PHYLOGENOMIC ANALYSIS OF GENE FAMILIES: PHENYLALANINE AMMONIA LYASES IN GYMNOSPERMS AS A CASE STUDY

by

UJWAL RANJIT BAGAL

(Under the Direction of JEFFREY F. D. DEAN)

ABSTRACT

The field of high throughput sequencing has advanced at a tremendous pace in the last few years. This has opened up opportunities to understand more about the functional genomics of non-model species for which genome sequences are not yet available. In such situations, transcriptome sequence analysis using comparative methods has facilitated gene discovery and gene expression studies, as well as new understanding of the functional responsibilities of gene family members, the effect of gene duplications, how Darwinian selection affects genome complexities. These approaches have opened unprecedented opportunities to understand functional compositions, as well as overrepresentation / underrepresentation of mRNAs involved in specific biological functions.

This dissertation, applies computational approaches and experimental verification to the reevaluation of an earlier report of a single phenylalanine ammonia lyase (PAL) gene in *Pinus taeda*. This work is followed by a biological analysis of the PAL gene family members in gymnosperms with an eye toward determining their individual evolutionary trajectories and functional variability. The five *P. taeda* PAL genes revealed diverse evolutionary path for

gymnosperms compared to angiosperms starting from a series of ancient gene duplication events. This hypothesis was further supported by identification of codon sites under relaxed evolutionary constraints in lineages associated with duplication events. While gene expression analyses proved insufficient to identify physiological functions of individual genes, it highlighted tissue-specific expression and provided some insight into functional associations of individual PAL genes with biotic / abiotic stress conditions.

A relative efficiency analysis of the statistical models used to infer changes in the mode of selection acting on protein coding genes was performed using simulated data sets. Despite the advantage of having more realistic models, the likelihood ratio test (LRT) in Fitmodel was unable to detect shifts in selection pressure. Similarly, the Bayesian approach used to detect individual sites under adaptive selection yielded both high false-negative and false-positive rates.

The findings from P. *taeda* PAL gene family analysis will be useful for future pine tree improvement programs, while, the simulation based studies is expected to provide cautionary advice to researchers about the unreliability of the inferences estimated by the evolutionary tools.

INDEX WORDS: loblolly pine, gymnosperms, phenylalanine ammonia lyase, evolution, duplication, positive selection, power, maximum likelihood, accuracy, false positive rate, false negative rate,

PHYLOGENOMIC ANALYSIS OF GENE FAMILIES: PHENYLALANINE AMMONIA LYASES IN GYMNOSPERMS AS A CASE STUDY

by

UJWAL RANJIT BAGAL

B.S, UNIVERSITY OF PUNE, INDIA 1994

M.S, UNIVERSITY OF PUNE, INDIA 1996

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

PHYLOGENOMIC ANALYSIS OF GENE FAMILIES: PHENYLALANINE AMMONIA LYASES IN GYMNOSPERMS AS A CASE STUDY

by

UJWAL RANJIT BAGAL

Major Professor: Jeffrey F. D. Dean

Committee: James Leebens-Mack
Jan Mrázek
Paul Schliekelman

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

DEDICATION

To my beloved family members and friends

ACKNOWLEDGEMENTS

"Let yourself be silently drawn by the strange pull of what you really love. It will not lead you astray."

-Rumi, Essential Rumi

A few years back, I followed my heart and made the decision to pursue higher studies and joined the doctoral program at UGA. Today, as I look back, I am glad with my decision, as these graduation years have allowed me to continue evolving for better. Although this journey was my chosen path, I dare not say it had no challenges and excitement. But help awaited me at every nook and corner and hence I call it "my path with a heart".

I am more than thankful to have a mentor who patiently guided me and gave me all the freedom to do my research which led to this dissertation. Dr. Dean, your support and trust in me while I was trying to overcome the hurdles in my research has helped me immensely. It would have been an impossible journey, for me, without your support and encouragement. I am amongst the few fortunate students who will always cherish having a mentor who supported me during testing times.

I would equally like to thank my committee members: Jim Leebens-Mack, Jan Mrázek and Paul Schliekelmann for providing patient guidance and insightful comments. Dr. Leebens-Mack thanks for investing time for guiding me on my projects. Dr. Travis Glenn, I am thankful for financially supporting me during my last few years at UGA and for allowing me to work on a number of interesting projects.

John Michael Bordeaux, many thanks for being a supportive lab-mate and a sincere friend. I will always cherish the times when we spent in discussions innumerable subjects. Walt, I cannot forget to acknowledge my gratitude for helping me with your scripts and suggestions while working on NGS analyses.

I am profoundly indebted to all my friends who have helped me in times of great stress and adversity. My special thanks to Chandana, Arpan, Priyanka, Sanjeev, Biswajeet, Sivshankari, Vijay, Sindhuri, Dipesh, Anuj Sinha, Kamal, Albina and Jimmy without whom life in Athens would be dull and gloomy.

Lastly, I would like to thank my beloved family: my mother Padmaja Bagal (aai) and my siblings Shivanjalee and Jaideep. Aai, you sacrificed your life for us and spared no efforts to help us achieve our goals. Shiu and Abhi, your support comforted me all the time. I hope, I have made you all proud. My immeasurable thanks to Mrs. Meenakshi Vaidya (kaku) and Dr. Nandini Savant (ajji) for their encouraging support which gave me the strength to sustain through difficult times. My special thanks to Pradeep mama, Almitra, Rucha, Suchu, Shama didi, Nitu didi for your comforting phone calls.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER

1 INTRODUCTION AND LITERATURE REVIEW 1

1.1 Introduction to Phylogenomics 1

1.2 Phylogenetic Markers 4

1.3 Sequence Alignment 6

1.4 Methods for Reconstructing Phylogenies 9

1.5 Post Alignment Applications using Phylogenetic Frameworks 17

1.6 Objectives 25

1.7 References 27

2 PHENYLALANINE AMMONIA LYASE (PAL) GENE FAMILY SHOWS A GYMNOSPERM SPECIFIC LINEAGE 36

2.1 Abstract 37

2.2 Introduction 39

2.3 Results 41

2.4 Discussion 46

2.5 Conclusion 48

2.6 Materials and Methods....49
2.7 List of Abbreviations....53
2.8 Competing Interests....53
2.9 Authors Contributions....53
2.10 Acknowledgements....53
2.11 References....54
2.12 Supplementary Material....57
3 TISSUE SPECIFIC AND STRESS-INDUCED EXPRESSION OF PHENYLALANINE AMMONIA LYASE GENE FAMILY MEMBERS IN *PINUS TAEDA*....63
3.1 Abstract....64
3.2 Introduction....65
3.3 Materials and Methods....67
3.4 Results....71
3.5 Discussion....75
3.6 Conclusion....78
3.7 Funding....79
3.8 Acknowledgement....79
3.9 References....79
4 EVOLUTIONARY ANALYSIS OF PHENYLALANINE AMMONIA LYASE GENE FAMILY IN GYMNOSPERMS....96
4.1 Abstract....97
4.2 Introduction....98

4.3 Materials and Methods....101
4.4 Results....103
4.5 Discussion....107
4.6 Conclusion....109
4.7 References....109
5 ESTIMATING RELATIVE POWER AND ACCURACY OF MOLECULAR EVOLUTIONARY ANALYSIS OF PROTEIN CODING SEQUENCES....116
5.1 Abstract....117
5.2 Introduction....118
5.3 Materials and Methods....122
5.4 Results....126
5.5 Discussion....131
5.6 References....135
6 CONCLUSIONS....152
6.1 Phenylalanine ammonia lyase gene family....152
6.2 Efficiency of evolutionary models....156
6.3 References....157

LIST OF TABLES

Page

Table 2.1: PtPAL (1-5) *de novo* transcriptome assemblies of *P. taeda* 58

Table 2.2: *P. taeda* PAL inferred amino acid sequence percent identity / similarity 59

Table 3.1: Potential regulatory elements in the 5' flanking regions of *P. taeda* PAL gene family members 87

Table 3.2: Primer sequences used for quantitative RT-PCR 88

Table 4.1: Likelihood analysis of PAL gene sequence data 112

Table 4.2: Likelihood ratio test (LTR) results for different model comparisons 112

Table 5.1: Simulation Schemes 138

Table 5.2: Definition of accuracy, power, false positive rate and false negative rate 139

Table 5.3: Percentage of significant tests of Branch-X-site and switching models at 5% level when data simulated with positive selection 140

Table 5.4: Performance of Codeml in inferring sites under positive selection in simulated data set 142

Table 5.5: Performance of Fitmodel in inferring sites under positive selection in simulated data set 142

LIST OF FIGURES

Page

Figure 2.1: Alignment between the five PtPAL genes in *P. taeda* 60

Figure 2.2: Consensus tree of the Phenylalanine ammonia lyase gene family 61

Figure 2.3: NOTUNG: Reconciled gene tree 62

Figure 3.1: Alignment of the PtPAL1 and PtPAL4 promoters and localization of potential *cis*-regulatory elements 89

Figure 3.2: Alignment of the PtPAL4 and PtPAL5 promoters and localization of potential *cis*-regulatory elements 90

Figure 3.3: Expression of PAL family members in various tissues from a mature *P. taeda* tree..91

Figure 3.4: Expression of PAL family members in needle (C-N) and stem (C-S) tissues from *P. taeda* shoot tip explants under control conditions 93

Figure 3.5: Expression of PAL family members in needle (D-N) and stem (D-S) tissues from *P. taeda* shoot tip explants after abiotic stress (drought) treatment 94

Figure 3.6: Expression of PAL family members in needle (V-N) and stem (V-S) tissues from *P. taeda* shoot tip explants after biotic stress (venom) treatment 95

Figure 4.1: Selection regime patterns across sites on major branches of the PAL gene tree 113

Figure 4.2: Mapping of potential sites under adaptive selection on a PAL 3D structure 115

Figure 5.1: Eight taxa tree topologies used in the simulations 143

Figure 5.2: LRT: Number of significant tests in Codeml and Fitmodel 145

Figure 5.3: Codeml: Accuracy and Power of BEB prediction method.....148

Figure 5.4: Average false positive rate and false negative rates for trees T1, T2 and T3 in Codeml and Fitmodel 149

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

"Evolution is, so to speak, an experimentalist that has been running experiments for three and a half billion years, since the origin of life on Earth. And, wonderfully, the genomes of today's organisms retain the lab notes of these experiments, so that we can go back and reconstruct the events that took place."

- Eric Lander
Genomicist and founding director of the Broad Institute of MIT and Harvard ("Genetics" In: Biology, 7th Edition 2005)

1.1 Introduction to Phylogenomics

To help explain his theory of evolution, Charles Darwin, in his book *The Origin of Species* in 1849 used a tree structure to illustrate the idea that all species descended from a common ancestor, where modern species were placed at the leafy ends and ancestral species formed the branches and the trunk of the tree. Although tree-like graphic illustrations of organismal relationships predated Darwin (Hitchcock, 1840), the popularity of phylogeny grew with DNA established as the blue print of each organism's evolutionary inheritance (Zuckerkandl and Pauling, 1965). Such a study was meant to decode the information retained in genomes so as to trace the evolutionary history of organisms.

With the advent of next generation sequencing technology and the consequent high volume of data made available, the field of phylogenetics has rapidly evolved in many new directions and has become an indispensable component of genomic studies. Phylogenetics has nearly gained the status of a distinct interdisciplinary specialty where evolutionary biologists

team with statisticians and computer scientists to develop models and algorithms to answer both evolutionary questions of gene origin and function. These algorithms capture phylogenetic signals from a multiple sequence alignment of homologous sequences. This makes alignment the foremost and most important starting step. However, the type of sequences to align depends on the objective of the study; for example, whether the purpose is to determine the population history of a single species, or the evolutionary and epidemiological dynamics of a pathogen, or the reconstruction of an ancestral genome from modern genome sequences (Rausch and Reinert, 2011). Various types of sequences, ranging from non-coding segments in genomes to the protein-coding sequences, have been used to reconstruct genetic phylogenies. Understanding the purpose and the issues for using coding or non-coding regions within a genome in phylogeny reconstruction are equally important. For example, the nuclear ribosomal spacers, such as the ITS and the ETS regions as well as 26S RNA coding region, are frequently used because of their high copy number in plant genomes (Baldwin, et al., 1995). But, chloroplast DNA markers, due to their uni-parental inheritance, are not suitable for reconstructing phylogenies associated with understanding relationships between lineages involved in hybridization and allopolyploidization (McCauley, et al., 2007). On the other hand, low copy gene markers, owing to their rapid rate of evolutionary change, are more suitable for phylogenetic purposes (Glenn, 2011).

Modern phylogenetics may be considered to have started with the gene tree concept, where a phylogeny of functional gene data from closely related species provides a portrait of the genealogy of a single gene locus. With more transcriptome data made available, a greater understanding of the uncertainties in gene trees caused by gene duplication, horizontal gene transfer, or coalescence became clearer (Maddison, 1997). The coalescence theory proposed by Kingman (Kingman, 2000) is dependent on tree-based thinking in which ancestral lineages are

tracked back in time to trace most recent ancestors. As more genomes are sequenced, such phylogenetic studies are moving from single gene tree approaches towards whole genome approaches in which all available information is used to identify gene contents, gene loss events, and reconstruct gene movement in the genome (Sleator, 2011). Further refinement of these gene trees has enabled development of numerous codon-based statistical models to trace back the footprints of adaptive protein evolution (Goldman and Yang, 1994; Guindon, 2004; Rodrigue, et al., 2010; Pond, et al., 2011).

Apart from gene trees, another important task is to depict evolutionary relationships between organisms through construction of species trees. In a species tree approach, using gene trees as statistical quantities helps in integrating signals from the gene trees and also copes with their heterogeneity and the coalescence problem (Edwards, 2009). As DNA sequencing information becomes increasingly abundant, a key task for systematicists will be to develop methods that fully utilize this data for refining species trees. With complete genomes available, a deeper understanding of gene structure and function will be possible. This in turn, is expected to help in resolve relationships at the species level for reconstructing the history of life, which is the ultimate purpose of phylogenetic analysis.

Similarly, new emerging field such as phylo-geography have evolved by using mathematical models that reflect biological phenomenon. In this case, coalescence theory is used to provide statistical frameworks for estimating demographic parameters (Knowles and Maddison, 2009). In phylo-geography theory, variations in gene copy numbers have been used to identify rapidly evolving phylogenetic markers (Zimmer and Wen, 2012).

In the following section, I will recount a few of the many paths that can be taken in developing phylogenies after sequence alignment under various phylogenetic frameworks, and will highlight advantages and pitfalls that may be expected with each approach.

1.2 Phylogenetic Markers

Since a phylogeny represents a "tree of heredity," only genetically transmitted traits should be considered informative. In general, molecular markers of heredity can be categorized as coding sequences from the protein coding region of genes and the non-coding sequences, which includes intergenic regions, such as introns and untranslated regions (UTRs) of the mRNA transcripts, as well as the internal transcribed spacer (ITS) and the external transcribed spacer (ETS) regions of ribosomal RNA and simple sequence repeats (SSRs), also called microsatellites (Bowcock, et al., 1994; Chen, et al., 2002).

The different molecular marker types have specific characteristics that can impact phylogeny reconstruction and must be taken into consideration (Koch, et al., 2003; Volkov, et al., 2007). For example, in case of non-coding markers, because ITS regions are inherited from both parents in eukaryotes, they can be used to address questions concerning reticulate evolution, hybridization events, and parentage of polyploids, as well as the general evolutionary history of a species or lineage (Baldwin, 1992; Kim and Jansen, 1994; Baldwin, et al., 1995; Álvarez and Wendel, 2003). Simple sequence repeats (SSRs) are frequently used to unravel patterns of relatedness within populations (Zhu, et al., 2000; Zimmer and Wen, 2012).

Before the advent of NGS technology and low-cost DNA sequencing, non-coding markers, especially the ITS region which has high copy number and universal presence in nuclear DNA, were frequently used for phylogenetic tree construction because they were easy to amplify and sequence (Baldwin, et al., 1995). The rapidly increasing abundance of low copy

number gene sequences from high-throughput sequencing (HTS) systems is making coding sequence markers more appealing for tree-based inferences because of their longer lengths of homologous sequence and their varied substitution rates compared to non-coding markers (Wang, et al., 2004). These characteristics have been particularly useful for phylogenetic analysis of gene family structure and evolution (Bräutigam and Gowik, 2010).

No molecular marker has all the characteristics ideal for phylogenetic analysis. Consequently, it is important to consider the goals of the study in selecting the type of sequence that will be analyzed. For example, organellular DNA is often transmitted uni-parentally and is, therefore, not useful for tracking relationships in lineages that include cases of speciation through hybridization and subsequent allo-polyploidization (Harris and Ingram, 1991). Conversely, nuclear ribosomal RNA repeats are not useful for tracking specific parental genomes in hybrids and polyploids (Renny-Byfield, et al., 2011). In the case of SSRs for phylogeny reconstruction, it is critical to establish that the sequences being used are truly homologous as use of non-homologous sequences can bias results by breaking assumptions in a phylogenetic analysis (Goldstein and Pollock, 1997).

With respect to using transcriptome sequences for phylogenetic analysis, a frequent matter of debate is whether nucleotide or amino acid sequence should be the basis for analysis (Simmons, et al., 2002). Points to consider in making this choice include the fact that nucleotide substitutions occur with greater frequency than amino acid substitutions due to the degeneracy of the genetic code. When working with distantly related sequences, another subject of concern is that evolutionary footprints can be lost in cases where multiple substitutions confuse proper alignment of sequences. Species-specific codon biases and GC /AT-rich genomes can create further issues. Under these conditions, amino acids are often preferred for generating optimal

sequence alignments. However, the opposite is true in studies examining closely related organisms or highly conserved protein domains (Simmons, 2000). Overall, phylogenetic analyses of protein-coding sequences have been critically important for understanding the nature of genomic diversity, changes in gene family number and expression, and the frequency of genome duplication, as well as resolving relationships between closely related species (Barker, et al., 2009).

There are some issues with protein-coding sequences that make non-coding sequences preferred for certain types of analyses. For example, homologous coding sequences may represent paralogs, orthologs, or pseudo-genes, and it is critical to correctly type and bin these sequences across species under study (Zimmer and Wen, 2012). Where *de novo* sequencing is required, there can be difficulty in developing coding sequence primers that correctly amplify orthologous sequences across diverse lineages (Li, et al., 2008). Protein coding sequences have been used to trace the phylogenetic lineages of organelles, such as chloroplasts. However, interpretation of these phylogenies can be complicated by issues related to concerted evolution, maternal inheritance, gene loss and rearrangements, low mutation rates, high homoplasy and poorly analyzed paralogs (Koch, et al., 2003).

1.3 Sequence Alignment

For any application that applies comparative approaches using multiple sequences from closely related organisms, the key to making correct inferences is starting with a biologically optimal alignment. However, from the very start, proper sequence alignment has been the most fundamental problem for phylogenetic analysis. The purpose of alignment is to define a matrix for homology where each column in the alignment consists of the set of characters (amino acid or nucleotide) believed to have descended from a common ancestor (Boussau and Daubin, 2010).

It is the alignment matrix that is interpreted by the analysis algorithms to infer the steps of selection that yielded the existent set of sequences or capture signals to build a phylogeny. All steps in the process are interdependent and wrong decisions with respect to the alignment lead to an accumulation of errors and false inference. Additional complexity is introduced because in addition to substitutions, insertions and deletions may be introduced in each column as an alternative molecular character. To reduce alignment uncertainty, multiple alignment tools and manual curation methods are used to obtain optimal alignments before attempting further analysis (Wong, et al., 2008).

Satisfactory alignments of sequence pairs can be achieved using a variety of straightforward algorithms, such as dynamic programming (Kruskal, 1983). However, computing optimal alignments simultaneously for multiple sequences is a difficult task, and a wide variety of multiple alignment tools have been developed. Many of these incorporate alignment graph methods and use a variety of heuristic techniques, such as divide and conquer, or branch and cut for multiple sequence alignment (Reinert, et al., 2000; Althaus, et al., 2006). Optimal alignments are determined using scoring functions that use penalties for substitutions, insertions and deletions. The alignment yielding the best score based on a scoring matrix is generally considered the best alignment, but score maximization in this fashion is a very simplistic approach to obtaining the most biologically accurate alignment.

Heuristic methods based on progressive alignment construction have gained favor in recent years (Thompson, et al., 1994; Notredame, 2007). Under this approach, a guided tree in which more closely related sequences are aligned first followed by alignments incorporating more distantly related sequences are subsequently merged. This approach can be prone to erroneous alignments, especially where errors are introduced in the earlier steps and cannot be

corrected later when new sequences are added. Under such conditions, correction of misalignments can be done by applying iterative procedures (Lawrence, et al., 1993). Progressive alignment approaches have been extended using the concept of consistency, refinement and segmentation where by the guided tree and the multiple sequence alignment are re-estimated until convergence is achieved (Edgar, 2004; Rausch and Reinert, 2011). Another important component of the progressive algorithm is the scoring scheme used for pair-wise alignment. Of the two methods, namely the matrix-based method, which makes use of substitution matrices to calculate the cost of matching (Thompson, et al., 1994; Edgar, 2004), and the consistency-based method, which compiles a collection of local and global alignments for use as a position-specific substitution matrix during progressive alignment (Notredame, et al., 2000), the latter has proven more accurate as it incorporates more information (Blackshields, et al., 2006).

A number of new alignment tools utilize statistical approaches, such as Bayesian Markov chain procedures (Lunter, et al., 2005), as well as the Maximum Likelihood method, to align multiple sequences and construct phylogenies with better overall accuracy of the alignment and tree (Loytynoja and Goldman, 2009). Under conditions of distant homology, alignments based just on sequence information are shaky. In such cases, template-based alignment tools, which make use of structural or homology based information, are more likely to improve multiple sequence alignments (Armougom, et al., 2006). With the growth of structural motif databases, alignment tools that use protein structure information, although slow compared to sequence-based aligners, have shown the capability to deliver much more accurate alignments. However, they lag in their ability to utilize insertion/deletion information for constructing phylogenies where gaps are considered as missing data (Liu, et al., 2009).

Additional categories of aligners exist that are better suited for sequence-based inference analyses other than phylogenetic analysis. These include aligners for finding conserved sequence motifs (Schuler, et al., 1991), RNA sequences with high structural similarity (Gardner, et al., 2005), aligning genomic sequences using anchor-based methods (Sommer, et al., 2007) or alignment of sequences with different or shuffled domain organization where alignment is represented as a de-bruijn graph (Raphael, et al., 2004).

In the end, researchers need consider the degree of homogeneity between sequences under study. The effect of alignment uncertainties on evolutionary parameters is well studied and has pressed for further progress in multiple alignments methods (Wong, et al., 2008). In cases where sequences come from highly divergent species, reasonable multiple sequence alignments are difficult to optimize. Under such situations, use of consensus meta-methods that combine output from several methods provides better confidence in the estimation (Wallace, et al., 2006). Tools that implement global alignment approaches appear to be the best choice for sequences of similar length, while local algorithms are better for identifying conserved motifs in sequences with large insertions and deletions. Where protein sequences are available, domain and motif aligners provide robust support for creating biologically relevant alignments.

1.4 Methods for Reconstructing Phylogenies

Phylogenetic trees constructed using nucleic acid or protein sequence information can be used to infer relationships through comparison of sequence homology. Early methods employed simplistic, non-character-based methods that did not make full use of the information available in the data (Saitou and Nei, 1987). Character-based probabilistic methods, which can detect subtle patterns underlying the sequence data, were developed more recently (Whelan, et al., 2001). Methods using statistical principles are evolving rapidly, but still lack the capability to make full

use of certain types of information, such as insertions and deletions. Most current approaches assume that gaps in an alignment indicate missing data that can simply be ignored. However, such information can be useful for resolving difficult problems, such as ancient divergences, which can be further useful for rooting trees instead of using an out-group (Loytynoja and Goldman, 2009).

1.4.1 Distance-based Methods

Distance-based methods are agglomerative clustering methods used for creating phylogenetic trees by iteratively combining taxa. Under this method, a matrix of pair-wise genetic distances is used to construct a phylogenetic tree where the genetic distance is calculated as the fraction of positions where the two sequences differ (Felsenstein, 1988). Two of the important distance-based methods are the UPGMA (unweighted pair group method with arithmetic means) and the neighbor-joining method. The UPGMA method is a sequential clustering approach where sequences with the least genetic distance are grouped together. The drawback of the UPGMA method is that it assumes an ultrametric tree (branch lengths from the root to the tips are equal) with constant rates and is, thus, highly sensitive to unequal substitution rates between lineages (Huelsenbeck and Hillis, 1993).In contrast, the neighbor-joining method does not assume constant rate of evolution across lineages (Saitou and Nei, 1987) . It is a special case of the star decomposition method where raw data is provided in the form of a distance matrix such that the initial tree is star-shaped. By using a modified matrix, least-distance pairs of nodes are linked together until only two nodes remain, separated by a single branch. This method is often sufficient for recently diverged sequences, particularly when information on variable substitution rates (molecular clock) is available. It can be applied to divergent sequences provided correction for multiple substitutions is possible; however, the sensitivity of this technique to gaps and

insertions should not be forgotten (Yang and Rannala, 2012). The approach is also favored for situations where computational power is main concern. Tools such as MEGA (Kumar, et al., 2008) and PHYLIP (Felsenstein, 1989) build phylogenetic trees using distance-based method. The disadvantage of distance-based method is the loss of evolutionary information when the sequence alignments are converted to pairwise distances. This prevents use of evolutionary models containing parameters whose values cannot be known (Steel, et al., 1988). With the advent of probabilistic methods, which can incorporate optimality criteria, distance-based methods are being increasingly relegated to generation of 'rough draft' trees that can serve as starting points for more accurate and sophisticated phylogenetic reconstructions (Rosenberg and Kumar, 2001; Holder and Lewis, 2003).

1.4.2 Character-based Methods

Character-based methods, unlike distance-based methods, make use of all the evolutionary information available in the sequences. The word character can be interpreted as data in the form of morphological characters, nucleotides, amino acids or codons. The input data is in the form of a multiple alignment which is a (n*m) matrix with n number of species and m characters. The goal is to build a tree with the n species at the leaf and the internal nodes as the ancestral characters. The three character-based methods most highly used for reconstructing phylogenetic trees are the maximum parsimony method, the maximum likelihood method, and the Bayesian method. Except the for the maximum parsimony method, these approaches make use of evolutionary substitution models for correcting multiple mutation rates at each site (Holder and Lewis, 2003).

1.4.2.1 Maximum Parsimony Method

Unlike methods that use distance matrices, the maximum parsimony method uses all the information available in sequence data and maps it onto a putative species tree. The input data is in the form of characters (morphological, genetic, behavioral, etc.) that are divided into discrete character states. In general, this approach attempts to select a tree that has a minimal number of mutations based on observed data uncorrected for multiple substitutions (Steel and Penny, 2000). The disadvantage of this approach is that it does not allow assessment of all possible paths along the evolutionary tree, including paths that may require more than the minimum number of substitution events. Since the mutational path is unknown, exploring all possible paths that could explain the current data is a superior approach, but is more computationally intensive. Other drawbacks of maximum parsimony are that it does not make use of substitution models that can correct for evolutionary distance between sequences and that it is unable to take into consideration unequal numbers of changes across branches. As a consequence, it is susceptible to the long branch attraction effect, where similarity between two branches is the result of evolutionary convergence rather than sequence divergence (Felsenstein, 1978; Holder and Lewis, 2003). Despite these shortfalls, the maximum parsimony method can do a good job of identifying correct associations between sequences across moderately diverged lineages when taxon sampling is sufficiently dense (Junhyong, 1996).

1.4.2.2 Maximum Likelihood Method

As the name suggests, the maximum likelihood (ML) method not only uses a probabilistic approach, but also uses all information from the aligned sequence data in addition to explicit substitution models that attempt to correct estimates of evolutionary distance and minimize the

effect of multiple mutations occurring at a given site. The ML method relies on the likelihood function, which is the probability of observing the data conditioned on the parameters of the model. Hence, parameter values, including branch length and tree topology as well as the model parameters, that maximize the probability of observing the given data set are selected. The popularity of the ML method for phylogenetic analyses stems from its statistical stability, robustness against model violations, and convergence to the true tree as more data is added. However, as more data accumulates, the need to traverse the multi-dimensional parameter space places enormous demands on computational capacity. Consequently, most algorithms in the ML category use heuristic approaches to reduce the computational load imposed for evaluating non-optimal trees (Lewis, 1998; Guindon and Gascuel, 2003). These approaches approximate the maximum likelihood estimates by using heuristic techniques, such as branch-swapping, nearest-neighbor interchange, sub-tree pruning and re-grafting, as well as tree bisection and reconnection (Guindon and Gascuel, 2003; Stamatakis, 2006).

Although the ML method is often the first choice for tree construction, there are additional issues that must be considered in its application. Firstly, if the implementation uses heuristic techniques, there is a real chance that the output will represent a local maxima rather than a global optimum. Secondly, unlike Bayesian methods, the ML method uses a joint estimation approach, where a single parameter value can become overly influential for the final output (Holder and Lewis, 2003). Thirdly, the ML method uses bootstrapping as a measure of support, but, a high bootstrap value does not guarantee a correct phylogeny, only a degree of support for a particular model to the clades within a tree (Hillis and Bull, 1993).

The most popular ML tools use heuristic approaches to identify an optimal best tree, at the same time generating bootstrap values to support each node. Recently favored

implementations of the ML method have employed the 'greedy,' hill-climbing algorithm along with heuristic searches (Guindon and Gascuel, 2003; Stamatakis, 2006). RAxML (Randomized Accelerated Maximum Likelihood), a program developed by Stamataki et al. incorporates various algorithms and integrates heuristic searches, such as lazy sub-tree rearrangement, is a particularly speedy program that performs bootstrapping along with tree construction (Stamatakis, 2006). It can also be used in a parallelized format to compute ML across several nodes, and it incorporates methods to run the program on queued clusters with run time limits. There is also a RAxML-light version that can run on low-RAM computers (Stamatakis, et al., 2012).

1.4.2.3 Bayesian Methods

Bayesian methods are character-based, somewhat similar to the maximum likelihood method (Li, et al., 2000). The Bayesian approach not only incorporates complex evolutionary models, but also provides confidence values in the form of posterior probabilities. It incorporates the Bayes' theorem, which is the posterior probability of the hypothesis, and is proportional to the product of the maximum likelihood and the prior probability, divided by the unconditional probability of the data. With respect to phylogenetic analysis, the posterior probability is the uncertainty measure and represents the probability of all parameters given the observed data. The maximum likelihood is the probability of observing the data given the model parameters, while the prior probability is a means to integrate a hypothesis into determining the joint posterior probability distribution of the parameters. The prior probability is a way to convey scientific belief before having seen the data. In most of the cases, researchers specify prior probability distribution (flat prior probability) which conveys a lack of informativeness (Townsend, 2007) associated with the parameters so that only the ML values come into play for finding the tree with optimal

posterior probability. As summation over all possible trees and individual tree integrations over all possible parameter values become difficult to calculate, heuristic methods, such as Markov chain Monte Carlo (MCMC) procedures, are employed to approximate posterior probability distributions for the tree. Stochastic simulation is used for obtaining samples from posterior distribution of trees (Yang and Rannala, 1997). In such cases, the Markov chain is constructed using the state space for the model parameters and then randomly perturbing the system to calculate the relative posterior probability as the system moves through multidimensional space. Posterior probability is assumed from the parameter values when the system attains stationary distribution across the parameter space. Additional generations can be implemented to lengthen the chains and explore the multi-dimensional space for trees with higher posterior probabilities.

MrBayes (Huelsenbeck and Ronquist, 2001), a software tool that employs Bayesian methods, is widely used for constructing phylogenetic trees. MrBayes uses Metropolis-coupled Markov chain Monte Carlo methods (MC3) (Geyer, 1991) to approximate posterior probabilities, and under default conditions uses three heated chains with each cold chain to escape local peaks in the probability distribution. Over the years, MrBayes has been upgraded by incorporating models not only for nucleotide and amino acid substitutions as well as morphological data, but also models for relaxed molecular clocks and node dating. A mixed model approach allows for combining different datasets with user specifications related to linking or unlinking of specific parameters. The software provides a means for testing and comparing different models through calculated Bayes' factors for each model (Kass and Raftery, 1995). MrBayes also allows parallel processing across large computer clusters through implementation of the message-passing interface (MPI) to enable construction of gene trees, or a species tree from multiple gene trees using large partitioned datasets (Altekar, et al., 2004; Liu and Pearl, 2007).

Concerns for the attainment of convergence or stationary distribution in Bayesian approaches has led to the development of convergence assessment methods (Cowles and Carlin, 1996). MrBayes will calculate the average standard deviation of split frequencies, which is a measure of tree similarity between runs. Similarly, tools such as AWTY (Are we there yet?) compare the posterior probabilities from independent runs to detect convergence (Nylander, et al., 2008). Another issue in Bayesian methods is the exercise of prior assumptions. In general hypothesis, the use of flat priors (uniform prior), which assumes equal probability for observing all the possible values without any biasness while calculating posterior probabilities, is justifiable when no prior information is available. Also, use of the MCMC algorithm for approximation creates issues for arriving at convergence when dealing with very divergent sequences or partitioned datasets.

Overall, the fundamental approach of the Bayesian and maximum likelihood methods is toward full parameterization of evolutionary models. But, they differ in the manner they treat the parameters from these models. The Bayesian method treats every parameter as a random variable. This allows it to apply marginal estimation approach in which the parameters are integrated out to obtain the posterior probability of a tree. Marginalizing allows choosing a tree, which has a good support over a wide range of parameter values. Hence the final result is independent of the parameter value, unlike in the ML method.

On the other hand, the maximum likelihood method uses a joint estimation approach, which places more importance to the parameter values. The final tree has a single value for each of the parameter and the highest point in the parameter space. Overall, marginal estimates are preferred over joint estimates as a range of values over a single parameter value is more

convincing, as estimates of parameter values can be imperfect leading to a wrong tree (Holder and Lewis, 2003).

1.5 Post-Alignment Applications Using Phylogenetic Frameworks

1.5.1 Gene and species trees

To trace the ancestor-descendent history for a given locus (gene), the first step is to generate a gene tree based on the alignment of homologous sequence. The alignment is used to pick up variations in the sequence information which will inform the nodes representing coalescent events in the tree. The homology concept of orthologous and paralogous genes was born from gene tree analysis (Fitch, 1970). Conserved coding sequences from the same species constitute a gene family, with non-allelic sequences that generally perform separate and distinct functions from other family members called paralogous genes (paralogs). These are typically the product of gene duplication or polyploidization events. In contrast, orthologous genes (orthologs) are genes found in different species that derived from a common ancestor that predated the speciation event. Orthologs frequently retain ancestral functions and, thus, gene trees can be a cost-effective way to associate function between orthologs (Fitch and Margoliash 1967). This knowledge can be extended to improve understanding of gene interactions across different pathways as well as evolutionary patterns within gene families.

Procedures for identifying orthologs generally start with sequence-based approaches that utilize the BLAST algorithm to detect reciprocal best hits in cross-species comparisons (Altschul, et al., 1997). Many of the public databases implement programs that use unsupervised methods for detection of orthologs, such as Inparanoid (O'Brien, et al., 2005) or OrthoMCL (Li, et al., 2003). These methods perform iterative BLAST searches and then cluster the returned

sequences as reciprocal BLAST hits to increase sensitivity. OrthoMCL uses a sequence-based Markov clustering algorithm. In contrast to active searching, there are existing large-scale databases containing well recognized collections of orthologous genes. Similar to Clusters of Orthologous Groups (COG) (Tatusov, et al., 2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Mao, et al., 2005) is a manually curated database of ortholog groups that uses sequence similarity, phylogenetic analysis, and pathway interactions, among other criteria for ortholog detection. Additional programs capable of detecting orthologs with high sensitivity make use of information about protein-protein networks (Ogata, et al., 2000), metabolic pathway alignments (Bandyopadhyay, et al., 2006), or phylogenetic trees (Dufayard, et al., 2005). In cases where complete genomes are available, syntenic positioning of homologous sequences may be used to infer orthologous relationship (Fu, et al., 2007).

An equally important goal for phylogenetic analyses is to construct an accurate species tree in which individual genetic lineages from different populations and species are depicted in an ancestral-descendant relationship of the sampled individuals. Since reconstruction of a species tree requires identification of events such as duplication and gene loss and reconstruction of gene tree, and identification of duplication events and gene loss requires a species tree, estimation of a species tree becomes an equally important task as reconstruction of the gene tree (Boussau, et al., 2012). Reconciliation of incongruent gene trees is one of the biggest hurdles in constructing accurate species trees. Gene tree incongruence can arise from sequence alignment issues (Wong, et al., 2008) as well as a variety of anomalous evolutionary processes that act on the organism at the molecular level, such as gene duplication and loss events and horizontal gene transfer (HGT) as well as deep coalescence and branch length heterogeneity (Maddison, 1997; DeSalle, 2005). Gene duplication events are the most common contributors to gene tree incongruence, but HGT

is being increasingly identified as the force in the evolution of prokaryotes (Bapteste, et al., 2004). Deep coalescence or incomplete lineage discordance is when failure of ancestral gene copies to coalesce into a common ancestral copy until deeper than previous speciation events. This leads to the possibility of the lineage coalescing with lineages from distant population (Maddison, 1997). Similarly, branch length heterogeneity introduced by coalescence processes can generate unexpected phylogenetic signals when gene sequences are concatenated for constructing a species tree (Edwards, 2009). These anomalous processes vary for each gene in taxa under study and, thus, make every gene tree subtly different from the trees for other genes from the same collection of organisms.

Gene duplication and loss events within a gene family are inferred using reconciliation methods between the gene tree and the species tree (Goodman, et al., 1979). Incongruence between the two trees is interpreted as the evolution of the gene family through duplication and loss events. Tools such as Notung (Chen, et al., 2000) can identify gene duplication and loss events through reconciliation of species and gene trees, and thereby help identify likely orthologs and paralogs. Such methods can be utilized to their fullest extend only when a fully resolved species tree is available.

There are only a few widely accepted methods for inferring species trees, however, their performance when using large numbers of loci has not been well studied. The simplest method creates a species tree based on the most frequently observed gene tree topology (Ruvolo, 1997). The consensus and the concatenation methods for inferring species trees are considered more attractive because they do not explicitly model relationships between the gene tree and the species tree (Degnan, et al., 2009). Using the concatenation method, all the gene sequences are combined together and considered to belong to a single supergene (Edwards, et al., 2007). The

issue with the concatenation method is that it ignores the varying evolutionary rates applicable to each locus and thereby creates false support for internal nodes. Discordance resulting from incomplete taxon sampling has also been an issue in dealing with loci with varying evolutionary rates.

A variety of approaches to reduce species tree instability due to anomalous gene trees and varying evolutionary rates have emerged (Bryant, 2003; de Queiroz and Gatesy, 2007). With closely related species, sampling sequences from more species appears to be more effective than sampling more genes across the same species (Maddison and Knowles, 2006). Another commonly used approach is to infer a species tree that minimizes the number of deep coalescent events so as to be compatible with the gene tree (Maddison and Knowles, 2006). Similarly, implementation of ML or Bayesian methods has been used for inferring species trees, where maximum likelihood is use to infer a species tree by conditioning over all sets of gene trees, or an approximation is used by implementation of gene tree probabilities (Degnan and Rosenberg, 2009). BEST uses a Bayesian approach incorporating concordance factors to detect optimal species trees (Liu and Pearl, 2007). BEST estimates both species tree and the gene tree from multiple sequence alignments, and the concordance factors are used to estimates the degree of conflict between the gene trees (Ane, et al., 2007; Liu and Pearl, 2007). Additional methods are being developed that implement Bayesian approaches to improve the accuracy of gene trees and species trees by modeling gene duplication and loss, rate variations, and sequence substitution rates (Kumar, et al., 2012).

1.5.2 Detecting Diversifying and Directional Selection

Another major focus for phylogenomic analyses is to identify the tracks of molecular adaptation in evolutionary processes (Kimura, 1983; Nielsen, 2005). Such analyses are important to the

evolutionary biologist for their ability to increase our understanding of genetic changes that are critical for adaptations and for identifying those parts of current genetic diversity that have proven crucial to species survival in the past. Facile and low-cost sequencing and re-sequencing of complete genomes using high-throughput sequencing (HTS) methods has enabled application of comparative genomic approaches to identify codons within genes and genes in genomes that have been selected for their contributions to survival.

The implementation of codon-level substitution models has made it feasible to use both nucleotide and amino acid sequence information for assessing whether observed genetic changes are non-random, as would be the case when selection occurs (Goldman and Yang, 1994; Muse and Gaut, 1994). Importantly, these codon-based models take into consideration varying evolutionary rates at different codon positions as well as lack of independence at neighboring sites within a codon, both of which were typically ignored in earlier analyses. This makes it possible to compare rates of non-synonymous (K_a) versus synonymous (K_s) substitutions as a means for understanding the dynamics of sequence change across sites and through time (Goldman and Yang, 1994).

The simplest and the earliest approach was to generate an average ratio value (K_a/K_s or ω) for all sites and branches, but averaging over sites leads to power issues that limit correct identification of sites under positive selection (Golding and Dean, 1998) . In addition, the idea that the same selection pressure would apply across all sites and lineages was too simplistic.

The basic codon-level model proposed by Goldman and Yang works under a maximum likelihood framework and is applicable to multiple sequences (Goldman and Yang, 1994). It incorporates a Markov model of substitution which incorporates important evolutionary concepts, such as transition/transversion ratio and codon frequency biases, along with non-

synonymous and synonymous rate ratios. The basic model has been extended to detect selection pressure on an individual branch or set of branches (Yang, 1998) or at sites across all the lineages (Nielsen and Yang, 1998) as well as at sites within a specific lineage (Yang and Nielsen, 2002). Under the extended models, synonymous substitution is assumed to be neutral and hence a ratio of non-synonymous/synonymous substitution is used to infer adaptation at a codon level or along a lineage. These models are part of the widely used Codeml program in the Phylogenetic Analysis Library (PAML) software package (Yang, 2007). By categorizing each codon site on the basis of its ω value, where $\omega < 1$, $\omega = 1$ or $\omega > 1$ determines the purifying, positive or negative selection, respectively, direct estimation of the selection category acting on the codon is feasible. Other programs that can be used to detect dN/dS ratios include MrBayes and Hyphy (Ronquist and Huelsenbeck, 2003; Pond, et al., 2005).

The branch–site model in Codeml was implemented to detect relaxed constraints that act on sites from a pre-defined branch (Yang and Nielsen, 2002). The requirement of this model to specify *a priori* the branch under study is useful when adaptation related to functional changes after a gene duplication event in a given species is the question under study. The clade model was introduced to help detect selection footprints along clades within phylogenetic trees (Yang, 2007). These models are studied using hypothesis testing methods in which a null model that does not consider adaptive selection category ($\omega > 1$) is compared to the alternative model in which few sites are considered to be under relaxed constraint. Since these are nested models, a likelihood ratio test (LRT) (Stuart and Ord, 1999) of the hypothesis is conducted to determine whether positive selection is operating on the fore branch. The test statistic for the LRT is twice the difference between the likelihood values of the two models and the degree of freedom is the difference between the parameters of the two models. The Bayesian empirical-based (BEB)

method is implemented in Codeml to calculate posterior probabilities for site classes under the alternative model if the LRT suggests the presence of sites under adaptive selection on the predefined branch.

The Codeml package, although widely extensively, has been criticized for model inadequacies, such as the requirement for the branch –site model of *a priori* fore-branch which is assumed to have sites under relaxed constraint. There has also been criticism of the assumption of no positive selection and only neutral or purifying selection acting on lineages in background branches (Nozawa, et al., 2009; Pond, et al., 2011). Since ω is considered a function of the particular site in an alignment as well as the lineage in the phylogenetic tree, the assumption that selection forces will affect any two sites or any two branches in the same fashion is questionable (Kumar, et al., 2012). When branches under different selective pressure are incorrectly assigned to the same class because of less exhaustive options for selection profiles, the branch-site model can yield misleading results (Pond, et al., 2011). Also, the use of correction factors for multiple testing methods required when more than one branch is tested under the fore-branch criteria is problematic. Reevaluations of the branch- and site-models (Anisimova, et al., 2002) have identified deficiencies where depth and length of the sequences, number of sites affected, and size of the effect all play important roles in determining whether sites under relaxed constraint can be detected. Yet, both the site model and the branch-site model have been implemented in numerous phylogenetic studies (>1000) and have proven useful for biologists intent on detecting traces of adaptive evolution.

Although progress in this area has been slow, new models that try to address some of these deficiencies are being developed. When selection pressure fluctuates along a phylogenetic tree, the assumption of temporary constant ω for sites across the lineages is violated. This has led

to the development of models that can seek evidence of episodic selection (Tuffley and Steel, 1998). Guindon et al. came up with a covarion-based model that allowed selection pressure to change over time (Guindon, 2004). By incorporating three more switching parameters in the codon-level substitution model by Goldman and Yang (Goldman and Yang, 1994), they have introduced the concept of switching between selection classes at a given codon site and presumed it to be under the influence of external stimuli. This allowed for the nesting of models and use of LRT for testing the hypothesis. For detecting sites under positive selection pressure, this approach makes use of a Bayesian method. For the switching model, the assumption of a site remaining in the same selection class throughout its evolutionary history is considered false, and instead, the expected fraction of time the given selection process spends in a particular regime is calculated. This enables calculation of the expected frequency of positive selection classes for individual sites on a branch by branch basis. This approach can also detect the overall probability of the site being under positive selection over the entire phylogenetic history by taking a weighted average of all the possibilities conditioned on the observed codons at the tip of the tree and dividing that by the sum of the branch lengths.

Many of the common model assumptions are false and need to be relaxed so as to provide adequate discretion for analyzing complex biological data. For example, the assumption that synonymous substitution rates do not vary across genes and are selectively neutral has been shown to be false (Suzuki and Gojobori, 1999; Subramanian and Kumar, 2003) . Incorporation of synonymous rate variation into the models should help overcome some of the effects of such false assumptions. However, statistical inconsistencies continue to reduce detection rates for sites under adaptive selection. Evaluations comparing real data sets against simulated data have identified numerous issues and limitations associated with these models (Kumar, et al., 2012).

Thus, application of statistical inference to isolate candidate genes from the growing pool of sequenced genomes and checking them using experimental verification methods to understand their functional importance is a task that will continue for some time to come.

1.6 Objectives

This dissertation applies computational and experimental approaches to characterize the phenylalanine ammonia lyase (PAL) gene family, and understand the functions, expression patterns, and tissue specificity of its members in *Pinus taeda.*

In the initial work, assembled transcriptome data was used to identify PAL family members. The assembled contigs were experimentally verified using gene-specific primers to amplify PCR products that were characterized by DNA sequencing. Phylogenetic analysis was subsequently used to understand the evolutionary history of PAL genes and families across the gymnosperms. Two different approaches (ML and Bayesian) were used to construct a phylogenetic tree for the PAL gene family. Further evolutionary analysis incorporating the phylogenetic tree was performed using Fitmodel. The intention behind using Fitmodel was to detect lineages, as well as sites within those lineages, that have been under positive selection pressure. From the analysis, five PAL gene family members were identified in *P. taeda*, in contrast to earlier reports of a single PAL gene family member in this species of pine. The phylogenetic tree showed gymnosperm PAL genes to have experienced a very different evolutionary history from the angiosperm PAL genes. Only a single ancestral PAL gene member was retained in the angiosperms, and the recent gene duplication events were lineage-specific. In contrast, gymnosperms retained genes corresponding to multiple ancestral PAL gene family members.

Work described in the third chapter (Chapter 3) used PCR to study expression of the *P. taeda* PAL gene family members in different tissues from a mature tree, as well as under varying stress conditions. This work showed variations in PAL gene expression in different tissues and in response to stress treatments. Although we were unable to associate specific functions with individual gene family members, the results did suggest some level of redundancy in metabolic pathways in different tissues, and possibly some genotypic effects with respect to external perturbations. Along with it, promoter analysis of four of the five PAL gene family members using loblolly pine genome draft was conducted. The result showed number of sequence motifs observed in PAL genes in other species associated with expression in response to external stress as well as tissue-specific gene expression patterns.

In Chapter 4, a structure-function analysis informed by evolutionary analyses was performed. The evolutionary analysis highlighted a number of sites under apparent positive selection, and lineages harboring these sites were the product of duplication events. The sites were positioned on a model of the PAL protein structure, but further experimental work will be required to validate any functional importance for the detected sites.

Inconsistencies and power issues associated with the models implemented in Fitmodel provoked us to take a deeper look at the relative power and accuracy of this tool using simulation methods (Chapter 5). This analysis proved useful as Fitmodel has more realistic models, and a comparative study of the results generated with another highly used and studied tool, Codeml, highlighted the strengths and weaknesses of both tools.

Overall, these studies should prove useful to the pine community because PAL is the entry point for carbon flow into the phenylpropanoid biosynthetic pathway, which leads to lignin and a wide variety of secondary metabolites important for defense as well as growth and

development. Identification of the multi-gene family and the varied expression patterns in different tissues and conditions will provide a useful contribution for future work directed at pine tree improvement.

The simulation studies carried out as part of this work should help to inform evolutionary biologists of shortcomings in evolutionary analysis programs, such as Fitmodel. The high false-positive and false-negative rates for this program highlight the inefficiencies these tools face for correctly detecting sites under relaxed constraints.

1.7 References

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P. and Ronquist, F. (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference, *Bioinformatics*, **20**, 407-415.

Althaus, E., Caprara, A., Lenhof, H. and Reinert, K. (2006) A branch-and-cut algorithm for multiple sequence alignment, *Mathematical Programming*, **105**, 387-425.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25**, 3389-3402.

Álvarez, I. and Wendel, J.F. (2003) Ribosomal ITS sequences and plant phylogenetic inference, *Molecular Phylogenetics and Evolution*, **29**, 417-434.

Ane, C., Larget, B., Baum, D.A., Smith, S.D. and Rokas, A. (2007) Bayesian Estimation of Concordance among Gene Trees, *Molecular Biology and Evolution*, **24**, 412-426.

Anisimova, M., Bielawski, J.P. and Yang, Z. (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection, *Molecular Biology and Evolution*, **19**, 950-958.
Armougom, F. , Moretti , S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V. and Notredame, C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee, *Nucleic Acids Research*, **34**, W604-W608.

Baldwin, B.G. (1992) Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the Compositae, *Molecular Phylogenetics and Evolution*, **1**, 3-16.

Baldwin, B.G., Sanderson, M.J., Porter, J.M., Wojciechowski, M.F., Campbell, C.S. and Donoghue, M.J. (1995) The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny, *Annals of the Missouri Botanical Garden*, **82**, 247-277.

Bandyopadhyay, S., Sharan, R. and Ideker, T. (2006) Systematic identification of functional orthologs based on protein network comparison, *Genome Research*, **16**, 428-435.

Bapteste, E., Boucher, Y., Leigh, J. and Doolittle, W.F. (2004) Phylogenetic reconstruction and lateral gene transfer, *Trends in Microbiology*, **12**, 406-411.

Barker, M., S., Vogel, H. and Eric, S.M. (2009) Paleopolyploidy in the Brassicales: Analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales, *Genome Biology and Evolution*
1, 391-399.

Blackshields, G., Wallace, I.M., Larkin, M. and Higgins, D.G. (2006) Analysis and comparison of benchmarks for multiple sequence alignment, *In Silico Biology*, **6**, 321-339.

Boussau, B. and Daubin, V. (2010) Genomes as documents of evolutionary history, *Trends in Ecology and Evolution*, **25**, 224-232.

Boussau, B., Szollosi, G.J., Duret, L., Gouy, M., Tannier, E. and Daubin, V. (2012) Genome-scale co-estimation of species and gene trees, *Genome Research*, **23**, 323-330.

Bowcock, A.M., Ruiz Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R. and Cavalli-Sforza, L.L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites, *Nature*, **368**, 455-457.

Bräutigam, A. and Gowik, U. (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research, *Plant Biology*, **12**, 831-841.

Bryant, D. (2003) A classification of consensus methods for phylogenetics. In Janowitz, M.F. (ed), *Bioconcsensus*. American Mathematical society, 163-183.

Chen, K., Durand, D. and Farach-Colton, M. (2000) NOTUNG: A program for dating gene duplications and optimizinggene family trees, *Journal of Computational Biology*, **7**, 429-447.

Chen, X., Cho, Y. and McCouch, S. (2002) Sequence divergence of rice microsatellites in Oryza and other plant species, *Molecular Genetics and Genomics*, **268**, 331-343.

Cowles, M.K. and Carlin, B.P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association*, **91**, 883-904.

de Queiroz, A. and Gatesy, J. (2007) The supermatrix approach to systematics, *Trends in Ecology and Evolution*, **22**, 34-41.

Degnan, J.H., DeGiorgio, M., Bryant, D. and Rosenberg, N.A. (2009) Properties of consensus methods for inferring species trees from gene trees, *Systematic Biology*, **58**, 35-54.

Degnan, J.H. and Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent, *Trends in Ecology and Evolution*, **24**, 332-340.

DeSalle, R. (2005) Animal phylogenomics: multiple interspecific genome comparisons, *Methods in Enzymology*, **395**, 104 - 133.

Dufayard, J., Duret, L., Penel, S., Gouy, M., Rechenmann, F. and Perrière, G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases, *Bioinformatics*, **21**, 2596-2603.

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

Edwards, S.V. (2009) Is a new and general theory of molecular systematics emerging?, *Evolution*, **63**, 1-19.

Edwards, S.V., Liu, L. and Pearl, D.K. (2007) High-resolution species trees without concatenation, *Proceedings of the National Academy of Sciences*, **104**, 5936-5941.

Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading, *Systematic Zoology*, **27**, 401-410.

Felsenstein, J. (1988) Phylogenies from molecular sequences: inference and reliability, *Annual Review of Genetics*, **22**, 521-565.

Felsenstein, J. (1989) PHYLIP - Phylogeny inference package (Version 3.2). *Cladistics*, **5**, 163-166.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins, *Systematic Biology*, **19**, 99-113.

Fitch, W.M. and Margoliash , E. (1967) Construction of phylogenetic trees, *Science*, **155**, 279-284.

Fu, Z., Chen, X., Vacic, V., Nan, P., Zhong, Y. and Jiang, T. (2007) MSOAR: A high-throughput ortholog assignmentsystem based on genome rearrangement *Journal of Computational Biology.*, **14**, 1160-1175.

Gardner, P.P., Wilm, A. and Washietl, S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs, *Nucleic Acids Research*, **33**, 2433-2439.

Geyer, C.J. (1991) Markov chain monte carlo maximum likelihood *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Interface Foundation*, . In Keramidas (ed.), Fairfax Station, 156-163.

Glenn, T.C. (2011) Field guide to next-generation DNA sequencers, *Molecular Ecology Resources*, **11**, 759-769.

Golding, G.B. and Dean, A.M. (1998) The structural basis of molecular adaptation, *Molecular Biology and Evolution*, **15**, 355-369.

Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Molecular Biology and Evolution*, **11**, 725-736.

Goldstein, D.B. and Pollock, D.D. (1997) Launching Microsatellites: a review of mutation processes and methods of phylogenetic inference, *The journal of Heredity*, **88**, 335-342.

Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E. and Matsuda, G. (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences, *Systematic Zoology*, **28**, 132-163.

Guindon, S. and Gascuel, O. (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood, *Systematic Biology*, **52**, 696-704.

Guindon, S., Rodrigo, A, Dyer, Kelly A., Huelsenbeck, John P. (2004) Modeling the site-specific variation of selection patterns along lineages, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12957-12962.

Harris, S.A. and Ingram, R. (1991) Chloroplast DNA and biosystematics: the effect of intraspecific diversity and plastid transmission, *Taxon*, **40**, 393–412.

Hillis, D.M. and Bull, J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis, *Systematic Biology*, **42**, 182-192.

Hitchcock, E. (1840) *Elementary Geology*. Ivison and Phinney, New York.

Holder, M. and Lewis, P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches, *Nature*, **4**, 275.

Huelsenbeck, J.P. and Hillis, D.M. (1993) Success of phylogenetic methods in the four-taxon case, *Systematic Biology*, **42**, 247-264.

Huelsenbeck, J.P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees, *Bioinformatics*, **17**, 754-755.

Junhyong, K. (1996) General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa, *Systematic Biology*, **45**, 363-374.

Kass, R.E. and Raftery, A.E. (1995) Bayes Factors, *Journal of the American Statistical Association*, **90**, 773-795.

Kim, K.J. and Jansen, R.K. (1994) Comparisons of phylogenetic hypotheses among different data sets in dwarf dandelions (*Krigia, Asteraceae*): Additional information from internal

transcribed spacer sequences of nuclear ribosomal DNA, *Plant Systematics and Evolution*, **190**, 157-185.

Kimura, M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, 361.

Kingman, J.F.C. (2000) Origins of the coalescent: 1974-1982, *Genetics*, **156**, 1461-1463.

Knowles, L.L. and Maddison, W.P. (2009) Statistical phylogeography, *Annual Review of Ecology, Evolution and Systematics*, **40**, 593-612.

Koch, M.A., Dobe, C. and Mitchell-Olds, T. (2003) Multiple hybrid formation in natural populations: concerted evolution of the internal transcribed spacer of nuclear ribosomal DNA (ITS) in North American *Arabis divaricarpa* (Brassicaceae), *Molecular Biology and Evolution*, **20**, 338-350.

Kruskal, J.B. (1983) An overview of sequence comparison: time warps, string edits and macromolecules *Society for Industrial and Applied Mathematics*, **25** 201-237.

Kumar, S., Filipski, A.J., Battistuzzi, F.U., Kosakovsky Pond, S.L. and Tamura, K. (2012) Statistics and truth in phylogenomics, *Molecular Biology and Evolution*, **29**, 457-472.

Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences, *Briefings in Bioinformatics*, **9**, 299-306.

Lawrence, C.E., Altschul, S.F., Boguski, Mark S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science*, **262**, 208-214.

Lewis, P.O. (1998) A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data, *Molecular Biology and Evolution*, **15**, 277-283.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL:identification of ortholog groups for eukaryotic genomes, *Genome Research*, **13**, 2178-2189.

Li, M., Wunder, J., Bissoli, G., Scarponi, E., Gazzani, S., Barbaro, E., Saedler, H. and Varotto, C. (2008) Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species, *Cladistics*, **24**, 727-745.

Li, S., Pearl, D.K. and Doss, H. (2000) Phylogenetic tree construction using Markov chain Monte Carlo, *Journal of the American Statistical Association*, **95**, 493-508.

Liu, K., Raghavan, S., Nelesen, S., Linder, C.R. and Warnow, T. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees, *Science*, **324**, 1561-1564.

Liu, L. and Pearl, D.K. (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions, *Systematic Biology*, **56**, 504-514.

Loytynoja, A. and Goldman, N. (2009) Uniting alignments and trees, *Science*, **324**, 1528-1529.

Lunter, G., Miklos, I., Drummond, A., Jensen, J. and Hein, J. (2005) Bayesian coestimation of phylogeny and sequence alignment, *BMC Bioinformatics*, **6**, 83.

Maddison, W.P. (1997) Gene trees in species trees, *Systematic Biology*, **46**, 523-536.

Maddison, W.P. and Knowles, L.L. (2006) Inferring phylogeny despite incomplete lineage sorting, *Systematic Biology*, **55**, 21-30.

Mao, X., Cai, T., Olyarchuk, J.G. and Wei, L. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary, *Bioinformatics*, **21**, 3787-3793.

McCauley, D.E., Sundby, A.K., Bailey, M.F. and Welch, M.E. (2007) Inheritance of chloroplast DNA is not strictly maternal in *Silene vulgaris* (Caryophyllaceae): evidence from experimental crosses and natural populations, *American Journal of Botany*, **94**, 1333-1337.

Muse, S.V. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome, *Molecular Biology and Evolution*, **11**, 715-724.

Nielsen, R. (2005) *Statistical methods in molecular evolution.* Springer, New York, 495.

Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, *Genetics*, **148**, 929-936.

Notredame, C. (2007) Recent evolutions of multiple sequence alignment algorithms, *PLoS Computational Biology*, **3**, e123.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment, *Journal of Molecular Biology*, **302**, 205-217.

Nozawa, M., Suzuki, Y. and Nei, M. (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods, *Proceedings of the National Academy of Sciences*, **106**, 6700-6705.

Nylander, J.A.A., Wilgenbusch, J.C., Warren, D.L. and Swofford, D.L. (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics, *Bioinformatics*, **24**, 581-583.

O'Brien, K., Remm, M. and Sonnhammer, E. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs, *Nucleic Acids Research*, **33**, D476 - D480.

Ogata, H., Fujibuchi, W., Goto, S. and Kanehisa, M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucleic Acids Research*, **28**, 4021-4028.

Pond, K.S.L., Murrell, B., Fourment, M., Frost, S.D.W., Delport, W. and Scheffler, K. (2011) A random effects branch-site model for detecting episodic diversifying selection, *Molecular Biology and Evolution*, 3033-3043.

Pond, S.L.K., Frost, S.D.W. and Muse, S.V. (2005) HyPhy: hypothesis testing using phylogenies, *Bioinformatics*, **21**, 676-679.

Raphael, B., Zhi, D., Tang, H. and Pevzner, P. (2004) A novel method for multiple alignment of sequences with repeated and shuffled elements, *Genome Research*, **14**, 2336-2346.

Rausch, T. and Reinert, K. (2011) Practical Multiple Sequence Alignment. In Lenwood, S.H. and Ramakrishnana, N. (eds), *Problem solving handbook in computational biology and bioinformatics*. Springer New York, 21-40.

Reinert, K., Stoye, J. and Will, T. (2000) An iterative method for faster sum-of-pairs multiple sequence alignment, *Bioinformatics*, **16**, 808-814.

Renny-Byfield, S., Chester, M., Kova, A., Le Comber, S., Grandbastien, M., Deloger, M., Nichols, R., Macas, J., Novak, P., Chase, M.W. and Leitch, A.R. (2011) Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs, *Molecular Biology and Evolution*, **28**, 2843-2854.

Rodrigue, N., Philippe, H. and Lartillot, N. (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles, *Proceedings of the National Academy of Sciences*, **107**, 4629-4634.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models, *Bioinformatics*, **19**, 1572-1574.

Rosenberg, M.S. and Kumar, S. (2001) Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well, *Molecular Biology and Evolution*, **18**, 1823-1827.

Ruvolo, M. (1997) Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets, *Molecular Biology and Evolution*, **14**, 248-265.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, **4**, 406-425.

Schuler, G.D., Altschul, S.F. and Lipman, D.J. (1991) A workbench for multiple alignment construction and analysis, *Proteins: Structure, Function, and Bioinformatics*, **9**, 180-190.

Simmons, M.P. (2000) A fundamental problem with amino-acid-sequence characters for phylogenetic analyses, *Cladistics*, **16**, 274-282.

Simmons, M.P., Ochoterena, H. and Freudenstein, J.V. (2002) Amino acid vs. nucleotide characters: challenging preconceived notions, *Molecular Phylogenetics and Evolution*, **24**, 78-90.

Sleator, R.D. (2011) Phylogenetics, *Archives of Microbiology*, **193**, 235-239

Sommer, D., Delcher, A., Salzberg, S. and Pop, M. (2007) Minimus: a fast, lightweight genome assembler, *BMC Bioinformatics*, **8**, 64.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics*, **22**, 2688-2690.

Stamatakis, A., Aberer, A.J., Goll, C., Smith, S.A., Berger, S.A. and Izquierdo-Carrasco, F. (2012) RAxML-Light: a tool for computing terabyte phylogenies, *Bioinformatics*, **28**, 2064-2066.

Steel, M. and Penny, D. (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics, *Molecular Biology and Evolution*, **17**, 839-850.

Steel, M.A., Hendy, M.D. and Penny, D. (1988) Loss of information in genetic distances, *Nature*, **336**, 118-118.

Stuart, A. and Ord, K. (1999) *Kendall's advance theory of statistics*. Oxford University Press, New York.
Subramanian, S. and Kumar, S. (2003) Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes, *Genome Research*, **13**, 838-844.

Suzuki, Y. and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites, *Molecular Biology and Evolution*, **16**, 1315-1328.

Tatusov, R., Galperin, M., Natale, D. and Koonin, E. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution, *Nucleic Acids Research*, **28**, 33 - 36.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucl. Acids Res.*, **22**, 4673-4680.

Townsend, J.P. (2007) Profiling phylogenetic informativeness, *Systematic Biology*, **56**, 222-231.

Tuffley, C. and Steel, M.A. (1998) Modeling the covarion hypothesis of nucleotide substitution, *Mathematical Biosciences*, **147**, 63-91.

Volkov, R.A., Komarova, N.Y. and Hemleben, V. (2007) Ribosomal DNA in plant hybrids: inheritance, rearrangement, expression, *Systematics and Biodiversity*, **5**, 261-276.

Wallace, I.M., O'Sullivan, O., Higgins, D.G. and Notredame, C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee, *Nucleic Acids Research*, **34**, 1692-1699.

Wang, C., Moller, M. and Cronk, Q.C.B. (2004) Phylogenetic position of *Titanotrichum oldhamii* (Gesneriaceae) inferred from four different gene regions, *Systematic Botany*, **29**, 407-418.

Whelan, S., Lio, P. and Goldman, N. (2001) Molecular phylogenetics: state-of-the art methods for looking into the past, *Trends in Genetics*, **17**.

Wong, K., M., Suchard, M.A. and Huelsenbeck, J.P. (2008) Alignment uncertainty and genomic analysis *Science*, **319**, 473-476.

Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution, *Molecular Biology and Evolution*, **15**, 568-573.

Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood, *Molecular Biology and Evolution*, **24**, 1586-1591.

Yang, Z. and Nielsen, R. (2002) Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages, *Molecular Biology and Evolution*, **19**, 908-917.

Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method, *Molecular Biology and Evolution*, **14**, 717 - 724.

Yang, Z.H. and Rannala, B. (2012) Molecular phylogenetics: principles and practice, *Nature Reviews Genetics*, **13**, 303-314.

Zhu, Y., Queller, D.C. and Strassmann, J.E. (2000) A phylogenetic perspective on sequence evolution in microsatellite loci, *Journal of Molecular Evolution*, **50**, 324-338.

Zimmer, E.A. and Wen, J. (2012) Using nuclear gene data for plant phylogenetics: Progress and prospects, *Molecular Phylogenetics and Evolution*, **65**, 774-785.

Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history, *Journal of Theoretical Biology*, **8**.

CHAPTER 2

THE PHENYLALANINE AMMONIA LYASE (PAL) GENE FAMILY SHOWS A GYMNOSPERM SPECIFIC LINEAGE [1]

[1] Ujwal R Bagal, James H Leebens-Mack , W Walter Lorenz, Jeffrey F.D. Dean 2012, BMC Genomics 13(Suppl 3):S1, doi:10.1186/1471-2164-13-S3-S1, Reprinted here with permission of publisher

2.1 Abstract

Phenylalanine ammonia lyase (PAL) is a key enzyme of the phenylpropanoid pathway that catalyzes the deamination of phenylalanine to trans-cinnamic acid, a precursor for the lignin and flavonoid biosynthetic pathways. To date, PAL genes have been less extensively studied in gymnosperms than in angiosperms. Our interest in PAL genes stems from their potential role in the defense responses of *Pinus taeda*, especially with respect to lignification and production of low molecular weight phenolic compounds under various biotic and abiotic stimuli. In contrast to all angiosperms for which reference genome sequences are available, *P. taeda* has previously been characterized as having only a single PAL gene. Our objective was to re-evaluate this finding, assess the evolutionary history of PAL genes across major angiosperm and gymnosperm lineages, and characterize PAL gene expression patterns in *Pinus taeda.*

We compiled a large set of PAL genes from the largest transcript dataset available for *P. taeda* and other conifers. The transcript assemblies for *P. taeda* were validated through sequencing of PCR products amplified using gene-specific primers based on the putative PAL gene assemblies. Verified PAL gene sequences were aligned and a gene tree was estimated. The resulting gene tree was reconciled with a known species tree and the time points for gene duplication events were inferred relative to the divergence of major plant lineages.

In contrast to angiosperms, gymnosperms have retained a diverse set of PAL genes distributed among three major clades that arose from gene duplication events predating the divergence of these two seed plant lineages. Whereas multiple PAL genes have been identified in sequenced angiosperm genomes, all characterized angiosperm PAL genes form a single clade in the gene PAL tree, suggesting they are derived from a single gene in an ancestral angiosperm genome. The five distinct PAL genes detected and verified in *P. taeda* were derived from a

combination of duplication events predating and postdating the divergence of angiosperms and gymnosperms. Gymnosperms have a more phylogenetically diverse set of PAL genes than angiosperms. This inference has contrasting implications for the evolution of PAL gene function in gymnosperms and angiosperms

2.2 Introduction

Conifers have experienced large environmental and distributional changes during their evolution, dating back to the Mesozoic era (Eckert and Hall, 2006). To adapt to their diverse ecological habitats as well as the biotic and abiotic stresses associated with specific habitats, they have developed diverse and multi-layered chemical defense systems as a major component of their survival strategy (de Laubenfels, 1957). Conifer defense systems synthesize a wide range of secondary metabolites upon pathogen attack. Central to these chemical systems, a wide variety of phenolic compounds, both low molecular weight toxins and highly polymerized physical barriers, such as in lignin, serve to prevent invasion by pathogens (Bonello, et al., 2006). The precursors for many of these phenolic defense compounds are synthesized via the phenylpropanoid pathway (Somssich and Hahlbrock, 1998; Adomas, et al., 2007).

The phenylpropanoid pathway has been extensively studied with respect to production of natural products, such as flavonoids, isoflavonoids, hydroxy-cinnamic acids, lignin, coumarins, stilbenes and a wide variety of other phenolic compounds. These products serve diverse functions in plants, including protection against biotic and abiotic stresses, cellular signaling, and UV protection, as well as mechanical support and response to low levels of iron and phosphate (Dixon and Paiva, 1995).

Phenylalanine ammonia lyase (PAL; E.C 4.3.1.5), the key enzyme linking primary metabolism of aromatic amino acids with secondary metabolic products in plants, has been extensively studied since its discovery by Koukal and Conn (Koukal and Conn, 1961). PAL plays a key regulatory role in controlling biosynthesis of all phenylpropanoid products. As the entry point into the pathway, PAL catalyses the non-oxidative deamination of phenylalanine to trans-cinnamic acid and ammonia. Trans-cinnamic acid, in turn, is the common precursor for the

lignin and flavonoids biosynthetic pathways, which are highly complex and branched pathways (Ritter and Schulz, 2004). Increased activity of PAL has been correlated with increased production of phenylpropanoid products (Ozeki and Komamine, 1985), and levels of PAL activity vary with developmental stage, cell and tissue differentiation, and exposure to different stress stimuli (Jones, 1984; Lois , et al., 1989; Shufflebottom, et al., 1993). PAL has been reported to be stimulated by infection, mechanical wounding, UV irradiation, drought stress and drastic temperature changes (Edwards, et al., 1985; Campbell and Ellis, 1992; Lange, et al., 1995).

Until now, the gene content of conifer genomes has received less attention than angiosperm genomes despite the economic importance and ecological dominant of conifers in many terrestrial ecosystems (Stefanoviac, et al., 1998). Conifer genomes, at ca. 20 Gb on average, are larger than most angiosperm genomes. Yet in recent years, attempts to probe the genomic diversity of conifers have seen the development of such genomic resources as expressed sequence tag (EST) databases, cDNA microarray chips, and bacterial artificial chromosome (BAC) libraries, covering a handful of conifer species, notably loblolly pine (*Pinus taeda*) and white spruce (*Picea glauca*). Surprisingly, despite their large size, the structure of conifer genomes seems to be remarkably well conserved across well-diverged lineages. Chromosome number (12 or 13) is nearly the same in all conifer species (only three naturally occurring species of polyploidy conifer have been reported), and genetic mapping techniques have demonstrated substantial synteny across conifer species (Ritland, et al., 2006). Although the organization of large conifer genomes has not yet been deeply studied, some gene families have been reported as being substantially larger in conifers than in angiosperms for which reference genomes are available (Perry and Furnier, 1996), suggesting that gene duplication may be an important

mechanism for genome expansion in conifers. Large multigene families have been suggested to be correlated with conifer genome size (Ahuja and Neale, 2005).

In contrast to numerous reports of PAL gene families in angiosperms, as well as a few other gymnosperms, only a single gene copy was reported to exist in the *P. taeda* genome (Whetten and Sederoff, 1992). An initial objective of this study was to assess whether uncharacterized PAL genes existed in the genomes of *P. taeda* and other conifers. Moreover, we were interested in assessing the duplication history of PAL genes in angiosperms and gymnosperms. Specifically, we wanted to characterize the timing of PAL gene duplication events relative to the origin of the conifers and the divergence of gymnosperms and angiosperms. The timing of these duplication events has implications for hypotheses concerning functional evolution within the PAL gene family.

Our results indicate that *P. taeda* possesses at least five (5) distinct PAL genes, and expression was demonstrated for at least four of these inferred genes. Phylogenomic analysis identified a diverse set of gymnosperm-specific PAL genes, with at least three conifer lineage-specific duplication events and two ancient duplications events predating the divergence of gymnosperms and angiosperms. These ancient duplications suggest a very different evolutionary history for the gymnosperm PAL gene family from that experienced by the family in angiosperms.

2.3 Results

2.3.1 PAL genes in *Pinus taeda*

For *P. taeda*, five distinct PAL consensus sequences were identified in *de novo* transcriptome assemblies performed using three different assemblers (Table 1). Complete coding sequences of

ca. 725 amino acid residues were inferred for the pseudo-transcripts of all five PtPAL genes. The number of ESTs identified for each of the five PAL genes varied nearly 30-fold between genes and between tissue-specific libraries, suggesting very different levels and patterns of expression for the different gene family members (*data not shown*).

Because *de novo* assemblies generated in the absence of a reference genome sequence are susceptible to mis-assembly, we compared our contigs with sequences deposited in GenBank for conifer PAL genes that had been cloned and sequenced in previous studies. The lengths of the pseudo-transcripts returned from each of the three assemblers were found to be reasonable in comparisons with related sequences in GenBank. For example, the previously cloned loblolly pine PAL1 gene [GenBank: U39792.1] is 2435 bp in length and showed 100% sequence identity to our PtPAL1 assembly. This inferred transcript length also matched well with full-length cDNA transcripts for the four Arabidopsis PAL genes [GenBank: NM_129260, NM_115186.3, NM_120505.3, NM_111869.3], which ranged from 2463 to 2584 bp in length.

When compared to each other, PtPAL4 (Pteda28316) and PtPAL5 (Pteda34319) were quite similar at the amino acid level (93%), while PtPAL1 (Pteda1143311) and PtPAL2 (Pteda17307) exhibited just 86% similarity (Table 2). PtPAL3 (Pteda9006), the longest of the five sequences, showed the least identity to the other *P. taeda* PAL sequences (Figure 1).

The PtPAL1 sequence was found to be 98% identical to the genomic PAL gene sequence found on *P. taeda* BAC clone PT_7Ba2966L14 [GenBank AC241300.1]. Unlike angiosperm PAL genes, which include an intron, PtPAL1 and the PAL genes previously characterized in *P. banksiana(Butland, et al., 1998)* lack introns.

2.3.2 Validation of PAL cDNA sequence assemblies

Pine cDNA was amplified using gene-specific primer pairs corresponding to PtPAL1-PtPAL4. Amplification products of the expected sizes (300-450 bps) were detected as distinct bands on agarose gels (*data not shown*). These results confirmed expression of at least four members of the predicted PAL gene family in *P. taeda*. The sequence of the PtPAL5 proved too similar to PtPAL4 to allow for development of gene-specific primers that could discriminate between transcripts from the two genes. DNA sequencing of the amplified products confirmed the sequences inferred from our *in silico* assemblies.

2.3.3 Sequence conservation

To detect sequence conservation between PAL genes from distantly related plant species, the inferred amino acid sequences of PAL genes from 25 species were aligned. In the alignment some of the PAL genes from gymnosperms showed higher homology to genes from non-gymnosperm taxa, which was also reflected in the subsequent phylogenetic analysis. Active sites residues, including those imparting substrate specificity, as well as those for catalysis and formation of the MIO [4- methylidine-imidazole-5-one] prosthetic group were clearly conserved (Figure 1), and as were additional residues previously noted as conserved in PAL proteins (Calabrese, et al., 2004; Xu, et al., 2008). These observations strongly support the contention that all enzymes encoded by the genes included in these analyses bind and utilize the same substrate, phenylalanine.

2.3.4 Phylogenetic analysis

Phylogenetic analysis was performed to evaluate the evolutionary relationships among the 71 PAL sequences from 25 taxa selected for this analysis (Additional file 1). Trees were estimated from the multiple sequence alignment using Maximum Likelihood and Bayesian algorithms. In

both analyses a PAL gene from *Physcomitrella patens* was used for the out-group (Figure 2). The consensus trees obtained using either method showed similar organization, with gymnosperm genes distributed among three distinct clades. One gymnosperm-specific clade was placed just above the out-group branches in the PAL gene tree. A clade with the remaining genes split into another gymnosperm-specific clade and a second clade containing both angiosperm and gymnosperm PAL genes. The high bootstrap values and posterior probability evidence provided strong support for the organisation of the gymnosperm genes into these three distinct clusters. Within the angiosperm PAL gene clade, monocot and eudicot gene clusters were each monophyletic as described in a previous report (Hamberger, et al., 2007).

Because complete genome sequences are not yet available for pine and low gene expression levels often prevent sampling of particular mRNA sequences, the existence of additional PAL genes cannot be ruled out. It was clear from datasets for *Picea* cDNA sequences that additional PAL genes may exist in conifers since several homologous but incomplete *Picea* PAL gene sequences had to be removed from the collection prior to phylogenetic analysis because they were too short. PAL representation was similarly limited in the cDNA sets for other gymnosperms, but should improve as more sequences are added to the databases. Of particular interest for future studies will be functional analyses of gymnosperm PAL genes from all three gymnosperm-specific clades.

A species tree based on taxonomic information from the National Center for Biotechnology Information (NCBI) database was used to reconcile the gymnosperm section of the gene tree, keeping *P. patens* as the out-group (Figure 3). Notung version 2.6 (Chen, et al., 2000) was used to infer the relative timing of speciation and duplication events. At least five duplication events were successfully traced in the ancestral lineages and confirmed on the basis of strong bootstrap

support and posterior probability. Parsimony mapping suggests successive origins of three distinct gymnosperm PAL gene clades before the origin of the angiosperm clade. Ancestral seed plants had three distinct PAL genes which have been conserved in gymnosperms, but two of these ancestral genes were lost in the angiosperm lineage after divergence from the gymnosperms. In addition, PAL genes have also diversified more recently within the pines (Figure 2).

The oldest PAL gene duplication event evident in Figure 3 took place after the divergence of the vascular plants (Tracheophyta) and mosses, as represented by *Physcomitrella*. The second oldest duplication took place after divergence of the seed plants (Spermatophyta) and *Selaginella* (Lycopodiophyta). Following these duplication events, the duplicate copies of PAL were retained in the gymnosperms and all but one paralog was lost on the branch leading to the angiosperms. Further diversification of the PAL gene family from a single gene copy occurred within the angiosperms after the split of the dicots and monocots. The occurrences of independent lineage-specific duplications within the monocots and dicots have led to substantial elaboration of PAL gene families in various species of angiosperms.

At least three ancestral duplication events within the gymnosperms were suggested on the basis of high confidence values. Because of incomplete sampling and low branch support across the conifer species, duplication events close to the tips of the tree were not fully resolved. One duplication event was evident within the Pinaceae family, where one of the duplicate gene copies was found in closely related pine species (*P. lambertiana* and *P. palustris*), which had smaller EST datasets, but not in *P. taeda*.

2.4 Discussion

Phenylalanine ammonia lyase, which belongs to the lyase class I super-family of enzymes (Ritter and Schulz, 2004), is a primary control point for the phenylpropanoid pathway, which in part explains the multi-gene families seen for PAL in almost all plants studied to date (Lois , et al., 1989; Wanner, et al., 1995; Kumar and Ellis, 2001; Reichert, et al., 2009). This study is the most extensive phylogenomic study so far for the PAL gene family, particularly with respect to conifers.

De novo transcriptome assemblies without a reference genome can lead to misassembly of contigs where transcripts are inaccurately joined together or single transcript can be split into two (Surget-Groba and Montoya-Burgos, 2010). Three different programs were used to assemble the transcriptomes of *P. taeda* and 12 other conifers. We were able to identify five distinct PAL genes in all three *P. taeda* cDNA sequence assemblies. The contig lengths were comparable to those of cloned PAL genes available in the GenBank, suggesting no obvious errors in the assemblies.

The total number of sequences assembled to form each contig varied for all five PALs reflecting variation in their respective expression patterns [*data not shown*]. Differential expression patterns suggest that the various PtPAL gene products may be responsible for providing biosynthetic precursors to different phenylpropanoid branch pathways under different developmental conditions or in response to various external stimuli.

Apparently complete coding sequences were obtained for all five *P. taeda* PAL genes. Variability in the sequences was mostly associated with the terminal ends of the coding sequences. As PtPAL4 and PtPAL5 were 88% identical at the nucleotide level and clustered together on the same phylogenetic branch, they cannot be ruled out as allelic forms.

Gymnosperm PAL genes were clustered into three clades. The origin of the most ancient clade is estimated to predate the origin of vascular plants (including *Selaginella*) while the other two clades originated by gene duplication within a seed-plant ancestor before the divergence of angiosperms and gymnosperms. This result suggests that PAL genes were lost on the branch leading to angiosperms.

The phylogeny of the PAL gene family identified in this study showed distinctive branching patterns for the monocot, dicot, and gymnosperm clades. The monocot-dicot split has been described previously (Hamberger, et al., 2007). In addition to ancient duplication events in a common ancestor of vascular plants and seed plants, respectively, distinct PAL genes clades within the monocots and eudicots point to lineage-specific diversification events within each of these taxa. The gymnosperm PAL clade that is sister to the angiosperm clade may include genes encoding for PAL isoforms that have similar functions or are regulated by similar developmental control mechanisms (Pina and Errea, 2008).

The existence of two additional gymnosperm PAL gene clades indicates maintenance of PAL genes in gymnosperms and loss of diversity in angiosperms (Okada, et al., 2008). The branching patterns within the conifer genes within these clades are in accordance with patterns reported previously for these species (Eckert and Hall, 2006).

Duplication events have been an important theme in the evolution of the PAL gene family. At least five distinct duplication events can be identified in the PAL gene tree, with the oldest event following the divergence of *Physcomitrella*. Duplication events in the ancestral lineage, as well at the tip of the gymnosperm branch, suggest potential sources of functional variability (Pina and Errea, 2008). Multigene families can be formed for a variety of reasons. It may be for production of additional trans-cinnamic acid for downstream metabolic pathways in these

lineages; for instance, for increased expression of lignin biosynthesis in response to insect and pathogen attack (Okada, et al., 2008). Duplicate copies of these genes may encode different isoforms, or each duplicate copy may have a distinct expression pattern in terms of response to different physiological needs, such as tissue development or resistance to biotic and abiotic stresses (Logemann, et al., 1995). Thus, in artichoke, three different PAL genes were suggested to play different roles in defense responses (Paolis, et al., 2008). In Poplar, one PAL gene product was associated with formation of condensed tannins while another was associated with lignin production (Chen, et al., 2000). In tobacco, post-transcriptional regulation of one PAL gene in the family was reported, although the exact mechanism was not clear (Reddy, et al., 2000). Early duplication events within a gene family, when compared to recent divergence events where genes from same species cluster together, have shown distinguishable biochemical, molecular and catalytic properties (Kumar and Ellis, 2001). Based on this model, PtPAL4 and PtPAL5 may have resulted from a recent duplication event and may still serve overlapping functions (Figure 2). Likewise, as seen in other species, PtPAL genes that do not cluster together are more likely to encode PAL isozymes having unique functions, perhaps playing different metabolic role by producing different products under varying conditions.

2.5 Conclusions

Five PtPAL genes were identified in cDNA assemblies for loblolly pine. The phylogenetic tree constructed using PAL gene sequences from 25 species including angiosperms, gymnosperm and basal taxa shows a very different evolutionary history for PAL genes in the gymnosperms, which may suggest different functional regulation. Reconciliation suggests early duplication events in the evolutionary history of PAL gene family as the root cause of phylogenetically separated genes rather than recent duplication events, which would lead to gene clustering.

2.6 Materials and Methods

2.6.1 PAL in conifer assemblies

A Community Sequencing Project undertaken at the US DOE Joint Genome Institute (http://www.jgi.doe.gov/) used 454 pyrosequencing to produce EST datasets for 12 conifer species, namely *Cedrus atlantica* (SRA023736), *Cephalotaxus harringtonia* (SRA023613), *Gnetum gnemon* (SRA023615), *Picea abies* (SRA023567), *Pinus lambertiana* (SRA023577), *Pinus palustris* (SRA023739), *Pinus taeda* (SRA023533), *Podocarpus macrophyllus* (SRA023741), *Pseudotsuga menziesii* (SRA023776), *Sciadopitys verticilliata* (SRA023758), *Sequoia sempervirens* (SRA023765), *Taxus baccata* (SRA023771), *and Wollemia nobilis* (SRA023774).All sequences are available from the Short-Read Archive (SRA) at GenBank.

Along with previously generated Sanger EST sequences available in GenBank, five cDNA libraries representing various tissues, treatments and genotypes of *P. taeda* yielded over 4 million reads used in these studies. Elongating shoot tissue cDNA libraries for the remaining conifer species were sequenced to yield from 0.4 to 1.2 million reads per species. The sequences were all assembled using three different assembly algorithms, namely Newbler Version 2.3 (454 Life Sciences), miraEST (Mira) Version 3.0.5 (Chevreux, et al., 2004), and SeqMan NGen Version 3.0 (DNAStar). The consensus sequences along with their annotations from all the three assemblies, as well as such information as number of sequences aligned to form a contig and overall contig length, were retrieved from the Conifer DBMagic database (Lorenz, et al., 2011).

Existing PAL sequences from *P. taeda* and other angiosperms available in GenBank were used as seeds to perform BLAST searches against the Conifer DBMagic database for novel PAL sequences from *P. taeda* and the other 11 conifers. Contigs with complete or near-complete coding sequence was selected for further analyses, while shorter sequences were discarded.

2.6.2 Sequence verification

Since the assembled sequences were products of *de novo* assemblies, they were considered prone to error. To confirm that the sequences represented true gene products, experimental verification was performed by designing gene-specific primers for the PtPAL1-PtPAL4 consensus sequences and verifying the identity of amplified products by sequencing of the PCR amplimers.

The same assembled contigs used for the phylogenetic analysis were used as the basis for designing gene-specific oligo-nucleotide primers for PCR studies. A pair of PCR primers, Fwd ["AAGAACGCAGAAGGTGAGAAGG"] and Rev ["AGCATTTGAAGAGAGGGACTATGAC"], were designed to amplify 307 bp from PtPAL1 (Pteda1143311). In a similar fashion, Fwd ["CTGACTGAGACTGCCCAAATTC"] and Rev ["TCCTCCTGCCGTTTCCAATG"] primers amplified a 444 bp sequence from PtPAL2 (Pteda17307), Fwd ["TCAGAGTTGGGAACCGATTTG"] and Rev ["CTATTGATTCATTGTTGTTGGAACC"] primers amplified a 388 bp sequence from PtPAL3 (Pteda9006), and Fwd ["CCAATAACGACGCTTCTATCCTTAC"] and Rev ["CGCCGTTCCATCGCTCAAG"] primers amplified a 306 bp sequence from PtPAL4 (Pteda28316). The quality of these primers was assessed *a priori* using the program Beacon Designer 3 (PREMIER Biosoft International, Palo Alto, CA).

PCR amplification of PAL cDNAs synthesized from mRNA extracted from the stem tissues of *P. taeda* seedlings was performed in a 50 µl reaction volume. Reaction mixtures contained 1µl of Taq polymerase, 2 µl of 10mM dNTP, 4 µl of Optiprime 10x buffer, 3 µl of 5mM primer and 10 µl of 1 ng/µl cDNA template was used for each gene-specific amplification reaction. Amplification was performed using a GeneAMP PCR system 9700 thermocycler (Applied Biosystems, Culver City, CA). The cycling conditions were 1 cycle of 95°C for 3 min followed

by 40 cycles of 94°C for 30 secs, 55°C for 30 secs, 72°C for 90 sec, and 1 cycle of 72°C for 10 min. PCR products were purified using a DNA purification kit (Invitrogen Corporation, Carlsbad, CA) and dideoxy sequencing was performed using an Applied Biosystems 3730XL sequencer at the Georgia Genomics Facility (http://dna.uga.edu/).

2.6.3 Taxonomic representation

Based on preliminary phylogenetic analyses, 25 representative taxa were selected for compilation of PAL genes, sequence alignment and tree reconstruction. The selected taxa (the number of PAL genes used from each species is shown parenthetically) comprised five dicotyledonous angiosperms, namely, *Arabidopsis thaliana (4), Medicago truncatula (2), Nicotiana tabacum (2), Persea americana (1) and Populus trichocarpa (4), and one monocot, Oryza sativa (8).* Nineteen gymnosperm taxa were analyzed, including *Cupressus atlantica (2), Cephalotaxus harringtonia (4), Ginkgo biloba (2), Gnetum gnemon (1), Picea abies (4), Picea sitchensis (3), Pinus lambertiana (4), Pseudotsuga menziesii (4), Pinus palustris (2), Pinus pinaster (2), Pinus sylvestris (1), Pinus taeda (5), Podocarpus macrophyllus (1), Sciadopitys verticillata (3), Sequoia sempervirens (4), Taxus baccata (3), and Wollemia nobilis (3).* Two non-seed plant taxa, the moss, *Physcomitrella patens (2),* which also served as an out-group, and the lycopod, *Selaginella kraussiana (1),* were also used for these analyses.

2.6.4 Taxon Sampling and Phylogenetic analysis

The nucleotide sequences and corresponding amino acid sequences for the representative taxa were collected from various public databases, including GenBank, PlantGDB and PlantTribes (Benson, et al., 1999; Dong, et al., 2004; Wall, et al., 2008). The inferred transcript sequences for *C. atlantica, C. harringtonia, P. abies, P. lambertiani, P. macrophyllus, P. palustris, P. sylvestris,*

S. verticillata, *S. sempervirens*, *T. baccata*, and *W. nobilis* were contigs assembled from cDNA datasets obtained by pyrosequencing. Using different angiosperm and gymnosperm PAL genes as seeds, outputs with expect-values (e-value) of $1e^{-45}$ and below were selected for use in the study. The resulting dataset was further sorted and screened to remove possible contaminations resulting from assembly errors, sequences with length ≤50% of the complete CDS length, or putative allelic sequences sampled from the same species, i.e. those with nucleotide sequence identities ≥95%. Following the screening process, 71 sequences from 25 taxa remained for phylogenetic and molecular evolutionary analyses of the *PAL* gene family.

An initial multiple sequence alignment for the complete dataset was performed using MAFFT (Katoh, et al., 2002). Multiple codon alignment corresponding to protein sequences was performed using PAL2NAL (Suyama, et al., 2006). Molecular phylogeny estimates were derived using RAxML (Stamatakis, 2006) and MrBayes (Huelsenbeck and Ronquist, 2001) on a 2430 character sequence alignment. For the RAxML estimation, a generalized time-reversible (GTR) substitution model (Lanave, et al., 1984) with across-site rate variation modeled as a gamma distribution (Yang, 1993) and invariant sites (GTR+GAMMA+I), was used for nucleotide alignments. For amino acid alignment, the JTT [Jones, Taylor and Thornton] substitution model (Jones, et al., 1992) with gamma distribution was used. Clade support was evaluated using 100 bootstrap replicates. For the MrBayes analysis, the GTR model was used with GAMMA correction and eight discrete rate categories. Analyses with MrBayes were performed over two runs, including four chains and three million generations per run. After 750,000 (25%) burn-in generations, trees were sampled every 300 generations and used to estimate posterior probabilities for each clade.

2.7 List of abbreviations used

1 PAL: Phenylalanine ammonia lyase

2 PtPAL: PAL from *Pinus taeda*

2.8 Competing interests

The Authors declare that they have no competing interests.

2.9Authors' contributions

JLM and JFDD conceived the general idea of the study; WWL carried out the assembly work.URB acquired the PAL sequences and performed the analysis on them. URB and JLM interpreted the data.

2.10 Acknowledgements

This work was supported in part by a research assistantship to U.B. from the Warnell School of Forestry and Natural Resources. Much of the sequence data used in this study was produced by the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/) Community Sequencing Project “An Expanded EST Resource for Pines and Other Conifers.”

2.11 References

Adomas, A., Heller, G., Li, G., Olson, A., Chu, T., Osborne, J., Craig, D., Van Zyl, L., Wolfinger, R., Sederoff, R., Dean, R.A., Stenlid, J., Finlay, R. and Asiegbu, F.O. (2007) Transcript profiling of a conifer pathosystem: response of *Pinus sylvestris* root tissues to pathogen (*Heterobasidion annosum*) invasion, *Tree Physiology*, **27**, 1441-1458.

Ahuja, M.R. and Neale, D.B. (2005) Evolution of genome size in conifers, *Silvae Genetica*, **54**, 126-137.

Benson, D., Boguski, M., Lipman, D., Ostell, J., Ouellette, B., Rapp, B. and Wheeler, D. (1999) GenBank, *Nucleic Acids Research*, **27**, 12-17.

Bonello, P., Gordon, T.R., Herms, D.A., Wood, D.L. and Erbilgin, N. (2006) Nature and ecological implications of pathogen-induced systemic resistance in conifers: a novel hypothesis, *Physiological and Molecular Plant Pathology*, **68**, 95-104.

Butland, S.L., Chow, M.L. and Ellis, B.E. (1998) A diverse family of phenylalanine ammonia lyase genes expressed in pine trees and cell cultures, *Plant Molecular Biology*, **37**, 15-24.

Calabrese, J.C., Jordan, D.B., Boodhoo, A., Sariaslani, S. and Vannelli, T. (2004) Crystal structure of Phenylalanine Ammonia Lyase: multiple helix dipoles implicated in Catalysis, *Biochemistry*, **43**, 11403-11416.

Campbell, M.M. and Ellis, B.E. (1992) Fungal elicitor-mediated responses in pine cell cultures : purification and characterization of phenylalanine ammonia lyase, *Plant Physiology*, **98**, 62-70.

Chen, K., Durand, D. and Farach-Colton, M. (2000) NOTUNG: A program for dating gene duplications and optimizinggene family trees, *Journal of Computational Biology*, **7**, 429-447.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E.G., Wetter, T. and Suhai, S. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs, *Genome Research*, **14**, 1147-1159.

de Laubenfels, D.J. (1957) The status of "Conifers" in vegetation classifications, *Annals of the Association of American Geographers*, **47**, 145-149.

Dixon, R.A. and Paiva, N.L. (1995) Stress-induced phenylpropanoid metabolism, *Plant Cell*, **7**, 1085-1097.

Dong, Q., Schlueter, S.D. and Brendel, V. (2004) PlantGDB, plant genome database and analysis tools, *Nucleic Acids Research*, **32**, D354-359.

Eckert, A.J. and Hall, B.D. (2006) Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): phylogenetic tests of fossil-based hypotheses, *Molecular Phylogenetics and Evolution*, **40**, 166-182.

Edwards, K., Cramer, C.L., Bolwell, G.P., Dixon, R.A., Schuch, W. and Lamb, C.J. (1985) Rapid transient induction of phenylalanine ammonia lyase mRNA in elicitor-treated bean cells, *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 6731-6735.

Hamberger, B., Ellis, M., Friedmann, M., de Azevedo Souza, C., Barbazuk, B. and Douglas, C.J. (2007) Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*: the Populus lignin toolbox and conservation and diversification of angiosperm gene families, *Canadian Journal of Botany*, **85**, 1182-1201.

Huelsenbeck, J.P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees, *Bioinformatics*, **17**, 754-755.

Jones, D.H. (1984) Phenylalanine ammonia lyase: regulation of its induction, and its role in plant development, *Phytochemistry*, **23**, 1349-1359.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences, *Bioinformatics*, **8**, 275-282.

Katoh, K., Misawa, K., Kuma, K.-i. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform, *Nucleic Acids Research*, **30**, 3059-3066.

Koukal, J. and Conn, E.E. (1961) The metabolism of aromatic compounds in higher plans. Purification and properties of the phenylalanine deaminase of *Hordeum vulgare*, *The Journal of biological chemistry*, **236**, 2692-2698.

Kumar, A. and Ellis, B.E. (2001) The phenylalanine ammonia lyase gene family in raspberry. Structure, expression, and evolution, *Plant Physiology*, **127**, 230-239.

Lanave, C., Preparata, G., Sacone, C. and Serio, G. (1984) A new method for calculating evolutionary substitution rates, *Journal of Molecular Evolution*, **20**, 86-93.

Lange, B.M., Lapierre, C. and Sandermann Jr, H. (1995) Elicitor-induced spruce stress lignin (structural similarity to early developmental lignins), *Plant Physiology*, **108**, 1277-1287.

Logemann, E., Parniske, M. and Hahlbrock, K. (1995) Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley, *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 5905-5909.

Lois , R., Dietrich, A., Hahlbrock, K. and Schulz, W. (1989) A phenylalanine ammonia lyase gene from parsley: structure, regulation and identification of elicitor and light responsive *cis*-acting elements., *The EMBO Journal*, **8**, 1641-1648.

Lorenz, W.W., Ayyampalayam, S., Bordeaux, J.M., Howe, G.T., Jermstad, K.D., Neale, D.B., Rogers, D.L. and Dean, J.F.D. (2011) Conifer DBMagic: a database housing multiple de novo

transcriptome assemblies for 12 diverse conifer species, *Tree Genetics & Genomes*, **8**, 1477-1485.

Okada, T., Mikage, M. and Sekita, S. (2008) Molecular characterization of the phenylalanine ammonia lyase from *Ephedra sinica*, *Biological & Pharmaceutical Bulletin*, **31**, 2194-2199.

Ozeki, Y. and Komamine, A. (1985) Changes in activities of enzymes involved in general phenylpropanoid metabolism during the induction and reduction of anthocyanin synthesis in a carrot suspension culture as regulated by 2,4-D, *Plant Cell Physiology*, **26**, 903-911.

Paolis, A.D., Pignone, D., Morgese, A. and Sonnante, G. (2008) Characterization and differential expression analysis of artichoke phenylalanine ammonia lyase coding sequences, *Physiologia Plantarum*, **132**, 33-43.

Perry, D.J. and Furnier, G.R. (1996) *Pinus banksiana* has at least seven expressed alcohol dehydrogenase genes in two linked groups, *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13020-13023.

Pina, A. and Errea, P. (2008) Differential induction of phenylalanine ammonia lyase gene expression in response to in vitro callus unions of Prunus spp, *Journal of Plant Physiology*, **165**, 705-714.

Reddy, J.T., Korth, K.L., Wesley, S.V., Howles, P.A., Rasmussen, S., Lamb, C. and Dixon, R.A. (2000) Post-transcriptional regulation of phenylalanine ammonia lyase expression in tobacco following recovery from gene silencing, *Biological Chemistry*, **381**, 655-665.

Reichert, A.I., He , X.Z. and Dixon, R.A. (2009) Phenylalanine ammonia lyase(PAL) from tobacco(*Nicotiana tabacum)*: characterization of the four tobacco PAL genes and active hetrotetramer *Biochemistry*, **424**, 233-242.

Ritland, K., Ralph, S., Lippert, D., Rungis, D. and Bohlmann, J. (2006) A new direction for conifer genomics. In Williams, C.G. (ed), *Landscapes, Genomics and Transgenic Conifers*. Springer, Netherland, 75-84.

Ritter, H. and Schulz, G.E. (2004) Structural basis for the entrance into the phenylpropanoid metabolism catalyzed by phenylalanine ammonia lyase, *Plant Cell*, **16**, 3426-3436.

Shufflebottom, D., Edwards, K., Schuch, W. and Bevan, M. (1993) Transcription of two members of a gene family encoding phenylalanine ammonia lyase leads to remarkably different cell specificities and induction patterns, *The Plant Journal*, **3**, 835-845.

Somssich, I.E. and Hahlbrock, K. (1998) Pathogen defence in plants -- a paradigm of biological complexity, *Trends in Plant Science*, **3**, 86-90.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics*, **22**, 2688-2690.

Stefanoviac, S., Jager, M., Deutsch, J., Broutin, J. and Masselot, M. (1998) Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences, *American Journal of Botany*, **85**, 688-697.

Surget-Groba, Y. and Montoya-Burgos, J.I. (2010) Optimization of de-novo transcriptome assembly from next-generation sequencing data, *Genome Research*, **20**, 1432-1440.

Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, *Nucleic Acids Research*, **34**, W609-W612.

Wall, P.K., Leebens-Mack, J., Muller, K.F., Field, D., Altman, N.S. and de Pamphilis, C.W. (2008) PlantTribes: a gene and gene family resource for comparative genomics in plants, *Nucleic Acids Research*, **36**, D970-976.

Wanner, L.A., Guoqing, L., Ware, D., Somssich, I.E. and Davis, K.R. (1995) The phenylalanine ammonia lyase gene family in *Arabidopsis thaliana Plant Molecular Biology*, **27**, 327-338.

Whetten, R.W. and Sederoff, R.R. (1992) Phenylalanine ammonia lyase from loblolly pine: purification of the enzyme and isolation of complementary DNA clones, *Plant Physiology*, **98**, 380-386.

Xu, F., Cai, R., Cheng, S., Du, H., Wang, Y. and Cheng, S. (2008) Molecular cloning, characterization and expression of phenylalanine ammonia lyase gene from *Ginkgo biloba*, *African Journal of Biotechnology*, **7**, 721-729.

Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, *Molecular Biology and Evolution*, **10**, 1396-1401.

2.12 Supplementary Material

Supplementary data is available at

http://www.biomedcentral.com/bmcgenomics/supplements/13/S3

Tables

Table 2.1: PtPAL (1-5) *de novo* transcriptome assemblies of *P. taeda*

MIRA[1A]	**Uniscript [2]**	**Uniscript Length [3]**	**Total Seq [4]**
PAL1 / MIRA	P.taeda.JGI_rep_c1829	2081	295
PAL2 / MIRA	P.taeda.JGI_rep_c1015	2660	392
PAL3 / MIRA	P.taeda.JGI_rep_c9006	2826	142
PAL4 / MIRA	P.taeda.JGI_rep_c4552	2474	155
PAL5 / MIRA	P.taeda.JGI_rep_c10177	2269	62
Newbler [1B]	**Uniscript [2]**	**Uniscript Length [3]**	**Total Seq [4]**
PAL1 / Newb	contig57512	3573	2924
PAL2 / Newb	isotig35091	3022	606
PAL3 / Newb	isotig22550	3110	506
PAL4 / Newb	isotig41305	2538	279
PAL5 / Newb	isotig35702	2278	87
SeqMan NGen[1C]	**Uniscript [2]**	**Uniscript Length [3]**	**Total Seq [4]**
PAL1 / NGen	Contig347	3773	2746
PAL2 / NGen	Contig13311	2889	560
PAL3 / NGen	Contig5954	2370	223
PAL4 / NGen	Contig26398	1798	154
PAL5 / NGen	Contig50748	2277	75

[1A] miraEST (Mira) Version 3.0.5, [1B]: Newbler Version 2.3, [1C]: SeqMan NGen Version 3.0 (DNAStar), [2] Contig name, [3] Contig lengths, and [4] numbers of sequences assembled to form a contig.

Table 2.2: *P. taeda* PAL inferred amino acid sequence percent identity / similarity

Gene id ([1])	PtPAL1	PtPAL2	PtPAL3	PtPAL4	PtPAL5
PtPAL1(754)	#	76/87	64/79	68/81	64/79
PtPAL2(727)		#	65/80	67/79	63/78
PtPAL3(808)			#	64/77	60/75
PtPAL4(711)				#	88/93
PtPAL5(687)					#

[1] Amino acid length

Figures

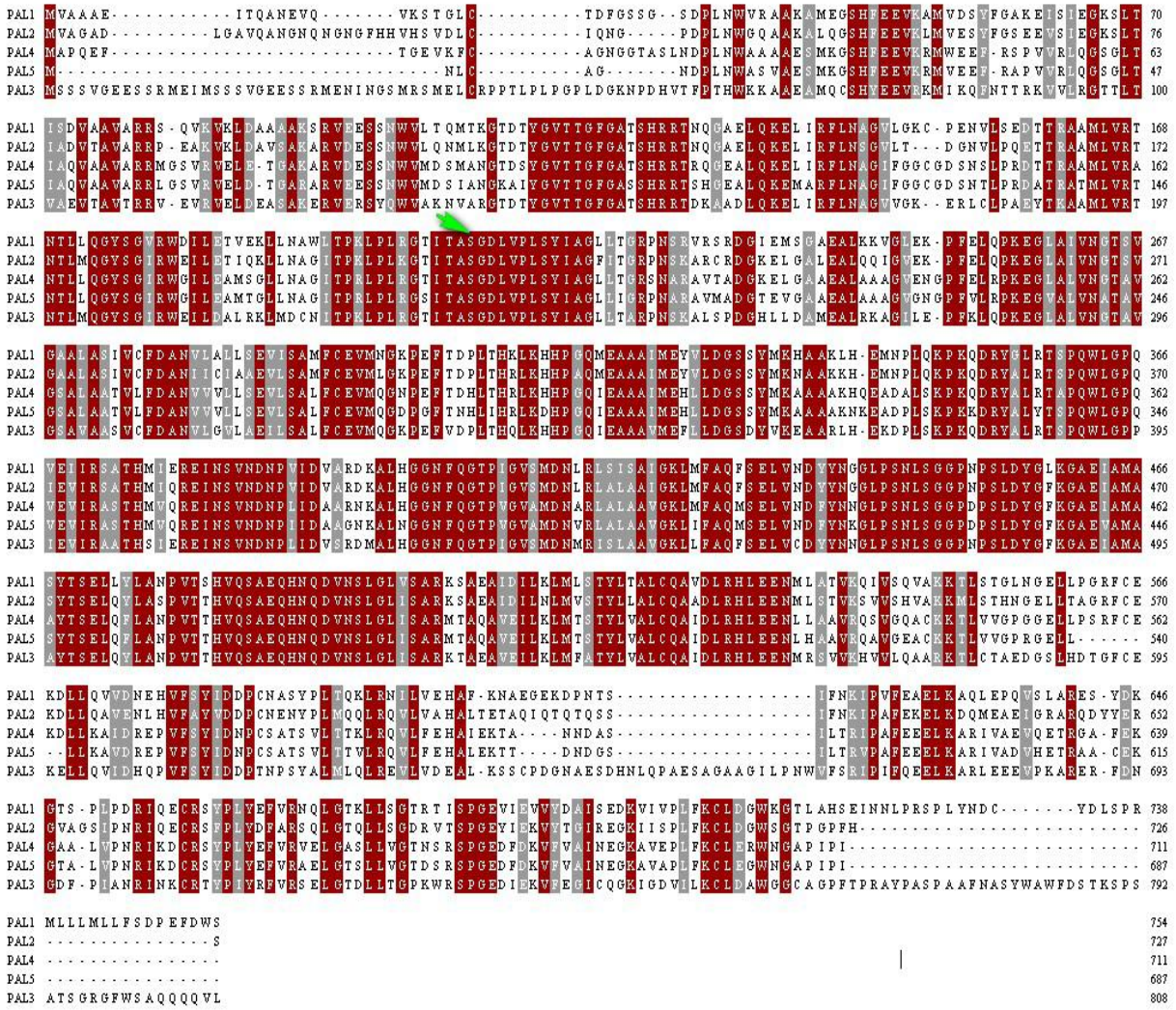


Figure 2.1: Alignment between the five PtPAL genes in *P. taeda*.

Amino acid sequences from five PAL genes from loblolly pine were aligned. The red shaded regions shows perfect conservation at sites while the grey shaded region shows partial conservation. Arrow indicates position of the conserved MIO region (Ala-Ser-Gly triad)

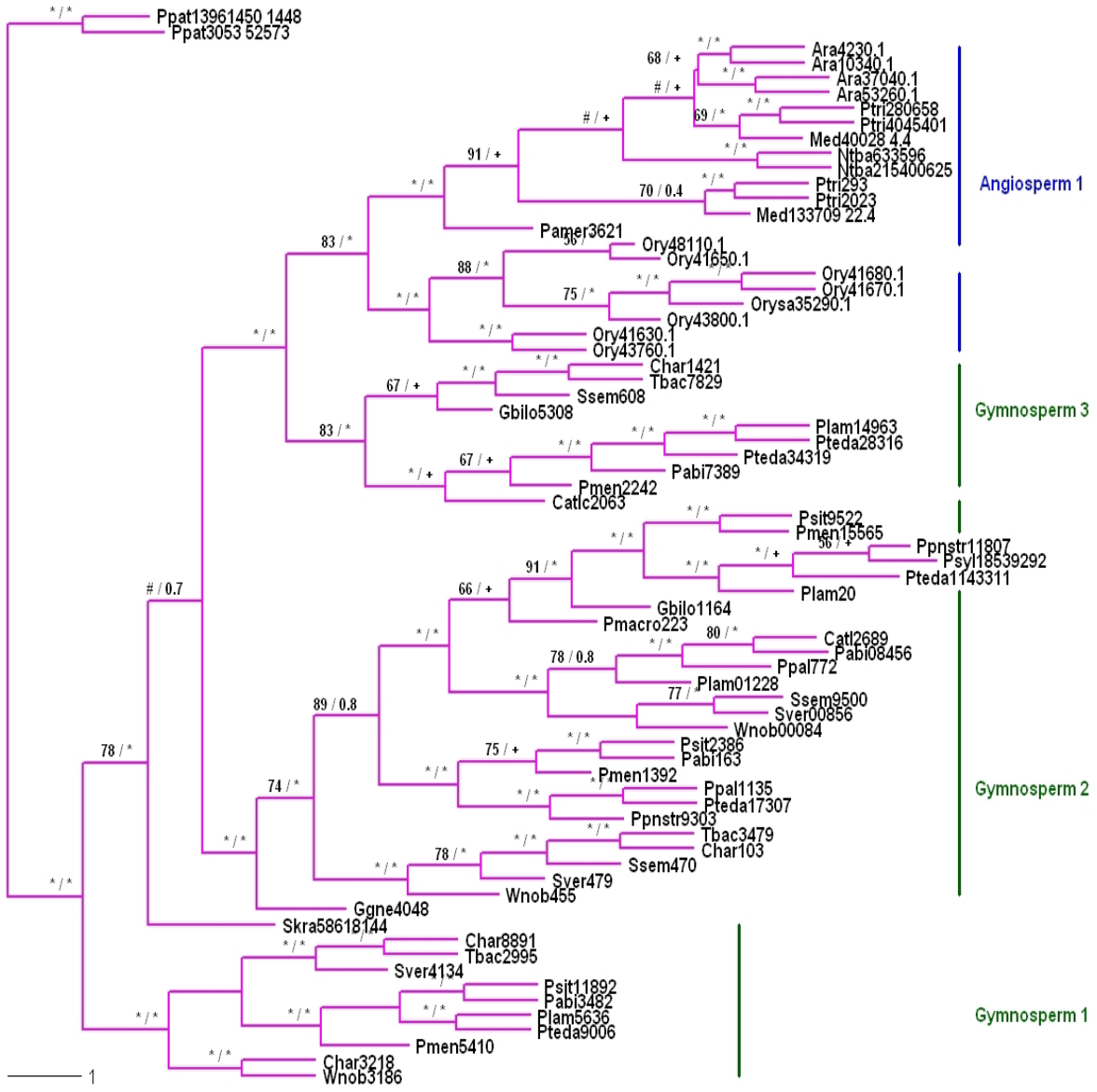


Figure 2.2: Consensus tree of the Phenylalanine ammonia lyase gene family

Numbers at nodes are nonparametric Bootstrap values (BS) from Maximum likelihood (ML) and Posterior Probabilities (PP) from Bayesian Inference (BI), respectively, separated by a slash. Asterisks (*/*) symbol indicates [90-100] / [0.9 - 1.00] support values. The # symbol indicates BS values lesser than 50%. Plus (+) symbol indicate variation in branching patterns between the ML and BI consensus trees.

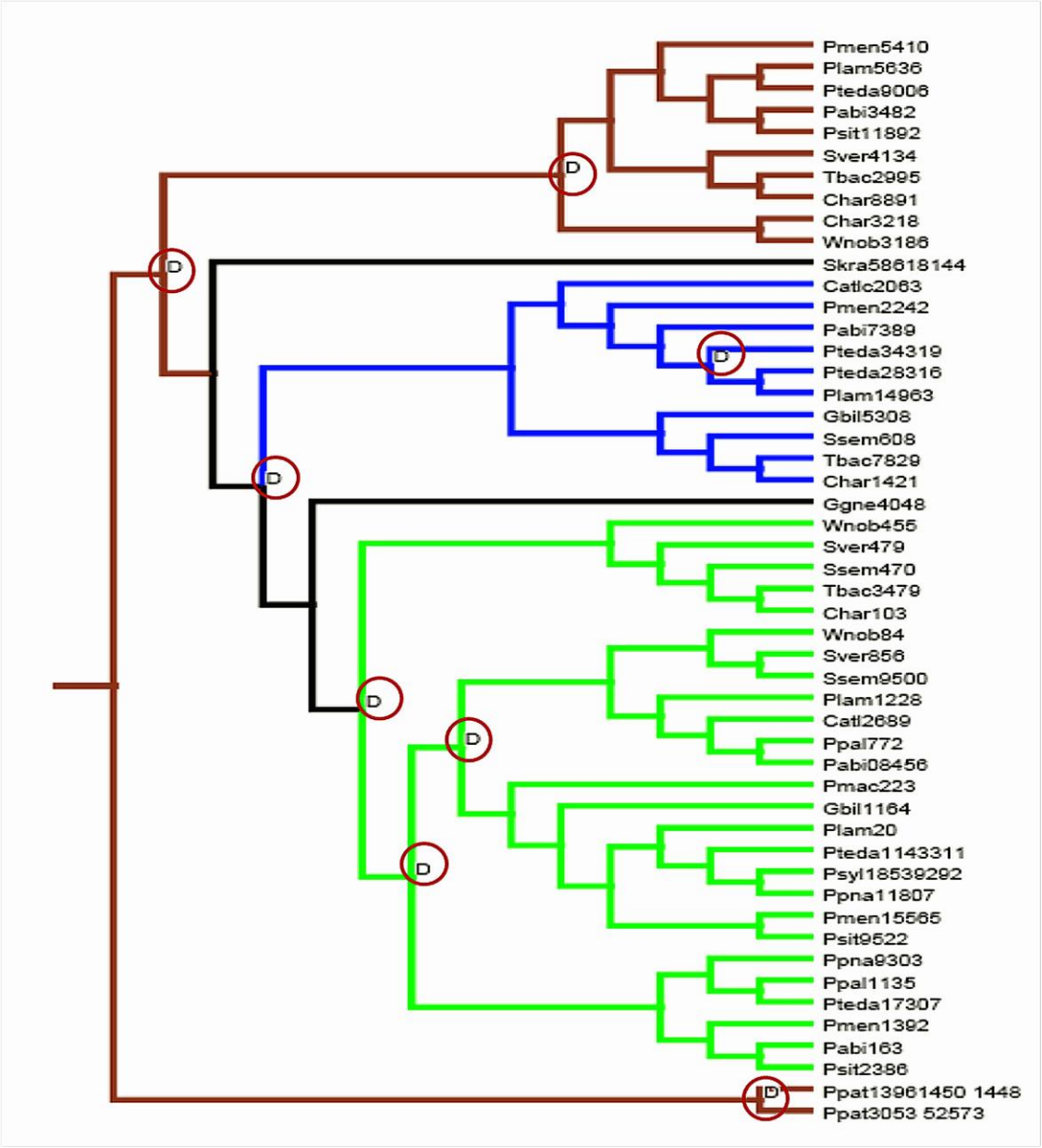


Figure 2.3: NOTUNG: Reconciled gene tree

A reconciled gene tree with duplication events as obtained from Notung is depicted. Duplication nodes are marked with circles. The branch shading corresponds to the pattern of gymnosperms branching. The blue branch indicates gymnosperm sequences that clustered with angiosperm PAL genes. The green branch indicates a unique gymnosperm branch, while the brown branch indicates gymnosperm sequences clustering with sequences from basal taxa.

CHAPTER 3

TISSUE SPECIFIC AND STRESS-INDUCED EXPRESSION OF PHENYLALANINE AMMONIA LYASE GENE FAMILY MEMBERS IN *PINUS TAEDA* [2]

[2] Ujwal R. Bagal, Jeffrey F.D. Dean. To be submitted to Tree Physiology

3.1 Abstract

Five phenylalanine ammonia lyase (PAL) gene family members previously identified as transcription products in *Pinus taeda* tissues were confirmed as the complete gene family for this species by analysis of the recently completed draft genome sequence. Promoter analyses for four of the genes identified a variety of sequence motifs associated with regulated gene expression. Some of these motifs have been found in the promoters of PAL genes in other species where they have been shown to contribute to tissue-specific gene expression patterns as well as expression in response to external stress. Quantitative RT-PCR analyses using four gene primer pairs were used to probe gene expression patterns for the *P. taeda* PAL (PtPAL) gene family members. PtPAL1, which had previously been associated with lignifying xylem tissues and wood formation in this species, was seen as the most highly expressed PtPAL gene under all of the test conditions. Expression data indicated that PtPAL 4 is also expressed in lignifying xylem. PtPAL2 is not expressed in xylem, but is relatively highly expressed in roots growing under control conditions. While PtPAL3 expression was detected under some conditions, levels were generally too low to indicate specific function for this family member. This work sets the stage for future studies to more precisely determine physiological functions for each PtPAL gene.

3.2 Introduction

To adapt to diverse ecological habitats, forest trees have evolved complex chemical responses that include the production of diverse secondary metabolites (Tsai, et al., 2006; Witzell and Martin, 2008; Vogt, 2010). In particular, the phenylpropanoid pathway is a major source of precursors for branch pathways that lead to a wide spectrum of secondary metabolites, including flavonoids, anthocyanins, stilbenes and other small molecules, as well as polymers like lignin and suberin (Edwards, et al., 1985; Kodan, et al., 2002). These secondary metabolites serve diverse biological functions including antimicrobial and anti-herbivory activity, cell wall reinforcement, and pigmentation to attract pollinators and seed-dispersal agents (Dixon and Paiva, 1995; Ververidis, et al., 2007).

Phenylalanine ammonia lyase (PAL; E.C 4.3.1.5) catalyzes the non-oxidative deamination of phenylalanine to form *trans*-cinnamic acid and ammonia, which is generally considered the initial step of the phenylpropanoid pathway (Dixon, et al., 2002). The *trans*-cinnamic acid produced by this reaction is subsequently used as the aromatic donor in many phenolic metabolites produced in plants. As the common entry point to the phenylpropanoid pathway PAL is well positioned to serve as a regulatory point for carbon flux through the pathway. Experiments with transgenic plants that either over-express or under-express PAL demonstrated the enzyme to be a rate-limiting step under some conditions and also that PAL enzyme activity can be regulated by allosteric effectors (Howles, et al., 1996; Blount, et al., 2000; Chang, et al., 2009).

Numerous studies have reported increased PAL activity in conjunction with increased levels of phenolic compounds in plants responding to various biotic stresses, including insect feeding and pathogen elicitors (Lange, et al., 1995; Chaman, et al., 2003), as well as abiotic

stresses, including low temperature, wounding, drought, and UV exposure (Dixon and Paiva, 1995; Costa, et al., 2003). The structural phenolic polymer, lignin, is a major constituent of secondary cell walls in plant vascular tissues and is particularly abundant in woody plants (Wang, et al., 2013). In fact, the word lignin derives from the Latin word for wood, "lignum." Despite its rather ubiquitous nature, lignin can also shows variability in content and composition in different tissues and cell types under different developmental stages and in response to external environmental stimuli (Douglas, 1996; Moura, et al., 2010). These variations in lignin content often coincide with changes in PAL activity (Sewalt, et al., 1997; Möller, et al., 2006).

Gene duplication and the expansion of functional gene families has been a primary route for the diversification of the pathways leading to secondary metabolites as individual gene family members provide the means to regulate carbon flux into pathways by differential expression in response to specific developmental cues or environmental stresses (Cramer, et al., 1985; Eckert and Hall, 2006). PAL has been reported to exist as a multi-gene family in virtually all the plants studied to date (Lois, et al., 1989; Wanner, et al., 1995; Kumar and Ellis, 2001; Bagal, et al., 2012). PAL gene family members frequently show differential expression with respect to tissue and cell type, so gene expression levels can integrate signals from developmental cues as well as environmental stimuli to produce a range of required phenolic compounds (Bevan, et al., 1989; Liang, et al., 1989). Thus, specific PAL gene family members having individual responsibilities for certain processes can be regulated by a specific set of signals in a specific tissue at a specific developmental stage (Logemann, et al., 1995; Paolis, et al., 2008).

Like other plants, conifers use phenolic compounds for disease resistance, either in the form of phytoalexins (small molecules) or as structural barriers like lignin (Kodan, et al., 2002;

Giberti, et al., 2012). Similarly, in adapting to their environment while growing over long time periods, conifers experience anatomical, structural and compositional changes to their xylem tissues that result in variations in wood properties, which has consequent effects on products subsequently made using wood (Plomion, et al., 2001). For example, the formation of compression wood, which has thicker cell walls and higher lignin content, forms on the underside of branches and leaning stems (Timmell 1986). In accordance with previous findings, these changes in lignin content and composition can be attributed to modulation of functional genes associated with the phenylpropanoid and lignin biosynthetic pathways (Westing, 1965; Yeh, et al., 2006; Brennan, et al., 2012; Villalobos, et al., 2012). Given the wide use of conifer wood for the production of paper and other consumer products, increased understanding of the factors influencing lignin content and composition has substantial economic importance.

The *P. taeda* reference genome sequence (Neale, et al., 2014) was used to verify the presence of five distinct PAL gene family members in loblolly pine detected in a previous project (Bagal, et al., 2012). The current study was undertaken to analyze PAL gene expression in specific tissues as well as in response to specific biotic and abiotic stresses. This information will give us new insight into the physiological roles played by individual members of the PAL gene family in loblolly pine.

3.3 Materials and Methods

3.3.1 Genomic sequence analysis

Genome sequences encoding *Pinus taeda* PAL (PtPAL) genes were identified using BLAST (Altschul, et al., 1990). The recovered sequences were assessed to identify allelic forms as well as pseudogenes and were then associated with the five genes identified in the previous study of

expressed genes. The genomic and cDNA sequences were then aligned to assess structural differences, such as the possible presence of introns.

From prior experience (Yuan and Dean, 2010), genomic DNA sequences stretching 2 Kbp upstream of the putative translational start site (ATG) were deemed sufficient for comparative promoter analyses. These upstream sequences for all five PAL genes were analyzed for putative *cis*-regulatory elements using PLACE (Higo, et al., 1998) and plantcare (Rombauts, et al., 1999). As the current draft reference genome sequence for loblolly pine is still highly fragmented, 2 Kbp of upstream sequence for PtPAL3 and PAL5 could not be recovered from scaffolds containing those coding gene sequences. For PtPAL3 and PtPAL5, sequences stretching only 154 bp and 724 bp upstream of the respective translational start sites were analyzed.

3.3.2 Plant materials

3.3.2.1 *Pinus taeda*

Tissue samples from a mature loblolly pine tree harvested from Whitehall Forest (Athens, GA) were prepared as previously described (Lorenz, et al., 2011). Seven tissue types, namely vertical trunk xylem, vertical crown xylem, branch compression xylem, branch opposite xylem, vertical trunk phloem, apical shoot tips, and 1st-year cones, were used for the assessment of tissue-specific expression. In addition to these aerial tissues, RNAs from the roots of young *P. taeda* trees subjected to drought or control conditions (Lorenz, et al., 2009) were also used in the analyses of tissue-specific expression.

To assess differential PAL gene expression in response to abiotic (drought) and biotic (*Sirex noctilio* venom) stresses we used shoot tip explants from three 72-month old loblolly pine

trees maintained as hedges in the greenhouse. The trees were all from a half-sibling family for which the elite *Pinus taeda* clone 7-56 was the maternal parent. The hedges were used to provide multiple clonal shoot tips (6-9 inches in length) for replicated experiments. A total of fourteen shoots from each tree were randomly assigned to the control treatment (6 shoots) or one of the two stresses (four shoots each to drought or venom treatment). This experimental setup was replicated three times with all samples run in parallel.

For the control treatment, excised shoot tips were placed in tubes containing deionized water for time intervals of 0, 4, 8, 12, 24 and 36 hrs. For the drought stress treatment, shoot tips placed upright in empty tube were collected after 8, 12, 24 and 36 hrs. For the biotic stress treatment, shoot tips were administered 50 µl of 20 mg/ml *S. noctilio* venom (Bordeaux, et al., 2012) and then incubated in deionized water for 4, 8, 12 and 24hrs. At each time point, the individual shoot tips were collected and separated into needle and stem tissues. These were subsequently flash-frozen separately in liquid nitrogen and samples segregated by individual shoot tip and tissue were placed in storage at -80°C until further use.

3.3.3 Gene expression analyses

3.3.3.1 Template preparation

The needles and stem tissues from each treated shoot tip were grinded separately to powder using either a liquid nitrogen-cooled mortar and pestle or a SPEX model 6850 freezer mill (SPEX, Metuchen, NJ). RNA was prepared by the protocol of Chang et al. (Chang, et al., 1993) as modified (Lorenz, et al., 2010). Briefly, 3 grams of ground tissue was suspended in 20 ml of RNA extraction buffer (2 % CTAB, 2% polyvinyl pyrrolidinone, 100 mM Tris-HCL pH 8.0, 25 mM EDTA, 2M NaCl, 0.5g/l spermidine and 2 % β-mercaptoethanol) and extracted twice with

chloroform before centrifugation. The aqueous layer was retained after each centrifugation step. RNA was precipitated from the pooled aqueous supernatants by adding LiCl to 10 M final concentration and incubating overnight at -20°C. After centrifugation, the pellet of precipitated RNA was re-suspending in a minimal volume of SSTE buffer (1M NaCl, 0.5% sodium dodecyl sulphate (SDS), 10 mM Ttris-HCl, pH 8.0, 1mM EDTA, sodium form pH 8.0) prior to further clean-up with phenol/chloroform, pH8 (PC8) and pure chloroform extraction steps. Following the final centrifugation and recovery of the aqueous layer, RNA was precipitated using ethanol, re-suspended in water and treated with RNAse-free DNAse (Turbo DNA-Free, Ambion, Austin, TX). Complementary DNAs were synthesized using reagents from a Superscript^{TM} Indirect labeling kit (Invitrogen, Carlsbad, CA). The resultant cDNAs were purified using a DNA Clean and Concentrate kit (Zymo Research Corp., Irvine, CA) and final DNA concentration was checked using a spectrophotometer.

3.3.3.2 Real-time quantitative PCR

For quantitative real-time PCR assays, primers were designed using Primer 3 v.0.4.0 (Table I). Gene-specific primer pairs were developed for PtPAL1, PtPAL2 and PtPAL3, PtPAL4, but we were unable to develop primers for PtPAL5 that could reliably distinguish the nearly identical PtPAL4 and PtPAL5 gene products. Primers for the ACT1 and GAPDH house-keeping genes identified in a previous project (Nairn, et al., 2008) were used as controls in this study (Table II). PCR reactions were performed in a 20 µl reaction volume containing 5 µl of 0.5 µM primer pair, 5 µl of 1 ng/µl cDNA template, and 10 µl of SybrGreen 2X Supermix (BioRad, Hercules, CA). Thermal cycling reactions were performed in an iCycler (BioRad) under the following conditions: 95° C for 30s, 65° for 40s, and 78° for 20s followed by 95° for 1 min, 55° for 1 min, and 81 cycles of 55°C for 10 s in 0.5° degree increments to assess product purity via melting

curve analysis. Three technical replicates were performed for each sample, and runs were considered acceptable when the replicates showed less than 0.5% relative standard deviation.

Biological replicates were not available for the samples used to examine tissue-specific expression of PAL genes. For these samples, two separate RNA extraction and cDNA synthesis procedures were performed. Three technical replicates were run for each of these two samples, and the reported expression levels for each tissue were the average for all six replicates. For the time-series experiments to examine the response of PAL genes to abiotic and biotic stresses, three independent biological replicates (shoot tips) from each tree genotype were generated for each treatment, and three technical replicates were run for each of these samples. In the final analyses the quantitative results were pooled for each treatment with the mean of the technical replicates used to calculate the mean and error values for each of the biological replicates.

3.4 Results

3.4.1 Genomic sequence analysis

Genomic sequences for all five PAL genes were identified using transcribed sequence to seed BLAST analyses of the *P. taeda* reference genome sequence (ver 0.9). Alignments of the transcribed coding sequences to the genomic sequences showed >99% identity for all five genes. None of the five loblolly pine PAL genes contained introns, in contrast to angiosperm PAL genes, which generally contain a single intron (Cramer, et al., 1989).

The 5'-flanking regions of four PtPAL genes (PtPAL1, PtPAL2, PtPAL4, and PtPAL5) were analyzed for regulatory sequence motifs that might give insight into possible physiological functions for individual gene family members. Putative TATA and CAAT box motifs, which are critical for transcription initiation, were found in the promoter regions of all four genes.

Predicted *cis*-elements common to all four promoters are listed in Table 3.1. Motifs linked to tissue-specific expression in mesophyll, pollen, root, and endosperm were identified, as were elements associated with response to plant hormones such as abscisic acid (ABA) (Seo and Koshiba, 2002). Additional elements associated with light responsiveness (GATA boxes, G-box, I box) (Giuliano, et al., 1988; Gilmartin, et al., 1990), water stress (MYB, MYC, ACGT-motif) (Abe, et al., 1997), and pathogen response (GT1, W-box) (Zhou, 1999; Eulgem, et al., 2000), and response to sulfur deficiency (SURE) (Maruyama-Nakashita, et al., 2005) were also present.

Motifs associated with calcium signaling (CGCG box) (Yang and Poovaiah, 2002) and phenylpropanoid biosynthesis (Box-L) (Logemann, et al., 1995) were detected in PtPAL1, PtPAL4 and PtPAL5. Copper-responsive elements (CURE) (Quinn and Merchant, 1995) were identified in the promoter regions of PtPAL1, PtPAL2 and PtPAL4. The PtPAL1 and PtPAL2 promoters contained CCAT box motifs that have been associated heat shock responses (Rieping and Schöffl, 1992), ARE motifs associated with anaerobic induction (Olive, et al., 1990), and the Dof transcription factor target (TAAAG) (Yanagisawa, 2000). The PtPAL1 and PtPAL4 promoters (Figure 3.1) shared the well conserved AC elements associated with lignin biosynthesis in vascular cells (Kawaoka, et al., 2000), low temperature-responsive elements (LTRE) (Dunn, et al., 1998) , TCA-elements associated with salicylic acid response (Goldsbrough, et al., 1993), ethylene responsive elements (EREs) (Itzhaki, et al., 1994), and ARR1-binding sites(Ross, et al., 2004) . The PtPAL4 and PtPAL5 promoters (Figure 3.2) both contained motifs associated with induction of gene expression by anaerobic conditions (ANAERO1) (Mohanty, et al., 2005), as well as the presence of pathogen elicitors (BIHD1OS) (Luo, et al., 2005) and UV-B light (BOXLCOREDCPAL) (Maeda, et al., 2005) . A functional motif associated with secondary xylem and wood formation (XYLAT) (Ko, et al., 2006) was

exclusively found in the PtPAL1 promoter, while only the PtPAL2 promoter contained elements associated with blue, white or UV light-induced gene expression (-10 promoter element) (Thum, et al., 2001) as well as response to auxin, salicylic acid, abscisic acid (AS1-binding site, DPBF site) or methyl jasmonate (TGACG-motif). The PtPAL5 promoter also contained a putative sugar-repressive element (SRE) motif(Tatematsu, et al., 2005) .

3.4.2 PAL gene expression

3.4.2.1 Tissue-specific expression

Expression levels of PtPAL genes was measured in a variety of above ground tissues of loblolly pine using four pairs of PCR primers specific for PtPAL1, PtPAL2, PtPAL3, and PtPAL4. In the juvenile and mature xylem tissues from crown, opposite wood and compression wood two important trends were seen from the relative gene expression (ΔC_T) values (Figure 3.3a). Firstly, PtPAL1 and PtPAL4 were the only gene family members expressed in all four of these tissues, with PtPAL1 always showing higher expression (4-9x) than PtPAL4. Secondly, PtPAL1 expression was highest in compression wood while PtPAL4 dropped to undetectable levels in compression wood. In phloem, apical shoot tips and 1st-year cones (Figure 3.3b), PtPAL2 expression was detected along with PtPAL1 and PtPAL4. PtPAL1 expression levels were consistently less in these tissues than what was observed in xylem tissues. PtPAL2 expression was highest in apical shoot tips, followed by 1st-year cones and phloem, but PtPAL1 remained the most highly expressed PAL gene in these tissues.

PtPAL1, PtPAL2 and PtPAL4 were all expressed in root tissues grown under control conditions as well roots subjected to drought and drought followed by recovery (Figure 3.3c). As was seen in aerial tissues, PtPAL1 was the most highly expressed PAL gene in root tissues under

all conditions, but it showed a 3-fold difference in expression between the drought and drought recovery conditions. Expression levels for PtPAL2 and PtPAL4 were not significantly different between roots that experienced different treatment conditions.

3.4.2.2 Differential gene expression in response to stress

To establish baselines for the stress response experiments, PAL gene expression levels were assessed separately for needle and stem tissues collected from excised shoot tips incubated in water 0, 4, 8, 12, 24 and 36 hr (Figure 3.4). Expression levels for PtPAL1, which showed the highest level of expression of all PAL genes in both tissues at all sampled time points, increased modestly up to 12 hrs post-excision and thereafter decreased in both needles and stems. PtPAL1 gene expression was also generally higher (ca. 2x) in stem tissues compared to needles. PtPAL2 gene expression was essentially undetectable in shoot tip stem tissues.

Water-stress was simulated by incubating excised shoot tips without water for 8, 12, 24 and 36 hrs prior to measuring PAL gene expression levels in the separated needle and stem tissues (Figure 3.5). Expression levels for PtPAL1 increased drastically in needles between 12 and 24 hr, and reached a16-fold increase over controls by 36 hr. Changes in PtPAL1 expression were much more modest in stem tissues with a < 2-fold increase by 36 hr. PtPAL2 was the only other PAL gene that showed detectable changes in expression under these conditions with increased expression in needles at 24 and 36 hr. However, PtPAL3 expression remained undetectable in water-stressed shoot tip stems.

PAL gene expression also responded in shoot tip tissues exposed to *Sirex noctilio* venom (Figure 3.6). PtPAL1 was again the most responsive family member, showing strong increases in needle tissues even after 8 hr. PtPAL2 expression also increased in venom-treated needles by 24

hr. PtPAL3 showed detectable expression levels in stem tissues after venom treatment, but in a reverse pattern to PtPAL1, with highest expression at the earliest sampling points after venom treatment followed by a gradual decline in expression.

3.5 Discussion

Our previous work identified at least five PAL genes expressed in various *P. taeda* tissues (Bagal, et al., 2012), and the recent release of a draft reference genome sequence for this species (Neale, et al., 2014) enabled us to confirm that the complete gene family comprises five members. Unlike angiosperm PAL genes, which generally contain a single intron, none of the *P. taeda* PAL genes contained an intron. This is similar to the intronless PAL genes from *Physcomitrella patens* (EMBL: PP1S500_4V6) and *Selaginella kraussiana* (EMBL: AAW80638.1), and suggests this to be the structure of the ancestral PAL gene.

The reference genome sequence also provided access to the upstream promoter regions for the five *P. taeda* PAL genes, enabling *in silico* identification of potential regulatory elements that could be associated with known PAL function (Song and Wang, 2009; Dong and Shang, 2013) (Table 3.1). Multiple copies of various regulatory-elements were observed at varying distances from the translational start sites in different PAL gene promoters. This observation implies functional redundancy for some regulatory motifs and almost certainly influences the ultimate expression pattern for each gene (Lapidot and Pilpel, 2003; Mehrotra, et al., 2005). The presence of multiple motifs also allows for functional independence of expression during different developmental stages and in response to varied external stimuli (Leyva, et al., 1992). The ubiquitous presence of certain motifs, such as light-responsive elements and Dof core sequences, highlights possible regulatory circuits for the expression of PAL genes that would be

consistent with observations across a wide range of plant species (Song and Wang, 2009). The presence of other motifs in specific subsets of the PAL gene family implies shared functionality or differential interactions in the spatial and temporal programs of gene expression (Wanke and Kolukisaoglu, 2010). Thus, motifs associated with gene expression in response to biotic-abiotic stress (W-box, Box-L, ABRE, SURE, LTRE), in specific cell types (e.g. guard cells), tissues (mesophyll, endosperm, roots), and organs (root nodule, pollen) suggest tissues and conditions that should be sampled for specific PAL gene expression. Studies of PAL genes in other plants have revealed differential expression of specific family members under some of these conditions (Cochrane, et al., 2004; Paolis, et al., 2008).

Tissue-specific expression of PtPAL gene family members was assessed in eight tissues from a mature pine tree using a qPCR approach. In the four different xylem tissues examined (Figure 3.3a) only two genes (PtPAL1 and PtPAL4) were expressed. In all cases, PtPAL1 showed far higher rates of expression compared to PtPAL4. In fact, PtPAL1 reached its highest levels of expression in compression wood xylem tissue while PtPAL4 expression dropped to undetectable levels in this tissue. This observation correlates with the increased levels of lignin deposited in compression wood (Timell, 1986; Bevan, et al., 1989; Whetten, et al., 2001). Besides providing mechanical support in tissues like compression wood, lignin is also used as a structural barrier to water loss. Accordingly, PtPAL1 was expressed at high levels in roots undergoing drought stress (Figure 3.3b), but expression was relatively low in roots that were recovering under well-hydrated conditions after drought stress. Also, the AC elements involved in regulation of genes associated with lignin biosynthesis were present exclusively in PtPAL1 and PtPAL4 (Patzlaff, et al., 2003). Taken together, this data supports an important role for PtPAL1 in lignin biosynthesis, as was previously surmised by Whetten and Sederoff (1992).

However, PtPAL1 expression was detected in variety of other tissues, including phloem, apical tips 1st-year cones, and root tissues, where lignin is present, but at highly varied levels during plant development (Yong, et al., 2011). This suggests that PtPAL1 expression is not strictly associated with lignification, but that it is a default form of the enzyme that provides cinnamic acid precursors for a variety of downstream pathways.

PtPAL2 expression was not detected in xylem tissues, but its expression was second only to PtPAL1 in phloem, apical tips, 1st-year cones and root tissues. Increased expression of PAL in apical meristem tissues has been reported in other species responding to changes in light conditions or mechanical injury (Liang, et al., 1989), and the presence of -10 promoter elements in the PtPAL2 promoter is consistent with an expectation for light-driven gene expression (Thum, et al., 2001). The aerial location of tissues sampled from mature field-grown *P. taeda* suggests that PtPAL2 expression may be associated with synthesis of flavonoids for protection against UV light (Schnitzler, et al., 1997) or pathogens (Adomas, et al., 2007).

PtPAL3 expression was only detected under a couple of conditions and, then, only at very low levels. However, the motif analyses of PtPAL3 promoters identified a number of regulatory elements associated with pollen-specific expression. Male strobili and pine pollen were not examined in this study, but the promoter analyses suggest that these tissues would be an appropriate place to look for PtPAL3 expression.

In the separated stems and needles from shoot tips exposed to biotic and abiotic stresses, PtPAL1 was again the most highly expressed member of the gene family. The trends for PtPAL1expression were similar in needles and stems responding to treatment. The expression of PtPAL1 increased over time and reached its highest level at 36hrs. The strong response of

PtPAL1 under stress can again be associated with lignin biosynthesis, which is a common and well-known response to biotic and abiotic stress conditions.

PtPAL2, PtPAL3 and PtPAL4 all showed variable levels of expression in response to the biotic and abiotic stress treatments. Needle-specific expression of PtPAL2 was clear under both control and stress conditions. PtPAL3 expression was detectable in stems and appeared to be responsive to the biotic stress treatment, albeit at very low levels compared to other PtPAL gene family members. Associations between gene family members and specific physiological function were not clear for the treatments under study. Overlapping metabolic networks leading to functional redundancy have previously been discussed with respect to PAL genes in other plants (Cochrane, et al., 2004). Plants have evolved substantial metabolic elasticity to enhance adaptive and protective responses by retaining multiple gene family members (Jung, et al., 2010). Such a mechanism has been proposed for *Arabidopsis* PAL gene family members along with mechanisms for channeling the products from specific PAL gene family members to different downstream biosynthetic pathways (Huang, et al., 2010), and similar mechanisms in *P. taeda* would help to explain the observed PtPAL gene expression patterns.

3.6 Conclusion

This study was an initial attempt to probe the expression patterns of PtPAL genes. Five PtPAL genes were confirmed from the recently announced draft genomic sequence, and *in silico* analysis of the 5' flanking region of each gene highlighted elements that could serve as possible regulators of gene expression. Gene expression in various tissues and under different treatment conditions showed different patterns of expression for the PtPAL family members. PtPAL1 remains the most highly expressed and most responsive member of the gene family under all

conditions. The results suggest overlapping functions or association with different downstream pathways if multiple PtPAL genes are expressed simultaneously in the same cell.

3.7 Funding

This research was supported in part by McIntire-Stennis project GEOZ-0154-MS.

3.8 Acknowledgements

The authors thank Dr. Campbell J. Nairn for providing actin and GAPDH primer pairs for qPCR, and Stephen Pettis for greenhouse services.

3.9 References

Abe, H., Yamaguchi-Shinozaki, K., Urao, T., Iwasaki, T., Hosokawa, D. and Shinozaki, K. (1997) Role of arabidopsis MYC and MYB homologs in drought and abscisic acid-regulated gene expression, *The Plant Cell Online*, **9**, 1859-1868.

Adomas, A., Heller, G., Li, G., Olson, A., Chu, T., Osborne, J., Craig, D., Van Zyl, L., Wolfinger, R., Sederoff, R., Dean, R.A., Stenlid, J., Finlay, R. and Asiegbu, F.O. (2007) Transcript profiling of a conifer pathosystem: response of *Pinus sylvestris* root tissues to pathogen (*Heterobasidion annosum*) invasion, *Tree Physiology*, **27**, 1441-1458.

Altschul, S., Gish, W., Miller, W., Meyers, E. and Lipman, D. (1990) Basic Local Alignment Search Tool, *Journal of Molecular Biology*, **215**, 403 - 410.

Bagal, U.R., Leebens-Mack, J.H., Lorenz, W.W. and Dean, J.F.D. (2012) The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage, *Bmc Genomics*, **13(Supp 3):S1**.

Bevan, M., Shufflebottom, D., Edwards, K., Jefferson, R. and Schuchl, W. (1989) Tissue and cell-specific activity of a phenylalanine ammonia-lyase promoter in transgenic plants, *The EMBO Journal*, **8**, 1899-1906.

Blount, J.W., Korth, K.L., Masoud, S.A., Rasmussen, S., Lamb, C. and Dixon, R.A. (2000) Altering expression of cinnamic acid 4-hydroxylase in transgenic plants provides evidence for a feedback loop at the entry point into the phenylpropanoid pathway, *Plant Physiology*, **122**, 107-116.

Bordeaux, J.M., Lorenz, W.W. and Dean, J.F.D. (2012) Biomarker genes highlight intraspecific and interspecific variations in the response of *Pinus taeda* L. and *Pinus radiate* D. Don to *Sirex noctilio* F. acid gland secretions, *Tree Physiology*, **32**, 1302-1312.

Brennan, M., McLean, J.P., Altaner, C.M., Ralph, J. and Harris, P.J. (2012) Cellulose microfibril angles and cell-wall polymers in different wood types of *Pinus radiata*, *Cellulose*, **19**, 1385-1404.

Chaman, M.E., Copaja, S.V. and Argandona, V.H. (2003) Relationships between salicylic acid content, phenylalanine ammonia-lyase (PAL) activity, and resistance of barley to aphid infestation, *Journal of Agricultural and Food Chemistry*, **51**, 2227-2231.

Chang, J., Luo, J. and He, G. (2009) Regulation of poly-phenols accumulation by combined overexpression/silencing key enzymes of phyenylpropanoid pathway, *Acta Biochimica et Biophysica Sinica*, **41**, 123-130.

Chang, S., Puryear, J. and Cairney, J. (1993) A simple and efficient method for isolating RNA from pine trees, *Plant Molecular Biology Reporter*, **11**, 113-116.

Cochrane, F.C., Davin, L.B. and Lewis, N.G. (2004) The Arabidopsis phenylalanine ammonia lyase gene family: kinetic characterization of the four PAL isoforms, *Phytochemistry*, **65**, 1557-1564.

Costa, M.A., Collins, R.E., Anterola, A.M., Cochrane, F.C., Davin, L.B. and Lewis, N.G. (2003) An in-silico assessment of gene function and organization of the phenylpropanoid pathway metabolic networks in *Arabidopsis thaliana* and limitations thereof, *Phytochemistry*, **64**, 1097-1112.

Cramer, C., Edwards, K., Dron, M., Liang, X., Dildine, S., Bolwell, G.P., Dixon, R., Lamb, C. and Schuch, W. (1989) Phenylalanine ammonia lyase gene organization and structure, *Plant Molecular Biology*, **12**, 367-383.

Cramer, C.L., Ryder, T.B., Bell, J.N. and Lamb, C.J. (1985) Rapid switching of plant gene-expressioninduced by fungal elicitor *Science*, **227**, 1240-1243.

Dixon, R.A., Achnine, L., Kota, P., Liu, C.J., Reddy, M.S.S. and Wang, L.J. (2002) The phenylpropanoid pathway and plant defence - a genomics perspective, *Molecular Plant Pathology*, **3**, 371-390.

Dixon, R.A. and Paiva, N.L. (1995) Stress-induced phenylpropanoid metabolism, *Plant Cell*, **7**, 1085-1097.

Dong, C. and Shang, Q. (2013) Genome-wide characterization of phenylalanine ammonia lyase gene family in watermelon (*Citrullus lanatus*), *Planta*, 1-15.

Douglas, C.J. (1996) Phenylpropanoid metabolism and lignin biosynthesis: from weeds to trees, *Trends in Plant Science*, **1**, 171-178.

Dunn, M.A., White, A.J., Vural, S. and Hughes, M.A. (1998) Identification of promoter elements in a low-temperature-responsive gene (blt4.9) from barley (*Hordeum vulgare* L.). *Plant Molecular Biology*, **38**, 551-564.

Eckert, A.J. and Hall, B.D. (2006) Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): phylogenetic tests of fossil-based hypotheses, *Molecular Phylogenetics and Evolution*, **40**, 166-182.

Edwards, K., Cramer, C.L., Bolwell, G.P., Dixon, R.A., Schuch, W. and Lamb, C.J. (1985) Rapid transient induction of phenylalanine ammonia lyase mRNA in elicitor-treated bean cells, *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 6731-6735.

Eulgem, T., Rushton, P.J., Robatzek, S. and Somssich, I.E. (2000) The WRKY superfamily of plant transcription factors, *Trends in Plant Science*, **5**, 199-206.

Giberti, S., Bertea, C.M., Narayana, R., Maffei, M.E. and Forlani, G. (2012) Two phenylalanine ammonia lyase isoforms are involved in the elicitor-induced response of rice to the fungal pathogen *Magnaporthe oryzae*, *Journal of Plant Physiology*, **169**, 249-254.

Gilmartin, P.M., Sarokin, L., Memelink, J. and Chua, N.H. (1990) Molecular light switches for plant genes, *The Plant Cell*, **2**, 369-378.

Giuliano, G., Pichersky, E., Malik, V.S., Timko, M.P., Scolnik, P.A. and Cashmore, A.R. (1988) An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene, *Proceedings of the National Academy of Sciences*, **85**, 7089-7093.

Goldsbrough, A.P., Albrecht, H. and Stratford, R. (1993) Salicylic acid-inducible binding of a tobacco nuclear protein to a 10 bp sequence which is highly conserved amongst stress-inducible genes, *The Plant Journal*, **3**, 563-571.

Higo, K., Ugawa, Y., Iwamoto, M. and Higo, H. (1998) PLACE: A database of plant *cis*-acting regulatory DNA elements, *Nucleic Acids Research*, **26**, 358-359.

Howles, P.A., Sewalt, V.J., Paiva, N.L., Elkind, Y., Bate, N.J., Lamb, C. and Dixon, R.A. (1996) Over-expression of L-phenylalanine ammonia lyase in transgenic tobacco plants reveals control points for flux into phenylpropanoid biosynthesis, *Plant Physiology*, **112**, 1617-1624.

Huang, J., Gu, M., Lai, Z., Fan, B., Shi, K., Zhou, Y., Yu, J. and Chen, Z. (2010) Functional analysis of the arabidopsis PAL gene family in plant growth, development, and response to environmental stress, *Plant Physiology*, **153**, 1526-1538.

Itzhaki, H., Maxson, J.M. and Woodson, W.R. (1994) An ethylene-responsive enhancer element is involved in the senescence-related expression of the carnation glutathione-S-transferase (GST1) gene., *Proceedings of the National Academy of Sciences*, **91**, 8925-8929.

Jung, K., Cao, P., Seo, Y., Dardick, C. and Ronald, P.C. (2010) The rice kinase phylogenomics database: a guide for systematic analysis of the rice kinase super-family, *Trends in Plant Science*, **15**, 595-599.

Kawaoka, A., Kaothien, P., Yoshida, K., Endo, S., Yamada, K. and Ebinuma, H. (2000) Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis, *The Plant Journal*, **22**, 289-301.

Ko, J., Beers, E. and Han, K. (2006) Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*, *Molecular Genetics and Genomics*, **276**, 517-531.

Kodan, A., Kuroda, H. and Sakai, F. (2002) A stilbene synthase from japanese red pine (*Pinus densiflora*): Implications for phytoalexin accumulation and down-regulation of flavonoid biosynthesis, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 3335-3339.
Kumar, A. and Ellis, B.E. (2001) The phenylalanine ammonia lyase gene family in raspberry. Structure, expression, and evolution, *Plant Physiology*, **127**, 230-239.

Lange, B.M., Lapierre, C. and Sandermann Jr, H. (1995) Elicitor-induced spruce stress lignin (structural similarity to early developmental lignins), *Plant Physiology*, **108**, 1277-1287.

Lapidot, M. and Pilpel, Y. (2003) Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription, *Nucleic Acids Research*, **31**, 3824-3828.

Leyva, A., Liang, X., Pintor-Toro, J.A., Dixon, R.A. and Lamb, C.J. (1992) *C*is-element combinations determine phenylalanine ammonia lyase gene tissue-specific expression patterns, *The Plant Cell Online*, **4**, 263-271.

Liang, X.W., Dron, M., Cramer, C.L., Dixon, R.A. and Lamb, C.J. (1989) Differential regulation of phenylalanine ammonia lyase genes during plant development and by environmental cues, *Journal of Biological Chemistry*, **264**, 14486-14492.

Liang, X.W., Dron, M., Schmid, J., Dixon, R.A. and Lamb, C.J. (1989) Developmental and environmental regulation of a phenylalanine ammonia lyase beta-glucuronidase gene fusion in transgenic tobacco plants, *Proceedings of the National Academy of Sciences*, **86**, 9284-9288.

Logemann, E., Parniske, M. and Hahlbrock, K. (1995) Modes of expression and common structural features of the complete phenylalanine ammonia lyase gene family in parsley, *Proceedings of the National Academy of Sciences*, **92**, 5905-5909.

Lois, R., Dietrich, A., Hahlbrock, K. and Schulz, W. (1989) A phenylalanine ammonia lyase gene from parsley: structure, regulation and identification of elicitor and light responsive *cis*-acting elements, *The EMBO Journal*, **8**, 1641-1648.

Lorenz, W.W., Ayyampalayam, S., Bordeaux, J.M., Howe, G.T., Jermstad, K.D., Neale, D.B., Rogers, D.L. and Dean, J.F.D. (2011) Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species, *Tree Genetics & Genomes*, **8**, 1477-1485.

Lorenz, W.W., Yu, Y.-S. and Dean, J.F.D. (2010) An improved method of RNA isolation from loblolly pine (P. taeda L.) and other conifer species, *Journal of Visualized Experiments*, e1751.

Lorenz, W.W., Yu, Y.-S., Simoes, M. and Dean, J.F.D. (2009) Processing the loblolly pine PtGen2 cDNA microarray, *Journal of Visualized Experiments*, e1182.

Luo, H., Song, F., Goodman, R.M. and Zheng, Z. (2005) Up-regulation of OsBIHD1, a rice gene encoding BELL homeodomain transcriptional factor, in disease resistance responses, *Plant Biology*, **7**, 459-468.

Maeda, K., Kimura, S., Demura, T., Takeda, J. and Ozeki, Y. (2005) DcMYB1 acts as a transcriptional activator of the carrot phenylalanine ammonia lyase gene (DcPAL1) inresponse to elicitor treatment, UV-B irradiation and the dilution effect, *Plant Molecular Biology*, **59**, 739-752.

Maruyama-Nakashita, A., Nakamura, Y., Watanabe-Takahashi, A., Inoue, E., Yamaya, T. and Takahashi, H. (2005) Identification of a novel *cis*-acting element conferring sulfur deficiency response in arabidopsis roots, *The Plant Journal*, **42**, 305-314.

Mehrotra, R., Kiran, K., Chaturvedi, C.P., Ansari, S.A., Lodhi, N., Sawant, S. and Tuli, R. (2005) Effect of copy number and spacing of the ACGT and GT *cis*-elements on transient expression of minimal promoter in plants, *Indian Academy of Sciences*, **84**.

Mohanty, B., Krishnan, S.P.T., Swarup, S. and Bajic, V.B. (2005) Detection and preliminary analysis ofmotifs in promoters of anaerobically induced genes of different plant species, *Annals of Botany*, **96**, 669 - 681.

Möller, R., Koch, G., Nanayakkara, B. and Schmitt, U. (2006) Lignification in cell cultures of *Pinus radiata*: activities of enzymes and lignin topochemistry, *Tree Physiology*, **26**, 201-210.

Moura, J.C.M.S., Bonine, C.A.V., De Oliveira Fernandes Viana, J., Dornelas, M.C. and Mazzafera, P. (2010) Abiotic and biotic stresses andchanges in the lignin content and composition in plants, *Journal of Integrative Plant Biology*, **52**, 360-376.

Nairn, C.J., Lennon, D.M., Wood-Jones, A., Nairn, A.V. and Dean, J.F.D. (2008) Carbohydrate-related genes and cell wall biosynthesis in vascular tissues of loblolly pine (*Pinus taeda*), *Tree Physiology*, **28**, 1099-1110.

Neale, D., Wegrzyn, J., Stevens, K., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Koriabine, M., Holtz-Morris, A., Liechty, J., Martinez-Garcia, P., Vasquez-Gross, H., Lin, B., Zieve, J., Dougherty, W., Fuentes-Soriano, S., Wu, L.-S., Gilbert, D., Marcais, G., Roberts, M., Holt, C., Yandell, M., Davis, J., Smith, K., Dean, J., Lorenz, W., Whetten, R., Sederoff, R., Wheeler, N., McGuire, P., Main, D., Loopstra, C., Mockaitis, K., deJong, P., Yorke, J., Salzberg, S. and Langley, C. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies, *Genome Biology*, **15**, R59.

Olive, M., Walker, J., Singh, K., Dennis, E. and Peacock, W.J. (1990) Functional properties of the anaerobic responsive element of the maize Adh1 gene, *Plant Molecular Biology*, **15**, 593-604.

Paolis, A.D., Pignone, D., Morgese, A. and Sonnante, G. (2008) Characterization and differential expression analysis of artichoke phenylalanine ammonia lyase coding sequences, *Physiologia Plantarum*, **132**, 33-43.

Patzlaff, A., McInnis, S., Courtenay, A., Surman, C., Newman, L.J., Smith, C., Bevan, M.W., Mansfield, S., Whetten, R.W., Sederoff, R.R. and Campbell, M.M. (2003) Characterisation of a pine MYB that regulates lignification, *The Plant Journal*, **36**, 743-754.

Plomion, C., Leprovost, G.g. and Stokes, A. (2001) Wood Formation in Trees, *Plant Physiology*, **127**, 1513-1523.

Quinn, J.M. and Merchant, S. (1995) Two copper-responsive elements associated with the Chlamydomonas Cyc6 gene function as targets for transcriptional activators, *Plant Cell*, **7** 623-628.

Rieping, M. and Schöffl, F. (1992) Synergistic effect of upstream sequences, CCAAT box elements, and HSE sequences for enhanced expression of chimaeric heat shock genes in transgenic tobacco, *Molecular Genetics and Genomics*, **231**, 226-232.

Rombauts, S., Déhais, P., Van Montagu, M. and Rouzé, P. (1999) PlantCARE, a plant *cis*-acting regulatory element database, *Nucleic Acids Research*, **27**, 295-296.

Ross, E.J.H., Stone, J.M., Elowsky, C.G., Arredondo-Peter, R., Klucas, R.V. and Sarath, G. (2004) Activation of the *Oryza sativa* non-symbiotic haemoglobin-2 promoter by the cytokinin-regulated transcription factor, ARR1, *Journal of Experimental Botany*, **55**, 1721-1731.

Schnitzler, J.P., Jungblut, T.P., Feicht, C., Köfferlein, M., Langebartels, C., Heller, W. and Sandermann Jr, H. (1997) UV-B induction of flavonoid biosynthesis in Scots pine (*Pinus sylvestris* L.) seedlings, *Trees*, **11**, 162-168.

Seo, M. and Koshiba, T. (2002) Complex regulation of ABA biosynthesis in plants, *Trends in Plant Science*, **7**, 41-48.

Sewalt, V., Ni, W., Blount, J.W., Jung, H.G., Masoud, S.A., Howles, P.A., Lamb, C. and Dixon, R.A. (1997) Reduced lignin content and altered lignin composition in transgenic tobacco down-regulated in expression of L-phenylalanine ammonia-lyase or cinnamate 4-hydroxylase, *Plant Physiology*, **115**, 41-50.
Song, J. and Wang, Z. (2009) Molecular cloning, expression and characterization of a phenylalanine ammonia lyase gene (SmPAL1) from *Salvia miltiorrhiza*, *Molecular Biology Reports*, **36**, 939-952.

Tatematsu, K., Ward, S., Leyser, O., Kamiya, Y. and Nambara, E. (2005) Identification of *cis*-elements that regulate gene expression during initiation of axillary bud outgrowth in Arabidopsis, *Plant Physiology*, **138**, 757-766.

Thum, K.E., Kim, M., Morishige, D.T., Eibl, C., Koop, H. and Mullet, J. (2001) Analysis of barley chloroplast psbD light-responsive promoter elements in transplastomic tobacco, *Plant Molecular Biology*, **47**, 353-366.

Timell, T.E. (1986) *Compression wood in gymnosperms*. Springer-Verlag, 2150.

Tsai, C., Harding, S.A., Tschaplinski, T.J., Lindroth, R.L. and Yuan, Y. (2006) Genome-wide analysis of the structural genes regulating defense phenylpropanoid metabolism in Populus, *New Phytologist*, **172**, 47-62.

Ververidis, F., Trantas, E., Douglas, C., Vollmer, G., Kretzschmar, G. and Panopoulos, N. (2007) Biotechnology of flavonoids and other phenylpropanoid-derived natural products. Part I: Chemical diversity, impacts on plant biology and human health, *Biotechnology Journal*, **2**, 1214-1234.

Villalobos, D.P., Diaz-Moreno, S.M., Said, E.S.S., Canas, R.A., Osuna, D., Van Kerckhoven, S.H.E., Bautista, R., Claros, M.G., Canovas, F.M. and Canton, F.R. (2012) Reprogramming of gene expression during compression wood formation in pine: Coordinated modulation of S-adenosylmethionine, lignin and lignan related genes, *Bmc Plant Biology*, **12**, 17.

Vogt, T. (2010) Phenylpropanoid Biosynthesis, *Molecular Plant*, **3**, 2-20.

Wang, Y., Chantreau, M., Sibout, R. and Hawkins, S. (2013) Plant cell wall lignification and monolignol metabolism, *Frontiers in Plant Science*, **4**.

Wanke, D. and Kolukisaoglu, H.U. (2010) An update on the ABCC transporter family in plants: many genes, many proteins, but how many functions?, *Plant Biology*, **12 Suppl 1**, 15-25.

Wanner, L.A., Guoqing, L., Ware, D., Somssich, I.E. and Davis, K.R. (1995) The phenylalanine ammonia lyase gene family in *Arabidopsis thaliana*, *Plant Molecular Biology*, **27**, 327-338.

Westing, A.H. (1965) Formation and function of compression wood in gymnosperms, *Botanical Review*, **31**, 381-480.

Whetten, R., Sun, Y.-H., Zhang, Y. and Sederoff, R. (2001) Functional genomics and cell wall biosynthesis in loblolly pine, *Plant Molecular Biology*, **47**, 275-291.

Witzell, J. and Martin, J.A. (2008) Phenolic metabolites in the resistance of northern forest trees to pathogens- past experiences and future prospects, *Canadian Journal of Forest Research*, **38**, 2711-2727.

Yanagisawa, S. (2000) Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize, *The Plant Journal*, **21**, 281-288.

Yang, T. and Poovaiah, B.W. (2002) A calmodulin-binding/CGCG box dna-binding protein family involved in multiple signaling pathways in plants, *Journal of Biological Chemistry*, **277**, 45049-45058.

Yeh, T., Morris, C.R., Goldfarb, B., Chang, H. and Kadla, J.F. (2006) Utilization of polar metabolite profiling in the comparison of juvenile wood and compression wood in loblolly pine (*Pinus taeda*), *Tree Physiology*, **26**, 1497-1503.

Yong, S., Choong, C., Cheong, P., Pang, S., Nor Amalina, R., Harikrishna, J., Mat-Isa, M., Hedley, P., Milne, L., Vaillancourt, R. and Wickneswari, R. (2011) Analysis of ESTs generated from inner bark tissue of an *Acacia auriculiformis* x *Acacia mangium* hybrid, *Tree Genetics and Genomes*, **7**, 143-152.

Yuan, S. and Dean, J.F.D. (2010) Differential responses of the promoters from nearly identical paralogs of loblolly pine (*Pinus taeda* L.) ACC oxidase to biotic and abiotic stresses in transgenic *Arabidopsis thaliana*, *Planta*, **232**, 873-886.

Zhou, D. (1999) Regulatory mechanism of plant gene transcription by GT-elements and GT-factors, *Trends in Plant Science*, **4**, 210-214.

Tables

Table 3.1: Common potential regulatory elements in the 5' flanking regions of *P. taeda* PAL gene family members

***Cis*-element**	**Sequence**	**PtPAL1**	**PtPAL2**	**PtPAL4**	**PtPAL5**	**Function**
ABRE element	ACGTG	4(2)	2(0)	1(0)	1(0)	Abscisic acid (ABA) responsive element
ACGT-motif	ACGT	2(4)	2(4)	1(2)	0(1)	induced expression under early response to dehydration
ARR1-element	NGATT	8(1)	22(14)	10(13)	4(6)	regulator of the cytokinin-regulated transcription factor ARR1
CAAT-box	CAAT/CCAAT	13(19)	10(20)	9(11)	0(4)	common cis-acting element in promoter and enhancer regions
CACT-element	YACT	16(15)	20(8)	11(15)	5(2)	cis-element for mesophyll-specific gene expression
DOF-core	AAAG	8(10)	13(1)	23(13)	3(4)	expression of multiple genes involved in carbon metabolism
G-box	CACGTG	3(4)	1(3)	1(1)	1(1)	cis-acting regulatory element involved in light responsiveness
GATA-box	GATA	7(3)	7(6)	7(9)	4(5)	pollen specific gene expression
GT1-motif	GRWAAW/GGTTAA/ GAAAAA	12(7)	6(2)	14(4)	5(6)	light responsive element pathogen responsive and salt induced gene expression
GTGA motif	GTGA	8(7)	5(6)	9(7)	1(5)	pollen specific gene expression
I box	GATAAG/GATAA	1(1)	3(0)	1(3)	2(3)	light responsive element
MYB	AACGG/CCWACC/ CNGTTR/GGATA	8(12)	9(11)	8(3)	3(2)	water stress responsive and flavonoids synthesis
MYC	CANNTG/CATGTG	9(9)	10(10)	5(5)	3(3)	early dehydration responsive element
OSE1_root nodule	AAGAT	1(0)	2(0)	2(1)	1(0)	organ-specific (infected cells of root nodules) element
OSE2_root nodule	CTCTT	2(2)	4(3)	3(0)	1(1)	organ-specific (infected cells of root nodules) element
POLLEN1LELAT52	AGAAA	4(5)	4(4)	11(4)	2(0)	pollen specific activation
Root motif	ATATT	5(6)	6(7)	4(7)	8(4)	Tissue specific(root) gene expression
Skn1-motif	GTCAT	1(4)	2(0)	5(0)	0(1)	Tissue specific (endosperm) expression
SURE element	GAGAC	1(2)	1(2)	2(1)	1(0)	sulfur-responsive element (SURE)
TATA box	TTATTT/TATTTAA/ TTTATATA	2(6)	4(2)	6(7)	4(7)	core promoter element
W box	TTGAC/TGAC	14(1)	5(2)	4(7)	2(2)	WRKY-protein binding site, disease resistance and fungal elicitor responsive element

** Values in brackets () are motif counts on the complementary strand

Table 3.2: Primer sequences used for quantitative RT-PCR

No.	Primer Name	Primer Sequence(5' ->3')
1	PtPAL1	5':CAAGAACGCAGAAGGTGAGAAG
		3':GGCTGGTCCCTTTGTCATAAC
2	PtPAL2	5':AGCATTGGAAACGGCAGGA
		3':CGCGAGCTGTGTTCATGCTA
3	PtPAL3	5':ACGCGCTTATGCTCCAACTC
		3':GGATTCAGCATTCCCGTCTG
4	PtPAL4	5':AGTGCCTTGAGCGATGGAAC
		3':CACGCAGCCAAACAACAT
5	ACT 1	5': AATGGTCAAGGCTGGATTTG
		3': AGGGCGACCAACAATACTTG
6	GAPDH	5': GAGGTTGGCGCAGAGTATGT
		3':TGGGCAGATGCTTTCTCTTT

Figures

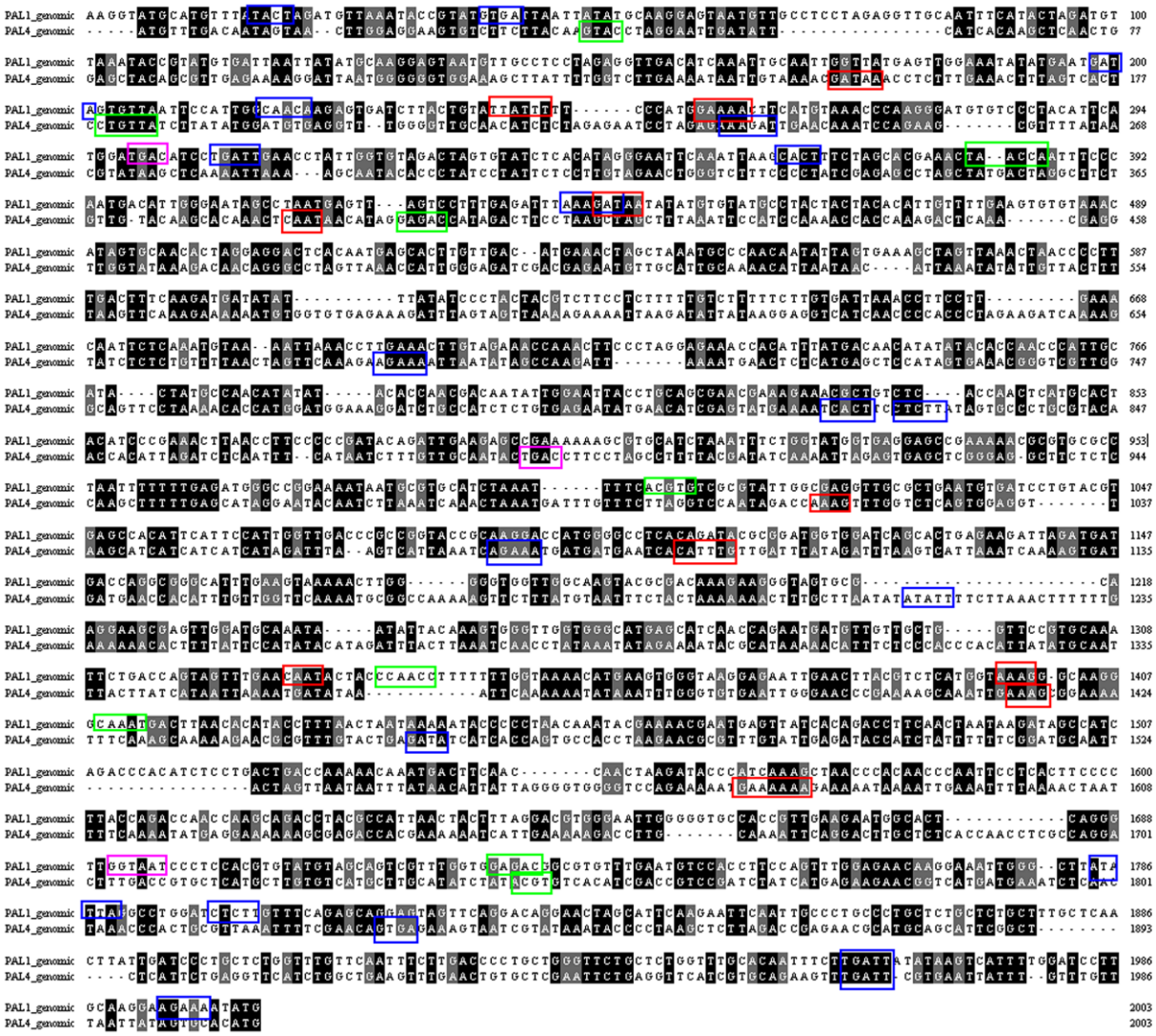


Figure 3.1: Alignment of the PtPAL1 and PtPAL4 promoters and localization of potential *cis*-regulatory elements

Identical nucleotides in both aligned sequences are shown against black boxes, while similar nucleotides are shown against gray boxes. Sequence motifs are highlighted with differently colored boxes enclosing specific classes of elements as follows. **Core elements (red)**: TATA-box, I-box, G-box, DOF core, GT1 consensus, E-box, CAAT motif. **Tissue-specific (blue)**: Root motif, pollen-specific, GTGA motif, GATA motif, CACT motif, ARR1, RAV1. **Pathogen-associated (pink):** W-box, GT1 motif, B1HD10S. E**nvironmental influences (blue):** SURE motif, MYC, MYB, ACGT-motif, ABRE element, Anaero consensus, CURE core element.

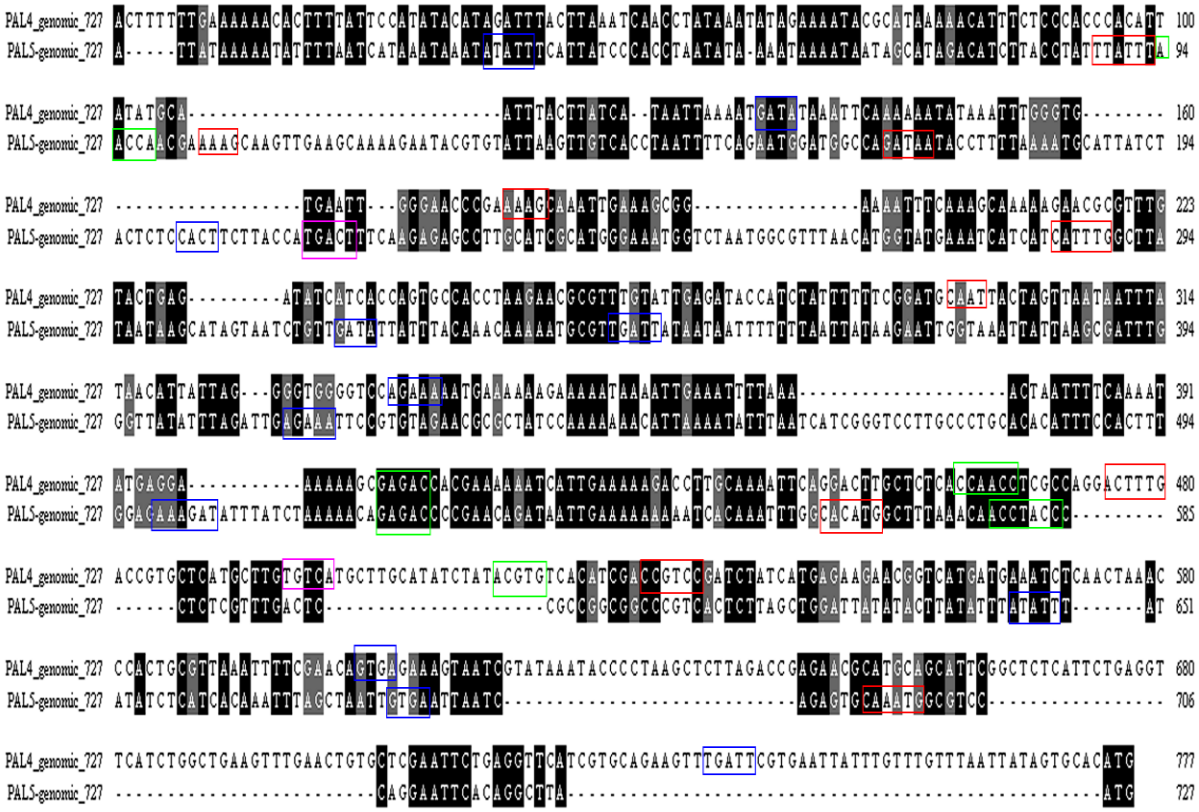
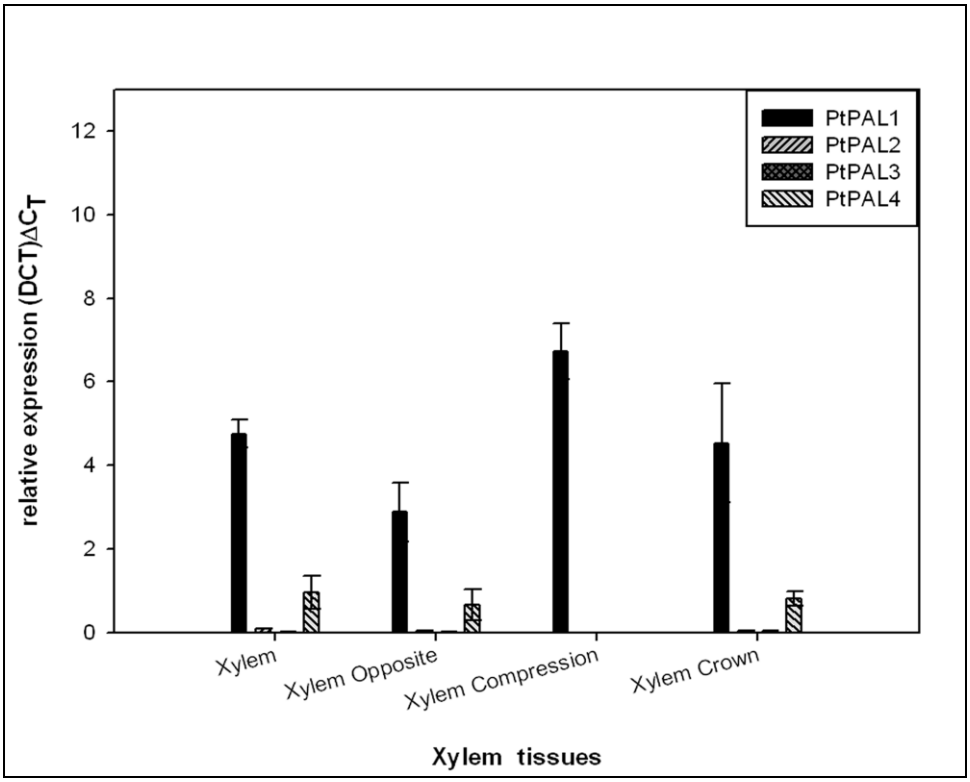


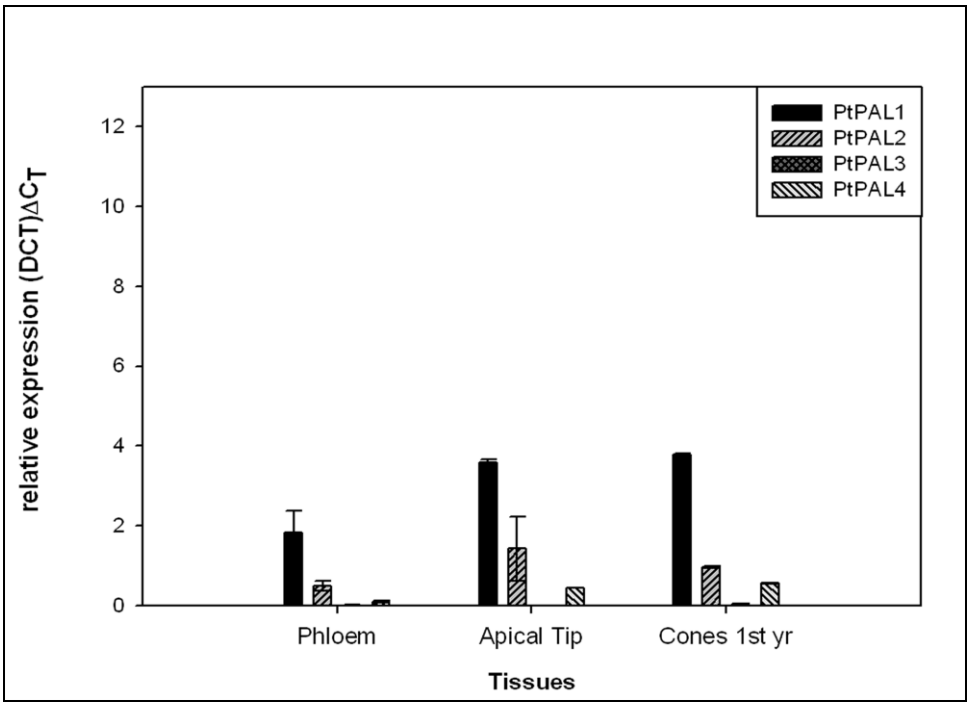
Figure 3.2: Alignment of the PtPAL4 and PtPAL5 promoters and localization of potential *cis*-regulatory elements

Identical nucleotides in both aligned sequences are shown against black boxes, while similar nucleotides are shown against gray boxes. Sequence motifs are highlighted with differently colored boxes enclosing specific classes of elements as follows. **Core elements (red)**: TATA-box, I-box, G-box, DOF core, GT1 consensus, E-box, CAAT motif. **Tissue-specific (blue)**: Root motif, pollen-specific, GTGA motif, GATA motif, CACT motif, ARR1, RAV1. **Pathogen-associated (pink):** W-box, GT1 motif, B1HD10S. E**nvironmental influences (blue):** SURE motif, MYC, MYB, ACGT-motif, ABRE element, Anaero consensus, CURE core element.

[a]


PtPAL1
PtPAL2
PtPAL3
PtPAL4
relative expression (DCT)ΔCT
0
2
4
6
8
10
12
Xylem
Xylem Opposite
Xylem Compression
Xylem Crown
Xylem tissues

[b]


PtPAL1
PtPAL2
PtPAL3
PtPAL4
relative expression (DCT)ΔCT
0
2
4
6
8
10
12
Phloem
Apical Tip
Cones 1st yr
Tissues

[c]

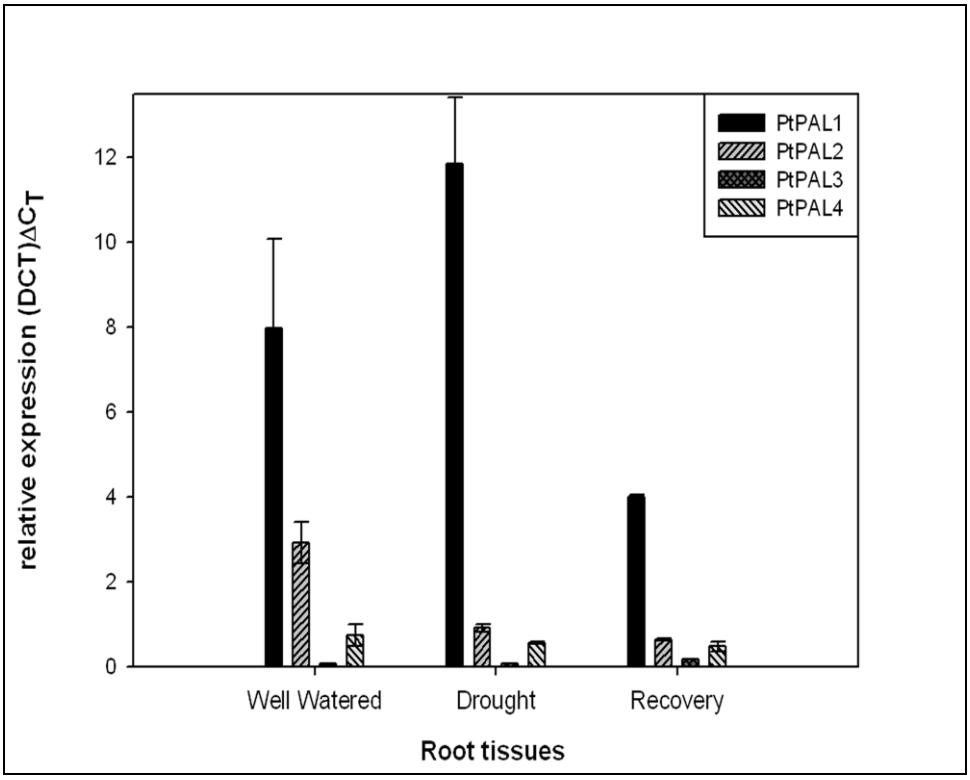


Figure 3.3: Expression of PAL family members in various tissues from a mature *P. taeda* tree

Expression levels of the PtPAL1, PtPAL2, PtPAL3 and PtPAL4 genes were measured by quantitative RT-PCR analyses of xylem tissues (**Panel a**) and other aerial tissues from a mature tree (**Panel b**), as well as root tissues from young trees involved in a water stress experiment (**Panel c**). The ΔC_T values were normalized against expressions of the ACT1 control gene. The reported values represented averages from two replicated extractions of RNA from the same tissues samples each measured in triplicate.

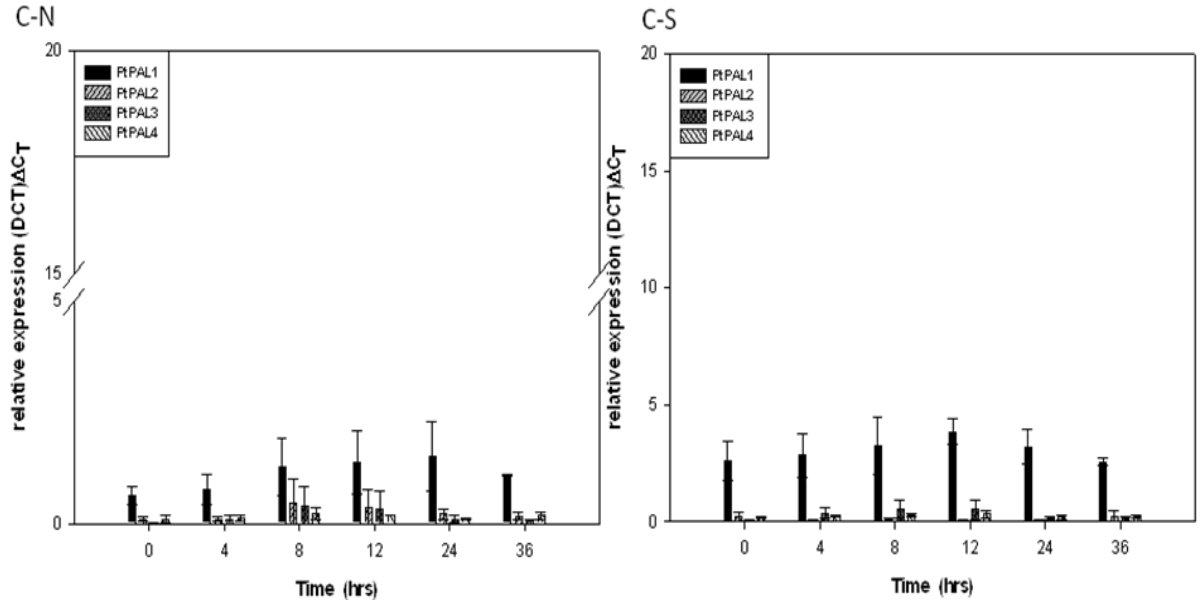


Figure 3.4: Expression of PAL family members in needle (C-N) and stem (C-S) tissues from *P. taeda* shoot tip explants under control conditions.

Quantitative RT-PCR analyses with the four PtPAL primer pair sets was used to measure gene expressions levels in the needle (C-N) and stem (C-S) tissues of *P. taeda* shoot tip explants incubated in water (control). Multiple shoot tip explants taken from three separate trees were collected at different time points. The treatment was run in three biological replicates. Needles and stems were separated and used to prepare RNA qRT-PCR measurements. The ΔC_T values were normalized against the GAPDH gene, which had the most stable expression for these tissues and conditions. Each biological replicate was measured in three technical replicates (n=3). The Y-axis scale for the needles (C_N) values was broken to improve visibility against a full-scale value selected to enable comparisons across the tree treatment experiments.

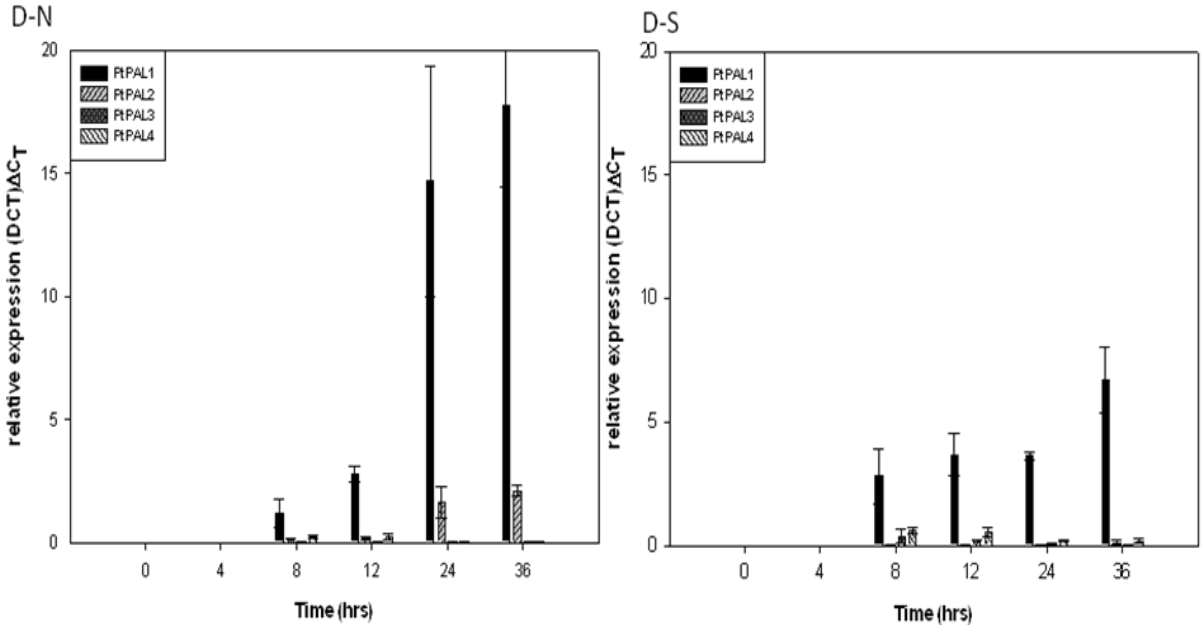


Figure 3.5: Expression of PAL family members in needle (D-N) and stem (D-S) tissues from *P. taeda* shoot tip explants after abiotic stress (drought) treatment

Quantitative RT-PCR analyses with the four PtPAL primer pair sets was used to measure gene expressions levels in the needle (D-N) and stem (D-S) tissues of *P. taeda* shoot tip explants incubated without water (abiotic stress). Multiple shoot tip explants taken from three separate trees were collected at different time points. The treatment was run in three biological replicates. Needles and stems were separated and used to prepare RNA qRT-PCR measurements. The ΔC_T values were normalized against the GAPDH gene, which had the most stable expression for these tissues and conditions. Each biological replicate was measured in three technical replicates (n=3).

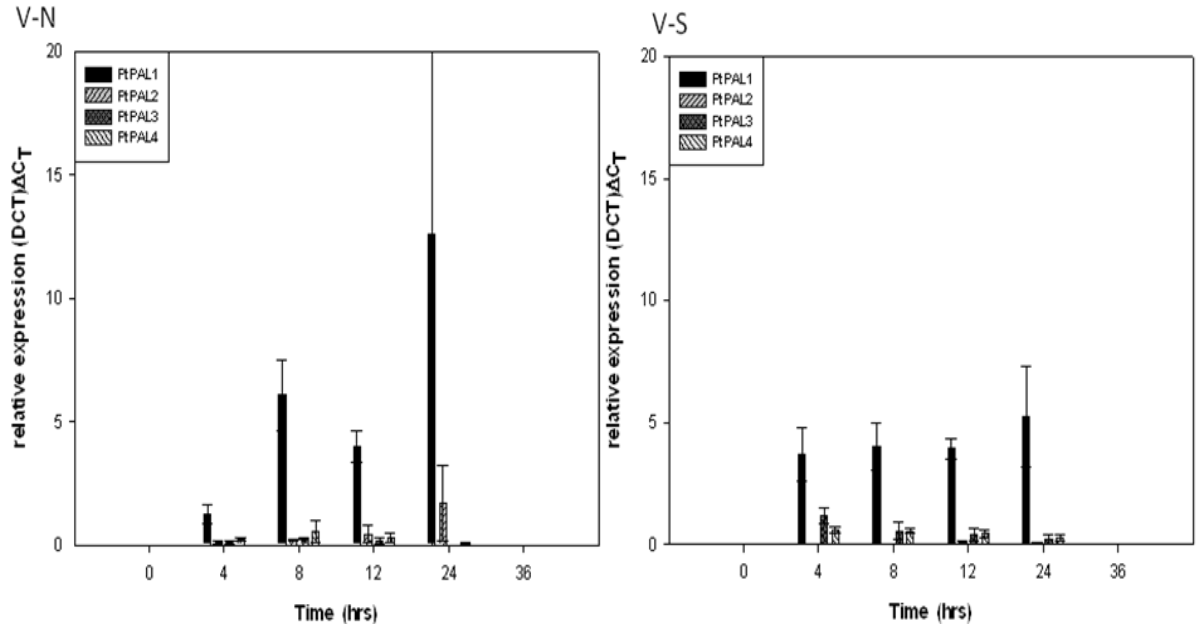


Figure 3.6: Expression of PAL family members in needle (V-N) and stem (V-S) tissues from *P. taeda* shoot tip explants after biotic stress (venom) treatment

Quantitative RT-PCR analyses with the four PtPAL primer pair sets was used to measure gene expressions levels in the needle (V-N) and stem (V-S) tissues of *P. taeda* shoot tip explants incubated in venom-diluted water (biotic stress). Multiple shoot tip explants taken from three separate trees were collected at different time points. The treatment was run in three biological replicates. Needles and stems were separated and used to prepare RNA qRT-PCR measurements. The ΔC_T values were normalized against the GAPDH gene, which had the most stable expression for these tissues and conditions. Each biological replicate was measured in three technical replicates (n=3).

CHAPTER 4

EVOLUTIONARY ANALYSIS OF THE PHENYLALANINE AMMONIA LYASE GENE FAMILY IN GYMNOSPERMS[3]

[3] Ujwal R. Bagal, James H Leebens-Mack, Jeffrey Dean. To be submitted to PLOS One

4.1 Abstract

Starting with the five phenylalanine ammonia lyase gene family members identified in *Pinus taeda*, a study was undertaken to examine the variability amongst gymnosperm PAL coding sequences from a functional perspective. Evolutionary analysis tools using maximum likelihood methods, such as Fitmodel, were used to detect adaptive sites in the protein sequence data. Selection was allowed to change over time in the model in order to detect sites under adaptive selection across all lineages in a gymnosperm phylogeny. Hypotheses were tested using the M3 and M2a site models in conjunction with switching models. A likelihood ratio test (LRT) identified the M3S2 model as most suitable for describing the data in this collection of sequences, but Bayesian methods did not agree and failed to detect sites under positive selection. These results suggested underlying power issues. A model with fewer parameters (M2aS1) was used to boost power in a subsequent analysis and the best fit for the data from that analysis was assessed within a biological context. This evolutionary analysis of gymnosperm PAL genes indicated a heterogeneous evolutionary history with shifting selection constraints associated with duplication events. By pinpointing suggestive sites under adaptive evolution and their location in the protein structure, we provide a framework for future experimental analyses.

4.2 Introduction

Conifers have experienced large environmental and distributional changes during their evolution, dating back to the Mesozoic era (Eckert and Hall, 2006).To adapt to their diverse ecological habitats, as well as the biotic and abiotic stresses associated with them, they have developed multifaceted chemical response systems as part of their biotic resistance strategy (de Laubenfels, 1957). These chemical defense systems produce a wide range of complex secondary metabolites in response to pathogen attack. Many of these secondary products are synthesized via the phenylpropanoid pathway (Adomas, et al., 2007).

The phenylpropanoid pathway has been extensively studied with respect to production of natural products, such as flavonoids, isoflavonoids, hydroxy cinnamic acids, lignin, coumarins, stilbenes and various other phenolic compounds. These products serve diverse functions in plants, including protection against biotic and abiotic stresses, cellular signaling, and UV protection, as well as mechanical support and response to low levels of iron and phosphate (Dixon and Paiva, 1995).

Phenylalanine ammonia lyase (PAL; EC 4.3.1.5), the key enzyme linking primary metabolism of aromatic amino acids with secondary metabolic products in plants, has been extensively studied since its discovery by Koukal and Conn (Koukal and Conn, 1961). PAL plays a key regulatory role in controlling biosynthesis of all phenylpropanoid products. As the entry point into the pathway, PAL catalyses the non-oxidative deamination of phenylalanine to *trans*-cinnamic acid and ammonia. *Trans*-cinnamic acid is the common precursor for the lignin and flavonoids biosynthetic pathways, which are both highly complex and branched pathways (Ritter and Schulz, 2004). *Trans*-cinnamic acid typically undergoes hydroxylation to produce *p*-coumaric acid, catalyzed by cinnamic acid 4-hydrolase (CH4; EC 1.14.13.11). *p*-Coumaric acid

is further transformed into *p*-coumaroyl-CoA by p-coumaroyl: CoA ligase (4CL). Either of the two products, *p*-coumaric acid or *p*-coumaroyl-CoA can enter the lignin monomer pathway, while *p*-coumaroyl-CoA is the preferred precursor of the flavonoids pathway, which produces the greatest diversity of secondary products.

Until now, conifer genomes have been given less attention than angiosperm genomes despite these trees being economically important and dominant in many terrestrial ecosystems (Stefanoviac, et al., 1998). Although little is known about the organization of large conifer genomes, large gene families compared to the angiosperms, have been reported (Perry and Furnier, 1996), suggesting gene duplication may be an important mechanism for genome expansion in conifers. Multigene families have been suggested to be correlated with conifer genome size (Ahuja and Neale, 2005). With the release of large EST datasets for *P. taeda* and other conifers (Lorenz, et al., 2011), phylogenetic methods applied to the assembled conifer transcriptomes have revealed PAL in *P. taeda* as a multi-gene family produced through ancient duplication events (Bagal, et al., 2012).

Gene duplication frequently leads to the development of new functions (*neo-functionalization*), the subdivision of functionality into smaller domains (*sub-functionalization*), or the complete loss of functionality (*non-functionalization*). These processes lead to development of gene families whose members have distinct but related functions, a phenomenon fundamental to adaptive evolution (Okada, et al., 2008). The various duplicated gene copies encoding similar proteins can subsequently follow different fates, with either structural or regulatory divergence leading to different isozymes occupying the same metabolic niche, but differentially regulated with respect to temporal or spatial expression (Ferris, et al., 1979).

Organismal complexity has been linked to such duplication events, which leads to increases in the number of loci performing different functions (Taylor and Raes, 2004).

In this project, under a phylogenetic framework, we have attempted to discern the evolutionary history of the PAL gene family in gymnosperms. We are interested in knowing whether gene duplication has been followed by shifts in gene function and adaptive molecular evolution.

Multispecies datasets are a great help in identifying particular gene sequences under selection because they make it possible to identify specific codon sites under selection, and these can then be tested to detect changes in function linked to these codon substitutions. To detect variation in selection constraints across sites and branches in the PAL gene family phylogeny, codon-based model approaches that allow site-specific selection processes to vary along lineages has been applied (Guindon, 2004). Rate ratios (ω) and switching rates were estimated under the maximum likelihood framework, and this was followed by posterior probabilities estimation in each selection class for assignment of sites to rate ratio classes. Since the pine genome is expected to have been under continuous adaptive selection during its evolution due to changes in environmental conditions (Eckert and Hall, 2006), the above approach should allow us to test the hypothesis that varying selection pressures across the gymnosperm branch were associated with early duplication events. Molecular evolutionary analysis reveals a heterogeneous evolutionary history within the gymnosperm PAL gene family with shifting selection constraints associated with duplication events.

4.3 Materials and Methods

4.3.1 Taxonomic representation and Alignment

PAL sequences were collected from nineteen gymnosperm taxa, namely, *Cupressus atlantica (2), Cephalotaxus harringtonia (4), Ginkgo biloba (2), Gnetum gnemon (1), Picea abies (4), Picea sitchensis (3), Pinus lambertiana (4), Pinus menzeii (4), Pinus palustris (2), Pinus pinaster (2), Pinus sylvestris (1), Pinus taeda (5), Podocarpus macrophyllus (1), Sciadopitys verticillata (3), Sequoia sempervirens (4), Taxus baccata (3), and Wollemia nobilis (3).* Two basal taxa, *Physcometrella patens (2),* which also served as an out-group, and *Selaginella kraussiana (1)* were used for reconstructing a gymnosperm gene tree (Bagal, et al., 2012). An initial multiple sequence alignment for the complete dataset was performed using MAFFT (Katoh, et al., 2002) . Multiple codon alignment corresponding to protein sequences was performed using pal2nal (Suyama, et al., 2006).

4.3.2 Molecular Evolutionary Analysis

To assess molecular evolutionary patterns, likelihood analysis was performed under the nested set of codon-substitution models (M0, M2a, M3, M2a+S1, M2a+S2, M3+S1, M3+S2) implemented in Fitmodel (Guindon, 2004). Fitmodel uses the maximum likelihood method to assess among-site and among-lineage variation in selective constraint without specifying in which branch (es) site-specific shifts in constraint might be expected. Under the M0 model, all sites in the sequence alignment are assumed under the same selection process and, consequently, have a constant dN/dS ratio over all the sites and branches. Under the M2a model, variation in the selective constraints across sites is modeled with rate-ratio classes $\omega<1$, $\omega1=1$ and $\omega2>1$. Under the M3 model variation in the selective constraints across sites is modeled as three rate-

ratio classes with $\omega 1<\omega 2<\omega 3$. Fitmodel implements site-specific shifts between rate ratio classes across the phylogeny. The S1 switching model imposes switching rate classes equally (e.g. shifts from rate class 0 ($\omega 0$) to rate class 2 ($\omega 2$) occur as frequently as shifts from rate class 1 to rate class 2 ($\omega 2$)), while the S2 model allows unequal rates of change from one rate class to another. By incorporating the switching models, only three parameters (overall rate of interchange among rate ratio classes (δ), the coefficient for shifts between $\omega 1$ and $\omega 3(\alpha)$, and the coefficient for shifts between $\omega 2$ and $\omega 3(\beta)$) are added to the models of Yang and Nielson (2002).

Using the ML method, estimates of parameters such as branch length, transition-transversion rate ratio (κ), substitution rate ratios ($\omega 1$, $\omega 2$, $\omega 3$), δ, β, α and their equilibrium frequencies (p1, p2, p3), are acquired. Nested likelihood ratio tests (LTR) were performed to examine whether additional parameters significantly improved the fit of the model to the data. LTR comparisons were made for the basic model assuming no heterogeneity against the model with variation across sites (M0 vs.M3); variation across sites without switching against model with switching rates ratio classes (M3 vs. M3+S1), and lastly, model with equal switching rates against class-dependent switching rates across sites (M3S1 vs.M3S2). Chi-square tests were used to estimate the significance of differences using degrees of freedom equal to the differences in the number of parameters for the models being compared. Posterior probabilities across the tree for assignment of sites to the third-rate class ($\omega 3$) were visualized using BASS (Bayesian Analysis of Selected Sites) (Huelsenbeck and Dyer, 2004).

4.4 Results

4.4.1 Shifting constraints within the gymnosperm PAL family

To investigate the substitution process within the gymnosperm branch of the PAL gene family, the tree constructed using RAxML with the highest log likelihood value (also referred as the 'Best tree') was selected. Likelihood analysis was performed under a nested set of codon-substitution models (M0, M2a, M3, M2a+S1, M2a+S2, M3+S1, M3+S2).

Table 4.1 shows that the log likelihood improved significantly as parameters were added to the nested substitution model [P-value <<0.001; Table 4.1]. The first null hypothesis of a single ω (rate ratio) for all the lineages and sites (M0 vs. M3) was rejected, indicating that the site-specific model was significantly better for the given data.

The second null hypothesis, that the site-specific selection pattern remains constant across lineages under the M3 model, was rejected (P- value << 0.001; Table 4.2). This suggests that site-specific variation amongst the selection classes may have played an important role during the evolution of the gymnosperm PAL gene family. (Note that the M3 model was compared against M3S1 model in this analysis).

A third null hypothesis, that unequal switching between selection categories did not occur during the history of these sequences (α= ß=1), was also tested. For this, the M3S1 model was compared to the M3S2 model. Under the hypothesis that the M3S1 model best described the substitution process, twice the difference between the log likelihoods obtained under the M3S1 and M3S2 models asymptotically followed a 50:50 mixture of the chi-square distribution. The M3S2 model had the largest log likelihood value of all the models, and the LTR test showed it to be statistically significant (p-value = 5.5×10^{-152}) compared to the M3S1 model.

These analyses suggested that the M3S2 codon substitution model provided the best fit for the PAL gene sequences. The substitution rate ratio estimates for the three classes under the M3S2 model were ω1 =0.006, ω2=0.16 and ω3=197.2 (Table 4.1). The switching rate between ω 1 and ω 2 (R12=0.065) was much smaller than ω1 to ω3 (R13=11.37) and ω2 to ω3 (R23=367.57), indicating that the site-specific shifts between moderate purifying selection and relaxed selection occurred more often that between highly constrained classes. However, the proportion of sites within ω3 (0.004) was much less than the proportion within ω1 (0.62) or ω2 (0.38). The ω3 values being > 1 in the M3S2 model strongly suggests that very few sites were subject to relaxed constraints.

4.4.2 Site Analysis

The codon alignments identified 802 codon sites amongst 50 sequences belonging to 17 gymnosperm and 2 basal taxa for site-specific analysis. Gaps were more frequent at the start and the ends of the alignment due in part to the incomplete nature of the EST assemblies, as well as less homology at the N- and C-terminal ends of the protein sequences. The estimated posterior probabilities for placing a site in the highest rate ratio class (ω3) were visualized using BASS (Bayesian Analysis of Selected Sites) (Huelsenbeck and Dyer, 2004). Relaxed selection could be associated with duplication branches within the gymnosperms, and was prominent in the MIO and the shielding domains. These sub-branches within the gymnosperms (Figure 4.1) were shown to be subject to relaxed selection at several sites under the M3S1 and M2aS1 models, but not under the M3S2 or M2aS2 models.

The number of sites shown to be under relaxed substitution fell into specific patterns. In these, certain duplication branches showed amino acid conservation, while others showed variability. For example, the oldest duplication event showed high conservation at sites 74, 103

and 704. A duplication branch after separation of mosses from vascular plants showed high conservation in one of the branches at sites (60, 65, 66, 89, 103, 592, 716, 719, and 738). The same branch showed relaxed constraints at sites 68, 92 and 657.

A second pattern was observed at sites 568, 603 and 658, where amino acids were highly conserved amongst the two gymnosperm clades; one that is the oldest branch (basal taxa) and the other which clusters with the angiosperms. While sites 98, 274, 658, 659, 662, 752 were highly conserved within the gymnosperms in the angiosperm branch, sites 135 and 701 were conserved on one of the duplicate branches within the angiosperm clade. Sites 55, 202 and 700 showed high conservation in one of the duplication branches specific to the gymnosperms. At a few sites under relaxed constraint, specific patterns of substitution associated with duplication events could not be discerned (46, 54, 56, 58, 67, 96, 97, 197, 279, 285, 656, 663, 686, and 708).

Since, both the M3S1 and M3S2 models showed different sites under relaxed constraint, use of a model other than M3 was considered. The M3 model uses more parameters and tests for variability, so testing with models using fewer parameters was done to check whether M3 was not an appropriate model for this application. Use of a simpler model could also help to boost the statistical power of Fitmodel by virtue of the lower number of parameters. The M2a model, which is a test of positive selection, was used on the dataset. Surprisingly, M2aS1 also detected positive selection at most of the sites detected by the M3S1 model. But the M2aS2 model found no relaxed constraint sites in common with M2aS1. Also, the log likelihood value and the likelihood ratio test indicated the M2aS1 results to be more significant than the M2aS2 results.

4.4.3 Location of Adaptive Sites in PAL

To look for structure-function relationships in sites identified as most likely undergoing adaptive evolution, the full-length *P. taeda* PAL coding sequence (Pteda1143311: Genbank: U39792.1) was aligned with the *Pertroselium crispum* PtPAL1 coding sequence whose crystal structure has been determined (Havir and Hanson, 1975). The result showed that residues showing relaxed selection were not concentrated in a single protein domain. The most conserved sequence common to the 47 gymnosperm sequences ran from codon 300 to codon 550.

The sites under relaxed selection constraint detected in the gymnosperm PALs using Fitmodel were mapped onto the 3D crystal structure of the *P. crispum* PAL [PDB ID: 1W27] using Rasmol (Figure 4.2). Based on the domain classification of Ritter and Schulz (Ritter and Schulz, 2004), non-synonymous sites were found within the MIO domain (residue 25 to 261), the core region (262 to 527 and 650 to 716) and the shielding region (528 to 649). Substitution at the N-terminus was not considered due to missing sequences. Most of the sites under relaxed constraint were present on the outer surface of the protein structure. Residues in specific regions of the protein shown to have functional importance did not show relaxed constraint and were highly conserved among all the 50 sequences from 21 species. Although the functional significance of residues under relaxed constraint on the outer surface cannot be ascertained at this stage, they appear to be positioned for interaction with other cellular components.

4.5 Discussion

Phenylalanine ammonia lyase belongs to the lyase class I super-family of enzymes (Ritter and Schulz, 2004). The PAL enzyme, catalyses the non-oxidative deamination of phenylalanine at the entry point to the phenylpropanoid pathway to produce *trans*-cinnamic acid. The phenylpropanoid pathway is important for biosynthesis of a wide range of natural products important for development and defense mechanisms in plants (Oh, et al., 2009). Due to its regulatory role, PAL is a primary control point for the phenylpropanoid pathway, which in part explains the multi-gene families seen for PAL in almost all plants studied to date (Lois, et al., 1989; Wanner, et al., 1995; Kumar and Ellis, 2001; Reichert, et al., 2009). This study is the most extensive evolutionary study so far of the PAL gene family, particularly with respect to conifers.

4.5.1 Site-specific Shifts in the gymnosperm branch of the PAL gene tree

Phylogenetic frameworks are highly favored as molecular evolutionary approaches for relating changes at the amino acid level with changes in function (Huelsenbeck and Dyer, 2004). In the present study, the evolutionary processes that have operated on the gymnosperm branch of the PAL gene tree were examined using the maximum likelihood method of Guindon *et al*. (Guindon, 2004). Overall, the results suggested purifying selection as the most common type of selection. Under the M3S2 model, shifts in selection constraints from moderate purifying selection to positive selection were evident from the switching rate Table 4.1: α (R13) value) from ω1 to ω3 (R13:11.37) as well as from (Table 4.1: ß (R23 value)) ω2 to ω3 class (R23: 367.57). Although the equilibrium frequency of sites in the ω3 rate ratio class (p3: 0.004) seems low, one cannot expect positive selection to occur at a high rate on all sites of the gene (Nielsen and Yang, 1998). The M3S1 model suggested relaxed constraints across specific sites, but due to

low statistical power of the Fitmodel, positive selections were not observed in the M3S2. Still, the substitution rate ratio of ω3 =197 and equilibrium frequency (p3=0.004) in the M3S2 model suggested a few sites with higher rates of non-synonymous substitutions, indicating heterogeneous evolution likely occurred in the PAL gene family history.

Our results indicate that relaxed selection pressure sites are specific to certain region of the coding sequence within gymnosperm PAL gene tree. Branches showing site-specific shifts to reduced constraints were associated with early gene duplication events in gymnosperm evolution. Relaxed selection sites observed in all species suggest that changes at these positions have minimal effect on protein function. But conservation among genes from duplicated branches suggests a different story with respect to expression and function. Although relaxed substitution at the N-terminus of the protein sequence was not considered due to missing sequences, one might expect these sites to be under relaxed constraint since the N-terminus has been proposed to anchor the enzyme to different cell components (Ritter and Schulz, 2004).

Information regarding patterns of conservation and variation among amino acid sites, and specific to duplication events can be used for further experimental studies seeking to decipher their functional importance. In this study, none of the sites under relaxed constraints in any of the models were associated with known active site residues. Amino acid substitutions specific to duplication events have been observed at the remaining sites suggesting that amino acid changes may have arisen early in divergence of the PAL genes and then were fixed separately in the two lineages. Elucidation of the functional importance of changes at these specific sites was not possible in this study.

4.6 Conclusions

Evolutionary analysis of the gymnosperm PAL genes showed codon sites under relaxed constraint associated with ancient duplication events. Such sites with non-synonymous substitutions were concentrated on the outer surface of the protein structure and may have functional importance suggesting interactions with other cellular components.

4.7 References

Adomas, A., Heller, G., Li, G., Olson, A., Chu, T., Osborne, J., Craig, D., Van Zyl, L., Wolfinger, R., Sederoff, R., Dean, R.A., Stenlid, J., Finlay, R. and Asiegbu, F.O. (2007) Transcript profiling of a conifer pathosystem: response of *Pinus sylvestris* root tissues to pathogen (*Heterobasidion annosum*) invasion, *Tree Physiology*, **27**, 1441-1458.

Ahuja, M.R. and Neale, D.B. (2005) Evolution of genome size in conifers, *Silvae Genetica*, **54**, 126-137.

Bagal, U.R., Leebens-Mack, J.H., Lorenz, W.W. and Dean, J.F.D. (2012) The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage, *BMC Genomics*, **13(Supp 3):S1**.

de Laubenfels, D.J. (1957) The status of "Conifers" in vegetation classifications, *Annals of the Association of American Geographers*, **47**, 145-149.

Dixon, R.A. and Paiva, N.L. (1995) Stress-induced phenylpropanoid metabolism, *Plant Cell*, **7**, 1085-1097.

Eckert, A.J. and Hall, B.D. (2006) Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): phylogenetic tests of fossil-based hypotheses, *Molecular Phylogenetics and Evolution*, **40**, 166-182.

Ferris, S.D., Portnoy, S.L. and Whitt, G.S. (1979) The roles of speciation and divergence time in the loss of duplicate gene expression, *Theoretical Population Biology*, **15**, 114-139.

Guindon, S., Rodrigo, A, Dyer, Kelly A., Huelsenbeck, John P. (2004) Modeling the site-specific variation of selection patterns along lineages, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12957-12962.

Havir, E.A. and Hanson, K.R. (1975) L-Phenylalanine ammonia lyase (maize, potato, and *Rhodotorula glutinis*). Studies of the prosthetic group with nitromethane, *Biochemistry*, **14**, 1620-1626.

Huelsenbeck, J.P. and Dyer, K.A. (2004) Bayesian estimation of positively selected sites, *Journal of Molecular Evolution*, **58**, 661-672.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform, *Nucleic Acids Research*, **30**, 3059-3066.

Koukal, J. and Conn, E.E. (1961) The metabolism of aromatic compounds in higher plans. Purification and properties of the phenylalanine deaminase of *Hordeum vulgare*, *The Journal of biological chemistry*, **236**, 2692-2698.

Kumar, A. and Ellis, B.E. (2001) The phenylalanine ammonia lyase gene family in raspberry. Structure, expression, and evolution, *Plant Physiology*, **127**, 230-239.

Lois, R., Dietrich, A., Hahlbrock, K. and Schulz, W. (1989) A phenylalanine ammonia lyase gene from parsley: structure, regulation and identification of elicitor and light responsive *cis*-acting elements., *The EMBO Journal*, **8**, 1641-1648.

Lorenz, W.W., Ayyampalayam, S., Bordeaux, J.M., Howe, G.T., Jermstad, K.D., Neale, D.B., Rogers, D.L. and Dean, J.F.D. (2011) Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species, *Tree Genetics & Genomes*, **8**, 1477-1485.

Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, *Genetics*, **148**, 929-936.

Oh, M.M., Carey, E.E. and Rajashekar, C.B. (2009) Environmental stresses induce health-promoting phytochemicals in lettuce, *Plant Physiology and Biochemistry*, **47**, 578-583.

Okada, T., Mikage, M. and Sekita, S. (2008) Molecular characterization of the phenylalanine ammonia lyase from *Ephedra sinica*, *Biological & Pharmaceutical Bulletin*, **31**, 2194-2199.

Perry, D.J. and Furnier, G.R. (1996) *Pinus banksiana* has at least seven expressed alcohol dehydrogenase genes in two linked groups, *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13020-13023.

Reichert, A.I., He , X.Z. and Dixon, R.A. (2009) Phenylalanine ammonia lyase(PAL) from tobacco(*Nicotiana tabacum)*: characterization of the four tobacco PAL genes and active hetrotetramer *Biochemistry*, **424**, 233-242.

Ritter, H. and Schulz, G.E. (2004) Structural basis for the entrance into the phenylpropanoid metabolism catalyzed by phenylalanine ammonia lyase, *Plant Cell*, **16**, 3426-3436.

Stefanoviac, S., Jager, M., Deutsch, J., Broutin, J. and Masselot, M. (1998) Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences, *American Journal of Botany*, **85**, 688-697.

Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, *Nucleic Acids Research*, **34**, W609-W612.

Taylor, J.S. and Raes, J. (2004) Duplication and Divergence: The Evolution of New Genes and Old Ideas, *Annual Review of Genetics*, **38**, 615-643.

Wanner, L.A., Guoqing, L., Ware, D., Somssich, I.E. and Davis, K.R. (1995) The phenylalanine ammonia lyase gene family in *Arabidopsis thaliana*, *Plant Molecular Biology*, **27**, 327-338.

Tables

Table 4.1: Likelihood analysis of PAL gene sequence data

P1= equilibrium frequency of sites in the ω 1 rate ratio class; P2 = equilibrium frequency of sites in the ω 2 rate ratio class; P3 = equilibrium frequency of sites in the ω3 rate ratio class; ω1 = substitution rate class with dN/dS < 1; ω2 = substitution rate class with dN/dS = 1; ω3= substitution rate class with dN/dS > 1; R12 = Switching rate between ω1 and ω2; R13 = Switching rate between ω1 and ω3; R23 = Switching rate between ω2 and ω3.

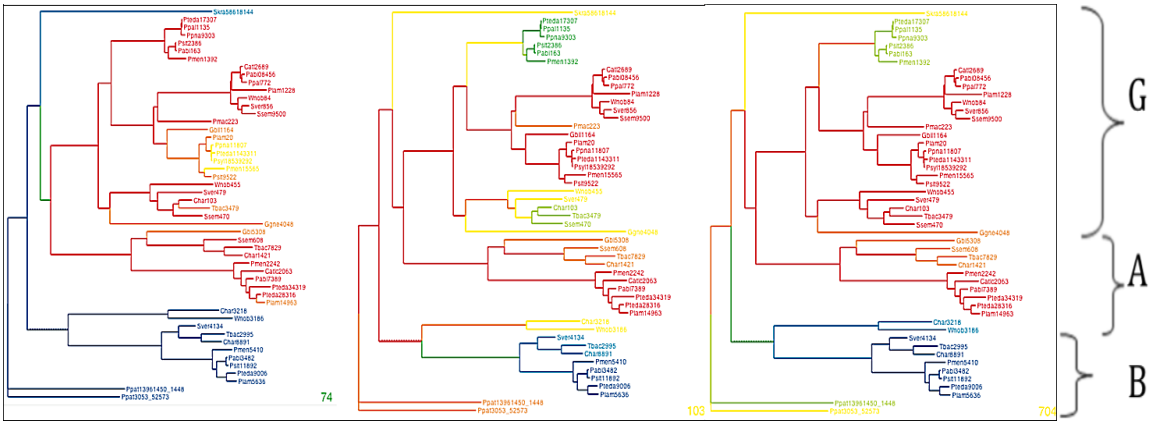
	M0	**M2a**	**M2aS1**	**M2aS2**	**M3**	**M3S1**	**M3S2**
Parameter	1	4	5	7	5	6	8
ln L	-45249.73	-43405.25	-43020.28	-43118.06	-43314.01	-43018.08	**-42674.79**
P1	1.0	0.5215	0.617	0.8836	0.4745	0.606	0.619
P2		0.07953	0.0923	0.1136	0.344	0.283	0.376
P3		0.3989	0.2897	0.0027	0.181	0.109	**0.0037**
ω 1	0.094	0.0152	0.0011	0.01688	0.011	0.0009	0.0057
ω 2		1	1.00	1	0.1240	0.1723	0.1639
ω 3		0.184530	0.190041	99.99	0.43	0.825	**197.22**
δ (R12)			0.2305	0.3313		0.22	0.065
α (R13)			1	26.61		1	11.37
ß (R23)			1	57.26		1	367.566

Table 4.2: Likelihood ratio test (LTR) results for different model comparisons

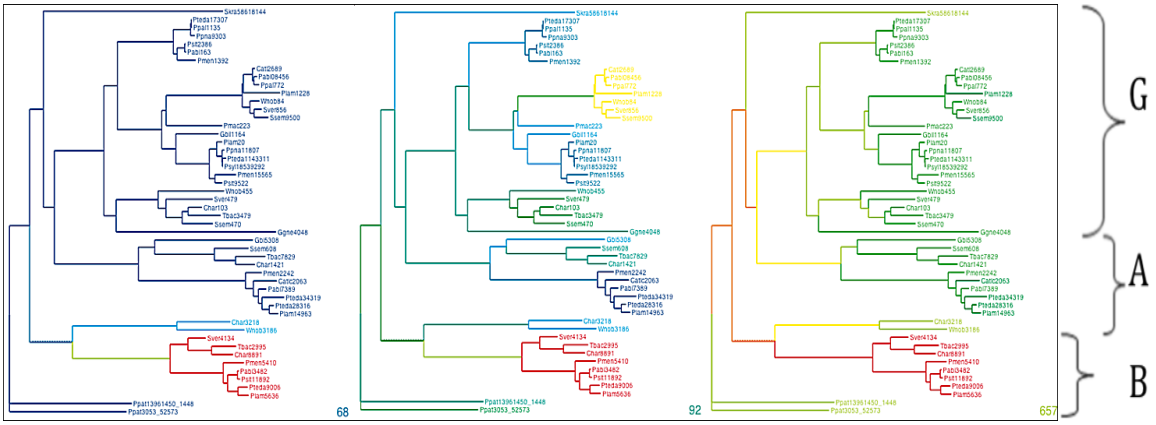
Model	**Test Statistics**	***P* value**
M0 vs. M2a	-3688.962	0 0
M0 vs. M3	-3871.444	0
M2a vs. M2aS1	-769.9296	1.8624E-169
M2aS1 vs. M2aS2	195.5556	1.94874E-44
M3 vs. M3S1	-591.858	9.8803E-131
M3S1 vs. M3S2	-686.58638	**8.122E-150**

Figures

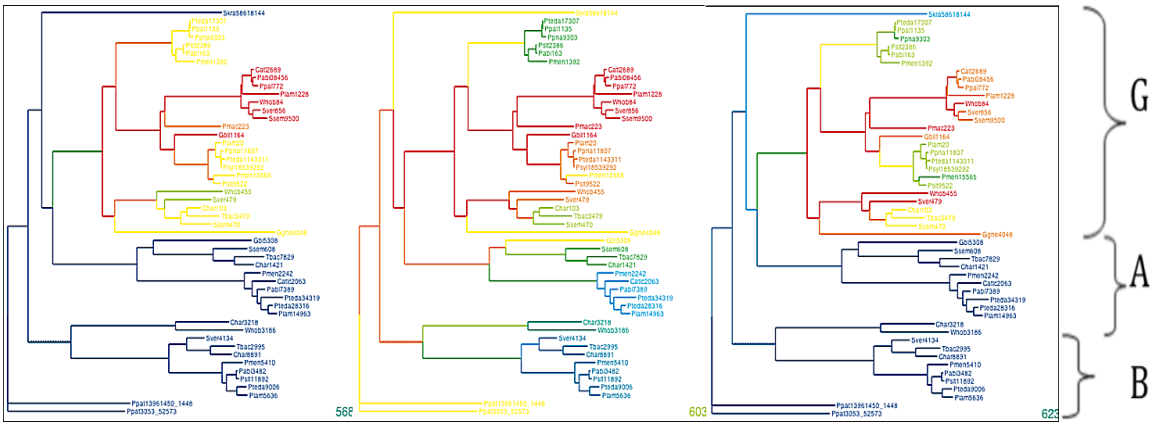
A



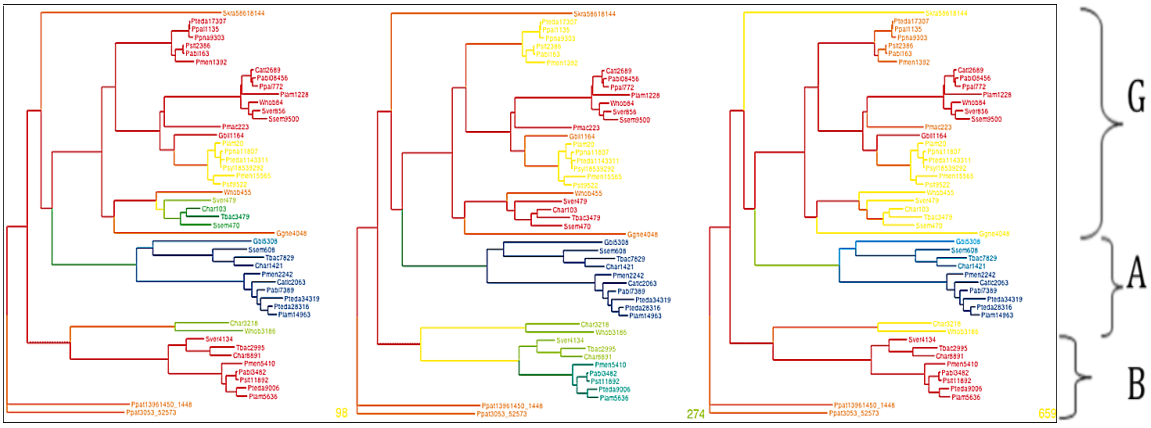
B



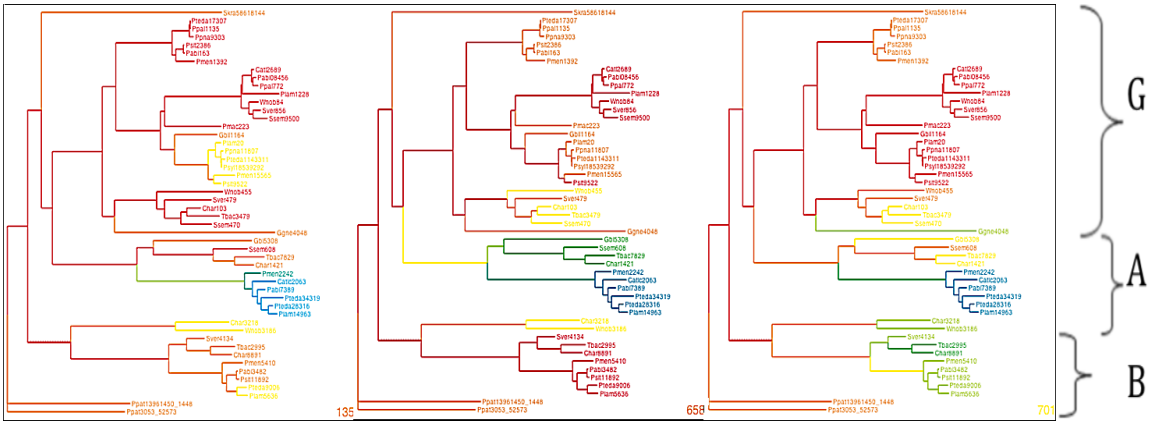
C



D



E



F

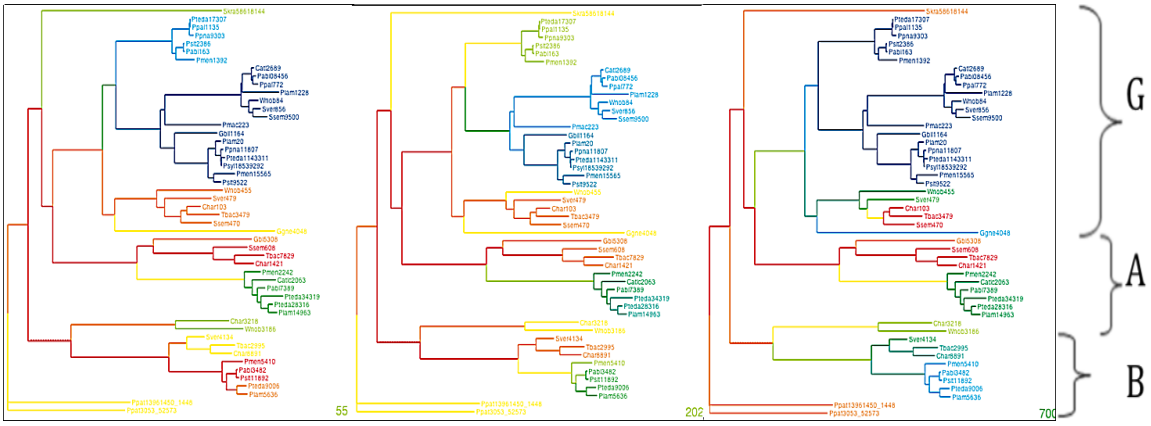


Figure 4.1: Selection regime patterns across select sites on major branches of the PAL gene tree

Figure 4.1: Selection regime patterns across select sites on major branches of the PAL gene tree

Panel A: High conservation pattern for the amino acids in the gymnosperm genes clustered with the basal taxa; **Panel B:** Relaxed selection pattern in the basal taxa branch with a highly constrained selection regime in the gymnosperm and angiosperm branches; **Panel C:** Highly constrained selection pattern in the angiosperm and basal taxa branches; **Panel D:** Constrained selection pattern in the gymnosperm gene sequences clustered with angiosperm sequences; **Panel E:** Conservation among amino acids of the sub-branch of gene sequences that clustered with angiosperm PAL genes; and **Panel F:** highly constrained selection regime within one of the duplicated branches of the gymnosperm branch. Branches under relaxed selection are shown in red, while those under constrained selection are shown in blue. The major branches are indicated by 'G' for gymnosperm-specific, 'A' for angiosperm, and 'B' for basal taxa.

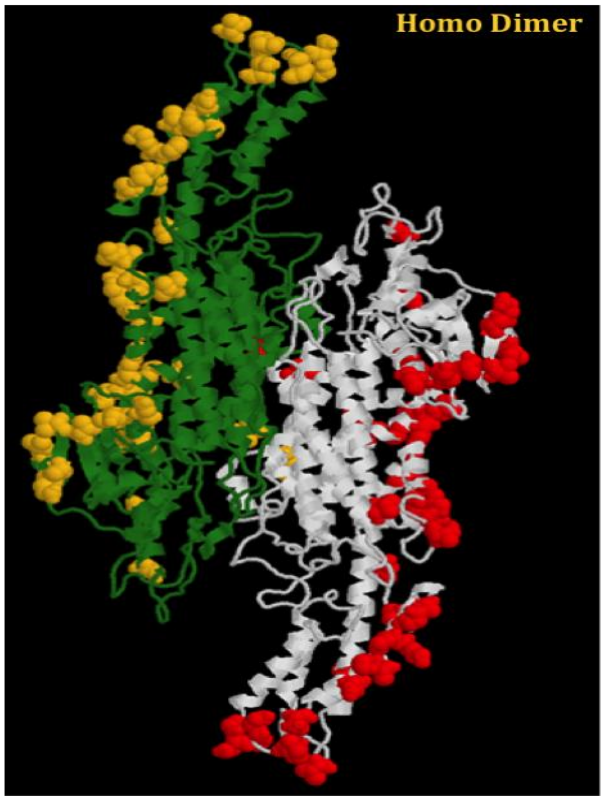


Figure 4.2 Mapping of potential sites under adaptive selection on a PAL 3D structure

Sites (yellow and red balloons) evolving under positive selection predicted by Fitmodel were mapped on a 3D crystallographic structure of *P. crispum* PAL [PDB ID: 1W27] using Rasmol

CHAPTER 5

ESTIMATING RELATIVE POWER AND ACCURACY OF MOLECULAR EVOLUTIONARY ANALYSES OF PROTEIN CODING SEQUENCES [4]

[4] Ujwal R. Bagal, James Leebens-Mack, Jeffrey F. D. Dean. To be submitted to PLOS One

5.1 Abstract

Adaptive evolution at the amino acid level is revealed through surplus non-synonymous substitutions over synonymous substitutions, with $\omega < 1$ indicating purifying selection, $\omega=1$ neutral selection and $\omega >1$ positive selection respectively. Maximum likelihood approaches for detecting shifts in selection processes in protein-coding genes suffers from either simplistic model assumptions or realistic but parameter-rich models. Such inadequacies in turn affect the model efficiency in inferring sites under the correct selection class. In this project, simulated data sets were used to evaluate and compare the efficiencies with which the branch-X-site models from Codeml and the switching models from Fitmodel infer change in mode of selection acting on protein coding sequences. Our results show Codeml to be more powerful than Fitmodel in detecting shifts in selection across datasets. Other factors such as alignment length, timing of shifts in selection and the magnitude of positive selection influenced the overall efficiency of these models. However, the Bayesian approaches used in Codeml and Fitmodel lacks the ability to assign sites to the correct rate ratio class with accurate posterior probabilities.

5.2 Introduction

As next generation sequencing techniques drive DNA sequencing cost down, the rapid accumulation of sequencing data has made possible new computational approaches for testing biological hypotheses. For example, abundant protein-coding gene sequences provide an opportunity to test our understanding of adaptive changes at the codon level and how this leads to shifts in gene functions (Dean and Thornton, 2007).

One approach to studying molecular evolution in protein coding sequences is through investigations of the ratio of non-synonymous to synonymous change, $d_N/d_S = \omega$. This method estimates selection pressure by comparing the rate of substitution at synonymous sites (d_S), which are presumed to be neutral, to the rate of substitution at non-synonymous sites (d_N) which may be influenced by selection acting on the amino acid sequences. Thus, when the d_N/d_S (ω) ratio exceeds unity ($\omega > 1$), the gene is considered to be evolving under positive selection (i.e. changes in amino acid sequences are adaptive). On the other hand, when the d_N/d_S ratio falls below one ($\omega < 1$), the gene is evolving under purifying selection as natural selection suppresses changes in the protein coding sequences (Kimura, 1977; Yang and Bielawski, 2000).

Averaging the d_N and d_S values over sites in a pair-wise comparison of protein coding sequence may reduce power, or the probability of detecting sites evolving under positive selection (Crandall, et al., 1999). In 1994, Goldman and Yang introduced the codon-based model under a maximum likelihood framework which was applicable to a joint comparison of multiple protein coding sequences (Goldman and Yang, 1994). The basic Markov model of codon substitution takes into account transition/transversion ratio as well as codon frequency biases. Under this model all codons in a multiples sequence alignment were modeled within the context

of a gene tree as evolving under purifying selection (ω<1), neutrality (ω=1) or positive selection (ω>1).

Building on this work, three main models namely the Branch models, the Site models and the branch-X-site models developed for testing variation in selection pressure at codon level were implemented in Codeml. The branch model was designed to quantify positive selection pressure along lineages in a phylogeny (Yang 1998). Under this model, various sub-models were developed to deal with variations in the d_N/d_S ratio among lineages in a phylogeny. For example the one-ratio model(M0) assumes a single ω for all the branches of the phylogeny, while the free-ratio model (M1a) allows different ω values for each branch of the phylogeny (Yang, 1998).

The sites model was constructed to model variations in d_N/d_S across codon positions in a multiple sequence alignment of protein-coding genes (Nielsen and Yang, 1998; Yang, et al., 2000). The neutral model (M1a) and the positive selection model (M2a) were based on the rate ratio classes of sites in the gene. The neutral model (M1a), has two site classes, purifying (ω0<1) and neutral (ω1=1), whereas the M2a model, allows some sites to evolve under positive selection: ω0<1, ω1=1, and ω2 >1. The M3 model is similar to the M2a model, but does not constrain some sites to be evolving under neutrality or positive selection: ω0 < ω1 < ω2.

The branch-X-site test models variations in selection across sites and branches in a phylogeny and has been applied extensively to test association between speciation or gene duplication events and subsequent positive selection (Yang and Nielsen, 2002; Zhang, et al., 2005). The test was designed to detect positive selection affecting a few sites along specified lineages in a gene tree (foreground branch). This model explains the evolution of codon sites using four predefined categories. Under the first two categories, the rate ratio (ω) does not vary across branches and a purifying (ω<0) or a neutral selection (ω=1) process applies to all the

branches in the phylogeny. Sites evolving under the third category have few branches known *a priori* (called as the foreground branch) to be evolving under positive selection while the rest of the tree (background branches) evolves under a negative selection. The fourth category allows sites to evolve under positive selection on the foreground branch while neutral selection on the background branches. For the above mixture model, the branches in the tree need to be classified as foreground or background class prior to running the model (Yang and Nielsen, 2002).

In the above models from Codeml, codon frequency can be calculated from the observed nucleotide frequencies while parameters including branch length, the transition/transversion rate ratio, and d_N/d_S ratios are estimated using a maximum likelihood approach. Along with these models, Codeml also provides a Bayes empirical Bayes (BEB) estimation that provides the posterior probability that each site on the foreground branch is evolving under positive selection ($\omega > 1$)(Yang, et al., 2005). All of the models and their associated tests are implemented in the Codeml program within the PAML software package (Yang, 2007).

The branch-X-site test from Codeml though extensively used, it has been criticized for model inadequacies (Nozawa, et al., 2009; Kumar, et al., 2012). With analytical advances and more sequence data becoming available, new methods to detect positive selection among codons in a gene as well as lineages in a phylogeny are being developed (Rodrigue, et al., 2010; Pond, et al., 2011; Murrell, et al., 2012) . One alternative to the Codeml implementation of the branch-X-site model applied in Fitmodel(Guindon, 2004), is based on the covarion model (Tuffley and Steel, 1998) where selection acting on a codon position is allowed to shift from one rate ratio class to another along all the lineages of a phylogeny.

Fitmodel implements the M2a and M3 site models from Codeml as the basic models, which allows sites to evolve under three rate ratio classes ($\omega < 1$), ($\omega = 1$) and ($\omega > 1$), but adds

parameters for the probabilities of sites switching from one rate-ratio class to another along the branches of a gene tree. The switching models are extensions of the widely used traditional codon substitution model (Nielsen and Yang, 1998; Yang, et al., 2000). This leads to the combined use of two Markov models; one for codon state and one for the selection regimes. By adding three additional parameters that measure the rate of change between selection regimes, namely, rate of interchange between selection classes (δ), relative rate of switching from class I to class III ((α) and relative rate of switching from class II to class III (β), Fitmodel accommodates variation in natural selection processes across a tree. Based on these relative rates of switching, two codon models are introduced. The "+S1" model is when the relative rate of switching ($\alpha=\beta$) is equal and the "+S2" model when the rates are allowed to vary freely ($\alpha \neq \beta$).

Both Codeml and Fitmodel implement models under a maximum likelihood framework. Since these model are nested, likelihood ratio test is used to evaluate significant differences between a simpler model in favor of a model with more parameters (Self and Liang, 1987).

For inferring positively selected sites, Codeml uses Bayesian empirical Bayes (BEB) estimation approach, which accommodates uncertainties in the maximum likelihood estimates parameters by assigning priors and averaging over them (Deely and Lindley, 1981). In Fitmodel, the posterior probability for a site being in a particular selection class on a particular branch is estimated from the probability distributions for the nodes at either end of the branch and the average time the site will be in a particular state given the length of the branch (Guindon, 2004).

In general, Fitmodel seems more realistic than the models implemented in Codeml due to its incorporation of switching between selection classes, which allows variation of selection intensities along lineages of the phylogeny (Messier, 1997; Ross and Rodrigo, 2002). However, in Codeml, with fewer parameters estimated from a given data set, it is expected to be more

powerful in detecting sites under positive selection and less computationally costly compared to Fitmodel. Also, in focusing on a particular lineage for detecting shifts in ω, the branch-X-site model implemented in Codeml is expected to be more powerful than models that average shifts over all branches (Messier, 1997; Zhang, et al., 1997).

Although the maximum likelihood method appears successful in real data analysis, the influence of sequence length, the level of sequence divergence, number of taxa and the strength of positive selection over the performance of these methods have also been explored (Anisimova, et al., 2001; Anisimova, et al., 2002; Zhang, 2004). Longer sequences with higher selection pressures always increased the power of the branch-X-site test, while highly divergent or highly similar sequences have shown a negative impact on the ability of the test to correctly reject the Null hypothesis when false and further correctly classify the sites evolving under positive selection.

In this study, computer simulations were conducted to obtain sequences along branches on a phylogeny. These sequences were used to evaluate and compare the efficiency of models from Codeml and Fitmodel for detecting shifts in selection pressures acting on sequences and accurately identifying sites evolving under positive selection.

5.3 Materials and Methods

Five artificial phylogenetic trees with similar topology and number of taxa (eight) were designed for generating sequence data sets. For each tree, a fraction of sites (10%) were allowed to evolve under positive selection. To test sensitivity and accuracy in Codeml as well as the ability of Fitmodel to detect shifts in selection pressure over evolutionary time, for sites from a terminal lineage, an ancient lineage or a combination of both**,** shifts were simulated to take place on specific branches of phylogeny (Fig. 5.1 a-e)**:** tree1 branch "a" (terminal branch); tree 2 branch

“abcdefgh” (tree root); tree 3 branch “a” and “abcdefgh” (tip and root); tree 4 clade (a, b) along with its stem lineage and tree 5 clade (a, b) and (c, d) and their stem lineages.

The objective of these simulations was to evaluate the model efficiencies from Codeml and Fitmodel under different conditions. For example, a comparison of outputs from tree 1 and tree 2 can test the influence of timing of shifts in selection, while a comparison of tree 1 and tree 2 with tree 3 would test more complicated circumstances where shifts were allowed on more than one branch; tree 4 and tree 5 were established to help test the detectibility of sustained shifts.

All data sets were simulated using the evolver program in the PAML package under the alternative branch-X-site model (evolverNSbranchsites) (Yang, 2007). For the selected lineages under positive selection, 10% of the sites were distributed among class ω3 and ω4 with proportions p3=0.07and p4=0.03. Of the remaining 90% of the sites, 60% (p0=0.6) were allowed to evolve under purifying selection with a rate ratio of 0.2 and 30 % (p1=0.3) under neutral selection with a rate ratio of 1.

For each tree, two parameters were varied: the ω values for class 3 and 4 on the foreground branch (2.5 or 5) and the codon alignment length (300 or 600). Further, for trees T1, T2 and T3, four schemes per tree were designed based on 1) codon alignment length (300, 600) and 2) strength of positive selection (ω=2.5/ 5). For trees T4 and T5, codon alignment length was fixed to 600 with a medium rate ratio (ω =2.5) (Table 5.1, column 5). Ten replicates for each of the 14 schemes were simulated for model efficiency detection.

For example, in Table 5.1, under scheme 1 for T1, the rate ratios were ω0=0.2, ω1=1, ω2=2.5 with proportions p0=0.6, p1=0.3, p2=0.07 and p3=0.03 for the foreground branch and ω0=0.2, ω1=1, ω0=0.2, ω1=1 with proportions p0=0.6, p1=0.3, p2=0.07 and p3=0.03 for the

background branches. The alignment length parameter was fixed to 300. In scheme 2, the alignment length parameter was set to 600. Under scheme 3, the parameters for the foreground branch were rate ratios (ω0=0.2, ω1=1, ω2=5) with proportions p0=0.6, p1=0.3, p2=0.07 and p3=0.03) and for the lineages on the background branch the rate ratio were (ω0=0.2, ω1=1, ω0=.2, ω1=1) with proportions p0=0.6, p1=0.3, p2=0.07 and p3=0.03 with alignment length of 300 codons. Under scheme 4, the alignment length was set to 600.

Since the models used in Codeml and Fitmodel are nested, a likelihood ratio test (LRT) is used to evaluate the difference between the competing models. The test statistics is defined as twice the difference between the log likelihood of the two models. If the Null hypothesis is true, $2\Delta l$ will asymptotically follow a χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters.

To determine the performance of the LRT for the models in Codeml and Fitmodel, the data were simulated under the alternative hypothesis (Table 5.1 column 3) with a proportion of sites under positive selection distributed on a single or multiple lineages. This allowed measurement of the number the times the Null hypothesis was rejected correctly when the alternative hypothesis was true. Bayesian posterior probabilities for the placement of each site into a positive selection rate ratio class were also compared with the true rate ratio class for each site.

For replicates of trees T1, T2, and T3, two different Null hypotheses were tested using Codeml, namely the branch-X-site Null model (Null BrNSite) and sites Null model, were applied in order to find if our results were comparable to earlier studies (Zhang, et al., 2005). The branch-X-site Null model is similar to the alternative branch-X-site model except ω2=1 is fixed on the foreground branch. The site Null model, which is the site neutral model (M1a) that

assumes two site classes: conserved sites with $0 < \omega 0 < 1$ and completely neutral sites with $\omega 1 = 1$ for all the branches (Yang and Bielawski, 2000). The significance of the test is expected to be caused by relaxed selection constraint or positive selection on the foreground branch.

Since T4 and T5 had clades in which sites experienced shifts in selection pressure, the M1a (nearly neutral Null model) with two site classes, $\omega 0 < 1$ and $\omega 1 = 1$ (Nielsen and Yang, 1998; Yang, et al., 2005) is compared against the clade model (Model C) (Bielawski and Yang, 2004), which detects divergent selective pressure between clades. The model assumes the tree partitioned into two clades and three site classes with $\omega 0 = 1$, $\omega 1 = 1$ and $\omega 2$, $\omega 3$ for each of the respective clades.

For assessment of Fitmodel, two hypotheses were framed for each set of simulated trees. The Null hypothesis of no switching between selection regimes was tested by comparing the M2a site model against the competing alternative M2aS1 model, which allows for switching among rate ratio classes with a single switching rate for transitions between rate ratio classes (i.e. $\omega_0 \Leftrightarrow \omega_1$, $\omega_0 \Leftrightarrow \omega_2$, and $\omega_1 \Leftrightarrow \omega_2$ are all equal) . A second test was performed with the M2aS1 model as the null hypothesis and the alternative hypothesis, M2aS2, which allowed variation in rates of switching between rate ratio classes.

For Codeml and Fitmodel, output from models that showed statistically significant differences in the log-likelihood ratios were further used to evaluate the accuracy of assignment of sites to the appropriate rate ratio class. For Codeml as well as Fitmodel, we used a 90% posterior probability cut off as used in earlier studies (Anisimova, et al., 2002). An average over all the replicates was used to calculate the power and accuracy of models to correctly identify sites evolving under positive selection.

The definition of the terms ‘accuracy’ and ‘power’ were coined according to earlier study by Anisimova (Anisimova, et al., 2002). Accuracy was calculated as the number of sites correctly identified by the Bayesian method as evolving under positive selection divided by the total number of sites identified to be under positive selection by the method (i.e. true positives and false positives). The false positive rate was calculated as 1-accuracy. Power was calculated as the total number of true positives identified by the Bayesian method divided by the total number of positive sites from evolver. The false negative rate was calculated as 1-power (Table 5.2).

5.4 Results

5.4.1 Likelihood ratio test

To determine the power of the likelihood-ratio test to correctly reject the Null hypothesis in Codeml and Fitmodel, the data were simulated under the alternative hypothesis. When the Null hypothesis was correctly rejected at the $\alpha \leq 5\%$ level in favor of the alternate model the result was considered significant. (Table 5.3 (a-e))

5.4.1.1 Codeml

The data simulated under the alternative hypothesis were tested using two different Null models, the branch-X-site Null model as well as the Sites Null model described earlier. The Sites Null test was less conservative than the branch-X-site Null test. The strength of selection pressure and longer coding sequence showed a positive impact on the power of the LRT (Table 5.3).

Under tree T1 and T2, the Null hypothesis was rejected for all replicates when the strength of selection pressure (ω2) was strong (5) and the complete coding sequence was used

(alignment length = 600) (Figure 5.2a, 5.2b). In case of T3, with two branches under positive selection, there were cases where the Null hypothesis was not rejected (Figure 5.2c).

Under the single branch scheme (T1 and T2), when the branch 'abcdefgh' (root) was assigned as the foreground branch, the Null hypothesis was rejected more often than when branch 'a' (tip) was designated as the foreground branch despite being under medium rate ratio strength and with 300 codons. In the case of T3, the opposite trend was observed for the tip ('a') and root ('abcdefgh') branches when the rate of positive selection (ω2) increased from 2.5 to 5 (Figure 5.2c).

For T4 and T5 the M1a Null model was compared with the clade model C. The Null hypothesis was rejected 70% percent of the time for T4 while for T5 it was only rejected 30% of the time (Figure 5.2 d, e).

5.4.1.2 Fitmodel

Under the tested combinations of codon alignment length (300,600) and strength of positive selection (ω2=2.5/5) in the simulated data sets, LRT seemed to be more conservative for the switching models in Fitmodel and showed less overall power than was seen in Codeml (table 5.3a-e). In the first experiment, where the Null hypothesis of no shift in selection classes was tested, the LRT showed varying levels of performance (from 0-40%). Strength of positive selection, sequence length and position and number of lineages under positive selection effect was clearly observed. For example, of all the trees tested, T1 showed the highest number of times the Null hypothesis was rejected and the rate of rejection increased with increase in ω2 values and codon length suggesting that the LRT was better able to detect sites under positive selection in terminal lineages compared to ancestral lineages as simulated for T2 (Figure 5.2b). For tree T3, where shifts to positive selection were imposed on ancestral and terminal lineages,

the overall number of significant test result was lower than found for trees T1 and T2. Under the tree T3 scenario, high selection strength (ω2 value) rather than longer codon sequence length boosted the power of the M2aS1 test.

In trees T4 and T5, which examined small and large clades under positive selection, the Null hypothesis was rejected 70% and 80% of the times, respectively. As expected the power of the LRT increased with the size of the clade with sites evolving under positive selection.
For the second hypothesis, which allowed variation in switching rates among selection classes ($\alpha \neq \beta$), the null model (M2aS1) was rejected for only 20% of the time showing the inability of the LRT to correctly reject the Null hypothesis (figure 5.2c). Though increase in strength of positive selection and longer sequence length had a positive impact on the power of the LRT, in general, the LRT was very conservative when comparing the +S1 and +S2 switching models. For T4, the Null model was rejected only 30% of the time and this dropped to 10% of the time for T5 suggesting significant power issues (figure 5.2d).

5.4.2 Bayesian Predictions of Sites Evolving under Positive Selection

Overall, the accuracy [TP/ (TP+FP)] of predicting a site under positive selection was better in Codeml than in Fitmodel (Table 5.4 and 5.5). Under Codeml which uses the BEB approach for inferring positively selected sites(Yang, et al., 2005), the combined impact of selection strength (**ω**2), codon sequence length and the position of the lineage under study were observed. Tree T2 showed the least accuracy under all the four combinations, suggesting that the Bayes predictions could not correctly identify sites evolving under positive selection in the ancestral lineages (in this case the root of the tree). Also, strength of selection pressure (ω**)** appeared to be having greater impact than did alignment length. Despite having more than one branch evolving under positive selection pressure, tree T3 outperformed analyses of T1 simulations and showed the

lowest FPR (28%). In the clade model test, the ω2 parameter values showed that the model could pick up the signals of the clade under positive selection in almost all the replicates, but the Bayesian prediction method was unable to identify sites under the influence of positive selection.

In Fitmodel, the accuracy of the Bayesian prediction method under the M2aS1 and M2aS2 model was measured using a posterior probability cutoff of 90%. The M2aS2 model performed poorly in all the 14 trees (*results not shown*). In case of M2aS1, similar to situation in Codeml, positioning of the branch under relaxed selection constraint seems to have the greatest effect followed by strength of selection pressure ω2, and sequence length. In tree T1, the current lineage under positive selection showed improved accuracy with increases in the ω2 value from 2.5 to 5. Sequence length did not greatly matter in this case. Similar to Codeml, tree T2 results exhibited the lowest accuracy suggesting an inability of the Bayesian method to predict sites correctly in the ancestral lineages. Tree T3 simulations gave the highest accuracy and improved with increased selection pressure and codon length. With sites under positive selection distributed amongst more lineages, the accuracy rate for tree T4 and T5 also improved.

In general, Codeml seemed to have more power [TP/ (TP+FN)] than Fitmodel. The power of the Bayesian prediction method under the alternate branch-X-site model seems to depend greatly on level of selection pressure and position of the lineage under positive selection. Relatively little apparent effect came with changes in sequence length as reported previously (Anisimova, et al., 2002). Comparing results for the three trees (T1, T2, and T3) analyses of simulations for tree T2 showed higher power than was found for the others. For tree T1, with increasing positive selection strength (from ω =2.5 to ω =5) acting on branch 'a', power increased substantially (Table 5.4). In contrast, tree T3 showed the least power in all the four categories and consequently increased false negative rates. This suggests that the Bayesian

prediction method shows diminishing power when model assumptions are violated by having sites on the background branches evolving under the effect of positive selection.

Under Fitmodel, where positive selection is predicted on a branch-by-branch basis for each site in an alignment, the strength of positive selection pressure, as well as its presence on the number of branches, seems to influence the power of the Bayesian method. Sequence length as seen in Codeml did not have much effect.

In a single branch evolving under relaxed constraint, the Bayesian prediction method had no power when the strength of positive selection was medium (2.5), but power increased when more than one lineage was simulated as evolving under positive selection. For example, in tree T3, when medium (ω =2.5) positive selection pressure and alignment length of 600 codons were used, the Bayesian prediction showed an increase in power from 0 (in T1 and T2) to 13%. Similar to tree T3, trees T4 and T5 showed higher power with small and large clades under relaxed constraint.

In Codeml, the average FPR for predicting sites under positive selection using the BEB method was close to 50%, which was much better than for Fitmodel under these same conditions. Along with selection strength and sequence alignment length, the effect of lineage position within a tree can be considered as a contributing parameter impacting the rate of false positives. The effect of data deviating from model assumption, where sites in the background branches are allowed to evolve under relaxed constraints was found to lead to high type I and type II errors (Pond, et al., 2011). The FPR for predicting sites with positive selection for tree T3 under a high selection strength (ω = 5) was less than for either T2 or T1, suggesting robustness of BEB method under strong selection strength (Table 5.4 and Figure 5.4a). The false negative rate varied 65% - 93%. High ω values, alignment length and single foreground branch reduced the

FNR for trees T1 and T2, but more than one branch under positive selection in tree T3 increased FNR.

A somewhat reverse effect on FPR was seen in Fitmodel under the M2aS1 switching model. Tree T3 which had more branches under adaptive selection, higher positive selection rate and longer coding sequence also had lower FPR. Similarly, for tree T5 where a large clade was simulated under positive selection, the FPR dropped to 50%.

5.5 Discussion

Fitmodel is an evolutionary analysis tool that applies a codon-based covarion model to detect site-specific selection along lineages (Guindon, 2004). The switching models from Fitmodel seem to be more realistic than the models from Codeml as they allow switches between selection patterns across lineages. Advantages of Fitmodel over the branch-X-site model implemented in Codeml include 1) no requirement to specify *a priori* the branch under positive selection and 2) its ability to identify on a branch by branch basis the selection class under which each site in an alignment evolves. Fitmodel results can be visualized in a graphical format with color codes to highlight the lineage for each site under positive selection (Huelsenbeck and Dyer, 2004; Shan, et al., 2009).

The current investigation included two important aspects: 1) assessment of power of the LRT to correctly detect positive selection affecting few sites on one or more lineages and 2) assessment of the accuracy of the Bayesian assignment method to correctly assign sites to their appropriate selection classes. Because the branch-X-site model has been explored extensively, it provided a good reference for comparison (Nozawa, et al., 2009; Pond, et al., 2011; Yang and dos Reis, 2011).

5.5.1 Likelihood ratio test

Despite the advantages of Fitmodel, due to the low power of the LRT, the efficiency of Fitmodel could not be fully utilized under the constraints imposed by data set sizes included in this study. The M2aS1 switching model was able to predict sites under positive selection ($\omega 3 > 1$) for almost 80% of the total replicates before the likelihood ratio test. This highlights the ability of the model to capture signals for adaptive selection. The performance of the LRT was influenced by the selection strength, longer sequence length, evolutionary timing (tip or root) and number of lineages under positive selection. For clades under positive selection, the power of the LRT increased measurably (Table 5.3d, 5.3e). This suggests the signal of positive selection improves when the total number of sites evolving under positive selection increases. More generally we would expect the performance of Fitmodel to improve with a larger number of taxa on the tree. On the other hand, the parameter-rich M2aS2 switching model suffered from optimization failure issues due to algorithmic complexity. Others have previously suggested that the underlying codon models are also inadequate (Nozawa, et al., 2009; Kumar, et al., 2012). The larger problem seems to be instability due to the large number of parameters. More systematic explorations of these apparent issues are needed to better judge the performance limitations of this model.

Results obtained for the branch-X-site model, agreed with the earlier findings, in which the Null site model made the LRT less conservative than the Null branch-X-site model (Zhang, et al., 2005). Additional factors such as sequence length, strength of positive selection and number of lineages under relaxed selection constraint affected the power of the LRT (Yang and dos Reis, 2011).

5.5.2 Bayesian classification of rate ratio class for each codon

Detecting sites under positive selection can be misleading and is not a simple task, especially when selection regimes are allowed to vary amongst sites as in Fitmodel. Under the conditions explored in our analysis, type I and type II error in both tools (Codeml, Fitmodel) occurred much more frequently that the expected 5%. The cumulative effect of tree topology, branch length, selection strength along with unrealistic model assumptions led to limited utility for detecting sites under selection (Zhang, 2004). All the models under evaluation restricted the number of classes under which a site could be allocated, and this restriction was likely a contributing factor to increasing the false positive rate (Pond, et al., 2011; Yang and dos Reis, 2011). In real data sets, bad alignment, codon bias, and high divergence between sequences also contribute to cause false positives (Nunney and Schuenzel, 2006). In our simulations, tree T3 was constructed to deliberately violate the assumptions of the branch-X-site model by having sites under relaxed selection constraints on the background branches. The alternative hypothesis assumes uniform selective pressure on the background branches, and when this assumption is violated the estimation of ω on foreground branch is biased (Pond, et al., 2011). Another questionable assumption of the models where the synonymous rate is kept unchanged, while the non-synonymous rate is allowed to vary as this leads to different ω values. Thus the low power of the Bayesian method can be attributed at least in part to the presence of high ω values on other branches other than the foreground branch.

Instability resulting from the large number of parameters can contribute to the inefficiency of the Bayesian prediction approach in Fitmodel. Other contributing factors such as codon sequence length, selection pressure and position of the lineage also impacted the ability of Bayesian method to predict sites under positive selection. Within the simulated data schemes,

trees with a clade or multiple branches under positive selection pressure performed better than a single branch under positive selection. More work needs to be done to understand the effective sample sizes, strength of selection pressure and the level of divergence required to increase the performance of Bayesian prediction. Thus in a real dataset of related sequences, we expect Fitmodel to demonstrate better performance than it did on the simulated data.

In our study, prediction of sites evolving under positive selection estimated by Codeml and Fitmodel showed very high FPR and FNR which suggested that application of the Bayesian prediction method is not always efficient. Inadequacies in these models can be judged by the number of times different models have been applied to the same biological data and interpreted differently (Guindon, 2004; Nunney and Schuenzel, 2006; Pond, et al., 2011). In a real data set of bird sequences, the site where relaxed constraint was present was not found by any of the methods, but other sites under neutral selection were falsely detected to be under positive selection (Nozawa, et al., 2009). While these statistical tools are frequently used for generating biological hypotheses for experimental verification, one should be always cautious in interpreting their results.

Reliable detection of positive selection acting on amino acid sequences along lineages within a phylogeny has been a subject of active discussion in the literature (Anisimova, et al., 2001; Anisimova, et al., 2002; Zhang, et al., 2005; Yang and dos Reis, 2011). The goal of this analysis was not to contrast sensitivity and specificity of the branch-X- site model in Codeml and the switching models in Fitmodel, but, rather to: 1) to estimate the efficiency of the switching models in Fitmodel; and 2) provide cautionary advice to researchers about the unreliability of the inferences estimated using these evolutionary analysis tools. As a precautionary measure, we

recommend using of more than one test to identify sites potentially under the influence of adaptive evolution.

5.6 References

Anisimova, M., Bielawski, J.P. and Yang, Z. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution, *Molecular Biology and Evolution*, **18**, 1585-1592.

Anisimova, M., Bielawski, J.P. and Yang, Z. (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection, *Molecular Biology and Evolution*, **19**, 950-958.

Bielawski, J. and Yang, Z. (2004) A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution, *Journal of Molecular Evolution*, **59**, 121-132.

Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, H.C. and Salzman, N.P. (1999) Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection, *Molecular Biology and Evolution*, **16**, 372-382.

Dean, A.M. and Thornton, J.W. (2007) Mechanistic approaches to the study of evolution: the functional synthesis, *Nature Reviews Genetics*, **8**, 675-688.

Deely, J.J. and Lindley, D.V. (1981) Bayes empirical Bayes, *Journal of the American Statistical Association*, **76**, 833-841.

Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Molecular Biology and Evolution*, **11**, 725-736.

Guindon, S., Rodrigo, A, Dyer, Kelly A., Huelsenbeck, John P. (2004) Modeling the site-specific variation of selection patterns along lineages, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12957-12962.

Huelsenbeck, J.P. and Dyer, K.A. (2004) Bayesian estimation of positively selected sites, *Journal of Molecular Evolution*, **58**, 661-672.

Kimura, M. (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution, *Nature*, **267**, 275-276.

Kumar, S., Filipski, A.J., Battistuzzi, F.U., Kosakovsky Pond, S.L. and Tamura, K. (2012) Statistics and truth in phylogenomics, *Molecular Biology and Evolution*, **29**, 457-472.

Messier, W. (1997) Episodic adaptive evolution of primate lysozymes, *Nature*, **385**, 151.

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K. and Kosakovsky Pond, S.L. (2012) Detecting individual sites subject to episodic diversifying selection, *PLoS Genetics*, **8**, e1002764.

Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, *Genetics*, **148**, 929-936.

Nozawa, M., Suzuki, Y. and Nei, M. (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods, *Proceedings of the National Academy of Sciences*, **106**, 6700-6705.

Nunney, L. and Schuenzel, E. (2006) Detecting natural selection at the molecular level: A reexamination of some "classic" examples of adaptive evolution, *Journal of Molecular Evolution*, **62**, 176-195.

Pond, K.S.L., Murrell, B., Fourment, M., Frost, S.D.W., Delport, W. and Scheffler, K. (2011) A random effects branch-site model for detecting episodic diversifying selection, *Molecular Biology and Evolution*, 3033-3043.

Rodrigue, N., Philippe, H. and Lartillot, N. (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles, *Proceedings of the National Academy of Sciences*, **107**, 4629-4634.

Ross, H.A. and Rodrigo, A.G. (2002) Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration, *Journal of Virology*, **76**, 11715-11720.

Self, S.G. and Liang, K. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association*, **82**, 605-610.

Shan, H., Zahn, L., Guindon, S., Wall, P.K., Kong, H., Ma, H., dePamphilis, C.W. and Leebens-Mack, J. (2009) Evolution of plant MADS Box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications, *Molecular Biology and Evolution*, **26**, 2229-2244.

Tuffley, C. and Steel, M. (1998) Modeling the covarion hypothesis of nucleotide substitution, *Mathematical Biosciences*, **147**, 63-91.

Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution, *Molecular Biology and Evolution*, **15**, 568-573.

Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood, *Molecular Biology and Evolution*, **24**, 1586-1591.

Yang, Z. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation, *Trends in Ecology & Evolution*, **15**, 496-503.

Yang, Z. and dos Reis, M. (2011) Statistical properties of the branch-site test of positive selection, *Molecular Biology and Evolution*, 1217-1228.

Yang, Z. and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages, *Molecular Biology and Evolution*, **19**, 908-917.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.-M.K. (2000) Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites, *Genetics*, **155**, 431-449.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.K. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites, *Genetics*, **155**, 431-449.

Yang, Z., Wong, W.S.W. and Nielsen, R. (2005) Bayes empirical Bayes inference of amino acid sites under positive selection, *Molecular Biology and Evolution*, **22**, 1107-1118.

Zhang, J. (2004) Frequent false detection of positive selection by the likelihood method with branch-site models, *Molecular Biology and Evolution*, **21**, 1332-1339.

Zhang, J., Kumar, S. and Nei, M. (1997) Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes, *Molecular Biology and Evolution*, **14**, 1335-1338.

Zhang, J., Nielsen, R. and Yang, Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level, *Molecular Biology and Evolution*, **22**, 2472-2479.

Tables

Table 5.1: Simulation Schemes

Five artificial trees simulated under the alternate branch-X-site model with sites on a single or multiple branches (column 4) under positive selection. Fourteen schemes were designed based on variable selection pressure ω_2 (2.5 /5) and sequence lengths (300/600 codons) (column 5).

Scheme	Tree	Simulation Model	Branch under positive selection	Simulation parameters in the ω Distribution and sequence length as number of codons (L_c)
1	T1	Alt Br & Site	a	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , k =2 $\omega_2 =$ **2.5** , $p_2 =$ 0.07 , $L_c =$ **300**
2	T1	Alt Br & Site	a	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **600** $\omega_2 =$ **2.5** , $p_2 =$ 0.07 , k = 2
3	T1	Alt Br & Site	a	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **300** $\omega_2 =$ **5** , $p_2 =$ 0.07 , k = 2
4	T1	Alt Br & Site	a	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **600** $\omega_2 =$ **5** , $p_2 =$ 0.07 , k = 2
5	T2	Alt Br & Site	beta*	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **300** $\omega_2 =$ **2.5** , $p_2 =$ 0.07 , k = 2
6	T2	Alt Br & Site	beta*	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **600** $\omega_2 =$ **2.5** , $p_2 =$ 0.07 , k = 2
7	T2	Alt Br & Site	beta*	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **300** $\omega_2 =$ **5** , $p_2 =$ 0.07 , k = 2
8	T2	Alt Br & Site	beta*	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **600** $\omega_2 =$ **5** , $p_2 =$ 0.07 , k = 2
9	T3	Alt Br & Site	a and beta*	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **300** $\omega_2 =$ **2.5** , $p_2 =$ 0.07 , k = 2
10	T3	Alt Br & Site	a and beta*	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **600** $\omega_2 =$ **2.5** , $p_2 =$ 0.07 , k = 2
11	T3	Alt Br & Site	a and beta*	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **300** $\omega_2 =$ **5** , $p_2 =$ 0.07 , k = 2
12	T3	Alt Br & Site	a and beta*	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03 $\omega_1 =$ 1 , $p_1 =$ 0.3 , $L_c =$ **600** $\omega_2 =$ **5** , $p_2 =$ 0.07 , k = 2
13	T4	Alt Br & Site	Clade (a, b, ab)	$\omega_0 =$ 0.2 , $p_0 =$ 0.6 , $p_3 =$ 0.03

				$\omega_{1=}$ 1 , $p_{1=}$0.3 , L_c =**600** $\omega_{2=}$ **2.5** , $p_{2=}$0.07 , k = 2
14	T5	Alt Br & Site	Clade (a, b, c, d, ab, cd, abcd)	$\omega_{0=}$ 0.2 , $p_{0=}$0.6 , p_3 = 0.03 $\omega_{1=}$ 1 , $p_{1=}$0.3 , L_c =**600** $\omega_{2=}$ **2.5** , $p_{2=}$0.07 , k = 2

beta*: branch abcdefgh

Table 5.2: Definition of accuracy, power, false positive rate and false negative rate

Evolver	**Predicted sites by Codeml and Fitmodel**		
	Present	**Absent**	**Total**
Positive	TP	FN	(TP+FN)
Non-positive	FP	TN	(FP+TN)
	(TP+FP)	(FN+TN)	(TP+FP+FN+TN)

Accuracy=TP / (TP+FP); Power=TP / (TP+FN); FPR = (1- Accuracy); FNR = (1-Power)

Table 5.3: Percentage of significant tests of Branch-X-site and switching models at 5% level when data simulated with positive selection

Table (**a-c**) shows LRT results for trees T1, T2 and T3 simulated with a single or multiple branches under shifting selection pressure. Two Null hypotheses were tested in Codeml (Null site model with Alternative branch-X-site model and Null branch-X-site model with Alternative branch-X-site model) and in Fitmodel (M2a with M2aS1, M2aS1 with M2aS2). Table (**d-e**) shows LRT results for trees T4 and T5 simulated with small and large clades under positive selection. In Codeml a single null hypothesis of no divergent selection pressure between clades (M1a with Clade model C) is tested while for Fitmodel two hypothesis similar to trees T1, T2 and T3 were tested.

a)

Tree	**Program**	**Tests**	**% of Significant test at 0.05 level**			
			$\omega_{=}$2.5; L_c =300	**$\omega_{=}$2.5; L_c =600**	**$\omega_{=}$5; L_c =300**	**$\omega_{=}$5; L_c =600**
T1	Codeml	NULL Site & Alt BrNSite	50	70	70	100
T1	Codeml	NULL BrNSite & Alt BrNSite	20	20	80	80
T1	Fitmodel	M2a & M2aS1	20	30	30	40
T1	Fitmodel	M2aS1 & M2aS2	0	0	20	40

b)

Tree	**Program**	**Tests**	**% of Significant test at 0.05 level**			
			$\omega_{=}$2.5; L_c =300	**$\omega_{=}$2.5; L_c =600**	**$\omega_{=}$5; L_c =300**	**$\omega_{=}$5; L_c =600**
T2	Codeml	NULL Site & Alt BrNSite	60	90	90	100
T2	Codeml	NULL BrNSite & Alt BrNSite	20	60	80	100
T2	Fitmodel	M2a & M2aS1	0	0	0	10
T2	Fitmodel	M2aS1 & M2aS2	0	20	0	20

c)

Tree	**Program**	**Test**	**% of Significant test at 0.05 level**			
			$\omega_{=}$2.5; L_c =300	**$\omega_{=}$2.5; L_c =600**	**$\omega_{=}$5; L_c =300**	**$\omega_{=}$5; L_c =600**
T3: Br a	Codeml	NULL Site & Alt BrNSite	30	20	10	60
T3: Br a	Codeml	NULL BrNSite & Alt BrNSite	20	10	10	80
T3: Br beta	Codeml	NULL Site & Alt BrNSite	10	40	80	80
T3: Br beta	Codeml	NULL BrNSite & Alt BrNSite	10	30	60	60
T3	Fitmodel	M2a & M2aS1	10	40	10	20
T3	Fitmodel	M2aS1 & M2aS2	0	10	0	20

d)

Tree	**Program**	**Test**	**% of Significant test at 0.05 level**
			$\omega_{=}$2.5; L_c =600
T4	Codeml	M1a & Clade Model C	70
T4	Fitmodel	M2a & M2aS1	70
T4	Fitmodel	M2aS1 & M2aS2	30

e)

Tree	**Program**	**Test**	**% of Significant test at 0.05 level**
			$\omega_{=}$2.5; L_c =600
T5	Codeml	M1a & Clade Model C	30
T5	Fitmodel	M2a & M2aS1	80
T5	Fitmodel	M2aS1 & M2aS2	10

Table 5.4: Performance of Codeml in inferring sites under positive selection in simulated data set

	Tree	ω, L_C	Accuracy	Power	FPR*	FNR**
1	1	2.5, 300	0.43	0.15	0.57	0.85
2	1	2.5, 600	0.53	0.11	0.47	0.89
3	1	5.0, 300	0.54	0.32	0.46	0.68
4	1	5.0, 600	0.60	0.32	0.40	0.68
5	2	2.5, 300	0.27	0.29	0.73	0.71
6	2	2.5, 600	0.14	0.18	0.86	0.82
7	2	5.0, 300	0.39	0.36	0.61	0.64
8	2	5.0, 600	0.42	0.35	0.58	0.65
9	3	2.5, 300	0.35	0.13	0.65	0.87
10	3	2.5, 600	0.60	0.08	0.40	0.92
11	3	5.0, 300	0.63	0.15	0.37	0.85
12	3	5.0, 600	0.72	0.07	0.28	0.93

* FPR=False positive rate; ** FNR=False negative rate

*$L_{C:}$ - Sequence length as number of codons; ω= selection strength

Table 5.5: Performance of Fitmodel in inferring sites under positive selection in simulated data set

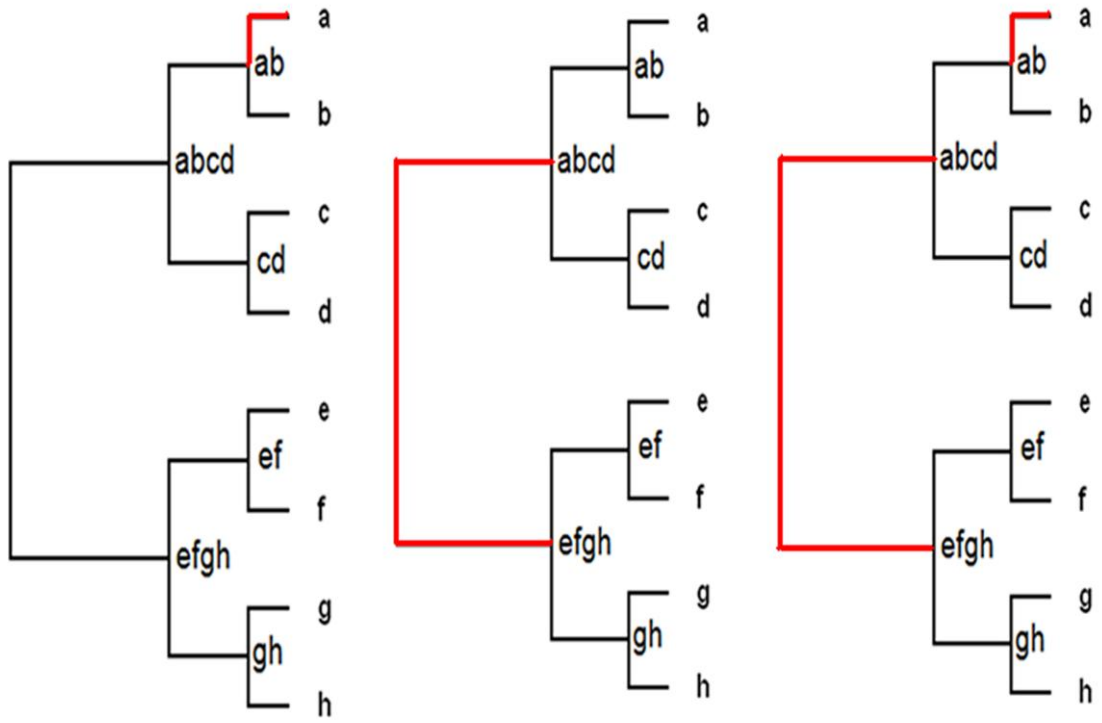
	tree	ω, L_C	Accuracy	Power	FPR	FNR
1	1	2.5, 300	0.00	0.00	1.00	1.00
2	1	2.5, 600	0.00	0.00	1.00	1.00
3	1	5.0, 300	0.18	0.06	0.82	0.94
4	1	5.0 600	0.16	0.06	0.84	0.94
5	2	2.5, 300	0.00	0.00	1.00	1.00
6	2	2.5, 600	0.00	0.00	1.00	1.00
7	2	5.0, 300	0.00	0.00	1.00	1.00
8	2	5.0, 600	0.12	0.13	0.88	0.87
9	3	2.5, 300	0.00	0.00	1.00	1.00
10	3	2.5, 600	0.17	0.13	0.83	0.87
11	3	5.0, 300	0.32	0.35	0.68	0.65
12	3	5.0, 600	0.26	0.19	0.74	0.81
13	4	2.5, 600	0.29	0.08	0.71	0.92
14	5	2.5, 600	0.46	0.08	0.54	0.92

* FPR=False positive rate; ** FNR=False negative rate

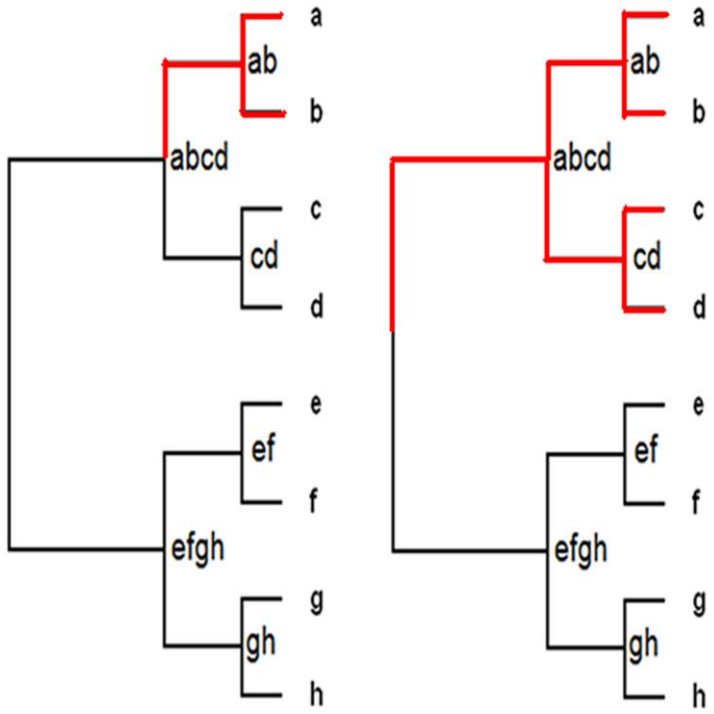
*$L_{C:}$ - Sequence length as number of codons; ω= selection strength

Figures

T1 **T2** **T3**



T4 T5

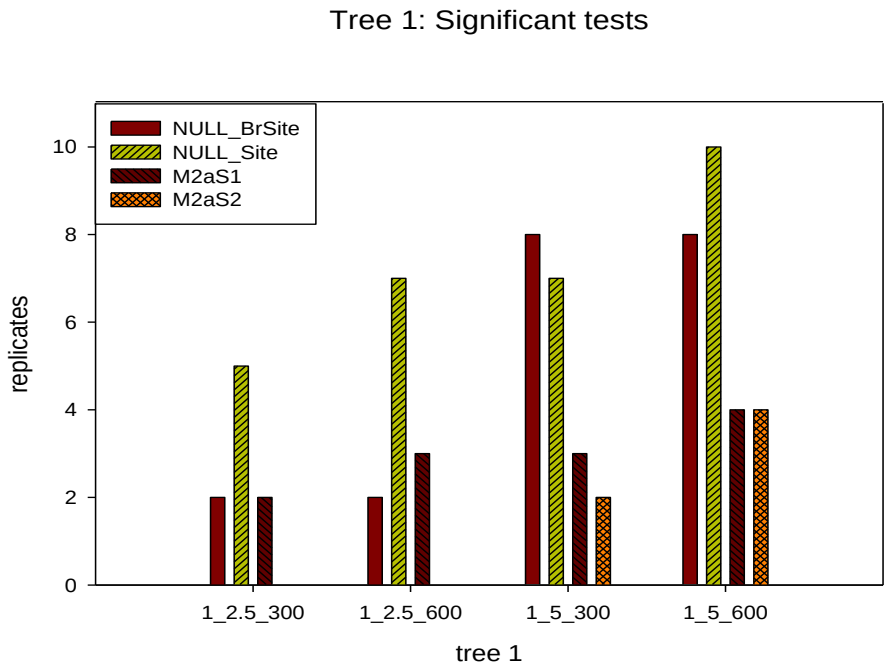


* Branches colored in red are simulated to have 10% of the sites evolving under positive selection

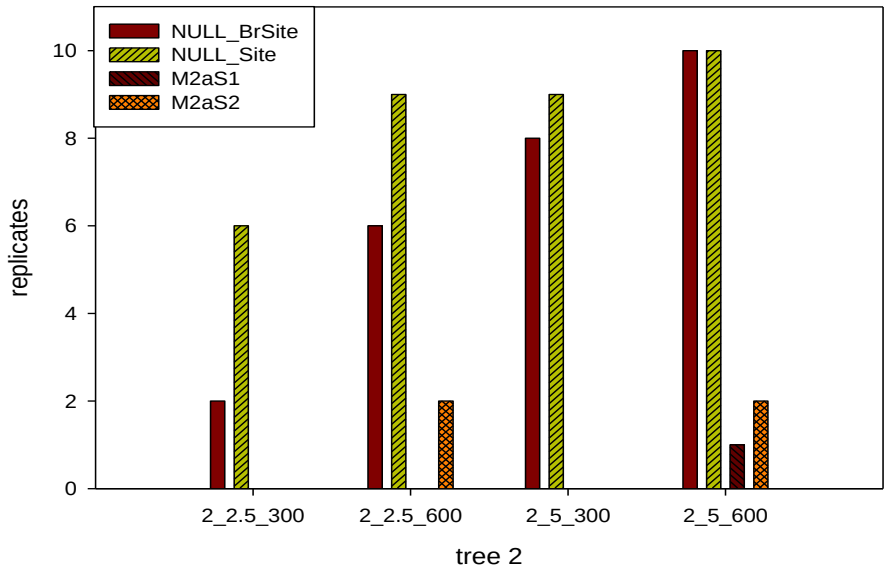
Figure 5.1: Eight taxa tree topologies used in the simulations

Five artificial trees each with eight taxa and equal branch length were simulated. Few lineages (colored in red) per tree (Tree T1: branch "a", tree T2: branch "abcdefgh" (tree root); tree 3 branch "a" and "abcdefgh" (tip and root); tree 4 clade (a, b) along with its stem lineage and tree 5 clade (a, b) and (c, d) and their stem lineages) with 10% of sites evolving under positive selection were simulated.

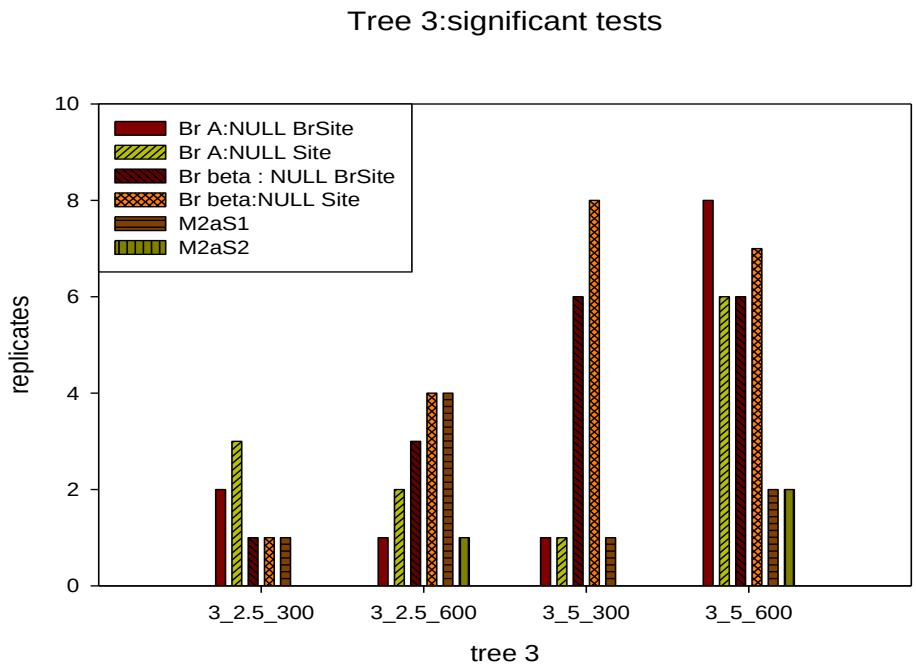
a)


Tree 1: Significant tests
NULL_BrSite
NULL_Site
M2aS1
M2aS2
replicates
0
2
4
6
8
10
1_2.5_300
1_2.5_600
1_5_300
1_5_600
tree 1

b)


Tree 2:Significant tests
NULL_BrSite
NULL_Site
M2aS1
M2aS2
replicates
0
2
4
6
8
10
2_2.5_300
2_2.5_600
2_5_300
2_5_600
tree 2

c)



d)

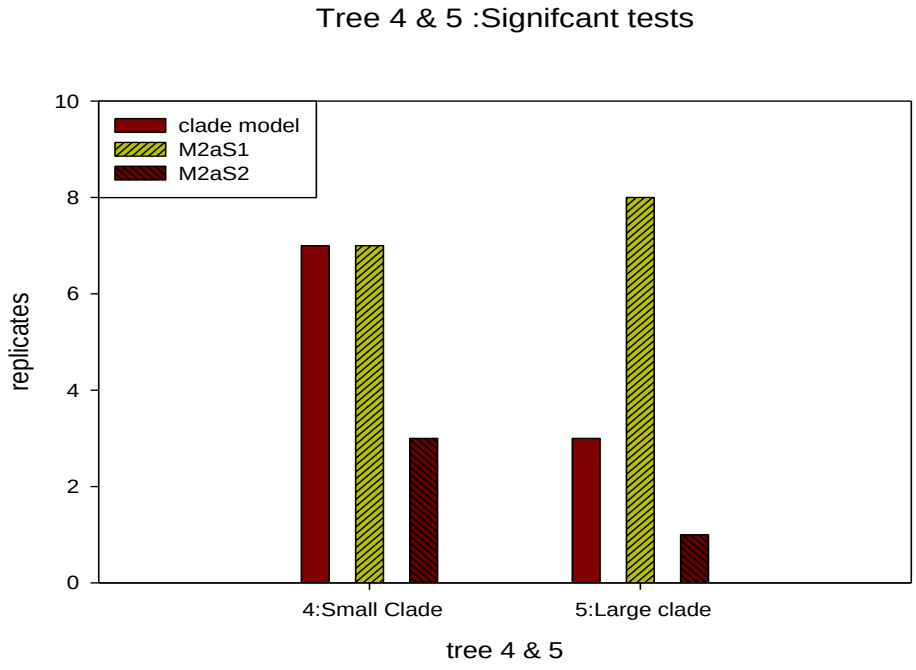
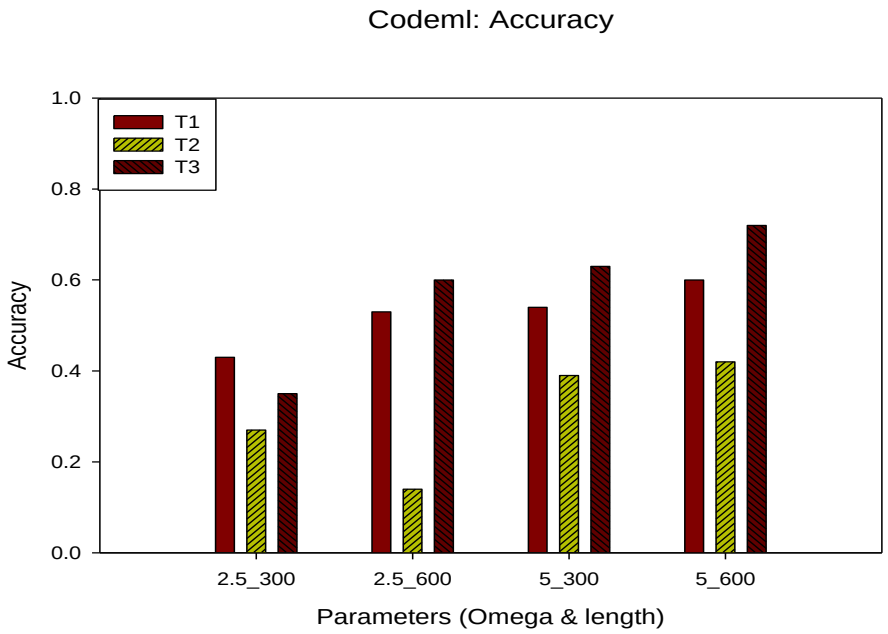


Figure 5.2: LRT: Number of significant tests in Codeml and Fitmodel

Figure 5.2: LRT: Number of significant tests in Codeml and Fitmodel

Panel a) shows number of significant replicates at 5% level when data simulated with sites on branch "a" evolving under positive selection. **Panel b)** shows number of significant replicates at 5% level when data simulated with sites on branch "abcdefgh" evolving under positive selection. **Panel c)** shows number of significant replicates at 5% level when data simulated with sites on branch "a" and "abcdefgh" (tip and root) evolving under positive selection. **Panel d)** shows number of significant replicates at 5% level when data simulated with sites on a small clade (a, b) along with its stem lineage as well as a large clade clade (a, b) and (c, d) along with its stem lineage evolving under positive selection.

a)



b)

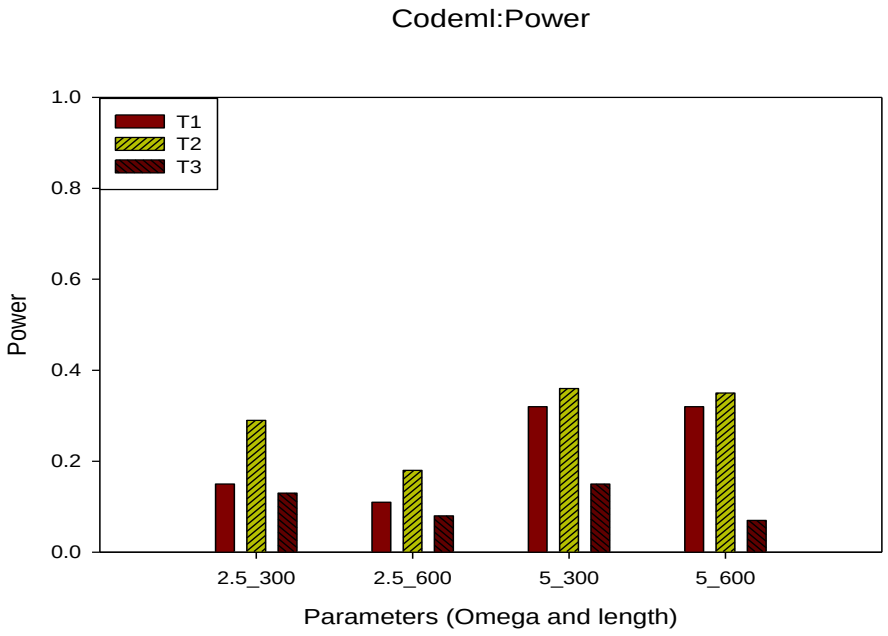
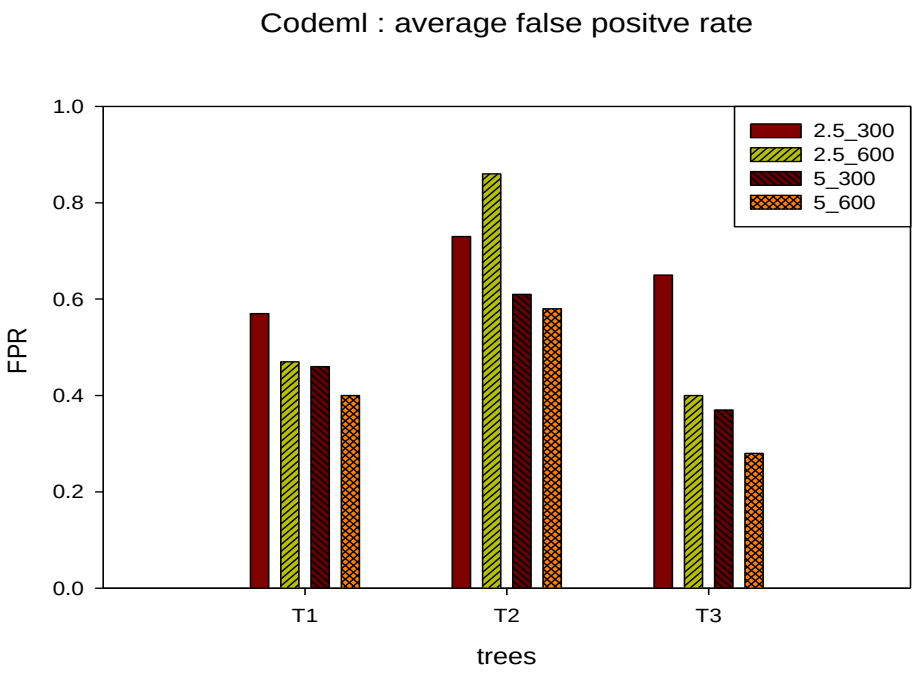


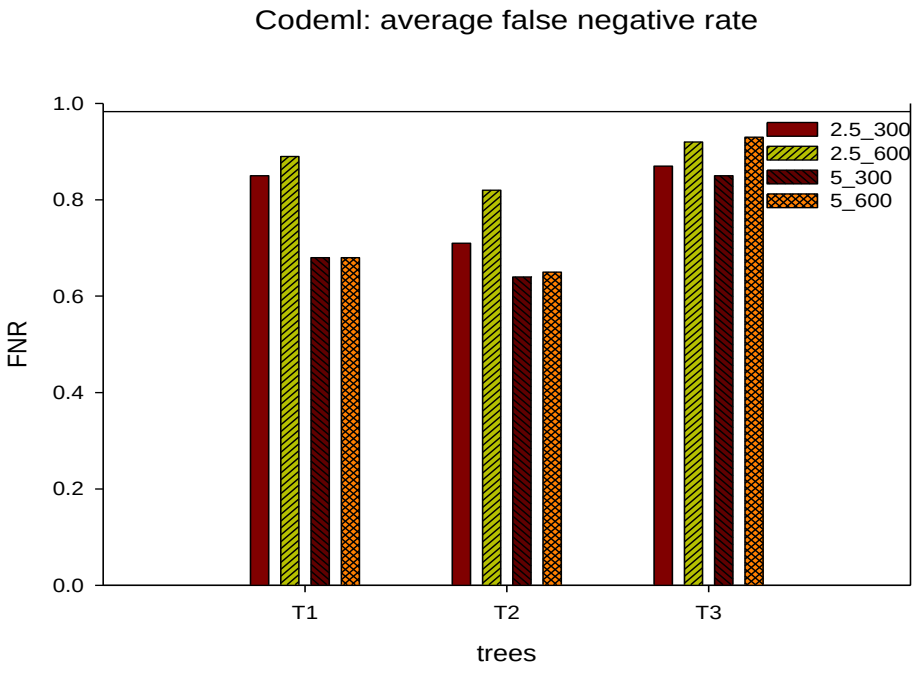
Figure 5.3: Codeml: Accuracy and Power of BEB prediction method

Panel a) shows accuracy of predicting sites under positive selection for trees T1, T2 and T3 under varying rate ratio (2.5/5) and sequence length (300/600 codons). **Panel b)** shows power of the BEB method for trees T1, T2 and T3 under varying rate ratio (2.5/5) and sequence length (300/600 codons)

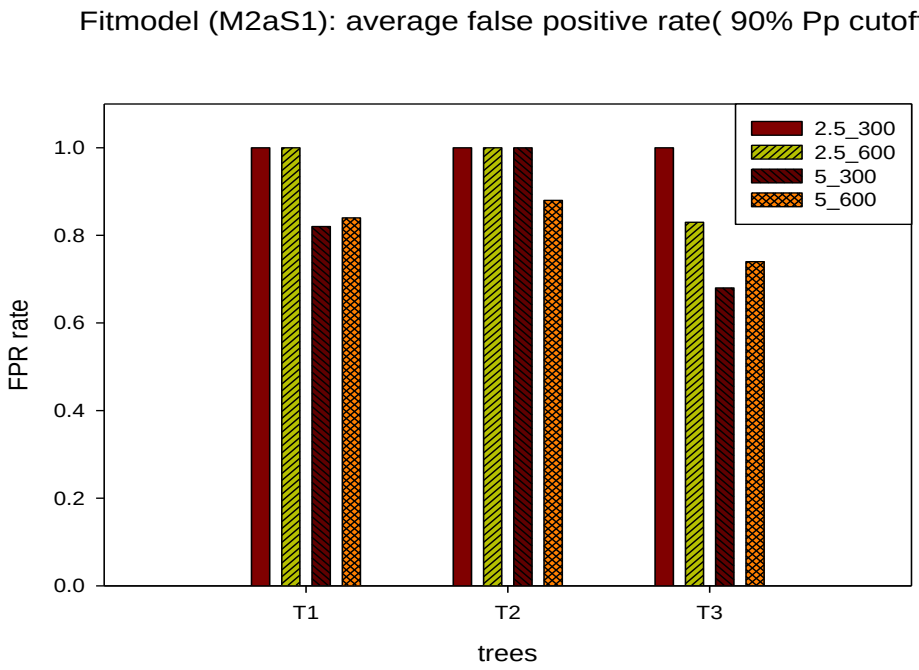
a)


Codeml : average false positve rate
1.0
0.8
0.6
0.4
0.2
0.0
FPR
2.5_300
2.5_600
5_300
5_600
T1
T2
T3
trees

b)


Codeml: average false negative rate
1.0
0.8
0.6
0.4
0.2
0.0
FNR
2.5_300
2.5_600
5_300
5_600
T1
T2
T3
trees

c)



d)

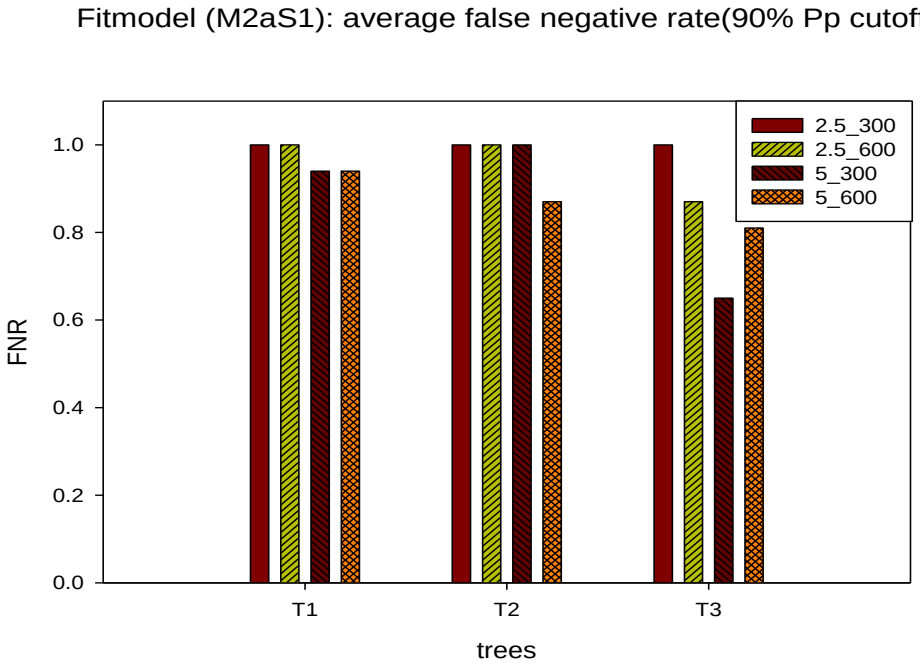


Figure 5.4: Average false positive rate and false negative rate for T1, T2 and T3 in Codeml and Fitmodel

Figure 5.4: Average false positive rate and false negative rate for T1, T2 and T3 in Codeml and Fitmodel

Panel a) shows the average false positive rate of predicting sites correctly by the BEB method in Codeml for schemes under trees T1, T2 and T3 respectively. **Panel b)** shows the average false negative rate of predicting sites correctly by the BEB method in Codeml for schemes under trees T1, T2 and T3 respectively. **Panel c)** shows the average false positive rate of predicting sites correctly by the posterior probability method in Fitmodel for schemes under trees T1, T2 and T3 respectively. **Panel d)** shows the average false negative rate of predicting sites correctly by the posterior probability method in Fitmodel for schemes under trees T1, T2 and T3 respectively. Four schemes per tree with combinations of rate ratio and sequence length were simulated.

CHAPTER 6

CONCLUSIONS

6.1 Phenylalanine ammonia lyase gene family

For this dissertation project, the phenylalanine ammonia lyase (PAL) gene family was selected for study. There were two principal reasons for selecting the PAL gene family. First, PAL plays a key role in regulating production of numerous chemically diverse phenolic compounds in plants, including lignin. These compounds contribute to a multilayered chemical defense system that plants use to adapt to changing ecological conditions (Bonello, et al., 2006). Secondarily, previous reports claimed that the PAL in loblolly pine existed as a single gene family (Whetten and Sederoff, 1992), which flew in the face of observations of PAL gene families in virtually all of the plants. As we had access to the largest transcriptome database for loblolly pine and several other conifer species, we were in an excellent position to confirm or overturn that previous report. Overall, the current work provides a more comprehensive understanding of the structural and functional aspects of the PAL gene family within gymnosperms, especially the conifers.

Deep transcriptome data from *Pinus taeda* and other conifers revealed that, as has been seen in angiosperms, PAL gene families are typically multigenic in the conifers. The five PAL members identified in the loblolly pine genome by computational analyses were subjected to experimental testing to verify that they were distinct expressed genes. For initial efforts to gain insight into functional roles of individual family members, a phylogenomic approach was utilized. It revealed diverse evolutionary trajectories of PAL genes in gymnosperms compared to angiosperms. This diverse set of PAL genes in loblolly pines was shown to be a product of

ancient duplication events, both predating and postdating divergence of angiosperms and gymnosperms.

The observation that ancient duplications played an important role in shaping evolution of PAL gene family in gymnosperms, strongly suggested the potential for functional variability between family members. With this in mind, experiments to detect and quantify differential expression of PAL gene family members were undertaken as a means to possibly associate specific physiological functions with individual gene family members. Consistent with the diverse biosynthetic pathways fed by with products from the PAL reaction, the gene family members displayed patterns of tissue-specific expression at different developmental stages and were differentially induced by various biotic and abiotic stresses. However, the analyses failed to provide conclusive evidence of specific functionality associated with one specific gene, so a level of functional redundancy cannot be ruled out.

Phylogenomic analyses coupled with reconciliation methods were used in an effort to detect potential sites under adaptive evolution. A software tool that evaluates different models of sequence evolution within a phylogenetic framework, Fitmodel (Guindon, 2004), was used to infer changes in mode of selection at the level of individual codon sites. Output from these analyses revealed a handful of sites evolving under shifting selection constraints. The output from Fitmodel, visualized using BASS (Huelsenbeck and Dyer, 2004), revealed these sites to be associated with duplication events.

For better understanding of the functional importance of site conservation and variation patterns observed at the amino acid level, a structure-function analysis was performed. The sites under adaptive selection or relaxed constraint were mapped onto the crystallographic structure of a PAL gene from parsley (Lois , et al., 1989). All of the sites under shifting selection constraints

mapped to the outer surface of the protein structure. This observation did not signify any special functional importance at this stage. However, since evolution typically occurs at amino acid residues that interact with external bio-molecules, these results provide a starting point for future studies of the relationships between specific candidate substitutions and functional variability of the PAL gene products. For example, application of *in silico* functional profiling procedures can be applied to determine the molecular effects of specific amino acid substitutions and quantify the genetic variation and functional consequences of these variations on the expression pattern of a gene (Horger, et al., 2012). And since this study was based on a limited set of cDNA sequences, expectations are that as additional whole genome sequences become available for gymnosperms and lower plants these approaches will increasingly highlight specific amino acid residue changes that are linked with different physiological functions for PAL gene family members. Thus, this work can be readily extended to perform more detailed analyses by extending the gene family studies to use genomic data (several projects to complete reference genome sequences for conifer species are nearing completion) (Mackay, et al., 2012; Neale, et al., 2014), or by utilizing additional high-throughput transcriptome sequencing in species selected for specific information content due to their phylogenetic position (Trapnell, et al., 2013).

As previously noted, phenylpropanoid derivatives are an extremely diverse class of metabolites in plants, but not all classes of phenylpropanoid compounds are present in all plant species. The elaboration of biosynthetic pathways to produce specific phenylpropanoids has been an important driving force behind the expansion of gene families associated with secondary metabolism. Comparative phylogenomic analyses of gene family evolution across species of interest is an increasingly important approach for filling gaps in our understanding of the roles

and functions of gene family paralogs and orthologs, particularly when we are faced with such metabolic diversity. Disparities between observed differences in gene family size can reflect differences in the complexity or regulation of divergent pathways in different species (Winkel-Shirley, 2001). Comprehensive lists of gene family members across different species can be used as the basis for detailed phylogenetic analyses highlighting the relatedness of the genes across species (Xu, et al., 2009). By extending such analyses to a variety of the gene families associated with phenylpropanoid metabolism (example Cinnamate-4-hydroxylase (C4H), 4-coumarate:CoA ligase (4CL), Ferulate 5-hydroxylase (F5H),Cinnamyl-alcohol dehydrogenase (CAD), Peroxidases and laccase), we can gain a better appreciation for the evolution of various conifer species while at the same time gaining further insight into the ways these pathways are regulated and diversified (Tsai, et al., 2006). Similar types of analyses in angiosperms have been useful for understanding the relationships between individual gene products and responses to specific stimuli (Dixon, et al., 2002).

Finally, work undertaken in this project to analyze the loblolly pine gene promoter sequences for *cis*-acting regulatory sequences that might act as transcription factors binding sites was an additional approach in the effort to link individual PAL gene family members to specific developmental processes or defense responses (Chiang, et al., 2012). The study identified a number of conserved transcription factor binding sites among co-expressed gene family members, as well as some sites unique to particular gene family members. Conserved sites may indicate overlapping functions and physiological redundancy, which may be reinforced under selected conditions (Wagner, 2005). These observations provide a foundation for designing specific experimental tests to determine whether, indeed, specific PAL family members respond to stimuli associated with the individual transcription factors from studies in other systems.

6.2 Efficiency of evolutionary models

To detect adaptive evolution, codon-based models under the maximum likelihood framework are frequently used (Guindon, 2004; Pond, et al., 2005; Yang, 2007). Fitmodel is one such tool that implements statistical models, which are considered to be more realistic than the widely used branch-X-site model implemented in Codeml (Zhang, et al., 2005). The application of switching models implemented under Fitmodel to investigate the substitution processes within the gymnosperm branch of the PAL gene family revealed problems with the statistical power. To develop a better understanding of the limitations this imposes for use of Fitmodel, I undertook a simulation-based study to compare and evaluate the branch-x-site model from Codeml against the switching models from Fitmodel. This served two purposes: 1) to understand the efficiency of the models to correctly detect selection pressure acting on individual codon sites; and 2) the ability of the Bayesian method to correctly identify sites evolving under positive selection. The simulated sequences were obtained from an artificial phylogenetic tree in which shifts in selection pressure were implemented in specific lineages at different evolutionary times (Yang, 2007). Other conditions tested included varying codon alignment length as well as strength of positive selection acting on specific sites.

The power of the models was tested by determining the performance of the likelihood ratio test under the various model conditions. Under the conditions tested, the branch-X-site model seemed to correctly reject the Null hypothesis as the alternative hypothesis was correct. However, the Bayesian approach, which identifies sites under adaptive evolution, performed poorly in both Codeml as well as Fitmodel. Overall, the effect of sequence alignment length as well as strength and timing of selection pressures, negatively impacted results from the Bayesian evaluations. Future extensions of this work should test whether other parameters, such as tree

topology, branch length and number of taxa (sequences) in an alignment, have similar ramifications for the current Bayesian implementations. As an example of why such analyses might be important, in one site-based analysis of population data, more than 1000 simulated sequences were necessary to generate results with sufficient statistical power to detect sites under selection constraint (Anisimova, 2003). Similarly, high sequence divergence has been an issue in previous studies using the branch-X-site model, leading to large number of false positives, frequently due to alignment difficulties and differences in codon usage patterns (Mallick, et al., 2009; Fletcher and Yang, 2010).

Besides Codeml and Fitmodel, new tools implementing more parameter-rich substitution models, such as the random effect Branch-Site Model distributed as a part of HYPHY package (Pond, et al., 2011) or the mixed effect model of evolution (Murrell, et al., 2012) are released and will need to be benchmarked for performance. These new models generally allow substitution rates to vary from branch to branch and site to site, similar to Fitmodel. Testing of both simulated and real datasets, particularly where convincing experimental evidences is available to help related to gene function to specific codon changes, will improve our understanding of these tools and better demonstrate their efficiencies and shortcomings.

6.3 References

Anisimova, M. (2003) Detecting positive selection in protein coding genes. *Department of Biology*. University College, London.

Bonello, P., Gordon, T.R., Herms, D.A., Wood, D.L. and Erbilgin, N. (2006) Nature and ecological implications of pathogen-induced systemic resistance in conifers: a novel hypothesis, *Physiological and Molecular Plant Pathology*, **68**, 95-104.

Chiang, Y.-H., Zubo, Y.O., Tapken, W., Kim, H.J., Lavanway, A.M., Howard, L., Pilon, M., Kieber, J.J. and Schaller, G.E. (2012) Functional Characterization of the GATA Transcription

Factors GNC and CGA1 Reveals Their Key Role in Chloroplast Development, Growth, and Division in Arabidopsis, *Plant Physiology*, **160**, 332-348.

Dixon, R.A., Achnine, L., Kota, P., Liu, C.J., Reddy, M.S.S. and Wang, L.J. (2002) The phenylpropanoid pathway and plant defence - a genomics perspective, *Molecular Plant Pathology*, **3**, 371-390.

Fletcher, W. and Yang, Z. (2010) The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection, *Molecular Biology and Evolution*, **27**, 2257-2267.

Guindon, S., Rodrigo, A, Dyer, Kelly A., Huelsenbeck, John P. (2004) Modeling the site-specific variation of selection patterns along lineages, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12957-12962.

Horger, A.C., Ilyas, M., Stephan, W., Tellier, A., van der Hoorn, R.A.L. and Rose, L.E. (2012) Balancing Selection at the Tomato RCR3 Guardee Gene Family Maintains Variation in Strength of Pathogen Defense, *PLoS Genetics*, **8**, e1002813.

Huelsenbeck, J.P. and Dyer, K.A. (2004) Bayesian estimation of positively selected sites, *Journal of Molecular Evolution*, **58**, 661-672.

Lois, R., Dietrich, A., Hahlbrock, K. and Schulz, W. (1989) A phenylalanine ammonia-lyase gene from parsley: structure, regulation and identification of elicitor and light responsive cis-acting elements., *The EMBO Journal*, **8**, 1641-1648.

Mackay, J., Dean, J.F.D., Plomion, C., Peterson, D.G., Cànovas, F., Pavy, N., Ingvarsson, P., Savolainen, O., Guevara, M.Ã., Fluch, S., Vinceti, B., Abarca, D., DÃaz-Sala, C. and Cervera, M.-T. (2012) Towards decoding the conifer giga-genome, *Plant Molecular Biology*, **80**.

Mallick, S., Gnerre, S., Muller, P. and Reich, D. (2009) The difficulty of avoiding false positives in genome scans for natural selection, *Genome Research*, **19**, 922-933.

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K. and Kosakovsky Pond, S.L. (2012) Detecting individual sites subject to episodic diversifying selection, *PLoS Genetics*, **8**, e1002764.

Neale, D., Wegrzyn, J., Stevens, K., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Koriabine, M., Holtz-Morris, A., Liechty, J., Martinez-Garcia, P., Vasquez-Gross, H., Lin, B., Zieve, J., Dougherty, W., Fuentes-Soriano, S., Wu, L.-S., Gilbert, D., Marcais, G., Roberts, M., Holt, C., Yandell, M., Davis, J., Smith, K., Dean, J., Lorenz, W., Whetten, R., Sederoff, R., Wheeler, N., McGuire, P., Main, D., Loopstra, C., Mockaitis, K., deJong, P., Yorke, J., Salzberg, S. and Langley, C. (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies, *Genome Biology*, **15**, R59.

Pond, K.S.L., Murrell, B., Fourment, M., Frost, S.D.W., Delport, W. and Scheffler, K. (2011) A random effects branch-site model for detecting episodic diversifying selection, *Molecular Biology and Evolution*, 3033-3043.

Pond, S.L.K., Frost, S.D.W. and Muse, S.V. (2005) HyPhy: hypothesis testing using phylogenies, *Bioinformatics*, **21**, 676-679.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L. and Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq, *Nature Biotechnology*, **31**, 46-53.

Tsai, C.-J., Harding, S.A., Tschaplinski, T.J., Lindroth, R.L. and Yuan, Y. (2006) Genome-wide analysis of the structural genes regulating defense phenylpropanoid metabolism in Populus, *New Phytologist*, **172**, 47-62.

Wagner, A. (2005) Distributed robustness versus redundancy as causes of mutational robustness, *BioEssays*, **27**, 176-188.

Whetten, R.W. and Sederoff, R.R. (1992) Phenylalanine ammonia lyase from loblolly pine: purification of the enzyme and isolation of complementary DNA clones, *Plant Physiology*, **98**, 380-386.

Winkel-Shirley, B. (2001) Flavonoid biosynthesis. A colorful model for Genetics, Biochemistry, Cell Biology, and Biotechnology, *Plant Physiology*, **126**, 485-493.

Xu, Z., Zhang, D., Hu, J., Zhou, X., Ye, X., Reichel, K., Stewart, N., Syrenne, R., Yang, X., Gao, P., Shi, W., Doeppke, C., Sykes, R., Burris, J., Bozell, J., Cheng, Z.-M., Hayes, D., Labbe, N., Davis, M., Stewart, C.N. and Yuan, J. (2009) Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom, *BMC Bioinformatics*, **10**, S3.

Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood, *Molecular Biology and Evolution*, **24**, 1586-1591.

Zhang, J., Nielsen, R. and Yang, Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level, *Molecular Biology and Evolution*, **22**, 2472-2479.