Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory

Dapeng Li[a], Sven Heiling[a], Ian T. Baldwin[a], and Emmanuel Gaquerel[a,b,1]

[a]Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, 07745 Jena, Germany; and [b]Centre for Organismal Studies, University of Heidelberg, 69120 Heidelberg, Germany

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved October 11, 2016 (received for review June 23, 2016)

Secondary metabolite diversity is considered an important fitness determinant for plants' biotic and abiotic interactions in nature. This diversity can be examined in two dimensions. The first one considers metabolite diversity across plant species. A second way of looking at this diversity is by considering the tissue-specific localization of pathways underlying secondary metabolism within a plant. Although these cross-tissue metabolite variations are increasingly regarded as important readouts of tissue-level gene function and regulatory processes, they have rarely been comprehensively explored by nontargeted metabolomics. As such, important questions have remained superficially addressed. For instance, which tissues exhibit prevalent signatures of metabolic specialization? Reciprocally, which metabolites contribute most to this tissue specialization in contrast to those metabolites exhibiting housekeeping characteristics? Here, we explore tissue-level metabolic specialization in *Nicotiana attenuata*, an ecological model with rich secondary metabolism, by combining tissue-wide nontargeted mass spectral data acquisition, information theory analysis, and tandem MS (MS/MS) molecular networks. This analysis was conducted for two different methanolic extracts of 14 tissues and deconvoluted 895 nonredundant MS/MS spectra. Using information theory analysis, anthers were found to harbor the most specialized metabolome, and most unique metabolites of anthers and other tissues were annotated through MS/MS molecular networks. Tissue–metabolite association maps were used to predict tissue-specific gene functions. Predictions for the function of two UDP-glycosyltransferases in flavonoid metabolism were confirmed by virus-induced gene silencing. The present workflow allows biologists to amortize the vast amount of data produced by modern MS instrumentation in their quest to understand gene function.

secondary metabolism | mass spectrometry | metabolomics | information theory | *Nicotiana attenuata*

Plants are elegant synthetic chemists making use of their metabolic prowess to produce complex blends of structurally diverse chemicals. Commonly quoted estimates state that plants produce somewhere on the order of 200,000 chemical structures. Secondary metabolites, also referred to as specialized metabolites or natural products, contribute to the largest fraction of this structural diversity. Compared with their counterparts in central metabolism (primary metabolites), secondary metabolite groups have diversified to the extreme in plant lineages, likely as a result of the multiple ecological roles they fulfill (1). The high degree of plasticity of secondary metabolism pathways is consistent with the existence of large families of metabolism-related genes such as cytochrome P450s and UDP-glycosyltransferases in plant genomes that can create structural and chemical modifications almost without limits. The majority of metabolic gene functions remain unknown, however, either because the metabolites that they produce are unknown or significant associations remain to be identified between the expression of specific metabolic genes and characterized metabolic groups.

The biosynthesis of particular secondary metabolites or of complete metabolic groups is frequently taxonomically restricted (2). For this reason, certain secondary metabolite classes have been used as signature characters for biochemical investigation of specific plant families: for instance, quinolizidine alkaloids for Fabaceae (3), tropane and steroidal alkaloids for Solanaceae (4), and iridoids for Lamiaceae (5). Another way of looking at plant secondary metabolism diversity is to consider the precise tissue-specific localization of pathways responsible for their production. Compositional differences are, for instance, readily apparent across floral tissues that produce metabolic blends very different from their vegetative counterparts (6). In the most extreme cases, the accumulation of secondary metabolites can be restricted to specific cell types. For instance, plant defense metabolites are frequently produced in specialized tissues/cell types as a means of minimizing autotoxicity reactions in the surrounding tissues and/or of maximizing the defensive function of these metabolites toward aggressors that attack in a spatially specific manner (7, 8). A better exploration of tissue-level metabolic specialization is therefore particularly helpful in understanding the contribution of a given tissue to an organism's fitness. Deep biological insight based on single-cell metabolomics

Significance

Population geneticists have educated molecular biologists in how to harness the statistical power of variance arising from interindividual natural variation to elucidate gene function in plants. The metabolic differences among tissues within a plant provide another source of variance that can be harnessed in the quest to understand gene function. We combine the power of information theory statistics and computational metabolomics to parse metabolic diversity within an ecological model plant, *Nicotiana attenuata*, to reveal intriguing patterns of metabolic specialization in floral limb and anthers, the responsible mechanisms of which we parse further by detecting and silencing the expression of two UDP-glycosyltransferases involved in floral flavonoid metabolism. The workflow defines a framework for future evolutionary studies on plant tissue metabolic specialization.

Author contributions: D.L. and E.G. designed research; D.L., S.H., and E.G. performed research; D.L., S.H., I.T.B., and E.G. contributed new reagents/analytic tools; D.L. and E.G. analyzed data; and D.L., I.T.B., and E.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The MS/MS dataset has been deposited in the European Molecular Biology Laboratory European Bioinformatics Institute open metabolomics database MetaboLights, www.ebi.ac.uk (accession no. MTBLS335). The RNA sequencing dataset is available at the *Nicotiana attenuate* Data Hub database (nadh.ice.mpg.de/NaDH) and at the National Center for Biotechnology Information Sequence Read Archive (SRA) database, https://www.ncbi.nlm.nih.gov/sra (accession nos. NA1498ROT, NA1500LET, NA1717LEC, NA1504STT, NA1505COE, NA1515COL, NA1506STI, NA1507POL, NA1508SNP, NA1509STO, NA1510STS, NA1511NEC, NA1512ANT, NA1513OVA, NA1514PED, NA1516OFL, NA1517FLB, NA1501SES, NA1502SEW, and NA1503SED).

[1]To whom correspondence should be addressed. Email: emmanuel.gaquerel@cos.uni-heidelberg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1610218113/-/DCSupplemental.

has remained technically challenging; however, important steps have been taken toward this goal in the field of microbial metabolomics, but not yet in plant science, and the technique has been proven as an excellent indicator of phenotypic heterogeneity in this field.

From a mechanistic standpoint, the accumulation of secondary metabolites in a given tissue requires the spatial-temporal coordination of a vast array of cellular processes in which systems controlling biosynthesis, storage, and degradation are of central importance. Regulatory mechanisms coordinating these processes are only beginning to be uncovered for some model metabolic pathways such as the metabolic pathways of the family of glucosinolates in Brassicaceae (9). Coexpression analysis using information about gene and secondary metabolite cross-tissue expression patterns has been applied successfully to infer biosynthetic genes in secondary metabolism (4, 10, 11). In *Arabidopsis*, several "-omics"-based tissue atlases (e.g., for gene expression, alternative splicing, proteome) are publicly accessible to conduct such types of analysis (12).

Tissue-level nontargeted metabolomics of downstream metabolic readouts are more challenging to implement. Notably, the potential of mass spectrometry (MS)-based metabolomics and of the large-scale acquisition of tandem MS (MS/MS) spectra for as many metabolites as possible within a metabolic profile is severely constrained by the absence of straightforward classification and visualization pipelines that enable facile pathway interpretations. Metabolite annotation and identification are the obvious bottlenecks that thwart the metabolomics analysis of secondary metabolism (13, 14). Ideally, we need approaches that combine the strengths of state-of-the-art statistical methods currently emerging from the genomics field with the recent advances in metabolomics data mining, such as the method of MS/MS molecular networking, which allow unknown metabolites to be readily classified based solely on their fragmentation patterns.

Here, we developed a pipeline combining tissue-wide nontargeted MS data acquisition and information theory to mine patterns of tissue-specific structural diversity. With this pipeline, we analyzed a compendium of 14 dissected tissues of *Nicotiana attenuata*, an ecological model for chemically mediated adaptive traits in the wild. The analysis resulted in the deconvolution of 895 nonredundant MS/MS spectra, of which 565 exhibited preferential tissue specificity. Using information theory analysis, we asked whether certain tissues exhibited a higher degree of tissue metabolic specialization and which MS/MS data were linked to these patterns. From all this information, tissue–metabolite association maps were created to provide predictions about the tissue-level analysis of gene functions, some of which were tested by gene silencing techniques.

Results

A Compendium of MS Profiles Obtained from Isolated *N. attenuata* Tissues. Here, we isolated 14 tissues from 28- and 50-d-old *N. attenuata* plants growing under controlled growth conditions in the glasshouse (Fig. 1*A*). Pools of 100-mg tissues were extracted using independent extractions with 80% or 20% (vol/vol) methanol to increase the coverage of the metabolome with polar to semipolar compounds not efficiently extracted by 80% (vol/vol) methanol. We used an optimized ultrahigh-performance liquid chromatography (UHPLC) electrospray ionization (ESI)/quadrupole time-of-flight (qTOF) MS method to analyze the metabolome profiles of these tissues. Identical chromatographic conditions were used for the analysis of these two extraction types because retention time consistency for identical mass features (with mass features being *m/z* signals detected at a given retention by the peak picking method) is one of the criteria implemented in our bioinformatics workflow.

The dataset (Dataset S1) was processed using the R package XCMS utilizing optimized parameters and analyzed by principal component analysis (PCA), which confirmed that extensive

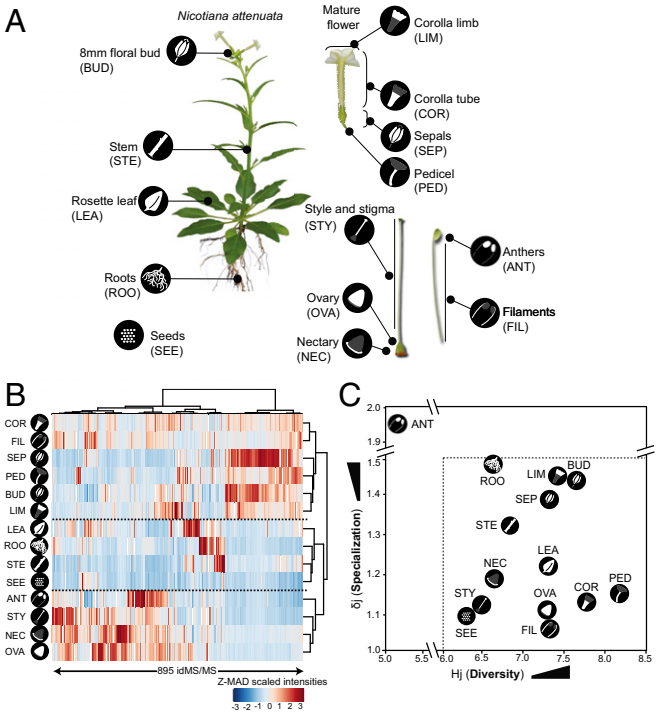

Fig. 1. Integration of MS-based metabolomics and information theory analysis highlights tissue-specific metabolome specialization. (*A*) Tissues were collected and analyzed separately for metabolomic profiling. Detailed explanations of the tissue collection procedure are provided in *Materials and Methods*. (*B*) Hierarchical clustering, using the Euclidean distance as the clustering metric, of tissue-specific idMS/MS relative expression profiles. The heat map coloring depicts the scaled intensities. Z-score–normalized median absolute distances captured the cross-tissue variations for idMS/MS intensities (895 idMS/MSs) obtained for each tissue. (*C*) Information theory analysis of tissue-level idMS/MS composition δj and Hj based on idMS/MS cross-tissue distributions is displayed in a 2D space to reveal gradients of metabolic specialization. ANT, anthers; BUD, floral bud; COR, corolla tube; FIL, filaments; LEA, rosette leaves; LIM, corolla limb; PED, floral pedicel; ROO, root; SEE, seeds; SEP, floral sepals; STE, stem; STY, floral style and stigma.

variations in the composition of mass features exist among the different tissue profiles (*SI Appendix*, Fig. S1). The XCMS × PCA processing procedure is a very common one and is, together with a priori knowledge annotation of prominent mass features and of in-source fragmentation patterns, frequently considered as the central mining step in most metabolomics studies. However, patterns revealed from this type of data mining provide little to no information with respect to compound diversity among samples. This type of biological interpretation critically requires an analysis at the level of metabolites described by deconvoluted spectra, and not at the level of individual mass features, which is what prior work has used.

Creating a Multitissue Indiscriminant MS/MS Library for Metabolite Structural Analysis. To collect a holistic repertoire of structural information on the metabolic diversity in our tissue compendium, we implemented a tissue-wide analytical pipeline for indiscriminant (data-independent) MS/MS (idMS/MS) analysis. Compared with data-dependent acquisition methods involving the preselection of a restricted list of precursor ions for collision-induced dissociation (CID) fragmentation, this approach considers for fragmentation analysis all signals within an *m/z* range set as large as possible (15). In recent years, the idMS/MS technique, sometimes referred to as shotgun or broad-scale MS/MS, has gained considerable interest as an exploratory method for metabolomics measurements. In a previous study, we showed that idMS/MS can

be efficiently implemented to most qTOF instruments by running replicated measurements of the same sample using idMS/MS at different CID voltages to maximize fragment coverage (16). Furthermore, because the idMS/MS method has the disadvantage of being uninformative about precursor-to-fragment relationships, we optimized a computational pipeline based on cross-sample correlation calculations to perform fragment relationship assignments with high confidence (16).

Here, we improved the previous computational pipeline for exploiting cross-tissue metabolic variations to gain statistical power in precursor-to-fragment assignments. Briefly, for each CID voltage, precursor and fragment relationships were assigned using Pearson correlation coefficient (PCC) analysis across all tissues (*Materials and Methods*). The idMS/MS spectra reconstructed at each CID voltage were merged into a composite idMS/MS spectrum, and some of redundant sub-idMS/MSs were grouped simultaneously via the calculation of spectral similarity. Notably, not all putatively redundant sub-idMS/MSs could be merged into respective compound-specific idMS/MS spectra using the single spectral similarity threshold value applied to the dataset; hence, metabolites prone to particularly intense in-source fragmentation frequently produced several idM/MS spectra by the analysis. This possible challenge was likely minimal in our study, however, because these different sub-idMS/MSs are expected to covary across the tissue dataset and to form tight clusters during the structural clustering analysis applied later on in the workflow (*SI Appendix*, Fig. S2). The discrimination of nearly coeluting isobaric peaks, resulting from compounds with the same molecular weight but different structures, is a challenge inherent to all large-scale metabolomics studies and one that can only be partly resolved via technical advances such as enhanced ion mobility MS. From a data processing standpoint, if two nearly coeluting isobaric species return overlapping fragmentation patterns, the correlation score for the precursor-fragment assignment will consequently be affected. Such a scenario could explain challenges encountered during the assembly of certain spectra (which were therefore not included in subsequent analyses). However, an advantage of the precursor-fragment assignment method in our pipeline is that it does not rely solely on the chromatography behavior of candidate *m/z* signals but also on their coregulated behavior across the tissue dataset, which, to a certain extent, improves the assembly of nearly coeluting metabolites. The deconvolution efficiency was tested by comparing idMS/MS spectra with previously reported MS/MS spectra at optimized CID voltages for major *N. attenuata* secondary metabolites as well as unknowns (Dataset S1). Altogether, the computational pipeline (merging of CID voltage-specific data and partial redundancy filtering) retrieved a library of 895 nonredundant idMS/MSs (Dataset S1); these idMS/MSs were used as the data for all subsequent analyses presented here.

Tissues Differ in Their Degree of Metabolic Specialization. For a first perspective into tissue metabolic relationships, we normalized idMS/MS spectra intensities (precursor intensities in MS mode) using a modified Z-score method, termed ZMAD (Z-score normalized median absolute distance) (*Materials and Methods*) and used hierarchical clustering analysis (HCA) (Fig. 1*B*). When merging the datasets obtained from the two extraction procedures, three main clusters appeared from the HCA based on Euclidean distance calculations (Fig. 1*B*): one cluster with most nonreproductive tissues of flowers (corolla limb, corolla tube, sepal, pedicel, bud, and filament), one with the reproductive parts (anther, nectary, ovary, style, and stigma), and a last one with vegetative tissues (leaf, root, stem, and seed). Interestingly, tissues that connect reproductive and nonreproductive parts in flowers, namely, filaments and stamens, exhibited strongly divergent idMS/MS compositional profiles, demonstrating that relatively fine-scale spatial modulations of metabolism can be analyzed by this approach. It should be noted that the upstream computational procedure used to deconvolute idMS/MS spectra was performed tissue-wide (and not at the individual tissue level) and relied on matrix alignment and noise filtering steps to produce an idMS/MS tissue-wide matrix that was of a consistent size (Dataset S1). A drawback of this computational approach is that no information about the number of idMS/MSs per tissue is readily available to explore tissue-level metabolic specialization. Intuitively, the presence of few high-intensity idMS/MSs in a given tissue compared with the average calculated across all tissues could be indicative of a high degree of metabolic specialization, whereas the presence of a large number of average-intensity idMS/MSs could reflect a low metabolic specialization. Such interpretations are linked to the frequency distribution of each idMS/MS within the dataset. Information theory, which was pioneered by Shannon (17) in a seminal article in 1948, provides the statistical framework to cope with this type of analysis. In defining tissue metabolic diversity and specialization, we therefore considered idMS/MS spectra as symbols, in the sense of information theory, and estimated for each tissue's metabolome its diversity based on the Shannon entropy of its frequency distribution. In other words, tissue-level metabolome specialization was measured as the average specificity of each of its idMS/MS components. Using previously implemented formulae (18), we retrieved values for the following indexes: diversity (Hj) reflecting the tissue-level idMS/MS diversity and specialization (δj) for the tissue-level idMS/MS specialization as inferred from the average idMS/MS specificity in the dataset.

Visualizing tissue metabolic profiles in a 2D space using these two indexes as coordinates revealed a number of interesting patterns (Fig. 1*C*). A most obvious one was that tissues significantly vary in their degree of δj and Hj. When extraction types were merged to achieve a more comprehensive view, anthers emerged as the tissue with the least diverse, most specialized metabolome (Hj = 5.16, δj = 1.95). In other words, several idMS/MSs exhibited relative high-intensity levels concomitant with low-frequency distributions across tissues. Root (Hj = 6.66, δj = 1.49), stem (Hj = 6.91, δj = 1.32), and sepal (Hj = 7.39, δj = 1.38) samples followed anthers in terms of low idMS/MS diversity and middle to high idMS/MS specialization. In the case of roots, the relatively high specialization index value retrieved for this tissue was especially supported by idMS/MS spectra collected from the 20% (vol/vol) methanol extraction (*SI Appendix*, Fig. S3). The signature for low diversity and low specialization detected in seeds (Hj = 6.34, δj = 1.09) was in line with the low density of chromatographic peaks seen for this tissue. Style and stigma (Hj = 6.47, δj = 1.11), filaments (Hj = 7.32, δj = 1.08), ovary (Hj = 7.31, δj = 1.10), corolla tube (Hj = 7.81, δj = 1.12), and pedicel (Hj = 8.17, δj = 1.16) were the tissues exhibiting lowest specialization indexes. The pedicel had the most diverse idMS/MS profile of all tissues analyzed.

Tissue-Level Differentiations in 17-Hydroxygeranyllinalool Diterpene Glycosides and Phenolamines. As a first step toward mining metabolite compositional variations across tissues, we amortized previous chemical knowledge acquired from *N. attenuata* leaves and evaluated whether the distribution across tissues was differentially modulated at different levels of know secondary metabolic pathways. For this analysis, we selected as a case study the 17-hydroxygeranyllinalool diterpene glycosides (17-HGL-DTGs) pathway that produces abundant acyclic diterpenes with antiherbivore functions (19). For this metabolic group, we retrieved the corresponding idMS/MSs for the most abundant metabolites and investigated cross-tissue modulations as visualized by plotting individual tissue ZMAD-normalized values (*SI Appendix*, Fig. S4). The 17-HGL-DTGs can be subcategorized based on their sugar/malonyl decorations as follows: the precursor molecule (lyciumoside I), core structures with a higher degree of glycosylation but no malonyl groups (nicotianoside III, lyciumoside IV, and attenoside), and monomalonylated (nicotianosides IV, Ia, and VI) and dimalonylated (nicotianosides V, II, and VII) structures. Lyciumoside I and

lyciumoside IV, its direct rhamnosylation product, were detected in young and photosynthetically active rosette leaves and at lower normalized levels in certain floral organs. The analysis revealed that 17-HGL-DTGs varied significantly among tissues, and the variance was organized by biosynthetic sequence in the pathway. The general trend was that greater tissue-specific variation was found in the downstream steps of the pathway. This trend was particularly apparent for monomalonylated 17-HGL-DTGs and suggests an increased translocation from source to sink tissues that increased with 17-HGL-DTG glycosylation and malonylation. Dimalonylated 17-HGL-DTGs were more abundant in certain reproductive organs relative to all other tissues. Another pathway monitoring example is provided for the phenolamide pathway, for which an apparent greater specificity toward certain reproductive organs was detected for polyacylated spermidine conjugates (*SI Appendix*, Fig. S4).

Large-Scale Inference of IdMS/MS Tissue Specificity Reveals Basic Principles of Tissue Interdependencies. The above descriptions confirmed that tissue-based differentiations are detectable for characterized metabolic pathways. This finding is consistent with the idea that specific groups of metabolites specifically accumulate in one or several tissues, albeit being detectable at lower levels in almost all other tissues. The specificity index of information theory of a given idMS/MS serving for δj calculation tends to be stringent and excludes features exhibiting a significant degree of specificity (association) with more than one tissue (Fig. 2*A*, *Center*). In an attempt to assess the degree of association of an idMS/MS toward one or several tissues statistically, we analyzed idMS/MS expression distribution across tissues using reduction of kurtosis as developed by Li et al. (20). The kurtosis analysis measures expression distribution patterns rather than frequencies and skirts the restriction of the number of tissues with which a given idMS/MS can be associated. As such, the method has been found to be highly successful in detecting tissue specificity from large-scale data. Briefly, idMS/MS spectra that exhibit high tissue specificity are characterized by high kurtosis values with either right- or left-tailed leptokurtic distributions, whereas idMS/MS spectra that exhibit low tissue specificity have low kurtosis values with normal distributions (Fig. 2 *A* and *B*). A total of 595 of 895 idMS/MSs exhibited preferential tissue associations ($Q < 0.05$), with the rest of the idMS/MSs being considered as non–tissue-associated features. For ease of interpretation, *SI Appendix*, Fig. S5 reports the statistical support via mapping of kurtosis Q values and inferred tissue associations for previously discussed tissue-level differentiations in the 17-HGL-DTG and phenolamide pathways.

To retrieve tissue idMS/MS–specific associations, we defined a tissue relative expression threshold Z ($Z = 2$) through the calculation of a reduction of kurtosis according to the rationale proposed by Li et al. (20) (Fig. 2*C* and *SI Appendix*, Fig. S6). Seeds harbored again the smallest associated metabolome had 99 specifically expressed idMS/MSs, followed by stem (158 idMS/MSs) and root (278 idMS/MSs), whereas a general trend was that floral organs had the largest numbers of associated idMS/MSs. An interesting level of analysis was therefore to look at the percentage of tissue-specific idMS/MSs compared with nonspecific ones per tissue. For instance, a number, albeit small (278 idMS/MSs), of idMS/MSs specific to roots represented 72.3% of the total detectable root metabolome, indicating the relatively high metabolic specialization of this tissue.

In agreement with the importance of not restricting the analysis only to single tissue-specific idMS/MSs and considering different degrees of specificity based on the number of tissues in which a given idMS/MS accumulates, we detected that idMS/MSs specifically associated with more than one tissue were highly prevalent (97%) in the tissue-specific idMS/MS pool (*SI Appendix*, Fig. S7). The relative strength of the metabolic interdependencies between two tissues based on the number of shared tissue-specific idMS/MSs was scored using the Jaccard index (Fig. 2*D*). Clustering based on this score again supported the fact that vegetative tissues such as leaf, stem, and root cluster apart from floral counterparts in terms of secondary metabolite profiles. The "floral" cluster subdivided into three smaller clusters: one with tight connections between tissues not directly involved in reproductive tissues (besides the complete bud); one comprising tissues with mostly reproductive functions (filament, style, ovary, but also the nectary); and, finally, anthers. The individualized positioning of anthers in this clustering analysis is in agreement with the information theory specialization signature detected in this tissue as discussed above.

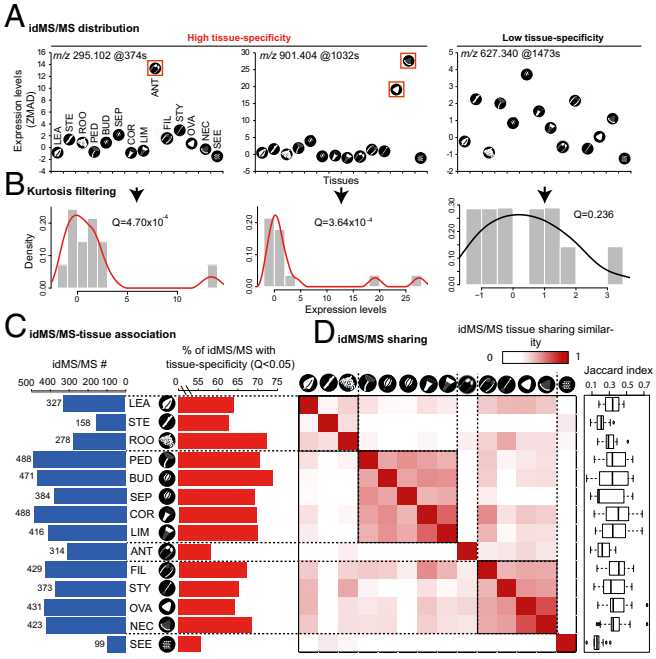

Fig. 2. Large-scale analysis of idMS/MS tissue specificity. (*A*) Cross-tissue distribution patterns for three idMS/MS examples. Z-score–normalized median absolute distances captured cross-tissue variations for idMS/MS intensities. idMS/MSs deconvoluted for *m/z* 295.102 @ 374 s and 901.404 @ 1,032 s revealed clear tissue specificity for one and two tissue types, respectively. idMS/MS for *m/z* 627.340 @ 1473 s was not associated with a particular tissue. (*B*) Density of intensity levels of each idMSMS across all analyzed tissues is computed and filtered using a reduction of kurtosis method to determine idMS/MS with significant tissue specificity. (*C*) Bar chart showing the number of idMS/MSs per tissue using an intensity threshold of 2 (*Left*), and bar chart showing the percentage of idMS/MSs illustrating tissue specificity per tissue (*Right*). (*D*) Heat map matrix visualizing idMSMS sharing among tissues as measured using the Jacquard index. The idMS/MS classifications to main compound classes in *N. attenuata* as obtained by idMS/MS alignments to public libraries and manual curation are shown in Dataset S1).

MS/MS Structural Analysis of IdMS/MS Associations. Examples of annotated tissue-specific idMS/MS spectra shared by different tissues are presented in Dataset S1. Metabolite annotation remains a bottleneck in metabolomics studies because public spectral databases are poorly populated with plant-specialized metabolites, with many of them being taxa-specific and frequently species-specific. The MS/MS molecular network method pioneered by Dorrestein and coworkers (21) circumvents the limitation of spectral databases via the analysis of within-dataset MS/MS similarities to accelerate hypothesis generation about the identity of unknown MS/MS (21). This approach also has the advantage of being amenable to the visualization of putative biochemical relationships among metabolites corresponding to highly similar MS/MS spectra (16). In a recent study, we improved the scoring and classification of MS/MS similarities for plant secondary metabolites notably by the implementation of a biclustering method that

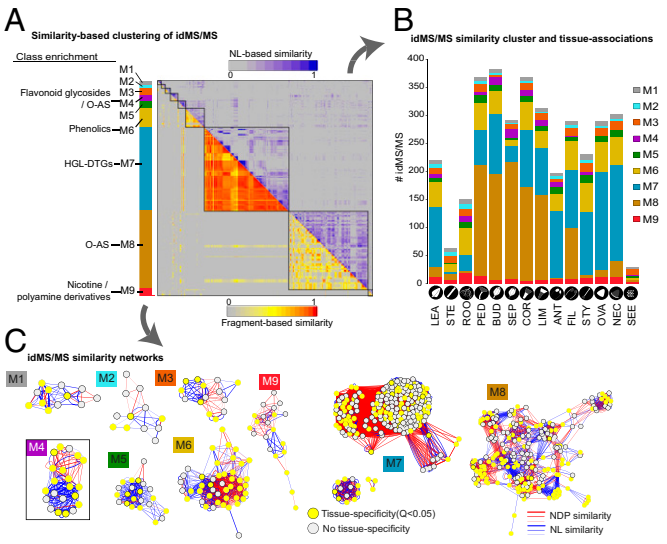

Fig. 3. Combination of structural classifications of idMS/MS and tissue specificity of expression. (*A*) Biclustering analysis to classify idMS/MSs according to structural similarities. The analysis used two scoring methods: one based on shared fragments among spectra, whereas the other scored shared common NLs among spectra. Using biclustering, which favors clustering based on iterative alignments of spectra based on the two scoring methods, produces large modules (M) with structurally related idMS/MSs. Some of these modules were congruent with known compound families, whereas others were composed of yet unknown or poorly characterized metabolites. Module annotation and idMS/MS intensity distribution are reported in Dataset S1. (*B*) Relative contribution of each module to the idMS/MSs associated with a given tissue. The visualization highlights the complete absence of specific metabolic groups, corresponding here to particular modules, such as *O*-acyl sugars (*O*-AS), in anthers. (*C*) Molecular networks constructed for each module. Nodes represent idMS/MSs and edges represent similarity values based on the two scoring types. Tissue specificity can easily be mapped to the molecular networks.

detects possible compound familial groupings according to fragment and neutral loss (NL)-based similarities (16). Applying this method to the total pool of idMS/MSs from the present study resulted in the formation of nine modules within which idMS/MSs are expected to share high structural similarities (Fig. 3*A*). Modules 4, 6, 7, 8, and 9 largely corresponded to previously identified compound classes: flavonoid glycosides, phenolics, 17-HGL-DTGs, acyl sugars, and nicotine and polyamine derivatives, respectively. As expected, uncharacterized metabolites likely belonging to these groups and not previously thoroughly investigated were also detected as sub-idMS/MSs that did not merge during the redundancy filtering step. A comprehensive view is provided in Dataset S1 that concatenates MS/MS spectral content and NLs, as well as their clustering and association with tissues. A critical consideration was whether tissues differ in their module relative composition as depicted in the stacked bar chart of Fig. 3*B*. For instance, it is clearly visible that complete *O*-acyl sugar metabolism (M8) is absent from anthers and the style, that the stem and seeds lack 17-HGL-DTG metabolism (M7), and that the flavonoid module (M4) is overrepresented in certain flower tissues. Molecular networks can be constructed for each module to visualize structural relationships among idMS/MSs better (Fig. 3*C*). The case of module M4 is presented (Fig. 4). A subpart of this flavonoid-enriched module contains *O*-acyl sugar type II due to shared NLs with flavonoid glycosides; those two groups are still discriminated according to the edge density and by the careful inspection of idMS/MS tissue coexpression scores. By simply mapping the relative expression of idMS/MSs onto nodes, it is possible to pinpoint metabolites that are characteristic of a given tissue rapidly, for instance, the dramatic overrepresentation of kaempferol-3-*O*-glucoside (KG) at the limb level (808.684 ZMAD scaled intensity). Also, the idMS/MS for *m/z* 295.102 specific to anthers and not coexpressed across tissues with any other flavonoid glycosides from module M4 is putatively annotated by our method as a glucose ester with C4 side chains, depicted here as 6-tuliposide B (22).

Exploring Metabolite and Gene Coassociations Across Tissues Facilitates Metabolic Gene Pathway Assignment. In this last section, we illustrate the power of first determining tissue–metabolite associations in generating predictions about the assignment of unknown genes to particular pathways. In the case of a unimodal regulation (with cross-tissue transport being minimal), the logic behind these predictions is that a gene responsible for the production of a given set of metabolites will share maximal tissue associations with these metabolites. As for gene expression data, we used an RNA-sequencing (RNAseq) transcriptome dataset (*SI Appendix*, Table S2) in which tissues and developmental stages largely overlap with those tissues and developmental stages used for metabolomics but that also included treatment responses to account for the fact that certain genes are expressed constitutively at low levels but the metabolites can accumulate without turnover to high levels. Similar to idMS/MS data, the kurtosis filtering allowed us to filter out genes with quasiconstant expression and focus on the genes exhibiting leptokurtic distributions (*SI Appendix*, Figs. S6 and S8). Thirty-seven percent of the total genes expressed exhibited leptokurtic distributions (i.e., expressed specifically in one or several tissues) (Fig. 5*A*). Subsequent analysis steps followed the steps presented above for the analysis of metabolites. Overrepresented gene ontologies (GOs) within the complete set of genes with preferential tissue associations corresponded to general processes such as chloroplast thylakoid activity, monocarboxylic acid biosynthetic process, anion transport, and metal ion transport (*SI Appendix*, Fig. S9). This GO overrepresentation analysis was also conducted on a module basis for a gene set specifically associated with a given idMS/MS module using an Ochiai similarity index (*SI Appendix*, Figs. S9 and S10). For this calculation, emphasis is placed on tissue specificity rather than on the characterization of a trend of coexpression across the complete tissue set such as is the case when using simple PCC analysis. Through this approach, it is now possible to target specific metabolic gene families, UDP-glycosyltransferases here, and to predict their importance for the metabolic group enriched within a given idMS/MS module (Fig. 5*B*). For mining this latter gene family, an additional filtering criterion is the presence within the coassociated idMS/MSs of NLs corresponding to glucose or rhamnose moieties. Modules 4, 7, and 8 are made up of idMS/MSs corresponding to glycosylated secondary metabolites, and hence enriched in the presence of these latter NLs (Fig. 5*B*). We extracted 10 members of this gene family that had cotissue specificities with members of M4. Next, we tested, by transient gene silencing using virus-induced gene silencing (*SI Appendix*, Fig. S11), the pathway assignment of two of these UDP-glycosyltransferases highlighted by the Ochiai similarity analysis: *UDP-glycosyltransferase-A* (*UGT-A*) [Ochiai similarity = 0.71 with quercetin-3-*O*-glucose (QG)] and *UDP-glycosyltransferase-B* (*UGT-B*) (Ochiai similarity = 0.71 with rutin). Briefly, when silencing *UGT-A*, a majority of the flavonoid glycosides in flower buds were significantly decreased in their accumulations, namely, KG, QG, kaempferol-3-*O*-glucose-rhamnose (KGR), and quercetin-3-*O*-glucose-glucose (QGG). On the other hand, silencing *UGT-B* translated into significant decreases in the levels of rhamnose-containing KGR and rutin, whereas QGG, QG, and KG accumulated to higher levels compared with the empty vector control. This result is consistent with the conclusion that *UGT-B* likely controls the rhamnosylation of these flavonoid glycosides and that the higher accumulations of nonrhamnose flavonoid glycosides reflect the metabolic tension existing with the *UGT-A*–mediated glucosylation process (Fig. 5*C* and *SI Appendix*, Fig. S11). Future work

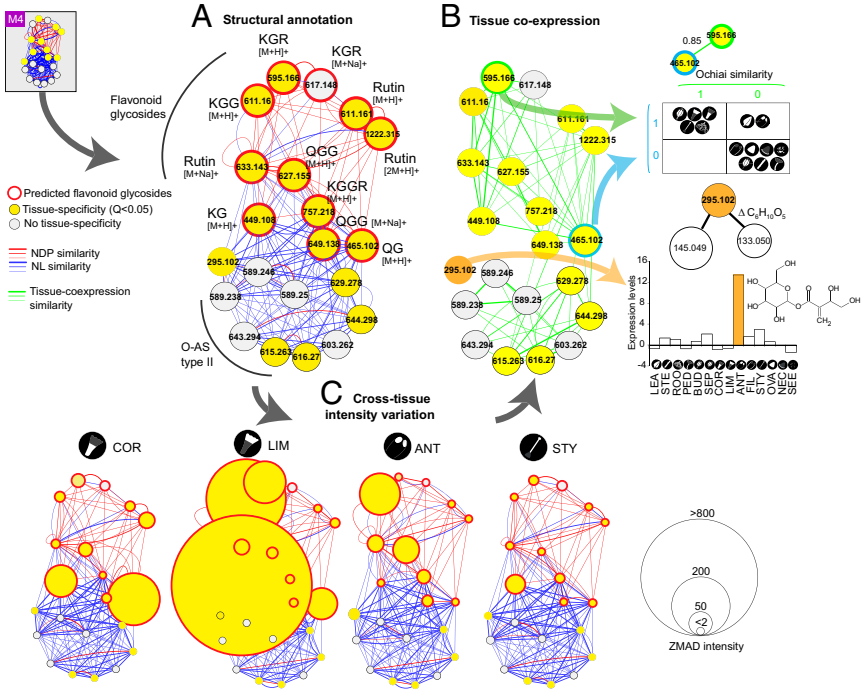

Fig. 4. Distribution of a flavonoid-enriched module among different flower parts. (*A*) Network representation and annotation of module M4 from the biclustering analysis. Nodes correspond to idMS/MS spectra, and edges correspond to their pairwise similarity as measured according to the fragment (NDP; >0.6) and NL (>0.6) similarity. Many of the spectra correspond to flavonoid glycosides, albeit *O*-acyl sugars of type II are also present due to shared NLs. (*B*) Cross-tissue coexpression (based on an Ochiai score > 0.6) between idMS/MS spectra discriminates flavonoid glycosides from *O*-acyl sugar. The analysis reveals metabolites within the M4 module with high tissue specificity, such as idMS/MS at *m/z* 295.102, predicted to be a tuliposide derivative, which is abundant in anthers (Fig. 3*A*). (*C*) Examples of visualization of cross-tissue variations for idMS/MSs of M4. Node size is proportional to the cross-tissue relative intensity of each idMS/MS. Color mapping denotes rules presented in *A*. Gray nodes do not exhibit tissue specificity, whereas yellow nodes were detected as tissue-specific. Red-circled nodes are annotated as flavonoid glycosides. Identifications of KG, kaempferol-3-*O*-sophoroside (glucosyl(1-2)glucoside) (KGG), KGR, QG, QGG, and kaempferol-3-*O*-rutinoside (glucosyl(1-2)rhamnoside) [QGR (Rutin)] are according to Snook et al. (53).

could test these predictions and examine the enzymatic properties of these two UDP-glycosyltransferases.

Discussion

In this study, we investigated tissue-level variations in secondary metabolism in an ecological model plant using computational metabolomics and information theory statistics. Information theory has been used for multivariate data generated in a broad scope of biological contexts ranging from plant ecology (23) to microbiome diversity (24), but, to our knowledge, it has never been used to summarize trends in MS-based metabolomics data. Previous studies identified preferential tissue-based redirectionalities in secondary metabolism, for instance, during the maturation of tomato fruits for which the green, turning, and red developmental stages are characterized by rearrangements in pathways related to flavonoids, phenolics, and glycoalkaloids (25). However, to our knowledge, no unbiased metabolomics study, other than a study of the AtMetExpress database (6), has been applied with rigorous statistical analysis to such a broad range of tissues as in the present study.

This statistical portfolio revealed that tissues exhibit distinct states of secondary metabolism activity but also that they differ in their degree of specialization. An extracted feature illustrative of the explorative power of this approach was that connecting tissues of flowers such as the anthers and filaments, on one hand, and the corolla limb and tube, on the other hand, differed dramatically in their metabolite specialization signatures. For the corolla, this contrast highlights the fact that limbs are functionally specialized for attracting and guiding pollinators, and likely require a highly specialized metabolome to fulfill this function. The latter is especially expected in a species such as *N. attenuata* whose main pollinator, the hawkmoth *Manduca sexta*, is also a voracious folivore during its larval stage (26), requiring that the plant critically fine-tune its blend of secondary metabolites to solve the dilemma imposed by these two contrasting interactions (27, 28). As expected, the green tissues of our dataset display the most prototypic and undifferentiated metabolic profiles, as highlighted by both targeted and nontargeted analyses.

N. attenuata is a pioneer plant in postfire habitats (29, 30), and as such, it represents one of the primary food sources for herbivorous insects (31). It is well established that the photosynthetically active tissues of this plant mount a very strong specialized metabolic response locally and systemically during biotic challenges such as insect herbivory (32). It would therefore be very interesting to reassess how the specialization indices readjust during stress adaptation, taking advantage of preexisting knowledge on antiherbivory function of many secondary metabolite classes (6, 19, 33). Also, the pools of many of these defensive secondary metabolites are rearranged during ontogeny in the form of quantitative gradients established across tissues (19, 34). The optimal defense theory provides a conceptual framework that links these quantitative patterns with the fitness of different tissues for the plant's fitness (35, 36). Even though the developmental stages of multiple tissues would need to be separately analyzed using our analytical approach to evaluate this theory thoroughly, several defense-related metabolites exhibited higher relative levels in reproductive tissues than in vegetative counterparts, a central prediction of the optimal defense theory. A last remark concerns the extremely low metabolic diversity detected in seeds, a result that could possibly be due to the fact that most apolar metabolites present in the seed endosperm were poorly recovered with our extraction systems. This result speaks to the need to use a more sophisticated combination of extraction and chromatographic

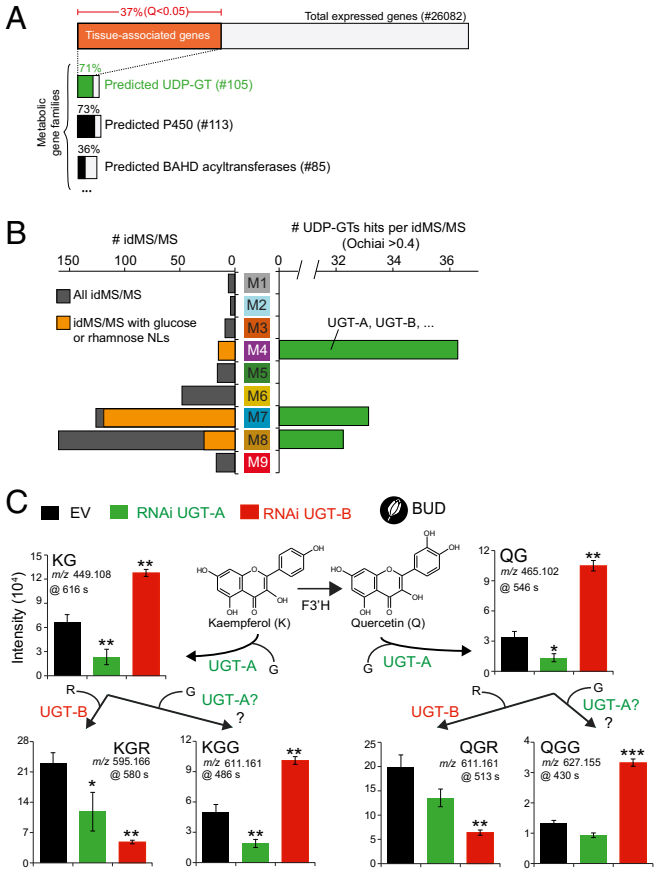

Fig. 5. Silencing *UGT-A* and *UGT-B* reveals their involvement as UDP-glucosyltransferase and UDP-rhamnosyltransferase, respectively, in floral flavonoid glycoside metabolism, two predictions of the tissue coexpression analysis. (*A*) Results of the kurtosis filtering analysis for preferential tissue–gene associations. Examples are provided for the tissue specificity of members of large metabolic gene families. Notably, 71% of all predicted UDP-glycosyltransferases (GT) exhibit tissue specificity in the transcriptome dataset. (*B*, *Left*) Number of tissue-specific idM/MS spectra containing glucose or rhamnose NLs, and therefore predicted to be glycosylated secondary metabolites, compared with the total number of tissue-specific idMS/MSs per biclustering module. M4 is enriched in flavonoid glycosides, M7 in 17-HGL-DTGs, and M8 in *O*-acyl sugars. (*B*, *Right*) Number of UDP-GT coassociated across tissues (Ochiai score > 0.4) with at least one idMS/MS containing glucose (G) or rhamnose (R) NL of each module. (*C*) Relative levels of precursors corresponding to idMS/MSs referred to in *B* after analysis of flower buds of plants inoculated with empty vector and gene silencing constructs for *UGT-A* and *UGT-B* (*SI Appendix*, Fig. S11). As supported by the annotation of idMS/MS spectra, silencing *UGT-A* decreases the glucosylation of flavonols, whereas silencing *UGT-B* decreases their additional rhamnosylation. Identifications of KG, KGG, KGR, QG, QGG, and QGR (Rutin) are according to Snook et al. (53). $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

systems in future experiments to capture the behavior of a broader range of compound classes.

It is tempting to consider that signatures of high metabolic specializations observed for certain tissues correlate with their highly specialized physiological functions. Previous tissue-level -omics analyses in plants and animals are consistent with the expectation that physiological differentiation is accompanied by qualitative variations of metabolic capacities (37, 38). As previously noted, this claim is difficult to support only with metabolomics due to the sparse knowledge about secondary metabolite biosynthetic schemes. The GO enrichment analysis conducted in this study supports the fact that the kurtosis-based method is able to discriminate gene signatures involved in the tissue-specialized physiological processes from housekeeping ones (e.g., pollen tube growth) (*SI Appendix*, Figs. S9 and S10), so it is reasonable to propose that metabolites extracted by the kurtosis method also reflect tissue-level functions, even if transport processes may obfuscate some of these trends. In this regard, the case of anthers exhibiting a prevalent signature of high idMS/MS specialization, greater than all other reproductive organs, is particularly germane. In line with our observation of the manufacture of a specific set of metabolites in this tissue (Dataset S1), previous studies have shown that metabolites, notably certain phenolic derivatives, are abundant and highly specific for the tapetum (specialized layer of nutritive cells and source of precursors for the pollen coat within anthers) of anthers and pollen grains (39–41). The biosynthesis of these phenolic derivatives, with potential roles in pollen coat composition and establishment of fertilization barriers, has been linked to rapid metabolic gene evolution through retroposition and neofunctionalization (42). In a cross-species study on transcriptome evolutionary divergence, the fastest rates of gene expression divergence and signatures of transcriptome specialization were detected in anthers, whereas the lowest rates of evolution were detected in roots (43). Our metabolomics study therefore suggests that transcriptome and metabolome specialization may be coupled patterns in anthers, likely as a result of strong reproduction-related selection pressures exerted at this tissue level. More broadly, it would therefore be very interesting to analyze whether such kinds of metabolic specialization patterns are consistent across species for homologous tissues.

Navigating large datasets in such a way that knowledge can be made more efficiently accessible for hypothesis formulation is one of the challenges that thwart the routine application of certain -omics technologies to nonmodel systems organisms. In this study, we used data-independent MS/MS acquisition. This approach, albeit suffering from redundant data collection for certain metabolites prone to intense in-source fragmentation, maximizes the comprehensiveness of fragment data collection, which forms the foundation of such unbiased analysis. The present study also speaks to the power of the previously described molecular network method for plant samples and identifies directions for its integration with genomics data. The latter is illustrated by our functional studies on UDP-glycosyltransferases and the assignment of two previously uncharacterized genes of *N. attenuata* to the glucosylation and rhamnosylation steps in floral flavonoid glycoside metabolism (44, 45) (Fig. 5). Importantly, the data platform generated here can also be mined for additional metabolic gene families (e.g., P450, BAHD acyltransferases).

Tissue-level PCC coexpression analysis among genes and metabolites has traditionally been shown to be an efficient way forward for gene function analysis in secondary metabolism (6, 44, 46, 47). However, one important message of the present study is that because PCC-based coexpression analysis relies on trends inferred from gene/metabolite expression levels, certain tissue-level gene–metabolite associations are difficult to capture via this approach because they take place only in a few of the analyzed tissues, thereby resulting in a poor coexpression output. Consistent with this finding, a recent study on gene-sharing analysis in plants and animals demonstrated that an approach that puts emphasis on gene expression tissue specificities is significantly more efficient in identifying functional gene clusters than one that relies on the complete tissue-level expression dataset (20). Our kurtosis analyses show that this inference is likely to be more pronounced when incorporating metabolomics as another -omics dimension, because up to 97% of detected secondary metabolites exhibit tissue-specific expression in only a few of the tissue atlases. As such, we concluded that relying on expression levels monitored across the overall tissue set would decrease rather than increase the statistical power to discover biologically meaningful gene–metabolite associations. We thus adopted for gene-to-metabolite analysis a modified Ochiai similarity analysis in which the emphasis is placed on tissue specificity. Comparison of performance between this

Ochiai similarity analysis and the PCC-based coexpression analysis revealed that the PCC analysis returned poor coexpression values (PCC for the association QG/*UGT-A* is only 0.09, whereas the Ochiai similarity for the same association is 0.71) and failed to associate *UGT-A* and *UGT-B* specifically to flavonoids (*SI Appendix*, Table S3). Similar comparisons of performance between these two approaches for a compendium of 70 previously characterized gene–metabolite associations also confirmed that the Ochiai similarity analysis systemically outperforms the PCC-based approach, especially when metabolites exhibited high tissue specificity (*SI Appendix*, Table S3). Taken together, this study reinforces the power of applying approaches combining large-scale metabolomics and information theory analysis to accelerate hypothesis generation on metabolic gene function.

Conclusion

In summary, a major strength of this unique study is that it synergistically combines, using a three-pronged approach, the strengths of (*i*) information theory to capture signatures of diversity and specialization in the dataset, (*ii*) computational MS to accelerate the structural annotation of the diversity of compounds collected, and (*iii*) experimental gene silencing to falsify hypotheses regarding metabolic gene functions from metabolomics–transcriptomics integration. A recent breakthrough study on mammals' metabolomes has highlighted the power of metabolomics to predict markers associated with organ specialization in a phylogenetic context (48). Future directions will make use of genomics resources existing for related species of *N. attenuata* to extend the approach to the diagnosis of gene divergence effects contributing the most to tissue-metabolic specialization.

Materials and Methods

Tissue-Level Metabolite Extraction. Here, we extracted 14 different tissues from 28- and 50-d-old *N. attenuata* plants growing in the glasshouse (Fig. 1*A*). For nonreproductive tissues, the sample collection included a pool of all nonsenescing rosette leaves; combined lower, middle, and higher segments of the stem; the complete root system; and matured seeds. Reproductive parts were harvested as follows. Complete floral buds of 8-mm length, a stage at which the corolla has not yet protruded from the sepals and for which important gene expression and metabolic reconfigurations have been detected in previous work (49), were harvested. Mature flowers at anthesis (5 d after 8-mm stage, 7:00 PM), were carefully separated into the following parts: pedicel, complete sepal ring, nectary, ovary (not including the nectary), style, anthers, filaments (not including anthers), corolla tube (not including the limb), and corolla limb.

Pools of 100 mg of isolated tissues (*SI Appendix, Materials and Methods*) were extracted as follows using extraction buffers containing either 20% or 80% (vol/vol) methanol to increase the coverage of chemically diverse metabolite classes. One milliliter of extraction buffer [50 mM acetate buffer (pH 4.8) containing 20% or 80% (vol/vol) methanol] per 100 mg of tissue was added, and samples were homogenized in a ball mill (Genogrinder 2000; SPEX CertiPrep) for 45 s at a rate of 1× and at 250 strokes per minute. Homogenized samples were centrifuged at 16,000 × *g* at 4 °C for 30 min, and supernatants were transferred into 1.5-mL microcentrifuge tubes and recentrifuged as before. Supernatants of 400 μL were transferred to 2-mL glass vials for MS-based metabolomics. To prevent the discarding of tissue-specific metabolites from the XCMS analysis due to poor grouping across samples (*SI Appendix, Materials and Methods*), five mixed extracts containing all 14 tissues at different ratios were generated and processed simultaneously with all other tissue samples.

UHPLC-ESI/qTOF-MS Conditions for IdMS/MS Data Acquisition. Data-independent or idMS/MS fragmentation analysis was conducted to gain structural information on the overall detectable metabolic profile. Injection and UHPLC binary gradient-based separation conditions used for the MS and MS/MS mode analyses are described in *SI Appendix, Materials and Methods*. For all MS analyses, the column eluent was infused into a MicrOTOF-Q II (Bruker Daltonics) equipped with quadrupole and TOF analyzers and fitted with an electrospray source operated in positive ionization mode (capillary voltage = 4,500 V, capillary exit = 130 V, dry temperature = 180 °C, dry gas flow = 8 L·min^{-1}). The concept of the idMS/MS approach relies on the fact that the quadrupole is operated with a very large mass isolation window (so that quasi all *m/z* signals are considered for fragmentation). For this determination, several independent analyses are performed with increasing CID collision energy (CE) values because the MicrOTOF-Q II instrument can operate neither alternated scans collected in MS and MS/MS mode nor CE ramping. Briefly, samples were first analyzed by UHPLC-ESI/qTOF-MS using the single-MS mode (low-fragmentation condition derived from in-source fragmentation) by scanning from *m/z* 50–1,400 at a rate of 5,000 scans per second. MS/MS analyses were conducted using nitrogen as collision gas and involved independent measurements at the following four different CID voltages: 20, 30, 40, and 50 eV. The quadrupole was operated throughout the measurement with the largest mass isolation window, from *m/z* 50–1,400. This mass range is automatically activated by the operating software of the instrument when the precursor *m/z* and the isolation width are set to 400 and 300 Da, respectively. Mass fragments were scanned in the single-MS mode between *m/z* 50 and 1,400 at a rate of 5,000 scans per second. Mass calibration was performed using sodium formate (50 mL of isopropanol, 200 μL of formic acid, 1 mL of 1 M NaOH in water). Data files were calibrated postrun on the average spectrum from this time segment, using the Bruker high-precision calibration algorithm. The idMS/MS dataset has been deposited in the open metabolomics database Metabolights (www.ebi.ac.uk) under accession no. MTBLS335.

Assembly of Compound-Specific IdMS/MS. We used a previously designed precursor-to-product assignment pipeline (15) using the output results from processing with the R packages XCMS and CAMERA. The idMS/MS assembly was achieved via correlational analysis between MS1 and idMS/MS mass signals for low- and high-CEs and newly implemented rules (*SI Appendix, Materials and Methods*). The correlation analysis for precursor-to-product assignment was implemented using an R script, and rules were operated using a C# script available at GitHub (https://github.com/PlantDefenseMetabolism).

Defining Tissue Metabolic Diversity and Specialization Using Information Theory. Tissue metabolic diversity, the Hj index, was calculated using Shannon entropy of idMS/MS tissue-level frequency distribution. Tissue metabolic specialization, the δj index, was measured by the average idMS/MS specificity of each of the tissue idMS/MS components. Framework details are described in *SI Appendix, Materials and Methods*.

IdMS/MS Similarity Scoring. The idMS/MS spectra were aligned in a pairwise manner, and their similarity was calculated according to two scores. First, a standard normalized dot product (*NDP*), also referred to as cosine correlation method, was used to score fragment similarity among spectra using the following equation:

$$NDP = \frac{\left(\sum_i^{S1\&S2} W_{S1,i} W_{S2,i}\right)^2}{\sum_i W_{S1,i}^2 \sum_i W_{S2,i}^2},$$

where *S*1 and *S*2 correspond, respectively, to spectrum 1 and spectrum 2 and $W_{S1,i}$ and $W_{S2,i}$ indicate peak intensity-based weights given to *i*th common peaks differing by less than 0.01 Da between the two spectra. Weights were calculated as follows:

$$W = [Peak\ intensity]^m [Mass]^n,$$

with $m = 0.5$ and $n = 2$ as suggested by MassBank (50).

A second scoring method involving the analysis of shared NLs among individual idMS/MSs was implemented as described in *SI Appendix, Materials and Methods*. For this analysis, we used a list of 52 NLs commonly encountered during MS/MS fragmentation (Dataset S1) as well as more specific ones that had been previously annotated for MS/MS spectra of *N. attenuata* secondary metabolite classes.

IdMS/MS Tissue-Specificity Inference Using Kurtosis Filtering. We used an outlier-insensitive Z-score measure, generally considered preferable for the statistical description of sample groups containing extreme differences in values, by using median and median absolute deviation (*MAD*) instead of mean and SD for the normalization of both idMS/MS and RNAseq datasets to obtain relative expressions within tissues, as calculated using the following equation described by Birmingham et al. (51):

$$Z_i = (E_i - Median(E))/MAD(E),$$

where E_i is the expression level of a metabolite or a gene in tissue *i*. *E* is a vector of a metabolite or a gene in all tissue samples.

Kurtosis (K) was calculated for each metabolite and gene using an R package (moments) utilizing the following equation:

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2},$$

where X_i stands for the expression level of a metabolite or a gene in the ith tissue and $\bar{X}$ is the mean of the same metabolite or gene. The P value of the kurtosis was calculated using Anscombe.test function in the R "moments" package.

Tissue specificity for a metabolite or a gene was defined using the reduction of kurtosis method as previously described (20). When a leptokurtic expressed metabolite or gene removes high expression values for certain tissues, the kurtosis of the metabolite or the gene will be reduced. Threshold Z filtering of the data from a particular tissue was obtained by plotting the cumulative reductions in the kurtosis curves for any given kurtosis threshold using different Z threshold values (*SI Appendix*, Fig. S6). When defining the false discovery rate-adjusted P value as Q, we chose a Z threshold of 2 for metabolite datasets, where 98.3% (the highest) of the metabolites with $Q < 0.01$ exhibit reduced kurtosis after applying the threshold cutoff. Similarly, a threshold of 3 was applied for the RNAseq dataset.

RNAseq Dataset of Different Tissues and Data Mining. A detailed overview of the RNAseq dataset and National Center for Biotechnology Information database accession numbers are available in *SI Appendix*, Table S2. A list of the tissues collected for RNAseq analysis and of metadata related to this experiment is also available at the *Nicotiana attenuata* Data Hub database web site (nadh.ice.mpg.de/NaDH). This experiment had been conducted and involved additional tissues and physiological conditions in addition to those tissues and physiological conditions reported in the metabolomics study presented here.

Virus-Induced Gene Silencing. Vector construction, plant growth, and inoculation conditions were as described by Saedler and Baldwin (52) and are described in *SI Appendix, Materials and Methods*.

ACKNOWLEDGMENTS. We thank Dr. Mathias Schöttner for technical support in establishing the idMS/MS acquisition method and Dr. Klaus Gase for help with gene silencing construct design. D.L. and I.T.B. are funded by the Max Planck Society, by Advanced Grant 293926 of the European Research Council (to I.T.B.), and by the Collaborative Research Centre "Chemical Mediators in Complex Biosystems" (Grant SFB 1127). E.G.'s research in Heidelberg is supported within the framework of the Deutsche Forschungsgemeinschaft Excellence Initiative to the University of Heidelberg.

1. Weng JK, Philippe RN, Noel JP (2012) The rise of chemodiversity in plants. *Science* 336(6089):1667–1670.
2. Wink M (2003) Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry* 64(1):3–19.
3. Wink M, Carey DB (1994) Variability of quinolizidine alkaloid profiles of *Lupinus argenteus* (Fabaceae) from North-America. *Biochem Syst Ecol* 22(7):663–669.
4. Itkin M, et al. (2011) GLYCOALKALOID METABOLISM1 is required for steroidal alkaloid glycosylation and prevention of phytotoxicity in tomato. *Plant Cell* 23(12):4507–4525.
5. vonPoser GL, Toffoli ME, Sobral M, Henriques AT (1997) Iridoid glucosides substitution patterns in Verbenaceae and their taxonomic implication. *Plant Syst Evol* 205(3-4):265–287.
6. Matsuda F, et al. (2010) AtMetExpress development: A phytochemical atlas of Arabidopsis development. *Plant Physiol* 152(2):566–578.
7. Tissier A (2012) Glandular trichomes: What comes after expressed sequence tags? *Plant J* 70(1):51–68.
8. Schilmiller AL, et al. (2010) Studies of a biochemical factory: tomato trichome deep expressed sequence tag sequencing and proteomics. *Plant Physiol* 153(3):1212–1223.
9. Zang YX, et al. (2009) Genome-wide identification of glucosinolate synthesis genes in *Brassica rapa*. *FEBS J* 276(13):3559–3574.
10. Hirai MY, et al. (2007) Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proc Natl Acad Sci USA* 104(15):6478–6483.
11. Rajniak J, Barco B, Clay NK, Sattely ES (2015) A new cyanogenic metabolite in *Arabidopsis* required for inducible pathogen defence. *Nature* 525(7569):376–379.
12. Sakurai T, et al. (2013) PRIMe Update: Innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. *Plant Cell Physiol* 54(2):e5.
13. Bowen BP, Northen TR (2010) Dealing with the unknown: Metabolomics and metabolite atlases. *J Am Soc Mass Spectrom* 21(9):1471–1476.
14. Allard PM, et al. (2016) Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication. *Anal Chem* 88(6):3317–3323.
15. Broeckling CD, Heuberger AL, Prince JA, Ingelsson E, Prenni JE (2013) Assigning precursor-product ion relationships in indiscriminant MS/MS data from non-targeted metabolite profiling studies. *Metabolomics* 9(1):33–43.
16. Li D, Baldwin IT, Gaquerel E (2015) Navigating natural variation in herbivory-induced secondary metabolism in coyote tobacco populations using MS/MS structural analysis. *Proc Natl Acad Sci USA* 112(30):E4147–E4155.
17. Shannon CE (1948) A mathematical theory of communication. *AT&T Tech J* 27(3): 379–423.
18. Martínez O, Reyes-Valdés MH (2008) Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proc Natl Acad Sci USA* 105(28):9709–9714.
19. Heiling S, et al. (2010) Jasmonate and ppHsystemin regulate key Malonylation steps in the biosynthesis of 17-Hydroxygeranyllinalool Diterpene Glycosides, an abundant and effective direct defense against herbivores in *Nicotiana attenuata*. *Plant Cell* 22(1):273–292.
20. Li S, et al. (2012) Gene-sharing networks reveal organizing principles of transcriptomes in *Arabidopsis* and other multicellular organisms. *Plant Cell* 24(4):1362–1378.
21. Watrous J, et al. (2012) Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci USA* 109(26):E1743–E1752.
22. Nomura T, Murase T, Ogita S, Kato Y (2015) Molecular identification of tuliposide B-converting enzyme: A lactone-forming carboxylesterase from the pollen of tulip. *Plant J* 83(2):252–262.
23. Ulanowicz RE (2001) Information theory in ecology. *Comput Chem* 25(4):393–399.
24. Eren AM, Borisy GG, Huse SM, Mark Welch JL (2014) Oligotyping analysis of the human oral microbiome. *Proc Natl Acad Sci USA* 111(28):E2875–E2884.
25. Moco S, et al. (2007) Tissue specialization at the metabolite level is perceived during the development of tomato fruit. *J Exp Bot* 58(15-16):4131–4146.
26. Kessler D, Diezel C, Baldwin IT (2010) Changing pollinators as a means of escaping herbivores. *Curr Biol* 20(3):237–242.
27. Euler M, Baldwin IT (1996) The chemistry of defense and apparency in the corollas of *Nicotiana attenuata*. *Oecologia* 107(1):102–112.
28. Kessler D, Baldwin IT (2007) Making sense of nectar scents: The effects of nectar secondary metabolites on floral visitors of *Nicotiana attenuata*. *Plant J* 49(5):840–854.
29. Baldwin IT, Staszak-Kozinski L, Davidson R (1994) Up in smoke: I. Smoke-derived germination cues for postfire annual, *Nicotiana attenuata* torr. Ex. Watson. *J Chem Ecol* 20(9):2345–2371.
30. Baldwin IT, Morse L (1994) Up in smoke: II. Germination of *Nicotiana attenuata* in response to smoke-derived cues and nutrients in burned and unburned soils. *J Chem Ecol* 20(9):2373–2391.
31. Baldwin IT (2001) An ecologically motivated analysis of plant-herbivore interactions in native tobacco. *Plant Physiol* 127(4):1449–1458.
32. Gulati J, Kim SG, Baldwin IT, Gaquerel E (2013) Deciphering herbivory-induced gene-to-metabolite dynamics in *Nicotiana attenuata* tissues using a multifactorial approach. *Plant Physiol* 162(2):1042–1059.
33. Weinhold A, Baldwin IT (2011) Trichome-derived *O*-acyl sugars are a first meal for caterpillars that tags them for predation. *Proc Natl Acad Sci USA* 108(19):7855–7859.
34. Onkokesung N, et al. (2012) MYB8 controls inducible phenolamide levels by activating three novel hydroxycinnamoyl-coenzyme A:polyamine transferases in *Nicotiana attenuata*. *Plant Physiol* 158(1):389–407.
35. McKey D (1974) Adaptive patterns in alkaloid physiology. *Am Nat* 108(961):305–320.
36. McKey D (1979) The distribution of secondary compounds within plants. *Herbivores: Their Interaction with Secondary Plant Metabolites* (Academic, New York), pp 55–133.
37. Uhlén M, et al. (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347(6220):1260419.
38. Schmid M, et al. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat Genet* 37(5):501–506.
39. Bassard JE, Ullmann P, Bernier F, Werck-Reichhart D (2010) Phenolamides: Bridging polyamines to the phenolic metabolism. *Phytochemistry* 71(16):1808–1824.
40. Werner C, Hu WQ, Lorenziriatsch A, Hesse M (1995) Di-coumaroylspermidines and tri-coumaroylspermidines in anthers of different species of the genus *Aphelandra*. *Phytochemistry* 40(2):461–465.
41. Meurer B, Wiermann R, Strack D (1988) Phenylpropanoid patterns in *Fagales* pollen and their phylogenetic relevance. *Phytochemistry* 27(3):823–828.
42. Matsuno M, et al. (2009) Evolution of a novel phenolic pathway for pollen development. *Science* 325(5948):1688–1692.
43. Yang R, Wang X (2013) Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. *Plant Cell* 25(1):71–82.
44. Yonekura-Sakakibara K, et al. (2008) Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in *Arabidopsis*. *Plant Cell* 20(8):2160–2176.
45. Tohge T, Fernie AR (2010) Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. *Nat Protoc* 5(6):1210–1227.
46. Ginglinger JF, et al. (2013) Gene coexpression analysis reveals complex metabolism of the monoterpene alcohol linalool in *Arabidopsis* flowers. *Plant Cell* 25(11):4640–4657.
47. Mintz-Oron S, et al. (2008) Gene expression and metabolism in tomato fruit surface tissues. *Plant Physiol* 147(2):823–851.
48. Ma S, et al. (2015) Organization of the mammalian metabolome according to organ function, lineage specialization, and longevity. *Cell Metab* 22(2):332–343.
49. Stitz M, Hartl M, Baldwin IT, Gaquerel E (2014) Jasmonoyl-L-isoleucine coordinates metabolic networks required for anthesis and floral attractant emission in wild tobacco (*Nicotiana attenuata*). *Plant Cell* 26(10):3964–3983.
50. Horai H, et al. (2010) MassBank: A public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 45(7):703–714.
51. Birmingham A, et al. (2009) Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods* 6(8):569–575.
52. Saedler R, Baldwin IT (2004) Virus-induced gene silencing of jasmonate-induced direct defences, nicotine and trypsin proteinase-inhibitors in *Nicotiana attenuata*. *J Exp Bot* 55(395):151–157.
53. Snook ME, Chortyk OT, Sisson VA, Costello CE (1992) The flower flavonols of *Nicotiana* species. *Phytochemistry* 31(5):1639–1647.

Supporting Information

Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory

Dapeng Li, Sven Heiling, Ian T. Baldwin and Emmanuel Gaquerel

Materials and Methods

Conditions for UHPLC-ESI/qTOF-MS analysis

An Acclaim column (150×2.1 mm, particle size 2.2 µm) with a 4 mm×4 mm guard column of the same material was used for the analysis. The following binary gradient was used with a Dionex Ultimate 3000 UHPLC system: 0 to 1 min, isocratic 90% A (de-ionized water, 0.1% [vol/vol] acetonitrile and 0.05% formic acid), 10% B (acetonitrile and 0.05% formic acid); 1 to 40 min, gradient phase to 15% A, 85% B; 40 to 45 min, isocratic 15% A, 85% B. Flow rate was 300 µL/min. Eluted compounds were detected by a high-resolution micrOTOF-Q II mass spectrometer (Bruker Daltonics, Bremen, Germany) equipped with an electrospray ionization source operating in positive ionization mode. Typical instrument settings were as follows: capillary voltage 4500 V, capillary exit 130 V, dry gas temperature 180°C, dry gas flow of 8 L/min. Ions were detected from *m/z* 50 to 1400 at a repetition rate of 1 Hz. Mass calibration was performed using sodium formate clusters (10 mM solution of NaOH in 50/50% vol/vol isopropanol/water containing 0.2% formic acid). Raw data files were converted to netCDF format using the export function of the Data Analysis v4.0 software (Bruker Daltonics, Bremen, Germany).

Additional rules for the assembly of compound-specific idMS/MS

To reduce false positive errors resulting from spurious correlations from background noise due to the fact that some *m/z* features are only detected in a few samples, we compared data processing results obtained with and without the “fill peaks” function of XCMS (use for

background noise correction) and calculated a background noise value from the average correction estimate used by this function to replace “NA” not detected peak intensities. When the “fill peaks” function is used, there still were many “0” intensity values in the dataset which affect the calculation of correlations, and these were replaced with the calculated background value. We also only considered features with intensities that were more than 3 times the background value and considered these as “true peaks”. Only *m/z* signals with at least six “true peaks” for the 28 samples precursors (MS1) and fragments datasets were considered for PCC calculation.
A precursor mass feature is further defined if its intensities across samples significantly correlate with the decreased intensities of the same mass feature subjected to low or high collision energies and that this feature is not annotated as an isotope peak by CAMERA. The correlation analysis was then conducted by calculating all possible precursor-to-product pairs within 9s – estimated maximum retention time window for peak deviation. Logically, *m/z* values for fragments should be lower than that of the precursor and MS/MS fragmentation should occur in the same sample position within the 28 sample dataset as the precursor from which it is derived.

Based on these two simple rules, we excluded assigned fragments at *m/z* values larger than that of the identified precursor as well as fragments mismatched with precursor sample position. Many in-source-fragmentation-generated mass features produced in the MS1 mode can also be selected as candidate precursors resulting in redundant compound idMS/MSs. To reduce such data redundancy, we merged spectra if their NDP similarity exceeded 0.9 and they belong to the chromatographic “pcgroup” annotated by CAMERA. Finally we merged all the 4 collision energy results for precursor-to-fragment associations into a final deconvoluted spectrum by choosing the highest intensity peak among all candidate peaks of the same *m/z* value at the different collision energies. This latter processing step is based on the composite spectrum concept and accounts for the different collision energy conditions required to maximize fragmentation possibilities since certain fragments are detected only at specific collision energies. After applying the entire pipeline and set of rules, 895 deconvoluted non-redundant spectra were reconstructed from the tissue-wide analysis.

Information theory framework for defining tissue metabolic diversity and specialization

Tissue metabolic diversity was calculated using Shannon entropy of idMS/MS tissue-level frequency distribution by the following equation as described in Martinez et al., (2008) (1):

$$Hj = -\sum_{i=1}^{m} P_{ij} \log_2(P_{ij})$$

where P_{ij} correspond to relative frequency of the ith idMS/MS (i = 1, 2, …, m) in the jth tissue (j = 1, 2, …, t).

The average frequency of the ith idMS/MS among tissues was calculated as:

$$P_i = \frac{1}{t}\sum_{j=1}^{t} P_{ij}$$

idMS/MS specificity was calculated as:

$$S_i = \frac{1}{t}\left(\sum_{j=1}^{t} \frac{P_{ij}}{P_i}\log_2 \frac{P_{ij}}{P_i}\right)$$

The tissue specialization δj index was measured for each jth tissue, the average of the idMS/MS specificities using the following formula:

$$\delta_j = \sum_{i=1}^{m} P_{ij}\, S_i$$

idMS/MS molecular networking by bi-clustering

To perform this clustering, we used the R package DiffCoEx which is based an extension of the Weighted Gene Coexpression Analysis (WGCNA). Using NDP and NL-scoring matrices for 895 idMS/MS spectra, we computed a comparative correlation matrix using DiffCoEx with the parameters of “cutreeDynamic” set to method="hybrid", cutHeight = 0.999, deepSplit = T, minClusterSize = 10. The R source code of DiffCoEx is downloaded from additional file 1 in Tesson et al. (2010) (2), the required R WGCNA package can be found at http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA.

Gene-to-metabolite tissue-association similarities

Cross-tissue gene-to-metabolite associations provide valuable clues in formulating functional hypothesis about metabolic genes. To do so, we used as data input the idMS/MS and RNAseq binary data-sets computed separately for the following Z-scores: a Z-score of 1 indicating tissue-specificity and 0 indicating no tissue-specificity for a given feature. Similarities in gene and metabolite tissue-specificity were calculated using Ochiai coefficient calculated as following:

$$Ochiai = \frac{a}{\sqrt{(a+b)}\sqrt{(a+c)}}$$

where a is the number of tissue-associations for which both a metabolite and a gene exhibit a Z-score of 1, b is the number of tissue-associations where a metabolite is 1 and the gene is 0, c is the number of tissue-associations where the metabolite is 0 and the gene is 1.

Virus-induced gene silencing (VIGS)

Vector construction, plant growth, and inoculation conditions were as described by Saedler and Baldwin (2004) (3). Briefly, 200- to 300-bp fragments of *N. attenuata* target genes were amplified by PCR using specific primer pairs as listed in ***SI Appendix,* Table S1**. Amplified fragments were cloned into pTV00 vector, and plasmids were transformed by electroporation into *Agrobacterium tumefaciens* strain GV3101. A pTV00 plasmid without insert (EV) was used as a negative control in all experiments. Three leaves of 24- to 25-d-old *N. attenuata* plants were infiltrated with a 1:1 mixture of *A. tumefaciens* transformed with pBINTRA and one of the gene fragment-containing construct or the pTV00 construct. *Phytoene desaturase* (pTVPDS) causing bleaching of tobacco leaves due to the depletion of carotenoids was used as a positive control to monitor the progression of VIGS in a separate set of inoculated plants. VIGS-silenced plants were used for treatment after PDS-VIGS leaves developed a strong bleaching phenotype. Silencing efficiency was verified by RT-qPCR of target gene transcripts after RNA extraction and cDNA synthesis.

RT-qPCR analysis of gene silencing efficiency

Total RNA was extracted by adding Trizol reagent (Invitrogen; http://www.invitrogen.com) to approximately 150 mg of powdered leaf material ground in liquid nitrogen following the manufacturer's protocol. A total of 500 ng of DNA-free RNA samples was reverse transcribed using oligo(dT)18 primers and SuperScript II enzyme (Invitrogen) following the manufacturer's recommendations. All RT-qPCR assays were performed with a Stratagene MX3005P instrument (http://www.stratagene.com) as recommended by the manufacturer. To normalize transcript levels, primers specific for the elongation factor-1α gene from *Nicotiana tabacum* (EF1-α; accession no. D63396) were used. Specific primers in the 5′ to 3′ direction used for SYBR Green-based analyses are listed in ***SI Appendix,* Table S1**.

Accession numbers

For construction of the UDP-Glycosyltransferase tree in different species, we used the following Genbank accessions for *Allium cepa* (UGT73G1, AAP88406.1; UGT73J1, AAP88407.1), *Antirrhinum majus* (AmC4GT, BAE48239; UGT73E2 (Amugt36), BAG16513.1;

UGT73N1 (Amugt38), BAG16514.1; UGT88D4, BAG31945), *Arabidopsis thaliana* (AtF3G7GT, Q9ZQ95; AtF3GT, AAM91339; AtF5GT, AAM91686; AtF7GT, AAL90934; AtUGT89C1, AAP31923; DOGT1, NP_181218; UGT75D1, AAB58497.1), *Aralia cordata* (AcGAT, BAD06514), *Avena sativa* (AvUGT80A1, CAB06081), *Bellis perennis* (UGT94B1 (BpUGAT), BAD77944), *Beta vulgaris* (BvUGT71F1, AAS94330; BvUGT73A4, AAS94329.1), *Brassica napus* (UGT84A9 (BnSGT1), AF287143_1), *Catharanthus roseus* (CaUGT3, BAH80312; CaUGT1, BAD29721; CaUGT2, BAD29722; UGT85A2a (CrUGT6), BAK55749; UGT709C2 (CrUGT8),BAO01109), *Celosia cristata* (CcCDOPA5GT, BAD91804), *Citrus maxima* (CmF7G12RT, AAL06646), *Citrus sinensis* (CsUFGT, AAS00612), *Citrus unshiu* (CuLGT, BAA93039), *Crocus sativus* (CsGT45, ACM66950.1; CsUGT707B1, CCG85331; Glt2 (UGTCs2), AAP94878.1), *Dianthus caryophyllus* (DcF3GT, BAD52004; DicGT1, BAD52003; DicGT2, BAD52005; DicGT4 (DcC2GT), BAD52006; DicGT5, BAD52007), *Dorotheanthus bellidiformis* (DbB5GT, CAB56231; DbB6GT, AAL57240), *Forsythia x intermedia* (FiF3GT, AAD21086), *Fragaria x ananassa* (FaFGT, AAU12367; FaGT2, AAU09443), *Gentiana triflora* (Gt5GT7, BAG32255; GtF3GT, BAA12737; GtGTX, BAC54092), *Glycine max* (GmF3G6R, BAN91401; GmIF7GT, BAF64416), *Glycyrrhiza echinata* (GeIF7GT, BAC78438), *Hordeum vulgare subsp. vulgar* (HvF3GT, CAA33729), *Ipomoea nil* (In3GGT(InA32GT), BAD95885; InGTase1, BAF75917), *Ipomoea purpurea* (Ip3GGT(IpA32GT), BAD95882), *Iris x hollandica* (Ih3GT, BAD83701; Ih5GT, BAD06874), *Lamium galeobdolon* (LgF3GT, AEB61487), *Linaria vulgaris* (LvC4GT, BAE48240), *Lycium barbarum* (Ugt73a10, BAG80536), *Maclura pomifera* (MpUGT75L4, ABL85474; MpUGT88A4, ABL85471), *Medicago trunculata* (MtUGT73C8, ABI94020; MtUGT73K1, AAW56091; MtUGT73P1, ABI94026; MtUGT78G1, ABI94025; MtUGT84F1, ABI94023; MtUGT85H2, ABI94024.1; MtUGT88E1, ABI94021; MtUGT88E2, ABI94025; UGT71G1, AAW56092), *Mirabilis jalapa* (CDOPA5GT, BAD91804), *Nicotiana tabacum* (NTGT1A, BAB60720.1; NTGT1b, BAB60721.1; NtGT2, BAB88935; NtGT3, BAB88934; NtSAGT, AAF61647; TOGT 1, AAK28303; TOGT 2, AAK28304), *Perilla frutescens* (PfA5GT, BAA36421; PfF3GT, BAA19659; PfUGT88D7 (F7GAT), BAG31948), *Petunia x hybrida* (PhA3ART, CAA50376; PhA3GT, BAA89008; PhA5GT, BAA89009; PhF3GalTase, AF165148_1), *Phaseolus lunatus* (PlZOG1, AAD04166); *Phaseolus vulgaris* (PvZOX1, AF116858_1), *Phytolacca americana* (PaGT2, BAG71125; PaGT3, BAG71127), *Pilosella officinarum* (PoUGT95A1, ACB56927), *Prunus dulcis* (PdUGT85A19, ABV68925), *Pyrus communis* (PcF7GT, AAY27090), *Quercus robur* (QrUGT84A13, AHA54051), *Rauwolfia serpentine* (RsAS, CAC35167), *Rhodiola sachalinensis* (RsUGT73B6, AAS55083; RsUGT74R1, ABP49574; RsUGT72B14, ACD87062), *Rosa hybrida* (RhA53GT, BAD99560), *Scutellaria baicalensis* (SbB7GAT, BAD99560; SbF7GT, BAA83484), *Scutellaria laeteviolacea var. yakusimensis* (SlUGT88D5, BAG31946), *Sesamum indicum* (SiUGT88D6, BAG31947),

Solanum aculeatissimum (SaGT4A, BAD89042.1), Solanum berthaultii (SbGT, AAB62270.1), Solanum lycopersicum (Gtsatom, CAI62049.1), *Solanum melongena* (SmUGT76, CAA54558.1), *Solanum tuberosum* (Sgt2.1, ABB29873.1; Sgt2.2, ABB29874.1; StSgt1, AAB48444.1; StSgt3, ABB84472.1), *Stevia rebaudiana* (SrUGT74G1, AY345982; SrUGT76G1, AY345974; SrUGT85C2, AY345978), *Torenia hybrida* (ThA5GT, BAC54093), *Verbena hybrida* (VhA5GT, BAA36423), *Vigna angularis* (VaABAGT, BAB83692), Vigna mungo (VmUF3GaT, BAA36972; VmUFGlyT, BAA36410), *Vitis labrusca* (VlGT, ABR24135), *Vitis vinifera* (VvGT1, AAB81682), *Withania somnifera* (WsFGT, FJ560880; WsUGT73A16, FJ654696/ACO44747.1), *Zea mays* (Zm-BX8, AF331854_1; Zm-BX9, CAX02221; ZmcisZog1, AAK53551; ZmcisZog2, AAL92460; ZmIaglu, AAA59054; ZmUFGT, CAA30760; ZmUGT71A1, CAA31856).

Tables

Table S1. List of primers used for qRT-PCR and gene fragment cloning for VIGS

Name	For VIGS cloning (5' to 3')	For qRT-PCR (5' to 3')
UGT-A forward	GCGGCGCTCGAGGTTGAGCATTATACTAAGGTGC	GACGCTAGAAGGAGTTTCAGG
UGT-A reverse	GCGGCGGGATCCCAGGCAACCAATCTTCGTTGTC	CCACTGAATCGAACCAACAC
UGT-B forward	GCGGCGCTCGAGGTGGTCCTACTGTATATGACCG	GGGAATTATTCATTCAGGTTGGG
UGT-B reverse	GCGGCGGGATCCGGTAGCCCAGTTTGCTCCAGAC	GGCAACATAACCACTTGACAG
Elongation factor forward		TGGTATGGTTAAGATGCTTCCC
Elongation factor reverse		TGTCAACGCTCTTGATAACAC

Table S2. Overview of RNAseq transcriptome data

NCBI accession number	Tissue type	Treatment/development stage	Additional note on sampling procedure
NA1498ROT	Root (ROT)	Rosette stage plants, treated with 5 µL 1:1 diluted *M. sexta* oral secretion three times in leaves	Roots of rosette stage plants that were treated three times on leaves were collected for RNA extraction. The treatments were performed at 10 am and 6pm on the day before sampling and 10 am on the day of sampling. Samples were collected at 11 am.
NA1500LET	Leaf (LET)	Rosette stage plants, treated with 5 µL 1:1 diluted *M. sexta* oral secretion three times in leaves	Local leaves of rosette stage plants that were treated three times on leaves were collected for RNA extraction. The treatments were performed at10 am, 6pm on the day before sampling and 10 am on the day of sampling. Samples were collected at 11 am.
NA1717LEC	Leaf (LEC)	Rosette stage plants, no treatment	Rosette stage leaves were collected for RNA extraction. Samples were collected at 11 am.

NA1504STT	Stem (STT)	Rosette stage plants, treated with 5 µL 1:1 diluted *M. sexta* oral secretion three times in leaves	Stems of rosette stage plants that were treated three times on leaves were collected for RNA extraction. The treatments were performed at 10 am, 6pm on the day before sampling and 10 am on the day of sampling. Samples were collected at 11 am.
NA1505COE	Corolla (COE)	Early developmental stage, no treatment	Samples were collected in the afternoon, 60 samples were pooled.
NA1515COL	Corolla (COL)	Late developmental stage, no treatment	Samples were collected at 6 pm (open flowers) and 9am (closed flower after opening), 4-10 samples were pooled.
NA1506STI	Stigma (STI)	Mature stigma, no treatment	Stigma samples were collected in the afternoon, 40 samples were pooled.
NA1507POL	Pollen tube (POL)	No treatment	Pollen tubes were pooled.
NA1508SNP	Style (SNP)	Mature style without pollination	Styles were collected at 7 am, anthers were removed one day before, and 50 samples were pooled.
NA1509STO	Style (STO)	Mature style, pollinated with pollens from different genotype	Styles were collected at two hours after pollination, at 7 am. Anthers were removed one day before, and 30 samples were pooled.
NA1510STS	Style (STS)	Mature style, self-pollinated	Styles were collected at two hours after pollination, at 7 am. Anthers were removed one day before, and 30 samples were pooled.
NA1511NEC	Nectary (NEC)	Mature nectary, no treatment	Samples were collected in the afternoon, 60 samples were pooled.
NA1512ANT	Anther (ANT)	Mature anther no treatment	Samples were collected in the afternoon, 60 samples were pooled.
NA1513OVA	Ovary (OVA)	Mature ovary, no treatment	Samples were collected in the afternoon, 60 samples were pooled.
NA1514PED	Pedicel (PED)	Mature pedicel, no treatment	Samples collected at 9 am (heading down) and 4 pm (heading up) were pooled.
NA1516OFL	Flower (OFL)	Fully opened flowers, no treatment	Both morning (7 am) and evening (6 pm) flowers were collected, 1 sample of each were pooled.
NA1517FLB	Flower bud (FLB)	Two early developmental stages of flowers, no treatment	Samples were collected at 6pm, 1 bud and 1 middle stage flower were collected. Sepals were removed from the samples.
NA1501SES	Seed (SES)	Treated with liquid smoke	100 mg seeds treated with 1:50 diluted liquid smoke solution for 9-15 min were used for RNA extraction.
NA1502SEW	Seed (SEW)	Treated with water	100 mg seeds treated with water for 9-15 min were used for total RNA extraction.
NA1503SED	Seed (SED)	Dry seeds	100 mg dried seeds directly used for total RNA extraction.

Table S3. Performance of the Ochiai distance-based integration of genes and idMS/MS data for known gene-metabolite associations

Gene name	Function	Reference		Metabolites	PCC		Ochiai	
					coeff.	*P*-value	coeff.	*P*-value
		1,2		N-caffeoylputrescine (mz 251.14 @127s)	0.00	1.0	0.63	0.0
MYB8	transcription factor	1,2		N-feruloylputrescine (mz 265.15 @280s)	-0.21	0.5	0.35	0.3
		1,2		N-Caffeoylspermidine (mz 308.20 @418s)	-0.01	1.0	0.45	0.1
		1,2		N-caffeoylputrescine (mz 251.14 @127s)	0.07	0.8	0.71	0.0
AT1	hydroxycinnamoyl-CoA: putrescine transferase	1,2		N-feruloylputrescine (mz 265.15 @280s)	-0.14	0.7	0.47	0.1
		1,2		Unidentified putrescine-based phenolamide (mz 350.21 @221s)	-0.17	0.6	0.57	0.1
		1,2		N-Caffeoylspermidine (mz 308.20 @418s)	0.16	0.6	0.68	0.0
		1,2		N',N''-Di-caffeoylspermidine (mz 470.23 @396s)	0.55	0.1	0.72	0.0
DH29	hydroxycinnamoyl-CoA: spermidine transferase	1,2		N,N'-Coumaroyl,caffeoylspermidine (mz 454.23 @480s)	-0.03	0.9	0.72	0.0
		1,2		N',N''-Caffeoyl,feruloylspermidine (mz 484.24 @506s)	0.03	0.9	0.72	0.0
		1,2		N',N''-Di-feruloyl-spermidine (mz 498.26 @591s)	0.23	0.5	0.76	0.0
		1,2		Unidentified spermidine-based phenolamide (mz 468.21 @228s)	0.12	0.7	0.71	0.0
		1,2		N',N''-Di-caffeoylspermidine (mz 470.23 @396s)	0.44	0.2	0.63	0.0
CV86	hydroxycinnamoyl CoA:hydroxycinnamoylspermidine-conjugating activity	1,2		N,N'-Coumaroyl,caffeoylspermidine (mz 454.23 @480s)	0.43	0.2	0.63	0.0
		1,2		N',N''-Caffeoyl,feruloylspermidine (mz 484.24 @506s)	0.28	0.4	0.63	0.0
		1,2		N',N''-Di-feruloyl-spermidine (mz 498.26 @591s)	0.45	0.1	0.67	0.0
PAL	Phenylalanine ammonia lyase	1,2	Other phenylpropanoids	Phenylalanine (mz 166.09 @132s)	-0.17	0.6	0.35	0.3
THT	Tyramine N-hydroxycinnamoyltransferase	3		N-coumaroyltyramine (mz 284.10 @95s)	-0.01	1.0	0.71	0.0
HQT	hydroxycinnamoyl CoA quinate transferase	4		Chlorogenic acid (mz 355.10 @285s)	0.36	0.2	0.44	0.2
		This study		Rutin (mz 611.16 @513s)	0.75	0.0	0.52	0.1
		This study		Quercetin-3-O-sophoroside (mz 627.16 @430s)	-0.13	0.7	0.26	0.4
		This study		Quercetin-3-O-glucose (mz 465.10 @546s)	0.09	0.8	0.71	0.0
UGT-A	Flavonoid UDP-Glucosyltransferase	This study		Kaempferol-3-O-sophoroside (mz 611.16 @486s)	0.01	1.0	0.26	0.4
		This study		Kaempferol-3-O-glucose-rhamnose (mz 595.17 @580s)	0.07	0.8	0.52	0.1
		This study		Kaempferol-3-O-glucose (mz 449.11 @616s)	0.08	0.8	0.52	0.1
		This study		Kaempferol-3-O-rutinoside-glucoside (mz 757.22 @342s)	-0.16	0.6	0.58	0.0
		This study		Rutin (mz 633.14 @513s)	0.46	0.1	0.71	0.0
		This study		Quercetin-3-O-sophoroside (mz 627.16 @430s)	-0.44	0.1	0.45	0.1
UGT-B	Flavonoid UDP-Rhamnosyltransferase	This study		Quercetin-3-O-glucose (mz 465.10 @546s)	-0.22	0.5	0.61	0.0
		This study		Kaempferol-3-O-sophoroside (mz 611.16 @486s)	-0.38	0.2	0.45	0.1
		This study		Kaempferol-3-O-glucose-rhamnose (mz 595.17 @580s)	-0.27	0.4	0.45	0.1
		This study		Kaempferol-3-O-glucose (mz 449.11 @616s)	-0.26	0.4	0.45	0.1
		5,6		Lyciumoside I (mz 653.35 @1099s)	0.91	0.0	0.75	0.0
GGPPS	geranylgeranyl pyrophosphate synthase	5,6		Lyciumoside IV (mz 799. 41 @1087s)	0.52	0.1	0.53	0.1
		5,6		Lyciumoside II (mz 815.10 @999s)	0.63	0.0	0.76	0.0
		7		Lyciumoside I (mz 653.35 @1099s)	0.89	0.0	0.89	0.0
GLS	Geranyllinalool Synthase	7		Lyciumoside IV (mz 799. 41 @1087s)	0.51	0.1	0.79	0.0
		7		Lyciumoside II (mz 815.10 @999s)	0.55	0.1	0.85	0.0
		in preparation		Lyciumoside IV (mz 799. 41 @1087s)	0.50	0.1	0.72	0.0
UGT74P3	HGL-DTG UDP-Glycosyltransferase	in preparation		Attenoside (mz 939.48 @990s)	0.07	0.8	0.72	0.0
		in preparation		Nicotianoside III (mz 940.49 @992s)	-0.01	1.0	0.72	0.0
		in preparation		Lyciumoside II (mz 815.10 @999s)	0.58	0.0	0.77	0.0
		in preparation		Lyciumoside IV (mz 799. 41 @1087s)	0.52	0.1	0.79	0.0
UGT91T1	HGL-DTG UDP-Glycosyltransferase	in preparation		Attenoside (mz 939.48 @990s)	0.25	0.4	0.79	0.0
		in preparation		Nicotianoside III (mz 940.49 @992s)	0.15	0.7	0.79	0.0
		in preparation		Lyciumoside II (mz 815.10 @999s)	0.63	0.0	0.68	0.0
		8		O-AS#9 Class II (mz 631.29 @1471s)	-0.22	0.5	0.63	0.0
		8		O-AS#2 Class III (mz 687.32 @1771s)	-0.22	0.5	0.89	0.0
ASAT1	acylsucrose acyltransferase	8		O-AS#8 Class III (mz 654.33 @1550s)	-0.13	0.7	0.77	0.0
		8		O-AS#19 Class IV (mz 715.315 @1856s)	-0.20	0.5	0.77	0.0
		8		O-AS#45 Class IV (mz 701.30 @1738s)	-0.18	0.6	0.63	0.0
		8		O-AS#9 Class II (mz 631.29 @1471s)	0.77	0.0	0.53	0.1
ASAT2	acylsucrose acyltransferase	8		O-AS#2 Class III (mz 687.32 @1771s)	0.92	0.0	0.76	0.0
		8		O-AS#8 Class III (mz 654.33 @1550s)	0.55	0.1	0.65	0.0
		8		O-AS#4 Class IV (mz 679.35 @1892s)	0.81	0.0	0.85	0.0
ASH1	acylsugar acylhydrolase	9		O-AS#2 Class III (mz 687.32 @1771s)	0.45	0.1	0.50	0.1
		9		O-AS#4 Class IV (mz 679.35 @1892s)	0.40	0.2	0.67	0.0
		9		O-AS#9 Class II (mz 631.29 @1471s)	-0.05	0.9	0.35	0.3
ASH2	acylsugar acylhydrolase	9		O-AS#2 Class III (mz 687.32 @1771s)	0.06	0.9	0.50	0.1
		9		O-AS#4 Class IV (mz 679.35 @1892s)	-0.07	0.8	0.45	0.1

References for Table S3: 1, Onkokesung et al. Plant Physiology 158.1 (2012): 389-407; **2,** Jillian et al. Planta 221.6 (2005): 904-914; **3,** Gaquerel et al. Journal of Agricultural and Food Chemistry 58.17 (2010): 9418-9427; **4,** Ricarda et al. Nature Biotechnology 22.6 (2004): 746-754; **5,** Heiling et al. The Plant Cell 22.1 (2010): 273-292; **6,** Heiling, et al. The Plant Journal 85.4 (2016): 561-577; **7,** Falara, et al. Plant Physiology 166.1 (2014): 428-441; **8,** Fan, et al. Proc. Natl. Acad. Sci. U. S. A. 113.2 (2015): E239-E248; **9,** Schilmiller, et al. Plant Physiology 170.3 (2016): 1331-1344

Datasets

Dataset S1. idMS/MS data analysis.

Spreadsheet S1.1, XCMS-processed data-set; **Spreadsheet S1.2,** cross-tissue ZMAD scaled data and results of the Kurtosis analysis; **Spreadsheet S1.3,** deconvoluted idMS/MS spectra; **Spreadsheet S1.4,** structural and tissue annotation of the idMS/MS bi-clustering.

Figures

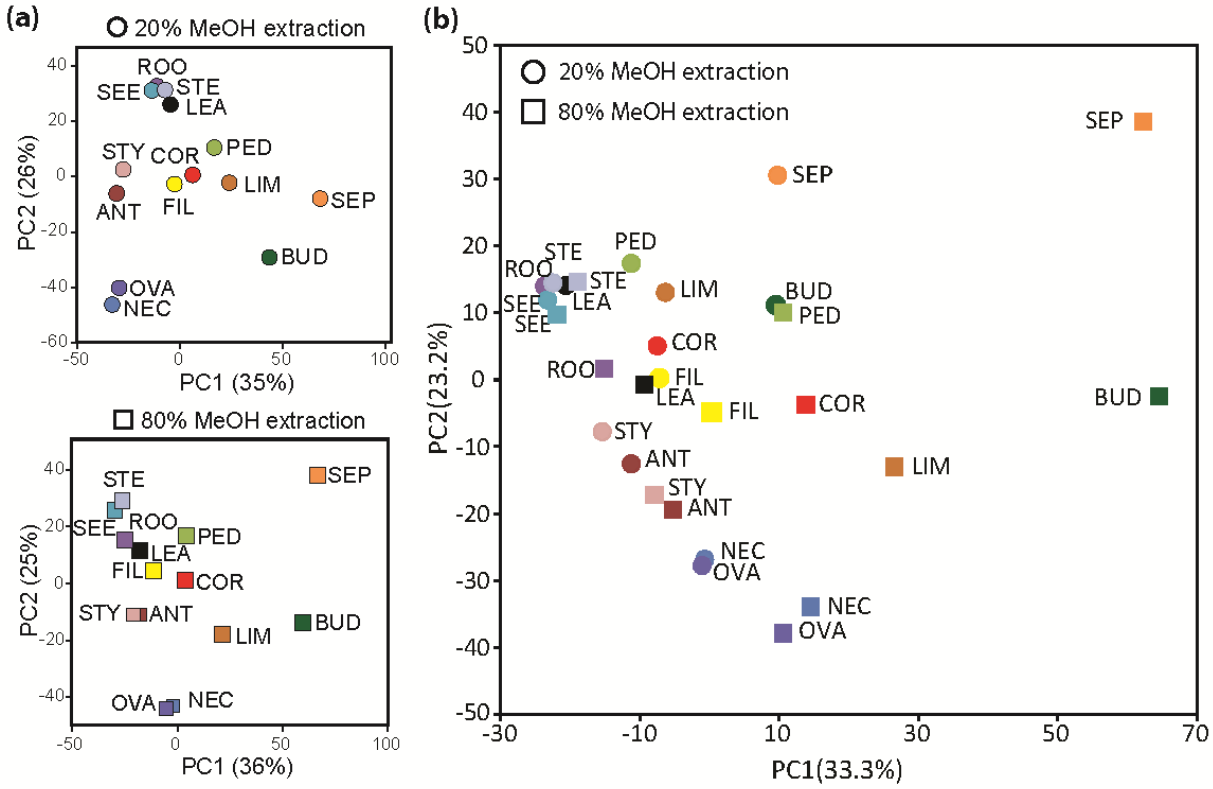

Figure S1. Principal component analysis (PCA) of the UHPLC-ESI/qTOF-MS metabolic profiles for all tissue and extraction types. The PCA score plot was conducted on the auto-scaled complete mass feature matrix resulting from XCMS processing of the samples extracted by 20% and 80% methanol (vol/vol) which are represented as circle and rectangular shapes respectively. Detailed explanations of the tissue collection procedure are provided in the Method section. **(a)** Separate PCA score plot of 20% and 80% methanolic (vol/vol) extracts. **(b)** PCA of the combined data-set.

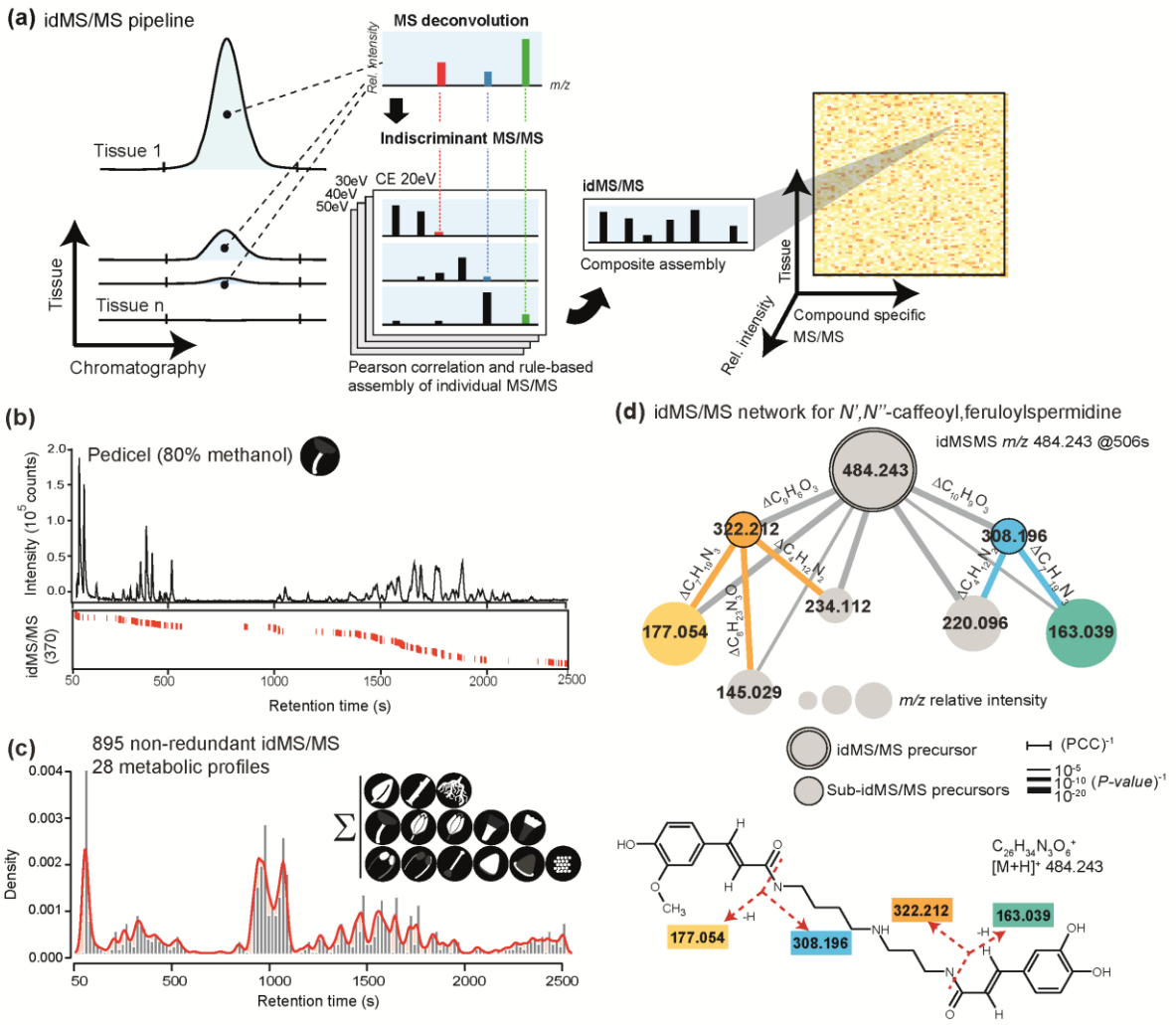

Figure S2. Tissue-wide indiscriminant acquisition and assembly of metabolite MS/MS spectral information. (a) Optimized pipeline to achieve indiscriminant acquisition of MS/MS data from tissue-wide metabolic variations. Indiscriminant MS/MS targets every mass signal within a range of 50 to 1400 *m/z* for fragmentation using 4 increasing collision energies (CE). Each tissue sample is processed using idMS/MS via individual analyses performed at different CID voltages in order to maximize fragment diversity. Information regarding a fragment's assignment to a given precursor mass is lost during indiscriminant MS/MS (idMS/MS) but can be computationally-retrieved for each CID voltage run, based on mathematical and chromatographic correlation analyses (see Method section). The pipeline harnesses the important chemical diversity among tissue samples. Reciprocally, cross-tissue quantitative variation provides the statistical power required for computing high confidence Pearson correlation coefficients (PCC) among the variation in intensities of precursors and candidate fragments. idMS/MSs assembled at each CID voltage for a given precursor *m/z* occurring at a

given retention time are then merged together into a composite MS/MS which displays the complete fragment diversity. The resulting output is a three-dimensional entry matrix with non-redundant idMS/MS spectra across tissues and their relative intensities. **(b)** idMS/MS coverage for a representative 80% methanol (vol/vol) pedicel sample. The lower heatmap displays the retention time position, aligned to the total ion current chromatogram, of the 370 idMS/MS spectra recorded for this sample. **(c)** Density plot summarizing the idMS/MS coverage across all 14 analyzed tissues denoted by their symbols. Bars represent the relative density of collected idMS/MSs for a 9 s retention time window. Red lines represent smoothed density curves. **(d)** Example of the idMS/MS obtained for *m/z* 484.243 @ 506 s corresponds to the $[M+H]^+$ of an *N',N''*-caffeoyl,feruloylspermidine isomer – ' and '' denote that the exact position of the caffeoyl and feruloyl moieties on the spermidine backbone cannot be assigned by MS analysis alone. idMS/MSs were also assembled for *m/z* 322.212 and *m/z* 308.196 as these *m/z* signals are already present as in-source ionization derived fragments in the MS mode analysis.

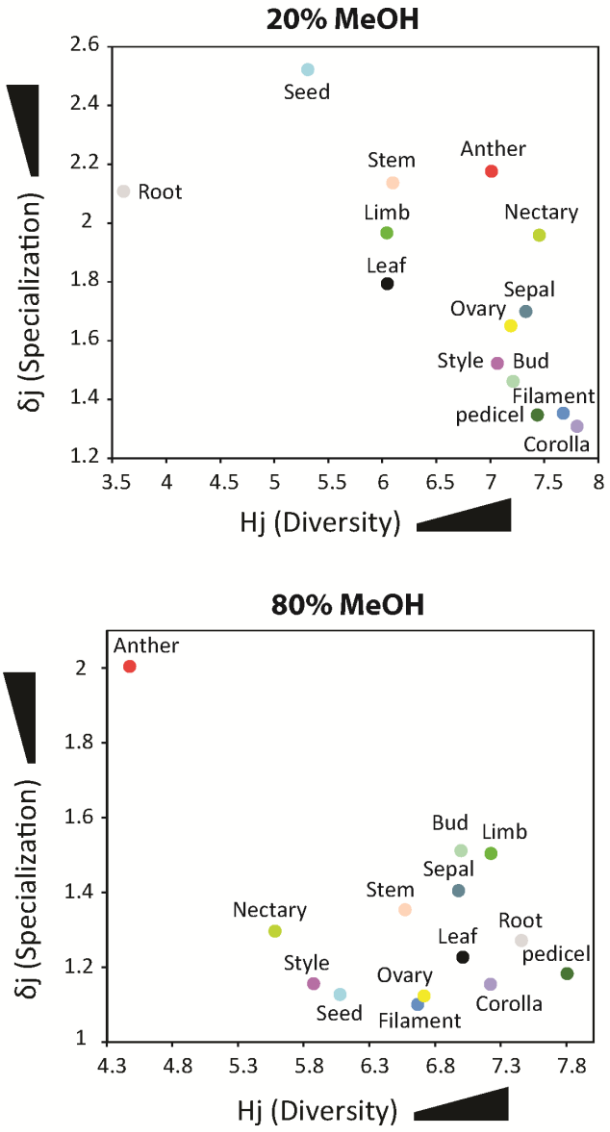

Figure S3. Information theory-based analysis of the degree of specialization and diversity in the idMS/MS composition across tissues. Tissue specialization (δj) and diversity (Hj) are mapped in a two-dimensional space using two indexes, where Hj, tissue-level idMS/MS diversity is calculated by Shannon entropy of idMS/MS frequency distribution of each tissue and δj, tissue-level idMS/MS specialization is measured by the average specificity of each idMS/MS component by taking into consideration its frequency among tissues. Tissue metabolomes extracted for two different extraction conditions (20% or 80% methanol (vol/vol)) were separately analyzed and visualized into two panels and tissue types are differentiated by different colors.

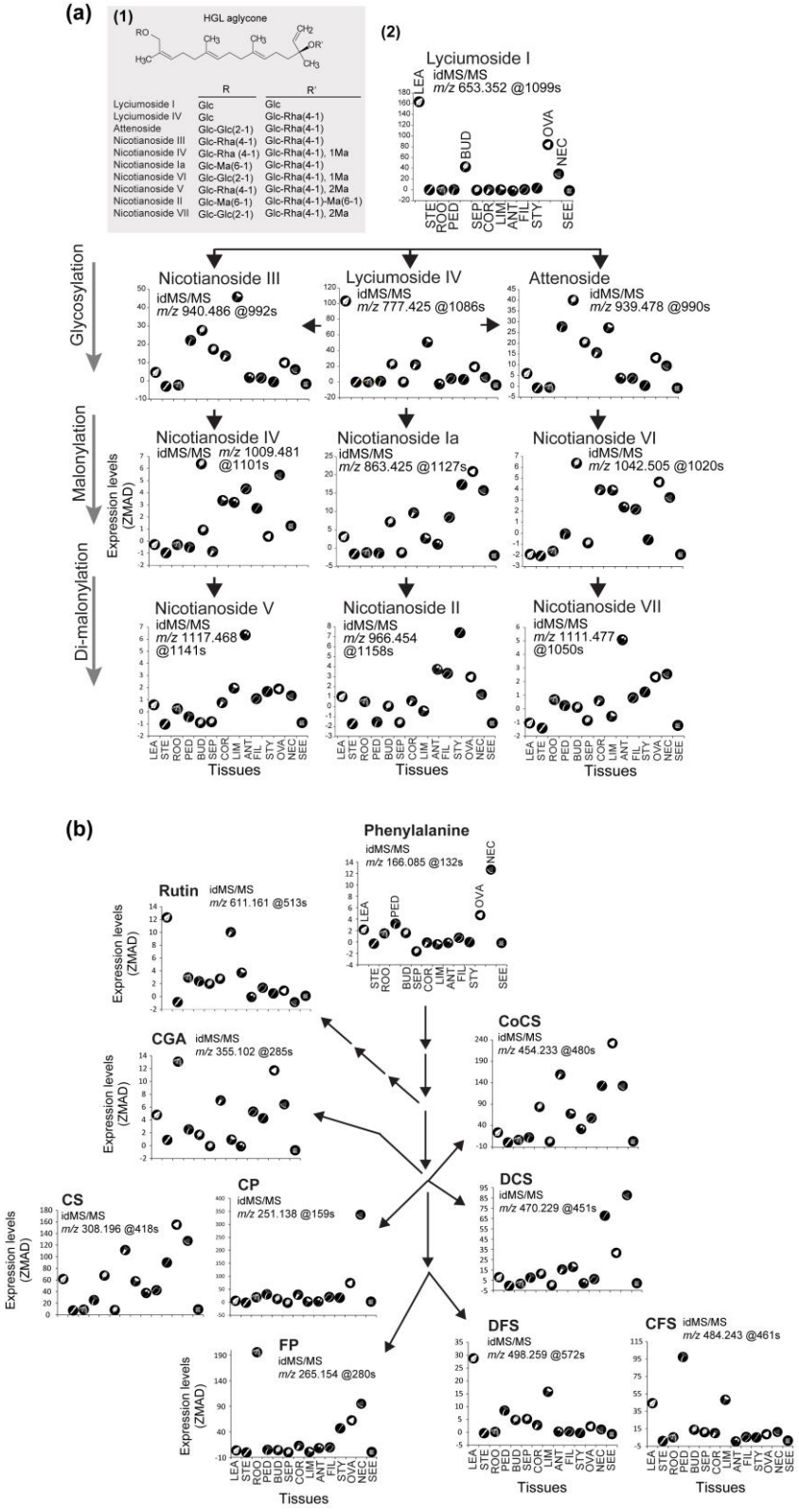

Figure S4. Tissue-based variations in the accumulation of HGL-DTGs and phenolic derivatives. 17-hydroxygeranyllinalool (HGL) diterpene glycosides are abundant secondary

metabolites in *N. attenuata* which differ in the number and types of sugar (Glc, glucose; Rha, Rhamnose; with or without malonyl, Ma, groups) decorations added to the acyclic HGL backbone which is characteristic of this compound family **(a.1)**. Z-score-normalized median absolute distances are employed to visualize cross-tissue variations in the idMS/MSs corresponding to the main **(a.2)** HGL-DTG intermediates as well as **(b)** major phenylpropanoid-quinate and -polyamine conjugates. For both metabolic groups, important changes in cross-tissue variations are detected. In the case of the HGL-DTG metabolic pathway, a progressive enrichment of certain metabolites is noticeable in reproductive floral tissues. CFS, *N',N''*-caffeoyl,feruloyl-spermidine; CGA, Chlorogenic acid; CoCS, *N',N''*-coumaraoyl,caffeoyl-spermidine; CP, *N*-caffeoylputrescine; CS, *N*-caffeoylspermine; FP, *N*-feruloylputrescine; DCS, *N', N''*-dicaffeoylspermidine; DFS, *N',N''*-diferuloyl-spermidine.

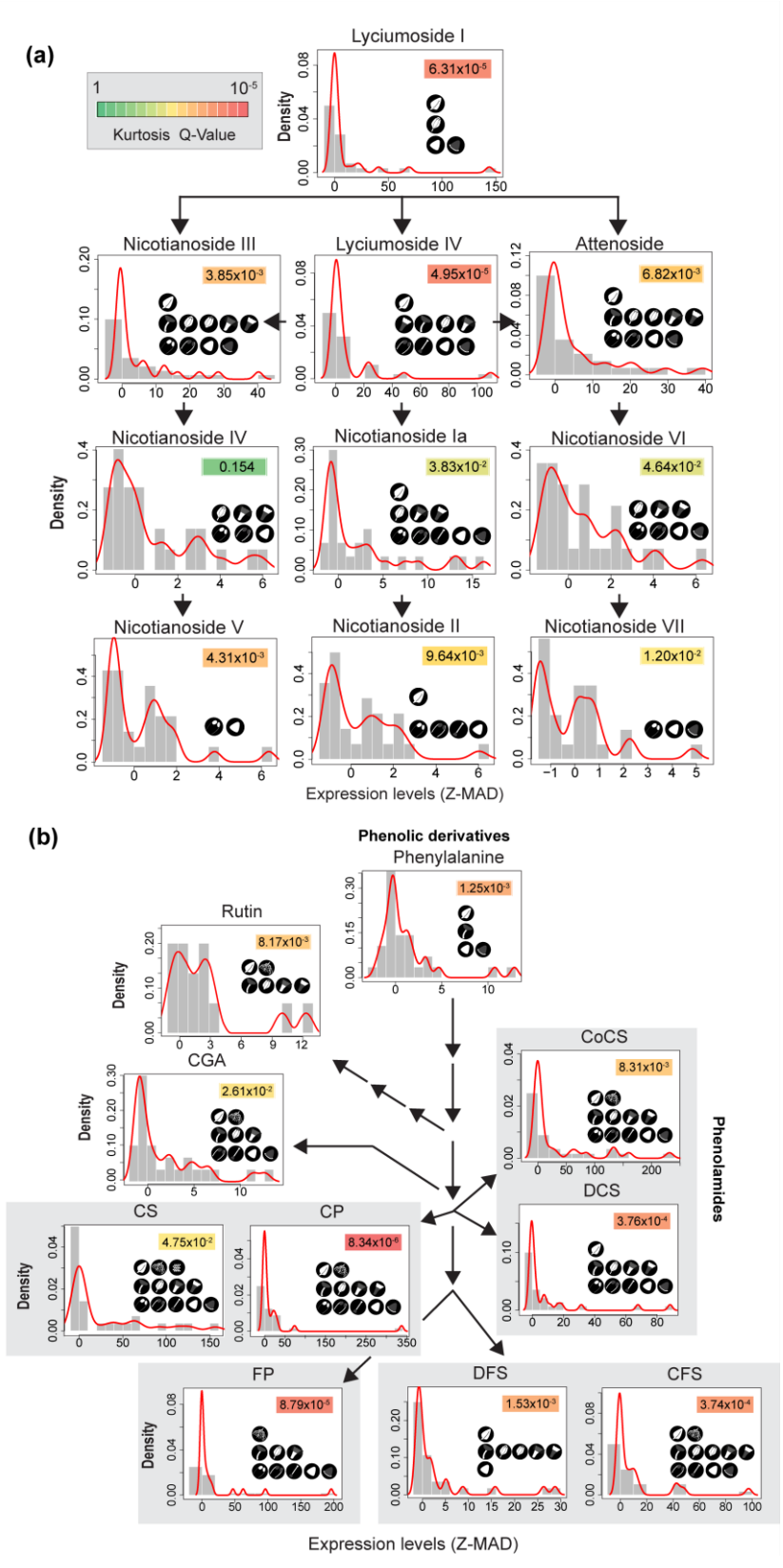

Figure S5. Significant levels of tissue-specificity derived from Kurtosis filtering for particular metabolic steps in the HGL-DTG and phenolic pathway. Q-values present in color

boxes derived from Kurtosis filtering analysis (see Method section) for the distribution of idMS/MSs corresponding to particular HGL-DTGs (17-hydrogeranyllinalool diterpene glycosides) **(a)** and phenolic derivatives **(b)**. Associated tissues for each idMS/MS are presented as tissue icons below the Q-value box. Low Q-values indicate strong tissue-specificity in the accumulation of particular HGL-DTGs and phenolic derivatives. Strongest tissue-associations are detected for upstream steps in the pathway indicating clear tissue-specificities in the biosynthesis of these upstream intermediates from which the complete HGL-DTG chemotype derives from. Complete results of the Kurtosis analysis are presented in ***SI Appendix,* Dataset S1**. CFS, *N',N''*-caffeoyl,feruloyl-spermidine; CGA, Chlorogenic acid; CoCS, *N',N''*-coumaraoyl,caffeoyl-spermidine; CP, *N*-caffeoylputrescine; CS, *N*-caffeoylspermine; FP, *N*-feruloylputrescine; DCS, *N', N''*-dicaffeoylspermidine; DFS, *N',N''*-diferuloyl-spermidine.

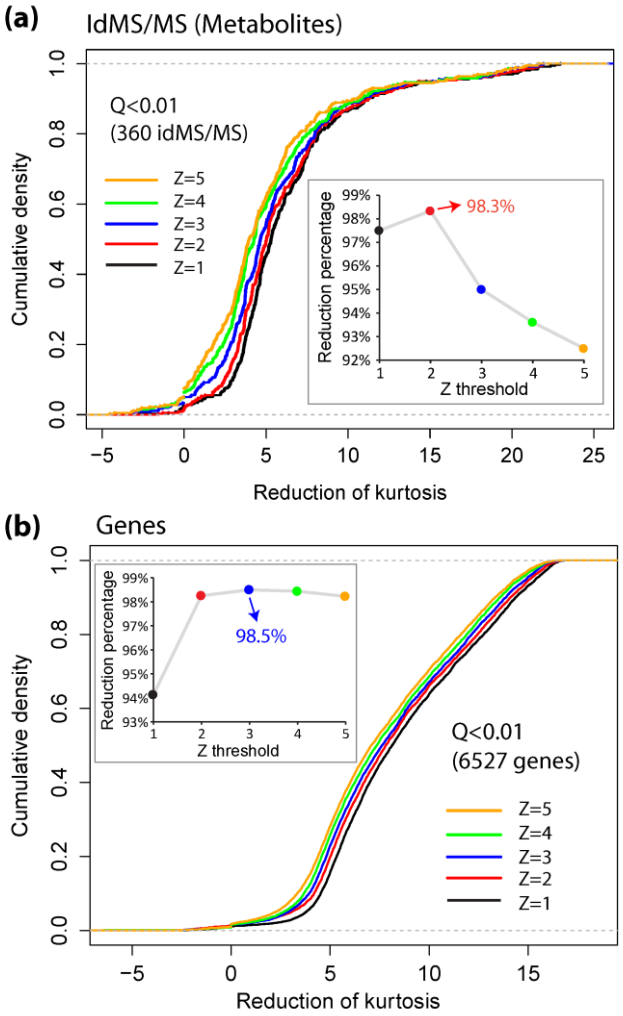

Figure S6. Implementing the reduction of Kurtosis analysis to infer idMS/MS and gene expression with high tissue-specificity. Cumulative distribution of idMS/MS **(a)** and gene **(b)** cross-tissue expression generated for q-value (Q) <0.01 obtained from R qvalue package (6). The analysis is based on the Anscombe test for kurtosis using the Anscombe.test function in the R "moments" package as described in Li et al. (7) and in the Method section. The x axis reports on the Kurtosis reduction when a certain Z threshold (ZMAD-normalized expression values) was applied, the y axis reports on the corresponding cumulative density. The insert panels correspond to the Kurtosis reduction percentage of idMS/MSs or genes when a given Z threshold was selected. Different colors denote for different Z thresholds and the corresponding cumulative curves. Z=2 was selected as the threshold to extract idMS/MSs with leptokurtic behaviors and Z=3 for genes.

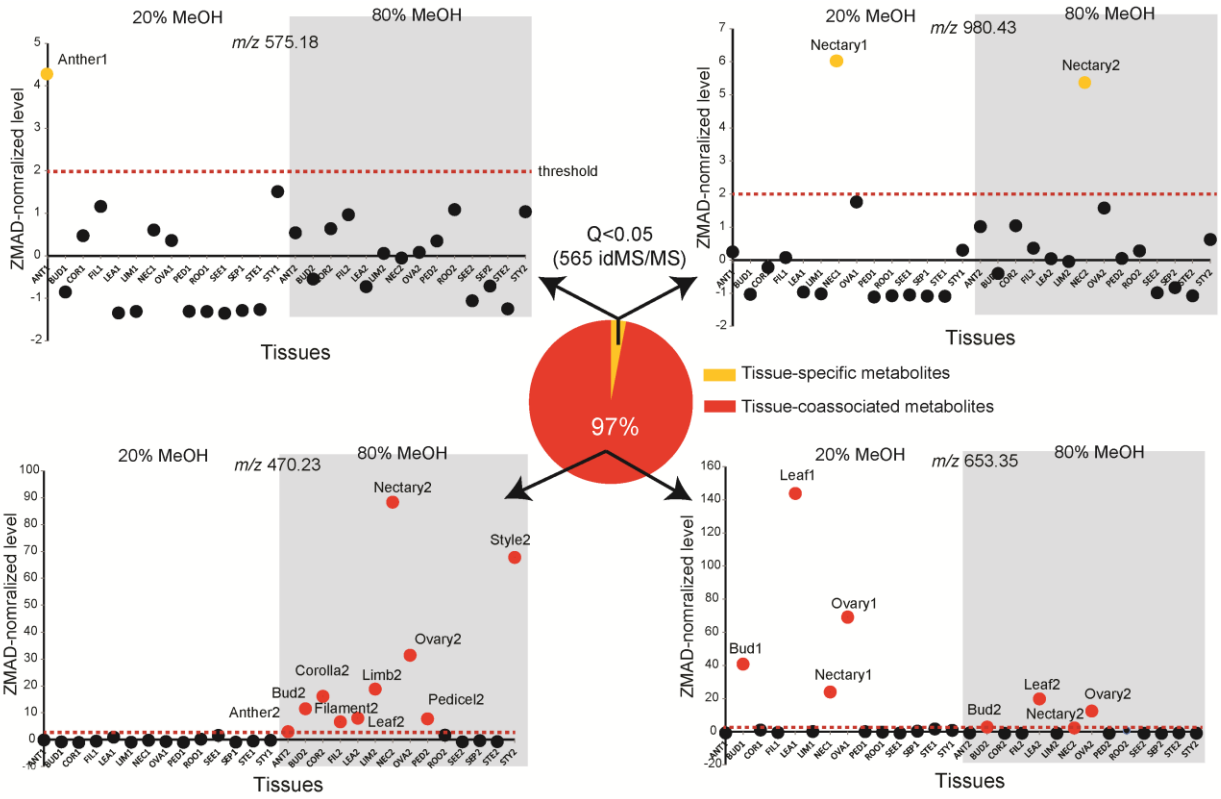

Figure S7. Different degrees of idMS/MS tissue-specificity inferred from the Kurtosis reduction analysis. The kurtosis analysis was used to discriminate leptokurtic idMS/MSs from those distributed in all tissues and the Z threshold of 2 that obtained from the reduction of Kurtosis analysis enabled the detection of single-tissue-specific and tissue-co-associated metabolites. The approach not only extracts single-tissue-specific idMS/MSs (3% of the 565 idMS/MSs with a Q<0.05 for the Kurtosis analysis; two upper panel as examples) but also those preferentially accumulating in several tissues (97% of idMS/MSs; two lower panels as examples). 1 and 2 indices after tissue names refer to the 20 % and 80 % methanol (vol/vol) extractions respectively.

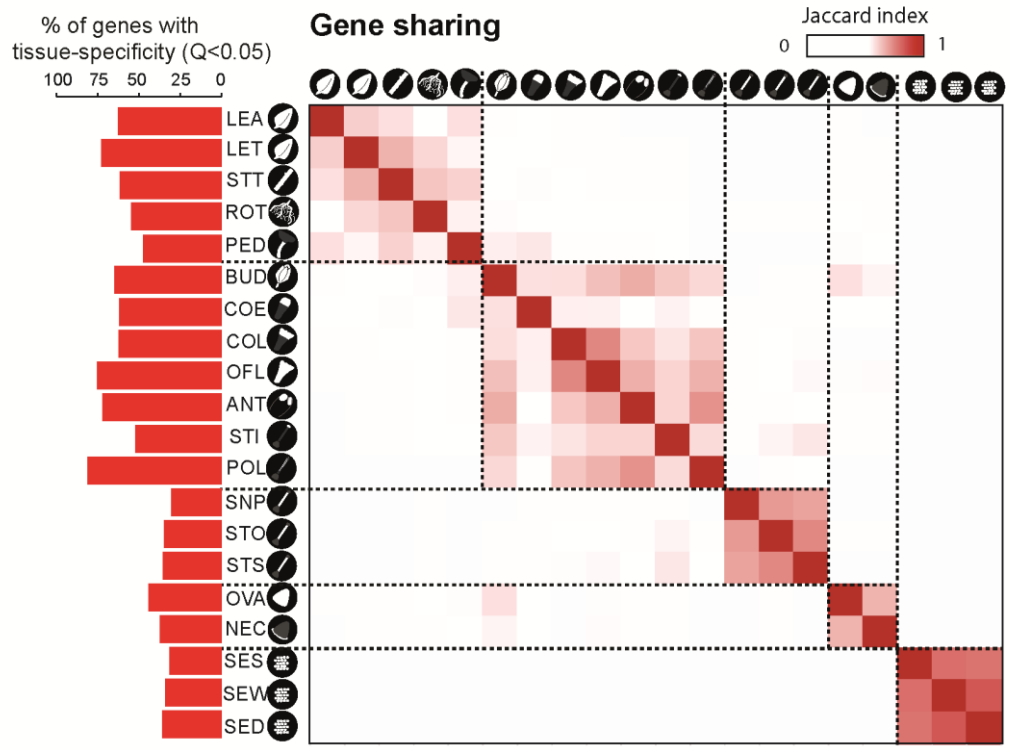

Figure S8. Gene-tissue specificity and gene-sharing among tissues of the transcriptome data-set. Left bar chart, % of genes showing tissue-specificity (Q<0.05) per tissue. Right Heatmap matrix visualizes genes sharing between tissues as measured using Jaccard index. LEA, rosette stage leaves; LET, rosette stage leaves treated with *Manduca sexta* oral secretion (OS); STT, stem from plants with leaves treated with OS; ROT, root from plants with leaves treated with OS; PED, pedicels; BUD, flower buds; COE, non-matured corolla collected 3 days after protrusion from the calyx; COL, matured corolla collected 5 days after protrusion from the corolla; OFL, open flowers; ANT, anthers; STI, stigma; POL, pollen tubes; SNP, style without pollination; STO, style outcross-pollinated (2h after pollination); STS, style self-pollinated (2h after pollination); OVA, ovary; NEC, nectary; SES, seeds treated with liquid smoke; SEW, watered seeds; SED, matured seeds. Detailed overview of RNAseq dataset is available in ***SI Appendix,* Table S2.**

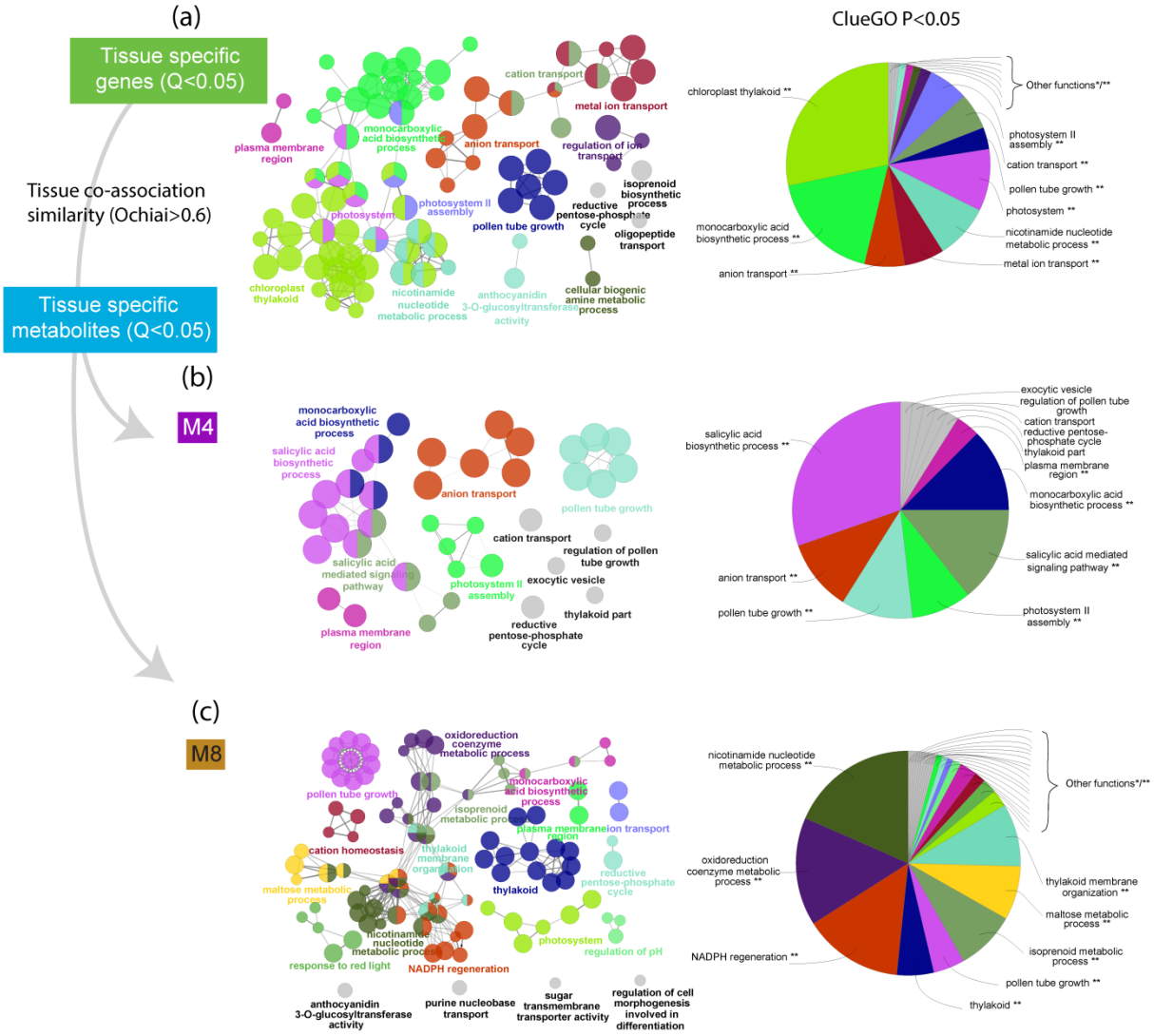

Figure S9. GO enrichment analysis for the pool of genes exhibiting tissue-specificity and co-tissue association with idMS/MSs classified as parts of modules M4 and M8. GO enrichment was generated via GlueGO (8). The significant GO enrichment was depicted in overview pie charts as well as functionally grouped networks with terms as nodes and edges linked based on their kappa score level (>0.4), where only the label of the most significant term per group is shown. Node size is in proportion to the term enrichment significance and functionally related groups partially overlap. **(a)** Significant GO enrichment obtained for all tissue-specific genes (Q<0.05) from transcriptomic data. **(b)** and **(c)** GO enrichment from transcriptomic data sharing tissue co-associations (Ochiai>0.6) with the tissue-specific metabolites (Q<0.05) in module 4 and 8.

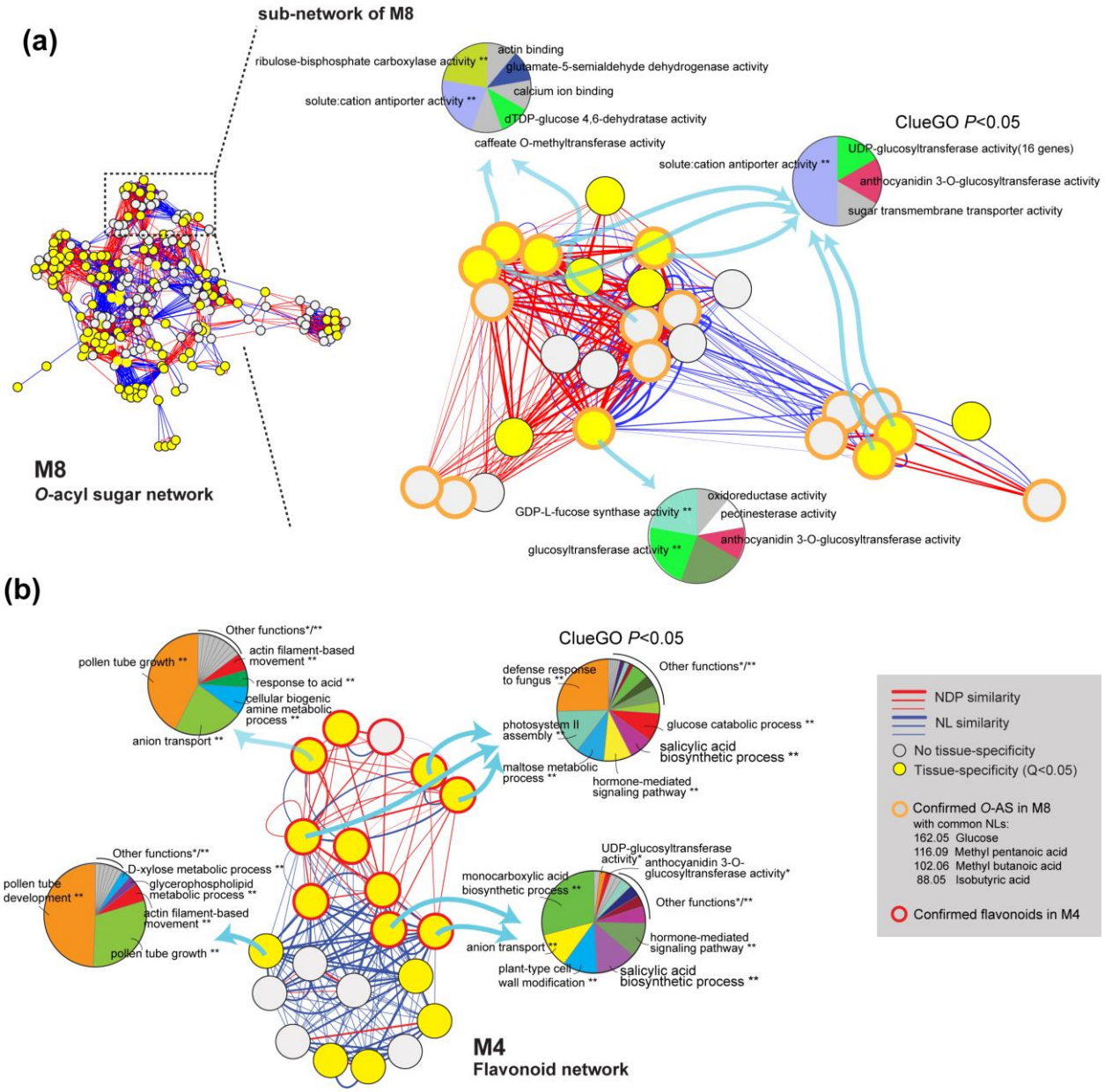

Figure S10. GO terms from transcriptomic data sharing significant tissue co-associations with targeted metabolites of modules 8 and 4. Molecular networks constructed from modules M8 **(a)** and M4 **(b)** are enriched in *O*-acyl sugars and flavonoids respectively. Nodes represent idMS/MS spectra and edges correspond to structural similarities based on the two score types (NDP similarity calculated based on shared fragments between spectra and NL similarity calculated based on shared common neutral losses between spectra). Tissue-specificity is mapped onto the molecular network with node colors. Nodes in the sub-network of module M8 that share typical *O*-acyl sugars neutral losses of glucose, methyl pentanoic acid, methyl butanoic acid and isobutyric acid are additionally circled in apricot. Identified flavonoids in module M4 are circled in red. A zoom-in of the network depicts metabolite-to-gene tissue co-

association calculated as Ochiai similarity. GO terms were generated by GlueGO (8) from transcriptomic data sharing significant tissue co-associations with targeted metabolites.

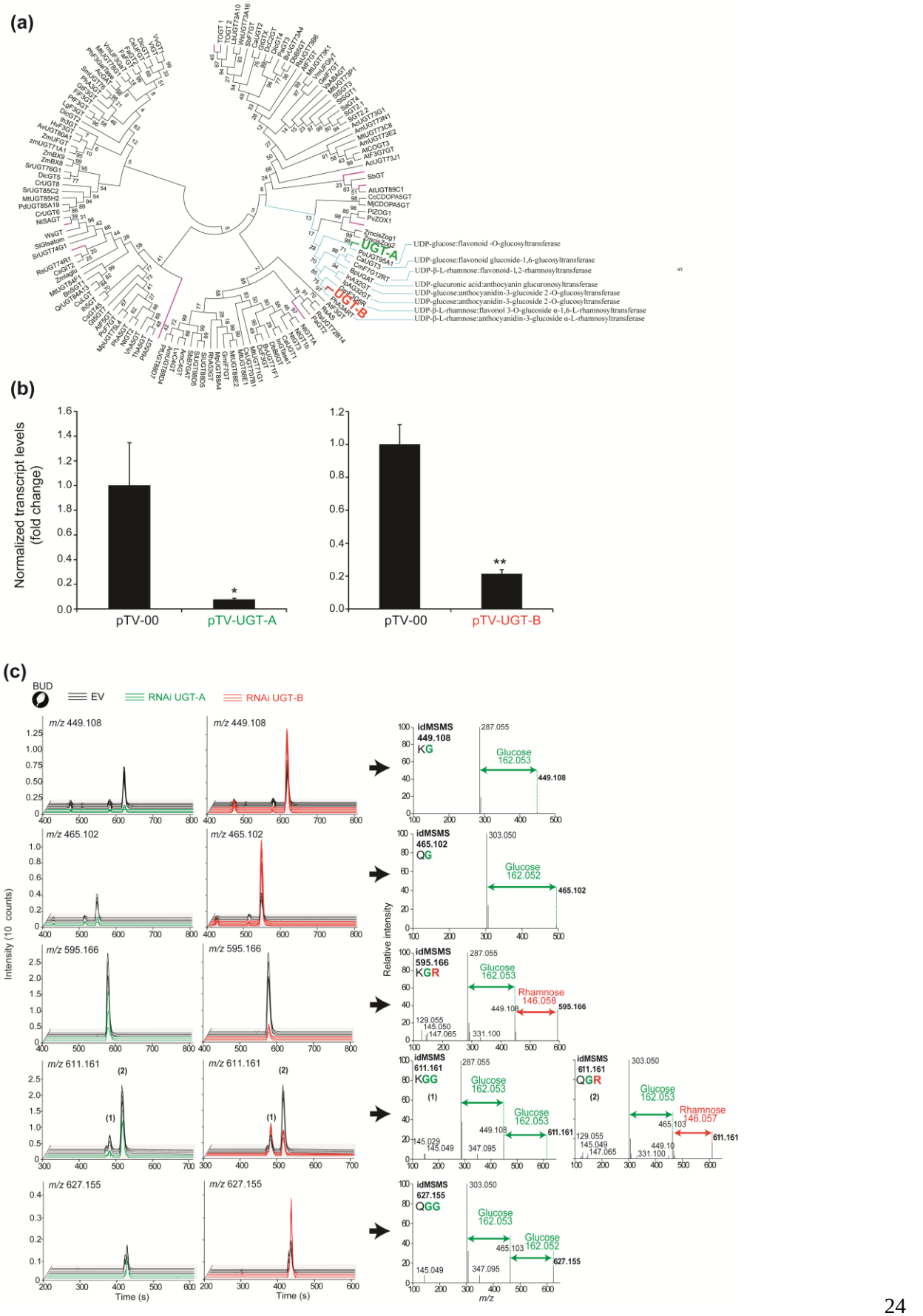
(a)
UGT-A
UGT-B
UDP-glucose:flavonoid -O-glucosyltransferase
UDP-glucose:flavonoid glucoside-1,6-glucosyltransferase
UDP-β-L-rhamnose:flavonoid-1,2-rhamnosyltransferase
UDP-glucuronic acid:anthocyanin glucuronosyltransferase
UDP-glucose:anthocyanidin-3-glucoside 2-O-glucosyltransferase
UDP-glucose:anthocyanidin-3-glucoside 2-O-glucosyltransferase
UDP-β-L-rhamnose:flavonol 3-O-glucoside α-1,6-L-rhamnosyltransferase
UDP-β-L-rhamnose:anthocyanidin-3-glucoside α-L-rhamnosyltransferase
(b)
Normalized transcript levels (fold change)
pTV-00
pTV-UGT-A
pTV-UGT-B
(c)
BUD
EV
RNAi UGT-A
RNAi UGT-B
Intensity (10 counts)
Relative intensity
Time (s)
m/z
idMSMS 449.108 KG
idMSMS 465.102 QG
idMSMS 595.166 KGR
idMSMS 611.161 KGG
idMSMS 611.161 QGR
idMSMS 627.155 QGG
Glucose 162.053
Rhamnose 146.058

Figure S11. Characterization of UGT-A and UGT-B association with flavonoid metabolism. **(a)** The tree was obtained by aligning characterized glycosyltransferases of the GT superfamily 1 (GT 1, 136 GT amino acid sequences) and inferring their phylogenetic relationship using the Maximum Likelihood method (bootstrap = 1000) based on the JTT matrix-based model(9). Evolutionary analyses were conducted in MEGA5 (10). UGT-A (UGT-02515) and UGT-B (UGT-02184) analyzed in the present study are highlighted with purple branches. The red marked branches represent 8 other putative GTs (no names reported) initially considered for virus-induced gene-silencing in *N. attenuata*. Accession numbers for construction of the UDP-Glycosyltransferase tree in different species are list in ***SI Appendix, Materials and Methods.*** **(b)** Gene silencing efficiency for the UDP-glycosyltransferases tested during VIGS experiments. Transcript levels (left panel, *pTV-UGT-A*; right panel, *pTV-UGT-B*) normalized to those of *ELONGATION FACTOR1* were determined in flower buds. Asterisks denote for significant differences between pTV-00 and UGT silenced lines (t-test, **P* < 0.05, ** < 0.01). pTV-00, empty vector; pTV-UGT-A, UDP-glycosyltransferase-A silencing VIGS construct; pTV-UGT-B, UDP-glycosyltransferase-B silencing VIGS construct. **(c)** UHPLC-MS analysis of flower buds of plants inoculated with empty vector and gene silencing constructs for *UGT-A* and *UGT-B*. UHPLC-MS analysis of flower buds of plants inoculated with empty vector and gene silencing constructs for *UGT-A* and *UGT-B.* Chromatograms are traces corresponding to idMS/MS signals with strong co-association with these two UGTs. As supported by the annotation of idMS/MS spectra, silencing *UGT-A* decreases the glucosylation of flavonols while silencing *UGT-B* decreases their additional rhamnosylation (**Figure 5c**). Asterisks denote significant differences between empty vector (EV) and UGT silenced lines (t-test, **P* < 0.05, ** < 0.01). KG, kaempferol-3-*O*-glucoside; KGG, kaempferol-3-*O*-sophoroside (glucosyl(1-2)glucoside); KGR, kaempferol-3-*O*-rutinoside (glucosyl(1-2)rhamnoside); QG, quercetin-3-*O*-glucoside; QGG, quercetin-3-*O*-sophoroside (glucosyl(1-2)glucoside); QGR (Rutin), kaempferol-3-*O*-rutinoside (glucosyl(1-2)rhamnoside).

References

1. Martinez O & Reyes-Valdes MH (2008) Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *P Natl Acad Sci USA* 105(28):9709-9714.
2. Tesson BM, Breitling R, & Jansen RC (2010) DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC bioinformatics* 11.
3. Saedler R & Baldwin IT (2004) Virus-induced gene silencing of jasmonate-induced direct defences, nicotine and trypsin proteinase-inhibitors in *Nicotiana attenuata*. *J Exp Bot* 55(395):151-157.
4. Gaquerel E, Heiling S, Schoettner M, Zurek G, & Baldwin IT (2010) Development and validation of a liquid chromatography-electrospray ionization-time-of-flight mass

spectrometry method for induced changes in *Nicotiana attenuata* leaves during simulated herbivory. *Journal of agricultural and food chemistry* 58(17):9418-9427.
5. Heiling S, Khanal, S., Barsch, A., Zurek, G., Baldwin, I. T., Gaquerel, E. (2016) Using the knowns to discover the unknowns: MS-based dereplication uncovers structural diversity in 17-hydroxygeranyllinalool diterpene glycoside defense production in the Solanaceae. *The Plant Journal* 85(4):561-577.
6. Storey JD (2002) A direct approach to false discovery rates. *J Roy Stat Soc B* 64:479-498.
7. Li S*, et al.* (2012) Gene-sharing networks reveal organizing principles of transcriptomes in *Arabidopsis* and other multicellular organisms. *The Plant cell* 24(4):1362-1378.
8. Bindea G*, et al.* (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25(8):1091-1093.
9. Jones DT, Taylor WR, & Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS* 8(3):275-282.
10. Tamura K*, et al.* (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731-2739.