5-8-2018

Investigation of the role of gene clusters in terpene biosynthesis in Sorghum bicolor

Rebecca Hay
rfhvh8@mail.umsl.edu

Investigation of the role of gene clusters in terpene biosynthesis in *Sorghum bicolor*

by

Rebecca F. K. Hay

B. S. Biology, University of Guelph 2012

A Thesis/Dissertation

Submitted to The Graduate School of the

University of Missouri-St. Louis
in partial fulfillment of the requirements for the degree

Master of Science

in

Biology

May 2018

Advisory Committee

Bethany Zolman, Ph.D.
Chairperson

Toni M. Kutchan, Ph.D.

Elizabeth Kellogg, Ph.D.

Wendy Olivas, Ph.D.

Abstract

The staple crop *Sorghum bicolor* shows potential as a source of secondary metabolite-based biofuels due to its diverse phenotype and chemical profile. *S. bicolor* produces a variety of high-energy metabolites, including terpenes which are a potential renewable source of fuel additives. Information on the biosynthetic and genetic pathways by which *S. bicolor* terpenes are produced is limited and these pathways must be better understood before they can be engineered for human applications. Recent work on plant biosynthetic pathways has shown that terpenes can be modified by the products of clustered genes. Identification of biosynthetic gene clusters may accelerate the elucidation of complete pathways, but few have been characterized in *S. bicolor*. The aims of this thesis were to identify a putative terpene biosynthetic gene cluster in *S. bicolor*, characterize the terpene synthase in the cluster, and express the terpene synthase alongside clustered enzymes to determine if they modify the terpene skeleton structure. The terpene synthase Sobic.001G339000 was found to produce a novel sesquiterpene product. Mass spectra analysis suggested that the novel product was similar to guaiol and β-eudesmol and possibly shared a mass (222.2 Da) and chemical formula ($C_{15}H_{26}O$) with these compounds. Transient expression of the putative gene cluster in *N. benthamiana* produced a metabolite of a significantly higher mass than anticipated based on the hypothesized mass of the unknown terpene. Elucidation of a structure by NMR spectroscopy will be required to characterize the unknown terpene product. Once the structure of the terpene is known, analysis of the metabolic profile of transfected *N. benthamiana* will be simplified and the effect of clustered enzymes on the terpene product can be better explored.

Acknowledgements

I would first like to thank Dr. Toni Kutchan for providing me the resources to conduct this research in her lab as well as all her guidance. I would also like to thank Dr. Elizabeth Kellogg for providing me with the expertise and experience to design this project. I would like to thank Dr. Bethany Zolman, my academic advisor, for all her assistance with my academic career. I would also like to thank Dr. Wendy Olivas for her contributions to my graduate committee. In addition, I want to thank Dr. Michael McKain for his invaluable assistance with the computational and bioinformatics portion of this project and Megan Augustin for her repeated and essential assistance with laboratory work and experimental design. This project could not have been completed without their input. I am also grateful to Dr. Fuzhong Zhang from Washington University in St. Louis for providing the plasmid containing the mevalonate pathway, pBbA5c-Mev, and Julie Gauthier from the Donald Danforth Plant Science Center STARS research associate program for donating her time to assist with the transfection of tobacco plants. Finally, I would like to thank Dr. Bradley Evans and Jon Mattingly for providing their expertise in the use of the Proteomics and Mass Spectrometry facility at the Donald Danforth Plant Science Center

Table of Contents

Chapter 1: Introduction to gene clustering in terpene biosynthesis

1.1 Production and applications of terpenes

1.1.1 Terpene biosynthesis

Secondary metabolites are compounds produced by an organism that are not essential to the organism's function (Osbourn, 2010). In plants, secondary metabolites have many roles, including as chemical deterrents to predation and disease, allelopathic competition against neighbouring plants, protection from environmental stresses and as pollinator attractants (Boutanaev et al., 2015; Nützmann et al., 2016). The study of secondary metabolites is essential for improving the natural defences of crop plants against pests and diseases and for investigating compounds that have applications in industry and medicine. The complete biosynthetic pathway of a metabolite, including all genetic and biochemical components, must be understood in order to manipulate the pathway for human use. The majority of secondary metabolite biosynthetic pathways in the plant kingdom remains uncharacterized, so little can be done to improve endogenous metabolite production or develop alternative production platforms to manufacture useful compounds.

The largest class of secondary metabolites is terpenes, which until the 1970's were thought to be waste products from other metabolic processes (Gershenzon and Dudareva, 2007). Terpenes are compounds comprised of five-carbon isoprene units that can be found in all domains of life, including Archaea, Bacteria, and Eukarya (Chen et al., 2016). In Archaea, isoprene is an essential component of the phospholipid cellular membrane, though genome analysis has determined that Archaea possess no terpene synthases and do not convert isoprene into more complex compounds (Naparstek et al., 2012; Yamada et al., 2015). Animals produce a small number of highly complex terpenes, primarily steroids derived from squalene such as lanosterol and cholesterol (Rozman et al., 1996). Some insect species produce defensive secretions that contain a mixture of terpenes which deter predators (Gershenzon and Dudareva, 2007). Bacteria

produce a small range of terpenes. One notable example is the degraded sesquiterpene alcohol geosmin, which was first isolated in 1891. The volatile compound is emitted by soil-dwelling bacteria and is responsible for petrichor, the earthy smell of soil after rain (Cane and Ikeda, 2012). The terpene profiles of fungi and plants are significantly more diverse than those of any other kingdoms. Thousands of fungal terpenes have been identified as being involved in communication and signalling, and as essential for growth and development (Schmidt-Dannert, 2015). In plants, terpenes play active roles in attracting and repelling insects, tissue toxicity, hormonal regulation, and photosynthesis (Gershenzon and Dudareva, 2007). Over 40, 000 unique plant-based terpenes have been identified, though few have been examined in depth (Boutanaev et al., 2015). Since all terpenes are formed from similar processes, inferences can be made about the structure and function of even understudied members of this metabolite family.

Terpenes are classified based on the number of 5-carbon isoprene subunits in their structure (Nagegowda, 2010). These structural variations convey specific functions to each terpene type. The most basic types are isoprenes, composed of 5-carbon chains, monoterpenes (C_{10}), and sesquiterpenes (C_{15}) (Degenhardt et al., 2003). These terpenes are typically volatile compounds that give fruits and flowers their scent and provide indirect defense against herbivores. These compounds are usually released only when specific circumstances, such as insect feeding, trigger emission (Degenhardt et al., 2003). Another major class is the diterpenes (C_{20}) which include phytoalexins involved in herbivore defense and the precursors to many essential compounds such as gibberellins, phytol, and taxanes (Zerbe et al., 2013). Other terpene types include the triterpenes (C_{30}) which are the precursor to sterols, and the tetraterpenes (C_{40}) which form several photosynthetic pigments (Phillips et al., 2006). Terpenes of every type range in complexity from simple carbon chains to large ring formations, corresponding to their diverse bioactivities (Fig. 1).

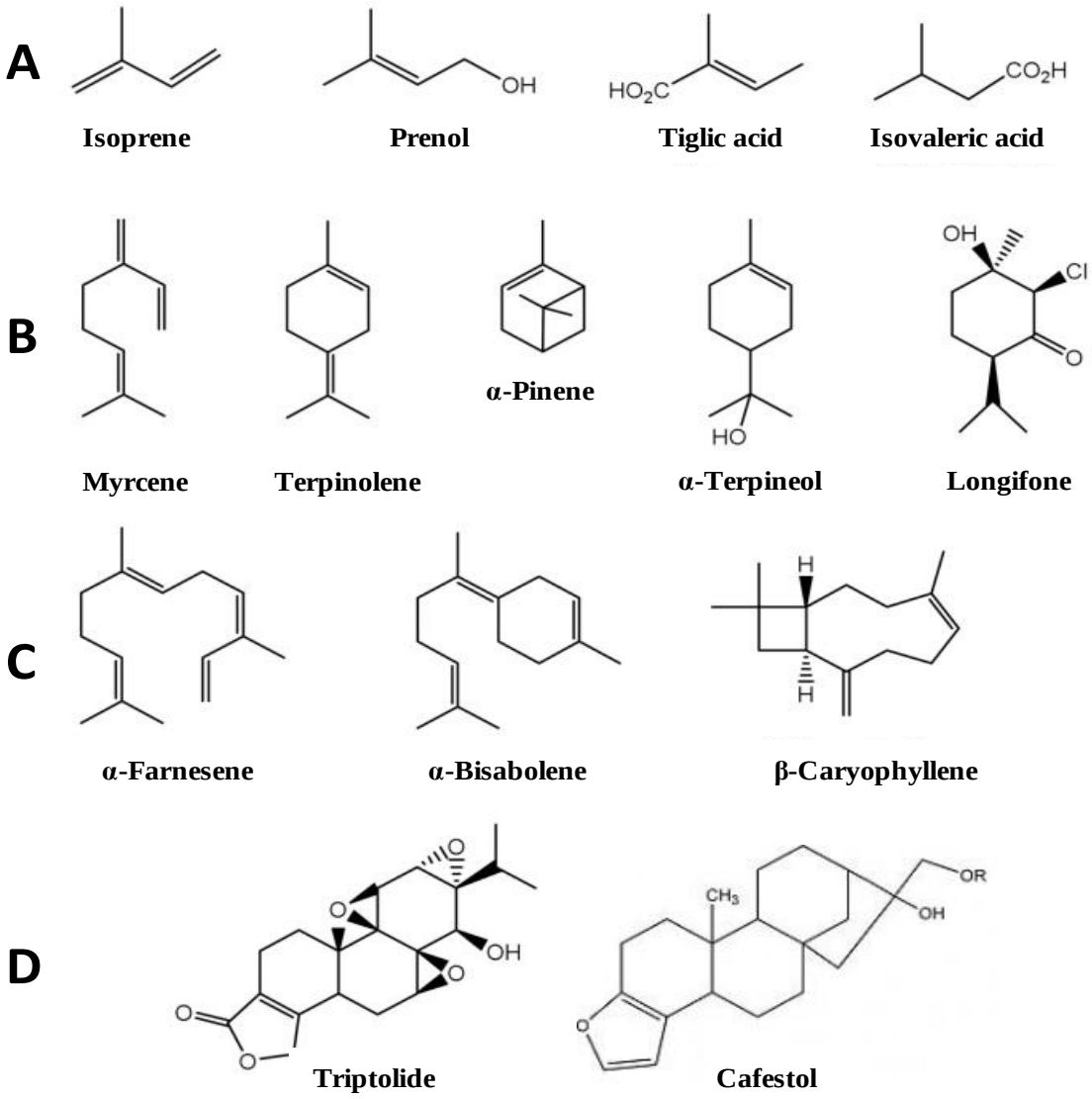


Figure 1: Representative terpene structures. *A) Isoprenes are the most basic C_5 chains used to construct more complex terpenes. B) Monoterpenes are C_{10} compounds composed of 2 isoprene subunits. C) Sesquiterpenes are C_{15} compounds composed of 3 isoprene subunits. D) Diterpenes are C_{20} compounds composed of 4 isoprene subunits. Figure adapted from Lima et al., 2016 (permission number 4340931284126) and Higdon and Frei, 2006.*

The terpene type also indicates how the structure was produced. Sesquiterpenes and triterpenes are produced by the mevalonate (MEV) pathway in the cytosol (Fig. 2). In this pathway, two molecules of isopentenyl diphosphate (IPP) and one of dimethylallyl diphosphate (DMAPP) are condensed into the terpene precursor farnesyl pyrophosphate (FPP). Monoterpenes, diterpenes, and tetraterpenes are formed by the

methylerthritol-4-phosphate pathway (MEP), also referred to as the mevalonate-independent pathway, which takes place in the plastid. This follows a similar condensation of one molecule of IPP and DMAPP into the precursors geranyl pyrophosphate (GPP), from which monoterpenes are formed, and three IPP to one DMAPP to create geranylgeranyl pyrophosphate (GGPP), from which diterpenes are formed (Degenhardt et al., 2003). FPP, GPP and GGPP are then acted upon by terpene synthases and other modifying enzymes to form a terpene, terpenoid or sterol product. There are thousands of terpene synthases which pair with a combination of modifying enzymes to produce unique structures from these three precursors. For example, the sesquiterpene artemisinin is formed in the cytosol from FPP by the activity of an initial amorpha-4,11-diene synthase, followed by augmentations by cytochromes P450 and dehydrogenases into a final functional product (Fig. 3).

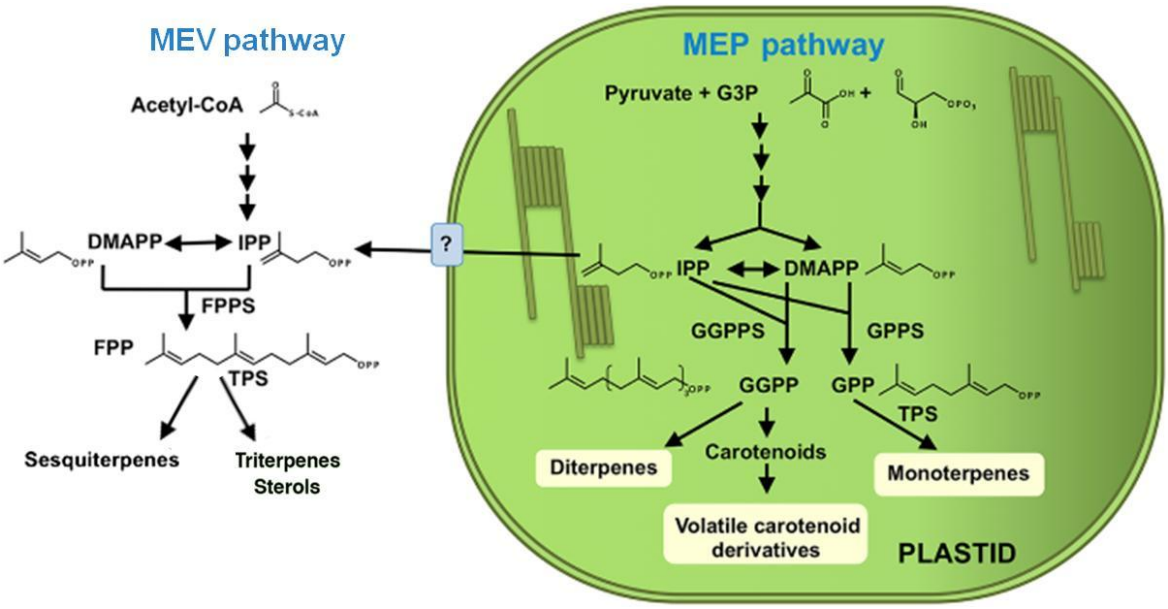


Figure 2: Terpene biosynthetic pathway. *The MEV and MEP pathways produce terpene precursors from DMAPP and IPP in either the plastid or the cytosol. These precursors are acted upon by terpene synthases (TPS) and other modifying enzymes to produce an active terpene product. Figure adapted from Muhlemann et al., 2014 (permission number 4340940542375).*

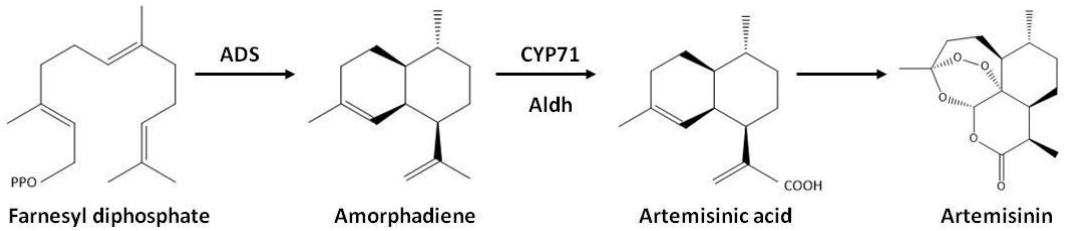


Figure 3: Simplified artemisinin biosynthetic pathway demonstrating the actions of a terpene synthase and modifying enzymes. *The FPP precursor is acted upon by a terpene synthase (amorpha-4,11-diene synthase, ADS) and then further modified by cytochromes P450 (CYP71) and aldehyde dehydrogenases (Aldh) and UV light into artemisinin. Figure adapted from Bohlmann and Keeling, 2008 (permission number 4340940705257).*

1.1.2 Terpenes for human use

Many terpenes have properties that are valuable for human use. Commercial applications of terpenes include use in food additives, medicine, fragrances, industrial solvents, insecticides, and alternative fuels (Augustin et al., 2015a; Brückner and Tissier, 2013). One of the most well known terpene-based products is turpentine, a solvent extracted from pine resin containing the monoterpenes pinene, carene, and myrcene among many others (Belgacem and Gandini, 2011). Turpentine is primarily associated with paint thinners and industrial solvents, but is also purified into its base monoterpenes for use in flavours, fragrances, and oils (Belgacem and Gandini, 2011). The monoterpene limonene is a common flavour and fragrance additive that is also used as an industrial solvent. Derived from the discarded rinds of citrus fruits, over 70, 000 tons of limonene are produced annually to provide an environmentally-friendly alternative to the harsh chemicals used in industrial cleaning (Ciriminna et al., 2014).

The pharmacological uses of terpenes are well established. The two most well-studied pharmaceutical terpenes are the diterpene taxol and sesquiterpene lactone artemisinin, which are used in the treatment of cancer and malaria, respectively (Sheludko, 2010). The bioactivity of complex terpene structures is correlated to their architecture. Taxol, which is extracted from the bark of *Taxus* species, inhibits cancerous tumor growth and

is an important component of chemotherapy treatments for breast, lung, and ovarian cancer (Fig. 4A) (Gershenzon and Dudareva, 2007; Juyal et al., 2014). Taxol binds to a specific region of β-tubulin and disrupts the microtubule construction and degradation cycle, preventing mitotic spindle formation and cell division (Kingston, 2007). The binding ability of taxol is related to the position of functional groups around its ring structure. Removal of functional groups heavily reduces cytotoxicity, while substitution of novel groups in place of native ones can increase tubulin binding efficiency (Kingston, 2007).

Artemisinin is derived from the annual herb *Artermisia annua*, and is an endoperoxide that can kill the malaria parasite *Plasmodium falciparum* during the early stages of its lifecycle (Fig. 4B) (Gershenzon and Dudareva, 2007). Artemisinin contains a peroxide bridge that reacts with hemin, a waste product of hemoglobin that has been digested by the parasite. This interaction produces free radicals which are thought to kill *P. falciparum* (Cheng et al., 2002). Though taxol and artemisinin are products of the same biosynthetic pathway, they have vastly different bioactivities and modes of action. The unique biosynthetic pathways by which terpenes are derived have led to a diversity of bioactive terpene structures.

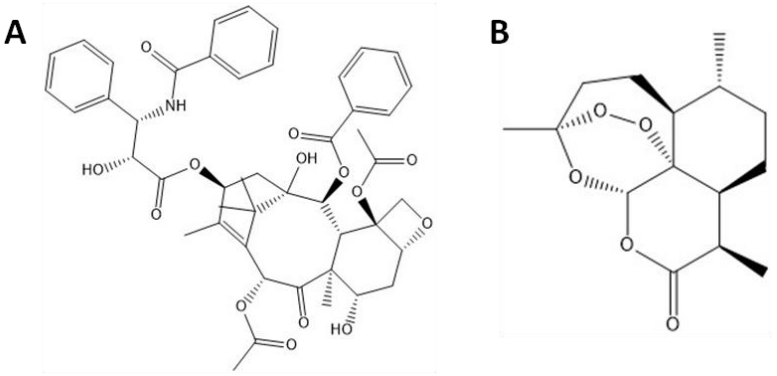


Figure 4: *A) Structure of taxol, a diterpene used in chemotherapy treatments. The positioning of oxygen and functional groups around the central ring structure coordinates binding of taxol to β-tubulin, affecting microtubule dynamics. B) Structure of artemisinin, a sesquiterpene valuable in the treatment of malaria. The peroxide bridge (O-O bond) interacts with by-products of the malaria parasite and generates damaging free radicals which inhibit parasite spread. Figure adapted from NCBI Resource Coordinators, 2017.*

While terpene bioactivities are useful in medicine, their structural features have led to their use as a renewable source of high-energy fuel additives. Terpene metabolites have been shown to have an energy content comparable to diesel fuel (Peralta-Yahya et al., 2011) and jet fuel (Chuck and Donnelly, 2014). Diesel fuel contains 16-carbon alkane chains that can be cyclical, linear, or branched. Sesquiterpenes (C_{15}) have a similar structure, composed of acyclic hydrocarbon chains or cyclized compounds that are energy dense (Fig. 5). Many terpenes have methyl branches which lower the freezing point and increase the stability of the fuel mixture under pressure, both benefits over diesel (George et al., 2015). However, heavily branched terpene structures often have lower combustion quality than what is required for diesel engines. The sesquiterpenes farnesane and bisabolane produce comparable combustion levels to diesel. While the mostly-linear farnesane has a lower energy content than cyclic bisabolane, it burns more efficiently due to its structure (George et al., 2015). A fuel source that contains a mixture of terpene structures could provide a balance between the beneficial features of chain and cyclic compounds.

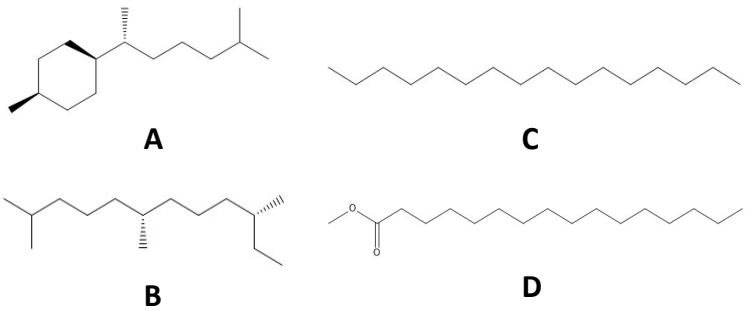


Figure 5: Comparison of fuel structures. *The chemical structure of the terpenes A) bisabolane and B) farnesane are similar to the diesel fuel component C) hexadecane and D) methyl palmitate, a component of plant-based oils. Figure adapted from Peralta-Yahya et al., 2011 and NCBI Resource Coordinators, 2017.*

Jet fuel requires a different energy profile than diesel and is graded based on energy content and function (George et al., 2015). To be suitable for aviation, an alternative fuel source must have high energy density, the ability to be ignited safely at both high altitudes and sea level, low viscosity and freezing point, and an economical production method (Chuck and Donnelly, 2014). Jet-A fuel has a mid-level energy content

comparable to limonene (Brennan et al., 2012). As limonene can be produced from waste rinds from the citrus industry, it is a sustainable source of a Jet-A fuel supplement (Ciriminna et al., 2014). Higher-energy fuels, such as JP-10, have strained ring structures that are energy dense and difficult to replicate (George et al., 2015). The monoterpenes α-pinene and β-pinene have similar energy contents to JP-10, but are difficult to sustainably produce from natural or synthetic sources at the volumes required for the fuel industry (Fig. 6) (George et al., 2015).

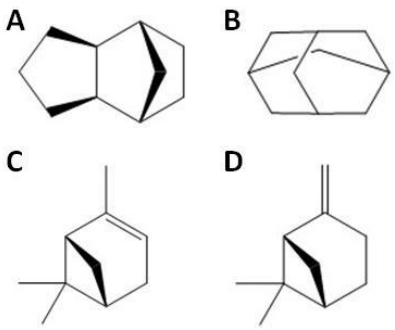


Figure 6: Structural comparison of JP-10 fuel components and pinene isomers*. JP-10 aviation fuel consists of strained ring structures such as A) exo-tetrahydrodicyclopentadiene and* B) *adamantane. The monoterpenes C) α-pinene and D) β-pinene have comparable structures and energy content. Figure adapted from Gao et al., 2015 (permission number 4340941292149) and NCBI Resource Coordinators, 2017.*

1.1.3 Current issues in terpene production

The availability of terpenes for use in pharmaceuticals and biofuels is often limited by the production system. Often, plants do not naturally produce enough of these secondary metabolites to make harvesting and extraction from plant tissues sustainable (Degenhardt et al., 2003). Collection of plant materials from nature can endanger the species and ecosystem, exemplified by the reduced range of the yew tree caused by the overharvesting of *Taxus* species for the production of taxol (Juyal et al., 2014). Alternative production methods must be developed to overcome the challenge of harvesting essential terpenes from endangered or low-yielding plants.

Engineering a biosynthetic pathway to produce more terpenes in the native species or transforming it into a species amenable to mass production could improve terpene availability. While the MEV pathway and several terpenes, including precursors to the pharmaceutical taxol, can be produced in microbial systems, commercially suitable

amounts of product often cannot be generated due to the toxicity of terpene intermediates (Ajikumar et al., 2010; Peralta-Yahya et al., 2011). For many terpenes, microbial production requires that the yield increase beyond the limit of toxicity for the cell culture in order to be profitable (Dunlop et al., 2011). There is currently no efficient commercial production system for key terpenes used as biofuel additives. Farnesane and bisabolane have been produced in microbial systems but in insufficient volumes (George et al., 2015; Peralta-Yahya et al., 2011). In order to compete with traditional fuels, a terpene production system must be developed that is cost effective and that meets the volume requirements of the fuel industry. Engineering a terpene biosynthetic pathway to increase production in the native plant species or an alternative production system could improve the economic viability of terpenes as a renewable fuel source.

1.2 Gene clustering in secondary metabolite production

1.2.1 History of gene clusters

A gene cluster is a portion of the genome that contains three or more non-homologous genes that function in the same biosynthetic pathway, often encoding sequential steps (Nützmann et al., 2016). Clusters have been identified in animals, fungi, bacteria, and plants. In humans, several essential proteins such as the β-globin subunit of haemoglobin and the human growth hormone somatomammotropin are produced by tightly regulated gene clusters (Chakravarti et al., 1984; Gusella et al., 1979). In fungi, plants, and bacteria, clusters are more often involved in non-essential metabolic pathways (Ballouz et al., 2010; Keller and Hohn, 1997). The first biosynthetic gene cluster was discovered in the bacterium *Streptomyces coelicolor* and was involved in production of the polyketide antibiotic actinorhodin (Malpartida and Hopwood, 1984). A wide array of fungal and bacterial metabolites have since been found to be produced by gene clusters (Ballouz et al., 2010; Brakhage and Schroeckh, 2011).

The physical clustering of genes involved in metabolite biosynthesis is seen in both monocots and dicots (Nützmann and Osbourn, 2014). The first described plant gene

cluster was found in *Zea mays* in the synthesis of 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA), an insect defence compound (Frey et al., 1997). Many clusters producing secondary metabolites have since been identified, several of them in monocot species. Clustering has been confirmed in the production of phytoalexins in *Oryza sativa* (Wilderman et al., 2004), avenacins in *Avena strigosa* (Qi et al., 2004), and the cyanogenic glucoside dhurrin in *Sorghum bicolor* (Darbani et al., 2016).

The structure and composition of gene clusters varies both between and within species. In plants, clusters have been found which contain 3-10 genes in varying arrangements (Nützmann and Osbourn, 2014). Clusters can be collinear, where the genes are positioned on the chromosome in the order in which they act, or randomly arrayed around a founder gene. The founder gene of a biosynthetic cluster encodes an enzyme that creates the initial skeleton structure of the metabolite, while the remaining genes are enzymes that modify the skeleton structure (Osbourn, 2010). These modifying enzymes can include, but are not limited to, cytochromes P450, reductases, methyl- and acyltransferases, sugar transferases, dioxygenases, carboxylesterases, transaminases, and polyketide synthases (Nützmann et al., 2016). Clusters can also contain transcription factors that regulate the expression of the gene cluster, or compounds that convey resistance to a toxic final product (Darbani et al., 2016; Osbourn, 2010).

The distance between individual genes in a cluster, and therefore overall cluster size, is highly variable. Clustered genes can be immediately adjacent to one another or separated by thousands of base pairs. The smallest described gene clusters, spanning 35 kb, are in *Arabidopsis thaliana* and encode the triterpenes thalianol and marneral (Field and Osbourn, 2008; Field et al., 2011) while the largest is the approximately 270 kb DIMBOA biosynthetic gene cluster (Dutartre et al., 2012). The variability of cluster size and composition complicates the identification of clustered genes. Rigorous testing of gene expression patterns and interactions between genes is required to confirm clustering.

It is unclear how gene clusters initially arose. Current hypotheses involve gene duplication events giving rise to chromosomal rearrangements and subfunctionalization of duplicated genes with clusters persisting due to novel advantages or protective adaptations (Boutanaev et al., 2015; Boycheva et al., 2014). There is also no definitive answer as to why gene clusters occur with such apparent frequency in secondary metabolite production. It is possible that the close proximity of the genes allows for more efficient co-regulation of expression, or that it increases the probability that all of the genes necessary for metabolite production will be inherited by the next generation (Osbourn, 2010). The intermediate compounds of a biosynthetic pathway can have negative effects when accumulated, and both the co-expression of all genes involved and the inheritance of a complete pathway reduces the likelihood that toxic intermediates will damage plant tissues (Nützmann et al., 2016). It is possible that biosynthetic gene clusters are widespread due to cluster movement via horizontal gene transfer. Genes involved in secondary metabolite production can be transferred from bacteria to fungi (Wenzl et al., 2005) and from fungi to plants (Richards et al., 2009), and complete gene clusters are capable of being transferred between different species of fungi and bacteria (Khaldi et al., 2008; Lawrence and Roth, 1996). It is possible that the horizontal gene transfer from bacteria to higher organisms is the origin of some plant clusters for secondary metabolite production, though this hypothesis has yet to be proven (Nützmann et al., 2016).

1.2.2 Gene clusters in terpene biosynthesis

Over 40,000 terpenes have been identified in plants but the majority of the genes involved in their biosynthesis are unknown (Boutanaev et al., 2015). Terpene biosynthesis is typically studied by identifying the products of a single terpene synthase (Zhuang et al., 2012). However, since terpene synthases can produce a skeleton structure that undergoes further modification, the product of a terpene synthase alone may not be a biologically active compound. In order to fully understand terpene biosynthesis, all enzymes in the pathway must be identified. The application of gene

cluster identification for terpene production can simplify the elucidation of terpene biosynthetic pathways.

Gene clustering in terpene biosynthesis has been reported in several bacteria, fungi, and plant species (Nützmann and Osbourn, 2014; Wilderman et al., 2004). Research on terpene production in bacteria is limited. Only a handful of terpene or terpenoid structures have been identified and few biosynthetic pathways characterized (Yamada et al., 2015). Phenalinolactone, a terpene glycoside, is one of the few terpenes known to be produced in *Streptomyces* species. It is encoded by a 35-gene cluster that contains all the biosynthetic and regulatory genes necessary for terpene production (Dürr et al., 2006). Other bacterial terpenes, such as the diterpenes terpenticin and brasilicardin A, are known to be produced in clusters and have partially characterized biosynthetic pathways that require further investigation to complete (Dairi et al., 2001; Hayashi et al., 2008). Fungi produce thousands of terpenes and gene clustering is widespread in fungal secondary metabolite biosynthetic pathways. Several fungal terpenes have fully elucidated biosynthetic pathways and are confirmed to be produced by gene clusters. The fungus *Cephalosporium aphidicola* synthesizes the diterpene aphidicolin from a 6-gene cluster composed of a terpene synthase, two cytochromes P450 and transcription factors (Toyomasu et al., 2004). Terpenoids such as terretonin, fumagillin, and austinol, among many others, are also produced by biosynthetic gene clusters of varying sizes and composition (Lin et al., 2013; Schmidt-Dannert, 2015). In plants, metabolic gene clusters involved in terpene biosynthesis typically contain the terpene synthase as the first committed step of the biosynthetic pathway, paired with a cytochrome P450 from the CYP71 family (Boutanaev et al., 2015). From these two classes of enzymes arise a huge diversity of terpene structures that can be further modified by additional enzymes.

Several gene clusters involved in terpene production have been identified in monocot species. In *Oryza sativa* the diterpenes momilactone A and B are produced from a gene cluster containing multiple terpene synthases that perform consecutive reactions in the momilactone biosynthetic pathway (Fig. 7) (Wilderman et al., 2004). *O. sativa* also

contains a gene cluster composed of 6 cytochromes P450 and four terpene synthases that produces both phytocassanes and oryzalides (Wilderman et al., 2004). In *Avena strigosa* the triterpene avenacin is produced from a cluster which contains the initial terpene synthase and the cyclase, oxidase, glycosyltransferase and acyltransferase required to convert the terpene precursor into an functional final product (Qi et al., 2004). These known clusters indicate the possible composition of terpene-producing clusters in other monocot species.

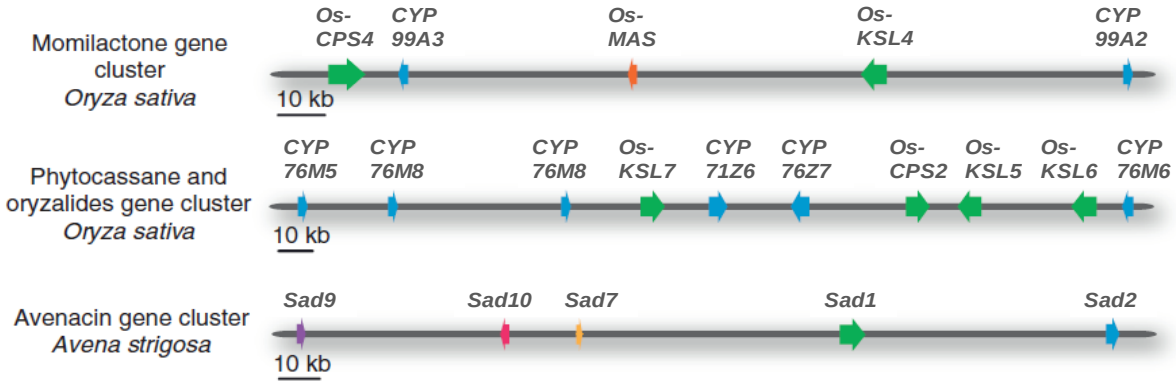


Figure 7: Terpene biosynthetic gene clusters in monocot species. O. sativa *has clusters producing the diterpene momilactones, phytocassanes, and oryzalides.* A. strigosa *has a cluster producing the triterpene avenacin. Green arrow = terpene synthase, blue = cytochrome P450, orange = dehydrogenase/reductase, red = glycosyltransferase, yellow = acyltransferase. Figure adapted from Nützmann and Osbourn, 2014 (permission number 4340941469816).*

1.2.3 Importance of cluster identification

Identification of gene clusters is essential for the discovery of new metabolic pathways and the engineering of these pathways for human use. Advances in sequencing technology and bioinformatic analysis have accelerated the identification of natural product pathways, but fewer than 50 biosynthetic pathways with genetic annotations have been completed in plants (Nützmann et al., 2016). Production of sustainable amounts of desirable metabolites requires a better understanding of how the metabolites are produced.

The discovery of gene clusters has simplified the process of identifying and characterizing the genes involved in biosynthetic pathways. It was previously thought that gene organization was mostly random, making it difficult to predict which genes are involved in related processes (Nützmann et al., 2016). Improvements in bioinformatic analytical techniques and subsequent identification of gene clusters has allowed for mining of large plant genomes for candidate biosynthetic clusters. The clusters can then be screened for their predicted functions and interactions and even engineered to produce more potent or higher volumes of product. Such investigations have been conducted on bacterially-produced compounds, such as the antitumor macrolides from actinobacteria. Mutation of a cytochrome P450 in the macrolide biosynthetic gene cluster resulted in a final product that was more cytotoxic than the native compound (Salcedo et al., 2016). Gene cluster identification can also be used to complete partially characterized biosynthetic pathways. The critical anti-cancer monoterpenes vinblastine and vincristine are partially derived from several small clusters of biosynthetic genes (Kellner et al., 2015). Investigation of these clusters revealed novel enzymes that are responsible for some of the missing steps of the biosynthetic pathway (Nützmann et al., 2016). A better understanding of gene clusters could lead to similar advances in plant systems for production of biofuels and industrial solvents.

1.3 Terpenes in *Sorghum bicolor*

1.3.1 S. bicolor *as a food and fuel*

Sorghum bicolor is a staple crop that is harvested worldwide for its grain, leaves, and sugar-rich stems, which are converted into feed, flour, sugar, syrup, alcohol, and biomass. It is predominately grown in semi-arid regions that cannot support corn or wheat (Zhuang et al., 2012). As rising temperatures increase the incidence of droughts worldwide, the C_4 photosynthetic pathway and drought tolerance of *S. bicolor* will become more advantageous.

The *S. bicolor* phenotype is highly varied and depends on the product the line has been bred to produce. Grain varieties have been bred to be significantly more compact than

sweet varieties, which are harvested for sugar or feedstock. The sweet varieties accumulate large amounts of biomass even under stress conditions such as drought or high salinity (Prasad et al., 2007). This genetic diversity, along with its small genome, available sequence data, hardiness, and variability in both photoperiod and biomass allocation makes *S. bicolor* an ideal crop for biofuel production (Turner et al., 2016).

For a crop to be an efficient source of renewable fuels, it must require low resource inputs and have a high fuel to biomass ratio (Fesenko and Edwards, 2014). *S. bicolor* is a low-input crop that is currently cultivated worldwide, with a framework for growth and harvest well established. *S. bicolor* has been documented as a source of ethanol but is not a competitive alternative to *Zea mays*. While *S. bicolor* requires lower inputs of resources such as fertilizer and water, current varieties produce as little as a third of the total ethanol per hectare extracted from maize (Elbehri et al., 2013; Wortmann et al., 2010). However, ethanol has only two-thirds the energy content of gasoline and, ideally, a renewable biofuel would have a high energy content comparable to traditional fossil fuels (Peralta-Yahya et al., 2011). Alternatives to ethanol are under investigation, including several terpene production systems (Augustin et al., 2015a; Peralta-Yahya et al., 2011). *S. bicolor,* which is already known to produce terpenes, could potentially become a production platform of these more energy-efficient biofuels.

1.3.2 Secondary metabolites and gene clustering in S. bicolor

Despite the genetic diversity and economic importance of *S. bicolor*, only limited information is available on the secondary metabolites produced by this crop. Extracts of plant tissues have been found to contain phenolic acids and flavonoids, such as cinnamic acid, anthocyanins, and tannins, all of which have antioxidant potential and other health benefits, as well as various phytosterols and policosanols (Awika and Rooney, 2004). However, the biosynthetic pathways and genetic regulation involved in the production of these metabolites remain largely unknown.

Like the phenotype, the chemical profile of *S. bicolor* varies depending on genotype and environmental conditions. In some cases the phytochemical variations are actively selected for, such as in South Africa where a variety with high grain tannin content is cultivated to decrease bird predation, despite reduced digestibility for livestock (Awika and Rooney, 2004). In most cases, phytochemical variation is a by-product of selection for agriculturally relevant traits. This variation can be exploited to find genotypes that produce high volumes of compounds of interest.

The release of the full *S. bicolor BTx623* genome has accelerated the characterization of unknown genes, some of which produce valuable or novel compounds (Paterson et al., 2009). Several of these compounds are produced by gene clusters, such as the immune response to *Setosphaeria turica*, the pathogen responsible for northern leaf blight. This activity is controlled by a cluster of 6 genes which encode a resistance protein that triggers host response upon pathogen detection (Martin et al., 2011). Additional *S. bicolor* gene clusters include ten repeated genes spanning 35 kb which produce the seed storage protein kafarin (Song et al., 2004) and a large cluster spanning 104 kb containing 5 genes encoding the cyanogenic glucoside dhurrin (Fig. 8) (Takos et al., 2011). The dhurrin gene cluster was initially thought to contain two cytochromes P450 and a UDP-glucosyltransferase, which act upon the precursor, tyrosine. The cluster was later found to include a glutathione *S*-transferase and a transporter from the multidrug and toxic compound extrusion (MATE) protein family which were located in the intervening region between genes of interest (Darbani et al., 2016). The MATE transporter functions in removing synthesized dhurrin to the vacuolar membrane, confirming that clusters in *S. bicolor* can contain non-biosynthetic genes.

The diversity of the known clusters in *S. bicolor* illustrates that even within a single species there is large variability in cluster size and composition. Currently no studies are available on the clustering of genes that produce terpenes in *S. bicolor,* but it is expected that they will be similar to those identified in other monocots.

Figure 8: Gene cluster producing the cyanogenic glucoside dhurrin*. The dhurrin cluster contains five genes, two cytochromes P450 (CYP71, CYP79), UDP-glucosyltransferase (UGT), glutathione* S*-transferase (GST) and a MATE transporter. Chromosomal position is indicated by grey genes. Figure adapted from Darbani et al., 2016 and Kristensen et al., 2005.*

1.3.3 Terpene synthesis in S. bicolor

In monocot crops, terpenes are often produced as an indirect defense in response to insect attacks. Both the model crops *Z. mays* and *O. sativa* emit a combination of terpenes that attract predators when tissues are damaged by herbivory (Yuan et al., 2008). While many volatiles have been identified, only a handful of terpene synthase genes have been characterized in these species. The enzymes which have been analyzed produce a diverse range of terpene and terpenoid structures in leaves, roots, stems, and reproductive plant organs (Zhuang et al., 2012). Analysis of *S. bicolor* leaf volatiles shows that sorghum produces a similar chemical profile to maize and rice. Terpenoids produced when *S. bicolor* tissues are damaged by insects are predominately β-caryophyllene, α-bergamotene, and β-farnesene (Zhuang et al., 2012). Most of these leaf volatiles are produced by five sesquiterpene synthases, which are currently the only confirmed terpene biosynthetic genes in *S. bicolor* (Zhuang et al., 2012).

Over 47 potential sesquiterpene synthases have been identified in *S. bicolor,* but the majority has not been investigated (Zhuang et al., 2012). In addition, there has been little investigation into how the physical clustering of biosynthetic genes influences terpene production in this species. This lack of information illustrates the need for a protocol that identifies potential terpene synthases and modifying genes and that confirms both clustering and gene activity. The elucidation of a complete clustered pathway for terpene production in *S. bicolor* would accelerate gene discovery by confirming the presence of terpene biosynthetic gene clusters and their role in

metabolism. A better understanding of metabolite synthesis would allow for augmentation of the pathway through genetic engineering to increase innate or alternative host terpene production for human use.

1.4 Hypothesis and Objectives

An ongoing issue in secondary metabolite analysis is elucidation of complete biosynthetic pathways. Thorough understanding of a pathway is necessary for engineering sustainable production of metabolites for human use. Applying current knowledge of gene clustering to terpene biosynthetic genes in *S. bicolor* could simplify pathway identification. This thesis project addresses the following goals: (1) identification of potential terpene biosynthetic gene clusters in *S. bicolor* based on sequence similarity to terpene synthases and associated enzymes in other species, (2) characterization of the terpene synthase present in the selected gene cluster, and (3) determination if clustered genes act upon the product of the terpene synthase. It was hypothesized that terpene biosynthetic gene clusters could be identified based on the spatial distribution of potentially related genes, and that clustering could be confirmed by gene expression and characterization in a heterologous system.

Chapter 2: Identification of putative terpene biosynthetic gene clusters

2.1 Summary

Gene clustering of secondary metabolite biosynthetic pathways is a common phenomenon in plants. Identification of clusters has improved as more complete genomes have been released. Previous studies have shown that some terpenes, a class of secondary metabolite explored as renewable fuel additives, industrial solvents, and pharmaceuticals, are produced by gene clusters. *Sorghum bicolor* produces a variety of terpenes but information on their biosynthesis and potential application is limited. The metabolic profile of *S. bicolor* and its suitability as a source of alternative fuel makes it an ideal crop for the study of gene clustering in terpene biosynthetic pathways. Herein, putative gene clusters involved in terpene synthesis are identified and a single cluster containing a terpene synthase, cytochrome P450, and a galacturonosyltransferase is selected for further study.

2.2 Significance

Few terpene synthase genes in *S. bicolor* have been characterized or investigated for linkages to clustered genes. In this study, a novel method of identifying putative terpene biosynthetic gene clusters was applied to the *S. bicolor* genome and many potential clusters were successfully located.

2.3 Contributions

The Kellogg lab curated the database of known plant terpene synthases, cyclotides, cysteins, glycosyltransferases, methyltransferases, polyketide synthases, reductases and cytochromes P450 that was used in this project. Dr. Michael McKain developed the novel cluster identification method used to identify putative terpene biosynthetic clusters. Collaborative analysis of putative clusters between the Kutchan and Kellogg labs advanced three clusters for further characterization.

2.4 Introduction

Characterization of biosynthetic pathways is hindered by the difficulty of identifying related biosynthetic genes. The sequencing of complete plant genomes and the

development of bioinformatic analytical techniques have vastly accelerated the gene identification process. Currently, complete genomes from nearly 100 plant species are available (Kautsar et al., 2017). Sequence analysis of these genomes can identify potential gene clusters and elucidate unknown genes involved in critical biosynthetic pathways.

When analyzing a genome for novel biosynthetic genes, mass searches featuring local or global alignment tools can quickly identify genomic regions that are similar to known biosynthetic genes. A commonly-used local alignment software is BLAST (Basic Local Alignment Search Tool) which is used to compare a query sequence to a database and identify regions of characterized sequences that are statistically similar to the query (Altschul et al., 1990). A global alignment, or multiple sequence alignment, does not involve a database and requires that sequences be manually input for comparison (Pearson, 2013). Tools such as MAFFT or MUSCLE align the entire length of the query sequences to each other to determine total sequence similarity (Edgar, 2004; Katoh and Standley, 2013). Multiple sequence alignment results are most accurate when applied to sequences that are of similar length and suspected to share a function. Multiple sequence alignments are more time consuming than local alignments and require that all sequences of interest be identified prior to comparison; as such, they are often used as a second alignment tool following a local alignment (Pearson, 2013).

Bioinformatic analysis has been used to identify novel biosynthetic genes in several plant species, such as triterpene saponin synthases in *Medicago truncatula* (Achnine et al., 2005), and a valerenadiene synthase in *Valeriana officinalis*, which is the sesquiterpene responsible for the sedative properties of valerian root (Yeo et al., 2013). Once a putative biosynthetic gene is identified by an alignment tool, it must be characterized to confirm function. Due to the massive amounts of genetic information available, many putative biosynthetic genes remain uncharacterized even after identification by sequence analysis (Zhuang et al., 2012).

Gene clusters can also be identified by sequence analysis using genomic data. Alignment tools can be used to identify multiple types of biosynthetic genes and determine their proximity to one another on the chromosome. Genes that are involved in similar pathways and localized to the same region of the chromosome form putative gene clusters, which can then be tested for similar patterns of expression and gene interactions. Dozens of gene clusters thought to be involved in terpene biosynthesis have been identified in over 40 plant species (Kautsar et al., 2017). Analysis of terpene biosynthetic gene clusters has led to a better understanding of how elusive compounds such as the anticancer agents vinblastine and vincristine are synthesized (Kellner et al., 2015). Research has primarily focussed on terpene biosynthetic clusters that produce antibiotics and antiviral agents or other compounds with medicinal benefits (Dürr et al., 2006; Toyomasu et al., 2004). The carbon density and ring structure of terpenes also show promise as an alternative fuel source, though information on the effect of gene clustering on the biosynthesis of these types of terpenes is limited. Understanding the prevalence of gene clustering in the production of terpenes that have potential as biofuels is an essential step in mass-producing these high-energy compounds.

In order to investigate gene clustering in terpene biosynthesis, a putative cluster containing a terpene synthase must first be identified and confirmed to produce a functional terpene. The terpene synthase gene, as well as the type of terpene it produces, can be identified by sequence similarities between terpene synthases in the plant kingdom. Terpene synthases are typically between 550 and 850 amino acids in length, contain 6-12 introns, and have a characteristic terpene synthase fold and DDXXD metal ion binding motif (Alquézar et al., 2017; Zhou and Peters, 2009). In plants, terpene type is indicated by the presence of a subcellular localization signal. Monoterpene and diterpene synthases are 50-70 amino acids longer than sesquiterpene synthases due to the N-terminus transit peptide which locates these proteins to the chloroplast (Nagegowda, 2010). These conserved regions allow for identification of putative terpene synthases by sequence analysis. The presence of these regions does not guarantee that

the protein is functional, nor does it necessarily guarantee the type, so further testing is required to confirm terpene synthase activity.

When localized to a gene cluster, the terpene synthase is considered the founding gene. It converts the precursor molecule FPP, GPP, or GGPP into a terpene skeleton structure, which can be further modified by other enzymes, such as cytochromes P450, reductases, methyltransferases, and glycosyltransferases, into a final terpene product (Degenhardt et al., 2003). The identification of these modifying genes in close proximity to a terpene synthase indicates a putative gene cluster. Gene function, expression and interaction with other members of the cluster must be analyzed before cluster identity is confirmed.

2.5 Experimental Procedures

2.5.1 Cluster identification

The annotated cDNA and protein sequences of *S. bicolor* (v3.1) were obtained from the Joint Genome Institute (Nordberg et al., 2014). A database of nucleotide sequences from plant-based terpene synthases, cyclotides, cystine knots, glycosyltransferases, methyltransferases, polyketide synthases, reductases, and cytochromes P450 was developed in house by Dr. Michael McKain at the Donald Danforth Plant Science Center (St. Louis, MO, USA). A BLAST search (BLASTX) comparing the *S. bicolor* protein sequence to the plant-based enzyme database identified *S. bicolor* proteins similar to those of the database with an e-value of 10^{-8} (Altschul et al., 1990). The resulting *S. bicolor* sequences were filtered based on their bidirectional overlap with the targeted gene families. A minimum of 85% overlap was used to accept a putative *S. bicolor* member of a gene family. As an additional filtering step to confirm identity, *S. bicolor* genes were aligned to the initial database of plant-based enzymes using the MAFFT multiple sequence alignment tool (Katoh and Standley, 2013).

The clustering of identified gene families was determined using a novel Perl script (2.8 Supplementary Information) that searched for targeted gene families within a specific distance from each other. Clusters consisting of at least one terpene synthase and one

other target gene family, and having a range of 0-20 intervening genes between targeted genes, were reserved while the others were discarded. The clusters were manually curated based on size, the proximity of targeted genes, the number of gene families contained, and gene family composition. Putative clusters larger than the current largest documented plant biosynthetic gene cluster, 270 kb, were discarded (Frey et al., 1997). Plant biosynthetic gene clusters are considered to contain three or more genes, so clusters containing only a gene pair were also removed from consideration (Nützmann et al., 2016). Terpene biosynthetic gene clusters typically contain a terpene synthase/cytochrome P450 pair and so putative clusters lacking a cytochrome P450 were not advanced (Boutanaev et al., 2015). The final filtering step was selecting the clusters with the fewest total number of genes to facilitate analysis. From these criteria, three clusters were selected for further analysis. A BLASTX search was used to determine the putative function of all genes contained in the cluster and the intervening regions.

2.5.2 Determination of terpene synthase type in candidate clusters

The types of terpenes produced by the selected terpene synthases were determined using their phylogenetic relationship to known terpene synthases. A database of protein sequences from 533 known mono-, di-, tri- and sesquiterpene synthases and FPP, GPP and GGPP synthases was collected from the National Center for Biotechnology Information (NCBI Resource Coordinators, 2017). The protein sequence of each candidate terpene synthase was aligned to this database using MAFFT alignment software (Katoh and Standley, 2013). Trees were constructed in Mega 7 using maximum likelihood analysis with 500 bootstrap replicates (Kumar et al., 2016). The protein sequence of each terpene synthase was also analyzed using SignalP and TargetP to identify potential subcellular localization signals (Emanuelsson et al., 2000; Petersen et al., 2011).

2.5.3 Modification of gene cluster based on expression data

Expression data of RNA transcripts is available for the majority of *S. bicolor* genes (Makita et al., 2015). As Cluster 1 was accepted for advancement, the expression patterns of all genes identified as part of the cluster as well as the intervening genes were compared. Genes that were not expressed in the same tissues as the founding terpene synthase gene of Cluster 1 were removed from further analysis.

2.6 Results

2.6.1 Cluster identification

The initial analysis yielded 315 unique putative terpene biosynthetic clusters, each containing at least one terpene synthase. Only 23 of these putative clusters contained 3 or more gene families (Table S2.1). These 23 clusters were further reduced to 3 candidates for further analysis based on cluster size and the total number of intervening genes (Table 1). The candidate clusters spanned between 104 kb and 156 kb, and contained no more than seven intervening genes. Cluster 1 contained three gene families of interest - a reductase, cytochrome P450, and a terpene synthase most similar to the sesquiterpene α-copaene synthase found in *Oryza brachyantha*. The terpene synthase, Sobic.001G339000, was previously identified as a putative sesquiterpene synthase, though gene function was not confirmed (Zhuang et al., 2012). Cluster 2 contained two cytochromes P450, a methyltransferase, and a putative terpene synthase that was annotated as a farnesyl pyrophosphate synthase, the precursor to sesquiterpenes. Cluster 3 contained the terpene synthase identified as Sobic.007G035800, which was characterized by Zhuang *et al*., 2012 as a sesquiterpene synthase, as well as a cytochrome P450 and a reductase.

Table 1: Candidate Terpene Biosynthetic Gene Clusters. *Three putative gene clusters were selected based on size and gene family composition. Genes with a confirmed Gene Family are members of the cluster. All other functions are based on BLAST searches. The putative terpene synthase of each cluster is indicated in bold.*

Cluster 1: 137.3 kb		
Gene ID	Gene Family	Putative Function
Sobic.001G338000	Reductase	Similar to dehydrogenase/reductase [*Dichanthelium oligosanthes*]
Sobic.001G338100		Similar to hypothetical protein [*Dichanthelium oligosanthes]*
Sobic.001G338200		Similar to hypothetical protein [*Oryza sativa*]
Sobic.001G338400		Similar to galacturonosyltransferase [*Setaria italica*]
Sobic.001G338500		Similar to ACT domain-containing protein [*Setaria italica*]
Sobic.001G338600		Similar to hypothetical protein [*Zea mays*]
Sobic.001G338700		Similar to transcription factor CBF/NF-YB/HAP3 [*Triticum aestivum*]
Sobic.001G338800		Similar to myosin-2 heavy chain [*Zea mays*]
Sobic.001G338900	Cytochrome P450	Similar to cytochrome P450 [*Zea mays*]
Sobic.001G339000	**Terpene Synthase**	**Similar to α-copaene synthase [*Oryza brachyantha*]**
Cluster 2: 104.3 kb		
Gene ID	Gene Family	Putative Function
Sobic.003G269500	Cytochrome P450	Similar to cytochrome P450 [*Dichanthelium oligosanthes]*
Sobic.003G269600	Cytochrome P450	Similar to cytochrome P450 [*Setaria italica*]
Sobic.003G269700	Methyltransferase	Similar to carboxymethyltransferase [*Dichanthelium oligosanthes*]
Sobic.003G269800		Similar to phosphatidylcholine transfer protein [*Zea mays*]
Sobic.003G269900		Similar to subtilisin-like protease [*Setaria italica*]
Sobic.003G270000		No putative function
Sobic.003G270100		Similar to chaperone protein dnaJ 10 [*Zea mays*]
Sobic.003G270200		Similar to dehydrin Rab25 [*Setaria italica*]
Sobic.003G270300		Similar to DNA binding protein Hv33 [*Setaria italica*]
Sobic.003G270400		Similar to ubiquitin-protein ligase [*Setaria italica*]
Sobic.003G270500	**Terpene Synthase**	**Similar to farnesyl phosphate synthase [*Setaria italica*]**
Cluster 3: 155.9 kb		
Gene ID	Gene Family	Putative Function
Sobic.007G034900	Cytochrome P450	Similar to indole-2-monozygenase [*Zea mays*]
Sobic.007G035000		Similar to aromatic-L-amino-acid decarboxylase [*Triticum urartu*]
Sobic.007G035100		Similar to cyclin protein [*Zea mays*]
Sobic.007G035200		Similar to cobyrinic acid a,c-diamide synthase [*Aquincola tertiaricarbonis*]
Sobic.007G035300		Similar to aromatic-L-amino-acid decarboxylase [*Setaria italica*]
Sobic.007G035500		Similar to aromatic-L-amino-acid decarboxylase [*Zea mays*]
Sobic.007G035600		Similar to vesicle-associated protein 4-2 [*Setaria italica*]
Sobic.007G035700	Reductase	Similar to NAD(P)H-ubiquinone oxidoreductase B2 [*Setaria italica*]
Sobic.007G035800	**Terpene Synthase**	**α-bergamotene and γ-bisabolene synthase [*Sorghum bicolor*]**

2.6.2 Determination of terpene type from candidate gene clusters

Phylogenetic analysis of the potential terpene synthases confirmed the BLAST results detailed in Table 1. Each candidate terpene synthase clustered strongly with various sesquiterpene synthases, or, in the case of Cluster 2, farnesyl pyrophosphate synthases (Fig. 9). Sobic.001G339000 clustered among α-copaene synthase-like proteins from *Oryza brachyantha*, as predicted by the BLAST search, and the sesquiterpene aristolochene and vetispiradiene synthases from *Triticum* and *Oryza* species. The terpene synthase from Cluster 2, Sobic.003G270500, is most closely related to the farnesyl pyrophosphate synthases from *Triticum, Aegilops, Oryza,* and *Zea* species. Sobic.007G035800 clustered most closely with β-sesquiphellandrene and β-farnesene synthases from *Setaria italica*, a zizaene synthase, and a (*S*)-β-macrocarpene synthase from *Zea mays*. The position of Sobic.007G035800, previously characterized as a sesquiterpene synthase that produces β-farnesene, among other sesquiterpene synthases supports the effectiveness of this phylogenetic method in determining terpene type. These results indicate that the candidate genes are sesquiterpene synthases or enzymes that form the precursor farnesyl pyrophosphate and are therefore localized to the cytosol. No subcellular localization signal is present in this type of terpene synthase. The absence of a subcellular localization signal was supported by analysis using both SignalP and TargetP software (Emanuelsson et al., 2000; Petersen et al., 2011). As the terpene synthase from Cluster 3 has been previously characterized, and Cluster 2 encodes a farnesyl pyrophosphate synthase rather than a true terpene synthase, Cluster 1 is the best candidate for further analysis.

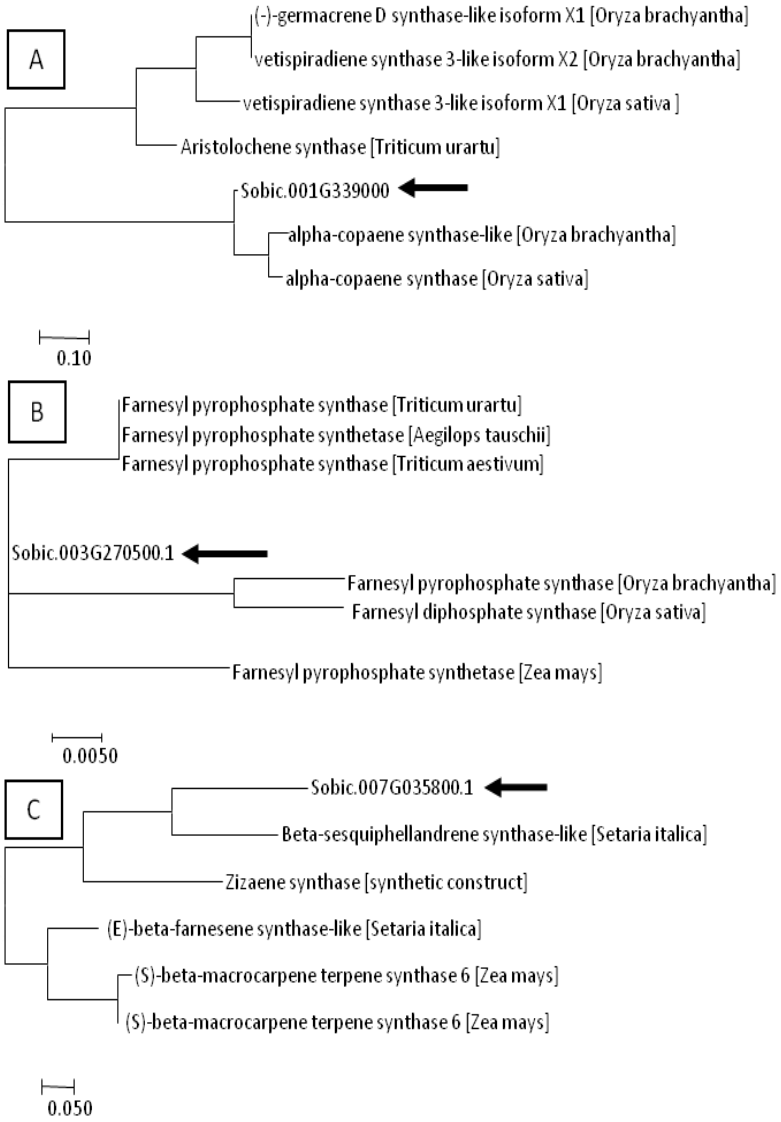


Figure 9: Phylogenetic analysis of candidate terpene synthases. *Each candidate terpene synthase (indicated by arrow) was positioned among known sesquiterpene synthases or FPP synthases, indicating that the proteins are active in the cytosol. A) Sobic.001G339000 is most similar to the α-copaene synthase found in* Oryza *sp. B) Sobic.003G270500 is found among farnesyl pyrophosphate synthases of monocots. C) Sobic.007G035800 is most similar to β-sesquiphellandrene synthase and is also known to produce β-farnesene.*

2.6.3 Modification of gene cluster based on expression data

Expression analysis of Cluster 1 determined that the terpene synthase, the founding gene of the cluster, was most highly expressed in the root tissues (Fig. 10). The cytochrome P450 of interest exhibited a similar expression pattern. However, the reductase that was identified as a potential cluster member was expressed in shoot tissues rather than the roots. As such, it was removed from further investigations of the cluster. The expression pattern of intervening genes was also examined. Sobic.001G338400, a putative galacturonosyltransferase, was expressed in all examined

tissues with high expression in roots. Sobic.001G338400 was hereafter included in further cluster analysis.

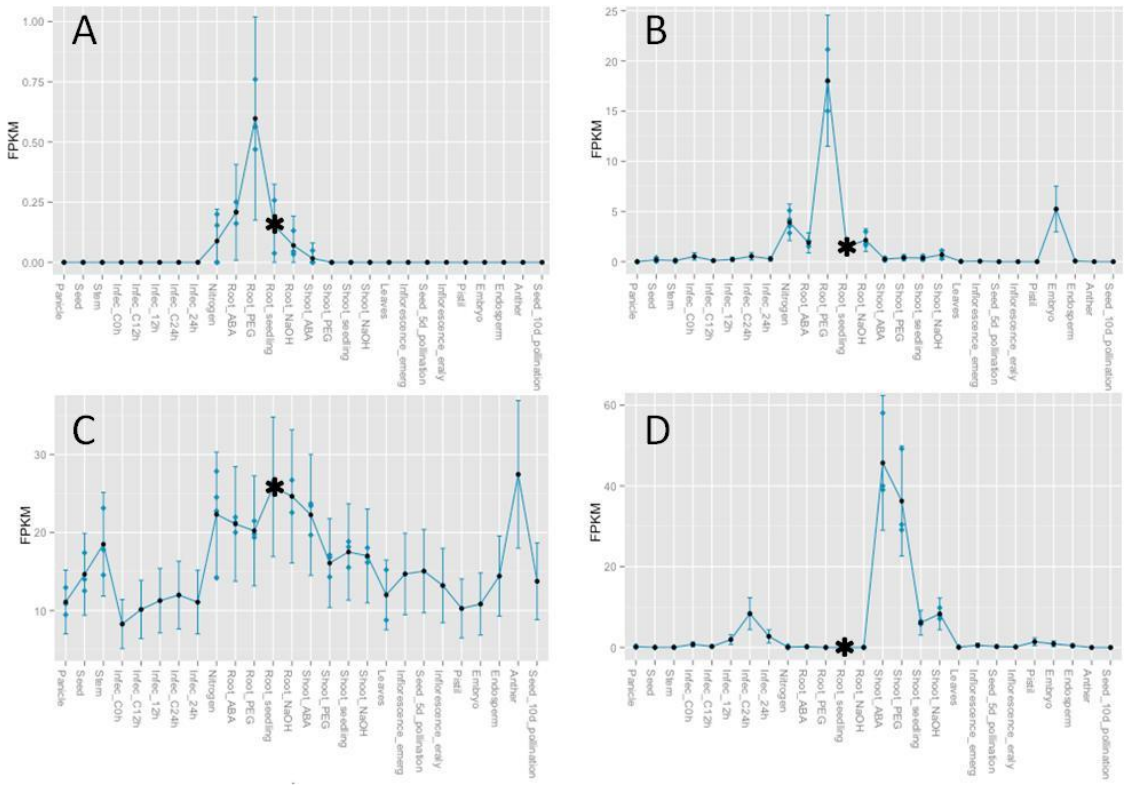


Figure 10: Expression patterns of the genes of interest. *RNA expression patterns for the S. bicolor genes A) Sobic.001G339000 (terpene synthase) B) Sobic.001G338900 (cytochrome P450) C) Sobic.001G338400 (galacturonosyltransferase) D) Sobic.001G338000 (reductase). In untreated conditions, TPS expression (A) is highest in seedling roots; this expression point is indicated with an asterisk. The CYP450 (B) and galacturonosyltransferase (C) are also expressed in untreated roots. The reductase (D) is not expressed in root tissues. Figure adapted from Makita et al., 2015.*

2.7 Discussion

The novel method of cluster identification used in this study yielded 315 potential terpene biosynthetic gene clusters in *S. bicolor*. Many of these putative clusters contained only two genes, which is smaller than most clusters described in plants. Therefore, while the paired genes may be linked, they were not suitable for this study and were not included in further analysis. Here, in addition, putative clusters were required to contain three or more gene families, since most documented terpene

biosynthetic gene clusters contain multiple enzymes types (Qi et al., 2004; Wilderman et al., 2004). Twenty-three putative clusters containing three or more gene families were identified. Previous research identified a total of 54 putative biosynthetic gene clusters in *S. bicolo*r (Kautsar et al., 2017), though these had different compositional requirements than those in this study. It is likely only a fraction of the 23 identified here are true gene clusters.

Further analysis of cluster size and composition reduced the pool of 23 potential gene clusters to three candidates for biochemical analysis. The majority of the 23 putative clusters contained a terpene synthase paired with a cytochrome P450, which is the gene pair most commonly found in terpene biosynthetic clusters (Boutanaev et al., 2015). Eight of the putative clusters did not contain a cytochrome P450 and were not considered for advancement, as it is unlikely for a terpene biosynthetic cluster to lack this gene based upon current published data. The remaining enzyme types found in clusters were, in order of decreasing frequency, reductases, polyketide synthases, glycosyltransferases, and methyltransferases. No cyclotides or cystine knots were found in any cluster. Of the fifteen putative clusters containing the terpene synthase/cytochrome P450 pair, only three of these contained fewer than 12 total genes and spanned less than 270 kb. Therefore, these three clusters were selected for phylogenetic analysis.

Cluster 1 contained the terpene synthase Sobic.001G339000 (formerly annotated as Sb01g032610), which is most closely related to the putative α-copaene synthase identified in *Oryza sativa*. Copaene is a tricyclic sesquiterpene, indicating that Sobic.001G339000 is likely a sesquiterpene synthase as well. Previous sequence analysis results also indicate that Sobic.001G339000 is a sesquiterpene synthase, though no attempt has been made to characterize this gene (Priya et al., 2018; Zhuang et al., 2012). Cluster 1 contained two more gene families of interest, a reductase and a cytochrome P450, both homologous to enzymes in *Zea mays*. The intervening region of the cluster also contained several promising genes that were not specifically targeted in

the cluster search. A putative galacturonosyltransferase was identified in this region and, based on expression data, could potentially be involved in terpene biosynthesis.

Cluster 2 contained the putative terpene synthase Sobic.003G270500. Both a BLAST search and phylogenetic analysis indicated that Sobic.003G270500 is not a terpene synthase, but more likely forms the terpene precursor farnesyl pyrophosphate. Cluster 2 contains three other genes of interest, two cytochromes P450 and a methyltransferase, that may be involved in FPP synthesis. As this cluster does not form a specific terpene product, it is not ideal for this study. Examination of this putative cluster is better-suited for a study on gene clustering in the MEV biosynthetic pathway.

Cluster 3 contains the terpene synthase Sobic.007G035800 (Sb07g003080) which has previously been characterized (Zhuang et al., 2012). It produces seven sesquiterpenes, including 7-*epi*-sesquithujene, (*E*)-α-bergamotene, sesquisabinene A, (*E*)-β-farnesene, β-bisabolene, (*Z*)-γ-bisabolene, and (*E*)-γ-bisabolene (Zhuang et al., 2012). This terpene synthase was advanced for further analysis to verify the effectiveness of our methods of cluster identification. The detection of putative terpene synthases and associated gene families in *S. bicolor* is validated by the identification of Sobic.007G035800 using this protocol. Sobic.007G035800 was also used to validate our method of terpene categorization. The phylogenetic tree used to ascertain terpene type correctly placed Sobic.007G035800 among sesquiterpene synthases, including a (*E*)-β-farnesene synthase and β-sesquiphellandrene synthase, a compound structurally similar to sesquisabinene. This validates our current terpene identification method and suggests that the placement of the other two terpene synthases, Sobic.001G339000 and Sobic.003G270500, is accurate.

In addition, cluster 3 contained two other genes of interest, a cytochrome P450 and a reductase homologous to enzymes in *Zea mays* and *Setaria italica*. These genes potentially act upon the product of Sobic.007G035800 to form novel sesquiterpenes in addition to the seven the terpene synthase alone is known to produce. However, since

Sobic.007G035800 has been previously characterized, cluster 3 was not an ideal candidate for this study.

All of the investigated terpene synthases were involved in the biosynthesis of sesquiterpenes, and were therefore active in the cytosol. No subcellular localization signal is present. Of the three clusters selected, cluster 1 containing the putative terpene synthase Sobic.001G339000 was best suited for further study. Sobic.003G270500 is likely involved in synthesis of a terpene precursor rather than a terpene, and investigation of gene clustering in the MEV pathway was beyond the scope of this study. Cluster 3 contained the terpene synthase Sobic.007G035800 which had already been thoroughly studied (Zhuang et al., 2012). Cluster 1 was the best candidate for advancement to both characterize a novel terpene synthase and validate our cluster identification methods.

The founding member of Cluster 1, the terpene synthase, is most highly expressed in root tissues. While most of the well-studied terpenes are involved in herbivore defense or pollinator attraction and are produced in above ground plant tissues, several terpenes are produced in roots. For instance, (E)-β-caryophyllene is produced in maize roots in response to insect attack (Rasmann et al., 2005). Other, non-volatile, terpenes are released by roots to assist development of hyphae by soil mycorrhizal fungi, creating a symbiotic relationship which improves nutrient acquisition by plant roots (Akiyama et al., 2005). Root-specific terpene synthases have also been found in *Arabidopsis,* producing the monoterpene 1,8-cineole and the sesquiterpene (Z)-γ-bisabolene (Chen et al., 2004; Ro et al., 2006), as well as *Orzya sativa,* which produces momilactones to suppress the growth of neighbouring plants (Wilderman et al., 2004).

Expression analysis of Cluster 1 shifted cluster composition from what was predicted based on genomic data. As the founding terpene synthase is highly expressed in roots, all other members of the gene cluster are expected to have a similar expression pattern. The putative CYP450 located in cluster 1 has a similar expression profile to the terpene synthase, with peak expression localized to the roots. This CYP450 also has some

expression in embryonic tissues, a trait not shared with the terpene synthase. Sobic.001G338000, a putative reductase identified as a cluster member by our sequence analysis, is expressed in shoots rather roots. It is unlikely that the reductase is a member of the gene cluster and it should not be included in further analysis. One of the intervening genes in the cluster, Sobic.001G338400, is a putative galacturonosyltransferase that is highly expressed throughout *S. bicolor*, including in roots. As such, the galacturonosyltransferase is a candidate member of the gene cluster and should be included in further analysis.

2.8 Supplementary Information

2.8.1 Supporting Resources

The script used to identify putative gene clusters was developed by Dr. Michael McKain and can be found at: https://github.com/mrmckain/Secondary_Metabolite_Clustering

2.8.2 Supplementary Tables

Table S2.1: Summary of putative terpene biosynthetic gene clusters. *Experimental ID indicates which clusters were advanced. Cluster selection was based on cluster size, number of intervening genes, and gene family composition. Gene family acronyms: TPS = terpene synthase, CYP450 = cytochrome P450, Red = reductase, GT = glycosyltransferase, MT = methyltransferase, PKS = polyketide synthase.*

					Gene Families					
Experimental ID	**Cluster ID**	**Cluster Size (kb)**	**Total Genes**	**No. Intervening Genes**	**TPS**	**CYP450**	**Red**	**GT**	**MT**	**PKS**
Cluster 1	69	137.3	10	7	1	1	1			
Cluster 2	263	104.3	11	7	1	2			1	
Cluster 3	517	155.9	9	6	1	1	1			
	429	134.3	23	8	1		4			3
	43	918.7	22	10	1		1		1	
	63	314.1	20	10	1	1	1			
	371	127.2	21	10	2		1		1	
	162	187.1	15	11	1			1		1
	361	220.2	50	11	2	1	1		1	3
	264	433.3	36	12	3			1	2	
	53	337.5	24	13	1	1	2			
	145	286.4	30	14	1		1	1		1
	339	562.1	25	14	1	3	1			1
	434	709.7	22	14	1	1		1		
	211	389.2	27	15	1	2		1	1	
	326	112.2	54	15	2	5		2		1
	463	446.0	23	15	1	1				3
	132	318.1	31	16	1		3			1
	135	544.8	58	16	1	4	1	1		1
	55	495.0	36	17	1	5	1			
	412	518.5	36	17	1	2	1			
	92	667.5	36	20	1	2	1			
	305	229.2	44	20	1		5			3

Chapter 3: Characterization of a novel *S. bicolor* terpene synthase

3.1 Summary

Sesquiterpene synthases convert the precursor molecule farnesyl pyrophosphate (FPP) into a skeleton structure that can undergo further enzymatic modifications to produce a variety of terpene products. The putative gene cluster identified in Chapter 2 contained an uncharacterized terpene synthase hypothesized to produce a sesquiterpene. In this chapter, protein functionality and terpene structure were determined. The terpene synthase was expressed in *Escherichia coli*, purified, and combined with FPP in an *in vitro* enzyme assay. Protein function was confirmed using gas chromatography - mass spectrometry (GC-MS) for identification of an unknown terpene product. In order to determine the structure of the unknown terpene, the terpene synthase was co-expressed in *E. coli* alongside a plasmid containing the mevalonate pathway to produce the product *in vivo*. The terpene product was isolated from the culture by solid-phase extraction (SPE) and analyzed by direct infusion mass spectroscopy.

3.2 Significance

Despite the economic importance and metabolic diversity of *S. bicolor*, few natural product biosynthetic genes have been characterized in this species. This study expresses a novel sesquiterpene synthase from *S. bicolor* and determines a potential molecular formula of the terpene product.

3.3 Contributions

Development of protocols for the Triversa Nanomate/Q Exactive direct infusion system was done by Dr. Bradley Evans.

3.4 Introduction

Five terpene synthase genes, each of which produces multiple terpenes, have been characterized in *S. bicolor* (Zhuang et al., 2012). These genes are part of the biosynthetic pathways of the leaf volatiles β-bisabolene, zingiberene, β-sesquiphellandrene, α-bergamotene, β- farnesene, and β-caryophyllene and minor products such as

sesquisabinene and α-humulene. At least 42 other putative terpene synthases have been identified by genome analysis in *S. bicolor*, but they remain uncharacterized (Zhuang et al., 2012). The gene of interest in this study, Sobic.001G339000, is a putative sesquiterpene synthase whose products have not been investigated. Sobic.001G339000 is the founding gene of a possible gene cluster containing several enzyme classes that are known to modify terpene skeleton structures. Before gene clustering can be confirmed, it must be determined if the founding gene encodes a functional protein.

Putative terpene synthases are often characterized by heterologous expression in an alternative host. The resulting protein is then extracted and its activity assessed by enzyme assay with the predicted substrate. Heterologous expression systems are selected based on their ease of use and ability to produce large quantities of product at a low cost. Plant species that are easy to cultivate and amenable to transformation, such as *Nicotiana benthamiana,* are the ideal host for synthesizing plant-based terpenes. They possess all the required cellular machinery and naturally produce terpene precursors, removing the need for exogenous supply of FPP, GPP, and GGPP. However, the yield of plant-based systems is variable and they are costly and time-consuming to maintain (Leavell et al., 2016). Microbial systems such as *E. coli* or *Saccharomyces cerevisiae* are better-suited for generating large volumes of protein, despite the extra steps required for purification and combination of the protein with an exogenous source of substrate.

A signal peptide on the terpene synthase will affect the subcellular location of the protein and therefore its functionality in a heterologous expression system. Sesquiterpene and triterpene synthases act upon FPP, which is synthesized in the cytosol by the mevalonate (MEV) pathway (Degenhardt et al., 2003). Monoterpene and diterpene synthases act on GPP and GGPP, respectively, which are synthesized in the plastid by the non-mevalonate (MEP) pathway (Degenhardt et al., 2003). Terpene synthases which act upon GPP and GGPP possess a subcellular localization signal that directs the protein to the plastid where the pools of precursor are stored. While some

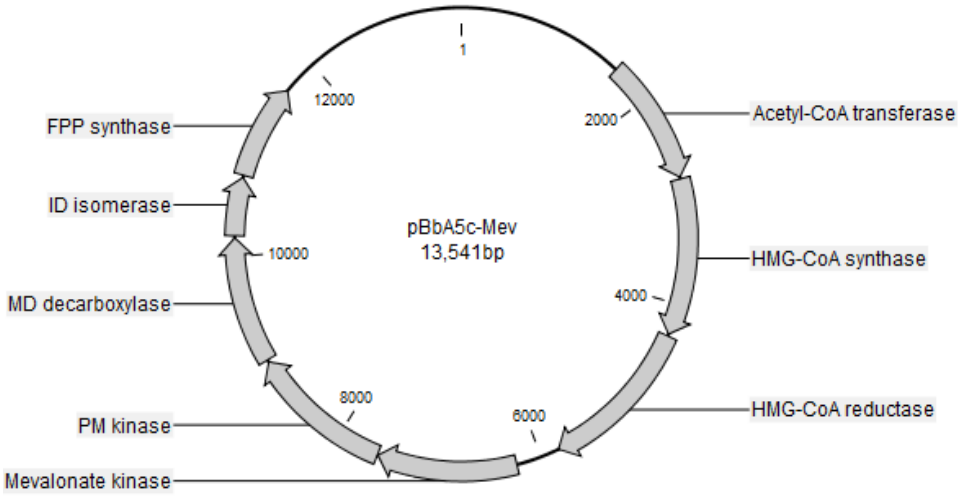
terpene synthases are promiscuous and can be found in both the plastid and the cytosol, typically they remain segregated (Aharoni et al., 2004). Expression systems such as *E. coli*, *S. cerevisiae,* and *Spodoptera frugiperda* (Sf9 cells) do not possess the correct cellular machinery to process compounds with this subcellular localization signal, so either a plant-based production system must be used to express plastidial terpene synthases or the transit peptide must be cleaved (Brückner and Tissier, 2013). Since Sobic.001G339000 is a putative sesquiterpene synthase and therefore active in the cytosol, no signal peptide is present and the protein is likely to be functional in a non-plant expression system.

E. coli is a common secondary metabolite production system as it is easy to transform and can produce high titres of product in small volumes of culture. Proteins can be synthesized in *E. coli*, purified, and then combined with an external source of substrate to determine protein function and product identity. The techniques for cloning and expressing plant-derived terpene synthases in *E. coli* are well established (Augustin et al., 2015a; Zhuang et al., 2012). *E. coli* cultures have been used to produce terpene synthases for a variety of terpenes, such as the defense compound β-caryophyllene (Norris, 2013) and several *S. bicolor* sesquiterpenes (Zhuang et al., 2012).

A limiting factor of sesquiterpene production in *E. coli* is the cost of the substrate. Pure FPP is costly to manufacture, and large quantities are required to create measurable amounts of sesquiterpenes. An alternative to exogenous supplementation with FPP is producing it in a microbial expression system alongside a sesquiterpene synthase. An *E. coli* vector containing the mevalonate pathway (pBbA5c-MevT(CO)-TI-MBIS(CO, ispA), referenced as pBbA5c-Mev) has been used to produce large volumes of sesquiterpenes in *E. coli* cultures (Peralta-Yahya et al., 2011). This vector contains all the genes of the MEV pathway necessary to convert acetyl-CoA into FPP, including acetyl-CoA transferase, 3-hydroxy-3-methylglutaryl (HMG)-CoA synthase, HMG-CoA reductase, mevalonate kinase, phosphomevalonate kinase, mevalonate diphosphate decarboxylase, isoprenyl diphosphate isomerase, and farnesyl pyrophosphate synthase

(Fig. 11). Overexpression of this pathway results in an accumulation of FPP in the cytosol of *E. coli* cells which can then be acted upon by a co-expressed terpene synthase, producing terpenes directly in the cell culture. Co-expression of the MEV pathway and a terpene synthase has been used to produce the alternative fuels farnesol and bisabolane and the artemisinin precursor amorphadiene (Martin et al., 2003; Peralta-Yahya et al., 2011; Wang et al., 2010).

Figure 11*: *pBbA5c-Mev contains the mevalonate pathway for FPP biosynthesis. *HMG = 3-*



hydroxyl-3-methylglutaryl, PM = phosphomevalonate, MD = mevalonate diphosphate, ID = isoprenyl diphosphate. Plasmid map adapted from sequence information from Peralta-Yahya et al., 2011 and assembled in CLC Sequence Viewer 7.

In this study, I present the characterization of a novel terpene synthase found in *S. bicolor*. Enzymatic activity of Sobic.001G339000 was confirmed by expressing the gene in *E. coli* both alone and alongside pBbA5c-Mev. The terpene product of Sobic.001G339000 was analyzed for chemical formula and structure using GC-MS and liquid chromatography - mass spectrometry (LC-MS) techniques.

3.5 Experimental Procedures

3.5.1 Insertion of the terpene synthase gene into E. coli

The terpene synthase Sobic.001G339000 was commercially synthesized (Genewiz) with *E. coli*-optimized codons in the vector pUC57. The gene was amplified using the primers and PCR parameters detailed in Sup. Tables S3.1 and S3.2 and purified by gel extraction (Qiagen). The gene was ligated into the expression vector pET28a(+) using NotI and NheI restriction sites and then transformed into the *E. coli* DH5α cell line by the heat shock method. The insertion in pET28a(+) was confirmed using Sanger sequencing and the vector was transformed into the *E. coli* expression strain BL21 Star (DE3) by the heat shock method.

3.5.2 Production of the terpene synthase in E. coli

Lysogeny broth (LB media) containing 50 µg/ml kanamycin was inoculated with BL21 Star (DE3) *E. coli* cells containing Sobic.001G339000 inside the pET28a(+) vector. BL21 Star (DE3) was also transformed with the empty vector pET28a(+) to serve as a negative control and cultured. Cultures were incubated for 12 h at 37˚C, 200 rpm, prior to inducing protein expression by the addition of 0 mM, 0.5 mM, 1 mM, 5 mM, and 10 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Induced cultures were incubated for 24 h at 15˚C, 200 rpm. Samples from induced *E. coli* were analyzed by SDS-PAGE on a 10% mini-PROTEAN TGX gel (Bio-Rad) to confirm that protein of the correct size was produced and to determine the ideal concentration of IPTG. Control cultures containing untransformed pET28a(+) underwent the same treatment. Once expression was confirmed, the above induction cycle was repeated using the optimal concentration of 1 mM IPTG. Protein was isolated using TALON metal affinity resins as described by Augustin et al., 2015. Protein stocks were frozen and stored at -80˚C.

3.5.3 Assessment of terpene synthase activity by in vitro *enzyme assay*

Enzyme assays were performed using the terpene-specific substrates FPP, GPP, and GGPP at a concentration of 10 mM along with the isolated protein. The substrate and

enzyme were combined in a buffer solution containing 500 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 1 M $MgCl_2$, and 200 mM dithiothreitol (DTT) with a 500 µL hexane overlay. Each assay was replicated three times with two negative controls, one with protein rendered non-functional by boiling and one with protein isolated from BL21 Star (DE3) transformed with the empty vector pET28a(+). Solutions were incubated at 30˚C for 2 hours and the reaction was halted using 500 mM ethylenediaminetetraacetic acid (EDTA). The hexane overlay was removed, subjected to centrifugation, and concentrated under N_2 to 100 µl for GC-MS analysis.

3.5.4 Assessment of terpene synthase activity by in vivo *enzyme assay*

The transformed plasmid pET28a(+) containing Sobic.001G339000 was transformed into the *E. coli* strain BL21 Star (DE3) alongside the vector pBbA5c-Mev, which contains all the genes involved in the MEV pathway. The transformation was confirmed using the primers and PCR protocol detailed in Sup. Tables S3.1 and S3.2. The two-vector strain of *E. coli* was inoculated into 5 mL of Terrific Broth (TB) media containing 1% v/v glycerol, 50 µg/ml kanamycin, and 32 µg/ml chloramphenicol. Control cultures, including pET28a(+) containing Sobic.001G339000, empty vector pET28a(+), and a two-vector strain containing empty vector pET28a(+) and pBbA5c-Mev, were also inoculated. Cultures were grown for 12 hours at 37˚C, 200 rpm. A 1 mL hexane overlay was added to each culture prior to induction of expression with 1 mM IPTG. Induced cultures were grown for 72 h at 28˚C, 200 rpm. The hexane overlay was removed, concentrated to 100 µl under N_2, and analyzed by GC-MS. The cell culture was extracted in 1 mL increments using a solution of 5 mL 3:2 hexanes:isopropanol and 20 µl acetic acid. The extracted cell culture was concentrated to 100 µl using N_2 and analyzed by GC-MS.

3.5.5 Optimization of the terpene extraction protocol

The two vector strain of BL21 Star (DE3) containing Sobic.001G339000 in pET28a(+) alongside pBbA5c-Mev was inoculated into 5 mL TB media containing 1% v/v glycerol, 50 µg/ml kanamycin, and 32 µg/ml chloramphenicol. Control cultures consisted of untransformed pET28a(+) alongside pBbA5c-Mev. Cultures were grown for 12 h at 37˚C,

200 rpm. Protein expression was induced with 1 mM IPTG and cultures were grown for 72 h at 28˚C, 200 rpm. Extraction of the *in vivo* enzyme assay cell culture was conducted using three different solvent systems: 3:2 hexanes:isopropanol, ethyl acetate, and chloroform. Culture was combined with solvent in glass culture tubes at a 1:1 ratio, vortexed, and sonicated in a sonication bath for 10 minutes. Tubes were then centrifuged for 5 minutes at 1000 rpm to remove cell debris. The organic layer was removed and concentrated to 100 µl using N_2.

Cell pellets were extracted using two solvent systems: methanol and 3:2 hexanes:isopropanol. 5 mL of cell culture was collected in a glass culture tube and centrifuged for 5 minutes at 3000 rpm to collect cells. All media was removed and 1 mL of solvent was added. The cell pellet was resuspended, briefly vortexed, and then sonicated for 10 minutes in a sonication bath. Tubes were then centrifuged for 5 minutes at 1000 rpm and the supernatant filtered by syringe though a 0.22 µm filter to remove cell debris. The filtered product was then concentrated to 100 µl using N_2 and analyzed by GC-MS.

3.5.6 GC-MS analysis of enzyme assay products and extraction method tests

The prepared samples were injected in 1 µl aliquots by a 7683B autosampler into a 7890A Agilent gas chromatograph coupled with a 5975C Agilent mass spectrometer (Agilent Technologies). A full scan method was run with helium as a carrier gas at a flow rate of 1.1 mL/min. The scan measured masses between 50 and 500 amu. Ion separation was conducted with a Phenomenex Zebron ZB-5MSi column (30 m x 250 µm internal diameter x 0.25 µm film). Inlet temperature was 250˚C. Samples were run with a temperature gradient starting at an initial temperature of 50˚C held for 3 minutes, then increasing to 160˚C at a rate of 35˚C/min, then increasing to 170˚C at a rate of 1.4˚C/min, and lastly ramping to 300˚C at a rate of 120˚C/min and holding for 2 minutes. Detected products were analyzed using MassHunter (Agilent Technologies) and the National Institute of Standards and Technology mass spectral database v2.0 (NIST).

Quantitation of enzyme assay products and cell extracts was conducted using a guaiol (Sigma-Aldrich) standard curve (curve range: 0.224 ng/µL to 4000 ng/µL).

3.5.7 Mass production of the terpene synthase

One litre cell cultures containing the dual plasmid system were prepared for extraction of metabolites with a protocol adapted from Peralta-Yahya et al., 2011. Cultures containing 5 mL of TB media, 50 µg/ml kanamycin and 32 µg/ml chloramphenicol, and 1% v/v glycerol were inoculated with the dual plasmid strain of BL21 Star (DE3) containing both pBbA5c-Mev and Sobic.001G339000 in pET28a(+). Cultures were grown for 12 h at 37˚C, 200 rpm, and used to inoculate a 50 mL culture of TB media with the same antibiotic concentrations. The 50 mL culture was grown for 12 h at 37˚C, 200 rpm, and used to inoculate a 1 L culture of TB media containing the same antibiotic conditions and grown for 12 h at 37˚C, 200 rpm. The 1 L culture was divided into two 500 mL cultures and an additional 500 mL of TB media supplemented with antibiotics was added. Protein expression was induced with 1 mM IPTG and cultures were grown for 72 hours at 28˚C, 200 rpm.

For metabolite extraction, two volumes of 3:2 hexanes:isopropanol were added to one volume of cell culture which was then vortexed and sonicated in a sonication bath for 10 minutes. The hexanes:isopropanol overlay was removed, centrifuged at 3000 rpm for 10 minutes, and dried down to 5 mL under N_2 gas. Samples were then stored at -20˚C.

3.5.8 Purification of terpene product by reversed-phase solid phase extraction

The SPE column was prepared by inserting a small glass wool frit (Ohio Valley Speciality) into the base of a 20 mm x 500 mm column, topped with a 1.5 cm layer of white quartz sand (Thermo Fisher Scientific). Air pockets in the glass and sand were removed by adding 10 mL of 50% methanol to the column and allowing it to flow through. Once clear of air, 60 mL of 50% methanol was slowly added to the column. 25 g of C-18 reverse phase silica gel with a 90 Å pore size (Sigma-Aldrich) was poured in small increments into the methanol and allowed to settle at the base of the column, with

frequent tapping to remove air pockets. The solvent was drained until it was 5 mm above the silica gel.

Aliquots (~100 µL) of extracted cell culture were diluted to 10 mL in 50% methanol and added to the SPE column by glass pipette. Solvent was drained until the entire sample was loaded into the gel. The column was then rinsed with approximately 1 column volume (50 mL) of 50% methanol followed by 75% methanol. These were collected as complete fractions and discarded. The column was then rinsed with 50 mL of 100% methanol, which was collected in 10 mL fractions. These fractions were re-extracted with 10 mL of pure hexanes to remove any residual glycerol, and then analyzed by GC-MS as described in section 3.5.6. Fractions containing the product of interest were dried completely under N_2 gas for long-term product storage. The above process was repeated until all the cell culture extract was purified.

3.5.9 Determination of terpene product chemical formula by direct infusion

Purified terpene product was analyzed by high resolution mass spectrometry using a Triversa Nanomate (Advion Biosciences) to infuse samples into a Q Exactive mass spectrometer (Thermo Fisher Scientific). The terpene product and standard guaiol were prepared in 80% methanol with either no supplementation, 0.1% formic acid, or 0.1% 100 mM ammonium acetate. Data were collected using a full MS scan in positive ion mode at a resolution of 140, 000 at m/z 50 to 750. Spray voltage was 1.7 kV and gas pressure was 0.5 psi. The capillary temperature was held at 250˚C and the S-lens radio frequency at 60. Collision energy was increased from 0 to 30 V as peaks were detected. Data was analyzed using Xcalibur 3.1.

3.6 Results

3.6.1 Expression of Sobic.001G339000 in E. coli

Induction testing confirmed that *E. coli* containing Sobic.001G339000 synthesized a protein distinct from control cultures when expression was induced with IPTG. Protein expression peaked at concentrations of 1 mM, with no observable increase at 5 mM or

10 mM concentrations. Once expression was confirmed, protein was purified from 1 L cultures induced at 1 mM IPTG. Based on the predicted amino acid sequence, the terpene synthase was expected to be 66.74 kDa. A protein of this size was isolated from induced cultures (Fig. 12). Control and pre-induction samples contained the expected range of proteins but no detectable quantity of the terpene synthase. Post-TALON purification samples contained a HIS-tagged protein of the expected size.

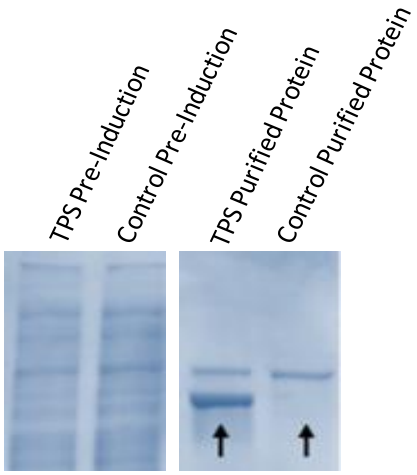


Figure 12: Confirmation of protein expression and isolation. *Protein expression was induced at 1 mM IPTG in control cultures and cultures containing Sobic.001G339000 (TPS). Cultures underwent TALON purification to isolate the TPS protein, mass 66.74 kDa (arrow).*

.

3.6.2 GC-MS analysis of in vitro *enzyme assay products*

The candidate terpene synthase produced detectable compounds when combined in an enzyme assay with FPP as a substrate (Sup. Fig. S3.1). No compounds were detected when the terpene synthase was combined with the substrates GPP or GGPP (Sup. Fig. S3.2, S3.3). Based on mass spectra and retention time analysis, the detected compound was most similar to the sesquiterpene alcohol guaiol. The retention time of the pure guaiol standard differed from that of the unknown terpene by 0.8 minutes, indicating that the compounds are not identical (Sup. Fig. S3.4). The second most-similar compound to the unknown terpene was the sesquiterpene alcohol β-eudesmol, though the retention time differed by 0.2 minutes and the fragmentation patterns were dissimilar (Fig. 13). The mass spectra data indicated that the terpene synthase product shared a precursor ion mass of 222.2 Da with guaiol and β-eudesmol, though that may also be a fragment of a larger molecule. All three compounds share molecular ions at

m/z = 59.0, characteristic of sesquiterpene alcohols. The terpene produced by Sobic.001G339000 could not be conclusively identified using either the NIST database or comparison to available standards.

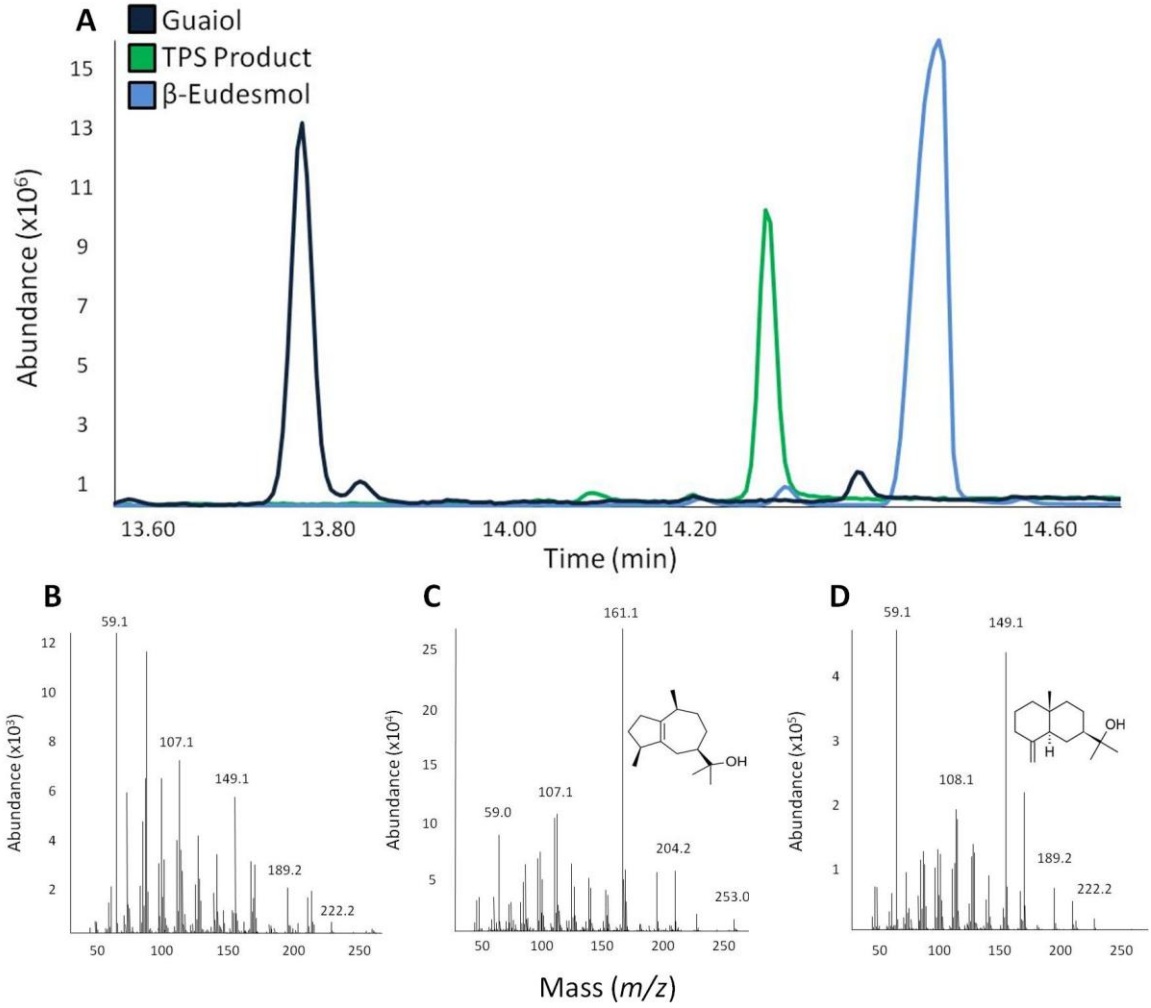


Figure 13: Comparison of unknown terpene synthase product (TPS) to standards by GC-MS. *The retention time (A) and mass spectra of the unknown terpene synthase product (B) were compared to the standards guaiol (C) and β-eudesmol (D) by GC-MS analysis. Standards were purchased from Sigma-Aldrich. Mass spectra after background subtraction are shown.*

3.6.3 GC-MS analysis of in vivo *enzyme assay products*

The expression of Sobic.001G339000 alongside pBbA5c-Mev produced compounds similar to those produced in the FPP-supplemented *in vitro* enzyme assay. The *in vivo* product had a retention time and fragmentation pattern identical to that of the *in vitro*

product (Fig. 14). The fragmentation pattern of both compounds contained molecular ions at *m/z* = 59.0, 107.0, 148.9, 189.0 and 222.1 (Fig 14B, 14C). The detection of the ion *m/z* = 59.0 supports the classification of the unknown compound as a terpene alcohol, as this peak is associated with the 2-hydroxyisopropyl group commonly seen in terpene alcohols (Dickschat et al., 2017). It can be concluded that an *in vivo* enzyme assay conducted in live *E. coli* cultures produces the same compound as purified protein combined with substrate in a controlled enzyme assay.

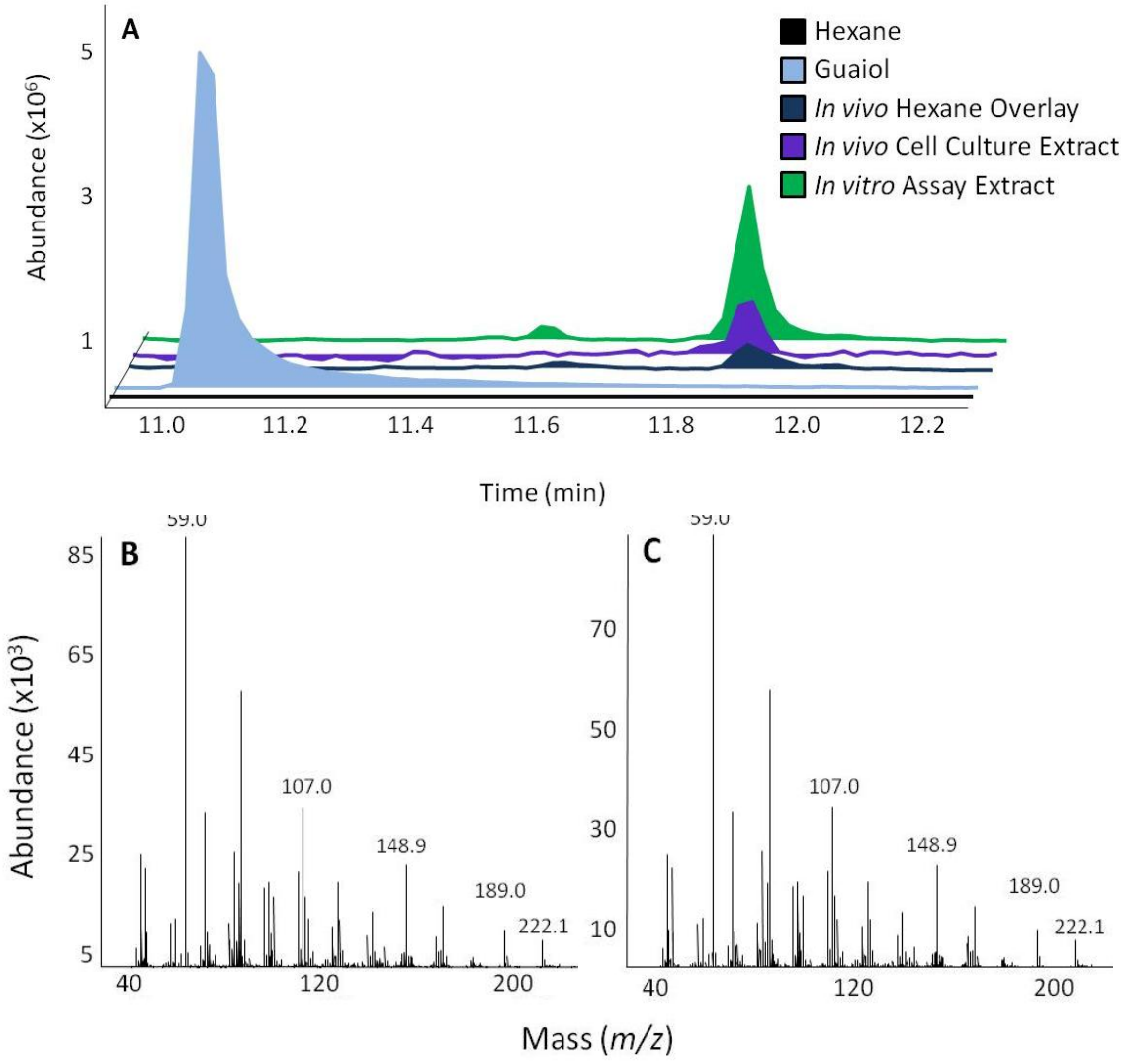


***Figure 14: Comparison of* in vivo *and* in vitro *enzyme assay products by GC-MS*.** *In vitro and in vivo enzyme assays were extracted using hexanes. (A) Comparison of the retention time of* in vivo *and* in vitro *enzyme assay products. (B, C) The mass spectra of* in vivo *(B) and* in vitro *(C) enzyme assay products were compared. Mass spectra after background subtraction are shown.*

3.6.4 Optimization of terpene extraction protocol

Extraction of the cell culture for the unknown terpene was successful with 3:2 hexanes:isopropanol, chloroform, and ethyl acetate. Hexanes:isopropanol was the most efficient extraction solvent, yielding approximately 107 ng/µL of terpene product, more than either ethyl acetate (91 ng/µL) or chloroform (51 ng/µL) (Fig. 15). Cell pellet extracts were found to contain less of the unknown terpene than cell culture extracts and to be an inefficient source of product. Hexanes:isopropanol extraction of cell pellets yielded approximately 47 ng/µL, while methanol only yielded 1 ng/µL.

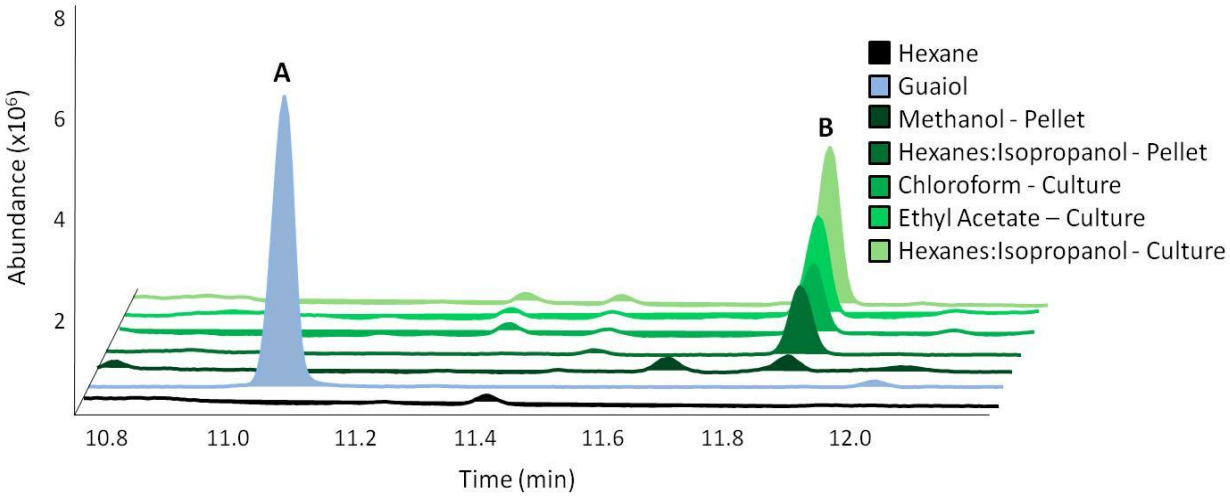


Figure 15: Analysis by GC-MS of extraction method testing for maximum terpene yield. *Guaiol (A) was used as a standard to quantitate the concentration of the unknown terpene product (B) after extraction of cell cultures and pellets with chloroform, ethyl acetate, methanol, and hexanes:isopropanol.*

3.6.5 Purification of terpene product by reversed-phase solid phase extraction

Application of concentrated cell culture extract to a SPE column resulted in significant removal of contaminants. As illustrated in Fig. 16, background products of *E. coli* cellular metabolism, such as indole, were washed from the column prior to collection of the terpene synthase product in 100% methanol. The compound eluted in fractions 4 and 5, which is between 30 mL and 50 mL of the 100% methanol wash.

The terpene synthase product was prone to degradation when stored in less than 75% MeOH. Samples which were prepared for chromatography by dilution in 50% MeOH and stored at either -20˚C or -80˚C overnight showed either no presence or severally depleted concentrations of the terpene synthase product when analysed by GC-MS the following day (data not shown).

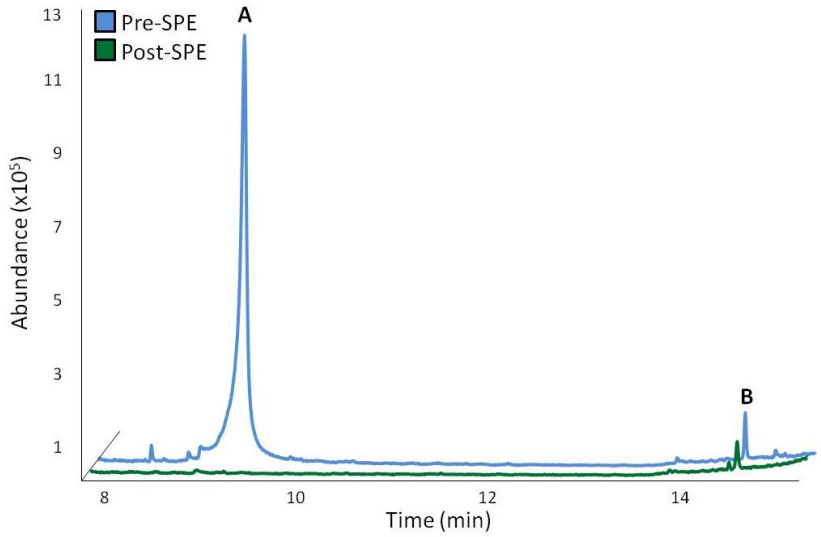


Figure 16: GC-MS analysis of purification of cell extracts by reversed-phase solid phase extraction. *Contaminants such as indole (A) were removes from cell culture by SPE, isolating the terpene synthase product (B).*

3.6.6 Chemical formula and structure determination by direct infusion

No peaks of interest were detected in unsupplemented or ammonium acetate-supplemented terpene synthase product samples using direct infusion mass spectrometry. Samples supplemented with 0.1% formic acid had a possible precursor ion peak at 223.2043 Da, identical to the precursor ion peak of the guaiol standard supplemented with 0.1% formic acid (Fig. 17). From this it could be concluded that either 1) the terpene synthase product had a mass identical to guaiol or 2) the terpene synthase product was larger than guaiol and the 223.2043 Da ion peak is a fragment of the true precursor ion. Based on previous GC-MS analysis the terpene synthase product had a precursor ion mass of 222.2, which would be equivalent to 223.2 when the compound becomes protonated during infusion.

Both guaiol and the terpene synthase product formed fragment ions, particularly at 217.1033 Da, 205.1939 Da, 195.0867 Da, 173.0566 Da, 173.0775 Da, and 157.0827 Da (Sup. Fig. 3.5). Once collision energy was increased above 0 V, the 223.2043 Da peak in

the guaiol standard and terpene synthase product was undetectable while the fragment ion peaks remained strong (data not shown).

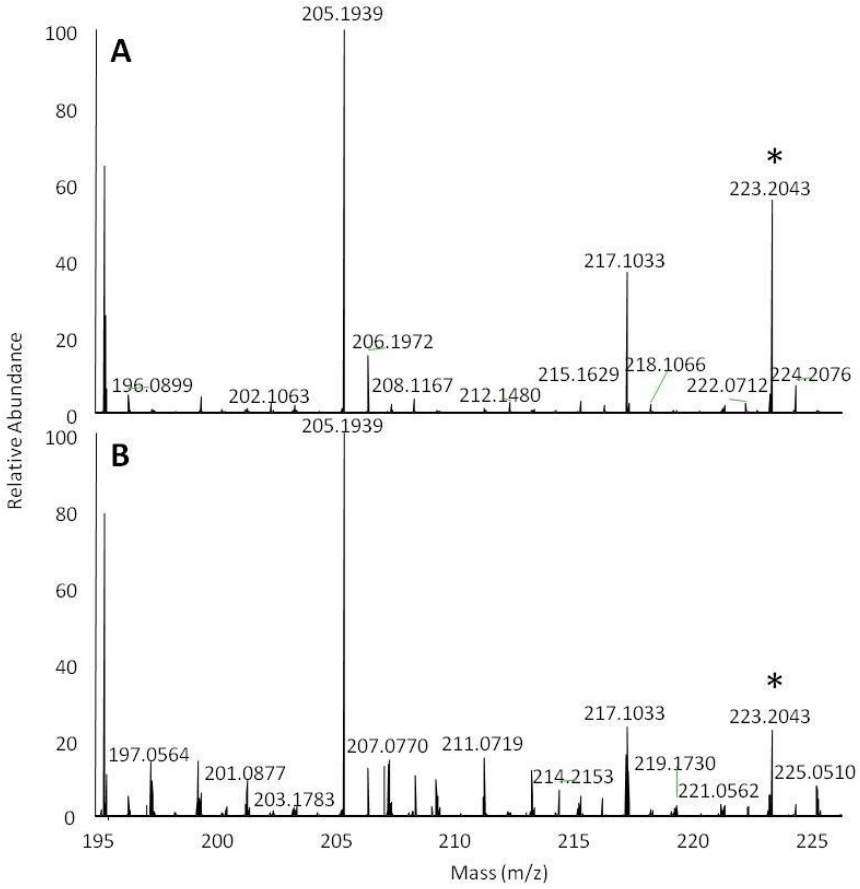


Figure 17: Mass spectra of guaiol and the terpene synthase product during Q Exactive direct infusion. *Guaiol (A) and the terpene synthase product (B) were analyzed by direct infusion MS using a Q Exactive. A precursor ion peak at 223. 2043 Da (*) is apparent when no collision energy is applied.*

Long-term exposure to formic acid, the acidification agent used during direct infusion, resulted in degradation of the terpene product.

3.7 Discussion

The putative terpene synthase Sobic.001G339000 acted upon FPP and produced a compound that was not identifiable by the NIST database or by comparison to available standards. Sobic.001G339000 had no activity when combined with GPP or GGPP, though it is possible that a terpene was produced in quantities below the detection limits of the GC-MS. The action of the terpene synthase upon FPP alone supported the conclusion that Sobic.001G339000 was a sesquiterpene synthase.

The terpene product was unidentifiable using the NIST database though it appeared to be similar to the sesquiterpene alcohols guaiol and β-eudesmol, which both have the chemical formula $C_{15}H_{26}O$ and exact mass of 222.1984 Da. The difference in the retention times and fragmentation patterns between guaiol, β-eudesmol and the unknown terpene product were significant and the identity of the unknown terpene could not be concluded from GC-MS analysis alone. Repeated GC-MS analysis of the terpene product consistently resulted in a possible precursor ion peak of 222.2 Da, the approximate mass of guaiol, β-eudesmol and a plethora of other sesquiterpenes. If 222.2 Da is the mass of the unknown terpene, it may share a chemical formula with guaiol and β-eudesmol. A search of the online NIST Chemistry WebBook yielded 400 compounds with the chemical formula $C_{15}H_{26}O$, though none of the available mass spectra were identical to that of the unknown terpene (NCBI Resource Coordinators, 2017).

The extreme degradation of the unknown terpene product when stored in less than 75% methanol or in a solution with formic acid suggested that it was highly unstable. This instability made characterization difficult as acidic and polar environments were often required for analysis, such as during direct infusion MS, LC-MS or HPLC. Direct infusion MS of the terpene product had results similar to GC-MS analysis, with a clear ion peak at mass 223.2043 Da identical to the standard guaiol. While this further supports the conclusion that the unknown terpene product shares a mass and molecular formula with guaiol, the labile nature of the compound and its propensity for degradation in polar and acidic solvents could mean that this detected peak is a fragment of a larger molecule which had fractured before entering the mass spectrometer. Further analysis, such as nuclear magnetic resonance spectroscopy, would be required to confirm that the peak is in fact the precursor ion and that the unknown terpene product has a mass of 222.2 Da and a chemical formula of $C_{15}H_{26}O$.

3.8 Supplementary Information

3.8.1 Supplementary Figures

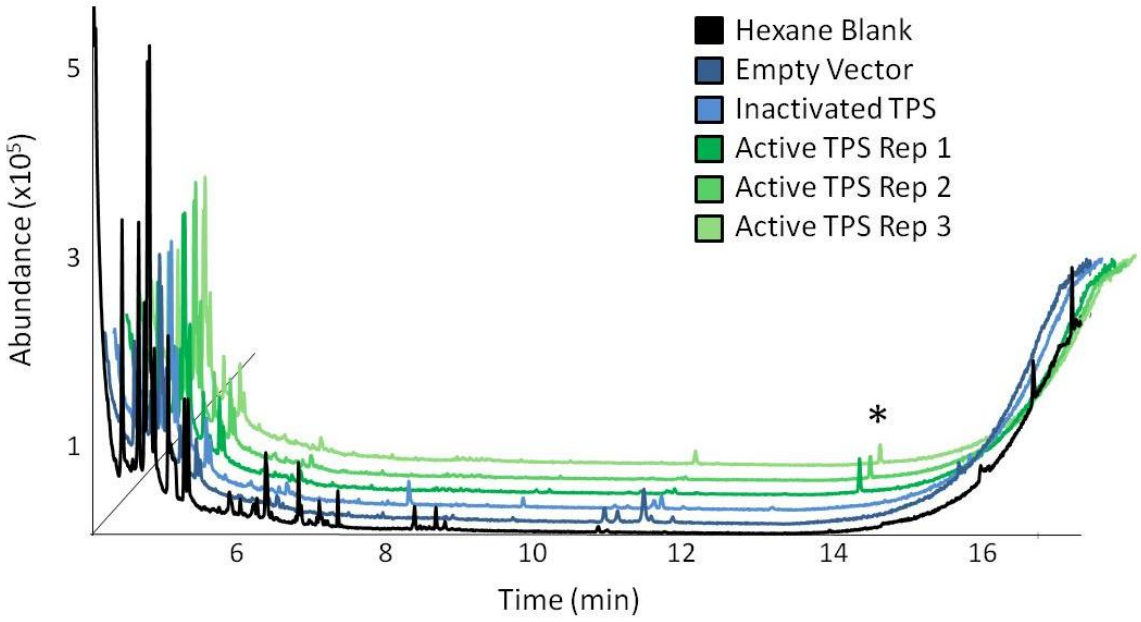


Figure S3.1: GC-MS analysis of the products of FPP-supplement enzyme assays. *Enzyme assays of His-tag purified Sobic.001G339000 were incubated with FPP and buffer solution for 2 hours and extracted with hexanes. A unique peak (*) was found in assays with active protein which is not found in the inactivated protein or negative controls. TPS = terpene synthase.*

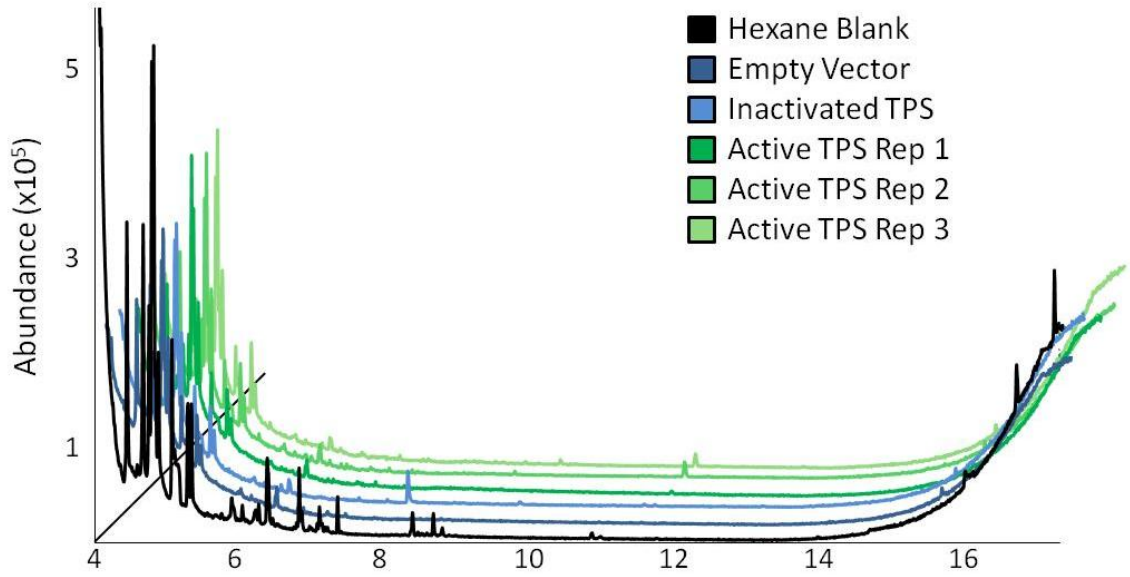


Figure S3.2: GC-MS analysis of the products of GPP-supplement enzyme assays. *Enzyme assays of His-tag purified Sobic.001G339000 were incubated with GPP and buffer solution for 2 hours and extracted with hexanes. No unique peaks were detected. TPS = terpene synthase.*

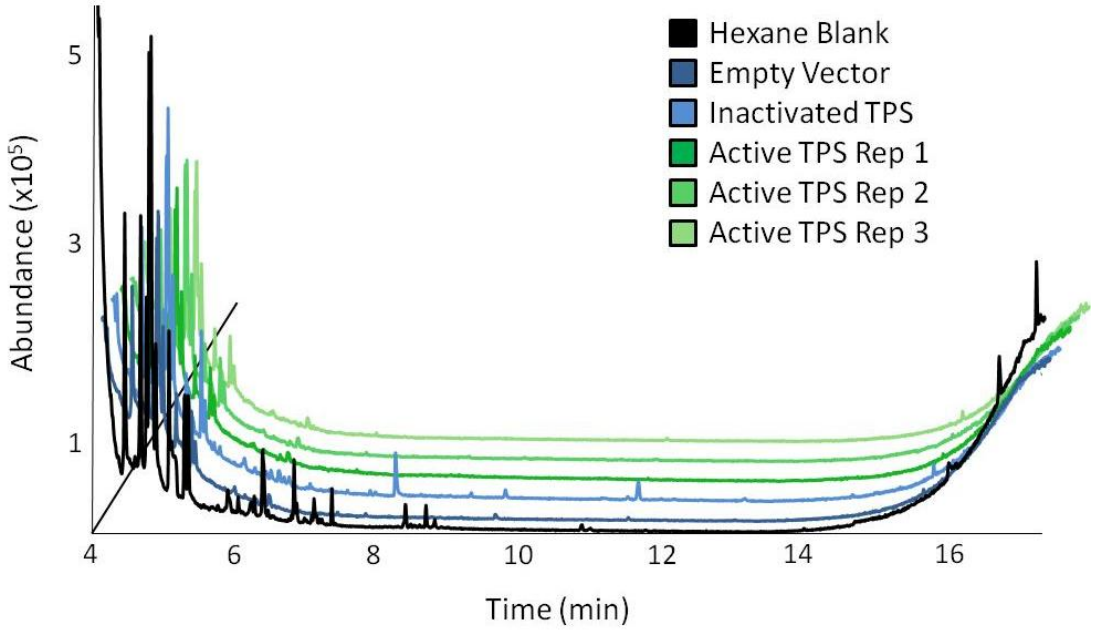


Figure S3.3: GC-MS analysis of the products of the GGPP-supplement enzyme assays. *Enzyme assays of His-tag purified Sobic.001G339000 were incubated with GGPP and buffer solution for 2 hours and extracted with hexanes. No unique peaks were detected. TPS = terpene synthase.*

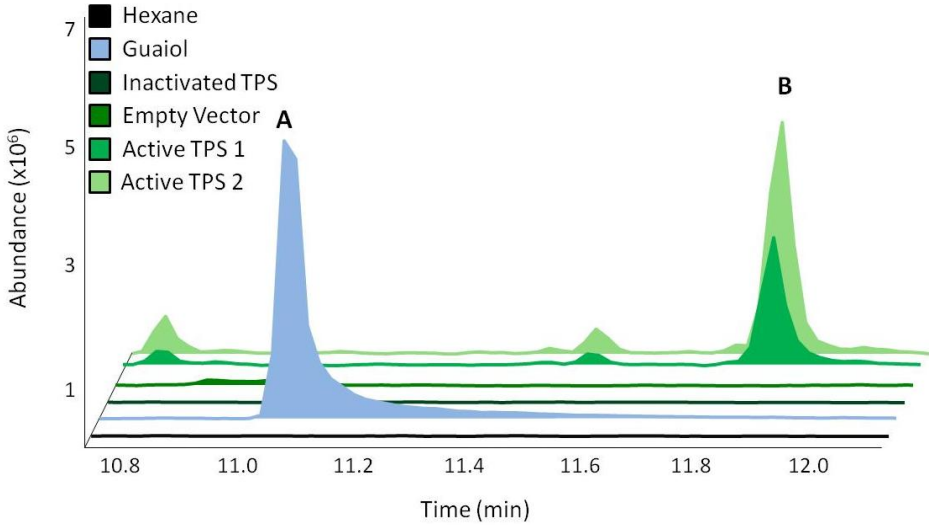


Figure S3.4: GC-MS analysis of the products of FPP-supplemented enzyme assays*. The retention time of guaiol (A) and the product of enzyme assays with active terpene synthase (TPS) (B) were compared by GC-MS analysis.*

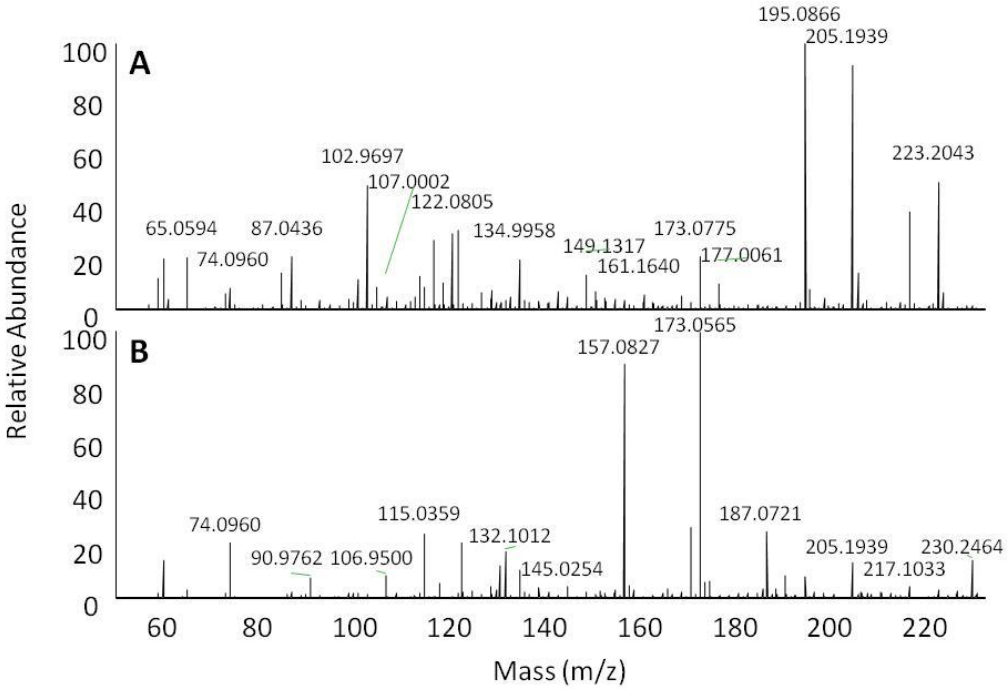


Figure S3.5: Direct infusion MS analysis of the terpene product. *The standard guaiol (A) and the unknown terpene product (B) were analyzed by direct infusion into a Q Exactive MS/MS system. The resulting fragmentation patterns were compared in search of the unknown terpene product precursor ion.*

3.8.2 Supplementary Tables

Table S3.1: Primer sequences

Insertion into pET28a(+)		
Gene/Plasmid	**Primer Position**	**5' - 3' Sequence**
Sobic.001G339000	5' with added NheI restriction site	TATATAGCTAGCATGGCCGCCGCACGTGAGGT
Sobic.001G339000	3' with added NotI restriction site	CACACAGCGGCCGCTTAAAAAGGAATCGGTTCGTCCAGC
Confirmation of insertion into *E. coli*		
Gene/Plasmid	**Primer Position**	**5' - 3' Sequence**
pBbA5c-Mev	5' Trc promoter	CACTGCATAATTCGTGTCGCTCAA
pBbA5c-Mev	3' Trc promoter	GGTTGAAGCTGATGTCCGGAAAGT

Table S3.2: PCR parameters

Application	**Polymerase**	**PCR Parameters**
Amplification of *E. coli*-optimized genes for insertion into pET28a(+) expression vector	Q5 DNA polymerase	98˚C for 30 sec, then 35 cycles of 98˚C for 10 sec, 52˚C for 20 sec, 72˚C for 1 min, and a final 5 min at 72˚C
Confirmation of gene/vector insertion	Taq DNA Polymerase	95˚C for 5 min, then 30 cycles of 94˚C for 30 sec, 52˚C for 30 sec, and 72˚C for 2 min, and a final 5 min at 72˚C

Chapter 4: The effect of clustered genes on the terpene synthase product

4.1 Summary

Terpene biosynthetic gene clusters can contain a variety of enzyme-producing genes and regulatory elements. The putative *S. bicolor* terpene biosynthetic gene cluster under investigation in this study includes a sesquiterpene synthase, a cytochrome P450, and a galacturonosyltransferase. The gene cluster was transiently expressed in *Nicotiana benthamiana* and *Spodoptera frugiperda* (Sf9) insect cell lines for protein production. Products of both heterologous expression systems were analyzed by GC-MS and LC-MS to determine if the clustered genes interact to produce a terpene product that is unique from the actions of the terpene synthase alone. A potential modified product of the gene cluster was identified in transfected *N. benthamiana* by Q Exactive LC-MS analysis.

4.2 Significance

In this study a putative terpene biosynthetic gene cluster was expressed in two heterologous expression systems and the metabolic profiles were compared using GC-MS and LC-MS analysis. A potential product of the gene cluster was identified and a protocol was developed for future gene cluster expression and chemical analysis.

4.3 Contributions

Development of the LC-MS protocols for the Eksigent/Q Exactive LC-MS system and compilation of resulting data was done by Dr. Bradley Evans. Development of protocols for the QTRAP 6500 LC-MS system was done by Megan Augustin. The transfection and RNA extraction of the first replicate of *N. benthamiana* transient expression was done by Julie Gauthier.

4.4 Introduction

Gene clustering cannot be concluded based on the physical proximity of biosynthetic genes alone. Their coordinated function must be verified. One way to confirm a putative gene cluster is expression in a heterologous system to determine if they are involved in the same biosynthetic pathway. Multiple expression systems exist, however host

selection is a critical factor of terpene biosynthetic cluster analysis. When expressing a eukaryotic gene in a prokaryotic system such as *E. coli*, proteins can be misfolded and sequestered in inclusion bodies to protect the cell from damage (Brückner and Tissier, 2013). Codon usage issues also affect synthesis of plant-based proteins in non-plant systems, as microbial hosts lack the necessary pools of tRNA (Moriyama and Powell, 1998). As such, eukaryotic systems including *N. benthamiana* and *S. frugiperda* (Sf9) insect cells are better suited to expressing genes with plant-optimized codons. These expression platforms are more costly to maintain than their microbial counterparts but offer a reduced likelihood of protein synthesis errors.

N. benthamiana is an effective terpene synthase expression system as it contains all the necessary cellular machinery to process plant-based compounds, ensuring that expressed proteins are properly folded and active (Bach et al., 2014). A secondary benefit of this expression system is that terpenes can be synthesized directly in plant tissues, as *N. benthamiana* naturally produces the terpene precursor FPP (Brückner and Tissier, 2013). It is also possible to transfect *N. benthamiana* with multiple genes at once by bacterial infiltration (Bach et al., 2014). An entire pathway can be expressed in a single leaf, provided the required substrate is present in sufficient quantities, which eliminates the need to isolate each protein individually and combine them in an *in vitro* enzyme assay. Complete metabolic pathways have been transfected into *N. benthamiana* for expression of terpenes such as artemisinin (Farhi et al., 2011), casbene, levopimaradiene, and cembratienol (Brückner and Tissier, 2013). Regulatory elements from the *S. bicolor* gene cluster producing the cyanogenic glucoside dhurrin have been transiently expressed in tobacco, and complete *S. bicolor* clusters have been transfected into other plant species such as *Zea mays* (Darbani et al., 2016; Song et al., 2004). The protocols for transfection and regulation of gene expression in *N. benthamiana* are well-established and make it a suitable host for *S. bicolor* gene cluster expression (Fischer et al., 2004).

An alternative to a tobacco expression system is *S. frugiperda* Sf9 insect cells. Sf9 cells are another eukaryotic expression system widely used to produce enzymes that are not amenable to other, more cost-effective systems. For expression in Sf9 cells, a gene of interest is inserted into a baculovirus transfer vector that homologously recombines with a linearized baculovirus, resulting in the expression of the recombinant protein (Chang et al., 2018). The protein can then be harvested and used for downstream analysis.

S. frugiperda can be used to express plant-based proteins as it has a low codon-usage bias and is able to supply the necessary tRNAs (Landais et al., 2003). The Sf9 insect cell expression system has been used repeatedly in the Kutchan lab to produce proteins such cytochromes P450 (Díaz Chávez et al., 2011; Gesell et al., 2009; Kilgore et al., 2016) and transaminases (Augustin et al., 2015b). Sf9 cells can also be infected with multiple recombinant viruses and express a complete pathway in a single culture, as has been done with the verazine biosynthetic pathway (Augustin et al., 2015b).

In this chapter, the putative *S. bicolor* gene cluster containing a terpene synthase (Sobic.001G339000), cytochrome P450 (Sobic.001G338900) and galacturonosyltransferase (Sobic.001G338400) was expressed in both *N. benthamiana* and Sf9 cells. Gene activity was compared both within and between systems.

4.5 Experimental Procedures

4.5.1 Extraction of gene cluster from S. bicolor *tissues*

S. bicolor was grown in the Donald Danforth Plant Science Center greenhouses under 14 hours of light, 28˚C in the day and 22˚C at night, and 40-100% humidity. Leaf, stem, and root tissues were collected from plants aged 2, 4, 6, 8, 10, and 12 weeks and frozen in liquid nitrogen. As the gene cluster of interest showed the highest expression in seedling root tissues (Section 2.6.3), RNA was extracted from 2-week old roots using the RNeasy Plant Mini Kit (Qiagen). RNA quality was confirmed using a Bioanalyzer 2100 (Agilent Technologies) and quantity was determined using a NanoDrop 2000 (Thermo Fisher

Scientific). RNA was converted to cDNA using the SuperScript III First Strand Synthesis System (Invitrogen).

Fulllength genes were amplified from cDNA using nested PCR. Primers listed in Sup. Table S4.1 were designed from the *S. bicolor* BTx623 genome, and PCR parameters were as detailed in Sup. Table S4.2. PCR products were purified by gel extraction (Qiagen). The extracted genes were transformed into a pJet vector using the restriction sites BglII and EcoRI (Sobic.001G339000 and Sobic.001G338900) and BglII and XmaI (Sobic.001G338400) and transformed into DH5α by heat shock. Insertion was confirmed by Sanger sequencing.

4.5.2 Expression of the gene cluster in N. benthamiana

The transfer vector pMDC32 was prepared for transfection without the use of the Gateway cloning system (Invitrogen) by removal of the toxic cassette. The cassette was flanked by the restriction sites AscI and SacI and digestion of pMDC32 with these enzymes followed by blunt end ligation (CloneJeET PCR cloning kit, Thermo Fisher Scientific) resulted in a functional vector with no toxic component. The multiple cloning site from pET28a(+) was amplified using the primers and PCR parameters described in Sup. Tables S4.1 and 4.2. The PCR product was digested using the KpnI restriction enzymes and purified by gel extraction (Qiagen). The pMDC32 vector without the toxic cassette was digested with KpnI, purified by phenol:chloroform extraction, ligated to the multiple cloning site from pET28a(+),and transformed into the *E. coli* DH5α cell line by the heat shock method. The insertion of the multiple cloning site was confirmed with Sanger sequencing.

Sobic.001G339000, Sobic.001G338900, and Sobic.001G338400 were PCR amplified from the pJet cloning vector, digested, purified by gel extraction (Qiagen), inserted into the modified pMDC32 vector using the restriction enzymes SalI and ApaI, and then transformed into the *E. coli* DH5α cell line by the heat shock method. The primers and PCR parameters used are detailed in Sup. Tables S4.1 and S4.2. The insertion was

confirmed using Sanger sequencing. Vectors containing the genes of interest were transformed into the *Agrobacterium tumefaciens* strain GV3101 pMP90 (Koncz and Schell, 1986) by electroporation using a 0.2 cm cuvette at a resistance of 400 Ω, capacitance of 25 µF, and a voltage of 2 kV. The silencing suppressor plasmid p19HA, derived from the p19 protein of the tomato bushy stunt virus with a hemaglutinin tag, was also independently inserted into GV3101 pMP90 (Dunoyer et al., 2004). After electroporation, 600 µL of LB media was added to each sample which was then incubated at 28˚C, 200 rpm, for 3 hours. Transformed cultures were spread on solid LB media plates containing 25 µg/mL rifampicin, 40 µg/mL gentamycin, and 50 µg/mL kanamycin. The gene combinations transformed into *A. tumefaciens* are detailed in Table 2. The silencing suppressor p19HA was transfected alongside each combination of vectors to reduce the likelihood of gene silencing. Transformation was confirmed using the primers and PCR protocols detailed in Sup. Tables S4.1 and S4.2. The bacteria were then added to infiltration media containing 1 M 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 5.6, 1 M $MgCl_2$, and 100 mM acetosyringone to a density of OD_{600}. Once mixed, the solution was incubated at room temperature for 2 hours. 1 mL of each gene combination was injected with a needleless syringe into the abaxial surface of 14-day old *N. benthamiana* leaves.

Each gene combination was injected in triplicate and the entire experiment was repeated four times. Transfected plants were transferred to a growth chamber under 16 hours of light, 22˚C and 50% humidity for 72 h. 50 mg of tissue was harvested and flash frozen in liquid nitrogen for RT-PCR. 200 mg of tissue was harvested and frozen for metabolite analysis.

***Table 2: Combination of vectors used to transfect* N. benthamiana.** *TPS = terpene synthase (Sobic.001G339000), P450 = cytochrome P450 (Sobic.001G338900), GT = galacturonosyltransferase (Sobic.001G338400). p19HA = silencing suppressor. All genes are in the modified pMDC32 vector.*

Infection Number	Vector Combination
1	Empty pMDC32 + p19HA
2	TPS, P450, GT, p19HA
3	TPS, P450, p19HA
4	TPS, GT, p19HA
5	P450, GT, p19HA
6	TPS, p19HA
7	P450, p19HA
8	GT, p19HA

4.5.3 Confirmation of gene expression in N. benthamiana *by RT-PCR*

RNA was extracted from 50 mg of frozen transfected tissues using the RNeasy Plant Mini Kit (Qiagen) and treated with DNAse using the Ambion TURBO DNA-free kit. RNA was then converted to cDNA using the SuperScript III First Strand Synthesis System (Invitrogen) or M-MLV reverse transcriptase (Thermo Fisher Scientific). Expression of the genes of interest was confirmed using the gene-specific primers and PCR parameters detailed in Sup. Tables S4.1 and S4.2. RNA quality was confirmed using the *N. benthamiana* tubulin gene TUA6.

4.5.4 Metabolite analysis of transfected N. benthamiana *by GC-MS and LC-MS*

Metabolites from replicates 1 and 2 were extracted from transfected leaf tissue by grinding frozen tissue in 70% ethanol, thoroughly vortexing, boiling for ten minutes, and then precipitating at -20˚C for 12 hours. Samples were then prepped for GC-MS and LC-MS by filtration through a 0.22 µm syringe filter. Metabolites from replicates 3 and 4 were prepared for analysis by suspending 200 mg frozen leaf tissue in 200 µL 80%

methanol in 1.5 mL round-bottom Safe-Lock eppendorf tubes. A stainless steel bead was added and samples were disrupted in a TissueLyser II (Retsch) at 20 Hz for 10 minutes. Samples were centrifuged at 16 000 g for 3 minutes and the supernatant was filtered through 0.8 µm Vivaclear Mini clarifying filters (Sartorius Stedim Biotech).

The prepared samples from replicates 1 and 2 were first analysed using a 7890A Agilent gas chromatograph coupled with a 5975C Agilent mass spectrometer (Agilent Technologies) as described in section 3.4.6. Detected products were analyzed using MassHunter (Agilent Technologies) and the NIST database (v2.0).

The samples from replicates 1 and 2 were also analyzed using a Prominence UFLC-XR (Shimadzu) LC system coupled with a QTRAP 6500 (AB Sciex Instruments) for MS/MS analysis. Samples were injected into a Phenomenex Gemini C-18 NX column (150 x 2.00 mm, 5 µM, 110 Å). Solvents were injected with a flow rate of 0.5 ml/min on the following binary gradient (Solvent A: 10% methanol, 0.1% acetic acid v/v in H_2O. Solvent B: 0.1% acetic acid v/v in 100% methanol): Solvent B was held at 0% for 2 min, then 0-95% from 3-20 min, then 95-0% from 21-28 min. Scan parameters included a Turbo Ion Spray at 550˚C and 5500 volts. Analysis included an Enhanced Product Ion (EPI) scan in the positive mode for masses of 223, 225, 239, 416, and 438 Da and a Neutral Loss scan for 176 and 194 Da. Masses were selected based on the possible modifications to the standard guaiol that could occur due to actions of a cytochrome P450 and galacturonosyltransferase (Sup. Table 4.3). Structures were predicted using ChemDraw Standard 12.0 (PerkinElmer). Retention times and fragmentation patterns of transfected leaves were compared to the guaiol standard (Sigma-Aldrich) and control tobacco leaves using Analyst 1.6 (AB Sciex Instruments).

The samples from replicates 3 and 4 were analyzed using an Eksigent LC (AB Sciex Instruments) coupled to a Q Exactive mass spectrometer (Thermo Fisher Scientific) to scan with high mass accuracy. Replicate 3 was analyzed using the solvents 10 mM ammonium bicarbonate (Solvent A) and 95% acetonitrile:10mM ammonium bicarbonate (Solvent B) and replicate 4 was analyzed using the solvents H2O:0.1%

formic acid (Solvent A) and 95% acetonitrile:10mM ammonium bicarbonate (Solvent B). Both replicates were run on the same gradient: Solvent A held at 98% from 0 to 120 seconds, decreased from 98% to 0% from 120 to 1020 seconds, held at 0% from 1020 to 1140 seconds, and increased to 98% from 1140 to 1200 seconds, and held at 98% until 1800 seconds and the end of the cycle. Replicate 3 was analyzed using a hydrophilic interaction liquid chromatography (HILIC) column (Higgins Analytical, Inc), 150 mm x 0.5 mm, and a polymeric reversed phase (PLRP-S) column (Higgins Analytical, Inc), 100 Å, 100 mm x 0.5 mm. Replicate 4 used the same HILIC column and a TARGA C-18 column (Higgins Analytical, Inc), 3 µm, 150 mm x 0.5 mm. Data was analyzed using Scaffold Elements 1.4.2 (Proteome Software) and Xcalibur (Thermo Fisher Scientific).

4.5.5 Expression of the complete cluster in S. frugiperda

Sobic.001G339000, Sobic.001G338900, and Sobic.001G338400 were amplified from the pJet vector using the primers and PCR parameters detailed in Sup. Tables S4.1 and S4.2. PCR amplified genes were digested using the restriction enzymes BglII and EcoRI (Sobic.001G339000, Sobic.001G338900) and BglII and XmaI (Sobic.001G338400), purified by gel extraction (Qiagen), and ligated into pVL1392, which was then transformed into the *E. coli* cell line DH5α by the heat shock method. The insertion was confirmed by Sanger sequencing.

The insect cell line produced from *S. frugiperda* ovarian tissues (Sf9) was grown in TC-100 media supplemented with 10% fetal bovine serum until they reached the log phase of growth at a density of 2 x 10^6 cells/mL with over 95% viability. Cells were removed from confluent growth flasks and seeded on a 24 well plate with approximately 8 x 10^5 cells/well. Sobic.001G339000, Sobic.001G338900, and Sobic.001G338400 in pVL1392 were then introduced dropwise into the insect cell line in the combinations described in Table 3, alongside the linearized baculovirus *Autographa californica* nuclear polyhedrosis virus (AcMNPV), using the Bac-N-Blue Transfection Kit (Thermo Fisher Scientific). Viral amplification occurred for 15 days at 27˚C, with the viral supernatant added to a fresh plate of Sf9 cells every 4 days. The presence of the gene of interest in

the viral supernatant was confirmed by PCR using the primers and parameters listed in Sup. Tables S4.1 and S4.2.

Table 3: Viral combinations used to infect Sf9 cells*. TPS = terpene synthase (Sobic.001G339000), P450 = cytochrome P450 (Sobic.001G338900), GT = galacturonosyltransferase (Sobic.001G338400), CPR = cytochrome P450 reductase from* Eschscholzia californica*.*

Infection Number	Vector Combination	Substrate
1	TPS	FPP
2	P450, CPR	TPS Product
3	GT	FPP
4	TPS, P450, CPR	FPP
5	TPS, GT	FPP
6	TPS, P450, GT, CPR	FPP

Viral supernatant (2.5 mL) was then added to fresh Sf9 cells which had been grown in 50 mL of suspension media (TC-100, 10% FBS, 0.1% Poloxamer 188 solution) at 27˚C, 140 rpm to a density of 2 x 10^6 cells/mL, centrifuged (900 rcf for 10 minutes) to collect, and resuspended in 7.5 mL of suspension media. The mixture was incubated for 1 hour at 27˚C, 80 rpm. Suspension media was added to a final volume of 50 mL. The viral stock multiplied for 4 days at 27˚C, 140 rpm before supernatant was collected by centrifugation (3000 rcf for 10 minutes) and stored at 4˚C.

Sf9 cells were prepared for protein expression in 50 mL of suspension media as described above and collected by centrifugation (900 rcf for 10 minutes). Cells were resuspended in 7.5 mL of suspension media to which viral supernatant was added. For single infections, 2.5 mL of viral supernatant was added. For double infections, 1.25 mL was added, for triple 0.83 mL, and 0.625 mL for quadruple infections. For every infection involving a cytochrome P450 a virus containing the *E. californica* cytochrome P450 reductase (CPR) and 2 mg/L hemin stock was added. The 10 mL solutions were incubated for 1 hour at 27˚C, 80 rpm. Suspension media was added to a total of 50 mL

and cultures were grown for 4 days at 27 ˚C, 140 rpm. Protein collection occurred as described in Gesell et al., 2009 using a solution of 100 mM Tricine and 5 mM thioglycolic acid at pH 7.5. Protein expression was confirmed by SDS-PAGE on a 10% mini-PROTEAN TGX gel (Bio-Rad). Protein was stored at -80˚C.

4.5.6 Assessment of gene clustering by in vitro *enzyme assay*

Enzyme assays were performed using the terpene precursor FPP at a concentration of 10 mM or the purified Sobic.001G339000 terpene product at a concentration of 1500 pmol. The isolated protein from Sf9 cultures was used as an enzyme source. Enzyme assays included the combinations of proteins and substrate described in Table 2. The substrate and enzyme were combined in a buffer solution containing 500 mM HEPES, 1 M $MgCl_2$, and 200 mM DTT with a 500 µL hexane overlay. Each combination was replicated five times. A protein sample rendered non-functional by boiling and the CPR from *E. californica* served as a negative control. The entire enzyme assay was repeated three times. Solutions were incubated at 30˚C for 4 hours and the reaction was halted using 500 mM ethylenediaminetetraacetic acid (EDTA). The hexane overlay was removed, purified by centrifugation, and all five replicates of each gene combination were concentrated under N_2 gas to 100 µl. Samples were then analyzed by GC-MS and LC-MS.

4.5.7 Analysis of enzyme assay products

Prepared enzyme assay samples were analyzed by GC-MS using a 7890A Agilent gas chromatograph coupled with a 5975C Agilent mass spectrometer (Agilent Technologies) as described in section 3.5.6. Mass spectra and retention times of detected products were analyzed using MassHunter (Agilent Technologies) and the NIST database (v2.0).

Samples were prepared for LC-MS analysis by filtration through a 0.22 µm syringe filter. Samples were injected in 5 µL aliquots into a QTRAP 6500 LC-MS/MS system using the same specifications and binary gradient as detailed in section 4.5.4. The retention times

and fragmentation patterns of active protein samples were compared to thenon-functional protein and CPR controls using Analyst 1.6 (AB Sciex Instruments).

4.6 Results

4.6.1 Confirmation of expression in N. benthamiana *by RT-PCR*

Transfected leaf extracts were processed to obtain RNA. RT-PCR was performed on each extract using primers for Sobic.001G339000, Sobic.001G338900, and Sobic.001G3384000. A tubulin gene (TUA6) from *N. benthamiana* was used as a control. RT-PCR confirmed that the transfected genes were expressed in the expected *N. benthamiana* leaf tissue. Each combination of vectors was expressed correctly in multiple replicates (Fig. 18). A faint band was visible in samples that were not expected to express the galacturonosyltransferase. *N. benthamiana* contains several putative galacturonsytransferases that have sequence similarities to the transfected gene which may be detected by RT-PCR (Bombarely et al., 2012)

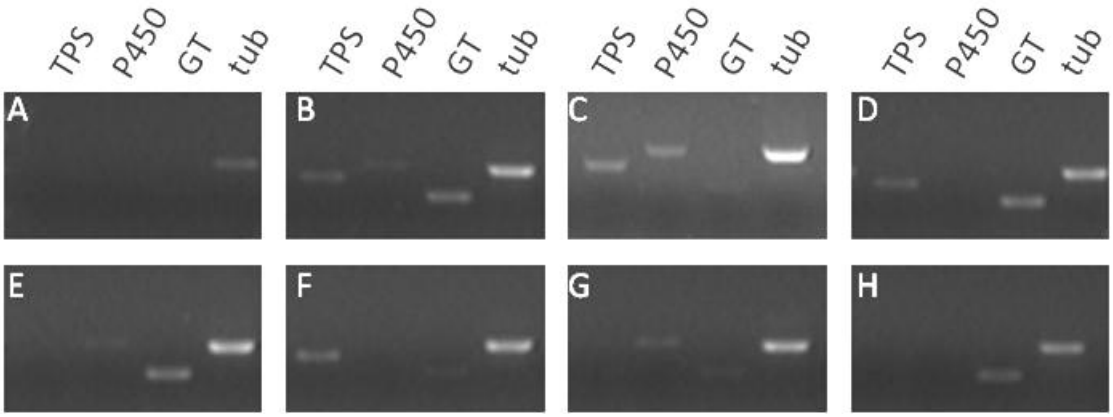


Figure 18: Confirmation of expression in* N. benthamiana *by RT-PCR of transfected leaf tissues. *Gene is indicated at the top of the figure, with TPS = terpene synthase, P450 = cytochrome P450, GT = galacturonosyltransferase, tub = tubulin positive control. Samples were ground under liquid nitrogen prior to RNA extraction and cDNA synthesis. Expected products for each transfection event are: A) Tubulin control only B) TPS, P450, GT C) TPS, P450 D) TPS, GT E) P450, GT F) TPS alone G) P450 alone H) GT alone.*

4.6.2 Metabolite analysis of transfected **N. benthamiana** *by GC-MS and LC-MS*

Transfected leaf tissues expressing the gene combinations described in Table 2 were extracted using 80% methanol and analyzed by GC-MS. GC-MS analysis provided limited results. All transfected tissues had similar metabolite profiles to both empty vector controls and untransfected tobacco leaves (Sup. Fig. S4.1). While several compounds present in tobacco leaves were detected, such as catechol and its precursor phenol, pyridine, ethanone, and the nicotine isomer anabasine (Xu et al., 2004), no sesquiterpene alcohols or similar compounds were detected. LC-MS analysis using the QTRAP 6500 system provided similar results with no unique peaks detected in transfected tissues compared to the controls (data not shown).

Samples from the third and fourth transient expression experiments were extracted using 80% methanol and analyzed by LC-MS. Whole metabolome analysis using the Eksigent/Q Exactive LC-MS system detected over 3,600 metabolites using the HILIC/TARGA C-18 column combination and over 7,600 using the HILIC/PLRP-S columns. Comparison of expressed genes to control tissues found a metabolite unique to tissues expressing the terpene synthase in conjunction with either the cytochrome P450 alone, the galacturonosyltransferase alone, or the cytochrome P450 and galacturonosyltransferase together. Similar peaks at low intensity were detected in tissues expressing the terpene synthase alone. The metabolite had an exact mass of 631.1999 Da and eluted through the HILIC column (Fig. 19). Analysis in Scaffold Elements predicted that this compound is ikarisoside F with a mass accuracy score of 0.86 and an identity score of 0.798. A carbon composition of 27 +/- 4 was determined based on the intensity of the ^{12}C peak at 631.1999 Da compared to the intensity of the ^{13}C isotope peak at 632.2031 (Fig. 20).

Attempts to repeat the analysis of the leaf extracts were futile as the majority of observed peaks were sharply degraded or lost within days of collection. Further analysis of the observed 631.1999 Da peak by LC-MS/MS was not possible as the peak was too degraded to obtain useful data.

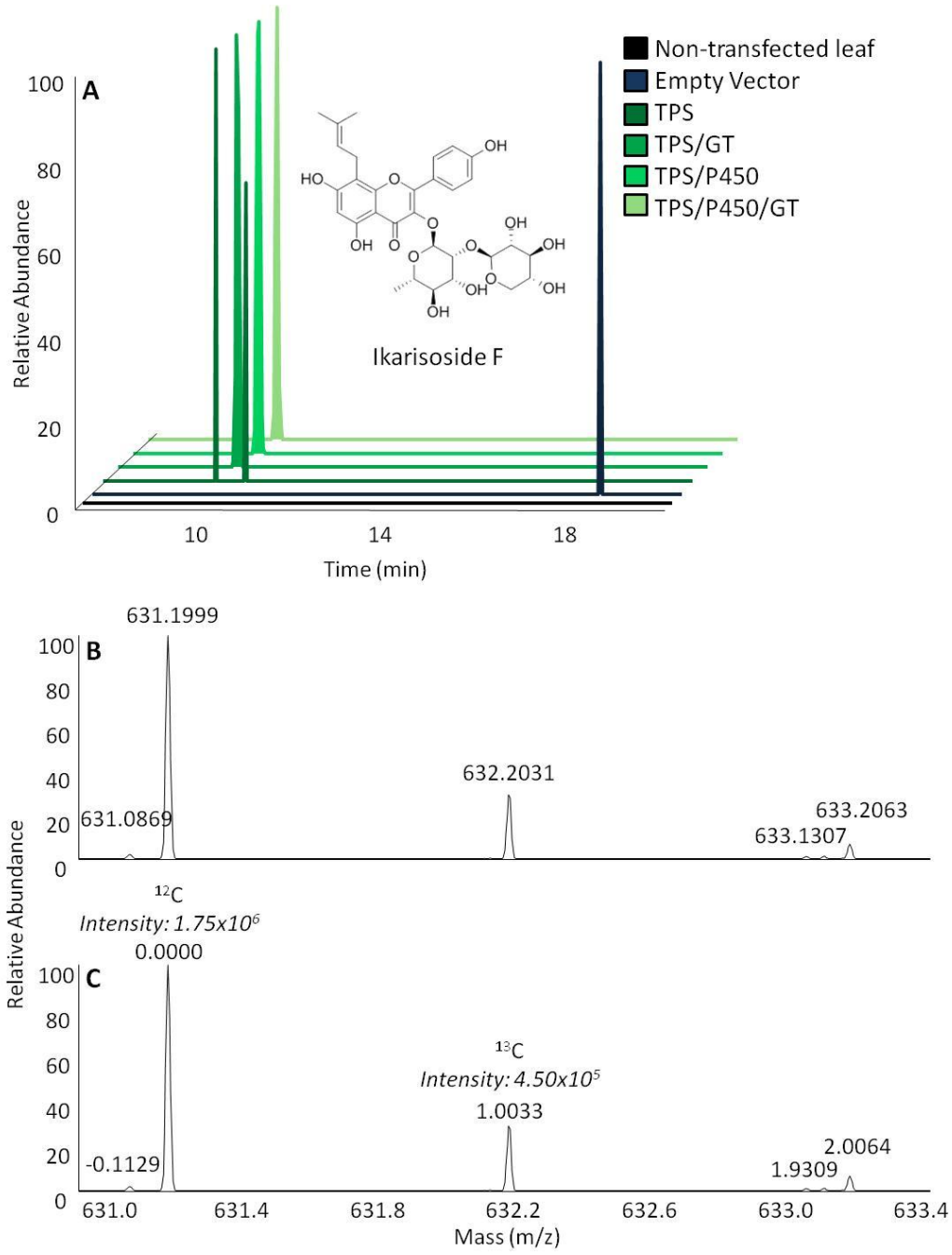


Figure 20: Metabolite analysis of transfected* N. benthamiana *by LC-MS detected a unique metabolite of 631.1999 Da. *Transfected leaf tissues were analyzed using an Q Exactive LC-MS system. Peaks of interest (A) and the exact mass of these peaks (B) were screened for using Scaffold Elements and Xcalibur. The ratio of C_{12} to C_{13} isotopes was used to calculate a carbon count of 27 +/- 4 (C).TPS = terpene synthase product, P450 = cytochrome P450, GT = galacturonosyltransferase.*

4.6.3 Confirmation of protein expression in Sf9 cells

The expression of protein in Sf9 cells was confirmed by SDS-PAGE. The expected mass of the terpene synthase was 66.75 kDa, the cytochrome P450 was 57.59 kDa, the galacturonosyltransferase was 62.63 kDa, and the CPR used as a control and co-expressed with the cytochrome P450 was 78.77 kDa. As Sf9 cells often do not produce enough recombinant protein to detect using SDS-PAGE, several of the samples did not show clear bands at the appropriate sizes (Fig. 21). A band of approximately 66.75 kDa is faintly visible in all samples. A band around the same size as the 63 kDa ladder marker is noticeably thicker in samples which are contain the 62.63 kDa galacturonosyltransferase. Poorly separated bands appear between 55 and 60 kDa, possibly attributed to the cytochrome P450. The CPR expressed alongside the cytochrome P450 was not detected outside control samples.

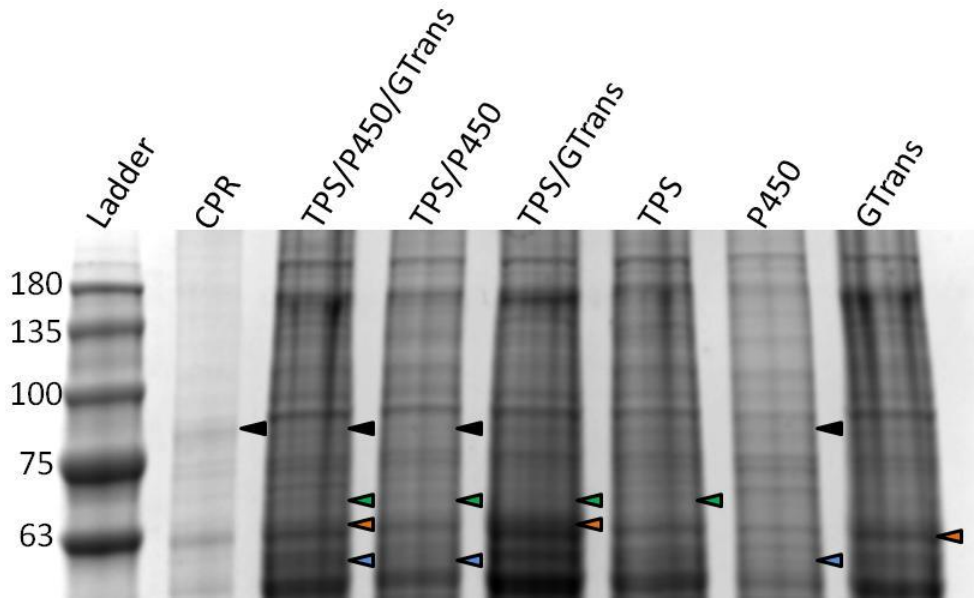


Figure 21: SDS-PAGE analysis of recombinant protein expression in Sf9 cells. *The areas where bands are expected are indicated by arrows. Black = CPR, 78.77 kDa, green = TPS (Sobic.001G339000, terpene synthase), 66.75 kDA, Orange = GTrans (Sobic.001G338400, galacturonosyltransferase), 62.63 kDa, and blue = cytochrome P450 (Sobic.001G338900), 59.59 kDa.*

4.6.4 Metabolite analysis of Sf9 cell enzyme assay products

The metabolite content of enzyme assays produced from recombinant proteins was similar to those of control CPR and inactivated protein under both GC-MS (data not shown) and LC-MS analysis (Sup. Fig. 4.2-4.7). No unique peaks were detected that could be attributed to the actions of the recombinant proteins.

4.7 Discussion

None of the expected products of enzyme assays using recombinant protein produced by Sf9 insect cells were detected by GC-MS or LC-MS. SDS-PAGE was conducted to confirm protein expression but the majority of expected bands were unclear. Sf9 cells often do not produce enough protein for detection by SDS-PAGE, especially when expressing multiple proteins in a single culture, so negative SDS-PAGE results do not always indicate that no protein was produced. Sf9 cells contain several proteases that can degrade recombinant proteins before they are collected (Gotoh et al., 2001). Adjusting the time of protein harvest could reduce the percentage of protein that is degraded. It is also possible that the plant-derived proteins may be misfolded and non-functional when synthesized in insect cells, or that protein is not being produced in large enough quantities for an enzyme assay. In this case, concentrating the protein and using alternative isolation methods could improve detection. Detection could also be improved by elucidating compound structure prior to analysis of enzyme assays by targeted LC-MS, as an accurate mass is needed to ensure assay products are detected.

While genes transfected into *N. benthamiana* were expressed, as confirmed by RT-PCR, no measurable amount of the unknown terpene product was detected using GC-MS or QTRAP 6500 LC-MS analysis. Several secondary metabolites were detected by GC-MS in leaf tissues extracted using the both the ethanol and methanol extraction protocols, and the product of the terpene synthase is known to be soluble in these solvents, so it is unlikely that the lack of product is due to extraction methodology. A possible explanation is that the transfected *N. benthamiana* did not synthesize enough terpene product to be detectable by GC-MS, as this equipment is significantly less sensitive than

LC-MS systems with electrospray ionization capabilities (Alder et al., 2006). Low accumulation of a sesquiterpene product could be due to plants having an inadequate amount of the FPP precursor in transfected tissues (Yu and Utsumi, 2009). Previous studies have overcome the issue of limited precursor availability by localizing enzyme activity to the mitochondria, where there is a larger store of FPP (Kappers et al., 2005), or by overexpressing a FPP synthase alongside the targeted genes (Wu et al., 2006). A more labor-intensive solution would be to generate transgenic lines of *N. benthamiana* to express the desired gene combinations. Transgenic plants would be able to grow and express the genes of interest for a longer period of time than transfected plants, and large tissue samples would be extracted, increasing the metabolite content for GC-MS (Ali et al., 2017).

Another potential reason that no sesquiterpene product was detected in transfected tissues is that tobacco plants may be converting the terpene product to another metabolite. While GC-MS analysis may not be able to detect low quantities of metabolites, LC-MS is several orders of magnitude more sensitive and should have no such issues. However, the LC-MS protocol used in this study is a highly targeted analysis that detects ions of a specific mass. If the terpene product is not within the programmed range of masses, it is unlikely it will be detected. Plant-based enzymes are known to produce variable products when expressed in different systems due to variations in native enzymes and cellular machinery. A geraniol synthase from *Ocimum basilicum* produced different minor products (citronellol, linalool, and nerol) when expressed in *Vitis vinifera*, *N. benthamiana*, *E. coli*, and *S. cerevisiae* (Fischer et al., 2013). Enzyme activity depends on the cellular environment, and plant-based expression systems contain a tremendous number of native enzymes which can interfere with transfected gene activity. A potential solution to this issue is to express the targeted genes in both *N. benthamiana* and an alternative expression system, such as Sf9 cells as was done in this study, and compare metabolite content.

The whole metabolome analysis of leaf extracts by Q Exactive LC-MS using a HILIC column detected over 3,600 metabolites. Screening for differences between controls and tissues expressing the genes of interest found one metabolite that could be the result of transfected genes. The predicted identity of this metabolite was ikarisoside F, a flavonoid from the roots of *Epimedium* species with the chemical formula $C_{31}H_{36}O_{14}$ (Zhang et al., 2005). However, the mass of ikarisoside F differs by 1.008 Da from the mass of the detected metabolite and the identity score of the match was less than 80% accurate. It is likely that the identified metabolite is a compound which is similar in mass to ikarisoside F, but not identical. Mass spectra analysis determined that the identified metabolite contains 27 carbon atoms, +/- 4, which would be expected for a base sesquiterpene structure modified by the addition of a sugar such as glucose (180.16 Da) and a sugar acid such as galacturonic acid (194.14 Da) (NCBI Resource Coordinators, 2017).

Identification of this metabolite by a more targeted Q Exactive LC-MS analysis was inhibited by degradation of the product. Many samples which had clear peaks immediately post-collection showed degraded peaks or an absence of peaks within days, despite being stored at -80˚C.While a methanol-water solution is commonly used during plant tissue extractions, up to 8% of compounds from extracts stored in methanol could be derived from a reaction between the solvent and extracted metabolites (Sauerschnig et al., 2017). The reaction between extract and solvent can occur immediately upon contact or over time, even when stored at -80˚C. The previous observation that the unknown terpene product degraded when stored in less than 75% MeOH or in 0.1% formic acid suggested that the compound of interest was highly labile and likely subject to degradation over time when in the presence of both a reactive solvent and the chemical activities of other extracted plant compounds. Generation of fresh tissues expressing the genes of interest is necessary for further analysis. Transgenic *N. benthamiana* may be a more efficient and consistent source of tissue than transfected leaves.

4.8 Supplementary Information

4.8.1 Supplementary Figures

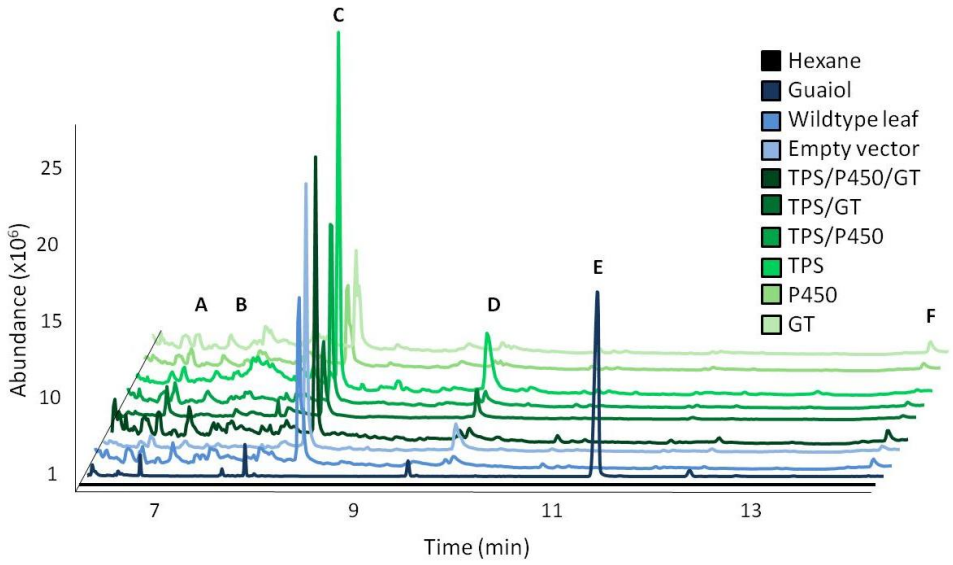


Figure S4.1: GC-MS analysis of transfected leaf tissues. *Leaf tissues expressing the genes indicated in the legend were extracted with 70% ethanol, ground in liqud nitrogen, and analyzed by GC-MS. Selected metabolites observed included phenol (A), catechol (B), pyridine (C), anabasine (D), and ethanone (F). Guaiol standard is indicated by (E). TPS = terpene synthase product, P450 = cytochrome P450, GT = galacturonsyltransferase.*

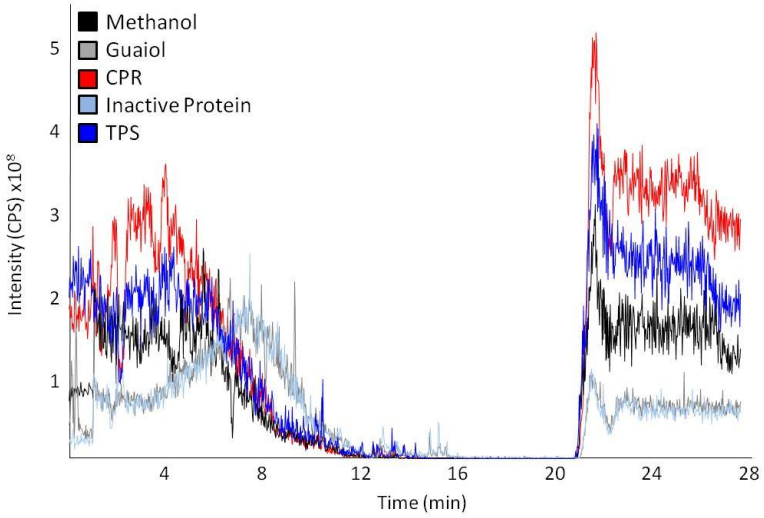


Figure S4.2: LC-MS analysis of Sf9 cell enzyme assay products. EPI scan at mass 223 Da. *Protein collected from Sf9 cells were combined in an* in vitro *enzyme assay. The activity of the terpene synthase (TPS) on FPP was compared to the control cytochrome P450 reductase (CPR) and inactivated protein by targeted LC-MS analysis. CPS = counts per second.*

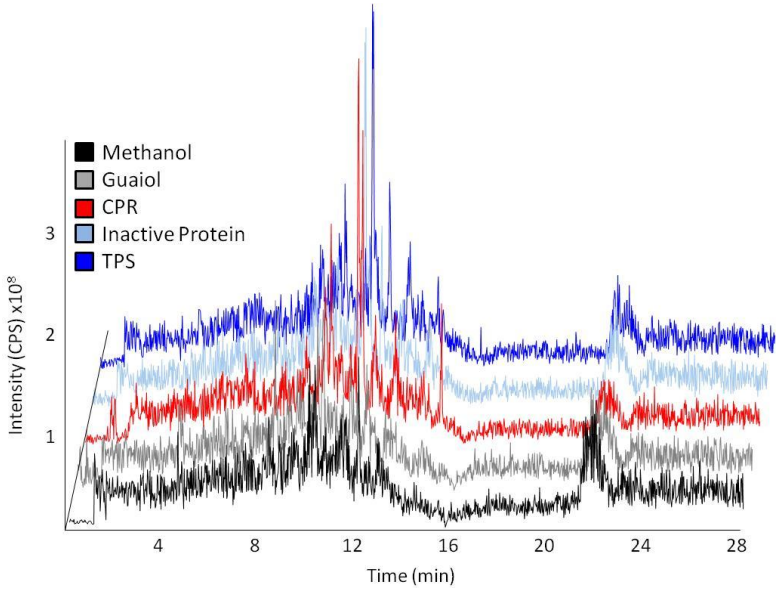


Figure S4.3: LC-MS analysis of Sf9 cell enzyme assay products. EPI scan at mass 225 Da. *Protein collected from Sf9 cells were combined in an* in vitro *enzyme assay. The activity of the terpene synthase (TPS) on FPP was compared to the control cytochrome P450 reductase (CPR) and inactivated protein by targeted LC-MS analysis. CPS = counts per second.*

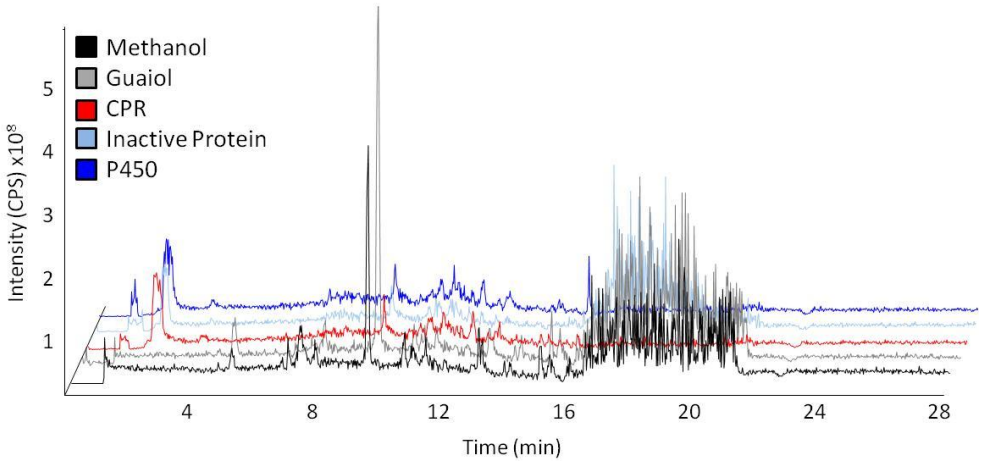


Figure S4.4: LC-MS analysis of Sf9 cell enzyme assay products. EPI scan at mass 238 Da. *Protein collected from Sf9 cells were combined in an* in vitro *enzyme assay. The activity of the cytochrome P450 (P450) on purified terpene product was compared to the activity of the control cytochrome P450 reductase (CPR) and inactivated protein. CPS = counts per second.*

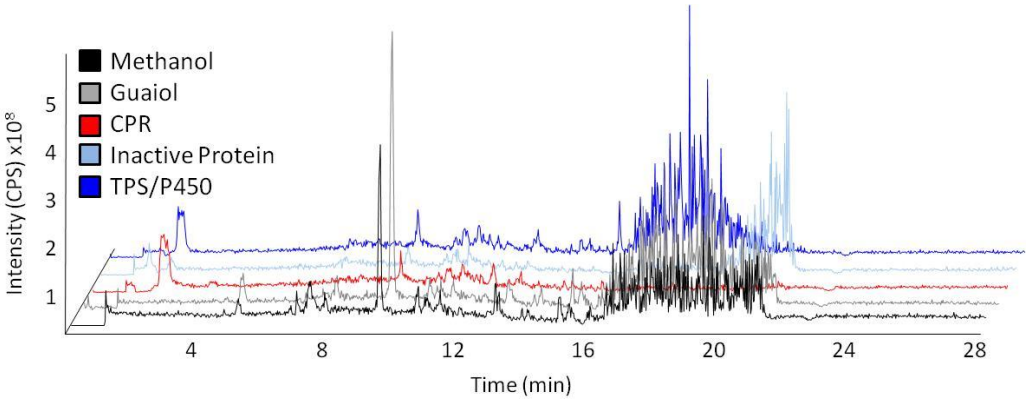


Figure S4.5: LC-MS analysis of Sf9 cell enzyme assay products. EPI scan at mass 238 Da. *Protein collected from Sf9 cells were combined in an* in vitro *enzyme assay. The activity of the terpene synthase (TPS) and cytochrome P450 (P450) on FPP was compared to the activity of the control cytochrome P450 reductase (CPR) and inactivated protein. CPS = counts per second.*

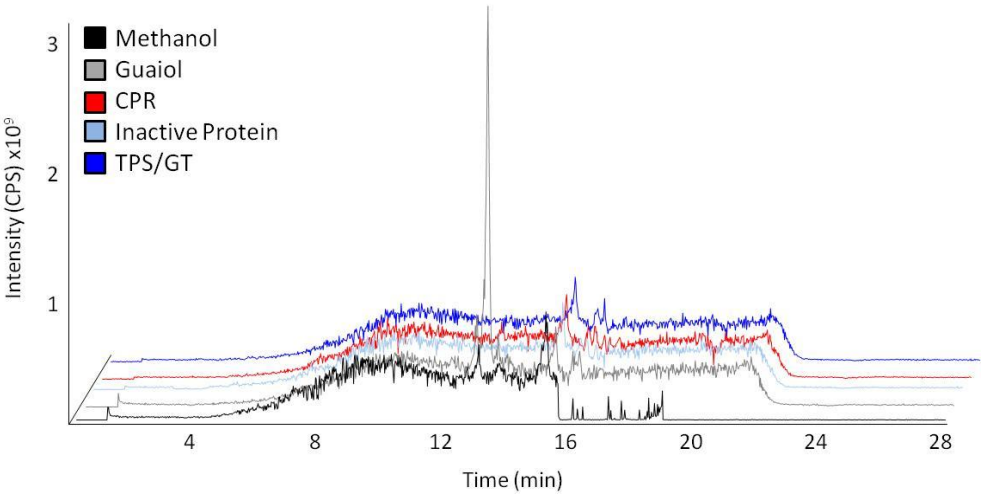


Figure S4.6: LC-MS analysis of Sf9 cell enzyme assay products. EPI scan at mass 415 Da. *Protein collected from Sf9 cells were combined in an* in vitro *enzyme assay. The activity of the terpene synthase (TPS) and galacturonosyltransferase (GT) on FPP was compared to the activity the control cytochrome P450 reductase (CPR) and inactivated protein. CPS = counts per second.*

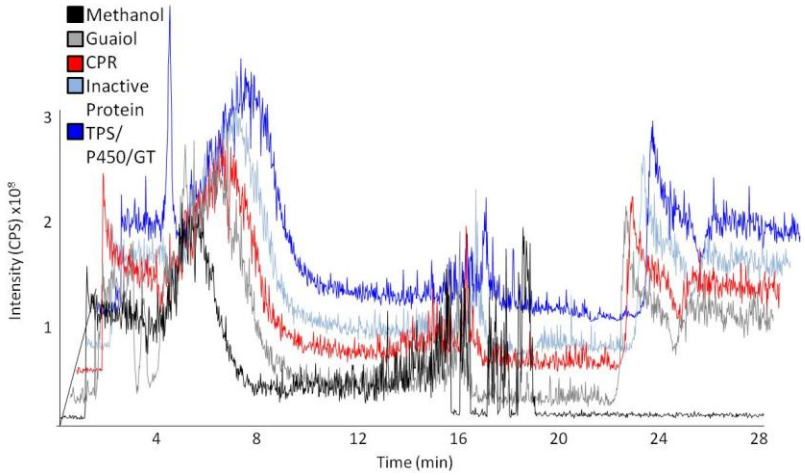


Figure S4.7: LC-MS analysis of Sf9 cell enzyme assay products. EPI scan at mass 431 Da. *Protein collected from Sf9 cells were combined in an* in vitro *enzyme assay. The activity of the terpene synthase (TPS), cytochrome P450 (P450), and galacturonosyltransferase (GT) on FPP was compared to the activity of the control cytochrome P450 reductase (CPR) and inactivated protein. CPS = counts per second.*

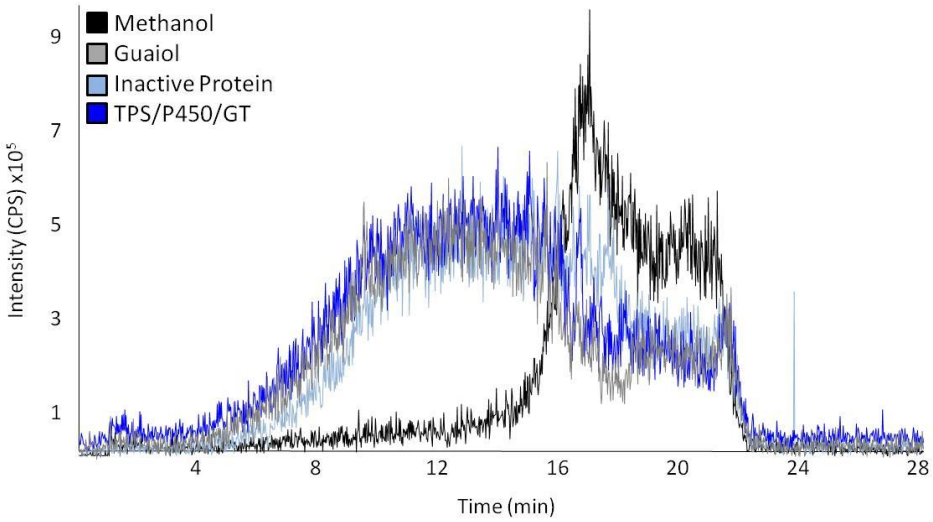


Figure S4.8: LC-MS analysis of Sf9 cell enzyme assay products. Neutral Loss scan at mass 194 Da. *Protein collected from Sf9 cells were combined in an* in vitro *enzyme assay. The activity of the terpene synthase (TPS), cytochrome P450 (P450), and galacturonosyltransferase (GT) on FPP was compared to the control inactivated protein. CPS = counts per second.*

4.8.2 Supplementary Tables

Table S4.1: Primer sequences

Gene Isolation from *S. bicolor* RNA		
Gene	**Primer Position**	**5' - 3' Sequence**
Sobic.001G339000	5' UTR	TGTGAGAGAGAGATGGCTGCTGCGA
Sobic.001G339000	3' UTR	AGGATGGTACCCTAGAAGGGAATCG
Sobic.001G339000	5' gene-specific with added BglII restriction site	AGAGAGATCTATGGCTGCTGCGAGAGAGGTTGATG
Sobic.001G339000	3' gene-specific with added EcoRI restriction site	AGAGGAATTCCTAGAAGGGAATCGGTTCATC
Sobic.001G338900	5' UTR	TCCATTTCCTCCAGGATCGGAATAG
Sobic.001G338900	3' UTR	CTTGGCTCAACTTGGTTGCATCTGC
Sobic.001G338900	5' gene-specific with added BglII restriction site	ATATATAGATCTATGGAGCTAATAAGCACAACCACTG
Sobic.001G338900	3' gene-specific with added EcoRI restriction site	CACACAGAATTCCTACATGGGTACTTGATACGGCGAG
Sobic.001G338400	5' UTR	GCTCTGGTTTAGTTCTTGCGT
Sobic.001G338400	3' UTR	TGAAGCATCCAGCTTCAATGCCAT
Sobic.001G338400	5' gene-specific with added BglII restriction site	ATATATAGATCTATGCTTCGTGGGGCGGGGCA
Sobic.001G338400	3' gene-specific with added XmaI restriction site	GTGTGTCCCGGGCTAATGCAACATACACTCTCTAACAT
Amplification of pET28a(+) multiple cloning site		
Gene	**Primer Position**	**5' - 3' Sequence**
MCS	5' with added KpnI restriction site	ATATAAGGTACCATCATCACAGCAGCGGCCTGGT
MCS	3' with added KpnI restriction site	CACACGGTACCTAGCAGCCGGATCTCAGTGGT
Amplification for insertion into pMDC32		
Gene	**Primer Position**	**5' - 3' Sequence**
Sobic.001G339000	5' with added SalI restriction site	TATATAGTCGACATGGCTGCTGCGAGAGAGGTTGATG
Sobic.001G339000	3' with added ApaI restriction site	GAGAGAGGGCCCCTAGAAGGGAATCGGTTCATCAAGC
Sobic.001G338900	5' with added SalI restriction site	ATATATGTCGACATGGAGCTAATAAGCACAACCACTG
Sobic.001G338900	3' with added ApaI restriction site	CACACAGGGCCCCTACATGGGTACTTGATACGGCGAG
Sobic.001G338400	5' with added SalI restriction site	ATATATGTCGACATGCTTCGTGGGGCGGGGCA
Sobic.001G338400	3' with added ApaI	CACACAGGGCCCCTAATGCAACATACACTCTCTA

	restriction site	ACAT
Confirmation of RNA expression/insertion into *A. tumefaciens/* baculovirus recombination		
Gene	**Primer Position**	**5' - 3' Sequence**
Sobic.001G339000	5' Mid-gene	AAGGATGAAGAGATTGGAG
Sobic.001G339000	3' Mid-gene	AGTTGACATTCCTTCTTCC
Sobic.001G338900	5' Mid-gene	ATCATCTGCTACGGCAAC
Sobic.001G338900	3' Mid-gene	GTGTTGTTGATTGGGATC
Sobic.001G338400	5' Mid-gene	GTTTCATGTGGTCACTGAC
Sobic.001G338400	3' Mid-gene	TTTCCACTGCTCCAATGAC
N. benthamiana TUA6	5' Mid-gene	ACAATTTACCCTTCACCACAG
N. benthamiana TUA6	3' Mid-gene	GGCTGATAGTTAATACCACACTTG
Amplification for insertion into pVL1392		
Gene	**Primer Position**	**5' - 3' Sequence**
Sobic.001G339000	5' with added BglII restriction site	ACACACAGATCTATGGCTGCTGCGAGAGAGGTTGATG
Sobic.001G339000	3' with added EcoRI restriction site	AGAGGAATTCCTAGAAGGGAATCGGTTCATC
Sobic.001G338900	5' with added BglII restriction site	ATATATAGATCTATGGAGCTAATAAGCACAACCACTG
Sobic.001G338900	3' with added EcoRI restriction site	CACACAGAATTCCTACATGGGTACTTGATACGGCGAG
Sobic.001G338400	5' with added BglII restriction site	ATATATAGATCTATGCTTCGTGGGGCGGGGCA
Sobic.001G338400	3' with added XmaI restriction site	GTGTGTCCCGGGCTAATGCAACATACACTCTCTAACAT

Table 4.2: PCR Parameters

Application	**Polymerase**	**PCR Parameters**
Amplification of gene cluster from *S. bicolor* cDNA	Q5 High-Fidelity DNA polymerase	98˚C for 30 sec, then 35 cycles of 98˚C for 10 sec, 58˚C for 20 sec, 72˚C for 1 min, and a final 5 min at 72˚C
Amplification of gene cluster from pJet cloning vector for transformation into pMDC32 and pVL1392	Q5 High-Fidelity DNA polymerase	98˚C for 30 sec, then 35 cycles of 98˚C for 10 sec, 58˚C for 20 sec, 72˚C for 1 min, and a final 5 min at 72˚C
Confirmation of Agrobacterium transformation	Taq DNA Polymerase	95˚C for 5 min, then 30 cycles of 94˚C for 30 sec, 52˚C for 30 sec, and 72˚C for 2 min, and a final 5 min at 72˚C
Confirmation of gene expression in *N. benthamiana*	Taq DNA Polymerase	95˚C for 5 min, then 30 cycles of 94˚C for 30 sec, 52˚C for 30 sec, and 72˚C for 2 min, and a final 5 min at 72˚C
Confirmation of baculovirus recombination	Q5 High-Fidelity DNA polymerase	98˚C for 30 sec, then 35 cycles of 98˚C for 10 sec, 58˚C for 20 sec, 72˚C for 1 min, and a final 5 min at 72˚C

Table 4.3: Masses targeted for LCMS and predicted structural modifications

Mass (Da)	Modification
223	H^+
225	Loss of a double bond
239	OH, H^+
416	OH+galacturonic acid, H^+
438	OH+OH+galacturonic acid, H^+
NL 176	Loss of galacturonic acid from glycosidic bond
NL 194	Loss of complete galacturonic acid

Chapter 5: Concluding remarks and future directions

Upon completion of this project, a putative *S. bicolor* terpene synthase was expressed in *E. coli* and found to produce an unidentified terpene product when combined with FPP. The activity of the terpene synthase on FPP suggested that the enzyme produced a sesquiterpene. GC-MS and direct infusion MS analysis determined that the unknown terpene product was similar in composition to sesquiterpene alcohols, such as guaiol and β-eudesmol, with a molecular weight of 222.2 Da and the molecular formula $C_{15}H_{26}O$. Further structural analysis, such as NMR, would confirm the composition of the unknown terpene product.

The terpene synthase gene was also expressed in two heterologous expression systems, *N. benthamiana* and Sf9 insect cells, alongside a cytochrome P450 and galacturonosyltransferase hypothesized to be clustered with the terpene synthase. The goal of these experiments was to determine whether the clustered genes modify the terpene product. Enzyme assays of protein produced in Sf9 cells generated no observable metabolites. Replication of this experiment with modified procedures, such as an earlier harvest of cells, could improve protein activity. This would be an ideal system for identifying the products of the putative gene cluster as Sf9 cells have less interference in analysis by background metabolism than *N. benthamiana*.

An unknown metabolite detected in transfected *N. benthamiana* tissues was a potential result of the activity of the cluster. The compound, which was only detectable using Q Exactive LC-MS with a HILIC column, most closely annotated as the flavonoid ikarisoside F. However, the difference in exact mass indicated that the unknown metabolite and ikarisoside F are not identical. Mass spectra analysis determined that the unknown metabolite likely contains between 23 and 31 carbons, which is within the expected range for a sesquiterpene modified by a sugar group. Further analysis was limited by the rapid degradation of the metabolite post-extraction and replication of the experiment coupled with a targeted MS approach could potentially better elucidate the structure.

Transgenic *N. benthamiana* that expresses the gene cluster and overexpresses the precursor FPP may be necessary to generate large volumes of the metabolite of interest. Transgenic *N. benthamiana* has been used to characterize terpenoids in other species such as *Saliva officinalis* , and while the time from infection to collection is much longer than during transient expression, stable transgenic lines can be tested multiple times and provide more consistent data (Ali et al., 2017).

Long-term goals for this project could include determining where in the plant this terpene synthase is expressed. Gene expression data indicates that it is highly expressed in root tissues of young plants (Makita et al., 2015), but more specific information about timing of expression and location could be used to determine terpene function *in planta*. One possible avenue to pursue is the use of matrix-assisted laser desorption ionization (MALDI) mass spectrometry to target specific masses in root tissues. MALDI imaging has been used to localize the activity of *Vitex agnus-castis* diterpenoids to fruit and leaf trichomes, resulting in functional characterization of the genes and identification of additional enzymes involved in their biosynthetic pathway (Heskes et al., 2018).

The protocols developed in this study for terpene characterization and gene cluster analysis can be applied to other clusters and metabolites. There are many uncharacterized terpene synthases present in *S. bicolor* and determination of their products and biosynthesis pathway could lead to advancements in terpene production for fuels and other purposes. Engineering of *S. bicolor* as a better source of food or fuel will not be possible unless more resources are devoted to the elucidation and manipulation of its biosynthetic pathways.

References

Achnine, L., Huhman, D.V., Farag, M.A., Sumner, L.W., Blount, J.W., and Dixon, R.A. (2005). Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume Medicago truncatula. Plant J. *41*, 875–887.

Aharoni, A., Giri, A.P., Verstappen, F.W.A., Bertea, C.M., Sevenier, R., Sun, Z., Jongsma, M.A., Schwab, W., and Bouwmeester, H.J. (2004). Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species. Plant Cell *16*, 3110–3131.

Ajikumar, P.K., Xiao, W.-H., Tyo, K.E.J., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T.H., Pfeifer, B., and Stephanopoulos, G. (2010). Isoprenoid Pathway Optimization for Taxol Precursor Overproduction in Escherichia coli. Science *330*, 70–74.

Akiyama, K., Matsuzaki, K., and Hayashi, H. (2005). Plant sesquiterpenes induce hyphal branching in arbuscular mycorrhizal fungi. Nature *435*, 824–827.

Alder, L., Greulich, K., Kempe, G., and Vieth, B. (2006). Residue analysis of 500 high priority pesticides: Better by GC–MS or LC–MS/MS? Mass Spectrom. Rev. *25*, 838–865.

Ali, M., Li, P., She, G., Chen, D., Wan, X., and Zhao, J. (2017). Transcriptome and metabolite analyses reveal the complex metabolic genes involved in volatile terpenoid biosynthesis in garden sage (Salvia officinalis). Sci. Rep. *7*.

Alquézar, B., Rodríguez, A., de la Peña, M., and Peña, L. (2017). Genomic Analysis of Terpene Synthase Family and Functional Characterization of Seven Sesquiterpene Synthases from Citrus sinensis. Front. Plant Sci. *8*.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Augustin, J.M., Higashi, Y., Feng, X., and Kutchan, T.M. (2015a). Production of mono- and sesquiterpenes in Camelina sativa oilseed. Planta *242*, 693–708.

Augustin, M.M., Ruzicka, D.R., Shukla, A.K., Augustin, J.M., Starks, C.M., O'Neil-Johnson, M., McKain, M.R., Evans, B.S., Barrett, M.D., Smithson, A., et al. (2015b). Elucidating steroid alkaloid biosynthesis in Veratrum californicum: production of verazine in Sf9 cells. Plant J. *82*, 991–1003.

Awika, J.M., and Rooney, L.W. (2004). Sorghum phytochemicals and their potential impact on human health. Phytochemistry *65*, 1199–1221.

Bach, S.S., Bassard, J.-É., Andersen-Ranberg, J., Møldrup, M.E., Simonsen, H.T., and Hamberger, B. (2014). High-Throughput Testing of Terpenoid Biosynthesis Candidate Genes Using Transient Expression in Nicotiana benthamiana. SpringerLink 245–255.

Ballouz, S., Francis, A.R., Lan, R., and Tanaka, M.M. (2010). Conditions for the Evolution of Gene Clusters in Bacterial Genomes. PLOS Comput. Biol. *6*, e1000672.

Belgacem, M.N., and Gandini, A. (2011). Monomers, Polymers and Composites from Renewable Resources (Elsevier).

Bohlmann, J., and Keeling, C.I. (2008). Terpenoid biomaterials. Plant J. Cell Mol. Biol. *54*, 656–669.

Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L.A., and Martin, G.B. (2012). A Draft Genome Sequence of Nicotiana benthamiana to Enhance Molecular Plant-Microbe Biology Research. Mol. Plant. Microbe Interact. *25*, 1523–1530.

Boutanaev, A.M., Moses, T., Zi, J., Nelson, D.R., Mugford, S.T., Peters, R.J., and Osbourn, A. (2015). Investigation of terpene diversification across multiple sequenced plant genomes. Proc. Natl. Acad. Sci. *112*, E81–E88.

Boycheva, S., Daviet, L., Wolfender, J.-L., and Fitzpatrick, T.B. (2014). The rise of operon-like gene clusters in plants. Trends Plant Sci. *19*, 447–459.

Brakhage, A.A., and Schroeckh, V. (2011). Fungal secondary metabolites – Strategies to activate silent gene clusters. Fungal Genet. Biol. *48*, 15–22.

Brennan, T.C.R., Turner, C.D., Krömer, J.O., and Nielsen, L.K. (2012). Alleviating monoterpene toxicity using a two-phase extractive fermentation for the bioproduction of jet fuel mixtures in Saccharomyces cerevisiae. Biotechnol. Bioeng. *109*, 2513–2522.

Brückner, K., and Tissier, A. (2013). High-level diterpene production by transient expression in Nicotiana benthamiana. Plant Methods *9*, 46.

Cane, D.E., and Ikeda, H. (2012). Exploration and mining of the bacterial terpenome. Acc. Chem. Res. *45*, 463–472.

Chakravarti, A., Phillips, J.A., Mellits, K.H., Buetow, K.H., and Seeburg, P.H. (1984). Patterns of polymorphism and linkage disequilibrium suggest independent origins of the human growth hormone gene cluster. Proc. Natl. Acad. Sci. *81*, 6085–6089.

Chang, J.C., Lee, S.J., Kim, J.S., Wang, C.H., and Nai, Y.S. (2018). Transient Expression of Foreign Genes in Insect Cells (sf9) for Protein Functional Assay. J. Vis. Exp. JoVE.

Chen, F., Ro, D.-K., Petri, J., Gershenzon, J., Bohlmann, J., Pichersky, E., and Tholl, D. (2004). Characterization of a Root-Specific Arabidopsis Terpene Synthase Responsible for the Formation of the Volatile Monoterpene 1,8-Cineole. Plant Physiol. *135*, 1956–1966.

Chen, X., Köllner, T.G., Jia, Q., Norris, A., Santhanam, B., Rabe, P., Dickschat, J.S., Shaulsky, G., Gershenzon, J., and Chen, F. (2016). Terpene synthase genes in eukaryotes beyond plants and fungi: Occurrence in social amoebae. Proc. Natl. Acad. Sci. *113*, 12132–12137.

Cheng, F., Shen, J., Luo, X., Zhu, W., Gu, J., Ji, R., Jiang, H., and Chen, K. (2002). Molecular docking and 3-D-QSAR studies on the possible antimalarial mechanism of artemisinin analogues. Bioorg. Med. Chem. *10*, 2883–2891.

Chuck, C.J., and Donnelly, J. (2014). The compatibility of potential bioderived fuels with Jet A-1 aviation kerosene. Appl. Energy.

Ciriminna, R., Lomeli-Rodriguez, M., Carà, P.D., Lopez-Sanchez, J.A., and Pagliaro, M. (2014). Limonene: a versatile chemical of the bioeconomy. Chem. Commun. *50*, 15288–15296.

Dairi, T., Hamano, Y., Kuzuyama, T., Itoh, N., Furihata, K., and Seto, H. (2001). Eubacterial diterpene cyclase genes essential for production of the isoprenoid antibiotic terpentecin. J. Bacteriol. *183*, 6085–6094.

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

Degenhardt, J., Gershenzon, J., Baldwin, I.T., and Kessler, A. (2003). Attracting friends to feast on foes: engineering terpene emission to make crop plants more attractive to herbivore enemies. Curr. Opin. Biotechnol. *14*, 169–176.

Díaz Chávez, M.L., Rolf, M., Gesell, A., and Kutchan, T.M. (2011). Characterization of two methylenedioxy bridge-forming cytochrome P450-dependent enzymes of alkaloid formation in the Mexican prickly poppy Argemone mexicana. Arch. Biochem. Biophys. *507*, 186–193.

Dickschat, J.S., Rinkel, J., Rabe, P., Beyraghdar Kashkooli, A., and Bouwmeester, H.J. (2017). 18-Hydroxydolabella-3,7-diene synthase – a diterpene synthase from Chitinophaga pinensis. Beilstein J. Org. Chem. *13*, 1770–1780.

Dunlop, M.J., Dossani, Z.Y., Szmidt, H.L., Chu, H.C., Lee, T.S., Keasling, J.D., Hadi, M.Z., and Mukhopadhyay, A. (2011). Engineering microbial biofuel tolerance and export using efflux pumps. Mol. Syst. Biol. *7*, 487.

Dunoyer, P., Lecellier, C.-H., Parizotto, E.A., Himber, C., and Voinnet, O. (2004). Probing the MicroRNA and Small Interfering RNA Pathways with Virus-Encoded Suppressors of RNA Silencing. Plant Cell *16*, 1235–1250.

Dürr, C., Schnell, H.-J., Luzhetskyy, A., Murillo, R., Weber, M., Welzel, K., Vente, A., and Bechthold, A. (2006). Biosynthesis of the Terpene Phenalinolactone in Streptomyces sp. Tü6071: Analysis of the Gene Cluster and Generation of Derivatives. Chem. Biol. *13*, 365–377.

Dutartre, L., Hilliou, F., and Feyereisen, R. (2012). Phylogenomics of the benzoxazinoid biosynthetic pathway of Poaceae: gene duplications and origin of the Bx cluster. BMC Evol. Biol. *12*, 64.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Elbehri, A., Segerstedt, A., and Liu, P. (2013). Biofuels and the sustainability challenge: a global assessment of sustainability issues, trends and policies for biofuels and related feedstocks. Biofuels Sustain. Chall. Glob. Assess. Sustain. Issues Trends Policies Biofuels Relat. Feedstock.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. *300*, 1005–1016.

Farhi, M., Marhevka, E., Ben-Ari, J., Algamas-Dimantov, A., Liang, Z., Zeevi, V., Edelbaum, O., Spitzer-Rimon, B., Abeliovich, H., Schwartz, B., et al. (2011). Generation of the potent anti-malarial drug artemisinin in tobacco. Nat. Biotechnol. *29*, 1072–1074.

Fesenko, E., and Edwards, R. (2014). Plant synthetic biology: a new platform for industrial biotechnology. J. Exp. Bot. *65*, 1–1.

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

Field, B., Fiston-Lavier, A.-S., Kemen, A., Geisler, K., Quesneville, H., and Osbourn, A.E. (2011). Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc. Natl. Acad. Sci. U. S. A. *108*, 16116–16121.

Fischer, M.J.C., Meyer, S., Claudel, P., Perrin, M., Ginglinger, J.F., Gertz, C., Masson, J.E., Werck-Reinhardt, D., Hugueney, P., and Karst, F. (2013). Specificity of Ocimum basilicum geraniol synthase modified by its expression in different heterologous systems. J. Biotechnol. *163*, 24–29.

Fischer, R., Stoger, E., Schillberg, S., Christou, P., and Twyman, R.M. (2004). Plant-based production of biopharmaceuticals. Curr. Opin. Plant Biol. *7*, 152–158.

Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grün, S., Winklmair, A., Eisenreich, W., Bacher, A., Meeley, R.B., Briggs, S.P., et al. (1997). Analysis of a Chemical Plant Defense Mechanism in Grasses. Science *277*, 696–699.

Gao, C.W., Vandeputte, A.G., Yee, N.W., Green, W.H., Bonomi, R.E., Magoon, G.R., Wong, H.-W., Oluwole, O.O., Lewis, D.K., Vandewiele, N.M., et al. (2015). JP-10 combustion studied with shock tube experiments and modeled with automatic reaction mechanism generation. Combust. Flame *162*, 3115–3129.

George, K.W., Alonso-Gutierrez, J., Keasling, J.D., and Lee, T.S. (2015). Isoprenoid drugs, biofuels, and chemicals--artemisinin, farnesene, and beyond. Adv. Biochem. Eng. Biotechnol. *148*, 355–389.

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

Gesell, A., Rolf, M., Ziegler, J., Chávez, M.L.D., Huang, F.-C., and Kutchan, T.M. (2009). CYP719B1 Is Salutaridine Synthase, the C-C Phenol-coupling Enzyme of Morphine Biosynthesis in Opium Poppy. J. Biol. Chem. *284*, 24432–24442.

Gotoh, T., Miyazaki, Y., Sato, W., Kikuchi, K., and Bentley, W.E. (2001). Proteolytic activity and recombinant protein production in virus-infected Sf-9 insect cell cultures supplemented with carboxyl and cysteine protease inhibitors. J. Biosci. Bioeng. *92*, 248–255.

Gusella, J., Varsanyi-Breiner, A., Kao, F.T., Jones, C., Puck, T.T., Keys, C., Orkin, S., and Housman, D. (1979). Precise localization of human beta-globin gene complex on chromosome 11. Proc. Natl. Acad. Sci. U. S. A. *76*, 5239–5242.

Hayashi, Y., Matsuura, N., Toshima, H., Itoh, N., Ishikawa, J., Mikami, Y., and Dairi, T. (2008). Cloning of the gene cluster responsible for the biosynthesis of brasilicardin A, a unique diterpenoid. J. Antibiot. (Tokyo) *61*, 164–174.

Heskes, A.M., Sundram, T.C.M., Boughton, B.A., Bjerg Jensen, N., Hansen, N.L., Crocoll, C., Cozzi, F., Rasmussen, S., Hamberger, B., Hamberger, B., et al. (2018). Biosynthesis of bioactive diterpenoids in the medicinal plant Vitex agnus-castus. Plant J. n/a-n/a.

Higdon, J.V., and Frei, B. (2006). Coffee and health: a review of recent human research. Crit. Rev. Food Sci. Nutr. *46*, 101–123.

Juyal, D., Thawani, V., Thaledi, S., and Joshi, M. (2014). Ethnomedical Properties of Taxus Wallichiana Zucc. (Himalayan Yew). J. Tradit. Complement. Med. *4*, 159–161.

Kappers, I.F., Aharoni, A., van Herpen, T.W.J.M., Luckerhoff, L.L.P., Dicke, M., and Bouwmeester, H.J. (2005). Genetic engineering of terpenoid metabolism attracts bodyguards to Arabidopsis. Science *309*, 2070–2072.

Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780.

Kautsar, S.A., Suarez Duran, H.G., Blin, K., Osbourn, A., and Medema, M.H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. *45*, W55–W63.

Keller, N.P., and Hohn, T.M. (1997). Metabolic Pathway Gene Clusters in Filamentous Fungi. Fungal Genet. Biol. *21*, 17–29.

Kellner, F., Kim, J., Clavijo, B.J., Hamilton, J.P., Childs, K.L., Vaillancourt, B., Cepela, J., Habermann, M., Steuernagel, B., Clissold, L., et al. (2015). Genome-guided investigation of plant natural product biosynthesis. Plant J. *82*, 680–692.

Khaldi, N., Collemare, J., Lebrun, M.-H., and Wolfe, K.H. (2008). Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome Biol. *9*, R18.

Kilgore, M.B., Augustin, M.M., May, G.D., Crow, J.A., and Kutchan, T.M. (2016). CYP96T1 of Narcissus sp. aff. pseudonarcissus Catalyzes Formation of the Para-Para' C-C Phenol Couple in the Amaryllidaceae Alkaloids. Front. Plant Sci. *7*.

Kingston, D.G.I. (2007). The Shape of Things to Come: Structural and Synthetic Studies of Taxol and Related Compounds. Phytochemistry *68*, 1844–1854.

Koncz, C., and Schell, J. (1986). The promoter of T<Subscript>L</Subscript>-DNA gene <Emphasis Type="Italic">5</Emphasis> controls the tissue-specific expression of chimaeric genes carried by a novel type of <Emphasis Type="Italic">Agrobacterium</Emphasis> binary vector. Mol. Gen. Genet. MGG *204*, 383–396.

Kristensen, C., Morant, M., Olsen, C.E., Ekstrøm, C.T., Galbraith, D.W., Lindberg Møller, B., and Bak, S. (2005). Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal

inadvertent effects on the metabolome and transcriptome. Proc. Natl. Acad. Sci. U. S. A. *102*, 1779–1784.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. *33*, 1870–1874.

Landais, I., Ogliastro, M., Mita, K., Nohata, J., López-Ferber, M., Duonor-Cérutti, M., Shimada, T., Fournier, P., and Devauchelle, G. (2003). Annotation pattern of ESTs from Spodoptera frugiperda Sf9 cells and analysis of the ribosomal protein genes reveal insect-specific features and unexpectedly low codon usage bias. Bioinformatics *19*, 2343–2350.

Lawrence, J.G., and Roth, J.R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics *143*, 1843–1860.

Leavell, M.D., McPhee, D.J., and Paddon, C.J. (2016). Developing fermentative terpenoid production for commercial usage. Curr. Opin. Biotechnol. *37*, 114–119.

Lima, P.S.S., Lucchese, A.M., Araújo-Filho, H.G., Menezes, P.P., Araújo, A.A.S., Quintans-Júnior, L.J., and Quintans, J.S.S. (2016). Inclusion of terpenes in cyclodextrins: Preparation, characterization and pharmacological approaches. Carbohydr. Polym. *151*, 965–987.

Lin, H.-C., Chooi, Y.-H., Dhingra, S., Xu, W., Calvo, A.M., and Tang, Y. (2013). The Fumagillin Biosynthetic Gene Cluster in Aspergillus fumigatus Encodes a Cryptic Terpene Cyclase Involved in the Formation of β-trans-Bergamotene. J. Am. Chem. Soc. *135*, 4616–4619.

Makita, Y., Shimada, S., Kawashima, M., Kondou-Kuriyama, T., Toyoda, T., and Matsui, M. (2015). MOROKOSHI: transcriptome database in Sorghum bicolor. Plant Cell Physiol. *56*, e6.

Malpartida, F., and Hopwood, D.A. (1984). Molecular cloning of the whole biosynthetic pathway of a Streptomyces antibiotic and its expression in a heterologous host. Nature *309*, 462–464.

Martin, T., Biruma, M., Fridborg, I., Okori, P., and Dixelius, C. (2011). A highly conserved NB-LRR encoding gene cluster effective against Setosphaeria turcica in sorghum. BMC Plant Biol. *11*, 151.

Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., and Keasling, J.D. (2003). Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. Nat. Biotechnol. *21*, 796–802.

Moriyama, E.N., and Powell, J.R. (1998). Gene length and codon usage bias in Drosophila melanogaster , Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. *26*, 3188–3193.

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

Nagegowda, D.A. (2010). Plant volatile terpenoid metabolism: Biosynthetic genes, transcriptional regulation and subcellular compartmentation. FEBS Lett. *584*, 2965–2973.

Naparstek, S., Guan, Z., and Eichler, J. (2012). A predicted geranylgeranyl reductase reduces the ω-position isoprene of dolichol phosphate in the halophilic archaeon, Haloferax volcanii. Biochim. Biophys. Acta *1821*, 923–933.

NCBI Resource Coordinators (2017). Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. *45*, D12–D17.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I.V., and Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. *42*, D26-31.

Norris, A. (2013). Genomic Analysis and Functional Characterization of the Terpene Synthase Gene Family in Brachypodium distachyon. Masters Theses.

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

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

Osbourn, A. (2010). Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends Genet. *26*, 449–457.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature *457*, 551–556.

Pearson, W.R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *0 3*.

Peralta-Yahya, P.P., Ouellet, M., Chan, R., Mukhopadhyay, A., Keasling, J.D., and Lee, T.S. (2011). Identification and microbial production of a terpene-based advanced biofuel. Nat. Commun. *2*, 483.

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods *8*, 785–786.

Phillips, D.R., Rasbery, J.M., Bartel, B., and Matsuda, S.P. (2006). Biosynthetic diversity in plant triterpene cyclization. Curr. Opin. Plant Biol. *9*, 305–314.

Prasad, S., Singh, A., Jain, N., and Joshi, H.C. (2007). Ethanol Production from Sweet Sorghum Syrup for Utilization as Automotive Fuel in India. Energy Fuels *21*, 2415–2420.

Priya, P., Yadav, A., Chand, J., and Yadav, G. (2018). Terzyme: a tool for identification and analysis of the plant terpenome. Plant Methods *14*.

Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R., and Osbourn, A. (2004). A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants. Proc. Natl. Acad. Sci. U. S. A. *101*, 8233–8238.

Rasmann, S., Köllner, T.G., Degenhardt, J., Hiltpold, I., Toepfer, S., Kuhlmann, U., Gershenzon, J., and Turlings, T.C.J. (2005). Recruitment of entomopathogenic nematodes by insect-damaged maize roots. Nature *434*, 732–737.

Reade, L. (2011). Festive fragrances. Chem. Ind. *75*, 17–20.

Richards, T.A., Soanes, D.M., Foster, P.G., Leonard, G., Thornton, C.R., and Talbot, N.J. (2009). Phylogenomic Analysis Demonstrates a Pattern of Rare and Ancient Horizontal Gene Transfer between Plants and Fungi. Plant Cell Online *21*, 1897–1911.

Ro, D.-K., Ehlting, J., Keeling, C.I., Lin, R., Mattheus, N., and Bohlmann, J. (2006). Microarray expression profiling and functional characterization of AtTPS genes: duplicated Arabidopsis thaliana sesquiterpene synthase genes At4g13280 and At4g13300 encode root-specific and wound-inducible (Z)-gamma-bisabolene synthases. Arch. Biochem. Biophys. *448*, 104–116.

Rozman, D., Strömstedt, M., Tsui, L.-C., Scherer, S.W., and Waterman, M.R. (1996). Structure and Mapping of the Human Lanosterol 14α- Demethylase Gene (CYP51) Encoding the Cytochrome P450 Involved in Cholesterol Biosynthesis; Comparison of Exon/Intron Organization with other Mammalian and Fungal CYP Genes. Genomics *38*, 371–381.

Salcedo, R.G., Olano, C., Gómez, C., Fernández, R., Braña, A.F., Méndez, C., de la Calle, F., and Salas, J.A. (2016). Characterization and engineering of the biosynthesis gene cluster for antitumor macrolides PM100117 and PM100118 from a marine actinobacteria: generation of a novel improved derivative. Microb. Cell Factories *15*, 44.

Sauerschnig, C., Doppler, M., Bueschl, C., and Schuhmacher, R. (2017). Methanol Generates Numerous Artifacts during Sample Extraction and Storage of Extracts in Metabolomics Research. Metabolites *8*, 1.

Schmidt-Dannert, C. (2015). Biosynthesis of terpenoid natural products in fungi. Adv. Biochem. Eng. Biotechnol. *148*, 19–61.

Sheludko, Y.V. (2010). Recent advances in plant biotechnology and genetic engineering for production of secondary metabolites. Cytol. Genet. *44*, 52–60.

Song, R., Segal, G., and Messing, J. (2004). Expression of the sorghum 10-member kafirin gene cluster in maize endosperm. Nucleic Acids Res. *32*, e189–e189.

Takos, A.M., Knudsen, C., Lai, D., Kannangara, R., Mikkelsen, L., Motawia, M.S., Olsen, C.E., Sato, S., Tabata, S., Jørgensen, K., et al. (2011). Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in Lotus japonicus and suggests the repeated evolution of this chemical defence pathway. Plant J. *68*, 273–286.

Toyomasu, T., Nakaminami, K., Toshima, H., Mie, T., Watanabe, K., Ito, H., Matsui, H., Mitsuhashi, W., Sassa, T., and Oikawa, H. (2004). Cloning of a Gene Cluster Responsible for the Biosynthesis of Diterpene Aphidicolin, a Specific Inhibitor of DNA Polymerase α. Biosci. Biotechnol. Biochem. *68*, 146–152.

Turner, M.F., Heuberger, A.L., Kirkwood, J.S., Collins, C.C., Wolfrum, E.J., Broeckling, C.D., Prenni, J.E., and Jahn, C.E. (2016). Non-targeted Metabolomics in Diverse Sorghum Breeding Lines Indicates Primary and Secondary Metabolite Profiles Are Associated with Plant Biomass Accumulation and Photosynthesis. Front. Plant Sci. *7*, 953.

Wang, C., Yoon, S.-H., Shah, A.A., Chung, Y.-R., Kim, J.-Y., Choi, E.-S., Keasling, J.D., and Kim, S.-W. (2010). Farnesol production from Escherichia coli by harnessing the exogenous mevalonate pathway. Biotechnol. Bioeng. *107*, 421–429.

Wenzl, P., Wong, L., Kwang-won, K., and Jefferson, R.A. (2005). A functional screen identifies lateral transfer of beta-glucuronidase (gus) from bacteria to fungi. Mol. Biol. Evol. *22*, 308–316.

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

Wortmann, C.S., Liska, A.J., Ferguson, R.B., Lyon, D.J., Klein, R.N., and Dweikat, I. (2010). Dryland Performance of Sweet Sorghum and Grain Crops for Biofuel in Nebraska. Agron. J. *102*, 319–326.

Wu, S., Schalk, M., Clark, A., Brandon Miles, R., Coates, R., and Chappell, J. (2006). Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. Nat. Biotechnol. *24*, 1441–1447.

Xu, X., Iba, M.M., and Weisel, C.P. (2004). Simultaneous and Sensitive Measurement of Anabasine, Nicotine, and Nicotine Metabolites in Human Urine by Liquid Chromatography–Tandem Mass Spectrometry. Clin. Chem. *50*, 2323–2330.

Yamada, Y., Kuzuyama, T., Komatsu, M., Shin-ya, K., Omura, S., Cane, D.E., and Ikeda, H. (2015). Terpene synthases are widely distributed in bacteria. Proc. Natl. Acad. Sci. *112*, 857–862.

Yeo, Y.-S., Nybo, S.E., Chittiboyina, A.G., Weerasooriya, A.D., Wang, Y.-H., Góngora-Castillo, E., Vaillancourt, B., Buell, C.R., DellaPenna, D., Celiz, M.D., et al. (2013). Functional identification of valerena-1,10-diene synthase, a terpene synthase catalyzing a unique chemical cascade in the biosynthesis of biologically active sesquiterpenes in Valeriana officinalis. J. Biol. Chem. *288*, 3163–3173.

Yu, F., and Utsumi, R. (2009). Diversity, regulation, and genetic manipulation of plant mono- and sesquiterpenoid biosynthesis. Cell. Mol. Life Sci. *66*, 3043–3052.

Yuan, J.S., Köllner, T.G., Wiggins, G., Grant, J., Degenhardt, J., and Chen, F. (2008). Molecular and genomic basis of volatile-mediated indirect defense against insects in rice. Plant J. Cell Mol. Biol. *55*, 491–503.

Zerbe, P., Hamberger, B., Yuen, M.M.S., Chiang, A., Sandhu, H.K., Madilao, L.L., Nguyen, A., Hamberger, B., Bach, S.S., and Bohlmann, J. (2013). Gene Discovery of Modular Diterpene Metabolism in Nonmodel Systems. Plant Physiol. *162*, 1073–1091.

Zhang, X., Li, Y., Yang, X., Wang, K., Ni, J., and Qu, X. (2005). Inhibitory effect of Epimedium extract on S-adenosyl-L-homocysteine hydrolase and biomethylation. Life Sci. *78*, 180–186.

Zhou, K., and Peters, R.J. (2009). Investigating the Conservation Pattern of a Putative Second Terpene Synthase Divalent Metal Binding Motif in Plants. Phytochemistry *70*, 366–369.

Zhuang, X., Köllner, T.G., Zhao, N., Li, G., Jiang, Y., Zhu, L., Ma, J., Degenhardt, J., and Chen, F. (2012). Dynamic evolution of herbivore-induced sesquiterpene biosynthesis in sorghum and related grass crops. Plant J. *69*, 70–80.