DITERPENOID BIOSYNTHESIS IN *CALLICARPA AMERICANA* AND THE LAMIACEAE

By

Emily Rose Lanier

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biochemistry and Molecular Biology – Doctor of Philosophy
Molecular Plant Sciences – Dual Major

2023

ABSTRACT

Terpenoids are the largest and most diverse group of plant specialized metabolites. Within the plant, these metabolites often function as tools for biological interactions, such as pollinator attraction, herbivore repellants, mediation of the microbial community, and plant-plant communication. Additionally, recent work has uncovered evidence that terpenoids may also be important in how plants handle abiotic stresses such as drought. Humans have recognized the extensive bioactivities of plant-derived terpenoids for thousands of years and have put them to use as fragrances, flavorings, medicines, insect repellants, and psychoactives. With modern genetic sequencing tools, the past two decades have ushered in a large-scale effort to unravel the biosynthetic pathways towards terpenoids, especially those with industrially relevant bioactivities. Since specialized terpenoid biosynthesis is often lineage-specific, this typically involves non-model plant species. In addition to contributing to our basic understanding of plant specialized metabolism, biosynthetic pathway studies can also identify enzymes which are prime candidates for biotechnological production of useful terpenoids. The numerous chiral centers and oxidations of these compounds often prevent facile access via traditional organic synthesis methods. In this dissertation, I investigate diterpenoid biosynthesis in the mint (Lamiaceae) family plant *Callicarpa americana* (American Beautyberry), which has several reported bioactive diterpenoids. Many of the findings in this plant become a starting point for further elucidation of diterpenoid biosynthesis in related species across the mint family. First, I elucidate the gateway enzymes in diterpenoid biosynthesis based on the generation of a high-quality genome for *C. americana*. This laid the foundation for further study of diterpenoid biosynthesis as well as leading to the discovery of a large diterpenoid biosynthetic gene cluster (BGC) in the genome. Next, I

investigate this BGC along with my coauthor Abby Bryson and we find that a version of this BGC is present in at least six additional Lamiaceae species genomes. Based on the evolutionary distance of these species, we conclude that this BGC has been conserved from an early Lamiaceae ancestor and has played an important role in the evolution of diterpene biosynthesis in this plant family. Functional characterization of the BGC genes in *C. americana* reveals the pathway to a previously inaccessible terpene backbone, (+)-kaurene, in addition to diterpene synthases and cytochrome P450s which catalyze formation of abietane diterpenoids. In the last two chapters, I search for enzymes involved in the clerodane-type diterpene pathways in *C. americana*. I discover along with coauthor Nick Schlecht that a set of orthologous P450s classified as CYP76BK1s are key to formation of furanoclerodanes in all but one Lamiaceae subfamily, including *C. americana*. Second, I generate additional transcriptomic resources for two additional *Callicarpa* species as well as a trichome-specific dataset for *C. americana.* This enables discovery of three short-chain dehydrogenases which also play a key role in furanoclerodane biosynthesis in *C. americana.* Together, this work represents an important advance in understanding diterpenoid biosynthesis in *C. americana* as well as the wider Lamiaceae family.

ACKNOWLEDGEMENTS

This dissertation would not be possible without the support of all of those who have been my teachers, mentors, and friends. First a big thank you to my advisor Dr. Bjoern Hamberger, who has shared his immense knowledge and love of terpenoids while being an incredibly supportive mentor. To all past and present Hamberger lab members, thank you for creating a fun and friendly scientific community. The Molecular Plant Science community, started by Brad Day, has also been a wonderful group for making friends and talking science. Special thanks to my collaborating authors, Abby Bryson, Nick Schlecht, and Dr. Trine Andersen, for your hard work and partnership in accomplishing our work together. And to Trine, you been a great mentor and friend and I couldn't have managed without your expertise, sanity, and willingness to help me through this degree and especially writing this dissertation.

I would also like to thank the collaborators who have lent their knowledge and help to this work. Dr. Robin Buell, Dr. John Hamilton, and the rest of the Buell lab have been indispensable in sharing their expertise with mint genomes. The staff at the MSU Core Facilities have also been incredibly gracious and helpful. Dr. Anthony Schilmiller and Dr. Casey Johnny supported all of my metabolomics work and kept the instruments running; Dr. Kevin Childs assisted with RNA sequencing and genomics questions; Dr. Daniel Holmes assisted with NMR sample analysis and structure elucidations; and Amy Albin at the microscopy facility gave us gorgeous trichome images. I would also like to thank Jessica Lawrence, who has been so helpful over the years in navigating all of the administrative hurdles of the PhD. To my advisory committee, Dr. Michaela TerAvest, Dr. Tom Sharkey, Dr. Erich Grotewold, and Dr. Claire Vieille, thank you for all of your input and support in getting my work to this point. This work would not have been possible

without my funding sources: The National Science Foundation Graduate Research Fellowship, and the University Enrichment Fellowship.

Finally, I would like to thank my family for supporting me in every way from the beginning. To my dad, the original Dr. Lanier, thank you for teaching me to love science and to always question everything. To my mom, thank you for giving me your love of plants. I never guessed that the Beautyberry you planted in our driveway would one day be the foundation of my academic career. And to my husband Chris Testerman, you deserve an honorary degree for all the many hours you have spent listening to me dissect my experiments and cry over the failed ones. Thank you for loving me so well.

TABLE OF CONTENTS

CHAPTER 1: PLANT TERPENE SPECIALIZED METABOLISM: COMPLEX NETWORKS OR SIMPLE LINEAR PATHWAYS?

Emily R. Lanier[†1], Trine Bundgaard Andersen[†1] and Björn Hamberger[*1]

[1] Department of Biochemistry & Molecular Biology
Michigan State University
Molecular Plant Sciences Building
1066 Bogue Street, East Lansing, MI 48824, USA
Phone 517-884-6964;
[†]These authors contributed equally.
*Corresponding author: hamberge@msu.edu

This chapter, except the final section "Thesis Rationale and Overview", was first published in:

Lanier, E.R., Andersen, T.B., and Hamberger, B. Plant terpene specialized metabolism: Complex networks or simple linear pathways? *The Plant Journal*, 2023, doi: 10.1111/tpj.16177

Author contributions:

ERL, TBA and BH conceived and wrote the manuscript. ERL and BH wrote the introduction, ERL and BH wrote the Diterpenoids section, TBA and BH wrote the mono/sesquiterpenoids section. ERL generated figures for the diterpenoid section, and TBA generated the figures for the mono/sesquiterpenoid section. All authors contributed to revisions.

Abstract

From the perspectives of pathway evolution, discovery, and engineering of plant specialized metabolism, the nature of the biosynthetic routes represents a critical aspect. Classical models depict biosynthesis typically from an end-point angle and as linear, e.g., connecting central and specialized metabolism. As the number of functionally elucidated routes increased, the enzymatic foundation of complex plant chemistries became increasingly well understood. The perception of linear pathway models has been severely challenged. With a focus on plant terpenoid specialized metabolism, we review here illustrative examples supporting that plants have evolved complex networks driving chemical diversification. The completion of several diterpene, sesquiterpene and monoterpene routes shows complex formation of scaffolds and their subsequent functionalization. These networks show that branch points, including multiple sub-routes, mean that metabolic grids are the rule rather than the exception. This concept presents significant implications for biotechnological production.

Key words: Terpene synthase, cytochrome P450, metabolic network, terpene specialized metabolism, UDP dependent glycosyl-transferases, biosynthetic gene cluster, enzyme promiscuity.

Significance statement: Traditionally plant specialized metabolism was regarded as a series of linear routes. Here we investigated the emerging theme of complex networks through representative examples across the major classes of terpenoids. All pathways covered in this review show branch points diversifying their chemistries. Consequently, pathway discovery should consider enzyme promiscuity, alternative routes, and how family expansion and enzyme repurposing can drive the evolution of new phytochemistries.

Introduction

Reasoning and scope of this review

In the wider metabolic network that contains the entirety of plant metabolism, specialized metabolism has generally been viewed as a separate branch of enzymatic reactions producing metabolites which are not critical for the "central" functions needed for plant survival. The main roles of specialized metabolites appear to be adaptation to the environment, i.e., biotic, and abiotic interactions. These include on the one side plant defense against herbivores and pathogens or attraction of pollinators, natural predators of herbivores, seed dispersers, and microbial symbionts. On the other side they also play critical roles in abiotic interactions such as drought stress and protection from UV light [1]. The importance and nature of these traits was already recognized in the second half of the 19th century, long before the beginning of what we now know as plant chemical ecology [2]. In modern times the boundaries between specialized metabolites and hormones, signaling molecules required for regulation and lastly metabolites vital for growth and development are becoming increasingly hazy. Specialized metabolites are emerging as multifunctional, i.e., playing roles in regulation of growth (flavonoids), or defense (callose deposition, glucosinolates, benzoxazinoids) or have examples of catabolism and re-integration into central metabolism [3].

Humans have recognized the bioactive nature of medicinal plants for thousands of years, extracting and combining numerous plant components into medicines [4]. With the advent of modern chemical techniques, it became possible to isolate and characterize single bioactive components [5]. This idea of single-compound medicine underpins the modern pharmaceutical paradigm, and thus it has also driven the initial approach to plant biosynthetic pathway discovery.

In this approach, a linear pathway is often presumed, or at least sought, with the goal of producing a single molecule of interest with high specificity. Yet the past few decades of biosynthetic pathway research have shown that many compounds are in fact the result of metabolic grids governed by promiscuous enzymes leading to complex mixtures of related natural products. As a result, it appears plausible that many plant biosynthetic pathways are geared towards chemical diversity rather than singular end products. Likewise, pharmaceutical research has now recognized the limits of single-molecule treatments in the face of complex, multi-target human diseases, and multi-component treatments such as those designed based on traditional Chinese medicinal recipes are now finding traction in modern combination therapy [6,7]. This network hypothesis is also supported by the sheer diversity of specialized metabolites reported from plants, notwithstanding the many thousands of compounds yet to be reported. Recent work investigating the ecological underpinnings of phytochemical diversity suggest that plants generate a chemical array to target the variety of generalist and specialist pests in their environment [8–10]. Linear pathways would require separate enzymatic and regulatory structures for each compound, but metabolic grids of promiscuous enzymes can generate tens of products per enzyme. This is far more efficient both in terms of metabolic cost as well as the ability to tweak chemistry quickly on an evolutionary time scale. In stark contrast to this strategy of generating chemical diversity stands the general, i.e. primary metabolism, where networks of often functionally redundant but differentially expressed genes lead to single end products. This principle is exemplified by the shikimate pathway yielding the amino acid phenylalanine and the subsequent core phenylpropanoid pathway gate keeper of routes to lignin, soluble and wall-bound phenolics and flavonoids [11–13]. In this review, we will focus on terpenoids, the largest class

of plant specialized metabolites. We give a brief overview of the history of terpene biosynthetic pathway discovery and how rapidly evolving technologies have enabled probing of metabolic networks to an extent never before possible. Examination of select examples from the field of terpenoid pathway discovery, provides evidence that generally, plant specialized metabolism is composed of complex interacting networks and not simple linear pathways.

Biochemistry of terpene biosynthesis, early pathways to the scaffolds

Several excellent reviews and studies cover the nature and evolution of plant terpene biosynthesis, and the rise of chemical diversity, so we will cover this only briefly [14–19]. Terpenoids (terpenes or isoprenoids) are derived from the five-carbon building blocks isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). These are synthesized through the mevalonate (MVA) pathway and the more recently discovered methylerythritol phosphate (MEP) pathway. The MVA pathway is localized to the cytosol and shares common ancestry with the animal, fungal, and archaeal kingdoms [20]. The MEP pathway is localized to the plastid and thus is of prokaryotic origin. Prenyl transferases are enzymes which condense IPP and/or DMAPP into longer prenyl diphosphates utilized by terpene synthases (TPSs). Typically, this is a head to tail condensation with double bonds retaining *trans*- configuration, but there are examples of irregular and *cis*-prenyl transferases [21–24]. TPSs then catalyze complex carbocation cascade reactions on the prenyl diphosphate substrate, resulting in cyclic or linear terpene backbones. The carbocation cascade is initiated by either removal of the diphosphate (class I TPSs), or by protonation of the distal double bond (class II TPSs). The shape (dictating substrate and intermediate conformations) and electronics (stabilization of carbocations) of the active site residues then control the sequence of the carbocation cascade. The final product is determined

by which potential energy minima is reached on the reaction path as well as the method of carbocation quenching, which may be through water addition, protonation, or deprotonation (Tantillo, 2011). The nature of terpene formation relying on this delicate energy control has great importance to the evolution of terpene biosynthesis. The control of reaction product can be altered by few or even a single amino acid switch [26–29]. There are also many examples of multi-product TPSs, where TPS-directed carbocation cascades may be biased towards one product but allow for several minor products to result. This phenomenon is outlined in more detail in the following sections.

Increasing complexity, functionalization, and decoration of terpenes

The nature of terpene biosynthesis lends itself naturally to biosynthetic networks. A single or few changes in the amino acid sequence of a TPS can profoundly change its product profile [26,30,31]. TPSs may also produce arrays of dozens of products [32]. After cyclization, the terpene backbone is often further decorated by other biosynthetic enzymes, such as cytochrome P450 monooxygenases (P450s), 2-oxoglutarate dependent dioxygenases (2-ODDs), acyl transferases and glycosyl transferases [33]. This increases the polarity of the molecules and often adds the most potent bioactive moieties. Specifically, P450s, drivers of the evolution of terpene specialized metabolites, are known to be promiscuous and may oxidize their substrates in multiple positions as well as oxidizing multiple related backbones (reviewed in Hamberger and Bak 2013; Bathe and Tissier 2019). This promiscuity with related intermediates in the same pathway is key to creating complex chemical networks. The ability for one enzyme to accept multiple related substrates increases flexibility of the network and exponentially expands the number of products that can be made by just a few enzymes. P450s are also recognized for their role as gatekeepers in

pathway bifurcations of industrially relevant bioproducts [36]. Thus, it is common in phytochemical studies for many related compounds to be identified from a single plant.

History of terpene pathway discovery

Recognizing their nutritional value, the biosynthesis of carotenoid pigments were among the first terpenes which attracted attention in the mid 50s [37,38]. Early studies relied on labelling studies, cell free extracts and purified enzymes [39] in a field that gained traction. The terpenes investigated expanded significantly, including mint oil terpenes, insecticidal pyrethrins and remarkably the identification of the early committed steps in cyclization and oxidation of the growth hormone gibberellin [40–43]. The mint family (Lamiaceae), exceptionally rich in terpenes, represents an illustrative example of how the terpene research in plants evolved over the next two decades with citation records increasing from 48 annual articles in 1968 to 365 in 1999 (https://pubmed.ncbi.nlm.nih.gov/; terpene, biosynthesis, plant). Pioneered by the lab of Rodney Croteau, this period saw the enzymatic cyclization from common acyclic precursors to key terpenes in peppermint, sage and thyme [44–46], elucidation of multistep routes to more complex products [47,48], and demonstration of the function of a cytochrome P450 in a multistep pathway [49]. This emerging theme of multistep pathways was applied to the discovery of the first committed step in the biosynthesis of paclitaxel (taxol®) through isolation of a cDNA clone from a cDNA library and heterologous expression of the encoded taxadiene synthase and demonstration of a P450 activity involved in the first hydroxylation [50,51]. With that, Croteau's lab groundbreaking work arguably gave birth to the new field of genomics driven terpene pathway discovery and bioengineering of isoprenoid biosynthesis, including modification of the mint oil composition [52–54].

Technologies accelerating pathway discovery

The next two decades in pathway discovery which are in focus of this review are driven by emerging genomic technologies, notably the early cloning and sequencing of cDNA libraries (or their fragments, expressed sequence tag (EST) libraries), hybridization of labelled cDNA to arrayed sequences (i.e., microarrays), whole genome shotgun sequencing and recently new high-throughput sequencing technologies (i.e., 454, illumina, nanopore, pacbio). The vast amount of data generated creates two pillars for the discovery of pathways: (1) the genetic underpinning and composition of genomes and transcriptomes and (2) organ, cell-type and condition-specific expression. Sequencing, assembly and release of the Arabidopsis genome in 2000 [55] was a harbinger of the explosion in plant genome and transcriptome sequencing capabilities that exponentially followed. The emerging framework of the genome led already in 1997 to the first report of a terpene synthase-like gene with similarity to cDNA sequences from gymnosperms, Lamiaceae and Solanaceae on a 23.9-kb fragment from the long arm of chromosome IV [56]. Cloning and heterologous expression of an identified cDNA clone showed activity as multiproduct monoterpene synthase [57]. Mapping of terpene synthase (TPS) sequences in the genome identified a small family in Arabidopsis and allowed to propose a model for the evolutionary history of plant TPSs from general to specialized metabolism [58]. The complement of 40 TPS genes was reported briefly thereafter, including their classification into six phylogenetic subfamilies TPS-a throughout -f, following an earlier proposed scheme [59,60]. This work also reported the first gene clusters in terpene specialized metabolism. The next year showed the dramatic advances with two milestones achieved, a functional link between biosynthesis and floral emission of terpenes from Arabidopsis flowers and the first metabolically engineered transgenic Arabidopsis

lines [61,62]. Findings in these early studies deploying functional genomics remained instructive for the coming decades.

A functional linkage between non-homologous genes which together form metabolic pathways can be inferred from co-expression across tissues and cell-types, in time course experiments, or under other conditions that elicit pathway activity. An important tool accelerating the discovery of pathways and establishing these links between genes with unknown function became the Affymetrix *A. thaliana* ATH1 microarray. Thousands of global transcript datasets were deposited in public databases and represented a treasure trove. The large family of P450s represented an ideal target, as they are highly diverse and at that time only a fraction was functionally assigned. Ehlting and co-workers re-annotated over 4,000 genes, and built and mined the extensive 'CYPedia' database for co-expressed pathways [63]. This revealed, among other findings, highly coordinated expression of P450s from the CYP71 clan under pathogen stress and elicitation and in triterpene pathways. The team led by Anne Osbourn fully validated the bioinformatic predictions with characterization of the thalianol pathway and the report of the corresponding biosynthetic gene cluster, while the above manuscript was under evaluation [64]. The well-cited database provided a highly valuable tool in the next decade for identification of cryptic and highly complex monoterpene pathways, with an example for formation of flower volatiles described in detail in the respective section below [65]. It is intriguing that while co-expression analyses corroborated numerous instances of genomically clustered functionally characterized pathways, such as those in *A. thaliana* or *O. sativa* [66,67] it did not support the inverse, pathway discovery based on genomic clustering alone, highlighting that the combination of both technologies are critical [68]. This early strategy of global co-expression analyses also provided the blueprint for

similar work approaching non-model plants with increasingly accessible transcriptome panels, but without accessible genomes [68]. Together, genomic technologies and co-expression strategies have enabled characterization of numerous terpenoid biosynthetic pathway enzymes over the past decade. In the following sections we explore some examples of these discoveries to evaluate the evidence for network-like organization of specialized terpenoid pathways.

Diterpenoids

Diterpenoids are derived from an acyclic 20-carbon prenyl diphosphate precursor. This is most commonly geranylgeranyl diphosphate (GGPP), though an all-*cis* precursor, nerylneryl diphosphate, has been identified in a few plants [24,69,70]. Both the prenyl diphosphate synthase and the diterpene synthase (diTPS) contain N-terminal targeting peptides for the chloroplast, so the precursors for these compounds are derived primarily from the MEP pathway. Some diterpenes are made by a single class I diTPS. However, the bicyclic labdane-type diterpenes require the sequential action of a class II followed by a class I diTPS [71]. This is exemplified by the gibberellins, central metabolites which proceed from GGPP via *ent*-copalyl diphosphate (*ent*-CPP) to *ent*-kaurene. Labdane-type diTPSs in specialized metabolism have evolved from this role in central metabolism through duplication and neofunctionalization. There are now 20 diterpene diphosphates identified [72] made by the class II, or copalyl diphosphate synthase (CPS)-type diTPSs (TPS-c family). These can be converted to over 7,000 individual scaffolds through the action of the class I, or kaurene synthase-like (KSL) diTPSs (TPS e/f family) [71]. Class I diTPSs introduce the first layer of promiscuity into diterpene biosynthesis, as several have been shown to accept multiple diterpene backbones [73–75]. This can contribute to network-type biosynthesis within a single plant, though sometimes specific diterpene pathways are separated by tissue specific

expression. This promiscuity has also enabled production of new-to-nature diterpene backbones by combining class II and class I diTPSs from different plants together in a heterologous host[74, 75]. Though the genetic sequencing advances of the past two decades have enabled a surge in diterpene biosynthesis studies, only a tiny fraction of the over 23,000 diterpenes so far identified from plants (Dictionary of Natural Products, March 2022) have known pathways.

Tanshinones

The abietanes are a widely present class of tricyclic diterpenoids with significant and numerous bioactivities reported [76]. One of the most well-studied pathways is for the tanshinones, a class of bioactive diterpenoid components of the Chinese medicinal herb Danshen (*Salvia miltiorrhiza*). Because tanshinone biosynthesis has been recently reviewed [77], we will not detail this body of work in-depth here. However, the network aspects of key enzymes identified so far are notable (Fig. 1.1). The biosynthesis of tanshinones begins with a classic labdane pathway. A class II diTPS catalyzes formation of (+)-copalyl diphosphate ((+)-CPP), and a class I diTPS cyclizes this into miltiradiene, a tricyclic abietane-type backbone 5/30/23 2:04:00 PM. This backbone is capable of auto-oxidation to the aromatic abietatriene. From there, three P450s (CYP76AH1, CYP76AH3, and CYP76AK1) have been identified that together catalyze oxidations on four different carbons. Initially, CYP76AH1 catalyzes hydroxylation of C12 on abietatriene to afford ferruginol [78].

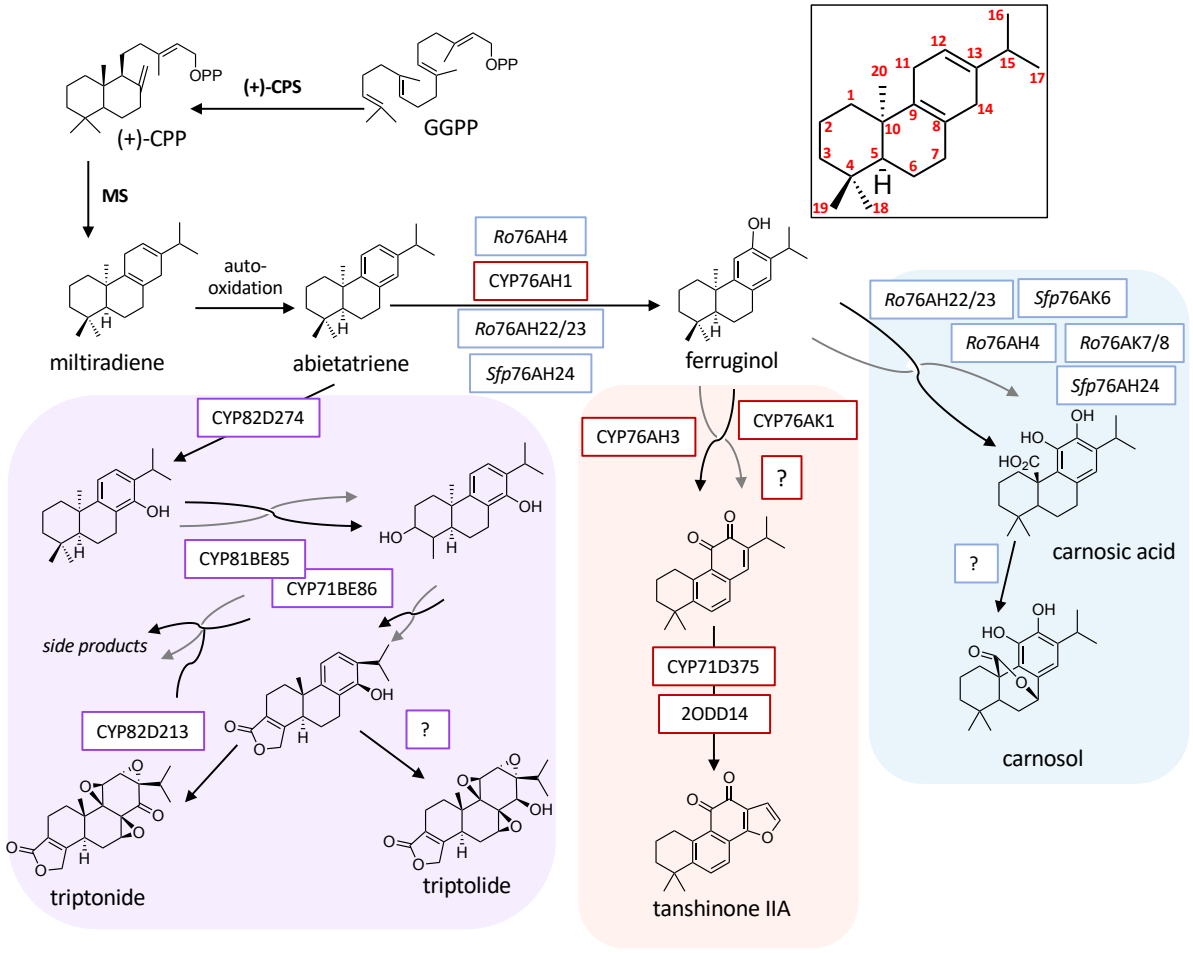


Figure 1.1. Overview of tanshinone, carnosic acid, and triptolide biosynthesis. All three of these angiosperm abietane pathways begin with a class II diTPS ((+)-CPS) that converts GGPP to (+)-copalyl diphosphate followed by a miltiradiene synthase (MS). Sequential steps are represented by a single arrow Curved double arrows indicate steps where multiple enzymes can interact with multiple substrates in a network fashion, i.e., a P450 may oxidize multiple positions and/or the same position on multiple substrates. For clarity, not all enzyme products from network steps are shown. Question marks represent steps that have not yet been elucidated. Red boxes indicate tanshinone biosynthesis in *S. miltiorrhiza*, blue boxes indicate carnosic acid biosynthesis, and purple boxes indicate triptolide/triptonide biosynthesis in *T. wilfordii*. Species names are included where the pathway enzymes are from multiple species (*Sfp*, *Salvia fruticosa/pomifera*; *Ro*, *Rosemary officinalis*). Inset shows standard numbering for the abietane backbone. Not all side products and intermediates are shown.

This step is required for additional oxidations. Subsequently, CYP76AH3 and CYP76AK1 represent bifurcations and can oxidize their substrate either in various positions, including multiple oxidation steps, or accept a range of different substrates, yielding an array of oxidation

products[79]. Another set of enzymes, CYP71D375 (and CYP71D373) and a 2-oxoglutarate dehydrogenase (Sm2ODD14), has been shown to catalyze furan ring formation for the classic tetracyclic tanshinone backbone [80,81]. Additional oxidation and demethylation steps have not yet been elucidated. Decades of phytochemical studies have shown that there are at least 40 different tanshinone diterpenoids present in *S. miltiorrhiza* [82]. Many of these contain the oxidations and ring closures so far elucidated, but there is additional variation as well. Based on the work so far presented, the most likely explanation is that a sizeable fraction of the hundreds of P450s and other biosynthetic type enzymes predicted in the *S. miltiorrhiza* genome are involved in a larger network [83]. While the pathway proceeds in a linear fashion up to ferruginol, it afterwards winds circuitously towards many end products, facilitated by the promiscuity of the P450s. Some steps mingle while others require specific substrates. The most abundant and/or bioactive tanshinones are the goal of biosynthetic elucidation, but minor products, as well as specific metabolic intermediates along the way may represent a core part of the plant's defense strategy.

The earlier stage of the oxidation network in tanshinone biosynthesis is present in similar form in other species of the mint (Lamiaceae) family. Another important abietane diterpenoid, carnosic acid, is found in high abundance in rosemary (*Rosmarinus officinalis*) and sage (*Salvia spp.*)[84,85]. An orthologous CYP76AH enzyme produces ferruginol in carnosic acid biosynthesis (*R. officinalis, S. pomifera, S. fruticosa*). Additional CYP76AH and CYP76AK family members catalyze promiscuous oxidations leading towards a network of carnosic acid and related compounds [86,87] (Fig. 1.1). Carnosol is another abundant and related compound in these plant species, but the biosynthetic link between carnosic acid and carnosol has not yet been elucidated.

Resin acids

Another subset of bioactive abietane diterpenoids are conifer diterpene resin acids (Fig. 1.2). Although the product array is less complex than observed in the tanshinone and carnosic acid chemical families, the network has been more completely mapped and demonstrates a key example of a multi-product diterpenoid pathway. In gymnosperms, the branch of labdane scaffolds is primarily synthesized by bi-functional diterpene synthases which contain both the class II and class I active sites (TPS-d family). The grand fir (*Abies grandis*) abietadiene synthase was the first of the diTPSs identified over 25 years ago, together with a proposed pathway to the carboxylic acid and remains a model enzyme for a broad range of studies with focus on structure-function relationship, catalytic mechanism and is notably one of the very limited number of plant dips crystallized to date [88,89]. As the biochemistry of the pathway has been reviewed extensively, we will focus here on two noteworthy milestones defining the metabolic grid of resin acid biosynthesis. The enzyme from Norway spruce (*Picea abies*) received attention for its capacity (among other orthologs) to afford a mixture of abietadiene, levopimaradiene, neoabietadiene, and palustradiene, all double bond isomers of the shared labdane skeleton [90]. Through an extensive series of careful examinations Keeling and co-workers demonstrated that the reaction mechanism of *Pa*LAS proceeds via water quenching of the abieta-8(14)-en-13-yl carbocation, resulting in 13-hydroxy-8(14)-abietene, detected by cold on-column injection and a lower final oven temperature. Dehydration of this product during work-up, or conventional GC-MS analysis then results in the non-enzymatically derived spectrum of products. P450s of the CYP720B subfamily (CYP720B1 and CYP720B4) had been shown to catalyze the three-step oxidation of the diterpene olefins to their carboxylic acids [91,92].

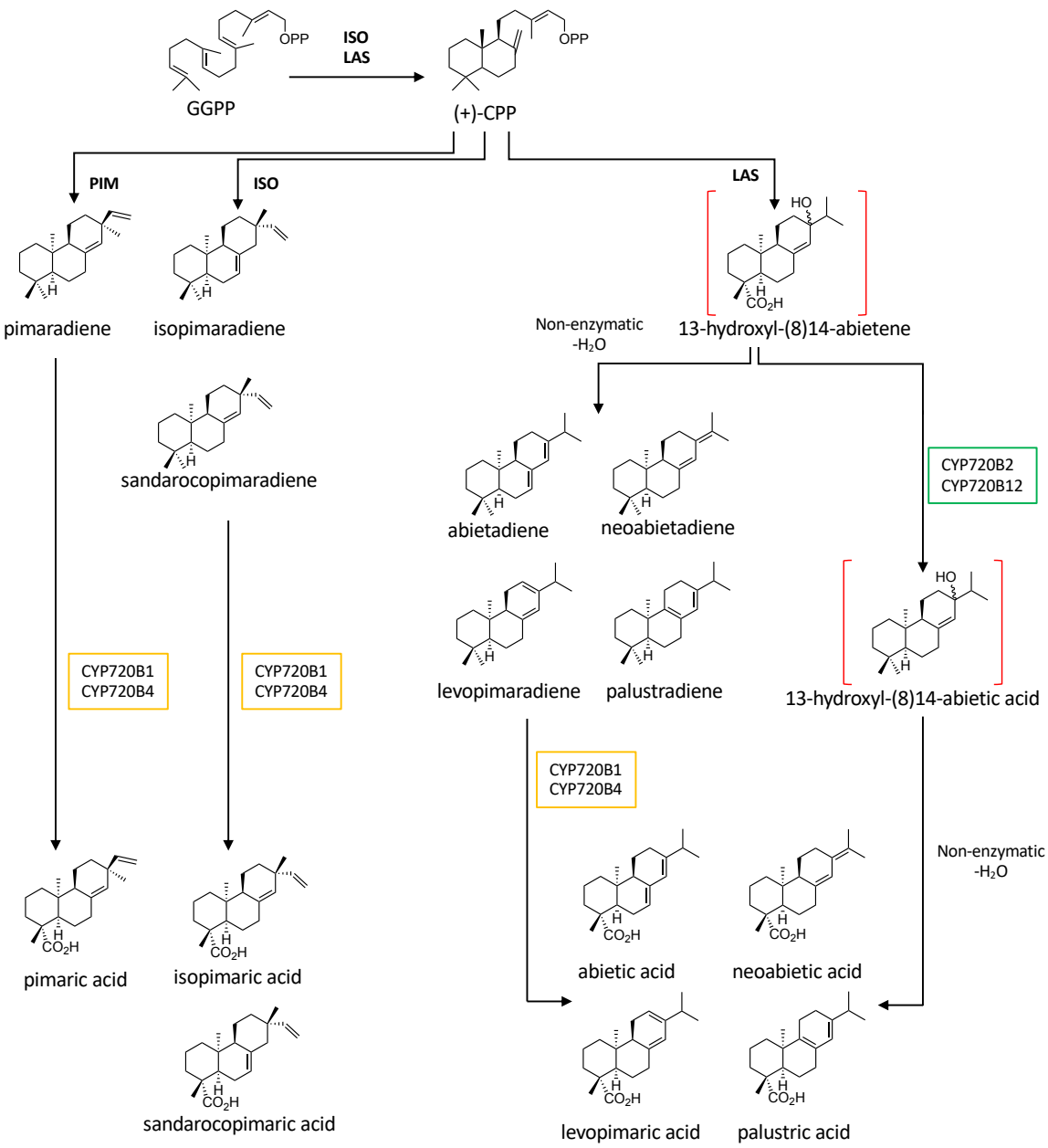


Figure 1.2. Overview of diterpene resin acid biosynthesis. Adapted from Geisler *et al.* (2016). The bifunctional diTPSs isopimaradiene synthase (ISO), pimaradiene synthase (PIM), and levopimaradiene/abietadiene synthase (LAS) convert GGPP via the intermediate (+)-CPP into the various diterpene olefins. In the case of LAS, the initial and unstable diTPS product is 13-hydroxy-8(14)-abietene (indicated by red brackets). Studies of the CYP720B subfamily showed that one clade, CYP720B1 and CYP720B4 (outlined in orange), converts a range of diterpene olefins to the corresponding acid. However, another clade, CYP720B2 and CYP720B12 (outlined in green), can only accept 13-hydroxy-8(14)-abietene to the corresponding 13-hydroxy-8(14)-abietic acid, which then undergoes dehydration to form abietic acid, levopimaric acid, neoabietic acid, and palustric acid.

However, a fascinating facet of the biological relevance of the unstable intermediates was discovered later, when a different clade of CYP720Bs (containing CYP720B2 and CYP720B12) was found inactive with the olefin scaffolds. These were ultimately shown to accept the unstable diterpene synthase product 13-hydroxy-8(14)-abietene, affording the same diterpene resin acids characteristic for the profiles in conifers [93]. In this case a network is created in two ways: In the first, two P450s promiscuously oxidize the known diterpene olefins, where in the second, a single unstable intermediate is oxidized before decomposing into a similar set of diterpene resin acids. These findings emphasize the modularity and combinatorial nature of this diterpene specialized pathway.

Forskolin

The biochemistry leading to the cyclic AMP activator and epoxy-labdane type forskolin has been reviewed [36]. Hence, we will here only briefly summarize the route and focus instead on the relevance of promiscuous enzymes in creating chemical diversity through a metabolic grid, with significant implications for biotechnological applications using those plant enzymes. Early signs for more complex, rather than linear routes, were found with the complement of enzymes catalyzing formation of key diterpene scaffolds in Indian coleus (*Coleus forskohlii*). Conspicuously, the epoxy-labdane manoyl oxide and its derivative forskolin accumulate in the roots of the plant, while diterpenoids derived from the common olefin miltiradiene are detected in aerial parts of the plant. This was shown to be seemingly controlled by differently expressed modules, consisting of two distinct and tissue-specific class II diTPSs coupled with a pair of broadly

expressed class I diTPS homologs. Swapping out the enzyme catalyzing the initial cyclization can then give rise to the different terpene scaffolds [94].

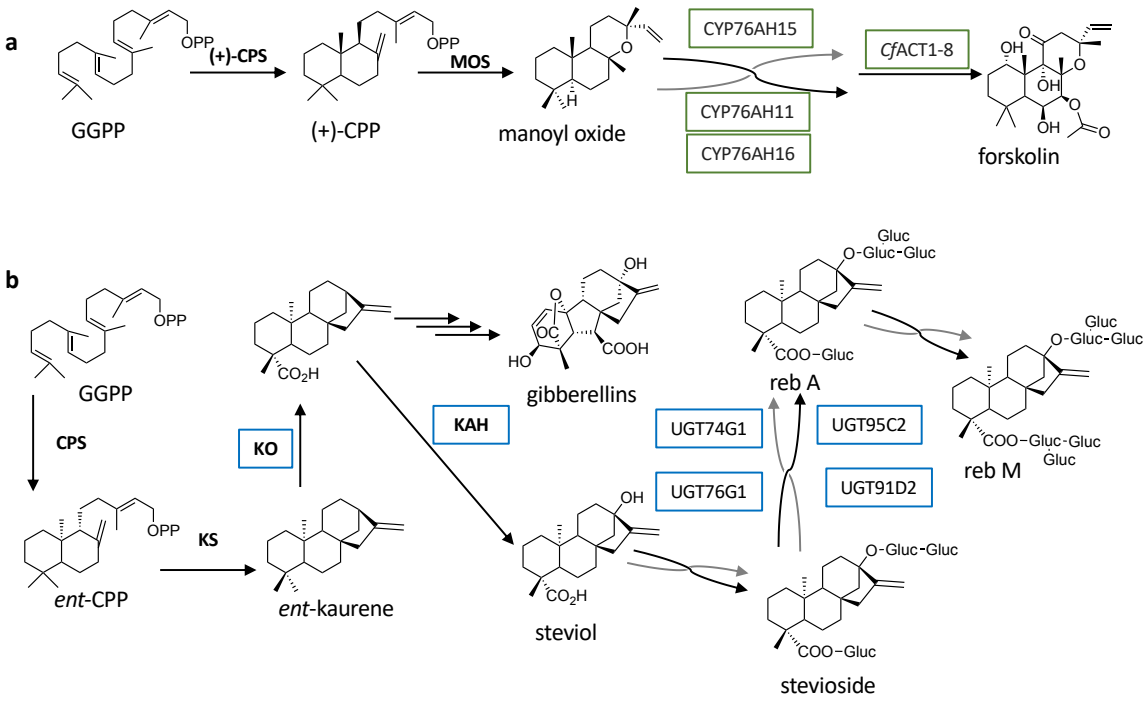


Figure 1.3. Overview of (a) forskolin and (b) steviol glycosides biosynthesis. Sequential steps are represented by a single arrow. Curved double arrows indicate steps where multiple enzymes can interact with multiple substrates in a network fashion, i.e., a P450 may oxidize multiple positions and/or the same position on multiple substrates. For clarity, only a few representative enzyme products from network steps are shown. CPS, *ent*-CPP synthase; KO, kaurene oxidase; KAH, kaurenoic acid 13-hydroxylase.

In the roots, the subsequent conversion of manoyl oxide to forskolin requires six regio- and stereospecific monooxygenations and a single regiospecific acetylation. Following the hypothesis that the corresponding enzymes catalyzing formation of forskolin should be co-expressed in the root, a panel of seven P450s of the CYP76AH-subfamily and two BAHD-type acyltransferases were found [95]. Upon testing of the P450s nearly all were found to oxidize manoyl oxide, together giving rise to a broad spectrum of products, including multi-oxygenated forms clearly not intermediates

of the hypothesized (linear) route to forskolin. One prolific P450 yielded 11 discrete products, including three hydroxyl groups and a keto group. In biotechnological applications such promiscuity would translate to near-complete loss of control, yet in plants this may be an important contribution to chemical diversity. In contrast, a second P450 catalyzed regioselective formation of keto-manoyl oxide in a configuration present in both forskolin and many forskolin/manoyl oxide derived diterpenoids. Using this P450 as an anchor, all other P450s were tested first in pairs, then in triplets in all permutations. Again, up to 19 different products were detected in these assays, indicating that *in planta* diterpenoid biosynthesis is anything but linear. However, and at this stage unexpected, a single combination of three P450s yielded a small and single peak of deacetyl-forskolin, the second last intermediate towards forskolin. Of ten acyl-transferases identified and tested, two resulted in acetylation of deacetyl-forskolin. These enzymes are notoriously promiscuous, so it was again not surprising to see a broad range of acetylated products with one, forskolin representing a minor fraction. In contrast, the second enzyme exhibited high activity and specificity with efficient conversion of the intermediate to forskolin and lacking detectable side products. With that, a minimal set of enzymes was reported constituting the entire and specific linear route from GGPP to forskolin and opening the door for biotechnological production (Fig. 1.3a) [95].

Triptolide

A final example of abietane diterpenoids is triptolide (Fig. 1.1). The suggested biosynthetic routes to rearranged (4→3) abeo-abietane type diterpenes in the root and root cultures of *Tripterygium wilfordii* represent an instructive case leading to a diverse portfolio of complex decorated related products [96,97]. A hallmark of this group is the bioactive triepoxide triptolide [98]. While a linear route

was early suggested, a metabolic grid appeared plausible [96,97]. Functional characterization of all candidate diTPSs from the TPS-e/f subfamily afforded specific combinations of class II/I diTPSs giving access to the kaurane, manoyl-oxide and kolavenyl scaffolds, characteristic for *T. wilfordii*. Markedly, functional redundancy was detected in both class II and class I diTPSs, which together form a metabolic grid to the diterpene scaffolds. The lack of functional combinations yielding abietane-type diterpenes prompted analysis of TPSs outside the canonical TPS-e/f subfamily. This led ultimately to the discovery of TwTPS27, a class I TPS of the TPS-b subfamily, typically harboring monoterpene synthase activity [99]. Like its other class I counterparts, this enzyme was shown to accept multiple substrates and may contribute abietane and manoyl-oxide scaffolds in specific stereochemistry to the biosynthetic capacity of the *T. wilfordii* diTPSs [99]. Even though production of the postulated triptolide intermediate dehydroabietic acid was demonstrated in yeast, using the proxy CYP720B4 from Sitka spruce [100], it took another three years until the corresponding gene encoding CYP728B70 was identified from the *T. wilfordii* genome [101]. The breakthrough came with the elegant demonstration of a cascade of four different P450s of the two different subfamilies, 82D and CYP71BE, identified from over 60 candidates with activity on the abietane-type miltiradiene. Drawing on parallels with the earlier reported forskolin pathway, the team now led by Hansen and Andersen-Ranberg who identified the diterpene cyclases in *T. wilfordii* focused on an anchor P450 (*Tw*CYP82D274) catalyzing the first step in the pathway, which afforded the aromatic hydroxylated intermediate 14-hydroxy-dehydroabietadiene [102]. Guided by promiscuous activities of *C. forskohlii* enzymes, homology, and the recognition that most P450s oxidizing diterpene specialized metabolites reside in the CYP71 clan, they identified two P450s (*Tw*CYP71BE85 and *Tw*CYP71BE86) to oxygenate the A-ring of 14-hydroxy-

dehydroabietadiene resulting in a C4→C3 methyl shift and lactone formation. Notably, both were found to additionally produce a broad panel of oxygenated and polyoxygenated diterpenes with the starting diterpene scaffold miltiradiene or the first oxidized intermediate. Testing of orthologs of the first P450 established *Tw*CYP82D213 as the last step in the route giving access to the triptolide derivative triptonide, next to a plethora of oxidized miltiradiene products. Indeed, the authors acknowledge that the striking substrate and product promiscuity of the *Tw*P450s could suggest that instead of a simplistic model of a linear pathway, triptonide may be one of the many products of a metabolic grid in *T. wilfordii* and engineered biotechnological hosts [102].

Macrocyclic diterpenes

In addition to the two-step class II/class I diTPS pathways leading to the labdane-type diterpenes, there are single step class I diTPSs capable of catalyzing the transformation of GGPP into unique diterpene scaffolds [14]. These are often members of the TPS-a family, which are typically sesquiterpene synthases, but can acquire a transit peptide and evolve to convert GGPP in the plastid [103]. One important example of this phenomenon are the macrocyclic diterpenes found extensively in the Euphorbiaceae plant family. Several of these compounds have strong bioactivity and pharmaceutical applications [104]. Investigations into the diTPSs across multiple Euphorbiaceae species determined that casbene synthase likely produces the precursor to most macrocyclic diterpenes in this plant family [105–107]. Casbene is a 14:3 bicyclic diterpene that was hypothesized to proceed through the 5:11:3 lathyrane skeleton to other ring configurations that require additional oxidations and ring closures [106]. The highly oxidized nature of the macrocyclic diterpenoids found in Euphorbiaceae suggested involvement of P450s. With a combination of

genomic and transcriptomic sequencing, the CYP726 subfamily was found to be key in conversion of casbene to jolkinol C, the first lathyrane intermediate, demonstrating the ability of a P450 to catalyze carbon-carbon bond formation in a ring closure (Fig. 1.4). King *et al.* [108] found the first active P450s in a biosynthetic gene cluster (BGC) containing diTPSs, P450s, and other terpene associated genes in castor bean (*Ricinus communis*), later confirmed by Boutanaev *et al.* [109]. In characterizing these enzymes, they found that multiple P450s (CYP726A14, CYP726A17, and CYP726A18) could oxidize the 5-position of casbene to either a hydroxy or keto group. CYP726A16 could then add a 7,8-epoxy group to 5-keto casbene, but not casbene, requiring linear and sequential oxidation. Later, this group investigated a similar BGC in another Euphorbiaceae species, *Jatropha curcas*. In this BGC, they found orthologous CYP726As capable of oxidation at the 5 and 6 positions and another P450, CYP71D495, that catalyzed production of 9-keto casbene [110]. The combination of the 5,6- and 9-hydroxylases led to the production of jolkinol C, likely through non-enzymatic tautomerisation of the enol group followed by an intramolecular aldol reaction that creates the key C-C bond. Similarly, Luo and co-workers found in the same year a CYP71D/CYP726A pair from *Euphorbia peplus* that catalyzed the same oxidation reactions [111]. However, in this instance a short-chain alcohol dehydrogenase (ADH) was also required for formation of jolkinol C. Careful *in vitro* experiments in this study demonstrated that the hydroxylations and oxidations from hydroxy to ketone could proceed in any order and combination of P450s and the ADH, network fashion. However, the presence of all three invariably led to jolkinol C production when expressed *in planta*, though not in yeast, ostensibly due to the different pH than plant cells [112]. Wong and co-workers optimized expression of both the *E. peplus* and *J. curcas* P450s in yeast and demonstrated production of jolkinol C in titers up

to ~800 mg/L, and validated that while the ADH was not critical for jolkinol C formation, it did significantly increase the titer [113]. Fattahian and co-workers recently reviewed the extensive literature on jatrophane and other cembrene-derived diterpenoids, which contains numerous hypotheses from synthetic studies as to how subsequent ring transformations may occur [114]. The heavily oxidized structures isolated from Euphorbiaceae species indicate that these biosynthetic pathways likely rely on additional oxidative enzymes, such as ADHs, to attain these configurations. Further variation exists in the acylation of these compounds, as the polyester form is typically isolated from the plant. With over 500 structures reportedly isolated, it is apparent that some degree of network exists in the biosynthesis. Like other pathways examined thus far, certain steps require exact precursors while others may occur in any order before crossing another committed step. A single plant species often accumulates multiple end products based on the same diterpene backbone, such as casbene in the case of the Euphorbiaceae [104,115]. In this case, the pathway is linear through the biosynthesis of casbene, the common precursor, and then differs in progressively divergent and complex oxidations.

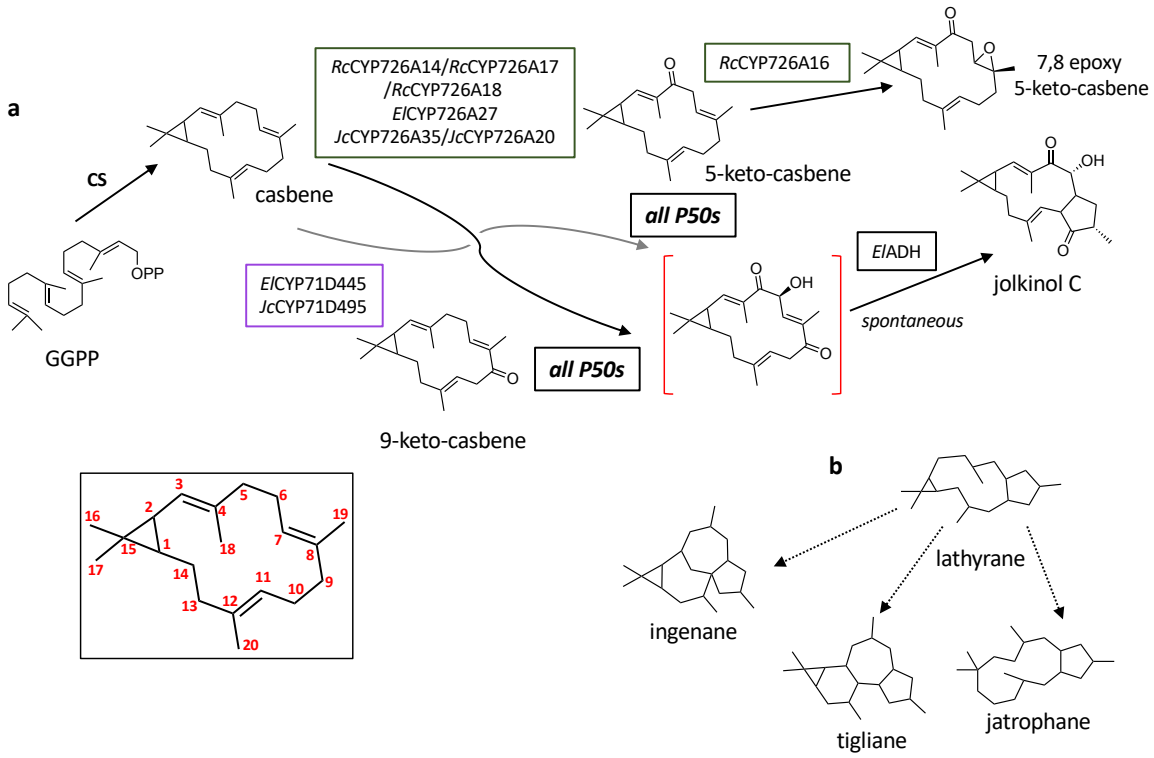


Figure 1.4. (a) Pathway to jolkinol C and (b) hypothesized route to additional macrocyclic diterpene backbones. Sequential steps are represented by a single arrow, while curved double arrows indicate steps where multiple enzymes can interact with multiple substrates in a network fashion. In this pathway, the CYP71Ds (purple box) hydroxylate the 9 position while the CYP726s hydroxylate the 5 position (green box), but this can occur in any order. Both can hydroxylate the 6 position, together forming the unstable intermediate 6-hydroxy-5,9-diketocasbene (in red brackets) which can non-enzymatically rearrange to jolkinol C. The addition of *El*ADH was found to significantly increase jolkinol C production by improving conversion of hydroxy- moieties to keto- groups. For clarity, additional side products of these enzymes are not shown. (b) Conversion of jolkinol C to additional rearranged terpenes has not yet been elucidated and is indicated by the dotted arrows. CS, casbene synthase. Species names are abbreviated as *Rc, Ricinus communis; El, Euphorbia lathyris; Jc, Jatropha curcas.* Inset shows standard casbene numbering.

Beyond TPSs and P450s

Up to this point, the various pathways discussed have relied on TPS or P450 networks to create an array of diterpenoids. Yet some plants may not have expanded P450 and diTPS gene families and rely on other enzyme types to create chemical diversity. One example of this is the steviol glycosides, which accumulate *in planta* to an astounding 25% of leaf dry weight. While over 30

different steviol glycosides have been reported, stevioside and rebaudioside A (Reb A) make up the bulk while the rest accumulate in low to trace amounts [116]. The diterpene core, steviol, is derived from a linear diTPS-CYP pathway that mirrors biosynthesis of the hormone gibberellin until the last step. An *ent*-CPP synthase cyclizes GGPP, and the class I diTPS kaurene synthase catalyzes transformation of *ent*-CPP to *ent*-kaurene [117]. The first oxidation at C19 is catalyzed by kaurene oxidase (CYP701A5) [118]. A recent high quality genome assembly of *Stevia rebaudiana* as well as early pathway discovery work indicate that these first three enzymes are likely present as multiple copies in the genome, allowing for separate regulation for gibberellin versus steviol biosynthesis [119]. From there the two pathways diverge, and the final oxidation at C13 is catalyzed by kaurenoic acid 13-hydroxylase (KAH) resulting in steviol. While the definitive identity of this enzyme in *S. rebaudiana* has not yet been verified, at least two P450s of the CYP716 and CYP714 subfamilies have been reported to have low to moderate KAH activity [116,120]. Additionally, four tandem duplicates in the CYP716 subfamily were found to be highly expressed in *S. rebaudiana* leaves, but have not been functionally verified [119]. In attempts to engineer high titer steviol glycoside production in microbial hosts, native enzymes were replaced in favor of a higher performing KAH (CYP714A2) from *Arabidopsis thaliana* [120,121]. The pathway then diverges into production of at least 14 steviol glycosides via a metabolic grid of four functionally characterized UDP-dependent glycosyl-transferases (UGTs; UGT76G1, UGT74G1, UGT95C2, UGT91D2) [120,122,123]. Hundreds of additional UGT candidates were identified in the genome analysis, of which 86 are expressed in relevant leaf tissues for steviol glycoside biosynthesis [119]. UGTs are known to be quite promiscuous. For example, crude extracts of *A. thaliana* and tobacco have been shown to glycosylate steviol and other pathway intermediates [118]. This raises difficulty in pathway

engineering for a specific glycosyl variant, as is the case for steviol glycosides. Reb M and Reb D are two minor products of the steviol glycoside pathway that are preferred for their superior organoleptic properties to the more naturally abundant stevioside and Reb A [123,124]. However, co-expression of the UGTs needed to produce these compounds results in numerous unwanted side products. Intensive mutational analysis of the relevant UGT (UGT76G1) led to a variant with moderate improvements in product profile, but still multiple products [123]. As with promiscuity of the UGTs exponentially increases the diversity of the product profile, creating maximum output for the plant but inhibiting facile biotechnological use of the pathway.

Monoterpenoids/sesquiterpenoids

Monoterpenoids and sesquiterpenoids share many characteristics and functions with diterpenoids, but also have quite different and distinct roles *in planta*. Monoterpenoids are biosynthesized in the plastids from primarily MEP precursors as seen for diterpenoids, while sesquiterpenoids are generated in the cytosol from the MVA precursor pathway. Despite the two distinct pathways, together monoterpenoids and sesquiterpenoids constitute a large portion of the volatiles that plants emit [125]. Both classes often contribute to floral scents used to attract pollinators or to deter florivores. Volatile terpenoids also play an important part in defense against pathogens and general stress tolerance both above and below ground [126]. One example is caryophyllene, which is emitted from maize roots and attracts beneficial nematodes that feed on the root pest corn rootworm (*Diabortica virgifera*) [127,128]. α-farnesene, which is emitted from apples, is an example of a sesquiterpene that functions as an insecticide while also attracting birds that feed on the apples and aid in seed dispersal [129]. Additionally, some of these terpenoids can also modulate the microbiome of the flower or vice versa [130–133] (reviewed by [134]). While both

hydrocarbons and terpenes with a single hydroxylation are important volatiles, it is interesting to note that 92% of the known monoterpenoids are indeed hydroxylated according to the Dictionary of Natural Products [135]. Despite their generally more volatile nature, mono- and sesquiterpenoids are often also retained in plant tissue and many plants have specialized storage cells, such as trichomes, laticifers and secretory ducts [136–138]. Storage compartments, whether internal or external, are excellent ways to store toxic terpenoids for potential usage against herbivores during tissue rupturing.

Sesquiterpenoids

The majority of sesquiterpene synthases (sesquiTPSs) are soluble proteins located in the cytosol that use (*E,E*)-FPP as a substrate. Nonetheless, a few sesquiTPSs have been found to be exceptions to this rule. In *Nicotiana tabacum* a sesquiTPS (tobacco 5-*epi*-aristolochene synthase, TEAS) was characterized and found to accept (*Z,E*)-FPP in addition to (*E,E*)-FPP as a substrate [139]. In addition, a sesquiTPS from sandalwood (*Santalum album*) produces sesquiterpenes with similar types of structures using both (*E,E*)-FPP and (*Z,Z*)-FPP as substrates [140]. An example of a *Z,Z*-FPP prenyl transferase was discovered in *Solanum habrochaites*. This correlated with a sesquiTPS also identified in *S. habrochaites* that uses (*Z,Z*)-FPP but not (*E,E*)-FPP. The localization of both these enzymes was interestingly found to be in the chloroplasts [141]. SesquiTPS enzymes can produce an astonishing variety of core terpene structures despite being limited in the variation of substrates. In fact, much of the diversity of sesquiterpenoids can be accounted for by the array of sesquiTPS products rather than extensive modifications, as seen for diterpenoids. These products may be cyclized, such as the guaienes (a C5- and C7-membered ring), eudesmanes (two C6-membered rings), germacrenes (a C10-membered ring), and humulenes (a

C11-membered ring); however, many sesquiterpenoids are also linear or monocyclic such as farnesene and bisabolene [142]. Some backbones already contain a hydroxy group due to water quenching of the carbocation cascade, such as kunzeaol produced by TgTPS2 in *Thapsia garganica* [143]. In *Zea mays* (maize) a TPS, *Zm*EDS, was even found that produced eudesmanediol, which as the names indicates contains 2 alcohol groups [144]. The multitude of backbones generated by the sesquiTPSs are due to various reactions that can occur during product formation such as hydride shifts, methyl shifts, rearrangements, re- and de-protonations, in addition to the presence of three double bonds in FPP. Extensive reviews of sesquiTPS product formation can be found in [142,145–147]. One of the most astounding sesquiTPSs is from *Abies grandis* (*Ag*TPS) where 52 different sesquiterpenoid products were identified, though the product profile of the TPS was dominated by γ-humulene while most other products were found in negligible amounts [32]. Unlike most TPSs, *Ag*TPS has two rather than one divalent cation (Mg2+) binding sites (the aspartate rich DDxxD domains) which are located at the entrance of the active site. This may help the enzyme to achieve a much more complex product profile than seen for other multiproduct sesquiTPSs. Due to the sensitive energy balance of carbocation cascades, altering a single amino acid in the active site can significantly impact product profile. A study using a β-sesquiphellandrene synthase from *Persicaria minor* showed that a single mutation changed the product profile to include several hydroxylated sesquiterpenoids [148]. Additional examples of multiproduct TPSs (both mono- and sesquiterpenoids) have been reviewed by Vattekkatte and colleagues [149]. Though multiproduct TPSs are common, many sesquiTPSs produce either one sole product or one major product with only very minor side products. In kiwi fruit (*Actinidia deliciosa*) two examples of such single -product sesquiTPSs were described with either α-farnesene and (+)-

germacrene D as the sole product [150]. Three sesquiTPSs from *Hyoscyamus muticus* exemplified this further, each producing a minimum of 93 % of a single compound compared to the total amount of sesquiterpenoids detected.

Considering the multitude of sesquiTPS products, it is not surprising that sesquiTPSs that share high sequence similarity may in fact produce different compounds. This is seen for two *T. garganica* TPSs, *Tg*TPS1 and *Tg*TPS2, which share 91% sequence identity but produce two different types of sesquiterpenoids, δ-cadinene and kunzeaol respectively. δ-cadinene is a eudesmane with the typical two 6-membered rings while kunzeaol is a hydroxylated germacrene with a 10-membered open ring structure. Additionally, *Tg*TPS1 is a single-product TPS while *Tg*TPS2 produces several sesquiterpenoids in addition to kunzeaol [143]. Various mutation studies show that with just one or two amino acid changes in the active site of a sesquiTPS the product profiles change substantially [151,152]. In *Gossypium arboreum* (cotton) a (+)-δ-cadinene synthase was subjected to a mutation study and it was shown that changing a single amino acid can result in the production of a germacrene D-4-ol in addition to the original compound δ-cadinene [153].

Many plants produce a variety of sesquiterpenoids, and it is common to see a portion of these based solely on one type of backbone structure. These backbones may have been modified by various hydroxylations as seen in the variety of caryophyllanes in *Eremophila spathulata* or acetylations as seen in *T. garganica* [154,155]. Most of the pathways for known sesquiterpenoids have not yet been elucidated, though in some species partial pathways have been found though several further modifications are expected to occur [156,157]. In some of these instances the pathway is linear up until a certain point but branches out during the final modifications, as seen in the diterpenoid pathways.

Like with diterpenoids, P450s are often the next class of enzymes to add further diversity to the sesquiterpenes. P450s often perform a single oxidation on the sesquiterpene backbone though in several cases a cascade of oxidations occurs on the same carbon or same backbone. This is often seen in the formation of sesquiterpene lactones, described below. The promiscuity of P450s also renders them in some cases able to perform several consecutive hydroxylations on various positions on the same backbone. For example, in *S. habrochaites* CYP71D184 can oxidize 7-epi-zingiberene into 9-hydroxy-zingiberene and 9-hydroxy-10,11 epoxyzingiberene [158]. One of the earliest modified sesquiterpenoid pathways to be studied was capsidiol biosynthesis in *N. tabacum*. The first step in the pathway is production of 5-epi-aristolochene by the TPS 5-epiaristolochene synthase (EAS). This is followed by hydroxylations at C1 and C3 by CYP71D20 to form capsidiol [159,160]. Capsidiol biosynthesis is further reviewed in [36]. Another interesting example of P450 promiscuity is found in *S. album*. Sandalwood is known for having an essential oil with fragrant sesquiterpenoids. These pathways have been reviewed previously and are therefore only mentioned in brief here [36]. A network of nine P450s from the CYP76F sub-family was found to hydroxylate santalene and bergamotene backbones into corresponding santalols and bergamotols showing both a large and promiscuous P450 subfamily [161,162].

Some of the most studied sesquiterpenoids belong to the sesquiterpene lactone subgroup due to their bioactivity and pharmaceutical activity [163]. A few examples include sesquiterpenoids from *Tanacetum parthenium*, *T. garganica* and *Artemisia annua* [164–166]. The enzymes responsible for lactone ring formation have been described in all three species but the complexity and the type of enzymes involved in lactone ring formation depend on the pathway and the species in question, each having a unique approach [167–170]. Kauniolide biosynthesis in *T. parthenium* and

artemisinin biosynthesis in *A. annua* are briefly highlighted. While several of the biosynthetic steps for lactone ring formation have been described in various Asteraceae species, most progress has been made in *T. parthenium* which is described here (Fig. 1.5). Kauniolide biosynthesis is initiated by converting FPP to germacrene A, catalyzed by the sesquiTPS germacrene A synthase (GAS) [171]. Germacrene A oxidase (CYP71, GAO) performs a three-step oxidation of germacrene A to yield germacrene A acid (germacra-1(10),4,11(13)-trien-12-oic acid). A second P450, costunolide synthase, (CYP71BL2/COS) catalyzes a hydroxylation of germacrene A acid on position C6. The interesting outcome of these hydroxylations was seen regarding the formation of the lactone ring. The 6-α-germacrene A acid can perform a spontaneous lactonization to yield costunolide, a reoccurring theme. Costunolide is one of the possible precursors of many germacranolide, eudesmanolide and guaianolide sesquiterpenoids. GAS, GAO and COS are found in several Asteraceae including *Cichorium intybus*, *T. parthenium* and *Helianthus annuus*, but were all described in *T. parthenium* [168]. Further steps modifying costunolide were elucidated in *T. parthenium*. In 2014 Liu and co-workers[168] found a CYP71 clan P450, *Tp*PTS/CYP71CA1, forming an epoxide and converting costunolide into parthenolide. This was followed in 2018 by the first discovery of a sesquiterpenoid P450 (*Tp*KLS, CYP71BZ6X) able to perform the ring closure of a germacrene (parthenolide) thereby yielding kauniolide, a member of the diverse guaianolides [172]. Though the pathway as described up until kauniolide appears to be linear, a P450 responsible for two potential side branches was also discovered.

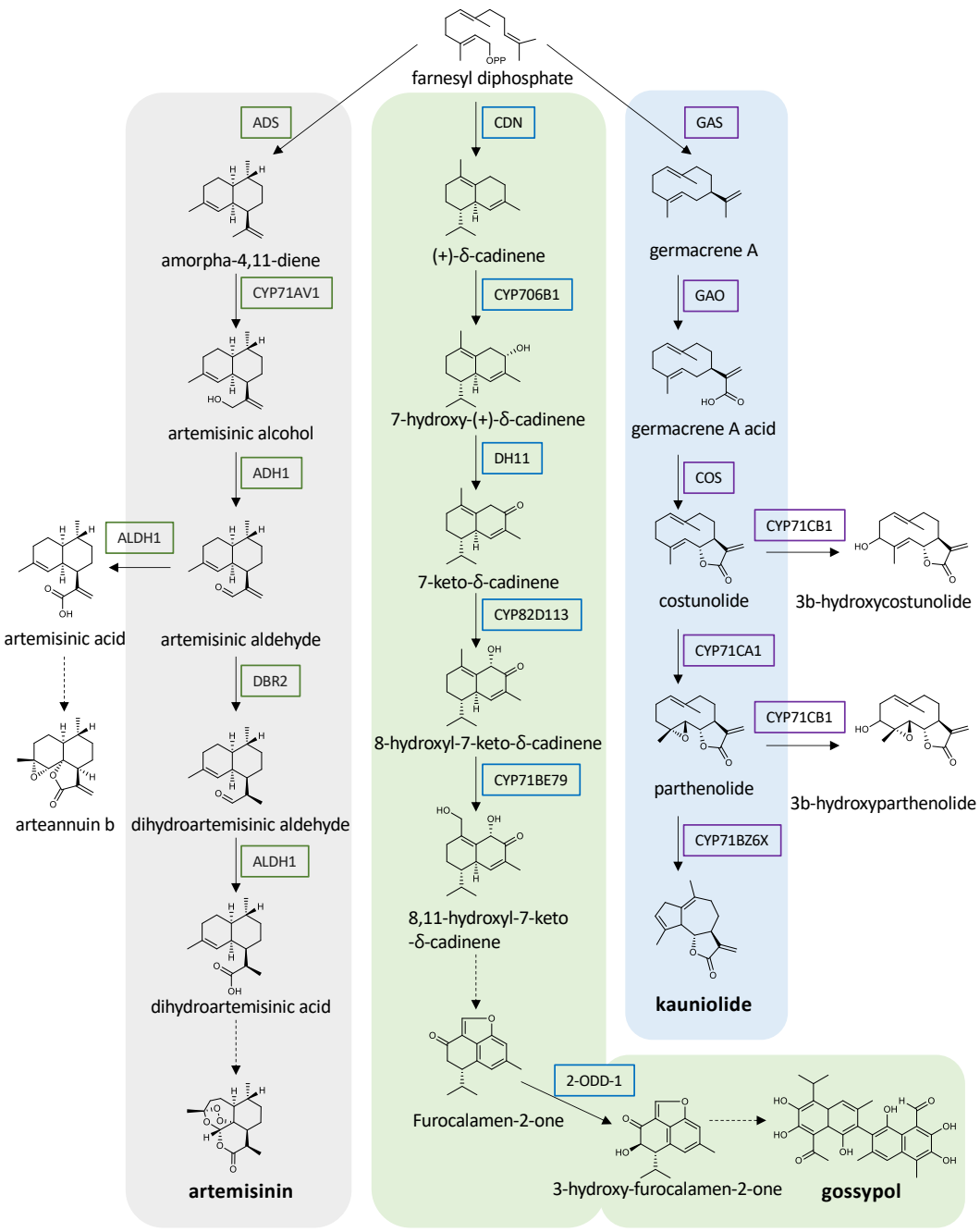


Figure 1.5. Three examples of distinct sesquiterpenoid pathways. The grey box shows the artemisinin biosynthetic pathway in *A. annua*. The green box shows the gossypol pathway in *G. hirsutum*. The blue box shows the kauniolide pathway in *T. parthenium*. Dashed arrows denoted missing steps in the pathways or non-enzymatic steps. For clarity, additional side products of these DH1 in the gossypol pathway are not shown. ADS, amorpha-4,11-diene synthase; ADH1, alcohol dehydrogenase 1; DRB2, artemisinic aldehyde Δ11(13) reductase; ALDH1, aldehyde dehydrogenase 1; CDN, (+)-δ-cadinene synthase; DH1, alcohol dehydrogenase; 2-ODD-1, 2-oxoglutarate/Fe(II)-dependent dioxygenase; GAS, germacrene A synthase; GAO, germacrene A oxidase; COS, costunolide synthase.

Using yeast microsomes expressing CYP71CB1, feeding studies with costunolide and parthenolide found these converted to 3β-hydroxycostunolide and 3β- hydroxyparthenolide, respectively. Both metabolites have indeed been detected in *T. parthenium*, alluding to a role *in planta* awaiting discovery. The genus *Tanacetum* in general produces a variety of different sesquiterpene lactones, though for *T. parthenium* at least costunolide and parthenolide are found in quantities sufficient for detection, parthenolide is furthermore one of the most abundant sesquiterpenoid lactones found in the plant [173]. *T. parthenium* also produces a variety of further decorated guaianolides and it is likely that these are based on modifications of kauniolide sesquiterpenoids and are Artemisinin biosynthesis is an instructive example of a different set of enzymes generating the precursor steps needed for lactone ring formation. Depending on the expression system used for characterizing the individual enzymes involved in artemisinin biosynthesis different activities have been reported [170,174,175]. Here we only describe the pathway overview given by Judd, Xie and coworkers (Fig. 1.5) [176,177]. The first step in artemisinin biosynthesis is catalyzed by the sesquiTPS, ADS, which converts FPP to amorpha-4,11-diene [178,179]. CYP71AV1 then performs a hydroxylation of amorpha-4,11-diene to artemisinic alcohol. Alcohol dehydrogenase 1 (ADH1) converts artemisinic alcohol to artemisinic aldehyde. From here a reduction of artemisinic aldehyde to dihydroartemisinic aldehyde is catalyzed by artemisinic aldehyde Δ11(13) reductase (DRB2). Aldehyde dehydrogenase 1 (ALDH1) is the final enzyme involved and forms dihydroartemisinic acid. The lactone ring is then expected to form spontaneously by photo-oxidation based reactions. Further characterization of enzymes including *in planta* studies conversely showed additional complexity. Several side branches are found with intermediates being converted away from the artemisinin path. These side branches

include the intermediate dihydroartemisinic aldehyde being converted to dihydroartemisinic alcohol by dihydroartemisinic aldehyde reductase (RED1) [180]. Another branch is the conversion of artemisinic aldehyde to artemisinic acid by aldehyde dehydrogenase 1 (ALDH1) from which arteannuin A is formed.

Gossypol, hemigossypolone and heliocides are cadinene type sesquiterpene aldehydes found in cotton (*G. hirsutum*). Significant progress has been made in the discovery of the enzymes responsible for the central backbone, though the final steps to gossypol are still awaiting discovery (Fig. 1.5). The initial biosynthetic step is performed by a (+)-δ-cadinene TPS (CDN) followed by oxidation by CYP706B1 to 7-hydroxy-(+)-δ-cadinene. An alcohol dehydrogenase, DH1, converts 7-hydroxy-(+)-δ-cadinene to 7-keto-δ-cadinene. Two P450s, CYP82D113 and CYP71BE7, yield single oxidations generating 8-hydroxy-7-keto-δ-cadinene and 8,11-dihydroxy-7-keto-δ-cadinene, respectively. While there are several steps missing mid-pathway, virus induced gene silencing (VIGS) led to identification of an additional step further downstream namely a 2-oxoglutarate/Fe(II)-dependent dioxygenase (2-OGD-1) which converted furocalamen-2-one to a new compound, 3-hydroxy-furocalamen-2-one [156]. Additionally, DH1 was found to be promiscuous and uses 8-hydroxy-7-keto-δ-cadinene and 8,11-dihydroxy-7-keto-δ-cadinene as substrates.

As seen for other classes of terpenoids, various sesquiterpenoids have also been found as glycosides. One example is *Dendrobium nobile* where eight guaiene and eudesmane type sesquiterpene glycosides have been isolated to date [181,182]. The sesquiterpenoid field has not progressed as far as the diterpenoid field when it comes to elucidating long and complex pathways especially when it comes to acetylated and glycosylated sesquiterpenoids. Recently,

however, a breakthrough was made in tea plant (*Camellia sinensis* (L.) O. Kuntze). During cold stress in tea plants, the linear and hydroxylated sesquiterpenoid nerolidol accumulated in a glycosylated form and is thought to play a role in cold adaptation. A glucosyltransferase, UGT91Q2, found to be induced by cold stress was indeed shown to produce nerolidol glucoside[183].

Monoterpenoids

The majority of known monoterpenoids are derived from geranyl diphosphate (GPP), produced from one IPP and one DMAPP by geranyl diphosphate synthase (GPPS) in the plastids. In tomato (*Solanum lycopersicum*) a neryl diphosphate synthase (NDPS1) was found to produce neryl diphosphate (NPP) [23]. Monoterpenoids derived from two DMAPP units have also been described and constitute a small subclass known as irregular monoterpenoids [184]. As seen for other classes of terpenoids, monoterpenoids are also present with different types of backbones such as acyclic, monocyclic and bicyclic structures [142]. For regular monoterpenoids, the first committed step is performed by a monoterpene synthase (monoTPS). As seen for many sesquiTPSs, it is not uncommon for monoTPSs to produce multiple products. [149] lists several examples of multi-product monoTPSs, but a few well-known examples include a 1,8-cineole synthase from *Nicotiana suaveolens* that produces six out of the eight flower volatiles found in the plant [185] and a limonene synthase from lavender (*Lavandula angustifolia*) that produces six monoterpenes also found in flowers [186].

The oxidative metabolic networks giving rise to a vast diversity of monoterpenes of biotechnological, commercial, and agricultural relevance has been reviewed for menthol (*Mentha x piperita*), linalool derivatives (*A. thaliana*) and iridoid precursor geraniol (*Catharanthus*

roseus, *A. thaliana*) [36,187]. For an update and elaboration on the vast network of monoterpene functionalization in glandular trichomes of members of the Lamiaceae, we refer the reader to [188]. In the following we will focus on advances in network discovery in this family, which is exceptionally rich and chemically diverse in monoterpene metabolites [189]. A few examples of remarkable monoterpenoid biosynthesis in other families are also highlighted. Increasingly available high-quality transcriptomes have in recent years fueled the discovery and functional annotation of genes that span networks in monoterpenoid metabolism in the mints [190].

Specifically in thyme, three distinct chemotypes were explored to identify the routes to thymol, carvacrol and the further oxidized thymoquinone [191]. Like the pathway to the diterpene resin acids and the macrocyclic jolkinol C described above, thyme provides an example of a route involving unstable intermediates. In thyme, the activity of seven members of the CYP71 subfamily were shown to yield specifically two dienol intermediates from gamma-terpinene, which can spontaneously dehydrate to the aromatic *p*-cymene. However, when combined with a short-chain dehydrogenase two allylic ketone intermediates were detected. These can isomerize via keto-enol tautomerism into the aromatic alcohols thymol and carvacrol. Two P450s of the CYP76S and CYP736A families then can intercept and oxidize these products to the shared product thymohydroquinone. It is currently unclear whether the final oxidation to thymoquinone occurs enzymatically, or spontaneous (Fig. 1.6, adapted from Krause *et al.*, 2021) [191]. In depth analysis of the recently published genome of Chinese native thyme *T. quinquecostatus* genome [192] may provide further insights into the evolution of these pathways. Furthermore, the reported presence of ten and structural elucidation of seven monoterpene glycosides indicates that the

networks are more expansive than just the routes to the oxidized scaffolds, and with yet to be discovered enzymes plausibly of the UGT family [193].

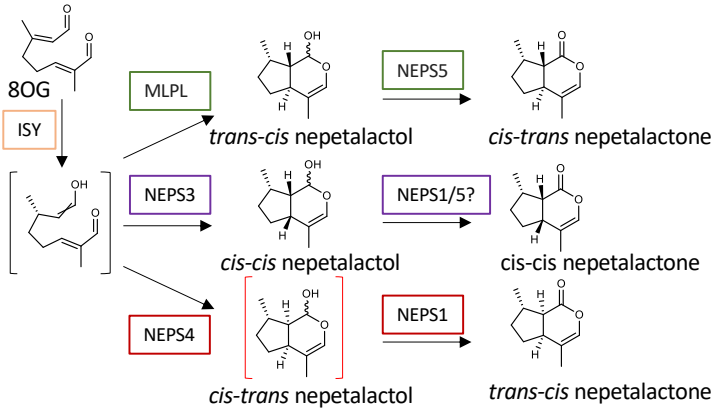


Figure 1.6. Nepetalactone biosynthesis in *Nepeta spp*. Undefined stereochemistry is depicted by a crossed double bond. 8OG, 8-oxogeranial; ISY, iridoid synthase; NEP, nepetalactol-related short-chain reductase/dehydrogenase.

Two illustrative examples of pathway discovery using chromosome-scale genome assemblies are found in nepetalactone catnip (*Nepeta* spp.) [194] and *Mentha longifolia* L. [195]. To investigate the repeated evolution of the unusual volatile nepetalactone iridoid-type monoterpenes, Lichman and co-workers combined a comparative genomic approach, with phylogenetic analysis and enzyme characterization. This specific example highlights that complex networks can arise through enzymatic activities in the pathways beyond TPSs and P450s. Specifically, the biosynthesis of nepetalactones involves dephosphorylation of geranyl diphosphate by geraniol synthase, hydroxylation by geraniol-8-hydroxylase and oxidation to the ketone 8-oxogeranial. This linear route was then suggested to be expanded by action of iridoid synthase yielding nepetalactol a decade ago [196]. The pathway was refined by the demonstration that short-chain reductases/dehydrogenases are required for stereoselective formation of nepetalactone in either *cis-trans* or *cis-cis* configuration by nepetalactone synthase 1 (NEPS1) and *cis-cis-*

nepetalactol (NEPS3) in *Nepeta* spp. (catmint) [197]. Intriguingly, the genomes of two *Nepeta* species revealed the presence of BGCs which, together with co-expression analysis finally clarified the routes to three of the four possible stereoisomers of nepetalactone. The authors reported specific sequences including four NEPS and a major latex protein-like (MLPL) enzyme spanning the metabolic network (Fig. 1.7, adapted from Lichman *et al.*, 2020) [194].

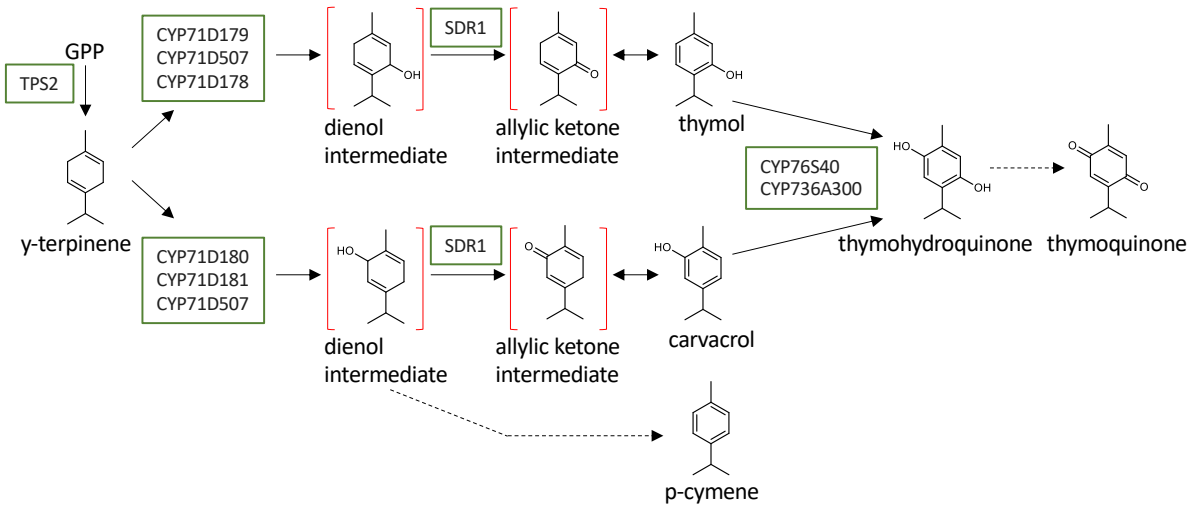


Figure. 1.7 Biosynthetic pathways of thymol, carvacrol, p-cymene, and thymohydroquinone. Pathways are from thyme and oregano. SDR, short-chain dehydrogenase/reductase.

Like thyme, *M. longifolia* accessions were reported to have highly diverse chemotypes, which can open the door to the discovery of enzymes contributing to characteristic profiles. Even though an earlier draft genome of *M. longifolia* existed [198], the chromosome scale reference genome published in 2022 impressively set the stage to identify and characterize the genetic underpinning for the monoterpene composition[195]. Integration of transcriptomic data, specifically from the glandular trichomes, and putative orthology to the large number of

functionally characterized homologs in other mint species (i.e., peppermint) involved in the pathway to (-)-mentone was used to mine the genome.

This approach identified some, but not all, menthone biosynthetic candidate genes including those encoding the key TPS for limonene (LS), the P450 catalyzing 3-hydroxylation of limonene (L3H), isopiperitenone dehydrogenase (ISPD), isopiperitenone reductase (ISPR) and pulegone reductase (PulR). The latter three were experimentally validated as recombinant enzymes. It is unclear, however, why the chemotype of the sequenced accession accumulates predominantly pulegone and only traces of mentone (79% and 0.06%, respectively), despite high expression and apparent functionality of the PulR. Yet, with detection of metabolites of the non-mentone, piperitenone-type, which originate from limonene as well, mint monoterpenes follow suit to the examples above and are formed through complex metabolic networks. It is noteworthy that the genomic mapping of the candidate genes indicated presence of at least one cluster of non-homologous genes of monoterpene metabolism (LS, L3H) [195].

In *Brassicaceae, A. thaliana* is a good example of a species with an active monoterpenoid network, specifically linalool metabolism. This has been extensively reviewed previously and therefore only a few enzymes are highlighted here. Two monoTPSs, TPS10 and TPS14, were characterized as (*R*)- and (*S*)-linalool synthases, respectively [65]. *A. thaliana* harbors several P450s from both the CYP71 and CYP76 families that can hydroxylate linalool. These include CYP76C3 and CYP71B31 that produce multiple linalool derivatives from (3S) and (3R) linalool, however not in equivalent quantities [65]. Additional CYP76s were found to also participate in linalool oxidation with CYP76C1 recognized to be the enzyme mainly responsible for linalool oxidation in flowers[199,200].

Grapevine (*Vitis vinifera*) from the Vitaceae family is another producer of linalool with several TPSs responsible for this [201]. Recently CYP76F14 was found to oxidize linalool in several steps to (E)-8-carboxylinalool [202]. (E)-8-carboxylinalool is present as glucose ester in grape berries, with expected involvement of a UGT in the biosynthesis. Of further importance to the wine industry is that linalool is a precursor of the wine odorant wine lactone.

Pyrethrins are an important and unique type of metabolites that are valued for their anti-insecticidal activity and found in several species in Asteraceae including *Tanacetum cinerariifolium*. The various pyrethrin structures consist of two parts derived from two different pathways, an acid moiety from an irregular monoterpenoid pathway and an alcohol moiety derived from jasmonic acid. The monoterpenoid acid moiety is either pyrethric acid or chrysanthemic acid and several different metabolites of the class of pyrethrin have been described [203]. The pathway to pyrethric acid was recently discovered and includes four enzymes. The initial enzyme, TcCDS, involved in the monoterpenoid branch was discovered in 2001 and found to utilize two DMAPP to produce chrysanthemyl diphosphate [21]. Later additional studies showed that TcCDS could also form chrysanthemol, which is the actual precursor of pyrethrins[204]. The following steps up until pyrethric acid was only discovered recently [205]. Chrysanthemol is converted to chrysanthemic acid by alcohol dehydrogenase (2 TcADH2), and aldehyde dehydrogenase (1 TcALDH1) [206]. CYP71BZ is a chrysanthemol 10-hydroxylase (TcCHH) that oxidizes C10 of chrysanthemol to form a carboxylic group thereby producing 10-carboxychrysanthemic acid. The final step to pyrethric acid is performed by 10-carboxychrysanthemic acid 10-methyltransferase (TcCCMT). A GDSL lipase (TcGLIP) is responsible for linking the acid moiety to the alcohol moiety [207].

Mono- and sesquiterpenoids – blurred lines

One of the most promiscuous plant P450s discovered so far, CYP706A3, has been found in *A. thaliana* [208]. CYP706A3 is a key player in the sesquiterpenoid metabolic network but can also oxidize monoterpenoids. CYP706A3 was found to cluster with TPS11 on the Arabidopsis genome. TPS11 is a sesquiterpene multiproduct enzyme, with (+)-a-barbatene and (+)-thujopsene amongst the major products [209]. The substrates of CYP706A3 are hydrocarbons and the hydroxylation performed on these is a key factor in switching the affected classes from volatiles released to retained and stored in flowers plausibly due to increased polarity. Interestingly insect larvae were observed to avoid feeding on flowers expressing CYP706A3. As seen before with promiscuous P450s, CYP706A3 did not oxidize all TPS11 products with the same efficiency but seemed to favor the 'bulky multi-cyclic sesquiterpenoids'. The promiscuity of CYP706A3 was further supported by not only single oxidations on multiple substrates, but also successive oxidations of the same substrates. Despite the clustering with a sesquiterpene producing TPS, additional screening in microsome assays with pure standards of several monoterpenoids, α-pinene, (+)-sabinene, β-pinene, and α-phellandrene, corresponding to the product of *A. thaliana* TPS24, also showed oxidation of these.

Conclusion

The evidence for complex networks in the pathways presented here demonstrates how plants often rely on a small set of enzymes to generate distinct, yet variegated, chemical profiles. Similar strategies are applicable beyond terpenoids and are often observed in other specialized metabolite families. In particular, the reliance on promiscuous P450s and other modifying and conjugating enzymes is common to the biosynthesis of alkaloids, flavonoids and glucosinolates.

In our survey of terpenoid biosynthesis, it is apparent that within a particular plant lineage, a single or few terpene backbones can give rise to the majority of terpenoid diversity. Within the diterpenoids, most of this diversity is generated through the activity of modifying enzymes such as P450s and other oxidoreductases as well as glycosyl and acyl transferases. In particular, the tendency of these enzymes to act on multiple substrates and thus create multiple network edges exponentially diversifies the product profile. Some multiproduct diTPSs exist, but they tend to be the exception, whereas volatile mono- and sesquiterpene pathways appear to rely largely on diversification through TPSs (both by multi-product single enzymes and large gene families), consistent with biological functions based on volatility. In the cases where mono- and sesquiterpenes are heavily modified, they reflect diterpenoid pathways in that one backbone is used to generate a class of related compounds.

The human bias in approaching specialized metabolism is to seek a linear pathway leading to single compounds of interest. Though this is the case in limited plant species where a single or few compounds predominate, there are typically many related compounds present in various quantities. Moreover, “intermediate” compounds of some pathways can accumulate to detectable levels and may also be bioactive. The terpenoid pathways presented here support the concept that plants benefit from a network of specialized metabolism that generates a diverse array of bioactive compounds, enabling defense against a variety of environmental threats. Those studied now are simply a snapshot of a dynamic evolutionary process that changes along with constantly shifting selective pressures. Moreover, although plants rely on a repertoire of both redundant and promiscuous enzymes for these networks, they have also developed extensive regulatory mechanisms to limit interactions and control product outcomes. Integrating study of

the native host metabolism through precise transcriptomics, proteomics, and metabolomics approaches along with biochemical enzyme characterization will improve our understanding of how plants control such complex network interactions. As we continue to learn how plants accomplish sophisticated biosynthesis of specialized metabolites, this knowledge will also serve as a lesson for biotechnological control of plant enzymes when engineering production of a single desired compound.

Acknowledgements

This work was supported by the Michigan State University Strategic Partnership Grant program ('Evolutionary-Driven Genome Mining of Plant Biosynthetic Pathways') to BH through Georgia Research Alliance funds to C. Robin Buell. BH gratefully acknowledges the US Department of Energy Great Lakes Bioenergy Research Center Cooperative Agreement DE-SC0018409, startup funding from the Department of Biochemistry and Molecular Biology, and support from AgBioResearch (MICL02454). EL is supported by the NSF Graduate Research Fellowship Program (DGE-1848739). BH is in part supported by the National Science Foundation under Grant Number 1737898. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Michigan State University occupies the ancestral, traditional, and contemporary Lands of the Anishinaabeg–Three Fires Confederacy of Ojibwe, Odawa, and Potawatomi peoples. The University resides on Land ceded in the 1819 Treaty of Saginaw.

Thesis Rationale and Overview

From both a fundamental and applied science perspective, mapping plant terpene biosynthetic pathways is an important part of understanding plant specialized metabolism. Due to the often

lineage-specific nature of specialized metabolism, different species generate unique chemistry to answer the demands of their native environment. Through investigation of non-model species, we can learn more about the evolution and genetic underpinnings of specialized metabolism while discovering the enzymes needed for biotechnological production of valuable fragrance, flavor, medicinal, and agricultural compounds which are difficult to synthesize by traditional methods. In this work, I explore the diterpenoid biosynthetic networks of American beautyberry (*Callicarpa americana)* within the wider context of orthologous pathways in other members of the mint (Lamiaceae) family. The mint family is a rich source of terpenes, especially bioactive diterpenes [73]. The genus *Callicarpa* contains over 150 recognized species [210] and has numerous documented uses in traditional medicines across Asian and native American cultures [70,211]. Nearly 1000 papers have been published in the past two decades documenting the phytochemical components and bioactivities of *Callicarpa* species, highlighting the presence of nearly 100 unique diterpenoid structures[212]. However, excepting one recent paper on iridoids (monoterpenoids)[213], there have been no reports on the biosynthetic pathways of specialized metabolites in *Callicarpa*. In Chapter 2, I characterize the set of diTPS genes in *C. americana*, using genomic and transcriptomic data developed by our collaborators in the Buell lab. This leads to work in Chapter 3, where I investigate along with my coauthor Abby Bryson the possible function of a large BGC of diterpenoid related genes. We also find syntenic BGCs present in the genomes of at least 7 other divergent mint family species. This work highlights the role of genomic organization in the evolution of diterpenoid metabolism, which has only recently been recognized as an important mechanism in plants. Moreover, I find that the diTPSs in the BGC are instrumental in making both a unique diterpene backbone without a previously reported

biosynthetic route as well as a class of diterpenes found in the roots of *C. americana* that has never been reported in this species. In Chapter 4, I discover a P450 that is key to the biosynthesis of bioactive furano-clerodane type diterpenes in *C. americana*. Along with coauthor Nick Schlecht, I investigate this subfamily of P450s in 7 additional mint family species, finding a shared enzymatic function that is key to numerous biosynthetic pathways for important insect antifeedant diterpenes. In Chapter 5, I continue to investigate the enzymes involved in the biosynthesis of bioactive clerodanes. I generate transcriptomic resources for two additional *Callicarpa* species, Japanese beautyberry (*C. japonica*) and Mexican beautyberry (*C. acuminata*), as well as for glandular trichome tissue in *C. americana.* After testing over 40 candidate enzymes, I find three active short chain dehydrogenases (SDRs) that catalyze formation of another intermediate step towards the desired bioactive clerodane compounds. Together, the work presented here contributes to a deep understanding of the diterpenoid networks present and how they came to be in American beautyberry and more widely, the mint family. These findings will help to enable biotechnological production of valuable diterpenoids while also providing an entry point for further biosynthetic discoveries in other *Callicarpa* and mint family species.

REFERENCES

1. Weng, J.-K., Lynch, J. H., Matos, J. O. & Dudareva, N. Adaptive mechanisms of plant specialized metabolism connecting chemistry to function. *Nat. Chem. Biol.* **17**, 1037–1045 (2021).

2. Hartmann, T. The lost origin of chemical ecology in the late 19th century. *Proc. Natl. Acad. Sci.* **105**, 4541–4546 (2008).

3. Erb, M. & Kliebenstein, D. J. Plant Secondary Metabolites as Defenses, Regulators, and Primary Metabolites: The Blurred Functional Trichotomy. *Plant Physiol.* **184**, 39–52 (2020).

4. Leroi-Gourhan, A. The Flowers Found with Shanidar IV, a Neanderthal Burial in Iraq. *Science* **190**, 562–564 (1975).

5. Patridge, E., Gareiss, P., Kinch, M. S. & Hoyer, D. An analysis of FDA-approved drugs: natural products and their derivatives. *Drug Discov. Today* **21**, 204–207 (2016).

6. Li, F.-S. & Weng, J.-K. Demystifying traditional herbal medicine with modern approach. *Nat. Plants* **3**, 1–7 (2017).

7. Zhao, Q. *et al.* Synergistic Mechanisms of Constituents in Herbal Extracts during Intestinal Absorption: Focus on Natural Occurring Nanoparticles. *Pharmaceutics* **12**, 128 (2020).

8. Salazar, D. *et al.* Origin and maintenance of chemical diversity in a species-rich tropical tree lineage. *Nat. Ecol. Evol.* **2**, 983–990 (2018).

9. Volf, M. *et al.* Community structure of insect herbivores is driven by conservatism, escalation and divergence of defensive traits in Ficus. *Ecol. Lett.* **21**, 83–92 (2018).

10. Whitehead, S. R., Bass, E., Corrigan, A., Kessler, A. & Poveda, K. Interaction diversity explains the maintenance of phytochemical diversity. *Ecol. Lett.* **24**, 1205–1214 (2021).

11. Hamberger, B., Ehlting, J., Barbazuk, B. & Douglas, C. J. Chapter Four - Comparative Genomics of The Shikimate Pathway in *Arabidopsis*, *Populus Trichocarpa* and *Oryza Sativa*: Shikimate Pathway Gene Family Structure and Identification of Candidates for Missing Links in Phenylalanine Biosynthesis. in *Recent Advances in Phytochemistry* (ed. Romeo, J. T.) vol. 40 85–113 (Elsevier, 2006).

12. Hamberger, B. & Hahlbrock, K. The 4-coumarate:CoA ligase gene family in Arabidopsis thaliana comprises one rare, sinapate-activating and three commonly occurring isoenzymes. *Proc. Natl. Acad. Sci.* **101**, 2209–2214 (2004).

13. Leong, B. J. & Last, R. L. Promiscuity, impersonation and accommodation: evolution of plant specialized metabolism. *Curr. Opin. Struct. Biol.* **47**, 105–112 (2017).

14. Karunanithi, P. S. & Zerbe, P. Terpene Synthases as Metabolic Gatekeepers in the Evolution of Plant Terpenoid Chemical Diversity. *Front. Plant Sci.* **10**, 1166 (2019).

15. Jia, Q. *et al.* Origin and early evolution of the plant terpene synthase family. *Proc. Natl. Acad. Sci.* **119**, e2100361119 (2022).

16. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229 (2011).

17. Zhou, F. & Pichersky, E. More is better: the diversity of terpene metabolism in plants. *Curr. Opin. Plant Biol.* **55**, 1–10 (2020).

18. Zi, J., Mafu, S. & Peters, R. J. To Gibberellins and Beyond! Surveying the Evolution of (Di)Terpenoid Metabolism. *Annu. Rev. Plant Biol.* **65**, 259–286 (2014).

19. Weng, J.-K., Philippe, R. N. & Noel, J. P. The Rise of Chemodiversity in Plants. *Science* **336**, 1667–1670 (2012).

20. Yoshida, R., Yoshimura, T. & Hemmi, H. Reconstruction of the “Archaeal” Mevalonate Pathway from the Methanogenic Archaeon *Methanosarcina mazei* in *Escherichia coli* Cells. *Appl. Environ. Microbiol.* **86**, e02889-19 (2020).

21. Rivera, S. B. *et al.* Chrysanthemyl diphosphate synthase: Isolation of the gene and characterization of the recombinant non-head-to-tail monoterpene synthase from *Chrysanthemum cinerariaefolium*. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4373–4378 (2001).

22. Demissie, Z. A., Erland, L. A. E., Rheault, M. R. & Mahmoud, S. S. The Biosynthetic Origin of Irregular Monoterpenes in *Lavandula*. *J. Biol. Chem.* **288**, 6333–6341 (2013).

23. Akhtar, T. A. *et al.* The tomato cis–prenyltransferase gene family. *Plant J.* **73**, 640–652 (2013).

24. Miller, G. P. *et al.* The biosynthesis of the anti-microbial diterpenoid leubethanol in *Leucophyllum frutescens* proceeds via an all-cis prenyl intermediate. *Plant J.* **104**, 693–705 (2020).

25. J. Tantillo, D. Biosynthesis via carbocations: Theoretical studies on terpene formation. *Nat. Prod. Rep.* **28**, 1035–1053 (2011).

26. Xu, M., Wilderman, P. R. & Peters, R. Following evolution's lead to a single residue switch for diterpene synthase product outcome. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7397–401 (2007).

27. Wilderman, P. R. & Peters, R. J. A Single Residue Switch Converts Abietadiene Synthase into a Pimaradiene Specific Cyclase. *J. Am. Chem. Soc.* **129**, 15736–15737 (2007).

28. Jia, M. & Peters, R. J. Extending a Single Residue Switch for Abbreviating Catalysis in Plant *ent*-Kaurene Synthases. *Front. Plant Sci.* **7**, (2016).

29. Xu, J. *et al.* Converting S-limonene synthase to pinene or phellandrene synthases reveals the plasticity of the active site. *Phytochemistry* **137**, 34–41 (2017).

30. Irmisch, S. *et al.* One amino acid makes the difference: the formation of *ent*-kaurene and 16α-hydroxy-ent-kaurane by diterpene synthases in poplar. *BMC Plant Biol.* **15**, 262 (2015).

31. Köllner, T. G., Degenhardt, J. & Gershenzon, J. The Product Specificities of Maize Terpene Synthases TPS4 and TPS10 Are Determined Both by Active Site Amino Acids and Residues Adjacent to the Active Site. *Plants* **9**, 552 (2020).

32. Steele, C. L., Crock, J., Bohlmann, J. & Croteau, R. Sesquiterpene synthases from grand fir (*Abies grandis*). Comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of delta-selinene synthase and gamma-humulene synthase. *J. Biol. Chem.* **273**, 2078–2089 (1998).

33. Wang, S., Alseekh, S., Fernie, A. R. & Luo, J. The Structure and Function of Major Plant Metabolite Modifications. *Mol. Plant* **12**, 899–919 (2019).

34. Hamberger, B. & Bak, S. Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120426 (2013).

35. Bathe, U. & Tissier, A. Cytochrome P450 enzymes: A driving force of plant diterpene diversity. *Phytochemistry* **161**, 149–162 (2019).

36. Banerjee, A. & Hamberger, B. P450s controlling metabolic bifurcations in plant terpene specialized metabolism. *Phytochem. Rev.* **17**, 81–111 (2018).

37. James, W. H. & Hollinger, M. E. The utilization of carotene. II. From sweet potatoes by young human adults. *J. Nutr.* **54**, 65–74 (1954).

38. Kramer, M. & Tarjan, R. Studies on carotene metabolism. IV. The biological value of carotene in various plants. *Int. Z. Vitaminforschung Int. J. Vitam. Res. J. Int. Vitaminol.* **30**, 49–59 (1959).

39. Modi, V. V. & Patwa, D. K. Biosynthesis of Carotenes in Carrot Extracts. *Nature* **184**, 983–984 (1959).

40. Battaile, J., Dunning, R. L. & Loomis, W. D. Biosynthesis of terpenes. I. Chromatography of peppermint oil terpenes. *Biochim. Biophys. Acta* **51**, 538–544 (1961).

41. Crowley, M. P., Godin, P. J., Inglis, H. S., Snarey, M. & Thain, E. M. The biosynthesis of the 'pyrethrins'. I. The incorporation of 14C-labelled compounds into the flowers of *Chrysanthemum cinerariaefolium* and the biosynthesis of chrysanthemum monocarboxylic acid. *Biochim. Biophys. Acta* **60**, 312–319 (1962).

42. Dennis, D. T. & West, C. A. Biosynthesis of gibberellins. 3. The conversion of (-)-kaurene to (-)-kauren-19-oic acid in endosperm of *Echinocystis macrocarpa* Greene. *J. Biol. Chem.* **242**, 3293–3300 (1967).

43. Upper, C. D. & West, C. A. Biosynthesis of gibberellins. II. Enzymic cyclization of geranylgeranyl pyrophosphate to kaurene. *J. Biol. Chem.* **242**, 3285–3292 (1967).

44. Croteau, R., Burbott, A. J. & Loomis, W. D. Enzymatic cyclization of neryl pyrophosphate to -terpineol by cell-free extracts from peppermint. *Biochem. Biophys. Res. Commun.* **50**, 1006–1012 (1973).

45. Croteau, R. & Karp, F. Biosynthesis of monoterpenes: partial purification and characterization of 1,8-cineole synthetase from Salvia officinalis. *Arch. Biochem. Biophys.* **179**, 257–265 (1977).

46. Poulose, A. J. & Croteau, R. Biosynthesis of aromatic monoterpenes: conversion of gamma-terpinene to p-cymene and thymol in *Thymus vulgaris* L. *Arch. Biochem. Biophys.* **187**, 307–314 (1978).

47. Kjonaas, R. & Croteau, R. Demonstration that limonene is the first cyclic intermediate in the biosynthesis of oxygenated p-menthane monoterpenes in *Mentha piperita* and other Mentha species. *Arch. Biochem. Biophys.* **220**, 79–89 (1983).

48. Kjonaas, R. B., Venkatachalam, K. V. & Croteau, R. Metabolism of monoterpenes: Oxidation of isopiperitenol to isopiperitenone, and subsequent isomerization to piperitenone by soluble enzyme preparations from peppermint (*Mentha piperita*) leaves. *Arch. Biochem. Biophys.* **238**, 49–60 (1985).

49. Karp, F., Harris, J. L. & Croteau, R. Metabolism of monoterpenes: demonstration of the hydroxylation of (+)-sabinene to (+)-*cis*-sabinol by an enzyme preparation from sage (*Salvia officinalis*) leaves. *Arch. Biochem. Biophys.* **256**, 179–193 (1987).

50. Wildung, M. R. & Croteau, R. A cDNA clone for taxadiene synthase, the diterpene cyclase that catalyzes the committed step of taxol biosynthesis. *J. Biol. Chem.* **271**, 9201–9204 (1996).

51. Hefner, J. *et al.* Cytochrome P450-catalyzed hydroxylation of taxa-4(5),11(12)-diene to taxa-4(20),11(12)-dien-5alpha-ol: the first oxygenation step in taxol biosynthesis. *Chem. Biol.* **3**, 479–489 (1996).

52. McCaskill, D. & Croteau, R. Prospects for the bioengineering of isoprenoid biosynthesis. *Adv. Biochem. Eng. Biotechnol.* **55**, 107–146 (1997).

53. Lange, B. M. & Croteau, R. Genetic engineering of essential oil production in mint. *Curr. Opin. Plant Biol.* **2**, 139–144 (1999).

54. Mahmoud, S. S. & Croteau, R. B. Metabolic engineering of essential oil yield and composition in mint by altering expression of deoxyxylulose phosphate reductoisomerase and menthofuran synthase. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 8915–8920 (2001).

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

56. Aubourg, S., Takvorian, A., Chéron, A., Kreis, M. & Lecharny, A. Structure, organization and putative function of the genes identified within a 23.9-kb fragment from *Arabidopsis thaliana chromosome* IV. *Gene* **199**, 241–253 (1997).

57. Bohlmann, J., Martin, D., Oldham, N. J. & Gershenzon, J. Terpenoid secondary metabolism in Arabidopsis thaliana: cDNA cloning, characterization, and functional expression of a myrcene/(E)-beta-ocimene synthase. *Arch. Biochem. Biophys.* **375**, 261–269 (2000).

58. Trapp, S. C. & Croteau, R. B. Genomic Organization of Plant Terpene Synthases and Molecular Evolutionary Implications. *Genetics* **158**, 811–832 (2001).

59. Bohlmann, J., Meyer-Gauen, G. & Croteau, R. Plant terpenoid synthases: Molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci.* **95**, 4126–4133 (1998).

60. Aubourg, S., Lecharny, A. & Bohlmann, J. Genomic analysis of the terpenoid synthase (AtTPS) gene family of *Arabidopsis thaliana*. *Mol. Genet. Genomics* **267**, 730–745 (2002).

61. Chen, F. *et al.* Biosynthesis and emission of terpenoid volatiles from *Arabidopsis* flowers. *Plant Cell* **15**, 481–494 (2003).

62. Aharoni, A. *et al.* Terpenoid metabolism in wild-type and transgenic *Arabidopsis* plants. *Plant Cell* **15**, 2866–2884 (2003).

63. Ehlting, J. *et al.* An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in *Arabidopsis thaliana*. *BMC Plant Biol.* **8**, 47 (2008).

64. Field, B. & Osbourn, A. E. Metabolic diversification--independent assembly of operon-like gene clusters in different plants. *Science* **320**, 543–547 (2008).

65. Ginglinger, J.-F. *et al.* Gene Coexpression Analysis Reveals Complex Metabolism of the Monoterpene Alcohol Linalool in *Arabidopsis* Flowers. *Plant Cell* **25**, 4640–4657 (2013).

66. Shimura, K. *et al.* Identification of a biosynthetic gene cluster in rice for momilactones. *J. Biol. Chem.* **282**, 34013–34018 (2007).

67. Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B. & Peters, R. J. CYP76M7 is an *ent*-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell* **21**, 3315–3325 (2009).

68. Wisecaver, J. H. *et al.* A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **29**, 944–959 (2017).

69. Gericke, O. *et al.* Nerylneryl diphosphate is the precursor of serrulatane, viscidane and cembrane-type diterpenoids in *Eremophila* species. *BMC Plant Biol.* **20**, 91 (2020).

70. Matsuba, Y., Zi, J., Jones, A. D., Peters, R. J. & Pichersky, E. Biosynthesis of the Diterpenoid Lycosantalonol via Nerylneryl Diphosphate in *Solanum lycopersicum*. *PLOS ONE* **10**, e0119302 (2015).

71. Peters, R. J. Two rings in them all: the labdane-related diterpenoids. *Nat. Prod. Rep.* **27**, 1521–1530 (2010).

72. Wang, Z., Nelson, D. R., Zhang, J., Wan, X. & Peters, R. J. Plant (di)terpenoid evolution: from pigments to hormones and beyond. *Nat. Prod. Rep.* (2022) doi:10.1039/D2NP00054G.

73. Johnson, S. R. *et al.* A database-driven approach identifies additional diterpene synthase activities in the mint family (Lamiaceae). *J. Biol. Chem.* **294**, 1349–1362 (2019).

74. Andersen-Ranberg, J. *et al.* Expanding the Landscape of Diterpene Structural Diversity through Stereochemically Controlled Combinatorial Biosynthesis. *Angew. Chem. Int. Ed.* **55**, 2142–2146 (2016).

75. Jia, M., Mishra, S. K., Tufts, S., Jernigan, R. L. & Peters, R. J. Combinatorial biosynthesis and the basis for substrate promiscuity in class I diterpene synthases. *Metab. Eng.* **55**, 44–58 (2019).

76. González, M. A. Aromatic abietane diterpenoids: their biological activity and synthesis. *Nat. Prod. Rep.* **32**, 684–704 (2015).

77. Wang, Z. & Peters, R. J. Tanshinones: Leading the way into Lamiaceae labdane-related diterpenoid biosynthesis. *Curr. Opin. Plant Biol.* **66**, 102189 (2022).

78. Zi, J. & Peters, R. J. Characterization of CYP76AH4 clarifies phenolic diterpenoid biosynthesis in the Lamiaceae. *Org. Biomol. Chem.* **11**, 7650–7652 (2013).

79. Guo, J. *et al.* Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. *New Phytol.* **210**, 525–534 (2016).

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

81. Song, J.-J. *et al.* A 2-oxoglutarate-dependent dioxygenase converts dihydrofuran to furan in *Salvia* diterpenoids. *Plant Physiol.* **188**, 1496–1506 (2022).

82. Pang, H. *et al.* Chemical Analysis of the Herbal Medicine *Salviae miltiorrhizae* Radix et Rhizoma (Danshen). *Molecules* **21**, 51 (2016).

83. Song, Z. *et al.* A high-quality reference genome sequence of *Salvia miltiorrhiza* provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome* **13**, e20041 (2020).

84. Birtić, S., Dussort, P., Pierre, F.-X., Bily, A. C. & Roller, M. Carnosic acid. *Phytochemistry* **115**, 9–19 (2015).

85. Loussouarn, M. *et al.* Carnosic Acid and Carnosol, Two Major Antioxidants of Rosemary, Act through Different Mechanisms. *Plant Physiol.* **175**, 1381–1394 (2017).

86. Ignea, C. *et al.* Carnosic acid biosynthesis elucidated by a synthetic biology platform. *Proc. Natl. Acad. Sci.* **113**, 3681–3686 (2016).

87. Scheler, U. *et al.* Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast. *Nat. Commun.* **7**, 12942 (2016).

88. Vogel, B. S., Wildung, M. R., Vogel, G. & Croteau, R. Abietadiene synthase from grand fir (*Abies grandis*). cDNA isolation, characterization, and bacterial expression of a bifunctional diterpene cyclase involved in resin acid biosynthesis. *J. Biol. Chem.* **271**, 23262–23268 (1996).

89. Zhou, K. *et al.* Insights into Diterpene Cyclization from Structure of Bifunctional Abietadiene Synthase from *Abies grandis*. *J. Biol. Chem.* **287**, 6840–6850 (2012).

90. Keeling, C. I., Madilao, L. L., Zerbe, P., Dullat, H. K. & Bohlmann, J. The Primary Diterpene Synthase Products of Picea abies Levopimaradiene/Abietadiene Synthase (PaLAS) Are Epimers of a Thermally Unstable Diterpenol. *J. Biol. Chem.* **286**, 21145–21153 (2011).

91. Ro, D.-K., Arimura, G.-I., Lau, S. Y. W., Piers, E. & Bohlmann, J. Loblolly pine abietadienol/abietadienal oxidase PtAO (CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase. *Proc. Natl. Acad. Sci.* **102**, 8060–8065 (2005).

92. Hamberger, B., Ohnishi, T., Hamberger, B., Séguin, A. & Bohlmann, J. Evolution of diterpene metabolism: Sitka spruce CYP720B4 catalyzes multiple oxidations in resin acid biosynthesis of conifer defense against insects. *Plant Physiol.* **157**, 1677–1695 (2011).

93. Geisler, K., Jensen, N. B., Yuen, M. M. S., Madilao, L. & Bohlmann, J. Modularity of Conifer Diterpene Resin Acid Biosynthesis: P450 Enzymes of Different CYP720B Clades Use Alternative Substrates and Converge on the Same Products. *Plant Physiol.* **171**, 152–164 (2016).

94. Pateraki, I. *et al.* Manoyl oxide (13R), the biosynthetic precursor of forskolin, is synthesized in specialized root cork cells in *Coleus forskohlii*. *Plant Physiol.* **164**, 1222–1236 (2014).

95. Pateraki, I. *et al.* Total biosynthesis of the cyclic AMP booster forskolin from *Coleus forskohlii*. *eLife* **6**, e23001 (2017).

96. Kutney, J. P. *et al.* Cytotoxic diterpenes triptolide, tripdiolide, and cytotoxic triterpenes from tissue cultures of *Tripterygium wilfordii*. *Can. J. Chem.* **59**, 2677–2683 (1981).

97. Inabuy, F. S. *et al.* Biosynthesis of Diterpenoids in Tripterygium Adventitious Root Cultures. *Plant Physiol.* **175**, 92–103 (2017).

98. Kupchan, S. M., Court, W. A., Dailey, R. G., Gilmore, C. J. & Bryan, R. F. Triptolide and tripdiolide, novel antileukemic diterpenoid triepoxides from *Tripterygium wilfordii*. *J. Am.Chem. Soc.* **94**, 7194–7195 (1972).

99. Hansen, N. L. *et al.* The terpene synthase gene family in *Tripterygium wilfordii* harbors a labdane-type diterpene synthase among the monoterpene synthase TPS-b subfamily. *Plant J. Cell Mol. Biol.* **89**, 429–441 (2017).

100. Forman, V., Callari, R., Folly, C., Heider, H. & Hamberger, B. Production of Putative Diterpene Carboxylic Acid Intermediates of Triptolide in Yeast. *Mol. J. Synth. Chem. Nat. Prod. Chem.* **22**, 981 (2017).

101. Tu, L. *et al.* Genome of Tripterygium wilfordii and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat. Commun.* **11**, 971 (2020).

102. Hansen, N. L. *et al. Tripterygium wilfordii* cytochrome P450s catalyze the methyl shift and epoxidations in the biosynthesis of triptonide. *Nat. Commun.* **13**, 5011 (2022).

103. Johnson, S. R. *et al.* Promiscuous terpene synthases from *Prunella vulgaris* highlight the importance of substrate and compartment switching in terpene synthase evolution. *New Phytol.* **223**, 323–335 (2019).

104. Kemboi, D., Siwe-Noundou, X., Krause, R. W. M., Langat, M. K. & Tembu, V. J. Euphorbia Diterpenes: An Update of Isolation, Structure, Pharmacological Activities and Structure–Activity Relationship. *Molecules* **26**, 5055 (2021).

105. Mau, C. J. & West, C. A. Cloning of casbene synthase cDNA: evidence for conserved structural features among terpenoid cyclases in plants. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 8497–8501 (1994).

106. Kirby, J. *et al.* Cloning of casbene and neocembrene synthases from Euphorbiaceae plants and expression in Saccharomyces cerevisiae. *Phytochemistry* **71**, 1466–1473 (2010).

107. Zerbe, P. *et al.* Gene Discovery of Modular Diterpene Metabolism in Nonmodel Systems. *Plant Physiol.* **162**, 1073–1091 (2013).

108. King, A. J., Brown, G. D., Gilday, A. D., Larson, T. R. & Graham, I. A. Production of Bioactive Diterpenoids in the Euphorbiaceae Depends on Evolutionarily Conserved Gene Clusters. *Plant Cell* **26**, 3286–3298 (2014).

109. Boutanaev, A. M. *et al.* Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci.* **112**, E81–E88 (2015).

110. King, A. J. *et al.* A Cytochrome P450-Mediated Intramolecular Carbon–Carbon Ring Closure in the Biosynthesis of Multidrug-Resistance-Reversing Lathyrane Diterpenoids. *Chembiochem* **17**, 1593–1597 (2016).

111. Luo, D. *et al.* Oxidation and cyclization of casbene in the biosynthesis of Euphorbia factors from mature seeds of *Euphorbia lathyris* L. *Proc. Natl. Acad. Sci.* **113**, E5082–E5089 (2016).

112. Nguyen, T.-D., MacNevin, G. & Ro, D.-K. De novo synthesis of high-value plant sesquiterpenoids in yeast. *Methods Enzymol.* **517**, 261–278 (2012).

113. Wong, J. *et al.* High-titer production of lathyrane diterpenoids from sugar by engineered *Saccharomyces cerevisiae*. *Metab. Eng.* **45**, 142–148 (2018).

114. Fattahian, M., Ghanadian, M., Ali, Z. & Khan, I. A. Jatrophane and rearranged jatrophane-type diterpenes: biogenesis, structure, isolation, biological activity and SARs (1984–2019). *Phytochem. Rev.* **19**, 265–336 (2020).

115. Esposito, M. *et al. Euphorbia dendroides* Latex as a Source of Jatrophane Esters: Isolation, Structural Analysis, Conformational Study, and Anti-CHIKV Activity. *J. Nat. Prod.* **79**, 2873–2882 (2016).

116. Ceunen, S. & Geuns, J. M. C. Steviol Glycosides: Chemical Diversity, Metabolism, and Function. *J. Nat. Prod.* **76**, 1201–1228 (2013).

117. Richman, A. S., Gijzen, M., Starratt, A. N., Yang, Z. & Brandle, J. E. Diterpene synthesis in *Stevia rebaudiana*: recruitment and up-regulation of key enzymes from the gibberellin biosynthetic pathway. *Plant J.* **19**, 411–421 (1999).

118. Humphrey, T. V., Richman, A. S., Menassa, R. & Brandle, J. E. Spatial Organisation of Four Enzymes from *Stevia rebaudiana* that are Involved in Steviol Glycoside Synthesis. *Plant Mol. Biol.* **61**, 47–62 (2006).

119. Xu, X. *et al.* The chromosome-level *Stevia* genome provides insights into steviol glycoside biosynthesis. *Hortic. Res.* **8**, 129 (2021).

120. Wang, J., Li, S., Xiong, Z. & Wang, Y. Pathway mining-based integration of critical enzyme parts for de novo biosynthesis of steviolglycosides sweetener in *Escherichia coli*. *Cell Res.* **26**, 258–261 (2016).

121. Xu, Y. *et al.* De novo biosynthesis of rubusoside and rebaudiosides in engineered yeasts. *Nat. Commun.* **13**, 3040 (2022).

122. Richman, A. *et al.* Functional genomics uncovers three glucosyltransferases involved in the synthesis of the major sweet glucosides of *Stevia rebaudiana*. *Plant J.* **41**, 56–67 (2005).

123. Olsson, K. *et al.* Microbial production of next-generation stevia sweeteners. *Microb. Cell Factories* **15**, 207 (2016).

124. Prakash, I., Markosyan, A. & Bunders, C. Development of Next Generation Stevia Sweetener: Rebaudioside M. *Foods* **3**, 162–175 (2014).

125. Muhlemann, J. K., Klempien, A. & Dudareva, N. Floral volatiles: from biosynthesis to function. *Plant Cell Environ.* **37**, 1936–1949 (2014).

126. Gershenzon, J. & Dudareva, N. The function of terpene natural products in the natural world. *Nat. Chem. Biol.* **3**, 408–414 (2007).

127. Robert, C. A. M. *et al.* Herbivore-induced plant volatiles mediate host selection by a root herbivore. *New Phytol.* **194**, 1061–1069 (2012).

128. Rasmann, S. *et al.* Recruitment of entomopathogenic nematodes by insect-damaged maize roots. *Nature* **434**, 732–737 (2005).

129. Theis, N., Barber, N. A., Gillespie, S. D., Hazzard, R. V. & Adler, L. S. Attracting mutualists and antagonists: Plant trait variation explains the distribution of specialist floral herbivores and pollinators on crops and wild gourds. *Am. J. Bot.* **101**, 1314–1322 (2014).

130. Junker, R. R., Gershenzon, J. & Unsicker, S. B. Floral Odor Bouquet Loses its Ant Repellent Properties After Inhibition of Terpene Biosynthesis. *J. Chem. Ecol.* **37**, 1323–1331 (2011).

131. Hammer, K. A., Carson, C. F. & Riley, T. V. Antifungal activity of the components of *Melaleuca alternifolia* (tea tree) oil. *J. Appl. Microbiol.* **95**, 853–860 (2003).

132. Huang, M. *et al.* The major volatile organic compound emitted from Arabidopsis thaliana flowers, the sesquiterpene (E)-β-caryophyllene, is a defense against a bacterial pathogen. *New Phytol.* **193**, 997–1008 (2012).

133. Peñuelas, J. *et al.* Removal of floral microbiota reduces floral terpene emissions. *Sci. Rep.* **4**, 6727 (2014).

134. Dudareva, N., Klempien, A., Muhlemann, J. K. & Kaplan, I. Biosynthesis, function and metabolic engineering of plant volatile organic compounds. *New Phytol.* **198**, 16–32 (2013).

135. Pateraki, I., Heskes, A. M. & Hamberger, B. Cytochromes P450 for terpene functionalisation and metabolic engineering. *Adv. Biochem. Eng. Biotechnol.* **148**, 107–139 (2015).

136. Tissier, A., Morgan, J. A. & Dudareva, N. Plant Volatiles: Going 'In' but not 'Out' of Trichome Cavities. *Trends Plant Sci.* **22**, 930–938 (2017).

137. Tissier, A. Plant secretory structures: more than just reaction bags. *Curr. Opin. Biotechnol.* **49**, 73–79 (2018).

138. Fahn, A. Secretory tissues in vascular plants. *New Phytol.* **108**, 229–257 (1988).

139. O'Maille, P. E., Chappell, J. & Noel, J. P. Biosynthetic potential of sesquiterpene synthases: Alternative products of tobacco 5-epi-aristolochene synthase. *Arch. Biochem. Biophys.* **448**, 73–82 (2006).

140. Jones, C. G. *et al.* Sandalwood Fragrance Biosynthesis Involves Sesquiterpene Synthases of Both the Terpene Synthase (TPS)-a and TPS-b Subfamilies, including Santalene Synthases. *J. Biol. Chem.* **286**, 17445–17454 (2011).

141. Sallaud, C. *et al.* A Novel Pathway for Sesquiterpene Biosynthesis from Z,Z-Farnesyl Pyrophosphate in the Wild Tomato Solanum habrochaites. *Plant Cell* **21**, 301–317 (2009).

142. Degenhardt, J., Köllner, T. G. & Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **70**, 1621–1637 (2009).

143. Pickel, B. *et al.* Identification and characterization of a kunzeaol synthase from *Thapsia garganica*: implications for the biosynthesis of the pharmaceutical thapsigargin. *Biochem. J.* **448**, 261–271 (2012).

144. Liang, J. *et al.* Probing Enzymatic Structure and Function in the Dihydroxylating Sesquiterpene Synthase ZmEDS. *Biochemistry* **59**, 2660–2666 (2020).

145. Xu, H. & Dickschat, J. S. Germacrene A–A Central Intermediate in Sesquiterpene Biosynthesis. *Chem. – Eur. J.* **26**, 17318–17341 (2020).

146. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).

147. Durairaj, J. *et al.* Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLOS Comput. Biol.* **17**, e1008197 (2021).

148. Ker, D.-S., Chan, K. G., Othman, R., Hassan, M. & Ng, C. L. Site-directed mutagenesis of β sesquiphellandrene synthase enhances enzyme promiscuity. *Phytochemistry* **173**, 112286 (2020).

149. Vattekkatte, A., Garms, S., Brandt, W. & Boland, W. Enhanced structural diversity in terpenoid biosynthesis: enzymes, substrates and cofactors. *Org. Biomol. Chem.* **16**, 348–362 (2018).

150. Nieuwenhuizen, N. J. *et al.* Two terpene synthases are responsible for the major sesquiterpenes emitted from the flowers of kiwifruit (*Actinidia deliciosa*). *J. Exp. Bot.* **60**, 3203–3219 (2009).

151. Drew, D. P. *et al.* Two key polymorphisms in a newly discovered allele of the *Vitis vinifera* TPS24 gene are responsible for the production of the rotundone precursor α-guaiene. *J. Exp. Bot.* **67**, 799–808 (2016).

152. Rising, K. A., Starks, C. M., Noel, J. P. & Chappell, J. Demonstration of Germacrene A as an Intermediate in 5-Epi-aristolochene Synthase Catalysis. *J. Am. Chem. Soc.* **122**, 1861–1866 (2000).

153. Yoshikuni, Y., Martin, V. J. J., Ferrin, T. E. & Keasling, J. D. Engineering cotton (+)-delta-cadinene synthase to an altered function: germacrene D-4-ol synthase. *Chem. Biol.* **13**, 91–98 (2006).

154. Bredahl, E. K. *et al.* Isolation and structure elucidation of caryophyllane sesquiterpenoids from leaves of *Eremophila spathulata*. *Phytochem. Lett.* **47**, 156–163 (2022).

155. Drew, D. P., Krichau, N., Reichwald, K. & Simonsen, H. T. Guaianolides in apiaceae: perspectives on pharmacology and biosynthesis. *Phytochem. Rev.* **8**, 581–599 (2009).

156. Tian, X. *et al.* Characterization of gossypol biosynthetic pathway. *Proc. Natl. Acad. Sci.* **115**, E5410–E5418 (2018).

157. Mao, H., Liu, J., Ren, F., Peters, R. J. & Wang, Q. Characterization of CYP71Z18 indicates a role in maize zealexin biosynthesis. *Phytochemistry* **121**, 4–10 (2016).

158. Zabel, S. *et al.* A single cytochrome P450 oxidase from *Solanum habrochaites* sequentially oxidizes 7-epi-zingiberene to derivatives toxic to whiteflies and various microorganisms. *Plant J.* **105**, 1309–1325 (2021).

159. Takahashi, S. *et al.* Kinetic and molecular analysis of 5-epiaristolochene 1,3-dihydroxylase, a cytochrome P450 enzyme catalyzing successive hydroxylations of sesquiterpenes. *J. Biol. Chem.* **280**, 3686–3696 (2005).

160. Facchini, P. J. & Chappell, J. Gene family for an elicitor-induced sesquiterpene cyclase in tobacco. *Proc. Natl. Acad. Sci.* **89**, 11088–11092 (1992).

161. Diaz-Chavez, M. L. *et al.* Biosynthesis of Sandalwood Oil: *Santalum album* CYP76F cytochromes P450 produce santalols and bergamotol. *PloS One* **8**, e75053 (2013).

162. Celedon, J. M. *et al.* Heartwood-specific transcriptome and metabolite signatures of tropical sandalwood (*Santalum album*) reveal the final step of (Z)-santalol fragrance biosynthesis. *Plant J. Cell Mol. Biol.* **86**, 289–299 (2016).

163. Chadwick, M., Trewin, H., Gawthrop, F. & Wagstaff, C. Sesquiterpenoids lactones: benefits to plants and people. *Int. J. Mol. Sci.* **14**, 12780–12805 (2013).

164. Mathema, V. B., Koh, Y.-S., Thakuri, B. C. & Sillanpää, M. Parthenolide, a Sesquiterpene Lactone, Expresses Multiple Anti-cancer and Anti-inflammatory Activities. *Inflammation* **35**, 560–565 (2012).

165. Thastrup, O., Cullen, P. J., Drøbak, B. K., Hanley, M. R. & Dawson, A. P. Thapsigargin, a tumor promoter, discharges intracellular Ca2+ stores by specific inhibition of the endoplasmic reticulum Ca2(+)-ATPase. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 2466–2470 (1990).

166. Tu, Y. The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nat. Med.* **17**, 1217–1220 (2011).

167. Andersen, T. B. *et al.* Localization and in-Vivo Characterization of *Thapsia garganica* CYP76AE2 Indicates a Role in Thapsigargin Biosynthesis. *Plant Physiol.* **174**, 56–72 (2017).

168. Liu, Q. *et al.* Elucidation and *in planta* reconstitution of the parthenolide biosynthetic pathway. *Metab. Eng.* **23**, 145–153 (2014).

169. Sy, L.-K. & Brown, G. D. The mechanism of the spontaneous autoxidation of dihydroartemisinic acid. *Tetrahedron* **58**, 897–908 (2002).

170. Teoh, K. H., Polichuk, D. R., Reed, D. W. & Covello, P. S. Molecular cloning of an aldehyde dehydrogenase implicated in artemisinin biosynthesis in *Artemisia*. *Botany* **87**, 635–642 (2009).

171. Majdi, M. *et al.* Biosynthesis and localization of parthenolide in glandular trichomes of feverfew (*Tanacetum parthenium* L. Schulz Bip.). *Phytochemistry* **72**, 1739–1750 (2011).

172. Liu, Q. *et al.* Kauniolide synthase is a P450 with unusual hydroxylation and cyclization-elimination activity. *Nat. Commun.* **9**, 4657 (2018).

173. Abad, M. J., Bermejo, P. & Villar, A. An approach to the genus *Tanacetum* L. (Compositae): Phytochemical and pharmacological review. *Phytother. Res.* **9**, 79–92 (1995).

174. Teoh, K. H., Polichuk, D. R., Reed, D. W., Nowak, G. & Covello, P. S. *Artemisia annua* L. (Asteraceae) trichome-specific cDNAs reveal CYP71AV1, a cytochrome P450 with a key role in the biosynthesis of the antimalarial sesquiterpene lactone artemisinin. *FEBS Lett.* **580**, 1411–1416 (2006).

175. Zhang, Y. *et al.* The molecular cloning of artemisinic aldehyde reductase and its role in glandular trichome-dependent biosynthesis of artemisinin in *Artemisia annua*. *J. Biol. Chem.* **283**, 21501–21508 (2008).

176. Judd, R. *et al.* Artemisinin Biosynthesis in Non-glandular Trichome Cells of *Artemisia annua*. *Mol. Plant* **12**, 704–714 (2019).

177. Xie, D.-Y., Ma, D.-M., Judd, R. & Jones, A. L. Artemisinin biosynthesis in *Artemisia annua* and metabolic engineering: questions, challenges, and perspectives. *Phytochem. Rev.* **15**, 1093–1114 (2016).

178. Bouwmeester, H. J. *et al.* Amorpha-4,11-diene synthase catalyses the first probable step in artemisinin biosynthesis. *Phytochemistry* **52**, 843–854 (1999).

179. Mercke, P., Bengtsson, M., Bouwmeester, H. J., Posthumus, M. A. & Brodelius, P. E. Molecular Cloning, Expression, and Characterization of Amorpha-4,11-diene Synthase, a Key Enzyme of Artemisinin Biosynthesis in *Artemisia annua* L. *Arch. Biochem. Biophys.* **381**, 173–180 (2000).

180. Rydén, A.-M. *et al.* The molecular cloning of dihydroartemisinic aldehyde reductase and its implication in artemisinin biosynthesis in *Artemisia annua*. *Planta Med.* **76**, 1778–1783 (2010).

181. Zhao, W. *et al.* Three new sesquiterpene glycosides from *Dendrobium nobile* with immunomodulatory activity. *J. Nat. Prod.* **64**, 1196–1200 (2001).

182. Tan, D. *et al.* Identification of sesquiterpene glycosides from *Dendrobium nobile* and their α-glycosidase and α-amylase inhibitory activities. *Food Sci. Technol.* **43**, (2023).

183. Zhao, M. *et al.* Sesquiterpene glucosylation mediated by glucosyltransferase UGT91Q2 is involved in the modulation of cold stress tolerance in tea plants. *New Phytol.* **226**, 362–372 (2020).

184. Epstein, W. W. & Poulter, C. D. A survey of some irregular monoterpenes and their biogenetic analogies to presqualene alcohol. *Phytochemistry* **12**, 737–747 (1973).

185. Roeder, S., Hartmann, A.-M., Effmert, U. & Piechulla, B. Regulation of simultaneous synthesis of floral scent terpenoids by the 1,8-cineole synthase of *Nicotiana suaveolens*. *Plant Mol. Biol.* **65**, 107–124 (2007).

186. Landmann, C. *et al.* Cloning and functional characterization of three terpene synthases from lavender (*Lavandula angustifolia*). *Arch. Biochem. Biophys.* **465**, 417–429 (2007).

187. Ilc, T., Parage, C., Boachon, B., Navrot, N. & Werck-Reichhart, D. Monoterpenol Oxidative Metabolism: Role in Plant Adaptation and Potential Applications. *Front. Plant Sci.* **7**, 509 (2016).

188.Lange, B. M. & Srividya, N. Enzymology of monoterpene functionalization in glandular trichomes. *J. Exp. Bot.* **70**, 1095–1108 (2019).

189. Lange, B. M. The Evolution of Plant Secretory Structures and Emergence of Terpenoid Chemical Diversity. *Annu. Rev. Plant Biol.* **66**, 139–159 (2015).

190. Mint Evolutionary Genomics Consortium. Electronic address: buell@msu.edu & Mint Evolutionary Genomics Consortium. Phylogenomic Mining of the Mints Reveals Multiple Mechanisms Contributing to the Evolution of Chemical Diversity in Lamiaceae. *Mol. Plant* **11**, 1084–1096 (2018).

191. Krause, S. T. *et al.* The biosynthesis of thymol, carvacrol, and thymohydroquinone in Lamiaceae proceeds via cytochrome P450s and a short-chain dehydrogenase. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2110092118 (2021).

192. Sun, M. *et al.* Chromosome-level assembly and analysis of the *Thymus* genome provide insights into glandular secretory trichome formation and monoterpenoid biosynthesis in thyme. *Plant Commun.* **3**, 100413 (2022).

193. Kitajima, J., Ishikawa, T., Urabe, A. & Satoh, M. Monoterpenoids and their glycosides from the leaf of thyme. *Phytochemistry* **65**, 3279–3287 (2004).

194. Lichman, B. R. *et al.* The evolutionary origins of the cat attractant nepetalactone in catnip. *Sci. Adv.* **6**, eaba0721 (2020).

195. Vining, K. J. *et al.* Chromosome-level genome assembly of *Mentha longifolia* L. reveals gene organization underlying disease resistance and essential oil traits. *G3 GenesGenomesGenetics* **12**, jkac112 (2022).

196. Geu-Flores, F. *et al.* An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. *Nature* **492**, 138–142 (2012).

197. Lichman, B. R. *et al.* Uncoupled activation and cyclization in catmint reductive terpenoid biosynthesis. *Nat. Chem. Biol.* **15**, 71–79 (2019).

198. Vining, K. J. *et al.* Draft Genome Sequence of *Mentha longifolia* and Development of Resources for Mint Cultivar Improvement. *Mol. Plant* **10**, 323–339 (2017).

199. Höfer, R. *et al.* Dual function of the cytochrome P450 CYP76 family from *Arabidopsis thaliana* in the metabolism of monoterpenols and phenylurea herbicides. *Plant Physiol.* **166**, 1149–1161 (2014).

200. Boachon, B. *et al.* CYP76C1 (Cytochrome P450)-Mediated Linalool Metabolism and the Formation of Volatile and Soluble Linalool Oxides in Arabidopsis Flowers: A Strategy for Defense against Floral Antagonists. *Plant Cell* **27**, 2972–2990 (2015).

201. Martin, D. M. *et al.* Functional Annotation, Genome Organization and Phylogeny of the Grapevine (*Vitis vinifera*) Terpene Synthase Gene Family Based on Genome Assembly, FLcDNA Cloning, and Enzyme Assays. *BMC Plant Biol.* **10**, 226 (2010).

202. Ilc, T. *et al.* A grapevine cytochrome P450 generates the precursor of wine lactone, a key odorant in wine. *New Phytol.* **213**, 264–274 (2017).

203. Matsuda, K. Pyrethrin Biosynthesis and Its Regulation in *Chrysanthemum cinerariaefolium* in *Pyrethroids: From Chrysanthemum to Modern Industrial Insecticide* (eds. Matsuo, N. & Mori, T.) 73–81 (Springer, 2012). doi:10.1007/128_2011_271.

204. Yang, T. *et al.* Chrysanthemyl Diphosphate Synthase Operates in Planta as a Bifunctional Enzyme with Chrysanthemol Synthase Activity*. *J. Biol. Chem.* **289**, 36325–36335 (2014).

205. Xu, H. *et al.* Pyrethric acid of natural pyrethrin insecticide: complete pathway elucidation and reconstitution in *Nicotiana benthamiana*. *New Phytol.* **223**, 751–765 (2019).

206. Xu, H. *et al.* Coexpression Analysis Identifies Two Oxidoreductases Involved in the Biosynthesis of the Monoterpene Acid Moiety of Natural Pyrethrin Insecticides in *Tanacetum cinerariifolium*. *Plant Physiol.* **176**, 524–537 (2018).

207. Kikuta, Y. *et al.* Identification and characterization of a GDSL lipase-like protein that catalyzes the ester-forming reaction for pyrethrin biosynthesis in *Tanacetum cinerariifolium*– a new target for plant protection. *Plant J.* **71**, 183–193 (2012).

208. Boachon, B. *et al.* A Promiscuous CYP706A3 Reduces Terpene Volatile Emission from Arabidopsis Flowers, Affecting Florivores and the Floral Microbiome. *Plant Cell* **31**, 2947–2972 (2019).

209. Tholl, D., Chen, F., Petri, J., Gershenzon, J. & Pichersky, E. Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from *Arabidopsis* flowers. *Plant J.* **42**, 757–771 (2005).

210. Callicarpa L. | Plants of the World Online | Kew Science. *Plants of the World Online* http://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:30044566-2.

211. Tu, Y., Sun, L., Guo, M. & Chen, W. The medicinal uses of *Callicarpa* L. in traditional Chinese medicine: An ethnopharmacological, phytochemical and pharmacological review. *J. Ethnopharmacol.* **146**, 465–481 (2013).

212. Dictionary of Natural Products 31.2. Accessed March 2023. https://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml. Accessed 22nd February 2023.

213. Rodríguez-López, C. E. *et al.* Phylogeny-Aware Chemoinformatic Analysis of Chemical Diversity in Lamiaceae Enables Iridoid Pathway Assembly and Discovery of Aucubin Synthase. *Mol. Biol. Evol.* **39**, msac057 (2022).

CHAPTER 2: IDENTIFICATION OF KEY DITERPENE SYNTHASES USING A CHROMOSOME-SCALE GENOME ASSEMBLY OF THE INSECT-REPELLENT TERPENOID-PRODUCING LAMIACEAE SPECIES, *CALLICARPA AMERICANA*

Emily R. Lanier, Wajid W. Bhat, John P. Hamilton, Bjoern R. Hamberger, Robin C. Buell.

This chapter is excerpted from the following publication:

Hamilton, J. P. *et al.* Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, *Callicarpa americana*. *GigaScience* **9**, giaa093 (2020).

Author contributions:

The publication associated with the work in this chapter with this chapter describes the genome assembly of *Callicarpa americana*, which was generated and described by Dr. Robin Buell, Dr. John Hamilton, and the rest of the Buell lab group. The abstract and introduction from the original publication, adapted below, was written by the Buell lab along with Figure 2.1. My contribution to this publication is the characterization of key diterpenoid enzymes using gene models described as part of the reported assembly. Together Dr. Wajid Bhat and I analyzed the set of diterpene synthases identified in the genome assembly. I then characterized 3 of these diterpene synthase enzymes. I wrote the "Specialized Metabolism Analyses" section as well as creating Figures 2.4 and 2.5. Wajid Bhat created Figures 2.2 and 2.3. The Methods section was written by myself and Wajid Bhat.

Abstract

Plants exhibit wide chemical diversity due to the production of specialized metabolites which function as pollinator attractants, defensive compounds, and signaling molecules. Lamiaceae (mints) are known for their chemodiversity and have been cultivated for use as culinary herbs as well as sources of insect repellents, health-promoting compounds, and fragrance. *Callicarpa americana* (American beautyberry) is a species within the early-diverging Callicarpoideae clade of Lamiaceae, known for its metallic purple fruits and use as an insect repellent due to its production of terpenoids. The diterpene intermediate kolavenyl diphosphate is a gateway to many of *C. americana*'s bioactive terpenoids. Using the chromosome-scale genome assembly generated by our collaboratoes as part of this project, we identified 53 terpene synthase gene candidates and mapped their expression in 8 tissue types. Experimental validation confirmed that *CamTPS2* encodes a kolavenyl diphosphate synthase, demonstrating that access to the *C. americana* genome provides a roadmap for rapid discovery of genes encoding plant-derived agrichemicals and a key resource for understanding the evolution of chemical diversity in Lamiaceae.

Introduction

Mints (Lamiaceae) are the sixth largest family of flowering plants and include many species grown for use as culinary herbs (basil, rosemary, thyme), food additives and flavorings (peppermint, spearmint), pharmaceuticals and health-promoting activities (skullcap, bee balm), feline euphoria induction (catnip), wood (teak), fragrance (lavender, patchouli), insect repellents (peppermint, rosemary), and ornamentals (coleus, chaste tree, beautyberry). This diverse set of uses for Lamiaceae is due in part to their production of specialized metabolites, primarily terpenes (monoterpenes, sesquiterpenes, diterpenes) and iridoids (irregular terpenes). Through an integrated phylogenetic-genomic-chemical approach, the evolutionary basis of Lamiaceae chemical diversity was shown to involve gene family expansion, differential gene expression, diversion of metabolic flux, and parallel evolution[1]. Genome sequences are currently available for a number of Lamiaceae species and are providing new insights into these phenomena, yet are primarily limited to members of Nepetoideae[2-5], the most species- and monoterpene-rich of the 12 major mint clades (= traditional subfamilies). As for the remaining major clades, a genome sequence is available only for *Tectona grandis* L. f. (teak; Tectonoideae)[6]. To expand our knowledge of the genome evolution underlying chemodiversity in this important family, we generated a chromosome-scale assembly of *Callicarpa americana* (American beautyberry), a species renowned for its charismatic purple fruits (Fig. 2.1). *Callicarpa* occupies a pivotal phylogenetic position as a representative from the early-diverging mint lineage, Callicarpoideae[1]. The species is native to North America (southern U. S. A., northern Mexico), North Atlantic (Bermuda, Bahamas), and Cuba, and has known insect repellent activity[7,8] due to production of spathulenol, intermedeol, and callicarpenal[9]. Access to its genome will enable discovery of the

genes encoding the biosynthetic pathways for these terpenes and the potential for heterologous expression of botanical-derived insect repellents; the genome is also an important evolutionary reference for the mint family.

Figure 2.1. ***Callicarpa americana.*** American beautyberry plant with fruit.

Results

Specialized metabolite analyses

Callicarpa americana produces a range of bioactive diterpenoids derived from the C_{20} clerodane skeleton[10], a less common instance of the labdanoid diterpenes. These include the C_{16} nor-diterpenoid (-)-callicarpenal with a range of mosquito, tick, and arthropod repellent activities[8]. Clerodane-type diterpenes are derived from the precursor kolavenyl diphosphate (KPP), which is formed by class II diterpene synthases (diTPS) of the terpene synthase c (TPS-c) subfamily[11-13]. Here, we describe the annotation and validation of the KPP synthase in *C. americana*, a gateway to many of its bioactive terpenoids. Using the assembled genomic and transcriptomic data, we performed a sequence similarity search with BLASTP comparing the *C. americana* peptide models against a set of reference terpene synthases (TPSs). Peptides shorter than 350 amino acids or

having less than 30% identity to the most similar reference sequence were filtered out, yielding a total of 53 candidate TPSs. We used phylogenetic clustering (Fig. 2.2) with known TPSs to identify and classify candidates most likely to catalyze the formation of KPP. The complement and distribution of TPSs discovered was found in accordance with plant species[14], reflecting general metabolism and species-specific evolution of specialized metabolism in *C. americana*. Specifically, our study resulted in eight putative diTPSs from the TPS-c subfamily; class II diTPSs are typically involved in formation of the necessary diphosphate intermediates of the labdane-type chemistry. Of the eight candidates, four were successfully cloned from cDNA and transferred into the plant expression vector pEAQ[6] as described previously. The others were not further pursued due to low expression levels or a lack of expression in tissues relevant for callicarpenal formation. Expression analysis of tissue-specific accumulation of transcripts for the diTPSs (Fig. 2.3) showed the highest expression in young leaves and flowers for *CamTPS2*, consistent with the presence of callicarpenal in leaves.

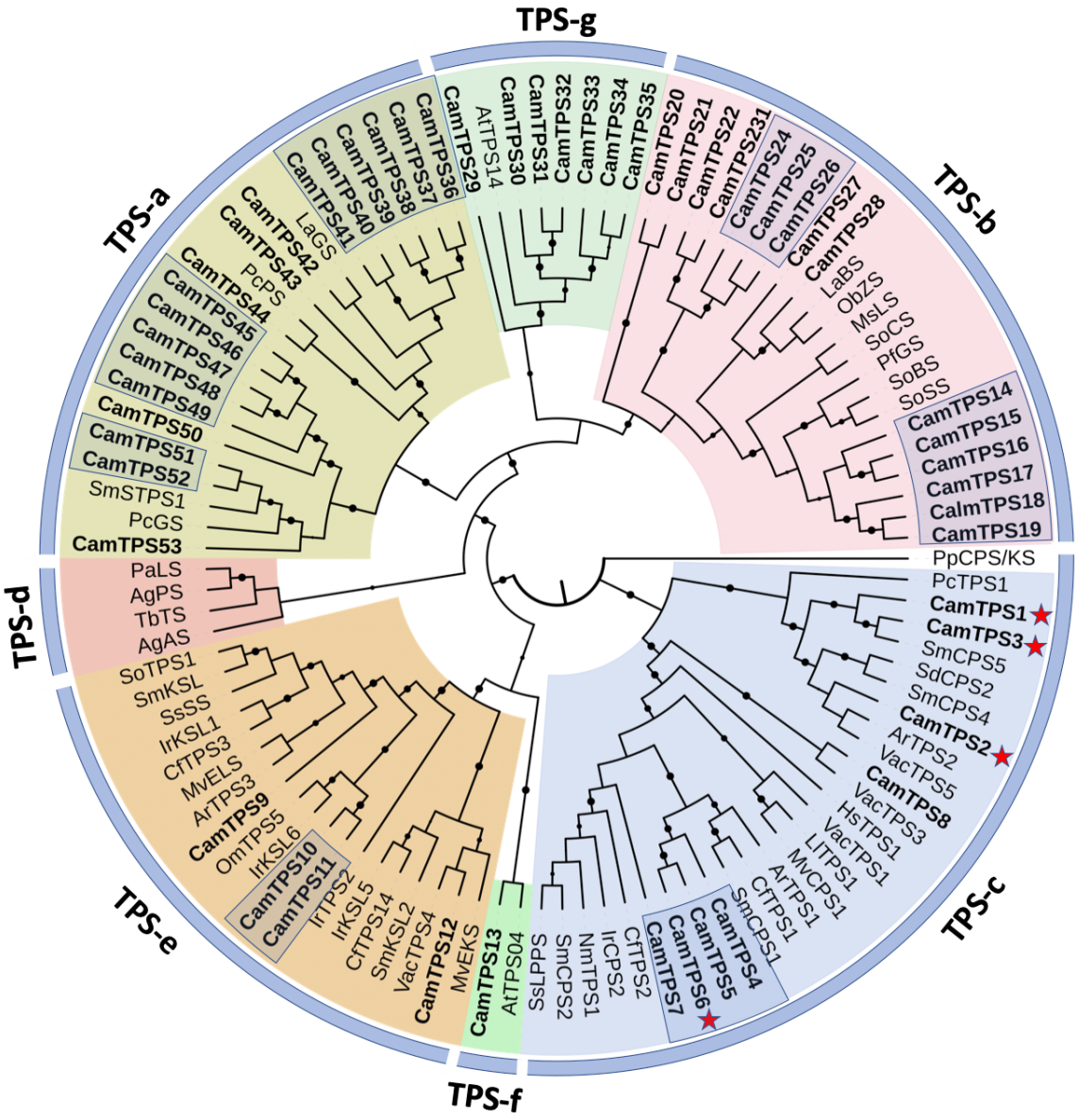


Figure 2.2. Phylogenetic analysis and classification of the *Callicarpa americana* terpene synthase family. Shown are the distinct terpene synthase gene families TPS-a through TPS-g. Highlighted in boxes are TPSs clustered in proximity on the genomic pseudomolecules. *C. americana* TPSs are in boldface; red stars indicate functionally characterized members of the TPS-c subfamily; dots on branches indicate bootstrap support ≥80%. The phylogeny was rooted with the bifunctional *Physcomitrella patens* (moss) PpCPS/EKS.

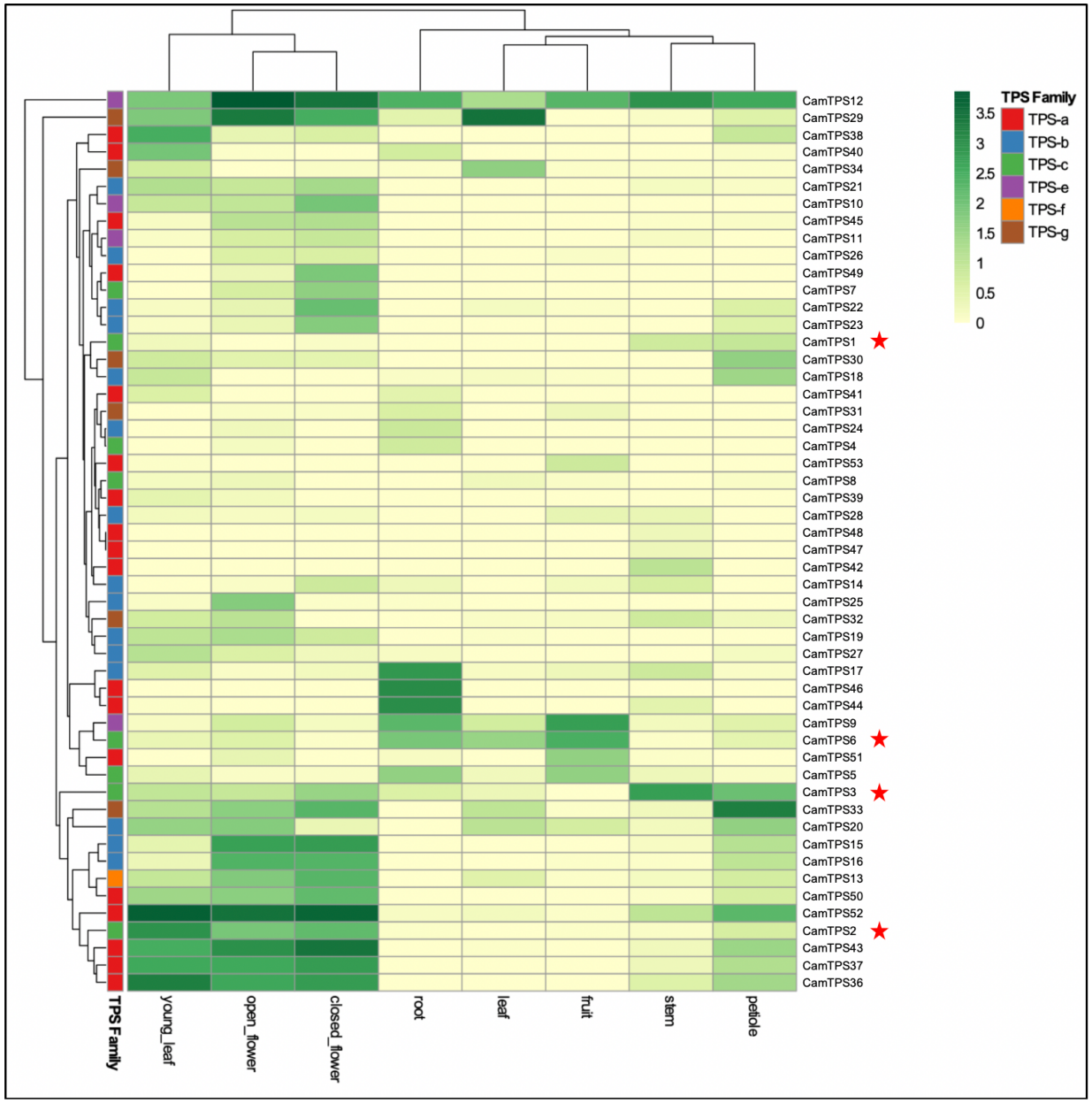


Figure 2.3. Tissue-specific expression of the *Callicarpa americana* terpene synthase gene family. Expression is in transcripts per million. Red stars indicate functionally characterized members of the TPS-c subfamily.

Characterization of the candidates through transient expression in *Nicotiana benthamiana* and gas chromatography–mass spectrometry (GC-MS) analysis showed that CamTPS1 and CamTPS3 catalyze the formation of *ent*-copalyl diphosphate (Fig. 2.4), the first step in the biosynthesis of

the ubiquitous *ent*-kaurane type plant growth hormone gibberellic acid (GA) (Fig. 2.5) and specialized metabolites in the *ent*-configuration found in this genus. CamTPS6 yielded (+)-copalyl diphosphate, precursor of calliterpenone, a rare (+)-kaurane type diterpene found across several species of *Callicarpa*[10]. (+)-Copalyl diphosphate is also the intermediate to the common diterpene miltiradiene, precursor to many defense related diterpenoids found in other Lamiaceae and previously identified in other *Callicarpa* species[10]. Finally, CamTPS2 was confirmed to yield the possible precursor of callicarpenal, KPP. All products were confirmed by comparison with reference combinations of diTPS (Fig. 2.4). Together, this set of diTPSs yielded all plausible precursors to the known chemical diversity of diterpene scaffolds in *C. americana* (Fig. 2.5).

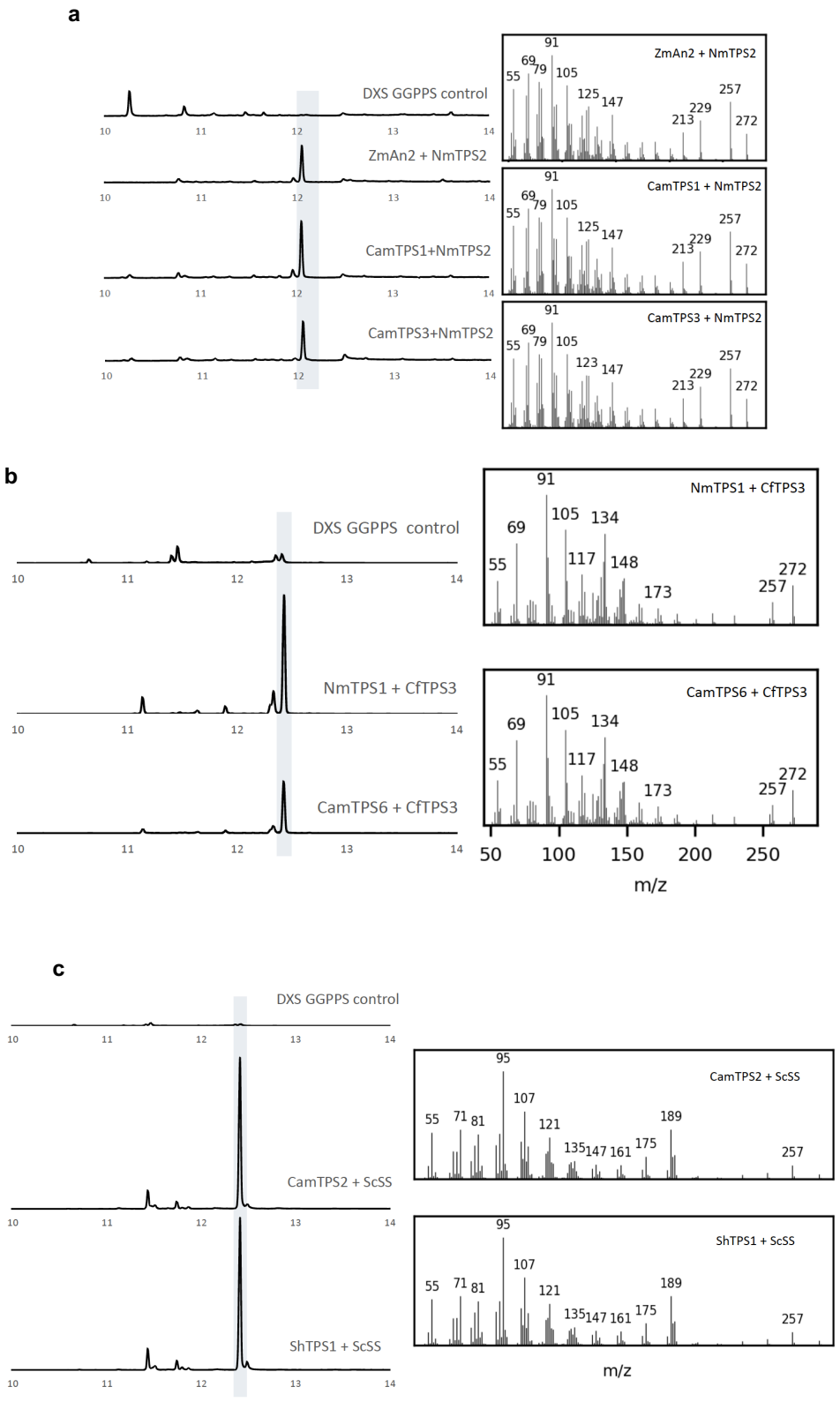


Figure 2.4. GC-MS data for TPS enzymes investigated alongside reference diTPS enzymes.

Figure 2.4 (cont'd).
Each class II diTPS is paired with a characterized class I diTPS and elution time/mass spectra compared to a pair of reference diTPS. (a) *CamTPS1* and *CamTPS3* paired with *NmTPS2* catalyze formation of *ent*-kaurene, thus confirming that *CamTPS1* and *CamTPS3* both catalyze formation of *ent*-CPP. The reference pair of *ZmAn2* + *NmTPS2* makes *ent*-kaurene from *ent*-CPP. (b) *CamTPS6* paired with *CfTPS3* catalyzes formation of miltiradiene, confirming activity as a (+)-CPP synthase. The reference pair NmTPS1 + CfTPS3 makes miltiradiene from (+)-CPP. C, CamTPS2 paired with ScSS makes kolavelool, confirming CamTPS2 as a KPP synthase. The reference pair ShTPS1 + ScSS makes kolavelool from KPP. Reference enzymes NmTPS1, NmTPS2 *Nepeta mussini*; CfTPS3, *Coleus forskohlii*; ZmAN2, *Zea mays*; SsSCS, *Salvia sclarea*[15-19].

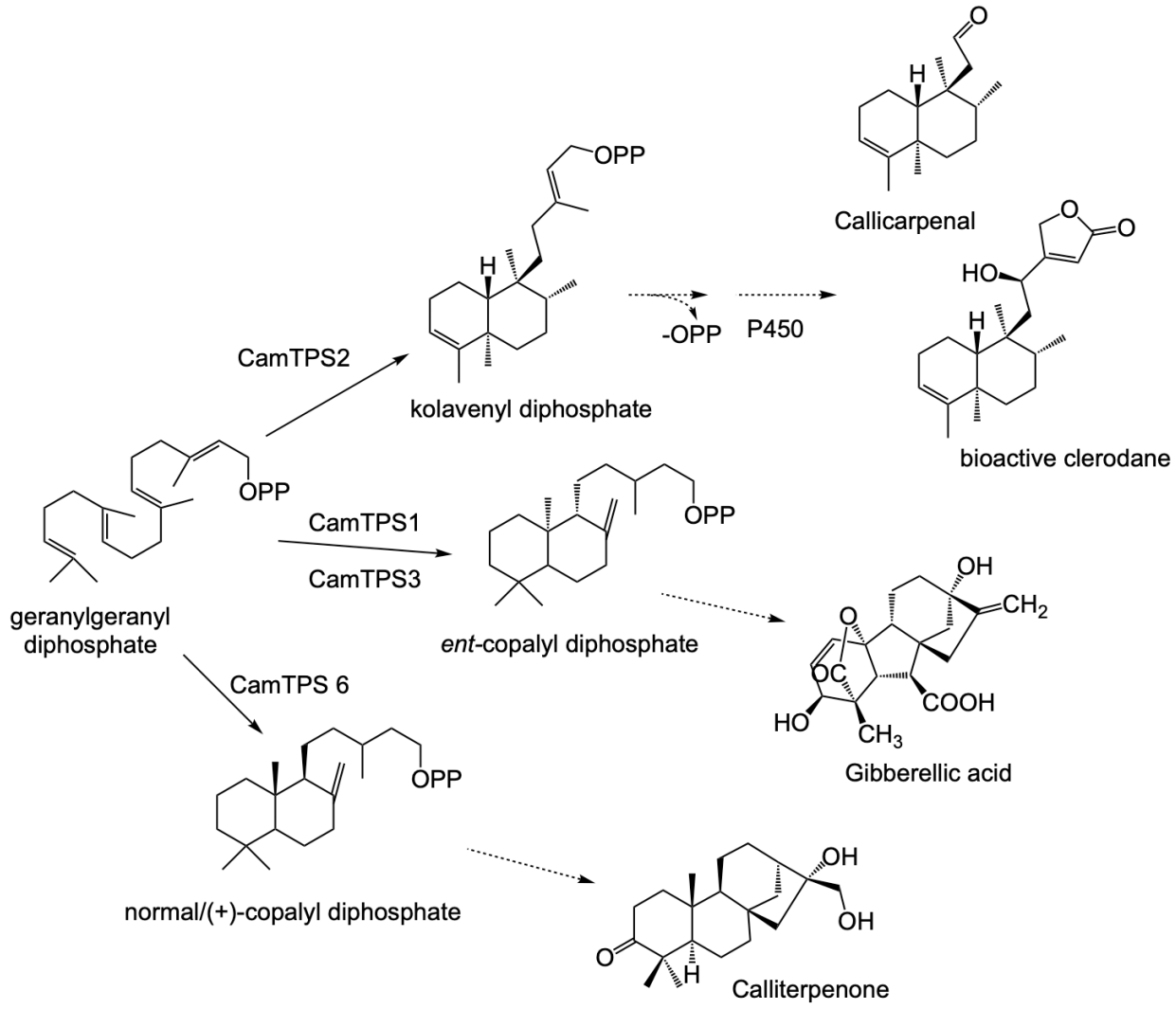


Figure 2.5. Activities of functionally characterized *Callicarpa americana* L. TPS-c. Dotted arrows indicate putative further functionalization by class I diTPS and cytochromes P450 to diterpene products accumulating in *C. americana.*

Conclusion

The insect repellent activity of *C. americana* is due to the production of the terpenoids spathulenol, intermedeol, and callicarpenal[9], and access to a chromosome-scale genome assembly of *C. americana* permitted identification of *CamTPS2* which synthesizes kolavenyl diphosphate, a precursor to callicarpenal. As the sixth largest angiosperm family, and with extensive chemical diversity, Lamiaceae are an ideal group for application of phylogenomic data-mining, a powerful approach for biosynthetic pathway discovery. Generation of the genome of *C. americana*, of the early-diverging Callicarpoideae clade of Lamiaceae, provides a roadmap for rapid discovery of genes encoding plant-derived agrichemicals and a key resource for understanding the evolution of both chemical diversity and mint genomes.

Methods

Phylogenetic tree

C. americana TPSs were identified by Blastp (v. 2.2.31+)[20] using a set of reference terpene synthases across all TPS-subfamilies against the gene models. Hits with less than 350 amino acids or less than 30% identity to the reference sequences were filtered out. Sequences were aligned using the MUSCLE program from MEGA[21], using default parameters and the alignment was manually verified for consistency. A maximum likelihood tree was generated using Jones-Taylor-Thornton model with MEGA X[21] with 1,000 bootstrap repetitions. The tree figure was generated using FigTree v1.4.3[22].

Heatmap generation

Gene expression heat maps were generated by using ClustVis web tool[23] with the default routine, using TPM values of the TPS gene expression in different tissues of *C. americana*.

RNA extraction and transcriptome sequencing

For transcriptome analyses, RNA was isolated from mature and young leaves, stems, petioles, roots, flowers (open and closed), and ripened whole fruits (denoted by the deep purple color) from growth chamber–grown plants using a hot phenol method[24]. Illumina TruSeq Stranded mRNA (polyA mRNA) libraries were constructed and sequenced on an Illumina HiSeq 4000 to 150 nt in paired-end mode. All Illumina sequencing was performed at the Research Technology Support Facility at Michigan State University.

Cloning

From RNA, cDNA was prepared using the Invitrogen SuperScript™ IV One-Step RT-PCR System. After cloning into pJET1.2 (Thermo Fisher Scientific, Waltham, MA, USA), TPSs were transferred into pEAQ-HT[25] using In-Fusion® HD Cloning Plus (Takara Bio, Mountain View, California, USA) for transient expression in *Nicotiana benthamiana*. Oligonucleotides for cloning of *C. americana* TPS candidates (given in 5' to 3'):

Cam_TPS1_For	AAGCTCTCCTCTGCCGTTAAA
Cam_TPS1_Rev	CACAACTTTCATGTACATACTATACC
Cam_TPS2_For	ATGTCATTTGCTTCCCATGCCA
Cam_TPS2_Rev	CAGAACAGGAAGTGTAACTCTACC
Cam_TPS3_For	TCCAATCACACCAACGTTAATTTC
Cam_TPS3_Rev	GATTTACATGTACGTACATGGTCAGAG
Cam_TPS6_For	CTTTGCTACACTGCAGACAAC
Cam_TPS6_Rev	AGTTCGACCGAATTGCGGAAACA

***Functional characterization of diTPSs by transient expression in* N. benthamiana**

DiTPS candidates and reference genes were transiently expressed in *N. benthamiana* leaves as previously described[16]. To increase product accumulation, diTPSs were co-expressed with genes from the upstream pathway providing the substrate, CfDXS and CfGGPPS (Cf, *Coleus forskohlii*)[17, 18]. Cultures containing different constructs were mixed in equal ratios to yield the appropriate combinations before infiltration into 4-5 weeks old plants. Plants were grown for an additional five days before metabolite extraction. Leaf discs of 2 cm diameter (approximately 0.1 g fresh weight) were cut from the infiltrated leaves. Diterpenes were extracted in 1 mL n-hexane with 1 mg/L 1-eicosene as internal standard (IS) at room temperature overnight in an orbital shaker at 200 rpm. Plant material was collected by centrifugation and the organic phase transferred to GC vials for analysis.

GC-MS Analysis

GC-MS analyses were performed on an Agilent 7890A GC with an Agilent VF-5ms column (30 m x 250 µm x 0.25 µm, with 10m EZ-Guard) and an Agilent 5975C detector. The inlet was set to 275°C splitless injection, He carrier gas with column flow of 1 mL/min. The oven program was 40°C hold 1 min, 40 °C/min to 200°C and hold 4.5 min, 20°C/min to 240°C, 10°C/min to 280°C, 40°C/min to 320°C hold 3 min. The detector was activated after a four-minute solvent delay. All analyses were done in duplicate.

Availability of supporting information

All sequences used in this study are available in the NCBI SRA under BioProject PRJNA529675. The genome assembly, annotation files, expression matrix, and other supporting data presented in the published version of this work can be accessed at the GigaScience GigaDB database[26].

Genbank accession identifiers for cloned TPSs are MT083919–MT083922. Original raw GC-MS data were deposited to Zenodo[27] and Metabolights[28] under accession MTBLS1983. Supplemental files are available with the online publication[29].

REFERENCES

1. Mint Evolutionary Genomics Consortium. Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. *Mol. Plant* **11**(8), 1084–96 (2018).

2. Xu H., Song J., Luo H., *et al.* Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*. *Mol Plant* **9**(6), 949–952 (2016).

3. Malli R.P.N., Adal A.M., Sarker L.S., *et al.* De novo sequencing of the *Lavandula angustifolia* genome reveals highly duplicated and optimized features for essential oil production. *Planta* **249**(1), 251–256 (2019).

4. Dong A.X., Xin H.B., Li Z.J., *et al.* High-quality assembly of the ref- erence genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. *Gigascience* **7**(7), doi:10.1093/gigascience/giy068 (2018).

5. Zhao Q., Yang J., Cui M.Y., *et al.* The reference genome sequence of *Scutellaria baicalensis* provides insights into the evolution of wogonin biosynthesis. *Mol. Plant.* **12**(7), 935–950 (2019).

6. Zhao D., Hamilton J.P., Bhat W.W., *et al.* A chromosomal-scale genome assembly of *Tectona grandis* reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *Gigascience* **8**(3), doi:10.1093/gigascience/giz005 (2019).

7. Krajick K. Medical entomology. Keeping the bugs at bay. *Science* **313**(5783), 36–8 (2006).

8. Cantrell C.L., Klun J.A. Callicarpenal and intermedeol: two natural arthropod feeding deterrent and repellent compounds identified from the southern folk remedy plant, *Callicarpa americana*. *Recent Developments in Invertebrate Repellents*. Washington, DC: American Chemical Society; 47– 58 (2011).

9. Cantrell C.L., Klun J.A., Bryson C.T., *et al.* Isolation and identification of mosquito bite deterrent terpenoids from leaves of American (*Callicarpa americana*) and Japanese (*Callicarpa japonica*) beautyberry. *J. Agric. Food Chem.* **53**(15), 5948–53 (2005).

10. Jones W.P., Kinghorn A.D. Biologically active natural products of the genus Callicarpa. *Curr. Bioact. Compd.* **4**, 15–32 (2008).

11. Hansen N.L., Heskes A.M., Hamberger B., *et al.* The terpene synthase gene family in *Tripterygium wilfordii* harbors a labdane-type diterpene synthase among the monoterpene synthase TPS-b subfamily. *Plant J.* **89**, 429–441 (2017).

12. Chen X., Berim A., Dayan F.E., *et al.* A (–)-kolavenyl diphosphate synthase catalyzes the first step of salvinorin A biosynthesis in *Salvia divinorum*. *J. Exp. Bot*. **68**, 1109–1122 (2017).

13. Pelot K.A., Mitchell R., Kwon M., *et al.* Biosynthesis of the psychotropic plant diterpene salvinorin A: discovery and characterization of the *Salvia divinorum* clerodienyl diphosphate synthase. *Plant J.* **89**, 885–897 (2017).

14. Jiang S.Y., Jin J., Sarojam R., *et al.* A comprehensive survey on the terpene synthase gene family provides new insight into its evolutionary patterns. *Genome Biol. Evol.* **11**(8), 2078–2098 (2019).

15. Pelot K.A., Mitchell R., Kwon M., *et al.* Biosynthesis of the psychotropic plant diterpene salvinorin A: discovery and characterization of the *Salvia divinorum* clerodienyl diphosphate synthase. *Plant J.* **89**, 885–897 (2017).

16. Johnson S.R., Bhat W.W., Bibik J., Mint Evolutionary Genomics Consortium, *et al.* A database-driven approach identifies additional diterpene synthase activities in the mint family (Lamiaceae). *J. Biol. Chem.* **25**, 1349–62 (2018).

17. Andersen-Ranberg J., Kongstad K.T., Nielsen M.T., *et al.* Expanding the landscape of diterpene structural diversity through stereochemically controlled combinatorial biosynthesis. *Angew. Chem. Int. Ed. Engl.* **55**(6), 2142–6 (2016).

18. Pateraki I., Andersen-Ranberg J., Hamberger B., *et al.* Manoyl oxide (13R), the biosynthetic precursor of forskolin, is synthesized in specialized root cork cells in *Coleus forskohlii*. *Plant Physiol.* **164**(3), 1222–36 (2014).

19. Harris L.J., Saparno A., Johnston A., *et al.* The maize An2 gene is induced by Fusarium attack and encodes an *ent*-Copalyl diphosphate synthase. *Plant. Mol. Biol.* **59**, 881–894 (2005).

20. Camacho C., Coulouris G., Avagyan V., *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

21. Kumar S., Stecher G., Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**(7), 1870–1874 (2016).

22. Figtree. http://tree.bio.ed.ac.uk/software/figtree/. Accessed January 2020.

23. ClustVis web tool. https://biit.cs.ut.ee/clustvis/. Accessed November 2019.

24. Davidson R.M., Gowda M., Moghe G., *et al.* Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J*. **71**(3), 492–502 (2012).

25. Sainsbury F., Thuenemann E.C., Lomonossoff G.P. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* 7, 682–93 (2009).

26. Hamilton J.P., Godden G.T., Lanier E., *et al.* Supporting data for "Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, Callicarpa americana." *GigaScience Database* doi.org/10.5524/100777 (2020).

27. Hamilton J.P., Godden G.T., Lanier E., *et al.* GC-MS data set for generation of a chromosome-scale genome assembly of the insect-repellant terpenoid-producing Lamiaceae species, Callicarpa americana. *Zenodo* doi.org/10.5281/zenodo.3672159 (2020).

28. MetaboLights. https://www.ebi.ac.uk/metabolights/. Accessed August 2020.

29. Hamilton, J.P., Godden G.T., Lanier E., *et al.* Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, *Callicarpa americana. GigaScience* **9**, giaa093 (2020).

CHAPTER 3: UNCOVERING A MILTIRADIENE BIOSYNTHETIC GENE CLUSTER IN THE LAMIACEAE REVEALS A DYNAMIC EVOLUTIONARY TRAJECTORY

Abigail E. Bryson[†1], Emily R. Lanier[†1], Kin H. Lau[‡2], John P. Hamilton[2,3], Brieanne Vaillancourt[2,3], Davis Mathieu[1], Alan E. Yocca[2,4], Garret P. Miller[1], Patrick P. Edger[4], C. Robin Buell[2,5], Björn Hamberger[*1]

[1]Department of Biochemistry, Michigan State University, East Lansing, USA; [2]Department of Plant Biology, Michigan State University, East Lansing, USA; [3]Center for Applied Genetic Technologies, University of Georgia, Athens, USA; [4]Department of Horticulture, Michigan State University, East Lansing, USA; [5]Plant Resilience Institute, Michigan State University, East Lansing, USA.

[†]These authors contributed equally.

*Corresponding author: hamberge@msu.edu

[‡]Currently: Bioinformatics and Biostatistics Core, Van Andel Institute, Grand Rapids, USA.

This chapter was first published in:

Bryson A. B. Lanier, E.R., Lau K.H., Hamilton J.P., Vaillancourt B., Mathieu D., Yocca A.E., Miller G.P., Edger P.P., Buell C.R., and Hamberger B. Uncovering a miltiradiene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory. *Nat. Commun.* **14**, 343 (2023). doi: 10.1038/s41467-023-35845-1

Author contributions:

AEB, ERL, and BH conceived and designed the study; AEB and ERL performed the experiments; AEB and DM performed and analyzed the synteny; KHL assembled and annotated the genomes; BV and JPH performed genome analyses; AEY performed ancestral state reconstruction; ERL and GPM analyzed the experimental data; AEB, ERL, and PPE generated and analyzed the phylogenetic relationships; AEB, ERL, and BH wrote the manuscript; BH and CRB supervised the project; all authors contributed to revisions.

Abstract

The spatial organization of genes within plant genomes can drive evolution of specialized metabolic pathways. Terpenoids are important specialized metabolites in plants with diverse adaptive functions that enable environmental interactions. Here, we report the genome assemblies of *Prunella vulgaris*, *Plectranthus barbatus*, and *Leonotis leonurus*. We investigate the origin and subsequent evolution of a diterpenoid biosynthetic gene cluster (BGC) together with other seven species within the Lamiaceae (mint) family. Based on core genes found in the BGCs of all species examined across the Lamiaceae, we predict a simplified version of this cluster evolved in an early Lamiaceae ancestor. The current composition of the extant BGCs highlights the dynamic nature of its evolution. We elucidate the terpene backbones generated by the *Callicarpa americana* BGC enzymes, including miltiradiene and the terpene (+)-kaurene, and show oxidization activities of BGC cytochrome P450s. Our work reveals the fluid nature of BGC assembly and the importance of genome structure in contributing to the origin of metabolites.

Introduction

Plants are renowned for their incredible diversity of specialized metabolites, which function in interactions with their environment. These biosynthetic pathways are dynamic, facilitating continual evolution of novel compounds. The rising number of high-quality plant genomes published in recent years has led to the discovery that some metabolic pathways are organized into biosynthetic gene clusters (BGCs). A BGC is a group of two or more different classes of non-homologous genes which are physically clustered, transcriptionally linked, and functionally related[1–6]. Over 30 plant BGCs have been functionally validated to date[7] since the discovery of the first BGC in maize[8]. The BGCs found in plants are predominately involved in specialized rather than central metabolism[9] and occur in multiple classes of compounds including benzylisoquinoline alkaloids in poppy[10,11], triterpenoid cucurbitacins in Cucurbitaceae[12,13], and diterpenoid momilactones in Poaceae and other cereals[14–18].

How and why BGCs form is still a topic of discussion, although several hypotheses are emerging. In bacteria and fungi, BGCs are common and aid in transference of the entire pathway during horizontal gene transfer[19,20]. While there is no evidence of horizontal gene transfer of plant BGCs reported thus far, BGCs still offer advantages in vertical inheritance of biosynthetic pathways[5,21]. The genetic linkage conveyed by BGCs facilitates coinheritance, which can protect the integrity of the entire pathway[22–24]. In some pathways, such as momilactone biosynthesis, loss of a single gene would result in a buildup of toxic intermediates[23]. Another fitness benefit of BGCs is the possibility of coregulation, such as by a single transcription factor or regulatory region. This can provide an energetically favorable control of the metabolite production in a tissue or

developmental stage-specific manner[5,16,21,25–28]. Regulation may also take place at the chromatin level, with DNA and histone methylation regulating transcription of the entire cluster[25,29–31].

Since the study of plant BGCs is still in its infancy, their origins and evolution are also not well understood. So far, evidence supports that plant BGCs have likely arisen from gene or genome duplication and/or genomic rearrangements[5]. BGC formation may be enhanced in highly active regions of the genome, such as the recent work detailing assembly of the oat avenacin BGC in a sub-telomeric region[32]. The birth of a gene cluster may begin with a single colocalized gene pair. Subsequent colocalization of additional classes of enzymes can occur through chromosomal remodeling or transposition[5,21,30,33]. Expansion of the cluster can also continue through tandem, local, or whole genome duplication[4,6,33–35]. The inherent promiscuity of enzymes involved in specialized metabolism enables rapid neofunctionalization, promoting functional divergence of BGCs as they evolve through different plant lineages[34,36–38]. Recent work has shown conservation of core genes and diversification into new functions/pathways when comparing BGCs across different plant families[6,39].

Terpenoids are a class of specialized metabolites that are well represented among the studied BGCs. Plant terpenoids are incredibly diverse and encompass over 65,000 structures[40], making them the largest known class of plant natural products. Plants rely on terpenoids for many interactions including pathogen and herbivore defense, signaling, and pollinator attraction[41–43]. Terpene synthases (TPSs) catalyze formation of terpene backbones from diphosphate isoprenoid precursors and are classified into seven subfamilies (a-h) based on their phylogenetic relationships[41,44,45]. The bicyclic labdane-type diterpenes are typically formed by the sequential activity of a class II (TPS-c) followed by a class I (TPS-e) diterpene synthase (diTPS). Class II diTPSs

catalyze a proton mediated cyclization of a 20-carbon isoprenoid diphosphate, usually geranylgeranyl diphosphate (GGPP), to form the characteristic decalin core. A class I diTPS then cleaves the diphosphate and may further differentiate the diterpene backbone. Diterpene backbones are functionalized by other enzyme classes through oxidation and subsequent conjugation to increase bioactivity. Cytochromes P450 (CYPs), particularly in the expansive CYP71 clan, often oxidize terpenes and have been found colocalized with TPSs either as pairs or as expanded BGCs[46, 47].

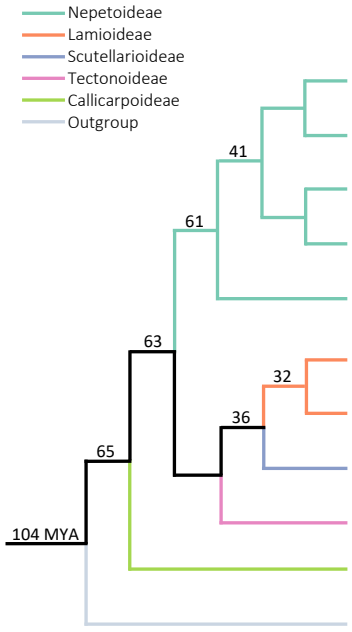
Terpenoid diversity is particularly rich in the Lamiaceae (mint) family[48,49]. Genome assemblies for 22 different Lamiaceae species (Supplementary Table 1) have been published to date, revealing BGCs for at least two classes of terpenoids: monoterpene-derived nepetalactones from catnip (*Nepeta* sp.)[50] and diterpenoid tanshinones in the Chinese medicinal herb Danshen (*Salvia miltiorrhiza*)[24,51,52]. Tanshinones are studied for their potent pharmacological activities, and as a result much of the biosynthetic pathway has been elucidated (Supplementary Fig. 1)[24,51–60]. The terpene backbone of the tanshinones is miltiradiene, a labdane diterpene formed by a class II (+)-copalyl diphosphate ((+)-CPP) synthase followed by the class I miltiradiene synthase. The abietane-type diterpenoid miltiradiene is the likely terpene precursor to a wide array of bioactive diterpenoids that are common throughout the Lamiaceae and beyond[61]. The antimicrobial effects demonstrated for many of these terpenoids suggest a native role in plant defense[61–65]. Carnosic acid is another abietane diterpenoid found in several Lamiaceae species with powerful antioxidant and anticancer properties[66]. The biosynthesis of carnosic acid and related diterpenoids has been elucidated in *Rosmarinus officinalis, Salvia pomifera* and *Salvia fruticosa*

(rosemary and sages)[67,68] and involves many CYPs orthologous to those involved in tanshinone biosynthesis (Supplementary Fig. 1).

Previous studies of the *S. miltiorrhiza* genome have found two BGCs that together contain the genes encoding miltiradiene diTPSs and two CYP76AHs involved in tanshinone biosynthesis[24,51,52]. A third locus containing an array of CYP71Ds includes the two genes for the enzymes (CYP71D375 and CYP71D373) responsible for the D-ring heterocycle of the tanshinones. Recent publication of additional Lamiaceae genomes revealed syntenic BGCs in four other species: *Tectona grandis, Salvia splendens* and *Scutellaria baicalensis* (teak, scarlet sage and Chinese skullcap, respectively)[24,59,69]. Additionally, we previously reported the presence of a large cluster in *Callicarpa americana* (American beautyberry) which contains orthologs of the miltiradiene diTPS genes as well as those encoding multiple CYP76AHs and CYP71Ds[70]. The divergence of these five species indicates that this BGC may be present ubiquitously throughout the Lamiaceae.

In this work, to explore the prevalence and evolution of the miltiradiene BGC, we survey a representative panel of 10 Lamiaceae genome assemblies (Fig. 3.1). We focus on synteny with the BGC in *C. americana*, which is one of the largest yet discovered, spanning approximately 400 Kb and encompassing seven diTPSs and twelve CYPs. Our syntenic analysis shows conservation of core miltiradiene biosynthetic genes throughout all species studied while highlighting lineage-specific diversification of the BGC in five subfamilies. Phylogenetic analysis supports common ancestry of each enzyme class and enables reconstruction of a minimal ancestral cluster. We find that the BGC in *C. americana* has evolved bifunctionality, providing the scaffold of the formerly unidentified diterpene (+)-kaurene in addition to miltiradiene. This opens biosynthetic avenues towards previously inaccessible diterpenes in addition to highlighting an instance of BGC

bifunctionality, which is rarely observed in plants[10,71]. We also discover complex miltiradiene BGCs in four additional species, laying the foundation for the elucidation of previously unknown diterpenoid pathways. Comparing the evolutionary trajectory of a BGC across a plant family illustrates how genomic organization can serve as a basis for expanding metabolic diversity.



Species	Common Name	Subfamily (Tribe)	Publication	Genome Size (as reported)
Hyssopus officinalis	Hyssop	Nepetoideae (Nepetinae)	Lichman *et.al.*, 2020	341 Mb
Prunella vulgaris	Common self-heal	Nepetoideae (Prunellinae)	**This work**	820 Mb
Rosmarinus officinalis	Rosemary	Nepetoideae (Salviinae)	Bornowski *et. al.*, 2020	1.1 Gb
Salvia miltiorrhiza	Danshen	Nepetoideae (Salviinae)	Song *et. al.*, 2020 Ma *et. al.*, 2021	600 Mb
Plectranthus barbatus	Coleus	Nepetoideae (Ocimeae)	**This work**	1.25 Gb
Leonotis leonurus	Lion's tail	Lamioideae	**This work**	585 Mb (4x)
Pogostemon cablin	Pachouli	Lamioideae	He *et.al.*, 2018	2.1 Gb (8x)
Scutellaria baicalensis	Chinese skullcap	Scutellariodeae	Xu *et. al.*, 2020	440 Mb
Tectona grandis	Teak	Tectonoideae	Zhao *et. al.*, 2019	325 Mb
Callicarpa americana	American beautyberry	Callicarpoideae	Hamilton *et. al.*, 2020	538 Mb
Erythranthe lutea	Monkey flower	Phrymaceae, Lamiales outgroup	Cooley *et. al.*, preprint	600 Mb (4x)

Figure 3.1. Species and genome assemblies used in this study. The cladogram shows evolutionary relationships between the species studied. Numbers at the nodes represent estimations of clade age in millions of years (MYA)[79–81]. Ploidy level of species not assumed to be diploid are shown in parenthesis next to their genome size (Supplementary Table 2).

Results

Genome assembly and annotation of L. leonurus, P. barbatus, *and* P. vulgaris

To increase the diversity of representatives across the Lamiaceae family, we sequenced three additional genomes, *Leonotis leonurus*, *Plectranthus barbatus*, and *Prunella vulgaris*, using the 10x Genomics linked read approach. High molecular weight DNA was isolated, 10x Genomics

libraries constructed and Supernova was used to assemble the genomes generating pseudohaplotype assemblies; pseudohaplotype-1 was selected for downstream analyses resulting in 585 Mb (*L. leonurus*), 1.25 Gb (*P. barbatus*), and 820 Mb (*P. vulgaris*) assemblies (Table 3.1). For *P. barbatus* and *P. vulgaris*, the assembled genome size is consistent with the estimations of genome size from both flow cytometry, 1.53 Gb and 786 Mb, respectively, as well as from a k-mer-based estimation from Supernova, 1.29 Gb and 871 Mb, respectively (Supplementary Table 2). However, for *L. leonurus*, there was a discrepancy in genome size estimation between flow cytometry (1042 Mb), k-mers (688 Mb) and genome assembly (585 Mb). Coupled with the large distance between heterozygous SNPs in *L. leonurus* outputted from Supernova (16.9 Kb), it is most likely that *L. leonurus* is an autotetraploid and the Supernova assembly is representative of all homologous chromosomes.

Table 3.1. Statistics for the 10x Genomics assemblies of *Leonitis leonurus, Plectranthus barbatus,* and *Prunella vulgaris.*

	Number of scaffolds	**Total size of scaffolds (bp)**	**N50 scaffold length (bp)**	**Number of Ns (Percent Ns)**	**Totals haps (consecutive Ns)**	**Largest scaffold (bp)**
Leonotis leonurus	23,651	585,264,293	1,094,942	40,883,810 (7.0%)	15,483	11,593,990
Plectranthus barbatus	62,959	1,249,907,925	258,138	70,313,430 (5.6%)	30,507	3,093,914
Prunella vulgaris	46,736	820,275,670	444,240	38,970,920 (4.8%)	20,293	5,268,047

Benchmarking Universal Single-Copy Ortholog (BUSCO)[72] of pseudohaplotype-1 assemblies revealed >97% complete BUSCOs in the three genomes (Table 3.2) with 18.5% and 13.4% duplicated BUSCOs present in *L. leonurus* and *P. barbatus*, respectively, suggesting of retained haplotigs in pseudohaplotype-1. Annotation of protein-coding genes with the unmasked genome using Lamiaceae-trained AUGUSTUS[73] matrices yielded 148,846 (*L. leonurus*), 413,222 (*P.*

barbatus), and 229,613 (*P. vulgaris*) genes (Supplementary Table 3). Assessment of the completeness of the annotation using BUSCO with the predicted proteomes revealed 94.4% (*L. leonurus*), 92.2% (*P. barbatus*) and 91.2% (*P. vulgaris*) complete BUSCO orthologs, suggesting that the annotation provided a robust gene set. A total of 57.9% (*L. leonurus*), 74.4% (*P. barbatus*), and 68.3% (*P. vulgaris*) of the genomes were repetitive with retroelements rather than DNA transposons dominating the genome space (Supplementary Table 4).

Table 3.2. Benchmarking Universal Single Copy Orthologs (BUSCOs) for *Leonotis leonurus*, *Plectranthus barbatus*, and *Prunella vulgaris* pseudohaplotype-1 genomes and predicted proteomes.

	Species	**Complete BUSCOs (C)**	**Complete single-copy BUSCOs (S)**	**Complete duplicate BUSCOs (D)**	**Fragmented BUSCOs (F)**	**Missing BUSCOs (M)**
Genome	*Leonotis leonurus*	98.5%	80.0%	18.5%	0.5%	1.0%
	Plectranthus barbatus	97.8%	84.4%	13.4%	1.0%	1.2%
	Prunella vulgaris	97.1%	91.8%	5.3%	1.5%	1.4%
Predicted proteome	*Leonotis leonurus*	94.4%	79.6%	14.8%	4.2%	1.4%
	Plectranthus barbatus	92.2%	80.7%	11.5%	5.4%	2.4%
	Prunella vulgaris	91.2%	86.8%	4.4%	6.1%	2.7%

Syntenic analysis reveals ubiquity of the miltiradiene biosynthetic gene cluster

C. americana provided a unique opportunity to investigate the evolution of a family-wide diterpenoid BGC since it is in a sister lineage to the rest of the Lamiaceae and has a large, dense BGC. We analyzed nine Lamiaceae genomes against our anchor species, *C. americana,* to determine synteny with its miltiradiene BGC. We chose our genome panel based on their

assembly quality and contiguity as well as subfamily representation (i.e., phylogenetic placement). We chose three species with previously reported syntenic BGCs and available genomes (*S. miltiorrhiza*[24,52]*, T. grandis*[69]*,* and *S. baicalensis*[75]), the three species we assembled in this study (*L. leonurus, P. barbatus,* and *P. vulgaris),* and three species with published genomes (*Hyssopus officinalis*[76]*, R. officinalis*[77]*,* and *Pogostemon cablin*[78]). In total, these ten species represent five of the twelve currently recognized subfamilies with a most recent common ancestor estimated at 60-70 million years ago (Fig. 3.1)[79–81]. As a close Lamiales outgroup, we also analyzed *Erythranthe lutea* (Monkey flower; formerly *Mimulus luteus*)[82].

Out of the 10 Lamiaceae species sampled, all contained diTPS genes orthologous to known (+)-CPP and miltiradiene synthases. In seven species these diTPSs were within syntenic BGCs (Fig. 3.2). The genomes of *P. vulgaris, P. barbatus,* and *R. officinalis* were too fragmented to determine whether they were part of a larger cluster. Four of the BGCs in this analysis have not been previously reported, showing that this cluster is even more conserved than originally described. All BGCs except that in *S. baicalensis* contain multiple CYP76AH genes. Five species, *C. americana*, *T. grandis, S. miltiorrhiza, H. officinalis*, and *L. leonurus,* also had at least one copy of a CYP71D gene.

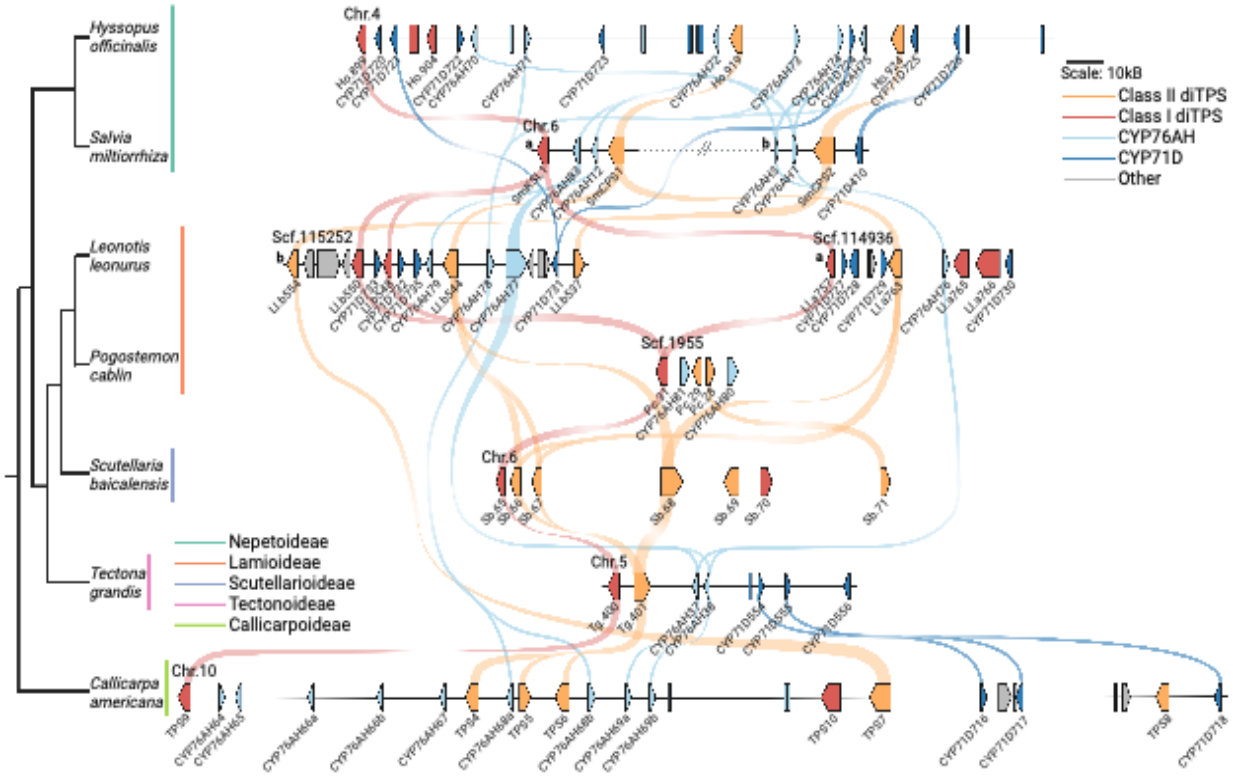


Figure 3.2. Syntenic relationships of a miltiradiene biosynthetic gene cluster present across the Lamiaceae. Genes are represented with arrows and pseudogenes are represented with boxes. A core set of genes are common to many species examined, including a diTPS class II (+)-CPP synthase, a diTPS class I miltiradiene synthase, and CYP450s in the 76AH and 71D subfamilies. Notably, there is divergence in gene number, cluster length, and unique genes, indicating lineage-specific evolution. Synteny between each species is shown here with colored curves. Species tree adapted from Mint Evolutionary Genomics Consortium 2018[126]. Figure created using BioRender.com.

Comparison of the BGCs provides insight into the formation and maintenance of this cluster in divergent lineages (Fig. 3.2). The *S. baicalensis* BGC uniquely contains no CYPs but appears to have tandem duplications of a class II diTPS and an additional non-syntenic class I diTPS. Non-syntenic diTPS and CYP genes are present in most of the BGCs, pointing toward dynamic assembly and independent refinement in each species. There are also several diTPS and CYP pseudogenes. Interestingly, there are few interrupting genes in these BGCs. The *H. officinalis* and *C. americana* BGCs encompass large genomic regions with more intergenic space, while others such as *P. cablin*

and *L. leonurus* are compact and gene dense. We speculate that the presence of two clusters in *L. leonurus* is due to its tetraploidy and is not a true duplication. Similarly, octoploid *P. cablin* showed some evidence of multiple copies of the BGC (Supplementary Fig. 2). It is evident that each BGC, while maintaining the core miltiradiene genes, has undergone some lineage-specific independent evolution.

Phylogenetic evidence of an ancestral miltiradiene cluster in Lamiaceae

To better understand evolution of genes from each BGC, we estimated phylogenetic relationships for each enzyme subfamily in the BGCs along with a set of functionally characterized reference genes from Lamiaceae, except in the CYP71D clade where few characterized Lamiaceae sequences are available (Fig. 3.3). Consistent with other angiosperm labdane-type diTPSs, those diTPSs with class II function cluster in the TPS-c subfamily while those with class I function cluster in the TPS-e subfamily.

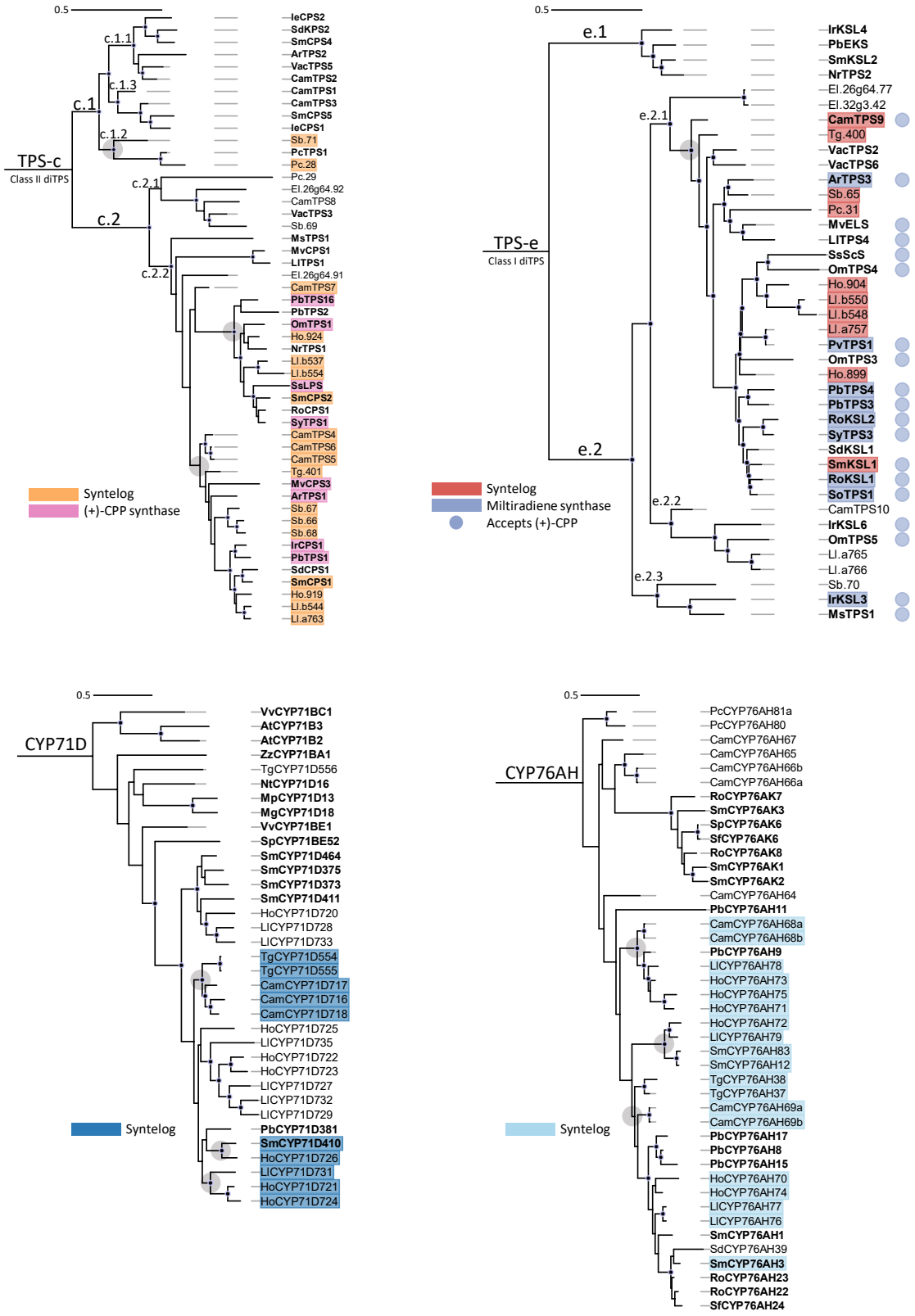


Figure 3.3. Phylogenetic evidence shows the relatedness of each gene class in the clusters.

Figure 3.3 (cont'd).
Enzymes present in each cluster with syntenic support from MCScanX and sequence identity from BLASTp are highlighted in red (TPS-e), orange (TPS-c), light blue (CYP76AH), and dark blue (CYP71D). DiTPSs characterized in previous reports are highlighted in pink and periwinkle ((+)-CPP synthases for TPS-c and miltiradiene synthases for TPS-e, respectively). Reference enzymes are bolded. Black solid dots at the base of the nodes represent 80% bootstrap confidence. Gray circles around clade nodes represent hypothetical expansion points for syntelogs and share approximately 70% or more sequence similarity. DiTPS trees are rooted to *Physcometrium patens ent*-kaurene synthase (PpEKS), and CYP trees are rooted to *Arabidopsis thaliana* AtCYP701A3.

As expected, syntenic diTPSs in both subfamilies have common ancestry. Recent tandem duplications in the TPS-c family are evident in *C. americana* and *S. baicalensis* and contribute to lineage-specific BGC expansion (Fig. 3.3, Fig. 3.4). The phylogenies also highlight the more distant origins of several non-syntenic diTPSs. The presence of divergent class I and II sequences points to independent acquisition as part of the diversification that occurred during speciation. Close inspection of phylogenetic relationships with characterized diTPSs can offer clues to likely functions. All class II diTPSs syntenic to *CamTPS6* phylogenetically cluster in clade TPS-c.2.2, which contains all known Lamiaceae (+)-CPP synthases as well as some diTPS enzymes which yield labdanes in the (+)-configuration. The two divergent class II enzyme sequences, Sb.71 and Pc.28, cluster in TPS-c.1 which produces compounds in the *ent*- rather than (+)-configuration, so it is likely that these two enzymes follow suit.

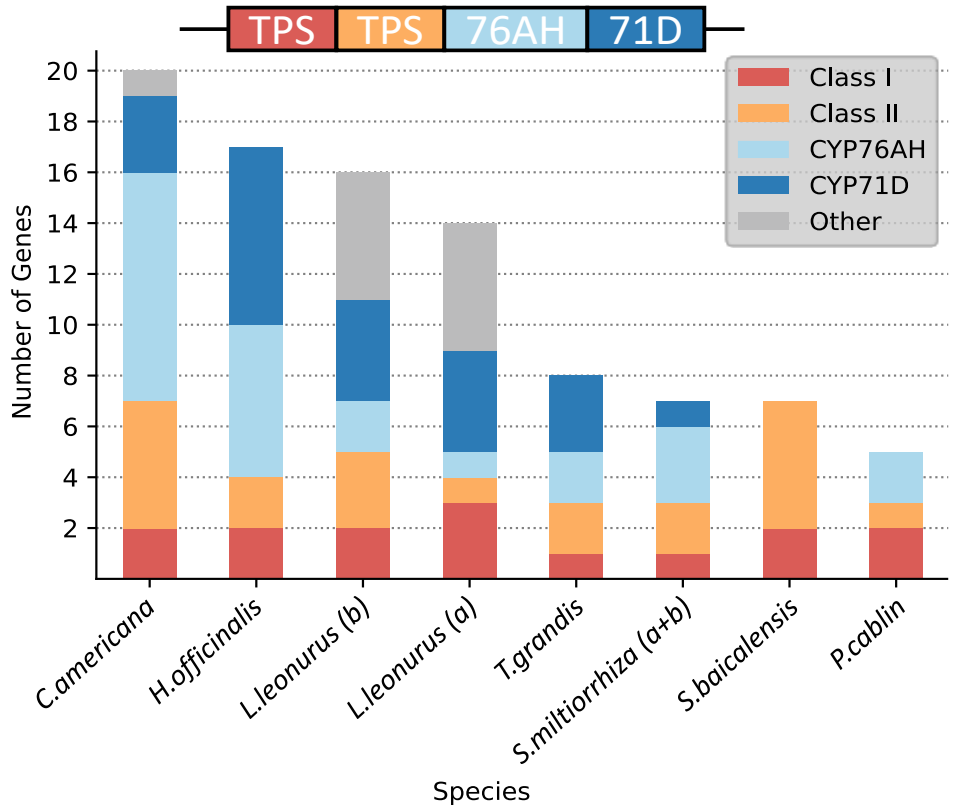


Figure 3.4. Predicted Lamiaceae minimal ancestral BGC and species-specific expansion of each. Based on maximum parsimony, we suggest that a cluster containing a class II diTPS, class I diTPS, CYP76AH, and CYP71D gene was formed in an early Lamiaceae ancestor. Lineage-specific expansion and refinement are evident from the number and composition of genes in each gene family present in the extant species studied.

None of the class I enzymes encoded in the BGCs clustered in clade TPS-e.1, consistent with their expected role in specialized metabolism. The TPS-e.1 clade primarily contains enzymes that convert *ent*-CPP to the gibberellin intermediate *ent*-kaurene. All BGC class I diTPSs cluster in TPS-e.2, which mostly encodes enzymes that accept (+)-CPP as a substrate. The presence of a presumed (+)-CPP synthase encoded in every BGC supports the likelihood that these class I diTPSs can all utilize (+)-CPP. Genes syntenic with *CamTPS9* are grouped in clade TPS-e.2.1, which contains all but one of the Lamiaceae sequences encoding enzymes known to catalyze formation of miltiradiene. Notably, every BGC contains at least one of these presumed miltiradiene

synthase sequences. Also characteristic of the TPS-e.2.1 clade is the loss of the internal γ domain, which is retained in most diTPSs but lost in mono- and sesqui-TPSs. The three non-syntenic enzyme sequences are split between clades TPS-e.2.2 and TPS-e.2.3, which encompass only a few characterized sequences with unique functions. The functional heterogeneity of these clades makes it difficult to draw conclusions as to the likely function of these BGC encoded enzymes but does offer intriguing possibilities for discovery of novel terpene backbones.

While phylogenetic classification is not a perfect predictor of TPS function[37,83], previous work has demonstrated a high level of clade specific consistency that allows us to draw tentative conclusions about the function of the BGC diTPSs[49]. Phylogenetic evidence supports that these BGCs likely have at minimum a (+)-CPP synthase and a miltiradiene synthase, enabling production of miltiradiene in each plant (Fig. 3.3). Moreover, several BGCs contain diTPSs from clades that may offer distinctive chemistries.

CYPs in the 76AH enzyme subfamily exhibit close phylogenetic clustering across the species analyzed. Several functionally characterized CYP76AHs have been found to oxidize miltiradiene in critical steps towards tanshinone and carnosic acid biosynthesis[55,56]. Although we were unable to identify a BGC in *R. officinalis* due to a fragmented assembly, the close relationship between the RoCYP76AH enzymes and those the other BGCs supports common ancestry. Nearly all CYP76AHs in the BGCs have paralogs within each cluster, highlighting the role of tandem duplication in expanding this subfamily[47,84]. However, there are several BGC CYP76AHs that are highly divergent from the syntelogs. The *C. americana* enzymes CYP76AH65, CYP76AH66, and CYP76AH67 are phylogenetically distinct, showing only 50-60% sequence similarity to other BGC

CYP76AHs. These enzymes are more related to the clade of CYP76AKs, which have not been found in this BGC but are part of the tanshinone and carnosic acid oxidation networks.

CYPs in the 71D subfamily similarly show phylogenetic clustering with others in the BGCs. Three CYP71D sequences from *H. officinalis* and *L. leonurus* are in the same clade as the CYP71D gene array from *S. miltiorrhiza*, which was implicated in furan ring formation for the tanshinones[24]. SmCYP71D410 is a previously unrecognized member of the BGC Sm-b that phylogenetically clusters with HoCYP71D724 and PbCYP71D381 enzymes. PbCYP71D381 can oxidize the forskolin precursor (13*R*) manoyl oxide, a close structural relative of miltiradiene[85]. One enzyme from *T. grandis* stands out as much less related than the rest, with only 40-50% sequence similarity to other BGC CYP71Ds. This enzyme is likely another recent independent acquisition, although it is the only one observed in the CYP71D subfamily. All BGCs containing CYP71D genes also have at least one duplication, once again highlighting the importance of duplication in the diversification of these pathways[86].

Close phylogenetic clustering of most enzymes in all four subfamilies provides compelling evidence for a common ancestral origin and subsequent lineage-specific duplications. We analyzed presence/absence of syntelogs and proposed a model for a minimal cluster using ancestral state reconstruction (Fig. 3.4; Supplementary Figs. 3, 4). High levels of sequence conservation between syntelogs supports a minimal ancestral cluster that contains genes encoding a (+)-CPP synthase, a miltiradiene synthase, a CYP76AH, and a CYP71D. The dynamic nature of this BGC over millions of years of evolution is evident through the gene loss, presence of pseudogenes, and addition of non-syntenic genes observed in these extant Lamiaceae. Despite these differences, the high degree of conservation of the ancestral cluster is notable.

Since the miltiradiene BGC was present in nearly every Lamiaceae species sampled, we also investigated the synteny in *E. lutea*, a closely related Lamiales outgroup[79,82,87]. We found a region syntenic to the *C. americana* BGC which contains class II and class I diTPSs but no CYPs (Supplementary Fig. 5). The genes encoding class II enzymes, *El.26g64.91* and *El.26g64.92*, are in clade TPS-c.2, showing some similarity with other (+)-CPP synthases (Fig. 3.3). The class I sequence, *El.26g64.77*, is within TPS-e.2.1, but distinct from the rest of the clade and surprisingly retains the γ domain. This domain loss has occurred multiple times in the evolution of plant TPSs[88], so it is conceivable that the class I enzymes in *E. lutea* represent the three-domain miltiradiene synthase shared by the most recent common ancestor in the Lamiales. While the *E. lutea* partial cluster may provide a glimpse into an ancestral state of the Lamiaceae BGC, a more widespread examination of additional Lamiales genomes would be an interesting avenue for future work and could more firmly establish the timeline of gene acquisition and loss in this cluster.

Functional characterization of the* C. americana *BGC reveals two metabolic modules and a previously inaccessible terpene backbone

Though increasing numbers of computationally predicted BGCs have been identified in plants, only a few are functionally characterized. So far, coregulation has proven to be a greater predictor of functional relationship in BGCs than colocalization alone[89]. Previous analysis of the two BGCs in *S. miltiorrhiza*, Sm-a and Sm-b, found that each had divided expression between root and aerial tissues. The diTPSs from Sm-a and *CYP76AH*s from Sm-b were expressed exclusively in root tissues and found to be vital steps in the root tanshinone biosynthetic pathway[51]. Additionally, an array of root-specific CYP71D encoded enzymes were also integral to tanshinone

biosynthesis but located elsewhere in the genome[24]. Another example where differentially expressed diTPSs and CYPs were reported in distinct specialized metabolite pathways despite being colocalized is the bifunctional gene clusters of phytocassanes/oryzalides found in *Oryza sativa* (rice)[71] and the noscapine/morphinan biosynthesis in *Papaver ssp.* (poppy)[11,58]. Divergence in expression may be one way in which plants exploit some of the benefits of genomic organization while creating unique pathways based on regulation.

Given the unprecedented size and complexity of the BGC identified in *C. americana*, we sought to investigate whether it is a metabolically unified BGC. We first analyzed RNA expression in 8 tissue types to determine the expression pattern of the BGC (Fig. 3.5; Supplementary Fig. 6)[70]. This revealed a clear divergence between the first and second halves of this BGC. The first half is preferentially expressed in fruit and root tissue and contains a (+)-CPP synthase (*CamTPS6*)[70], the predicted miltiradiene synthase (*CamTPS9*), and several CYP76AHs. The second half is more strongly expressed in flower and young leaf tissues and contains a non-orthologous class I diTPS (*CamTPS10*), another predicted (+)-CPP synthase (*CamTPS7*), and two CYP71Ds as well as partial fragments of a CYP76AH (*Ca.26-27*). The presence of a diTPS class II/class I pair as well as CYPs in each module suggests that this BGC may have evolved divergent diterpenoid pathways. Additionally, we investigated expression of each of BGC in the other species with published transcriptomic data but found no overarching expression trends, unlike in *C. americana* (Supplementary Fig. 6).

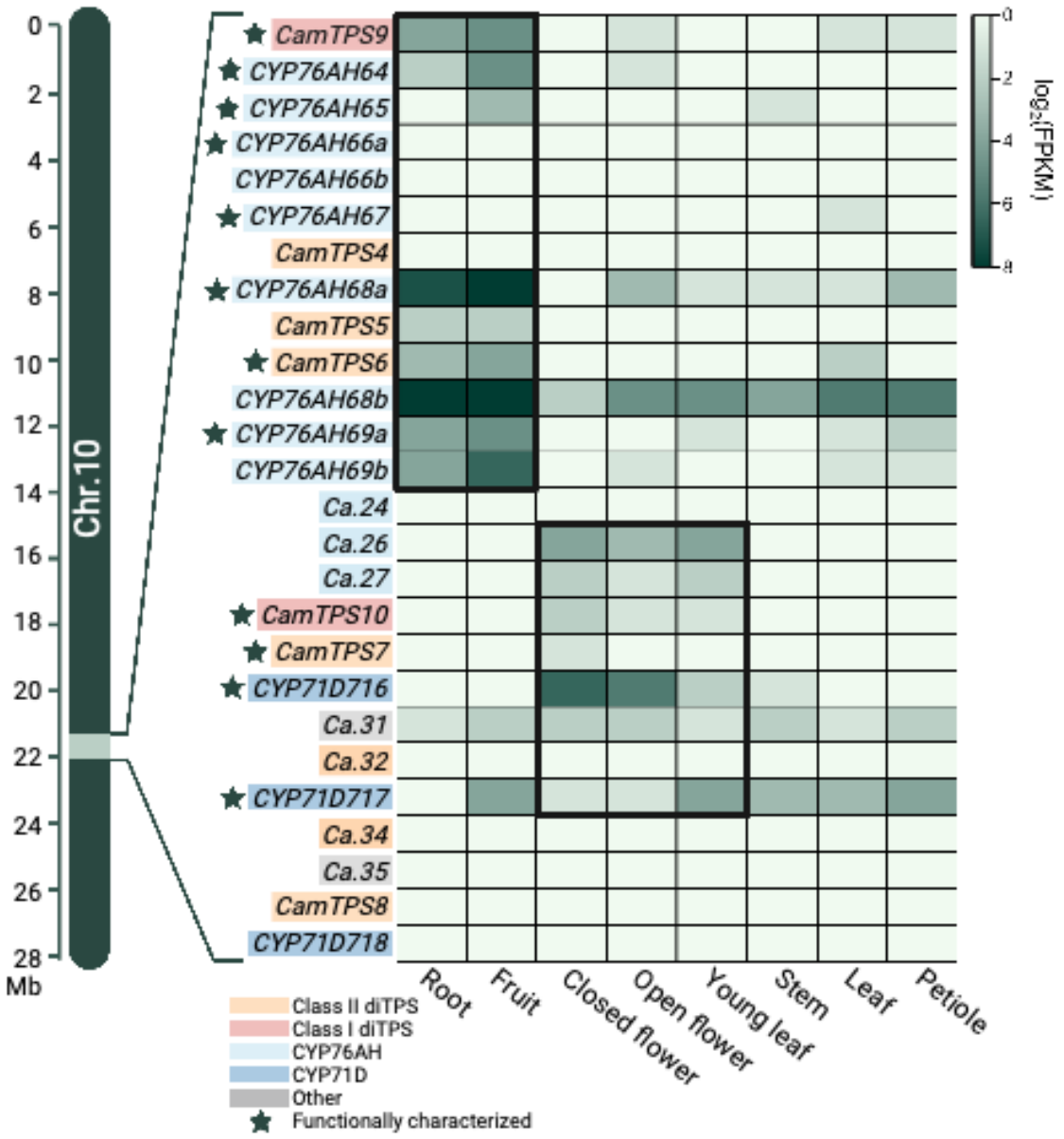


Figure 3.5. Tissue specific expression of a miltiradiene BGC in *C. americana* obtained from RNA sequencing. Functional characterization of these enzymes refers to this study. This figure represents Chr10:21.92-22.33 Mb. Approximate location on the chromosome is indicated. Two differentially expressed metabolic clusters are boxed to highlight similar expression patterns. Colors indicate diTPS, CYP, or unrelated gene family, including pseudogenes (unnamed). Data obtained from Hamilton *et al*, 2020[70].

We investigated enzyme activity for the following members of the *C. americana* cluster: *CamTPS7, CamTPS8, CamTPS9, CamTPS10, CamCYP76AH64, CamCYP76AH65, CamCYP76AH67, CamCYP76AH68, CamCYP76AH69, CamCYP71D716,* and *CamCYP71D717*. Combinations of all genes were transiently expressed in *Nicotiana benthamiana* to evaluate enzyme function. DiTPS functions were determined by comparison of mass spectra and retention time by GC-MS with published diTPS activities or using NMR for previously unpublished activity (Fig 3.6). CamTPS7 was confirmed to be a (+)-CPP synthase (Supplementary Fig. 6). CamTPS9 is a miltiradiene (**1**) synthase, with some abietatriene (**2**; *ent*-abieta-8,11,13-triene) resulting from spontaneous aromatization *in plantae* consistent with previous observations[90]. CamTPS10, when paired with a (+)-CPP synthase, forms (+)-kaurene (**4**), a previously unknown diTPS activity (Supplementary Fig. 8-10). The biological relevance of this activity is supported by the structure of the diterpenoid calliterpenone, which is derived from the (+)-kaurene backbone and has been documented in multiple *Callicarpa* species[91]. Calliterpenone has been investigated for its potential as a plant growth promoting agent[92], and thus represents an interesting biosynthetic target. Discovery of this (+)-kaurene synthase will enable biosynthetic access to this group of metabolites as well as to non-natural diterpenoids that may have useful bioactivities[93]. The physical grouping and similar expression patterns of *CamTPS10* and *CamTPS7* supports that this cluster has diverged into two metabolically distinct modules through the duplication of a (+)-CPP synthase, the recruitment of an additional class I diTPS, and a shift in tissue-specific gene expression.

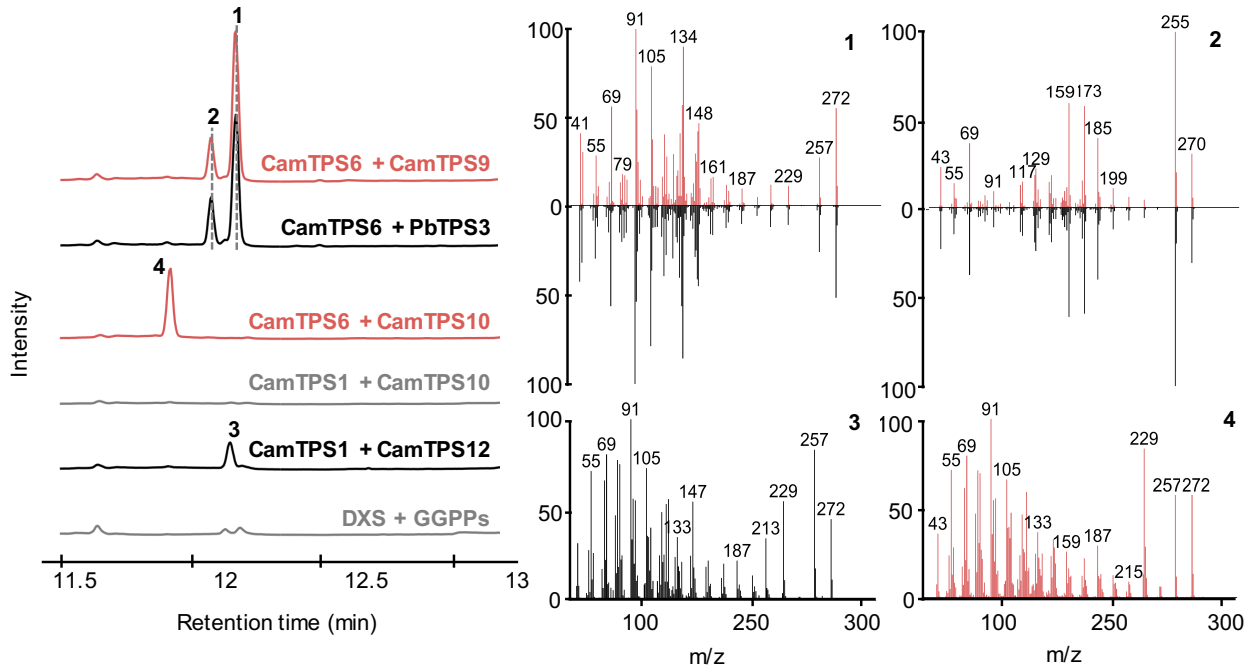


Figure 3.6. GC-MS analysis of *C. americana* BGC diTPS products. CamTPS9 was confirmed to be a miltiradiene synthase by comparison with the retention time and mass spectra of PbTPS3[127–130] products when both were expressed with the (+)-CPP synthase CamTPS6[70], forming miltiradiene (**1**) and abietatriene (**2**). CamTPS10 was found to make **4** from (+)-CPP but not *ent*-CPP (CamTPS1)[70]. This product has a different retention time but similar mass spectrum to *ent*-kaurene (**3**), made by the combination of CamTPS1 and CamTPS 12 (Supplementary Fig. 11). All chromatograms shown are total ion chromatograms. Red and black traces correspond to combinations yielding **1**, **2**, **3** and **4** respectively, as indicated in the mass spectra. Each combination includes *P. barbatus* 1-deoxy-D-xylulose-5-phosphate synthase (*DXS*) and GGPP synthase (*GGPPS*), shown as a control in gray.

After establishing routes to the formation of the *C. americana* diterpene backbones, we tested each CYP against all possible diterpene intermediates found in this plant (Fig. 3.7): *ent*-kaurene (CamTPS12; Supplementary Fig. 11) and kolavenol[70] formed by diTPSs outside the cluster, and (+)-kaurene and miltiradiene from the BGC. No activity was detected with kolavenol or *ent*-kaurene. With miltiradiene, CamCYP76AH67 formed six different oxidation products (**1a-d**, **2a-b**, Fig. 3.7a). Based on m/z of the molecular ions and comparison of mass spectra with each other and the NIST database, two match oxidations of abietatriene and the other four of miltiradiene

(Supplementary Fig. 12). Most of these products proved difficult to separate by column chromatography, preventing complete structural elucidation. However, we were able to purify **2a**, and NMR experiments support the assignment as 15-hydroxy-*ent*-abieta-8,11,13-triene (Supplementary Fig. 13-15). Oxidation in this position on an abietane diterpene has only been reported twice before: by a 2-oxoglutarate dehydrogenase in *S. miltiorrhiza*[94] and by CYP81AM1 in *Tripterygium wilfordii*[95]. CamCYP76AH68 also showed activity with miltiradiene, dramatically shifting the product profile towards abietatriene and affording a small amount of oxidized abietatriene (**2c**; Supplementary Fig. 12). This indicates that CamCYP76AH68 may be hydroxylating the c-ring of miltiradiene, which then undergoes water loss to form abietatriene more readily than the spontaneous aromatization of miltiradiene alone (Fig. 3.8a). In previous work characterizing enzymes involved in tanshinone and carnosic acid biosynthesis, the ferruginol synthases showed a preference for abietatriene, but enzymatic conversion of miltiradiene to abietatriene was not observed. It was suggested that the aromatization is spontaneous and possibly driven by sunlight[90]. The discovery of CamCYP76AH68 indicates that at least in *C. americana* an enzyme may assist in the conversion of miltiradiene to abietatriene. When we expressed each CYP with *CamTPS6* and *CamTPS10* to evaluate CYP activity with the (+)-kaurene backbone, we observed a peak with expression of *CamCYP71D717*. Upon further investigation, however, we realized this enzyme apparently catalyzes formation of (+)-manool (**6**) from (+)-copalol (**5**), the dephosphorylation product of (+)-CPP (Fig. 3.7b, Supplementary Fig. 16). Each CYP/TPS enzyme combination that resulted in observable products was then expressed in combination with all other CYPs. *CamCYP76AH67* combined with *CamCYP76AH68* and miltiradiene yielded at least one previously undetected oxidized compound (**2d**, Fig. 3.7a; Supplementary Fig. 12). The combination

of *CamTPS6* with *CamCYP71D716* and *CamCYP71D717* resulted in full conversion of (+)-manool (**6**) to 3(*S*)-hydroxy-(+)-manool (**7**), which was confirmed by NMR (Fig. 3.7b, Fig. 3.8b; Supplementary Figs. 17-19).

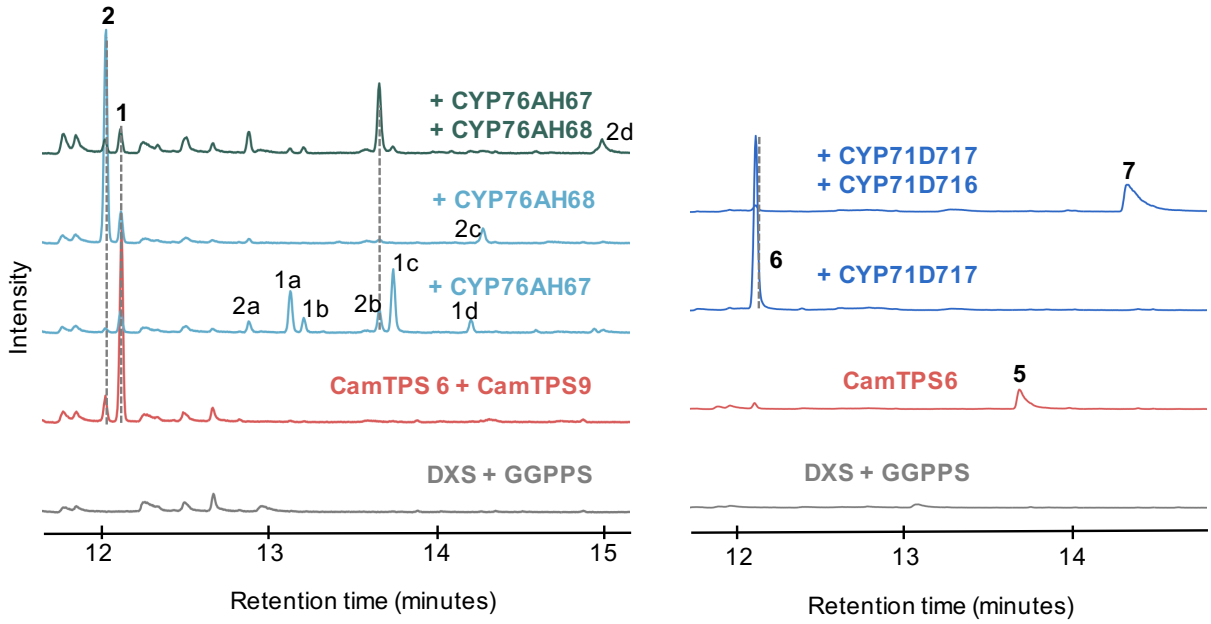


Figure 3.7. GC-MS chromatograms showing oxidation products of *C. americana* BGC CYPs. a Oxidation products of the CamCYP76AHs from **1** and **2**, assigned based on analysis of mass spectra (Supplementary Fig. 12). **b** CamCYP71D717 catalyzes the production of (+)-manool (**6**)**,** likely from (+)-copalol (**5**) (Supplementary Fig. 16) and the addition of CamCYP71D716 results in 3(*S*)-hydroxy-(+)-manool (**7**). Each combination includes *P. barbatus* 1-deoxy-D-xylulose-5-phosphate synthase (*DXS*) and GGPP synthase (*GGPPS*), shown as a control in gray. CamTPS6 and CamTPS6 + CamTPS9 controls given in red.

To the best of our knowledge, no abietane-type diterpenoids were previously reported in *C. americana,* which has been primarily studied for clerodane diterpenoids produced in leaves[96–98]. However, other *Callicarpa* species, including *C. bodinieri* and *C. macrophylla*[99], produce a wide variety of medicinally relevant abietane diterpenoids (Fig. 3.8c), indicating that the abietane skeleton is a key intermediate for at least some plants in this genus[65,99]. We analyzed a whole root extract of *C. americana* by GC-MS and found compounds with matching retention time and

mass spectra to abietatriene and the oxidized product (**2c**) produced by CYP76AH68. This supports the biological relevance of enzyme activities elucidated in *N. benthamiana* (Supplementary Fig. 20).

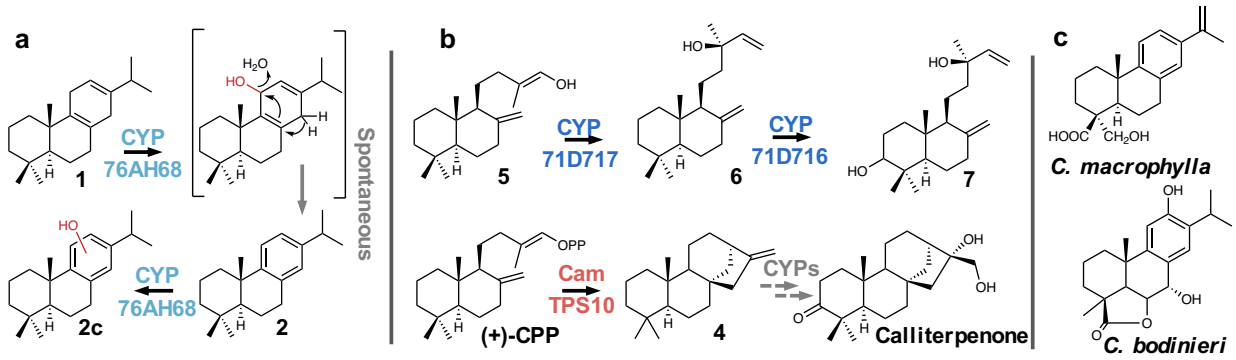


Figure 3.8. Pathway schematic for CYP oxidations in *C. americana*. (a) Proposed mechanism for enzyme-assisted conversion of **1** to **2**, followed by an additional oxidation of **2** to form **2c**. Mass spectra supports assignment of the hydroxy group in **2c** to the c-ring (Supplementary Fig. 12, Supplementary Data 3). (b) Proposed conversion of **5** to **6** by CamCYP71D717, and oxidation of **6** by CamCYP71D716. This occurs in the same position as a keto group on calliterpenone, which is derived from **4**. (c) Structures of abietane diterpenoids found in two other species of *Callicarpa*.

C. americana contains over 600 predicted CYPs, and it is likely that the BGC CYPs are part of a larger metabolic network with peripheral modifying enzymes elsewhere within the genome[70]. However, the functional activities we report here validate the biological significance of the BGC and its divergent modules. The CYPs showed a marked preference for the (+)-copalol and miltiradiene backbones over other diterpenes present in the plant. Within the two modules, the miltiradiene and (+)-kaurene synthases were differentially expressed along with their respective (+)-CPP synthases. The CYP76AHs were more active towards miltiradiene, whereas the CYP71Ds utilized (+)-copalol. Functionalization of (+)-kaurene may require oxidations catalyzed by non-clustered enzymes.

Discussion

In this study we found that the miltiradiene BGC, previously identified in only a few species, is present across five divergent Lamiaceae subfamilies. The preserved enzyme sequences and gene order in the cluster provide strong evidence for an ancestral cluster in an early Lamiaceae ancestor. From this core cluster, these species have retained the diTPSs necessary to form the signature miltiradiene backbone but tailored their chemical diversity through gene duplication, sequence divergence, gene acquisition, and gene loss. We can speculate that the metabolic products from the ancestral cluster have diversified as the Lamiaceae family diverged and populations adapted to new environments. Gene duplication appears to be a major driver of the evolution and expansion of the vast diversity of TPSs and CYPs in plants[2,41,100,101], and the Lamiaceae miltiradiene cluster exemplifies this. This is notable in the *C. americana* cluster where tandem duplication has generated five sequential, highly similar CYP76AH genes. However, every species examined had at least one apparent duplication event, supplying the material for evolution toward metabolic diversification. There is also a striking example of cluster expansion through the apparent recruitment of *CamTPS10* in *C. americana.* The discovery of the (+)-kaurene synthase showcases another example of a bifunctional BGC with divergent transcription patterns. The presence of phylogenetically distinct diTPSs in other miltiradiene BGCs discovered here similarly suggests multifunctionality.

Conservation of the miltiradiene backbone suggests strong selective pressures for retention in the Lamiaceae and beyond, as illustrated by the recently discovered clustered pair of diTPSs forming the same backbone in *Tripterygium wilfordii* in the distant Celastraceae[102]. Surprisingly little is known about how plants use abietane diterpenoids, but they are mostly thought to be

involved in pathogen responses due to their antibacterial activities[61,103]. However, abietanes have been extensively studied for their importance to human health. They exhibit a range of bioactivities from anti-tumor to antimicrobial to anti-inflammatory, among others[61–64,104]. Nearly 500 abietane diterpenoids have been reported to date in Lamiaceae species[40,105]. Earlier investigations of these diterpenoids in Lamiaceae have taken a metabolite-guided approach, which has yielded much progress towards the biosynthesis of tanshinones, carnosic acid, and related compounds. The findings of this study establish a framework for a genomics-guided investigation of additional abietane diterpenoids throughout the Lamiaceae. The functional characterization of part of the *C. americana* BGC as well as the root metabolite data support the presence of a miltiradiene diterpenoid network in this plant despite the lack of previously documented abietanes. Further characterization of the other identified miltiradiene BGCs in *H. officinalis, P. cablin,* and *L. leonurus* could similarly lead to the discovery of yet unknown chemistries.

A deeper understanding of the enzymatic activities encompassed by BGC genes will also help to elucidate how BGCs drive expansion of metabolic diversity. It is clear from the conservation of the miltiradiene BGC in at least five extant Lamiaceae subfamilies that gene colocalization is an important contributor to plant specialized metabolism. Genomic organization is also of special interest in synthetic biology, as understanding natural BGCs can provide a blueprint for the construction and control of synthetic clusters in heterologous systems[106]. This study presents one of the currently limited examples of a BGC present throughout an entire family. With the increasing quality and quantity of plant genomes available, future large-scale BGC investigations

may find that plants frequently rely on BGCs as a toolbox for adaptability through metabolic diversity.

Methods

Collinearity analysis

The BLAST function makeblastdb (E-value of $1e^{-10}$, 5 alignments)[107] was used to create protein databases between *C. americana* and each other species examined. Peptide sequences and genome annotation files were obtained through respective data repositories. Syntenic analysis between *C. americana* and every other species discussed was performed using the standard MCScanX pipeline (Match score = 50; Match size = 5; Gap penalty = -1; Overlap window = 5; E-value = $1e^{-5}$; Max gaps = 25)[108]. Results were visualized using SynVisio[109]. Orthologs and syntenic lines were manually curated using 70% sequence identity cutoff determined by the BLASTp alignment function (Threshold = 0.05, Word Size = 3, Matrix = BLOSUM62, Gap Costs = Existence:11 Extension:1).

Ancestral state reconstruction

Extant character states were collected into a single document coded as 1 for presence and 0 for absence of each gene. Ancestral state analysis was performed using the phytools R package (version 0.7-80)[110]. Evolutionary models were selected using information from the `fitMK()` function. Ancestral states were determined with the `ace()` function.

Phylogenetic trees

Sequences used in all protein phylogenies were obtained from annotated peptide sequences from their respective species. A list of reference sequences used can be found in Supplementary Data 1. CYP annotation was kindly provided by David Nelson (University of Tennessee). Full-

length protein coding sequences were used, however plastidial targeting sequences present in diTPSs were removed from alignments. Multiple sequence alignments were generated using ClustalOmega (version 1.2.4; default parameters)[111] and phylogenetic trees were generated by RAxML (version 8.2.12; Model = protgammaauto; Algorithm = a)[112] with support from 1000 bootstrap replicates. All alignments are available in our dryad repository (https://doi.org/10.5061/dryad.w9ghx3frg). The tree graphic was rendered using the Interactive Tree of Life (version 6.5.2)[113].

Genome sequencing, assembly, and annotation of three Lamiaceae species

High molecular weight DNA was isolated from mature leaves from *L. leonurus, P. barbatus,* and *P. vulgaris* and used to construct a 10x Genomics library using the Genome and Gel Bead Kit v2 (10x Genomics, Pleasanton, CA). Libraries were sequenced on an Illumina NovaSeq 6000 (Illumina, San Diego, CA) in paired end mode, 150 nt. Libraries were made and sequenced by the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign. The genomes were assembled using 10x Supernova (version 2.1.1)[114]. The script 'supernova run' was run with default settings except --maxreads was set to 360000000 (*P. vulgaris*), 531000000 (*P. barbatus*) or 297550000 (*L. leonurus*), which yielded the best results for genome contiguity and percent of estimated genome size after testing multiple coverage levels. To obtain fasta files, 'supernova mkoutput' was run with the parameters, '--style=pseudohap2' and '--headers=full'. Genes were predicted on the non-repeat-masked pseudohaplotype-1 assemblies using AUGUSTUS (version 3.3)[73] with the parameter, '--UTR=off', and the '--species' and 'c--extrinsicCfgFile' parameters to use training results from closely related species, *H. officinalis* (*P. barbatus*, *P. vulgaris)* or *T. grandis* (*L. leonurus*). Assembly statistics were calculated using the

tool assembly-stats (version 1.0.1)[115]. The AUGUSTUS default gene annotations were converted to GFF3 format using the gtf2gff.pl in the AUGUSTUS repository (version 3.4.0) and gene annotation metrics were generated using GAG (version 2.0.1)[116]. BUSCO (version 5.2.2)[72] was run in genome mode using the lineage dataset 'embryophyta_odb10.' To identify repetitive sequences in the three *de novo* assembled genomes, a custom repeat library (CRL) for each assembly was created with RepeatModeler (version 2.0.3)[117]. Protein-coding genes were removed from each CRL using ProtExcluder (version 1.2)[118] and RepBase Viridiplantae repeats from RepBase (version 20150807)[119] were added to create a final CRL. Each assembly was repeat masked with its corresponding CRL using RepeatMasker (version 4.1.2-p1)[120] using the parameters -e ncbi -s -nolow -no_is -gff.

Transcriptomic analysis

All transcriptomic datasets used in Fig. 3.5 and Supplementary Fig. 6 were downloaded from the SRA database. Raw reads were trimmed using fastp (version 0.23.2)[121], mapped to respective coding sequence files using Salmon 'index' (version 1.8.0)[122], and quantified using Salmon 'quant' (libtype=A, validate mappings). Genes specific to each respective cluster were parsed out to compare expression levels between tissues. Data was transformed by a factor of log2(X+1), where the quantified expression, X, had a value of 1 added to all genes in an unbiased fashion to account for occurrences of 0 expression and to remove negative log values due to lowly expressed genes, which would exaggerate differences between genes. The caveat to this transformation is lower expressed genes appear to have expression closer to 0 while more highly expressed genes are comparatively unaffected. Genes were clustered based on order of

appearance within the genome, while tissues were clustered based on similarity between tissue groups. Heatmaps were generated using ggplot2 (version 3.1.1)[123].

PCR and cloning

Synthetic oligonucleotides are given in Supplementary Table 5, GenBank accession numbers, and sequences of all enzymes characterized or discussed in this study are listed in the source data of Figs. 3.2 and 3.3. Candidate enzymes were PCR-amplified from root, fruit, leaf, and flower cDNA, and coding sequences were cloned and sequence-verified with respective gene models (Supplementary Table 6). Constructs were then cloned into the plant expression vector pEAQ-HT [124] and used in transient expression assays in *N. benthamiana*.

***Transient expression in* N. benthamiana**

Transient expression assays in *N. benthamiana* were carried out based on a published protocol[49]. *N. benthamiana* plants were grown for 5 weeks in a controlled growth room under 16 H light (24 °C) and 8 H dark (17 °C) cycle before infiltration. Constructs for co-expression were separately transformed into *Agrobacterium tumefaciens* strain LBA4404. Cultures were grown overnight at 30 °C in LB with 50 µg/mL kanamycin and 50 µg/mL rifampicin. Cultures were collected by centrifugation and washed twice with 10 mL water. Cells were resuspended and diluted to an OD_{600} of 1.0 in water with 200 µM acetosyringone and incubated at 30 °C for 1-2 H. Separate cultures were mixed in a 1:1 ratio for each combination of enzymes, and 4-5 week old plants were infiltrated with a 1 mL syringe into the underside (abaxial side) of *N. benthamiana* leaves. All gene constructs were co-infiltrated with two genes encoding rate-limiting steps in the upstream 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway*: P. barbatus* 1-deoxy-D-xylulose-5-phosphate synthase (*PbDXS*) and GGPP synthase (*PbGGPPS*) to boost production of the

diterpene precursor GGPP[93,125]. Plants were returned to the controlled growth room (76 °C, 12 H diurnal cycle) for 5 days. Approximately 200 mg fresh weight from infiltrated leaves was extracted with 1 mL hexane (diTPS products) or ethyl acetate (CYP products) overnight at 18 °C. Plant material was collected by centrifugation, and the organic phase was removed for GC-MS analysis. Each experiment was performed in triplicate. Data shown are from single experiments representative of the replicates.

Root metabolite extraction

The entire root system of a healthy 3 year old *C. americana* plant grown under greenhouse conditions was collected, washed, and blended with water to break down the tissue. The mixture was then combined with 500 mL ethyl acetate and allowed to extract for 24 H. The organic layer was then separated from the aqueous layer, filtered, concentrated via rotary evaporator, and stored at -20 °C. This extract was diluted 1:500 in ethyl acetate and analyzed by GC-MS. All GC-MS analyses were performed in Michigan State University's Mass Spectrometry and Metabolomics Core Facility on an Agilent 7890A GC with an Agilent VF-5ms column (30 m × 250 µm × 0.25 µm, with 10 m EZ-Guard) and an Agilent 5975C detector. The inlet was set to 250 °C splitless injection of 1 µL and He carrier gas (1 mL/min), and the detector was activated following a 3 min solvent delay. All assays and tissue analysis used the following method: temperature ramp start 40 °C, hold 1 min, 40 °C/min to 200 °C, hold 4.5 min, 20 °C/min to 240 °C, 10 °C/min to 280 °C, 40 °C/min to 320 °C, and hold 5 min. MS scan range was set to 40-400.

Product scale-up and NMR

For NMR analysis, production in the *N. benthamiana* system was scaled up to 1 L. A vacuum-infiltration system was used to infiltrate *A. tumefaciens* strains in bulk. *N. benthamiana* leaves.

Approximately 80 g of leaf tissue was extracted overnight in 600 mL hexane at 4 °C and 150 rpm. The extract was dried down on a rotary evaporator. Each product was purified by silica gel flash column chromatography with a mobile phase of 100% hexane for (+)-kaurene and successive column washes from 100% hexane to 95/5 hexane/ethyl acetate for 3(*S*)-hydroxy-(+)-manool. NMR spectra were measured in Michigan State University's Max T. Rogers NMR Facility on a Bruker 800 MHz or 600 MHz spectrometer equipped with a TCl cryoprobe using $CDCl_3$ as the solvent. $CDCl_3$ peaks were referenced to 7.26 and 77.00 ppm for ^{1}H and ^{13}C spectra, respectively.

Availability of supporting information

The data supporting the findings of this work are available within the Supplementary Information files, which are available in the published online version of this work. The raw genomic reads generated in this study have been deposited in the NCBI BioSample database under the following accession codes *Plectranthus barbatus* (SAMN26547115), *Leonotis leonurus* (SAMN26547116), and *Prunella vulgaris* (SAMN26547117). The genome assemblies have been deposited in NCBI with accession codes *Plectranthus barbatus* (JAPKLW000000000), *Leonitis leonurus* (JAPKLX000000000), and *Prunella vulgaris* (JAPLKY000000000). Sequences for the functionally characterized enzymes from *Callicarpa americana* can be found in the NCBI GenBank database: ON260868-ON260876. Additional Supplementary materials including genome assemblies and annotations, GC-MS raw data, NMR raw data, phylogenetic alignments, cluster sequences, and collinearity files can be found in our Dryad Repository: https://doi.org/10.5061/dryad.w9ghx3frg.

Acknowledgements

This work was supported in part through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University and the Georgia Advanced Computing Resource Center. We would like to thank Dr. Cassandra Johnny of Michigan State University's Mass Spectrometry and Metabolomics Core Facility for their help in obtaining and interpreting GC-MS data, and Dr. Daniel Holmes and the Max T. Rogers NMR Facility for their help in obtaining and interpreting NMR data. We would also like to thank Dr. David Nelson for the naming of all CYP sequences presented in this work, Dr. Kevin Childs for his advice and guidance, Dr. Wajid Bhat for extracting DNA for genomic sequencing, and Malik Sankofa for assistance with plants, media, and general lab preparation. This work was supported by the Michigan State University Strategic Partnership Grant program ('Evolutionary-Driven Genome Mining of Plant Biosynthetic Pathways') to BH and CRB and through Georgia Research Alliance funds to CRB. BH gratefully acknowledges the US Department of Energy Great Lakes Bioenergy Research Center Cooperative Agreement DE-SC0018409, startup funding from the Department of Biochemistry and Molecular Biology, and support from AgBioResearch (MICL02454). GM is supported by a fellowship from Michigan State University under the Training Program in Plant Biotechnology for Health and Sustainability (T32-GM110523), EL is supported by the NSF Graduate Research Fellowship Program (DGE-1848739), and AB is supported by NSF-IMPACTS Training Grant (DGE-1828149). BH is in part supported by the National Science Foundation under Grant Number 1737898. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Postnikova, O. A., Minakova, N. Y., Boutanaev, A. M. & Nemchinov, L. G. Clustering of pathogen-response genes in the genome of *Arabidopsis thaliana*. *J. Integr. Plant Biol.* **53**, 824–834 (2011).

2. Boutanaev, A. M. *et al.* Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E81–E88 (2015).

3. Medema, M. H. *et al.* Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).

4. Nützmann, H.-W., Huang, A. & Osbourn, A. Plant metabolic clusters – from genetics to genomics. *New Phytol.* **211**, 771–789 (2016).

5. Nützmann, H.-W., Scazzocchio, C. & Osbourn, A. Metabolic gene clusters in eukaryotes. *Annu. Rev. Genet.* **52**, 159–183 (2018).

6. Liu, Z. *et al.* Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. *New Phytol.* **227**, 1109–1123 (2020).

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

8. Frey, M. *et al.* Analysis of a Chemical Plant Defense Mechanism in Grasses. *Science* **277**, 696–699 (1997).

9. Chu, H. Y., Wegel, E. & Osbourn, A. From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *Plant J.* **66**, 66–79 (2011).

10. Winzer, T. *et al.* A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science* **336**, 1704–1708 (2012).

11. Yang, X. *et al.* Three chromosome-scale Papaver genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway. *Nat. Commun. 2021 121* **12**, 1–14 (2021).

12. Shang, Y. *et al.* Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084–1088 (2014).

13. Dai, L. *et al.* Functional Characterization of Cucurbitadienol Synthase and Triterpene Glycosyltransferase Involved in Biosynthesis of Mogrosides from *Siraitia grosvenorii.* *Plant Cell Physiol.* **56**, 1172–1182 (2015).

14. Sakamoto, T. *et al.* An Overview of Gibberellin Metabolism Enzyme Genes and Their Related Mutants in Rice. *Plant Physiol.* **134**, 1642–1653 (2004).

15. Wilderman, P. R., Xu, M., Jin, Y., Coates, R. M. & Peters, R. J. Identification of Syn-Pimara-7,15-Diene Synthase Reveals Functional Clustering of Terpene Synthases Involved. *Plant Physiol.* **135**, 2098–2105 (2004).

16. Schmelz, E. A. *et al.* Biosynthesis, elicitation and roles of monocot terpenoid phytoalexins. *Plant J.* **79**, 659–678 (2014).

17. Kitaoka, N. *et al.* Interdependent evolution of biosynthetic gene clusters for momilactone production in rice. *Plant Cell* (2020) doi:10.1093/plcell/koaa023.

18. Liang, J. *et al.* Rice contains a biosynthetic gene cluster associated with production of the casbane-type diterpenoid phytoalexin *ent* -10-oxodepressin. *New Phytol.* nph.17406 (2021) doi:10.1111/nph.17406.

19. Slot, J. C. & Hibbett, D. S. Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: A phylogenetic study. *PLOS ONE* **2**, e1097 (2007).

20. Slot, J. C. & Rokas, A. Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc. Natl. Acad. Sci.* **107**, 10136–10141 (2010).

21. Field, B. *et al.* Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci.* **108**, 16116–16121 (2011).

22. Takos, A. M. & Rook, F. Why biosynthetic genes for chemical defense compounds cluster. *Trends Plant Sci.* **17**, 383–388 (2012).

23. Zhang, J. & Peters, R. J. Why are momilactones always associated with biosynthetic gene clusters in plants? *Proc. Natl. Acad. Sci.* **117**, 13867–13869 (2020).

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

25. Hurst, L. D., Pál, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310 (2004).

26. Qi, X. *et al.* A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci.* **101**, 8233–8238 (2004).

27. Okada, A. *et al.* OsTGAP1, a bZIP transcription factor, coordinately regulates the inductive production of diterpenoid phytoalexins in rice. *J. Biol. Chem.* **284**, 26510–26518 (2009).

28. Mugford, S. T. *et al.* Modularity of plant metabolic gene clusters: A trio of linked genes that are collectively required for acylation of triterpenes in oat. *Plant Cell* **25**, 1078–1092 (2013).

29. Yu, N. *et al.* Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Res.* **44**, 2255–2265 (2016).

30. Rokas, A., Wisecaver, J. H. & Lind, A. L. The birth, evolution and death of metabolic gene clusters in fungi. *Nat. Rev. Microbiol.* **16**, 731–744 (2018).

31. Nützmann, H.-W. *et al.* Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. *Proc. Natl. Acad. Sci.* **117**, 13800–13809 (2020).

32. Li, Y. *et al.* Subtelomeric assembly of a multi-gene pathway for antimicrobial defense compounds in cereals. *Nat. Commun.* **12**, 2563 (2021).

33. Liu, Z. *et al.* Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nat. Commun.* **11**, 5354 (2020).

34. Field, B. & Osbourn, A. E. Metabolic Diversification—Independent Assembly of Operon-Like Gene Clusters in Different Plants. *Science* **320**, 543–547 (2008).

35. Itkin, M. *et al.* Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**, 175–179 (2013).

36. Matsuba, Y. *et al.* Evolution of a complex locus for terpene biosynthesis in *Solanum*. *Plant Cell* **25**, 2022–2036 (2013).

37. Johnson, S. R. *et al.* Promiscuous terpene synthases from *Prunella vulgaris* highlight the importance of substrate and compartment switching in terpene synthase evolution. *New Phytol.* **223**, 323–335 (2019).

38. Schenck, C. A. & Last, R. L. Location, location! Cellular relocalization primes specialized metabolic diversification. *FEBS J.* **287**, 1359–1368 (2020).

39. Fan, P. *et al.* Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity. *eLife* **9**, e56717 (2020).

40. Dictionary of Natural Products 30.2. Accessed March 20th, 2023. https://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml.

41. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229 (2011).

42. Tholl, D. Biosynthesis and Biological Functions of Terpenoids in Plants. *Adv. Biochem. Eng. Biotechnol.* **148**, 63–106 (2015).

43. Gershenzon, J. & Dudareva, N. The function of terpene natural products in the natural world. *Nat. Chem. Biol. 2007 37* **3**, 408–414 (2007).

44. Bohlmann, J., Steele, C. L. & Croteau, R. Monoterpene synthases from grand fir (*Abies grandis*): cDNA isolation, characterization, and functional expression of myrcene synthase, (-)-(4S)- limonene synthase, and (-)-(1S,5S)-pinene synthase. *J. Biol. Chem.* **272**, 21784–21792 (1997).

45. Karunanithi, P. S. & Zerbe, P. Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Front. Plant Sci.* **10**, 1166 (2019).

46. Boutanaev, A. M. *et al.* Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci.* **112**, E81–E88 (2015).

47. Bathe, U. & Tissier, A. Cytochrome P450 enzymes: A driving force of plant diterpene diversity. *Phytochemistry* **161**, 149–162 (2019).

48. Lange, B. M. The Evolution of Plant Secretory Structures and Emergence of Terpenoid Chemical Diversity. *Annu. Rev. Plant Biol.* **66**, (2015).

49. Johnson, S. R. *et al.* A database-driven approach identifies additional diterpene synthase activities in the mint family (Lamiaceae). *J. Biol. Chem.* **294**, 1349–1362 (2019).

50. Sherden, N. H. *et al.* Identification of iridoid synthases from *Nepeta species*: Iridoid cyclization does not determine nepetalactone stereochemistry. *Phytochemistry* **145**, 48–56 (2018).

51. Xu, H. *et al.* Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*. *Mol. Plant* **9**, 949–952 (2016).

52. Song, Z. *et al.* A high-quality reference genome sequence of *Salvia miltiorrhiza* provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome* **13**, e20041 (2020).

53. Gao, W. *et al.* A functional genomics approach to tanshinone biosynthesis provides stereochemical insights. *Org. Lett.* **11**, 5170–5173 (2009).

54. Ma, Y. *et al.* Genome-wide identification and characterization of novel genes involved in terpenoid biosynthesis in *Salvia miltiorrhiza*. *J. Exp. Bot.* **63**, 2809–2823 (2012).

55. Guo, J. *et al.* CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts. *Proc. Natl. Acad. Sci.* **110**, 12108–12113 (2013).

56. Guo, J. *et al.* Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. *New Phytol.* **210**, 525–534 (2016).

57. Cui, G. *et al.* Functional divergence of diterpene syntheses in the medicinal plant *Salvia miltiorrhiza*. *Plant Physiol.* **169**, 1607–1618 (2015).

58. Bai, Z. *et al.* The ethylene response factor SmERF6 co-regulates the transcription of SmCPS1 and SmKSL1 and is involved in tanshinone biosynthesis in *Salvia miltiorrhiza* hairy roots. *Planta* **248**, 243–255 (2018).

59. Wang, Z. & Peters, R. J. Tanshinones: Leading the way into Lamiaceae labdane-related diterpenoid biosynthesis. *Curr. Opin. Plant Biol.* **66**, 102189 (2022).

60. Song, J.-J. *et al.* A 2-oxoglutarate-dependent dioxygenase converts dihydrofuran to furan in *Salvia* diterpenoids. *Plant Physiol.* **188**, 1496–1506 (2022).

61. González, M. A. Aromatic abietane diterpenoids: Their biological activity and synthesis. *Nat. Prod. Rep.* **32**, 684–704 (2015).

62. Smith, E. C. J., Wareham, N., Zloh, M. & Gibbons, S. 2β-Acetoxyferruginol—A new antibacterial abietane diterpene from the bark of *Prumnopitys andina*. *Phytochem. Lett.* **1**, 49–53 (2008).

63. Machumi, F. *et al.* Antimicrobial and antiparasitic abietane diterpenoids from the roots of *Clerodendrum eriophyllum*. *Nat. Prod. Commun.* **5**, 1934578X1000500605 (2010).

64. Abdissa, N., Frese, M. & Sewald, N. Antimicrobial abietane-type diterpenoids from *Plectranthus punctatus*. *Molecules* **22**, 1919 (2017).

65. Gao, J. *et al.* Anti-NLRP3 inflammasome abietane diterpenoids from *Callicarpa bodinieri* and their structure elucidation. *Chin. Chem. Lett.* **31**, 427–430 (2020).

66. Birtić, S., Dussort, P., Pierre, F.-X., Bily, A. C. & Roller, M. Carnosic acid. *Phytochemistry* **115**, 9–19 (2015).

67. Ignea, C. *et al.* Carnosic acid biosynthesis elucidated by a synthetic biology platform. *Proc. Natl. Acad. Sci.* **113**, 3681–3686 (2016).

68. Scheler, U. *et al.* Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast. *Nat. Commun.* **7**, 12942 (2016).

69. Zhao, D. *et al.* Chromosomal-scale genome assembly of *Tectona grandis* reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *Giga Science* doi.org/10.1093/gigascience/giz005 (2019).

70. Hamilton, J. P. *et al.* Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, *Callicarpa americana*. *GigaScience* **9**, giaa093 (2020).

71. Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B. & Peters, R. J. CYP76M7 is an *ent*-cassadiene C11α-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell* **21**, 3315–3325 (2009).

72. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

73. Stanke, M. *et al.* AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).

74. Xu, Z. *et al.* Comparative genome analysis of *Scutellaria baicalensis* and *Scutellaria barbata* reveals the evolution of active flavonoid biosynthesis. *Genomics Proteomics Bioinformatics* **18**, 230–240 (2020).

75. Lichman, B. R. *et al.* The evolutionary origins of the cat attractant nepetalactone in catnip. *Sci. Adv.* **6**, (2020).

76. Bornowski, N. *et al.* Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae. *DNA Res.* **27**, dsaa016 (2020).

77. He, Y. *et al.* Building an octaploid genome and transcriptome of the medicinal plant *Pogostemon cablin* from Lamiales. *Sci. Data* **5**, 180274 (2018).

78. Godden, G. T., Kinser, T. J., Soltis, P. S. & Soltis, D. E. Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints. *Genome Biol. Evol.* **11**, 3393–3408 (2019).

79. Yao, G. *et al.* Phylogenetic relationships, character evolution and biogeographic diversification of *Pogostemon* s.l. (Lamiaceae). *Mol. Phylogenet. Evol.* **98**, 184–200 (2016).

80. Li, P. *et al.* Molecular phylogenetics and biogeography of the mint tribe Elsholtzieae (Nepetoideae, Lamiaceae), with an emphasis on its diversification in East Asia. *Sci. Rep.* **7**, 2057 (2017).

81. Cooley, A. M. *et al.* Genetic architecture of spatially complex color patterning in hybrid *Mimulus*. 2022.04.29.490035 Preprint at https://doi.org/10.1101/2022.04.29.490035 (2022).

82. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).

83. Bak, S. *et al.* Cytochromes P450. *Arab. Book Am. Soc. Plant Biol.* **9**, e0144 (2011).

84. Pateraki, I. *et al.* Total biosynthesis of the cyclic AMP booster forskolin from *Coleus forskohlii*. *eLife* **6**, e23001 (2017).

85. Kliebenstein, D. J. A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PloS One* **3**, e1838 (2008).

86. Liu, B. *et al.* Phylogenetic relationships of *Cyrtandromoea* and *Wightia* revisited: A new tribe in Phrymaceae and a new family in Lamiales. *J. Syst. Evol.* **58**, 1–17 (2020).

87. Hillwig, M. L. *et al.* Domain loss has independently occurred multiple times in plant terpene synthase evolution. *Plant J.* **68**, 1051–1060 (2011).

88. Wisecaver, J. H. *et al.* A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell* **29**, 944–959 (2017).

89. Zi, J. & Peters, R. J. Characterization of CYP76AH4 clarifies phenolic diterpenoid biosynthesis in the Lamiaceae. *Org. Biomol. Chem.* **11**, 7650 (2013).

90. Jones, W. P. & Kinghorn, A. D. Biologically active natural products of the genus *Callicarpa*. *Curr. Bioact. Compd.* **4**, 15–32 (2008).

91. Bose, S. K. *et al.* Effect of gibberellic acid and calliterpenone on plant growth attributes, trichomes, essential oil biosynthesis and pathway gene expression in differential manner in *Mentha arvensis* L. *Plant Physiol. Biochem.* **66**, 150–158 (2013).

92. Andersen-Ranberg, J. *et al.* Expanding the landscape of diterpene structural diversity through stereochemically controlled combinatorial biosynthesis. *Angew. Chem. Int. Ed Engl.* **55**, 2142–2146 (2016).

93. Hu, Z. *et al.* Functional Characterization of a 2OGD Involved in Abietane-Type Diterpenoids Biosynthetic Pathway in *Salvia miltiorrhiza*. *Front. Plant Sci.* **13**, (2022).

94. Wang, J. *et al.* A cytochrome P450 CYP81AM1 from *Tripterygium wilfordii* catalyses the C-15 hydroxylation of dehydroabietic acid. *Planta* **254**, 95 (2021).

95. Cantrell, C. L., Klun, J. A., Bryson, C. T., Kobaisy, M. & Duke, S. O. Isolation and identification of mosquito bite deterrent terpenoids from leaves of American (*Callicarpa americana*) and Japanese (*Callicarpa japonica*) beautyberry. *J. Agric. Food Chem.* **53**, 5948–5953 (2005).

96. Jones, W. P. *et al.* Cytotoxic constituents from the fruiting branches of *Callicarpa americana* collected in southern Florida. *J. Nat. Prod.* **70**, 372–377 (2007).

97. Dettweiler, M. *et al.* A clerodane diterpene from *Callicarpa americana* resensitizes methicillin-resistant staphylococcus aureus to β-lactam antibiotics. *ACS Infect. Dis.* **6**, 1667–1673 (2020).

98. Wang, Z.-H., Niu, C., Zhou, D.-J., Kong, J.-C. & Zhang, W.-K. Three New Abietane-Type Diterpenoids from *Callicarpa macrophylla* Vahl. *Molecules* **22**, 842 (2017).

99. Hillwig, M. L. *et al.* Domain loss has independently occurred multiple times in plant terpene synthase evolution. *Plant J.* **68**, 1051–1060 (2011).

100. Jiang, S.-Y., Jin, J., Sarojam, R. & Ramachandran, S. A Comprehensive Survey on the Terpene Synthase Gene Family Provides New Insight into Its Evolutionary Patterns. *Genome Biol. Evol.* **11**, 2078–2098 (2019).

101. Tu, L. *et al.* Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat. Commun.* **11**, 971 (2020).

102. Chaturvedi, R. *et al.* An abietane diterpenoid is a potent activator of systemic acquired resistance. *Plant J.* **71**, 161–172 (2012).

103. Smirnova, I. E., Tret'yakova, E. V., Baev, D. S. & Kazakova, O. B. Synthetic modifications of abietane diterpene acids to potent antimicrobial agents. *Nat. Prod. Res.* **0**, 1–9 (2021).

104. Zeng, T. *et al.* TeroKit: A Database-Driven Web Server for Terpenome Research. *J. Chem. Inf. Model.* **60**, 2082–2090 (2020).

105. Nützmann, H.-W. & Osbourn, A. Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* **26**, 91–99 (2014).

106. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

107. Wang, Y. *et al.* MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).

108. Bandi, V. & Gutwin, C. SynVisio: An interactive multiscale synteny visualization tool for MCScanX. in *In Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI'20)* (Canadian Human-Computer Communications Society, 2020).

109. Revell, L. J. Phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

110. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

111. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

112. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

113. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).

114. Assembly-stats. (2022). https://github.com/sanger-pathogens/assembly-stats.

115. Geib, S. M. *et al.* Genome Annotation Generator: a simple tool for generating and correcting WGS annotation tables for NCBI submission. *GigaScience* **7**, giy018 (2018).

116. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).

117. Campbell, M. S. *et al.* MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. *Plant Physiol.* **164**, 513–524 (2014).

118. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).

119. Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **5**, 4.10.1-4.10.14 (2004).

120. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

121. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat. Methods* **14**, 417–419 (2017).

122. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Pringer-Verlag New York, 2016).

123. Sainsbury, F., Thuenemann, E. C. & Lomonossoff, G. P. pEAQ: Versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* **7**, 682–693 (2009).

124. Englund, E., Andersen-Ranberg, J., Miao, R., Hamberger, B. & Lindberg, P. Metabolic engineering of *Synechocystis* sp. PCC 6803 for production of the plant diterpenoid manoyl oxide. *ACS Synth. Biol.* **4**, 1270–1278 (2015).

125. Boachon, B. *et al.* Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. *Mol. Plant* **11**, 1084–1096 (2018).

126. Guenard, D., Gueritte-Voegelein, F. & Potier, P. Taxol and taxotere: Discovery, chemistry, and structure-activity relationships. *Acc. Chem. Res.* **26**, 160–167 (1993).

127. Croteau, R., Ketchum, R. E. B., Long, R. M., Kaspera, R. & Wildung, M. R. Taxol biosynthesis and molecular genetics. *Phytochem. Rev.* **5**, 75–97 (2006).

128. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nat. 2013 4967446* **496**, 528–532 (2013).

129. Pateraki, I. *et al.* Manoyl oxide (13R), the biosynthetic precursor of forskolin, is synthesized in specialized root cork cells in *Coleus forskohlii*. *Plant Physiol.* **164**, 1222–1236 (2014).

CHAPTER 4: CYP76BK1 ORTHOLOGS CATALYZE FURAN AND LACTONE RING FORMATION IN CLERODANE DITERPENOIDS ACROSS THE MINT FAMILY

Emily R. Lanier*, Nicholas J. Schlecht*, Trine B. Andersen, Bjoern R. Hamberger

*These authors contributed equally to this work

Author contributions:

ERL, NJS, and BRH conceived and designed the study; NJS, TBA, and ERL performed the experiments; NJS analyzed the DNP data; ERL, TBA, and NJS analyzed the experimental data; ERL wrote the manuscript; BH and TBA supervised the project; all authors contributed to revisions.

Abstract

The mint family (Lamiaceae) produces an abundance of specialized diterpenoid metabolites which facilitate plant-environment interactions. Furanoclerodanes comprise a subset of diterpenoids from the mint family which are particularly bioactive as insect antifeedants. Here, we report a set of orthologous cytochrome P450 enzymes from 8 different mint family species which are key pathway enzymes for furanoclerodane biosynthesis. These CYP76BK1s catalyze formation of a furan and lactone derivative of both the neo-clerodane substrates kolavenol and isokolavenol. Additionally, we investigate the diterpene synthases in these plants and identify likely participants in clerodane biosynthesis. The conservation of this P450 across 50M years of evolution suggests its importance in the biosynthesis of furanoclerodanes, and discovery of this set of enzymes provides important biotechnological access to both furan and lactone clerodane derivatives.

Introduction

Diterpenoids are a widespread and diverse group of mostly specialized metabolites which are found throughout all kingdoms of life but are particularly prevalent in plants. The variety of their structures embodies the various roles they play in plants, serving in pathogen and herbivory defense, abiotic stress responses, signaling for symbiotic relationships and plant development. In addition to their native function, these compounds can also be coopted as agriculturally and pharmaceutically relevant natural products.

All terpenoids in plants are made of precursors from either the cytosolic mevalonate (MVA) or the plastidial methylerythritol phosphate (MEP) pathway, which generate dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP). Within the plastids, IPP and DMAPP are condensed to the canonical diterpenoid precursor geranylgeranyl diphosphate (GGPP) by the prenyl transferase GGPP synthase. From here there is a layered route towards diterpenoids that allows for generation of hundreds of distinct backbones. The most prevalent class of diterpenes are the labdane-type, with a characteristic decalin core [1]. For these, a class II diterpene synthase (diTPS) catalyzes proton-mediated cyclization of GGPP into a bicyclic diphosphate intermediate, which is typically followed by a class I diTPS that cleaves the diphosphate and leads to further rearrangements of the distal carbon chain. Alternately, class I diTPSs can act independently of class II diTPSs to yield irregular and non-labdane diterpenoids [2]. The resulting diterpene backbones are often oxidized by cytochrome P450s, which are typically within the CYP71 clan [3].

The Lamiaceae (mint) family is one of the greatest sources for diterpenoid structures. Of the over 23,000 distinct diterpenoids currently reported in the Dictionary of Natural Products (DNP)[4] from all kingdoms of life, roughly a fifth are from the Lamiaceae family alone. This can, in part, be

explained by the various whole genome duplication events that have occurred throughout the Lamiaceae family's evolutionary history, which leads to opportunities for neofunctionalization of existing diTPS genes [5].

Clerodanes are a class of labdane-type diterpenes which are most prevalent among Lamiaceae species, though they are also present in several other plant families. They are distinguished by their unique arrangement of the methyl groups, which is brought about through methyl and hydride shifts during cyclization of GGPP by the class II diTPS. Within the mint family, this backbone gives rise to powerful opioid receptor agonists in the psychedelic plant *Salvia divinorum*, potent insect antifeedant and insecticidal compounds in species of the *Scutelleroideae* and *Ajugoideae* subfamilies, and a recently identified MRSA-active antibacterial clerodane in *Callicarpa americana* [6,7]. Additional therapeutic bioactivities have also been reported for a range of other clerodane diterpenoids [8]. Most clerodanes can be classified into 7 types based on the presence of different oxygenated rings on this sidechain, typically some variation of furan and lactone moieties (Fig. 4.1) [8]. In rare cases, other heteroatoms can be included in side chains or even as pyrrolidine rings [9]. The widespread presence of furan and lactone moieties and variations thereof appear to be the main drivers of the biological activities of clerodanes [7,8,10].

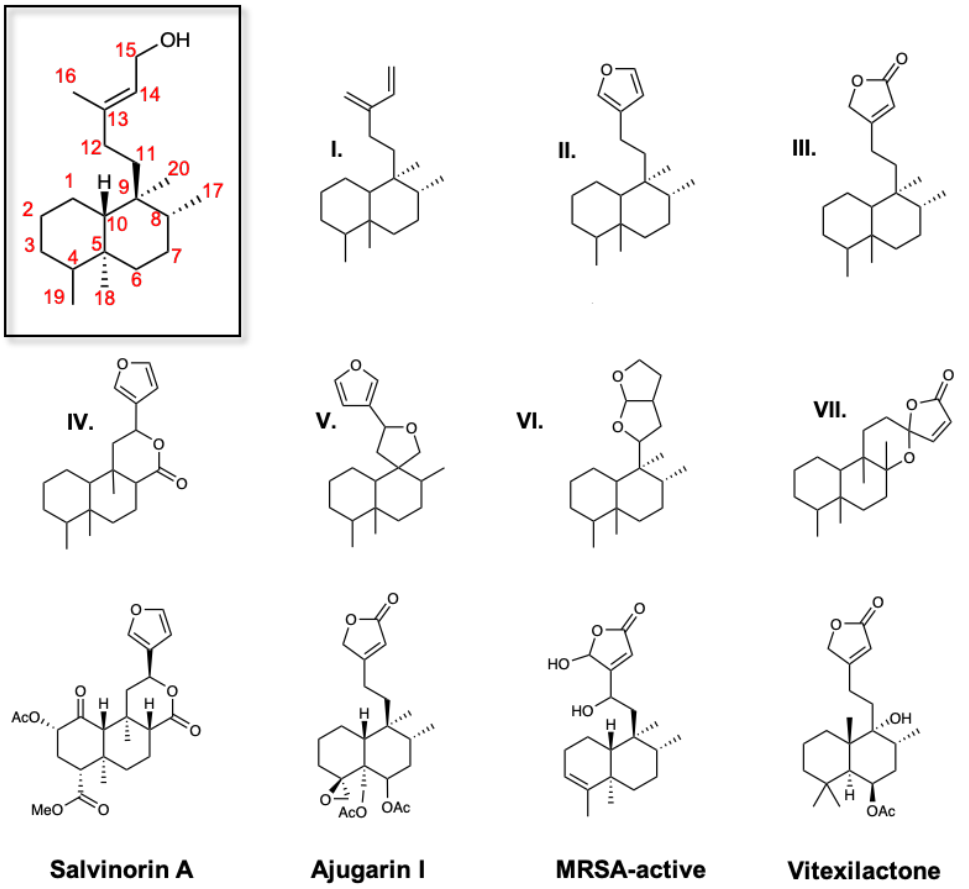


Figure 4.1. Structures relevant clerodanes and furanoditerpenoids. Inset, traditional carbon numbering of the neoclerodane backbone. **I.**-**VII.**, seven types of clerodane backbones. Salvinorin A is from *S. divinorum*; Ajugarin I is from *A. reptans*, the MRSA-active furanoclerodane is from *C. americana*, and Vitexilactone is an example of a furanolabdane from *V. agnus-castus.*

Although furanoditerpenoids are primarily derived from clerodanes, there are also those with other labdane backbones [10]. Data from the DNP shows that furanoditerpenoids, and furanoclerodanes in particular, have been reported more in the Lamiaceae family than any other plant family (Fig. 4.2). Within the Lamiaceae, seven subfamilies have reported furanoditerpenoids. More furanoclerodane structures have been reported in the Ajugoideae subfamily than any other clade, and clerodane backbones dominate over other labdanes.

According to this dataset, the labdane and clerodane furanoditerpenoids appear to be mutually exclusive at the genus level, demonstrating typical lineage-specificity of diterpenoid types [11].

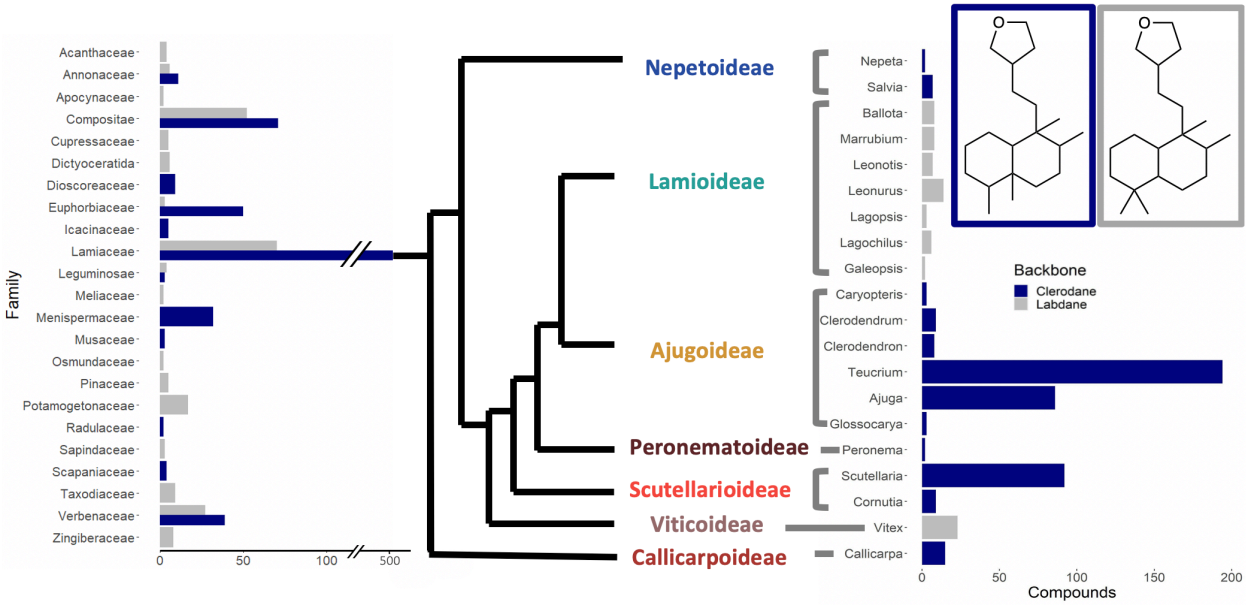


Figure 4.2. Distribution of furanoditerpenoids reported in the DNP [12]. Navy bars indicate clerodane backbones and gray bars indicate those with labdane backbones. Specific substructures used for this search are boxed. Lamiaceae species are grouped according to subfamily.

The past seven years have brought the first progress towards elucidation of the enzymes involved in clerodane and furanoditerpenoid biosynthesis in plants. TwTPS14, from *Trypterygium wilfordii* of the family Celestraceae, was the first reported class II enzyme from plants capable of catalyzing formation of the neo-clerodane precursor kolavenyl diphosphate (KPP) [13]. Within the Lamiaceae, KPP synthase activity has been identified from *Salvia divinorum* (SdKPS1), *Vitex agnus-castus* (VacTPS5), *C. americana* (CamTPS2) and most recently *Scutellaria baicalensis* (SbdiTPS 2.8) and *Salvia splendens* (SspdiTPS 2.1) [14–17]. A double-bond isomer of KPP, iso-KPP, is the major product of ArTPS2 from *Ajuga reptans* as well as diTPSs from both *S. baicalensis* (SbdiTPS 2.7) and *Scutellaria barbata* (SbbdiTPS 2.1, 2.3) [16,18]. One intriguing finding is that in most of these plants, no class I diTPS could be found to convert KPP and iso-KPP into their corresponding alcohols,

kolavenol and iso-kolavenol, which are the likely precursors to the furan and lactone derivatives. Class I diTPS typically catalyze hydride and alkyl shifts. Formation of the complex clerodanes in contrast starts with a simple cleavage of the diphosphate and deprotonation, or water quenching of the carbocation (Fig. 4.1). The structural diversity emerges thereafter through non-TPS catalyzed ring formations. It has been postulated that phosphatase activity could contribute to this dephosphorylation, but no evidence has been presented yet supporting the role of phosphatases in the native plant. However, a recent study of clerodane biosynthesis in *Scutellaria spp.* and *S. splendens* identified class I diTPSs capable of this activity, suggesting that at least in some plants this pathway follows the traditional class II-class I labdane pathway [16].

Recent work has also uncovered the first P450s involved in furanoditerpenoid biosynthesis. In switchgrass (*Panicum virgatum*), several P450s in the monocot-specific CYP71Z family were shown to catalyze furan ring formation from the diterpene alcohols of both clerodane and labdane backbones [19]. In *S. divinorum*, CYP76AH39 was found to catalyze production of a dihydrofuran moiety when combined with SdKPS1 [20]. In *V. agnus-castus*, CYP76BK1 was shown to hydroxylate C16 of the labdane peregrinol when tested in yeast. While this could plausibly lead to furan formation consistent with furanolabdanes found in this plant, no furan product was observed [15]. As with many of the studied clerodane pathways, a class I diTPS converts the class II diTPS product peregrinol diphosphate to peregrinol, could not be identified.

Building on our previous work identifying KPP and iso-KPP synthases in *C. americana* and *A. reptans,* respectively, in this study we seek to identify additional pathway enzymes responsible for the oxidations that transform these terpene intermediates into bioactive clerodanes. We discover that *CYP76BK1* orthologs in both *A. reptans* and *C. americana* are capable of catalyzing

addition of a furan ring on both kolavenol and isokolavenol. Based on the conserved function between these evolutionarily divergent species, we further investigate a representative set of mint family transcriptomes and find that *CYP76BK1* orthologs are present in 6 additional species, representing in total 6 of the 11 Lamiaceae subfamilies and accounting for all but one subfamily (Nepetoideae) known to produce furanoclerodanes. Phylogenetic analysis suggests a common ancestral origin for furanoclerodane biosynthesis in all Lamiaceae subfamilies except the *Salvia* (Nepetoideae), echoing the conclusion of a recent evolutionary analysis of the Lamiaceae clerodane diTPSs [16]. Additionally, we find that the CYP76BK1 enzymes characterized here have a range of activity. Some can catalyze production of a lactone ring, as well as the furan, when paired with (iso)kolavenol substrates. This enables biosynthetic access to a wider range of industrially important furanoditerpenoid targets as the first set of enzymes known to catalyze this transformation.

Results & Discussion

***Identification of 2 P450s catalyzing production of furanoclerodanes in* A. reptans *and* C. americana**

We began this work seeking enzymes involved in the biosynthesis of furanoclerodanes from *A. reptans* and *C. americana*. Based on previous work in diterpene pathway elucidation, we hypothesized that at least the first oxidation would be carried out by a P450, likely in the CYP71 clan. Homology with reference CYP71s was used to identify candidates from transcriptomic (*A. reptans*) and genomic (*C. americana*) data, yielding around 250 candidates from each plant. To rank candidates for testing, in *C. americana* we used previously generated tissue-specific expression data to correlate expression of candidates with that of the KPP synthase CamTPS2

(Fig. 4.3). In *A. reptans* only expression data from leaf tissue was available, so candidates were chosen based on strength of expression and clustering with the CYP76 subfamily, which is prominent in Lamiaceae diterpenoid metabolism [3]. This approach yielded 8 candidates from each species for functional characterization.

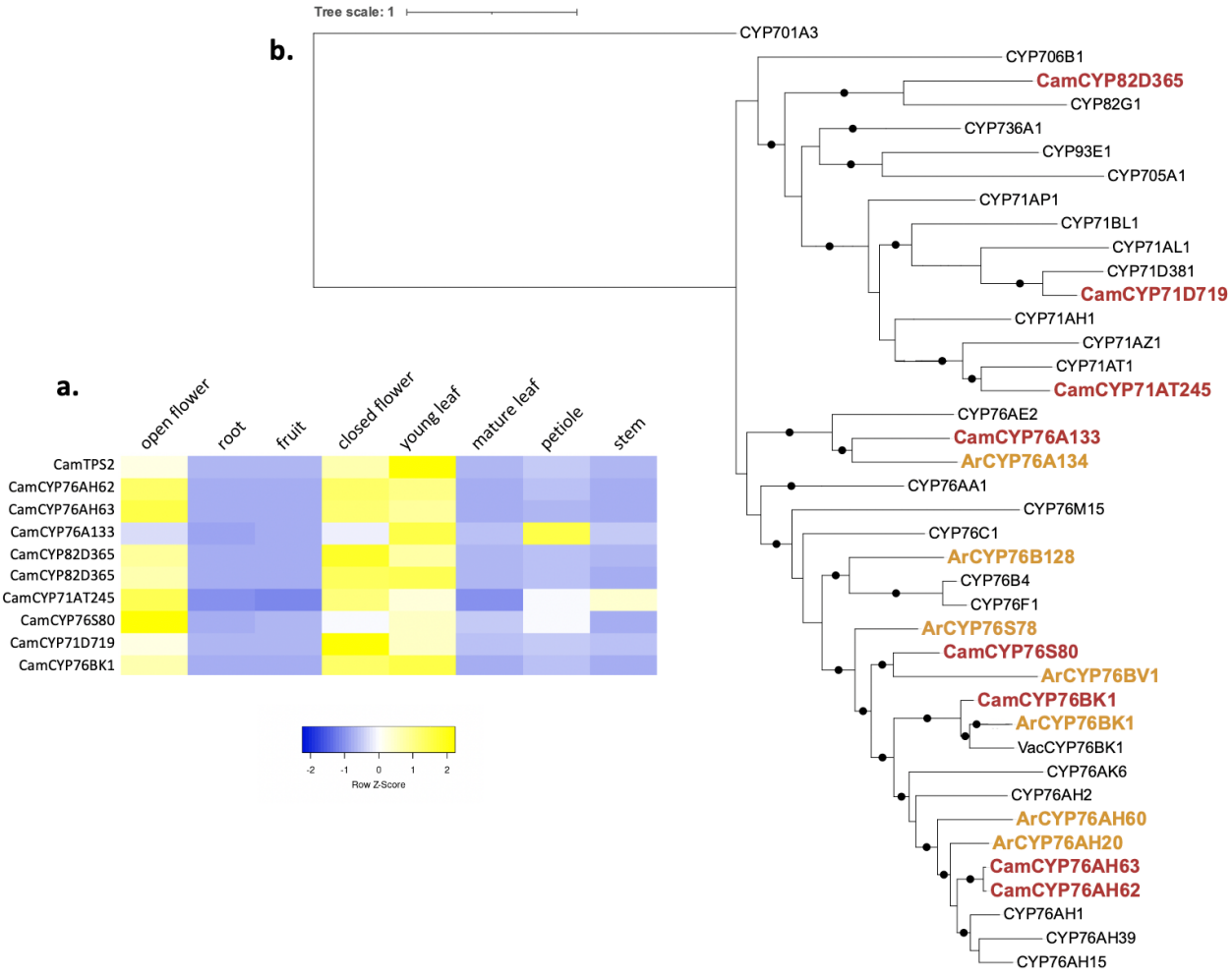


Figure 4.3. P450 candidates selected for functional characterization. (a) Heatmap of *C. americana* candidates showing comparison of expression profile with CamTPS2, the KPP synthase. (b) Maximum-likelihood tree of all cloned candidates from *C. americana* (Cam) and *A. reptans* (Ar). Black circles indicate bootstrap support of 70% or greater (1000 replicates). Sequence names in black are previously reported sequences for reference. Tree is rooted to *Arabidopsis thaliana* CYP701A3.

Candidate genes were cloned and transiently expressed in *Nicotiana benthamiana* along with upstream terpene precursor genes and either *CamTPS2* or *ArTPS2*. As with previous studies using kolavenol and isokolavenol as a substrate, we found that expression of the class II (I)KPP synthases in *N. benthamiana* yielded sufficient quantities of (iso)kolavenol without a class I diTPS due to nonspecific activities of native enzymes (plausibly phosphatases). One candidate P450 from each species was found to convert the clerodane substrate to a new product based on GC-MS analysis. These genes were classified as orthologs, *CamCYP76BK1* and *ArCYP76BK1*.

NMR confirmed that both structures contained a furan moiety (Supplementary Figs. 1-4). This is a plausible intermediate in the pathway towards the bioactive clerodanes and echoes activity of the previously identified CYP71Zs from switchgrass and CYP76AH39 in *S. divinorum*. The peptide sequences exhibit high sequence identity (71-74%) as well as functional similarity to VacCYP76BK1, despite 30-50M years of evolutionary distance between the Callicarpoideae, Viticoideae, and Ajugoideae [21]. This led us to investigate the CYP76BK subfamily further across other Lamiaceae species.

Exploration of CYP76BK1 orthologs across the mint family

A representative set of 48 Lamiaceae species transcriptomes generated by the Mint Evolutionary Genomics Consortium was used to identify additional *CYP76BK1* orthologs (Fig. 4.4a). Peptide sequences with over 45% identity to VacCYP76BK1, CamCYP76BK1, and ArCYP76BK1 were combined for phylogenetic analysis along with reference sequences from the CYP76 family (Fig. 4.4b). In total, we found an additional six species with *CYP76BK1* orthologs. These represent all subfamilies with reported furanoclerodanes except for the Nepetoideae (Fig. 4.1), and includes the subfamily Viticoideae (*VacCYP76BK1*), which has only furanolabdanes reported. These were

named according to species: *CpCYP76BK1* (*Cornutia pyramidata*), *PbCYP76BK1* (*Petraeovitex bambusetorum*), *HsCYP76BK1* (*Holmskioldia sanguinea*), *SbaiCYP76BK1* (*S. baicalensis*), *CbCYP76BK1* (*Clerodendrum bungei*), and *TchCYP76BK1* (*Teucrium chamaedrys*). *T. chamaedrys* (not part of the Mint Consortium study) was substituted with *Teucrium canadense* due to plant availability. Two species containing a *CYP76BK1* ortholog, *H. sanguinea* and *P. bambusetorum*, have no reported diterpenoids. While there are a few phytochemical studies in *H. sanguinea,* identifying bioactive properties of extracts along with identification of flavonoids and iridoids [22,23], the *Petraeovitex* genus lacks phytochemical studies. Yet, both species are members of subfamilies (Scutellariodeae and Peronematoideae, respectively) in which furanoclerodanes have been found.

The lack of a *CYP76BK1* ortholog in the Nepetoideae supports conclusions from previous studies. Our earlier work in *C. americana* and *A. reptans* showed phylogenetic divergence between the peptides of SdKPS1 from *S. divinorum* and CamTPS2, ArTPS2, and VacTPS5 from *V. agnus-castus* [14]. Recently, detailed analysis of two *Scutellaria* genomes provided strong evidence for two evolutionary events leading to clerodane diTPS biosynthesis in the Lamiaceae, both originating with an ancestral *ent*-CPP synthase [16]. One lineage is found in both Callicarpoideae and the clade containing the Viticoideae, Scutellarioideae, and Ajugoideae, indicating an early Lamiaceae ancestor, while the other has been confirmed only in *Salvia* species (Nepetoideae) so far. Likewise, it appears that the transformation from (iso)kolavenol to a furan intermediate has also evolved convergently in separate clades of P450s, once with *CYP76AH39* in *S. divinorum* and again with *CYP76BK1* orthologs conserved in 6 subfamilies. The fact that this activity evolved at

least twice within the CYP76 subfamily is consistent with the prevalence of the CYP76 subfamily in Lamiaceae diterpenoid metabolism.

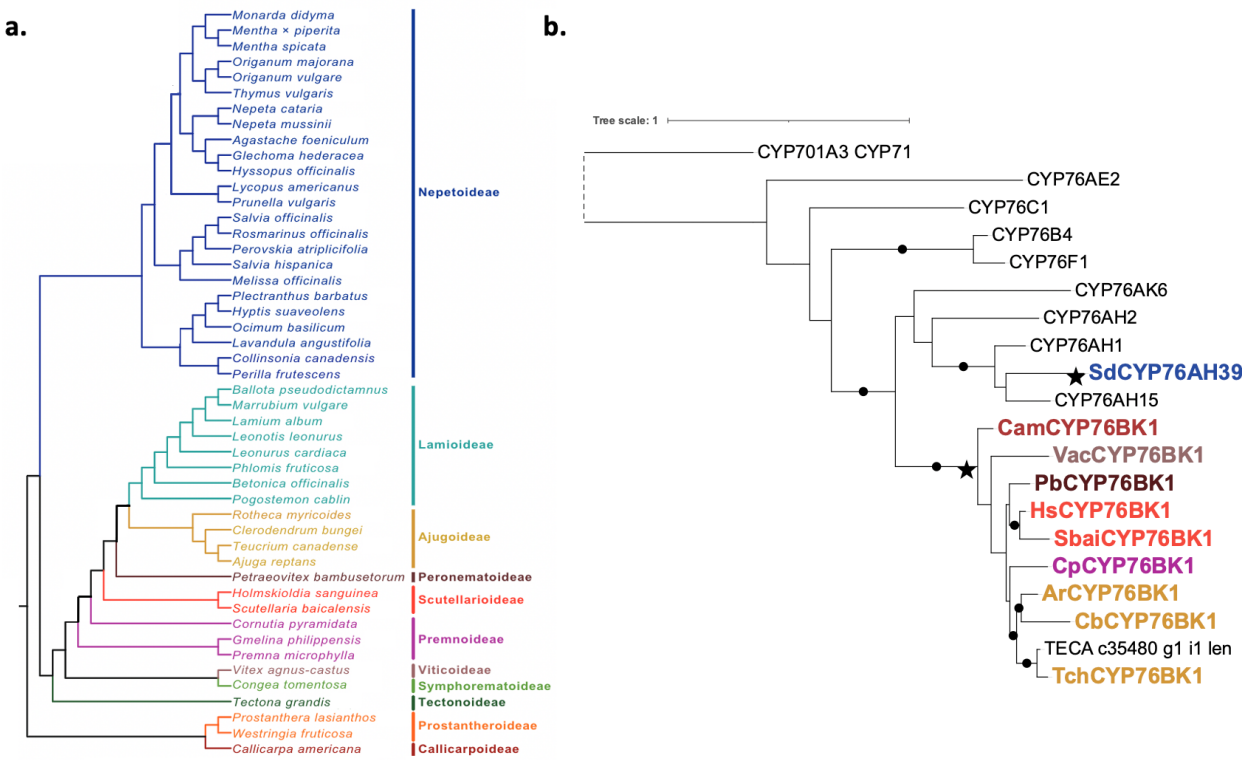


Figure 4.4. Identification of additional CYP76BK1 orthologs from the Lamiaceae family. (a) Species phylogeny of the 48 transcriptomes used to search for CYP76BK1 orthologs. Figure adapted from[5]. (b) Maximum likelihood tree of all identified CYP76BK1 orthologs along with reference sequences of the CYP76 family (in black). Black stars indicate evolutionary events leading to furanoclerodane/furanoditerpenoid catalytic function.

Functional characterization of CYP76BK1 orthologs reveals lactone functionality

CYP76BK1 orthologs were cloned from leaf tissue cDNA and tested against reference P450s *VacCYP76BK1*, *CamCYP76BK1*, and *ArCYP76BK1* using transient expression in *N. benthamiana*. Each *CYP76BK1* ortholog was express with the respective diTPSs for both kolavenol and isokolavenol, with the exception of *TchCYP76BK1*, which was tested only with kolavenol (Fig. 4.5). The ortholog from *C. pyramidata* is still in the cloning stage but will be included in the final data set for publication. In the final experiment for publication, every *CYP76BK1* sequence will be

expressed with diTPSs that catalyze formation of kolavenol, isokolavenol and peregrinol, the native substrate for VacCYP76BK1.

Characterization of these new CYP76BK1 enzymes highlighted an additional catalytic capability for this enzyme class. While previously we had noticed the presence of small side products with express of *CamCYP76BK1* and *ArCYP76BK1*, none were produced in sufficient quantities for purification and structural elucidation. With this expanded set of enzymes, we found that a few could catalyze formation of significant quantities of a second product when combined with either KPP synthase or iso-KPP synthase. Structural elucidation by NMR confirmed it as a lactone derivative (Supplementary Figs. 5-8). The position of the ketone does not match the orientation of most of the bioactive lactones reported in *C. americana*, *A. reptans*, and other mint species. However there is biological support for this product, as there are reports of five clerodane structures with this lactone configuration from species in the Scutellaroideae, Ajugoideae, and Callicarpoideae, and one reported labdane from Viticoideae[4]. The hypothesized mechanism (Fig. 4.6) shows how three successive oxidations could plausibly lead to the lactone product, while just two yield the furan. It appears that certain isoforms of this enzyme are far more active with the third oxidation, yielding appreciable quantities of the lactone in addition to the furan.

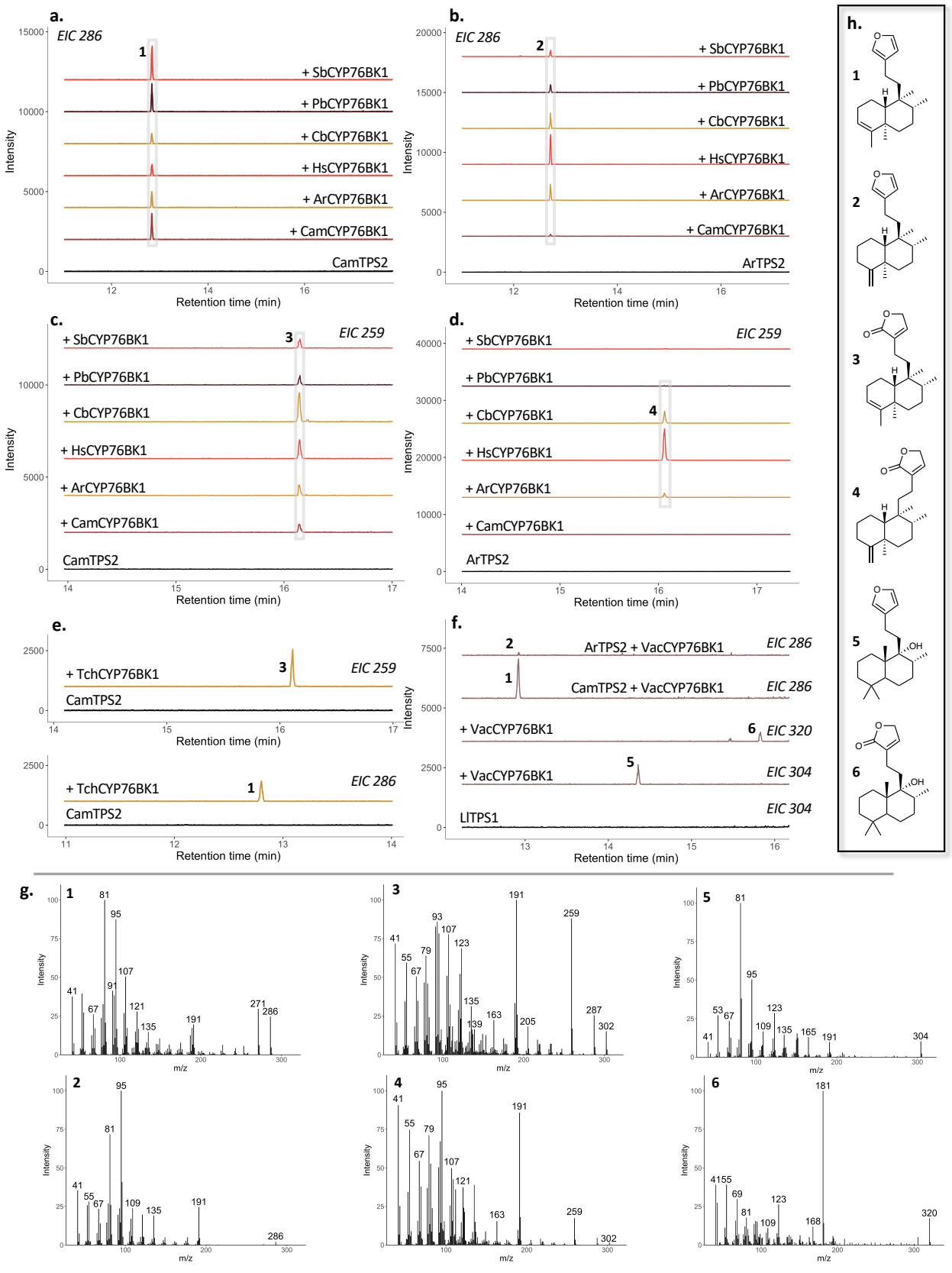


Figure 4.5. Functional characterization of CYP76BK1 orthologs.

Figure 4.5 (cont'd).
(a-f) Extracted ion GC-MS chromatograms of extracts of *N. benthamiana* leaves transiently expressing each P450 candidate with different diTPS. CamTPS2, kolavenol; ArTPS2, isokolavenol; and LlTPS1, peregrinol [1]. (a) Presence of furan product when combined with kolavenol (b) presence of furan product when combined with isokolavenol; (c) presence of lactone product when combined with kolavenol; (d) presence of lactone product when combined with isokolavenol; (e) presence of furan and lactone when expressing TchCYP76BK1 with kolavenol (separate experiment); (f) activity of VacCYP76BK1 with peregrinol, kolavenol, and isokolavenol. (g) Mass spectra of labeled furanoditerpenoid products. (h) Structures of identified enzyme products. Structures of **1**-**4** were verified by NMR analysis and absolute stereochemistry assigned based on the corresponding class II diterpene synthase products.

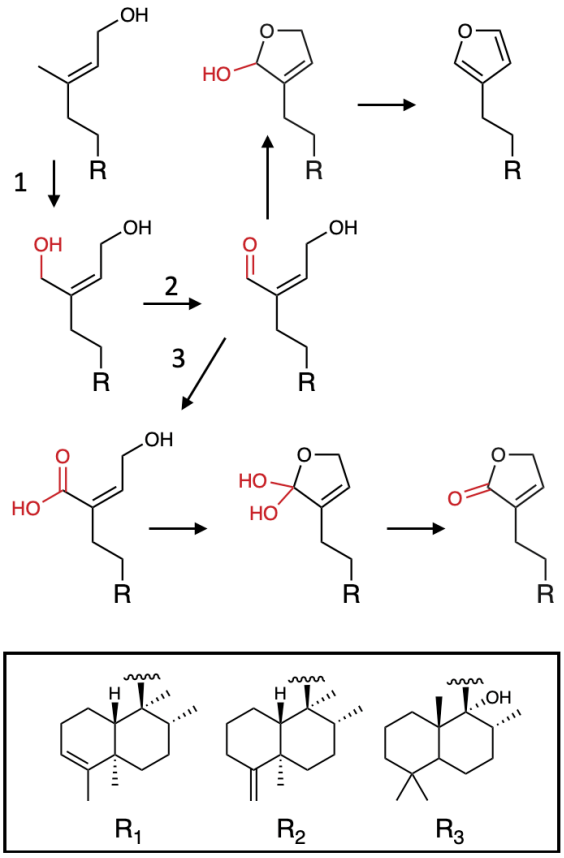


Figure 4.6. Proposed enzyme mechanism. 3 oxidations of C16, indicated by 1, 2, and 3, can lead to the furan or the lactone derivatives of the clerodane and labdane backbones. Non-oxidative steps may be autocatalytic.

Our findings also confirm that VacCYP76BK1 can catalyze formation of two products with mass spectra consistent with furan and lactone derivatives of peregrinol when expressed in *N.*

benthamiana. It also has some limited activity with the kolavenol substrate, although it appears less active than the orthologues from other species. The low activity may explain why VacCYP76BK1 was previously observed to catalyze only one hydroxylation of C16 when tested in yeast. Due to the miniscule yields of the peregrinol derivatives, NMR analysis could not be carried out to confirm these structures.

While the lactone derivatives identified here do have biological relevance, the dominance of the C15 over C16 ketone isomers in the set of reported compounds implies that there is another enzyme in play in these plants. It remains to be seen whether the CYP76BK1 enzymes are working in tandem or in competition with this additional pathway enzyme.

We also note that while all orthologs show promiscuity between the two neo-clerodane scaffolds, activity levels differ when each ortholog is presented with kolavenol vs. isokolavenol. This substrate preference may be a result of positive selection for the clerodane isomer present in each respective species, although that cannot be confirmed without identifying the major diTPS backbones present in each species.

Analysis of plant extracts

In limited cases, the presence of specific classes of diterpenes in a plant can be confirmed by identification of heterologous enzymatic products matching compounds from plant extracts. Hence, we analyzed all leaf extracts of plants with newly identified *CYP76BK1* orthologs by GC-MS for evidence of the furan and lactone intermediates. However, only the leaf extract of *C. americana* had peaks with the same mass spec and retention time as **1** and **3** (Fig. 4.7). Other extracts had peaks which were likely diterpenoids based on mass fragmentation patterns, but did not match the CYP76BK1 products (Supplementary Fig. 9).

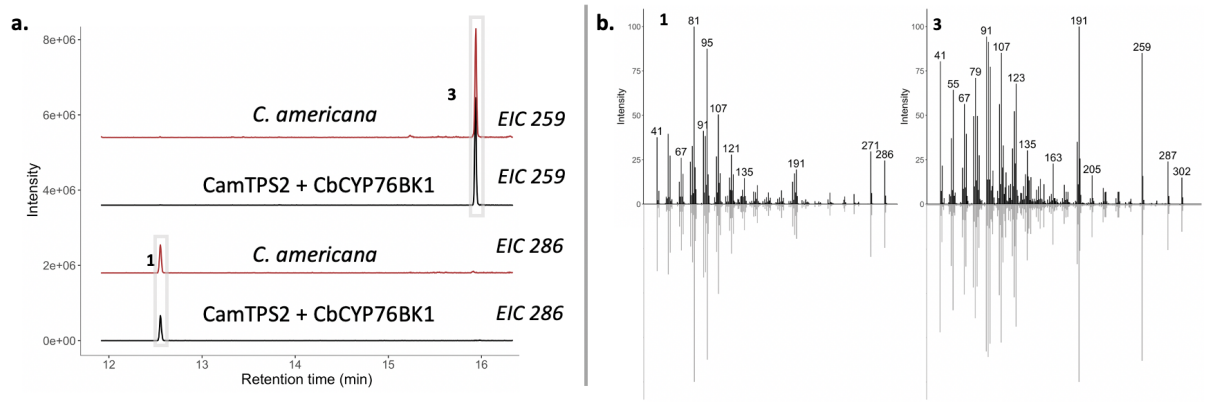


Figure 4.7. GC-MS analysis of plant extracts. (a) EIC showing the presence of the furan (286) and lactone (259) derivatives of kolavenol in both the *C. americana* leaf extract as well as the *N. benthamiana* leaf extract expressing CamTPS2 and CbCYP76BK1. (b) Mass spectra of **1** and **3**. Top (black) from the *C. americana* extract, bottom (gray) from the *N. benthamiana* extract.

***Identification of an (iso)kolavenol synthase orthologue in* V. agnus-castus**

Characterization of enzymes in a heterologous system can be useful to determining their *in planta* function. However, confirming whether tested substrates are present in the plant can further support the conclusions of heterologous testing. While all identified CYP76BK1 enzymes demonstrated the ability to convert two clerodane backbones to their furan and sometimes lactone derivatives, most of these species lack functionally characterized diTPSs. We surveyed transcripts clustering with a set of reference class II diTPS (TPS-c) sequences to identify (I)KPP synthase orthologs, which would support the biological relevance of the CYP76BK1 furanoclerodane catalytic function. The phylogenetic relationships among the predicted class II diTPSs showed that all but one species with a *CYP76BK1* ortholog was also found to have a transcript clustering in the same clade as the other non-*Salvia* (I)KPP synthases (Fig. 4.8). The exception was *P. bambusetorum*. It is possible that within the plant, PbCYP76BK1 utilizes a

different labdane substrate. Alternatively, single-tissue transcriptomes can be incomplete and thus lack evidence for genes which are present in a plant. Although functional characterization would be needed to confirm diTPS activities, this finding supports the likelihood that the CYP76BK1 enzymes have access to clerodane substrates within the plant.

Although most species with characterized clerodane pathways have no reported class I diTPSs in this pathway, one recent study reported class I (iso)kolavenol synthases in *Scutellaria* and Salvia. This led us to investigate the class I labdane-type diTPSs (TPS-e subfamily) across these mint transcriptomes as well (Fig. 4.8). We focused on putative orthologs to the *Scutellaria* sequences in this larger set of transcriptomes given the demonstrated divergence of the *Salvia* clerodane pathway. Three Lamiaceae species were found to have close homologs which cluster in the same clade as Sbb 1.2, Sbb 1.4, and Sb 1.3. Two (*Westringia fruticosa* and *Premna microphylla*) are members of subfamilies for which there are no reported furanoditerpenoids, nor did we find any *CYP76BK1* orthologs. The other is from *V. agnus-castus.* While 6 diTPS were previously cloned and characterized from this plant, this peptide (VacTPS7) does not cluster with any of the *V. agnus-castus* reference sequences. We chose to functionally characterize this representative TPS-e to see if VacTPS7 shares a similar function with the (iso)kolavenol synthases, possibly as a peregrinol synthase. Sbb 1.4 was used for comparison as a reference (iso)kolavenol synthase. This experiment is currently in process, and results will be included for publication.

The lack of evidence for class I diTPSs in all characterized clerodane pathways except *Scutellaria*, including those outside of the Lamiaceae such as switchgrass and *T. wilfordii*, is a fascinating addition to the story of the canonical labdane biosynthesis pathways. It is curious that these pathways both lack a class I diTPS and so often converge towards furanoditerpenoids, as

documented in at least three cases so far: switchgrass, the Nepetoideae, and the CYP76BK1 containing subfamilies of Lamiaceae, which all rely on different P450 subfamilies to accomplish the same transformation. We speculate on two possible explanations for this observation. Either there is a non-canonical class I diTPS that has not yet been discovered, or the molecular structure lends itself to dephosphorylation by phosphatases, such as nudix hydrolases, which have been implicated in other terpenoid pathways [24]. Simple dephosphorylation may then facilitate evolution of furanoditerpenoids due to position of the terminal alcohol on the acyclic sidechain, which is poised to act as a nucleophile and form a myriad of ring structures depending on the location and order of subsequent oxidations.

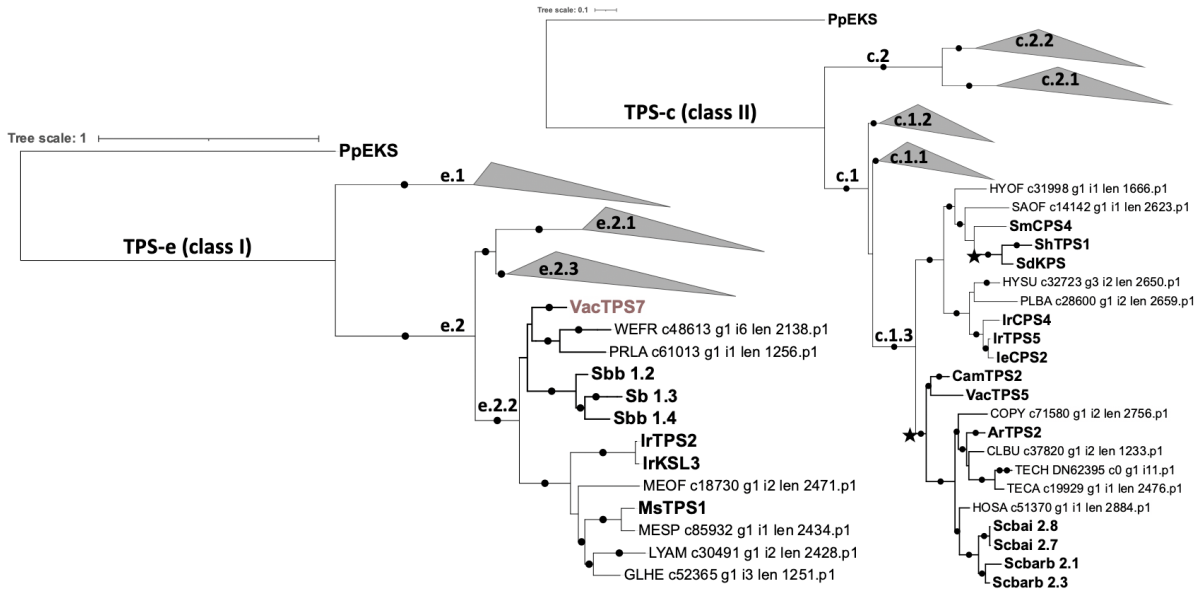


Figure 4.8. Maximum likelihood trees of class I and class II diTPSs. DiTPS trees are rooted to *Physcometrium patens ent*-kaurene synthase (PpEKS). Clade labels are consistent with those previously reported [18]. Clades not known to include clerodane diTPSs are collapsed for clarity. Reference sequences are bolded, and VacTPS7 (brown) is the only transcript which was functionally characterized in this study. Black stars in the TPS-c tree indicate two evolutionary events for acquisition of clerodane diphosphate synthase function. Species are indicated for each transcript by the initial four-letter code, which can be found in Supplementary Table 1.

Conclusion

The results presented here demonstrate a widely conserved purpose for CYP76BK1 enzymes, which have retained sequence similarity and function over 50M years of evolution in the Lamiaceae family. We have also shown that some these orthologs can catalyze not only production of a furan ring but also a lactone, which provides biosynthetic access to a wider range of bioactive diterpenoids. This work has demonstrated that CYP76BK1 is likely a key pathway enzyme in the formation of a large number of industrially valuable furanoditerpenoids in several Lamiaceae species as well as a powerful tool for biotechnological access to the lactone moiety.

Methods

Survey of diterpenoids from the DNP

The DNP was datamined for relevant diterpenoids using the following search criteria. The Type of Compound was either V.S.55000 or V.S.54000, which correspond to clerodanes and labdane diterpenoids respectively. Additionally, lactone, furan, and furanofuran containing carbon backbones were input for each respective backbone and were subsequently searched for any relevant substructures. CSV files were exported containing the Chemical Name, Molecular formula, Accurate mass, Type of Compound, Type of Organism, and Biological Source. The various CSVs were imported, given a column to represent whether they contained a lactone, furanfuranofuran and then each file was combined. Duplicate chemicals found within a given genus were removed. The phylogenetic families were extracted from the 'Type Of Organism' column and made into new columns along with the information on other taxonomic ranks. The data was then grouped by their Chemical names and genera and summed up on a genus level. This process was repeated for family and higher taxonomic ranks as well where each was plotted.

Candidate gene selection

Previously assembled genomic and tissue-specific expression data 14 were used to identify candidate genes in C. americana. The heatmap of candidate gene expression was generated using Heatmapper 25. For all other species, previously assembled transcriptomic data were used 5. Candidate diTPSs and P450s were identified based on 45% identity using BLASTP against a set of reference sequences (Supplementary File S1). For gene expression analysis, trimmed reads were mapped to respective peptides using Salmon 'index' (version 1.8.0), and quantified using Salmon 'quant' (libtype=A, validate mappings)26 to obtain TPM values.

Phylogenetic trees

Reference sequences used in all protein phylogenies were obtained from GenBank (Supplementary Table S3). Full-length peptide sequences were used. Multiple sequence alignments were generated using ClustalOmega (version 1.2.4; default parameters) and phylogenetic trees were generated by RAxML (version 8.2.12; Model = protgammaauto; Algorithm = a) with support from 1000 bootstrap replicates 27,28. Tree graphics were rendered using the Interactive Tree of Life (version 6.5.2) 29.

Plant material

Plants were grown in a greenhouse under ambient photoperiod and 24 °C day/17 °C night temperature and obtained from the following sources: *A. reptans:* Horizon Herbs (Williams, OR)*; C. americana:* Garden Gate Nursery (Gainesville, FL); *C. bungeii:* UF Campus (Gainesville, FL); *Cornutia pyramidata:* Kartuz Greenhouses (Vista, CA); *H. sanguinea:* (McCarthy Gardens, UF Campus, Gainesville, FL); *P. bambusetorum:* Peppers Greenhouses (Milton, DE); *S. baicalensis:* Richters Herbs (Canada); *T. chamaedrys:* Mountain Valley Growers (Yokuts Valley, CA).

Cloning

Synthetic oligonucleotides for all enzymes used in this study are given in Supplementary Table 2. GenBank accession numbers for all cloned genes are given in Supplementary Table 3. RNA was prepared from young leaf material using the Spectrum Total Plant RNA kit (Sigma) with on-column DNAse digest. CDNA was prepared with the SuperScript Double-Stranded cDNA synthesis kit (Thermo-Fisher). Candidate enzymes were PCR-amplified from cDNA, and coding sequences were cloned into the sequencing vector pjET using the CloneJET PCR Cloning kit (Thermo-Fisher) and sequence-verified with respective gene models. Constructs were then cloned into the plant expression vector pEAQ-HT[30] and used in transient expression assays in *N. benthamiana*. Three constructs (*VacTPS7*, *VacCYP76BK1*, and *Sbb 1.4*) were synthesized by Twist Bioscience before cloning into pEAQ-HT.

Transient expression for functional characterization in N. benthamiana

N. benthamiana plants were grown for 5 weeks in a controlled growth room with 25 °C/18 °C day/night and 16h/8h light/dark before infiltration. Constructs for co-expression were separately transformed into *Agrobacterium tumefaciens* strain LBA4404. 20 mL cultures were grown overnight at 30 °C in LB with 50 µg/mL kanamycin and 50 µg/mL rifampicin. Cultures were collected by centrifugation and washed twice with 10 mL water. Cells were resuspended and diluted to an OD_{600} of 1.0 in 200 µM acetosyringone/water and incubated at 30 °C for 1–2 H. Separate cultures were mixed in a 1:1 ratio for each combination of enzymes, and a 1 mL syringe was used to infiltrate 3 mL of culture into the the underside (abaxial side) of *N. benthamiana* leaves. All gene constructs were co-infiltrated with two genes from *Plectranthus barbatus* encoding rate-limiting steps in the upstream (MEP) pathway: 1-deoxy-D-xylulose-5-

phosphate synthase (*PbDXS*) and GGPP synthase (*PbGGPPS*) to boost production of the diterpene precursor GGPP[13,31]. Plants were returned to the controlled growth room for 5 days. Approximately 200 mg fresh weight from infiltrated leaves was extracted with 1 mL hexane overnight at 18 °C. Plant material was removed by centrifugation, and the organic phase was transferred to a fresh vial for GC-MS analysis.

Plant extract metabolomics

Leaves from *C. pyramidata*, *P. bambusetorum*, *H. sanguinea*, *S. baicalensis*, *C. bungei*, *T. chamaedrys, A. reptans,* and *C. americana* were harvested for metabolite analysis. Leaves were frozen in liquid nitrogen, crushed, and extracted for three hours in ethyl acetate. Leaf material was collected by centrifugation and the organic phase was removed and concentrated for GC-MS analysis.

GC-MS analysis

All GC-MS analyses were performed in Michigan State University's Mass Spectrometry and Metabolomics Core Facility on an Agilent 7890 A GC with an Agilent VF-5ms column (30 m × 250 µm × 0.25 µm, with 10 m EZ-Guard) and an Agilent 5975 C detector. The inlet was set to 250 °C splitless injection of 1 µL and He carrier gas (1 mL/min), and the detector was activated following a 3 min solvent delay. All assays and tissue analysis used the following method: temperature ramp start 40 °C, hold 1 min, 40 °C/min to 200 °C, hold 4.5 min, 20 °C/min to 240 °C, 10 °C/min to 280 °C, 40 °C/min to 320 °C, and hold 5 min. MS scan range was set to 40–400.

Product scale-up and NMR

For NMR analysis, production in the *N. benthamiana* system was scaled up to 1 L infection culture. A vacuum-infiltration system was used to infiltrate *A. tumefaciens* strains into whole *N.*

benthamiana plants, with approximately 40 plants used for each enzyme combination. The furan and lactone derivatives of CamTPS2 were identified from the combination of CamTPS2 and CbCYP76BK1. The furan derivative of ArTPS2 was identified from the combination of ArTPS2 and ArCYP76BK1, while the lactone derivative was identified from ArTPS2 with HsCYP76BK1. After 5 days, all leaf tissue was harvested and extracted overnight in 600 mL hexane at room temperature. The extract was concentrated by rotary evaporator. Each product was purified by silica gel flash column chromatography with a mobile phase of 98% hexane/2% ethyl acetate. NMR spectra were measured in Michigan State University's Max T. Rogers NMR Facility on a Bruker 800 MHz or 600 MHz spectrometer equipped with a TCl cryoprobe using $CDCl_3$ as the solvent. $CDCl_3$ peaks were referenced to 7.26 and 77.00 ppm for ^{1}H and ^{13}C spectra, respectively.

Availability of supporting information

Supplementary information is included as a separate folder with the following files.

Figures S1, S2. NMR analysis of **1**.

Figures S3, S4. NMR analysis of **2**.

Figures S5, S6. NMR analysis of **3**.

Figures S7, S8. NMR analysis of **4**.

Figure S9. GC-MS analysis of additional plant extracts.

Table S1. List of mint transcriptomes analyzed along with species abbreviations and subfamily.

Table S2. List of primers used in this study.

Table S3. GenBank accession numbers for genes cloned in this study.

Data S1. Reference diTPS and P450 sequences used in phylogenetic analysis.

REFERENCES

1. Peters, R. J. Two rings in them all: the labdane-related diterpenoids. *Nat. Prod. Rep.* **27**, 1521–1530 (2010).

2. Johnson, S. R. *et al.* Promiscuous terpene synthases from Prunella vulgaris highlight the importance of substrate and compartment switching in terpene synthase evolution. *New Phytol.* **223**, 323–335 (2019).

3. Bathe, U. & Tissier, A. Cytochrome P450 enzymes: A driving force of plant diterpene diversity. *Phytochemistry* **161**, 149–162 (2019).

4. CHEMnetBASE. https://dnp.chemnetbase.com/chemical/ChemicalSearch.xhtml?dswid=-6885.

5. Boachon, B. *et al.* Phylogenomic Mining of the Mints Reveals Multiple Mechanisms Contributing to the Evolution of Chemical Diversity in Lamiaceae. *Mol. Plant* **11**, 1084–1096 (2018).

6. Dettweiler, M. *et al.* A Clerodane Diterpene from Callicarpa americana Resensitizes Methicillin-Resistant Staphylococcus aureus to β-Lactam Antibiotics. *ACS Infect. Dis.* **6**, 1667–1673 (2020).

7. Klein Gebbinck, E. A., Jansen, B. J. M. & de Groot, A. Insect antifeedant activity of clerodane diterpenes and related model compounds. *Phytochemistry* **61**, 737–770 (2002).

8. Li, R., Morris-Natschke, S. L. & Lee, K.-H. Clerodane diterpenes: sources, structures, and biological activities. *Nat. Prod. Rep.* **33**, 1166–1226 (2016).

9. Kobayashi, J., Sekiguchi, M., Shimamoto, S., Shigemori, H. & Ohsaki, A. Echinophyllins C−F, New Nitrogen-Containing Clerodane Diterpenoids from Echinodorus macrophyllus. *J. Nat. Prod.* **63**, 1576–1579 (2000).

10. Bao, H., Zhang, Q., Ye, Y. & Lin, L. Naturally occurring furanoditerpenoids: distribution, chemistry and their pharmacological activities. *Phytochem. Rev.* **16**, 235–270 (2017).

11. Jia, Q. *et al.* Origin and early evolution of the plant terpene synthase family. *Proc. Natl. Acad. Sci.* **119**, e2100361119 (2022).

12. CHEMnetBASE. https://dmnp.chemnetbase.com/chemical/ChemicalSearch.xhtml?dswid=-7966.

13. Andersen-Ranberg, J. *et al.* Expanding the Landscape of Diterpene Structural Diversity through Stereochemically Controlled Combinatorial Biosynthesis. *Angew. Chem. Int. Ed.* **55**, 2142–2146 (2016).

14. Hamilton, J. P. *et al.* Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, Callicarpa americana. *GigaScience* **9**, giaa093 (2020).

15. Heskes, A. M. *et al.* Biosynthesis of bioactive diterpenoids in the medicinal plant Vitex agnus-castus. *Plant J.* **93**, 943–958 (2018).

16. Li, H. *et al.* The genomes of medicinal skullcaps reveal the polyphyletic origins of clerodane diterpene biosynthesis in the family Lamiaceae. *Mol. Plant* **0**, (2023).

17. Pelot, K. A. *et al.* Biosynthesis of the psychotropic plant diterpene salvinorin A: Discovery and characterization of the Salvia divinorum clerodienyl diphosphate synthase. *Plant J.* **89**, 885–897 (2017).

18. Johnson, S. R. *et al.* A database-driven approach identifies additional diterpene synthase activities in the mint family (Lamiaceae). *J. Biol. Chem.* **294**, 1349–1362 (2019).

19. Muchlinski, A. *et al.* Cytochrome P450-catalyzed biosynthesis of furanoditerpenoids in the bioenergy crop switchgrass (Panicum virgatum L.). *Plant J.* **108**, 1053–1068 (2021).

20. Kwon, M. *et al.* Cytochrome P450-Catalyzed Biosynthesis of a Dihydrofuran Neoclerodane in Magic Mint (*Salvia divinorum*). *ACS Catal.* **12**, 777–782 (2022).

21. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

22. Chaudhuri, P. K., Srivastava, R., Kumar, S. & Kumar, S. Phytotoxic and antimicrobial constituents of Bacopa monnieri and Holmskioldia sanguinea. *Phytother. Res.* **18**, 114–117 (2004).

23. Helfrich, E. & Rimpler, H. Iridoid glycosides and phenolic glycosides from Holmskioldia sanguinea. *Phytochemistry* **50**, 619–627 (1999).

24. Bergman, M. E., Bhardwaj, M. & Phillips, M. A. Cytosolic geraniol and citronellol biosynthesis require a Nudix hydrolase in rose-scented geranium (Pelargonium graveolens). *Plant J.* **107**, 493–510 (2021).

25. Babicki, S. *et al.* Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Res.* **44**, W147–W153 (2016).

26. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

27. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

28. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

29. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

30. Sainsbury, F., Thuenemann, E. C. & Lomonossoff, G. P. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* **7**, 682–693 (2009).

31. Englund, E., Andersen-Ranberg, J., Miao, R., Hamberger, B. & Lindberg, P. Metabolic Engineering of Synechocystis sp. PCC 6803 for Production of the Plant Diterpenoid Manoyl Oxide. *ACS Synth. Biol.* **4**, 1270–1278 (2015).

CHAPTER 5: KEY ENZYMATIC STEPS IN THE BIOSYNTHETIC ROUTE TOWARDS BIOACTIVE CLERODANES IN *CALLICARPA AMERICANA*

Emily R. Lanier, Trine B. Andersen, Nicholas J. Schlecht, Katheryn Van Winkle, Anh Phan, Bjoern R. Hamberger

Author Contributions:

ERL and BRH conceived and designed the study; ERL, TBA, NJS, AP, and KV performed the experiments; ERL and TBA analyzed the experimental data; ERL wrote the manuscript; BH and TBA supervised the project; all authors contributed to revisions.

Abstract

Callicarpa americana (American Beautyberry) is a plant in the mint (Lamiaceae) family, which is known for a diverse array of terpenoid compounds with uses including as fragrances, in culinary herbs, and for medicinal applications. The diterpenoids produced by *C. americana* are primarily clerodanes, and they have been shown to have bioactivities including the potent insect repellant callicarpenal, an antibiotic, and other compounds with cytotoxic activity. In this work we use two transcriptomics approaches to identify key pathway genes in the production of clerodanes. First we explore callicarpenal production in other *Callicarpa* species, identifying Mexican Beautyberry (*C. acuminata*) as a new source of this compound. Second we find that glandular trichomes in *C. americana* are an important tissue for clerodane production. Using trichome-specific transcriptomic data, we identify three short-chain dehydrogenases, CamOxr01, CamOxr03 and CamOxr08, as key pathway enzymes in catalyzing formation of furanoclerodane intermediates.

Introduction

The genus *Callicarpa*, which translates literally to its common name, "Beautyberry", is a rich source of bioactive compounds. There are over 150 recognized species in this genus whose native range extends from southeast Asia, where the majority are found, to Australia and the mid latitudes of the Americas [1]. Traditional Chinese medicine and other ethnobotanical traditions have incorporated roots, fruits, and leaves from many of these species into treatments for inflammation, hemorrhage, infection, pain, rheumatism, and other ailments [2,3]. Over the past two decades, phytochemical studies have catalogued hundreds of natural products from at least 20 different species and investigated potential therapeutic uses.

One of the most prevalent metabolite classes among reported bioactive constituents is the isoprenoids, or terpenes. This finding is in line with the classification of Callicarpa within the mint (Lamiaceae) family, known to be a prodigious source of terpenoids with a wide range of therapeutic and industrial uses. According to the Dictionary of Natural Products, there are 88 different diterpenoids (derived from 20-carbon backbone) reported from 16 different *Callicarpa* species [4]. This is over double the number reported in a 2008 review [3] and 20 more than reported in the most recent review of *Callicarpa* constituents in 2013 [2]. The most common types of diterpenes reported are the abietanes, phyllocladanes ((+)-kaurene derived), and clerodanes (Fig. 5.1), although some more unusual ring structures are present as well. Based on the common bicyclic core structure, most of these are likely products of labdane-type diterpene synthases (diTPSs). Labdanes are the most common class of diterpenes and typically proceed through a two-step route from the precursor geranyl geranyl diphosphate (GGPP) [5]. Cyclization is catalyzed by a class II diTPS, followed by dephosphorylation and often further structural rearrangement by

a class I diTPS. As is common with specialized diterpenoids, the compounds isolated from the plant generally contain multiple oxidations. These are catalyzed by cytochrome P450s and other oxidoreductases.

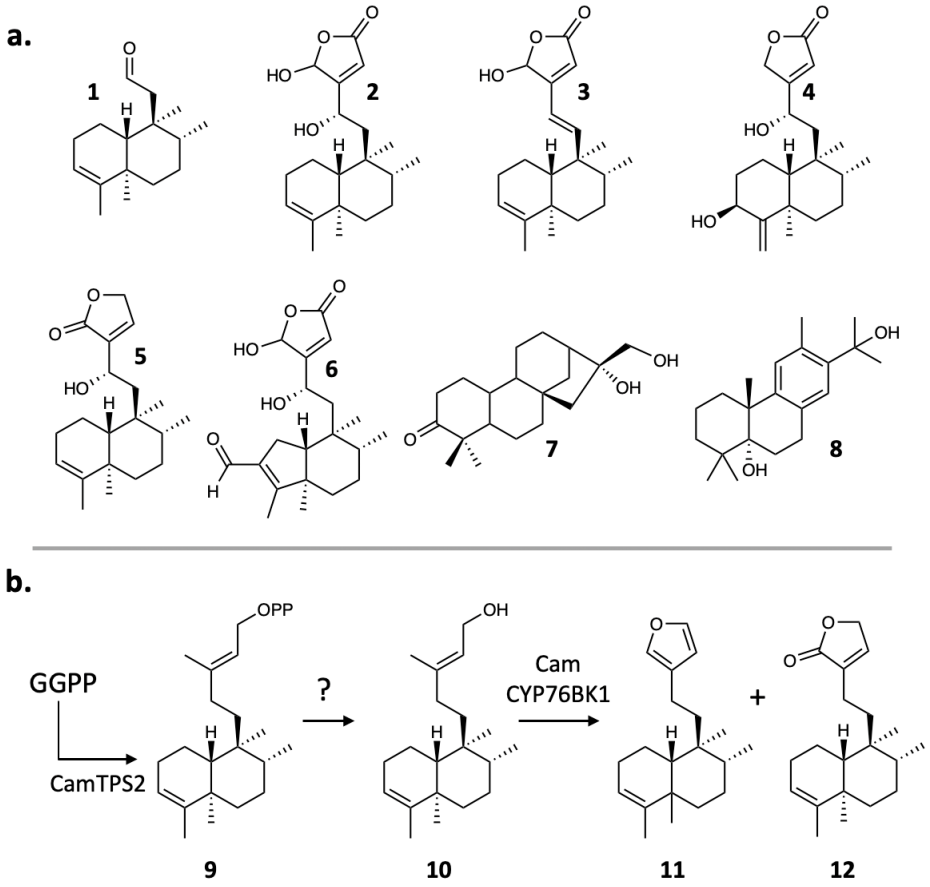


Figure 5.1. Representative set of diterpenoids in *C. americana.* (a) (**1-6**) clerodanes, of note are **1**, callicarpenal, and **2,** a clerodane found to be active against MRSA. **7** is the phyllocladane calliterpenone and **8** is an abietane identified in our previous work in *C. americana* [6]. (b) Previous work identified CamTPS2 as a kolavenyl diphosphate synthase (**9**), which is converted to kolavenol (**10**) by an unknown enzyme. CamCYP76BK1 converts **10** to a mixture of **11** (major product) and **12** (minor product)**.**

While the phytochemical studies have provided an excellent foundation for studying the specialized metabolites of the genus, little has been reported regarding the biosynthetic pathways towards these compounds. Until recently, such work was restricted by a lack of genetic

data. The publication of a chromosome-scale genome assembly for American beautyberry (*C. americana*) in 2020 provided the first significant genetic resource for this genus and enabled our work characterizing the four diterpene scaffold-forming enzymes, CamTPS1-3 and CamTPS6 [7]. This genome assembly and accompanying tissue-specific transcriptomic data also facilitated biosynthetic studies towards iridoids (a class of monoterpene derived compounds) [8]. In our most recent work, we investigated a biosynthetic gene cluster in *C. americana* containing 7 diTPSs and 12 cytochrome P450s, elucidating complete pathways to the abietane and phyllocladane backbones as well as identifying P450s involved in abietane pathways [6]. Genetic resources for other species however are still limited, as *C. bodinieri* is currently the only other species with available transcriptomic data on the NCBI short read archive (SRA) database [9].

From *C. americana* specifically, there are nine clerodane-type diterpenes and one phyllocladane reported (Fig. 5.1)[10], although our recent work suggests that there are also abietanes present in the roots of this species[6]. Useful bioactivities have been reported for at least three of these diterpenoids so far. One clerodane resensitizes methicillin-resistant *Staphylococcus aureus* (MRSA) to beta-lactam antibiotics, such as penicillin[11]. Another clerodane-derived diterpenoid, callicarpenal, has potent mosquito and tick repellant activities[12] and a unique 16-carbon structure. Calliterpenone, a phyllocladane, may have plant growth-promoting properties[13]. Only a few other species in the *Callicarpa* genus have reported clerodane diterpenoids. One study found callicarpenal in *C. japonica*[12], and others have documented clerodanes in *C. pentandra*[14] and *C. cathayana*[15].

In our initial investigation of the clerodane pathways in *C. americana* and other mint family species, we first identified CamTPS2 as a class II diTPS catalyzing formation of the clerodane

precursor kolavenyl diphosphate (KPP)[7]. Uniquely among labdane-type diterpene pathways, multiple reports in different species have found that clerodanes may proceed from the bicyclic labdane diphosphate precursor to a dephosphorylation performed by an as yet unidentified enzyme, possibly a phosphatase, rather than a canonical class I diTPS[16,17]. Consistent with these findings, no class I kolavenol synthase was found in *C. americana* among the labdane-type class I diTPS candidates (TPS-e subfamily). We then identified a P450, CamCYP76BK1, which is capable of catalyzing cyclization of a furan (major product) and lactone (minor product) ring from kolavenol (Chapter 4). While the furan is a plausible intermediate for the majority of the clerodanes reported from this plant, the lactone orientation matched that of **5**, and not the more industrially relevant **2.** Determining the full pathways for callicarpenal and the MRSA-active antibiotic clerodane would enable heterologous production of these potentially valuable compounds. Additionally, identifying these enzymes will expand the toolkit for biosynthesis of a range of oxidized furanoditerpenoid scaffolds.

In this work, we sought to elucidate additional enzymatic steps towards the biosynthesis of bioactive clerodanes in *C. americana*. To improve selection of gene candidates, we used metabolomic analysis to guide generation of additional transcriptomic data in a two-pronged approach. First, metabolomics supported RNA sequencing of two additional species based on evidence of callicarpenal and furanoclerodane production. These data were used for comparative transcriptomics, which has previously assisted with pathway elucidation[18,19]. Next, we identified glandular trichomes as a rich source of callicarpenal and other clerodanes in *C. americana*. Previously, the only trichomes reported in *Callicarpa* were stellate trichomes, which are structural hairs, rather than the glandular type which often produce terpenes[20]. In other

investigations of diterpene biosynthetic pathways, transcriptomic data for a highly productive tissue such as trichomes[21,22] or root cork cells in the case of forskolin[22] (Pateraki *et al.* 2017) was key for narrowing down candidate enzymes. The findings in our study supported trichome-specific RNA sequencing in *C. americana* as a second transcriptomic approach.

Assisted primarily by the trichome transcriptomic data, coexpression with previously identified pathway genes *CamTPS2* and *CamCYP76BK1* enabled identification and characterization of three oxidoreductase enzymes (Oxr) which catalyze formation of key intermediates in the clerodane pathway of *C. americana* when coexpressed alongside *CamCYP76BK1.* The first, CamOxr08, increases production of the lactone over the furan derivative. The others, CamOxr01 and CamOxr07, both catalyze formation of a furanoclerodane intermediate which is apparent in a leaf extract of *C. americana*. This compound is currently being analyzed by NMR for structural elucidation. We also found that an enzyme from *Clerodendrum bungeii*, CbCYP76BK1, can catalyze formation of callicarpenal in addition to **11** and **12** when expressed with *CamTPS2* in *Nicotiana benthamiana.* Together these findings enhance our understanding of clerodane biosynthesis in *C. americana* while expanding biotechnological access to the widely bioactive furanoclerodanes.

Results & Discussion

Clerodane production in other Callicarpa species

Metabolomics

Two cultivars of *C. japonica*, 'Leucocarpa' (L.C.) and 'Heavy Berry' (H.B.); *C. bodinieri* ('Profusion'), a native to China; and Mexican beautyberry (*C. acuminata*) were obtained for comparative analysis. *C. bodinieri* was chosen due to publicly available transcriptome data, *C. japonica* for the

report of callicarpenal, and Mexican beautyberry is the only other species native to the American continents, which we speculated could lead to similar specialized metabolite pathways. In the phytochemical literature, these species have only abietane diterpenoids reported aside from callicarpenal in *C. japonica*[23–25]. To determine which species produce callicarpenal, ethyl acetate extracts of young leaves from each plant were analyzed by GC-MS (Fig. 5.2). Because of its unique 16-carbon structure, callicarpenal can be identified by the presence of its molecular ion (234) and a similar fragmentation pattern to clerodane precursors. Comparison with the *C. americana* extract revealed that callicarpenal and other clerodanes (including **11**) are present in *C. acuminata,* but only a trace of callicarpenal could be detected in *C. japonica* H.B. and none in *C. japonica* L.C. No evidence for clerodanes could be found in the *C. bodinieri* extract.

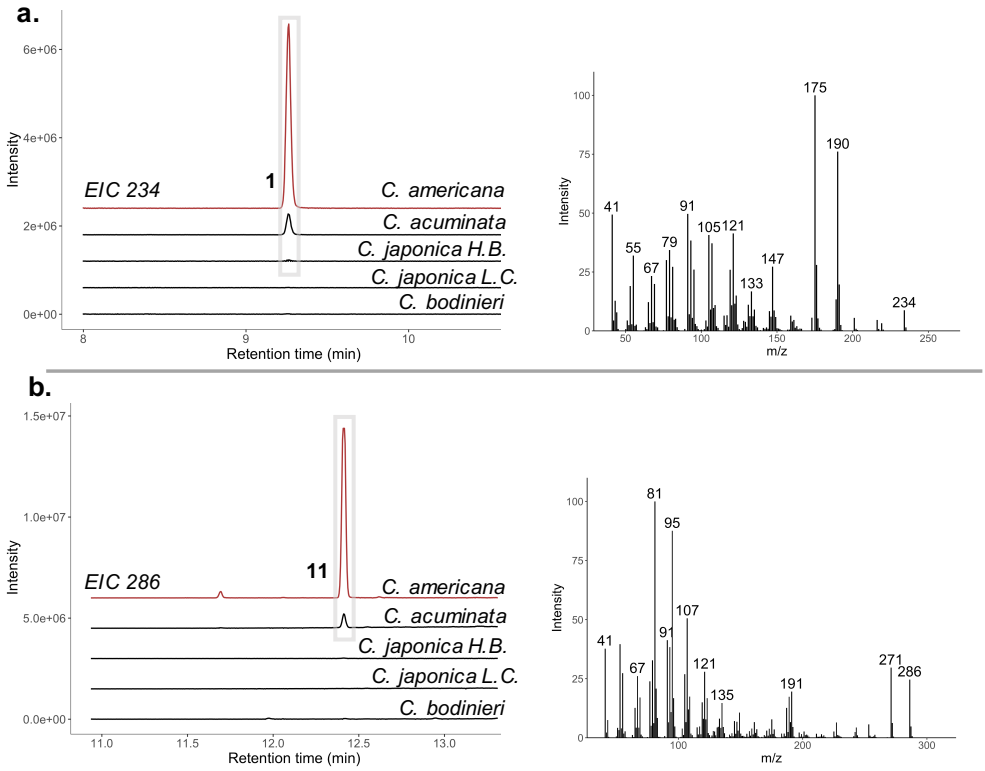


Figure 5.2. Metabolomics comparison between *Callicarpa* species. (a) Comparison of EIC 234, callicarpenal, and its mass spectra. (b) Comparison of EIC 286, **11**, and its mass spectra.

Transcriptomic data evaluation

On this basis, transcriptomic data were generated from young leaves and roots of *C. acuminata* and leaves of *C. japonica* H.B. The predicted peptides were compared with known biosynthetic pathway enzymes from *C. americana*, using homology with reference sequences to identify predicted diTPSs and P450s (Figure 5.3). Phylogenetic analysis identified possible orthologs to CamTPS2 in both *C. acuminata* and *C. japonica* H.B. Sequence alignment showed 99% similarity between CamTPS2 and the closest *C. acuminata* peptide. From *C. japonica* H.B., the longest assembled transcript orthologous to CamTPS2 was only 311 amino acids, making comparison with CamTPS2 (810 aa) difficult. *C. acuminata* also had a sequence matching *CamCYP76BK1* with 100% identity, though no ortholog was apparent from the *C. japonica* H.B. transcriptome.

As a method of further ensuring accuracy of the transcriptome assemblies, primers based on *CamTPS2* and *CamCYP76BK1* were used to clone orthologs from cDNA of *C. acuminata* and *C. japonica* H.B. This confirmed predicted sequences for *C. acuminata.* Despite the lack of clear orthologs in the *C. japonica* H.B. transcriptome, a *CYP76BK1* sequence (100% identical) and a diTPS (97% identical) were also identified. Functional characterization was carried out using agrobacterium-mediated transient expression in *N. benthamiana*, confirming the diTPS ortholog in *C. acuminata* as a KPP synthase while the *C. japonica* diTPS was inactive (Fig. 3).

The comparative transcriptomics approach relies on two plants having conserved biosynthetic pathways, while at the same time being evolutionarily distant enough that enzymes of other pathways are less conserved. While *C. japonica* and *C. americana* are more divergent, the ambiguous results of metabolomic and pathway gene analysis do not support definitively shared biosynthetic pathway enzymes. Conversely, *C. acuminata* has a similar metabolomic profile to *C.*

americana and shares the two known clerodane pathway enzymes. However, the phylogenetic analysis suggested a high degree of similarity between the two species, which fits with their overlapping geographical ranges and likely recent evolutionary divergence. This similarity limits the usefulness of transcriptomic comparison as a method for identifying shared biosynthetic pathway genes. The limitations of these two species lead us to pursue a second approach for better identification of clerodane pathway genes in *C. americana*.

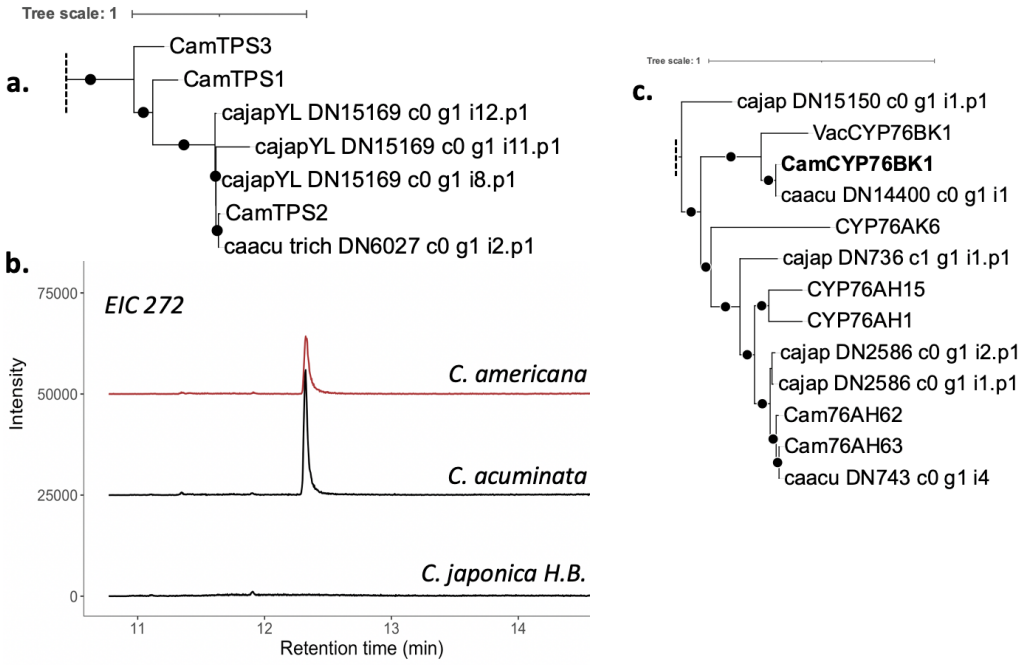


Figure 5.3. Analysis of clerodane pathway orthologs. Maximum likelihood trees of predicted peptides identify (a) orthologs of CamTPS2 and (c) orthologs of CamCYP76BK1. Bootstrap support of >70% (1000 repetitions) indicated by black dots. (b) GC-MS analysis of CamTPS2 orthologs, expressed in *N. benthamiana* along with the class I diTPS sclareol synthase [26] to yield kolavelool, an isomer of kolavenol.

Identification of glandular trichomes as a source of callicarpenal and other clerodanes

The leaves of *C. americana* were investigated under a light microscope for evidence of glandular trichomes. In addition to structural stellate trichomes, all had clear evidence of glandular trichomes as well. Scanning electron microscope images were generated for the glandular

trichomes of *C. americana*, revealing both capitate and peltate glandular trichomes on the cell surface (Fig. 5.4).

To determine whether these glandular trichomes are a source of callicarpenal, the leaf surface was washed briefly with ethyl acetate, and this extract was compared with a whole leaf extract. GC-MS analysis confirmed that callicarpenal is easily extracted from the leaf surface, and certain clerodane pathway intermediates are more prominent in the leaf surface wash than the whole leaf extract (Fig. 5.4). Based on these findings, trichomes were used to generate an additional tissue-specific transcriptomic dataset for *C. americana*. Analysis of this data confirmed the results of the metabolomic analysis, as both *CamTPS2* and *CamCYP76BK1* are highly expressed in the trichomes (Fig. 5.5).

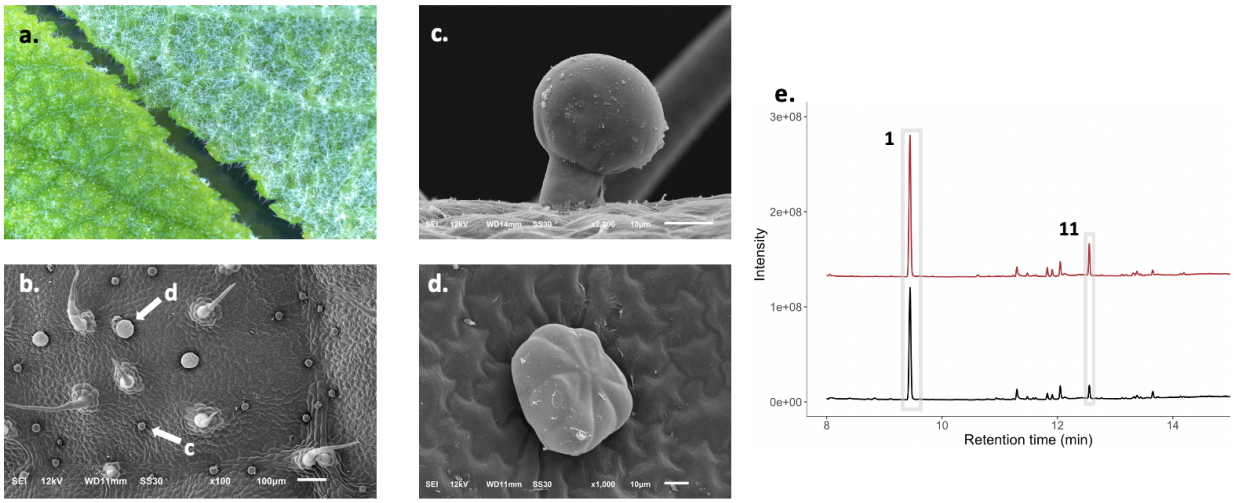


Figure 5.4. Microscopy of the *C. americana* leaf surface. (a) Light microscope images of top (left) and bottom (right) of a *C. americana* leaf, showing the fuzzy white stellate trichomes and small yellow dots of glandular trichomes. (b) SEM image of leaf surface showing two types of glandular trichomes, shown in detail in (c), capitate and (d), peltate. (e) GC-MS chromatogram (TIC) of leaf surface wash (top, red) and the extract of the washed leaf (bottom, black).

Evaluation of candidate P450s

Pathway candidate genes were ranked based on similarity of expression with *CamTPS2* and *CamCYP76BK1*. Nine tissue-specific datasets were available: mature leaf, young leaf, open flower, closed flower, petiole, stem, root, and fruit from a previous study [7], and trichome data from this work. Given the prevalence of P450s in diterpenoid oxidation pathways, this was the first enzyme class to be considered for testing. Candidate P450s were chosen based on Pearson's correlation coefficient (PCC)>0.5 for both bait genes as well as moderate trichome expression (TPM>10), low expression in non-target tissues (i.e., root and fruit), and a minimum predicted peptide length of 400 amino acids. Four candidates which fell slightly outside these criteria but had PCC>0.7 were also selected. This selection process yielded 18 candidate P450s. Furthermore, of the nine candidates originally picked in Chapter 3 during the identification of *CamCYP76BK1*, seven met the selection criteria with the addition of the trichome RNA data. For thoroughness, all nine were included. In total, 27 candidates from the CYP71, CYP72, and CYP96 clans were cloned for functional characterization (Fig. 5.5).

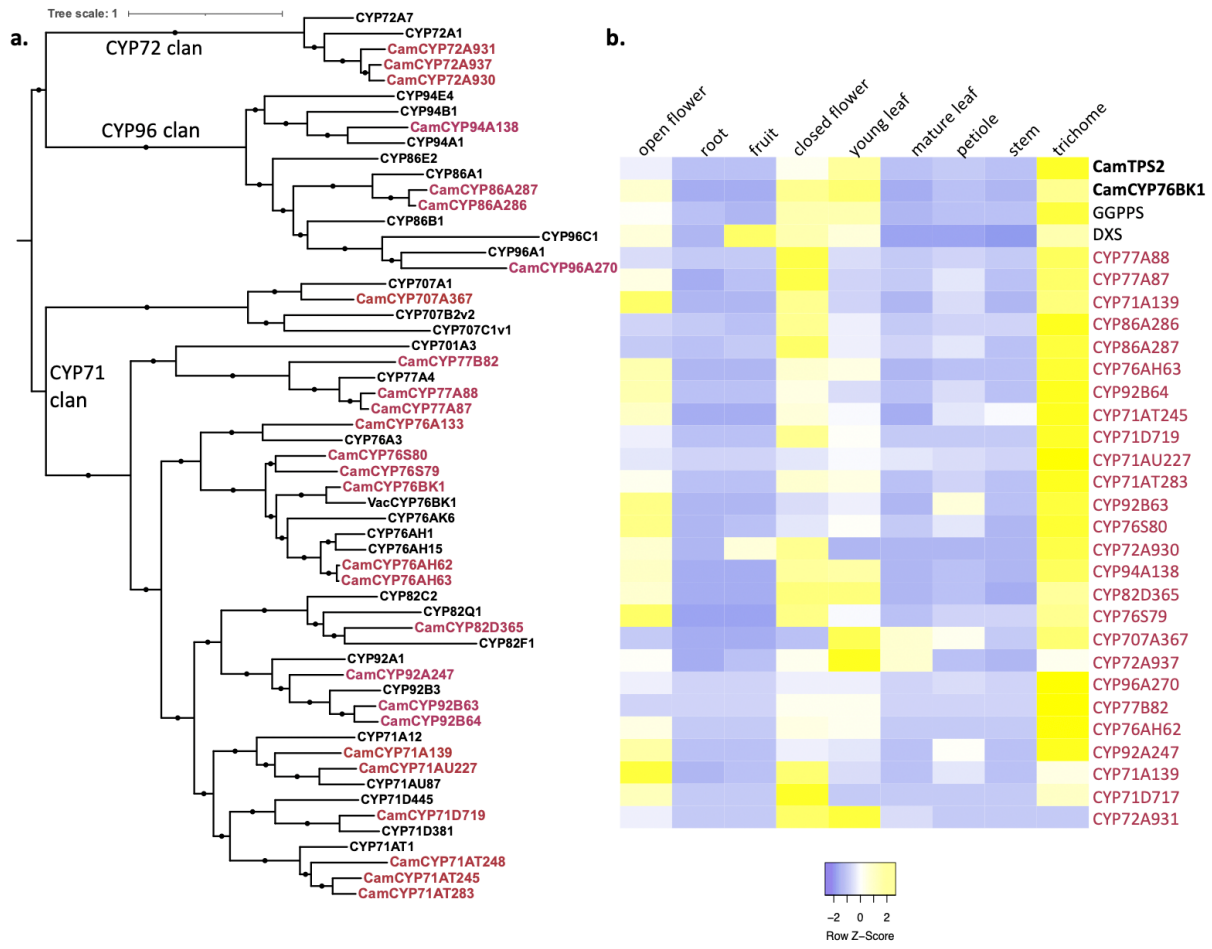


Figure 5.5. Selection of P450 candidates. (a) Maximum likelihood tree of candidate enzymes along with reference sequences to show clade and subfamily topology. Bootstrap support of >70% (1000 repetitions) indicated by black dots. Tree is rooted to *Arabidopsis thaliana* AtCYP701A3. (b) Heatmap of selected candidates along with previously identified clerodane pathway genes *CamTPS2*, *CamCYP76BK1*, and the upstream diterpenoid pathway genes *GGPP synthase* and *DXS* (deoxylulose 5-phosphate synthase).

Candidates were evaluated for activity with substrates kolavenol (*CamTPS2*) and its furan derivative (**11**, *CamTPS2* + *CamCYP76BK1*) in *N. benthamiana*. However, based on GC-MS analysis of leaf extracts, none of these P450s showed any activity with either combination. Previously, we observed that a *CYP76BK1* ortholog in *C. bungeii* could more effectively catalyze formation of a lactone (**12**) rather than furan ring when coexpressed with *CamTPS2* (Chapter 4). Given the additional hydroxylation present on **5**, we hypothesized that **12** may be a preferred substrate

over **11** for some P450s. All candidates were additionally tested with *CbCYP76BK1* + *CamTPS2*, but no activity was observed with this combination either. However, a closer inspection of the *CbCYP76BK1* + *CamTPS2* metabolite profile revealed a peak with the same mass spectrum and retention time as the callicarpenal peak in the plant extract (Fig. 5.6a). There are no reports of callicarpenal in *C. bungeii,* nor does analysis of the *C. bungeii* leaf extract yield any evidence of callicarpenal production. Despite this, CbCYP76BK1 is apparently capable of cleaving the bond between C12 and C13, possibly through oxidative cleavage such as is seen with CYP72A1, secologanin synthase[27]. This result offers the first biosynthetic access to callicarpenal, though *C. americana* may be relying on enzymes in addition to CamCYP76BK1 to complete this transformation.

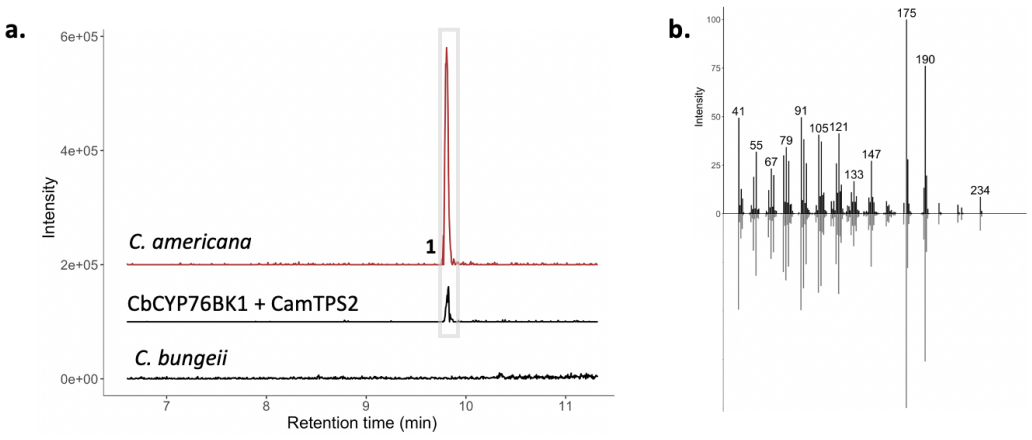


Figure 5.6. Formation of callicarpenal. (a) GC-MS chromatogram showing alignment of callicarpenal from *C. americana* leaf extract and an enzyme product of CbCYP76BK1 + CamTPS2. (b) Mass spectra match between callicarpenal from *C. americana* (bottom, gray) and from CbCYP76BK1 (top, black).

Oxidoreductase candidates beyond P450s

While P450s are typically the first oxidoreductase to react with hydrophobic terpene backbones due to their membrane-bound location, other oxidoreductases, such short chain

dehydrogenases/reductases (SDRs) and 2-oxoglutarate dehydrogenases (2OGDs), have also been identified [28,29]. Thus we expanded our testing to include other types of oxidoreductases. Candidate genes were selected by first filtering for PCC>0.7 with respect to the *CamCYP76BK1* expression profile. The higher cutoff was used based on the wider range of gene families to sort through. Predicted oxidoreductases were identified based on functional gene model annotations reported previously[7]. Discarding those with low trichome expression or high non-target tissue expression, 13 candidates were selected for functional characterization (*CamOxr01*-*CamOxr13*).

Identification of three SDRs with clerodane biosynthetic activity

All but one candidate (*CamOxr04*) were successfully cloned from trichome cDNA. As before, each candidate was expressed in *N. benthamiana* along with *CamTPS2* alone or in combination with *CamCYP76BK1*. GC-MS analysis showed that three candidates catalyzed production of a new product when coexpressed with *CamTPS2* + *CamCYP76BK1* (Fig. 5.7). Expression of *CamOxr08* mirrored the product profile of CbCYP76BK1, significantly increasing the amount of **12** produced relative to expression of *CamCYP76BK1* alone. The other two, CamOxr01 and CamOxr07, catalyzed production of the same product (**13**). This compound is also visible in the *C. americana* and *C. acuminata* leaf extract, confirming that this is a biologically relevant metabolite in the plant. Product scale-up and NMR analysis are currently being carried out to elucidate the structure. These sequences, annotated as "NAD(P)-binding Rossman-fold superfamily protein", can be classified as SDRs. The classification according to the 'SDR Nomenclature Initiative' is that CamOxr01 is part of the SDR 110C family, CamOxr07 is a member of the SDR7C family, and CamOxr08 is part of the SDR108E family [30]. All of these families are implicated in specialized metabolism in vascular plants [28].

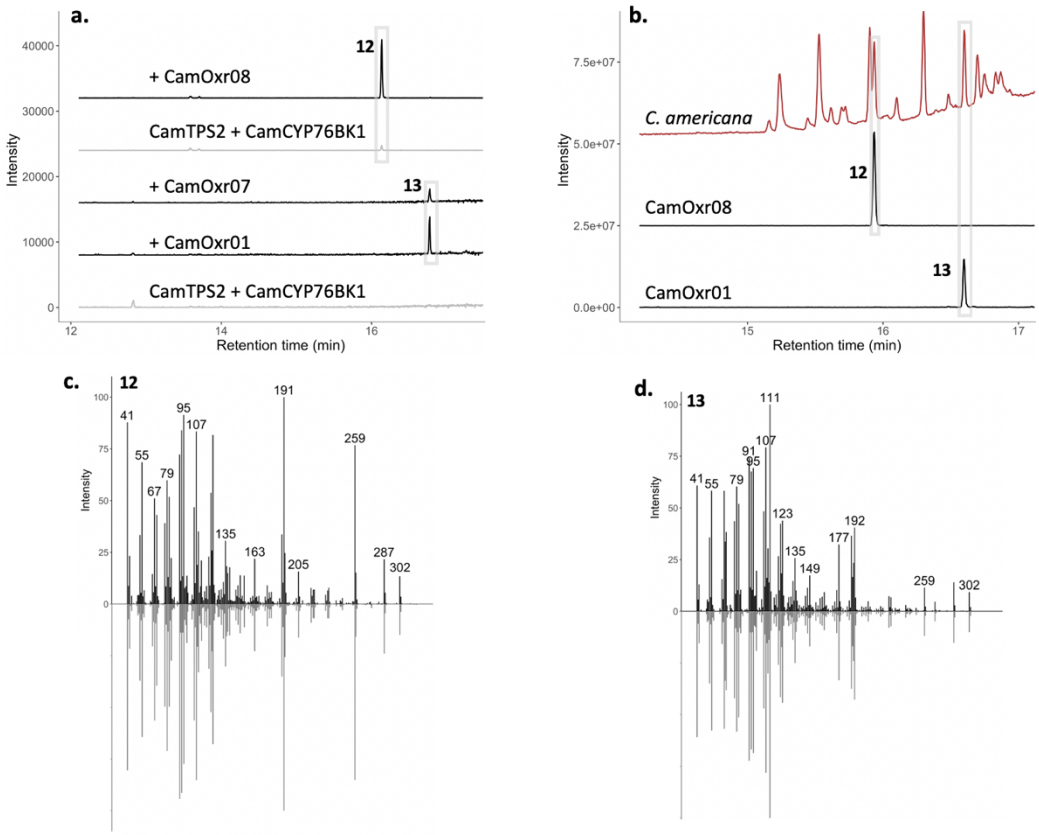


Figure 5.7. GC-MS analysis of oxidoreductase candidates. (a) GC-MS chromatograms showing new products formed by expression of *CamOxr01*, *CamOxr07*, or *CamOxr08* alongside *CamCYP76BK1* + *CamTPS2* (b) *C. americana* leaf extract compared with extracts from *N. benthamiana* expressing *CamCYP76BK1* + *CamTPS2* + *CamOxr01* or *CamOxr08*. Retention times match with **12** and **13,** and mass spectra (c) and (d) also match between peaks from *N. benthamiana* (top, black) and *C. americana* (bottom, gray).

Hierarchical clustering analysis

To verify that all high-priority candidates were included, hierarchical clustering analysis was applied to the set of genes with PCC>0.7 relative to *CamCYP76BK1* (Fig. 5.8). Hierarchical clustering analysis can provide insight into gene coexpression better than PCC alone by analyzing expression in an "all vs. all" approach and therefore identifying "clusters" of related genes[31]. Of 25 genes clustered with *CamTPS2*, there was one P450 pseudogene and two oxidoreductases, both of which had been tested. The cluster with *CamCYP76BK1* contained 33 genes, and of these

four were tested oxidoreductases (including *CamOxr01* and *CamOxr07*), two were tested P450s, and there was one P450 pseudogene. In the larger clade containing both the *CamTPS2* and *CamCYP76BK1* clusters, there were 177 enzymes total including 3 tested oxidoreductases, four untested oxidoreductases (one of which was *CamOxr04*), and 3 tested P450 candidates. In total, 6 of 27 P450s tested and 9 of 12 oxidoreductases tested clustered in close proximity to *CamTPS2* and *CamCYP76BK1*.

The set of 15 cloned candidate genes identified through hierarchical clustering analysis were expressed together along with *CamTPS2* and *CamCYP76BK1* in *N. benthamiana.* Some studies have shown that when working in *N. benthamia*, pathway flux may be increased with coexpression of a more complete pathway, enabling easier identification of end products [32–34]. However, analysis of this experiment did not show production of any additional oxidized clerodane products.

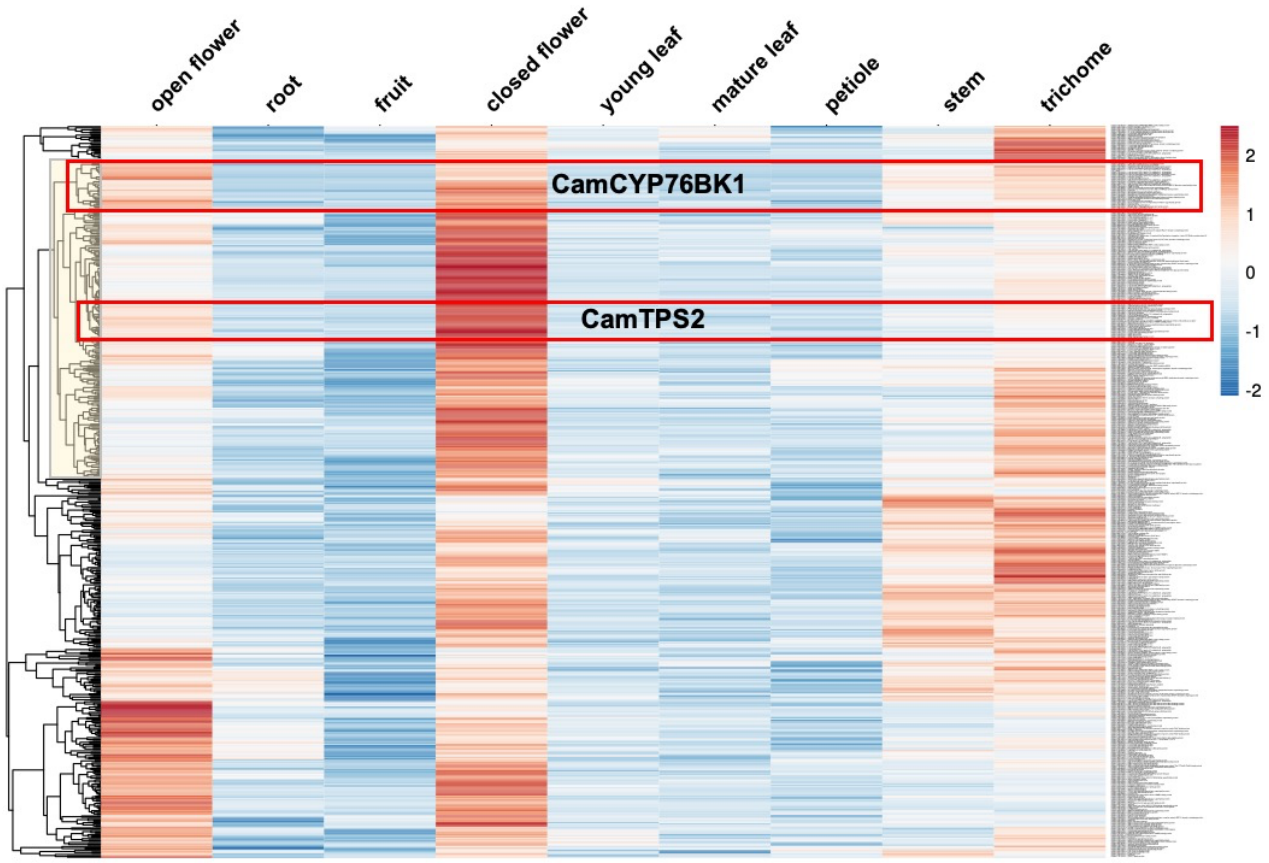


Figure 5.8. Hierarchical cluster analysis. All gene candidates with a PCC > 0.7 were subjected to hierarchical clustering analysis. The clusters containing *CamCYP76BK1* and *CamTPS2* are boxed in red. The larger cluster containing both sequences is highlighted in yellow.

Conclusion

The identification of trichomes as a source of clerodanes enabled discovery of three SDRs key to forming furanoclerodane products in *C. americana*. While the comparative transcriptomic approach was less helpful for clerodane pathway gene discovery, this work generated new data for future exploration of biosynthetic pathways in *C. japonica* and *C. acuminata*. Although extensive cloning of genes with similar tissue-specific expression patterns has not resulted in elucidation of the complete pathways to callicarpenal and the MRSA-active clerodane, it is still possible that one or more of the tested genes may be key in these pathways. Certain pathway

genes may be difficult to identify either due to differences in the cellular environment of *N. benthamiana* or because another key gene responsible for generation of the next intermediate is lacking. Finally, these experiments also found callicarpenal as an unexpected product of the enzyme CbCYP76BK1. While not confirming the pathway in the native *C. americana,* this discovery does demonstrate the first biosynthetic route to a powerful tick and mosquito repelling compound. Overall, this work was able to advance understanding of clerodane biosynthesis in *C. americana* while providing resources for future pathway discovery work.

Methods

Plant material

Plants were grown in a greenhouse under ambient photoperiod and 24 °C day/17 °C night temperatures. *C. acuminata* and *C. japonica* H.B. were ordered from Nurseries Caroliniana (North Augusta, SC), *C. japonica* L.C. was obtained from Joy Creek Nursery (Scappoose, OR), *C. bodinieri* was obtained from Garden Goods Direct (Bowie, MD) , and *C. americana* was obtained from Garden Gate Nursery (Gainesville, FL).

Transcriptomic Sequencing

For RNA-sequencing of *C. acuminata* and *C. japonica* H.B., 100 mg young leaf tissue from both and root tissue from *C. acuminata* were harvested and frozen in liquid nitrogen. RNA was isolated using Spectrum Plant Total RNA kit (Sigma) with on-column DNAse digest. TruSeq stranded mRNA (polyA mRNA) libraries were constructed and sequenced on an Illumina Novaseq 6000 to 150 nt in paired-end mode. Sequencing was performed at the Research Technology Support Facility at Michigan State University.

For RNA-sequencing of trichomes in *C. americana*, scotch tape was used to remove trichomes from the leaf surface for extraction. A 2 inch piece of tape was gently pressed to the underside of 3 fresh young leaves to remove trichomes (verified under light microscope), then cut into pieces and placed into 1 mL of lysis solution of the Invitrogen RNAqueous-Micro kit (Thermo-Fisher). The pieces were extracted for 30 minutes at room temperature before removing the tape and proceeding with the extraction. The solution was then loaded 500 µL at a time and washed with kit buffers, 500 µL each and eluted in 15 µL. Invitrogen Turbo DNAse (Thermo-Fisher) was used to remove trace DNA contamination. RNA was further concentrated using Genejet RNA cleanup (Thermo Scientific) and concentration micro kit. Library preparation and sequencing was performed by Novogene company[35], using Illumina Novaseq 6000 to 150 nt in paired-end mode.

Transcriptome assembly

Raw reads were trimmed and corrected using rCorrector (v. 1.0.4)[36] to remove erroneous k-mers TrimGalore (v. 0.6.6)[1,2][37] to trim the reads. Trimmed and corrected reads were assembled using Trinity (v. 2.1.1)[38]. Peptide sequences were predicted using TransDecoder (v. 5.5.0). For gene expression analysis, trimmed reads were mapped to respective peptides using Salmon 'index' (version 1.8.0), and quantified using Salmon 'quant' (libtype=A, validate mappings)[39] to obtain TPM values.

Candidate gene selection

Assembled transcriptomic and previously assembled genomic data[7] were used to identify candidate genes in *C. americana*. The heatmap of candidate gene expression was generated using Heatmapper[40]. Candidate P450s were identified based on 45% identity using BLASTP against a set of reference sequences (Supplementary File S1). Candidate oxidoreductases were identified

based on annotations from genomic assembly. Hierarchical clustering analysis was performed using the ClustVis web tool[41]. PCC was calculated using the 'CORREL' function in Excel.

Phylogenetic trees

Reference sequences used in all protein phylogenies are listed in Supplementary File S1. Full-length peptide sequences were used. Multiple sequence alignments were generated using ClustalOmega (v. 2.1) and phylogenetic trees were generated by RAxML (version 8.2.12; Model = protgammaauto; Algorithm = a) with support from 1000 bootstrap replicates (Sievers *et al.*, 2011; Stamatakis, 2014). Tree graphics were rendered using the Interactive Tree of Life (version 6.5.2) (Letunic and Bork, 2021).

Cloning

Synthetic oligonucleotides for all enzymes used in this study are given in Supplementary Table 1 along with GenBank accession numbers in Supplementary Table 2. Candidate enzymes were PCR-amplified from leaf or trichome cDNA, obtained from RNA using the SuperScript Double-Stranded cDNA synthesis kit (Thermo-Fisher). Coding sequences were cloned into the sequencing vector pJET using the CloneJET PCR Cloning kit (Thermo-Fisher) and sequence-verified with respective gene models. Constructs were then cloned into the plant expression vector pEAQ-HT[42].

Transient expression for functional characterization in N. benthamiana

N. benthamiana plants were grown for 5 weeks in a controlled growth room with 25 °C/18 °C day/night and 16h/8h light/dark before infiltration. Constructs for co-expression were separately transformed into *Agrobacterium tumefaciens* strain LBA4404. 20 mL cultures were grown overnight at 30 °C in LB with 50 µg/mL kanamycin and 50 µg/mL rifampicin. Cultures were collected by centrifugation and washed twice with 10 mL water. Cells were resuspended and

diluted to an OD_{600} of 1.0 in 200 µM acetosyringone/water and incubated at 30 °C for 1–2 H. Separate cultures were mixed in a 1:1 ratio for each combination of enzymes, and a 1 mL syringe was used to infiltrate 3 mL of culture into the the underside (abaxial side) of *N. benthamiana* leaves. All gene constructs were co-infiltrated with two genes from *Plectranthus barbatus* encoding rate-limiting steps in the upstream (MEP) pathway: 1-deoxy-D-xylulose-5-phosphate synthase (*PbDXS*) and GGPP synthase (*PbGGPPS*) to boost production of the diterpene precursor GGPP[26,43]. Plants were returned to the controlled growth room for 5 days. Approximately 200 mg fresh weight from infiltrated leaves was extracted with 1 mL hexane overnight at 18 °C. Plant material was removed by centrifugation, and the organic phase was transferred to a fresh vial for GC-MS analysis.

Plant extract metabolomics

For comparison between species, young leaves were harvested and placed in 5mL ethyl acetate for overnight extraction at room temperature. Leaf material was removed by centrifugation and organic layer was collected and concentrated for GC-MS analysis. For the trichome/leaf surface wash, a young leaf was held with forceps and washed with 2 mL ethyl acetate by repeatedly pipetting the solvent over the top and bottom leaf surfaces, 30 seconds each side. The same leaf was then placed in a fresh vial with another 2 mL ethyl acetate for 3 hours at room temperature. Organic layers were collected and analyzed by GC-MS without further concentration.

GC-MS analysis

All GC-MS analyses were performed in Michigan State University's Mass Spectrometry and Metabolomics Core Facility on an Agilent 7890 A GC with an Agilent VF-5ms column (30 m × 250 µm × 0.25 µm, with 10 m EZ-Guard) and an Agilent 5975 C detector. The inlet was set to

250 °C splitless injection of 1 µL and He carrier gas (1 mL/min), and the detector was activated following a 3 min solvent delay. All assays and tissue analysis used the following method: temperature ramp start 40 °C, hold 1 min, 40 °C/min to 200 °C, hold 4.5 min, 20 °C/min to 240 °C, 10 °C/min to 280 °C, 40 °C/min to 320 °C, and hold 5 min. MS scan range was set to 40–400.

Product scale-up and NMR analysis

For product purification and NMR analysis, production in the *N. benthamiana* system was scaled up to 1 L infection culture. A vacuum-infiltration system was used to infiltrate *A. tumefaciens* strains into whole *N. benthamiana* plants, with approximately 40 plants used for each enzyme combination. After 5 days, all leaf tissue was harvested and extracted overnight in 600 mL hexane at room temperature. The extract was concentrated by rotary evaporator. **13** was purified by silica gel flash column chromatography with a mobile phase of 98% hexane/2% ethyl acetate. NMR spectra will be measured in Michigan State University's Max T. Rogers NMR Facility on a Bruker 800 MHz or 600 MHz spectrometer equipped with a TCl cryoprobe using $CDCl_3$ as the solvent.

Microscopy

Light microscope images were taken of a young leaf at 32x magnification using a Leica S9i instrument with a 16x/15B Leica eye lens. For scanning electron microscopy, all sample preparation and image generation was carried out by the MSU Center for Advanced Microscopy. Briefy, cut into pieces, fixed, rinsed, dehydrated with ethanol, and critical point dried before coating with gold. Images were generated using a JEOL JSM-6610LV SEM instrument.

Availability of supporting information

The transcriptomic data generated in this study can be found in the NCBI SRA database under Bioproject PRJNA947623.

Supporting information is included as a separate folder with the following files.

Table S1. List of primers used in this study.

Table S2. GenBank accession numbers for genes cloned in this study.

Data S1. Reference diTPS and P450 sequences used in phylogenetic analysis.

REFERENCES

1. Callicarpa L. | Plants of the World Online | Kew Science. *Plants of the World Online* http://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:30044566-2.

2. Tu, Y., Sun, L., Guo, M. & Chen, W. The medicinal uses of Callicarpa L. in traditional Chinese medicine: An ethnopharmacological, phytochemical and pharmacological review. *J. Ethnopharmacol.* **146**, 465–481 (2013).

3. Jones, W. P. & Kinghorn, A. D. BIOLOGICALLY ACTIVE NATURAL PRODUCTS OF THE GENUS CALLICARPA. *Curr. Bioact. Compd.* **4**, 15 (2008).

4. CHEMnetBASE. https://dnp.chemnetbase.com/chemical/ChemicalSearch.xhtml?dswid=-2358.

5. Peters, R. J. Two rings in them all: the labdane-related diterpenoids. *Nat. Prod. Rep.* **27**, 1521–1530 (2010).

6. Bryson, A. E. *et al.* Uncovering a miltiradiene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory. *Nat. Commun.* **14**, 343 (2023).

7. Hamilton, J. P. *et al.* Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, Callicarpa americana. *GigaScience* **9**, giaa093 (2020).

8. Rodríguez-López, C. E. *et al.* Phylogeny-Aware Chemoinformatic Analysis of Chemical Diversity in Lamiaceae Enables Iridoid Pathway Assembly and Discovery of Aucubin Synthase. *Mol. Biol. Evol.* **39**, msac057 (2022).

9. Kodama, Y., Shumway, M., Leinonen, R., & on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).

10. Jones, W. P. *et al.* Cytotoxic Constituents from the Fruiting Branches of Callicarpa americana Collected in Southern Florida,1. *J. Nat. Prod.* **70**, 372–377 (2007).

11. Dettweiler, M. *et al.* A Clerodane Diterpene from Callicarpa americana Resensitizes Methicillin-Resistant Staphylococcus aureus to β-Lactam Antibiotics. *ACS Infect. Dis.* **6**, 1667–1673 (2020).

12. Cantrell, C. L., Klun, J. A., Bryson, C. T., Kobaisy, M. & Duke, S. O. Isolation and Identification of Mosquito Bite Deterrent Terpenoids from Leaves of American (Callicarpa americana) and Japanese (Callicarpa japonica) Beautyberry. *J. Agric. Food Chem.* **53**, 5948–5953 (2005).

13. Pandey, P. *et al.* Calliterpenone, a natural plant growth promoter from a medicinal plant Callicarpa macrophylla, sustainably enhances the yield and productivity of crops. *Front. Plant Sci.* **13**, (2022).

14. Xu, J., Harrison, L. J., Vittal, J. J., Xu, Y.-J. & Goh, S.-H. Four New Clerodane Diterpenoids from *Callicarpa pentandra*. *J. Nat. Prod.* **63**, 1062–1065 (2000).

15. Gong, S. *et al.* Cathayanalactone G and other constituents from leaves and twigs of Callicarpa cathayana. *Chin. Herb. Med.* **14**, 332–336 (2022).

16. Muchlinski, A. *et al.* Cytochrome P450-catalyzed biosynthesis of furanoditerpenoids in the bioenergy crop switchgrass (Panicum virgatum L.). *Plant J.* **108**, 1053–1068 (2021).

17. Kwon, M. *et al.* Cytochrome P450-Catalyzed Biosynthesis of a Dihydrofuran Neoclerodane in Magic Mint (*Salvia divinorum*). *ACS Catal.* **12**, 777–782 (2022).

18. Miller, G. P. *et al.* The biosynthesis of the anti-microbial diterpenoid leubethanol in Leucophyllum frutescens proceeds via an all-cis prenyl intermediate. *Plant J.* **104**, 693–705 (2020).

19. Wisecaver, J. H. *et al.* A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **29**, 944–959 (2017).

20. Martin, C. O. & Mott, S. P. American Beautyberry (Callicarpa americana). *U.S Army Corps of Engineers Wildlife Resources Management Manual* **Section 7.5.8**, (1997).

21. Heskes, A. M. *et al.* Biosynthesis of bioactive diterpenoids in the medicinal plant Vitex agnus-castus. *Plant J.* **93**, 943–958 (2018).

22. Pelot, K. A. *et al.* Biosynthesis of the psychotropic plant diterpene salvinorin A: Discovery and characterization of the Salvia divinorum clerodienyl diphosphate synthase. *Plant J.* **89**, 885–897 (2017).

23. Anaya, A. L. *et al.* Allelochemical Potential of Callicarpa acuminata. *J. Chem. Ecol.* **29**, 2761–2776 (2003).

24. Ono, M. *et al.* A new diterpenoid and a new triterpenoid glucosyl ester from the leaves of Callicarpa japonica Thunb. var. luxurians Rehd. *J. Nat. Med.* **63**, 318–322 (2009).

25. Gao, J.-B. *et al.* Isolation, Characterization, and Structure–Activity Relationship Analysis of Abietane Diterpenoids from Callicarpa bodinieri as Spleen Tyrosine Kinase Inhibitors. *J. Nat. Prod.* **81**, 998–1006 (2018).

26. Andersen-Ranberg, J. *et al.* Expanding the Landscape of Diterpene Structural Diversity through Stereochemically Controlled Combinatorial Biosynthesis. *Angew. Chem. Int. Ed.* **55**, 2142–2146 (2016).

27. Irmler, S. *et al.* Indole alkaloid biosynthesis in Catharanthus roseus: new enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. *Plant J.* **24**, 797–804 (2000).

28. Moummou, H., Kallberg, Y., Tonfack, L. B., Persson, B. & van der Rest, B. The Plant Short-Chain Dehydrogenase (SDR) superfamily: genome-wide inventory and diversification patterns. *BMC Plant Biol.* **12**, 219 (2012).

29. Song, J.-J. *et al.* A 2-oxoglutarate-dependent dioxygenase converts dihydrofuran to furan in Salvia diterpenoids. *Plant Physiol.* **188**, 1496–1506 (2022).

30. Persson, B. *et al.* The SDR (Short-Chain Dehydrogenase/Reductase and Related Enzymes) Nomenclature Initiative. *Chem. Biol. Interact.* **178**, 94–98 (2009).

31. Delli-Ponti, R., Shivhare, D. & Mutwil, M. Using Gene Expression to Study Specialized Metabolism—A Practical Guide. *Front. Plant Sci.* **11**, (2021).

32. Dudley, Q. M. *et al.* Reconstitution of monoterpene indole alkaloid biosynthesis in genome engineered Nicotiana benthamiana. *Commun. Biol.* **5**, 1–12 (2022).

33. Pateraki, I. *et al.* Total biosynthesis of the cyclic AMP booster forskolin from Coleus forskohlii. *eLife* **6**, e23001 (2017).

34. Hansen, N. L. *et al.* Tripterygium wilfordii cytochrome P450s catalyze the methyl shift and epoxidations in the biosynthesis of triptonide. *Nat. Commun.* **13**, 5011 (2022).

35. Research Services. *Novogene* https://www.novogene.com/us-en/services/research-services/.

36. Song, L. & Florea, L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* **4**, 48 (2015).

37. Babraham Bioinformatics - Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

38. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

39. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

40. Babicki, S. *et al.* Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Res.* **44**, W147–W153 (2016).

41. Metsalu, T. & Vilo, J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Res.* **43**, W566–W570 (2015).

42. Sainsbury, F., Thuenemann, E. C. & Lomonossoff, G. P. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* **7**, 682–693 (2009).

43. Englund, E., Andersen-Ranberg, J., Miao, R., Hamberger, B. & Lindberg, P. Metabolic Engineering of Synechocystis sp. PCC 6803 for Production of the Plant Diterpenoid Manoyl Oxide. *ACS Synth. Biol.* **4**, 1270–1278 (2015).

CHAPTER 6: SYNTHESIS

Emily R. Lanier

Summary

The biosynthesis pathways of plant specialized metabolites are incredibly complex, and our scientific understanding of them has only just begun to uncover the full story. In this dissertation, I focus on uncovering the enzymes involved in diterpenoid biosynthesis in *Callicarpa americnana* and the wider Lamiaceae (mint) family. As the Lamiceae is the largest reservoir of plant terpenoids[1], studying these pathways is important to improve both understanding of and biotechnological access to an exceptionally wide range of diterpenoids. In this work we identify crucial pathway enzymes. We also find, as outlined in Chapter 1, that these pathways resemble a complex network far more than straightforward linear pathways. Our findings here advance our knowledge of a class of compounds which are crucial for the survival and flourishing of both plants and humans.

Future Directions

Chapter 2

This work begins with the assembly of a chromosome scale genome of *C. americana*, the first in this genus and the subfamily Callicarpoideae. I used the gene models generated here to identify four class II diterpene synthases (diTPSs) which are the entry point to all labdane-type diterpenes in this plant. I found that these enzymes catalyzed formation of *ent*-copalyl diphosphate (*ent*-CPP), the precursor to giberellins in primary metabolism; (+)-CPP, the precursor to the plant growth promoter calliterpenone; and kolavenyl diphosphate (KPP), the precursor to the clerodanes which are the most well studied diterpenoids in this plant. The genome assembly, annotations, and tissue-specific transcriptome data generated in this chapter facilitated the remainder of my work studying *C. americana.* In particular, the high

quality of the genome assembly enabled discovery of an exceptionally large diterpene biosynthetic gene cluster (BGC), which is the basis for Chapter 3. This genome assembly will continue to be a resource for future studies of all plant specialized metabolism, such as recent work investigating iridoid biosynthesis in *C. americana* [2]. The class II diTPSs characterized here contribute to the remainder of my work while also serving as a point of reference for continued study of the evolution of diterpenoid biosynthesis in the Lamiaceae[3].

Chapter 3

The expansion of genomic sequencing beyond model species has resulted in surprising discoveries about the organization of plant genomes. BGCs were first thought to exist only in microbes and fungi, where they allow sharing of complete metabolic pathways during horizontal gene transfer. However, evidence is growing that pathway gene clustering is important for plants, too. We found that the BGC in *C. americana* shared synteny with similar BGCs throughout the Lamiaceae family and predicted a minimal gene cluster that likely originated with an ancient Lamiaceae progenitor. Functional characterization of the BGC revealed bifunctionality, with two diterpene backbones generated by genes in differently expressed modules. This work demonstrated the importance of BGCs as an evolutionary strategy for diterpene biosynthesis across the Lamicaeae. While we also found evidence that this cluster may be conserved in at least one other species in the order Lamiales, a more in-depth investigation of other Lamiales genomes could reveal the BGC as even more widely conserved than just the family level. Further study of plant BGCs is needed to understand both their prevalence as well as how and why they form. Identification of regulatory factors controlling expression of BGC genes could shed light on how they contribute to plant fitness.

Additionally, BGCs are a new starting point for biosynthetic studies. While in the past most pathway investigations began with a known target compound, the identification of BGCs presents an opportunity for genomics-led pathway discovery. In this chapter, functional characterization of the BGC led to finding abietane diterpenoids in *C. americana*, which had not previously been reported. The syntentic BGCs we found in other species also represent a rich resource for additional pathway elucidation studies for compounds as yet undiscovered.

Chapter 4

Furanoditerpenoids are a group of widely bioactive diterpenoids which are also prevalent in Lamiaceae species [4]. Most are derived from the clerodane backbone, like the bioactive diterpenoids found in *C. americana*. Using transcriptomic and genomic data from Chapter 2, I found that the cytochrome P450 CamCYP76BK1 catalyzes formation of furan and lactone derivatives of kolavenol, the dephosphorylated product of CamTPS2. We further investigated a representative set of 48 Lamiaceae transcriptomes and found that CYP76BK1 enzyme orthologs are present in 9 other Lamiaceae species, representing all subfamilies which are known to produce furanoclerodanes as well as one (Viticoideae) which makes furanolabdanes. This set of enzymes is able to catalyze furan and lactone formation from both kolavenol and isokolavenol, although substrate preference differs between enzymes. Further work is needed to understand the enzyme mechanism as well as how sequence and structure dictate functional differences. Labeling studies tracing the incorporation of $^{18}O_2$, such as performed with the biosynthesis of a dihydrofuan by the *Salvia divinorum* enzyme CYP76AH39, could clarify the mechanism. Structural modeling of these enzymes coupled with site-directed mutagenesis could assist with identification of key residues which determine substrate and catalytic promiscuity. Finally,

although this enzyme is likely key in the biosynthesis of a wide array of furanoditerpenoids, additionally pathway enzymes are needed to reach the most potent bioactive products. In the Ajugoideae family, my coauthors will continue this work. I address this gap in *C. americana* in Chapter 5.

Chapter 5

The bioactive clerodanes reported from *C. americana* are attractive targets for determining methods of heterologous biosynthesis. The 16 carbon structure of callicarpenal is also a unique structural target as it implies a biosynthetic enzyme capable of oxidative cleavage, a rare but not unprecedented activity[5–7]. Building on previous work from Chapters 2 and 4, I sought additional biosynthetic enzymes involved in the formation of these bioactive clerodanes. In the process, I found that callicarpenal and other clerodanes are present in Mexican Beautyberry, which has not previously been reported to contain clerodanes. Moreover, I identified trichomes as a rich source of clerodanes within *C. americana.* Trichome-specific transcriptomic data enabled discovery of three short-chain dehydrogenases involved in furanoclerodane biosynthesis. Additionally, I found that the CYP76BK1 enzyme from *Clerodendrum bungeii* is able to catalyze formation of callicarpenal, which was unexpected given the lack of callicarpenal in the plant extract. While this work advanced our understanding of furanoclerodane biosynthesis in *C. americana*, the full pathways have not yet been fully elucidated. Within the scope of this study, a few more steps can be taken to ensure all possible progess has been made. Once the structure has been elucidated for the product of CamOxr01/CamOxr07, possible next steps can be hypothesized if this product fits with the structures of the bioactive clerodanes. All previously cloned enzymes can be retested with this product to identify next

steps. Additionally, the oxidoreductases identified during hierarchical clustering analysis can be cloned and tested. Beyond this work, additional transcriptomic experiments may be useful in further elucidating the pathway. Time-course experiments in response to metabolite elicitation have often proven useful in pathway discovery projects. The callicarpenal activity of CbCYP76BK1 could also be explored further. Other labdane backbones with this same oxidation pattern have been identified as potent fragrance molecules[8]. CbCYP76BK1 is potentially a tool for heterologous production of valuable terpene derivatives if it could be engineered for high activity with other diterpene substrates.

REFERENCES

1. Boachon, B. *et al.* Phylogenomic Mining of the Mints Reveals Multiple Mechanisms Contributing to the Evolution of Chemical Diversity in Lamiaceae. *Mol. Plant* **11**, 1084–1096 (2018).

2. Rodríguez-López, C. E. *et al.* Phylogeny-Aware Chemoinformatic Analysis of Chemical Diversity in Lamiaceae Enables Iridoid Pathway Assembly and Discovery of Aucubin Synthase. *Mol. Biol. Evol.* **39**, msac057 (2022).

3. Li, H. *et al.* The genomes of medicinal skullcaps reveal the polyphyletic origins of clerodane diterpene biosynthesis in the family Lamiaceae. *Mol. Plant* **0**, (2023).

4. Bao, H., Zhang, Q., Ye, Y. & Lin, L. Naturally occurring furanoditerpenoids: distribution, chemistry and their pharmacological activities. *Phytochem. Rev.* **16**, 235–270 (2017).

5. Irmler, S. *et al.* Indole alkaloid biosynthesis in Catharanthus roseus: new enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. *Plant J.* **24**, 797–804 (2000).

6. Guengerich, F. P. Mechanisms of Cytochrome P450-Catalyzed Oxidations. *ACS Catal.* **8**, 10964–10976 (2018).

7. Doungchawee, J. *et al.* Volatile Chemical Composition, Antibacterial and Antifungal Activities of Extracts from Different Parts of Globba schomburgkii Hook.f. *Chem. Biodivers.* **16**, e1900057 (2019).

8. De la Torre, M. C., García, I. & Sierra, M. A. Straightforward synthesis of the strong ambergris odorant γ-bicyclohomofarnesal and its endo-isomer from R-(+)-sclareolide. *Tetrahedron Lett.* **43**, 6351–6353 (2002).