Multi-tissue transcriptome analysis using hybrid-sequencing reveals potential genes and biological pathways associated with azadirachtin A biosynthesis in neem (*Azadirachta indica*)

Huiyan Wang
Beijing University of Technology https://orcid.org/0000-0002-1068-5854
Ning Wang (wangning@bit.edu.cn)
https://orcid.org/0000-0002-9228-6872
Yixin Huo
Beijing Institute of Technology

Research article

Keywords: azadirachtin A, natural insecticides, secondary metabolism, triterpenoid biosynthesis, transcriptome, neem

Posted Date: August 1st, 2020

DOI: https://doi.org/10.21203/rs.2.23446/v3

Version of Record: A version of this preprint was published at BMC Genomics on October 28th, 2020. See the published version at https://doi.org/10.1186/s12864-020-07124-6.

Abstract

Background: Azadirachtin A is a triterpenoid from neem tree exhibiting excellent activities against over 600 insect species in agriculture. The production of azadirachtin A depends on extraction from neem tissues, which is not an eco-friendly and sustainable process. The low yield and discontinuous supply of azadirachtin A impedes further applications. The biosynthetic pathway of azadirachtin A is still unknown and is the focus of our study.

Results: We attempted to explore azadirachtin A biosynthetic pathway and identified the key genes involved by analyzing transcriptome data from five neem tissues through the hybrid-sequencing (Illumina HiSeq and Pacific Biosciences Single Molecule Real-Time (SMRT)) approach. Candidates were first screened by comparing the expression levels between the five tissues. After phylogenetic analysis, domain prediction, and molecular docking studies, 22 candidates encoding 2,3-oxidosqualene cyclase (OSC), alcohol dehydrogenase, cytochrome P450 (CYP450), acyltransferase, and esterase were proposed to be potential genes involved in azadirachtin A biosynthesis. Among them, two unigenes encoding homologs of MaOSC1 and MaCYP71CD2 were identified. A unigene encoding the complete homolog of MaCYP71BQ5 was reported. Accuracy of the assembly was verified by quantitative real-time PCR (qRT-PCR) and full-length PCR cloning.

Conclusions: By integrating and analyzing transcriptome data from hybrid-seq technology, 22 differentially expressed genes (DEGs) were finally selected as candidates involved in azadirachtin A pathway. The obtained reliable and accurate sequencing data provided important novel information for understanding neem genome. Our data shed new light on understanding the biosynthesis of other triterpenoids in neem trees and provides a reference for exploring other valuable natural product biosynthesis in plants.

Background

With the increasing concern on the threat of chemical pesticides to global crop protection programs, more attention is being paid towards bioactive and biodegradable plant or microbial-based biopesticides. Azadirachtin A, the major insecticidal ingredient in neem-based products, exhibits excellent bioactivity against over 600 insect species [1] in agricultural areas [2]. It is processed by insects as a natural hormone and it induces antifeedant, repellent, and growth inhibiting behavior in insects [3]. Azadirachtin A-based pesticides are biodegradable, environment-friendly, and non-toxic to mammals, plants, and birds. Due to these general superior characteristics of azadirachtin A, the agriculture segment accounts for the highest share (40%) in total revenues of neem extract product market. This market is expected to grow from $653 million in 2015 to $1.8 billion in 2022 with a high annual growth rate of 16.3% [4]. The current supply of azadirachtin A mostly depends on extracts from neem seeds [5] and leaves. Due to the limitation of the distribution of neem materials along with the complex and low-yield azadirachtin A-extracting approach, the production of azadirachtin A is far from meeting the market demand.

Being a potential insecticide, the synthesis of azadirachtin has been investigated over the last few decades. However, the complexity in its molecular architecture [6] was the main obstacle that held back the advances in azadirachtin A biosynthesis. After 20 years of investigation, the complete chemical synthesis of azadirachtin A was finally accomplished [7] in 2007. However, the low productivity (0.00015%) of the 71 step-reaction still limits its applications in the industry.

Being a triterpenoid from neem, azadirachtin A is derived from 2,3-oxidosqualene. The downstream biosynthesis pathway, that is, the pathway from the biosynthesis of scaffold to azadirachtin A remains unclear after decades of investigation. One of the attempts at the investigation of its biosynthesis was the feeding experiments in 1971; [^{3}H] euphol and [^{3}H] tirucallol were incubated with neem leaf crude extract. Euphol, rather than tirucallol was more efficiently incorporated into nimbolide (a limonoid with structural similarity to azadirachtin A) [8] . However, no enzyme producing tirucallol or euphol was isolated or characterized from neem.

Several resources on neem genome, such as the complementary DNA (cDNA) library [9], expressed sequence tag library [10], draft genome [11, 12] and transcriptome data [13, 14, 15] are available in public databases such as the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI). Genes involved in the mevalonate (MVA) pathway, methylerythritol phosphate (MEP) pathway [14, 16], or 2,3-oxidosquanlene biosynthesis [11, 15, 16, 17] in neem have been identified. Several neem-specific genes were found after comparative genome analysis of neem with *Arabidopsis thaliana*, *Oryza sativa,* and *Citrus sinensis* [11]. Upon comparison of gene expression within different neem tissues, genes encoding cytochromes P450 (CYP450s) were identified from fruit [16] and leaf [17]. However, none of the isolated genes had been functionally characterized. It was not until the key enzyme that converts 2,3-oxidosqualene into tirucalla-7,24-dien-3β-ol, AiOSC1, was isolated from neem transcriptome [18]. This was the first functionally characterized AiOSC1 reported to be related to azadirachtin scaffold biosynthesis. Two full-length CYP450s subsequently catalyzing tirucalla-7,24-dien-3β-ol into melianol were also isolated from *C. sinensis* var. Valencia and *Melia azedarach*. The homolog of MaCYP71CD2 in neem (AiCYP71CD2) was identified; the homolog of MaCYP71BQ5 (a characterized CYP450 from *M. azedarach*) [18] in neem was fragmented.

Although the genome and transcriptome data were easy to download from the Sequence Read Archive (SRA) database, the submitted files included all raw data without genome assembly or annotation. Besides, samples used in the previous reports [11, 12, 14-16] were from India. Genetic information within the same species found in different areas or developmental stages were different; this caused differences in gene expression. Hence, five neem tissues (fruit, leaf, stem, flower, and root) were sampled for transcriptome sequencing. Among the five tissues, fruit with the green hard seed has been reported to contain the highest amount of azadirachtin A throughout fruit development [14, 19] Leaf sample from a neem tree in China contained azadirachtin A at a concentration of 969.9 μg/g [20]. The percentage azadirachtin content in different tissues was consistent with that in a previous report (seed kernels, 0.03%; leaves, 0.9x10^{-3}%; bark, 0.5x10^{-3}%; root, 0.3x10^{-3}%; stem 0.2x10^{-3}%) [21]. Hence, these samples

were chosen as the higher azadirachtin A group and used for mining genes for azadirachtin A biosynthesis.

With the advances in high-throughput sequencing technologies, the third-generation sequencing represented by PacBio Single Molecule Real-Time (SMRT) technology [22] and Oxford Nanopore sequencing has been applied in academic research. Due to the high error rate of sequencing longer reads (15%) in third-generation sequencing as well as the accuracy of reads from Illumina sequencing, a new method called hybrid-seq [23] has been generated, that brings together the best of two sequencing technologies; the longer reads obtained are corrected by short but accurate reads from Illumina. As reported by Koren [24], only 0.1% (41) of the PacBio reads exactly matched the annotated exon structure before correction during genome assembly. This percentage will rise to 24.1% (12,065) after correction with short reads. Error correction method was used for *de novo* genome assembly of *Saccharomyces cerevisiae*. The generated contig N50 length is more than ten times greater than an Illumina-only assembly (678 kb versus 59.9 kb) and has >99.88% consensus identity when compared to the reference genome [25]. The successful applications of hybrid-seq in genome refining and isoform identification laid strong foundations for our study.

While the three steps involved in azadirachtin A biosynthesis have been characterized, the rest of the downstream pathway is still unexplored. Based on the metabolites in the Neem Metabolite Structure Database [26] as well as the distribution of metabolites (Table S1) in neem tissues, a putative biosynthetic pathway for azadirachtin A (Figure 1) was proposed. Five kinds of putative enzymes, oxidosqualene cyclase (OSC), cytochrome P450 (CYP450) [27], alcohol dehydrogenase (ADH) [28], acyltransferase (ACT), and esterase (EST) [29] were proposed in the putative biosynthetic pathway. Candidate mining was performed through the workflow of gene mining (Figure S1). Extensive bioinformatic analysis of unigenes involving phylogenetic analysis, domain prediction, and molecular docking further provided 22 candidates (Table 1) for the putative biosynthetic pathway, including 1 OSC unigene, 2 ADH unigenes, 12 CYP450 unigenes, 2 ACT unigenes, and 5 EST unigenes. Among them, 3 transcripts encoding the complete AiOSC1 and homologs of MaCYP71CD2 and MaCYP71BQ5 were also found in our study. Unigene containing the complete open reading frame (ORF) encoding the homolog of MaCYP71BQ5 [18] was first reported. Quantitative real-time PCR and full-length PCR cloning were used for verifying unigene expression level (Table S2) (Fragments per kilobase of transcript per million mapped reads (FPKM)) and transcript sequence accuracy. The obtained candidates could be used as an important resource to investigate the catalysts responsible for essential biochemical reactions in azadirachtin A biosynthesis as well as triterpenoid metabolism in closely related species of neem.

Table 1. Summary of candidate genes involved in azadirachtin A biosynthesis in neem

Classification	Unigene	Gene accession number	Homologs
OSC	transcript/14449	gi|443299067	AiOSC1*
ADH	transcript/18833	gi|572153023	cinnamyl-alcohol dehydrogenase
	transcript/19291	gi|572153023	cinnamyl-alcohol dehydrogenase
CYP450	transcript/17636	gi|225458053	CYP83B1
	transcript/17854	gi|641826901	CYP77A3
	transcript/16057	gi|567902124	flavonoid 3'-monooxygenase
	transcript/17284	gi|567889747	CYP89A2
	transcript/16777	gi|590722535	brassinosteroid-6-oxidase
	transcript/17001	gi|590722535	brassinosteroid-6-oxidase
	transcript/16577	gi|568834016	flavonoid 3'-monooxygenase
	transcript/16950	gi|645239614	cytokinin trans-hydroxylase
	transcript/17057	gi|567868115	flavonoid 3'-monooxygenase
	transcript/16971	gi|568825869	MaCYP71CD2*
	transcript/16742	gi|568830413	MaCYP71BQ5*
	transcript/16198	gi|568868580	CYP51G1
ACT	transcript/17792	gi|567873443	BAHD acyltransferase
	transcript/18214	gi|641830965	BAHD acyltransferase
EST	transcript/19188	gi|568822600	SGNH_plant_lipase
	transcript/19882	gi|568867507	Acetyl esterase/lipase
	transcript/19697	gi|225440163	Acetyl esterase/lipase
	transcript/19748	gi|641846434	Acetyl esterase/lipase
	transcript/18100	gi|568835134	pectin acetylesterase

Homologs with * represent the identified gene in the previous report [18].

Methods

Plant materials

Fresh and healthy tissues (from root, leaf, stem, flower, and fruit containing seed) were randomly sampled from a neem (*Azadirachta indica, A. indica*) tree at the park at Hainan University, China, followed by their

transcriptome analysis. All samples for RNA extraction and transcriptome sequencing were harvested from three plants. Tissues were gently rinsed and subsequently cut into small pieces. All materials were immediately frozen in liquid nitrogen and stored at −80°C before use.

RNA extraction

Total RNA was extracted using the RNA plant Plus Reagent (TianGen, Beijing, China) according to the manufacturer's protocol. The extracted RNA concentration and integrity were assessed using the RNA Nano 6000 assay kit with the Agilent Bioanalyzer 2100 system (Agilent, CA, USA). For PacBio sequencing, the RNA concentration and integrity were assessed using the Fragment Analyzer system (Agilent, CA, USA). The A260/A280 ratio ranging from 1.9 to 2.0, concentration above 285 ng/μL, and RNA integrity number (RIN) greater than 8.0 were used for subsequent cDNA library construction.

cDNA library construction and hybrid-sequencing

A total amount of 5 μg RNA was used for cDNA library construction. For Illumina HiSeq sequencing, oligo (dT) beads were used to isolate poly(A)$^+$ mRNA. The paired-end libraries were constructed with an insert size of approximately 250 bp. All libraries were sequenced commercially on the Illumina HiSeq 2000 sequencing platform (HiSeq 2000 V3) by the Beijing Genomics Institute (BGI-Shenzhen, China) according to the manufacturer' protocol to generate paired-end reads of an average length of 150 bp.

For PacBio SMRT sequencing, 1000 ng of mRNA from each tissue was pooled for cDNA library construction. Double-strand cDNA was synthesized according to SMARTer PCR cDNA synthesis kit (Clontech). DNA fragments were selected by BluePippin^{TM} (Sage Science, MA, USA) and ranged over four sizes: 1–2, 2–3, 3–6, and 5–10 kb. DNA fragments after the second large-scale PCR were used as template for SMRTbell library for sequencing. The throughput was about 12 Gb and covered all transcripts in the sample.

HiSeq reads were filtered by discarding the reads with adaptor sequences, reads with more than 5% ambiguous “N” bases, and low-quality reads. The filtered reads were then assembled using Trinity (v2.0.6) with default parameters to generate contigs. These contigs were then processed by sequencing clustering software TGICL (v2.0.6) for redundant Trinity assembled contig removal. Raw PacBio SMRT reads were processed using SMRT analysis server (v2.3) for full-length transcript generation. The obtained transcripts were corrected with filtered HiSeq reads using the LSC error correction tool [30] and subsequently filtered with CD-HIT-EST [31] for the removal of redundant Trinity generated fragments. Finally, the calibrated transcripts were assembled into unique putative transcripts (including contigs and singletons) and unigenes were characterized for subsequent analysis.

Annotation and differential gene expression analysis

The unigenes were annotated based on sequence similarity using BLASTX against five databases, including non-redundant protein database (Nr), SwissProt, COG, and KEGG protein database. The Pfam

annotation for unigenes was done using the HMMER 3.0 package. Sequence description for each unigene was transferred from homologous BLAST hits with E-value<10^{-5}. GO terms were assigned based on the top BLAST hit using Blast2GO. Genes were obtained by BLASTN using non-redundant nucleotide sequence database (Nt). Functional enrichment of the assigned GO terms was calculated and analyzed by the WEGO software. The distribution of gene functions was illustrated by the GO terms for biological process, cellular component, and molecular function.

Clean reads were mapped to unigenes using Bowtie2 (v2.2.5). The gene expression level was calculated with RSEM (v1.1.12). To compare the difference of gene expression among different samples, the FPKM (Fragments per kilobase of transcript per million mapped reads) method was used for normalization [32]. DESeq2 was used to identify differentially expressed genes (DEGs) (absolute value of $\log_2$ fold change≥1) after correction of *p*-values (adjusted<0.05) using the Benjamini-Hochberg procedure (false discovery rate, FDR≤0.001). Highly expressed unigenes characterized from leaf and fruit (high azadirachtin A tissues) libraries were used for candidate mining.

Analysis of phylogeny and domain architecture of unigenes

The SwissProt database was queried to retrieve all reviewed sequences of alcohol dehydrogenase, CYP450, acyltransferase, and esterase. These sequences were downloaded in FASTA format and aligned with the four kinds of candidates using the ClustalW algorithm with default parameters. Phylogenetic analysis based on multiple alignments of protein sequences was done using the Neighbor Joining [33] method as implemented in MEGA7 and the phylogenetic trees were visualized on iTOL [34]. Accessions of these protein sequences used in phylogenetic analysis are provided in Additional file 1. The protein sequences of the candidates were also searched against the Pfam database in order to get the domain architecture information complementary to that provided by SwissProt.

Molecular modelling and docking for enzyme-substrate analysis

Modelling of five CYP450 proteins encoded by unigenes was performed using the Phyre2 web portal using the fold recognition method [35]. To characterize the potential active site of binding sites in the protein, we used the web server, 3DLigandSite -Ligand binding site prediction Server [36]. Next, molecular docking was performed with Autodock 4.0 [37] to predict the interactions of four triterpenoids (tirucalla-7,24-dien-3β-ol, azadirone, nimbin, and nimbolide) as substrates for the CYP450 proteins. For covering acting domains present in CYP450 protein, grid spacing was maintained at 0.375 Å. Genetic algorithm (GA) was applied as the searching parameter with 10 GA runs; population size was set to 150; energy evaluations was set to maximum 25,00,000, considering the maximum number of generations as 27,000. The most favorable docking conditions were in the form of lowest binding energy conformations with H-bonds in cluster. Phymol 2.3 software was used for better analysis of interactions in the protein-ligand complexes obtained from Autodock 4.0 software.

Validation of hybrid-seq by quantitative real-time PCR and full-length PCR cloning

Ten transcripts were randomly selected to validate their expression from hybrid-seq by quantitative real-time PCR (qRT-PCR). Total RNA was processed with RNase-free-DNase I (TianGen, Beijing, China) following the manufacturer's instructions, to eliminate potential DNA contamination. First strand cDNA was synthesized using GoScript^{TM} Reverse Transcription System (Promega, Canada). The reactions were performed in triplicate using 2 μL diluted cDNA template in 20 μL total volume. qRT-PCR was performed in 96-well plates on a Bio-Rad CFX96 real-time PCR system (Bio-Rad, CA, USA) using SYBR Green Mix (Bio-Rad, CA, USA). A two-step cycling program was performed, comprising an initial 95℃ polymerase activation for 3 min, followed by 40 cycles of 95℃ for 10 s and 60℃ for 30 s. The melting curve was obtained by heating the amplicon from 65℃ to 95℃ at increments of 0.5℃ per 5s. The actin gene was used as an internal control to normalize all data. The relative quantitation ($\Delta\Delta Ct$) method was used to evaluate differences between the tissues for each gene examined. Data analysis was performed using GraphPad Prism version 5 for Windows (GraphPad Software, Inc. La Jolla, CA, USA). The primers for qRT-PCR reactions were listed in Additional file 2.

Ten transcripts were selected for cloning PCR verification; 4 out of 10 (transcript/14449, transcript/14554, transcript/16971, and transcript/16742) were specifically selected as they contained a complete ORF and were significant candidates in the azadirachtin A biosynthetic pathway. The others were randomly selected from the 10 transcripts used in qRT-PCR verification. The cDNA obtained (2 μL) was used in a 50 μL PCR reaction containing 2 μL of forward primer (10 μM), 2 μL of reverse primer (10 μM), and 25 μL of I-5™ 2×High-Fidelity Master Mix (TsingKe Biotech, Beijing, China). The PCR product was purified using GeneJET PCR Purification Kit (Thermo, USA) and assembled into pJET vector using CloneJET PCR Clone Kit (Thermo, USA). The constructs were then transformed into Trelief 5α chemically competent cells (TsingKe Biotech, Beijing, China) and the amplified constructs were sequenced by Genewiz Company (Genewiz, Beijing, China). The primers for full-length PCR cloning are listed in Additional file 2.

To generate a comprehensive overview of neem transcriptome, total RNAs were extracted from leaves, fruits (containing seeds), roots, stems, and flowers. To obtain the transcriptome data, we used the hybrid-seq technology that combines Illumina HiSeq and PacBio SMRT sequencing data and corrects the errors in long reads with short reads [24]. Different neem tissues (leaves, flowers, stems, fruits (containing seeds) and roots) were sequenced separately using Illumina HiSeq platform and generated 41.14, 41.35, 40.60, 40.59 and 40.64 million clean reads, respectively. PacBio SMRT platform produced 6.75 million clean reads. After calibration with short reads from Illumina platform, the assembled unigenes were corrected with an N50 of 5076 bp and mean length of 3607 bp (Table 2). The obtained assembly was 2.5 times longer than in the previous report [14]. The length of these unigenes ranged from 500 to 6001 bp; the majority (over 55.5%) of reads were distributed in the range of 4501 bp and above (Figure 2a).

To confirm the accuracy of the hybrid-seq (FPKM) results, we selected 10 unigenes and used qRT-PCR to determine their relative expression (Figure S2). The qRT-PCR and FPKM results were consistent except for transcript/19882 and transcript/16577. Their expression from RNA-seq (FPKM) and relative expression level by qRT-PCR in root and stem were inconsistent. The inconsistency between qRT-PCR and FPKM in transcriptome analysis has also been reported by other researchers. In the study by Zhang [38], the FPKM

expression levels of 8 out of 58 genes from transcriptome were inconsistent with their qRT-PCR results. In another study comparing the gene expression fold changes in the samples of *Metarhizium acridum* CQMa102, approximately 86.2% of the genes showed consistent results between RNA-sequencing and qRT-PCR data [39]. The inconsistency between FPKM and qRT-PCR may result from multiple reasons. Although qRT-PCR and RNA-seq are both used to measure gene expression, the unit of measurement [40] as well as the computing method are different for FPKM and qRT-PCR. Many factors affect the accuracy of FPKM and qRT-PCR. Bias in PCR amplification [41] and RNA-seq library preparation [42] and sequencing adds noise to the RNA-seq data. The quality of the mRNA, amplification efficiency, and the choice of reliable internal controls referred to as reference genes affect the accuracy of qRT-PCR [43]. Therefore, the consistency of transcript/16577 and transcript/19882 between qRT-PCR and FPKM was acceptable. Further examinations need to be performed on these two unigenes.

Ten unigenes were cloned by PCR using primers listed in Additional file 2. According to the sequencing results of the cloned unigenes, each of them was 100% identical to the sequences obtained from hybrid-seq platform. Among them, 4 unigenes attracted our attention since they contained a complete ORF of the genes. One of them (transcript/14554) was annotated as the neem NADPH-cytochrome P450 reductase 2. The cloned transcript/14449 was found to encode an OSC consisting of 760 amino acids. The length of two CYP450 unigenes (transcript/16971 and transcript/16742) was 1536 bp and 1527 bp; they encoded proteins consisting of 511 and 508 amino acids, respectively. Detailed sequencing results indicated that the unigenes in our study had good accuracy and were therefore reliable for further analyses.

Functional annotation and classification of unigenes

A total of 19,907 unigenes (98.54% of 20,201 unigenes) were annotated in at least one database (Table S3). The annotated unigenes were compared to known nucleotide sequences of other plant species. They best matched to the known nucleotide sequences from *C. sinensis* (52.43%), *Citrus clementine* (23.49%), *Theobroma cacao* (2.44%), *Vitis vinifera* (2.4%), and others (19.23%).

Clusters of Orthologous Groups of proteins (COG) and Gene Ontology (GO) classification were used to further evaluate the completeness and effectiveness of the neem annotation. All 17,634 unigenes (87.3% of 20,201 unigenes) were classified into 25 functional COGs (Figure 2b); of those, 2610 unigenes (14.8% of the total 17,634 classified unigenes) were categorized into general function prediction only cluster, which formed the largest group, whereas the clusters for replication, transcription, recombination, and repair followed closely. Although only 378 unigenes were categorized into the “Secondary metabolites biosynthesis transport and catabolism” cluster, they may play important roles in providing precursors for secondary metabolite biosynthesis.

A total of 13,453 assembled unigenes (66.6% of 20,201 unigenes) were assigned at least one of the 55 GO terms (Figure 2c); these unigenes were predominantly assigned to metabolic process (GO:0008152) and cellular process (GO:0009987). The unigenes categorized in the molecular function category were predominantly associated with catalytic activity (GO:0003824) and binding functions (GO:0005488). The

unigenes categorized in the cellular component category were predominantly associated with membrane (GO: 0016020), cell part (GO: 0044464), and cell (GO: 0005623). These findings showed that the main COG and GO classifications for the fundamental biological processes were identified.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database was used to systematically evaluate the gene biological functions in different pathways. A total of 16,778 unigenes were matched to the database and assigned to 135 KEGG pathways (Table S4); metabolic pathways (3590 unigenes, 21.4%) and biosynthesis of secondary metabolites (1560 unigenes, 9.3%) were the two dominant categories. In the biosynthesis of secondary metabolites category (Table S5) in neem, subcategories of flavonoid biosynthesis, terpenoid backbone biosynthesis (Table S6), steroid biosynthesis, sesquiterpenoid and triterpenoid biosynthesis (Table S7), diterpenoid metabolism, and carotenoid biosynthesis were included. There were 242 unigenes involved in the metabolic pathways of terpenoids and polyketides.

Differential expression analysis of unigenes in neem

In order to find candidate genes in the azadirachtin A biosynthesis pathway, all transcripts had been identified, annotated, and mapped in different pathways. Genes with a higher expression in the tissues with high azadirachtin A content, such as fruits and leaves, are more likely to be the involved in azadirachtin A biosynthesis. The expression level of unigenes was calculated by the FPKM method. The upregulated DEGs in leaf and fruit were unigenes which higher-expressed in leaf or fruit compared to the other tissues. Up-regulated DEGs in fruit and leaf were 219 and 397, respectively. These DEGs were used for mining candidates involved in azadirachtin A biosynthesis.

First screening of candidate genes involved in azadirachtin A biosynthesis

According to the putative azadirachtin A pathway in Figure 1, tirucalla-7,24-dien-3β-ol is assumed as the scaffold formed from 2,3-oxidosqualene. A few steps such as hydroxylation and furan ring formation occur after scaffold formation. The hydroxyl groups are then either oxidized to acid or acylated or esterified to esters, forming limonoid compounds like azadirone or nimbin. Azadirachtin A is finally obtained after modifications on azadirone or nimbin. ADH, CYP450, ACT, and EST are supposed to be involved in azadirachtin downstream pathway and thus their encoding-unigenes were chosen as candidates.

As a triterpenoid, the first step of azadirachtin A biopathway was the formation of its scaffold catalyzed by OSC. Among all unigenes involved in terpenoid biosynthesis, 8 detected unigenes were annotated as OSC, including transcript/1784, transcript/1866, transcript/8176, transcript/8892, transcript/9751, transcript/14584, transcript/19700, and transcript/14449. Among them, only transcript/14449 expressed higher in fruit. After phylogenetic analysis (Figure 3) of transcript/14449 with several characterized OSCs from other plants, transcript/14449 was grouped with AiOSC1 [18], a newly characterized OSC from neem, catalyzing the formation of tirucalla-7,24-dien-3β-ol; other unigenes were grouped with cycloartenol synthase. After DNA sequence analysis with AiOSC1(Figure S3), transcript/14449 is 100% identical to

AiOSC1, which means that these two genes were the same gene. It indicated that transcript/14449 could be a candidate gene for producing the azadirachtin A scaffold.

Among all the DEGs in fruit and leaf transcriptome data, DEGs encoding ADH, CYP450, ACT and EST were discovered. There were sixteen DEGs encoding ADH. Four up-regulated DEGs encoding ADH were selected for further screening. As for CYP450, sixteen DEGs encoded CYP450 and ten DEGs were selected. Similarly, thirteen and nineteen DEGs encoded ACT and EST respectively and there were four and twelve DEGs up-regulating in leaf and fruit tissues, respectively. DEGs with same sequence or sequences within 150 amino acids were excluded in further screening.

Among all the DEGs in fruit and leaf transcriptome data, DEGs encoding ADH, CYP450, ACT, and EST were identified. There were 16 DEGs encoding ADH; 4 upregulated DEGs encoding ADH were selected for further screening. There were 16 DEGs encoding CYP450; 10 out of these DEGs were selected. Similarly, 13 and 19 DEGs encoded ACT and EST, respectively; there were 4 and 12 DEGs upregulated in leaf and fruit tissues, respectively. DEGs with the same sequence or sequences coding for proteins of length less than 150 amino acids were excluded from further screening.

Further screening of the other four enzymes through phylogenetic analysis and domain prediction

Phylogenetic analysis (Figure 4) and protein domain prediction were used for further screening of these DEGs. Both, transcript/22186 and transcript/18833, were grouped with cinnamyl-alcohol dehydrogenase 4 (CADH4) [44] and transcript/19291 was grouped with CADH1; CADHs catalyze the biosynthesis of cinnamaldehyde from cinnamyl alcohol. Transcript/18482 was grouped with ADHX [45] which showed activity to primary and secondary alcohols. Transcript/18833 and transcript/19291 contained the PLN02514 domain that was also found in CADH [46]. Transcript/22186 contained the nsLTP2 [47] domain that is present in non-specific lipid-transfer protein. Transcript/18482 contained the GxGxxG motif [48] found in S-(hydroxymethyl) glutathione dehydrogenase. Transcript/18833 and transcript/19291 were selected as candidates for further examination.

According to the phylogenetic analysis of unigenes encoding CYP450, 5 transcripts (transcript/16057, transcript/16577, transcript/16950, transcript/16777, and transcript/17001) were grouped with members in CYP71 [49] and CYP72 [50] clades, whose members are reported to be involved in terpenoid biosynthesis. Transcript/16971 was grouped with CYP94B1 [51], an enzyme catalyzing the hydroxylation at C12 of jasmonyl-L-amino acid. Transcript/17284, transcript/17057, transcript/17636, and transcript/17854 were grouped in another clade. Transcript/17057 fell into a group with CYP82C4 [52], an enzyme hydrolyzing xanthotoxin (8-methoxypsoralen) into 5-hydroxyxanthotoxin. Transcript/17284 was grouped with CYP94B3 [53], this revealed that transcript/17284 may act as a hydroxylase. Transcript/17636 and transcript/17854 were classified into a group with uncharacterized CYP98A1.

According to CYP450 domain analysis, transcript/17636 and transcript/16971 contained the same CypX domain [54] as CYP81B1. The P450-cyclo_AA_1 domain in transcript/16950 was also found within cytokinin trans-hydroxylase [55]. The PLN00168 domain [56] was found in transcript/17824 and

transcript/19854. The PLN02687 domain (a domain in flavonoid 3'-monooxygenase [57]) was also contained by transcript/16057, transcript/16577, and transcript/17057. Transcript/16777 and transcript/17001 contained the PLN02774 domain, which is often found in brassinosteroid-6-oxidase [58]. Therefore, 5 DEGs (transcript/16057, transcript/ 16577, transcript/17057, transcript/16777, and transcript/17001) contained domains found within CYP450 oxidases and the others contained domain found within CYP450 hydroxylase.

Among all ACT DEGs, transcript/18186 was grouped with ARE1, the enzyme encoding sterol *O*-acyltransferase [59]. Transcript/18214 fell into a subclade with a DCR, a member of BAHD acyltransferase from *A. thaliana* involved in cutin biosynthesis [60]. Transcript/17792 was grouped with ARE2, a sterol O-acyltransferase from *Candida albicans* [61]. Transcript/19132 and TSM1 [62] were in a group which reveal that transcript/19132 was likely to be methyltransferase. Through domain analysis, transcript/17792 and transcript/18214 were shown to contain the HXXXD domain that is often found in the BAHD ACT family [60]. Transcript/19132 contained the domain PLN02177, also found in glycerol-3-phosphate acyltransferase [63]. Transcript/18186 and eukaryotic initiation factor 4B [64] had the same eIF-4B domain. Therefore, after combining phylogenetic analysis and domain prediction, transcript/17792 and transcript/18214 were selected as candidates of ACT involved in azadirachtin A biosynthesis.

Upon phylogenetic analysis of all EST DEGs, transcript/19998 and transcript/16750 were grouped with KAI2 [65]. KAI2 has been reported to be involved in seed germination and did not show esterase activity. Transcript/19188 was divided into a subclade with *A. thaliana* GDL15, that belonged to GDSL-like [66] lipase/ acylhydrolase superfamily and displayed hydrolytic activity with esters. Transcript/18100 was grouped with PME3, a pectinesterase catalyzing the hydrolysis of (1,4)-α-D-galacturonosyl methyl ester [67]. Transcript/19748 formed a tight subclade with TGL1, which is a sterol esterase mediating the hydrolysis of steryl esters [68]. Transcript/19882 and transcript/19697 were in the same group as HIDH [69] and CXE18, respectively and these two enzymes show activity to carboxylic esters. The conserved domain analysis of EST candidates presented that transcript/19188 contained the Ser-His-Asp (Glu) triad found in an SNGH plant lipase [70]. Transcript/19882, transcript/19697, and transcript/19748 had the AES domain [71] which is also contained within acetyl esterase/lipase. PAE domain [72] in transcript/18100 was also found in pectin acetylesterase while transcript/16750 and transcript/19998 contained the plant pectinesterase inhibitor domain PLN02201. Therefore, transcript/16750 and transcript/19998 were removed from the list of candidates after phylogenetic and domain analysis.

Molecular Docking analysis of CYP450s

The active site prediction in the 5 CYP450s was performed and the results are listed in Table S8. To further analyze the interactions between CYP450s and the four ligands, molecular docking was performed with Autodock 4.0. The details of docking are listed in Table S9 and specific interactions are displayed in Figure 5. Analysis revealed that binding energy was lowest in case of CYP16057 docked with tirucalla-7,24-dien-3β-ol forming zero hydrogen bond. However, azadirone and nimbolide formed stable

complexes with CYP16057 with one hydrogen bond with −7.20 and −6.40 kcal/mol of binding energies (Figure 5 and Table S9), respectively. The docking analysis for CYP16577 revealed that among all the ligands, binding energy was lowest for tirucalla-7,24-dien-3β-ol and nimbin with −10.07 and −9.83 kcal/mol, respectively, forming zero and two hydrogen bonds, respectively. Docking of CYP16577 with triterpenoids showed interaction through one hydrogen bond with binding energy of −9.42 and −9.16 kcal/mol for azadirone and nimbolide, respectively (Figure 5 and Table S9). The binding energy of CYP16777 docked with azadirone and tirucalla-7,24-dien-3β-ol was −9.96 and −9.75 kcal/mol, respectively, forming one hydrogen bond each. However, nimbolide and nimbin formed stable complexes with three and two hydrogen bonds respectively, indicating that the conformation of nimbolide is best suitable for CYP16777 when the number of hydrogen bonds formed between protein and ligand is set as the criterion. The hydrogen bonds between nimbolide and CYP16777 are formed at PHE354, ARG355, and ARG419, with ligand moiety at different positions (Table S9) with varying bond length (Figure 5). The docking analysis for CYP16950 revealed that among all the ligands, binding energy was lowest for tirucalla-7,24-dien-3β-ol and azadirone with −8.31 and −7.90 kcal/mol without forming hydrogen bond. However, nimbin and nimbolide formed stable complexes with CYP16950 with one hydrogen bond with −6.32 and −6.66 kcal/mol binding energies, respectively (Figure 5 and Table S9). Docking of CYP17001 with triterpenoids showed interaction through only one hydrogen bond with the lowest binding energy of −10.11 kcal/mol for tirucalla-7,24-dien-3β-ol. Both, azadirone and nimbolide, formed stable complexes with CYP17001 through two hydrogen bonds. The hydrogen bonds between nimbin and CYP17001 are formed at ARG355, ARG419, and GLY423 with the binding energy of −9.82 kcal/mol (Figure 5 and Table S9).

Tirucalla-7,24-dien-3β-ol was confirmed to be the scaffold of azadirachtin A. Nimbin, nimbolide, and azadirone are three important compounds isolated from neem. They were proposed as intermediates in azadirachtin A pathway. According to the docking analysis of five CYP450s with all the ligands, CYP16057, CYP16577, CYP16950, and CYP17001 showed strongest binding with tirucalla-7,24-dien-3β-ol. CYP16777 could more easily bind with azadirone. Three residues in CYP16777 and CYP17001 formed stable hydrogen bonds with nimbolide and nimbin, respectively. All the docking results indicate the priority of reactions between five CYP450s and four ligands. It also provided a theoretical basis for further functional assays of these CYP450s. The residues in proteins forming hydrogen bond with ligands also led to identification of sites for mutation analysis to improve the catalytic ability of these CYP450s.

Measurement of expression of unigenes in the neem secondary metabolite pathways

The expression levels of unigenes involved in secondary metabolite pathways including three terpenoid and two sterols, and putative azadirachtin A downstream pathway were analyzed based on the KEGG annotation and the FPKM method (Figure 6 and Table S10). Within all the unigenes, 13 were found to be related to the MVA pathway and 38 unigenes were found to be related to the MEP pathway (Table S6). Some of them (unigenes encoding mevalonate kinase (MVK), 1-deoxy-D-xylulose-5-phosphate synthase (DXPS), and 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS)) were expressed at higher levels in leaf and fruit. Geranyl pyrophosphate (GPP), catalyzed by geranyl diphosphate synthase (GPPS)

was the common intermediate of monoterpenoids. Four enzymes involved in myrcene, limonene, and terpineol biosynthetic pathways were highly expressed in flower and fruit. Unigenes encoding geranylgeranyl diphosphate synthase (GGPPS) and CYP82G1, involved in (E, E)-4, 8, 12-trimethyltrideca-1, 3, 7, 11-tetraene (TMTT) biosynthesis, were expressed at higher levels in flower and stem.

Farnesyl diphosphate synthase (FDS) catalyzes the formation of farnesyl pyrophosphate (FPP) from IPP; the unigene encoding FDS is expressed at the highest level in fruit followed by root and flower. Solavetivol pathway is one of the sesquiterpenoid pathways; the unigene encoding solavetivol synthase (CYP71D55) is expressed at higher levels in fruit and leaf. In triterpenoid biosynthesis, two FPPs under the continuous catalysis of squalene synthase (SQS) and squalene epoxidase (SQLE) form 2,3-oxidosqualene. SQS and SQLE encode unigenes that are expressed the highest in leaf. The expression level of the unigene encoding cycloartenol synthase (CAS1) in different tissues was in the order of fruit > stem > root. Methylsterol monooxygenase (SMO1) and sterol-4-alpha-carboxylate 3-dehydrogenase (NSDHL) were expressed at the highest level in flower. Unigenes encoding delta(14)-sterol reductase (TM7SF2) and CYP51G1 were highly expressed in leaf and fruit, respectively.

The expression level of unigenes involved in putative azadirachtin A downstream pathway has been presented. Transcript/14449 encoding the first enzyme in azadirachtin A downstream pathway is expressed at the highest level in fruit followed by in leaf. After scaffold synthesis, the methyl group at C14 is removed; this is catalyzed by the enzyme encoded by transcript/16198, that is highly expressed in leaf. Alcohol at C3 is continuously oxidized into C3 ketone group by transcript/18725 and transcript/17679 and forms common compounds [73] isolated from Meliaceae family; these two unigenes expressed highest in root and fruit, respectively. Next important step involved in azadirachtin A is the formation of the furan ring; the two CYP450s catalyzing its formation were isolated from *M. azedarach* and *C. sinensis* [18]. Two transcripts (transcript/16971 and transcript/16742) in our study were found to be homologs of the two identified CYP450s and might produce melianone [74] from its precursor. Transcript/16971 and transcript/16742 expressed highest in leaf and fruit, respectively. With some unknown enzymes, melianone is further modified into a compound with furan ring and C7-OH. The insecticidal C7-hydroxylated compound [75] is esterized by transcript/18100 (highly expressed in fruit and leaf) and further forms compounds like nimbin or nimbolide after some modifications. However, reactions between nimbin or nimbolide and azadirachtin A are still unclear.

Upon bioinformatic analysis of these DEGs, 2 transcripts (transcript/18833 and transcript/19291) encoding ADH, 12 transcripts encoding CYP450, 2 ACT transcripts (transcript/17792 and transcript/18214), and 5 transcripts encoding EST, were selected as candidates in the azadirachtin A downstream pathway. Some DEGs were removed from the library after phylogenetic analysis and domain prediction; some non-encoding DEGs were also deleted. DEGs were also selected even though they were not expressed at higher levels in fruit and leaf, for example, transcript/16198, transcript/16742, and transcript/16950. Transcript/16198 was annotated to encode sterol 14-demethylase, that removes the methyl group from C14 of sterol. Transcript/16742 encodes a protein with 509 amino acids and it was 98% identical to MaCYP71BQ5 (Figure S4). MaCYP71BQ5 is the CYP450 found to be involved in melianol

formation. Researchers could only get a fragment of its homolog (AiCYP71BQ5) from neem [18], whereas, our transcript/16742 contained the complete ORF of AiCYP71BQ5.

The increase in terpenoid precursor leads to the higher production of terpenoid. In the case of artemisinic acid production, improvement of terpenoid precursor by engineering the MVA pathway resulted in an increase in yield by 500 times [76]. Thus, the upregulated DEGs in the MVA or MEP pathway and the 2,3-oxidosqualene biosynthetic pathway could be used as building blocks in the construction of azadirachtin A precursor biosynthetic pathway in future.

Although reaction types and key enzymes were partially proposed based on the structural differences between intermediates in our putative azadirachtin A pathway, some information was still missing. For instance, we could not find the enzyme that catalyzes the hydroxylation reaction at C7 site. Further, the order of reactions downstream of azadirachtin A was not clear. Neither the number of reactions nor the catalysis type were characterized. These limitations of pathway lead to insufficient mining of the neem transcriptome data. This might be one of the reasons for slow progress in azadirachtin pathway exploration even though numerous neem genome and transcriptome data are available.

Conclusions

In conclusion, the multi-tissue transcriptome analysis revealed five types of genes potentially involved in azadirachtin A downstream pathway and their respective transcript levels. It also indicated that the neem tree genome encodes a high number of terpene or limonoid biosynthetic genes. Finally, 22 unigenes encoding enzymes including OSC, ADH, CYP450, ACT, and EST were selected as candidates involved in azadirachtin A downstream pathway. This is the first report on hybrid-seq transcriptome profiling analysis of *A. indica*. The obtained unigenes may provide a valid and diverse candidate pool for the study of selective modification by the functional groups in the triterpenoid or limonoid skeleton as well as for the study of convergent evolution in secondary metabolism.

Declarations

Ethics approval and consent to participate

The locations of material collected here are neither privately owned lands nor protected areas. No specific permits were required for our research.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Prof. Zhiwei Wang, Dr. Huangying Shu, and Yaqing Wei from the Hainan Key Laboratory for Sustainable Utilization of Tropical Bioresources, College of Horticulture, Hainan University, for neem identification and sampling. We would like to thank Editage (www.editage.cn) for English language editing.

Funding

This work was supported by National Key R&D Program of China(2017YFD0201400) and the Fundamental Research Funds for the Central Universities.

Availability of data and materials

All data generated or analyzed during the current study are included in this article and its supplementary information files. The raw reads have been deposited in the NCBI Sequence Read Archive (SRA) database under BioProject ID PRJNA590058.

Author's contributions

HW and YH conceived this study and designed the experiments. HW conducted all the experiments and analyzed the data. HW wrote the manuscript, NW and YH reviewed and edited the manuscript. All authors read and approved the final manuscript.

Abbreviations

cDNA: complementary DNA; DEGs: differentially expressed genes; OSC: 2,3-oxidosqualene cyclase; ADH: alcohol dehydrogenase; CYP450: cytochrome P450; ACT: acyltransferase; EST: esterase; PacBio SMRT: Pacific Biosciences Single Molecule Real Time; FPKM: Fragments per kilobase of transcript per million mapped reads; Nr: non-redundant protein database, Nt: nucleotide sequences; COG: Clusters of Orthologous Groups of proteins; KEGG: Kyoto Encyclopedia of Genes and Genomes protein database; GO: Gene Ontology; qRT-PCR: quantitative real-time PCR; NGS: next-generation sequencing; IDI: Isopentenyl-diphosphate δ-isomerase; GGPS: Geranyl diphosphate synthase; FDS: Farnesyl diphosphate synthase; SQLE: Squalene epoxidase; G3P: 3-phosphate glyceraldehyde; MVA: mevalonate; MEP: methylerythritol phosphate; IPP: isopentenyl pyrophosphate; DMAPP: γ-dimethylallyl pyrophosphate; GPP: geranyl pyrophosphate; FPP: farnesyl pyrophosphate; DXP: 1-Deoxy-D-xylulose 5-phosphate; MVP: 5-phosphomevalonate; MVPP: (R)-5-Diphosphomevalonate; GGPP: geranylgeranyl pyrophosphate; HMG-CoA: 3-hydroxy-3-methyl-glutaryl-CoA; PCME: 2-phospho-4-2-C-methyl-D-erythritol; MECP: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate; HMBDP: 1-hydroxy-2-methyl-2-butenyl 4-diphosphate; TMTT: (E,E)-4,8,12-trimethyltrideca-1,3,7,11-tetraene; DXPS: 1-deoxy-D-xylulose-5-phosphate synthase; MDS: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HMGR: hydroxymethyl-glutaryl-CoA reductase; MVK: Mevalonate kinase; PMVK: Phosphomevalonate kinase; GPPS: Geranyl diphosphate synthase; MCS: Myrcene/ocimene synthase; LMS: (R)-limonene synthase; ATNS: (-)-alpha-terpineol synthase; CNS: 1,8-

cineole synthase; TM7SF2: Delta(14)-sterol reductase; NSDHL: Sterol-4-alpha-carboxylate 3-dehydrogenase; SMO1: Methylsterol monooxygenase 1; GGPPS: Geranylgeranyl diphosphate synthase; SQS: Squalene synthase; CAS1: Cycloartenol synthase.

References

1. Hummel H, Langner S, Leithold G, Schmutterer H. NEEM: UNUSUALLY VERSATILE PLANT GENUS AZADIRACHTA WITH MANY USEFUL AND SO FAR INSUFFICIENTLY EXPLOITED PROPERTIES FOR AGRICULTURE, MEDICINE, AND INDUSTRY. Commun Agric Appl Biol Sci. 2014; 79:211-228.
2. Agbo B, And A, Ajaba M. A REVIEW ON THE USE OF NEEM (Azadirachta indica) AS A BIOPESTICIDE. Journal of Biopesticide and Environment. 2015; 2.
3. Oulhaci CM, Denis B, Kilani-Morakchi S, Sandoz J-C, Kaiser L, Joly D, et al. Azadirachtin effects on mating success, gametic abnormalities and progeny survival in Drosophila melanogaster (Diptera). Pest Manag Sci. 2018; 74(1):174-180.
4. P&S Market Research. Global Neem Extract Market Size, Share, Development, Growth and demand Forecast to 2022 [https://www.psmarketresearch.com/market-analysis/neem-extract-market]
5. Ambrosino P, Fresa R, Fogliano V, Monti SM, Ritieni A. Extraction of Azadirachtin A from Neem Seed Kernels by Supercritical Fluid and Its Evaluation by HPLC and LC/MS. J Agric Food Chem. 1999; 47(12):5252-5256.
6. Bilton JN, Broughton HB, Jones PS, Ley SV, Rzepa HS, Sheppard RN, et al. An x-ray crystallographic, mass spectroscopic, and NMR study of the limonoid insect antifeedant azadirachtin and related derivatives. Tetrahedron. 1987; 43(12):2805-2815.
7. Veitch GE, Beckmann E, Burke BJ, Boyer A, Maslen SL, Ley SV. Synthesis of Azadirachtin: A Long but Successful Journey. Angew Chem, Int Ed. 2007; 46(40):7629-7632.
8. Ekong DEU, Ibiyemi SA, Olagbemi EO. The meliacins (limonoids). Biosynthesis of nimbolide in the leaves of Azadirachta indica. J Chem Soc D. 1971(18):1117-1118.
9. Narnoliya LK, Rajakani R, Sangwan NS, Gupta V, Sangwan RS. Comparative transcripts profiling of fruit mesocarp and endocarp relevant to secondary metabolism by suppression subtractive hybridization in Azadirachta indica (neem). Mol Biol Rep. 2014; 41(5):3147-3162.
10. Rajakani R, Narnoliya L, Sangwan NS, Sangwan RS, Gupta V. Subtractive transcriptomes of fruit and leaf reveal differential representation of transcripts in Azadirachta indica. Tree Genetics & Genomes. 2014; 10(5):1331-1351.
11. Krishnan NM, Pattnaik S, Jain P, Gaur P, Choudhary R, Vaidyanathan S, et al. A draft of the genome and four transcriptomes of a medicinal and pesticidal angiosperm Azadirachta indica. BMC Genom. 2012; 13:464-464.
12. Krishnan NM, Jain P, Gupta S, Hariharan AK, Panda B. An Improved Genome Assembly of Azadirachta indica A. Juss. G3 (Bethesda). 2016; 6(7):1835-1840.

13. Wang S, Zhang H, Li X, Zhang J. Gene expression profiling analysis reveals a crucial gene regulating metabolism in adventitious roots of neem (Azadirachta indica). RSC Adv. 2016; 6(115):114889-114898.
14. Pandreka A, Dandekar DS, Haldar S, Uttara V, Vijayshree SG, Mulani FA, et al. Triterpenoid profiling and functional characterization of the initial genes involved in isoprenoid biosynthesis in neem (Azadirachta indica). BMC Plant Biol. 2015; 15:214-214.
15. Krishnan N, Pattnaik S, Sa D, K Hariharan A, Gaur P, Chaudhary R, et al. De novo sequencing and assembly of Azadirachta indica fruit transcriptome. Curr Sci. 2011; 101:1553.
16. Bhambhani S, Lakhwani D, Gupta P, Pandey A, Dhar YV, Kumar Bag S, et al. Transcriptome and metabolite analyses in Azadirachta indica: identification of genes involved in biosynthesis of bioactive triterpenoids. Scientific Reports. 2017; 7(1):5043.
17. Wang Y, Chen X, Wang J, Xun H, Sun J, Tang F. Comparative analysis of the terpenoid biosynthesis pathway in Azadirachta indica and Melia azedarach by RNA-seq. Springerplus. 2016; 5(1):819-819.
18. Hodgson H, De La Peña R, Stephenson MJ, Thimmappa R, Vincent JL, Sattely ES, et al. Identification of key enzymes responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production. Proc Natl Acad Sci. 2019; 116(34):17096-17104.
19. Kurimoto S-i, Takaishi Y, Ahmed FA, Kashiwada Y. Triterpenoids from the fruits of Azadirachta indica (Meliaceae). Fitoterapia. 2014; 92:200-205.
20. Song L, Wang J, Gao Q, Ma X, Wang Y, Zhang Y, et al. Simultaneous determination of five azadirachtins in the seed and leaf extracts of Azadirachta indica by automated online solid-phase extraction coupled with LC–Q-TOF–MS. Chem Cent J. 2018; 12(1):85.
21. Sundaram KMS. Azadirachtin biopesticide: A review of studies conducted on its analytical chemistry, environmental behaviour and biological effects. J Environ Sci Health B. 1996; 31(4):913-948.
22. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015; 13(5):278-289.
23. Bashir A, Klammer A, Robins WP, Chin C-S, Webster D, Paxinos E, et al. A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol. 2012; 30(7):701-707.
24. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012; 30(7):693-700.
25. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. Oxford Nanopore Sequencing, Hybrid Error Correction, and de novo Assembly of a Eukaryotic Genome. bioRxiv. 2015:013490.
26. Hatti KS, Muralitharan L, Hegde R, Kush A. NeeMDB: Convenient Database for Neem Secondary Metabolites. Bioinformation. 2014; 10(5):314-315.
27. Xu J, Wang X-y, Guo W-z. The cytochrome P450 superfamily: Key players in plant development and defense. J Integr Agric. 2015; 14(9):1673-1686.

28. Paddon CJ, Keasling JD. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. Nat Rev Microbiol. 2014; 12(5):355-367.
29. Legrand G, Delporte M, Khelifi C, Harant A, Vuylsteker C, Mörchen M, et al. Identification and Characterization of Five BAHD Acyltransferases Involved in Hydroxycinnamoyl Ester Metabolism in Chicory. Front Plant Sci. 2016; 7(741).
30. Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio long read accuracy by short read alignment. PLoS One. 2012; 7(10):e46679-e46679.
31. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England). 2012; 28(23):3150-3152.
32. Hu Z, Li G, Sun Y, Niu Y, Ma L, He B, et al. Gene transcription profiling of Aspergillus oryzae 3.042 treated with ergosterol biosynthesis inhibitors. Braz J Microbiol. 2019; 50(1):43-52.
33. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4(4):406-425.
34. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research. 2019; 47(W1):W256-W259.
35. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015; 10(6):845-858.
36. Wass MN, Kelley LA, Sternberg MJ. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic acids research. 2010; 38(Web Server issue):W469-473.
37. Pandey V, Dhar YV, Gupta P, Bag SK, Atri N, Asif MH, et al. Comparative interactions of withanolides and sterols with two members of sterol glycosyltransferases from Withania somnifera. BMC Bioinform. 2015; 16(1):120-120.
38. Zhang W, Chen J, Keyhani NO, Zhang Z, Li S, Xia Y. Comparative transcriptomic analysis of immune responses of the migratory locust, Locusta migratoria, to challenge by the fungal insect pathogen, Metarhizium acridum. BMC Genom. 2015; 16:867-867.
39. Everaert C, Luypaert M, Maag JLV, Cheng QX, Dinger ME, Hellemans J, et al. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. Sci Rep. 2017; 7(1):1559-1559.
40. Pombo MA, Zheng Y, Fei Z, Martin GB, Rosli HG. Use of RNA-seq data to identify and validate RT-qPCR reference genes for studying the tomato-Pseudomonas pathosystem. In: Sci Rep. vol. 7; 2017: 44905.
41. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. 2012; 109(4):1347-1352.
42. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012; 30(8):777-782.

43. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clin Chem. 2009; 55(4):611-622.
44. Kim S-J, Kim M-R, Bedgar DL, Moinuddin SGA, Cardenas CL, Davin LB, et al. Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in Arabidopsis. Proc Natl Acad Sci U S A. 2004; 101(6):1455-1460.
45. Achkor H, Díaz M, Fernández MR, Biosca JA, Parés X, Martínez MC. Enhanced formaldehyde detoxification by overexpression of glutathione-dependent formaldehyde dehydrogenase from Arabidopsis. Plant Physiol. 2003; 132(4):2248-2255.
46. Jin Y, Zhang C, Liu W, Qi H, Chen H, Cao S. The cinnamyl alcohol dehydrogenase gene family in melon (Cucumis melo L.): bioinformatic analysis and expression patterns. PLoS One. 2014; 9(7):e101730-e101730.
47. Hoh F, Pons JL, Gautier MF, de Lamotte F, Dumas C. Structure of a liganded type 2 non-specific lipid-transfer protein from wheat and the molecular basis of lipid binding. Acta Crystallogr D Biol Crystallogr. 2005; 61(Pt 4):397-406.
48. Kavanagh KL, Jörnvall H, Persson B, Oppermann U. Medium- and short-chain dehydrogenase/reductase gene and protein families : the SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. Cell Mol Life Sci. 2008; 65(24):3895-3906.
49. Nagegowda DA, Gupta P. Advances in biosynthesis, regulation, and metabolic engineering of plant specialized terpenoids. Plant Sci. 2020; 294:110457.
50. Zheng X, Li P, Lu X. Research advances in cytochrome P450-catalysed pharmaceutical terpenoid biosynthesis in plants. J Exp Bot. 2019; 70(18):4619-4630.
51. Koo AJ, Thireault C, Zemelis S, Poudel AN, Zhang T, Kitaoka N, et al. Endoplasmic reticulum-associated inactivation of the hormone jasmonoyl-L-isoleucine by multiple members of the cytochrome P450 94 family in Arabidopsis. J Biol Chem. 2014; 289(43):29728-29738.
52. Rajniak J, Giehl RFH, Chang E, Murgia I, von Wirén N, Sattely ES. Biosynthesis of redox-active metabolites in response to iron deficiency in plants. Nat Chem Biol. 2018; 14(5):442-450.
53. Heitz T, Widemann E, Lugan R, Miesch L, Ullmann P, Desaubry L, et al. Cytochromes P450 CYP94C1 and CYP94B3 catalyze two successive oxidation steps of plant hormone Jasmonoyl-isoleucine for catabolic turnover. J Biol Chem. 2012; 287(9):6296-6306.
54. Cryle MJ, Bell SG, Schlichting I. Structural and Biochemical Characterization of the Cytochrome P450 CypX (CYP134A1) from Bacillus subtilis: A Cyclo-l-leucyl-l-leucyl Dipeptide Oxidase. Biochemistry. 2010; 49(34):7282-7296.
55. Takei K, Yamaya T, Sakakibara H. Arabidopsis CYP735A1 and CYP735A2 Encode Cytokinin Hydroxylases That Catalyze the Biosynthesis of trans-Zeatin. The Journal of biological chemistry. 2004; 279:41866-41872.

56. Lam PY, Liu H, Lo C. Completion of Tricin Biosynthesis Pathway in Rice: Cytochrome P450 75B4 Is a Unique Chrysoeriol 5'-Hydroxylase. Plant Physiol. 2015; 168(4):1527-1536.
57. Rice Annotation P, Itoh T, Tanaka T, Barrero RA, Yamasaki C, Fujii Y, et al. Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. Genome Res. 2007; 17(2):175-183.
58. Shimada Y, Fujioka S, Miyauchi N, Kushiro M, Takatsuto S, Nomura T, et al. Brassinosteroid-6-Oxidases from Arabidopsis and Tomato Catalyze Multiple C-6 Oxidations in Brassinosteroid Biosynthesis. Plant Physiol. 2001; 126(2):770-779.
59. Ivashov VA, Zellnig G, Grillitsch K, Daum G. Identification of triacylglycerol and steryl ester synthases of the methylotrophic yeast Pichia pastoris. Biochim Biophys Acta. 2013; 1831(6):1158-1166.
60. Panikashvili D, Shi JX, Schreiber L, Aharoni A. The Arabidopsis DCR encoding a soluble BAHD acyltransferase is required for cutin polyester formation and seed hydration properties. Plant physiology. 2009; 151(4):1773-1789.
61. Kim K-Y, Shin Y-K, Park J-C, Kim J-H, Yang H, Han D-M, et al. Molecular cloning and biochemical characterization of Candida albicans acyl-CoA:sterol acyltransferase, a potential target of antifungal agents. Biochemical and Biophysical Research Communications. 2004; 319(3):911-919.
62. Fellenberg C, Milkowski C, Hause B, Lange PR, Vogt T. Tapetum-specific location of a cation-dependent O-methyltransferase in Arabidopsis thaliana. Plant Journal. 2008; 56(1):132-145.
63. Yu J, Loh K, Song Z-y, Yang H-q, Zhang Y, Lin S. Update on glycerol-3-phosphate acyltransferases: the roles in the development of insulin resistance. Nutrition & Diabetes. 2018; 8(1):34.
64. Metz AM, Wong KCH, Malmström SA, Browning KS. Eukaryotic Initiation Factor 4B from Wheat and Arabidopsis thaliana Is a Member of a Multigene Family. Biochem Biophys Res Commun. 1999; 266(2):314-321.
65. Guo Y, Zheng Z, La Clair JJ, Chory J, Noel JP. Smoke-derived karrikin perception by the α/β-hydrolase KAI2 from Arabidopsis. Proc Natl Acad Sci U S A. 2013; 110(20):8284-8289.
66. Chepyshko H, Lai C-P, Huang L-M, Liu J-H, Shaw J-F. Multifunctionality and diversity of GDSL esterase/lipase gene family in rice (Oryza sativa L. japonica) genome: new insights from bioinformatics analysis. BMC Genom. 2012; 13:309-309.
67. Christensen TMIE, Nielsen JE, Kreiberg JD, Rasmussen P, Mikkelsen JD. Pectin methyl esterase from orange fruit: characterization and localization by in-situ hybridization and immunohistochemistry. Planta. 1998; 206(4):493-503.
68. Köffel R, Tiwari R, Falquet L, Schneiter R. The Saccharomyces cerevisiae YLL012/YEH1, YLR020/YEH2, and TGL1 genes encode a novel family of membrane-anchored lipases that are required for steryl ester hydrolysis. Mol Cell Biol. 2005; 25(5):1655-1668.
69. Akashi T, Aoki T, Ayabe S-I. Molecular and biochemical characterization of 2-hydroxyisoflavanone dehydratase. Involvement of carboxylesterase-like proteins in leguminous isoflavone biosynthesis. Plant Physiol. 2005; 137(3):882-891.

70. Mølgaard A, Kauppinen S, Larsen S. Rhamnogalacturonan acetylesterase elucidates the structure and function of a new family of hydrolases. Structure. 2000; 8(4):373-383.
71. Pereira EO, Tsang A, McAllister TA, Menassa R. The production and characterization of a new active lipase from Acremonium alcalophilum using a plant bioreactor. Biotechnol Biofuels. 2013; 6:111-111.
72. Philippe F, Pelloux J, Rayon C. Plant pectin acetylesterase structure and function: new insights from bioinformatic analysis. BMC Genom. 2017; 18(1):456.
73. Paal C. Ueber die Derivate des Acetophenonacetessigesters und des Acetonylacetessigesters. Ber Dtsch Chem Ges. 1884; 17(2):2756-2767.
74. Polonsky J, Varon Z, Rabanal RM, Jacquemin H. 21,20-Anhydromelianone and Melianone from Simarouba amara (Simaroubaceae); Carbon-13 NMR Spectral Analysis of Δ7-Tirucallol-Type Triterpenes. Isr J Chem. 1977; 16(1):16-19.
75. Huidana F, Conghaia Z, Shengjiaoa Y, Jun L. Advances of synthesis and structure modification and bioactivity of azadirachtin. Chinese J Org Chem. 2009; 29:20-33.
76. Paddon CJ, Westfall PJ, Pitera DJ, Benjamin K, Fisher K, McPhee D, et al. High-level semi-synthetic production of the potent antimalarial artemisinin. Nature. 2013; 496:528.

Table 2

Table 2. Summary of Illumina and PacBio output data quality and assembled sequences of libraries of neem.

Sequencing Platform	Sample	Reads Number (M)	Clean Bases (G)	Q20 (%)	Q30 (%)	Total number of unigene	Total length of unigene (bp)	Mean length of unigene (bp)	N50	GC (%)
Ilumina HiSeq	**Leaf**	41.14	6.17	96.95	92.30	50394	82508501	1380	2201	40.64
	Flower	41.35	6.20	97.29	93.06	62426	95564612	1530	2335	40.13
	Stem	40.60	6.09	97.29	93.11	65762	98612970	1499	2372	40.39
	Fruit	40.59	6.09	97.45	93.39	66668	81799852	1226	2077	41.93
	Root	40.64	6.10	97.44	93.43	45459	57083335	1255	2100	41.10
Hiseq summary	—	—	—	—	—	113008	175268545	1550	2599	40.99
PacBio SMRT	**Mixed tissue**	6.75	11.28	—	—	22884	82035635	3584	5068	43.55
Calibrated PacBio by Illumina		—	—	—	—	20201	72872459	3607	5076	43.55

Q20 and Q30 on Illumina platform correspond to the predicted base call error rate of 1 % and 0.1%, respectively.

N50: The minimum contig length needed to cover 50% of the transcriptome.

Figures

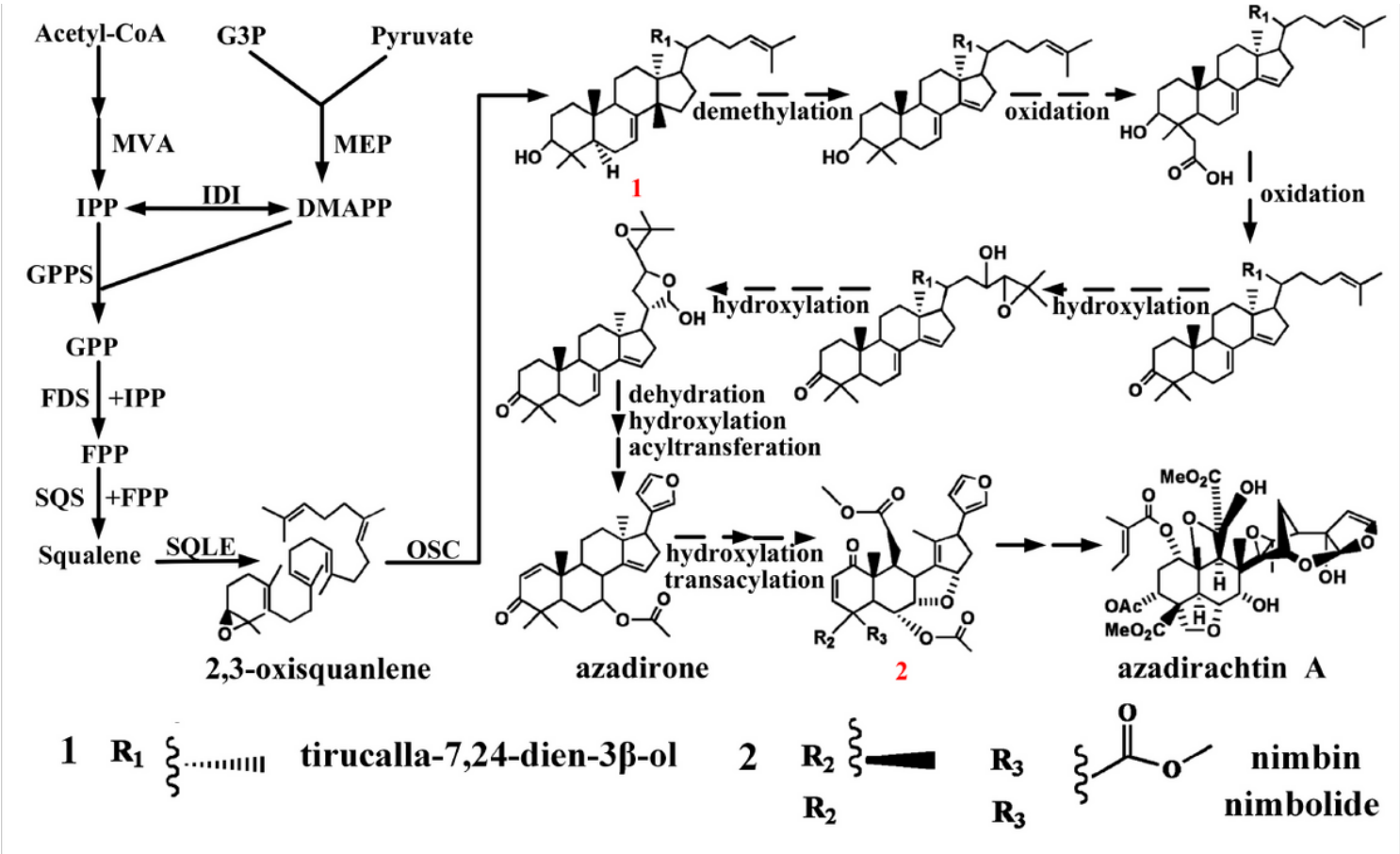

Figure 1

Hypothetical pathway of azadirachtin biosynthesis in A. indica. Isopentenyl-diphosphate δ-isomerase (IDI); Geranyl diphosphate synthase (GPPS); Farnesyl diphosphate synthase (FDS); Squalene epoxidase (SQLE); 2,3-oxidosqualene cyclase (OSC); G3P: 3-phosphoglyceraldehyde; MVA: mevalonate; MEP: methylerythritol phosphate; IPP: isopentenyl pyrophosphate; DMAPP: γ-dimethylallyl pyrophosphate; GPP: geranyl pyrophosphate; FPP: farnesyl pyrophosphate. The pathway of 2,3-oxidosqualene biosynthesis is connected by solid lines and arrows. The putative biosynthetic pathway of azadirachtin A is connected by dashed lines and arrows. Isomers are numbered in red and detailed structures are shown in the figure.

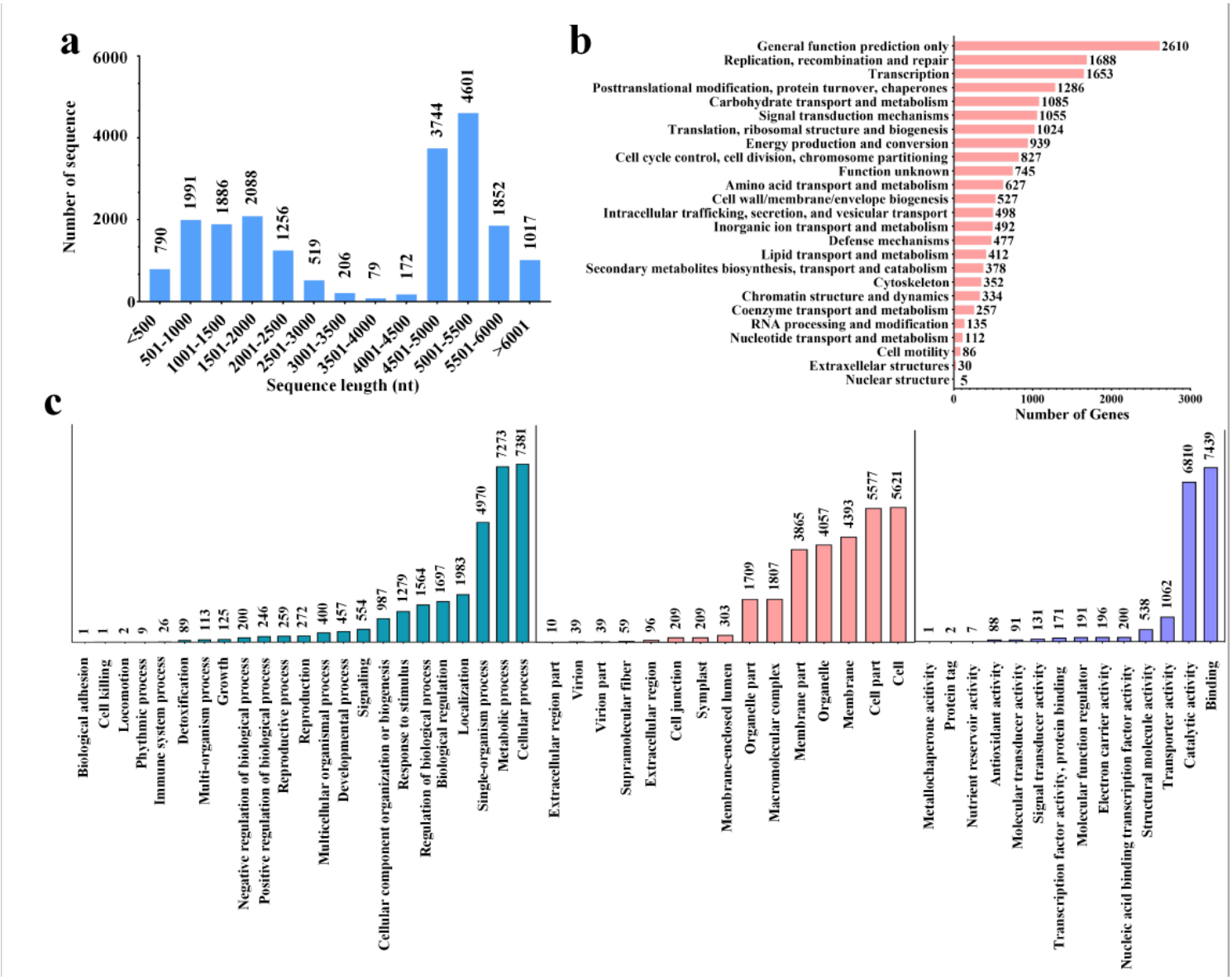

Figure 2

Analysis of size distribution and gene classification. a) the size distribution of de novo assembled unigenes of A. indica, b) COG functional classification, and c) distribution of GO terms assigned to unigenes of assembled cDNA library.

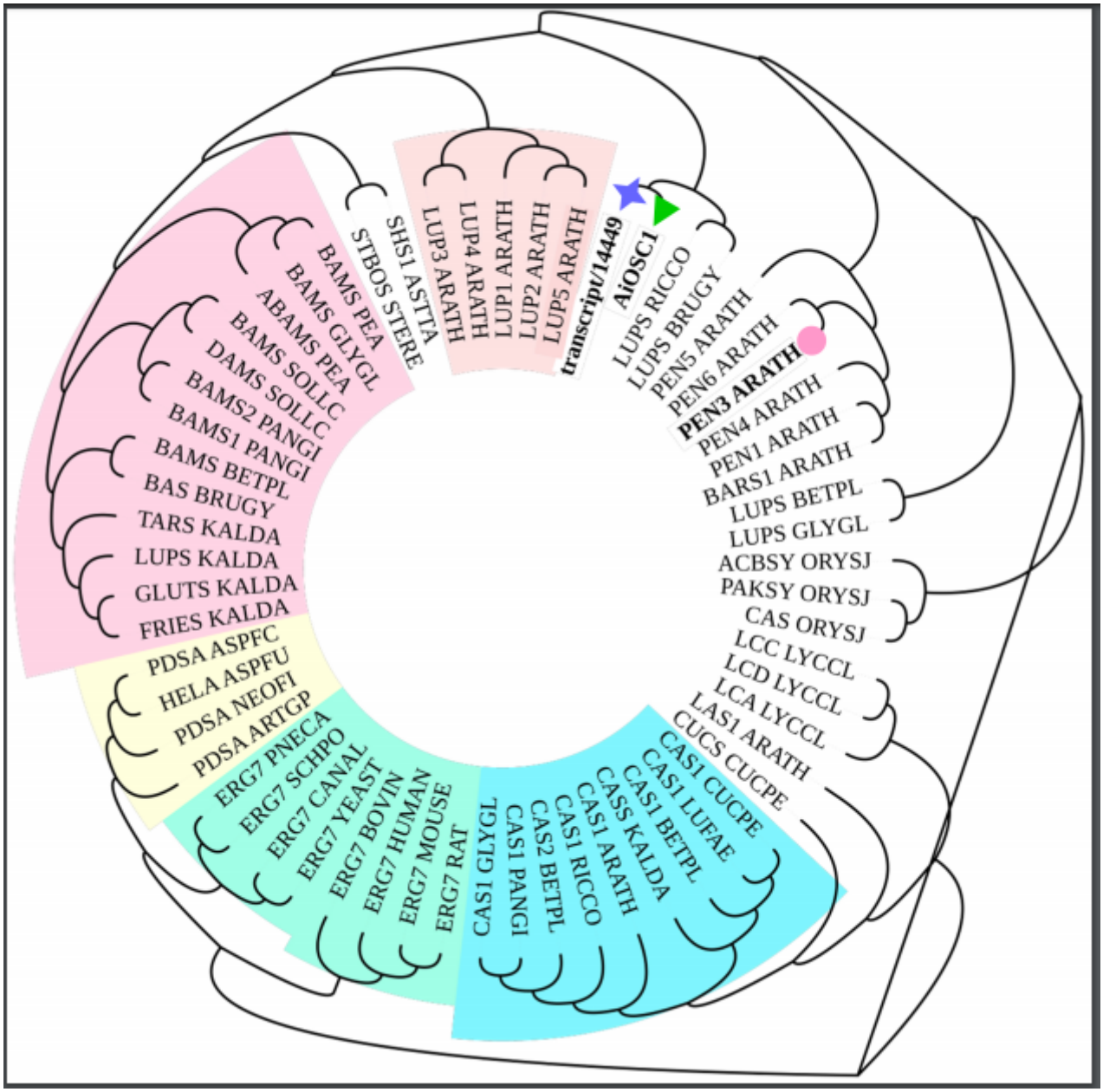

Figure 3

Phylogenetic tree of candidate OSC from neem transcriptome. Functionally characterized OSCs from other plant species including the previously characterized tirucalla-7,24-dien-3β-ol synthases from A. thaliana (PEN3) (bold, marked with pink circle) and AiOSC1 (bold, marked with green triangle) identified previously. Candidate OSC chosen for further analysis is displayed in bold and marked with a purple star. The phylogenetic tree was constructed by MEGA V7 and formatted using iTOL.

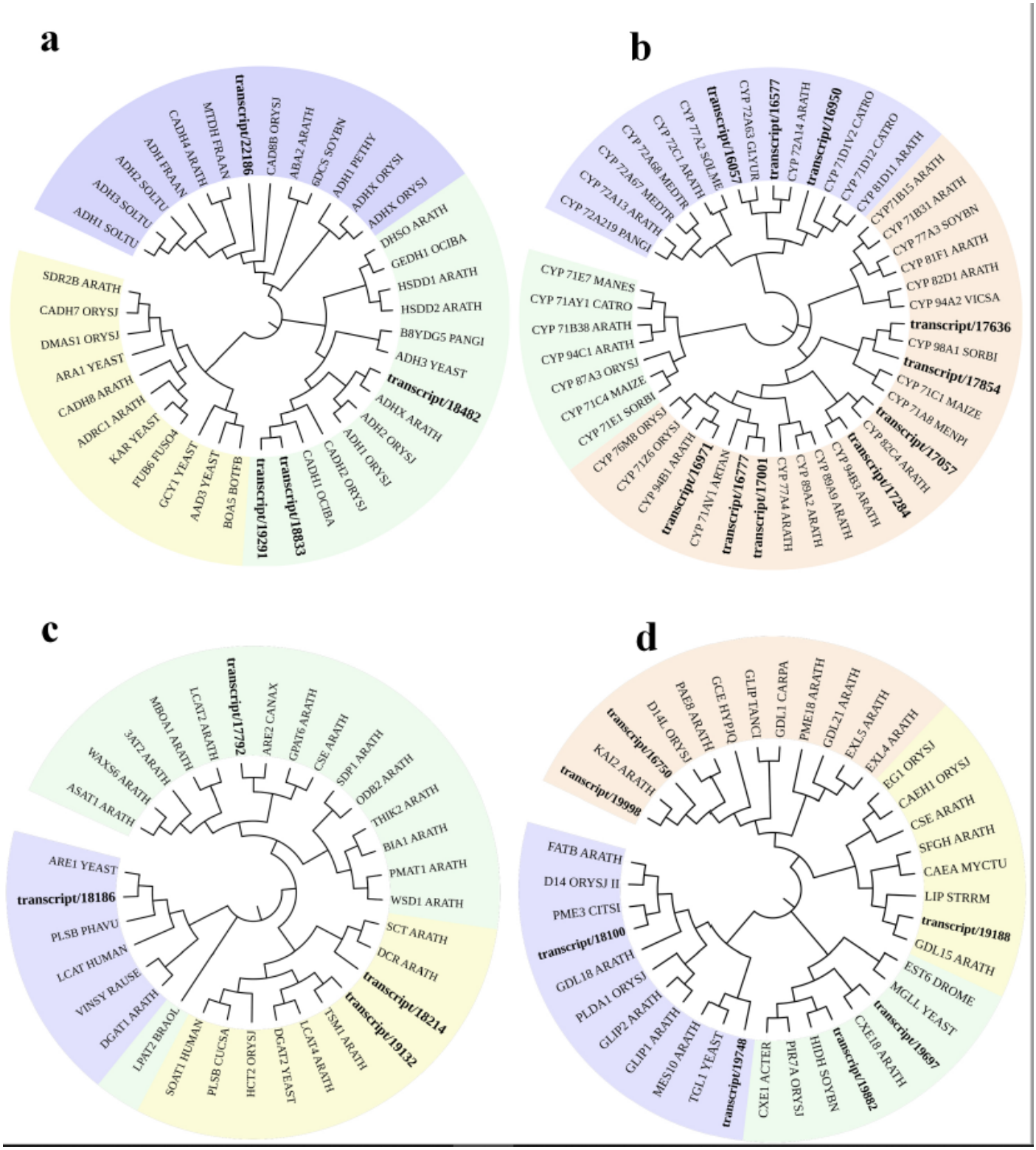

Figure 4

Phylogenetic analysis of a) alcohol dehydrogenase, b) CYP450, c) acyltransferase, and d) esterase candidates. Neighbor-joining trees were constructed for four types of A. indica candidate enzymes along with corresponding proteins identified from other plant species. Candidates are marked in bold.

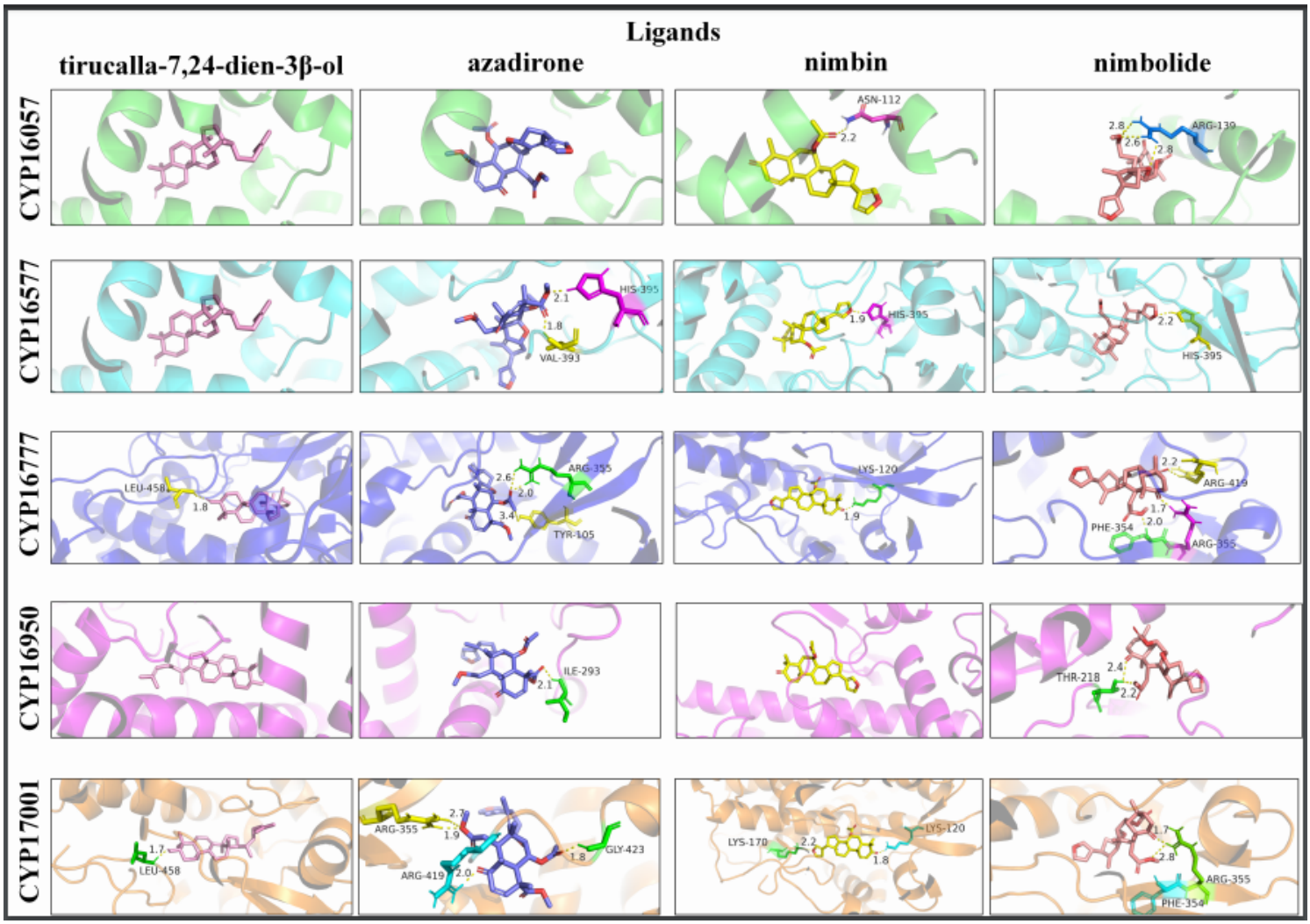

Figure 5

Molecular docking analysis of CYP450s putatively involved in azadirachtin A biosynthesis. Interaction of 5 CYP450s' docked regions with four ligands (tirucalla-7,24-dien-3β-ol, azadirone, nimbin, and nimbolide) are shown in the figure. CYP450s and the four ligands are represented in different colors. The zoomed blocks show the interacting orientation of functional moieties of tirucalla-7,24-dien-3β-ol, azadirone, nimbin, and nimbolide within the cavity of protein. The hydrogen bonds formed between residue and ligand are displayed in yellow and the distances (Å) are also shown.

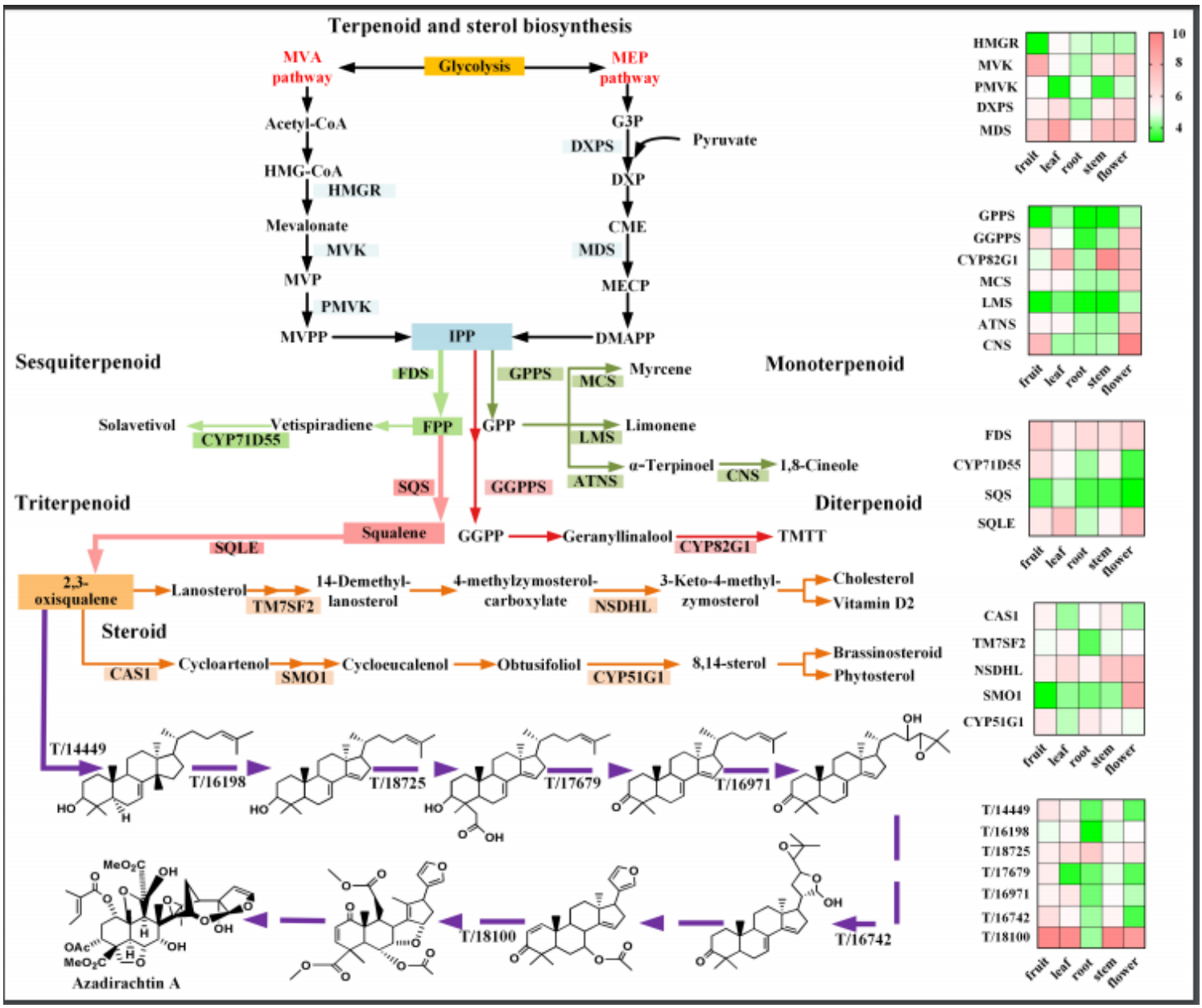

Figure 6

Mapping of unigenes related to secondary metabolites in A. indica. Tissue abundance and relative expression patterns are color coded and represented by bars, respectively. G3P: 3-phosphate glyceraldehyde; DXP: 1-Deoxy-D-xylulose 5-phosphate; MVP: 5-phosphomevalonate; MVPP: (R)-5-diphosphomevalonate; IPP: isopentenyl pyrophosphate; FPP: farnesyl pyrophosphate; GPP: geranyl pyrophosphate; GGPP: geranylgeranyl pyrophosphate; HMG-CoA: 3-hydroxy-3-methyl-glutaryl-CoA; PCME: 2-phospho-4-2-C-methyl-D-erythritol; MECP: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate; HMBDP: 1-hydroxy-2-methyl-2-butenyl 4-diphosphate; TMTT: (E,E)-4,8,12-trimethyltrideca-1,3,7,11-tetraene; DXPS: 1-deoxy-D-xylulose-5-phosphate synthase; MDS: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HMGR: Hydroxymethylglutaryl-CoA reductase; MVK: Mevalonate kinase; PMVK: Phosphomevalonate kinase; GPPS: Geranyl diphosphate synthase; MCS: Myrcene/ocimene synthase; LMS: (R)-limonene synthase; ATNS: (-)-alpha-terpineol synthase; CNS: 1,8-cineole synthase; TM7SF2: Delta(14)-sterol

reductase; NSDHL: Sterol-4-alpha-carboxylate 3-dehydrogenase; SMO1: Methylsterol monooxygenase 1; FDS: Farnesyl diphosphate synthase; GGPPS: Geranylgeranyl diphosphate synthase; SQS: Squalene synthase; SQLE: Squalene monooxygenase; CAS1: Cycloartenol synthase.

Supplementary Files

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

- Additionalfile1.xlsx
- FigureS1.pdf
- FigureS3.tif
- FigureS5.tif
- TableS1.xlsx
- TableS11.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS5.xlsx
- TableS4.xlsx
- TableS7.xlsx
- TableS9.xlsx
- TableS6.xlsx
- TableS10.xlsx
- FigureS2.tif
- TableS8.xlsx
- FigureS4.tif