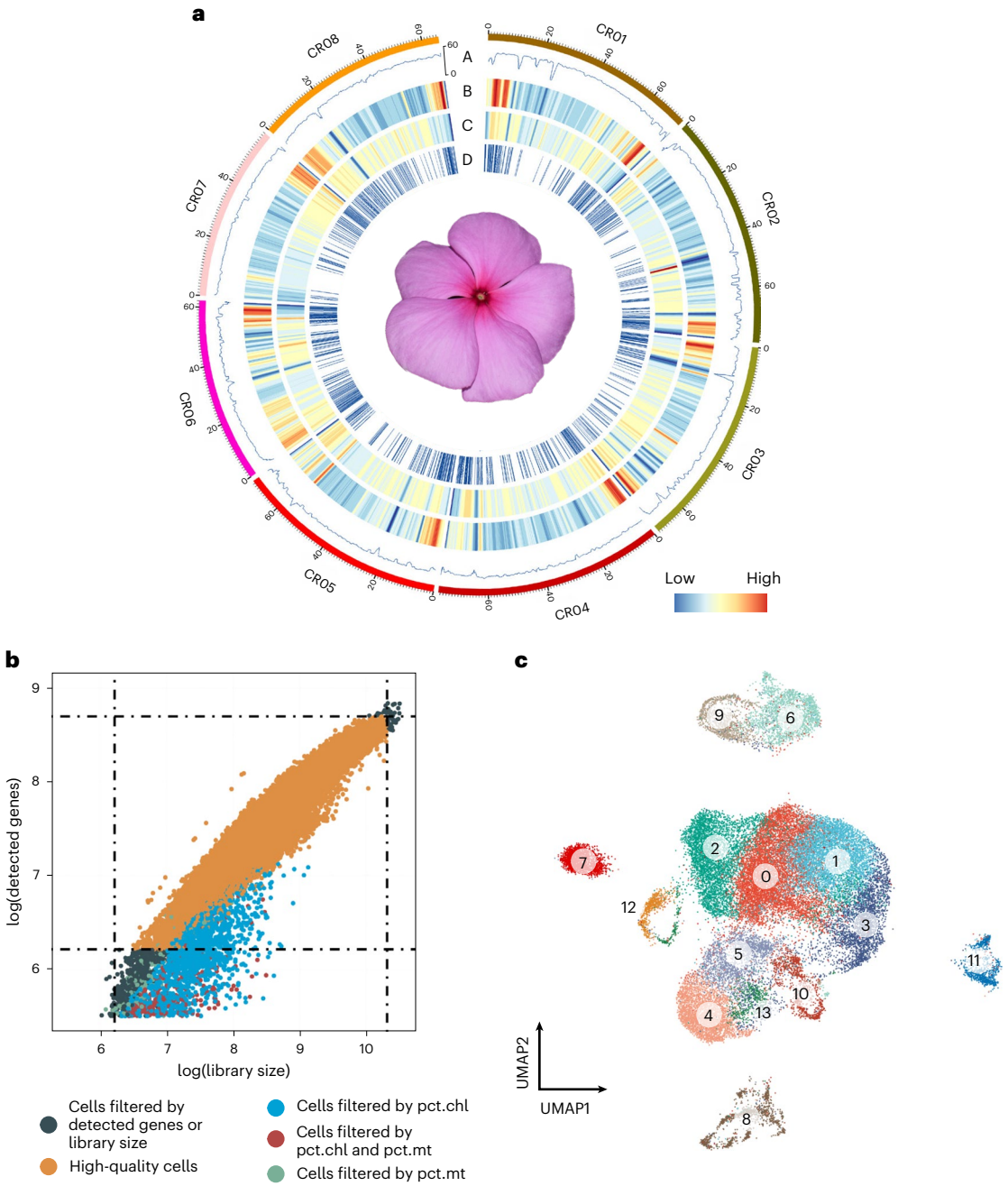



**Fig. 2 | Chromosomal distribution of genomic features and cell clustering.** **a**, Genome features of *C. roseus*: A, GC content; B, gene density; C, repeat density; D, tandem repeat density. **b**, Quality control of expression matrices. pct.chl, percentage of chloroplast genes; pct.mt, percentage of mitochondrial genes. Dashed lines show the thresholds of detected genes (500 and 6,000) and library size (500 and 30,000). Cells were marked as high-quality and retained if the number of detected genes ranged from 500 to 6,000, the library size was between 500 and 30,000 UMIs, the percentage of chloroplast genes was less than 40%, and the percentage of mitochondrial genes was less than 3%. **c**, UMAP visualization of cell clusters based on gene expression matrices of high-quality cells.

AtPUP3-like genes also experience expansion in opium poppy and evolve to generate a benzylisoquinoline alkaloid transporter through neofunctionalization and subfunctionalization[28], suggesting that the gene expansion mediated by tandem duplication played an important role in the origin of alkaloid transporters. Because all the characterized PUPs transporting specialized metabolites are importers localized in the plasma membrane, these PUP candidates were presumed to mediate MIA uptake in ICs and ECs. Two IC-specific MATE candidates were identified, of which CrMATE16 was a close neighbour of CjMATE1 in the phylogenetic tree. CjMATE1 is localized in tonoplasts and imports berberine into vacuoles in *Coptis japonica*[29]. Therefore, we propose that CrMATE16 might be a tonoplast-localized importer of MIAs in ICs. The identification of cell-specific candidates for trafficking MIA intermediates represents an important starting point for obtaining a better understanding of the MIA transport machinery and its roles in pathway compartmentation.

## Reconstruction of developmental trajectories

A common application of scRNA-seq to developing tissues is to reconstruct the developmental trajectories of individual cell populations

https://doi.org/10.1038/s41477-022-01291-y

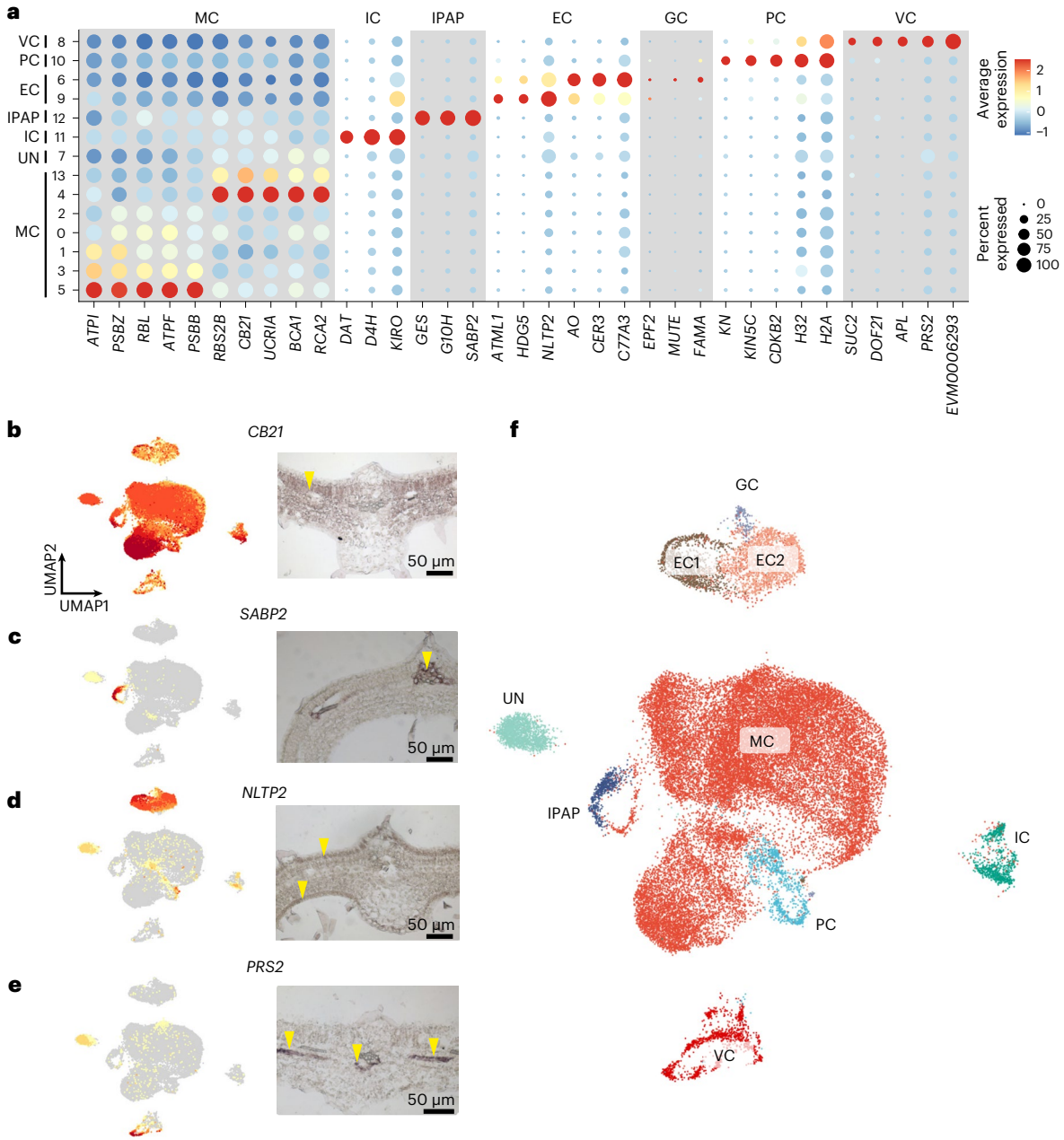



**Fig. 3 | Annotation of the *C. roseus* leaf cell types. a**, The expression patterns of representative cell type-specific marker genes. The dot diameter represents the proportion of cells expressing a given gene in each cluster, whereas the colour indicates the scaled average expression. The full names of the selected genes are given in Supplementary Table 5. **b**–**e**, UMAP visualization of the transcript accumulation of cell type-specific marker genes and RIH validating the annotated cell identities in *C. roseus* leaves. **b**, MC (*CB21*); **c**, IPAP (*SABP2*); **d**, EC (*NLTP2*); **e**, VC (*PRS2*). On the UMAP plot, the colour intensity represents the relative transcript expression level for the indicated gene in each cell. In the hybridized sections, the identified cell types are indicated by yellow arrows. **f**, Visualization of nine broad populations using UMAP: IPAP, EC, IC, MC, PC, UN, VC and GC.

through pseudotime analysis. Pseudotime analysis was applied to the EC population to further examine variation in the expression of MIA biosynthetic genes during EC development. The developmental trajectory started from cluster EC_2, followed by EC_3, EC_4 and EC_1, whereas EC_0 cells were distributed at the end of the trajectory according to the expression patterns of *AMTL1* and *CER3* along the trajectory (Fig. 5a–c). Genes predominantly expressed at the beginning (modules 2 and 4) of the developmental trajectory showed an overrepresentation of genes involved in meristem growth regulation and cell growth, as is expected for early ECs. At the end of the trajectory, module 1 captured the expression of genes involved in protein phosphorylation and the abiotic stress response, consistent with the hypersensitive nature of the mature epidermis to environmental stimuli (Fig. 5d and Supplementary Tables 9 and 10). Once we established the correct pseudotime trajectory, we used it to ascertain the dynamic expression of MIA genes during epidermal development. Interestingly, most EC-specific MIA genes were expressed at higher levels in younger ECs and continuously downregulated along the trajectory (Supplementary Fig. 9 and Extended Data Fig. 7), among which the expression of *TDC*, *CS*, *16OMT* and *T3O* decreased substantially and were rarely expressed at the late stage. However, the expression of genes that were expressed in multiple cell types, such as *LAMT*, *STR*, *PAS* and *DPAS*, decreased slowly along

https://doi.org/10.1038/s41477-022-01291-y

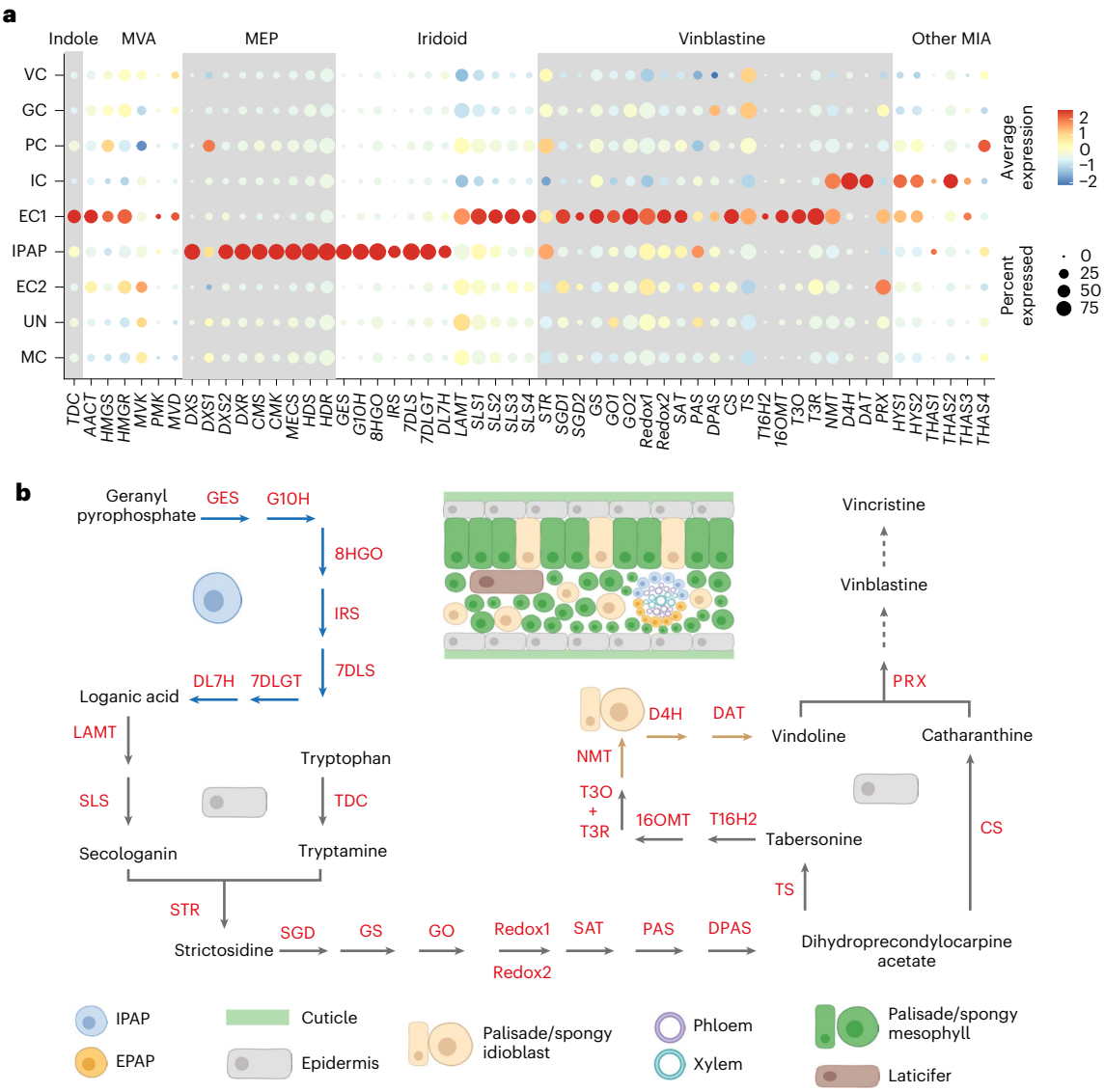



**Fig. 4 | Cell type-specific expression of genes involved in the MIA biosynthesis pathway and the spatial organization of MIA metabolism in *C. roseus* leaves.** **a**, Expression pattern of genes for MIA biosynthesis. The full names of the selected genes are given in Supplementary Table 7. **b**, Schematic of the cellular organization of MIA biosynthesis. The colour of the arrows indicates that the catalysing reactions are distributed in corresponding cell types. Schematics were created with BioRender.com.

the trajectory. Only *PRX* was expressed at slightly higher levels in the late stage of EC development. This phenomenon suggested that the activities of EC-specific MIA genes were tightly modulated by developmental cues.

The vasculature plays critical roles in the transport and biosynthesis of some plant alkaloids[30]. However, the development of the *C. roseus* vasculature remains largely unexamined. We constructed a putative developmental trajectory of VC cells to understand the cell differentiation states within the VC cluster of *C. roseus*. First, VC cells were reassigned into six subclusters (Fig. 5e), among which cluster 8_1 mainly comprised procambium/protophloem cells expressing orthologues to *Arabidopsis* identity markers, including *APL* and *DOF5.6*, whereas clusters VC_2 and VC_5 were termed companion cells (CCs), in which genes orthologous to *Arabidopsis* CC identity markers, including *SUC2*, *FTIP1*, *CEPR1*, *SCL15* and *Aux/IAA26* (refs. [31–34]), were expressed at high levels (Fig. 5f). A pseudotime trajectory based on the cells in clusters 8_1, 8_2, and 8_5 was used to visualize continuous differentiation trajectories for CCs from the procambium/protophloem to CCs (Fig. 5e,g). Gene Ontology (GO) enrichment analysis showed that the enriched gene signal at the beginning of the pseudotime axis (modules 4 and 2) was consistent with cell proliferation and the nucleosome assembly function of procambium/protophloem, whereas sugar uptake and ion transportation, as well as the energy necessary for CCs, were enriched at the end of the trajectory (modules 3 and 1) (Fig. 5h and Supplementary Tables 11 and 12). Pseudotime analysis provided novel insights into the dynamic process of CC development and variations in essential gene expression during the transition of cell states.

## Interspecies comparison of cell types

We compared the scRNA-seq data sets from *C. roseus* and *A. thaliana* leaves to explore the evolutionary conservation of dicotyledonous leaves. We first downloaded the published *A. thaliana* scRNA-seq data sets[31] and quantified them with Alevin. After cell clustering and cell type annotation, we obtained similar cell type clusters to those previously described, including MC, EC, VC, CC and GC clusters (Supplementary Fig. 10). Two distinct methods were used to profile the conserved and

https://doi.org/10.1038/s41477-022-01291-y

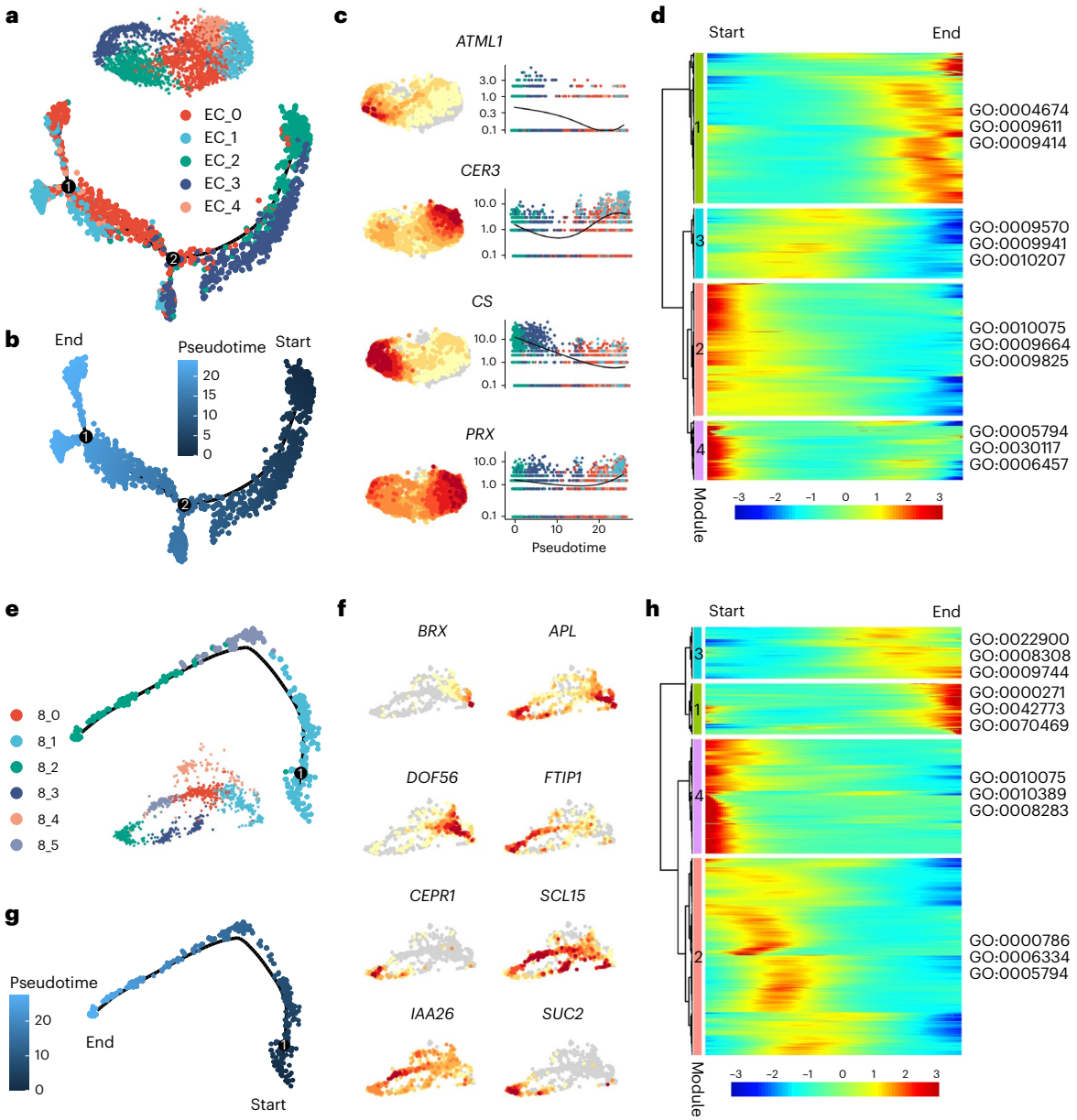



**Fig. 5 | Reconstruction of the developmental trajectories of ECs and VCs.** **a**, UMAP visualization and cluster distribution of EC subclusters along the pseudotime trajectory. Each dot indicates a single cell, and the colour of the cells indicates cell identities. **b**, Distribution of cells in the EC subclusters on the pseudotime trajectory. The colour on the dots indicates the pseudotime score. **c**, UMAP visualization of transcript accumulation and gene expression kinetics during pseudotime progression for representative genes. The colour of the cells indicates cell identities (right). **d**, Heatmap showing the expression of pseudotime-dependent genes over the pseudotime trajectory of EC subclusters. Representative GO terms of modules are labelled on the right. **e**, VC subcluster distribution along the pseudotime trajectory and UMAP visualization of subclusters. **f**, Visualization of the selected VC marker genes by the UMAP algorithm. Colour intensity indicates the relative transcript level for the indicated gene in each cell. **g**, Distribution of cells in the VC cluster on the pseudotime trajectory. **h**, Heatmap showing the expression of pseudotime-dependent genes over the pseudotime trajectory of VC subclusters. Representative GO terms of modules are labelled on the right. The full names of the selected genes are provided in Supplementary Table 13.

divergent cell types between two species: cross-species pairwise cluster correlation analysis (Fig. 6a) and cell clustering analysis with integrated data sets of the two species (Fig. 6b). The results from the two methods corroborate each other and support the same conclusion. For example, the EC clusters of the two species had the highest correlation with each other, and the two clusters were grouped together in the UMAP plot. Both results suggested molecular similarities and homologies between the corresponding clusters of *C. roseus* and *A. thaliana*. Interestingly, the *C. roseus* VC cluster showed the highest correlation with the *A. thaliana* CC cluster, supporting that the *C. roseus* VC cluster consisted mainly of CCs, partly because CCs were located at the outer layer of the vasculature and easily released from the tissue.

The PC cluster did not show a close relationship with any clusters in *A. thaliana*. We suspected that, as proliferating cells, PCs may exist in the shoot apex. Therefore, we integrated our data set with *A. thaliana* shoot apex and leaf data sets and then repeated the analytical procedure described above (Supplementary Fig. 11). The PC clusters of both species exhibited the strongest correlation in the correlation analysis (Fig. 6c) and overlapped with each other in the UMAP plot (Fig. 6d), suggesting that the PCs in *C. roseus* were more similar to those in the

https://doi.org/10.1038/s41477-022-01291-y

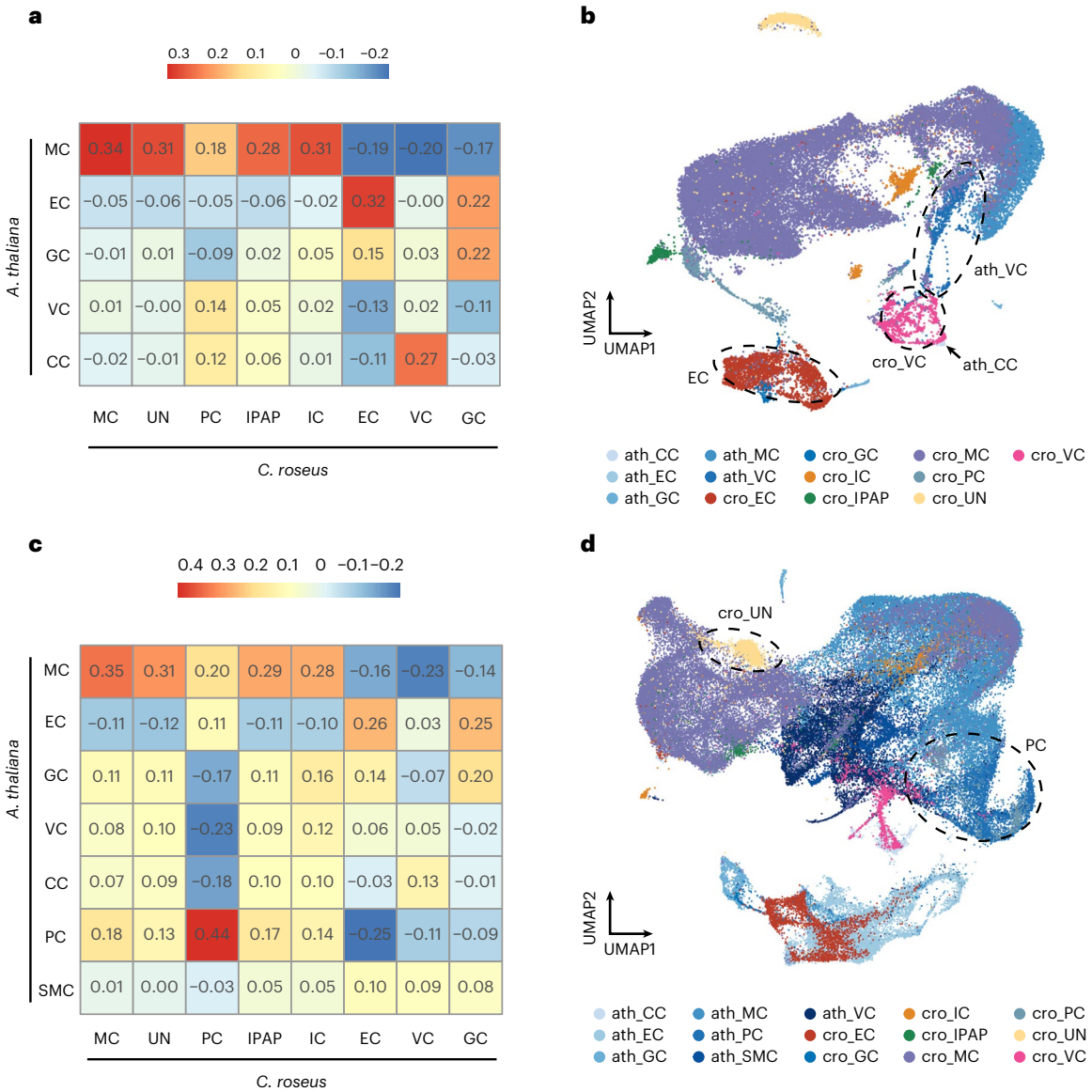



**Fig. 6 | Comparison of *C. roseus* and *A. thaliana* data sets at single-cell resolution. a**, Spearman's correlation coefficients between *C. roseus* and *A. thaliana* leaf cell clusters. **b**, UMAP visualization of the clusters from integrated leaf data sets; cells are coloured by clusters from different species. ath, *A. thaliana*; cro, *C. roseus*. **c**, Spearman's correlation coefficients between *C. roseus* (leaf) and *A. thaliana* (shoot apex and leaf) cell clusters. **d**, UMAP visualization of the clusters from integrated leaf and shoot apex data sets.

*A. thaliana* shoot apex. The UN cluster of *C. roseus* was strongly correlated with MCs (Fig. 6a,c) and adjacent to MCs in Fig. 6d, indicating that UNs were more likely to be a type of MC. Collectively, the interspecies comparison not only confirmed the evolutionary conservation of dicotyledonous leaves, but also validated the robustness of our cell clustering and annotation method.

## Discussion

scRNA-seq technology has revolutionized molecular and cellular biology by improving the spatiotemporal resolution of transcriptomic analysis to the level of the individual cell, which is rapidly expanding our ability to elucidate cell types, states, origins and differentiation[35]. Here, we built the first single-cell atlas of *C. roseus* leaves and filled the knowledge gap in the spatial organization of the MIA biosynthesis pathway. Compared with previous localization technologies, such as RIH and immunocytochemistry, scRNA-seq has a higher throughput, resolution and efficiency, and is able to determine the expression of thousands of genes at single-cell resolution in one experiment. This study represents the successful application of this method in elucidating multicellular compartmentation of primary or specialized metabolism in plants.

Compared with that in model plants, scRNA-seq research in *C. roseus* faces much greater challenges. Protoplast isolation is the first necessary step in the scRNA-seq procedure. Owing to their extremely long tubular shapes, LCs were not released from the leaf tissue, which meant that we could not obtain expression information for MIA genes in all cell types using scRNA-seq. Therefore, for these cell groups difficult to obtain by protoplasting, single-nucleus RNA-seq or spatial transcriptomic technology may be a better choice to investigate their gene expression profile at the single-cell level. In addition, we found that the recovery levels varied among different cell groups. As seen in Supplementary Table 14, the recovery of ECs was the lowest, mainly because ECs are easily lost during centrifugation, probably owing to a lower sedimentation coefficient. Many factors can change the percentage of cell groups during protoplasting, such as cell size, cell shape,

the relative position of cells within tissues (surface or inner layer), the biochemical compositions of the cell wall and the developmental stage of the cells. However, as described in our study and previous reports[36], the variation in the percentage of each cell group introduced by protoplasting cannot be avoided but has little effect on the downstream analysis, as long as each cell group contains a sufficient number of cells.

Cell clustering and annotation are the next crucial steps in scRNA-seq. Three biological replicates were employed to diminish the bias introduced by random biological variation. Partial inconsistency in cell clustering and pathway gene expression between the replicates indicated that replicates were necessary for scRNA-seq. By integrating the data sets of three replicates and correcting batch effects, we obtained better cell clustering and annotation, as well as MIA gene expression profiles that were more consistent with previously reported RIH experiments[10–12] (Supplementary Figs. 12 and 13, Supplementary Table 15 and Supplementary Data 2). Finally, all detected cells were divided into nine broad populations, and the cell identities were validated by performing RIH experiments and an interspecies comparative analysis of *C. roseus* and *A. thaliana* single-cell data sets. MCs consisted of seven clusters, showing the highest heterogeneity among all populations. According to the expression of proliferation-related genes in MCs, such as *H32* and *H2A*, cells in clusters 0–3 seemed much younger than those in other clusters (Supplementary Fig. 4). In addition, chloroplast-related genes were expressed at high levels in clusters 4, 5 and 13, consistent with the high level of photosynthesis in mature leaves. Interestingly, cells in cluster 5 contained a high percentage of mitochondrial genes and the highest expression of chloroplast-related genes, which may represent most photosynthetically active cells with mitochondrial proliferation to support photorespiration (Supplementary Figs. 14 and 15). All these phenomena together suggested that cells in clusters 4, 5 and 13 may be more mature than other MCs. Therefore, we proposed that the differences in the maturity of cells represented the major contribution to MC heterogeneity in developing leaves, which was also mentioned in previous research[37].

Cell cluster correlation analysis supplied additional information on the relationship of different clusters (Supplementary Fig. 16). IC and IPAP populations were strongly correlated with MCs and expressed chloroplast-related genes at high levels, implying their possible origination from MCs. However, IPAP cells appear to be a part of the vasculature in the anatomical structure of *C. roseus* leaves; therefore, their origin remains unknown. In addition, we estimated that the UN population may represent a group of cells at a specific development stage. All anatomically known cell types in *C. roseus* leaves were discovered in our scRNA-seq data, except for LCs that could not be protoplasted. Somatic polyploidy through endoreduplication generates a considerable number of multinuclear cells in developing leaves. Therefore, we cannot exclude the possibility that the UN population might be composed of some polyploid cells.

By quantifying the expression of all the characterized MIA pathway genes in different cell types, we provided an initial high-resolution picture of the spatial organization of MIA biosynthesis in *C. roseus* leaves. Among all genes in the pathway, the cell type-specific expression of 20 genes was investigated for the first time. Our results for these genes whose transcript or protein locations have been investigated using RIH or immunocytochemical experiments are consistent with previous reports. The scRNA-seq analysis suggested that most of the genes in the MIA pathway are preferentially expressed in younger tissues, consistent with a basipetal gradient of the expression pattern of these genes in *C. roseus* leaves detected using RIH (Extended Data Fig. 7) because the leaf tissue gradually matures from base to tip[10,38]. However, the localization of MIA biosynthetic enzymes inferred by the scRNA-seq analysis and RIH experiments was not completely consistent with the localization of the corresponding products estimated by imaging MS or a single-cell MS analysis. For instance, *CS*, *TS* and *16OMT* were mainly expressed in ECs, whereas their products, catharanthine, tabersonine and 16-methoxytabersonine, were detected in ICs, suggesting that these products experienced a translocation from ECs to ICs.

Because of the highly complex architecture of the MIA pathway, multiple intermediate translocation processes are predicted to be essential for the biosynthesis of end products. Transporters are presumed to play central roles in the shuttling of MIA intermediates between different inter- and intracellular compartments, representing key switches connecting sequential metabolic modules. However, the MIA transporters remain largely unknown. In this study, eight cell-specific candidate transporters involved in MIA translocation were screened out, and their further functional identification will greatly improve our knowledge about the compartmentation of MIA biosynthesis and its roles in plant development and survival. If we can fully understand the higher-order organization of MIA biosynthesis, metabolic flux and MIA production may be maximized and increased in *C. roseus* or heterologous hosts through metabolic engineering and synthetic biology, ultimately facilitating the research and development of MIA-derived drugs.

## Methods

### Plant material and protoplast isolation

*Catharanthus roseus* (L.) G. Don plants were grown at 25 °C in a growth chamber under a 16:8 h light/dark photoperiod. Bright-field microscopy and epifluorescence microscopy were performed using a Leica M205FA microscope. A UV longpass filter was used as the epifluorescence filter for the observation of ICs and LCs. Leaf protoplasts were isolated as described by Carqueijeiro et al.[39]. Briefly, 0.15 g of young healthy leaves (1.8–2.0 cm long) was harvested and digested for 2 h at room temperature in an enzyme solution (1% cellulase R-10, 0.15% macerozyme R-10, 0.45 M mannitol and 20 mM MES, pH 5.6–5.8). After passing through 70- and 40-μm strainers, protoplasts were centrifuged at 100*g* for 7 min and washed once with protoplasting solution without enzymes. Protoplast viability was determined by trypan blue (C0040, Solarbio) and fluorescein diacetate (IF0160, Solarbio) staining, and cells were counted with a haemocytometer. The final suspension volume was adjusted to a density of 1,500 to 1,800 cells $\mu l^{-1}$. Protoplasts were placed on ice until further processing.

### scRNA-seq library preparation and sequencing

The scRNA-seq libraries were prepared with the Chromium Single Cell 3′ Gel Beads-in-emulsion (GEM) Library and Gel Bead Kit v.3 (16 rxns PN-1000268, 10x Genomics) according to the user's manual supplied with the kit. Sequencing was performed with an Illumina HiSeq 4000 according to the manufacturer's instructions.

### Genome assembly and gene model prediction

Contig-level assembly was performed on the full set of PacBio clean reads using Canu (v.1.5, https://github.com/marbl/canu). Single molecules labelled with DLE1 (DNA motif CTTAAG) were assembled into BioNano optical maps using the BioNano Solve pipeline (v.3.5, BioNano Genomics), and contigs were hybrid scaffolded into superscaffolds based on the BioNano optical maps. The final genome assembly with better assembly continuity was obtained by mapping the paired-end Hi-C reads onto the superscaffolds using BWA (v.0.7.17, https://github.com/lh3/bwa) with the default parameters. Finally, pseudochromosomes were constructed from superscaffolds using LACHESIS (https://github.com/shendurelab/LACHESIS) with tuned parameters, and the chromosome number was set to eight based on karyotype[40]. The completeness of assembly was evaluated using BUSCO (v.2.0, https://busco-archive.ezlab.org/v2) with Embryophyta odb9 as a reference. In addition, Illumina shotgun paired-end reads were mapped to the pseudochromosomes using BWA (v.0.7.17) with the default parameters. For the gene model prediction, three ab initio gene prediction tools were used: AUGUSTUS (v.3.1, https://github.com/Gaius-Augustus/Augustus), SNAP (v.2006, https://github.com/KorfLab/SNAP) and GlimmerHMM (v.1.2, https://ccb.jhu.edu/software/glimmerhmm).

Homology-based prediction was performed using GeMoMa (v.1.3.1, www.jstacs.de/index.php/GeMoMa). The PASA pipeline (v.2.0.2, https://github.com/PASApipeline/PASApipeline) was used to generate the RNA-seq prediction. Finally, we employed EVM (v.1.1.1, http://evidencemodeler.github.io) to combine all the evidence types into a single set of gene structure annotations.

### Preprocessing of scRNA-seq data

For the Cell Ranger pipeline, Cell Ranger (v.6.0.0, 10x Genomics) was used to map reads to the genome (BioProject: PRJNA841429) using the STAR aligner, and UMIs were counted to construct digital expression matrices. For the Alevin pipeline, Salmon (v.1.4) (ref. [16]) was used to map reads to the genome and construct matrices. Using the R package Seurat[41] (v.4.0.3), we first merged three biological replicates and removed low-quality cells based on the criteria of gene number (500–6,000), UMI count (500–30,000), and percentages of mitochondrial (3%) and chloroplast genes (40%). Doublets in the merged scRNA-seq data set were filtered out with DoubletFinder (v.2.0.3). Data were normalized using the NormalizeData function in Seurat[41] with the LogNormalize method and a scaling factor of 10,000. Highly variable genes were detected using the FindVariableGenes function with the 'vst' method and 2,000 features. The batch effect was eliminated with the RunHarmony function. The cell-cycle score was calculated by the CellCycleScoring function in Seurat, and the variations due to the cell cycle, mitochondrial genes, chloroplast genes and library size were removed by regression using the ScaleData function. The genes used for cell-cycle scoring are listed in Supplementary Table 16.

### Cell clustering and annotation

PCA was performed to reduce the dimensionality with the RunPCA function in Seurat[41]. Graph-based clustering was performed to cluster cells according to their gene expression profile using the FindClusters function. The EC population was divided into subclusters using the FindSubCluster function. Cells were visualized using two- and three-dimensional UMAP algorithms with the RunUMAP function. We used the FindAllMarkers function to identify cell type-specific genes. For a particular cluster, FindAllMarkers identified positive markers compared with all other cells, where pct.2 < 0.1 and logfc.threshold = 0.25 were set as the thresholds for significantly different expression. GO enrichment analyses for cell type-specific genes were performed using ClusterProfiler (v.3.18.1) (ref. [42]). Feature plots of marker genes were drawn using the FeaturePlot function, and the gene expression matrix for drawing the feature plot was imputed by MAGIC software (v.3.0.0)[43]. The cluster-averaged expression was extracted using the AverageExpression function in Seurat, and Spearman's correlation coefficients were calculated using the cor function with the 'spearman' method in R. The correlation heatmap was drawn using the pheatmap package (v.1.0.12, https://CRAN.R-project.org/package=pheatmap).

### Pseudotime analysis

Pseudotime trajectory analysis was performed using the Monocle[44] R package (v.2.8.0). The log-normalized data from the Seurat object were imported into Monocle using the as.CellDataSet function. Cells were ordered along the trajectory and visualized in a reduced dimensional space. The root of the trajectory was chosen according to cell subcluster identities. Genes that changed significantly along the pseudotime were identified using the differentialGeneTest function with a $q$ value < 0.1 and clustered using plot_pseudotime_heatmap functions with the default parameters. For EC clusters, genes that were dynamically expressed along pseudotime were visualized using the plot_genes_in_pseudotime function.

### Interspecies scRNA-seq data comparison

The scRNA-seq data sets of *A. thaliana* (BioProject: PRJCA003094) were downloaded from the Beijing Institute of Genomics Data Center (http://bigd.big.ac.cn) and quantified with Alevin. Cell clustering and annotation were performed as described previously[31]. A set of one-to-one homologous genes between *C. roseus* and *A. thaliana* defined by the reciprocal best hit method[45] was used for interspecies scRNA-seq analysis (Supplementary Table 17). Two methods were used to compare the difference between cell types in *C. roseus* and *A. thaliana*: cross-species pairwise cluster correlation analysis and cell clustering analysis with integrated data sets. In the former analysis, the correlation coefficients between cell populations of *C. roseus* and *A. thaliana* were calculated as described previously[45]. Briefly, the average gene expression level of each cell type was calculated using the AverageExpression function in Seurat[41]. Gene expression matrices were transformed into gene specificity matrices based on the gene specificity index equation before calculating pairwise cell type correlations. The resulting gene specificity matrices were used to calculate pairwise Spearman rank order correlations. For the latter analysis, if more than one *A. thaliana* data set was used, an integrated *A. thaliana* cell atlas was constructed by merging downloaded data sets and performing batch effect correction using the Harmony algorithm. If only one *A. thaliana* data set was used, we bypassed the integration and correction steps and directly performed the subsequent procedures. After cell clustering and annotation, FindIntegrationAnchors and IntegrateData were used to integrate the data sets of *C. roseus* and *A. thaliana*. PCA and UMAP were then performed to visualize the clusters.

### Phylogenetic analysis

Protein sequences of MATE, PUP and NPF transporters from *C. roseus* and *A. thaliana* were aligned with functionally characterized alkaloid-transport related transporters of these three families. The alignment was performed using ClustalW with default settings, and the neighbour-joining trees were constructed using Geneious (v.2022, https://www.geneious.com).

### Bulk RNA-seq analysis

The same pooled protoplasts and intact leaves used for scRNA-seq were subjected to bulk RNA-seq analysis. Sequencing libraries were constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB #E7770, NEB) and then sequenced on the HiSeq 4000 platform. The clean reads were aligned to the *C. roseus* genome and quantified using Salmon (v.1.4)[16]. Spearman's correlation coefficients for mean expression across leaf and protoplast replicates were calculated.

### Real-time PCR

Total RNA was isolated using an RNAprep Plant Kit (DP441, TIANGEN). The qualitative and quantitative estimation of the total extracted RNA was determined using agarose gel electrophoresis and a NanoDrop spectrophotometer (Thermo Fisher Scientific). First-strand complementary DNA was synthesized according to the manufacturer's protocol using a GoScript Reverse Transcription System Kit (Promega, catalogue no. A5001). Primers were designed using Primer Premier (v.6.0) and are listed in Supplementary Table 18. PCR amplification was performed under the following conditions: 95 °C for 30 s; 40 cycles of 95 °C for 5 s, 60 °C for 30 s, 72 °C for 15 s and 95 °C for 10 s. Products were verified by melting curve analysis. The relative transcript expression levels of 29 MIA pathway genes were determined using the $2^{-\Delta\Delta Ct}$ method. The cycle threshold value for each gene was normalized against the cycle threshold value from the *C. roseus GAPDH* gene.

### RNA in situ hybridization

The unique fragments of selected genes were cloned into the pGEM-T easy vector (Promega, catalogue no. A1360). The primers for amplification are listed in Supplementary Table 19. For RNA probe synthesis, linearized vectors were added as templates. In vitro transcription was performed with a DIG RNA Labeling Kit (Sp6/T7) (Roche, catalogue no. 11175025910). Tissue embedding and RIH were essentially performed as

described in a previous study[10], with a few modifications. Briefly, leaves in the same growth period as those used to prepare protoplasts were fixed with formaldehyde-acetic acid-ethanol fixative (50%) at 4 °C for 24 h and embedded manually. Paraffin-embedded samples were sectioned to a thickness of 10 µm with a sliding microtome (Leica). The sections were spread on 3-aminopropyltriethoxysilane (AES)-coated slides (WHITE 12-550-15, Thermo Fisher Scientific) overnight at 40 °C, and paraffin was removed using xylene (twice for 15 min) before rehydration in an ethanol gradient up to diethypyrocarbonate (DEPC)-treated water. The rehydrated sections were digested with Proteinase K (Sigma, catalogue no. P2308), dehydrated with a gradient series of ethanol solutions and hybridized with RNA probes. After washing, sections were incubated with anti-digoxigenin-AP Fab fragments (Roche, catalogue no. 11093274910). For colour development, sections were immersed in Milli-Q water in which an NBT/BCIP tablet (Sigma, catalogue no. B5655) was dissolved at room temperature until the target colour was clear. After development, sections were washed with TE buffer, dried, mounted in immersion oil under cover slips, and observed under a microscope (Zeiss).

### Reporting summary

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

## Data availability

The single-cell and bulk RNA sequencing data generated in this study have been deposited to NCBI with the accession number PRJNA759937. The scRNA-seq data sets of *A. thaliana* were downloaded from the Beijing Institute of Genomics Data Center (http://bigd.big.ac.cn) with the accession number PRJCA003094. The genome used in this study has been deposited to NCBI with the accession number PRJNA841429. The quantification results have been deposited to figshare with https://doi.org/10.6084/m9.figshare.20255094.

## References


1. Brown, S., Clastre, M., Courdavault, V. & O'Connor, S. E. De novo production of the plant-derived alkaloid strictosidine in yeast. *Proc. Natl Acad. Sci. USA* **112**, 3205–3210 (2015).
2. Pan, Q., Mustafa, N. R., Tang, K., Choi, Y. H. & Verpoorte, R. Monoterpenoid indole alkaloids biosynthesis and its regulation in *Catharanthus roseus*: a literature review from genes to metabolites. *Phytochem. Rev.* **15**, 221–250 (2016).
3. Facchini, P. J. & De Luca, V. Opium poppy and Madagascar periwinkle: model non-model systems to investigate alkaloid biosynthesis in plants. *Plant J.* **54**, 763–784 (2008).
4. Zhu, X., Zeng, X., Sun, C. & Chen, S. Biosynthetic pathway of terpenoid indole alkaloids in *Catharanthus roseus*. *Front. Med.* **8**, 285–293 (2014).
5. Courdavault, V. et al. A look inside an alkaloid multisite plant: the *Catharanthus* logistics. *Curr. Opin. Plant Biol.* **19**, 43–50 (2014).
6. Caputi, L. et al. Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science* **360**, 1235–1239 (2018).
7. Miettinen, K. et al. The seco-iridoid pathway from *Catharanthus roseus*. *Nat. Commun.* **5**, 3606 (2014).
8. Qu, Y. et al. Completion of the seven-step pathway from tabersonine to the anticancer drug precursor vindoline and its assembly in yeast. *Proc. Natl Acad. Sci. USA* **112**, 6224–6229 (2015).
9. Costa, M. M. et al. Molecular cloning and characterization of a vacuolar class III peroxidase involved in the metabolism of anticancer alkaloids in *Catharanthus roseus*. *Plant Physiol.* **146**, 403–417 (2008).
10. St-Pierre, B., Vazquez-Flota, F. A. & De Luca, V. Multicellular compartmentation of *Catharanthus roseus* alkaloid biosynthesis predicts intercellular translocation of a pathway intermediate. *Plant Cell* **11**, 887–900 (1999).
11. Irmler, S. et al. Indole alkaloid biosynthesis in *Catharanthus roseus*: new enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. *Plant J.* **24**, 797–804 (2008).
12. Burlat, V., Oudin, A., Courtois, M., Rideau, M. & St-Pierre, B. Co-expression of three MEP pathway genes and geraniol 10-hydroxylase in internal phloem parenchyma of *Catharanthus roseus* implicates multicellular translocation of intermediates during the biosynthesis of monoterpene indole alkaloids and isoprenoid-derived primary metabolites. *Plant J.* **38**, 131–141 (2004).
13. Yamamoto, K. et al. Cell-specific localization of alkaloids in *Catharanthus roseus* stem tissue measured with Imaging MS and Single-cell MS. *Proc. Natl Acad. Sci. USA* **113**, 3891–3896 (2016).
14. Yamamoto, K. et al. The complexity of intercellular localisation of alkaloids revealed by single-cell metabolomics. *New Phytol.* **224**, 848–859 (2019).
15. Mahroug, S., Burlat, V. & St-Pierre, B. Cellular and sub-cellular organisation of the monoterpenoid indole alkaloid pathway in *Catharanthus roseus*. *Phytochem. Rev.* **6**, 363–381 (2007).
16. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol.* **20**, 65 (2019).
17. Mizutani, M. & Ohta, D. Diversification of P450 genes during land plant evolution. *Annu. Rev. Plant Biol.* **61**, 291–315 (2010).
18. Wilson, A. E. & Tian, L. Phylogenomic analysis of UDP-dependent glycosyltransferases provides insights into the evolutionary landscape of glycosylation in plant metabolism. *Plant J.* **100**, 1273–1288 (2019).
19. Lu, P., Porat, R., Nadeau, J. A. & O'Neill, S. D. Identification of a meristem L1 layer-specific gene in *Arabidopsis* that is expressed during embryonic pattern formation and defines a new class of homeobox genes. *Plant Cell* **8**, 2155–2168 (1996).
20. Bernard, A. et al. Reconstitution of plant alkane biosynthesis in yeast demonstrates that *Arabidopsis* ECERIFERUM1 and ECERIFERUM3 are core components of a very-long-chain alkane synthesis complex. *Plant Cell* **24**, 3106–3118 (2012).
21. Vranova, E., Coman, D. & Gruissem, W. Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annu. Rev. Plant Biol.* **64**, 665–700 (2013).
22. Oudin, A. et al. Spatial distribution and hormonal regulation of gene products from methyl erythritol phosphate and monoterpene–secoiridoid pathways in *Catharanthus roseus*. *Plant Mol. Biol.* **65**, 13–30 (2007).
23. Guirimand, G. et al. Cellular and subcellular compartmentation of the 2*C*-methyl-D-erythritol 4-phosphate pathway in the Madagascar periwinkle. *Plants* **9**, 462 (2020).
24. Halkier, B. A. & Xu, D. The ins and outs of transporters at plasma membrane and tonoplast in plant specialized metabolism. *Nat. Prod. Rep.* **39**, 1483–1491 (2022).
25. Larsen, B. et al. Identification of iridoid glucoside transporters in *Catharanthus roseus*. *Plant Cell Physiol.* **58**, 1507–1518 (2017).
26. Yu, F. & De Luca, V. ATP-binding cassette transporter controls leaf surface secretion of anticancer drug components in *Catharanthus roseus*. *Proc. Natl Acad. Sci. USA* **110**, 15830–15835 (2013).
27. Payne, R. M. et al. An NPF transporter exports a central monoterpene indole alkaloid intermediate from the vacuole. *Nat. Plants* **3**, 16208 (2017).
28. Dastmalchi, M. et al. Purine permease-type benzylisoquinoline alkaloid transporters in opium poppy. *Plant Physiol.* **181**, 916–933 (2019).
29. Takanashi, K. et al. A multidrug and toxic compound extrusion transporter mediates berberine accumulation into vacuoles in *Coptis japonica*. *Phytochemistry* **138**, 76–82 (2017).

30. Ozber, N. & Facchini, P. J. Phloem-specific localization of benzylisoquinoline alkaloid metabolism in opium poppy. *J. Plant Physiol.* **271**, 153641 (2022).
31. Zhang, T. Q., Chen, Y. & Wang, J. W. A single-cell analysis of the *Arabidopsis* vegetative shoot apex. *Dev. Cell* **56**, 1056–1074 e1058 (2021).
32. Kim, J. Y. et al. Distinct identities of leaf phloem cells revealed by single cell transcriptomics. *Plant Cell* **33**, 511–530 (2021).
33. Rodriguez-Villalon, A. Wiring a plant: genetic networks for phloem formation in *Arabidopsis thaliana* roots. *New Phytol.* **210**, 45–50 (2016).
34. Otero, S. & Helariutta, Y. Companion cells: a diamond in the rough. *J. Exp. Bot.* **68**, 71–78 (2017).
35. Seyfferth, C. et al. Advances and opportunities in single-cell transcriptomics for plant research. *Annu. Rev. Plant Biol.* **72**, 847–866 (2021).
36. Shaw, R., Tian, X. & Xu, J. Single-cell transcriptome analysis in plants: advances and challenges. *Mol. Plant* **14**, 115–126 (2021).
37. Tenorio Berrio, R. et al. Single-cell transcriptomics sheds light on the identity and metabolism of developing leaf cells. *Plant Physiol.* **188**, 898–918 (2022).
38. Dale, J. E. The control of leaf expansion. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **39**, 267–295 (1988).
39. Carqueijeiro, I. et al. Isolation of vacuoles from the leaves of the medicinal plant *Catharanthus roseus*. *Methods Mol. Biol.* **1789**, 81–99 (2018).
40. Guimaraes, G. et al. Cytogenetic characterization and genome size of the medicinal plant *Catharanthus roseus* (L.) G. Don. *AoB Plants* **2012**, pls002 (2012).
41. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
42. Yu, G., Wang, L., Han, Y. & He, Q. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
43. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 e727 (2018).
44. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
45. Tosches, M. A. et al. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).

## Acknowledgements

This work was supported by the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (grant no. 2021-I2M-1-032). The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Author contributions

C.S., J.W. and S.C. conceived and designed the project. Yi Li, X.S., S.W., R.L. and H.Z. performed the experiments. S.S., Ying Li, J.X. and G.S. analysed the data. Yi Li, S.S. and X.S. wrote the manuscript draft. C.S., B.S.-P., B.G., J.W. and S.C. revised the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

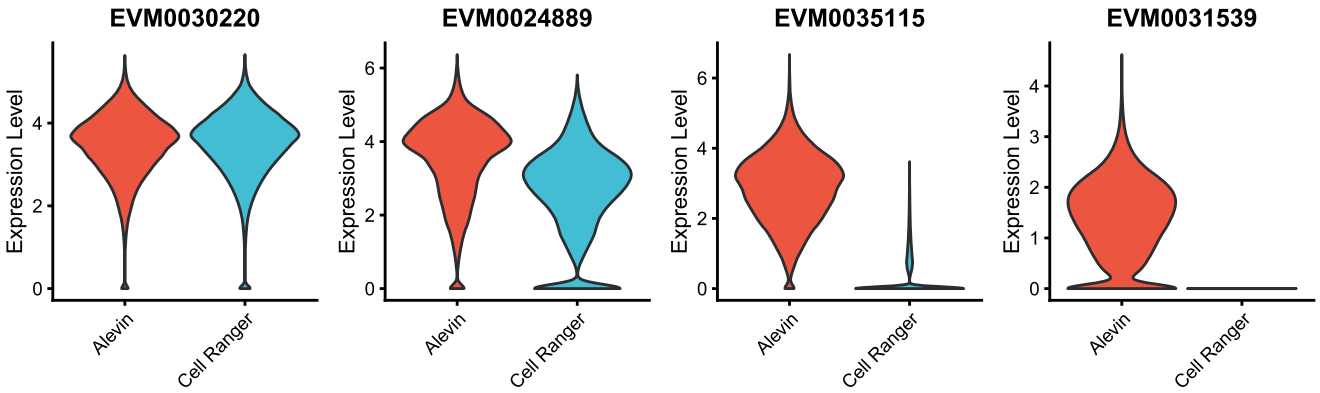
**Extended data** is available for this paper at https://doi.org/10.1038/s41477-022-01291-y.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41477-022-01291-y.

**Correspondence and requests for materials** should be addressed to Benoit St-Pierre, Shilin Chen or Chao Sun.

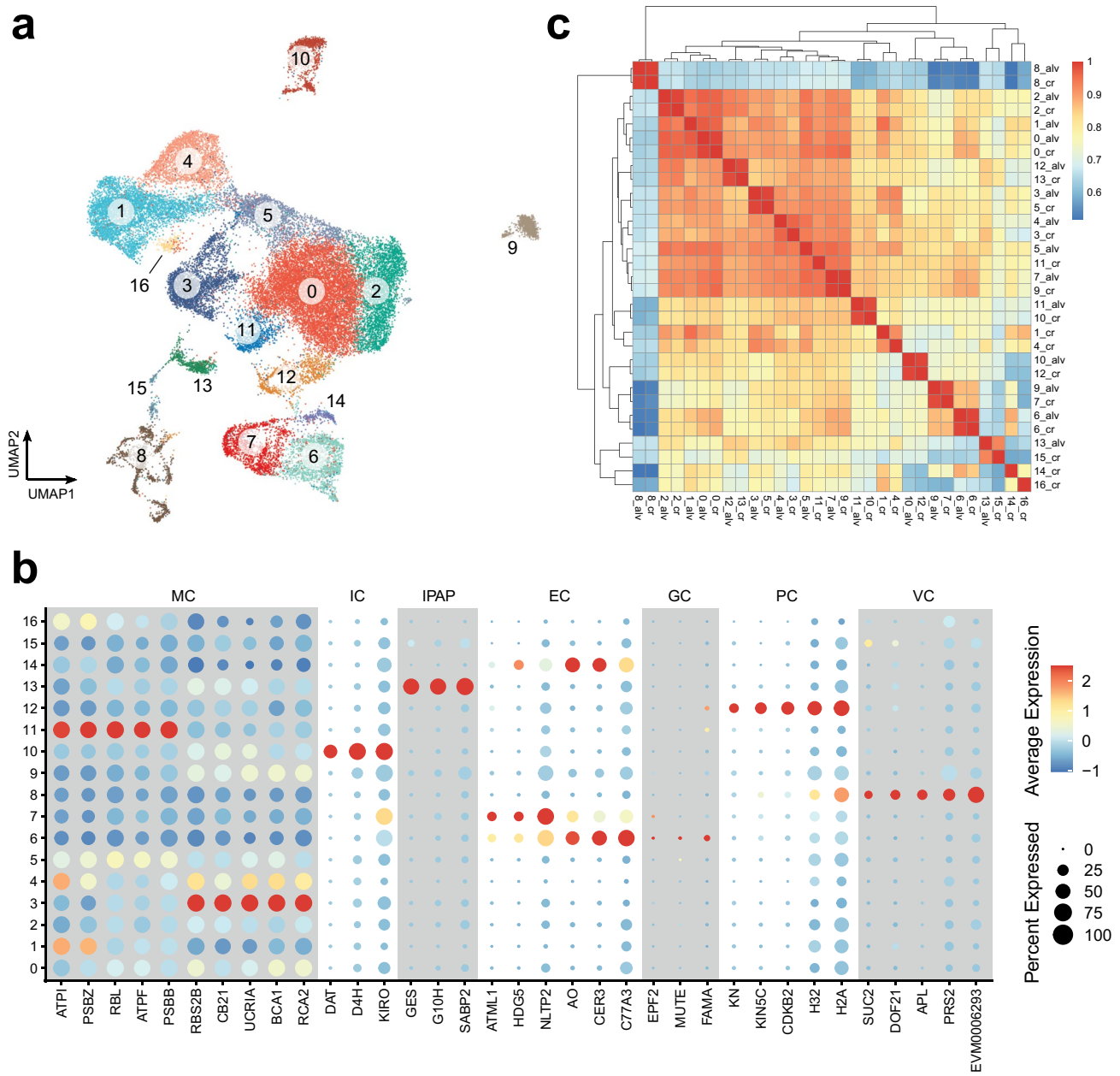
**Peer review information** *Nature Plants* thanks Tetsuro Mimura, Silin Zhong, Tomáš Pluskal, Kenneth Birnbaum and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41477-022-01291-y



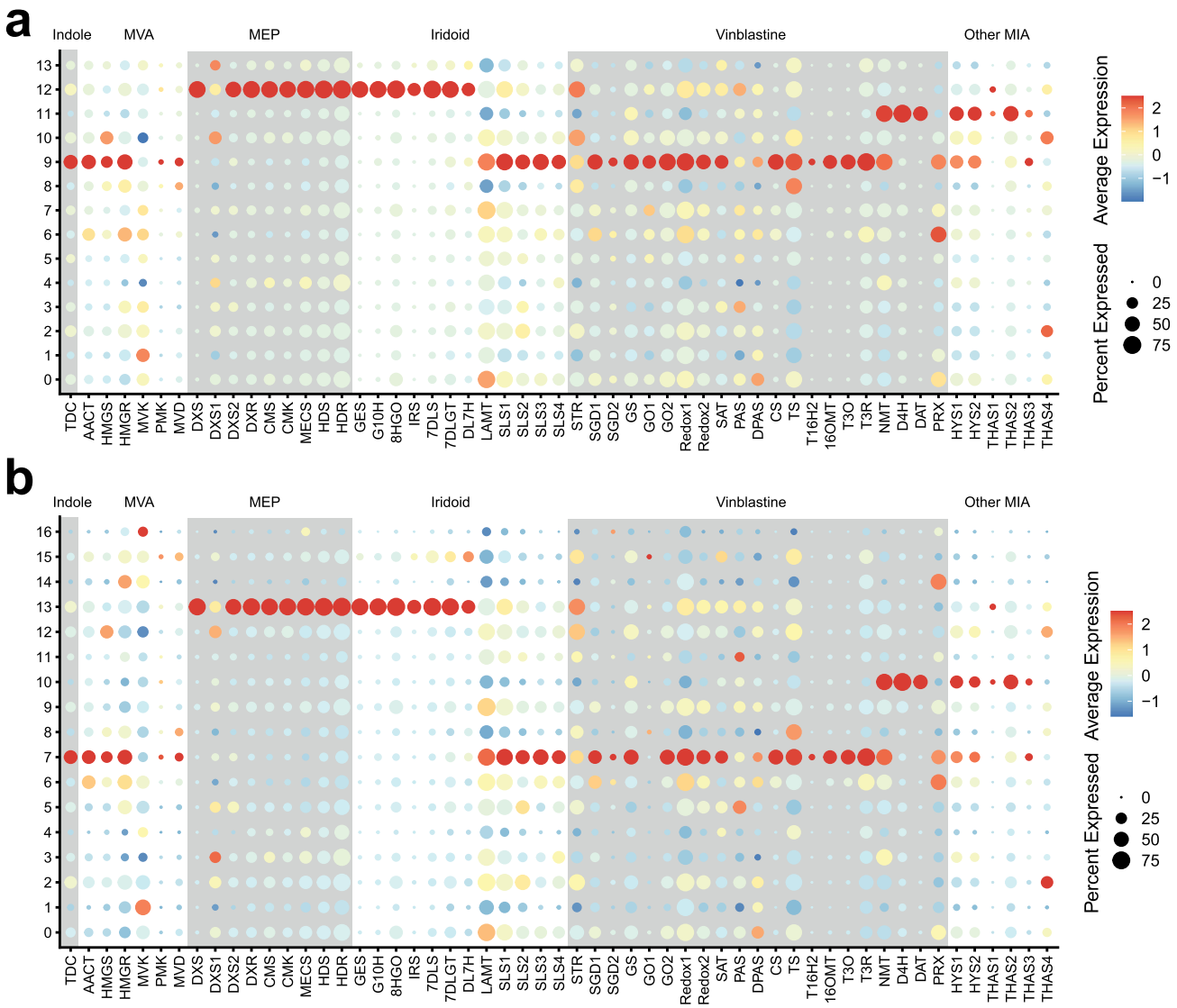

**Extended Data Fig. 1 | The expression of genes quantified by Alevin and Cell Ranger.** *EVM0030220*, which was unique among genes, had similar expression, as estimated by Alevin or Cell Ranger. *EVM0024889*, which shared 77% identity with *EVM0001360*, showed slightly diminished expression using Cell Ranger compared with that obtained using Alevin. *EVM0035115*, which shared 99.00% identity with *EVM0010963*, exhibited remarkably reduced expression with Cell Ranger, and the expression of *EVM0031539*, which was 100% identical to *EVM0012710*, was relatively high according to Alevin but was quantified close to zero using Cell Ranger.

https://doi.org/10.1038/s41477-022-01291-y



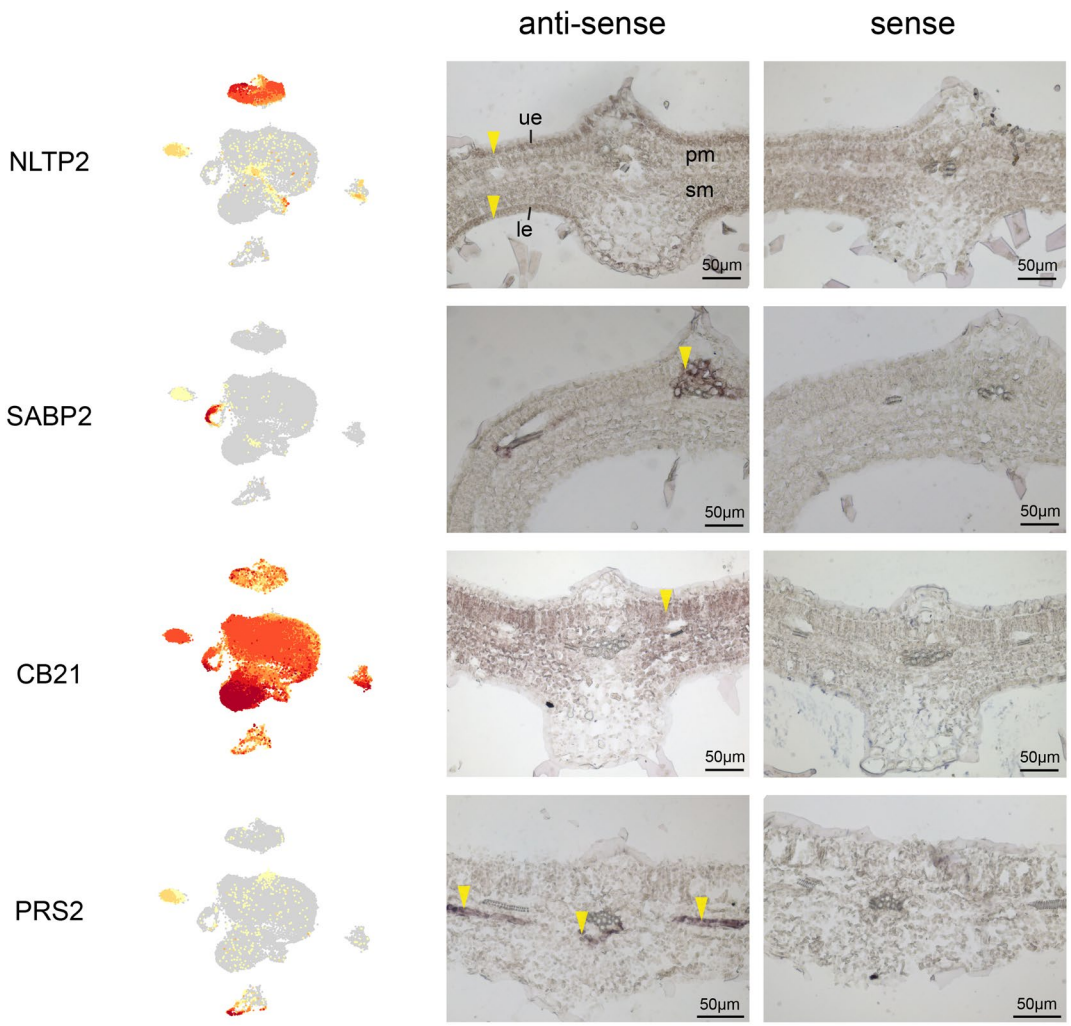

**Extended Data Fig. 2 | Cell cluster assignment of Cell Ranger expression matrices and correlation of cell cluster expression patterns derived from different quantification methods. a**, UMAP visualization of cell clusters based on gene expression matrices of high-quality cells. **b**, The expression patterns of representative cell type-specific marker genes. The dot diameter represents the proportion of cells expressing a particular gene in each cluster, whereas the color indicates the scaled average expression. The full names of selected genes are provided in Supplementary Table 5. **c**, Heatmap showing Spearman's correlation between clusters from two quantification pipelines: Alevin (alv) and Cell Ranger (cr). IPAP: 12_alv/13_cr; VC: 8_alv/8_cr; IC: 11_alv/10_cr; PC: 10_alv/12_cr; UN: 7_alv/9_cr; MC: 0–5, 13_alv/0–5, 11, 15, 16_cr; EC: 6, 9_alv/6, 7, 14_cr.

https://doi.org/10.1038/s41477-022-01291-y



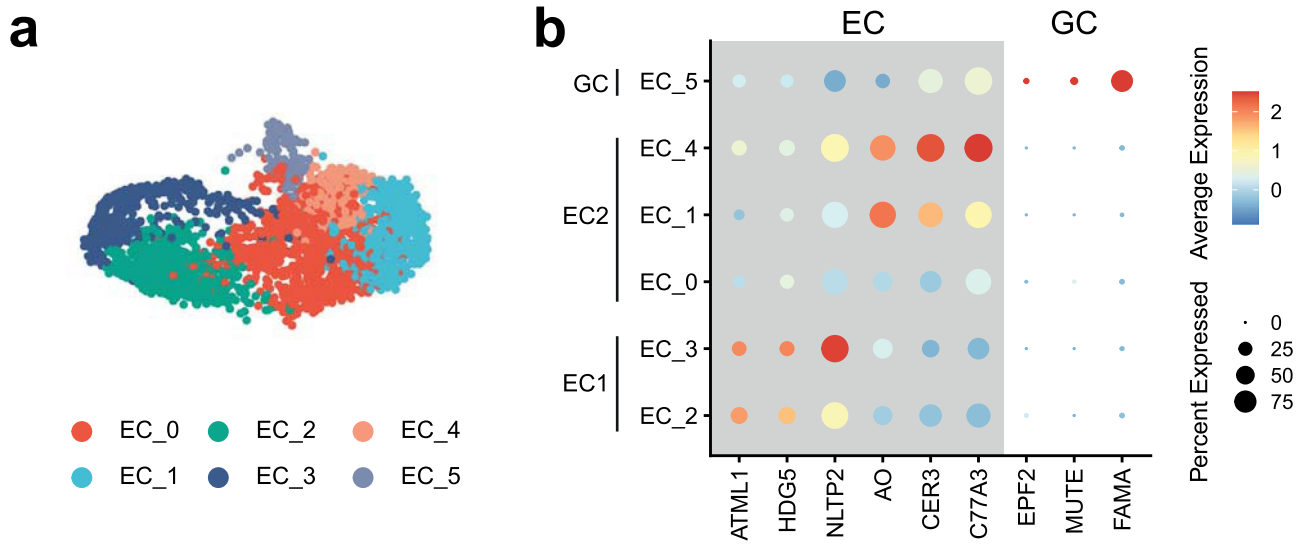

**Extended Data Fig. 3 | Expression patterns of MIA genes in cell clusters derived from Alevin (a) and Cell Ranger (b).** The expression of *GO1* was dramatically underestimated by Cell Ranger. The full names of the selected genes are provided in Supplementary Table 7.

https://doi.org/10.1038/s41477-022-01291-y



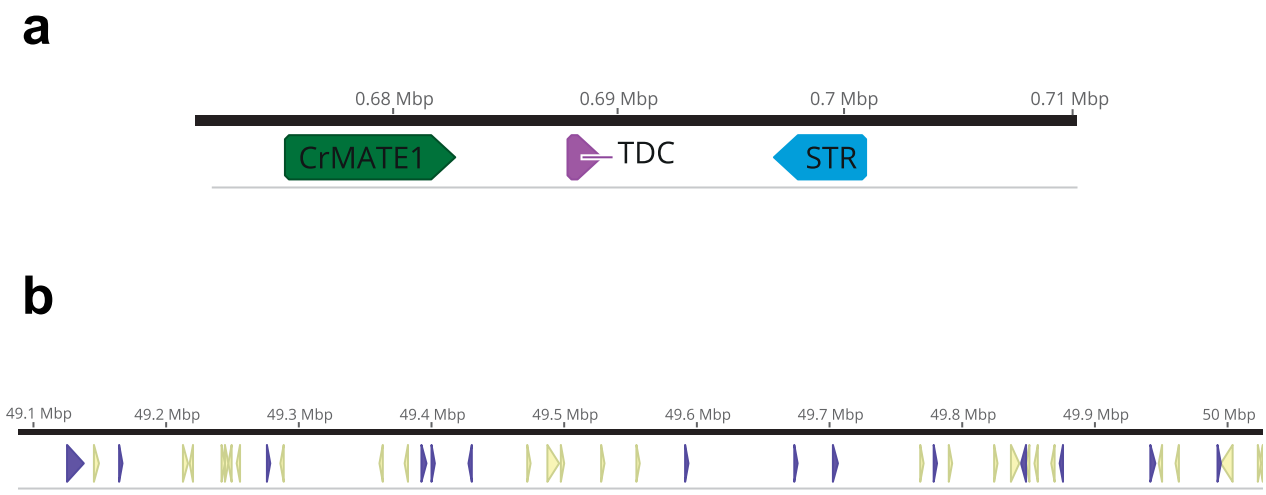

**Extended Data Fig. 4 | RIH validation of marker genes used for cell type annotation.** Paraffin-embedded serial cross-sections from 1.8–2.0 cm leaves were hybridized with digoxigenin-labeled transcripts. Sections were hybridized with sense and antisense RNA probes for *NLTP2*, *SABP2*, *CB21* and *PRS2* to localize their mRNAs in *C. roseus* leaves. The identified cell types are indicated by yellow arrows: *CB21*, mesophyll cell; *SABP2*, internal phloem-associated parenchyma cell; *NLTP2*, epidermal cell; *PRS2*, vascular cell. le, lower epidermis; ue, upper epidermis; pm, palisade mesophyll cells; and sm, spongy mesophyll cells.

https://doi.org/10.1038/s41477-022-01291-y



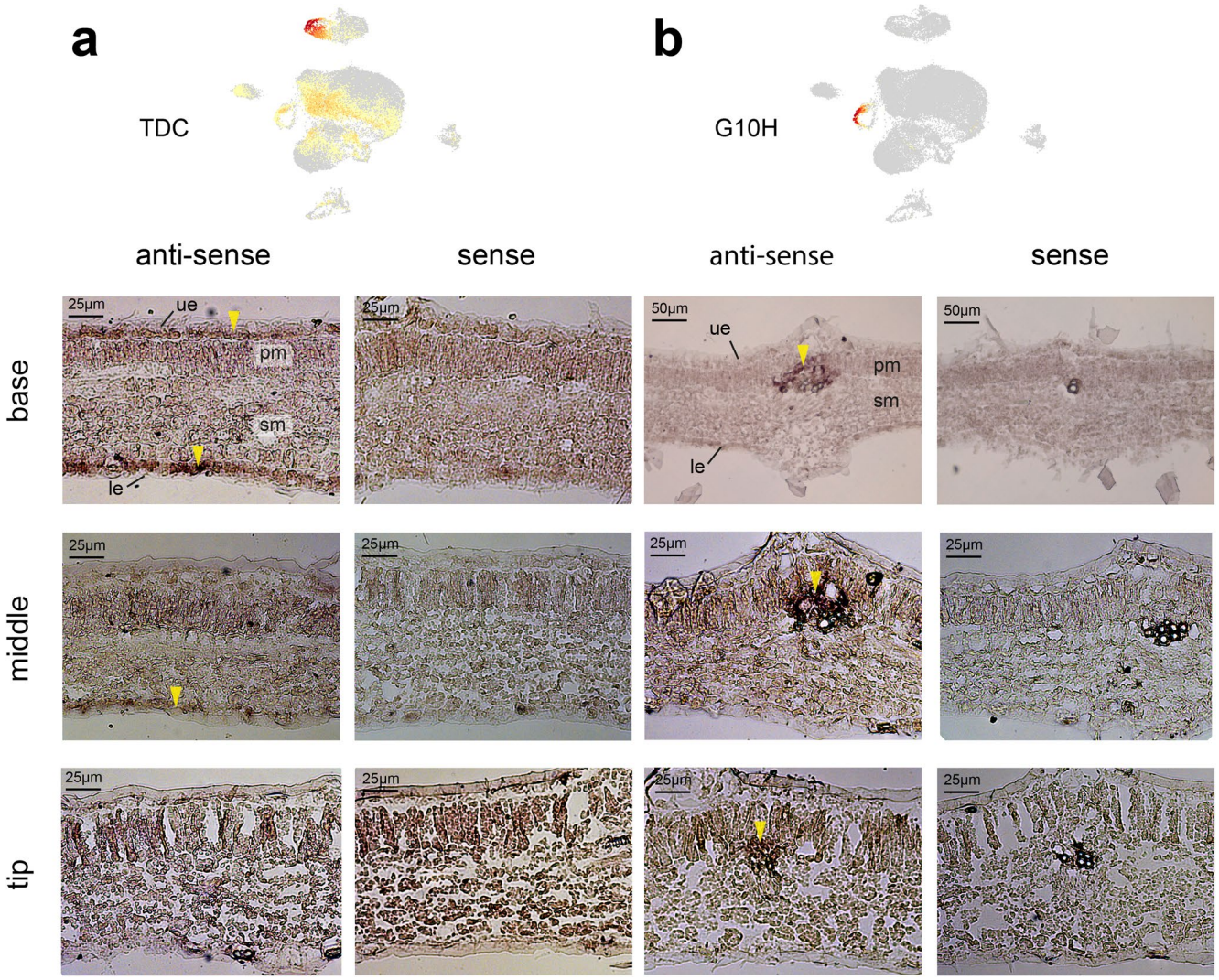

**Extended Data Fig. 5 | Reassignment of the EC population. a**, UMAP visualization of subclusters in the EC population. **b**, Dot plot showing the expression patterns of EC and GC marker genes in EC subclusters. Dot diameter indicates the proportion of cells expressing a given gene in each cluster, whereas the color indicates the scaled average expression. The full names of the selected genes are given in Supplementary Table 5.

https://doi.org/10.1038/s41477-022-01291-y




**Extended Data Fig. 6 | Gene clusters containing transporter genes that are possibly involved in the shuttling of MIA intermediates. a**, A gene cluster containing *STR*, *TDC* and *CrMATE1* on Pseudo-Chr5. **b**, A PUP cluster on Pseudo-Chr7. The PUPs are highlighted in purple.

https://doi.org/10.1038/s41477-022-01291-y




**Extended Data Fig. 7 | Localization of *TDC* (a) and *G10H* (b) mRNAs in developing leaves using RIH.** Paraffin-embedded serial cross-sections from 1.8–2.0 cm leaves were hybridized with digoxigenin-labeled transcripts. Sections were hybridized with antisense and sense RNA probes. base, leaf base; middle, the middle area of the leaf at a distance of 6 mm from the base; tip, the tip portion of the leaf at 11 mm from the base; le, lower epidermis; ue, upper epidermis; pm, palisade mesophyll cells; sm, spongy mesophyll cells.

nature portfolio

Corresponding author(s): Benoit St-Pierre, Shilin Chen and Chao Sun

Last updated by author(s): Oct 8, 2022

# 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<br>*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<br>*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 | No software was used for data collection |
| Data analysis | Cell Ranger v6.0.0, Salmon v1.4, Circos v0.69.9, Adobe Illustrator 25.0, Geneious v2022, Primer Premier v6.0<br><br>Genome assembly and gene model prediction<br>Canu v1.5, BioNano Solve v3.5, BWA v0.7.17, LACHESIS, BUSCO v2.0, AUGUSTUS v3.1, SNAP v2006, GlimmerHMM v1.2, GeMoMa v1.2.1, PASA v2.0.2, EVM v1.1.1<br><br>R and R packages:<br>R v4.0.5, Seurat v4.0.3, Monocle v2.8.0, clusterProfiler v3.18.1, dplyr v1.0.7, ggplot2 v3.3.5, ggsci v2.9, plotly v4.9.4.1, pheatmap v1.0.12, DoubletFinder v2.0.3, MAGIC v3.0.0 |

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 single-cell and bulk RNA sequencing data generated in this study have been deposited to NCBI with the accession number PRJNA759937. The scRNA-seq datasets of A. thaliana were downloaded from the Beijing Institute of Genomics Data Center (http://bigd.big.ac.cn) with the accession number PRJCA003094. The genome used in this study has been deposited to NCBI with the accession number PRJNA841429. The quantification results have been deposited to figshare with DOI: 10.6084/m9.figshare.20255094.

# 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 | We sequenced three libraries. Sampling was randomized. For each sample, we collected a large number of leaves and pooled them together. |
| Data exclusions | We removed low quality cells based on the criteria of gene number (500–6,000), UMI count (500–30,000), and percentages of mitochondrial (3%) and chloroplast genes (40%). |
| Replication | Three biological replicates were supplied and all attempts at replication were successful. Single cell data were used to predict distinct cell clusters. The cluster annotation were validated by comparing with published datasets. |
| Randomization | Sampling was randomized. We collected a large number of leaves and pooled them together. |
| Blinding | Not applicable - no treatment/control groups used |

# 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 |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

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