Comparative transcriptome analysis of root, stem, and leaf tissues of *Entada phaseoloides* reveals potential genes involved in triterpenoid saponin biosynthesis

Weifang Liao (leesalwf89@126.com)
Wuhan Polytechnic University https://orcid.org/0000-0002-2239-566X
Zhinan Mei
South-Central University for Nationalities
Lihong Miao
Wuhan Polytechnic University
Pulin Liu
Wuhan Polytechnic University
Ruijie Gao
Wuhan Polytechnic University

Research article

Keywords: Entada phaseoloides, Transcriptome, Triterpenoid saponins, Secondary metabolites

Posted Date: September 3rd, 2020

DOI: https://doi.org/10.21203/rs.2.20018/v3

Version of Record: A version of this preprint was published at BMC Genomics on September 15th, 2020.
See the published version at https://doi.org/10.1186/s12864-020-07056-1.

Abstract

Background Entada phaseoloides (L.) Merr. is an important traditional medicinal plant. The stem of Entada phaseoloides is popularly used as traditional medicine because of its significance in dispelling wind and dampness and remarkable anti-inflammatory activities. Triterpenoid saponins are the major bioactive compounds of Entada phaseoloides. However, genomic or transcriptomic technologies have not been used to study the triterpenoid saponin biosynthetic pathway in this plant. Results We performed comparative transcriptome analysis of the root, stem, and leaf tissues of Entada phaseoloides with three independent biological replicates and obtained a total of 53.26 Gb clean data and 116,910 unigenes, with an average N50 length of 1218 bp. Putative functions could be annotated to 42,191 unigenes (36.1%) based on BLASTx searches against the Non-redundant, Uniprot, KEGG, Pfam, GO, KEGG and COG databases. Most of the unigenes related to triterpenoid saponin backbone biosynthesis were specifically upregulated in the stem. A total of 26 cytochrome P450 and 17 uridine diphosphate glycosyltransferase candidate genes related to triterpenoid saponin biosynthesis were identified. The differential expressions of selected genes were further verified by qPT-PCR. Conclusions The dataset reported here will facilitate the research about the functional genomics of triterpenoid saponin biosynthesis and genetic engineering of Entada phaseoloides.

Background

Entada phaseoloides (L.) Merr. is a liana belonging to Fabaceae family. It grows in Southern China and other tropical countries. The stem of *Entada phaseoloides* is popularly used in traditional medicine because of its significant pharmacological activities [1-3]. The stem of *Entada Phaseoloides,* also called “Guo Gang Long,” produces curative effects that dispel wind and dampness and exhibits remarkable anti-inflammatory activity. Its main bioactive ingredients are triterpenoid saponins compounds [3]. Various types of triterpene saponins have been isolated from *E. phaseoloides*. The representative saponins of *Entada phaseoloides* are oleanane-type triterpene saponins which contain seven sugar chains.

The mevalonic acid (MVA) pathway is an important metabolic pathway in plants [4, 5]. Triterpenoid saponins comprise six isoprene units and are derived from a C-30 hydrocarbon precursor, squalene. Squalene is synthesized from isopentenyl diphosphate (IPP) via the MVA pathway. Subsequently, squalene epoxidase (SQE) catalyzes the conversion of squalene to 2,3-oxidosqualene. The diversifying step in triterpenoid backbone biosynthesis is the cyclization of 2,3-oxidosqualene catalyzed by a class of oxidosqualene cyclases (OSCs) [6, 7]. Cytochrome P450 monooxygenases (CYP450s) and UDP-glycosyltransferases (UGTs) govern the hydroxylation, oxidation, and glycosylation steps, yielding triterpenoid saponins [8-10]. However, the key genes related to triterpenoid saponin biosynthesis in *Entada phaseoloides* have not been identified.

High-throughput sequencing analysis is a useful method to clarify the molecular mechanism of plant secondary metabolism [11, 12]. Recently, transcriptome assay with next-generation sequencing has been extensively used to explore the novel genes underlying active-ingredient biosynthesis pathways in

medicinal plants. Some include the excavation of genes encoding enzymes that catalyze distinct steps related to the biosynthetic pathway of ginsenosides in *Panax ginseng* [13], triterpenoid saponin biosynthesis in *Bacopa monnieri* [14], artemisinin in *Artemisia annua* [15, 16], flavonoid biosynthesis in safflower [17], glycyrrhizin in *Glycyrrhiza uralensis* [18], rubber in *Parthenium argentatum* [19], cardiac glycoside in *Calotropis procera* [20], terpenoid in *Cinnamomum camphora* [21], cannabinoids in *Cannabis sativa* [22], withanolide in *Withania somnifera* [23], picrosides in *Picrorhiza kurrooa* [24], paclitaxel in *Taxus chinensis* [25] and steroidal saponins in *Asparagus racemosus* [26].

Because the synthesis and accumulation of specific metabolites in different tissues depends on the age of the plant, and are greatly affected by the different developmental stages. It was found that the content of triterpenes accumulated in the leaves of *P. ginseng* are higher in the early growth stage, while the content of triterpenes in the roots of old plants are higher [27]. Comparative transcriptome analysis including root and leaf tissues to excavate transcripts related to saponin biosynthesis is already reported for many plants, such as *Hedera helix* [28], *Panax notoginseng* [29] and *Asparagus racemosus* [26]. The maximum triterpenoid saponin content was identified in stem. So, in this study, comparative transcriptome analysis of root, stem, and leaf tissues of *Entada phaseoloides* was performed to identify genes related to triterpenoid saponin. We obtained thousands of putative genes, including a series of genes related to triterpene saponin biosynthesis. Moreover, different expression patterns of CYP450s and UGTs in the three tissues were analyzed. This work was established to functionally research the genes related to triterpene saponin biosynthesis and provide more information about this species.

Results

Illumina sequencing and *de novo* assembly

To characterize the transcriptomes of *Entada phaseoloides*, we sequenced nine cDNA libraries prepared from the root, stem, and leaf tissues with three biological repeats by using the Illumina Hiseq 2500 platform. A total of 53.26 Gb clean data were obtained after removing adaptors, poly-A tails, and primer sequences, short (< 50 bp), and low-quality sequences. A total of 57–60 million for each tissue were generated (Additional file 1). The high-quality reads were assembled using the Trinity program [30] and the TGI clustering tool (TGICL) [31] to remove redundant sequences. Finally, 116,910 unigenes were identified, with an average N50 length of 1,218 bp (Additional file 1). The correlation indices between repeated samples were > 0.9 (Additional file 2), indicating that the Illumina sequencing results are credible.

Functional annotation

All assembled unigenes were searched against the Non-redundant (Nr), Uniprot, Kyoto Encyclopedia of Genes and Genomes (KEGG), Pfam, Gene Ontology (GO), and Clusters of Orthologous Groups (COG) databases using the BLASTx program with *E*-value < 1e-5. Among the 116,910 sequences, 42,191

(36.1%), 41,228 (35.3%), 28,126 (24.1%), 26,874 (23.0%), 15,119 (13.0%) and 11,812 (10.1%) unigenes showed significant similarity to known proteins in NR, Uniprot, GO, Pfam, KEGG and COG database, respectively. The result of BLASTX with different databases and their annotation were listed in Additional file 3. Based on the Nr database, the E-value distribution indicated that 70.67% of the matched unigenes ranged from 1e-5 to 1e-100 (Fig. 1a). For the similarity distribution, 46.43% unigenes exhibited a similarity of above 80%, whereas 50.79% of the unigenes showed a similarity of 40%–80% (Fig. 1b). Furthermore, 14.63% of the *Entada phaseoloides* unigenes shared high similarity with the genes of *Cicer arietinum*, 13.46% similarity with *Cajanus cajan*, 12.84% similarity with *Glycine max*, and 9.07% similarity with *Medicago truncatula* (Fig. 1c).

GO analysis included three main domains that describe biological processes, cellular components, and molecular functions. When GO was used to classify gene functions, 28,126 unigenes were assigned to 60 functional categories (Additional files 4 and 5). Within the biological process domain, the three most enriched categories were “biosynthetic process,” “cellular nitrogen compound metabolic process,” and “response to stress.” In the cellular component domain, the three most matched categories were “cellular component.” “nucleus,” and “protein complex.” In the molecular function domain, the three most common categories were “ion binding,” “molecular function,” and “kinase activity.”

To better understand the functions of specific metabolic pathways in *Entada phaseoloides*, we mapped the annotated unigenes to the reference biological pathways in the KEGG database. A total of 15,119 unigenes (13.0 %) could be assigned to five main categories and 31 sub-categories (Fig. 2, Additional file 6). These enzymes feature assigned functions in 28 secondary metabolic pathways in KEGG (Table 1). Among these unigenes, 147 encode key enzymes are related to the pathways for terpenoid biosynthesis, including the synthesis of the terpenoid backbone (63 unigenes), monoterpenoids (7 unigenes), diterpenoids (18 unigenes), sesquiterpenoids and triterpenoids (16 unigenes), and other terpenoid-quinone complexes (43 unigenes). Fifty unigenes are involved in alkaloid biosynthesis, including isoquinoline alkaloid (24 unigenes) and tropane, piperidine, and pyridine alkaloid biosynthesis (26 unigenes). Exactly 210 unigenes were associated with the flavonoid biosynthesis pathway, including the phenylpropanoid (162 unigenes), flavonoid (36 unigenes), flavone and flavonol (7 unigenes), and isoflavonoid (5 unigenes) biosynthesis pathways. Unigenes involved in these pathways should be further identified to understand their functions in the biosynthesis of active ingredients in leguminous plants.

Differentially expressed gene (DEG) analysis

The clean reads were mapped back onto the assembled unigenes by using the alignment via Burrows–Wheeler aligner (BWA) program to analyze the DEGs among different tissues [32]. The Fragments per Kilobase Million (FPKM) value was calculated for each unigene in each tissue of *Entada phaseoloides*. The DEGs were identified (Additional file 7) using FDR≤0.001 and |log2Ratio|≥1 [33]. The lowest number of DEGs was observed between the stem and leaf tissues, and the highest was noted between the root and leaf. Furthermore, the DEGs in one tissue were studied and compared with those in the other two

tissues. The stem contained the largest number of highly expressed unigenes, having 8,962 unigenes more abundant in the stem. Fig. 3 shows the other expression differences among various tissues.

KEGG enrichment analyses were performed with the DEGs among the root, stem, and leaf tissues to investigate the genes regulating the distribution of triterpenoid saponin. These DEGs were evidently enriched in specific pathways. Meanwhile, the top 20 significant pathways were analyzed based on the FDR≤0.01. Between the leaf and stem, plant hormone signal transduction, phenylpropanoid biosynthesis, photosynthesis, terpenoid backbone biosynthesis, and phosphatidyllnositol signaling system showed significant enrichment (Fig. 4a). Between the leaf and root, plant hormone signal transduction, phenylpropanoid biosynthesis, photosynthesis, ubiquinone and other terpenoid-quinone biosynthesis, and cyanoamino acid metabolism pathways showed visible differential expression (Fig. 4b). Between the stem and root, plant hormone signal transduction, phenylpropanoid biosynthesis, photosynthesis, cyanoamino acid metabolism, and terpenoid backbone biosynthesis showed significant enrichment (Fig. 4c). In addition to the common pathways of primary metabolism, enriched secondary metabolic pathways, including terpenoid and phenylpropanoid biosynthesis, were also found between different tissues, indicating the possible distinct distribution of secondary metabolites in different tissues.

Putative genes involved in triterpenoid saponin backbone biosynthesis

Triterpenes are synthetized from a five-carbon isoprene unit through the cytosolic MVA pathway. Triterpenoid saponins are composed of six isoprene units and are derived from the C-30 hydrocarbon precursor, squalene. Squalene is synthesized from isopentenyl diphosphate (IPP) via the MVA pathway. All genes encoding the enzymes associated with the upstream regions of triterpenoid biosynthesis were successfully detected in the *Entada phaseoloides* transcriptome. Their expression value was monitored in three biological replicates along with their mean values (Table 2, Fig. 5). Most unigenes related to MVA pathway were specifically upregulated in the stem tissue. Hydroxymethylglutaryl-CoA reductase showed the highest expression, which is the rate limiting step MVA pathway for saponin biosynthesis.

The diversifying step in triterpenoid backbone biosynthesis is the cyclization of 2,3-oxidosqualene catalyzed by a class of OSCs. The major saponins in *Entada phaseoloides* are oleanane-type triterpenoid saponins derived from β-amyrin. The Illumina sequencing of *Entada phaseoloides* revealed 21 OSC sequences, among which eight unigenes were putative β-amyrin synthases. A full-length OSC sequence (EpBAS) with high identity to β-amyrin synthase was obtained (Additional file 8). The EpBAS cDNA included a 2,289 bp full open reading frame fragment. The deduced amino acid sequence of EpBAS (762 amino acids) shared 89.37% and 89.34% similarity with β-amyrin synthase in *Abrus precatorius* (ApBAS) and GiBAS in *Glycyrrhiza inflata* (Fig. 6), respectively. The relatively high similarities of the EpBAS protein with other β-amyrin synthases suggest that this gene encodes β-amyrin synthase in *Entada phaseoloides*.

CYP450s and UGTs

Earlier studies suggested that CYP450s and UGTs may account for the biosynthesis and accumulation of triterpene saponins in specific organs [34]. Tissue-specific transcriptome analysis of *Entada phaseoloides* suggests that the enzymes involved in triterpenoid saponin backbone are present in all the three tissues. Based on DEG analysis using transcriptome data, there is the possibility of further modifications such as oxidation and glycosylation using CYP450s and UGTs occur in the stem. Although the enzymes related to precursor biosynthesis are also present in root and leaf tissues which suggest the involvement of all the three tissues in the metabolic pathway. However, the metabolic analysis has indicated that it is the stem which mostly contains higher triterpenoid saponin content and utilized widely for its excellent pharmacological activity. In this study, in total of 326 CYP450s and 148 UGTs were found. Among the DEGs, 26 CYP450s and 17 UGTs were upregulated in the stem compared with the root and leaf tissues (Fig. 7a and b).

qRT-PCR validation of candidate genes involved in triterpenoid saponin biosynthesis

To verify the expression profiles obtained from Illumina sequencing, we performed qRT-PCR on nine selected genes related to triterpene saponin biosynthesis (Fig. 8). Consistent with the Illumina data, most of these genes showed strong expression levels in the stem compared with the root and leaf, and acetyl-CoA acetyltransferase, hydroxymethylglutaryl-CoA synthase, hydroxymethylglutaryl-CoA reductase and SQE genes were expressed abundantly. The expression fold changes were also close to the RNA-seq results. qRT-PCR results indicate that the RNA-seq data in this studty were reliable.

Discussion

Entada phaseoloides is an important traditional medicinal plant with various pharmaceutical activities. Although this plant is pharmacologically important, its genomic or transcriptomic information is highly limited. In NCBI, only 38 protein sequences are accessible for *Entada phaseoloides*. We revealed the comparative transcriptome analysis of the root, stem, and leaf tissues of *Entada phaseoloides*. The dataset reported here is useful in understanding the biosynthetic pathway of pharmacodynamic triterpenoid saponin and genetic engineering of this species.

In this study, a total of 53.26 Gb clean data were generated from nine RNA-seq libraries of the root, stem, and leaf. De novo assembly acquired 116,910 unigenes, with an average N50 length of 1,218 bp, which is similar to that of previously reported non-model plants, such as *Raphanus sativus* [35] and *Isodon Amethystoides* [36]. The best match for each unigene search against the Nr and KEGG databases was of help to assign GO functional annotation under biological process, cellular component, and molecular function categories. The varied GO assignments to unigenes represented the possible assortment of genes in the *Entada phaseoloides* transcriptome. Several unigenes mapped onto KEGG are related to distinct secondary metabolic pathways. Most unmatched unigenes are short sequence proteins with no

domain, untranslated regions, non-coding RNA or assembly mistakes. In support of the annotation, all the unigenes encoding enzymes related to the upstream regions of the MVA pathway for saponin biosynthesis from acetyl-CoA to squalene were found.

SQE enzymes catalyze the oxidation of squalene to 2,3-oxidosqualene. In our transcriptomic analysis, sequences encoding SQE represented the highest number (16) of unigenes associated with the MVA pathway. Single copies of SQE were identified in mouse and yeast, and the destruction of SQE in these species is lethal [37]. However, two or more copies of SQE are usually found in plants. Hwang et al. [38] examined 17 SQE sequences in *Eleutherococcus senticosus.* In *Arabidopsis thaliana*, six SQE enzymes have been identified, and three of them encode functional SQEs [39]. The expression of PgSQE1 regulates the biosynthesis of ginsenoside in *Panax ginseng* [40]. Thus, SQE is possibly an important enzyme in the saponin biosynthetic pathway. The SQE enzyme responsible for the saponin biosynthesis in the 16 SQE sequences in *Entada phaseoloides* remains to be identified.

The cyclization of 2,3-oxidosqualene is a branch point of saponin synthesis. The major saponins in the stem of *Entada phaseoloides* are oleanane-type triterpenoids. Oleanolic acid sapogenin was derived from β-amyrin after hydroxylation by CYP450s and glycosylation by UGTs [41, 42]. A total of 8 β-amyrin synthase, 326 CYP450, and 148 UGT sequences were observed in our transcriptome. The high expression of β-amyrin synthase, an important enzyme related to triterpenoid sapogenin biosynthesis at later stages, further reveals the high concentration of sapogenins in the stem of *Entada phaseoloides*. Similar tissue-specific concentrations of triterpenoid sapogenins have already been reported in other plants [43-45]. Moreover, 26 CYP450s and 17 UGTs were found to be upregulated in the stem. Further characterization of these candidate enzymes is needed to confirm the pathway of triterpenoid saponin biosynthesis in *Entada phaseoloides.*

Conclusions

In the present study, the comparative transcriptome analysis of root, stem and leaf tissues of *Entada phaseoloides* was performed to investigate the putative genes involved in triterpenoid saponin biosynthetic pathway of an important medicinal plant. The differential expression pattern of pathway genes suggest tissue-specific synthesis. The identified data will help the further discovery and functional genomics and transcriptomics analysis of *Entada phaseoloides*.

Methods

Plant materials

Three-year-old healthy wild-type Entada phaseoloides plants were collected from the experimental farm of South China Botanical Garden, Guangzhou City, Guangdong Province, P.R. China, in May 2019. After cleaning with ultrapure water, the roots, stems, and leaves were collected separately, immediately frozen in liquid nitrogen, and stored at −80 °C.

RNA extraction, cDNA synthesis, and sequencing

Total RNA from approximately 1.0 g of each tissue was extracted using TRIzol (Invitrogen, Canada) following the manufacturer's instructions. Three replicates were employed for each experiment. mRNA was isolated from total RNA by using Oligo(dT) magnetic beads. By mixing with fragmentation buffer, the mRNA was broken into short fragments.

The short fragments were purified and resolved with EB buffer. After end repair and single base "A"addition, adapters were ligated to the cDNA molecules. To select suitable cDNA fragments for PCR amplification, we purified the sample library with the AMPure XP system (Beckman Coulter, USA). Finally, PCR products were purified, and library quality was were determined using an Agilent Bioanalyzer 2100 system (Agilent Technologies, USA) and a Qubit 3.0 fluorometer (Invitrogen, USA). Each cDNA library was sequenced in a single lane of the Illumina Hiseq 2500 platform.

Data filtering and *de novo* assembly

The raw reads were first filtered to exclude the reads containing adaptors or with ambiguous nucleotides ('N'). Next, the low-quality reads having more than 20% Q < 20 bases were also trimmed. The yielded high-quality clean reads were used to the develop sequence assembly by using the Trinity software. After the removal of redundant Trinity-generated sequences by using the TGICL, clusters and unigenes were finally obtained.

Functional annotation and classification

All assembled unigenes were annotated by BLASTx analysis against the Nr (http://www.ncbi.nlm.nih.gov/), UniProt (http://www.uniprot.org/downloads), Pfam (http://pfam.xfam.org/), COG (http://www.ncbi.nlm.nih.gov/COG/) databases with an E-value <1e-5. Only the top hit results were extracted for each unigene. GO (http://www. geneontology. org) terms were functionally classified based on Nr annotations by using the Blast2go program (http://www.Blast2 go.de/). KEGG (http://www.genome.jp/kegg/) was used to draw metabolic maps. The KEGG analysis results included KEGG orthology (KO) numbers and enzyme commission (EC) numbers.

DEG analysis

Clean reads were mapped back onto the assembled unigenes by using the BWA program. The FPKM value was calculated for each unigene in each tissue of *Entada phaseoloides*. The expression difference

was analyzed by Fisher's exact test, and the FDR for each gene were obtained. DEGs were required to have thresholds of FDR≤0.001 and |log2Ratio|≥1. KEGG pathways were then reconstructed on DEGs.

Verification of gene expression by using qRT-PCR

Nine genes related to triterpene saponin biosynthesis were selected for validation by qRT-PCR. The primers for qRT-PCR analysis are listed in Additional file 9. All reactions were performed on the CFX96 real-time PCR system (Bio-Rad, USA) with a SYBR® Premix Ex Taq™ kit (Takara, China). The *Actin* gene was used as an internal control (Additional file 10). Each qRT-PCR experiment was performed with three biological repeats. The relative gene expression was calculated using the $2^{-\Delta\Delta CT}$ method.

Abbreviations

MVA: mevalonic acid; IPP: isopentenyl diphosphate; SQE: squalene epoxidase; OSC: oxidosqualene cyclase; CYP450: Cytochrome P450 monooxygenase; UGT: UDP-glycosyltransferase; TGICL: The TGI clustering tool; Nr: Non-redundant; KEGG: Kyoto Encyclopedia of Genes and Genomes; GO: Gene Ontology; COG: Clusters of Orthologous Groups; DEG: Differentially expressed gene; BAS: β-amyrin synthases; qRT-PCR: Quantitative reverse transcription-polymerase chain reaction

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Final sequences obtained from all the three tissues were submitted to the SRA database of NCBI with accession number PRJNA597694.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was financially supported by the Hubei Provincial Natural Science Foundation of China (Grant no. 2019CFB265), the Research and Innovation Initiatives of WHPU (Grant no. 2019Y03), and the National Natural Science Foundation of China (Grant no. 31700100). The Funding bodies were not involved in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

WFL has performed the transcriptome analysis and did the qRT-PCR experiments. RJG has also performed the qRT-PCR analysis with WFL. LHM and PLL have helped WFL in transcriptome analysis. ZNM conceived the experiments. All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge Prof. Zulin Ning (South China Botanical Garden) for providing the root, stem, and leaf tissues of *Entada phaseoloides.*

References

1. Xiong H, Ding X, Yang XZ, Yang GZ, Mei ZN. Triterpene Saponins from the Stems of *Entada phaseoloides*. Planta Med. 2014; 80: 710-18.
2. Dong Y, Shi H, Yang H, Peng Y, Wang M, Li X. Antioxidant phenolic compounds from the stems of *Entada phaseoloides*. Chem Biodivers. 2012; 9: 68-79.
3. Xiong H, Zheng YN, Yang GZ, Wang HX, Mei ZN. Triterpene saponins with anti-inflammatory activity from the stems of *Entada phaseoloides*. Fitoterapia. 2015; 103: 33-45.
4. Haralampidis K, Trojanowska M, Osbourn AE. Biosynthesis of triterpenoid saponins in plants. Adv Biochem Eng Biotechnol. 2002; 75: 31-49.
5. Contin A, Collu G, Van der Heijden R, Verpoorte R. The effects of phenobarbital and ketoconazole on the alkaloid biosynthesis in *Catharanthus roseus* cell suspension cultures. Plant Physiol Biochem. 1999; 37: 139-44.
6. Abe I, Rohmer M, Prestwich GD. Enzymatic cyclization of squalene and oxidosqualene to sterols and triterpenes. Chem Rev. 1993; 93: 2189-
7. Xu R, Fazio GC, Matsuda SPT. On the origins of triterpenoid skeletal diversity. Phytochemistry. 2004; 65: 261-91.

8. Luo H, Sun C, Sun Y, Wu Q, Li Y, Song J, et al. Analysis of the transcriptome of *Panax notoginseng* root uncovers putative triterpene saponin-biosynthetic genes and genetic markers. BMC Genomics. 2011; 12 Suppl 5: S5.
9. Zhao YJ, Li C. Biosynthesis of plant triterpenoid saponins in microbial cell factories. Agr Food Chem. 2018; 66:12155-65.
10. Seki H, Tamura K, Muranaka P450s and UGTs: key players in the structural diversity of triterpenoid saponins. Plant Cell Physiol. 2015; 56: 1463-71.
11. Wang Z, Gerstein M, Snyder M. RNA-Seq a revolutionary tool for Nat Rev Genet. 2009; 10: 57-63.
12. Belair CD, Hu T, Chu B, Freimer JW, Cooperberg MR, Blelloch High-throughput, Efficient, and Unbiased Capture of Small RNAs from Low-input Samples for Sequencing. Sci Rep. 2019; 9: 2262.
13. Li C, Zhu Y, Guo X, Sun C, Luo H, Song J, et al. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng*. BMC Genomics. 2013; 14:
14. Jeena GS, Fatima S, Tripathi P, Upadhyay S, Shukla RK. Comparative transcriptome analysis of shoot and root tissue of *Bacopa monnieri* identifies potential genes related to triterpenoid saponin biosynthesis. BMC genomics. 2017; 18: 1-15.
15. Soetaert SS, Van Neste CM, Vandewoestyne ML, Head SR, Goossens A, Van Nieuwerburgh FC, et al. Differential transcriptome analysis of glandular and filamentous trichomes in *Artemisia annua*. BMC Plant Biol. 2013; 13:
16. Nair P, Misra A, Singh A, Shukla AK, Gupta MM, Gupta AK, et al. Differentially expressed genes during contrasting growth stages of *Artemisia annua* for artemisinin content. PLoS One. 2013; 8:
17. Ramilowski JA, Sawai S, Seki H, Mochida K, Yoshida T, Sakurai T, et al. *Glycyrrhiza uralensis* transcriptome landscape and study of phytochemicals. Plant Cell Physiol. 2013; 54: 697-
18. Chen J, Tang XH, Ren CX, Wei B, Wu YY, Wu QH, et al. Full-length transcriptome sequences and the identification of putative genes for flavonoid biosynthesis in safflower. BMC genomics. 2018; 19:
19. Stonebloom SH, Scheller HV. Transcriptome analysis of rubber biosynthesis in guayule (*Parthenium argentatum* gray). BMC Plant Biol. 2019; 19:
20. Pandey A, Swarnkar V, Pandey T, Srivastava P, Kanojiya, Mishra DK, et al. Transcriptome and metabolite analysis reveal candidate genes of the cardiac glycoside biosynthetic pathway from *Calotropis procera*. Sci Rep. 2016; 6:
21. Chen CH, Zheng YJ, Zhong YD, Wu YF, Li ZT, Xu LA, et al. Transcriptome analysis and identification of genes related to terpenoid biosynthesis in *Cinnamomum camphora*. BMC genomics. 2018; 19:
22. Gagne SJ, Stout JM, Liu E, Boubakir Z, Clark SM, Page JE. Identification of olivetolic acid cyclase from *Cannabis sativa* reveals a unique catalytic route to plant polyketides. Proc Natl Acad Sci. 2012; 109: 12811-
23. Gupta P, Goel R, Pathak S, Srivastava A, Singh SP, Sangwan RS, et al. De novo assembly, functional annotation and comparative analysis of *Withania somnifera* leaf and root transcriptomes to identify

putative genes involved in the withanolides biosynthesis. PLoS One. 2013; 8:

24. Gahlan P, Singh HR, Shankar R, Sharma N, Kumari A, Chawla V, et al. De novo sequencing and characterization of *Picrorhiza kurrooa* transcriptome at two temperatures showed major transcriptome adjustments. BMC 2012; 13: 126.
25. Liao WF, Zhao SY, Zhang M, Dong KG, Chen Y, Fu CH, et al. Transcriptome assembly and systematic identification of novel cytochrome P450s in *Taxus chinensis*. Front Plant Sci. 2017; 8:
26. Upadhyay S, Phukan UJ, Mishra S, Shukla RK. De novo leaf and root transcriptome analysis identified novel genes involved in steroidal sapogenin biosynthesis in *Asparagus racemosus*. BMC Genomics. 2014; 15:
27. Kim YJ, Zhang D, Yang DC. Biosynthesis and biotechnological production ofginsenosides. Biotechnol A 2015; 33: 717-35.
28. Sun HP, Li F, Xu ZJ, Sun ML, Cong HQ, Qiao F, et al. De novo leaf and root transcriptome analysis to identify putative genes involved in triterpenoid saponins biosynthesis in Hedera helix L. PLoS One. 2017; 12:
29. Liu MH, Yang BR, Cheung WF, Yang KY, Zhou HF, Kwok JS, et al. Transcriptome analysis of leaves, roots and flowers of *Panax notoginseng* identifies genes involved in ginsenoside and alkaloid biosynthesis. BMC Genomics. 2015; 16: 265.
30. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011; 29: 644-52.
31. Perteal G, Huang XQ, Liang F, Antonescu V, Sultanal R, Karamycheval S, et al. TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 2003; 19: 651-
32. Li H, Durbin Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25: 1754-60.
33. Wang LK, Feng ZX, Wang X, Wang XW, Zhang XG. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. 2010; 26: 136-8.
34. Yendo AC, de Costa F, Gosmann G, Fett-Neto AG. Production of plant bioactive triterpenoid saponins: elicitation strategies and target genes to improve yields. Mol Biotechnol. 2010; 46: 94-104.
35. Zhang L, Jia H, Yin Y, Wu G, Xia H, Wang X, et al. Transcriptome analysis of leaf tissue of *Raphanus sativus* by RNA sequencing. PLoS One. 2013; 8:
36. Zhao F, Sun M, Zhang W, Jiang C, Teng J, Sheng W, et al. Comparative transcriptome analysis of roots, stems and leaves of *Isodon amethystoides* reveals candidate genes involved in Wangzaozins biosynthesis. BMC plant biology. 2018; 18:
37. Landl KM, Klösch B, Turnowsky F. ERG1, encoding squalene epoxidase, is located on the right arm of chromosome VII of *Saccharomyces cerevisiae*. 1996; 12: 609-13.
38. Hwang HS, Lee H, Choi, YE. Transcriptomic analysis of *Siberian ginseng* (*Eleutherococcus senticosus*) to discover genes involved in saponin biosynthesis. BMC genomics. 2015; 16:

39. Rasbery JM, Shan H, LeClair RJ, Norman M, Matsuda SP, Bartel B. *Arabidopsis thaliana* squalene epoxidase 1 is essential for root and seed development. J Biol Chem. 2007; 282: 17002-
40. Han JY, In JG, Kwon YS, Choi YE. Regulation of ginsenoside and phytosterol biosynthesis by RNA interferences of squalene epoxidase gene in *Panax ginseng*. Phytochemistry. 2010; 71: 36-
41. Fukushima EO, Seki H, Ohyama K, Ono E, Umemoto N, Mizutani M, et al. CYP716A subfamily members are multifunctional oxidases in triterpenoid Plant Cell Physiol. 2011; 52: 2050–61.
42. Han JY, Kim MJ, Ban YW, Hwang HS, Choi YE. The involvement of β-amyrin 28-oxidase (CYP716A52v2) in oleanane-type ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Physiol. 2013; 54: 2034-
43. Perez SL, Scossa F, Proost S, Bitocchi E, Papa R, Tohge T, et al. Multi-tissue integration of transcriptomic and specialized metabolite profiling provides tools for assessing the common bean (*Phaseolus vulgaris*) metabolome. Plant J. 2019; 97: 1132-
44. Wu Q, Ma X, Zhang K, Feng X. Identification of reference genes for tissue-specific gene expression in *Panax notoginseng* using quantitative real-time PCR. Biotechnol Lett. 2015; 37: 197-204.
45. Alonso-Serra J, Safronov O, Lim KJ, Fraser-Miller SJ, Blokhina OB, Campilho A, et al. Tissue-specific study across the stem reveals the chemistry and transcriptome dynamics of birch bark. New Phytol. 2019; 222: 1816-

Tables

Table 1 Secondary metabolism pathways in *Entada phaseoloides*

Pathway ID	Pathways	Unigene number
ko00100	Steroid biosynthesis	33
ko00130	Ubiquinone and other terpenoid-quinone biosynthesis	43
ko00232	Caffeine metabolism	6
ko00254	Aflatoxin biosynthesis	1
ko00261	Monobactam biosynthesis	19
ko00332	Carbapenem biosynthesis	1
ko00400	Phenylalanine, tyrosine and tryptophan biosynthesis	58
ko00401	Novobiocin biosynthesis	2
ko00521	Streptomycin biosynthesis	23
ko00524	Butirosin and neomycin biosynthesis	12
ko00860	Porphyrin and chlorophyll metabolism	49
ko00900	Terpenoid backbone biosynthesis	63
ko00902	Monoterpenoid biosynthesis	7
ko00903	Limonene and pinene degradation	21
ko00904	Diterpenoid biosynthesis	18
ko00905	Brassinosteroid biosynthesis	13
ko00906	Carotenoid biosynthesis	26
ko00908	Zeatin biosynthesis	18
ko00909	Sesquiterpenoid and triterpenoid biosynthesis	16
ko00940	Phenylpropanoid biosynthesis	162
ko00941	Flavonoid biosynthesis	36
ko00943	Isoflavonoid biosynthesis	5
ko00944	Flavone and flavonol biosynthesis	7
ko00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	11
ko00950	Isoquinoline alkaloid biosynthesis	24
ko00960	Tropane, piperidine and pyridine alkaloid biosynthesis	26
ko00965	Betalain biosynthesis	1
ko00966	Glucosinolate biosynthesis	1

Table 2 List of transcripts related to triterpenoid saponin backbone biosynthesis

Annotation	Enzyme code	Number of Unigenes	Expression value (Root)	Expression value (Stem)	Expression value (Leef)
Acetyl-CoA acetyltransferase	EC: 2.3.1.9	3	75.867	198.683	101.921
Hydroxymethylglutaryl-CoA synthase	EC: 2.3.3.10	4	91.369	386.235	56.173
Hydroxymethylglutaryl-CoA reductase	EC:1.1.1.34	10	450.327	760.578	72.126
Mevalonate kinase	EC:2.7.1.36	6	5.834	15.325	0.428
Phosphomevalonate kinase	EC: 2.7.4.2	1	11.324	40.568	20.753
Mevalonate-5-pyrophosphate decarboxylase	EC: 4.1.1.33	3	96.023	51.040	93.265
Isopentenyl pyrophosphate isomerase	EC: 5.3.3.2	1	20.335	32.018	35.736
Farnesyl pyrophosphate synthase	EC: 2.5.1.10	1	95.416	70.236	99.372
Squalene synthase	EC: 2.5.1.21	5	68.731	84.359	73.837
Squalene epoxidase	EC:1.14.99.7	16	163.378	206.325	90.231
β-Amyrin synthase	EC: 5.4.99.39	8	60.329	91.661	32.965
Lupeol Synthase	EC: 5.4.99.41	1	5.368	2.197	3.562
Cycloartenol synthase	EC: 5.4.99.8	5	10.312	15.385	2.214

Supplementary Information

Additional file 1. Summary of transcriptome sequencing and assembly results.

Additional file 2. Correlation indices between different samples.

Additional file 3. Functional annotation of *Entada phaseoloides* unigenes.

Additional file 4. Frequencies of unigenes matching GO terms.

Additional file 5. GO enrichment of unigenes.

Additional file 6. Unigenes for KEGG analysis.

Additional file 7. DEGs in different tissues.

Additional file 8. Sequence of β-amyrin synthase (EpBAS) gene in *Entada phaseoloides*.

Additional file 9. List of qRT-PCR primer sequences.

Additional file 10. Melting curves of reference gene *Actin* for qRT-PCR amplification.

Figures

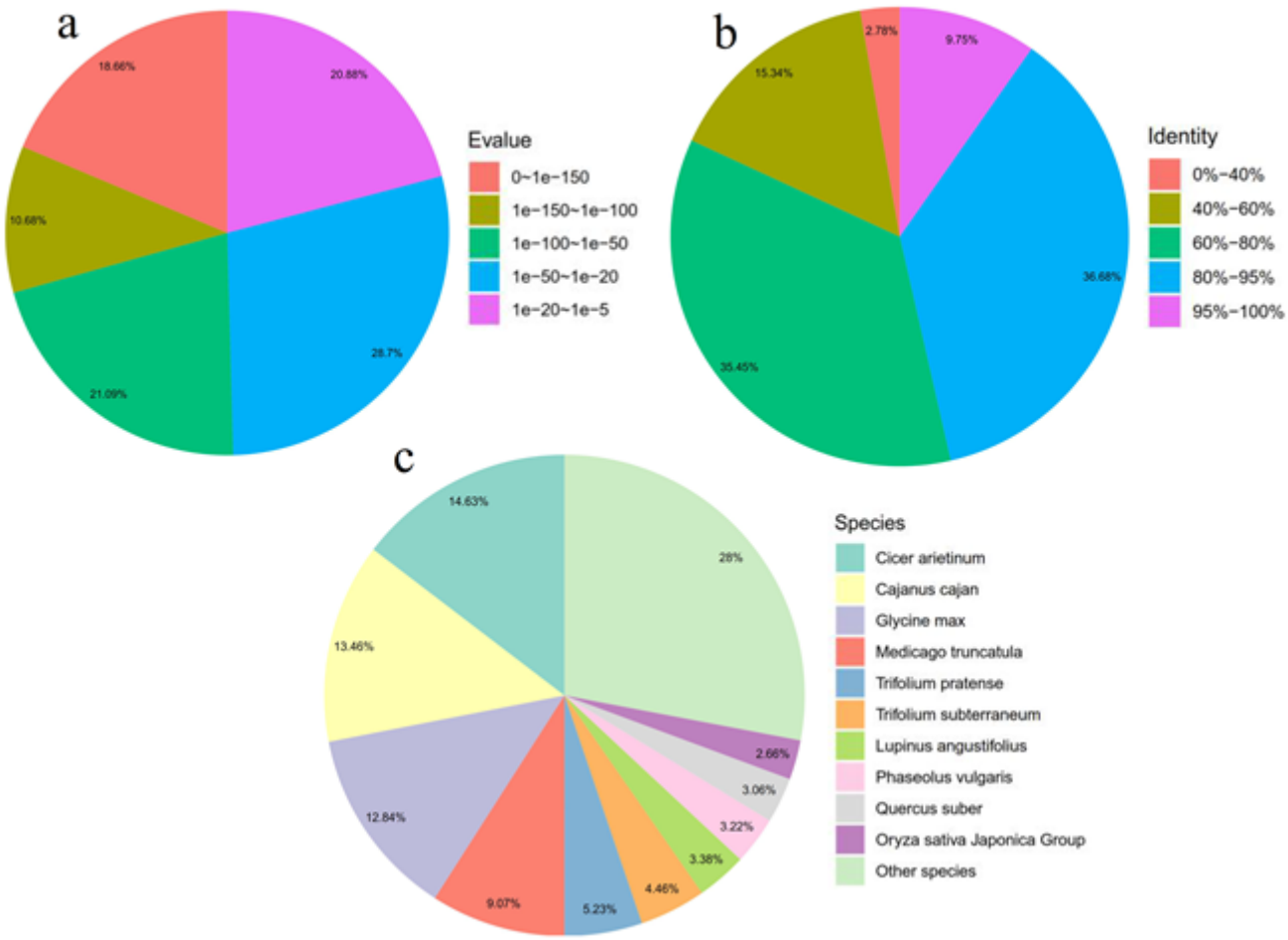

Figure 1

Similarity of unigenes annotated using the Nr database. (a) E-value distribution of best BLAST hits for each unigene (E-value < 1e-5). (b) Similarity distribution of top BLAST hits for each unigene. (c) Distribution of the most homologous sequence results for each unigene by species (E-value < 1e-5).

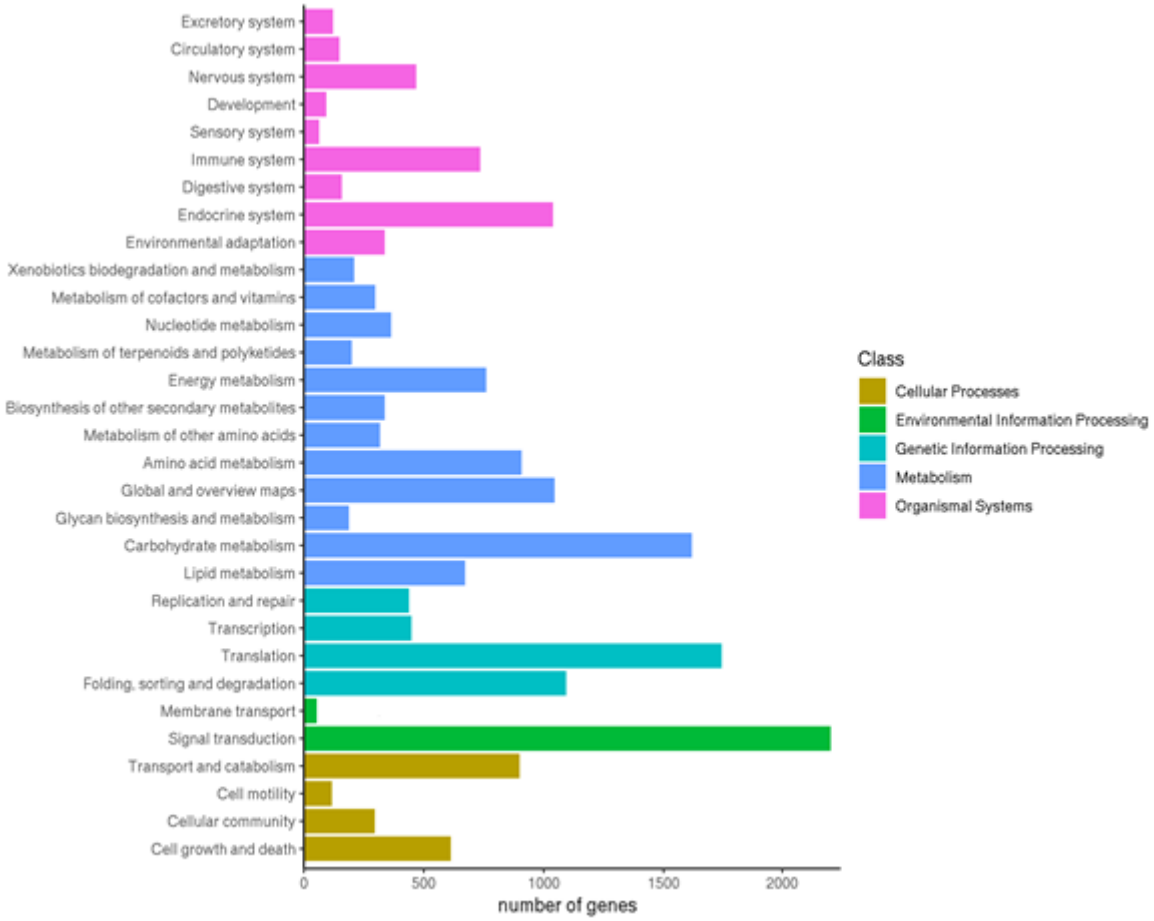

Figure 2

KEGG pathway classification of Entada phaseoloides unigenes. Exactly 15,119 unigenes were divided into five main categories in accordance with the corresponding pathways.

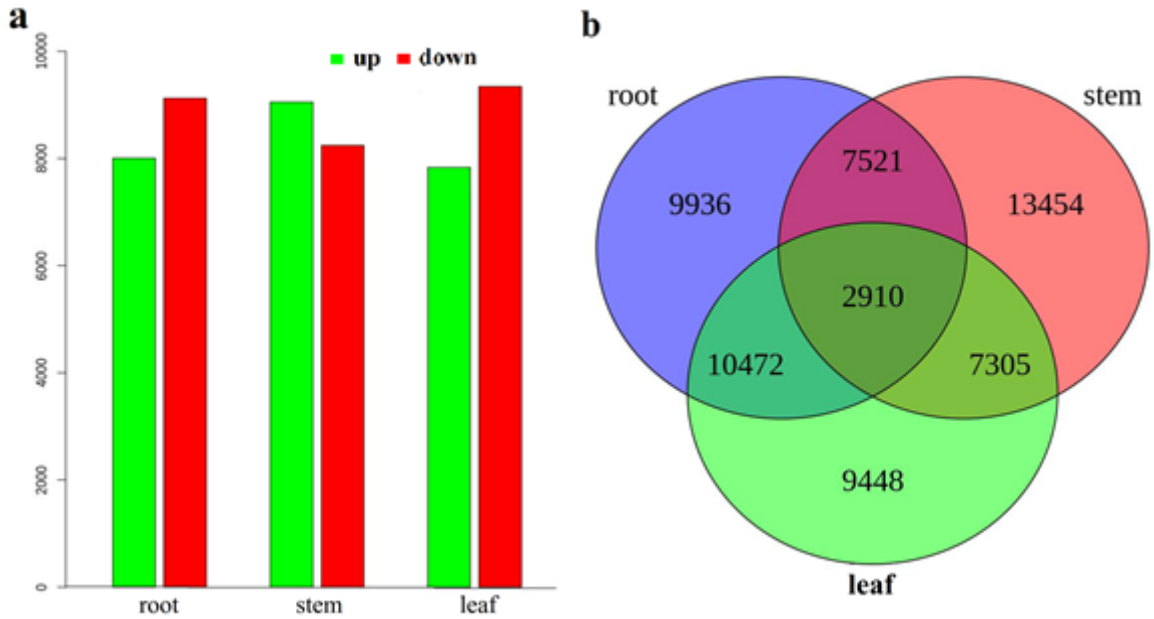

Figure 3

Differential expression analysis of unigenes. (a) Number of DEGs in each tissue compared with the other two tissues; (b) Venn diagram representing the number of DEGs among Entada phaseoloides tissues.

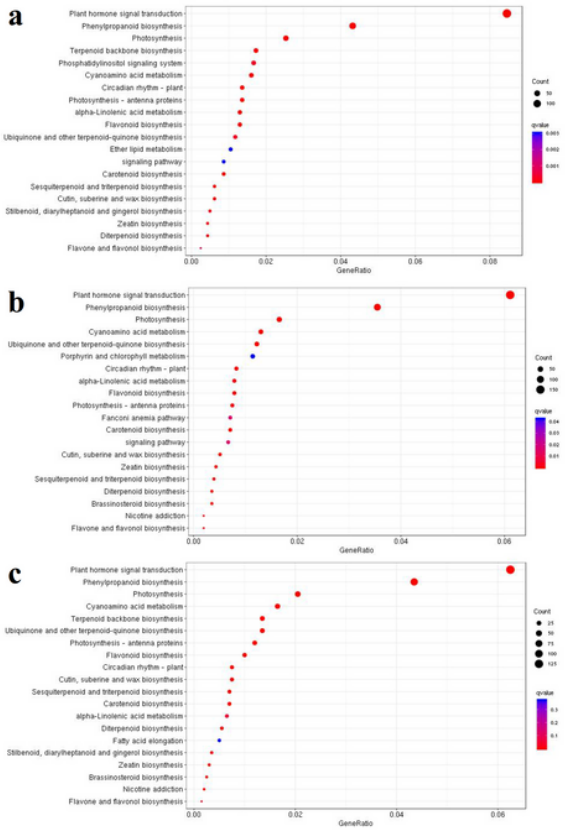

Figure 4

KEGG enrichment analyses of DEGs in different tissues. (a) between leaf and stem; (b) between leaf and root; (c) between stem and root.

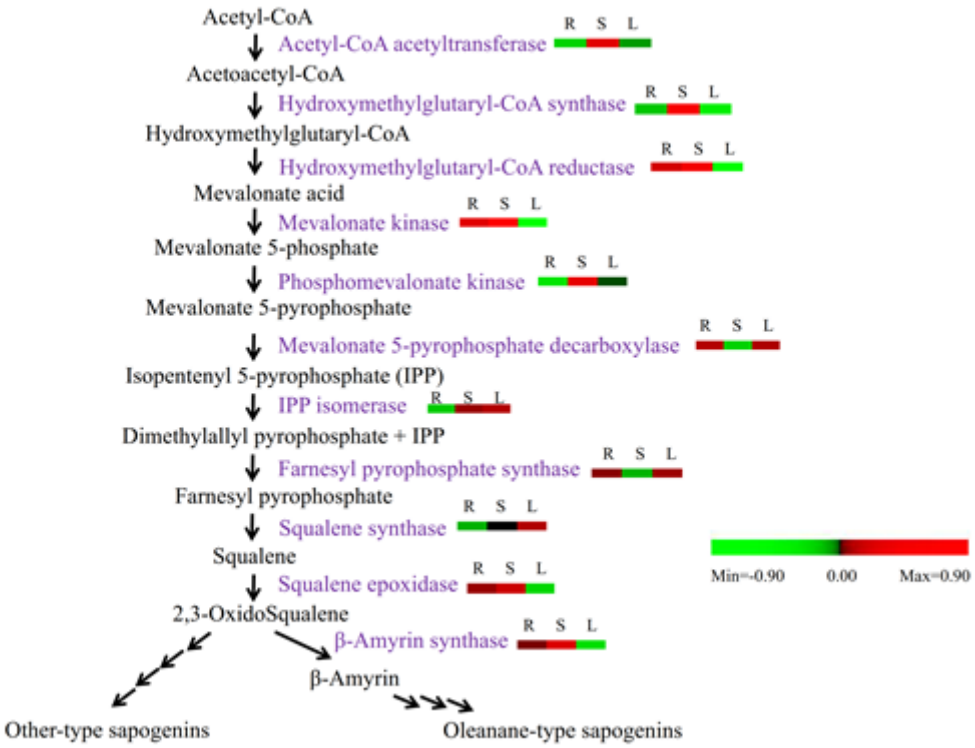

Figure 5

Schematic representation of the potential triterpenoid saponin biosynthesis pathway. Transcriptomic data (lg FPKM) for each gene represent the expression in the root (R), stem (S), and leaf (L) on heat map.

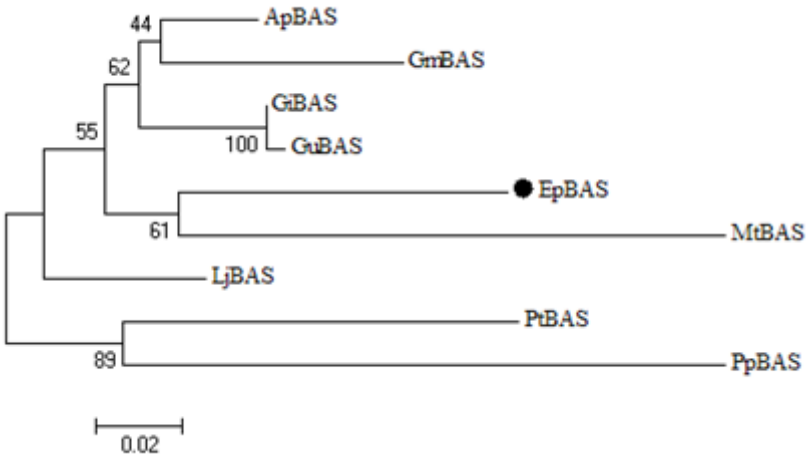

Figure 6

Phylogenetic analysis of the EpBAS and other plant BASs. The distances between each clone and group were calculated using CLUSTAL W. Bootstrap values are shown. Ap: Abrus precatorius; Gm: Glycine max; Gi: Glycyrrhiza inflata; Gu: Glycyrrhiza uralensis; Ep: Entada phaseoloides; Mt: Medicago truncatula; Lj: Lotus japonicus; Pt: Polygala tenuifolia; Pp: Prunus persica.

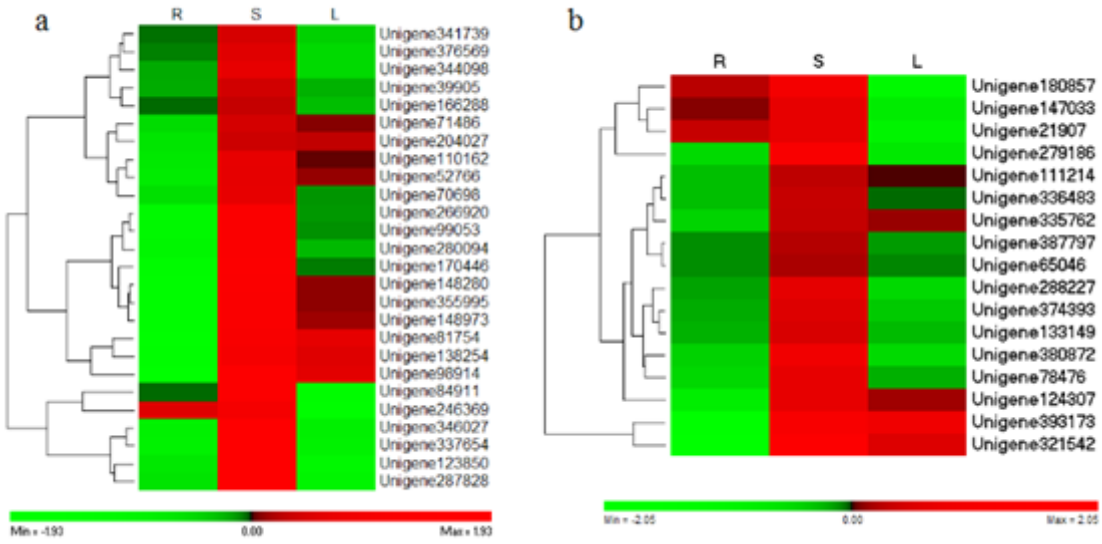

Figure 7

Heat map representing the upregulated unigenes of CYP450s (a) and UGTs (b) in stem (S) compared with the root (R) and leaf (L). The fold change expression data were obtained after three biological replicates.

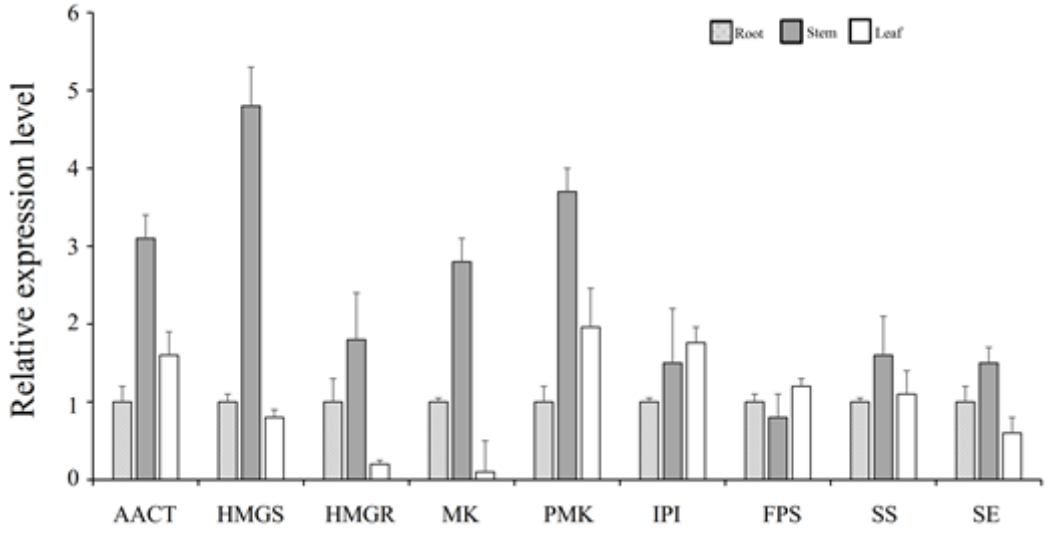

Figure 8

qRT-PCR validation of selected genes related to triterpene saponin biosynthesis. AACT: acetyl-CoA acetyltransferase; HMGS: hydroxymethyl- glutaryl-CoA synthase; HMGR: hydroxymethylglutaryl-CoA reductase; MK, mevalonate kinase; PMK: Phosphomevalonate kinase; IPI: isopentenyl pyrophosphate isomerase; FPS: farnesyl pyrophosphate synthase; SS: squalene synthase; SE: squalene epoxidase.

Supplementary Files

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

- Additionalfile1.doc
- Additionalfile2.pdf
- Additionalfile3.txt
- Additionalfile4.pdf
- Additionalfile5.xlsx
- Additionalfile6.xlsx
- Additionalfile7.txt
- Additionalfile8.txt
- Additionalfile9.docx
- Additionalfile10.tif