BMC Genomics

PROCEEDINGS Open Access

The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage

Ujwal R Bagal[1], James H Leebens-Mack[1,2], W Walter Lorenz[3], Jeffrey FD Dean[1,3,4*]

From IEEE International Conference on Bioinformatics and Biomedicine 2011
Atlanta, GA, USA. 12-15 November 2011

Abstract

Background: 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*.

Methods: 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.

Results: 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.

Conclusions: 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.

Background

Conifers have experienced large environmental and distributional changes during their evolution, dating back to the Mesozoic era [1]. 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 [2]. 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 [3]. The precursors for many of these phenolic defense compounds are synthesized via the phenylpropanoid pathway [4].

* Correspondence: jeffdean@uga.edu
[1]Institute of Bioinformatics, The University of Georgia, Davison Life Sciences Bldg, Athens, GA 30602-7229, USA
Full list of author information is available at the end of the article

The phenylpropanoid pathway has been extensively studied with respect to production of natural products, such as flavonoids, isoflavonoids, hydroxycinnamic 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 signalling, and UV protection, as well as mechanical support and response to low levels of iron and phosphate [5].

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 [6]. 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 [7]. Increased activity of PAL has been correlated with increased production of phenylpropanoid products [8], and levels of PAL activity vary with developmental stage, cell and tissue differentiation, and exposure to different stress stimuli [9-11]. PAL has been reported to be stimulated by infection, mechanical wounding, UV irradiation, drought stress and drastic temperature changes [12-14].

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 [15]. 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 [16]. 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 [17], 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 [18].

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 [19]. 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.

Results

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 pseudotranscripts 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 misassembly, 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 pseudotranscripts 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,

Table 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.

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* [20] lack introns.

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.

Table 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.

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 [21,22]. These observations strongly support the contention that all enzymes encoded by the genes included in these analyses bind and utilize the same substrate, phenylalanine.

Phylogenetic analysis

Phylogenetic analysis was performed to evaluate the evolutionary relationships among the 71 PAL sequences

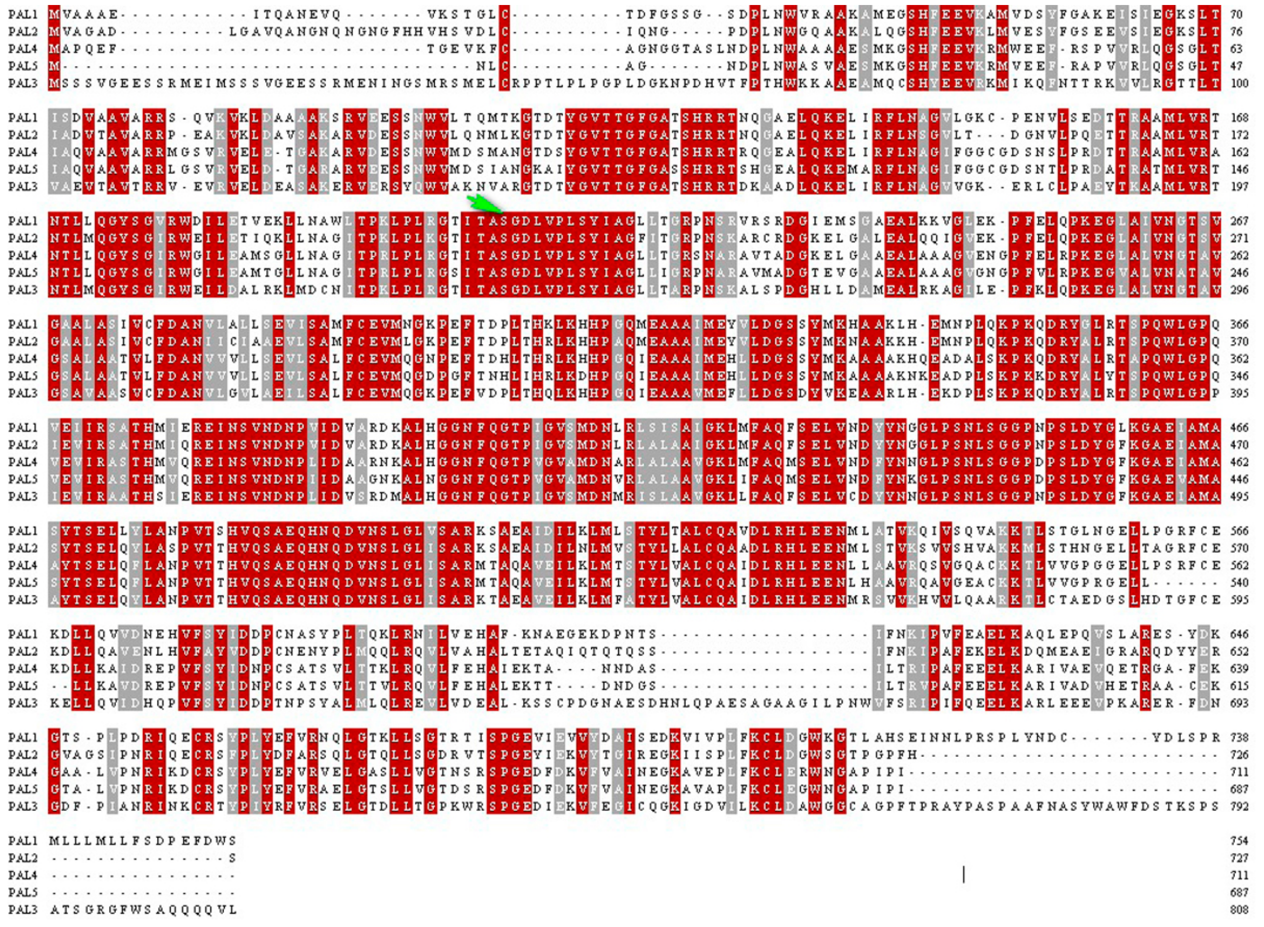

Figure 1 Alignment between the five PtPAL genes in P. taeda. Arrow indicates position of the conserved MIO region (Ala-Ser-Gly triad).

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 [23].

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 outgroup (Figure 3). Notung version 2.6 [24] 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

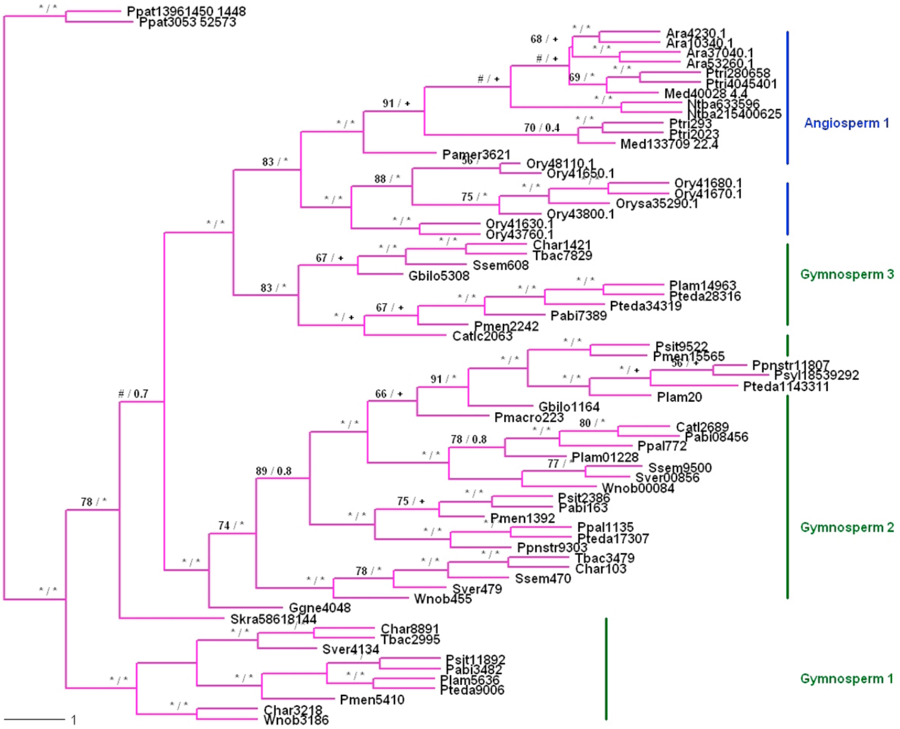

Figure 2 Consensus tree of the Phenylalanine ammonia lyase gene family. Numbers at nodes are nonparametric Bootstrap values (BS) from Maximum Likelihood (ML) and Posterior Probabilities (PP) from Bayesian Inference (BI), respectively, separated by a slash. Asterisks (*/*) symbol indicates [90-100]/[0.9 - 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.

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*.

Discussion

Phenylalanine ammonia lyase, which belongs to the lyase class I super-family of enzymes [7], is a primary control point for the phenylpropanoid pathway, which in part explains the multigene families seen for PAL in almost all plants studied to date [10,25-27]. 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 [28]. 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.

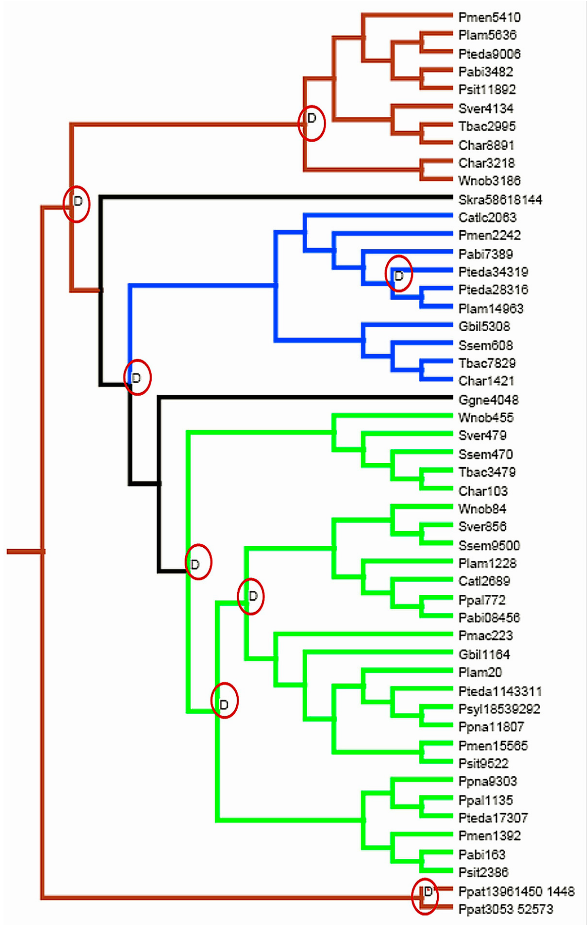

Figure 3 NOTUNG: reconciled gene tree. A reconciled gene tree with duplication events as obtained from Notung is depicted. Duplication nodes are marked with circles. The branch shading corresponds to the pattern of gymnosperms branching. The blue branch indicates gymnosperm sequences that clustered with angiosperm PAL genes. The green branch indicates a unique gymnosperm branch, while the brown branch indicates gymnosperm sequences clustering with sequences from basal taxa.

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 [23]. 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 [29].

The existence of two additional gymnosperm PAL gene clades indicates maintenance of PAL genes in gymnosperms and loss of diversity in angiosperms [30]. The branching patterns within the conifer genes within these clades are in accordance with patterns reported previously for these species [1].

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 [29]. 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 [30]. 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 [31]. Thus, in artichoke, three different PAL genes were suggested to play different roles in defense responses [32]. In Poplar, one PAL gene product was associated with formation of condensed tannins while another was associated with lignin production [23]. In tobacco, post-transcriptional regulation of one PAL gene in the family was reported, although the exact mechanism was not clear [33]. 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 [26]. 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.

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.

Methods

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 [34], 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 [35].

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.

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 oligonucleotide 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 ["TCCTCCTGCC GTTTCCAATG"] primers amplified a 444 bp sequence from PtPAL2 (Pteda17307), Fwd ["TCAGAGTTGGGAACCGATTTG"] and Rev ["CTATTGATTCATTGT TGTTGGAACC"] primers amplified a 388 bp sequence from PtPAL3 (Pteda9006), and Fwd ["CCAATAACGACGCTTCTATCCTTAC"] and Rev ["CGCCGTTCC ATCGCTCAAG"] 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 1 µl of Taq polymerase, 2 µl of 10 mM dNTP, 4 µl of Optiprime 10× buffer, 3 µl of 5 mM 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/).

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.

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 [36-38]. 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 [39]. Multiple codon alignment corresponding to protein sequences was performed using PAL2NAL [40]. Molecular phylogeny estimates were derived using RAxML [41] and MrBayes [42] on a 2430 character sequence alignment. For the RAxML estimation, a generalized time-reversible (GTR) substitution model [43] with across-site rate variation modelled as a gamma distribution [44] and invariant sites (GTR +GAMMA+I), was used for nucleotide alignments. For amino acid alignment, the JTT [Jones, Taylor and Thornton] substitution model [45] 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.

Additional material

Additional file 1: Nucleotide sequences used for analysis. The given file contains sequences downloaded from the public database for angiosperms and few of the gymnosperms. It also contains assembled consensus sequences for gymnosperms. These sequences were used for getting amino acid as well as codons for evolutionary analysis. Primers specific to PtPAL were designed using PtPAL1 to PtPAL4.

List of abbreviations used

PAL: Phenylalanine ammonia lyase; PtPAL: PAL from *Pinus taeda*.

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."
This article has been published as part of *BMC Genomics* Volume 13 Supplement 3, 2012: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2011: Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S3.

Author details

[1]Institute of Bioinformatics, The University of Georgia, Davison Life Sciences Bldg, Athens, GA 30602-7229, USA. [2]Department of Plant Biology, Miller Plant Sciences, The University of Georgia, Athens, GA 30602-7271, USA. [3]Warnell School of Forestry and Natural Resources, The University of Georgia, Athens, GA 30602-2152, USA. [4]Department of Biochemistry and Molecular Biology, The University of Georgia, Davison Life Sciences Bldg, Athens, GA 30602-7229, USA.

Authors' 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.

Competing interests

The authors declare that they have no competing interests.

Published: 11 June 2012

References

1. Eckert AJ, Hall BD: **Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): phylogenetic tests of fossil-based hypotheses.** *Mol Phylogenet Evol* 2006, **40**:166-182.
2. de Laubenfels DJ: **The status of "conifers" in vegetation classifications.** *Ann Assoc Am Geogr* 1957, **47**:145-149.
3. Bonello P, Gordon TR, Herms DA, Wood DL, Erbilgin N: **Nature and ecological implications of pathogen-induced systemic resistance in conifers: a novel hypothesis.** *Physiol Mol Plant Pathol* 2006, **68**:95-104.
4. Adomas A, Heller G, Li G, Olson A, Chu T, Osborne J, Craig D, Zyl LV, Wolfinger R, Sederoff R, Dean RA, Stenlid J, Finlay R, Asiegbu FO: **Transcript profiling of a conifer pathosystem: response of *Pinus sylvestris* root tissues to pathogen (*Heterobasidion annosum*) invasion.** *Tree Physiol* 2007, **27**:1441-1458.
5. Dixon RA, Paiva NL: **Stress-induced Phenylpropanoid metabolism.** *Plant Cell* 1995, **7**:1085-1097.
6. Koukal J, Conn EE: **The metabolism of aromatic compounds in higher plants. IV. Purification and properties of the phenylalanine deaminase of Hordeum vulgare.** *J Biol Chem* 1961, **236**:2692-2698.

7. Ritter H, Schulz GE: **Structural basis for the entrance into the phenylpropanoid nucleic metabolism catalyzed by phenylalanine ammonia-lyase.** *Plant Cell* 2004, **16**:3426-3436.
8. Ozeki Y, Komamine A: **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 Physiol* 1985, **26**:903-911.
9. Jones DH: **Phenylalanine ammonia-lyase: regulation of its induction, and its role in plant development.** *Phytochemistry* 1984, **23**:1349-1359.
10. Lois R, Dietrich A, Hahlbrock K, Schulz W: **A phenylalanine ammonia-lyase gene from parsley: structure, regulation and identification of elicitor and light responsive cis-acting elements.** *EMBO J* 1989, **8**:1641-1648.
11. Shufflebottom D, Edwards K, Schuch W, Bevan M: **Transcription of two members of a gene family encoding phenylalanine ammonia-lyase leads to remarkably different cell specificities and induction patterns.** *Plant J* 1993, **3**:835-845.
12. Edwards K, Cramer CL, Bolwell GP, Dixon RA, Schuch W, Lamb CJ: **Rapid transient induction of phenylalanine ammonia-lyase mRNA in elicitor-treated bean cells.** *Proc Natl Acad Sci USA* 1985, **82**:6731-6735.
13. Lange BM, Lapierre C, Sandermann H Jr: **Elicitor-induced spruce stress lignin (structural similarity to early developmental lignins).** *Plant Physiol* 1995, **108**:1277-1287.
14. Campbell MM, Ellis BE: **Fungal elicitor-mediated responses in pine cell cultures: III. Purification and characterization of phenylalanine ammonia-lyase.** *Plant Physiol* 1992, **98**:62-70.
15. Stefanoviac S, Jager M, Deutsch J, Broutin J, Masselot M: **Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences.** *Am J Bot* 1998, **85**:688-697.
16. Ritland K, Ralph S, Lippert D, Rungis D, Bohlmann J: **A new direction for conifer genomics.** *Landscapes, Genomics and Transgenic Conifers* Springer; 2006, 75-84.
17. Perry DJ, Furnier GR: **Pinus banksiana has at least seven expressed alcohol dehydrogenase genes in two linked groups.** *Proc Natl Acad Sci USA* 1996, **93**:13020-13023.
18. Ahuja MR, Neale DB: **Evolution of genome size in conifers.** *Silvae Genetica* 2005, **54**:126-137.
19. Whetten RW, Sederoff RR: **Phenylalanine ammonia-lyase from loblolly pine: purification of the enzyme and isolation of complementary DNA clones.** *Plant Physiol* 1992, **98**:380-386.
20. Butland SL, Chow ML, Ellis BE: **A diverse family of phenylalanine ammonia-lyase genes expressed in pine trees and cell cultures.** *Plant Mol Biol* 1998, **37**:15-24.
21. Xu F, Cai R, Cheng S, Du H, Wang Y, Cheng S: **Molecular cloning, characterization and expression of phenylalanine ammonia-lyase gene from Ginkgo biloba.** *African J of Biotech* 2008, **7**:721-729.
22. Calabrese JC, Jordan DB, Boodhoo A, Sariaslani S, Vannelli T: **Crystal structure of phenylalanine ammonia lyase: multiple helix dipoles implicated in catalysis.** *Biochemistry* 2004, **43**:11403-11416.
23. Hamberger B, Ellis M, Friedmann M, Souza C, Barbazuk B, Douglas CJ: **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.** *Can J Bot* 2007, **85**:1182-1201.
24. Chen K, Durand D, Farach-Colton M: **NOTUNG: a program for dating gene duplications and optimizing gene family trees.** *J Comput Biol* 2000, **7**:429-447.
25. Reichert AI, He X, Dixon RA: **Phenylalanine ammonia-lyase (PAL) from tobacco (*Nicotiana tabacum*): characterization of the four tobacco PAL genes and active hetrotetrameric enzymes.** *Biochemistry* 2009, **424**:233-242.
26. Kumar A, Ellis BE: **The phenylalanine ammonia-lyase gene family in raspberry. Structure, expression, and evolution.** *Plant Physiol* 2001, **127**:230-239.
27. Wanner LA, Li G, Ware D, Somssich IE, Davis KR: **The phenylalanine ammonia-lyase gene family in *Arabidopsis thaliana*.** *Plant Mol Biol* 1995, **27**:327-338.
28. Surget-Groba Y, Montoya-Burgos JI: **Optimization of de novo transcriptome assembly from next-generation sequencing data.** *Genome Res* 2010, **20**:1432-1440.
29. Pina A, Errea P: **Differential induction of phenylalanine ammonia-lyase gene expression in response to in vitro callus unions of Prunus spp.** *J Plant Physiol* 2008, **165**:705-714.
30. Okada T, Mikage M, Sekita S: **Molecular characterization of the phenylalanine ammonia-lyase from *Ephedra sinica*.** *Biol Pharm Bull* 2008, **31**:2194-2199.
31. Logemann E, Parniske M, Hahlbrock K: **Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley.** *Proc Natl Acad Sci USA* 1995, **92**:5905-5909.
32. De Paolis A, Pignone D, Morgese A, Sonnante G: **Characterization and differential expression analysis of artichoke phenylalanine ammonia-lyase-coding sequences.** *Physiol Plant* 2008, **132**:33-43.
33. Reddy JT, Korth KL, Wesley SV, Howles PA, Rasmussen S, Lamb C, Dixon RA: **Post-transcriptional regulation of phenylalanine ammonia-lyase expression in tobacco following recovery from gene silencing.** *Biol Chem* 2000, **381**:655-665.
34. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller W, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res* 2004, **14**:1147-1159.
35. **Conifer DBMagic Database.** [http://ancangio.uga.edu/ng-genediscovery/ptaeda.jnlp].
36. Benson DA, Boguski MS, Lipman DJ, Ostell J: **GenBank.** *Nucleic Acids Res* 1997, **25**:1-6.
37. Dong Q, Schlueter SD, Brendel V: **PlantGDB, plant genome database and analysis tools.** *Nucleic Acids Res* 2004, **32**:D354-D359.
38. Wall PK, Leebens-Mack J, Muller KF, Field D, Altman NS, dePamphilis CW: **PlantTribes: a gene and gene family resource for comparative genomics in plants.** *Nucleic Acids Res* 2008, **36**:D970-D976.
39. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res* 2002, **30**:3059-3066.
40. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Res* 2006, **34**:W609-W612.
41. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**:2688-2690.
42. Huelsenbeck JP, Ronquist F: **MrBayes: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**:754-755.
43. Lanave C, Preparata G, Sacone C, Serio G: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20**:86-93.
44. Yang Z: **Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites.** *Mol Biol Evol* 1993, **10**:1396-1401.
45. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8**:275-282.

doi:10.1186/1471-2164-13-S3-S1
Cite this article as: Bagal *et al.*: The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage. *BMC Genomics* 2012 **13**(Suppl 3):S1.