Single cell omics extends metabolic regulon via orthologous transcription factors from a pair of medicinal plant species

Van-Hung Bui[1], Joshua C. Wood[2], Brieanne Vaillancourt[2], John P. Hamilton[2,3], Sophia L. Jones[2], Lemor H. Carlton[2], Anh Hai Vu[4], C. Robin Buell[2,3,5,6], Lorenzo Caputi[4*], Thu-Thuy T. Dang[1*], Chenxin Li[2,3*]

[1]Department of Chemistry, University of British Columbia, Okanagan, BC, Canada
[2]Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA
[3]Department of Crop and Soil Sciences, University of Georgia, Athens, GA, USA
[4]Department of Natural Product Biosynthesis, Max Planck Institute for Chemical Ecology, Jena, Germany
[5]Institute of Plant Breeding, Genetics, and Genomics, University of Georgia, Athens, Georgia, USA
[6]The Plant Center, University of Georgia, Athens, Georgia, USA

*Co-Corresponding authors

Abstract

The medicinal plant *Camptotheca acuminata* is a source of camptothecin, the starting material for several chemotherapeutic agents. Camptothecin is derived from the monoterpene indole alkaloid (MIA) strictosidinic acid, yet most biosynthetic enzymes downstream of strictosidinic acid are unknown. *Catharanthus roseus* is a medicinal plant that produces MIAs such as the anti-cancer agent vinblastine currently in use in the clinic. MIA biosynthesis in *C. roseus* is cell type specific, with the final stage of vinblastine biosynthesis restricted to the rare cell type idioblast. However, the cell type specificity of camptothecin biosynthesis is unknown. Through single cell multi-omics experiments, we detected rare MIA biosynthetic cells in *C. acuminata* leaf and stem. Metabolite feeding and single cell metabolomics identified stem as the active site for camptothecin biosynthesis. While early intermediates were detected in most cells, late-stage biosynthetic intermediates were detected specifically in a small number of cells within the stem, suggesting strict compartmentalization of the camptothecin biosynthetic pathway. We discovered MYB and bHLH transcription factors (TFs) co-expressed in the same cell type as biosynthetic genes in *C. acuminata* stem. Interestingly, the *C. roseus* orthologs of these TFs are idioblast specific and activated MIA biosynthetic genes in *C. roseus* upon overexpression. Integrating multi-omics and phylogenetic analyses, we generated an extended gene regulatory network for the idioblast metabolic regulon in *C. roseus*. This study demonstrates that the same clades of TFs have been co-opted to regulate cell type specific MIA biosynthesis across different species. Investigating cell type specific TFs that are co-expressed with biosynthetic genes across multiple species is a powerful strategy to increase the power for discovering plant metabolic regulators.

Introduction

Camptotheca acuminata Decne., known as 'happy tree', has been used to treat psoriasis and skin ulcers in Chinese traditional medicine (Lorence and Nessler 2004). *C. acuminata* contains about a hundred bioactive compounds, including 28 alkaloids, 14 triterpenoids, 11 ellagic acids among other flavonoids, sterols, and polyphenols (Lorence and Nessler 2004). Camptothecin, a natural product derived from monoterpene indole alkaloids (MIAs) was first extracted from the stems of *C. acuminata* in 1966 (Wall et al. 1966) and has been used since then as starting material to semi-synthesize many leading anticancer drugs, such as irinotecan (Camptosar®) and topotecan (Hycamtin®) (Lorence and Nessler 2004). Camptothecin and its derivatives are potent inhibitors of DNA topoisomerase I and are widely used for the treatment of lung, cervix, ovarian, and colon cancers (Giovanella et al. 1989). Traditional production strategies of camptothecin and derivatives based on recovery from natural sources and chemical synthesis present many challenges associated with the purity, scale, and complexity of the compounds, contributing to their rising costs and inaccessibility. Synthetic approaches rely on precursors from plants, which require more than fifteen steps with low yield and harsh chemical conditions (K. Li, Ou, and Gao 2016).

Camptothecin contains a highly conjugated pentacyclic skeleton with quinoline (A/B rings), indolizinone (C/D rings), and α-hydroxylactone moieties (E ring). Camptothecin biosynthesis in *C. acuminata* involves complicated catalytic steps, most of which remain undeciphered. In this pathway, tryptamine and secologanic acid, rather than secologanin like in the camptothecin pathway of *Ophiorrhiza pumila* (Yamazaki et al. 2003) and other MIAs (O'Connor and Maresh 2006), are coupled via a Pictet-Spengler reaction, leading to strictosidinic acid. The formation of strictosidinic acid is catalyzed by strictosidine/strictosidine acid synthase enzymes (STR). Intramolecular dehydrative cyclization of strictosidinic acid yields strictosamide. Conversion of the indole ring system to the quinoline ring is postulated to occur via multi-step oxidation to strictosamide ketolactam followed by condensation and water elimination to yield the quinolinone ring of pumiloside. Subsequent reduction to deoxypumiloside, deglycosylation, and other metabolic conversions lead to camptothecin (Fig. 1A) (Sadre et al. 2016). Currently, there are 12 enzymes that have been identified and characterized in *C. acuminata* that are involved in the camptothecin biosynthetic pathway or act on the camptothecin scaffold. Most enzymes that have been identified are involved in the early iridoid biosynthetic pathway or decorating the final camptothecin scaffold, including Isopentyl diphosphate isomerase (Pan et al. 2008), Geranyl diphosphate synthase (L. Yang et al. 2017), Geraniol synthase (Chen et al. 2016), Geraniol-10-hydroxylase (Sadre et al. 2016), 10-Hydroxygeraniol Oxidoreductase (Awadasseid et al. 2020), Iridoid synthase, Secologanic acid synthase (Y. Yang et al. 2019), Camptothecin hydroxylases (Nguyen et al. 2021), and 10-Hydroxycamptothecin *O*-methyl transferase (Salim, Jones, and DellaPenna 2018).

Like *C. acuminata*, *Catharanthus roseus* G. Don (Madagascar Periwinkle) also produces MIAs with clinical relevance, including vinblastine and vincristine (also known as vinca alkaloids) that are used to treat various cancers in the clinics (O'Connor and Maresh 2006). *C. acuminata* (with-

in the Nyssaceae family) and *C. roseus* (within the Apocynaceae family) shared a common ancestor ~115 million years ago (Kumar et al. 2017). The biosynthetic pathways of vinblastine and camptothecin are largely conserved up to *STR*, as both alkaloids originate from a strictosidine or strictosidinic acid scaffold. While *C. roseus* is primarily known as a medicinal plant, in recent years, it has become a model for studying how specialized metabolism is restricted to specific cell types (C. Li et al. 2023; 2025; Vu et al. 2024), since the MIA biosynthetic pathway exhibited strong cell type specificity in *C. roseus*. The early steps of the pathway, including the methylerythritol phosphate (MEP) pathway and iridoid biosynthetic genes, are specifically expressed in the internal phloem associated parenchyma (IPAP) (Burlat et al. 2004; Miettinen et al. 2014; Simkin et al. 2013). Subsequent biosynthetic steps occur in the epidermis (Burlat et al. 2004; Guirimand et al. 2011; Miettinen et al. 2014), while the final known steps of the pathway are restricted to a rare and specialized cell type termed idioblast (Guirimand et al. 2011; C. Li et al. 2023), which are interspersed among mesophyll cells (Yamamoto et al. 2019; 2016). Cell type specificity of this pathway is in part generated by cell type specific transcription factors (TFs). For instance, a group basic helix-loop-helix (bHLH) TFs termed "*bHLH Iridoid Synthesis*" (*BIS*) TFs (Van Moerkercke et al. 2015; 2016; Singh et al. 2021) regulate the IPAP-expressed biosynthetic genes in *C. roseus* (C. Li et al. 2025). Furthermore, an R2-R3 MYB TF *Idioblast MYB1* (*CrIDM1*) was recently reported to be idioblast specific and activates late-stage vinca alkaloid biosynthetic genes (C. Li et al. 2025). Other than the above-mentioned TFs, factors contributing to cell type specific biosynthesis in *C. roseus* remained undiscovered. It is also unclear whether conserved regulatory components exist across species to control the spatial expression of MIA biosynthetic genes.

Our understanding of plant metabolic pathways in plants has, until very recently, relied on whole tissue or whole organ-derived metabolome and transcriptome datasets. However, biosynthetic pathways of natural products are localized not only to distinct organs but also to distinct cell types within these organs. The emergence of single cell omics technologies has demonstrated enormous potential to revolutionize metabolic pathway gene discovery in plants by offering unprecedented resolution (C. Li et al. 2023; Zhan et al. 2023; Vu et al. 2024; Wu et al. 2024; McClune et al. 2024). Linking single-cell transcriptomics and metabolomics holds a great promise to (1) reveal how metabolic pathways are partitioned across cell types, (2) empower new discoveries in the alkaloid biosynthetic pathway in *C. acuminata*, and (3) identify transcriptional regulators that control cell type specificity of biosynthesis. By using a pair of medicinal plant species that produces structurally related MIAs, it is also possible to leverage regulatory relationships in one species to inform gene regulation in the other species, or vice versa, through phylogenetic analyses.

Results

Single cell multi-omics of *Camptotheca acuminata* revealed rare biosynthetic cells.

We performed organ-level metabolite profiling on key biosynthetic intermediates (Fig. 1A, Fig. S1A) across a panel of *C. acuminata* organs to identify organs relevant to camptothecin accumulation. We found that camptothecin was detected across all organs tested, with above-ground organs generally having higher abundance (Fig. S1A). We also sampled stem fluids and detected both loganic acid and strictosidinic acid in the xylem and phloem sap, suggesting these metabolites could be transported across organs over long distance, implying the complete biosynthetic pathway of camptothecin might be partitioned across organs (Fig. S1B). To investigate the cell type specificity of camptothecin biosynthesis, we chose young leaves for our first single cell omics experiments (Table S1) due to the relative ease of nuclei isolation, which also permits the simultaneous profiling of gene expression and chromatin accessibility using nuclei with the 10x Genomics Multiome Kit. To aid downstream bioinformatic analyses such as read mapping and quantification of gene expression, we produced a new version of the *C. acuminata* genome assembly and annotation (Table S2-S5), in which we used Nanopore full-length cDNAs to refine gene model predictions. This version (v4) of the genome assembly and annotation was highly colinear with the previous version of *C. acuminata* annotation (Fig. S2A) (Kang et al. 2021), except for a small chromosomal translocation. We isolated nuclei from young leaves (Fig. S2B-D) and used fluorescence activated nuclei sorting to remove organelles and debris. The sorted nuclei were then used to generate single nuclei RNA-seq and ATAC-seq libraries. For the gene expression assay of the multi-omics experiment, we obtained gene expression profiles for a total of 4,012 high quality nuclei and 26,074 expressed genes (Table S6, Fig. S3), which includes homologs of a few previously studied *C. roseus* MIA biosynthetic genes (Table S7-S8) (C. Li et al. 2023), as well as previously reported camptothecin decorating enzymes, *CPT11H* and *CPT10-OMT* (Nguyen et al. 2021; Salim, Jones, and DellaPenna 2018).

We first examined the gene expression data of the multi-omics dataset. Using unsupervised clustering and previously established leaf cell type marker genes (Table S9) (J.-Y. Kim et al. 2021), we identified the major cell types of leaf (i.e., mesophyll, epidermis, and vasculature) from a total of 18 cell clusters (Fig. 1B and Fig. S4A). Interestingly, among the expressed MIA genes (Table S7), *STR* genes were highly specific to a rare cell type that accounts for only 4.4% of all leaf cells, which we termed “STR+ cells” (Fig. 1C). We next performed a joint RNA-ATAC analysis by matching the cell barcodes from both assays (Fig. S4B) and thus transferring the cell type identity from the RNA-seq assay to the ATAC-seq assay. The ATAC-seq signals were strongly enriched at transcriptional start and end sites (Fig. S5A) and highly enriched at ATAC-seq peaks (Fig. S5B, Table S10). Among 48,456 ATAC-seq peaks, 199 (0.41%) were detected as STR+ marker peaks (ATAC-seq peaks that were specifically enriched in STR+ cells, see also Methods) (Fig. 1D). We found that genes adjacent (within 2-kb) to these STR+ marker peaks were highly enriched for expression in STR+ cells (Fig. 1E). Among the 121 genes that were within 2-kb to a STR+ marker peak, 34 (28%) of them were most highly expressed in STR+ cells, which is 9.3-fold over the expected background ratio ($p < 2.2 \times 10^{-16}$, Chi-squared test). The enrichment of STR+-expressed genes suggests that STR+ marker peaks could act as cell type specific enhanc-

ers. By performing *de novo* motif enrichment on STR+ marker peaks, we detected the MYB binding motif (Fornes et al. 2019) was overrepresented among these accessible chromatin regions (Fig 1F-G), suggesting MYB family TFs may be involved in cell type specific expression of *STR* genes in *C. acuminata*. Together, these results showcased that single cell omics techniques are powerful in detecting rare biosynthetic cell types or cell populations and the gene regulatory landscape within.

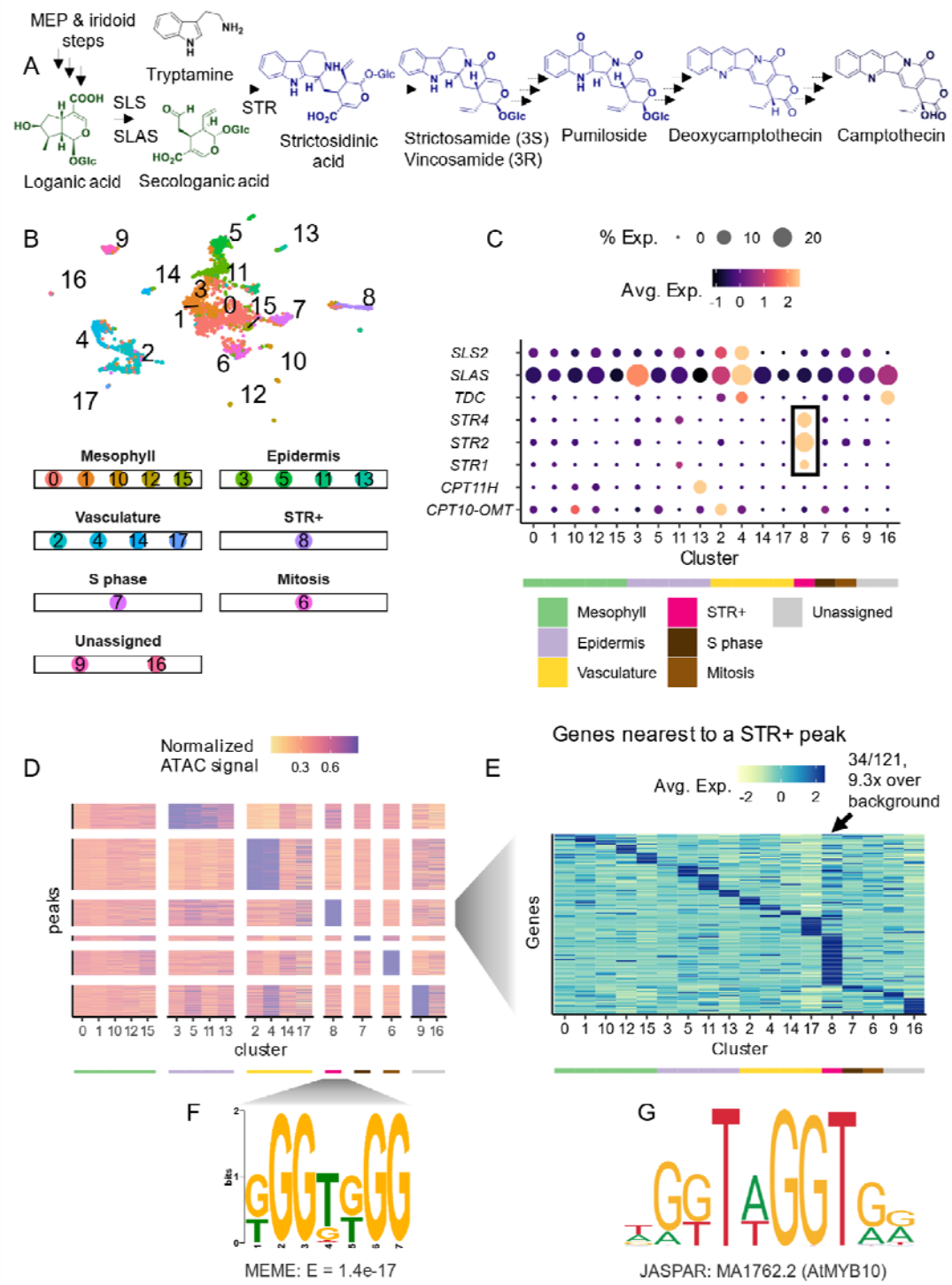


Fig. 1. Single cell multiome of *Camptotheca acuminata* reveals rare biosynthetic cells.

(A). The proposed biosynthetic pathway for camptothecin. Solid arrows indicate previously characterized enzymatic steps. Dashed arrows indicate proposed enzymatic steps. MEP: methylerythritol phosphate. SLS/SLAS: secologanin and secologanic acid synthases. STR: Strictosidine and strictosidinic acid synthases.

(B). Uniform Manifold Approximation and Projection (UMAP) of nuclei of the single nuclei RNA-seq dataset (n = 4,012), color coded by cell clusters.

(C). Gene expression heatmap of MIA biosynthetic genes across cell clusters detected in (A). Rows are expressed biosynthetic genes, which are ordered from upstream to downstream (see also Table S7-S8). 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 (see also Fig. S4A and Table S9). Box highlights expression of *STR* genes in STR+ cells.

(D). Heat map showing accessibility of cell type 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, with the same color palette as (B). Grey trapezoids highlight STR+ marker peaks, aspects of which were elaborated in (E) and (F).

(E). Heat map showing gene expression across cell clusters. Each row is a gene within 2-kb to a STR+ marker peak. Each column is a cell cluster. The predicted cell type for each cell cluster is annotated by the color strip below the x-axis, with the same color palette as (B). Arrow highlights genes most highly expressed in STR+ cells.

(F)-(G). DNA motifs enriched in STR+ marker peaks, as well as a reference MYB motif. E value $= 1.4 \times 10^{-17}$.

Stem is the active site of camptothecin biosynthesis.

Bulk metabolomic profiles of *C. acuminata* tissues did not reveal clear abundance patterns for metabolites previously proposed to be involved in camptothecin biosynthesis (Fig. S1A). Moreover, xylem and phloem sap extracted from *C. acuminata* stems, in which pathway intermediates, especially loganic acid and strictosidinic acid were detected, suggested that MIA biosynthesis likely involves the inter-organ transport of metabolic intermediates (Fig. S1B), further complicating the identification of biosynthetic genes. While STR+ cells were detected from single cell sequencing of young leaf tissues (Fig. 1), it remained unclear which cell type(s) camptothecin biosynthesis is localized, if it occurs in the leaf at all.

To clarify the localization of metabolites and potentially their corresponding biosynthetic genes, isotope feeding experiments were performed using labeled *d5*-tryptamine (Fig. 2A) on various *C. acuminata* tissues, including meristem, young and mature leaf disks, young and mature stem cuttings, and root segments (Fig. 2B). Labeled metabolites, including *d4*-strictosidinic acid, *d4*-

strictosamide, *d4*-vincosamide, *d4*-pumiloside, *d4*-deoxycamptothecin, and *d4*-camptothecin were monitored. All tested tissues produced substantial amounts of labeled MIAs derived from *d5*-tryptamine, particularly *d4*-strictosidinic acid, *d4*-strictosamide, and *d4*-vincosamide, confirming the presence of activities responsible for the early stages of camptothecin biosynthesis. Roots accumulated only minimal amounts of labeled intermediates, indicating limited biosynthetic activity. Notably, labeled pumiloside was detected exclusively in mature leaves, with no detectable levels in other tissues. Additionally, *d4*-deoxycamptothecin was undetectable across all tissues analyzed. Interestingly, only young and mature stems contained trace amounts of labeled camptothecin. The absence of *d4*-pumiloside in stems might suggest its consumption to produce labeled camptothecin. Based on these findings, we inferred that the stem is the primary site of camptothecin biosynthesis in *C. acuminata*.

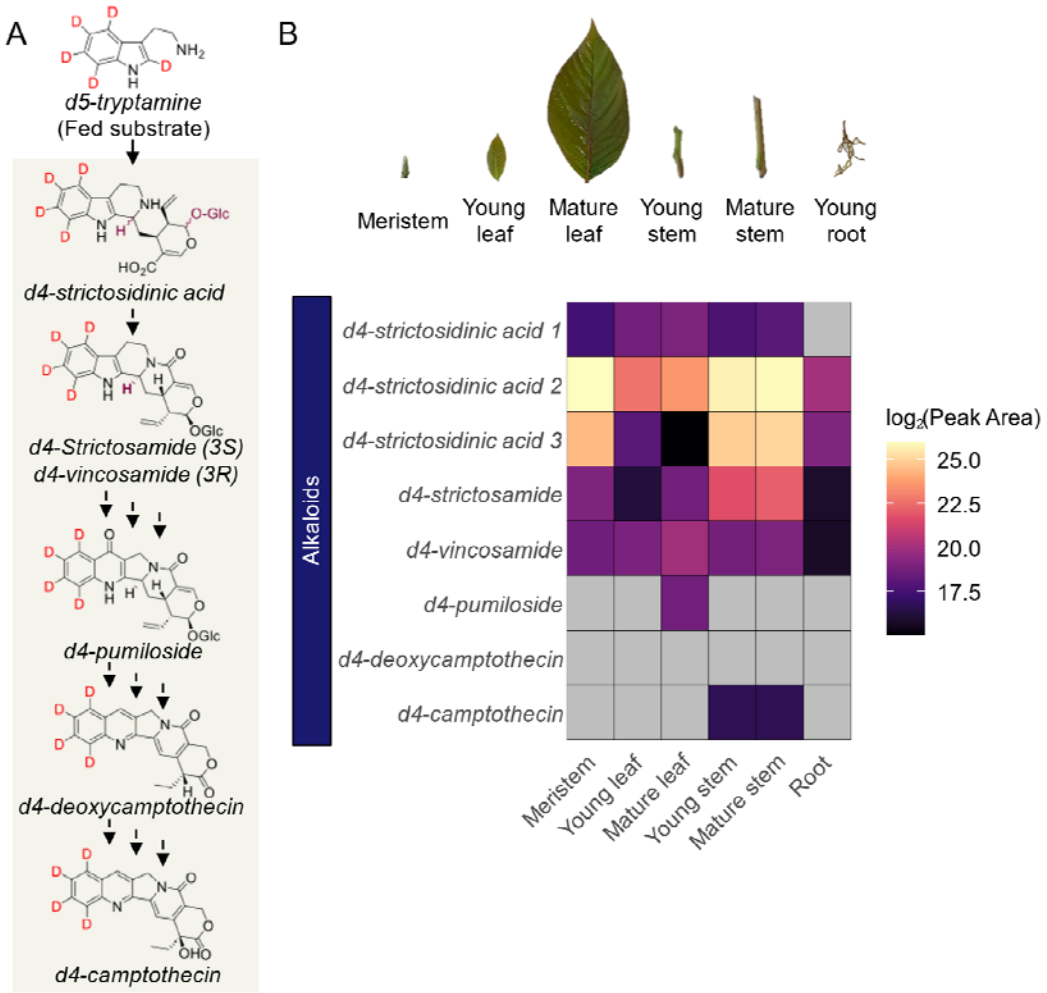


Fig. 2. Feeding experiments to reveal the active tissue(s) of camptothecin biosynthesis.
(A). Structures of labelled substrate and metabolites being monitored.
(B). Heat map showing abundances ($\log_2$ peak area) of labelled metabolites across organs. Grey color indicates lack of detection. Photos of organs are shown on top of the heatmap.

Complementary single cell omics detected differences in iridoid and alkaloid accumulation across organs and cells.

To better understand the cell type specificity of metabolites involved in camptothecin biosynthesis, we performed single cell metabolomic analysis using protoplasts (C. Li et al. 2023; Vu et al. 2024) (Fig. 3A). First, we optimized protocols to generate healthy, viable protoplasts from the meristem, leaf, and stem of *C. acuminata* (see also Methods). Individual protoplasts were then isolated using an automated single-cell picking system and transferred to 96-well plates. After addition of solvent and internal standard, the cells were analyzed by ultra-high-performance liquid chromatography coupled with high resolution mass spectrometry (UHPLC-HRMS). The analytical method was set up using authentic MIA reference compounds (Table S11). A total of 576 single cells per tissue were initially collected across six 96-well plates, each were subjected individually to targeted metabolomic analysis. To minimize batch-to-batch variation resulting from experimental or instrumental variability, each 96-well plate was treated as an independent experiment. After removing data from empty wells and doublets (two or more cells picked into the same well), the final dataset included 495 cells from meristem, 495 from leaf, and 493 from stem (Supplementary Datasets 3-5). External calibration with authentic standards enabled accurate quantification of specific metabolites within individual cells (Table S12), and analyte concentrations were computed based on cell volumes, calculated from cell size measured during collection.

Major iridoids reached remarkable intra-cellular concentrations in the millimolar range (Fig. 3B). Although significant variability was noted at the single cell level, the average loganic acid concentrations were highest in stem (128 mM), followed by meristem (74 mM) and leaf (17 mM). Notably, cells accumulating loganic acid represented a substantial proportion of the population in stems (57%) and meristems (32%), while they constituted only 0.6% in leaves. Secologanic acid, the immediate downstream of loganic acid, also reached millimolar concentrations across organs. Strictosidinic acid, a key intermediate in camptothecin biosynthesis, was detected across all tissues, exhibiting similar levels in stem and meristem but higher in leaves. Cell populations containing strictosidinic acid were most abundant in meristem (53%), followed by stem (38%), and were notably low in leaf tissue (3.4%). Interestingly, we detected low levels of strictosamide, pumiloside, and camptothecin exclusively in stem tissue with the population of 1% of the total stem protoplast population, further supporting our hypothesis that the stem is the primary site for camptothecin biosynthesis. Further analysis of metabolite co-occurrence within stem tissues revealed five distinct cell groups: Group A cells contained only loganic acid; Group B cells contained only secologanic acid; Group C cells contained loganic acid and secologanic acid (with or without strictosidinic acid); Group D cells contained only strictosidinic acid; and lastly Group E cells contained camptothecin and its proposed late stage intermediates (strictosamide, pumiloside, camptothecin) (Fig. S6). Notably, most cells accumulating loganic acid also contained secologanic acid. Additionally, very few cells exhibited co-occurrence of loganic acid/secologanic acid with proposed late stage camptothecin intermediates, highlighting that late

stage camptothecin biosynthesis is restricted to rare cell populations. Similar cell grouping patterns were observed in leaf and meristem tissues, where the co-occurrence of loganic acid and secologanic acid were prevalent (Fig. S6). Overall, this approach holds significant promise for investigating how metabolites are synthesized and transported within *C. acuminata* plants.

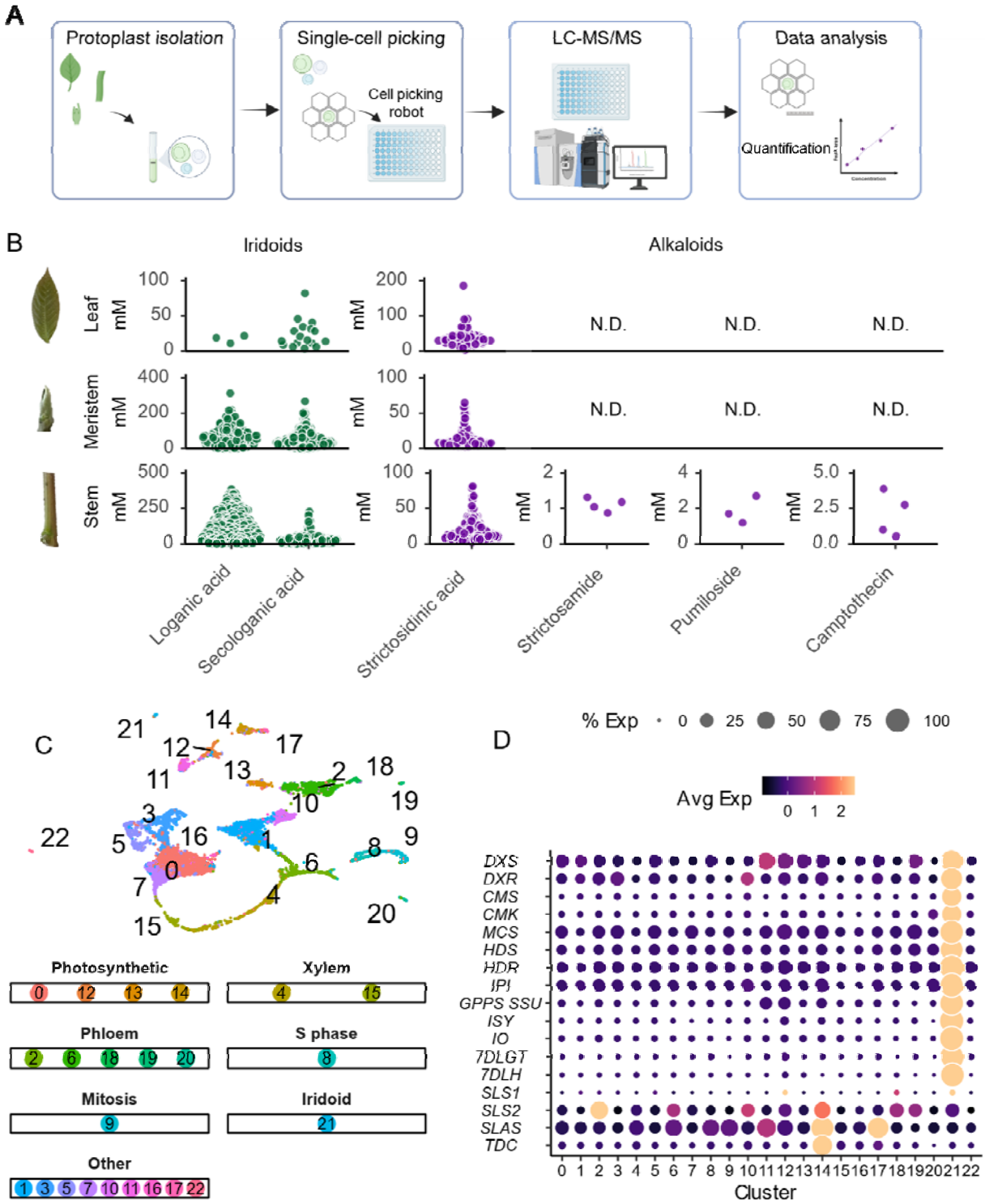


Fig. 3. Complementary single cell omics detected differences in iridoid and alkaloid accumulation across organs and cells.

(A). A schematic overview of single cell metabolomics experiments.

(B). Dot plots showing concentrations of key metabolites in single cells across organs. Each data point is a protoplast. For each metabolite, dots were offset from the center based on the distribution of concentrations among single cells. N.D. not detected.

(C). Uniform Manifold Approximation and Projection (UMAP) of stem protoplast containing high-quality RNA-seq data (n = 11,321), color coded by cell clusters. Photosynthetic cells refer to green cortex cells.

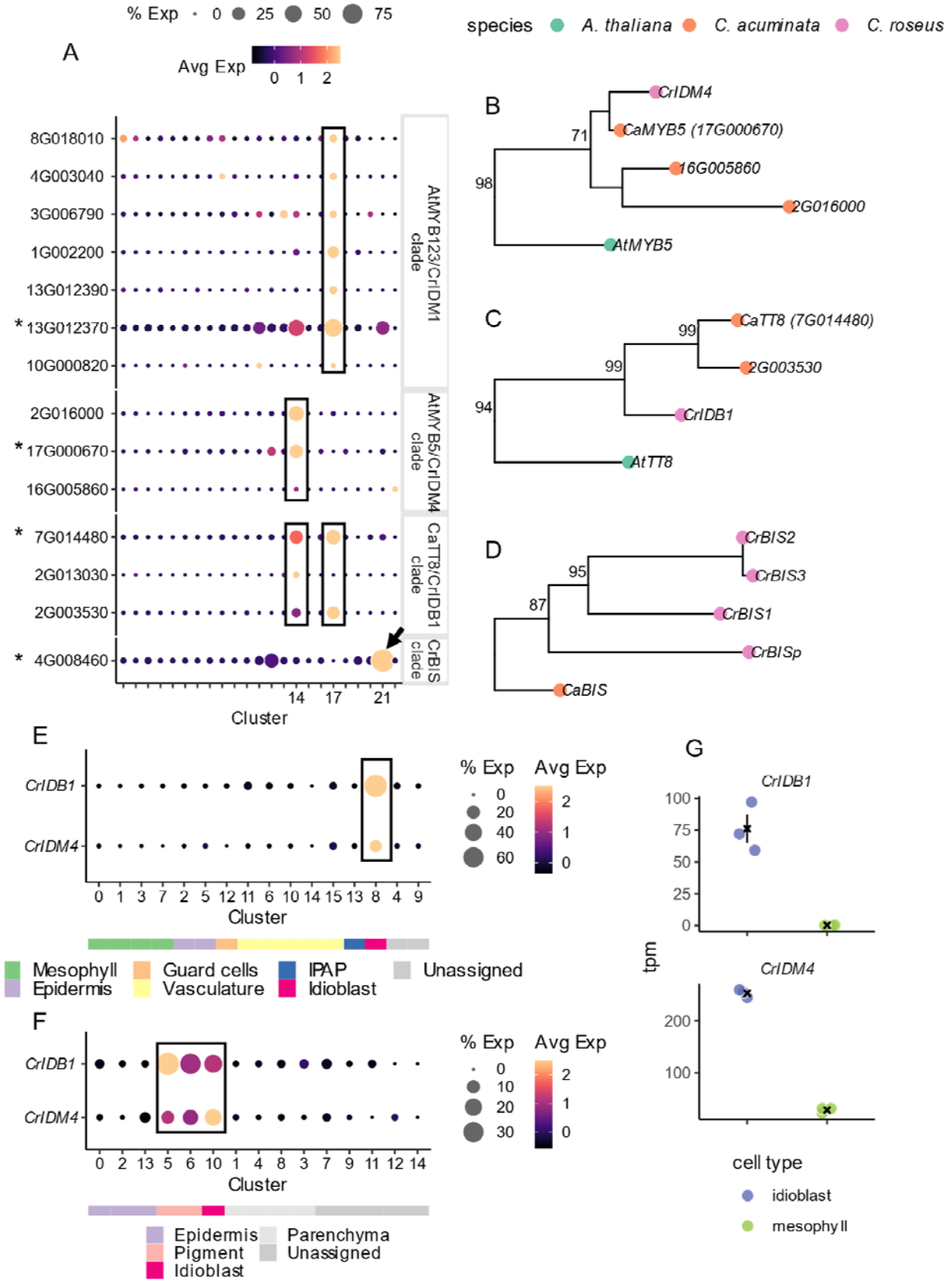
(D). Gene expression heatmap of MIA biosynthetic genes across cell clusters detected in (C). Rows are expressed biosynthetic genes, which are ordered from upstream to downstream (see also Table S7-S8). Color scale shows the average scaled expression of each gene at each cell cluster. Dot size indicates the percentage of cells where a given gene is detected.

Since we detected key late-stage metabolites (namely strictosamide, pumiloside, and camptothecin) in stem protoplasts but not the other organs, we next performed single cell RNA-seq using stem samples to investigate the cell type specificity of the biosynthetic pathway, as well as to identify potential regulators involved in organ or cell type specificity. The woody nature of stems rendered nuclei isolation difficult, and thus we opted to perform protoplast-based single cell RNA-seq using the PIP-seq platform (Clark et al. 2023). We obtained gene expression profiles of 11,321 cells and detected 29,002 expressed genes (Table S13, Fig. S7). With unsupervised clustering, we detected 23 cell clusters and identified cell types using previously established cell type marker genes for stem (Fig.3C, Fig. S8, Table S14) (Conde et al. 2022). We found that the MEP pathway and iridoid biosynthetic genes (Table S7-S8) are highly specific to a rare cell type (Fig. 3D) that accounts for only 0.68% of all stem protoplasts, which we termed "iridoid cells". In *C. roseus*, MEP and iridoid biosynthetic genes are specific to internal phloem associated parenchyma (IPAP) (Miettinen et al. 2014). The iridoid cells we detected in *C. acuminata* stems could be homologous to IPAP in *C. roseus*. Similar to *C. roseus*, the rest of the biosynthetic pathway appeared to be expressed in different cell types (Clusters 14 and 17, Fig 3D) (C. Li et al. 2023). Considering loganic acid was detected in a much larger proportion of cells (Fig. 3B), these results suggest that loganic acid may be synthesized in one cell type and subsequently transported to another cell type for downstream biosynthetic steps, which is consistent with the fact that downstream biosynthetic steps (e.g., *SLAS* and *TDC*) were expressed in a different cell cluster (Fig. 3D). These results showcased that single cell omics techniques are powerful in detecting exceptionally rare biosynthetic cells that account for a fraction of a percent of the total cell population.

Transcription factors co-expressed with MIA biosynthetic genes in *C. acuminata* have orthologs that are idioblast specific in *C. roseus*.

We sought to identify the transcriptional regulators that are involved in the cell type specific expression of biosynthetic genes in the stem, since stem is the primary site of camptothecin biosynthesis. We performed a gene co-expression analysis to search for TFs that are co-expressed with biosynthetic genes in Clusters 14, 17, and 21, where MIA biosynthetic genes were expressed (Fig. 3D, Fig. 4A). Four clades of TFs caught our attention (Fig. 4A), which belong to MYB or bHLH TF families (Supplemental Datasets 7-8). In Cluster 17, we detected MYB TFs that are in the same clade as *C. roseus IDM1*, which we recently reported to activate late-stage vinca alkaloid biosynthesis in *C. roseus* and found to be in the same clade as Arabidopsis MYB TFs *MYB123*, *GL1*, and *PAP1/2* (Supplemental Dataset 7) (C. Li et al. 2025). In Cluster 14, we detected another clade of MYB TFs, whose phylogeny contains Arabidopsis *MYB5* (S. F. Li et al. 2009) and a single *C. roseus* MYB ortholog (Fig. 4B). In both Clusters 14 and 17, we discovered a clade of bHLH TFs that belong to the *TT8* clade (Fig 4A and C), which contains *AtTT8* (Nesi et al. 2001) and a single *C. roseus* ortholog. Lastly, we detected a single bHLH TF in Cluster 21 belonging to the *bHLH Iridoid Synthesis* (*BIS*) clade (Fig. 4A and D), which were well-characterized TFs regulating iridoid biosynthesis in *C. roseus* (Van Moerkercke et al. 2015). This bHLH TF was named *CaBIS*. For each clade of TFs discussed above, we designated one member as the representative *C. acuminata* homolog for further characterization using overexpression (see below). In stark contrast to their expression patterns in the stem, none of these *C. acuminata* homologs were cell type specific in the leaf (for example, Fig. S9A-C). In the case of *CaBIS*, its transcript was undetected in the leaf, despite cell type specific expression in the stem (Fig. 4A).

We then investigated the expression patterns of *C. roseus* orthologs of *CaMYB5* and *CaTT8*, which have not been previously characterized. Querying *C. roseus* leaf and petal single cell RNA-seq datasets (C. Li et al. 2025; Vu et al. 2024), we found that these orthologs were highly specific to the idioblast in the leaf, with enriched expression in pigmented cells and idioblast in petals (Fig. 4E-F). In light of their idioblast specific/enriched expression in two organs, these *C. roseus* orthologs were named *Idioblast bHLH 1* (*IDB1*) and *Idioblast MYB4* (*IDM4*), respectively, as we previously used this naming scheme to describe three idioblast specific MYBs (*CrIDM1/2/3*) and one idioblast specific WRKY TF (*CrIDW1*) (C. Li et al. 2025). In a cell sorting experiment (Guedes et al. 2024), both *CrIDB1* and *CrIDM4* were highly expressed in the idioblast fraction of sorted protoplasts, but undetected in the mesophyll fraction (Fig. 4G). Like *CrBIS1* and *CrIDM1*, *CrIDB1* was inducible by a methyl jasmonate treatment (Van Moerkercke et al. 2013) ($\log_2$FC = 1.3), whereas *CrIDM4* displayed a lesser degree of induction ($\log_2$FC = 0.75, Fig. S9D). Lastly, both *CrIDB1* and *CrIDM4* loci were highly accessible in the idioblast (Fig. S9E-F), consistent with their expression in the idioblast (C. Li et al. 2025). These analyses suggest that the same clade of TFs has been co-opted to regulate biosynthetic genes in a cell type specific manner across multiple species.


% Exp 0 25 50 75
Avg Exp 0 1 2
A
8G018010
4G003040
3G006790
1G002200
13G012390
* 13G012370
10G000820
2G016000
* 17G000670
16G005860
* 7G014480
2G013030
2G003530
* 4G008460
AtMYB123/CrIDM1 clade
AtMYB5/CrIDM4 clade
CaTT8/CrIDB1 clade
CrBIS clade
14 17 21
Cluster
species A. thaliana C. acuminata C. roseus
B
CrIDM4
CaMYB5 (17G000670)
16G005860
2G016000
AtMYB5
71
98
C
CaTT8 (7G014480)
2G003530
CrIDB1
AtTT8
99
99
94
D
CrBIS2
CrBIS3
CrBIS1
CrBISp
CaBIS
95
87
E
CrIDB1
CrIDM4
0 1 3 7 2 5 12 11 6 10 14 15 13 8 4 9
Cluster
% Exp 0 20 40 60
Avg Exp 2 1 0
Mesophyll
Epidermis
Guard cells
Vasculature
IPAP
Idioblast
Unassigned
F
CrIDB1
CrIDM4
0 2 13 5 6 10 1 4 8 3 7 9 11 12 14
Cluster
% Exp 0 10 20 30
Avg Exp 2 1 0
Epidermis
Pigment
Idioblast
Parenchyma
Unassigned
G
CrIDB1
CrIDM4
tpm
100 75 50 25 0
200 100
cell type
idioblast
mesophyll

357

Fig. 4. TFs co-expressed with alkaloid biosynthetic genes and their *Catharanthus roseus* orthologs.

(A). Gene expression heatmap of *C. acuminata* TFs co-expressed in Clusters 14, 17, or 21. Rows are TFs, which are grouped by clade. 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. Boxes highlight co-expression. Arrow highlights cell type specific expression of *CaBIS*. Asterisks represent representative homologs that were used in transient overexpression experiments.

(B)-(D). Phylogeny of the *AtMYB5/CrIDM4* clade (B), the *AtTT8/CrIDB1* clade (C), and the *BIS* clade (D). Branch lengths represent amino acid substitutions. Numbers on nodes represent bootstrap support (only nodes with > 70% bootstrap support were labelled). Colors of tips represent species.

(E)-(F). Gene expression heatmap of *CrIDB1* and *CrIDM4* across cell clusters detected in *C. roseus* leaf (E) and petal (F). 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. Boxes highlighted cell cluster(s) where these TFs were expressed. Data from Li et al. (C. Li et al. 2025).

(G). Mean separation plots showing expression levels of *CrIDB1* and *CrIDM4* (in units of transcripts per million) in idioblast or mesophyll fractions of sorted cells. Each data point is a biological replicate. Black × indicates arithmetic means. Error bars indicate standard error of mean. Data from Guedes et al. (Guedes et al. 2024).

Overexpression of alkaloid co-expressed TFs activates biosynthetic genes in *C. roseus*.

Co-expression of *CaMYB123*, *CaMYB5*, *CaTT8*, and *CaBIS* with MIA biosynthetic genes in a cell type specific manner suggests that these TFs might regulate camptothecin biosynthetic gene expression. Their orthologs (*CrIDB1* and *CrIDM4*) in *C. roseus* are co-expressed in the idioblast, a rare cell type that accumulates several MIAs and the site of vindoline biosynthesis, suggestive of orthologous function regulating MIA biosynthesis in rare cell types in these two species. We hypothesized that these TFs may be functional homologs in regulating cell type specific MIA biosynthesis in diverse organs of *C. roseus* and *C. acuminata*. While transient overexpression for a tree species such as *C. acuminata* proved difficult, *C. roseus* petals have been established as a robust platform for assessing gene function for MIAs (Colinas and Goossens 2022). Thus, we tested not only *C. acuminata* TFs but also *C. roseus* orthologs (Table 1) using agrobacterium mediated transient overexpression in *C. roseus* petals. We devised 13 overexpression treatments (Treatments A-M) (Table 1, Supplementary Dataset 9, Fig. 5A), that included *CrIDB1*, *CrIDM4*, *CrBIS1*, and their *C. acuminata* orthologs. For example, Treatments A and B included overexpression of the ortholog pair *CaTT8* and *CrIDB1*, respectively (Fig. 5A, Fig. 4C). Since *AtTT8* and closely related bHLH TFs (e.g., *AtGL3* and *AtEGL3*) are known to interact with MYB TFs in

the *AtMYB123*, *AtGL1*, and *AtPAP1/2* clade (Millard et al. 2019), we also tested co-overexpression treatments of *CaTT8*/*CrIDB1* with *CrIDM1* and its putative *C. acuminata* ortholog *CaMYB123*. This was based on the fact that *CrIDM1* belongs to a clade of MYB TFs known to interact with bHLH TFs. Each overexpression plasmid also contained an *AmCyan* (a bright green/cyan fluorescent protein) expression cassette as an indicator of successful transient expression. A plasmid containing only the *AmCyan* expression cassette was used as the negative control (Treatment M, Fig. 5A).

Treatment	Transgene	Notes
A	*CaTT8*	Ortholog of *CrIDB1*
B	*CrIDB1*	Ortholog of *CaTT8*
C	*CaBIS*	Ortholog of *CrBIS1*
D	*CrBIS1*	Ortholog of *CaBIS*
E	*CrIDM1*	Ortholog of *CaMYB123*
F	*CaMYB123*	Ortholog of *CrIDM1*
G	*CrIDM4*	Ortholog of *CaMYB5*
H	*CaMYB5*	Ortholog of *CrIDM4*
I	*CrIDB1* + *CrIDM1*	Combination treatment of B & E
J	*CrIDB1* + *CaMYB123*	Combination treatment of B & F
K	*CaTT8* + *CrIDM1*	Combination treatment of A & E
L	*CaTT8* + *CaMYB123*	Combination treatment of A & F
M	*AmCyan* only	Negative control

Table 1. Layout of overexpression treatments. Treatments A-L also contained an *AmCyan* expression cassette. Light pink: ortholog pair of *CaTT8* and *CrIDB1*. Yellow: ortholog pair of *CaBIS* and *CrBIS1*. Green: ortholog pair of *CrIDM1* and *CaMYB123*. Blue: ortholog pair of *CrIDM4* and *CaMYB5*. Hot pink: combination treatments. Grey: negative control.

We found that six out of the 13 treatments led to upregulation of biosynthetic genes in the MIA pathway (Fig. 5B). We found examples of both divergent and conserved activities between ortholog pairs. For example, of all the *C. acuminata* TFs tested, *CaMYB5* and its *C. roseus* ortholog *CrIDM4*, activated late vinca alkaloid biosynthetic genes, much like *CrIDM1* (Treatments G and H, respectively, Fig. 5B-C), suggesting this pair of orthologs can substitute for each other in heterologous systems. However, *CaBIS* and *CaMYB123* did not elicit the same response as their *C. roseus* orthologs. Specifically, *CrBIS1* activated early biosynthetic genes (i.e., MEP and iridoid pathways) (Treatment D, Fig. 5B), but not later steps of the pathway. This finding is consistent with previous reports regarding *BIS* family TFs in *C. roseus* (Van Moerkercke et al. 2015). In contrast, the *C. acuminata* ortholog *CaBIS* did not activate these genes (Treatment C, Fig 5B). *CrIDM1* activated late vinca alkaloid biosynthetic genes, namely *DAT*, *D4H*, and *THAS2* (Treatment E, Fig 5B-C), which is consistent with our recent findings (C. Li et al. 2025). However, its putative ortholog *CaMYB123*, did not elicit the same response (Treatment F, Fig

5B). Interestingly, *CaTT8* and*CrIDB1* in combination with *CrIDM1*, but not *CaMYB123*, activated late-stage vinca alkaloid biosynthetic genes (Treatments I-L), with stronger effects than *CrIDM1* alone (Fig. 5B-C). Surprisingly, *CaTT8/ CrIDB1+ CrIDM1* also activated upstream biosynthetic genes *T16H2* and *16OMT* (Fig. 5B-C), which are normally not co-expressed with late-stage enzymes in the idioblast (C. Li et al. 2025). Lastly, without *CrIDM1*, *CaTT8* and its ortholog *CrIDB1* alone did not result in the activation of biosynthetic genes (Treatments A-B, Fig 5B), suggesting their regulatory activities require a specific MYB partner, such as *CrIDM1*.

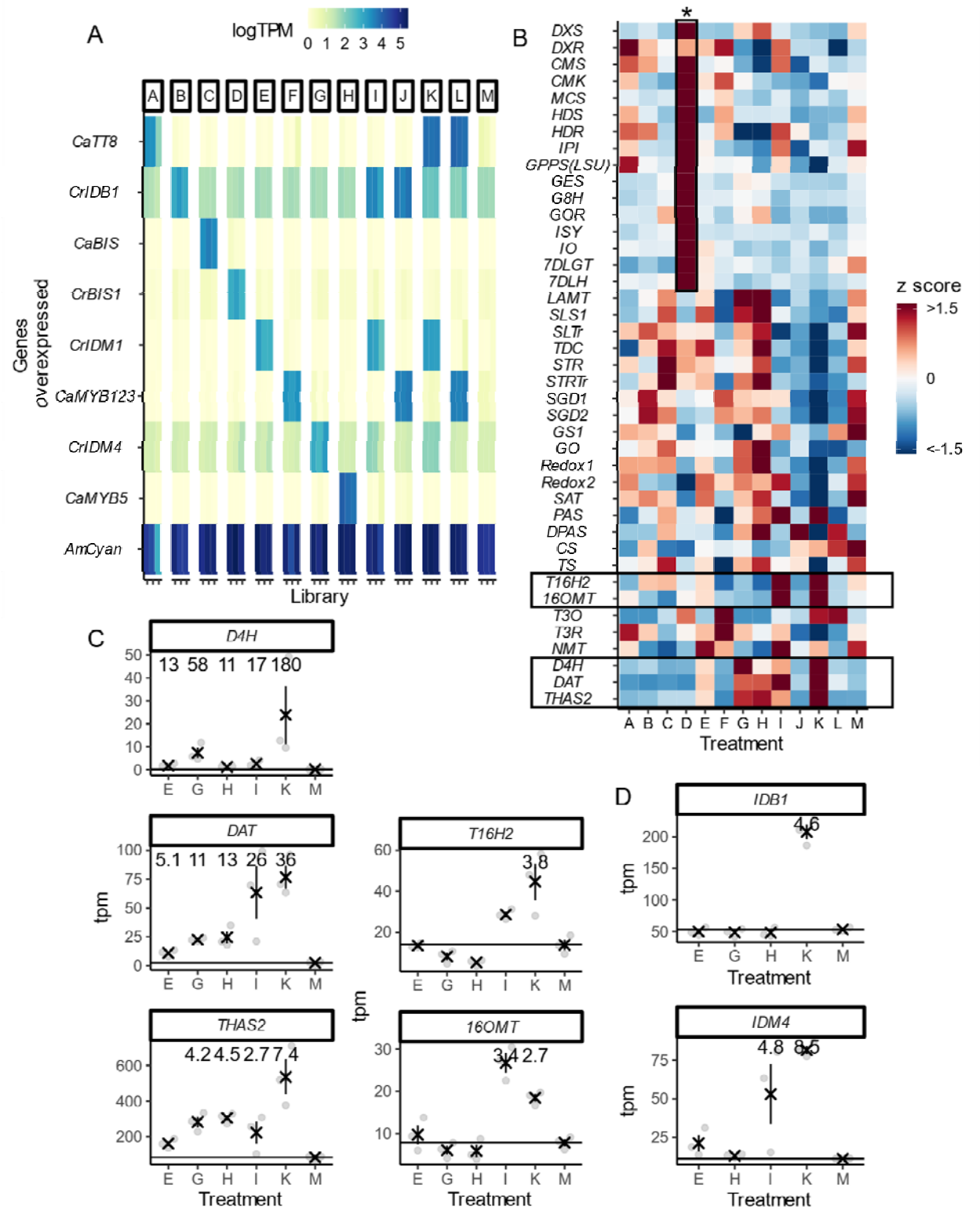


Fig. 5. Overexpression of alkaloid co-expressed TFs activates biosynthetic genes in *C. roseus*.

(A). Heat map showing the layout of the overexpression experiment. Each row is a transgene being overexpressed. Each column is a biological replicate, grouped by treatments (A-M) (See also Table 1). Color scale represents expression level, in units of $\log_{10}$ transcripts per million.

(B). Heatmap of the MIA 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). Boxes represent biosynthetic genes activated by overexpression treatments. Asterisk represents MEP and iridoid biosynthetic genes being activated by *CrBIS1*.

(C)-(D). Mean separation plots showing expression levels of biosynthetic genes (C) or TFs (D) (in units of transcripts per million). Each data point is a biological replicate. Black × indicates arithmetic means. Error bars indicate standard error of mean. Numbers on top of the graph represent fold-change values (only displayed if adjusted p-value < 0.05). Horizontal lines represent baseline expression in the negative control.

In addition to biosynthetic genes, overexpression of TFs also activated other TFs. For example, when *CaTT8* and *CrIDM1* were co-overexpressed, *CrIDB1* was activated (Fig. 5D). *CrIDB1* was likely also upregulated by the combination of *CrIDB1* and *CrIDM1*, but any upregulation of the endogenous locus was masked by the overexpressing transgene. Given that in Arabidopsis, *TT8* can reinforce its own expression in a MYB-dependent manner (Baudry, Caboche, and Lepiniec 2006), it is likely that *CrIDB1* can self-activate via *CrIDM1*. We also found that the combination of *CaTT8/CrIDB1* with *CrIDM1*, but not *CrIDM1* alone, can activate *CrIDM4* (Fig. 5D), suggesting the stronger effects of the combined overexpression could in part be due to the downstream TF *CrIDM4*. We previously described three other idioblast specific TFs, *CrIDM2/3* and *CrIDW1* (C. Li et al. 2025), among which *CrIDM2* was activated by *CaMYB5*/*CrIDM4*, whereas *CrIDM3* and *CrIDW1* were activated by the combined overexpression of *CaTT8*/*CrIDB1* and *CrIDM1* (Fig. S10A). Taken together, these results suggest *CrIDM1* occupies a higher position in the hierarchy of idioblast specific TFs.

An extended idioblast regulon based on overexpression, co-expression at the cell type level, and orthology.

When we compared the target genes of *CaTT8* or *CrIDB1* in combination of *CrIDM1*, where a target gene was defined as upregulated by overexpression and co-expressed in the idioblast, we found that the majority of target genes were shared between these combinations ($p < 2.2 \times 10^{-16}$, Fisher Exact test) (Fig. 6A). The same was true for *CaMYB5* and *CrIDM4* ($p < 2.2 \times 10^{-16}$, Fisher Exact test), which also shared a significant number of target genes. (Fig. 6B). In addition, we found that genes activated by either the *CrIDM1* + *CrIDB1* combination or by *CrIDM4* alone are highly enriched for idioblast expression (Fig. 6C-D). Specifically, among all differentially upregulated genes in the *CrIDM1* + *CrIDB1* combination treatment, 147 out of 796 (18%) genes were most highly expressed in the idioblast, representing a 4.5-fold enrichment over the expected

background ratio ($p < 2.2 \times 10^{-16}$, Chi-squared test) (Fig. 6C). The enrichment was even stronger for *CrIDM4*, since 43 out of 144 (30%) differentially upregulated genes were most highly expressed in the idioblast, representing a 7.5-fold enrichment over the expected background ratio ($p < 2.2 \times 10^{-16}$, Chi-squared test) (Fig. 6D). Among the target genes were putative biosynthetic genes encoding alcohol dehydrogenase, FAD/NAD(P)-binding oxidoreductase, and cytochrome P450 family enzymes (Fig. S10B-C), which also included previously characterized idioblast-specific enzymes *CrTHAS1* (01G033230), *CrADH32* (06G024590) (C. Li et al. 2023), and *Serpentine Synthase* (03G003540) (Yamamoto et al. 2021).

Using the overexpression and single cell omics datasets, combined with phylogenetic analyses and previously reported TF-promoter transactivation assays (C. Li et al. 2025), we identified TF candidates from *C. accuminata* and produced an extended gene regulatory network for late-stage vinca alkaloid biosynthetic genes in *C. roseus* (Fig. 6E). The extension of such a network is facilitated by the orthologous relationships between *C. acuminata* and *C. roseus*. The network contained 14 edges, which were decorated by the type(s) of supporting evidence: 1) activated by overexpression, 2) co-expressed across single cells, 3) orthologous between *C. acuminata and C. roseus*, and 4) promoter activated in a transactivation assay. Notably, we discovered a feed-forward activation network motif: while *CrIDM1* and *CrIDB1* in combination activate late-stage biosynthetic genes, they also activate *CrIDM4* that in turn activates the same late-stage biosynthetic genes targeted by *CrIDM1* and *CrIDB1* (Fig. 6E). These results showcased that by investigating TFs co-expressed with biosynthetic genes in a different species (e.g., *CaTT8* and *CaMYB5*), we gain power for discovering cell type specific metabolic regulators, in this case *CrIDB1* and *CrIDM4*.

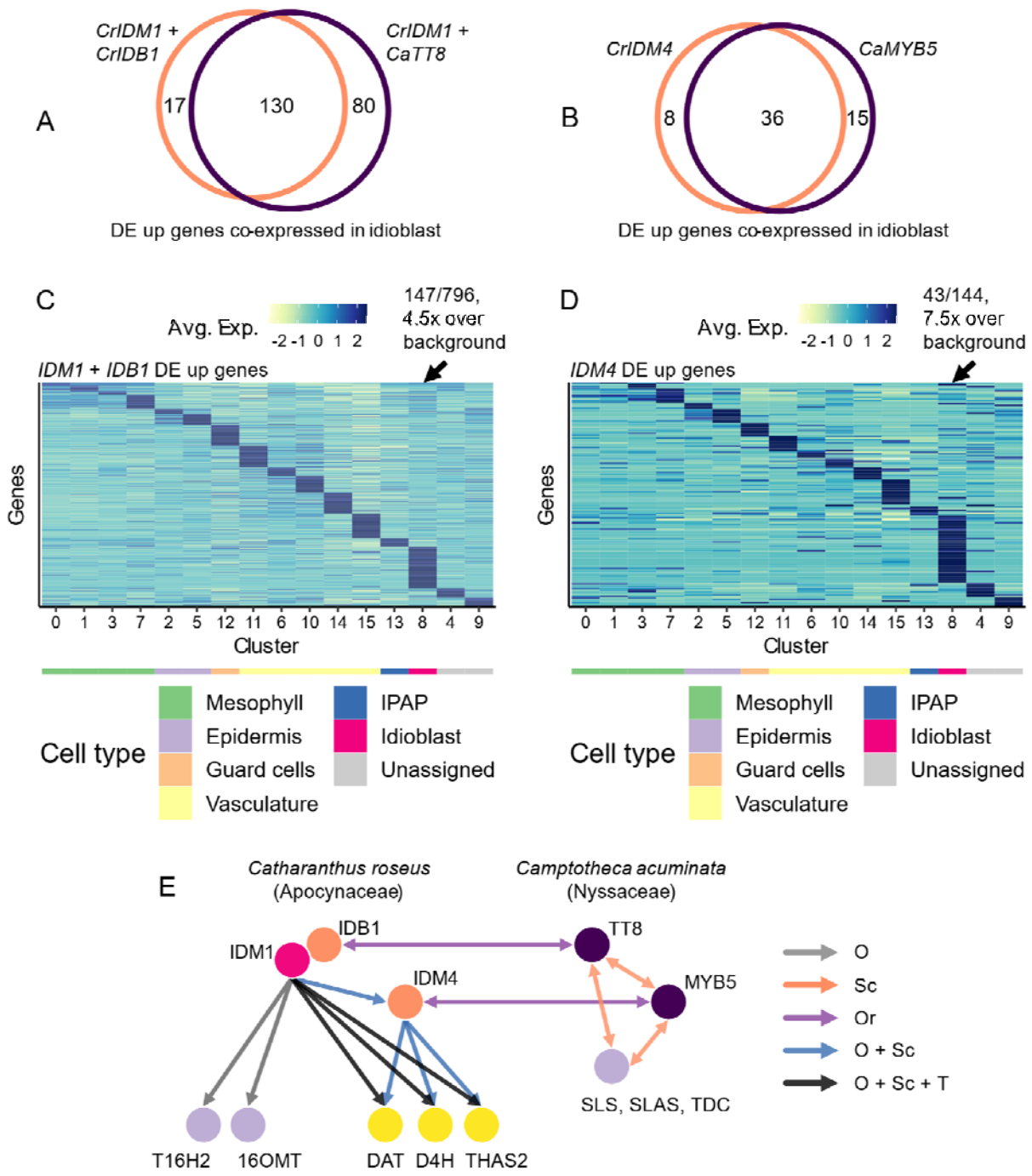


Fig. 6. An extended idioblast regulon based on overexpression, co-expression at the cell type level, and orthology.

(A)-(B). Venn diagrams showing the overlap of target genes for *CrIDM1* + *CrIDB1* vs *CrIDM1* + *CaTT8* (A) and *CrIDM4* vs *CaMYB5* (B). DE up genes: differentially expressed genes that are upregulated by overexpression.

(C)-(D). Heat map showing gene expression across cell clusters in *C. roseus* leaf. Each row is a differentially upregulated gene by *CrIDM1* + *CrIDB1* (C) or *CrIDM4* (D). Each column is a cell cluster. The cell type for each cell cluster is annotated by the color strip below the x-axis. Arrows highlight genes most highly expressed in idioblast. Data from Li et al. (C. Li et al. 2025).

(E). A gene regulatory network integrating multiple modules of omics and experimental data. Each node is a gene. Each edge represents a regulatory or phylogenetic relationship, color coded by the type of evidence supporting it. O: upregulated when the TF is overexpressed; Sc: co-expressed across single cells; Or: orthologous TFs; T: promoter activated in a transactivation assay.

Discussion

Both *C. acuminata* and *C. roseus*, separated by ~115 million years of evolution (Kumar et al. 2017), produce MIAs with camptothecin and vinblastine being the respective key natural products of interest to human health. The respective biosynthetic pathways are largely shared until the strictosidine/strictosidinic acid scaffold. Cell type specificity of plant specialized metabolism has been well-documented (Weng et al. 2021), including MIA biosynthesis in *C. roseus*. In this study, we showed that camptothecin biosynthesis is also cell type specific. Using single cell/single nuclei sequencing, we identified rare biosynthetic cell types, such as STR+ cells in the leaf (Fig. 1) and iridoid cells in the stem (Fig. 3C-D). While late-stage camptothecin biosynthetic genes are yet to be discovered, the observation that strictosamide, pumiloside, and camptothecin were only detected in a small number of cells in the stem (Fig. 3B) suggests that the late-stage biosynthesis of camptothecin is likely highly cell type specific as well. Consistent with recent single cell sequencing and metabolomics studies in *C. roseus* (C. Li et al. 2023; 2025; Vu et al. 2024), our findings support a potentially generalizable principle that a small percentage of cells could account for the vast majority of biosynthesis, both at the level of transcription and at the level of metabolite accumulation. Such a principle may extend beyond MIAs to other classes of plant natural products other than MIAs. The reasons for cell type specificity of specialized metabolism are unknown, but it is plausible that restricting specialized metabolism to rare, specialized cell types helps mitigate autotoxicity/interference with other metabolic processes, such as photosynthesis.

We showed that the accumulation of metabolites does not always correspond to active biosynthesis, which highlights the potential roles of long-distance metabolite transport between organs. While camptothecin was detected across all organs analyzed (Fig. S1B), we found that stem is the primary site of camptothecin biosynthesis (Fig. 2). Using single cell transcriptomics from two organs, we showcased that the expression patterns of biosynthetic genes are distinct across organs. Stem-expressed iridoid biosynthetic genes and leaf-expressed *STR* genes (Fig. 3D and Fig. 1C) pose a possibility that iridoids (e.g., loganic acid) are produced in iridoid cells in the stem and transported to the leaf, where it was converted to strictosidinic acid in STR+ cells in the

leaf. Strictosidinic acid can then be transported back to the stem, where the rest of the camptothecin biosynthesis occurs. This inter-organ partitioning of the pathway is also supported by the detection of loganic acid and strictosidinic acid in stem fluids (Fig. S1C). The observed spatial separation of biosynthetic activities likely reflects differences in transcriptional regulation across organs, driven by organ-specific expression of TFs. For example, TFs that were highly co-expressed with biosynthetic genes in the stem (Fig. 4A) showed either low expression or lacked cell type specificity in the leaf (Fig. S9A-C).

We discovered four clades of TFs co-expressed with MIA biosynthetic genes in the *C. acuminata* stem (Fig. 4), whose phylogenies include previously reported TFs that regulate MIA biosynthesis in *C. roseus*, such as *CrIDM1* and *CrBIS1*, suggesting cell type specific regulation may be a conserved feature. Among these newly described TFs, *CaTT8* and *CaMYB5* were able to activate similar sets of target genes as their *C. roseus* orthologs (Fig. 6A-B) when heterologously expressed, either in combination with *CrIDM1* (in the case of *CaTT8*), or on its own (in the case of *CaMYB5*). Surprisingly, *CrIDB1* (*CaTT8* ortholog) and *CrIDM4* (*CaMYB5* ortholog) were specific to the idioblast in *C. roseus*, suggesting they play a regulatory role in controlling biosynthetic genes which are also idioblast-specific. We demonstrated that C*rIDM1*, in combination with *CrIDB1*, occupies a higher position within the hierarchy of idioblast specific TFs, where the combination of *CrIDM1* and *CrIDB1* not only activates biosynthetic genes, but also activates other TFs, including *CrIDM4* that in turn activates biosynthetic genes (Fig. 5B-D, Fig. S10A), forming a feed-forward activation loop. We infer that *CrIDB1* reinforces its own expression via *CrIDM1* (Fig. 5D), reminiscence of how *AtTT8*, the Arabidopsis ortholog, activates its own transcription in a MYB-dependent manner as a feedback activation loop, which contributes to strong and cell-specific accumulation of flavonoids in Arabidopsis (Baudry, Caboche, and Lepiniec 2006). *CrIDM1*, *CrIDB1*, and *CrIDM4* activate an idioblast metabolic regulon, which includes all currently known idioblast specific enzymes (Fig. 6, Fig. S10). The lack of a convenient transient expression platform for *C. acuminata* makes similar *in planta* experiments for *C. acuminata* very difficult, but it is likely that *CaTT8* and *CaMYB5* play a role in regulating calprotectin biosynthesis as well.

Our experiments established *CrIDM1*, *CrIDB1*, and *CrIDM4* as idioblast metabolic activators that activate late-stage vinca alkaloid biosynthetic genes. Additional mechanisms, such as chromatin accessibility (in the case of *T16H2* and *16OMT*), might be involved in restricting biosynthetic genes expression to specific cell types. In *C. roseus* leaves, *T16H2* and *16OMT* are specific to the epidermis, not idioblast (C. Li et al. 2025). The lack of *T16H2* and *16OMT* expression in the idioblast can be explained by the fact that the promoter regions of these biosynthetic genes are only accessible in the epidermis (C. Li et al. 2025), such that CrIDM1 and CrIDB1 cannot activate *T16H2* and *16OMT* transcription in the idioblast.

Lastly, the same clade of TFs (e.g., the *TT8* clade and the *MYB5* clade) might have been co-opted to regulate biosynthetic genes in a cell type specific manner across multiple species. Investigating cell type specific TFs that are co-expressed with biosynthetic genes across multiple species is a powerful strategy for discovering metabolic activators. A common strategy for identifying TFs relevant to a biosynthetic pathway is co-expression analysis, either across organs and/or stress conditions, and more recently, across cell types (C. Li et al. 2025). However, such analyses often yield an unwieldly number of candidate TFs, making downstream functional characterization challenging. Taking advantage of phylogenetic analyses, we demonstrated that if a clade of TF includes orthologs that are cell type specific and co-expressed with biosynthetic genes across multiple species, other members of the same clade are also likely to be metabolic regulators, as shown in the case of *CrIDB1* and *CrIDM4* (Fig. 6E).

Methods

Plant growth conditions.

One-year-old *C. acuminata* plants were bought from Planten Tuin Esveld, Netherlands. Plants were cultivated in a standard soil mix in a greenhouse at 20-25°C (daytime) and 17-20°C (nighttime) with 40-70% humidity under the minimum artificial light of 12 hrs at Jena Germany. At Athens, U.S.A., *C. acuminata* trees were grown in a greenhouse at 27-29°C (daytime) and 17-20°C (nighttime) with 15 hrs light and natural humidity. *Catharanthus roseus* (cultivar "Little Bright Eyes") plants were grown under a 15-hr photoperiod on a light shelf in the lab.

C. acuminata metabolite profiling.

Plant tissues (meristem, axillary bud, young leaf, mature leaf, young stem, mature stem, root) were collected and snap frozen in liquid nitrogen. Mortar and pestle were used to grind the tissues. Metabolites were extracted from the tissues using extraction solvent (85% methanol, 15% water, 0.1% formic acid) with the ratio of 100 mg tissues:1 ml extraction solvent. The extracts were filtered through 0.22 µm PTFE syringe filters. For xylem and phloem sap extraction, a sterile needleless syringe adapted with parafilm and modified micropipette tips formed an airtight cartridge enclosing a 2-cm stem segment, allowing negative pressure generation (Longchar et al. 2020). Finally, extracts were diluted 100-fold with extraction buffer containing 50 nM catharanthine as an internal standard before LC-MS analysis (Fig. 1, Fig. S1).

DNA isolation, sequencing, and genome assembly.

High molecular weight DNA was isolated from immature leaf tissue using the Takara Nucleobond HMW DNA kit (Takara, Kusatsu, Shiga, Japan). Immature leaves were harvested from a tree two meters in height, grown in a greenhouse under the following conditions: 28ºC day, 19ºC night, with 15 hours of light at an intensity of 600 µM. Using the PacBio SMRTbell Template Prep Kit 3.0 (Menlo Park, CA), a single library was prepared and sequenced by the University of Minnesota Genomics Center. Sequencing was performed on a PacBio Revio instrument producing 47.4 Gb of HiFi data with a N50 of 18.5-kb, and an estimated coverage of

117x. PacBio HiFi reads were assembled using hifiasm v0.19.9-r616 (Cheng et al. 2024) with the homozygous coverage parameter set to 119. Contigs less than 50-kb were discarded using SeqKit (v2.8.2) (Shen et al. 2016) followed by two rounds of scaffolding using the scaffold tool in RagTag (v2.1.0) (Alonge et al. 2019) with the previously published *C. acuminata* v3.0 assembly (Kang et al. 2021). Assembly metrics were calculated at each step using assembly-stats (v1.0.1) (github.com/sanger-pathogens/assembly-stats). PacBio HiFi reads and the genome assembly were checked for contamination using Kraken 2 (v2.1.3) (Wood, Lu, and Langmead 2019) with the PlusPFP database (k2_pluspfp_20240605, https://benlangmead.github.io/aws-indexes/k2). BUSCO (v5.7.1) (Manni et al. 2021) was run at each step in the assembly process with the embryophyta_odb10 database to verify completeness of the genome assembly. Additional QC was performed by aligning the genome assembly to the *C. acuminata* v3.0 assembly (Kang et al. 2021) using minimap2 (v2.17-r941) (H. Li 2018) and the alignments visualized using D-Genies (Cabanettes and Klopp 2018).

cDNA library construction, long reads sequencing, and processing.

Mature leaf, shoot apex, stem, and root were used for RNA isolation using a modified hot borate method (Wan and Wilkins 1994). mRNA was isolated using the Dynabead mRNA Purification Kit (Thermo Fisher Scientific, 61011, Waltham, MA) and used as input into the Oxford Nanopore Technologies (ONT) SQK-PCB109 kit (ONT, Oxford, UK) to construct cDNA libraries which were then sequenced on R9 FLO-MIN106 Rev D flowcells on a MinION. Dorado (v0.7.3) github.com/nanoporetech/dorado) employing the SUP model (dna_r9.4.1_e8_sup@v3.6) to base-call the data with the following parameters also enabled; --no-trim --min-qscore 10 --no-classify.

Genome Annotation.

The genome assembly (Table S2 and Table S3) was repeat masked by first creating a custom repeat library (CRL) for the genome. Repeats were first identified with RepeatModeler (v2.03) (Flynn et al. 2020) and the version 3 genome assembly (Kang et al. 2021) and protein coding genes were filtered from the repeat database using ProtExcluder (v1.2) (Campbell et al. 2014) to create a CRL. The CRL was then combined with Viridiplantae repeats from RepBase (v20150807) (Bao, Kojima, and Kohany 2015) to generate the final CRL. The genome assembly was repeat-masked using the final CRL and RepeatMasker (v4.1.2-p1) (Tarailo□Graovac and Chen 2009) using the parameters -e ncbi -s -nolow -no_is -gff (Table S4).

RNA-seq libraries were processed for genome annotation by first cleaning with Cutadapt (v2.10) (Martin 2011) using a minimum length of 50-nt and quality cutoff of 10 then aligning the cleaned reads to the genome using HISAT2 (2.1.0) (D. Kim et al. 2019). Oxford Nanopore (ONT) cDNA reads were processed with Pychopper (v2.5.0) (github.com/epi2me-labs/pychopper) and trimmed reads greater than 500-nt were aligned to the genome using minimap2 (v2.17-r941) (H. Li 2018) with a maximum intron length of 5,000-nt. The aligned RNA-seq and ONT cDNA reads were

each assembled using Stringtie (v2.2.1) (Kovaka et al. 2019) and transcripts less than 500-nt were removed.

Initial gene models were created using BRAKER2 (v2.1.6) (Hoff et al. 2019) using the soft-masked genome assembly and the aligned RNA-seq libraries as hints. The gene models were then refined using two rounds of PASA2 (v2.5.2) (Haas 2003) to create a working gene model set (Table S5). High-confidence gene models were identified from each working gene model by filtering out gene models without expression evidence, or a PFAM domain match, or were a partial gene model or contained an interior stop codon. Functional annotation was assigned by searching the working gene model proteins against the TAIR (v10) (Lamesch et al. 2012) database and the Swiss-Prot plant proteins (release 2015_08) database using BLASTP (v2.12.0) (Camacho et al. 2009) and the PFAM (v35.0) (El-Gebali et al. 2019) database using PfamScan (v1.6) and assigning the annotation based on the first significant hit. Visualization of synteny between v3 and v4 annotations were performed using GENESPACE (Lovell et al. 2022) (Fig. S2A).

Nuclei isolation.

Nuclei isolation was performed as described previously (C. Li et al. 2022) with 0.0375% Triton-X-100 and no EDTA in the nuclei isolation buffer. Half a gram of immature leaves (~3 cm in length) were 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, which was then passed through a 20 µm sieve 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 150,000 nuclei were sorted into 100 µL of nuclei buffer (part of the 10x Genomics Single Cell Multiome Kit). Nuclei were then centrifuged 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. S2B-D). Single cell multiome libraries were constructed according to manufacturer's recommendation (10x Genomics, Single Cell Multiome ATAC + Gene Expression Kit).

Single nuclei RNA-seq processing.

Single nuclei RNA-seq libraries were processed using Cutadapt (v3.5) (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 (v0.16.1) *pair* (Shen et al. 2016). Paired reads were aligned and quantified using STARsolo (Kaminow, Yunusov, and Dobin 2021), with the following parameters: --alignIntronMax 5000 --soloUMIlen 12 --soloCellFilter EmptyDrops_CR --soloFeatures GeneFull --soloMultiMappers EM --soloType CB_UMI_Simple, and --soloCBwhitelist using the latest 10x Genomics whitelist of multiome barcodes. Gene-barcode matrices were analyzed with Seurat (v4) (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. S3, Table S6).

Single nuclei RNA-seq analyses.

Biological replicates were integrated using the `IntegrateData()` function in Seurat using the top 3,000 variable genes. Uniform manifold approximation and projection (UMAP) were performed after a principal component analysis (PCA) using the following parameters: dims = 1:30, min.dist = 0.001, repulsion.strength = 1, n.neighbors = 30, spread = 1. Clustering of cells was performed with a resolution of 0.5 (Fig. 1B). Expression of previously studied biosynthetic genes were plotted (Table S7-S8) across each cell cluster (Fig. 1C). For cell type classification, we used a manually curated marker gene list for mesophyll, epidermis, and vasculature (Table S9), using previously established marker genes from Arabidopsis (J.-Y. Kim et al. 2021; Lopez-Anido et al. 2021) (Fig. S4A). 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 were performed using the `FindAllMarkers()` function in Seurat using the following options: only.pos = T, min.pct = 0.25, logfc.threshold = 0.25. Expression matrix (logCPM values, Supplemental Dataset 1) was 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 10x Genomics Cell Ranger ARC pipeline (https://www.10xgenomics.com/software). The peak bed files for biological replicates were sorted and merged using BEDTools (v2.30) *merge* (Quinlan and Hall 2010). This common set of peaks was used to process both biological replicates. The 'atac_fragments.tsv.gz' files from the Cell Ranger ARC output were used for downstream analyses using Signac (v1.6.0) (Stuart et al. 2021) and Seurat (v4) (Hao et al. 2021). Nuclei were filtered for > 1000 peaks/nuclei, > 2000 fragments/nuclei, and fraction of fragments in peaks > 0.25 (Table S10). 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 (Fig. S4B). 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 and peaks (Fig. S5) was calculated and visualized using deepTools (v3.5.1) (Ramírez et al. 2014).

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 (Fig. 1D, also reported in Supplementary Dataset2). Genes adjacent to STR+ marker peaks were defined using BEDTools (v.2.30) *closest*, with -d option. Only genes within 2-kb to a STR+ marker peak were used in visualization (Fig. 1E). DNA sequence of STR+ maker peaks were extracted using BEDTools (v.2.30) *getfasta* and subjected to *de novo* motif discovery using MEME (v5.4.1) (Bailey et al. 2009): using the following parameters: -dna -revcomp -mod anr -nmotifs 10 -minw 5 -maxw 12 -evt 0.01 (Fig. 1F). Similar motifs were identified from JASPAR2020 collection (Fornes et al. 2019) (Fig. 1G).

Feeding of *C. acuminata* tissue.

All solvents used in this study were of UPLC/MS grade. Information about chemicals and reagents is listed in Table S11. *d5*-tryptamine was synthesized according to McDonald et al. (McDonald et al. 2024) . For isotopologue feeding experiments, fresh plant tissues (10-mm leaf disks, 5-mm meristem segments, 5-mm stem segments, and 5-mm root segments) were excised and incubated individually in 200 µL of aqueous solution containing 500 µM *d5*-tryptamine for 5 days in a growth chamber under conditions of 16-hour daylight at 23°C, followed by night conditions at 21°C, with humidity levels maintained between 40% and 60%. After incubation, residual liquid was discarded, and metabolites were extracted from the tissues using extraction solvent (85% methanol, 15% water, 0.1% formic acid). The extracts were filtered through 0.22 µm PTFE syringe filters. Any remaining liquid was removed, and the samples were extracted with extraction solvent (85% methanol, 15% water, 0.1% formic acid), filtered, and then diluted 10-fold with extraction buffer containing 50 nM catharanthine as an internal standard before LC-MS (Fig. 2).

Leaf and meristem protoplast isolation for single cell metabolomics.

Young leaves (100 mg) measuring ~3 cm in length from the first and second nodes were selected, rinsed gently with water, and cut into 1 mm strips using a sterile surgical blade. Leaf strips were immediately transferred to a Petri dish containing 10 mL of digestion medium composed of 2% (w/v) Cellulase Onozuka R-10, 0.3% (w/v) Macerozyme R-10, and 0.1% (v/v) pectinase in Mannitol-MES (MM) buffer (0.4 M mannitol, 20 mM MES, pH 5.7–5.8 adjusted with 1 M KOH). The open Petri dish was placed in a desiccator and subjected to a vacuum of 100 mBar for 15 min, with periodic vacuum disruption (10s per minute), facilitating infiltration of the digestion medium into the leaf tissue. Following vacuum infiltration, the leaf strips were incubated in the digestion medium at room temperature for 2.5 hours. Post-incubation, the mixture was gently agitated on an orbital shaker at approximately 70 rpm for 30 minutes to facilitate protoplast release. To remove cellular debris, the protoplast suspension was filtered sequentially through nylon sieves of 70 µm and 40 µm mesh sizes. The filtrate was centrifuged at 200×g for 5 minutes at 4°C with gentle acceleration and deceleration to pellet the protoplasts. The supernatant was removed, and the protoplast pellet was washed once with 5 mL of MM buffer. Proto-

plasts were then pooled and resuspended in 2 mL of MM buffer. For protoplast purification via OptiPrep density gradient, a working solution (WS) was prepared by dissolving 6 mg KCl in 1 mL OptiPrep. Layer A was created by mixing 500 µL WS with 2 mL protoplast suspension. Layer B consisted of 266 µL WS combined with 1 mL resting solution, and Layer C contained 200 µL resting solution alone. Layers were sequentially and carefully layered into a 5 mL tube (A, followed by B and C) without disturbing interfaces. The gradient was centrifuged at 200×g for 5 minutes at 4°C, allowing viable protoplasts to float to the top and debris to settle at the bottom. The uppermost 200 µL fraction containing purified protoplasts was collected into a new tube. Protoplast viability and concentration were determined using fluorescein diacetate staining and hemocytometer counting. The final concentration was adjusted to 1 × 10□ protoplasts per mL.

<u>Stem protoplast isolation for single cell metabolomics and single cell RNA-seq.</u>

Young stems from the 2nd to 4th nodes below the shoot apex were sampled in ice-cold water in a 50 mL tube. Using a sterile razor blade, an incision was made longitudinally along the stem segments, and the outer stem layers were carefully peeled away. Both outer and inner stem tissues were immediately transferred into 20 mL of protoplast enzyme cocktail. The cocktail contained 0.4 M sorbitol, 20 mM KCl, 20 mM MES, 2% (w/v) Cellulase Onozuka R-10, 0.3% (w/v) Macerozyme R-10, 0.1% (v/v) pectinase, 1 mM CaCl□, 0.2% (v/v) 2-mercaptoethanol, and 0.1% (w/v) bovine serum albumin (BSA), adjusted to pH 5.7 and heated at 50°C for 10 minutes before cooling to room temperature. Vacuum infiltration of stem segments in the protoplast cocktail was performed for 10 minutes. Following vacuum infiltration, the mixture was shaken at 80 rpm for 70 minutes at room temperature. Subsequently, the mixture was briefly shaken at 100 rpm for 5 minutes to facilitate protoplast release. The suspension was then filtered through sequential nylon sieves of 100 µm and 70 µm mesh sizes to remove debris. Filtrate was centrifuged at 200 × g for 5 minutes at 4°C, after which the protoplast pellet was gently resuspended in 2 mL of resting solution. For protoplast purification via OptiPrep density gradient, the procedure was the same as the leaf/meristem protoplast isolation. Protoplast viability and concentration were determined using fluorescein diacetate staining and hemocytometer counting. The final protoplast concentration was adjusted to 1 × 10□ protoplasts per mL.

<u>Single-cell picking for metabolite profiling.</u>

This procedure (Fig. 3A) was modified based on the method described by Vu et al. (Vu et al. 2024). Single-cell isolation was performed using SIEVEWELL™ chips (Sartorius) containing 90,000 nanowells (50 µm diameter × 50 µm depth). The chips were initially primed with 100% ethanol, rinsed twice with MM buffer, and incubated with 5% BSA in MM buffer for 30 min at room temperature. After removing the BSA solution through side ports, chips were replenished with fresh MM buffer. Approximately 10,000 diluted protoplasts suspended in 1 mL of MM buffer were then gently dispensed across each chip in a Z-shaped pattern. Excess solution was drained via the side ports. Chips were subsequently mounted on a CellCelector™ Flex (Sartorius)

equipped with an optical imaging system comprising a fluorescence microscope (Spectra X Lumencor) and a CCD camera (XM-10). Nanowells were imaged under bright-field or fluorescence (DAPI filter) illumination. Single protoplasts, along with 20 nL of nanowell solution, were individually collected using a 30 μm glass capillary and transferred into SureSTART™ WebSeal™ 96-well microtiter plates (Thermo Fisher Scientific) containing 8 μL MilliQ water with 0.1% formic acid. Images confirming successful picking were captured before and after cell collection. Following single-cell collection, 8 μL methanol containing 20 nM cathanrantine as an internal standard was added to each well. Pooled quality-control (QC) samples, prepared by combining 2 μL from each sample, were included in each experimental batch for QC purposes and MS^2 fragmentation analyses.

LC-MS analysis.

UHPLC-HRMS analysis was conducted using a Vanquish UHPLC system coupled to a Q-Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific). Metabolites were separated on a Waters™ ACQUITY UPLC BEH C18 130 Å column (1.7 μm particle size, 1 mm × 50 mm) maintained at 40 °C. The mobile phase consisted of a binary gradient system with solvent A (MilliQ water with 0.1% formic acid) and solvent B (acetonitrile, ACN). The gradient started at 1% solvent B and increased linearly to 70% solvent B over 5 min, followed by a washing step at 99% solvent B for 0.5 min. The column was subsequently re-equilibrated at 1% solvent B for 1.5 min, resulting in a total run time of 7 min. The flow rate was maintained at 0.3 mL min^{-1} with an injection volume of 4 μL. During the analysis, samples were held at 10 °C in the autosampler, and the autosampler needle was cleaned for 20 s after each injection using a 1:1 (v/v) mixture of methanol and MilliQ water at a washing speed of 50 μL s^{-1}. The Q-Exactive Plus Orbitrap mass spectrometer was operated with a heated electrospray ionization (HESI) source, calibrated using Pierce positive and negative ion mass calibration solutions (Thermo Fisher Scientific). The following HESI parameters were applied based on the UHPLC flow rate of 0.3 mL min^{-1} in source auto-default mode: sheath gas flow rate of 48, auxiliary gas flow rate of 11, sweep gas flow rate of 2, spray voltage of +3,500 V, capillary temperature of 256 °C, auxiliary gas heater temperature of 413 °C, and S-lens RF level of 50. Data acquisition was performed in positive mode using full-scan MS with a resolution of 70,000 (FWHM at 200 Da), scanning across a mass range of *m/z* 120–1,000. Quality control (QC) pooled samples were analyzed using data-dependent MS/MS (full MS/dd-MS^2, Top10), enabling simultaneous acquisition of precursor and fragmentation spectra. Additionally, an inclusion list-based dd-MS^2 analysis was conducted on QC samples to confirm fragmentation patterns of targeted precursors. Parameters for dd-MS^2 analyses included a resolution of 17,500, an isolation window of 0.7 Da, and normalized collision energies (NCE) at three levels: 15%, 30%, and 45%. Spectra were acquired in centroid mode, and all instrument parameters were controlled via Xcalibur software (version 4.3.73.11, Thermo Fisher Scientific). Standard solutions of alkaloids, and iridoids were prepared at approximately 1 mM in methanol, with exact concentrations recorded. UHPLC-MS analyzed serial dilutions down to 0.01 nM to establish the limit of quantification (LOQ) and calibration ranges (Table S12), with

each calibration concentration measured in triplicate. Chromatographic peak areas from extracted ion chromatograms (EIC) were integrated and quantified using Xcalibur Quan Browser software (version 4.3.73.11, Thermo Fisher Scientific). Concentration of key metabolites across single cells (Fig. 3B, Fig. S5) were also reported as Supplementary Datasets 3-5, for leaf, meristem, and stem, respectively.

PIP-seq library construction and data processing.

Stem protoplasts were isolated and used as input into the PIPseq v4.0 Plus kit (Illumina, San Diego, CA) to generate scRNA-seq libraries per manufacturer recommendations. The libraries were sequenced on an Aviti instrument (Element Biosciences, San Diego, CA). The subsequent reads were processed using pipseeker (v3.2.0) (Illumina) which trimmed read two, generated a barcode whitelist and prepared the reads for downstream analysis. Gene-barcode matrices were analyzed with Seurat (v4) (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. S7, Table S13).

Single cell RNA-seq analyses using PIP-seq data.

Biological replicates were integrated using the `IntegrateData()` function in Seurat using the top 3,000 variable genes. Uniform manifold approximation and projection (UMAP) were performed after a principal component analysis (PCA) using the following parameters: dims = 1:30, min.dist = 0.001, repulsion.strength = 1, n.neighbors = 30, spread = 1. Clustering of cells was performed with a resolution of 0.5 (Fig. 3C). Expression of previously studied biosynthetic genes were plotted (Table S7-S8) across each cell cluster (Fig. 3D). For cell type classification, we used a manually curated marker gene list stem (Table S14), using previously established marker genes from a poplar stem single cell RNA-seq dataset (Conde et al. 2022) (Fig. S8). Expression matrix (logCPM values, Supplemental Dataset 6) was generated at the cell cluster level, where each row is a gene, and each column is a cell cluster.

Gene co-expression analyses.

Gene co-expression analysis by graph-based clustering was performed as previously described (C. Li, Deans, and Buell 2023). Graph-based clustering was performed using the Leiden algorithm (Traag, Waltman, and van Eck 2019), which was implemented in R by the igraph package (Csardi and Nepusz 2006). Expression of co-expressed TFs were visualized in Fig. 4A.

Phylogenetic analysis of MYB and bHLH domain TFs.

We used “MYB annotator” (Pucker 2022) and “bHLH annotator” (Thoben and Pucker 2023) to automatically identify MYB domain proteins in *C. acuminata* and *C. roseus* (Supplementary Datasets 7-8). We aligned the sequences of MYB domains or bHLH domains using MAFFT (v7.505) (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. We used ggtree (Yu et al. 2017) to perform visualization of phylogenies (Fig. 4B-D). Expression patterns of *C. roseus* across single cell datasets (Fig. 4E-F) were visualized like Fig. 1B. Expression of stem MIA co-expressed TFs in leaf was visualized using FeaturePlot() function of Seurat (Fig. S9A-C). Expression data of methyl jasmonate treatment experiments (Van Moerkercke et al. 2013), sorted idioblast, and sorted mesophyll (Guedes et al. 2024) were trimmed using Cutadapt (v3.5) (Martin 2011) and pseudo-aligned and quantified using kallisto (v0.48) (Bray et al. 2016), with the --plaintext option turned on (Fig. 4G, Fig. S9D). ATAC-seq signal across cell types at *CrIDB1* and *CrIDM4* loci in leaf (Fig. S9E-F) were visualized using CoveragePlot() function of Signac in R.

<u>Overexpression assays.</u>

Coding sequences of TFs of interest were cloned in between the 35S promoter and 35S terminator and transformed into *Agrobacterium tumefaciens* strain GV3101. Transient expression experiments were done on *C. roseus* (cultivar “Little Bright Eyes”) petals. Infiltration was done as previously described (Colinas and Goossens 2022). Two days before the infiltration, all open flowers were removed. We performed infiltrations at 0.4 optical density (OD) which is the highest OD that can be used without resulting in wilting of the petals after infiltration (Fig. 5A). Two days after the infiltration, infiltrated petals were harvested and stored in a -80 freezer until RNA extraction. RNA extraction was performed with a Monarch Total RNA Miniprep Kit.

<u>RNA-seq analysis for overexpression samples.</u>

Sequencing adapters were trimmed from petal RNA-seq libraries using Cutadapt (v3.5) (Martin 2011). Adapter trimmed libraries were pseudo-aligned and quantified using kallisto (v0.48) (Bray et al. 2016), with the --plaintext option turned on (Fig. 5B). Differential gene expression analyses were performed using DESeq2 (v.1.34.0) (Love, Huber, and Anders 2014), using the *AmCyan*-only treatment (Treatment M) as control. Genes with adjusted p values < 0.05 were taken as differentially expressed genes (Fig. 5C-D). Target genes of TFs (Fig. 6, Fig. S10) were defined as genes differentially upregulated and have the highest expression in the idioblast. Gene expression matrix was also reported as Supplementary Dataset 9.

Data Availability

All sequencing data associated with this study are available at the National Center for Biotechnology Institute Sequence Read Archive BioProject PRJNA1243444 (to be made available upon publication). Genome assembly, annotation, and Seurat objects for single nuclei/cell RNA-seq and/or ATAC-seq experiments are available via the online digital repository figshare (to be made public upon publication). Plasmid maps and sequences of MYB and bHLH domains are available at Zenodo (to be made available upon publication).

Code Availability

Custom codes used to generate figures can be found at https://github.com/cxli233/Camptotheca_single_cell.

Acknowledgements

This project was supported by the University of Georgia, Georgia Research Alliance (C.R.B.) and Georgia Seed Development (C.R.B.), National Science Foundation MCB-2309665 (C.R.B and C.L.), the Max Planck Gesellschaft (L.C.). T.T.T.D. receives funding from NSERC Alliance Collaboration (ALLRP 571673 – 21) and Catalyst (ALLRP 579871 – 22). T.T.T. Dang is grateful for the Michael Smith Health BC Scholar Award (SCH-2020-0401). V.H.B. received an International Four-Year Doctoral Partial Tuition Award and Graduate Research Scholarship. Sequencing was performed at the University of Minnesota Genomics Center (UMGC), Texas A&M AgriLife Research: Genomics and Bioinformatics Service, and UC Davis DNA Technologies and Expression Analysis Core. We thank Kathrine Mailloux for high molecular weight DNA isolation. We thank Keira Mchugh and Sreeja Challa for assistance in molecular cloning and RNA isolation. We thank Dr. Jon Cody and Dr. Daniel Voytas for providing *AmCyan* plasmid. We thank Dr. Maritta Kunert and Sarah Heinicke for their assistance with mass spectrometry analyses, and Jens Wurlitzer for support with the cell-picking robot. We thank Dr. Allwin McDonald for generously providing d5-tryptamine. We thank Eva Rothe and the greenhouse team for plant care. Additionally, we thank Dr. Moonyoung Kang and Tingan Zhou for their insightful comments on the protoplast isolation protocol.

Author Contributions

CL, C.R.B., and T.T.T.D. designed the study. V.H.B generated single cell metabolomics and isotope feeding data with assistance from A.H.V. and L.C. C.L. generated single nuclei multiome and single cell RNA-seq libraries. L.H.C. performed cDNA sequencing. J.C.W. performed single cell library construction and quality control. B.V. performed genome assembly and data management. J.P.H. performed genome annotation. C.L. and S.J. performed overexpression assays. C.L., J.C.W., B.V., J.P.H., and V.H.B performed data analyses. C.L. and V.H.B wrote the manuscript with input from all authors.

Conflict of Interest Statement

The authors have declared no conflict of interest.

Supplementary figures, tables, and datasets

Supplementary Figures:

Fig. S1. Key monoterpene indole alkaloids of *Camptotheca acuminata*.

Fig. S2. *Camptotheca acuminata* v4 genome assembly and annotation and nuclei preparation for single nuclei omics.

Fig. S3. Characterization of leaf single nuclei RNA-seq libraries.

Fig. S4. Maker genes and multimodal integration for leaf single nuclei omics datasets.
Fig. S5. Characterization of leaf single nuclei ATAC-seq libraries.
Fig. S6. Metabolite concentrations measured in individual leaf/meristem/stem-derived cells.
Fig. S7. Characterization of stem single cell RNA-seq libraries.
Fig. S8. Marker gene for stem single cell RNA-seq dataset.
Fig. S9. Expression of transcription factors of interest in leaf and the expression of *Catharanthus roseus* orthologs.
Fig. S10. Expression of the idioblast regulon.

<u>Supplementary Tables:</u>
Table S1. List of libraries and sequencing data generated by this study.
Table S2. Assembly statistics
Table S3. BUSCO for assembly and annotation
Table S4. Repetitive sequence content
Table S5. Gene annotation summary
Table S6. Details of leaf single nuclei RNA-seq libraries
Table S7. Gene names of known monoterpene indole alkaloid biosynthetic genes in *Campotheca acuminata*.
Table S8. Gene names of known monoterpene indole alkaloid biosynthetic genes in *Catharanthus roseus*.
Table S9. Leaf cell type marker genes.
Table S10. Details of leaf single nuclei ATAC-seq libraries.
Table S11. List of available standards used for the scMS method.
Table S12. Validation of the targeted scMS analysis.
Table S13. Details of stem single cell RNA-seq libraries.
Table S14. Stem cell type marker genes.

<u>Supplementary Datasets:</u>
1. Gene expression matrix (log-normalized counts) across leaf cell clusters.
2. Genome coordinates of leaf cell type marker peaks.
3. Quantification of iridoids and alkaloids in single cells from leaf
4. Quantification of iridoids and alkaloids in single cells from meristem
5. Quantification of iridoids and alkaloids in single cells from stem
6. Gene expression matrix (log-normalized counts) across stem cell clusters.
7. Phylogeny of Arabidopsis, *Catharanthus roseus*, and *Camptotheca acuminata* MYB TFs.
8. Phylogeny of Arabidopsis, *Catharanthus roseus*, and *Camptotheca acuminata* bHLH TFs.
9. Gene expression matrix (transcripts per million) of overexpression experiment.

Reference

Alonge, Michael, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J. Sedlazeck, Zachary B. Lippman, and Michael C. Schatz. 2019. "RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes." *Genome Biology* 20 (1): 224. https://doi.org/10.1186/s13059-019-1829-6.

Awadasseid, Annoor, Wei Li, Zhan Liu, Chong Qiao, Jing Pang, Guolin Zhang, and Yinggang Luo. 2020. "Characterization of *Camptotheca Acuminata* 10-Hydroxygeraniol Oxidoreductase and Iridoid Synthase and Their Application in Biological Preparation of Nepetalactol in *Escherichia Coli* Featuring NADP+ - NADPH Cofactors Recycling." *International Journal of Biological Macromolecules* 162 (November):1076–85. https://doi.org/10.1016/j.ijbiomac.2020.06.223.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. "MEME SUITE: Tools for Motif Discovery and Searching." *Nucleic Acids Research* 37 (Web Server): W202–8. https://doi.org/10.1093/nar/gkp335.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. 2015. "Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes." *Mobile DNA* 6 (1): 11. https://doi.org/10.1186/s13100-015-0041-9.

Baudry, Antoine, Michel Caboche, and Loïc Lepiniec. 2006. "TT8 Controls Its Own Expression in a Feedback Regulation Involving TTG1 and Homologous MYB and bHLH Factors, Allowing a Strong and Cell-specific Accumulation of Flavonoids in *Arabidopsis Thaliana*." *The Plant Journal* 46 (5): 768–79. https://doi.org/10.1111/j.1365-313X.2006.02733.x.

Bray, Nicolas L, Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Burlat, Vincent, Audrey Oudin, Martine Courtois, Marc Rideau, and Benoit St-Pierre. 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 (1): 131–41. https://doi.org/10.1111/j.1365-313X.2004.02030.x.

Cabanettes, Floréal, and Christophe Klopp. 2018. "D-GENIES: Dot Plot Large Genomes in an Interactive, Efficient and Simple Way." Edited by Thomas Tullius. *PeerJ* 6 (June):e4958. https://doi.org/10.7717/peerj.4958.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Campbell, Michael S., MeiYee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, et al. 2014. "MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations." *Plant Physiology* 164 (2): 513–24. https://doi.org/10.1104/pp.113.230144.

Chen, Fei, Wei Li, Liangzhen Jiang, Xiang Pu, Yun Yang, Guolin Zhang, and Yinggang Luo. 2016. "Functional Characterization of a Geraniol Synthase-Encoding Gene from Camptotheca Acuminata and Its Application in Production of Geraniol in

Escherichia Coli.” *Journal of Industrial Microbiology and Biotechnology* 43 (9): 1281–92. https://doi.org/10.1007/s10295-016-1802-2.

Cheng, Haoyu, Mobin Asri, Julian Lucas, Sergey Koren, and Heng Li. 2024. “Scalable Telomere-to-Telomere Assembly for Diploid and Polyploid Genomes with Double Graph.” *Nature Methods* 21 (6): 967–70. https://doi.org/10.1038/s41592-024-02269-8.

Clark, Iain C., Kristina M. Fontanez, Robert H. Meltzer, Yi Xue, Corey Hayford, Aaron May-Zhang, Chris D’Amato, et al. 2023. “Microfluidics-Free Single-Cell Genomics with Templated Emulsification.” *Nature Biotechnology* 41 (11): 1557–66. https://doi.org/10.1038/s41587-023-01685-z.

Colinas, Maite, and Alain Goossens. 2022. “Transient Gene Expression in Catharanthus Roseus Flower Petals Using Agroinfiltration.” In *Catharanthus Roseus*, edited by Vincent Courdavault and Sebastien Besseau, 2505:281–91. Methods in Molecular Biology. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-2349-7_20.

Conde, Daniel, Paolo M. Triozzi, Wendell J. Pereira, Henry W. Schmidt, Kelly M. Balmant, Sara A. Knaack, Arturo Redondo-López, Sushmita Roy, Christopher Dervinis, and Matias Kirst. 2022. “Single-Nuclei Transcriptome Analysis of the Shoot Apex Vascular System Differentiation in *Populus*.” *Development* 149 (21): dev200632. https://doi.org/10.1242/dev.200632.

Csardi, Gabor, and Tamas Nepusz. 2006. “The Igraph Software Package for Complex Network Research,” 9.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, et al. 2019. “The Pfam Protein Families Database in 2019.” *Nucleic Acids Research* 47 (D1): D427–32. https://doi.org/10.1093/nar/gky995.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. “RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families.” *Proceedings of the National Academy of Sciences* 117 (17): 9451–57. https://doi.org/10.1073/pnas.1921046117.

Fornes, Oriol, Jaime A Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A Richmond, Bhavi P Modi, et al. 2019. “JASPAR 2020: Update of the Open-Access Database of Transcription Factor Binding Profiles.” *Nucleic Acids Research*, November, gkz1001. https://doi.org/10.1093/nar/gkz1001.

Giovanella, B C, J S Stehlin, M E Wall, M C Wani, A W Nicholas, L F Liu, R Silber, and M Potmesil. 1989. “DNA Topoisomerase I--Targeted Chemotherapy of Human Colon Cancer in Xenografts.” *Science (New York, N.Y.)* 246 (4933): 1046–48. https://doi.org/2555920.

Guedes, Joana G, Rogério Ribeiro, Inês Carqueijeiro, Ana Luísa Guimarães, Cláudia Bispo, John Archer, Herlander Azevedo, Nuno A Fonseca, and Mariana Sottomayor. 2024. “The Leaf Idioblastome of the Medicinal Plant *Catharanthus Roseus* Is Associated with Stress Resistance and Alkaloid Metabolism.” Edited by Sonia Osorio. *Journal of Experimental Botany* 75 (1): 274–99. https://doi.org/10.1093/jxb/erad374.

Guirimand, Grégory, Anthony Guihur, Pierre Poutrain, François Héricourt, Samira Mahroug, Benoit St-Pierre, Vincent Burlat, and Vincent Courdavault. 2011. “Spa-

tial Organization of the Vindoline Biosynthetic Pathway in Catharanthus Roseus." *Journal of Plant Physiology* 168 (6): 549–57. https://doi.org/10.1016/j.jplph.2010.08.018.

Haas, B. J. 2003. "Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies." *Nucleic Acids Research* 31 (19): 5654–66. https://doi.org/10.1093/nar/gkg770.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Hoff, Katharina J, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2019. "Whole-Genome Annotation with BRAKER." *Gene Prediction: Methods and Protocols*, 65–95.

Kaminow, Benjamin, Dinar Yunusov, and Alexander Dobin. 2021. "STARsolo: Accurate, Fast and Versatile Mapping/Quantification of Single-Cell and Single-Nucleus RNA-Seq Data." *bioRxiv*, May. https://doi.org/10.1101/2021.05.05.442755.

Kang, Minghui, Rao Fu, Pingyu Zhang, Shangling Lou, Xuchen Yang, Yang Chen, Tao Ma, Yang Zhang, Zhenxiang Xi, and Jianquan Liu. 2021. "A Chromosome-Level Camptotheca Acuminata Genome Assembly Provides Insights into the Evolutionary Origin of Camptothecin Biosynthesis." *Nature Communications* 12 (1): 3531. https://doi.org/10.1038/s41467-021-23872-9.

Katoh, K. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37 (8): 907–15. https://doi.org/10.1038/s41587-019-0201-4.

Kim, Ji-Yun, Efthymia Symeonidi, Tin Yau Pang, Tom Denyer, Diana Weidauer, Margaret Bezrutczyk, Manuel Miras, et al. 2021. "Distinct Identities of Leaf Phloem Cells Revealed by Single Cell Transcriptomics." *The Plant Cell* 33 (3): 511–30. https://doi.org/10.1093/plcell/koaa060.

Kovaka, Sam, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea. 2019. "Transcriptome Assembly from Long-Read RNA-Seq Alignments with StringTie2." *Genome Biology* 20 (1): 278. https://doi.org/10.1186/s13059-019-1910-1.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times." *Molecular Biology and Evolution* 34 (7): 1812–19. https://doi.org/10.1093/molbev/msx116.

Lamesch, Philippe, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, et al. 2012. "The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools." *Nucleic Acids Research* 40 (D1): D1202–10. https://doi.org/10.1093/nar/gkr1090.

Li, Chenxin, Maite Colinas, Joshua C. Wood, Brieanne Vaillancourt, John P. Hamilton, Sophia L. Jones, Lorenzo Caputi, Sarah E. O'Connor, and C. Robin Buell. 2025. "Cell-type-aware Regulatory Landscapes Governing Monoterpene Indole Alkaloid

Biosynthesis in the Medicinal Plant *Catharanthus Roseus*." *New Phytologist* 245 (1): 347–62. https://doi.org/10.1111/nph.20208.

Li, Chenxin, Natalie C. Deans, and C. Robin Buell. 2023. "'Simple Tidy GeneCoEx': A Gene Co-expression Analysis Workflow Powered by Tidyverse and Graph-based Clustering in R." *The Plant Genome* 16 (2): e20323. https://doi.org/10.1002/tpg2.20323.

Li, Chenxin, Joshua C. Wood, Natalie C. Deans, Anne Frances Jarrell, Dionne Martin, Kathrine Mailloux, Yi-Wen Wang, and C. Robin Buell. 2022. "Nuclei Isolation Protocol from Diverse Angiosperm Species." *bioRxiv*, November. https://doi.org/10.1101/2022.11.03.515090.

Li, Chenxin, Joshua C. Wood, Anh Hai Vu, John P. Hamilton, Carlos Eduardo Rodriguez Lopez, Richard M. E. Payne, Delia Ayled Serna Guerrero, et al. 2023. "Single-Cell Multi-Omics in the Medicinal Plant Catharanthus Roseus." *Nature Chemical Biology* 19 (8): 1031–41. https://doi.org/10.1038/s41589-023-01327-0.

Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." Edited by Inanc Birol. *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Ke, Jinjie Ou, and Shuanhu Gao. 2016. "Total Synthesis of Camptothecin and Related Natural Products by a Flexible Strategy." *Angewandte Chemie International Edition* 55 (47): 14778–83. https://doi.org/10.1002/anie.201607832.

Li, Song Feng, Olga Nicolaou Milliken, Hanh Pham, Reg Seyit, Ross Napoli, Jeremy Preston, Anna M. Koltunow, and Roger W. Parish. 2009. "The *Arabidopsis* MYB5 Transcription Factor Regulates Mucilage Synthesis, Seed Coat Development, and Trichome Morphogenesis." *The Plant Cell* 21 (1): 72–89. https://doi.org/10.1105/tpc.108.063503.

Longchar, Bendangchuchang, Tarinee Phukan, Sarita Yadav, and Muthappa Senthil-Kumar. 2020. "An Efficient Low-cost Xylem Sap Isolation Method for Bacterial Wilt Assays in Tomato." *Applications in Plant Sciences* 8 (4): e11335. https://doi.org/10.1002/aps3.11335.

Lopez-Anido, Camila B., Anne Vatén, Nicole K. Smoot, Nidhi Sharma, Victoria Guo, Yan Gong, M. Ximena Anleu Gil, Annika K. Weimer, and Dominique C. Bergmann. 2021. "Single-Cell Resolution of Lineage Trajectories in the Arabidopsis Stomatal Lineage and Developing Leaf." *Developmental Cell* 56 (7): 1043-1055.e4. https://doi.org/10.1016/j.devcel.2021.03.014.

Lorence, Argelia, and Craig L. Nessler. 2004. "Camptothecin, over Four Decades of Surprising Findings." *Phytochemistry* 65 (20): 2735–49. https://doi.org/10.1016/j.phytochem.2004.09.001.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Lovell, John T, Avinash Sreedasyam, M Eric Schranz, Melissa Wilson, Joseph W Carlson, Alex Harkess, David Emms, David M Goodstein, and Jeremy Schmutz. 2022. "GENESPACE Tracks Regions of Interest and Gene Copy Number Variation across Multiple Genomes." *eLife* 11 (September):e78526. https://doi.org/10.7554/eLife.78526.

Manni, Mosè, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. 2021. “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.” *Molecular Biology and Evolution* 38 (10): 4647–54. https://doi.org/10.1093/molbev/msab199.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet* 17 (1): 3.

McClune, Conor James, Jack Chun-Ting Liu, Chloe Wick, Ricardo De La Peña, Bernd Markus Lange, Polly M. Fordyce, and Elizabeth S. Sattely. 2024. “Multiplexed Perturbation of Yew Reveals Cryptic Proteins That Enable a Total Biosynthesis of Baccatin III and Taxol Precursors.” *bioRxiv*, January, 2024.11.06.622305. https://doi.org/10.1101/2024.11.06.622305.

McDonald, Allwin, Yoko Nakamura, Carsten Schotte, Kin Lau, Ryan Alam, Adriana A Lopes, C Robin Buell, and Sarah O’Connor. 2024. “Enzymatic Epimerization of Monoterpene Indole Alkaloids in Kratom.” *bioRxiv*, 2024–12.

Miettinen, Karel, Lemeng Dong, Nicolas Navrot, Thomas Schneider, Vincent Burlat, Jacob Pollier, Lotte Woittiez, et al. 2014. “The Seco-Iridoid Pathway from Catharanthus Roseus.” *Nature Communications* 5 (1): 3606. https://doi.org/10.1038/ncomms4606.

Millard, Peter S, Konrad Weber, Birthe B Kragelund, and Meike Burow. 2019. “Specificity of MYB Interactions Relies on Motifs in Ordered and Disordered Contexts.” *Nucleic Acids Research* 47 (18): 9592–9608. https://doi.org/10.1093/nar/gkz691.

Nesi, Nathalie, Clarisse Jond, Isabelle Debeaujon, Michel Caboche, and Loïc Lepiniec. 2001. “The Arabidopsis TT2 Gene Encodes an R2R3 MYB Domain Protein That Acts as a Key Determinant for Proanthocyanidin Accumulation in Developing Seed.”

Nguyen, Tuan-Anh M., Trinh-Don Nguyen, Yuen Yee Leung, Matthew McConnachie, Oleg Sannikov, Zhicheng Xia, and Thu-Thuy T. Dang. 2021. “Discovering and Harnessing Oxidative Enzymes for Chemoenzymatic Synthesis and Diversification of Anticancer Camptothecin Analogues.” *Communications Chemistry* 4 (1): 177. https://doi.org/10.1038/s42004-021-00602-2.

O’Connor, Sarah E., and Justin J. Maresh. 2006. “Chemistry and Biology of Monoterpene Indole Alkaloid Biosynthesis.” *Natural Product Reports* 23 (4): 532. https://doi.org/10.1039/b512615k.

Pan, Xichun, Min Chen, Yan Liu, Qiang Wang, Lingjiang Zeng, Lianqiang Li, and Zhihua Liao. 2008. “A New Isopentenyl Diphosphate Isomerase Gene from Camptotheca Acuminata: Cloning, Characterization and Functional Expression in Escherichia Coli: Full Length Research Paper.” *DNA Sequence* 19 (2): 98–105. https://doi.org/10.1080/10425170701446509.

Pucker, Boas. 2022. “Automatic Identification and Annotation of MYB Gene Family Members in Plants.” *BMC Genomics* 23 (1): 220. https://doi.org/10.1186/s12864-022-08452-5.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Ramírez, Fidel, Friederike Dündar, Sarah Diehl, Björn A. Grüning, and Thomas Manke. 2014. “deepTools: A Flexible Platform for Exploring Deep-Sequencing Data.” *Nucleic Acids Research* 42 (W1): W187–91. https://doi.org/10.1093/nar/gku365.

Sadre, Radin, Maria Magallanes-Lundback, Sujana Pradhan, Vonny Salim, Alex Mesberg, A. Daniel Jones, and Dean DellaPenna. 2016. “Metabolite Diversity in Alkaloid Biosynthesis: A Multilane (Diastereomer) Highway for Camptothecin Synthesis in *Camptotheca* Acuminata.” *The Plant Cell* 28 (8): 1926–44.

Salim, Vonny, A Daniel Jones, and Dean DellaPenna. 2018. “Camptotheca Acuminata 10-Hydroxycamptothecin O-Methyltransferase: An Alkaloid Biosynthetic Enzyme Co-Opted from Flavonoid Metabolism.” *The Plant Journal* 95 (1): 112–25.

Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. 2016. “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.” Edited by Quan Zou. *PLOS ONE* 11 (10): e0163962. https://doi.org/10.1371/journal.pone.0163962.

Simkin, Andrew J., Karel Miettinen, Patricia Claudel, Vincent Burlat, Grégory Guirimand, Vincent Courdavault, Nicolas Papon, 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 (January):36–43. https://doi.org/10.1016/j.phytochem.2012.09.014.

Singh, Sanjay Kumar, Barunava Patra, Priyanka Paul, Yongliang Liu, Sitakanta Pattanaik, and Ling Yuan. 2021. “*BHLH IRIDOID SYNTHESIS 3* Is a Member of a bHLH Gene Cluster Regulating Terpenoid Indole Alkaloid Biosynthesis in *Catharanthus Roseus*.” *Plant Direct* 5 (1). https://doi.org/10.1002/pld3.305.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Stuart, Tim, Avi Srivastava, Shaista Madad, Caleb A. Lareau, and Rahul Satija. 2021. “Single-Cell Chromatin State Analysis with Signac.” *Nature Methods* 18 (11): 1333–41. https://doi.org/10.1038/s41592-021-01282-5.

Tarailo-Graovac, Maja, and Nansheng Chen. 2009. “Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences.” *Current Protocols in Bioinformatics* 25 (1). https://doi.org/10.1002/0471250953.bi0410s25.

Thoben, Corinna, and Boas Pucker. 2023. “Automatic Annotation of the bHLH Gene Family in Plants.” *BMC Genomics* 24 (1): 780. https://doi.org/10.1186/s12864-023-09877-2.

Traag, V. A., L. Waltman, and N. J. van Eck. 2019. “From Louvain to Leiden: Guaranteeing Well-Connected Communities.” *Scientific Reports* 9 (1): 5233. https://doi.org/10.1038/s41598-019-41695-z.

Van Moerkercke, Alex, Michele Fabris, Jacob Pollier, Gino J.E. Baart, Stephane Rombauts, Ghulam Hasnain, Heiko Rischer, Johan Memelink, Kirsi-Marja Oksman-Caldentey, and Alain Goossens. 2013. “CathaCyc, a Metabolic Pathway Database Built from Catharanthus Roseus RNA-Seq Data.” *Plant and Cell Physiology* 54 (5): 673–85. https://doi.org/10.1093/pcp/pct039.

Van Moerkercke, Alex, Priscille Steensma, Ivo Gariboldi, Javiera Espoz, Purin C. Purnama, Fabian Schweizer, Karel Miettinen, 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 (1): 3–12. https://doi.org/10.1111/tpj.13230.

Van Moerkercke, Alex, Priscille Steensma, Fabian Schweizer, Jacob Pollier, Ivo Gariboldi, Richard Payne, Robin Vanden Bossche, 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* 112 (26): 8130–35. https://doi.org/10.1073/pnas.1504951112.

Vu, Anh Hai, Moonyoung Kang, Jens Wurlitzer, Sarah Heinicke, Chenxin Li, Joshua C. Wood, Veit Grabe, C. Robin Buell, Lorenzo Caputi, and Sarah E. O'Connor. 2024. "Quantitative Single-Cell Mass Spectrometry Provides a Highly Resolved Analysis of Natural Product Biosynthesis Partitioning in Plants." *Journal of the American Chemical Society*, August, jacs.4c06336. https://doi.org/10.1021/jacs.4c06336.

Wall, Monroe E., M. C. Wani, C. E. Cook, Keith H. Palmer, A. T. McPhail, and G. A. Sim. 1966. "Plant Antitumor Agents. I. The Isolation and Structure of Camptothecin, a Novel Alkaloidal Leukemia and Tumor Inhibitor from Camptotheca Acuminata 1,2." *Journal of the American Chemical Society* 88 (16): 3888–90. https://doi.org/10.1021/ja00968a057.

Wan, C. Y., and T. A. Wilkins. 1994. "A Modified Hot Borate Method Significantly Enhances the Yield of High-Quality RNA from Cotton (Gossypium Hirsutum L.)." *Analytical Biochemistry* 223 (1): 7–12. https://doi.org/10.1006/abio.1994.1538.

Weng, Jing-Ke, Joseph H. Lynch, Jason O. Matos, and Natalia Dudareva. 2021. "Adaptive Mechanisms of Plant Specialized Metabolism Connecting Chemistry to Function." *Nature Chemical Biology* 17 (10): 1037–45. https://doi.org/10.1038/s41589-021-00822-6.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. "Improved Metagenomic Analysis with Kraken 2." *Genome Biology* 20 (1): 257. https://doi.org/10.1186/s13059-019-1891-0.

Wu, Song, Ana Luisa Malaco Morotti, Jun Yang, Ertao Wang, and Evangelos C. Tatsis. 2024. "Single-Cell RNA Sequencing Facilitates the Elucidation of the Complete Biosynthesis of the Antidepressant Hyperforin in St. John's Wort." *Molecular Plant* 17 (9): 1439–57. https://doi.org/10.1016/j.molp.2024.08.003.

Yamamoto, Kotaro, Dagny Grzech, Konstantinos Koudounas, Emily Amor Stander, Lorenzo Caputi, Tetsuro Mimura, Vincent Courdavault, and Sarah E O'Connor. 2021. "Improved Virus-Induced Gene Silencing Allows Discovery of a Serpentine Synthase Gene in Catharanthus Roseus." *Plant Physiology* 187 (2): 846–57.

Yamamoto, Kotaro, Katsutoshi Takahashi, Lorenzo Caputi, Hajime Mizuno, Carlos E. Rodriguez-Lopez, Tetsushi Iwasaki, Kimitsune Ishizaki, et al. 2019. "The Complexity of Intercellular Localisation of Alkaloids Revealed by Single-cell Metabolomics." *New Phytologist* 224 (2): 848–59. https://doi.org/10.1111/nph.16138.

Yamamoto, Kotaro, Katsutoshi Takahashi, Hajime Mizuno, Aya Anegawa, Kimitsune Ishizaki, Hidehiro Fukaki, Miwa Ohnishi, Mami Yamazaki, Tsutomu Masujima, and Tetsuro Mimura. 2016. "Cell-Specific Localization of Alkaloids in Catharanthus Roseus Stem Tissue Measured with Imaging MS and Single-Cell MS." *Proc. Natl. Acad. Sci. U. S. A.* 113 (14): 3891–96. https://doi.org/10.1073/pnas.1521959113.

Yamazaki, Yasuyo, Hiroshi Sudo, Mami Yamazaki, Norio Aimi, and Kazuki Saito. 2003. "Camptothecin Biosynthetic Genes in Hairy Roots of Ophiorrhiza Pumila: Cloning, Characterization and Differential Expression in Tissues and by Stress Compounds." *Plant and Cell Physiology* 44 (4): 395–403. https://doi.org/10.1093/pcp/pcg051.

Yang, Lixia, Liangzhen Jiang, Wei Li, Yun Yang, Guolin Zhang, and Yinggang Luo. 2017. "A Homomeric Geranyl Diphosphate Synthase-Encoding Gene from Camptotheca Acuminata and Its Combinatorial Optimization for Production of Geraniol in Escherichia Coli." *Journal of Industrial Microbiology and Biotechnology* 44 (10): 1431–41. https://doi.org/10.1007/s10295-017-1967-3.

Yang, Yun, Wei Li, Jing Pang, Liangzhen Jiang, Xixing Qu, Xiang Pu, Guolin Zhang, and Yinggang Luo. 2019. "Bifunctional Cytochrome P450 Enzymes Involved in Camptothecin Biosynthesis." *ACS Chemical Biology* 14 (6): 1091–96. https://doi.org/10.1021/acschembio.8b01124.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. "GGTREE: An R Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data." Edited by Greg McInerny. *Methods in Ecology and Evolution* 8 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.

Zhan, Xiaori, Tian Qiu, Hongshan Zhang, Kailin Hou, Xueshuang Liang, Cheng Chen, Zhijing Wang, et al. 2023. "Mass Spectrometry Imaging and Single-Cell Transcriptional Profiling Reveal the Tissue-Specific Regulation of Bioactive Ingredient Biosynthesis in Taxus Leaves." *Plant Communications* 4 (5). https://doi.org/10.1016/j.xplc.2023.100630.