Identification and analysis of CYP450 supergene family members from the transcriptome of *Aralia elata* (Miq.) Seem reveal candidate genes for triterpenoid saponin biosynthesis

Yao Cheng
Northeast Agricultural University https://orcid.org/0000-0001-5561-9343
Hanbing Liu
Northeast Agricultural University
Xuejiao Tong
Northeast Agricultural University
Zaimin Liu
Northeast Agricultural University
Xin Zhang
Northeast Agricultural University
Dalong Li
Northeast Agricultural University
Xinmei Jiang
Northeast Agricultural University
Xihong Yu (yxhong001@163.com)
Northeast Agricultural University

Research article

Keywords: Aralia elata (Miq.) Seem, Cytochrome P450, Transcriptome-wide identification, Triterpenoid saponin, Subcellular localization

Posted Date: February 21st, 2020

DOI: https://doi.org/10.21203/rs.2.20462/v2

Version of Record: A version of this preprint was published at BMC Plant Biology on May 13th, 2020. See the published version at https://doi.org/10.1186/s12870-020-02411-6.

Abstract

Background: Members of the cytochrome P450 (CYP450) gene superfamily have been shown to play essential roles in regulating secondary metabolites biosynthesis. However, the systematic identification and bioinformatics analysis of CYP450s have not been reported in Aralia elata (Miq.) Seem , a highly valued medicinal plant.

Results: In the present study we conducted the RNA-sequencing (RNA-seq) analysis of the leaves, stems, and roots of A. elata, yielding 66,713 total unigenes. Following the annotation and classification of these unigenes, we were able to identify two pathways and 19 putative genes associated with the synthesis of triterpenoid saponins in these plants, with qRT-PCR subsequently being used to validate these gene expression patterns. Scanning with the CYP450 model from Pfam resulted in the identification of 111 full-length and 143 partial-length CYP450s, with the full-length CYP450s being further clustered into 7 clans and 36 families. Through phylogenetic and conserved motif analyses, we were further able to group these CYP450 proteins into two primary branches: A-type (53%) and non-A type (47%). We further conducted representative protein sequence alignment for these CYP450 family members, with secondary elements being assigned in light of the recently published Arabidopsis CYP90B1 structure. Using the available sequence information, we further identified predicted substrate recognition sites (SRSs) and substrate binding sites within these putative proteins.We further assessed the expression patterns of these 111 CYP450 genes across A. elata tissues, with 12 members of this gene family being selected at random for qRT-PCR validation. From these data, we identified CYP716A295 and CYP716A296 as the candidate genes most likely to be associated with oleanolic acid synthesis, while CYP72A763 was identified as being the most likely to play a role in hederagenin biosynthesis. Finally, we assessed the subcellular localization of these CYP450 proteins within Arabidopsis protoplasts, highlighting the fact that they localize to the endoplasmic reticulum.

Conclusions: This study presents a systematic analysis of the CYP450 gene family in A. elata and provided a foundation for further functional characterization of CYP450 genes.

Background

Plant cytochrome P450 (CYP450) supergene family proteins are key enzymes involved in a wide range of metabolic processes in plants, including the synthesis of sterols, flavonoids, terpenoids, and other secondary metabolites [1, 2]. To date, over 5,100 distinct CYP450 sequences have been defined [3], however, owing to the substantial diversity of these CYP450 proteins with respect to their reactivity and substrate/product accessibility, they have proven difficult to fully functionally characterize [4]. It is essential that efforts be taken to systematically identify and characterize the CYP450 members in a given plant using genomic technologies. The systematic genome-wide classification of CYP450s has, to date, only been conducted in some model species, such as *Arabidopsis* [5], *Medicago truncatula* [6] and rice [7]. The advent of next-generation sequencing technologies such as RNA-Seq has allowed for the more rapid and cost-effective identification of CYP450s [8]. Recent transcriptomic studies have, for example,

detected 116 full-length and 135 partial-length CYP450s in *Salvia miltiorrhiza* [9], with similar work being performed in *Lonicera japonica* [10] and *Taxus chinensis* [11].

The vast majority of CYP450s are membrane-localized proteins that are predictably retained in the endoplasmic reticulum (ER) by N-terminal transmembrane helix [12], while very few CYP74s and CYP97s have been reported to localize to the chloroplast membrane [13, 14]. The crystal structures of a variety of mammalian and bacterial CYP450 have been produced, but very few plant CYP450 structures have been reported to date. A recent study detailed the first plant CYP450 protein structure, providing a crystal structure for *Arabidopsis thaliana* CYP90B1 that offers great value as a template for homology-based studies of the structures of other plant CYP450s [15]. Most CYP450 proteins exhibit conservation in four regions: a heme-binding region, an I-helix, a K-helix, and a PERF motif, with the greatest variability having been observed in the substrate recognition sites (SRSs) of these proteins [16].

Aralia elata (Miq.) Seem is a member of the *Araliaceae* family that grows widely throughout Korea, Japan, Russia, and China, where it is used both as a food and as a medicinal plant [17]. Owing to their unique taste, young *A. elata* shoots are commonly eaten in many regions of Asia [18]. In addition, the roots and bark of these plants are often incorporated into the traditional Chinese medicine known as “cilaoya”. Previous phytochemical studies have determined that triterpenoid saponins are the primary bioactive substances within *A. elata*, and these compounds have been employed for the treatment of neurasthenia [19], diabetes mellitus [20], hepatitis [21], and gastrospasm [22]. A number of distinct triterpene saponins (chikusetsusaponins Iva and IV and aralosides A, B, V, VII, and X) have been isolated from the leaves [23, 24] and root bark [25, 26] of *A. elata*. Therefore, *A. elata* is ideal for the study on the biosynthesis of triterpenoid saponins, and in particular those of hederagenin and oleanane-types.

Triterpenoids and steroids are a highly diverse group of natural products and they largely share a metabolic pathway that can be divided into three parts [27] (Fig. 1A). First, terpenoids are constructed from C5 units, isopentenyl diphosphate (IPP), which is supplied either from the cytosolic mevalonic acid (MVA) pathway or from the plastidal methylerythritol phosphate (MEP) pathway. Triterpenoids are biosynthesized via the MVA pathway. In addition, IPP can be converted into its isomer, DMAPP (dimethylallyl diphosphate) by IDI (isopentenyl diphosphate isomerase) [28]. IPP and DMAPP are then finally converted into 2,3-oxidosqualene by a series of enzymes, including GPS (geranyl diphosphatesynthase), FPS (farnesyl diphosphate synthase), squalene synthase (SS) and squalene epoxidase (SE). The cyclization of 2,3-oxidosqualene is then catalyzed by a class of oxidosqualene cyclases (OSCs) to form a variety of triterpenoid backbones [29], including β-amyrin, phytosterol, dammarane and lupane [27]. This step is thus a critical branching point for triterpenoid and phytosterol biosynthesis [30] (Fig. 1A). Finally, CYP450s and UDP-glycosyltransferases (UGTs) govern oxidation, hydroxylation, and glycosylation steps so as yield triterpenoid saponins and phytosterol [31]. In the context of pentacyclic triterpenoid saponin biosynthesis, CYP450s introducing a carboxyl group at C-28 and hydroxyl groups at C-2β, C-16α, C-23 and C-24 of the β-amyrin skeleton are predicted to form multiple sapogenins, such as oleanolic acid, hederagenin and glycyrrhetinic acid [32] (Fig. 7B). In contrast, UGTs

that can glycosylate the sapogenins at the C-3 and C-28 position are predicted to form monodesmosidic or bisdesmosidic saponins with specific structures and activities [33].

In this study, we performed RNA-sequencing in order to analyze the transcriptomic profiles of three different *A. elata* tissues. We then further sought to identify those genes associated with triterpenoid saponin biosynthesis, using pathway enrichment analyses in order to identify unigenes predicted to be involved in the MEP and MVA pathways. We further conducted a systematic analysis of *A. elata* CYP450 family members by identifying full-length CYP450-encoding sequences in our RNA-seq datasets and then conducting pathway enrichment, phylogenetic, structural, and expression pattern-based analyses of these identified genes. Lastly, we mined this CYP450 family member gene set in an effort to identify members involved in regulating triterpenoid saponin biosynthesis, leading us to identify three candidate CYP450s that were then subjected to subcellular localization analyses. The results of this study will help to foster further research aimed at better understanding the role of CYP450 genes in *A. elata*.

Results

Quantitative analysis of A. elata aralosides

The saponins present within *A. elata* primarily contain oleanolic acid and hederagenin aglycone [19], with total and monomer araloside accumulation varying widely between different tissues in these plants. Specifically, the leaves of *A. elata* have been found to contain the largest quantity of these saponins, with progressively lower levels found in root and stem tissues. The roots of these plants contained higher oleanolic acid levels than did the other tested tissues, whereas hederagenin levels were highest in leaves relative to samples roots and stems (Table 1; Additional file 1: Figure S1). Two selected oleanane-type saponins (chikusetsusaponin IV and araloside X) and one hederagenin-type saponin (araloside VII) were also detected in these *A. elata* samples, with root chikusetsusaponin IV levels being fairly high whereas they were minimal in leaf and stem tissues. In contrast, we detect aralosides VII, and X showed a high level in the leaves of these plants, suggesting that different glycosyltransferases were responsible for their generation. These tissue-specific saponin distribution results offer significant value as a reference source when identifying those CYP450s and other proteins involved in araloside production in *A. elata*.

De novo A. elata sequence assembly

In order to identify genes pertaining to saponin biosynthesis in these *A. elata* plants, we next employed an Illumina HiSeq 4000 platform to sequence the total RNA transcriptome in root, leaf, and stem tissue samples. In total this approach yielded 448,112,618 reads that were assembled into 82,238 contigs, with the longest being 16,016 bp, and with an average contig length of 1,058 bp. We were then able to assemble these contigs into 66,713 unigenes with a 1,846 bp N50 length (Table 2). Next, these unigenes were annotated with the KEGG, UniProt, NCBI nonredundant nucleotide (Nt), and Nr databases via use of the BLASTN and BLASTX algorithms, leading to the annotation of 35,232 (52.81%) unigenes (Additional file 2: Figure S2). These transcriptome sequence data have been deposited in the NCBI Short Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under the accession number PRJNA555256.

Enrichment of terpenoid backbone and triterpenoid biosynthetic pathways

Following the annotation of 7,291 *A. elata* unigenes with the KEGG database, we were able to assign unigenes to the terpenoid backbone and sesquiterpenoid/triterpenoid biosynthesis pathways, which contained the upstream MVA and MEP pathways and the 2,3-oxidosqualene biosynthesis pathway (Fig. 1A). In total, we mapped 79 unigenes to the terpenoid backbone biosynthesis pathway, while 47 were mapped to the sesquiterpenoid and triterpenoid biosynthesis pathways. As shown in Fig 1A, 6 putative genes (AACT, HMGS, HMGR, MVK, PMVK, and MVD) and 8 putative genes (DXS, DXR, MEP-CT, COP-MEK, MECDPS, HMBPPS, HMBPPR, and IDI) were associated with the MVA and MEP pathways, respectively, while 5 putative genes (GPS, FPS, SS, SE, and bAS) were found to be associated with carbocyclic biosynthesis (Additional file 3: Table S1). For the majority of these unigenes, we were able to map >1 unigene to a given gene or gene family, suggesting that these sequences may correspond to different fragments and/or isoforms of a given gene [34]. Those unigenes that were expressed at the highest levels in each of these enzymatic steps were next selected and arranged into a heat map showing the differentially expressed genes associated with triterpene saponin biosynthesis (Fig. 1A). The expression of these genes differed substantially in a tissue-dependent manner, as confirmed via qRT-PCR, with those genes involved in the MEP synthesis pathway being highly expressed in leaf samples (Fig. 1B), and those involved in the MVP synthesis pathway being expressed at higher levels in different tissues (Fig. 1B).

A. elata CYP450 identification and classification

Through our transcriptome analysis, we were able to identify 111 full-length and 143 partial CYP450 genes in these *A. elata* samples. This number of total CYP450 unigenes (254) was lower than the number identified in a study of *Panax. ginseng* (484). We next aligned the 111 full-length CYP450s with the CYP450 database, using allelic, subfamily, and family variant cutoff values of 97%, 55%, and 40%, respectively [35]. Based on these sequence similarity findings, we were able to classify these CYP450s into 7 clades, 36 families and 64 subfamilies, with 53% being A-type CYP450s and 47% being non-A-type CYP450s (Additional file 4: Table S2). The CYP71 clan was the most highly represented in these samples, containing 59 genes belonging to 16 families (CYP71, CYP73, CYP75-CYP78, CYP80-CYP82, CYP84, CYP89, CYP92, CYP98, CYP701, CYP706, and CYP736). The next largest clan was CYP85, which contained 18 genes belonging to 8 families.

We next analyzed the primary characteristics of each of these CYP450 genes, including their predicted coding sequence (CDS) length, protein sequence length, protein molecular weight (MW), isoelectric point (pI), and subcellular localization (Additional file 4: Table S2). These 111 CYP450 proteins were predicted to be 410 - 620 amino acids long (average: 508), with MWs ranging from 46.8 - 69.5 kDa, and with pI values ranging from 6.01 - 9.55. We additionally calculated the instability index (II) of these proteins, revealing 59 of them to be unstable (stability factor >40), while 52 proteins were stable (stability factor < 40). The GRAVY values were negative for these proteins, indicating them to be hydrophilic. Predictive analyses of the subcellular localization of these CYP450s suggested that they were all localized

exclusively to the ER in *A. elata*. Overall, the results of these predictive analyses were in line with prior studies of CYP450s and their subcellular localization.

A.elata CYP450 pathway enrichment analyses

To further understand the functional importance of the identified CYP450s in *A. elata*, we next conducted a KEGG pathway enrichment analysis of these 111 genes. Individual CYP450s were assigned to multiple KEGG pathways, with 34 (30.63%) being assigned to 18 metabolism pathways and six classes, including global and overview, biosynthesis of other secondary metabolites, metabolism of terpenoids and polyketides, lipid metabolism, amino acid metabolism, and metabolism of cofactors and vitamins (Additional file 5: Figure S3).The pathways most enriched for these CYP450s were the biosynthesis of secondary metabolites and metabolic pathways, which contained 21 and 19 CYP450s, respectively. In total, 14 CYP71 clan members were represented in 14 distinct pathways, suggesting that members of the CYP71 clan exhibit highly diverse biological activities. We further found that 6 CYP450s, all of which were in the CYP86 clan, were involved in cutin, suberine, and wax biosynthesis. Additionally, the CYP71D603 and CYP71D610 proteins from the CYP71 clan were involved in sesquiterpenoid and triterpenoid biosynthesis (Additional file 5: Figure S3).

CYP450 family member phylogenetic and conserved motif analyses

To further explore the functional roles and evolutionary relationships for these *A. elata* CYP450s, we next constructed a phylogenetic tree incorporating these 111 CYP540s along with 59 CYP450s from other plants, including *A. thaliana* (37), *P. ginseng* (20), *Daucus carota* (2) and *Polygala tenuifolia* (1). The resultant tree (Fig. 2) divided these CYP450s into 7 clades, including three single-family clans (CYP51, CYP97 and CYP711) and four multifamily clans (CYP85, CYP86, CYP72 and CYP71). Genes within the same clan clustered in a single clade, with, for example, the 59 CYP71 members that were assigned to 16 families forming a single clade along with 28 representative CYP450s, while the CYP85 clan, which contained 18 CYP450s in 3 families, clustered with 11 representative CYP450s. This analysis additionally identified two clans, CYP711 and CYP51, containing only a single member each. These 7 clades were further grouped into the A-type and non-A type clusters. All A-type CYP450s were grouped into the CYP71 clan, with all other clans being of the non-A-type.

We further sought to identify conserved motifs within these CYP450s using the MEME web server. This analysis identified 10 different conserved motifs, with motif 1, motif 2, motif 3, and motif 4 being evident within all *A. elata* CYP450s. The most similar CYP450s typically had the most similar motif composition (Additional file 6: Figure S4). For example, the adjacent CYP72 and CYP97 clans exhibited similar motif compositions. Members of the CYP71 clan contained 10 conserved motifs, with motifs 5 and 10 being unique to the members of this clan, suggesting that these motifs may have functional roles that are specific to these proteins. Interestingly, we found that motif 6 and motif 9 were present in both the CYP71 and CYP72 clans, potentially highlighting the functional divergence of CYP450 genes. Together the results of these two analyses served to reaffirm the accuracy of the classification of these *A. elata* family proteins into defined clans and families.

CYP450 multiple sequence alignment and secondary structural element assignment

The prototypical CYP450 heme-binding domain, K-helix region, PERF motif, and I-helix region were detected in all of these *A. elata* CYP450s, with the residues in some of these regions being identical across these CYP450s (Fig. 3). Overall, the majority of conserved residues were located in these found different conserved motif regions. For example, C residues in the heme-binding region were highly conserved, as were R residues in the PXRX motif and E and R residues in the EXXR motif. We further used ESPript to conduct homology alignments and secondary structure predictions for these CYP450s, revealing that overall the sequence homology of these proteins was fairly low (< 30%), whereas their secondary structures were very similar (Fig. 3). This analysis revealed that α-helices were the primary elements composing these proteins, with random coils, and β-sheets and turns being more dispersed throughout the protein. We additionally mapped Gotoh's SRSs and putative substrate binding sites onto these CYP450s through alignment with the P450cam sequence [36], leading to the identification of 6 SRSs in these *A. elata* CYP450s, with the majority of these being distributed in different structural elements and associated with particular substrates, such as helices 4 and 6, and β-sheets 11 and 12.

A. elata CYP450 gene expression patterns

In order to understand the tissue-specific expression of CYP450s in *A. elata*, we next conducted an analysis of their expression in the three different tissues that had been subjected to RNA-seq analysis. Of these 111 CYP450s, we found that 97 (87.39%) exhibited patterns of differential expression across these tissues. Leaves expressed a high proportion (41.44%) of highly expressed CYP450s, whereas stem samples contained the smallest number of these genes (20.72%). A hierarchical clustering analysis was used to assess CYP450 coexpression patterns in these different analyses, with an expression profile heat map for these genes being constructed according to their RPKM normalized expression values. This clustering analysis led to the assignment of 111 CYP450s to six clustered (C1-C6; Fig. 4). The CYP450s that were most highly expressed in roots (17 genes), leaves (38 genes), and stems (7 genes) were grouped into clusters C1, C4, and C6, respectively. In addition, those CYP450s in clusters C2 (29 genes), C3 (3 genes), and C5 (17 genes) were expressed at the lowest levels in leaf, stem, and root tissues, respectively. A qRT-PCR approach was further used to validate these transcriptomic findings, with the expression of 12 randomly selected CYP450s representative of 7 clans being quantified (Fig. 5A, B).

Previous research has shown that members of the CYP72A and CYP716A subfamilies are the primary CYP450s involved in pentacyclic triterpenoid saponin biosynthesis. As such, we next specifically focused on the co-expression patterns of the 3 CYP716A and 6 CYP72A genes identified in the *A. elata* transcriptome with β-amyrin synthase (bAS), which is a oxidosqualene cyclase that catelyzes 2,3-oxidosqualene to form the β-amyrin skeleton (Fig. 6). The qRT-PCR profiles for these genes revealed that two CYP716A genes (CYP716A295 and CYP716A296) and two CYP72A genes (CYP72A762 and CYP72A764) exhibited a similar expression patterns to that of bAS, being expressed at high levels in leaf tissues relative to stems and roots. In contrast, CYP72A759 and CYP72A763 being expressed at higher

levels in leaves and at lower levels in stems and roots, whereas CYP716A306, CYP72A760, and CYP72A761 were expressed at the highest levels in stems.

Identification of candidate CYP450s involved in triterpenoid biosynthesis

As previously described, hederagenin aglycone and oleanolic acid were the major sapogenins in *A. elata,* so we specially focused on CYP450s related to the biosynthesis of these two sapogenins. Up to date, a total of 36 CYP450s have been found to play roles in triterpenoid biosynthesis (Fig. 7A; Additional file 7: Table S3). The CYP716A and CYP72A subfamilies are the primary CYP450 gene families involved in pentacyclic triterpenoid saponin diversification, with the CYP716A family being the largest multifunctional C28-oxidase family involved in such oleanane-type triterpenoid saponins biosynthesis [32, 37, 38] (Fig. 7A). We were able to identify 3 CYP716A genes (CYP716A295, CYP716A296, and CYP716A306) and 6 CYP72A genes (CYP72A759-764) in the *A. elata* transcriptome. In order to identify the most relevant unigenes involved in pentacyclic triterpenoid saponin biosynthesis for further analysis, we conducted BLASTx searches that compared these *A. elata* CYP450s to those 36 CYP450s known to be involved in triterpenoid biosynthesis. This analysis revealed that the *A. elata* CYP716A295 and CYP716A296 exhibited 93.97% and 94.39% sequence identity with *P. ginseng* CYP716A52v2 respectively, which is a β-amyrin 28-oxidase enzyme involved in oleanolic acid production [39] (Fig. 7B). Moreover, CYP716A295 and CYP716A296 were highly expressed in leaves relative to stems and roots, cosistent with bAS expression pattern and oleanolic acid contents (Fig. 6, Table 1). As such, we selected CYP716A295 and CYP716A296 as the best candidate CYP450s likely to be involved in oleanolic acid biosynthesis in *A. elata.* Two CYP450s (CYP72A397 and CYP72A68v2) have thus far been identified as oleanolic acid 23-oxidases, catalyzing oleanolic acid oxidation into hederagenin [40, 41] (Fig. 7B). In the present study, we observed higher expression of CYP72A759 and CYP72A763 in leaf tissues, with progressively lower levels in stems and roots, consistent with the observed hederagenin content distribution. A BLASTx analysis further indicated that CYP72A759 encoded a protein with <50% identity to CYP72A397 and CYP72A68v2, and that CYP72A763 shared 52.49% identity with CYP72A397. Given these results, we further selected CYP72A763 as the CYP450 most likely to be involved in hederagenin biosynthesis, although further functional validation will be necessary.

Phylogenetic analyses revealed CYP716A295 and CYP716A296 to be grouped in the CYP716A subfamily and to be most closely related to *P. grandiflorus* CYP716A140v2 and *P. ginseng* CYP716A52v2, respectively, both of which encode β-amyrin 28-oxidase enzymes involved in oleanolic acid production [33, 39]. CYP72A763 clustered in the CYP72A group and was most closely related to *M. truncatula* CYP72A61v2, which can transform 24-OH-β-amyrin into soyasapogenol B [34] (Fig. 7).

Assessment of the subcellular localization of three CYP450:GFP fusion proteins

Almost all CYP450s are membrane-associated proteins that localize to the ER, with relatively few localizing to chloroplasts and mitochondria [12]. As detailed above, all 111 of the *A. elata* CYP450s identified in this study were predicted to localize to the ER. To confirm this prediction, we therefore conducted the PEG-mediated transient expression of *Arabidopsis* protoplasts co-transformed with

CYP450-GFP reporter proteins and GHD7-RFP (a nuclear marker), with CYP716A295, CYP716A296, and CYP72A763 all being selected for this subcellular localization analysis. We observed no overlap between the CYP716A295, CYP716A296, and CYP72A763 GFP fluorescent proteins and GHD7-RFP fluorescence and , with these CYP450-GFPe proteins instead appearing as a reticular ribbon upon microscopic examination, consistent with their likely localization to the ER (Fig. 8).

Discussion

While both *A. elata* and *P. ginseng* are triterpenoid saponin rich members of the *Araliaceae* family that are commonly used in traditional medicinal contexts, *P. ginseng* has been far better-studied to date, with its genome having been released in the Ginseng Genome Database (http://ginsengdb.snu.ac.kr/). In contrast, there have been minimal molecular biology studies conducted to date focusing on *A. elata*, with no corresponding genomic or transcriptomic data being available for this species in the NCBI database. The present study was the first to conduct a de novo transcriptome analysis of *A. elata*, analyzing three replicates each of root, stem, and leaf tissue samples from these plants. An Illumina HiSeq 4000 platform was used to sequence the libraries prepared from these 9 samples, yielding 66,713 unigenes, of which over half were well-annotated within public databases. The N50 length and average length of the unigenes were consistent with the effective and high-quality assembly of these sequencing results [42]. The results of this deep sequencing analysis have immense value as a means of identifying those genes involved in the biosynthesis of pharmacologically-relevant secondary metabolites in *A. elata*, as evidenced by our identification of CYP450s involved in triterpenoid saponins in these plants.

After the assembly and annotation of the *A. elata* transcriptome in the present study, we were able to begin examining the terpenoid biosynthesis backbone and sesquiterpenoid and triterpenoid biosynthesis pathways in these samples, leading to the identification of 19 functional genes. While these two pathways are well-known to be important for the biosynthesis of terpenoids and sterols in *P. notoginseng* [43] and *Hedera helix* L. [34], the results of the present analysis are the first such comprehensive analysis of these pathways in *A. elata*. Triterpenoids and sesquiterpenoids are biosynthesized via the MVA pathway, that takes place primarily in the cytoplasm, whereas monoterpenoids, diterpenoid, and tetraterpenoids are biosynthesized via the MEP pathway, that takes place primarily in the plastid [23]. Chloroplasts are abundant in leaf samples less in stem but not in root samples, this may explain why we found that MEP pathway-associated genes were primarily upregulated in leaf samples, as has also previously been observed in *Periploca sepium* Bunge and *Cymbopogon winterianus* [44, 45]. This fact may also explain why we observed a higher abundance of triterpenoid saponins in the leaves of *A. elata* relative to the root and stem tissues.

CYP450s compose one of the largest enzymatic families, catalyzing irreversible oxidation reactions and being subject to complex functional classification [46]. Over 5,100 play CYP450 sequences have been identified to date (http://drnelson.uthsc.edu/CytochromeP450.html), with hundred of these proteins being encoded in the genome of a given plant. For example, 246 functional CYP450s have been identified in *Arabidopsis*, while 355 have been identified in rice [7], and 484 have been identified in *P. ginseng*. In the

present study, we conducted systematic identification, nomenclature assignments, and structural and functional analyses of CYP450 genes in *A. elata*. In total we identified 111 full-length and 143 partial CYP450 genes, and all which were novel and had not previously been reported in *A. elata*. We further classified the 111 full-length CYP450s into 7 clan, with the CYP74, CYP710, CYP727, and CYP746 clans not being identified in the present analysis owing to the absence of an available whole-genome sequence for *A. elata.* The CYP51 clan member genes are thought to be the evolutionarily oldest CYP450s, having evolved from a sterol-metabolizing CYP51 ancestor [7]. We identified only a single CYP51 clan member in the present analysis (CYP51G1). The CYP71 family is the largest CYP450 clan, being composed of 16 different families and making up the entirety of A-type CYP450 genes, with two unique conserved motifs that are involved in the biosynthesis of most secondary metabolites in plants [47].

Despite having distinct substrates and < 30% sequence identity in many cases, virtually all CYP450s have been found to adopt similar secondary structures [48]. An X-ray structure analysis of substrate-bound CYP90B1 [15] recently allowed investigators to conduct a sequence alignment and to conduct secondary structural element prediction analysis, leading to the mapping of six putative SRSs to the aligned sequences according to Gotoh's prediction models [36]. All *A. elata* CYP450s were found to contain 4 conserved motifs, including certain highly conserved amino acids, with these domains being important for protein folding and assembly [12]. By analyzing SRSs, it is possible to gain further insight into the structure and function of these CYP450s. We were able to identify 6 SRSs in *A. elata* CYP450s, with the majority being located in regions of variable structure, consistent with previous studies conducted in *M. truncatula* [6]. This suggests that these variable regions are the primary determinants of CYP450 substrate specificity. Together these results may aid in the molecular design of engineered CYP450 enzymes in *Araliaceae* plants that have novel substrate specificities.

Ever since CYP716A12 was first described as a triterpenoid-oxidizing enzyme that catalyzes α-amyrin, β-amyrin, and lupeol at the C-28 position to ursolic acid, oleanolic acid, and betulinic acid, respectively [38, 48], several other members of this CYP716 family have also been characterized and found to play complex roles in different plants. In the context of pentacyclic triterpenoid synthesis, CYP716 family enzymes have been shown to exhibit oxidation activity at the C-28, C-22α, C-3, and C-16β positions, respectively (Fig. 7). In dammarane-type triterpenoid synthesis, CYP716 family enzymes also exhibit catalytic activity at the C-12 and C-6 positions (Fig. 7). In *A. elata,* the primary saponins are oleanane-type pentacyclic triterpenoids, which are catalyzed by CYP450 genes from β-amyrin at the C-28 position [19]. This CYP716A subfamily is closely associated with oleanane-type triterpene biosynthesis in several plant species [31, 49]. In the present study, we were able to identify three CYP716A family members (CYP716A295, CYP716A296, and CYP716A306), and we further found that CYP716A295 and CYP716A296 exhibited expression patterns similar to that of β-amyrin synthase and their expression levels in different tissues were also consistent with oleanolic acid contents. We therefore postulated that CYP716A295 and CYP716A296 are the best candidate genes involved in the synthesis of oleanolic acid in *A. elata*.

We also detected significant levels of the hederagenin aglycone sapogenin in *A. elata*, with this compound being produced via the C-23 oxidation of oleanolic acid (Fig. 1A). Members of the CYP72A subfamily have been shown to be involved in a variety of sapogenin biosynthesis reactions, with four *M. truncatula* CYP450 genes (CYP72A63, CYP72A61v2, CYP72A67, and CYP72A68v2), one *Glycyrrhiza uralensis* CYP450 gene (CYP72A154), and one *K. septemlobus* gene (CYP72A397) having been shown to be involved in sapogenin biosynthesis [40, 41, 50, 51]. The CYP72A68v2 enzyme in *M. truncatula* has been shown to catalyze the oleanolic acid to gypsogenic acid conversion via intermediate hederagenin formation, while CYP72A397 in *K. septemlobus* produces hederagenin as a single compound [40, 41]. These CYP450s were proposed to have hydroxylation activity at the C-23 position of the enzyme on the oleanolic acid substrate. A BLASTx analysis suggested that the *A. elata* CYP72A760-CYP72A763 amino acid sequences shared >50% sequence identity with *K. septemlobus* CYP72A397. Given that its expression pattern aligned well with the tissue-specific distribution of hederagenin in *A. elata*, we therefore identified CYP72A763 as a candidate gene involved in hederagenin biosynthesis.

Conclusion

In this study, leaf, root, and stem transcriptomes from *A. elata* were sequenced for the first time. The resultant large dataset of transcripts and unigenes provided a robust genetic basis for discovering important genes and secondary metabolic pathways in these plants. Based on this transcriptomic data and available databases, two pathways and 19 putative genes related to triterpenoid saponin biosynthesis were discovered. We systematically identified CYP450 superfamily genes, identifying 111 full-length CYP450s for the first time in *A.elata*. Analyses of CYP450 genes with respect to their phylogeny, conserved motifs, gene structures, gene functions, and expression patterns in different tissues were further conducted based on bioinformatics and qRT-PCR methods. Finally, three candidate CYP450s related to triterpenoid saponin biosynthesis were identified and their subcellular localization was analyzed. Together, this study provides comprehensive insight into the CYP450 gene family in *A. elata* and will aid in determining CYP450 gene functions in this and related species.

Methods

Plant materials

A. elata cultivar (Plant Materials No. CH02-1-03 in Heilongjiang Crop Committee) was used in this study, which was provided by Prof. Hengtian Zhao (Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences), was grown in the Wild Plant Germplasm Resources Nursery of Northeast Agricultural University (Harbin, Heilongjiang, China; 45°44’21”N, 126°43’22”E). Two-year-old plants were used for this study, with samples of roots, leaves, and stems from three biological replicates of these plants being collected, snap frozen, and stored at −80 °C.

Saponin content quantification

A slightly modified version of the vanillin-glacial acetic colorimetric approach designed by Huang et al. [52] was used to quantify saponin contents in *A. elata* samples, with oleanolic acid serving as an analytical standard. Briefly, we ground 200 mg of each freeze-dried tissue samples into a fine powder, after which a slightly modified ultrasonic-assisted method [53] was used to achieve total saponin extraction from these samples. Briefly, each sample was suspended using a 6 mL volume of 80% ethanol, and the extraction procedure was allowed to proceed for 1 h at 25 °C. Extracts were then filtered before being diluted in a 10 mL volume of 80% ethanol. Next, 100 uL of the filtrate was collected and evaporated in a 70 °C water bath until dry, at which time 400 uL 5% vanillin-glacial acetic acid and 1.6 mL perchloric acid were added to the sample, which was then heated for 15 minutes in a 60 °C water bath, after which it was to 25 °C before 8 mL ethyl acetate was added. Samples were then mixed thoroughly, and absorbance at 560 nm (A560) was quantified via microplate reader (Biotek Elx800, USA). The total saponin content of samples was calculated using the regression equation, Y=5.31X-0.036 (R^2=0.9995), with Y indicating the A560 and X for corresponding to the amount of oleanolic acid (μg).

Ultra-performance liquid chromatography–quadrupole time-of-flight–mass spectrometry (UPLC–QTOF–MS) was used to identify two main sapogenin and three selected araloside monomers which were isolated before in *A. elata,* with their retention times and MS data being compared to those of standards in order to facilitate their identification. For this analysis, a 10 mL volume of 80% methanol was used to ultrasonically extract 100 mg of each freeze-dried tissue samples at 22°C for 60 min, followed by extract filtration via 0.22 μm Econofilter. A Waters I Class UPLC–QTOF mass spectrometer (Waters, MA, USA) was used for UPLC-QTOF-MS, with a UPLC C18 analytical column (100 mm×2.1 mm, ACQUITY UPLC BEH) being used for separation at 40 °C. For this separation, the mobile phase was composed of (A) 0.1% formic acid in water and (B) acetonitrile. The linear gradient conditions were as follows [54]: 0-5.0 min, 5-95% B; 5-11 min, 95% B; 11-12 min, 95-5% B; 12-15 min, 5% B. For each sample, a 5 μl injection volume was used, with a 0.4 mL/min flow rate. The mass spectrometer conducted a full scan in a negative ion mode. N_2 was used as the desolvation gas. The scanning ranges of the primary mass spectrometer and secondary mass spectrometer were 100-1200 m/z and 50-1200 m/z, respectively. Data analysis was performed using the Peakview 2.0/Masterview 1.0 software (AB SCEIX, USA). Five compounds were separated well in 5 min (Additional file 1: Figure S1) and they were identified based on the library of the Peakview 2.0/Masterview 1.0 software containing information pertaining to the molecular formula and retention time (RT) (Additional file 8: Table S4). For quantitative analysis, each compound was identified repeatedly (n = 3), and the height of peaks was used to measure the intensity. Next, standard curves for five standards were prepared and used to calculate saponin contents based on the regression equation (Additional file 8: Table S4).

RNA-sequencing

Roughly 100 mg of frozen tissue was then used for total RNA extraction with an OmniPlant RNA Kit based upon provided directions. For RNA-seq analyses, NEBNext Oligo(dT)25 beads (NEB, USA) were used to specifically enrich for the mRNA present within a 50 μl total RNA sample, after which a NEBNext Ultra RNA Library Prep Kit for Illumina (NEB) was used to prepare an mRNA library from this enriched

samples according to provided directions. An Illumina HiSeq^{TM} 4000 platform was then used for sequencing. The resultant raw reads then underwent quality filtering in order to remove those reads that were of low-quality, contained poly-N sequences, or contained adapter sequences. Trinity was then used for de novo assembly of clean reads [55], yielding a transcriptomic reference database.

Functional annotation and pathway analyses

A BLASTx analysis that compared the identified putative unigenes from our transcriptomic database to the nonredundant protein (Nr) database of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov), the Swiss-Prot protein database (http://www.expasy.ch/sprot), and the Clusters of Orthologous Groups (COG)/EuKaryotic Orthologous Groups (KOG) databases (http://www.ncbi.nlm.nih.gov/COG) was next conducted. Furthermore, each unigene was assigned to defined KEGG pathways according to its similarity to genes within the KEGG database (http://www.genome.jp/kegg) as determined via BLAST search, with $1e^{-5}$ as the cut-off value. The output of this pathway analysis yielded both enzyme commission (EC) and KEGG orthology (KO) numbers.

Full-length A. elata CYP450 genes identification and classification

A hidden Markov model (HMM) was retrieved from the Pfam database (http://pfam.sanger.ac.uk) and used for CYP450 family member identification, with HMMER being used to search the *A. elata* deduced amino acid database for the P450.hmm (PF00067) sequence. Those unigenes identified via this initial analysis were then subjected to additional validation with the Simple Modular Architecture Research Tool (SMART; http://smart.embl-heidelberg.de), and open reading frames (ORFs) for these genes were identified using the ORF Finder software (http://bioinf.ibun.unal.edu.co/servicios/sms/orf_find.html). As whole-genome sequencing data for these plants was unavailable, we instead utilized the strict CYP450 gene criteria previously outlined by Chen et al. [9] were used. First, the proteins that started with the amino acid “M” and that terminated before a position corresponding to a stop codon were selected. Next, the amino acid sequence before the predicted starting amino acid “M” was found to not be conserved compared to the corresponding regions of the homologous sequences (http://www.herbalgenomics.org/samicyp450/msa/index_msa.html). In addition, a BLAST search for previously published CYP450 genes, including 484 CYP450 family members encoded by the *A. elata* homolog *P. ginseng,* was also conducted in an effort to facilitate more comprehensive CYP450 gene identification. The names of these CYP450s were defined by Prof. David Nelson based on reference sequences contained within a thoroughly-annotated CYP450 reference database [35]. ExPASy (http://www.expasy.org/tools/) was further used to assess the physicochemical properties of these putative CYP450s, while the Cell-PLoc 2.0 software (http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc/) was used to assess their potential subcellular localization.

CYP450 conserved motif identification and phylogenetic analysis

The online MEME program (http://meme.nbcr.net/meme/intro.html) was used to identify conserved motifs within putative CYP450s, with the motif number being set to ‘10’ and all other parameters being

set to their default values.

Of the 111 identified CYP450 sequences in *A. elata*, 59 were selected to serve as representative sequences for use in phylogenetic analyses, which were conducted using the ClustalX 2.1 software [56]. A neighbor-joining algorithm with a Poisson model and pairwise deletion was used to generate a phylogenetic tree with the MEGAX software [57], with 1,000 replicates being used for bootstrap testing to validate this tree. EvolView (http://www.evolgenius.info/evolview/) was used for modification of the bootstrap consensus tree, which was exported in the Newick format file [58].

Secondary structural element assignment

We selected two *A. elata* CYP450 protein sequences from each of seven clans (the CYP51 and CYP711 clans contained only one member), using this resultant CYP450 set as a representative sample for secondary structural element assignments. As a template for these assignments, *Arabidopsis* CYP90B1 (PDB ID: 6A15) was used, with its crystal structure being downloaded from the RCSB data bank (http://www.rcsb.org/). ClustalX 2.1 was used to conduct multiple sequence alignments of CYP90B1 with the representative *A. elata* CYP450s, and ESPript3.0 (http://espript.ibcp.fr/ESPript/ESPript/) was used to assign secondary structural elements to these aligned sequences. Gotoh's SRSs were mapped onto the aligned *A. elata* P450s based on alignment with the P450cam sequence (P00183).

Assessment of gene expression patterns

The reads per kb per million mapped reads (RPKM) method was used to quantify CYP450 gene expression in the root, stem, and leaf tissues from *A. elata* in this study. TBtools (Toolbox for Biologist, v0.6652) was used for hierarchical clustering analyses. In addition, qRT-PCR was used to validate the RNA-seq results for 12 randomly selected CYP450s from across the 7 identified clades as follows:

An RNAprep Pure Plant Kit (TianGen, Beijing) was used to isolate RNA from plant tissue samples, after which a ReverTra Ace qPCR RT Master Mix with gDNA Remover (TOYOBO) was used to conduct first-strand cDNA synthesis. A qTOWER real-time PCR system (Analytik Jena, Germany) was then used for qRT-PCR analyses, together with the THUNDERBIRD SYBR qPCR Mix (TOYOBO). As normalization control, *A. elata GAPDH* was also measured. Thermocycler settings were as follows: 95 °C for 30 s; 40 cycles of 95 °C for 10 s, 55 °C for 10 s, and 72 °C for 15 s. Three biological replicates per sample were analyzed, and the $2^{-\triangle\triangle CT}$ method was used to quantify gene expression results. Primers used in this study are compiled in Additional file 9Table S5.

Subcellular localization analysis

We selected the CYP716A295, CYP716A296, and CYP72A763 genes to assess representative CYP450 subcellular localization by PCR-amplifying the ORFs for these genes without a stop coding using specific primers with corresponding enzyme sites (Additional file 6: Table S3). Sangon Biotech (Shanghai, China) then conducted sequence validation of the isolated PCR products, after which they were inserted

upstream of enhanced green fluorescent protein (GFP) at appropriate restriction enzyme digestion site in the pAN580-35S-GFP vector, yielding pAN580-35S-CYP450::GFP vectors. These recombinant plasmids were transformed into *Arabidopsis* protoplasts along with the GHD7-RFP plasmid using a polyethylene glycol (PEG)-mediated transient transformation system [59]. Protoplasts expressing the resultant GFP fusion proteins were then visualized via Airyscan confocal laser scanning microscope (ZEISS710, Carl Zeiss, Jena, Germany).

Abbreviations

A.elata：*Aralia elata* (Miq.) Seem; RNA-seq: RNA-sequencing; CYP450：Cytochrome P450; qRT-PCR: Quantiative real-time reverse transcription PCR; SRSs: Substrate recognition sites; ER: Endoplasmic reticulum; IPP: Isopentenyl diphosphate; DMAPP: Dimethylallyl pyrophosphate; MVA: Mevalonate; MEP: Methylerythritol 4-phosphate; OSC: Oxidosqualene cyclase; UGTs: UDP-glycosyltransferases; AACT: Acetyl-CoA acetyltransferase; HMGS: Hydroxymethyl glutaryl CoA synthase; HMGR: Hydroxymethyl gutryl CoA reductase; MVK: Mevalonate kinase; PMK: Phosphomevalonate kinase; MVD: Diphosphosphate decarboxylase; IDI: Isopentenyl pyrophosphate; DXS:1-deoxy-D-xylulose-5-phosphate synthase; DXR: 1-deoxy-D-xylulose-5-phosphate reductoisomerase; MEP-CT: 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase; CDP-MEK: 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MECDPS: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HMBPPS: (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase; HMBPPR: 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase; GPS: Geranyl diphosphate synthase; FPS: Farnesyl diphosphate synthase; SS: Squalene synthase; SE: Squalene monooxygenase; bAS: β-amyrin synthase.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All RNA-seq reads generated by this study are publicly available at the NCBI Short Read Archive (SRA) under accession numbers PRJNA555256.

Competing interests

We declare that we have no conflict of interest.

Funding

This work was supported by the National Key Research and Development Program of China (2016YFC0500307-06). The funders had no rule in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

YC, XY and XJ conceived and designed the experiments. YC, XT, HL and DL performed the experiments. YC and HL performed the data analysis and wrote the manuscript. All authors read and approved the final manuscript. It's worth noting that YC and HL are co-first authors.

Acknowledgments

We thank Prof. Hengtian Zhao for providing plant material, we also thank Dr. David R. Nelson (University of Tennessee) for the naming of P450s.

Author details

[1] College of Horticulture and Landscape Architecture, Northeast Agricultural University, Harbin, Heilongjiang 150030, China.

[2] Key Laboratory of Biology and Genetic Improvement of Horticulture Crops (Northeast Region), Ministry of Agriculture, Northeast Agricultural University, Harbin, Heilongjiang 150030, China.

References

1. Nelson DR. Cytochrome P450 and the individuality of species. Arch Biochem Biophys 1999; 369(1):1-10. doi:10.1006/abbi.1999.1352.
2. Morant M, Bak S, Møller BL, Werck-Reichhart D. Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. Curr Opin Biotechnol 2003; 14(2):151-62. doi:10.1016/s0958-1669(03)00024-7.
3. Paquette SM, Bak S, Feyereisen R. Intron–exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. DNA Cell Biol 2000; 19(5):307-17. doi:10.1089/10445490050021221.
4. Augustin JM, Kuzina V, Andersen SB, Bak S. Molecular activities, biosynthesis and evolution of triterpenoid saponins. Phytochem 2011; 72(6):435-57. doi:10.1016/j.phytochem.2011.01.015.
5. Nelson DR, Schuler MA, Paquette SM, Werck-Reichhart D, Bak S. Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. Plant Physi 2004; 135(2):756-72
6. Li L, Cheng H, Gai J, Yu D. Genome-wide identification and characterization of putative cytochrome P450 genes in the model legume *Medicago truncatula*. Planta 2007; 226(1):109-23. doi:10.1007/s00425-006-0473-z.

7. Wei K, Chen H. Global identification, structural analysis and expression characterization of cytochrome P450 monooxygenase superfamily in rice. BMC Genomics 2018; 19(1). doi:10.1186/s12864-017-4425-8.
8. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet 2014; 30(9):418-26. doi:10.1016/j.tig.2014.07.001.
9. Chen H, Wu B, Nelson DR, Wu K, Liu C. Computational identification and systematic classification of novel cytochrome P450 genes in *Salvia miltiorrhiza*. PLoS ONE 2014; 9(12):e115149. doi:10.1371/journal.pone.0115149.
10. Qi X, Yu X, Xu D, Fang H, Dong K, Li W et al. Identification and analysis of CYP450 genes from transcriptome of *Lonicera japonica* and expression analysis of chlorogenic acid biosynthesis related CYP450s. PeerJ 2017; 5:e3781. doi:10.7717/peerj.3781.
11. Liao W, Zhao S, Zhang M, Dong K, Chen Y, Fu C et al. Transcriptome assembly and systematic identification of novel cytochrome P450s in *Taxus chinensis*. Front Plant Sci 2017; 8. doi:10.3389/fpls.2017.01468.
12. Schuler MA. Plant cytochrome P450 monooxygenases. Crit Rev Plant Sci 1996; 15(3):235-84
13. Chehab E, Raman G, Walley J, Perea J, Banu G, Theg S et al. Rice HYDROPEROXIDE LYASES with unique expression patterns generate distinct aldehyde signatures in Arabidopsis. Plant physiol 2006; 141(1):121-34
14. Quinlan RF, Shumskaya M, Bradbury LM, Beltrán J, Ma C, Kennelly EJ et al. Synergistic interactions between carotene ring hydroxylases drive lutein formation in plant carotenoid biosynthesis. Plant physiol 2012; 160(1):204-14
15. Fujiyama K, Hino T, Kanadani M, Watanabe B, Jae Lee H, Mizutani M et al. Structural insights into a key step of brassinosteroid biosynthesis and its inhibition. Nat Plants 2019; 5(6):589-94. doi:10.1038/s41477-019-0436-6.
16. Hasemann CA, Ravichandran KG, Peterson JA, Deisenhofer J. Crystal structure and refinement of cytochrome P450terp at 2·3 Å resolution. J Mol Biol 1994; 236(4):1169-85. doi:10.1016/0022-2836(94)90019-1.
17. Wang M, Xu X, Xu H, Wen F, Zhang X, Sun H et al. Effect of the total saponins of *Aralia elata* (Miq) Seem on cardiac contractile function and intracellular calcium cycling regulation. J Ethnopharm 2014; 155(1):240-7. doi:10.1016/j.jep.2014.05.024.
18. Zhang M, Liu G, Tang S, Song S, Yamashita K, Manabe M et al. Effect of five triterpenoid compounds from the buds of *Aralia elata* on stimulus-induced superoxide generation, tyrosyl phosphorylation and translocation of cytosolic compounds to the cell membrane in human neutrophils. Planta Med 2006; 72(13):1216-22. doi:10.1055/s-2006-951679.
19. Zhang Y, Wang W, He H, Song X-y, Yao G-d, Song S-j. Triterpene saponins with neuroprotective effects from a wild vegetable *Aralia elata*. J Funct Foods 2018; 45:313-20. doi:10.1016/j.jff.2018.04.026.
20. Xi S, Zhou G, Zhang X, Zhang W, Cai L, Zhao C. Protective effect of total aralosides of *Aralia elata* (Miq) Seem (TASAES) against diabetic cardiomyopathy in rats during the early stage, and possible

mechanisms. Exp Mol Med 2009; 41(8):538. doi:10.3858/emm.2009.41.8.059.

21. Hwang K-A, Hwang Y-J, Kim GR, Choe J-S. Extracts from *Aralia elata* (Miq) Seem alleviate hepatosteatosis via improving hepatic insulin sensitivity. BMC Complement Altern Med 2015; 15(1). doi:10.1186/s12906-015-0871-5.
22. Chen R-C, Wang J, Yu Y-L, Sun G-B, Sun X-B. Protective effect of total saponins of *Aralia elata* (Miq) Seem on lipopolysaccharide-induced cardiac dysfunction via down-regulation of inflammatory signaling in mice. RSC Adv 2015; 5(29):22560-9. doi:10.1039/c4ra16353b.
23. Saito S, Sumita S, Tamura N, Nagamura Y, Nishida K, Ito M et al. Saponins from the leaves of *Aralia elata* Seem. (Araliaceae). Chem Pharm Bull 1990; 38(2):411-4. doi:10.1248/cpb.38.411.
24. Kuang H-X, Sun H, Zhang N, Okada Y, Okuyama T. Two New Saponins, Congmuyenosides A and B, from the Leaves of *Aralia elata* Collected in Heilongjiang, China. Chem Pharm Bull 1996; 44(11):2183-5. doi:10.1248/cpb.44.2183.
25. Kang SS, Kim JS, Kim OK, Lee EB. Triterpenoid saponins from the root barks of *Aralia elata*. Arch Pharmacal Res 1993; 16(2):104-8. doi:10.1007/bf03036855.
26. Sakai S, Katsumata M, Satoh Y, Nagasao M, Miyakoshi M, Ida Y et al. Oleanolic acid saponins from root bark of *Aralia elata*. Phytochem 1994; 35(5):1319-24. doi:10.1016/s0031-9422(00)94846-5.
27. Sawai S, Saito K. Triterpenoid biosynthesis and engineering in plants. Front Plant Sci 2011; 2:25
28. Wen L, Yun X, Zheng X, Xu H, Zhan R, Chen W et al. Transcriptomic comparison reveals candidate genes for triterpenoid biosynthesis in two closely related Ilex species. Front Plant Sci 2017; 8:634
29. Cordoba E, Porta H, Arroyo A, San Román C, Medina L, Rodríguez-Concepción M et al. Functional characterization of the three genes encoding 1-deoxy-D-xylulose 5-phosphate synthase in maize. J Exp Bot 2011; 62(6):2023-38
30. Aharoni A, Jongsma MA, Kim T-Y, Ri M-B, Giri AP, Verstappen FWA et al. Metabolic engineering of terpenoid biosynthesis in plants. Phytochem Rev 2006; 5(1):49-58. doi:10.1007/s11101-005-3747-3.
31. Seki H, Tamura K, Muranaka T. P450s and UGTs: Key players in the structural diversity of triterpenoid saponins. Plant Cell Physiol 2015; 56(8):1463-71. doi:10.1093/pcp/pcv062.
32. Tamura K, Teranishi Y, Ueda S, Suzuki H, Kawano N, Yoshimatsu K et al. Cytochrome P450 monooxygenase CYP716A141 is a unique β-amyrin C-16β oxidase involved in triterpenoid saponin biosynthesis in *Platycodon grandiflorus*. Plant Cell Physiol 2017; 58(6):1119-. doi:10.1093/pcp/pcx067.
33. Augustin JM, Drok S, Shinoda T, Sanmiya K, Nielsen JK, Khakimov B et al. UDP-glycosyltransferases from the UGT73C subfamily in *Barbarea vulgaris* catalyze sapogenin 3-O-glucosylation in saponin-mediated insect resistance. Plant physiol 2012; 160(4):1881-95
34. Sun H, Li F, Xu Z, Sun M, Cong H, Qiao F et al. De novo leaf and root transcriptome analysis to identify putative genes involved in triterpenoid saponins biosynthesis in Hedera helix L. PLoS ONE 2017; 12(8):e0182243. doi:10.1371/journal.pone.0182243.
35. Nelson DR. The cytochrome p450 homepage. Human Genomics 2009; 4(1):59-65

36. Gotoh O. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. J Biol Chem 1992; 267(1):83-90
37. Yasumoto S, Seki H, Shimizu Y, Fukushima EO, Muranaka T. Functional characterization of CYP716 family P450 enzymes in triterpenoid biosynthesis in tomato. Front Plant Sci 2017; 8. doi:10.3389/fpls.2017.00021.
38. Fukushima EO, Seki H, Ohyama K, Ono E, Umemoto N, Mizutani M et al. CYP716A subfamily members are multifunctional oxidases in triterpenoid biosynthesis. Plant Cell Physiol 2011; 52(12):2050-61. doi:10.1093/pcp/pcr146.
39. Han J-Y, Hwang H-S, Choi S-W, Kim H-J, Choi Y-E. Cytochrome P450 CYP716A53v2 catalyzes the formation of protopanaxatriol from protopanaxadiol during ginsenoside biosynthesis in *Panax Ginseng*. Plant Cell Physiol 2012; 53(9):1535-45. doi:10.1093/pcp/pcs106.
40. Han JY, Chun J-H, Oh SA, Park S-B, Hwang H-S, Lee H et al. Transcriptomic analysis of *Kalopanax septemlobus* and characterization of KsBAS, CYP716A94 and CYP72A397 genes involved in hederagenin saponin biosynthesis. Plant Cell Physiol 2017; 59(2):319-30. doi:10.1093/pcp/pcx188.
41. Fukushima EO, Seki H, Sawai S, Suzuki M, Ohyama K, Saito K et al. Combinatorial biosynthesis of legume natural and rare triterpenoids in engineered yeast. Plant Cell Physiol 2013; 54(5):740-9. doi:10.1093/pcp/pct015.
42. Zhang X, Allan A, Li C, Wang Y, Yao Q. De Novo assembly and characterization of the transcriptome of the Chinese medicinal herb, *Gentiana rigescens*. Int J Mol Sci 2015; 16(12):11550-73. doi:10.3390/ijms160511550.
43. Liu M-H, Yang B-R, Cheung W-F, Yang KY, Zhou H-F, Kwok JS-L et al. Transcriptome analysis of leaves, roots and flowers of *Panax notoginseng* identifies genes involved in ginsenoside and alkaloid biosynthesis. BMC Genomics 2015; 16(1). doi:10.1186/s12864-015-1477-5.
44. Zhang J, Li X, Lu F, Wang S, An Y, Su X et al. De novo sequencing and transcriptome analysis reveal key genes regulating steroid metabolism in leaves, roots, adventitious roots and calli of *Periploca sepium* Bunge. Front Plant Sci 2017; 8. doi:10.3389/fpls.2017.00594.
45. Devi K, Mishra SK, Sahu J, Panda D, Modi MK, Sen P. Genome wide transcriptome profiling reveals differential gene expression in secondary metabolite pathway of Cymbopogon winterianus. Sci Rep 2016; 6(1). doi:10.1038/srep21026.
46. Rasool S, Mohamed R. Plant cytochrome P450s: nomenclature and involvement in natural product biosynthesis. Protoplasma 2015; 253(5):1197-209. doi:10.1007/s00709-015-0884-4.
47. Morant M, Jørgensen K, Schaller H, Pinot F, Møller BL, Werck-Reichhart D et al. CYP703 is an ancient cytochrome P450 in land plants catalyzing in-chain hydroxylation of lauric acid to provide building blocks for sporopollenin synthesis in pollen. Plant Cell 2007; 19(5):1473-87. doi:10.1105/tpc.106.045948.
48. Carelli M, Biazzi E, Panara F, Tava A, Scaramelli L, Porceddu A et al. *Medicago truncatula* CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. Plant Cell 2011;

23(8):3070-81. doi:10.1105/tpc.111.087312.

49. Miettinen K, Pollier J, Buyst D, Arendt P, Csuk R, Sommerwerk S et al. The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis. Nat commu 2017; 8:14153
50. Seki H, Sawai S, Ohyama K, Mizutani M, Ohnishi T, Sudo H et al. Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. Plant Cell 2011; 23(11):4112-23. doi:10.1105/tpc.110.082685.
51. Biazzi E, Carelli M, Tava A, Abbruscato P, Losini I, Avato P et al. CYP72A67 catalyzes a key oxidative step in *Medicago truncatula* hemolytic saponin biosynthesis. Mol Plant 2015; 8(10):1493-506. doi:10.1016/j.molp.2015.06.003.
52. Huang F, Zhao H, Zhou K, Li F, Zhang K. Study on distribution characteristics of the total aralosides content in *Aralia elata* (Miq.) Seem. Chinese Wild Plant Resources 2014; 33:1-8
53. Ma N, Gao M-j, Cui X-m, Chen Z-j. Studies on ultrasonic extracting saponins of *Panax notoginseng*. LiShiZhen Med Materia Medica Res 2005; 16:854-5
54. Song H-H, Kim D-Y, Woo S, Lee H-K, Oh S-R. An approach for simultaneous determination for geographical origins of Korean *Panax ginseng* by UPLC-QTOF/MS coupled with OPLS-DA models. J Ginseng Res 2013; 37(3):341
55. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013; 8(8):1494-512. doi:10.1038/nprot.2013.084.
56. Mahé S, Duhamel M, Le Calvez T, Guillot L, Sarbu L, Bretaudeau A et al. PHYMYCO-DB: A curated database for analyses of fungal diversity and evolution. PLoS ONE 2012; 7(9):e43117. doi:10.1371/journal.pone.0043117.
57. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evo 2018; 35(6):1547-9. doi:10.1093/molbev/msy096.
58. Zhang H, Gao S, Lercher MJ, Hu S, Chen W-H. EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic Acids Res 2012; 40(W1):W569-W72. doi:10.1093/nar/gks576.
59. Yoo S-D, Cho Y-H, Sheen J. Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nat Protoc 2007; 2(7):1565-72. doi:10.1038/nprot.2007.199.

Tables

Table 1 The aralosides content in different tissues of *A. elata*

Aralosides contents	Roots	Stems	Leaves
Total saponins (mg/g)	37.27±0.70b	6.26±1.23c	52.74±3.40a
Oleanolic acid (ug/g)	44.70±5.23a	11.95±1.68b	8.57±1.94b
Hederagenin(ug/g)	9.05±0.34c	88.71±12.10b	349.97±24.52a
Chikusetsusaponin IV(ug/g)	573.89±12.96a	26.18±1.19b	10.181±0.95c
Araloside VII(ug/g)	ND	ND	97.54±15.08a
Araloside X(ug/g)	9.63±0.28b	9.09±0.49b	114.99±5.02a

ND not detected Values were mean ± SD. Different letters within a row indicated significant differences at $P < 0.05$

Table 2 Summary of the RNA-Seq analysis of *A.elata*

Total of raw reads	448,112,618
Total assembled bases	66,367,722,247
GC percentage	38.83
Number of contigs	82,238
Maximum length of contigs (bp)	16,016
Minimum length of contigs (bp)	201
Average length of contigs (bp)	1,058
N50 of contigs (bp)	1,846
Number of unigenes	66,713

Supplementary Information

Additional file 1: Figure S1. Typical ion current (TIC) chromatograms for aralosides in leaves (A), stems (B)and roots (C) of *A. elata* and of the mixed reference substance (D) as identified via UPLC–QTOF–MS. Peak number correspond to these different aralosides, including: araloside VII (1), araloside X (2), chikusetsusaponin IV (3), hederagenin (4) and oleanolic acid (5).

Additional file 2: Figure S2. Venn diagram indicating annotated genes by the KEGG, KOG, Nr and Swissprot databases. The number of genes annotated is listed in each diagram component.

Additional file 3: Table S1. Unigenes related to saponin skeleton biosynthesis obtained after three independent biological replicates along with their mean values.

Additional file 4: Table S2. List of full-length CYP450s of *A. elata* identified in this study.

Additional file 5: Figure S3. KEGG pathway analysis of predicted CYP450s in *A elata*.

Additional file 6: Figure S4. Phylogenetic relationships and distribution of conserved motifs in *A. elata* CYP450s.

Additional file 7: Table S3. A list of 36 previously reported plant CYP450s involved in triterpenoid biosynthesis.

Addtional file 8: Table S4. List of five standards analyzed by UPLC-QTOF-MS.

Additional file 9: Table S5. Sequences of the primers used in this study.

Figures

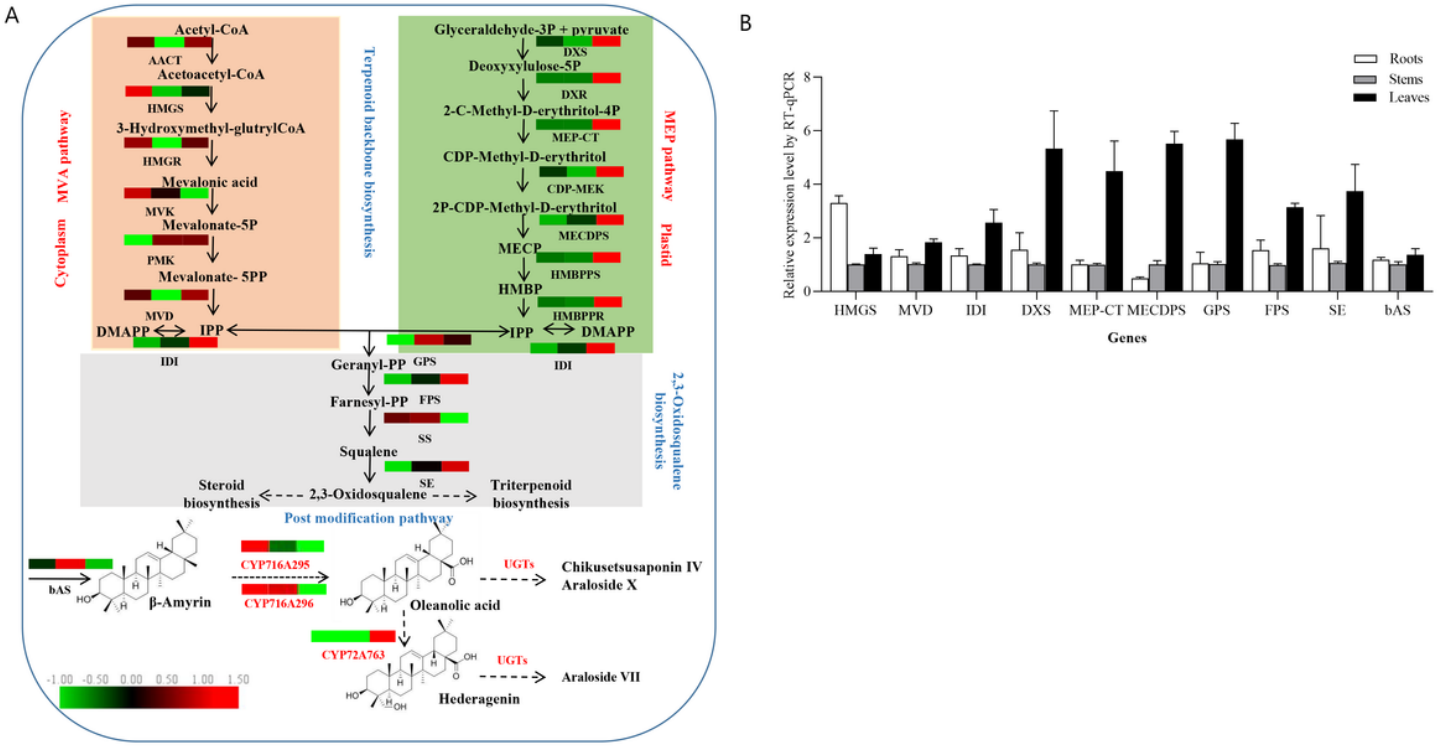

Figure 1

(A) Putative pathways of triterpenoid biosynthesis in A.elata, with the enzymes identified herein included in this diagram. The heatmap highlights the patterns of expression for these genes in the root, stem, and leaf tissues, with RPKM values used for normalization and color-coding conducted accordingly. Broken arrows indicate putative araloside biosynthesis steps that involve CYP450s and UGTs. (B) The selected genes putatively involved in triterpene saponin backbone biosynthesis were quantified via qRT-PCR, with the $2^{-\triangle\triangle CT}$ approach used to assess gene expression levels relative to those in stem tissues. GAPDH was used for normalization, and data are included with standard deviations.

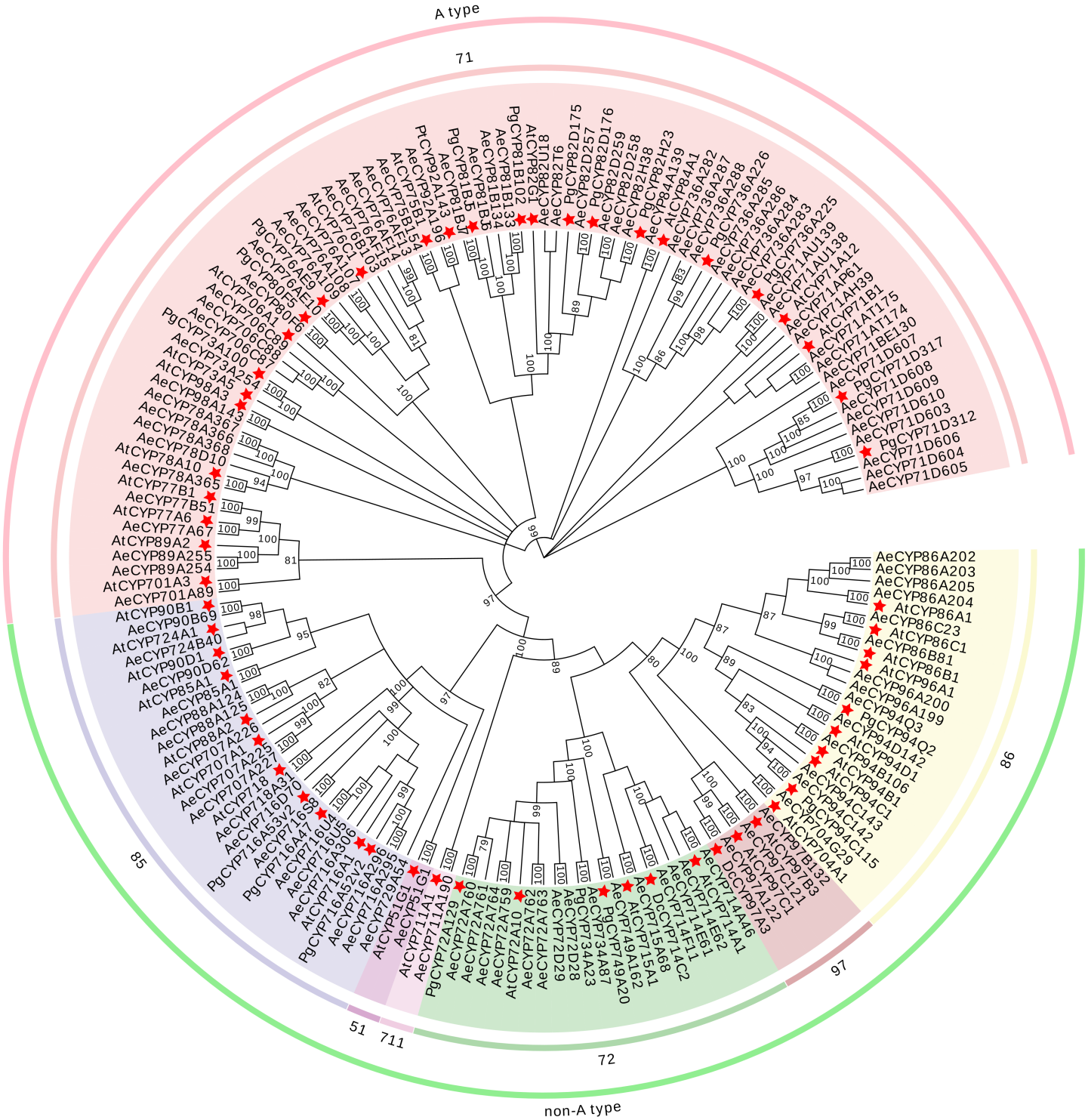

Figure 2

Phylogenetic analysis of predicted CYP450s in A. elata and representative CYP450 family members.

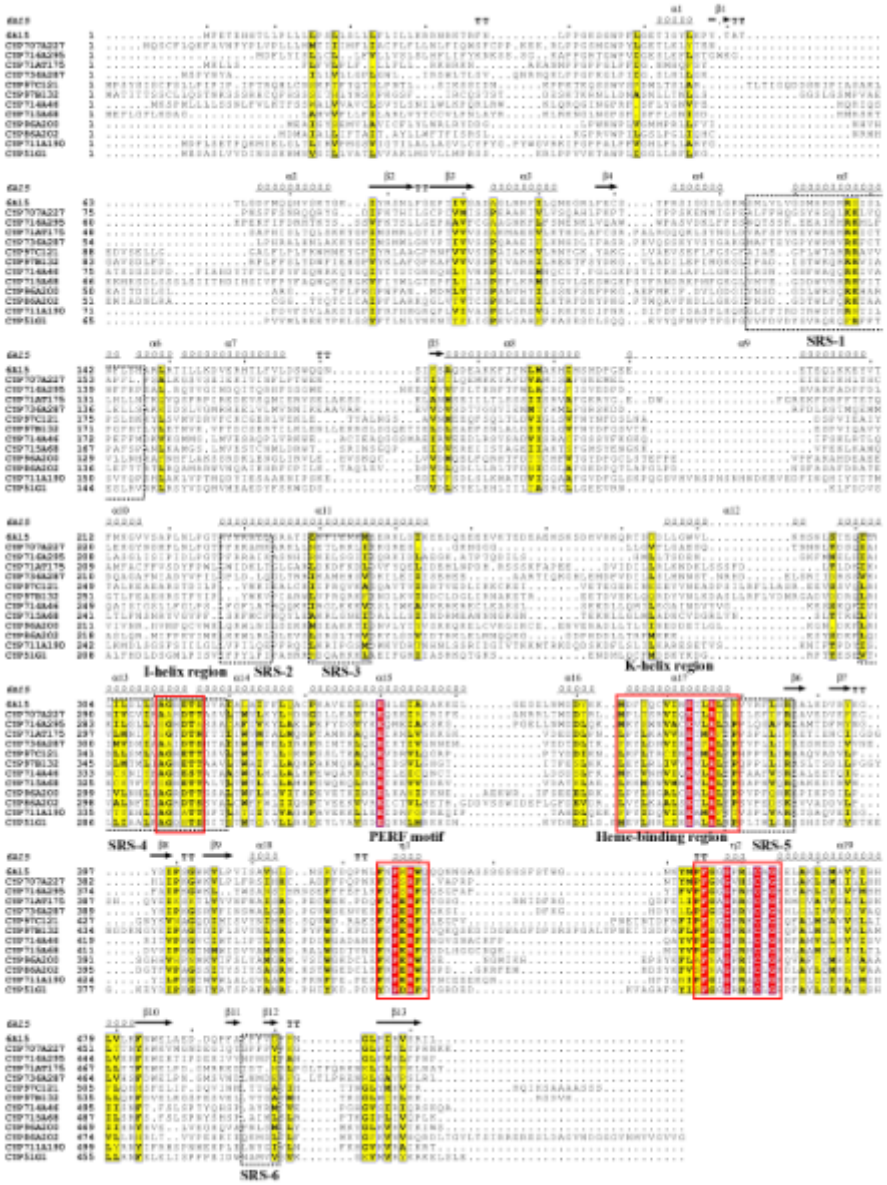

Figure 3

Alignment of representative CYP450s in A.elata with Arabidopsis CYP90B1. Secondary structural elements were assigned according to the CYP90B1 (6A15) structure. The four conserved motifs in these CYP450s are circles in red, while the black boxes enclose Gotoh's SRSs as identified via sequence alignment with P450cam.

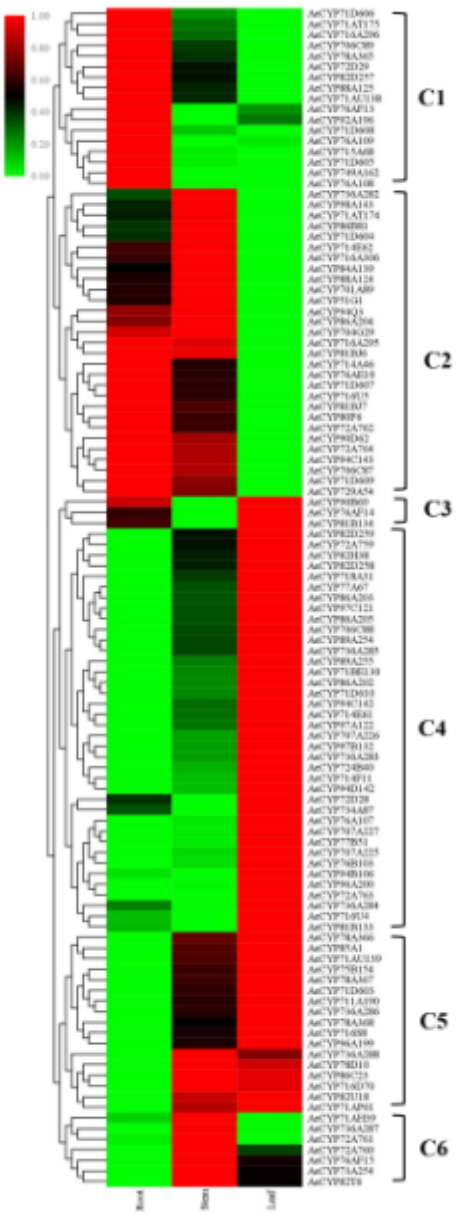

Figure 4

A.elata CYP450 expression profiles. Hierachical clustering for these 111 full-length CYP450s was conducted based upon RNA-seq results.

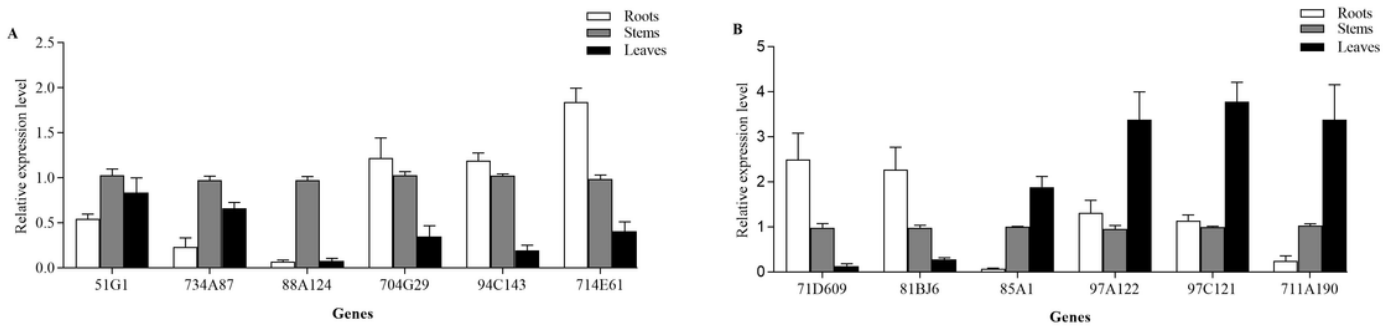

Figure 5

The expression levels of the indicated randomly chosen CYP450s from the RNA-seq analysis were validated via qRT-PCR.

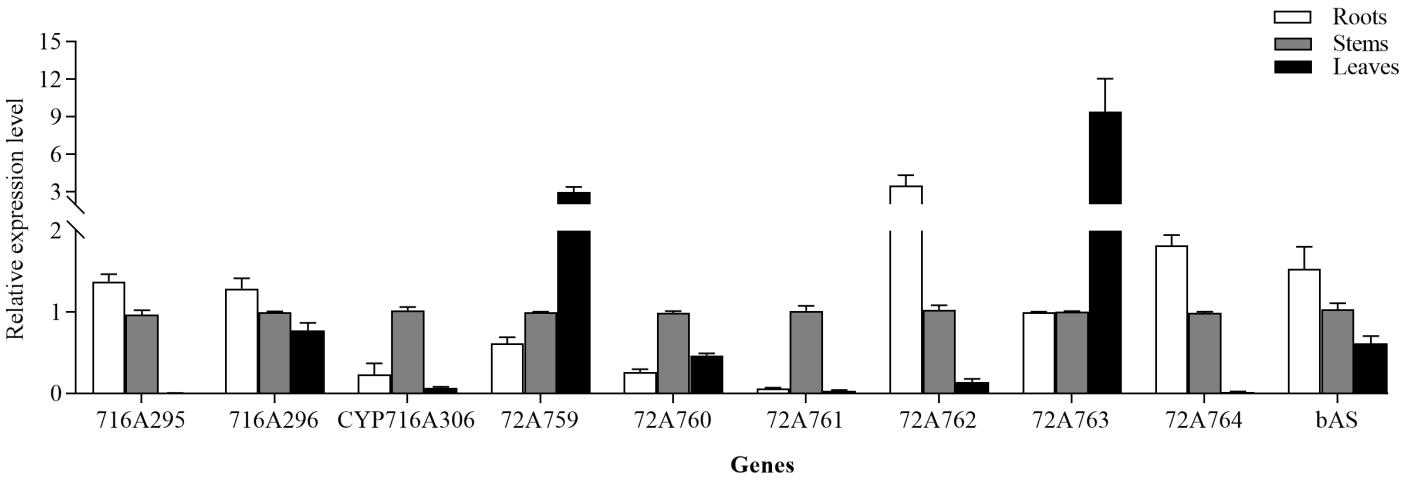

Figure 6

Comparison of the expression levels of 3 CYP716A and 6 CYP72A genes in different tissues.

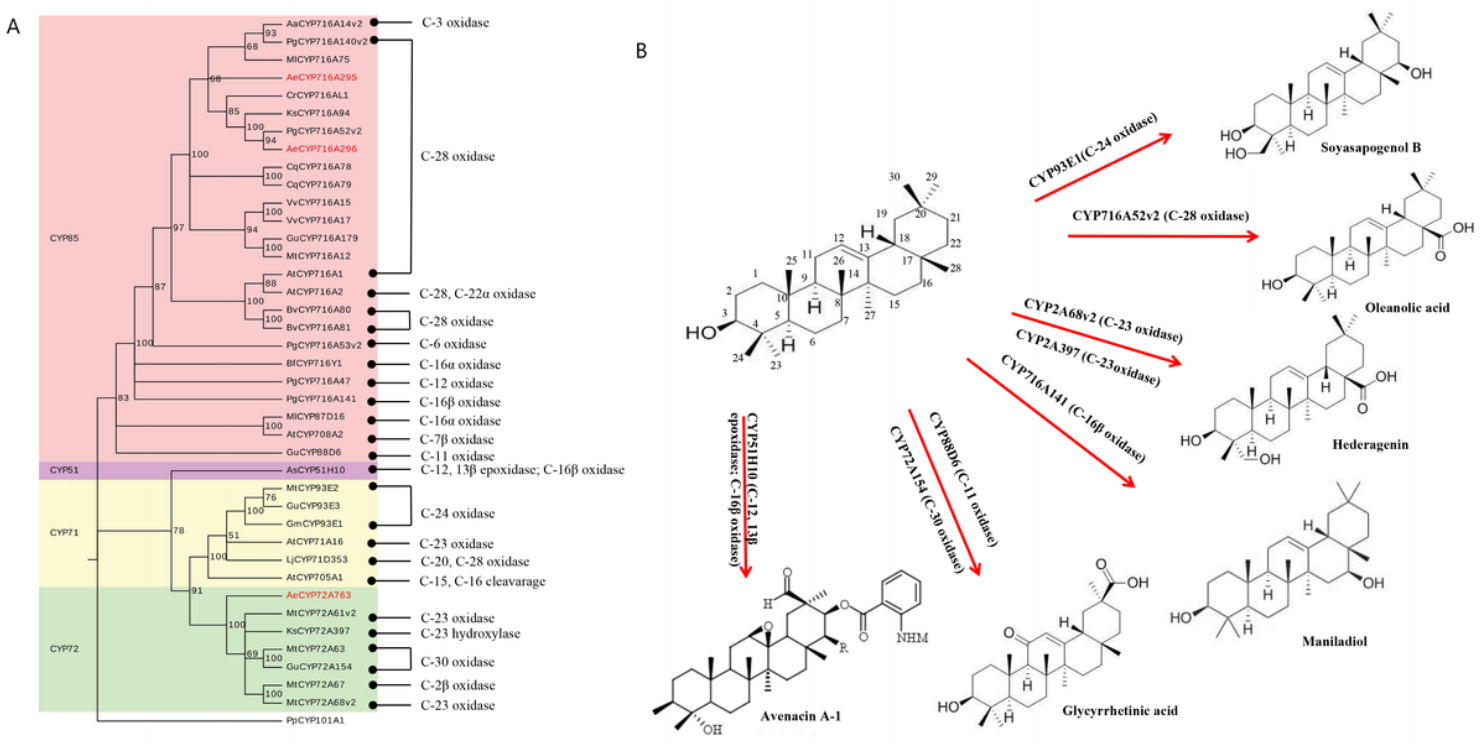

Figure 7

(A) A phylogenetic tree of previously characterized triterpenoid biosynthesis CYP450s and those A. elata CYP450s isolated in this study (in red). The known biochemical activities of these P450s are indicated on the right. CYP101A1 from Pseudomonas putida (accession No. 2L8M_A) was used as an outgroup in the

phylogenetic tree. (B) Representative CYP450s involved in the post-modification of β-amyrin at different positions.

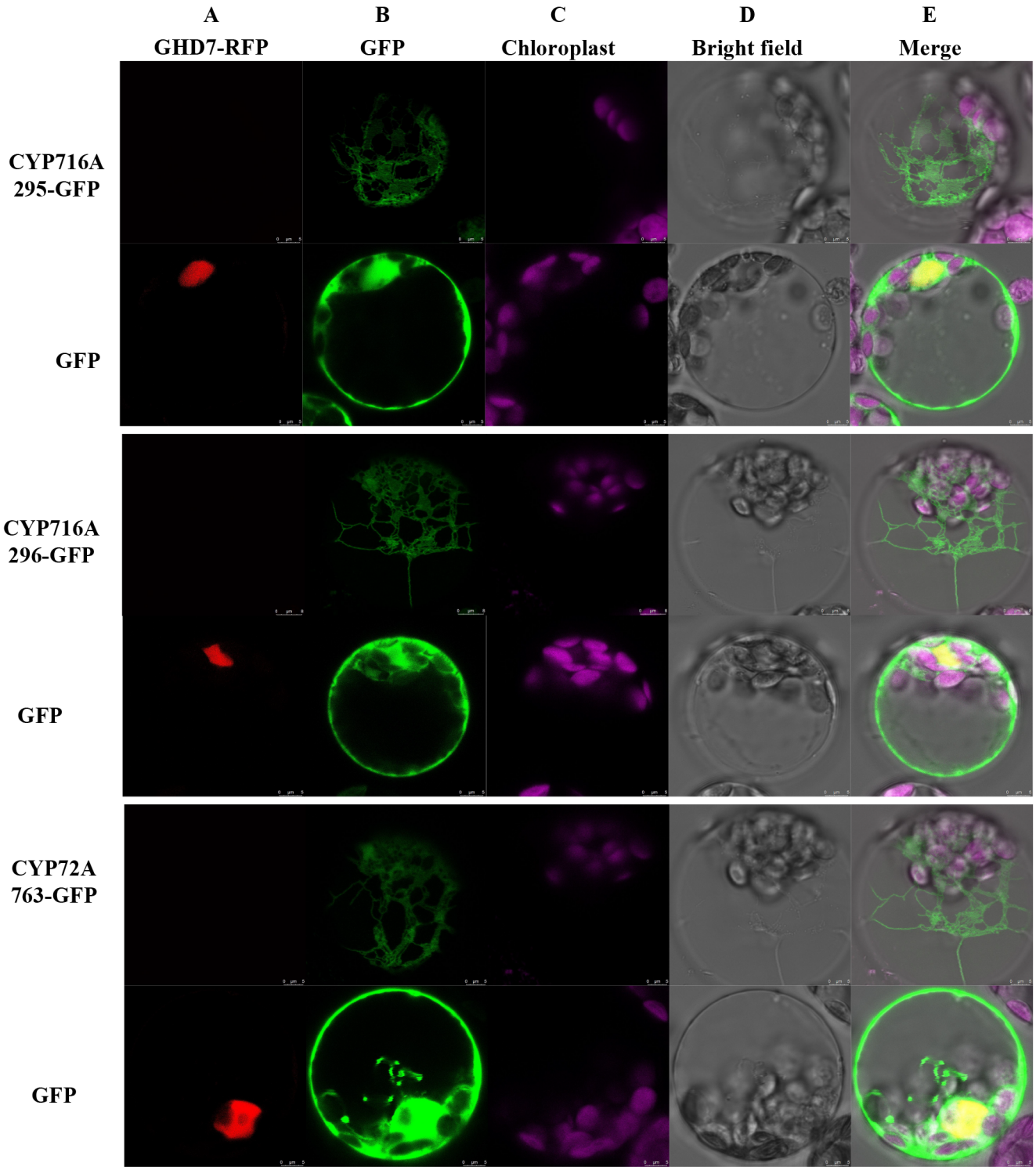

Figure 8

CYP716A295, CYP716A296, and CYP72A763 subcellular localization in Arabidopsis protoplasts co-transformed with CYP450CDS-YFP and the nuclear marker GHD7-RFP as analyzed via confocal microscopy. (A) The nuclei are marked by red fluorescence; (B) CYP450s are indicated by green; (C)

Chloroplasts are indicated by purple fluorescence; (D) Bright field illumination is shown in white; (E) A merged image of (A, B, C) indicates that these CYP450s are localized to the ER.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile4.xlsx
- Additionalfile6.tiff
- Additionalfile8.docx
- Additionalfile5.tif
- Additionalfile2.jpeg
- Additionalfile1.jpeg
- Additionalfile7.docx
- Additionalfile3.docx
- Additionalfile9.xlsx