Article:

Lichman, Benjamin R., Godden, Grant T. and Buell, C. Robin (2020) Gene and genome duplications in the evolution of chemodiversity:perspectives from studies of Lamiaceae. CURRENT OPINION IN PLANT BIOLOGY. pp. 74-83. ISSN: 1369-5266

https://doi.org/10.1016/j.pbi.2020.03.005

Gene and genome duplications in the evolution of chemodiversity:

perspectives from studies of Lamiaceae

Benjamin R. Lichman[1], Grant T. Godden[2], C. Robin Buell[3,4,5]*

[1]Centre for Novel Agricultural Products, Department of Biology, University of York, York YO10 5DD, UK.

[2]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA.

[3]Department of Plant Biology, Michigan State University, 612 Wilson Road, East Lansing, MI 48824, USA.

[4]Plant Resilience Institute, Michigan State University, 612 Wilson Road, East Lansing, MI 48824, USA.

[5]MSU AgBioResearch, Michigan State University, 446 West Circle Drive, East Lansing, MI 48824, USA.

*Correspondence to: buell@msu.edu

Summary

Plants are reservoirs of extreme chemical diversity, yet biosynthetic pathways remain underexplored in the majority of taxa. Access to improved, inexpensive genomic and computational technologies has recently enhanced our understanding of plant specialized metabolism at the biochemical and evolutionary levels. Genomics has led to the elucidation of pathways leading to key metabolites. Furthermore, it has provided insights into the mechanisms of chemical evolution, including neo- and sub-functionalization, structural variation, and modulation of gene expression. The broader utilization of genomic tools across the plant tree of life, and an expansion of genomic resources from multiple accessions within species or populations, will improve our overall understanding of chemodiversity. It will also lead to greater insight into the selective pressures contributing to and maintaining this diversity, which in turn will enable the development of more accurate predictive models of specialized metabolism in plants.

Introduction

As with other aspects of biology, genomics has enabled a revolution in plant specialized metabolism [1] that has been driven by rapid advancements in sequencing technologies [2-4], access to high-performance computing resources [5], and increased access to genomic and transcriptomic datasets from species with diverse specialized metabolism (Fig. 1) [6, 7]. Through genome assemblies, signatures of specialized metabolism such as coexpression, gene duplication, and clustered biosynthetic pathways greatly facilitate biosynthetic pathway discovery (Fig. 2) [8-11]. In a complementary manner, the power of comparative genomics permits resolution of mechanisms that underpin the evolution of metabolism across taxa [12-14]. This is highlighted by phylogenomic investigations that have revealed the pivotal roles of gene and genome duplication in the diversity of specialized metabolism [6, 12-16]. In this review, we discuss the general impact of genomics on plant specialized metabolism research, with a focus on examples from the mint family (Lamiaceae) for which a multitude of genome sequences have been generated with a primary focus on understanding specialized metabolism. We conclude by highlighting future directions in genome-enabled studies of plant specialized metabolism.

Genomic approaches to plant specialized metabolism

(Phylo-) transcriptomics

Despite rapid increases in genome sequence availability and quality, transcriptome sequencing (RNA-sequencing, RNA-seq) remains a key approach in the investigation of plant-specialized metabolism across diverse taxa (see article on Transcriptomics in this issue for more detail). Candidate genes involved in metabolic pathways can be identified through differential gene expression across temporal and spatial dimensions [17] and through homology with known genes [18, 19]. With its low cost and ease of data generation, transcriptome sequencing can enable generation of comparative data sets for large numbers of species [7], providing multipurpose datasets for gene discovery [20] and applied phylogenomic analyses [17, 21-23].

There is mounting evidence that gene and whole-genome duplications (WGDs) are prevalent across land plants and are drivers of trait innovation, including the diversity of specialized metabolism [24]. In the absence of genome sequences from species spanning multiple related lineages, researchers have relied on phylotranscriptomic analyses to characterize genome dynamics (e.g. duplications) in a phylogenomic context. Such studies have enabled the discovery and resolution of gene expansions and WGDs within lineages [17, 21-23], and provided a macroevolutionary understanding of the origins of chemical diversity [6, 14, 25].

Genome sequencing, assembly, and annotation

Over the past decade, access to inexpensive next-generation sequencing technologies has enabled a wide range of genome-enabled studies of plant metabolism [1]. With the maturity of third-generation sequencing technologies, this pace of change promises to continue [reviewed here: [2, 3, 26]]. Two categories of technologies have led to recent improvements in plant genome assemblies: long-read single molecule sequencing platforms and methods that identify long-range genomic connections. Sequencing platforms such as the Pacific Biosciences and Oxford Nanopore Technologies platforms [4, 27-29] provide long contiguous reads that result in longer, more contiguous genome assemblies generated with a suite of new assembly algorithms. Complementary methods such as 10X Chromium linked reads, Hi-C proximity ligation, and optical mapping facilitate generation of longer, chromosome-scale assemblies [27, 28, 30] from long-read generated assemblies. Coupled with the platform advances are major improvements in genome sequence assembly and annotation tools. With the advent of Oxford Nanopore Technology, new long-read genome assembly software have been developed including Shastac [31] and Flye [32], which can assemble genomes rapidly. Plant species are not always homozygous inbreds, and improvements in addressing heterozygous genomes have been made including the ability to purge haplotigs [33]. Software for scaffolding the genomes into higher order assemblies such as pseudomolecules or pseudochromosomes includes SALSA2 [34] and Juicebox [35]. For genome annotation, transcript evidence is key to accurate gene model construction and multiple software for generation of genome-guided transcripts using long-read cDNA sequences have been released including Stringtie2 [36] and FLAIR [37]. Collectively, parallel advancements in algorithms and software for third-generation sequencing platforms will significantly improve the quality of not only genome assemblies, but also their annotation.

Some genes involved in specialized metabolism are physically clustered within the genome as tandem repeats or metabolic gene clusters, loci containing multiple non-paralogous genes involved in a pathway [10]]. Long read length technologies can resolve tandem duplications and extended repetitive sequences typical of such regions, revealing complex loci associated with specialized metabolism [38, 39]. Syntenic analyses are vital for understanding genome evolution and dynamics, including the origins of gene duplications [15] and gene clusters [13, 40]. Both the quantity and quality of genome sequences are essential for the improved resolution of metabolic pathway genes.

Population-scale studies

Large-scale genome resequencing-based approaches are emerging as powerful tools for the understanding of plant specialized metabolism. Metabolite-associated genome wide association studies (mGWAS) involve the identification of genomic loci associated with metabolic traits through identification of single nucleotide polymorphisms and selective sweeps in large-scale genome resequencing efforts [41-43]. Structural variants such as copy number and presence/absence variants have been identified in numerous plant species; functional analyses of a subset of these structural variants have revealed roles in adaptation, including biosynthesis of anti-insecticidal methyl ketones in *Solanum tuberosum* L. (potato) [44] and noscapine in *Papaver somniferum* L. (opium poppy) [45].

The evolution and genomics of specialized metabolism in Lamiaceae

Mints are a species-rich angiosperm clade (~7000 species) with a cosmopolitan distribution, and exhibit a wide range of chemical diversity, including iridoids, polypropanoids, and canonical terpenoids. This chemical richness plays functionally significant roles in nature, facilitating complex interactions among mints and insects (e.g., plant-pollinator and plant-herbivore interactions [82]), phytopathogens (e.g., antimicrobial activity [83]), and other co-occurring plants (e.g., allelopathy [84, 85]). It has also led to the use of numerous mint species as herbal medicines (*Salvia* L., *Scutellaria* L.), culinary herbs (*Mentha* L., *Origanum* L., *Thymus* L.), and sources for perfume oils (*Lavandula* L., *Pogostemon* Desf.) and health-promoting or therapeutic bioactive compounds and phytochemicals (reviewed in [86]). These uses have motivated recent multidisciplinary and genome-based research efforts to elucidate their specialized metabolism and identify the origins of their chemical diversity and complexity. At the time of writing, there are twelve published genomes from eight Lamiaceae species [46-57]. The highest quality genomes, *Tectona grandis* L.f. (teak) [56] and *Scutellaria baicalensis* Georgi (Chinese skullcap) [57], represent state-of-the-art plant genomes sequenced using a hybrid approach and assembled into pseudochromosomes. A recent phylotranscriptomic analysis of Lamiacaeae, consisting of 48 mint leaf transcriptomes, led to insights into the evolution of metabolism [6] and the occurrence of whole-genome duplications [58] across the clade.

Lineage-specific gene family expansions

Orthogroup (gene family) expression and occupancy (gene number), and their association with chemical traits, were examined across Lamiaceae in a phylogenomic context [6]. The distribution of monoterpene diversity and terpene synthase b (TPSb; primarily monoterpene synthases) orthogroup occupancy data in Lamiaceae suggests that lineage-specific gene expansions (LSEs) may have contributed to chemical novelty or diversity in their respective clades [6].

Comparisons of genome assemblies have also implicated LSEs in Lamiaceae, specifically in terpenoid related genes. For example, the monoterpene producer lavender had multiple copies of 1-Deoxy-d-xylulose-5-phosphate synthase (DXS) and 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase (HDR), genes that control the accumulation of monoterpenes [49]. Patchouli, grown for its high sesquiterpene content, had an increased copy number of TPSa genes (primarily sesquiterpene synthases) [47, 48]. Interestingly, these TPSa genes showed divergent gene expression patterns, including some that were not expressed in any tissue, indicating that they may have undergone functional divergence or pseudogenization.

Whole-Genome, Tandem, and Dispersed Duplications

Ancient WGD events have been associated with increased specialized metabolite diversity and species diversification rates in Brassicales, where they may have contributed to the proliferation of novel genes and gene interactions and triggered species diversification rates, especially within Brassicaceae [12]. Ancient WGDs might explain similar patterns within and among major mint clades, which exhibit asymmetrical levels of species richness and chemical diversity [6, 58]. In Lamiaceae, phylotranscriptomic analyses have revealed widespread but asymmetrical gene duplication dynamics and signatures of ancient polyploidy [58]. As many as 28 putative ancient WGD events were inferred across the clade (Fig. 3), with pronounced levels of gene and whole-genome duplications in the species- and monoterpene- rich Nepetoideae, relative to other major mint subclades [58].

LSEs identified in phylotranscriptomic studies may originate from WGDs, followed by selective retention of genes, or from small-scale duplication processes such as tandem duplication. The teak genome showed a high prevalence of tandem repeats, with over 60% of TPSs and cytochrome P450 (CYP) genes occurring in tandem repeats [56]. In the majority of cases, tandem copies had differential expression patterns, implying neo-, sub-, or non- functionalization processes were at play. In the root-specific flavone biosynthesis pathway in *S. baicalensis*, root-specific flavone synthase II (FNSII-2) and O-methyltransferase 5 (OMT5) originated from tandem duplications followed by neofunctionalization or subfunctionalization, respectively [57].

Duplications in which paralogs are in distant, non-syntenic, genomic locations are referred to as dispersed duplications, and are thought to arise through transposition, segmental duplication, or retroduplication [15]. Through synteny, it is possible to identify which paralog is the ancestral locus. In *S. baicalensis*, the root-specific chalcone-synthase (CHS2) and flavone 8-hydroxylase F8H neofunctionalized after dispersed duplication events. Curiously, it appears that the new enzyme activities evolved in copies present at the ancestral locus, whilst the paralogs that relocated to new genomic locations maintained their ancestral function [57]. Fragments of retrotransposons surrounding duplicated paralogs point to duplication mechanisms such as unequal crossing-over or (retro)transposition. The reason why a specific paralog neofunctionalizes after duplication is unclear, though the genomic context of the paralog, including neighboring genes, regulatory elements, and chromatin modifications may be influential.

Gene pairs represent non-random genomic association of two non-paralogous, functionally related, genes. These genomic signatures have emerged as a common feature of plant specialized metabolism, especially in terpenoid biosynthesis. These associations often echo the early stages of terpenoid chemical diversification, with associations observed between TPSs and CYPs [8] or prenyltransferases (PTs) [59]. Associations between CYPs and methyltransferases (MTs), enzyme pairs involved in late-stage modifications, have also been observed [60]. The genomes of *Salvia splendens* Sellow ex J.A. Shultes (scarlet sage) and *S. miltiorrhiza* Bunge (Chinese sage or danshen) show signatures of association between TPSs and CYPs [46, 53]. The clustered TPS and CYP genes in *S. miltiorrhiza* have roles in tanshinone (diterpene) biosynthesis [61, 62]. The phylogenetic relationships between the component genes of the clusters reveals their probable origins from an ancestral TPS/CYP pair that underwent cluster duplication prior to CYP tandem duplication. These genes do not show clear coexpression patterns within each cluster, indicating that, after duplication, the expression patterns diverged due to drift or neo-functionalization.

Pairwise association of functionally related biosynthetic genes may represent the starting point in plant gene cluster formation. These minimal clusters can grow and diversify through tandem repetition of individual components or by recruitment of non-homologous genes. However, gene cluster size and prevalence may be underrepresented in poorly assembled genomes.

The highly contiguous teak genome contains a large terpenoid cluster (~700 kb) that appears to be syntenic to the *S. miltiorrhiza* and *S. splendens* tanshinone clusters [56]. This cluster features TPSs and CYPs from multiple clades, perhaps indicating gene recruitment. Although this cluster is not coexpressed across all tissues, four of the clustered genes are expressed in roots and may therefore be functionally related. Similar clusters can be observed in patchouli [47], a tantalizing indication of conserved loci responsible for (di)terpenoid biosynthesis across Lamiaceae. Analysis of mint genomes using predictive software (e.g. PlantiSmash) may reveal a wealth of uncharacterized gene clusters and, consequently, new biosynthetic pathways [9].

As more genome sequences are obtained, and greater numbers of conserved clusters are identified, our understanding of the origin and function of metabolic gene clusters will improve. Computational identification and functional characterization of gene clusters, coupled with phylogenomic and syntenic analyses, will enable identification of key patterns and allow for inference of evolutionary and genomic events with greater resolution [13, 63].

Evolution without duplication

The *S. baicalensis* genome contains an unusual example of an enzyme that evolved without duplication [57]. The enzyme catalyzing the first committed step into the 4-deoxyflavone pathway, cinnamate-CoA ligase (CLL-7), does not appear to have undergone any duplication or transposition in *Scutellaria* L.. Instead, the CLL-7 enzyme from *S. baicalensis* acquired mutations that enable it to turn over the cinnamic acid substrate, whereas orthologs from *S. miltiorrhiza, S. splendens* and *Sesamum indicum* L. (Pedaliaceae) cannot. It is unusual to observe evolution of new enzyme function without gene duplication, though it is possible CLL-7 represents a pre-duplication state in a subfunctionalization regime [64].

Localization of gene expression

Gene co-expression as determined by transcriptomics continues to be a powerful method for detecting metabolic gene candidates, with the hypothesis being that genes in the same metabolic pathway have similar expression patterns across tissue types. For example, in *Coleus forskohlii, Vitex agnus-castus* and *Salvia militiorrhiza*, TPS and P450 genes responsible for diterpenoid biosynthesis were discovered as they were highly expressed in tissues accumulating the diterpenes (root cork, trichomes and hair root respectively) [65-67]. Interestingly, *S. militiorrhiza* copalyl diphosphate synthases (CPSs) show divergent tissue expression patterns despite similar activities, indicating that different paralogs have evolved to contribute separately to tanshinone (diterpenoid) biosynthesis in specific tissues [61].

Regulation of metabolism

Access to a genome rather than a transcriptome assembly permits identification of regulatory regions that control gene expression and regulation. For example, two novel glandular-trichome specific promoters were identified in the genome of *Mentha longifolia* (L.) Huds. (horse mint) and used to modify the essential oil composition of *M.* × *piperita* L. (peppermint) [52]. Analysis of the promoter regions of genes involved in tanshinone (diterpene) and phenolic acid biosynthesis in *S. miltiorrhiza* led to the putative identification of transcription factor binding regions [68-70]. Experimental investigations into transcription factor binding and gene expression has led to a detailed and complex picture of specialized metabolite regulation in *S. miltiorrhiza*, where phenolic acids and tanshinones are regulated antagonistically [71]

Phylotranscriptomic analysis of Lamiaceae indicated that iridoid biosynthesis presence/absence was primarily controlled by expression of a single gene, geraniol synthase (GES), a dedicated terpene synthase that occupies the branch point between monoterpene and iridoid biosynthesis (Fig. 3). Loss of iridoids occurred in the common ancestor of Nepetoideae, the largest of twelve major mint lineages, though *Nepeta* L. regained iridoid biosynthesis [72]. Outside of Nepetoideae there have been at least eight independent losses of iridoid biosynthesis in Lamiaceae, each coupled with reduction of GES gene expression (Fig. 3).

The teak genome assembly provides insight into mechanisms of iridoid biosynthesis loss. Whilst no GES expression was detected in the leaf transcriptome (Fig. 3), a GES gene (Tg14g06840) and a gene encoding iridoid synthase (Tg03g18820), which catalyzes the first committed step into the iridoid pathway, are both present in the genome. This indicates that reduction of gene expression is a mechanism for silencing of iridoid biosynthesis. Pseudogenization and gene loss may follow if there is no selective pressure to maintain these genes (i.e., they are not involved in other pathways).

This highlights the role of regulation in the evolution of metabolism. Not only is the gain and loss of genes important for metabolic evolution, but similar selective forces also act on promoters, regulatory regions, and transcription factor networks. WGDs enabled the emergence of a Nicotiana unique co-expression network which supplements the jasmonic acid signalling systems for mediating the defence response to herbivory [73]. The fate of promoter regions in gene duplication and gene cluster formation is largely unknown, but may be key to understanding metabolic evolution [74].

Future perspectives

With access to inexpensive, high quality genome sequencing platforms and advancements in computational tools (Fig. 2), an explosion in our knowledge of both the components and the evolutionary origins of chemodiversity will occur. Genome-wide gene duplication dynamics and their links to chemical evolution have not been explored in a phylogenomic context in any plant group, representing an area of opportunity for specialized metabolism research that may shed light on important evolutionary processes operating at the genome level. The comparative contributions of ancient WGDs and other small-scale gene duplication events to gene family expansions, particularly those important to specialized metabolism, will be revealed through these studies [75]. Large-scale resequencing can lead to the detection of selective sweeps, which can facilitate discovery of candidate genes or biosynthetic gene clusters underlying chemical variation at microevolutionary scales. Dedicated resequencing projects that annotate structural variants enable construction of pan-genomes and identification of core and dispensable genes, including genes responsible for genotype-specific metabolism [76]. Inclusion of epigenomics will augment our understanding of gene regulation and specialized metabolism, expanding our understanding from a simple transcription factor-target gene level to a global, genome landscape view of regulation. Paradigm-altering approaches to gene annotation such as implementation of machine learning algorithms will enable improved predictions of gene function [77] and, as a consequence, the ease of functional genomics.

The identification and analysis of conserved gene clusters in a phylogenomic framework will reveal details of their birth, growth, and death [13][40]. Moreover, detection of cluster variants within species will provide insights into how population-level (microevolutionary) processes influence cluster formation and loss and contribute to patterns at macroevolutionary scales. The effects of gene duplication and gene clustering on gene regulation remains largely unknown. A recent “recruitment model” of plant metabolic evolution implicates post-duplication promoter evolution as a key step in the recruitment of genes into new metabolic pathways under the control of conserved transcription factors [74]. Comparative genomic analyses, coupled with genome editing experiments, will enable investigations of this compelling hypothesis.

Acknowledgements

Funding for work on specialized metabolism was provided in part by a grant to CRB from the Michigan State University Foundation, the National Science Foundation (IOS-1444499 and IOS-1546657), and the USDA Hatch Act. BRL is supported by a UKRI Future Leaders Fellowship.

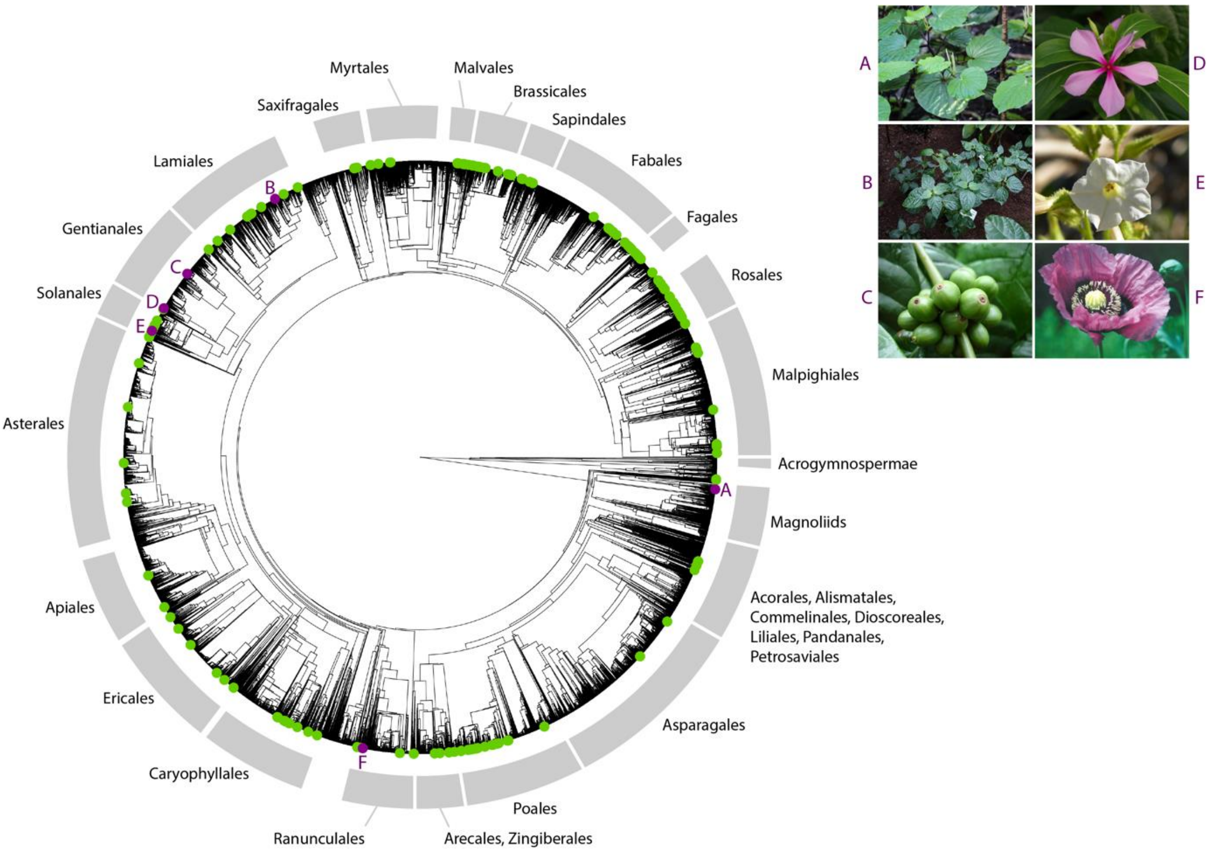

Figure 1. Phylogenetic diversity of publicly available genome sequences. Angiosperm species with genome sequences reported by [78]are indicated with colored circles on the seed plant phylogeny of Smith and Brown [22]. A selection of major clade annotations is provided to aid visualization. Examples of genomes recently used in investigations of specialized metabolism are denoted with magenta circles (A–F) and correspond to the following photos, shown as an inset: (A) *Piper methysticum* G. Forst. (kava), source of kavalactones [79]; (B) Pogostemon cablin (Blanco) Benth. (patchouli), source of sesquiterpenes [48]]; (C) *Coffea canephora* Pierre ex A. Forehner (Robusta coffee), source of caffeine [16]; (D) *Catharanthus roseus (L.) G. Don* (Madagascar periwinkle), source of vincristine [80]; (E) *Nicotiana attenuata* Torr. ex S. Watson (coyote tobacco), source of nicotine [81]; (F) *Papaver somniferum* L. (opium poppy), source of noscapine and morphine [39]. All photos were sourced from the Encyclopedia of Life and Wikimedia Commons and available under a creative commons (CC) public domain license, except for the following: (A) *P. methysticum* by Arthur Chapman (CC BY-NC 2.0); (E) *N. attenuata* by Stan Shebs (CC BY-SA 3.0).

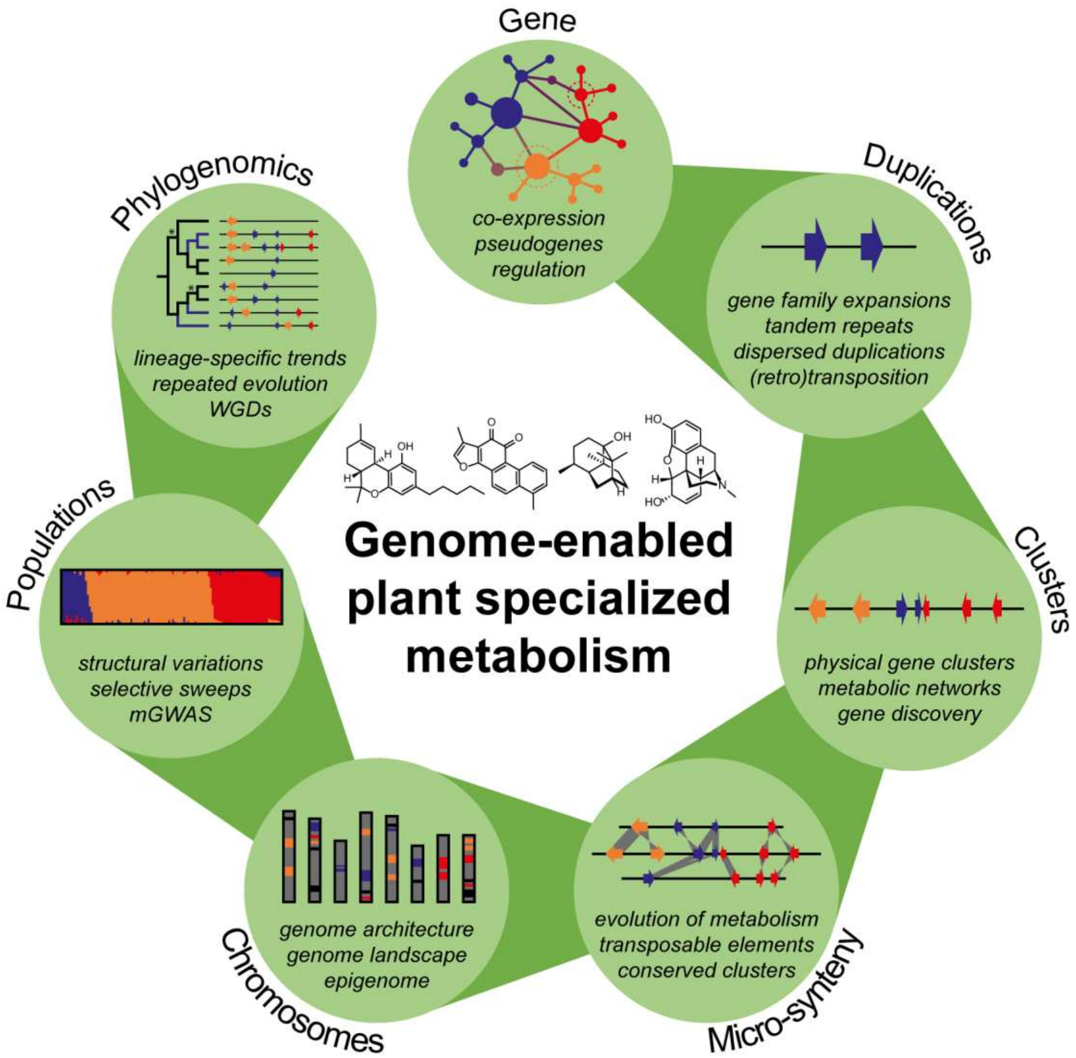

Figure 2. Overview of genome-enabled plant specialized metabolism research. This scheme highlights how genomics can inform understanding of plant biology and plant specialized metabolism at different resolutions. Acronyms: mGWAS (metabolic genome-wide association studies) and WGDs (whole genome duplications). Plant specialized metabolites (L-R): tetrahydrocannabinol (*Cannabis sativa* L., Cannabaceae), tanshinone I (*Salvia miltiorrhiza* Bunge*,* Lamiaceae), patchoulol (*Pogostemon cablin* (Blanco) Benth., Lamiaceae) and morphine (*Papaver somniferum* L., Papaveraceae).

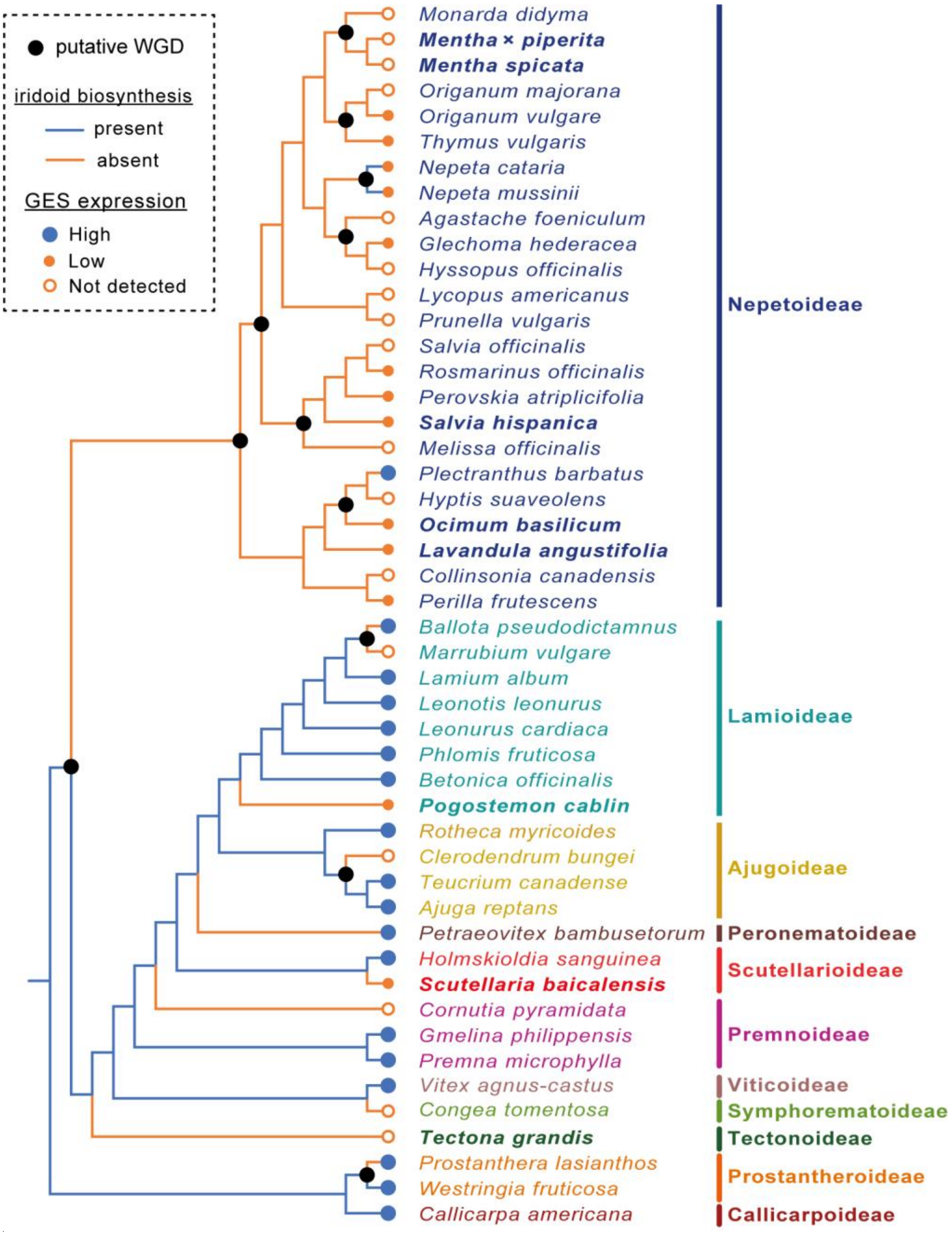

Figure 3. Iridoid biosynthesis in the Lamiaceae. Species tree of Lamiaceae, adapted from [6] and [58]. Putative whole genome duplication (WGD) events (black circles) are those supported by >2 of 3 methods as reported in [58]. Notice the increased prevalence of WGDs in the species-rich Nepetoideae clade. Presence of iridoid metabolites depicted with colored branches (blue=present, orange=absent). Geraniol synthase (GES) orthogroup expression levels from leaf transcriptomes [6] and depicted as circles on the tips (blue = high expression [z-score > 1], orange solid = low expression [z-score < 1], orange outline = gene not detected). Notice how absence of iridoid metabolites is associated with reduction or loss of GES expression. Despite lack of GES expression in *Tectona grandis*, a gene encoding GES remains on the genome. Species names in bold have had their genome, or another species' genome in their genus, sequenced.

References

1. Kim J, Buell CR. A Revolution in Plant Metabolism: Genome-Enabled Pathway Discovery. Plant physiology2015. p. 1532-9.

2. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV. Current strategies of polyploid plant genome sequence assembly. Frontiers in Plant Science2018. p. 1-15.

3. Li FW, Harkess A. A guide to sequence your favorite plant genomes. Applications in Plant Sciences2018. p. 1-7.

4. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nature Communications: Springer US; 2018. p. 541.

5. Schmidt B, Hildebrandt A. Next-generation sequencing: big data meets high performance computing. Drug Discovery Today: Elsevier Ltd; 2017. p. 712-7.

6. Consortium MEG. Phylogenomic Mining of the Mints Reveals Multiple Mechanisms Contributing to the Evolution of Chemical Diversity in Lamiaceae. Molecular Plant: Chinese Society for Plant Biology; 2018. p. 1084-96.

7. Initiative OTPT. One thousand plant transcriptomes and the phylogenomics of green plants. Nature2019.

8. Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, et al. Investigation of terpene diversification across multiple sequenced plant genomes. Proceedings of the National Academy of Sciences2015. p. E81-E8.

9. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Research2017. p. W55-W63.

10. Nützmann HW, Huang A, Osbourn A. Plant metabolic clusters – from genetics to genomics. New Phytologist2016. p. 771-89.

11. Chae L, Kim T, Nilo-Poyanco R, Rhee SY. Genomic signatures of specialized metabolism in plants. Science. 2014;344(6183):510-3. doi: 10.1126/science.1252076. PubMed PMID: 24786077.

12. Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, et al. The butterfly plant arms-race escalated by gene and genome duplications. Proceedings of the National Academy of Sciences of the United States of America2015. p. 8362-6.

13. Miyamoto K, Fujita M, Shenton MR, Akashi S, Sugawara C, Sakai A, et al. Evolutionary trajectory of phytoalexin biosynthetic gene clusters in rice. The Plant journal2016. p. 293-304.

14. Sheehan H, Feng T, Walker-Hale N, Lopez-Nieves S, Pucker B, Guo R, et al. Evolution of L-DOPA 4,5-dioxygenase activity allows for recurrent specialisation to betalain pigmentation in Caryophyllales. New Phytologist2019.

15. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of Gene Duplication in Plants. Plant physiology2016. p. 2294-316.

16. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science. 2014;345(6201):1181-4. Epub 2014/09/06. doi: 10.1126/science.1255274. PubMed PMID: 25190796.

17. Yang Y, Moore MJ, Brockington SF, Mikenas J, Olivieri J, Walker JF, et al. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. New Phytologist2018. p. 855-70.

18. Hodgson H, De La Peña R, Stephenson MJ, Thimmappa R, Vincent JL, Sattely ES, et al. Identification of key enzymes responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production. Proceedings of the National Academy of Sciences2019. p. 201906083.

19. Sherden NH, Lichman B, Caputi L, Zhao D, Kamileen MO, Buell CRR, et al. Identification of iridoid synthases from <i>Nepeta</i> species: Iridoid cyclization does not determine nepetalactone stereochemistry. Phytochemistry: Elsevier Ltd; 2018. p. 48-56.

20. Johnson SR, Bhat WW, Bibik J, Turmo A, Hamberger B, Consortium EMG, et al. A database-driven approach identifies additional diterpene synthase activities in the mint family (Lamiaceae). Journal of Biological Chemistry2019. p. 1349-62.

21. Cai L, Xi Z, Amorim AM, Sugumaran M, Rest JS, Liu L, et al. Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. New Phytologist2019. p. 565-76.

22. Smith SA, Brown JW, Yang Y, Bruenn R, Drummond CP, Brockington SF, et al. Disparity, diversity, and duplications in the Caryophyllales. New Phytologist2018. p. 836-54.

23. Unruh SA, McKain MR, Lee YI, Yukawa T, McCormick MK, Shefferson RP, et al. Phylotranscriptomic analysis and genome evolution of the Cypripedioideae (Orchidaceae). American Journal of Botany2018. p. 631-40.

24. Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. Current Opinion in Plant Biology: Elsevier Ltd; 2016. p. 159-65.

25. Brockington SF, Yang Y, Gandia-Herrero F, Covshoff S, Hibberd JM, Sage RF, et al. Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. New Phytologist2015. p. 1170-80.

26. Kersey PJ. Plant genome sequences: past, present, future. Current Opinion in Plant Biology: Elsevier Ltd; 2019. p. 1-8.

27. Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nature Plants: Springer US; 2018. p. 879-87.

28. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, et al. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nature communications2018. p. 4844.

29. Schmidt MHW, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, et al. De novo assembly of a new Solanum pennellii accession using nanopore sequencing. Plant Cell2017. p. 2336-48.

30. Hoang PNT, Michael TP, Gilbert S, Chu P, Motley ST, Appenroth KJ, et al. Generating a high-confidence reference genome map of the Greater Duckweed by integration of cytogenomic, optical mapping, and Oxford Nanopore technologies. Plant Journal2018. p. 670-84.

31. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Efficient <em>de novo</em> assembly of eleven human genomes using PromethION sequencing and a novel nanopore toolkit. bioRxiv. 2019:715722. doi: 10.1101/715722.

32. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540-6. Epub 2019/04/03. doi: 10.1038/s41587-019-0072-8. PubMed PMID: 30936562.

33. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020. Epub 2020/01/24. doi: 10.1093/bioinformatics/btaa025. PubMed PMID: 31971576.

34. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15(8):e1007273. Epub 2019/08/23. doi: 10.1371/journal.pcbi.1007273. PubMed PMID: 31433799; PubMed Central PMCID: PMCPMC6719893 Nanopore Technologies conferences. Anthony Schmitt and Siddarth Selvaraj are employees of Arima Genomics, a company commercializing Hi-C DNA sequencing technologies.

35. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, et al. The Juicebox Assembly Tools module facilitates <em>de novo</em> assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv. 2018:254797. doi: 10.1101/254797.

36. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20(1):278. Epub 2019/12/18. doi: 10.1186/s13059-019-1910-1. PubMed PMID: 31842956; PubMed Central PMCID: PMCPMC6912988.

37. Tang AD, Soulette CM, Baren MJv, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of <em>SF3B1</em> mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. bioRxiv. 2018.

38. Grassa CJ, Wenger JP, Dabney C, Poplawski SG, Motley ST, Michael TP, et al. A complete Cannabis chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. bioRxiv2018. p. 458083.

39. Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, et al. The opium poppy genome and morphinan production. Science2018. p. eaat4096.

40. Liu Z, Suarez Duran HG, Harnvanichvech Y, Stephenson MJ, Schranz ME, Nelson D, et al. Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. New Phytol. 2019. Epub 2019/11/27. doi: 10.1111/nph.16338. PubMed PMID: 31769874.

41. Fang C, Luo J. Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. Plant Journal2019. p. 91-100.

42. Wu S, Tohge T, Cuadros-Inostroza Á, Tong H, Tenenboim H, Kooke R, et al. Mapping the Arabidopsis Metabolic Landscape by Untargeted Metabolomics at Different Environmental Conditions. Molecular Plant2018. p. 118-34.

43. Zhou S, Kremling KA, Bandillo N, Richter A, Zhang YK, Ahern KR, et al. Metabolome-scale genome-wideassociation studies reveal chemical diversity and genetic control of maize specialized metabolites. Plant Cell2019. p. 937-55.

44. Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, et al. Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated *Solanum tuberosum*. Plant Cell. 2016;28(2):388-405. doi: 10.1105/tpc.15.00538. PubMed PMID: 26772996; PubMed Central PMCID: PMCPMC4790865.

45. Winzer T, Gazda V, He Z, Kaminski F, Kern M, Larson TR, et al. A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. Science. 2012;336(6089):1704-8. Epub 2012/06/02. doi: 10.1126/science.1220757. PubMed PMID: 22653730.

46. Dong AX, Xin HB, Li ZJ, Liu H, Sun YQ, Nie S, et al. High-quality assembly of the reference genome for scarlet sage, Salvia splendens, an economically important ornamental plant. GigaScience: Oxford University Press; 2018. p. 1-10.

47. He Y, Peng F, Deng C, Xiong L, Huang ZY, Zhang RQ, et al. Data descriptor: Building an octaploid genome and transcriptome of the medicinal plant pogostemon cablin from lamiales. Scientific Data: The Author(s); 2018. p. 1-11.

48. He Y, Xiao H, Deng C, Xiong L, Nie H, Peng C. Survey of the genome of Pogostemon cablin provides insights into its evolutionary history and sesquiterpenoid biosynthesis. Scientific Reports: Nature Publishing Group; 2016. p. 1-10.

49. Malli RPN, Adal AM, Sarker LS, Liang P, Mahmoud SS. De novo sequencing of the Lavandula angustifolia genome reveals highly duplicated and optimized features for essential oil production. Planta: Springer Berlin Heidelberg; 2019. p. 251-6.

50. Rastogi S, Kalra A, Gupta V, Khan F, Lal RK, Tripathi AK, et al. Unravelling the genome of Holy basil: An "incomparable" "elixir of life" of traditional Indian medicine. BMC Genomics: ???; 2015.

51. Upadhyay AK, Chacko AR, Gandhimathi A, Ghosh P, Harini K, Joseph AP, et al. Genome sequencing of herb Tulsi (Ocimum tenuiflorum) unravels key genes behind its strong medicinal properties. BMC Plant Biology: BMC Plant Biology; 2015. p. 1-20.

52. Vining KJ, Johnson SR, Ahkami A, Lange I, Parrish AN, Trapp SC, et al. Draft Genome Sequence of <i>Mentha longifolia</i> and Development of Resources for Mint Cultivar Improvement. Molecular Plant: Elsevier Ltd; 2017. p. 323-39.

53. Xu H, Song J, Luo H, Zhang Y, Li Q, Zhu Y, et al. Analysis of the Genome Sequence of the Medicinal Plant <i>Salvia miltiorrhiza</i>. Molecular Plant2016. p. 949-52.

54. Yasodha R, Vasudeva R, Balakrishnan S, Sakthi AR, Abel N, Binai N, et al. Draft genome of a high value tropical timber tree, Teak (Tectona grandis L. f): insights into SSR diversity, phylogeny and conservation. DNA Research2018. p. 409-19.

55. Zhang G, Tian Y, Zhang J, Shu L, Yang S, Wang W, et al. Hybrid de novo genome assembly of the Chinese herbal plant danshen (<i>Salvia miltiorrhiza</i> Bunge). GigaScience: GigaScience; 2015. p. 62.

56. Zhao D, Hamilton JP, Bhat WW, Johnson SR, Godden GT, Kinser TJ, et al. A chromosomal-scale genome assembly of <i>Tectona grandis</i> reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. GigaScience2019. p. 1-10.

57. Zhao Q, Yang J, Cui MY, Liu J, Fang Y, Yan M, et al. The Reference Genome Sequence of Scutellaria baicalensis Provides Insights into the Evolution of Wogonin Biosynthesis. Molecular Plant2019. p. 935-50.

58. Godden GT, Kinser TJ, Soltis PS, Soltis DE. Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints. Genome Biology and Evolution2019. p. X.

59. Huang AC, Kautsar SA, Hong YJ, Medema MH, Bond AD, Tantillo DJ, et al. Unearthing a sesterterpene biosynthetic repertoire in the Brassicaceae through genome mining reveals convergent evolution. Proceedings of the National Academy of Sciences of the United States of America2017. p. E6005-E14.

60. Franke J, Kim J, Hamilton JP, Zhao D, Pham GM, Wiegert-Rininger K, et al. Gene Discovery in Gelsemium Highlights Conserved Gene Clusters in Monoterpene Indole Alkaloid Biosynthesis. ChemBioChem2019. p. 83-7.

61. Cui G, Duan L, Jin B, Qian J, Xue Z, Shen G, et al. Functional divergence of diterpene syntheses in the medicinal plant Salvia miltiorrhiza. Plant Physiology2015. p. 1607-18.

62. Guo J, Zhou YJ, Hillwig ML, Shen Y, Yang L, Wang Y, et al. CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts. Proceedings of the National Academy of Sciences of the United States of America2013. p. 12108-13.

63. Matsuba Y, Nguyen TTH, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, et al. Evolution of a Complex Locus for Terpene Biosynthesis in <i>Solanum</i>. The Plant Cell2013. p. 2022-36.

64. Escape from Adaptive Conflict follows from weak functional trade-offs and mutational robustness, (2012).

65. Guo J, Ma X, Cai Y, Ma Y, Zhan Z, Zhou YJ, et al. Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. New Phytol. 2016;210(2):525-34. Epub 2015/12/20. doi: 10.1111/nph.13790. PubMed PMID: 26682704; PubMed Central PMCID: PMCPMC4930649.

66. Heskes AM, Sundram TCM, Boughton BA, Jensen NB, Hansen NL, Crocoll C, et al. Biosynthesis of bioactive diterpenoids in the medicinal plant Vitex agnus-castus. Plant J. 2018;93(5):943-58. Epub 2018/01/10. doi: 10.1111/tpj.13822. PubMed PMID: 29315936; PubMed Central PMCID: PMCPMC5838521.

67. Pateraki I, Andersen-Ranberg J, Jensen NB, Wubshet SG, Heskes AM, Forman V, et al. Total biosynthesis of the cyclic AMP booster forskolin from Coleus forskohlii. Elife. 2017;6. Epub 2017/03/16. doi: 10.7554/eLife.23001. PubMed PMID: 28290983; PubMed Central PMCID: PMCPMC5388535.

68. Du T, Niu J, Su J, Li S, Guo X, Li L, et al. SmbHLH37 functions antagonistically with smMYC2 in regulating jasmonate-mediated biosynthesis of phenolic acids in salvia miltiorrhiza. Frontiers in Plant Science2018. p. 1-13.

69. Zhang X, Luo H, Xu Z, Zhu Y, Ji A, Song J, et al. Genome-wide characterisation and analysis of bHLH transcription factors related to tanshinone biosynthesis in Salvia miltiorrhiza. Scientific Reports: Nature Publishing Group; 2015. p. 1-10.

70. Zhang Y, Xu Z, Ji A, Luo H, Song J. Genomic survey of bZIP transcription factor genes related to tanshinone biosynthesis in Salvia miltiorrhiza. Acta Pharmaceutica Sinica B: Elsevier B.V.; 2018. p. 295-305.

71. Ding K, Pei T, Bai Z, Jia Y, Ma P, Liang Z. SmMYB36, a Novel R2R3-MYB Transcription Factor, Enhances Tanshinone Accumulation and Decreases Phenolic Acid Content in Salvia miltiorrhiza Hairy Roots. Scientific Reports2017. p. 1-15.

72. Lichman BR, Kamileen MO, Titchiner GR, Saalbach G, Stevenson CEM, Lawson DM, et al. Uncoupled activation and cyclisation in catmint reductive terpenoid biosynthesis. Nature Chemical Biology: Springer US; 2019. p. 71-9.

73. Zhou W, Brockmoller T, Ling Z, Omdahl A, Baldwin IT, Xu S. Evolution of herbivore-induced early defense signaling was shaped by genome-wide duplications in Nicotiana. Elife. 2016;5. Epub 2016/11/05. doi: 10.7554/eLife.19531. PubMed PMID: 27813478; PubMed Central PMCID: PMCPMC5115867.

74. Shoji T. The Recruitment Model of Metabolic Evolution: Jasmonate-Responsive Transcription Factors and a Conceptual Model for the Evolution of Metabolic Pathways. Frontiers in Plant Science2019. p. 1-12.

75. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. Genome Biology: Genome Biology; 2019. p. 1-23.

76. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nature Genetics: Springer US; 2019. p. 1044-51.

77. Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, et al. Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci U S A. 2019;116(6):2344-53. doi: 10.1073/pnas.1817074116. PubMed Central PMCID: PMCPMC6369796.

78. Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, et al. The sequenced angiosperm genomes and genome databases. Frontiers in Plant Science2018. p. 1-14.

79. Pluskal T, Torrens-Spence MP, Fallon TR, De Abreu A, Shi CH, Weng JK. The biosynthetic origin of psychoactive kavalactones in kava. Nat Plants. 2019;5(8):867-78. Epub 2019/07/25. doi: 10.1038/s41477-019-0474-0. PubMed PMID: 31332312.

80. Kellner F, Kim J, Clavijo BJ, Hamilton JP, Childs KL, Vaillancourt B, et al. Genome-guided investigation of plant natural product biosynthesis. The Plant journal : for cell and molecular biology. 2015;82(4):680-92. Epub 2015/03/12. doi: 10.1111/tpj.12827. PubMed PMID: 25759247.

81. Xu S, Brockmoller T, Navarro-Quezada A, Kuhl H, Gase K, Ling Z, et al. Wild tobacco genomes reveal the evolution of nicotine biosynthesis. Proc Natl Acad Sci U S A. 2017;114(23):6133-8. doi: 10.1073/pnas.1700073114. PubMed PMID: 28536194; PubMed Central PMCID: PMCPMC5468653.

82. Zhang, F. P., Yang, Q. Y., Wang, G., & Zhang, S. B. (2016). Multiple functions of volatiles in flowers and leaves of *Elsholtzia rugulosa* (Lamiaceae) from southwestern China. *Scientific reports*, *6*, 27616.

83. Ahmad, H., & Matsubara, Y. I. (2019). Antifungal effect of Lamiaceae herb water extracts against Fusarium root rot in Asparagus. *Journal of Plant Diseases and Protection*, 1-8.

84. Linhart, Y. B., Gauthier, P., Keefover-Ring, K., & Thompson, J. D. (2015). Variable phytotoxic effects of Thymus vulgaris (Lamiaceae) terpenes on associated species. *International Journal of Plant Sciences*, *176*(1), 20-30.

85. Sakai, A., & Yoshimura, H. (2012). Monoterpenes of *Salvia leucophylla*. *Current bioactive compounds*, *8*(1), 90-100.

86. Mishra, L. K., Sarkar, D., & Shetty, K. (2019). Human Health-Relevant Bioactives and Associated Functionalities of Herbs in the Lamiaceae Family. *Functional Foods and Biotechnology: Sources of Functional Foods and Ingredients*.