Investigations into the Molecular Evolution of Plant Terpene, Alkaloid, and Urushiol Biosynthetic Enzymes

Alexandra J. Weisberg

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Bioinformatics, and Computational Biology

John G. Jelesko, Chair

David R. Bevan

Florian D. Schubot

Dorothea B. Tholl

Liqing Zhang

May 7th, 2014

Blacksburg, Virginia

Keywords: molecular evolution, plant specialized metabolism, urushiol, caffeine, nicotine, N-methyltransferase, alkaloid biosynthesis, protein engineering, terpene synthase, natural selection

Investigations into the Molecular Evolution of Plant Terpene, Alkaloid, and Urushiol Biosynthetic Enzymes

Alexandra J. Weisberg

ABSTRACT

Plants produce a vast number of low-molecular-weight chemicals (so called secondary or specialized metabolites) that confer a selective advantage to the plant, such as defense against herbivory or protection from changing environmental conditions. Many of these specialized metabolites are used for their medicinal properties, as lead compounds in drug discovery, or to impart our food with different tastes and scents. These chemicals are produced by various pathways of enzyme-mediated reactions in plant cells. It is suspected that enzymes in plant specialized metabolism evolved from those in primary metabolism. Understanding how plants evolved to produce these diverse metabolites is of primary interest, as it can lead to the engineering of plants to be more resistant to both biotic and abiotic stress, or to produce more complex small molecule compounds that are difficult to derive.

To that end, the first objective was to develop a schema for rational protein engineering using meta-analyses of a well-characterized sesquiterpene synthase family encoding two closely-related but different types of enzymes, using quantitative measures of natural selection on amino-acid positions previously demonstrated as important for neofunctionalization between two terpene synthase gene families. The change in the nonsynonymous to synonymous mutation rate ratio (d_N/d_S) between these two gene families was large at the sites known to be responsible for interconversion. This led to a metric ($\Delta d_N/d_S$) that might have some predictive power. This natural selection-oriented approach was tested on two related enzyme families involved in either nicotine/tropane alkaloid biosynthesis (putrescine N-methyltransferase) or primary metabolism (spermidine synthase) by attempting to interconvert a spermidine synthase to encode putrescine N-methyltransferase activity based upon past patterns of natural selection. In contrast to the *HPS*/*TEAS* system, using $\Delta d_N/d_S$ metrics between *SPDS* and *PMT* and site directed mutagenesis of SPDS did not result in the desired neofunctionalization to PMT activity.

Phylogenetic analyses were performed to investigate the molecular evolution of plant N-methyltransferases involved in three alkaloid biosynthetic pathways. The results from these studies indicated that unlike O-MTs that show monophyletic origins, plant N-MTs showed patterns indicating polyphyletic origins.

To provide the foundation for future molecular-oriented studies of urushiol production in poison ivy, the complete poison ivy root and leaf transcriptomes were sequenced, assembled, and analyzed.

This work was supported in part by the Genetics, Bioinformatics, and Computational Biology program, the department of Plant Pathology, Physiology, and Weed Science, and the Molecular Plant Sciences program at Virginia Tech; as well as a Fralin Life Science Institute and Virginia Bioinformatics Institute NextGen DNA sequencing award.

Acknowledgments

There are numerous people I would like to thank for their guidance and support over the past 5 years, and if I were to list them all here this dissertation would be a lot longer. First of all I would like to thank my advisor for teaching, guiding, and taking a chance on a student like me. John has taught me so much about molecular biology, evolution, lab technique, and most of all how to be a better scientist. His encouragement made it much easier to keep going when experiments didn't work or when I was feeling particularly pessimistic, and I am so thankful for his constant support and for being able to learn and grow in his lab.

I am also grateful to my advisory committee for their insights and helpful discussions, and I would also like to thank Dr. Beth Grabau for her support of my graduate studies.

I would of course also like to thank my family for all the love and support they have shown me, in both the good times and not so good times. I never would have made it this far without you. My parents and brothers have been anchors in my life, and I greatly admire their moral standing, their patience, and their capacity for love and kindness. Thank you for for pushing me to continue to grow as a person and for sticking with me from day one.

Finally, it would be impossible for me to not mention many of the members of the Molecular Plant Sciences community who have been there for me as friends, colleagues, and mentors; I would particularly like to thank Phoebe, Nicole, Megan, Gunjune, members of the McDowell lab (John, Dev, Ryan, Mike, Kevin, Bhadra), Shelley, Padma, Sherry, Ritesh, and Elise. I would also like to thank many of the friends I have made along the way who have all made a difference in my life: Sean, Susan, Eli, Jandi, Caroline, Andrew, Kara, Erik, Ellie, Jeremy, and Gabrielle. Words cannot express how grateful I am to have shared this time with you all.

Attribution

This dissertation is previously unpublished work by Alexandra J. Weisberg.

The experiments described were jointly designed by A. Weisberg and John G. Jelesko, and were mainly performed by A. Weisberg. The UPLC analysis in Chapter 3 was performed by Eva Collakova (Assistant Professor, Department of Plant Pathology, Physiology, and Weed Science at Virginia Tech) with assistance by A. Weisberg in performing recombinant enzyme assays. Plants used in Chapter 5 were germinated and grown by Elise Benhase (master's student, PPWS Department, Virginia Tech) and RNA was extracted and purified by J. Jelesko. Gunjune Kim (PhD candidate, PPWS Department, Virginia Tech) and James Westwood (Professor, PPWS Department, Virginia Tech) assisted with developing a *de novo* assembly software pipeline in Chapter 5. The *de novo* assembly pipeline in Chapter 5 was adapted from work published by Haas et al. 2013.

Contents

1 Introduction **1**

1.1 Plant chemodiversity 1

1.2 The evolution of plant secondary metabolism from primary metabolism 3

1.3 Enzymatic diversity 5

1.4 Enzyme evolutionary models 5

1.5 Protein structure and dynamics 11

1.6 Convergent evolution 12

1.7 Protein engineering 13

1.8 Quantifying natural selection 15

1.9 Research objectives 16

2 Identifying effectively-mutable protein space using measures of natural selection: a

meta-analysis of two terpene synthase gene families **18**

2.1 Abstract 19

2.2 Introduction 19

2.3 Methods 26

2.4 Results 27

2.5 Discussion 33

3 Investigations into molecular evolution-guided protein engineering of putrescine N-methyltransferase and spermidine synthase **39**

3.1 Abstract 40

3.2 Introduction 41

3.3 Methods 46

3.3.1 Phylogenetic analyses 46

3.3.2 Differing selective pressure on amino acid sites 47

3.3.3 Protein purification and expression 48

3.3.4 PMT enzyme assays 50

3.4 Results 51

3.5 Discussion 61

4 The polyphyletic molecular evolution of N-methyltransferases in plant alkaloid metabolism 67

4.1 Abstract 67

4.2 Introduction 68

4.2.1 Evolution of N-methyltransferases 68

4.2.2 Caffeine biosynthesis 70

4.3 Methods 73

4.3.1 Reciprocal homology searches 73

4.3.2 Phylogenetic analyses 73

4.4 Results 75

4.4.1 Polyphyly of plant N-methyltransferases 75

4.4.2 Molecular evolution of caffeine biosynthetic enzymes 78

4.5 Discussion 84

4.5.1 Polyphyletic origins of N-methyltransferases in alkaloid metabolism 84

4.5.2 Molecular evolution of caffeine biosynthesis 85

5 *De novo* assembly of the *Toxicodendron radicans* (poison ivy) root and leaf transcriptomes 88

5.1 Abstract 88

5.2 Introduction 89

5.3 Methods 93

5.3.1 Axenic cultured seedlings and plant tissue 93

5.3.2 RNA purification and sequencing 94

5.3.3 *De novo* assembly of transcripts 95

5.3.4 Comparison with mango transcriptome 96

5.3.5 Molecular cloning of type III polyketide synthase-like transcripts 96

5.3.6 Protein expression and purification 99

5.3.7 *Agrobacterium*-mediated transformation of *T. radicans* 100

5.4 Results 102

5.4.1 Assembly analysis/quality 102

5.4.2 Cloning of type III polyketide synthases 108

5.4.3 *Agrobacterium tumefaciens*-mediated poison ivy transient transformation . 115

5.5 Discussion 116

5.5.1 Future work 121

5.5.2 Conclusions 123

5.5.3 Acknowledgements 123

6 Conclusion **124**

6.1 Molecular evolution in plant specialized metabolism 124

6.2 Urushiol biosynthesis in *Toxicodendron radicans* 128

Bibliography **130**

List of Figures

2.1 TEAS and HPS pathways. 24

2.2 ML search tree of TEAS and HPS-like homologs with non-parametric bootstrap support values mapped to branches. 28

2.3 Phylogenetic analyses of TEAS and HPS-like sequences. 31

2.4 Distribution of d_N/d_S values across three datasets (*TEAS*-like genes, *HPS*-like genes, and both combined) calculated using the PAML M7 model. 32

2.5 3D crystal structure of Tobacco 5EAT (PDB: 5EAT) with $\Delta d_N/d_S$ values mapped to residues. 34

3.1 Putrescine N-methyltransferase and spermidine synthase enzyme activity. 44

3.2 PMT/SPDS maximum likelihood phylogeny. 52

3.3 Change in selective pressure mapped to SPDS structure. 56

3.4 Alignment of amino acid sequences for StPMT, StSPDS, and $StSPDS^{PMT14}$. 57

3.5 Purified recombinant protein for a. NtPMT with MBP, b. StSPDS, and c. StSPDS^{PMT14}. 58

3.6 Enzyme assays quantified by UPLC. 60

3.7 Maximum likelihood non-parametric bootstrap phylogenies. 64

4.1 The major caffeine biosynthetic pathway in coffee and tea plants. 71

4.2 Reciprocal similarity searches to identify distantly related homologs of N-methyltransferases. 77

4.3 Unrooted xanthosine N-methyltransferase maximum likelihood phylogeny. 79

4.4 Unrooted maximum likelihood phylogenies of clades containing xanthosine/xanthine N-methyltransferases. 81

4.5 PAML site model d_N/d_S values across codons in *Coffea* and *Camellia*. 83

5.1 Alkylphenols produced in plants. 92

5.2 Process for *de novo* RNA seq assembly and annotation. 103

5.3 Assembly length distributions. 104

5.4 Blast searches of *T. radicans* transcripts against NCBI nr. 107

5.5 Phylogenetic tree and expression levels of poison ivy probable type III polyketide synthases. 109

5.6 Unified urushiol and alkylresorcinol biosynthetic pathways. 110

5.7 Multiple sequence alignment of PKS-like1 and PKS-like2 protein sequences with other type III polyketide synthases. 111

5.8 Long and accurate PCR of *T. radicans* predicted type III polyketide synthases. . . . 113

5.9 Recombinant protein expression and purification for PKS-like1 and PKS-like2. . . 114

5.10 Transient expression of luciferase *in vivo*. 117

List of Tables

2.1 Likelihood ratio test of PAML site models for *TEAS* and *HPS*. 29

2.2 PAML site model analysis results for nine sites important for differentiating TEAS and HPS activity. 33

3.1 *PMT* and *SPDS* PAML site model results. 54

4.1 XMT alternate topology hypothesis testing. 80

4.2 PAML site model likelihood ratio tests (LRT) on xanthosine/xanthine N-methyltransferases from coffee and tea. 82

5.1 Paired read quality per sample. 105

5.2 *De novo* assembly statistics. 105

Chapter 1

Introduction

1.1 Plant chemodiversity

Plants produce an enormous number of low-molecular weight chemicals. The total number of plant metabolites has been estimated at over 200,000 [1], but may even number in the millions, as many more are constantly being discovered with additional surveys providing greater sensitivity [2] (E. Pichersky (personal communication)). The number of plant terpenes alone is estimated to be greater than 55,000 [3], while the number of plant alkaloids is somewhere around 12,000 [4]. Many of these chemicals have historically been used by humans for their medicinal properties, or as lead compounds in searches for novel drugs [5]. Others play a major role in the taste and smell of food and drink, as well as determine the color and scent of flowers in ornamental plants [6, 7].

While we know of the existence of many phytochemicals, their function in the plant is often far

less clear. Metabolites involved in plant growth and development are called primary metabolites. Given the fact that most eukaryotic cells utilize very similar molecular machinery for assembling viable cells, the spectrum of metabolites involved in primary metabolism is rather narrow and estimated to be roughly 10,000 known primary metabolites [8]. This constitutes a tiny fraction of the chemical diversity found among plants so far.

The majority of plant chemical diversity is believed to exist because they give some additional adaptive advantage to the organism, and therefore are called secondary or specialized metabolites [9]. The potential adaptive roles that secondary metabolites take are as varied as the ecological habitats and niches that plants can occupy. They may play an adaptive role in plant defense against herbivory, microbial pathogens, temperature regulation, or assist in allopathy against the growth of other plants growing nearby [10, 11, 12, 13, 14]. Other secondary metabolites protect the plant from UV radiation or desiccation [15]. Alternatively, secondary metabolites may promote favorable ecological interactions by attracting insect or animal pollinators through use of flower color, scent, or mimicking insect pheromones [7, 16, 17, 18].

The evolution of lignin and phenylpropanoid biosynthesis greatly assisted or even enabled the transition of life from aquatic to terrestrial habitats hundreds of millions of years ago [19, 20]. The evolution of phenylpropanoids, which absorb energy in the UV-B spectra, allowed organisms to protect themselves against UV radiation once out of the protective water environment [21]. The production of lignin then allowed these early plants to grow larger and transport water more efficiently by reinforcing the cell wall [19].

1.2 The evolution of plant secondary metabolism from primary metabolism

It has long been suspected that the genes involved in plant specialized metabolism arose from those responsible for primary metabolism [22]. This may have occurred through various processes, such as gene duplication followed by divergence. That is, a primary metabolic gene may have been duplicated in the genome, which reduced the selective pressure to maintain the initial activity in one of the copies [23]. The duplicated copy (paralog) then acquired mutations that changed the enzyme to perform some new activity (neofunctionalization) that provided a selective advantage to the plant.

Evidence for this process is plentiful, most notably in the large enzyme families that exist in plants. Duplicated genes often form adjacent tandem repeats, which may have similar functions or different expression patterns [24, 25]. Having multiple copies of a gene allows for more opportunities for neofunctionalization by allowing various enzymes to diverge to recognize either new substrates or perform different reactions with the original substrate. Examples of several large gene families that seem to have taken this evolutionary trajectory are the terpene synthase, O-methyltransferase, and cytochrome P450 families [26, 27, 28]. The large number of gene duplicates in plant specialized metabolism may also be due to the fact that secondary metabolic enzymes tend to have drastically slower activity (~30-fold less) than primary metabolic enzymes [29]. Some secondary metabolic enzymes may be expressed at higher levels to compensate for low activity using gene duplicates, while others are only expressed in specific tissues at certain times, often in response to external

stimuli [30]. The metabolites produced by such enzymes can accumulate in tissues at up to several percent of the total dry weight [31, 32]. There are numerous examples of these principles, but only a few exemplars will be outlined below because they provide especially clear insights into these processes.

A well-characterized example of a secondary metabolite pathway evolving from a primary metabolite one is in pyrrolizidine alkaloid biosynthesis. The enzyme encoding homospermidine synthase (HSS) appears to have diverged from the primary metabolic enzyme deoxyhypusine synthase (DHS) after a duplication of the ancestral DHS gene [33, 34]. DHS is a bifunctional enzyme performing both HSS activity and the modification of eukaryotic initiation factor 5A (eIF-5A), an activity performed in all forms of life (Ober, 2003). Both enzymes have homospermidine synthase activity, but only DHS can catalyze the deoxyhypusine synthase activity. HSS is presumed to have formed as a duplicated DHS enzyme that subsequently lost the initiation factor modification activity, but retained HSS activity.

Genes involved in secondary metabolism may also give rise to yet other specialized metabolites through duplication of previously existing secondary metabolic genes followed by divergence of paralogs. For example, three tandem clustered terpene synthase genes in *Arabidopsis thaliana* share high cDNA sequence identity (78-100%), with two of the copies being identical to each other. The identical copies both express in roots and catalyze the production of 1,8-cineole, while the third is only found in flowers and produces myrcene and (E)-β-ocimene. All three share identical intron structure, suggesting that they arose through recent duplication of paralogs followed by divergence [35].

1.3 Enzymatic diversity

Plant genomes contain anywhere from 20,000-60,000 genes. Of these, roughly 20% of a given plant genome encodes enzymes that appear to be involved in plant specialized metabolism [36, 37]. This suggests that all of the enzymes used to produce even the 200,000 known secondary metabolites are not present in any one plant genome. Rather, plant chemical diversity overall is large because the genetic diversity of plant enzymes involved in specialized metabolism is distributed across many different plant species, each containing different metabolite profiles [22, 38, 39]. Due to the principle of one enzyme- one mechanism, it is estimated that there are hundreds of thousands of unique enzymes involved in plant specialized metabolism among all plants [22]. While the number of individual enzymes in specialized metabolism is large, the total number of unique protein folds is in actuality much smaller [15].

1.4 Enzyme evolutionary models

Many of the reactions producing metabolites in plants are known to be catalyzed by at most one enzyme each [19]. Enzymes are proteins that catalyze particular chemical reactions in a cell. Enzymes are comprised of a primary sequence of amino acids (AAs) synthesized into a polypeptide that subsequently folds into a three dimensional structure. This structure facilitates the interaction of groups of AAs in the enzyme to interact with specific small molecule chemical substrates, bringing them together in close-proximity (or modify a single bound substrate) thereby lowering the activation energy of specific chemical reactions and enhancing the rate of that chemical reaction

[40]. The shape of the active site cavity in these proteins, along with specific amino acids within that cavity, often determines the type of substrates that can be acted upon. These proteins are encoded by individual genes. Specifically, each amino acid (AA) is coded by at least one nucleotide triplet sequence, called a codon. The universal genetic code is degenerate, since there are 4^3 (64) possible codons, yet there are only 21 possible amino acids and three stop codons. This redudancy of the universal genetic code allows for some protection against mutations (particularly at the third nucleotide in a codon), as each amino acid may be coded for by up to six codons. Genetic changes/mutations in these codons may give rise to changes in the corresponding encoded amino acid in the protein, providing the basis by which evolutionary change occurs.

There are several evolutionary models to explain the origins of plant secondary metabolism. Each of these models predicts how changes in DNA sequence results in changes affecting enzyme function. Gene duplication is an important aspect of many of these models, providing DNA under less strong selective pressure than the original gene in which neofunctionalization can occur. Mutations in genes may occur that confer new enzyme function, such as the recognition of new substrates or the production of new products from the original substrate. Where these models differ is the relative order in which these events happen, and the relative importance of when and how natural selection drives these changes to either fixation or balancing selection. Adaptive or Neo-Darwinian models place great importance on Darwinian natural selection driving neofunctionalization, particularly adaptive/positive selection [28]. Other models instead claim that genetic drift (mutations randomly fixing in small populations) is responsible for most genetic diversity and the main driving force in the rise of new enzyme function [41, 42, 43]. Yet other models suggest some combination

of the two, such as those developed by Tawfik [44, 45, 46, 47, 48].

Under adaptive/Darwinian models of evolution, a gene must acquire a mutation providing a new function relatively quickly after duplication in order to prevent the much more common deleterious mutations from rapidly destroying protein stability or otherwise inactivating the protein. If this new function provides a phenotype that is selectively advantageous, Darwinian (positive) natural selection will cause the underlying mutation to become fixed in a population. Anywhere from 33-40% of mutations may be selectively deleterious [45, 49, 50]; and experiments have shown that protein stability declines exponentially with the addition of more and more mutations [45, 51]. These deleterious mutations primarily affect protein stability, resulting in mis-folded or denatured protein aggregates that reduce the amount of soluble, correctly folded protein in a cell [52]. In these models, the rare mutation that is not deleterious but instead selectively advantageous could cause the encoded protein to perform a new catalytic function, albeit at very low levels. If this function provides a selective advantage to the plant, natural selection will cause more mutations to come to fixation that gradually increase the new activity of the encoded enzyme or increase gene expression.

Kimura's neutral model of evolution instead suggests that genetic drift is responsible for the majority of evolutionary change [41]. Ohta furthered the idea, suggesting that most amino acid substitutions are instead slightly deleterious rather than completely neutral, and are influenced by both natural selection and genetic drift [43]. Under this model, a small number of mutations may either be selectively deleterious (under purifying or negative selection), or may be advantageous and selected for (positive or adaptive selection); however most mutations are selectively neutral

or nearly-neutral (slightly deleterious). Such nearly-neutral mutations are neither extremely deleterious nor extremely advantageous, but rather have only slightly negative effects on structure or function. Kimura initially showed that the observed rate of amino acid substitution is too high for mammalian species to tolerate if the vast majority of mutations are deleterious; therefore most mutations must be neutral [41]. It is known that many proteins could have regions with many substitutions with almost no negative effect on function [42]. A selective sweep may also fix certain variations in a population if they are located near a site under positive selection [53]. In many cases there is a balancing-act aspect of natural selection. That is, there may be conditions or environments where a particular allele is advantageous, but in other environments a different allele is more advantageous. In these cases selective sweeps can be a disadvantage, as the organism may need both enzyme functions depending upon which of several different habitats it might potentially occupy. Not selecting entirely for either function (and thus lose the other) may be a compromise for withstanding changing environmental conditions.

The traditional notion of ”one enzyme- one mechanism- one chemical product” or highly specialized enzymes may not be an entirely accurate model for understanding protein evolution, especially the evolution of plant specialized metabolism. Neo-Darwinian models tend to put great importance on small changes with big effects, whereas neutral evolutionary models are exceedingly gradual, making it difficult to determine exactly how or when neofunctionalization occurs.

Many enzymes in plant specialized metabolism are catalytically promiscuous when compared with primary metabolic enzymes [53, 29, 54]. That is, they can catalyze a reaction type using a variety of different substrates, or they may produce a variety of different products from a single substrate.

On the other hand, enzymes in primary metabolism are often highly specialized for a specific substrate(s) and product, and form relatively ordered, linear pathways [15]. In contrast, secondary metabolic pathways tend to form a web or grid-like structure, with many enzymes feeding several metabolites into an array of alternative pathways [15].

Protein stability is also very important to enzyme activity. Proteins fold into a three-dimensional structure, typically driven by hydrophobic amino acid residues forming a hydrophobic core, and hydrophilic residues interacting with solvent on the outer surface. Mutations in genes that affect this stability, such as switching a hydrophobic residue with a hydrophilic one and vice versa, can cause a protein to mis-fold and form non-functional protein aggregates, reducing the level of soluble catalytically-active protein in a cell [52]. These mutations are typically not tolerated and selected against (by negative/purifying selection).

One apparent issue in our understanding of protein evolution lies in an apparent conflict between neutrality and enzyme neofunctionalization [55, 47]. That is, according to the theory of nearly-neutral evolution, (neutral or nearly-neutral) mutations can accumulate with little or no change in protein function. However, mutations are also expected to have large effects on protein function or stability (plasticity) due to the fact that they generally occur one at a time [47]. One suggested answer to this problem is that mutations can be acquired that are only mildly-deleterious to the native function, yet greatly increase the level of an altered enzyme activity, i.e. a more promiscuous enzyme. [44]. Similarly, some enzymes may be comprised of a highly-stable ”backbone” structure that can withstand multiple neutral mutations without losing overall protein stability, as well as a more flexible active site that is quickly evolvable [47].

One model proposed by Khersonsky et al. bridges the extremes of a few strong neo-Darwinian mutations with neutral mutations that actually lead to additional activities, without breaking the original activity. This model suggests that an enzyme can go from being a specialist in one enzymatic activity to become more of a generalist, expanding its repertoire of catalytic activities without losing the original function, often through a small number of mutations [44]. Changes in the organism's environment (and thus selective pressure) may later cause one of these new activities to be selectively advantageous, causing it to specialize, favoring the new activity and eventually lose the original enzyme activity [19]. Large increases in this new function (~1,000-fold) can be gained through a handful of mutations, with only a small (~3.2-fold) loss in the original function [44]. If gene duplication were to have occurred before this change of function, the original ortholog may have been under selective pressure to retain the original enzyme function, eventually resulting in two related genes (paralogs) encoding non-overlapping enzyme functions.

The range (or landscape) of acceptable mutations for protein stability, as pictured by DePristo et al. as well as Tokuriki and Tawfik, is comprised of a "window" of neutral space in which mutations can occur that are not strongly deleterious to organism fitness [56, 46]. As long as long as an enzyme stays within this neutral space, mutations that do not massively reduce stability are allowed to accumulate and fix in a population. A mutation that greatly reduces the stability of a protein may also cause it to form aggregates or lose function, and thus is selected against. Likewise, a mutation conferring too much stability can also be deleterious, as too much protein rigidity may decrease the enzyme's ability to move to accommodate substrates or catalyze reactions, and can also negatively affect metabolic regulation [46]. Some mutations may also be compensatory for

other more destabilizing mutations [57, 58, 59]. These mutations can occur before a destabilizing mutation to provide the necessary structural stability to remain in neutral space, or they can occur afterwards and ”rescue” the stability of the protein.

1.5 Protein structure and dynamics

More recently, the idea of a protein has changed from that of a single stable structure, to a collection of multiple, dynamic conformational states that includes the primary native state [60]. The native state represents that which has been previously selected for, while other, more promiscuous states may be a consequence of the structural flexibility needed to maintain the native function [61, 62]. The structure of these proteins can be affected by different things, such as the presence of the native substrate(s) or other potential substrates/binding targets, as well as physiological conditions. This flexibility may be constrained only to active site residues, such as loops important for catalytic activity [63, 64], or it may be a property of the entire enzyme fold [47]. While the amino acids in direct contact with the substrate may not be changed, other residues nearby may affect the overall turnover rate of the reaction by speeding up the process of product release [47]. Other proteins can shift entirely between multiple unrelated folded states, rearranging the entire structure [65]; such as has been observed in the protein lymphotactin [66].

It has also been noted that protein flexibility and enzyme activity promiscuity appear to be linked in some way [47]. This promiscuity tends to come at a cost, as more flexible enzymes are also inherently more unstable [67, 68]. For example, the cytochrome P450 CYP2B4 has a flexible active

site that can accommodate and act upon a wide range of substrates; forming an open conformation when binding large substrates, and a closed one when binding smaller substrates [69]. Other P450s may be fairly rigid and only recognize few substrates (such as CYP2A6) while others are much more flexible and accommodate a wide range of substrates (CYP3A4) [70]. Enzymes in primary metabolism tend to have very low measurable promiscuous activity, if any, and are usually structurally stable [15].

Many mutations that increase one promiscuous function may also increase the activity of other promiscuous functions. For example, in a study of two closely-related sesquiterpene synthases from Tobacco and Henbane (5-epi-aristolochene synthase, TEAS and premnaspirodiene synthase, HPS respectively), pairwise substitutions between the two not only changed activity from one enzyme type to the other in *in vitro* assays, but also resulted in new products not produced by either wild-type enzyme [71]. The native TEAS also produces up to 24 minor products in addition to 5-epi-aristolochene [72]. This promiscuity could be caused by more flexibility in the active site [67].

1.6 Convergent evolution

The above models are typical descent with modification of a basic chemical transformation with different substrates. That is, enzymes with one activity giving rise to enzymes performing similar activities. There is also growing evidence that some biochemical pathways have evolved several times independently in different plant lineages [8]. These pathways may evolve from enzymes per-

forming very different functions (convergent evolution) [73, 74]. These plants may use different, unrelated enzymes to catalyze the same type of chemical reaction, or they may use different substrates but arrive at the same product [8]. For example, hydroxynitrile lyase enzymes, responsible for the biosynthesis and release of the toxic chemical hydrogen cyanide in response to herbivory, have been shown to have evolved independently at least three times in different plant lineages [75].

In other cases, some plants may evolve the ability to synthesize metabolites that, while chemically or structurally different, produce a similar phenotype [8]. An example of this would be in metabolites responsible for flower pigmentation. Various anthocyanins, which (when accumulated) give flowers a reddish or blueish color, are found in many plant lineages [6]. Most of the members of one plant lineage do not contain anthocyanins, but achieve a similar flower color phenotype by producing betacyanins instead [76].

The convergent evolution in plant specialized metabolism may be a result of, or consequence of, the generally-more promiscuous nature of enzymes in specialized metabolic pathways. Pathways in specialized metabolism commonly form a grid (or web) of interacting enzyme-mediated reactions, rather than relatively-linear pathways as is found in primary metabolism [19].

1.7 Protein engineering

The principles of past molecular evolution events in plant specialized metabolism has broad applicability to rational protein engineering for desired improved characteristics. A common approach to evaluate the evolution of one enzyme activity to that of another utilizes comparative protein

multiple sequence alignments between two closely related but distinct enzyme families, each with a different substrate or product profile. The objective is to identify amino acids that are highly conserved (i.e. under purifying selection and presumably conferring enzyme specificity) within each group of enzymes, but that differ between the two groups. This two-dimensional analysis can be refined somewhat by examining the three-dimensional protein structural information to determine if these separately conserved amino acids are part of the substrate binding or active sites for the two differing enzymes. Site-directed mutagenesis and in vitro biochemical assays of recombinant protein can then be used to determine if these identified sites are important for enzyme specificity.

This approach can lead to the desired outcome [77]. However in a great many cases, usually unreported, these methods result in either destabilized or otherwise nonfunctional protein, indicating that the modified residues are necessary for one activity but are not the sole determinates necessary for the other [78, 79, 80, 81, 82]. Other approaches, such as three-dimensional protein structure contact-mapping strategies, have been somewhat more successful [72], but a comprehensive rule for successfully identifying these sites in all or many enzyme families has not yet been elucidated.

Since proteins found in plants are their modern-day incarnations, attempts to interconvert these enzymes may not be successful due to unknown important adaptive intermediates between extant and ancestral forms of the enzymes. These ancestral enzymes may have been more promiscuous than their modern day forms, which later came under selective pressure to perform a single activity at the expense of the other. Other groups have attempted to resurrect these ancient proteins through ancestral state reconstruction analyses [58]. The amino acid sequence at different splits in a phylogeny can be predicted using statistical models of protein evolution, with the aim of elucidating

the structure and/or catalytic activity of the ancestral form [78].

1.8 Quantifying natural selection

Natural selection is a common aspect to all evolutionary models. Ironically, natural selection is also typically the least characterized aspect of the evolution of a given gene family, for which there is only indirect quantifiable evidence. While this important attribute is generally unknown, it can be quantitatively modeled and thus measured. Methods of statistically quantifying the forces of natural selection on a gene family have been developed that measure the ratio of nonsynonymous to synonymous mutation rates (d_N/d_S) at individual codon triplets in a multiple sequence alignment (site models), at specific branches in a phylogenetic tree (branch models), or a combination of both (branch-site models) [83]. If there is a higher rate of synonymous mutations than nonsynonymous ones ($d_N/d_S < 1$), then fewer mutations are tolerated at that site, and that site can be said to be under purifying or negative selection. This site If the rates of nonsynonymous and synonymous mutations are relatively equal ($d_N/d_S = 1$), then that site is tolerating a variety of mutations and is undergoing neutral selection. Sites with a d_N/d_S greater than 1 are differentially acquiring new nonsynonymous mutations and are said to be under positive/adaptive/Darwinian selection. These methods have been used to detect positive selection [84] or purifying selection [85] in a population. One caveat is that these methods (calculating d_N/d_S ratios) are most relevant for detecting extant signals of natural selection pressure, but not necessarily ancient ones. Past selective sweeps or strong purifying selection in ancient genes may ”erase” sequence diversity in the extant genes,

limiting our ability to trace the path of evolution.

1.9 Research objectives

A major objective in this research was to develop a method for rational protein engineering that uses quantitative measures of the forces of natural selection [83] combined with 3D protein structural information to identify sites important for neofunctionalization within divergent gene families. This schema was developed using a meta-analysis of a well-characterized sesquiterpene synthase enzyme family (tobacco 5-epi-aristolochene synthase and henbane premnaspirodiene synthase) to determine important patterns of natural selection at nine sites previously known to be important for catalytic activity specificity [72, 71]. We then applied and attempted to refine our methodology on a gene family encoding a tropane alkaloid biosynthesis pathway enzyme (putrescine N-methyltransferase, PMT) and it's progenitor from primary metabolism (spermidine synthase, SPDS). The goal of this method was to identify sites that may not have been previously chosen based on naive examination of conserved differences in multiple sequence alignments, but still play a role in determining substrate or product specificity. The predictions from this approach were tested using assays of modified recombinant protein in an attempt to convert a SPDS to have PMT activity.

Another research objective focused on phylogenetic investigations into the possibility that plant N-methyltransferase enzymes involved in different alkaloid biosynthetic pathways evolved by convergent evolutionary trajectories. In order to thoroughly test this hypothesis, deep phylogenetic

searches were performed with the intent of finding any possible ancient common homologs between several different N-methyltransferases involved in nicotine, tropane, and nortropane alkaloid (nicotine/scopolamine/calystegines, etc), purine alkaloid (caffeine), and benzylisoquinoline alkaloid (sanguinarine/morphine) biosynthetic pathways.

The final objective, while somewhat separate from the previous objectives, focused on another plant secondary metabolite. Poison ivy is infamous for causing delayed allergenic contact dermatitis symptoms in humans. The small molecule secondary metabolite responsible for causing the characteristic poison ivy skin rash is called urushiol. Although the chemical composition of urushiol has been known for decades, the gene enzyme systems responsible for urushiol synthesis have not been identified. To this end, a fourth research objective focused on utilizing Next-Generation DNA sequencing to develop genetic resources suitable for investigating urushiol biosynthesis in poison ivy.

Chapter 2

Identifying effectively-mutable protein space using measures of natural selection: a meta-analysis of two terpene synthase gene families

Alexandra J Weisberg and John G Jelesko.

2.1 Abstract

Plant secondary (or specialized) metabolism consists of an enormous array of metabolites that likely provide a selectable benefit to the plant. Many of these plants have been used for thousands of years for medicinal purposes due to their ability to affect specific physiologies of various animals and microbes. Understanding how plant enzymes evolve to produce these complex small-molecule therapeutics and other economically important metabolites can benefit in the quest to rationally modify existing enzymes to produce novel compounds. To this end, a meta-analysis of two members of a biochemically well characterized terpene synthase gene family was performed to reveal patterns of past natural selection responsible for enzyme reaction product specificity. These studies showed that eight of the nine sites previously identified as important for product-specificity were under stronger negative selective pressure (had a smaller d_N/d_S) in members encoding premnaspirodiene synthase (HPS) activity compared with those encoding 5-epi-aristolochene synthase (TEAS) activity. These sites had a $\Delta d_N/d_S$ greater than 0.2. We propose that sites under different levels of purifying selection between two gene families (large $\Delta d_N/d_S$) may be good predictors of which sites impart neofunctionalization.

2.2 Introduction

Plants produce a vast number of small-molecule metabolites. While many metabolites are directly involved in plant growth and development (so-called primary metabolites), an even greater number of phytochemicals are presumably produced because they confer some selective advantage (known

as specialized or secondary metabolites). With currently known specialized metabolites numbering over 200,000 [8], and estimated to be over 1 million (Eran Pichersky, personal communication), these chemicals play a multitude of possible roles in plant survival. Specialized metabolites, such as alkaloids [86, 87, 88, 89, 90], cyanogenic glucosides [91], and terpenes [92], have been shown to provide defense against herbivores. They also play a role in temperature regulation and are responsible for flower color and scent [93]. Natural products compose of over 70% of anti-cancer drugs to date [94], and many plant metabolites have long been used by humans for their medicinal properties [95]. Many of these metabolites are structurally complex and are very difficult to synthesize synthetically because of many chiral centers. These plant natural products have been gradually shaped by evolutionary pressure over long time scales to keep up with changing targets and environments [96]. Understanding how plants evolved to produce these diverse chemicals has applications in many fields, particularly protein engineering.

There have been numerous attempts at engineering new enzyme activities or converting activities between closely related enzymes so far, with some more successful than others [97]. Various strategies have been employed, including ancestral state reconstruction [78, 98, 99], directed evolution experiments [100], contact mapping/3D structural analysis [72], as well as examination of comparative multiple sequence alignments (MSA) [82, 101]. Attempts often result in ”dead” or inactive enzymes with no new activity and loss of the original activity. Protein stability is often negatively affected as well, resulting in the formation of insolube aggregates and inclusion bodies. Additionally, when enzymes are successfully engineered to have new activity, their catalytic efficiency is often much less than that of wild-type enzymes.

Much work has gone into elucidating the path of neofunctionalization in proteins, and several models have been put forward in an attempt to explain it. Enzymes can gain new activities (albeit at much lower levels relative to the original) from relatively few mutations and with little loss of the original function [44]. However, increasing numbers of mutations causes protein instability [45, 51]. This dichotomy has made protein engineering difficult [55, 47]. One model of enzyme evolution proposes that enzymes transition from being more specialist, only recognizing one or few substrates, to being more generalist, by recognizing more substrates or perhaps showing an altered catalytic reaction profile[44]. If those new functions are selectively advantageous, the gene encoding that enzyme may undergo further mutations to then become a specialist for the new enzyme activity, possibly at the expense of losing the original activity. Gene duplication at a number of time points before or after this occurring can act to preserve the original function in one enzyme (ortholog) while allowing the other paralog to acquire new activities with little negative effect on the original ortholog's role in plant physiology [23]. Ohta's nearly neutral model of evolution states that most mutations are selectively neutral or very mildly deleterious (nearly neutral) [43, 102]. It is thought that ancestral enzymes had broad specificity, and later gene duplication and divergence is responsible for the large quantity and variety of current enzymes [103]. This ability for enzymes to rapidly acquire new activities may come from an inherent flexibility in protein structure [67].

There is a range of protein stability that is acceptable for enzyme activity. Mutations causing a protein to become too unstable will cause it to lose all activity and/or aggregate, while mutations causing too much stability will affect regulation as well as may reduce activity dependent on structural flexibility, and will also be selected against [47]. This window of selectively neu-

tral mutation space allows for several possibilities for acquiring new enzyme activity. Mutations causing a change of function may be mildly destabilizing. These mutations could be preceded or followed by compensatory mutations to regain stability. Chaperone proteins may also effectively "widen" the range of acceptable mutations by allowing more destabilizing mutations to occur and be tolerated [47].

In plants, the terpene synthases (TPS) form a large group of related secondary metabolic enzymes that produce upwards of 55,000 unique terpenes, all of which are derived from different numbers of conjugated isoprene units as substrate [3]. Most of these complex chemicals are involved in plant defense in various ways [104, 105]. Volatile terpenes may deter insects feeding on the plant or attract the predators of these insects [106, 107, 108], while others may also act in inter-plant communications, warning other plants of herbivory [109].

The initial substrates for TPSs are formed from different combinations of two metabolites, isopentyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). These substrates can be derived from either of two distinct pathways, the methylerythritol phosphate (MEP) pathway and the mevalonic acid (MVA) pathway. The MEP pathway is performed in the plastid and generates GPP and GGPP, which are used by monoterpene and diterpene synthases [110]. The MVA pathway is located primarily in the peroxisome [111] and is responsible for the production of FPP, used by sesquiterpene synthases.

Terpene synthases are grouped according to their common substrates. Monoterpenes all utilize (*E*)-geranyl diphosphate (GPP) as substrate, which is comprised of two isoprene units; while sesquiterpene synthases all catalyze reactions using (*E*,*E*)-farnesyl diphosphate (FPP), which is comprised

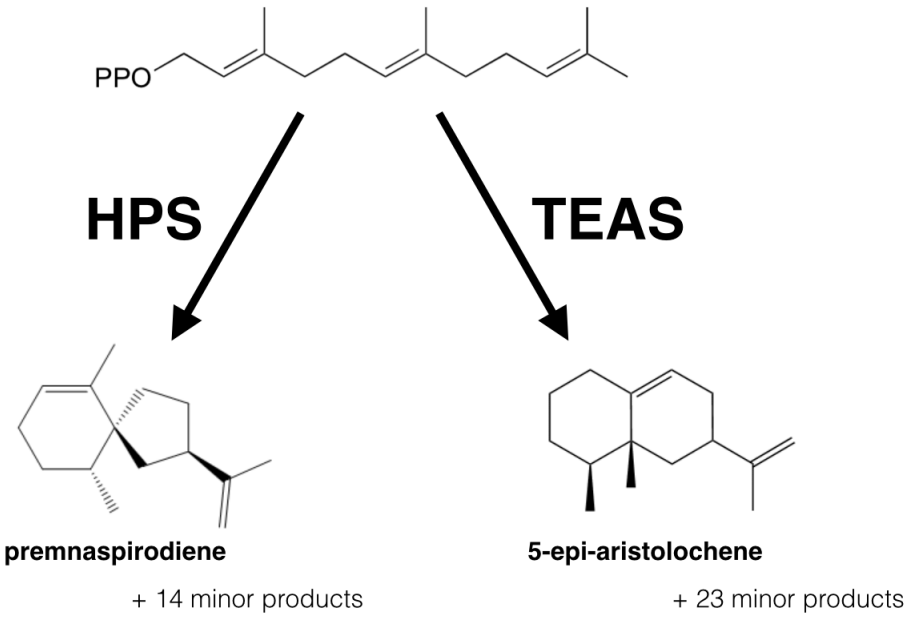
of three conjugated isoprenes. Plant diterpenes all utilize (*E*,*E*,*E*)-geranylgeranyl diphosphate (GGPP), which is comprised of four isoprene units. These enzymes have similar reaction mechanisms and may perform relatively complex reactions on the initial substrates, ionizing, cyclizing, and aromatizing all in one enzyme [72, 112]. In each case, divalent metals bound to the TPS enzyme contribute to the ionization of the isoprenoid substrate, forming carbocation intermediates which can then become cyclized or further rearranged to form the terpene products [113].

Plant hemi-, mono-, sesqui-, and diterpene synthases all appear to share a common ancestor, most likely a diterpene synthase involved in primary metabolism [114, 115, 93]. More than 100 terpene synthases have been identified in plants, forming a large superfamily and contributing to the enormous variety of terpenes [93].

Two closely related plant sesquiterpene synthases, *N. tabacum* 5-epi-aristolochene synthase (TEAS) and *H. muticus* premnaspirodiene synthase (HPS), have been extensively studied and characterized. TEAS and HPS share 72% amino acid (AA) identity [72]. Both utilize farnesyl diphosphate (FPP) as substrate, yet each yields different reaction product profiles (Figure 2.1). Once FPP is ionized, it can form a variety of different carbocation intermediates that can lead to a diverse array of alternative cyclization products. Wild-type TEAS catalyzes the ionization of FPP, and further cyclizes it twice, eventually forming mostly 5-epi-aristolochene as well as 24 other minor products [72]. The synthesis of 5-epi-aristolochene is the first committed step in the biosynthesis of the plant phytoalexin capsidiol, which serves as an antifungal agent [116].

HPS performs similar functions, but produces premnaspirodiene as it's main product as well as 14 other minor products [15]. Both 5-epi-aristolochene synthases (5EAS) and premnaspirodiene

Figure 2.1: TEAS and HPS pathways. Both TEAS and HPS utilize farnesyl diphosphate as substrate, but produce either 5-epi-aristolochene or premnaspirodiene, respectively, along with other minor products.



synthases (PS) share a similar protein fold (Figure 2.5) with two major domains: an inactive N-terminal domain with structure similar to the ”class II terpenoid synthase fold” [116, 117], and a C-terminal domain containing the active site [118], which is comprised largely of hydrophobic residues [119], matching a well-known ”class I terpenoid synthase” fold [116]. Several terpene cyclase motifs have been identified in TEAS, particularly a DDxxD dual Mg^{2+} binding domain [120] at the entrance of the active site, and a ”NSE/DTE” domain that binds a third Mg^{2+} [119]. These three Mg^{2+} ions bind and position the diphosphate moiety within the active site.

Based on 3D crystal structures of TEAS (pdb:5EAT and 5EAU) and a homology model of HPS, it was found that presumed active site residues in contact with the substrate are the same in both HPS and TEAS and therefore residues controlling product specificity must be outside the active site [118]. Domain swapping experiments between TEAS and HPS yielded enzymes with mixed

product profiles [121]. Using a contact mapping strategy, Greenhagen et al. identified 9 residues in so-called ”second tier” sites (residues in contact with active site residues) that differ between TEAS and HPS. When these ”second tier” site amino acids are interconverted between TEAS and HPS, the nine substitutions result in a switch in product profiles to the alternate enzyme [72]. Stepwise mutagenesis of these residues in TEAS shows an additive effect, and a mutant containing all nine mutations results in a near total conversion to HPS. Corresponding substitutions in HPS also result in a product profile similar to TEAS. They then created a mutant library containing all 512 (2^9) permutations of these nine substitutions and measured the product profile of each to examine the catalytic landscape [71]. These detailed structure function studies established which sites are responsible for determining HPS and TEAS enzyme reaction product specificity.

The objective of this study was to identify patterns of natural selection in TEAS and HPS, specifically at the nine sites known to be important for product profile specificity. There are methods that can quantitatively estimate the forces of natural selection on a gene family or at specific codon sites in a given gene. One such method, as performed by the program PAML [83], fits several models (with and without parameters for positive/adaptive selection) to a dataset to calculate the ratio of the nonsynonymous to synonymous mutation rates (d_N/d_S, or ω). These d_N/d_S values can be inferred for whole genes (averaged across sites) at specific branches in a phylogenetic tree, or at individual codon sites in a multiple sequence alignment (MSA). A d_N/d_S value less than one means the rate of synonymous mutations is higher than the rate of nonsynonymous ones, inferring that amino acid changes are not being permitted at that site (purifying or negative selection). A d_N/d_S equal to one indicates equal rates of nonsynonymous and synonymous mutations, implying

that there is no bias whatsoever in which AAs are being substituted, and therefore an indication of neutral selection. Positive (or adaptive) selection can be inferred as sites with a d_N/d_S greater than one, indicating that there are significant differential biases in AA substitutions that are occurring in the protein. This pattern is associated with positive/adaptive/Darwinian natural selection acting on the gene family.

2.3 Methods

A protein BLASTP search of the NCBI nr database (as of June 29, 2010) for amino acid sequences similar to tobacco 5EAS (Accession Q40577.3) using the BLOSUM45 matrix and an e-value cut-off of 10^{-72} yielded 524 sequences. Crystal structure sequences were removed (as the corresponding protein sequence was already represented), and the remaining sequences were aligned using MAFFT E-INS-I [122]. ProtTest 2.2 [123] identified the JTT+G+F model as the best fitting amino acid model. A 1024 replicate Maximum Likelihood (ML) bootstrap phylogenetic analysis, as well as a 100 replicate ML search were performed on this dataset using the program GARLI [124], and a majority-rule ML search consensus tree was produced using PAUP* [125] and sumtrees.py [126]. Bootstrap values were also mapped to the best overall ML search tree. The sequences in the clade containing known TEAS-like and HPS-like sequences were then aligned separately with MAFFT E-INS-I. Using ProtTest 2.2, The JTT+I+G+F model was identified as the best fitting amino acid model for the smaller dataset. A 100 replicate ML search phylogenetic analysis was performed using GARLI, and the best overall likelihood tree was selected for further analysis. Nucleotide

sequences corresponding to each protein sequence in the MSA were downloaded from NCBI, and mapped to the alignment using PAL2NAL [127]. PAML [83] was used to test various models of evolution over the codon alignment as well as datasets containing only aligned *TEAS*-like sequences or *HPS*-like sequences. The F61 model was identified as having the best likelihood and was used for all branch and site codon model analyses.

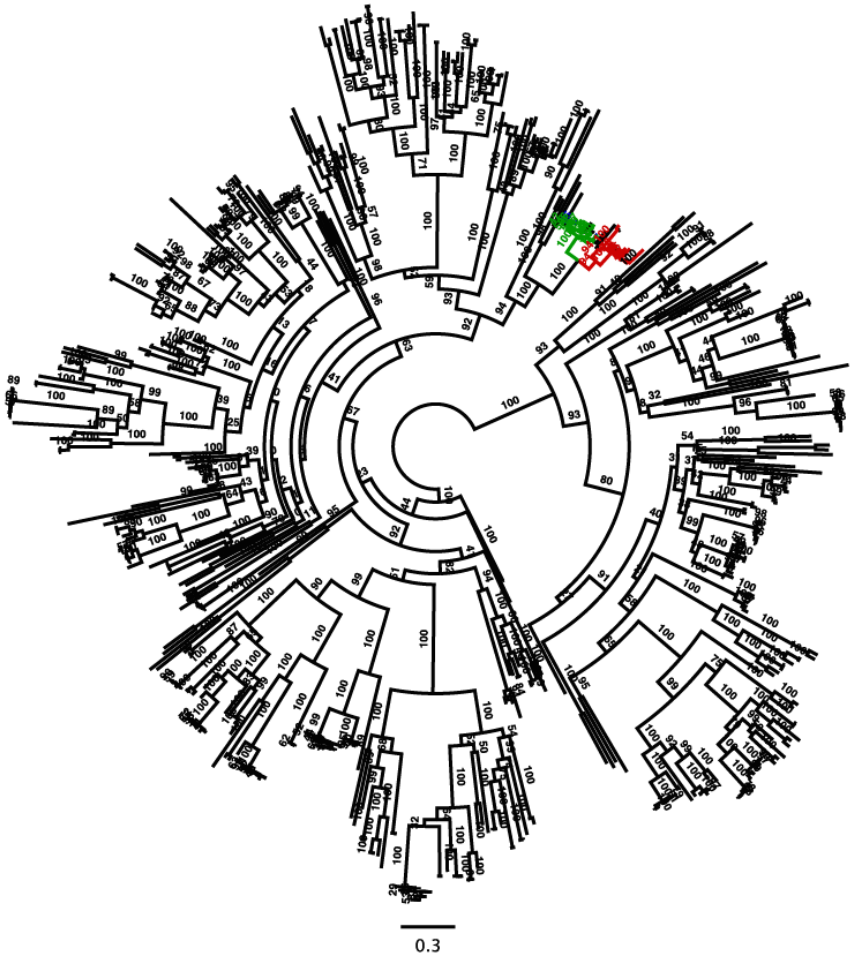
A ML phylogenetic analysis of TEAS and HPS-like amino acid sequences showed a close sister-grouping relationship between these two terpene synthases (Figure 2.3a). A similarity search of *N. tabacum* 5EAS yielded both TEAS-like sequences and HPS-like sequences, as well as numerous plant sesquiterpene synthases. All biochemically characterized TEAS-like and HPS-like sequences formed a monophyletic clade. Proteins with 5EAS activity clustered together separately from those with PS activity (Figure 2.3b), as identified previously [71]. The sequences closest to the 5EAS/PS split had various sesquiterpene synthase activities, catalyzing the formation of germacrene, delta-cadinene, or (*E*)-B-farnesene. This suggests that TEAS and HPS likely share a recent common ancestor in the sesquiterpene synthases, and that they likely diverged relatively early in the Solanaceae.

2.4 Results

The evolutionary trajectories of TEAS and HPS presuppose natural selection forces that promoted the differentiation of these two enzyme specificities. Therefore, we became interested in measuring contemporary natural selection acting upon these enzymes and especially the nine sites known to

be important for enzyme product profile specificity as a means of gaining a clearer understanding how natural selection affects protein evolution. A thorough meta-analysis of this model dataset using quantitative measures of past natural selection should reveal important patterns for neofunctionalization in enzyme evolution.

Figure 2.2: ML search tree of TEAS and HPS homologs with non-parametric bootstrap support values mapped to branches. The tree with the best overall likelihood score out of 100 ML search replicates was selected and bootstrap support values were mapped to branches. Bootstrap values are labeled as percentages out of 1024 bootstrap replicates. Taxa labeled red have proven TEAS activity. Taxa labeled green have proven HPS activity. Blue taxa are probable HPS-like enzymes.



PAML was run on the original dataset containing known *TEAS*-like and *HPS*-like sequences, as

Table 2.1: Likelihood ratio test of PAML site models for *TEAS* and *HPS*. PAML site analyses were performed on the three indicated datasets. The M1a and M2a models bin sites into two or three categories, with ω values restricted to less than 1 or equal to 1, as well as ω values greater than one (positive selection) in the M2a model. The M7 and M8 models bin sites in a beta distribution from 0 to 1; as well as a bin for ω values greater than 1 in the M8 model (positive selection). The M1a and M2a models likelihood ratio test (LRT) is a more conservative measure of positive selection than the M7 and M8 models LRT. A hyphen ("-") indicates the model does not include that parameter.

TEAS/*HPS* combined

Model	lnL	LRT p-value	ω <1	ω = 1	ω >1
M1a	-11532.6	1	0.1214	1	-
M2a	-11532.6		0.1214	1	1
M7	-11486.1	<0.01	-	-	-
M8	-23768.0		-	-	1.6382

TEAS alone

Model	lnL	LRT p-value	ω <1	ω = 1	ω >1
M1a	-5988.38	1	0.1198	1	-
M2a	-5988.38		0.1198	1	1
M7	-5988.82	<0.01	-	-	-
M8	-5980.09		-	-	3.6263

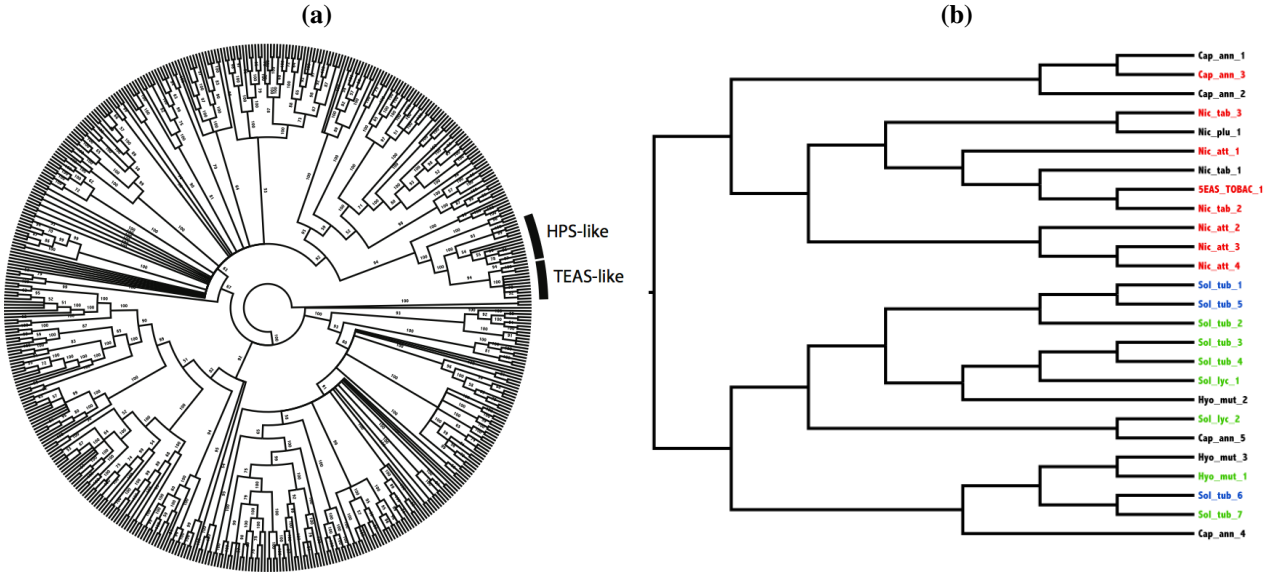
HPS alone

Model	lnL	LRT p-value	ω <1	ω = 1	ω >1
M1a	-6779.92	1	0.0826	1	-
M2a	-6779.92		0.0826	1	1
M7	-6770.40	1	-	-	-
M8	-6767.71		-	-	1.2697

well as datasets containing only *TEAS*-like or *HPS*-like sequences based on topology in the phylogenetic tree (Figure 2.3b) and important differences in the protein multiple sequence alignment (MSA) as described in Greenhagen et al. 2006. Using a likelihood ratio test (LRT) on the strict M1a/M2a models (which model purifying and neutral selection or purifying, neutral, and positive selection), PAML did not detect significant positive selection in any of the combined or individual datasets (Table 2.1). Using the more relaxed M7/M8 models (which model purifying and neutral selection as a beta distribution, and may (M8) or may not (M7) allow for positively selected sites) the LRT found support for positive selection in both the combined *TEAS-HPS* dataset and the *TEAS*-like datasets.

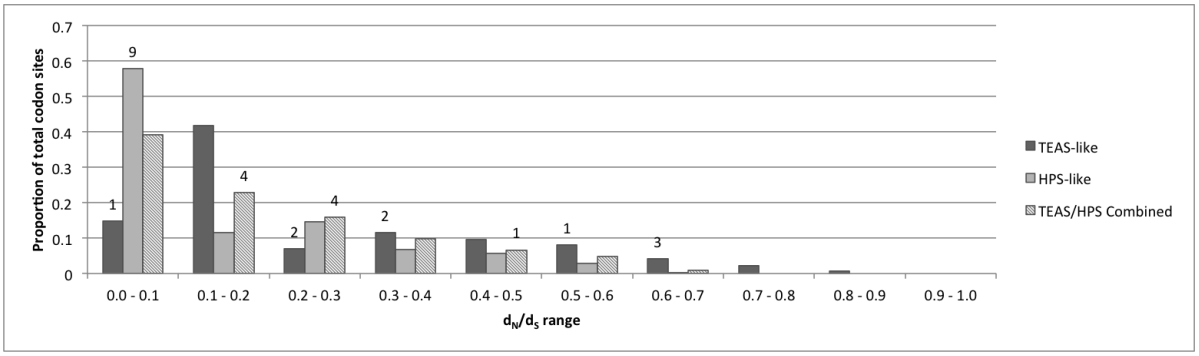
The nine known amino acid sites affecting TEAS and HPS product-specificity determined by the Noel lab [72] were then examined for distinct patterns of natural selection using the neutral M7 codon model in the three datasets since none of the nine sites were found to have significant positive selection in any of the analyses. Since there was no detected positive selection, we were interested in what character of natural selection was occurring at these sites, whether purifying or neutral. Graphing the distribution of d_N/d_S values across the three datasets (Figure 2.4) revealed different profiles of natural selection in *TEAS* and *HPS*. In *HPS*, approximately 60% of all sites had d_N/d_S values between 0 and 0.1, indicating strong purifying selection. In contrast, only 14.8% of sites in *TEAS* were under strong purifying selection (d_N/d_S less than 0.1). While sites in *HPS* skewed towards strong purifying selection, sites in *TEAS* were under apparently more relaxed (but not neutral) selection (d_N/d_S between 0.1 and 1). The consistent selective differences between these two enzymes led an interest in the differences between the two gene families at the nine

Figure 2.3: Phylogenetic analyses of TEAS and HPS-like sequences. a. A 1024-replicate non-parametric bootstrap maximum likelihood (ML) consensus tree of amino acid sequences similar to tobacco 5EAS (BLAST e-value cutoff 10^{-72}). Sequences with known TEAS or HPS activity are marked. Bootstrap scores are reported as percentages. b. Best overall ML search tree of TEAS and HPS homologs. The optimal likelihood tree from a 100-replicate ML search using GARLI was selected and used for later analyses. Sequences colored red have proven TEAS activity. Sequences colored green have biochemically characterized HPS activity. Blue sequences are annotated as probable HPS enzymes.



catalytically-important sites. At eight of the nine sites responsible for product-specificity, purifying selection was stronger in *HPS* than in *TEAS*. These eight sites each had a difference in d_N/d_S value ($\Delta d_N/d_S$) of 0.2 or greater (Table 2.2). This may be indicative of a past selective sweep in *HPS* as this activity became selectively advantageous in the plants producing *HPS*. Strong differences in selective pressure ($\Delta d_N/d_S$) may be indicative of sites important for one enzyme activity but not necessarily the other.

Figure 2.4: Distribution of d_N/d_S values across three datasets (*TEAS*-like genes, *HPS*-like genes, and both combined) calculated using the PAML M7 model. The numbers above bars indicate the count of the nine sites identified as important for product-profile specificity as identified by Greenhagen et al. with d_N/d_S falling in that range. Sites with d_N/d_S values close to zero indicate strong purifying selection. Values around one indicate neutral selection. Values greater than one indicate positive selection. In these analyses, the likelihood ratio test between PAML models M7 and M8 supported positive selection in the *TEAS*-like and *TEAS*/*HPS* combined datasets, but not in the *HPS*-like dataset. Comparisons were therefore made using the M7 model for all three datasets.



Mapping $\Delta d_N/d_S$ values onto their corresponding amino acid in the 3D crystal structure of *TEAS* with substrate analog (PDB:5EAT) revealed a map of differences in natural selection between these two enzymes (Figure 2.5). Sites with a $\Delta d_N/d_S$ greater than 0.2 comprised 194 out of 548 (35.4%) of total amino acid positions, and just 51 of 177 (28.8%) of sites within 15Å of the active site.

Table 2.2: PAML site model analysis results for nine sites important for differentiating TEAS and HPS activity. The M7 model was used for estimating d_N/d_S values.

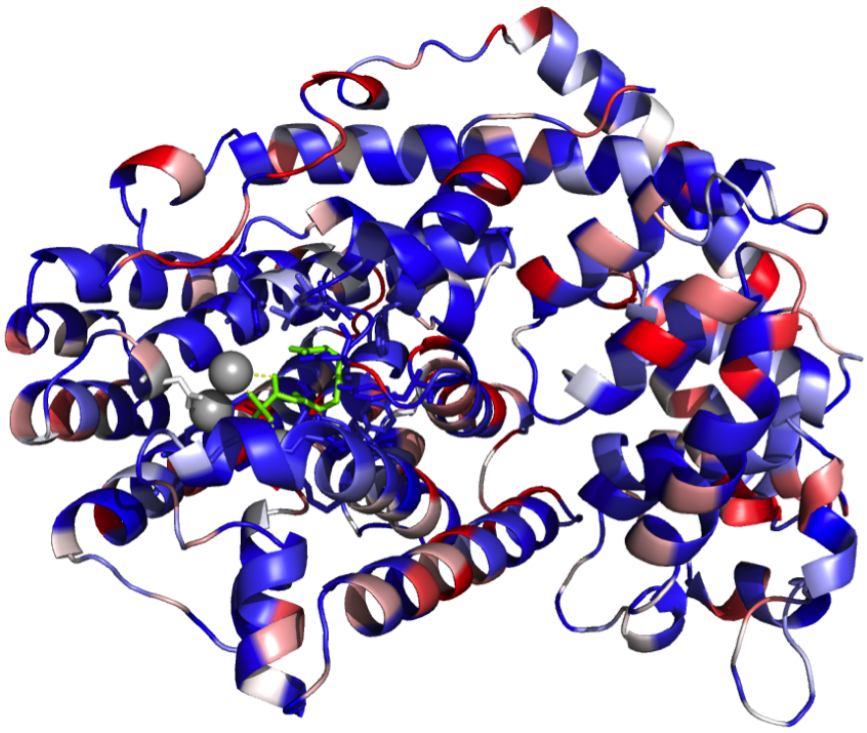
AA position responsible for altered activity	*TEAS*/*HPS* combined d_N/d_S	*TEAS* d_N/d_S	*HPS* d_N/d_S	*TEAS*vs.*HPS* $\Delta d_N/d_S$
TEAS Ala-274	0.266	0.364	0.072	0.292
TEAS Val-291	0.192	0.285	0.052	0.233
TEAS Val-372	0.261	0.559	0.051	0.508
TEAS Thr-402	0.299	0.406	0.074	0.332
TEAS Thr-406	0.475	0.597	0.075	0.552
TEAS Ser-436	0.235	0.537	0.032	0.505
TEAS Ile-438	0.139	0.336	0.059	0.277
TEAS Ile-439	0.159	0.092	0.080	0.012
TEAS Val-516	0.188	0.288	0.049	0.239

2.5 Discussion

Signals of positive selection have previously been used to identify sites important for neofunctionalization [128], however in this study there was very little support for contemporary positive selection acting on the extant *TEAS* and *HPS* gene families. While the likelihood ratio test identified the PAML M8 model as significantly better than the M7 model in the whole gene *TEAS* (p-value <0.001) and *TEAS-HPS* combined (p-value <0.01) datasets (Table 2.1), very few individual sites were under positive selection. When d_N/d_S values were mapped to the crystal structure of *TEAS* bound with a substrate analog (PDB:5EAT), the few sites identified as under positive selection were either on the surface of the protein or were far from the active site, and thus unlikely

to directly influence catalytic activity. However, it is formally possible that these more distant sites promoted overall enzyme stability and thus were important for manifesting an adaptive trait overall. Thus, positive selection did not seem to be operating on these extant gene families at amino acid positions critical for either substrate binding or at second tier sites responsible for determining reaction product specificity [71]. Instead, using the PAML M7 model for comparisons between datasets, all sites in both *TEAS* and/or *HPS* had d_N/d_S values less than 0.9. Many of these sites were under strong purifying selection (<0.2) and presumably essential for overall enzyme activity. However, there were also a number of sites that do not clearly fall into positive selection nor purifying selection.

Figure 2.5: 3D crystal structure of Tobacco 5EAT (PDB: 5EAT) with $\Delta d_N/d_S$ values mapped to residues. Differences in selective pressure between *TEAS* and *HPS* are indicated by a gradient from blue to red (0.0 through 0.5). Substrate farnesyl diphosphate (FPP) is colored green. Sites with a large $\Delta d_N/d_S$ indicate different selective pressure between *TEAS* and *HPS*.



Interestingly, a large proportion of sites had somewhat intermediate d_N/d_S values between 0.1 and 0.6 (60.9% of sites in *TEAS-HPS* combined dataset), indicating some intermediate form of relaxed natural selection between strong purifying selection (d_N/d_S less than 0.1) and neutral selection (d_N/d_S equal to ~1). It is unclear exactly what value or range of d_N/d_S values equates to Ohta's nearly-neutral selection, that is, sites that are only somewhat selectively deleterious and not undergoing pure neutral selection.

HPS appears to be under overall stronger selective pressure than *TEAS* (Figure 2.4), with a large proportion (57.8% versus 14.8% in *TEAS*) of sites under very strong purifying selection (d_N/d_S values less than 0.1). In contrast, *TEAS* appeared to be under more relaxed selection with a majority of sites (85.1%) having d_N/d_S values greater than 0.1 but less that 1.0. The combined *TEAS*/*HPS* dataset had a distribution of d_N/d_S values somewhat in between those of the *TEAS* or *HPS* alone datasets.

The TEAS and HPS enzymes have active sites very similar to one another, with many of the same residues in the active site. Product specificity appears to be contingent on the overall shape of the active site rather than specific amino acids differences at the interface of substrate binding. Many of these active site residues and residues in other catalytically important regions are under strong purifying selection (d_N/d_S <0.1) in both *TEAS* and *HPS*, reflecting the selection to maintain the overall sesquiterpene synthase fold. The first two aspartate residues in the metal-binding DDXXD conserved motif [116] appear to be under strong purifying selection in both *TEAS* and *HPS*, though the third is under slightly less strong selection in *TEAS*, as well as the two non specific residues in the motif. Residues further from the active site tend to be under less strong selection. As long as

the general shape of the protein fold is maintained, more mutations are tolerated at those sites.

Of the nine catalytically important sites identified by Greenhagen, all nine were under strong purifying selection in *HPS* ($d_N/d_S<0.1$), while in *TEAS* eight of the nine sites were under less strong selection, with d_N/d_S values between 0.2 and 0.7. Comparing equivalent sites in *TEAS* and *HPS*, these eight sites had a difference in d_N/d_S, or $\Delta d_N/d_S$, of 0.2 or greater (Table 2.2). These sites may be under stronger selective pressure in *HPS* to maintain its high specificity towards the premnaspirodiene product. This may also be reflected in the relatively smaller number of minor products produced by HPS (14) compared to TEAS (24). In the regions previously identified as catalytically important for 5EAS and PS activity (*TEAS* sites 261-443) [121], 56 sites had a $\Delta d_N/d_S$ of 0.2 or greater, out of 182 total sites. Of these, only 10 sites within 12.5Å of the active site and differed in *TEAS* and *HPS* had a $\Delta d_N/d_S$ of 0.2 or more. Eight of these positions were those identified by Greenhagen et al.

The phylogenetic analyses clearly demonstrate that HPS and TEAS arose from a recent common ancestor. It is possible that the ancestral terpene synthase-like enzyme to TEAS and HPS was catalytically promiscuous and produced a variety of products, and thus was under less strong selective pressure towards producing one specific product. This would reflect the hypothesis that ancestral enzymes were more promiscuous and/or structurally disorganized [15]. The contemporary signals of natural selection acting on the *TEAS* gene family were consistent with this pattern, because there were both weak signals of positive selection overall and at specific sites distant from the catalytic center. Moreover, there were proportionally more sites showing intermediary signals of natural selection (>0.2 but less than 1.0), suggesting some balance between neutral selection/genetic drift

and purifying selection. This spectrum of natural selection operating between purifying and purely neutral was likely important for the evolution of the *TEAS* gene family. This pattern of natural selection can be thought of as ”effectively mutable protein space”. Thus, contemporary *TEAS* gene family appears to be rather tolerant to natural selection operating within this ”effectively mutable” or nearly neutral evolutionary trajectory. On the other hand, natural selection acting on the contemporary *HPS* gene family seems to be on a different evolutionary trajectory.

In contrast to *TEAS* and combined *TEAS-HPS* datasets, the *HPS* dataset did not show any evidence of positive selection. Moreover, the proportion of sites showing purifying selection was substantially greater. For these reasons, it is likely that the extant *HPS* gene family underwent an episode of strong purifying selection typified by a small population undergoing a selective sweep because it conferred a strong adaptive advantageous to the plant in response to a fungal pathogen.

At this juncture it is difficult to determine if the ancestral terpene synthase gene that gave rise to *HPS* and *TEAS* was a generalist or specialist. In one scenario, the ancestral enzyme had HPS-like activity, and a gene duplication event allowed for one copy to become more generalist and eventually encode TEAS-like activity. This would explain the relaxed selection in the *TEAS* dataset. Alternatively, the ancestral enzyme may have been more of a generalist enzyme tolerant of many alleles that contributed to catalytic diversity, from which one unique allele provided HPS activity that conferred a strong selective advantage in a particular ecological habitat, thereby giving rise to the *HPS* gene family.

Based on this meta-analysis of existing *TEAS*/*HPS* data, we hypothesize that sites in or relatively close to the active site with large $\Delta d_N/d_S$ values might be used to identify a relatively small number

of critical amino acid sites that, when altered, have a high probability of resulting in productive enzyme neofunctionalization. Combining these quantitative measures of natural selection with 3D structural data provides a visualization of past and present selection on a gene family, and may have applications in engineering other enzyme systems.

Chapter 3

Investigations into molecular evolution-guided protein engineering of putrescine N-methyltransferase and spermidine synthase

Alexandra J Weisberg, Eva Collakova, and John G Jelesko.

3.1 Abstract

Determining the evolutionary path to neofunctionalization is of special interest to rational protein engineering. Current protein engineering approaches use either comparative protein multiple sequence alignments, or high frequency random mutagenesis to obtain enzymes with modified properties. Specific patterns in quantitative measures of the forces of natural selection, particularly the nonsynonymous to synonymous mutation rate ratio (d_N/d_S) might provide new opportunities to define "effective mutable space". We propose that the combination of (3D) protein structural data with quantitative measures of natural selection (EVO) may provide a novel method for rational protein engineering (3D-EVO). This methodology was first examined with a meta-analysis of a well-studied model system of two closely-related terpene synthases, and then applied to two other closely-related gene families (spermidine synthase and putrescine N-methyltransferase) to predict which residues are responsible for determining substrate specificity of these two closely related enzyme families. These predictions were tested using a synthetic gene expressing a recombinant spermidine synthase protein engineered with the objective of expressing putrescine N-methyltransferase activity. Biochemical analysis of the engineered SPDS recombinant protein did not indicate neofunctionalization with putrescine N-methyltransferase activity. Possible deficiencies with this approach are discussed.

3.2 Introduction

Plants produce a vast number of small molecular-weight chemicals, many of which are not necessary for plant growth and development, but rather confer some other selective advantage (so-called secondary or specialized metabolism). These metabolites may play a role in defense against herbivory, attract pollinators, defend against pathogens, or regulate temperature, among others [10, 11, 12, 13, 14]. It has been estimated that these metabolites number over 200,000 [8] and may even number in the millions (Eran Pichersky, personal communication). There are over 12,000 known alkaloids [129] and more than 55,000 identified terpenes [3].

Despite what is known about the existence of phytochemical diversity, relatively little is known about how plants evolved to produce them. There has been much debate over whether neofunctionalization arises primarily through adaptive/Darwinian/positive selection, nearly-neutral genetic drift, or some combination of the two. The adaptive or Neo-Darwinian models of evolution claim that the vast majority of amino acid-changing mutations in codons are deleterious and are selected against (aka purifying selection), while a select few are drastically advantageous and are immediately selected for [34, 22]. Neofunctionalization thus occurs when a gene acquires a mutation that changes the function of the encoded protein, producing new metabolites that are of immediate selective advantage to the plant.

Mutations may instead be primarily neutral or nearly-neutral in nature, implying that the majority of neofunctionalization is through genetic drift along with some level of selection. Mutations may be deleterious, nearly-neutral (slightly deleterious), neutral, or adaptive; though the vast majority

are assumed to be neutral or nearly so [102, 43]. As long as protein stability is not compromised, enzyme activity is preserved, and regulation is stable, enzymes can withstand a large number of changes [56, 46]. The rate of amino acid substitutions in mammals has been shown to be too high for species to tolerate if the majority are selectively deleterious [41]. However, it has also been shown that increasing numbers of mutations may totally destabilize a protein, often in as a few as 10 mutations [45, 51]. Some mutations may be compensatory for others that confer new activity yet are destabilizing [57, 58, 59]. These compensatory mutations may occur before or after a destabilizing mutation, either providing the stability to allow for a more unstable mutation or by ”rescuing” stability afterwards. The existence of chaperone proteins may also broaden the types and number of destabilizing mutations that are tolerated [130]. This range of acceptable mutations can be described as enzyme ”effectively-mutable space.”

Identifying target sites conferring neofunctionalization has been attempted through various methods, some more successful than others. One common approach utilizes naive comparisons between protein multiple sequence alignments containing sequences from two closely related but different enzyme families; combined with location within the active site cavity using 3D protein crystal structure information. This method is sometimes successful, but often results in denatured protein or protein lacking any measurable enzyme activity [78, 79, 80, 81, 82]. Other approaches, such as active site residue contact-mapping strategies [72] have been more successful.

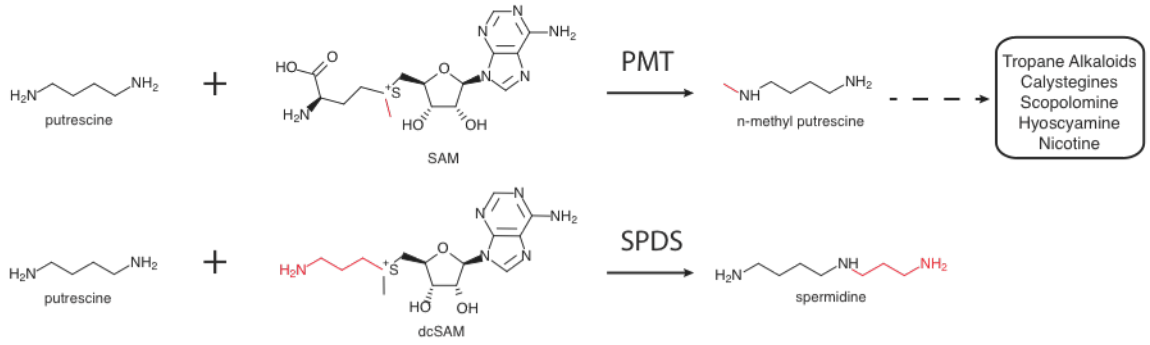
In the previous chapter, I described a meta-analyses of a model gene family of sesquiterpene synthases (Chapter 2), it was found that codons encoding amino acid (AA) residues at sites known to be important for product specificity [72] were under differing degrees of selective pressure in 5-

epi-aristolochene (TEAS) and premnaspirodiene synthase (HPS). These residues were not directly in the active site, which was nearly identical between TEAS and HPS, but rather at second-tier sites that, that when modified changed the reaction product-profile specificity [72]. Substitutions from the opposing enzyme at any one of these sites (and vice versa) caused the proteins to produce a mixture of the original metabolite, the metabolite of the opposing enzyme, and sometimes new metabolites produced by neither wild-type enzyme. Product profile-changing effects were additive in nature, but mutation of all nine residues resulted in a near-complete shift in activity to the opposing enzyme.

Interestingly, neither gene was under strong adaptive selection. In the case of *TEAS*, most of the sites previously identified as important for product specificity were under intermediate to weak purifying selection, suggesting that neofunctionalization may have occurred through a combination of selection and genetic drift. Moreover, using a comparative approach between one gene family showing more relaxed/weak purifying selection and another showing mostly purifying selection, sites that were previously identified as important for conferring enzyme reaction specficity were identified.

The present chapter aimed to use comparative measures of extant natural selection ($\Delta d_N/d_S$) metrics to identify critical amino acid positions responsible for the catalytic specificities of putrescine N-methyltransferase (PMT) and spermidine synthase (SPDS). This was done with the end goal of developing a novel method for rational protein engineering using quantitative measures of natural selection complemented by 3D protein structural data. Our aim was to quantitatively assess natural selection in order to identify effectively mutable protein space in two other closely related enzyme

Figure 3.1: Putrescine N-methyltransferase and spermidine synthase enzyme activity.



families and thereby validate or reject the general applicability of this methodology.

Plant putrescine N-methyltransferase and spermidine synthase are two closely-related enzymes (64-68% identity [131] that perform different catalytic functions (Figure 3.1). Both enzymes utilize putrescine as a substrate, and either of two closely related substrates: either S-adenosylmethionine (SAM) or decarboxylated S-adenosylmethionine (dcSAM), respectively. SPDS acts as an amino-propyltransferase to produce spermidine, a polyamine important to primary metabolism. Spermidine is further acted on by spermine synthase (SPMS) to form spermine, another polyamine necessary in primary metabolism to regulate cell differentiation [132]. Spermidine synthase and spermine synthase are found in all branches of the tree of life. In contrast, PMT catalyzes the transfer of a methyl group from SAM to putrescine to form N-methylputrescine and thus is an N-methyltransferase. N-methylputrescine is an important precursor in certain plant alkaloid biosynthetic pathways; and has only been identified in plants.

PMT performs the first committed step in the biosynthesis of nicotine and tropane/nortropane alkaloids such as atropine, hyoscyamine, and scopolamine. Nicotine acts as an important chemical

defensive compound against herbivory [86, 87, 88, 89, 90, 92]. While *PMT* was originally cloned from *N. tabacum* [133, 134], it has also been cloned from other members of the Solanaceae that produce tropane and/or nortropane alkaloids (e.g. henbane and potato) [135, 131, 136]. While X-ray crystal structures of SPDS from a variety of organisms have been resolved [137, 138, 139, 140], none have been successfully determined for PMT. PMT is surprisingly difficult to crystalize and measure (despite its high sequence similarity to SPDS), as crystals fall apart upon the application of an X-ray beam [141]. The *N. tabacum* PMT protein contains an N-terminus with different numbers of an 11 amino acid repeating section that is not necessary for enzyme activity but provides additional stability [142]. These results suggesting PMT may be inherently more unstable than SPDS. Chimeric proteins of N- and C-terminus halves of PMT and SPDS revealed that enzyme activity and thus substrate recognition specificity is primarily controlled by the N-terminus half of each protein [141].

It is postulated that PMT evolved from an ancestral plant SPDS, as it shares a sequence identity with SPDS, but not with other methyltransferases [134]. This is intriguing because SAM-dependent methylation is a relatively common reaction, whereas amino-propyltransferase activities are not a common reaction in cellular metabolism.

Several attempts have previously been made to interconvert a SPDS to a PMT or vice versa, with little success [101, 141, 82]. The amino acid sites changed in these experiments were identified through individual conservation in multiple sequence alignments and PMT structural homology modeling to identify amino acids in contact with either putrescine or SAM. In each case, the modified *Datura stramonium* SPDS or *D. stramonium* PMT had either reduced or no original

activity, and no measurable desired enzyme activity. We hypothesized that substrate recognition specificity may be dependent not on amino acids directly in the active site, but rather those in second-tier sites that might change the overall shape of the active site. To test this hypothesis and a natural selection-guided protein-engineering methodology, the same computational methods used in the meta-analysis of the *TEAS*/*HPS* gene family were applied to the *PMT*/*SPDS* family. Based upon differences in apparent natural selection between the *SPDS* and the *PMT* gene families, a synthetic *SPDS* gene was designed with the goal of creating a neofunctional PMT enzyme activity. The recombinant reengineered SPDS protein was isolated and failed to show PMT activity.

3.3 Methods

3.3.1 Phylogenetic analyses

Nicotiana tabacum PMT1 (NCBI accession Q42963.1) was used as a query in a BLASTP search of the NCBI nr database using the BLOSUM45 matrix and an e-value cutoff of 1e-20. Amino acid (AA) sequences were aligned using MAFFT E-INS-I [122]. The program ProtTest 2.2 [123] identified the WAG+I+G+F model as the best-fitting amino acid model. A 100 search replicate Maximum Likelihood (ML) phylogenetic analysis was performed on this dataset using the program GARLI 0.961b [124], and a ML search majority-rule consensus tree was generated using PAUP* [125] (Figure 3.2a). Similar analyses were performed on a reduced dataset containing only PMT and plant SPDS/SPMS sequences (Figure 3.2b). A 1024-replicate non-parametric bootstrap

analysis was also performed using GARLI [124]. The sumtrees.py program from the DendroPy package [126] was used to generate a majority-rule non-parametric bootstrap consensus tree (Figure 3.7a) and to map bootstrap support values onto the best ML search tree (Figure 3.7b).

DNA corresponding to each plant protein sequence was aligned using the plant ML search consensus AA alignment as a guide using PAL2NAL [127]. The program PAML [83] was used to perform site-model analyses to calculate the rate of non-synonymous to the rate of synonymous mutation ratios (d_N/d_S) at each codon site in a multiple sequence alignment. The best pairwise likelihood codon model was found to be F61 and was used for each analysis. Individual site-model analyses were performed on datasets containing only *PMT* sequences, only plant *SPDS*/*SPMS* sequences, or *PMT* and *SPDS*/*SPMS* sequences combined.

In order to putatively define the active site in an *Arabidopsis thaliana* SPDS crystal structure (PDB: 1XJ5), a pairwise structure alignment was performed using DaliLite [143] with a *SPDS* crystal structure from *Plasmodium falciparum* (PDB: 2I7C). *P. falciparum* SPDS was crystalized with the putrescine and dcSAM substrate analog AdoDATO (S-adenosyl-3-thio-1,8-diaminooctane); and had the best structure alignment with 1XJ5. Amino acids in "second tier-sites" were defined as being within 15Å of the predicted active site, in this case centered around catalytically-important residue 131D [144].

3.3.2 Differing selective pressure on amino acid sites

Site-wise d_N/d_S ratios from the M7 model were compared between the *PMT* and *SPDS* datasets to identify differential selective pressure at each site ($\Delta d_N/d_S$). Individual site $\Delta d_N/d_S$ values were mapped to their corresponding amino acid residue in the *A. thaliana* SPDS crystal structure, and those within 15Å of residue 131D were selected for further analysis. Of these sites, 9 had a $\Delta d_N/d_S$ greater than 0.2. In order to increase the number of targets, sites with a $\Delta d_N/d_S$ greater than 0.1 were considered. There were 20 sites with a $\Delta d_N/d_S > 0.2$ within 15Å of the putative active site, 14 of which differed between *Solanum tuberosum* SPDS and *S. tuberosum* PMT.

DNA encoding St*SPDS* with the 14 identified sites replaced with their corresponding AA in StPMT (StSPDS^{PMT14}) along with *Nco*I and *Xho*I restriction sites was custom synthesized with codons optimized for expression in *E. coli* (Genscript). Custom synthesized DNA was then cloned into the pET21d expression vector for expression with a C-terminus 6xHIS tag. Wild-type StSPDS in pET21d [136] was kindly provided by Birgit Dräger. Wild-type NtPMT cloned into the pET32a+ vector for expression as a gene fusion with N-terminus thioredoxin and 6xHIS tag [145] was used as a positive control.

3.3.3 Protein purification and expression

Recombinant NtPMT, StSPDS, and StSPDS^{PMT14} were expressed and purified in *E. coli* for use in *in vitro* enzyme assays using the following protocol. A 50mL LB culture with appropriate antibiotics was inoculated with *E. coli* BL21 (DE3) containing each individual plasmid and grown

overnight at 37°C 250 rpm. Up to six 1L cultures of ZYM-5052 autoinduction media [146] were inoculated with 12mL of the 50mL overnight culture and grown at 37°C 250 rpm for 7 hours, then moved to 18°C for a total of 24 hours. Cells were pelleted in a Sorvall RC 6+ centrifuge (Thermo Scientific, Waltham, MA) at 4000rpm for 20min at 4°C. Cell pellets were frozen at -80°C.

Cell pellets were thawed on ice and resuspended in 3x volume Buffer A (50mM sodium phosphate pH 7.4, 300mM NaCl, 10% w/v glycerol). Lysozyme (25µg/mL) was added and the mixture was incubated at 4°C for 1 hour with constant stirring. Cells were then lysed on ice using a Sonic Dismembrator Model 500 (Fisher Scientific, Hampton, New Hampshire) for a total of 5 minutes of sonication (50% amplitude; 5 seconds on, 15 seconds off). The insoluble fraction was pelleted by centrifugation in a tabletop microcentrifuge at 14,000rpm for 30min at 4°C. The soluble fraction was then run over a column containing HisPur Ni-NTA resin (Thermo Scientific, Waltham, MA) using either a BioRad Econo Pump peristaltic pump connected to a BioRad Econo UV Monitor (BioRad, Hercules, CA) or an ʹAKTA Prime FPLC system (GE Healthcare, Buckinghamshire, United Kingdom). The column was then washed with Buffer A, then a mixture of 90% Buffer A, 10% Buffer B (50mM sodium phosphate pH 7.4, 300mM NaCl, 10% v/v glycerol, 250mM imidazole) to elute non-specifically binding proteins. Protein was eluted from the column using Buffer B. Elution fractions collected during a peak on the UV monitor were pooled and concentrated using a 15mL Vivaspin concentrator with a 10,000MW cutoff (Sartorius Stedim Biotech, Göttingen, Germany), then buffer exchanged into 100mM Glycine-NaOH pH 9 using a PD-10 column (GE Healthcare, Buckinghamshire, United Kingdom). Glycerol was added to a concentration of 10% and the samples were stored at -80°C. Protein concentration was quantified by Bradford assay

[147].

A plasmid containing DNA encoding recombinant 5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase (MTAN/SAHN) under control of the T7 promoter was kindly provided by Sunny Zhou (Washington State University)[148]. Recombinant SAHN protein was expressed in *E. coli* strain BL21 (DE3) and purified using Ni-NTA resin using a similar protocol to the above, with the exception of being grown in 1L LB media at 37°C 250rpm until an OD600 of 0.6 followed by induction with isopropyl β-D-1-thiogalactopyranoside (IPTG), and then growth overnight. SAHN was included in enzyme assays as it reduces the pool of S-adenosylhomocysteine, which inhibits PMT activity [135].

3.3.4 PMT enzyme assays

PMT enzyme activity was quantified by ultra performance liquid chromatography (UPLC) using an AccQ-Tag Ultra Derivatization kit (Waters, Milford, MA) to derivatize putrescine, N-methylputrescine, SAM, and SAH. Each reaction contained 3.8mM putrescine, 127 µM SAM, 0.3 µM SAHN, 23.4 µg either NtPMT, StSPDS, or StSPDS^{PMT14}, and 50mM HEPES pH 7 in a total volume of 315 µL. Assays were incubated for 60min at 30°C and were stopped by addition of 1x volume acetonitrile. Reaction products and substrates were derivatized by addition of 10µL AccQ-Tag solution to 10µL assay solution and 30µL borate buffer, followed by vortexing. A Waters Acquity H-class UPLC chromatograph with a reverse-phase AccQ-Tag Ultra Column (1.7µm particles) (Waters, Milford, MA) was used for all measurements. Standards of putrescine,

N-methylputrescine, and SAM were also run individually to test separation (Figure 3.6a). Assays were repeated in buffer sodium phosphate pH 7 after buffer exchanging recombinant proteins to remove glycerol, which was found to also be derivatized by the AccQ-Tag kit.

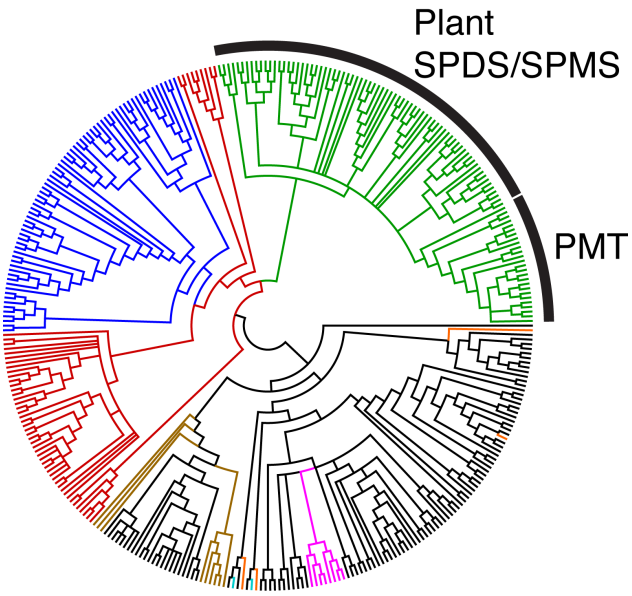
3.4 Results

Using model oriented phylogenetic methods and extensive sampling of homologs across the tree of life, plant PMTs formed a single clade within a larger, poorly resolved clade of plant SPDS homologs. As previously noted in less rigorous phylogenetic studies, the closest relatives to PMT were not other methyltransferases, but rather aminopropyltransferases. Both PMT and SPDS utilize putrescine but differ in their use of SAM or dcSAM, respectively, and it has previously been posited that substrate specificity may be due to the shape of the active site (Biastoff2009). PMT and SPDS/SPMS homologs were not well resolved, forming a large polytomy at the base of the clade in the majority rule ML search consensus tree (Figure 3.2a), suggesting that PMT and SPDSs diverged very recently in plants. More focused phylogenetic analyses comprising only plant SPDS and PMTs revealed a well supported split between PMT and SPDS from higher plants (Figure 3.2b), suggesting PMT arose from a plant SPDS. These more detailed phylogenetic studies are consistent with previous studies that suggest that PMT (an N-methyltransferase) arose from an ancestral SPDS (a propylaminotransferase).

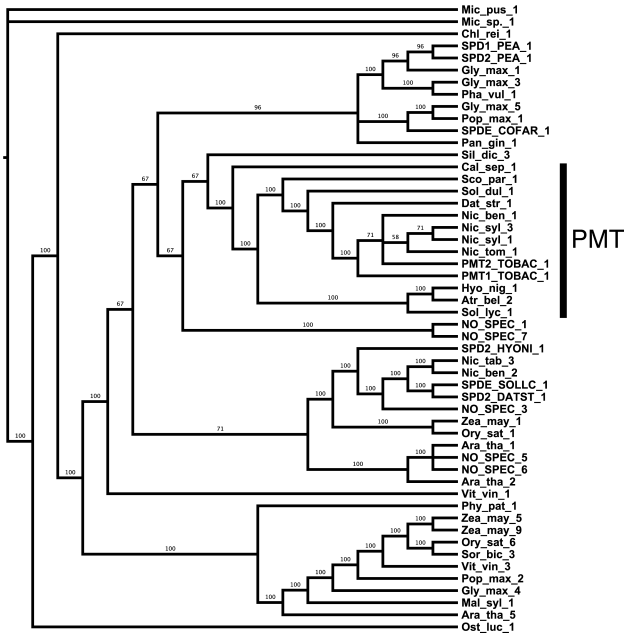
Quantitative measures of positive selection are often associated with protein/enzyme neofunctionalization in other gene families [149, 150]. Therefore, we implemented a test for positive selection

Figure 3.2: PMT/SPDS maximum likelihood phylogeny. a. Maximum likelihood (ML) search consensus tree (100 ML trees) of amino acid (AA) sequences for PMT and SPDS homologs. Plants are colored green, Fungi blue, Animals/Protists/Diatoms red, Cyanobacteria cyan, α-proteobacteria brown, Chlorobi/Bacteroidetes/Flavobacterium orange, Archaebacteria brown. b. ML search consensus tree (100 ML trees) of plant PMT and SPDS homologs. Branches are labeled with how many times that split occurred in 100 ML searches.

(a)



(b)



in plant *PMT* and/or *SPDS*/*SPMS* gene families using the program PAML and implementing the Likelihood ratio test to confirm or reject models of natural selection. The stringent M1a/M2a likelihood ratio test (LRT) compares models with either two or three categories: sites with a d_N/d_S (ω) <1 (purifying/negative selection), sites with $\omega = 1$ (neutral selection), or in M2a, sites with additional $\omega > 1$ (positive/adaptive selection). In tests of just *PMT* sequences, just *SPDS*/*SPMS* sequences, or combined *PMT*/*SPDS*/*SPMS* sequences there was no statistical support for positive selection (Table 3.1). The less stringent site models M7 and M8 model sites with]textomega values in a beta distribution from 0 to 1, with the M8 model including a separate category for sites with an $\omega > 1$. In tests using the *SPDS*/*SPMS* or *PMT*/*SPDS*/*SPMS* combined, the M8 model was significantly better than the M7 model, showing there was evidence of some significant support for positive selection. This might have been due to spermidine synthases and spermine synthases being included in the same dataset, as the phylogenetic analyses could not resolve them independently (Figure 3.2).

Since there were no strong signals of positive selection in the extant *PMT* gene family, we proceeded with the assumption that PMT neofunctionalization may have arose through neutral or nearly-neutral selection. To test this hypothesis, we compared d_N/d_S values at each site between *PMT* and *SPDS* to establish whether differential levels of natural selection were affecting *PMT* vs. *SPDS* at the same sites, and therefore might be important for neofunctionalization of PMT from SPDS. In the previous analyses of 5-epi-aristolochene synthase (*TEAS*) and premnaspirodiene synthase (*HPS*) (Chapter 2), it was shown that nine sites conferring enzyme product specificity were under different selective pressure, and most had a $\Delta d_N/d_S > 0.2$. Therefore, this approach

Table 3.1: *PMT* and *SPDS* PAML site model results. PAML site models were compared using the likelihood ratio test (LRT) on datasets containing *PMT* and plant *SPDS* sequences, just *PMT* sequences, or just *SPDS* genes. Both the *PMT*/*SPDS* combined model and the *SPDS* alone model shows support for positive selection under the M8 model, but not in the more strict M2a model. The combined *PMT*/*SPDS* M8 model has a]textomega >1 category value of 1.0, exhibiting essentially neutral evolution.

PMT/*SPDS* combined

Model	lnL	LRT p-value	ω <1	ω = 1	ω >1
M1a	-24120.6	1	0.0682	1	-
M2a	-24120.6		0.0682	1	1
M7	-23773.5	0.01	-	-	-
M8	-23768.0		-	-	1

PMT alone

Model	lnL	LRT p-value	ω <1	ω = 1	ω >1
M1a	-5278.27	1	0.0336	1	-
M2a	-5278.27		0.0336	1	27.54
M7	-5270.09	1	-	-	-
M8	-5268.91		-	-	1

SPDS alone

Model	lnL	LRT p-value	ω <1	ω = 1	ω >1
M1a	-19238.0	1	0.0659	1	-
M2a	-19238.0		0.0659	1	33.55
M7	-18950.9	0.01	-	-	-
M8	-18945.5		-	-	1.359

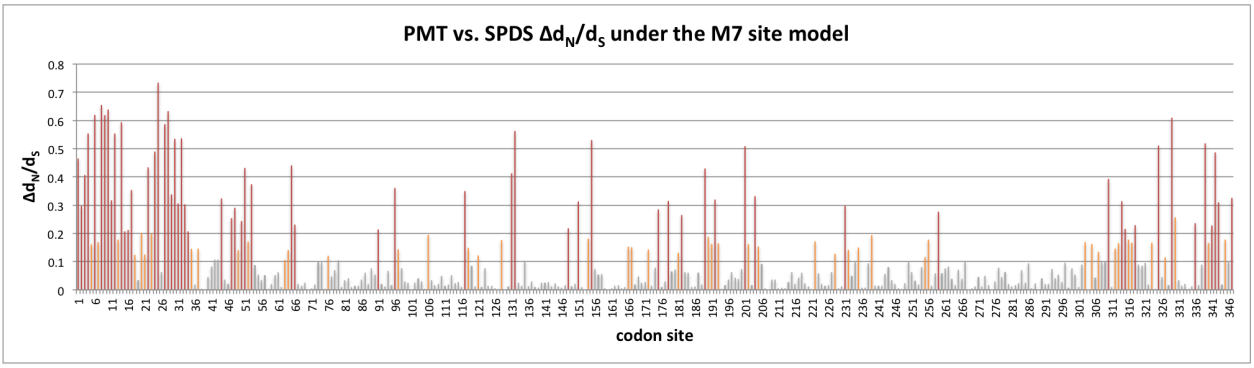
was applied to *PMT* and *SPDS* gene families to identify putative positions in SPDS that could be modified resulting in PMT activity (i.e. rationale protein engineering using signals of past natural selection). In *PMT* and *SPDS*, nine out of a total of 126 sites within 15Å of the presumed active site had a $\Delta d_N/d_S$ greater than 0.2. In order to increase the number of targets in a preliminary study, we considered sites with a $\Delta d_N/d_S$ greater than 0.1. In *PMT* and *SPDS*, 20 sites near the active site met this expanded criteria, 14 of which were nonsynonomous between *StSPDS* and *StPMT*.

Given the hypothesis that PMT likely arose from SPDS and the identification of 14 sites showing $\Delta d_N/d_S$ values between *PMT* and *SPDS*, the next objective was to engineer a modified *SPDS* gene to express a recombinant enzyme with PMT activity. Another rationale for modifying a *SPDS* gene to encode a PMT enzyme activity stemmed from the fact that relevant substrates for PMT activity are readily available, whereas the normal substrate for SPDS dcSAM must be custom synthesized and is prohibitively expensive. A recombinant *StSPDS* gene with nucleotide changes resulting in the 14 nonsynonymous codon substitutions from *StPMT* (St*SPDS*^{PMT14}) was custom synthesized and cloned into an expression vector. Recombinant StSPDS, NtPMT, and StSPDS^{PMT14} proteins were expressed and purified from *E. coli* to perform *in vitro* PMT enzyme assays.

PMT enzyme assays were initially performed using two linked-enzyme assays approaches. The first was a dual-enzyme (SAHN and adenine deaminase) spectrophotometric assay [148]. This assay measures SAM-dependent methylation using two recombinant enzymes to convert the PMT reaction product SAH into hypoxanthine, accumulation of which can be measured as a decrease in absorbance at 265nm using a UV spectrophotometer over time. The second PMT linked enzyme assay (SAHN and LuxS) was a colorimetric assay that uses SAHN and LuxS to sequentially con-

Figure 3.3: Change in selective pressure mapped to SPDS structure. a. Change in the strength of natural selection ($\Delta d_N/d_S$) between PMT and plant SPDS at each amino acid site. Sites with a $\Delta d_N/d_S$ >0.1 are colored orange, sites with a $\Delta d_N/d_S$ >0.2 are colored red. b. Structural 3D homology model of StSPDS. Sites with a $\Delta d_N/d_S$ >0.1 and within 15Å of the presumed active site (residue 131D) are colored orange; sites with a $\Delta d_N/d_S$ >0.2 and within 15Å of the active site are colored red. Substrate analog S-adenosyl-1,8-diamino-3-thiooctane with coordinates from PDB: 2I7C colored green.

(a)



(b)

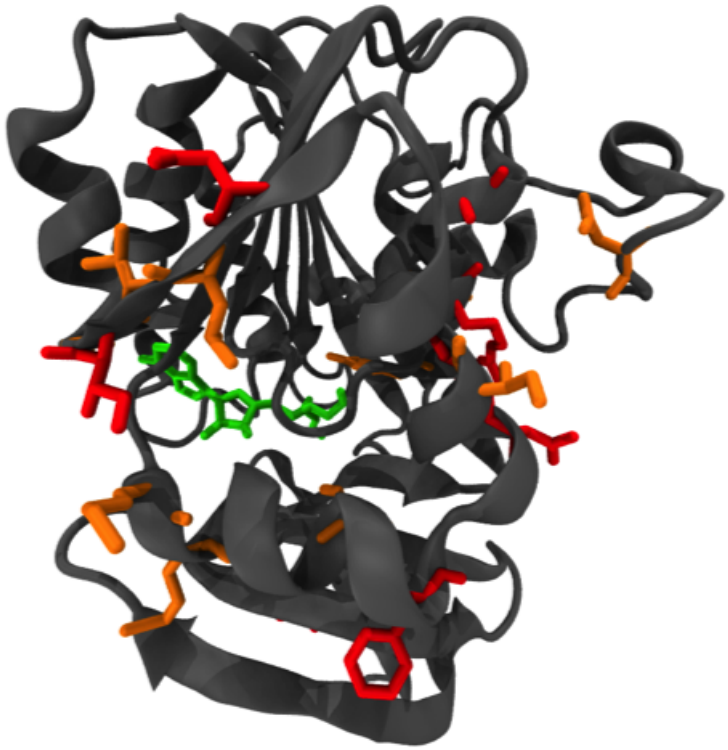
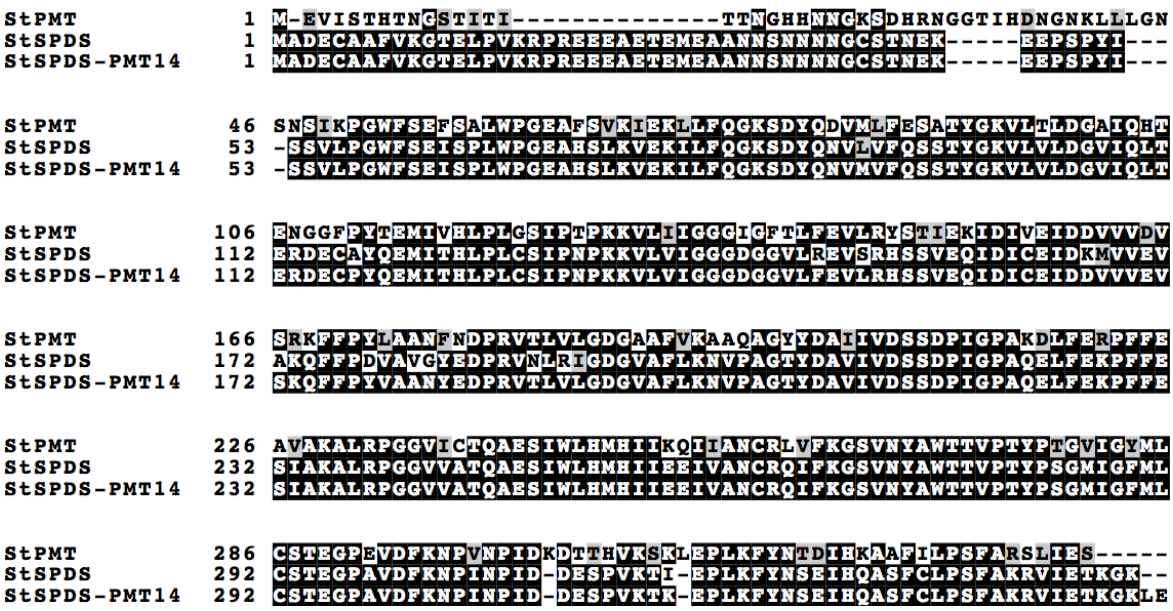


Figure 3.4: Alignment of amino acid sequences for StPMT, StSPDS, and StSPDS^{PMT14}. Sites differing in StSPDS and StSPDS^{PMT14} are those substituted in wild type StSPDS with their equivalent in StPMT to form StSPDS^{PMT14}.

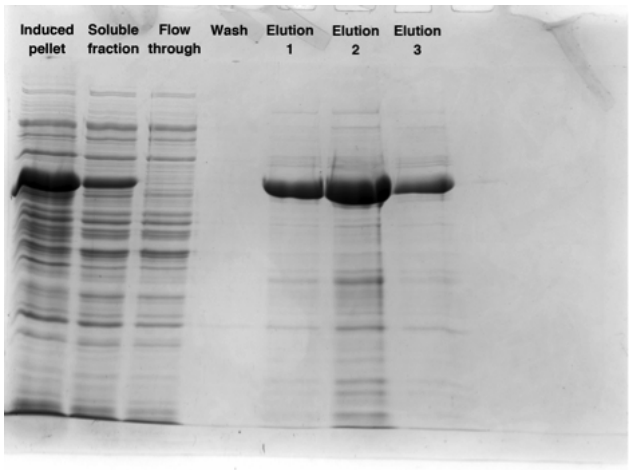


vert SAH into homocysteine, which can be quantified by addition of 5,5'-dithiobis-2-nitrobenzoic acid (DTNB) which produces a yellow product upon spontaneous reaction with thiol groups [151]. While control recombinant NtPMT activity was readily measured in both linked enzyme assays, neither assay was particularly sensitive in terms of signal to noise ratio (data not shown). Recombinant StSPDS^{PMT14} enzyme did not demonstrate reproducible PMT activity in either of these linked enzyme assays. Given the overall low signal to noise observed with the positive control recombinant NtPMT enzyme, together with the possibility that nascent neofunctional PMT activity in the recombinant StSPDS^{PMT14} might be substantially less than an extant PMT [82], a more sensitive PMT assay was required.

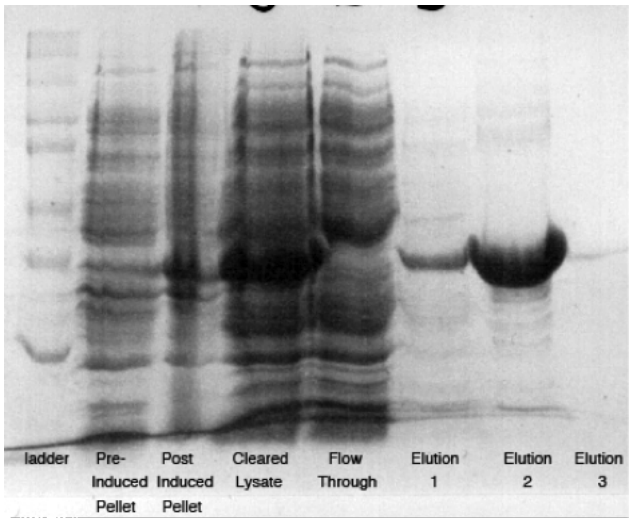
To test for very low levels of N-methylputrescine, a UPLC-based approach which utilizes a fluorescent tag to identify very low concentrations of primary amine containing metabolites was used.

Figure 3.5: Purified recombinant protein for a. NtPMT with MBP, b. StSPDS, and c. StSPDS^{PMT14}. Recombinant protein was expressed in *E. coli* with a C-terminus 6xHIS tag for purification on Ni-NTA resin.

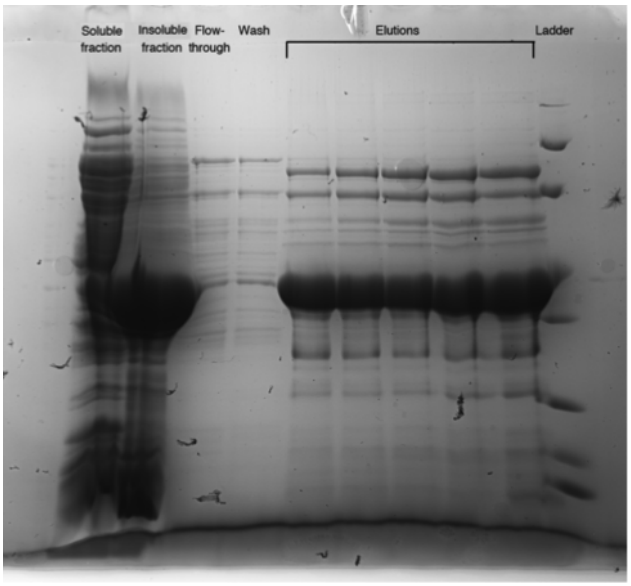
(a)



(b)

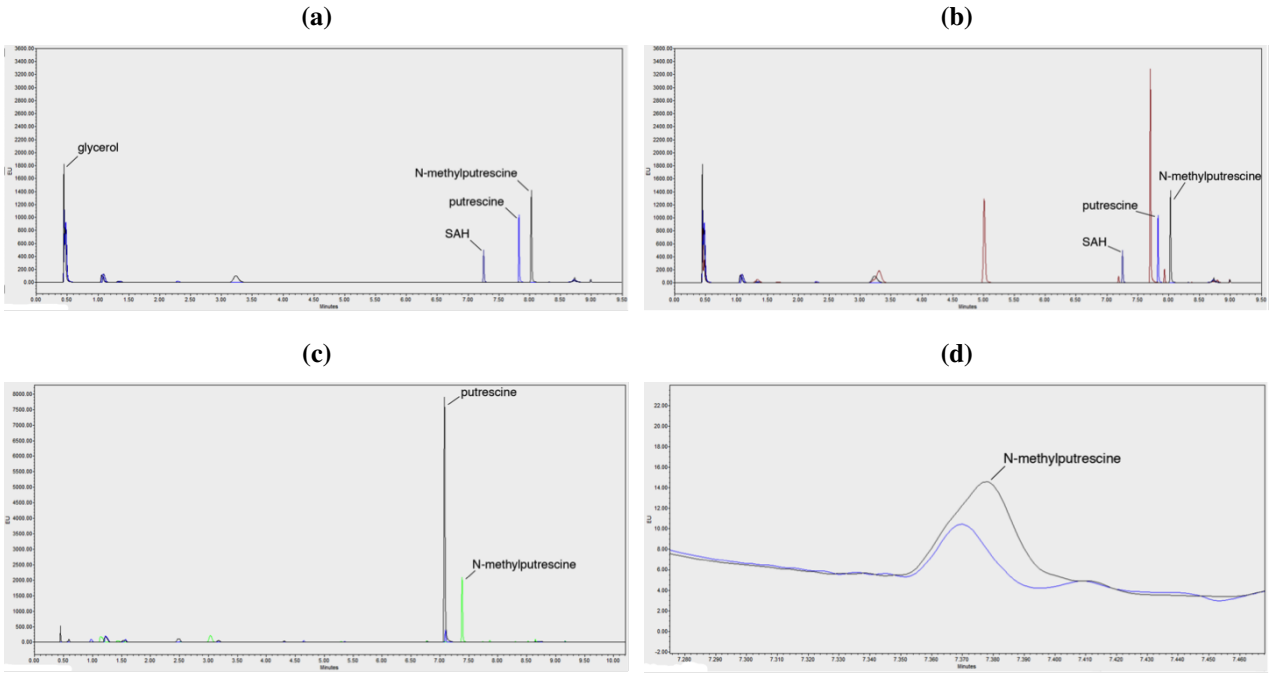


(c)



An AccQ-Tag Ultra Derivatization Kit was used to derivatize primary amines present upon completion of standard PMT enzyme assays. These were then separated on a UPLC reverse phase chromatography system using a fluorescent detector. When derivatized and run individually on the UPLC, each substrate exhibited clear separation from other substrates. A peak corresponding to derivatized SAM was never identified, however derivatized forms of SAH, putrescine, and N-methylputrescine each eluted from the column in non-overlapping peaks (Figure 3.6a). Assays containing recombinant NtPMT with putrescine and SAM resulted in a peak corresponding to derivatized N-methylputrescine; though all derivatized products eluted from the column earlier (shifted to the left) than the controls of individual substrates, most likely due to residual protein binding to the column (Figure 3.6b). Assays performed as a negative control, containing recombinant StSPDS, putrescine, and SAM, did not produce a peak corresponding to derivatized N-methylputrescine. Choosing concentration amounts for SAM was difficult, as large amounts have been found to inhibit PMT activity [82], while dilute amounts may not be enough for producing measurable amounts of N-methylputrescine in an enzyme with weak PMT activity. Multiple assays were performed for each recombinant enzyme. In assays of recombinant StSPDS^{PMT14} (Figure 3.6c), a small peak was observed close to the elution time for N-methyputrescine (Figure 3.6d). However, spiking this assay with purified N-methylputrescine revealed that these formed two distinct peaks, and thus it was not N-methylputrescine but some other derivatized molecule with a primary amine. Speculatively, if the modifications in StSPDS^{PMT14} allow it to still perform SPDS-like activity but recognize SAM as substrate instead of dcSAM, it may be catalyzing aminocarboxyl-propyltransferase activity and produce a spermidine-like molecule with a carboxyl group

Figure 3.6: Enzyme assays quantified by UPLC. a. Separation of derivatized standards. b. NtPMT assay (red), putrescine standard (blue), N-methylputrescine standard (black), SAH standard (dark blue). c. StSPDS^{PMT14} assay (black), putrescine standard (blue), N-methylputrescine standard (green). d. StSPDS^{PMT14} assay (blue), same assay spiked with N-methylputrescine (black).



attached the same (first) carbon as one of the amine groups. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) could be used to identify this unknown reaction product. Regrettably, the UPLC results did not indicate any detectable PMT activity in the recombinant StSPDS^{PMT14}, even at very low product detection levels afforded by the UPLC assay.

3.5 Discussion

The molecular evolution of PMT from SPDS was originally proposed [152] based upon rather simplistic neighbor-joining phylogenetic analyses [134, 152, 101]. Here we present a deep model-

oriented maximum likelihood phylogeny of PMT and SPDS proteins from across the tree of life (Figure 3.7). As expected, spermidine synthases separated according the standard tree of life. A clade containing all known plant PMTs branched off from plant SPDS/SPMS, implying a plant specific eukaryotic origin for PMT, rather than a cyanobacterial origin during the endosymbiotic origins of the chloroplast. The BLAST search to identify the closest homologs of *N. tabacum* PMT1 (used to create this tree) did not find any other N-methyltransferases (besides PMTs), or any O-methyltransferases. This phylogenetic evidence more fully supports previous assertions that an N-methyltransferase (PMT) evolved from an animopropyltransferase (SPDS), rather than from a pre-existing N-methyltransferase.

This study used quantitative measures of differential signals of extant natural-selection to guide a rational redesign of a SPDS enzyme with PMT activity. There are several possible reasons why this approach did not result in the neofunctionalization of PMT activity in the recombinant $StSPDS^{PMT14}$ protein. Changing amino acids in second-tier locations relative to the active site may have altered the shape of the active site, which interrupted the interactions between catalytically important residues and the substrates. These substitutions may also have changed the active site cavity so that putrescine or SAM can no longer fit within or enter the active site. These mutations did not completely denature $StSPDS^{PMT14}$, as the recombinant protein was soluble after purification under native conditions, though an admittedly large amount was insoluble during the over-expression in *E. coli* (Figure 3.5c). In this case, the $\Delta d_N/d_S$ metric identified sites that indirectly affect enzyme function, but not those directly responsible for enzyme activity specificity.

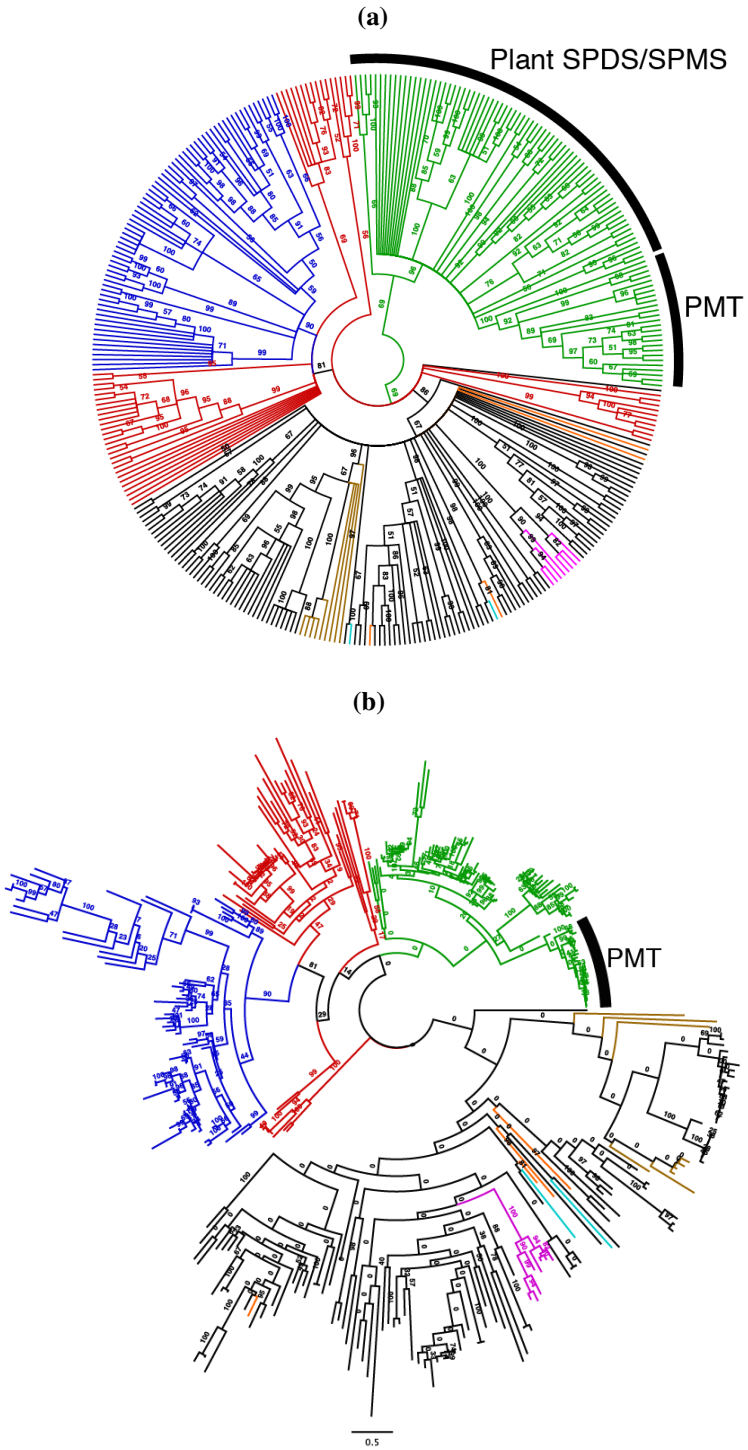
Recently, the Dräger lab published results showing that they were able to modify *D. stramonium*

SPDS to have PMT activity with only one site-directed substitution (D103I), using homology modeling and ligand docking methods [77]. This enzyme had no SPDS activity and extremely low N-methyltransferase activity (kcat [1 s^{-1}] 0.84 x 10^{-3}). The addition of two more changes (Q79T and V106T) increased *PMT* activity (kcat [1 s^{-1}] 0.15). In similar experiments with *A. thaliana* SPDS (*Arabidopsis* does not produce N-methylputrescine), substitutions from PMT at three sites corresponding to those in DsSPDS resulted in a bifunctional SPDS/PMT enzyme. Our quantitative measure of natural selection computational analyses indicated that the sites used by the Dräger lab were under strong purifying selection in both *PMT* and *SPDS* ($\Delta d_N/d_S$ <0.1) and therefore were not identified by our approach emphasizing different patterns of natural selection.

These results indicate that the evolutionary history of *PMT* and *SPDS* was most likely markedly different than that of *TEAS* and *HPS*. The *TEAS*/*HPS* gene family appears to show two different patterns of natural selection: *TEAS* showing more signals of relaxed/weak purifying selection, whereas *HPS* appears to have gone through a recent selective sweep as evidenced by more sites under strong purifying selection. Different patterns of natural selection at important sites for neofunctionalization may make the $\Delta d_N/d_S$ metric a useful tool for identifying important sites responsible for enzyme function. On the other hand, extant *PMT* and *SPDS* gene families both appear to be under strong purifying selection (albeit for different amino acid composition) at sites important for catalytic activity, in which case the $\Delta d_N/d_S$ metric would not indicate the importance of these sites for conferring different substrate specificity and reaction products.

There are other differences between these two model enzyme systems that are worth noting. The active site responsible for substrate binding in TEAS and HPS are almost identical, because the

Figure 3.7: Maximum likelihood non-parametric bootstrap phylogenies. a. Non-parametric bootstrap consensus tree of PMT and SPDS homologs. b. Best overall maximum likelihood phylogeny of PMT and SPDS homologs with non-parametric bootstrap values mapped to branches.



substrates are in fact identical, so it is the shape of the overall catalytic site that confers reaction product specificity. This is not so in SPDS and PMT, as different substrates (dcSAM and SAM, respectively) require different amino acids in the active site to both provide substrate specificity as well as reaction pathway selectivity (between the transfer of an aminopropyl group or a methyl group, respectively). PAML codon site models did not find evidence for positive selection in the extant *PMT* gene family. Given the phylogenetic evidence it is likely that an ancestral *SPDS* acquired a few mutations conferring low levels of N-methyltransferase activity that were immediately advantageous (and thus selected for), consistent with adaptive evolution. Further mutations then increased this activity or provided additional enzyme stability as both enzymes diverged, gradually erasing the signal of positive selection in modern genes as strong selection or selective sweeps removed less advantageous mutations. The mutations required to increase PMT activity may also be detrimental to protein stability, as seen by the apparent relative structural instability of the modern PMT protein compared with SPDS [82]. Computational and biochemical experiments using ancestral state reconstruction could shed light on the more ancient versions of these proteins to further resolve how this may have originally occurred.

One of the limitations of the PAML codon site models is that they only allow d_N/d_S values to vary at individual sites in an alignment, but not over branches in a tree. On the other hand, the PAML branch models allow for d_N/d_S values to vary at individual branches on a phylogenetic tree, but not at individual sites. The PAML branch/site models allow for d_N/d_S values to vary at individual sites, but also can detect positive selection at different points in evolutionary history at specific splits in a tree. These models are still somewhat unstable [153], but may be more condusive to identifying

how and when neofunctionalization occurred in *SPDS*.

Additionally, phylogenetic analyses using a Bayesian Metropolis-coupled Markov chain Monte Carlo (MCMCMC) approach performed by MrBayes [154] with 6 runs of 12 chains each did not converge on a single topology, despite running for 30 million generations (data not shown), further exemplifying the uncertainty in this dataset. At the beginning of this analysis we pruned every other sequence from the initial BLAST search to reduce the computational power needed while still representing the overall topology of taxa. Some plant PMT or SPDS/SPMS taxa may have been lost that could further resolve the large, basal polytomies we observed in the clade containing Viridiplantae.

These experiments were all performed using a maximum likelihood search consensus tree as the basis for all phylogenetic analyses. We later learned that this method does not represent the most likely tree, but that rather the single best ML search tree with non-parametric bootstrap support for branches (Figure 3.7b) is a more likely representation. This tree resolves the large polytomy in plants, and shows that the PMT/SPDS split most likely occurred very early on in plants, as only SPDS from single-cell green algae (*Micromonas pusilla*, *Ostreococcus lucimarinus*, and *Chlamydomonas reinhardtii*) are found basal to the clade containing PMT. However, non-parametric bootstrap branch support for these splits are small or non-existent, and collapsing branches with low bootstrap support would result in a tree with large polytomies similar to the ML search consensus tree.

Nucleotide and protein database have been growing quickly in recent years. As the BLAST search used as a basis for these experiments was performed in 2010, additional SPDS or PMT proteins

may have been identified and submitted to NCBI databases that could provide more signals of past natural selection.

Chapter 4

The polyphyletic molecular evolution of N-methyltransferases in plant alkaloid metabolism

Alexandra J Weisberg and John G Jelesko.

4.1 Abstract

Plants produce a vast array of diverse small-molecule chemicals. However, our understanding of the evolutionary origins of the underlying enzymatic and metabolic diversity is becoming a topic of increasing interest. In most cases, enzymes responsible for a particular chemical transformation

diversified from a common ancestor given rise to large superfamilies. For example, O-methyltransferases typically have similarity to other O-methyltransferases. Unlike the O-methyltransferases, initial observations of plant N-methyltransferases suggest a different pattern: one of possible polyphyletic origins. However, no systematic study has focused on this topic. Therefore, molecular evolution analyses on several different families of N-methyltransferases were performed to determine whether they share common ancestry. Additional analyses of N-methyltransferases involved in the caffeine biosynthesis pathway in coffee (*Coffea arabica*) and tea (*Camellia sinensis*) plants found signatures of adaptive selection, as well as provided statistical support for the independent molecular evolution of caffeine biosynthesis.

4.2 Introduction

4.2.1 Evolution of N-methyltransferases

Plants produce a tremendous number of small-molecule products, most with an as yet unknown function. While the total number of primary metabolites is around 10,000 [8], the number of known plant specialized metabolites (those not strictly necessary for growth and development) is enormous, with total estimates of over 200,000 unique metabolites among plants [1]. These metabolites presumably confer some selective advantage to the plant, such as defense against herbivory or pathogens, attraction of pollinators, or temperature regulation. The large number of enzymes used to produce these molecules have been observed to have evolved from enzymes in

primary metabolism through gene duplication and divergence [28, 23, 155], or from other specialized metabolism enzymes through divergence of orthologs in different species [106, 72, 156]. This conventional descent with modification should give rise to enzymes performing similar chemistries (albeit with different substrates) and these should organize as large monophyletic clades in phylogenetic analysis. In cases of rapidly diversifying enzymes this will give rise to large superfamilies.

The large plant O-methyltransferase superfamily is an example of this. These enzymes catalyze the transfer of a methyl group from S-adenosylmethionine (SAM) to hydroxyl groups on a wide variety of secondary metabolites. They participate in the methylation of flavonoids, caffeic acids, lignin precursors, benzylisoquinoline alkaloids, phenylpropanoids, and many other substrates [157, 4]. Phylogenetic analyses of representatives of these various O-methyltransferases suggest that they all form a single monophyletic clade, and that all likely share a single common ancestor [157]

The independent convergent evolution of some enzyme activities in plant specialized metabolism has also been suspected [8]. Unlike the O-methyltransferases, which all appear to share a single common ancestor, plant N-methyltransferases appear to have arisen independently from various other genes encoding enzymes with different catalytic activities [158]. For example, putrescine N-methyltransferase (PMT) is an enzyme that catalyzes the SAM-dependent methylation of putrescine to form N-methylputrescine, and is the first committed step in nicotine, tropane, and nortropane alkaloid pathways found in various members of the Solanaceae [135, 136]. It has common ancestry with spermidine synthase, an aminopropyltransferase (Figure 3.7b; Chapter 3). Phylogenetic analyses and *in vitro* enzyme assays of modified recombinant protein have shown that PMT most likely evolved from a spermidine synthase, a primary metabolic enzyme [77].

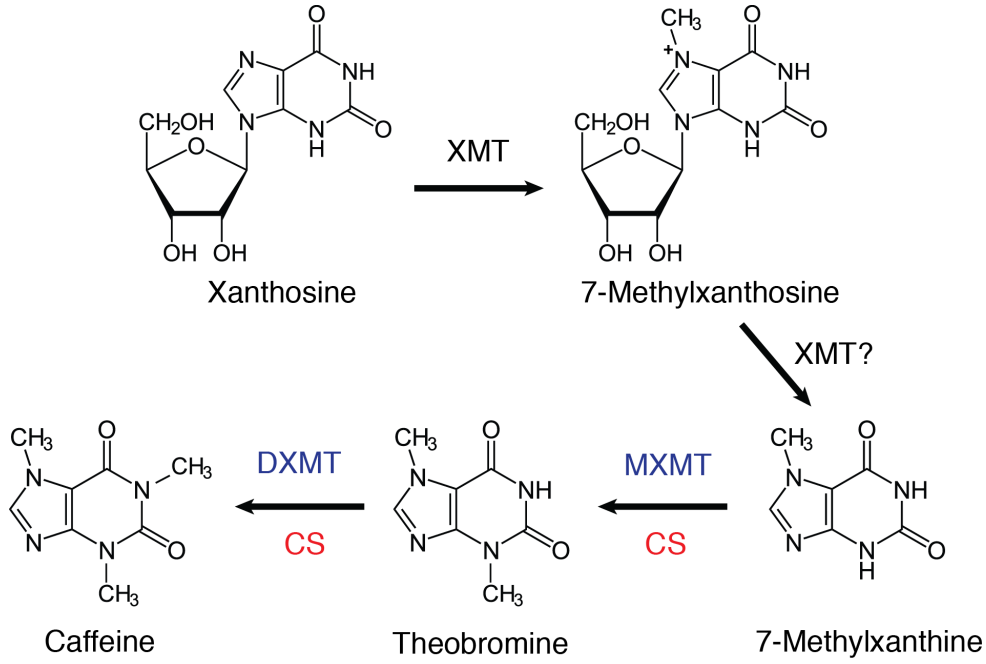
Coclaurine N-methyltransferase (CNMT) and tetrahydroprotoberberine N-methyltransferase (TNMT) are two plant secondary metabolic enzymes involved in benzylisoquinoline biosynthetic pathways in the Papaveraceae, and in the order Ranunculales [4]. CNMT is found in the morphine biosynthetic pathway in opium poppy (*Papaver somniferum*) and in the berberine biosynthetic pathway in *Coptis japonica* [159, 160]. TNMT catalyzes the formation of intermediates in the sanguinarine biosynthetic pathway and has been cloned from opium poppy [158]. TNMT is similar to CNMT (45-48% amino acid identity) indicating a likely common ancestor with N-MT activity [4].

4.2.2 Caffeine biosynthesis

A number of plants have evolved to produce the purine alkaloid caffeine (1,3,7-methylxanthine), including those used in the making of coffee (*Coffea arabica*) and tea (*Camellia sinensis*), among others. This secondary metabolite provides a defense against herbivory by interfering with the central nervous system of insects, slugs, and other potential herbivores [161, 162, 163, 164]. Caffeine released by seeds into soil has also been hypothesized to play an allopathic role in preventing competing seeds from germinating [165]. Theobromine (3,7-methylxanthine), the immediate precursor of caffeine, is also found in the fruit of the cocoa tree (*Theobroma cacao*) and is responsible for the mild stimulatory effect of chocolate [166, 167]. Caffeine comprises upwards of 2.4% of coffee bean dry weight [31], and around 2.3% of the dry weight of young tea leaves [32]. Caffeine has been ingested for thousands of years by various cultures for its stimulant properties and is possibly the most widely-used drug in the world today [168], consumed by upwards of 90% of North

Americans.

Figure 4.1: The major caffeine biosynthetic pathway in coffee and tea plants. Both coffee and tea contain xanthosine N-methyltransferase (XMT), which may also remove the ribose group in the second step. *Coffea arabica* contains 7-methylxanthine N-methyltransferase (MXMT) as well as a bifunctional dimethylxanthine N-methyltransferase (DXMT), which can also perform the second methylation. *Camellia sinensis* contains only a dual-function caffeine/theobromine synthase (CS) that performs both the second and third methylations of xanthine. This figure adapted from Ziegler and Faccini, 2008.



The caffeine biosynthetic pathway in plants is comprised of four distinct steps, including three N-methylations performed by varying numbers of enzymes (Figure 4.1). The initial substrate, xanthosine, is derived from either purine alkaloids, S-adenosylmethionine (SAM), or *de novo* biosynthesis [32]. A methyl group is transferred from SAM to xanthosine by xanthosine N-methyltransferase (XMT) to form 7-methylxanthosine [169, 170]. In *C. sinensis* N-methylnucleosidase cleaves 7-methylxanthosine to form 7-methylxanthine [171, 172, 173]. This step can also be catalyzed by XMT in *Coffea* in *in vitro* enzyme assays [174]. Either theobromine synthase [172, 173] or a

dual-function theobromine/caffeine synthase [175] further methylates 7-methylxanthine to form theobromine, which is then converted to caffeine by caffeine synthase.

A gene encoding *C. sinensis* caffeine synthase (*TCS1*) was first cloned by the Ashihara lab at Ochanomizu University, Tokyo in the late 90s [175, 176]. A paralog of *TCS1* was later identified and called *TCS2*, but it has not yet been characterized [177]. Dual-function caffeine/theobromine synthases (CaDXMT, CCS1, CtCS7) have since been cloned from various *Coffea* species [170, 169]. Single-function theobromine synthases, that catalyze the methylation of 7-methylxanthine to theobromine but lack caffeine synthase activity, are biochemically characterized (CTS1, CTS2, CaMXMT1 and CaMXMT2) [172, 173]. A gene encoding the first and possibly second step in the pathway, xanthosine N-methyltransferase (*XMT*), is cloned from *C. arabica* [169]. The x-ray crystal structures of XMT (PDB: 2EG5) and DXMT (PDB: 2EFJ) are solved [174].

Caffeine biosynthetic proteins in *Coffea* have high similarity among themselves (~80% AA identity), but only have 40% AA identity with caffeine synthases from tea. This relatively low similarity, coupled with neighbor-joining phylogenetic trees of representatives of each enzyme type, suggest that caffeine biosynthesis in tea and coffee arose independently [177, 8]. The present study represents a coordinated systematic investigation of the deep molecular evolution of plant N-MTs specifically focused on the question of monophyletic or polyphyletic origins of plant N-MTs.

4.3 Methods

4.3.1 Reciprocal homology searches

Representatives of putrescine N-methyltransferase (*Nicotiana tabacum PMT1*, accession Q42963.1), xanthosine N-methyltransferase (*Coffea arabica* XMT1, accession Q9AVK0.1), coclaurine N-methyltransferase (*Coptis japonica* CNMT, accession BAB71802.1), and tetrahydroprotoberberine N-methyltransferase (*Papaver bracteatum* TNMT1, accession C3SBU5.1) were each used as queries in Smith-Waterman searches [178] of a local NCBI nr database using the program SSEARCH 35.04 [179] with the BLOSUM50 matrix and an e-value cutoff of 10. These searches were performed on the NCBI nr database as of February 27, 2013. The result of each similarity search was compared to the three other searches, and shared hits were identified using a custom Perl script utilizing the BioPerl library [180].

4.3.2 Phylogenetic analyses

Phylogenetic analyses of putrescine N-methyltransferase were performed as described in Chapter 3.

Coffea arabica XMT1 (accession BAC75665.1) was used as a query in a BLASTP search of the NCBI nr database using the BLOSUM62 matrix and an e-value cutoff of 10^{-20}. The program MAFFT e-INS-i [181] was used to assemble a multiple sequence alignment (MSA) of all BLAST hits. The program ProtTest 3.2 [123] identified the JTT+G+F substitution model as the best model

for this alignment. A 100 replicate maximum likelihood (ML) search was performed using GARLI 0.961b [124]. An additional 100 replicate ML search with *Coffea* and *C. sinensis* sequences constrained together was also performed using GARLI.

The approximately unbiased (AU) test [182] as performed by the program Consel [183] was used to compare the best likelihood XMT unconstrained ML search tree with the best tree with *Coffea* and *Camellia* constrained together.

Sequences from the clades containing XMT-like proteins from *Coffea* or *Camellia* (as identified in the first phylogenetic tree) were also used in individual phylogenetic analyses. The JTT+G+F model was again found to be the best model for each of these datasets, and 100 replicate ML searches were performed using GARLI.

A PAML [83] codon site model analysis was performed on datasets of DNA coding sequences corresponding to the XMT-like proteins identified in the above phylogenies. These sequences were aligned to the original protein alignment using the PAL2NAL web server [127]. The best-fitting codon frequency model for both the *Coffea* dataset and the *Camellia* dataset was identified by PAML as F61 (individual frequency parameters for each codon). Additional analyses using the F3X4 model (codon frequencies estimated from nucleotide frequencies at all three codon positions) yielded comparable results. The PAML site models M0, M3, M1a, M2a, M7, and M8 were used on each dataset and likelihood ratio tests (LRT) were performed between the M0 and M3 models (degrees of freedom = 4, the M1a and M2a models (df = 2), and the M7 and M8 models (df = 2).

4.4 Results

4.4.1 Polyphyly of plant N-methyltransferases

Due to the transitive nature of homology (*sensu stricto*, shared common ancestry), shared hits in reciprocal similarity searches can indicate shared common ancestry, even if the two query sequences have very low sequence similarity [184]. For example, if protein A is homologous to protein B, and B is homologous to protein C, then by definition A is homologous to C (Figure 4.2a), even if A does not identify C in a similarity search. Most sequence similarity searches are interested in identifying close homologs and the BLAST algorithm is well suited for this purpose. However, BLAST is not as sensitive as the Smith-Waterman algorithm at accurately finding very distantly related sequences. The virtue of BLAST is that it is fast at finding closely related sequences, whereas the Smith-Waterman algorithm is quite slow, but is better at aligning and scoring more distantly related sequences. The expectation values of all similarity search algorithms are influenced by the size of the sequence database, so all Smith-Waterman searches were performed on the same local database. The key to this study was using criteria that enables the assignment of homology based upon sequences identified in reciprocal similarity search hits, regardless of expectation value [184, 185]. The four search query sequences used in the Smith-Waterman searches were PMT, XMT, CNMT, and TNMT with an expectation cutoff value of 10 representing different plant N-methyltransferases involved in specialized metabolism (Figure 4.2b). Searches returned anywhere from 742 to 9241 hits with an expectation value of 10 or lower. None of the hits identified using PMT as query were shared by queries using XMT, CNMT, or TNMT (Figure 4.2c).

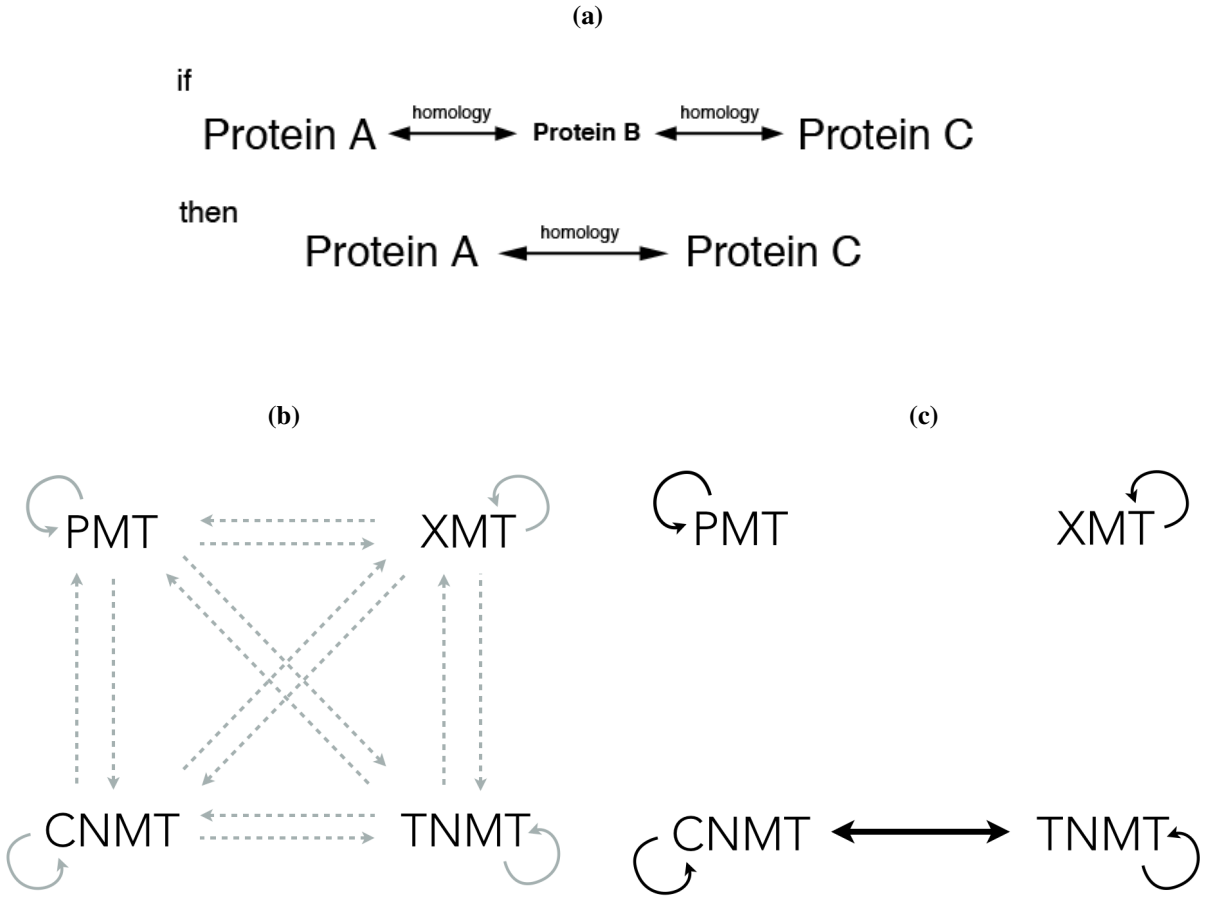
As was previously described in Chapter 3, PMT is most similar to plant spermidine and spermine synthases, and PMT searches found no other N-MTs used as queries in this study.

Similarly, none of the hits identified using XMT as query were found in the CNMT or TNMT query results. XMT is most similar to SABATH family methyltransferases, such as jasmonate O-methyltransferase, as well as salicylic acid and benzoate carboxyl-methyltransferases, as described elsewhere [32]. A single shared hit was found in common in the XMT, CNMT, and TNMT searches: a cyclopropane-fatty-acyl-phospholipid synthase from the cyanobacteria *Gloeobacter violaceus* PCC 7421 (accession NP_924885.1). This 669 amino acid protein had very low AA identity (13.9%-16.3%) to either XMT, CNMT, or TNMT; and appears to be an unusual gene fusion. Alignments of XMT or CNMT with this protein revealed that CNMT only aligned with the N-terminus half while XMT only aligned to the C-terminal half. BLASTP searches of the NCBI nr database using the N- and C-terminus halves of this protein found only CNMT-like sequences or only XMT-like sequences, respectively (data not shown). This suggests that this prokaryotic protein may have originated through a gene fusion event and is most likely not indicative of distant homology between XMT and CNMT/TNMT within the Plantae.

CNMT and TNMT appear to be very closely related, with searches of one finding the other, as well as 4,766 shared search hits between the two queries. The most similar proteins to CNMT and TNMT were each other, as well as pavine N-methyltransferase and cyclopropane-fatty-acyl-phospholipid synthases from all branches of the tree of life.

Thus, using the broadest criteria for identifying presumed homologs (common hits in reciprocal sensitive similarity searches of a shared database) there was no evidence supporting common an-

Figure 4.2: Reciprocal similarity searches to identify distantly related homologs of N-methyltransferases. a. Shared hits in reciprocal similarity searches can indicate shared ancestry due to the transitive nature of homology. b. If all plant N-methyltransferases involved in secondary metabolism are monophyletic, reciprocal searches using representatives of putrescine N-methyltransferase (PMT), xanthosine N-methyltransferase (XMT), cochlaurene N-methyltransferase (CNMT), and tetrahydroprotoberberine N-methyltransferase (TNMT) should all find shared hits. c. Significant reciprocal shared hits were only found between CNMT and TNMT N-methyltransferases.



cestry between PMT, XMT, and CNMT/TNMT type N-methyltransferases.

4.4.2 Molecular evolution of caffeine biosynthetic enzymes

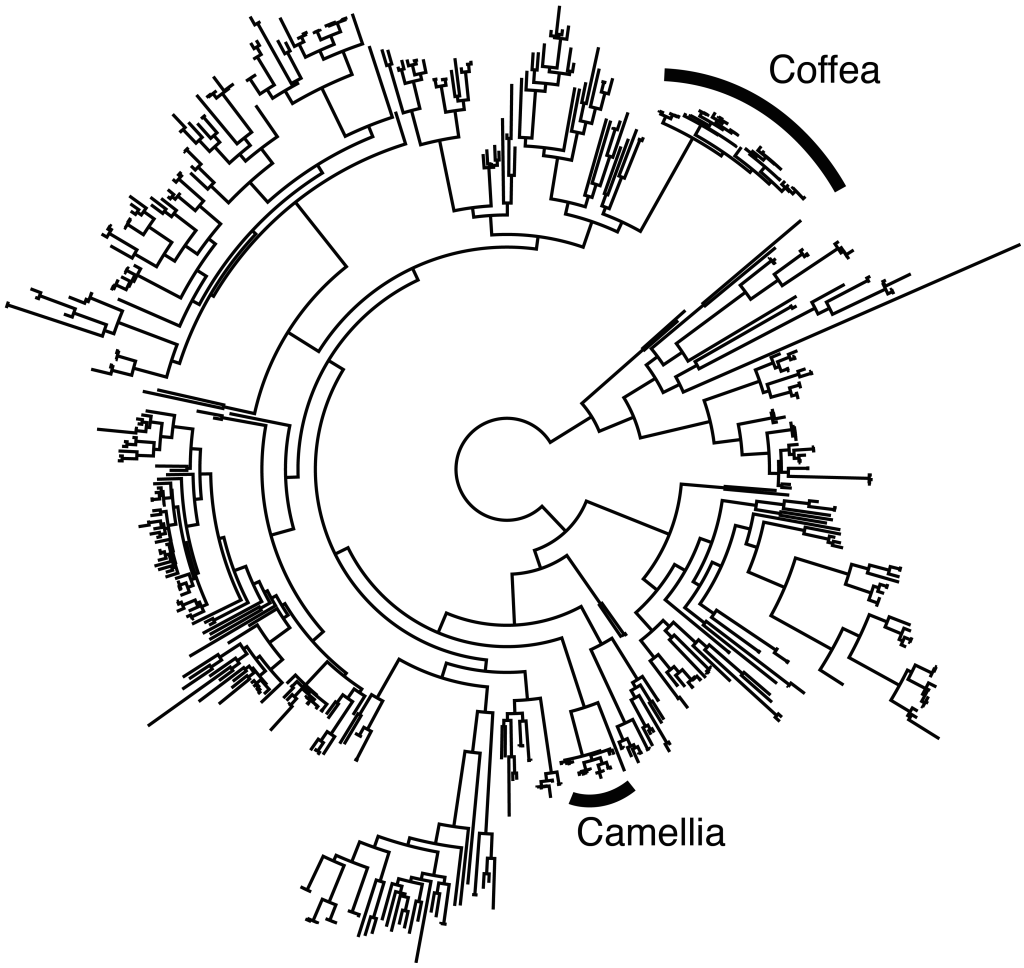
It has previously been suggested that due to the relatively low sequence identity between enzymes with the same catalytic activity and their separate clustering in neighbor-joining trees, the enzymes involved in the caffeine biosynthetic pathway evolved independently in coffee and tea plants [177, 8]. This hypothesis was further investigated using a deep maximum likelihood phylogeny of XMT protein sequences with homologs from a variety of organisms (Figure 4.3).

Caffeine biosynthetic proteins from *Coffea* and *Camellia sinensis* clustered in two distinct clades with multiple branches between them (Figure 4.3). Enzymes with equivalent function in tea and coffee did not group together as sister clades as would be expected for truly orthologous enzymes, but rather all caffeine biosynthesis proteins grouped together by organism. When the clades containing orthologous *Coffea* or *Camellia* proteins were constrained together in ML searches, the resulting best log-likelihood tree had a significantly worse log-likelihood than the best unconstrained tree (Table 4.1), suggesting XMT-like proteins arose independently in tea and coffee plants.

In separate, individual phylogenies of proteins from clades containing XMT in either *Coffea* (Figure 4.4a) or *C. sinensis* (Figure 4.4b), proteins representing each step in the pathway formed individual groups, particularly in *Coffea*.

In *Coffea*, several proteins with unknown function are most basal in the tree (Figure 4.4a). Recombinant proteins of *C. arabica* CS3 and CS4 do not exhibit N-methyltransferase activity in

Figure 4.3: Unrooted xanthosine N-methyltransferase maximum likelihood phylogeny. XMT and other caffeine biosynthetic proteins from *Coffea* or *Camellia* are most closely related to SABATH family member proteins, particularly carboxyl- and O-methyltransferases. Proteins from coffee and tea plants group separately, suggesting caffeine biosynthesis arose independently at least twice, through the convergent evolution of caffeine biosynthetic enzymes. Theobromine synthase from *Theobroma cacao* (cocoa tree) has also been characterized and groups with the clade containing *C. sinensis* proteins.



in vitro enzyme assays [169]. The next deepest branch contains dual-function dimethylxanthine N-methyltransferases (DXMT1, CCS1), which can produce both theobromine and caffeine from 7-methylxanthine or theobromine, respectively. Finally, branches containing either the single function theobromine synthases (MXMT, MXMT2, CTS2) or XMT appear to be the most derived and

Table 4.1: XMT alternate topology hypothesis testing. An alternate phylogeny where XMT-like proteins from *Coffea* and *C. sinensis* are constrained together (suggesting a single origin) was compared to the unconstrained phylogeny where coffee and tea caffeine biosynthetic genes arose independently.

Topology	AU test p-value	standard error
Unconstrained (polyphyly)	1.000	0.000
XMT forced monophyly	10^{-4}	0.000

had arisen after the dual-function caffeine synthases. This is somewhat surprising, because XMT performs the first step in the pathway and possibly the second, yet appears to have evolved more recently in evolutionary history. This step may have evolved to increasingly divert metabolic substrates towards the production of the now selectively-advantageous metabolite caffeine. It is also possible that the ancestral caffeine synthase was more promiscuous and performed this activity as well.

There are fewer characterized caffeine biosynthetic enzymes in *C. sinensis* than in *Coffea*. The tea plant contains two known dual-function caffeine synthases (CsTCS1 and CsTCS2) that catalyze the transfer of a methyl group to either 7-methylxanthine or theobromine. A functional equivalent of XMT has not yet been cloned from tea. These two caffeine synthases form two separate branches comprised of apparent paralogs (Figure 4.4b). Theobromine synthase from *T. cacao* grouped closest to the clade containing *C. sinensis* proteins. However, this may be due to the lack of *T. cacao* protein sequences in the nr database at the time of the initial BLAST search.

Figure 4.4: Unrooted maximum likelihood phylogenies of clades containing xanthosine/xanthine N-methyltransferases. a. 100 ML search replicate consensus tree of members of the clade containing *Coffea* XMT/MXMT/DXMT. b. 100 ML search replicate consensus tree of members of the clade containing *Camellia* caffeine/theobromine synthases; with *Theobroma cacao* theobromine synthase as an outgroup. Branch labels indicate the number of search replicates out of 100 containing that split.

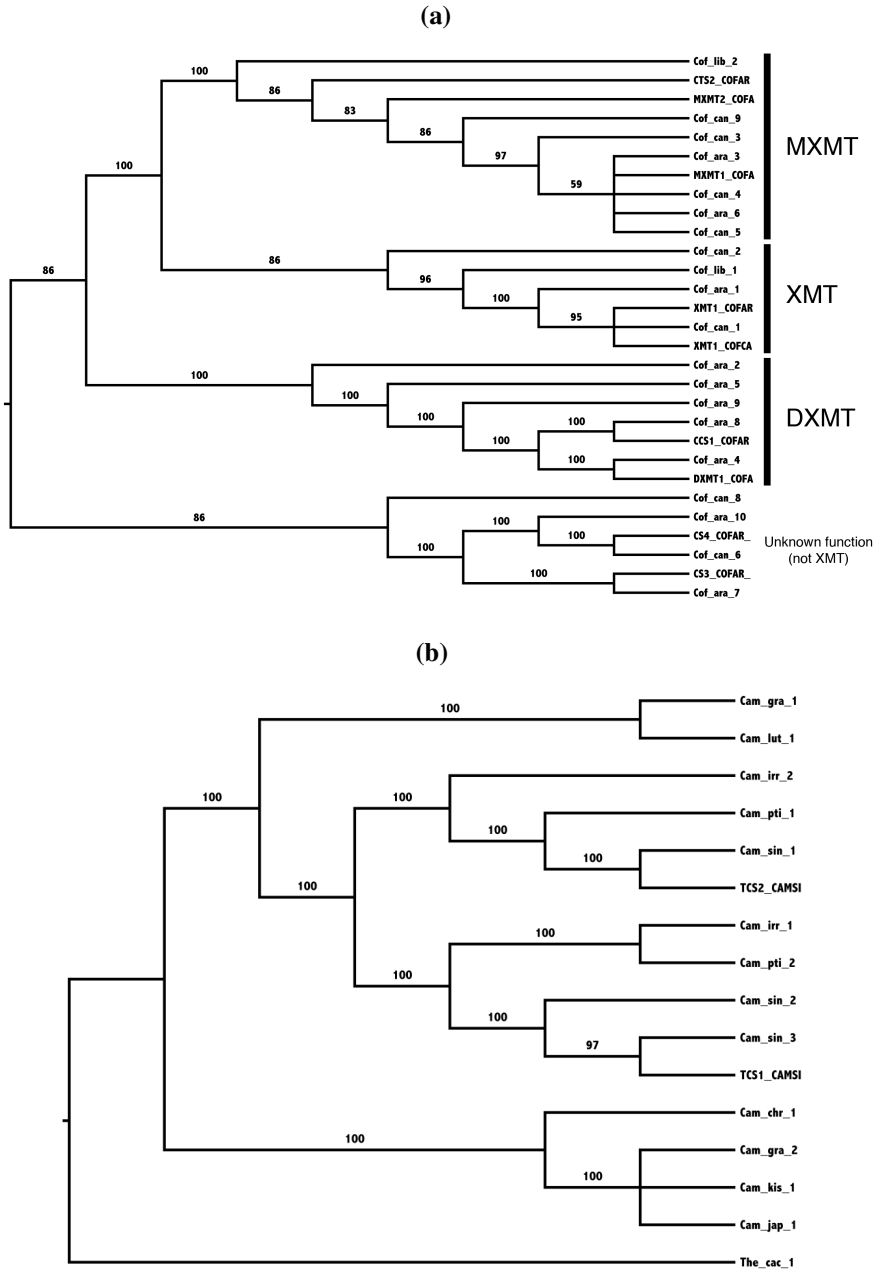


Table 4.2: PAML site model likelihood ratio tests (LRT) on xanthosine/xanthine N-methyltransferases from coffee and tea. In *Coffea*, LRTs of PAML site models found strong support for positive selection at individual sites. All three of the model tests, including the strict M1a/M2a test, found support for some sites with $\omega > 1$, all with p-values less than 0.0001. In tea, the M7/M8 model LRT found support for positive selection, but the more strict M1a/M2a LRT did not.

Coffea XMT PAML site model testing

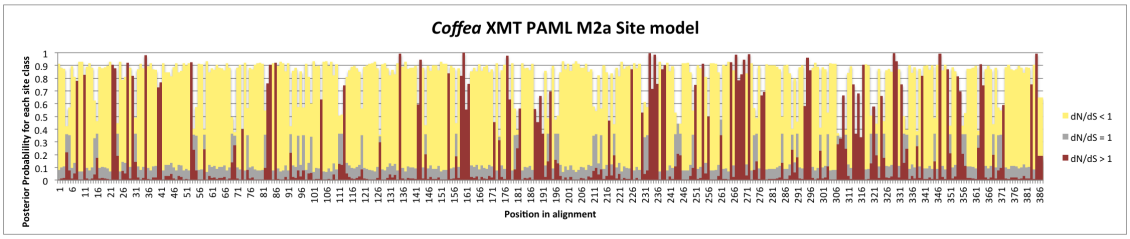
Model	lnL	LRT p-value	$\omega < 1$	$\omega = 1$	$\omega > 1$
M1a	-4211.9	<0.0001	0.01459	1	-
M2a	-4181.7		0.020133	1	3.073
M7	-4212.1	<0.0001	-	-	-
M8	-4181.8		-	-	3.07737

Camellia XMT PAML site model testing

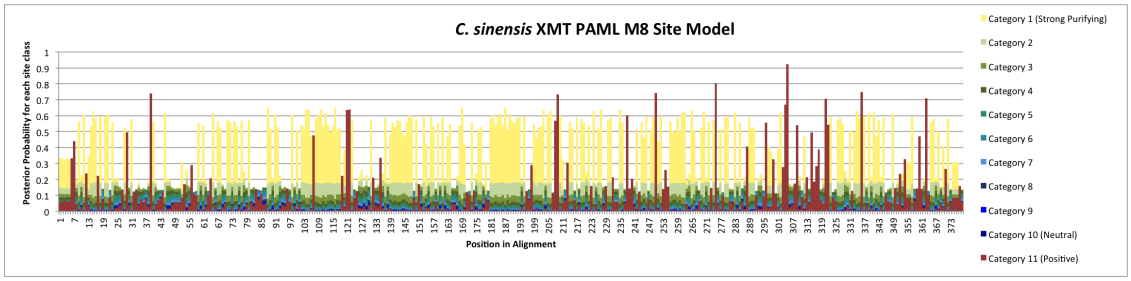
Model	lnL	LRT p-value	$\omega < 1$	$\omega = 1$	$\omega > 1$
M1a	-3953.94	0.3164	0.0131	1	-
M2a	-3952.79		0.0151	1	2.017
M7	-3956.48	0.0051	-	-	-
M8	-3951.22		-	-	1.954

Figure 4.5: PAML site model d_N/d_S values across codons in *Coffea* and *Camellia*. The strict M2a codon site model found support for positive selection at a number of sites in *Coffea XMT* coding regions (a.). There was less support for positive selection, as only the M8 model was significantly better than the neutral-only M7 model in *Camellia sinensis XMT* genes (b.). The M7 and M8 models approximate a beta distribution from 0 to 1 by binning sites into one of 10 categories, while the M8 model also includes a separate category for positive selection (sites with an $\omega > 1$). PAML calculates a Bayesian posterior probability of the site belonging to each category.

(a)



(b)



The program PAML was used to perform computational analyses measuring quantitative measures of the forces of natural selection on individual codon sites in genes encoding *Coffea* or *C. sinensis* caffeine biosynthetic pathway N-methyltransferases by calculating the ratio of the rate of non-synonymous mutations to the rate of synonymous mutations (d_N/d_S values, or ω)(Table 4.2). A d_N/d_S of one indicates equal rates of non-synonymous and synonymous mutations, or neutral selection. A d_N/d_S less than one indicates purifying or negative selection. Sites with a d_N/d_S greater than one can indicate positive or adaptive selection.

In *Coffea*, PAML found strong significant support for positive selection in specific sites with both

the strict M1a/M2a models likelihood ratio test (LRT) and the M7/M8 models LRT (Table 4.2, Figure 4.5a). The M1a model has parameters for two categories, sites with $\omega < 1$ (purifying selection), and sites with $\omega = 1$ (neutral selection). The M2a model includes both of these categories, but also includes a category for sites with $\omega > 1$ (positive or adaptive selection). The M7 and M8 models bin sites according to an approximation of a beta distribution from 0 to 1. The M8 model also includes a category for sites with $\omega > 1$. In tea, the M0/M3 LRT (which tests some selection vs. no selection) was significant. In tests for positive selection, only the M8 model was significantly better than the M7 model, the M1a/M2a LRT was not significant, suggesting possible adaptive selection at lower levels than in *Coffea*. In both coffee and tea, sites with detected positive selection varied across the length of the alignment (Figure 4.5).

4.5 Discussion

4.5.1 Polyphyletic origins of N-methyltransferases in alkaloid metabolism

The molecular evolution of plant enzymes involved in the methylation of nitrogen groups of small molecule metabolites appears to have occurred independently at least three times. The N-methylation of putrescine by PMT appears to be derived from a polyamine biosynthetic enzyme SPDS with aminopropyltransferase activity modified to perform N-methyltranerase reactions [134, 152, 131, 142, 136, 82, 77]. The XMT activity appears to be derived from the SABATH carboxyl-O-methyltransferase superfamily to carry out N-methylation reactions in purine alkaloid

biosynthesis. Lastly, CNMT and TNMT appear to be derived from yet a different category of enzymes, namely cyclopropane-fatty-acyl-phospholipid synthases, among others. SAM-dependent N-methylation appears to be relatively simple to acquire, often only requiring changes in the positioning of SAM in the active site [77]. In summary, the molecular evolution of N-methylation reactions in plant secondary metabolism independently arose at least three different times within the Plantae. This is a good example of convergent evolution of different ancestral enzymes with different enzymatic specficities independently evolving the capacity to use SAM as a substrate to methylate primary amines on small molecules. In contrast, O-methylation seems to have taken a different evolutionary trajectory in which all plant O-methyltransferases were derived from a common O-methyltransferase ancestor.

4.5.2 Molecular evolution of caffeine biosynthesis

It was previously hypothesized that caffeine biosynthetic enzymes arose independently in coffee and tea plants; however this hypothesis was not previously rigorously tested using statistically oriented phylogenetic methods. ML model-oriented phylogenetic methods also indicated that purine alkaloid N-MTs did not show orthologous groupings. Moreover, when orthologous grouping was constrained in the ML phylogenetic analyses, it resulted in significantly poorer fit of the model to the data (Table 4.1). Thus, the present studies provide a rigorous test that rejected the hypothesis of *Coffea* and *Camellia* caffeine biosynthetic genes showing orthologous relationships and thus close common ancestry. These phylogenetic analyses provide statistical support for the independent evolution of those N-methyltransferases involved in the caffeine biosynthetic pathway. If these

proteins arose once, we would expect to see orthologs from tea and coffee cluster together (such as TCS1 and CaDXMT), rather than group by paralogs within species, as we see in our phylogenies. Enzymes performing similar functions in other plants that produce caffeine, such as the kola tree [186], yerba mate [187], and guarana [188], may also have also evolved independently; though caffeine synthases have yet to be cloned from any of them. While the N-methyltransferases in caffeine biosynthesis in *Coffea* and *C. sinensis* arose independently of one another, they all were derived from the plant O- and carboxyl-methyltransferase superfamily, and thus are monophyletic.

Of particular interest was the apparent order of enzymes encoding different steps in the caffeine biosynthetic pathway in *Coffea*. Dual-function caffeine synthases, which perform the final steps in the pathway, appear most basal in the phylogeny, while enzymes performing earlier steps, specifically xanthosine N-methyltransferases, appear to have evolved from enzymes carrying out later steps in the pathway. This raises the obvious question of how 7-methylxanthine formed prior to the origin of extant XMT enzymes. Presumably, there were less efficient mechanisms to produce 7-methylxanthine prior to the evolution of extant XMT enzymes. XMT performs both the first methylation as well as the nucleosidase step in *Coffea*. This enzyme may have later evolved in order to consistently direct more of the initial substrates toward the further steps in the pathway in a controlled manner, rather than rely on the likely low levels of activity by other more promiscuous nucleosidases. The continued diversification and specialization of XMT enzymes performing yet other catalytic activities may be responsible for the rather strong signals of positive selection in *Coffea XMT* genes.

While *C. sinensis* appears to lack the enzyme specialization seen in *Coffea*, caffeine biosynthetic

genes in tea were also identified as having some codons under positive selection. An ortholog of XMT has not yet been characterized in tea, and would be particularly informative in providing a clearer picture of whether XMTs evolved from dual function caffeine synthases in both *Coffea* and *Camellia*.

Chapter 5

De novo assembly of the *Toxicodendron radicans* (poison ivy) root and leaf transcriptomes

Alexandra J Weisberg, Elise Benhase, Gunjune Kim, James Westwood, and John G Jelesko.

5.1 Abstract

Poison ivy (*Toxicodendron radicans*) is a native ”neo-invasive” plant common to much of North America that causes contact dermatitis in humans. Poison ivy and other members of the Anacardiaceae family (such as poison oak and poison sumac) produce urushiol, a catechol with a 15

or 17-carbon side chain that is responsible for the characteristic allergic skin rash. The alkyl side chain may have several degrees of unsaturation, with more double bonds equating with a stronger allergic reaction. There is not much currently known about the genes and enzymes used by poison ivy to synthesize urushiol. To develop resources suitable for molecular-oriented investigations of urushiol biosynthesis, I sequenced the transcriptome of several tissues from *T. radicans*. Poison ivy drupes were collected and axenic seedlings were used as a source of leaves and roots. RNA was extracted and subject to high-throughput Illumina sequencing. The poison ivy leaf and root transcriptomes were then assembled and annotated using a published pipeline for *de novo* assembly. Several *T. radicans* transcripts encoding type III polyketide synthase homologs were identified as candidate genes responsible for the proposed first step in urushiol biosynthesis.

5.2 Introduction

Poison ivy (*Toxicodendron radicans* subsp. *radicans*) is a fast-growing weed common to much of North America and east Asia [189, 190, 191]. Skin contact with any part of the poison ivy plant (as well as other members of the Anacardiaceae, including poison oak, poison sumac, cashew, and mango) results in painful allergic contact dermatitis in most of the population [192, 193]. The rash symptoms manifest as red, blistering, inflamed skin, that lasts anywhere from several days to several weeks, and in severe cases requires hospitalization. In poison ivy/oak/sumac, the natural product responsible for this allergic contact dermatitis is called urushiol [194, 195, 196], based upon its chemical similarity to nearly identical compounds produced by the Japanese lacquer tree

Toxicodendron verniciflua [197]. Annually, 10 to 50 million Americans will suffer an allergic reaction from contact with urushiol [198], and contact with poison ivy, oak, or sumac accounted for 8% of incident reports of USDA Forest Service employees in 2005 and 2006 [199]. Two studies have also linked poison ivy with global climate change. Controlled-forest experiments have demonstrated that increasing levels of carbon dioxide in the atmosphere result in poison ivy plants that grow faster, produce more biomass, and produce a larger amount of the more allergenic forms of urushiol [200, 201]. Based on these experiments, it is estimated that poison ivy plants have doubled their growth rate since the 1960s, and grow faster than other woodland lianas [201].

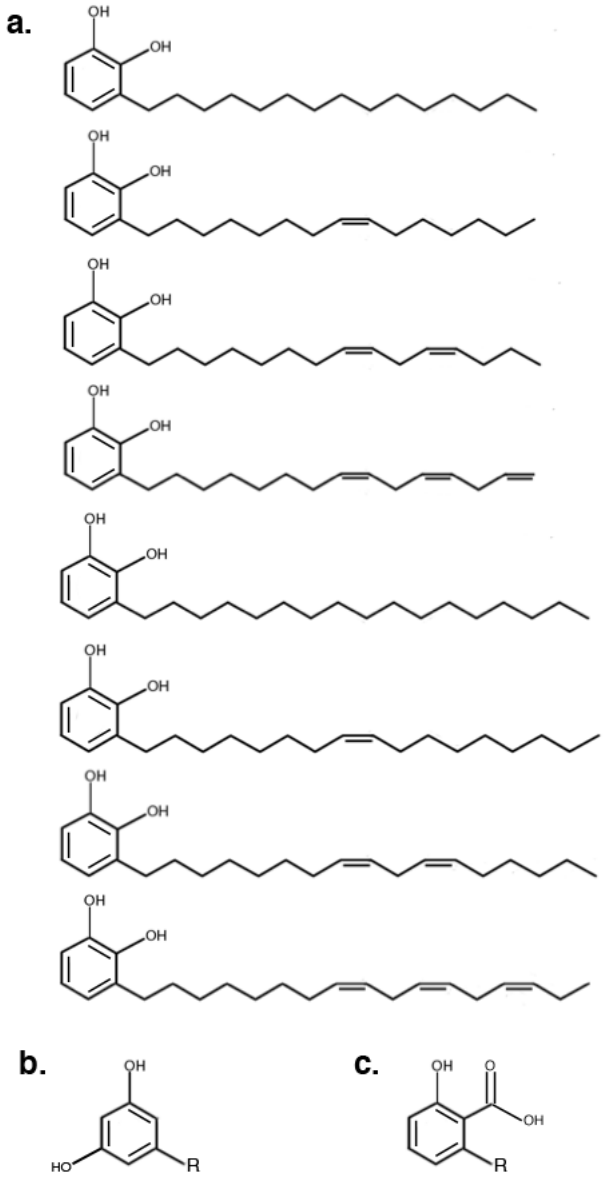
Urushiol is comprised of a catechol ring with a 15 or 17-carbon alkyl side chain [202, 203, 196] (Figure 5.1a). Poison ivy primarily produces the C15 form, while poison oak primarily produces the C17 form. Up to three of the carbon-carbon bonds in the alkyl side chain may be unsaturated [204, 205, 206], and urushiols with greater degrees of unsaturation are more allergenic [207]. Urushiol is very chemically reactive due to the ortho hydroxyl groups [208], and this reactivity has been used for thousands of years in the making of Asian (Chinese/Korean/Japanese) lacquerware, as it readily polymerizes into a non-allergenic hard clear shellac on wooden objects. Urushiol biosynthesis appears to be constitutive, as it is found in all growing plant tissues [195, 205, 209, 210, 211, 206]. Current methods to detect urushiol in plants use gas chromatography-mass spectrometry (GC-MS) as a means of identifying the various alkyl-phenols in cell extracts [205, 209, 210, 211, 206, 212]. An alternative means using fluorescence detection has been developed that does not require time-consuming extraction steps for the qualitative detection of urushiols [213]. It is not currently known what the function of urushiol is in the plant. With

the notable exception of humans, most mammals do not react to urushiol exposure as humans do. Ruffed grouse and two species of squirrel eat poison ivy drupes with no apparent ill effect [214]. Goats can also be used as a form of poison ivy weed control, as they too readily eat the plant with no ill effect, and urushiol is not found in their milk after ingesting the plant [215]. Treating poison ivy drupes with a strong acid increases the germination rate [216], suggesting that partial digestion in the gut of birds or other mammals may promote seed dispersal and germination.

Very little is currently known about the biosynthetic pathway for urushiol in poison ivy and other related species. Dewick (1997) proposed that urushiol biosynthesis is derived from intermediates involved in fatty-acid metabolism. The first step in the pathway is postulated to be a type III a polyketide synthase using malonyl-CoA and a fatty acid-CoA substrate [217, 218] to generate an alkyl-tetraketide intermediate that eventually gives rise to the catechol moiety of urushiol. Alkylresorcinol synthases and other type III polyketide synthases utilize long-chain fatty acids as substrates *in vitro* [219, 218] to produce structurally similar chemicals in a number of plants [220, 221, 222]. Dewick's proposed pathway for urushiol biosynthesis was presented in a very general way, and did not explain important details such as the relative order of certain metabolic transformations and what enzymes are likely to carry out those proposed reactions. Nevertheless, the initial type III polyketide synthase tetraketide formation is well supported in the alkylresorcinol field and is very likely to be the case in urushiol biosynthesis.

Understanding how poison ivy produces urushiol is a necessary first step for the development of novel weed control measures for this noxious neo-invasive plant. To the best of my knowledge, there is no published genomic or transcriptomic data for poison ivy. Sequencing the leaf and root

Figure 5.1: Alkylphenols produced in plants. a. Urushiol is comprised of a catechol ring with a 15 or 17-carbon alkyl side chain. The side chain can have up to 3 unsaturated bonds. Poison ivy plants contain a mix of these urushiols in varying amounts, though primarily in the 15-carbon forms. The more unsaturated forms of urushiol are more allergenic. b. alkylresorcinol c. anacardic acid. Side chains (R) for alkylresorcinol and anacardic acid are similar to those for urushiol, comprised of an alkyl chain with possibly several double bonds.



poly-(A) transcriptome of poison ivy provides the coding sequence without introns (and thus predicted protein sequence) of mRNA expressed in those tissues. These predicted protein sequences, and their relative mRNA expression within tissues, are an excellent means of generating hypotheses focused on which proteins may be involved in urushiol biosynthesis. Towards this end, the sequenced leaf and root transcriptome of *T. radicans* was used to identify three candidate transcripts possibly encoding type III polyketide synthases.

5.3 Methods

5.3.1 Axenic cultured seedlings and plant tissue

T. radicans subsp. *radicans* drupes from the RoaCo-1 liana were germinated and grown under sterile culture conditions as previously described [216]. In short, drupes were mechanically scarified using a 3lb rock tumbler (Chicago Electric Power Tools, Chicago) for 4 days to remove exocarp and much of the mesocarp tissues. Seeds were further chemically scarified by treatment with 13N sulfuric acid (SA) (Fisher Scientific, Waltham, MA). Scarified seeds were washed three times with 25mL of sterile deionized distilled water (ddH_2O) in a laminar flow sterile cabinet (Labconco, Kansas City, MO) to remove residual sulfuric acid (SA). Seeds were then treated with a solution of 50% liquid bleach, 3% sodium hypochlorite final, (Clorox Co., Oakland, CA.), 50% ddH_2O, 0.5% Tween 20 using the same protocol as the SA treatment, and then washed three times with 25 mL sterile ddH_2O. Floating seeds were found to be non-viable and were discarded. Seeds that did not

float were plated 20 seeds per 150 mm x 15 mm plastic Petri plate on sterile 0.5x MS Basal Salts media [223] containing 0.3% w/v Phytagel (Sigma-Aldrich Co., St. Louis, MO) pH 5.7. Sterile forceps were changed after handling every second seed. Plates were stored in the dark at room temperature for 7 days, and then placed in a Percival CU-36l4 growth chamber (Percival Scientific, Perry, IA) at 28°C and 16 h light. After 4 days, closed plates were examined under a Leica Zoom 2000 illuminated stereo microscope (20x magnification). Seedlings without apparent bacterial or fungal contamination were transferred to individual Magenta boxes or Phytatray II boxes (Sigma-Aldrich Co., St. Louis) containing 0.5x MS Basal Salts media with 0.3% w/v Phytagel and returned to the growth chamber.

5.3.2 RNA purification and sequencing

Total RNA was extracted from two replicates each of true leaves or roots harvested from sterile-cultured plants using a phenol/chloroform/SDS extraction protocol [224]. The root and leaf material were harvested from two separate biological replicates, with each tissue type combining material from $>$10 plants. A polysaccharide precipitation step using 0.2 M KOAc was also included to remove polysaccharides prior to the precipitation of total nucleic acids. RNA was selectively purified using a 2 M $LiCl_2$ precipitation step, followed by a subsequent purification using a Qiagen RNeasy Mini Kit (Qiagen, Venlo, Netherlands) according to the manufacturers instructions until an $A_{260/280}$ of 1.8 was reached.

RNA samples were submitted to the Virginia Bioinformatics Institute, Blacksburg, VA for analysis

using a BioAnalyzer, library preparation, and sequencing on an Illumina HiSeq 2000. Each of the four samples was selected for poly-(A) sequences and run on individual lanes of the HiSeq.

5.3.3 *De novo* assembly of transcripts

The quality of the sequenced paired reads produced in each sample was observed using FastQC 0.10.1 [225] before and after trimming using Trimmomatic 0.3 [226] (Table 5.1). Paired reads were trimmed to remove Illumina sequencing adaptors as well as portions of poor quality reads using a sliding window analysis (Trimmomatic settings: "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36"). Reads from all samples were combined prior to assembly to allow for differential expression analyses. The combined paired reads were *de novo* assembled into a reference transcriptome using Trinity RNA Seq release 2013/08/14 [227]. Reads were then aligned to transcripts using Bowtie 2.1.0 [228] to assess overall assembly quality and analyze individual transcripts.

Assembled transcripts were annotated using the Trinotate package of Trinity as well as several other programs, including HMMER [229] to predict PFAM domains [230], SignalP [231], tmHMM [232], and RNAmmer [233]. The Trinity package TransDecoder was used to predict open reading frames (ORFs) and predicted translated protein sequences from assembled transcripts. Each transcript was used as a query in a BLASTX search against the NCBI nr database; predicted protein sequences were also used as queries in BLASTP searches against the NCBI nr database. The program Blast2GO [234] was used to map GO terms to transcripts. All annotations were then

combined together using Trinotate and visualized with TrinotateWeb.

To estimate transcript abundance and compare expression across all samples, assembled transcripts were TMM (trimmed mean of M) normalized using Trinity, and FPKM values (fragments per kilobase of transcript per million mapped reads) were calculated using RSEM [235]. Differentially expressed transcripts between the two leaf samples and the two roots samples were identified using the Bioconductor package edgeR [236].

5.3.4 Comparison with mango transcriptome

The assembled transcriptome for mango (Mangifera indica) was downloaded from NCBI (SRA: SRX331487). Predicted peptide sequences were generated from transcript nucleotide sequences using the Trinity package Transdecoder. Predicted protein sequences for mango and poison ivy were combined and compared in a BLAST all vs. all analysis using NCBI Blast+ 2.2.28. Transcripts were then clustered using OrthoMCL [237].

5.3.5 Molecular cloning of type III polyketide synthase-like transcripts

Transcripts encoding polyketide synthases possibly involved in the first step of urushiol biosynthesis were identified using sequence similarity to known type III polyketide synthases. Four known type III polyketide synthases (two alkylresorcinol synthases from *Oryza sativa* refs: AAP52217.1 and AAP52307.1, olivetol synthase from *Cannabis sativa* ref: BAG14339.1, and alkylresorcinol synthase from *Physcomitrella patens* ref:ABU87504.1) were each used as queries in GGSEARCH

36.3.6 [238] global-global similarity searches (e-value cutoff of 10) against a database containing predicted proteins from all sequenced transcripts. Hits from all four searches were combined along with the four query sequences, aligned using MAFFT [122], and assembled into a neighbor-joining tree using PAUP* [125] (Figure 5.5a). Based on high similarity to the query sequences, annotation information, and relatively high expression in *T. radicans* roots and leaves, three transcripts from different Trinity components were selected for molecular cloning (Figure 5.5b) as cDNAs generated from RNA samples used in the RNA-seq analyses. RNA from leaves sample A was chosen, as expression of all 3 transcripts was relatively high in both leaf samples (Table 5.1). RNA was treated with DNase I using an Ambion DNase kit (Life Technologies, Carlsbad, CA) and cDNA synthesis was performed using an Omniscript RT kit (Qiagen, Venlo, Limburg) according to package instructions and then quantified using a NanoDrop 3300 fluorospectrometer (Thermo Scientific, Wilmington, DE).

Primers were designed for the predicted protein coding region of each of the three transcripts, including stop codons. Restriction sites *Nde*I and *Xho*I were also included for cloning into the pET15b vector (Novagen) for recombinant expression in *E. coli* with an N-terminus 6xHIS tag. Four extra nucleotides were added to the 5' end of each primer for efficient digest of restriction sites. Primers were as follows: ACTGCATATGGTTAGCGTGAATCAAATTCGCAAGG (forward) and AATCCTCGAGCTAAGCACTGTTAACACTGTGAAGGACC (reverse) for comp74373_c0_seq1, ACTGCATATGGTGAGCGTTGATGAAGTTCGC (forward) and AATCCTCGAGTTATGCAGTAGCAACGCTGTGAAGG (reverse) for comp87518_c0_seq4, and ACTGCATATGGCATCCGTATCCGTTGAAGAGATTAG

(forward), AATCCTCGAGTCAGTGGGCGGCTGC (reverse 1), and TATCCTCGAGTTTAATGATGATTTTATGGGATTCAGTGG (reverse 2) for comp77034_c0_seq1. Restriction cut sites for *Nde*I and *Xho*I are underlined in primer sequences.

Initial PCR for each of the three chosen transcripts was performed using GoTaq®Green Master Mix (Promega, Madison, WI). PCR reactions in 20μL contained 10 ng cDNA and primers at a concentration of 0.5 μM each. Reactions were performed in an Eppendorf Mastercycler pro thermocycler (Eppendorf, Hamburg, Germany) as follows: 95°C for 10 min; 35 cycles of 95°C for 30 sec, 60°C for 45 sec, 72°C for 1 min 30 sec; followed by a final extension at 72°C for 5 min and then held at 4°C. Samples were run on a 1% agarose TAE gel along with 1 kb DNA ladder (New England Biolabs, Ipswich, MA).

Long and accurate PCR for target component transcripts was performed using a TaKaRa LA PCR kit version 2.1 (Takara Bio, Shiga, Japan). PCR reactions in 50 μL contained 400 ng cDNA and primer at a concentration of 0.25 μM each. Thermocycler settings were as follows: 94°C 1 min, 35 cycles of 94°C 30 sec, 60°C 30 sec, 72°C 1 min 30 sec, a final extension for 5 min at 72°C; and then held at 4°C. PCR products were visualized on 1% agarose gels (Figure 5.8). PCR products were purified by ethanol precipitation [224]. DNA pellets were resuspended in ddH_2O and quantified using a NanoDrop 1000.

LA-PCR products were first subcloned into the pGEM-T plasmid (Promega, Madison, WI) and transformed into *E. coli* strain DH5α [239]. Recombinant plasmid was purified using a Qiagen Plasmid Mini Prep Kit (Qiagen, Venlo, Limburg) and digested with *Nde*I and *Xho*I (New England Biolabs, Ipswich, MA). Digests were run on a 1% low-melt agarose gel, and bands corresponding

to the transcripts were cut out and purified using a Qiaquick gel purification kit (Qiagen, Venlo, Limburg). Digested PCR fragment and pET15b vector were then ligated using T4 DNA ligase overnight at 4°C and transformed into DH5αas well as BL21(DE3) [240] along with the pRARE plasmid (Novagen, EMD Millipore, Billerica, MA).

5.3.6 Protein expression and purification

For each cloned cDNA of interest, 5 colonies from each transformation plate were inoculated in 50 mL LB cultures with the appropriate antibiotics and grown overnight in a shaker at 37°C at 250 rpm. A 12 mL sample of each overnight culture was used to inoculate three 1L ZYM-5052 autoinduction media [146] cultures. Cultures were grown for 7 hours at 37°C at 250 rpm, then moved to 18°C overnight for a total of 24 hours. Cultures were centrifuged in 1L flasks at 3500 rpm for 20min at 4°C in a Sorvall RC 6+ centrifuge (Thermo Scientific, Waltham, MA) and cell pellets were stored at -80°C. Pellets were thawed and then resuspended in 3x volume Buffer A (50 mM sodium phosphate pH 7.4, 300 mM NaCl, 10% w/v glycerol). Lysozyme was added to a concentration of 25 µg/mL and incubated at 4°C for 1 hour with a magnetic stir bar. Cells were then lysed using a sonicator at 50% amplitude for 5 seconds at 15 second intervals, for a total of 5 min of sonication. Lysed cells were then centrifuged in a tabletop centrifuge at 14,000 rpm for 20 min at 4°C. Soluble lysate was collected and flowed over a column containing HisPur Ni-NTA resin (Promega, Fitchburg, WI) using a BioRad Econo EP1 peristaltic pump connected to a BioRad EP1 UV light meter (BioRad, Hercules, CA). The column was washed with Buffer A, and non-specifically-binding proteins were eluted with a mixture of 90% Buffer A, 10% Buffer

B (50 mM sodium phosphate pH 7.4, 300 mM NaCl, 10% v/v glycerol, 250mM imidazole). Recombinant protein was eluted using Buffer B, and 2 mL samples were collected corresponding to significant peaks in the UV light meter. Samples were run on a 9% SDS-PAGE gel (4% stacking gel). Elution fractions containing recombinant protein of the expected size were concentrated using a VivaSpin 6 mL spin concentrator (Sartorius AG, Göttingen, Germany), and then buffer exchanged into 100 mM potassium phosphate pH 7 using a PD10 column (GE Healthcare Life Sciences, Little Chalfont, United Kingdom). Glycerol was added to purified recombinant protein at a 10% concentration and frozen in aliquots at -80°C.

5.3.7 *Agrobacterium*-mediated transformation of *T. radicans*

Sterile-cultured poison ivy plants were vacuum-infiltrated with *Agrobacterium tumefaciens* according to a modified protocol for agroinfiltration of tomato [241]. *A. tumefaciens* strain GV3101 containing pJGJ204 [242], which consists of pSLJ7292 with a gene fusion of *Arabidopsis thaliana* RUBISCO small subunit *1B* gene fused in frame to firefly luciferase (*LUC*). *A. tumefaciens* GV3101 containing pJGJ204 or pSLJ7292 (as negative control) was grown in 5 mL LB containing 50 µg/mL gentamycin and 2 µg/mL tetracycline overnight at 28°C 250rpm. A 50 mL culture of LB, 10 mM MES, 20 µM acetosyringone, and the same antibiotics was inoculated with the 5 mL culture and grown overnight at 28°C at 250 rpm. Cell pellet was resuspended in MMA media (10 mM MES, 10 mM MgCl2, 200 µM acetosyringone) to an OD600 of 2.0 and allowed to sit at room temperature for 3 hours. Whole poison ivy plants at 20 days or harvested leaflets from the first emerged true leaves and cotyledons from plants at 35 days were submerged in the suspension and subjected

to four rounds of vacuuming, then grown in Promix potting soil (Premier Tech Horticulture, Quebec, Canada) or plated on 0.5x MS phytagel plates, respectively, and placed in a growth chamber (28°C and 16 h light). Cut leaflets and cotyledons were assayed for luciferase activity in three 1 hour time periods each at 24, 48, and 72 hours using an intensified charge-coupled device video camera (model C2400 47; Hamamatsu Photonics, Hamamatsu City, Japan), an Image Intensifier Controller (model M4314; Hamamatsu Photonics), Image Processor (Argus 50; Hamamatsu Photonics), and Hamamatsu Photonics imaging chamber (model A417) mounted with a Xenon CM 120 lens (Schneider, Bad Kreuznach, Germany). Whole plants or plates were sprayed with 1 mM D-luciferin (Biosynth, Basel) 0.01% Triton X-100 and incubated for 20 min prior to imaging. *In vivo* luciferase activity was measured using the photon counting function of the Hamamatsu imaging system and superimposed over images of illuminated leaves and cotyledons taken using the integration function. Whole plants were assayed at 24 hours post infiltration. Photon counts were measured as 10 technical replicates using a sliding window over regions containing first true leaves and cotyledons agroinfiltrated with either pJGJ204 or pSLJ7292. Photon counts were also taken over two 1 hour intervals of a 96 well plate containing 30µL *A. tumefaciens* with either pJGJ204 or pSLJ7292 suspended in MMA media with 30 µL 1 mM D-luciferin 0.01% Triton X-100.

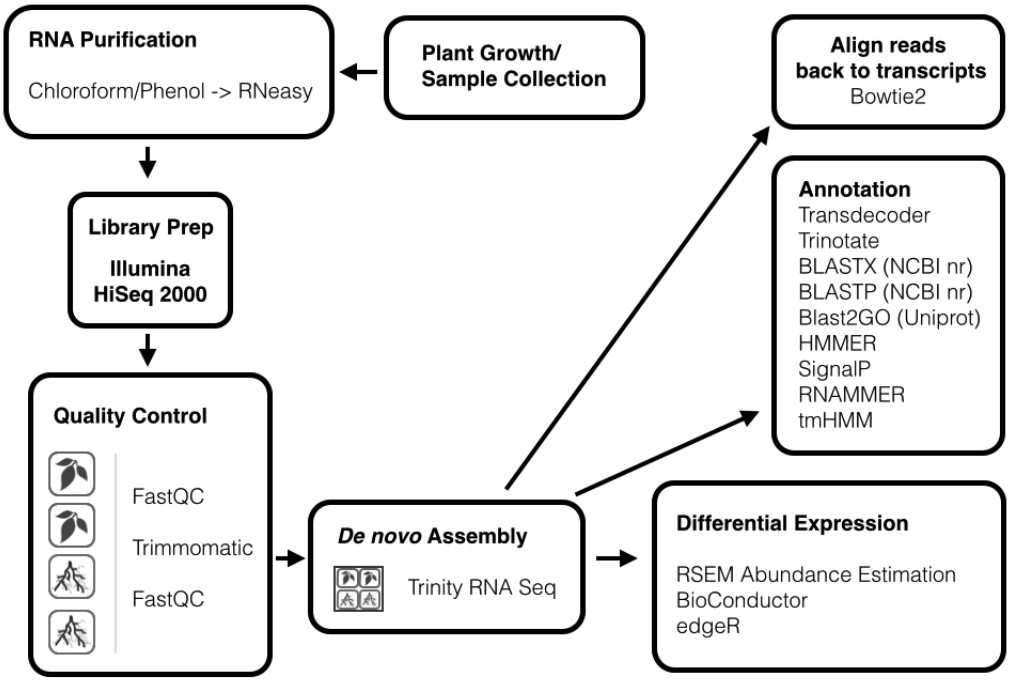
5.4 Results

5.4.1 Assembly analysis/quality

There are currently very few publically available sequences from *T. radicans*. As of March 6 2014, there are 44 nucleotide sequences and 17 protein sequences in NCBI's public databases for *T. radicans*. Most of these nucleotide sequences are gene fragments or intergenic regions used in phylogenetic population studies [189]. In order to comprehensively identify all expressed genes in *T. radicans* leaves and roots producing urushiol, we sequenced the transcriptomes of two biological replicates comprised of roots and true leaves, containing material from >10 axenic poison ivy plants. Each of the four samples was submitted for RNA-seq sequencing on an individual lane on an Illumina HiSeq 2000 in order to maximize sequencing depth of each sample. Three of the four samples produced ~100 million 100 bp paired reads (Table 5.1). However, one root sample was of poor quality, producing only 30 million paired reads, over half of which were removed during read trimming due to poor quality. Trimmed reads from all four datasets were combined (289,325,296 paired reads in total, ~100 bp each) and *de novo* assembled into a reference transcriptome using Trinity RNA seq [227] using an adapted form of a published pipeline [243] for this analysis (Figure 5.2).

In total, Trinity assembled 139,925 components (1 component is defined as all transcripts derived from a single de Bruijn graph) comprised of 211,453 transcript isoforms (Table 5.2). Of these, 32,535 unigenes were identified (Trinity components with length >200 and FPKM >0.5). The average contig length was 1,068 bp and the N50 was 2,002 bp. A large proportion of assembled

Figure 5.2: Process for *de novo* RNA-seq assembly and annotation. RNA from two replicates each of roots and leaves was sequenced with an Illumina HiSeq 2000. Trinity RNA seq was used to *de novo* assemble a reference transcriptome from the combined paired reads from all 4 *T. radicans* leaves and roots samples. Assembled transcripts were then annotated using a variety of programs.

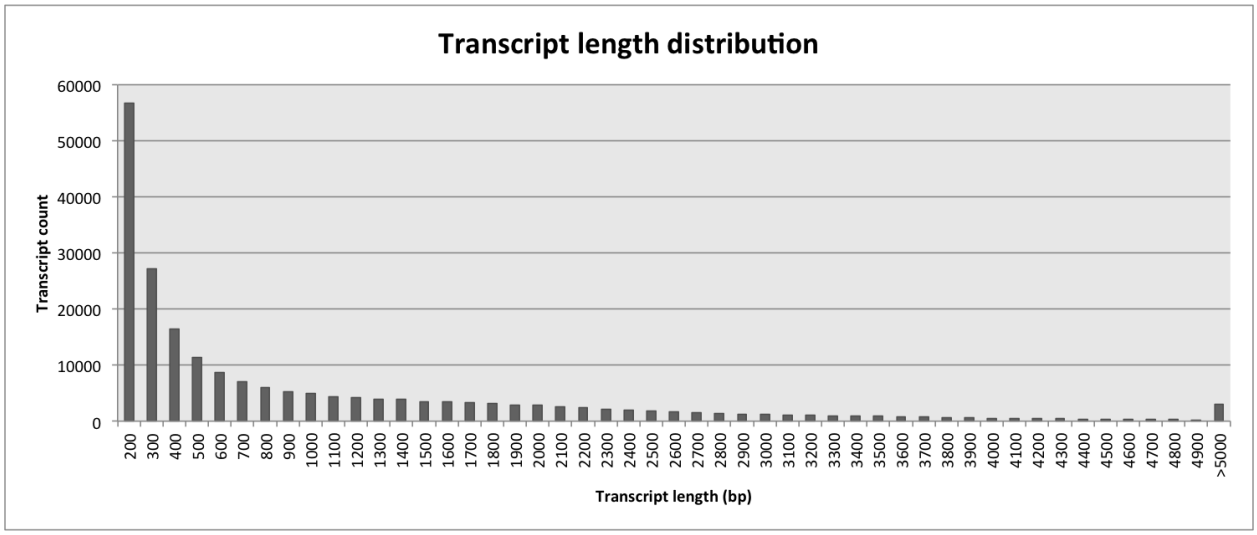


transcripts were between 200 and 400 bases long, close to the minimum length (200bp) of transcripts reported by Trinity (Figure 5.3a), while unigenes appear to be less skewed towards shorter transcripts (Figure 5.3b). Aligning paired reads to transcripts using Bowtie2 to assess overall assembly quality revealed that 76.14% of reads formed proper pairs, while 22.58% formed improper pairs. Less than 1.3% of paired reads had only one of the paired reads map to transcripts, and 2.6% of reads did not map to any transcript.

BLASTX searches of each transcript and BLASTP searches of predicted proteins from transcripts

Figure 5.3: Assembly length distributions. These graphs show the counts of the lengths of a. individual transcripts as well as b. unigenes (trinity components with length >200 and FPKM >0.5); represented by 100 bp bins. Sequences with length greater than 5000 were binned together. The minimum length transcript reported by Trinity is 200 bp.

(a)



(b)

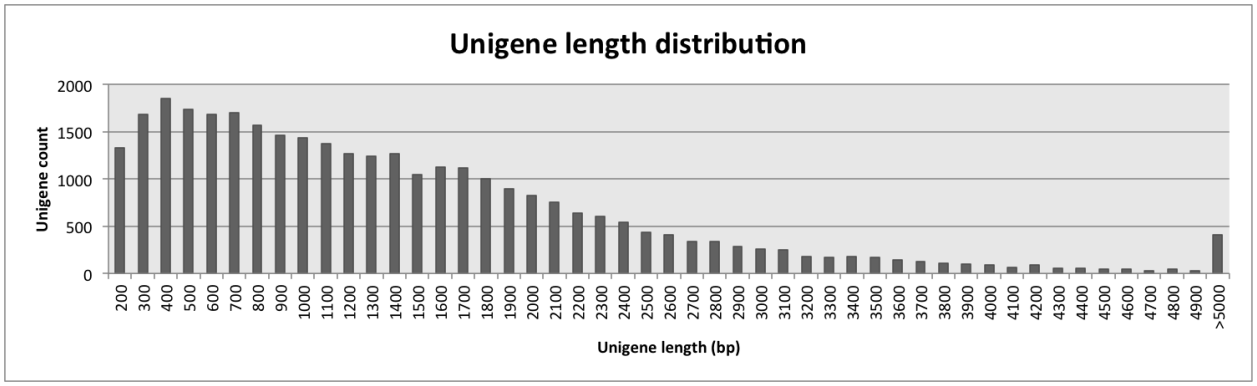


Table 5.1: Paired read quality per sample. The read quality of each of the four samples was checked using FastQC before and after quality trimming with Trimmomatic. Leaves replicates A and B as well as Roots replicate B each had ~100 million paired reads, while roots replicate A had relatively much fewer (~30 million). Most reads in Leaves A and B and Roots B samples were of high quality, with over 80% surviving trimming, while more than half of the Roots A reads were removed due to poor quality. Trimmed reads from all samples were combined (289,325,296 paired reads) to assemble a reference transcriptome.

Sample	Leaves A	Leaves B	Roots A	Roots B	Total
Total # of paired reads	97,494,163	93,384,797	30,270,702	128,496,278	349,645,940
Total # of paired reads after trimming	87,835,193	81,637,895	14,779,168	105,073,040	289,325,296
% of paired reads surviving trimming	90.09%	87.42%	48.82%	81.77%	
GC content	46%	46.5%	40%	47%	

Table 5.2: *De novo* assembly statistics. The *de novo* assembly of a reference transciptome for poison ivy was performed by Trinity RNA Seq using the combined paired reads of all four samples. Overall, 289,325,296 paired reads were assembled into 9,022,555 contigs (using a k-mer size of 25 and minimum length of 25 bases), which were resolved into 211,453 unique transcripts. Out of all Trinity components, 32,535 unigenes were identified using a minimum expression cutoff of 0.5 FPKM and a minimum length of 200bp.

	Sequences	Mean length (bp)	N50	Assembled bases
Total contigs	9,022,555	1,068	2,002	225,861,945
Total Trinity components	139,925			
Total transcripts	211,453			
Total unigenes	32,535			

were performed to annotate individual transcripts based on similarity to known protein sequences in the NCBI nr database. Of the 50,651 Trinity components with BLAST hits, 24,744 components encoded by 83,165 transcripts had a top hit to a member of the Viridiplantae (Figure 5.4a). When looking at only unigenes, 16,825 unigenes had at least one transcript with a top BLAST hit to a member of the Viridiplantae (Figure 5.4b). The organisms with the most BLAST hits from poison ivy transcript queries were *Theobroma cacao* (10,883 BLASTX hits) and *Vitis vinifera* (6,302 BLASTX hits).

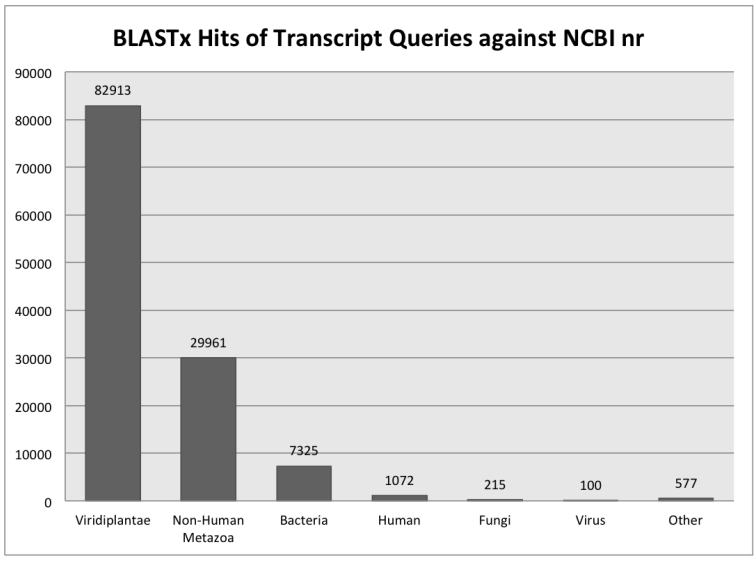
An unexpectedly large number of transcripts had top BLAST hits to Metazoa (particularly various insects) and Bacteria. This was surprising, because the poison ivy plants were grown under sterile culture conditions. The main observable contamination of sterile-cultured plants are fungi and bacteria [216], yet relatively few BLAST hits were to either bacterial or fungal species.

Comparisons were made between the relative expression levels of transcripts in the four root and leaf samples. Differential expression analysis of TMM-normalized FPKM values using edgeR found zero differentially expressed transcripts with p-value cutoff equal to 0.001 and min abs($\log_2$(a/b)) fold change equal to 2. Similarly there were no identified differentially expressed Trinity components with the same cutoffs; and manual observation of FPKM values for each sample showed a very different expression pattern for the root A sample compared with the rest of the samples. This may be due to the relatively poor quality of the root A sample. Removing this sample from the analysis resulted in 6,347 transcripts differentially expressed, however with no statistical power due to only two leaf replicates and only one reliable root sample.

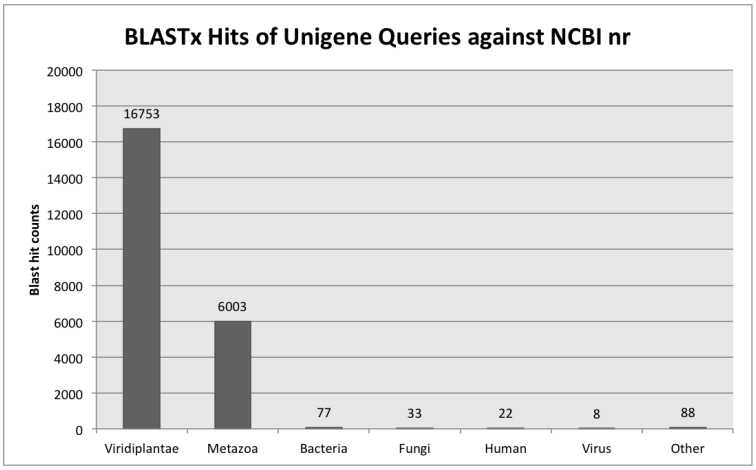
Another method for analyzing transcriptome assembly data involves clustering transcripts with

Figure 5.4: Blast searches of *T. radicans* transcripts against NCBI nr. The results of BLASTX and BLASTP searches were sorted by classification. In both cases, BLAST hits to known plant sequences were the single largest group, however over 30,000 transcripts were most similar to proteins from Metazoa, in particular various insect species. This may be due to contamination during any of plant growth, RNA extraction, or library preparation.

(a)



(b)



those from a closely-related species. Currently the only member of Anacardiaceae with publicly available transcriptome data is the mango tree (*Mangifera indica*). Transcripts from the mango leaf transcriptome [244] (13,558 transcripts) and plant-annotated transcripts from poison ivy (92,108 transcripts) were combined and clustered using OrthoMCL [237, 245]. OrthoMCL identified 13,558 groups using an inflation index of 1.5. Out of 103,122 total sequences, 87,883 formed groups, while 15,239 transcripts from either *T. radicans* or *M. indica* did not cluster with any other transcript.

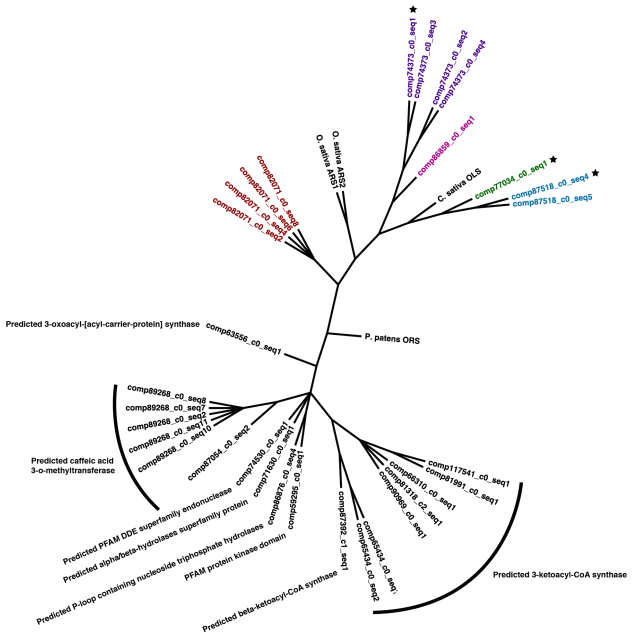
5.4.2 Cloning of type III polyketide synthases

The first proposed biosynthetic step in urushiol synthesis is believed to be similar to the first step in alkylresorcinol biosynthesis. Specifically, that a type III polyketide synthase elongates a fatty-acid-CoA starter molecule to form a tetraketide-alkyl intermediate [217]. Type III polyketide synthases such as chalcone synthase can utilize various length fatty acid-CoA substrates *in vitro* in addition to their presumed primary substrate coumaroyl-CoA [218] and appear to be fairly promiscuous *in vitro*, though it is not known if this occurs *in planta*. Alkylresorcinols, plant natural products structurally similar to urushiol (though not nearly as reactive), are found in several plant species. Alkylresorcinols are also comprised of an aromatic ring with two hydroxyl groups, but they are in the meta position (in urushiol they are in the ortho position) (Figure 5.1b).

Alkylresorcinol synthases have been identified in *Oryza sativa* [221] and *Physcomitrella patens* [222] that take a fatty acid-CoA conjugate substrate, and extend it to a tetraketide, cycilize, and

Figure 5.5: Phylogenetic tree and expression levels of poison ivy probable type III polyketide synthases. Known type III polyketide synthases that utilize varying length fatty acid-CoA substrates were used as queries in global:global similarity searches (GGSEARCH) of *T. radicans* predicted proteins. A Neighbor-Joining phylogenetic tree (a.) of the top search hits and query sequences revealed two distinct groups. Taxa that are colored are annotated as putative type III polyketide synthases/chalcone synthases. Each color represents an individual unigene. b. Relative expression levels for each of the putative type III polyketide synthases from leaves and the good quality roots sample. Taxa labeled with stars (⋆) represent those transcripts selected for molecular cloning.

(a)



(b)

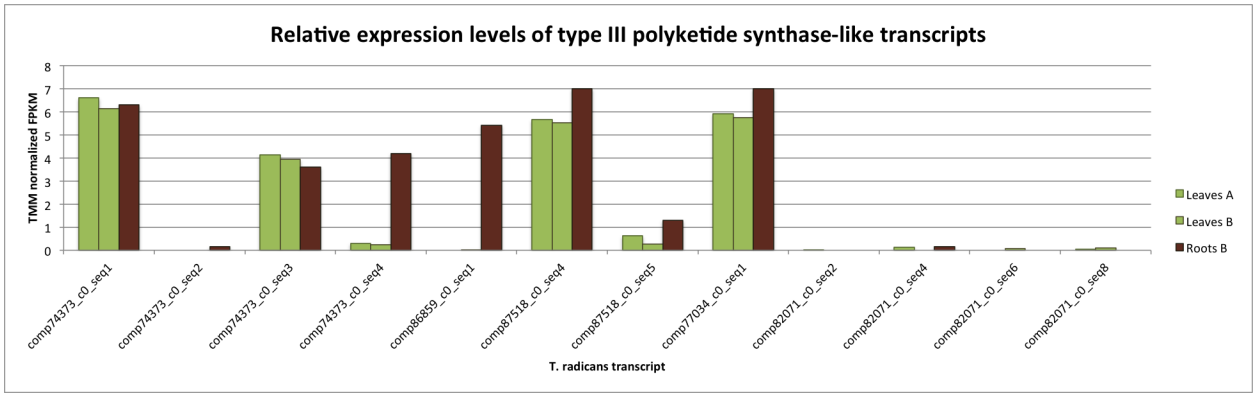
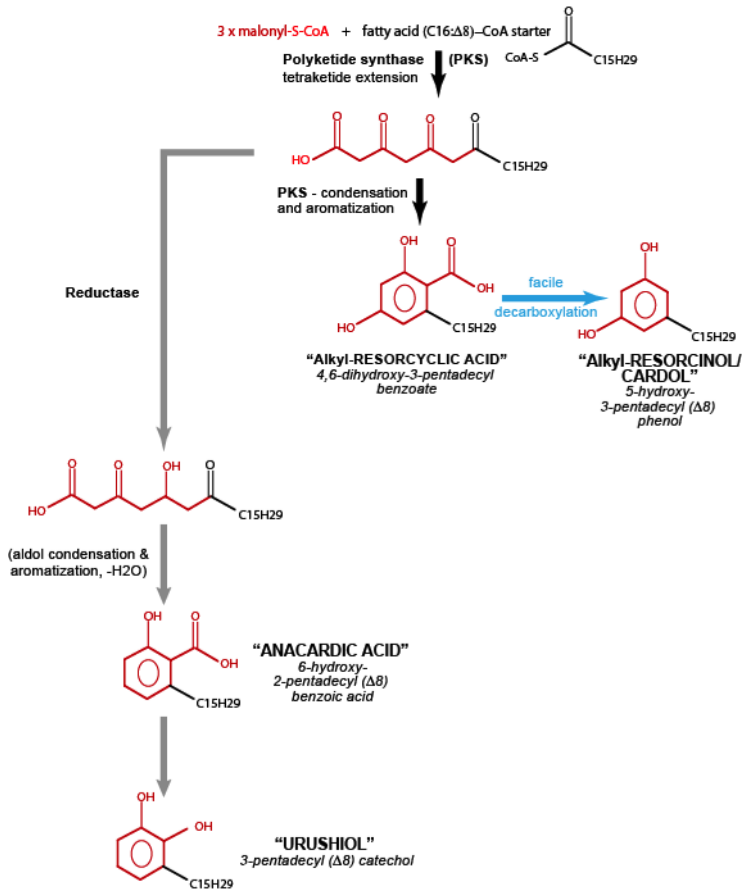


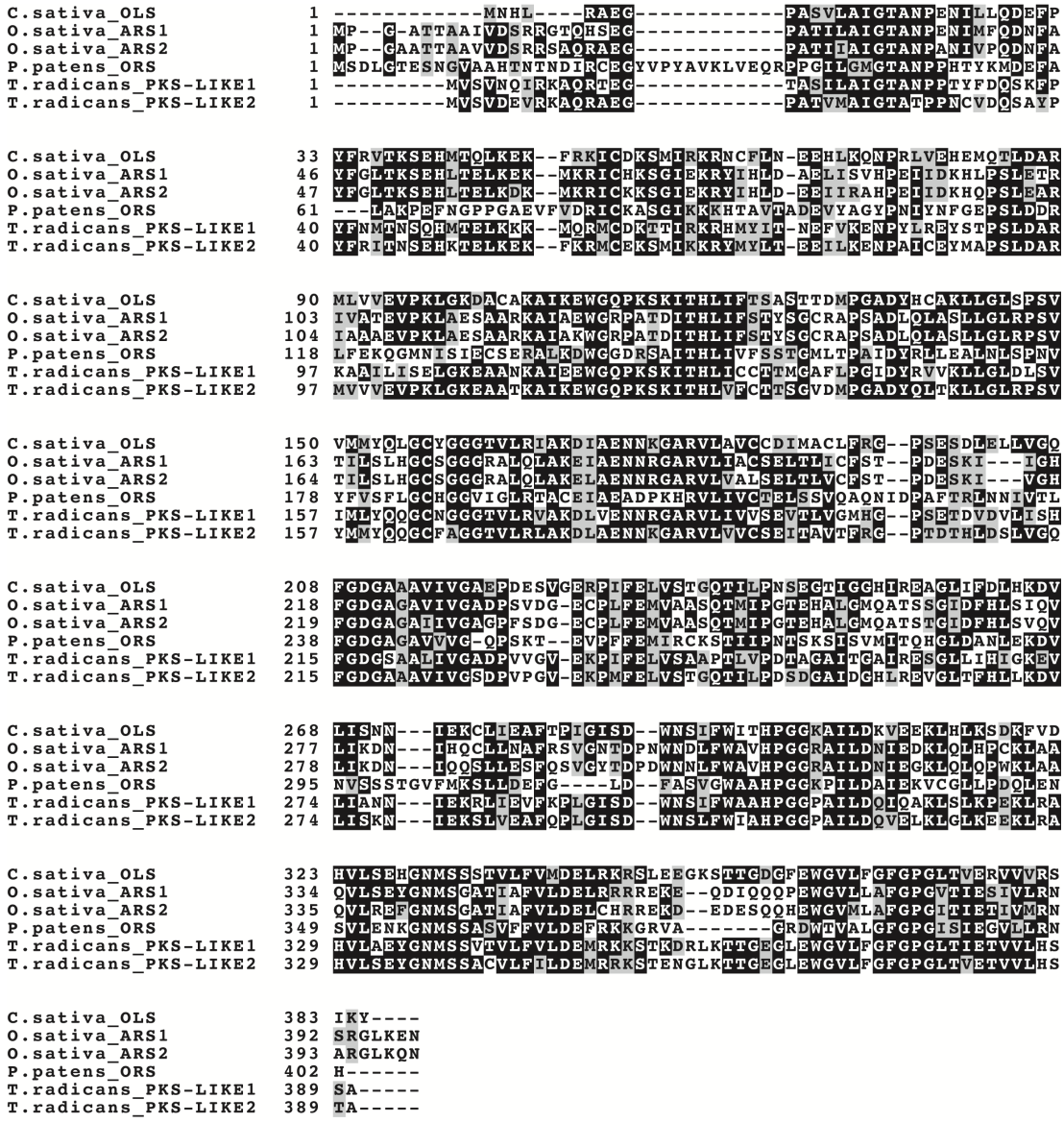
Figure 5.6: Unified urushiol and alkylresorcinol biosynthetic pathways. The first step has been previously proposed [217, 218] as a type III polyketide synthase utilizing malonyl-CoA and a long chain fatty acid-CoA substrate.



aromatize to form various alkylresorcinols. Another type III polyketide synthase, olivetol synthase, performs a similar reaction in *Cannabis sativa* [220], though a separate polyketide cyclase performs the aromatization step to produce olivetol, an intermediate in the tetrahydrocannabinol (THC) biosynthetic pathway [246]. The alkylresorcinol-producing type III PKS-like enzymes were used as a starting point to search for similar proteins in poison ivy.

Similarity searches of poison ivy predicted proteins using four type III polyketide synthases previously demonstrated to be involved in alkylresorcinol biosynthesis identified a variety of differ-

Figure 5.7: Multiple sequence alignment of PKS-like1 and PKS-like2 protein sequences with other type III polyketide synthases. PKS-like1 and PKS-like2 are amino acid sequences encoded by *PKS-like1* and *PKS-like2*, respectively. C.sativa_OLS is *Cannabis sativa* olivetol synthase (NCBI accession BAG14339.1), O.sativa_ARS1 and O.sativa_ARS2 are alkylresorcinol synthases from *Oryza sativa* (AAP52217.1 and AAP52307.1, respectively), and P.patens_ORS is 2'-oxoalkylresorcinol synthase from *Physcomitrella patens* (ABU87504.1). Amino acid similarity between *T.radicans* PKSs and other species ranged from 61.3%-84.9%.

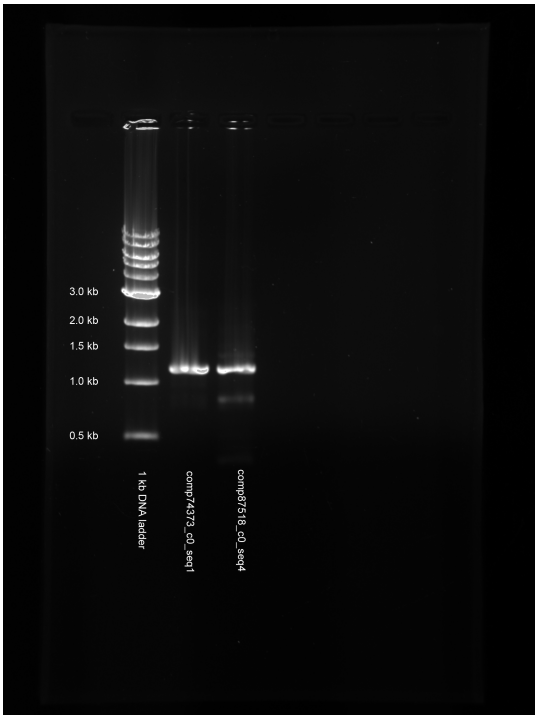


ently annotated poison ivy transcripts. The most similar were annotated as chalcone synthase-like, though other less similar transcripts were annotated as caffeic acid o-methyltransferases, ketoacyl-

CoA synthases, and alpha/beta hydrolases, among others. A Neighbor-Joining phylogenetic tree of the search sequences and the poison ivy homologs (Figure 5.5a) revealed a split, with the type III polyketide synthase/chalcone synthase-like annotated transcripts grouping closest to the search sequences. Since urushiol is found in all growing poison ivy tissues (Elise Benhase, unpublished results), the expectation was that transcripts encoding urushiol biosynthetic enzymes should accumulate to relatively high levels in both leaves and roots. Three transcripts (comp74373_c0_seq1, comp87518_c0_seq4, and comp77034_c0_seq1) with sequence similarity to known type III polyketide synthases (33.4%-66.3% identity) accumulated to relatively higher steady-state expression levels in leaves and roots relative to other *PKS*-like transcripts (Figure 5.5b) and therefore were selected for further analysis. Transcript comp74373_c0_seq3 was also expressed at high steady-state levels but encoded the same protein as comp74373_c0_seq1 and therefore was not selected for further characterization..

The next objective was to clone cDNAs corresponding to the three selected predicted transcripts to validate the quality of each assembly. PCR was performed using primers designed for the predicted coding region of each of the three assembled transcripts using cDNA synthesized from poison ivy leaf RNA. Two cDNAs corresponding to comp74373_c0_seq1 and comp87518_c0_seq4 amplified successfully (hereafter described as *PKS-like1* and *PKS-like2*, respectively) (Figure 5.8). However, several attempts to amplify a cDNA corresponding to the predicted comp77034_c0_seq1 transcript did not produce an amplfied PCR fragment despite using various annealing temperatures and two different sets of primer pairs. This was surprising, as the same reverse transcribed cDNA batch was used as template for all three PCR reactions. The predicted transcript comp77034_c0_seq1

Figure 5.8: Long and accurate PCR of *T. radicans* predicted type III polyketide synthases. Primers specific to PKS-like1 and PKS-like2 were used to amplify the coding sequence of these unigenes from cDNA.

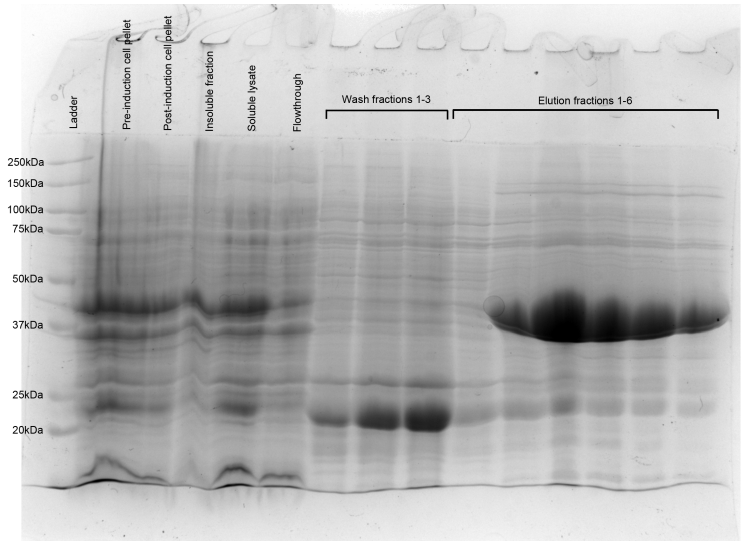


had a deep read coverage across the predicted protein-coding region, suggesting it was not a *de novo* assembly artifact. Sequencing or assembly error may have occurred, resulting in a predicted mRNA transcript that did not exist in the plant.

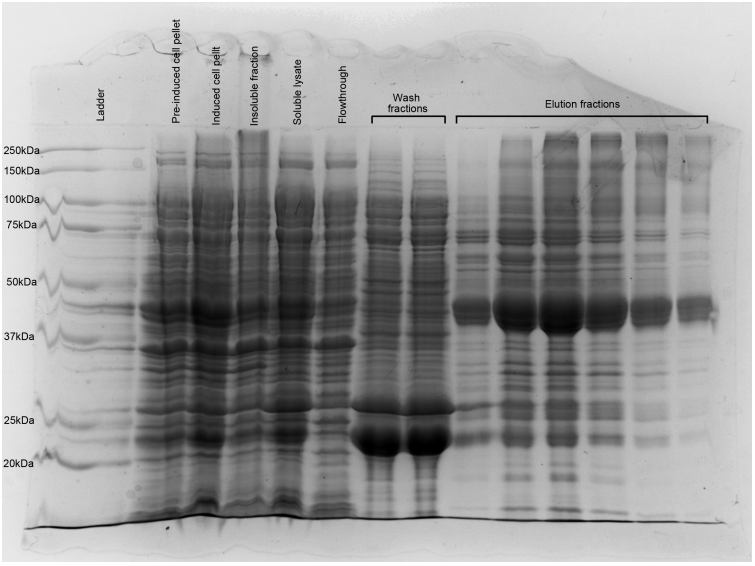
In order to assay the biochemical activity of the two predicted PKS-like proteins, recombinant protein encoded by *PKS-like1* and *PKS-like2* was expressed in *E. coli* and highly enriched (Figure 5.9) for use in future *in vitro* recombinant enzyme assays using fatty acid CoA and malonyl-CoA as substrates. Liquid chromotography MS-MS will be used to determine if these recombinant enzymes produce a released alkyl tetraketide or alkyl resorcyclic/resornicol reaction products.

Figure 5.9: Recombinant protein expression and purification for PKS-like1 and PKS-like2. Recombinant protein encoded by a. *PKS-like1* and b. *PKS-like2* was expressed in E. coli with a N-terminus 6xHIS tag and purified using HisPur Ni-NTA resin. Greatly enriched protein of the expected sizes (43.55 and 43.49 kDa, respectively) can be observed in the elution fractions on the 9% SDS-PAGE gel.

(a)



(b)



5.4.3 *Agrobacterium tumefaciens*-mediated poison ivy transient transformation

Definitive proof that a predicted *T. radicans* gene product is involved in urushiol biosynthesis will require a functional genomics platform in poison ivy. A necessary step in validating *T. radicans* gene function in *T. radicans* will require the introduction of recombinant DNA (rDNA) into poison ivy tissues. Towards that goal, *Agrobacterium tumefaciens* infiltration experiments were performed with the goal of transiently transforming leaf cells with recombinant DNA resulting in the transient expression of a reporter gene *in planta*. Using a modified published protocol for *Agrobacterium* infiltration of tomato leaves [241], both whole poison ivy plants and detached true leaves and cotyledons were infiltrated with *A. tumefaciens* strains containing either a T-DNA binary vector control pSLJ7292, or pSLJ7292 with an in-frame gene fusion of the *A. thaliana RBCS1B-LUC* previously shown to result in stable transformed transgenic plants expressing firefly luciferase activity (i.e. bioluminescence) under control of the *RBCS1B* promoter (*RBCS1B-LUC*, pJGJ204)[242]. Using a single photon imaging system, there was a significant increase in photons emitted from detached true leaves and cotyledons infiltrated with *A. tumefaciens*/*RBCS1B-LUC* gene fusion compared with the empty pSLJ7292 plasmid control (Figure 5.10a-b). However, when *Agrobacterium* containing either pJGJ204 or pSLJ7292 cultures were imaged after the addition of luciferin (no plant infiltration), significantly more photons were observed for the well containing *Agrobacterium* with pJGJ204 compared with the *Agrobacterium* with pSLJ7292 negative control (Figure 5.10c). While the difference in photon emission was quite small (~30 photons over 1 hour,

2 replicates), the higher photon emission from *Agrobacterium* strains containing the *RBCS1B-LUC* construct relative to the vector control strains introjected sufficient uncertainty about the specificity of photo emission resulting from the plant cells.

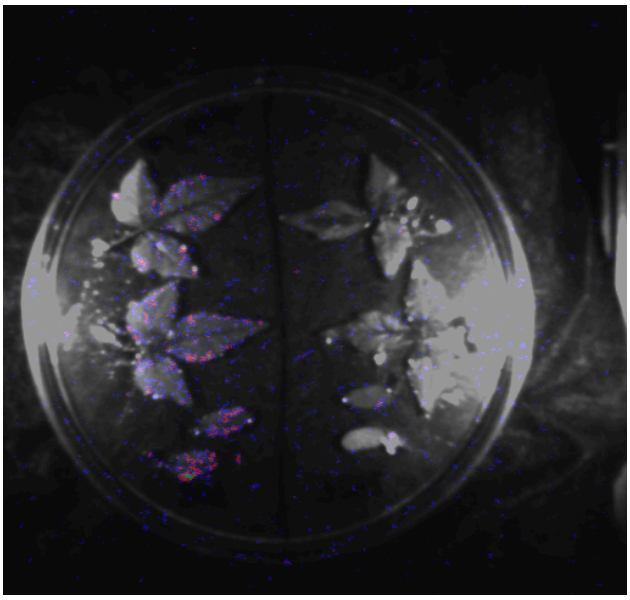
When whole plants were infiltrated with *Agrobacterium* harboring pJGJ204 we did not observe a significant increase in photon counts with the luciferase assay. This may be due to the age of the plants, as they were younger (20 days vs. 35 days since plating on 0.5XMS/phytagel) than the plants used in the detached cotyledons and first true leaves infiltrations and therefore may not have been infiltrated as well. Plants that had been vacuum-infiltrated and then potted in soil dropped many of their leaves and cotyledons, however they continued to grow and form secondary true leaves.

5.5 Discussion

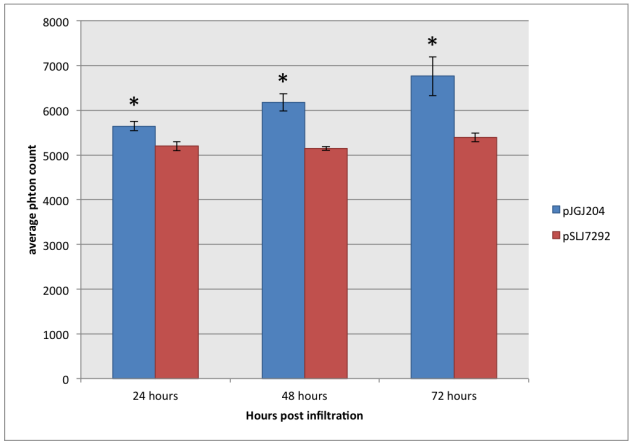
De novo assembled Illumina and/or Roche 454 high-throughput RNA-seq data for non-model plants and other organisms is rapidly replacing previous methods for gene discovery. Transcriptomes for safflower [247], parasitic weed *Cuscuta pentagona* [248], dogwood tree [249], *Chlorophytum borivilianum* [250], alfalfa [251], and many other non-model organisms are recently published. Most studies used one or more of Trinity [227], Oases [252], or SOAPdenovo-Trans [253] assemblers, with comparable results [249]. Trinity was chosen for this analysis due to its excellent reputation and built-in support for downstream analyses; including management of annotations as well as differential expression analysis. A published Trinity pipeline [243] provided a good starting

Figure 5.10: Transient expression of luciferase *in vivo*. Poison ivy first true leaves and cotyledons were infiltrated with A. tumefaciens for transient expression of a RBCS1B-LUC gene fusion and assayed with D-luciferin. a. Photon counting image superimposed over illuminated plate at 48 hours post infiltration. True leaves and cotyledons infiltrated with Agrobacterium containing pJGJ204 are on the left, those containing pSLJ7292 are on the right. b. Photon counts at different timepoints across two replicates. Leaves infiltrated with Agrobacterium containing pJGJ204 had significantly higher photon counts than those with the empty plasmid at 24, 48, and 72 hours post infiltration (1-tail T-test p-values 4.1e-20, 6.4e-26, 7.2e-20, respectively. c. Photon counts of luciferase assays of Agrobacterium containing either pJGJ204 or pSLJ7292 in a 96-well plate over two 1-hour intervals. Lines over bars indicate no significant differences among members (one-way ANOVA F(3,4) = 60.69, p = 0.00086; post-hoc Tukey's HSD).

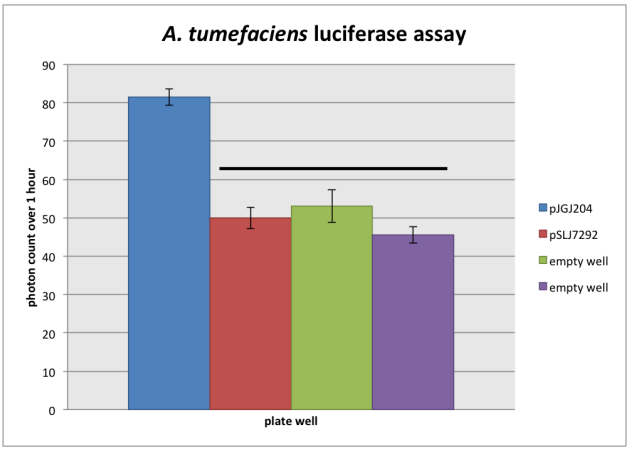
(a)



(b)



(c)



reference with the addition of a few modifications (Figure 5.2).

Illumina sequencing is very sensitive and can produce reads corresponding to mRNAs at very low expression levels. Therefore, minimizing contamination of samples becomes increasingly important, as is the correct identification of assembled transcripts stemming from contamination. It is very difficult to germinate *T. radicans* seedlings without fungal contamination. Molecular analysis of the contaminating fungus identified it as *Colletotrichum fioriniae* [254], which can infect both plant and insect hosts [255]. Even with repeated treatments of drupes with concentrated sulfuric acid and bleach, a large proportion of germinated seedlings still formed visible fungal or bacterial contamination. As such, acquiring axenic plants for sequencing becomes exceedingly challenging, and goals shift to minimizing and identifying potential contamination rather than completely eliminating it. It was expected that many fungal-annotated transcripts would be found in the *T. radicans* transcriptomes, however they were not the predominant form of contamination. Instead, the predominant contaminating sequences seemed to be derived from metazoans.

The metazoan contamination in our dataset may have come from a variety of sources. This foreign RNA (or DNA) may have been introduced during either poison ivy RNA purification or during library preparation. The plants were grown in sterile-culture conditions in closed phytatrays without apparent insect contamination. One possible source of insect nucleic acid contamination could have been from the tap water used to clean the reusable polypropylene Oakridge screw capped tubes that were used during the initial plant RNA extractions. Contamination may also have occurred during library preparation, as multiple samples from a variety of organisms were being simultaneously prepared for sequencing at VBI. Whatever the source of metazoan nucleic acid

contamination, the absolute levels of any given metazoan transcript was typically very small, as evidenced by their consistently very low FPKM values in the expression data. In addition, the proportion of metazoan contaminating sequences in the poor quality Root A sample was greater, and the Root A sample was used in the original *de novo* contig assembly and thereby contributed to the *de novo* assembly.

Some transcripts annotated as coming from a non-plant source may also be novel plant genes not previously sequenced in other members of Viridiplantae. Therefore, their closest sequenced homologs could be from fungi or other Eukarya. Sequencing the *T. radicans* genome would reveal if these transcripts are fungal or metazoan in origin or if they are novel plant genes in poison ivy.

The sequenced poison ivy transcriptome can be used as a tool for generating hypotheses and for gene discovery. The two pathways in Figure 5.6 are composed of enzyme activities proposed by Dewick or in the case of alkylresorcinol synthesis known to occur in plants. One or both of these pathways may exist in poison ivy. Similarity searches using representatives of each enzyme class were used to query the *T. radicans* transcriptome, providing a list of targets for genes possibly involved in the urushiol biosynthetic pathway.

Starting with the first likely step in the urushiol pathway, I attempted to identify a type III polyketide synthase (PKS) that can utilize malonyl-CoA and a 16 or 18 carbon fatty acid-CoA as substrate, as proposed previously in the literature [217, 218]. In several distantly-related plants, alkylresorcinol synthases (ARSs) utilize varying length fatty acid-CoA substrates and extend, cyclize, and aromatize within the ARS (PKS-like) enzymes producing alkylresorcinols, structurally similar but not identical to urushiol [221, 222, 220]. Alkylresorcinols are comprised of an aromatic ring

with two hydroxyl groups (in the meta position, as opposed to ortho in urushiol) and a varying-length alkyl side chain. A similar enzyme in poison ivy may perform one or more of these steps, and therefore these ARSs were used to mine the poison ivy transcriptome for similar proteins (Figure 5.5a; Figure 5.7). Urushiol is found in all growing poison ivy tissues; therefore we expected to see relatively high levels of expression in these tissues for genes involved in urushiol biosynthesis. However, It is also possible that urushiol is not produced constitutively in all cells, but rather restricted to specific cell types. There have been reports of resin ducts or canals in poison ivy that may contain urushiol [256]. If urushiol is only produced in the cells lining these ducts, and these cell types comprise a small percentage of total cells, then in this scenario it is likely that urushiol PKS-encoding transcripts might alternatively be expressed at relatively low overall transcript accumulation levels because of the paucity of cell types producing urushiol. The present study assumed that PKS transcripts responsible for the first step of urushiol biosynthesis would accumulate to high steady state levels in leaves. Relative expression levels narrowed down the potential targets to three unigenes (Figure 5.5b). Reverse transcription PCR (RT-PCR) of two of the three unigenes produced nearly identical coding regions to the *de novo* assembled transcripts with relatively few (one or two) differences. This difference was most likely due to PCR error, as the mapped read coverage at each of these sites in assembled transcripts was ~5000x. This also validated the accuracy of the *de novo* assembly in correctly assembling these transcripts.

The effectiveness of vacuum agroinfiltration of poison ivy cells of leaves/cotyledons was somewhat unclear due to the very low-level expression of luciferase in *Agrobacterium* (Figure 5.10c). Detached leaves and cotyledons that had likely been transformed and expressing luciferase had

significantly higher light emission than empty vector transformed controls in luciferin assays at 24, 48, and 72 hours post-infiltration (Figure 5.10a-b). However, this may be due to expression in *Agrobacterium* located in the inter-cellular regions in the plant or on the surface of leaves. More experiments will need to be taken, particularly the transformation of genes encoding green or red fluorescent protein (GFP or RFP) with cellular organelle targeting, to determine if stable transformation has occurred. Alternatively, transient expression using a particle bombardment/gene gun approach would circumvent all concerns about expression of the reporter gene constructs in an *Agrobacterium*-based transformation system.

5.5.1 Future work

This pilot study was performed with the intent of examining the feasibility of RNA sequencing for poison ivy. A significant limiting factor was the number of sequencing runs (4) available for this project. Since the number of possible mRNA transcripts in *T. radicans* was initially unknown, each sample was sequenced on an individual lane of the HiSeq flow cell in order to maximize the breadth of sequencing. The small number of quality sequenced samples for leaves (2 replicates), and even more so for roots (1 high-quality replicate) severely limited the statistical power of differential expression analyses between tissues. Sequencing more replicates of both *T. radicans* leaves and roots will provide statistical support for any differentially expressed transcripts/unigenes in poison ivy; as well as validate current assembled transcripts, particularly from the single high-quality roots sample.

We observed a high level of read duplication (77.67%-85.86%) in our three high quality samples that had been run on individual lanes of the HiSeq. An assembly of RNA-seq data of these same samples mistakenly multiplexed on a single lane on the HiSeq produced 190,002 transcripts (data not shown), or 90% of the number of transcripts from running on four lanes. Predicted amino acid sequences for *PKS-like1* and *PKS-like2* (as well as comp77034_c0_seq1) remained identical in both assemblies. This suggests that in the future multiple samples can be multiplexed on the same flow cell lane with little loss of sequencing coverage.

A functional genetics platform will need to be developed in poison ivy in order to more accurately assess the function of these proteins *in planta*. Both gene over-expression and targeted knockdowns are important technologies will enable the testing of various hypotheses about urushiol biosynthesis. Over-expression of genes putatively involved in the urushiol biosynthetic pathway followed by quantification of urushiol steady-state levels *in planta* will provide more substantial support than *in vitro* enzyme assays. For example, these methods were recently used to overexpress benzylisoquinoline biosynthesis pathway gene codeinone reductase in opium poppy, resulting in an increase in morphine steady state levels in leaves, which typically do not produce morphine [149]. Likewise, Viral-induced gene silencing (VIGS) using modified tobacco rattle virus (TRV) constructs will be optimized for more comprehensive transformation in poison ivy. This will enable similar experiments to assess the effectiveness of gene expression knockdowns of genes potentially involved in the urushiol biosynthetic pathway (such as *PKS-like1* and *PKS-like2*) and their effect on urushiol accumulation levels. Methods for stable transformation and regeneration also need to be developed to enable the long-term genetic modification of poison ivy.

5.5.2 Conclusions

This work provides a genomic framework for further understanding this under-studied organism. The poison ivy transcriptome was sequenced in two tissues and *de novo* assembled for the first time, representing the first member of *Toxicodendron* to be sequenced *en mass*.

Advances in genomic techniques have spurred novel means of understanding how aggressive and/or invasive plants grow and interact with their environment. Manipulation of weeds at the molecular level may soon replace, or guide and supplement, herbicide-based methods of weed control, and prompt a move towards "molecular weed science" approaches. More research will need to be undertaken to understand what urushiol is doing in the natural environment.

5.5.3 Acknowledgements

We would like to thank Dr. Charles Frazier (Virginia Tech) for generously sharing purified urushiol. We would also like to thank Drs. Pablo Sobrado and Isabel deFonseca for helpful discussion of protein purification and use of lab equipment. This research was funded by a VBI/Fralin Institute Next-Gen DNA sequencing grant award and MPS Student grant and travel awards. Support for AJW was from the PPWS department and GBCB program at Virginia Tech.

Chapter 6

Conclusion

6.1 Molecular evolution in plant specialized metabolism

The results of these varied but inter-related studies illustrated important topics in the molecular evolution in plant secondary metabolism. For example, despite the recognized importance of natural selection during enzyme evolution, there were generally only very weak signals of positive selection (if any) in all the gene families investigated. More commonly, purifying selection was observed at many, but depending upon the gene family, not most sites. The studies in this dissertation provided evidence of intermediary levels of selection somewhere between strong purifying selection to neutral selection. What were conspicuously absent were codons showing d_N/d_S values approximating the theoretical construct of "neutral selection" ($d_N/d_S = 1.0$). This raises the important matter of what d_N/d_S values best approximates the "neutral" and n̈early-neutral" selec-

tion models of Kimura and Ohta? While it is intuitively reasonable to define a d_N/d_S value of 0.9 as "nearly-neutral", it is a far more tenuous proposition to assign d_N/d_S values from 0.2 - 0.5 as "nearly-neutral" selection. In the case of *TEAS*, a considerable proportion of codons showed d_N/d_S values greater than 0.2 but less than 1.0. These intermediate degrees of natural selection constitute what might operationally defined as effectively mutable protein space because natural selection seems to be most variable at these sites. Whether these quantitative metrics constitute "nearly-neutral" defined by Kimura [41] and Ohta [102, 43] is not well understood, and remains an interesting area of future investigation.

It is clear that the enzymes involved in these pathways evolved through often very different means, and that what causes neofunctionalization to emerge in one gene family may not be the same for others. In TEAS and HPS, a change in the overall shape of the active site is sufficient to change enzyme product-profile specificity. These catalytically important second-tier sites were under strong purifying selection in *HPS* and under more relaxed selection in *TEAS*. On the other hand, the evolution of PMT from SPDS likely required important changes in the amino acid sites in direct contact with the substrates putrescine and/or dcSAM/SAM. These sites were under strong purifying selection in both *SPDS* and *PMT*. Differences like these make it less likely that any one specific pattern of natural selection, such as significant positive selection or differences in selective pressure at a given site, will accurately predict all current catalytically important residues in an enzyme.

The diverging evolution of TEAS and HPS from an ancestral terpene synthase may also be a special case due to the nature of the terpene synthase fold. These enzymes evolved from ancestral

secondary metabolic enzymes, rather than one involved in primary metabolism. What is unusual about this enzyme family is the diversity of the different potential carbocation reactions that can be generated from a single FPP substrate. In this case, enzyme specificity does not just hinge on lowering the activation energy for catalysis, but rather guiding which direction the alternative carbocation chemistries that will be permitted to occur by constraining the shape of the active site. Additionally, most of the mutant versions of TEAS and HPS could still initiate carbocation formation, so they were not "dead" enzymes. In contrast, PMT and SPDS mutants often have no measurable catalytic activity of any kind [101, 82, 141]. The corresponding amino acids in the TEAS or HPS active sites are identical, therefore small changes in the shape of the active site of either TEAS or HPS (such as those caused by amino acid substitutions located nearby) can change the product profile due to the already promiscuous nature of either enzyme (24 minor products with TEAS, 12 with HPS), and that catalytically important residues for intiating terpene synthase activity remain unchanged. Previous studies established that second tier sites were responsible for HPS and TEAS reaction specificity. Our $\Delta d_N/d_S$ metric seemed to apply to natural selection acting in second tier sites that were previously established as conferring enzyme product-profile specificity. However, this meant modifying the overall shape of the catalytic active site, and not necessarily the amino acid-substrate contacts required for catalysis between two different substrates.

In SPDS versus PMT the reaction is quite different. This reaction favors either a SAM-dependent methylation, or an aminopropyl transfer from decarboxylated SAM. These are not intra-molecular reactions but inter-molecular reactions in which substrate-enzyme contacts in the active site play a different role than the second tier sites that affect the intra-molecular reactions of terpene syn-

thases. This may partially explain why we did not see adequate conversion of SPDS into having putrescine N-methyltransferase activity. As was later shown by research by the Dräger lab, a single, very separately conserved site in the active site was identified in PMT/SPDS that could interconvert between either activity, albeit at low levels. Their work further showed that both enzymes were showing strong purifying selection (albeit for different AAs) at the key sites responsible for conferring neofunctionalization, and thus would not generate a large $\Delta d_N/d_S$ value indicating different patterns of natural selection. This revealed an inherent deficiency in the $\Delta d_N/d_S$ metric.

An important issue that was raised is just how far back in time the genetic diversity the extant *TEAS*/*HPS* and *PMT*/*SPDS* gene families represent, and how much of that diversity is represented in the d_N/d_S ratios. Since fossil records do not capture phytochemical composition, it is currently impossible to calibrate a molecular clock for these gene families. Was the natural selection event that promoted these sequences to diverge 10, 100, 1000, 10,000, or 1 million years ago? It probably matters a lot, but we do not know. These natural selection analyses were all performed on extant genes. These sequences represent only the current forms of the TEAS/HPS or SPDS/PMT proteins. The mutations that occurred to give rise to new function occurred in ancient genes that were most likely significantly different from their current forms. Adaptive mutations in the ancestral *SPDS* gene responsible for neofunctionalization to a *PMT*, but as evolutionary time went on, purifying selection and/or selective sweeps eliminated the allelic diversity that would indicate positive selection in the extant *PMT* and *SPDS* genes. Therefore, d_N/d_S ratios at these sites would not show this past positive selection, but instead exhibit purifying selection, as we see in our analyses. Evolutionary models that can detect positive selection both at individual sites in an alignment

as well as over evolutionary time (at specific, important nodes in a phylogenetic tree, such as the split between PMT and SPDS) may be able to better infer this past positive selection. The sites identified by this method, however, may still not cause inter-conversion of activity between two closely-related enzymes, as important interactions between residues in a protein may not still exist in the modern enzyme. This is may be the case for PMT, as shown by the work of Junker et al., as N-methyltransferase activity in *D. stramonium* SPDS with modified residue D103I was extremely low, and increased with the addition of two more substitutions from *D. stramonium* PMT (Q79T, V106T).

6.2 Urushiol biosynthesis in *Toxicodendron radicans*

Most molecular studies of plant secondary metabolism have used plants that have been cultivated by humans for thousands of years. These species underwent at least some modest forms of domestication (constrained chemical composition, seed germination, and genetic homogeneity) that facilitated their use as model systems for molecular study. However, "NextGen" DNA sequencing is greatly expanding the scope of molecular studies to non-domesticated plants, such as poison ivy. Sequencing the leaf and root transcriptome of poison ivy is only a first step in understanding this non-model organism, but it provides a platform from which to generate hypotheses about gene discovery/expression (such as which genes are involved in the urushiol biosynthetic pathway), and the biochemistry of recombinant proteins (such as PKS-like1 and PKS-like2). The development of other basic molecular biology tools, such as those important for *in planta* genetic manipulation

(whether through reverse genetic technologies like transient gene over-expression or viral-induced gene silencing, VIGS), will enable the characterization of various genes acting in the poison ivy plant itself and other non-model species. The end goal of these experiments could have applications in novel means of poison ivy weed control or even improvements in lacquer polymer science.

Bibliography

[1] R. A. Dixon and D. Strack, "Phytochemistry meets genome analysis, and beyond." *Phytochemistry*, vol. 62, no. 6, pp. 815–6, Mar. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12590109

[2] K. Yonekura-Sakakibara and K. Saito, "Functional genomics for plant natural product biosynthesis." *Natural product reports*, vol. 26, no. 11, pp. 1466–87, Nov. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19844641

[3] D. W. Christianson, "Unearthing the roots of the terpenome." *Current opinion in chemical biology*, vol. 12, no. 2, pp. 141–50, Apr. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2430190&tool=pmcentrez&rendertype=abstract

[4] J. Ziegler and P. J. Facchini, "Alkaloid biosynthesis: metabolism and trafficking." *Annual review of plant biology*, vol. 59, pp. 735–69, Jan. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18251710

[5] G. T. Carter, "Natural products and pharma 2011: strategic changes spur new opportunities." *Natural product reports*, vol. 28, no. 11, pp. 1783–9, Oct. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21909580

[6] Y. Tanaka, N. Sasaki, and A. Ohmiya, "Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids." *The Plant journal : for cell and molecular biology*, vol. 54, no. 4, pp. 733–49, May 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18476875

[7] J. Mol, E. Grotewold, and R. Koes, "How genes paint flowers and seeds," *Trends in Plant Science*, vol. 3, no. 6, pp. 212–217, Jun. 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1360138598012424

[8] E. Pichersky and E. Lewinsohn, "Convergent evolution in plant specialized metabolism." *Annual review of plant biology*, vol. 62, pp. 549–66, Jan. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21275647

[9] G. S. FRAENKEL, "The raison d'tre of secondary plant substances; these odd chemicals arose as a means of protecting plants from insects and now guide insects to food." *Science (New York, N.Y.)*, vol. 129, no. 3361, pp. 1466–70, May 1959. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/13658975

[10] R. N. Bennett and R. M. Wallsgrove, "Secondary metabolites in plant defence mechanisms," *New Phytologist*, vol. 127, no. 4, pp. 617–633, Aug. 1994. [Online]. Available: http://doi.wiley.com/10.1111/j.1469-8137.1994.tb02968.x

[11] R. A. Dixon, C. J. Lamb, S. Masoud, V. J. Sewalt, and N. L. Paiva, "Metabolic engineering: prospects for crop improvement through the genetic manipulation of phenylpropanoid biosynthesis and defense responses–a review." *Gene*, vol. 179, no. 1, pp. 61–71, Nov. 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8955630

[12] J. Harborne, "The comparative biochemistry of phytoalexin induction in plants," 1999. [Online]. Available: http://www.ingentaconnect.com/content/els/03051978/1999/00000027/00000004/art00095

[13] T. Mitchell-Olds, J. Gershenzon, I. Baldwin, and W. Boland, "Chemical ecology in the molecular era," *Trends in Plant Science*, vol. 3, no. 9, pp. 362–365, Sep. 1998. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1360138598012965

[14] M. F. Vicente, A. Basilio, A. Cabello, and F. Peláez, "Microbial natural products as a source of antifungals." *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, vol. 9, no. 1, pp. 15–32, Jan. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12691539

[15] J. K. Weng and J. P. Noel, "The remarkable pliability and promiscuity of specialized metabolism." *Cold Spring Harbor Symposia on Quantitative Biology*, pp. –, 2012. [Online]. Available: http://pubget.com/site/paper/23269558?institution=vt.eduhttp://symposium.cshlp.org/content/77/309.full.pdf

[16] N. Dudareva and E. Pichersky, "Biochemical and molecular genetic aspects of floral scents." *Plant physiology*, vol. 122, no. 3, pp. 627–33, Mar. 2000. [On-

line]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1539243&tool=pmcentrez&rendertype=abstract

[17] D. R. Gang, "Evolution of flavors and scents." *Annual review of plant biology*, vol. 56, pp. 301–25, Jan. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15862098

[18] E. A. Shank and R. Kolter, "New developments in microbial interspecies signaling." *Current opinion in microbiology*, vol. 12, no. 2, pp. 205–14, Apr. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2709175&tool=pmcentrez&rendertype=abstract

[19] J.-K. Weng and C. Chapple, "The origin and evolution of lignin biosynthesis." *The New phytologist*, vol. 187, no. 2, pp. 273–85, Jul. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20642725

[20] J. A. Banks, T. Nishiyama, M. Hasebe, J. L. Bowman, M. Gribskov, C. DePamphilis, V. A. Albert, N. Aono, T. Aoyama, B. A. Ambrose, N. W. Ashton, M. J. Axtell, E. Barker, M. S. Barker, J. L. Bennetzen, N. D. Bonawitz, C. Chapple, C. Cheng, L. G. G. Correa, M. Dacre, J. DeBarry, I. Dreyer, M. Elias, E. M. Engstrom, M. Estelle, L. Feng, C. Finet, S. K. Floyd, W. B. Frommer, T. Fujita, L. Gramzow, M. Gutensohn, J. Harholt, M. Hattori, A. Heyl, T. Hirai, Y. Hiwatashi, M. Ishikawa, M. Iwata, K. G. Karol, B. Koehler, U. Kolukisaoglu, M. Kubo, T. Kurata, S. Lalonde, K. Li, Y. Li, A. Litt, E. Lyons, G. Manning, T. Maruyama, T. P. Michael, K. Mikami, S. Miyazaki, S.-i. Morinaga, T. Murata, B. Mueller-Roeber, D. R. Nelson, M. Obara, Y. Oguri, R. G. Olmstead, N. Onodera, B. L. Petersen,

B. Pils, M. Prigge, S. A. Rensing, D. M. Riaño Pachón, A. W. Roberts, Y. Sato, H. V. Scheller, B. Schulz, C. Schulz, E. V. Shakirov, N. Shibagaki, N. Shinohara, D. E. Shippen, I. Sø rensen, R. Sotooka, N. Sugimoto, M. Sugita, N. Sumikawa, M. Tanurdzic, G. Theissen, P. Ulvskov, S. Wakazuki, J.-K. Weng, W. W. G. T. Willats, D. Wipf, P. G. Wolf, L. Yang, A. D. Zimmer, Q. Zhu, T. Mitros, U. Hellsten, D. Loqué, R. Otillar, A. Salamov, J. Schmutz, H. Shapiro, E. Lindquist, S. Lucas, D. Rokhsar, and I. V. Grigoriev, "The selaginella genome identifies genetic changes associated with the evolution of vascular plants." *Science (New York, N.Y.)*, vol. 332, no. 6032, pp. 960–3, May 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3166216&tool=pmcentrez&rendertype=abstract

[21] J. B. Lowry, D. W. Lee, and C. Hébant, *the Origin of Land Plants: A New Look at an Old Problem*, 1980. [Online]. Available: http://books.google.com/books/about/The_Origin_of_Land_Plants.html?id=Oo0QHQAACAAJ&pgis=1

[22] E. Pichersky and D. R. Gang, "Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective." *Trends in plant science*, vol. 5, no. 10, pp. 439–45, Oct. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11044721

[23] S. Ohno, *Evolution by Gene Duplication*, 1970. [Online]. Available: http://books.google.com/books/about/Evolution_by_gene_duplication.html?hl=nl&id=sxUDAAAAMAAJ&pgis=1

[24] C. M. Fraser, M. G. Thompson, A. M. Shirley, J. Ralph, J. A. Schoenherr, T. Sinlapadech, M. C. Hall, and C. Chapple, "Related arabidopsis serine carboxypeptidase-like sinapoylglucose acyltransferases display distinct but overlapping substrate specificities." *Plant physiology*, vol. 144, no. 4, pp. 1986–99, Aug. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1949888&tool=pmcentrez&rendertype=abstract

[25] D. J. Kliebenstein, "A role for gene duplication and natural variation of gene expression in the evolution of metabolism." *PloS one*, vol. 3, no. 3, p. e1838, Jan. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2263126&tool=pmcentrez&rendertype=abstract

[26] J. Bohlmann, G. Meyer-Gauen, and R. Croteau, "Plant terpenoid synthases: molecular biology and phylogenetic analysis." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 8, pp. 4126–33, Apr. 1998. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=22453&tool=pmcentrez&rendertype=abstract

[27] T. Hashimoto and Y. Yamada, "New genes in alkaloid metabolism and transport." *Current opinion in biotechnology*, vol. 14, no. 2, pp. 163–8, Apr. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12732317

[28] D. Ober, "Seeing double: gene duplication and diversification in plant secondary metabolism." *Trends in plant science*, vol. 10, no. 9, pp. 444–9, Sep. 2005. [Online].

Available: http://www.ncbi.nlm.nih.gov/pubmed/16054418

[29] A. Bar-Even, E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S. Tawfik, and R. Milo, "The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters." *Biochemistry*, vol. 50, no. 21, pp. 4402–10, May 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21506553

[30] W. White, D. Arias-Garzon, J. McMahon, and R. Sayre, "Cyanogenesis in cassava. the role of hydroxynitrile lyase in root cyanide production," *Plant physiology*, vol. 116, no. 4, pp. 1219–25, Apr. 1998. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=35028&tool=pmcentrez&rendertype=abstract

[31] P. Mazzafera and A. Carvalho, "Breeding for low seed caffeine content of coffee (Coffea L.) by interspecific hybridization," *Euphytica*, vol. 59, no. 1, pp. 55–60, Nov. 1991. [Online]. Available: http://link.springer.com/article/10.1007/BF00025361

[32] H. Ashihara, M. Kato, and A. Crozier, "Distribution, biosynthesis and catabolism of methylxanthines in plants." *Handbook of experimental pharmacology*, no. 200, pp. 11–31, Jan. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20859792

[33] D. Ober and T. Hartmann, "Homospermidine synthase, the first pathway-specific enzyme of pyrrolizidine alkaloid biosynthesis, evolved from deoxyhypusine synthase." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 26, pp. 14 777–82, Dec. 1999. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24724&tool=pmcentrez&rendertype=abstract

[34] D. Ober and E. Kaltenegger, "Pyrrolizidine alkaloid biosynthesis, evolution of a pathway in plant secondary metabolism." *Phytochemistry*, vol. 70, no. 15-16, pp. 1687–95, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19545881

[35] F. Chen, D.-k. Ro, J. Petri, J. Gershenzon, J. Bohlmann, E. Pichersky, and D. Tholl, "Characterization of a root-specific arabidopsis terpene synthase responsible for the formation of the volatile," *Plant physiology*, vol. 135, no. August, pp. 1956–1966, 2004.

[36] M. Bevan, I. Bancroft, E. Bent, K. Love, H. Goodman, C. Dean, R. Bergkamp, W. Dirkse, M. Van Staveren, W. Stiekema, L. Drost, P. Ridley, S. A. Hudson, K. Patel, G. Murphy, P. Piffanelli, H. Wedler, E. Wedler, R. Wambutt, T. Weitzenegger, T. M. Pohl, N. Terryn, J. Gielen, R. Villarroel, R. De Clerck, M. Van Montagu, A. Lecharny, S. Auborg, I. Gy, M. Kreis, N. Lao, T. Kavanagh, S. Hempel, P. Kotter, K. D. Entian, M. Rieger, M. Schaeffer, B. Funk, S. Mueller-Auer, M. Silvey, R. James, A. Montfort, A. Pons, P. Puigdomenech, A. Douka, E. Voukelatou, D. Milioni, P. Hatzopoulos, E. Piravandi, B. Obermaier, H. Hilbert, A. Düsterhöft, T. Moores, J. D. Jones, T. Eneva, K. Palme, V. Benes, S. Rechman, W. Ansorge, R. Cooke, C. Berger, M. Delseny, M. Voet, G. Volckaert, H. W. Mewes, S. Klosterman, C. Schueller, and N. Chalwatzis, "Analysis of 1.9 mb of contiguous sequence from chromosome 4 of arabidopsis thaliana." *Nature*, vol. 391, no. 6666, pp. 485–8, Jan. 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9461215

[37] C. Somerville and S. Somerville, "Plant functional genomics." *Science (New York, N.Y.)*, vol. 285, no. 5426, pp. 380–3, Jul. 1999. [Online]. Available: http:

//www.ncbi.nlm.nih.gov/pubmed/10411495

[38] S. E. O'Connor and J. J. Maresh, "Chemistry and biology of monoterpene indole alkaloid biosynthesis." *Natural product reports*, vol. 23, no. 4, pp. 532–47, Aug. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16874388

[39] R. Croteau, R. E. B. Ketchum, R. M. Long, R. Kaspera, and M. R. Wildung, "Taxol biosynthesis and molecular genetics." *Phytochemistry reviews : proceedings of the Phytochemical Society of Europe*, vol. 5, no. 1, pp. 75–97, Feb. 2006. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2901146&tool=pmcentrez&rendertype=abstract

[40] P. Cook and W. Cleland, *Enzyme Kinetics and Mechanism*, 1st ed. Garland Science, 2007. [Online]. Available: http://www.garlandscience.com/product/isbn/9780815341406?fromSearchResults=fromAlphaSearchResults

[41] M. Kimura, "Evolutionary rate at the molecular level," *Nature*, vol. 217, no. 5129, pp. 624–626, 1968. [Online]. Available: http://www.eebweb.arizona.edu/Courses/Ecol426_526/Kimura_1968.pdf

[42] J. L. King and T. H. Jukes, "Non-darwinian evolution," *Science*, vol. 164, no. 881, pp. 788–798, 1969. [Online]. Available: http://www.sciencemag.org/cgi/reprint/sci;164/3881/788.pdf

[43] T. Ohta, "The nearly neutral theory of molecular evolution," *Annual Review of Ecology and Systematics*, vol. 23, pp. 263–286, 1992. [Online]. Available: http://www.jstor.org/stable/2097289

[44] O. Khersonsky, C. Roodveldt, and D. Tawfik, "Enzyme promiscuity: evolutionary and mechanistic aspects," *Current Opinion in Chemical Biology*, vol. 10, no. 5, pp. 498–508, 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1367593106001189http://ac.els-cdn.com/S1367593106001189/1-s2.0-S1367593106001189-main.pdf?_tid=fc8466e0-c0cd-11e3-bfd9-00000aacb360&acdnat=1397147849_3b8d8b5e46b19b07046d7707f2512d35

[45] S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, and D. S. Tawfik, "Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein." *Nature*, vol. 444, no. 7121, pp. 929–32, Dec. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17122770

[46] N. Tokuriki and D. S. Tawfik, "Stability effects of mutations and protein evolvability," *Current Opinion in Structural Biology*, vol. 19, no. 5, pp. 596–604, 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0959440X09001249http://ac.els-cdn.com/S0959440X09001249/1-s2.0-S0959440X09001249-main.pdf?_tid=fa0a6b80-c0cd-11e3-9243-00000aacb35e&acdnat=1397147844_796d19f4f2faab8fad5f7b66ca8faf93

[47] ——, "Protein dynamism and evolvability." *Science*, vol. 324, no. 5924, pp. 203–207, 2009. [Online]. Available: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19359577&retmode=ref&cmd=prlinkshttp://www.sciencemag.org/content/324/5924/203.full.pdf

[48] O. K. Tawfik and D. S, "Enzyme promiscuity: A mechanistic and evolutionary perspective," *Annual Review Biochemistry*, vol. 79, no. 1, pp. 471–505, 2010. [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev-biochem-030409-143718http://www.annualreviews.org/doi/pdf/10.1146/annurev-biochem-030409-143718

[49] M. Camps, A. Herman, E. Loh, and L. A. Loeb, "Genetic constraints on protein evolution." *Critical reviews in biochemistry and molecular biology*, vol. 42, no. 5, pp. 313–26, 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3825456&tool=pmcentrez&rendertype=abstract

[50] B. D. Smith and R. T. Raines, "Genetic selection for critical residues in ribonucleases." *Journal of molecular biology*, vol. 362, no. 3, pp. 459–78, Sep. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16920150

[51] S. Bershtein and D. S. Tawfik, "Advances in laboratory evolution of enzymes," *Current Opinion in Chemical Biology*, vol. 12, no. 2, pp. 151–158, 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1367593108000173http://ac.els-cdn.com/S1367593108000173/1-s2.

0-S1367593108000173-main.pdf?_tid=005e9b46-c0ce-11e3-a4f6-00000aab0f27&acdnat=1397147855_d12840082371d8f56432b8d8f13eb34f

[52] P. Yue, Z. Li, and J. Moult, "Loss of protein structure stability as a major causative factor in monogenic disease." *Journal of molecular biology*, vol. 353, no. 2, pp. 459–73, Oct. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16169011

[53] J.-K. Weng, R. N. Philippe, and J. P. Noel, "The rise of chemodiversity in plants," *Science*, vol. 336, no. 6089, pp. 1667–1670, 2012. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1217411http://www.sciencemag.org/content/336/6089/1667.full.pdf

[54] H. Nam, N. E. Lewis, J. A. Lerman, D.-H. Lee, R. L. Chang, D. Kim, and B. O. Palsson, "Network context and selection in the evolution to enzyme specificity." *Science (New York, N.Y.)*, vol. 337, no. 6098, pp. 1101–4, Aug. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3536066&tool=pmcentrez&rendertype=abstract

[55] E. Ferrada and A. Wagner, "Protein robustness promotes evolutionary innovations on large evolutionary time-scales." *Proceedings. Biological sciences / The Royal Society*, vol. 275, no. 1643, pp. 1595–602, Jul. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2602809&tool=pmcentrez&rendertype=abstract

[56] M. A. DePristo, D. M. Weinreich, and D. L. Hartl, "Missense meanderings in sequence space: a biophysical view of protein evolution." *Nature reviews. Genetics*, vol. 6, no. 9, pp.

678–87, Sep. 2005. [Online]. Available: http://dx.doi.org/10.1038/nrg1672

[57] X. Wang, G. Minasov, and B. K. Shoichet, "Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs." *Journal of molecular biology*, vol. 320, no. 1, pp. 85–95, Jun. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12079336

[58] E. A. Ortlund, J. T. Bridgham, M. R. Redinbo, and J. W. Thornton, "Crystal structure of an ancient protein: evolution by conformational epistasis." *Science (New York, N.Y.)*, vol. 317, no. 5844, pp. 1544–8, Sep. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2519897&tool=pmcentrez&rendertype=abstract

[59] L. Zhang and L. T. Watson, "Analysis of the fitness effect of compensatory mutations." *HFSP journal*, vol. 3, no. 1, pp. 47–54, Jan. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2689613&tool=pmcentrez&rendertype=abstract

[60] A. M. Dean and J. W. Thornton, "Mechanistic approaches to the study of evolution: the functional synthesis." *Nature reviews. Genetics*, vol. 8, no. 9, pp. 675–88, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1038/nrg2160

[61] L. C. James and D. S. Tawfik, "Conformational diversity and protein evolution–a 60-year-old hypothesis revisited." *Trends in biochemical sciences*, vol. 28, no. 7, pp. 361–8, Jul. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12878003

[62] S. Meier and S. Ozbek, "A biological cosmos of parallel universes: does protein structural plasticity facilitate evolution?" *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 29, no. 11, pp. 1095–104, Nov. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17935152

[63] K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern, "A hierarchy of timescales in protein dynamics is linked to enzyme catalysis." *Nature*, vol. 450, no. 7171, pp. 913–6, Dec. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18026087

[64] N. M. Goodey and S. J. Benkovic, "Allosteric regulation and catalysis emerge via a common route." *Nature chemical biology*, vol. 4, no. 8, pp. 474–82, Aug. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18641628

[65] A. Andreeva and A. G. Murzin, "Evolution of protein fold in the presence of functional constraints." *Current opinion in structural biology*, vol. 16, no. 3, pp. 399–408, Jun. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16650981

[66] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman, "Interconversion between two unrelated protein folds in the lymphotactin native state." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 13, pp. 5057–62, Apr. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2278211&tool=pmcentrez&rendertype=abstract

[67] Y. Chen, J. Delmas, J. Sirot, B. Shoichet, and R. Bonnet, "Atomic resolution structures of ctx-M beta-lactamases: extended spectrum activities from increased mobility and decreased

stability." *Journal of molecular biology*, vol. 348, no. 2, pp. 349–62, Apr. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15811373

[68] N. Tokuriki, F. Stricher, L. Serrano, and D. S. Tawfik, "How protein stability and new functions trade off." *PLoS computational biology*, vol. 4, no. 2, p. e1000002, Mar. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265470&tool=pmcentrez&rendertype=abstract

[69] B. K. Muralidhara, L. Sun, S. Negi, and J. R. Halpert, "Thermodynamic fidelity of the mammalian cytochrome p450 2b4 active site in binding substrates and inhibitors." *Journal of molecular biology*, vol. 377, no. 1, pp. 232–45, Mar. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18241887

[70] J. Skopalík, P. Anzenbacher, and M. Otyepka, "Flexibility of human cytochromes p450: molecular dynamics reveals differences between cyps 3a4, 2c9, and 2a6, which correlate with their substrate preferences." *The journal of physical chemistry. B*, vol. 112, no. 27, pp. 8165–73, Jul. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18598011

[71] P. E. O'Maille, A. Malone, N. Dellas, B. Andes Hess, L. Smentek, I. Sheehan, B. T. Greenhagen, J. Chappell, G. Manning, and J. P. Noel, "Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases." *Nature chemical biology*, vol. 4, no. 10, pp. 617–23, Oct. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2664519&tool=pmcentrez&rendertype=abstract

[72] B. T. Greenhagen, P. E. O'Maille, J. P. Noel, and J. Chappell, "Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9826–31, Jun. 2006. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1502538&tool=pmcentrez&rendertype=abstract

[73] J. Wang and V. De Luca, "The biosynthesis and regulation of biosynthesis of concord grape fruit esters, including 'foxy' methylanthranilate." *The Plant journal : for cell and molecular biology*, vol. 44, no. 4, pp. 606–19, Nov. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16262710

[74] T. G. Köllner, C. Lenk, N. Zhao, I. Seidl-Adams, J. Gershenzon, F. Chen, and J. Degenhardt, "Herbivore-induced sabath methyltransferases of maize that methylate anthranilic acid using s-adenosyl-L-methionine." *Plant physiology*, vol. 153, no. 4, pp. 1795–807, Aug. 2010. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923889&tool=pmcentrez&rendertype=abstract

[75] A. Hickel, M. Hasslacher, and H. Griengl, "Hydroxynitrile lyases: functions and properties," *Physiologia Plantarum*, vol. 98, no. 4, pp. 891–898, Dec. 1996. [Online]. Available: http://doi.wiley.com/10.1111/j.1399-3054.1996.tb06700.x

[76] D. Strack, T. Vogt, and W. Schliemann, "Recent advances in betalain research." *Phytochemistry*, vol. 62, no. 3, pp. 247–69, Feb. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12620337

[77] A. Junker, J. Fischer, Y. Sichhart, W. Brandt, and B. Dräger, "Evolution of the key alkaloid enzyme putrescine N-methyltransferase from spermidine synthase." *Frontiers in plant science*, vol. 4, p. 260, Jan. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3725402&tool=pmcentrez&rendertype=abstract

[78] M. J. Harms and J. W. Thornton, "Analyzing protein structure and function using ancestral gene reconstruction," *Current Opinion in Structural Biology*, vol. 20, no. 3, pp. 360–366, 2010. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0959440X10000588http://ac.els-cdn.com/S0959440X10000588/1-s2.0-S0959440X10000588-main.pdf?_tid=40a52c24-c0ce-11e3-bfdb-00000aacb361&acdnat=1397147963_4edf38bd18cd4bf99a066cee5a2343eb

[79] A. K. Datta, "Comparative sequence analysis in the sialyltransferase protein family: analysis of motifs." *Current drug targets*, vol. 10, no. 6, pp. 483–98, Jun. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19519350

[80] T. A. Desai, D. A. Rodionov, M. S. Gelfand, E. J. Alm, and C. V. Rao, "Engineering transcription factors with novel dna-binding specificity using comparative genomics." *Nucleic acids research*, vol. 37, no. 8, pp. 2493–503, May 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2677863&tool=pmcentrez&rendertype=abstract

[81] N. Maita, J. Nyirenda, M. Igura, J. Kamishikiryo, and D. Kohda, "Comparative structural biology of eubacterial and archaeal oligosaccharyltransferases." *The*

Journal of biological chemistry, vol. 285, no. 7, pp. 4941–50, Feb. 2010. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2836098&tool=pmcentrez&rendertype=abstract

[82] S. Biastoff, N. Reinhardt, V. Reva, W. Brandt, and B. Dräger, "Evolution of putrescine N-methyltransferase from spermidine synthase demanded alterations in substrate binding." *FEBS letters*, vol. 583, no. 20, pp. 3367–74, Oct. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19796640

[83] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in BioSciences*, vol. 13, pp. 555–556, 1997.

[84] W. Yang, J. P. Bielawski, and Z. Yang, "Widespread adaptive evolution in the human immunodeficiency virus type 1 genome." *Journal of molecular evolution*, vol. 57, no. 2, pp. 212–21, Aug. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14562964

[85] L. Muggia, I. Schmitt, and M. Grube, "Purifying selection is a prevailing motif in the evolution of ketoacyl synthase domains of polyketide synthases from lichenized fungi." *Mycological research*, vol. 112, no. Pt 2, pp. 277–88, Mar. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0953756207002043

[86] H. W. Kircher and F. V. Lieberman, "Toxicity of tobacco smoke to the spotted alfalfa aphid therioaphis maculata (Buckton)." *Nature*, vol. 215, no. 5096, pp. 97–8, Jul. 1967. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6053423

[87] T. Richardson, R. Robinson, and E. Seijo, "171. synthetical experiments in the chelidonine?sanguinarine group of the alkaloids. part I," *Journal of the Chemical Society (Resumed)*, p. 835, Jan. 1937. [Online]. Available: http://pubs.rsc.org/en/content/articlehtml/1937/jr/jr9370000835

[88] F. Saitoh, M. Noma, and N. Kawashima, "The alkaloid contents of sixty nicotiana species," *Phytochemistry*, vol. 24, no. 3, pp. 477–480, Jan. 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031942200807517

[89] F. Smith and D. Goodhue, "Toxicity of nicotine aerosols to the green peach aphid, under greenhouse conditions," *Journal of Economic Entomology*, vol. 36, pp. 911–914, 1943. [Online]. Available: http://scholar.qsensei.com/content/9zq4d

[90] M. Tomizawa and J. E. Casida, "Selective toxicity of neonicotinoids attributable to specificity of insect and mammalian nicotinic receptors." *Annual review of entomology*, vol. 48, pp. 339–64, Jan. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12208819

[91] A. Nahrstedt, "Relationships between the defensive systems of plants and insects," *Recent Advances in Phytochemistry*, vol. 30, pp. 217–230, 1996.

[92] P. Pare and J. Tumlinson, "Plant volatiles as a defense against insect herbivores," *Plant physiology*, vol. 121, no. 2, pp. 325–32, Oct. 1999. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1539229&tool=pmcentrez&rendertype=abstract

[93] D. Tholl, "Terpene synthases and the regulation, diversity and biological roles of terpene metabolism." *Current opinion in plant biology*, vol. 9, no. 3, pp. 297–304, Jun. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16600670

[94] D. J. Newman and G. M. Cragg, "Natural products as sources of new drugs over the last 25 years." *Journal of natural products*, vol. 70, no. 3, pp. 461–77, Mar. 2007. [Online]. Available: http://dx.doi.org/10.1021/np068054v

[95] M. Wink, "A short history of alkaloids," in *Alkaloids: Biochemistry, Ecology, and Medicinal Applications*, M. Roberts and M. Wink, Eds. New York: Plenum Press, 1998, p. 486.

[96] J. Chappell, "Production platforms for the molecular pharming of alkaloid diversity," *Proceedings of the National Academy of Sciences*, vol. 105, no. 23, pp. 7897–7898, 2008. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0803930105http://www.pnas.org/content/105/23/7897.full.pdf

[97] J. A. Gerlt and P. C. Babbitt, "Enzyme (re)design: lessons from natural evolution and computation," *Current Opinion in Chemical Biology*, vol. 13, no. 1, pp. 10–18, 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1367593109000040http://ac.els-cdn.com/S1367593109000040/1-s2.0-S1367593109000040-main.pdf?_tid=3d6d144a-c0ce-11e3-be04-00000aab0f27&acdnat=1397147957_e25be74ed859a4030da1e34f26120e6e

[98] S. Yokoyama, T. Tada, H. Zhang, and L. Britt, "Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates." *Proceedings of the National*

Academy of Sciences of the United States of America, vol. 105, no. 36, pp. 13 480–5, Sep. 2008. [Online]. Available: http://www.pnas.org/content/105/36/13480

[99] S. F. Field and M. V. Matz, "Retracing evolution of red fluorescence in gfp-like proteins from faviina corals." *Molecular biology and evolution*, vol. 27, no. 2, pp. 225–33, Mar. 2010. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2877551&tool=pmcentrez&rendertype=abstract

[100] J. E. Vick and J. A. Gerlt, "Evolutionary potential of (beta/alpha)8-barrels: stepwise evolution of a "new" reaction in the enolase superfamily." *Biochemistry*, vol. 46, no. 50, pp. 14 589–97, Dec. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18020459

[101] M. Teuber, M. E. Azemi, F. Namjoyan, A.-C. Meier, A. Wodak, W. Brandt, and B. Dräger, "Putrescine N-methyltransferases–a structure-function analysis." *Plant molecular biology*, vol. 63, no. 6, pp. 787–801, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17221359

[102] T. Ohta, "Slightly deleterious mutant substitutions in evolution." *Nature*, vol. 246, no. 5428, pp. 96–8, Nov. 1973. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/4585855

[103] R. A. Jensen, "Enzyme recruitment in evolution of new function." *Annual review of microbiology*, vol. 30, pp. 409–25, Jan. 1976. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/791073

[104] J. H. Langenheim, "Higher plant terpenoids: A phytocentric overview of their ecological roles." *Journal of chemical ecology*, vol. 20, no. 6, pp. 1223–80, Jun. 1994. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24242340

[105] J. Gershenzon and W. Kreis, "Biosynthesis of monoterpenes, sesquiterpenes, diterpenes, sterols, cardiac glycosides and steroid saponins," in *Biochemistry of Plant Secondary Metabolism*, M. Wink, Ed. Sheffield: Sheffield Academic Press, 1999, pp. 222–299. [Online]. Available: http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:439009

[106] E. Pichersky and J. Gershenzon, "The formation and function of plant volatiles: perfumes for pollinator attraction and defense." *Current opinion in plant biology*, vol. 5, no. 3, pp. 237–43, Jun. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11960742

[107] S. B. Unsicker, G. Kunert, and J. Gershenzon, "Protective perfumes: the role of vegetative volatiles in plant defense against herbivores." *Current opinion in plant biology*, vol. 12, no. 4, pp. 479–85, Aug. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19467919

[108] M. Dicke and I. T. Baldwin, "The evolutionary context for herbivore-induced plant volatiles: beyond the 'cry for help'." *Trends in plant science*, vol. 15, no. 3, pp. 167–75, Mar. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20047849

[109] G. Arimura, R. Ozawa, T. Shimoda, T. Nishioka, W. Boland, and J. Takabayashi, "Herbivory-induced volatiles elicit defence genes in lima bean leaves." *Nature*, vol.

406, no. 6795, pp. 512–5, Aug. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10952311

[110] H. K. Lichtenthaler, "THE 1-deoxy-D-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants." *Annual review of plant physiology and plant molecular biology*, vol. 50, pp. 47–65, Jun. 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15012203

[111] M. Sapir-Mir, A. Mett, E. Belausov, S. Tal-Meshulam, A. Frydman, D. Gidoni, and Y. Eyal, "Peroxisomal localization of arabidopsis isopentenyl diphosphate isomerases suggests that part of the plant isoprenoid mevalonic acid pathway is compartmentalized to peroxisomes." *Plant physiology*, vol. 148, no. 3, pp. 1219–28, Nov. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2577245&tool=pmcentrez&rendertype=abstract

[112] J. Degenhardt, T. G. Köllner, and J. Gershenzon, "Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants." *Phytochemistry*, vol. 70, no. 15-16, pp. 1621–37, Jan. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031942209003057

[113] R. Croteau, "Biosynthesis and catabolism of monoterpenoids," *Chemical Reviews*, vol. 87, no. 5, pp. 929–954, Oct. 1987. [Online]. Available: http://dx.doi.org/10.1021/cr00081a004

[114] D. M. Martin, J. Fäldt, and J. Bohlmann, "Functional characterization of nine norway spruce tps genes and evolution of gymnosperm terpene synthases of the

tps-d subfamily." *Plant physiology*, vol. 135, no. 4, pp. 1908–27, Aug. 2004. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=520763&tool=pmcentrez&rendertype=abstract

[115] S. C. Trapp and R. B. Croteau, "Genomic organization of plant terpene synthases and molecular evolutionary implications." *Genetics*, vol. 158, no. 2, pp. 811–32, Jun. 2001. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461696&tool=pmcentrez&rendertype=abstract

[116] D. W. Christianson, "Structural biology and chemistry of the terpenoid cyclases." *Chemical reviews*, vol. 106, no. 8, pp. 3412–42, Aug. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16895335

[117] K. U. Wendt and G. E. Schulz, "Isoprenoid biosynthesis: manifold chemistry catalyzed by similar enzymes." *Structure (London, England : 1993)*, vol. 6, no. 2, pp. 127–33, Mar. 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9519404

[118] C. M. Starks, "Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase," *Science*, vol. 277, no. 5333, pp. 1815–1820, 1997. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.277.5333.1815

[119] J. P. Noel, "Chemical biology: synthetic metabolism goes green," *Nature*, vol. 468, no. 7322, pp. 380–381, 2010. [Online]. Available: http://www.nature.com/doifinder/10.1038/468380ahttp://www.nature.com/nature/journal/v468/n7322/pdf/468380a.pdf

[120] P. Marrero, C. Poulter, and P. Edwards, "Effects of site-directed mutagenesis of the highly conserved aspartate residues in domain ii of farnesyl diphosphate synthase activity," *J. Biol. Chem.*, vol. 267, no. 30, pp. 21 873–21 878, Oct. 1992. [Online]. Available: http://www.jbc.org/content/267/30/21873.abstract?ijkey=a84c34a4b7c0ea884d59849af34c24575f280c9f&keytype2=tf_ipsecsha

[121] K. Back and J. Chappell, "Identifying functional domains within terpene cyclases using a domain-swapping strategy." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 13, pp. 6841–5, Jun. 1996. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=39115&tool=pmcentrez&rendertype=abstract

[122] Katoh, Misawa, Kuma, and Miyata, "MAFFT: a novel method for rapid multiple sequence alignment base on fast fourier transform," *Nucleic Acids Res.*, vol. 30, pp. 3059–3066, 2002.

[123] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[124] D. Zwickl, "Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion," Ph.D. dissertation, The University of Texas at Austin, 2006.

[125] D. L. Swofford, "PAUP*. phylogenetic analysis using parsimony (*and other methods)," Sunderland, Massachusetts, 2003.

[126] J. Sukumaran and M. T. Holder, "DendroPy: A python library for phylogenetic computing," *Bioinformatics*, vol. 26, pp. 1569–1571, 2010.

[127] M. Suyama, D. Torrents, and P. Bork, "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments." *Nucleic acids research*, vol. 34, no. Web Server issue, pp. W609–12, Jul. 2006. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1538804&tool=pmcentrez&rendertype=abstract

[128] B. Moury, "A new lineage sheds light on the evolutionary history of potato virus Y." *Molecular plant pathology*, vol. 11, no. 1, pp. 161–8, Jan. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20078785

[129] P. J. Facchini, "ALKALOID biosynthesis in plants: biochemistry, cell biology, molecular regulation, and metabolic engineering applications." *Annual review of plant physiology and plant molecular biology*, vol. 52, pp. 29–66, Jun. 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11337391

[130] S. L. Rutherford, "Between genotype and phenotype: protein chaperones and evolvability." *Nature reviews. Genetics*, vol. 4, no. 4, pp. 263–74, Apr. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12671657

[131] T. Hashimoto, T. Shoji, T. Mihara, H. Oguri, K. Tamaki, K.-i. Suzuki, and Y. Yamada, "Intraspecific variability of the tandem repeats in nicotiana putrescine N-methyltransferases." *Plant molecular biology*, vol. 37, pp. 25–37, 1998. [Online]. Available: http://media.proquest.com/media/pq/classic/doc/2188865391/fmt/pi/rep/

NONE?hl=&cit:auth=Hashimoto,+Takashi;Shoji,+Tsubasa;Mihara,+Taku;Oguri,+Hideo;Tamaki,+Katsutomo;Suzuki,+Ken-ichi;Yamada,+Yasuyuki&cit:title=Intraspecific+variability+of+the+tandem+repeats+in+Nicotiana+putrescine+N-methyltransferases&cit:pub=Plant+Molecular+Biology&cit:vol=37&cit:iss=1&cit:pg=25&cit:date=May+1998&ic=true&cit:prod=ProQuest+Biological+Science+Collection&_a=ChgyMDE0MDQxMDE2MzM0NDUyMjoxNzUzNjkSBTk1MjE2GgpPTkVfU0VBUk

[132] A. E. Pegg, "Recent advances in the biochemistry of polyamines in eukaryotes." *The Biochemical journal*, vol. 234, no. 2, pp. 249–62, Mar. 1986. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1146560&tool=pmcentrez&rendertype=abstract

[133] S. Mizusaki, Y. Tanabe, M. Noguchi, and E. Tamaki, "Phytochemical studies on tobacco alkaloids xiv. the occurrence and properties of putrescine N-methyltransferase in tobacco roots," *Plant and Cell Physiology*, vol. 12, pp. 633–640, 1971.

[134] N. Hibi, S. Higashiguchi, T. Hashimoto, and Y. Yamada, "Gene expression in tobacco low-nicotine mutants." *The Plant cell*, vol. 6, pp. 723–735, 1994. [Online]. Available: http://www.plantcell.org/content/6/5/723.full.pdf

[135] N. Hibi, T. Fujita, M. Hatano, T. Hashimoto, and Y. Yamada, "Putrescine N-methyltransferase in cultured roots of hyoscyamus albus: n-butylamine as a potent inhibitor of the transferase both in vitro and in vivo." *Plant Physiol*, vol. 100and, pp. 826–835, 1992.

[136] O. Stenzel, M. Teuber, and B. Dräger, "Putrescine N-methyltransferase in solanum tuberosum L., a calystegine-forming plant." *Planta*, vol. 223, pp. 200–212, 2006. [Online]. Available: http://download.springer.com/static/pdf/908/art%3A10.1007%2Fs00425-005-0077-z.pdf?auth66=1397320385_f97c066d24812069807d537de2d39695&ext=.pdf

[137] S. Korolev, Y. Ikeguchi, T. Skarina, S. Beasley, C. Arrowsmith, A. Edwards, A. Joachimiak, A. E. Pegg, and A. Savchenko, "The crystal structure of spermidine synthase with a multisubstrate adduct inhibitor." *Nature structural biology*, vol. 9, pp. 27–31, 2002. [Online]. Available: http://www.nature.com/nsmb/journal/v9/n1/pdf/nsb737.pdf

[138] H. Wu, J. Min, Y. Ikeguchi, H. Zeng, A. Dong, P. Loppnau, A. E. Pegg, and A. N. Plotnikov, "Structure and mechanism of spermidine synthases." *Biochemistry*, vol. 46, no. 28, pp. 8331–9, Jul. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17585781

[139] M. Ohnuma, T. Ganbe, Y. Terui, M. Niitsu, T. Sato, N. Tanaka, M. Tamakoshi, K. Samejima, T. Kumasaka, and T. Oshima, "Crystal structures and enzymatic properties of a triamine/agmatine aminopropyltransferase from thermus thermophilus." *Journal of molecular biology*, vol. 408, no. 5, pp. 971–86, May 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21458463

[140] H. Wu, J. Min, H. Zeng, D. E. McCloskey, Y. Ikeguchi, P. Loppnau, A. J. Michael, A. E. Pegg, and A. N. Plotnikov, "Crystal structure of human spermine synthase: implications of substrate binding and catalytic mechanism." *The Journal of biological chemistry*, vol. 283,

no. 23, pp. 16 135–46, Jun. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3259631&tool=pmcentrez&rendertype=abstract

[141] S. Biastoff, W. Brandt, and B. Dräger, "Putrescine N-methyltransferase–the start for alkaloids." *Phytochemistry*, vol. 70, no. 15-16, pp. 1708–18, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19651420

[142] D. E. Riechers and M. P. Timko, "Structure and expression of the gene family encoding putrescine N-methyltransferase in nicotiana tabacum: new clues to the evolutionary origin of cultivated tobacco." *Plant molecular biology*, vol. 41, pp. 387–401, 1999.

[143] L. Holm and J. Park, "DaliLite workbench for protein structure comparison," *Bioinformatics*, vol. 16, no. 6, pp. 566–567, Jun. 2000. [Online]. Available: http://europepmc.org/abstract/MED/10980157/reload=0

[144] Y. Ikeguchi, M. C. Bewley, and A. E. Pegg, "Aminopropyltransferases: function, structure and genetics." *Journal of biochemistry*, vol. 139, no. 1, pp. 1–9, Jan. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16428313

[145] S. Hildreth, "Investigation of protein protein interactions among nicotine biosynthetic enzymes and characterization of a nicotine transporter," Dissertation, Virginia Polytechnic Institute and State University, 2009.

[146] F. W. Studier, "Protein production by auto-induction in high density shaking cultures." *Protein expression and purification*, vol. 41, no. 1, pp. 207–34, May 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15915565

[147] M. M. Bradford, "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding." *Analytical biochemistry*, vol. 72, pp. 248–54, May 1976. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/942051

[148] K. M. Dorgan, W. L. Wooderchak, D. P. Wynn, E. L. Karschner, J. F. Alfaro, Y. Cui, Z. S. Zhou, and J. M. Hevel, "An enzyme-coupled continuous spectrophotometric assay for S-adenosylmethionine-dependent methyltransferases." *Analytical biochemistry*, vol. 350, pp. 249–255, 2006. [Online]. Available: http://ac.els-cdn.com/S0003269706000078/1-s2.0-S0003269706000078-main.pdf?_tid=ca239c52-c0cd-11e3-97d8-00000aab0f01&acdnat=1397147764_7efa7db0c6d97d6a546afd019b0aad61

[149] B. Hosseini, F. Shahriari-Ahmadi, H. Hashemi, M.-H. Marashi, M. Mohseniazar, A. Farokhzad, and M. Sabokbari, "Transient expression of cor gene in papaver somniferum." *BioImpacts : BI*, vol. 1, no. 4, pp. 229–35, Jan. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3648975&tool=pmcentrez&rendertype=abstract

[150] G. Yin, H. Xu, S. Xiao, Y. Qin, Y. Li, Y. Yan, and Y. Hu, "The large soybean (Glycine max) wrky tf family expanded by segmental duplication events and subsequent divergent selection among subgroups." *BMC plant biology*, vol. 13, p. 148, Jan. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3850935&tool=pmcentrez&rendertype=abstract

[151] C. L. Hendricks, J. R. Ross, E. Pichersky, J. P. Noel, and Z. S. Zhou, "An enzyme-coupled colorimetric assay for S-adenosylmethionine-dependent methyltransferases." *Analytical biochemistry*, vol. 326, no. 1, pp. 100–5, Mar. 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14769341

[152] T. Hashimoto, K. Tamaki, K.-i. Suzuki, and Y. Yamada, "Molecular cloning of plant spermidine synthases." *Plant & cell physiology*, vol. 39, pp. 73–79, 1998.

[153] Z. Yang and R. Nielsen, "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages," *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 908–917, Jun. 2002. [Online]. Available: http://mbe.oxfordjournals.org/content/19/6/908.full

[154] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, Aug. 2003. [Online]. Available: http://www.bioinformatics.oupjournals.org/cgi/doi/10.1093/bioinformatics/btg180

[155] E. Pichersky, "_USNomad dna ? A model for movement and duplication of dna sequences in plant genomes," *_USPlant Molecular Biology*, vol. 15, no. 3, pp. 437–448, Sep. 1990. [Online]. Available: http://deepblue.lib.umich.edu/handle/2027.42/43425

[156] A. L. Schilmiller, I. Schauvinhold, M. Larson, R. Xu, A. L. Charbonneau, A. Schmidt, C. Wilkerson, R. L. Last, and E. Pichersky, "Monoterpenes in the glandular trichomes of tomato are synthesized from a neryl diphosphate precursor rather than geranyl

diphosphate." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 26, pp. 10 865–70, Jun. 2009. [Online]. Available: http://www.pnas.org/content/106/26/10865.long

[157] R. K. Ibrahim, A. Bruneau, and B. Bantignies, "Plant O-methyltransferases: molecular analysis, common signature and classification." *Plant molecular biology*, vol. 36, no. 1, pp. 1–10, Jan. 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9484457

[158] D. K. Liscombe and P. J. Facchini, "Molecular cloning and characterization of tetrahydroprotoberberine cis-N-methyltransferase, an enzyme involved in alkaloid biosynthesis in opium poppy." *The Journal of biological chemistry*, vol. 282, no. 20, pp. 14 741–51, May 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17389594

[159] P. J. Facchini and S.-U. Park, "Developmental and inducible accumulation of gene transcripts involved in alkaloid biosynthesis in opium poppy." *Phytochemistry*, vol. 64, no. 1, pp. 177–86, Sep. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12946416

[160] K.-B. Choi, T. Morishige, N. Shitan, K. Yazaki, and F. Sato, "Molecular cloning and characterization of coclaurine N-methyltransferase from cultured cells of coptis japonica." *The Journal of biological chemistry*, vol. 277, no. 1, pp. 830–5, Jan. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11682473

[161] P. M. Frischknecht, J. Ulmer-Dufek, and T. W. Baumann, "Purine alkaloid formation in buds and developing leaflets of coffea arabica: expression of an optimal defence strategy?" *Phytochemistry*, vol. 25, no. 3, pp. 613–616, Jan. 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0031942286880098

[162] R. G. Hollingsworth, J. W. Armstrong, and E. Campbell, "Caffeine as a repellent for slugs and snails." *Nature*, vol. 417, no. 6892, pp. 915–6, Jun. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12087394

[163] H. Uefuji, Y. Tatsumi, M. Morimoto, P. Kaothien-Nakayama, S. Ogita, and H. Sano, "Caffeine production in tobacco plants by simultaneous expression of three coffee N-methyltrasferases and its potential as a pest repellant." *Plant molecular biology*, vol. 59, no. 2, pp. 221–7, Sep. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16247553

[164] Y.-S. Kim, H. Uefuji, S. Ogita, and H. Sano, "Transgenic tobacco plants producing caffeine: a potential new strategy for insect pest control." *Transgenic research*, vol. 15, no. 6, pp. 667–72, Dec. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17091387

[165] J. Friedman and G. R. Waller, "Caffeine hazards and their prevention in germinating seeds of coffee (Coffea arabica L.)." *Journal of chemical ecology*, vol. 9, no. 8, pp. 1099–106, Aug. 1983. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24407803

[166] U. M. Senanayake and R. O. B. Wijesekera, "Theobromine and caffeine content of the cocoa bean during its growth," *Journal of the Science of Food and Agriculture*, vol. 22, no. 5, pp.

262–263, May 1971. [Online]. Available: http://doi.wiley.com/10.1002/jsfa.2740220512

[167] G. Duthie and A. Crozier, "Beverages," in *Plants: Diet and Health*, G. Goldberg, Ed. Oxford, UK: Blackwell Science Ltd, Jul. 2008, ch. 9. [Online]. Available: http://doi.wiley.com/10.1002/9780470774465

[168] P. Nawrot, S. Jordan, J. Eastwood, J. Rotstein, A. Hugenholtz, and M. Feeley, "Effects of caffeine on human health." *Food additives and contaminants*, vol. 20, no. 1, pp. 1–30, Jan. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12519715

[169] K. Mizuno, M. Kato, F. Irino, N. Yoneyama, T. Fujimura, and H. Ashihara, "The first committed step reaction of caffeine biosynthesis: 7-methylxanthosine synthase is closely homologous to caffeine synthases in coffee (Coffea arabica L.)." *FEBS letters*, vol. 547, no. 1-3, pp. 56–60, Jul. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12860386

[170] H. Uefuji, S. Ogita, Y. Yamaguchi, N. Koizumi, and H. Sano, "Molecular cloning and functional characterization of three distinct N-methyltransferases involved in the caffeine biosynthetic pathway in coffee plants." *Plant physiology*, vol. 132, no. 1, pp. 372–80, May 2003. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=166982&tool=pmcentrez&rendertype=abstract

[171] O. Negishi, T. Ozawa, and H. Imagawa, "N-methylnucleosidase from tea leaves," *Agric. Biol. Chem.*, vol. 52, pp. 169–175, 1988.

[172] K. Mizuno, M. Kato, H. Ashihara, and T. Fujimora, "cDNA cloning of caffeine (theobromine) synthase from coffee (Coffea arabica L.)," *International scientific colloquium on coffee 19 ASIC, Paris*, vol. 19, pp. 815–818, 2001.

[173] M. Ogawa, Y. Herai, N. Koizumi, T. Kusano, and H. Sano, "7-methylxanthine methyltransferase of coffee plants. gene isolation and enzymatic properties." *The Journal of biological chemistry*, vol. 276, no. 11, pp. 8213–8, Mar. 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11108716

[174] A. A. McCarthy and J. G. McCarthy, "The structure of two N-methyltransferases from the caffeine biosynthetic pathway." *Plant physiology*, vol. 144, no. 2, pp. 879–89, Jun. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1914188&tool=pmcentrez&rendertype=abstract

[175] M. Kato, K. Mizuno, T. Fujimura, M. Iwama, M. Irie, A. Crozier, and H. Ashihara, "Purification and characterization of caffeine synthase from tea leaves." *Plant physiology*, vol. 120, no. 2, pp. 579–86, Jun. 1999. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=59297&tool=pmcentrez&rendertype=abstract

[176] M. Kato, K. Mizuno, A. Crozier, T. Fujimura, and H. Ashihara, "Caffeine synthase gene from tea leaves." *Nature*, vol. 406, no. 6799, pp. 956–7, Aug. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10984041

[177] N. Yoneyama, H. Morimoto, C.-X. Ye, H. Ashihara, K. Mizuno, and M. Kato, "Substrate specificity of N-methyltransferase involved in purine alkaloids synthesis is

dependent upon one amino acid residue of the enzyme." *Molecular genetics and genomics : MGG*, vol. 275, no. 2, pp. 125–35, Feb. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16333668

[178] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences." *Journal of molecular biology*, vol. 147, no. 1, pp. 195–7, Mar. 1981. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7265238

[179] W. R. Pearson, "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the smith-waterman and fasta algorithms." *Genomics*, vol. 11, no. 3, pp. 635–50, Nov. 1991. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1774068

[180] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney, "The bioperl toolkit: perl modules for the life sciences." *Genome research*, vol. 12, no. 10, pp. 1611–8, Oct. 2002. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=187536&tool=pmcentrez&rendertype=abstract

[181] K. Katoh and H. Toh, "Recent developments in the mafft multiple sequence alignment program." *Briefings in bioinformatics*, vol. 9, no. 4, pp. 286–98, Jul. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18372315

[182] H. Shimodaira, "An approximately unbiased test of phylogenetic tree selection," *Syst. Biol.*, vol. 51, pp. 492–508, 2002.

[183] H. Shimodaira and M. Hasegawa, "CONSEL: for assessing the confidence of phylogenetic tree selection," *Bioinformatics*, vol. 17, pp. 1246–1247, 2001.

[184] W. R. Pearson, "Effective protein sequence comparison," *Methods in Enzymology*, vol. 266, pp. 227–258, 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8743688

[185] ——, "Protein sequence comparison and protein evolution tutorial," in *International Symposium of Molecular Biology*, San Diego, 2000.

[186] K. Kiple and O. K, "Kola nut," in *the Cambridge World History of Food*. New York, NY: Cambridge University Press, 2000, pp. 684–692.

[187] M. D. Saldaña, R. S. Mohamed, M. G. Baer, and P. Mazzafera, "Extraction of purine alkaloids from maté (Ilex paraguariensis) using supercritical co(2)." *Journal of agricultural and food chemistry*, vol. 47, no. 9, pp. 3804–8, Sep. 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10552725

[188] D. K. Bempong, P. J. Houghton, and K. Steadman, "The xanthine content of guarana and its preparations," *Pharmaceutical Biology*, vol. 31, no. 3, pp. 175–181, Jan. 1993. [Online]. Available: http://informahealthcare.com/doi/abs/10.3109/13880209309082937

[189] Z.-L. Nie, H. Sun, Y. Meng, and J. Wen, "Phylogenetic analysis of toxicodendron (Anacardiaceae) and its biogeographic implications on the evolution of north temperate and tropical intercontinental disjunctions," *Journal of Systematics and Evolution*, vol. 47, no. 5, pp. 416–430, Sep. 2009. [Online]. Available: http://doi.wiley.com/10.1111/j.1759-6831.2009.00045.x

[190] C. Wu and T. Ming, "Toxicodendron hookeri (Sahni & bahadur) C. Y. wu & T. L. ming var. microcarpum (C. C. huang ex T. L. ming)," *Fl. Reipubl. Popularis Sin.*, vol. 45, no. 1, p. 110, 1980.

[191] W. T. Gillis, *the Systematics and Ecology of Poison-ivy and the Poison-oaks (Toxicodendron, Anacardiaceae)*. Michigan State University. Department of Botany and Plant Pathology, 1969. [Online]. Available: http://books.google.com/books/about/The_Systematics_and_Ecology_of_Poison_iv.html?id=AyFFAAAAYAAJ&pgis=1

[192] W. L. Epstein, "Plant-induced dermatitis." *Annals of emergency medicine*, vol. 16, no. 9, pp. 950–5, Sep. 1987. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3307555

[193] K. D. Gayer and J. W. Burnett, "Toxicodendron dermatitis." *Cutis*, vol. 42, no. 2, pp. 99–100, Aug. 1988. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/2970948

[194] J. Khittel, "Analysis of the leaves of poison-oak (Rhus toxicodendron)," *Am. J. Pharm*, p. 542, 1858.

[195] F. Pfaff, "On the active principle of rhus toxicodendron and rhus venenata," *The Journal of experimental medicine*, pp. 348–359, 1897.

[196] G. A. Hill, V. Mattacotti, and W. D. Graham, "The toxic principle of the poison ivy," *Journal of the American Chemical Society*, vol. 56, no. 12, pp. 2736–2738, Dec. 1934. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ja01327a064

[197] R. Majima, "Über den hauptbestandteil des japan-lacks, viii. mitteilung: stellung der doppelbindungen in der seitenkette des urushiols und beweisführung, daßdas urushiol eine mischung ist," *Berichte der deutschen chemischen Gesellschaft (A and B Series)*, vol. 55, no. 1, pp. 172–191, Jan. 1922. [Online]. Available: http://doi.wiley.com/10.1002/cber.19220550123

[198] D. Pariser, R. Ceilley, A. Lefkovits, B. Katz, and A. Paller, "Poison ivy, oak and sumac," *Derm. Insights*, vol. 4, no. 1, pp. 26–28, 2003.

[199] M. A. Davies, "Outsmarting poison ivy and its relatives," USDA United States Forest Service, Tech. Rep., 2007. [Online]. Available: http://www.fs.fed.us/t-d/pubs/htmlpubs/htm07672313/

[200] J. E. Mohan, L. H. Ziska, W. H. Schlesinger, R. B. Thomas, R. C. Sicher, K. George, and J. S. Clark, "Biomass and toxicity responses of poison ivy (Toxicodendron radicans) to elevated atmospheric co2." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 24, pp. 9086–9, Jun. 2006. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1474014&tool=pmcentrez&rendertype=abstract

[201] L. H. Ziska, R. C. Sicher, K. George, and J. E. Mohan, "Rising atmospheric carbon dioxide and potential impacts on the growth and toxicity of poison ivy (Toxicodendron radicans)," *Weed Science*, vol. 55, pp. 288–292, 2007.

[202] R. Majima and J. Tahara, "Über den hauptbestandteil des japanlacks. vi. mitteilung: Über die synthese des hydro-urushiols," *Berichte der deutschen chemischen Gesellschaft*, vol. 48, no. 2, pp. 1606–1611, Jul. 1915. [Online]. Available: http://doi.wiley.com/10.1002/cber.19150480253

[203] J. B. McNair, "LobinolA dermatitant from rhus diversiloba (Poison oak)." *Journal of the American Chemical Society*, vol. 43, no. 1, pp. 159–164, Jan. 1921. [Online]. Available: http://dx.doi.org/10.1021/ja01434a021

[204] W. F. Symes and C. R. Dawson, "Poison ivy urushiol," *Journal of the American Chemical* . . . , vol. 1, pp. 2959–2963, 1954.

[205] M. D. Corbett and S. Billets, "Characterization of poison oak urushiol." *Journal of pharmaceutical sciences*, vol. 64, no. 10, pp. 1715–8, Oct. 1975. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1185545

[206] M. A. ElSohly, P. D. Adawadkar, C.-Y. Ma, and C. E. Turner, "Separation and characterization of poison ivy and poison oak urushiol components," *Journal of Natural Products*, vol. 45, no. 5, pp. 532–538, Sep. 1982. [Online]. Available: http://dx.doi.org/10.1021/np50023a004

[207] T. McGovern and T. Barkley, "Review botanical dermatology," *International Journal of Dermatology*, vol. 37, pp. 321–334, 1998.

[208] M. a. ElSohly, P. D. Adawadkar, D. a. Benigni, E. S. Watson, and T. L. Little, "Analogues of poison ivy urushiol. synthesis and biological activity of disubstituted n-alkylbenzenes." *Journal of medicinal chemistry*, vol. 29, pp. 606–611, 1986.

[209] S. Billets, J. Craig Jr, M. Corbett, and J. Vickery, "Component analysis of the urushiol content of poison ivy and poison oak," *Phytochemistry*, vol. 15, no. 4, pp. 533–535, 1976.

[210] J. C. Craig, C. W. Waller, S. Billets, and M. A. Elsohly, "New glc analysis of urushiol congeners in different plant parts of poison ivy, toxicodendron radicans." *Journal of pharmaceutical sciences*, vol. 67, no. 4, pp. 483–5, Apr. 1978. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/641754

[211] H. Baer, M. Hooton, H. Fales, A. Wu, and F. Schaub, "Catecholic and other constituents of the leaves of toxicodendron radicans and variation of urushiol concentrations within one plant," *Phytochemistry*, vol. 19, pp. 799–802, 1980.

[212] F. K. Tadjimukhamedov, G. Huang, Z. Ouyang, and R. G. Cooks, "Rapid detection of urushiol allergens of toxicodendron genus using leaf spray mass spectrometry." *The Analyst*, vol. 137, no. 5, pp. 1082–4, Mar. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22249516

[213] R. Braslau, F. Rivera, E. Lilie, and M. Cottman, "Urushiol detection using a profluorescent nitroxide." *The Journal of organic chemistry*, vol. 78, no. 2, pp. 238–45, Jan. 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22900824

[214] R. Penner, G. E. E. Moodie, and R. J. Staniforth, "The dispersal of fruits and seeds of poison-ivy, toxicodendron radicans, by ruffed grouse, bonasa umbellus, and squirrels, tamiasciurus hudsonicus and sciurus carolinensis," *Canadian Field-Naturalist*, vol. 113, pp. 616–620, 1999.

[215] I. Popay and R. Field, "Grazing animals as weed control agents," *Weed Technology*, vol. 10, pp. 217–231, 1996.

[216] E. B. Benhase and J. G. Jelesko, "Germinating and culturing axenic poison ivy seedlings," *HortScience*, vol. 48, no. 12, pp. 1525–1529, Dec. 2013. [Online]. Available: http://hortsci.ashspublications.org/content/48/12/1525.abstract?related-urls=yes&legid=hortsci;48/12/1525

[217] P. M. Dewick, *Medicinal Natural Products: A Biosynthetic Approach*. West Sussex, England: John Wiley & Sons Ltd, 1997.

[218] I. Abe, T. Watanabe, and H. Noguchi, "Enzymatic formation of long-chain polyketide pyrones by plant type iii polyketide synthases." *Phytochemistry*, vol. 65, no. 17, pp. 2447–53, Sep. 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031942204003796

[219] P. Saxena, G. Yadav, D. Mohanty, and R. S. Gokhale, "A new family of type iii polyketide synthases in mycobacterium tuberculosis." *The Journal of biological chemistry*, vol. 278, no. 45, pp. 44 780–90, Nov. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12941968

[220] F. Taura, S. Tanaka, C. Taguchi, T. Fukamizu, H. Tanaka, Y. Shoyama, and S. Morimoto, "Characterization of olivetol synthase, a polyketide synthase putatively involved in cannabinoid biosynthetic pathway." *FEBS letters*, vol. 583, no. 12, pp. 2061–6, Jun. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19454282

[221] M. Matsuzawa, Y. Katsuyama, N. Funa, and S. Horinouchi, "Alkylresorcylic acid synthesis by type iii polyketide synthases from rice oryza sativa." *Phytochemistry*, vol. 71, no. 10, pp. 1059–67, Jul. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20451227

[222] S. Y. Kim, C. C. Colpitts, G. Wiedemann, C. Jepson, M. Rahimi, J. R. Rothwell, A. D. McInnes, M. Hasebe, R. Reski, B. T. Sterenberg, and D.-Y. Suh, "Physcomitrella ppors, basal to plant type iii polyketide synthases in phylogenetic trees, is a very long chain 2'-oxoalkylresorcinol synthase." *The Journal of biological chemistry*, vol. 288, no. 4, pp. 2767–77, Jan. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3554942&tool=pmcentrez&rendertype=abstract

[223] T. Murashige and F. Skoog, "A revised medium for rapid growth and bio assays with tobacco tissue cultures," *Physiologia Plantarum*, vol. 15, no. 3, pp. 473–497, Jul. 1962. [Online]. Available: http://doi.wiley.com/10.1111/j.1399-3054.1962.tb08052.x

[224] F. Asubel, R. Brent, R. Kingston, D. Moore, J. Seidman, J. Smith, and K. Struhl, "Phenol/SDS method for plant rna preparation." *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, vol. Chapter 4, p. Unit4.3, May 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18265237

[225] S. Andrews, "FastQC A quality control tool for high throughput sequence data," 2010. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[226] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for illumina sequence data," *Bioinformatics*, Apr. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24695404

[227] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from rna-seq data without a reference genome." *Nature biotechnology*, vol. 29, no. 7, pp. 644–52, Jul. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3571712&tool=pmcentrez&rendertype=abstract

[228] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2." *Nature methods*, vol. 9, no. 4, pp. 357–9, Apr. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3322381&tool=pmcentrez&rendertype=abstract

[229] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching." *Nucleic acids research*, vol. 39, no. Web Server issue, pp. W29–37, Jul. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125773&tool=pmcentrez&rendertype=abstract

[230] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, "The pfam protein families database." *Nucleic acids research*, vol. 40, no. Database issue, pp. D290–301, Jan. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245129&tool=pmcentrez&rendertype=abstract

[231] T. N. Petersen, S. r. Brunak, G. von Heijne, and H. Nielsen, "SignalP 4.0: discriminating signal peptides from transmembrane regions." *Nature methods*, vol. 8, no. 10, pp. 785–6, Jan. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21959131

[232] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes." *Journal of molecular biology*, vol. 305, no. 3, pp. 567–80, Jan. 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11152613

[233] K. Lagesen, P. Hallin, E. A. Rø dland, H.-H. Staerfeldt, T. r. Rognes, and D. W. Ussery, "RNAmmer: consistent and rapid annotation of ribosomal rna genes." *Nucleic acids research*, vol. 35, no. 9, pp. 3100–8, Jan. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1888812&tool=pmcentrez&rendertype=abstract

[234] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research."

Bioinformatics (Oxford, England), vol. 21, no. 18, pp. 3674–6, Sep. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16081474

[235] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from rna-seq data with or without a reference genome." *BMC bioinformatics*, vol. 12, p. 323, Jan. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3163565&tool=pmcentrez&rendertype=abstract

[236] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics (Oxford, England)*, vol. 26, no. 1, pp. 139–40, Jan. 2010. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2796818&tool=pmcentrez&rendertype=abstract

[237] L. Li, C. J. Stoeckert, and D. S. Roos, "OrthoMCL: identification of ortholog groups for eukaryotic genomes." *Genome research*, vol. 13, no. 9, pp. 2178–89, Sep. 2003. [Online]. Available: http://genome.cshlp.org/content/13/9/2178.long

[238] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Science*, vol. 85, no. 8, pp. 2444–2448, 1988. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3162770

[239] Bethesda Research Laboratories, "BRL puc host: E. coli dh5α competent cells," *Focus*, vol. 8, no. 2, p. 9, 1986.

[240] B. Miroux and J. E. Walker, "Over-production of proteins in escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels."

Journal of molecular biology, vol. 260, no. 3, pp. 289–98, Jul. 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8757792

[241] Y. Liu, M. Schiff, and S. P. Dinesh-Kumar, "Virus-induced gene silencing in tomato," *The Plant Journal*, vol. 31, pp. 777–786, 2002. [Online]. Available: http://onlinelibrary.wiley.com/store/10.1046/j.1365-313X.2002.01394.x/asset/j.1365-313X.2002.01394.x.pdf?v=1&t=htuarnyj&s=625727f71f210b6345988658fcc4e4caea0d5cdc

[242] J. G. Jelesko, R. Harper, M. Furuya, and W. Gruissem, "Rare germinal unequal crossing-over leading to recombinant gene formation and gene duplication in arabidopsis thaliana." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 18, pp. 10 302–7, Aug. 1999. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=17883&tool=pmcentrez&rendertype=abstract

[243] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman, and A. Regev, "De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis." *Nature protocols*, vol. 8, no. 8, pp. 1494–512, Aug. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3875132&tool=pmcentrez&rendertype=abstract

[244] M. K. Azim, I. a. Khan, and Y. Zhang, "Characterization of mango (Mangifera indica L.) transcriptome and chloroplast genome." *Plant molecular biology*, Feb. 2014. [Online].

Available: http://www.ncbi.nlm.nih.gov/pubmed/24515595

[245] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families." *Nucleic acids research*, vol. 30, no. 7, pp. 1575–84, Apr. 2002. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101833&tool=pmcentrez&rendertype=abstract

[246] S. J. Gagne, J. M. Stout, E. Liu, Z. Boubakir, S. M. Clark, and J. E. Page, "Identification of olivetolic acid cyclase from cannabis sativa reveals a unique catalytic route to plant polyketides." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 31, pp. 12 811–6, Jul. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3411943&tool=pmcentrez&rendertype=abstract

[247] H. Lulin, Y. Xiao, S. Pei, T. Wen, and H. Shangqin, "The first illumina-based de novo transcriptome sequencing and analysis of safflower flowers." *PloS one*, vol. 7, p. e38653, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3378585/pdf/pone.0038653.pdf

[248] A. Ranjan, Y. Ichihashi, M. Farhi, K. Zumstein, B. Townsley, R. David-Schwartz, and N. R. Sinha, "De novo assembly and characterization of the transcriptome of the parasitic weed cuscuta pentagona identifies genes associated with plant parasitism." *Plant physiology*, Jan. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24399359

[249] J. Zhang, T. a. Ruhlman, J. P. Mower, and R. K. Jansen, "Comparative analyses of two geraniaceae transcriptomes using next-generation sequencing." *BMC plant biology*, vol. 13, p. 228, Jan. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3880972&tool=pmcentrez&rendertype=abstract

[250] S. Kalra, B. L. Puniya, D. Kulshreshtha, S. Kumar, J. Kaur, S. Ramachandran, and K. Singh, "De novo transcriptome sequencing reveals important molecular networks and metabolic pathways of the plant, chlorophytum borivilianum." *PloS one*, vol. 8, no. 12, p. e83336, Jan. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3871651&tool=pmcentrez&rendertype=abstract

[251] S. Liu, W. Li, Y. Wu, C. Chen, and J. Lei, "De novo transcriptome assembly in chili pepper (Capsicum frutescens) to identify genes involved in the biosynthesis of capsaicinoids." *PloS one*, vol. 8, no. 1, p. e48156, Jan. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3551913&tool=pmcentrez&rendertype=abstract

[252] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney, "Oases: robust de novo rna-seq assembly across the dynamic range of expression levels." *Bioinformatics (Oxford, England)*, vol. 28, no. 8, pp. 1086–92, Apr. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3324515&tool=pmcentrez&rendertype=abstract

[253] Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T.-W. Lam, Y. Li, X. Xu, G. K.-S. Wong, and J. Wang, "SOAPdenovo-trans: de

novo transcriptome assembly with short rna-seq reads." *Bioinformatics (Oxford, England)*, Mar. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24532719

[254] M. Kasson, J. R. Pollok, E. B. Benhase, and J. G. Jelesko, "First report of seedling blight of eastern poison ivy (Toxicodendron radicans) by colletotrichum fioriniae in virginia," *Plant Disease "First Look"*, vol. Accepted f, 2014.

[255] J. Marcelino, R. Giordano, S. Gouli, V. Gouli, B. L. Parker, M. Skinner, D. TeBeest, and R. Cesnik, "Colletotrichum acutatum var. fioriniae (teleomorph: glomerella acutata var. fioriniae var. nov.) infection of a scale insect," *Mycologia*, vol. 100, no. 3, pp. 353–374, May 2008. [Online]. Available: http://www.mycologia.org/cgi/doi/10.3852/07-174R

[256] H. Solereder and D. H. Scott, *Introduction. Polypetalae. Gamopetalae*. Clarendon Press, 1908. [Online]. Available: http://books.google.com/books?id=VagUAAAAYAAJ&pgis=1