Unlocking plant metabolic diversity: A (pan)-genomic view

Xuan Zhou1,2 and Zhenhua Liu1,2,*

1Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China
2Shanghai Collaborative Innovation Center of Agri-Seeds, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China

*Correspondence: Zhenhua Liu (zhenhua.liu@sjtu.edu.cn)

https://doi.org/10.1016/j.xplc.2022.100300

ABSTRACT

Plants produce a remarkable diversity of structurally and functionally diverse natural chemicals that serve as adaptive compounds throughout their life cycles. However, unlocking this metabolic diversity is significantly impeded by the size, complexity, and abundant repetitive elements of typical plant genomes. As genome sequencing becomes routine, we anticipate that links between metabolic diversity and genetic variation will be strengthened. In addition, an ever-increasing number of plant genomes have revealed that biosynthetic gene clusters are not only a hallmark of microbes and fungi; gene clusters for various classes of compounds have also been found in plants, and many are associated with important agronomic traits. We present recent examples of plant metabolic diversification that have been discovered through the exploration and exploitation of various genomic and pan-genomic data. We also draw attention to the fundamental genomic and pan-genomic basis of plant chemodiversity and discuss challenges and future perspectives for investigating metabolic diversity in the coming pan-genomics era.

Key words: metabolic diversity, gene and genome duplication, biosynthetic gene clusters, genomic structural variations, pan-genomics

Zhou X. and Liu Z. (2022). Unlocking plant metabolic diversity: A (pan)-genomic view. Plant Comm. 3, 100300.

INTRODUCTION

A hallmark of plants is their ability to synthesize an extraordinarily large number (~0.2–1 million) of diverse small molecules (Dixon and Strack, 2003; Fang et al., 2019). Although a small portion of these serve as primary metabolites that are essential for plant growth and reproduction, the majority are recognized as secondary or specialized metabolites (SMs) that play important ecological roles in plant defense against various biotic and abiotic stresses (Hartmann, 2007; Chae et al., 2014; Weng, 2014). In addition, although the functions of SMs were initially thought to be “secondary,” they can sometimes play important “primary” roles, as in the regulation of plant growth and development (Erb and Klübenstein, 2020). From a biochemical viewpoint, plant SMs are mainly classified as phenylpropanoids, terpenoids, and alkaloids. Different classes of compounds can also be included, adding enormous chemical diversity to plant repertoires. Other types of plant SMs, such as unusual peptides and highly modified fatty acids, have drawn increasing attention in recent years (Kersten and Weng, 2018; Jeon et al., 2020).

SMs are often found in limited groups of plant lineages or clades, and they can therefore reveal how plants have adapted to various ecological niches. For example, the sulfur-containing glucosinolate defense compounds are found exclusively in the Brassicales (which evolved roughly 90 million years ago), and their diversification is thought to reflect co-evolution with butterflies in an evolutionary arms-race (Edger et al., 2015). Thalianin (T10), a triterpene found only in the model plant Arabidopsis thaliana, has been shown to play important roles in modulating root growth and shaping a healthy rhizosphere bacterial community (Huang et al., 2019; Bai et al., 2021). In addition to the large number of metabolites synthesized by plants and their patchy distributions across phylogenetic taxa, plant metabolic diversity is also reflected in metabolic variation within plant organs and developmental stages. This variation may be even greater when examined at the level of specific cell types, single cells, or organelles (Misra et al., 2014; de Souza et al., 2020). Although the obstacles mentioned above challenge our ability to dissect and understand plant metabolic diversity, combinatorial approaches, including comparative genomics and transcriptomics, appear to offer powerful strategies for revealing the DNA fingerprints that lead to metabolite divergence in related species (Fernie and Tohge, 2017; van der Hoof et al., 2020). For example, comparing the genomes of the sister species A. thaliana and Arabidopsis lyrata has revealed...
that pathway genes for the biosynthesis of thalianin (T10) are also present in the *A. lyrata* genome. However, *A. lyrata* lacks two additional reductase genes, resulting in the production of the new triterpenoid epi-thalianin (T17, an epimer of thalianin) in its roots (Liu et al., 2020a).

Genome sequencing is now becoming routine, revolutionizing our interpretation of the genomic basis of metabolic diversification in a wide range of plant taxa. Conventionally, gene and genome duplications are the main types of genomic variants observed in surveys of related plant genomes (Moore and Purugganan, 2005; Lichman et al., 2020). Genetic drift and natural selection then drive the retention and diversification of these genetic foundations (Ohno, 1970; Panchy et al., 2016), translating into a spectacular diversity of SMs among plant species. In addition, re-sequencing and de novo assembly of representative accessions within a species (referred to as pan-genomics) have enabled the identification of other, underexplored genomic structural variations, such as presence–absence variations and gene copy-number variations (Ho et al., 2020), enabling us to uncover the molecular mechanisms that underlie plant metabolic diversity with greater resolution. Pan-genomic variants found at the intra-species level are more evolutionarily recent than genomic variants found in inter-species comparisons. Therefore, the investigation of both genomic and pan-genomic variations associated with a metabolic trait can shed light on the comprehensive evolutionary history of variations in plant chemical profiles.

Here, we review recent advances linking phytochemical innovations to genetic variations at both shallow (genomic) and deep (pan-genomic) phylogenetic scales, with a particular emphasis on genome structural variations, some of which have only recently become available following the construction of plant pan-genomes. Another key development in the field of specialized metabolism (BGCs) that serve as the genomic basis for various types of natural products (Polturak and Osbourn, 2021). Metabolic diversity generated by diversification of gene clusters will therefore also be covered. Finally, we summarize and discuss challenges and perspectives on genomic and pan-genomic variation and their connections to the pan-metabolomes they encode.

**CLASSICAL DRIVERS OF METABOLIC DIVERSITY: GENE AND GENOME DUPLICATION**

Since the first plant genome was sequenced from the model plant *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000), the genomes of several hundred plant species have been sequenced and published according to the National Center for Biotechnology Information, although they may differ in sequencing depth and assembly quality. These high-quality genomes were then analyzed by comparative genomics, often coupled with transcriptomics and metabolomics, methods that are key to revealing the phylogenomic basis for chemical innovation and metabolite distribution among species. It is widely accepted that gene and genome duplication provide the basic genomic foundations that support the expansion of plant metabolism (Ober, 2005, 2010; Lichman et al., 2020). For example, the expansion of the cytochrome P450 family, one of the largest enzyme families in plants, is well correlated with the diversification of metabolites in the plant kingdom (Nelson and Werck-Reichhart, 2011; Hansen et al., 2021). Duplications of genes encoding other types of enzymes, such as acyltransferases, methyltransferases, and glycosyltransferases, have all been shown to be excellent markers of plant metabolic diversification (Barakat et al., 2011; Leong and Last, 2017; Louveau and Osbourn, 2019). Recently, high-quality genomes have been produced in some non-model plants, and the expansion of specific functional enzymes has been identified. For example, analysis of the *Coptis chinensis* genome revealed amplification of the Ranunculales clade-specific CYP719 family, which is associated with the diversification of protoberberine-type alkaloids (Liu et al., 2021). Similarly, amplification of chalcone synthase (CHS) genes appears to be associated with urushiol biosynthesis in *Magnifera indica* (mango) (Wang et al., 2020), and the expansion of the CYP71F1 family is proposed to have driven the diversification of tanshinone biosynthesis in *Salvia miltiorrhiza* (Ma et al., 2021).

Although genome sequencing has undergone impressive advances in recent years, the sequencing and assembly of non-model plant genomes can be expensive and time consuming. Alternatively, RNA-seq analysis can be used to identify biosynthetic genes and reveal the evolutionary basis of targeted metabolic pathways. For instance, based on biochemical data on kavalactones (with potent anxiolytic and analgesic activity) isolated from kava (*Piper methysticum*), Pluskal et al. (2019) hypothesized that a CHS-like enzyme (a type of polyketide synthase) probably participates in the biosynthesis of the styrylpyrone backbone of kavalactone. They subsequently identified three CHS-like genes in a kava transcriptome assembled de novo from leaf and root tissues. One of these genes was characterized as a chalcone synthase for flavokavains, but the other two appear to be recently duplicated and have undergone neofunctionalization to catalyze the formation of the styrylpyrone backbone of kavalactones specific to kava. Metabolic pathways can also be elucidated via comparative transcriptomics in related plant species or tissues/organs of the same species by correlating differential gene expression patterns with chemical profiles in related samples (Tzfadia et al., 2015; Delli-Ponti et al., 2020; Nett et al., 2020; Song et al., 2021). In addition to singleton gene duplication (e.g. tandem duplication and dispersed duplication), whole-genome duplication (WGD), or polyploidization, also plays important roles in plant function and evolution (Clark and Donoghue, 2018). As genome sequencing becomes more affordable, more and more genomes representing a wider variety of plant families have been sequenced, and the resulting data consistently support the finding that WGD has occurred frequently throughout plant evolutionary history (Qiao et al., 2019; Soltis and Soltis, 2021). Commonly used methods for inferring WGD in plants include (often in combination) synteny analysis (Wang et al., 2012; Qiao et al., 2019), analysis of synonymous substitutions per synonymous site (Ks) among paralogs (Cui et al., 2006), phylogenomic (gene tree) reconciliation (Jiao et al., 2011), and a likelihood-based gene-count method (Rabier et al., 2014). Although WGD is pervasive in plant genomes, assessing its effect on plant metabolism remains a challenge. WGD is often associated with an ancient polyploidization event, followed by
loss of most duplicated genes over a few million years (Lynch and Conery, 2000). The complexity of metabolic pathways, including those that produce very complex chemical structures or that interact with other pathways, adds additional barriers to investigating WGD-generated metabolic diversification. Only a few examples of WGD-driven metabolic diversity have been reported in plants (Figure 1). Representative examples include WGD-generated CYP79s for glucosinolate biosynthesis in Brassicaceae (e.g. Arabidopsis thaliana) (Hofberger et al., 2013; Edger et al., 2015). The differential accumulation of oleic and linoleic acids produced in olive (Olea europaea) compared with the closely related oil crop sesame is due in part to the divergence of the fatty acid desaturase FAD2 gene, generated by WGD (Unver et al., 2017). N-methylputrescine oxidase (MPO1), required for the biosynthesis of nicotine in Nicotiana attenuata, is proposed to have been generated via whole-genome triplication in Solanaceae (Xu et al., 2017). In Papaver somniferum, three genes (T6ODM, CODM, and COR) are likely to have been duplicated via WGD, and their encoded enzymes catalyze the final steps of the morphine biosynthetic pathway (Li et al., 2020). Retention and co-option of WGD-generated sister pathways underlie the high levels of bioactive triterpenes in Eriobotrya japonica (loquat) (Su et al., 2021).

Figure 1. Whole genome duplication (WGD) underlies metabolic diversification in plants.
WGD-associated metabolic diversification has been found for various classes of compounds in plants. The CYP79 gene family expansion that originated from the second major WGD of Brassicales laid the foundation for glucosinolate diversity in Brassicaceae (e.g. Arabidopsis thaliana) (Hofberger et al., 2013; Edger et al., 2015). The differential accumulation of oleic and linoleic acids produced in olive (Olea europaea) compared with the closely related oil crop sesame is due in part to the divergence of the fatty acid desaturase FAD2 gene, generated by WGD (Unver et al., 2017). N-methylputrescine oxidase (MPO1), required for the biosynthesis of nicotine in Nicotiana attenuata, is proposed to have been generated via whole-genome triplication in Solanaceae (Xu et al., 2017). In Papaver somniferum, three genes (T6ODM, CODM, and COR) are likely to have been duplicated via WGD, and their encoded enzymes catalyze the final steps of the morphine biosynthetic pathway (Li et al., 2020). Retention and co-option of WGD-generated sister pathways underlie the high levels of bioactive triterpenes in Eriobotrya japonica (loquat) (Su et al., 2021).

Plant Communications 3, 100300, March 14 2022 © 2022 The Author(s).
SYNERGISTIC DRIVERS OF CHEMICAL INNOVATION: BGCs

Genes were once thought to be randomly organized in plant genomes. The clustering in close proximity of multiple non-homologous genes that collectively encode metabolic pathway(s) offers a clear rebuttal of this notion (Figure 2) (Field and Osbourn, 2008; Nützmann et al., 2016). The so-called BGCs are no longer a hallmark of only microbes and fungi: they have also been found for various classes of compounds in plants, including terpenes, alkaloids, benzoxazinoids, cyanogenic glucosides, polyketides, fatty acids, and phenylpropanoids (Polturak and Osbourn, 2021). Although the clustering of metabolic genes in plants is an exception to the more common scenario of non-clustered pathways (Wisecaver et al., 2017), the study of plant BGCs has hugely facilitated the functional identification of previously unknown enzymes and pathways, thus advancing our understanding of plant genome evolution (Nützmann et al., 2016; Polturak et al., 2022). Plant BGCs are commonly composed of signature and tailoring genes (which generate and modify the backbone, respectively), although atypical types have also been reported (Polturak and Osbourn, 2021). The distinct genomic features of BGCs have led to the development of computer algorithms for mining of metabolic genes and pathways, some of which use only plant genomic data (Kautsar et al., 2017; Töpfer et al., 2017). Combining genome sequences with other types of data, such as transcriptomic and metabolomic data, further amplifies the power of genome mining for the investigation of metabolic diversity (Medema et al., 2021).

A recent emerging theme for plant BGCs is that auxiliary genes not directly involved in biosynthesis per se are, in some cases, clustered with the core cluster (Figure 2A). Examples include cognate transporters in the noscapine cluster and dhurrin cluster (Darbani et al., 2016; Dastmalchi et al., 2019) and a cofactor that is used to support the biosynthesis of hydroxycinnamoyl tyramine in rice (Shen et al., 2021). Clustering of genes with auxiliary functions thus extends the boundaries of gene clustering in plant genomes and has opened up new windows for mining BGC-related metabolic diversity (Figure 2B). Liu et al. (2020b) carried out a systematic genomic neighborhood (GN) association analysis across 13 Brassicaceae genomes, centered on oxidosqualene cyclase (OSC) genes that generate the triterpene skeleton. Pfam domains in the OSC genomic region (5- or 10-gene distance on each side of an OSC gene) were retrieved, and their presence was compared with that at the whole-genome level by fitting to a hypergeometric distribution. The OSC-centric GN analysis involved various statistical analyses to capture significant co-occurrence of OSCs with any type of coding gene, either enzymatic or non-enzymatic. This analysis identified in the OSC-GNs not only clustered metabolic genes (some of which were later
Figure 3. (pan)-genomic basis underlying metabolic diversification in plants.
We summarize the major types of (pan)-genomic variations associated with metabolic diversity in plants by showing representative examples. (A) Gene and genome duplications. SGD, single gene duplication. In A. thaliana, CYP98A8 and CYP98A9 are tandemly duplicated from a common ancestor CYP98A8/9 whose product has dual 3'-hydroxylase and 5'-hydroxylase activity on phenolamide N1, N5, N10-tri-coumaryl-spermidine. During the course (legend continued on next page)
families for the discovery of new metabolic pathways. In angiosperms (Zhao et al., 2021), and it may offer a great gene content to discover novel metabolism-related genes or pathways. Indeed, microsynteny-based phylogeny has recently been successfully tested across 50 different plant families and 30 orders in angiosperms (Zhao et al., 2021), and it may offer a great opportunity for the application of GN analysis to wider plant families for the discovery of new metabolic pathways.

**PAN-GENOMICS: A NEW FRONTIER FOR MINING METABOLIC DIVERSITY IN PLANTS**

The pan-genome, which refers to the entire conserved (core genome) and diversified (dispensable genome) genetic information of a species, has recently drawn significant attention (Bayer et al., 2020; Lei et al., 2021). Previously, population-level genomic variation was limited mainly to single-nucleotide polymorphisms (SNPs) and small insertion and deletion (<50 bp) variations, which could be picked up by short-read sequencing (Ikegawa, 2012; 1001 Genomes Consortium, 2016). Recent advances in long-read genome sequencing and de novo assembly have enabled long-overlooked genomic structural variations, such as presence–absence variations (PAVs) (>50 bp), gene copy-number variations, chromosomal inversions (CIs), and genomic translocations, to be included in pan-genomes (Ho et al., 2020). These structural variations have significantly revolutionized our understanding of the molecular mechanisms that underlie plant survival and adaptation (Li et al., 2014; Tao et al., 2019; Barragan and Weigel, 2021). To date, about 30 plant pan-genomes have been created, although there are large differences in the number of sequenced accessions in individual pan-genomes (Bayer et al., 2020; Jiao and Schneeberger, 2020; Song et al., 2020; Qin et al., 2021). Li et al. (2020) sequenced 10 opium poppy cultivars and identified gene copy-number variations that underlie marked differences in benzylisoquinoline alkaloid production among the sequenced cultivars. Interestingly, benzylisoquinoline alkaloid pathway genes within BGCs are more likely to exhibit gene copy-number variations than those outside of BGCs. Gao et al. (2019) generated a pan-genome for tomato by sequencing 725 accessions in the promoter region of TomLoxC (solyC01g006540), a 13-lipoxygenase gene linked to tomato flavor. Based on eight high-quality genomes for rapeseed (Brassica napus), Song et al. (2020) identified a large number of PAVs that were suitable for a PAV-based genome-wide association study (PAV-GWAS). This enabled the direct identification of causal structural variations for several important crop traits, including silique length, seed weight, and flowering time. PAV-GWAS has also been applied to pigeon pea (Cajanus cajan) (Zhao et al., 2020). It is thus tempting to speculate that PAV-GWAS may be widely used in the near future to identify causative genes not only for qualitative traits (as mentioned above) but also for metabolic diversification in natural populations. This is also likely to complement the widely used SNP-based mgWAS (SNP-mGWAS) approach (Luo, 2015), leading to greater resolution in the dissection of plant metabolic plasticity and complexity.

CIs play important roles in evolution (Kirkpatrick, 2010). Although it has been observed that CIs are ubiquitous across plant genomes (Huang and Rieseberg, 2020), CI-associated metabolic diversity in plants is rarely reported. Liu et al. (2020a) retrieved 22 genomic regions for the thalianol cluster from an ongoing pan-genome project in Arabidopsis thaliana. Comparative analysis revealed that CI acted as a shuffling mechanism for relocating distant metabolic genes into the core cluster region, thus forming a more compact gene cluster. The compacted thalianol cluster was present in ~80% of the examined genomes, indicating that there are advantages to tighter metabolic gene clustering (Liu et al., 2020a). Further investigations are required, however, to clarify whether these CIs lead to metabolic diversification in natural accessions of A. thaliana.

**CONCLUDING REMARKS AND PERSPECTIVES**

Great advances in genome sequencing have propelled biology and biological research into a new digital era, leading to the identification of numerous genetic and genomic structural variations among and within species (Figure 3). The rapidly growing number of sequenced plant genomes has strengthened our appreciation of the importance of gene and genome duplications for the diversification of plant metabolism. In a phylogenomic context, metabolic gene duplicates derived from random, dispersed, or whole-genome duplications can be distinguished (Moore et al., 2019; Qiao et al., 2019), and the selective forces driving these variants in shaping metabolic flexibility and variability can be inferred with greater resolution. Powerful statistical methods and phylogeny-guided approaches that can be used to detect signatures of selection at the gene and population level have been the subject of recent extensive reviews (Scosea and Fernie, 2020; Schenck and Busta, 2021). However, the application of comparative genomics beyond closely related plant species remains a challenge, particularly when the same or similar compounds are synthesized convergently or independently by distant or early diverged species (Pichersky and Lewinsohn, 2011; Lou et al., 2021). New tools are therefore required that can delineate conserved (i.e., sharing a common ancestor) and independent evolution of metabolic traits at the genomic level.
Unlocking plant metabolic diversity

High-quality genome sequences provide an entrance for functional genomics and subsequent characterization of metabolic diversity between plant species. However, for many recently sequenced non-model plants, genome annotations are incomplete and sometimes even incorrect, thereby hindering functional analysis. Integrating genome sequencing with other omics techniques (e.g. gene co-expression analysis, proteomics, and metabolomics) and additional synergistic approaches (e.g. phylogenetic analysis, gene enrichment analysis, evolutionary principles, and functional validation) is likely to enable effective and efficient annotation and characterization of metabolic diversity in the near future.

It is also becoming evident that metabolic gene clustering is widely distributed across different classes of natural products and, putatively, in all sequenced plant genomes. On average, approximately 30 BGC candidates have been identified in each sequenced plant genome (a rough estimate based on plantiSMASH) (Kautsar et al., 2017), but the vast majority have not been functionally characterized and are therefore only putative BGCs. Functional characterization of more *in silico*-predicted BGCs will in turn be used to adjust methods for the genome mining of BGCs. The accuracy of BGC prediction also depends on the contiguity of the host genome assembly, and assembly currently faces challenges in resolving the rich repetitive sequences that are commonly identified in plant BGC regions (Field et al., 2011; Shen et al., 2021). Improving long-read sequencing with high accuracy is likely to conquer this problem. In addition to the localization of gene clusters, co-expression of the clustered genes is another important feature of BGCs and has been leveraged for the discovery of biosynthetic pathways (Nützmann et al., 2016; Polturak and Osbourn, 2021). However, the molecular mechanisms that underlie the co-regulation of BGC genes are not fully understood, impeding investigations of their encoded metabolic diversity across various developing tissues of host plants (Yu et al., 2016; Nützmann et al., 2020).

Although plant genome sequencing is becoming routine, the creation and visualization of plant pan-genomes is still challenging (Lei et al., 2021). Calling variants in pan-genomes is heavily reliant on the degree of their contiguity, completeness, and accuracy. Association studies based on structural variants are further constrained by other factors, including (1) the size and highly repetitive nature of plant pan-genomes; (2) the pervasive occurrence of polyploidization and gene duplication events; and (3) the need to sequence an adequate number of accessions that are sufficiently variable for association analysis at a reasonable time and cost. As metabolic traits are often quantitative, the challenges mentioned above are likely to be magnified in mGWAS analysis, e.g. PAV-mGWAS enabled by pan-genomes. Compared with the pan-genomes of microbes, in which the concept of pan-genome was first conceived (Tettelin et al., 2005), the pan-genomes of plants have developed much more slowly because of the challenges mentioned above. However, progress in plant pan-genomics can be achieved by taking inspiration from lower organisms; for instance, the most frequently used software for plant BGCs, plantiSMASH, is derived from antiSMASH, which was initially developed to mine gene clusters from bacterial and fungal genomes (Medema et al., 2011).

In summary, the surge in the genome sequencing field has greatly expanded our genomic and pan-genomic understanding of metabolic diversification in plants. Massive progress in the identification of genomic structural variations will provide rich materials for the interpretation of complex metabolic variations that differ widely among plants and will facilitate the elucidation of their largely uncharacterized ecological roles in the relationships between plants and other associated organisms.

**FUNDING**
The Z.L. laboratory is supported by a startup grant provided by Shanghai Jiao Tong University, School of Agriculture and Biology and the Shanghai Pujiang Program (20PJ1405900).

**ACKNOWLEDGMENTS**
We apologize to those authors whose work could not be discussed owing to space limitations. We thank Dr. Guy Polturak (John Innes Centre) for useful comments. No conflict of interest is declared.

**REFERENCES**

1001 Genomes Consortium. (2016). 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell 166:481–491.

Bai, Y., et al. (2021). Modulation of Arabidopsis root growth by specialized triterpenes. New Phytol. 230:228–243.

Barakat, A., Choi, A., Yassin, N.B.M., Park, J.S., Sun, Z., and Carlson, J.E. (2011). Comparative genomics and evolutionary analyses of the O-methyltransferase gene family in Populus. Gene 479:37–46.

Barragan, A.C., and Weigel, D. (2021). Plant NLR diversity: the known unknowns of pan-NLRomes. Plant Cell 33:814–831.

Bayer, P.E., Golicz, A.A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. Nat. Plants 6:914–920.

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S.Y. (2014). Genomic signatures of specialized metabolism in plants. Science 344:510–513.

Clark, J.W., and Donoghue, P.C.J. (2018). Whole-genome duplication and plant macroevolution. Trends Plant Sci. 23:933–945.

Cui, L., et al. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:738–749.

Darbani, B., Motawia, M.S., Olsen, C.E., Nour-Eldin, H.H., Møller, B.L., and Rook, F. (2016). The biosynthetic gene cluster for the cyanogenic glucoside dhurrin in Sorghum bicolor contains its co-expressed vacuolar MATE transporter. Sci. Rep. 6:37079.

Dastmalchi, M., Chang, L., Chen, R., Yu, L., Chen, X., Hagel, J.M., and Facchini, P.J. (2019). Purine permease-type benzyloquinoline alkaloid transporters in opium poppy. Plant Physiol. 181:916–933.

Delli-Porti, R., Shihvare, D., and Mutwil, M. (2020). Using gene expression to study specialized metabolism—a practical guide. Front. Plant Sci. 11:625035.

Dixon, R.A., and Strack, D. (2003). Phytochemistry meets genome analysis, and beyond. Phytochemistry 62:815–816.

Edger, P.P., et al. (2015). The butterfly plant arms-race escalated by gene and genome duplications. Proc. Natl. Acad. Sci. U S A 112:8362–8366.

Erb, M., and Kliewenstein, D.J. (2020). Plant secondary metabolites as defenses, regulators, and primary metabolites: the blurred functional trichotomy. Plant Physiol. 184:39–52.

Fang, C., Fernie, A.R., and Luo, J. (2019). Exploring the diversity of plant metabolism. Trends Plant Sci. 24:83–98.
Unlocking plant metabolic diversity

Li, Y., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32:1045–1052.

Li, Q., Ramasamy, S., Singh, P., Hagel, J.M., Dunemann, S.M., Chen, X., Chen, R., Yu, L., Tucker, J.E., Facchin, P.J., et al. (2020). Gene clustering and copy number variation in alkaloid metabolic pathways of opium poppy. Nat. Commun. 11:1190.

Lichman, B.R., Godden, G.T., and Buell, C.R. (2020). Gene and genome duplications in the evolution of chemodiversity: perspectives from studies of Lamiaceae. Curr. Opin. Plant Biol. 55:74–83.

Liu, Z., Tavares, R., Forsythe, E.S., André, F., Lungan, R., Jonasson, G., Boutet-Mercuy, S., Tohge, T., Beilstein, M.A., Werck-Reichhart, D., et al. (2016). Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. Nat. Commun. 7:1–10.

Liu, Z., Cheema, J., Vigouroux, M., Hill, L., Reed, J., Paajanen, P., Yant, L., and Osbourn, A. (2020a). Formation and diversification of a paradigm biosynthetic gene cluster in plants. Nat. Commun. 11:5354.

Liu, Z., Suarez Duran, H.G., Harmanvichyech, Y., Stephenson, M.J., Schranz, M.E., Nelson, D., Medema, M.H., and Osbourn, A. (2020b). Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. New Phytol. 227:1109–1123.

Liu, Y., et al. (2021). Analysis of the Coptis chinensis genome reveals the diversification of protoberberine-type alkaloids. Nat. Commun. 12:3278.

Lou, Y.-R., Anthony, T.M., Fiesel, P.D., Arking, R.E., Christensen, E.M., Jones, A.D., and Last, R.L. (2021). It happened again: convergent evolution of acetylglucose specialized metabolism in black nightshade and wild tomato. Sci. Adv. 7:eabj8726.

Louveau, T., and Osbourn, A. (2019). The sweet side of plant-specialized metabolism. Cold Spring Harb. Perspect. Biol. 11:a034744.

Luo, J. (2015). Metabole-based genome-wide association studies in plants. Curr. Opin. Plant Biol. 24:31–38.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

Ma, Y., et al. (2021). Expansion within the CYP71D subfamily drives the heterocyclization of tanshinones synthesis in Salvia miltiorrhiza. Nat. Commun. 12:685.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Ma, Y., et al. (2021). Expansion within the CYP71D subfamily drives the heterocyclization of tanshinones synthesis in Salvia miltiorrhiza. Nat. Commun. 12:685.

Misra, B.B., Assmann, S.M., and Chen, S. (2014). Plant single-cell and single-cell-type metabolomics. Trends Plant Sci. 19:637–646.

Moore, R.C., and Purugganan, M.D. (2005). The evolutionary dynamics of plant duplicate genes. Curr. Opin. Plant Biol. 8:122–128.

Moore, B.M., Wang, P., Fan, P., Leong, B., Schenck, C.A., Lloyd, J.P., Lehti-Shiu, M.D., Last, R.L., Pichersky, E., and Shiu, S.-H. (2019). Robust predictions of specialized metabolism genes through machine learning. Proc. Natl. Acad. Sci. U S A 116:2344–2353.

Nelson, D., and Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. Plant J. 66:194–211.

Nett, R.S., Lau, W., and Sattely, E.S. (2020). Discovery and engineering of colchicine alkaloid biosynthesis. Nature 584:148–153.

Nutzmann, H.W., Huang, A., and Osbourn, A. (2016). Plant metabolic clusters—from genetics to genomics. New Phytol. 211:771–789.

Nutzmann, H.-W., Doerr, D., Ramirez-Colmenero, A., Sotelo-Fonseca, J.E., Wegel, E., Di Stefano, M., Wingett, S.W., Fraser, P., Hurst, L., Fernandez-Valverde, S.L., et al. (2020). Active and
Unlocking plant metabolic diversity

repressed biosynthetic gene clusters have spatially distinct chromosome states. Proc. Natl. Acad. Sci. U S A 117:13800–13809.

Ober, D. (2005). Seeing double: gene duplication and diversification in plant secondary metabolism. Trends Plant Sci. 10:444–449.

Ober, D. (2010). Gene duplications and the time thereafter—examples from plant secondary metabolism. Plant Biol. 12:570–577.

Ohno, S. (1970). Evolution by Gene Duplication (Springer-Verlag).

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. Ober, D. (2005). Seeing double: gene duplication and diversification in plant secondary metabolism. Annu. Rev. Plant Biol. 62:549–566.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiol. 171:2294–2316.

Pichersky, E., and Lewinsohn, E. (2016). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biol. 20:38.

Pluskal, T., Torrens-Spence, M.P., Fallon, T.R., De Abreu, A., Shi, C.H., and Weng, J.-K. (2019). The biosynthetic origin of psychoactive kavalactones in kava. Nat. Plants 5:867–878.

Polturak, G., and Osbourn, A. (2021). The emerging role of biosynthetic gene clusters in plant defense and plant interactions. Plos Pathog. 17:e1009698.

Polturak, G., Liu, Z., and Osbourn, A. (2022). New and emerging concepts in the evolution and function of plant biosynthetic gene clusters. Curr. Opin. Green. Sustain. Chem. 33:100568.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., and Paterson, A.H. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biol. 20:38.

Qin, P., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell 184:3542–3558.e16.

Rabier, C.-E., Ta, T., and Ané, C. (2014). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. Mol. Biol. Evol. 31:750–762.

Schenck, C.A., and Busta, L. (2021). Using interdisciplinary, phylogeny-guided approaches to understand the evolution of plant metabolism. Plant Mol. Biol.

Scossa, F., and Fernie, A.R. (2020). The evolution of metabolism: how to test evolutionary hypotheses at the genomic level. Comput. Struct. Biotechnol. J. 18:482–500.

Shen, S., et al. (2021). An Oryza-specific hydroxycinnamoyl tyramine gene cluster contributes to enhanced disease resistance. Sci. Bull. 66:2360–2380.

Soltis, P.S., and Soltis, D.E. (2021). Plant genomes: markers of evolutionary history and drivers of evolutionary change. Plants, People, PLANET. 3:74–82.

Song, J.-M., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat. Plants 6:34–45.

Song, J.-J., et al. (2021). A 2-oxoglutarate dependent dioxygenase converts dihydrofuran to furan in Salvia diterpenoids. Plant Physiol. 190:567.

de Souza, L.P., Borghi, M., and Fernie, A. (2020). Plant single-cell metabolomics-challenges and perspectives. Int. J. Mol. Sci. 21:9987.

Su, W., et al. (2021). Polyploidy underlies co-option and diversification of biosynthetic triterpene pathways in the apple tribe. Proc. Natl. Acad. Sci. U. S. A. 118, e2101767118.

Tao, Y., Zhao, X., Mace, E., Henry, R., and Jordan, D. (2019). Exploring and exploiting pan-genomics for crop improvement. Mol. Plant 12:156–169.

Tettelin, H., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome. Proc. Natl. Acad. Sci. U S A 102:13950–13955.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815.

Topfer, N., Fuchs, L.-M., and Aharoni, A. (2017). The PhytoClust tool for metabolic gene clusters discovery in plant genomes. Nucleic Acids Res. 45:7049–7063.

Tzfadia, O., Diels, T., De Meyer, S., Vandepoele, K., Aharoni, A., and Van de Peer, Y. (2015). CoExpNetViz: comparative co-expression networks construction and visualization tool. Front. Plant Sci. 6:1194.

Unver, T., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. Proc. Natl. Acad. Sci. U S A 114:E9413–E9422.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., et al. (2012). MCSmapX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40:e49.

Wang, P., et al. (2020). The genome evolution and domestication of tropical fruit mango. Genome Biol. 21:60.

Weng, J.K. (2014). The evolutionary paths towards complexity: a metabolic perspective. New Phytol. 201:1141–1149.

Wisecaver, J.H., Borowsky, A.T., Tzin, V., Jander, G., Kliebenstein, D.J., and Rokas, A. (2017). A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. Plant Cell 29:944–959.

Xu, S., et al. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. Proc. Natl. Acad. Sci. U S A 114:6133–6138.

Yu, N., Nutzmann, H.-W., MacDonald, J.T., Moore, B., Field, B., Berriri, S., Trick, M., Rosser, S.J., Kumar, S.V., Freemont, P.S., et al. (2016). Delineation of metabolic gene clusters in plant genomes by chromatin signatures. Nucleic Acids Res. 44:2255–2265.

Zhan, C., et al. (2020). Selection of a subspecies-specific diterpene gene cluster implicated in rice disease resistance. Nat. Plants 6:1447–1454.

Zhao, J., Bayer, P.E., Ruperao, P., Saxena, R.K., Khan, A.W., Golicz, A.A., Nguyen, H.T., Batley, J., Edwards, D., and Varshney, R.K. (2020). Trait associations in the pangeneome of pigeon pea (Cajanus cajan). Plant Biotechnol. J. 18:1946–1954.

Zhao, T., Zwaenepoel, A., Xue, J.-Y., Kao, S.-M., Li, Z., Schranz, M.E., and Van de Peer, Y. (2021). Whole-genome microsynteny-based phylogeny of angiosperms. Nat. Commun. 12:3498.