

Computational Identification of Terpene Synthase Genes and Their Evolutionary Analysis

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Qidong Jia

May 2016

© by Qidong Jia, 2016

All Rights Reserved.

dedicated to my beloved family

Acknowledgements

First, I would like to thank my advisor Dr. Feng Chen for his consistent support, guidance and confidence in me during my dissertation work. He is a person with great passion and I sincerely appreciate his help in developing my scientific attitude. I also sincerely appreciate Dr. Albrecht Von Arnim for his tremendous mentorship, encouragement and support. I would also like to express my sincere gratitude to my statistics advisor Dr. Robert Mee for his detailed guidance and advice.

Second, I'm very grateful to my committee members, Dr. Brian C. O'Meara, Dr. Gerald A. Tuskan and Dr. Xiaohan Yang for their time, guidance and expertise throughout my dissertation.

Finally, and most importantly, my upmost gratitude goes to my parents, parents-in-law and my wife, none of this would have been possible without their unconditional support, sacrifice, love and faith in me.

Abstract

Terpenoids, the largest and most structurally and functionally diverse class of natural compounds on earth, are mostly synthesized by plants to be involved in various plant-environment interactions. Some terpenoids are classified as primary metabolites essential for plant growth and development. Terpene synthases (TPSs), the key enzymes for terpenoid biosynthesis, are the major determinant of the tremendous diversity of terpenoid carbon skeletons. The TPS genes represent a mid-size family of about 30-100 functional genes in almost all major sequenced plant genomes. TPSs are also found in fungi and bacteria, but microbial TPS genes share low levels of sequence similarity and different patterns of gene structure with their plant counterparts. Although one common ancestor theory has been suggested and supported by studies from model species, the evolution of plant terpene synthase genes and the evolutionary relationships among terpene synthase genes in plants, bacteria and fungi are still unclear. The recent discovery of microbial-type TPS genes in *Selaginella muellendorffii* makes it even more confusing. The goal of this dissertation project is to study the mechanisms that govern the dynamic evolution of terpene synthase genes using comparative genomics methods. Here, we carried out

a large-scale screen to identify terpene synthases in plants (transcriptomes for over 1000 plant species sequenced by the OneKP project and selected genomes from non-seed plants), fungi and bacteria species (sequenced genomes in JGI). Several important discoveries were made by analyzing the data: (1) the microbial-type TPS genes are widely and specifically distributed in non-seed land plants; (2) HGT from bacterial to fungi in TPS gene family is identified; (3) a new subfamily x is identified and new insights into the subfamily classification of TPSs are reported by including TPSs identified from large-scale non-seed plant species; (4) the distribution and genomic organization of four types of TPSs in fungi are characterized. These important findings will help us better understand the evolution of plant secondary metabolism, especially for basal land plants.

Table of Contents

1	Introduction and Literature Review	1
1.1	Terpenoids - Natural Functions and Industrial Uses	2
1.2	Biosynthesis of Terpenoids	3
1.3	Terpene Synthases	6
1.4	Origin and Evolution of Terpene Synthase Genes	8
1.5	Objectives	10
1.6	Bibliography	11
1.7	Appendix	16
2	Evolution of Typical Plant Terpene Synthase Genes in Non-Seed Plants	19
2.1	Abstract	20
2.2	Introduction	21
2.3	Results and Discussion	23
2.3.1	Identification of Terpene Synthase Genes from Genomes of Horn- wort, Moss, Liverwort and Two Ferns	23

2.3.2	Identification of Terpene Synthases from Non-Seed Plants Transcriptomes	24
2.3.3	Search for Terpene Synthase Genes in Green Algae	24
2.3.4	Phylogenetic Analysis of TPSs from Non-Seed Plants with TPSs from Selected Seed Plants	25
2.4	Conclusions	28
2.5	Materials and Methods	30
2.5.1	Data Retrieval, Management and Taxonomy Classification	30
2.5.2	Identification of Typical Plant Terpene Synthases from Transcriptomes of 324 Non-Seed Plants	31
2.5.3	Assembly of Hornwort <i>Anthoceros punctatus</i> Genome and Identification of Terpene Synthases	31
2.5.4	Identification of Terpene Synthases from <i>Marchantia polymorpha</i> and <i>Sphagnum fallax</i>	32
2.5.5	Identification of Terpene Synthases from <i>Salvinia cucullata</i> and <i>Azolla superorganism</i>	33
2.5.6	Identification of Terpene Synthases from <i>Lygodium japonicum</i>	33
2.5.7	Search for Terpene Synthase Genes in Twelve Algal Genomes and Transcriptomes	34
2.5.8	TPSs on JBrowse	34
2.5.9	Phylogenetic Analyses of Terpene Synthases	34
2.6	Bibliography	36

2.7	Appendix	42
3	Microbial Type Terpene Synthase Genes Occur Widely and Specifically in	
	Non-seed Land Plants	60
3.1	Abstract	62
3.2	Introduction	63
3.3	Results and Discussion	65
3.3.1	Terpene Synthase Genes of Microbial Type are Highly Enriched in the Transcriptomes of Non-Seed Land Plants	65
3.3.2	The Majority of <i>MTPSL</i> Genes Identified from Plant Transcrip- tomes Forms Four Groups Clustered with Either Fungal or Bac- terial Terpene Synthases	66
3.3.3	The Majority of <i>MTPSL</i> Genes Identified from Plant Transcrip- tomes are Plant Genes	67
3.3.4	Evolutionary Implications of Non-Seed Plant-Specific MTPSLs . . .	71
3.3.5	Biochemical Function of Selected MTPSLs: Diversity of Activities .	74
3.4	Conclusions	75
3.5	Materials and Methods	76
3.5.1	Identification of Terpene Synthases of Microbial Type from Tran- scriptomes and Sequenced Genomes	76
3.5.2	Assembly of Hornwort <i>Anthoceros punctatus</i> Genome and Identi- fication of <i>MTPSL</i> Genes	78

3.5.3	Phylogenetic Analyses of Terpene Synthases	78
3.5.4	Plant Material, Genomic DNA Isolation and PCR	79
3.6	Acknowledgements	80
3.7	Bibliography	81
3.8	Appendix	88
4	Horizontal Gene Transfer of Terpene Synthase Genes from Bacteria to Fungi	103
4.1	Abstract	105
4.2	Introduction	106
4.3	Results	109
4.3.1	Analysis of Bacterial and Fungal Terpene Synthase Genes Suggests Possible HGT Events from Bacteria to Fungi	109
4.3.2	The Apparent Orthologs of <i>BTPSL</i> were Identified in a Group of Entomopathogenic Fungi	110
4.3.3	Collinearity for the Genome Region Containing the <i>BTPSL</i> and the Identification of Neighbor Genes	112
4.3.4	Experimental Verification of <i>BTPSL</i> from <i>M. Robertsii</i> as a Fungal Gene	114
4.3.5	The Presence of Typical Fungal TPS Genes in Relevant Fungal Species	114

4.3.6	Phylogenetic Analysis of BTPSLs, Related Bacterial TPSs and Other TPSs	115
4.3.7	The <i>BTPSL</i> Gene of HGT Origin in <i>M. Robertsii</i> Encodes a Functional Enzyme	116
4.3.8	Molecular Evolutionary Analysis of <i>BTPSL</i> Genes of HGT Origin in <i>Metarhizium</i>	116
4.4	Discussion	117
4.5	Materials and Methods	121
4.5.1	Data Sources and Analysis	121
4.5.2	Multiple Sequence Alignments and Phylogenetic Inference	122
4.5.3	Multiple Genome Alignment	124
4.5.4	Analysis of Selective Pressure	124
4.5.5	Fungal Culture	124
4.5.6	<i>MAA_08668</i> Gene Cloning and Verification of Its Position	125
4.5.7	Biochemical Characterization of <i>MAA_08668</i>	126
4.6	Acknowledgements	127
4.7	Bibliography	128
4.8	Appendix	136
5	Identification, Characterization, and Evolution of Fungal Terpene Synthases	147
5.1	Abstract	148

5.2	Introduction	148
5.3	Results	151
5.3.1	Distribution of TPSs in Fungal Genomes	151
5.3.2	Intron-Exon Structure	155
5.3.3	Motif Analysis	155
5.4	Discussion	157
5.5	Methods	159
5.5.1	Genome Mining of Fungal Terpene Synthases	159
5.5.2	Multiple Sequence Alignments and Phylogenetic Inference	160
5.6	Bibliography	161
5.7	Appendix	168
6	Summary and Conclusions	183
	Appendix	186
A	List of species	187
A.1	List of OneKP Species	187
A.2	List of 519 Fungi Species	198
Vita		202

List of Tables

2.1	Total number of TPS genes identified in six non-seed plant genomes. . . .	42
2.2	Typical plant terpene synthase genes in <i>Salvinia cucullata</i>	43
2.3	Typical plant terpene synthase genes in <i>Azolla superorganism</i>	44
2.4	Assembly statistics for <i>Anthoceros punctatus</i>	45
2.5	Typical plant terpene synthase genes in <i>Anthoceros punctatus</i>	46
2.6	Typical plant terpene synthase genes in <i>Marchantia polymorpha</i>	47
2.7	Assembly statistics for <i>Sphagnum fallax</i>	48
2.8	Typical plant terpene synthase genes in <i>Sphagnum fallax</i>	49
2.9	Number of TPSs genes from non-seed plant transcriptomes.	50
2.10	List of charophyta transcriptomes/genomes.	51
3.1	List of screened plant species and the number of MTPSLs in each species. All these 1103 transcriptomes were from OneKP (www.onekp.com). MTP- SLs are found from 146 species.	88
3.2	Summary statistics of MTPSLs in 9 plant lineages.	90

3.3	22 <i>MTPSL</i> genes outside of the four groups (high similarity to microbial TPSs).	91
3.4	<i>MTPSL</i> genes from the hornwort <i>Anthoceros punctatus</i> .	92
3.5	Typical plant terpene synthase genes in <i>Sphagnum fallax</i> .	93
3.6	A list of sequenced plants searched for <i>MTPSL</i> genes.	94
4.1	Presence/Absence of BTPSLs for each entomopathogenic fungi examined in this study.	136
4.2	List of typical fungal TPSs for each entomopathogenic fungi.	137
4.3	Presence/Absence of BTPSLs for each entomopathogenic fungi examined in this study.	138
4.4	Presence/Absence of BTPSLs for each entomopathogenic fungi examined in this study.	139
5.1	Summary statistics of exon numbers of four TPS types.	168
5.2	Summary statistics of mRNA lengths of four TPS types.	169
A.1	List of screened transcriptomes and the number of PTPS and MTPS in each sample. All these 1175 transcriptomes were from OneKP (www.onekp.com). The unique four letter codes are the transcriptome identifiers assigned by 1KP.	188
A.2	List of 519 screened fungi genomes and the number of PTPS and MTPS in each species. These data were downloaded from the fungal genomics portal (http://jgi.doe.gov/fungi).	198

List of Figures

1.1	Two compartmentalized pathways of isoprenoid biosynthesis. Dashed arrows indicate more than one step.	16
1.2	The mevalonic acid (MVA) pathway and the methylerythritol phosphate (MEP) pathway in plants (Vranova et al., 2013).	17
1.3	A schematic diagram of terpene synthases with annotated sequence features (Keeling and Bohlmann, 2006).	18
2.1	Distribution of protein lengths for typical plant terpene synthases in non-seed plants. The green dash line indicates the median length, while the read line indicates the average protein length.	52
2.2	Phylogenetic relationships of terpene synthases from non-seed plants and selected TPSs from seed plants. Weblogs show the conserved motifs found in each sunfamily.	53
2.3	Alignment of the conserved motifs found in 339 terpene synthases.	54
3.1	Distribution of terpene synthase genes of microbial type (MTPSL) identified from the transcriptomes of 1103 plant species.	96

3.2	Phylogeny of terpene synthases of microbial type (MTPSL) identified from OneKP.	97
3.3	Validation of representative <i>MTPSL</i> genes to be plant genes.	98
3.4	Motif analysis of MTPSLs.	99
3.5	Biochemical activity of selected <i>MTPSL</i> genes.	100
3.6	Alignment of one MTPSL of putative contamination with one fungal TPS.	101
3.7	Phylogenetic tree of <i>MTPSL</i> genes identified from <i>Anthoceros punctatus</i> genome with those identified from hornworts transcriptomes.	102
4.1	Maximum Likelihood phylogenetic tree of 341 Terpene Synthase C-domain sequences from bacteria (Blue) and fungi (Magenta).	140
4.2	Amino acid sequence alignment of 8 BTPSL TPSs, and representative TPSs from bacteria and fungi.	141
4.3	Alignment of the orthologous genomic regions containing the <i>BTPSL</i> genes from seven fungi species.	142
4.4	Confirmation of <i>MAA_08668</i> , a fungal TPS gene of putative horizontal gene transfer origin, integration into <i>Metarhizium robertsii</i> genome.	143
4.5	Maximum Likelihood phylogenetic tree and exon-intron structure of both types of TPS genes.	144
4.6	Biochemical activity of terpene synthase <i>MAA_08668</i> of putative horizontal gene transfer origin in <i>Metarhizium robertsii</i>	145

4.7	A species tree redrawn based on previous studies showing the evolutionary relationship of several entomopathogenic fungi.	146
5.1	Phylogenetic tree of major fungal lineages and corresponding number of terpene synthases in each of the four types.	170
5.2	Phylogeny of all four types of fungal terpene synthases.	171
5.3	Maximum likelihood phylogeny of 1753 α type terpene synthases.	172
5.4	Maximum likelihood phylogeny of 545 Trichodiene synthases.	173
5.5	Phylogenetic relationship among 7 Basidiomycota species containing bifunctional CPS/KSs.	174
5.6	Maximum likelihood phylogeny of 230 bifunctional CPS/KSs.	175
5.7	Multiple sequence alignment of 11 Basidiomycota CPS/KSs and 3 Ascomycota CPS/KSs.	176
5.8	Maximum likelihood phylogeny of 114 $\alpha\alpha 1$ type terpene synthases.	177
5.9	Multiple sequence alignment of 2 Basidiomycota and 3 Ascomycota $\alpha\alpha 1$ type terpene synthases.	178
5.10	Species distribution in the four types of fungal Terpene synthases.	179
5.11	Comparisons of the terpene synthase gene length and exon number.	180
5.12	Sequence logos of conserved terpene synthase family motifs observed in each type of fungal terpene synthases.	181
5.13	Venn diagram of identified CPS/KSs by known CPS/KSs from three species.	182

Chapter 1

Introduction and Literature Review

1.1 Terpenoids - Natural Functions and Industrial Uses

Terpenoids, also known as isoprenoids, constitute the largest and most diverse class of natural products, consisting of over 50,000 different terpenoid structures. They are chemically modified, such as by the addition of oxygen or rearrangement of the carbon skeleton, from terpenes, most of which are pure hydrocarbons. The term “terpene” is derived from the word “turpentine”, which is a fluid distilled from resin produced mainly by pine trees. This fluid is a variable mixture of terpenes. All classes of terpenoids are biosynthesized from two fundamental C₅ isoprene building blocks: isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). They are condensed by prenyltransferases in a typical head-to-tail manner to form various prenyl diphosphates that serve as the precursors for terpenoid biosynthesis. According to the number of isoprene building blocks they contain, terpenoids can be classified as hemiterpenoids (C₅), monoterpenoids (C₁₀), sesquiterpenoids (C₁₅), diterpenoids (C₂₀), triterpenoid (C₃₀), tetraterpenoid (C₄₀) or polyterpenoids (C_{5n}) (Tholl and Lee, 2011).

Although terpenoids are more abundant and extensively studied in plants, they are ubiquitous and involved in a wide range of biological functions in all living organisms. Many of them are produced as primary metabolites, where they function as pigments in photosynthesis (carotenoids, chlorophylls and plastoquinones), plant hormones (gibberellic acid, abscisic acid and cytokinins), electron transport carriers (ubiquinone and plastoquinone), carbohydrate carriers (bactoprenol and dolichols), membrane stabilizers (sterols), etc. (Tholl, 2006). However, the most diverse and abundant isoprenoids are

discovered in plants as secondary metabolites that play vital roles in direct and indirect plant defenses in response to herbivores and pathogens, as well as in communicating with their environment, among which volatile terpenoids released by different plant tissues, such as flowers, leaves and fruits, are particular important (Christianson, 2008; Pichersky and Lewinsohn, 2011; Pichersky et al., 2006). Besides their importance in plant physiology and ecology, many terpenoids are economically important as fragrance, pigments, polymers, fibers, natural insecticide and natural rubber. There also have been many important applications of terpenoids in pharmaceutical industry such as the two most valuable terpenoids, the antineoplastic agent TAXOL® (paclitaxel) and the antimalarial drug artemisinin that has been used to treat malaria for more than 2000 years.

1.2 Biosynthesis of Terpenoids

The universal isoprene building blocks of all isoprenoids, IPP and DMAPP, can be synthesized via two distinct pathways: the mevalonic acid (MVA) pathway and the methylerythritol phosphate (MEP) pathway (Figure 1.1). For many years, the classical MVA pathway had been widely accepted as the only route in all organisms. However, several discordant results suggesting the existence of another pathway had been reported in certain bacteria. These led to the discovery of the MEP pathway by Rohmer in 1993 (Rohmer et al., 1993). Although there are some exceptions, it is now well established that the MVA pathway is present in eukaryotes and archaea and located in the cytosol; in contrast, the MEP pathway, which operates in plastids, is used by eubacteria, plant

chloroplasts and green algae but absent in archaea, fungi and animals (Vranova et al., 2013). In plants, these two pathways coexist but differ in their localization (Roberts, 2007). The MVA pathway provides the precursors for the biosynthesis of sesquiterpenes and triterpenes, while monoterpenes, diterpenes, carotenoids, chlorophylls and gibberellins are generated via the MEP pathway. Although interactions between them have been reported in plants, very little is known about the exchange mechanism by which they are finely cooperated (Hemmerlin et al., 2003).

The MVA pathway uses seven enzymes to transform acetyl-CoA to IPP. In the initial reaction of the MVA pathway (Figure 1.2), three molecules of acetyl-CoA are condensed successively to produce 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) by the enzymes acetoacetyl-CoA thiolase (AACT) and HMG-CoA synthase (HMGS). HMG-CoA is further converted to mevalonic acid (MVA) in two successive reduction steps by HMG-CoA reductase (HMGR), a key rate-limiting enzyme in the MVA pathway. The subsequent conversion of MVA into mevalonate-5-diphosphate (MVADP) involves two-step phosphorylation reactions, catalyzed by mevalonate kinase (MK or MVK) and phosphomevalonate kinase (PMK or PMVK). Then, MVADP is decarboxylated by mevalonate-5-pyrophosphate decarboxylase (MPDC or MVD) to finally form IPP, which is then converted to its isomer, DMAPP, in a reaction catalyzed by the IPP/DMAPP isomerase (IDI).

The MEP pathway (Figure 1.2), also named 1-deoxy-D-xylulose-5-phosphate (DXP) pathway, starts with the formation of DXP by condensation of glyceraldehyde-3-phosphate (GAP) with (hydroxyethyl) thiamine diphosphate derived from pyruvate. This reaction is

catalyzed by 1-deoxy-D-xylulose 5-phosphate synthase (DXS). In the second step, DXP is rearranged and reduced to generate 2-C-Methylerythritol 4-phosphate (MEP) by DXP reductoisomerase (DXR) in the presence of NADPH. The next step is the conversion of MEP into 4-Diphosphocytidyl-2-C-methylerythritol (CDP-ME) in a CTP-dependent reaction mediated by the enzyme 4-diphosphocytidyl-2C-methyl-D-erythritol synthase (MCT or IspD). Subsequently, CDP-ME is phosphorylated to yield 4-Diphosphocytidyl-2-C-methyl-D-erythritol 2-phosphate (CDP-MEP), which is catalyzed by 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase (CMK). The following two steps are the sequential conversion of CDP-MEP into (E)-4-Hydroxy-3-methyl-but-2-enyl pyrophosphate (HMBPP) under the catalysis of 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS) and 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (HDS). In the final step, the enzyme HMBPP reductase (HDR) converts HMBPP into IPP and DMAPP with an approximate ratio of 5.

In the last stage of terpenoid biosynthesis, prenyltransferases catalyze the head-to-tail fusion of DMAPP with one, two, or three units of IPP to form linear prenyl diphosphate precursors: geranyl diphosphate (GPP, C₁₀, monoterpenoids), farnesyl diphosphate (FPP, C₁₅, sesquiterpenoids) and geranylgeranyl diphosphate (GGPP, C₂₀, diterpenoids). Next, these direct precursors of terpenoids are catalyzed by terpene synthases to produce diverse terpenoid compounds, which can be further modified by hydroxylation, methylation, glycosylation, acylation or peroxidation.

1.3 Terpene Synthases

As described above, all terpenoids are produced from allylic diphosphates by the action of terpene synthases (TPSs), which constitute a superfamily and contribute to the tremendous diversity of terpenoid carbon skeletons. They are present in multiple domains of life, but their protein sequences show a relatively low level of conservation even within the same domain (Bohlmann et al., 1997). However, the elucidated protein structures of several TPSs (Hyatt et al., 2007; Köksal et al., 2011b, 2010; McAndrew et al., 2011; Zhou et al., 2012) revealed a high similarity in their tertiary structures consisting of an N-terminal domain and a catalytically active C-terminal domain, which implies a similarity in their catalytic mechanisms. Insights obtained from structural studies helped us identify two distinct chemical strategies employed for initiating cyclization reactions, based on which TPSs can be divided into two major categories. Typically, class I TPSs has a C-terminal domain (also referred to as α -domain or class I fold) and catalyzes an ionization-initiated cyclization of the substrate diphosphate group. The α -domain adopts the α -helical protein fold and contains two metal binding motifs, highly conserved DDXXD and less well conserved NSE/DTE motifs located on opposing helices near the entrance of the active site. Both of them bind a trinuclear magnesium cluster to trigger the ionization of the isoprenoid substrate and initiate the cyclization reaction. In contrast, class II TPSs possesses a functional N-terminal domain (or β -domain), which together with a third “insertion” γ -domain forming a vestigial class II fold. Enzymes in this class contain a conserved DXDD motif, which is located in the β -domain and responsible for

the protonation-initiated cyclization. The γ -domain carries a highly acidic EDXXD-like Mg^{2+} /diphosphate binding motif that also contributes to the activity of class II TPSs (Cao et al., 2010).

Typically, microbial class I and II TPSs are monofunctional and display only α -domain and $\beta\gamma$ -domain, respectively. Most plant monoterpene and sesquiterpene synthases contain both α -domain and β -domain, but only have α -domain activity. The β -domain is always rendered inactive due to the loss of conserved DXDD motif. The available protein crystal structures suggest that most of plant diterpene and some sesquiterpene synthases possess all these three domains (γ - β - α). However, usually only one domain is functional. For example, sesquiterpene synthase only retain class I TPSs activity as seen in α -bisabolene synthase (AgBIS) from *Abies grandis* (McAndrew et al., 2011). In contrast, Diterpene synthases use either one of the two active sites, such as the class I TPS Taxadiene synthase from the Pacific yew (Köksal et al., 2011b), the class II TPS ent-copalyl diphosphate synthase (CPS) from *Arabidopsis thaliana* (Köksal et al., 2011a). However, there are always exceptions. Some diterpene synthases with both class I and class II activities, such as the abietadiene synthase (AgAS) (Zhou et al., 2012) from *Abies grandis*, ent-kaurene synthase (FCPS/KS) (Kawaide et al., 1997) from the fungus *Phaeosphaeria* sp. L487 and ent-kaurene synthase (PpCPS/KS) (Hayashi et al., 2006) from the moss *Physcomitrella patens*, have been reported.

Terpene synthases are approximately 550–860 amino acids long and 50–100 kDa in their molecular masses (Figure 1.3), differentiated by the combinations of domains and motifs. In general, sesquiterpene synthases are 550–600 aa long and 50–70 amino acids

shorter than monoterpene synthases, which contain the N-terminal plastid-targeting peptides. Diterpene synthases are even longer than monoterpene synthases due to the additional insertion sequences (γ -domain) in their N-terminal. Many terpene synthases also carry a highly conserved RR(x)8W motif downstream of the N-terminal transit peptide. Previous studies shown that the activity of monoterpene synthase doesn't require this motif, and thus it is considered as a landmark separating the transit peptide from the mature protein sequence (Bohlmann et al., 1998; Keeling and Bohlmann, 2006).

1.4 Origin and Evolution of Terpene Synthase Genes

Although the TPS represents a mid-size gene family, they are too few in numbers when compared to the great number of compounds in this group, which is mainly due to the functional plasticity of most TPSs. Many terpene synthases can produce more than one compound via only very few substitutions. In order to survive in the rapidly changing environment, TPS genes must evolve quickly to generate the required terpenoid profile. It is widely accepted that TPS genes have undergone extensive lineage-specific gene duplication, followed by sub-functionalization or even neo-functionalization to perform different roles, the so-called divergent evolutionary process, which is indicated by the fact that a significant number of TPS genes are clustered into tandem arrays. On the other hand, instances of convergent evolution can also be observed. For example, TPS23, a sesquiterpene synthase for producing (E)- β -caryophyllene from farnesyl diphosphate (FPP) in maize (*Zea mays*), is more similar to other maize TPSs than (E)- β -caryophyllene

synthases from other dicot species (Köllner et al., 2008). Also, the low level of sequence similarity among remotely related species and distinct patterns of genomic organizations of plant and microbial TPSs support a convergent evolution.

However, where did plant TPS genes come from and do plant and microbial TPS genes share a common ancestor? Previous studies proposed that plant terpene synthase gene ancestor, resembling a conifer diterpene synthase associated with Gibberellin biosynthesis, emerged prior to the divergence of angiosperms and gymnosperms (Trapp and Croteau, 2001). Later, the bifunctional diterpene synthase (PpCPS/KS) identified in the bryophyte *Physcomitrella patens*, was hypothesized to be the common ancestor (Hayashi et al., 2006; Keeling et al., 2010). Interestingly, in bacterium *Bradyrhizobium japonicum*, the ent-copalyl diphosphate synthase (BjCPS) and ent-kaurene synthase (BjKS) share some similarity with the β -domain (CPS activity) and α -domain (KS activity) of the plant and fungal TPS genes, respectively. This, together with the fact that plant terpene synthase genes are longer and the two types of monoterpene synthases (CPS/KS) are of roughly equal length, suggested a common ancestral diterpene synthase gene shared by plants, fungi, and bacteria (Morrone et al., 2009). Results gained by analyzing the structures of diterpene synthase genes supported this hypothesis and showed that all modern plant TPS genes evolved from a three-domain (γ - β - α) diterpene synthase gene, which is the fusion of ancient bacterial α -domain and $\beta\gamma$ -domain terpene synthase genes (Cao et al., 2010). Recently, a new class of plant terpene synthase gene, the microbial-type class I terpene synthase genes containing only a α -domain, has been discovered in the spikemoss *Selaginella muellendorffii* (Li et al., 2012). This class of TPS genes is more similar to

microbial TPS genes than other plant TPS genes and is probably integrated into plant genomes from microbes via horizontal gene transfer, indicating a different evolutionary path rather than evolving from the three-domain diterpene synthase gene.

1.5 Objectives

As more genome and transcriptome sequences from non-model species are available, particular for non-seed plants, it is possible to obtain more TPS genes and examine them within a broader range of taxa. This analysis holds true for constructing a comprehensive evolutionary history of this gene family. Generally speaking, we are interested to trace the origin and evolutionary history of terpene synthase genes, and then to confirm or refine the evolutionary paths. The overall project is designed to address broad questions about the evolution of terpene synthase genes.

1.6 Bibliography

- Bohlmann, J., Meyer-Gauen, G., and Croteau, R. (1998). Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proceedings of the National academy of Sciences of the United States of America*, 95(8):4126–33. [8](#)
- Bohlmann, J., Steele, C. L., and Croteau, R. (1997). Monoterpene synthases from grand fir (*Abies grandis*): cDNA isolation, characterization, and functional expression of myrcene synthase, (-)-(4S)-limonene synthase, and (-)-(1S,5S)-pinene synthase. *Journal of Biological Chemistry*, 272(35):21784–21792. [6](#)
- Cao, R., Zhang, Y., Mann, F. M., Huang, C., Mukkamala, D., Hudock, M. P., Mead, M. E., Pristic, S., Wang, K., Lin, F. Y., Chang, T. K., Peters, R. J., and Oldfield, E. (2010). Diterpene cyclases and the nature of the isoprene fold. *Proteins*, 78(11):2417–32. [7](#), [9](#)
- Christianson, D. W. (2008). Unearthing the roots of the terpenome. *Current Opinion in Chemical Biology*, 12(2):141–50. [3](#)
- Hayashi, K. i., Kawaide, H., Notomi, M., Sakigi, Y., Matsuo, A., and Nozaki, H. (2006). Identification and functional analysis of bifunctional ent-kaurene synthase from the moss *Physcomitrella patens*. *FEBS Letters*, 580(26):6175–6181. [7](#), [9](#)
- Hemmerlin, A., Hoeffler, J. F., Meyer, O., Tritsch, D., Kagan, I. A., Grosdemange-Billiard, C., Rohmer, M., and Bach, T. J. (2003). Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells. *Journal of Biological Chemistry*, 278(29):26666–26676. [4](#)

- Hyatt, D. C., Youn, B., Zhao, Y., Santhamma, B., Coates, R. M., Croteau, R. B., and Kang, C. (2007). Structure of limonene synthase, a simple model for terpenoid cyclase catalysis. *Proceedings of the National academy of Sciences of the United States of America*, 104(13):5360–5. [6](#)
- Kawaide, H., Imai, R., Sassa, T., and Kamiya, Y. (1997). Ent-kaurene synthase from the fungus *phaeosphaeria* sp. l487. cDNA isolation, characterization, and bacterial expression of a bifunctional diterpene cyclase in fungal gibberellin biosynthesis. *Journal of Biological Chemistry*, 272(35):21706–12. [7](#)
- Keeling, C. I. and Bohlmann, J. (2006). Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defence of conifers against insects and pathogens. *New Phytologist*, 170(4):657–75. [xv](#), [8](#), [18](#)
- Keeling, C. I., Dullat, H. K., Yuen, M., Ralph, S. G., Jancsik, S., and Bohlmann, J. (2010). Identification and functional characterization of monofunctional ent-copalyl diphosphate and ent-kaurene synthases in white spruce reveal different patterns for diterpene synthase evolution for primary and secondary metabolism in gymnosperms. *Plant Physiology*, 152(3):1197–1208. [9](#)
- Köksal, M., Hu, H., Coates, R. M., Peters, R. J., and Christianson, D. W. (2011a). Structure and mechanism of the diterpene cyclase ent-copalyl diphosphate synthase. *Nature Chemical Biology*, 7(7):431–433. [7](#)

- Köksal, M., Jin, Y., Coates, R. M., Croteau, R., and Christianson, D. W. (2011b). Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature*, 469(7328):116–122. [6](#), [7](#)
- Köksal, M., Zimmer, I., Schnitzler, J. P., and Christianson, D. W. (2010). Structure of isoprene synthase illuminates the chemical mechanism of teragram atmospheric carbon emission. *Journal of Molecular Biology*, 402(2):363–373. [6](#)
- Köllner, T. G., Held, M., Lenk, C., Hiltbold, I., Turlings, T. C. J., Gershenzon, J., and Degenhardt, J. (2008). A maize (e)- β -caryophyllene synthase implicated in indirect defense responses against herbivores is not expressed in most american maize varieties. *Plant Cell*, 20(2):482–494. [9](#)
- Li, G., Kollner, T. G., Yin, Y., Jiang, Y., Chen, H., Xu, Y., Gershenzon, J., Pichersky, E., and Chen, F. (2012). Nonseed plant selaginella moellendorffi [corrected] has both seed plant and microbial types of terpene synthases. *Proceedings of the National academy of Sciences of the United States of America*, 109(36):14711–5. [9](#)
- McAndrew, R. P., Peralta-Yahya, P. P., DeGiovanni, A., Pereira, J. H., Hadi, M. Z., Keasling, J. D., and Adams, P. D. (2011). Structure of a three-domain sesquiterpene synthase: a prospective target for advanced biofuels production. *Structure*, 19(12):1876–84. [6](#), [7](#)
- Morrone, D., Chambers, J., Lowry, L., Kim, G., Anterola, A., Bender, K., and Peters, R. J. (2009). Gibberellin biosynthesis in bacteria: Separate ent-copalyl diphosphate and ent-kaurene synthases in bradyrhizobium japonicum. *FEBS Letters*, 583(2):475–480. [9](#)

- Pichersky, E. and Lewinsohn, E. (2011). Convergent evolution in plant specialized metabolism. *Annual Review of Plant Biology*, 62:549–66. [3](#)
- Pichersky, E., Noel, J. P., and Dudareva, N. (2006). Biosynthesis of plant volatiles: nature's diversity and ingenuity. *Science*, 311(5762):808–11. [3](#)
- Roberts, S. C. (2007). Production and engineering of terpenoids in plant cell culture. *Nature Chemical Biology*, 3(7):387–95. [4](#)
- Rohmer, M., Knani, M., Simonin, P., Sutter, B., and Sahm, H. (1993). Isoprenoid biosynthesis in bacteria: a novel pathway for the early steps leading to isopentenyl diphosphate. *Biochemical Journal*, 295 (Pt 2):517–24. [3](#)
- Tholl, D. (2006). Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Current Opinion in Plant Biology*, 9(3):297–304. [2](#)
- Tholl, D. and Lee, S. (2011). Terpene specialized metabolism in arabidopsis thaliana. *Arabidopsis Book*, 9:e0143. [2](#)
- Trapp, S. C. and Croteau, R. B. (2001). Genomic organization of plant terpene synthases and molecular evolutionary implications. *Genetics*, 158(2):811–32. [9](#)
- Vranova, E., Coman, D., and Grusissem, W. (2013). Network analysis of the mva and mep pathways for isoprenoid synthesis. *Annual Review of Plant Biology*, 64:665–700. [xv](#), [4](#), [17](#)

Zhou, K., Gao, Y., Hoy, J. A., Mann, F. M., Honzatko, R. B., and Peters, R. J. (2012). Insights into diterpene cyclization from structure of bifunctional abietadiene synthase from *Abies grandis*. *Journal of Biological Chemistry*, 287(9):6840–6850. [6](#), [7](#)

1.7 Appendix

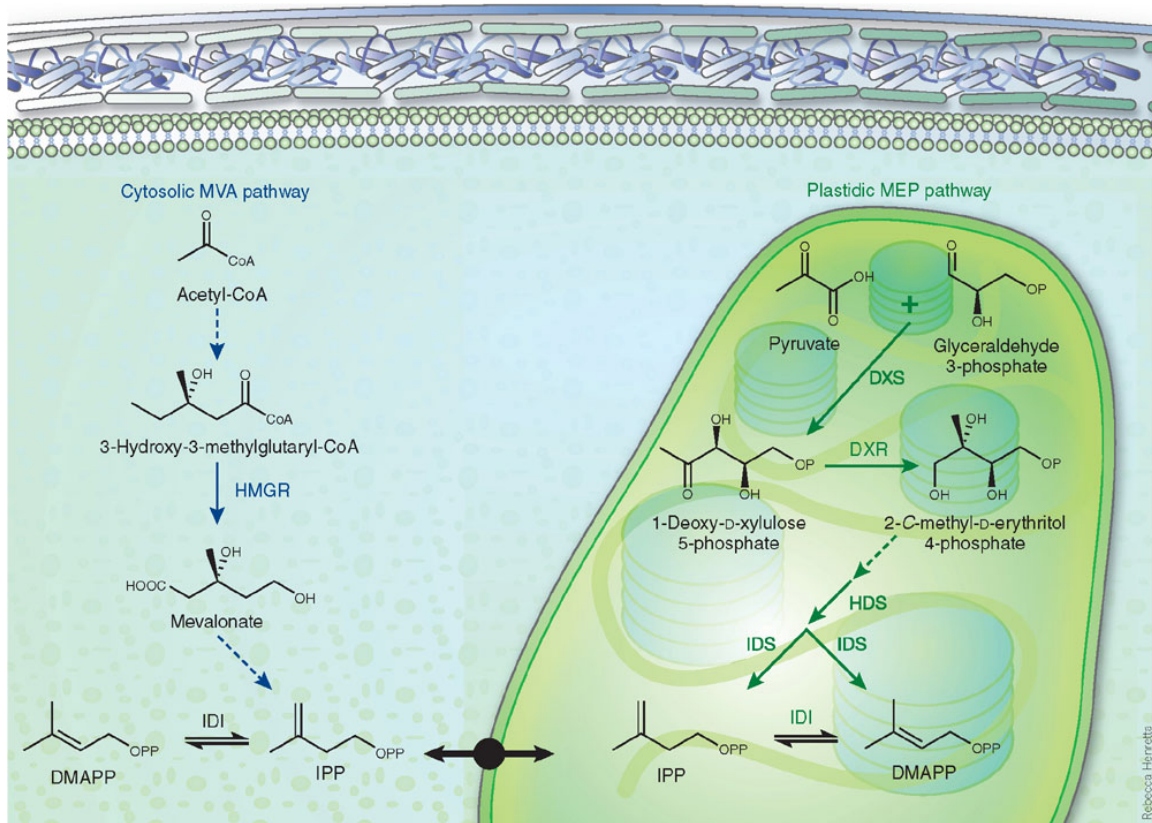


Figure 1.1: Two compartmentalized pathways of isoprenoid biosynthesis. Dashed arrows indicate more than one step.

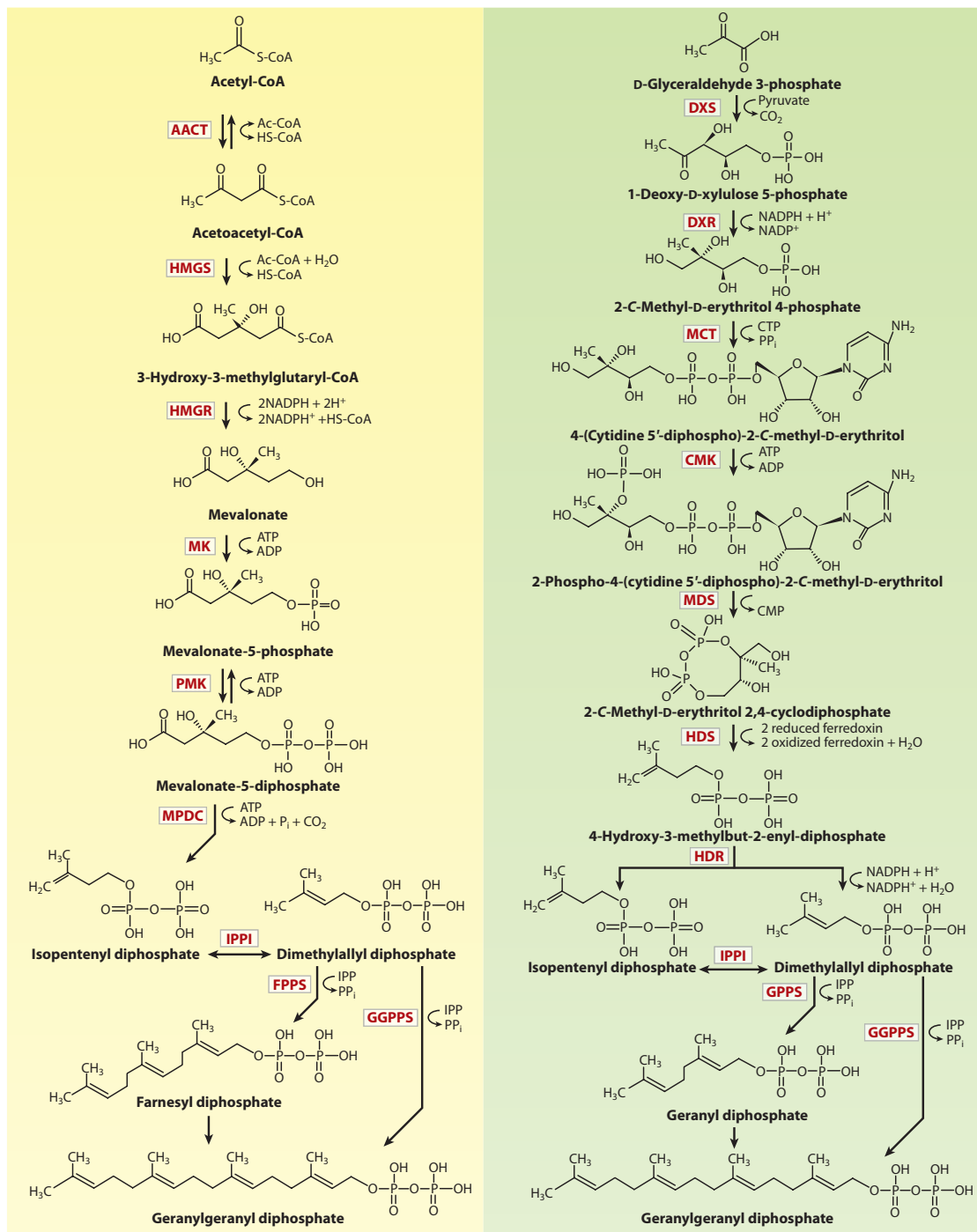


Figure 1.2: The mevalonic acid (MVA) pathway and the methylerythritol phosphate (MEP) pathway in plants (Vranova et al., 2013).

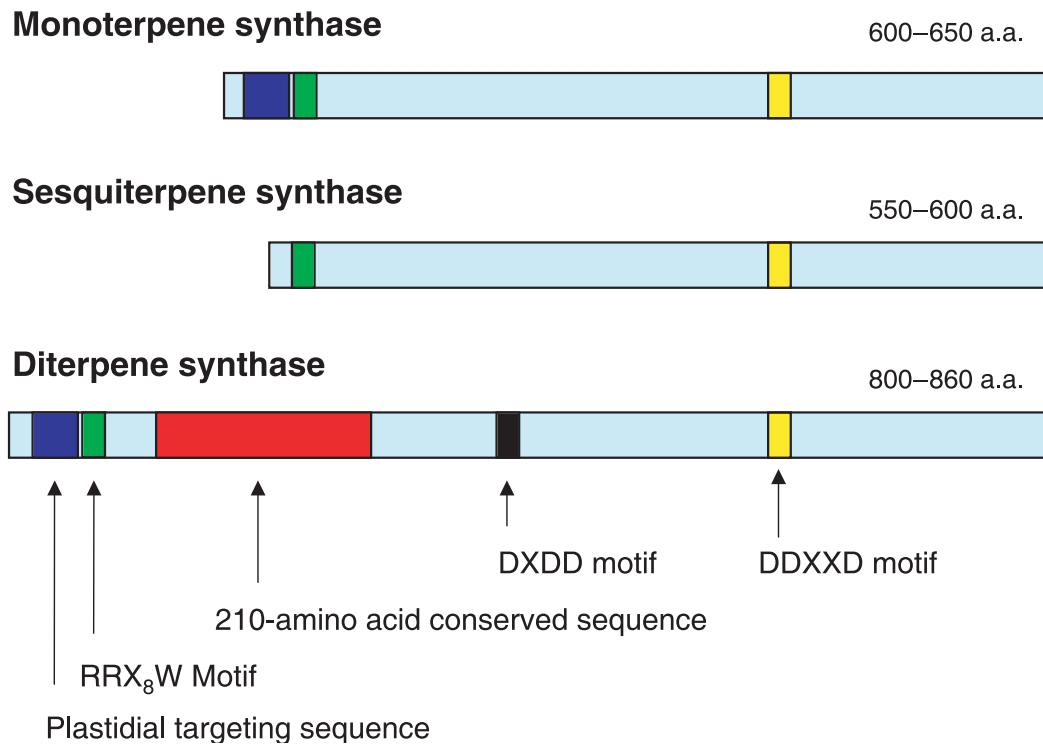


Figure 1.3: A schematic diagram of terpene synthases with annotated sequence features (Keeling and Bohlmann, 2006).

Chapter 2

Evolution of Typical Plant Terpene Synthase Genes in Non-Seed Plants

2.1 Abstract

The terpene synthases (TPSs) are widely encoded in plants and the primary enzymes responsible for producing diverse natural compounds. Previous studies, which typically focus on sequenced plant species, classify this mid-size gene family into seven subfamilies: a, b, c, d, g, e/f and h. However, TPSs from non-seed plant species are very limited in this classification. In this study, by adding more TPSs identified from large-scale non-seed plant transcriptomes and several sequenced non-seed plant genomes, we were able to increase the resolution of TPS subfamily classification. The new phylogenetic analysis recognized seven previous known subfamilies and a new subfamily x. The tree topographies are dominated by taxonomic relatedness and placements of some TPSs from *Physcomitrella patens* (PpCPS/KS) and *Selaginella moellendorffii* are changed. Subfamilies x and h are specific to non-seed plants. The new subfamily x contains TPSs from all major lineages of non-seed plants except Lycophytes, while the subfamily h has no TPSs from Hornworts. Although TPSs in subfamilies x and h share similar motif composition, TPSs in subfamily h are more closely related to some sequences in the subfamily d, which is specific to gymnosperm. This suggests that sequences in the subfamily h are more similar to the ancestor three-domain gene in plants.

2.2 Introduction

Terpenoids, the largest and most structurally and functionally diverse class of natural compounds on earth, are mostly synthesized by plants to be involved in various plant-environment interactions. Some terpenoids are classified as primary metabolites essential for plant growth and development. Terpene synthases (TPSs), the key enzymes for terpenoid biosynthesis, are the major determinant of the tremendous diversity of terpenoid carbon skeletons. The TPS genes represent a mid-size family of 30-100 functional genes in almost all major sequenced plant genomes (Chen et al., 2011).

Terpene synthases can be further divided into seven subfamilies: a, b, c, d, g, e/f and h (Bohlmann et al., 1998; Chen et al., 2011; Dudareva et al., 2003; Martin and Bohlmann, 2004). The a, b and g subfamilies are angiosperm specific, while the d subfamily is gymnosperm-specific. The subfamily c encodes monoterpene synthases and contains the single Terpene Synthase gene in the moss *Physcomitrella patens*, genes from angiosperms and gymnosperms and also genes in *Selaginella moellendorffii*. The subfamily e/f contains ent-kaurene synthases (KSs) from angiosperms and gymnosperms. The new subfamily h is specific to *Selaginella moellendorffii* and members of this subfamily contain both DxDD and DDxxD motifs, which are signatures of bifunctional diterpene synthases.

However, most of our current knowledge of terpene synthase genes is based on studies in model species of plants. Among the 50 plants with genomes, the vast majority (94%) of them belong to angiosperms, while only one from each of the three clades: gymnosperm, bryophyte and lycophyta (Michael and Jackson, 2013). The inferred subfamilies are based

on the phylogenetic analysis of terpene synthases from seven sequenced plant genomes and selected sequences from gymnosperms. Among these seven plant species, *Selaginella moellendorffii* is the only non-seed plant analyzed and some members of its terpene family form the new subfamily h. Given the high diversity of terpenoids, taxonomically specialized TPS genes should exist in some plant species or families, in particular for non-seed plants. So, in order to have a comprehensive understanding of the evolutionary history of this gene family, studies within a more comprehensive coverage of plant species are indispensable, which are also suggested in previous studies (Trapp and Croteau, 2001). In the last years, new advances in high throughput sequencing techniques have allowed us to discover genes and profile their expression patterns more quickly and cheaply. Lots of studies (Bleeker et al., 2011; Drew et al., 2013; Hall et al., 2013; Han et al., 2013; Keeling et al., 2011) have been made by transcriptome mining to identify terpene synthase genes in non-model species. Interestingly, a project, called PhytoMetaSyn with the aim of sequencing 75 non-model plants that generate high-value metabolites such as polyketides, terpenes and alkaloids, has already identified a list of new biosynthetic genes involved in the metabolic pathways (Xiao et al., 2013). Another project is the One Thousand Plant Transcriptomes Project (OneKP), which generates transcriptomes for over 1000 plant species that span almost all major plant clades (from green algae to flowering plants), aiming to increase the taxonomic diversity of available genes in GenBank. Also, genomes of more non-seed plants are emerging, such as a moss (*Sphagnum fallax*), a liverwort (*Marchantia polymorpha*), a Hornwort (*Anthoceros Punctatus*) and two ferns (*Salvinia cucullata* and *Azolla superorganism*). Analysis of terpene synthases from these genomic

datasets of non-seed plants will provide valuable insights into the evolutionary history of terpene synthases.

2.3 Results and Discussion

2.3.1 Identification of Terpene Synthase Genes from Genomes of Hornwort, Moss, Liverwort and Two Ferns

Six non-seed plant species (Table 2.1) with available genomes were screened for typical plant terpene synthase genes. These six species span all five lineages of non-seed plants. The two fern genomes, *Salvinia cucullata* and *Azolla superorganism*, contain 7 (Table 2.2) and 1 (Table 2.3) terpene synthase genes, respectively. Five of these seven terpene synthase genes in *Salvinia cucullata* have DxDD or DDxxD motifs, while the other two, ScuTPS5 and ScuTPS7, have no these two functional motifs. The longest one, ScuTPS1, has the DxTx and DDxxD motifs. All these seven genes lack the NDSE motif. The only gene in the *Azolla superorganism* genome encodes a protein of 894 amino acids with the DxDD and DDxxD motifs. The genome of the hornwort *Anthoceros punctatus* is not complete and contains more than 15,000 contigs (Table 2.4). 22 terpene synthases were identified and the protein lengths range from 186 to 880 amino acids (Table 2.5). 14 of them contain at least one functional motif. ApuTPS5 and ApuTPS16 have both the DxDD and DDxxD motifs. The liverwort genome, *Marchantia polymorpha*, contains seven terpene synthases (Table 2.6), all of which are around 800 amino acids. The genome of the moss *Sphagnum*

fallax has four terpene synthases and all of them have both the DxDD and DDxxD motifs except the shortest one (Table 2.7 and 2.8). For the lycophyte *Selaginella moellendorffii*, 18 terpene synthases were identified in our previous study and three of them are putative bifunctional terpene synthases.

2.3.2 Identification of Terpene Synthases from Non-Seed Plants

Transcriptomes

By mining transcriptomes of 324 non-seed plant species, we found 253 archetypical plant terpene synthases (at least 150 amino acids long) from 88 species (Table 2.9). The median and average length of these 253 terpene synthases is 370 and 433 amino acids, respectively (Figure 2.1). They are present in all major clades of non-seed plants: monilophytes, lycophytes, hornworts, mosses and liverworts. However, they are completely absent in two groups of green algae: Chlorophyte and Charophyte. It is also clear from table 2.9 that all lycophytes and majority of liverworts screened encode terpene synthase genes, while only 27% (17 out of 90) of monilophytes contain terpene synthase genes.

2.3.3 Search for Terpene Synthase Genes in Green Algae

To confirm the complete absence of terpene synthase genes in green algae, we further analyzed additional available genomic datasets obtained from other sources. In total, 8 algal genomes (1 from Charophyta and 7 from Chlorophyta) and 4 transcriptomes (3 from

Charophyta and 1 from Chlorophyta) were screened for the presence of terpene synthase genes (Table 2.10). The blastp searches didn't find any significant hits to the known terpene synthase genes. This, at least now, further confirmed the absence of terpene synthase genes in green algae and is also consistent with the findings of one early study (Sasso et al., 2012).

2.3.4 Phylogenetic Analysis of TPSs from Non-Seed Plants with TPSs from Selected Seed Plants

Previous studies mainly in seed plants classified terpene synthases into seven sub-families (TPS-a to TPS-h) by means of phylogenetic analysis. In this study, we extended this analysis further by adding terpene synthases identified from large-scale non-seed plant transcriptomes and genomes. 253 plant terpene synthases identified from 88 oneKP species, 60 from 7 non-seed plant genomes (7 in *Salvinia cucullata*, 1 in *Azolla superorganism*, 18 in *Selaginella moellendorffii*, 22 in *Anthoceros punctatus*, 4 in *Sphagnum fallax*, 7 in *Marchantia polymorpha* and 1 in *Physcomitrella patens*), 1 bifunctional TPS from the transcriptome of *Lygodium japonicum* and 149 selected TPSs from seed plants (*Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Sorghum* and Gymnosperm) were collected (463 in total) to conduct the phylogenetic analyses. An initial multiple sequence alignment based on the 463 terpene synthase protein sequences was used to detect the conserved functional DxDD and DDxxD motifs in each sequence. Sequences containing at least one of the two motifs were selected.

The aforementioned curation and criteria resulted in a final dataset of 339 terpene synthases (200 from non-seed plants and 139 from seed plants). The phylogenetic analysis based on alignment of these 339 terpene synthases enabled us to classify them into eight subfamilies: the seven currently recognized subfamilies (a, b, c, d, e/f, g and h) and one newly identified subfamily x (Figure 2.2 and Figure 2.3). The sequence motifs observed in each subfamily and each terpene synthase were displayed in Figure 2.2 and Figure 2.3, respectively. As shown in Figure 2.2, TPS-a, b, g and d subfamilies are specific to seed plants. The RR(x)8W motif that is essential for the isomerization of substrates (Williams et al., 1998) is conserved in both b and d subfamilies. Its variations are found in subfamily a and it is completely absent from subfamily g. Within these four subfamilies, the SAYDTAWVA motif, which is shared by both CPS and KS proteins (Rodriguez et al., 2012; Sakamoto et al., 2004; Smith et al., 1998), is only observed in subfamily d, especially in sequences having both DxDD and DDxxD motifs. All four subfamilies have conserved RW and RWD motifs around the DDxxD motif. The functions of both motifs are unknown. Another motif whose function is known and shared by all four subfamilies is the RxR motif, which plays a role in directing the diphosphate anion away from the reactive carbocation after ionization (Kollner et al., 2004; Starks et al., 1997; Tholl et al., 2005). In subfamily g, this motif is changed to RxQ. The typical motifs, DDxxD and NSE/DTE, are conserved among all four subfamilies.

The subfamily h, which is specific to non-seed plants and previously includes only sequences from *Selaginella moellendorffii*, now has members from all five lineages of non-seed plants except the Hornworts. The RR(x)8W motif is completely absent and all other

motifs investigated are observed in almost all members of this subfamily, the exceptions being the ten members from Liverworts. They lack both the DDxxD and NSE/DTE motifs, but the RxR and RWD motifs are intact. One possible explanation is that sequences in Liverworts lost this functional motif during the evolution, and only keep the DxDD motif and function as class II terpene synthases. This is contrary to the general feature of this subfamily being that sequences have both functional motifs and are putative bifunctional terpene synthases.

The subfamily e/f contains the class I terpene synthases with the DDxxD signature. All members possess the conserved SAYDTAWVA motif and modified RxR motif, but lack the RR(x)8W and DxDD motifs. Members are from both seed and non-seed plants. Interestingly, no sequences from Monilophytes and Mosses are found in this subfamily.

Another subfamily that includes sequences from both seed and non-seed plants is the subfamily c, which contains CPS proteins from all four angiosperm species, gymnosperms and Lycophytes. Members of this subfamily have motif features similar to that shared by ten members from Liverworts in the subfamily h. They contain the DxDD motif, but the NSE/DTE motif is missing. They also share an almost completely conserved SAYDTAWVA motif.

The subfamily x that is the largest subfamily with 109 members is specific to non-seed plants. Members are from all five lineages of non-seed plants except the Lycophytes and are clustered together according to the lineages to which they belong. The single terpene synthase in the genome of *Physcomitrella patens*, PpCPS/KS, belongs to the subfamily c according to previous studies. However, it is now in the subfamily x and clusters with

sequences from other Mosses. Its new placement is strongly supported by the obvious differences in motif features between subfamilies c and x. Members in the subfamily x contain all conserved motifs except the RR(x)8W motif and are putative bifunctional terpene synthases, while members in the subfamily c only have the DxDD functional motif and are *bona fide* CPS synthases.

2.4 Conclusions

In this chapter, typical plant terpene synthases in non-seed plant transcriptomes and genomes were identified and characterized. A phylogenetic analysis of these newly identified terpene synthases from non-seed plants and selected terpene synthases from seed plants recognized seven previous known subfamilies and a new subfamily x. Members in different subfamilies usually have different motif features, and sequences from same plant lineage are more inclined to cluster together. Subfamilies a, b, g and d are specific to seed plants, while subfamilies h and x are specific to non-seed plants. Subfamilies e/f and c include members from both seed plant and non-seed plants.

Although the results obtained here are largely consistent with previous studies, they still provide new findings and insights into the evolution of terpene synthase gene family, for example, the discovery of new subfamily x. These results can be attributed to the increased number of terpene synthases identified from non-seed plants. In previous studies, sequences from non-seed plants are limited to two sequenced genomes: *Physcomitrella patens* and *Selaginella moellendorffii*. When more sequences from non-seed

plants added, placements of some sequences found in these two genomes were changed. For example, the PpCPS/KS in *Physcomitrella patens* is now in the subfamily x and some *Selaginella moellendorffii* genes previous clustered with gymnosperms KS proteins now cluster with sequences from Liverworts and Lycophytes (Figure 2.3).

Previous study suggested that modern plant terpene synthases originated from the bifunctional diterpene synthase, PpCPS/KS in *Physcomitrella patens* (Hayashi et al., 2006). However, from my results, it is in the largest subfamily x and seems a very common gene in non-seed plants. It is very likely that members in the subfamily x are one of the oldest existing terpene synthases in land plants, but not actual ancestors of modern plant terpene synthases, at least not the immediate ancestors. In another study (Trapp and Croteau, 2001), an extant relative of a conifer diterpene synthase was suggested to be the ancestor. It seems that sequences in the subfamily h are more closely related to the ancestor three-domain gene in plants, in view of their closer sequence similarity with the gymnosperm terpene synthases and almost same motif feature with members in the subfamily x. Also, some members in subfamily h, such as the ten sequences in Liverworts, function as mono-functional enzymes, even though they still have the three-domain structure. However, the loss of DDxxD and NSE/DTE motifs makes them as CPS proteins. The motif feature for gymnosperm diterpene synthases is also similar to that in the subfamily h. These seven diterpene synthases (Figure 2.3) are the only group of sequences containing the SAYDTAWVA motif in seed plants.

2.5 Materials and Methods

2.5.1 Data Retrieval, Management and Taxonomy Classification

Transcriptomes assembled by SOAPdenovo-trans are downloaded from the repository at Westgrid (<http://onekp.westgrid.ca/1kp-data>). Each sample typically has about 10k scaffolds with lengths of greater than 1kb. For species having multiple samples from different tissues or developmental stages, only one sample (the merged one) is kept for further analysis, resulting in a total of 1175 representative species (Appendix A.1), 324 of which are non-seed plants. For each of these species, both assembled sequences and predicted protein sequences are pre-processed so as to make the datasets more uniform and meet the various requirements of different tools, such as shortening the sequence name or assigning a unique and easy-to-distinguish sequence id. All original information and the newly assigned sequence information is integrated into an SQLite database for future reference. The flat sequence files are then loaded into the local searchable blast database.

Although the oneKP initiative provides basic taxonomic assignments for each species, not all of them are indexed by NCBI taxonomy database (<http://www.ncbi.nlm.nih.gov/taxonomy>) due to the mismatching names between the two sources. After manually correcting these errors, each species is assigned a unique taxonomic identifier and the species tree, which displays the taxonomic relationships among these species, is generated using the taxident function in the NCBI C++ toolkit.

2.5.2 Identification of Typical Plant Terpene Synthases from Transcriptomes of 324 Non-Seed Plants

For all the assembled contigs, the longest regions without stop codons were annotated and translated using the getorf program from the EMBOSS package (Rice et al., 2000) with a minimum length of 50 amino acids. The resulting peptides were searched against the Pfam-A database locally using HMMER 3.0 hmmsearch (Finn et al., 2011) with an E value of $1e^{-5}$. Only sequences with best hits from the following four HMM profiles were considered as putative terpene synthases: Terpene_synth_C (PF03936) and Terpene synthase N terminal domain (PF01397), TRI5 (PF06330). For sequences from the same species that have 100% identity, only the longest one was kept as the representative sequence to reduce redundancy. All the putative TPS sequences were subjected to BLASTP (Altschul et al., 1990) search against the NCBI's non-redundant database using default parameters.

2.5.3 Assembly of Hornwort *Anthoceros punctatus* Genome and Identification of Terpene Synthases

The Illumina paired-end whole genome sequencing data (access number: SRR1278954) was retrieved from NCBI's Sequence Read Archive (SRA) database. The reads were assembled using SPAdes-3.1.1 (Bankevich et al., 2012; Nurk et al., 2013) and the resulting contigs and singletons were further assembled by CAP3 (Huang and Madan, 1999). The final CAP3 assembly contains 34448 sequences (16272 contigs and 18176 singletons), of

which 15596 sequences have a minimum length of 500 bp. The assembly statistics based on these 15596 sequences are described in Table 2.4 generated by QQuality ASessment Tool (QUAST) (Gurevich et al., 2013).

The final assembly was then searched for occurrences of terpene synthases using homology-based methods and *ab initio* predictors. TBLASTN searches were performed with an E-value cutoff of $1e^{-30}$ using the known genes typical plant terpene synthases from *Arabidopsis thaliana* and *Oryza sativa*, and typical plant terpene synthases identified from 324 transcriptomes of non-seed plant species. We also run SNAP (Korf, 2004) trained for *Arabidopsis thaliana* and softberry (Solovyev et al., 2006) with specific parameters for *Physcomitrella patens* on the assembly. The resulting protein sequences of predicted genes were subsequently subjected to a HMMER search against three HMM profiles (PF03936, PF01397 and PF06330).

2.5.4 Identification of Terpene Synthases from *Marchantia polymorpha* and *Sphagnum fallax*

To identify typical plant TPSs, the annotated protein sequences of *Marchantia polymorpha* genome (version 3.1) and *Sphagnum fallax* (version 0.5, download from JGI Phytozome, <https://phytozome.jgi.doe.gov/pz/portal.html>) were subjected to HMMER searches against three HMM profiles (PF03936, PF01397, PF06330). Each TPS gene was manually checked and the exon-intron information was obtained from the genome annotation.

2.5.5 Identification of Terpene Synthases from *Salvinia cucullata* and *Azolla superorganism*

The genome and transcriptome assemblies for both species were shared by Fay-Wei Li. The annotated protein sequences from both species were searched using HMMER. One TPS was found in *Azolla superorganism* and nothing was found in *Salvinia cucullata*. To make sure of not losing any terpene synthases, tBLASTn was used to search the genome and transcriptome assemblies with terpene synthases identified from non-seed plants and from *Lygodium japonicum* (GI: 747018826) and *Physcomitrella patens* (GI: 146325986) as queries. The tBLASTn searches confirmed the only one typical plant terpene synthase in *Azolla superorganism* and found 7 terpene synthases in *Salvinia cucullata*, of which 3 were identified from its transcriptome assembly and the other 4 were from its genome assembly.

2.5.6 Identification of Terpene Synthases from *Lygodium japonicum*

The *Lygodium japonicum* Transcriptome data was downloaded from <http://bioinf.mind.meiji.ac.jp/kanikusa/download.php> and searched using HMMER to identify typical plant terpene synthases. Two short sequences, isotig20305_F_T_0.4 (131 amino acids long) and isotig20305_N_P_0.4 (114 amino acids long), were identified. Both sequences are part of the known ent-kaurene synthase (GI: 747018826) in *Lygodium japonicum*.

2.5.7 Search for Terpene Synthase Genes in Twelve Algal Genomes and Transcriptomes

Typical plant TPS genes from *Arabidopsis thaliana*, *Oryza sativa* and *Physcomitrella patens* were used as blast queries to search against the genomics datasets for the following twelve algal species: *Klebsormidium flaccidum*, *Nitella mirabilis*, *Coleochaete orbicularis*, *Mesostigma viride*, *Chlamydomonas reinhardtii* v5.5, *Ostreococcus tauri*, *Ostreococcus lucimarinus* v2.0, *Micromonas pusilla* CCMP1545 v3.0, *Micromonas* sp. RCC299 v3.0, *Volvox carteri* v2.1, *Coccomyxa subellipsoidea* C-169 v2.0 and *Spirogyra pratensis* (Table 2.10). The blast searches were performed at <http://marchantia.info/blast/blast.html> and JGI Phytozome.

2.5.8 TPSs on JBrowse

For *Anthoceros Punctatus*, *Sphagnum fallax* and *Salvinia cucullata*, all the TBLASTN results and predicted protein-coding genes by SNAP and softberry are displayed on local JBrowse. The transcriptome data from non-seed plant species were mapped to the corresponding draft genome by BLASTN and the results are also displayed on JBrowse.

2.5.9 Phylogenetic Analyses of Terpene Synthases

Sequences were aligned using MAFFT (einsi) with 1000 iterations of improvement. ProtTest (Darriba et al., 2011) was used to selection of the most appropriate protein evolution model for the alignment under the Akaike Information Criterion. For the

maximum likelihood analyses, we used RAxML ([Stamatakis, 2014](#)) with 1000 bootstrap replicates under the best substitution model (JTT+G+F).

2.6 Bibliography

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–10. [31](#)
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–77. [31](#)
- Bleeker, P. M., Spyropoulou, E. A., Diergaarde, P. J., Volpin, H., De Both, M. T., Zerbe, P., Bohlmann, J., Falara, V., Matsuba, Y., Pichersky, E., Haring, M. A., and Schuurink, R. C. (2011). Rna-seq discovery, functional characterization, and comparison of sesquiterpene synthases from *solanum lycopersicum* and *solanum habrochaites* trichomes. *Plant Molecular Biology*, 77(4-5):323–36. [22](#)
- Bohlmann, J., Meyer-Gauen, G., and Croteau, R. (1998). Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proceedings of the National academy of Sciences of the United States of America*, 95(8):4126–33. [21](#)
- Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant Journal*, 66(1):212–29. [21](#)

- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–5. [34](#)
- Drew, D. P., Dueholm, B., Weitzel, C., Zhang, Y., Sensen, C. W., and Simonsen, H. T. (2013). Transcriptome analysis of *Thapsia laciniata* Rouy provides insights into terpenoid biosynthesis and diversity in Apiaceae. *International Journal of Molecular Sciences*, 14(5):9080–98. [22](#)
- Dudareva, N., Martin, D., Kish, C. M., Kolosova, N., Gorenstein, N., Fäldt, J., Miller, B., and Bohlmann, J. (2003). (E)- β -ocimene and myrcene synthase genes of floral scent biosynthesis in snapdragon: Function and expression of three terpene synthase genes of a new terpene synthase subfamily. *Plant Cell*, 15(5):1227–1241. [21](#)
- Finn, R. D., Clements, J., and Eddy, S. R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–37. [31](#)
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–5. [32](#)
- Hall, D. E., Yuen, M. M., Jancsik, S., Quesada, A. L., Dullat, H. K., Li, M., Henderson, H., Arango-Velez, A., Liao, N. Y., Docking, R. T., Chan, S. K., Cooke, J. E., Breuil, C., Jones, S. J., Keeling, C. I., and Bohlmann, J. (2013). Transcriptome resources and functional characterization of monoterpene synthases for two host species of the mountain pine beetle, lodgepole pine (*Pinus contorta*) and jack pine (*Pinus banksiana*). *BMC Plant Biology*, 13:80. [22](#)

- Han, X. J., Wang, Y. D., Chen, Y. C., Lin, L. Y., and Wu, Q. K. (2013). Transcriptome sequencing and expression analysis of terpenoid biosynthesis genes in *litsea cubeba*. *PLOS ONE*, 8(10):e76890. [22](#)
- Hayashi, K. i., Kawaide, H., Notomi, M., Sakigi, Y., Matsuo, A., and Nozaki, H. (2006). Identification and functional analysis of bifunctional ent-kaurene synthase from the moss *physcomitrella patens*. *FEBS Letters*, 580(26):6175–6181. [29](#)
- Huang, X. and Madan, A. (1999). Cap3: A dna sequence assembly program. *Genome Research*, 9(9):868–77. [31](#)
- Keeling, C. I., Weisshaar, S., Ralph, S. G., Jancsik, S., Hamberger, B., Dullat, H. K., and Bohlmann, J. (2011). Transcriptome mining, functional characterization, and phylogeny of a large terpene synthase gene family in spruce (*picea* spp.). *BMC Plant Biology*, 11:43. [22](#)
- Kollner, T. G., Schnee, C., Gershenzon, J., and Degenhardt, J. (2004). The variability of sesquiterpenes emitted from two *zea mays* cultivars is controlled by allelic variation of two terpene synthase genes encoding stereoselective multiple product enzymes. *Plant Cell*, 16(5):1115–31. [26](#)
- Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5:59. [32](#)
- Martin, D. M. and Bohlmann, J. (2004). Identification of *vitis vinifera* (-)- α -terpineol synthase by in silico screening of full-length cDNA ESTs and functional characterization of recombinant terpene synthase. *Phytochemistry*, 65(9):1223–1229. [21](#)

- Michael, T. P. and Jackson, S. (2013). The first 50 plant genomes. *Plant Gen.*, 6(2):-. [21](#)
- Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., Clingenpeel, S. R., Woyke, T., McLean, J. S., Lasken, R., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric mda products. *Journal of Computational Biology*, 20(10):714–37. [31](#)
- Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends in Genetics*, 16(6):276–7. [31](#)
- Rodriguez, M. V., Mendiondo, G. M., Cantoro, R., Auge, G. A., Luna, V., Masciarelli, O., and Benech-Arnold, R. L. (2012). Expression of seed dormancy in grain sorghum lines with contrasting pre-harvest sprouting behavior involves differential regulation of gibberellin metabolism genes. *Plant and Cell Physiology*, 53(1):64–80. [26](#)
- Sakamoto, T., Miura, K., Itoh, H., Tatsumi, T., Ueguchi-Tanaka, M., Ishiyama, K., Kobayashi, M., Agrawal, G. K., Takeda, S., Abe, K., Miyao, A., Hirochika, H., Kitano, H., Ashikari, M., and Matsuoka, M. (2004). An overview of gibberellin metabolism enzyme genes and their related mutants in rice. *Plant Physiology*, 134(4):1642–53. [26](#)
- Sasso, S., Pohnert, G., Lohr, M., Mittag, M., and Hertweck, C. (2012). Microalgae in the postgenomic era: a blooming reservoir for new natural products. *FEMS Microbiology Reviews*, 36(4):761–85. [25](#)

- Smith, M. W., Yamaguchi, S., Ait-Ali, T., and Kamiya, Y. (1998). The first step of gibberellin biosynthesis in pumpkin is catalyzed by at least two copalyl diphosphate synthases encoded by differentially regulated genes. *Plant Physiology*, 118(4):1411–9. [26](#)
- Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology*, 7 Suppl 1:S10 1–12. [32](#)
- Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–3. [35](#)
- Starks, C. M., Back, K., Chappell, J., and Noel, J. P. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. *Science*, 277(5333):1815–20. [26](#)
- Tholl, D., Chen, F., Petri, J., Gershenzon, J., and Pichersky, E. (2005). Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from arabidopsis flowers. *Plant Journal*, 42(5):757–71. [26](#)
- Trapp, S. C. and Croteau, R. B. (2001). Genomic organization of plant terpene synthases and molecular evolutionary implications. *Genetics*, 158(2):811–32. [22](#), [29](#)
- Williams, D. C., McGarvey, D. J., Katahira, E. J., and Croteau, R. (1998). Truncation of limonene synthase preprotein provides a fully active 'pseudomature' form of this monoterpene cyclase and reveals the function of the amino-terminal arginine pair. *Biochemistry*, 37(35):12213–20. [26](#)
- Xiao, M., Zhang, Y., Chen, X., Lee, E. J., Barber, C. J., Chakrabarty, R., Desgagne-Penix, I., Haslam, T. M., Kim, Y. B., Liu, E., MacNevin, G., Masada-Atsumi, S., Reed, D. W., Stout,

J. M., Zerbe, P., Zhang, Y., Bohlmann, J., Covello, P. S., De Luca, V., Page, J. E., Ro, D. K., Martin, V. J., Facchini, P. J., and Sensen, C. W. (2013). Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. *Journal of Biotechnology*, 166(3):122–34. [22](#)

2.7 Appendix

Table 2.1: Total number of TPS genes identified in six non-seed plant genomes.

Lineage	Species	TPS Count
Ferns	<i>Salvinia cucullata</i>	7
	<i>Azolla superorganism</i>	1
Lycophytes	<i>Selaginella moellendorffii</i>	18
Hornworts	<i>Anthoceros Punctatus</i>	22
Mosses	<i>Sphagnum fallax</i>	4
Liverworts	<i>Marchantia polymorpha</i>	7

Table 2.2: Typical plant terpene synthase genes in *Salvinia cucullata*.

ID	Contig / Gene_loc	Strand	Start	End	Length
ScuTPS1	XLOC_018097 g.99102	+	119	2299	727
ScuTPS2	XLOC_008367 g.46493	+	3	620	206
ScuTPS3	XLOC_008367 g.46489	+	81	1586	502
ScuTPS4	scf7180000009411	+	140260	147974	447
ScuTPS5	scf7180000010110	-	1257296	1259664	260
ScuTPS6	scf7180000010110	-	1251971	1256640	270
ScuTPS7	scf7180000010334	-	568868	581876	611

Table 2.3: Typical plant terpene synthase genes in *Azolla superorganism*.

ID	Contig	Strand	Start	End	Length
AsuTPS1	scf7180000009599	-	1241006	1248428	894

Table 2.4: Assembly statistics for *Anthoceros punctatus*.

Statistics without reference	scaffolds.fasta.cap3
# contigs	15596
# contigs (≥ 0 bp)	34448
# contigs (≥ 1000 bp)	13144
Largest contig	94143
Total length	94470290
Total length (≥ 0 bp)	97989643
Total length (≥ 1000 bp)	92634135
N50	12462
N75	5614
L50	2103
L75	4939
GC (%)	49.52
Mismatches	
# N's	89463
# N's per 100 kbp	94.7

Table 2.5: Typical plant terpene synthase genes in *Anthoceros punctatus*.

ID	Contig	Strand	Start	End	Length
ApuTPS1	Contig1003	-	5384	6147	186
ApuTPS2	Contig11875	-	33079	36795	780
ApuTPS3	Contig1639	-	116	5200	880
ApuTPS4	Contig1919	+	1570	2527	225
ApuTPS5	Contig2366	-	4054	7915	841
ApuTPS6	Contig3848	+	585	1607	238
ApuTPS7	Contig4529	-	79	1572	497
ApuTPS8	Contig5467	-	219	1859	343
ApuTPS9	Contig5884	+	49	2024	379
ApuTPS10	Contig654	+	419	3724	676
ApuTPS11	Contig68	+	12275	14825	655
ApuTPS12	Contig799	+	4488	7388	422
ApuTPS13	Contig9716	+	118	5299	880
ApuTPS14	NODE_10797_length_1803_cov_7.30435_ID_21593	+	223	1267	159
ApuTPS15	NODE_11022_length_1747_cov_7.20567_ID_22043	-	240	1708	192
ApuTPS16	NODE_2742_length_9352_cov_7.29709_ID_5483	-	3261	7292	796
ApuTPS17	NODE_4282_length_6121_cov_7.00363_ID_8563	+	106	2462	615
ApuTPS18	NODE_4542_length_5765_cov_6.30368_ID_9083	+	30	2098	381
ApuTPS19	NODE_50_length_49024_cov_6.82293_ID_99	+	15844	19514	868
ApuTPS20	NODE_6403_length_3849_cov_17.8529_ID_12805	-	333	3138	600
ApuTPS21	NODE_717_length_21934_cov_18.8625_ID_1433	-	1739	5575	858
ApuTPS22	NODE_8602_length_2556_cov_7.03039_ID_17203	+	1289	2499	403

Table 2.6: Typical plant terpene synthase genes in *Marchantia polymorpha*.

scaffold	ID	length	DXDD	DDXXD	NSE/DTE	Subfamily
scaffold_15	Mapoly0015s0008.1	882	+	-	-	C
scaffold_46	Mapoly0046s0122.1	879	+	-	-	h
scaffold_50	Mapoly0050s0107.1	791	+	+	+	C
scaffold_97	Mapoly0097s0049.1	858	-	+	+	e/f
scaffold_167	Mapoly0167s0025.1	795	-	+	+	e/f
scaffold_207	Mapoly0207s0001.1	892	+	+	+	C
scaffold_221	Mapoly0221s0001.1	779	+	+	+	C

Table 2.7: Assembly statistics for *Sphagnum fallax*.

Statistics without reference	sphag_f_v0.1
# contigs	1929
# contigs (≥ 0 bp)	1929
# contigs (≥ 1000 bp)	1432
Largest contig	9831565
Total length	397359651
Total length (≥ 0 bp)	397359651
Total length (≥ 1000 bp)	396998157
N50	1820331
N75	926644
L50	58
L75	134
GC (%)	36.44
Mismatches	
# N's	1372291
# N's per 100 kbp	345.35

Table 2.8: Typical plant terpene synthase genes in *Sphagnum fallax*.

Id	Contig	Start	End	Strand	Protein_Length
Sphfalx0014s0037.1	super_14	437280	444282	+	858
Sphfalx0136s0053.1	super_136	830109	836630	-	720
Sphfalx0136s0055.1	super_136	857468	868275	+	894
Sphfalx0790s0001.1	super_790	29	3476	-	379

Table 2.9: Number of TPSs genes from non-seed plant transcriptomes.

Lineage	TPS Count	Species with TPSs	Total Species
Monilophytes	47	19	70
Lycophytes	54	21	21
Hornworts	23	3	7
Mosses	40	22	41
Liverworts	89	23	27
Charophyta	0	0	47
Chlorophyta	0	0	111
Total	253	88	324

Table 2.10: List of charophyta transcriptomes/genomes.

Phylum	Species	Data_type
Charophyta	<i>Klebsormidium flaccidum</i>	Genome
	<i>Nitella mirabilis</i>	Transcriptome
	<i>Coleochaete orbicularis</i>	Transcriptome
	<i>Mesostigma viride</i>	Transcriptome
Chlorophyta	<i>Chlamydomonas reinhardtii</i> v5.5	Genome
	<i>Ostreococcus tauri</i>	Genome
	<i>Ostreococcus lucimarinus</i> v2.0	Genome
	<i>Micromonas pusilla</i> CCMP1545 v3.0	Genome
	<i>Micromonas</i> sp. RCC299 v3.0	Genome
	<i>Volvox carteri</i> v2.1	Genome
	<i>Coccomyxa subellipsoidea</i> C-169 v2.0	Genome
	<i>Spirogyra pratensis</i>	Transcriptome

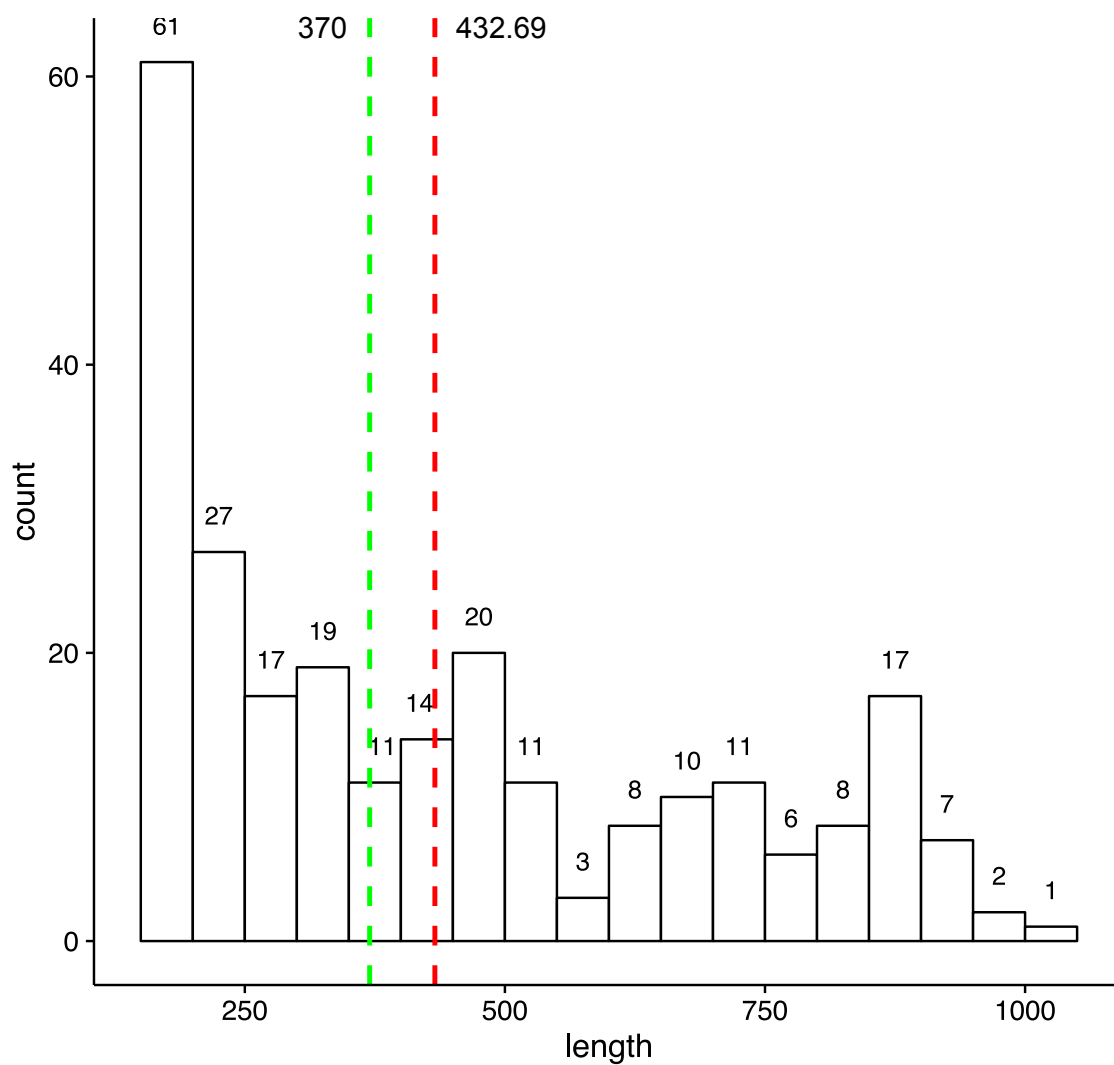


Figure 2.1: Distribution of protein lengths for typical plant terpene synthases in non-seed plants. The green dash line indicates the median length, while the red line indicates the average protein length.

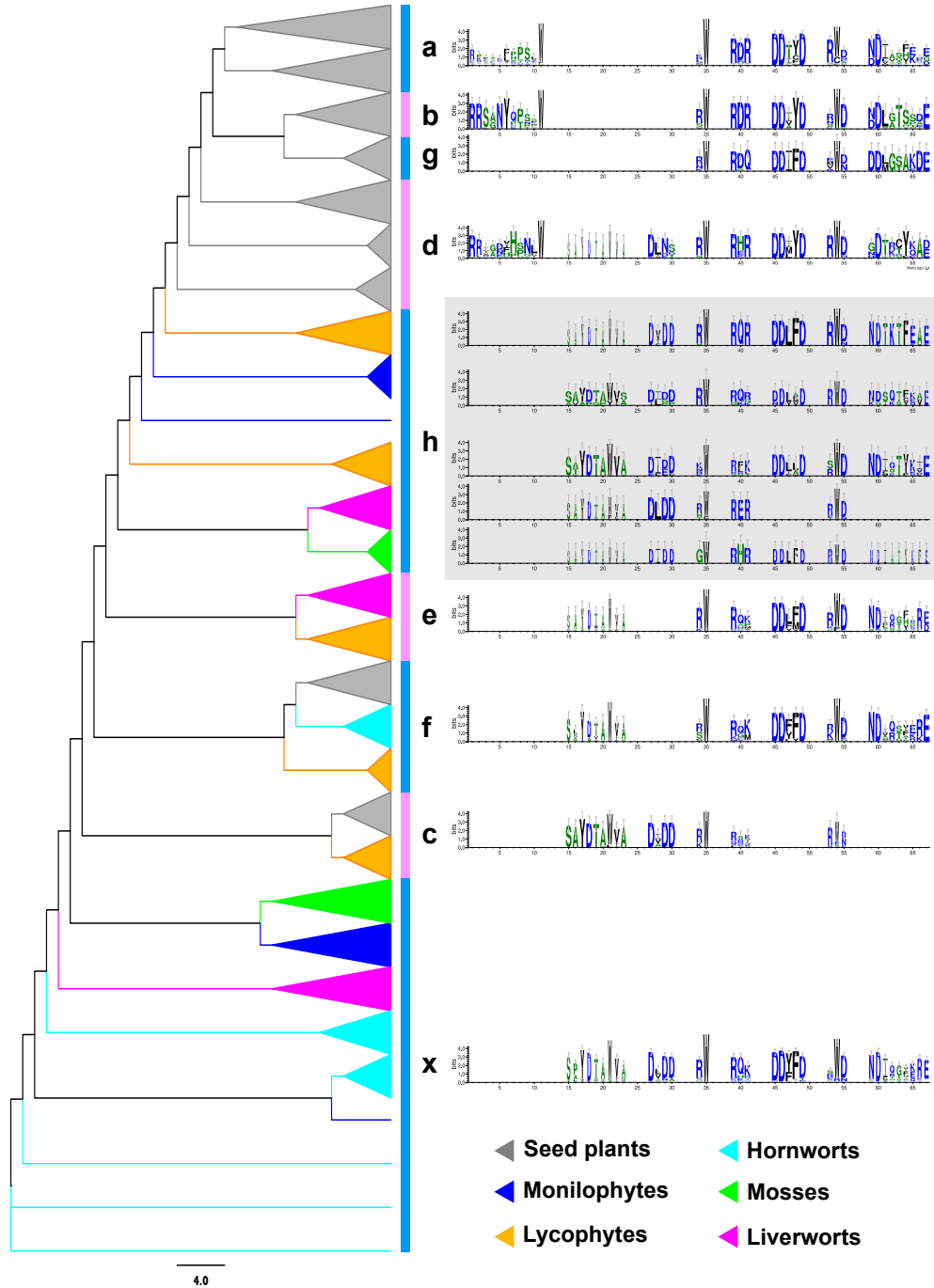


Figure 2.2: Phylogenetic relationships of terpene synthases from non-seed plants and selected TPSs from seed plants. Weblogs show the conserved motifs found in each sunfamily.

Figure 2.3: Alignment of the conserved motifs found in 339 terpene synthases. Each motif is separated by three consecutive dots (...).

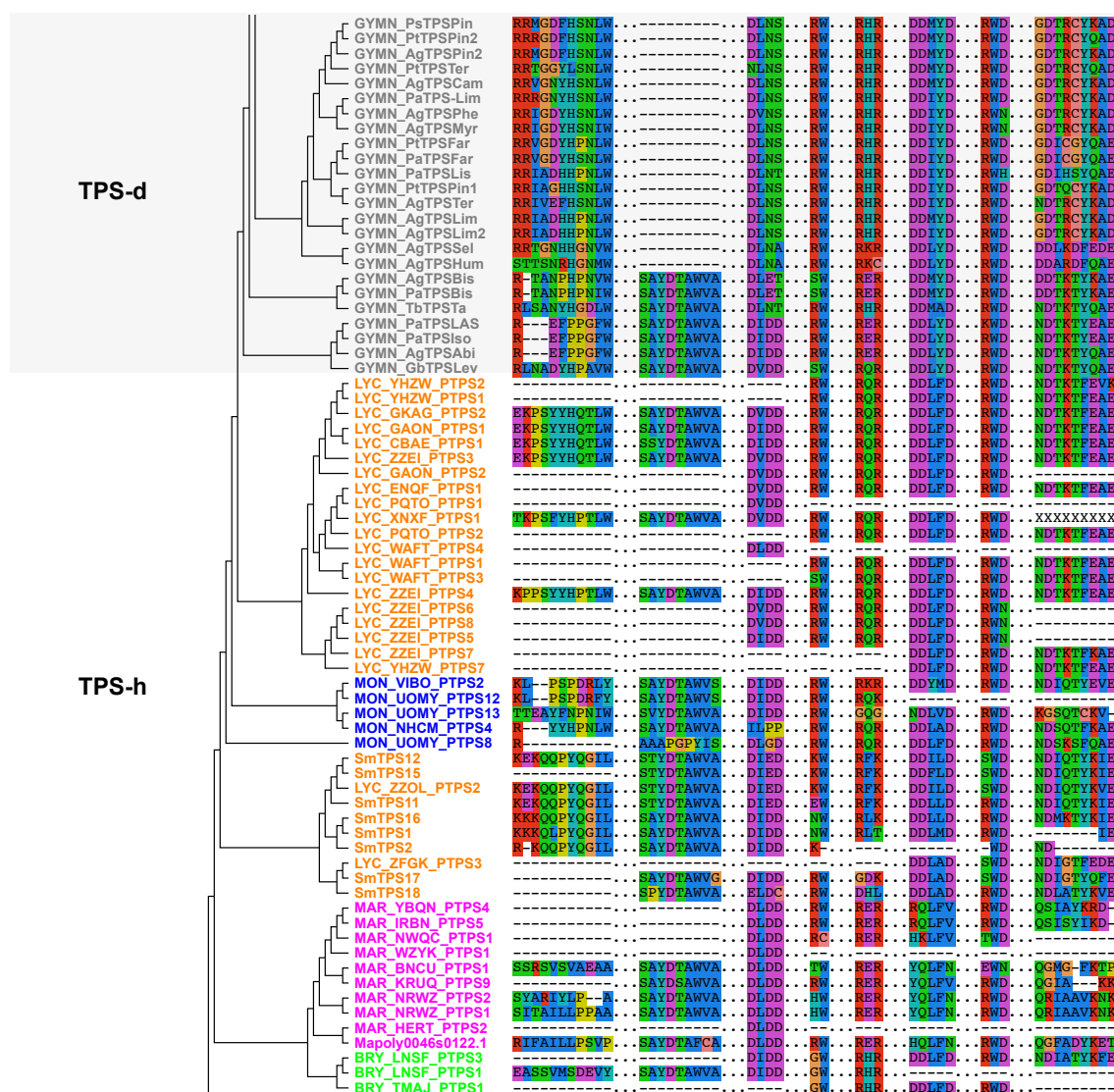


Figure 2.3: Continued.

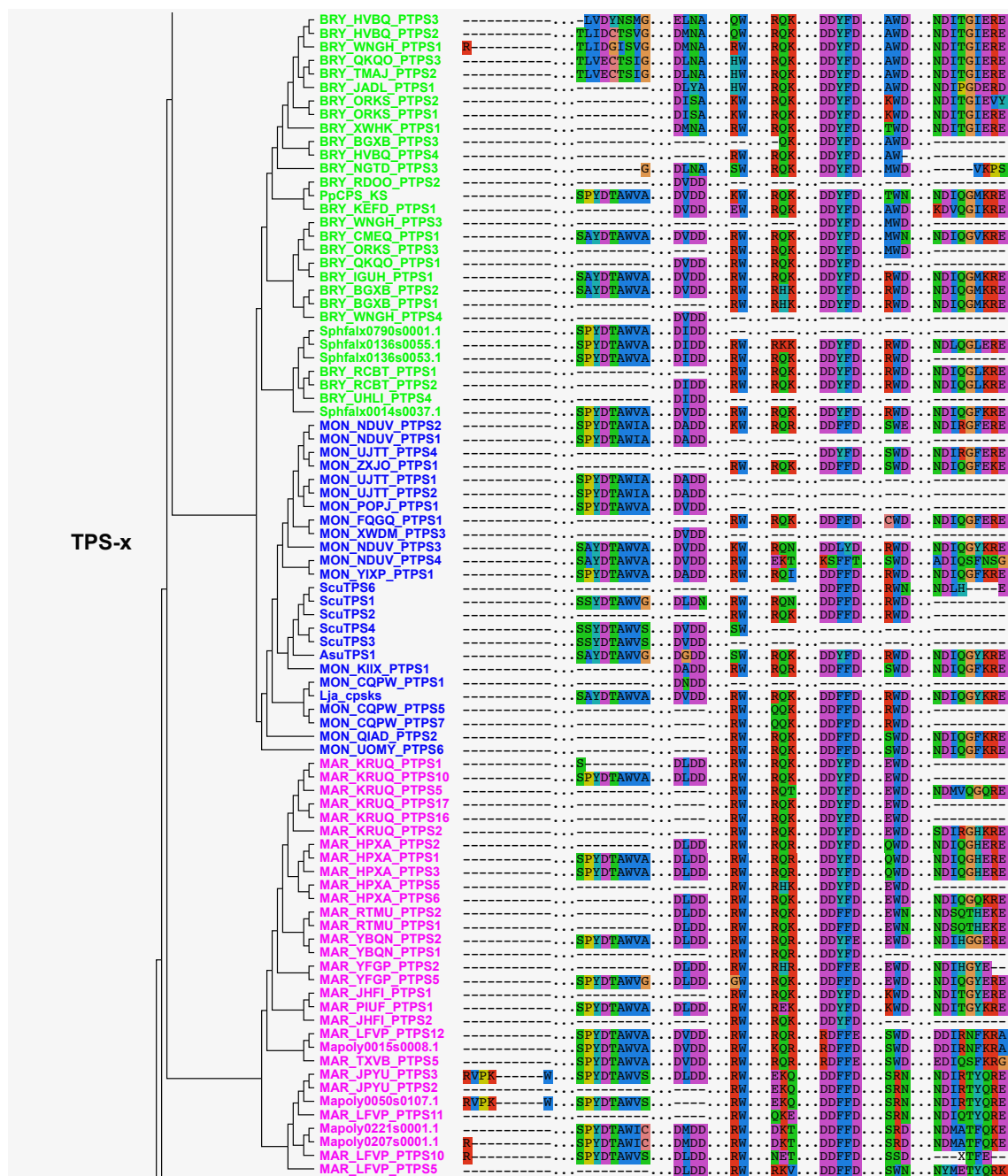


Figure 2.3: Continued.

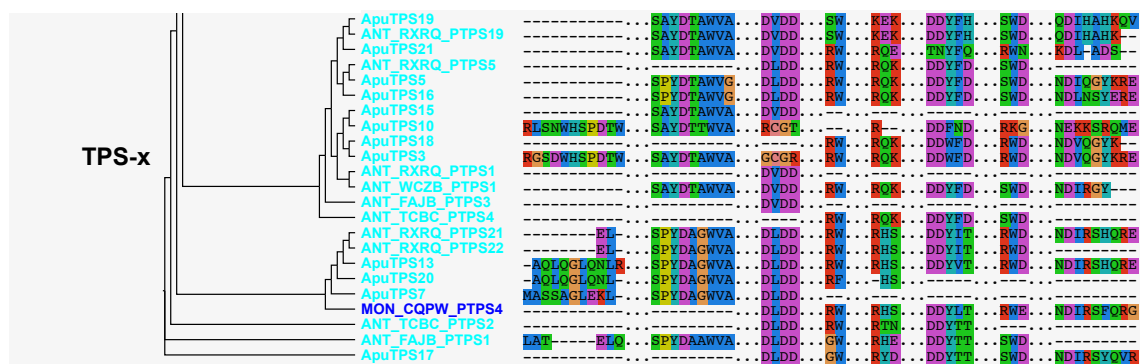


Figure 2.3: Continued.

Chapter 3

Microbial Type Terpene Synthase

Genes Occur Widely and Specifically

in Non-seed Land Plants

Author Contributions: Qidong Jia conceptualized and conducted all the bioinformatics analyses, and co-wrote the manuscript.

3.1 Abstract

Microbial terpene synthases-like genes (*MTPSLs*) are a novel type of plant terpene synthase genes that have been previously identified only in the lycophyte *Seleginella moellendorffii*. *MTPSLs* are more closely related to bacterial and fungal terpene synthases than typical plant terpene synthases. The goal of this study is to investigate the distribution, evolution and biochemical functions of plant *MTPSL* genes. By analyzing the transcriptomes of 1103 plant species ranging from green algae to flowering plants, putative *MTPSL* genes were identified predominantly from non-seed land plants including bryophytes, lycophytes and ferns. Based on phylogenetic analyses and the confirmation of representative members in sequenced plant genomes, the vast majority of these genes (97% and all from non-seed plants) were assigned with confidence to be plant genes rather than microbial contamination. The general absence of *MTPSL* genes in seed plants was further supported by the negative result of searching for *MTPSL* genes from the sequenced genomes of a wide range of seed plants. Thus, it was concluded that *MTPSL* genes occur widely and specifically in non-seed plants. *MTPSLs* from non-seed plants are divergent, forming four major groups. Two of the four groups showed closer relatedness to bacterial terpene synthase, while the other two groups are more closely related to fungal terpene synthases. Two of the four groups contain canonical aspartate-rich 'DDxxD' motif. The third group has 'DDxxxD' motif and the fourth group has only the first two DD conserved in this motif. Representative members from each of the four distinct groups with varying aspartate-rich motif were characterized for terpene synthase

activities. These MTPSLs displayed diverse catalytic functions as monoterpene synthases and sesquiterpene synthases.

3.2 Introduction

Terpenoids are the largest class of secondary metabolites made by plants (Connolly and Hill, 1991). Particularly rich are monoterpenes (C_{10}), sesquiterpenes (C_{15}) and diterpenes (C_{20}). Seed plants (angiosperms and gymnosperms) are well known for their rich production of terpenoids (Tholl, 2006). For non-seed plants, only liverworts are known to be rich producer of terpenoids (Asakawa, 2008). Terpenoids have diverse biological and ecological functions (Gershenzon and Dudareva, 2007). Some of the functions are specific to certain plant lineages. For instance, volatile terpenoids as floral scent compounds are involved in attracting pollinators for pollination. Many of the functions of terpenoids especially in plant defenses against biotic and abiotic stresses seem to be universal for all plant lineages (Mithofer and Boland, 2012). The biosynthesis of terpenoids in all plant lineages, particularly in basal land plants, is therefore the key to understand the roles of terpenoids for adaptation of terrestrial plants to the land.

Terpene synthases (TPSs) are pivotal enzymes for terpenoid biosynthesis. From both the primary proteins sequences and structural perspective, terpene synthases form a superfamily. TPS genes have been found in plants, fungi and bacteria (Chen et al., 2011). While all TPSs share certain structural similarities (Köksal et al., 2011), typical plant TPSs and microbial TPSs from fungi and bacteria have very limited sequence similarities and are

therefore only distantly related evolutionarily (Li et al., 2012). Much progress has been made in our understanding of the function and evolution of typical plant TPSs, which are conserved in land plants (Chen et al., 2011). Typical plant TPSs form subfamilies. Monoterpene synthases and sesquiterpene synthases have been proposed to have evolved independently in gymnosperms and angiosperms, they are absent in the non-seed plants with sequenced genome, both monoterpene synthases and sesquiterpenes appear to be absent (Degenhardt et al., 2009), and their typical plant TPSs are only of diterpene synthase type. Therefore, the molecular basis underlying the biosynthesis of monoterpenes and sesquiterpenes identified from non-seed plants remains elusive.

Recently, an important discovery made in the field of plant terpene biosynthesis is the discovery of microbial type terpene synthase genes in plants. Such genes have been named microbial terpene synthase-like genes (MTPSLs) (Li et al., 2012). Unlike typical plant TPSs, which contains either α - β - γ three domains or α - β two domains (Köksal et al., 2011), MTPSLs contain only α domain. Phylogenetic analysis showed that MTPSLs from *Selaginella moellendorffii* have closer relatedness to microbial TPSs, especially fungal TPSs, than to typical plant TPSs (Li et al., 2012). So far, MTPSL has only been identified in a lycophyte *S. moellendorffii*, not in the moss *Physcomitrella patens* and 15 species of flowering plants with sequenced genome (Li et al., 2012). This finding raises many intriguing questions about the origin, evolution and function of this new type of terpene synthase genes in the plant kingdom. The goal of this study is to determine the distribution of MTPSL in the plant kingdom, to infer their evolution and to determine their biochemical activities.

3.3 Results and Discussion

3.3.1 Terpene Synthase Genes of Microbial Type are Highly Enriched in the Transcriptomes of Non-Seed Land Plants

To determine the distribution of *MTPSL* genes in the plant kingdom, we took a transcriptomic approach. The transcriptomes for 1103 species (Table 3.1) of green plants (779 species of seed plants, 166 species of non-seed land plants, 47 species of Charophyta and 111 species of Chlorophyte) generated from the OneKP initiative (<https://sites.google.com/a/uAlberta.ca/onekp/>) were searched for microbial type terpene synthase genes using a HMMER method as previously described (Li et al., 2012). A total of 712 *MTPSL* genes were identified from the transcriptomes of a total of 146 species. Strikingly, the vast majority of *MTPSL* genes (706 out of the 712 *MTPSL* genes or 99.2%) were found from the transcriptomes of non-seed land plants (Figure 3.1).

Bryophytes consist of three groups: hornworts, mosses and liverworts, which have 7, 41 and 26 species in the onekp dataset, respectively. The number of species whose transcriptomes contain *MTPSL* genes for hornworts, mosses and liverworts are 3, 30 and 24. Among the 22 species of lycophytes, 21 were found to contain *MTPSL* genes in their transcriptomes. The Monilophytes group contains 70 species, of which 65 species were found to contain *MTPSL* genes in their transcriptomes. The median number of *MTPSL* genes from the transcriptome of each species for hornworts, mosses, liverworts, lycophytes and ferns are 0, 1, 8, 3 and 4.5, respectively (Figure 3.1 and Table 3.2). Among

all species, the fern *Cystopteris utahensis* (a tetraploid) was found to contain the most *MTPSL* genes with 20 members (Table 3.1).

Except the genes found from non-seed land plants, extremely low occurrence of *MTPSL* genes was also found from the transcriptomes of seed plants and Charophyta. Among the 779 species of seed plants, only two species: *Phytolacca bogotensis* and *Opuntia* *sp.* were found to contain *MTPSL* genes in their transcriptomes with one and four members respectively. Among the 47 species of Charophyta, only one species *Micrasterias fimbriata* contained *MTPSL* gene (one member) in its transcriptome. No *MTPSL* genes were found from the transcriptomes of 111 species of Chlorophyta.

3.3.2 The Majority of *MTPSL* Genes Identified from Plant Transcriptomes Forms Four Groups Clustered with Either Fungal or Bacterial Terpene Synthases

Phylogenetic analysis was performed for the 712 *MTPSL* genes identified from plant transcriptomes together with 48 known *MTPSL* genes from *S. moellendorffii* and selected terpene synthase genes from bacteria and fungi. The resulting phylogenetic tree shown that the distribution of *MTPSL* genes in non-seed plants exhibit lineage-specific characteristic and the majority of them (690 out of 712) were clustered into four major groups (Figure 3.2).

The group I, the second largest *MTPSL* gene group, contains about 86% of *MTPSL* genes (152 out of 177) from 23 species of liverworts, 34% of *MTPSL* genes (27 out of 79) from 10

species of mosses and 28% of *MTPSL* genes (23 out of 83) from 9 species of lycophytes. The group II was composed of *MTPSL* genes mainly from mosses (about 66% of *MTPSL* genes from 24 species) and hornworts (50% of *MTPSL* genes from all 3 species that have *MTPSL* genes found). There was also one *MTPSL* gene found in a liverwort species, *Scapania nemorosa*. Members of *MTPSL* genes in this species were also present in group I and III. The group III, the smallest group, contains 4 *MTPSL* genes from 3 species of hornworts and 14 *MTPSL* genes from 7 species of liverworts. These 18 genes were clustered with the fungal Trichodiene synthase (TRI5) genes. The group IV contain almost all *MTPSL* genes found from 65 species of ferns (Monilophytes, 352 out of 353) and about 70% of *MTPSL* genes in Lycophytes (58 out of 83). The known *MTPSL* genes from *S. moellenhoffii* were closely related with *MTPSL* genes from the transcriptomes of other Lycophyte species.

There are 22 *MTPSL* genes that locate outside of the four major groups and disperse within the fungal or bacterial clade (Figure 3.2 and Table 3.3). They are categorized in group V. *MTPSL* genes found in two seed plant species, *Opuntia* sp. (4 members) and *Phytolacca bogotensis* (1 member), and one green algae, *Micrasterias fimbriata* (1 member), are in this list.

3.3.3 The Majority of *MTPSL* Genes Identified from Plant Transcriptomes are Plant Genes

Putative *MTPSL* genes identified in the plant transcriptomes could have one of the two origins: plants and microbes. The reason why some of these genes can be of microbial

origin is that all plants are known to be associated with endophytic fungi and bacteria (Faeth and Fagan, 2002; Gaiero et al., 2013; Rosenblueth and Martinez-Romero, 2006; Ryan et al., 2008). Some of the plant transcriptomes may unavoidably contain transcripts from microbes that were associated with the plants. The 22 *MTPSL* genes from group V (Table 3.3) have a high probability of being derived from plant-associated microbes. The similarities of these 22 *MTPSL* genes with microbial terpene synthase genes are extremely high. For Instance, IRBN_*MTPSL*8, one of the 22 *MTPSL* genes identified from the transcriptome of liverwort *S. Nemorosa*, is 91% identical to a terpene synthase from a fungus *Serendipita vermifera* (Figure 3.6). *S. Vermifera* is a member of the ubiquitous fungal order Sebaciales, which has the greatest diversity of mycorrhizal-types (Weiss et al., 2004). The association of liverworts with sebaciales as endophytes has been documented (Weiss et al., 2011). Therefore, IRBN_*MTPSL*8 is most likely a fugal terpene synthase gene derived from *S. nemorosa*-associated endophyte. Certainly, at this time it is impossible to rule out the possibility that some of the 22 *MTPSL* genes are plant genes obtained from microbes very recently through horizontal gene transfer, which however will require extensive investigations. Consequently, these 22 *MTPSL* genes are not being further considered in this study. Three lines of evidence were used to judge that the most majority, if not all, of the remaining 690 *MTPSL* genes in the four major clades are plant genes.

The first line of evidence is the clade patterns of these genes based on phylogenetic analysis. Forming four groups, the *MTPSL* genes in each group were identified from multiple plant species from multiple plant lineages (Figure 3.2). This pattern supports

that the MTPSLs in each group share a common evolutionary origin. The plant species associated with each clade are highly divergent, have diverse geographic distribution, and their plant materials were acquired in diverse fashion (from axenic culture to field-collected sample). Plant-associated microbes are highly diverse and some exhibit species-specificity (Opelt et al., 2007). Terpene synthase genes from different lineages of bacteria can be highly diverged (Yamada et al., 2015). A single clade of these genes if from microbes would suggest all these microbes from diverse plants are closely related, which seems to be very unlikely. Therefore, we can rule out that possibility that the plants in the same group are associated with closely related endophytes to give out MTPSL transcripts in the plant transcriptomes. Within each group, the MTPSLs from the same lineages showed closer relatedness. For Instance, in group I, the MTPSLs from liverworts, mosses and lycophytes form three distinct subclades (Figure 3.2). The MTPSLs from closely related species are also most closely related. For instance, the analyzed mosses include three species from the same genus *Sphagnum*: *S. lescurii*, *S. palustre* and *S. recurvatum*. In group I, the MTPSLs from these three species clustered together (Figure 3.2).

The second line of evidence came from the analysis of putative *MTPSL* genes from axenic culture. The Liverwort *S. nemorosa* was selected for this purpose. A total of 8 putative *MTPSL* genes were identified in the transcriptome of *S. nemorosa* (Table 3.1). They belong to group I (5 genes), group II (1 gene) and group III (1 gene). The eighth gene belongs to group V. Axenic culture of *S. nemorosa* was initiated by germinating isolated spore. Its culture is therefore free of contamination of endophytic microbes. We extracted genomic DNA from *S. nemorosa* and did PCR to amplify DNA fragment for the

eight putative *MTPSL* genes. Six of the eight *MTPSL* genes (five from group I and one from group III) were amplified and confirmed by sequencing. However, the amplification of the members from groups II and V failed. The member from group V has been suggested to be contamination from endophytic fungi (Figure 3.6). analysis of the onekp transcriptomes has suggested that *s. nemorosa* was contaminated with unknown source of plant (<https://pods.iplantcollaborative.org/wiki/display/iptol/Sample+source+and+purity>). The putative *MTPSL* from *S. nemorosa* in group II is the only *MTPSL* from liverworts to be assigned in this group. Together with contamination analysis, our experimental data suggested this single gene is derived contaminated plant material. Overall, This experimental study confirmed that the *MTPSL*s from *S. nemorosa* in group I and III are plant genes.

The third line of evidence came from the confirmation of representative *MTPSL* genes together with their neighboring genes from each of the four groups to be bona fide plant genes. The genes in group IV are clustered with *SmMTPSL* genes from *S. moellendorffii*, which have been determined to be plant genes (Li et al., 2012). Therefore, no additional analysis is needed. For obtain similar evidence for group II and III, we analyzed the hornwort *Anthoceros punctatus*, whose draft genome has been recently sequenced (Li et al., 2014). By assembling the raw genome sequence and then identify *MTPSL* genes using the same HMMER approach, 16 *MTPSL* genes were identified, of which 11 were considered full-length (Table 3.4). Phylogenetic analysis showed that three of the 11 genes belong to group II and the rest eight to the group III (Figure 3.7). *ApMTPSL1* from group II and *ApMTPSL2* from group III were selected as representative for the following study.

In the assembled genome of *A. punctatus*, *ApMTPSL1* is neighbored with a Cytochrome P450 gene, to which the most homologous is from the plant *Oryza Brachyantha* (Figure 3.3A). Similarly *ApMTPSL2* is neighbored with a leucine-rich receptor-like kinase gene, to which the most homologous gene is from the plant *Theobroma cacao* (Figure 3.3B). We extracted genome DNA from *A. punctatus* grown in axenic culture and did PCR to amplify the coding sequence of *ApMTPSL1/2* gene and their respective neighboring gene. The amplified DNA fragment was cloned and fully sequenced. It was confirmed that both *ApMTPSL1* (Figure 3.3A) and *ApMTPSL2* genes (Figure 3.3B) reside in the *A. punctatus* genome and neighbors with a plant gene. The similar evidence for the group I was obtained with a moss *Sphagnum fallax*, whose genome sequence is available (JGI site). The sequenced *S. fallax* genome contains 21 *MTPSL* genes, All of which belong to group I (Table 3.5). Representative *MTPSL* genes from *S. fallax* were confirmed to reside in its genome and neighbor with a copia-like retrotransposable element, to which the most homologous is from the plant *Arabidopsis Thaliana* (Figure 3.3C). The confirmation of representative *MTPSL* genes in each group to reside in plant genome supports that their apparent orthologs in each group are plant genes.

3.3.4 Evolutionary Implications of Non-Seed Plant-Specific MTPSLs

The confirmation of the vast majority of *MTPSL* genes identified from the OneKP transcriptomes to be plant genes indicates that *MTPSL* genes occur widely in non-seed

plants. The *MTPSL* genes in group I appear to be most conserved. It contains *MTPSL* genes from bryophytes (liverworts and mosses) and lycophytes (Figure 3.2). This implies the presence of *MTPSL* genes in the common ancestor of land plants. It is interesting to notice that *MTPSL* genes are generally absent in the transcriptomes of green algae (Figure 3.1). To provide further evidence about the presence/absence of *MTPSL* genes in green algae, *MTPSL* genes were searched in the sequenced genomes for six species of Chlorophyte and one Species of Charophyta (Table 3.6). No *MTPSL* genes were detected from these sequenced green algae. It will be interesting to determine whether *MTPSL* genes originated in algal ancestor of land plants or in the common ancestral land plants. The absence of *MTPSL* genes in the transcriptomes of a wide range of Charophyta and the genome of *Klebsormidium flaccidum* (Hori et al., 2014) does not support the evolution of *MTPSL* genes in the algal ancestor of land plants. If so, the evolution of *MTPSL* genes may be associated with plants' transition from freshwater to land. The pioneer land plants faced harsh environment and had to overcome many biotic and abiotic stresses. Many products of terpene synthases are volatile hydrocarbons. It is possible that such compounds are not so useful for an aquatic environment but become important in a terrestrial ecosystem.

It is interesting to observe that *MTPSL* genes from non-seed plants showed different relatedness to bacterial terpene synthase and fungal terpene synthases. The groups III and IV are most closely related to fungal terpene synthases, while the group I are most closely related to bacterial terpene synthases. The group II is most related to bacterial terpene synthases, which however resides in the fungal clade. These clade patterns suggest

complex evolutionary history of microbial type terpene synthases in these kingdoms of life. While it is possible that microbial type terpene synthase genes are ancestral universal genes and evolved differently, the absence of terpene synthase genes in other lineages of life expect bacteria, fungi and plants does imply that such a distribution pattern may be a result of horizontal gene transfer. It is premature to make strong claims about the donor and receipts of such horizontal gene transfer events. If the terpene synthase genes in bacteria and fungi are presumed to be ancestral to MTPSLs, then *MTPSL* genes can be explained by multiple hgt events from bacteria and fungi. Unfortunately, our understating of terpene synthase genes in bacteria and fungi is very much limited (Quin et al., 2014; Schmidt-Dannert, 2015; Yamada et al., 2015). We just showed the first case of hgt of terpene synthase genes from bacteria to fungi, which will undoubtedly confound our inference of the evolution of microbial type terpene synthase genes.

Equally interesting is the absence of *mtpls* genes in seed plants (Figure 3.1). The absence of *MTPSL* genes in transcriptomes could be due to the lack of expression. To gain additional evidence about the presence/absence of *MTPSL* genes in seed plants, we analyzed the genome sequences of 48 species of seed plants (Table 3.6), but no *mtpls* genes were identified. As mentioned previously, land plants contain typical plant terpene synthases, which catalyzes the similar biochemical reactions for the production of terpenoids as microbial type terpene synthases but are only distantly related to MTPSLs (Li et al., 2012). There is a distinct difference in typical plant terpene synthases between those from seed and non-seed plants. From the non-seed plants that have been studied, their typical plant terpene synthases are diterpene synthases. In contrast, typical

terpene synthases from seed plants include monoterpene synthases and sesquiterpene synthases in addition to diterpene synthases (Chen et al., 2011). Several *MTPSL* genes from *S. moellendorffii* have been demonstrated to function as monoterpene synthases and sesquiterpene synthases (Li et al., 2012). We hypothesize that most *MTPSL* genes in non-seed plants function as monoterpene synthases and sesquiterpene synthases. After HGT, the transferred gene has to be established and maintained in the population. Since seed plants evolved their own plant-like tps that produce terpenes for various functions, the pressure for maintaining *MTPSL* from current hgt events in higher plants is probably not high enough (it is just one more tps in an already existing family of about 10-30 members) – thus this would be one of the reasons that there are no *MTPSL* in seed plants.

3.3.5 Biochemical Function of Selected *MTPSL*s: Diversity of Activities

The presence of *MTPSL* genes only in basal land plants poses an intriguing question about their biological functions. In seed plants, TPSs are responsible for the production of a diversity of terpenoids important for plant-environment interactions, especially as defenses. Some basal land plants, such as liverworts (Nagashima et al., 2001) and mosses (Saritas et al., 2001), also produce a vast diversity of terpenoids. However, essentially nothing is known about how such terpenoids in basal land plants are synthesized and their biological functions. Based on this knowledge, we hypothesize that *MTPSL*s in basal land plants make diverse terpenoids for plant defense against various biotic and

abiotic stresses. As a first step, in this study, we chose to characterize selected *MTPSL* gene biochemically.

Terpene synthases contain two highly conserved motifs: the ‘DDxxD’ motif and ‘NSE’ motif, which are involved in substrate binding (Hyatt et al., 2007; Koksai et al., 2011; Starks et al., 1997). *MTPSL* genes in all four groups contain the highly conserved “NSE” motif in their C-terminal (Figure 3.4). In contrast, these four groups of proteins show difference in the aspartate-rich motif (Figure 3.4). The conserved ‘DDxxD’ motif is present in genes in group I and IV. Genes in group II display a conserved ‘DDxxxD’ motif, while genes in groups III have no clear ‘DDxxD’ existed. Although the ‘DDxxD’ motif is prevalent in group IV, many members of *MTPSL* genes in this group display a ‘DDDDD’ or ‘DxDDD’ motif instead of the ‘DDxxD’ motif.

To determine whether the four groups of *MTPSL*s with varying aspartate-rich motifs have terpene synthase activities, representative members from each group were chosen and synthesized. The synthesized cDNAs were cloned into protein expression vector. Recombinant proteins were produced in *E. coli* and tested for terpene synthase activities. Two types of substrates were tested: farnesyl diphosphate and geranyl diphosphate, which are substrates for sesquiterpenes and monoterpenes, respectively.

3.4 Conclusions

In this study, the *MTPSL* genes were systematically mined from large-scale transcripts. Out of 779 seed plant species, only 5 *MTPSL* genes were found from 2 species,

while 706 *MTPSL* genes were found from 143 non-seed land plant species. So, besides the genome of *Selaginella moellendorffii*, *MTPSL* genes are widely and specifically distributed in non-seed land plants. Based on the phylogenetic analysis and experimental validation of selected *MTPSL* genes, the vast majority of them are believed to be real plant genes, although a small percentage of them are most likely from contamination. *MTPSL* genes form four lineage specific groups that exhibit diverse structural features, which indicates multiple evolutionary origins of plant *MTPSL* genes. Biochemical studies of selected genes shown that *MTPSL* genes function as monoterpene synthase and sesquiterpene synthases. However, the biological functions of *MTPSL* genes are largely unknown. This study will pave the way for characterization of the secondary metabolites produced by these *MTPSL* genes and investigation the biological functions of *MTPSL* genes in non-seed land plants. This study also provokes interest in inferring the evolutionary origins of *MTPSL* genes.

3.5 Materials and Methods

3.5.1 Identification of Terpene Synthases of Microbial Type from Transcriptomes and Sequenced Genomes

The One Thousand Plants Project (OneKP; <http://www.onekp.com>) has sequenced transcriptomes for over 1000 non-model plant species spanning almost all major plant clades (from green algae to flowering plants). All transcriptomes were pre-assembled with the SOAPdenovo-Trans assembler (Xie et al., 2014). Transcriptomes of 1103 representative

species (Table 3.1) were analyzed in this study. For all the assembled contigs, the longest regions without stop codons were annotated and translated using the getorf program from the EMBOSS package (Rice et al., 2000) with a minimum length of 150 amino acids. The resulting peptides were searched against the Pfam-a database locally using HMMER 3.0 Hmsearch (Finn et al., 2011) With an E value of $1e^{-5}$. Only sequences with best hits from the following four HMM profiles were considered as putative terpene synthases: Terpene_synth_C(PF03936) and Terpene Synthase N terminal domain (PF01397), TRI5 (PF06330) and SmMTPSLs (a profile created by using 48 microbial type tpss identified from *Selaginella moellendorffii*). For sequences from the same species that have 100% identity, only the longest one was kept as the representative sequence to reduce redundancy.

All the putative tps sequences were subjected to blastp search against the NCBI's non-redundant database using default parameters and the "microbial tps-like protein" annotation was based on the taxonomies of the top 10 best hits for each sequence, heavily weighted on the taxonomy of best BLASTP hit. Since *S. moellendorffii* is the only species reported having both plant and microbial types of terpene synthases, for sequences with the best hits from this species, another criteria taking into account the hits' type of terpene synthases or the types they are mostly close to was applied.

3.5.2 Assembly of Hornwort *Anthoceros punctatus* Genome and Identification of *MTPSL* Genes

For *Anthoceros punctatus*, the illumina paired-end whole genome sequencing data (Access Number: SRR1278954) was retrieved from NCBI's Sequence Read Archive (SRA) Database. The reads were assembled using SPAdes-3.1.1 (Bankevich et al., 2012) and the resulting contigs and singletons were further assembled by CPA3 (Huang and Madan, 1999). The final CAP3 assembly contains 34448 sequences (16272 Contigs and 18176 Singletons) with a total length of 97Mb, of which 15596 sequences have a minimum length of 500 bp. The N50 contig length based on these 15596 sequences is 12,462 bp. The assemblies were searched for occurrences of terpene synthases using homology-based methods and *ab initio* predictors. A TBLASTN search was performed with an E-value cutoff of $1e^{-30}$ using the 716 *MTPSL* genes identified from OneKP transcriptomes. We also run SNAP (Korf, 2004) trained for *Arabidopsis thaliana* on the assembly. The resulting protein sequences of predicted genes were subsequently subjected to a HMMER search against four HMM profiles (PF03936, PF01397, PF06330 and SmMTPSLs generated by using 48 microbial type tpss from spikemoss).

3.5.3 Phylogenetic Analyses of Terpene Synthases

Bacterial and fungal terpene synthases were obtained from Pfam (Version 27.0). To reduce ambiguities in sequence alignment, only the terpene synthase c terminal domains were included. Sequences were aligned using MAFFT (Linsi) (Katoh et al., 2002) with

1000 iterations of improvement. ProtTest (Darriba et al., 2011) was used to selection of the most appropriate protein evolution model for the protein alignment under the Akaike Information Criterion. For the maximum likelihood analyses, we used RAxML (Stamatakis, 2014) with 1000 bootstrap replicates under the best substitution model (LG+G+F) selected by ProtTest.

3.5.4 Plant Material, Genomic DNA Isolation and PCR

Anthoceros punctatus was cultured in Knop medium containing 0.025 g/L K₂HPO₄, 0.025 g/L KH₂PO₄, 0.025 g/L KCl, 0.025 g/L MgSO₄•7H₂O, 0.1g/L Ca(NO₃)₂•4H₂O, 0.037 g/L FeSO₄•7H₂O, 7 g agar and pH was adjusted to 6.5 before autoclave. *Sphagnum fallax* was cultured in BCD medium (Ashton and Cove, 1977). *A. punctatus* and *S. fallax* were grown at 22 °C under 16 h light/8 h dark photoperiods.

Genome DNA from *A. punctatus* and *S. fallax* was isolated using the VIOGENE plant genomic isolation kit protocol. A fragment of genomic DNA covering partial *MTPSL* genes and its neighbor gene were amplified and subcloned onto the Teasy vector. The amplification was carried out using Phusion high-fidelity DNA polymerase with the following protocol: 2 min at 98 °C, followed by 35 cycles of 10s at 98 °C, 15s at T_m temperature, 45s at 72 °C and a final elongation step of 7 min at 72 °C. The cloned DNA fragment was fully sequenced.

3.6 Acknowledgements

This project was supported by a University of Tennessee AgResearch Innovation Grant (to F.C.). The 1000 Plants (1KP) initiative, led by GKSW, is funded by the Alberta Ministry of Innovation and Advanced Education, Alberta Innovates Technology Futures (AITF), Innovates Centres of Research Excellence (iCORE), Musea Ventures, BGI-Shenzhen and China National Genebank (CNGB). We also acknowledge the contributions of Dr. James Leebens-Mack's lab in obtaining material for RNA Seq and developing the infrastructure for distribution of the 1kp data.

3.7 Bibliography

- Asakawa, Y. (2008). Liverworts-potential source of medicinal compounds. *Current Pharmaceutical Design*, 14(29):3067–88. [63](#)
- Ashton, N. W. and Cove, D. J. (1977). The isolation and preliminary characterisation of auxotrophic and analogue resistant mutants of the moss, *physcomitrella patens*. *Molecular and General Genetics MGG*, 154(1):87–95. [79](#)
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–77. [78](#)
- Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant Journal*, 66(1):212–29. [63](#), [64](#), [74](#)
- Connolly, J. D. and Hill, R. A. (1991). *Dictionary of Terpenoids*. Chapman and Hall/CRC. [63](#)
- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–5. [79](#)

- Degenhardt, J., Kollner, T. G., and Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, 70(15-16):1621–37. [64](#)
- Faeth, S. H. and Fagan, W. F. (2002). Fungal endophytes: common host plant symbionts but uncommon mutualists. *Integrative and Comparative Biology*, 42(2):360–8. [68](#)
- Finn, R. D., Clements, J., and Eddy, S. R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–37. [77](#)
- Gaiero, J. R., McCall, C. A., Thompson, K. A., Day, N. J., Best, A. S., and Dunfield, K. E. (2013). Inside the root microbiome: bacterial root endophytes and plant growth promotion. *American Journal of Botany*, 100(9):1738–50. [68](#)
- Gershenzon, J. and Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nature Chemical Biology*, 3(7):408–14. [63](#)
- Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., Moriyama, T., Ikeuchi, M., Watanabe, M., Wada, H., Kobayashi, K., Saito, M., Masuda, T., Sasaki-Sekimoto, Y., Mashiguchi, K., Awai, K., Shimojima, M., Masuda, S., Iwai, M., Nobusawa, T., Narise, T., Kondo, S., Saito, H., Sato, R., Murakawa, M., Ihara, Y., Oshima-Yamada, Y., Ohtaka, K., Satoh, M., Sonobe, K., Ishii, M., Ohtani, R., Kanamori-Sato, M., Honoki, R., Miyazaki, D., Mochizuki, H., Umetsu, J., Higashi, K., Shibata, D., Kamiya, Y., Sato, N., Nakamura, Y., Tabata, S., Ida, S., Kurokawa, K., and

- Ohta, H. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, 5:3978. [72](#)
- Huang, X. and Madan, A. (1999). Cap3: A dna sequence assembly program. *Genome Research*, 9(9):868–77. [78](#)
- Hyatt, D. C., Youn, B., Zhao, Y., Santhamma, B., Coates, R. M., Croteau, R. B., and Kang, C. (2007). Structure of limonene synthase, a simple model for terpenoid cyclase catalysis. *Proceedings of the National academy of Sciences of the United States of America*, 104(13):5360–5. [75](#)
- Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–66. [78](#)
- Köksal, M., Jin, Y., Coates, R. M., Croteau, R., and Christianson, D. W. (2011). Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature*, 469(7328):116–20. [75](#)
- Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5:59. [78](#)
- Köksal, M., Jin, Y., Coates, R. M., Croteau, R., and Christianson, D. W. (2011). Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature*, 469(7328):116–122. [63](#), [64](#)
- Li, F. W., Villarreal, J. C., Kelly, S., Rothfels, C. J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E. M., Der, J. P., Pittermann, J., Burge, D. O., Pokorny, L., Larsson, A., Chen, T.,

- Weststrand, S., Thomas, P., Carpenter, E., Zhang, Y., Tian, Z., Chen, L., Yan, Z., Zhu, Y., Sun, X., Wang, J., Stevenson, D. W., Crandall-Stotler, B. J., Shaw, A. J., Deyholos, M. K., Soltis, D. E., Graham, S. W., Windham, M. D., Langdale, J. A., Wong, G. K., Mathews, S., and Pryer, K. M. (2014). Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proceedings of the National academy of Sciences of the United States of America*, 111(18):6672–7. [70](#)
- Li, G., Kollner, T. G., Yin, Y., Jiang, Y., Chen, H., Xu, Y., Gershenzon, J., Pichersky, E., and Chen, F. (2012). Nonseed plant *selaginella moellendorffi* [corrected] has both seed plant and microbial types of terpene synthases. *Proceedings of the National academy of Sciences of the United States of America*, 109(36):14711–5. [64](#), [65](#), [70](#), [73](#), [74](#)
- Mithofer, A. and Boland, W. (2012). Plant defense against herbivores: chemical aspects. *Annual Review of Plant Biology*, 63:431–50. [63](#)
- Nagashima, F., Suzuki, M., Takaoka, S., and Asakawa, Y. (2001). Sesqui- and diterpenoids from the japanese liverwort *jungermannia infusca*. *Journal of Natural Products*, 64(10):1309–17. [74](#)
- Opelt, K., Berg, C., Schonmann, S., Eberl, L., and Berg, G. (2007). High specificity but contrasting biodiversity of sphagnum-associated bacterial and plant communities in bog ecosystems independent of the geographical region. *ISME Journal*, 1(6):502–16. [69](#)
- Qiu, Y. L., Li, L., Wang, B., Chen, Z., Knoop, V., Groth-Malonek, M., Dombrowska, O., Lee, J., Kent, L., Rest, J., Estabrook, G. F., Hendry, T. A., Taylor, D. W., Testa, C. M.,

- Ambros, M., Crandall-Stotler, B., Duff, R. J., Stech, M., Frey, W., Quandt, D., and Davis, C. C. (2006). The deepest divergences in land plants inferred from phylogenomic evidence. *Proceedings of the National academy of Sciences of the United States of America*, 103(42):15511–6. [96](#)
- Quin, M. B., Flynn, C. M., and Schmidt-Dannert, C. (2014). Traversing the fungal terpenome. *Natural Product Reports*, 31(10):1449–73. [73](#)
- Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends in Genetics*, 16(6):276–7. [77](#)
- Rosenblueth, M. and Martinez-Romero, E. (2006). Bacterial endophytes and their interactions with hosts. *Molecular Plant-Microbe Interactions*, 19(8):827–37. [68](#)
- Ryan, R. P., Germaine, K., Franks, A., Ryan, D. J., and Dowling, D. N. (2008). Bacterial endophytes: recent developments and applications. *FEMS Microbiology Letters*, 278(1):1–9. [68](#)
- Saritas, Y., Sonwa, M. M., Iznaguen, H., Konig, W. A., Muhle, H., and Mues, R. (2001). Volatile constituents in mosses (musci). *Phytochemistry*, 57(3):443–57. [74](#)
- Schmidt-Dannert, C. (2015). Biosynthesis of terpenoid natural products in fungi. *Advances in Biochemical Engineering/Biotechnology*, 148:19–61. [73](#)
- Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–3. [79](#)

- Starks, C. M., Back, K., Chappell, J., and Noel, J. P. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-*epi*- aristolochene synthase. *Science*, 277(5333):1815–1820. [75](#)
- Tholl, D. (2006). Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Current Opinion in Plant Biology*, 9(3):297–304. [63](#)
- Weiss, M., Selosse, M. A., Rexer, K. H., Urban, A., and Oberwinkler, F. (2004). Sebaciniales: a hitherto overlooked cosm of heterobasidiomycetes with a broad mycorrhizal potential. *Mycological Research*, 108(Pt 9):1003–10. [68](#)
- Weiss, M., Sykorova, Z., Garnica, S., Riess, K., Martos, F., Krause, C., Oberwinkler, F., Bauer, R., and Redecker, D. (2011). Sebaciniales everywhere: previously overlooked ubiquitous fungal endophytes. *PLOS ONE*, 6(2):e16793. [68](#)
- Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., Rothfels, C. J., Pokorny, L., Shaw, A. J., DeGironimo, L., Stevenson, D. W., Surek, B., Villarreal, J. C., Roure, B., Philippe, H., dePamphilis, C. W., Chen, T., Deyholos, M. K., Baucom, R. S., Kutchan, T. M., Augustin, M. M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G. K., and Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National academy of Sciences of the United States of America*, 111(45):E4859–68. [96](#)

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T. W., Li, Y., Xu, X., Wong, G. K., and Wang, J. (2014). Soapdenovo-trans: de novo transcriptome assembly with short rna-seq reads. *Bioinformatics*, 30(12):1660–6.

76

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. *Proceedings of the National academy of Sciences of the United States of America*, 112(3):857–62. 69, 73

Table 3.1: Continued.

[illegible]

Table 3.2: Summary statistics of MTPSLs in 9 plant lineages.

Lineage	Species Count	MTPSL Count	Mean	Median	St. Dev.	Min	Max
Angiosperms	699	5	0.01	0	0.16	0	4
Gymnosperms	80	0	0	0	0	0	0
Monilophytes	70	353	5.04	4.50	3.81	0	20
Lycophytes	22	83	3.77	3	2.65	0	10
Hornworts	7	14	2	0	3.32	0	9
Mosses	41	79	1.93	1	1.82	0	7
Liverworts	26	177	6.81	8	4.15	0	16
Charophyta	47	1	0.02	0	0.15	0	1
Chlorophyta	111	0	0	0	0	0	0

Table 3.3: 22 *MTPSL* genes outside of the four groups (high similarity to microbial TPSs).

Group	MTPSL_id	NR_top_hit	evaluate	bit_score	Scientific_name	Kingdom
V	ANT_WCZB_MFTPS7	gi 751680917 gb KIM31075.1	0	1019	Serendipita vermifera MAFF 305830	Fungi
V	ANT_WCZB_MFTPS8	gi 353240956 emb CCA72799.1	0	981	Piriformospora indica DSM 11827	Fungi
V	ANT_WCZB_MFTPS9	gi 353240956 emb CCA72799.1	7.00E-130	391	Piriformospora indica DSM 11827	Fungi
V	CHO_MCHJ_MFTPS1	gi 913451420 ref WP_050430829.1	0.001	50.8	Chondromyces crocatus	Bacteria
V	EUD_MRKX_MFTPS1	gi 588255974 ref XP_006957163.1	0	558	Wallemia mellicola CBS 633.66	Fungi
V	EUD_QAIR_MFTPS1	gi 751178290 gb KIL64254.1	5.00E-177	508	Amanita muscaria Koide BX008	Fungi
V	EUD_QAIR_MFTPS2	gi 751175026 gb KIL61028.1	0	620	Amanita muscaria Koide BX008	Fungi
V	EUD_QAIR_MFTPS3	gi 751174784 gb KIL60790.1	7.00E-143	427	Amanita muscaria Koide BX008	Fungi
V	EUD_QAIR_MFTPS5	gi 751181225 gb KIL67171.1	0	691	Amanita muscaria Koide BX008	Fungi
V	LYC_JKAA_MFTPS1	gi 927407765 ref XP_013949969.1	0	541	Trichoderma virens Gv29-8	Fungi
V	LYC_ZYCD_MFTPS1	gi 238496645 ref XP_002379558.1	1.00E-130	390	Aspergillus flavus NRRL3357	Fungi
V	MAR_AEXY_MFTPS1	gi 389636521 ref XP_003715910.1	6.00E-178	513	Magnaporthe oryzae 70-15	Fungi
V	MAR_IRBN_MFTPS6	gi 751680917 gb KIM31075.1	0	934	Serendipita vermifera MAFF 305830	Fungi
V	MAR_JHFI_MFTPS20	gi 751680917 gb KIM31075.1	6.00E-101	313	Serendipita vermifera MAFF 305830	Fungi
V	MAR_LGOW_MFTPS3	gi 629725325 ref XP_007822988.1	7.00E-56	196	Metarhizium robertsii	Fungi
V	MAR_NWQC_MFTPS8	gi 629725325 ref XP_007822988.1	6.00E-36	144	Metarhizium robertsii	Fungi
V	MAR_OFTV_MFTPS5	gi 751680917 gb KIM31075.1	0	931	Serendipita vermifera MAFF 305830	Fungi
V	MAR_RTMU_MFTPS4	gi 667838359 ref XP_007783348.1	3.00E-128	386	Coniosporium apollinis CBS 100218	Fungi
V	MAR_WJLO_MFTPS3	gi 549052256 emb CCX30236.1	4.00E-68	231	Pyronema omphalodes CBS 100304	Fungi
V	MAR_WJLO_MFTPS4	gi 549052256 emb CCX30236.1	1.00E-71	240	Pyronema omphalodes CBS 100304	Fungi
V	MAR_YBQN_MFTPS7	gi 648165817 gb KDR79494.1	0	596	Galerina marginata CBS 339.88	Fungi
V	MON_QIAD_MFTPS2	gi 629662947 ref XP_007805277.1	2.00E-77	251	Endocarpon pusillum Z07020	Fungi

Table 3.4: *MTPSL* genes from the hornwort *Anthoceros punctatus*.

Gene	Length (bp)	Location	Introns	Protein size (aa)	Group
ApMTPSL1	1227	NODE_4150_length_6347_cov_5.87285_ID_8299:1863..3089	0	408	II
ApMTPSL2	1293	Contig14756:5881..7173	0	430	III
ApMTPSL3	1311	Contig1451:3817..5127	0	436	II
ApMTPSL4	1206	Contig5434:17076..18281	0	401	III
ApMTPSL5	1242	Contig969:424..1665	0	413	III
ApMTPSL6	1266	NODE_4646_length_5616_cov_12.4068_ID_9291:3111..4376	0	421	III
ApMTPSL7	1284	NODE_2278_length_10879_cov_6.65983_ID_4555:7120..8403	0	427	III
ApMTPSL8	927	Contig1090:645..1571	0	308	III
ApMTPSL9	1026	Contig4580:3013..4038	0	341	III
ApMTPSL10	1149	Contig8980:4454..5685	1	382	II
ApMTPSL11	822	NODE_112_length_40181_cov_9.86101_ID_223:37545..38965	3	273	III
ApMTPSL12	873	NODE_2080_length_11745_cov_6.2947_ID_4159:3313..4244	1	290	III
ApMTPSL13*		Contig13088:2558..5572			III
ApMTPSL14	1026	Contig969:7202..8227	0	341	III
ApMTPSL15	297	Contig8303:306..602	0	98	
ApMTPSL16	330	NODE_9289_length_2299_cov_6.55749_ID_18577:52..381	0	109	

Table 3.5: Typical plant terpene synthase genes in *Sphagnum fallax*.

Gene	Length	Group	NR_top_hit	evalue	bit_score	Scientific_name	Kingdom
Sphfalx0057s0049.1	490	I	gi 288346547 gb EFC80872.1	2.00E-26	122	Frankia sp. EUN1f	Bacteria
Sphfalx0363s0003.1	478	I	gi 664165251 ref WP_030699689.1	2.00E-27	122	Streptomyces griseus	Bacteria
Sphfalx0008s0258.1	482	I	gi 827012632 ref WP_047174997.1	3.00E-25	119	Streptomyces sp. MNU77	Bacteria
Sphfalx0141s0067.1	478	I	gi 664165251 ref WP_030699689.1	7.00E-26	117	Streptomyces griseus	Bacteria
Sphfalx0021s0008.1	482	I	gi 491462970 ref WP_005320742.1	2.00E-25	116	Streptomyces pristinaespiralis	Bacteria
Sphfalx0097s0020.1	473	I	gi 664054675 ref WP_030593937.1	9.00E-24	115	Streptomyces anulatus	Bacteria
Sphfalx0455s0001.1	228	I	gi 664054675 ref WP_030593937.1	5.00E-24	110	Streptomyces anulatus	Bacteria
Sphfalx0420s0001.1	488	I	gi 972355343 ref WP_059010379.1	3.00E-22	110	Streptomyces specialis	Bacteria
Sphfalx0043s0071.1	483	I	gi 491462970 ref WP_005320742.1	1.00E-22	108	Streptomyces pristinaespiralis	Bacteria
Sphfalx0199s0005.1	456	I	gi 740130670 ref WP_037978644.1	5.00E-21	106	Streptomyces sp. TAA486	Bacteria
Sphfalx0045s0015.1	458	I	gi 740130670 ref WP_037978644.1	2.00E-19	101	Streptomyces sp. TAA486	Bacteria
Sphfalx0442s0005.1	456	I	gi 740130670 ref WP_037978644.1	2.00E-18	98.6	Streptomyces sp. TAA486	Bacteria
Sphfalx0197s0010.1	456	I	gi 926397772 ref WP_053725884.1	2.00E-18	98.6	Streptomyces sp. WM6378	Bacteria
Sphfalx0197s0012.1	454	I	gi 927089120 ref WP_053788880.1	9.00E-19	97.1	Streptomyces sp. XY332	Bacteria
Sphfalx0012s0019.1	457	I	gi 902778075 ref WP_049649135.1	9.00E-19	96.7	Kitasatospora sp. MY 5-36	Bacteria
Sphfalx0280s0009.1	379	I	gi 664165251 ref WP_030699689.1	1.00E-13	81.6	Streptomyces griseus	Bacteria
Sphfalx0280s0009.2	378	I	gi 664165251 ref WP_030699689.1	4.00E-13	79.7	Streptomyces griseus	Bacteria
Sphfalx0308s0010.1	164	I	gi 644654594 ref WP_025348809.1	3.00E-12	73.2	Nocardia nova	Bacteria
Sphfalx0128s0030.1	87	I	gi 644654594 ref WP_025348809.1	1.00E-09	63.2	Nocardia nova	Bacteria
Sphfalx0043s0070.1	124	I	gi 702805978 ref WP_033270350.1	0.003	46.2	Streptomyces lydicus	Bacteria
Sphfalx0128s0029.1	107	I	gi 652906113 ref WP_027160120.1	0.21	40	Methylobacter luteus	Bacteria

Table 3.6: A list of sequenced plants searched for *MTPSL* genes.

Species	Data version
<i>Amaranthus hypochondriacus</i>	v1.0
<i>Amborella trichopoda</i>	v1.0
<i>Ananas comosus</i>	v3
<i>Aquilegia coerulea</i>	v1.1
<i>Aquilegia coerulea</i>	v3.1
<i>Arabidopsis halleri</i>	v1.1
<i>Arabidopsis lyrata</i>	v1.0
<i>Arabidopsis thaliana</i>	TAIR10
<i>Boechera stricta</i>	v1.2
<i>Brachypodium distachyon</i>	v3.1
<i>Brachypodium stacei</i>	v1.1
<i>Brassica rapa</i>	FPsc v1.3
<i>Capsella grandiflora</i>	v1.1
<i>Capsella rubella</i>	v1.0
<i>Carica papaya</i>	ASGPBv0.4
<i>Chlamydomonas reinhardtii</i>	v5.5
<i>Citrus clementina</i>	v1.0
<i>Citrus sinensis</i>	v1.1
<i>Coccomyxa subellipsoidea</i> C-169	v2.0
<i>Cucumis sativus</i>	v1.0
<i>Eucalyptus grandis</i>	v2.0
<i>Eutrema salsugineum</i>	v1.0
<i>Fragaria vesca</i>	v1.1
<i>Glycine max</i>	Wm82.a2.v1
<i>Gossypium raimondii</i>	v2.1
<i>Kalanchoe marnieriana</i>	v1.1
<i>Klebsormidium flaccidum</i>	v1.0
<i>Linum usitatissimum</i>	v1.0
<i>Malus domestica</i>	v1.0
<i>Manihot esculenta</i>	v6.1
<i>Medicago truncatula</i>	Mt4.0v1
<i>Micromonas pusilla</i> CCMP1545	v3.0
<i>Micromonas</i> sp. RCC299	v3.0
<i>Mimulus guttatus</i>	v2.0
<i>Musa acuminata</i>	v1
<i>Oryza sativa</i>	v7_JGI
<i>Ostreococcus lucimarinus</i>	v2.0
<i>Panicum hallii</i>	v2.0
<i>Panicum virgatum</i>	v1.1

Table 3.6: Continued

Species	Data version
<i>Phaseolus vulgaris</i>	v1.0
<i>Physcomitrella patens</i>	v3.3
<i>Populus trichocarpa</i>	v3.0
<i>Prunus persica</i>	v2.1
<i>Ricinus communis</i>	v0.1
<i>Salix purpurea</i>	v1.0
<i>Selaginella moellendorffii</i>	v1.0
<i>Setaria italica</i>	v2.2
<i>Setaria viridis</i>	v1.1
<i>Solanum lycopersicum</i>	iTAG2.3
<i>Solanum tuberosum</i>	v3.4
<i>Sorghum bicolor</i>	v3.1
<i>Sphagnum fallax</i>	v0.5
<i>Spirodela polyrhiza</i>	v2
<i>Theobroma cacao</i>	v1.1
<i>Triticum aestivum</i>	v2.2
<i>Vitis vinifera</i>	Genoscope.12X
<i>Volvox carteri</i>	v2.1
<i>Zea mays</i>	6a
<i>Zostera marina</i>	v2.2

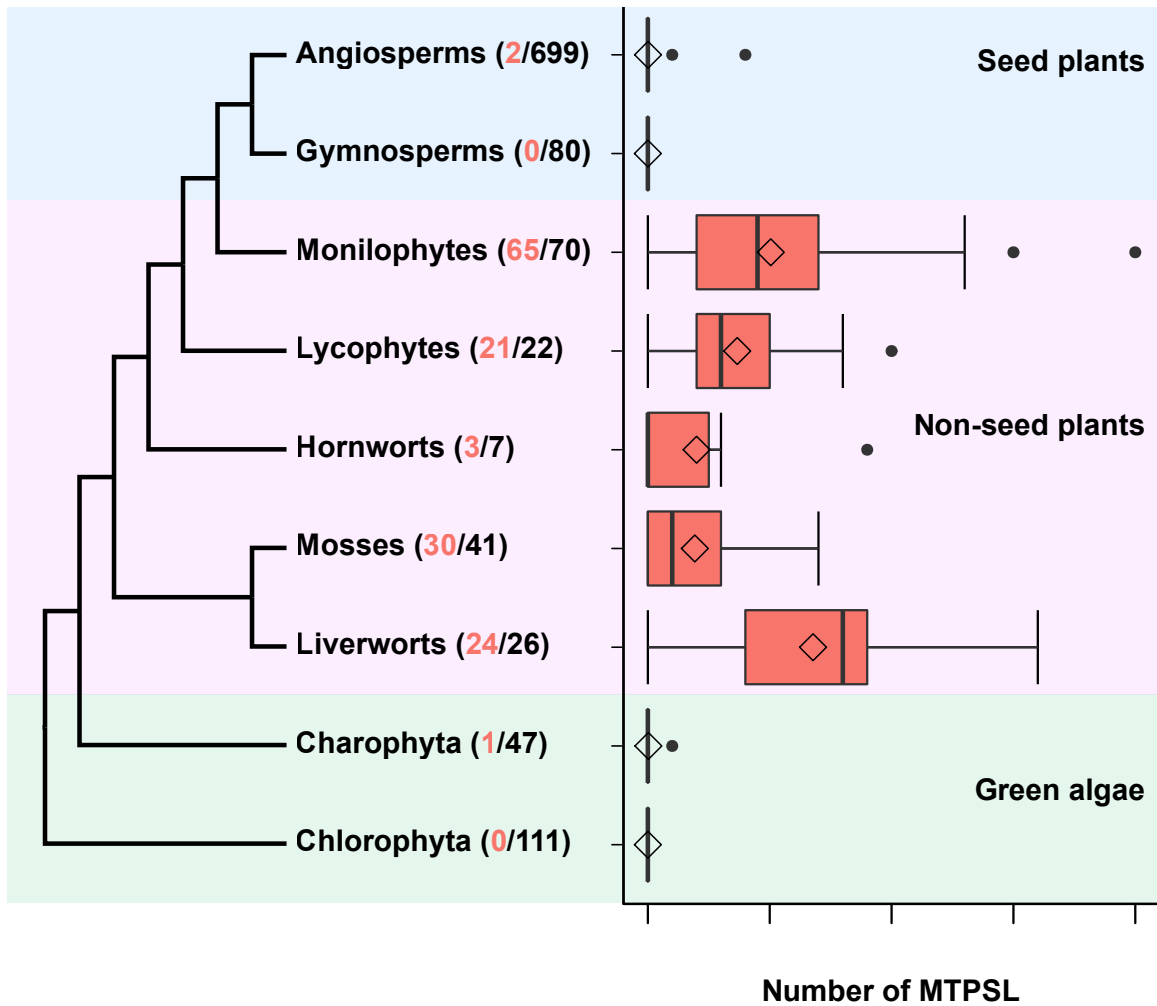


Figure 3.1: Distribution of terpene synthase genes of microbial type (MTPSL) identified from the transcriptomes of 1103 plant species. The phylogeny of green plants was adapted from (Qiu et al., 2006) and (Wickett et al., 2014). The numbers in parentheses represent the number of transcriptomes containing putative MTPSL (in red) and total transcriptomes analyzed in each lineage (in black). Each boxplot represents the number of MTPSL found for individual species in each plant lineage. The solid black lines denote the median number of MTPSL from each species. Whiskers represent 1.5 times the quantile of the data. Points outside of the range of the whiskers are outliers.

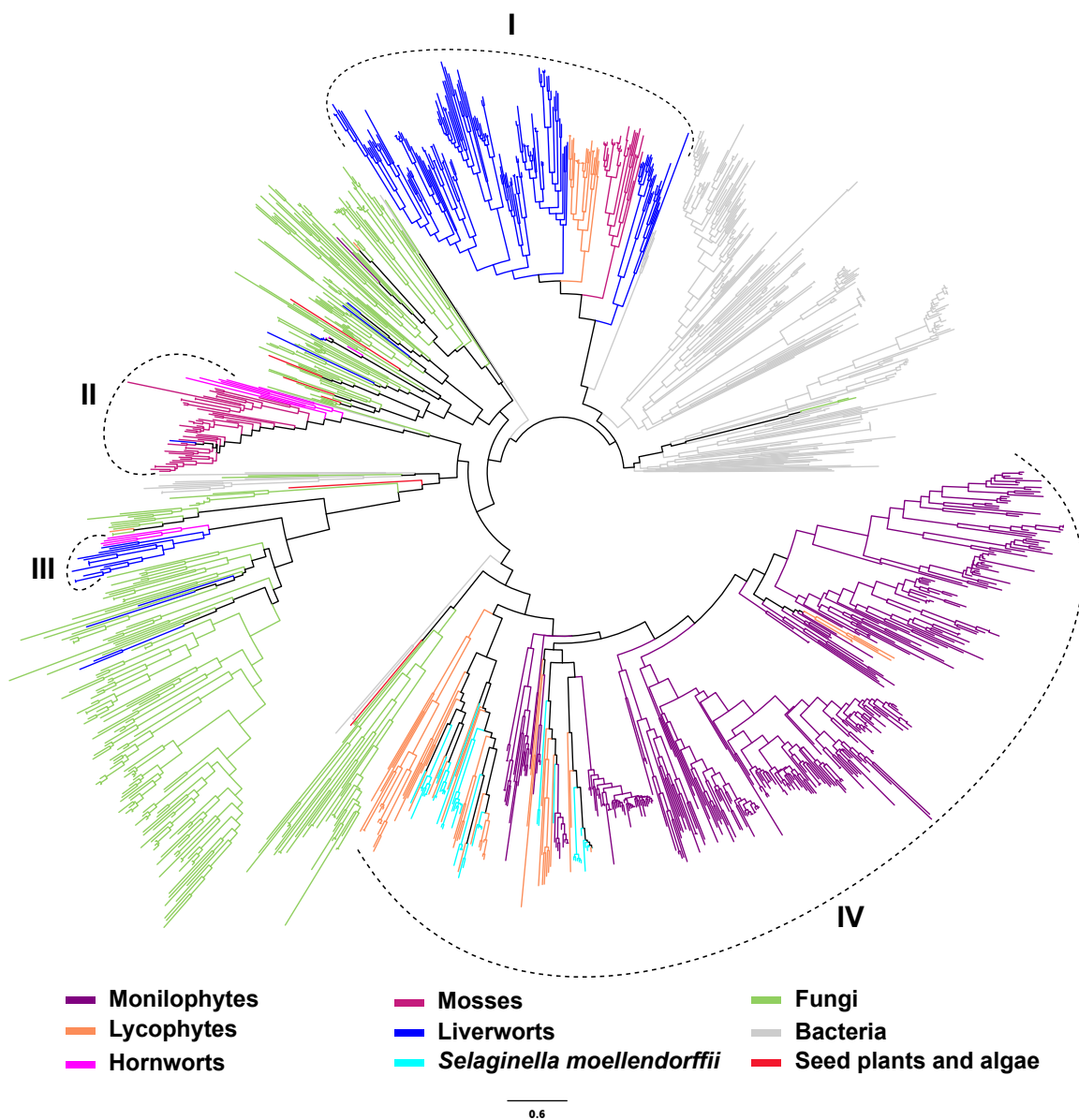


Figure 3.2: Phylogeny of terpene synthases of microbial type (MTPSL) identified from OneKP with known *MTPSL* genes from *S. moellendorffii*, bacterial terpene synthases and fungal terpene synthases. Terpene synthases are color-coded based on their source. The majority of *MTPSL* genes are clustered into four major groups (I to IV).

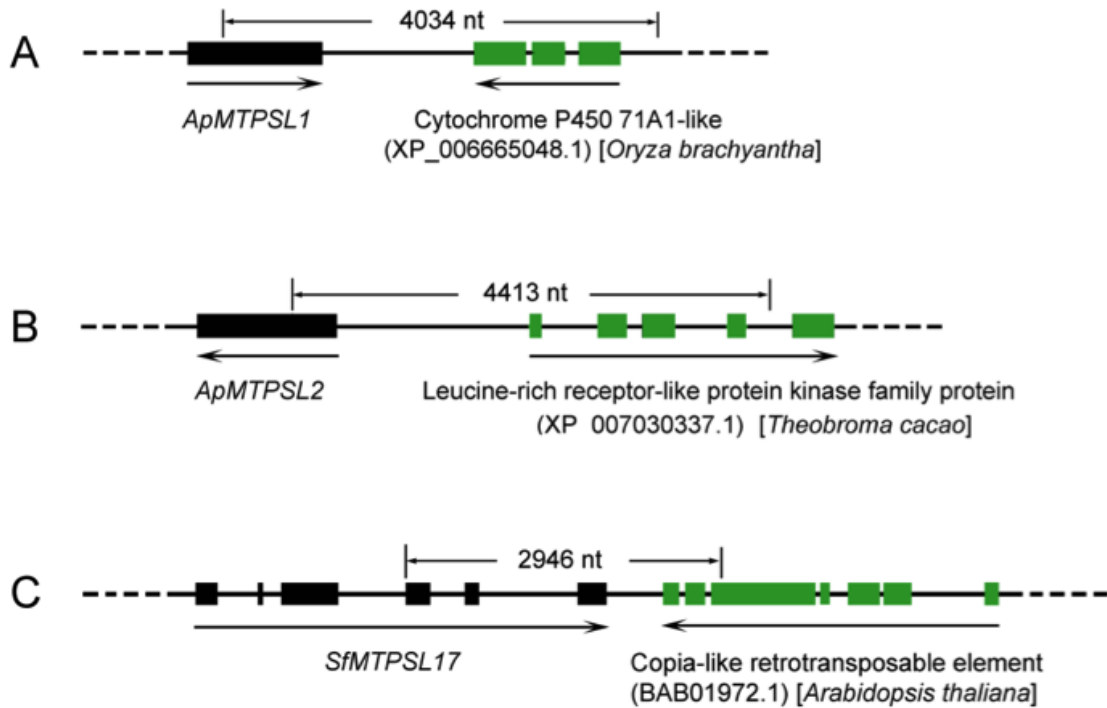


Figure 3.3: Validation of representative *MTPSL* genes to be plant genes. The figures show the schematic genomic organization of representative *MTPSL* genes with their neighboring genes. The genomic region spanning each *MTPSL* gene and its neighboring gene and the intergenic region was amplified using PCR and confirmed by sequencing. A. *ApMTPSL1* from the hornwort *Anthoceros punctatus*. B. *ApMTPSL2* from the hornwort *A. punctatus*. C. *SfMTPSL17* from the moss *Sphagnum fallax*. In each schematic figure, the neighboring gene was annotated as a functional protein. The species that contains a protein with the highest homology to the respective neighbor gene was indicated. The group number indicates that the specific groups as listed in Figure 3.2 that the respective *MTPSL* represents.

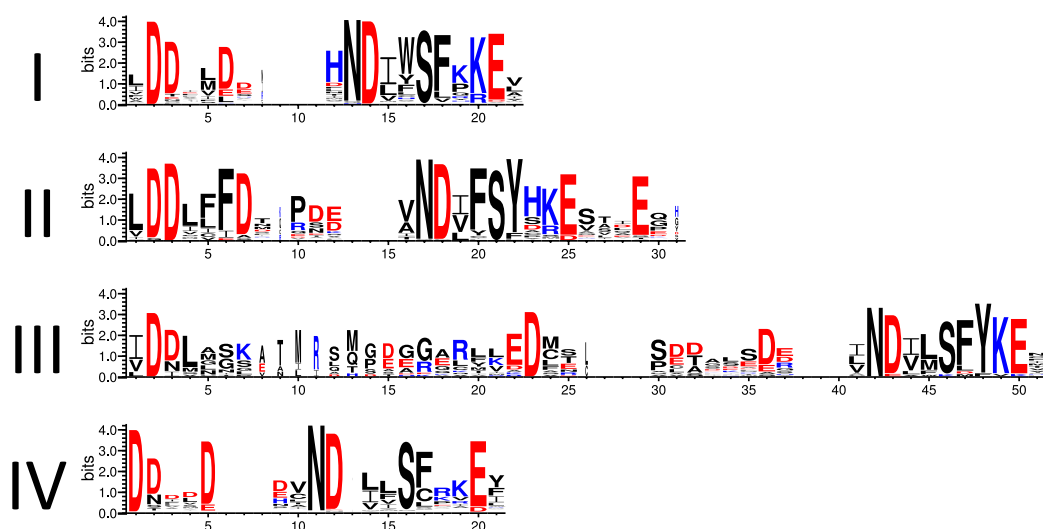


Figure 3.4: Motif analysis of MTPSLs. 'DDxxD' and 'NSE' are two highly conserved motifs found in terpene synthases. Sequence motif logos made using weblogo 3.0, showing the conserved motifs found in each group of terpene synthase genes of microbial type.

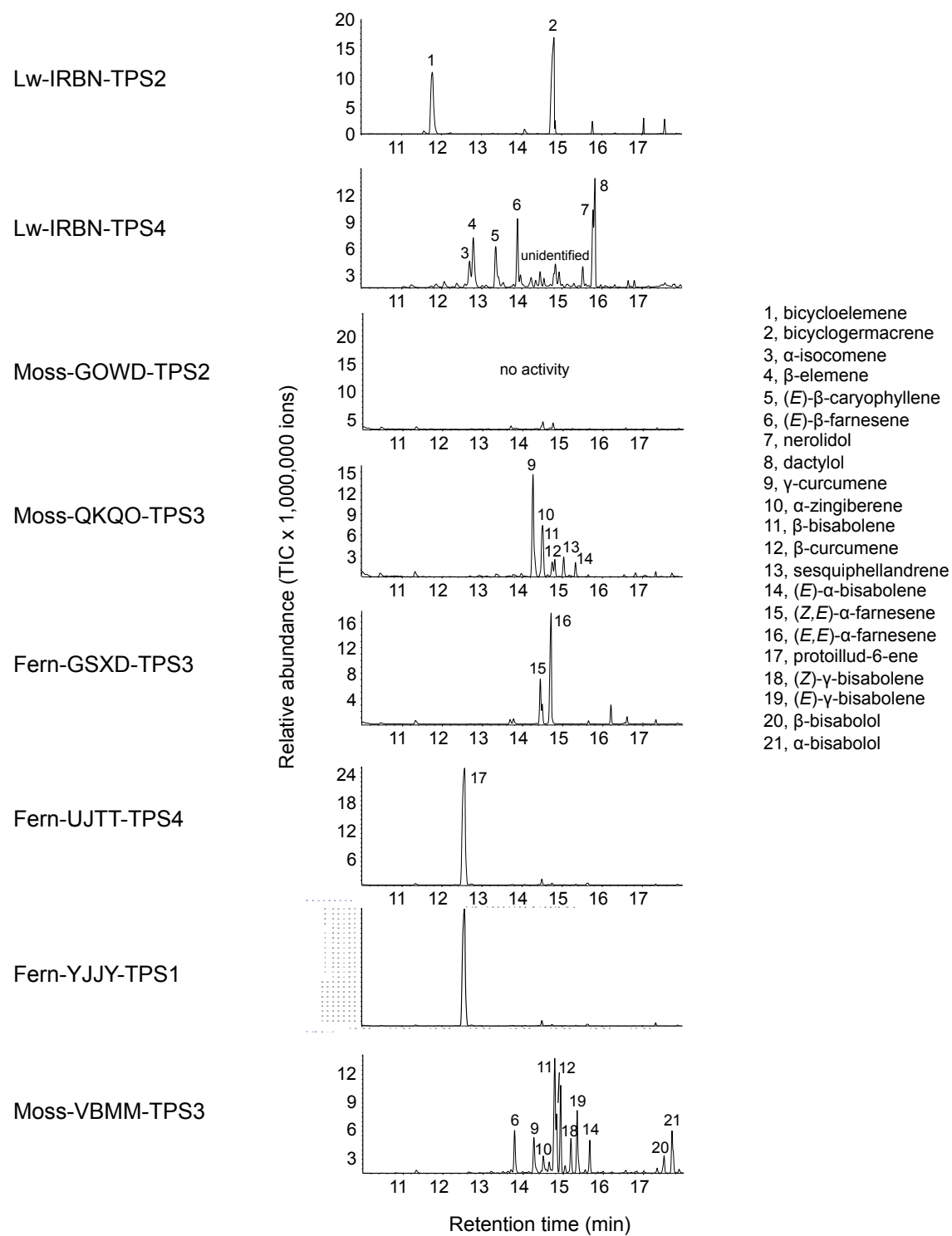


Figure 3.5: Biochemical activity of selected *MTPSL* genes.

```

      *      20      *      40      *      60
IRBN_MTPSL8 : MA--SPATIRLPDILSAMDRFELRTHPDEREVTRASNEWFNSYNMMPAPIFEKFVKCDFGLMTAM : 63
gi|751680917 : MASPSPATIRLPDILSAMDKFELRTHPDEREVTRASNEWFNSYNMMPAPIFEKFVKCDFGLMTAM : 65
      MA SPATIRLPDILSAM4FELRTHPDEREVTRASNEWFNSYNMMPAPIFEKFVKC FGLMTAM

      *      80      *      100      *      120      *
IRBN_MTPSL8 : SYPDTDATRLRITADYMSILFAYDDLMDLPSSDLMDHRIASSKAAKIMMQVLTHPHKFKVPVGLP : 128
gi|751680917 : SYPDTDATRLRITADYMSILFAYDDLMDLPSSDLMDHRIASSKAAKIMMQVLTHPHKFKVPVGLP : 130
      SYPDTDATRLRITADYMSILFAYDDLMDLPSSDLMDHRIASSKAAKIMMQVLTHPHKFKVPVGLP

      140      *      160      *      180      *
IRBN_MTPSL8 : VATAFHDFWTRFCATSTPSMQKRFTETTYEYVMAVKNQVGNRASSVCPSIEEYVSLRRDTSIAIKV : 193
gi|751680917 : VATAFHDFWTRFCATSTPSMQKRFTETTYEYVMAVKNQVGNROSSVCPSIEEYVSLRRDTSIAIKV : 195
      VATAFHDFWTRFCATST SMQKRFTETTYEYVMAVKNQVGNR SSVCPSEIEEYVSLRRDTSIAIKV

      200      *      220      *      240      *      260
IRBN_MTPSL8 : TYACIEYCLNIDCPDEAFYHPSLAALQEAGNDILSWANDVYSFDNEQCSDCHNLIAVVAINKNI : 258
gi|751680917 : TYACIEYCLNIDVPDEAFYHPSLAALQEAGNDILSWANDVYSFDNEQCSDCHNLIAVVAINKNI : 260
      TYACIEYCLNID PDEAFYHPSLAALQEAGNDILSWANDVYSFDNEQCSDCHNLIAVVAINKNI

      *      280      *      300      *      320
IRBN_MTPSL8 : TVQAAMEYAMGMIDSAINRFFEECSNVPSFGPDVDPKVQAYIKGVELYLSGSVFWHLESERYFGP : 323
gi|751680917 : TVQAAMEYAMGMIDSAINRFFEECANVPSFGPDVDPKVQAYIKGVELYLSGSVFWHLESERYFGP : 325
      TVQAAMEYAMGMIDSAINRFFEECSNVPSFGPDVDPKVQAYIKGVELYLSGSVFWHLESERYFGP

      *      340      *      360      *      380      *
IRBN_MTPSL8 : RVKHVKDTLMVELRPLDEGAKPAFDLIYKLPSNLTSNVLAAVSNRTPTP-PAPVEAAP-AAPSPP : 386
gi|751680917 : RVKHVKDTLMVELRPLDEGAKPAFNLIYKLPSNLTSNVLAAVTPTTKTPPEVPVAAAPTVAAPSPP : 390
      RVKHVKDTLMVELRPLDEGAKPAFDLIYKLPSNLTSNVLAAV3 T TP P PV AAP APSPP

      400      *      420      *      440      *
IRBN_MTPSL8 : PRTRRGTP-----PAHHAPETIHAPVPISPFNPNFPTVSPTSVPPPSYEHQRAFAQYMAAQLDE : 445
gi|751680917 : PRGSSNSSTGTVRASPVQH--EIHAPTPISPFNPNFPTSNPNMPPPSYEQQRVFAQFMAAQLDE : 453
      PR 3 3 T P H EIHAP PISPFNPNFPT P 6PPPSYE QR FAQ5MAAQL

      460      *      480      *      500      *      520
IRBN_MTPSL8 : KMRAEQYYNQAPQYYSAPQSPYQDQQQ---KLQNLSLMEVLLSRPTSELTNILVIASVLMASSP : 506
gi|751680917 : KMRAEQQW-QVPOYYSAPQSPYQPOQQQQLTKARQNLSLMEAILNRPTSELTNILVIASVLMASSP : 517
      KMRAEQ 5 Q PQYYSAPQSPYQ QQQ K QNLSLME 6L RPTSELTNILVIASVLMASSP

      *      540
IRBN_MTPSL8 : LALVPFVPLLVLVLLFPEAPAVLLS : 530
gi|751680917 : LALIPFVPLLVLVLLYPEAPAVLLA : 541
      LAL6PFVPLLVLVLL5PEAPAVLL

```

Figure 3.6: Alignment of one MTPSL of putative contamination with one fungal TPS.

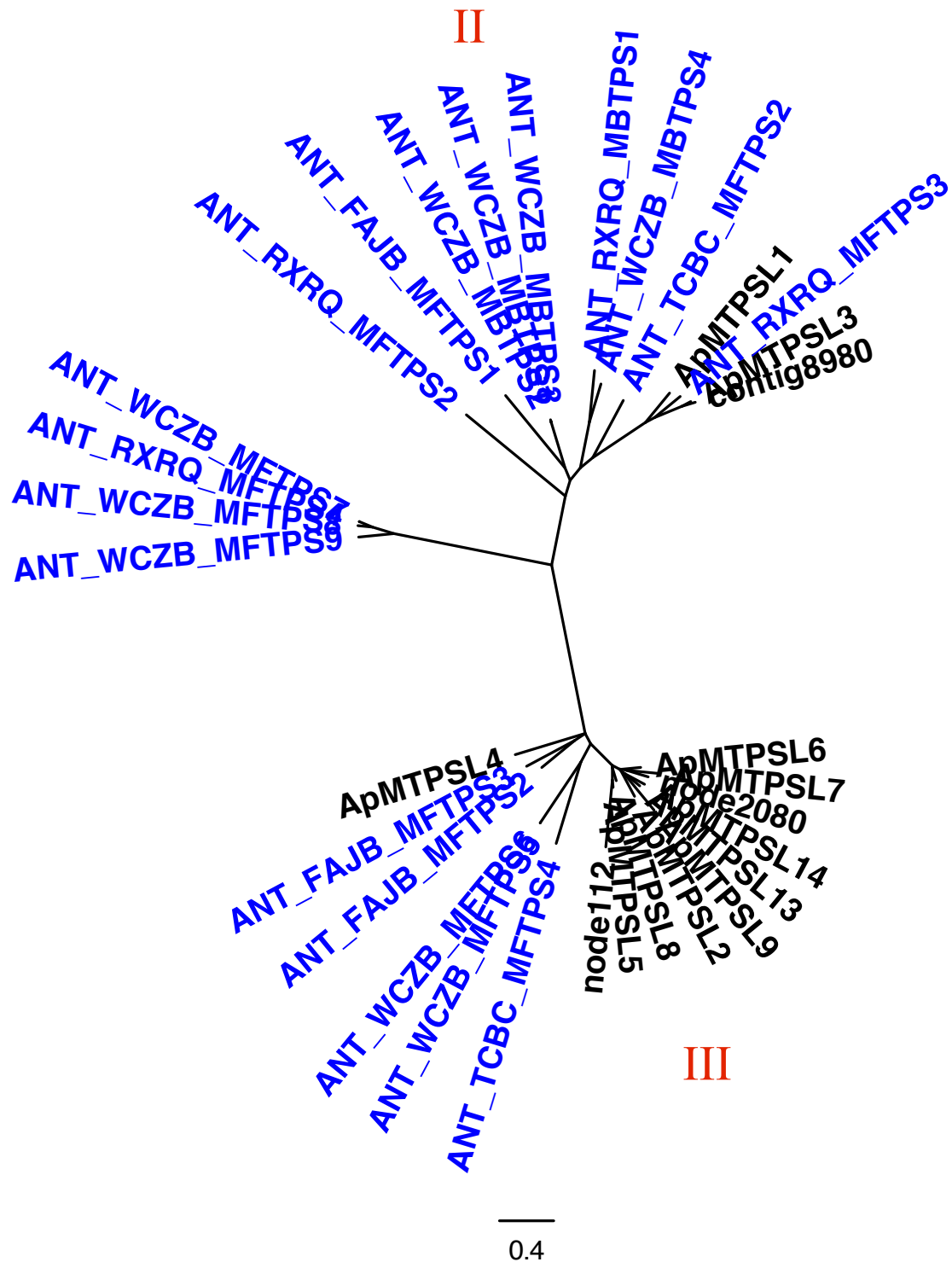


Figure 3.7: Phylogenetic tree of MTPSL genes identified from *Anthoceros punctatus* genome with those identified from hornworts transcriptomes.

Chapter 4

Horizontal Gene Transfer of Terpene

Synthase Genes from Bacteria to Fungi

Author Contributions: Qidong Jia conducted the phylogenetic analyses, identified the HGT events, performed all comparative genomics and bioinformatics analyses, and wrote the manuscript.

4.1 Abstract

Terpenoids constitute the largest class of secondary metabolites. The vast diversity of terpenoids is partly achieved through the continued creation of novel terpene synthase (TPS) genes, which encode key enzymes for terpenoid biosynthesis, through gene duplication followed by function divergence. In contrast, little is known about the contribution of horizontal gene transfer (HGT) of *TPS* genes for the diversity of terpenoids. The goal of this study was to investigate HGT of *TPS* genes from bacteria to fungi. By phylogenetic analysis of TPSs from bacteria and fungi, several fungal TPSs were found to be nested within bacterial TPSs, implying HGT from bacteria to fungi. These TPSs were renamed *BTPSL* (bacterial TPS-like). We then focused our study on a group of entomopathogenic fungi with sequenced genomes. Of the eleven species of fungi analyzed, eight *BTPSL* genes were found from seven species, of which the majority are *Metarhizium* species. In most fungal species containing *BTPSL*, collinearity could be identified for *BTPSL* and neighbor genes. In addition to *BTPSL* genes, each of the fungal species was found to contain typical fungal *TPS* genes, suggesting that terpenoids produced in each fungus are determined by both *BTPSL* and typical fungal *TPS*s. We also performed biochemical studies on one of the identified *BTPSL*s (MAA_08668) and showed that it has sesquiterpene synthase activity. Molecular evolutionary analysis of *BTPSL* genes implied purifying selection, suggesting that the novel chemistry brought about by the acquisition of *BTPSL* genes via HGT may have important and conserved functions for the receipt fungi.

4.2 Introduction

Metabolites can be divided into two groups: primary metabolites and secondary metabolites. While primary metabolites are generally conserved and essential for cellular functions, secondary metabolites tend to be lineage- or even species-specific and are important for ecological functions. Collectively, secondary metabolites are highly diverse, mounting to more than 200,000 compounds identified so far (Hartmann, 2007). Species-specificity and diversity of secondary metabolites are believed to be the result of natural selection. Therefore, understanding the mechanisms of evolution of secondary metabolite biosynthesis is an important avenue to our understanding of the adaptation of different living organisms to their unique niches. At the molecular level, duplication followed by functional divergence of genes of secondary metabolism is an important mechanism for achieving diversity and specificity (Kliebenstein et al., 2001; Ober, 2005). The impact of this mechanism is manifested by the presence of families of genes of secondary metabolites of varied sizes. Another mechanism for acquiring genetic novelty for secondary metabolism is horizontal gene transfer (HGT), which is defined as transfer of genes between organisms in a manner other than traditional reproduction (Gogarten and Townsend, 2005; Keeling and Palmer, 2008). In fungi, a few HGT events for genes of secondary metabolism transferred from bacteria have been documented. For example, Schmitt reported a clade of fungal type 1 polyketide synthase (PKS) genes, termed 6-MSAS-type PKS, likely originated from Actinobacteria (Schmitt and Lumbsch, 2009). The presence of bacterial type hybrid non-ribosomal peptide synthase/PKS proteins in several

fungal species was also reported to be the result of a HGT event from bacteria to fungi (Lawrence et al., 2011). Since biosynthetic genes for secondary metabolism are often organized into clusters and co-expressed under particular conditions, these metabolic gene clusters can also be transferred between fungi species (Campbell et al., 2012; Khaldi and Wolfe, 2011; Proctor et al., 2013).

Terpenoids constitute the largest class of secondary metabolites. Despite their vast diversity, all terpenoids are synthesized from the same five-carbon precursor isopentenyl diphosphate (IPP). Depending on the number of IPP units, terpenoids can be further categorized as monoterpenes (C_{10}), sesquiterpenes (C_{15}), diterpenes (C_{20}), triterpenes (C_{30}), tetraterpenes (C_{40}) and polyterpenes. The biosynthesis of monoterpenes, sesquiterpenes and diterpenes from their respective substrates geranyl diphosphate, farnesyl diphosphate and geranylgeranyl diphosphate is catalyzed by the same type of enzymes, termed terpene synthases (TPSs). TPSs have been identified in bacteria, fungi and plants (Chen et al., 2011; Keller et al., 2005; Yamada et al., 2015). Despite some structural homology, bacterial and fungal TPSs are only distantly related to typical plant TPSs judged on protein sequence similarity. Structural studies showed that TPSs are modular in nature, consisting one or more of the three modular structures called α -domain, β -domain and γ -domain (Köksal et al., 2011). Typical plant TPSs have either α - β or α - β - γ structure. The majority of TPSs from bacteria and fungi are only formed by the α -domain. Interestingly, a recent study showed that microbial type terpene synthase genes are also present in plants (Li et al., 2012). The sporadic distribution of microbial type TPSs in the three domains of life suggests that the early evolution of microbial type TPSs genes may have involved HGT.

Despite the enormous diversity of terpenoids and the implication of HGT of *TPS* genes to achieve such diversity, to the best of our knowledge, a clear case of HGT of *TPS* genes has not been reported.

In principal, HGT of *TPS* genes can rapidly result in novel chemistry in the recipient organism. There are two biochemical pathways for the production of the terpenoid precursor IPP, the mevalonic acid pathway and the methylerythritol 4-phosphate pathway, and all living organisms contain one or both of these pathways (Vranova et al., 2013). Consequently, the substrates for monoterpenes, sesquiterpenes and diterpenes are being produced by essentially all living organisms (also for the production of primary terpene metabolites). Therefore, *TPS* genes acquired through HGT can be functional in the recipient immediately after being acquired with its substrate readily available. One important feature of many TPSs is that they often produce multiple products from one substrate (Degenhardt et al., 2009). Therefore, acquisition of a single TPS may lead to the production of a suite of novel terpene metabolites. Terpenoids have many ecological functions, ranging from defense to beneficial communications (Chen et al., 2011; Tholl, 2006). If the acquisition of a *TPS* gene and consequently the ability of producing a suite of terpenoids provide fitness benefit for the recipient, such *TPS* genes of HGT would likely be retained and fixed in the recipient. Therefore, HGT of *TPS* genes most likely have made important contribution to the vast diversity of terpenoids observed in plants, bacteria, and fungi. In recent studies it has been shown that many fungi acquired genes from bacteria through HGT (Bushley and Turgeon, 2010; Kroken et al., 2003; Lawrence et al., 2011; Schmitt and Lumbsch, 2009). To test our hypothesis concerning *TPS* genes, the first

objective of this study was to detect events of HGT of *TPS* genes from bacteria to fungi. Once identified, our second objective was to analyze such *TPS* genes in a specific biological context to understand their biochemical function and evolution.

4.3 Results

4.3.1 Analysis of Bacterial and Fungal Terpene Synthase Genes Suggests Possible HGT Events from Bacteria to Fungi

The list of sequenced bacterial and fungal species is growing rapidly, as evidenced by the growing number of sequencing projects listed in Joint Genome Institute's Genomes Online Database (GOLD, <https://gold.jgi.doe.gov>). However, considering the fact that some of the sequenced genomes are not well annotated and the challenge that a huge amount of data we would have to deal with, in this study, we restricted our initial analysis to well-defined *TPS* genes to explore the evolutionary relationships of TPSs in fungi and bacteria. Therefore, we chose the TPSs included in the Pfam database (<http://pfam.xfam.org>, version 27.0).

The 'gold standard' for HGT identification is through phylogenetic analysis by identifying incongruent relationships. Since the low degree of overall sequence identity between bacterial and fungal TPSs, the C-terminal domain of each TPS, which is a more conserved region, was extracted and used for the construction of phylogenetic trees. An unrooted maximum likelihood phylogenetic tree of 341 TPS C-domain protein sequences

is shown in Figure 4.1. According to the tree topology, the majority of bacterial and fungi TPSs clustered into two separate groups. However, some sequences were not clustered within their own kingdom, although most of them were poorly supported by the respective bootstrap values and therefore ignored for further analysis. Among them, there were two clades in which fungi TPSs were embedded within bacterial TPSs with high bootstrap values (100%, Figure 4.1). This nested relationship implies that these genes were transferred horizontally from bacteria to fungi. One clade contained a fungi TPS from *Arthrobotrys oligospora* (Nematode-trapping fungus) and a bacterial TPS from *Sphingobacterium sp.* strain 21. The other one included four TPSs, two fungi TPSs MAA_08668 and MAC_05714 from the two *Metarhizium* species *M. robertsii* and *M. acridum*, respectively, and two bacterial TPSs from *Granulicella mallensis* and *Kitasatospora setae*. The overall sequence similarity among these four domain sequences was more than 50%. The sequence identity of MAA_08668 and MAC_05714 is 77%. In this study, the fungal TPSs nested in bacterial TPSs were designated BTPSL (bacterial TPS-like). We chose to do detailed follow-up study on the BTPSL genes from putative HGT event initially identified in *Metarhizium*.

4.3.2 The Apparent Orthologs of BTPSL were Identified in a Group of Entomopathogenic Fungi

Because the TPS dataset used in our initial analysis was restricted to the pfam database, to determine the distribution of BTPSL genes in the fungi kingdom, we searched the

NCBI's non-redundant protein sequence database (NR) with blastp using MAA_08668 and MAC_05714 as query respectively. Due to the high sequence homology between MAA_08668 and MAC_05714, essentially same results were obtained. There was a clear sequence identity drop from 55% to 29%, so sequences showing at least 50% sequence identity to the query sequences were chosen for further analysis. Using this criterion, in addition to *M. robertsii* and *M. acridum*, BTPSLs were identified in *M. acridum* (MAC), *M. majus* (MAJ), *M. guizhouense* (MGU), *M. brunneum* (MBR), *M. anisopliae* (MAN) and *Ophiocordyceps sinensis* (OCS) (Table 4.1). Interestingly, all these are entomopathogenic fungi and *O. sinensis* is closely related to *Metarhizium*. *Metarhizium album* (MAM) is another species in the *Metarhizium* genus that have been fully sequenced (Hu et al., 2014). We specifically searched its genome but no homolog of *BTPSL* genes was detected. In MBR, two distinct paralogs (MBR_10393 and MBR_09977) with 75% global amino acid sequence identity were present. The phylogeny of the group of entomopathogenic fungi including *Metarhizium* and *O. sinensis* has been well resolved (Hu et al., 2013). *Beauveria bassiana* and *Cordyceps militaris* are in a lineage sister to the common ancestor of *Metarhizium* and *O. sinensis* (Figure 4.7). Because their genome has been fully sequenced, we purposely searched genome sequences of *B. bassiana* nor *C. militaris* but BTPSL homologs were not detected in neither species.

From the same blast search of the NCBI's NR database, seven TPSs were identified from bacterial species (Table 4.3), five of which are species of the Betaproteobacteria *Burkholderia* sp, one Acidobacteria and one Actinobacteria.

These eight BTPSLs from fungi range in size from 310 to 355 amino acids. They show higher sequence homology to bacterial TPSs than typical fungal TPSs at the amino acid level. All BTPSLs contain a highly conserved aspartate-rich “DDxxxD” motif located approximately 90 aa of their N-terminus and “NDxxSxxxE” motif at the C-terminus (Figure 4.2)

4.3.3 Collinearity for the Genome Region Containing the BTPSL and the Identification of Neighbor Genes

In order to rule out the possibility of bacterial contamination, we aligned the corresponding scaffolds containing *BTPSLs* from seven entomopathogenic fungi using Mauve, which draws each collinear set of matching regions as a contiguously colored local collinear block (LCB). The alignment shows that the surrounding regions of *BTPSLs* were collinear among all the *Metarhizium* species although some assembly was of low quality (Figure 4.3). For example, in MGU, scaffold_187 (AZNH01000187.1) is only 24845 bp in size and showed high sequence similarity to others, but the expected downstream sequence (scaffold_81, AZNH01000081.1) by comparing with others was not joined together. All *BTPSLs* were located in a single LCB with two exceptions, MBR_09977 and OCS_03958. Neighboring genes in the recipient genomes were displayed below blocks as rectangles (Figure 4.3). It is not surprising that the gene order in the region surrounding *BTPSL* was found to be highly conserved across all members of the *Metarhizium* genus. In each genome, the *BTPSL* gene’s nearest upstream gene was a short hypothetical gene encoding

a protein of 148 aa. The next gene upstream was identified as a gene that encodes a C6 zinc finger protein (458 aa), one type of zinc finger proteins found exclusively in fungal species, particularly in ascomycete fungus. The nearest downstream neighbor gene of the *BTPSL* was also a C6 zinc finger, Zn(2)-Cys(6) type (Todd and Andrianopoulos, 1997). Downstream of this zinc finger was a member of the major/yellow royal jelly family of proteins that are present in all insects studied to date, bacteria and fungal species in the phyla ascomycota and basidiomycota but absent in non-insect Metazoa. In insects, bacterial origins of this gene family by horizontal gene transfer have been suggested (Ferguson et al., 2011). Members of this protein family can be found in most of the *Metarhizium* species, but not in MAC, MAM and OCS. In MAC, downstream of the *BTPSL* was a secreted aspartic proteinase, instead of the zinc finger and major/yellow royal jelly proteins. Further down, ADAM and bzip transcription factor genes could be identified. ADAMs are zinc-dependent proteases and mainly present in animal species from protozoa to mammals (Edwards et al., 2008; Huovila et al., 2005; Pollheimer et al., 2014). They have also been found in fungi but are absent in plants (Lavens et al., 2005). The bZIP transcription factors are found in all eukaryotes and play an import role in controlling gene activity (Amoutzias et al., 2007). For TPS MBR_09977, its nearest upstream gene was identified as a zinc finger protein of C2H2 type, which is almost exclusively found in eukaryotes (Iuchi, 2001). Taken together, although major/yellow royal jelly proteins are also present in bacteria, all other flanking genes are non-bacterial on either side of the *BTPSL*. It is interesting to note that there were three types of transcription factors in this

region and the rate of transcription of the *BTPSL* gene might be regulated by a few or all of them.

4.3.4 Experimental Verification of BTPSL from *M. Robertsii* as a Fungal Gene

It is always a concern that a putative HGT is actually due to DNA contamination during genome sequencing. To verify that these *BTPSL* genes were really originated from fungi and not the result of bacterial DNA contamination, *M. robertsii* was selected as a model species for experimental study. First, genomic DNA was isolated and used as template for the amplification of the coding sequence of MAA_08668. Genomic DNA sequences covering MAA_08668 and its upstream and downstream genes as well as its intergenic regions were amplified using PCR. Amplified DNA sequences were fully sequenced and its position was confirmed in the published genomic sequence (Figure 4.4).

4.3.5 The Presence of Typical Fungal TPS Genes in Relevant Fungal Species

With an increasing number of fungal species being sequenced, *TPS* genes were found to be widely distributed in the fungi kingdom especially in the phyla ascomycota and basidiomycota (Quin et al., 2014). We were interested to ask whether the group of entomopathogenic fungi examined in this study contain other, i.e. typical fungal, *TPS* genes. To answer this question, we searched their genome sequence for the presence of

putative *TPS* genes. The number of fungal TPS found in these genomes ranged from 1 to 6 (Table 4.2). Most of them were around 330 aa in size, which is similar to that of TPSs in fungi, with several exceptions (Figure 4.5 and Table 4.2).

4.3.6 Phylogenetic Analysis of BTPSLs, Related Bacterial TPSs and Other TPSs

Phylogenetic analysis was performed to understand the evolutionary relatedness of BTPSLs, related bacterial TPSs listed in Table 4.3 and typical fungal TPSs identified from entomopathogenic fungi listed in Table 4.2. It is evident that BTPSLs from fungi are monophyletic. Their sister clade are TPSs identified from bacteria listed in Table 4.3. When the intron-exon structures for all the TPS genes used in the phylogenetic analysis were compared, it is interesting to notice that as is typical in bacterial genes, none of the 8 *BTPSL* genes in entomopathogenic fungi had introns except the one in *Ophiocordyceps sinensis* (OCS_03958), which contained a single 56 bp intron (Figure 4.5). Since the phylogenetic analysis indicated their prokaryotic origin, the intron in gene OCS_03958 was probably acquired later on after transferring horizontally from bacteria. Most typical fungal *TPS* genes contain 2 introns, but some of the fungal TPSs, in particular those possessing fungal trichodiene synthase (TRI5: PF06330) domain, had no intron.

4.3.7 The *BTPSL* Gene of HGT Origin in *M. Robertsii* Encodes a Functional Enzyme

MAA_08668 was heterologously expressed in *E. coli* and the recombinant protein was tested for TPS enzyme activity. In vitro assays revealed that MAA_08668 was able to convert the common TPS substrate (E,E)-FPP into a mixture of 20 sesquiterpenes, dominated by the two alcohols epi- α -cadinol and nerolidol (Figure 4.6). γ -Cadinene, germacrene D, δ -cadinene, β -elemene, (E)- β -caryophyllene, α -cadinene, epi-bicyclosesquiphellandrene, cadina-1,4-diene and 10 unidentified sesquiterpenes were detected as minor products. Besides (E,E)-FPP, MAA_08668 also accepted (Z,E)-FPP, producing β -bisabolene, α -longipinene, α -ylangene and at least 25 further so far unidentified sesquiterpenes (Figure 4.6). Fed with (Z,Z)-FPP, the enzyme produced only trace amounts of an unidentified sesquiterpene. GPP and GGPP, the precursors for monoterpenes and diterpenes, respectively, were not accepted as substrates.

4.3.8 Molecular Evolutionary Analysis of *BTPSL* Genes of HGT Origin in *Metarhizium*

The selective forces that have shaped the evolution of the *BTPSL* genes in fungal recipients were inferred by examining the ratios of non-synonymous to synonymous nucleotide substitution rates (dN/dS) using codeml of PAML (Yang, 2007). When applied the one-ratio model that assumes the same ω ratio for all branches to our data, we got a log likelihood of -6363.8 with an estimate $\omega = 0.16093$ (Table 4.4), indicating that the major

selective force acting on all *BTPSL* genes and bacterial TPS genes in the tree (Figure 4.5) was that of purifying selection. We then applied the free-ratio model, which assumes a different ω ratio for each branch, to test whether rate heterogeneity exists among lineages. This model gave a log likelihood of -6309.81 (Table 4.4). The LRT test showed that the free-ratio model did fit the data significantly better than the one-ratio model ($2\Delta\ln L = 107.99$, $p = 0$ with $df = 26$). This indicates that significantly different ω ratios indeed existed at specific lineages. Therefore, we used PAML's branch model to test whether significant shifts in selective pressures exist following HGT. Two different ω ratios were assigned for the *BTPSL* genes in fungal recipients (ω_1 for foreground branches) and for all bacterial genes in donor groups (ω_0 for background branches). The log likelihood under this model was -6348.66 with $\omega_0 = 0.18$ and $\omega_1 = 0.00075$ (Table 4.4). The likelihood ratio test in comparing this model to one-ratio model was significant ($2\Delta\ln L = 30.28$, $p = 0$ with $df = 1$) (Table 4.4). This significantly decreased ω ratio compared to that in donor groups indicates more stringent purifying selection acting on genes in fungal recipients, demonstrating clearly a great adaptive significance of these HGT-acquired genes.

4.4 Discussion

Our comprehensive phylogenetic analysis of bacterial and fungi terpene synthases showed that a few *TPS* genes designated as bacterial terpene synthase-like (*BTPSL*) found in several entomopathogenic fungi were embedded within the bacteria TPS genes. This nested relationship indicates a bacterial origin of fungi *BTPSL* (Figure 4.1 and 4.5). To the

best of our knowledge, this is the first time HGT has been reported in TPS gene family although HGT is not uncommon in other classes of secondary metabolite genes (Bushley and Turgeon, 2010; Kroken et al., 2003; Lawrence et al., 2011; Schmitt and Lumbsch, 2009). Several independent lines of evidence suggest that fungal *BTPSL*s were achieved by HGT from bacteria. First, the regions surrounding the *BTPSL* genes in *Metarhizium* species are highly conserved and all flanking genes in this conserved syntenic region appear to be fungal sequences. If the *BTPSL* genes were from a bacteria contaminant, chances of their integration into this region would be very unlikely. Second, all these genes except the one found in OCS have no introns, which is consistent with HGT from bacteria. Third, in contrast to the typical “DDxxD” motif shared by almost all eukaryotic TPSs, all these *BTPSL*s possess a “DDxxxD” motif (Figure 4.2), which is another strong indicator of HGT from bacteria. Fourth, the presence of at least two transcription factors in the *BTPSL* surrounding region is consistent with the observation that many fungal secondary metabolite gene clusters often possess one or more transcription factors (Hoffmeister and Keller, 2007; Osbourn, 2010; Yin and Keller, 2011), indicating that the special gene cluster identified in this work is very likely the result of natural selection over time following HGT rather than a random process. Taken together, these results provide circumstantial evidence for HGT of *TPS* genes from bacteria to fungi. With this first case, we speculate that the occurrence of HGT of *TPS* genes from bacteria to fungi may not be an extremely rare event. As shown in Figure 4.1, one *TPS* gene from *A. oligospora*, a predacious fungus of nematodes, is most likely originated from an HGT event from a bacterium donor. Comprehensive phylogenetic analysis of *TPS* genes from a most current list of sequenced

bacteria and fungi will provide a better assessment on the occurrence of HGT of *TPS* genes from bacteria to fungi.

The presence of *BTPSL* genes in six of the seven *Metarhizium* species and *O. sinensis* and the absence in *B. bassiana* and *C. militaris* implies that the HGT event occurred in the immediate common ancestor of *Metarhizium* species and *O. sinensis* and after their divergence from the common ancestor of *B. bassiana* and *C. militaris*. In this scenario, the absence of *BTPSL* gene in MAM would be explained by lineage-specific gene loss. *Metarhizium* species can be classified into three groups based on their host ranges: specialist species (MAM and MAC), transitional species (MAJ and MGU) and generalist species (MAA, MBR and MAN). Previous studies have shown that specialists MAM and MAC have fewer gene clusters involved in secondary metabolite biosynthesis than other *Metarhizium* species or plant-associated fungi (Hu et al., 2014). It is tempting to speculate that the loss of *BTPSL* gene in MAM is neutral or even adaptive. On the other hand, it is interesting to observe that for typical fungal terpene synthases, the number of genes varied little among entomopathogenic fungi species studied except in two generalist species OCS (6) and BBA (5) (Table 4.2). Further functional studies of all these *TPS* genes in each of these entomopathogenic fungi will help explain how these genes contribute to their success of living in a more diverse environment and invasion of varying range of insect hosts.

The actual donor bacterium species is hard to determine due to the insufficient taxon sampling. Bacterial sequences which shared more than 50% sequence identity with *BTPSL* genes identified in fungi are from three different bacterial groups: Betaproteobacteria,

Acidobacteria and Actinobacteria, which are all possible sources of HGT. Most of them are from the genus *Burkholderia* of Betaproteobacteria, containing a wide variety of Gram-negative species that occupy a wide range of ecological niches, such as soil, water, plant rhizosphere and fungi. Many *Burkholderia* species contain so-called genomic islands (GIs) that are the main contributors to the great genomic plasticity within this genus (Holden et al., 2004; Kim et al., 2005; Tuanyok et al., 2008). These GIs can perform a broad range of functions and are more likely to arise from horizontal gene transfer between different bacteria. Their genomes are open as evidenced by the previous finding that only about 3% of protein-coding gene families are conserved across all *Burkholderia* species (Ussery et al., 2009). Previous studies also have shown that *Metarhizium* species are prone to accept exogenous DNA, particular for nonspecialist species (Hu et al., 2014). Therefore it is possible that the *BTPSL* genes originated from a *Burkholderia* species. The mechanism of the HGT is not clear, but one pre-condition must be met for HGT to occur, that is the close contact between donor and recipient. Given the diverse ecological niches (particular in or on fungal mycelia) *Burkholderia* are occupying, contact between donor and recipient is guaranteed and routes for HGT into fungi are more likely to establish.

Compared to vertical inheritance, the acquisition of new genes by HGT is random and dangerous. So, gain of beneficial genes is not guaranteed. The majority of horizontally transferred genes get lost in the recipient genomes very quickly except those increasing the fitness of the recipient organism (Hao and Golding, 2006; Novozhilov et al., 2005). In *Metarhizium* species, the fact that these *BTPSL* genes are arranged in clusters along with other genes like transcription factors (a typical characteristic of fungal secondary

metabolism) indicates that they are under strong selection and their expression are probably fine-tuned by these transcription factors. Previous studies have shown that in *Metarhizium* species transcription factors are under strong positive selection, particularly the Zn(2)-Cys(6) type and C2H2 type (Hu et al., 2014), as compared to the *BTPSL* genes that are under strong purifying selection. The changes in transcription factors and presence of *BTPSL* genes in *Metarhizium* species except the basal species MAM could, in part, explain the expanded host range. Within *Metarhizium* species, there are other differences in gene contents and transcriptional regulations contributing to the differences in pathogenicity and host range, such as different number of proteases, cytochrome P450 enzymes, polyketide synthases and dehydrogenases (Gao et al., 2011; Hu et al., 2014). Although MAA_08668 showed TPS enzyme activity and catalyzed the formation of multiple sesquiterpenes *in vitro*, the *in vivo* function and identity of the secondary metabolites produced by this horizontally transferred gene remains unknown. Further experimental studies, such as mutational analysis, are needed to confirm the requirement of this type of *TPS* gene for pathogenicity and to determine the active molecules that directly cause disease in insects.

4.5 Materials and Methods

4.5.1 Data Sources and Analysis

Bacterial and fungal terpene synthases (TPSs) were obtained from Pfam (version 27.0). Since the majority of TPSs from these two kingdoms contains only the C terminal domain

of the TPS protein, only sequences that included in Pfam entry PF03936 were retrieved. The original dataset contained 287 bacterial TPSs in 140 species and 202 fungal TPSs in 69 species. In order to remove spurious sequences, the downloaded sequences were subjected to a HMMER (Finn et al., 2011) search against the Pfam-A database. Only sequences with the PF03936 domain as the best-matched domain and active in UniProt database were kept in our analysis. This resulted in a final dataset of 275 bacterial TPSs in 133 species and 181 fungal TPSs in 64 species.

TPSs in 7 *Metarhizium* species, *Ophiocordyceps sinensis*, *Cordyceps militaris*, *Beauveria bassiana* ARSEF 2860 and *Beauveria bassiana* D1-5 were mined from available protein sequences in NCBI protein database using HMMER at an e-value of $1e^{-2}$. For detailed information about the species and the number of TPSs in each species, see Table 4.1 and 4.2.

Sequences similar to the BTPSLs in *Metarhizium* species were retrieved from NCBI nr database based on blastp searches with the BTPSLs from *M. robertsii* (MAA_08668) and *M. acridum* (MAC_05714) as the queries. The blastp hits with at least 50% sequence identity to the query were chosen.

4.5.2 Multiple Sequence Alignments and Phylogenetic Inference

TPSs obtained from Pfam were first clustered at 80% sequence identity in each kingdom (bacteria and fungi) using CD-HIT (Li and Godzik, 2006) to eliminate highly similar sequences within each kingdom. After clustering, the number of TPSs for bacteria and fungi were reduced to 165 and 138, respectively. The corresponding C

terminal domain sequences from bacterial and fungal TPSs were retrieved based on the coordinates predicted by HMMER. All multiple sequence alignments were made using MAFFT (v7.130b) (Katoh and Standley, 2013) in a highly accurate setting (L-INS-i) with 1000 iterations of improvement. The appropriate amino acid substitution model was determined using ProtTest version 3.4 (Darriba et al., 2011) for each alignment according to Akaike information criterion (AIC) and Bayesian information criterion (BIC). The improved general amino acid substitution matrix with empirical base frequencies along with a gamma distribution (LG+G+F) was obtained as the most appropriate model for all protein datasets. Maximum likelihood analyses were performed using RAxML version 8.1.11 (Stamatakis, 2014) with 1000 bootstrap replicates under the best substitution model for each dataset via the online CIPRES Science gateway portal (Miller et al., 2010).

The maximum-likelihood tree shown in Figure 4.5 was inferred from the codon alignment of TPS genes identified from several entomopathogenic fungi and TPS genes in bacteria that showing great similarity to BTPSLs. The codon alignment was generated using PAL2NAL (Suyama et al., 2006) from the MAFFT protein alignment (L-INS-I method with 1000 iterations of improvement) and the corresponding nucleotide sequences. The maximum-likelihood analysis was performed using PhyML 3.1 (Guindon et al., 2010) with GTR+I+G nucleotide substitution model chosen by jModeltest2 (Darriba et al., 2012) based on the AIC and BIC criteria. The robustness of the phylogenetic tree was estimated by bootstrapping with 1000 replicates.

4.5.3 Multiple Genome Alignment

For comparisons of sequences at the genome level, homologous contig sequences from each of seven species were aligned with sequence of MAA as the reference using the progressive alignment algorithm of the MAUVE Multiple Genome Aligner (version 2.4.0) at default settings (Darling et al., 2004).

4.5.4 Analysis of Selective Pressure

Selection analyses were performed using the codeml program implemented in the PAML 4.8 (Yang, 2007). To detect the different selective pressure acting on the *TPS* gene in fungal recipients (foreground branches) and genes in bacterial donor groups (background branches), we used PAML's branch models that allow ω to vary among lineages. The likelihood ratio test (LRT) was carried out to determine if foreground branches are under significantly different selection pressure from the background branches by comparing one-ratio against two-ratios.

4.5.5 Fungal Culture

Metarhizium robertsii (MAA) isolate, ARSEF 23, was ordered from USDA ARS Collection of Entomopathogenic Fungal Cultures (ARSEF), Ithaca, New York. Lyophilized isolates were resuspended with 500 μ l sterilized distilled water and suspension solution was cultured on Difco potato dextrose agar medium (PDA) at 28°C for one week.

4.5.6 MAA_08668 Gene Cloning and Verification of Its Position

Mycelia of *M. robertsii* were used to isolate genomic DNA. The genomic DNA was extracted using Plant Genomic DNA Extraction Miniprep System (<http://www.viogene.com>), according to the protocol recommended by the manufacturer. The MAA_08668 gene was amplified from genomic DNA using the forward primer 5'-ATGGAAAAACAAAGACTGAAAGCTC-3' and the reverse primer 5'-CTAGACCAAGCTGCTCGTTGACTCAG-3' and the resulting PCR product was cloned into pEXP5-CT/TOPO vector (www.lifetechnologies.com), according to the protocol provided by the manufacturer. MAA_08668 gene cloned into pEXP5-CT/TOPO was confirmed by sequencing.

Position of MAA_08668 in the published *Metarhizium* genomic sequences was verified by PCR. 10591 bp of DNA sequences covering MAA_08668 and its neighbor genes were amplified using four pairs of primers: 11039F 5'-ATGCCGCTGGCAGACTTGCGGTAC - 3', 6768R1 5'-GAGGTGAGACACCAGCGTTATG-3', 6768F2 5'-CTACGTAGCCATTAGCG AGAGGGAC-3', 8868R1 5'-GACTCAAGTACTGGATACGAGGTTAC-3', 8868F2 5'-GTCAG TCACCAGCAGATTATGCTC-3', 6869R1 5'-CTAATTGTCGACACTACTCTGGCCAG-3', 6869F2 5'-CAACTTCATGTGCACTACTATAGCAAG-3', 8869R 5'-TACACAAGGCCGGCA AGAACAGTG 3'. The PCR products were subcloned into pGEM T-EASY vector and fully sequenced.

4.5.7 Biochemical Characterization of MAA_08668

The *E.coli* BL21 codon plus strain (<http://www.lifetechnologies.com>), transformed with the plasmid containing MAA_08668, was used for protein expression. The BL21 culture was grown in liquid LB at 37°C until the culture reached OD₆₀₀ of 0.6. Protein expression was induced by addition of 1 M isopropylthio- β -galactoside (IPTG) to a final concentration of 1 mM. After 20 hours incubation at 18°C, cells were harvested by centrifugation at 6000 g, resuspended in protein extraction buffer (50 mM Mopso (pH7.0), 5 mM MgCl₂, 5 mM Sodium ascorbate, 5 mM dithiothreitol, 0.5 mM PMSF and 10% (v/v) glycerol) and disrupted by a 4 × 30 sec treatment with a sonicator (Microson XL 2000; Misonix, Farmingdale, New York). Cell debris was removed by centrifugation at 14,000 rpm (20 min, 4°C) and the supernatant was desalted by passage through a PD-10 Desalting Column (<http://www.gelifesciences.com>) into assay buffer (10 mM Mopso, pH7.0, 1 mM dithiothreitol, 10% (v/v) glycerol).

To determine the catalytic activity of MAA_08668, enzyme assays were conducted in a Teflon-sealed, screw-capped 1 ml GC glass vial containing 40 μ l of the bacterial extract and 60 μ l assay buffer containing 10 μ M substrate [(E,E)-FPP, (Z,E)-FPP, (Z,Z)-FPP and GPP, respectively], 10 mM MgCl₂, 0.2 mM NaWO₄ and 0.1 mM NaF. A SPME (solid phase microextraction) fiber consisting of 100 μ m polydimethylsiloxane (<http://www.sigmaaldrich.com>) was placed into the headspace of the vial for 45 min incubation at 30°C to adsorb the TPS reaction products. Product analysis was conducted using an Agilent 6890 Series gas chromatograph (GC) coupled to an Agilent 5973

quadrupole mass selective detector (interface temp, 250°C; quadrupole temp, 150°C; source temp, 230°C; electron energy, 70 eV). The GC was operated with a DB-5MS column (Agilent, Santa Clara, USA, 30 m x 0.25 mm x 0.25 µm). The sample (SPME) was directly injected without split at an initial oven temperature of 80°C. The temperature was held for 3 min, then increased to 240°C with a gradient of 7°C min⁻¹, and further increased to 300°C with a gradient of 60°C min⁻¹ and a hold of 2 min. To determine potential diterpene synthase activity, assays were set up as described above, containing 50 µM (E,E,E)-GGPP as substrate, and were overlaid with 100 µl hexane. After incubation for 60 min at 30°C, the hexane phase was collected and analyzed using GC-MS.

4.6 Acknowledgements

We thank USDA ARS Collection of Entomopathogenic Fungal Cultures (ARSEF), Ithaca, New York for providing us with the *Metarhizium robertsii* (MAA) isolate, ARSEF

23.

4.7 Bibliography

- Amoutzias, G. D., Veron, A. S., Weiner, J., r., Robinson-Rechavi, M., Bornberg-Bauer, E., Oliver, S. G., and Robertson, D. L. (2007). One billion years of bzip transcription factor evolution: conservation and change in dimerization and dna-binding site specificity. *Molecular Biology and Evolution*, 24(3):827–35. [113](#)
- Bushley, K. E. and Turgeon, B. G. (2010). Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evolutionary Biology*, 10:26. [108](#), [118](#)
- Campbell, M. A., Rokas, A., and Slot, J. C. (2012). Horizontal transfer and death of a fungal secondary metabolic gene cluster. *Genome Biol Evol*, 4(3):289–93. [107](#)
- Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant Journal*, 66(1):212–29. [107](#), [108](#)
- Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–403. [124](#)
- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–5. [123](#)
- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jmodeltest 2: more models, new heuristics and parallel computing. *Nature Methods*, 9(8):772. [123](#)

- Degenhardt, J., Kollner, T. G., and Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, 70(15-16):1621–37. [108](#)
- Edwards, D. R., Handsley, M. M., and Pennington, C. J. (2008). The adam metalloproteinases. *Molecular Aspects of Medicine*, 29(5):258–89. [113](#)
- Ferguson, L. C., Green, J., Surridge, A., and Jiggins, C. D. (2011). Evolution of the insect yellow gene family. *Molecular Biology and Evolution*, 28(1):257–72. [113](#)
- Finn, R. D., Clements, J., and Eddy, S. R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–37. [122](#)
- Gao, Q., Jin, K., Ying, S. H., Zhang, Y., Xiao, G., Shang, Y., Duan, Z., Hu, X., Xie, X. Q., Zhou, G., Peng, G., Luo, Z., Huang, W., Wang, B., Fang, W., Wang, S., Zhong, Y., Ma, L. J., St Leger, R. J., Zhao, G. P., Pei, Y., Feng, M. G., Xia, Y., and Wang, C. (2011). Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi metarhizium anisopliae and m. acridum. *PLoS Genetics*, 7(1):e1001264. [121](#)
- Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews. Microbiology*, 3(9):679–87. [106](#)
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology*, 59(3):307–321. [123](#)

- Hao, W. and Golding, G. B. (2006). The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Research*, 16(5):636–43. [120](#)
- Hartmann, T. (2007). From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry*, 68(22-24):2831–2846. [106](#)
- Hoffmeister, D. and Keller, N. P. (2007). Natural products of filamentous fungi: enzymes, genes, and their regulation. *Natural Product Reports*, 24(2):393–416. [118](#)
- Holden, M. T., Titball, R. W., Peacock, S. J., Cerdeno-Tarraga, A. M., Atkins, T., Crossman, L. C., Pitt, T., Churcher, C., Mungall, K., Bentley, S. D., Sebaihia, M., Thomson, N. R., Bason, N., Beacham, I. R., Brooks, K., Brown, K. A., Brown, N. F., Challis, G. L., Cherevach, I., Chillingworth, T., Cronin, A., Crossett, B., Davis, P., DeShazer, D., Feltwell, T., Fraser, A., Hance, Z., Hauser, H., Holroyd, S., Jagels, K., Keith, K. E., Maddison, M., Moule, S., Price, C., Quail, M. A., Rabinowitsch, E., Rutherford, K., Sanders, M., Simmonds, M., Songsivilai, S., Stevens, K., Tumapa, S., Vesaratchavest, M., Whitehead, S., Yeats, C., Barrell, B. G., Oyston, P. C., and Parkhill, J. (2004). Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proceedings of the National academy of Sciences of the United States of America*, 101(39):14240–5. [120](#)
- Hu, X., Xiao, G., Zheng, P., Shang, Y., Su, Y., Zhang, X., Liu, X., Zhan, S., St Leger, R. J., and Wang, C. (2014). Trajectory and genomic determinants of fungal-pathogen speciation and host adaptation. *Proceedings of the National academy of Sciences of the United States of America*, 111(47):16796–801. [111](#), [119](#), [120](#), [121](#)

- Hu, X., Zhang, Y., Xiao, G., Zheng, P., Xia, Y., Zhang, X., St Leger, R. J., Liu, X., and Wang, C. (2013). Genome survey uncovers the secrets of sex and lifestyle in caterpillar fungus. *Chinese Science Bulletin*, 58(23):2846–2854. [111](#)
- Huovila, A. P., Turner, A. J., Peltto-Huikko, M., Karkkainen, I., and Ortiz, R. M. (2005). Shedding light on adam metalloproteinases. *Trends in Biochemical Sciences*, 30(7):413–22. [113](#)
- Iuchi, S. (2001). Three classes of c2h2 zinc finger proteins. *Cellular and Molecular Life Sciences*, 58(4):625–35. [113](#)
- Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–80. [123](#)
- Keeling, P. J. and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews. Genetics*, 9(8):605–18. [106](#)
- Keller, N. P., Turner, G., and Bennett, J. W. (2005). Fungal secondary metabolism - from biochemistry to genomics. *Nature Reviews. Microbiology*, 3(12):937–47. [107](#)
- Khalidi, N. and Wolfe, K. H. (2011). Evolutionary origins of the fumonisin secondary metabolite gene cluster in fusarium verticillioides and aspergillus niger. *Int J Evol Biol*, 2011:423821. [107](#)
- Kim, H. S., Schell, M. A., Yu, Y., Ulrich, R. L., Sarria, S. H., Nierman, W. C., and DeShazer, D. (2005). Bacterial genome adaptation to niches: divergence of the potential virulence

- genes in three burkholderia species of different survival strategies. *BMC Genomics*, 6:174. [120](#)
- Kliebenstein, D. J., Lambrix, V. M., Reichelt, M., Gershenzon, J., and Mitchell-Olds, T. (2001). Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in arabidopsis. *Plant Cell*, 13(3):681–93. [106](#)
- Kroken, S., Glass, N. L., Taylor, J. W., Yoder, O. C., and Turgeon, B. G. (2003). Phylogenomic analysis of type i polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proceedings of the National academy of Sciences of the United States of America*, 100(26):15670–5. [108](#), [118](#)
- Köksal, M., Jin, Y., Coates, R. M., Croteau, R., and Christianson, D. W. (2011). Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature*, 469(7328):116–122. [107](#)
- Lavens, S. E., Rovira-Graells, N., Birch, M., and Tuckwell, D. (2005). Adams are present in fungi: identification of two novel adam genes in aspergillus fumigatus. *FEMS Microbiology Letters*, 248(1):23–30. [113](#)
- Lawrence, D. P., Kroken, S., Pryor, B. M., and Arnold, A. E. (2011). Interkingdom gene transfer of a hybrid nps/pks from bacteria to filamentous ascomycota. *PLOS ONE*, 6(11):e28231. [107](#), [108](#), [118](#)

- Li, G., Kollner, T. G., Yin, Y., Jiang, Y., Chen, H., Xu, Y., Gershenzon, J., Pichersky, E., and Chen, F. (2012). Nonseed plant *selaginella moellendorffi* [corrected] has both seed plant and microbial types of terpene synthases. *Proceedings of the National academy of Sciences of the United States of America*, 109(36):14711–5. [107](#)
- Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–9. [122](#)
- Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). Creating the cipres science gateway for inference of large phylogenetic trees. In *Gateway Computing Environments Workshop (GCE), 2010*, pages 1–8. [123](#)
- Novozhilov, A. S., Karev, G. P., and Koonin, E. V. (2005). Mathematical modeling of evolution of horizontally transferred genes. *Molecular Biology and Evolution*, 22(8):1721–32. [120](#)
- Ober, D. (2005). Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends in Plant Science*, 10(9):444–9. [106](#)
- Osbourn, A. (2010). Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics*, 26(10):449–57. [118](#)
- Pollheimer, J., Fock, V., and Knofler, M. (2014). Review: the adam metalloproteinases - novel regulators of trophoblast invasion? *Placenta*, 35 Suppl:S57–63. [113](#)
- Proctor, R. H., Van Hove, F., Susca, A., Stea, G., Busman, M., van der Lee, T., Waalwijk, C., Moretti, A., and Ward, T. J. (2013). Birth, death and horizontal transfer of the fumonisins

- biosynthetic gene cluster during the evolutionary diversification of fusarium. *Molecular Microbiology*, 90(2):290–306. [107](#)
- Quin, M. B., Flynn, C. M., and Schmidt-Dannert, C. (2014). Traversing the fungal terpenome. *Natural Product Reports*, 31(10):1449–73. [114](#)
- Schmitt, I. and Lumbsch, H. T. (2009). Ancient horizontal gene transfer from bacteria enhances biosynthetic capabilities of fungi. *PLOS ONE*, 4(2):e4437. [106](#), [108](#), [118](#)
- Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–3. [123](#)
- Suyama, M., Torrents, D., and Bork, P. (2006). Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(Web Server issue):W609–12. [123](#)
- Tholl, D. (2006). Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Current Opinion in Plant Biology*, 9(3):297–304. [108](#)
- Todd, R. B. and Andrianopoulos, A. (1997). Evolution of a fungal regulatory gene family: the zn(ii)2cys6 binuclear cluster dna binding motif. *Fungal Genetics and Biology*, 21(3):388–405. [113](#)
- Tuanyok, A., Leadem, B. R., Auerbach, R. K., Beckstrom-Sternberg, S. M., Beckstrom-Sternberg, J. S., Mayo, M., Wuthiekanun, V., Brettin, T. S., Nierman, W. C., Peacock, S. J., Currie, B. J., Wagner, D. M., and Keim, P. (2008). Genomic islands from five strains of burkholderia pseudomallei. *BMC Genomics*, 9:566. [120](#)

- Ussery, D. W., Kiil, K., Lagesen, K., Sicheritz-Ponten, T., Bohlin, J., and Wassenaar, T. M. (2009). The genus burkholderia: analysis of 56 genomic sequences. *Genome Dyn*, 6:140–57. [120](#)
- Vranova, E., Coman, D., and Grussem, W. (2013). Network analysis of the mva and mep pathways for isoprenoid synthesis. *Annual Review of Plant Biology*, 64:665–700. [108](#)
- 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. *Proceedings of the National academy of Sciences of the United States of America*, 112(3):857–62. [107](#)
- Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–91. [116](#), [124](#)
- Yin, W. and Keller, N. P. (2011). Transcriptional regulatory elements in fungal secondary metabolism. *J Microbiol*, 49(3):329–39. [118](#)

4.8 Appendix

Table 4.1: Presence/Absence of BTPSLs for each entomopathogenic fungi examined in this study.

Organism	BTPSLs	Length (aa)
Metarhizium album (MAM)	ND ^a	
Metarhizium acridum (MAC)	MAC_05714	349
Metarhizium majus (MAJ)	MAJ_08936	355
Metarhizium guizhouense (MGU)	MGU_11447	348
Metarhizium brunneum (MBR)	MBR_10393	332
	MBR_09977	342
Metarhizium anisopliae (MAN)	MAN_10655	355
Metarhizium robertsii (MAA)	MAA_08668	355
Ophiocordyceps sinensis (OCS)	OCS_03958	310
Cordyceps militaris (CCM)	ND	
Beauveria bassiana (BBA)	ND	
Beauveria bassiana D1-5 (B15)	ND	

^a not detected.

Table 4.2: List of typical fungal TPSs for each entomopathogenic fungi.

Organism	Typical fungal TPSs	Length (aa)
Beauveria bassiana D1-5 (B15)	B15_g10715	320
	B15_g11707	410
	B15_g9145	241
	B15_g8322	308
	B15_g8139	805
Beauveria bassiana (BBA)	BBA_00396	193
	BBA_03632	320
	BBA_08696	347
	BBA_08783	805
	BBA_08897	308
Cordyceps militaris (CCM)	CCM_03050	228
Metarhizium acridum (MAC)	MAC_09823	215
	MAC_06344	194
Metarhizium anisopliae (MAN)	MAN_07035	331
	MAN_02977	327
Metarhizium album (MAM)	MAM_08406	1009
	MAM_06660	321
Metarhizium brunneum (MBR)	MBR_03882	331
	MBR_00969	326
Metarhizium guizhouense (MGU)	MGU_07289	331
	MGU_07703	326
	MGU_08368	256
	MGU_10775	472
Metarhizium majus (MAJ)	MAJ_05691	331
	MAJ_07611	248
	MAJ_10664	459
	MAJ_11006	181
Metarhizium robertsii (MAA)	MAA_06799	331
	MAA_07176	326
	MAA_06544	320
	MAA_06581	328
Ophiocordyceps sinensis (OCS)	OCS_00486	389
	OCS_04474	424
	OCS_00133	281
	OCS_01341	440
	OCS_00754	387
	OCS_06263	360

Table 4.3: Presence/Absence of BTPSLs for each entomopathogenic fungi examined in this study.

Gene name	Organism	Strain	Taxonomy ID	Length (aa)
B_MSh2	Burkholderia sp.	MSh2	1506588	340
B_MSh1	Burkholderia sp.	MSh1	1506587	340
B_AU4i	Burkholderia sp.	AU4i	1335308	347
B_ce	Burkholderia cepacia complex	NA ^a	87882	342
B_ceJBK9	Burkholderia cepacia	JBK9	1395570	311
G_ma	Granulicella mallensis	NA	940614	333
K_se	Kitasatospora setae	NA	2066	320

^a not available.

Table 4.4: Presence/Absence of BTPSLs for each entomopathogenic fungi examined in this study.

Model	Estimates of Parameters	np	lnL	$2\Delta\ln L(\text{LRT})$
One-ratio	$\omega^a = 0.16093$	29	-6363.800274	Free-ratio vs. One-ratio
Free-ratio		55	-6309.807230	107.986088 (df=26; p=5.824e-12)
Two-ratios				Two-ratios vs. One-ratio
BTPSLs	$\omega_0^b = 0.18221; \omega_1^c = 0.00075$	30	-6348.661135	30.278278 (df=1; p=3.743e-08)

^a ω = nonsynonymous/synonymous rate ratio (dN/dS).

^b ω_0 = ω for bacterial TPSs branch.

^c ω_1 = ω for BTPSLs branch.

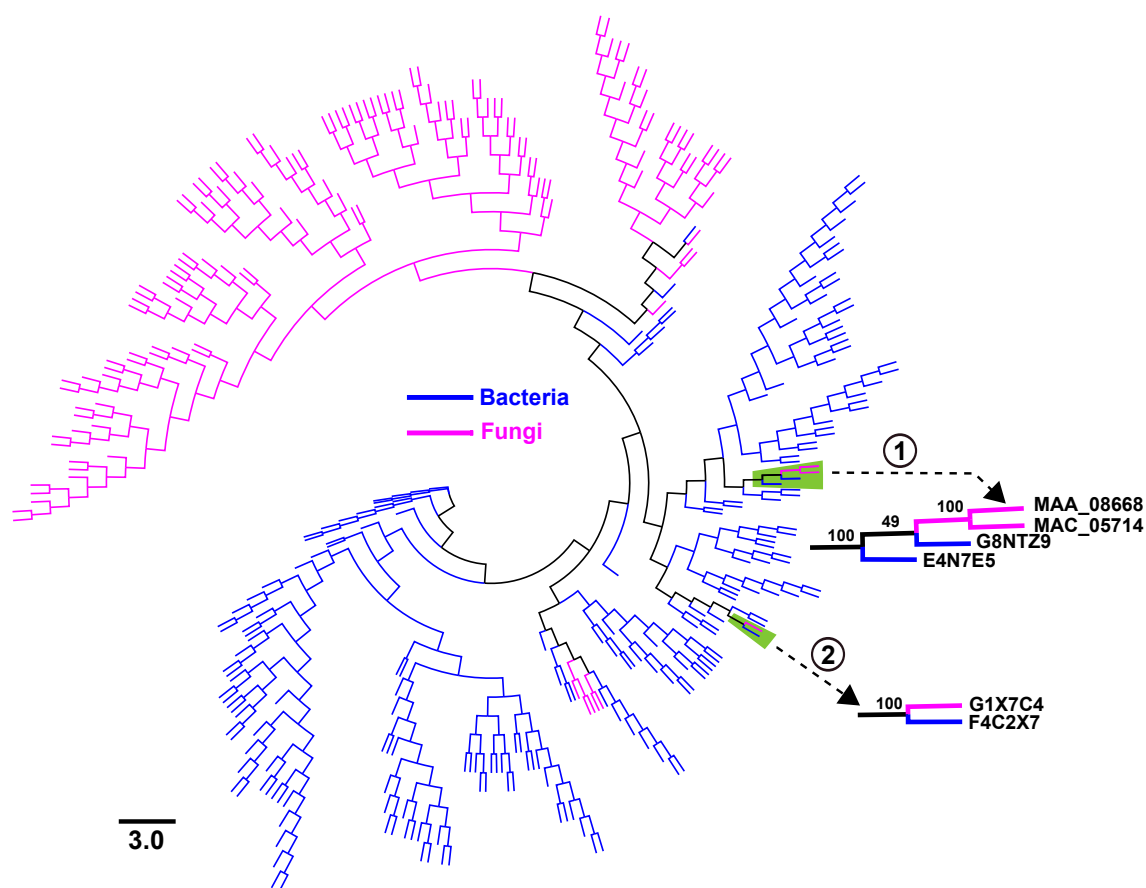


Figure 4.1: Maximum Likelihood phylogenetic tree of 341 Terpene Synthase C-domain sequences from bacteria (Blue) and fungi (Magenta) collected in the Pfam database. Branches with putative horizontal gene transfer events are labeled in green. Enlarged views of the two clusters are depicted on the right. Scale bars indicate the numbers of substitutions per site.

```

      *          20          *          40          *          60          *          80
MAN_10655 : ME-----KQRIKAQLSSLRVLELSVPWFQGCNSKAEVVEARMWKWAEENLLVTDEYRNRVIRTRYGL : 63
MAA_08668 : ME-----KQRIKAQLSSLRVLELSVPWFQGCNSKAEVVEARMWKWAEENLLVTDEYRNRVIRTRYGL : 63
MGU_11447 : ME-----KQRIKAQLSSLRVLELSVPWFQGCNSKAEVVEARMWKWAEENLLVTDEYRNRVIRTRYGL : 63
MBR_10393 : ME-----KQRIKAQLSSLRVLELSVPWFQGCNSKAEVVEARMWKWAEENLLVTDEYRNRVIRTRYGL : 63
MAJ_08936 : ME-----KQRIKAQLSSLRVLELSVPWFQGCNSKAEVVEARMWKWAEENLLVTDEYRNRVIRTRYGL : 63
MAC_05714 : ME-----KQRIKAQLSSLRVLELSVPWFQGCNSKAEVVEARMWKWAEENLLVTDEYRNRVIRTRYGL : 63
MBR_09977 : ME-----KQRIKAQLSSLRVLELSVPWFQGCNSKAEVVEARMWKWAEENLLVTDEYRNRVIRTRYGL : 63
OCS_03958 : ME-----KQRIKAQLSSLRVLELSVPWFQGCNSKAEVVEARMWKWAEENLLVTDEYRNRVIRTRYGL : 63
B_MSh1 : MK-----TDAAKIIDAAALRIEVDVFWAGACSPHQRIETRMLEWADDEGLVNDMYRKRVMMTRYGW : 63
CC1G_10050 : MP-----SP-AGALP[SFI]LPDLVN-----DC[F]FLRVNPLCDE[V]GRLSEQ[FLR]A-NYSPRAVAF[V]ALK[A]E : 63

      *          100          *          120          *          140          *          160
MAN_10655 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVSGSPPVFGEIAWLDVQCRLRR : 141
MAA_08668 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVAGSPPVFGEIAWLDVQCRLRR : 141
MGU_11447 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVAGSPPVFGEIAWLDVQCRLRR : 141
MBR_10393 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVAGSPPVFGEIAWLDVQCRLRR : 141
MAJ_08936 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVAGSPPVFGEIAWLDVQCRLRR : 141
MAC_05714 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVAGSPPVFGEIAWLDVQCRLRR : 141
MBR_09977 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVAGSPPVFGEIAWLDVQCRLRR : 141
OCS_03958 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVAGSPPVFGEIAWLDVQCRLRR : 141
B_MSh1 : LAARCYPNAGEELLOVIADLVWFFITDDLFVDRVEVADEIIRNLT--AMVDVLDLNVAGSPPVFGEIAWLDVQCRLRR : 141
CC1G_10050 : LTAACYPDADAFELRVSDDMNLEENADD--LDDFDIEDTYGLA[C]VRAIRLPVNFTDKRAG[MTK]SYE----S[FI]LK : 138

      *          180          *          200          *          220          *          240
MAN_10655 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
MAA_08668 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
MGU_11447 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
MBR_10393 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
MAJ_08936 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
MAC_05714 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
MBR_09977 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
OCS_03958 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
B_MSh1 : LLQVEAFERFAQGMRLWATTAAQILNHLRPTSVGIREYOTIRRHISGNNPCTSLADAANKGSVQACEFYADAVQTLVQ : 221
CC1G_10050 : TAGPRCTERFIQTLALYFESVVTQKQARNNGTLPLDESMTIRRRNSGCKPCYALIEFCAGIDLPEDEVINHPIIQSLDA : 218

      *          260          *          280          *          300          *          320
MAN_10655 : TNNIVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
MAA_08668 : TNNIVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
MGU_11447 : TNNIVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
MBR_10393 : TNNIVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
MAJ_08936 : TNNIVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
MAC_05714 : TNNIVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
MBR_09977 : ANNVVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
OCS_03958 : ANNVVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
B_MSh1 : ANNVVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 299
CC1G_10050 : SNLLIAMSNDIFSFNREQSRHDSFNMVSI[MH]KCFAL[EA]NVFVGLCKKAMER[Q]ADKRN[PSW]GPEIDG[V]AMVVDG : 298

      *          340          *          360          *
MAN_10655 : LKYWIRGYLDWVVDH[IR]VADQ[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
MAA_08668 : LKYWIRGYLDWVVDH[IR]VADH[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
MGU_11447 : LKYWIRGYLDWVVDH[IR]VADQ[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
MBR_10393 : LKYWIRGYLDWVVDH[IR]VADQ[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
MAJ_08936 : LKYWIRGYLDWVVDH[IR]VADQ[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
MAC_05714 : LKYWIRGYLDWVVDH[IR]VADQ[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
MBR_09977 : LKYWIRGYLDWVVDH[IR]VADQ[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
OCS_03958 : LKYWIRGYLDWVVDH[IR]VADQ[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
B_MSh1 : LKYWIRGYLDWVVDH[IR]VADQ[IESD]ADDRFSA[PD]SLLNNSC[SSV]TEST[RS]LV : 355
CC1G_10050 : LQNVVWANDIOSLRTEIHQPGQFRNMVITYAQOGQSLQDAVETATRVNKLIASTCELANAFATARP--ISDELHGLIDG : 298

```

Figure 4.2: Amino acid sequence alignment of 8 BTPSL TPSs, and representative TPSs from bacteria and fungi. 8 BTPSL TPSs: MAN_10655 from *Metarhizium anisopliae*, MAA_08668 from *Metarhizium robertsii*, MGU_11447 from *Metarhizium guizhouense*, MBR_10393 and MBR_09977 from *Metarhizium brunneum*, MAJ_08936 from *Metarhizium majus*, MAC_05714 from *Metarhizium acridum*, OCS_03958 from *Ophiocordyceps sinensis*; Representative TPSs from bacteria and fungi: B_MSh1 from bacteria, *Burkholderia sp* MSh1; CC1G_10050 from fungi, *Coprinosia cinerea* okayama7#130. The positions of DDxx(x)D and NDxxSxxxE motifs are indicated by black bars.

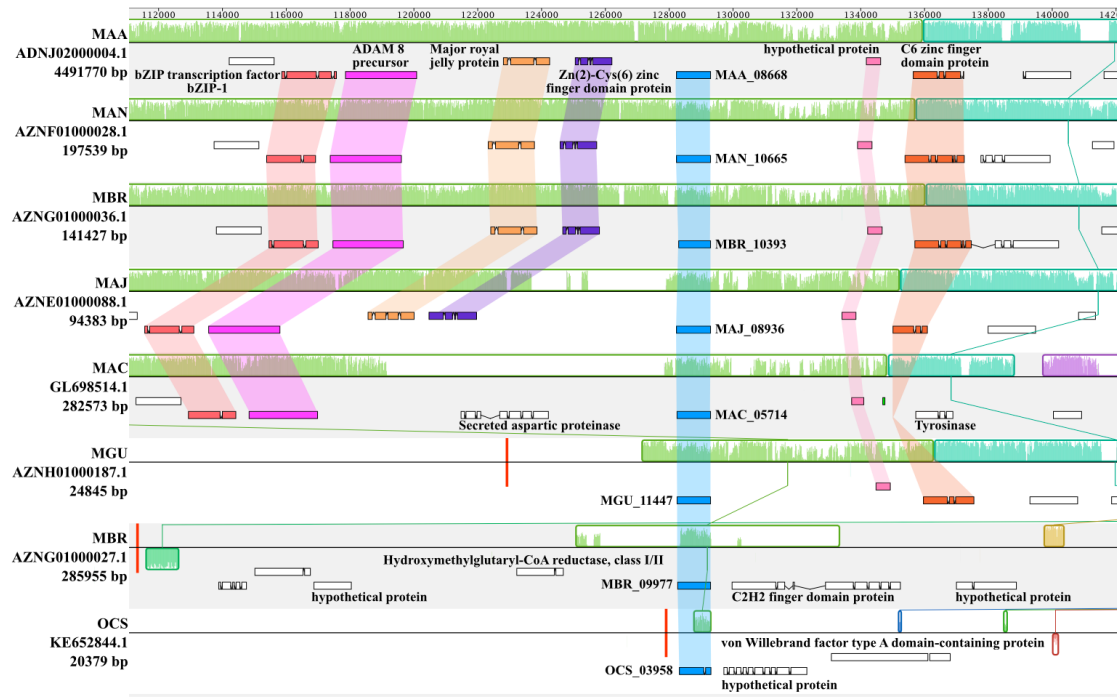


Figure 4.3: Alignment of the orthologous genomic regions containing the *BTPSL* genes from seven fungi species. Each horizontal profile indicates one genomic region with the height as the indicator of degree of sequence conservation. The white regions correspond to unaligned sequences. The regions that share the same color are locally collinear blocks. Below each profile are gene models and orthologous genes are colored and connected by shaded boxes in the same color. Red vertical lines delineate the ends of contigs. Three letter codes correspond to species: MAA, *Metarhizium robertsii*; MAN, *Metarhizium anisopliae*; MBR, *Metarhizium brunneum*; MAJ, *Metarhizium majus*; MAC, *Metarhizium acridum*; MGU, *Metarhizium guizhouense*; OCS, *Ophiocordyceps sinensis*. The scaffold names with length information are displayed under species codes.

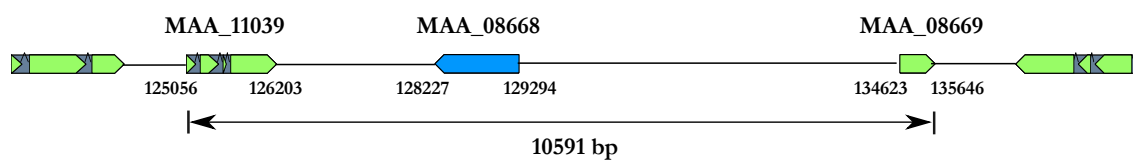


Figure 4.4: Confirmation of *MAA_08668*, a fungal TPS gene of putative horizontal gene transfer origin, integration into *Metarhizium robertsii* genome. The genomic region spanning *MAA_08668*, its neighboring genes (*MAA_11039*: Zn(2)-Cys(6) zinc finger domain protein; *MAA_08669*: hypothetical protein) and the intergenic region were amplified using PCR and confirmed by sequencing.

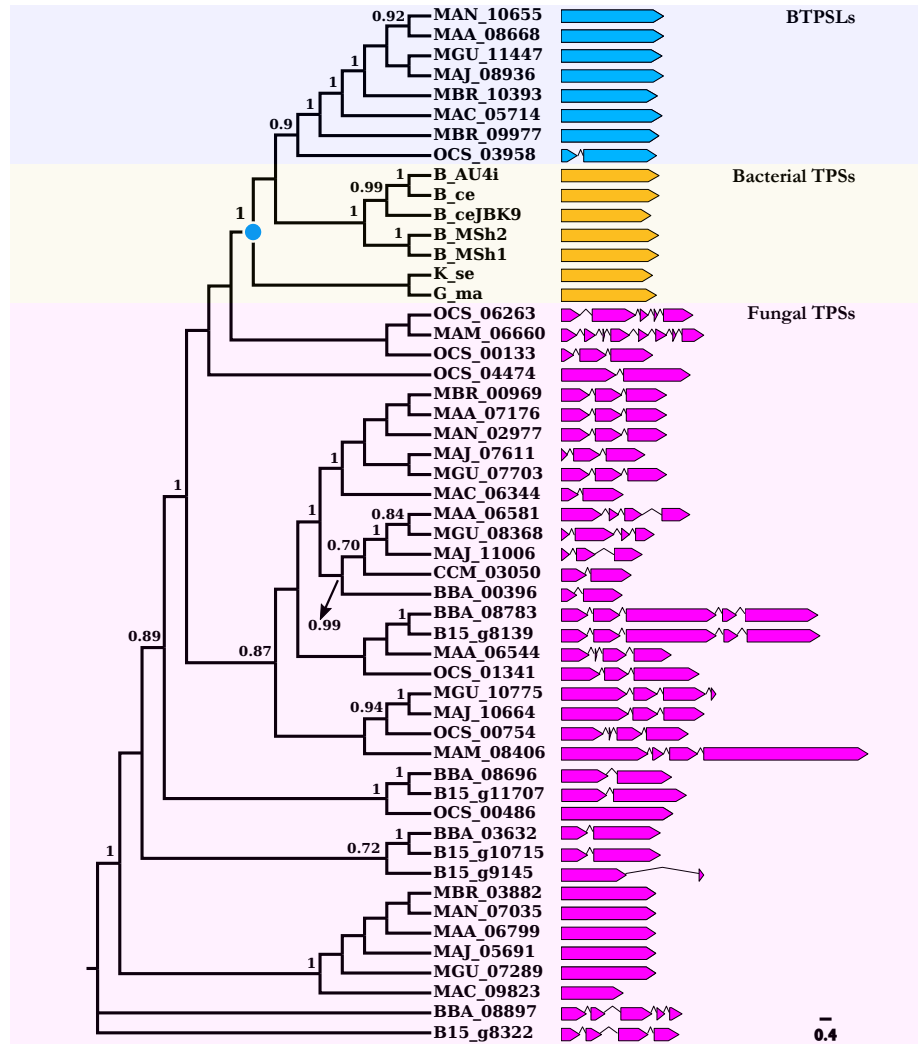


Figure 4.5: Maximum Likelihood phylogenetic tree and exon-intron structure of both types of TPS genes from entomopathogenic fungi and those bacterial TPS genes similar to the *BTPSL* genes in fungi. The phylogenetic tree was constructed using PhyML with GTR + I + G substitution model based on the codon alignment that is generated from the corresponding protein alignment. Bootstrap values greater than 70% are shown beside branches. The *BTPSL* genes (Table 4.1), typical fungal TPS genes (Table 4.2) and bacterial TPS genes (Table 4.3) are colored in cyan, magenta and orange, respectively. Exons and introns are represented by boxes with arrowheads and lines. Scale bars indicate the numbers of substitutions per site. Gene models are drawn to scale.

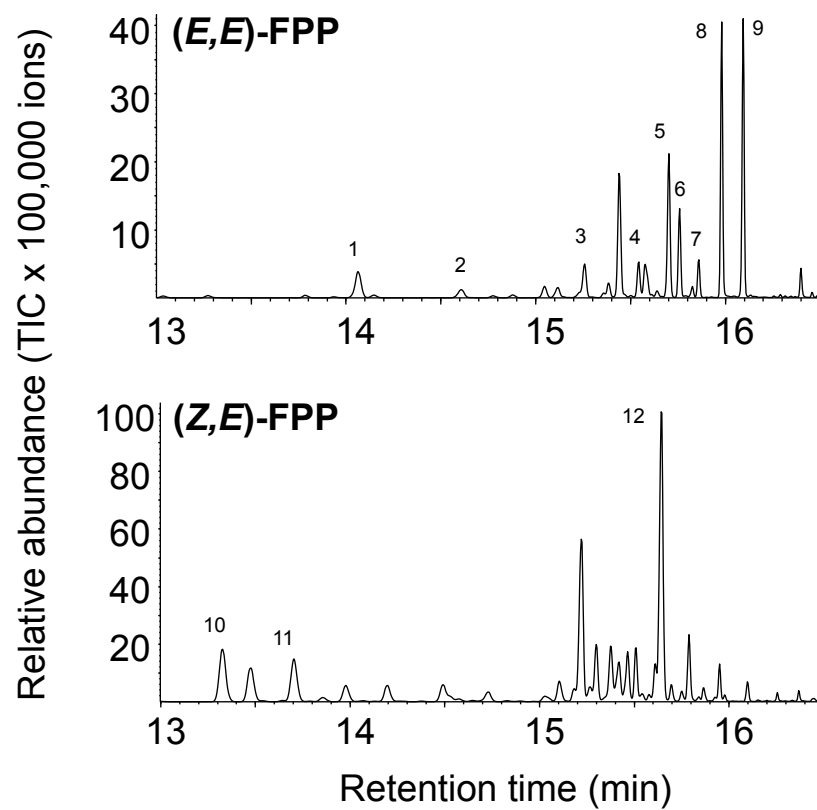


Figure 4.6: Biochemical activity of terpene synthase *MAA_08668* of putative horizontal gene transfer origin in *Metarhizium robertsii*. The gene was expressed in *Escherichia coli* and recombinant protein was incubated with the potential substrates (E,E)-farnesyl diphosphate (FPP) and (Z,E)-FPP. Enzyme products were analyzed using GC-MS. The total ion current (TIC) chromatograms are shown. 1, β -elemene*; 2, (E)- β -caryophyllene*; 3, epi-bicyclosesquiphellandrene; 4, cadi-1,4-diene; 5, γ -cadinene; 6, δ -cadinene; 7, α -cadinene; 8, nerolidol*; 9, epi- α -cadinol; 10, α -longipinene; 11, α -ylangene; 12, β -bisabolene*. Compounds marked with * were identified using authentic standards. All other compounds were tentatively identified using commercial mass spectra libraries. Unlabeled peaks represent unidentified sesquiterpenes.

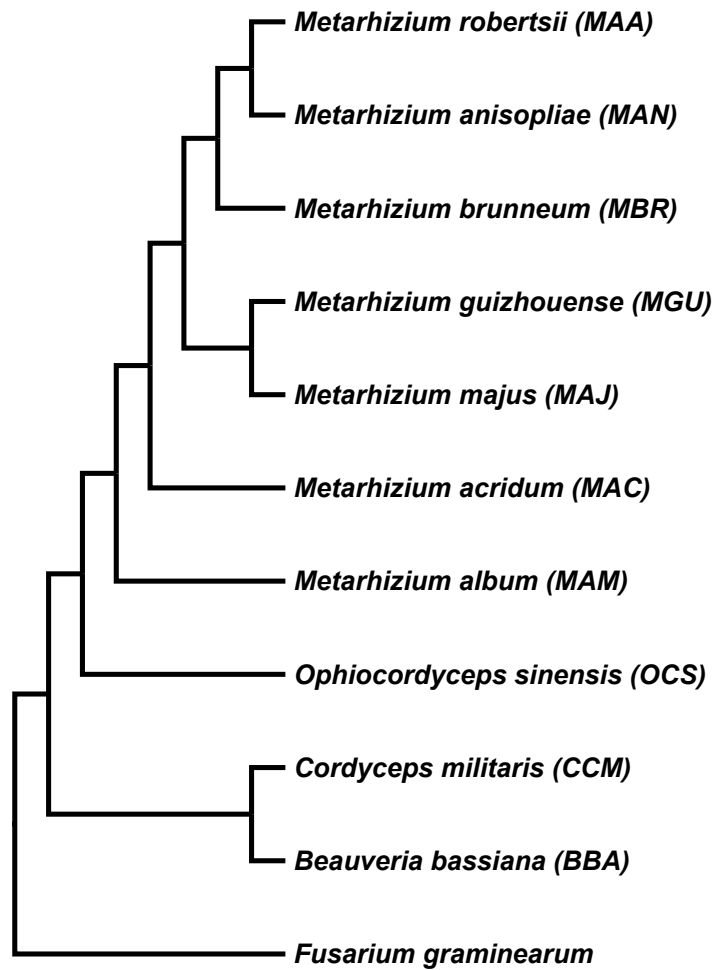


Figure 4.7: A species tree redrawn based on previous studies showing the evolutionary relationship of several entomopathogenic fungi. *Fusarium graminearum* is included as an outgroup.

Chapter 5

Identification, Characterization, and Evolution of Fungal Terpene Synthases

5.1 Abstract

Terpenes, one important class of fungi secondary metabolites, are the largest and most structurally diverse natural compounds. Fungi are known for producing a wide range of bioactive terpene products that play important roles in fungal survival and development. Many of them are also used by humans as medicines, such as antibiotics and anticancer drugs. Terpene synthases (TPSs) are the key enzymes to produce these terpenes. Recent advances in sequencing technologies have made available a large number of fungi genomes, so this provides a great opportunity to systematically investigate the distribution of fungal terpene synthase genes. In this chapter, four types of terpene synthase genes were identified from 519 fungi genomes. All of them are from higher fungi (Ascomycota and Basidiomycota). The two groups of fungi have evolved different sets of terpene synthases. The α and TRI5 types of terpene synthases are mainly found in species of Basidiomycota, while the bifunctional CPS/KS and the chimera diterpene synthases are present in Ascomycota. The results also suggest possible HGT between Ascomycota and Basidiomycota.

5.2 Introduction

During evolution, fungi have successfully adapted to an extraordinarily wide range of environments, from soil and human body to the deep-sea sediments and even underwater volcano. Survival in such a wide diversity of habitats needs employ various strategies for obtaining nutrients. Among these strategies, such as producing almost indestructible

spores and developing trapping structures by predatory fungi (Yang et al., 2007), the ability to produce vast diversity of natural secondary metabolites in response to complex, extreme and changing environments also ensure their survival and reproduction. Fungi use these nature products as chemical signals for communication (Brakhage and Schroeckh, 2011; Netzker et al., 2015), to protect them from predators. Many of these biologically active compounds are also used as beneficial drugs by humans, including antibiotics (Demain and Fang, 2000), plant growth hormones (gibberellins) (Bomke and Tudzynski, 2009) and anticancer drugs (Balde et al., 2010; Evidente et al., 2014).

Terpenoids are one class of secondary metabolites widely produced by fungi (Keller et al., 2005). They are derived from a single isoprene unit and are the most diverse class of natural compounds on earth. This huge diversity is partially achieved by the terpene synthases, key enzymes involved in the synthesis of terpenoids. Based on different substrate activation mechanisms, two classes of terpene synthases are recognized currently. The Class I terpene synthases, which has the α -domain, catalyzes an ionization-dependent cyclization of the substrate diphosphate group and contains two conserved motifs: DDxxD and NSE. The Class II terpene synthases contains a functional β -domain and catalyzes a protonation-dependent cyclization. The catalytic site of this type is a DxDD motif. Based on the types of terpenoids they produce, terpene synthases can be classified as isoprene synthases, monoterpene synthases, sesquiterpene synthases and diterpene synthases (Chen et al., 2011). In fungi, monoterpene synthases are rare and the first one, 1,8-cineole synthase, was recently reported from *Hypoxylon* sp. (Shaw et al., 2015). Sesquiterpene synthases, such as the trichodiene synthases, convert the

farnesyl pyrophosphate into different compounds and have the Class I terpene synthases signatures. Diterpene synthases that use geranylgeranyl pyrophosphate as substrates have both α -domain and β -domain. As opposed to majority plant diterpene synthases, fungal counterparts are believed to be bifunctional. The first diterpene synthase was identified from the Fungus *Phaeosphaeria* sp. L487 (Kawaide et al., 1997) and produces the ent-Kaurene, a precursor of Gibberellin. A special type of diterpene synthases, called chimeric fungal terpene synthases that can catalyze both chain elongation and cyclization reactions, was also existed in fungi (Matsuda et al., 2015; Qin et al., 2016; Toyomasu et al., 2009; Ye et al., 2015). Sequences of this type have terpene synthase domains and prenyltransferase domains.

Fungi terpene synthases are often found within biosynthetic clusters. For example, the Tri5 gene cluster found in *Fusarium* species has 12 genes including terpene synthase gene, the cytochrome P450 monooxygenase genes and the regulatory genes (Kimura et al., 2007; Proctor et al., 2009). The GA biosynthetic cluster in *Gibberella fujikuroi* contains six genes: a GGPP synthase, a bifunctional diterpene synthase and four cytochrome P450 monooxygenase genes (Tudzynski and Höltér, 1998). These gene clusters are key sources of novel high-value compounds (Cacho et al., 2014; Keller, 2015; Weber et al., 2015).

Although fungi are rich source of terpenoids, when compared to plant terpene synthases, our knowledge of fungal terpene synthases is very limited. So, identification and characterization of terpene synthase genes and biosynthetic gene clusters will pave the way to a deeper understanding of fungal terpene synthases. Recently, the JGI 1000 fungal genome project provides us a great data set to look at the fungal terpene

synthases. Currently, more than 500 fungal genome sequences are already publicly available through MycoCosm (<http://genome.jgi.doe.gov/programs/fungi>). In this chapter, four types of fungal terpene synthases are identified from 519 fungal genomes. Their distributions, sequence features and evolutionary relationships are investigated.

5.3 Results

5.3.1 Distribution of TPSs in Fungal Genomes

By mining 519 fungal genomes from the Joint Genome Institute (JGI), we identified 1753 class I terpene synthases (Figure 5.1 and Appendix A.2), 545 Trichodiene synthases, 230 bifunctional ent-copalyl diphosphate synthase/ent-kaurene synthase (CPS/KS) enzymes (all sequences of length less than 300 amino acids are eliminated) and 114 chimera diterpene synthases (sequences with both Terpene_synth_C and polyprenyl_synt domains). All four types of terpene synthases are found exclusively in Ascomycota and Basidiomycota (Figure 5.1). For any type, it is clear that there is no terpene synthase found in the basal lineages of fungi, such as Glomeromycota, Mucoromycotina and Zoopagomycotina. Terpene synthases identified in this study clearly cluster into four groups in the phylogenetic tree (Figure 5.2) according to the type, suggesting significant variation within the four types of Terpene synthases.

About 70% of the class I terpene synthases (1220 out of 1753) are present in Agaricomycotina, which is one of three major subphyla of the fungal division Basidiomycota

and produce mushrooms, but completely absent in Ustilaginomycotina, a group of non-mushroom forming fungi in the Basidiomycota and most of which are plant parasites. In another subphyla, Pucciniomycotina, only 7 terpene synthases are found from two species, *Atractiellales* sp. 95 and *Cronartium quercuum* f. *sp. fusiforme* G11. The remaining 30% are found in the largest subphylum of Ascomycota, Pezizomycotina, which includes almost all filamentous fungi. It appears that terpene synthases are present in all the eight studied classes of this subphylum and are evenly distributed across species. However, no terpene synthases are found from 40 species of Saccharomycotina and 7 species of Taphrinomycotina. Species of Saccharomycotina and Taphrinomycotina are yeasts and yeast-like fungi in the Ascomycota and they don't form fruiting bodies during the sexual phase of their life cycle, while species of Pezizomycotina are characterized by their fruiting bodies. It is clear that the presence or absence of terpene synthases in the species of Ascomycota depends on different morphological characters, such as whether possessing fruiting bodies. It also needs to point out that the class I terpene synthases found in Ascomycota and Basidiomycota cluster into two major groups according to the fungal phylogeny (Figure 5.3), although there are a small number of genes scattered over the two groups.

Similar to the α type of terpene synthases, more than half of Trichodiene synthases (359 out of 545) are identified from species of Agaricomycotina. This type is also completely absent in Ustilaginomycotina, although there are 16 genomes available. Two (*Tritirachium* sp. CBS 265.96 and *Septobasidium* sp. PNB30-8B) of 22 Pucciniomycotina species examined have Trichodiene synthase genes and the numbers of genes in the two

species are 18 and 9, respectively. The remaining are present in the following five classes of Ascomycota: Eurotiomycetes, Sordariomycetes, Dothideomycetes, Pezizomycetes and Leotiomyces. Although there is no gene identified from the other three classes, it not yet known whether this type is absent from species of these three classes with the consideration of limited available genomes. Compared to the α type, the phylogenetic relationship of Trichodiene synthases is less clear (Figure 5.4). To a certain extent, the phylogenetic analysis indicates the lineage-specific distribution of this type of genes (Ascomycota specific and Basidiomycota specific). However, there are more genes showing higher sequence similarity with those outside of their specific lineages. For example, more than half of Trichodiene synthases identified from species of Sordariomycetes cluster together with those in Agaricomycotina. In addition, all 18 Trichodiene synthase genes found in *Tritirachium sp.* CBS 265.96 reside on the branch of Ascomycota species. These incongruities indicate possible horizontal transfer events among species in these two phyla of Dikarya.

The bifunctional CPS/KSs are mainly found in three clades of the Ascomycota division: Eurotiomycetes, Dothideomycetes and Sordariomycetes. Although Agaricomycotina has the largest number of genomes available (136 out of 519), only 11 CPS/KSs are found from seven species. All of these seven species belong to class Agaricomycetes, but reside in three different orders: Agaricales, Boletales and Corticiales (Figure 5.5). These 11 CPS/KSs form a single group and cluster with members from the Ascomycota division with strong support in the phylogenetic tree (Figure 5.6). The multiple sequence alignment also indicates the high sequence similarity among these sequences and the conserved motifs

(Figure 5.7). This indicates the possible existence of horizontal gene transfer events from species of Ascomycota to species of Basidiomycota. The possibility that these transfers occur in the opposite direction also exists. If this is true, the CPS/KS gene has been lost in vast majority of Basidiomycota species during evolution.

The $\alpha\alpha 1$ type of diterpene synthases contains the smallest number of genes compared to the other three types. Its distribution is very similar to the distribution of CPS/KSs. Out of 136 available genomes in the Basidiomycota division, only two sequences are found from a single species, *Gymnopus luxurians*, which also has the largest number of CPS/KSs among the seven Basidiomycota species containing CPS/KSs. These two sequences are very similar to three members in Class Dothideomycetes (Figure 5.8 and Figure 5.9).

The distribution of these four types of terpene synthases among species is also investigated within each fungal lineage (Figure 5.10). Species containing TRI5 type terpene synthases are more likely to have α type (100% in Agaricomycotina, 83% in Dothideomycetes, 92% in Sordariomycetes, 84% in Eurotiomycetes, 50% in Pezizomycetes and 100% in Leotiomycetes). The chances of having CPS/KS and TRI5 types are 1%, 17%, 21% and 53% in Agaricomycotina, Dothideomycetes, Sordariomycetes and Eurotiomycetes, respectively. The chances of having CPS/KS and $\alpha\alpha 1$ types are 40%, 28% and 53% in Dothideomycetes, Sordariomycetes and Eurotiomycetes, respectively. The class Eurotiomycetes has the largest number of species (19) that have all four types of terpene synthases and the numbers are 1, 3 and 8 in Agaricomycotina, Dothideomycetes and Sordariomycetes, respectively.

5.3.2 Intron-Exon Structure

The mRNA sequence lengths and exon numbers of four types of terpene synthases were investigated (Figure 5.11, Table 5.1 and Table 5.2). The pairwise Wilcoxon rank sum test with Bonferroni's correction was used to test the mRNA length difference between any two types of terpene synthases and show that all four groups are significantly different from each other ($p < 2e-16$). The median mRNA lengths of the four types are, respectively, 1041 (α), 987 (TRI5), 2808 (CPS/KS) and 2142 ($\alpha\alpha1$).

The number of exons of all four types appears to be no unique pattern and ranges from a single exon to more than 10. The same test was also used to test the exon number difference between any two types and reveal that all, except the exon numbers between TRI5 and CPS/KS ($p = 1$) and those between α and $\alpha\alpha1$ ($p = 0.055$), are significantly different at 0.01 significance level or above. The median exon numbers of the four types are 5 (α), 4 (TRI5), 4 (CPS/KS) and 5.5 ($\alpha\alpha1$), respectively. As shown in Figure 5.11A, the number of exons is not correlated with the length of the gene (Pearson correlation coefficient $r = 0.02$). However, genes of CPS/KS type tend to have long exon sequences as indicated by its longest average gene length and low exon number, while the α type tend to have short exon sequences.

5.3.3 Motif Analysis

Potential functional and structural motifs that are important for substrate binding and catalysis were checked for each of the four types of terpene synthases. The α type

contains two aspartate-rich motifs, D(D,E)xxD and NDxxS(Y,F)xxE (Figure 5.12), which are signatures of class I terpene synthases. The two motifs are believed to form a trinuclear metal cluster that trigger the Mg^{2+} dependent ionization of isoprenoid diphosphate group. Terpene synthases of this type are putative sesquiterpene synthases.

The Trichodiene synthases here have no typical aspartate-rich motif DDxxD, although the DD motif is obvious. It is possible that a small number of sequences possess the motif, but most of them don't. The additional metal cofactor binding motif, NDxxSFYKE, is observed in this type.

The CPS/KS type sequences are diterpene synthases and have both α and β domains. Four conserved motifs, D(D,E)xxE, DxDD, DExxE and NDxGSxxRDxxE, are observed. The DXDD motif is a signature of class II terpene synthases and is responsible for catalyzing the protonation dependent cyclization reaction (Prisic et al., 2007). The DExxE motif is similar to the DDXXD motif, but is more common observed in fungi (Fischer et al., 2015). As opposed to most plant CPS/KS sequences that have only one active domain (CPS-type or KS-type), the fungal CPS/KSs are believed to be bifunctional enzymes, which synthesize ent-kaurene, a precursor for gibberellins (GAs), from geranylgeranyl diphosphate (GGDP) using single enzymes (Schmidt-Dannert, 2015).

As expected, the chimera diterpene synthases, $\alpha\alpha 1$ type, contain two DDxxD motifs, one from the terpene synthase domain and the other one from the prenyltransferase domain. Both these two DDxxD motifs have KS activities and are expected to be involved in binding the substrate diphosphate group via an Mg^{2+} . The third motif, NDxxxxxxE, is also observed in the terpene synthase domain. All these chimera diterpene synthases

have their terpene domain at the N terminus and the prenyltransferase domain at the C terminus, except the one AlAl1_EUR_RTHQ_7399, of which the two domains are overlapped mostly. This conserved domain architecture indicates a common origin and conserved functions performed by these unusual multifunctional enzymes.

5.4 Discussion

Like plants, fungi also produce a wide range of diverse terpenoids to help them adapt and survive in different ecological niches. Current research on fungal terpenome is mainly limited to the phylum Ascomycota (Quin et al., 2014). The results obtained by mining such a wide range of fungi species, in particular for species in the phylum Basidiomycota, greatly expanded our knowledge and understanding of fungal terpenome. Our results show that the phyla Ascomycota and Basidiomycota have evolved different sets of terpene synthases and both phyla use α and TRI5 types as their predominant terpene synthases. For species of Basidiomycota, the α and TRI5 types of terpene synthases are commonly encoded in their genomes, while the two types of diterpene synthases are almost completely lacking. The α and TRI5 types are putative sesquiterpene synthases and expected to produce various sesquiterpenoids from farnesyl diphosphate (FPP). In contrast, species of Ascomycota encode almost all the diterpene synthases identified in this study, although they also encode about 30% of α and TRI5 types terpene synthases. It seems that species of Ascomycota produce a wider range of terpene natural products than species of Basidiomycota.

The seemingly lack of diterpene synthases in Basidiomycota doesn't mean that they don't exist. Over 30000 estimated species (Morrow and Fraser, 2009) in this phylum and only 174 species are investigated in this study and majority of them are not selected for the discovery of terpene synthases (Schmidt-Dannert, 2015). So, it is possible that more species in this phylum contain diterpene synthases. It is also possible that our current search method is not power enough to discover these genes and is also limited by our current knowledge of fungal diterpene synthases. It is worth mentioning that a HMMER search using known terpene synthase domains from Pfam doesn't pick up any fungal CPS/KSs because of lacking these domains as opposed to their plant counterparts. That's why we used four known CPS/KSs as blast queries (one from a species of Ascomycota, the other three from plant species). About 75% of CPS/KSs found in this study can be found by all these four sequences, while the rest 25% was exclusively found by the fungal sequence (Figure 5.13). In view of the low sequence similarity shared by terpene synthases and no CPS/KSs from Basidiomycota used as queries, Basidiomycota species probably have their CPS/KSs that show little sequence similarity to those in species of Ascomycota. Another possibility is that Basidiomycota acquired CPS/KSs from Ascomycota via horizontal gene transfer, which is also speculated by a recent study (Fischer et al., 2015).

Currently, five $\alpha\alpha 1$ type terpene synthases have been characterized. Three of them (Chiba et al., 2013; Qin et al., 2016; Ye et al., 2015) are from eurotiomycetes fungi and the other two from a species of Sordariomycetes (Toyomasu et al., 2008, 2007). This is consistent with our results that the vast majority of $\alpha\alpha 1$ type sequences are found from Sordariomycetes, Eurotiomycetes and Dothideomycetes. The close relationship between

α type and $\alpha\alpha 1$ type sequences in the phylogenetic tree indicates that the $\alpha\alpha 1$ type actually is a result of gene fusion. Terpene synthases might also fuse with other genes, such as the cytochrome P450 monooxygenase genes. Sequences with this architecture are observed in this study. So, more types of terpene synthases are waiting to be found.

We didn't find any terpene synthases in species of early diverging fungal lineages. There are some short sequences have been found, but at current stage, they are removed for further analysis because of low quality. However, we can't rule out the possibility that species belong to this group may contain terpene synthases.

5.5 Methods

5.5.1 Genome Mining of Fungal Terpene Synthases

519 fungal genomes were obtained from JGI (<http://genome.jgi.doe.gov/programs/fungi/index.jsf>, date last accessed October 2, 2015). Putative terpene synthases were identified by a search of all predicted amino acid sequences with HMMER against the Pfam database (<http://pfam.xfam.org>, version 28.0). Sequences with TRI5 domain (PF06330) as the best matching domain are classified as Trichodiene synthases, while sequences with Terpene_synth_C (PF03936) domain but without polyprenyl_synt (PF00348) domain are classified as class I terpene synthases. For sequences with both Terpene_synth_C and polyprenyl_synt domains, we named them as chimera diterpene synthases.

In order to recover the sequences with both ent-copalyl diphosphate synthase/ent-kaurene synthase (CPS/KS) activities, selected known CPS/KS sequences (SmCPSKS

from *Sphaceloma manihoticola*, GI: 197724597 (Bömke et al., 2008); PpCPSKS from *Physcomitrella patens*, GI: 146325986 (Hayashi et al., 2006); AtGA1 and AtGA2 from *Arabidopsis thaliana*, GI: 15235504 and 15235504 (Mayer et al., 1999), respectively) were used as blastp queries to identify CPS/KS homologs in the available JGI fungal proteins (e-value cutoff of 10^{-5}).

5.5.2 Multiple Sequence Alignments and Phylogenetic Inference

All multiple sequence alignments were made using MAFFT in a highly accurate setting (L-INS-i) with 1000 iterations of improvement. Maximum-likelihood phylogenetic trees were inferred with FastTree (version 2.1.7) using a more accuracy setting (-spr 4 -mlacc 2 -slownni). All trees were rendered using FigTree (version 1.4.2).

5.6 Bibliography

- Balde, E. S., Andolfi, A., Bruyere, C., Cimmino, A., Lamoral-Theys, D., Vurro, M., Damme, M. V., Altomare, C., Mathieu, V., Kiss, R., and Evidente, A. (2010). Investigations of fungal secondary metabolites with potential anticancer activity. *Journal of Natural Products*, 73(5):969–71. [149](#)
- Bomke, C. and Tudzynski, B. (2009). Diversity, regulation, and evolution of the gibberellin biosynthetic pathway in fungi compared to plants and bacteria. *Phytochemistry*, 70(15-16):1876–93. [149](#)
- Brakhage, A. A. and Schroeckh, V. (2011). Fungal secondary metabolites - strategies to activate silent gene clusters. *Fungal Genetics and Biology*, 48(1):15–22. [149](#)
- Bömke, C., Rojas, M. C., Gong, F., Hedden, P., and Tudzynski, B. (2008). Isolation and characterization of the gibberellin biosynthetic gene cluster in *sphaceloma manihoticola*. *Applied and Environmental Microbiology*, 74(17):5325–5339. [160](#)
- Cacho, R. A., Tang, Y., and Chooi, Y. H. (2014). Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi. *Frontiers in Microbiology*, 5:774. [150](#)
- Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant Journal*, 66(1):212–29. [149](#)

- Chiba, R., Minami, A., Gomi, K., and Oikawa, H. (2013). Identification of ophiobolin f synthase by a genome mining approach: A sesterterpene synthase from *aspergillus clavatus*. *Organic Letters*, 15(3):594–597. [158](#)
- Demain, A. L. and Fang, A. (2000). The natural functions of secondary metabolites. *Advances in Biochemical Engineering/Biotechnology*, 69:1–39. [149](#)
- Evidente, A., Kornienko, A., Cimmino, A., Andolfi, A., Lefranc, F., Mathieu, V., and Kiss, R. (2014). Fungal metabolites with anticancer activity. *Natural Product Reports*, 31(5):617–27. [149](#)
- Fischer, M. J., Rustenhloz, C., Leh-Louis, V., and Perriere, G. (2015). Molecular and functional evolution of the fungal diterpene synthase genes. *BMC Microbiology*, 15:221. [156](#), [158](#)
- Hayashi, K. i., Kawaide, H., Notomi, M., Sakigi, Y., Matsuo, A., and Nozaki, H. (2006). Identification and functional analysis of bifunctional ent-kaurene synthase from the moss *physcomitrella patens*. *FEBS Letters*, 580(26):6175–6181. [160](#)
- Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., Huhndorf, S., James, T., Kirk, P. M., Lucking, R., Thorsten Lumbsch, H., Lutzoni, F., Matheny, P. B., McLaughlin, D. J., Powell, M. J., Redhead, S., Schoch, C. L., Spatafora, J. W., Stalpers, J. A., Vilgalys, R., Aime, M. C., Aptroot, A., Bauer, R., Begerow, D., Benny, G. L., Castlebury, L. A., Crous, P. W., Dai, Y. C., Gams, W., Geiser, D. M., Griffith, G. W., Gueidan, C., Hawksworth, D. L., Hestmark, G., Hosaka, K., Humber, R. A., Hyde, K. D.,

- Ironside, J. E., Koljalg, U., Kurtzman, C. P., Larsson, K. H., Lichtwardt, R., Longcore, J., Miadlikowska, J., Miller, A., Moncalvo, J. M., Mozley-Standridge, S., Oberwinkler, F., Parmasto, E., Reeb, V., Rogers, J. D., Roux, C., Ryvarde, L., Sampaio, J. P., Schussler, A., Sugiyama, J., Thorn, R. G., Tibell, L., Untereiner, W. A., Walker, C., Wang, Z., Weir, A., Weiss, M., White, M. M., Winka, K., Yao, Y. J., and Zhang, N. (2007). A higher-level phylogenetic classification of the fungi. *Mycological Research*, 111(Pt 5):509–47. [170](#)
- Kawaide, H., Imai, R., Sassa, T., and Kamiya, Y. (1997). ent-kaurene synthase from the fungus *phaeosphaeria* sp. 1487: cDNA isolation, characterization, and bacterial expression of a bifunctional diterpene cyclase in fungal gibberellin biosynthesis. *Journal of Biological Chemistry*, 272(35):21706–21712. [150](#)
- Keller, N. P. (2015). Translating biosynthetic gene clusters into fungal armor and weaponry. *Nature Chemical Biology*, 11(9):671–7. [150](#)
- Keller, N. P., Turner, G., and Bennett, J. W. (2005). Fungal secondary metabolism - from biochemistry to genomics. *Nature Reviews. Microbiology*, 3(12):937–947. [149](#)
- Kimura, M., Tokai, T., Takahashi-Ando, N., Ohsato, S., and Fujimura, M. (2007). Molecular and genetic studies of fusarium trichothecene biosynthesis: Pathways, genes, and evolution. *Bioscience, Biotechnology and Biochemistry*, 71(9):2105–2123. [150](#)
- Matsuda, Y., Mitsunashi, T., Quan, Z., and Abe, I. (2015). Molecular basis for stellate acid biosynthesis: A genome mining approach for discovery of sesterterpene synthases. *Organic Letters*, 17(18):4644–7. [150](#)

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K. D., Terryn, N., Harris, B., Ansorge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Muller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-Alonso, M., Boutry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P., Langham, S. A., McCullagh, B., Bilham, L., Robben, J., Van der Schueren, J., Grymonprez, B., Chuang, Y. J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiaens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., Ramsperger, U., Hilbert, H., Braun, M., Holzer, E., Brandt, A., Peters, S., van Staveren, M., Dirske, W., Mooijman, P., Klein Lankhorst, R., Rose, M., Hauf, J., Kotter, P., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Van den Daele, H., De Keyser, A., Buysschaert, C., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Rogers, J., Cronin, A., Quail, M., Bray-Allen, S., Clark, L., Doggett, J., Hall, S., Kay, M., Lennard, N., McLay, K., Mayes, R., Pettett, A., Rajandream, M. A., Lyne, M., Benes, V., Rechmann, S., Borkova, D., Blocker, H., Scharfe, M., Grimm, M., Lohnert, T. H., Dose, S., de Haan, M., Maarse, A., Schafer, M., et al. (1999). Sequence and analysis of chromosome 4 of the plant *arabidopsis thaliana*. *Nature*, 402(6763):769–77. [160](#)

Morrow, C. A. and Fraser, J. A. (2009). Sexual reproduction and dimorphism in the pathogenic basidiomycetes. *FEMS Yeast Research*, 9(2):161–77. [158](#)

Netzker, T., Fischer, J., Weber, J., Mattern, D. J., Konig, C. C., Valiante, V., Schroeckh, V., and Brakhage, A. A. (2015). Microbial communication leading to the activation of silent

- fungus secondary metabolite gene clusters. *Frontiers in Microbiology*, 6:299. [149](#)
- Prisic, S., Xu, J., Coates, R. M., and Peters, R. J. (2007). Probing the role of the dxdd motif in class ii diterpene cyclases. *Chembiochem*, 8(8):869–74. [156](#)
- Proctor, R. H., McCormick, S. P., Alexander, N. J., and Desjardins, A. E. (2009). Evidence that a secondary metabolic biosynthetic gene cluster has grown by gene relocation during evolution of the filamentous fungus *Fusarium*. *Molecular Microbiology*, 74(5):1128–1142. [150](#)
- Qin, B., Matsuda, Y., Mori, T., Okada, M., Quan, Z., Mitsuhashi, T., Wakimoto, T., and Abe, I. (2016). An unusual chimeric diterpene synthase from *Emericella variecolor* and its functional conversion into a sesterterpene synthase by domain swapping. *Angewandte Chemie. International Edition in English*, 55(5):1658–61. [150](#), [158](#)
- Quin, M. B., Flynn, C. M., and Schmidt-Dannert, C. (2014). Traversing the fungal terpenome. *Natural Product Reports*, 31(10):1449–73. [157](#)
- Schmidt-Dannert, C. (2015). Biosynthesis of terpenoid natural products in fungi. *Advances in Biochemical Engineering/Biotechnology*, 148:19–61. [156](#), [158](#)
- Shaw, J. J., Berbasova, T., Sasaki, T., Jefferson-George, K., Spakowicz, D. J., Dunican, B. F., Portero, C. E., Narvaez-Trujillo, A., and Strobel, S. A. (2015). Identification of a fungal 1,8-cineole synthase from *Hypoxylon* sp. with specificity determinants in common with the plant synthases. *Journal of Biological Chemistry*, 290(13):8511–26. [149](#)

- Toyomasu, T., Kaneko, A., Tokiwano, T., Kanno, Y., Kanno, Y., Niida, R., Miura, S., Nishioka, T., Ikeda, C., Mitsuhashi, W., Daiiri, T., Kawano, T., Oikawa, H., Kato, N., and Sassa, T. (2009). Biosynthetic gene-based secondary metabolite screening: A new diterpene, methyl phomopsenonate, from the fungus *phomopsis amygdali*. *Journal of Organic Chemistry*, 74(4):1541–1548. [150](#)
- Toyomasu, T., Niida, R., Kenmoku, H., Kanno, Y., Miura, S., Nakano, C., Shiono, Y., Mitsuhashi, W., Toshima, H., Oikawa, H., Hoshino, T., Daiiri, T., Kato, N., and Sassa, T. (2008). Identification of diterpene biosynthetic gene clusters and functional analysis of labdane-related diterpene cyclases in *phomopsis amygdali*. *Bioscience, Biotechnology and Biochemistry*, 72(4):1038–1047. [158](#)
- Toyomasu, T., Tsukahara, M., Kaneko, A., Niida, R., Mitsuhashi, W., Daiiri, T., Kato, N., and Sassa, T. (2007). Fusicoccins are biosynthesized by an unusual chimera diterpene synthase in fungi. *Proceedings of the National academy of Sciences of the United States of America*, 104(9):3084–3088. [158](#)
- Tudzynski, B. and Hölter, K. (1998). Gibberellin biosynthetic pathway in *Gibberella fujikuroi*: Evidence for a gene cluster. *Fungal Genetics and Biology*, 25(3):157–170. [150](#)
- Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Brucoleri, R., Lee, S. Y., Fischbach, M. A., Muller, R., Wohlleben, W., Breitling, R., Takano, E., and Medema, M. H. (2015). antismash 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, 43(W1):W237–43. [150](#)

- Yang, Y., Yang, E., An, Z., and Liu, X. (2007). Evolution of nematode-trapping cells of predatory fungi of the orbiiliaceae based on evidence from rRNA-encoding DNA and multiprotein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 104(20):8379–84. [149](#)
- Ye, Y., Minami, A., Mandi, A., Liu, C., Taniguchi, T., Kuzuyama, T., Monde, K., Gomi, K., and Oikawa, H. (2015). Genome mining for sesterterpenes using bifunctional terpene synthases reveals a unified intermediate of di/sesterterpenes. *Journal of the American Chemical Society*, 137(36):11846–53. [150](#), [158](#)

5.7 Appendix

Table 5.1: Summary statistics of exon numbers of four TPS types.

Type	Count	Mean	Median	St. Dev.	Min	Max
α	1753	5.08	5	2.12	1	14
TRI5	545	3.90	4	1.89	1	10
CPS/KS	230	4.04	4	1.70	1	13
$\alpha\alpha 1$	114	6.05	5.5	2.90	1	11

Table 5.2: Summary statistics of mRNA lengths of four TPS types.

Type	Count	Mean	Median	St. Dev.	Min	Max
α	1753	1071.03	1041	132.78	903	3117
TRI5	545	1009.87	987	106.01	900	2499
CPS/KS	230	2581.29	2808	544.81	948	4269
$\alpha\alpha 1$	114	2112.45	2142	148.11	1605	2418

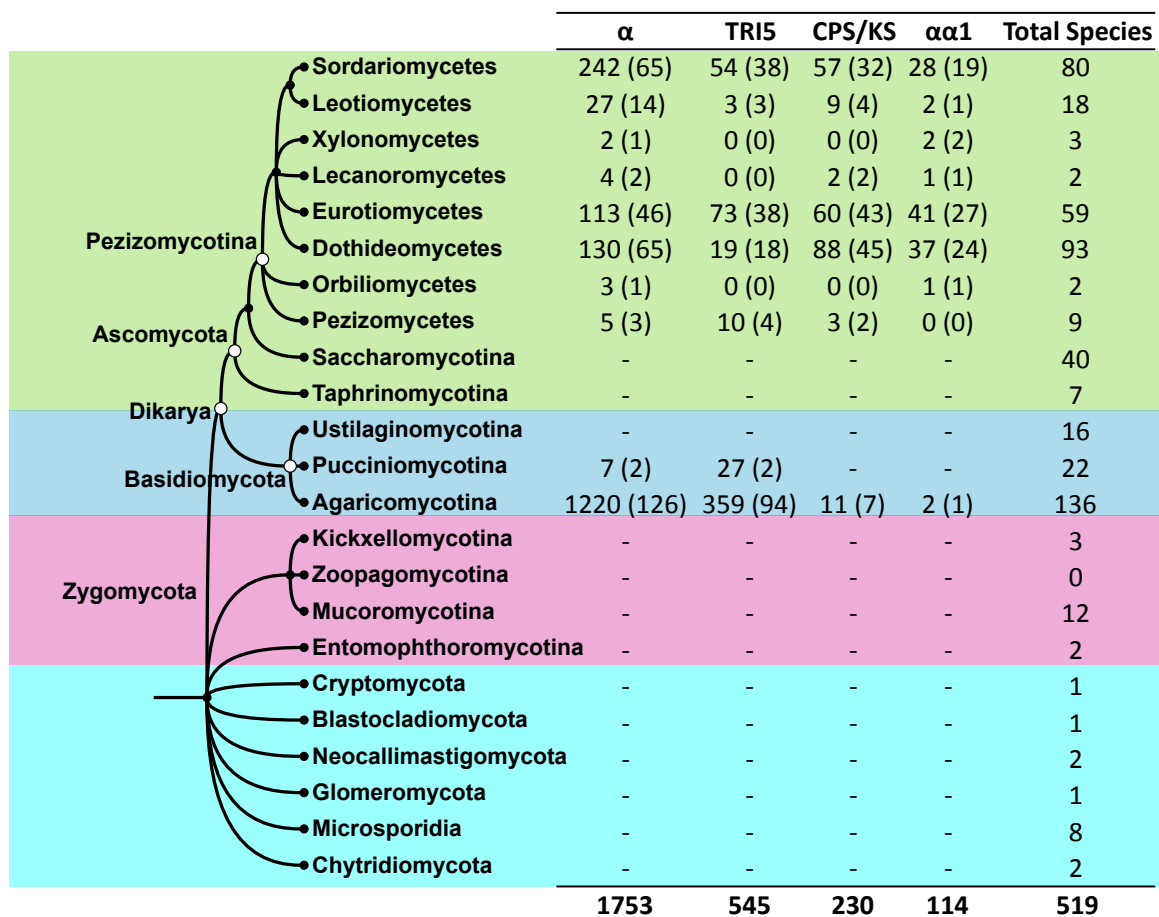


Figure 5.1: Phylogenetic tree of major fungal lineages and corresponding number of terpene synthases in each of the four types. The numbers of species is shown in parentheses. The depicted tree is approximately based on paper (Hibbett et al., 2007). For the full terpene synthase gene distribution for each of 519 fungal species analyzed, see Appendix A.2.

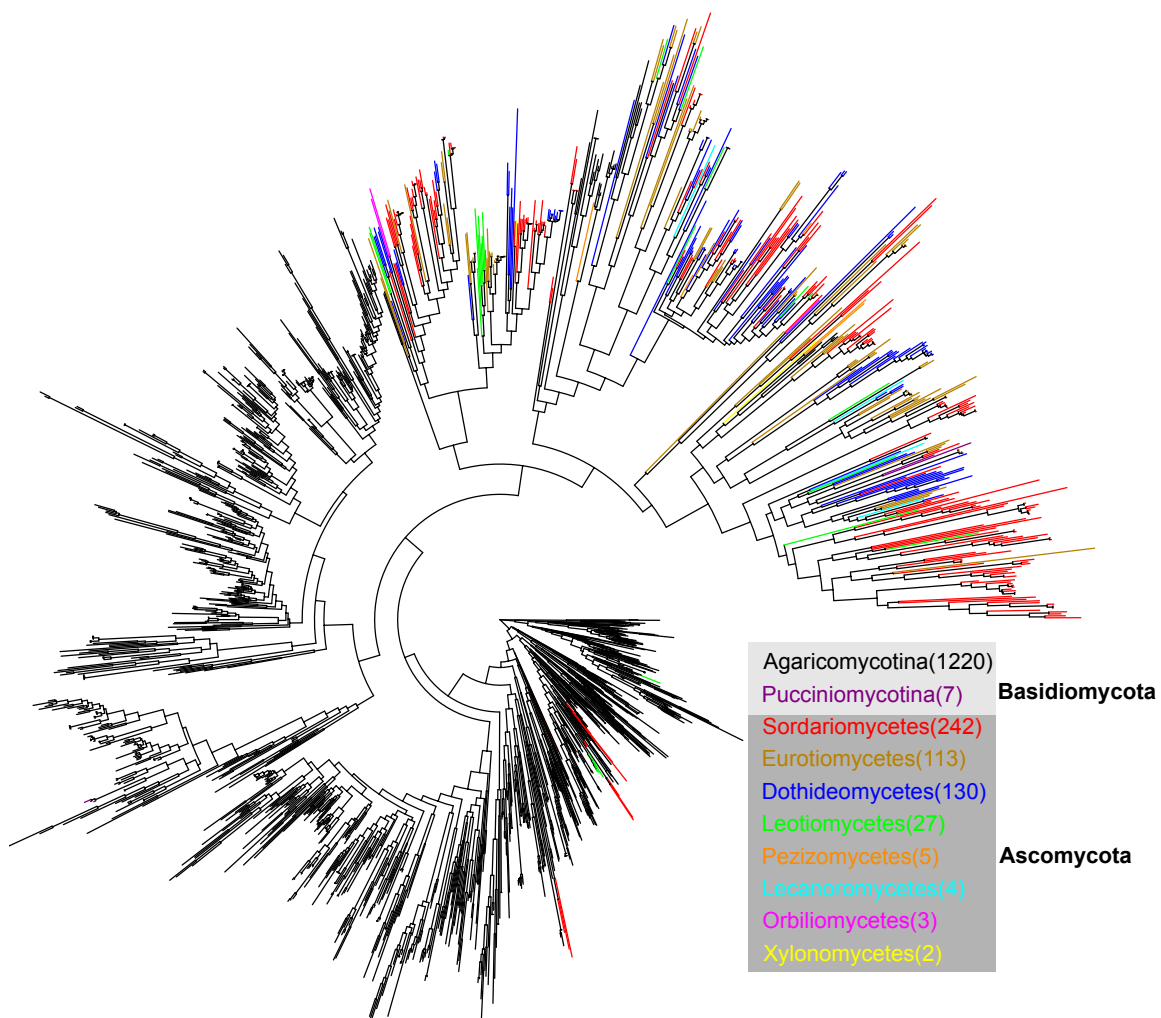


Figure 5.3: Maximum likelihood phylogeny of 1753 α type terpene synthases.

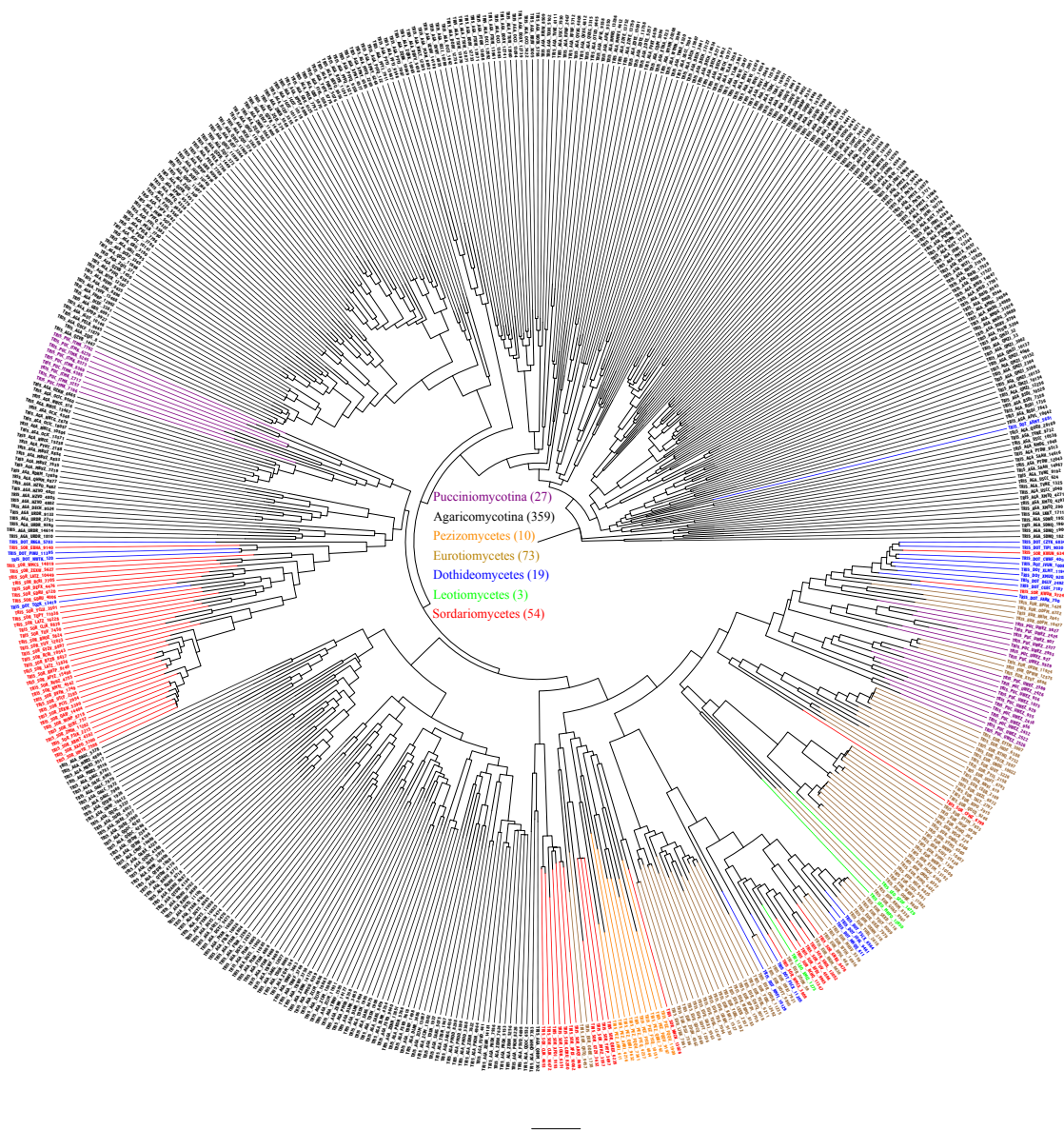


Figure 5.4: Maximum likelihood phylogeny of 545 Trichodiene synthases.

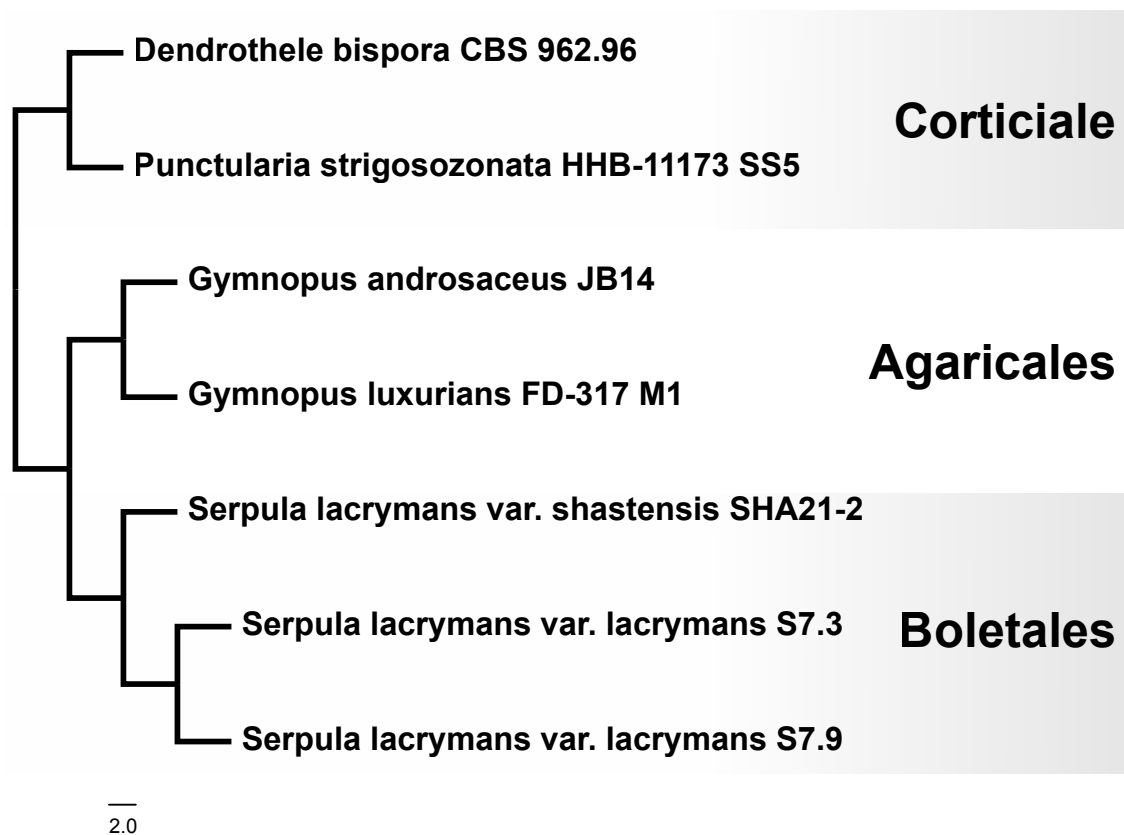


Figure 5.5: Phylogenetic relationship among 7 Basidiomycota species containing bifunctional CPS/KSs.

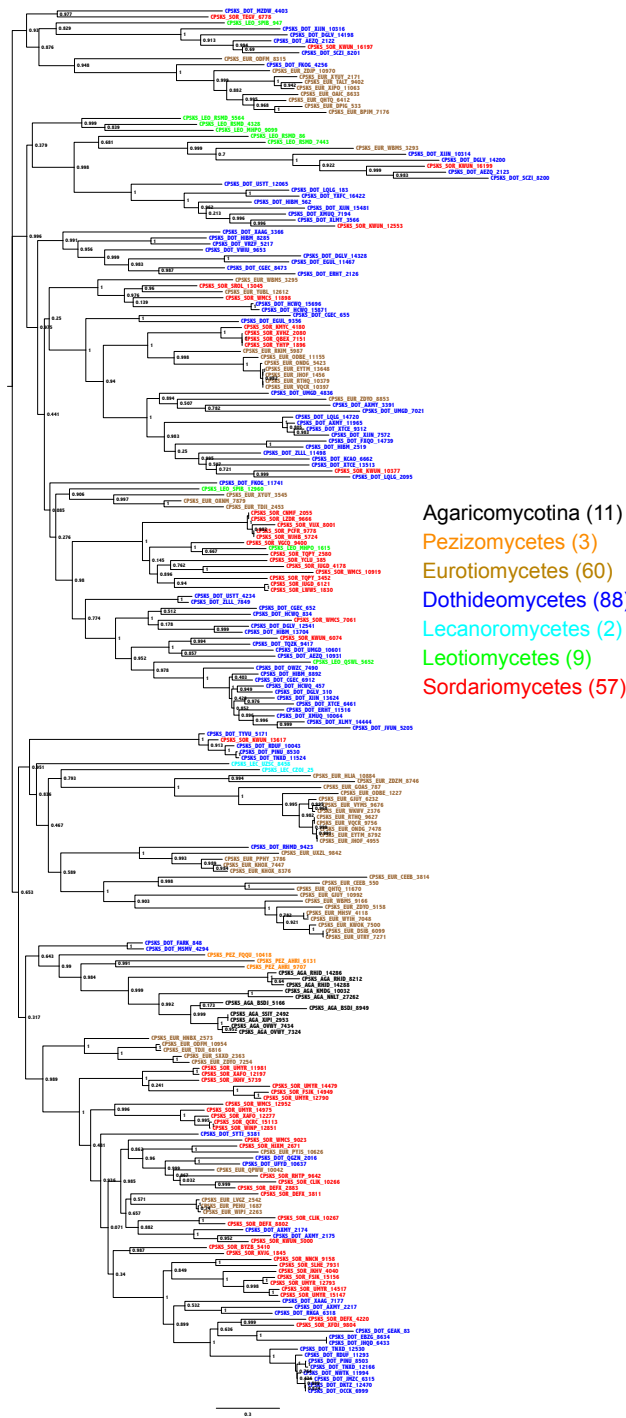


Figure 5.6: Maximum likelihood phylogeny of 230 bifunctional CPS/KSs.

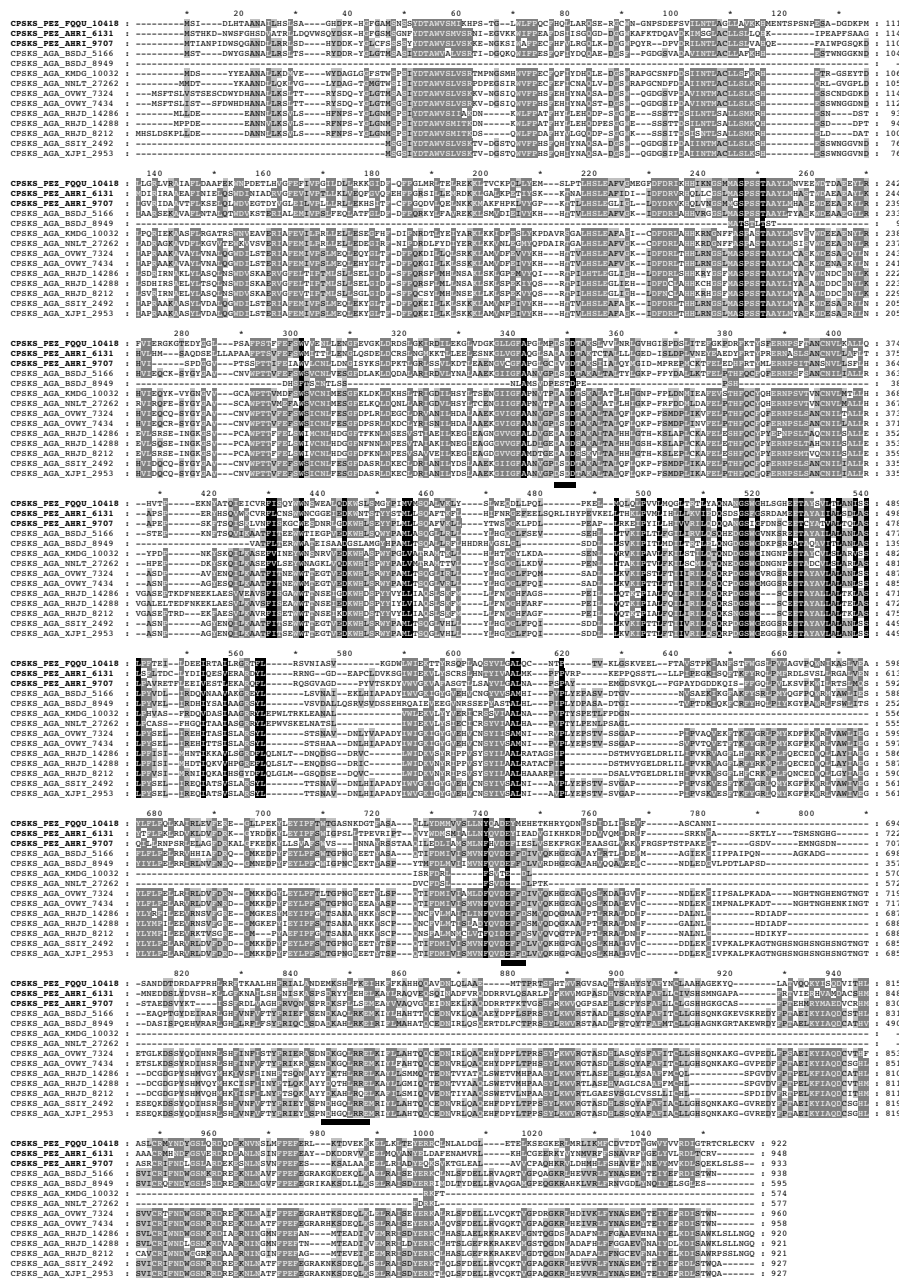


Figure 5.7: Multiple sequence alignment of 11 Basidiomycota CPS/KSs and 3 Ascomycota CPS/KSs.

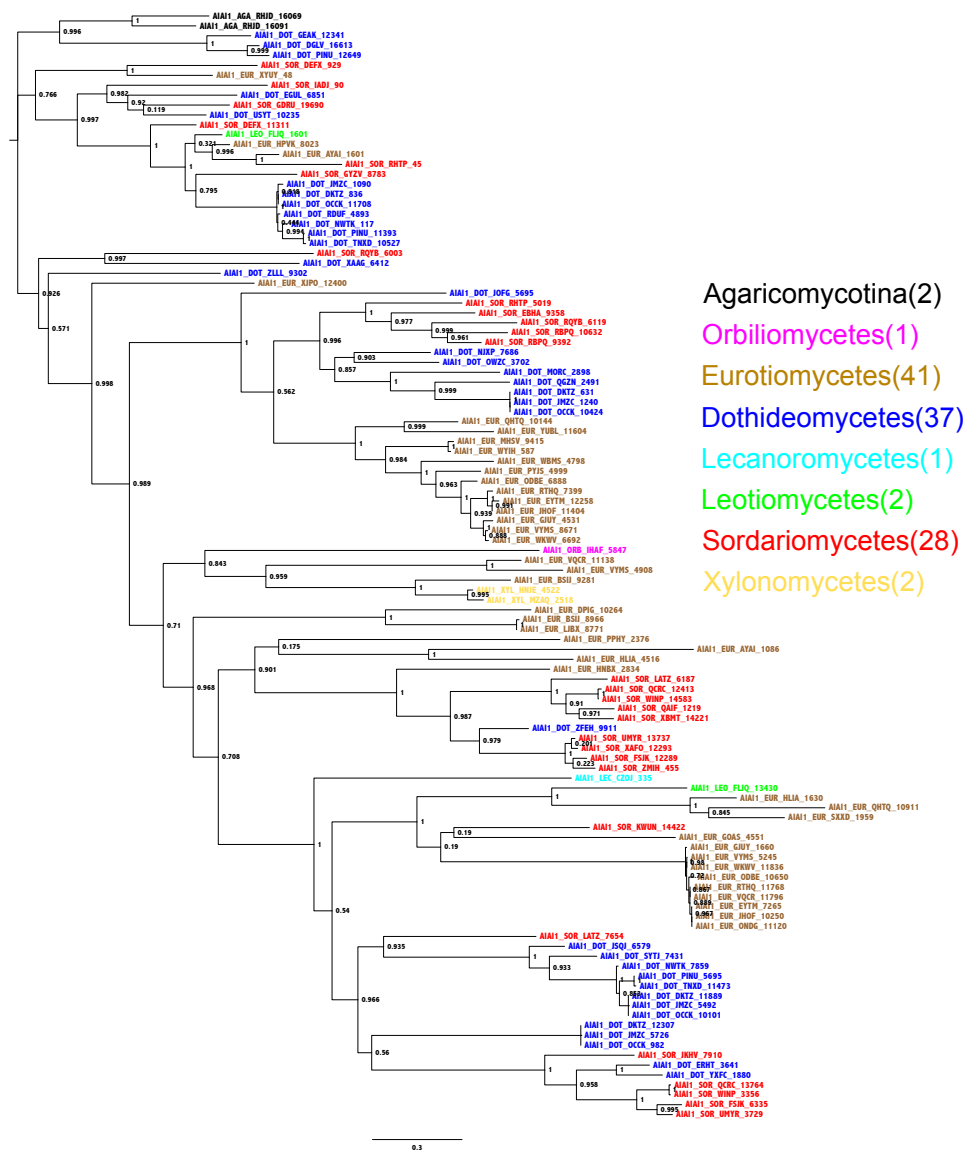


Figure 5.8: Maximum likelihood phylogeny of 114 α 1 type terpene synthases.

```

Ala11_AGA_RHJD_16069 : MELLYSTIVDPSEYPSGGLCGIDLRKSNFTWLEDRGAIIRAQODWRRKYSFQFEFRGTLGPEYSELVLLPECPRLRLVLYANEFAPL : 90
Ala11_AGA_RHJD_16091 : MEFOYSTVIDPSTYDTGLCGIDLRKSNFTWLEDRGAIIRAQODWRRKYSFQFEFRGTLGPEYSELVLLPECPRLRLVLYANEFAPL : 90
Ala11_DOT_GEAK_12341 : MIQFSTIVDPKTYDNEGLSKGIDLRKSNFTWLEDRGAIIRAQODWRRKYSFQFEFRGTLGPEYSELVLLPECPRLRLVLYANEFAPL : 90
Ala11_DOT_DGLV_16613 : MKQFSTIVDPATYDNEGLSKGIDLRKSNFTWLEDRGAIIRAQODWRRKYSFQFEFRGTLGPEYSELVLLPECPRLRLVLYANEFAPL : 90
Ala11_DOT_PINU_12649 : MRQFSTIVDLAAYDNEGLSKGIDLRKSNFTWLEDRGAIIRAQODWRRKYSFQFEFRGTLGPEYSELVLLPECPRLRLVLYANEFAPL : 90
M q5ST66dp YD eGL GIDLR4nNFT EDRGAIIRAQ DW 4 6 P f4gtLG 55f63v 6PEC P RLE66sYAnEFaF6

Ala11_AGA_RHJD_16069 : YDNFIYFVDKEQSTIEYDITGOAFLECARIGKIFTOITDAETKRAGRRKMSOMVLEMTADREGAIAIMRSWVGFAEAS---NHKEFA : 177
Ala11_AGA_RHJD_16091 : LDDVINATDEQGTIESNDMMQAFLEVOVGKI---TKNDAQTKREGRRKIOSLDLEMFSDIRERAIKAWAEFAEVGSGROHHEFA : 178
Ala11_DOT_GEAK_12341 : YDDATELETEDNMNSENDKMGFLTDALSSRP-PKLDLS---SGKMRILITOMVSEMIAIDKKCAVVTMRWSEFLRVGSSROHGTIET : 175
Ala11_DOT_DGLV_16613 : YDDDELDNENNTSAENDKLGIFLACTOGWSS-LOQSS---SGKTRILKOLFSEMVEIDKCAIVTMKAWAEFLRVGSSROHGTIET : 175
Ala11_DOT_PINU_12649 : YDDDELDTDDTSAENDKLGIFLAPESRSP-SQOQSS---SGKTRILKOLFSEMVAVDKCAITRAMKAWAEFLRVGSSROHGTIET : 175
ydl e E 1 6 FL g d GK 46 Q6 EM 6D4e A6 64aW eF gS rqh F

Ala11_AGA_RHJD_16069 : TLEDEYLPFRILNAGAVVLMQFLFLGMNLKIPENSKDKCYLMOALFVLAALNDLCSEWSEYIAKNCQDVHINALNVLRLRYNTDVP : 267
Ala11_AGA_RHJD_16091 : TLEDEYLPYRIVDAGHALWYTFITFGMGUNIPQWREKRCDELLOSATATVLENDLRSWQBYATASNHQSHVYNALRVLMRBNIGQ : 268
Ala11_DOT_GEAK_12341 : RLEDYLPYRIVDVGEMRWGVVTFGMALHIPDHMEACHRLMEPAVAVGLANDVSWPKERDNOKSGRSHIINAVWILMOBKGLSEEC : 265
Ala11_DOT_DGLV_16613 : RLEDYLPYRIVDVGEMRWGVVTFGMALHIPDHMEACHRLMEPAVAVGLANDVSWPKERDNOKSGRSHIINAVWILMOBKGLSEEC : 265
Ala11_DOT_PINU_12649 : RLEDYLPYRIVDVGEMRWGVVTFGMALHIPDHMEACHRLMEPAVAVGLANDVSWPKERDNOKSGRSHIINAVWILMOBKGLSEEC : 265
Le YLP5R6 1 G m W 6tFGM L IP E C L 2Pa a6 L ND6fSW kE a q H6 NA6w66M E

Ala11_AGA_RHJD_16069 : AOEICRNLIKKTISFYVQVEDAKODESPDARKFVEASKYSTIAGMAWSTTCPRYQCVSENEROLEWNRNGVFNK---PGSFEF : 352
Ala11_AGA_RHJD_16091 : AOCMCRRLIKTHVSDYITQIVENKKNESADIRRYTEAMQYITIGNIAMSNCPRYPPOASLNEHOLEWMSVDPKLTSLPSPPAP : 358
Ala11_DOT_GEAK_12341 : AOCYCREBAAYVARYVENVQRKODESSADIRRYTEAMQYITIGNIAMSNCPRYPPOCOHENOTQVDMWHLNGVDPPTGFDSS---SS : 352
Ala11_DOT_DGLV_16613 : AOCYCREBAAYVSYVDSENRNRNEESSPOLRHYEAMQYISIGNVMSKFCPRYNPEKRENOTQVDMWHLNGVDPPTGFDSS---SS : 355
Ala11_DOT_PINU_12649 : ANCYCRKLAAYVYQYDNRNRNRNEESSPOLRHYEAMQYISIGNVMSKFCPRYNPEKRENOTQVDMWHLNGVDPPTGFDSS---SS : 355
A 2 CR L qy6 Y6 6 K ES S DLR 56EAmqY3tsGN WS CTRY P fn2tQ6 WM ng6P s

Ala11_AGA_RHJD_16069 : -CAGAESRFSNLGAST-----EPP-----SNPAPEEEKS-LSAIGQFVVEAPYQYISLSPSKGVRDFIDAVNWLV : 420
Ala11_AGA_RHJD_16091 : EVLGLSPFSWSVHSDSRSSSTPPLEPPITATKNLNLNLEPPPPPESE-----IWEAPYQYISLSPSKGVRDFIDAVNWLV : 437
Ala11_DOT_GEAK_12341 : YDSTSTNGSEPKPESTVESRS---QGGDDISGIMSSLLNCSLEPLSHKVRSS---TFGAPYQYISLSPSKGVRDFIDAVNWLV : 436
Ala11_DOT_DGLV_16613 : FLSTSTHGSFVSGSOATIESK---DGTADSSGIMSSLLNCSLEPLSHKVRSS---IWEAPYQYISLSPSKGVRDFIDAVNWLV : 432
Ala11_DOT_PINU_12649 : FLSTSTHGSFVSGSOATIESK---DGTADSSGIMSSLLNCSLEPLSHKVRSS---IWEAPYQYISLSPSKGVRDFIDAVNWLV : 432
s s s 3 e t 1 pP AP Y6 SLPSKG RD F6DA6N WL V

Ala11_AGA_RHJD_16069 : PHNVVKQIKAAINRHHASLLDDDDQSSLLRRCKPAATMFGAPQATNSAGYCHIKAEQIQAGNAIITNKL-----LSLYKQ : 502
Ala11_AGA_RHJD_16091 : PENIVEQIKALTNRLHOSALLDDDESSLLRRCKPAATMFGAPQATNSAGYCHIKAEQIQAGNAIITNKL-----LSLYKQ : 527
Ala11_DOT_GEAK_12341 : GRSTSSQVKVVRMLHNASLMDDDDQSSLLRRCKPAATMFGAPQATNSAGYCHIKAEQIQAGNAIITNKL-----LSLYKQ : 521
Ala11_DOT_DGLV_16613 : DGQRASQVKMATRMLHNASLMDDDDQSSLLRRCKPAATMFGAPQATNSAGYCHIKAEQIQAGNAIITNKL-----LSLYKQ : 517
Ala11_DOT_PINU_12649 : DEQKVSQVKMATRMLHNASLMDDDDQSSLLRRCKPAATMFGAPQATNSAGYCHIKAEQIQAGNAIITNKL-----LSLYKQ : 517
Q6K LH ASL6LD 2D SpLRR Kp ah 6FG Qt NSA 5 6 I 1 q 6 L SL5 GQ

Ala11_AGA_RHJD_16069 : ALDLEHTYNGIGFTPAEYIQMIDCKTGAQDFVIELMLAHSNLSKDF--LSKLTTLGRYFOIADDDYRNIVSADYAKOKGFCDDLDGK : 590
Ala11_AGA_RHJD_16091 : ALDLEHTYNGIGFTPAEYIQMIDCKTGAQDFVIELMLAHSNLSKDF--LSKLTTLGRYFOIADDDYRNIVSADYAKOKGFCDDLDGK : 615
Ala11_DOT_GEAK_12341 : AODLHSSRNLSRPSLEYIQMIDCKTGAQDFVIELMLAHSNLSKDF--LSKLTTLGRYFOIADDDYRNIVSADYAKOKGFCDDLDGK : 611
Ala11_DOT_DGLV_16613 : AODLHSSRNLSRPSLEYIQMIDCKTGAQDFVIELMLAHSNLSKDF--LSKLTTLGRYFOIADDDYRNIVSADYAKOKGFCDDLDGK : 607
Ala11_DOT_PINU_12649 : AODLHSSRNLSRPSLEYIQMIDCKTGAQDFVIELMLAHSNLSKDF--LSKLTTLGRYFOIADDDYRNIVSADYAKOKGFCDDLDGK : 607
A DLH 3 N P3 EYIQ ID KT A F L 6M s a p L 4 LLGR5FQI DD5 N6 S Y KQKGFC DL1 G

Ala11_AGA_RHJD_16069 : YSLPLIHLMSQDFDNLQRLNLSLSTRNEGRMMYEHKLVLYKYKEAKSLEYTYSTLADLHARTQOIDEELVGESNTEFLRLWEFLRV : 680
Ala11_AGA_RHJD_16091 : YSLPLIHLQSHPENLQRLNLSLSTRNEGRMMYEHKLVLYKYKEAKSLEYTYSTLADLHARTQOIDEELVGESNTEFLRLWEFLRV : 705
Ala11_DOT_GEAK_12341 : YTIPLIYTAQEPHNILQRLNLSLSTRNEGRMMYEHKLVLYKYKEAKSLEYTYSTLADLHARTQOIDEELVGESNTEFLRLWEFLRV : 701
Ala11_DOT_DGLV_16613 : YTIPLIYTAQEPHNILQRLNLSLSTRNEGRMMYEHKLVLYKYKEAKSLEYTYSTLADLHARTQOIDEELVGESNTEFLRLWEFLRV : 697
Ala11_DOT_PINU_12649 : YTIPLIYTAQEPHNILQRLNLSLSTRNEGRMMYEHKLVLYKYKEAKSLEYTYSTLADLHARTQOIDEELVGESNTEFLRLWEFLRV : 697
Y36PL6 6 N6 L N6LSTR aeg 6 qK L L 6 3 5 6L L 6 26 6 f n 6 6 L 6

```

Figure 5.9: Multiple sequence alignment of 2 Basidiomycota and 3 Ascomycota α 1 type terpene synthases.

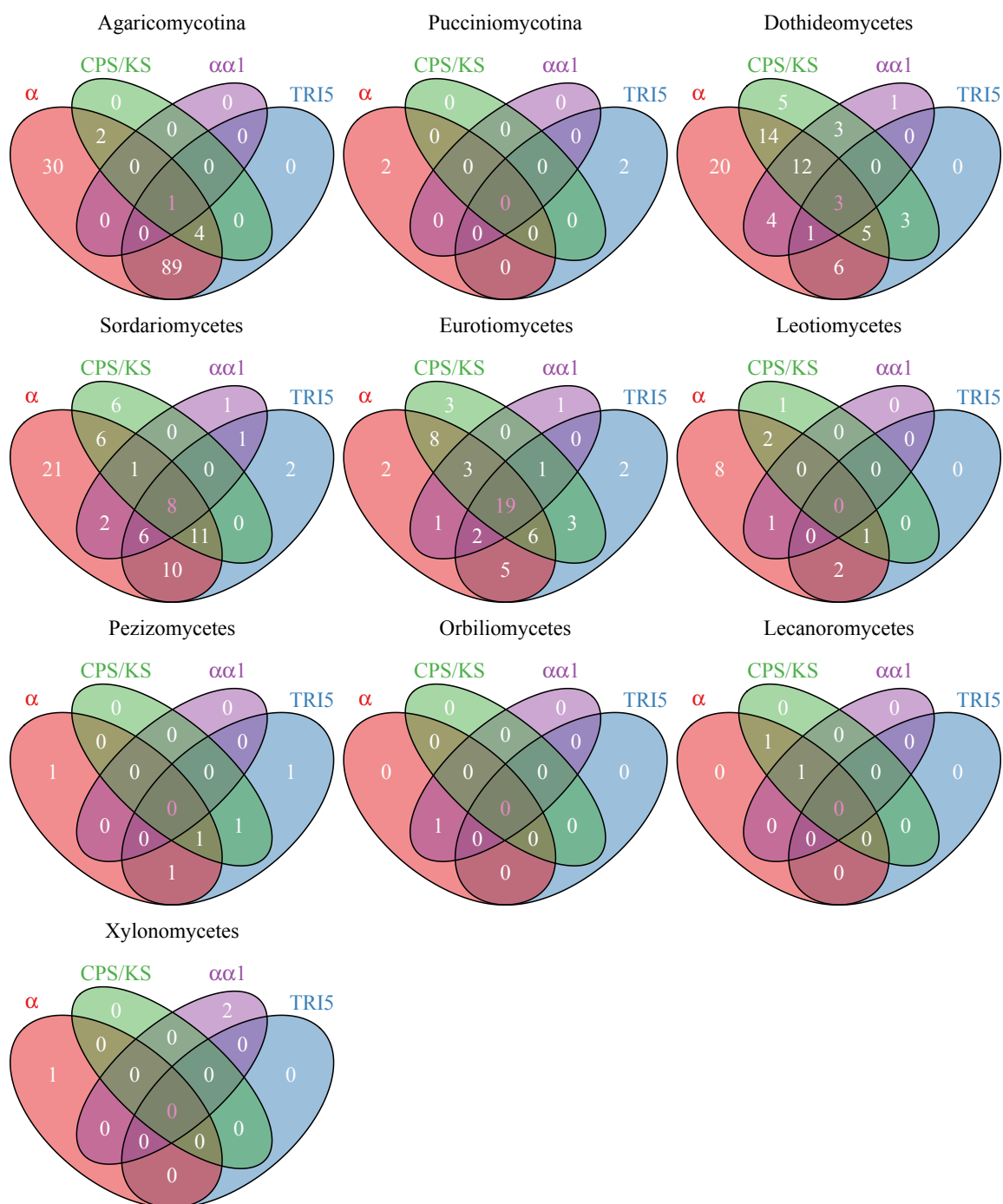


Figure 5.10: Species distribution in the four types of fungal Terpene synthases. Each venn diagram represents a fungal lineage and the number of species that are unique to and shared between four Terpene synthases types.

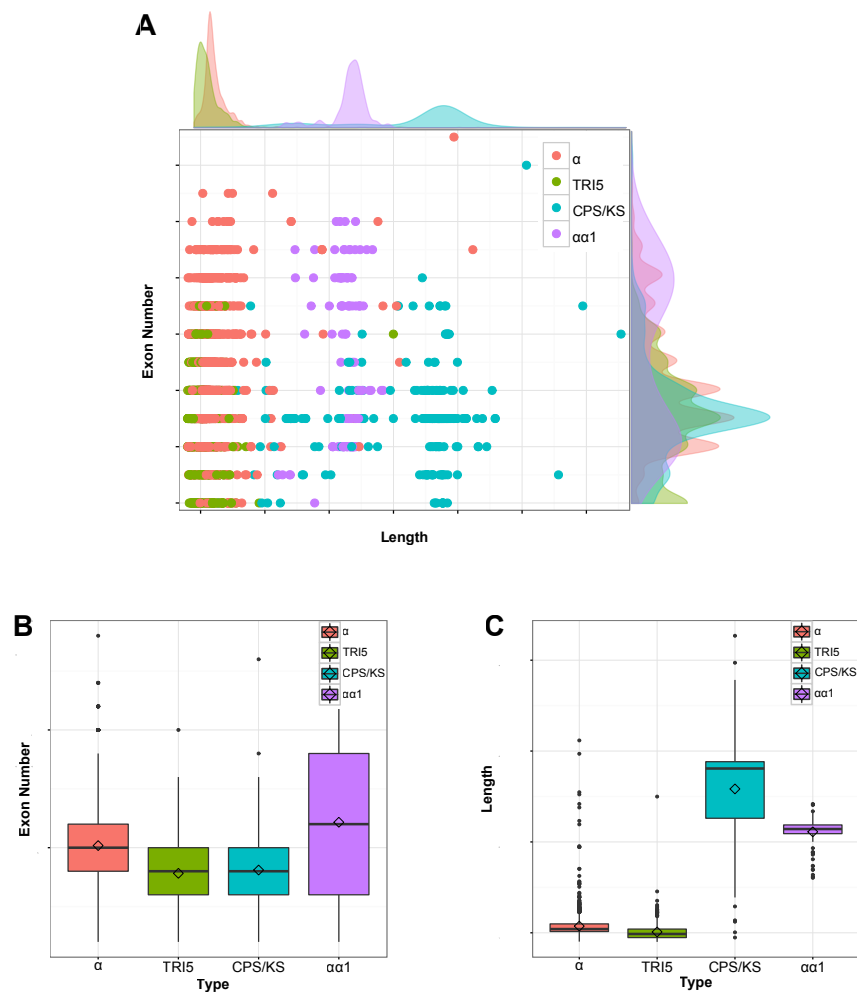


Figure 5.11: Comparisons of the terpene synthase gene length and exon number. (A) A scatter plot of exon number (y-axis) compared to the mRNA length (x-axis) for α (red points), TRI5 (forestgreen points), CPS/KS (darkturquoise points) and $\alpha\alpha 1$ (darkorchid1 points). The overlaid density distributions of exon number and terpene synthase gene length for each type are shown along the axes. (B) Boxplot shows the distribution of counts of exons across the four different terpene synthase genes. (C) Length of four types of terpene synthase genes shown in boxplot. (The horizontal line within boxplot shows the median and the diamond inside each boxplot shows the mean. Outliers are shown as black dots.)

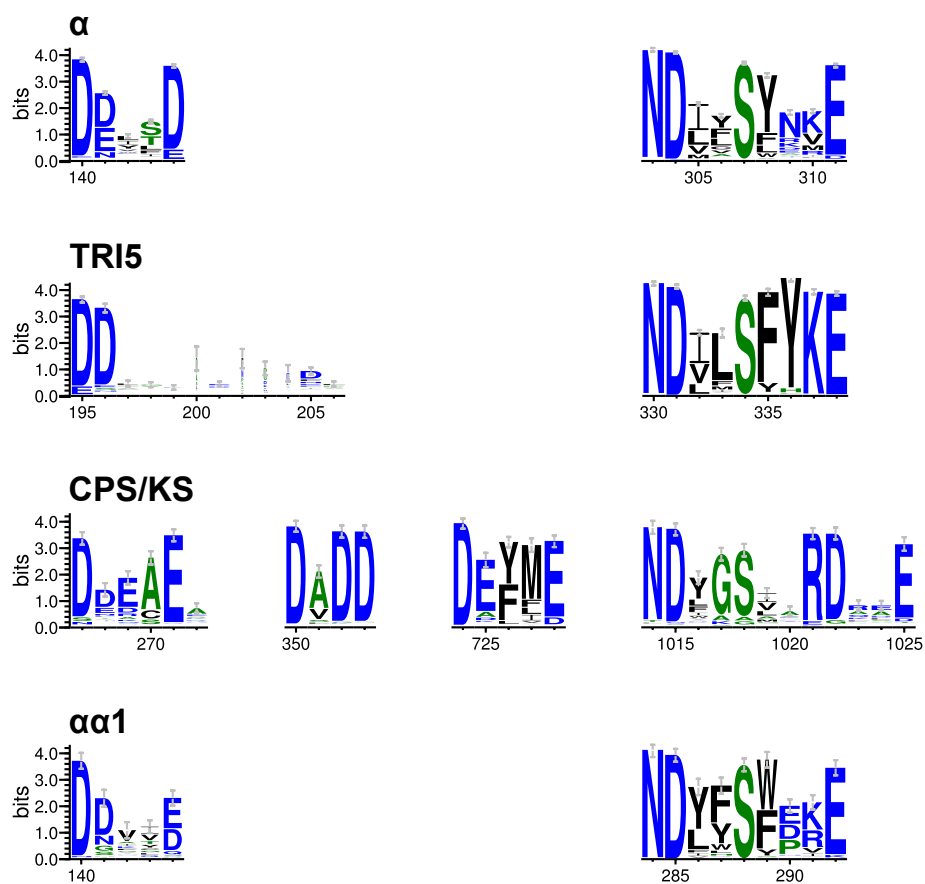


Figure 5.12: Sequence logos of conserved terpene synthase family motifs observed in each type of fungal terpene synthases. Bits indicate the conservation of amino acids at specific positions and the height of letters within each stack reflects the relative frequency of the corresponding amino acid. The logos were created on the WebLogo 3 server (<http://weblogo.threeplusone.com/create.cgi>).

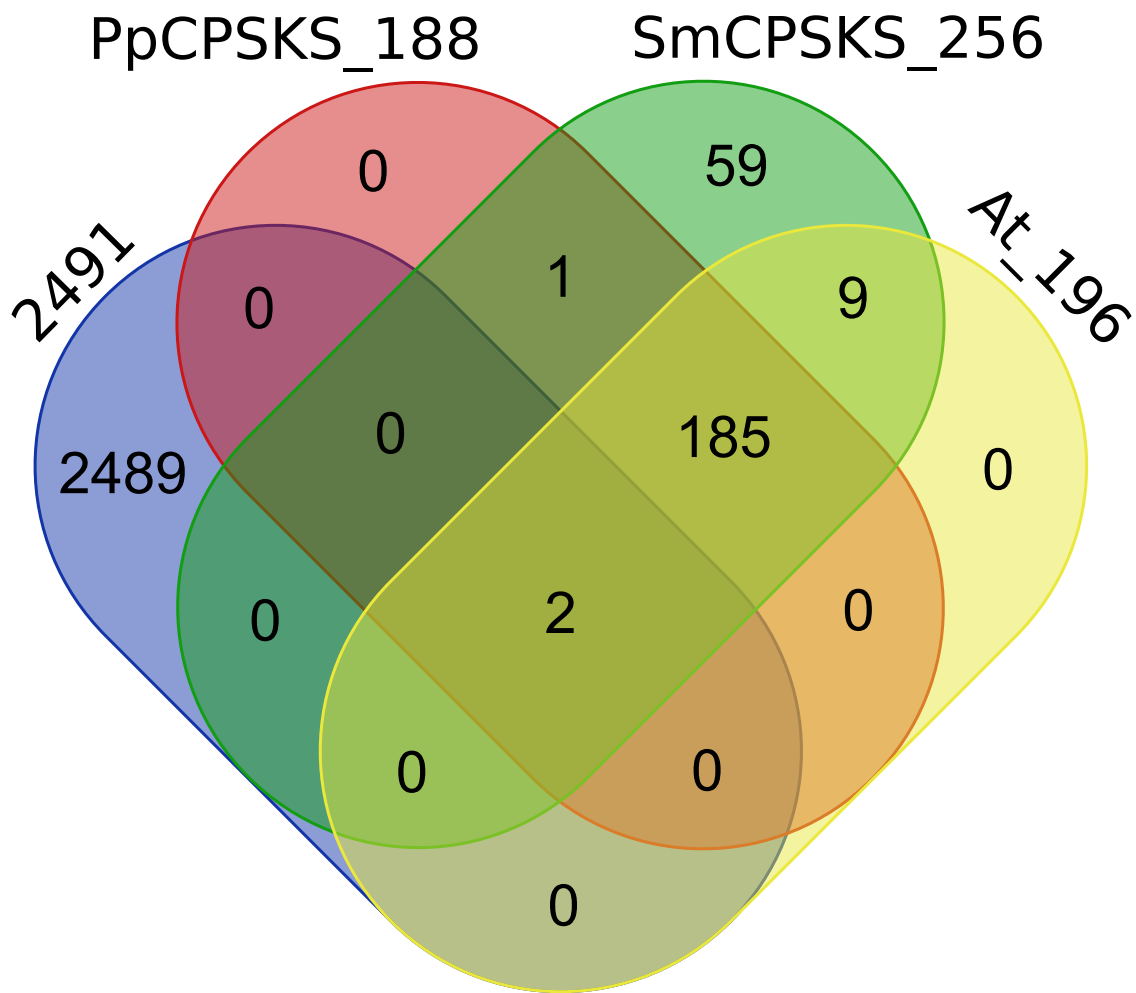


Figure 5.13: Venn diagram of identified CPS/KSs by known CPS/KSs from three species. SmCPSKS from *Sphaceloma manihoticola*; PpCPSKS from *Physcomitrella patens*; AtGA1 and AtGA2 from *Arabidopsis thaliana*.

Chapter 6

Summary and Conclusions

Recent advances in sequencing technology greatly increase the width, depth and accessibility of genomics data, and thus propelled the functional and comparative genomics studies at large scales. The research presented in this dissertation applies modern computational techniques into the study of terpene synthases in plants, fungi and bacteria. One aspect of this research is to attempt to trace the evolutionary history of this gene family by phylogenetic analysis. This benefits to a great extent from analyzing terpene synthases in a wide range of species.

In Chapter 1, an overview of the state-of-the-art knowledge on terpene synthases and their evolution was given. Also, specific aims were outlined: (1) identify and characterize the archetypical plant terpene synthases and microbial-type terpene synthases in the plant kingdom via comparative transcriptomic analysis of 1000 plants, (2) identify the terpene synthases from fungi and bacteria species, (3) infer the evolutionary history of this gene family by comparative analysis of terpene synthases found in plants, fungi and bacteria.

In Chapter 2, typical plant terpene synthases were identified and characterized from more than 1000 plant transcriptomes and six genomes of non-seed plants. The phylogenetic analysis based on sequences in non-seed plants and selected seed plants provides new insights into the evolution of this gene family. A new subfamily x was discovered and new hypotheses about the ancestor of modern plant terpene synthases was proposed.

In Chapter 3, by mining the transcriptomes of 1103 plant species ranging from green algae to flowering plants, 712 microbial type terpene synthases, a new type that is first found in the genome of *Selaginella moellendorffii* in 2012, were identified from 146

plant species, almost all of which are non-seed land plants. They are dominant only in major clades of non-seed land plants: Monilophytes, Lycophytes, Hornworts, Mosses and Liverworts. A phylogenetic tree built on these 712 microbial type terpene synthases and selected sequences from bacteria and fungi enables us classify them into four major groups and the biochemical analysis reveal that they functions as monoterpene synthases and sesquiterpene synthases.

In Chapter 4, two putative horizontal gene transfer events from bacterial to fungi in the terpene synthase gene family were identified. To the best of our knowledge, this is the first case of HGT of terpene synthase genes. These HGT-acquired genes were restricted to a group of entomopathogenic fungi and they were under stringent purifying selection in fungal recipients.

In Chapter 5, the distribution and genomic organization of four types of terpene synthase genes in 519 fungi genomes are characterized. The results show that two main groups of fungi, Ascomycota and Basidiomycota, have evolved different sets of terpene synthases. Compared to the α and TRI5 types of terpene synthases, two types of diterpene synthases, the bifunctional CPS/KS and the chimera diterpene synthases, were much less common in both groups. Also, almost all the diterpene synthases were encoded in the species of Ascomycota.

Appendix

Appendix A

List of species

A.1 List of OneKP Species

Table A.1: List of screened transcriptomes and the number of PTPS and MTPS in each sample. All these 1175 transcriptomes were from OneKP (www.onekp.com). The unique four letter codes are the transcriptome identifiers assigned by 1KP.

Code	Species	P	M
Anthocerotophyta (ANT)			
RXRQ	Phaeoceros carolinianus	17	4
WCZB	Phaeoceros carolinianus	2	9
FAJB	Paraphymatoceros hallii	2	3
TCBC	Megaceros vincentianus	2	2
DXOU	Nothoceros aenigmaticus	0	0
FANS	Leiosporoceros dussii	0	0
IQJU	Anthoceros formosae	0	0
UCRN	Megaceros tosanus	0	0
Basal Magnoliophyta (BMA)			
VZCI	Illicium floridanum	20	0
FZJL	Austrobaileya scandens	9	0
URDJ	Amborella trichopoda	5	0
ROAP	Illicium parviflorum	4	0
NWMY	Kadsura heteroclita	2	0
WTKZ	Nuphar advena	2	0
NPND	Ceratophyllum demersum	0	0
Bryophyta (BRY)			
HVBQ	Tetraphis pellucida	4	0
LNSF	Hypnum subimponens	3	5
ORKS	Philonotis fontana	3	5
QKQO	Pseudotaxiphyllum elegans	3	3
BGXB	Plagiomnium insignis	3	2
WNGH	Aulacomnium heterostichum	3	1
KEFD	Encalypta streptocarpa	3	0
RCBT	Sphagnum palustre	2	7
NGTD	Dicranum scoparium	2	3
TMAJ	Neckera douglasii	2	2
JADL	Rhynchostegium serrulatum	1	4
RDOO	Racomitrium varium	1	4
VBMM	Anomodon rostratus	1	4
IGUH	Schwetschkeopsis fabronia	1	3
RGKI	Leucobryum glaucum	1	3
UHLI	Sphagnum recurvatum	1	3
HRWG	Buxbaumia aphylla	1	1
QMWB	Anomodon attenuatus	1	1
WSPM	Rhytidiadelphus loreus	1	1
XWHK	Funaria	1	1
CMEQ	Orthotrichum lyellii	1	0
WOGB	Andreaea rupestris	1	0
EEMJ	Thuidium delicatulum	0	5
GOWD	Sphagnum lescurii	0	5
AWOI	Diphyscium foliosum	0	3
SKQD	Takakia lepidiozoides	0	2
VMXJ	Leucobryum albidum	0	2
ZACW	Leucodon sciurioides	0	2
ZQRI	Timmia austriaca	0	2
ABCD	Niphotrichum elongatum	0	1
GRKU	Syntrichia princeps	0	1
MIRS	Climacium dendroides	0	1
SZYG	Polytrichum commune	0	1
ZTHV	Atrichum angustatum	0	1
BPSG	Scouleria aquatica	0	0

Table A.1: Continued

Code	Species	P	M
DHWX	Fontinalis antipyretica	0	0
FFPD	Ceratodon purpureus	0	0
JMXW	Bryum argenteum	0	0
TAVP	Calliergon cordifolium	0	0
YEPO	cf. Physcomitrium sp.	0	0
YWNF	Hedwigia ciliata	0	0
Chloranthales (CHA)			
OSHQ	Sarcandra glabra	13	0
WZFE	Ascarina rubricaulis	8	0
Chlorophyta (CHO)			
FPCO	Interfilum paradoxum	2	0
MCHJ	Micrasterias fimbriata	0	1
ACRY	Pteromonas sp.	0	0
AEKF	Penium margaritaceum	0	0
AJAU	Helicodictyon planctonicum	0	0
AJUW	Chloromonas rosa	0	0
AKCR	Parachlorella kessleri	0	0
ALZF	Halochlorococcum marinum	0	0
AYPS	Unidentified species CCMP 1205	0	0
AZZW	Chlorokybus atmophyticus	0	0
BAZF	Chaetopeltis orbicularis	0	0
BCYF	Chlamydomonas cribrum	0	0
BFIK	Entransia fimbriat	0	0
BHBK	Cosmarium tinctum	0	0
BILC	Prototheca wickerhamii	0	0
BTFM	Monomastix opisthostigma	0	0
BZSH	Golenkinia longispicula	0	0
CBNG	Planophila laetevirens	0	0
CQQP	Ochlochaete sp.	0	0
DFDS	Desmidium aptogonum	0	0
DRFX	Closterium lunula	0	0
DRGY	Chaetosphaeridium globosum	0	0
DUMA	Tetraselmis cordiformis	0	0
DVYE	Oedogonium cardiacu	0	0
DZPJ	Cylindrocapsa geminella	0	0
EATP	Microthamnion kuetzigianum	0	0
EEJO	Neochloris oleoabundans	0	0
EGNB	Scourfieldia sp.	0	0
ENAU	Spermatozopsis similis	0	0
ETGN	Botryococcus braunii	0	0
FFGR	Netrium digitus	0	0
FMRU	Zygnema sp.	0	0
FMVB	Scherffelia dubia	0	0
FOYQ	Microspora cf. tumidula	0	0
FQLP	Klebsormidium subtile	0	0
FXHG	Hafniomonas reticulata	0	0
GBGT	Xanthidium antilopaeum	0	0
GFUR	Chloromonas subdivisa	0	0
GGWH	Onychonema laeve	0	0
GJIY	Neochloris sp.	0	0
GUDB	Brachiomonas submarina	0	0
GXBM	Coccomyxa pringsheimii	0	0
GYBH	Codium fragile	0	0
GYRP	Euastrum affine	0	0
HAOX	Spirogyra sp.	0	0
HHXJ	Tetraselmis striata	0	0
HIDG	Cosmarium broomei	0	0
HJVM	Cosmarium ochthodes	0	0
HKZW	Mesotaenium caldariorum	0	0
HVNO	Tetraselmis chui	0	0
HYHN	Prasinoderma coloniale	0	0

Table A.1: Continued

Code	Species	P	M
IHOI	Chloromonas oogama	0	0
IJMT	Aphanochaete repens	0	0
IRYH	Heterochlamydomonas inaequalis	0	0
ISGT	Hormidiella sp.	0	0
ISHC	Staurostrum sebaldi	0	0
ISIM	Nephroselmis pyriformis	0	0
ISPU	Volvox globator	0	0
JIWJ	Acrosiphonia sp.	0	0
JKKI	Lobomonas rostrata	0	0
JMTE	Pseudoscourfieldia marina	0	0
JMUI	Stigeoclonium helveticum	0	0
JOJQ	Cylindrocystis cushleackae	0	0
JRDV	Pediastrum duplex	0	0
JRGZ	Chlamydomonas moewusii	0	0
JTIG	Bryopsis plumosa	0	0
KADG	Ignatius tetrasporus	0	0
KEYW	Gonatozygon kinahanii	0	0
KFEB	Haematococcus pluvialis	0	0
KMNX	Nucleotaenium eifelense	0	0
KSFK	Chlorosarcinopsis halophila	0	0
KUJU	Gonium pectorale	0	0
KYIO	Mesostigma viride	0	0
LETF	Planophila terrestris	0	0
LNIL	Pteromonas angulosa	0	0
LSHT	Bolbocoleon piliferum	0	0
MCPK	Bathycoccus prasinos	0	0
MFYC	Nannochloris atomus	0	0
MFZO	Zygnemopsis sp.	0	0
MMKU	Nephroselmis olivace	0	0
MNCB	Eremosphaera viridi	0	0
MNNM	Cosmarium granatum	0	0
MOYY	Pleurotaenium trabecul	0	0
MWAN	Chlorella minutissima	0	0
MWXT	Chara vulgaris	0	0
MXDS	Spermatozopsis exsultans	0	0
MXEZ	Pycnococcus provasolii	0	0
NATT	Trentepohlia annulata	0	0
NBYP	Mesotaenium kramstei	0	0
NDPQ	Dunaliella salina	0	0
NKXU	Trebouxia arboricola	0	0
NNHQ	Spirotaenia minuta	0	0
NQYP	Pirula salina	0	0
NSTT	Oltmannsiellopsis viridis	0	0
OAEZ	Persursaria percursa	0	0
ODXI	Haematococcus pluvialis	0	0
OFUE	Lobochlamys segnis	0	0
OQON	Entocladia endozoica	0	0
OTQG	Ankistrodesmus sp.	0	0
PFUD	Geminella sp.	0	0
POIR	Volvox aureus	0	0
PRIQ	Pleurastrum insigne	0	0
PUAN	Pedinomonas tuberculata	0	0
PZIF	Scenedesmus dimorphus	0	0
QPDY	Coleochaete irregularis	0	0
QRTH	Chloromonas perforata	0	0
QWVW	Bambusina borrieri	0	0
QWRA	Vitreochlamys sp.	0	0
QXSZ	Mantoniella squamata	0	0
QYXY	Botryococcus terribilis	0	0
RAWF	Uronema belka	0	0
RNAT	Eudorina elegans	0	0

Table A.1: Continued

Code	Species	P	M
RPQV	Phymatodocis nordstedtiana	0	0
RPRU	Staurodesmus omearii	0	0
RQFE	Cosmocladium cf. constrictum	0	0
RRSV	Pedinomonas minor	0	0
RUIF	Carteria obtusa	0	0
RYJX	Pandorina morum	0	0
SDPC	Oedogonium foveolatum	0	0
SNOX	Planotaenium ohtanii	0	0
SRGS	Chlamydomonas bilatus	0	0
TGNL	Picocystis salinarum	0	0
TNAW	Pyramimonas parkeae	0	0
TPHT	Spirotaenia sp.	0	0
USIX	Neochlorosarcina sp.	0	0
UTRE	Chloromonas tughillensi	0	0
VALZ	Chlamydomonas noctigama	0	0
VAZE	Cylindrocystis sp.	0	0
VBLH	Cladophora glomerata	0	0
VFIV	Frittschiella tuberosa	0	0
VHIJ	Blastophysa cf. rhizopus	0	0
VIAU	Carteria crucifera	0	0
VJDZ	Botryococcus sudeticus	0	0
VQBJ	Coleochaete scutata	0	0
WCLV	Prasiola crispa	0	0
WCQU	Staurodesmus convergens	0	0
WDCW	Mesotaenium endlicherianum	0	0
WDGV	Cosmarium subumidum	0	0
WDWX	Dunaliella primolecta	0	0
WSJO	Mesotaenium braunii	0	0
WXRI	Stichococcus bacillaris	0	0
XDLL	Oogamochlamys gigantea	0	0
XIVI	Cymbomonas sp.	0	0
XJGM	coccoid prasinophyte	0	0
XMCL	Prasinococcus capsulatus	0	0
XOAL	Dolichomastix tenuilepi	0	0
XOZZ	Chlamydomonas sp.	0	0
XRTZ	Roya obtusa	0	0
YDCQ	Cephaleuros virescens	0	0
YLBK	Cylindrocystis brebissonii	0	0
YSQT	Penium exiguum	0	0
ZDIZ	Dunaliella tertiolecta	0	0
ZFXU	Asteromonas gracilis	0	0
ZIVZ	Phacotus lenticularis	0	0
ZLBP	Chloromonas reticulata	0	0
ZLQE	Stephanosphaera pluvialis	0	0
ZNUM	Leptosira obovata	0	0
ZRMT	Mougeotia sp.	0	0
Chromista (CHR)			
APTP	Ishige okamurai	0	0
ASZK	Punctaria latifolia	0	0
BAJW	Isochrysis sp.	0	0
BAKF	Cryptomonas curvata	0	0
BOGT	Mallomonas sp.	0	0
DBYD	Synura petersenii	0	0
EBWI	Ochromonas sp.	0	0
FIDQ	Undaria pinnatifida	0	0
FIKG	Sargassum henslowianum	0	0
FOMH	Sargassum integerrimum	0	0
FSQE	Desmarestia viridis	0	0
HFIK	Sargassum vachellianum	0	0
IAYV	Rhodomonas sp.	0	0
IRZA	Proteomonas sulcata	0	0

Table A.1: Continued

Code	Species	P	M
JCXF	Scytosihon lomentaria	0	0
JGGD	Sargassum muticum	0	0
JQFK	Nannochloropsis oculata	0	0
LDRY	Hizikia fusiforme	0	0
LIRF	Dictyopteris undulata	0	0
LXRN	Prymnesium parvum	0	0
MJMQ	Hemiselmis virescens	0	0
QLMZ	Colpomenia sinuosa	0	0
RAPY	Kjellmaniella crassifolia	0	0
ROZZ	Chroomonas sp.	0	0
RWXW	Sargassum horneri	0	0
SRSQ	Laminaria japonica	0	0
ULXR	Scytosiphon dotyo	0	0
VJED	Pavlova lutheri	0	0
VKVG	Synura sp.	0	0
VRGZ	Petalonia fascia	0	0
VYER	Sargassum hemiphyllum	0	0
YRMA	Sargassum thunbergii	0	0
Coniferophyta (CON)			
GMHZ	Cryptomeria japonica	46	0
BUWV	Platycladus orientalis	40	0
NKIN	Thujopsis dolabrata	40	0
YYPE	Austrocedrus chilensis	40	0
QNGJ	Cupressus dupreziana	39	0
AIGO	Chamaecyparis lawsoniana	37	0
XIRK	Athrotaxis cupressoides	34	0
CDFR	Manoao colensoi	33	0
QSNJ	Taiwania cryptomerioides	33	0
GKCZ	Diselma archeri	32	0
FRPM	Calocedrus decurrens	31	0
ZQVF	Cunninghamia lanceolata	31	0
NRXL	Metasequoia glyptostroboides	30	0
OWFC	Halocarpus bidwillii	30	0
ETCJ	Pilgerodendron uviferum	29	0
IIOL	Pinus parviflora	29	0
IOVS	Pseudotsuga menziesii	29	0
XQSG	Microbiota decussata	29	0
EFMS	Torreya taxifolia	28	0
GAMH	Tsuga heterophylla	28	0
GGEA	Cedrus libani	28	0
AUDE	Widdringtonia cedarbergensis	27	0
FMWZ	Dacrycarpus compactus	27	0
PLYX	Falcatifolium taxoides	27	0
DZQM	Pinus radiata	26	0
NVGZ	Cephalotaxus harringtonia	26	0
QCGM	Saxegothaea conspicua	26	0
XMGP	Juniperus scopulorum	26	0
JDQB	Neocallitropsis pancheri	25	0
VGSX	Retrophyllum minus	25	0
VSRH	Abies lasiocarpa	25	0
HILW	Acmopyle pancheri	23	0
IAJW	Amentotaxus argotaenia	23	0
JBND	Pinus ponderosa	23	0
AWQB	Picea engelmannii	22	0
JRNA	Phyllocladus hypophyllus	22	0
MHGD	Microcachrys tetragona	22	0
MIXZ	Agathis robusta	22	0
XLGK	Podocarpus rubens	22	0
FHST	Taxodium distichum	21	0
OXGJ	Glyptostrobus pensilis	21	0
RMMV	Callitris macleayana	21	0

Table A.1: Continued

Code	Species	P	M
CGDN	Tetraclinis sp.	20	0
EGLZ	Prumnopitys andina	20	0
HQOM	Torreya nucifera	20	0
IFLI	Callitris gracilis	20	0
RSCE	Wollemia nobilis	20	0
JUWL	Keteleeria evelyniana	19	0
MFTM	Pinus jeffreyi	19	0
UEVI	Fokienia hodginsii	19	0
UUJS	Nageia nagi	19	0
XTZO	Araucaria rulei	19	0
IZGN	Dacrydium balansae	18	0
KLGF	Sundacarpus amarus	18	0
SCEB	Podocarpus coriaceus	17	0
OVIJ	Papuacedrus papuana	16	0
VFYZ	Thuja plicata	16	0
QFAE	Sequoiadendron giganteum	15	0
ZQWM	Lagarostrobos franklinii	15	0
AQFM	Pseudolarix amabilis	14	0
NPRL	Cathaya agryrophylla	14	0
YFZK	Sciadopitys verticillata	14	0
YLPM	Pseudotaxus chienii	14	0
AREG	Nothotsuga longibracteata	13	0
BTTS	Austrotaxus spicata	12	0
ACWS	Arucaria sp.	11	0
JZVE	Parasitaxus usta	11	0
WVWN	Larix speciosa	9	0
WWSS	Taxus baccata	9	0
BBDD	Microstrobos fitzgeraldii	6	0
ZYAX	Taxus cuspidata	5	0
HBGV	Sequoia sempervirens	0	0
Cycadales (CYC)			
XZUY	Cycas micholitzii	7	0
WLIC	Dioon edule	4	0
KAWQ	Stangeria eriopus	3	0
GNQG	Encephalartos barteri	2	0
Eudicotyledons (EUD)			
MHYG	Conzya canadensis	43	0
LRRR	Teucrium chamaedrys	39	0
GUMF	Anthemis tinctoria	36	0
BMSE	Senecio rowleyanus	33	0
XQRV	Ipomoea purpurea	32	0
BAHE	Solenostemon scutellarioides	31	0
EQDA	Salvia spp.	31	0
DESP	Erigeron speciosus	30	0
BEKN	Papaver rhoeas	28	0
CPOC	Convolvulus arvensis	28	0
ERIA	Heuchera sanguinea	26	0
FCCA	Boswellia sacra	26	0
OLXF	Hibiscus cannabinus	26	0
PVGM	Oncotheca balansae	26	0
UBLN	Xanthium strumarium	26	0
WOHL	Schlegelia parasitica	26	0
YGCX	Geranium maculatum	26	0
CAQZ	Symphoricarpos sp.	25	0
DSUV	Asclepia curassavica	25	0
OKEF	Hibbertia grossulariifolia	25	0
TQOO	Loropetalum chinense	25	0
WOBD	Limonium spectabile	25	0
EYKJ	Silybum marianum	24	0
FFFY	Valeriana officianalis	24	0
LKKX	Talinum sp.	24	0

Table A.1: Continued

Code	Species	P	M
NBMW	Sanchezia sp.	24	0
NHAG	Ipomoea nil	24	0
QCOU	Papaver setigerum	24	0
TEZA	Solidago canadensis	24	0
UHBV	Oresitrophe rupifraga	24	0
BNDE	Hypericum perforatum	23	0
CWLL	Schlegelia parasitica B	23	0
JETM	Bauhinia tomentosa	23	0
JYMN	Flaveria brownii	23	0
SMMC	Chenopodium quinoa	23	0
TXMP	Linum strictum	23	0
VKJD	Maesa lanceolata	23	0
AJFN	Mydocarpus sp.	22	0
AYMT	Eucalyptus leucoxylon	22	0
CYVA	Cimicifuga racemosa	22	0
EMBR	Ipomoea pubescens	22	0
ERWT	Ipomoea coccinea	22	0
GETL	Pogostemon sp.	22	0
GRFT	Buddleja sp.	22	0
JEPE	Flaveria bidentis	22	0
TQKZ	Angelica archangelica	22	0
ZBPY	Alternanthera brasileana	22	0
DMLT	Vitex agnus castus	21	0
HTIP	Paonia lactiflora	21	0
LQJY	Solanum xanthocarpum	21	0
PUCW	Agastache rugosa	21	0
WHNV	Micromeria fruticosa	21	0
YKZB	Plantago maritima	21	0
BJSW	Cannabis sativa	20	0
CWYJ	Heracleum lanatum	20	0
DUQG	Tanacetum parthenium	20	0
EAAA	Marrubium vulgare	20	0
EDEQ	Wrightia natalensis	20	0
GHLF	Solanum dulcamara	20	0
IANR	Rosa palustris	20	0
QJXB	Wikstroemia indica	20	0
RQNK	Papaver somniferum	20	0
SNNC	Leonurus japonicus	20	0
VXOD	Linum tenuifolium	20	0
WQUF	Nepenthes alata	20	0
ZCUA	Flaveria trinervia	20	0
ARYD	Oenothera laciniata	19	0
EDXZ	Schlegelia violacea	19	0
HLJG	Viburnum odoratissimum	19	0
KFZY	Tragopogon castellanus	19	0
MQIV	Phyla dulcis	19	0
OHAE	Polygala lutea	19	0
SCAO	Mollugo nudicaulis	19	0
SUVN	Hedera helix	19	0
VKGP	Geranium carolinianum	19	0
YFQX	Apocynum androsaemifolium	19	0
ZINQ	Oenothera grandiflora	19	0
BIDT	Flaveria cronquistii	18	0
BVOF	Linum perenne	18	0
CKDK	Ochna serrulata	18	0
GGJD	Strychnos spinosa	18	0
IDGE	Heliotropium racemosum	18	0
IZNU	Ipomoea lobata	18	0
NJLF	Viola canadensis	18	0
PJSX	Acacia pycnantha	18	0
PXYR	Euphorbia pekinensis	18	0

Table A.1: Continued

Code	Species	P	M
QZZU	Pyrenacantha malvifolia	18	0
RHAU	Chamaseyca mesebyranthemum	18	0
THHD	Elaeocarpus sylvestris	18	0
YADI	Asclepias syriaca	18	0
YKQR	Hamamelis virginiana	18	0
ZUHO	Anemone hupenhensis	18	0
AQZD	Flaveria pringlei	17	0
CLRW	Bacopa caroliniana	17	0
EDBB	Polyscias fruticosa	17	0
FDMM	Rosmarinus officinalis	17	0
GCFE	Verbena hastata	17	0
JBGU	Amaranthus palmeri	17	0
JEXA	Impatiens balsamifera	17	0
JTRM	Dipsacus sativum	17	0
LLQV	Portulaca cryptopetala	17	0
LYPZ	Flaveria pubescens	17	0
OAGK	Matricaria matricariodes	17	0
OQBM	Ipomoea indica	17	0
OSMU	Lycium sp.	17	0
RTTY	Salvadora sp.	17	0
TLCA	Oenothera speciosa	17	0
UTQR	Tabebuia umbellata	17	0
XVRU	Heliotropium calcicola	17	0
YGAT	Phyllanthus sp.	17	0
ZJRC	Sambucus canadensis	17	0
BFJL	Cornus floridana	16	0
BHYC	Linum lewisii	16	0
BZDF	Oenothera biennis	16	0
COBX	Polypremum procumbens	16	0
INSP	Myrica cerifera	16	0
JSZD	Bursera simaruba	16	0
LPGY	Viola tricolor	16	0
PCGJ	Anisacanthus quadrifidas	16	0
QSKP	Polansia trachysperma	16	0
QSLH	Ipomoea hederacea	16	0
RKLL	Copaifera officianalis	16	0
RQUG	Hakea prostrata	16	0
WBXY	Oenothera elata	16	0
XPBC	Galphimia gracilis	16	0
ZBTA	Boerhavia cf. spiderwort	16	0
ALUC	Ipomoea quamoclit	15	0
BLWH	Portulaca mauui	15	0
EILE	Rhamnus japonica	15	0
EJCM	Lindenbergia philippensis	15	0
FUMQ	Nepeta cataria	15	0
FYUH	Lavandula angustifolia	15	0
KPTE	Bixa orellana	15	0
LWCK	Lycium barbarum	15	0
OEKO	Heliotropium filiforme	15	0
OHKC	Alternanthera caracasana	15	0
PDQH	Aerva lanata	15	0
PMTB	Oenothera gaura	15	0
QUTB	Aextoxicon punctatum	15	0
SALZ	Pittosporum resiniferum	15	0
WAXR	Litchi chinensis	15	0
WEQK	Centella asiatica	15	0
WXVX	Ledum palustre	15	0
AEPI	Linum leonii	14	0
CFRN	Carthamus lanatus	14	0
DOVJ	Leontopodium alpinum	14	0
FUPX	Tragopogon pratensis	14	0

Table A.1: Continued

Code	Species	P	M
HNCF	<i>Linum hirsutum</i>	14	0
IMZV	<i>Oenothera grandis</i>	14	0
JNKW	<i>Cicerbita plumieri</i>	14	0
KWGC	<i>Crassula perforata</i>	14	0
MXFG	<i>Pinguicula agnata</i>	14	0
MZLD	<i>Ligustrum sinense</i>	14	0
NEBM	<i>Syzygium macranthum</i>	14	0
NHUA	<i>Castanea crenata</i>	14	0
QKMG	<i>Phoradendron serotinum</i>	14	0
TOKV	<i>Aphanopetalum resinosum</i>	14	0
TORX	<i>Olea europaea</i>	14	0
UDUT	<i>Larrea tridentata</i>	14	0
UPZX	<i>Cleome viscosa</i>	14	0
VLNB	<i>Gompholobium polymorphum</i>	14	0
WEAC	<i>Strobilanthes dyerianus</i>	14	0
WRPP	<i>Synsepalum dulcificum</i>	14	0
XRCX	<i>Aster tataricus</i>	14	0
YNUE	<i>Punica granatum</i>	14	0
YRHD	<i>Antirrhinum braun</i>	14	0
AXPJ	<i>Linum flavum</i>	13	0
BOLZ	<i>Atropa belladonna</i>	13	0
EZGR	<i>Portulaca oleracea</i>	13	0
FCBJ	<i>Saxifraga stolonifera</i>	13	0
GDZS	<i>Byblis gigantea</i>	13	0
HJMP	<i>Astragalus membranaceus</i>	13	0
IHPC	<i>Platycodon grandiflorus</i>	13	0
IRAF	<i>Argemone mexicana</i>	13	0
IYDF	<i>Thymus vulgaris</i>	13	0
JCMU	<i>Pinguicula caudata</i>	13	0
KTAR	<i>Chionanthus retusus</i>	13	0
MKZR	<i>Nicotiana sylvestris</i>	13	0
NGRR	<i>Ternstroemia gymnanthera</i>	13	0
NTEO	<i>Pittosporum sahnianum</i>	13	0
NVSO	<i>Flaveria palmeri</i>	13	0
OINM	<i>Hydrocotyle umbellata</i>	13	0
PTBJ	<i>Plantago virginica</i>	13	0
TVSH	<i>Bituminaria bituminosa</i>	13	0
UAXP	<i>Gyrostemon ramulosus</i>	13	0
XGUF	<i>Exocarpos cupressiformis</i>	13	0
XVJB	<i>Morus nigra</i>	13	0
ZCDJ	<i>Acacia argyrophylla</i>	13	0
AFQQ	<i>Inula helenium</i>	12	0
AUIP	<i>Phelline lucida</i>	12	0
AYIY	<i>Ruellia brittoniana</i>	12	0
CBJR	<i>Atriplex rosea</i>	12	0
CLMX	<i>Escallonia rubra</i>	12	0
CLNU	<i>Escallonia</i> sp. cv. Newport	12	0
COAQ	<i>Malesherbia fasciculata</i>	12	0
CTYH	<i>Basella alba</i>	12	0
DDRL	<i>Tragopogon dubius</i>	12	0
DIHD	<i>Heliotropium tenellum</i>	12	0
FWBF	<i>Alangium chinense</i>	12	0
HENI	<i>Quercus shumardii</i>	12	0
JPDJ	<i>Symplocos</i> sp.	12	0
KCPT	<i>Linum macraei</i>	12	0
KJAA	<i>Mollugo pentaphylla</i>	12	0
LAPO	<i>Draba oligosperma</i>	12	0
LWDA	<i>Alnus serrulata</i>	12	0
NUZN	<i>Cercidiphyllum japonicum</i>	12	0
PCNH	<i>Psychotria marginata</i>	12	0
QAUE	<i>Actinidia chinensis</i>	12	0

Table A.1: Continued

Code	Species	P	M
RFSO	<i>Lactuca graminifolia</i>	12	0
RTNA	<i>Oxera pulchella</i>	12	0
SUAK	<i>Codariocalyx motorius</i>	12	0
TJLC	<i>Nothofagus obliqua</i>	12	0
TJQY	<i>Kerria japonica</i>	12	0
UCNM	<i>Ajuga reptans</i>	12	0
UEEN	<i>Forestiera segregata</i>	12	0
UJGI	<i>Oenothera rosea</i>	12	0
UZWG	<i>Castanea pumila</i>	12	0
VNMY	<i>Bischofia javanica</i>	12	0
WFBF	<i>Podophyllum peltatum</i>	12	0
WKCY	<i>Urtica dioica</i>	12	0
WQRD	<i>Galium boreale</i>	12	0
XXYA	<i>Verbascum</i> sp.	12	0
YZGX	<i>Cyrilla racemiflora</i>	12	0
ZSSR	<i>Xanthicercis zambesiaca</i>	12	0
MRKX	<i>Phytolacca bogotensis</i>	11	1
AIOU	<i>Brugmansia sanguinea</i>	11	0
ATYL	<i>Scutellaria montana</i>	11	0
AXNH	<i>Oenothera filiformis</i>	11	0
CCID	<i>Akebia trifoliata</i>	11	0
DLJZ	<i>Solanum ptychanthum</i>	11	0
EVOD	<i>Eschscholzia californica</i>	11	0
FXGI	<i>Stylidium adnatum</i>	11	0
GBVZ	<i>Thalictrum thalictroides</i>	11	0
HAEU	<i>Berberidopsis beckleri</i>	11	0
HMFE	<i>Allionia incarnata</i>	11	0
IWIS	<i>Portulaca pilosa</i>	11	0
JGAB	<i>Mirabilis jalapa</i>	11	0
JHCN	<i>Oxalis</i> sp.	11	0
JKNQ	<i>Oenothera suffulta suffulta</i>	11	0
KPUM	<i>Exacum affine</i>	11	0
LDEL	<i>Portulaca amilis</i>	11	0
MDJK	<i>Heliotropium texanum</i>	11	0
MYVH	<i>Linum grandiflorum</i>	11	0
MYZV	<i>Tropaeolum peregrinum</i>	11	0
NAUM	<i>Ipomoea lindheimeri</i>	11	0
ODDO	<i>Ardisia humilis</i>	11	0
ODDC	<i>Linum bienne</i>	11	0
PEZP	<i>Glycyrrhiza glabra</i>	11	0
POZS	<i>Linum usitatissimum</i>	11	0
RJNQ	<i>Lagerstroemia indica</i>	11	0
RSPO	<i>Santalum acuminatum</i>	11	0
SBZH	<i>Pennantia corymbosa</i>	11	0
SIZE	<i>Passiflora caerulea</i>	11	0
SVVG	<i>Fagus sylvatica</i>	11	0
SWGX	<i>Tetrazygia bicolor</i>	11	0
TPEM	<i>Platyspermation crassifolium</i>	11	0
UPOG	<i>Anemone pulsatilla</i>	11	0
XRLM	<i>Buddleja lindleyana</i>	11	0
YHFG	<i>Nandina domestica</i>	11	0
YRBQ	<i>Flaveria angustifolia</i>	11	0
AAAX	<i>Atriplex prostrata</i>	10	0
AHRN	<i>Cuscuta pentagonia</i>	10	0
BJKT	<i>Delosperma echinatum</i>	10	0
CIAC	<i>Bergenia</i> sp.	10	0
DAYQ	<i>Mitella pentandra</i>	10	0
DCCI	<i>Calceolaria pinifolia</i>	10	0
DXQW	<i>Juglans nigra</i>	10	0
FGDU	<i>Syzygium paniculatum</i>	10	0
FYTP	<i>Daphniphyllum macropodium</i>	10	0

Table A.1: Continued

Code	Species	P	M
GNPX	<i>Oxera neriifolia</i>	10	0
GSZA	<i>Lonicera japonica</i>	10	0
GVCB	<i>Oenothera affinis</i>	10	0
HBUQ	<i>Licania michauxii</i>	10	0
JAFJ	<i>Bougainvillea spectabilis</i>	10	0
KYAD	<i>Celtis occidentalis</i>	10	0
MGVU	<i>Allamanda cathartica</i>	10	0
MYMP	<i>Astragalus propinquus</i>	10	0
OMYK	<i>Trianthemum portulacastrum</i>	10	0
SSDU	<i>Papaver bracteatum</i>	10	0
SXCE	<i>Physocarpus opulifolius</i>	10	0
TAGM	<i>Melissa officinalis</i>	10	0
UHJR	<i>Citrus x paradisi</i>	10	0
UQCB	<i>Portulaca molokaiensis</i>	10	0
VGVI	<i>Dendropemon caribaeus</i>	10	0
VMNH	<i>Sinapis alba</i>	10	0
VYDM	<i>Orobancha fasciculata</i>	10	0
XSSD	<i>Amaranthus cruentus</i>	10	0
AAAL	<i>Meliosma cuneifolia</i>	9	0
ASMV	<i>Ilex vomitoria</i>	9	0
AWJM	<i>Edgeworthia papyrifera</i>	9	0
BEFC	<i>Manilkara zapota</i>	9	0
DFYF	<i>Ilex</i> sp.	9	0
DKFZ	<i>Mertensia paniculata</i>	9	0
INQX	<i>Salix acutifolia</i>	9	0
JTQQ	<i>Glycyrrhiza lepidota</i>	9	0
KDCH	<i>Portulaca umbraticola</i>	9	0
KEGA	<i>Glycine soja</i>	9	0
LVNW	<i>Cocculus laurifolius</i>	9	0
OBTI	<i>Peganum harmala</i>	9	0
OUER	<i>Heliotropium convolvulaceum</i>	9	0
PIYM	<i>Galax urceolata</i>	9	0
PKMO	<i>Cistus inflatus</i>	9	0
PPPZ	<i>Rhododendron scopulorum</i>	9	0
PUDI	<i>Daphne geraldii</i>	9	0
PZRT	<i>Nelumbo nucifera</i>	9	0
QNOC	<i>Sanguisorba minor</i>	9	0
QTJY	<i>Euptelea pleiosperma</i>	9	0
RKFX	<i>Cercis canadensis</i>	9	0
RPPC	<i>Erythroxylum coca</i>	9	0
TKEK	<i>Mansoa alliacea</i>	9	0
UYED	<i>Flaveria sonorensis</i>	9	0
VFFP	<i>Acer negundo</i>	9	0
VUSY	<i>Nyssa ogeche</i>	9	0
WGET	<i>Kochia scoparia</i>	9	0
WWQZ	<i>Medinilla magnifica</i>	9	0
XMVD	<i>Chelidonium majus</i>	9	0
ZHMB	<i>Krameria lanceolata</i>	9	0
AQGE	<i>Humulus lupulus</i>	8	0
DAAD	<i>Ardisia revoluta</i>	8	0
DUNJ	<i>Helenium autumnale</i>	8	0
DZLN	<i>Oenothera picensis</i>	8	0
ECTD	<i>Gentiana acaulis</i>	8	0
EHNF	<i>Dillenia indica</i>	8	0
EZZT	<i>Passiflora edulis</i>	8	0
FAMO	<i>Conopholis americana</i>	8	0
FNEN	<i>Phlox</i> sp.	8	0
GIPR	<i>Aucuba japonica</i>	8	0
GIWN	<i>Sarcobatus vermiculatus</i>	8	0
GLVK	<i>Salix eriocephala</i>	8	0
HKMQ	<i>Oenothera villaricae</i>	8	0

Table A.1: Continued

Code	Species	P	M
HUQC	<i>Scaevola</i> sp.	8	0
IWMW	<i>Buxus sempervirens</i>	8	0
IZLO	<i>Lobelia siphilitica</i>	8	0
JGYZ	<i>Holarrhena pubescens</i>	8	0
KNMB	<i>Lathyrus sativus</i>	8	0
LRTN	<i>Monotropa uniflora</i>	8	0
NNGU	<i>Coriaria nepalensis</i>	8	0
ONLQ	<i>Atriplex hortensis</i>	8	0
OOVX	<i>Boykinia jamesii</i>	8	0
OQHZ	<i>Quillaja saponaria</i>	8	0
OTAN	<i>Deutzia scabra</i>	8	0
QEHE	<i>Rauvolfia tetraphylla</i>	8	0
QICX	<i>Ailanthus altissima</i>	8	0
QURC	<i>Dichroa febrifuga</i>	8	0
QXWF	<i>Flaveria kochiana</i>	8	0
RUUB	<i>Physena madagascariensis</i>	8	0
RVGH	<i>Drypetes deplanchei</i>	8	0
SIBR	<i>Celsia arcturus</i>	8	0
SIK	<i>Hakea drupaceae</i>	8	0
SJAN	<i>Oenothera serrulata</i>	8	0
SLOI	<i>Tiarella polyphylla</i>	8	0
UWUF	<i>Itea virginica</i>	8	0
XMBA	<i>Poliomintha bustamanta</i>	8	0
XNLP	<i>Manihot grahamii</i>	8	0
ZBVT	<i>Chrysobalanus icaco</i>	8	0
ZETY	<i>Hydrangea quercifolia</i>	8	0
ZSGF	<i>Psychotria ipecacuanha</i>	8	0
ZTLR	<i>Rhizophora mangle</i>	8	0
ABEH	<i>Heliotropium greggii</i>	7	0
BGZG	<i>Cissus quadrangularis</i>	7	0
BSEY	<i>Daenikera</i> sp.	7	0
CTSS	<i>Tellima breviflora</i>	7	0
CVDF	<i>Simmondsia chinensis</i>	7	0
DTNC	<i>Sinningia tuberosa</i>	7	0
EDIT	<i>Sessuvium ventricosum</i>	7	0
EJBY	<i>Anticharis glandulosa</i>	7	0
EMAL	<i>Ehretia acuminata</i>	7	0
FEDW	<i>Epilobium</i> sp.	7	0
FZQN	<i>Silene latifolia</i>	7	0
GNRI	<i>Digitalis purpurea</i>	7	0
HDSY	<i>Aerva persica</i>	7	0
HGSM	<i>Gelsemium sempervirens</i>	7	0
HPNZ	<i>Oenothera longituba</i>	7	0
IUSR	<i>Myriophyllum aquaticum</i>	7	0
KGJF	<i>Tragopogon porrifolius</i>	7	0
NMGG	<i>Hypocoum procumbens</i>	7	0
PSHB	<i>Lantana camara</i>	7	0
QIKZ	<i>Griselinia racemosa</i>	7	0
STKY	<i>Balanophora fungosa</i>	7	0
SWPE	<i>Reseda odorata</i>	7	0
SYHW	<i>Ribes</i> aff. <i>giraldii</i>	7	0
TIUZ	<i>Cunonia capensis</i>	7	0
UVDC	<i>Azadirachta indica</i>	7	0
VVPY	<i>Croton tiglium</i>	7	0
XHKT	<i>Sanguinaria canadensis</i>	7	0
XISJ	<i>Anthriscum majus</i>	7	0
XMQO	<i>Gunnera manicata</i>	7	0
YSRZ	<i>Fouquieria macdougalli</i>	7	0
YUOM	<i>Rhus radicans</i>	7	0
AVJK	<i>Cavendishia cuatrecasasii</i>	6	0
BCAA	<i>Kirkia wilmsii</i>	6	0

Table A.1: Continued

Code	Species	P	M
BERS	Zaleya pentandra	6	0
BKQU	Phytolacca americana	6	0
BLVL	Sorbus koehneana	6	0
CSUV	Cochlearea officinalis	6	0
CWZU	Betula pendula	6	0
DRIL	Kalanchoe crenato diagremontiana	6	0
EGOS	Allionia spp.	6	0
EQYT	Oenothera berlandieri	6	0
FONV	Greyia sutherlandii	6	0
FROP	Epifagus virginiana	6	0
FWCQ	Garcinia oblongifolia	6	0
GJNX	Cypselea humifusum	6	0
HDWF	Francoa appendiculata	6	0
HXCD	Flaveria vaginata	6	0
HYZL	Akania lucens	6	0
IDAU	Oenothera elata hookeri	6	0
KVAY	Trubulus eichlerianus	6	0
MFEA	Stackhousia spathulata	6	0
MZOB	Heliotropium mendocinum	6	0
NCVK	Prunus prostrata	6	0
OLES	Schiedea membranacea	6	0
OSIP	Garcinia livingstonei	6	0
PHCE	Prunella vulgaris	6	0
QACK	Helwingia japonica	6	0
QIEH	Cotoneaster transcaasicus	6	0
RXEN	Polycarpaea repens	6	0
SKNL	Saponaria officianalis	6	0
TTRG	Lupinus angustifolius	6	0
TZWR	Arabis alpina	6	0
UOYN	Catharanthus roseus	6	0
VHZV	Gleditsia sinensis	6	0
VTIJ	Caiohpora chuquitensis	6	0
WMLW	Amaranthus retroflexus	6	0
WMUK	Schizolaena sp.	6	0
WVEF	Rhamnus caroliniana	6	0
WYIG	Dombeya burgessiae	6	0
XAYK	Kigelia africana	6	0
YNFJ	Microtea debilis	6	0
ZHEE	Ziziphus jujuba	6	0
ZRIN	Uncarina grandidieri	6	0
ACFP	Boehmeria nivea	5	0
ATFX	Muntingia calabura	5	0
AZBL	Petiveria alliacea	5	0
BNTL	Souroubea exauriculata	5	0
BXBF	Draba sachalinensis	5	0
CPKP	Lophophora williamsii	5	0
DHAW	Geum quellyon	5	0
EAVM	Amelanchier canadensis	5	0
FYSJ	Polygonum convolvulus	5	0
GCYL	Portulaca suffruticosa	5	0
IEPQ	Salix dasyclados	5	0
KKDQ	Salix viminalis	5	0
NGS	Heliotropium karwinsky	5	0
OCTM	Begonia sp.	5	0
OWAS	Rehmannia glutinosa	5	0
RYC	Elaeagnus pungens	5	0
SFKQ	Hillieria latifolia	5	0
SQCF	Dryas octopetala	5	0
SWOH	Trochodendron araliodes	5	0
SXML	Ilex paraguariensis	5	0
TJES	Spergularia media	5	0

Table A.1: Continued

Code	Species	P	M
UMUL	Paulownia fargesii	5	0
VQFW	Platanus occidentalis	5	0
XKPS	Ximenia americana	5	0
YHLF	Oenothera rhombipetala	5	0
ZSAB	Hoheria angustifolia	5	0
BWRK	Alternanthera sessilis	4	0
CKKR	Astilbe chinensis	4	0
CPLT	Portulaca grandiflora	4	0
CUTE	Blutoparon vermiculare	4	0
CZPV	Moringa oleifera	4	0
GEHT	Gleditsia triacanthos	4	0
GRRW	Grevillea robusta	4	0
IHCQ	Crossopetalum rhacoma	4	0
IXVJ	Menyanthes trifoliata	4	0
JNVS	Datura metel	4	0
JWEY	Heliotropium sp.	4	0
KBRW	Oenothera clelandii	4	0
KTWL	Kaliphora madagascariensis	4	0
LNER	Casuarina glauca	4	0
LVUS	Cleome violaceae	4	0
MVSE	Griselinia littoralis	4	0
PTFA	Jacquinia sp.	4	0
PTLU	Staphylea trifolia	4	0
RJIM	Melia azedarach	4	0
RWKR	Saintpaulia ionantha	4	0
SHEZ	Dianthus sp.	4	0
TVCU	Ochna mossambicensis	4	0
VCIN	Malus baccata	4	0
WVMY	Phlox drummondii	4	0
XFFT	Cercocarpus ledifolius	4	0
YKFU	Peltoboykinia watanabei	4	0
YQIJ	Phacelia campanularia	4	0
ZJUL	Rhodiola rosea	4	0
CJGZ	Oenothera nana	3	0
CQMG	Ulmus alata	3	0
DLAI	Solanum lasiophyllum	3	0
EYRD	Alternanthera tenella	3	0
FVXD	Beta maritima	3	0
HBHB	Aesculus pavia	3	0
HUSX	Roridula gorgonias	3	0
IPWB	Brassica nigra	3	0
LFOG	Salix purpurea	3	0
NFXV	Mumea americana	3	0
NMDZ	Solanum sisymbriifolium	3	0
NXOH	Apios americana	3	0
NXTS	Mollugo verticillata	3	0
OXYP	Sideroxylon reclinatorum	3	0
PAZJ	Ricinus communis	3	0
RZTJ	Salix fargesii	3	0
SERM	Sarcodes sanguinea	3	0
SZPD	Tetrastigma voinierianum	3	0
UDHA	Ceratocarpus vesicaria	3	0
UGJI	Lycopersicon cheesmanii	3	0
ULGV	Morinda citrifolia	3	0
UVQL	Draba magellanica	3	0
VDKG	Cleome gynandra	3	0
VGGH	Hydrastis canadensis	3	0
VYGG	Stachyurus praecox	3	0
WWKL	Tapiscia sinensis	3	0
AJJE	Terminalia neotaliala	2	0
AXBO	Sinojackia xylocarpa	2	0

Table A.1: Continued

Code	Species	P	M
BTZX	Philadelphus inodorus	2	0
CMFF	Lupinus polyphyllus	2	0
DZTK	Batis maritima	2	0
EDHN	Ficus religiosa	2	0
HANM	Pholisma arenarium	2	0
HZTS	Sessuvium portulacastrum	2	0
JLOV	Pereskia aculeata	2	0
KVFU	Diospyros malabarica	2	0
NJJO	Pilostyles thunbergii	2	0
QZXQ	Gymnocladus dioicus	2	0
SLYR	Cladrastis lutea	2	0
SZUO	Eucommia ulmoides	2	0
TDTF	Salix sachalinensis	2	0
VJPU	Boerhavia domininii	2	0
YZVJ	Cephalotus follicularis	2	0
ZGQD	Corydalis linstowiana	2	0
ZPKK	Aruncus dioicus	2	0
AUGV	Capnoides sempervirens	1	0
CRNC	Limnanthes douglassii	1	0
FAKD	Nelumbo sp.	1	0
HABV	Draba aizoides	1	0
IKFD	Quassia amara	1	0
JOIS	Koeberlina spinosa	1	0
KZED	Senna hebecarpa	1	0
LQUX	Gylcine soja	1	0
RNBN	Mollugo cerviana	1	0
SMUR	Lennoa madreporoides	1	0
QAIR	Opuntia sp.	0	4
AKRH	Gylcine soja	0	0
AXAF	Thladiantha villosula	0	0
BBBA	Tetrastigma obtectum	0	0
CGGO	Plumbago auriculata	0	0
DCVZ	Plantago coronopis	0	0
DNQA	Psychotria douarrei	0	0
DYFF	Pycnanthemum tenuifolium	0	0
FPLR	Gylcine soja	0	0
GIOY	Corokia cotoneaster	0	0
GTSV	Draba hispida	0	0
HRUR	Utricularia sp.	0	0
HSXO	Ancistrocladus tectorius	0	0
HTDC	Tamarix chinensis	0	0
JLLY	Melaleuca quinquenervia	0	0
LJQF	Draba ossetica	0	0
LMVB	Liquidambar styraciflua	0	0
NIJU	Heteropyxis natalensis	0	0
PFSA	Gylcine soja	0	0
PGKL	Phellodendron amurense	0	0
RIDD	Gylcine soja	0	0
RMWJ	Wisteria floribunda	0	0
TJMB	Gylcine soja	0	0
UZNH	Curtisia dentata	0	0
VWIP	Carya glabra	0	0
XONJ	Camptotheca acuminata	0	0
XOOE	Desmanthus illinoensis	0	0
	Euglenida (EUG)		
UNBZ	Euglena sp.	0	0
	Ginkgoales (GIN)		
SGTW	Ginkgo biloba	0	0
	Glaucophyta (GLA)		
JKHA	Cyanoptyche gloeocystis	0	0
POOW	Glaucocystis cf. nostochinearum	0	0

Table A.1: Continued

Code	Species	P	M
PQED	Gloeochaete wittrockiana	0	0
WLZP	Cyanophora paradoxa	0	0
	Gnetales (GNE)		
TOXE	Welwitschia mirabilis	5	0
GTHK	Gnetum montanum	4	0
VDAO	Ephedra sinica	2	0
	Liliopsida (LIL)		
EFCZ	Eleusine coracana	37	0
LEMW	Curcuma olena	22	0
WTDE	Johnsonia pubescens	18	0
ZENX	Neurachne alopecuroidea	18	0
TRRQ	Narcissus viridiflorus	17	0
XZME	Drakea elastica	17	0
YPIC	Microstegium vimineum	17	0
MTII	Acorus americanus	15	0
RCUX	Maianthemum sp.	15	0
SOHV	Panicum miliaceum A	15	0
WCOR	Thyridolepis multiculmis	15	0
RQZP	Nolina bigelorii	14	0
SILJ	Talbotia elegans	14	0
LTZF	Sisyrinchium angustifolium	13	0
MUMD	Lomandra longifolia	13	0
PRFO	Agapanthus africanus	13	0
CMCY	Hesperaloe parviflora	12	0
DGXS	Freyinetia multiflora	12	0
QNPH	Colchicum autumnale	12	0
VQYB	Neurachne lanigera	12	0
WBIB	Lepidosperma gibsonii	12	0
XHHU	Heliconia sp.	12	0
XUAB	Paraneurachne muelleri	12	0
BLAJ	Hemerocallis spp.	11	0
CNTZ	Oncidium sphacelatum	11	0
JHUL	Hemerocallis sp.	11	0
LELS	Haemaria discolor	11	0
MTHW	Platanthera clavellata	11	0
QOXT	Xerophyta villosa	11	0
TZNS	Canna sp.	11	0
UZXL	Disporopsis pernyi	11	0
YBML	Yucca brevifolia	11	0
DPFW	Zephyranthes treatiae	10	0
KOFB	Urginea maritima	10	0
LSKK	Orchidantha maxillarioides	10	0
NNOK	Neurachne tenuifolia	10	0
YXNR	Triodia aff. bynoei	10	0
BPKH	Neurachne annularis	9	0
HOKG	Nolina atopocarpa	9	0
HXJE	Serenoa repens	9	0
ICNN	Yucca filamentosa	9	0
YJUG	Curculigo sp.	9	0
BDJQ	Zingiber officinale	8	0
BXAY	Neurachne minor	8	0
DMIN	Phycella aff. cyrtanthoides	8	0
HATH	Aristida stricta	8	0
IXEM	Brodiaea sierrae	8	0
JDTY	Rhodophiala pratensis	8	0
MFIN	Pistia stratioides	8	0
ROEI	Cymbopogon nardus	8	0
UOEL	Strelitzia reginae	8	0
XBKS	Thyridolepis mitchelliana	8	0
AFLV	Xerophyllum asphodeloides	7	0
JNUB	Maranta leuconeura	7	0

Table A.1: Continued

Code	Species	P	M
JSAG	Masdevallia yuangensis	7	0
JVBR	Aloe vera	7	0
KYNE	Cyanella orchidofromis	7	0
MWYQ	Smilax bona nox	7	0
OOSO	Helonias bullata	7	0
THDM	Vanilla planifolia	7	0
VVVV	Ludovia sp.	7	0
XFJG	Maianthemum canadense	7	0
YMES	Typhonium blumei	7	0
ZMGN	Neurachne munroi	7	0
BSTR	Chondropetalum tectorum	6	0
COCF	Triglochin maritimum	6	0
GZSF	Uniola paniculata	6	0
IADP	Deschampsia cespitosa	6	0
KBXS	Allium commutatum	6	0
MVRF	Sansevieria trifasciata	6	0
OCWZ	Dioscorea villosa	6	0
SVTS	Drimys altissima	6	0
XPAF	Mapania palustris	6	0
FGRF	Asparagus densiflorus	5	0
KXSK	Agave tequilana	5	0
LDME	Amaryllis belladonna	5	0
LSJW	Ruscus sp.	5	0
TCYS	Peliosanthese minor	5	0
HWUP	Sabal bermudana	4	0
PLBZ	Chlorogalum pomeridianum	4	0
SART	Xeronema callistemon	4	0
BRUD	Typha latifolia	3	0
BYQM	Posidonia australis	3	0
CIEA	Juncus inflexus	3	0
EMJJ	Borya sphaerocephala	3	0
FCEL	Phormium tenax	3	0
ONBE	Eriosperrum lancifolia	3	0
VTUS	Goodyera pubescens	3	0
BYPY	Brocchinia reducta	2	0
GDKK	Gloriosa superba	2	0
NSPR	Nypa fruticans	2	0
PPQR	Typha angustifolia	2	0
PWSG	Cyperus papyrus	2	0
RDYY	Cyanastrum cordifolium	2	0
VBHQ	Stemona tuberosa	2	0
WXNT	Joinvillea ascendens	2	0
DWZT	Sagittaria latifolia	1	0
GJPF	Allium sativum	1	0
THEW	Lilium sargentiae	1	0
ZKPF	Traubia modesta	1	0
QGLJ	Cocos nucifera	0	0
RMVB	Avena fatua	0	0
Lycopodiophyta (LYC)			
ZZEI	Phylloglossum drummondii	8	4
ZFGK	Selaginella kraussiana	5	8
PYHZ	Isoetes sp.	5	7
ZZOL	Selaginella stauntoniana	4	2
YHZZ	Huperzia selago	3	7
PKOX	Isoetes tegetiformans	3	5
UPMJ	Pseudolycopodiella caroliniana	3	3
ZYCD	Selaginella acanthonota	3	3
WAFT	Diphasiastrum digitatum	3	0
JKAA	Selaginella wallacei	2	6
PQTO	Lycopodium deuterodensum	2	4
GAON	Huperzia squarrosa	2	2

Table A.1: Continued

Code	Species	P	M
KUXM	Selaginella selaginoides	2	2
LGDQ	Selaginella apoda	2	2
ABIJ	Selaginella lepidophylla	1	10
KJYC	Selaginella willdenowii	1	5
CBAE	Huperzia myrsinites	1	2
XNXF	Dendrolycopodium obscurum	1	2
ENQF	Lycopodium annotinum	1	1
GKAG	Huperzia lucidula	1	1
ULKT	Lycopodiella apressa	1	1
Magnoliidae (MAG)			
WBOD	Magnolia grandiflora	25	0
XQWC	Michelia maudiae	21	0
XSZI	Peperomia fraseri	20	0
FALI	Calycanthus floridus	19	0
OBPL	Myristica fragrans	19	0
WAIL	Laurelia sempervirens	17	0
MAQO	Gomortega keule	16	0
MUNP	Piper auritum	16	0
ABSS	Sassafras albidum	15	0
PSJT	Uvaria microcarpa	14	0
WKSU	Drimys winteri	12	0
PAWA	Aristolochia elegans	11	0
SJEV	Canella winterana	11	0
WPHN	Idiospermum australiense	10	0
YZRI	Annona muricata	9	0
VYLQ	Cassytha filiformis	8	0
BCGB	Cinnamomum camphora	7	0
BSVG	Gyrocarpus americanus	7	0
WIGA	Persea borbonia	7	0
CSSK	Houttuynia cordata	6	0
DHPO	Eupomatia bennettii	6	0
QDVV	Saruma henryi	5	0
WPYJ	Frankenia laevis	4	0
KRJP	Peumus boldus	3	0
OPDF	Saururus cernuus	1	0
KAYP	Lindera benzoin	0	0
Marchantiophyta (MAR)			
KRUQ	Porella navicularis	14	8
LFVP	Marchantia paleacea	11	8
JPYU	Marchantia polymorpha	7	7
RTMU	Calypogeia fissa	7	4
HPXA	Ptilidium pulcherrimum	6	11
TXVB	Lunularia cruciata	5	9
YBQN	Odontoschisma prostratum	4	10
IRBN	Scapania nemorosa	4	8
UUHD	Porella pinnata	4	8
JHFI	Pellia neesiana	3	16
LGOW	Schistochila sp.	3	3
CHJJ	Lejeuneaceae sp.	2	14
ILBQ	Conocephalum conicum	2	9
PIUF	Pellia cf. Epiphylla	2	9
TGKW	Frullania	2	8
WJLO	Riccia berychiana	2	4
YFGP	Pallavicinia lyellii	2	4
HERT	Sphaerocarpos texanus	2	3
NRWZ	Metzgeria crassipilis	2	1
WZYK	Bazzania trilobata	2	0
NWQC	Plagiochila asplenoides	1	9
BNCU	Radula lindenbergia	1	8
FITN	Treubia lacunose	1	6
OFTV	Barbilophozia barbata	0	11

Table A.1: Continued

Code	Species	P	M
TFYI	Marchantia emarginata	0	4
AEXY	Blasia sp.	0	1
TFDQ	Monoclea gottschei	0	0
Moniliformopses (MON)			
UOMY	Osmunda sp.	8	4
CQPW	Anemia tomentosa	6	12
MEKP	Dipteris conjugata	5	5
NDUV	Vittaria appalachiana	4	1
UJTT	Pityrogramma trifoliata	3	7
NHCM	Angiopteris evecta	3	5
VIBO	Osmunda javanica	3	2
GSXD	Myriopteris eatonii	2	11
ALVQ	Tmesipteris parva	2	4
XDVM	Sticherus lobatus	2	0
POPJ	Pteris vittata	1	10
YLJA	Polypodium amorphum	1	9
XWDM	Cystopteris fragilis	1	8
FQGG	Polystichum acrostichoides	1	5
WCLG	Adiantum aleuticum	1	5
QIAD	Hymenophyllum bivalve	1	4
ZXJO	Hemionitis arifolia	1	3
YIXP	Lindsaea microphylla	1	2
KIIX	Pilularia globulifera	1	1
HNDZ	Cystopteris utahensis	0	20
RWYZ	Blechnum spicant	0	15
NWWI	Nephrolepis exaltata	0	13
XDDT	Argyrochosma nivea	0	10
YJJY	Woodsia scopulina	0	10
UJWU	Pleopeltis polypodioides	0	9
BMJR	Adiantum tenerum	0	8
FCHS	Deparia lobato	0	8
OQWW	Davallia fejeensis	0	8
WGTU	Leucostegia immersa	0	8
CJNT	Polypodium glycyrrhiza	0	7
DCDT	Cheilanthes arizonica	0	7
DJSE	Ophioglossum vulgatum	0	7
JBLI	Danaea sp.	0	7
KJZG	Asplenium platyneuron	0	7
NOKI	Lindsaea linearis	0	7
BEGM	Botrypus virginianus	0	6
HEGQ	Gymnocarpium dryopteris	0	6
OCZL	Homalosorus pycnocarpus	0	6
PSKY	Asplenium nidus	0	6
UFJN	Diplazium wichurae	0	6
CVEG	Azolla cf. caroliniana	0	5
ORJE	Cibotium glaucum	0	5
PBUU	Lygodium japonicum	0	5
YOWV	Cystopteris protrusa	0	5
EEAQ	Sceptridium dissectum	0	4
YCKE	Notholaena montieliae	0	4
ZRAV	Polypodium hesperium	0	4
EWXK	Thyrsopteris elegans	0	3
MROH	Thelypteris acuminata	0	3
MTGC	Dennstaedtia	0	3
SKYV	Vittaria lineata	0	3
UXCS	Marattia sp.	0	3
WQML	Cryptogramma acrostichoides	0	3
YQEC	Woodsia ilvensis	0	3
ZQYU	Polypodium plectolens	0	3
CAPN	Equisetum diffusum	0	2
FLTD	Pteris ensiformis	0	2

Table A.1: Continued

Code	Species	P	M
PNZO	Culcita macrocarpa	0	2
QVMR	Psilotum nudum	0	2
RFRB	Didymochlaena truncatula	0	2
URCP	Athyrium filix femina	0	2
UWOD	Plagiogyria japonica	0	2
GANB	Cyathea spinulosa	0	1
HTFH	Onoclea sensibilis	0	1
RICC	Cystopteris reevesiana	0	1
TRPJ	Hymenophyllum cupressiforme	0	1
BIVQ	Osmundastrum cinnamomeum	0	0
JVSZ	Equisetum hymale	0	0
TWFZ	Crepidomanes venosum	0	0
YKSS	Osmunda regalis	0	0
Prorocentraceae (PRO)			
TZJQ	Prorocentrum micans	0	0
Rhodophyta (RHO)			
IKWM	Gracilaria lemaneiformis	1	0
FTRP	Gracilaria chouae	0	1
OBUE	Porphyridium cruentum	0	1
PVGP	Porphyridium purpureum	0	1
BWVJ	Betaphycus gelatinae	0	0
CKXF	Gymnogongrus ftabelliformis	0	0
IEHF	Dumontia simplex	0	0
IHYJ	Kappaphycus alvarezii	0	0
IKIZ	Grateloupia livida	0	0
JEBK	Eucheuma denticulatum	0	0
JJZR	Rhodochaete parvula	0	0
LJPN	Gracilaria blodgettii	0	0
LLXJ	Chroodactylon ornatu	0	0
PWKQ	Prionitis divaricata	0	0
PYDB	Sinotubimorpha guangdongensis	0	0
RSOF	Glaucosphaera vacuolata	0	0
RTLC	Rhodella violacea	0	0
SBLT	Gloeopeltis furcata	0	0
UGPM	Chondrus crispus	0	0
URSB	Grateloupia turuturu	0	0
UYFR	Symphyocladia latiuscula	0	0
VNAL	Gracilaria asiatica	0	0
VZWX	Ceramium kondoi	0	0
WEJN	Mazzaella japonica	0	0
XAXW	Polysiphonia japonica	0	0
YSBD	Heterosiphonia pulchra	0	0
ZJOJ	Grateloupia filicina	0	0
ZULJ	Porphyra yezoensis	0	0

A.2 List of 519 Fungi Species

Table A.2: List of 519 screened fungi genomes and the number of PTPS and MTPS in each species. These data were downloaded from the fungal genomics portal (<http://jgi.doe.gov/fungi>).

Code	JGI_id	Species	Clade	α	TRI5	α - α 1	CPS/KS	Total
KMDG	Denbi1	Dendrothele bispora CBS 962.96 v1.0	Agaricomycotina	33	5	0	1	39
QOEU	Sphst1	Sphaerobolus stellatus v1.0	Agaricomycotina	28	1	0	0	29
FTUW	Armga1	Armillaria gallica 21-2 v1.0	Agaricomycotina	20	7	0	0	27
RHJD	Gymlu1	Gymnopus luxurians v1.0	Agaricomycotina	13	7	2	3	25
DECH	Marfi1	Marasmius fiardii PR-910 v1.0	Agaricomycotina	17	8	0	0	25
PQBN	Suihi1	Suillus hirtellus EM16 v1.0	Agaricomycotina	21	3	0	0	24
DCEF	Leisp1	Leiotrametes sp BRFM 1775 v1.0	Agaricomycotina	17	6	0	0	23
CHIK	Galma1	Galerina marginata v1.0	Agaricomycotina	23	0	0	0	23
RSVJ	Corgl3	Cortinarius glaucopus AT 2004 276 v2.0	Agaricomycotina	17	5	0	0	22
MKIJ	Clapy1	Clavicornia pyxidata HHB10654 v1.0	Agaricomycotina	21	0	0	0	21
PFFQ	Gansp1	Ganoderma sp. 10597 SS1 v1.0	Agaricomycotina	13	7	0	0	20
ZCJM	Laesu1	Laetiporus sulphureus var. sulphureus v1.0	Agaricomycotina	13	7	0	0	20
LKKY	Exigl1	Exidia glandulosa v1.0	Agaricomycotina	19	1	0	0	20
PMTX	Rhivi1	Rhizopogon vinicolor AM-OR11-026 v1.0	Agaricomycotina	11	8	0	0	19
OAGC	Cerun2	Cerrena unicolor v1.1	Agaricomycotina	13	6	0	0	19
PMOD	Antsi1	Antrodia sinuosa v1.0	Agaricomycotina	14	5	0	0	19
QHHM	Mutel1	Mutinus elegans ME.BST v1.0	Agaricomycotina	16	3	0	0	19
NDTR	Leumo1	Leucogyrophana mollusca KUC20120723A-06 v1.0	Agaricomycotina	17	2	0	0	19
EHFU	Suigr1	Suillus granulatus EM37 v1.0	Agaricomycotina	17	2	0	0	19
FYAN	Cersu1	Ceriporiopsis (Gelatorporia) subvermispora B	Agaricomycotina	8	10	0	0	18
DGYS	Obbri1	Obba rivulosa 3A-2 v1.0	Agaricomycotina	10	8	0	0	18
OTUO	Xerba1	Xerocomus badius 84.06 v1.0	Agaricomycotina	12	6	0	0	18
URDR	Gymch1	Gymnopilus chrysopellus PR-1187 v1.0	Agaricomycotina	13	5	0	0	18
NMWS	Conpu1	Coniophora puteana v1.0	Agaricomycotina	16	2	0	0	18
QDCF	Pycsa1	Pycnoporus sanguineus BRFM 1264 v1.0	Agaricomycotina	10	7	0	0	17
GBII	Trave1	Trametes versicolor v1.0	Agaricomycotina	11	6	0	0	17
ZDTE	Amamu1	Amanita muscaria Koide v1.0	Agaricomycotina	12	5	0	0	17
DOXV	Guyne1	Guyanagaster necrorhiza MCA 3950 v1.0	Agaricomycotina	12	5	0	0	17
FJVL	Lenti7_1	Lentinus tigrinus ALCF2SS1-7 v1.0	Agaricomycotina	13	4	0	0	17
QMZJ	Plicr1	Plicaturopsis crispa v1.0	Agaricomycotina	5	11	0	0	16
ZQJS	Pycco1662_1	Pycnoporus coccineus CIRM1662	Agaricomycotina	9	7	0	0	16
SAKG	Fompi3	Fomitopsis pinicola FP-58527 SS1 v3.0	Agaricomycotina	11	5	0	0	16
NWCG	Ricme1	Peniophora aff. cinerea v1.0	Agaricomycotina	11	5	0	0	16
FSVD	PosplRSB12_1	Postia placenta MAD-698-R-SB12 v1.0	Agaricomycotina	12	4	0	0	16
XPNY	Dicsq1	Dichomitus squalens v1.0	Agaricomycotina	12	4	0	0	16
JWAZ	Fomme1	Fomitiporia mediterranea v1.0	Agaricomycotina	16	0	0	0	16
PXUU	Sisni1	Sistotremastrum niveocremaum HHB9708 ss-1 1.0	Agaricomycotina	16	0	0	0	16
HMXP	Sissu1	Sistotremastrum suecicum v1.0	Agaricomycotina	16	0	0	0	16
CKWM	Triab1_1	Trichaptum abietinum v1.0	Agaricomycotina	16	0	0	0	16
PELB	Artel1	Artolenzites elegans CIRM1663 v1.0	Agaricomycotina	8	7	0	0	15
OCCJ	Armme1_1	Armillaria mellea	Agaricomycotina	9	6	0	0	15
TDSN	Daequ1	Daedalea quercina v1.0	Agaricomycotina	10	5	0	0	15
USCC	PleosPC15_2	Pleurotus ostreatus PC15 v2.0	Agaricomycotina	10	5	0	0	15
JGOT	Traci1	Trametes cingulata BRFM 1805 v1.0	Agaricomycotina	10	5	0	0	15
MBVZ	Hetan2	Heterobasidion annosum v2.0	Agaricomycotina	11	4	0	0	15
OCJC	Lopni1	Peniophora sp. CONTA v1.0	Agaricomycotina	11	4	0	0	15
JEWI	Suibr1	Suillus brevipes v1.0	Agaricomycotina	11	4	0	0	15
FHPK	Lenti6_1	Lentinus tigrinus ALCF2SS1-6 v1.0	Agaricomycotina	12	3	0	0	15
YVJV	Sisbr1	Lentinus tigrinus v1.0	Agaricomycotina	12	3	0	0	15
BCXG	Schpa1	Schizopora paradoxa KUC8140 v1.0	Agaricomycotina	15	0	0	0	15
BSDJ	Punst1	Punctularia strigosozonata v1.0	Agaricomycotina	8	4	0	2	14
YBWF	Wolco1	Wolfiporia cocos MD-104 SS10 v1.0	Agaricomycotina	8	6	0	0	14
PKON	Pospl1	Postia placenta MAD 698-R v1.0	Agaricomycotina	9	5	0	0	14
TVNE	PleosPC9_1	Pleurotus ostreatus PC9 v1.0	Agaricomycotina	9	5	0	0	14
QZVB	Pycco1	Pycnoporus coccineus BRFM 310 v1.0	Agaricomycotina	9	5	0	0	14

Table A.2: Continued

Code	JGI_id	Species	Clade	α	TRI5	α - α 1	CPS/KS	Total
MRRS	Panru1	Panus rudis PR-1116 ss-1 v1.0	Agaricomycotina	11	3	0	0	14
HEVO	Polbr1	Polyporus brumalis BRFM 1820 v1.0	Agaricomycotina	11	3	0	0	14
ZQBK	Glotr1_1	Gloeophyllum trabeum v1.0	Agaricomycotina	12	2	0	0	14
PCAN	Stehi1	Stereum hirsutum FP-91666 SS1 v1.0	Agaricomycotina	13	1	0	0	14
MWLD	Aurvu1	Auriscalpium vulgare FP105234-Sp v1.0	Agaricomycotina	14	0	0	0	14
SDNQ	Aurde3_1	Auricularia subglabra v2.0	Agaricomycotina	9	4	0	0	13
DLNR	Fibra1	Fibroporia radiculosa TFFH 294	Agaricomycotina	9	4	0	0	13
SUPT	Boled1	Boletus edulis v1.0	Agaricomycotina	12	1	0	0	13
DBVV	Neole1	Neolentinus lepideus v1.0	Agaricomycotina	12	1	0	0	13
QQOC	Phlbr1	Phlebia brevispora HHB-7030 SS6 v1.0	Agaricomycotina	7	5	0	0	12
PFYJ	Tulca1	Tulasnella calospora AL13/4D v1.0	Agaricomycotina	8	4	0	0	12
TDAC	Hypsu1	Hypholoma sublateralitium v1.0	Agaricomycotina	9	3	0	0	12
XPDP	Tra1j1	Trametes ljubarskyi CIRM1659 v1.0	Agaricomycotina	9	3	0	0	12
RIYR	Polar1	Polyporus arcularius v1.0	Agaricomycotina	10	2	0	0	12
DKBC	Jaar1	Jaapia argillacea v1.0	Agaricomycotina	12	0	0	0	12
HZVO	Ramac1	Ramaria rubella (R. acris) UT-36052-T v1.0	Agaricomycotina	7	4	0	0	11
XMTQ	Elmca1	Aporpium caryae L-13461	Agaricomycotina	8	3	0	0	11
RJGY	Suilu1	Suillus luteus UH-Slu-Lm8-n1 v1.0	Agaricomycotina	8	3	0	0	11
RFWF	Cylto1	Cylindrobasidium torrendii v1.0	Agaricomycotina	9	2	0	0	11
NNLT	Gyman1	Gymnopus androsaceus JB14 v1.0	Agaricomycotina	9	1	0	1	11
STXZ	Hydpi2	Hydnomerulius pinastris v2.0	Agaricomycotina	10	1	0	0	11
WLYJ	Macfu1	Macrolepiota fuliginosa v1.0	Agaricomycotina	10	1	0	0	11
RYSH	Rhisa1	Rhizopogon salebrosus TDB-379 v1.0	Agaricomycotina	10	1	0	0	11
NCZF	Onnsc1	Onnia scaura P-53A v1.0	Agaricomycotina	11	0	0	0	11
SUQO	Paxam1	Paxillus ammoniavirescens Pou09.2 v1.0	Agaricomycotina	4	6	0	0	10
QTDW	Phchr2	Phanerochaete chrysosporium RP-78 v2.2	Agaricomycotina	4	6	0	0	10
SUXE	Trace1	Trametopsis cervina CIRM-BRFM 1824 v1.0	Agaricomycotina	6	4	0	0	10
OVWY	Serla_varsha1	Serpula lacrymans var shastensis SHA21-2 v1.0	Agaricomycotina	7	1	0	2	10
AFVL	Volvo1	Volvariella volvacea V23	Agaricomycotina	7	3	0	0	10
ZNEB	Pycci1	Pycnoporus cinnabarinus BRFM 137	Agaricomycotina	8	2	0	0	10
XJPI	SerlaS7_3_2	Serpula lacrymans S7.3 v2.0	Agaricomycotina	9	0	0	1	10
SSIY	SerlaS7_9_2	Serpula lacrymans S7.9 v2.0	Agaricomycotina	9	0	0	1	10
RAHU	Phaca1	Phanerochaete carnosa HHB-10118-Sp v1.0	Agaricomycotina	4	5	0	0	9
JWFA	Paxin1	Paxillus involutus ATCC 200175 v1.0	Agaricomycotina	5	4	0	0	9
CNHF	Bjead1_1	Bjerkandera adusta v1.0	Agaricomycotina	6	3	0	0	9
YLAB	Ompol1	Omphalotus olearius	Agaricomycotina	7	2	0	0	9
PTDW	Lacam1	Laccaria amethystina LaAM-08-1 v1.0	Agaricomycotina	7	2	0	0	9
SAAH	Lacbi2	Laccaria bicolor v2.0	Agaricomycotina	7	2	0	0	9
LNRR	Thega1	Thelephora ganbajun P2 v1.0	Agaricomycotina	9	0	0	0	9
RRJW	Agabi_varbur_1	Agaricus bisporus var. burnettii JB137-S8	Agaricomycotina	6	2	0	0	8
QVXG	Fibsp1	Fibulorhizoctonia sp. CBS 109695 v1.0	Agaricomycotina	4	3	0	0	7
XWAP	Paxru1	Paxillus rubicundulus Ve08.2h10 v1.0	Agaricomycotina	4	3	0	0	7
UZKW	Panst_KUC8834_1_1	Panellus stipticus KUC8834 v1.1	Agaricomycotina	5	2	0	0	7
WOTK	Agabi_varbisH97_2	Agaricus bisporus var bisporus (H97) v2.0	Agaricomycotina	6	1	0	0	7
CQIM	Copci1	Coprinopsis cinerea	Agaricomycotina	6	1	0	0	7
ETNA	Copci_AmutBmut1	Coprinopsis cinerea AmutBmut pab1-1 v1.0	Agaricomycotina	6	1	0	0	7
OABX	Phlgi1	Phlebiopsis gigantea v1.0	Agaricomycotina	6	1	0	0	7
RDKM	Trima3	Tricholoma matsutake 945 v3.0	Agaricomycotina	6	1	0	0	7
NUVL	Hebcy2	Hebeloma cylindrosporum h7 v2.0	Agaricomycotina	7	0	0	0	7
YDNI	Pilcr1	Piloderma croceum F 1598 v1.0	Agaricomycotina	7	0	0	0	7
ONZG	Pismi1	Pisolithus microcarpus 441 v1.0	Agaricomycotina	7	0	0	0	7
DJWQ	Gyrli1	Gyrodont lividus BX v1.0	Agaricomycotina	5	1	0	0	6
ARTW	Botbo1	Botryobasidium botryosum v1.0	Agaricomycotina	6	0	0	0	6
ZKED	Pisti1	Pisolithus tinctorius Marx 270 v1.0	Agaricomycotina	6	0	0	0	6
CWSV	Rhiso1	Rhizoctonia solani AG-1 IB	Agaricomycotina	6	0	0	0	6
PDMQ	Selci1	Scleroderma citrinum Foug A v1.0	Agaricomycotina	6	0	0	0	6
FVKT	Amath1	Amanita thiersii Skay4041 v1.0	Agaricomycotina	4	1	0	0	5
OMBC	Fishe1	Fistulina hepatica v1.0	Agaricomycotina	4	0	0	0	4
YXNI	Leugo1_1	Leucoagaricus gongylophorus	Agaricomycotina	2	1	0	0	3
KXQT	Calco1	Calocera cornea v1.0	Agaricomycotina	3	0	0	0	3
AEOE	Calvi1	Calocera viscosa v1.0	Agaricomycotina	3	0	0	0	3

Table A.2: Continued

Code	JGI_id	Species	Clade	α	TRI5	α - α 1	CPS/KS	Total
KNQF	Dacsp1	Dacryopinax sp. DJM 731 SSP1 v1.0	Agaricomycotina	3	0	0	0	3
FDSM	Schco3	Schizophyllum commune H4-8 v3.0	Agaricomycotina	3	0	0	0	3
WSPI	Schco_LoeD_1	Schizophyllum commune Loenen D v1.0	Agaricomycotina	3	0	0	0	3
XIIL	Schco_TatD_1	Schizophyllum commune Tattone D v1.0	Agaricomycotina	3	0	0	0	3
PYHT	Pirin1	Piriformospora indica DSM 11827 from MPI	Agaricomycotina	1	1	0	0	2
JYHG	CerAGI	Ceratobasidium sp. (anastomosis group I, AG-I) v1.0	Agaricomycotina	2	0	0	0	2
BFSU	ClaPMI390	Clavulina sp. PMI_390 v1.0	Agaricomycotina	2	0	0	0	2
EPBK	Monpe1_1	Moniliophthora perniciosa FA553	Agaricomycotina	1	0	0	0	1
UAVJ	Sebve1	Sebacina vermifera MAFF 305830 v1.0	Agaricomycotina	1	0	0	0	1
TEXE	Walse1	Wallemia sebi v1.0	Agaricomycotina	1	0	0	0	1
YDCR	Basun1	Basidioascus undulatus	Agaricomycotina	0	0	0	0	0
IBEU	Cryne_JEC21_1	Cryptococcus neoformans var neoformans JEC21	Agaricomycotina	0	0	0	0	0
FWMR	Cryne_H99_1	Cryptococcus neoformans var. grubii H99	Agaricomycotina	0	0	0	0	0
BBWI	Cryvi1	Cryptococcus vishniacii v1.0	Agaricomycotina	0	0	0	0	0
MSAX	Diocr1	Dioszegia cryoxerica v1.0	Agaricomycotina	0	0	0	0	0
HMCi	Hydru2	Hydnum rufescens UP504 v2.0	Agaricomycotina	0	0	0	0	0
GZLM	Treme1	Tremella mesenterica Fries v1.0	Agaricomycotina	0	0	0	0	0
GHXA	Trich1	Trichosporon chialellii MYA-4694 v1.0	Agaricomycotina	0	0	0	0	0
KWWA	Triol1	Trichosporon oleaginosus IBC0246 v1.0	Agaricomycotina	0	0	0	0	0
CYOI	Walic1	Wallemia ichthyophaga EXF-994	Agaricomycotina	0	0	0	0	0
BXWZ	Catan1	Catenaria anguillulae PL171 v1.0	Blastocladiomycota	0	0	0	0	0
SKYA	Batde5	Batrachochytrium dendrobatidis JAM81 v1.0	Chytridiomycota	0	0	0	0	0
QRGY	Ganpr1	Gonapodya prolifera v1.0	Chytridiomycota	0	0	0	0	0
LQKN	Rozal1_1	Rozella allomycis CSF55	Cryptomycota	0	0	0	0	0
DGLV	Corca1	Corynespora cassicola CCP v1.0	Dothideomycetes	2	1	1	5	9
PINU	Cocvi1	Cochliobolus victoriae FI3 v1.0	Dothideomycetes	3	1	3	2	9
DKTZ	CocheC4_1	Cochliobolus heterostrophus C4 v1.0	Dothideomycetes	4	0	4	1	9
OCCK	CocheC5_3	Cochliobolus heterostrophus C5 v2.0	Dothideomycetes	4	0	4	1	9
AXMY	Bysci1	Byssothecium circinans CBS 675.92 v1.0	Dothideomycetes	2	1	0	5	8
TNXD	Cocca1	Cochliobolus carbonum 26-R-13 v1.0	Dothideomycetes	3	0	2	3	8
JMZC	CocheC5_1	Cochliobolus heterostrophus C5	Dothideomycetes	3	0	4	1	8
XJJN	Amnl1i	Amniculicola lignicola CBS 123094 v1.0	Dothideomycetes	2	0	0	5	7
HIBM	Melpu1	Melanomma pulvis-pyrius v1.0	Dothideomycetes	2	0	0	5	7
NWTK	Coca1	Cochliobolus sativus ND90Pr v1.0	Dothideomycetes	3	1	2	1	7
XAAG	Psehy1	Pseudovirgaria hyperparasitica CBS 121739 v1.0	Dothideomycetes	4	0	1	2	7
FKOG	Tryvi1	Trypethelium eluteriae v1.0	Dothideomycetes	5	0	0	2	7
ZFEH	Botdo1_1	Botryosphaeria dothidea	Dothideomycetes	6	0	1	0	7
HCWQ	Linin1	Lindgomyces ingoldianus ATCC 200398 v1.0	Dothideomycetes	2	0	0	4	6
AEZQ	Aaoar1	Aaosphaeria arxii CBS 175.79 v1.0	Dothideomycetes	3	0	0	3	6
UMGD	Rhyru1_1	Rhytidhysterium rufulum	Dothideomycetes	3	0	0	3	6
SYTJ	Neopa1	Neofusicoccum parvum UCRNP2	Dothideomycetes	4	0	1	1	6
FXQO	Ophdi1	Ophiobolus disseminans CBS 113818 v1.0	Dothideomycetes	5	0	0	1	6
CGEC	Zoprh1	Zopfia rhizophila v1.0	Dothideomycetes	0	1	0	4	5
ERHT	Lopma1	Lophiostoma macrostomum v1.0	Dothideomycetes	2	0	1	2	5
RDUF	Cocmi1	Cochliobolus miyabeanus ATCC 44560 v1.0	Dothideomycetes	2	0	1	2	5
LQLG	Karrh1	Karstenula rhodostoma CBS 690.94 v1.0	Dothideomycetes	2	0	0	3	5
FSCX	Aplpr1	Aplosporella prunicola CBS 121.167 v1.0	Dothideomycetes	3	2	0	0	5
XLMY	Lenfl1	Lentithecium fluviale v1.0	Dothideomycetes	1	1	0	2	4
XMUQ	Trepe1	Trematosphaeria pertusa CBS 122368 v1.0	Dothideomycetes	1	1	0	2	4
USYT	Glost2	Glomium stellatum CBS 207.34 v1.0	Dothideomycetes	1	0	1	2	4
EGUL	Plesi1	Pleomassaria siparia v1.0	Dothideomycetes	1	0	1	2	4
RKGA	Dotsy1	Dothidotthia symphoricarpi v1.0	Dothideomycetes	2	1	0	1	4
TOST	Altbr1	Alternaria brassicicola	Dothideomycetes	3	0	1	0	4
CWHF	Stasp1	Stagonospora sp. SRC1lsM3a v1.0	Dothideomycetes	3	1	0	0	4
ZLLL	Cenge3	Cenococcum geophilum 1.58 v2.0	Dothideomycetes	0	0	1	2	3
XTCE	Polfu1	Polyplospora fusca CBS 125425 v1.0	Dothideomycetes	0	0	0	3	3
QGZN	Pyrrt1	Pyrenophora teres f. teres	Dothideomycetes	1	0	1	1	3
JOFG	Bauco1	Baudoinia compniacensis UAMH 10762 (4089826) v1.0	Dothideomycetes	1	1	1	0	3
YXFC	Bimnz1	Bimuria novae-zelandiae CBS 107.79 v1.0	Dothideomycetes	1	0	1	1	3
JVUN	Stano2	Stagonospora nodorum SN15 v2.0	Dothideomycetes	1	1	0	1	3
MORC	Coclu2	Cochliobolus lunatus m118 v2.0	Dothideomycetes	2	0	1	0	3

Table A.2: Continued

Code	JGI_id	Species	Clade	α	TRI5	α - α 1	CPS/KS	Total
JSQJ	Macph1	Macrophomina phaseolina MS6	Dothideomycetes	2	0	1	0	3
WIWL	Altal1	Alternaria alternata SRC1lrK2f v1.0	Dothideomycetes	3	0	0	0	3
KROL	Sepmu1	Septoria musiva SO2202 v1.0	Dothideomycetes	3	0	0	0	3
CYWZ	Seppo1	Septoria populicola v1.0	Dothideomycetes	3	0	0	0	3
OWZC	Delco1	Delitschia confertaspora ATCC 74209 v1.0	Dothideomycetes	0	0	1	1	2
GEAK	Mycfi2	Mycosphaerella fijiensis v2.0	Dothideomycetes	0	0	1	1	2
TQZK	Setho1	Setomelanomma holmii CBS 110217 v1.0	Dothideomycetes	0	1	0	1	2
SCZI	Spofi1	Sporormia fimetaria v1.0	Dothideomycetes	0	0	0	2	2
FARK	Venpi1	Venturia pirina	Dothideomycetes	0	1	0	1	2
KMFJ	Aurpu_var_mel1	Aureobasidium pullulans var. melanogenum CBS 110374	Dothideomycetes	1	1	0	0	2
KCAO	Clael1	Clathrospora elyanae CBS 161.51 v1.0	Dothideomycetes	1	0	0	1	2
RHMD	Decga1	Decorospora gaudefroyi v1.0	Dothideomycetes	1	0	0	1	2
MHQQ	Disac1	Dissoconium aciculare v1.0	Dothideomycetes	1	1	0	0	2
VWIU	Leppa1	Lepidopterella palustris v1.0	Dothideomycetes	1	0	0	1	2
VRZF	Maseb1	Massarina eburnea CBS 473.64 v1.0	Dothideomycetes	1	0	0	1	2
EBZG	Mycgr3	Mycosphaerella graminicola v2.0	Dothideomycetes	1	0	0	1	2
TIPI	Settu1	Setosphaeria turcica Et28A v1.0	Dothideomycetes	1	1	0	0	2
CZYK	Settur1	Setosphaeria turcica NY001 v1.0	Dothideomycetes	1	1	0	0	2
UFYD	Ternu1	Teratosphaeria nubilosa CBS 116005 v1.0	Dothideomycetes	1	0	0	1	2
WTQZ	Hyspu1_1	Hysterium pulicare	Dothideomycetes	2	0	0	0	2
CIVI	Lepmu1	Leptosphaeria maculans	Dothideomycetes	2	0	0	0	2
XDMT	Lopmy1	Lophium mytilinum CBS 269.34 v1.0	Dothideomycetes	2	0	0	0	2

Vita

Qidong Jia was born in Zhengzhou, the provincial capital of Henan Province. He obtained his Bachelor's degree in plant protection and Master's degree in bioinformatics at the Nanjing Agricultural University, China. He was then accepted to the Graduate School of Genome Science and Technology, a multidisciplinary joint program between the University of Tennessee Knoxville and Oak Ridge National Laboratory, to further pursue his research in the field of bioinformatics and computational biology. He also earned a Master's degree in Statistics and a Minor in Computational Science at the University of Tennessee.