New Phytologist

Research

Cell-type-aware regulatory landscapes governing monoterpene indole alkaloid biosynthesis in the medicinal plant *Catharanthus roseus*

Chenxin Li[1,2], Maite Colinas[3], Joshua C. Wood[1], Brieanne Vaillancourt[1], John P. Hamilton[1,2], Sophia L. Jones[1], Lorenzo Caputi[3], Sarah E. O'Connor[3] and C. Robin Buell[1,2,4]

[1]Center for Applied Genetic Technologies, University of Georgia, Athens 30602 GA, USA; [2]Department of Crop and Soil Sciences, University of Georgia, Athens 30602 GA, USA; [3]Department of Natural Product Biosynthesis, Max Planck Institute for Chemical Ecology, Jena, 07745, Germany; [4]Institute of Plant Breeding, Genetics, and Genomics, University of Georgia, Athens 30602 GA, USA

Authors for correspondence:
C. Robin Buell
Email: robin.buell@uga.edu

Sarah E. O'Connor
Email: oconnor@ice.mpg.de

Lorenzo Caputi
Email: lcaputi@ice.mpg.de

Received: *9 July 2024*
Accepted: *3 October 2024*

New Phytologist (2025) **245:** 347–362
doi: 10.1111/nph.20208

Key words: idibioblast, single-cell ATAC-seq, single-cell RNA-seq, transcription factors, vinca alkaloids.

Summary

- In plants, the biosynthetic pathways of some specialized metabolites are partitioned into specialized or rare cell types, as exemplified by the monoterpenoid indole alkaloid (MIA) pathway of *Catharanthus roseus* (Madagascar Periwinkle), the source of the anticancer compounds vinblastine and vincristine. In the leaf, the *C. roseus* MIA biosynthetic pathway is partitioned into three cell types with the final known steps of the pathway expressed in the rare cell type termed idioblast. How cell-type specificity of MIA biosynthesis is achieved is poorly understood.
- We generated single-cell multi-omics data from *C. roseus* leaves. Integrating gene expression and chromatin accessibility profiles across single cells, as well as transcription factor (TF)-binding site profiles, we constructed a cell-type-aware gene regulatory network for MIA biosynthesis.
- We showcased cell-type-specific TFs as well as cell-type-specific *cis*-regulatory elements. Using motif enrichment analysis, co-expression across cell types, and functional validation approaches, we discovered a novel idioblast-specific TF (*Idioblast MYB1*, *CrIDM1*) that activates expression of late-stage MIA biosynthetic genes in the idioblast.
- These analyses not only led to the discovery of the first documented cell-type-specific TF that regulates the expression of two idioblast-specific biosynthetic genes within an idioblast metabolic regulon but also provides insights into cell-type-specific metabolic regulation.

Introduction

An emerging feature of plant-specialized metabolism is the spatial and temporal partitioning of biosynthetic gene expression (Jacobowitz & Weng, 2020), where biosynthetic pathways can be confined to rare and specialized cells within an organ, and/or partitioned across multiple cell types (Weng *et al.*, 2021). For example, in opium poppy, morphine biosynthesis is portioned between the sieve element and laticifer cells (Onoyovwe *et al.*, 2013). The medicinal plant *Catharanthus roseus* produces monoterpene indole alkaloids (MIAs), including vinblastine and vincristine (also known as vinca alkaloids), which are clinically used to treat various cancers (O'Connor & Maresh, 2006). The MIA biosynthetic pathway can be conceptually divided into four stages: the methyl erythritol phosphate (MEP) pathway that provides the precursor to the monoterpene moiety of MIAs, the iridoid stage that generates the monoterpene moiety of MIAs, the alkaloid scaffolding stage, and finally the late alkaloid stage that further decorates these MIA scaffolds (Supporting Information Table S1). The MIA pathway genes in *C. roseus* display intricate cell-type-specific expression patterns. The MEP and iridoid stages of the pathway are exclusively expressed in a specialized vasculature-associated cell type, the inner phloem-associated parenchyma (IPAP) (Burlat *et al.*, 2004; Simkin *et al.*, 2013; Miettinen *et al.*, 2014). The alkaloid scaffolding steps are expressed in the epidermis (Burlat *et al.*, 2004; Guirimand *et al.*, 2011; Miettinen *et al.*, 2014), and the final known steps of the pathway are restricted to a rare and specialized cell type termed idioblast (Guirimand *et al.*, 2011; Li *et al.*, 2023), which are scattered throughout the leaf (Yamamoto *et al.*, 2016, 2019). While *C. roseus* is primary known as a source of chemotherapeutic medications, the intricate partitioning of the MIA pathway into multiple cell types makes *C. roseus* an important model system for investigating cell-type-specific regulation of plant-specialized metabolism.

Even though many biosynthetic genes have been cloned and characterized, overall, we have a limited understanding of how

cell-type specificity of biosynthetic gene expression is achieved in plants. This is especially true for MIA biosynthetic genes expressed in the idioblast. Since the idioblast is where clinically relevant alkaloids accumulate (Li *et al.*, 2023b), understanding the biology of idioblast and how biosynthetic genes are activated in the idioblast has practical implications regarding breeding or engineering *C. roseus* with high alkaloid content. Biosynthesis of specialized metabolites from other species is sometimes also restricted to rare cell types. For example, aliphatic glucosinolate biosynthesis is restricted to phloem parenchyma in Arabidopsis (Tenorio Berrío *et al.*, 2022), and the antidepressant hyperforin is restricted to rare cells termed 'hyper cells' in St. John's wort (Wu *et al.*, 2024). Understanding how biosynthetic genes are activated in specialized cell types may produce generalizable principles across species, which may in turn facilitate metabolic engineering efforts concerning high-value plant natural products.

Several transcription factors (TFs) have been identified as regulators of the MIA biosynthetic pathway in *C. roseus* (Menke, 1999; van der Fits, 2000; van der Fits & Memelink, 2000; Peebles *et al.*, 2009; Zhang *et al.*, 2011; Van Moerkercke *et al.*, 2015, 2016; Bahieldin *et al.*, 2018; Schweizer *et al.*, 2018; Singh *et al.*, 2021), primarily in the context of jasmonate (JA) induction of this pathway. Major known regulators of the MIA pathway include *MYC2* (Zhang *et al.*, 2011; Schweizer *et al.*, 2018), bHLH iridoid synthesis (*BIS*) family TFs (Van Moerkercke *et al.*, 2015, 2016; Colinas *et al.*, 2021; Singh *et al.*, 2021), and Octadecanoid-derivative Responsive Catharanthus AP2-domain (*ORCA*) family TFs (Menke, 1999; van der Fits & Memelink, 2000; Peebles *et al.*, 2009; Colinas *et al.*, 2021), all of which mediate JA induction of the MIA pathway. Historically, the regulation of MIA biosynthesis has been studied in the context of herbivory or JA elicitation, and more recently in the context of light induction (Liu *et al.*, 2019). However, it is clear that *C. roseus* produces MIAs in the baseline (or unelicited) condition, and even in the baseline condition, MIA biosynthetic genes exhibit cell-type specificity (Burlat *et al.*, 2004; Simkin *et al.*, 2013; Miettinen *et al.*, 2014; Guirimand *et al.*, 2011; Li *et al.*, 2023b). As all currently available studies on transcriptional regulation of the MIA pathway have relied on whole organ (bulk) samples, how the pathway is regulated at the cell-type level remains an enigma. Furthermore, *MYC2*, *BIS*, and *ORCA* families TFs have been shown to activate the pathway up to the alkaloid scaffolding stage of the pathway, and to date, no cell-type-specific regulators for the late-stage portion of the pathway have been identified.

Here, we investigated cell-type-specific regulation of MIA biosynthetic genes with a focus on the baseline (or unelicited) state. We applied single-cell multiome (gene expression and accessible chromatin profiles from the same nucleus) to investigate the regulatory landscapes of the MIA biosynthetic pathway in mature *C. roseus* leaves at the cell-type level. Using co-expression across single cells, TF-binding site (TFBS) profiles, and cell-type-aware TFBS accessibility, we constructed a knowledge-based gene regulatory network (GRN) for this biosynthetic pathway. Our analyses uncovered a new idioblast-specific MYB TF, which regulates the expression of two idioblast-specific biosynthetic genes that are coregulated within an idioblast metabolic regulon. This study discovered a new regulatory component pertinent to the final steps of vinblastine and vincristine biosynthesis in *C. roseus* and furthers our understanding of cell-type-specific regulation of plant-specialized metabolism.

Materials and Methods

Nuclei isolation and single-cell library preparation

Catharanthus roseus (L.) G.Don (cultivar 'Sunstorm Apricot') plants were grown under a 14-h photoperiod at 22°C, with 350 μmol m^{-2} light intensity, potted in Sungro mix soil. Plants were watered every Monday morning, Wednesday at noon, and Friday afternoon. Plants were fertilized with Jack's 20–20–20 fertilizer on Wednesday. Mature, fully expanded leaves were sampled from 8- to 10-wk-old plants. Nuclei isolation was performed as described previously (Li *et al.*, 2022) with 0.01% Triton-X-100 in the nuclei isolation buffer. Around 0.3–0.5 g of leaves was chopped vigorously on ice on a Petri dish in nuclei isolation buffer for exactly 2 min. The lysate was filtered through 100 μm and 40 μm sieves, before passing through a 20 μm strainer twice. Nuclei were stained with 4′,6-diamidino-2-phenylindole (DAPI) and sorted using a Moflo Astrios EQ flow cytometer at the UGA Cytometry Shared Resource Laboratory. At least 100 000 nuclei were sorted into 500 μl of nuclei buffer (part of 10× Genomics Single Cell Multiome Kit). Sorted nuclei were pelleted by centrifugation at 200 ***g*** for 5 min and resuspended in 50 μl nuclei buffer. The integrity of the nuclei was visually inspected using a fluorescence microscope (Fig. S1a–i). Multiome libraries were constructed using the 10x Genomics Single Cell Multiome Kit, according to the manufacturer's instruction. Multiome libraries were sequenced on an Illumina NovaSeq 6000 (Table S2).

Single-nuclei RNA-seq processing

Single-nuclei RNA-seq libraries were processed using CUTADAPT (v.3.5) (Martin, 2011) with the following parameters: -q 30 -m 30 --trim-n -n 2 -g AAGCAGTGGTATCAACGCAGAGTACATGGG -a 'A{20}'. The pairing of the reads was restored using SeqKit (v.0.16.1) *pair* (Shen *et al.*, 2016). Paired reads were aligned and quantified using STARSOLO (Kaminow *et al.*, 2021), with the following parameters: --runThreadN 24 --alignIntronMax 5000 --soloUMIlen 12 --soloCellFilter EmptyDrops_CR --soloFeatures GeneFull --soloMultiMappers EM --soloType CB_UMI_Simple, and --soloCBwhitelist using the latest 10× Genomics whitelist of multiome barcodes. Gene-barcode matrices were analyzed with SEURAT (v.4) (Hao *et al.*, 2021) for downstream analysis. Removal of low-quality nuclei and suspected multiplets was performed using the distributions of UMI counts and detected genes (Fig. S2; Table S3).

Single-nuclei RNA-seq analyses

Biological replicates were integrated using the 'IntegrateData()' function in Seurat using the top 3000 variable genes. Uniform

manifold approximation and projection (UMAP) was performed after a principal component analysis (PCA) using the following parameters: dims = 1:30, min.dist = 0.001, repulsion.strength = 1, n.neighbors = 15, spread = 5. Clustering of cells was performed with a resolution of 0.5. For cell-type classification, we used a manually curated marker gene list for mesophyll, epidermis, guard cells, and vasculature (Table S4), using previously established marker genes from Arabidopsis (Kim *et al.*, 2021; Lopez-Anido *et al.*, 2021) and *C. roseus* (Guirimand *et al.*, 2011; Li *et al.*, 2023b; Miettinen *et al.*, 2014; Simkin *et al.*, 2013) (Fig. S3). For subclustering idioblast cells, the Seurat object was first filtered for idioblast cells (Cluster 8), rescaled, and reclustered. Differential gene expression was done between Subcluster 0 and Subcluster 1 (Fig. S4) using 'FindMarkers()' function in Seurat with the following parameters: only.pos = F, logfc.threshold = 0. For dot-plot style expression heat maps, average expression of genes was calculated as the average *Z*-score of log-transformed normalized expression values across cell clusters and cell types. Dot sizes indicated the percentage of cells where a given gene is expressed (> 0 reads) in each cell type or cell cluster. *De novo* marker identification was performed using the 'FindAllMarkers()' function in Seurat using the following options: only.pos = T, min.pct = 0.25, logfc.threshold = 0.25. Expression matrices (logCPM values (Dataset S1) and *Z* scores (Dataset S2)) were generated at the cell cluster level, where each row is a gene, and each column is a cell cluster.

Single-nuclei ATAC-seq processing

Single-nuclei ATAC-seq data were processed using the 10× Genomics Cell Ranger ARC pipeline (https://www.10xgenomics.com/software). For initial quality control and nuclei filtering, the 'atac_peaks.bed' files from the Cell Ranger ARC output were used. The peak bed files for the three biological replicates were sorted and merged using BEDTOOLS (v.2.30) *merge* (Quinlan & Hall, 2010). This common set of peaks was used to process all three biological replicates. The 'atac_fragments.tsv.gz' files from the Cell Ranger ARC output were used for downstream analyses using SIGNAC (v.1.6.0) (Stuart *et al.*, 2021) and SEURAT (v.4) (Hao *et al.*, 2021). Nuclei were filtered for > 1000 peaks/nuclei, > 2000 fragments/nuclei, and fraction of fragments in peaks > 0.25 (Table S5). For data integration, the replicates were merged first, then integrated using the 'IntegrateEmbeddings()' function in Signac using the 'lsi' dimension reduction. Integration with the gene expression assay was performed by first filtering for shared nuclei in both gene expression and chromatin assays, after which the integrated ATAC-seq object was adjoined to the integrated RNA-seq object as a chromatin assay. By doing so, the cell cluster and cell-type assignment information is transferred to the ATAC-seq assay. Fragment files were split into separate files for each cell cluster and converted to bed files. Peak calling at each cell cluster performed using MACS2 (v.2.2.7.1) (Zhang *et al.*, 2008) using the following parameters: -f BED -g 444 800 000 (80% of the genome assembly size was set as the effective mappable genome size) --nomodel --broad. The resultant peak files were sorted and merged to be used as the features in the chromatin accessibility assay. These peaks were used as 'ATAC-seq peaks' in all downstream analyses. UMAP visualization for ATAC-seq was performed using the following parameters: reduction = 'lsi', dims = 2 : 30, min.dist = 0.001, repulsion.strength = 1, n.neighbors = 30, spread = 1. Joint UMAP visualization was done using the 'FindMultiModalNeighbors()' functions in Signac. ATAC-seq coverage around genes (Fig. S5) and peaks (Fig. S6) was calculated and visualized using DEEPTOOLS (v.3.5.1) (Ramírez *et al.*, 2014).

DAP-seq library construction and processing

The coding sequence of *ORCA3* and *ORCA4* was synthesized and cloned into pIX-Halo (O'Malley *et al.*, 2016), downstream and in frame with the halo tag. *In vitro* gene expression was performed using Promega TnT SP6 High-Yield Wheat Germ Protein Expression System. Each *in vitro* gene expression reaction was spiked with 200 ng of a pIX-RFP plasmid, such that the gene expression reaction can be monitored using RFP fluorescence. Genomic data libraries were constructed from genomic DNA isolated from mature leaves of 8- to 10-wk-old *C. roseus* plants using a KAPA HyperPrep Kit, after the genomic DNA was sheared to 200 bp with a Covaris ultrasonicator at the UGA Genomics and Bioinformatics Core. The full volume of gene expression reaction was combined with 40 ng of gDNA library and 10 µl of Promega Halo-beads for each affinity reaction. Bead-bound DNA was recovered by heating the affinity reaction to 95°C for 5 min. Indexing PCR was performed with 13 cycles, and the libraries were sequenced in paired-end 50-bp format on an Illumina NextSeq 2000 instrument (Tables S2, S6).

Sequencing adapters were trimmed with CUTADAPT (v.3.5) (Martin, 2011), after which reads were aligned to the *C. roseus* v.3 genome (Li *et al.*, 2023b) using BWA MEM (v.0.1.17) (Li, 2013). Peak calling was performed with MACS2 (v.2.2.7.1) using the following parameters: -g 444800000 (80% of the genome assembly size was set as the effective mappable genome size), -q 0.05 (adjusted *P* value cutoff < 0.05), using the bam file of the halo tag control as the background file. DAP-seq coverage around peaks (Fig. S7) was calculated and visualized using DEEPTOOLS (v.3.5.1). Putative target genes were assigned using BEDTOOLS (v.2.30) *closest*, with the -d parameter selected. Genes overlapping or within 2 kb of a DAP-seq peak were designated as a putative target gene. Accessible DAP-seq peaks were defined as DAP-seq peaks overlapping or within 100-bp to either ends of an ATAC-seq peak. DNA sequence of DAP-seq peaks was extracted using BEDTOOLS (v.2.30) *getfasta* and subjected to *de novo* motif discovery using MEME (v.5.4.1) (Bailey *et al.*, 2009): using the following parameters: -dna -revcomp -mod anr -nmotifs 10 -minw 5 -maxw 12 -evt 0.01.

Marker peak and motif overrepresentation analyses

Marker peaks for epidermis and idioblast were detected using the 'FindMarkers()' function in Seurat after setting the default assay of the multiome object to chromatin accessibility, using the following parameters: only.pos = T, test.use = 'LR',

min.pct = 0.05, latent.vars = 'nCount_peaks', group.by = 'cell_type'. Only peaks with adjusted p values < 0.05 were used for downstream analyses. For motif enrichment analysis, position weight matrices were obtained using the 'getMatrixSet()' function in Signac, using the following parameters: x = JASPAR2020 (Fornes *et al.*, 2019), opts = list(collection = 'CORE', tax_group = 'plants', all_versions = FALSE). These motifs were added to the multiome object using the 'AddMotifs()' function in Signac. Overrepresented motifs were identified using the 'FindMotifs()' function in Signac.

Gene co-expression analyses

Gene co-expression analysis by graph-based clustering was performed as previously described (Li *et al.*, 2023a). The top 3000 most variable genes were used for gene-wise correlation. Pairwise Pearson's correlation was performed to generate an edge table, which was filtered for $r > 0.75$. Graph-based clustering was performed using the Leiden algorithm (Traag *et al.*, 2019) with a resolution parameter of 4, which was implicated in R by the IGRAPH package (Csardi & Nepusz, 2006).

Overexpression assays

Coding sequences of *IDW1* and *IDM1/2/3* were cloned in between the 35S promoter and 35S terminator and transformed into *Agrobacterium tumefaciens* strain GV3101. We used previously published MYC2 and ORCA3 overexpression constructs (Schweizer *et al.*, 2018). Transient expression experiments were done on *C. roseus* (cultivar 'Little Bright Eyes') petals. Infiltration was done as previously described (Colinas & Goossens, 2022). Two days before the infiltration, all open flowers were removed. Two sets of experiments were performed. We performed infiltrations at two agrobacterium titers, 0.1 optical density (OD) and 0.4 OD, which is the highest titer that can be used without resulting in wilting of the petals after infiltration. In the first set, individual strains were infiltrated at 0.1 OD, and the total OD was adjusted to 0.4 using the control agrobacterium strain carrying GUS. In the second set, all strains were infiltrated at 0.4 OD. Two days after the infiltration, infiltrated petals were harvested and stored in a − 80 freezer until RNA extraction. RNA extraction was performed with a Qiagen RNeasy Plant Kit.

RNA-seq analysis for overexpression samples

Sequencing adapters were trimmed from petal RNA-seq libraries using CUTADAPT (v.3.5) (Martin, 2011). Adapter trimmed libraries were pseudo-aligned and quantified using KALLISTO (v.0.48) (Bray *et al.*, 2016), with the --plaintext option turned on. When the appropriate strandedness parameter was used, pseudo-alignment rate ranged from 86.2% to 89%. Differential gene expression analyses were performed using DESEQ2 (v.1.34.0) (Love *et al.*, 2014), using the GUS treatment with of the corresponding experiment as control. Genes with adjusted *P* values <0.05 were taken as differentially expressed genes.

Phylogenetic analysis of MYB domain TFs

We used 'MYB annotator' (Pucker, 2022) to automatically identify MYB domain proteins in *C. roseus*. We aligned the sequences of MYB domain using MAFFT (v.7.505) (Katoh, 2002) using the following parameters: --anysymbol --maxiterate 1000 –localpair. The alignment file generated by MAFFT was used to generate maximum likelihood phylogenetic trees using RAXML (Stamatakis, 2014) with the following parameters: -f a -m PROTGAMMAAUTO -x 666 -N 250 -p 666 -T 4, with 250 iterations for bootstrapping.

Reporter transactivation assays

The reporter transactivation assays were performed in a two-component format: a reporter component and an overexpression component. Genetic parts used in reporter assays were amplified from a vector tool kit for plant molecular biology (Chamness *et al.*, 2023). The accessible chromatin regions immediately upstream of *D4H* and *DAT* were cloned upstream of a 35S minimal promoter, which controls the expression of DsRed reporter. On the same plasmid, a GFP internal control driven by the Arabidopsis UBQ1 promoter was also included. The overexpression component was an agrobacterium GV3101 strain carrying 35S:*IDM1*, the same construct used in overexpression assays. As in the transient expression assays described above, experiments were done on *C. roseus* (cultivar 'Little Bright Eyes'). Two days after the infiltration, petals were imaged using a fluorescent microscope. Pixel intensity was quantified using IMAGEJ (Schneider *et al.*, 2012).

Results

The cell-type-specific expression patterns of MIA biosynthetic genes are reflected in single-cell multiome profiles

To investigate the regulation of MIA biosynthetic genes (Table S1; Fig. S1a) at the single-cell resolution, we generated dual gene expression and chromatin accessibility profiles across single cells. We first isolated intact nuclei (Fig. S1b–i) from mature *C. roseus* leaves and constructed replicated single-cell multiome (RNA-seq and assay for transposase accessible chromatin followed by sequencing (ATAC-seq)) libraries using the 10x Genomics Multiome Kit (Table S2). For gene expression, we obtained gene expression profiles for a total of 8803 high-quality nuclei and 18 532 expressed genes (Figs 1a, S2; Table S3).

We first examined the gene expression data of this multiome dataset (Fig. 1a). Cell clustering patterns are highly similar across the three biological replicates (Fig. S3a). Using previously established marker genes (Table S4), we identified major cell types of leaf (e.g. mesophyll, epidermis, and vasculature) as well as two rare cell types in which MIA biosynthetic genes were expressed (i.e. IPAP and idioblast) (Fig. S3b). Mesophyll and epidermis were the most abundant cell types, accounting for 54% and 18%

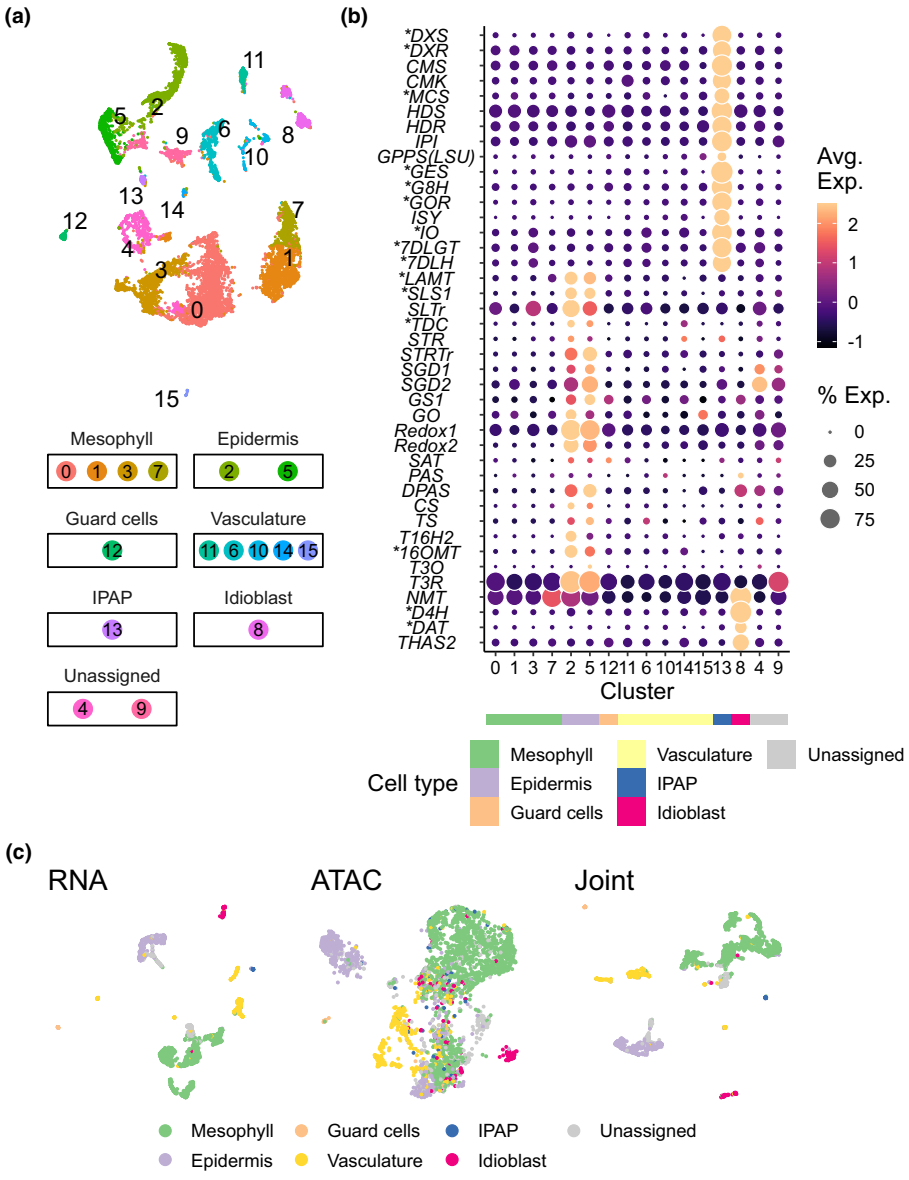

Fig. 1 Cell-type-specific expression of monoterpene indole alkaloid (MIA) biosynthetic genes is recapitulated in a *Catharanthus roseus* leaf single-cell multiome dataset. (a) Uniform Manifold Approximation and Projection (UMAP) of nuclei containing high-quality RNA-seq data (n = 8803), color coded by cell clusters. (b) Gene expression heatmap of *C. roseus* MIA biosynthetic genes across cell clusters detected in (a). Rows are biosynthetic genes and transporters, which are ordered from upstream to downstream. Asterisks denote matching cell-type specificity with previously reported RNA *in situ* hybridization results. Color scale shows the average scaled expression of each gene at each cell cluster. Cell clusters are sorted by cell types. Dot size indicates the percentage of cells where a given gene is detected. The predicted cell type for each cell cluster is annotated by the color strip below the x-axis. (c) UMAP of nuclei containing both high-quality RNA-seq and ATAC-seq data (n = 3542 nuclei for all three UMAP), color coded by cell types. From left to right: UMAP based on gene expression assay, chromatin accessibility assay, and joint analysis.

of assayed nuclei, respectively. Consistent with their rare nature, IPAP and idioblast accounted for only 1% and 4% of assayed nuclei, respectively (Fig. S3c). We found that the MIA biosynthetic pathway was organized into three discrete cell types (Fig. 1b). The MEP and iridoid stages (up to *7-DLH*, Fig. S1a) of the pathway were exclusively expressed in the IPAP cells. The following stage, which includes most of the alkaloid steps, was expressed in the epidermis. Finally, the last four known steps of the pathway were only expressed in the idioblast. These data were highly consistent with recently published single-cell RNA-seq results using protoplasts (Li *et al.*, 2023b) and were fully supported by previously reported RNA *in situ* hybridization results (marked with asterisk) (Burlat *et al.*, 2004; Miettinen *et al.*, 2014).

We noticed that idioblast cells were composed of two subpopulations in UMAP space (Fig. 1a), which were present across all three replicates (Fig. S3a), suggestive of heterogeneity among idioblast cells. We performed subclustering of idioblast cells (Fig. S4). In both PCA and UMAP space, two major subclusters can be observed (Fig. S4a). We performed differential gene expression analysis and found that the differences between the major subclusters are overall subtle, as the vast majority of $\log_2$FC values are between −1 and 1 (Fig. S4b). We found that one subcluster has slightly higher expression of genes associated with photosynthesis (e.g. LHC protein and Rubisco Activase), while the other subcluster has higher expression of alkaloid biosynthetic genes *THAS2* and *Serpentine Synthase* (*SS*) (Yamamoto *et al.*, 2021) (Fig. S4b,c). Despite the observed heterogeneity, both subclusters expressed the canonical idioblast-specific biosynthetic genes *D4H* and *DAT* (Fig. S4c).

We next proceeded to analyze chromatin accessibility data to investigate how biosynthetic genes were regulated to generate cell-type-specific expression patterns. For the chromatin accessibility assay, high-quality ATAC-seq nuclei have fraction of

fragments in peaks > 0.25, greater than 2000 ATAC fragments per nuclei, and greater than 1000 peaks per cell (Table S5), resulting in accessibility profiles for a total of 3765 high-quality nuclei and 43 630 accessible chromatin peaks (Fig. 1c). We performed a joint analysis by matching the cell barcodes from both assays. Matching 8803 high-quality nuclei from the RNA-seq assay with 3765 high-quality nuclei for the ATAC-seq assay, the joint analysis resulted in an intersecting set of 3542 nuclei containing both high-quality RNA-seq and ATAC-seq data (Fig. 1c).

A gene regulatory network for MIA biosynthetic genes integrating co-expression, chromosome accessibility, and transcription factor-binding site (TFBS) profiles

To investigate the regulation of MIA biosynthetic genes, we examined chromatin accessibility landscapes across cell types. ATAC-seq fragments were highly enriched at transcription start and end sites (Fig. S5). Among the three biological replicates, 45.9%, 40.9%, and 41.3% of ATAC-seq fragments overlapped transcriptional start sites (Fig. S5; Table S5). We then defined ATAC-seq peaks using MACS2 (Zhang *et al.*, 2008); among biological replicates, 48.9%, 48.5%, and 48.8% of fragments are within ATAC-seq peaks (Fig. S6; Table S5). The median length of ATAC-seq peaks was 566 bp (Fig. S7a). The chromatin accessibility landscapes were complemented TFBS profiles of ORCA3, a well-known regulator of MIA biosynthesis (van der Fits & Memelink, 2000), and its tandemly duplicated paralog ORCA4 (Kellner *et al.*, 2015; Paul *et al.*, 2017; Colinas *et al.*, 2021) (Fig. 2a,b; Table S6). We determined TFBS profiles for ORCA3/4 using DNA affinity purification sequencing (DAP-seq) (O'Malley *et al.*, 2016). Average DAP-seq peak lengths were similar (*c.* 300-bp) between ORCA3 and ORCA4 (Fig. S7b,c) with *c.* 10% of DAP-seq peaks intersecting with ATAC-seq peaks (Fig. S7d), consistent with the *in vitro* nature of the DAP-seq assay (O'Malley *et al.*, 2016). Signal-to-noise ratios at DAP-seq peaks were strong (Fig. S7e,f; Table S6), comparable to the most high-quality DAP-seq datasets that have been published (O'Malley *et al.*, 2016).

ORCA TFs are known to activate both the alkaloid steps of the biosynthetic pathway (e.g. *Strictosidine Synthase* (*STR*) and *Tryptophan Decarboxylase* (*TDC*)) (van der Fits & Memelink, 2000; Peebles *et al.*, 2009) and the upstream iridoid steps (e.g. *7-DLGT*) (Colinas *et al.*, 2021). *7-DLGT* is exclusively expressed in the IPAP cells (Fig. 1b) and consistent with its expression specificity, and the *7-DLGT* locus has a unique chromatin accessibility signal in IPAP cells at both 5′ and 3′ ends of the gene (Fig. 2a). Strong DAP-seq peaks were observed for both ORCA3 and ORCA4 at the *7-DLGT* locus, but not for the affinity-tag control (Fig. 2a). These DAP-seq peaks also overlapped with an ATAC-seq peak that was accessible across all cell types. Together with previously reported data that overexpression of *ORCA3* or *ORCA4* led to the upregulation of *7-DLGT* (Colinas *et al.*, 2021), *7-DLGT* is a direct target of both *ORCA3* and *ORCA4*.

ORCA3 has been reported to bind to the promoters of *STR* and *TDC* and activate their expression (van der Fits & Memelink, 2000; Colinas *et al.*, 2021). *STR* and *TDC* are physically clustered on chromosome 3, along with *Secologanin Transporter SLTr* (Li *et al.*, 2023b). Multiple ATAC-seq peaks were detected within this 25-kb biosynthetic gene cluster, all of which were accessible across multiple cell types (Fig. 2b). ORCA3 and ORCA4 displayed similar binding profiles at this biosynthetic gene cluster. Each TF binds a total of four DAP-seq peaks in this region. Consistent with *STR* and *TDC* being direct targets of ORCA3, DAP-seq peaks were detected in the promoters of both *STR* and *TDC*. Overexpression of ORCA4 has also been reported to lead to enhanced expression of *STR* and *TDC* (Paul *et al.*, 2017; Colinas *et al.*, 2021), and the presence of ORCA4-binding sites suggests ORCA4 can directly activate both *STR* and *TDC*, consistent with previously reported promoter transactivation activity (Paul *et al.*, 2017). Lastly, since binding sites for ORCA3/4 were detected at the promoter of *Secologanin Transporter* (*SLTr*), and as *SLTr* is highly co-expressed with *STR* and *TDC* in the epidermis (Fig. 1b), *SLTr* is likely a direct target for ORCA3/4 as well.

We performed *de novo* motif discovery to identify the DNA-binding motifs of ORCA3/4. We found that the GCC box motif (Brown *et al.*, 2003) was enriched among ORCA3/4-binding sites (E value = 2.1×10^{-75} for ORCA3 and E value = 1.8×10^{-96} for ORCA4) (Fig. 2c). The same GCC box motif was detected regardless of whether we used all DAP-seq peaks as input or only accessible DAP-seq peaks as input. The GCC box is recognized by ethylene-responsive factors (ERFs) (Fig. 2c) (Fujimoto *et al.*, 2000) consistent with ORCA family TFs being within the broader AP2/ERF family.

Integrating gene co-expression across single cells, TF-binding sites, binding site chromatin accessibility, as well as previously reported overexpression (Schweizer *et al.*, 2018; Colinas *et al.*, 2021), and reporter transactivation data (Peebles *et al.*, 2009; Van Moerkercke *et al.*, 2015, 2016; Singh *et al.*, 2021), we generated a knowledge-based GRN for the MIA biosynthetic pathway (Fig. 2d). We first queried the expression patterns of previously studied TFs (Table S7) (Menke, 1999; van der Fits, 2000; van der Fits & Memelink, 2000; Peebles *et al.*, 2009; Suttipanta *et al.*, 2011; Zhang *et al.*, 2011; Van Moerkercke *et al.*, 2015, 2016; Bahieldin *et al.*, 2018; Schweizer *et al.*, 2018; Liu *et al.*, 2019; Singh *et al.*, 2021) in our single-cell dataset and found that only *ORCA4* and *BIS1/2/3* displayed cell-type-specific expression patterns relevant to iridoid and alkaloid biosynthetic genes (Fig. S8a). *BIS1/2/3* were expressed specifically in the IPAP cells, highly concordant with the iridoid biosynthetic genes that they regulate (Fig. 1d). *ORCA4*, but not *ORCA3*, was expressed specifically in the epidermis, albeit only in a small fraction of cells. Thus, *ORCA4*, but not *ORCA3*, may contribute to the epidermal-specific expression of alkaloid biosynthetic genes such as *STR*, *TDC*, and *SLTr* (Fig. 2b). All other TFs reported in the literature to be associated with MIA biosynthesis were expressed broadly across cell types or were not expressed in IPAP, epidermis, or idioblast cells (Fig. S8).

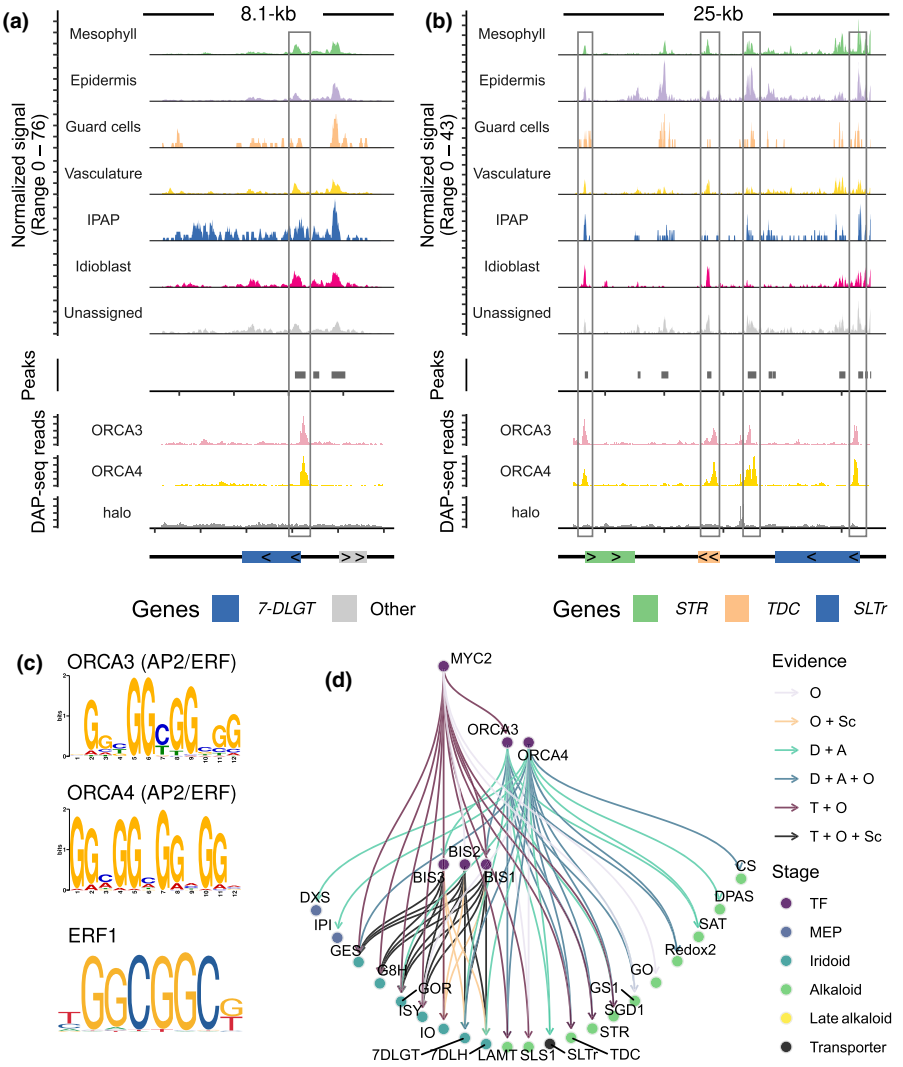

Fig. 2 A gene regulatory network for *Catharanthus roseus* monoterpene indole alkaloid biosynthetic genes integrating chromosome accessibility landscapes and transcription factor binding site profiles. (a, b) Coverage plot showing ATAC-seq (upper panels) and DAP-seq (lower panels) signals at the *7-DLGT* locus (a) and *STR-TCD-SLTr* biosynthetic gene cluster (b). Gray unfilled boxes highlight DAP-seq peaks that overlap with ATAC-seq peaks. Bottom track indicates the location and length of genes, where the direction of carets (> or <) indicates the strand of a gene. Halo: control DAP-seq experiment using the halo tag (affinity tag) alone. (c) DNA motifs enriched in ORCA3/4-binding sites, as well as a reference GCC box/ERF motif. *E* value = 2.1×10^{-75} for ORCA3 and *E* value = 1.8×10^{-96} for ORCA4. (d) A gene regulatory network integrating multiple modules of omics data and experimental data. Each node is a gene, color-coded by the stage of the biosynthetic pathway. Each edge represents a regulatory relationship, color-coded by the type of evidence supporting it. O, upregulated when the TF is overexpressed; Sc, co-expressed across single cells; D, overlapping or within 2 kb to a DAP-seq peak; A, DAP-seq peak accessible; T, promoter activated in a transactivation assay.

Based on their co-expression with target genes at the cell-type level, *BIS1/2/3* and *ORCA4* were selected as TF nodes for the GRN. Co-overexpression of *MYC2* and *ORCA3* was previously reported to strongly activate the iridoid and alkaloid stages of the pathway (Schweizer *et al.*, 2018), and thus, *MYC2* and *ORCA3* were also included in this network (Fig. 2d). The GRN contains 66 edges (Table S8), which were decorated by the types of evidence: (1) activated by overexpression of TF; (2) co-expressed at the single-cell level; (3) overlapping or within 2 kb of a DAP-seq peak; (4) DAP-seq peak accessible; and (5) promoter activated in a transactivation assay (Fig. 2d). We found that the combined actions of *MYC2*, *ORCA3/4*, and *BIS1/2/3* activate a large section of the MIA pathway, up to the biosynthetic gene encoding *Catharanthine Synthase*. Evidence also supported multiple feed-forward regulatory loops, where an upstream TF activates other TFs in addition to biosynthetic genes. The downstream TFs in turn activate the same target biosynthetic genes. For example, *ORCA3/4* activates iridoid and alkaloid biosynthetic genes, as well as *BIS* TFs that in turn activate iridoid biosynthetic genes. However, we also found that no regulatory relationships were detected beyond *Catharanthine Synthase* for these six TFs, consistent with previous reports where overexpression of *MYC2*, *ORCA*, and/or *BIS* TFs led to an increase in early-stage alkaloid metabolites (e.g. strictosidine), but not late-stage alkaloid such as vinblastine (Schweizer *et al.*, 2018; Colinas *et al.*, 2021). These observations prompted us to investigate components involved in the regulation of the late MIA pathway.

Cell-type-specific accessible chromatin regions mark late-stage MIA biosynthetic genes

MIA biosynthetic pathway genes downstream of *Catharanthine Synthase* are sequentially expressed in epidermis (*TS*, *T16H2*, *16OMT*, *T3O*, and *T3R*) and then in idioblast cells (*NMT*, *D4H*, *DAT*, and *THAS2*) (Fig. 1b; Table S1). *T16H2* and *16OMT* are consecutive steps of the late MIA pathway, expressed exclusively in the epidermis (Fig. 1b), and physically linked as a biosynthetic gene cluster (Fig. 3a), between which there is another gene encoding a cytochrome P450 that was not expressed in the leaf. At the *T16H2/16OMT* locus, there are four ATAC-

New Phytologist

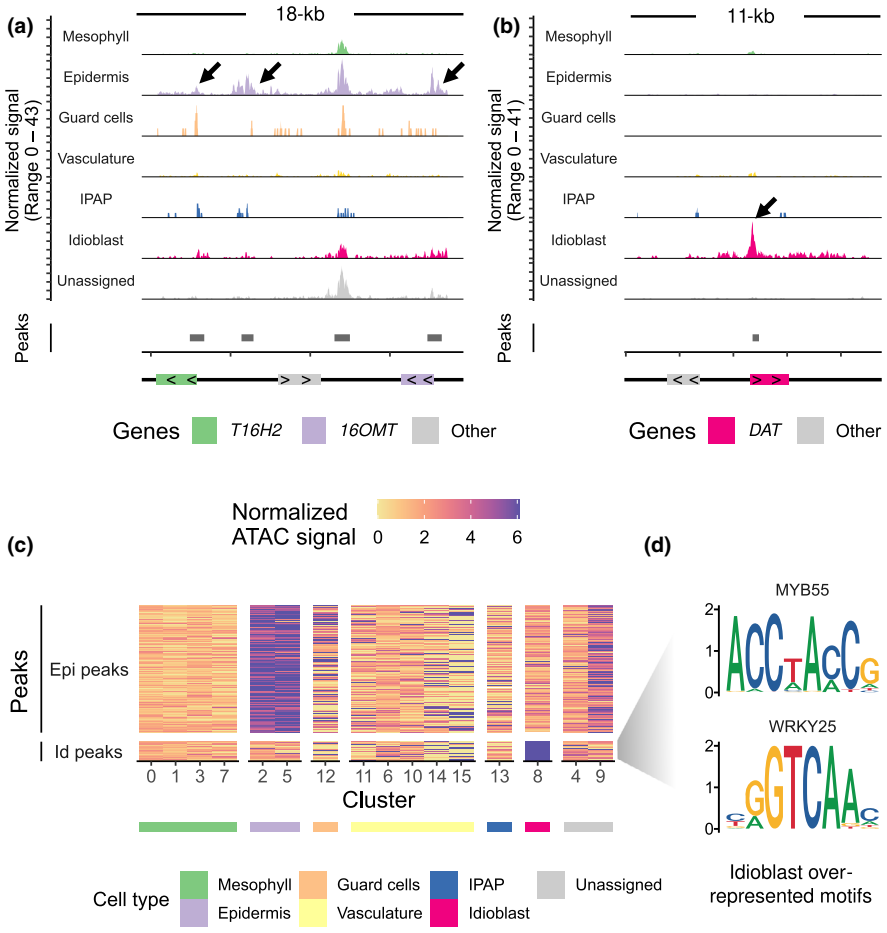

Fig. 3 Cell-type-specific accessible chromatin regions mark *Catharanthus roseus* late-stage monoterpene indole alkaloid biosynthetic genes. (a, b) Coverage plot showing ATAC-seq signals at the *T16H2-16OMT* gene pair (a) and *DAT* locus (b). Arrows highlight cell-type-specific ATAC-seq peaks. Bottom track indicates the location and length of genes, where the direction of carets (> or <) indicates the strand of a gene. Grey boxes along the 'Peaks' track represent ATAC-seq peaks. (c) Heat map showing accessibility of epidermis (Epi) and idioblast (Id) ATAC-seq marker peaks across cell clusters. Each row is an ATAC-seq peak. Each column is a cell cluster. Color scale is maxed out at 90th percentile of normalized ATAC-seq signal. The predicted cell type for each cell cluster is annotated by the color strip below the *x*-axis. (d) Transcription factor binding motifs overrepresented among idioblast marker peaks.

seq peaks. All but one of the peaks were preferentially accessible in the epidermis, consistent with the cell-type-specific expression of this gene pair (Fig. 3a). *DAT*, one of the final known steps of the MIA pathway, is only expressed in the idioblast (Fig. 1b), and its promoter was also specifically accessible in the idioblast (Fig. 3b).

To identify novel regulators for late-stage MIA biosynthetic genes downstream of *Catharanthine Synthase*, we first detected epidermis and idioblast marker peaks, which are ATAC-seq peaks preferentially accessible in the epidermis or idioblast, but not in any other cell types (Fig. 3c; Table S9). We detected 1050 epidermis marker peaks and 163 idioblast marker peaks. We next performed a motif enrichment analysis on epidermis marker peaks against the JASPAR plant TF-binding motif collection (Fornes *et al.*, 2019). We found that homeodomain, ERF, and MYB motifs were overrepresented among epidermis marker peaks (Fig. S9a). Homeodomain (e.g. *ANTHOCYANINLESS2* (Kubo *et al.*, 1999)), AP2/ERF (e.g. *WAX INDUCER1* (Kim *et al.*, 2022)), and MYB TFs (e.g. *WEREWOLF* (Tominaga *et al.*, 2007)) have been reported to control metabolic and developmental processes such as anthocyanin biosynthesis, cuticle development, and trichome development, respectively. Enrichment of these motifs suggests that additional TFs in the homeodomain, ERF, or MYB families may play a role in the regulation of MIA biosynthesis in the epidermis.

We also performed motif enrichment analysis on idioblast marker peaks and found that MYB- and WRKY-type motifs were overrepresented (Fig. 3c), for which we followed up with additional analyses and experiments.

Candidate WRKY and MYB TFs specifically expressed in the idioblast discovered by gene co-expression analysis

To further understand gene regulation in idioblast cells, we focused our attention on potential metabolic regulators in the idioblast. We performed gene co-expression analysis across cell clusters using graph-based clustering and detected tightly co-expressed modules (Fig. S9b). We queried co-expression modules containing MIA biosynthetic genes and detected a single co-expression module for epidermis, IPAP, and idioblast, respectively (Table S10). For example, *SLS1*, which was specifically expressed in the epidermis, was a member of the epidermis co-expression module, whereas the final known steps of the pathway, namely *NMT*, *D4H*, *DAT*, and *THAS2*, were all members of the idioblast co-expression module (Fig. S9b; Table S10). The partitioning of MIA biosynthetic genes into three distinct co-expression modules is similar to a co-expression network constructed from single-cell RNA-seq data generated from protoplasts (Li *et al.*, 2023b).

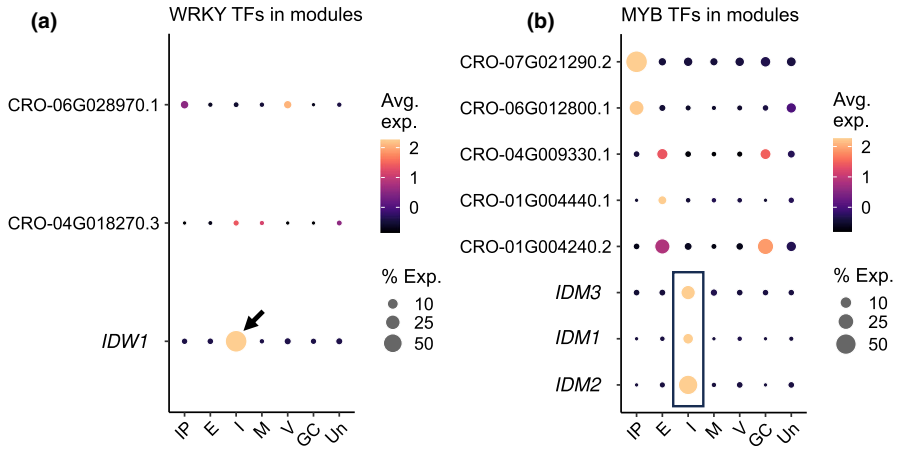

Fig. 4 Gene co-expression analysis across *Catharanthus roseus* leaf cell clusters discovered candidate WRKY and MYB transcription factors (TFs) specifically expressed in the idioblast. (a, b) Gene expression heatmap of WRKY TFs (a) and MYB TFs (b) across cell types. Color scales show the average scaled expression of each gene for each cell type. Dot size indicates the percentage of cells where a given gene is detected in each cell type. Only WRKY and MYB TFs detected in epidermis, internal phloem-associated parenchyma (IPAP), or idioblast co-expression modules are presented. Arrow indicates a single WRKY candidate (*Idioblast WRKY1*, *IDW1*: CRO_03G000120) specifically expressed in the idioblast. Box highlights three MYB candidates (*Idioblast MYB1*, *IDM1*: CRO_05G006800, *Idioblast MYB2*, *IDM2*: CRO_04G033370, *Idioblast MYB3*, *IDM3*: CRO_07G002170) specifically expressed in the idioblast.

Since WRKY and MYB motifs were overrepresented among idioblast ATAC-seq marker peaks, we queried WRKY and MYB family TFs within the gene co-expression modules. We identified a single WRKY TF (Fig. 4a) as well as three strong candidates of R2-R3 MYB TFs (Fig. 4b) that were exclusively expressed in the idioblast. We named these candidates *Idioblast WRKY1* (*IDW1*) and *Idioblast MYB1/2/3* (*IDM1/2/3*), respectively. All four candidates were induced by a methyl-jasmonate treatment (Van Moerkercke *et al.*, 2013) (Fig. S9c), among which *IDM1* displayed the highest level of induction ($\log_2$FC = 5.4, or 42-fold increase over control). Since the entire vinblastine biosynthetic pathway is elicited by methyl-jasmonate (van der Fits & Memelink, 2000; Van Moerkercke *et al.*, 2016), the MeJA-responsiveness displayed by these TF candidates suggests they might be transcriptional activators of the pathway. A recent study applied fluorescence activated cell sorting to enrich for idioblast cells before RNA-seq (Guedes *et al.*, 2024). Consistent with their idioblast specificity, all four TF candidates were detected at high levels in the idioblast fraction of sorted cells, but not in the mesophyll fraction (Fig. S9d).

To investigate the phylogenetic relationship among the three MYB candidates, we performed genome-wide identification of MYB domain proteins (Pucker, 2022) in the *C. roseus* genome (Li *et al.*, 2023b) and detected 92 MYB domain proteins (Fig. S10). We aligned the MYB domains from MYB TFs to produce a phylogeny that includes *C. roseus*, the model species *Arabidopsis thaliana*, *Solanum lycopersicum* (tomato), and *Solanum tuberosum* (potato) MYBs (Fig. S10). Tomato and potato MYBs were included to distinguish Solanaceae-specific MYBs against Asterid-specific (encompassing Apocynaceae species that include *C. roseus* and Solanaceae species) MYBs. We found that the three *IDM* candidates were not closely related to each other (Fig. S10). Their MYB domains are more similar to MYB TFs in other species than to each other, although they share the same expression pattern. Due to low bootstrap support for some nodes in the phylogeny, it was impossible to definitively identify the ortholog of each *IDM* in Arabidopsis.

IDM1 directly activates the expression of *D4H* and *DAT*

To test the functions of *IDW1* and *IDM1/2/3*, we performed overexpression assays followed by RNA-seq to investigate whether overexpression of these TFs affects the expression of the MIA biosynthetic pathway. Coding sequences of *IDW1* and *IDM1/2/3* were cloned downstream of the 35S promoter. The overexpression vectors were transformed into *Agrobacterium tumefaciens* and infiltrated into *C. roseus* petals. Petals were chosen as the model for transient expression for the following reasons: (1) *C. roseus* petals are much more amendable to agrobacterium-mediated transient expression than leaves, and a highly efficient transient expression protocol has been established for petals (Colinas *et al.*, 2021). We can reliably achieve transient expression in petals, but not in leaves, despite extensive optimization. (2) Petals are metabolically active. A recent study demonstrated that heterologous expression of *Tabernanthe iboga* biosynthetic genes in *C. roseus* petals resulted in the accumulation of *T. iboga* MIAs (Kamileen *et al.*, 2024). (3) Petals, similar to leaves, contain idioblast cells (Fig. S11). Much like the leaf, petal idioblast cells also accumulate vindoline, anhydrovinblastine, and vinblastine (Vu *et al.*, 2024). Reanalysis of a recently published petal single-cell RNA-seq dataset (Vu *et al.*, 2024) revealed the presence of a rare cell type in the petal with the characteristics of idioblast (Fig. S11a,b). Notably, late-stage MIA genes, namely *NMT*, *D4H*, *DAT*, and *THAS2*, were all strongly upregulated in the petal idioblast (Fig. S10b). Among 16 709 genes expressed in both leaf and petal, 147 (or 0.8%) were detected as leaf idioblast

marker genes and 106 were detected as petal idioblast marker genes (see the Materials and Methods section). A total of 33 genes (14.7% of the union set) were shared among leaf and petal idioblast marker genes (Fig. S11c), representing a strong enrichment in the intersect (18.4-fold over background, $P < 2.2 \times 10^{-16}$, χ-squared test). Finally, *IDM1/2/3* and *IDW1*

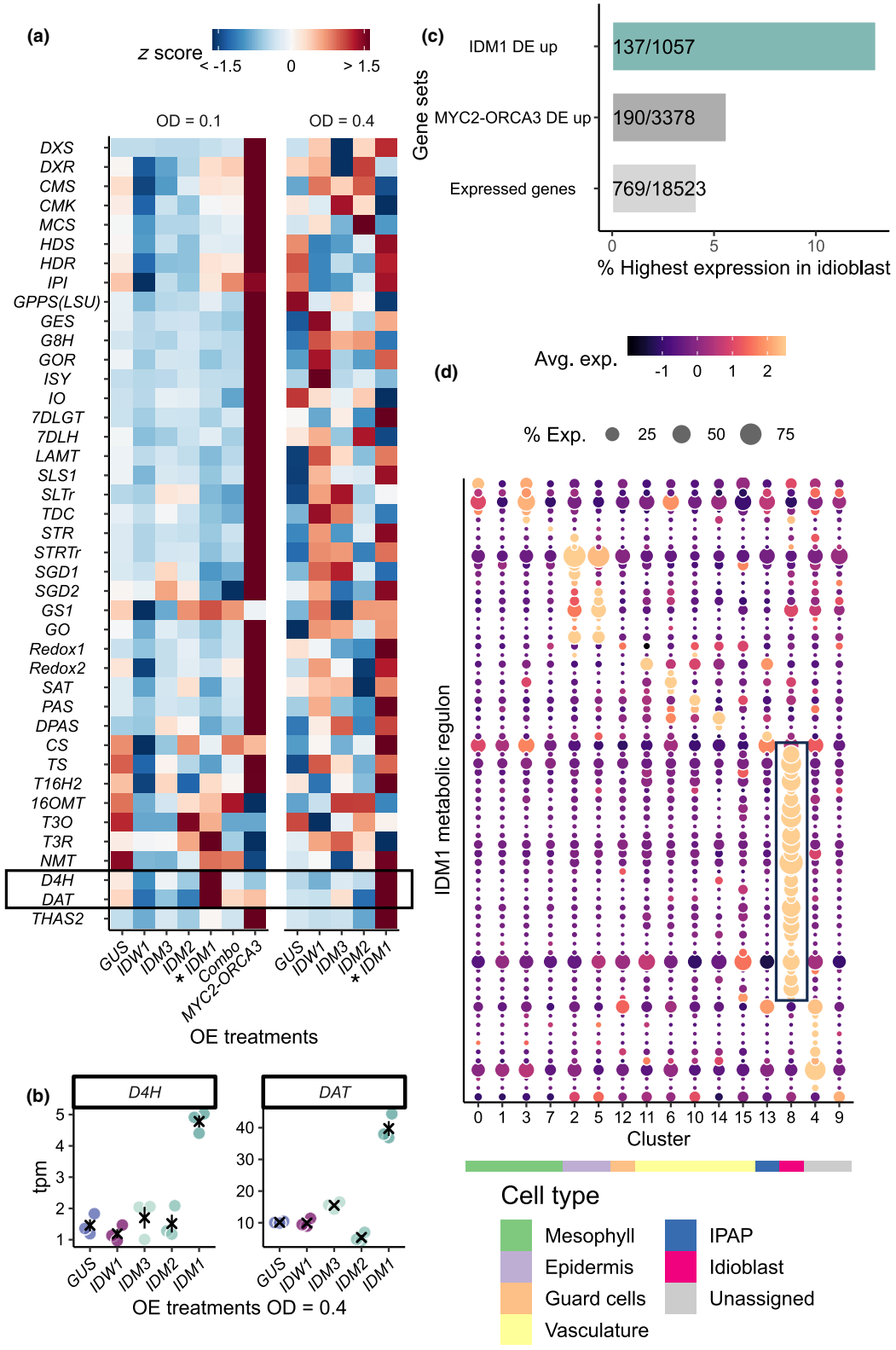

Fig. 5 Idioblast MYB1 (IDM1) activates the expression of *D4H* and *DAT*, as well as an idioblast-specific transcriptional program in *Catharanthus roseus*. (a) Gene expression heatmap of the monoterpene indole alkaloid biosynthetic genes across overexpression treatments. Each row is a biosynthetic gene or transporter, ordered from upstream to downstream. Color scale represents scaled expression (z-score). Combo: the combinatory treatment in which *IDW1* and *IDM1/2/3* are co-infiltrated. (b) Mean separation plots showing expression levels of *D4H* and *DAT* (in units of transcripts per million) in the 0.4 optical density treatments. Each data point is a biological replicate. Error bars represent average and SE. Black × indicates average. (c) Bar graph showing percentage of genes that are most highly expressed in the idioblast. Expressed genes, all 18 523 expressed genes in this single-cell multiome dataset; *MYC2-ORCA3*, 3378 differentially expressed genes that are upregulated in the *MYC2-ORCA3* overexpression treatment; *IDM1*, 1057 differentially expressed genes that are upregulated in the 0.4 OD overexpression *IDM1* treatment. (d) Gene expression heatmap of *IDM1* metabolic regulon. Color scale shows the average scaled expression of each gene at each cell cluster. Dot size indicates the percentage of cells where a given gene is detected. The predicted cell type for each cell cluster is annotated by the color strip below the *x*-axis. Box highlights genes specifically expressed in the idioblast.

all exhibited notable idioblast expression and were co-expressed with late-stage MIA genes, with *IDM1* detected most specifically in the petal idioblast. Thus, the petal is a suitable organ for functional genomics experiments including idioblast-specific regulation.

As infiltrating agrobacterium affects the expression of the pathway, we used GUS as a negative control. As a positive control, an engineered *MYC2* TF (Schweizer *et al.*, 2018) and *ORCA3* (van der Fits & Memelink, 2000) were co-infiltrated, which have been previously shown to strongly activate the MIA pathway (Schweizer *et al.*, 2018). The *MYC2* coding sequence was previously engineered to carry the D126N mutation, such that it could no longer be post-translationally repressed by the JAZ repressor protein. A combined overexpression treatment of *IDW1* and *IDM1/2/3* was also tested; a total of seven treatments including controls were assayed.

Using triplicated overexpression treatments at 0.1 OD (Figs S11e, S12a; Table S2), we found that the *MYC2-ORCA3*-positive control strongly activates the MIA pathway up to the DPAS step (Fig. 5a), consistent with previous reports that these known regulators do not activate late-stage biosynthetic genes downstream of *Catharanthine Synthase* (Fig. 2d) (Schweizer *et al.*, 2018). We discovered that one of the MYB candidates, *IDM1*, activated the expression of both *D4H* and *DAT* (Fig. 5a), resulting in a 2.8-fold and 1.68-fold increase in expression relative to the GUS control, respectively. All other overexpression treatments, including the combination of all candidates, did not activate the pathway relative to the GUS control (Fig. 5a). Encouraged by the initial result for *IDM1*, we examined gene expression profiles at 0.4 OD (Fig. S12b; Table S2). To control for batch-to-batch variation between experiments, independent GUS controls were included across both 0.1 OD and 0.4 OD experiments (Fig. 5a). We found that *IDM1* continued to activate both *D4H* and *DAT* at 0.4 OD (Fig. 5a,b), resulting in even higher fold changes (3.2-fold and 3.9-fold increase relative to GUS control of the corresponding experiment, respectively).

In addition to *D4H* and *DAT*, we found that genes differentially upregulated by IDM1 were enriched for idioblast expression (Fig. 5c). Among all 18 523 expressed genes in the single-cell multiome dataset, only 769 (4% of 18 523) were most highly expressed in the idioblast. Similarly, among 3378 differentially upregulated genes in the *MYC2-ORCA3* treatment, 190 (5.6% of 18 523) were most highly expressed in the idioblast. I contrast, among 1057 differentially upregulated genes in the *IDM1* treatment at 0.4 OD, 137 (13% of 1057) were most highly expressed in the idioblast, representing a 3.3-fold enrichment over the background of all expressed genes ($P < 2.2 \times 10^{-16}$, χ-squared test). IDM1 also activated *IDW1* and *IDM2/3*, the three other idioblast-specific WRKY and MYB TF candidates described above (Fig. S12c). Gene set enrichment analyses revealed that, similar to *MYC2-ORCA3*, *IDM1* upregulated genes were enriched for gene families relevant to specialized metabolism (transporters, cytochrome P450s, alcohol dehydrogenases, and 2-OG-dependent oxygenases). For example, among all expressed genes, 0.27% of them are annotated as alcohol dehydrogenases, whereas 0.65% and 1.3% of *MYC2-ORCA3* and *IDM1* upregulated genes were annotated as alcohol dehydrogenase, respectively. These *IDM1* upregulated genes that are potentially relevant to specialized metabolism were designated as the *IDM1* metabolic regulon ($n = 61$ genes, Table S11). We found that 44% (27/61) of the *IDM1* metabolic regulon were specifically expressed in the idioblast (Fig. 5d), more than 10-fold enrichment over the background of all expressed genes (background = 4% of 18 523 expressed genes, $P < 2.2 \times 10^{-16}$, χ-squared test). Taken together, these observations suggest *IDM1* regulates an idioblast-specific transcriptional program, which includes the MIA biosynthetic genes *D4H* and *DAT*, as well as other gene families potentially involved in natural product biosynthesis.

We next tested whether IDM1 could transactivate the promoters of *D4H* and *DAT* using reporter transactivation assays in *C. roseus* (Figs 6, S13). We first confirmed that IDM1 is localized to the nucleus (Fig. S13a–d). To construct reporters, we fused accessible chromatin regions upstream of *DAT* (Fig. 3b) and *D4H* (Fig. S13e) to a minimal 35S promoter driving an DsRed reporter. On the same plasmid, a GFP internal control was included, which is driven by the constitutive Arabidopsis UBQ1 promoter (Fig. S13f). We then performed the transactivation assay by co-infiltrating an agrobacterium strain carrying 35S:*IDM1* (Fig. S13g) and a strain carrying the reporter construct for either *DAT* or *D4H*. We observed conspicuous $DsRed^+$ cells in infiltrated petals for both *DAT* and *D4H* reporters (Fig. 6a,b,e,f). By contrast, no $DsRed^+$ cells in petals could be observed when either reporter was infiltrated alone (Fig. 6c,d,g,h). These observations were confirmed by pixel intensity quantifications. High red-to-green pixel intensity ratio was only detected when 35S:*IDM1* and one of *DAT* or *D4H* reporters were co-infiltrated (Fig. 6i). By contrast, low red-to-green ratio was detected when the reporter was infiltrated without 35S:*IDM1*. *IMD1* could not transactivate a reporter construct that did not contain the *DAT* or *D4H* accessible

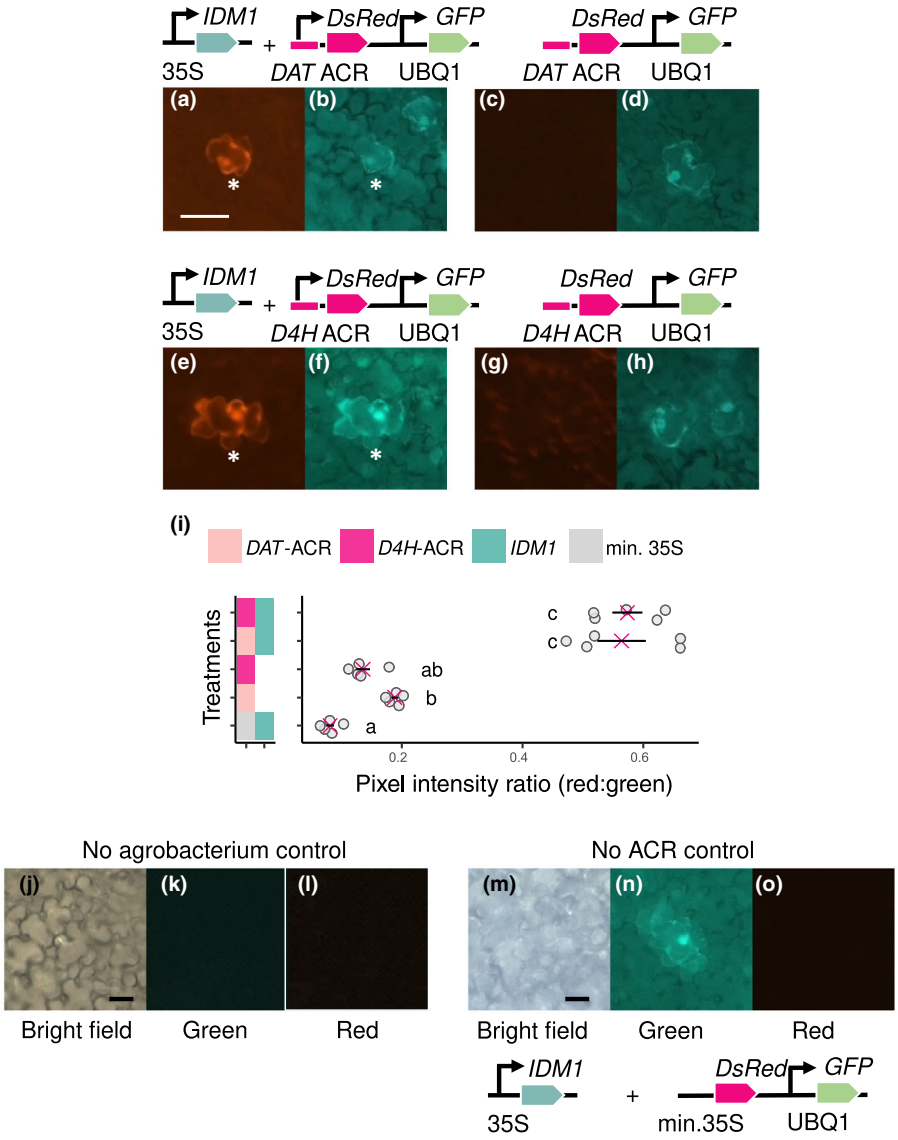

Fig. 6 Transactivation assays for *IDM1* against DAT and D4H promoters in *Catharanthus roseus* petals. (a, b) Reporter transactivation when agrobacterium strain carrying 35S:*IDM1* and *DAT* reporter are co-infiltrated. (c, d) Control samples when only the agrobacterium strain carrying the *DAT* reporter is infiltrated. (e, f) Reporter transactivation when agrobacterium strain carrying 35S:*IDM1* and *D4H* reporter are co-infiltrated. (g, h) Control samples when only the agrobacterium strain carrying the *D4H* reporter is infiltrated. (i) Quantification of microscope pixel intensity ratios. Color boxes on the right represent agrobacterium strains infiltrated. The first column indicates the reporter constructs (*DAT*, *D4H*, or minimal 35S promoter control). The second column indicates whether 35S:*IDM1* is co-infiltrated. Error bars represent average and SE. Pink × indicates average. (j–l) Microscope images of petals infiltrated with infiltration buffer (no agrobacterium control). No fluorescent signal can be detected in either the green or red channel. (m–o) Microscope images of petals co-infiltrated with 35S:*IDM1* and a control reporter without accessible chromatin region. No fluorescent signal can be detected in the red channel. Bars, 20 μm. ACR, accessible chromatin region.

chromatin regions (Fig. 6j–o), which was confirmed by pixel intensity quantifications (Fig. 6i). Taken together, these results suggest that the idioblast-specific expression of *IDM1* contributes to the idioblast-specific expression of *D4H* and *DAT*.

Discussion

The cell-type-specific expression patterns of MIA biosynthetic genes in *C. roseus* are well-documented (Guirimand *et al.*, 2011; Miettinen *et al.*, 2014; Li *et al.*, 2023b). In this study, using single-cell multi-omics datasets, we discovered the first reported idioblast-specific TF (*CrIDM1*) that regulates late-stage vinblastine biosynthetic genes (*D4H* and *DAT*). Although several TFs that regulate MIA biosynthesis have been characterized (Menke, 1999; van der Fits, 2000; van der Fits & Memelink, 2000; Peebles *et al.*, 2009; Suttipanta *et al.*, 2011; Zhang *et al.*, 2011; Van Moerkercke *et al.*, 2015, 2016; Bahieldin *et al.*, 2018; Schweizer *et al.*, 2018; Liu *et al.*, 2019; Singh *et al.*, 2021), how the exquisite cell-type-specific regulation is achieved for this pathway remains unclear. We generated the first single-cell multiome dataset for *C. roseus* leaves to investigate gene regulation of the MIA pathway at single-cell resolution. Not only did we recapitulate the cell-type-specific expression pattern of the pathway, but also we catalogued a dictionary of *cis*-regulatory elements associated with MIA biosynthetic genes. In this experiment, we not only captured the major cell types of the leaf (mesophyll, epidermis, and vasculature) but also the rare cell types (IPAP and idioblast) associated with MIA biosynthesis. We also noted that the single-nuclei RNA-seq assay is more powerful in distinguishing different cell types compared with the single-nuclei ATAC-seq array (Fig. 1c), since the majority of the ATAC-seq peaks are accessible across multiple cell types, if not all cell types. This is a known phenomenon for plants, based on single-cell ATAC-seq datasets from multiple species, including

maize (Marand *et al.*, 2021), rice (Yan *et al.*, 2024), and soybean (Zhang *et al.*, 2024).

We showed that among previously studied TFs pertinent to the MIA pathway, only *BIS1/2/3* and *ORCA4* were co-expressed with their target genes at the cell-type level (Figs 2d, S8a), suggesting *BIS1/2/3* and *ORCA4* contribute to the cell-type-specific expression pattern of the MIA biosynthetic pathway. The late-stage MIA biosynthetic genes were marked with cell-type-specific ATAC-seq peaks, suggestive of coordinated regulation at the chromatin level (Fig. 3a). Epidermis marker peaks (Fig. 3b) were enriched for homeodomain, ERF-, and MYB-binding motifs (Fig. S9a). Members of the abovementioned TF families have been reported to regulate other specialized metabolism pathways, such as anthocyanin (Kubo *et al.*, 1999), cuticle (Skaliter *et al.*, 2023), suberin (Cantó-Pastor *et al.*, 2024), and glucosinolate (Sønderby *et al.*, 2010) in other species. We speculate that yet unidentified homeodomain, ERF, and MYB TFs may contribute to the cell-type-specific expression of MIA biosynthetic genes in epidermis. The dataset generated in this study can be used to mine and characterize additional metabolic regulators that operate specifically in the epidermis.

We found that WRKY and MYB motifs were overrepresented among idioblast marker peaks (Fig. 3c). Paired with gene co-expression analyses across cell clusters, we narrowed down our candidates to a single WRKY (*IDW1*) and three MYB TFs (*IDM1/2/3*) (Fig. 4). While candidate TFs can be identified from gene expression data alone (Li *et al.*, 2023b; Guedes *et al.*, 2024), we demonstrated that cell-type-specific chromatin accessibility profiles allowed us to identify putative cell-type-specific *cis*-regulatory elements and the corresponding TF families using motif enrichment, which in turn pinpointed TF candidates that most likely activate target genes in a cell-type-specific manner.

Overexpression and reporter transactivation assays demonstrated that *IDM1* is a novel idioblast-specific regulator for *D4H* and *DAT* (Fig. 5). Recently, a GATA family TF, *GATA1*, was reported to activate the expression of late vinblastine biosynthetic genes in de-etiolating seedlings, including *T16H2*, *T3O*, *T3R*, *D4H*, and *DAT* (Liu *et al.*, 2019). However, we found that *GATA1* was only expressed in the mesophyll of the leaf in our single-cell dataset (Fig. S8a), suggesting that in the baseline (or unelicited) state of leaf, *GATA1* is likely not responsible for the cell-type-specific patterns of the late-stage pathway. However, it is possible that *GATA1* does exhibit cell-type-specific patterns in de-etiolating seedlings. In comparison, *IDM1* is expressed exclusively in the idioblast, and thus, it contributes to the idioblast-specific expression of *D4H* and *DAT*. Since *IDM1* is also JA-inducible (Fig. S9c), *IDM1* may also mediate JA-dependent activation of *D4H* and *DAT*.

In addition to *D4H* and *DAT*, *IDM1* activates an idioblast metabolic regulon (Fig. 5c,d). Gene sets such as transporters, cytochrome P450, alcohol dehydrogenase, and 2-OG-dependent oxygenase are strongly enriched in IDM1 upregulated genes, suggesting that *IDM1* is a *bona fide* metabolic regulator. The *IDM1* metabolic regulon is highly enriched for idioblast-specific expression (Fig. 5d), suggesting other targets of *IDM1* may play a role in the biosynthesis of vinblastine or other alkaloids in the idioblast. *IDM1* activates *IDW1* and *IDM2/3*, which did not appear to activate the MIA pathway, at least in the experimental conditions we tested (Fig. 5a). *IDW1* and *IDM2/3* might regulate other biological processes in the idioblast, which may be important for the specialization of these rare cells.

Mechanisms of gene regulation in rare cell types associated with specialized metabolism are not well-understood. The identification of *IDM1* will provide new opportunities for future investigations on idioblast development and function. Insights into the differentiation and function of rare biosynthetic cells may contribute to studies on the evolution of plant-specialized metabolism and cloning of biosynthetic pathways, in addition to engineering or breeding plants with enhanced biosynthetic capabilities. Even after decades of focused research, the final steps of the *C. roseus* MIA biosynthetic pathway remains an enigma. The discovery of *IDM1* as a regulator of the late stages of MIA biosynthesis and access to an idioblast-specific gene regulatory network will expedite completion of this 40-plus step biosynthetic pathway with important human-health implications.

Acknowledgements

This project was supported by the University of Georgia, Georgia Research Alliance (CRB) and Georgia Seed Development (CRB), National Science Foundation MCB-2309665 (CRB and CL), and the Max Planck Gesellschaft (SEOC and LC). Sequencing was performed at Biomarker Technologies (BMK) GmbH, Max-Planck Institute for Biochemistry Next Generation Sequencing Facility, Novogene, Texas A&M AgriLife Research: Genomics and Bioinformatics Service, and University of Georgia Genomics and Bioinformatics Core (GGBC, UG Athens, GA, RRID:SCR_010994). The author would like to acknowledge Julie Nelson at University of Georgia Cytometry Shared Resource Laboratory for assistance with flow cytometry, Dr Alexandre P. Marand for suggestions on nuclei isolation protocol development, Dr Robert J. Schmitz for advice on ATAC-seq analyses, and Dr Alain Goossens for providing plasmids.

Competing interests

None declared.

Author contributions

CRB, SEOC and CL designed the study. CL generated single-cell multiome and DAP-seq datasets. JCW assisted with single-cell library preparation and quality control. CL and SLJ performed molecular cloning and transactivation assays. MC performed molecular cloning and overexpression assays, and together with LC generated overexpression samples and RNA-seq datasets. CL, JCW, BV and JPH performed data analyses. CL wrote the manuscript with input from all authors.

ORCID

C. Robin Buell https://orcid.org/0000-0002-6727-4677

Lorenzo Caputi https://orcid.org/0000-0002-7783-6733
Maite Colinas https://orcid.org/0000-0001-7053-2983
John P. Hamilton https://orcid.org/0000-0002-8682-5526
Chenxin Li https://orcid.org/0000-0002-6276-7712
Sarah E. O'Connor https://orcid.org/0000-0003-0356-6213
Brieanne Vaillancourt https://orcid.org/0000-0002-6795-5173
Joshua C. Wood https://orcid.org/0000-0002-7691-6088

Data availability

All sequencing data associated with this study are available at the National Center for Biotechnology Institute Sequence Read Archive BioProject PRJNA1098712. Seurat objects for single-nuclei multiome experiment and gene expression matrices are available via the online digital repository figshare (doi: 10.6084/m9.figshare.25586979). Plasmid maps and sequences of MYB domains are available at Zenodo (https://zenodo.org/records/13366861). All custom codes used to generate figures can be found at https://github.com/cxli233/Catharanthus_multiome.

References

Bahieldin A, Atef A, Edris S, Gadalla NO, Al-matary M, Al-Kordy MA, Ramadan AM, Bafeel S, Alharbi MG, Al-Quwaie DAH *et al.* 2018. Stepwise response of MeJA-induced genes and pathways in leaves of *C. roseus*. *Comptes Rendus Biologies* **341**: 411–420.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**: W202–W208.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-Seq quantification. *Nature Biotechnology* **34**: 525–527.

Brown RL, Kazan K, McGrath KC, Maclean DJ, Manners JM. 2003. A role for the GCC-Box in jasmonate-mediated activation of the *PDF1.2* gene of Arabidopsis. *Plant Physiology* **132**: 1020–1032.

Burlat V, Oudin A, Courtois M, Rideau M, St-Pierre B. 2004. 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. *The Plant Journal* **38**: 131–141.

Cantó-Pastor A, Kajala K, Shaar-Moshe L, Manzano C, Timilsena P, De Bellis D, Gray S *et al.* 2024. A suberized exodermis is required for tomato drought tolerance. *Nature Plants* **10**: 118–130.

Chamness JC, Kumar J, Cruz AJ, Rhuby E, Holum MJ, Cody JP, Tibebu R, Gamo ME, Starker CG, Zhang F *et al.* 2023. An extensible vector toolkit and parts library for advanced engineering of plant genomes. *The Plant Genome* **16**: e20312.

Colinas M, Goossens A. 2022. Transient gene expression in *Catharanthus roseus* flower petals using agroinfiltration. In: Courdavault V, Besseau S, eds. Catharanthus roseus. *Methods in molecular biology, vol. 2505*. New York, NY, USA: Springer US, 281–291.

Colinas M, Pollier J, Vaneechoutte D, Malat DG, Schweizer F, De Milde L, De Clercq R *et al.* 2021. Subfunctionalization of paralog transcription factors contributes to regulation of alkaloid pathway branch choice in *Catharanthus roseus*. *Frontiers in Plant Science* **12**: 687406.

Csardi G, Nepusz T. 2006. *The IGRAPH software package for complex network research*, 9.

van der Fits L. 2000. A *Catharanthus roseus* BPF-1 homologue interacts with an elicitor-responsive region of the secondary metabolite biosynthetic gene Str and is induced by elicitor via a JA-independent signal transduction pathway. *Plant Molecular Biology* **44**: 675–685.

van der Fits L, Memelink J. 2000. ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science* **289**: 295–297.

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D *et al.* 2019. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48**: gkz1001.

Fujimoto SY, Ohta M, Usui A, Shinshi H, Ohme-Takagi M. 2000. Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. *Plant Cell* **12**: 393–404.

Guedes JG, Ribeiro R, Carqueijeiro I, Guimarães AL, Bispo C, Archer J, Azevedo H, Fonseca NA, Sottomayor M. 2024. The leaf idioblastome of the medicinal plant *Catharanthus roseus* is associated with stress resistance and alkaloid metabolism. *Journal of Experimental Botany* **75**: 274–299.

Guirimand G, Guihur A, Poutrain P, Héricourt F, Mahroug S, St-Pierre B, Burlat V, Courdavault V. 2011. Spatial organization of the vindoline biosynthetic pathway in *Catharanthus roseus*. *Journal of Plant Physiology* **168**: 549–557.

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M *et al.* 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573–3587.

Jacobowitz JR, Weng J-K. 2020. Exploring uncharted territories of plant specialized metabolism in the postgenomic era. *Annual Review of Plant Biology* **71**: 631–658.

Kamileen MO, Nakamura Y, Luck K, Heinicke S, Hong B, Colinas M, Lichman BR, O'Connor SE. 2024. Streamlined screening platforms lead to the discovery of pachysiphine synthase from *Tabernanthe iboga*. *New Phytologist*. doi: 10.1101/2024.06.30.601415.

Kaminow B, Yunusov D, Dobin A. 2021. STARSOLO: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*. doi: 10.1101/2021.05.05.442755.

Katoh K. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**: 3059–3066.

Kellner F, Kim J, Clavijo BJ, Hamilton JP, Childs KL, Vaillancourt B, Cepela J, Habermann M, Steuernagel B, Clissold L *et al.* 2015. Genome-guided investigation of plant natural product biosynthesis. *The Plant Journal* **82**: 680–692.

Kim J-Y, Symeonidi E, Pang TY, Denyer T, Weidauer D, Bezrutczyk M, Miras M, Zöllner N, Hartwig T, Wudick MM *et al.* 2021. Distinct identities of leaf phloem cells revealed by single cell transcriptomics. *Plant Cell* **33**: 511–530.

Kim RJ, Lee SB, Pandey G, Suh MC. 2022. Functional conservation of an AP2/ERF transcription factor in cuticle formation suggests an important role in the terrestrialization of early land plants. *Journal of Experimental Botany* **73**: 7450–7466.

Kubo H, Peeters AJM, Aarts MGM, Pereira A, Koornneef M. 1999. ANTHOCYANINLESS2, a homeobox gene affecting anthocyanin distribution and root development in Arabidopsis. *Plant Physiology* **121**: 399–410.

Li C, Deans NC, Robin Buell C. 2023a. 'Simple Tidy GeneCoEx': a gene co-expression analysis workflow powered by tidyverse and graph-based clustering in R. *The Plant Genome* **16**: e20323.

Li C, Wood JC, Anh Hai V, Hamilton JP, Lopez CER, Payne RME, Guerrero DAS, Gase K, Yamamoto K, Vaillancourt B *et al.* 2023b. Single-cell multi-omics in the medicinal plant *Catharanthus roseus*. *Nature Chemical Biology* **19**: 1031–1041.

Li C, Wood JC, Deans NC, Jarrell AF, Martin D, Mailloux K, Wang Y-W, Robin Buell C. 2022. Nuclei isolation protocol from diverse angiosperm species. *bioRxiv*. doi: 10.1101/2022.11.03.515090.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. http://arxiv.org/abs/1303.3997.

Liu Y, Patra B, Pattanaik S, Wang Y, Yuan L. 2019. GATA and phytochrome interacting factor transcription factors regulate light-induced vindoline biosynthesis in *Catharanthus roseus*. *Plant Physiology* **180**: 1336–1350.

Lopez-Anido CB, Vatén A, Smoot NK, Sharma N, Guo V, Yan Gong M, Gil XA, Weimer AK, Bergmann DC. 2021. Single-cell resolution of lineage

trajectories in the Arabidopsis stomatal lineage and developing leaf. *Developmental Cell* **56**: 1043–1055.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.

Marand AP, Chen Z, Gallavotti A, Schmitz RJ. 2021. A *cis*-regulatory atlas in maize at single-cell resolution. *Cell* **184**: 3041–3055.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**: 3.

Menke FLH. 1999. A novel jasmonate- and elicitor-responsive element in the periwinkle secondary metabolite biosynthetic gene Str interacts with a jasmonate- and elicitor-inducible AP2-domain transcription factor, ORCA2. *EMBO Journal* **18**: 4455–4463.

Miettinen K, Dong L, Navrot N, Schneider T, Burlat V, Pollier J, Woittiez L, van der Krol S, Lugan R, Ilc T *et al.* 2014. The seco-iridoid pathway from *Catharanthus roseus*. *Nature Communications* **5**: 3606.

O'Connor SE, Maresh JJ. 2006. Chemistry and biology of monoterpene indole alkaloid biosynthesis. *Natural Product Reports* **23**: 532–547.

O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. 2016. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **165**: 1280–1292.

Onoyovwe A, Hagel JM, Chen X, Khan MF, Schriemer DC, Facchini PJ. 2013. Morphine biosynthesis in opium poppy involves two cell types: sieve elements and laticifers. *Plant Cell* **25**: 4110–4122.

Paul P, Singh SK, Patra B, Sui X, Pattanaik S, Yuan L. 2017. A differentially regulated AP 2/ERF transcription factor gene cluster acts downstream of a MAP kinase cascade to modulate terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *New Phytologist* **213**: 1107–1123.

Peebles CAM, Hughes EH, Shanks JV, San K-Y. 2009. Transcriptional response of the terpenoid indole alkaloid pathway to the overexpression of ORCA3 along with jasmonic acid elicitation of *Catharanthus roseus* hairy roots over time. *Metabolic Engineering* **11**: 76–86.

Pucker B. 2022. Automatic identification and annotation of MYB gene family members in plants. *BMC Genomics* **23**: 220.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* **42**: W187–W191.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH image to ImageJ: 25 years of image analysis. *Nature Methods* **9**: 671–675.

Schweizer F, Colinas M, Pollier J, Van Moerkercke A, Vanden Bossche R, de Clercq R, Goossens A. 2018. An engineered combinatorial module of transcription factors boosts production of monoterpenoid indole alkaloids in *Catharanthus roseus*. *Metabolic Engineering* **48**: 150–162.

Shen W, Le S, Li Y, Fuquan H. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**: e0163962.

Simkin AJ, Miettinen K, Claudel P, Burlat V, Guirimand G, Courdavault V, Papon N, Meyer S, Godet S, St-Pierre B *et al.* 2013. Characterization of the plastidial geraniol synthase from Madagascar Periwinkle which initiates the monoterpenoid branch of the alkaloid pathway in internal phloem associated parenchyma. *Phytochemistry* **85**: 36–43.

Singh SK, Patra B, Paul P, Liu Y, Pattanaik S, Yuan L. 2021. *BHLH IRIDOID SYNTHESIS 3* is a member of a bHLH gene cluster regulating terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Direct* **5**: e00305.

Skaliter O, Bednarczyk D, Shor E, Shklarman E, Manasherova E, Aravena-Calvo J, Kerzner S, Cna'ani A, Jasinska W, Masci T *et al.* 2023. The R2R3-MYB transcription factor EVER controls the emission of petunia floral volatiles by regulating epicuticular wax biosynthesis in the petal epidermis. *Plant Cell* **36**: 174–193.

Sønderby IE, Burow M, Rowe HC, Kliebenstein DJ, Halkier BA. 2010. A complex interplay of three R2R3 MYB transcription factors determines the profile of aliphatic glucosinolates in Arabidopsis. *Plant Physiology* **153**: 348–363.

Stamatakis A. 2014. RAxML v.8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. 2021. Single-cell chromatin state analysis with Signac. *Nature Methods* **18**: 1333–1341.

Suttipanta N, Pattanaik S, Kulshrestha M, Patra B, Singh SK, Yuan L. 2011. The transcription factor CrWRKY1 positively regulates the terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Physiology* **157**: 2081–2093.

Tenorio Berrío R, Verstaen K, Vandamme N, Pevernagie J, Achon I, Van Duyse J, Van Isterdael G, Saeys Y, De Veylder L, Inzé D *et al.* 2022. Single-cell transcriptomics sheds light on the identity and metabolism of developing leaf cells. *Plant Physiology* **188**: 898–918.

Tominaga R, Iwata M, Okada K, Wada T. 2007. Functional analysis of the epidermal-specific MYB genes *CAPRICE* and *WEREWOLF* in *Arabidopsis*. *Plant Cell* **19**: 2264–2277.

Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**: 5233.

Van Moerkercke A, Fabris M, Pollier J, Baart GJ, Rombauts S, Hasnain G, Rischer H, Memelink J, Oksman-Caldentey KM, Goossens A. 2013. CathaCyc, a metabolic pathway database built from *Catharanthus roseus* RNA-seq data. *Plant and Cell Physiology* **54**: 673–685.

Van Moerkercke A, Steensma P, Gariboldi I, Espoz J, Purnama PC, Schweizer F, Miettinen K, Vanden Bossche R, De Clercq R, Memelink J *et al.* 2016. The basic helix-loop-helix transcription factor BIS2 is essential for monoterpenoid indole alkaloid production in the medicinal plant *Catharanthus roseus*. *The Plant Journal* **88**: 3–12.

Van Moerkercke A, Steensma P, Schweizer F, Pollier J, Gariboldi I, Payne R, Vanden Bossche R, Miettinen K, Espoz J, Purnama PC *et al.* 2015. The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in *Catharanthus roseus*. *Proceedings of the National Academy of Sciences, USA* **112**: 8130–8135.

Vu AH, Kang M, Wurlitzer J, Heinicke S, Li C, Wood JC, Veit G, Robin Buell C, Caputi L, O'Connor SE. 2024. Quantitative single cell mass spectrometry reveals the dynamics of plant natural product biosynthesis. *bioRxiv*. doi: 10.1101/2024.04.23.590720.

Weng J-K, Lynch JH, Matos JO, Dudareva N. 2021. Adaptive mechanisms of plant specialized metabolism connecting chemistry to function. *Nature Chemical Biology* **17**: 1037–1045.

Wu S, Morotti ALM, Yang J, Wang E, Tatsis EC. 2024. Single cell RNA sequencing facilitates the elucidation of the complete biosynthesis of the antidepressant hyperforin in St. John's Wort. *Molecular Plant* **17**: 1439–1457.

Yamamoto K, Grzech D, Koudounas K, Stander EA, Caputi L, Mimura T, Courdavault V, O'Connor SE. 2021. Improved virus-induced gene silencing allows discovery of a serpentine synthase gene in *Catharanthus roseus*. *Plant Physiology* **187**: 846–857.

Yamamoto K, Takahashi K, Caputi L, Mizuno H, Rodriguez-Lopez CE, Iwasaki T, Ishizaki K, Fukaki H, Ohnishi M, Yamazaki M *et al.* 2019. The complexity of intercellular localisation of alkaloids revealed by single-cell metabolomics. *New Phytologist* **224**: 848–859.

Yamamoto K, Takahashi K, Mizuno H, Anegawa A, Ishizaki K, Fukaki H, Ohnishi M, Yamazaki M, Masujima T, Mimura T. 2016. Cell-specific localization of alkaloids in *Catharanthus roseus* stem tissue measured with imaging MS and single-cell MS. *Proceedings of the National Academy of Sciences, USA* **113**: 3891–3896.

Yan H, Mendieta JP, Zhang X, Marand AP, Liang Y, Luo Z, Minow MAA, Roulé T, Wagner D, Tu X *et al.* 2024. Evolution of plant cell-type-specific *cis*-regulatory elements. *bioRxiv*. doi: 10.1101/2024.01.08.574753.

Zhang H, Hedhili S, Montiel G, Zhang Y, Chatel G, Pré M, Gantet P, Memelink J. 2011. The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis in *Catharanthus roseus*: CrMYC2 controls JA-responsive ORCA gene expression. *The Plant Journal* **67**: 61–71.

Zhang X, Luo Z, Marand AP, Yan H, Jang H, Bang S, Mendieta JP, Minow MAA, Schmitz RJ. 2024. A spatially resolved multiomic single-cell atlas of soybean development. *bioRxiv*. doi: 10.1101/2024.07.03.601616.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al.* 2008. Model-based analysis of ChIP-seq (MACS). *Genome Biology* **9**: R137.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Dataset S1 Gene expression matrix (in unit of logCPM) across cell clusters.

Dataset S2 Gene expression matrix (Z scores) across cell clusters.

Fig. S1 Schematics of the pathway and isolated nuclei.

Fig. S2 Evaluation and quality control for the gene expression assay of single-cell multiome dataset.

Fig. S3 Cell clustering and cell-type marker gene analyses.

Fig. S4 Heterogeneity of idioblast cells.

Fig. S5 Evaluation of chromatin accessibility profiles around genes.

Fig. S6 Evaluation of chromatin accessibility profiles around ATAC-seq peaks.

Fig. S7 Evaluation of DAP-seq peaks.

Fig. S8 Gene expression analyses regarding known transcription factors.

Fig. S9 Epidermis overrepresented motifs, gene co-expression modules, and idioblast candidate transcription factors.

Fig. S10 Phylogeny of *IDM* factors.

Fig. S11 Establishing petals as a model for functional genomics.

Fig. S12 Gain-of-function experiments on idioblast-specific transcription factor candidates using overexpression assay.

Fig. S13 Characterization of *IDM1*.

Table S1 Gene names and functions of monoterpenoid indole alkaloid biosynthetic genes.

Table S2 List of libraries and sequence data used in this study.

Table S3 Details of single-nuclei RNA-seq libraries.

Table S4 Marker genes used to assign cell types.

Table S5 Details of single nuclei ATAC-seq libraries.

Table S6 Details of DAP-seq libraries.

Table S7 Gene names and coordinates of transcription factors.

Table S8 Edge table of gene regulatory network.

Table S9 Genome coordinates of epidermis and idioblast marker peaks.

Table S10 Membership of inner phloem-associated parenchyma, epidermis, and idioblast co-expression module.

Table S11 Membership of IDM1 metabolic regulon.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.