DECIPHERING DITERPENOID BIOSYNTHESIS AND GENOME ARCHITECTURE IN THE LAMIACEAE FAMILY: INSIGHTS FROM *CALLICARPA AMERICANA* AND *TEUCRIUM CHAMAEDRYS*

By

Abigail Elizabeth Bryson

A DISSERTATION

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

Genetics and Genome Sciences – Doctor of Philosophy
Molecular Plant Sciences – Dual Major

2025

ABSTRACT

The Lamiaceae is one of the largest plant families incorporating over 7,000 species. It is home to a variety of ornamental and medicinal plants, but perhaps the most well-known are the culinary herbs, such as basil, oregano, rosemary, and thyme. The often-fragrant leaves of these plants are in part thanks to their terpenoid content. Terpenes are a class of natural compounds that are highly prolific in plants, and specifically the Lamiaceae family. These specialized metabolites can serve a variety of functions in the plant, including communication and defense. Terpenoids have been co-opted for human applications, as many can be used as medicines, pesticides, flavors, and fragrances. Occasionally, specialized metabolism, or specifically terpenoid metabolism, can be physically clustered in the genome. Biosynthetic gene clusters (BGCs) are a set of genes in the genome that are biochemically associated, phylogenetically distinct, and are often transcriptionally linked. One piece of the puzzle making the discovery of BGCs more feasible is the increase in robustness and quality of genome sequencing today. The following dissertation integrates the ideas of terpenoid metabolism and BGCs in the Lamiaceae family. Specifically, I explore the genomes of several Lamiaceae species, including *Callicarpa americana* and *Teucrium chamaedrys*, in search of BGCs and to explore terpenoid metabolism. I discovered a Lamiaceae family-wide gene cluster in *C. americana* that appears to predate the family at around 65 million years old. Additionally, I sequence and analyze the genome of the medicinally relevant plant *T. chamaedrys*, and find not only the same large BGC, but also discover the most putative diTPSs identified in a single species to date. I also functionally characterize several diTPSs and associated Cytochrome P450s. My work presented here aims to lay another brick in the larger building of understanding plant genomes and elucidating plant specialized metabolism.

ACKNOWLEDGEMENTS

I would first like to thank my incredible PI, Dr. Björn Hamberger, for all his mentoring and support over these last several years as I learn and grow as a scientist and as a person. In the same vein, I would like to thank my undergraduate mentor, Dr. Gregory Bonito, for starting me on the path to research. Thank you to Dr. C Robin Buell, who generously hosted me at the University of Georgia, mentoring me and helping me develop my genomics skills. Additionally, I would like to thank the rest of my guidance committee, Dr. Patrick Edger and Dr. Marjorie Weber, for always answering my questions and challenging me to learn outside my discipline. Dr. Claire Vielle and Alaina Burghardt have provided me with many resources and answered all my frantic emails as my Genetics and Genome Sciences program team. Dr. Jianping Hu and the Association of Molecular Plant Sciences Students have also been a great source of resources and companionship in the Molecular Plant Sciences program. I would like to thank the other Hamberger Lab graduate students, past and present, for the encouragement, advice, and friendship over the years: Dr. Jacob Bibik, Dr. Garret Miller, Dr. Emily Lanier, Dr. Davis Mathieu, Nick Schlecht, Lucas Reist, and Angel McKay Whiteman. Thank you to my family, especially my siblings, Elise, Matt, and Josh, who have been there to encourage and advise from the very beginning. I want to especially thank my friends, who joined me in both celebrations and failures: my college roomies, my MPS floor mates, my non-scientist people, and the rest. Too many to name, but you know who you are, and you mean the world to me. Lastly, I could not have done this without the encouragement, support, and love from my partner Davis and our sweet dog Mango. Thank you for being with me every step of the way, through thick and thin.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

2OGD	2-oxoglutarate dependent oxygenases
BGC	biosynthetic gene cluster
bp	base pair
BUSCO	Benchmarking Universal Single-Copy Ortholog
CPP/CDP	copalyl diphosphate
CTAB	cetyltrimethylammonium bromide
CYP	cytochrome P450
diTPS	diterpene synthase
DMAPP/DMADP	dimethylalyl diphosphate
DNA	dioxyribonucleic acid
DXS	1-deoxy-D-xylulose-5-phosphate synthase
EKS	*ent*-kaurene synthase
Gbp	giga base pairs
GC-MS	gas chromatography mass spectrometry
GGPP/GGDP	geranyl geranyl diphosphate
GGPPS/GGDPS	geranyl geranyl diphosphate synthase
HGT	horizontal gene transfer
HMW	high molecular weight
Kbp	kilo base pairs
KDP	kolavenyl diphosphate
Mbp	mega base pairs
MEP	methylerythritol 4-phosphate
MVA	mevalonate

NMR	nuclear magnetic resonance
ONT	Oxford Nanopore Technology
RNA	ribonucleic acid
RTA	real time analysis
SS	sclereol synthase
TPS	terpene synthase
WGD	whole genome duplication

CHAPTER 1: AN OVERVIEW OF DITERPENOID METABOLISM AND THE LAMIACEAE FAMILY

Abigail E. Bryson, Björn R. Hamberger

Abstract

This chapter provides an overview of diterpenoid metabolism and genomic structure within the context of the Lamiaceae family. The Lamiaceae family plays a significant role in both traditional medicine and genomic research. Recent advancements in sequencing technologies have led to the publication of genomes for 46 species within this family, enhancing our understanding of their phylogenetic relationships and evolutionary biology. These genomes allow us to study biosynthetic gene clusters that can, in some rare cases, facilitate the production of specialized metabolites in plants. Gene clusters are crucial for efficient co-regulation of metabolic pathways and evolutionary adaptation. This chapter further delves into the diverse class of terpenoids, highlighting their biosynthesis, functional diversification, and significance in both plant defense and human applications. The importance of gene duplication and neofunctionalization in the evolution of terpene synthases is emphasized, particularly in relation to diterpenoid biosynthesis. Additionally, the role of cytochrome P450 oxygenases and other enzymes in modifying terpenoid backbones to enhance their bioactivity is explored. The chapter sets the stage for detailed investigations in subsequent chapters on the genomic landscape and diterpenoid metabolism in selected Lamiaceae species, underscoring the family's vast potential for yielding novel and valuable bioactive compounds.

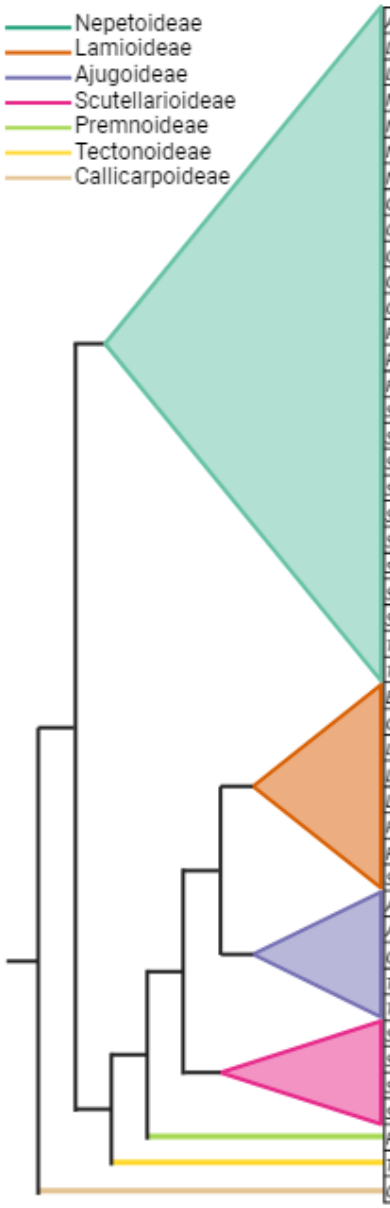
Overview of the Lamiaceae Family and Its Role in Genomic Research

The Lamiaceae family is the sixth largest angiosperm family with around 7,000 species and is estimated to have diverged around 65-70 million years ago (Yao *et al.*, 2016; Li *et al.*, 2017; Godden *et al.*, 2019). Some of the most well-known members of this family include peppermint (*Mentha x piperita*), lavender (*Lavandula angustifolia*), basil (*Ocimum basilicum*), and catnip (*Nepeta cataria*). Species in this family are native to nearly every region on earth, and most are naturalized globally. Many species in this family have been used for thousands of years in traditional medicines, serving as the primary source of health care in numerous traditions (Pieroni, Quave and Santoro, 2004; di Tizio *et al.*, 2012; Arı *et al.*, 2015; Jarić, Mitrović and Pavlović, 2020).

The prevalence of the Lamiaceae in ethnobotanical studies underscores the importance they can serve in medicine and beyond. And, in the last several decades, there has been a burst of genomic data becoming available in an effort to study these species. This has largely been driven by sequencing technology advancements as well as more powerful computing resources. The first Lamiaceae species to have its genome sequenced was pachuli (*Pogostemon cablin*) in 2016 (He *et al.*, 2016). As of July 2024, now 46 species in the Lamiaceae have published genomes, most of which have been published by 2018 or later (Figure 1.1). Of these, 27 can be considered chromosome-scale assemblies, although telomere and centromere sequences remain poorly studied in this family.

The families of Lamiaceae and Verbenaceae were joined until the 1990's, so genera are still being reclassified today. Although plastidial loci and genes have been used to classify familial relationships (Zhao *et al.*, 2021), a more direct approach is also used, with single-copy nuclear genes originating largely from transcriptomes used to separate the major clades of the Lamiaceae (Figure 1.1; Boachon *et al.*, 2018). These trees originating from plastidial and nuclear genes are mostly congruent, with only a few topology differences between them. An accurate phylogeny is an important resource for evolutionary biology, as it provides context for trait evolution. It is also useful for tracking and inferring

whole genome duplication (WGD) events. WGD is a major driver of evolution in plants (Clark and Donoghue, 2018; Ren *et al.*, 2018), especially in the realm of specialized metabolism where it creates genetic redundancy with decreased selective pressure that can more easily neofunctionalize (Ohno, 1970; Birchler and Yang, 2022).



Agastache rugosa	Korean mint	Chromosome
Dracocephalum rupestre	Dragonhead	Chromosome
Lavandula angustifolia	English lavender	Chromosome
Mentha longifolia	Horse mint	Chromosome
Monarda citriodora	Lemon beebalm	Scaffold
Nepeta cataria	Catnip	Scaffold
Nepta racemosa (mussinii)	Dwarf catnip	Scaffold
Ocimum basilicum	Basil	Scaffold
Ocimum sanctum	Holy basil	Scaffold
Ocimum tenuiflorum	Holy basil	Scaffold
Origanum marjorana	Marjoram	Scaffold
Origanum vulgare	Oregano	Scaffold
Perilla citriodora	Wild Perilla	Chromosome
Perilla frutescens	Beefsteak plant	Scaffold
Prunella vulgaris	Heal-all	Scaffold
Salvia bowleyana	Southern Danshen	Chromosome
Salvia elegans	Pineapple Sage	Scaffold
Salvia hispanica	Chia	Chromosome
Salvia miltiorrhiza	Danshen	Chromosome
Salvia officinalis	Common sage	Chromosome
Salvia rosmarinus	Rosemary	Chromosome
Salvia sclarea	Clary sage	Scaffold
Salvia splendens	Scarlet Sage	Chromosome
Schizonepeta tenuifolia	Jing Jie	Chromosome
Thymus mandschuricus	Manchurian thyme	Chromosome
Thymus quinquecostatus	Five-rib Thyme	Chromosome
Ballota nigra	Black horehound	Chromosome
Galeopsis tetrahit	Common help-nettle	Chromosome
Lamium amplexicaule	Henbit	Chromosome
Leonotis leonurus	Lion's mane	Scaffold
Leonurus japonicus	Chinese motherwort	Chromosome
Plectranthus barbatus	Indian coleus	Scaffold
Pogostemon cablin	Patchouli	Chromosome
Stenogyne calaminthoides	Oahu Stenogyne	Scaffold
Ajuga chamaepitys	Yellow bugle	Chromosome
Ajuga reptans	Bugleweed	Scaffold
Caryopteris x clandonesis	Bluebeard	Scaffold
Teucrium chamaedrys	Wall Germander	Scaffold
Teucrium marum	Cat thyme	Chromosome
Scutellaria baicalensis	Chinese Skullcap	Chromosome
Scutellaria barbata	Barbed Skullcap	Chromosome
Scutellaria galericulata	Common Skullcap	Chromosome
Scutellaria minor	Lesser Skullcap	Chromosome
Premna serratifolia	Headache Tree	Scaffold
Tectona grandis	Teak	Chromosome
Callicarpa americana	American Beautyberry	Chromosome

Figure 1.1 Genomes in the Lamiaceae. Cladogram shows evolutionary relationships between the represented subfamilies (Boachon *et al.*, 2018). The second column denotes the common name, and the third column denotes assembly level. Species listed are those with genomes published as of July 2024.

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 (Xu *et al.*, 2012; Lu *et al.*, 2018). The pathways that produce these compounds are evolutionarily dynamic, allowing plants to evolve novel compounds. The rising number of published, high-quality plant genomes in recent years has led to the discovery that some specialized metabolism in plants form biosynthetic gene clusters (BGCs). A BGC is a group of two or more different classes of non-homologous, functionally related genes which are physically close and transcriptionally linked (Postnikova *et al.*, 2011; Boutanaev *et al.*, 2015; Medema *et al.*, 2015; Nützmann, Huang and Osbourn, 2016; Nützmann, Scazzocchio and Osbourn, 2018; Liu *et al.*, 2020). This type of genomic structure is common in bacteria, where many pathways are organized into operons. BGCs are common in fungi where a notable examples include the biosynthesis of penicillin (Díez *et al.*, 1990; Fierro *et al.*, 1993) and aflatoxin (Yu *et al.*, 1995). Some well-known, functionally related, clustered genes found in eukaryotes arose from tandem duplication and divergence, such as the animal *Hox* genes which are responsible for proper embryonic development. However, repeats of paralogous genes are fundamentally different from BGCs because they do not contain multiple classes of enzymes and thus were not created in the same way. The discovery of BGCs is uncommon in plants compared to bacteria and fungi. Yet increased genome quality and availability has led to the discovery of over 30 BGCs in plants to date (Polturak and Osbourn, 2021) since their first discovery in maize (Frey *et al.*, 1997). The BGCs found in plants are predominately involved in specialized rather than primary metabolism (Chu, Wegel and Osbourn, 2011). Examples include noscapine alkaloid synthesis in poppy (Winzer *et al.*, 2012), cucurbitacin in cucurbits (Dai *et al.*, 2015), and terpenoid phytoalexins and momilactones in Poaceae and other cereals (Sakamoto *et al.*, 2004; Wilderman *et al.*, 2004; Mylona *et al.*, 2008; Kitaoka *et al.*, 2021; Liang *et al.*, 2021).

The direct benefit of BGCs in plants remains mostly a mystery. However, genomic colocalization of some pathways has been hypothesized to offer fitness benefits for a plant. One advantage is the possibility of efficient co-regulation. In some instances, one inducer can control an entire pathway, providing an energetically favorable control of the metabolite(s) produced (Field *et al.*, 2011; Yu *et al.*, 2016; Nützmann, Scazzocchio and Osbourn, 2018). For example, a pathway can be activated in a tissue-specific or developmental stage-specific manner (Qi *et al.*, 2004; Mugford *et al.*, 2013). Regulation may also take place at the chromatin level, with DNA and histone methylation regulating transcription of the entire cluster (Yu *et al.*, 2016). Thus, an entire section of DNA containing the BGC can be unwound and exposed to transcription factors and RNA Polymerase all at once, which increases the efficiency of regulation. Additionally, pathways organized into BGCs are genetically linked and thus have a greater likelihood of co- inheritance. Therefore, a BGC could promote survival of the metabolic pathway during homologous recombination, decreasing the chance of disruption of the pathway through loss of a single gene (Takos and Rook, 2012). Likewise, if homologous recombination occurs near a BGC, it is more likely to occur with entire BGC at once (Slot and Rokas 2010). Having an intact pathway simultaneously regulated has been hypothesized to decrease the instances of toxic intermediate products building up and disrupting normal cellular function (Kristensen *et al.*, 2005; Mylona *et al.*, 2008; Nützmann, Huang and Osbourn, 2016). When subsequent enzymes in a pathway are immediately available, the amount of time a compound exists before being acted upon by the next enzyme can be significantly reduced. Thus, a BGC can both promote the survival of a complete metabolic pathway and provide efficient means of regulating the pathway.

Perhaps even more mysterious than the fitness advantage of BGCs is how they arise in plants. Fungal and bacterial BGCs are thought to be shared through horizontal gene transfer (HGT) based on sequence similarity of clusters across vastly divergent fungal species (Slot and Hibbett, 2007; Slot and Rokas, 2010). Physical proximity of genes is directly related to the chance of a successful HGT event, making BGCs

especially useful for these organisms. However, plant clusters are likely restricted to vertical transmission, and thus their BGCs probably arise *de novo* within the plant instead of coming from a bacterial donor (Chu, Wegel and Osbourn, 2011; Field *et al.*, 2011; Nützmann, Scazzocchio and Osbourn, 2018; Kitaoka *et al.*, 2021). It is thought that a single gene or gene pair could serve as the seed for a plant BGC, encouraging expansion of the cluster through duplication and neofunctionalization or recruitment of other genes in the pathway (Field and Osbourn, 2008; Itkin *et al.*, 2013; Nützmann, Huang and Osbourn, 2016). However, there is little evidence in the literature examining how a target gene may be recruited to a BGC. Duplication can occur as tandem, local, or whole genome, and provides genetic redundancy leading to functional evolution through enzymatic mutations or enzyme relocation (Matsuba *et al.*, 2013; Johnson, Bhat, Sadre, *et al.*, 2019; Schenck and Last, 2020). Changing the catalytic function of the enzyme can produce novel compounds, as can changing the environment of the enzyme and thus substrates available to it (Johnson, Bhat, Sadre, *et al.*, 2019). BGC formation and divergence can be especially plastic in cases of specialized metabolism where enzymes are often noted for their catalytic promiscuity (Field and Osbourn, 2008; Schenck and Last, 2020).

The Importance of Specialized Diterpenoid Metabolism

The most abundant class of specialized metabolites in plants are the terpenoids. Terpenoids encompass nearly 95,000 structures (*Dictionary of Natural Products 30.2*), making them the largest class of known natural plant products. Terpenoids are natively used by the plant for a variety of responses, such as microbial defense, signaling, anti-herbivory, and pollinator attraction (Trapp and Croteau, 2001; Gershenzon and Dudareva, 2007; Chen *et al.*, 2011; Tholl, 2015). As well as their natural role in plant metabolism, terpenoids can be harnessed for a variety of medicinal applications (Pateraki *et al.*, 2014; Zager *et al.*, 2019). For example, the study of terpenoids has led to the discovery of a widely used chemotherapy drug, Taxol (Guenard, Gueritte-Voegelein and Potier, 1993; Croteau *et al.*, 2006), and anti-malarial drug, artemisinin (Klayman, 1985). Terpenoids are also used in industry as flavors, fragrances,

and perfumes (Daviet and Schalk, 2010; Lange *et al.*, 2011; Philippe *et al.*, 2014; Celedon and Bohlmann, 2016). While terpenes have provided practical uses in humanity's daily life since recorded history, we have only scratched the surface of terpene discovery and research, with many yet to be discovered and functionally characterized.

Terpenes are derived from two five-carbon precursor molecules, isopentyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). Production of terpenes is spatially separated, with isoprene diphosphate units produced by plants in both the mevalonate pathway (MVA) in the cytosol and the methylerythritol 4- phosphate (MEP) pathway in the plastids (Figure 1.2). Typically, triterpenes (30C) and sesquiterpenes (15C) are produced in the cytosol by the MVA pathway, and diterpenes (20C) and monoterpenes (10C) are produced in the plastid by the MEP pathway (Figure 1.2). Terpene synthases (TPSs) are classified into clades TPS-a—h based on phylogenetic relationships. Diterpene synthases (diTPSs), the main TPS focus of this project, are differentiated by their enzymatic function into class I or class II. Class II diTPSs typically act first, with a proton mediated cyclization, followed by cleavage of the diphosphate by a class I diTPSs (Figure 1.2). In some cases, one enzyme can be bifunctional, performing both the class II/I functions (Chen *et al.*, 2011). Nearly all land plants contain two diTPS, as the ubiquitous plant growth promoting hormone gibberellic acid is a diterpenoid. The bifunctional *Physcomitrium patens* (formerly *Physcomitrella patens*) PpEKS enzyme has been hypothesized to be most conserved to the first diTPS, with all other diTPSs arising via duplication and neofunctionalization (Bohlmann, Meyer-Gauen and Croteau, 1998; Trapp and Croteau, 2001; Jiang *et al.*, 2019). Although, new research suggests a bacterial origin for plant terpene synthases (Chen *et al.*, 2024).

Gene duplication has been hypothesized to be a major driver of the evolution and expansion of all TPSs in plants (Chen *et al.*, 2011; Karunanithi and Zerbe, 2019). This is because neo- or subfunctionalization which may cause one or a few amino acid alterations can completely change the product (Potter *et al.*, 2016). One interesting type of neofunctionalization seen in terpene evolution is the addition or removal

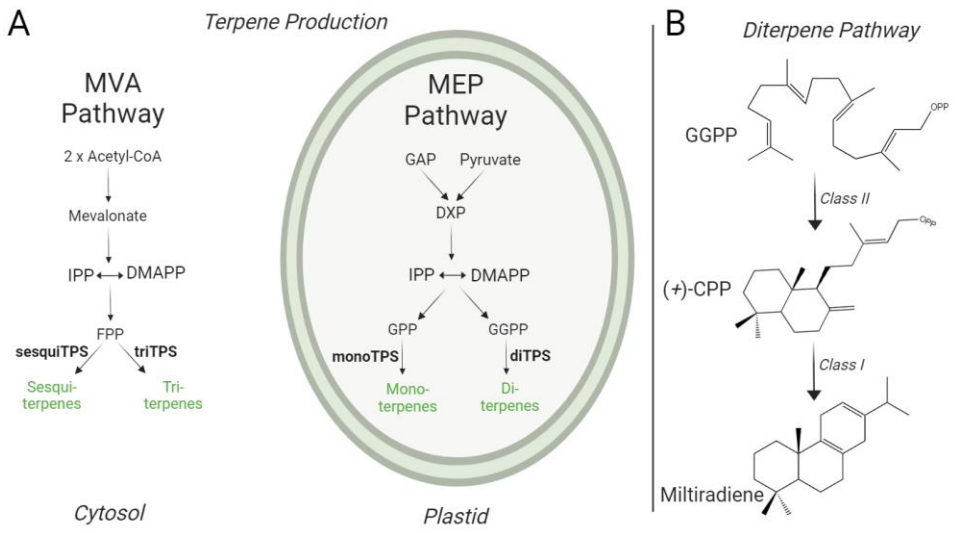


Figure 1.2 Terpenoid metabolism in plants. A) The mevalonate pathway is present in the cytosol and produces sesquiterpenes and triterpenes, while monoterpenes and diterpenes are usually produced by the methylerythritol 4-phosphate pathway in the plastid. **B)** Many diterpenes begin from the 20-carbon precursor geranyl geranyl diphosphate. They are then often acted upon by a class II and class I enzyme, respectively, forming a diterpene.

of a plastidial targeting sequence. Since terpene production is spatially separated, a change in an enzyme's environment can expose it to new substrates and it may evolve new chemistry. It has been noted that compartment switching of TPSs can promote or take advantage of enzyme promiscuity, resulting in the production of novel. It is proposed that there have been at least five independent evolutions of compartment switching of TPSs in dicots based on phylogenetic evidence terpenoids (Mau and West, 1994; Ennajdaoui *et al.*, 2010; Vaughan *et al.*, 2013; Johnson, Bhat, Sadre, *et al.*, 2019; Miller *et al.*, 2020). It is also theorized that some of the TPS-a sesquiTPSs that gained a plastidial targeting sequence evolved back into diTPSs, which explains the presence of diTPSs in the evolutionarily distinct lineage of TPS-a (Johnson, Bhat, Sadre, *et al.*, 2019). Additionally, in *Tripterygium wilfordii*, there is an instance of a monoterpene TPS-b enzyme which has neofunctionalized into a diTPS (Hansen *et al.*, 2017). TPSs are often observed catalyzing more than one reaction, with substrate promiscuity being an opportunity for plant metabolic evolution (Pateraki *et al.*, 2014; Boutanaev *et al.*, 2015; Johnson, Bhat, Bibik, *et al.*, 2019). Their affinity for diversification via duplication, their spatially separated chemistry,

and their substrate promiscuity all provide an excellent opportunity for terpenoid metabolism to be organized into a BGC.

Other enzyme classes further functionalize terpenes to increase bioactivity. Cytochrome P450 oxygenases (CYPs), particularly in the expansive CYP71 clan, frequently oxidize terpene backbones (Bathe and Tissier, 2019). The CYP71 clan is unique to land plants and includes over half of the CYPs in plants (Nelson and Werck-Reichhart, 2011; Hamberger and Bak, 2013). Specifically, CYPs in the Lamiaceae family often belong to the subfamilies CYP71BE, CYP71D, CYP76AH, CYP76AK and CYP76BK (Bathe and Tissier, 2019). CYPs often have stereo-and regiospecific on the terpene backbone where they are most likely to act. However, they are also known for their substrate promiscuity, regularly acting on a wide variety of products both *in vivo* and *in vitro*. CYPs can perform a variety of reactions, including oxidation, dehydration, reduction, ring extensions, C-C cleavage, and more (Bathe and Tissier, 2019). Besides CYPs, other oxidoreductases that can act on terpene products include 2-oxoglutarate dependent oxygenases (2OGDs) and short chain dehydrogenases. After oxygenation, additional functional groups can then be added by enzymes such as BAHD acyl transferases and glycosyl transferases. All these modifications serve to enhance bioavailability of terpenoids as well as increasing the sheer number of terpenoids in nature.

Project Goals and Significance

In this dissertation, I investigate diterpenoid metabolism and genomic landscape in the Lamiaceae family. In Chapter 2, I surveyed a representative panel of Lamiaceae species and found a syntenic diterpenoid BGC relating to miltiradiene synthesis. By homing in on *Callicarpa americana*, my co-author Emily Lanier and I elucidated the synthesis of a novel compound, (+)-kaurene. I also date this cluster as originating before the Lamiaceae by comparing it to the Lamiales outgroup *Erythranthe lutea.* In Chapter 3, I sequenced the genome of wall germander, *Teucrium chamaedrys*. Its large genome (3 Gbp) and tetraploidy afforded this species evolutionary flexibility, especially reflected in the prevalence of its diTPS

content. With around 74 putative diTPS sequences present, it is one of the most diTPS-rich species known. Additionally, I found that most of its diTPS content is present within four copies of a large biosynthetic gene cluster. Using comparative genomics, genomic sequencing, transcriptomics and biochemical analysis of diterpenoids, this work demonstrates the importance of the understudied Lamiaceae family and its wealth of important diterpenoid compounds.

REFERENCES

Arı, S. *et al.* (2015) 'Ethnobotanical survey of plants used in Afyonkarahisar-Turkey', *Journal of Ethnobiology and Ethnomedicine*, 11(1), pp. 1–15. Available at: https://doi.org/10.1186/s13002-015-0067-6.

Bathe, U. and Tissier, A. (2019) 'Cytochrome P450 enzymes: A driving force of plant diterpene diversity', *Phytochemistry*, 161, pp. 149–162. Available at: https://doi.org/10.1016/j.phytochem.2018.12.003.

Birchler, J.A. and Yang, H. (2022) 'The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation', *The Plant Cell*, 34(7), pp. 2466–2474. Available at: https://doi.org/10.1093/plcell/koac076.

Boachon, B. *et al.* (2018) 'Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae', *Molecular Plant*, 11(8), pp. 1084–1096. Available at: https://doi.org/10.1016/j.molp.2018.06.002.

Bohlmann, J., Meyer-Gauen, G. and Croteau, R. (1998) 'Plant terpenoid synthases: Molecular biology and phylogenetic analysis', *Proceedings of the National Academy of Sciences*, 95(8), pp. 4126–4133. Available at: https://doi.org/10.1073/pnas.95.8.4126.

Boutanaev, A.M. *et al.* (2015) 'Investigation of terpene diversification across multiple sequenced plant genomes', *Proceedings of the National Academy of Sciences*, 112(1), pp. E81–E88. Available at: https://doi.org/10.1073/pnas.1419547112.

Celedon, J.M. and Bohlmann, J. (2016) 'Chapter Three - Genomics-based discovery of plant genes for synthetic biology of terpenoid fragrances: A case study in sandalwood oilbiosynthesis', in S.E. O'Connor (ed.) *Methods in Enzymology*. Academic Press (Synthetic Biology and Metabolic Engineering in Plants and Microbes Part B: Metabolism in Plants), pp. 47–67. Available at: https://doi.org/10.1016/bs.mie.2016.03.008.

Chen, F. *et al.* (2011) 'The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom', *The Plant Journal*, 66(1), pp. 212–229. Available at: https://doi.org/10.1111/j.1365-313X.2011.04520.x.

Chen, X. *et al.* (2024) 'Discovery of bifunctional diterpene cyclases/synthases in bacteria supports a bacterial origin for the plant terpene synthase gene family', *Horticulture Research*, p. uhae221. Available at: https://doi.org/10.1093/hr/uhae221.

Chu, H.Y., Wegel, E. and Osbourn, A. (2011) 'From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants', *The Plant Journal: For Cell and Molecular Biology*, 66(1), pp. 66–79. Available at: https://doi.org/10.1111/j.1365-313X.2011.04503.x.

Clark, J.W. and Donoghue, P.C.J. (2018) 'Whole-Genome Duplication and Plant Macroevolution', *Trends in Plant Science*, 23(10), pp. 933–945. Available at: https://doi.org/10.1016/j.tplants.2018.07.006.

Croteau, R. *et al.* (2006) 'Taxol biosynthesis and molecular genetics', *Phytochemistry Reviews*, 5(1), pp. 75–97. Available at: https://doi.org/10.1007/s11101-005-3748-2.

Dai, L. *et al.* (2015) 'Functional characterization of cucurbitadienol synthase and triterpene glycosyltransferase involved in biosynthesis of mogrosides from *Siraitia grosvenorii*', *Plant and Cell Physiology*, 56(6), pp. 1172–1182. Available at: https://doi.org/10.1093/pcp/pcv043.

Daviet, L. and Schalk, M. (2010) 'Biotechnology in plant essential oil production: progress and perspective in metabolic engineering of the terpene pathway', p. 6.

Dictionary of Natural Products 30.2 (no date). Available at: https://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml (Accessed: 11 March 2022).

Díez, B. *et al.* (1990) 'The cluster of penicillin biosynthetic genes. Identification and characterization of the pcbAB gene encoding the alpha-aminoadipyl-cysteinyl-valine synthetase and linkage to the pcbC and penDE genes', *The Journal of Biological Chemistry*, 265(27), pp. 16358–16365.

Ennajdaoui, H. *et al.* (2010) 'Trichome specific expression of the tobacco (*Nicotiana sylvestris*) cembratrien-ol synthase genes is controlled by both activating and repressing cis-regions', *Plant Molecular Biology*, 73(6), pp. 673–685. Available at: https://doi.org/10.1007/s11103-010-9648-x.

Field, B. *et al.* (2011) 'Formation of plant metabolic gene clusters within dynamic chromosomal regions', *Proceedings of the National Academy of Sciences*, 108(38), pp. 16116–16121. Available at: https://doi.org/10.1073/pnas.1109273108.

Field, B. and Osbourn, A.E. (2008) 'Metabolic Diversification—Independent Assembly of Operon-Like Gene Clusters in Different Plants', *Science*, 320(5875), pp. 543–547. Available at: https://doi.org/10.1126/science.1154990.

Fierro, F. *et al.* (1993) 'Resolution of four large chromosomes in penicillin-producing filamentous fungi: the penicillin gene cluster is located on chromosome II (9.6 Mb) in *Penicillium notatum* and chromosome 1 (10.4 Mb) in *Penicillium chrysogenum*', *Molecular and General Genetics MGG*, 241(5), pp. 573–578. Available at: https://doi.org/10.1007/BF00279899.

Frey, M. *et al.* (1997) 'Analysis of a chemical plant defense mechanism in grasses', *Science*, 277(5326), pp. 696–699. Available at: https://doi.org/10.1126/science.277.5326.696.

Gershenzon, J. and Dudareva, N. (2007) 'The function of terpene natural products in the natural world', *Nature Chemical Biology*, 3(7), pp. 408–414. Available at: https://doi.org/10.1038/nchembio.2007.5.

Godden, G.T. *et al.* (2019) 'Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints', *Genome Biology and Evolution*, 11(12), pp. 3393–3408. Available at: https://doi.org/10.1093/gbe/evz239.

Guenard, D., Gueritte-Voegelein, F. and Potier, P. (1993) 'Taxol and taxotere: Discovery, chemistry, and structure-activity relationships', *Accounts of Chemical Research*, 26(4), pp. 160–167. Available at: https://doi.org/10.1021/ar00028a005.

Hamberger, B. and Bak, S. (2013) 'Plant P450s as versatile drivers for evolution of species-specific chemical diversity', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1612), p. 20120426. Available at: https://doi.org/10.1098/rstb.2012.0426.

Hansen, N.L. *et al.* (2017) 'The terpene synthase gene family in *Tripterygium wilfordii* harbors a labdane-type diterpene synthase among the monoterpene synthase TPS-b subfamily', *The Plant Journal*, 89(3), pp. 429–441. Available at: https://doi.org/10.1111/tpj.13410.

He, Y. *et al.* (2016) 'Survey of the genome of *Pogostemon cablin* provides insights into its evolutionary history and sesquiterpenoid biosynthesis', *Scientific Reports*, 6(1), p. 26405. Available at: https://doi.org/10.1038/srep26405.

Itkin, M. *et al.* (2013) 'Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes', *Science (New York, N.Y.)*, 341(6142), pp. 175–179. Available at: https://doi.org/10.1126/science.1240230.

Jarić, S., Mitrović, M. and Pavlović, P. (2020) 'Ethnobotanical Features of *Teucrium* Species', in M. Stanković (ed.) *Teucrium Species: Biology and Applications*. Cham: Springer International Publishing, pp. 111–142. Available at: https://doi.org/10.1007/978-3-030-52159-2_5.

Jiang, S.-Y. *et al.* (2019) 'A comprehensive survey on the terpene synthase gene family provides new insight into its evolutionary patterns', *Genome Biology and Evolution*, 11(8), pp. 2078–2098. Available at: https://doi.org/10.1093/gbe/evz142.

Johnson, S.R., Bhat, W.W., Bibik, J., *et al.* (2019) 'A database-driven approach identifies additional diterpene synthase activities in the mint family (Lamiaceae)', *Journal of Biological Chemistry*, 294(4), pp. 1349–1362. Available at: https://doi.org/10.1074/jbc.RA118.006025.

Johnson, S.R., Bhat, W.W., Sadre, R., *et al.* (2019) 'Promiscuous terpene synthases from *Prunella vulgaris* highlight the importance of substrate and compartment switching in terpene synthase evolution', *New Phytologist*, 223(1), pp. 323–335. Available at: https://doi.org/10.1111/nph.15778.

Karunanithi, P.S. and Zerbe, P. (2019) 'Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity', *Frontiers in Plant Science*, 10, p. 1166. Available at: https://doi.org/10.3389/fpls.2019.01166.

Kitaoka, N. *et al.* (2021) 'Interdependent evolution of biosynthetic gene clusters for momilactone production in rice', *The Plant Cell*, 33(2), pp. 290–305. Available at: https://doi.org/10.1093/plcell/koaa023.

Klayman, D.L. (1985) 'Qinghaosu (Artemisinin): an Antimalarial Drug from China', *Science*, 228(4703), pp. 1049–1055. Available at: https://doi.org/10.1126/science.3887571.

Kristensen, C. *et al.* (2005) 'Metabolic engineering of dhurrin in transgenic *Arabidopsis* plants with marginal inadvertent effects on the metabolome and transcriptome', *Proceedings of the National Academy of Sciences of the United States of America*, 102(5), pp. 1779–1784. Available at: https://doi.org/10.1073/pnas.0409233102.

Lange, B.M. *et al.* (2011) 'Improving peppermint essential oil yield and composition by metabolic engineering', *Proceedings of the National Academy of Sciences*, 108(41), pp. 16944–16949. Available at: https://doi.org/10.1073/pnas.1111558108.

Li, P. *et al.* (2017) 'Molecular phylogenetics and biogeography of the mint tribe Elsholtzieae (Nepetoideae, Lamiaceae), with an emphasis on its diversification in East Asia', *Scientific Reports*, 7(1), p. 2057. Available at: https://doi.org/10.1038/s41598-017-02157-6.

Liang, J. *et al.* (2021) 'Rice contains a biosynthetic gene cluster associated with production of the casbane-type diterpenoid phytoalexin ent-10-oxodepressin', *New Phytologist*, 231(1), pp. 85–93. Available at: https://doi.org/10.1111/nph.17406.

Liu, Z. *et al.* (2020) 'Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae', *New Phytologist*, 227(4), pp. 1109–1123. Available at: https://doi.org/10.1111/nph.16338.

Lu, X. *et al.* (2018) 'Inferring Roles in Defense from Metabolic Allocation of Rice Diterpenoids', *The Plant Cell*, 30(5), pp. 1119–1131. Available at: https://doi.org/10.1105/tpc.18.00205.

Matsuba, Y. *et al.* (2013) 'Evolution of a complex locus for terpene biosynthesis in *Solanum*', *The Plant Cell*, 25(6), pp. 2022–2036. Available at: https://doi.org/10.1105/tpc.113.111013.

Mau, C.J. and West, C.A. (1994) 'Cloning of casbene synthase cDNA: evidence for conserved structural features among terpenoid cyclases in plants.', *Proceedings of the National Academy of Sciences*, 91(18), pp. 8497–8501. Available at: https://doi.org/10.1073/pnas.91.18.8497.

Medema, M.H. *et al.* (2015) 'Minimum information about a biosynthetic gene cluster', *Nature Chemical Biology*, 11(9), pp. 625–631. Available at: https://doi.org/10.1038/nchembio.1890.

Miller, G.P. *et al.* (2020) 'The biosynthesis of the anti-microbial diterpenoid leubethanol in *Leucophyllum frutescens* proceeds via an all-cis prenyl intermediate', *The Plant Journal*, 104(3), pp. 693–705. Available at: https://doi.org/10.1111/tpj.14957.

Mugford, S.T. *et al.* (2013) 'Modularity of plant metabolic gene clusters: A trio of linked genes that are collectively required for acylation of triterpenes in oat', *The Plant Cell*, 25(3), pp. 1078–1092. Available at: https://doi.org/10.1105/tpc.113.110551.

Mylona, P. *et al.* (2008) 'Sad3 and Sad4 Are Required for Saponin Biosynthesis and Root Development in Oat', *The Plant Cell*, 20(1), pp. 201–212. Available at: https://doi.org/10.1105/tpc.107.056531.

Nelson, D. and Werck-Reichhart, D. (2011) 'A P450-centric view of plant evolution', *The Plant Journal*, 66(1), pp. 194–211. Available at: https://doi.org/10.1111/j.1365-313X.2011.04529.x.

Nützmann, H.-W., Huang, A. and Osbourn, A. (2016) 'Plant metabolic clusters – from genetics to genomics', *New Phytologist*, 211(3), pp. 771–789. Available at: https://doi.org/10.1111/nph.13981.

Nützmann, H.-W., Scazzocchio, C. and Osbourn, A. (2018) 'Metabolic gene clusters in eukaryotes', *Annual Review of Genetics*, 52(1), pp. 159–183. Available at: https://doi.org/10.1146/annurev-genet-120417-031237.

Ohno, S. (1970) 'Duplication for the Sake of Producing More of the Same', in S. Ohno (ed.) *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer, pp. 59–65. Available at: https://doi.org/10.1007/978-3-642-86659-3_11.

Pateraki, I. *et al.* (2014) 'Manoyl oxide (13R), the biosynthetic precursor of forskolin, is synthesized in specialized root cork cells in *Coleus forskohlii*', *Plant Physiology*, 164(3), pp. 1222–1236. Available at: https://doi.org/10.1104/pp.113.228429.

Philippe, R.N. *et al.* (2014) 'Biotechnological production of natural zero-calorie sweeteners', *Current Opinion in Biotechnology*, 26, pp. 155–161. Available at: https://doi.org/10.1016/j.copbio.2014.01.004.

Pieroni, A., Quave, C.L. and Santoro, R.F. (2004) 'Folk pharmaceutical knowledge in the territory of the Dolomiti Lucane, inland southern Italy', *Journal of Ethnopharmacology*, 95(2), pp. 373–384. Available at: https://doi.org/10.1016/j.jep.2004.08.012.

Polturak, G. and Osbourn, A. (2021) 'The emerging role of biosynthetic gene clusters in plant defense and plant interactions', *PLOS Pathogens*, 17(7), p. e1009698. Available at: https://doi.org/10.1371/journal.ppat.1009698.

Postnikova, O.A. *et al.* (2011) 'Clustering of pathogen-response genes in the genome of *Arabidopsis thaliana*', *Journal of Integrative Plant Biology*, 53(10), pp. 824–834. Available at: https://doi.org/10.1111/j.1744-7909.2011.01071.x.

Potter, K.C. *et al.* (2016) 'Blocking Deprotonation with Retention of Aromaticity in a Plant ent-Copalyl Diphosphate Synthase Leads to Product Rearrangement', *Angewandte Chemie International Edition*, 55(2), pp. 634–638. Available at: https://doi.org/10.1002/anie.201509060.

Qi, X. *et al.* (2004) 'A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants', *Proceedings of the National Academy of Sciences*, 101(21), pp. 8233–8238. Available at: https://doi.org/10.1073/pnas.0401301101.

Ren, R. *et al.* (2018) 'Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms', *Molecular Plant*, 11(3), pp. 414–428. Available at: https://doi.org/10.1016/j.molp.2018.01.002.

Sakamoto, T. *et al.* (2004) 'An overview of gibberellin metabolism enzyme genes and their related mutants in rice', *Plant Physiology*, 134(4), pp. 1642–1653. Available at: https://doi.org/10.1104/pp.103.033696.

Schenck, C.A. and Last, R.L. (2020) 'Location, location! Cellular relocalization primes specialized metabolic diversification', *The FEBS Journal*, 287(7), pp. 1359–1368. Available at: https://doi.org/10.1111/febs.15097.

Slot, J.C. and Hibbett, D.S. (2007) 'Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: A phylogenetic study', *PLOS ONE*, 2(10), p. e1097. Available at: https://doi.org/10.1371/journal.pone.0001097.

Slot, J.C. and Rokas, A. (2010) 'Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi', *Proceedings of the National Academy of Sciences*, 107(22), pp. 10136–10141. Available at: https://doi.org/10.1073/pnas.0914418107.

Takos, A.M. and Rook, F. (2012) 'Why biosynthetic genes for chemical defense compounds cluster', *Trends in Plant Science*, 17(7), pp. 383–388. Available at: https://doi.org/10.1016/j.tplants.2012.04.004.

Tholl, D. (2015) 'Biosynthesis and biological functions of terpenoids in plants', in J. Schrader and J. Bohlmann (eds) *Biotechnology of Isoprenoids*. Cham: Springer International Publishing (Advances in Biochemical Engineering/Biotechnology), pp. 63–106. Available at: https://doi.org/10.1007/10_2014_295.

di Tizio, A. *et al.* (2012) 'Traditional food and herbal uses of wild plants in the ancient South-Slavic diaspora of Mundimitar/Montemitro (Southern Italy)', *Journal of Ethnobiology and Ethnomedicine*, 8(1), p. 21. Available at: https://doi.org/10.1186/1746-4269-8-21.

Trapp, S.C. and Croteau, R.B. (2001) 'Genomic organization of plant terpene synthases and molecular evolutionary implications', *Genetics*, 158(2), pp. 811–832. Available at: https://doi.org/10.1093/genetics/158.2.811.

Vaughan, M.M. *et al.* (2013) 'Formation of the Unusual Semivolatile Diterpene Rhizathalene by the Arabidopsis Class I Terpene Synthase TPS08 in the Root Stele Is Involved in Defense against Belowground Herbivory', *The Plant Cell*, 25(3), pp. 1108–1125. Available at: https://doi.org/10.1105/tpc.112.100057.

Wilderman, P.R. *et al.* (2004) 'Identification of syn-pimara-7,15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis', *Plant Physiology*, 135(4), pp. 2098–2105. Available at: https://doi.org/10.1104/pp.104.045971.

Winzer, T. *et al.* (2012) 'A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine', *Science*, 336(6089), pp. 1704–1708. Available at: https://doi.org/10.1126/science.1220757.

Xu, M. *et al.* (2012) 'Genetic evidence for natural product-mediated plant–plant allelopathy in rice (*Oryza sativa*)', *New Phytologist*, 193(3), pp. 570–575. Available at: https://doi.org/10.1111/j.1469-8137.2011.04005.x.

Yao, G. *et al.* (2016) 'Phylogenetic relationships, character evolution and biogeographic diversification of *Pogostemon* s.l. (Lamiaceae)', *Molecular Phylogenetics and Evolution*, 98, pp. 184–200. Available at: https://doi.org/10.1016/j.ympev.2016.01.020.

Yu, J. *et al.* (1995) 'Comparative mapping of aflatoxin pathway gene clusters in *Aspergillus parasiticus* and *Aspergillus flavus*', *Applied and Environmental Microbiology*, 61(6), pp. 2365–2371. Available at: https://doi.org/10.1128/aem.61.6.2365-2371.1995.

Yu, N. *et al.* (2016) 'Delineation of metabolic gene clusters in plant genomes by chromatin signatures', *Nucleic Acids Research*, 44(5), pp. 2255–2265. Available at: https://doi.org/10.1093/nar/gkw100.

Zager, J.J. *et al.* (2019) 'Gene networks underlying cannabinoid and terpenoid accumulation in *Cannabis*', *Plant Physiology*, 180(4), pp. 1877–1897. Available at: https://doi.org/10.1104/pp.18.01506.

Zhao, F. *et al.* (2021) 'An updated tribal classification of Lamiaceae based on plastome phylogenomics', *BMC Biology*, 19(1), p. 2. Available at: https://doi.org/10.1186/s12915-020-00931-z.

CHAPTER 2: 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 has been adapted from the following published article:

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 (Postnikova *et al.*, 2011; Boutanaev *et al.*, 2015a; Medema *et al.*, 2015; Nützmann, Huang and Osbourn, 2016; Nützmann, Scazzocchio and Osbourn, 2018; Liu, Suarez Duran, *et al.*, 2020). Over 30 plant BGCs have been functionally validated to date (Polturak and Osbourn, 2021) since the discovery of the first BGC in maize (Frey *et al.*, 1997). The BGCs found in plants are predominately involved in specialized rather than central metabolism (Chu, Wegel and Osbourn, 2011) and occur in multiple classes of compounds including benzylisoquinoline alkaloids in poppy (Winzer *et al.*, 2012; Yang *et al.*, 2021), triterpenoid cucurbitacins in Cucurbitaceae (Shang *et al.*, 2014; Dai *et al.*, 2015), and diterpenoid momilactones in Poaceae and other cereals (Sakamoto *et al.*, 2004; Wilderman *et al.*, 2004; Schmelz *et al.*, 2014; Kitaoka *et al.*, 2020; Liang *et al.*, 2021).

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 (Slot and Hibbett, 2007; Slot and Rokas, 2010). 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 (Field *et al.*, 2011; Nützmann, Scazzocchio and Osbourn, 2018). The genetic linkage conveyed by BGCs facilitates coinheritance, which can protect the integrity of the entire pathway (Takos and Rook, 2012; Zhang and Peters, 2020; Ma *et al.*, 2021). In some pathways, such as momilactone biosynthesis, loss of a single gene would result in a buildup of toxic intermediates (Zhang

and Peters, 2020). 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 (Hurst, Pál and Lercher, 2004; Qi *et al.*, 2004; Okada *et al.*, 2009; Field *et al.*, 2011; Mugford *et al.*, 2013; Schmelz *et al.*, 2014; Nützmann, Scazzocchio and Osbourn, 2018). Regulation may also take place at the chromatin level, with DNA and histone methylation regulating transcription of the entire cluster (Hurst, Pál and Lercher, 2004; Yu *et al.*, 2016; Rokas, Wisecaver and Lind, 2018; Nützmann *et al.*, 2020).

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 (Nützmann, Scazzocchio and Osbourn, 2018). 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 (Li *et al.*, 2021). 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 (Field *et al.*, 2011; Nützmann, Scazzocchio and Osbourn, 2018; Rokas, Wisecaver and Lind, 2018; Liu, Cheema, *et al.*, 2020). Expansion of the cluster can also continue through tandem, local, or whole genome duplication (Field and Osbourn, 2008; Itkin *et al.*, 2013; Nützmann, Huang and Osbourn, 2016; Liu, Cheema, *et al.*, 2020; Liu, Suarez Duran, *et al.*, 2020). The inherent promiscuity of enzymes involved in specialized metabolism enables rapid neofunctionalization, promoting functional divergence of BGCs as they evolve through different plant lineages (Field and Osbourn, 2008; Matsuba *et al.*, 2013; Johnson, Bhat, Sadre, *et al.*, 2019; Schenck and Last, 2020). Recent work has shown conservation of core genes and diversification into new functions/pathways when comparing BGCs across different plant families (Fan *et al.*, 2020; Liu, Suarez Duran, *et al.*, 2020).

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 (*Dictionary of Natural Products 30.2*), 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 (Gershenzon and Dudareva, 2007; Chen *et al.*, 2011; Tholl, 2015). 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 (Bohlmann, Steele and Croteau, 1997; Chen *et al.*, 2011; Karunanithi and Zerbe, 2019). 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 (Boutanaev *et al.*, 2015b; Bathe and Tissier, 2019).

Terpenoid diversity is particularly rich in the Lamiaceae (mint) family (Lange, 2015; Johnson, Bhat, Bibik, *et al.*, 2019). Genome assemblies for 22 different Lamiaceae species (Supplementary Table 2.1) have been published to date, revealing BGCs for at least two classes of terpenoids: monoterpene-derived nepetalactones from catnip (*Nepeta* sp.; Sherden *et al.*, 2018) and diterpenoid tanshinones in the Chinese medicinal herb Danshen (*Salvia miltiorrhiza*; Xu *et al.*, 2016; Song *et al.*, 2020; Ma *et al.*, 2021). Tanshinones are studied for their potent pharmacological activities, and as a result much of the biosynthetic pathway has been elucidated (Supplementary Figure 2.1; Gao *et al.*, 2009; Ma *et al.*, 2012, 2021; Guo *et al.*, 2013, 2016; Cui *et al.*, 2015; Xu *et al.*, 2016; Bai *et al.*, 2018; Song *et al.*, 2020, 2022;

Wang and Peters, 2022). 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 (González, 2015). The antimicrobial effects demonstrated for many of these terpenoids suggest a native role in plant defense (Smith *et al.*, 2008; Machumi *et al.*, 2010; González, 2015; Abdissa, Frese and Sewald, 2017; Gao *et al.*, 2020). Carnosic acid is another abietane diterpenoid found in several Lamiaceae species with powerful antioxidant and anticancer properties (Birtić *et al.*, 2015). The biosynthesis of carnosic acid and related diterpenoids has been elucidated in *Rosmarinus officinalis, Salvia pomifera* and *Salvia fruticosa* (rosemary and sages; Ignea *et al.*, 2016; Scheler *et al.*, 2016) and involves many CYPs orthologous to those involved in tanshinone biosynthesis (Supplementary Figure 2.1).

Previous studies of the *S. miltiorrhiza* genome have found two BGCs that together contain the miltiradiene diTPSs and two CYP76AHs involved in tanshinone biosynthesis (Xu *et al.*, 2016; Song *et al.*, 2020; Ma *et al.*, 2021). 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; Zhao *et al.*, 2019; Ma *et al.*, 2021; Wang and Peters, 2022). 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 (Hamilton *et al.*, 2020). 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 survey a representative panel of 10 Lamiaceae genome assemblies (Figure 2.1). We focus on synteny with the BGC in *C. americana*, which is

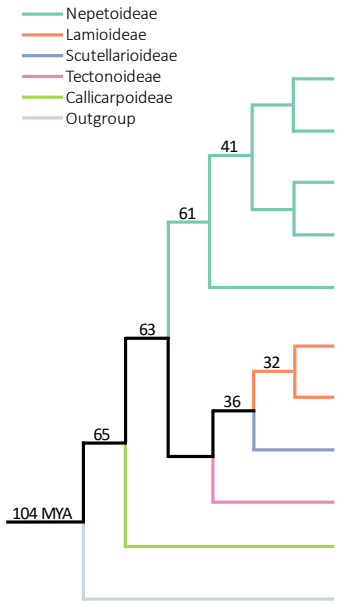
one of the largest yet discovered, spanning approximately 400 Kb and encompassing seven diTPSs and twelve CYPs. Our syntenic analysis shows conservation of core miltiradiene biosynthetic genes throughout all species studied while highlighting lineage-specific diversification of the BGC in five subfamilies. Phylogenetic analysis supports common ancestry of each enzyme class and enabled reconstruction of a minimal ancestral cluster. We find that the BGC in *C. americana* 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 (Swaminathan *et al.*, 2009; Winzer *et al.*, 2012). We also discover 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

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 new 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 2.1). 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 2.2). 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.



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 2.1 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; Yao *et al.*, 2016; Li *et al.*, 2017; Godden *et al.*, 2019). Ploidy level of species not assumed to be diploid are shown in parenthesis next to their genome size (Supplementary Table 2.2).

Benchmarking Universal Single-Copy Ortholog (BUSCO; Manni *et al.*, 2021) of pseudohaplotype-1 assemblies revealed >97% complete BUSCOs in the three genomes (Table 2.2) 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(Stanke *et al.*, 2006) matrices yielded 148,846 (*L. leonurus*), 413,222 (*P. barbatus*), and 229,613 (*P. vulgaris*) genes (Supplementary Table 2.3). 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 2.4).

	Number of Scaffolds	Total Size of Scaffolds (bp)	N50 Scaffold Length (bp)	Number of Ns (Percent Ns)	Totals Gaps (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

Table 2.1 Assembly statistics for the 10x genomics pseudohaplotype-1 of *Leonotis leonurus*, *Plectranthus barbatus*, and *Prunella vulgaris*.

	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%

Table 2.2 Benchmarking universal single copy orthologs for *Leonotis leonurus*, *Plectranthus barbatus*, and *Prunella vulgaris* pseudohaplotype-1 genomes and predicted proteomes.

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*; Song *et al.*, 2020; Ma *et al.*, 2021)*, T. grandis* (Zhao *et al.*, 2019)*,* and *S. baicalensis* (Xu *et al.*, 2020)), the three species we assembled in this

study (*L. leonurus, P. barbatus,* and *P. vulgaris),* and three species with published genomes (*Hyssopus officinalis* (Lichman *et al.*, 2020)*, R. officinalis* (Bornowski *et al.*, 2020)*,* and *Pogostemon cablin* (He *et al.*, 2018). 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 (Figure 2.1; Yao *et al.*, 2016; Li *et al.*, 2017; Godden *et al.*, 2019). As a close Lamiales outgroup, we also analyzed *Erythranthe lutea* (Monkey flower; formerly *Mimulus luteus;* Cooley *et al.*, 2022).

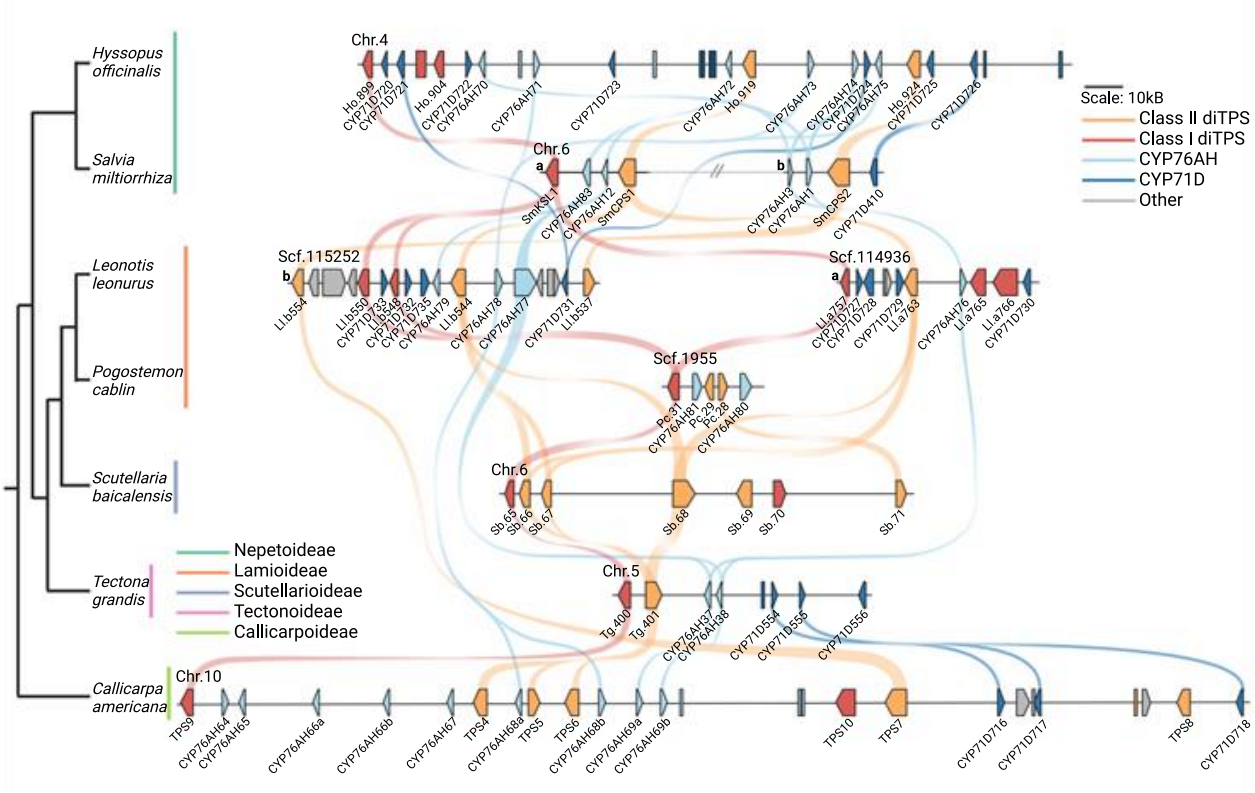


Figure 2.2 Syntenic relationships of a miltiradiene biosynthetic gene cluster present across the Lamiaceae. Genes are represented with arrows and pseudogenes are represented with boxes. A core set of genes are common to many species examined, including a diTPS class II (+)-CPP synthase, a diTPS class 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. Figure created using BioRender.com.

Out of the 10 Lamiaceae species sampled, all contained diTPSs orthologous to known (+)-CPP and miltiradiene synthases. In seven species these diTPSs were within syntenic BGCs (Figure 2.2). 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.

Comparison of the BGCs provides insight into the formation and maintenance of this cluster in divergent lineages (Figure 2.2). 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 Figure 2.2). 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

(Figure 2.3; Supplementary Table 2.5). 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

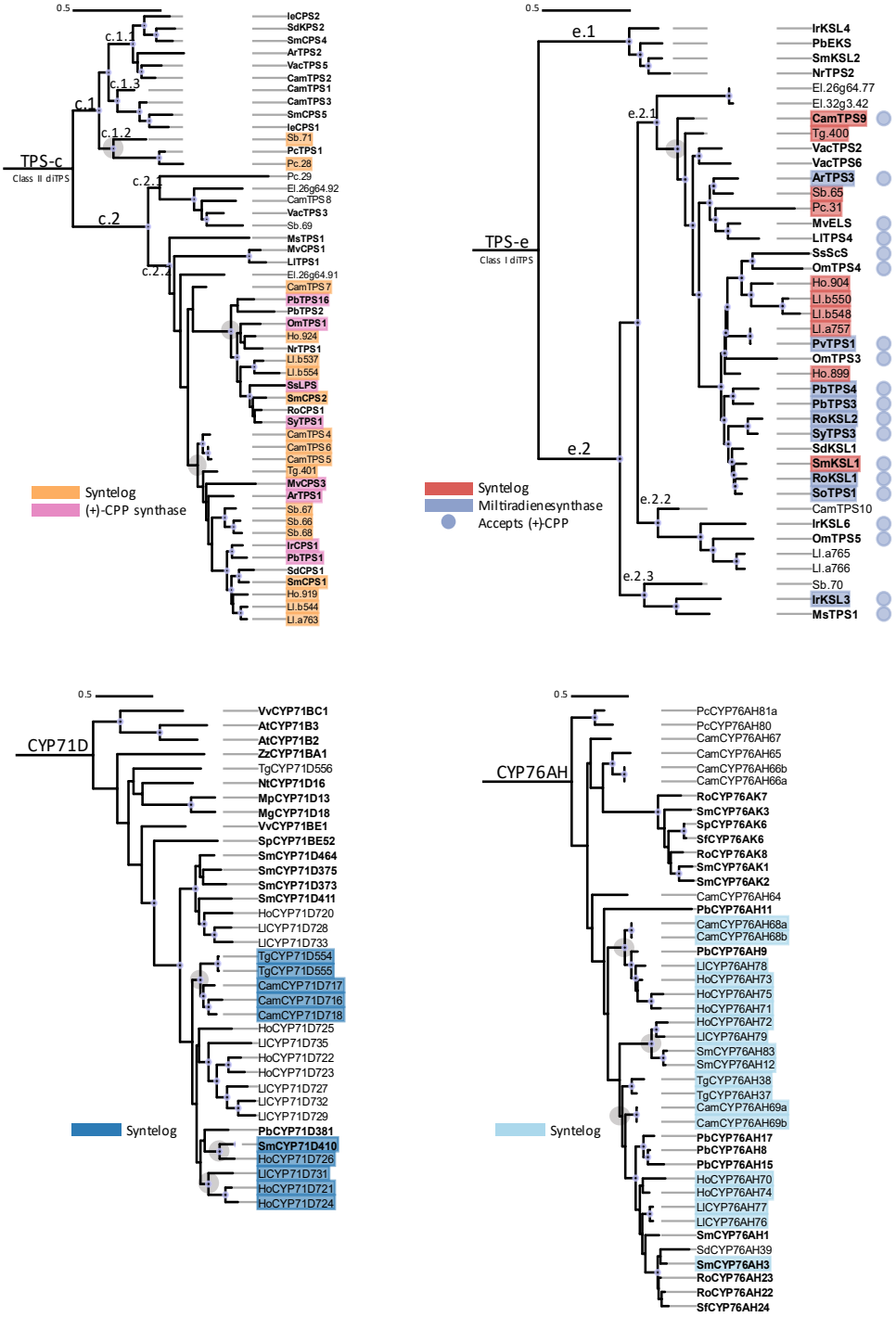


Figure 2.3 Phylogenetic evidence shows the relatedness of each gene class in the clusters. Enzymes present in each cluster with syntenic support from MCScanX and sequence identity from BLASTp are

Figure 2.3 (cont'd)
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.

TPS-e subfamily. 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 (Figure 2.3, Figure 2.4). 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 diTPSs 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.

None of the class I enzymes 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 contains enzymes that accept (+)-CPP as a substrate. The presence of a presumed (+)-CPP synthase in every BGC supports the likelihood that these class I diTPSs can all utilize (+)-CPP. 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. Notably, every BGC contains at least one of these presumed miltiradiene synthases. 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 enzymes but does offer intriguing possibilities for discovery of novel terpene backbones.

While phylogenetic classification is not a perfect predictor of TPS function (Durairaj *et al.*, 2019; Johnson, Bhat, Sadre, *et al.*, 2019), 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 (Johnson, Bhat, Bibik, *et al.*, 2019). Phylogenetic evidence supports that these BGCs likely have at minimum a (+)-CPP synthase and a miltiradiene synthase, enabling production of miltiradiene in each plant (Figure 2.3). 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 (Guo *et al.*, 2013, 2016). 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 (Bak *et al.*, 2011; Bathe and Tissier, 2019). 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 (Ma *et al.*, 2021).

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 (Pateraki *et al.*, 2017). 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 (Kliebenstein, 2008).

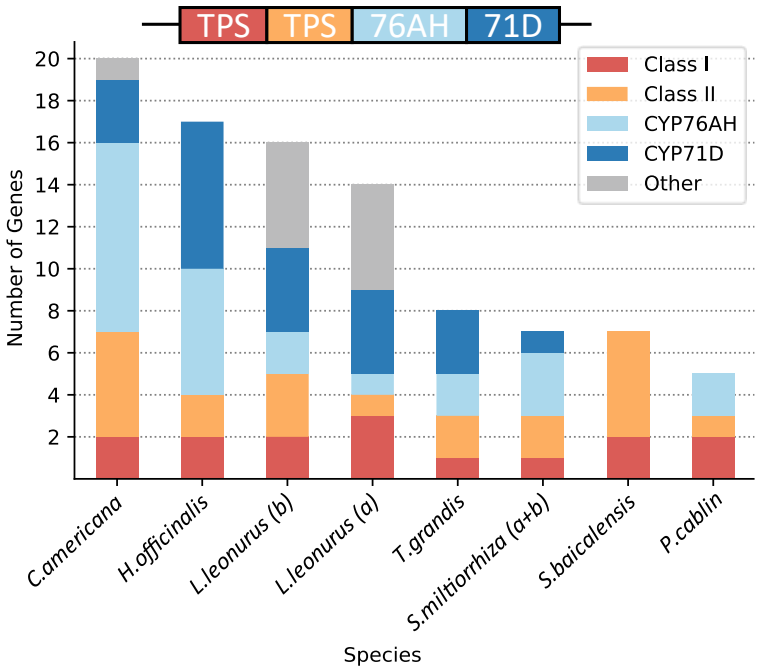


Figure 2.4 Predicted Lamiaceae minimal ancestral BGC and species-specific expansion of each. Based on maximum parsimony, we suggest that a cluster containing a class II diTPS, class I diTPS, CYP76AH, and CYP71D gene was formed in an early Lamiaceae ancestor. Lineage-specific expansion and refinement are evident from the number and composition of genes in each gene family present in the extant species studied.

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 (Figure 2.4; Supplementary Figure 2.3, Supplementary Figure 2.4). 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 *E. lutea*, a closely related Lamiales outgroup (Godden *et al.*, 2019; Liu, Tan, *et al.*, 2020; Cooley *et al.*, 2022). We found a region syntenic to the *C. americana* BGC which contains class II and class I diTPSs but no CYPs (Supplementary Figure 2.5). The class II enzymes, El.26g64.91 and El.26g64.92, are in clade TPS-c.2, showing some similarity with other (+)-CPP synthases (Figure 2.3). The class I enzyme, 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 (Hillwig *et al.*, 2011a), 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 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 (Wisecaver *et al.*, 2017). 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 (Xu *et al.*, 2016). Additionally, an array of root-specific CYP71Ds were also integral to tanshinone biosynthesis but located elsewhere in the genome (Ma *et al.*, 2021). 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; Swaminathan *et al.*, 2009) and the noscapine/morphinan biosynthesis in *Papaver ssp.* (poppy; Bai *et al.*, 2018; Yang *et al.*, 2021). 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.

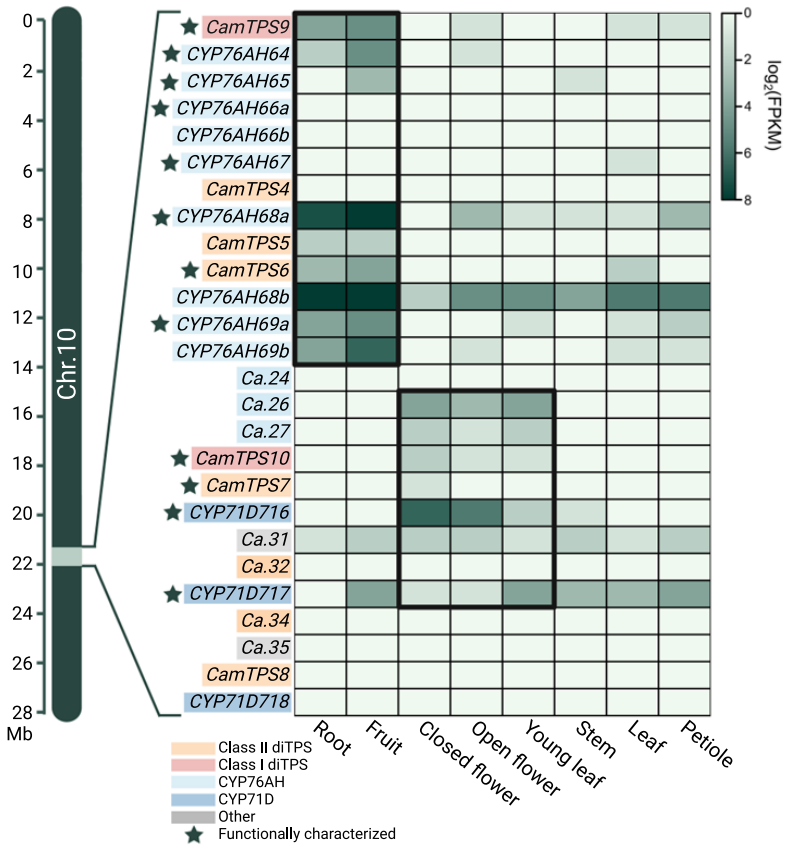


Figure 2.5 Tissue specific expression of a miltiradiene BGC in *C. americana* obtained from RNA sequencing. Functional characterization of these enzymes refers to this study. This figure represents

Figure 2.5 (cont'd)
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.

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 (Figure 2.5; Supplementary Table 2.7, Supplementary Figure 2.6)(Hamilton *et al.*, 2020). 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*; Hamilton *et al.*, 2020), 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 looked at expression of each of BGC in the other species with published transcriptomic data but found no overarching expression trends, unlike in *C. americana* (Supplementary Figure 2.6, Supplementary Table 2.7).

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 2.6). CamTPS7 was confirmed to be a (+)-CPP synthase (Supplementary Figure 2.6). CamTPS9 is a miltiradiene (**1**) synthase, with some abietatriene

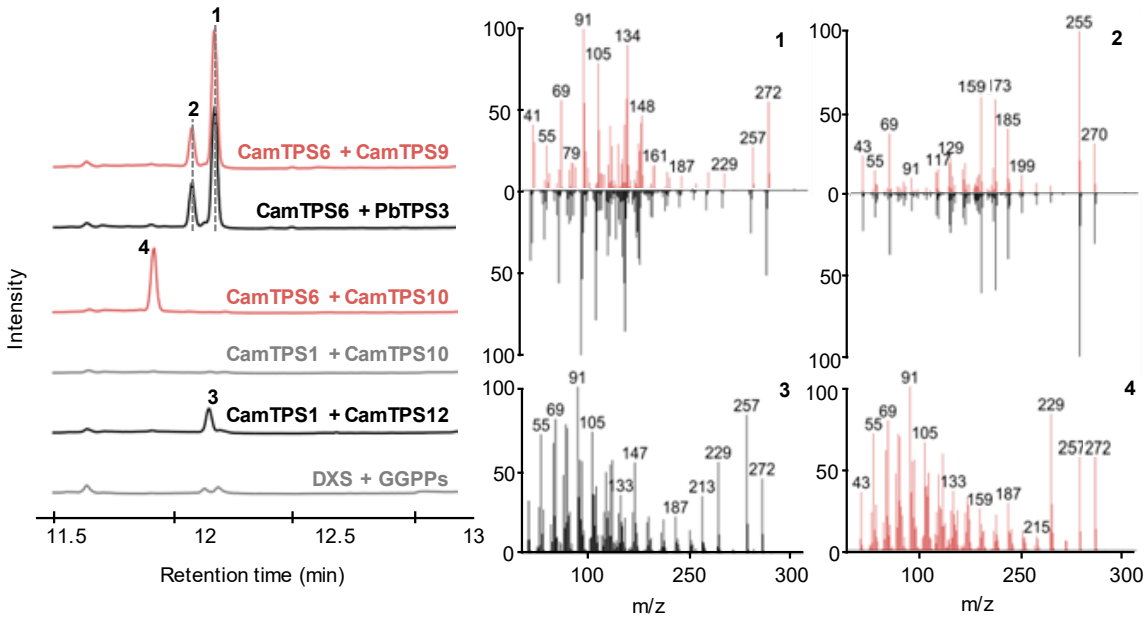


Figure 2.6 GC-MS analysis of *C. americana* BGC diTPS products. CamTPS9 was confirmed to be a miltiradiene synthase by comparison with the retention time and mass spectra of PbTPS3125–128 products when both were expressed with the (+)-CPP synthase CamTPS6[69], forming miltiradiene (1) and abietatriene (2). CamTPS10 was found to make 4 from (+)-CPP but not ent-CPP (CamTPS1)69. This product has a different retention time but similar mass spectrum to ent-kaurene (3), made by the combination of CamTPS1 and CamTPS12 (Supplementary Figure 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.

(**2**; *ent*-abieta-8,11,13-triene) resulting from spontaneous aromatization *in plantae* consistent with previous observations (Zi and Peters, 2013). CamTPS10, when paired with a (+)-CPP synthase, forms (+)-kaurene (**4**), a previously unknown diTPS activity (Supplementary Figure 2.8). 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 (Jones and Kinghorn, 2008). Calliterpenone has been investigated for its potential as a plant growth promoting agent (Bose *et al.*, 2013), 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 (Andersen-Ranberg *et al.*, 2016). 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.

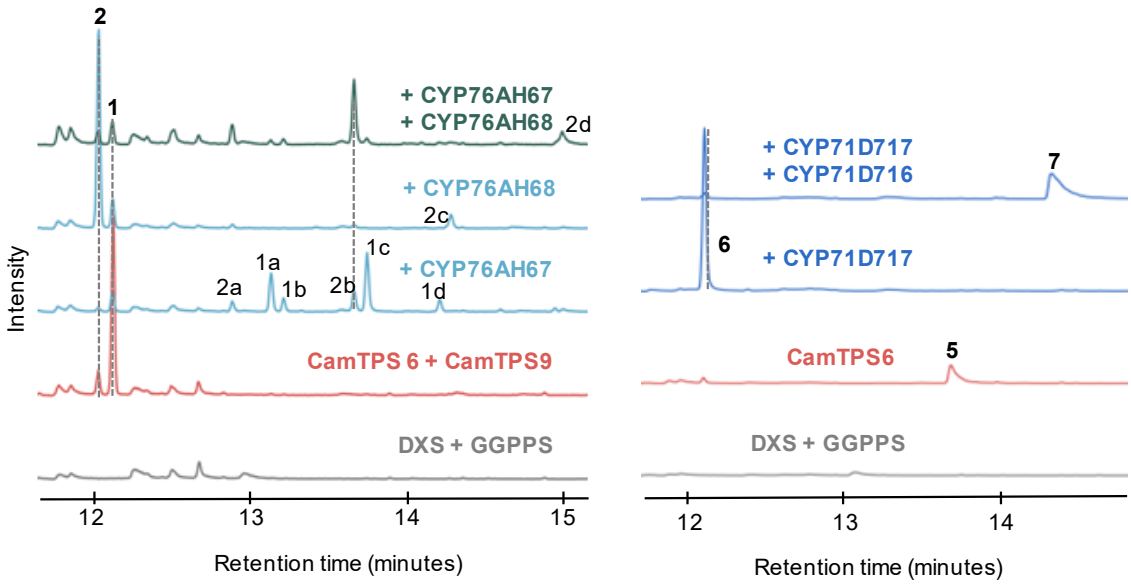


Figure 2.7 GC-MS chromatograms showing oxidation products of *C. americana* BGC CYPs. **a** Oxidation products of the CamCYP76AHs from 1 and 2, assigned based on analysis of mass spectra (Supplementary Figure 12). **b** CamCYP71D717 catalyzes the production of (+)-manool (6), likely from (+)-copalol (5) (Supplementary Figure 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.

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 (Figure 2.7): *ent*-kaurene (CamTPS12; Supplementary Figure 2.9) and kolavenol (Hamilton *et al.*, 2020) 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**, Figure 2.7a). 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 Figure 2.10). 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 Figure 2.11). Oxidation in this position on an abietane diterpene has only been reported twice before: by a 2-oxoglutarate dehydrogenase in *S. miltiorrhiza* (Hu *et al.*, 2022) and by CYP81AM1 in *Tripterygium wilfordii* (Wang *et al.*, 2021). CamCYP76AH68 also showed activity with miltiradiene, dramatically shifting the product profile towards abietatriene and affording a small amount of oxidized abietatriene (**2c**; Supplementary Figure 2.10). 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 (Figure 2.8a). 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 (Zi and Peters, 2013). 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 (Figure 2.7b, Supplementary Figure 12). 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 (**2d**, Figure 2.7a; Supplementary Figure 2.10). 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 (Figure 2.7b, Figure 8b; Supplementary Figure 13). No abietane-type diterpenoids were previously reported in *C. americana,* which has been primarily studied

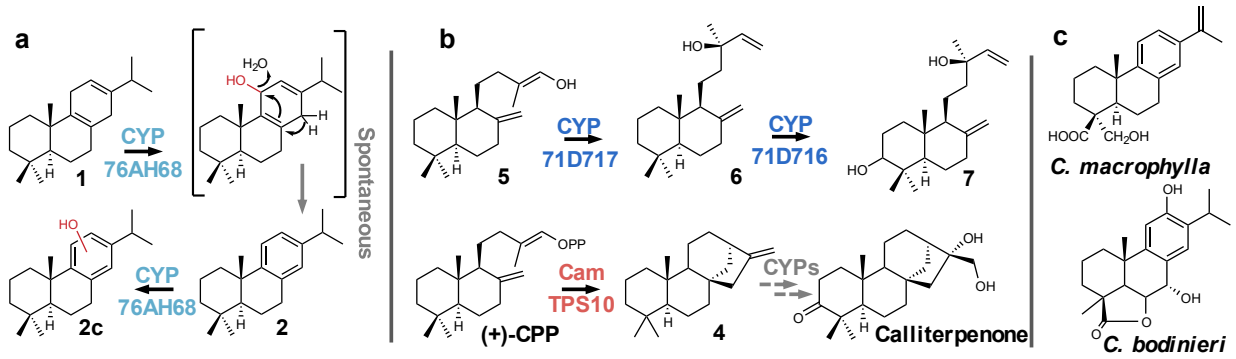


Figure 2.8 Pathway schematic for CYP oxidations in *C. americana*. **a** Proposed mechanism for enzyme-assisted conversion of **1** to **2**, followed by an additional oxidation of **2** to form **2c**. Mass spectra supports assignment of the hydroxy group in **2c** to the c-ring (Supplementary Figure 2.12). **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*.

for clerodane diterpenoids produced in leaves (Cantrell *et al.*, 2005; Jones *et al.*, 2007; Dettweiler *et al.*, 2020). However, other *Callicarpa* species, including *C. bodinieri* and *C. macrophylla* (Wang *et al.*, 2017), produce a wide variety of medicinally relevant abietane diterpenoids (Figure 2.8c), indicating that the abietane skeleton is a key intermediate for at least some plants in this genus(Wang *et al.*, 2017; Gao *et al.*, 2020). 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 Figure 2.14).

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(Hamilton *et al.*, 2020). 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(Chen *et al.*, 2011; Hillwig *et al.*, 2011b; Boutanaev *et al.*, 2015a; Jiang *et al.*, 2019), 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 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(Tu *et al.*, 2020). 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(Chaturvedi *et al.*, 2012; González, 2015).

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(Smith *et al.*, 2008; Machumi *et al.*, 2010; González, 2015; Abdissa, Frese and Sewald, 2017; Smirnova *et al.*, 2021). Nearly 500 abietane diterpenoids have been reported to date in Lamiaceae species(Zeng *et al.*, 2020; *Dictionary of Natural Products 30.2*, no date). 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 supportthe 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(Nützmann and Osbourn, 2014). 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; Camacho *et al.*, 2009) 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; Wang *et al.*, 2012). Results were visualized using SynVisio (Bandi and Gutwin, 2020). 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; Revell, 2012). 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 Table 2.5. 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; Sievers *et al.*, 2011) and phylogenetic trees were generated by RAxML (version 8.2.12; Model = protgammaauto; Algorithm = a; Stamatakis, 2014) 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; Letunic and Bork, 2021).

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; Weisenfeld *et al.*, 2017). 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; Stanke *et al.*, 2006) 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; 'Assembly-stats', 2022). 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; Geib *et al.*, 2018). BUSCO (version 5.2.2; Manni *et al.*, 2021) 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; Flynn *et al.*, 2020). Protein-coding genes were removed from each CRL using ProtExcluder (version 1.2; Campbell *et al.*, 2014) and RepBase Viridiplantae repeats from RepBase (version 20150807; Jurka *et al.*, 2005)

were added to create a final CRL. Each assembly was repeat masked with its corresponding CRL using RepeatMasker (version 4.1.2-p1; Chen, 2004) using the parameters -e ncbi -s -nolow -no_is -gff.

Transcriptomic analysis

All transcriptomic datasets used in Figure 2.5 and Supplementary Figure 2.6 were downloaded from the SRA database (Supplementary Table 7). Raw reads were trimmed using fastp (version 0.23.2; Chen *et al.*, 2018), mapped to respective coding sequence files using Salmon 'index' (version 1.8.0; Patro *et al.*, 2017), 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; Wickham, 2016).

PCR and cloning

Synthetic oligonucleotides, GenBank accession numbers, and sequences of all enzymes characterized or discussed in this study are listed in Supplementary Table 2.5. 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 2.6). Constructs were then cloned into the plant expression vector pEAQ-HT (Sainsbury, Thuenemann and Lomonossoff, 2009) 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 (Johnson, Bhat, Bibik, *et al.*, 2019). *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(Englund *et al.*, 2015; Andersen-Ranberg *et al.*, 2016). 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 TCI 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.

REFERENCES

Abdissa, N., Frese, M. and Sewald, N. (2017) 'Antimicrobial abietane-type diterpenoids from *Plectranthus punctatus*', *Molecules*, 22(11), p. 1919. Available at: https://doi.org/10.3390/molecules22111919.

Andersen-Ranberg, J. *et al.* (2016) 'Expanding the landscape of diterpene structural diversity through stereochemically controlled combinatorial biosynthesis', *Angewandte Chemie (International Ed. in English)*, 55(6), pp. 2142–2146. Available at: https://doi.org/10.1002/anie.201510650.

'Assembly-stats' (2022). Pathogen Informatics, Wellcome Sanger Institute. Available at: https://github.com/sanger-pathogens/assembly-stats (Accessed: 31 March 2022).

Bai, Z. *et al.* (2018) '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(1), pp. 243–255. Available at: https://doi.org/10.1007/s00425-018-2884-z.

Bak, S. *et al.* (2011) 'Cytochromes P450', *The Arabidopsis Book / American Society of Plant Biologists*, 9, p. e0144. Available at: https://doi.org/10.1199/tab.0144.

Bandi, V. and Gutwin, C. (2020) '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)*. *Interactive Exploration of Genomic Conservation*, Waterloo, CAN: Canadian Human-Computer Communications Society. Available at: https://synvisio.github.io/#/.

Bathe, U. and Tissier, A. (2019) 'Cytochrome P450 enzymes: A driving force of plant diterpene diversity', *Phytochemistry*, 161, pp. 149–162. Available at: https://doi.org/10.1016/j.phytochem.2018.12.003.

Birtić, S. *et al.* (2015) 'Carnosic acid', *Phytochemistry*, 115, pp. 9–19. Available at: https://doi.org/10.1016/j.phytochem.2014.12.026.

Bohlmann, J., Steele, C.L. and Croteau, R. (1997) 'Monoterpene synthases from grand fir (Abies grandis): cDNA isolation, characterization, and functional expression of myrcene synthase, (-)-(4S)- limonene synthase, and (-)-(1S,5S)-pinene synthase', *Journal of Biological Chemistry*, 272(35), pp. 21784–21792. Available at: https://doi.org/10.1074/jbc.272.35.21784.

Bornowski, N. *et al.* (2020) 'Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae', *DNA Research*, 27(3), p. dsaa016. Available at: https://doi.org/10.1093/dnares/dsaa016.

Bose, S.K. *et al.* (2013) '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 Physiology and Biochemistry*, 66, pp. 150–158. Available at: https://doi.org/10.1016/j.plaphy.2013.02.011.

Boutanaev, A.M. *et al.* (2015a) 'Investigation of terpene diversification across multiple sequenced plant genomes', *Proceedings of the National Academy of Sciences of the United States of America*, 112(1), pp. E81–E88. Available at: https://doi.org/10.1073/pnas.1419547112.

Boutanaev, A.M. *et al.* (2015b) 'Investigation of terpene diversification across multiple sequenced plant genomes', *Proceedings of the National Academy of Sciences*, 112(1), pp. E81–E88. Available at: https://doi.org/10.1073/pnas.1419547112.

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. Available at: https://doi.org/10.1186/1471-2105-10-421.

Campbell, M.S. *et al.* (2014) 'MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations', *Plant Physiology*, 164(2), pp. 513–524. Available at: https://doi.org/10.1104/pp.113.230144.

Cantrell, C.L. *et al.* (2005) 'Isolation and identification of mosquito bite deterrent terpenoids from leaves of American (*Callicarpa americana*) and Japanese (*Callicarpa japonica*) beautyberry', *Journal of Agricultural and Food Chemistry*, 53(15), pp. 5948–5953. Available at: https://doi.org/10.1021/jf0509308.

Chaturvedi, R. *et al.* (2012) 'An abietane diterpenoid is a potent activator of systemic acquired resistance', *The Plant Journal*, 71(1), pp. 161–172. Available at: https://doi.org/10.1111/j.1365-313X.2012.04981.x.

Chen, F. *et al.* (2011) 'The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom', *Plant Journal*, 66(1), pp. 212–229. Available at: https://doi.org/10.1111/j.1365-313X.2011.04520.x.

Chen, N. (2004) 'Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences', *Current Protocols in Bioinformatics*, 5(1), p. 4.10.1-4.10.14. Available at: https://doi.org/10.1002/0471250953.bi0410s05.

Chen, S. *et al.* (2018) 'fastp: an ultra-fast all-in-one FASTQ preprocessor', *Bioinformatics*, 34(17), pp. i884–i890. Available at: https://doi.org/10.1093/bioinformatics/bty560.

Chu, H.Y., Wegel, E. and Osbourn, A. (2011) 'From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants', *The Plant Journal*, 66(1), pp. 66–79. Available at: https://doi.org/10.1111/J.1365-313X.2011.04503.X.

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

Cui, G. *et al.* (2015) 'Functional divergence of diterpene syntheses in the medicinal plant *Salvia miltiorrhiza*', *Plant Physiology*, 169(3), pp. 1607–1618. Available at: https://doi.org/10.1104/pp.15.00695.

Dai, L. *et al.* (2015) 'Functional Characterization of Cucurbitadienol Synthase and Triterpene Glycosyltransferase Involved in Biosynthesis of Mogrosides from Siraitia grosvenorii', *Plant and Cell Physiology*, 56(6), pp. 1172–1182. Available at: https://doi.org/10.1093/PCP/PCV043.

Dettweiler, M. *et al.* (2020) 'A clerodane diterpene from *Callicarpa americana* resensitizes methicillin-resistant staphylococcus aureus to β-lactam antibiotics', *ACS Infectious Diseases*, 6(7), pp. 1667–1673. Available at: https://doi.org/10.1021/acsinfecdis.0c00307.

Dictionary of Natural Products 30.2 (no date). Available at: https://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml (Accessed: 11 March 2022).

Durairaj, J. *et al.* (2019) 'An analysis of characterized plant sesquiterpene synthases', *Phytochemistry*, 158, pp. 157–165. Available at: https://doi.org/10.1016/j.phytochem.2018.10.020.

Englund, E. *et al.* (2015) 'Metabolic engineering of *Synechocystis* sp. PCC 6803 for production of the plant diterpenoid manoyl oxide', *ACS Synthetic Biology*, 4(12), pp. 1270–1278. Available at: https://doi.org/10.1021/acssynbio.5b00070.

Fan, P. *et al.* (2020) 'Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity', *eLife*. Edited by D.J. Kliebenstein, C.S. Hardtke, and R. Peters, 9, p. e56717. Available at: https://doi.org/10.7554/eLife.56717.

Field, B. *et al.* (2011) 'Formation of plant metabolic gene clusters within dynamic chromosomal regions', *Proceedings of the National Academy of Sciences*, 108(38), pp. 16116–16121. Available at: https://doi.org/10.1073/pnas.1109273108.

Field, B. and Osbourn, A.E. (2008) 'Metabolic Diversification—Independent Assembly of Operon-Like Gene Clusters in Different Plants', *Science*, 320(5875), pp. 543–547. Available at: https://doi.org/10.1126/science.1154990.

Flynn, J.M. *et al.* (2020) 'RepeatModeler2 for automated genomic discovery of transposable element families', *Proceedings of the National Academy of Sciences*, 117(17), pp. 9451–9457. Available at: https://doi.org/10.1073/pnas.1921046117.

Frey, M. *et al.* (1997) 'Analysis of a Chemical Plant Defense Mechanism in Grasses', *Science*, 277(5326), pp. 696–699. Available at: https://doi.org/10.1126/SCIENCE.277.5326.696.

Gao, J. *et al.* (2020) 'Anti-NLRP3 inflammasome abietane diterpenoids from *Callicarpa bodinieri* and their structure elucidation', *Chinese Chemical Letters*, 31(2), pp. 427–430. Available at: https://doi.org/10.1016/j.cclet.2019.09.020.

Gao, W. *et al.* (2009) 'A functional genomics approach to tanshinone biosynthesis provides stereochemical insights', *Organic Letters*, 11(22), pp. 5170–5173. Available at: https://doi.org/10.1021/ol902051v.

Geib, S.M. *et al.* (2018) 'Genome Annotation Generator: a simple tool for generating and correcting WGS annotation tables for NCBI submission', *GigaScience*, 7(4), p. giy018. Available at: https://doi.org/10.1093/gigascience/giy018.

Gershenzon, J. and Dudareva, N. (2007) 'The function of terpene natural products in the natural world', *Nature Chemical Biology 2007 3:7*, 3(7), pp. 408–414. Available at: https://doi.org/10.1038/nchembio.2007.5.

Godden, G.T. *et al.* (2019) 'Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints', *Genome Biology and Evolution*, 11(12), pp. 3393–3408. Available at: https://doi.org/10.1093/gbe/evz239.

González, M.A. (2015) 'Aromatic abietane diterpenoids: Their biological activity and synthesis', *Natural Product Reports*, 32(5), pp. 684–704. Available at: https://doi.org/10.1039/c4np00110a.

Guo, J. *et al.* (2013) 'CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts', *Proceedings of the National Academy of Sciences*, 110(29), pp. 12108–12113. Available at: https://doi.org/10.1073/pnas.1218061110.

Guo, J. *et al.* (2016) 'Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones', *New Phytologist*, 210(2), pp. 525–534. Available at: https://doi.org/10.1111/nph.13790.

Hamilton, J.P. *et al.* (2020) 'Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, *Callicarpa americana*', *GigaScience*, 9(9), p. giaa093. Available at: https://doi.org/10.1093/gigascience/giaa093.

He, Y. *et al.* (2018) 'Building an octaploid genome and transcriptome of the medicinal plant *Pogostemon cablin* from Lamiales', *Scientific Data*, 5(1), p. 180274. Available at: https://doi.org/10.1038/sdata.2018.274.

Hillwig, M.L. *et al.* (2011a) 'Domain loss has independently occurred multiple times in plant terpene synthase evolution', *The Plant Journal*, 68(6), pp. 1051–1060. Available at: https://doi.org/10.1111/j.1365-313X.2011.04756.x.

Hillwig, M.L. *et al.* (2011b) 'Domain loss has independently occurred multiple times in plant terpene synthase evolution', *Plant Journal*, 68(6), pp. 1051–1060. Available at: https://doi.org/10.1111/j.1365-313X.2011.04756.x.

Hu, Z. *et al.* (2022) 'Functional Characterization of a 2OGD Involved in Abietane-Type Diterpenoids Biosynthetic Pathway in *Salvia miltiorrhiza*', *Frontiers in Plant Science*, 13. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2022.947674 (Accessed: 18 August 2022).

Hurst, L.D., Pál, C. and Lercher, M.J. (2004) 'The evolutionary dynamics of eukaryotic gene order', *Nature Reviews Genetics*, 5(4), pp. 299–310. Available at: https://doi.org/10.1038/nrg1319.

Ignea, C. *et al.* (2016) 'Carnosic acid biosynthesis elucidated by a synthetic biology platform', *Proceedings of the National Academy of Sciences*, 113(13), pp. 3681–3686. Available at: https://doi.org/10.1073/pnas.1523787113.

Itkin, M. *et al.* (2013) 'Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes', *Science (New York, N.Y.)*, 341(6142), pp. 175–179. Available at: https://doi.org/10.1126/science.1240230.

Jiang, S.-Y. *et al.* (2019) 'A Comprehensive Survey on the Terpene Synthase Gene Family Provides New Insight into Its Evolutionary Patterns', *Genome Biology and Evolution*. Edited by M. Alba, 11(8), pp. 2078–2098. Available at: https://doi.org/10.1093/gbe/evz142.

Johnson, S.R., Bhat, W.W., Bibik, J., *et al.* (2019) 'A database-driven approach identifies additional diterpene synthase activities in the mint family (Lamiaceae)', *Journal of Biological Chemistry*, 294(4), pp. 1349–1362. Available at: https://doi.org/10.1074/jbc.RA118.006025.

Johnson, S.R., Bhat, W.W., Sadre, R., *et al.* (2019) 'Promiscuous terpene synthases from *Prunella vulgaris* highlight the importance of substrate and compartment switching in terpene synthase evolution', *New Phytologist*, 223(1), pp. 323–335. Available at: https://doi.org/10.1111/nph.15778.

Jones, W.P. *et al.* (2007) 'Cytotoxic constituents from the fruiting branches of *Callicarpa americana* collected in southern Florida', *Journal of Natural Products*, 70(3), pp. 372–377. Available at: https://doi.org/10.1021/np060534z.

Jones, W.P. and Kinghorn, A.D. (2008) 'Biologically active natural products of the genus *Callicarpa*', *Current Bioactive Compounds*, 4(1), pp. 15–32. Available at: https://doi.org/10.2174/157340708784533393.

Jurka, J. *et al.* (2005) 'Repbase Update, a database of eukaryotic repetitive elements', *Cytogenetic and Genome Research*, 110(1–4), pp. 462–467. Available at: https://doi.org/10.1159/000084979.

Karunanithi, P.S. and Zerbe, P. (2019) 'Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity', *Frontiers in Plant Science*, 10, p. 1166. Available at: https://doi.org/10.3389/fpls.2019.01166.

Kitaoka, N. *et al.* (2020) 'Interdependent evolution of biosynthetic gene clusters for momilactone production in rice', *The Plant Cell* [Preprint]. Available at: https://doi.org/10.1093/plcell/koaa023.

Kliebenstein, D.J. (2008) 'A role for gene duplication and natural variation of gene expression in the evolution of metabolism', *PloS One*, 3(3), p. e1838. Available at: https://doi.org/10.1371/journal.pone.0001838.

Lange, B.M. (2015) 'The Evolution of Plant Secretory Structures and Emergence of Terpenoid Chemical Diversity', *Annual Review of Plant Biology*, 66(1). Available at: https://doi.org/10.1146/annurev-arplant-043014-114639.

Letunic, I. and Bork, P. (2021) 'Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation', *Nucleic Acids Research*, 49(W1), pp. W293–W296. Available at: https://doi.org/10.1093/nar/gkab301.

Li, P. *et al.* (2017) 'Molecular phylogenetics and biogeography of the mint tribe Elsholtzieae (Nepetoideae, Lamiaceae), with an emphasis on its diversification in East Asia', *Scientific Reports*, 7(1), p. 2057. Available at: https://doi.org/10.1038/s41598-017-02157-6.

Li, Y. *et al.* (2021) 'Subtelomeric assembly of a multi-gene pathway for antimicrobial defense compounds in cereals', *Nature Communications*, 12(1), p. 2563. Available at: https://doi.org/10.1038/s41467-021-22920-8.

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

Lichman, B.R. *et al.* (2020) 'The evolutionary origins of the cat attractant nepetalactone in catnip', *Science Advances*, 6. Available at: https://doi.org/10.1126/sciadv.aba0721.

Liu, B., Tan, Y.-H., *et al.* (2020) 'Phylogenetic relationships of *Cyrtandromoea* and *Wightia* revisited: A new tribe in Phrymaceae and a new family in Lamiales', *Journal of Systematics and Evolution*, 58(1), pp. 1–17. Available at: https://doi.org/10.1111/jse.12513.

Liu, Z., Suarez Duran, H.G., *et al.* (2020) 'Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae', *New Phytologist*, 227(4), pp. 1109–1123. Available at: https://doi.org/10.1111/nph.16338.

Liu, Z., Cheema, J., *et al.* (2020) 'Formation and diversification of a paradigm biosynthetic gene cluster in plants', *Nature Communications*, 11(1), p. 5354. Available at: https://doi.org/10.1038/s41467-020-19153-6.

Ma, Y. *et al.* (2012) 'Genome-wide identification and characterization of novel genes involved in terpenoid biosynthesis in *Salvia miltiorrhiza*', *Journal of Experimental Botany*, 63(7), pp. 2809–2823. Available at: https://doi.org/10.1093/jxb/err466.

Ma, Y. *et al.* (2021) 'Expansion within the CYP71D subfamily drives the heterocyclization of tanshinones synthesis in *Salvia miltiorrhiza*', *Nature Communications*, 12(1), p. 685. Available at: https://doi.org/10.1038/s41467-021-20959-1.

Machumi, F. *et al.* (2010) 'Antimicrobial and antiparasitic abietane diterpenoids from the roots of *Clerodendrum eriophyllum*', *Natural Product Communications*, 5(6), p. 1934578X1000500605. Available at: https://doi.org/10.1177/1934578X1000500605.

Manni, M. *et al.* (2021) 'BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes', *Molecular Biology and Evolution*, 38(10), pp. 4647–4654. Available at: https://doi.org/10.1093/molbev/msab199.

Matsuba, Y. *et al.* (2013) 'Evolution of a complex locus for terpene biosynthesis in *Solanum*', *The Plant Cell*, 25(6), pp. 2022–2036. Available at: https://doi.org/10.1105/tpc.113.111013.

Medema, M.H. *et al.* (2015) 'Minimum information about a biosynthetic gene cluster', *Nature Chemical Biology*, 11(9), pp. 625–631. Available at: https://doi.org/10.1038/nchembio.1890.

Mugford, S.T. *et al.* (2013) 'Modularity of plant metabolic gene clusters: A trio of linked genes that are collectively required for acylation of triterpenes in oat', *The Plant Cell*, 25(3), pp. 1078–1092. Available at: https://doi.org/10.1105/tpc.113.110551.

Nützmann, H.-W. *et al.* (2020) 'Active and repressed biosynthetic gene clusters have spatially distinct chromosome states', *Proceedings of the National Academy of Sciences*, 117(24), pp. 13800–13809. Available at: https://doi.org/10.1073/pnas.1920474117.

Nützmann, H.-W., Huang, A. and Osbourn, A. (2016) 'Plant metabolic clusters – from genetics to genomics', *New Phytologist*, 211(3), pp. 771–789. Available at: https://doi.org/10.1111/nph.13981.

Nützmann, H.-W. and Osbourn, A. (2014) 'Gene clustering in plant specialized metabolism', *Current Opinion in Biotechnology*, 26, pp. 91–99. Available at: https://doi.org/10.1016/j.copbio.2013.10.009.

Nützmann, H.-W., Scazzocchio, C. and Osbourn, A. (2018) 'Metabolic gene clusters in eukaryotes', *Annual Review of Genetics*, 52(1), pp. 159–183. Available at: https://doi.org/10.1146/annurev-genet-120417-031237.

Okada, A. *et al.* (2009) 'OsTGAP1, a bZIP transcription factor, coordinately regulates the inductive production of diterpenoid phytoalexins in rice', *The Journal of Biological Chemistry*, 284(39), pp. 26510–26518. Available at: https://doi.org/10.1074/jbc.M109.036871.

Pateraki, I. *et al.* (2017) 'Total biosynthesis of the cyclic AMP booster forskolin from *Coleus forskohlii*', *eLife*. Edited by J. Bohlmann, 6, p. e23001. Available at: https://doi.org/10.7554/eLife.23001.

Patro, R. *et al.* (2017) 'Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference', *Nature methods*, 14(4), pp. 417–419. Available at: https://doi.org/10.1038/nmeth.4197.

Polturak, G. and Osbourn, A. (2021) 'The emerging role of biosynthetic gene clusters in plant defense and plant interactions', *PLOS Pathogens*. Edited by C. Zipfel, 17(7), p. e1009698. Available at: https://doi.org/10.1371/JOURNAL.PPAT.1009698.

Postnikova, O.A. *et al.* (2011) 'Clustering of pathogen-response genes in the genome of *Arabidopsis thaliana*', *Journal of Integrative Plant Biology*, 53(10), pp. 824–834. Available at: https://doi.org/10.1111/j.1744-7909.2011.01071.x.

Qi, X. *et al.* (2004) 'A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants', *Proceedings of the National Academy of Sciences*, 101(21), pp. 8233–8238. Available at: https://doi.org/10.1073/pnas.0401301101.

Revell, L.J. (2012) 'Phytools: An R package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution*, 3(2), pp. 217–223. Available at: https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Rokas, A., Wisecaver, J.H. and Lind, A.L. (2018) 'The birth, evolution and death of metabolic gene clusters in fungi', *Nature Reviews Microbiology*, 16(12), pp. 731–744. Available at: https://doi.org/10.1038/s41579-018-0075-3.

Sainsbury, F., Thuenemann, E.C. and Lomonossoff, G.P. (2009) 'pEAQ: Versatile expression vectors for easy and quick transient expression of heterologous proteins in plants', *Plant Biotechnology Journal*, 7(7), pp. 682–693. Available at: https://doi.org/10.1111/j.1467-7652.2009.00434.x.

Sakamoto, T. *et al.* (2004) 'An Overview of Gibberellin Metabolism Enzyme Genes and Their Related Mutants in Rice', *Plant Physiology*, 134(4), pp. 1642–1653. Available at: https://doi.org/10.1104/PP.103.033696.

Scheler, U. *et al.* (2016) 'Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast', *Nature Communications*, 7(1), p. 12942. Available at: https://doi.org/10.1038/ncomms12942.

Schenck, C.A. and Last, R.L. (2020) 'Location, location! Cellular relocalization primes specialized metabolic diversification', *The FEBS Journal*, 287(7), pp. 1359–1368. Available at: https://doi.org/10.1111/febs.15097.

Schmelz, E.A. *et al.* (2014) 'Biosynthesis, elicitation and roles of monocot terpenoid phytoalexins', *The Plant Journal*, 79(4), pp. 659–678. Available at: https://doi.org/10.1111/TPJ.12436.

Shang, Y. *et al.* (2014) 'Biosynthesis, regulation, and domestication of bitterness in cucumber', *Science*, 346(6213), pp. 1084–1088. Available at: https://doi.org/10.1126/SCIENCE.1259215.

Sherden, N.H. *et al.* (2018) 'Identification of iridoid synthases from *Nepeta species*: Iridoid cyclization does not determine nepetalactone stereochemistry', *Phytochemistry*, 145, pp. 48–56. Available at: https://doi.org/10.1016/j.phytochem.2017.10.004.

Sievers, F. *et al.* (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology*, 7(1), p. 539. Available at: https://doi.org/10.1038/msb.2011.75.

Slot, J.C. and Hibbett, D.S. (2007) 'Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: A phylogenetic study', *PLOS ONE*, 2(10), p. e1097. Available at: https://doi.org/10.1371/journal.pone.0001097.

Slot, J.C. and Rokas, A. (2010) 'Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi', *Proceedings of the National Academy of Sciences*, 107(22), pp. 10136–10141. Available at: https://doi.org/10.1073/pnas.0914418107.

Smirnova, I.E. *et al.* (2021) 'Synthetic modifications of abietane diterpene acids to potent antimicrobial agents', *Natural Product Research*, 0(0), pp. 1–9. Available at: https://doi.org/10.1080/14786419.2021.1969566.

Smith, E.C.J. *et al.* (2008) '2β-Acetoxyferruginol—A new antibacterial abietane diterpene from the bark of *Prumnopitys andina*', *Phytochemistry Letters*, 1(1), pp. 49–53. Available at: https://doi.org/10.1016/j.phytol.2007.12.006.

Song, J.-J. *et al.* (2022) 'A 2-oxoglutarate-dependent dioxygenase converts dihydrofuran to furan in *Salvia* diterpenoids', *Plant Physiology*, 188(3), pp. 1496–1506. Available at: https://doi.org/10.1093/plphys/kiab567.

Song, Z. *et al.* (2020) 'A high-quality reference genome sequence of *Salvia miltiorrhiza* provides insights into tanshinone synthesis in its red rhizomes', *The Plant Genome*, 13(3), p. e20041. Available at: https://doi.org/10.1002/tpg2.20041.

Stamatakis, A. (2014) 'RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. Available at: https://doi.org/10.1093/bioinformatics/btu033.

Stanke, M. *et al.* (2006) 'AUGUSTUS: Ab initio prediction of alternative transcripts', *Nucleic Acids Research*, 34(suppl_2), pp. W435–W439. Available at: https://doi.org/10.1093/nar/gkl200.

Swaminathan, S. *et al.* (2009) 'CYP76M7 is an *ent*-cassadiene C11α-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice', *The Plant Cell*, 21(10), pp. 3315–3325. Available at: https://doi.org/10.1105/tpc.108.063677.

Takos, A.M. and Rook, F. (2012) 'Why biosynthetic genes for chemical defense compounds cluster', *Trends in Plant Science*, 17(7), pp. 383–388. Available at: https://doi.org/10.1016/j.tplants.2012.04.004.

Tholl, D. (2015) 'Biosynthesis and Biological Functions of Terpenoids in Plants', *Advances in Biochemical Engineering/Biotechnology*, 148, pp. 63–106. Available at: https://doi.org/10.1007/10_2014_295.

Tu, L. *et al.* (2020) 'Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis', *Nature Communications*, 11(1), p. 971. Available at: https://doi.org/10.1038/s41467-020-14776-1.

Wang, Jiadian *et al.* (2021) 'A cytochrome P450 CYP81AM1 from *Tripterygium wilfordii* catalyses the C-15 hydroxylation of dehydroabietic acid', *Planta*, 254(5), p. 95. Available at: https://doi.org/10.1007/s00425-021-03743-9.

Wang, Y. *et al.* (2012) 'MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity', *Nucleic Acids Research*, 40(7), pp. e49–e49. Available at: https://doi.org/10.1093/nar/gkr1293.

Wang, Z. and Peters, R.J. (2022) 'Tanshinones: Leading the way into Lamiaceae labdane-related diterpenoid biosynthesis', *Current Opinion in Plant Biology*, 66, p. 102189. Available at: https://doi.org/10.1016/j.pbi.2022.102189.

Wang, Z.-H. *et al.* (2017) 'Three New Abietane-Type Diterpenoids from Callicarpa macrophylla Vahl.', *Molecules*, 22(5), p. 842. Available at: https://doi.org/10.3390/molecules22050842.

Weisenfeld, N.I. *et al.* (2017) 'Direct determination of diploid genome sequences', *Genome Research*, 27(5), pp. 757–767. Available at: https://doi.org/10.1101/gr.214874.116.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Pringer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Wilderman, P.R. *et al.* (2004) 'Identification of Syn-Pimara-7,15-Diene Synthase Reveals Functional Clustering of Terpene Synthases Involved', *Plant Physiology*, 135(4), pp. 2098–2105. Available at: https://doi.org/10.1104/pp.104.045971.

Winzer, T. *et al.* (2012) 'A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine', *Science*, 336(6089), pp. 1704–1708. Available at: https://doi.org/10.1126/science.1220757.

Wisecaver, J.H. *et al.* (2017) 'A global coexpression network approach for connecting genes to specialized metabolic pathways in plants', *The Plant Cell*, 29(5), pp. 944–959. Available at: https://doi.org/10.1105/tpc.17.00009.

Xu, H. *et al.* (2016) 'Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*', *Molecular Plant*, 9(6), pp. 949–952. Available at: https://doi.org/10.1016/j.molp.2016.03.010.

Xu, Z. *et al.* (2020) 'Comparative genome analysis of *Scutellaria baicalensis* and *Scutellaria barbata* reveals the evolution of active flavonoid biosynthesis', *Genomics, Proteomics & Bioinformatics*, 18(3), pp. 230–240. Available at: https://doi.org/10.1016/j.gpb.2020.06.002.

Yang, X. *et al.* (2021) 'Three chromosome-scale Papaver genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway', *Nature Communications 2021 12:1*, 12(1), pp. 1–14. Available at: https://doi.org/10.1038/s41467-021-26330-8.

Yao, G. *et al.* (2016) 'Phylogenetic relationships, character evolution and biogeographic diversification of *Pogostemon* s.l. (Lamiaceae)', *Molecular Phylogenetics and Evolution*, 98, pp. 184–200. Available at: https://doi.org/10.1016/j.ympev.2016.01.020.

Yu, N. *et al.* (2016) 'Delineation of metabolic gene clusters in plant genomes by chromatin signatures', *Nucleic Acids Research*, 44(5), pp. 2255–2265. Available at: https://doi.org/10.1093/nar/gkw100.

Zeng, T. *et al.* (2020) 'TeroKit: A Database-Driven Web Server for Terpenome Research', *Journal of Chemical Information and Modeling*, 60(4), pp. 2082–2090. Available at: https://doi.org/10.1021/acs.jcim.0c00141.

Zhang, J. and Peters, R.J. (2020) 'Why are momilactones always associated with biosynthetic gene clusters in plants?', *Proceedings of the National Academy of Sciences*, 117(25), pp. 13867–13869. Available at: https://doi.org/10.1073/pnas.2007934117.

Zhao, D. *et al.* (2019) *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*. Available at: doi.org/10.1093/gigascience/giz005 (Accessed: 3 December 2021).

Zi, J. and Peters, R.J. (2013) 'Characterization of CYP76AH4 clarifies phenolic diterpenoid biosynthesis in the Lamiaceae', *Organic & Biomolecular Chemistry*, 11(44), p. 7650. Available at: https://doi.org/10.1039/c3ob41885e.

CHAPTER 3: DECIPHERING THE TETRAPLOID GENOME AND DITERPENOID METABOLISM OF *TEUCRIUM CHAMAEDRYS*

Abigail E Bryson[1], Kevin L Childs[2], Nicholas Schlecht[1], Davis Mathieu[1], John P Hamilton[4,5], Haoyang Xin[2], Jiming Jiang[2,3], C Robin Buell[4,5,6,7], Björn Hamberger[1].

[1]Department of Biochemistry; Michigan State University; East Lansing, MI, 48823, USA. [2]Department of Plant Biology; Michigan State University; East Lansing, MI, 48823, USA. [3]Department of Horticulture; Michigan State University; East Lansing, MI, 48823, USA. [4]Center for Applied Genetic Technology; University of Georgia; Athens, GA, 30602, USA. [5]Department of Crop & Soil Sciences; University of Georgia; Athens, GA, 30602, USA. [6]Institute of Plant Breeding, Genetics, & Genomics; University of Georgia; Athens, GA, 30602, USA. [7]The Plant Center; University of Georgia; Athens, GA, 30602, USA.

*Corresponding author: hamberge@msu.edu

Author contributions:

AEB, NS, and BH designed the study. AEB, KLC, and JH assembled the genome. AEB and KLC annotated the genome. AEB and DM performed the genomic analysis. AEB performed phylogenetic experiments. AEB performed biochemical experiments, and NS and DM analyzed the results. HX performed the chromosome squash. AEB and NS wrote the manuscript. JJ, CRB, and BH supervised. All authors edited the manuscript.

Abstract

Teucrium chamaedrys, also called wall germander, is a small woody shrub native to the Mediterranean region. Its name is derived from the Greek words meaning 'ground oak', since its tiny leaves resemble those of an oak tree. *Teucrium* species are proliferative producers of diterpenes, which afford them valuable properties widely co-opted in traditional and western medicines. Sequence and assembly of the 3 Gbp tetraploid *T. chamaedrys* revealed 74 diterpene synthase genes, with the vast representation of these diterpene synthases clustered along four genomic loci. Comparative genomics revealed that this cluster is mirrored in the closely related species, *Teucrium marum*. Along with the presence of several cytochrome p450 sequences, this region is the largest biosynthetic gene cluster identified. *Teucrium* is well known for making clerodane-type diterpenoids which are produced from a kolavanyl diphosphate precursor. To elucidate the complex biosynthetic pathways of these medicinal compounds, we identified and functionally characterized several kolavanyl diphosphate synthases from *T. chamaedrys*. Its remarkable chemistry and tetraploidy make *T. chamaedrys* an interesting and unique model for studying genomic evolution and adaptation in plants.

Introduction

The Lamiaceae (mint) family includes culturally and economically important plants such as peppermint, lavender, sage, rosemary, and teak. It is the third largest family of flowering plants with an estimated 7,000 species. However, representative genomes are limited, with only around 0.66% (46) being published to date. Sampling understudied clades in the Lamiaceae can reveal the underlying bases of specialized metabolism, as it is home to nearly 7,500 unique plant natural products with utility and application relevant to human interests (*Dictionary of Natural Products 30.2*). The subfamily Ajugoideae (syn. Teucrioideae) is one such understudied clade, with around 770 species and only three published genomes (Ritz *et al.*, 2023; Smit *et al.*, 2024). Within the Ajugoideae, the polyphyletic *Teucrium* is one of the largest genera with around 300 species. The genus *Teucrium* has been used for millennia, with notable historical applications such as treating asthma in ancient Greece (Menichini *et al.*, 2009). *Teucrium* species are also well known for their insect antifeedant activity, allelopathic growth inhibition of cosmopolitan weeds, anti-microbial, anti-viral, anti-inflammatory, hepatotoxic effects, and potential as a selective anti-cancer agent for colorectal cancer (Klein Gebbinck, Jansen and de Groot, 2002; Milutinović *et al.*, 2019; Candela *et al.*, 2020). *Teucrium chamaedrys*, or wall germander, is a woody shrub native to the Mediterranean area and is one of the most cited *Teucrium* species in folk medicine (Jarić, Mitrović and Pavlović, 2020). It is specifically recognized in ethnobotanical studies for treating a wide variety of health issues, including digestive problems, hypertension, and malaria (Pieroni, Quave and Santoro, 2004; di Tizio *et al.*, 2012; Arı *et al.*, 2015; Jarić, Mitrović and Pavlović, 2020).

Medicinal properties of the plant are typically consequences of its specialized metabolite profiles. Ajugoideae, and *Teucrium* specifically, is especially well known for its abundance of diterpenoids (*Dictionary of Natural Products 30.2*). Generally, diterpenoids are formed by the sequential activity of two diterpene synthases (diTPSs): first a class II diTPS (TPS-c) catalyzes a proton mediated cyclization of a 20-carbon isoprenoid diphosphate, usually geranylgeranyl diphosphate (GGDP); then a class I (often a

TPS-e) diTPS cleaves the diphosphate, further differentiating the diterpene structure. *Teucrium* is especially rich in clerodane-type diterpenoids (Li, Morris-Natschke and Lee, 2016; Schlecht *et al.*, 2024). Clerodane synthases usually form a class II product, either (-)-kolavenyl diphosphate ((-)-KDP), iso-KDP, or rarely, cis-trans-clerodienyl diphosphate. Characterized iso-KDP synthases so far have been limited to species in the Lamiaceae: A*juga reptans, Scutellaria barbata* and *Scutellaria baicalensis* (Johnson *et al.*, 2019; Qiu *et al.*, 2023). (-)-KDP synthases have been characterized in *Salvia divinorum, Salvia splendens, Vitex agnus-castus, Callicarpa americana, Scutellaria barbata, Scutellaria baicalensis,* and *Tripterygium wilfordii* (Andersen-Ranberg *et al.*, 2016; Hansen *et al.*, 2017; Pelot *et al.*, 2017; Heskes *et al.*, 2018; Hamilton *et al.*, 2020). Iso-KDP differs from (-)-KDP by the final deprotonation, with the double bond placed along the 4,18 bond rather than the 3,4 bond (Figure 3.1). The third and most uncommon structure has exclusively been found in the monocot species *Panicum virgatum*, and is a cis-trans-clerodienyl diphosphate, which while having the same final quenching as (-)-KDP, is a different stereoisomer (Pelot *et al.*, 2018). A variety of clerodane-derived products have been characterized specifically from *T. chamaedrys*, including various neo-clerodanes, chamaedryosides A–C, Teucrin, and more (Figure 3.1; Bedir, Manyam and Khan, 2003; Fiorentino *et al.*, 2009; Sadeghi *et al.*, 2022; *Dictionary of Natural Products 30.2*).

Diversity of plant natural products can be driven in part by duplication of genes via several mechanisms. When duplications occur, it can significantly increase the number of novel genes by dispersing selective pressure, which in turn can allow an explosion of metabolic diversity (Ren *et al.*, 2018). A duplication can occur via tandem or segmental duplication, both of which copy a region locally and can be the result of unequal DNA crossover events (Achaz *et al.*, 2000). Repeats can also be introduced via retrotransposition, which is evident by their lack of introns and presence of nearby inverted repeats (Hughes *et al.*, 2003; Field *et al.*, 2011). But perhaps the most radical duplication method is whole genome duplication (WGD). It is estimated that around 35% of all extant angiosperm species are

polyploids, having a history of WGDs (Wood *et al.*, 2009; Landis *et al.*, 2018; Godden *et al.*, 2019). Similarly, nearly one third (65%) of annotated plant genes are duplicated, with most derived from WGD events (Panchy, Lehti-Shiu and Shiu, 2016).

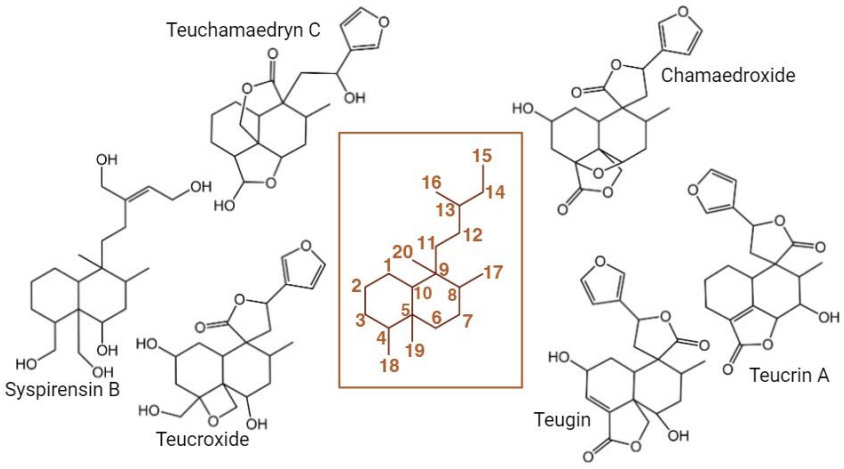


Figure 3.1 Clerodane skeleton and select clerodanes from *T. chamaedrys*. *Teucrium*, and specifically *T. chamaedrys*, is rich in clerodane-type diterpenoids. Middle box shows numbered carbons on a typical clerodane skeleton.

To better understand how polyploidy affects diTPSs and clerodane diversity in *Teucrium,* we determined the large (3 Gbp) tetraploid genome of *T. chamaedrys.* A recent whole genome duplication has afforded this species a rich diversity of diTPSs, including a massive clustering of most of the reported diTPSs. This clustering is also present in the closely related species, *Teucrium marum* (Smit *et al.*, 2024). The physical clustering of most of these diTPSs creates the largest biosynthetic gene cluster to date. Since *Teucrium* species are well known for their clerodane-derived products, we functionally characterized all four putative clerodane synthases present in *T. chamaedrys*, as well as a single representative from *Teucrium canadense*. Using comparative genomic and biochemical methods, we present the genetic underpinning of diterpenoid diversity within this species.

Results and Discussion

T. chamaedrys *genome reveals evidence for tetraploidy*

To create a high-quality genome assembly for *T. chamaedrys*, we generated 265 Gbp of Oxford Nanopore Technology (ONT) long reads and 95 Gbp of Illumina short reads. GenomeScope estimated the *T. chamaedrys* genome size at approximately 1.7 Gbp with low heterozygosity of 0.14% (Supplemental Figure 3.1). Assembling, polishing, and removing contigs less than 10 Kbp in length produced 3,162 contigs (Supplemental Table 3.1) with a final assembly size of 2.9 Gbp (Supplemental Figure 3.2). Benchmarking Universal Single-Copy Orthologs (BUSCO; Manni *et al.*, 2021) analysis with 2,326 total BUSCO genes (eudicots_odb10) revealed 2,274 (97.8%) complete orthologs, of which 60 (2.6%) were single-copy, 2,214 (95.2%) were duplicated, 8 (0.3%) were fragmented, and 44 (1.9%) were missing. Annotation of protein-coding genes revealed 128,111 high confidence genes. BUSCO analysis of the annotation revealed a similar set of statistics, with 2,210 (95.0%) complete orthologs, 88 (3.8%) single-copy, 2,122 (91.2%) duplicated, 21 (0.9%) fragmented, and 95 (4.1%) missing. Overall, this demonstrates a high-quality assembly and annotation of the *T. chamaedrys* genome.

The presence of a highly duplicated BUSCO score suggests a recent WGD event, which is additionally corroborated by Smudgeplot k-mer analysis of genome duplication (Figure 3.2C; Ranallo-Benavidez, Jaron and Schatz, 2020). Polyploids frequently have highly divergent subgenomes, which can lead to underestimation of shared k-mers (Supplemental Figure 3.3; Ranallo-Benavidez, Jaron and Schatz, 2020). The presence of a smudge at the ‘AAAB’ position, coupled with the trace presence of ‘AABB’ resulted in a relatively strong signal at 4n coverage, indicating *T. chamaedrys* is a tetraploid. This is consistent with OrthoFinder (Emms and Kelly, 2019) analysis comparing *T. chamaedrys* to *Arabidopsis thaliana* which revealed a majority of 4:1 ratio of orthologs, with lesser evidence for 2:1 and 3:1 ratios (Figure 3.2D). Furthermore, a chromosome count in dividing root tip cells and comparison to the closely related diploid

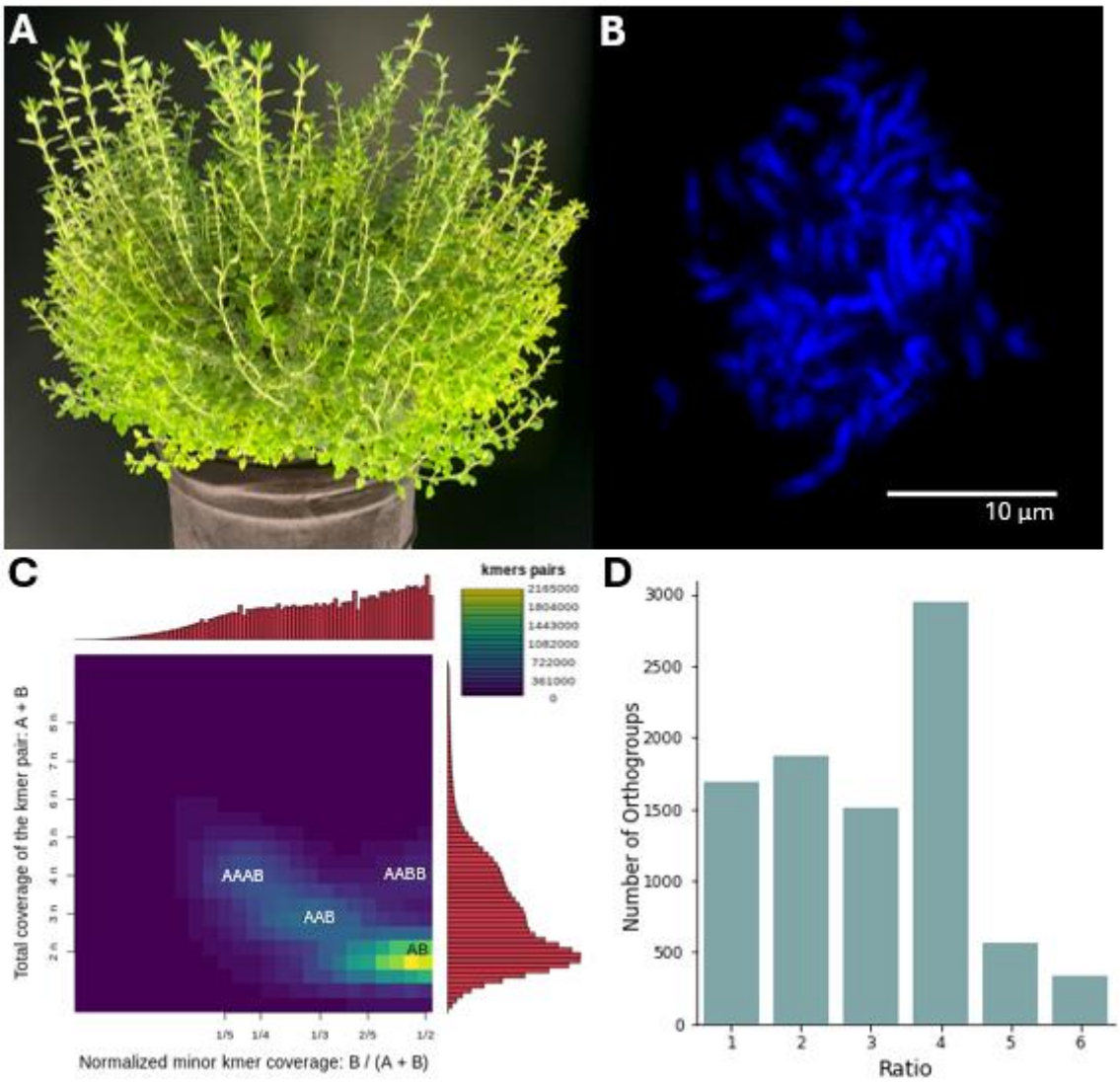


Figure 3.2 The tetraploid genome of *T. chamaedrys*. A) Image of mature *T. chamaedrys* shrub. **B)** Representative chromosome squash; a root tip in meiosis. **C)** Smudgeplot analysis showing evidence for genome duplication, with evidence at 4n 'AAAB' and 'AABB'. **D)** Orthogroup proportions between *T. chamaedrys* and *A. thaliana*. Approximately 3,000 orthogroups have four times as many orthologs of *T. chamaedrys* compared to *A. thaliana.*

relative *T. marum* (2n = 34; Smit *et al.*, 2024), reveals that most metaphase cells contained 2n = 60 for *T. chamaedrys* (Figure 3.2B), which is also in line with a recent WGD event leading to tetraploidy. Previous chromosome counting efforts have shown this species to be diverse in the base chromosome number (2n = 32-96; Ranjbar, Mahmoudi and Nazari, 2018); the genome k-mer analysis, orthology, and cytogenetic evidence supports tetraploidy.

The evolutionary split between *T. marum* and *T. chamaedrys* is estimated at approximately 4 million years ago (Salmaki *et al.*, 2016), meaning the WGD event in the *T. chamaedrys* lineage is relatively recent. Meiotic abnormalities, cell architecture changes, and genetic instability are a few of the detrimental side effects to WGD (Osborn *et al.*, 2003; De Storme and Mason, 2014; Wang *et al.*, 2021; Blasio *et al.*, 2022), and many polyploids re-diploidized to mitigate them (Li *et al.*, 2021; Wang *et al.*, 2021). Therefore, this recent tetraploid genome may be a fleeting snapshot representative of the many polyploidization events that are ubiquitously present in plant lineages. Therefore, the data we provide may fuel further studies of the effects of polyploidization.

Phylogenetic evidence shows clustering and expansion of diterpene synthases in Teucrium

We estimated phylogenetic relationships of the 90 putative diTPS sequences in three *Teucrium* species (*T. chamaedrys*, *T. marum*, and *T. canadense*) alongside a set of functionally characterized diTPSs from species in the Lamiaceae family and *A. thaliana* (Supplemental Table 3.2). One locus in *T. marum* (*Teum.10G004340.2—Teum.10G004860.4*) accounts for 11 of the 15 predicted diTPSs. Mirroring that, four loci within *T. chamaedrys* (*Tcha40759—Tcha40827*, *Tcha129821—Tcha129881, Tcha25933—Tcha25972*, and *Tcha102085—Tcha102138*) account for 53 of the 74 predicted diTPSs, and they cluster across the phylogeny with the orthologs in *T. marum* (Figure 3.3A).

There is clear synteny between the four genomic regions harboring these diTPSs in *T. chamaedrys* as compared to the region in *T. marum* (Figure 3B), where *Tcha40759—Tcha40827* are on contig Tc20548, *Tcha129821—Tcha129881* are on contig Tc17783, *Tcha25933—Tcha25972* are on contig Tc11061, and *Tcha102085—Tcha102138* are on contig Tc19693. This syntenic region contains predicted enzymes that include both class II and class I mechanisms, which is evidence for a large biosynthetic gene cluster (BGC). Interestingly, these clustered genes appear to be a part of a Lamiaceae-wide miltiradiene-producing BGC (Bryson *et al.*, 2023), as the clustered *Teucrium* genes are in the same phylogenetic clade.

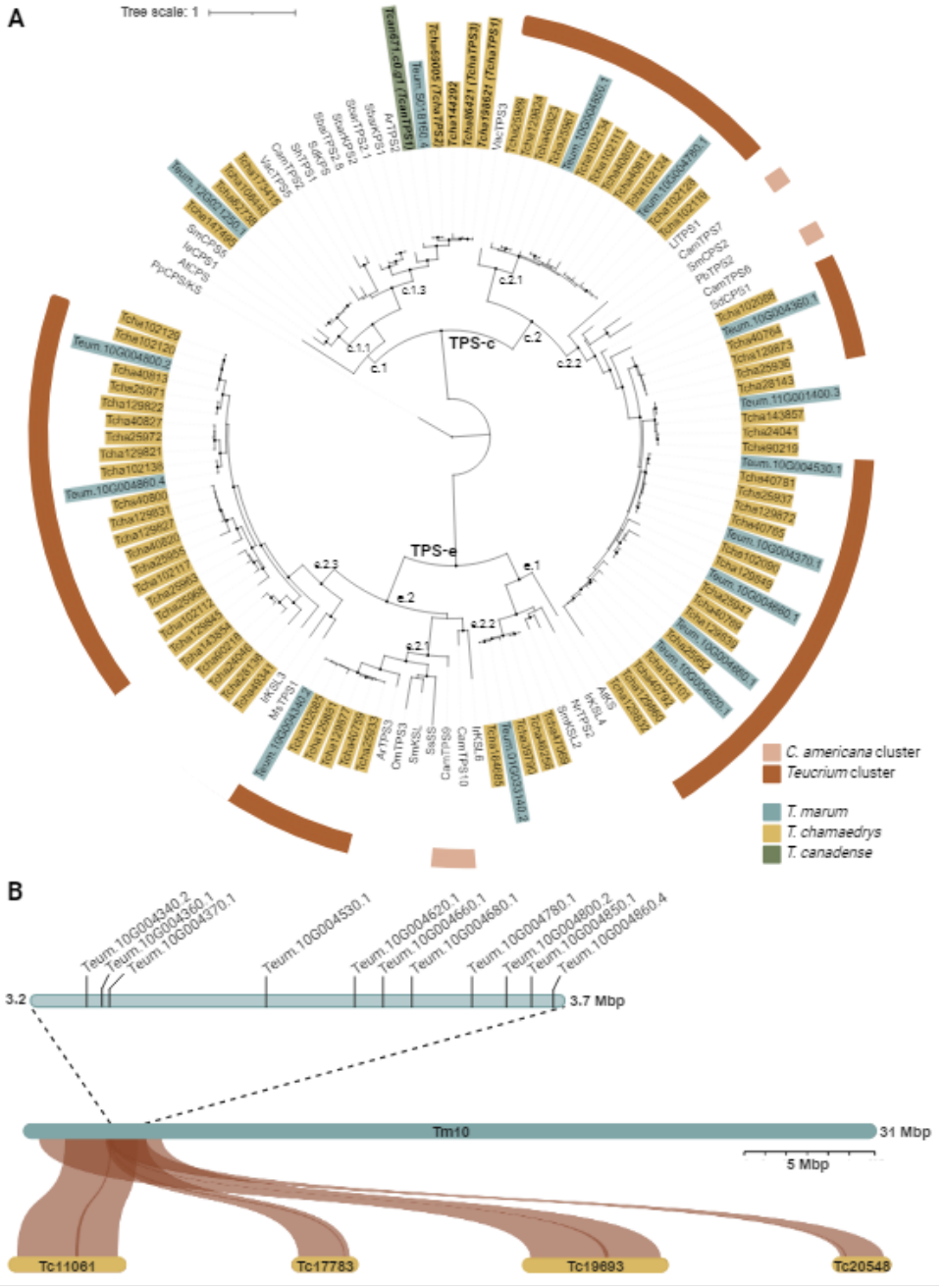


Figure 3. 3 Phylogenetic analysis of the diterpene gene content in three *Teucrium* species. A) This tree is rooted by the class II/class I bifunctional *ent*-kaurene synthase from *Physcometrium patens*. Genes from *T. chamaedrys* are in gold, *T. marum* in blue, and *T. canadense* in green. Those without highlights are previously characterized diTPSs from other Lamiaceae species and *A. thaliana*. Those bolded were functionally characterized in this work. Dots at the base of nodes denote 80% bootstrap confidence. Clades are labeled according to Johnson *et al.*, 2019. Figure was made using iTOL and BioRender.com. **B)** Syntenic analysis between closely related *T. marum* (blue) and *T. chamaedrys* (gold) show a 1:4 ratio in a

Figure 3.3 (cont'd)
specific region containing the vast majority of diTPSs in these species. Figure was made using SynVisio and BioRender.com.

Additionally, this BGC also contains around 15 predicted CYPs from the CYP71 clan which are often involved in diterpenoid metabolism and are also present in the Lamiaceae-wide BGC. This *T. marum* cluster appears to form what is now the largest diTPS BGC to date, spanning around 500 Kbp (Bryson *et al.*, 2023).

Introducing genetic redundancy can lead to diversity in specialized metabolic pathways since selective pressure is dispersed (Ohno, 1970; Birchler and Yang, 2022). It is evident by the high number of diTPS sequences presented here that there has been an explosion of specialized metabolism in *T. chamaedrys* and *T. marum* (Figure 3.1A). *Teucrium* is in the top five Lamiaceae producers of unique diterpene skeletons and compounds (Johnson *et al.*, 2019), and the sheer number of predicted diTPSs in *T. chamaedrys* is in alignment with this. Phylogenetic blooms in a species can be attributed to tandem duplication and neofunctionalization which appears to be the case here, with the large majority of diTPS diversity appearing to predate the speciation of *T. chamaedrys* and *T. marum*. The number of diTPS sequences in *Teucrium* illustrates the vast diversity of diterpenoids harbored in these species, especially *T. chamaedrys*. The sequences from *T. canadense* were derived from transcriptomic rather than genomic data, and therefore may not show a complete picture of diTPS diversity in this species. A higher proportion of diTPSs present in *T. chamaedrys* further suggests WGD, and syntenic analysis corroborates this.

Most plant species have two diTPSs which biosynthesize the initial pathway toward gibberellic acid; one in the phylogenetic clade TPS-e.1 and one in TPS-c.1, corresponding with the class I and class II enzymes, respectively. The same 1:4 ratio present in the BGC is evident in the genes involved in gibberellic acid synthesis as well (Figure 3.3A; TPS-c.1.1 and TPS-e.1). Where we usually see one TPS-e.1, there are four present in *T. chamaedrys*; and the same is true for TPS-c.1 (Figure 3A). Understanding that the *T. marum*

genome assembly is haploid (Smit *et al.*, 2024), this 1:4 ratio would be consistent with a WGD event present specifically in the lineage of *T. chamaedrys*.

Biochemical analysis reveals basis of clerodane metabolism in T. chamaedrys

In order to better understand clerodane representation in *Teucrium*, we investigated enzyme activity of four predicted clerodane synthase homologs in *T. chamaedrys* and one in *T. canadense*: TchaTPS1, TchaTPS2, TchaTPS3, and TcanTPS1. The fourth predicted clerodane synthase gene from *T. chamaedrys* was found to be inactive, which is also suggested by its low expression in the plant (Tcha144292; Supplemental Figure 3.4). We also combined each enzyme with sclareol synthase (SsSS), a promiscuous class I diTPS that in this case produces exclusively iso-kolavelool from iso-KDP (Caniard *et al.*, 2012), enabling us to determine the product at 11.5 minutes (**1**) to be iso-kolavelool (*neo*-cleroda-4(18),14-dien-13-ol), and the other major product at 13.5 minutes (**2**) was identified as iso-kolavenol, by comparison with the reference class II enzyme ArTPS2 (Figure 3.4; Supplemental Figure 3.5; Johnson *et al.*, 2019). The mixture of products in runs without SsSS occurs as a result of a dephosphorylation, catalyzed by non-specific endogenous enzymes in *Nicotiana benthamiana* (Supplemental Figure 3.6). SsSS specifically creates iso-kolavalool as opposed to promiscuous cleavage by the endogenous *N. benthamiana* enzymes. Therefore, all active enzymes were found to produce iso-KDP, and none of the enzymes yielded conclusive evidence of (-)-KDP.

The presence of iso-KDP synthases in *T. chamaedrys* is not surprising given an evolutionarily close relative, *A. reptans*, has an ortholog, ArTPS2. Additionally, there are various *Teucrium* furanoclerodanes reported with a 4,18 double bond, 4,18 epoxides, and C18 esters lacking a C3-C4 double bond, all which presumably come from an iso-KDP precursor. All reported *T. chamaedrys* clerodanes are heavily modified and lack either double bond, but do have various C18 ester linkages (*Dictionary of Natural Products 30.2*), suggesting they are likely formed by an iso-KDP precursor.

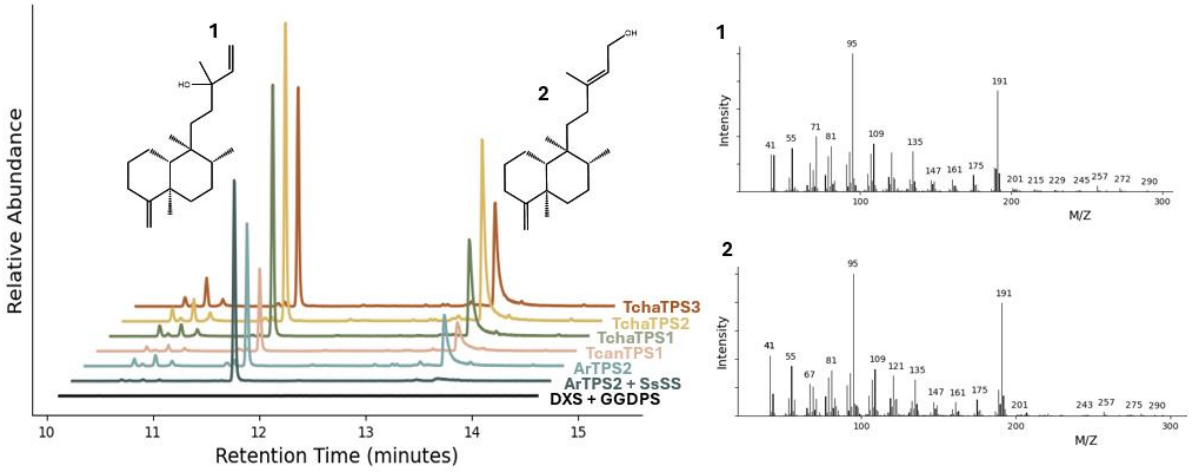


Figure 3.4 Extracted ion chromatogram (191 m/z) demonstrating iso-kolavenyl diphosphate synthase activity. Each extracted ion chromatogram was stacked and shifted to compare their products. Tested enzymes TchaTPS1, TchaTPS2, TchaTPS3, and TcanTPS1 were compared to the known iso-KDP producer, ArTPS2, and the negative control, DXS+GGDPS. DXS+GGDPS is present in all samples. Peak at ~11.5 min corresponds to iso-kolavalool (**1**) and the peak at ~13.5 min corresponds to iso-kolavanol (**2**). Representative mass spectra of ArTPS2 for iso-kolavalool (**1**) and iso-kolavenol (**2**) peaks. Mass spectra of all relevant peaks can be found in Supplemental Figure 3.6.

While we found no evidence for a dedicated (-)-KDP synthase in *T. chamaedrys*, some Lamiaceae species contain (-)-KDP synthases and accumulate furanoclerodanes, including *Teucrium* species with a 3,4 double bond (*Dictionary of Natural Products 30.2*). Given the presence of iso-KDP derived chemistries in *T. chamaedrys,* either there is a loss of the (-)-KDP enzyme in some species, or specific amino acid changes in the enzyme alter the products. Deprotonation of C3, as opposed to C18, likely only requires a shift of the base-acting residue by a few angstroms to alter which proton is being abstracted. It has been shown that blocking the deprotonation site of an *ent*-copalyl diphosphate (*ent*-CDP) synthase with a single amino acid shift can swap the *ent*-CDP synthase into a (-)-KDP synthase (Potter *et al.*, 2016). A third possibility is another unrelated TPS-c which may have convergently evolved (-)-KDP synthase activity.

This study represents the first functional characterization of diTPSs in *Teucrium*. This opens the avenue for characterization of enzymes in subsequent steps of diterpenoid metabolism, such as clerodane-

derived compounds like Teucrin, chamaedrosides, and neo-clerodanes. Understanding the natural pathways of these medicinally relevant compounds provides an important first step to biotechnological production and utilization of these terpenes in medicine and beyond.

Material and Methods

Plant growth conditions, tissue collection, and storage

The *T. chamaedrys* plant was purchased from Mountain Valley Growers (California, USA) and grown in a greenhouse. For DNA extraction, the plant was dark-adapted for 72 hours prior to harvesting. Healthy, mature leaves were collected, flash frozen in liquid Nitrogen, and stored at -80°C. For RNA extraction, healthy, mature leaves and rinsed roots were collected, flash frozen in liquid Nitrogen, and stored at -80°C.

Nucleotide isolation

High molecular weight (HMW) genomic DNA was extracted from *T. chamaedrys* leaves using a modified CTAB-based protocol (Li, Parris and Saski, 2020; Longley *et al.*, 2023). Briefly, frozen tissue was ground into a fine powder with a mortar and pestle in liquid Nitrogen and resuspended in nuclear isolation buffer. After nuclei were isolated, CTAB was added, HMW nucleic acids were extracted with chloroform and isoamyl alcohol, washed with isopropanol, and RNAse treated (Thermo Fisher Scientific, MA, USA). Short read DNA was extracted using DNeasy Plant Mini Kit (QIAGEN, Hilden, Germany).

Library preparation and sequencing

DNA libraries for long-read sequencing with Nanopore (Oxford Nanopore Technologies Ltd, USA) were prepared using the Oxford Nanopore SQK-LSK114 Ligation Sequencing Kit V14, and the library was loaded onto a PromethION FLO-PRO114M (R10.4.1) flow cell. Prior to long-read sequencing, the DNA was processed with the Standard Short Read Eliminator Kit (Circulomics Inc., MD, USA). The MinKNOW (v22.10.07) and base calling was performed using Guppy (v6.3.9) with the High Accuracy (HAC) model.

For short read-sequencing, library preparation was done using Roche Kapa HyperPrep DNA Library Kit with Unique Dual Index adapters (Sigma-Aldrich, MO, USA). The completed library was quality assessed and quantified using Qubit dsDNA HS, Agilent 4200 TapeStation HS DNA1000, and the Invitrogen Collibri Illumina Library Quantification qPCR assays. Sample was loaded onto one lane of an Illumina v1.5 S4 flow cell using the Xp Workflow. Sequencing was performed in a 2x150 bp paired end format using a NovaSeq 6000 (v1.5) 300 cycle reagent cartridge (Illumina Inc., CA, USA). Base calling was done by Illumina Real Time Analysis (RTA; v3.4.4) and output from the RTA software was demultiplexed and converted to FASTQ format with Illumina's Bcl2fastq (v2.20.0).

Genome size and heterozygosity estimation

Jellyfish (v.2.3.0; Marçais and Kingsford, 2011) was used to estimate the size and heterozygosity of the genome via k-mer analysis. 31-mers from the Illumina DNA libraries were used after trimming and filtering.

Ploidy analysis

KMC (Kokot, Długosz and Deorowicz, 2017) was also used to count k-mers in the genome using k-mer length of 31, yielding 4,876,867,453 unique k-mers. K-mer analysis was visualized using GenomeScope (v.1.0; Supplemental Figure 3.1; Vurture *et al.*, 2017). Subsequently, ploidy was measured using Smudgeplot analysis (Ranallo-Benavidez, Jaron and Schatz, 2020). The lower cutoff for coverage threshold was 12 and the upper cutoff was 2,800, estimated using 'cutoff' from the Smudgeplot suite. The Smudgeplot analysis output was hand annotated according to Ranallo-Benavidez, Jaron and Schatz, 2020, as the original output did not include the 'AAAB' annotation (Figure 3.2; Supplemental Figure 3.3).

Genome assembly

Raw Nanopore DNA reads with mean Q-scores greater than 7 were used and processed with Porechop (v.0.2.4) to remove adapters, Chopper (v.0.8.0-0; De Coster *et al.*, 2018) to filter reads less than 10 Kb, and Filtlong (v.0.2.0) to filter the out the worst 10% of reads based on quality. Sequences were then

assembled using Flye (v.2.9; Kolmogorov *et al.*, 2019) with minimum overlap of 5 Kbp, two iterations of polishing, and retaining the haplotypes. The draft assembly was polished once using Medaka (v.1.4.3) and the model 'r1041_e82_400bps_hac_g632'. BWA-MEM2 (v.2.0; Vasimuddin *et al.*, 2019) was used to align the Illumina paired-end reads to the draft assembly for error correction. The resulting draft assembly was polished with one round of Pilon (v.1.24; Walker *et al.*, 2014) using the diploid option. Contigs smaller than 100 Kbp were then removed from the assembly. In order to eliminate potential contamination present, Kraken2 (v. 2.1.3; Wood, Lu and Langmead, 2019) was used with the database 'PlusPFP' (https://benlangmead.github.io/aws-indexes/k2). Approximately 0.44% of the assembly was determined to be human and was subsequently removed. No further contamination was detected.

Genome annotation

The draft genome was first mined for *de novo* repeats using Repeat Modeler (v.2.0.2a; Flynn *et al.*, 2020). These *de novo* repeats along with Viridiplantae repeats from RepBase were used by Repeat Masker (v.4.1.1; Chen, 2004) to mask the draft genome. Next RNA-seq data from *T. chamaedrys* (PRJNA1124528) mature leaf and root tissues were aligned to the draft genome using HISAT2 (v.2.1.0; Kim *et al.*, 2019). In addition to this transcript evidence, protein evidence from the closely related species *T. marum* (Smit *et al.*, 2024) was used as an input for BRAKER (v.2.1.6; Altschul *et al.*, 1990; Stanke *et al.*, 2006, 2008; Camacho *et al.*, 2009; Quinlan, 2014; Kovaka *et al.*, 2019; Pertea and Pertea, 2020; Gabriel *et al.*, 2021; Bruna, Lomsadze and Borodovsky, 2023) with the flag ' --etpmode' to create initial gene models. Generated gene models were then fed into MAKER (Law *et al.*, 2015), along with RNA-seq evidence and protein evidence from *A. thaliana* (TAIR v.11; Cheng *et al.*, 2017) and *T. marum* (Smit *et al.*, 2024) to create a working gene model set (Supplemental Table 3.3).

This yielded 217,373 working gene models, which were later filtered down to 128,111 high confidence gene models. Of the original 217,373 gene models, 153,810 had an annotation edit distance (AED) score less than one and/or contained a protein domain, meaning there was evidence for transcripts or protein

homology (Yandell and Ence, 2012). Of those, we kept one gene model per locus, which generated 144,380 gene models. Although a repeat-masked genome was used initially, we found additional TE-related genes, which when removed, yielded 134,486 gene models. All genes less than 300 bp were then removed leaving 128,264 gene models. Finally, we removed non-plant contamination according to Kraken2 (Wood, Lu and Langmead, 2019) for a final high confidence gene model count of 128,111.

Chromosome counting

Root tips were harvested from greenhouse-grown rooted cuttings and pretreated with nitrous oxide at a pressure of 160 psi (approximately 10.9 atm) for 40 min. Subsequently, the root tips were fixed in a solution of three parts ethanol to one part acetic acid and maintained at 22°C until ready for enzymatic treatment. An enzymatic solution containing 4% cellulase (Yakult Pharmaceutical, Tokyo, Japan), 2% pectinase (Plant Media, Dublin, OH, USA), and 2% pectolyase (Sigma Chemical, St. Louis, MO, USA) was used to digest the root tips for 50 min at 37°C. Chromosomes were prepared using a stirring method as described by Xin *et al.*, 2020 and counterstained with 4',6-diamidino-2-phenylindole (DAPI) in VectaShield antifade solution (Vector Laboratories, Burlingame, CA, USA). Images were captured with a QImaging Retiga EXi Fast 1394 CCD camera (Teledyne Photometrics, Tucson, AZ, USA) attached to an Olympus BX51 epifluorescence microscope. Image processing was performed using Meta Imaging Series 7.5 software, and the final image contrast was adjusted using Adobe Photoshop (Adobe, San Jose, CA, USA). Chromosome counting was conducted on at least 10 metaphase spreads.

Phylogeny

T. canadense reads were downloaded from NCBI SRA database (SRR5150734) and split files were assembled into a *de novo* transcriptome using Trinity (v.2.9.1; Grabherr *et al.*, 2011). The resulting mRNA were filtered for the longest open reading frame and translated into protein sequences using TransDecoder (v. 2.1.0; Haas, BJ. https://github.com/TransDecoder/TransDecoder) *T. marum* gene models were downloaded from Figshare

(https://figshare.com/articles/dataset/Teucrium_marum_genome_assembly/25109411). The representative gene models from each of the three *Teucrium* species were blasted (BLAST+ v. 2.13.0, e-value = 1e-20; Camacho *et al.*, 2009) against a bait set of 34 functionally characterized diTPS (Supplemental Table 3.2). Resulting protein matches were identified and combined with the bait set. Multiple sequence alignments were generated using ClustalOmega (v. 1.2.4; Sievers *et al.*, 2011) and phylogenetic trees were generated using RAxML using the model 'protgammaauto', algorithm 'a', and 100 bootstrap replicates (v.8.2.12; Stamatakis, 2014).

Synteny

The BLAST function makeblastdb (E-value of 1e−10, 5 alignments) was used to create protein databases between *T. chamaedrys* and *T. marum* (Smit et al 2024). Syntenic analysis 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; Wang *et al.*, 2012). Results were visualized using SynVisio (Bandi and Gutwin, 2020).

Cloning and transient expression

Candidate enzymes from *T. chamaedrys* were synthesized (Twist Bioscience, CA, USA) and cloned into the plant expression vector pEAQ-HT (Sainsbury, Thuenemann and Lomonossoff, 2009) for use in transient expression in *N. benthamiana*. Sequences were validated via Sanger sequencing. *N. benthamiana* plants were grown for 4-5 weeks in a controlled growth room under 12 H light and 12 H dark (22°C) cycle before infiltration. Coexpression constructs were transformed separately 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 approximately 10 mL water before being resuspended and diluted to an OD600 of 1.0 in water with 200 µM acetosyringone. Cultures were incubated at 30°C for 1–2 hours, then separate cultures were mixed in an equal ratio for each combination of enzymes. *N. benthamiana* leaves were infiltrated into the

underside (abaxial side) with a 1 mL syringe. 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: *Plectranthus barbatus* 1-deoxy-D-xylulose-5-phosphate synthase (PbDXS) and GGDP synthase (PbGGDPS), in order to boost production of the diterpene precursor GGDP (Andersen-Ranberg *et al.*, 2016). Plants were returned to the controlled growth room for 5 days. Approximately 200 mg fresh weight from three separate infiltrated leaves was extracted with 1.5 mL hexane overnight at room temperature. Plant material was collected by centrifugation, and the organic phase was removed for GC-MS analysis.

GC-MS analysis

All GC-MS analyses were performed 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 with a He carrier gas (1 mL/min). The detector was activated following a 4 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.

REFERENCES

Achaz, G. *et al.* (2000) 'Analysis of Intrachromosomal Duplications in Yeast *Saccharomyces cerevisiae*: A Possible Model for Their Origin', *Molecular Biology and Evolution*, 17(8), pp. 1268–1275. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a026410.

Altschul, S.F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. Available at: https://doi.org/10.1016/S0022-2836(05)80360-2.

Andersen-Ranberg, J. *et al.* (2016) 'Expanding the landscape of diterpene structural diversity through stereochemically controlled combinatorial biosynthesis', *Angewandte Chemie (International Ed. in English)*, 55(6), pp. 2142–2146. Available at: https://doi.org/10.1002/anie.201510650.

Arı, S. *et al.* (2015) 'Ethnobotanical survey of plants used in Afyonkarahisar-Turkey', *Journal of Ethnobiology and Ethnomedicine*, 11(1), pp. 1–15. Available at: https://doi.org/10.1186/s13002-015-0067-6.

Bandi, V. and Gutwin, C. (2020) '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)*. *Interactive Exploration of Genomic Conservation*, Waterloo, CAN: Canadian Human-Computer Communications Society. Available at: https://synvisio.github.io/#/.

Bedir, E., Manyam, R. and Khan, I.A. (2003) '*Neo*-clerodane diterpenoids and phenylethanoid glycosides from *Teucrium chamaedrys* L.', *Phytochemistry*, 63(8), pp. 977–983. Available at: https://doi.org/10.1016/S0031-9422(03)00378-9.

Birchler, J.A. and Yang, H. (2022) 'The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation', *The Plant Cell*, 34(7), pp. 2466–2474. Available at: https://doi.org/10.1093/plcell/koac076.

Blasio, F. *et al.* (2022) 'Genomic and Meiotic Changes Accompanying Polyploidization', *Plants*, 11(1), p. 125. Available at: https://doi.org/10.3390/plants11010125.

Bruna, T., Lomsadze, A. and Borodovsky, M. (2023) 'GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistence with Extrinsic Data'. bioRxiv, p. 2023.01.13.524024. Available at: https://doi.org/10.1101/2023.01.13.524024.

Bryson, A.E. *et al.* (2023) 'Uncovering a miltiradiene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory', *Nature Communications*, 14(1), p. 343. Available at: https://doi.org/10.1038/s41467-023-35845-1.

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. Available at: https://doi.org/10.1186/1471-2105-10-421.

Candela, R.G. *et al.* (2020) 'A Review of the Phytochemistry, Traditional Uses and Biological Activities of the Essential Oils of Genus *Teucrium*', *Planta Medica*, 87, pp. 432–479. Available at: https://doi.org/10.1055/a-1293-5768.

Caniard, A. *et al.* (2012) 'Discovery and functional characterization of two diterpene synthases for sclareol biosynthesis in *Salvia sclarea* (L.) and their relevance for perfume manufacture', *BMC Plant Biology*, 12, p. 119. Available at: https://doi.org/10.1186/1471-2229-12-119.

Chen, N. (2004) 'Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences', *Current Protocols in Bioinformatics*, 5(1), p. 4.10.1-4.10.14. Available at: https://doi.org/10.1002/0471250953.bi0410s05.

Cheng, C.-Y. *et al.* (2017) 'Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome', *The Plant Journal*, 89(4), pp. 789–804. Available at: https://doi.org/10.1111/tpj.13415.

De Coster, W. *et al.* (2018) 'NanoPack: visualizing and processing long-read sequencing data', *Bioinformatics (Oxford, England)*, 34(15), pp. 2666–2669. Available at: https://doi.org/10.1093/bioinformatics/bty149.

De Storme, N. and Mason, A. (2014) 'Plant speciation through chromosome instability and ploidy change: Cellular mechanisms, molecular factors and evolutionary relevance', *Current Plant Biology*, 1, pp. 10–33. Available at: https://doi.org/10.1016/j.cpb.2014.09.002.

Dictionary of Natural Products 30.2. Available at: https://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml (Accessed: 11 March 2022).

Emms, D.M. and Kelly, S. (2019) 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome Biology*, 20(1), p. 238. Available at: https://doi.org/10.1186/s13059-019-1832-y.

Field, B. *et al.* (2011) 'Formation of plant metabolic gene clusters within dynamic chromosomal regions', *Proceedings of the National Academy of Sciences*, 108(38), pp. 16116–16121. Available at: https://doi.org/10.1073/pnas.1109273108.

Fiorentino, A. *et al.* (2009) 'Potential allelopathic effect of neo-clerodane diterpenes from *Teucrium chamaedrys* (L.) on stenomediterranean and weed cosmopolitan species', *Biochemical Systematics and Ecology*, 37(4), pp. 349–353. Available at: https://doi.org/10.1016/j.bse.2009.06.006.

Flynn, J.M. *et al.* (2020) 'RepeatModeler2 for automated genomic discovery of transposable element families', *Proceedings of the National Academy of Sciences*, 117(17), pp. 9451–9457. Available at: https://doi.org/10.1073/pnas.1921046117.

Gabriel, L. *et al.* (2021) 'TSEBRA: transcript selector for BRAKER', *BMC Bioinformatics*, 22(1), p. 566. Available at: https://doi.org/10.1186/s12859-021-04482-0.

Godden, G.T. *et al.* (2019) 'Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints', *Genome Biology and Evolution*, 11(12), pp. 3393–3408. Available at: https://doi.org/10.1093/gbe/evz239.

Grabherr, M.G. *et al.* (2011) 'Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data', *Nature biotechnology*, 29(7), pp. 644–652. Available at: https://doi.org/10.1038/nbt.1883.

Haas, B. *TransDecoder/TransDecoder: TransDecoder source*. Available at: https://github.com/TransDecoder/TransDecoder (Accessed: 15 August 2024).

Hamilton, J.P. *et al.* (2020) 'Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, *Callicarpa americana*', *GigaScience*, 9(9), p. giaa093. Available at: https://doi.org/10.1093/gigascience/giaa093.

Hansen, N.L. *et al.* (2017) 'The terpene synthase gene family in *Tripterygium wilfordii* harbors a labdane-type diterpene synthase among the monoterpene synthase TPS-b subfamily', *The Plant Journal*, 89(3), pp. 429–441. Available at: https://doi.org/10.1111/tpj.13410.

Heskes, A.M. *et al.* (2018) 'Biosynthesis of bioactive diterpenoids in the medicinal plant *Vitex agnus-castus*', *The Plant Journal*, 93(5), pp. 943–958. Available at: https://doi.org/10.1111/tpj.13822.

Hughes, A.L. *et al.* (2003) 'Non-random association of transposable elements with duplicated genomic blocks in *Arabidopsis thaliana*', *Molecular Phylogenetics and Evolution*, 29(3), pp. 410–416. Available at: https://doi.org/10.1016/S1055-7903(03)00262-8.

Jarić, S., Mitrović, M. and Pavlović, P. (2020) 'Ethnobotanical Features of *Teucrium* Species', in M. Stanković (ed.) *Teucrium Species: Biology and Applications*. Cham: Springer International Publishing, pp. 111–142. Available at: https://doi.org/10.1007/978-3-030-52159-2_5.

Johnson, S.R. *et al.* (2019) 'A database-driven approach identifies additional diterpene synthase activities in the mint family (Lamiaceae)', *Journal of Biological Chemistry*, 294(4), pp. 1349–1362. Available at: https://doi.org/10.1074/jbc.RA118.006025.

Kim, D. *et al.* (2019) 'Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype', *Nature Biotechnology*, 37(8), pp. 907–915. Available at: https://doi.org/10.1038/s41587-019-0201-4.

Klein Gebbinck, E.A., Jansen, B.J.M. and de Groot, A. (2002) 'Insect antifeedant activity of clerodane diterpenes and related model compounds', *Phytochemistry*, 61(7), pp. 737–770. Available at: https://doi.org/10.1016/S0031-9422(02)00174-7.

Kokot, M., Długosz, M. and Deorowicz, S. (2017) 'KMC 3: counting and manipulating k-mer statistics', *Bioinformatics*, 33(17), pp. 2759–2761. Available at: https://doi.org/10.1093/bioinformatics/btx304.

Kolmogorov, M. *et al.* (2019) 'Assembly of long, error-prone reads using repeat graphs', *Nature Biotechnology*, 37(5), pp. 540–546. Available at: https://doi.org/10.1038/s41587-019-0072-8.

Kovaka, S. *et al.* (2019) 'Transcriptome assembly from long-read RNA-seq alignments with StringTie2', *Genome Biology*, 20(1), p. 278. Available at: https://doi.org/10.1186/s13059-019-1910-1.

Landis, J.B. *et al.* (2018) 'Impact of whole-genome duplication events on diversification rates in angiosperms', *American Journal of Botany*, 105(3), pp. 348–363. Available at: https://doi.org/10.1002/ajb2.1060.

Law, M. *et al.* (2015) 'Automated Update, Revision, and Quality Control of the Maize Genome Annotations Using MAKER-P Improves the B73 RefGen_v3 Gene Models and Identifies New Genes', *Plant Physiology*, 167(1), pp. 25–39. Available at: https://doi.org/10.1104/pp.114.245027.

Li, R., Morris-Natschke, S.L. and Lee, K.-H. (2016) 'Clerodane diterpenes: sources, structures, and biological activities', *Natural product reports*, 33(10), pp. 1166–1226. Available at: https://doi.org/10.1039/c5np00137d.

Li, Z. *et al.* (2021) 'Patterns and Processes of Diploidization in Land Plants', *Annual Review of Plant Biology*, 72(Volume 72, 2021), pp. 387–410. Available at: https://doi.org/10.1146/annurev-arplant-050718-100344.

Li, Z., Parris, S. and Saski, C.A. (2020) 'A simple plant high-molecular-weight DNA extraction method suitable for single-molecule technologies', *Plant Methods*, 16(1), p. 38. Available at: https://doi.org/10.1186/s13007-020-00579-4.

Longley, R. *et al.* (2023) 'Comparative genomics of Mollicutes-related endobacteria supports a late invasion into Mucoromycota fungi', *Communications Biology*, 6(1), pp. 1–13. Available at: https://doi.org/10.1038/s42003-023-05299-8.

Manni, M. *et al.* (2021) 'BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes', *Molecular Biology and Evolution*, 38(10), pp. 4647–4654. Available at: https://doi.org/10.1093/molbev/msab199.

Marçais, G. and Kingsford, C. (2011) 'A fast, lock-free approach for efficient parallel counting of occurrences of k-mers', *Bioinformatics*, 27(6), pp. 764–770. Available at: https://doi.org/10.1093/bioinformatics/btr011.

Menichini, F. *et al.* (2009) 'Phytochemical composition, anti-inflammatory and antitumour activities of four *Teucrium* essential oils from Greece', *Food Chemistry*, 115(2), pp. 679–686. Available at: https://doi.org/10.1016/j.foodchem.2008.12.067.

Milutinović, M.G. *et al.* (2019) 'Potential of *Teucrium chamaedrys* L. to modulate apoptosis and biotransformation in colorectal carcinoma cells', *Journal of Ethnopharmacology*, 240, p. 111951. Available at: https://doi.org/10.1016/j.jep.2019.111951.

Ohno, S. (1970) 'Duplication for the Sake of Producing More of the Same', in S. Ohno (ed.) *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer, pp. 59–65. Available at: https://doi.org/10.1007/978-3-642-86659-3_11.

Osborn, T.C. *et al.* (2003) 'Understanding mechanisms of novel gene expression in polyploids', *Trends in Genetics*, 19(3), pp. 141–147. Available at: https://doi.org/10.1016/S0168-9525(03)00015-5.

Panchy, N., Lehti-Shiu, M. and Shiu, S.-H. (2016) 'Evolution of Gene Duplication in Plants', *Plant Physiology*, 171(4), pp. 2294–2316. Available at: https://doi.org/10.1104/pp.16.00523.

Pelot, K.A. *et al.* (2017) 'Biosynthesis of the psychotropic plant diterpene salvinorin A: Discovery and characterization of the *Salvia divinorum* clerodienyl diphosphate synthase', *The Plant Journal*, 89(5), pp. 885–897. Available at: https://doi.org/10.1111/tpj.13427.

Pelot, K.A. *et al.* (2018) 'Functional Diversity of Diterpene Synthases in the Biofuel Crop Switchgrass', *Plant Physiology*, 178(1), pp. 54–71. Available at: https://doi.org/10.1104/pp.18.00590.

Pertea, G. and Pertea, M. (2020) 'GFF Utilities: GffRead and GffCompare', *F1000Research*, 9, p. ISCB Comm J-304. Available at: https://doi.org/10.12688/f1000research.23297.2.

Pieroni, A., Quave, C.L. and Santoro, R.F. (2004) 'Folk pharmaceutical knowledge in the territory of the Dolomiti Lucane, inland southern Italy', *Journal of Ethnopharmacology*, 95(2), pp. 373–384. Available at: https://doi.org/10.1016/j.jep.2004.08.012.

Potter, K.C. *et al.* (2016) 'Blocking Deprotonation with Retention of Aromaticity in a Plant ent-Copalyl Diphosphate Synthase Leads to Product Rearrangement', *Angewandte Chemie International Edition*, 55(2), pp. 634–638. Available at: https://doi.org/10.1002/anie.201509060.

Qiu, T. *et al.* (2023) 'Tandem duplication and sub-functionalization of clerodane diterpene synthase originate the blooming of clerodane diterpenoids in *Scutellaria barbata*', *The Plant Journal*, 116(2), pp. 375–388. Available at: https://doi.org/10.1111/tpj.16377.

Quinlan, A.R. (2014) 'BEDTools: The Swiss-Army Tool for Genome Feature Analysis', *Current Protocols in Bioinformatics*, 47(1), p. 11.12.1-11.12.34. Available at: https://doi.org/10.1002/0471250953.bi1112s47.

Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. (2020) 'GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes', *Nature Communications*, 11(1), p. 1432. Available at: https://doi.org/10.1038/s41467-020-14998-3.

Ranjbar, M., Mahmoudi, C. and Nazari, H. (2018) 'An overview of chromosomal criteria and biogeography in the genus *Teucrium* (Lamiaceae)', *Caryologia*, 71(1), pp. 63–79. Available at: https://doi.org/10.1080/00087114.2017.1420587.

Ren, R. *et al.* (2018) 'Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms', *Molecular Plant*, 11(3), pp. 414–428. Available at: https://doi.org/10.1016/j.molp.2018.01.002.

Ritz, M. *et al.* (2023) 'Comparative Genome-Wide Analysis of Two *Caryopteris x Clandonensis* Cultivars: Insights on the Biosynthesis of Volatile Terpenoids', *Plants*, 12(3), p. 632. Available at: https://doi.org/10.3390/plants12030632.

Sadeghi, Z. *et al.* (2022) 'A review of the phytochemistry, ethnopharmacology and biological activities of *Teucrium* genus (Germander)', *Natural Product Research*, 36(21), pp. 5647–5664. Available at: https://doi.org/10.1080/14786419.2021.2022669.

Sainsbury, F., Thuenemann, E.C. and Lomonossoff, G.P. (2009) 'pEAQ: Versatile expression vectors for easy and quick transient expression of heterologous proteins in plants', *Plant Biotechnology Journal*, 7(7), pp. 682–693. Available at: https://doi.org/10.1111/j.1467-7652.2009.00434.x.

Salmaki, Y. *et al.* (2016) 'Phylogeny of non-monophyletic *Teucrium* (Lamiaceae: Ajugoideae): Implications for character evolution and taxonomy', *TAXON*, 65(4), pp. 805–822. Available at: https://doi.org/10.12705/654.8.

Schlecht, N.J. *et al.* (2024) 'CYP76BK1 orthologs catalyze furan and lactone ring formation in clerodane diterpenoids across the mint family'. bioRxiv, p. 2024.08.28.609960. Available at: https://doi.org/10.1101/2024.08.28.609960.

Sievers, F. *et al.* (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology*, 7(1), p. 539. Available at: https://doi.org/10.1038/msb.2011.75.

Smit, S.J. *et al.* (2024) 'The genomic and enzymatic basis for iridoid biosynthesis in cat thyme (*Teucrium marum*)', *The Plant Journal*, n/a(n/a). Available at: https://doi.org/10.1111/tpj.16698.

Stamatakis, A. (2014) 'RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. Available at: https://doi.org/10.1093/bioinformatics/btu033.

Stanke, M. *et al.* (2006) 'AUGUSTUS: Ab initio prediction of alternative transcripts', *Nucleic Acids Research*, 34(suppl_2), pp. W435–W439. Available at: https://doi.org/10.1093/nar/gkl200.

Stanke, M. *et al.* (2008) 'Using native and syntenically mapped cDNA alignments to improve de novo gene finding', *Bioinformatics*, 24(5), pp. 637–644. Available at: https://doi.org/10.1093/bioinformatics/btn013.

di Tizio, A. *et al.* (2012) 'Traditional food and herbal uses of wild plants in the ancient South-Slavic diaspora of Mundimitar/Montemitro (Southern Italy)', *Journal of Ethnobiology and Ethnomedicine*, 8(1), p. 21. Available at: https://doi.org/10.1186/1746-4269-8-21.

Vasimuddin, Md. *et al.* (2019) 'Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems', in. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 314–324. Available at: https://doi.org/10.1109/IPDPS.2019.00041.

Vurture, G.W. *et al.* (2017) 'GenomeScope: fast reference-free genome profiling from short reads', *Bioinformatics*, 33(14), pp. 2202–2204. Available at: https://doi.org/10.1093/bioinformatics/btx153.

Walker, B.J. *et al.* (2014) 'Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement', *PLoS ONE*, 9(11), p. e112963. Available at: https://doi.org/10.1371/journal.pone.0112963.

Wang, X. *et al.* (2021) 'Genome downsizing after polyploidy: mechanisms, rates and selection pressures', *The Plant Journal*, 107(4), pp. 1003–1015. Available at: https://doi.org/10.1111/tpj.15363.

Wang, Y. *et al.* (2012) 'MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity', *Nucleic Acids Research*, 40(7), pp. e49–e49. Available at: https://doi.org/10.1093/nar/gkr1293.

Wood, D.E., Lu, J. and Langmead, B. (2019) 'Improved metagenomic analysis with Kraken 2', *Genome Biology*, 20(1), p. 257. Available at: https://doi.org/10.1186/s13059-019-1891-0.

Wood, T.E. *et al.* (2009) 'The frequency of polyploid speciation in vascular plants', *Proceedings of the National Academy of Sciences*, 106(33), pp. 13875–13879. Available at: https://doi.org/10.1073/pnas.0811575106.

Xin, H. *et al.* (2020) 'An extraordinarily stable karyotype of the woody *Populus* species revealed by chromosome painting', *The Plant Journal: For Cell and Molecular Biology*, 101(2), pp. 253–264. Available at: https://doi.org/10.1111/tpj.14536.

Yandell, M. and Ence, D. (2012) 'A beginner's guide to eukaryotic genome annotation', *Nature Reviews Genetics*, 13(5), pp. 329–342. Available at: https://doi.org/10.1038/nrg3174.

CHAPTER 4: FUTURE DIRECTIONS

Abigail E Bryson, Björn Hamberger

Presence of the miltiradiene-containing biosynthetic gene cluster in the Lamiales

In chapter 2, I searched the genome of a single Lamiales representative, *E. lutea*, looking for the presence of the miltiradiene-containing BGC present in the Lamiaceae. My finding of a diTPS gene pair implicated the ancestral cluster to be older than the Lamiaceae; however, the lack of relevant CYPs in the corresponding genomic area, paired with the larger evolutionary distance (approx. 104 MY; Godden *et al.*, 2019) between the species being compared made determining the presence/absence of this cluster beyond the Lamiaceae inconclusive. A similar setup, using a panel of species with highly contiguous genomes in sister families, could be used to determine this BGC's full phylogenetic presence/absence beyond the Lamiaceae. I hypothesize that there may be a core set of genes, such as the predicted ancestral cluster, which are common across the higher core Lamiales (Lamiaceae, Phrymaceae, Mazaceae, Paulownaceae, Orbanchaceae). But, the overall function of the cluster and associated genes may have evolved in different ways, leading to alternative metabolites and phenotypes. If identified, instances of independent evolution and neofunctionalization from this cluster could also be associated with speciation and adaptation. *In planta*, it is believed that miltiradiene plays a role in plant defense (Machumi *et al.*, 2010; González, 2015), so species without this enzyme may have to compensate in other ways, such as deriving alternative metabolites with similar function.

Another feature that could affect the way in which miltiradiene is used *in planta* is the regulation of this BGC. Although not previously reported, there appears to be a regulatory element that is common amongst the miltiradiene BGCs throughout the Lamiaceae. Further investigation must be done to determine if it truly is part of the regulation of this cluster, but its conservation in the cluster across this family indicates that it may be critical for regulation by the plant.

The number and position of diTPSs in Teucrium species

In chapter 3, I found the majority (11/15) of diTPSs present in *Teucrium marum* were physically clustered on chromosome 10, suggesting that the diTPSs in this area may have been formed from tandem

duplication and subsequent neofunctionalization. This is not an uncommon phenomenon: we see large groupings of similar genes via tandem duplication or transposable elements in *Vitis vinifera* (Martin *et al.*, 2010) and *Triticum* species (Liu *et al.*, 2024). However, this is a stark example of concentrated tandem duplication and warrants a closer look.

In plants, it is apparent that one prevalent mechanism of BGC formation involves tandem duplication and subsequent neofunctionalization (Nützmann, Scazzocchio and Osbourn, 2018). Since *Teucrium* species studied here have clearly undergone this, it may be a unique place to study birth, or at least evolution, of clusters. Additionally, the availability of the genomes of the two closely related species, *T. marum* and *T. chamaedrys*, uniquely allows for an in-depth study of the last 4 MY (Salmaki *et al.*, 2016). In addition to the diTPSs being physically clustered, there are also many more in total than expected; for example, tomato has 6 diTPSs (Zhou and Pichersky, 2020) and *Arabidopsis* has 3 (Tholl and Lee, 2011). The 15 diTPSs in *T. marum* and 76 in *T. chamaedrys* stand out starkly as diTPS-rich species. RNA-seq could provide insight into whether these enzymatic products influence defense, tissue-specific and/or development-specific expression. Additionally, there is a chromosome-level assembly for *Ajuga chamaepitys* (Mian *et al.*, 2024) and *A. decumbens* (Gao *et al.*, 2024), more members of the Ajugoideae. Comparative genomics, such as syntenic analysis, transcriptomics and phylogenetics, could provide more context for the evolution of the enzymes in this BGC.

Polyphyletic origins of clerodane synthases in Lamiaceae

In chapter 3, I functionally characterized all the putative clerodane synthases found in *T. chamaedrys* and *T. canadense*. Clerodanes are widespread in biology, being present in plants, fungi, and bacteria. Within the Lamiaceae family, there appears to be polyphyletic origins of clerodane synthases, with those in Salvia forming one clade and those in the Scutellarioideae forming a second (Li *et al.*, 2023). Based on species-level phylogeny (Godden *et al.*, 2019), I hypothesize that the clerodane synthases present in *Teucrium* species are among the second clade, as Scutellarioideae and Ajugoideae are closely related sub

families. However, molecular analysis such as K_a/K_s and syntenic analysis could more firmly substantiate these claims.

REFERENCES

Gao, Y. *et al.* (2024) 'Chromosome-level genome assembly of *Ajuga decumbens*', *Frontiers in Plant Science*, 15. Available at: https://doi.org/10.3389/fpls.2024.1413468.

Godden, G.T. *et al.* (2019) 'Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints', *Genome Biology and Evolution*, 11(12), pp. 3393–3408. Available at: https://doi.org/10.1093/gbe/evz239.

González, M.A. (2015) 'Aromatic abietane diterpenoids: Their biological activity and synthesis', *Natural Product Reports*, 32(5), pp. 684–704. Available at: https://doi.org/10.1039/c4np00110a.

Li, H. *et al.* (2023) 'The genomes of medicinal skullcaps reveal the polyphyletic origins of clerodane diterpene biosynthesis in the family Lamiaceae', *Molecular Plant*, 16(3), pp. 549–570. Available at: https://doi.org/10.1016/j.molp.2023.01.006.

Liu, Yiyang *et al.* (2024) 'Genome-Wide Identification and Evolution-Profiling Analysis of TPS Gene Family in *Triticum* Plants', *International Journal of Molecular Sciences*, 25(15), p. 8546. Available at: https://doi.org/10.3390/ijms25158546.

Machumi, F. *et al.* (2010) 'Antimicrobial and antiparasitic abietane diterpenoids from the roots of *Clerodendrum eriophyllum*', *Natural Product Communications*, 5(6), p. 1934578X1000500605. Available at: https://doi.org/10.1177/1934578X1000500605.

Martin, D.M. *et al.* (2010) '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 Biology*, 10(1), p. 226. Available at: https://doi.org/10.1186/1471-2229-10-226.

Mian, S. *et al.* (2024) 'The genome sequence of yellow bugle, *Ajuga chamaepitys* (L.) Schreb. (Lamiaceae)', *Wellcome Open Research* [Preprint].

Nützmann, H.-W., Scazzocchio, C. and Osbourn, A. (2018) 'Metabolic gene clusters in eukaryotes', *Annual Review of Genetics*, 52(1), pp. 159–183. Available at: https://doi.org/10.1146/annurev-genet-120417-031237.

Salmaki, Y. *et al.* (2016) 'Phylogeny of non-monophyletic *Teucrium* (Lamiaceae: Ajugoideae): Implications for character evolution and taxonomy', *TAXON*, 65(4), pp. 805–822. Available at: https://doi.org/10.12705/654.8.

Tholl, D. and Lee, S. (2011) 'Terpene Specialized Metabolism in *Arabidopsis thaliana*', *The Arabidopsis Book / American Society of Plant Biologists*, 9, p. e0143. Available at: https://doi.org/10.1199/tab.0143.

Zhou, F. and Pichersky, E. (2020) 'The complete functional characterisation of the terpene synthase family in tomato', *New Phytologist*, 226(5), pp. 1341–1360. Available at: https://doi.org/10.1111/nph.16431.

APPENDIX

Supplemental documentation referenced in the text in Chapter 2 and Chapter 3 can be found on the Dryad Data Repository, available here: https://doi.org/10.5061/dryad.8w9ghx3wm.