Uncovering a miltiradiene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory

Abigail Bryson
Michigan State University https://orcid.org/0000-0003-4176-0084
Emily Lanier
Michigan State University https://orcid.org/0000-0002-1878-4566
Kin Lau
Michigan State University
Davis Mathieu
Michigan State University
Alan Yocca
Michigan State University https://orcid.org/0000-0002-0974-364X
Garret Miller
Michigan State University
Patrick Edger
Michigan State University
Carol Robin Buell
University of Georgia
Björn Robert Hamberger (hamberge@msu.edu)
Michigan State University https://orcid.org/0000-0003-1249-1807

Article

Keywords:

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1535494/v1

Version of Record: A version of this preprint was published at Nature Communications on January 20th, 2023. See the published version at https://doi.org/10.1038/s41467-023-35845-1.

Abstract

The spatial organization of genes within plant genomes can drive evolution of specialized metabolic pathways. In this study we investigated the origin and subsequent evolution of a diterpenoid biosynthetic gene cluster (BGC) present throughout the Lamiaceae (mint) family. Terpenoids are important specialized metabolites in plants with diverse adaptive functions that enable environmental interactions, such as chemical defense. 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 novel 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 novel metabolites.

Introduction

Significance of biosynthetic gene clusters

Plants are renowned for their incredible diversity of specialized metabolites, which they use to interact with and interpret their environment. These biosynthetic pathways are dynamic, facilitating constant 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 close, 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 range across 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].

Terpenoid metabolism

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]. People have harnessed them for their medicinal properties[44–47] as well as for flavors, fragrances, and perfumes[48–51]. 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[41,52,53]. The bicyclic labdane-type diterpenes are typically formed by the sequential activity of a class II followed by a class I 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[54,55].

Terpenoid diversity is particularly rich in the Lamiaceae (mint) family[56,57], one of the largest families of angiosperms including aromatic herbs such as peppermint, basil, and sage. Genome assemblies for approximately 20 different species (Supplemental Table 1) have been published to date, revealing BGCs for at least two classes of terpenoids: monoterpene-derived nepetalactones from catnip (*Nepeta* sp.)[58] and diterpenoid tanshinones in the Chinese medicinal herb Danshen (*Salvia miltiorrhiza*)[24,59,60]. Tanshinones are well studied for their potent pharmacological activities, and as a result much of the biosynthetic pathway has been elucidated (Fig. 1)[24,59–68]. 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 abietadiene-type miltiradiene is the likely terpene precursor to a wide array of bioactive diterpenoids that are common throughout the Lamiaceae and

beyond[69]. The antimicrobial effects demonstrated for many of these terpenoids suggest a native role in plant defense, although most studies have focused on their numerous medicinal uses[69–73]. Carnosic acid is another abietane diterpenoid found in several Lamiaceae species with powerful antioxidant and anticancer properties[74]. The biosynthesis of carnosic acid and related diterpenoids has been elucidated in *Rosmarinus officinalis, Salvia pomifera* and *Salvia fruticosa* (rosemary and sages)[75,76] and involves many CYPs orthologous to those involved in tanshinone biosynthesis (Fig. 1).

The miltiradiene gene cluster

Previous studies of the *S. miltiorrhiza* genome have found two BGCs that together contain the miltiradiene diTPSs and two CYP76AHs involved in tanshinone biosynthesis[24,59,60]. A third locus containing an array of CYP71Ds includes the two 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)[24,67,77]. Additionally, we previously reported the presence of a large cluster in *Callicarpa americana* (American beautyberry) which contains orthologs of the miltiradiene diTPSs as well as multiple CYP76AHs and CYP71Ds[78]. The divergence of these five species indicates that this BGC may be present ubiquitously throughout the Lamiaceae.

To explore the prevalence and evolution of the miltiradiene BGC, we surveyed a representative panel of 10 Lamiaceae genome assemblies (Table 1). We focused 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 showed 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 enabled reconstruction of a minimal ancestral cluster. Functional characterization of the *C. americana* BGC revealed the presence of two discretely regulated diterpene pathways. We found that this BGC has evolved bifunctionality, providing the scaffold of the previously unidentified diterpene (+)-kaurene in addition to miltiradiene. This opens new biosynthetic avenues towards novel diterpenes in addition to highlighting an instance of BGC bifunctionality, which is rarely observed in plants[10,79]. We also discovered complex miltiradiene BGCs in four additional species, laying the foundation for the elucidation of new 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.

Results

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 (Fig. 2). We chose our genome panel based on their assembly quality and contiguity as well as subfamily representation (i.e., phylogenetic placement). In addition to three species with previously reported syntenic BGCs, we selected four species with published genomes. To increase the diversity of representatives across the phylogeny, we sequenced three new genomes (Supplemental Table 2, Supplemental Table 3). 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[80–82].

Out of the 10 species sampled, all contained diTPSs orthologous to known (+)-CPP and miltiradiene synthases. In seven species these diTPSs were within syntenic BGCs (Fig. 3). The genomes of *Prunella vulgaris, Plectranthus 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, Hyssopus officinalis*, and *Leonotis leonurus*, also had at least one copy of a *CYP71D* gene.

Comparison of the BGCs provides insight into the formation and maintenance of this cluster in divergent lineages (Fig. 3). 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 presumably present from past tandem duplications. 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 *Pogostemon cablin* and *L. leonurus* are compact and gene dense. The presence of two related BGCs in both *S. miltiorrhiza* and *L. leonurus* may award the plant evolutionary flexibility with a duplicated pathway. It is evident that each BGC, while maintaining the core miltiradiene genes, has assembled and disassembled in a lineage-specific manner.

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 71D clade where few characterized Lamiaceae sequences are available (Fig. 4, Supplemental Table 4). 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.

As expected, syntenic diTPSs in both subfamilies have common ancestry. Recent tandem duplications in the TPS-e and TPS-c families are evident in all examined species and contribute to lineage-specific BGC

expansion (Fig. 4, Fig. 5). 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 diTPSs which yield labdanes in the (+)-configuration. The two divergent class I 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.

Consistent with their expected role in specialized metabolism, no BGC class I enzymes clustered in clade TPS-e.1. This clade contains mostly *ent*-kaurene synthases, which are integral to gibberellin metabolism. All BGC class I diTPSs cluster in TPS-e.2, which contains enzymes that generally accept (+)-CPP as a substrate. Enzymes syntenic with CamTPS9 are grouped in clade TPS-e.2.1, which contains all but one of the Lamiaceae enzymes known to catalyze formation of miltiradiene. Also characteristic of this clade is the loss of the internal γ domain, which is retained in most diTPSs but lost in mono- and sesqui-TPSs. The 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 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[57]. 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. 4). Moreover, several BGCs contain diTPSs from clades that may offer distinctive chemistries.

CYPs in the 76AH 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[63,64]. 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[55,84]. 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 enzymes from *H. officinalis* and *L. leonurus* are in the same clade as the CYP71D 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 (13R) 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 CYP71Ds 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. 5, Supplemental Fig. 1, Supplemental Fig. 2). High levels of sequence conservation between syntelogs supports a minimal ancestral cluster that contains 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 *Erythranthe lutea* (yellow monkeyflower; formerly *Mimulus luteus*), a closely related Lamiales outgroup[80,87,88]. We found a syntenic block which contains a class II diTPS as well as a class I diTPS but no CYPs. The class II diTPS, El.13152, is in clade TPS-c.2, showing some similarity with the (+)-CPP synthases. The class I enzyme, El.13874, is within TPS-e.2.1, but distinct from the rest of the clade and surprisingly retains the γ domain (Fig. 4). This domain loss has occurred multiple times in the evolution of plant TPSs[89], so it is conceivable that El.13874 represents the three-domain Lamiaceae miltiradiene synthase shared by the most recent common ancestor. While the *E. lutea* cluster provides 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.

Functional characterization of the C. americana *BGC reveals two metabolic modules and a novel 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[90]. 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 CYP76AHs from Sm-b were expressed exclusively in root tissues and found to be vital steps in the root tanshinone biosynthetic pathway[59]. Additionally, an array of root-specific CYP71Ds 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)[79] and the noscapine/morphinan biosynthesis in *Papaver ssp.* (poppy)[11,66]. 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. 6)[78]. 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*)[78], 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.

We successfully cloned from cDNA and tested 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 and potential promiscuity. 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[91,92]. 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. 7). CamTPS7 was confirmed to be a (+)-CPP synthase (Supplemental Fig. 3). CamTPS9 is a miltiradiene synthase, with some abietatriene resulting from spontaneous aromatization *in plantae* consistent with previous observations[93]. CamTPS10, when paired with a (+)-CPP synthase, forms (+)-kaurene, a previously unknown diTPS activity (NMR: Supplemental Fig. 4). 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[94]. Calliterpenone has been investigated for its potential as a plant growth promoting agent[95], 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[92]. 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.

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. 8): *ent*-kaurene (CamTPS12; Supplemental Fig. 5) and kolavenol[78] 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. 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 (Supplemental Fig. 6). We were unable to separate these products by column chromatography, preventing complete structural elucidation. CamCYP76AH68 dramatically shifted the product profile towards abietatriene and afforded a small amount of oxidized abietatriene (Supplemental Fig. 6). 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. 9). 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[93]. 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 new 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. 8, Supplemental Fig. 7). Each CYP/TPS 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 new oxidized compound (Fig. 8, Supplemental Fig. 6). The combination of *CamTPS6* with *CamCYP71D716* and *CamCYP71D717* resulted in full conversion of (+)-manool to 3-oxy-manool, which was confirmed by NMR (Fig. 8, NMR: Supplemental Fig. 8).

No abietane-type diterpenoids were previously found in *C. americana*, which has been primarily studied for clerodane diterpenoids produced in leaves[96–98]. Other *Callicarpa* species, including *C. bodinieri* and *C. macrophylla*[99], produce a wide variety of medicinally relevant abietane diterpenoids (Fig. 9), indicating that the abietane skeleton is a key intermediate for at least

some plants in this genus[73,99]. We analyzed methanol extracts of *C. americana* root, fruit, and leaf tissue for evidence of diterpenoids. All extracts showed peaks with distinct retention times and MS/MS fragmentation patterns consistent with diterpenoids (Supplemental Fig. 9).

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[78]. 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 tandem 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 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 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 newly discovered miltiradiene BGCs 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[69,103]. 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[69–72, 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 supports the presence of a miltiradiene diterpenoid network in this plant despite the lack of previously documented abietanes. Further characterization of the other newly identified miltiradiene BGCs in *H. officinalis, P. cablin*, and *L. leonurus* could similarly lead to the discovery of new 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 first 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.

Material And 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]. Synteny was manually curated using sequence identity from BLASTp and ClustalOmega alignments (version 1.2.4)[110].

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 [111]. 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 Supplemental Table 4. CYP annotation was kindly provided by David Nelson (University of Tennessee). All protein phylogenies were generated with the alignment software Clustal-Omega (version 1.2.4)[110] and phylogenetic support was provided by maximum likelihood by RAxML (version 8.2.12)[112] with support from 1000 bootstrap replicates. 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, 150nt. 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 with Augustus (version 3.3)[115] with the parameter, '–UTR = off', and the '–species' and 'c–extrinsicCfgFile' parameters to use training results from closely related species, *Hyssopus officinalis* (*P. barbatus*, *P. vulgaris)* or *Tectona grandis* (*L. leonurus*). Assembly statistics were calculated using the tool assembly-stats v1.0.1[116]. BUSCO v5.2.2[117] was run in genome mode using the lineage dataset 'embryophyta_odb10.'

PCR and cloning

Synthetic oligonucleotides, GenBank accession numbers, and sequences of all enzymes characterized or discussed in this study are listed in Supplemental Table 4. Candidate enzymes were PCR-amplified from root, fruit, leaf, and flower cDNA, and coding sequences were cloned and sequence-verified. Constructs were then cloned into the plant expression vector pEAQ-HT [118] and used in transient expression assays in *N. benthamiana*.

Transient expression in N. benthamiana

Transient expression assays in *N. benthamiana* were carried out as previously described[57]. *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. Each candidate was co-expressed with *Plectranthus barbatus* (formerly *Coleus forskohlii*) DXS and *P. barbatus* GGPPS[92] in pEAQ-HT. 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. Plants were returned to the controlled growth room (76°C, 12H diurnal cycle) for 5 days. Approximately 200 mg fresh weight from infiltrated leaves was extracted with 1 ml hexane (TPS 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.

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-hydroxy-(+)-kaurene. NMR spectra were measured in Michigan State University's Max T. Rogers NMR Facility on a Bruker 800 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.

Declarations

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; 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.

ACKNOWLEDGEMENTS

We would like to thank Drs. Dan Jones and Casey Johnny of Michigan State University's Mass Spectrometry and Metabolomics Core Facility for their help in obtaining and interpreting GC-MS data, and Dr. Dan 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, Brieanne Vaillancourt for assisting with the 10x Illumina sequencing data management and assembly statistics, 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'). 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. 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.

COMPETING INTERESTS

The authors declare that there is no conflict of interest.

DATA AVAILABILITY

Raw genomic reads are available on the NCBI BioSample database: *Plectranthus barbatus* (SAMN26547115), *Leonotis leonurus* (SAMN26547116), and *Prunella vulgaris* (SAMN26547117). Additional supplemental materials and genome assemblies and annotations can be found in the Dryad Repository: https://doi.org/10.5061/dryad.w9ghx3frg.

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.* <bvertical-align:super;>112</bvertical-align:super;>, 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* <bvertical-align:super;>12</bvertical-align:super;>, 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.* <bvertical-align:super;>107</bvertical-align:super;>, 10136–10141 (2010).
21. Field, B. *et al.* Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci.* <bvertical-align:super;>108</bvertical-align:super;>, 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.* <bvertical-align:super;>117</bvertical-align:super;>, 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.* <bvertical-align:super;>101</bvertical-align:super;>, 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.* <bvertical-align:super;>117</bvertical-align:super;>, 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. 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* <bvertical-align:super;>3</bvertical-align:super;>, 408–414 (2007).
44. Guenard, D., Gueritte-Voegelein, F. & Potier, P. Taxol and Taxotere: Discovery, chemistry, and structure-activity relationships. *Acc. Chem. Res.* **26**, 160–167 (1993).
45. 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).
46. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nat. 2013 4967446* **496**, 528–532 (2013).
47. 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).
48. Lange, B. M. *et al.* Improving peppermint essential oil yield and composition by metabolic engineering. *Proc. Natl. Acad. Sci.* <bvertical-align:super;>108</bvertical-align:super;>, 16944–16949 (2011).
49. Schalk, M. *et al.* Toward a biosynthetic route to sclareol and amber odorants. *J. Am. Chem. Soc.* **134**, 18900–18903 (2012).
50. Philippe, R. N., De Mey, M., Anderson, J. & Ajikumar, P. K. Biotechnological production of natural zero-calorie sweeteners. *Curr. Opin. Biotechnol.* **26**, 155–161 (2014).
51. Celedon, J. M. & Bohlmann, J. Genomics-based discovery of plant genes for synthetic biology of terpenoid fragrances: A case study in sandalwood oil biosynthesis. *Methods Enzymol.* **576**, 47–67 (2016).
52. 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).
53. Karunanithi, P. S. & Zerbe, P. Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Front. Plant Sci.* **10**, 1166 (2019).
54. Boutanaev, A. M. *et al.* Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci.* <bvertical-align:super;>112</bvertical-align:super;>, E81–E88 (2015).
55. Bathe, U. & Tissier, A. Cytochrome P450 enzymes: A driving force of plant diterpene diversity. *Phytochemistry* **161**, 149–162 (2019).
56. Lange, B. M. The Evolution of plant secretory structures and emergence of terpenoid chemical diversity. *Annu. Rev. Plant Biol.* **66**, (2015).
57. 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).
58. Sherden, N. H. *et al.* Identification of iridoid synthases from *Nepeta* species: Iridoid cyclization does not determine nepetalactone stereochemistry. *Phytochemistry* **145**, 48–56 (2018).

59. Xu, H. *et al.* Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*. *Mol. Plant* **9**, 949–952 (2016).
60. 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).
61. Gao, W. *et al.* A functional genomics approach to tanshinone biosynthesis provides stereochemical insights. *Org. Lett.* **11**, 5170–5173 (2009).
62. 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).
63. Guo, J. *et al.* CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts. *Proc. Natl. Acad. Sci.* <bvertical-align:super;>110</bvertical-align:super;>, 12108–12113 (2013).
64. Guo, J. *et al.* Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. *New Phytol.* **210**, 525–534 (2016).
65. Cui, G. *et al.* Functional divergence of diterpene syntheses in the medicinal plant *Salvia miltiorrhiza*. *Plant Physiol.* **169**, 1607–1618 (2015).
66. 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).
67. Wang, Z. & Peters, R. J. Tanshinones: Leading the way into Lamiaceae labdane-related diterpenoid biosynthesis. *Curr. Opin. Plant Biol.* **66**, 102189 (2022).
68. Song, J.-J. *et al.* A 2-oxoglutarate-dependent dioxygenase converts dihydrofuran to furan in *Salvia* diterpenoids. *Plant Physiol.* **188**, 1496–1506 (2022).
69. González, M. A. Aromatic abietane diterpenoids: Their biological activity and synthesis. *Nat. Prod. Rep.* **32**, 684–704 (2015).
70. 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).
71. Machumi, F. *et al.* Antimicrobial and antiparasitic abietane diterpenoids from the roots of *Clerodendrum eriophyllum*. *Nat. Prod. Commun.* **5**, 1934578X1000500605 (2010).
72. Abdissa, N., Frese, M. & Sewald, N. Antimicrobial abietane-type diterpenoids from *Plectranthus punctatus*. *Molecules* **22**, 1919 (2017).
73. Gao, J. *et al.* Anti-NLRP3 inflammasome abietane diterpenoids from *Callicarpa bodinieri* and their structure elucidation. *Chin. Chem. Lett.* **31**, 427–430 (2020).
74. Birtić, S., Dussort, P., Pierre, F.-X., Bily, A. C. & Roller, M. Carnosic acid. *Phytochemistry* **115**, 9–19 (2015).
75. Ignea, C. *et al.* Carnosic acid biosynthesis elucidated by a synthetic biology platform. *Proc. Natl. Acad. Sci.* <bvertical-align:super;>113</bvertical-align:super;>, 3681–3686 (2016).

76. Scheler, U. *et al.* Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast. *Nat. Commun.* **7**, 12942 (2016).
77. 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).
78. Hamilton, J. P. *et al.* Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, *Callicarpa americana*. *GigaScience* <bvertical-align:super;>9</bvertical-align:super;>, giaa093 (2020).
79. 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).
80. 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).
81. Yao, G. *et al.* Phylogenetic relationships, character evolution and biogeographic diversification of *Pogostemon* s.l. (Lamiaceae). *Mol. Phylogenet. Evol.* **98**, 184–200 (2016).
82. 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).
83. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).
84. Bak, S. *et al.* Cytochromes P450. *Arab. Book Am. Soc. Plant Biol.* **9**, e0144 (2011).
85. Pateraki, I. *et al.* Total biosynthesis of the cyclic AMP booster forskolin from *Coleus forskohlii*. *eLife* **6**, e23001 (2017).
86. Kliebenstein, D. J. A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PloS One* **3**, e1838 (2008).
87. 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).
88. BioArchive reference in prep.
89. Hillwig, M. L. *et al.* Domain loss has independently occurred multiple times in plant terpene synthase evolution. *Plant J.* **68**, 1051–1060 (2011).
90. 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).
91. 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).
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. Zi, J. & Peters, R. J. Characterization of CYP76AH4 clarifies phenolic diterpenoid biosynthesis in the Lamiaceae. *Org. Biomol. Chem.* **11**, 7650 (2013).
94. Jones, W. P. & Kinghorn, A. D. Biologically active natural products of the genus *Callicarpa*. *Curr. Bioact. Compd.* **4**, 15–32 (2008).
95. 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).
96. 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).
97. 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).
98. 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).
99. 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).
100. Hillwig, M. L. *et al.* Domain loss has independently occurred multiple times in plant terpene synthase evolution. *Plant J.* **68**, 1051–1060 (2011).
101. 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).
102. Tu, L. *et al.* Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat. Commun.* **11**, 971 (2020).
103. Chaturvedi, R. *et al.* An abietane diterpenoid is a potent activator of systemic acquired resistance. *Plant J.* **71**, 161–172 (2012).
104. 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).
105. Zeng, T. *et al.* TeroKit: A database-driven web server for terpenome research. *J. Chem. Inf. Model.* **60**, 2082–2090 (2020).
106. Nützmann, H.-W. & Osbourn, A. Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* **26**, 91–99 (2014).
107. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
108. Wang, Y. *et al.* MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
109. 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).

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. Revell, L. J. Phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
112. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
113. 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).
114. 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).
115. Stanke, M. *et al.* AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
116. Assembly-*stats*. (Pathogen Informatics, Wellcome Sanger Institute, 2022).
117. 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).
118. 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).
119. 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).

Figures

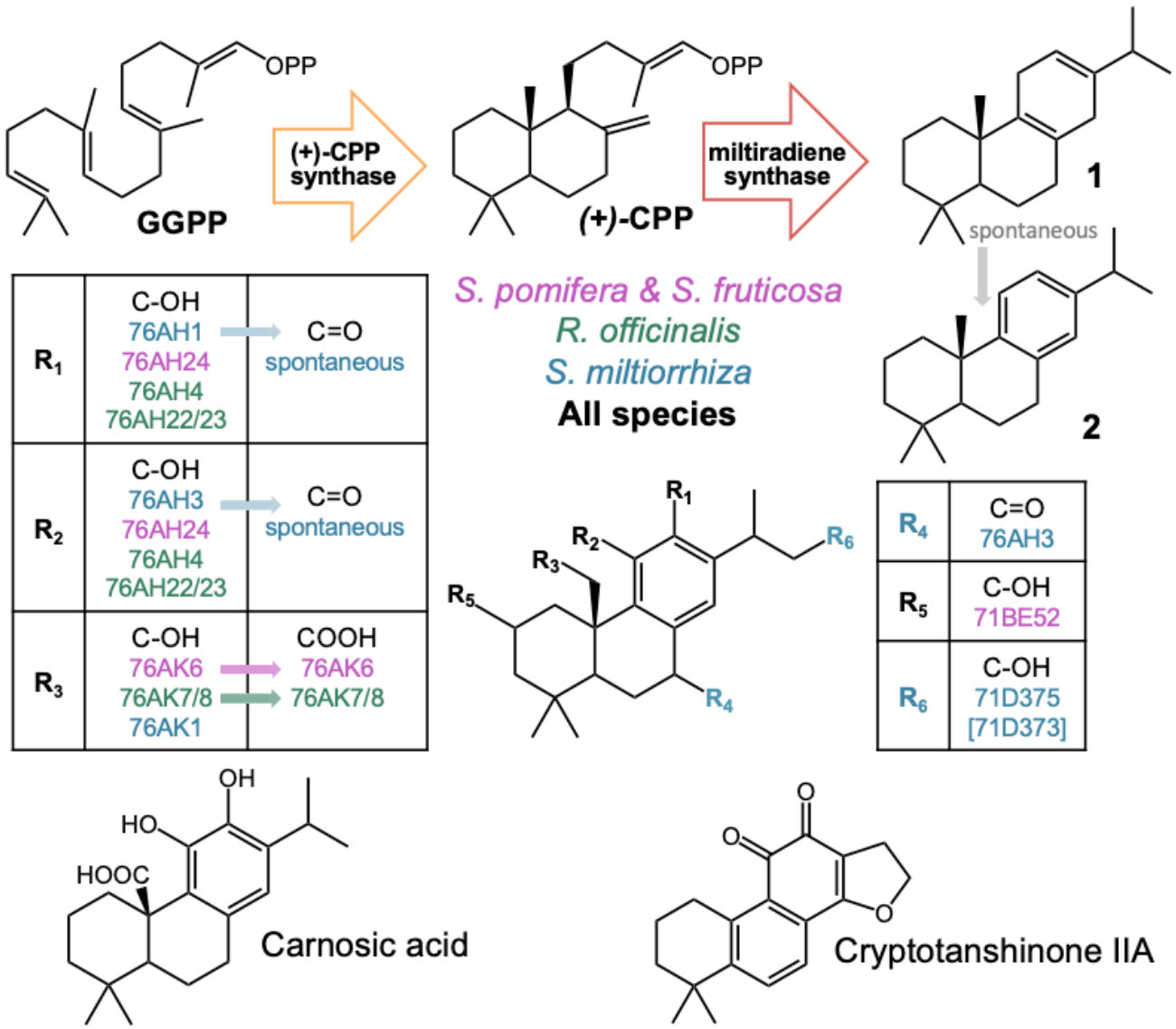
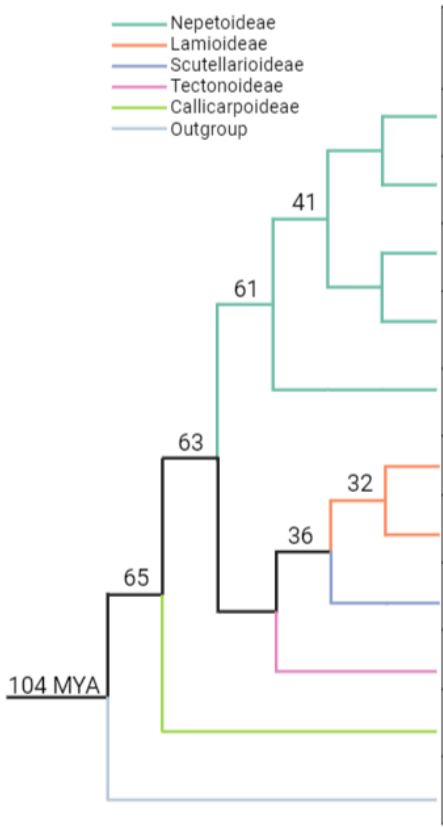


Figure 1

Oxidation positions for CYPs involved in the biosynthesis of tanshinones (*S. miltiorrhiza*), carnosic acid (*S. pomifera, S. fruticosa, R. officinalis*), and related diterpenoids. These CYPs have been shown to be part of a metabolic network, and intermediates required for sequential oxidation are condensed here for brevity. Aromatization of miltiradiene to abietatriene, the precursor for most oxidations shown, is presumed to be spontaneous[93]. Enzymes from different species that oxidize at the same position are also orthologous.



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**	
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**	
Leonotis leonurus	**Lion's tail**	**Lamioideae**	**This work**	
Pogostemon cablin	Pachouli	Lamioideae	He *et.al.*, 2018	1.9 Gb
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
Mimulus luteus	Monkey flower	Phrymaceae, Lamiales outgroup	BioArchive	680 Mb

Figure 2

Species and genome assemblies used in this study. The phylogram shows evolutionary relationships between the species studied. Numbers at the nodes represent estimations of clade age in millions of years (MYA)[80–82]. Figure created using BioRender.com.

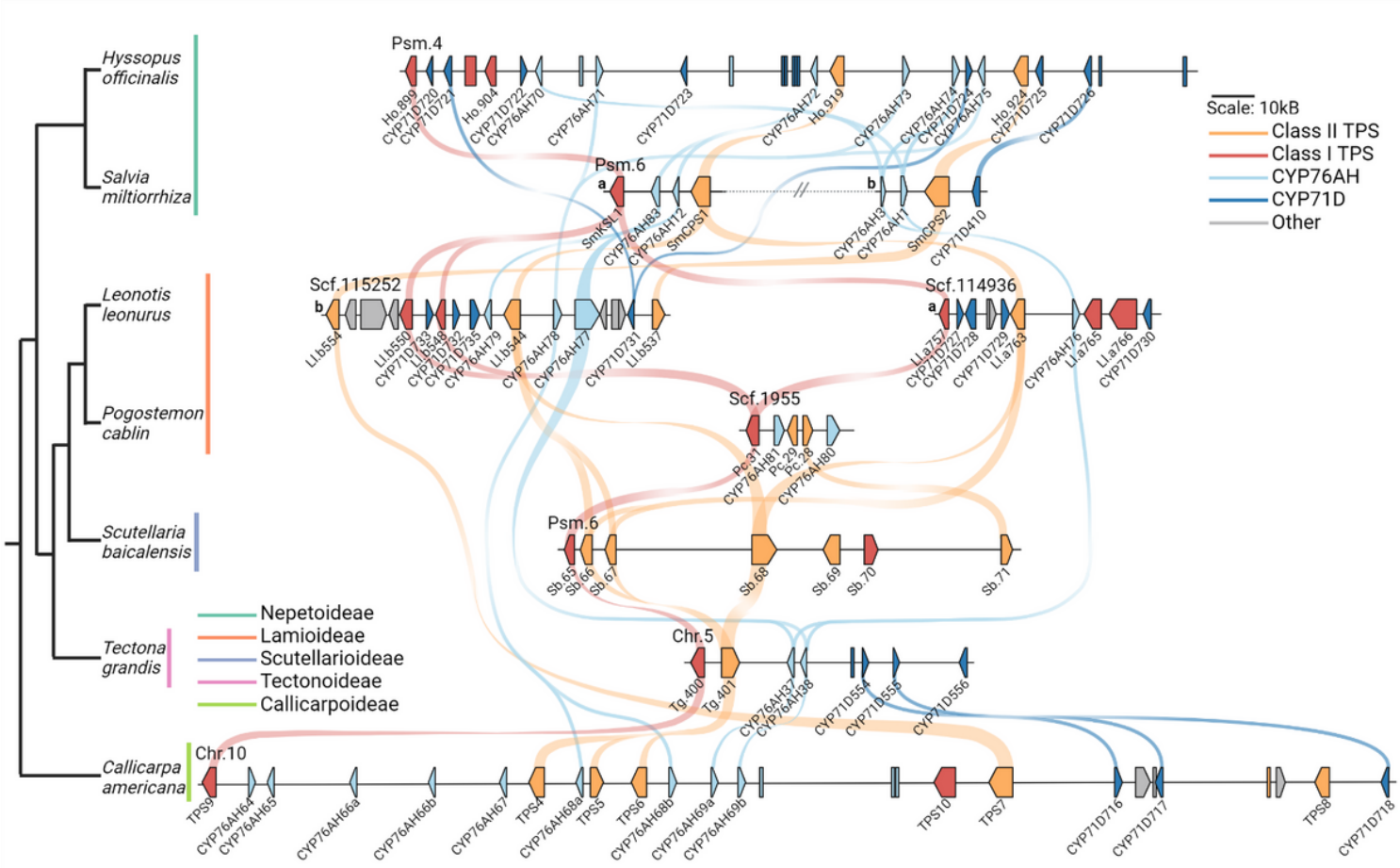


Figure 3

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 TPS class II (+)-CPP synthase, a TPS 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. MCScanX software was used to determine synteny between each species, shown here with colored curves. Species tree adapted from Mint Evolutionary Genomics Consortium 2018[119]. Figure created using BioRender.com.

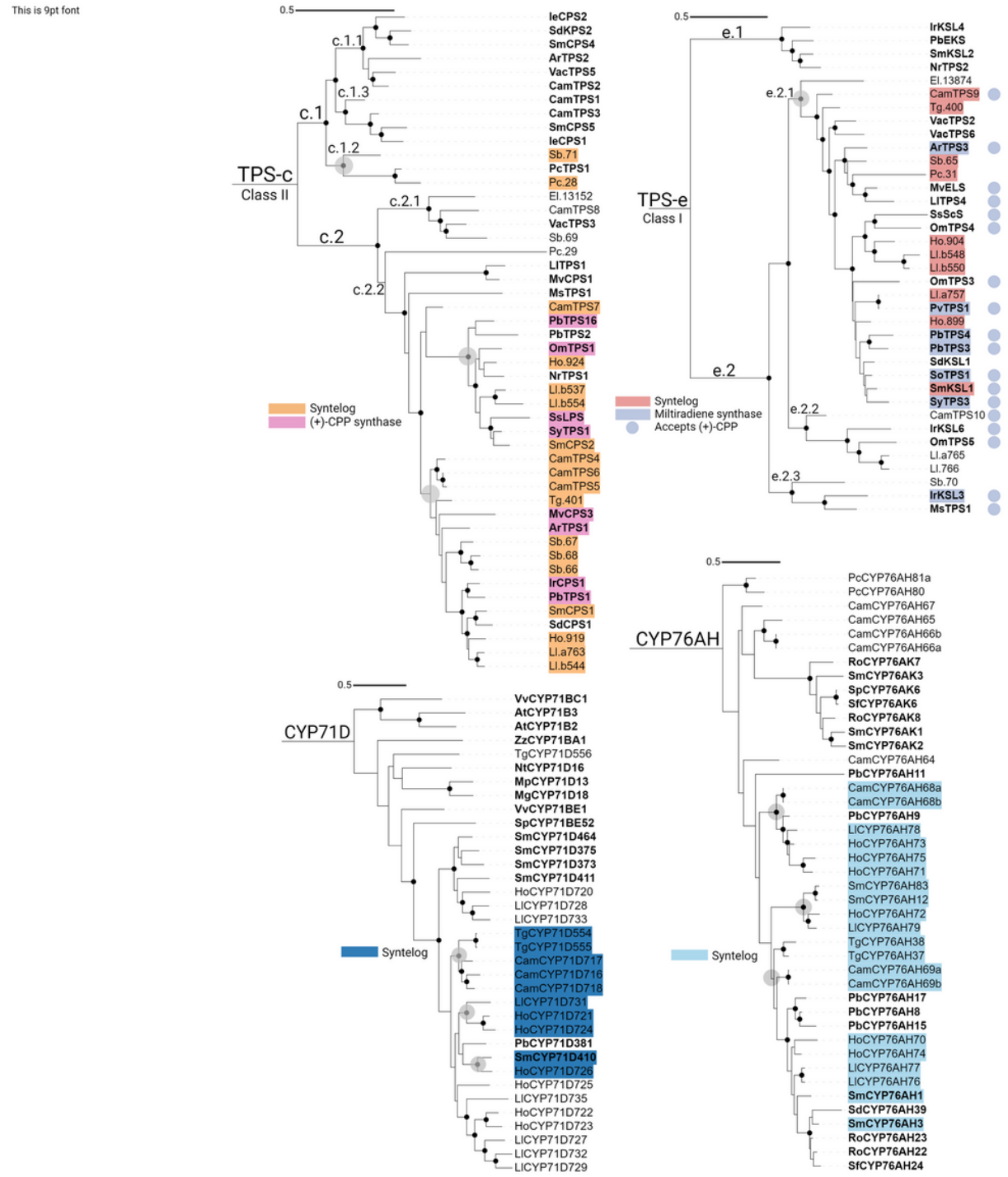


Figure 4

Phylogenetic evidence shows the relatedness of each gene class in the clusters. 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). TPSs 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. Circles around clade nodes represent the expansion point for syntelogs and

share approximately 70% or more sequence similarity. TPS trees are rooted to PpEKS, and CYP trees are rooted to AtCYP701A3.

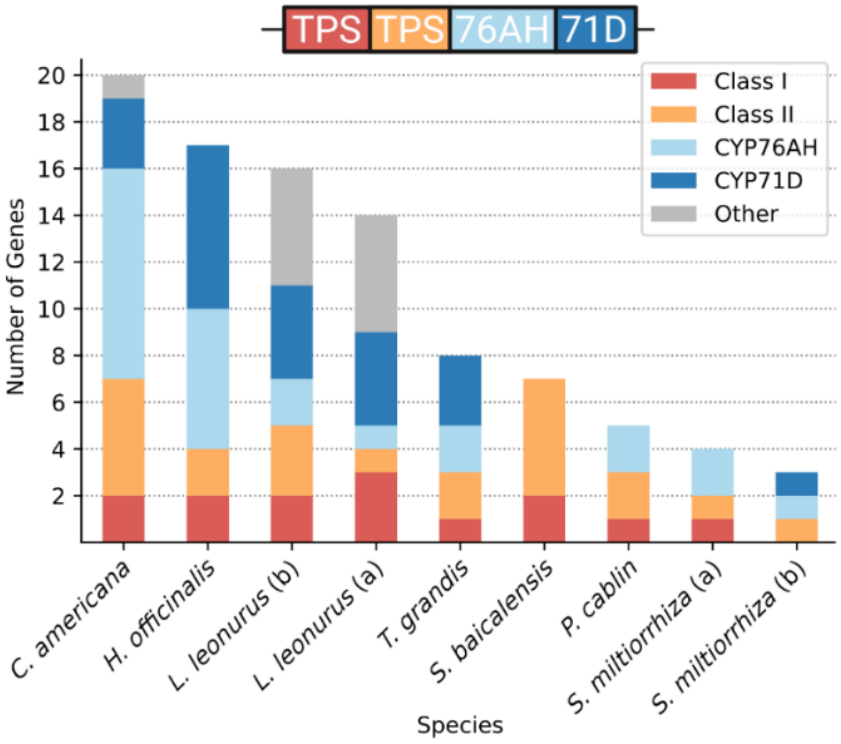


Figure 5

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. Linage-specific expansion and refinement are evident from the number and composition of genes in each gene family present in the extant species studied.

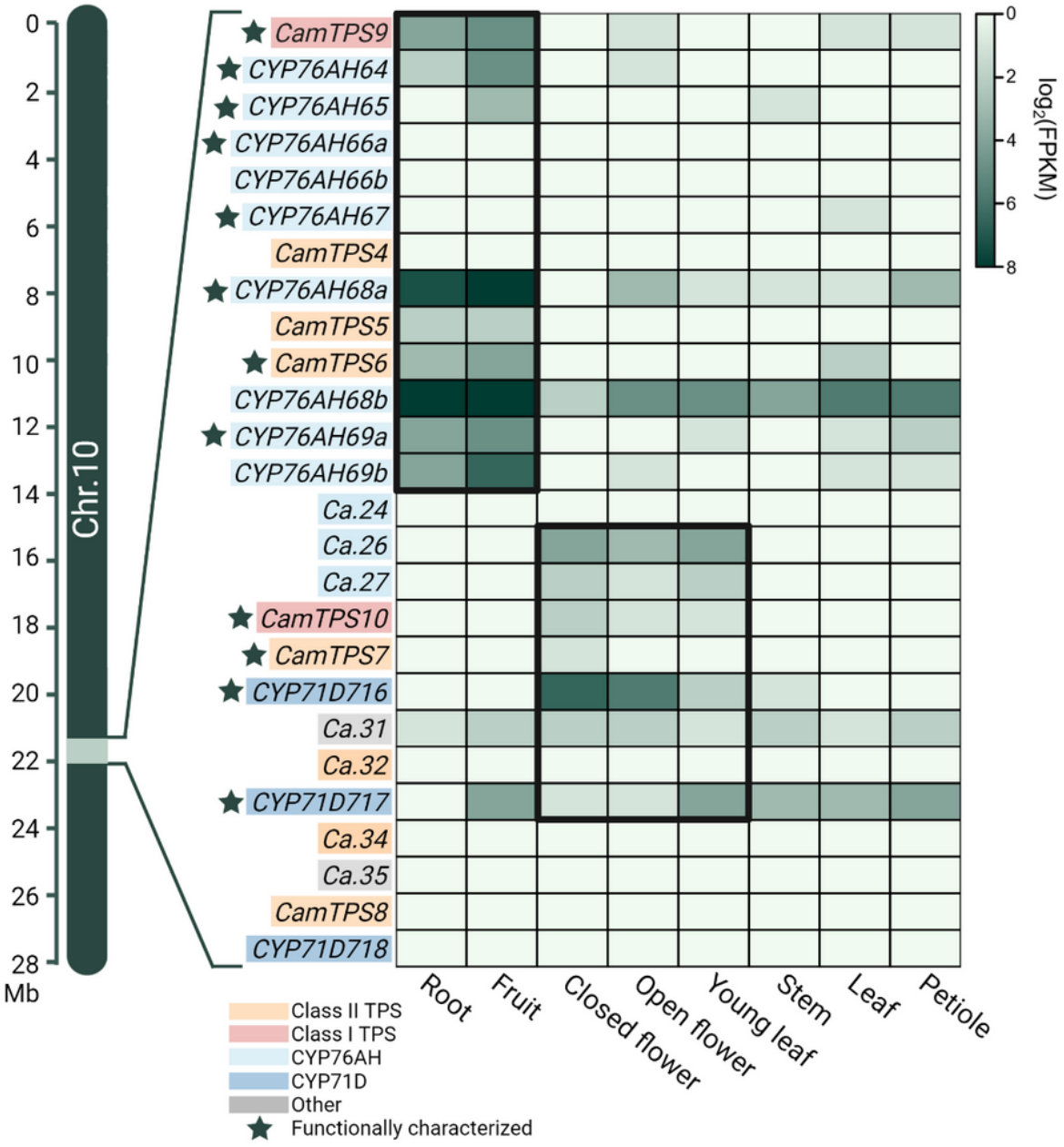


Figure 6

Tissue specific expression of a miltiradiene biosynthetic gene cluster 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 highlighted. Pseudogenes and non-related genes are colored respectively. Data obtained from Hamilton *et al*, 2020[78].

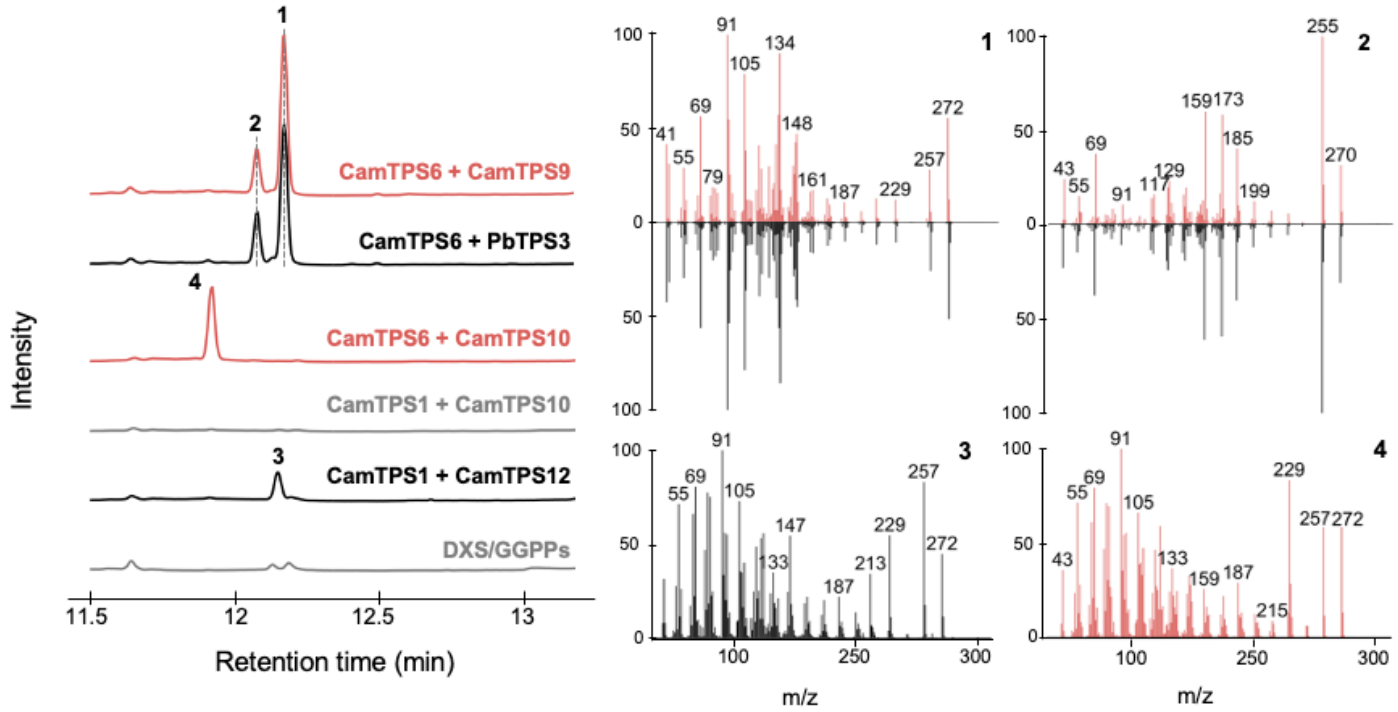


Figure 7

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[44–47] products when both were expressed with the (+)-CPP synthase CamTPS6[78], forming miltiradiene (**1**) and abietatriene (**2**). CamTPS10 was found to generate **4** from (+)-CPP but not *ent*-CPP (CamTPS1)[78]. This product has a different retention time but similar mass spectrum to *ent*-kaurene (**3**), formed by the combination of CamTPS1 and CamTPS 12 (Supplemental Fig. 5). All chromatograms shown are total ion chromatograms. Each combination includes DXS+GPPS, shown as a control in gray.

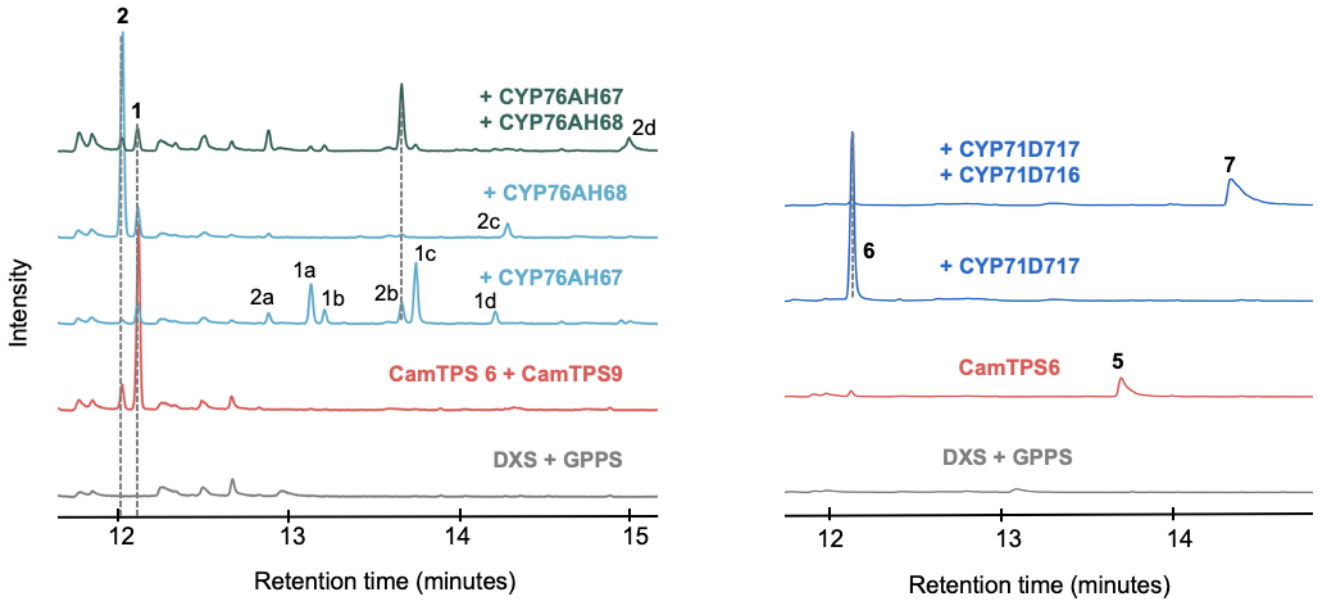


Figure 8

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. 6). (b) CamCYP71D717 catalyzes the production of (+)-manool **(6),** likely from (+)-copalol **(5)** (Supplementary Fig 7) and the addition of CamCYP71D716 results in 3-hydroxy-(+)-manool **(7)**.

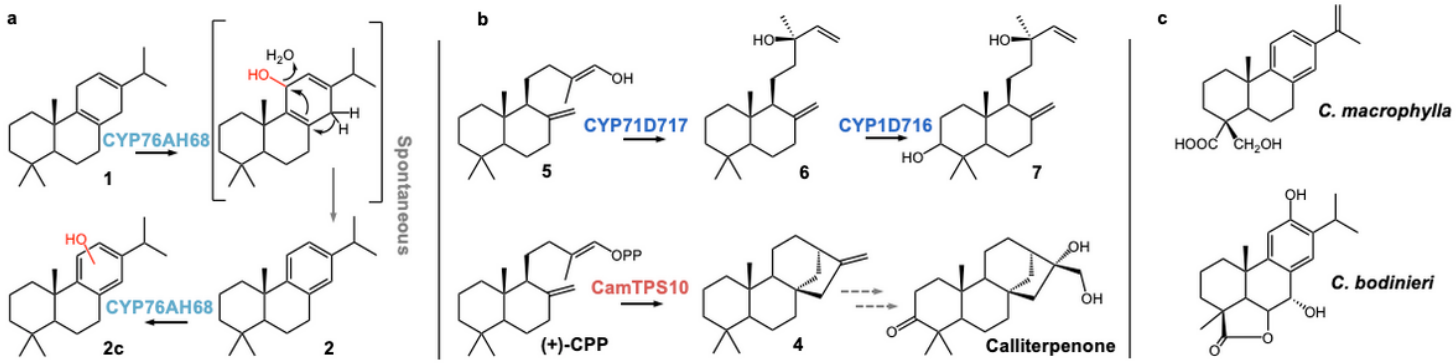


Figure 9

(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. 6). (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 (+)-kaurene. (c) Structures of abietane diterpenoids found in two other species of *Callicarpa*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BrysonLanieretalSupplementalFiguresandDataNatComms.docx