ANNUAL REVIEWS

Annual Review of Plant Biology

Next-Generation Mass Spectrometry Metabolomics Revives the Functional Analysis of Plant Metabolic Diversity

Dapeng Li[1] and Emmanuel Gaquerel[2]

[1]Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, 07745 Jena, Germany; email: dli@ice.mpg.de

[2]Institut de Biologie Moléculaire des Plantes du CNRS, Université de Strasbourg, 67084 Strasbourg, France; email: emmanuel.gaquerel@ibmp-cnrs.unistra.fr

Annu. Rev. Plant Biol. 2021. 72:867–91

First published as a Review in Advance on March 29, 2021

The *Annual Review of Plant Biology* is online at plant.annualreviews.org

https://doi.org/10.1146/annurev-arplant-071720-114836

Keywords

metabolomics, mass spectrometry, plant specialized metabolism, plant-insect interactions, plant defense theories

Abstract

The remarkable diversity of specialized metabolites produced by plants has inspired several decades of research and nucleated a long list of theories to guide empirical ecological studies. However, analytical constraints and the lack of untargeted processing workflows have long precluded comprehensive metabolite profiling and, consequently, the collection of the critical currencies to test theory predictions for the ecological functions of plant metabolic diversity. Developments in mass spectrometry (MS) metabolomics have revolutionized the large-scale inventory and annotation of chemicals from biospecimens. Hence, the next generation of MS metabolomics propelled by new bioinformatics developments provides a long-awaited framework to revisit metabolism-centered ecological questions, much like the advances in next-generation sequencing of the last two decades impacted all research horizons in genomics. Here, we review advances in plant (computational) metabolomics to foster hypothesis formulation from complex metabolome data. Additionally, we reflect on how next-generation metabolomics could reinvigorate the testing of long-standing theories on plant metabolic diversity.

Contents

1. INTRODUCTION

Plants are master synthetic chemists, making use of their metabolic prowess to produce complex blends of structurally diverse metabolites. The plant kingdom contains somewhere on the order of one hundred thousand to one million chemically unique structures (3, 26), with an estimated range of five thousand to fifteen thousand structures per plant species (38, 40). Plant specialized metabolites (PSMs, synonymously referred to as secondary metabolites or natural products), compared with their counterparts integrated in primary metabolic pathways that are broadly conserved across plant lineages, contribute to the majority of the pan-plant metabolome diversity. This diversity of PSMs is a central dimension of the functional traits that propelled plants' colonization of very diverse ecological niches, and the pronounced plasticity in PSM production provides plants with on-demand biochemical capabilities to cope with highly unpredictable fluctuations of biotic and abiotic stress conditions (144).

Metabolome: comprises the complete set of small molecules presented in a biological system of a given state

From a biosynthetic standpoint, the production of structurally diverse PSMs is supported by large metabolic gene families such as those of the cytochrome P450s, uridine diphosphate (UDP)-glycosyltransferases, and BAHD acyltransferases (25). Tailoring or decorating enzymes encoded by these large gene families can modify PSM scaffolds with various chemical groups that exponentially increase PSM structural diversity and often exhibit relatively low substrate specificity compared with enzymes in primary metabolism, which tend to be more restrictive (141). Varying degrees of enzyme promiscuity, i.e., the "coincidental catalysis of reactions other than the reaction(s) for which an enzyme evolved" (67, p. 473), are hence viewed as key mechanistic bases for the vast reaction space characteristic of PSM pathways. Gene duplication, via

tandem, segmental, and whole-genome duplications, and subsequent divergence in expression and amino acid sequences of the duplicated genes, is the central evolutionary motor to metabolic diversification (71, 72, 142). The expansion of gene families resulting from the retention of duplicated genes supported by neofunctionalization—selective retention by acquisition of a new beneficial function—and subfunctionalization—partitioning of ancestral functions between gene duplicates (88)—is indeed widely recognized as a major driving force of metabolic innovations over short evolutionary time scales. Genetic mechanisms for gene duplications, recruitments, and the evolution of new metabolic activities in PSM pathways have been recently discussed in several excellent reviews (93, 141).

Metabolite profiling: provides direct functional information by targeting a set of metabolites associated with specific metabolic pathways or compound classes in a biological system

1.1. Hierarchical Levels of Plant Specialized Metabolite Diversity

The compositional diversity of PSM profiles or phytochemical diversity can be examined at different hierarchical levels. A first hierarchical level of analysis considers the micro- and macroevolutionary diversification of PSM pathways among individuals of plant populations (i.e., intraspecific diversity) and among closely related species up to more evolutionarily distant plant species (i.e., interspecific diversity). For the latter comparison, researchers have noted that particular PSMs or complete metabolic classes are sparsely distributed among plant lineages. Consequently, certain PSM classes have been frequently targeted as idiosyncratic metabolic characters for the biochemical investigation of specific plant families, for instance, quinolizidine alkaloids and nonprotein amino acids for Fabaceae (145), tropane and steroidal alkaloids for Solanaceae (58), and iridoids and essential oils for Lamiaceae (136). This taxonomically restricted distribution of several PSM classes reflects that particular taxa-specific metabolic pathways have been preferentially selected during evolution when their final metabolic products provide fitness benefits in the ecological niches inhabited by plants of these taxonomic groups. For instance, antiherbivore pyrrolizidine alkaloid biosynthetic pathways have undergone repeated loss and gain transitions during evolution, resulting in patchy metabolic distribution among closely related species (101). At the population level, within- and among-population compositional heterogeneities in PSMs have been reported in a large body of targeted phytochemical studies (2, 71), more rarely using broadly targeted metabolite profiling (148). These population-level polymorphisms in PSMs most frequently appear in the form of quantitative variations in metabolite concentrations and their maintenance has been hypothesized to contribute to overall-population fitness in fluctuating and geographically dispersed environments (95) as well as in the context of rapid changes in herbivore regimes (150).

A second hierarchical level of exploration considers quantitative and qualitative metabolic variations within a plant that occur among different organ or tissue types as a result of the confined, and, in the most extreme cases, cell-type-specific, localization of the underlying biosynthetic pathways (59). Intertwined with the spatial heterogeneities of PSMs are temporal adjustments in PSM production according to ontogenic stages or circadian or annual cycles and during stress acclimation (55, 69). Modulations of PSM diversity detected at the intraspecies scale are further exacerbated as part of natural polymorphisms in induced signaling and responses to biotic stresses, in particular those resulting from interactions with herbivorous insects (61). The underlying time- and ecological context–dependent specialized metabolism plasticity contributes to the broad set of reconfigurations of a plant's defensive physiology when challenged by herbivores, pathogens, or competitors, and, when enemies are absent, it is thought to avoid the excessive allocation of valuable resources to chemical defenses at the expense of maintaining physiological investments (13) as well as collateral damages to mutualistic interactions (123). The evidence for such materialized trade-offs between constitutive and induced metabolic diversity in the context of insect herbivory is, however, equivocal (64).

Metabolomics: the comprehensive and systematic analysis (both qualitatively and quantitatively) of all measurable metabolites in a biological sample using high-throughput analytical means

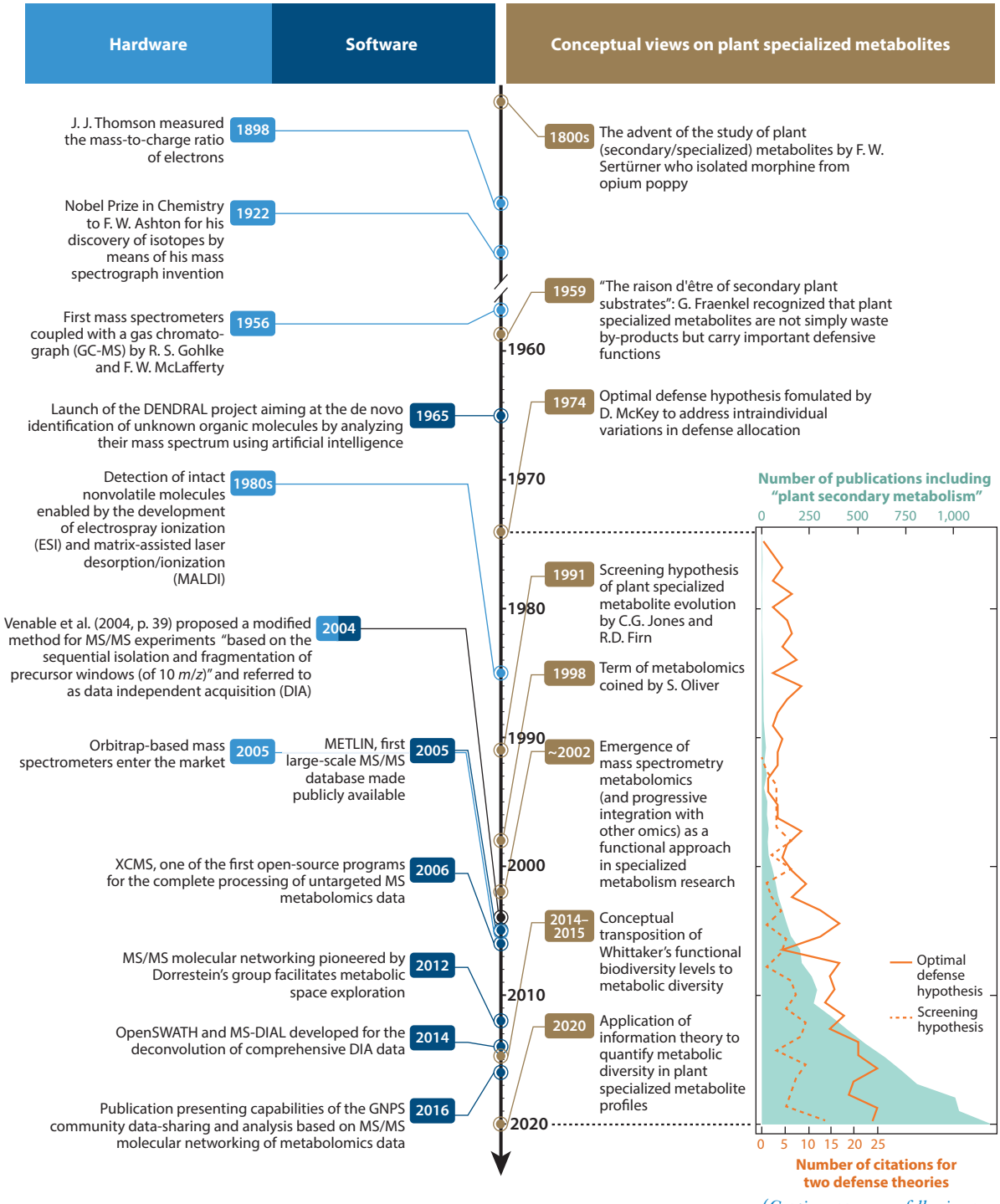
Chemical ecology: an interdisciplinary field between chemistry and biology that primarily focuses on the small molecules mediating interactions between living organisms and their environment

Beyond plant biology research, exploring metabolic diversity is relevant to all important theoretical advances in the study of organisms' adaptations to their environments and networks of organismic interactions in ecosystems. In ecology, three functional levels of biodiversity (i.e., species) over spatial scales are traditionally considered (143): α diversity, referring to the diversity of local species within a particular geographic area; β diversity, signifying the ratio between regional and local species; and γ diversity, describing the total species diversity in a landscape. These functional levels of biodiversity have recently been transposed to phytochemical diversity (66, 95). According to the above-described hierarchical levels of analysis, functional categorizations could translate in α metabolic diversity as the diversity of PSMs in a given plant and β metabolic diversity as the fundamental intraspecific PSM space for a (set of) population(s), while γ metabolic diversity would be the extension of the diversity analysis to a multispecies perspective. We advocate (see Section 3) that the comprehensive metabolic data necessary to assess these different levels of PSM diversity, among other variables of interest, require taking full benefit of recent advances in mass spectrometry (MS) metabolomics for large-scale data acquisition. Additionally, the implementation of adequate statistical approaches to score metabolic diversity from these data into simple currencies is another critical step. In Section 4, we formalize this view and discuss how predictions of seminal theories in the chemical ecology of PSMs could be revisited when they are posed at this new scale of analysis and when statistical currencies derived from the widely used information theory framework are implemented to compare metabolomes.

1.2. A Brief Historical Perspective on Parallel Developments in the Chemical Ecology and MS Analysis of Plant Specialized Metabolites

The advent of the study of PSMs dates back to 1806 when Friedrich Wilhelm Sertürner isolated the "salt of opium," the first alkaloid morphine, from opium poppy *principium somniferum* (50). For about 150 years, documentation of the vast small-molecule structural space in plant extracts, leading to the discovery of diverse compound classes, was mostly viewed as a strong argument for the incidental origin of PSMs. The diversity of PSMs was seen as "flotsam and jetsam on the metabolic beach," as waste or inner end products from central metabolism degradation pathways (51, p. 243), and referred to as serving no primary function in a plant's growth and reproduction (50). Interestingly, the research of entomologists in the middle of the twentieth century led to a remerging functional interpretation of PSMs in light of plants' interactions with insects. In his seminal article published in 1959, Gottfried Fraenkel recognized that PSMs were not the inert waste products of a plant's metabolic exuberance but carried intrinsic defenses against and host selection functions for insects (44). Core to Ehrlich & Raven's plant-insect coevolution theory formulated in 1964 was the prediction that variations in metabolic traits constrained, as part of an arms race, rates of lineage diversification in both producing plants and the insects they interact with (33). These paradigm shifts that emphasized the interorganismic functions of PSMs provided guidance to the formal establishment of (plant) chemical ecology in the 1970s. Additionally, bioassays on PSMs benefited, starting from the middle of the twentieth century, from impressive advances in the area of organic synthetic chemistry methods (32), including approaches enabled by the widespread application of Diels–Alder cyclization and retrosynthetic analysis by pioneers such as Corey and Woodward (discussed in 97).

Early plant chemical ecology studies were conducted with a focus on a few compounds and did not systematically absorb the merits from contemporary technical innovations in MS (**Figure 1**). In the 1980s, the development of two soft ionization techniques, electrospray ionization (ESI) (146) and matrix-assisted laser desorption/ionization (MALDI) (125) enabled for the first time the direct detection of intact nonvolatile molecules, thereby significantly expanding the detectable



(Caption appears on following page)

Figure 1 (*Figure appears on preceding page*)

Timeline of methodological breakthroughs in mass spectrometry and concepts to support plant specialized metabolism exploration. The left panel visualizes important technical advances in the implementation of broader-scale mass spectrometry–based plant metabolite analysis [e.g., the method proposed by Venable et al. (135)]. The right panel highlights selected key concepts and studies associated with a holistic and functional view of plant specialized metabolite diversity. The graph depicts bibliographic trends for published research articles on plant secondary and specialized metabolites (*teal area*) and citations for two key plant defense theories (the optimal defense hypothesis and screening hypothesis) addressing plant metabolic diversity (*orange lines*). Bibliographic information from the Web of Science online database.

chemical space beyond small volatile molecules analyzable by gas chromatography coupled with mass spectrometry (GC-MS) (48). The large-scale commercialization of these two ionization techniques made MS the dominant analytical technique to conduct more broadly targeted metabolite profiling. The term metabolomics was first coined by Steven Oliver et al. in an article published in 1998 (99), followed by several research articles adopting nontargeted MS-based metabolite profiling methods for the study of plant metabolism (42, 45, 107), leading to the emergence of metabolomics as a vigorous technical research field. Early plant metabolomics analyses were largely exploratory and relied either on GC-MS, liquid chromatography-MS (LC-MS), or direct-injection MS to profile, with relatively low throughput, genetically modified or environmentally challenged plants for metabolite fingerprints established from tens to hundreds of metabolites detected in a few samples (5, 42, 53, 107, 108, 137). In the past decade, steady improvements in the scanning speed and resolution of mass analyzers (measuring mass-to-charge [*m/z*] ratios based on differential ion behaviors in electric, magnetic, and electromagnetic fields) as well as in the sensitivity and sample throughput of MS instruments have revolutionized this research field. While a concomitant decrease in costs has not been observed for MS instruments equipped with the highest-resolution MS analyzers [Fourier-transform ion cyclotron resonance technology (FT-ICR)], medium-high-resolution quadrupole time-of-flight mass analyzer (QTOF) MS instruments for routing metabolomics work have continuously gained market share due to their relative affordability for single laboratories (6). Together, these hardware developments have enabled a fundamental transition, progressively adopted in a wide range of biological studies, from a reductionist to a more global analysis of small-molecule profiles. In chemical ecology, the relatively recent application of the latest advances in untargeted MS metabolomics has dramatically improved the efficiency of detection of PSMs mediating key chemical interactions between organisms and their environment, thereby overcoming one of the main limitations of traditional bioassay-guided structure elucidation workflows.

2. CORNERSTONE CHALLENGES IN THE UNTARGETED MS ANALYSIS OF PLANT SPECIALIZED METABOLITE DIVERSITY

Compared with other omics approaches, metabolomics offers direct and real-time readouts of the small-molecule determinants of an organism's physiology and ecological interactions (82). Major technical obstacles to metabolomics analyses are directly associated with the aforementioned highly diverse chemistry as well as the several orders of concentrations and complex spatiotemporal dynamics associated with small molecules. These challenges combined with the inherent specificities of each biological matrix render technically unfeasible the profiling of the complete metabolome of a given cell type, tissue, organ, or organism and thwart the routinization of untargeted metabolomics studies (39). DNA and RNA consist of only four bases that make up the backbone of the strands connected with phosphodiester bonds, whereas proteins are linear polymers consisting of long chains of amino acid residues. Although the chemical complexity of proteins expands to 20 amino acid types, creating the extremely vast chemical space in which proteins function, inferences on protein sequences can be derived from genome information according to

the central dogma of molecular biology. In clear contrast, unlike the polymeric nature of nucleic acids and proteins, precursors (building blocks) for the biosynthesis of PSMs are highly diverse. PSM scaffolds are not assembled in a template-based modular fashion and are further chemically modified and decorated, leading to significant heterogeneity in molecular composition, polarity, and stability, all of which render comprehensive chemical analyses challenging even for extracts obtained through multistep enrichments for specific compound classes. Obviously, nucleic acid and protein sequences may not be used to directly infer physicochemical properties of PSMs. But, conversely, metabolomics analysis can advantageously be applied as a stand-alone technique to nonmodel plant species for which genome and transcriptome data are not yet available. Rather than reviewing the entire analytical and processing method portfolio for MS metabolomics, we focus here on critical steps that affect the diversity of explored metabolites.

2.1. Illuminating the Dark Matter in MS Metabolomics Studies

All combinations of metabolomics methods per se are technically biased toward very small fractions of the true metabolome of plant specimens. Reciprocally, the dark metabolome represents all the metabolites present in a system that are either not extracted, lost and/or transformed during the extraction, or not detected using standard analytical methods and hence remain unknown or detected metabolites that cannot be structurally annotated and hence are considered as known unknowns (23). Recent innovations in MS technologies that allow real-time metabolome analysis of unicellular microorganisms (57, 84) and complete living multicellular organisms (46) have significantly expanded our capacity to inventory part of this dark metabolite space. Among the many frontiers in this research field, an important obstacle is to resolve the spatial heterogeneities in metabolites at cellular or subcellular levels that, in turn, result from heterogeneities in enzymatic activities and dynamic metabolon assemblies at these spatial scales (100). More generally, various technical factors, such as sample extraction, liquid chromatography separation, MS acquisition, detection, and metabolic feature extraction during postprocessing of the data (described in Section 2.2), influence qualitatively and quantitatively metabolite annotation outputs (147). Ultimately, metabolites measurable by MS metabolomics are frequently estimated to be less than 5% of the organism's metabolic space (131).

Optimization of extraction protocols that consider different solvent combinations constitutes an instrumental, but strikingly poorly exploited, means of achieving both efficiency and reproducibility with a better coverage of PSM diversity (49). A key remaining issue is to increase the sensitivity in the detection of metabolites with low abundance or poor ionization behavior. To this end, different metabolomics strategies using multiple orthogonal analysis platforms such as LC-MS, GC-MS, capillary electrophoresis-MS (CE-MS), and nuclear magnetic resonance (NMR), which are complementary techniques in terms of their expertise in the detection and characterization of metabolites of different physiochemical properties, are frequently combined to achieve improved comprehensiveness (8, 76, 78). In this respect, different MS ionization techniques including the commonly used ESI, atmospheric pressure chemical ionization (APCI) and atmospheric pressure photoionization (APPI) offer overlapping but also specific ranges of applicability dependent on metabolites' molecular weight and polarity (147). In addition, positive and negative ionization mode measurements are routinely conducted in untargeted metabolomics as complementarily parallel runs, with each of the two ionization modes detecting significant numbers of unique analytes that can be combined through dual processing of ionization mode–specific data.

2.2. Anatomy of Key Processing Steps of MS Metabolomics Data

A typical processing pipeline for untargeted high-resolution MS (HR-MS) metabolomics data is depicted in **Figure 2**. After data acquisition, an initial step to the computational processing of

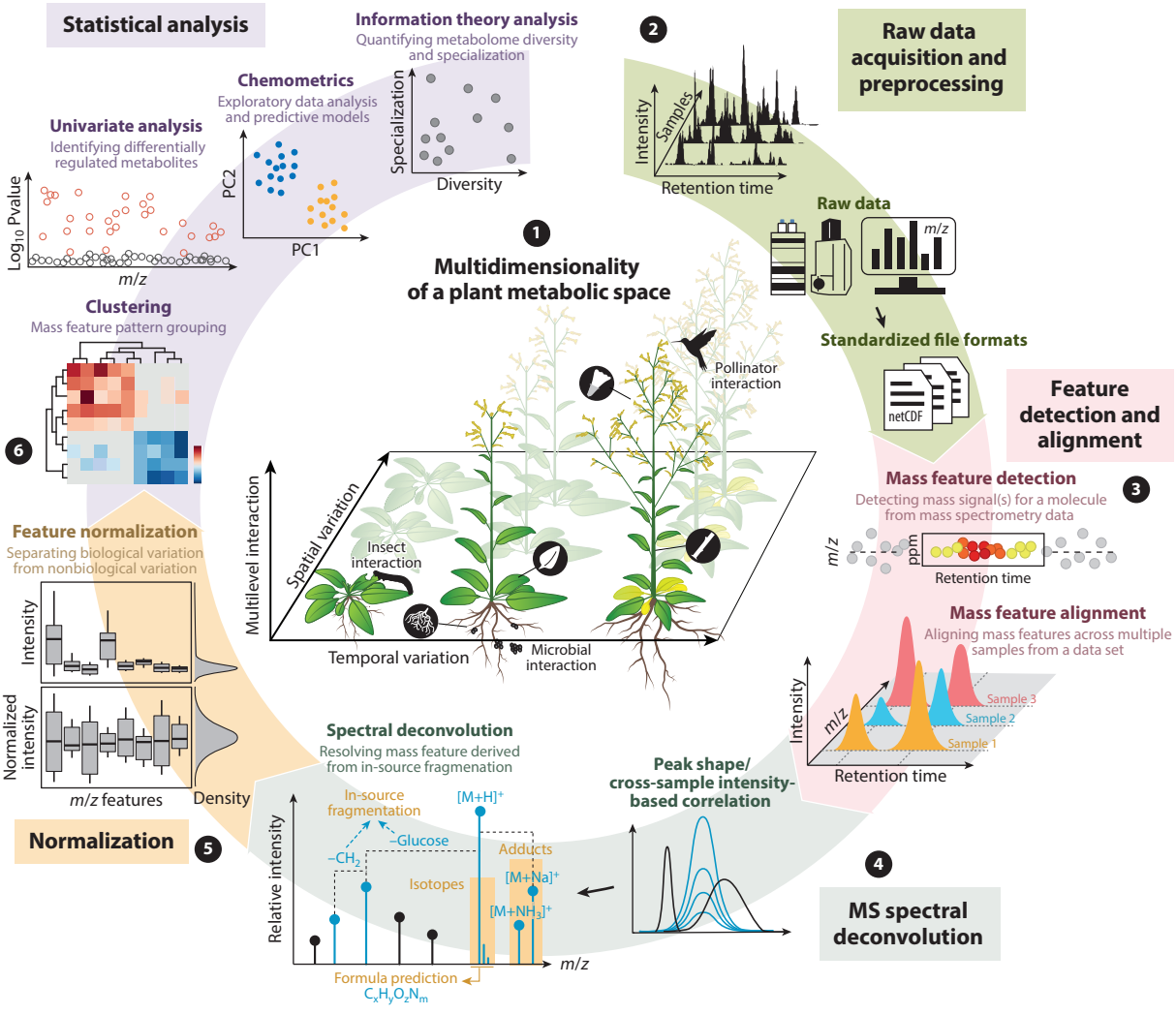


Figure 2

A schematic workflow for high-resolution mass spectrometry untargeted data processing and simple statistical mining. (❶) Spatiotemporal heterogeneities in a plant's metabolic space are detectable at different hierarchical scales and for different ecological interactions. Experimental design and plant specimen selections are to be critically considered according to explored levels of metabolic diversity. (❷) After metabolite extraction and measurement by LC coupled to HR-MS, vendor-specific raw data are converted to universal file formats for subsequent automated processing by a wide range of open-source software. The latter have been benchmarked in recent reviews (63, 124). (❸) Automated processing aims to unfold multidimensional HR-MS data into a mass matrix consistent in size for the data set and used as input for statistical mining. Automated processing steps critically include mass feature detection and alignment among samples as well as the (❹) MS spectral deconvolution by clustering of mass features belonging to a single metabolite and generated during in-source fragmentation. (❺) Feature normalization and missing value imputation are important for the next steps of statistical analysis. (❻) Multilevel statistical analyses are typically conducted for the interpretation of global metabolic trends and identification of biomarker metabolic features associated with those trends. More recently, statistical descriptors from the information theoretical framework were transposed to score indices of diversity and specialization from metabolome profiles, thereby allowing the quantification of the reprogramming of metabolome diversity according to ecological interactions (80). Abbreviations: HR, high-resolution; LC, liquid chromatography, MS, mass spectrometry, *m/z*, mass-to-charge ratio.

HR-MS data using noncommercial platforms consists of the conversion of the vendor-specific data format to the NetCDF, mzML or mzXML standardized file formats so that the HR-MS raw data can be read by the many widely used software tools, such as XCMS (121), MZmine (62), MetAlign (86), OpenMS (109), MS-DIAL (130), and others. HR-MS data are three-dimensional data sets composed of distributions of *m/z* signals and their retention times and corresponding intensities. Hence, data preprocessing solutions are ultimately required to unfold these multidimensional data sets and facilitate subsequent data interpretation. Computational solutions typically involve feature detection and alignment across multiple samples, noise filtering, and missing feature imputation in order to generate a concatenated matrix in which mass features associate to a unique *m/z* and a specific retention time. Rationales and algorithmic bases to built-in feature detection and preprocessing methods have been extensively reviewed (63, 124).

The extracted matrix consists of an uncharacterized pool of mass features amassed from the ionization of (plant) metabolites that correspond to different adduct types, isotopic ions, and in-source fragments extracted by data processing. Hence, an important step in untargeted HR-MS data processing consists of clustering mass features belonging to the same metabolite into so-called MS pseudospectra (MS1) using open-source software tools such as CAMERA (74), RAMClust (16), MSClust (128), xMSannotator (133), or AStream (10) with the aim of reducing data redundancy and facilitating detectable metabolite number estimations and further annotation steps. Along this line, the use of in-source MS fragments (from MS pseudospectra) to interrogate low-energy tandem mass spectrometry (MS/MS) spectra in public databases is often recognized as an underexploited approach that enhances metabolite annotation in mass feature lists from metabolomics studies (16, 89).

Mass feature: corresponds to an *m/z* signal, detected by automated processing of LC-MS data, for a compound being eluted at a particular retention time from the chromatographic separation

Tandem mass spectrometry (MS/MS): also referred to as MS2; an instrumental method to break down ions into fragments via the tandem coupling of mass analyzers

Deconvolution: the computational process of resolving coeluting mass features and creating compound-specific spectra

Dereplication: the identification of metabolites in an empirical study using existing knowledge of the known chemical structures of these metabolites, thus avoiding repeatedly characterizing the structures of known metabolites

2.3. Challenges of MS Metabolomics Data Annotation via Database Interrogation

Processing of HR-MS data typically translates into tens of thousands of mass features that correspond to up to several thousand metabolites, the majority of which remain unknown. Consequently, there is inherently an order of magnitude of difference between known and unknown features. Hence, a formidable challenge is to prioritize and confidently annotate mass features extracted during processing steps and possibly combine in-source fragmentation-derived features as MS pseudospectra. In GC-MS, automatic mass spectral deconvolution and straightforward spectral identification are routinely applied, owing to the highly reproducible ionization process of electron impact ionization (EI), historically standardized at 70 eV, together with the robustness of capillary columns, thus creating highly reproducible mass spectral and relative retention time information (41). Standardized spectral libraries such as that of the National Institute of Standards and Technology (NIST) (12) and the Golm Metabolome Database (73) further allow matching of mass spectral records with experimental data, and the many automated tools and multivariate computational approaches for high-throughput spectral extraction (27) have greatly streamlined the spectral annotation process in GC-MS-based metabolomics analysis.

Neither of these developments is readily applicable in LC-MS-based metabolomics, notably because mass spectra obtained from LC-MS are highly instrument- and condition-dependent and, probably more problematic, due to the scarcity of PSM spectral representation and lack of chemical knowledge capture in public databases (6). In model species for which specific databases are available, such as via the Platform for RIKEN Metabolomics (PRIMe) resources for *Arabidopsis* (110) or other well-studied plant species for which in-house libraries have been created, dereplication of metabolite identification knowledge (representing only a small fraction of the plant's PSM diversity) can be relatively straightforward via the matching of parent and fragment masses

to those of database entries. Yet, for most other species, such a specific knowledge base is unavailable. While the unambiguous assignment of a molecular formula can remain a challenging task for high-molecular-weight metabolites, MS analyzers such as Orbitrap and time-of-flight (TOF) are now capable of reaching 1 ppm (part per million) or sub-ppm mass accuracy levels (90) that drastically reduce the space of chemically plausible formulas for a given mass feature. This process can be further empowered by computational approaches that merge MS data of fully ^{13}C-labeled and unlabeled plant extracts in order to eliminate ambiguity in chemical formula assignment (131).

Global Natural Product Social Molecular Networking (GNPS): a community-based ecosystem for the archiving, sharing, data analysis, and exploration as well as knowledge capture of mass spectrometry, in particular, of MS/MS data

MS/MS (or MS2) experiments and the consecutive interrogation of publicly available spectral databases, such as METLIN (126), MassBank (56), ReSpect (111), WEIZMASS (117), the Global Natural Products Social Molecular Networking (GNPS) community library (138), and others, with target MS/MS spectra are central steps in HR-MS data analysis. Currently, there are approximately 2.4 million MS/MS spectra corresponding to less than 80,000 molecules readily accessible in the most-renowned MS/MS databases, which is likely an overestimate considering the large overlap of chemical entries among these libraries (6). In recent years, in silico (i.e., computer-generated) MS/MS spectral libraries have been employed as promising alternatives to overcome the restricted chemical space coverage of experimental MS/MS databases by generating simulated spectra from chemical databases (7), such as PubChem (68), the Chemical Abstracts Service (CAS) registry (85), and ChemSpider (102), in which large volumes of known chemical structures have been indexed. These computational approaches, as showcased in the Critical Assessment of Small Molecule Identification (CASMI) contest (9, 114), have ramified considerably in the last few years and, when combined with other orthogonal information such as retention time predictions (15) and chemical ontology inference tools (52), can greatly assist in annotation knowledge dereplication. If all of the previously mentioned annotation strategies are exhausted, structural elucidation has to be conducted de novo, with NMR remaining as the gold standard for the structural elucidation from close-to-pure metabolite fractions. Nonetheless, its low sensitivity and low throughput still preclude its application to large-scale exploratory MS metabolomics. Coming back to MS analysis, an early attempt at de novo interpretation of MS data was the DENDRAL project back in the 1960s, which applied artificial intelligence to identify unknown compounds by analyzing their mass spectra and using knowledge of chemistry (83). Recently, a computational approach has been developed to interpret MS/MS data of small molecules by searching molecular structure databases using fragmentation trees and machine learning techniques. This approach, termed CSI:FingerID, represents a powerful means of increasing identification rates of metabolites (30). Recently, a computational method, termed CANOPUS, has been developed to reinforce systematic compound class assignment and ontology prediction using high-resolution fragmentation mass spectra when neither spectral nor structural reference data are available in libraries or databases (29).

3. A NEXT-GENERATION OF MS METABOLOMICS COMBINING MASSIVE AND ALIGNMENT-BASED DATA EXPLORATION

The switch in 2008 from traditional Sanger sequencing to the so-called next-generation sequencing (NGS) techniques has revolutionized genomics research (119). A technological breakthrough of that magnitude has not yet occurred in MS metabolomics; instead, this field has witnessed more incremental technological advances in instrumental sensitivity, accuracy, and MS/MS data collection capacity. Nonetheless, profound advances are currently ongoing from simple univariate and/or multivariate statistical mining (11, 19) (to pinpoint phenotype-associated metabolites on which identification efforts are targeted) toward large-data exploratory approaches amortizing chemoinformatics as well as bioinformatics methods frequently inspired from the field of DNA

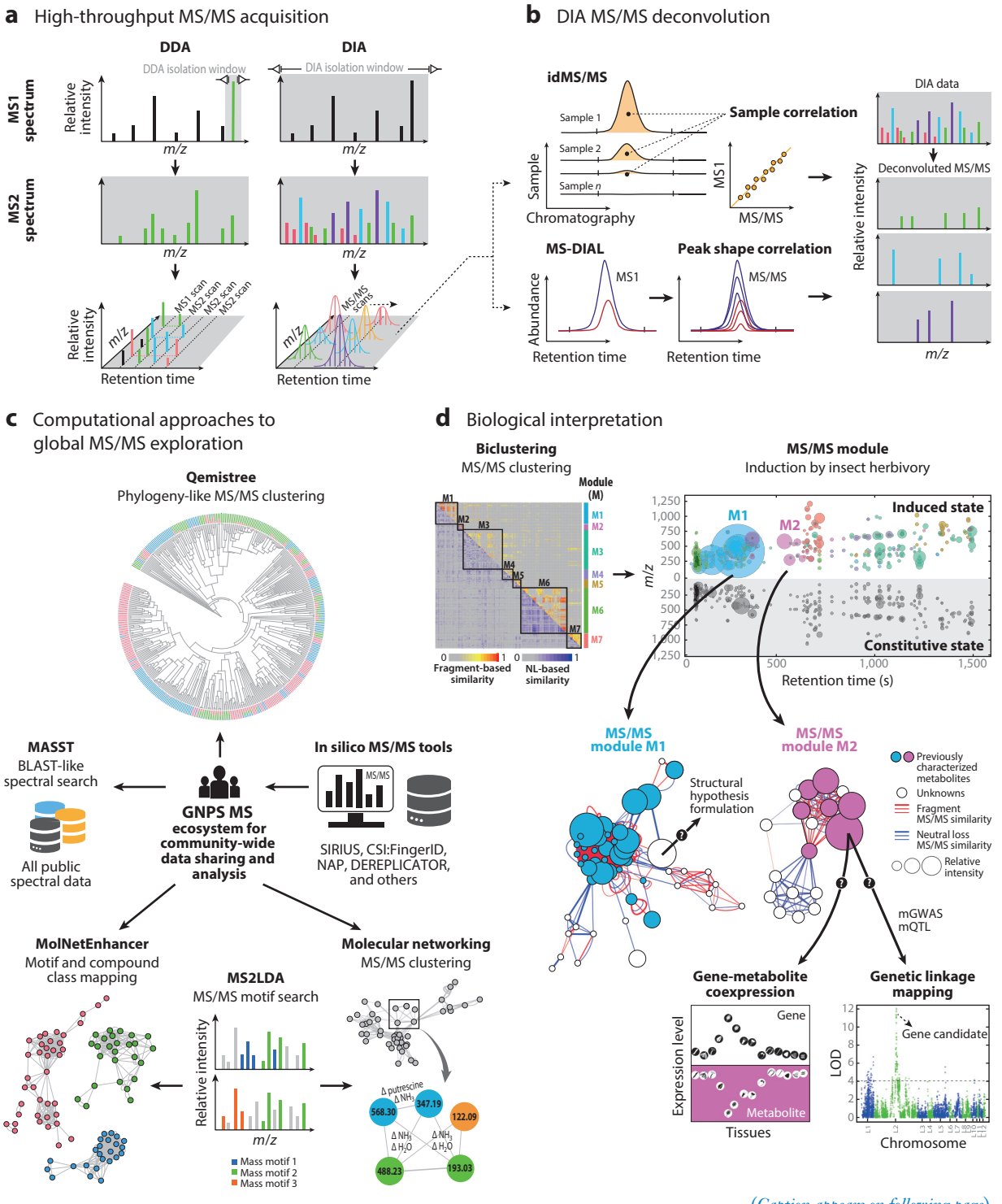
sequencing. Analogous to the large-scale exploration of biodiversity that has flourished with the advent of NGS, recent crowdsourced data and knowledge capture initiatives (138) as well as the transposition of sequence alignment-based classification to MS/MS data (6) are game-changers for large-scale phytochemical diversity exploration.

3.1. Implementing Massive MS/MS Data Acquisition

MS/MS data acquisition as a starting point for compound annotation is traditionally conducted using data-dependent MS/MS acquisition (DDA) methods. In the DDA method, a narrow mass isolation window is selected for which a few of the most intense precursor *m/z* features are targeted for collision-induced dissociation (CID) fragmentation to generate MS/MS spectra (**Figure 3**). In contrast, data-independent MS/MS acquisition (DIA) methods, such as the SWATH method, apply multiple acquisition cycles, and in each one a relatively large mass window (10–25 Da or more) is stepped across the entire mass range without precursor selection; these cycles are repeated over time during the entire chromatographic separation, thus collectively generating MS/MS data for all detectable signals (47). This method, however, requires high scan rate MS instruments to compensate for the short travel times of ions in mass spectrometers in order to cover the whole mass range when smaller isolation windows are applied. Alternatively, DIA methods can be operated indiscriminately by manipulating CID insource ionization fragmentation within an *m/z* range set as large as possible (17). Although traditional DDA is a powerful and versatile strategy, it suffers from several fundamental limitations as compared with DIA methods. First, the precursor isolation width in DDA methods must be set as a compromise between sensitivity and specificity to achieve decent MS/MS signals while avoiding contamination. Second, due to scan rate limitation, only a relatively restricted number of precursor ions can be selected for further CID fragmentation in each DDA acquisition cycle, and the stochastic nature of DDA methods frequently lead to biased acquisition of the same highly abundant metabolites, which reduces the comprehensiveness of the MS/MS analysis. In recent years, DIA methods that allow for a massively parallel collection of structural information on samples' metabolic diversity have therefore received considerable attention. The major disadvantage of DIA is a missing link between precursors and fragments, which requires computational approaches to reconstruct precursor-to-fragment relationships. These characteristics of DIA methods are reminiscent of the situation in NGS in which the computational assembly of short sequencing reads is a technological prerequisite enabling high-throughput sequencing of millions of DNA molecules simultaneously. Several open-source software programs are able to tackle this DIA precursor-to-fragment assembly issue, such as MS-DIAL, which gained momentum for the processing of SWATH data (130) (**Figure 3**). An alternative DIA data analysis approach, termed indiscriminant MS/MS, extracts the resolution required for spectral assemblies from cross-sample variance for the correlation calculations used in precursor-to-fragment assignments (17). Altogether, these DIA data-processing tools allow the deconvolution of large-scale MS/MS libraries, thereby maximizing structural information collection on a given phytochemical profile.

3.2. Alignment-Based Approaches to Navigate Plant Specialized Metabolite Diversity via MS/MS Molecular Networking

The basic local alignment search tool (BLAST) that performs comparisons between pairs of sequences, searching for regions of local similarity, is undoubtedly the most routinely used bioinformatics tool for genomics data mining. A breakthrough development in MS/MS data exploration is the molecular networking approach developed by Dorrestein's group (140) that is based on



(Caption appears on following page)

Figure 3 (*Figure appears on preceding page*)

Computational approaches exploiting large-scale MS/MS capacities to explore plant specialized metabolite diversity. (*a*) Method comparison for DDA and DIA high-throughput MS/MS acquisition. (*b*) DIA-generated MS/MS spectra deconvolution. Information regarding a fragment's assignment to a given precursor mass is lost during DIA acquisition but can be computationally retrieved using spectral deconvolution according to mass feature peak shape, such as in the universal program MS-DIAL, or based on mass feature intensity–based correlation methods exploiting sample-to-sample variance, as in the idMS/MS method. (*c*) A portfolio of MS/MS similarity scoring, clustering, and representation as molecular networks are integrated into the GNPS (**http://gnps.ucsd.edu**) community-wide MS data archiving and analysis platform. Optimized pipelines available in GNPS further allow the mapping of MS/MS molecular network visuals with shared mass motifs inferred from MS/MS data sets (MS2LDA) (134) or annotations of chemical families via a combination of in silico annotation and hierarchical chemical ontologies (MolNetEnhancer) (34), MS/MS clustering via the exploitation of phylogenetics-derived tools (Qemistree) (129), as well as MS/MS queries using MASST (139), which is conceptually similar to the NCBI BLAST search. The chemical clustering from Qemistree, available on iTOL (**https://itol.embl.de/tree/7095134164943815874325 76**), was generated using data from Reference 129, with permission from Pieter Dorrenstein. (*d*) In a closely related analytic approach to molecular networking, MS/MS spectra are clustered into modules in a biclustering approach, according to scores for shared fragments and neutral losses between pairs of MS/MS spectra. Molecular networks reconstructed from modules extracted from the biclustering approach or from the GNPS-based molecular networking are advantageous because they allow the formulation of annotation hypotheses based on previously known metabolites populating a network associated with a phenotypic response of interest [here, the strong response to insect feeding (induced state)] or can be explored based on gene-to-metabolite coexpression and quantitative genetics data to detect candidate genes underlying the production of specific metabolites. Abbreviations: BLAST, basic local alignment search tool; DDA, data-dependent MS/MS acquisition; DIA, data-independent MS/MS acquisition; GNPS, Global Natural Products Social Molecular Networking; idMS/MS, indiscriminant MS/MS; LOD, logarithm of odds; mGWAS, metabolic genome-wide association study; mQTL, metabolic quantitative trait locus; MS, mass spectrometry; MS/MS, tandem mass spectrometry; *m/z*, mass-to-charge ratio; NCBI, National Center for Biotechnology Information; NL, neutral loss.

multiscale pairwise alignments of MS/MS spectra from HR-MS analyses. Molecular networks are visual displays in which nodes represent spectra and edges represent spectrum-to-spectrum alignments (with similar fragmentation implying similar structure) (140). This approach is extremely useful for PSM diversity analysis, as it efficiently taps into preexisting structural knowledge for the considered species as well as for structural annotation based on MS/MS similarities for PSMs of shared biochemical origin (**Figure 3**). The molecular networking analysis is empowered by the vast crowdsourced compilation of metabolomics data sets and chemical knowledge within the GNPS data set (138). Further, software tools such as MS2LDA (which finds shared structural motifs) (134), MolNetEnhancer (which uses MS2LDA motifs for metabolite classification) (34), Qemistree (which computes MS/MS phylogenies) (129), and others are particularly useful for the efficient mining of the spectral interrelationships of PSMs within an investigated system. Additionally, these tools, together with in silico MS/MS approaches such as SIRIUS (28), CSI:FingerID (30), NAP (24), DEREPLICATOR (94), and others, are integrated in the GNPS community library. To complete this data analysis ecosystem, GNPS authors have developed an equivalent of the NCBI BLAST search to query MS/MS spectra, termed MASST (139). Recently, a feature-based molecular networking method that enables quantitative analysis and the resolution of isomers, including from ion mobility spectrometry, has been integrated with other annotation tools and is now becoming one of the most commonly used analysis pipelines within the GNPS environment (98).

Classical spectral similarity scoring algorithms to construct molecular networks include, among others, probability-based-matching algorithms and the normalized dot product, which calculates the cosine of the angle between two MS/MS spectra (70). These algorithms, which rely on accurate matching of common peaks in two spectra, however, do not take into account mass shifts caused by structural analogs whose structures differ in modifications such as methylation, hydroxylation, or glycosylation. Several recent studies have highlighted the importance of considering neutral losses for aligning MS/MS spectra (104, 134). Along these lines, a simple and sensitive unsupervised approach termed biclustering, which was originally designed to find differentially

coexpressed gene modules under two types of conditions, was recently implemented to improve the scoring and classification of MS/MS similarities for PSMs by considering both types of fragmentation variables (fragment and neutral loss similarity scores) in constructing possible compound familial groupings (79, 81, 96). The output of the biclustering modules can be transposed as molecular networks to visualize other orthogonal biological information (for instance, metabolite inducibility by herbivore attack) linked to a given MS/MS spectrum (**Figure 3**). Such molecular networking procedures were applied to detect pathway- and PSM-specific natural variation effects in *Nicotiana attenuata* native populations (79) or insect-species-specific defensive metabolites (80).

3.3. Linkage Analysis of MS Metabolomics Data with Genetics Data

Untargeted MS metabolomics has been extensively applied in combination with genome and transcriptome studies to understand plant gene functions. These multi-omics analyses either rely on coexpression analyses between transcriptomics and metabolomics expression data to identify gene-to-metabolite associations inferred from conditional transcriptional regulation or are used to exploit natural variation via metabolic quantitative trait loci analysis and metabolic genome-wide association studies (mGWASs) (36). As for the latter, linkage mapping analysis using structured populations such as multiparent advanced generation intercross (MAGIC) and recombinant inbred lines (RILs) and/or unstructured natural populations as used by GWASs together with the combination of NGS and MS metabolomics represent powerful tools to query genomic regions associated with specific PSM traits in both model and crop species (20, 148). The ability to statistically harness the multidimensional variance of PSM production (at tissue-, species-, and population-level, as well as according to ontogeny/phenology and (a)biotic factor interactions) using MS metabolomics data has been exploited on rare occasions to reveal divergent and convergent genetic regulation of plant metabolism (20, 21). Coexpression-based approaches have notably been employed to lead the identification of transcription factors regulating aliphatic glucosinolate (54), steroidal glycoalkaloid biosynthesis (58), and flavonoid pathways (81) and can be integrated with mGWASs and expression QTLs to more confidently impute gene functions to plant metabolism (148) (**Figure 3**). Interestingly, models used in GWASs have recently been employed to identify PSMs (analogous to loci) associated with insect resistance and ultraviolet radiation tolerance from the thousands of candidate metabolites (77, 103). In a recent study, we similarly parsed population-level intraspecific variations in PSMs to infer novel jasmonate-dependent metabolic traits potentially involved in defense against insect herbivory (79).

4. NEXT-GENERATION MS METABOLOMICS REINVIGORATES THE TESTING OF PLANT DEFENSE THEORY PREDICTIONS

Over the past six decades, plant defense theories have provided conceptual frameworks from which to infer predictions about the evolution and function of the considerable diversity of PSMs (113). These plant defense theories have been topics of excellent reviews (32, 113, 122). These theories were often posed at different levels of analysis (120), making it difficult to contrast their critical predictions and advance to the next cycle of theory development (113, 122). Additionally, as previously discussed, the lack of comprehensive metabolomics data and associated processing workflows to compare the metabolic space among different plant taxa in a common currency has thwarted the scientific maturation of the field, as predictions were made far beyond the reach of the available data (113). In this section, we briefly discuss seminal guiding theories in chemical ecology (**Table 1**) and how the previously mentioned next-generation metabolomics advances would allow rigorous testing of aspects of these important theories that inspired previous generations of researchers.

Table 1 Long-standing hypotheses or models addressing plant metabolism–centered evolutionary and ecological questions that would benefit from next-generation metabolomics analysis

Guiding plant defense theory	Year	Critical predictions for plant specialized metabolites (PSMs)
Screening hypothesis (43, 60)	1991	During evolutionary screening processes, PSMs that are being exapted increase in specificity in the phytochemical profile as a consequence of providing adaptive value for a given environmental condition
Escape-and-radiate hypothesis (33)	1964	Innovations in PSMs contribute to lineage diversification; hence, a strong association between rates of species-lineage diversification and character evolution of PSMs is predicted
Geographic mosaic hypothesis (127)	1994	PSM diversity is expected to be heterogeneously distributed in geographic scales, with the highest diversity in coevolutionary hotspots where strong reciprocal selection on the interacting species takes place
Optimal defense hypothesis (92, 105)	1974	Plants directionally allocate resources to PSM production according to their defensive value and the probability of whole-plant or tissue-level attack
Moving target hypothesis (1)	1994	When attacked by herbivores, PSM production is reconfigured nondirectionally as a way of creating random plant phenotypes that are hard for insects to adapt to
Information transfer hypothesis (65, 66)	2015	PSM diversity is conceptualized as information through which plants interact with their environment; inducibility and specificity in PSM landscapes are expected to be correlated with the specificity of information exchange
Plant apparency hypothesis (37)	1976	Apparent plants invest quantitative (high in richness, less toxic) defensive PSMs, whereas unapparent plants invest qualitative (low in richness, highly toxic) specialized metabolite defense
Carbon:nutrient balance hypothesis (18)	1983	PSM production is determined by the availability of carbon and nitrogen
Growth rate hypothesis (22)	1985	Slow-growing plants evolve quantitative specialized metabolite defense, whereas fast-growing plants evolve qualitative specialized metabolite defense
Growth:differentiation balance hypothesis (87)	1932	The allocation pattern of PSMs is expected to be curvilinear across a resource gradient, peaking at the intermediate resource level

4.1. Formulating Metabolomics-Level Predictions for Plant Defense Theories

Two of the most important concepts that have guided the functional interpretation of PSM diversity as an adaptive response to aggressors, in particular, phytophagous insects, are the synergy and screening hypotheses. The synergy hypothesis, which has not yet been addressed using large-scale possibilities offered by PSM metabolomics, postulates that PSM function is metabolic-context-dependent, since most individual PSMs exhibit biological activities only when part of a given metabolic landscape (31, 106), and hence not directly inferable through traditional bioassay-guided workflows. Molecular networking of MS/MS metabolomics data can identify metabolites associated with a given resistance phenotype in light of structurally related metabolite congeners with which functional synergies are likely to be established. Similarly, PSM covariance in natural populations could be mapped onto MS/MS molecular networks built on structural similarities as a new layer to examine PSM synergies.

The screening hypothesis formulated in 1991 complements the functional synergy perspective (60) (**Table 1**). Based on assumptions that most PSMs have no adaptive value at any given time and that the probability of generating novel PSMs that are biologically active is low, the screening hypothesis suggests that PSM diversity is maintained at different hierarchical scales in order to provide the raw material to exapt bioactivities of previously extraneous or alternatively adapted metabolites (43). In the genera *Bursera* and *Inga*, correlations between rates of species diversification and the structural diversity within target classes of PSMs have been interpreted by some as

experimental support of the screening hypothesis (14, 75). However, it remains untested whether these trends apply at broader PSM scales. As mentioned earlier, such analysis is now possible in the context of recent MS metabolomics advances. Additionally, while the vast majority of detected metabolites remain unannotated, molecular networking approaches partly circumvent this hurdle by providing a means of achieving, without a priori chemical knowledge, compound class-level clustering of known and unknown MS/MS spectra. It is then possible to examine how each of these computationally generated compound class clusters evolved with species diversification rates. The complexity of most PSM pathways coupled with enzyme promiscuity provides the biochemical underpinnings for the large number of analogs with small chemical differences that are thought to be screened or functioning in synergy. Again, the molecular networking–based ordination and exploration of these closely related structures in MS/MS metabolomics data sets could accelerate the detection of those subtle chemical modifications that vary in natural populations and confer adaptive value in a given ecological context.

Similarly, large-scale MS metabolomics data could be used to test some of the predictions of Ehrlich & Raven (33) as well as Thompson's (127) geographic mosaic hypothesis of coevolution theories (**Table 1**). Core to Ehrlich & Raven's (33) coevolution theory is the escape-and-radiate hypothesis, which predicts that interspecific variations in metabolic diversity are responsible for lineage diversification in plants (33). Key predictions at the level of PSM production have remained untested at taxonomic scales beyond closely related congeners (4) due to the difficulty of conducting comprehensive analyses on a large array of PSMs at a time and the lack of a metric with which to compare rates of PSM profile evolution. As elegantly discussed and convincingly exemplified by Sedio (115), the development of a chemical structural–compositional similarity metric that weights the structural similarity of every pair of compound-derived MS/MS spectra between species-level phytochemical profiles offers a rigorous test to central predictions of Ehrlich & Raven's coevolutionary theory regarding interspecific metabolic variations (116). Thompson's geographic mosaic hypothesis of coevolution emphasizes the importance of geographic variations in natural selection mosaics, resulting in the patchy distributions of PSMs across space and time (127). A comprehensive computational MS/MS metabolomics analysis of the cosmopolitan plant genus *Euphorbia* has revealed structural diversity patterns across geographically separated phylogenetic clades of this genus that are consistent with the geographic mosaic hypothesis (35).

Starting from the 1970s, much of the focus of plant defense theories has been on the costs and benefits of PSM production. Two core sets of hypotheses can be distinguished. One includes the optimal defense (OD) (92, 105), moving target (MT) (1), and apparency (37) hypotheses that place major emphasis on explaining temporal and spatial distributions of PSMs according to defensive function and probability of attack. Whereas the other group of hypotheses [including the carbon:nutrient balance hypothesis (18), the growth rate hypothesis (22), and the growth:differentiation balance hypothesis (87)] seeks mechanistic explanations for how variations in resource availability influence tradeoffs between plant resource investments in growth and PSM production. The OD hypothesis postulates that plants directionally adjust their preferential investments into costly defensive PSMs to maximize their fitness as a function of the probability of future attack (92, 105). The apparency hypothesis further defines apparency as the ecological predictability (in space and time) of entire plants or tissues and predicts that more apparent plants will invest more heavily in broadly defensive PSMs (37). In contrast, the MT hypothesis, posed at the same functional level of analysis, argues that the evidence of such a directional metabolic change is unsupported but rather that PSMs change randomly when plant tissues are consumed by herbivores, thereby creating a metabolic moving target, which could thwart herbivore adaptation (1). In Section 4.2, we discuss how information theory modeling can help to distinguish between the contrasting predictions of the OD and MT hypotheses regarding attack-induced trajectories of PSMs.

4.2. Decoding Plant Specialized Metabolite Diversity in the Form of Simple Statistical Currencies Using Information Theory

Several recent studies have demonstrated the use of information theory to quantify the chemical information content in phytochemical studies in a framework that can be conceptually connected to other types of information processing (80, 81, 149). Information theory was first introduced in a seminal article by Claude Shannon (118) in 1948 in which he describes how uncertainty can be represented, manipulated, and quantified using a probability-based model. This revolutionary theoretical framework not only laid the foundation for a mathematical analysis of information but also opened new avenues for almost every field of information-rich science and technology. In ecology, information theory has been particularly useful in assessing biodiversity and trophic flows by quantifying numbers of organisms and patterns of interactions of trophic processes, respectively (132). Information theory has been successfully employed in genomics to quantify sequence conservation information (112) and in multi-organ transcriptomics studies to decipher gene specifiers for organ-level transcriptome specialization (91).

In the case of MS metabolomics, a recent transposition of statistics derived from information theory analysis was employed to parse plant tissue-level metabolic specialization from large-scale MS/MS data (81). In a follow-up study (80), an integration of information theory statistics and MS/MS molecular networking was used to quantify consistencies in highly controlled herbivory elicitations of temporal modulations of PSM diversity (α, β, and γ metabolic diversity) at the three functional levels of phytochemical diversity (66, 143). In this analysis, trajectories of herbivory-elicited plasticity of PSM production are captured using information theory descriptors of metabolome diversity (Hj index) and specialization (δj index) calculated from MS/MS metabolomes. The ability to describe a complex information landscape in a few discrete indices allowed for tests of contrasting plant defense theory predictions posed at the level of herbivory-elicited metabolomes, such as those of the OD and MT hypotheses: namely unidirectional accumulations of metabolites with defense functions (OD) versus nondirectional metabolic changes (MT) (80) (**Table 1**). Consistent with the prediction of the OD hypothesis, herbivory-induced PSM changes were channeled toward the production of defensive metabolites, which, in the context of the above statistical descriptors, translated into an overall greater metabolite profile specialization (δj index) and lower metabolic diversity index (Hj index). In contrast, the MT hypothesis predicts nondirectional changes in the metabolome. Within an information theory framework, such nondirectional reconfigurations result in an overall increase of metabolic diversity as an indicator of greater metabolic information uncertainty as well as no consistent change in metabolome specialization, leading to a random distribution of the specificity indices of individual metabolites (no metabolite exhibits significant specificity in response to herbivore attack) (80).

In the case of the previously described screening hypothesis, one key prediction is that screening processes should give rise to adaptive PSM divergence across populations maintained in order to increase the probability that a plant contains specific PSMs that may eventually be effective against a particular type of natural enemy. In an information theory statistical framework, such PSM exaptations should materialize as an increase in the specificity indices of exapted metabolites whose bioactivity becomes adaptive for a given environmental context. Another application of transposing information theory to phytochemical communication is exemplified by a recent study (149) decoding conflicting volatile-mediated information processes between plants and herbivores as a test of the information transfer hypothesis (65, 66) (**Table 1**). In this elegant study, information theory modeling of volatile information certainty for insects' foraging is consistent with the premise that interspecific volatile redundancy is a key variable in the communication arms race between plants and their herbivores (149).

5. CONCLUDING REMARKS

When seminal hypotheses to explain the different levels of PSM diversity were developed, hypothesis-driven methods were the unique means of making scientific advances. Due to the technical inability at the time to conduct comprehensive metabolite profiling at appropriate taxonomical scales, key predictions of these hypotheses remained largely untested. The last decade of metabolomics hardware and computational breakthroughs has led to a next generation of metabolomics analyses for which the collection and data-driven ordination of structural information on up to several thousands of metabolites per measurement is now achievable. These advances provide unprecedented access to PSM diversity as a functional variable to revisit these seminal hypotheses (once their predictions are reformulated to match the new scales afforded by the new comprehensive analysis), examine genetic determinants of phytochemical diversity, and accelerate the identification of cryptic bioactive chemicals. In plant genomics, advances and standardization of NGS methods have recently culminated with the release of the 1,000 Plant Genomes Project (1KP), representing the most detailed catalogue of genomic variations in plants. The recent community-level efforts for metabolomics data archiving, reuse, and interoperability with other platforms should further foster additional opportunities to study the metabolic prowess of plants.

SUMMARY POINTS

1. Variations in plant specialized metabolite (PSM) diversity detected at different hierarchical levels ranging from intraindividual to intra- and interspecific levels are important functional dimensions of plants' adaptations to their environments but are challenging to examine.
2. Theoretical frameworks about phytochemical diversity remain largely untested, as their main predictions reach far beyond the data needed to fully test them (in particular, large-scale phytochemical analysis).
3. The next generation of untargeted metabolomics marrying advances in mass spectrometry (MS) resolution with streamlined methods for large volumes of tandem mass spectrometry (MS/MS) data deconvolution already allows for the transition from traditional reductionist approaches focusing on a few metabolites to a more holistic investigation of a plant's metabolome.
4. While still in their infancy in the metabolomics field, community-wide efforts [such as the Global Natural Products Social Molecular Networking (GNPS) ecosystem] for MS data sharing, analysis, and knowledge capture through bioinformatics solutions often inspired from successes in genomics contribute substantially to the global interpretation of metabolomics data and reinvigorate a community-wide interest in structural chemical diversity.
5. Analogous to the exploration of biodiversity revolutionized by the advent of next-generation sequencing (NGS), we recommend, in order to fully embrace the opportunities of the metabolomics information-rich era, that predictions of the seminal theories be posed at this new scale of analysis and that simple statistical currencies be implemented to compare and contrast metabolomes.

6. With the proliferation of data-intensive analyses in this field, a risk of marginalization of hypothesis-driven research on phytochemical diversity could exist. An exciting challenge should notably be to infuse a plant-natural-history-driven perspective into metabolomics data exploration in order to revisit important theories that inspired previous generations.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Ian T. Baldwin for many insightful discussions. D.L. was funded by the Max Planck Society and by the Collaborative Research Centre Chemical Mediators in Complex Biosystems (ChemBioSys) (SFB 1127). E.G. was supported within the framework of the Deutsche Forschungsgemeinschaft Excellence Initiative to the University of Heidelberg and by the CNRS and the Initiative d'Excellence (IdEx) via the University of Strasbourg.

LITERATURE CITED

1. Adler FR, Karban R. 1994. Defended fortresses or moving targets? Another model of inducible defenses inspired by military metaphors. *Am. Nat.* 144:813–32
2. Adler LS, Schmitt J, Bowers MD. 1995. Genetic variation in defensive chemistry in *Plantago lanceolata* (Plantaginaceae) and its effect on the specialist herbivore *Junonia coenia* (Nymphalidae). *Oecologia* 101:75–85
3. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, et al. 2012. KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.* 53:e1
4. Agrawal AA, Fishbein M, Halitschke R, Hastings AP, Rabosky DL, Rasmann S. 2009. Evidence for adaptive radiation from a phylogenetic study of plant defenses. *PNAS* 106:18067–72
5. Aharoni A, de Vos CHR, Verhoeven HA, Maliepaard CA, Kruppa G, et al. 2002. Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS* 6:217–34
6. Aksenov AA, da Silva R, Knight R, Lopes NP, Dorrestein PC. 2017. Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* 1:0054
7. Allard PM, Peresse T, Bisson J, Gindro K, Marcourt L, et al. 2016. Integration of molecular networking and *in-silico* MS/MS fragmentation for natural products dereplication. *Anal. Chem.* 88:3317–23
8. Allwood JW, Ellis DI, Goodacre R. 2008. Metabolomic technologies and their application to the study of plants and plant–host interactions. *Physiol. Plant* 132:117–35
9. Allwood JW, Weber RJM, Zhou J, He S, Viant MR, Dunn WB. 2013. CASMI—the small molecule identification process from a Birmingham perspective. *Metabolites* 3:397–411
10. Alonso A, Julià A, Beltran A, Vinaixa M, Díaz M, et al. 2011. AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics* 27:1339–40
11. Alonso A, Marsal S, Julià A. 2015. Analytical methods in untargeted metabolomics: state of the art in 2015. *Front. Bioeng. Biotechnol.* 3:23
12. Babushok VI, Linstrom PJ, Reed JJ, Zenkevich IG, Brown RL, et al. 2007. Development of a database of gas chromatographic retention properties of organic compounds. *J. Chromatogr. A* 1157:414–21
13. Baldwin IT. 1998. Jasmonate-induced responses are costly but benefit plants under attack in native populations. *PNAS* 95:8113–18
14. Becerra JX, Noge K, Venable DL. 2009. Macroevolutionary chemical escalation in an ancient plant–herbivore arms race. *PNAS* 106:18062–66

15. Blaženović I, Kind T, Ji J, Fiehn O. 2018. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* 8:31
16. Broeckling CD, Afsar FA, Neumann S, Ben-Hur A, Prenni JE. 2014. RAMClust: A novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Anal. Chem.* 86:6812–17
17. **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* 9:33–43**
18. Bryant JP, Chapin FS, Klein DR. 1983. Carbon/nutrient balance of boreal plants in relation to vertebrate herbivory. *Oikos* 40:357–68
19. Cambiaghi A, Ferrario M, Masseroli M. 2017. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Brief. Bioinformat.* 18:498–510
20. Chan EK, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ. 2011. Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLOS Biol.* 9:e1001125
21. Chen W, Wang W, Peng M, Gong L, Gao Y, et al. 2016. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat. Commun.* 7:12767
22. Coley PD, Bryant JP, Chapin FS 3rd. 1985. Resource availability and plant antiherbivore defense. *Science* 230:895–99
23. da Silva RR, Dorrestein PC, Quinn RA. 2015. Illuminating the dark matter in metabolomics. *PNAS* 112:12549–50
24. da Silva RR, Wang MX, Nothias LF, van der Hooft JJJ, Caraballo-Rodriguez AM, et al. 2018. Propagating annotations of molecular networks using *in silico* fragmentation. *PLOS Comput. Biol.* 14:e1006089
25. D'Auria JC, Gershenzon J. 2005. The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. *Curr. Opin. Plant Biol.* 8:308–16
26. Dixon RA, Strack D. 2003. Phytochemistry meets genome analysis, and beyond. *Phytochemistry* 62:815–16
27. Domingo-Almenara X, Brezmes J, Vinaixa M, Samino S, Ramirez N, et al. 2016. eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC/MS-based metabolomics. *Anal. Chem.* 88:9821–29
28. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, et al. 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* 16:299–302
29. Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, et al. 2020. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol.* **https://doi.org/10.1038/s41587-020-0740-8**
30. **Dührkop K, Shen H, Meusel M, Rousu J, Bocker S. 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *PNAS* 112:12580–85**
31. Dyer LA, Dodson CD, Stireman JO 3rd, Tobler MA, Smilanich AM, et al. 2003. Synergistic effects of three *Piper* amides on generalist and specialist herbivores. *J. Chem. Ecol.* 29:2499–514
32. Dyer LA, Philbin CS, Ochsenrider KM, Richards LA, Massad TJ, et al. 2018. Modern approaches to study plant–insect interactions in chemical ecology. *Nat. Rev. Chem.* 2:50–64
33. Ehrlich PR, Raven PH. 1964. Butterflies and plants: a study in coevolution. *Evolution* 18:586–608
34. Ernst M, Kang KB, Caraballo-Rodriguez AM, Nothias LF, Wandy J, et al. 2019. MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* 9:144
35. Ernst M, Nothias LF, van der Hooft JJJ, Silva RR, Saslis-Lagoudakis CH, et al. 2019. Assessing specialized metabolite diversity in the cosmopolitan plant genus *Euphorbia* L. *Front. Plant Sci.* 10:846
36. Fang C, Fernie AR, Luo J. 2019. Exploring the diversity of plant metabolism. *Trends Plant Sci.* 24:83–98
37. Feeny P. 1976. Plant apparency and chemical defense. In *Biochemical Interaction Between Plants and Insects*, ed. JW Wallace, RL Mansell, pp. 1–40. Boston: Springer
38. Fernie AR. 2007. The future of metabolic phytochemistry: larger numbers of metabolites, higher resolution, greater understanding. *Phytochemistry* 68:2861–80
39. Fernie AR, Stitt M. 2012. On the discordance of metabolomics with proteomics and transcriptomics: coping with increasing complexity in logic, chemistry, and network interactions scientific correspondence. *Plant Physiol.* 158:1139–45

17. Rationalizes a computational pipeline to infer precursor-to-fragment relationships in large-scale data-independent MS/MS experiments.

30. Combines fragmentation tree computation and machine learning methods for searching a molecular structure database using MS/MS data.

40. Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L. 2004. Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* 5:763–69
41. Fiehn O. 2016. Metabolomics by gas chromatography–mass spectrometry: combined targeted and untargeted profiling. *Curr. Protoc. Mol. Biol.* 114:30.4.1–32
42. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. 2000. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18:1157–61
43. Firn RD, Jones CG. 2003. Natural products—a simple model to explain chemical diversity. *Nat. Prod. Rep.* 20:382–91
44. Fraenkel GS. 1959. The raison d'être of secondary plant substances. *Science* 129:1466–70
45. Fraser PD, Pinto MES, Holloway DE, Bramley PM. 2000. Application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. *Plant J.* 24:551–58
46. Fujii T, Matsuda S, Tejedor ML, Esaki T, Sakane I, et al. 2015. Direct metabolomics for plant cells by live single-cell mass spectrometry. *Nat. Protoc.* 10:1445–56
47. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, et al. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteom.* 11:O111.016717
48. Gohlke RS, McLafferty FW. 1993. Early gas chromatography/mass spectrometry. *J. Am. Soc. Mass Spectrom.* 4:367–71
49. Gullberg J, Jonsson P, Nordstrom A, Sjostrom M, Moritz T. 2004. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. *Anal. Biochem.* 331:283–95
50. Hartmann T. 2007. From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry* 68:2831–46
51. Haslam E. 1986. Secondary metabolism—fact and fiction. *Nat. Prod. Rep.* 3:217–49
52. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, et al. 2013. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* 41:D456–63
53. Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, et al. 2005. Elucidation of gene-to-gene and metabolite-to-gene networks in *Arabidopsis* by integration of metabolomics and transcriptomics. *J. Biol. Chem.* 280:25590–95
54. Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, et al. 2007. Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *PNAS* 104:6478–83
55. Holeski LM, Hillstrom ML, Whitham TG, Lindroth RL. 2012. Relative importance of genetic, ontogenetic, induction, and seasonal variation in producing a multivariate defense phenotype in a foundation tree species. *Oecologia* 170:695–707
56. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, et al. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* 45:703–14
57. Hsu CC, ElNaggar MS, Peng Y, Fang JS, Sanchez LM, et al. 2013. Real-time metabolomics on living microorganisms using ambient electrospray ionization flow-probe. *Anal. Chem.* 85:7014–18
58. Itkin M, Rogachev I, Alkan N, Rosenberg T, Malitsky S, et al. 2011. GLYCOALKALOID METABOLISM1 as required for steroidal alkaloid glycosylation and prevention of phytotoxicity in tomato. *Plant Cell* 23:4507–25
59. **Jacobowitz JR, Weng JK. 2020. Exploring uncharted territories of plant specialized metabolism in the postgenomic era. *Annu. Rev. Plant Biol.* 71:631–58**
60. Jones CG, Firn RD. 1991. On the evolution of plant secondary chemical diversity. *Philos. Trans. R. Soc. B* 333:273–80
61. Karban R, Baldwin IT. 1997. *Induced Responses to Herbivory*. Chicago: Univ. Chicago Press
62. Katajamaa M, Miettinen J, Oresic M. 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22:634–36
63. Katajamaa M, Oresic M. 2007. Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A* 1158:318–28

59. Provides a thorough synthesis of the biochemical exploration of plant specialized metabolism diversity.

64. Kempel A, Schadler M, Chrobock T, Fischer M, van Kleunen M. 2011. Tradeoffs associated with constitutive and induced plant resistance against herbivory. *PNAS* 108:5685–89
65. Kessler A. 2015. The information landscape of plant constitutive and induced secondary metabolite production. *Curr. Opin. Insect Sci.* 8:47–53
66. Kessler A, Kalske A. 2018. Plant secondary metabolite diversity and species interactions. *Annu. Rev. Ecol. Evol. Syst.* 49:115–38
67. Khersonsky O, Tawfik DS. 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* 79:471–505
68. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, et al. 2016. PubChem substance and compound databases. *Nucleic Acids Res.* 44:D1202–13
69. Kim SG, Yon F, Gaquerel E, Gulati J, Baldwin IT. 2011. Tissue specific diurnal rhythms of metabolites and their regulation during herbivore attack in a native tobacco, *Nicotiana attenuata*. *PLOS ONE* 6:e26214
70. Kind T, Tsugawa H, Cajka T, Ma Y, Lai ZJ, et al. 2018. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* 37:513–32
71. Kliebenstein DJ, Kroymann J, Brown P, Figuth A, Pedersen D, et al. 2001. Genetic control of natural variation in *Arabidopsis* glucosinolate accumulation. *Plant Physiol.* 126:811–25
72. Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T. 2001. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate–dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. *Plant Cell* 13:681–93
73. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, et al. 2005. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21:1635–38
74. Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S. 2012. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* 84:283–89
75. Kursar TA, Dexter KG, Lokvam J, Pennington RT, Richardson JE, et al. 2009. The evolution of antiherbivore defenses and their contribution to species coexistence in the tropical tree genus *Inga*. *PNAS* 106:18073–78
76. Kusano M, Redestig H, Hirai T, Oikawa A, Matsuda F, et al. 2011. Covering chemical diversity of genetically-modified tomatoes using metabolomics for objective substantial equivalence assessment. *PLOS ONE* 6:e16989
77. Kuzina V, Nielsen JK, Augustin JM, Torp AM, Bak S, Andersen SB. 2011. *Barbarea vulgaris* linkage map and quantitative trait loci for saponins, glucosinolates, hairiness and resistance to the herbivore *Phyllotreta nemorum*. *Phytochemistry* 72:188–98
78. Lei ZT, Huhman DV, Sumner LW. 2011. Mass spectrometry strategies in metabolomics. *J. Biol. Chem.* 286:25435–42
79. Li D, Baldwin IT, Gaquerel E. 2015. Navigating natural variation in herbivory-induced secondary metabolism in coyote tobacco populations using MS/MS structural analysis. *PNAS* 112:E4147–55
80. Li D, Halitschke R, Baldwin IT, Gaquerel E. 2020. Information theory tests critical predictions of plant defense theory for specialized metabolism. *Sci. Adv.* 6:eaaz0381
81. Li D, Heiling S, Baldwin IT, Gaquerel E. 2016. Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. *PNAS* 113:E7610–18
82. Lindon JC, Holmes E, Bollard ME, Stanley EG, Nicholson JK. 2004. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers* 9:1–31
83. Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J. 1980. *Applications of Artificial Intelligence for Organic Chemistry: the DENDRAL Project*. New York: McGraw-Hill
84. Link H, Fuhrer T, Gerosa L, Zamboni N, Sauer U. 2015. Real-time metabolome profiling of the metabolic switch between starvation and growth. *Nat. Methods* 12:1091–97
85. Little JL, Cleven CD, Brown SD. 2011. Identification of "known unknowns" utilizing accurate mass data and chemical abstracts service databases. *J. Am. Soc. Mass Spectrom.* 22:348–59
86. Lommen A. 2009. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.* 81:3079–86

66. Outlines an information transfer hypothesis framework for understanding plant specialized metabolite diversity as a factor driving ecological interactions.

80. Applies information theory–derived statistics to quantify diversity and specialization in metabolome profiles and test seminal plant defense theories.

87. Loomis WF. 1932. Growth-differentiation balance vs carbohydrate-nitrogen ratio. *Proc. Am. Soc. Horticult. Sci.* 29:240–45
88. Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–73
89. Lynn KS, Cheng ML, Chen YR, Hsu C, Chen A, et al. 2015. Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information. *Anal. Chem.* 87:2143–51
90. Marshall AG, Hendrickson CL. 2008. High-resolution mass spectrometers. *Annu. Rev. Anal. Chem.* 1:579–99
91. Martinez O, Reyes-Valdes MH. 2008. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *PNAS* 105:9709–14
92. McKey D. 1974. Adaptive patterns in alkaloid physiology. *Am. Nat.* 108:305–20
93. Moghe GD, Last RL. 2015. Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiol.* 169:1512–23
94. Mohimani H, Gurevich A, Mikheenko A, Garg N, Nothias LF, et al. 2017. Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* 13:30–37
95. Moore BD, Andrew RL, Kulheim C, Foley WJ. 2014. Explaining intraspecific diversity in plant secondary metabolites in an ecological context. *New Phytol.* 201:733–50
96. Naake T, Gaquerel E. 2017. MetCirc: navigating mass spectral similarity in high-resolution MS/MS metabolomics data. *Bioinformatics* 33:2419–20
97. Nicolaou KC, Snyder SA, Montagnon T, Vassilikogiannakis G. 2002. The Diels–Alder reaction in total synthesis. *Angew. Chem. Int. Ed. Engl.* 41:1668–98
98. Nothias L-F, Petras D, Schmid R, Dührkop K, Rainer J, et al. 2020. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* 17:905–8
99. Oliver SG, Winson MK, Kell DB, Baganz F. 1998. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* 16:373–78
100. Pareek V, Tian H, Winograd N, Benkovic SJ. 2020. Metabolomics and mass spectrometry imaging reveal channeled de novo purine synthesis in cells. *Science* 368:283–90
101. Pelser PB, de Vos H, Theuring C, Beuerle T, Vrieling K, Hartmann T. 2005. Frequent gain and loss of pyrrolizidine alkaloids in the evolution of *Senecio* section *Jacobaea* (Asteraceae). *Phytochemistry* 66:1285–95
102. Pence HE, Williams A. 2010. ChemSpider: an online chemical information resource. *J. Chem. Educ.* 87:1123–24
103. Peng M, Shahzad R, Gul A, Subthain H, Shen S, et al. 2017. Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. *Nat. Commun.* 8:1975
104. Rasche F, Svatos A, Maddula RK, Bottcher C, Bocker S. 2011. Computing fragmentation trees from tandem mass spectrometry data. *Anal. Chem.* 83:1243–51
105. Rhoades DF. 1979. Evolution of plant chemical defense against herbivores. In *Herbivores: Their Interaction with Secondary Plant Metabolites*, ed. GA Rosenthal, DH Janzen, pp. 1–55. New York: Academic
106. Richards LA, Glassmire AE, Ochsenrider KM, Smilanich AM, Dodson CD, et al. 2016. Phytochemical diversity and synergistic effects on herbivores. *Phytochem. Rev.* 15:1153–66
107. Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, et al. 2001. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13:11–29
108. Roessner U, Willmitzer L, Fernie AR. 2001. High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* 127:749–64
109. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, et al. 2016. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* 13:741–48
110. Sakurai T, Yamada Y, Sawada Y, Matsuda F, Akiyama K, et al. 2013. PRIMe update: innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. *Plant Cell Physiol.* 54:e5
111. Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, et al. 2012. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry* 82:38–45

112. Schneider TD, Mastronarde DN. 1996. Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discrete Appl. Math.* 71:259–68
113. Schuman MC, Baldwin IT. 2016. The layers of plant responses to insect herbivores. *Annu. Rev. Entomol.* 61:373–94
114. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, et al. 2017. Critical assessment of small molecule identification 2016: automated methods. *J. Cheminform.* 9:22
115. Sedio BE. 2017. Recent breakthroughs in metabolomics promise to reveal the cryptic chemical traits that mediate plant community composition, character evolution and lineage diversification. *New Phytol.* 214:952–58
116. Sedio BE, Rojas Echeverri JC, Boya PC, Wright SJ. 2017. Sources of variation in foliar secondary chemistry in a tropical forest tree community. *Ecology* 98:616–23
117. Shahaf N, Rogachev I, Heinig U, Meir S, Malitsky S, et al. 2016. The WEIZMASS spectral library for high-confidence metabolite identification. *Nat. Commun.* 7:12423
118. Shannon CE. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27:379–423
119. Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat. Biotechnol.* 26:1135–45
120. Sherman PW. 1988. The levels of analysis. *Anim. Behav.* 36:616–19
121. **Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78:779–87**
122. Stamp N. 2003. Out of the quagmire of plant defense hypotheses. *Q. Rev. Biol.* 78:23–55
123. Strauss SY, Rudgers JA, Lau JA, Irwin RE. 2002. Direct and ecological costs of resistance to herbivory. *Trends Ecol. Evol.* 17:278–85
124. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M. 2012. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr. Bioinform.* 7:96–108
125. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T. 1988. Protein and polymer analyses up to *m/z* 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* 2:151–53
126. Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G. 2012. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat. Biotechnol.* 30:826–28
127. Thompson JN. 2005. *The Geographic Mosaic of Coevolution*. Chicago: Univ. Chicago Press
128. Tikunov YM, Laptenok S, Hall RD, Bovy A, de Vos RC. 2012. MSClust: A tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics* 8:714–18
129. Tripathi A, Vázquez-Baeza Y, Gauglitz JM, Wang M, Dührkop K, et al. 2021. Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. *Nat. Chem. Bio.* 17:146–51
130. **Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, et al. 2015. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* 12:523–26**
131. Tsugawa H, Nakabayashi R, Mori T, Yamada Y, Takahashi M, et al. 2019. A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nat. Methods* 16:295–98
132. Ulanowicz RE. 2001. Information theory in ecology. *Comput. Chem.* 25:393–99
133. Uppal K, Walker DI, Jones DP. 2017. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. *Anal. Chem.* 89:1063–67
134. van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S. 2016. Topic modeling for untargeted substructure exploration in metabolomics. *PNAS* 113:13738–43
135. Venable JD, Dong M-Q, Wohlschlegel J, Dillin A, Yates JR. 2004. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* 1:39–45
136. Von Poser GL, Toffoli ME, Sobral M, Henriques AT. 1997. Iridoid glucosides substitution patterns in *Verbenaceae* and their taxonomic implication. *Plant Syst. Evol.* 205:265–87
137. von Roepenack-Lahaye E, Degenkolb T, Zerjeski M, Franz M, Roth U, et al. 2004. Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* 134:548–59
138. **Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, et al. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* 34:828–37**

121. Describes one of the first examples of open-source software for untargeted metabolomics data processing.

130. Reports on a broadly applicable open-source program for deconvolution and annotation of data-independent MS/MS data.

138. Breakthrough report on the crowd-source Global Natural Products Social Molecular Networking (GNPS) infrastructure for MS/MS metabolomics data sharing, reanalysis and chemical knowledge capture.

139. Wang MX, Jarmusch AK, Vargas F, Aksenov AA, Gauglitz JM, et al. 2020. Mass spectrometry searches using MASST. *Nat. Biotechnol.* 38:23–26
140. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, et al. 2012. Mass spectral molecular networking of living microbial colonies. *PNAS* 109:E1743–52
141. Weng JK. 2014. The evolutionary paths towards complexity: a metabolic perspective. *New Phytol.* 201:1141–49
142. Weng JK, Philippe RN, Noel JP. 2012. The rise of chemodiversity in plants. *Science* 336:1667–70
143. Whittaker RH. 1972. Evolution and measurement of species diversity. *Taxon* 21:213–51
144. Wink M. 2003. Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry* 64:3–19
145. Wink M, Carey DB. 1994. Variability of quinolizidine alkaloid profiles of *Lupinus argenteus* (Fabaceae) from North America. *Biochem. Syst. Ecol.* 22:663–69
146. Yamashita M, Fenn JB. 1984. Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.* 88:4451–59
147. Zhou B, Xiao JF, Tuli L, Ressom HW. 2012. LC-MS-based metabolomics. *Mol. Biosyst.* 8:470–81
148. Zhu G, Wang S, Huang Z, Zhang S, Liao Q, et al. 2018. Rewiring of the fruit metabolome in tomato breeding. *Cell* 172:249–61.e12
149. Zu P, Boege K, Del-Val E, Schuman MC, Stevenson PC, et al. 2020. Information arms race explains plant-herbivore chemical communication in ecological communities. *Science* 368:1377–81
150. Züst T, Heichinger C, Grossniklaus U, Harrington R, Kliebenstein DJ, Turnbull LA. 2012. Natural enemies drive geographic variation in plant defenses. *Science* 338:116–19

140. Illustrates the exploratory power of a broadly applicable and simple MS/MS scoring approach, providing a foundation for spectral molecular networking.

149. Applies information theory to decode the volatile-mediated chemical communication arms race between plants and herbivores and to test the information transfer hypothesis.