Phylogenomic Analysis of the Phenylalanine Ammonia Lyase Gene Family in Loblolly Pine (Pinus taeda L.)

UJWAL R. BAGAL, JAMES H. LEEBENS-MACK, W. WALTER LORENZ AND JEFFREY F.D. DEAN
Institute of Bioinformatics, University of Georgia
Athens, GA 30602-7229, USA
e-mail: bujwal@uga.edu, jleebensmack@plantbio.uga.edu, wlorenz@uga.edu, jeffdean@uga.edu

Abstract - Phenylalanine ammonia lyase (PAL) is a key enzyme of the phenylpropanoid pathway. PAL genes are less extensively studied in gymnosperms than angiosperms. Our interest in the PAL genes in Pinus taeda stems from their potential role in pine defense mechanisms. Despite the number of characterized gymnosperm PAL genes, the functional diversity of these and the extensive transcriptome data available for Pinus taeda, only one PAL gene had been characterized in P. taeda. Here, we present an extensive phylogenomics analysis of the PAL gene family in P. taeda including identification of five distinct paralogs. Assembly coverage and expression profiles are variable among the five PAL genes in P. taeda. A phylogenetic analysis revealed a unique evolutionary history for the PAL genes in gymnosperms including origin of a distinct “gymnosperm” PAL gene clade. A reconciliation method was used to map ancestral duplication events and lineage-specific duplication. Molecular evolutionary analysis of the gymnosperm genes showed shifts in selective constraints following gene duplications.

I. INTRODUCTION

Conifers have experienced large environmental and distributional changes during their evolution [9]. To adapt to their diverse ecological habitats as well as the biotic and abiotic stresses associated with them, conifers have developed multifaceted chemical response systems as part of their biotic resistance strategy [3]. Many of the secondary products of these response systems are synthesized via the phenylpropanoid pathway [1]. 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 [16]. PAL plays a key regulatory role in controlling biosynthesis of all phenylpropanoid products.

In contrast to angiosperms and a few other gymnosperms for which PAL gene families have been characterized, before this study only a single gene copy had been reported for P. taeda [30]. With new and much larger EST datasets having been recently made available for P. taeda and other conifers, we have applied phylogenetic approaches to investigate the evolutionary history of the PAL gene family with emphasis on the diversification of conifer PAL genes, especially newly characterized genes identified from P. taeda transcriptome data. Specifically, a PAL gene phylogeny was used to test whether diversification of gymnosperm PAL genes occurred independently of diversification in the angiosperms.

Our preliminary results indicate that P. taeda possesses at least five (5) distinct PAL genes. Further, the gene tree elucidated a diverse set of gymnosperm specific PAL genes. At least three conifer lineage-specific duplication events amongst all the ancient duplication events in the PAL gene tree suggest an evolutionary history different from that in angiosperms. Lastly, gene-specific primers for four of the five PAL genes in P. taeda were used to amplify unique sequences from P. taeda cDNA using PCR.

II. MATERIALS AND METHODS

A. 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 Wolllemia 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.

B. Taxon Sampling and Phylogenetic analysis

The nucleotide sequences and corresponding amino acid sequences for the representative taxa were collected from various public databases [2, 8, 29]. The inferred transcript sequences for C. atlantica, C. harringtonia, P. abies, P. lambertiana, P. macrophyllus, P. palustris, P. sylvestris, S.
**A. Primer Design:**

The assembled contigs used for the phylogenetic analysis were also used as the basis for designing gene-specific, degenerate oligo-nucleotide primers for PCR studies. A pair of PCR primers, Fwd ["AAGAACAAGCAAGGTCAGAAG"] and Rev ["AGCATTTAAGAGAGGAGACTATGAC"], were designed to amplify 307 bp from PtPAL1 (Pteda143311). In a similar fashion, Fwd ["CTGACTGAGACTGCGCAAATTC"] and Rev ["ATCCTCCTGCCGTTTCAATG"] primers amplified a 444 bp sequence from PtPAL2 (Pteda17307), Fwd ["TCAGATTGGGGACGATTTG"] and Rev ["TTATTGATTCATTGTTGAAAC"] primers amplified a 388 bp sequence from PtPAL3 (Pteda9006), and Fwd ["CGAATACGAGGTCTTCATCCITTAC"] and Rev ["CGGCCCTCGATCGCCTAAAG"] 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).

**2. PCR amplification and sequencing of the PCR product:**

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. A reaction mixture containing 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 a Applied Biosystems 3730XL sequencer at the Georgia Genomics Facility (http://dna.uga.edu/).

**III. RESULT**

**A. PAL in Pinus taeda**

**TABLE 1:** *P. taeda* PAL inferred amino acid sequence percent identity/similarity

<table>
<thead>
<tr>
<th>Gene id (aa length)</th>
<th>PtPAL1</th>
<th>PtPAL2</th>
<th>PtPAL3</th>
<th>PtPAL4</th>
<th>PtPAL5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PtPAL1(754)</td>
<td>#</td>
<td>76/87</td>
<td>64/79</td>
<td>68/81</td>
<td>64/79</td>
</tr>
<tr>
<td>PtPAL2(727)</td>
<td>#</td>
<td>65/80</td>
<td>67/79</td>
<td>63/78</td>
<td></td>
</tr>
<tr>
<td>PtPAL3(808)</td>
<td></td>
<td>64/77</td>
<td>60/75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PtPAL4(711)</td>
<td>#</td>
<td>88/93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PtPAL5(687)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>#</td>
</tr>
</tbody>
</table>

DNA sequence data from 71 PAL genes representing 25 taxa representing Monocots, Dicots, Gymnosperms and basal taxa were selected after performing a preliminary phylogenetic analysis. For *P. taeda*, five distinct phenylalanine ammonia lyase consensus sequences were identified in our EST assemblies. Complete coding sequences were inferred for all five PtPAL genes. 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 angiosperms PAL genes which include an intron, PtPAL1 and the PAL genes previously characterized in *P. banksiana* [7] lack introns.

Counts of ESTs from various *P. taeda* cDNA libraries were used to assess differential expression among tissues and developmental stages. There was variation in read numbers across libraries. PtPAL1 sequences were found most frequently, and were detected in nearly all libraries. In general, PtPAL2 was found to be the second most highly expressed gene, followed by PtPAL3, PtPAL4 and PtPAL5. Unlike PtPAL1 and PtPAL2 the remaining three PtPALS were restricted to aerial tissues only.

**B. Active Sites**

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, as were additional residues, previously noted as conserved (Leu134, Hist137, Leu215, Leu266, Val269, Asn270, Tyr363, Arg366, Phe413, Lys468, Glu496, Gln500, Ile472) [5,10]. These observations strongly support the contention that all the products of the
genes included in these analyses bind the same substrate, phenylalanine.

C. PAL gene characterization

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 the presence of distinct PAL genes in *P. taeda*.

D. Phylogenetic analysis

Phylogenetic analysis was performed to evaluate the evolutionary relationships among the 71 PAL sequences from 25 taxa selected for this analysis. Both the maximum likelihood and Bayesian methods were used to ensure that the results were in congruence. In both analyses *Physcomitrella patens* was used as an out-group (Figure 1). The consensus trees obtained from both methods showed similar organization, with gymnosperm genes distributed among three distinct clades. One gymnosperm-specific clade is placed just above the out-group branches in the PAL gene tree. A clade with the remaining genes is split into a gymnosperm-specific clade and a second clade containing both angiosperm and gymnosperm PAL genes. The high bootstrap value and posterior probability evidence provided strong support for the organisation of the gymnosperm genes into these three distinct clusters. Within the angiosperm PAL genes, monocot and eudicot clades were recovered as described in a previous report [12]. The successive origins of three distinct gymnosperm PAL gene clades before the origin of the angiosperm clade suggests that 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 1).

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 that additional *Picea* cDNA sequences may yield additional PAL genes since several incomplete *Picea* PAL genes were removed 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 clades.

A species tree based on taxonomic information from NCBI (National Center for Biotechnology Information) was used to reconcile the gymnosperm section of the gene tree keeping *P. patens* as the out-group (Figure 2). Notung version 2.6 [6] 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. The oldest duplication event, 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.

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*.

Figure 2: 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.

IV. DISCUSSION

Phenylalanine ammonia lyase, which belongs to the lyase class I super-family of enzymes [27], 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 [21, 26, 17, 19]. This study is the most extensive phylogenomic study so far for the PAL gene family, particularly with respect to conifers. We were able to identify five distinct PAL genes in *P. taeda* cDNA sequence assemblies. 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*) and the other two clades originated by gene duplication within a seed-plant ancestor before the divergence of angiosperms and gymnosperms. This result suggests PAL genes were lost on the branch leading to angiosperms.

Analysis of *P. taeda* PAL gene sequences from libraries representing different tissues and different stress conditions uncovered differential expression patterns for the five PAL genes. The level of expression was highest for PtPAL1, which showed uniformly high expression in almost all conditions.
tissues and treatments. PtPAL4 and PtPAL5 showed the lowest levels of expression. There was some evidence for increased expression of PtPAL1 and PtPAL4 in xylem, which might suggest a role in lignin biosynthesis. Expression of PtPAL3, PtPAL4 and PtPAL5 appeared to be restricted primarily to aerial tissues. Other than PtPAL1, only PtPAL2 showed expression in roots under varying conditions. These 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 and in response to various external stimuli.

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 [12]. 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 regulated by similar developmental control mechanisms [24].

The existence of two additional gymnosperm PAL gene clades indicates maintenance of PAL genes in gymnosperms and loss of diversity in angiosperms [7, 22]. The branching patterns within the conifer genes within these clades are in accordance with patterns reported previously for these species [9].

Duplication events have been an important theme in the overall evolution of the PAL gene family [17]. At least five distinct duplication events can be identified in the PAL gene tree, with the oldest event following divergence of Physcomitrella. Duplication events in the ancestral lineage, as well as at the tip of the gymnosperm branch, suggest potential sources of functional variability [24]. 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 [22]. 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 development or resistance to biotic and abiotic stresses [20]. Thus, in artichoke, three different PAL genes were suggested to play different roles in defense responses [23]. In Populus, one PAL gene product was associated with formation of condensed tannins while another was associated with lignin production. In tobacco, post-transcriptional regulation of one PAL gene in the family was reported, although the exact mechanism was not clear [25]. 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 [17]. Based on this model, PtPAL4 and PtPAL5 may have resulted from a recent duplication event (Figure 1). Likewise, as seen in other species, PtPAL genes that do not cluster together may encode PAL isozymes having unique functions, perhaps playing different metabolic role by producing different products under varying conditions.

Preliminary study of the evolutionary processes of PAL genes has detected number of non-synonymous sites. When mapped on a PAL crystal structure these non-synonymous substitution events were concentrated at sites on the outer surface of the protein structure, which may have functional importance and suggest interactions with other cellular components.

**ACKNOWLEDGEMENTS**

This work was supported in part by a research assistancehip 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/].

**REFERENCES**


