American Journal of Bioinformatics

Original Research Paper

Illumina-Based *De Novo* Transcriptome Analysis and Identifications of Genes Involved in the Monolignol Biosynthesis Pathway in *Acacia koa*

Kazue Ishihara, Eric K.W. Lee, Isabel Rushanaedy and Dulal Borthakur

Department of Molecular Biosciences and Bioengineering, University of Hawaii, Honolulu, HI 96822, USA

Article history
Received: 09-06-2015
Revised: 29-06-2015
Accepted: 29-06-2015

Corresponding Author:
Dulal Borthakur
Department of Molecular Biosciences and Bioengineering, University of Hawaii, Honolulu, HI 96822, USA
Email: dulal@hawaii.edu

Abstract: *Acacia koa* is a leguminous timber tree endemic to the Hawaiian Islands. For breeding projects involved in improving wood quality of *A. koa*, understanding of genes influencing wood quality is crucial. Therefore, the objective of this study was to identify *A. koa* genes in the monolignol biosynthesis pathway, which is involved in wood formation and development. In this study, whole transcritpome sequencing of *A. koa* seedlings was performed through Illumina-based sequencing and over 88 million high-quality paired-end raw reads were generated. Trinity *de novo* assembly of those reads yielded 85,533 unigenes with an average length of 641 bp. Based on sequence similarity search with known proteins, we annotated 47,038 of the unigenes. Using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, 149 unigenes were assigned to ortholog groups of enzymes involved in the monolignol biosynthesis pathway. In addition, we identified complete coding sequences of genes for all the ten identified enzymes of the pathway. Future studies on expression levels of these genes in *A. koa* with different wood qualities will provide a tool for selection of desirable types. Comprehensive sequence resources of *A. koa* generated through this study will contribute to genomic studies and improvement programs for this tree.

Keywords: *Acacia koa*, Tree Legume, Transcriptome Analysis, Lignin, Monolignol Biosynthesis

Introduction

Acacia koa is an important leguminous tree endemic to the Hawaiian Islands. The native *A. koa* forests are broadly distributed across all five major Hawaiian Islands (Wagner *et al*., 1990). The *A. koa* populations in these islands are genetically diverse and can be divided into morphologically distinguishable groups of koa, koaia and an intermediate type (Adamski *et al*., 2012). *A. koa* serves as an ecologically and economically vital resource for the Hawaiian Islands. It provides a habitat for many native fauna and flora (Elevitch *et al*., 2006; Sakai, 1988; Whitesell, 1990). In addition, due to the beautiful texture, hardiness, and carving quality of the wood, the *A. koa* timber, referred to as Hawaiian mahogany, is a high priced commodity with a current market value of up to $125 per board foot (Baker *et al*., 2009). *A. koa* wood is used for fine furniture, decorative items, musical instruments, and jewelry. The gross value of the *A. koa* timber and wood products produced is estimated to be in the range of $20-$30 million annually (Baker *et al*., 2009; Yanagida *et al*., 2004).

Because of high value of *A. koa* wood, it is crucial to understand the factors affecting wood formation and development. There have been many studies to identify factors that affect wood qualities, including wood density, wood color, stiffness and orientation and morphology of fiber. Although wood quality is a highly complex trait, in recent years, technologies such as gene mapping, sequencing and microarrays have been developed to understand molecular mechanisms underlying it. Once candidate genes are identified, they can be used as markers for selection of seedlings with desirable traits at early stages. However, there are only a limited number of nucleotide sequences publicly available for *A. koa*. To identify genes for wood formation and development in *A. koa*, sequencing of the

genome will be necessary. Given that *A. koa* is an allotetraploid, it is substantially challenging to sequence and assemble its complex genome (Hamilton and Buell, 2012). For such cases, transcriptome sequencing provides an alternative to the whole genome sequencing.

In the present study, we utilized Illumina *de novo* sequencing technology to characterize the transcriptome of *A. koa*. Our objectives were to enrich the gene resource of *A. koa* with the sequencing data and to identify the transcripts involved in the monolignol biosynthesis pathway, which may be related to wood formation and development in *A. koa,* as lignin is one of the major constituents of wood. To the best of our knowledge, this study is the first exploration to characterize the transcriptome of *A. koa*. The transcriptome sequencing of *A. koa* will offer valuable sequence resources and contribute to further research on functional genomics and improvement of *A. koa*.

Materials and Methods

Plant Materials and RNA Extraction

Two and a half month old *A. koa* seedlings were obtained from the Maunawili sub-center of Hawaii Agriculture Research Center (HARC), Kailua, HI. Total RNA was extracted from the whole seedlings using RNeasy Plant Mini Kit (Qiagen) and purified with TURBO DNA-free Kit (Ambion). The quality and quantity of the RNA were assessed using NanoDrop Spectrophotometer (ND-1000).

Library Construction, Sequencing and Assembly

Cofactor Genomics, St. Louis, MO conducted cDNA library construction, sequencing and assembly. Sequencing was performed through the Illumina platform (Illumina Genome Analyzer IIx) with 60 bp paired-end reads. The quality of the raw reads were assessed through FASTQC to make sure more than 90% of the bases have Q20 or higher and were assembled using a *de novo* assembly program Trinity (http://trinityrnaseq.sourceforge.net/) (Grabherr *et al*., 2011). The resulting assembled sequences were defined as unigenes. Assembled sequences with lengths ≥200 bp were included in the downstream analysis.

Functional Annotations of Unigenes

The assembled unigene sequences were compared against multiple protein databases, including the NCBI non-redundant (nr) database, the Swiss-Prot database, the Translated European Molecular Biology Laboratory (TrEMBL) database, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and the Clusters of Orthologous Group (COG) database, through the Basic Local Alignment Search Tool (BLAST) algorithm with a cut-off E-value of 1E-3, using the *doblast* server of the Noble Foundation (http://bioinfo3.noble.org/doblast/) and the WebMGA server (http://weizhong-lab.ucsd.edu/metagenomic-analysis/) (Wu *et al*., 2011). Gene names were assigned to each query based on the highest sequence similarity. A Java program Blast2Go (Conesa *et al*., 2005) was utilized to assign Gene Ontology (GO) functional categories for the annotated unigenes. The COG database, which classifies orthologous gene products, was used to categorize the annotated unigenes into 26 general functional groups. With the KEGG database, which contains systematic analysis of biochemical pathways and functions of the gene products, unigenes involved in the monolignol biosynthesis pathway were identified. The BLASTX analysis was performed to confirm the sequence identities of some unigenes in the ortholog groups and to detect unigenes with a complete Open Reading Frame (ORF). NCBI ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/) was used to determine the ORFs and the protein sequences of the unigenes.

Putative SSR Molecular Markers

For development of new molecular markers, the annotated unigenes were used to identify potential simple sequence repeats (SSRs). With the MISA Perl script (http://pgrc.ipk-gatersleben.de/misa/), motifs of di- to hexanucleotide with a minimum of four repetitions and compound motifs, in this case, motifs interrupted by sequences of up to 100 bp, were also identified.

Results

Sequence Analysis and Assembly

In this study, a total of 88,983,363 paired-end raw reads were generated from a 250-bp insert library. These reads contained 97.66% Q20 bases (base quality 20) and were used for *de novo* assembly. The raw reads were deposited on the NCBI Sequence Read Archive (SRA) with an accession number SRR1686818. Using the Trinity *de novo* assembly software, 85,533 unigenes were generated with a total length of 45.82 Mb, an average length of 640.97 bp and an N50 length of 1,068 bp (Table 1). Of these, 15,022 (17.56%) were >1 kb, 14,090 (16.47%) were 500-999 bp and 56,421 (65.96%) were 200-499 bp (Table 2).

Table 1. Summarized assembly statistics for unigenes in *A. koa*

Statics	Number
Total number of paired-end reads	88,983,363
Total number of assembled unigenes	85,533
Total length of unigenes (bp)	54,824,004
Mean length of unigenes (bp)	641
Median length of unigenes (bp)	345
Max length of unigenes (bp)	13,405
N50 length of unigenes (bp)	1,068

Functional Annotation

The 85,533 unigenes were searched against diverse protein databases, including the nr database, the Swiss-Prot database, the TrEMBL database, the KEGG database and the COG database, using the BLAST algorithm (E-value <1E-3). The annotation with the TrEMBL database had the highest aligned unigenes (46,146 unigenes), followed by the annotation with the nr database (45,800 unigenes). With the two databases combined, a total of 46,782 unigenes (54.69%) were annotated. The number of unigenes that showed homology with sequences in the Swiss-Prot, KEGG and COG databases were 33,113, 26,024 and 20,288, respectively. Overall, a total of 47,038 unigenes (54.99%) were successfully annotated using nr, TrEMBL, Swiss-Prot, KEGG and COG (Table 3). Among the unigenes annotated in the nr and TrEMBL databases, 50.24% and 42.75%, respectively, had an E-value <1.0E-50, showing strong homology; however, only 33.53% of the unigenes annotated in the Swiss-Prot database had an E-value <1.0E-50 (Fig. 1).

Table 2. Length distribution of *de novo* assembled unigenes in *A. koa*

Length (bp)	Number of unigenes	Frequency (%)
200-299	35,240	41.20
300-499	21,181	24.76
500-999	14,090	16.47
1,000-1,499	6,259	7.32
1,500-1,999	3,948	4.62
2,000-2,499	2,185	2.55
2,500-2,999	1,127	1.32
3,000	1,503	1.76

Table 3. Summary for the annotation of unigenes of *A. koa (cutoff <1.0E-3)*

	Number of unigenes	Number of functional categories
Gene annotation against nr	45,800	-
Gene annotation against Swiss-Prot	33,113	-
Gene annotation against TrEMBL	46,146	-
Gene annotation against KEGG	26,024	-
Total gene annotation against nr and TrEMBL	46,782	-
GO annotation for nr and TrEMBL protein hits	20,884	52
KEGG pathway mapping for nr and TrEMBL protein hits	12,646	208
COG functional classification for nr and TrEMBL protein hits	18,320	26
Total annotated unigenes	46,891	-

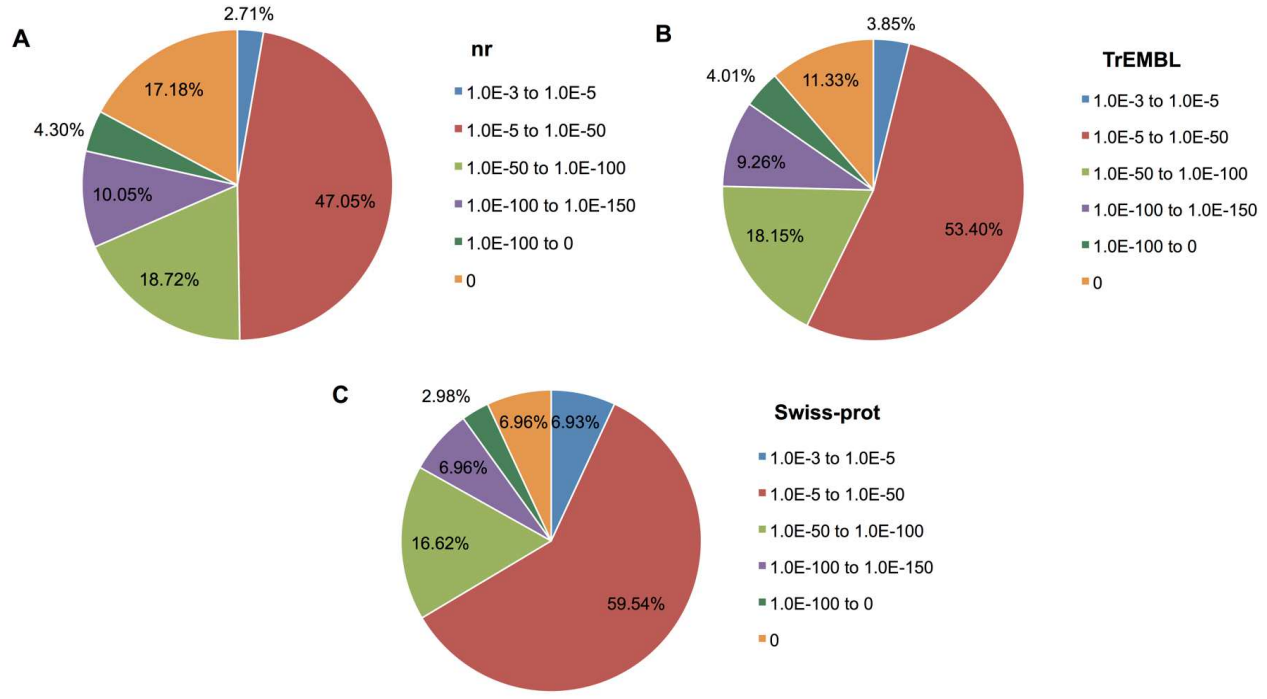

Fig. 1. E-value distributions of annotated *A. koa* unigenes. The E-values of the highest-scored BLAST hit was identified for each unigene by aligning against (A) the nr protein database, (B) the TrEMBL protein database and (C) the Swiss-Prot protein database for each unigene

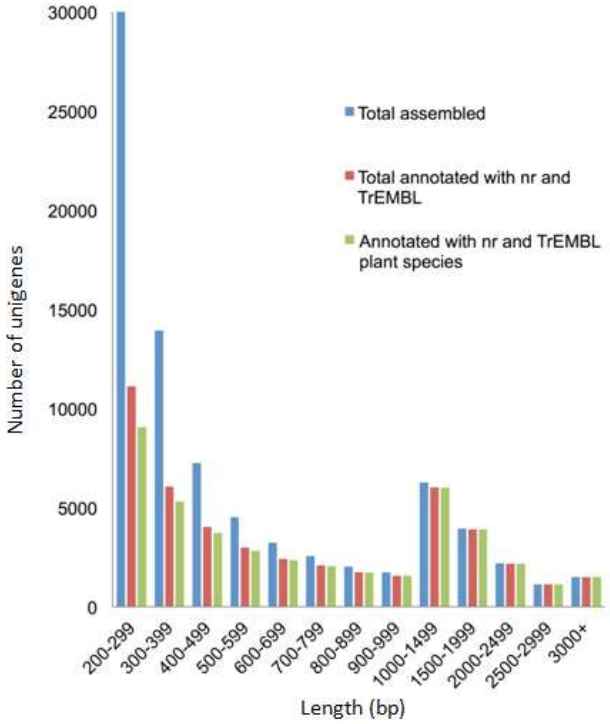

Fig. 2. Length distributions of assembled unigenes. Blue bars represent the total number of assembled unigenes. Red bars represent the total number of unigenes annotated by nr and TrEMBL. Green bars represent the total number of unigenes that have high similarities to known plant proteins. The two peak-pattern of the graph is due to two ranges of data used; the distribution range was 100 bp for unigenes with lengths of up to 1 kb and 500 bp for unigenes above 1 kb

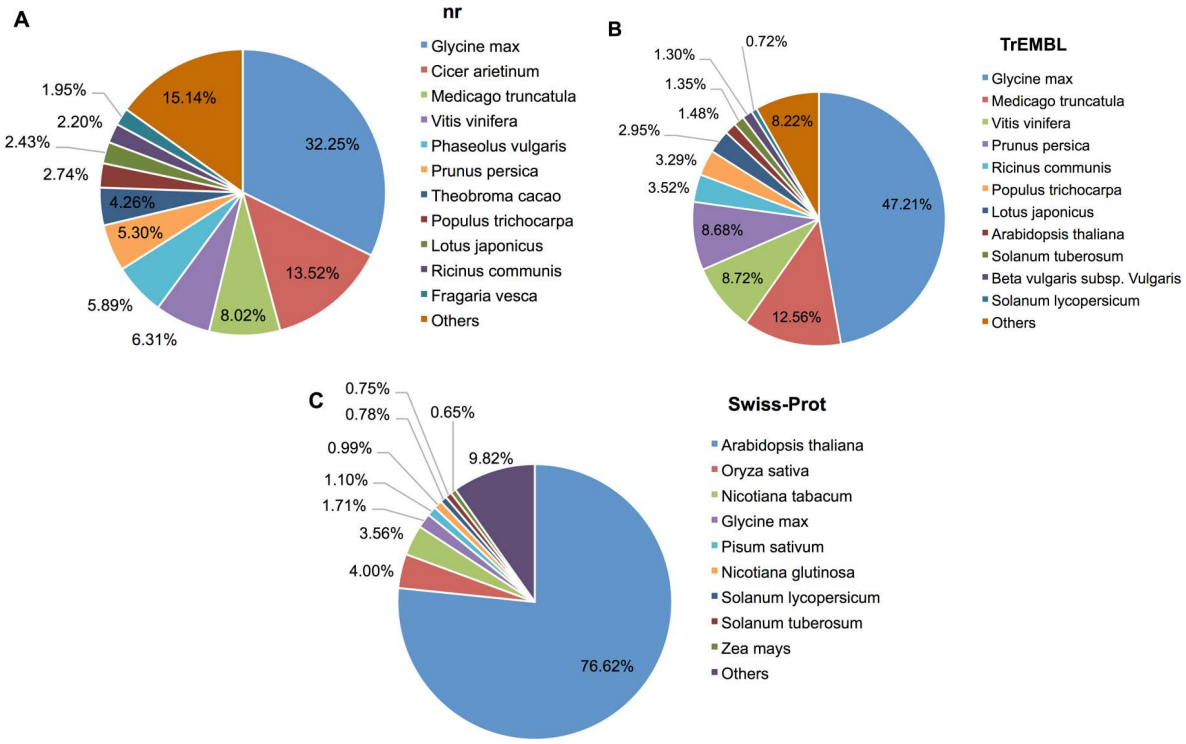

Fig. 3. Top hit plant species distribution of unigenes. The unigenes were annotated against (A) the nr protein database (B) the TrEMBL protein database and (C) the Swiss-Prot protein database

The annotation rate decreased as the lengths of unigenes decreased; 98.16% of unigenes with the length of ≥1,000 bp showed homologous matches, whereas the annotation rates for unigenes with a length of 500-999 bp and unigenes <500 bp were 78.78 and 37.60%, respectively (Fig. 2). The majority of the unigenes without annotations from the nr and TrEMBL databases were <500 bp, as 62.25 and 28.59% were 200-299 bp and 300-499 bp, respectively, totaling 90.84%. The reason for this was most likely their short sequence lengths, resulting statistically insignificant matches.

For the plant species distribution, the most represented species in the unigenes aligned in the nr database were all legumes: *Glycine max* (32.25%), *Cicer arietinum* (13.52%) and *Medicago truncatula* (8.02%) (A). Similarly, more than half of the unigenes that matched with sequence in the TrEMBL database showed homology with legumes, including *G. max* (47.21%) and *M. truncatula* (8.72%) (Fig. 3B). Since the Swiss-Prot database contains only manually reviewed protein sequences, a higher percentage of the unigenes (76.62%) showed homology with the well-studied *Arabidopsis thaliana* sequences (Fig. 3C). Considering the E-value and plant species distributions, the annotations of the unigenes with the nr and TrEMBL databases gave consistent results. The 46,782 unigenes annotated from the nr and TrEMBL databases were used for further analysis. Additionally, we found that a total of 3,473 unigenes (3,262 and 3,442 annotated with the nr and TrEMBL databases, respectively) had homology to sequences with non-plant origins, such as *Staphylococcus* and *Drosophila*. Approximately 90.09% (3,128 unigenes) of those were 200-499 bp in length; 9.04% (314 unigenes) were 500-999 bp and 0.86% (30 unigenes) were ≥1,000 bp (Fig. 2). These sequences were considered contaminants and removed and the remaining 43,309 unigenes were used for sequence classifications.

GO Classification

Of the 43,309 unigenes, 20,884 were classified into 3 GO functional categories: biological process, cellular component, and molecular function (Fig. 4). In the biological process category, the unigenes were further clustered into 20 subcategories. Of those, the largest was metabolic process (23.99%); the second was cellular process (19.62%), and the third was single-organism process (14.25%). Under the cellular component category, the unigenes were assigned to 16 subcategories; the most abundant classes were cell (21.85%), cell part (21.84%), and organelle (15.15%). The unigenes under the molecular function category were sorted into 6 subcategories; the most represented ones were binding activity (44.33%), catalytic activity (42.69%), and transporter activity (4.76%).

COG Classification

Using the COG database, 18,320 unigenes of the 43,309 annotated ones were classified into 26 functional categories (Table 3 and Fig. 5). Some of the unigenes were classified into more than one category. The category for 'signal transduction mechanisms' (3,224 unigenes) represented the largest group, followed by 'general function prediction only' (2,358 unigenes), 'posttranslational modification, protein turnover, chaperones' (1,979 unigenes), 'transcription' (1,242 unigenes), 'function unknown' (1,218 unigenes), 'carbohydrate transport and metabolism' (1,116 unigenes), 'intracellular trafficking, secretion, vesicular transport' (1,093 unigenes), and 'secondary metabolites biosynthesis, transport, and catabolism' (980 unigenes).

KEGG Pathway Classification

The KEGG database provides systemic functional information of biochemical pathways and functions of gene products. From the 43,309 annotated unigenes, 12,646 unigenes were grouped into 208 KEGG biochemical pathways (Fig. 6). Major KEGG biochemical pathway groups were metabolism (5,729 unigenes), genetic information processing (2,707 unigenes), environmental information and processing (507 unigenes), cellular processes (1,381 unigenes), and organismal system (1,306 unigenes). The largest metabolic pathway groups include carbohydrate metabolism (1,269 unigenes), nucleotide metabolism (1,100 unigenes), and amino acid metabolism (836 unigenes). The pathways related to genetic information processing involved folding, sorting, and degradation (1,179 unigenes), translation (939 unigenes), and replication (306 unigenes). Biochemical pathways for cellular processes were most represented by pathways for cell growth and death (787 unigenes) and transport (487 unigenes).

Identification of Genes Involved in the Monolignol Biosynthesis Pathway

Through the KEGG pathway analysis, we identified a total of 149 orthologs for all the ten enzymes involved in monolignol biosynthesis, namely phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), *p*-coumarate: CoA ligase (4CL), cinnamoyl CoA reductase (CCR), cinnamyl alcohol dehydrogenase (CAD), *p*-coumarate 3-hydroxylase (C3H), hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase (HCT), caffeic acid O-methyltransferase (COMT), ferulate 5-hydroxylase (F5H) and caffeoyl CoA 3-O-methyltransferase (CCoAOMT) (Table S1). Through BLASTX analysis, we confirmed the unigenes in these ortholog groups. We also identified the complete coding sequences

(CDSs) in 19 unigenes representing isoforms of the monolignol biosynthesis enzymes. They all had sequence similarities of >70% with other legume species (Table 4). The CDSs of the unigenes were deposited at NCBI Transcriptome Shotgun Assembly (TSA) under the accession number GBYE00000000.

SSR Identification

For development of new molecular markers, the 43,309 annotated unigenes were used to identify potential Simple Sequence Repeats (SSRs). With the MISA Perl script, we searched for di- to hexa-nucleotides with a minimum of four repetitions and identified 13,109 unigenes containing a total of 20,755 putative SSRs. Among them, 4,731 unigenes had more than one SSR (Table 5). Of those, 2,699 had SSRs in compound formation, having two or more consecutive SSRs interrupted by less than 100 bp. In total, we detected 111 different motifs. Di-nucleotide repeats except CG/CG (0.35%) were the most abundant (71.95%), and tri-nucleotide was the second abundant (26.95%) (Table 6). The dominant repeat motif was AG/CT (45.28%), followed by AT/AT (13.72%), AC/GT (12.45%), and AAG/CTT (8.78%) (Fig. 7).

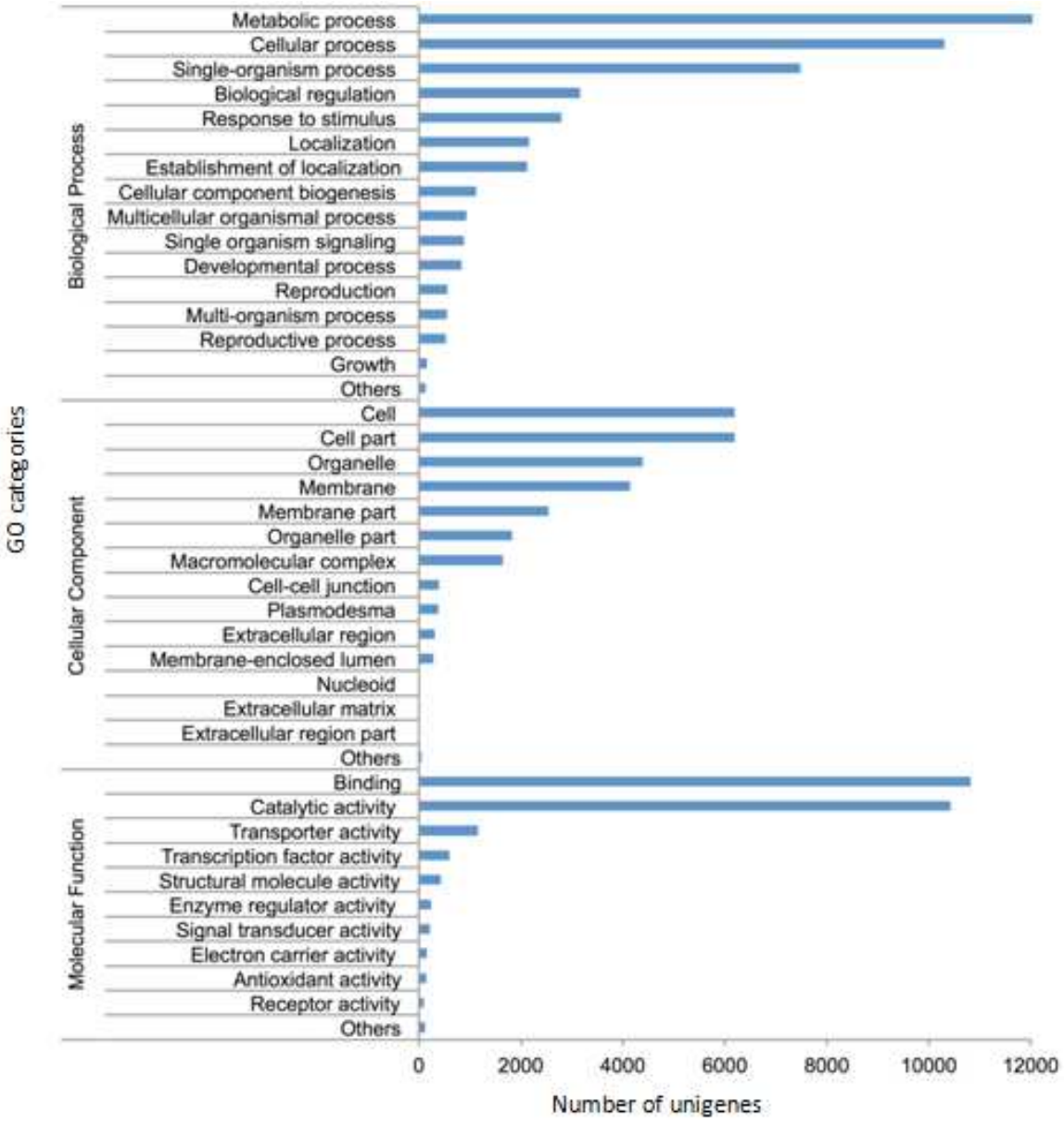

Fig. 4. Gene Oontology (GO) functional categorization of the unigenes annotated against the nr and TrEMBL databases. A total of 20,884 unigenes were classified into 3 main GO categories and 42 sub-categories

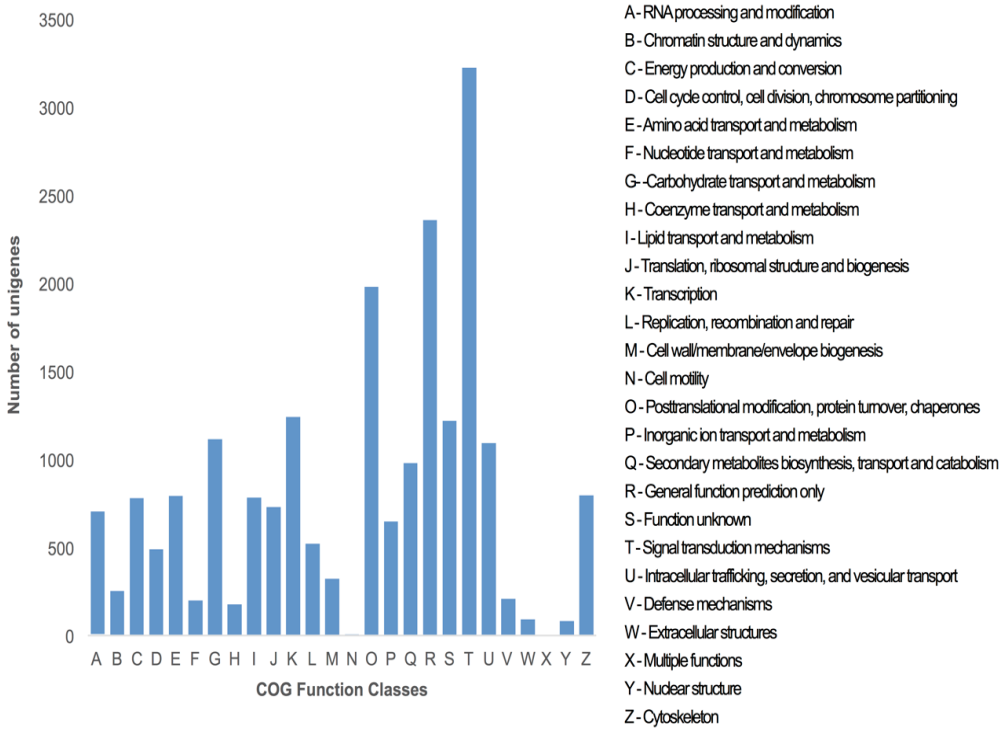

Fig. 5. Clusters of Orthologous Groups (COG) of unigenes annotated against the nr and TrEMBL databases. A total of 18,320 unigenes were classified into 26 functional categories. Some of the unigenes were assigned to more than one category

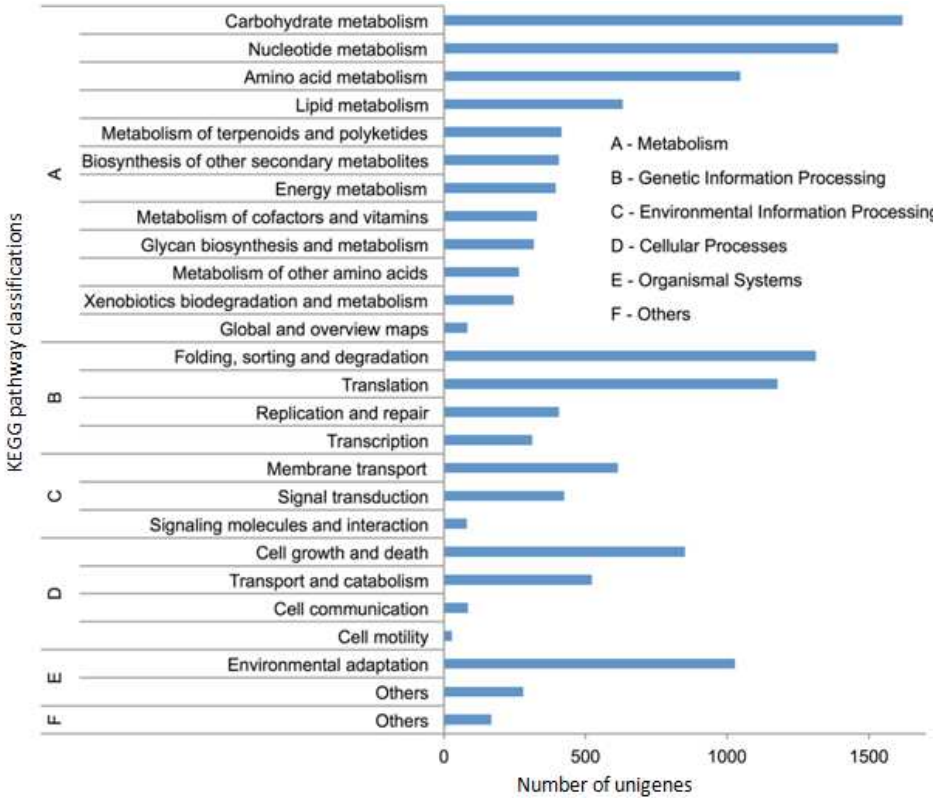

Fig. 6. KEGG pathway classification of unigenes annotated against the nr and TrEMBL databases. A total of 12,646 unigenes were grouped into 208 KEGG biochemical pathways

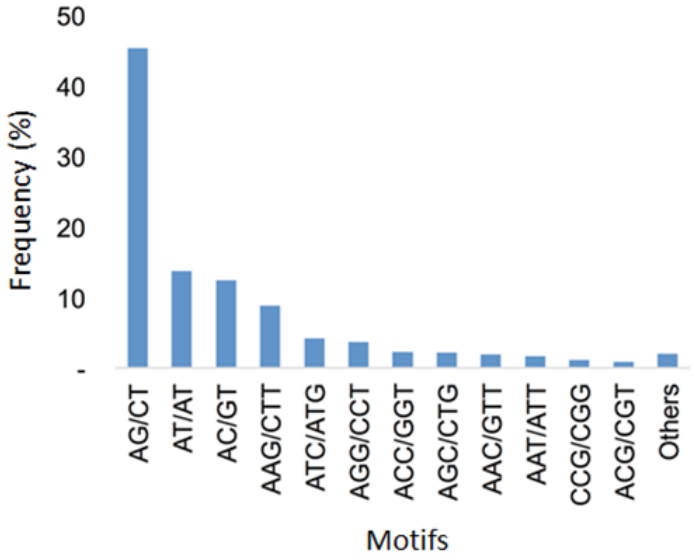

Fig. 7. Frequency distribution of SSR based on motif sequence types. A total of 111 motifs were identified

Table 4. Complete coding sequences of the monolignol biosynthesis pathway in *A. koa*. They were deposited at NCBI Transcriptome Shotgun Assembly (TSA) under the accession number GBYE00000000

Function	Blast hit organism	Sequence similarity*	Blast Hit Acc. No.
Phenylalanine ammonia-lyase (PAL)			
PAL1	*Acacia acuriculiformis x Acacia mangium*	95% (83%)	ABD42947.1
PAL2	*Glycine max*	90% (86%)	NP_001236956.1
Cinnamate 4-hydroxylase (C4H)			
C4H1	*Leucaena leucocephala*	95% (85%)	AEM63594.1
C4H2	*G max*	88% (59%)	XP_003555891.1
p-coumarate:CoA ligase (4CL)			
4CL1	*G max*	84% (68%)	NP_001237270.1
4CL2	*G max*	81% (72%)	XP_003545004.1
4CL3	*L. leucocephala*	92% (69%)	ACI23349.1
4CL4	*L. leucocephala*	89% (63%)	ACI23348.1
Cinnamoyl CoA reductase (CCR)			
CCR1	*L. leucocephala*	93% (70%)	CAK22319.1
CCR2	*Cicer arietinum*	71% (61%)	XP_004515542.1
Cinnamyl alcohol dehydrogenase (CAD)			
CAD1	*A. acuriculiformis x A. mangium*	97% (67%)	ABX75855.1
CAD2	*C. arietinum*	83% (75%)	XP_004485621.1
p-coumarate 3-hydroxylase (C3H)			
C3H	*Caragana korshinskii*	91% (80%)	AEV93473.1
Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase (HCT)			
HCT	*L. leucocephala*	94% (78%)	AGA20364.1
Caffeic acid O-methyltransferase (COMT)			
COMT	*A. acuriculiformis x A. mangium*	97% (74%)	AAY86361.1
Ferulate 5-hydroxylase (F5H)			
F5H	*L. leucocephala*	89% (72%)	ABS53040.1
Caffeoyl CoA 3-O-methyltransferase (CCoAOMT)			
CCoAOMT1	*Leucaena leucocephala*	98% (89%)	ABE60812.1
CCoAOMT2	*A. acuriculiformis x A. mangium*		ABX75853.1
CCoAOMT3	*Musa acuminate*	73% (54%)	XP_009413347.1

*Numbers in parentheses show homologies with *Arabidopsis thaliana*

Table 5. Summary of SSR searching results

Searching items	Number
Total number of sequences examined	43,309
Total size of examined sequences (bp)	41,338,838
Total number of identified SSRs	20,755
SSR containing sequences	13,109
Sequences containing more than 1 SSR	4,731
SSRs present in compound formation	2,699
Di- nucleotide	14,901
Tri- nucleotide	5,594
Tetra-nucleotide	153
Penta-nucleotide	35
Hexa-nucleotide	72

Table 6. Length distribution of SSR based on the number of repeat units

Number of repeat unit	Di-	Tri-	Tetra-	Penta-	Hexa-
4	11220	3885	131	28	56
5	1927	1143	14	7	16
6	705	411	6	0	0
7	399	127	2	0	0
8	247	21	0	0	0
9	172	7	0	0	0
10	148	0	0	0	0
11	72	0	0	0	0
12	10	0	0	0	0
13	1	0	0	0	0
Total	14901	5594	153	35	72

Discussion

Transcriptome Sequencing and Assembly

Despite numerous studies on genomes of various legume species, only a limited number of nucleotide or protein sequences of *A. koa* have been reported, and almost no genomic information is available in public databases. As a majestic timber tree, *A. koa* could be a rich source of genes for tree improvement programs. Thus, the objective of this study was to produce a global overview of the whole transcriptome of *A. koa*. After comparing with the five databases and filtering out all of the mostly small, unannotated sequences, a total of 43,309 unigenes were identified in this study. A large proportion of smaller unigenes obtained through Illumina sequencing may be due to the allotetraploid \genome of *A. koa* (2n = 52); homeologous or paralogous gene copies can be distinct yet highly similar, possibly causing incomplete assembly (Duan *et al*., 2012; Gruenheit *et al*., 2012; Nakasugi *et al*., 2014). In spite of being an allotetraploid species, both the average length and the N50 length obtained from *A. koa* in the present study were greater than those obtained from other related diploid legume species, such as *Acacia auriculifomis* (496 and 949 bp), *Acacia mangium* (498 and 938 bp) (Wong *et al*., 2011), and *Cicer arietinum* (523 and 900 bp) (Garg *et al*., 2011). Also, the total number and cumulative length of unigenes of *A. koa* were more than twice of those of *A. auriculiformis* (42,217 unigenes and 21.02 Mb) and *A. mangiu*m (35,759 unigenes and 17.84 Mb) (Wong *et al*., 2011). These differences may be also due to their ploidy levels, as *A. auriculiformis* and *A. mangium* are both diploid (2n = 26), in addition to the use of different assembly software in those studies.

Genes Involved in the Monolignol Biosynthesis Pathway in A. koa

In this study, we identified genes in the monolignol biosynthesis pathway in *A. koa* because the pathway is involved in wood formation and development. Monolignols, which consist of *p*-hydroxyphenyl (H), guaiacyl (G), and syringyl (S) units, are the building blocks of lignin. Lignin constitutes 25-35% of the secondary cell wall (Plomion *et al*., 2001) and lignin composition is one of the major determinants of physical characteristics of wood (Novaes *et al*., 2010). Various studies have shown that the downregulation of upstream biosynthesis genes PAL, C4H, and 4CL results in lower content and altered composition of lignin in *N. tabacum, A. thaliana* and *Populus tremuloides* (Bate *et al*., 1994; Elkind *et al*., 1990; Hu *et al*., 1999; Kajita *et al*., 1997; Lee *et al*., 1997; Sewalt *et al*., 1997). Silencing a 4CL gene in *Pinus taeda* also reduced the G/H ratio, making it similar to that of compression wood (Wagner *et al*., 2009). Other enzymes in the pathway also affect lignin content and composition of wood. For instance, in transgenic *Populus*, the downregulation of C3H decreased lignin

levels by half and highly increased the proportion of H units (Ralph *et al*., 2012). The repression of CCoAOMT also reduced lignin production and increased S/G ratio in *Zea mays* and *Populus* (Li *et al*., 2013; Meyermans *et al*., 2000). The downregulated COMT expression reduced content of S units in *N. tabacum* and *Populus* (Atanassova *et al*., 1995; Jouanin *et al*., 2000; Pincon *et al*., 2001). In *P. taeda,* a mutation in the CAD gene causes a decline in CAD protein, resulting in lower lignification, higher wood density, and increased stem-growth, thus affecting wood quality (Gill *et al*., 2003; Yu *et al*., 2005; Wu *et al*., 1999). Also, some monoligninol biosynthesis genes (4CL, C4H, C3H and CCoAOMT) matched with quantitative trait loci (QTL) for wood density in *P. taeda* (Brown *et al*., 2003). In addition, the monolignol biosynthesis pathway is part of the phenylpropanoid pathway, which generates a wide variety of secondary metabolites, such as flavonoids and tannins (Fig. 8), so regulations of monolignol biosynthesis enzymes can affect other metabolite production. For example, repression of a HCT gene in *A. thaliana* resulted in accumulation of flavonoids, such as anthocyanin (Besseau *et al*., 2007). Based on these published reports, we expect wood quality attributes, such as wood density and wood color, are determined through differential expression of the genes encoding enzymes in the monolignol biosynthesis pathway.

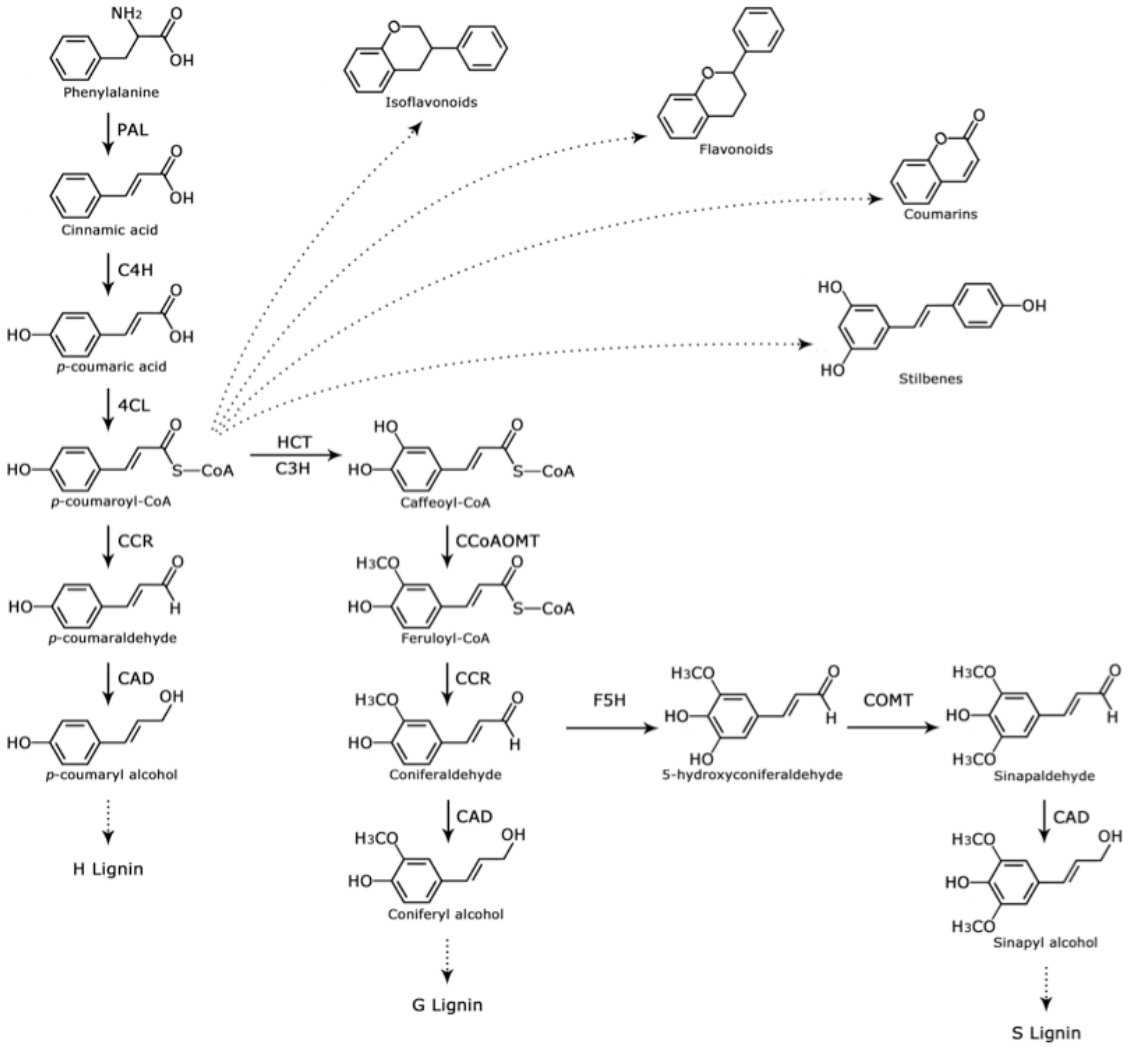

Fig. 8. Monolignol biosynthesis pathway showing synthesis of isoflavonoids, flavonoids, coumarins, stilbenes and lignins from *p*-coumaroyl-CoA. Because of the variety of isoenzymes and kinetic properties, alternative routes through the metabolic pathway may exist. Dashed arrows represent multiple reaction steps. PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydoxylase; 4CL, p-coumarate:CoA ligase; CCR, cinnamoyl-CoA reductase; CAD, cinnamyl alcohol dehydrogenase; C3H, p-coumarate 3-hydroxylase; HCT, hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase; COMT, caffeic acid O-methyltransferase; F5H, ferulate 5-hydroxylase; CCoAOMT, caffeoyl-CoA O-methyltransferase

In the present study, 149 unigenes were assigned as orthologs of the enzymes involved in the monolignol biosynthesis pathway (Fig. 8). However, there are many closely related superfamily members ("like" proteins) and some of them may be unrelated to the monolignol biosynthesis pathway, so we excluded unigenes with <50% homology with *A. thaliana* monolignol genes and unigenes without important conserved amino acid motifs identified in previous studies (Ehlting *et al*., 2001; Hoffmann *et al*., 2003; Joshi and Chiang, 1998; Larsen 2004; Lynch *et al*., 2002; Mckie *et al*., 1993; Zubieta *et al*., 2002; Schuler, 1996; Wanner *et al*., 1995) (Fig. S1). We identified complete CDSs of genes for all the ten enzymes involved in monolignol biosynthesis. There may exist more isoforms in *A. koa* because conserved motifs could not be determined in some of the unigenes due to incomplete assembly. Future studies of *A. koa* involving significant variations in wood quality attributes will determine the level of expressions of key monolignol biosynthesis genes that result in specific phenotypes. Therefore, determining the expression levels of those key genes will be useful for selection for improved wood quality.

Putative SSR Molecular Markers

Next-generation sequencing (NGS) is a rapid and effective approach to identify SSR molecular markers in non-model organisms without known genomic sequences. The traditional approach to develop SSR molecular markers involves fragmentation of DNA, construction of genomic DNA libraries in *Escherichia coli*, PCR amplification, and sequencing of the amplified fragments (Sahu *et al*., 2012; Song *et al*., 2005). These steps are time-consuming and labor-intensive. Using NGS technology and bioinformatics, identification of numerous SSRs from the sequence data can be rapid and cost-effective. Previously, through the traditional approach, Fredua-Agyeman *et al*. (2008) analyzed 96 sequences and developed 31 primer pairs that targeted microsatellite loci in *A. koa*. Some of these primers successfully identified polymorphic loci and were also used to measure genetic diversity in *A. koa* (Adamski *et al*., 2012); yet, only limited number of genetic markers exists in *A. koa*, so the identification of more SSRs with NGS technology will be useful.

In the present study, we predicted 13,109 unigenes containing a total of 20,755 putative SSRs. In *A. koa*, dinucleotide repeats were the predominant motif as in many other plants, such as *A. thaliana*, *Arachis hypogaea*, *Brassica napus*, *Beta vulgaris*, *Brassica oleracea*, *G. max*, *Vitis vinifera*, and *Sesamum indicum* (Kumpatla and Mukhopadhyay, 2005; Wei *et al*., 2014). The frequency of SSR repeat motifs in *A. koa* obtained in this study was consistent with that of other plant species. The AG/CT repeat (45.28%) was the most abundant dinucleotide motif group and the CG/CG repeat (0.35%) was the smallest dinucleotide motif group in *A. koa* just as in various species studied by Jayashree *et al*. (2006) and Kumpatla and Mukhopadhyay (2005). The AAG/CTT repeat (8.78%) was the predominant trinucleotide motif in *A. koa*, and it was also predominant in other plants, including three legume species, *G. max*, *M. truncatula* and *Lotus japonicus* (Jayashree *et al*., 2006; Kumpatla and Mukhopadhyay, 2005). Our results provide a substantial number of SSRs; in future studies, we may be able to identify SSR loci linked to genes associated with wood properties.

Conclusion

This is the first comprehensive transcriptome-wide analysis of *A. koa* using NGS technology. Illumina sequencing and Trinity *de novo* assembly generated 85,533 unigenes, and we successfully annotated 43,309 of them. With the KEGG database, we identified complete coding sequences of all the ten genes involved in the monolignol biosynthesis pathway, which could be highly associated with wood formation and development in *A. koa*. Further characterization of these genes will contribute to a deeper understanding of wood quality in *A. koa*. In addition, we predicted a significant number of potential SSR markers from our transcriptome data. Our results will be a valuable resource for future genetic studies and improvement programs of *A. koa*.

Acknowledgement

We would like to acknowledge Tyler Jones and Nick Dudley from HARC for providing samples, Bradley Porter for helpful discussion, and Patrick Zhao from Noble Foundation for his assistance in data annotation using the *doblast* sever. This research was supported primarily by the McIntire-Stennis Grant HAW00597-M. KI is supported by a Monsanto Graduate Fellowship.

Author Contributions

Kazue Ishihara: She contributed in analyzing data and writing of this manuscript.

Eric K. W. Lee: He contributed in the computational analyses of data and assist in writing of this manuscript.

Isabel Rushanaedy: She contributed in important initial experiments, including growing plants and extracting RNA to acquire raw data.

Dulal Borthakur: He contributed in conceptualization through a thorough discussion and also in assistance in writing of this manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Adamski, D.J., N.S. Dudley, C.W. Morden and D. Borthakur, 2012. Genetic differentiation and diversity of *Acacia koa* populations in the Hawaiian Islands. Plant Species Biol., 27: 181-190. DOI: 10.1111/j.1442-1984.2011.00359.x

Atanassova, R., N. Favet, F. Martz, B. Chabbert and M.T. Tollier *et al*., 1995. Altered lignin composition in transgenic tobacco expressing O-methyltransferase sequences in sense and antisense orientation. Plant J., 8: 465-477. DOI: 10.1046/j.1365-313X.1995.8040465.x

Baker, P.J., P.G. Scowcroft and J.J. Ewel, 2009. Koa (*Acacia Koa*) ecology and silviculture. General Technical Report - Pacific Southwest Research Station, USDA Forest Service.

Bate, N.J., J. Orr, W. Ni, A. Meromi and T. Nadler-Hassar *et al*., 1994. Quantitative relationship between phenylalanine ammonia-lyase levels and phenylpropanoid accumulation in transgenic tobacco identifies a rate-determining step in natural product synthesis. Proc. Nat. Acad. Sci. USA, 91: 7608-7612. DOI: 10.1073/pnas.91.16.7608

Besseau, S., L. Hoffmann, P. Geoffroy, C. Lapierre and B. Pollet *et al*., 2007. Flavonoid accumulation in Arabidopsis repressed in lignin synthesis affects auxin transport and plant growth. Plant Cell, 19: 148-162. DOI: 10.1105/tpc.106.044495

Brown, G.R., D.L. Bassoni, G.P. Gill, J.R. Fontana and N.C. Wheeler *et al*., 2003. Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda L.*). III. QTL verification and candidate gene mapping. Genetics, 164: 1537-1546. DOI: 10.1007/s001220100697

Conesa, A., S. Götz, J.M. García-Gómez, J. Terol and M. Talón *et al*., 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21: 3674-3676. DOI: 10.1093/bioinformatics/bti610

Duan, J., C. Xia, G. Zhao, J. Jia and X. Kong, 2012. Optimizing *de novo* common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genom., 13: 392-392. DOI: 10.1186/1471-2164-13-392

Ehlting, J., J.J. Shin and C.J. Douglas, 2001. Identification of 4-coumarate: Coenzyme A Ligase (4CL) substrate recognition domains. Plant J., 27: 455-465. DOI: 10.1105/tpc.109.072652

Elevitch, C.R., K.M. Wilkinson and J.B. Friday, 2006. *Acacia Koa* (Koa) and *Acacia Koaia* (Koai‘a). In: Species Profiles for Pacific Island Agroforestry, Elevitch, C.R. (Ed.), Permanent Agriculture Resources, Hōlualoa, HI, pp: 1-29.

Elkind, Y., R. Edwards, M. Mavandad, S.A. Hedrick and O. Ribak *et al*., 1990. Abnormal plant development and down-regulation of phenylpropanoid biosynthesis in transgenic tobacco containing a heterologous phenylalanine ammonia-lyase gene. Proc. Nat. Acad. Sci. USA, 87: 9057-9061. DOI: 10.1073/pnas.87.22.9057

Fredua-Agyeman, R., D. Adamski, R.J. Liao, C. Morden and D. Borthakur, 2008. Development and characterization of microsatellite markers for analysis of population differentiation in the tree legume *Acacia Koa* (*Fabaceae*: *Mimosoideae*) in the Hawaiian Islands. Genome, 51: 1001-1015. DOI: 10.1139/G08-087

Garg, R., R.K. Patel, A.K. Tyagi and M. Jain, 2011. *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res., 18: 53-63. DOI: 10.1093/dnares/dsq028

Gill, G.P., G.R. Brown and D.B. Neale, 2003. A sequence mutation in the cinnamyl alcohol dehydrogenase gene associated with altered lignification in loblolly pine. Plant Biotechnol. J., 1: 253-258. DOI: 10.1046/j.1467-7652.2003.00024.x

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin and D.A. Thompson *et al*., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol., 29: 644-652. DOI: 10.1038/nbt.1883

Gruenheit, N., O. Deusch, C. Esser, M. Becker and C. Voelckel *et al*., 2012. Cutoffs and k-mers: Implications from a transcriptome study in allopolyploid plants. BMC Genom., 13: 92-92. DOI: 10.1186/1471-2164-13-92

Hamilton, J.P. and C.R. Buell, 2012. Advances in plant genome sequencing. Plant J.: For Cell Molecular Biol., 70: 177-90. DOI: 10.1111/j.1365-313X.2012.04894.x

Hoffmann, L., S. Maury, F. Martz, P. Geoffroy and M. Legrand, 2003. Purification, cloning and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. J. Biol. Chem., 278: 95-103. DOI: 10.1074/jbc.M209362200

Hu, W.J., S.A. Harding, J. Lung, J.L. Popko and J. Ralph *et al*., 1999. Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nat. Biotechnol., 17: 808-812. DOI: 10.1038/11758

Jayashree, B., R. Punna, P. Prasad, K. Bantte and C.T. Hash *et al*., 2006. A database of simple sequence repeats from cereal and legume expressed sequence tags mined in silico: Survey and evaluation. Silico Boil., 6: 607-620.

Joshi, C.P. and V.L. Chiang, 1998. Conserved sequence motifs in plant S-adenosyl-L-methionine-dependent methyltransferases. Plant Molecular Biol., 37: 663-674. DOI: 10.1023/A:1006035210889

Jouanin, L., T. Goujon, V. de Nadaï, M.T. Martin and I. Mila *et al*., 2000. Lignification in transgenic poplars with extremely reduced caffeic acid O-methyltransferase activity. Plant Physiol., 123: 1363-1374. DOI: 10.1104/pp.123.4.1363

Kajita, S., S. Hishiyama, Y. Tomimura, Y. Katayama and S. Omori, 1997. Structural characterization of modified lignin in transgenic tobacco plants in which the activity of 4-coumarate: Coenzyme a ligase is depressed. Plant Physiol., 114: 871-879. DOI: 10.1104/pp.114.3.871

Kumpatla, S.P. and S. Mukhopadhyay, 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome, 48: 985-998. DOI: 10.1139/g05-060

Larsen, K., 2004. Molecular cloning and characterization of cDNAs encoding Cinnamoyl CoA Reductase (CCR) from barley (*Hordeum vulgare*) and potato (*Solanum tuberosum*). J. Plant Physiol., 161: 105-112. DOI: 10.1078/0176-1617-01074

Lee, D., K. Meyer, C. Chapple and C.J. Douglas, 1997. Antisense suppression of 4-coumarate: Coenzyme a ligase activity in *Arabidopsis* leads to altered lignin subunit composition. Plant Cell, 9: 1985-1998. DOI: 10.1105/tpc.9.11.1985

Li, X., W. Chen, Y. Zhao, Y. Xiang and H. Jiang *et al*., 2013. Downregulation of caffeoyl-CoA O-methyltransferase (CCoAOMT) by RNA interference leads to reduced lignin production in maize straw. Genetics Molecular Biol., 36: 540-546. DOI: 10.1590/S1415-47572013005000039

Lynch, D., A. Lidgett, R. McInnes, H. Huxley and E. Jones *et al*., 2002. Isolation and characterization of three cinnamyl alcohol dehydrogenase homologue cDNAs from perennial ryegrass (*Lolium perenne L*.). J. Plant Physiol., 159: 653-660. DOI: 10.1186/1471-2164-12-342

McKie, J.H., R. Jaouhari, K.T. Douglas, D. Goffner and C. Feuillet *et al*., 1993. A molecular model for cinnamyl alcohol dehydrogenase, a plant aromatic alcohol dehydrogenase involved in lignification. Biochim. Biophys. Acta., 1202: 61-69. DOI: 10.1016/0167-4838(93)90063-W

Meyermans, H., K. Morreel, C. Lapierre, B Pollet and A. De Bruyn *et al*., 2000. Modifications in lignin and accumulation of phenolic glucosides in poplar xylem upon down-regulation of caffeoyl-coenzyme A O-methyltransferase, an enzyme involved in lignin biosynthesis. J. Biol. Chem., 275: 36899-36909. DOI: 10.1074/jbc.M006915200

Nakasugi, K., R. Crowhurst, J. Bally and P. Waterhouse, 2014. Combining transcriptome assemblies from multiple *de novo* assemblers in the allo-tetraploid plant *Nicotiana Benthamiana.* PloS One, 9: e91776-e91776. DOI: 10.1371/journal.pone.0091776

Novaes, E., M. Kirst, V. Chiang, H. Winter-Sederoff and R. Sederoff, 2010. Lignin and biomass: A negative correlation for wood formation and lignin content in trees. Plant Physiol., 154: 555-561. DOI: 10.1104/pp.110.161281

Pincon, G., S. Maury, L. Hoffmann, P. Geoffroy and C. Lapierre *et al*., 2001. Repression of O-methyltransferase genes in transgenic tobacco affects lignin synthesis and plant growth. Phytochemistry, 57: 1167-1176. DOI: 10.1016/S0031-9422(01)00098-X

Plomion, C., G. Leprovost and A. Stokes, 2001. Wood formation in trees. Plant Physiol., 127: 1513-1523. DOI: 10.1104/pp.010816

Ralph, J., T. Akiyama, H.D. Coleman and S.D. Mansfield, 2012. Effects on lignin structure of coumarate 3-hydroxylase downregulation in poplar. Bioenergy Res., 5: 1009-1019. DOI: 10.1074/jbc.M511598200

Sahu, J., P. Sen, M.D. Choudhury, M. Barooah and M.K. Modi *et al*., 2012. Towards an efficient computational mining approach to identify EST-SSR markers. Bioinformation, 8: 550-560. DOI: 10.6026/97320630008201

Sakai, H.F., 1988. Avian response to mechanical clearing of a native rainforest in Hawaii. Condor, 90: 339-348.

Schuler, M.A., 1996. Plant cytochrome P450 monooxygenases. Critical Rev. Plant Sci., 15: 235-284. DOI: 10.1104/pp.108.130757

Sewalt, V., W. Ni, J.W. Blount, H.G. Jung and S.A. Masoud *et al*., 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 Physiol., 115: 41-50. DOI: 10.1104/pp.115.1.41

Song, Q.J., J.R. Shi, S. Singh, E.W. Fickus and J.M. Costa *et al*., 2005. Development and mapping of microsatellite (SSR) markers in wheat. Theoretical Applied Genet., 110: 550-560. DOI: 10.1007/s00122-004-1871-x

Wagner, A., L. Donaldson, H. Kim, L. Phillips and H. Flint *et al*., 2009. Suppression of 4-coumarate-CoA ligase in the coniferous gymnosperm *Pinus radiata*. Plant Physiol., 149: 370-383. DOI: 10.1104/pp.108.125765

Wagner, W.L., D.R. Herbst and S.H, Sohmer, 1990. Manual of the Flowering Plants of Hawaii. 1st Edn., University of Hawaii Press, Honolulu, HI, ISBN-10: 0824811526, pp: 1853.

Wanner, L.A., G. Li, D. Ware, I.E. Somssich and K.R. Davis, 1995. The phenylalanine ammonia-lyase gene family in *Arabidopsisthaliana*. Plant Molecular Biol., 27: 327-338. DOI: 10.1007/BF00020187

Wei, X., L. Wang, Y. Zhang, X. Qi and X. Wang *et al*., 2014. Development of Simple Sequence Repeat (SSR) markers of sesame (*Sesamum indicum*) from a genome survey. Molecules, 19: 5150-62. DOI: 10.3390/molecules19045150

Whitesell, C.D., 1990. Silvical Characteristics of *Acacia koa Gray*. In: Agriculture Handbook 654, Burns, H.R.M. and B.H. Honkala (Eds.), USDA Forest Service, Washington, D.C., pp: 17-28.

Wong, M.M.L., C.H. Cannon and R. Wickneswari, 2011. Identification of lignin genes and regulatory sequences involved in secondary cell wall formation in *Acacia auriculiformis* and *Acacia mangium* via *de novo* transcriptome sequencing. BMC Genom., 12: 342-342. DOI: 10.1186/1471-2164-12-342

Wu, R.L., D.L. Remington, J.J. MacKay, S.E. McKeand and D.M. O'Malley, 1999. Average effect of a mutation in lignin biosynthesis in loblolly pine. Theoretical Applied Genetics, 99: 705-710. DOI: 10.1007/s001220051287

Wu, S., Z. Zhu, L. Fu, B. Niu and W. Li, 2011. WebMGA: A customizable web server for fast metagenomic sequence analysis. BMC Genomics, 12: 444-444. DOI: 10.1186/1471-2164-12-444

Yanagida, J.F., J.B. Friday, P. Illukpitiya, R.J. Mamiit and Q Edwards, 2004. Economic value of Hawaii's forest industry in 2001. Economic Issues 7, College of Tropical Agriculture and Human Resources, University of Hawai'i at Manoa.

Yu, Q., S.E. McKeand, C.D. Nelson, B. Li and J.R. Sherrill *et al*., 2005. Differences in wood density and growth of fertilized and nonfertilized loblolly pine associated with a mutant gene, *cad-n1*. Can. J. Forest Res., 35: 1723-1730. DOI: 10.1139/x05-103

Zubieta, C., P. Kota, J. Ferrer, R.A. Dixon and J.P. Noel, 2002. Structural basis for the modulation of lignin monomer methylation by caffeic acid/5-hydroxyferulic acid 3/5-O-methyltransferase. Plant Cell, 14: 1265-1277. DOI: 10.1105/tpc.001412

Supplementary Material

Table S1. Number of unigenes categorized in the monolignol biosynthesis pathways

KO number	Definition	Number of unigenes
K10775	Phenylalanine ammonia-lyase	14
K00487	Cinnamate 4-hydroxylase	3
K01904	*p*-coumarate: CoA ligase	11
K09753	Cinnamoyl-CoA reductase	25
K00083	Cinnamyl-alcohol dehydrogenase	19
K09754	*p*-coumarate 3-hydroxylase	1
K13065	Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase	33
K13066	Caffeic acid 3-O-methyltransferase	1
K09755	Ferulate-5-hydroxylase	36
K00588	Caffeoyl-CoA O-methyltransferase	6

```
AtPAL1          SKATGPNGEALTAEEAFKLAGISSGFFDLQPKEGLALVNGTAVGSGMASMVLFETNVLSV 290
AtPAL2          SKATGPDGESLTAKEAFEKAGISTGFFDLQPKEGLALVNGTAVGSGMASMVLFEANVQAV 282
AkPAL1          SKSIGPDGRVLSPKEAFHLAGINGGFFELQPKEGLALVNGTAVGSALASIVLFETNILGV 280
AkPAL2          SKAVGPNGEALNPKEAFKLAGIESEFFELQPKEGLALVNGTAVGSGLASMVLFEANILAV 281
GmPAL1          SKAVGPSGEVLNAKEAFELASINSEFFELQPKEGLALVNGTAVGSGLASMVLFEANILAV 278
GmPAL2          SKAVGPSGEILNAKEAFELANIGAEFFELQPKEGLALVNGTAVGSGLASIVLFEANIIAV 282
AtPAL4          SKAVGPSGETLTASEAFKLAGVSS-FFELQPKEGLALVNGTAVGSGLASTVLFDANILAV 272
AtPAL3          SRSVGPSGEILTALEAFKLAGVSS-FFELRPKEGLALVNGTAVGSALASTVLYDANILVV 269
                *:: **.*. *.. ***. *.:   **:*:****************.:** **:::*:  *

AtPAL1          LAEILSAVFAEVMSGKPEFTDHLTHRLKHHPGQIEAAAIMEHILDGSSYMKLAQKLHEMD 350
AtPAL2          LAEVLSAIFAEVMSGKPEFTDHLTHRLKHHPGQIEAAAIMEHILDGSSYMKLAQKVHEMD 342
AkPAL1          LSEVMSAIFAEVMQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSSYMKAAQKLHEID 340
AkPAL2          LAEVISAIFAEVMQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSSYVKAAKKLHEMD 341
GmPAL1          LSEVLSAIFAEVMQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSSYMKAAKKLHEID 338
GmPAL2          LSEVISAIFAEVMQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILEGSSYVKAAKKLHEID 342
AtPAL4          LSEVMSAMFAEVMQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSSYVKEAQLLHEMD 332
AtPAL3          FSEVASAMFAEVMQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSSYVKEALHLHKID 329
                ::*: **:*****.***********:******************:****:* *  :*::*

AtPAL1          PLQKPKQDRYALRTSPQWLGPQIEVIRYATKSIEREINSVNDNPLIDVSRNKAIHGGNFQ 410
AtPAL2          PLQKPKQDRYALRTSPQWLGPQIEVIRQATKSIEREINSVNDNPLIDVSRNKAIHGGNFQ 402
AkPAL1          PLQKPKQDRYALRTSPQWLGPLIEVIRFSTKSIEREINSVNDNPLIDVSRNKALHGGNFQ 400
AkPAL2          PLQKPKQDRYALRTSPQWLGPLIEVIRFSTKSIEREINSVNDNPLIDVSRNKALHGGNFQ 401
GmPAL1          PLQKPKQDRYALRTSPQWLGPLIEVIRFSTKSIEREINSVNDNPLIDVSRNKALHGGNFQ 398
GmPAL2          PLQKPKQDRYALRTSPQWLGPLIEVIRFSTKSIEREINSVNDNPLIDVSRNKALHGGNFQ 402
AtPAL4          PLQKPKQDRYALRTSPQWLGPQIEVIRAATKMIEREINSVNDNPLIDVSRNKALHGGNFQ 392
AtPAL3          PLQKPKQDRYALRTSPQWLGPQIEVIRAATKMIEREINSVNDNPLIDVSRNKAIHGGNFQ 389
                ******************** ***** :** **********************:******


AtPAL1          GTPIGVSMDNTRLAIAAIGKLMFAQFSELVNDFYNNGLPSNLTASRNPSLDYGFKGAEIA 470
AtPAL2          GTPIGVSMDNTRLAIAAIGKLMFAQFSELVNDFYNNGLPSNLTASSNPSLDYGFKGAEIA 462
AkPAL1          GTPIGVSMDNTRLALASIGKLMFAQFSELVNDFYNNGLPSNLSAGRNPSLDYGFKGAEIA 460
AkPAL2          GTPIGVSMDNTRLAIASIGKLMFAQFSELVNDFYNNGLPSNLTASRNPSLDYGFKGAEIA 461
GmPAL1          GTPIGVSMDNTRLALASIGKLMFAQFSELVNDFYNNGLPSNLTASRNPSLDYGFKGAEIA 458
GmPAL2          GTPIGVSMDNTRLALASIGKLMFAQFSELVNDYYNNGLPSNLTASRNPSLDYGFKGAEIA 462
AtPAL4          GTPIGVAMDNSRLAIASIGKLMFAQFSELVNDFYNNGLPSNLSGGRNPSLDYGFKGAEIA 452
AtPAL3          GTPIGVAMDNTRLALASIGKLMFAQFTELVNDFYNNGLPSNLSGGRNPSLDYGLKGAEVA 449
                ******:***:***:*:*********:*****:*********:.. *******:****:*

AtPAL1          MASYCSELQYLANPVTSHVQSAEQHNQDVNSLGLISSRKTSEAVDILKLMSTTFLVAICQ 530
AtPAL2          MASYCSELQYLANPVTSHVQSAEQHNQDVNSLGLISSRKTSEAVDILKLMSTTFLVGICQ 522
AkPAL1          MASYCSEIQYLANPVTSHVQSAEQHNQDVNSLGLISSRKTNEAIEILKLMSSTYLIALCQ 520
AkPAL2          MASYCSELQYLANPVTTHVQSAEQHNQDVNSLGLISSRKTNEAIEILKLMSSTFLVALCQ 521
GmPAL1          MASYCSELQYLANPVTTHVQSAEQHNQDVNSLGLISSRKTNEAIEILKLMSSTFLIALCQ 518
GmPAL2          MASYCSELQYLANPVTSHVQSAEQHNQDVNSLGLISSRKTHEAIEILKLMSSTFLVALCQ 522
AtPAL4          MASYCSELQFLANPVTNHVQSAEQHNQDVNSLGLISSRKTAEAVDILKLMSTTYLVALCQ 512
AtPAL3          MASYCSELQFLANPVTNHVESASQHNQDVNSLGLISSRTTAEAVVILKLMSTTYLVALCQ 509
                *******:*:******.**:**.***************.* **: ******:*:*:.:**

AtPAL1          AVDLRHLEENLRQTVKNTVSQVAKKVLTTGVNGELHPSRFCEKDLLKVVDREQVYTYADD 590
AtPAL2          AVDLRHLEENLRQTVKNTVSQVAKKVLTTGINGELHPSRFCEKDLLKVVDREQVFTYVDD 582
AkPAL1          AIDLRHLEENLKSTVNSTVSQVAKRTLTTGVNGELHPSRFCEKDLLKVVDREHVFAYIDD 580
AkPAL2          AIDLRHLEENLKNTVKNTVSQVAKRTLTTGVNGELHPSRFCEKDLLKVVDREYVFAYADD 581
GmPAL1          AIDLRHLEENLKNSVKNTVSQVSKRILTTGVNGELHPSRFCEKDLLKVVDREYIFSYIDD 578
GmPAL2          AIDLRHLEENLKNTVKNVVSQVAKRTLTTGVNGELHPSRFCEKDLLKVVDREYTFAYIDD 582
AtPAL4          AVDLRHLEENLKKAVKSAVSQVAKRVLTVGANGELHPSRFTERDVLQVVDREYVFSYADD 572
AtPAL3          AFDLRHLEEILKKAVNEVVSHTAKSVLAI------EPFRKHD-DILGVVNREYVFSYVDD 562
                *.******* *:.:*:..**:.:*  *:       .* *  : *:* **:**  :** **

AtPAL1          PCSATYPLIQKLRQVIVDHALINGESEKNAVTSIFHKIGAFEEELKAVLPKEVEAARAAY 650
AtPAL2          PCSATYPLMQRLRQVIVDHALSNGETEKNAVTSIFQKIGAFEEELKAVLPKEVEAARAAY 642
AkPAL1          PTSATYPLMQKLRQVLVDHALENGDNEKNSSTSIFQKIAAFEEELKTLLPKEVERARTAY 640
AkPAL2          PCSALYPLMQKLRQVLVDHALANAENEKNTSTSIFQKIGTFEEELNNVLPKEVESARVAY 641
GmPAL1          PCSATYPLMQKLRQVLVDHALVNAECEKDVNSSIFQKIAIFEEELKNLLPKEVEGARAAY 638
GmPAL2          PCSGTYPLMQKLRQVLVDYALANGENEKNTSTSIFQKIATFEEELKTLLPKEVEGARVAY 642
AtPAL4          PCSLTYPLMQKLRHILVDHALADPEREANSATSVFHKIGAFEAELKLLLPKEVERVRVEY 632
AtPAL3          PSSLTNPLMQKLRHVLFDKALAEPEGETD---TVFRKIGAFEAELKFLLPKEVERVRTEY 619
                * *   **:*:**:::.* ** : : * :   ::*:**. ** **: :****** .*. *

AtPAL1          DNGTSAIPNRIKECRSYPLYRFVREELGTELLTGEKVTSPGEEFDKVFTAICEGKIIDPM 710
AtPAL2          GNGTAPIPNRIKECRSYPLYRFVREELGTKLLTGEKVVSPGEEFDKVFTAMCEGKLIDPL 702
AkPAL1          ENGNSSVPNKIKECRSYPLYKFVREDLGAGLLTGEKTRSPGEECDKVFTALCQGKIIDPL 700
AkPAL2          ETGTSEIPNRIKECRSYPLYKFVREELGTQLLTGERVISPGEECDKVFTALCQGKIIDPL 701
GmPAL1          ESGKAAIPNKIQECRSYPLYKFVREELGTGLLTGEKVRSPGEEFDKLFTAMCQGKIIDPL 698
GmPAL2          ENDQCAIPNKIKECRSYPLYKFVREELGTALLTGERVISPGEECDKVFTALCQGKIIDPL 702
AtPAL4          EEGTSAIANRIKECRSYPLYRFVRDELNTELLTGENVRSPGEEFDKVFLAISDGKLIDPL 692
AtPAL3          ENGTFNVANRIKKCRSYPLYRFVRNELETRLLTGEDVRSPGEDFDKVFRAISQGKLIDPL 679
                  .   :.*:*::*******:***::* : ***** . ****: **:* *:.:**:***:

AtPAL1          MECLNEWNGAPIPIC 725
AtPAL2          MDCLKEWNGAPIPIC 717
AkPAL1          LECLGEWNGAPLPIC 715
AkPAL2          LECVGEWNGAPLPIC 716
GmPAL1          MECLGEWNGAPLPIS 713
GmPAL2          LECLGEWNGAPLPIC 717
AtPAL4          LECLKEWNGAPVSIC 707
AtPAL3          FECLKEWNGAPISIC 694
                ::*: ******:.*.
```

2) Cinnamate 4-hydroxylase (C4H)

```
GmC4H          MDLLLLEKT------------LIGLFLAAVVAIAVSTLRG-------------------R 29
AkC4H1         MDLLLLEKT------------LIGLFVAVIVAIVVSKLRG-------------------R 29
AtC4H          MDLLLLEKS------------LIAVFVAVILATVISKLRG-------------------K 29
AkC4H2         MAPLLVHKSSFSLFSLFTLITTIALFSFIIVKSFSPSFSSPAAFVFPLTILLFLKHSPRH 60
               *  **:.*:              *.:*   ::    ..: .                   :

GmC4H          KFKLPPGPLPVPIFGNWLQVGDDLNHRNLTDLAKKFGDIFLLRMGQRNLVVVSSPELAKE 89
AkC4H1         RFKLPPGPLPVPIFGNWLQVGDDLNHRNLTDLAKKYGDIFLLRMGQRNLVVVSSPELAKD 89
AtC4H          KLKLPPGPIPIPIFGNWLQVGDDLNHRNLVDYAKKFGDLFLLRMGQRNLVVVSSPDLTKE 89
AkC4H2         SSKVPPGPLSVPIFGNWLQVGNDLNHRLLASMSQTYGPVFLLKLGSKNLVVVSDPELATQ 120
                 *:****:.:**********:***** *.. ::.:* :***::*.:******.*:*:.:

GmC4H          VLHTQGVEFGSRTRNVVFDIFTGKGQDMVFTVYGEHWRKMRRIMTVPFFTNKVVQQYRHG 149
AkC4H1         VLHTQGVEFGSRTRNVVFDIFTGKGQDMVFTVYGEHWRKMRRIMTVPFFTNKVVQQQREG 149
AtC4H          VLLTQGVEFGSRTRNVVFDIFTGKGQDMVFTVYGEHWRKMRRIMTVPFFTNKVVQQNREG 149
AkC4H2         VLHSQGVEFGSRPRNVVFDIFTGNGQDMVFTVYGDHWRKMRRVMTVPFFTNKVVQQYSVM 180
               ** :********.**********:**********:*******:**************

GmC4H          WESEAAAVVEDVKKNPDAAVSGTVIRRRLQLMMYNNMYRIMFDRRFESEEDPIFQRLRAL 209
AkC4H1         WENEVASVVEDVKKNPESATNGIVLRKRLQLMMYNNMYRIMFDRRFESEDDPLFQRLKAL 209
AtC4H          WEFEAASVVEDVKKNPDSATKGIVLRKRLQLMMYNNMFRIMFDRRFESEDDPLFLRLKAL 209
AkC4H2         WEQEMDLVVRDLKTNEAVRSRGIVIRKRLQLMLYNIMYRMMFDAKFESQDDPLFIEATRF 240
               ** *   **.*:*.*      * *:*:*****:** *:*:*** :***::**:* .   :

GmC4H          NGERSRLAQSFEYNYGDFIPILRPFLKGYLKICKEVKETRLKLFKDYFVDERKKLGSTKS 269
AkC4H1         NGERSRLAQSFEYNYGDFIPILRPFLRGYLKICKEVKETRLKLFKDYFVDERKKLGSTRS 269
AtC4H          NGERSRLAQSFEYNYGDFIPILRPFLRGYLKICQDVKDRRIALFKKYFVDERKQIASSKP 269
AkC4H2         NSERSRLAQSFDYNYGDFIPLLRPFLRGYLNKCRDLQTRRLAFFNDHYVQQRRKIMAANG 300
               *.*********:********:*****:***: *::::  *: :*:.::*::*::: ::.

GmC4H          TNNNNELKCAIDHILDAQRKGEINEDNVLYIVENINVAAIETTLWSIEWGIAELVNHPEI 329
AkC4H1         SSNG-ELKCAIDHILDAQKKGEINEDNVLYIVENINVAAIETTLWSIEWGVAELVNHPEI 328
AtC4H          TGSE-GLKCAIDHILEAEQKGEINEDNVLYIVENINVAAIETTLWSIEWGIAELVNHPEI 328
AkC4H2         EKHK--ISCAIDHIIDAEMKGEISEENVLYIVENINVAAIETTLWSMEWAIAELVRNPRV 358
                     :.******::*: ****.*:**********************:**.:****.:*.:

GmC4H          QQKLRDEIDRVLGAGHQVTEPDIQKLPYLQAVVKETLRLRMAIPLLVPHMNLHDAKLGGY 389
AkC4H1         QKKLRDEIDTVLGPGHQVTEPDTHKLPYLQAVVKETLRLRMAIPLLVPHMNLNDAKLGGY 388
AtC4H          QSKLRNELDTVLGPGVQVTEPDLHKLPYLQAVVKETLRLRMAIPLLVPHMNLHDAKLAGY 388
AkC4H2         QSKIREEISRVL-KGEAVTESNLQELPYLQAVVKETLRLHSPIPLLVPHMNLEEAELGGY 417
               *.*:*:*:. **  *  ***.: ::************: .**********.:*:*.**

GmC4H          DIPAESKILVNAWWLANNPAHWKKPEEFRPERFFEEESLVEANGN---DFRYLPFGVGRR 446
AkC4H1         DIPAESKILVNAWWLANNPAHWKNPEQFRPERFLEEESKVEANGN---DFRYLPFGVGRR 445
AtC4H          DIPAESKILVNAWWLANNPNSWKKPEEFRPERFFEEESHVEANGN---DFRYVPFGVGRR 445
AkC4H2         KIPKESKVVVNAWWLANNPAWWEKAEEFRPERFMEEESGTDAVAGGKVDFRYLPFGVGRR 477
               .** ***::********** *:::.*:******:**** .:* ..   ****:*******

GmC4H          SCPGIILALPILGITLGRLVQNFELLPPPGQSQIDTSEKGGQFSLHILKHSTIVAKPRSF 506
AkC4H1         SCPGIILALPILGVTLGRLVQNFELLPPPGQSKLDTAEKGGQFSLHILKHSTIVAKPRSF 505
AtC4H          SCPGIILALPILGITIGRMVQNFELLPPPGQSKVDTSEKGGQFSLHILNHSIIVMKPRNC 505
AkC4H2         SCPGIILALPILGLVIAKLVTNFEMEAPKG-TQIDVSEKGGQFSLHIANHSTVVFHPINA 536
               *************:.:.::* ***: .* * :::*.:********** :** :* :* .

GmC4H          -
AkC4H1         -
AtC4H          -
AkC4H2         S 537
```

3)*p*-coumarate:CoA ligase (4CL)

```
Ak4CL1          MISVESQQKPDLSDPSPKPASNEPNSANTAQTTHIFKSKLPDIPISNHLPLHTYCFENLA 60
Gm4CL2          MITLA---------PSLDTPKTDQNQVSDPQTSHVFKSKLPDIPISNHLPLHSYCFQNLS 51
At4CL3          MITAALHEP-QIHKPTDTSVVSDDVLPHSPPTPRIFRSKLPDIDIPNHLPLHTYCFEKLS 59
At4CL1          ---------MAPQEQAVSQVMEKQSNNNNS--DVIFRSKLPDIYIPNHLSLHDYIFQNIS 49
At4CL2          ---------MTTQDVIVNDQNDQKQCSN----DVIFRSRLPDIYIPNHLPLHDYIFENIS 47
At4CL5          -------MVLQQQTHFLTKKIDQEDEEEEPSHDFIFRSKLPDIFIPNHLPLTDYVFQRFS 53
Ak4CL3          --------------------MAATETTVNPQTDFIFRSKLPDIYIPKHLPLHSYCFENLA 40
Gm4CL1          --------------------MADDGS----RRELIFRSKLPDIYIPKHMPLHSYCFENLR 36
Ak4CL4          ---------------------METPPT------EFIFRSKLPDISIPTHLPLHSYVFENLS 34
Ak4CL2          --------------------MEQIDERSGFCKSNSIFYSKRKPLPLTQNYSLDVTTFISSR 41
                                                 :* *:   : :. : .*    *

Ak4CL1          ----QFADRPCLIVGSTGKIYSYAEAHLQSRKIAAGFSKLGVKKGDVIMILLQNSAEFAL 116
Gm4CL2          ----QFAHRPCLIVGPASKTFTYADTHLISSKIAAGLSNLGILKGDVVMILLQNSADFVF 107
At4CL3          ----SVSDKPCLIVGSTGKSYTYGETHLICRRVASGLYKLGIRKGDVIMILLQNSAEFVF 115
At4CL1          ----EFATKPCLINGPTGHVYTYSDVHVISRQIAANFHKLGVNQNDVVMLLLPNCPEFVL 105
At4CL2          ----EFAAKPCLINGPTGEVYTYADVHVTSRKLAAGLHNLGVKQHDVVMILLPNSPEVVL 103
At4CL5          GDGDGDSSTTCIIDGATGRILTYADVQTNMRRIAAGIHRLGIRHGDVVMLLLPNSPEFAL 113
Ak4CL3          ----EFGSRPCLINAPTGDIYTYYDVELTARRVASGLNKFGVGQGDVIMVLLPNSPEFVF 96
Gm4CL1          ----ECGSRPCLINAPTGDVYSYHEVDSTARKVARGLKKEGVEQGQVIMILLPNCPEFVF 92
Ak4CL4          ----QVKDRPCLINGDTGETFTYANVELTARRVAAGLAKIGIRQGDVIMLVLRNCPEFAL 90
Ak4CL2          ----AHHGKLAFIDAATGRHFTFPQLWRAVDAVATSLSDLGVRKGHVILLLSPNSIYFPV 97
                      .:* . :.   :: :       :* .:   *: : .*::::  *.  . .

Ak4CL1          SFFAASMIGAVVTTANPFYTADEIFKQFHAAKAKLIITQSAYVGKLRDHEAGKDIGEDQL 176
Gm4CL2          SFLAISMIGAVATTASPFYTAPEIFKQFTVSKEKLVITQAMYVDKLRNHDGAK-LGED-F 165
At4CL3          SFMGASMIGAVSTTANPFYTSQELYKQLKSSGAKLIITHSQYVDKLKN------LGEN-L 168
At4CL1          SFLAASFRGATATAANPFFTPAEIAKQAKASNTKLIITEARYVDKIKPLQ-----NDDGV 160
At4CL2          TFLAASFIGAITTSANPFFTPAEISKQAKASAAKLIVTQSRYVDKIKNLQ-----ND-GV 157
At4CL5          SFLAVAYLGAVSTTANPFYTQPEIAKQAKASAAKMIITKKCLVDKLTNLK-----ND-GV 167
Ak4CL3          SFLGASLRGAMTTAANPFFTAAEVAKQAKASKAKLIVTQSNFFEKVKDI------DVKLI 150
Gm4CL1          SFLGASHRGAMATAANPFFTPAEIAKQAHASNAKLLITQASYYDKVKDLR-----DIKLV 147
Ak4CL4          AFLGASFAGAAVTTANPLFTPAELTKQAAASKSKLIITQTAFVEKIKDFA-----HTHGI 145
Ak4CL2          VCLAVMSLGAIITTTNPLNTTREIAKQISDSKPSLAFTIRPLLPKITAAS------NLPI 151
                  :.    **  *::.*: *  *: **   :  .: .*      *:            .
```

```
Ak4CL1          KVITVDDP---------PEKCLHFSVVSEADENE---VPEVEIDPDDPVALPFSSGTTGL 224
Gm4CL2          KVVTVDDP---------PENCLHFSVLSEANESD---VPEVEIHPDDAVAMPFSSGTTGL 213
At4CL3          TLITTDEPT--------PENCLPFSTLITDDETNP-FQETVDIGGDDAAALPFSSGTTGL 219
At4CL1          VIVCIDDNES-----VPIPEGCLRFTELTQSTTEASEVIDSVEISPDDVVALPYSSGTTGL 216
At4CL2          LIVTTDSD-------AIPENCLRFSELTQSEEPRVDSIP-EKISPEDVVALPFSSGTTGL 209
At4CL5          LIVCLDDDGDNGVVSSSDDGCVSFTELTQADET---ELLKPKISPEDTVAMPYSSGTTGL 224
Ak4CL3          FVDSPP------------DGHSHFSELSQAD-E-KDVP-EVKIKPDDAVALPYSSGTTGL 195
Gm4CL1          FVDSCPPHTE-------EKQHLHFSHLCEDNGD-ADVDVDVDIKPDDVVALPYSSGTTGL 199
Ak4CL4          SLMCIDSTFP------EKEDISHFSLLTQSDEA---DMPAIKISPNDIVALPYSSGTSGV 196
Ak4CL2          VLMEEDSNTP----ITEPKIVTTLGKMMKTEPRAS--QVRDRVYQDDTATLLYSSGTTGP 205
                 :                .    :  :                :  :* .:: :****:*

Ak4CL1          PKGVILTHKSLATSVAQQVDGENPNLYLKPEDVLLCVLPLFHIFSLNSVLLCALRAGSAV 284
Gm4CL2          PKGVILTHKSLTTSVAQKIDGENPNLYLTTEDVLLCVLPLFHIFSLNSVLLCALRAGSAV 273
At4CL3          PKGVVLTHKSLITSVAQQVDGDNPNLYLKSNDVILCVLPLFHIYSLNSVLLNSLRSGATV 279
At4CL1          PKGVMLTHKGLVTSVAQQVDGENPNLYFHSDDVILCVLPMFHIYALNSIMLCGLRVGAAI 276
At4CL2          PKGVMLTHKGLVTSVAQQVDGENPNLYFNRDDVILCVLPMFHIYALNSIMLCSLRVGATI 269
At4CL5          PKGVMITHKGLVTSIAQKVDGENPNLNFTANDVILCFLPMFHIYALDALMLSAMRTGAAL 284
Ak4CL3          PKGVMLTHKGLVTSIAQQVDGENPNLYFHHEDVILCVLPLFHIYSLNSVLLCGLRAKAAI 255
Gm4CL1          PKGVMLSHKGLVTSIAQQVDGDNPNLYYHCHDTILCVLPLFHIYSLNSVLLCGLRAKATI 259
Ak4CL4          PKGVMLTHKNLVTTVAQLVDGENPNQYTSSDDVHICVLPMFHIYALNSILLCSIRAGAAI 256
Ak4CL2          SKGVVSSHRNLMAMVQIVLG----RFNIDKEETFICTVPMFHIYGLVAFATGLLASGSTI 261
                .***: :*:.* : :   :.    .     .:. :* :*:***:.* :.    :   :::

Ak4CL1          LLMHKFEIGALLELTQKHKVSVAMVVPPLVLALAK--NPMVADYDLSSIRLVLSGAAPLG 342
Gm4CL2          LLMQKFEIGTLLELIQRHRVSVAMVVPPLVLALAK--NPMVADFDLSSIRLVLSGAAPLG 331
At4CL3          LLMHKFEIGALLDLIQRHRVTIAALVPPLVIALAK--NPTVNSYDLSSVRFVLSGAAPLG 337
At4CL1          LIMPKFEINLLLELIQRCKVTVAPMVPPIVLAIAK--SSETEKYDLSSIRVVKSGAAPLG 334
At4CL2          LIMPKFEITLLLEQIQRCKVTVAMVVPPIVLAIAK--SPETEKYDLSSVRMVKSGAAPLG 327
At4CL5          LIVPRFELNLVMELIQRYKVTVVPVAPPVVLAFIK--SPETERYDLSSVRIMLSGAATLK 342
Ak4CL3          MLMPKFEINALLGLIQKYKVSIAPVVPPIVLAISK--SPDIDKYDLSSVRVLKSGGAPLS 313
Gm4CL1          LLMPKFDINSLLALIHKHKVTIAPVVPPIVLAISK--SPDLHKYDLSSIRVLKSGGAPLG 317
Ak4CL4          LTMSKYDITTLLKMIETYKVTMASFVPPILLSIVK--SEEVDRYDLSSIRVIVTGAAPVS 314
Ak4CL2          VVLSKFDMDNMLSAIQKYRATYLPLVPPILVAMMNGADAIKAKYDLSSLHSALSGGAPLS 321
                : : :::: ::   .  :.:   ..**::::: :  .     :****::   :*.*.:

Ak4CL1          KELEDALRSRVPQAVLGQGYGMTEAGPVLSMCLGFAREAFTTKSGSCGTVVRNAELKVLH 402
Gm4CL2          KELEEALRNRMPQAVLGQGYGMTEAGPVLSMCLGFAKQPFQTKSGSCGTVVRNAELKVVD 391
At4CL3          KELQDSLRRRLPQAILGQGYGMTEAGPVLSMSLGFAKEPIPTKSGSCGTVVRNAELKVVH 397
At4CL1          KELEDAVNAKFPNAKLGQGYGMTEAGPVLAMSLGFAKEPFPVKSGACGTVVRNAEMKIVD 394
At4CL2          KELEDAISAKFPNAKLGQGYGMTEAGPVLAMSLGFAKEPFPVKSGACGTVVRNAEMKILD 387
At4CL5          KELEDAVRLKFPNAIFGQGYGMTESG-TVAKSLAFAKNPFKTKSGACGTVIRNAEMKVVD 401
Ak4CL3          KELEDSVRAKFPKARLGQGYGMTEAGPVLTMSLAFAKEPMEVKAGACGTVVRNAEMKIVD 373
Gm4CL1          KELEDTLRAKFPNAKLGQGYGMTEAGPVLTMSLAFAKEPIDVKPGACGTVVRNAEMKIVD 377
Ak4CL4          VELVRALKAKLPQAILGQGYGMTEGG-PLAISLSFAKEPEEMKSGACGTVIRNAEMKIVD 373
Ak4CL2          KEVIEGFVDKFPNVTILQGYGLTESTGVGASTDSLEE---SRRYGTAGLLSPATEAKVVN 378
                 *:   .  :.*:. : ****:**.    :   .: .     : *:.* :   :* *::.

Ak4CL1          PETGLSLGYNQPGEICIRGHQIMKGYLNNEEATATTIDAEGWLHTGDIGYVDDDDELFIV 462
Gm4CL2          PETGRSLGYNQPGEICIRGQQIMKGYLNDEAATASTIDSEGWLHTGDVGYVDDDDEIFIV 451
At4CL3          LETRLSLGYNQPGEICIRGQQIMKEYLNDPEATSATIDEEGWLHTGDIGYVDEDDEIFIV 457
At4CL1          PDTGDSLSRNQPGEICIRGHQIMKGYLNNPAATAETIDKDGWLHTGDIGLIDDDDELFIV 454
At4CL2          PDTGDSLPRNKPGEICIRGNQIMKGYLNDPLATASTIDKDGWLHTGDVGFIDDDDELFIV 447
At4CL5          TETGISLPRNKSGEICVRGHQLMKGYLNDPEATARTIDKDGWLHTGDIGFVDDDDEIFIV 461
Ak4CL3          PETSLSLPRGHPGEICIRGDQIMKGYLNDPEATKNTIDEEGWLHTGDIGYIDEDDELFIV 433
Gm4CL1          PETGHSLPRNQSGEICIRGDQIMKGYLNDGEATERTIDKDGWLHTGDIGYIDDDDELFIV 437
Ak4CL4          IDTGASLPRNRAGEICIRGDQVMKGYLNDPEATKTTIDEEGWLHTGDIGYVDDEAQVFVV 433
Ak4CL2          PETGEALPPNRTGELWLRGPTIMKGYFSNEEATTSTLDSEGWLRTGDICYIDGDGYIFIV 438
                 :*  :*  .:.**: :**  :** *:.:  **  *:* :***:***:  :* :  :*:*

Ak4CL1          DRVKELIKFKGFQVPPAELEGLLVSHPSIADAAVVPQKDEAAGEVPVAFVVRSNG-FDLT 521
Gm4CL2          DRVKELIKYKGFQVPPAELEGLLVSHPSIADAAVVPQKDVAAGEVPVAFVVRSNG-FDLT 510
At4CL3          DRLKEVIKFKGFQVPPAELESLLINHHSIADAAVVPQNDEVAGEVPVAFVVRSNG-NDIT 516
At4CL1          DRLKELIKYKGFQVAPAELEALLIGHPDITDVAVVAMKEEAAGEVPVAFVVKSKD-SELS 513
At4CL2          DRLKELIKYKGFQVAPAELESLLIGHPEINDVAVVAMKEEDAGEVPVAFVVRSKD-SNIS 506
At4CL5          DRLKELIKFKGYQVAPAELEALLISHPSIDDAAVVAMKDEVADEVPVAFVARSQG-SQLT 520
Ak4CL3          DRLKELIKYKGFQVAPAELEALLLSHPHISDAAVVPMKDEAAGEVPVAFVVRSNGHTQTT 493
Gm4CL1          DRLKELIKYKGFQVAPAELEALLLTHPKISDAAVVPMKDEAAGEVPVAFVVISNGYTDTT 497
Ak4CL4          DRLKEIIKYKGFQVAPAELEALLISHPLISDAAVVPMTDEGAGELPVAFVVRSNG-SKIS 492
Ak4CL2          DRLKELIKYKGYQVPPAELEALLLTHPAISDAAVIPFPDKEVGQYPMAYVVRKSG-SHAS 497
                **:**:**:**:**.*****.**: *  * *.**:.  :  ..: *:*:*. ...  . :

Ak4CL1          EEAVKEYIAKRVVFYKRLHRVYFVHAIPKSPSGKILRKDLRAKLESNNIQTA 573
Gm4CL2          EEAVKEFIAKQVVFYKRLHKVYFVHAIPKSPSGKILRKDLRAKLETAATQTP 562
At4CL3          EEDVKEYVAKQVVFYKRLHKVFFVASIPKSPSGKILRKDLKAKLC------- 561
At4CL1          EDDVKQFVSKQVVFYKRINKVFFTESIPKAPSGKILRKDLRAKLANGL---- 561
At4CL2          EDEIKQFVSKQVVFYKRINKVFFTDSIPKAPSGKILRKDLRARLANGLMN-- 556
At4CL5          EDDVKSYVNKQVVHYKRIKMVFFIEVIPKAVSGKILRKDLRAKLETMCSK-- 570
Ak4CL3          EDDIKQFVSKQVVFYKRISRVFFIDAIPKSPSGKILRKDLRAKLAADAANH- 544
Gm4CL1          EDEIKQFISKQVVFYKRINRVFFIDAIPKSPSGKILRKDLRAKIAASVPK-- 547
Ak4CL4          EDDIKHFVSQQVVYYKRIHKVVFTDTIPKAASGKILRKDLKARLASDLGN-- 542
Ak4CL2          EKEIMDFVAAQVAPYKRIRRVTFISSIPKNPSGKILRKDLIKLATSKL---- 545
                *. :  ::  :*. ***:  * *   ***  *********
```

4) Cinnamoyl CoA Reductase (CCR)

```
AkCCR1          MPAAASPSGSGQTICVTGAGGFIASWIVKLLLERGYTVRGTVRNPDDSKNAHLRELEGAQ 60
GmCCR           MPTDTS-SVSGEIVCVTGAGGFIASWLVKLLLEKGYTVRGTVRNPDDPKNGHLKELEGGK 59
AtCCR1          MPVDVA-SPAGKTVCVTGAGGYIASWIVKILLERGYTVKGTVRNPDDPKNTHLRELEGGK 59
AtCCR2          MLVD------GKLVCVTGAGGYIASWIVKLLLERGYTVRGTVRNPTDPKNNHLRELQGAK 54
AkCCR3          --------MAATKVCVTGAGGFIASWLVKLLLSKSYIVHGTVRKPGDQKYDHLLKFEGAS 52
                         .  :*******:****:**:**.:.* *:****:* * *  ** :::*..

AkCCR1          ERLTLHKVDLLDLDSVKSVVNGCDGVIHTASPVTD----NPE-EMVEPAVNGTKNVIIAS 115
GmCCR           ERLTLHKVDLFDIASIKAALHGCHGVFHTASPVTD----NPE-EMVEPAVKGTKNVIIAA 114
AtCCR1          ERLILCKADLQDYEALKAAIDGCDGVFHTASPVTD----DPE-QMVEPAVNGAKFVINAA 114
AtCCR2          ERLTLHSADLLDYEALCATIDGCDGVFHTASPMTD----DPE-TMLEPAVNGAKFVIDAA 109
AkCCR3          ENLKLFKAELLDYESISSAIAGCRAVFHVACPVPSTVVSNPEVEMIEPSVTGTTNVLKAC 112
                *.* * ..:* *  :: :.: ** .*:*.*.*:..    :**  *:**:*.*:. *: *.
```

```
AkCCR1          AEAKVRRVVFTSSIGAVYMDPNRNIDEVVDESCWSNLEYCKNTKNWYCYGKAVAEEAAWD 175
GmCCR           AEAKVRRVVFTSSIGTVYMDPNTSRDALVDESFWSDLEYCKNTKNWYCYGKTVAEQAAWD 174
AtCCR1          AEAKVKRVVITSSIGAVYMDPNRDPEAVVDESCWSDLDFCKNTKNWYCYGKMVAEQAAWE 174
AtCCR2          AKAKVKRVVFTSSIGAVYMNPNRDTQAIVDENCWSDLDFCKNTKNWYCYGKMLAEQSAWE 169
AkCCR3          LEAKVERVVFVSSQGAVSVNPNLPKDKLVDESCWSDKEYCRRTKNWYSYSKTGAELQALE 172
                 :***.***:.** *:* ::**   : :***. **: ::*:.*****.*.*  **  * :

AkCCR1          EAKARGVDLVVVNPVLVLGPLLQS-TINASTIHVLKYLTGSAKTYANATQAYVHVKDVAL 234
GmCCR           VAKERGVDLVVVNPVLVIGPLLQP-TINASTIHILKYLTGSAKTYVNATQAYVHVRDVAL 233
AtCCR1          TAKEKGVDLVVLNPVLVLGPPLQP-TINASLYHVLKYLTGSAKTYANLTQAYVDVRDVAL 233
AtCCR2          TAKAKGVDLVVLNPVLVLGPPLQS-AINASLVHILKYLTGSAKTYANLTQVYVDVRDVAL 228
AkCCR3          FARNNGLSVVTICPTVVWGPILQASTVNASSLILLKLFKG-LESLENKHRWIVDVRDVAD 231
                 *: .*:.:*.: *.:* ** **. ::***   :** :.*  ::  *  :  *.*:***

AkCCR1          AHVLVYETPSASGRYLCSERSLHRGELVEILAKHFPEYPTPTKCSDEKNPRAKPHTFSNK 294
GmCCR           AHILVYETPSASGRFICAESSLHRGELVEILAKFFPEYPIPTKCSDEKNPRVKPYIFSNQ 293
AtCCR1          AHVLVYEAPSASGRYLLAESARHRGEVVEILAKLFPEYPLPTKCKDEKNPRAKPYKFTNQ 293
AtCCR2          GHVLVYEAPSASGRYILAETALHRGEVVEILAKFFPEYPLPTKCSDEKNPRAKPYKFTTQ 288
AkCCR3          AVLMAYEEPEAEGRYLCSAHAIMTDDLVEKLRRIYPNYSYPKNFVEVDD----CIKLSSE 287
                . ::.** *.*.**:: :  :   .::** * : :*:*. *.:  : .:       ::.:

AkCCR1          KLKDLGLEFTPVDQCLYDTVKSLQDKGHLPLP-------TKQAEESVQIKS 338
GmCCR           KLKDLGLEFTPVKQCLYDTVKNLQENGHLPVP-------PKQKDSY----- 332
AtCCR1          KIKDLGLEFTSTKQSLYDTVKSLQEKGHLAPPPPPPSASQESVENGIKIGS 344
AtCCR2          KIKDLGLEFKPIKQSLYESVKSLQEKGHLPLP-------QDSNQNEVIIES 332
AkCCR3          KLQRLGWRYRPFEETLVDSVESYKAARLIESIN------------------ 320
                *:: ** .: . .: * ::*:. :    :
```

5)Cinnamoyl alcohol dehydrogenase (CAD)

```
AtCAD4          -MGSVEAG---EKKALGWAARDPSGVLSPYSYTLRSTGADDVYIKVICCGICHTDIHQIK 56
AtCAD5          -MGIMEA----ERKTTGWAARDPSGILSPYTYTLRETGPEDVNIRIICCGICHTDLHQTK 55
AkCAD2          -MGSVEG----ERTTVGWAARDPSGILSPYTYNLRNTGPDDVYIKVHYCGVCHSDLHQIK 55
ATCAD2          -MVDQ-------NKAFGWAANDESGVLSPFHFSRRENGENDVTVKILFCGVCHSDLHTIK 52
AtCAD3          -MVDQ-------NRAFGWAANDESGVLSPFHFSRRENGENDVTVKILFCGVCHSDLHTIK 52
AtCAD9          -MAKSP-ETEHPNKVFGWGARDKSGVLSPFHFSRRDNGENDVTVKILFCGVCHTDLHTIK 58
AtCAD6          -MERLSGEKEQSVEAFGWAARDSSGHLSPFVFSRRKTGEEEVRVKVLYCGICHSDLHCLK 59
AkCAD3          -MSSKDAN----TDCLGWAARDSSGVLSPYQFSRRALGDNDVFIKIAHCGVCYADVIWSR 55
GmCAD           -MSSKGVG----EDCLGWAARDASGVLSPYKFSRRTLGNEDVHIKITHCGVCFADVVWTR 55
AtCAD1          MSSSESVE----NECMCWAARDPSGLLSPHTITRRSVTTDDVSLTITHCGVCYADVIWSR 56
                             *.*.* ** ***.  . *    ::* : :  **:*.:*:   :

AtCAD4          NDLGMSNYPMVPGHEVVGEVLEVGSDVSKFTVGDVVGVGVVVGCCGSCKPCSSELEQYCN 116
AtCAD5          NDLGMSNYPMVPGHEVVGEVVEVGSDVSKFTVGDIVGVGCLVGCCGGCSPCERDLEQYCP 115
AkCAD2          NDLGMSNYPMVPGHEVVGEVIEVGSNVTKFEVGEVVGAGLIVGSCRNCRACKSDIEQYCG 115
ATCAD2          NHWGFSRYPIIPGHEIVGIATKVGKNVTKFKEGDRVGVGVIIGSCQSCESCNQDLENYCP 112
AtCAD3          NHWGFSRYPIIPGHEIVGIATKVGKNVTKFKEGDRVGVGVIIGSCQSCESCNQDLENYCP 112
AtCAD9          NDWGYSYYPVVPGHEIVGIATKVGKNVTKFKEGDRVGVGVISGSCQSCESCDQDLENYCP 118
AtCAD6          NEWHSSIYPLVPGHEIIGEVSEIGNKVSKFNLGDKVGVGCIVDSCRTCESCREDQENYCT 119
AkCAD3          NKHGDSKYPLVPGHEIAGIVEKIGSNVHHFNVGDHVGVGTYVNSCRDCQYCNEGIEVYCM 115
GmCAD           NKHGDSKYPVVPGHEIAGIVTKVGANVHHFKVGDHVGVGTYINSCRDCEYCNDGQEVHCT 115
AtCAD1          NQHGDSKYPLVPGHEIAGIVTKVGPNVQRFKVGDHVGVGTYVNSCRECEYCNEGQEVNCA 116
                *.   * **::*****: * . ::* .* :*  *: **.*   ..*  *  *    *  *

AtCAD4          KR-IWSYNDVYTDGK-PTQGGFADTMIVNQKFVVKIPEGMAVEQAAPLLCAGVTVYSPLS 174
AtCAD5          KK-IWSYNDVYINGQ-PTQGGFAKATVVHQKFVVKIPEGMAVEQAAPLLCAGVTVYSPLS 173
AkCAD2          KK-IWNYNDVYTDGK-PTQGGFAETMIVDQHFVVKIPEGMSPEQVAPLLCAGVTVYSPLS 173
ATCAD2          KV-VFTYNSRSSDGTSRNQGGYSDVIVVDHRFVLSIPDGLPSDSGAPLLCAGITVYSPMK 171
AtCAD3          KV-VFTYNSRSSDGT-RNQGGYSDVIVVDHRFVLSIPDGLPSDSGAPLLCAGITVYSPMK 170
AtCAD9          QM-SFTYNAIGSDGT-KNYGGYSENIVVDQRFVLRFPENLPSDSGAPLLCAGITVYSPMK 176
AtCAD6          KA-IATYNGVHHDGT-INYGGYSDHIVVDERYAVKIPHTLPLVSAAPLLCAGISMYSPMK 177
AkCAD3          KGSVLTFNGVDYDGT-ITKGGYSSNIVVHERYCFLIPKGYPLASAAPLLCAGITVYSPMM 174
GmCAD           KGSVYTFNGVDFDGT-ITKGGYSSYIVVHERYCFMIPKSYPLASAAPLLCAGITVYSPMV 174
AtCAD1          KG-VFTFNGIDHDGS-VTKGGYSSHIVVHERYCYKIPVDYPLESAAPLLCAGITVYAPMM 174
                :    .:*    :*   . **::.  :*..::   :*   .  . *******:::*:*:

AtCAD4          HFGLM-ASGLKGGILGLGGVGHMGVKIAKAMGHHVTVISSSDKKKEEAIEHLGADDYVVS 233
AtCAD5          HFGLK-QPGLRGGILGLGGVGHMGVKIAKAMGHHVTVISSSNKKREEALQDLGADDYVIG 232
AkCAD2          HFGLR-KSGLRGGILGLGGVGHMGVKIAKAMGHHVTVISSSEKKKKEALEHLGADDYVVS 232
ATCAD2          YYGMTKESGKRLGVNGLGGLGHIAVKIGKAFGLRVTVISRSSEKEREAIDRLGADSFLVT 231
AtCAD3          YYGMTKESGKRLGVNGLGGLGHIAVKIGKAFGLRVTVISRSSEKEREAIDRLGADSFLVT 230
AtCAD9          YYGMT-EAGKHLGVAGLGGLGHVAVKIGKAFGLKVTVISSSSTKAEEAINHLGADSFLVT 235
AtCAD6          YFGLT-GPDKHVGIVGLGGLGHIGVRFAKAFGTKVTVVSSTTGKSKDALDTLGADGFLVS 236
AkCAD3          RHKMN-QPGKSLGVIGLGGLGHMAVKFGKAFGLNVTIFSTSASKKEEALTLLGADNFVVS 233
GmCAD           RHKMN-QPGKSLGVIGLGGLGHMAVKFGKAFGLSVTVFSTSISKKEEALSLLGADKFVVS 233
AtCAD1          RHNMN-QPGKSLGVIGLGGLGHMAVKFGKAFGLSVTVFSTSISKKEEALNLLGAENFVIS 233
                 . :   ..   *: ****:**:.*:::.**:*  **:.* :  * .:*:  ***: :::

AtCAD4          SDPAEMQRLADSLDYIIDTVPVFHPLDPYLACLKLDGKLILMGVINTPLQFVTPLVILGR 293
AtCAD5          SDQAKMSELADSLDYVIDTVPVHHALEPYLSLLKLDGKLILMGVINNPLQFLTPLLMLGR 292
AkCAD2          SDETQMQKIADSLDYIIDTVPVGHPLEPYLSLLKVDGKLILMGVINTPLQFVSPMVMLGR 292
ATCAD2          TDSQKMKEAVGTMDFIIDTVSAEHALLPLFSLLKVNGKLVALGLPEKPLDLPIFSLVLGR 291
AtCAD3          TDSQKMKEAVGTMDFIIDTVSAEHALLPLFSLLKVSGKLVALGLLEKPLDLPIFPLVLGR 290
AtCAD9          TDPQKMKAAIGTMDYIIDTISAVHALYPLLGLLKVNGKLIALGLPEKPLELPMFPLVLGR 295
AtCAD6          TDEDQMKAAMGTMDGIIDTVSASHSISPLIGLLKSNGKLVLLGATEKPFDISAFSLILGR 296
AkCAD3          SDQDQMMALARSFDFIVDTASGDHPFDPYMSLLKTSGVLTLVGFP-SEVKFSPASLNLGM 292
GmCAD           SNQEEMTALAKSLDFIIDTASGDHSFDPYMSLLKTYGVFVLVGFP-SQVKFIPASLNIGS 292
AtCAD1          SDHDQMKALEKSLDFLVDTASGDHAFDPYMSLLKIAGTYVLVGFP-SEIKISPANLNLGM 292
                ::  :*     ::* ::** .  *.: * :. **  *    :*   . ..:    : :*

AtCAD4          KVISGSFIGSIKETEEVLAFCKEKGLTSTIETVKIDELNIAFERLRKNDVRYRFVVDVAG 353
AtCAD5          KVITGSFIGSMKETEEMLEFCKEKGLSSIIEVVKMDYVNTAFERLEKNDVRYRFVVDVEG 352
AkCAD2          KTITGSFIGSIKETEEMLEFWKEKGLSSMIEMVKMDYINKALERLEKNDVRYRFVVDVAG 352
ATCAD2          KMVGGSQIGGMKETQEMLEFCAKHKIVSDIELIKMSDINSAMDRLAKSDVRYRFVIDVAN 351
AtCAD3          KMVGGSQIGGMKETQEMLEFCAKHKIVSDIELIKMSDINSAMDRLVKSDVRYRFVIDVAN 350
AtCAD9          KMVGGSDVGGMKETQEMLDFCAKHNITADIELIKMDEINTAMERLAKSDVRYRFVIDVAN 355
AtCAD6          KSIAGSGIGGMQETQEMIDFAAEHGIKAEIEIISMDYVNTAMDRLAKGDVRYRFVIDISN 356
AkCAD3          KTISGSITGGTKETQEMIDFCAANKIYPNIEVIPIEYINEALQRVVNKDVKYRFVIDIEN 352
GmCAD           KTVAGSVTGGTKDIQEMIGFCAANEIHPNIEVIPIEYANEALERLINRDVKYRFVIDVEN 352
AtCAD1          RMLAGSVTGGTKITQQMLDFCAAHKIYPNIEVIPIQKINEALERVVKKDIKYRFVIDIKN 352
                : : **  *. :  :::: *   : : . ** : :.  * *::*: : *::****:*: .
```

```
AtCAD4          SNLVEEAATTTN------------ 365
AtCAD5          SNLDA------------------- 357
AkCAD2          SKLDDQ------------------ 358
ATCAD2          SLLPESSAEILTEQVDHGVSITSRF 376
AtCAD3          SLLPESSAEILTEHVDHGVSITSRF 375
AtCAD9          SLSPP------------------- 360
AtCAD6          TLAATRS----------------- 363
AkCAD3          SLK--------------------- 355
GmCAD           SLKEK------------------- 357
AtCAD1          SLK--------------------- 355
                :
```

6) *p*-coumarate 3-hydroxylase (C3H)

```
AkC3H           MALLLIF-ISVIGVFLCYQLYQRLRFKLPPGPRPWPVVGNLYDIKPVRFRCFAEWAQSYG 59
CkC3H           MALFLIIPLSLITLFLCYTLFQRLRFKLPPGPRPWPVVGNLYDIKPVRFRCFAEWAQSYG 60
AtC3H           MSWFLIA-VATIAAVVSYKLIQRLRYKFPPGPSPKPIVGNLYDIKPVRFRCYYEWAQSYG 59
                *: :**  :: *  .:.* * ****:*:**** * *:***************: *******

AkC3H           PIISVWFGSTLNVIVSNSELAREVLKDHDQQLADRHRSRSAAKFSRDGKDLIWADYGPHY 119
CkC3H           PIISVWFGSTLNVIVSNSELAREVLKENDQLLADRHRSRSAAKFSRDGKDLIWADYGPHY 120
AtC3H           PIISVWIGSILNVVVSSAELAKEVLKEHDQKLADRHRNRSTEAFSRNGQDLIWADYGPHY 119
                ******:** ***:**.:***:****::** ******.**:  ***:*:***********

AkC3H           VKVRKVCTLELFSPKRLEALRPIREDEVTAMVESIYRDSTNPENEGKSVMVKQYLGTVAF 179
CkC3H           VKVRKVCTLELFSPKRIEALRPIREDEVTAMVESIFNDCTNPENMGKGILMRKYLGAVAF 180
AtC3H           VKVRKVCTLELFTPKRLESLRPIREDEVTAMVESVFRDCNLPENRAKGLQLRKYLGAVAF 179
                ************:***:*:***************::.*.. *** .*.: :::***:***

AkC3H           NNITRLAFGKRFVNSEGIMDEQGVEFKAVVTNGLKLGASLAMAEHIPWLRWMFPLEEGAF 239
CkC3H           NNITRLAFGKRFVNSEGVIDEQGVEFKAIVANGLKLGASLAMAEHIPWLRWMFPLEEEAF 240
AtC3H           NNITRLAFGKRFMNAEGVVDEQGLEFKAIVSNGLKLGASLSIAEHIPWLRWMFPADEKAF 239
                ************:*:**::****:****:*:*********::************ :* **

AkC3H           AKHGARRDRLTRAIMEEHTQARQRSGGAKQHFVDALLTLQDKYDLSEDTIIGLLWDMITA 299
CkC3H           AKHGARRDRLTRAIMEEHTQARKQSGGAKQHFVDALLTLQDKYDLSEDTIIGLLWDMITA 300
AtC3H           AEHGARRDRLTRAIMEEHTLARQKSSGAKQHFVDALLTLKDQYDLSEDTIIGLLWDMITA 299
                *:***************** **::*.*************:*:******************

AkC3H           GMDTTAISVEWAMAELIKNPRVQQKAQEELDRVIGFERIITETDFSSLPYLQCVAKEAMR 359
CkC3H           GMDTTAISVEWAMAELIKNPRVQQKAQEELDKVIGFERVMTEADFSSLPYLQCVAKEAMR 360
AtC3H           GMDTTAITAEWAMAEMIKNPRVQQKVQEEFDRVVGLDRILTEADFSRLPYLQCVVKESFR 359
                *******:.******:*********.***:*:*:*::*::**:*** *******.**:::*

AkC3H           LHPPTPLMLPHRANANVKIGGYDIPKGSNVHVNVWAVARDPAVWKDPSEFRPERFLEEDV 419
CkC3H           LHPPTPLMLPHRANANVKVGGYDIPKGSNVHVNVWAVARDPAVWKDPSEFRPERFLEEDV 420
AtC3H           LHPPTPLMLPHRSNADVKIGGYDIPKGSNVHVNVWAVARDPAVWKNPFEFRPERFLEEDV 419
                ************:**:**:****************************:* ************

AkC3H           DMKGHDFRLLPFGAGRRVCPGAQLGINLVTSMLGHLLHHFCWTPPGGIKPEEIDMSENPG 479
CkC3H           DMKGHDFRLLPFGAGRRVCPGAQLGINLVTSMLGHLLHHFCWAPHEGVKPEEIDMAENPG 480
AtC3H           DMKGHDFRLLPFGAGRRVCPGAQLGINLVTSMMSHLLHHFVWTPPQGTKPEEIDMSENPG 479
                *********************************:.****** *:*  * *******:****

AkC3H           LVTFMRTPLQAVATPRLPSHLYKRVPAEI 508
CkC3H           LVAYMRTPLQAVASPRLPSDLYKRVPADI 509
AtC3H           LVTYMRTPVQAVATPRLPSDLYKRVPYDM 508
                **::****:****:*****.****** ::
```

7) Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase (HCT)

```
GmHCT           MMINVKESTMVRPAEEVARRVVWNSNVDLVVPNFHTPSVYFYRSNGAPNFFDGKVMKEAL 60
AkHCT3          MIINVKASTMVRPAEETPRQALWNSNVDLVVPNFHTPSVYFYRPTGAADFFDAEVMKQAL 60
AtHCT           MKINIRDSTMVRPATETPITNLWNSNVDLVIPRFHTPSVYFYRPTGASNFFDPQVMKEAL 60
                •**:: ******* *..   :********:*.**********..**.:*** :***:**
                •

GmHCT           TKVLVPFYPMAGRLLRDDDGRVEIDCDGQGVLFVEADTGAVIDDFGDFAPTLELRQLIPA 120
AkHCT3          AKALVPFYPMAGRLRRDEDGRVEIDCNGEGVLFVEAETTSLIDDFGDFAPTLELRQLIPA 120
AtHCT           SKALVPFYPMAGRLKRDDDGRIEIDCNGAGVLFVVADTPSVIDDFGDFAPTLNLRQLIPE 120
                :*.*********** **:***:****:* ***** *:* ::***********:******

GmHCT           VDYSQGIASYPLLVLQVTHFKCGGVSLGVGMQHHVADGASGLHFINTWSDVARGLDVSIP 180
AkHCT3          VDYSGGIETYPLLVLQVTYFKCGGVSLGVGMQHHAADGFSGLHFINTWSDMARGLDLTLP 180
AtHCT           VDHSAGIHSFPLLVLQVTFFKCGGASLGVGMQHHAADGFSGLHFINTWSDMARGLDLTIP 180
                **:* ** ::********.*****.*********.*** ***********:*****:::*

GmHCT           PFIDRTILRARDPPRPIFDHIEYKPPPAMKTQQAT--------NASAAVSIFRLTRDQLN 232
AkHCT3          PFIDRTLLRARDPPQPVFEHIEYKPPPTMKSPVQTASKPGSDNNPPAAVSIFKMTRDQLN 240
AtHCT           PFIDRTLLRARDPPQPAFHHVEYQPAPSMKIPLDPSK----SGPENTTVSIFKLTRDQLV 236
                ******:*******:* *.*:**:*.*:**    .          ::****::*****

GmHCT           TLKAKSKEDGNTISYSSYEMLAGHVWRSVSKARALPDDQETKLYIATDGRSRLQPPTPPG 292
AkHCT3          ALKAKSKEAGNTITYSSYEMLAGHVWKSTCKARALPDDQETKLYIATDGRSRLQPPLPPG 300
AtHCT           ALKAKSKEDGNTVSYSSYEMLAGHVWRSVGKARGLPNDQETKLYIATDGRSRLRPQLPPG 296
                :******* ***::*************:*. ***.**:*****************:*  ***

GmHCT           YFGNVIFTTTPIAVAGDLMSKPTWYAASRIHNALLRMDNDYLRSALDYLELQPDLKALVR 352
AkHCT3          YFGNVIFTATPMAMAGDLMSKPTWFAASRIHNALSRMDNEYLKSALDYLELQPDLKALVR 360
AtHCT           YFGNVIFTATPLAVAGDLLSKPTWYAAGQIHDFLVRMDDNYLRSALDYLEMQPDLSALVR 356
                ********:**:*:****:*****:**.:**: * ***::**:*******:****.****
```

```
GmHCT           GAHTFKCPNLGITSWTRLPIHDADFGWGRPIFMGPGGIAYEGLSFIIPSSTNDGSLSVAI 412
AkHCT3          GAHTFRCPNLGITSWVRLPIHDADFGWGRPIFMGPGGIAYEGLSFILPHSNNDGSLSVAI 420
AtHCT           GAHTYKCPNLGITSWVRLPIYDADFGWGRPIFMGPGGIPYEGLSFVLPSPTNDGSLSVAI 416
                ****::*********.****:*****************.******::* ..*********

GmHCT           ALQPDHMKLFKDFLYDI 429
AkHCT3          ALQPEHMKVFKEVLYEI 437
AtHCT           ALQSEHMKLFEKFLFEI 433
                ***.:***:*:..*::*
```

8) Caffeic acid O-methyltransferase (COMT)

```
AkCOMT          MGSAGETQITPTHVNDEEANLFAMQLASASVLPMILKSALELDLLEIIAKAGPNAQLSSS 60
AaxAmCOMT       MGSAGETQITPTHVNDEEANLFAMQLASASVLPMILKSALELDLLEIIAKAGPNAQLSPS 60
AtCOMT          MGSTAETQLTPVQVTDDEAALFAMQLASASVLPMALKSALELDLLEIMAKNG--SPMSPT 58
                ***:.***:**.:*.*:** ************** ************:** *  : :*.:

AkCOMT          DIASQLPTKNPDAAVMLDRMMRLLACYNVLSSSLRTLPDGKIERLYGLAPVAKYLVKTED 120
AaxAmCOMT       DIASQLPTKNPDAAVMLDRMMRLLACYNVLSSSLRTLPDGKIERLYGLAPVAKYLVKNED 120
AtCOMT          EIASKLPTKNPEAPVMLDRILRLLTSYSVLTCSNRKLSGDGVERIYGLGPVCKYLTKNED 118
                :***:******:*.*****::***:.*.**:.* *.*... :**:***.**.***.*.**

AkCOMT          GVSIAPLSLMNQDKVLMESWYYLTEAVLEGGIPFNKAHGMTSFEYHGKDARFNKVFNKGM 180
AaxAmCOMT       GVSIAPLNLMNQDKVLMESWYYLTETVLEGGIPFNKAHGMTSFEYHGKDARFNKVFNKGM 180
AtCOMT          GVSIAALCLMNQDKVLMESWYHLKDAILDGGIPFNKAYGMSAFEYHGTDPRFNKVFNNGM 178
                *****.* **************:*.:::*:********:**::*****.*.*******:**

AkCOMT          ADHSTITMKKILETYTGFEGLKSLVDVGGGTGAVISMIVSKFPSIKGFNFDLPHVIEEAP 240
AaxAmCOMT       ADHSTITMKKILETYTGFEGLKSLVDVGGGTGAVINTIVSKYPSIKGINFDLPHVIEEAP 240
AtCOMT          SNHSTITMKKILETYKGFEGLTSLVDVGGGIGATLKMIVSKYPNLKGINFDLPHVIEDAP 238
                ::**************.*****.******** **.:. ****:*.:**:*********:**

AkCOMT          SYPGVEHVGGDMFVSVPKADAVFMKWICHDWSDEHCVKFLKNCYDALPENGKVIVAECIL 300
AaxAmCOMT       SFPGVEHVGGDMFVSVPKADAVFMKWICHDWSDEHCVKFLKNCYDALPENGKVIVAECIL 300
AtCOMT          SHPGIEHVGGDMFVSVPKGDAIFMKWICHDWSDEHCVKFLKNCYESLPEDGKVILAECIL 298
                *.**:**************.**:*********************:::***:****:*****

AkCOMT          PVAPDSSLATKGVVHIDVIMLAHNPGGKERTEKEFEALAKGAGFQGFRVCCSAFNSYIIE 360
AaxAmCOMT       PVAPDSSLATKGVVHIDVIMLAHNPGGKERTEKEFEALAKGAGFQGFRVCCSAFNSYIME 360
AtCOMT          PETPDSSLSTKQVVHVDCIMLAHNPGGKERTEKEFEALAKASGFKGIKVVCDAFGVNLIE 358
                * :*****:** ***:* **********************.:**:*::* *.**.  ::*

AkCOMT          FLKKP 365
AaxAmCOMT       FLKKP 365
AtCOMT          LL--- 360
                :*
```

9) Ferulate 5-hydroxylase (F5H)

```
AkF5H           MNIIPAMDSLLVHLQPLGMAMLFAIPLILLLGLLSRVISKRPPYPPGPKGLPIIGNMMMM 60
GmF5H           ---MANLD-----LDPFQTSILILVPIALLVALLSRTR-RRAPYPPGPKGLPIIGNMLMM 51
AtF5H           --MESSISQTLSKLSDPTTSLVIVVSLFIFISFITRRR--RPPYPPGPRGWPIIGNMLMM 56
                    . :.     *.    :::: :.: :::.:::*    *.******:* ******:**

AkF5H           DQLTHRGLAKLANKYGGVLHLRMGFLHMVAISDAEAARQVLQVHDNIFSNRPATIAISYL 120
GmF5H           EQLTHRGLANLAKHYGGIFHLRMGFLHMVAISDPVAARQVLQVQDNIFSNRPATIAISYL 111
AtF5H           DQLTHRGLANLAKKYGGLCHLRMGFLHMYAVSSPEVARQVLQVQDSVFSNRPATIAISYL 116
                :********:**::***: ********* *:*.. .*******:*.:**************

AkF5H           TYDRADMAFAHYGPFWRQMRKICVMKLFSRKRAESWQSVREEVESVVHTVAKNTGKEVNI 180
GmF5H           TYDRADMAFAHYGPFWRQMRKLCVMKLFSRKRAESWQSVRDEVDAAVRAVASSVGKPVNI 171
AtF5H           TYDRADMAFAHYGPFWRQMRKVCVMKVFSRKRAESWASVRDEVDKMVRSVSCNVGKPINV 176
                ********************:****:********* ***:**:  *::*: ..** :*:

AkF5H           GELVFSLTKNITYRAAFGSTSQEGQGEFIGILQEFSKLFGAFNMADFIPGLAWVDPQGLN 240
GmF5H           GELVFNLTKNIIYRAAFGSSSQEGQDEFIKILQEFSKLFGAFNIADFIPYLGCVDPQGLN 231
AtF5H           GEQIFALTRNITYRAAFGSACEKGQDEFIRILQEFSKLFGAFNVADFIPYFGWIDPQGIN 236
                ** :* **:** *******:.::**.*** *************:***** :. :****:*

AkF5H           DRLAKARGSLDSFIDKIIDAHMNNKKQP-------DKEGDMVDELLAFYGEGESKVSESD 293
GmF5H           SRLARARGALDSFIDKIIDEHVHKMKNDKSSEIV-DGETDMVDELLAFYSEEAKLNNESD 290
AtF5H           KRLVKARNDLDGFIDDIIDEHMKKKENQNAVDDGDVVDTDMVDDLLAFYSEEAKLVSETA 296
                .**.:**. **.***.*** *::: ::          : ****:*****.*  .  .*:

AkF5H           DLQNSIKLTKDNIKAIIMDVMFGGTETVASAIEWAMAELMRSPEDLKRVQQEMAEVVGLD 353
GmF5H           DLQNSIRLTKDNIKAIIMDVMFGGTETVASAIEWAMAELMRSPEDQKRVQQELADVVGLD 350
AtF5H           DLQNSIKLTRDNIKAIIMDVMFGGTETVASAIEWALTELLRSPEDLKRVQQELAEVVGLD 356
                ******:**:************************::**:***** ******:*:*****

AkF5H           RGVEESDIEKLTYLKCCLKETLRLHPPIPLLLHETAEDAVVSGYFVPKKSRVMINAWAIG 413
GmF5H           RRAEESDFEKLTYLKCALKETLRLHPPIPLLLHETAEDATVGGYFVPRKARVMINAWAIG 410
AtF5H           RRVEESDIEKLTYLKCTLKETLRMHPPIPLLLHETAEDTSIDGFFIPKKSRVMINAFAIG 416
                • .****:******** ******:**************: :.*:*:*:*:******:***
                •
```

```
AkF5H           RDRNAWDDADRFKPSRFLGEGVPDFKGSNFEFIPFGSGRRSCPGMQLGLYGLELSVAYLL 473
GmF5H           RDKNSWEEPETFKPARFLKPGVPDFKGSNFEFIPFGSGRRSCPGMVLGLYALELAVAHLL 470
AtF5H           RDPTSWTDPDTFRPSRFLEPGVPDFKGSNFEFIPFGSGRRSCPGMQLGLYALDLAVAHIL 476
                ** .:* :.: *:*:***  ************************ ****.*:*:**::*

AkF5H           HYFSWELPNGMKPSDMDMSDVFGLTAPRASRLIAVPRKRLVCPLF 518
GmF5H           HCFTWELPDGMKPSEMDMGDVFGLTAPRSTRLIAVPTKRVVCPLF 515
AtF5H           HCFTWKLPDGMKPSELDMNDVFGLTAPKATRLFAVPTTRLICAL- 520
                * *:*:**:*****::**.*********:::**:*** .*::*.*
```

10) Caffeoyl CoA 3-O-methyltransferase

```
AkCCoAOMT1      ------------------------------------------------------------
GmCCoAOMT       MPKPCCSMHKHYYKTNLANPGPDANRSNQTSHRFEPAFSFLISITLLHPPTSSSSNYQLF 60
AkCCoAOMT2      ------------------------------------------------------------
AtCCoAOMT1      ------------------------------------------------------------
AkCCoAOMT5      ------------------------------------------------------------
AtCCoAOMT7      ------------------------------------------------------------

AkCCoAOMT1      ------------------MADQNQ---SEAGRHQEVGHKSLLQSDALYQYILDTSVYPRE 39
GmCCoAOMT       QKGEEKERKQNAQRIIIAMAEQNQNQTTEAGRHQEVGHKSLLQSDALYQYILETSVYPRE 120
AkCCoAOMT2      ----------------MASSNGEEKKQSEAGRHQEVGHKSLLQSDALYQYILETSVYPRE 44
AtCCoAOMT1      ------MATTTTEATKTSSTNGEDQKQSQNLRHQEVGHKSLLQSDDLYQYILETSVYPRE 54
AkCCoAOMT5      ---------------MASDKEDAAPKSQASGPHKESSHKSLLQSDALYQYILETSVYPKE 45
AtCCoAOMT7      --------------------------------MAKDEAKGLLKSEELYKYILETSVYPRE 28
                                               :   *.**:*: **:***:*****:*

AkCCoAOMT1      PEPMKELREITAKHPWNIMTTSADEGQFLNMLLKLINAKNTMEIGVYTGYSLLATALALP 99
GmCCoAOMT       PESMKELRELTAKHPWNIMTTSADEGQFLNMLLKLINAKNTMEIGVYTGYSLLATALALP 180
AkCCoAOMT2      PEPMKELREITAKHPWNIMTTSADEGQFLNMLLKLINAKNTMEIGVYTGYSLLATALALP 104
AtCCoAOMT1      PESMKELREVTAKHPWNIMTTSADEGQFLNMLIKLVNAKNTMEIGVYTGYSLLATALALP 114
AkCCoAOMT5      HQCLKEMRDLTEKHPWNCIAISADEGQFLNMLVKLINAKNTIEIGVFTGYSLLATALALP 105
AtCCoAOMT7      PEVLRELRNITHNHPQAGMATAPDAGQLMGMLLNLVNARKTIEVGVFTGYSLLLTALTLP 88
                 : ::*:*::* :**   :: :.* **::.**::*:**::*:*:**:****** ***:**

AkCCoAOMT1      EDGKILAMDINKENYELGLPVIQKAGVAHKITFKEGPALPVLDELVKDEKNHGSYDFIFV 159
GmCCoAOMT       EDGKILAMDINRENYELGLPVIKKAGVDHKIEFREGPALPVLDEMVKDEKNHGSYDFIFV 240
AkCCoAOMT2      DDGKILAMDINRENYELGRPTLEKAGVAHKVDFREGPALPFLDELVKDEKNHGSFDFIFV 164
AtCCoAOMT1      EDGKILAMDVNRENYELGLPIIEKAGVAHKIDFREGPALPVLDEIVADEKNHGTYDFIFV 174
AkCCoAOMT5      QDGK----------------------------------------EKNKGAFDYIFV 121
AtCCoAOMT7      EDGKVIAIDMNRDSYEIGLPVIKKAGVEHKIDFKESEALPALDELLNNKVNEGGFDFAFV 148
                :***                                          : *.* :*: **

AkCCoAOMT1      DADKDNYLNYHRRLIDLVKVGGVIGYDNTLWNGSVVAPPDAPLRKYVRYYRDFVLELNKA 219
GmCCoAOMT       DADKDNYLNYHKRLIELVKVGGVIGYDNTLWNGSVVAPPDAPLRKYVRYYRDFVLELNKA 300
AkCCoAOMT2      DADKDNYLNYHERLLQLVKVGGVIGYDNTLWNGSVVAPDDAPLRKYVMYYREFVLKLNKA 224
AtCCoAOMT1      DADKDNYINYHKRLIDLVKIGGVIGYDNTLWNGSVVAPPDAPMRKYVRYYRDFVLELNKA 234
AkCCoAOMT5      DADKENYLNYHKRAIELVKVGGVIAYDNTLWNGSVADPPEVPQLKHIKHSLAFVQELNKA 181
AtCCoAOMT7      DADKLNYWNYHERLIRLIKVGGIIVYDNTLWGGSVAEP-DSSTPEWRIEVKKATLELNKK 207
                **** ** ***.* : *:*:**:* ******.***. * : .  :        . :***

AkCCoAOMT1      LAVDPRIEICMLPVGDGVTLCRRIS 244
GmCCoAOMT       LAVDPRIEICMLPVGDGITICRRIK 325
AkCCoAOMT2      LAVDPRIEICMLPVGDGITLCRRIK 249
AtCCoAOMT1      LAADPRIEICMLPVGDGITICRRIS 259
AkCCoAOMT5      LALDSRIEICQLPIADGITLCRRIT 206
AtCCoAOMT7      LSADQRVQISQAALGDGITICRRLY 232
                *: * *::*.  .:.**:*:***:
```

Fig. S1. Alignments of the monolignol biosynthesis enzymes