Trends in Plant Science

CellPress REVIEWS

Review

Exploring the Diversity of Plant Metabolism

Chuanying Fang,[1] Alisdair R. Fernie,[2,3,*] and Jie Luo[1,4,*]

Plants produce a huge array of metabolites, far more than those produced by most other organisms. Unraveling this diversity and its underlying genetic variation has attracted increasing research attention. Post-genomic profiling platforms have enabled the marriage and mining of the enormous amount of phenotypic and genetic diversity. We review here achievements to date and challenges remaining that are associated with plant metabolic research using multi-omic strategies. We focus mainly on strategies adopted in investigating the diversity of plant metabolism and its underlying features. Recent advances in linking metabotypes with phenotypic and genotypic traits are also discussed. Taken together, we conclude that exploring the diversity of metabolism could provide new insights into plant evolution and domestication.

Highlights

Plants produce a huge array of metabolites in spatiotemporal- and/or environment-dependent manner, which not only make it a challenge to understand plant metabolic diversity but also render plants ideal models for identifying metabolites and dissecting metabolic pathways.

In addition to reverse genetic approaches, forward genetic-based approaches combining genome sequences with population genetics provide clues for understanding biological mechanisms.

Genomic evolution provides the genetic basis for metabolic diversity, including gene duplication, gene loss, transposon insertion, and the evolution of substrate preference. Selective events during crop domestication and improvement have also played a vital role in the evolution of metabolism.

Analysis of the metabolome in genetically diverse populations can also facilitate the dissection of phenotypic traits, and will eventually lead to metabolite-assisted breeding of crops.

Simple Beginning: The Significance of Plant Metabolism

Sessile in nature, plants collectively produce a vast array of **metabolites** (see Glossary) with estimates ranging from 100 000 to 1 million, and many of these compounds are thought to play essential roles in resistance and tolerance to biotic and abiotic stresses, respectively [1–4]. Any single plant species only produces a subset of these metabolites, and current estimates range from 5000 to tens of thousands [5]. In addition, recent research has revealed that the extents of qualitative and quantitative variation of metabolism within a species are much larger than had previously been assumed [6]. Plant metabolites play vital roles in growth, cellular replenishment, and whole-plant resource allocation, as well as in adaptation of plants to a constantly changing environment [7,8]. Hence, the **metabolome** is consequently often regarded as the 'readout' of the physiological status and the bridge between the genotype and the phenotype of a plant [9–11]. In addition, natural products synthesized in plants provide indispensable resources for human health and survival. In excess of 30% of our drugs are sourced directly from plants, >60% of the drugs introduced in the past 20 years are based on plant extracts or their close derivatives [3]. Given the importance of plant metabolism to plant development and adaptation, and for human health, numerous studies have been performed to decipher the genetic regulation of plant metabolism [1–7]. Recently, the development of broad profiling approaches such as genomics, transcriptomics, and **metabolomics** has aided exploration of the diversity of plant metabolism as well as the underlying molecular mechanisms by which the plant cell controls its own chemical composition [6]. We review here our current understanding of the biochemistry and genetics which underlie the massive diversity apparent in plant metabolism, and provide a perspective on how this likely evolved. We also comment on the linkage between metabolite levels and morphological phenotypes, and conclude by outlining future challenges that need to be addressed. Specifically, we focus on how and why this extraordinary level of diversity has arisen, and why it currently persists or even expands.

Genetic and Biochemical Insights into Plant Metabolism

The Diversity of Plant Metabolism

Although most **primary metabolites** constitutively accumulate in plant cells [8], the majority of specialized metabolites are only detected in defined species, within particular tissues/organs,

[1]Hainan Key Laboratory for Sustainable Utilization of Tropical Bioresource, Institute of Tropical Agriculture and Forestry, Hainan University, Haikou 570288, China
[2]Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm 144776, Germany
[3]Center of Plant System Biology and Biotechnology, 4000 Plovdiv, Bulgaria
[4]National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China

*Correspondence:
fernie@mpimp-golm.mpg.de (A.R. Fernie) and
jie.luo@hainu.edu.cn (J. Luo).

at given developmental stages, or under specific environmental conditions. There are legion examples of the specificity of specialized metabolism: for example, glucosinolate defense compounds are largely confined to the *Brassicaceae* [12,13], and acyl sugar production is largely specific to glandular trichromes [14], whereas alkaloid-, terpene-, and phenlypropanoid-derived metabolites are much more widely spread [15–17]. Among the latter, trihydroxycinnamoyl spermidine derivatives were initially found to be synthesized in *Arabidopsis thaliana* and to accumulate in the pollen coat [18]. Moreover, although trihydroxycinnamoyl spermidine conjugates represented the major forms of spermidine derivatives in the pollen coat of most eudicots, they were undetectable in monocots [18]. That said, this class of compounds is illustrative of how recent research has revealed that both qualitative and quantitative variation of metabolism is much larger than had previously been thought. Leaves of Zhonghua11, a typical *japonica* rice accession display high levels of *N*′,*N*″-diferuloyl spermidine, whereas typical *indica* accessions do not [19]. Moreover, this accumulation in *japonica* follows a developmental gradient, being very high in young leaves but declining dramatically during leaf development [19] (Figure 1A). Furthermore, various metabolites display environmentally induced

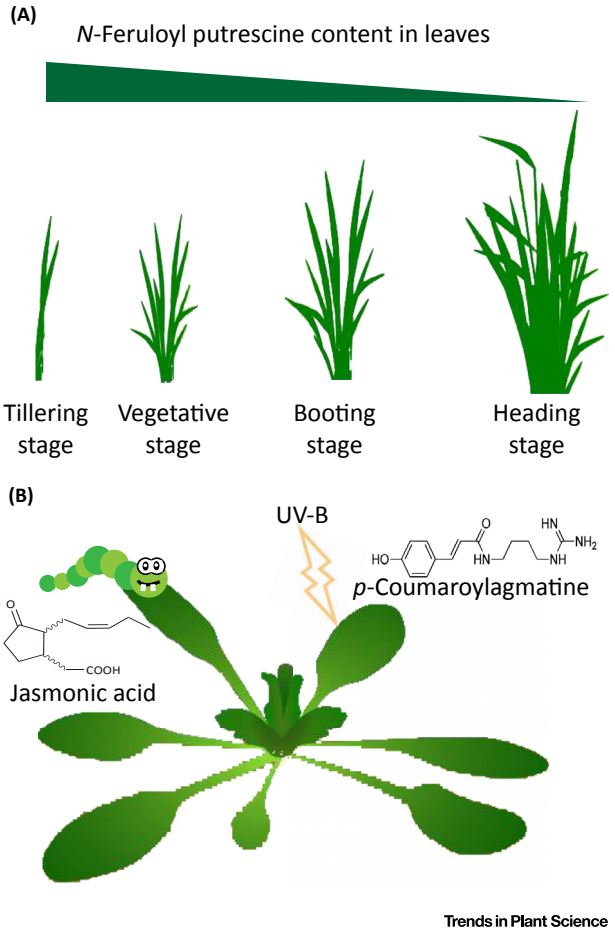

Figure 1. The Diversity of Plant Metabolism. (A) *N*′,*N*″-diferuloyl spermidine accumulation in *japonica* follows a developmental gradient with very high levels in young leaves but declining dramatically during leaf development [19]. (B) Metabolites display environmentally induced accumulation. Plants irradiated by UV-B light produce dramatically more *p*-coumaroylagmatine [20], whereas herbivory commonly leads to a burst in biosynthesis of jasmonic acid and its derivatives [21].

Glossary

Ancestral protein resurrection: advances in phylogenetics and DNA synthesis techniques have made it possible to infer the sequences of ancestral genes and then synthesize and express them in the laboratory. As a result, hypotheses about the functions of ancient genes – and the mechanistic basis for their evolution – can now be empirically tested using the reductionist power of experimental molecular biology.

Gene cluster: a gene family is composed of several genes which share similar features. A gene cluster is part of a gene family, and is a group of two or more genes in the DNA of an organism that encode for similar polypeptides, or proteins, which collectively share a generalized function and are located within a few kb of each other. The size of gene clusters can vary significantly, from a few genes to several hundred genes. Portions of the DNA sequence of each gene within a gene cluster are found to be identical; however, the protein encoded by each gene is distinct from the protein encoded by another gene within the cluster. Genes found in a gene cluster may be near one another on the same chromosome or on different but homologous chromosomes. Because of DNA sequence homology, the presence of gene clusters on the same chromosome suggests a close evolutionary relationship between two species. Therefore, a gene cluster may be used to assess the evolutionary relationship between organisms.

Metabolite: metabolites are the intermediates and products of metabolism. The term metabolite is usually restricted to small molecules. Metabolites have various functions, including fuel, structure, signaling, stimulatory and inhibitory effects on enzymes, catalytic activity of their own (usually as a cofactor to an enzyme), defense, and interactions with other organisms.

Metabolome: the complete set of small-molecule chemicals found within a biological sample.

Metabolomics: the scientific study of chemical processes involving metabolites. Specifically, metabolomics is the 'systematic study of the unique chemical

accumulation across species. For instance, plants irradiated by UV-B light produce dramatically more *p*-coumaroylagmatine [20], whereas herbivory commonly leads to a burst in biosynthesis of jasmonic acid and its derivatives [21] (Figure 1B). These examples highlight the dynamism of metabolism, but should also act as a cautionary note that it is dangerous to infer the specificity of metabolite production from studies that do not take into account analyses which include comprehensive tissue, developmental, and environmental components. That said, the fact that we now have access to a huge number of genomes of the green lineage [22] provides powerful resources for assessing the likely presence of metabolic pathways, as illustrated by recent surveys of shikimate, phenylpropanoid, and acyl sugar metabolism [17,23,24]. Similar surveys have also described the enormous diversity in metabolite transport mechanisms (Box 1). However, as we detail below, such studies ultimately need to be supported by detection of the metabolites in question to provide functional evidence that gene paralogs have maintained their initial function.

Dissection of the Genetic and Biochemical Bases of Plant Metabolic Diversity

In studying the diversity of plant metabolism it is important to elucidate how each metabolite is synthesized, transported, and degraded, and how these processes are regulated. Tremendous progress has been achieved using reverse genetic approaches to elucidate the biosynthesis and control of *Arabidopsis* metabolite accumulation, especially concerning **secondary metabolites** in this species [25]. Indeed, advances in profiling technologies have enabled analysis of the variation of metabolites both between species and within natural accessions of a single species [25,26]. Moreover, population genetic studies integrating metabolic profiling and quantitative genetics have begun to reveal the genetic regulation of the metabolome in both model and crop species [27–29]. Candidate genes responsible for the accumulation of specific secondary metabolites can be identified through correlative analysis of various omic datasets, including genomics, transcriptomics, proteomics, and metabolomics, combined with the use of forward and reverse genetics. This strategy is further facilitated by recent advances in next-generation sequencing technology.

Box 1. Structural Features Underpinning Plant Metabolic Diversity

When plant metabolic diversity is assessed through the lens of the genomes which guide it, considerable genomic architecture becomes apparent. This is evidenced by the frequent observation of metabolic quantitative trait loci (mQTL) hotspots both of multiple compound classes and within multiple species [31,35,37,47,115]. Moreover, tandem genes are over-represented in such hotspots [35,45,86]. However, it can additionally be seen merely in the genome structure itself. That is because a large number of specialized metabolic pathways are controlled by regulon-like gene clusters [62,116], although the genes encoding the enzymes of many pathways are randomly scattered throughout the genome. We focus our discussion here on the largest class of specialized metabolites – the terpenoids – detailing structural variations such as copy-number variations and presence/absence variations that are associated with chemodiversity. Several gene clusters relating to terpenoids have been reported, including those for arabidiol synthase in *Arabidopsis thaliana* [117] and ingenol mebutate in the Euphorbiaceae [118]. Zerbe and Böhlmann attempted a detailed phylogenic annotation of diterpene synthase gene functions, concluding that it was very difficult to faithfully assign gene function [119]. That said, given the ever-increasing volume of functional studies and species for which sequence data are available, this may prove easier and facilitate studies of their evolution in the future. Two recent studies elegantly demonstrate how such data enhance our understanding of evolution. In the first, investigations of the primary drivers of diversity, namely terpenoid synthase (TS) and cytochrome P450s (CYPs), identified different evolutionary routes in which either (i) microsyntenic blocks of TS/CYPs duplicate and provide templates for the evolution of new pathways, or (ii) new pathways arise by mixing and matching of individual TS and CYPs [15]. The second study used genome mining to identify terpene synthase and prenyltransferase gene pairs, followed by structural modeling studies to identify the residue responsible for the structural variation of the terpenoids across the Brassicaceae [120]. This study additionally suggests convergent evolution of plant and fungal sesqueterpene synthases, and suggests that the colocalized terpene synthase and prenyltransferase gene pairs likely originated from a common ancestral gene pair present before speciation. Asides from these detailed examples, two recent computational approaches aiming to identify gene clusters in genomes have provided important hints for the evolution of many more facets of plant secondary metabolism [121,122].

fingerprints that specific cellular processes leave behind' – the study of their small-molecule metabolite profiles.

Primary metabolites: a type of metabolite that is directly involved in normal growth, development, and reproduction. It usually performs a physiological function in the organism (i.e., has an intrinsic function). A primary metabolite is typically present in many organisms or cells. It is also referred to as a central metabolite, which has an even more restricted meaning (present in all autonomously growing cells or organisms). Common examples of primary metabolites include ethanol, lactic acid, and particular amino acids.

Secondary metabolites: organic compounds that are not directly involved in the normal growth, development, or reproduction of an organism. These are also referred to as specialized metabolites, often restricted to a narrow set of species within a phylogenetic group. Secondary metabolites often play an important role in plant defense against herbivory and other interspecies defenses.

Selective sweep: the reduction or elimination of variation among the nucleotides in neighboring DNA of a mutation as the result of recent and strong positive natural selection. A selective sweep can occur when a new mutation occurs that increases the fitness of the carrier relative to other members of the population. Natural selection favors individuals that have a higher fitness, and with time the newly mutated variant (allele) will increase in frequency relative to other alleles. As its prevalence increases, neutral and nearly neutral genetic variation linked to the new mutation will also become more prevalent. This phenomenon is called genetic hitchhiking. A strong selective sweep results in a region of the genome where the positively selected haplotype (the mutated allele and its neighbors) is essentially the only haplotype in the population, resulting in a large reduction in the total genetic variation in that chromosomal region.

For example, Sadre *et al*. recently identified two essential genes for camptothecin biosynthesis in *Camptotheca acuminate* by analyzing transcriptome and metabolome data combined with reverse genetics [30]. Being a monoterpene indole alkaloid, the biosynthesis of camptothecin is dependent on conversion of 8-oxogeranial to iridodial, and of tryptophan to tryptamine. Metabolic profiling revealed that tryptamine only accumulated in stem, shoot apex, and young leaf tissues of *Camptotheca acuminate*, indicating tissue-specific expression of the corresponding gene(s). There are two differentially expressed tryptophan decarboxylase genes, *TDC1* and *TDC2*, for the biosynthesis of tryptamine in *Camptotheca acuminate*. *TDC1* is expressed in tissues accumulating tryptamine, while *TDC2* is barely detectable. *CYCLASE 1* (*CYC1*), that functions in the conversion of 8-oxogeranial to iridodial, was additionally found to be coexpressed with *TDC1*, indicating that it might be an additional determinant of camptothecin biosynthesis. The *in vivo* function of *TDC1* and *CYC1* in camptothecin production were further validated by transgenic assays [30]. Combined analysis of transcriptome and metabolome data is a powerful tool to decipher genetic determinants of metabolic pathways, but lacks the power to unravel the genetic basis of natural variation in the metabolome. To access such information the analysis of natural variation using population genetics is now becoming widely adopted. This is usually investigated by linkage mapping, in other words quantitative trait locus (QTL) mapping using artificial breeding populations, and/or by genome-wide association studies (GWAS) using unrelated natural populations (Figure 2).

Recent advances in next-generation sequencing technology provide us with opportunities to identify metabolic quantitative trait loci (mQTLs) using ultra-high-density maps. For example, Gong *et al*. recently carried out mQTL mapping with an ultra-high-density map consisting of 1619 bins generated by population sequencing. Hundreds of mQTL in flag leaf or germinating seed were identified, with a significant deviation from a random distribution across the 12 chromosomes. A total of 44 and 16 potential mQTL 'hotspots' were identified in flag leaf and in germinating seed, respectively [27]. Tissue-specific accumulation of metabolites, especially secondary metabolites, is of special importance for the survival and adaptation of plant species. mQTL mapping with different tissues was able to decipher the divergent and convergent genetic regulation of metabolism across tissues. Comparative analysis of mQTLs of individual metabolite identified in two tissues revealed that the majority of QTLs are under different genetic control. Despite the overall tissue-specific regulation of metabolism, 23 loci for 19 metabolites were detected simultaneously in both tissues, suggesting considerable overlap of genetic control of metabolism between different tissues of rice [27], and a similar observation was also made for maize [31] and tomato [32].

Metabolic GWAS (mGWAS) in a broad number of plant species has shown that plant metabolism is generally moderately heritable [33,34], as would be anticipated for polygenic traits. For example, among the 840 metabolite features detected in rice leaf, >50% displayed broad-sense heritability >0.5, and >70% of the metabolic features displayed at least one significant association and an average of 4.9 associations per metabolite feature [35]. Complex traits, such as plant height and grain shape, are controlled by numerous loci of small effects [36], whereas metabolite content, especially of secondary metabolites, is generally determined by a small number of loci with large effects [35]. That said, natural variation in primary metabolites tends to more closely resemble the aforementioned physiological traits in being controlled by many loci of smaller effects [28,37]. A common feature in the genetic architecture of metabolism is the prevalence of hotspots of major genes/genome regions that determine the natural variation of large sets of metabolites [28,38]. More detailed evaluation in *Arabidopsis* revealed that some of the hotspots are within regions of the genome that were previously identified as being subject to recent strong positive selection (**selective sweeps**), and in regions showing

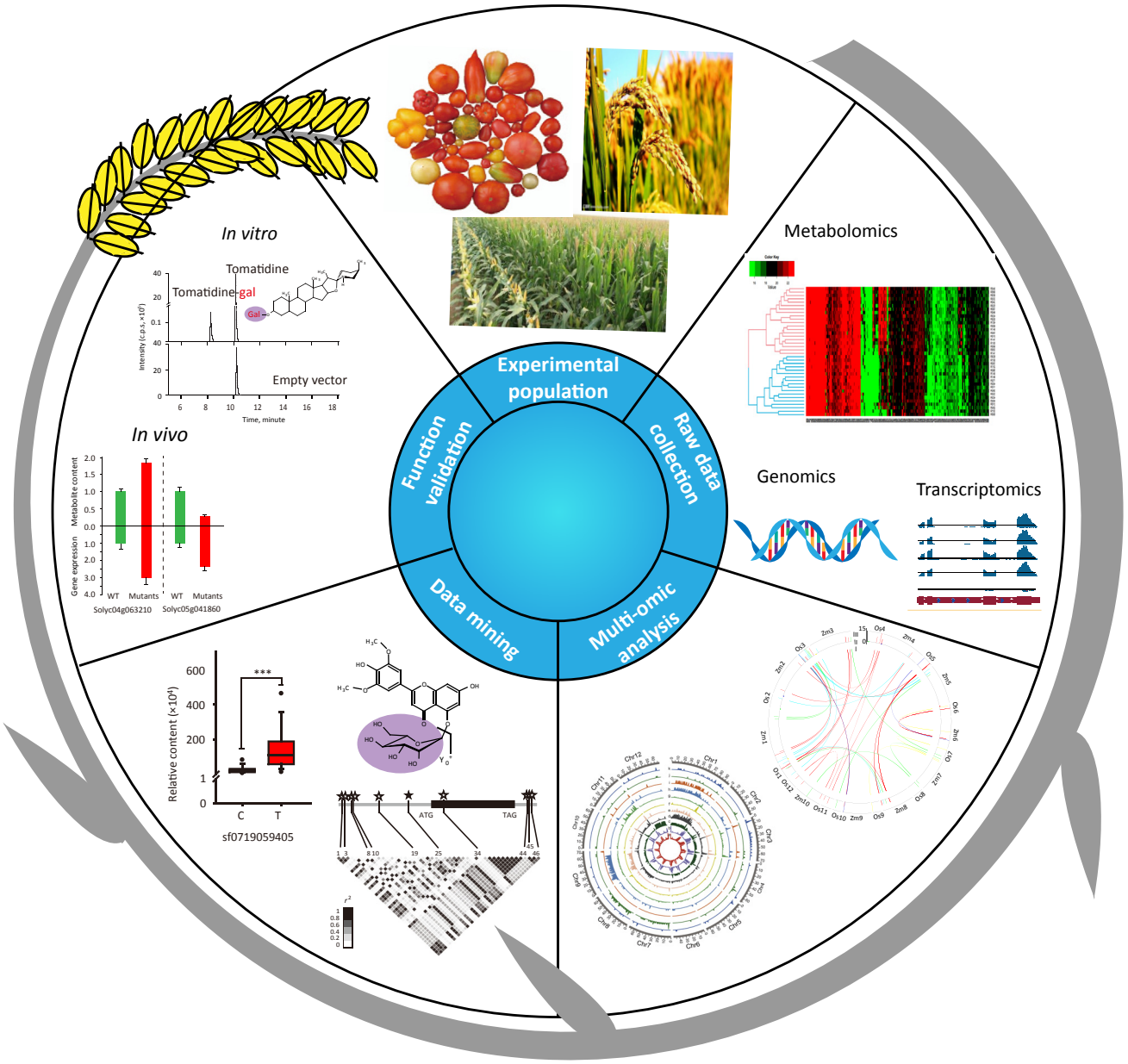

Figure 2. Metabolic GWAS (mGWAS)-Based Dissection of the Genetic Basis of the Plant Metabolome. An essential first step toward unveiling the genetic basis of plant metabolome via mGWAS is to collect suitable experimental populations. Subsequently, raw data of metabolome, genome, and transcriptome should be collected for further multi-omic analysis. Causative genes can be identified through intensive data mining and validated by *in vivo* and/or *in vitro* experiments.

trans-linkage to these putative sweeps, suggesting that selective forces have impacted on genome-wide control of *Arabidopsis* metabolism [28]. The effects of interactions between genotypes, environment, and development on the accumulation of secondary metabolites have been well documented within structured mapping populations. mGWAS on the naturally occurring variation of glucosinolate accumulation in *Arabidopsis* showed a significant bias toward identifying different causal genes for the glucosinolate phenotypes under different

developmental stages [29], which suggests that natural variation of glucosinolates is genetically controlled in a spatiotemporal manner. Interestingly, distinct genetic control of metabolism was also observed at subspecies level. mGWAS hotspots were located on different chromosomes in the *indica* and *japonica* subspecies of rice [35].

Correspondence among crop QTLs for agronomic performance has been characterized by comparative linkage mapping among crop plants such as wheat, maize, and rice [39]. To identify conserved regulators of metabolic trait(s) across species, the concept of comparative linkage mapping was modified and extended to mGWAS. Comparative mGWAS between rice and maize was performed by exploring the convergent genetic determinants for the metabolites detected in both species [40]. A total of 420 and 292 loci were detected for the 123 codetected metabolites in rice and maize, respectively. There were 42 homologous loci underlying the abundance of ~19% of codetected metabolites between the two species. Novel candidate genes for the codetected metabolites were subsequently identified in the comparative mGWAS. Taking advantage of the high resolution and SNP saturation in maize and rice mGWAS [31,41], Chen *et al*. performed comparative mGWAS and were therefore able to examine the convergent genetic loci that determine the natural variation of the same, or similar, metabolites [40]. It is important to note that this strategy is restricted *de facto* to the 'common' loci shared between plant species.

Association mapping by GWAS using a natural population is suitable for screening a large number of accessions for common variants within the population at relative high resolution [36], while linkage mapping using artificial populations such as recombinant inbred lines and introgression lines is likely to be more powerful in identifying alleles with low frequency or small effects in the population [42,43]. Joint linkage and association mapping has proved to be powerful not only in cross-validating results from one another but also in complementing each other in identifying new causative loci [44]. For example, association and linkage mapping of 11 phenolamides (PAs) was performed based on metabolic profiling in grain and leaf of rice [45]. Peng *et al*. identified significant associations for >80% of the 11 PAs, with an average of about three loci for each individual metabolite in leaf and grain. In general, those loci showed large effects, up to 43.1%, with an average of >12% in each tissue. This indicated that the levels of most PAs are controlled by a few major loci with large effects. To independently dissect the genetic basis of PA variation, biparental QTL analyses were performed with a recombinant inbred population generated from Zhenshan97 (ZS97) and Minghui63 (MH63) [27]. In total, seven significant chromosome regions for loci with LOD (logarithm of the odds) values >6.0 were identified, explaining 10.1–71.1% of the total variation in the population, and three of them were shared with the loci detected by GWAS. These loci showed, overall, relatively high resolution, ranging from 0.12 to 1.31 Mb, with an average of 0.54 Mb, possibly owing to the high resolution of the map generated by second-generation sequencing of the population [45]. Similarly, the combination of the two approaches has also been demonstrated to be highly informative in the study of primary metabolism in maize [46] and tomato [47].

Multidimensional analysis and multi-developmental stage analysis have been increasingly used to provide clues for understanding biological mechanisms because combining multiple datasets can compensate for missing or unreliable information in any single data type [48]. Metabolic profiling combined with transcriptome analysis has been utilized to dissect secondary metabolic pathways such as for steroidal glycoalkaloid (SGA), phenylpropanoid, and flavonoid biosynthesis, elucidating vital roles of **novel gene clusters** as well as of the GAME9 transcription factor [49,50]. Joint metabolomic and genomic data subsequently allowed a

comprehensive refinement of SGA biosynthesis [51]. In addition to the substantial inroads made in the targeted studies described above, global insights into metabolic regulation were also obtained. Multidimensional analysis of 100s of genomes, transcriptomes, and metabolomes was performed to provide new discovery leads for identifying genes controlling metabolic pathways in tomato fruits, including those for flavonoids and SGA metabolism. The overlap of mGWAS and expression QTL (eQTL) results generated >13 000 triple relationships (metabolite–SNP–gene), including 371 metabolites, 970 SNPs, and 535 genes [47]. This dataset thus facilitates both causal gene identification and metabolic pathway elucidation. For example, one mGWAS signal of the SGA hydroxytomatidenol (SlFM0964) was also supported by the eQTL of *Solyc03g118100*, an oxidoreductase gene that was previously reported to play an important role in SGA biosynthesis [47], indicating the power of this approach in evaluating metabolic control.

Furthermore, with the increasing number of genome sequences available from closely or more distantly related species, it is becoming possible to combine comparative genomic analysis with metabolomics in gene identification and pathway elucidation. Various types of cucurbitacins have been identified from cucurbit plants, such as cucurbitacin B from melon, cucurbitacin C from cucumber, and cucurbitacin E from watermelon [52]. A comparative genomic study revealed conserved genes encoding cytochrome P450s and acyltransferases for the biosynthesis of distinct cucurbitacins [53]. Comparative genomic analysis is thus able to unveil the genetic basis of the divergence of metabolite biosynthesis. Although abscisic acid (ABA) is present in numerous species, the conversion of ABA to phaseic acid and subsequently to dihydrophaseic acid has only been found in terrestrial plants. In a comparative genomic study, a seed plant-specific clade of DFR-like NAD(P)H-dependent reductases was identified, including *AtADH2*, a key regulator of phaseic acid accumulation [54].

The Evolution of Plant Metabolic Diversity

The study of the evolution of metabolism has blossomed in the past decade, having previously lagged considerably behind the evolution of development. This has been led primarily by advances in genome and exome sequencing [55,56], but latterly cross-species metabolic profiling has also significantly boosted this research field by providing functional evidence of gene neofunctionalization as well as for convergent and divergent evolution of metabolic functions [40,57]. Several mechanisms for the evolution of metabolism have been evidenced, with the most frequently described being (i) gene duplication and divergence, (ii) gene loss, and (iii) the evolution of substrate preference and promiscuity [23]. Local and whole-genome duplication (WGD) and subsequent sub- or neofunctionalization have contributed greatly to the metabolic diversity of land plants. Interestingly, gene duplication is far more prominent in plants than in other species [58], perhaps because, given their sessile nature, plant populations must be extremely adaptive to their environment. It is important to note that several isoforms exist for the enzymes in many of the major central metabolic pathways [59], and that for example the mitochondrial carrier family which catalyzes the transport of primary metabolites has also dramatically expanded by gene duplication [60], as has the number of sugar transporters [61]. However, gene duplications and gene clustering are considerably more prominently associated with plant secondary metabolism. Indeed many pathways for specialized metabolism in plants came about following duplication of genes of primary metabolism. Given that these aspects have been the subject of several previous reviews [62,63], in Box 1 we largely restrict ourselves to discussing the largest category of specialized metabolites, namely the terpenes, as well as providing a few recent examples in acyl sugar and phenylpropanoid metabolism.

Beyond these examples, recent studies on the evolution of nicotine biosynthesis [64] and the convergent evolution of caffeine in plants [65] provide interesting insight. The former was carried out via comparative analysis of two wild tobacco genomes, namely the considerably larger genome of *Nicotiana attenuata* and that of *Nicotiana obtusifolia*, which allowed the direct association of genome evolution with the assembly of the nicotine biosynthetic pathway. In doing so, the authors were able to conclude that both gene duplication and the insertion of transposable elements played an important role in the evolution of this chemical ecological trait [64]. The study concerning the evolution of caffeine followed up on earlier research of the same group that used enzymes of the salicylic acid/benzoic acid/theobromine (SABATH) family to demonstrate the potential of **ancestral protein resurrection**, in which non-preferred or even latent ancestral protein activities may be coopted at later times to become the primary or preferred protein activities [66] as an evolutionary driver of metabolic diversity. Application of this approach to the evolution of caffeine in plants revealed that convergent caffeine production surprisingly arose from two previously unknown biochemical pathways in coffee, tea, chocolate, citrus, and guarana plants. Furthermore, by resurrecting extinct enzymes that ancient plants once possessed, they revealed that the novel pathways would have evolved rapidly owing to cooption from their prior role to that in caffeine biosynthesis for which they were already primed [65].

Although the evolution of plant secondary/specialized metabolism is far too vast a subject to comprehensively review, there are four further areas of metabolism (in addition to the terpenoids detailed in Box 1) which provide important illustrations of how current experimental tools can provide insight into the mechanisms underlying the evolution of metabolism. Gene duplication and changes in substrate specificity are well characterized in the evolution of glucosinolate biosynthesis. Study of the enzyme crystal structures suggested that side-chain binding most likely underlies the difference in substrate specificity between the ancestral enzyme involved in leucine biosynthesis [67]. This hypothesis was further confirmed in glucosinolate biosynthesis by site-directed mutagenesis [67]. Similarly, a combination of sequence comparison and homology modeling demonstrated that the second enzyme of acylsugar biosynthesis in tomato is highly specific in cultivated tomato, but promiscuous in wild tomatoes, allowing the identification of the residue responsible for this difference [68]. Crystal structures of two 4-coumarate:CoA ligases also yielded insight into the evolution of their substrate preferences [69,70], and a recent study in *Arabidopsis* demonstrated that one of the four isoforms of the enzyme encoded in the genome catalyzes the formation of caffeoyl-CoA, and is thus important in syringyl lignin formation [70]. Finally, BAHD acyltransferases responsible for phenolamine biosynthesis have recently been characterized to display allelic variation for tissue specificity [45]. These combined examples illustrate that the evolution of metabolic novelty is driven by various factors, including changes in substrate promiscuity, enzyme activity, and gene expression, as well as by gene duplication [71].

Aside from the details of the evolution of specific pathways, however, it is important to understand the broad scope of the evolution of metabolism. This has been elegantly covered by Washburn *et al*. [63] who reviewed examples of convergent evolution across all domains of life, but they argue that the plant kingdom is particularly tractable for its study. They further state that the loss and retention of features of gene duplication, such as in the domestication syndrome, provide examples that strongly support this claim. In addition, the multiple appearances (and configurations) of C_4 and crassulacean acid metabolism (CAM) and hybrid vigor are highly pertinent aspects of plant evolution. Indeed, C_4 and CAM metabolism have been suggested to have evolved >60 times and at least 35 times, respectively [72,73], and considerable evidence links metabolic traits to hybrid vigor [74]. Washburn *et al*.

additionally highlight fascinating observations that aspects of the latter are conserved across the kingdoms of life [22], and that the Crabtree and Warburg effects of yeast and cancer cells, respectively, harbor considerable metabolic similarities [75,76], suggesting that the drive for efficient functionality means that convergent evolution has an amazingly broad scope. Staying at the general level, a recent modeling study evaluated *E. coli* metabolism with regard to what the authors refer to as 'diversity-generating biosynthesis', which they postulate evolved to produce large numbers of different metabolites [77]. With promiscuity increasing further down the specialized pathway studied, this study provides general principles for diversity-generating mechanisms underlying the expansion of specialized metabolism. It will thus be interesting to see how such models stand up against the vastness of plant secondary metabolism.

To conclude this section we discuss recent insight into the impact of crop domestication on metabolic diversity. Although a large number of exome studies have been performed in a range of crop species and their progenitors [55,78], so far very few studies have been conducted at the metabolite level in a manner that allows evaluation of the impact of the domestication syndrome on either metabolic diversity or the abundance of diverse metabolites [47,79]. Starting with the simpler study, deep evaluation of changes in primary metabolism revealed that a reduction in unsaturated fatty acids was evident during the (primary) domestication of emmer wheat, but that selection-driven changes in the amino acid content mark the domestication of durum wheat [79]. In the more extensive recent study, the impact of domestication on the fruit metabolome in tomato was investigated [47]. In this study the genomes, transcriptomes and metabolomes of between 399 and 610 genotypes were evaluated, and the question was addressed of how breeding has globally altered fruit chemical composition [47]. As can be seen in Figure 3, three independent selective events played a major role: (i) selection for larger fruits altered the metabolite profiles owing to linkage drag, (ii) selection for pink tomatoes preferred in the Asian market was associated with considerable metabolic changes, and (iii) introgression of resistance genes also resulted in major and unexpected changes. This study provided the first direct evaluation of the chemical compositional consequences of domestication at such a scale; however, as stated in the review by Giovannoni, it is highly likely that the findings are reflective of the mechanisms underpinning the evolution of metabolism in all our major crops [80].

Linking Metabolic Variation to Plant End-Phenotypes

In addition to elucidating the genetic and biochemical bases of plant metabolism, analysis of the metabolome in genetically diverse populations can also facilitate the dissection of phenotypic traits (Figure 4). At its simplest, the identification of common genomic regions which affect both metabolic and morphological traits can be established and physiological linkages determined [43]. Results from such studies in *Arabidopsis* and maize revealed considerable overlap between the levels of several primary metabolites and lignin precursors with biomass production [41,81]. Similarly, in a parallel QTL analysis in potato, loci for metabolites were found to be colocalized with those for starch- and cold sweetening-related traits [82]. This approach was taken a step further by evaluating correlation networks in a tomato introgression line population by subjecting the observation that amino acid content was negatively associated with the harvest index to experimental trials wherein the fruit load was experimentally manipulated [83]. More recently, combined studies on the metabolite composition of multiple tissues across broad metabolic populations have been carried out in both tomato and maize [31,32]. The study in tomato demonstrated that correlation between metabolites is affected by developmental stages [32], whereas that in maize revealed different genetic determinants of metabolic features across tissues [31]. These studies thus demonstrate the power of this approach to

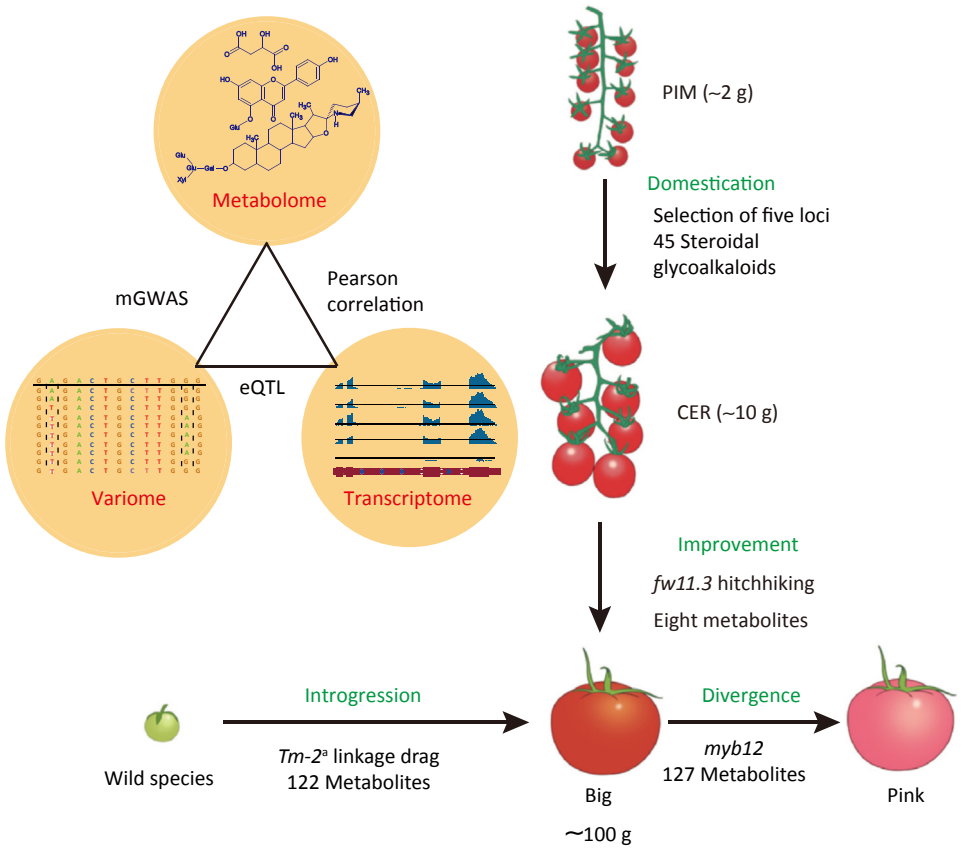

Figure 3. Evolution of Metabolic Diversity during Tomato Breeding. A multi-omic view of metabolic breeding history was proposed based on genomic, transcriptomic, and metabolic data from a large collection of tomato accessions [47]. The metabolome is shaped by three independent selective events: (i) selection for larger fruits altered the metabolite profiles as a result of linkage drag, (ii) selection for pink tomatoes preferred in the Asian market was associated with considerable metabolic changes, and (iii) introgression of resistance genes also resulted in major and unexpected changes. Abbreviations: BIG, *Solanum lycopersicum* group; CER, *lycopersicum* var. *cerasiforme* group; eQTL, expression quantitative trait locus; mGWAS, metabolic GWAS; PIM, *S. pimpinellifolium* group.

decipher physiological mechanisms with a higher resolution, which may aid in future metabolic engineering strategies.

However, perhaps surprisingly given the major insights recently obtained on source–sink interactions from a developmental perspective [84,85], relatively little research has been carried out on characterizing the metabolism of extreme natural variants in organ size. That said, considerable recent molecular analyses of natural variation have identified enzyme-encoding genes and have associated plant metabolism with variation in morphological and developmental traits. However, comparative genetic analyses of both metabolic and phenotypic traits at higher genetic resolution is necessary (especially for minor QTL) to allow better dissection of the causative factors underlying phenotypic traits [57,86–88]. That said, some powerful examples of the power of this approach come from studies demonstrating that metabolite profiling of young plants can offer good predictions as to their future growth potential [88], and in linking metabolism either to the clock [87] or to resistance to (a)biotic stress [57,86].

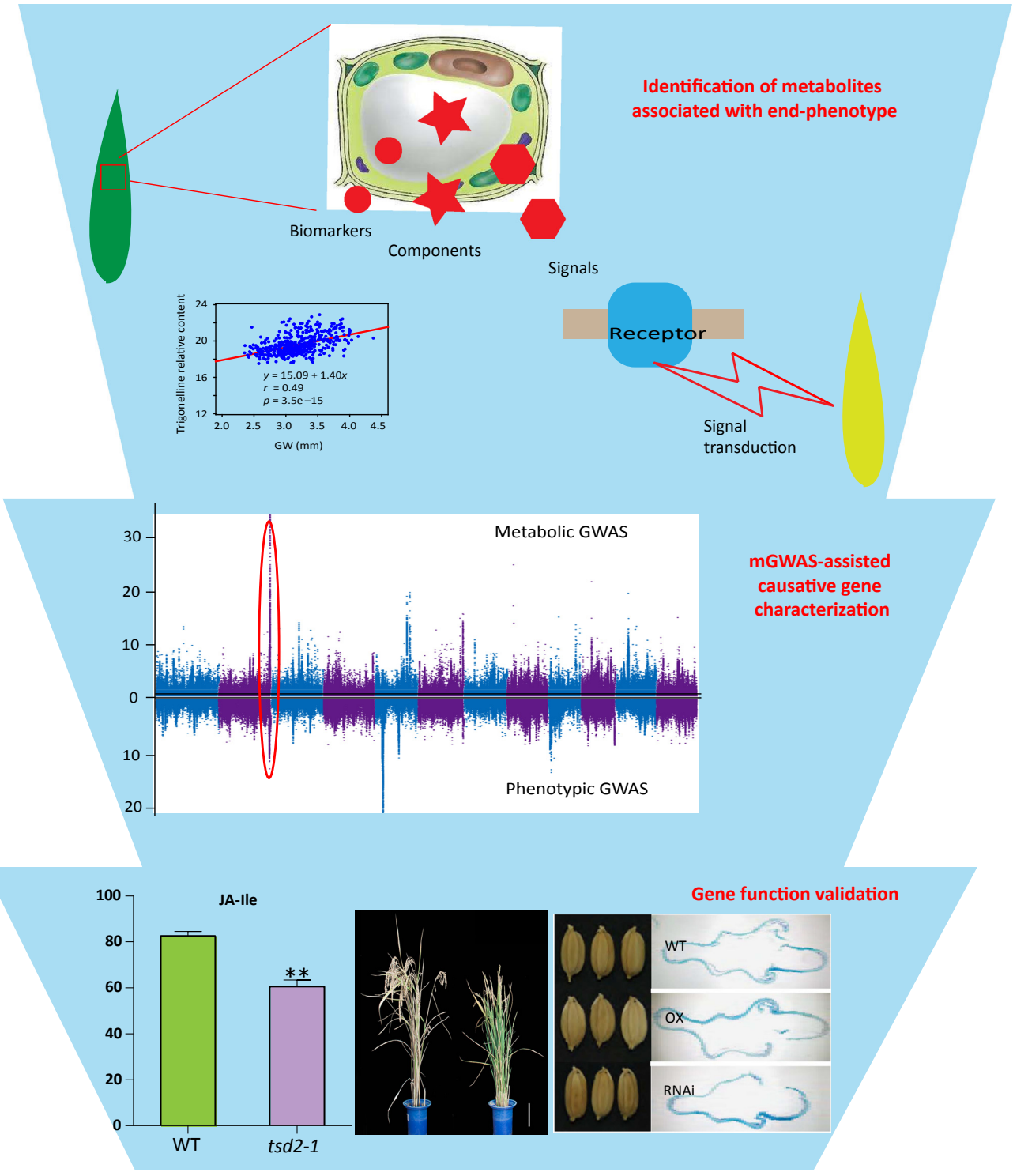

Figure 4. Metabolome-Facilitated Dissection of Phenotypic Traits. Three catalogs of metabolites contribute to facilitating the dissection of phenotypic traits, including compounds as components (represented by pentastars) or biomarkers (represented by circles) of phenotypic traits, and metabolites as signals leading to

(Figure legend continued on the bottom of the next page.)

Moreover, many recent studies have provided support for metabolite sensing, including considerable evidence that sugars, organic acids, and amino acids are sensed and that, once perceived, can dramatically impact on aspects of plant development [89,90]. Surprisingly, there is a lack of identified plant metabolite receptors which, by analogy to mammalian and microbial systems, would be activated to mediate such signaling [91]. This is arguably part of a more general problem facing the plant metabolic biologist – namely that of metabolite function. Although most metabolomic research is directed toward the more technical challenge of improving our coverage of the plant metabolome, the exact *in vivo* function of the majority of plant metabolites is poorly understood [91]. This fact is further compounded by the results of several recent studies indicating that a range of phytochemicals have roles additional to those traditionally ascribed to them. For example, the roles of flavonoids, in general, were demonstrated, in a suite of mutants of the core pathway, to be important both under conditions of oxidative and drought stress [92]. However, these studies were only able to draw general conclusions, and further studies will be necessary to dissect the specific quantitative contributions of the individual flavonoid species to the observed resistance. More precise identifications of function have been provided in the identification of modified flavonoids conferring UV tolerance [86] as well as metabolite-mediated defense [11]. However, in these cases it is not clear whether the defined function is the sole function of the metabolite or not. This caveat notwithstanding, our understanding of the metabolic events which are crucial in defining end-phenotypes has greatly benefited from the marriage of metabolomics, natural genetic diversity, and next-generation sequencing. With the increasing application of a wide range of profiling techniques to immortalized genetic populations, it would seem highly likely that further insights will be made in this research frontier that will be highly valuable for the rational design of future metabolic engineering strategies.

Outstanding Questions

How to detect and dissect diversity in transportation, storage, and even degradation of metabolites with multi-omic strategies?

How to unveil the biological functions of structually similar but distinct metabolites?

What are the effects of different modifications on the biochemical and biological functions of metabolites?

To date, causative genes underlying metabolic diversity identified by mGWAS mainly belong to structural genes and a few transcription factors. What could we do to comprehensively dissect metabolite pathways and characterize gene function in post-transcriptional, post-translational, and/or epigenetic manners?

How can we rationally design plant-based cell factories for the large-scale production of metabolites with commercial importance?

Concluding Remarks and Future Outlook

Past research has focused largely on enhancing our technical capacities to annotate ever more metabolites. This type of data evaluation was aided by several important recent developments, including the advent of high-resolution mass spectrometry (Box 2) and better computational methods. However, a far more exciting research front, towards the genetic control of metabolism, has been provided by the marriage of metabolomics, genetics, and next-generation sequencing. We have provided here a broad synthesis of recent advances with the aim of understanding how the immense metabolic diversity of the plant kingdom has evolved. Studies combining genome, transcriptome, and metabolome data clearly have immense potential in advancing our understanding of the metabolic networks underlying specialized metabolism. It is our opinion that genome and transcriptome data emanating from next-generation sequencing approaches need to be complemented more frequently with metabolomic data as well as with data from contemporary genetic techniques. Although the articles covered here have begun to address important questions, such as the elucidation of the structure of biosynthetic pathways, the impact of artificial selection on the metabolomes of our crop plants, and the link between the metabolome and end-phenotypes, many further questions remain to be tackled (see Outstanding Questions). Importantly, 20 years after the advent of metabolomics, the availability of an advanced toolkit now allows us to address these questions and expand our understanding of the forces driving plant metabolic diversity and the functionality that it confers to the host organism.

corresponding traits. Metabolite–trait interaction could be tested by various network or intercorrelation analyses. Parallel GWAS combined with metabolic GWAS and phenotypic GWAS or mGWAS could be subsequently performed to identify causative genes whose function in regulating metabolite accumulation and phenotype should be further validated by *in vivo* and/or *in vitro* experiments. Abbreviations: GW, grain width; JA-Ile, jasmonate-isoleucine; mGWAS, metabolic GWAS; OX, overexpression; WT, wild type.

Box 2. Mass Spectrometry (MS)-Based Metabolomics

The complexity of the metabolome has so far made it impossible to perform profiling of the entire metabolic complement with a single platform-based approach. We focus here on arguably the most developed platforms, namely MS-based analytical systems, which are attracting increasing attention owing to their high coverage, sensitivity, and resolution. MS-based metabolic profiling can be broadly categorized as being either untargeted or targeted. A general strategy for untargeted metabolome analysis is to characterize features that are different in the sample sets compared, and thereafter to elucidate their corresponding structures [93]. Current methods are frequently able to detect diverse mass signals in a single analysis and generate structural predictions [93,94]. Because untargeted studies are more challenging to interpret, more holistic and systematic approaches going beyond the current state of the art will be necessary to derive functional insights [95]. By contrast, in comparison to untargeted metabolomics [34,94,96], widely targeted metabolomics based on multiple reaction monitoring (MRM) is more sensitive and accurate. Current approaches define targets based on screening the samples using MRM conditions optimized from the available authentic standards, whereas endogenous metabolites within the samples are not subject to specific 'targeting'. Many efforts have been made to obtain MS/MS (MS2) acquisition for both known and unknown metabolites [97–100]. A novel liquid chromatography (LC)–MS2 method using multiple ion monitoring (MIM) as a survey scan to trigger the acquisition of enhanced product ions (EPI) has recently been developed [101]. This method is able to obtain MS2 acquisition for both known and unknown metabolites. This method was refined to develop a novel strategy called stepwise MIM–enhanced product ions (stepwise MIM–EPI) [102]. Integrating the data gathered from the resultant M22 spectral tag (MS2T) library and other available MRM information allows the quantification of hundreds of metabolites, and its use within GWAS has proved to be very powerful in dissecting the genetic and biochemical bases of metabolic diversity [35,40,47,103]. Mass accuracy and resolving power are vital for compound identification in MS-based metabolomic studies. Owing to the ultimate high resolution and high mass accuracy, Fourier transform ion cyclotron resonance MS (FT–ICR–MS) has been adopted for metabolome analysis in several species [104–107]. With the development of MS-based metabolomic studies, several MS2 databases are available, including METLIN [108], BinBase [109], HMDB [110], MMCD [111], MassBank [112], and ReSpect [113]. To effectively identify compounds with similar chromatographic behavior and UV absorption properties in plant metabolomic analyses, Lei *et al*. constructed an ultraperformance LC (UPLC)–MS/MS library [114].

Acknowledgments

Research in our laboratories was supported by the National Science Fund for Distinguished Young Scholars (grant 31625021), the State Key Program of National Natural Science Foundation of China (31530052), the Ministry of Science and Technology of the People's Republic of China (2016YFD0100500), and the Hainan University Startup Fund [KYQD (ZR)1866 to J.L., KYQD(ZR)1824 to C.F.], and the PlantaSYST project by the EU Horizon 2020 Research and Innovation Programme (SGA-CSA 664621 and 739582 under FPA 664620).

References

1. Afendi, F.M. *et al.* (2013) Data mining methods for omics and knowledge of crude medicinal plants toward big data biology. *Comput. Struct. Biotechnol. J.* 4, 1–14
2. Dixon, R.A. and Strack, D. (2003) Phytochemistry meets genome analysis, and beyond. *Phytochemistry* 62, 815–816
3. Rai, A. *et al.* (2017) Integrated omics analysis of specialized metabolism in medicinal plants. *Plant J.* 90, 764–787
4. Weng, J.K. (2013) The evolutionary paths towards complexity: a metabolic perspective. *New Phytol.* 201, 1141–1149
5. Fernie, A. *et al.* (2004) Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* 5, 763–769
6. Fernie, A.R. and Tohge, T. (2017) The genetics of plant metabolism. *Annu. Rev. Genet.* 51, 287–310
7. Obata, T. and Fernie, A.R. (2012) The use of metabolomics to dissect plant responses to abiotic stresses. *Cell. Mol. Life Sci.* 69, 3225–3243
8. Sulpice, R. and McKeown, P.C. (2015) Moving toward a comprehensive map of central plant metabolism. *Annu. Rev. Plant Biol.* 66, 187–210
9. Fang, C. *et al.* (2016) Control of leaf senescence by an MeOH-jasmonates cascade that is epigenetically regulated by OsSRT1 in rice. *Mol. Plant* 9, 1366–1378
10. Sanchez, D.H. *et al.* (2008) Metabolome–ionome–biomass interactions: what can we learn about salt stress by multiparallel phenotyping? *Plant Signal. Behav.* 3, 598–600
11. Yang, Q. *et al.* (2017) A gene encoding maize caffeoyl-CoA O-methyltransferase confers quantitative resistance to multiple pathogens. *Nat. Genet.* 49, 1364–1372
12. Agerbirk, N. and Olsen, C.E. (2012) Glucosinolate structures in evolution. *Phytochemistry* 77, 16–45
13. Grubb, C.D. and Abel, S. (2006) Glucosinolate metabolism and its control. *Trends Plant Sci.* 11, 89–100
14. Schilmiller, A.L. *et al.* (2008) Harnessing plant trichome biochemistry for the production of useful compounds. *Plant J.* 54, 702–711
15. Boutanaev, A.M. *et al.* (2015) Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci. U. S. A.* 112, E81–E88
16. Ober, D. and Hartmann, T. (2000) Phylogenetic origin of a secondary pathway: the case of pyrrolizidine alkaloids. *Plant Mol. Biol.* 44, 445–450
17. Tohge, T. *et al.* (2013) Shikimate and phenylalanine biosynthesis in the green lineage. *Front. Plant Sci.* 4, 62
18. Elejalde-Palmett, C. *et al.* (2015) Characterization of a spermidine hydroxycinnamoyltransferase in *Malus domestica* highlights the evolutionary conservation of trihydroxycinnamoyl spermidines in pollen coat of core eudicotyledons. *J. Exp. Bot.* 66, 7271–7285
19. Dong, X.K. *et al.* (2015) Spatiotemporal distribution of phenolamides and the genetics of natural variation of hydroxycinnamoyl spermidine in rice. *Mol. Plant* 8, 111–121

20. Kusano, M. *et al.* (2011) Metabolomics reveals comprehensive reprogramming involving two independent metabolic responses of *Arabidopsis* to UV-B light. *Plant J.* 67, 354–369
21. Wasternack, C. and Hause, B. (2013) Jasmonates: biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in Annals of Botany. *Ann. Bot.* 111, 1021–1058
22. Reddy, T.B. *et al.* (2015) The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* 43, D1099–D1106
23. Moghe, G.D. *et al.* (2017) Evolutionary routes to biochemical innovation revealed by integrative analysis of a plant-defense related specialized metabolic pathway. *eLife* 6, e28468
24. Tohge, T. *et al.* (2013) The evolution of phenylpropanoid metabolism in the green lineage. *Crit. Rev. Biochem. Mol. Biol.* 48, 123–152
25. Biais, B. *et al.* (2010) Metabolic acclimation to hypoxia revealed by metabolite gradients in melon fruit. *J. Plant Physiol.* 167, 242–245
26. Borevitz, J.O. *et al.* (2007) Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 12057–12062
27. Gong, L. *et al.* (2013) Genetic analysis of the metabolome exemplified using a rice population. *Proc. Natl. Acad. Sci. U. S. A.* 110, 20320–20325
28. Chan, E.K. *et al.* (2010) The complex genetic architecture of the metabolome. *PLoS Genet.* 6, e1001198
29. Chan, E.K. *et al.* (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLoS Biol.* 9, e1001125
30. Sadre, R. *et al.* (2016) Metabolite diversity in alkaloid biosynthesis: a multi-lane (diastereomer) highway for camptothecin synthesis in *Camptotheca acuminata*. *Plant Cell* 28, 1926–1944
31. Wen, W.W. *et al.* (2015) Genetic determinants of the network of primary metabolism and their relationships to plant performance in a maize recombinant inbred line population. *Plant Cell* 27, 1839–1856
32. Toubiana, D. *et al.* (2012) Metabolic profiling of a mapping population exposes new insights in the regulation of seed metabolism and seed, fruit, and plant relations. *PLoS Genet.* 8, e1002612
33. Matsuda, F. *et al.* (2015) Metabolome–genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *Plant J.* 81, 13–23
34. Matsuda, F. *et al.* (2012) Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. *Plant J.* 70, 624–636
35. Chen, W. *et al.* (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46, 714–721
36. Huang, X. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967
37. Rowe, H.C. *et al.* (2008) Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. *Plant Cell* 20, 1199–1216
38. Knoch, D. *et al.* (2017) Genetic dissection of metabolite variation in *Arabidopsis* seeds: evidence for mQTL hotspots and a master regulatory locus of seed metabolism. *J. Exp. Bot.* 68, 1655–1667
39. Ahn, S. and Tanksley, S.D. (1993) Comparative linkage maps of the rice and maize genomes. *Proc. Natl. Acad. Sci. U. S. A.* 90, 7980–7984
40. Chen, W. *et al.* (2016) Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat. Commun.* 7, 12767
41. Riedelsheimer, C. *et al.* (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc. Natl. Acad. Sci. U. S. A.* 109, 8872–8877
42. Fridman, E. *et al.* (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science* 305, 1786–1789
43. Schauer, N. *et al.* (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotechnol.* 24, 447–454
44. Luo, J. (2015) Metabolite-based genome-wide association studies in plants. *Curr. Opin. Plant Biol.* 24, 31–38
45. Peng, M. *et al.* (2016) Evolutionarily distinct BAHD N-acyltransferases are responsible for natural variation of aromatic amine conjugates in rice. *Plant Cell* 28, 1533–1550
46. Wen, W. *et al.* (2018) An integrated multi-layered analysis of the metabolic networks of different tissues uncovers key genetic components of primary metabolism in maize. *Plant J.* 93, 1116–1128
47. Zhu, G.T. *et al.* (2018) Rewiring of the fruit metabolome in tomato breeding. *Cell* 172, 249–261
48. Ritchie, M.D. *et al.* (2015) Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* 16, 85–97
49. Cardenas, P.D. *et al.* (2016) GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. *Nat. Commun.* 7, 10654
50. Itkin, M. *et al.* (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* 341, 175–179
51. Schwahn, K. *et al.* (2014) Metabolomics-assisted refinement of the pathways of steroidal glycoalkaloid biosynthesis in the tomato clade. *J. Integr. Plant Biol.* 56, 864–875
52. Shang, Y. *et al.* (2014) Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* 346, 1084–1088
53. Zhou, Y. *et al.* (2016) Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nat. Plants* 2, 16183
54. Weng, J.K. *et al.* (2016) Co-evolution of hormone metabolism and signaling networks expands plant adaptive plasticity. *Cell* 166, 881–893
55. Bellucci, E. *et al.* (2014) Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. *Plant Cell* 26, 1901–1912
56. Bolger, M.E. *et al.* (2014) Plant genome sequencing – applications for crop improvement. *Curr. Opin. Biotechnol.* 26, 31–37
57. Tohge, T. *et al.* (2016) Characterization of a recently evolved flavonol-phenylacyltransferase gene provides signatures of natural light selection in *Brassicaceae*. *Nat. Commun.* 7, 12399
58. Zhang, J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18, 292–298
59. Millar, A.H. *et al.* (2011) Organization and regulation of mitochondrial respiration in plants. *Annu. Rev. Plant Biol.* 62, 79–104
60. Palmieri, F. *et al.* (2011) Evolution, structure and function of mitochondrial carriers: a review with new insights. *Plant J.* 66, 161–181
61. Chen, L.Q. *et al.* (2015) Transport of sugars. *Annu. Rev. Biochem.* 84, 865–894
62. Boycheva, S. *et al.* (2014) The rise of operon-like gene clusters in plants. *Trends Plant Sci.* 19, 447–459
63. Washburn, J.D. *et al.* (2016) Convergent evolution and the origin of complex phenotypes in the age of systems biology. *Int. J. Plant Sci.* 177, 305–318
64. Xu, S.Q. *et al.* (2017) Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 114, 6133–6138
65. Huang, R. *et al.* (2016) Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes. *Proc. Natl. Acad. Sci. U. S. A.* 113, 10613–10618
66. Huang, R. *et al.* (2012) Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proc. Natl. Acad. Sci. U. S. A.* 109, 2966–2971
67. Leong, B. and Last, R. (2017) Promiscuity, impersonation and accommodation: evolution of plant specialized metabolism. *Curr. Opin. Struct. Biol.* 47, 105–112

68. Fan, P.X. *et al.* (2016) *In vitro* reconstruction and analysis of evolutionary variation of the tomato acylsucrose metabolic network. *Proc. Natl. Acad. Sci. U. S. A.* 113, E239–E248
69. Hu, Y. *et al.* (2010) Crystal structures of a *Populus tomentosa* 4-coumarate:CoA ligase shed light on its enzymatic mechanisms. *Plant Cell* 22, 3093–3104
70. Li, Y. *et al.* (2015) Four isoforms of *Arabidopsis thaliana* 4-coumarate:CoA ligase have overlapping yet distinct roles in phenylpropanoid metabolism. *Plant Physiol.* 169, 2409–2421
71. Blount, Z.D. *et al.* (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489, 513–518
72. Heyduk, K. *et al.* (2016) Evolution of a CAM anatomy predates the origins of crassulacean acid metabolism in the Agavoideae (Asparagaceae). *Mol. Phylogenet. Evol.* 105, 102–113
73. Sage, R.F. *et al.* (2012) Photorespiration and the evolution of C-4 photosynthesis. *Annu. Rev. Plant Biol.* 63, 19–47
74. Meyer, R.C. *et al.* (2012) Heterosis manifestation during early *Arabidopsis* seedling development is characterized by intermediate gene expression and enhanced metabolic activity in the hybrids. *Plant J.* 71, 669–683
75. Mayfield-Jones, D. *et al.* (2013) Watching the grin fade: tracing the effects of polyploidy on different evolutionary time scales. *Semin. Cell Dev. Biol.* 24, 320–331
76. Mordhorst, B.R. *et al.* (2016) Some assembly required: evolutionary and systems perspectives on the mammalian reproductive system. *Cell Tissue Res.* 363, 267–278
77. Tianero, M.D. *et al.* (2016) Metabolic model for diversity-generating biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 113, 1772–1777
78. Koenig, D. *et al.* (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc. Natl. Acad. Sci. U. S. A.* 110, E2655–E2662
79. Beleggia, R. *et al.* (2016) Evolutionary metabolomics reveals domestication-associated changes in tetraploid wheat kernel. *Mol. Biol. Evol.* 33, 1740–1753
80. Giovannoni, J. (2018) Tomato multiomics reveals consequences of crop domestication and improvement. *Cell* 172, 6–8
81. Lisec, J. *et al.* (2008) Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations. *Plant J.* 53, 960–972
82. Carreno-Quintero, N. *et al.* (2012) Untargeted metabolic quantitative trait loci analyses reveal a relationship between primary metabolism and potato tuber quality. *Plant Physiol.* 158, 1306–1318
83. Do, P.T. *et al.* (2010) The influence of fruit load on the tomato pericarp metabolome in a *Solanum chmielewskii* introgression line population. *Plant Physiol.* 154, 1128–1142
84. Sonnewald, U. and Fernie, A.R. (2018) Next-generation strategies for understanding and influencing source–sink relations in crop plants. *Curr. Opin. Plant Biol.* 43, 63–70
85. Soyk, S. *et al.* (2017) Bypassing negative epistasis on yield in tomato imposed by a domestication gene. *Cell* 169, 1142–1155 e12
86. Peng, M. *et al.* (2017) Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. *Nat. Commun.* 8, 1975
87. Kerwin, R.E. *et al.* (2011) Network quantitative trait loci mapping of circadian clock outputs identifies metabolic pathway-to-clock linkages in *Arabidopsis*. *Plant Cell* 23, 471–485
88. Lipka, A.E. *et al.* (2015) From association to prediction: statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* 24, 110–118
89. Hausler, R.E. *et al.* (2014) Amino acids – a life between metabolism and signaling. *Plant Sci.* 229, 225–237
90. Wahl, V. *et al.* (2013) Regulation of flowering by trehalose-6-phosphate signaling in *Arabidopsis thaliana*. *Science* 339, 704–707
91. Alseekh, S. and Fernie, A.R. (2018) Metabolomics 20 years on: what have we learned and what hurdles remain? *Plant J.* 94, 933–942
92. Yonekura-Sakakibara, K. *et al.* (2014) A flavonoid 3-O-glucoside:2-O-glucosyltransferase responsible for terminal modification of pollen-specific flavonols in *Arabidopsis thaliana*. *Plant J.* 79, 769–782
93. Bottcher, C. *et al.* (2008) Metabolome analysis of biosynthetic mutants reveals a diversity of metabolic changes and allows identification of a large number of new compounds in *Arabidopsis*. *Plant Physiol.* 147, 2107–2120
94. Wu, S. *et al.* (2018) Mapping the *Arabidopsis* metabolic landscape by untargeted metabolomics at different environmental conditions. *Mol. Plant* 11, 118–134
95. Sawada, Y. *et al.* (2009) Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant Cell Physiol.* 50, 37–47
96. Hu, C. *et al.* (2014) Metabolic variation between *japonica* and *indica* rice cultivars as revealed by non-targeted metabolomics. *Sci. Rep.* 4, 5067
97. Chen, S. *et al.* (2013) Pseudotargeted metabolomics method and its application in serum biomarker discovery for hepatocellular carcinoma based on ultra high-performance liquid chromatography/triple quadrupole mass spectrometry. *Anal. Chem.* 85, 8326–8333
98. Chen, Y. *et al.* (2017) Development of a data-independent targeted metabolomics method for relative quantification using liquid chromatography coupled with tandem mass spectrometry. *Anal. Chem.* 89, 6954–6962
99. Gu, H. *et al.* (2015) Globally optimized targeted mass spectrometry: reliable metabolomics analysis with broad coverage. *Anal. Chem.* 87, 12355–12362
100. Luo, P. *et al.* (2015) Multiple reaction monitoring-ion pair finder: a systematic approach to transform nontargeted mode to pseudotargeted mode for metabolomics study based on liquid chromatography–mass spectrometry. *Anal. Chem.* 87, 5050–5055
101. Yao, M. *et al.* (2008) Rapid screening and characterization of drug metabolites using a multiple ion monitoring-dependent MS/MS acquisition method on a hybrid triple quadrupole–linear ion trap mass spectrometer. *J. Mass Spectrom.* 43, 1364–1375
102. Chen, W. *et al.* (2013) A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. *Mol. Plant* 6, 1769–1780
103. Matsuda, F. *et al.* (2009) MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. *Plant J.* 57, 555–577
104. Aharoni, A. *et al.* (2002) Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS* 6, 217–234
105. Giavalisco, P. *et al.* (2011) Elemental formula annotation of polar and lipophilic metabolites using ^{13}C, ^{15}N and ^{34}S isotope labelling, in combination with high-resolution mass spectrometry. *Plant J.* 68, 364–376
106. Kueger, S. *et al.* (2012) High-resolution plant metabolomics: from mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. *Plant J.* 70, 39–50
107. Pollier, J. *et al.* (2013) The protein quality control system manages plant defence compound synthesis. *Nature* 504, 148–152
108. Smith, C.A. *et al.* (2005) METLIN – a metabolite mass spectral database. *Ther. Drug Monit.* 27, 747
109. Fiehn, O. *et al.* (2008) Quality control for plant metabolomics: reporting MSI-compliant studies. *Plant J.* 53, 691–704
110. Wishart, D.S. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* 37, D603–D610
111. Cui, Q. *et al.* (2008) Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.* 26, 162–164

112. Horai, H. *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* 45, 703–714

113. Sawada, Y. *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

114. Lei, Z. *et al.* (2015) Construction of an ultrahigh pressure liquid chromatography-tandem mass spectral library of plant natural products and comparative spectral analyses. *Anal. Chem.* 87, 7373–7381

115. Alseekh, S. *et al.* (2015) Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. *Plant Cell* 27, 485–512

116. Fernie, A.R. and Tohge, T. (2015) Location, location, location – no more! The unravelling of chromatin remodeling regulatory aspects of plant metabolic gene clusters. *New Phytol.* 205, 458–460

117. Sohrabi, R. *et al.* (2015) In planta variation of volatile biosynthesis: an alternative biosynthetic route to the formation of the pathogen-induced volatile homoterpene DMNT via triterpene degradation in *Arabidopsis* roots. *Plant Cell* 27, 874–890

118. King, A.J. *et al.* (2014) Production of bioactive diterpenoids in the euphorbiaceae depends on evolutionarily conserved gene clusters. *Plant Cell* 26, 3286–3298

119. Zerbe, P. and Bohlmann, J. (2015) Plant diterpene synthases: exploring modularity and metabolic diversity for bioengineering. *Trends Biotechnol.* 33, 419–428

120. Huang, A.C. *et al.* (2017) Unearthing a sesterterpene biosynthetic repertoire in the Brassicaceae through genome mining reveals convergent evolution. *Proc. Natl. Acad. Sci. U. S. A.* 114, E6005–E6014

121. Kautsar, S.A. *et al.* (2017) PlantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* 45, W55–W63

122. Topfer, N. *et al.* (2017) The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.* 45, 7049–7063