Multi-tissue transcriptome analysis using hybrid-seq revealed potential genes and biological pathways associated with azadirachtin A biosynthesis in neem (Azadirachtin indica)

CURRENT STATUS: UNDER REVIEW

Huiyan Wang  
Beijing University of Technology  
ORCiD: https://orcid.org/0000-0002-1068-5854

Ning Wang  
Corresponding Author  
ORCiD: https://orcid.org/0000-0002-9228-6872

Yixin Huo  
Beijing Institute of Technology

DOI:  
10.21203/rs.2.23446/v1

SUBJECT AREAS  
Epigenetics & Genomics

KEYWORDS  
azadirachtin A, natural insecticides, secondary metabolism, triterpenoid biosynthesis, transcriptome, neem
Abstract
Background: Azadirachtin A is a triterpenoid from neem tree exhibiting excellent activities against over 600 insect species in agriculture. The manufacture of azadirachtin A depends on extraction from neem tissues, which is not ecofriendly and sustainable. The low yield and discontinuous supply impeded the further application. The biosynthetic pathway of azadirachtin A is still well-known.

Results: We attempted to explore azadirachtin A biosynthetic pathway and identified key involved genes by analyzing transcriptome data of five neem tissues through hybrid-seq (Illumina HiSeq and Pacific Biosciences Single Molecule Real Time (PacBio SMRT)) technology. A total 219 and 397 up-regulated differentially expressed genes (DEGs) in leaf and fruit tissues than other tissues (root, stem and flower) were isolated. After phylogenetic analysis and domain prediction, 22 candidates encoding 2,3-oxidosqualene cyclase (OSC), alcohol dehydrogenase (ADH), cytochrome P450 (CYP450), acyltransferase (ACT) and esterase (EST) proposed to be involved in azadirachtin A biosynthesis were finally selected. De novo assembled sequences were verified by Quantitative Real-Time PCR (qRT-PCR) analysis.

Conclusions: By integrating and analysis data from Illumina HiSeq and PacBio SMRT platform, 22 DEGs were finally selected as candidates involved in azadirachtin A biosynthesis. The obtained reliable and accurate sequencing data provided important novel information for understanding neem genome. Our data shed new light on the understanding of other triterpenoids biosynthesis in neem trees and provide reference for exploring other valuable natural product biosynthesis in plants.

Background
With the increasing concern on chemical pesticides threat to crop protection program globally, more attention is being paid towards bioactive, biodegradable plant or microbial based biopesticide instead of neurotoxic, broad spectrum, synthetic pesticides. Azadirachtin A, the major insecticidal ingredient in neem-based products, exhibits excellent activities against over 600 kinds of insect species [1] in agricultural area and can offer a non-toxic alternative to most synthetic pesticides. It is processed by insects as natural hormone and introduces antifeedant, repellent and growth inhibiting behavior [2]. The possibility of adaptive resistance in next-generation insects can be completely eliminated.
Furthermore, azadirachtin A-based pesticides are biodegradable, environmentally friendly and non-toxic to mammals, plants and birds. Due to these general superiorities, the agriculture segment accounted for the highest share (40%) in total revenues of neem extract products market which is expected to grow from 653 million dollars in 2015 to 1.8 billion dollars in 2022 with high annual growth rate of 16.3% (https://www.psmarketresearch.com/market-analysis/neem-extract-market).

Azadirachtin A current supply depends on extraction of neem leaf and seed [3]. The extraction process is complicated, time consuming and limited by neem distribution, and these further limited application of azadirachtin A.

Many attempts have been made to explore the complete biosynthesis of azadirachtin A in neem. However, its molecular architecture [4] complexity was the main obstacle that held back the advances in azadirachtin A biosynthesis. Chemical synthesis of azadirachtin A parts, tricyclic dihydrofuran [5] and ABC ring [6] were chemically synthesized. In 2007, the complete synthesis of azadirachtin A was finally accomplished after the 71 steps-reaction [7]. While the low productivity of 0.00015% still limited the application in practical industry.

Being a triterpenoid from neem, azadirachtin A biosynthesis pathway is still unclear after half hundred years investigation. The first attempt about its biosynthesis investigation was the feeding experiments [8] in 1971. $^3$H euphol and $^3$H tirucallol were used to incubate with neem leaf crude proving that euphol rather than tirucallol could be more efficiently incorporated into nimbolide (a limonoid whose carbon skeleton was similar to azadirachtin A). It suggested that tirucallol or euphol could be one of intermediates in azadirachtin A biosynthesis.

In order to find genes involved in azadirachtin A biosynthesis, cDNA library construction, expressed sequence tag library construction [9], genome sequencing [10, 11] and transcriptome sequencing [10, 12, 13] had been used. Comparative genomic analysis [10, 11] identified several unique genes in azadirachtin A-producing neem compared with Arabidopsis thaliana, Oryza sativa and Citrus sinensis. The transcriptome sequencing samples mainly used different neem tissues [14] or tissues in different developmental stages [13]. Genes involved in 2,3-oxidosqualene biosynthesis from neem tissues and genes encoding cytochrome P450 (CYP450) had been isolated [14].
The formation of scaffold synthesis from 2,3-oxidosqualene is an important step in azadirachtin A biosynthesis. However, enzyme participating in this step has not been reported yet. Recently, AiOSC1 [15] has been isolated from neem fruit transcriptome data that converting 2,3-oxidosqualene into tirucalla-7,24-dien-3β-ol which can be subsequently catalyzed into melianol by two CYP450s. Tirucalla-7,24-dien-3β-ol has similar molecular structure with tirucallol. This was the first functional characterized AiOSC1 reported to be related to azadirachtin biosynthesis.

Although three steps of azadirachtin A pathway had been figured out, the rest part of its biosynthetic way was still unexplored. Based on the numerous metabolites provided in Neem Metabolite Structure Database ([http://www.vmsrfdatabase.org/index.php](http://www.vmsrfdatabase.org/index.php)) as well as the distribution of compounds (Table S1) similar with tirucalla-7,24-dien-3β-ol, a putative biosynthetic pathway for azadirachtin A (Figure 1). Reactions such as hydroxylation, continuous oxidations, acyltransferation and esterification occurred in azadirachtin A biosynthesis. And four kinds of putative enzymes that participate in these reactions were CYP450, ADH, ACT and EST. CYP450 participating in biosynthesis of secondary metabolites catalyze hydroxylation, epoxidation, C-C bond formation and breakage in terpenoids [16]. Alcohol dehydrogenase was supposed to catalyze the ketone synthesis of C3 in azadirachtin A as it was supported by ADH1(artemisinic alcohol dehydrogenase). ADH1 from Artemisia annua was found to be involved in the biosynthesis of artemisinic aldehyde from artemisinic alcohol [17]. BAHD acyltransferase and esterase are enzymes which catalyze the formation of ester bond in plants metabolism [18]. As a consequence, these four enzymes were more likely to catalyze the reactions in azadirachtin A pathway.

In order to get candidate involved azadirachtin A downstream pathway, five neem tissues (root, stem, leaf, flower and fruit [19]) were sampled for transcriptome sequencing using Illumina HiSeq and Pacific Biosciences Single Molecule Real Time (PacBio SMRT) sequencing platform. Unigenes higher expressed in leaf and fruit compared with other tissues were selected as a library for mining genes in azadirachtin A pathway. Unigenes encoding 2,3-oxidosqualene cyclase (OSC), alcohol dehydrogenase (ADH), cytochrome P450 (CYP450), acyltransferase (ACT) and esterase (EST) were isolated. Extensive bioinformatics analysis of these unigenes further provided twenty-two candidates (Table 1) including
one OSC unigene, three ADH unigenes, ten CYP450 unigenes, two ACT unigenes and six EST unigenes. Qualitative real-time PCR and full-length cloning PCR were used for verifying unigenes expression level (FPKM) and transcript sequence accuracy. The obtained candidate genes could be used as an important resource to investigate the catalysts responsible for essential biochemical reactions in azadirachtin A biosynthesis as well as triterpenoids metabolism in close species of neem.

Methods
Plant materials
All fresh and healthy tissues (root, leaf, stem, flower and fruit containing seed) from neem (Azadirachtin indica, A. indica) tree were randomly sampled from the park of Hainan University, followed by transcriptome analysis on September 26, 2018. All samples for transcriptome sequencing were harvested from three plants for RNA extraction. 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. The A. indica was authenticated by Prof. Zhiwei Wang.

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 extracted 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 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 with an insert size of approximately 250 bp were constructed. All libraries were sequenced commercially to generate paired-end reads of average length 150 bp on the Illumina HiSeq 2000 sequencing platform (Hiseq 2000 V3) according to the manufacturer’s protocol by Beijing Genomics Institute (BGI-Shenzhen,
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). DNAs were selected by BluePippin™ and ranged into four sizes: 1-2, 2-3, 3-6 and 5-10 kb. DNAs after the second large-scale PCR were used as template prepared for SMRTbell library for sequencing. The amount of generated base was about 12 GB to cover all transcripts in sample. HiSeq reads were filtered by discarding the reads with adaptor sequences, reads with ambiguous “N” bases larger than 5% and low-quality reads. and 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 transcripts generation. The obtained transcripts were calibrated with filtered HiSeq reads using LSC software and subsequently filtered with CD-HIT-EST 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, Clusters of Orthologous Groups of proteins (COG), and Kyoto Encyclopedia of Genes and Genomes protein database (KEGG). Pfam annotation for the unigenes was finished using the HMMER 3.0 package. Sequence description for each unigene were transferred from homologous BLAST hits with E-value<10^{-4}. Gene Ontology (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 via. WEGO software. The distribution of gene functions was illustrated by biological process, cellular component and molecular function.

Clean reads were mapped to unigenes using Bowtie2 (v2.2.5), and then 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 per transcript per million mapped reads) method was used [20]. DEseq2 was used to identify DEGs (absolute value of log₂ fold-change≥1) after correction of p-values (adjusted p≤0.05) using Benjamini-Hochberg (false discovery rate, FDR≤0.001). Higher expressional unigenes characterized from leaf and fruit (azadirachtin A stimulating tissues) libraries were used for candidate mining.

**Analysis of phylogeny, domain architecture and subcellular localization**

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 all four kinds of candidate using Clustal W algorithm with default parameters. Phylogenetic analysis based on multiple alignments of protein sequences was done using the Neighbor Joining [21] method as implemented in MEGA7 and the phylogenetic trees were visualized on iTOL (https://itol.embl.de/). Accessions of these protein sequences used in phylogenetic analysis were provided in Additional file 1. The protein sequences of the candidates were also searched against Pfam database in order to get domain architecture information complementary to that provided by SwissProt. Subcellular localizations of them were obtained using TargetP-2.0 server (http:// www. cbs. dtu. dk / services /TargetP/).

**Quantitative real-time PCR and full-length cloning PCR validation of RNA-seq**

To validate the findings of RNA-seq data, ten transcripts (transcript/18900, transcript/18482, transcript/19291, transcript/17792, transcript/18186, transcript/18214, transcript/19882, transcript/19751, transcript/16577 and transcript/16950) were selected to confirm their expression in different tissues 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™ Reverse Transcription System (Promega, Canada). The reactions were performed in triplicate using 2 μL diluted cDNA template in a 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°C polymerase activation for 3 min, followed by 40 cycles of 95°C for 10s, then 60°C for 30s. The melting curve was obtained by heating the amplicon from 65°C to 95°C at increments of 0.5°C per 5s. The actin gene was used as an internal control to normalize all data. The relative quantitation (ΔΔCt) 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.

The obtained cDNA was also used to verify the accuracy of sequence obtained from PacBio data. Obtained cDNA (2 μL) was used in a 50 μL PCR reaction containing 2 μL of forward primer (10 μM), and 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 cell (TsingKe Biotech, Beijing, China) and the amplified constructs were sequenced by Genewiz Company (GENEWIZ, Beijing, China).

**Results And Discussions**

**Transcriptome analysis and data validation**

Hybrid-seq based omics analysis had made great advances in the fusion gene discovery [22] and genome assembling [23]. The high quality of transcriptome sequencing and assembly represent a great resource for the study of non-model species. Neem genome data and transcriptome data were obtained using next-generation sequencing (NGS) technology, followed by a genome draft improvement using PacBio assembly. Given the complexity of neem genome, and just the complexity of secondary metabolites synthesis in general, both long- and short-read technologies at the highest level of accuracy were needed in order to propel the level of discover the elusive mechanisms lies behind. The sensitive and high-throughput Illumina HiSeq technique can add long-reads capabilities with PacBio SMRT assistance. The accuracy achieved with PacBio SMRT essentially is on par with what was achieved with sequencing by synthesis short-read technology, which was critically important for
de novo sequencing and highly homologous regions of plant genomes. The precursors, intermediates and even azadirachtin A were predominantly distributed and differentially concentrated in the seed kernels and leaves of neem [24]. These secondary metabolites with plant defense activities were accumulated in order to survive the damage by pests during the critical period of fruit development and dry matter accumulation. Therefore, tissues of leaf and fruit are very likely to reveal most functional genes participated in the production of azadirachtin A or intermediates related to biosynthetic process.

To generate a comprehensive overview of neem transcriptome, total RNAs were extracted from leaves, fruits (containing seeds), roots, stems and flowers, then the mRNA was isolated, and cDNA libraries were established and sequenced separately using Illumina Hiseq platform, which generated 41.14, 41.35, 40.60, 40.59 and 40.64 million clean reads. PacBio SMRT platform were applied to mixed tissue sample (composed of these five tissues) to assist the short-read technology and 6.75 million clean reads were generated. The clean reads were individually generated with the average GC percentage of 40.99% and 43.55% by Illumina Hiseq and PacBio platforms. After calibration with long reads data generated by PacBio platform, short reads were assembled into 20201 good quality unigenes with an N50 of 5076 bp and mean length of 3607 bp (Table 2). The length of these unigenes ranged from 500 to 6001bp, and the majority (over 55.5%) were disturbed in 4501 bp and above (Figure 4a). However, there are still 8987 unigenes whose lengths were less than 4501 bp.

To verify the reliability of DEGs from transcriptome analysis, ten candidates were selected and validated by their expressions. Gene expressions determined using RNAseq and qRT-PCR were compared. Good concordance was observed between the qRT-PCR and RNAseq results (Figure 3), which indicated the high reliability of Illumina Hiseq data. To verify the reads accuracy from PacBio SMRT, these ten candidates were also selected to clone the sequences. The PCR sequencing results of selected unigenes showed that they were 100% identical to the sequences obtained by PacBio SMRT. Detailed sequencing results indicated that the generated unigenes in our study were of good quality and therefore reliable for further DEG analysis and functional annotation.

**Functional annotation and classification**
In order to analyze the function of these assembled unigenes, the unigenes were aligned against the Nr database, Nt database, SwissProt protein database, KEGG database, COG and GO database using the BLASTX program with an E-value cut-off of $10^{-5}$. In total, 19907 unigenes (98.54% of 20201 unigenes) were annotated in at least one database (Table S2). The annotated unigenes were compared to known nucleotide sequences of other plant species, which were best matched to the known nucleotide sequences from *Citrus sinensis* (52.43%), followed by *Citrus clementine* (23.49%), *Theobroma cacao* (2.44%), *Vitis vinifera* (2.4%) and other (19.23%).

COG and GO classification were used to further evaluate the completeness and effectiveness of the neem annotation. All 17634 unigenes (87.3% of 20201 unigenes) were classified into 25 functional Clusters of Orthologous Groups (COG) clusters (Figure 4b). Of those, 2610 unigenes (14.8% of the total 17634 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 (Figure 4c), they may play important roles in providing precursors in secondary metabolite biosynthesis.

In total, the 74905 assembled unigenes (66.3% of 113008 unigenes obtained in Hiseq sequencing analysis) were assigned at least one of the 55 GO terms (Figure 4b). Of those, these unigenes were predominantly assigned to the 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 KEGG pathway database was used to systematically evaluate the gene biological functions participated in different pathways. A total of 16778 unigenes were matched to the database and assigned to 135 KEGG pathways (Table S3). Metabolic pathways (3590 unigenes, 21.4%) and biosynthesis of second metabolites (1560 unigenes, 9.3%) were the two dominant categories. In the
biosynthesis of secondary metabolites category (Table S4) in neem, subcategories of flavonoid biosynthesis, terpenoid backbone biosynthesis, steroid biosynthesis, diterpenoid metabolism and carotenoid biosynthesis were included. There were 242 unigenes involved in the metabolism pathways of terpenoids and polyketides.

Based on the KEGG database analysis, a total of 63 unigenes (0.38% of 16778 unigenes with pathway annotation) were associated with terpenoid backbone biosynthesis (Table S5), which included 13 mevalonate (MVA) pathway unigenes and 38 methyerythritol phosphate (MEP) pathway or 1-deoxy-D-xylulose-5-phosphate (DXP) pathway unigenes (Figure 6). Among them, three unigenes encoded squalene synthase and squalene epoxidase, respectively. Transcript/13723 and transcript/18081 encoding squalene synthase and they were up-regulated in both fruit and leaf. Transcript/17638 was found expressing highest in fruit. For unigenes encoding squalene epoxidase, transcript/15581 expressed high in flower and leaf while transcript/15615 expressed much higher in fruit and root. Transcript/21847 expressed the highest in fruit while expressed the lowest in leaf.

As a triterpenoid, the first step of azadirachtin A biopathway was the formation of its scaffold catalyzed by OSC. However, there was no exact information about the scaffold or key enzyme catalyzing its formation in neem. Among all unigenes involved in terpenoids biosynthesis, eight detected unigenes were annotated as OSC, including transcript/1784, transcript/1866, transcript/8176, transcript/8892, transcript/9751, transcript/14584 and transcript/19700 and transcript/14449. Among them, only transcript/14449 expressed higher in fruit and leaf than in other three tissues. While after phylogenetic analysis (Figure S1) of these OSCs with several identified OSCs from other plants, transcript/14449 was grouped with AiOSC1[15], a newly characterized OSC from neem catalyzing the formation of tirucalla-7,24-dien-3β-ol, while other unigenes were grouped with cycloartenol synthase. After amino acid sequence analysis with AiOSC1, 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 azadirachtin A scaffold.

**Differential expression analysis of unigenes in neem**

In order to find candidate genes in azadirachtin A biosynthesis pathway, all transcripts had been
identified and annotated and mapped in different pathways. Genes expressed higher in azadirachtin A stimulating-organ leaf and fruit would be more likely to be the involved-genes in azadirachtin A biosynthesis. The expression level of unigenes was calculated by FPKM. The number of up-regulated DEGs in leaf and fruit comparing to other tissues were 219 and 397, respectively. All these DEGs were used for mining candidate involved in azadirachtin A biosynthesis.

**First screening of candidate genes involved in azadirachtin A biosynthesis**

From the putative azadirachtin A pathway in Figure 1, tirucalla-7,24-dien-3β-ol assumed as scaffold and formed from 2,3-oxidisqualene. After scaffold formation, a few important steps occurred such as hydroxylation and furan ring formation. Formation of hydroxyl group at C7 of azadirachtin A was critical for maintaining insecticidal activity [25]. The hydroxyl groups were either oxidized to acid or acylated by acyltransferase or esterified by esterase/lipase and then further formed limonoid compounds like azadirone or nimbin. After multiple steps such as C ring clearage and endoperoxide biosynthesis, azadirachtin A was finally obtained. In our study, enzymes ADH, CYP450, ACT and EST were supposed to be involved in azadirachtin pathway and their encoded unigenes were chosen for mining candidates. According to the annotation of DEGs and their expression levels in five tissues, DEGs up-regulated in leaf and fruit tissues were used for mining azadirachtin A biosynthetic candidates.

Among all the DEGs in fruit and leaf transcriptome data, sixteen DEGs encoding ADH were discovered, and four DEGs (transcript/19291, transcript/18833, transcript/18482 and transcript/22186) were up-regulated in fruit and leaf. As for CYP450, sixteen DEGs encoding CYP450 and ten higher expressing DEGs (transcript/17636, transcript/17854, transcript/16057, transcript/17284, transcript/16777, transcript/17001, transcript/16577, transcript/16950, transcript/17057 and transcript/16971) were identified. Similarly, thirteen and nineteen DEGs encoding ACT and EST were isolated and there were four and twelve DEGs up-regulating in leaf and fruit tissues, respectively. There are 30 DEGs up-regulated in leaf and fruit. Sequences under 150 amino acids were excluded. Finally, the resulted 26 DEGs were selected and their amino acid sequences were used for further screening phylogenetic analysis and domain prediction.
Further screening of candidates through phylogenetic analysis and domain prediction

Phylogenetic trees were constructed (Figure 5) with MEGA7. Protein domains (Table 1) were predicted and assigned with HMMSCAN. Transcript/22186 was grouped with CADH4 [26]. Transcript/18833 and transcript/19291 were grouped with CADH1. CADH1 and CADH4 are enzymes catalyzing biosynthesis of cinnamaldehyde from cinnamyl alcohol. Transcript/18482 was grouped with ADHX [27] which had activity to primary and secondary alcohols. Transcript/18833 and transcript/19291 contained PLN02514 domain which was also found in cinnamyl-alcohol dehydrogenase [28]. Transcript/22186 contained nsLTP2 [29] domain which was contained in non-specific lipid-transfer protein. Transcript/18482 contained GxGxxG motif [30] found in S-(hydroxymethyl) glutathione dehydrogenase.

According to the phylogenetic analysis of CYP450 DEGs, they were divided into three clades. Among them, transcript/17001, transcript/16777 were grouped with CYP71AV1, and transcript/16971 was grouped with CYP94B1. CYP71AV1 was involved in artemisin biosynthesis and catalyzed three continuous oxidations of amorpha-4,11-diene to produce artemisinic acid [31]. CYP94B1 [32] catalyzed the hydroxylation at C12 of jasmonyl-L-amino acid. Transcript/17284, transcript/17057, transcript/17636 and transcript/17854 were contained in second clade. Transcript/17057 fell into a group with CYP82C4 [33], an enzyme hydrolyzed xanthotoxin (8-methoxypsoralen) into 5-hydroxyxanthotoxin. Transcript/17284 grouped with CYP94B3 [34] and it revealed that transcript/17284 may act as a hydroxylase. Transcript/17636 and transcript/17854 were divided into a group with uncharacterized CYP98A1. Transcript/16950 and transcript/16577 were grouped with a hydroxylase CYP71D1 [35] and CYP72A14, respectively. Transcript/16057 grouped with CYP72A63 which catalyzed three sequential oxidations at C-30 of 11-oxo-beta-amyrin [36]. Through phylogenetic analysis, transcript/17001, transcript/16777 and transcript/16057 may act as oxidases, while the others were likely to be hydroxylases.

According to CYP450 domain analysis, transcript/17636 and transcript/16971 contained the same CypX domain [37] with CYP81B1. P450-cyclo_AA_1 domain in transcript/16950 was also found within cytokinin trans-hydroxylase [38]. PLN00168 domain [39] was found in transcript/17824 and
transcript/19854, and PLN02687 domain (a domain in flavonoid 3'-monooxygenase [40]) was also contained by transcript/16057, transcript/16577 and transcript/17057. Transcript/16777 and transcript/17001 contained PLN02774 domain, which was often found in brassinosteroid-6-oxidase [41]. Therefore, four DEGs (transcript/16057, transcript/16577 and transcript/17057 and transcript/17001) contained domains within CYP450 oxidases and the others contained domain within CYP450 hydroxylase

Among all ACT DEGs, transcript/18186 was grouped with ARE1, the enzyme encoding sterol O-acyltransferase [42]. Transcript/18214 fell into a subclade with LPAT2, an acyl-sn-glycerol-3-phosphate acyltransferase of Brassica oleracea [43]. Transcript/17792 was grouped with HCT2 which catalyzing the transfer of an acyl from p-coumaroyl-CoA to various acyl acceptors. Transcript/19132 and TSM1 [44] were in a group which reveal that transcript/19132 was likely to be methyltransferase. Through domain analysis, transcript/17792 and transcript/18214 contained HXXXD domain that often found in BAHD family [45]. Transcript/19132 contained same domain PLN02177 with glycerol-3-phosphate acyltransferase [46]. Transcript/18186 and eukaryotic initiation factor 4B [47] had the same eIF-4B domain.

For EST DEGs after phylogenetic analysis, transcript/19998 and transcript/16750 grouped with KAI2 [48]. KAI2 was reported to be involved in seed germination and didn’t show activity as esterase. Transcript/19188 was divided into a subclade with A. thaliana GDL15, which belonged to GDSL [49]-like lipase/acylhydrolase superfamily and displayed hydrolytic activity with esters. Transcript/18100 was grouped with PME3, a pectinesterase catalyzed the hydrolysis of (1,4)-α-D-galacturonosyl methyl ester [50]. Transcript/19751 and transcript/19748 formed a tight subclade with TGL1 which was a sterol esterase and mediated the hydrolysis of steryl esters [51]. Transcript/19882 and transcript/19697 were in a group with HIDH [52] and CXE18, respectively and these two enzymes shew activity to carboxylic esters. The conserved domain analysis of EST candidates presented transcript/19188 contained Ser-His-Asp (Glu) triad found in a SNGH plant lipase [53]. Transcript/19882, transcript/19697, transcript/19748 and transcript/19751 had AES domain [54] which contained by acetyl esterase/lipase. PAE [55] domain in transcript/18100 was also found in
pectin acetyl esterase while transcript/16750 and transcript/19998 contained the plant pectinesterase inhibitor domain PLN02201.

According to results from phylogenetic analysis and domain prediction, three ADH DEGs (transcript/22186 was excluded) and ten CYP450 DEGs were finally selected as candidates although there were some differences in results from the phylogenetic tree and domain prediction. As for ACT DEGs, transcript/18214 and transcript/17792 were selected as candidates. Among DEGs for EST candidates, transcript/16750 and transcript/19998 were excluded from the candidate pool. Finally, 22 DEGs were isolated and chosen as candidates involved in azadirachtin A biosynthesis.

**Measurement of unigenes expression in secondary metabolite pathways in neem**

The expressions of unigenes involved in secondary metabolites including three terpenoids, two sterols and putative azadirachtin A downstream pathway in five neem tissues were analyzed based on the KEGG annotation and FPKM (Figure 6). There are two pathways for the biosynthesis of terpenoids precursor IPP (isopentenyl pyrophosphate), one is MVA pathway, and the other is MEP pathway. Among all unigenes, 13 unigenes were found to be related to MVA pathway and 38 unigenes related to MEP pathway (Table S5). They encoded some enzymes involved in IPP biosynthesis in MVA and MEP pathway. 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 expressing higher in leaf and fruit comparing with the other three tissues.

GPP (geranyl pyrophosphate) was the common intermediate of all monoterpenoids and it formed by IPP catalyzed by Geranyl diphosphate synthase (GPPS). Unigenes encoding four enzymes in monoterpenoids myrcene, limonene and terpineol biosynthetic pathways expressed high in flower and fruit. As for diterpenoid (E, E)-4, 8, 12-trimethyltrideca-1, 3, 7, 11-tetraene (TMTT) biosynthesis, unigene encoding Geranylgeranyl diphosphate synthase (GGPPS) and CYP82G1 expressed higher in flower and stem followed by in fruit and leaf, respectively.

Expression pattern of unigenes involved in in sesquiterpenoid and triterpenoid were also measured. Farnesyl diphosphate synthase (FDS) catalyzed formation of farnesyl pyrophosphate (FPP) from IPP and unigene encoding FDS expressed highest in fruit followed by in root and flower. Solavetivol
pathway is the sesquiterpenoid pathway detected in neem and unigene encoding solavetivol synthase (CYP71D55) expressed higher in fruit and leaf. As for triterpenoids biosynthesis, unigenes involved in 2,3-oxidosqualene biosynthesis from FPP were measured. Two FPPs were catalyzed by squalene synthase (SQS) and then formed into 2,3-oxidosqualene through squalene epoxidase (SQLE). Unigenes encoding SQS and SQLE were found to express the highest in leaf.

Plants contain various kinds of terpenoids, polyketides and sterols. Campesterol, stigmasterol and β-sitosterol had also been successfully isolated from neem [56]. They are important sources for the production of heterocyclic compounds. In this study, unigenes encoding cycloartenol synthase (CAS1) were identified. The expression level of the unigenes in different tissues were in the order of fruit > stem > root. Unigenes encoding enzymes ((delta14-sterol reductase (TM7SF2) and CYP51G1)) respectively were relatively high in fruit or leaf while the expression level was not the same for unigene encoding methylsterol monooxygenase (SMO1) and sterol-4alpha-carboxylate 3-dehydrogenase (NSDHL).

The expression level of unigenes encoding enzyme in putative azadirachtin A pathway had been presented. Transcript/14449 encoded the first enzyme in azadirachtin A downstream pathway and it expressed highest in fruit followed by in leaf. After scaffold synthesis, the deletion of a methyl group at C14 occurred and catalyzed by enzyme encoded by transcript/16198. Transcript/16198 shew similar expression pattern with transcript/14449. Secondary alcohol at C3 was continuously oxidized into ketone group by transcript/18725 and transcript/17679, two unigenes expressed highest in fruit tissue. Intermediates with C3 ketone were more common in the isolated compounds [57]. And next important step involved in azadirachtin A was the formation of furan ring. Transcript/16971 and transcript/16742 encoded two CYP450s producing compound melianone [58] which was common in members of Meliaceae family such as Melia azedarach [59] and Melia toosenda [60]. With some uncharacterized enzymes, melianone was modified into intermediate with furan ring as well as the alcohol at C7. The C7-hydroxylased compound was further modified by esterase which encoded by unigenes transcript/18100 (highest-expressed in fruit and leaf) and finally became azadirachtin A under several reactions.
After analysis of the expression pattern of unigenes in secondary metabolites pathway in neem, the DEGs involved in these pathways were obtained. As a kind of terpenoid, the synthesis of azadirachtin A starts from IPP. The increase of precursor leads to the higher production of terpenoid. In the case of artemisinic acid production, improvement of terpenoid precursor by engineering MVA pathway made 500 time increases in yield [61]. Thus, the up-regulated DEGs in MVA or MEP pathway and 2,3-oxidosqualene biosynthesis from IPP would be used as building blocks in the construction of azadirachtin A precursor biosynthetic pathway in the microorganism.

Unigenes up-regulated in both leaf and fruit (azadirachtin A producing tissues) were selected as candidates. While some unigenes which expressed the highest either in leaf or in fruit were also selected for further characterization. For example, one unigene transcript/16198 encoded sterol 14-demethylase catalyzing deletion of methyl group of C14 in sterols. It was chosen as a candidate and supposed to catalyzed removal of methyl group at C14 of tirucalla-7,24-dien-3β-ol because of the structural similarity between sterols and tirucalla-7,24-dien-3β-ol.

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 couldn’t find the enzyme in hydroxylation reaction at C7 site from neem. What’s more, the order of reactions during the synthesis from azadirone to azadirachtin A was not clear. Neither the numbers of reactions nor the catalysis type were characterized. These limitations of pathway lead to insufficient mining on neem transcriptome data. This might be one of the reasons for slow progress in azadirachtin pathway exploration although numerous available neem genome and transcriptome data.

Conclusions
In conclusion, the multi-tissue transcriptome analysis revealed four types of genes potentially involved in azadirachtin A biosynthesis and their respective transcription levels in these tissues. It also indicated that the neem tree genome encodes a high number of terpene or limonoid biosynthetic genes. Finally, 22 unigenes (Table 1) encoding enzymes include OSC, ADH, CYP450, ACT and EST were selected as candidates involved in biosynthesis of azadirachtin A. This is the first report about
transcriptome profiling analysis of A. indica using hybrid-sEq. The obtained unigenes may provide a valid and diverse candidate pool for the study of the selective modification by the functional groups at triterpenoid or limonoids skeleton as well as the convergent evolution in secondary metabolism.

Additional Files

**Figure S1.** Phylogenetic tree of candidate OSC from neem transcriptome.

**Table S1.** Concentration of metabolites in leaf, bark and seed extract of A. indica.

**Table S2.** Summary of functional annotations of A. indica.

**Table S3.** KEGG annotation of unigenes from A. indica.

**Table S4.** Pathways and number of unigenes related to secondary metabolites in A. india.

**Table S5.** Discovery and expression of unigenes involved in terpenoid backbone biosynthesis in A. indica.

**Table S6.** Unigenes involved in the sesquiterpenoid and triterpenoid biosynthesis in A. indica.

**Additional file 1:** Details of proteins used for phylogenetic analysis of four kinds of candidates.

**Additional file 2:** Genes primers used for qRT-PCR analysis and full-length cloning PCR.

**Abbreviations**

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 per 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: methylerthritol 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: Delta14-sterol reductase; NSDHL: Sterol-4alpha-carboxylate 3-dehydrogenase; SMO1: Methylsterol monooxygenase 1; GGPPS: Geranylgeranyl diphosphate synthase; SQS: Squalene synthase; CAS1: Cycloartenol synthase.

Declarations

Acknowledgements

We thank Zhiwei Wang and Huangying Shu (Hainan Key Laboratory for Sustainable Utilization of Tropical Bioresources, College of Horticulture, Hainan University) for neem identifying and sampling.

Funding

This work was supported by National Key R&D Program of China(2017YFD0201400). The funders did not have any role in the design of the study, collection, analysis, or interpretation of data or in the writing of the manuscript.

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.

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.

References

1. Chaudhary S, Kanwar RK, Sehgal A, Cahill DM, Barrow CJ, Sehgal R, Kanwar JR. Progress on Azadirachta indica Based Biopesticides in Replacing Synthetic Toxic Pesticides. Front Plant Sci. 2017; 8:610-610.

2. Oulhaci CM, Denis B, Kilani-Morakchi S, Sandoz J-C, Kaiser L, Joly D, Aribi N. Azadirachtin effects on mating success, gametic abnormalities and progeny survival in Drosophila melanogaster (Diptera). Pest Manag Sci. 2018; 74(1):174-180.

3. 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.

4. Bilton JN, Broughton HB, Jones PS, Ley SV, Rzepa HS, Sheppard RN, Slawin AMZ, Williams DJ, Lidert Z, David Morgan E. An x-ray crystallographic, mass spectroscopic, and NMR study of the limonoid insect antifeedant azadirachtin and related derivatives. Tetrahedron. 1987; 43(12):2805-2815.

5. Ishihara J, Fukuzaki T, Murai A. Synthetic studies on azadirachtin (Part 3): Asymmetric synthesis of the tricyclic dihydrofuran moiety of azadirachtin. Tetrahedron Lett. 1999; 40(10):1907-1910.

6. Nicolaou KC, Sasmal PK, Roecker AJ, Sun X-W, Mandal S, Converso A. Studies toward the Synthesis of Azadirachtin, Part 1: Total Synthesis of a Fully Functionalized ABC Ring Framework and Coupling with a Norbornene Domain. Angew Chem, Int Ed. 2005; 44(22):3443-3447.
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. Krishnan NM, Pattnaik S, Jain P, Gaur P, Choudhary R, Vaidyanathan S, Deepak S, Hariharan AK, Krishna PB, Nair J et al. A draft of the genome and four transcriptomes of a medicinal and pesticidal angiosperm Azadirachta indica. BMC genomics. 2012; 13:464-464.

11. 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.

12. 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.

13. Pandreka A, Dandekar DS, Haldar S, Uttara V, Vijayshree SG, Mulani FA, Aarthy T, Thulasiram HV. Triterpenoid profiling and functional characterization of the initial genes involved in isoprenoid biosynthesis in neem (Azadirachta indica). BMC Plant Biol. 2015; 15:214-214.

14. Bhambhani S, Lakhwani D, Gupta P, Pandey A, Dhar YV, Kumar Bag S, Asif MH, Kumar Trivedi P. Transcriptome and metabolite analyses in Azadirachta indica: identification of genes involved in biosynthesis of bioactive triterpenoids. Sci Rep. 2017;
15. Hodgson H, De La Peña R, Stephenson MJ, Thimmappa R, Vincent JL, Sattely ES, Osbourn A. 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.

16. 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.

17. 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.

18. Legrand G, Delporte M, Khelifi C, Harant A, Vuylsteker C, Mörchen M, Hance P, Hilbert J-L, Gagneul D. Identification and Characterization of Five BAHD Acyltransferases Involved in Hydroxycinnamoyl Ester Metabolism in Chicory. Front Plant Sci. 2016; 7(741).

19. Kurimoto S-i, Takaishi Y, Ahmed FA, Kashiwada Y. Triterpenoids from the fruits of Azadirachta indica (Meliaceae). Fitoterapia. 2014; 92:200-205.

20. Hu Z, Li G, Sun Y, Niu Y, Ma L, He B, Ai M, Han J, Zeng B. Gene transcription profiling of Aspergillus oryzae 3.042 treated with ergosterol biosynthesis inhibitors. Braz J Microbiol. 2019; 50(1):43-52.

21. Zhang W, Sun Z. Random local neighbor joining: A new method for reconstructing phylogenetic trees. Molecular Phylogenetics and Evolution. 2008; 47(1):117-128.

22. Weirather JL, Afshar PT, Clark TA, Tseng E, Powers LS, Underwood JG, Zabner J, Korlach J, Wong WH, Au KF. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. Nucleic Acids Res. 2015; 43(18):e116-e116.

23. Ning G, Cheng X, Luo P, Liang F, Wang Z, Yu G, Li X, Wang D, Bao M. Hybrid
sequencing and map finding (HySeMaFi): optional strategies for extensively
deciphering gene splicing and expression in organisms without reference genome.
Sci Rep. 2017; 7:43793.

24. Rangiah K, Agasimundin VB, Gowda M. UHPLC-MS/SRM Method for Quantification of
Neem Metabolites from Leaf Extracts of Meliaceae Family Plants. Analytical Methods.
2016; 8(9):2020-2031.

25. Ley SV, Anderson JC, Blaney WM, Jones PS, Lidert Z, Morgan ED, Robinson NG,
Santafianos D, Simmonds MSJ, Toogood PL. Insect antifeedants from azadirachta
indica (part 5): Chemical modification and structure-activity relationships of
azadirachtin and some related limonoids. Tetrahedron. 1989; 45(16):5175-5192.

26. Kim S-J, Kim M-R, Bedgar DL, Moinuddin SGA, Cardenas CL, Davin LB, Kang C, Lewis
NG. Functional reclassification of the putative cinnamyl alcohol dehydrogenase
multigene family in Arabidopsis. Proc Natl Acad Sci U S A. 2004; 101(6):1455-1460.

27. 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.

28. 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.

29. 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.

30. 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.

31. Komori A, Suzuki M, Seki H, Nishizawa T, Meyer JJM, Shimizu H, Yokoyama S, Muranaka T. Comparative functional analysis of CYP71AV1 natural variants reveals an important residue for the successive oxidation of amorpha-4,11-diene. FEBS Lett. 2013; 587(3):278-284.

32. Koo AJ, Thireault C, Zemelis S, Poudel AN, Zhang T, Kitaoka N, Brandizzi F, Matsuura H, Howe GA. 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.

33. 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.

34. Heitz T, Widemann E, Lugan R, Miesch L, Ullmann P, Desaubry L, Holder E, Grausem B, Kandel S, Miesch M 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.

35. Sun J, Zhao L, Shao Z, Shanks J, Peebles CAM. Expression of tabersonine 16-hydroxylase and 16-hydroxytabersonine-O-methyltransferase in Catharanthus roseus hairy roots. Biotechnol Bioeng. 2018; 115(3):673-683.

36. Seki H, Sawai S, Ohyama K, Mizutani M, Ohnishi T, Sudo H, Fukushima EO, Akashi T, Aoki T, Saito K et al. Triterpene Functional Genomics in Licorice for Identification of CYP72A154 Involved in the Biosynthesis of Glycyrrhizin. The Plant Cell. 2011; 23(11):4112-4123.

37. 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.

38. 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.

39. 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.

40. Rice Annotation P, Itoh T, Tanaka T, Barrero RA, Yamasaki C, Fujii Y, Hilton PB, Antonio BA, Aono H, Apweiler R et al. Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. Genome Res. 2007; 17(2):175-183.

41. Shimada Y, Fujioka S, Miyauchi N, Kushiro M, Takatsuto S, Nomura T, Yokota T, Kamiya Y, Bishop GJ, Yoshida S. Brassinosteroid-6-Oxidases from Arabidopsis and Tomato Catalyze Multiple C-6 Oxidations in Brassinosteroid Biosynthesis. Plant Physiol. 2001; 126(2):770-779.

42. 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.

43. Kim HU, Li Y, Huang AHC. Ubiquitous and Endoplasmic Reticulum–Located Lysophosphatidyl A cyltransferase, LPAT2, Is Essential for Female but Not Male Gametophyte Development in Arabidopsis. The Plant Cell. 2005; 17(4):1073-1089.

44. 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.
45. 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 Physiol. 2009; 151(4):1773-1789.

46. 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.

47. 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.

48. 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.

49. 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 genomics. 2012; 13:309-309.

50. 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.

51. 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.

52. 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.
53. Mølgaard A, Kauppinen S, Larsen S. Rhamnogalacturonan acetyesterase elucidates the structure and function of a new family of hydrolases. Structure. 2000; 8(4):373-383.

54. 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.

55. Philippe F, Pelloux J, Rayon C. Plant pectin acetyesterase structure and function: new insights from bioinformatic analysis. BMC Genomics. 2017; 18(1):456.

56. Momchilova S, Antonova D, Marekov I, Kuleva L, Nikolova-Damyanova B, Jham G. Fatty Acids, Triacylglycerols, and Sterols in Neem Oil (Azadirachta Indica A. Juss) as Determined by a Combination of Chromatographic and Spectral Techniques. J Liq Chromatogr Relat Technol. 2007; 30(1):11-25.

57. Paal C. Ueber die Derivate des Acetonphenonacetessigesters und des Acetonylacetessigesters. Ber Dtsch Chem Ges. 1884; 17(2):2756-2767.

58. 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.

59. Ntalli NG, Cottiglia F, Bueno CA, Alché LE, Leonti M, Vargiu S, Bifulco E, Menkissoglu-Spiroudi U, Caboni P. Cytotoxic Tirucallane Triterpenoids from Melia azedarach Fruits. Molecules. 2010; 15(9):5866-5877.

60. Zhang Y, Tang C-P, Ke C-Q, Yao S, Ye Y. Limonoids and Triterpenoids from the Stem Bark of Melia toosendan. Journal of Natural Products. 2010; 73(4):664-668.

61. Paddon CJ, Westfall PJ, Pitera DJ, Benjamin K, Fisher K, McPhee D, Leavell MD, Tai A, Main A, Eng D et al. High-level semi-synthetic production of the potent antimalarial artemisinin. Nature. 2013; 496:528.
## Tables

### Table 1. Summary of candidates involved in azadirachtin A biosynthesis in neem

| Classification | Unigene    | NCBI          | Family       | Domain/Motif |
|----------------|------------|---------------|--------------|--------------|
| OSC            | transcript/14449 | gi|443299067   | terpene cyclase | SQCY_1       |
| ADH            | transcript/18833  | gi|572153023   | ADH_N        | PLN02514     |
|                | transcript/19291  | gi|572153023   | ADH_N        | PLN02514     |
|                | transcript/18482  | gi|567902424   | MDR          | GxGxxG       |
| CYP450         | transcript/17636  | gi|225458053   | P450-cycloAA_1 | CpyX        |
|                | transcript/17854  | gi|641826901   | p450 superfamily | PLN00168    |
|                | transcript/16057  | gi|567902124   | p450 superfamily | PLN02687    |
|                | transcript/17284  | gi|567889747   | p450 superfamily | PLN00168    |
|                | transcript/16777  | gi|590722535   | p450 superfamily | PLN02774    |
|                | transcript/17001  | gi|590722535   | p450 superfamily | PLN02774    |
|                | transcript/16577  | gi|568834016   | p450 superfamily | PLN02687    |
|                | transcript/16950  | gi|645239614   | P450-cycloAA_1 | PLN02290    |
|                | transcript/17057  | gi|567868115   | p450 superfamily | PLN02687    |
|                | transcript/16971  | gi|568825869   | P450-cycloAA_1 | CpyX        |
| ACT            | transcript/17792  | gi|567873443   | transferase   | HXXXD        |
|                | transcript/18214  | gi|641830965   | transferase   | HXXXD        |
| EST            | transcript/19188  | gi|568822600   | SGNH-hydrolase | Ser-His-Asp (Glu) |
|                | transcript/19882  | gi|568867507   | Aes superfamily | AES         |
|                | transcript/19697  | gi|225440163   | Aes superfamily | AES         |
|                | transcript/19748  | gi|641846434   | Aes superfamily | AES         |
|                | transcript/19751  | gi|567898564   | Aes superfamily | AES         |
|                | transcript/18100  | gi|568835134   | PAE superfamily | PAE         |

Tissue flower, fruit, leaf, root and stem were represented by f, F, L, R and S.

### Table 2. Summary of Illumina Hiseq and PacBio data output quality and assembled sequences of six libraries of *Azadirachta indica*
| Sequencing Platform | Sample | Reads Number (M) | Clean Bases (G) | Q20 (%) | Q30 (%) |
|---------------------|--------|-----------------|----------------|---------|---------|
| Illumina HiSeq      | Leaf   | 41.14           | 6.17           | 96.95   | 92.30   |
|                     | Flower | 41.35           | 6.20           | 97.29   | 93.06   |
|                     | Stem   | 40.60           | 6.09           | 97.29   | 93.11   |
|                     | Fruit  | 40.59           | 6.09           | 97.45   | 93.39   |
|                     | Root   | 40.64           | 6.10           | 97.44   | 93.43   |
| Hiseq summary       | —      | —               | —              | —       | —       |
| PacBio SMRT         | Mixed tissue | 6.75     | 11.28          | —       | —       |
| Calibrated PacBio by Illumina | — | — | — | — | — |

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 route 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 was connected by solid lines and arrows and putative biosynthetic pathway of azadirachtin A was shown by dashed lines and arrows. Two kinds of isomers were numbered in red.
Figure 2

The work-flow of gene mining. Letters L, F, R, S and f represent leaf, fruit, root, stem and flower, respectively.
Figure 3

Real time quantitative PCR (RT-qPCR) validation of selected 10 unigenes from fruit, leaf,
root, stem and flower tissues of A. indica. The relative expression level of each selected
gene was determined by 2-ΔΔCT. Experiments have been conducted in triplicates.

Figure 4

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 mix tissue subtracted cDNA library.
Phylogenetic analysis of alcohol dehydrogenase, CYP450, acyltransferase and esterase candidates of A. indica. Neighbor-joining trees were constructed for four types of A. indica candidate enzymes along with corresponding identified proteins from other plant species. Candidates are marked in bold. a) alcohol dehydrogenase. b) CYP450. c) acyltransferase. d) esterase.
Figure 6

Mapping of unigenes related to secondary metabolites in A. indica. Tissue abundance and relative expression patterns are represented by color code and 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; 1-deoxy-D-xylulose-5-phosphate synthase (DXPS); 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS); Hydroxymethylglutaryl-CoA reductase (HMGR); Mevalonate kinase (MVK); Phosphomevalonate kinase (PMVK); Geranyl diphosphate
synthase (GPPS); Myrcene/ocimene synthase (MCS); (R)-limonene synthase (LMS); (-)-alpha-terpineol synthase (ATNS); 1,8-cineole synthase (CNS); Delta14-sterol reductase (TM7SF2); Sterol-4alpha-carboxylate 3-dehydrogenase (NSDHL); Methylsterol monooxygenase 1 (SMO1); Farnesyl diphosphate synthase (FDS); Geranylgeranyl diphosphate synthase (GGPPS); Squalene synthase (SQS); Squalene monooxygenase (SQLE); Cycloartenol synthase (CAS1).

Supplementary Files

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

Table S5.xlsx
Table S6.xlsx
Additional file 1.xlsx
Additional file 2.xlsx
Table S2.xlsx
Table S3.xlsx
Table S4.xlsx
figureS1 (1).pdf
Table S1.xlsx