Zhang X *et al.* (2020)
Notulae Botanicae Horti Agrobotanici Cluj-Napoca 48(4):1802-1831
DOI:10.15835/48412044
Research Article

Transcriptome analysis to identify genes involved in lignan, sesquiterpenoid and triterpenoid biosynthesis in medicinal plant *Kadsura heteroclita*

Xiaodong ZHANG[1,2a], Caixia LI[1b], Chonlong CHIO[2], Ayyappa K.S. KAMESHWAR[2], Tianxiao MA[2,3], Wensheng QIN[2*]

[1]*Xuchang University, Food and Pharmacy College, College of Chemical and Materials Engineering, 88 Baiyi Road, Xuchang 461000, China; zxd95@xcu.edu.cn; 20201009@xcu.edu.cn*
[2]*Lakehead University, Department of Biology, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada; cchio@lakeheadu.ca; asistak@lakeheadu.ca; wqin@lakeheadu.ca (*corresponding author)*
[3]*Huanghe Science and Technology University, Faculty of Engineering, 666 South Zijingshan Road, Zhengzhou 450063, China; tma5@lakeheadu.ca*
[a,b]*These authors contributed equally to the work*

Abstract

Stems and roots of *Kadsura* plant species were the significant ingredients of traditional Chinese medicine. *Kadsura heteroclita* is one of the popular medicinal plants used in Tujia and Yao nationalities of China. Antioxidant compounds like lignan, sesquiterpenoid and triterpenoid are the major active components of *K. hetroclita*. Mass cultivation and bio-manufacturing strategies were being proposed to meet the increasing demand of *Kadsura* species plant parts. Therefore, it is important to reveal the molecular networks involved in biosynthesis of these highly efficient medicinal compounds. Here, transcriptomes of roots, stems and leaves in *K. heteroclite* seedling were sequenced by Hiseq2000 and unigenes involved in biosynthesis of lignan, sesquiterpenoid and triterpenoid biosynthesis were mined. As a result, 472 million clean reads were obtained which after aligning resulted in 160,248 transcripts and 98,005 genes. 191 and 279 unigenes were expected to be involved in biosynthesis of lignan, sesquiterpenoid and triterpenoid biosynthetic pathways respectively. Lignan, sesquiterpenoid and triterpenoid biosynthesis pathway genes were highly significant and differentially upregulated in roots and stems and downregulated in leaves. Also, genes encoding for MYB and bHLH transcription factors possibly involved in regulation of lignan, sesquiterpenoid and triterpenoid biosynthesis were discovered. These results provide the fundamental genomic resources for dissecting of biosynthetic pathways of the active components in *K. hetroclita*.

Keywords: *Kadsura heteroclita*; transcriptome; lignan biosynthesis; sesquiterpenoid biosynthesis; triterpenoid biosynthesis

Introduction

Kadsura heteroclita (Roxb.) Craib, popularly known in China as jixueteng, dixuexiang, sanxuexiang, dazuangufeng, chuifengsan, dafengshateng, dahongzuan, is a medicinal plant belonging to genus *Kadsura* and

Received: 18 Aug 2020. Received in revised form: 05 Oct 2020. Accepted: 02 Dec 2020. Published online: 22 Dec 2020.

family Schisandraceae. It is mainly distributed in Hubei, Guangdong, Hainan, Guangxi, Guizhou, Yunnan provinces of China, Bangladesh, Vietnam, Laos, Myanmar, Thailand, India, Sri Lanka, etc. in Asia (Editorial Committee of Flora of China, 1996). *K. heteroclita* is one of the major ingredients of both ethnomedicines "Xuetong" and "dahongzuan" which is highly used among the Tujia and Yao nationalities (Zhang *et al.*, 2019; Cao *et al.*, 2020). Rattan and root parts of *K. heteroclita* exhibit several medicinal properties such as promoting Qi, relieving pain, expelling wind and dehumidification, which is mainly used to treat rheumatism and rheumatoid arthritis, low back pain, lumbar muscle strain, stomach pain, abdominal pain, hemiplegia, postpartum paralysis, dysmenorrhea, fracture and bruise (Editorial Committee of Flora of China, 1996; Zhang *et al.*, 2019; Cao *et al.*, 2020). Modern studies have reported that lignans, sesquiterpenes and triterpenes are main components of rattan and root in *K. heteroclita.* The lignans, sesquiterpenes and triterpenes exhibit various pharmacological functions such as anti-oxidation, anti-virus, anti-tumor, cardiovascular protection, liver protection, anti-inflammatory, analgesic, and in memory improvement respectively (Zhang *et al.*, 2019).

The biosynthesis of lignan consists of three stages in plant: (1) biosynthesis of lignan precursor coniferyl-alcohol through phenylpropanoid pathway, (2) biosynthesis of lignan with various structures through coniferyl-alcohol and (3) in final stage lignans are modified by glycosylation to form lignan glycosides (Zhang *et al.*, 2019; Kanehisa, 2020). Similarly, terpenoids are synthesized from isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMADP) which are derived from plastidial methylerythritol phosphate (MEP) pathway and cytosolic mevalonic acid (MVA) pathway (Zhang *et al.*, 2015). Later, one molecule of IPP and one molecule of DMADP are condensed into geranyl diphosphate (GPP), a C_{10}-compound, by geranyl diphosphate synthase (GPPS) (Zhang *et al.*, 2015). One molecule of GPP and one IPP are further condensed into farnesyl diphosphate (FPP), a C_{15}-compound, by farnesyl diphosphate synthase (FPPS) in the cytoplasm (Fidan and Zhan, 2018). The FPP is a branch-point precursor for the synthesis of both sesquiterpenoid and triterpenoid (Fidan and Zhan, 2018). The sesquiterpenoid solavetivone can be synthesized by vetispiradiene synthase (VES) and premnaspirodiene oxygenase (PO) from FPP, while germacrene D, 7-epi-α-selinene and valencene can be synthesized by germacrene D synthase (GDS), 7-epi-alpha-selinene synthase (SS) and valencene synthase (VS), separately (Kanehisa, 2020). Thus, briefly the biosynthesis of triterpenoid involves two molecules of FPP are condensed into one molecule of squalene by squalene synthase (SQS) in the endoplasmic reticulum membrane followed by conversion of squalene into 2,3-epoxy-2,3-dihydrosqualene by squalene monooxygenase or squalene epoxidase (SE). Thus, 2,3-epoxy-2,3-dihydrosqualene is a branch-point precursor for the synthesis of the linear and cyclic triterpenoid compounds (Kanehisa, 2020).

Till date research studies on *K. heteroclita* are majorly focused on isolation of chemical components, pharmacological effects, pharmacokinetics, fingerprints and DNA barcodes (Zhou *et al.*, 2016; Guo, 2017; Fan *et al.*, 2019; Cao *et al.*, 2020; Cao *et al.*, 2020; Liu *et al.*, 2020; Shehla *et al.*, 2020). Limited research has been conducted towards understanding the lignan, sesquiterpenoid and triterpenoid biosynthetic pathways. As of today, 69 different types of lignan compounds (including dibenzocyclooctadienes, aryltetralins, aryltetralins, diarylbutanes, and tetrahydrofurans), 64 different types of triterpenoids (including cycloartane type, lanolin type, oleanane type and schiartane type) and 15 different types of sesquiterpenoids has been isolated from *K. heteroclita* (Cao *et al.*, 2019; Zhang *et al.*, 2019; Cao *et al.*, 2020). As these bioactive components are sparsely produced in plant, it is highly important to implement and develop novel synthetic biological strategies for industrial production to meet the demand (Fidan and Zhan, 2018). Another major problem in production is usage of several names for the same medicinal plant and the same name for various other species which have caused the wrong usage of medicinal material (Editorial Committee of Flora of China, 1996). Therefore, presence of complete genome sequence of *K. heteroclita* would significantly benefit both scientific and industrial communities by separating it from other species. As of today, the complete genome sequence of *K. heteroclita* is unavailable and only chloroplast genome is publicly available (Guo, 2017).

In our present study, we have performed genome-wide transcriptome analysis of the roots, stems and leaves of *K. heteroclita* using the Illumina Hiseq 2000® platform. For the first time we have reported the complete genome-wide transcriptome analysis of *K. heteroclita* to reveal the gene expression patterns of roots,

stems and leaves of *K. heteroclita* seedlings, which are involved in the biosynthesis pathways of lignan, sesquiterpenoid and terpenoid and their regulatory mechanisms. We strongly believe that the results obtained in our study will lay a strong foundation for future studies on identification of medicinal compounds and synthetic biological studies on *K. heteroclita*.

Materials and Methods

Plant materials

The seedlings of *Kadsura heteroclita* (24°28'25"N, 100°8'33"E) were cultivated from the branch cuttings obtained from the *K. hetroclita*. All the branch cuttings of *K. hetroclita* were kindly provided by Lincang Biological Pharmaceutical Technology Company Limited, Yun County, China. The roots, stems and leaves were collected from five-month-old *K. heteroclita* seedlings (Figure 1). Materials from three individual plants were collected using scissors to yield 1 g of root, stem and leaf samples which were used in our further transcriptome studies (BioSample accession numbers: SAMN14883617, SAMN14883618, SAMN14883619, SAMN14883620, SAMN14883621, SAMN14883622, SAMN14883623, SAMN14883624, SAMN14883625). All the samples were immediately frozen in liquid nitrogen and then stored in -80 ℃ refrigerator after cutting and wrapping with tinfoil respectively.

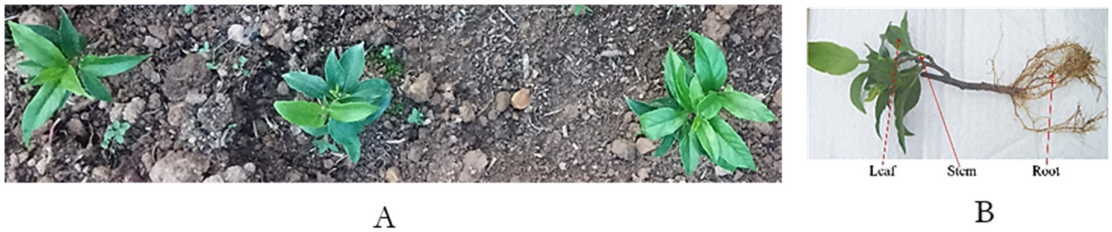

Figure 1. Plant materials of *K. heteroclita*
(A) *K. heteroclita* growing in the field; (B) Schematic diagram of roots, stems and leaves used in experiments for sequencing

RNA isolation and quality control

Total RNA was extracted by TRIzol® reagent (Invitrogen, Carlsbad, CA, USA), and DNase I (Takara, Dalian, China) was used to remove the residual genomic DNA. The RNA isolation was performed according to the manufacturer's protocols. The purity of the RNA was determined by the NanoPhotometer® (IMPLEN, Westlake Village, CA, USA). The integrity of the RNA was determined by the Bioanalyzer 2100 system (Agilent Technologies, Palo Alto, CA, USA). RNA concentration was precisely measured by the Qubit® 2.0 Flurometer (Life Technologies, Carlsbad, CA, USA). Thus, the obtained RNA was used for library construction.

Library construction, sequencing and quality control

Illumina TruSeq™ RNA Sample Preparation Kit v2 (Illumia, San Diego, USA) was used for the library construction according to the manufacturer's manual. mRNA was enriched by the magnetic beads with Oligo(dT). Fragmentation buffer is then added to break the mRNA into short fragments. The 1st-strand of cDNA is synthesized with random hexamers using pure mRNA as a template and then the 2nd-strand of cDNA was synthesized by adding the buffer, dNTPs, DNA polymerase I and RNase H. AMPure® XP beads (Beckman Coulter, High Wycombe, UK) were used to purify the double-stranded cDNA. Thus, obtained purified double-stranded cDNA is first subjected to end repair, A-tailing, adding the sequencing adapter, and later the DNA fragments were selected using the AMPure XP beads respectively. Finally, PCR amplification was performed, and the PCR product were purified using the AMPure XP beads to obtain the final library. Once the library construction is completed, we have used Qubit® 2.0 flurometer (Life Technologies, Carlsbad,

CA, USA) for preliminary quantification and the libraries were diluted to 1.5 ng/μl, and then Agilent 2100 bioanalyzer (Agilent Technologies, Palo Alto, California) was used to detect the insert size of the library. The fragments meeting the ideal insert size were used, the fragments with effective concentration above 2 nM of library was accurately quantified using the Q-PCR method. The libraries obtained from the quality control analysis (effective concentration and the target data volume) were pooled and further used for the transcriptome sequencing which was achieved using Illumina HiSeq® sequencing platform.

Transcriptome assembly

The original sequence image files obtained from Illumina HiSeq were converted to raw sequenced reads using Casava base calling software. Thus, these obtained raw reads were stored in fastq files, and they were later processed to remove the ones containing adapter, ploy(N) and low-quality. We have calculated the Q20, Q30, GC content and sequence duplication level of the clean data. Then these clean reads were assembled by Trinity® v2.4.0 software (Grabherr *et al.*, 2011) with following parameters min_kmer_cov set to 3 and all other parameters were set to default.

Gene functional annotation, coding sequence (CDS) and transcription factor (TF) prediction

The assembled transcriptome was annotated to obtain the functional annotations by mapping against BLAST databases: Nr (NCBI non-redundant protein sequences, diamond v0.8.22, e-value = 10^{-5}), Nt (NCBI nucleotide sequences, NCBI blast 2.2.28+, e-value = 10^{-5}), PFAM (Protein family, HMMER 3.0 package, hmmscan e-value = 10^{-2}), SwissProt (A manually annotated and reviewed protein sequence database, diamond v0.8.22, e-value = 10^{-5}), KOG (eukaryotic Ortholog Groups, diamond v0.8.22, e-value = 10^{-5}), KEGG (Kyoto Encyclopedia of Genes and Genomes, KEGG Automatic Annotation Server, e-value= 10^{-10}) and GO (Gene Ontology, Blast2GO v2.5, e-value = 10^{-6}). For the CDS prediction, unigenes were first aligned to the NR database, ORF (open reading frame) information of transcripts was extracted from the alignment results, and the sequence of the coding region was translated into amino acid sequence according to the standard codon table. For the transcripts without a possible match to the NR, ESTSCAN 3.0.3 software was used to predict the CDS. We have used iTAK (https://github.com/kentn/iTAK/) software for the prediction, identification and classification of the TF as previously reported by (Paulino *et al.*, 2010).

Differential expression analysis

We have used RSEM software package to analyse the differential gene expression profiles (Li and Dewey, 2011). Also, we have used DESeq 1.18.0 R package to perform differential gene expression analysis. Unigenes with an adjusted p-value < 0.05 reported by DESeq were defined as differentially expressed genes (DEGs). The heatmaps were generated using the R package pheatmap.

GO enrichment and KEGG pathway enrichment analysis

GOseq R packages was used to perform GO enrichment analysis on the above obtained DEGs (Young *et al.*, 2010). KOBAS software was used to perform the pathway enrichment of DEGs (Mao *et al.*, 2005).

Phylogenetic analysis

The gene sequences coding for KhDIR, KhPLR, KhSDH, KhSQS, KhSE, KhMYB and KhbHLH TFs were retrieved from the *K. heteroclita* CDS, and all the sequences used were listed in supplementary file. The above-mentioned sequences coding for TFs were aligned using Clustal X 2.1 to generate the .phy format files. Then PhyML 3.0 (*http://www.atgc-montpellier.fr/phyml/*) was used to construct the phylogenetic tree by using the default parameters with the boostrap value set at = 100 respectively (Guindon *et al.*, 2010). We have used the iTOL 5.5.1 (*https://itol.embl.de/*) software to visualize the trees (Letunic and Bork, 2019).

Sequence alignment and building of three-dimensional models

For the sequence alignment of DIR and SDH, we have used DNAMAN 10.0.2.8 software with the default parameters. And three-dimensional models of DIR proteins were built using the SWISS-MODEL online sever (*https://swissmodel.expasy.org/*) with the default parameters.

Results

Sequencing and assembly

In order to understand and reveal the biosynthesis pathway of lignan, sesquiterpenoid and triterpenoid in *K. heteroclita*, nine sequencing libraries including roots (K_R1, K_R2, K_R3), stems (K_S1, K_S2, K_S3) and leaves (K_L1, K_L2, K_L3) were prepared and sequenced using the Illumina Hiseq 2000® platform. More than 6.58 G clean reads per library was obtained after quality control analysis (Table 1). The error rate of all libraries was found to be 0.03%, while Q20 and Q30 were over 95.19% and 88.62% respectively, which indicated that these data were suitable for the downstream analysis. The raw datasets from the nine sequencing libraries have been deposited in the Short Reads Archive (SRA) database under the accession numbers: SRR11747016-SRR11747024. The clean reads were combined and assembled using the Trinity v2.4.0, with min_kmer_cov set to 3 and for all the other settings we have used default parameters (Grabherr *et al.*, 2011). Assembled sequences were subjected to cluster using the trinity algorithm. Finally, a total of 160,248 transcripts and 98,005 genes were assembled, with 49,623 (30.97%) transcripts and 48,659 (49.65%) genes being longer than 1 Kb in length (Figure 2). The average length of transcripts and genes were found to be 927 bp and 1306 bp (Table 2), and the N50 for transcripts and genes were 1609 bp and 1838 bp respectively (Table 2).

Table 1. Quality assessment of the sequencing data

Sample	Raw Reads	Clean Reads	Clean Bases	Error (%)	Q20 (%)	Q30 (%)	GC (%)
K_R1	66225100	65006078	9.75G	0.03	95.33	88.65	45.96
K_R2	51515170	50789876	7.62G	0.03	97.16	92.49	46.85
K_R3	51515170	50868564	7.63G	0.03	95.80	89.45	46.85
K_S1	51515170	50630252	7.59G	0.03	95.76	89.39	45.95
K_S2	59818694	59151942	8.87G	0.03	97.72	93.51	46.69
K_S3	51515170	50861982	7.63G	0.03	97.42	92.75	46.86
K_L1	50867794	50385662	7.56G	0.03	95.92	89.61	46.55
K_L2	51515170	50959816	7.64G	0.03	97.23	92.31	47.76
K_L3	44509274	43883852	6.58G	0.03	95.19	88.62	47.35

Note: K_R1, K_R2 and K_R3: three root samples; K_S1, K_S2 and K_S3: three stem samples; K_L1, K_L2 and K_L3: three leaf samples

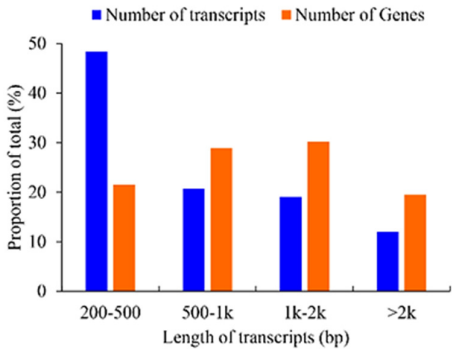

Figure 2. Length distribution frequency of spliced transcripts and unigenes

Table 2. Length distribution frequency of the spliced transcripts and unigenes

	Min length	Mean length	Median length	Max length	N50	N90	Total nucleotides
Transcript	201	927	525	17760	1609	356	148564511
Unigene	201	1306	992	17760	1838	630	127993489

Gene function annotation and classification

A total of 98,005 unigenes were annotated against seven classic databases which included NR, NT, PFAM, SwissProt, KOG, KEGG and GO databases with an E-value cutoff of 10^{-5}, 10^{-5}, 10^{-2}, 10^{-5}, 10^{-3}, 10^{-10} and 10^{-6} respectively. The 7458 (7.60%) unigenes were found to be commonly annotated to all the above seven databases, while 76,101 unigenes (77.65%) were annotated at least to one database (Table 3). Totally 54,930 unigenes (56.04%) exhibited high similarity with sequences in the NR database, 36,225 unigenes (36.96%) similarity to protein sequences in NT, and finally 56,998 unigenes (56.15%) exhibited similarity with known genes in SwissProt. Results obtained from the species distribution based on their similarity results against NR databases showed that highest number of transcripts were annotated to *Nelumbo nucifera* (20.42%), followed by *Macleaya cordata* (11.36%) and *Amborella trichopoda* (11.30%) respectively (Figure 3).

Table 3. Statistical results of gene annotation

Item	Number of Unigenes (n)	Percentage (%)
Annotated in NR	54930	56.04
Annotated in NT	36225	36.96
Annotated in KO	28234	28.8
Annotated in SwissProt	56998	58.15
Annotated in PFAM	55662	56.79
Annotated in GO	55662	56.79
Annotated in KOG	20480	20.89
Annotated in all Databases	7458	7.6
Annotated in at least one Database	76101	77.65
Total Unigenes	98005	100

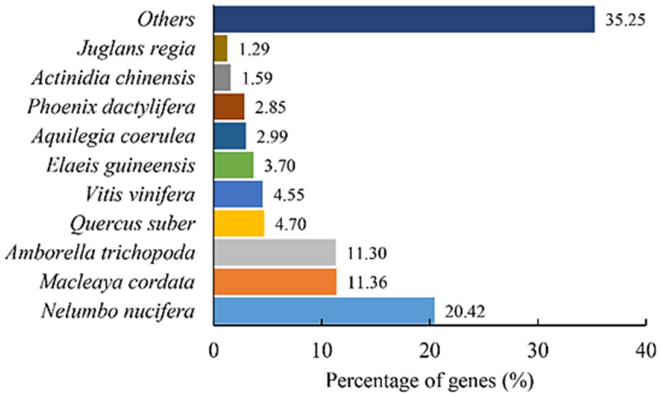

Figure 3. Species distribution of top10 BLASTx hits against NR database

Coding sequences were also predicted using NR database and ESTSCAN software respectively. A total of 61,074 peptides were predicted by using BLASTp with the peptide length ranging between 15 to 1254 (Figure 4a), whereas 37,045 peptides were predicted by ESTSCAN with peptide length ranging from 15 to 734 respectively (Figure 4b).

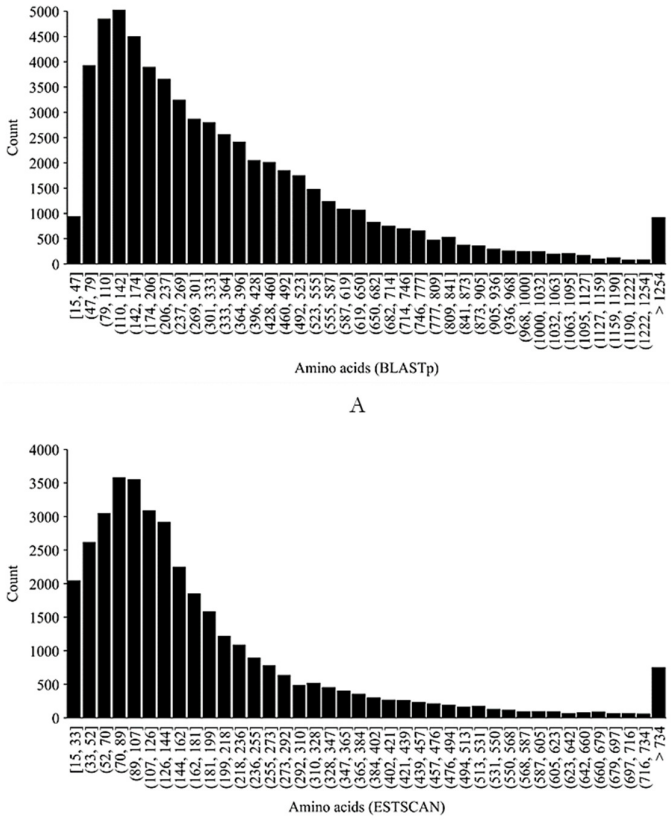

Figure 4. Length distributions of predicted peptides. (A) Predicted by BLASTp; (B) Predicted by ESTSCAN

A total of 55,662 (56.80%) unigenes were successfully classified into 56 functional groups (Figure 5). Out of which 26, 20 and 10 groups belong to biological process, cellular component and molecular function respectively. Results obtained from these biological contextualization studies have shown that the majority of the cellular and metabolic processes are predicted to be involved in biosynthesis of lignan, sesquiterpenoid and triterpenoid. The gene ontology distribution of unigenes can be categorized as: "cellular process" (30,693, 55.14%), "metabolic process" (28,461, 51.13%) and "single-organism process" (22,794, 40.95%), "binding" (33,337, 59.89%) and "catalytic activity" (24,598, 44.19%) (Figure 5) respectively. Results obtained in this study are in accordance with previous genome-wide annotation results of *Dendrobium huoshanense* and *Arisaema heterophyllum* (Wang *et al.*, 2018; Zhou *et al.*, 2020).

Annotating against KOG database has resulted in a total of 20,480 unigenes classified among 26 KOG (Figure 6). Amongst the 26 KOG groups, the majority of genes were distributed among "post-translational modification, protein turnover, and chaperon" (3126, 15.26%), followed by "general function prediction only" (2671, 13.04%), "translation, ribosomal structure and biogenesis" (2407, 11.75%) and "intracellular trafficking, secretion, and vesicular transport" (1468, 7.16%) respectively. Interestingly, the genes distributed among the top-three groups in *Arisaema heterophyllum* were "general function prediction only (9,277, 26.57%)", "transcription (5,784, 16.56%)" and "translation, ribosomal structure and biogenesis (5,667, 16.23%)" (Wang *et al.*, 2018; Zhao *et al.*, 2020). These results show that protein synthesis, processing and transport are more active in *K. heteroclita* seedling.

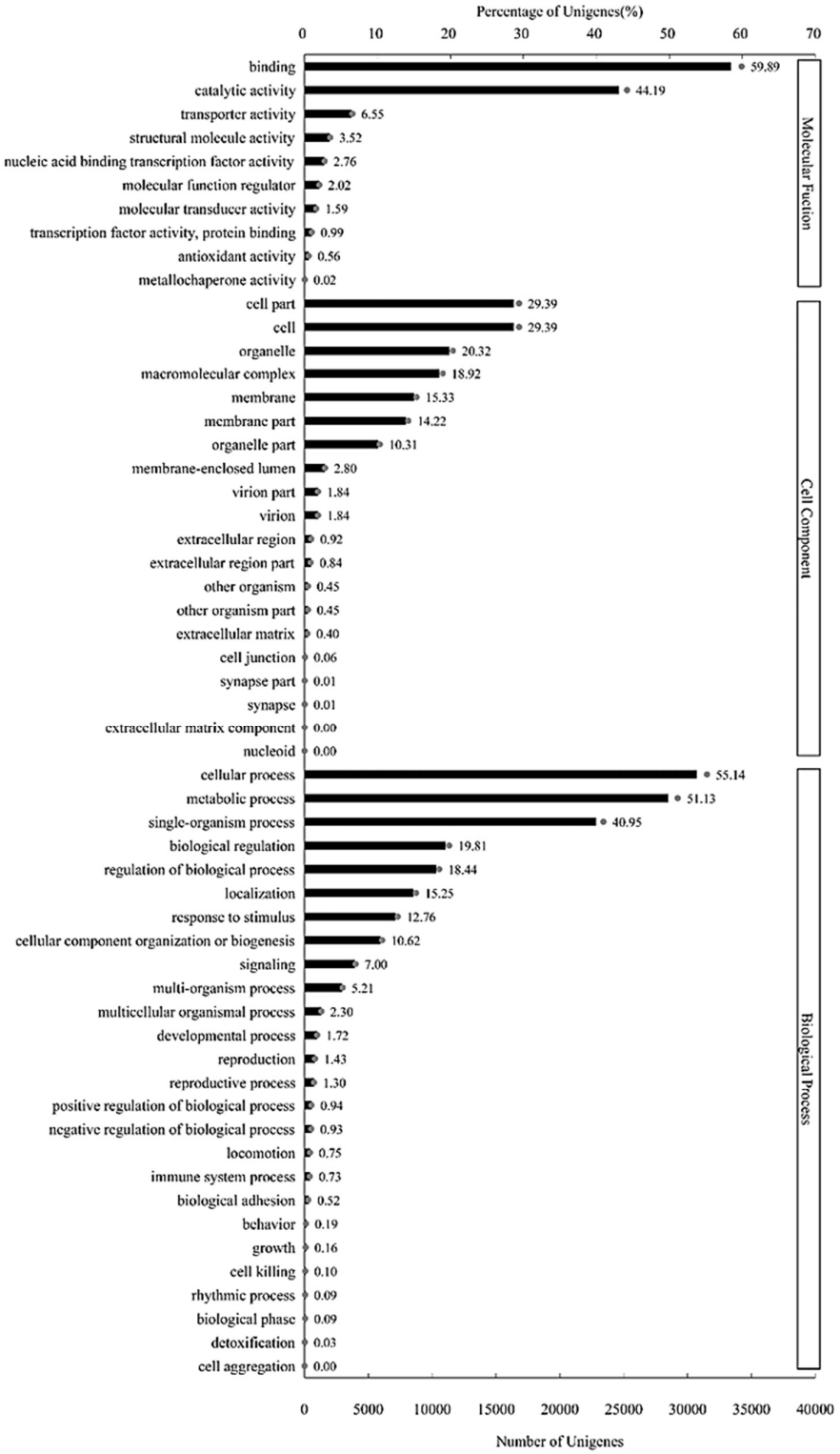

Figure 5. GO classification map. The ordinate represents the GO terms of the three GO categories while the abscissa represents the number and percentage of genes annotated into the corresponding term

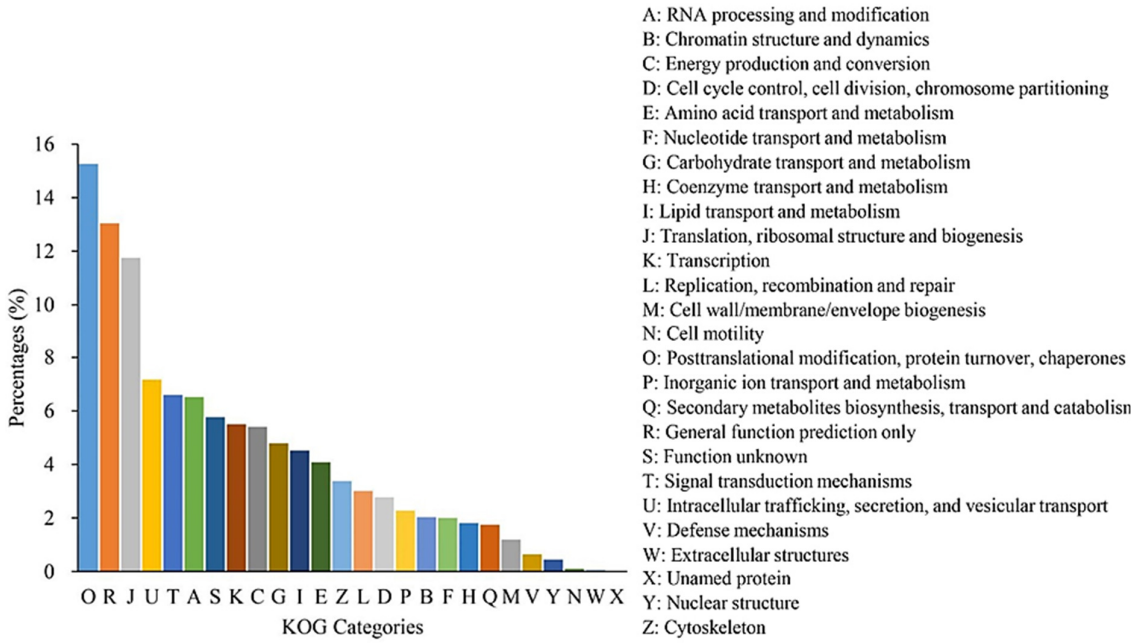

Figure 6. KOG classification map. The abscissa represents KOG groups, while the ordinate represents the percentage of annotated genes

Finally, to understand and reveal the active metabolic pathways differentially expressed in *K. heteroclita* seedlings, we have annotated the transcriptome against the KEGG database. A total of 98,005 unigenes were assigned into five categories including cellular processes, environmental information processing, genetic information processing, metabolism and organismal systems, 19 sub-categories (Figure 7) and 130 pathways respectively. Especially, the following pathways: "Translation" (2794, 9.90%), "carbohydrate metabolism" (2282, 8.08%), "folding, sorting and degradation" (2048, 7.25%) and "overview" (1572, 5.57%) were found to be the top-four pathways (Figure 7). Especially, we have observed that a total of 1359, 933 and 890 genes were involved in "amino acid metabolism", "metabolism of terpenoids and polyketides" and "biosynthesis of other secondary metabolites" respectively. These results conveyed that the amino acid pathway and terpenoid pathways were also highly active in *K. heteroclita*, and the corresponding genes would be good candidate genes for lignan, sesquiterpenoid and triterpenoid biosynthesis respectively.

Identification of DEGs, GO and KEGG enrichment analysis

Results obtained from the statistical analysis have revealed a list of highly significant and DEGs ($p < 0.05$ and $\log_2$ foldchange > 1). As our current study is focused on understanding the DEGs involved biosynthesis of lignan, sesquiterpenoid and triterpenoid pathways among the leaves, roots and stems of *K. heteroclite*, we have designed our analysis as K_L vs. K_R, K_S vs. K_R and K_L vs. K_S respectively. The statistical analysis has resulted in total of 41,497 significant DEGs. Among these significant DEGs, 22,468 (22.92%) genes were found to be significantly expressed in leaves and roots samples, out of which 9,511 (42.33%) genes were upregulated in leaves samples and 12,957 (57.67%) genes were found to be downregulated in leaves (Figure 8A). Similarly, 7,373 genes (7.52% of all genes) were identified as significantly DEGs between leaves and stems samples with 3,950 (53.57%) genes upregulated and 3,423 (46.43%) genes downregulated in leaves (Figure 8B). Finally, 11,656 genes (11.89%) were identified as significantly DEGs between stems and roots with 4,302 (36.91%) genes upregulated and 7,354 (63.09%) genes downregulated in stems (Figure 8C). In order to compare the commonly expressed genes among the K_L vs. K_R, K_S vs. K_R and K_L vs. K_S conditions, Venn diagram for different comparison groups was used. Among these three groups, 1761 DEGs were identified in common among all the datasets (Figure 8D). Specifically, 9594 DEGs were commonly identified

among the "K_L vs. K_R" and "K_S vs. K_R"; 6032 DEGs were commonly identified among the "K_L vs. K_R" and "K_L vs. K_S" groups, while 2348 DEGs were identified commonly among the "K_L vs. K_S" and "K_S vs. K_R" groups respectively (Figure 8D).

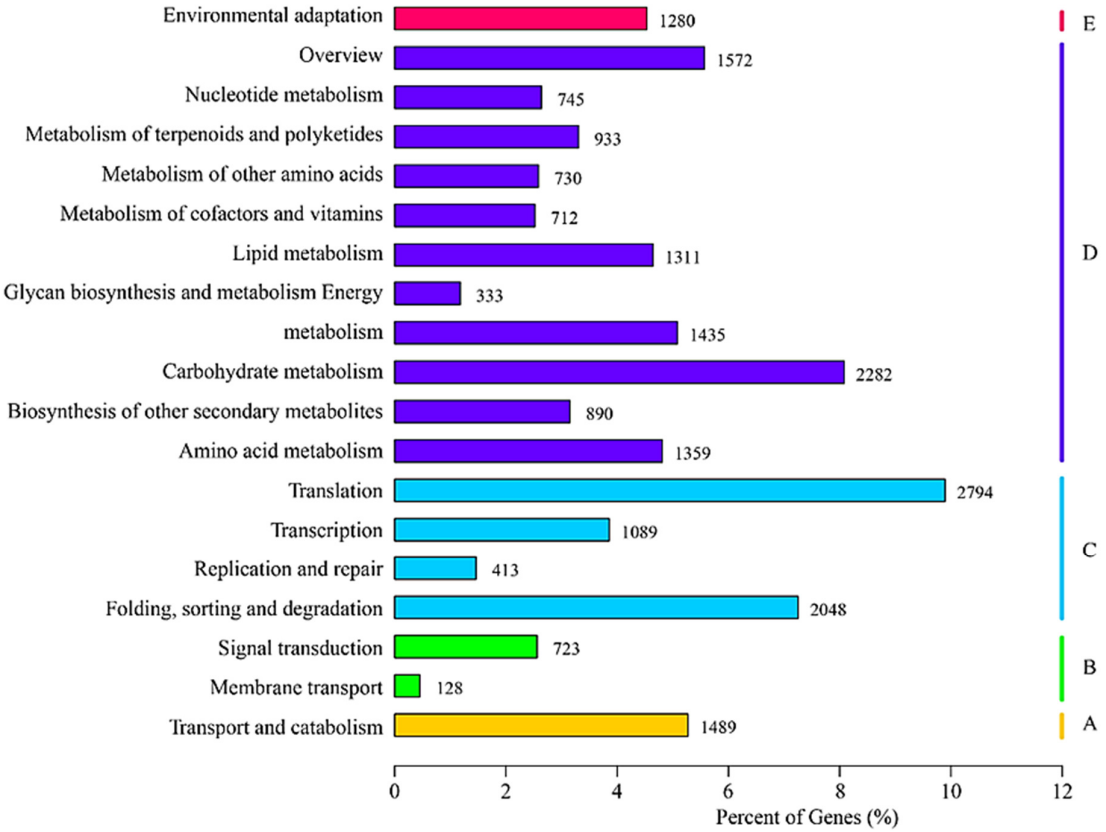

Figure 7. KEGG classification map

The ordinate is the pathway, and the abscissa is the proportion of genes belonging to the corresponding pathway. These genes were divided into five categories: A. Cellular Processes; B. Environmental Information Processing; C. Genetic Information Processing; D. Metabolism; E. Organismal Systems.

The above obtained significant DEGs were further analyzed using the GO and KEGG databases for understanding and revealing the functional involvement and biological contextualization of these genes in biosynthesis of lignan, sesquiterpenoid and triterpenoids in *K. heteroclita* (Table S1-S6). The results of gene enrichment analysis using KEGG pathway database showed that 83 DEGs (number of DEGs ranked 4th) fell into sesquiterpenoid and triterpenoid biosynthesis pathway, while 192 DEGs (ranked 10th) fell into phenylpropanoid biosynthesis pathway in the K_L vs. K_R comparison (Figure 9A, Table S4). Whereas in the comparison of K_L vs. K_S, 18 DEGs (ranked 4th) were assigned into sesquiterpenoid and triterpenoid biosynthesis pathway and 79 DEGs (ranked 2nd) were assigned into phenylpropanoid biosynthesis pathways, respectively (Figure 9B, Table S5). Finally, in the comparison of K_S vs. K_R, 60 (ranked 4th) and 107 (ranked 14th) significant DEGs were expected to be involved in the sesquiterpenoid and triterpenoid biosynthesis pathway and phenylpropanoid biosynthesis pathway respectively (Figure 9C, Table S6). Also, the top 20 KEGG enrichment pathways obtained from the above enrichment analysis were shown in Figure 9.

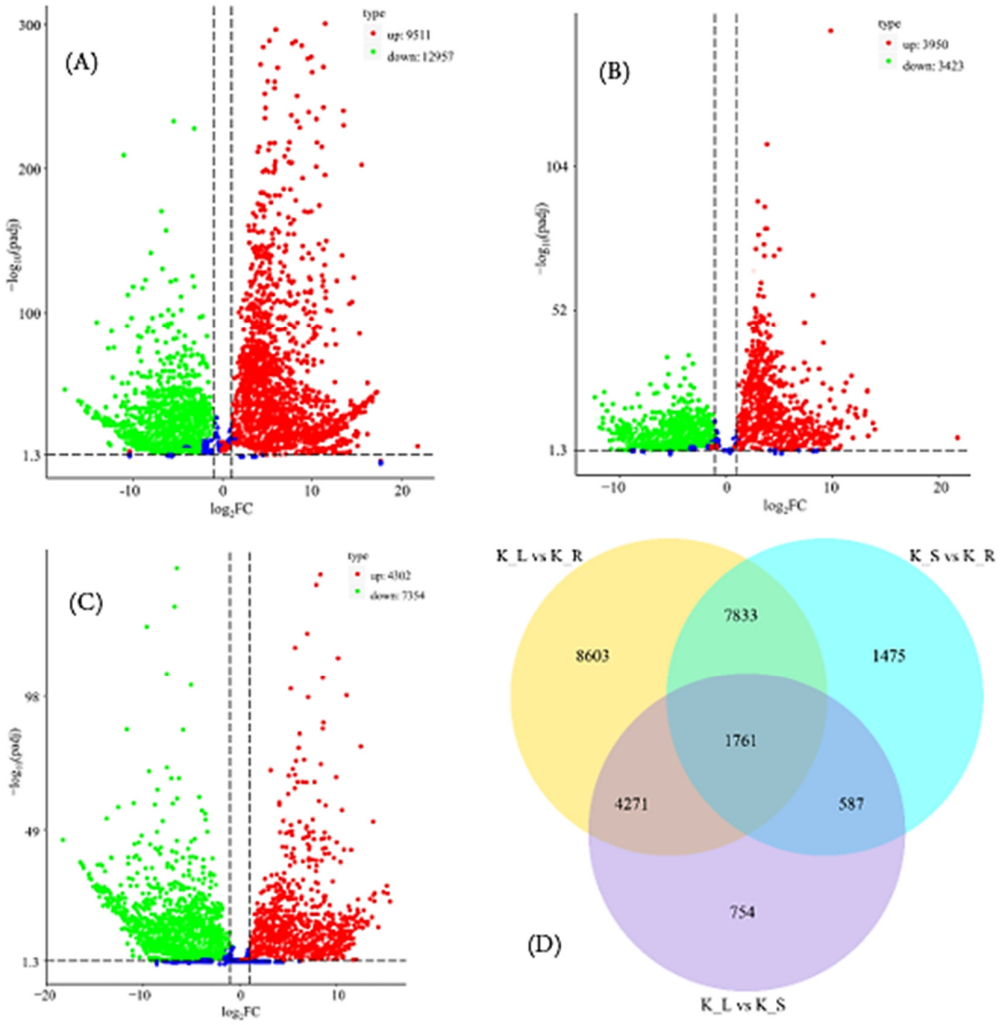

Figure 8. DEGs in different comparisons

(A)-(C) Volcano plots of the DEGs in different comparisons. The red and green dots denote significantly upregulated and the downregulated genes, separately, while the blue ones represent non-DEGs. (A) K_L vs. K_R volcano; (B) K_L vs. K_S volcano; (C) K_S vs. K_R volcano. (D) Venn diagram of DEGs in different comparisons. All DEGs are clustered into three comparison groups represented by three circles. Overlapping parts of the different circles represent the number of DEGs in common among those groups.

Biosynthetic genes of lignan in K. heteroclita

The annotation of *K. heteroclita* genes against the KEGG database has resulted in a total of 191 unigenes assigned to the lignan biosynthesis pathway (Table 4). Interestingly, out of these 191 genes, 28 genes code for DIR (dirigent protein), 25 genes code for CAD (cinnamyl alcohol dehydrogenase), 24 genes code for SDH (shikimic acid dehydrogenase), 24 genes code for 4CL (4-Coumarate-CoA Ligase), and only 1 gene codes for CSE (Caffeoylshikimate esterase) respectively. Results obtained from our gene expression study reports that genes involved in lignan biosynthesis pathways were highly expressed in root samples. Specifically, we have observed that genes encoding for PAL (phenylalanine ammonia-lyase), C4H(Cinnamate-4-Hydroxylase), 4CL, CAD, CCOMT (caffeoyl CoA 3-O-methyltransferase), DIR and SDH were significant and highly expressed in root samples compared to leaf and stem (Figure 10, Table 4).

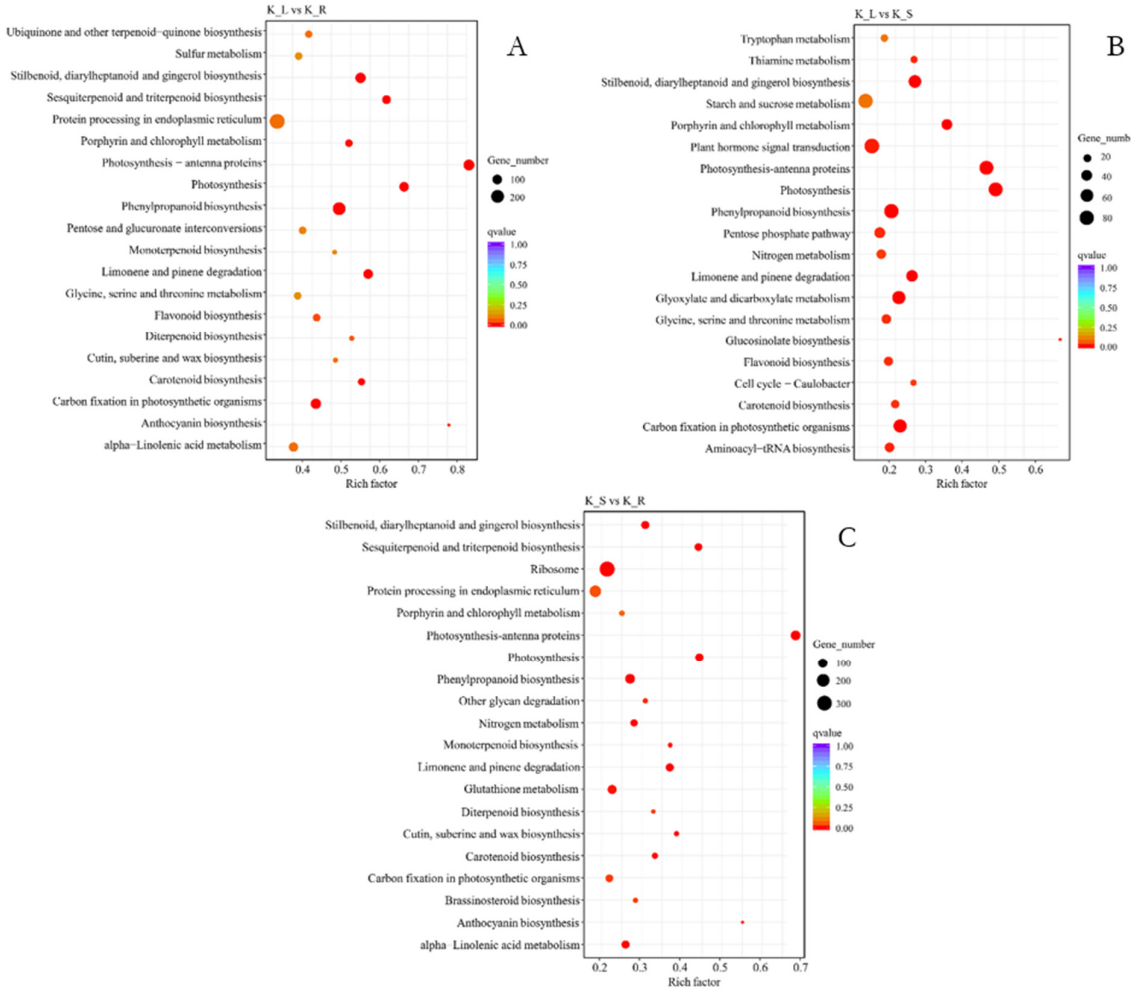

Figure 9. Top 20 KEGG pathway enrichment of DEGs
(A) C_L vs. C_R; (B) C_L vs. C_S; (C). C_S vs. C_R. The ordinate represents the pathway name, and the abscissa denotes the enrichment factor corresponding to the pathway. The q-value is represented by the color of the dot. The number of DEGs is represented by the size of the dots.

Interestingly, we have observed two genes coding for *Coq6* (Coenzyme Q6, Monooxygenase) were slightly up-regulated in root and stem samples with zero expression in leaf samples. These results indicated that caffeoyl-CoA was synthesized from p-cinnamoyl-CoA by HCT and C3'H during developmental stages of *K. heteroclita* (Figure 10, Table 4). We also predicted that *K. heteroclita* employs *CCOMT* instead of *COMT* gene for the synthesis of feruloyl-CoA based on the gene expression ratios of *CCOMT/COMT* in roots (4.77), stems (3.19) and leaves (1.07). Using the information obtained from our current transcriptome study and by using the pre-existing literature we have developed the tentative lignan biosynthesis pathway in *K. heteroclita* (Figure 10a).

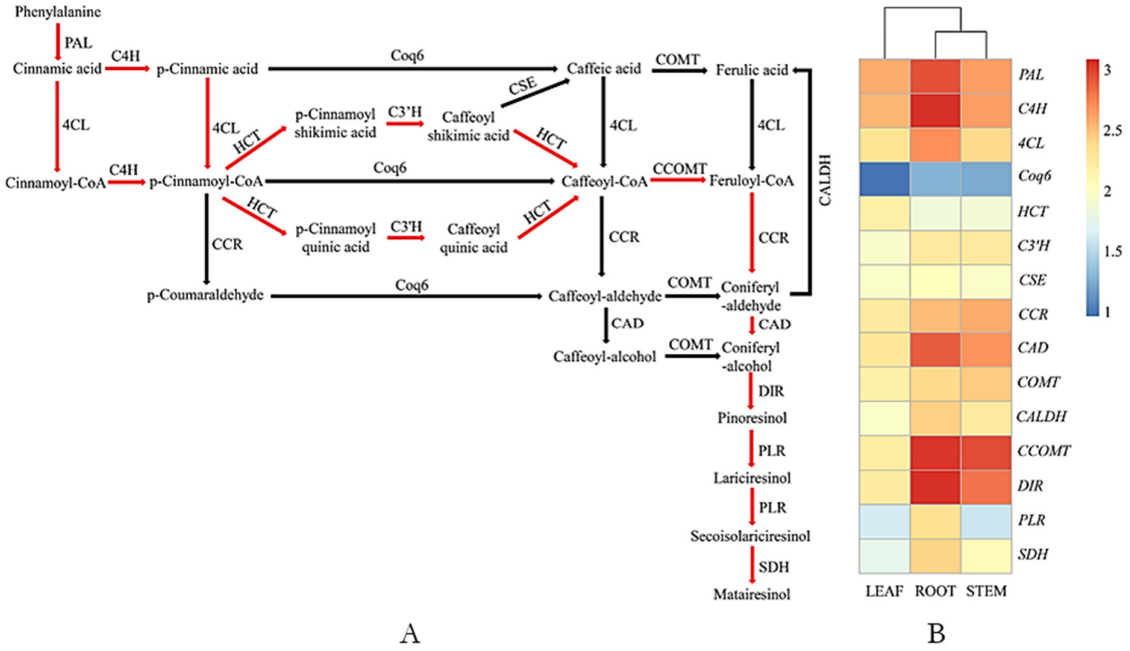

Figure 10. Putative pathway of lignan biosynthesis and expression of unigenes in *K. heteroclita*
(A) Proposed biosynthetic pathway of lignan. The arrows denote the deduced biosynthetic pathway of lignan. (B) Heatmap based on the expression level of unigenes involving in lignan biosynthesis across three tissues in *K. heteroclita*. The expression level is the sum of all the unigenes for each gene, and $\log_{10}$ (sum(FPKM) + 1) was used to plot the heatmap. Candidate unigenes come from the annotation.

Table 4. Putative genes of lignan biosynthesis pathway

Gene	Gene Name	EC	Number	Up (L/R)	Down (L/R)	Up (L/S)	Down (L/S)	Up (S/R)	Down (S/R)
PAL	Phenylalanine ammonia-lyase	4.3.1.24	12	-	5	7	-	-	2
C4H	Cinnamic acid 4-hydroxylase	1.14.13.11	16	-	1	-	-	-	-
4CL	4-coumarate-CoA ligase	6.2.1.12	24	2	11	-	-	-	5
Coq6	Ubiquinone biosynthesis monooxygenase Coq6	1.14.13.-	2	-	-	-	-	-	-
HCT	Shikimate-O-hydroxycinnamoyl transferase	2.3.1.133	22	10	1	2	-	1	-
C3'H	5-O-(4-coumaroyl)-D-quinate-3'-monooxygenase	1.14.14.96	7	1	1	-	-	-	-
CSE	Caffeoyl shikimate esterase	3.1.1.-	1	-	1	-	-	-	-
CCR	Cinnamoyl-CoA reductase	1.2.1.44	3	-	-	-	-	-	-
CAD	Cinnamyl-alcohol dehydrogenase	1.1.1.195	25	2	8	-	-	2	4
COMT	Caffeate O-methyltransferase	2.1.1.68	5	-	-	-	-	-	-
CALDH	Coniferyl-aldehyde dehydrogenase	1.2.1.68	6	-	-	-	-	-	1
CCOMT	Caffeoyl-CoA O-methyltransferase	2.1.1.104	12	1	7	1	-	-	3
DIR	Dirigent protein	-	28	-	17	-	-	-	14
PLR	Pinoresinol lariciresinol reductase	-	2	-	1	-	-	-	1
SDH	Secoisolariciresinol dehydrogenase	-	24	1	8	4	17	1	7
	Total		191	18	61	14	17	4	37

Note: up and down are refer to the value of $\log_2$ (FoldChange) is above or below 2.

Among the list of significant DEGs, 28 genes code for DIRs (Table S7). Interestingly, the *K. heteroclita* genome codes double the number of DIR genes compared to its closely related species *Schisandra chinensis* (Chen *et al.*, 2020). Earlier studies have revealed and classified plant DIR proteins into six subfamilies : DIR-a, DIR-b/d, DIR-c, DIR-e, DIR-f and DIR-g (Ralph *et al.*, 2007), interestingly these subfamilies exhibits very low sequence similarity between them (Song and Peng, 2019). Kim *et al.* (2002) have reported that only the DIR-a subfamily can guide the cellular machinery to synthesis correct three-dimensional structure of pinoresinol (Kim *et al.*, 2002). While the functional involvement of other DIR-like subfamily proteins are unclear (Kim *et al.*, 2002). Among these 28 significant *KhDIR* genes: 17 genes belong to DIR-b/d, 4 genes belong to DIR-c, 6 genes belong to DIR-e and only 1 gene belongs to DIR-a subfamilies respectively (Figure 11A). These results suggest that the only one gene coding for KhDIR4 which belongs to DIR-a subfamily is responsible for the formation of pinoresinol in *K. heteroclita*. However, the expression levels of KhDIR4 gene are comparatively lesser than other DIR encoding genes (Figure 11).

Kim et al. (2012) has reported that DIR genes expressed in different phylogenetically related species of *K. heteroclita* such as *SchDIR1, FiDIR, TpDIR5* and *TpDIR8* transcribed for (+)-pinoresinol-forming proteins, and two other candidate AtDIR5 and AtDIR6 transcribed for *(-)*-pinoresinol formation (Kim *et al.*, 2012). Research studies based on site directed mutagenesis, protein modelling and docking experiments have revealed that three phenylalanines F90, F113, F163 and amino acid sequence starting from Asn98 to Pro146 in SchDIR1 protein are critical for its (+)-pinoresinol-forming activity (Kim *et al.*, 2012).

Results obtained from the phylogenetic analysis of DIR protein sequences have revealed that KhDIR4 and SchDIR1 shares same clade exhibiting 92.82 % similarity (Figure 11). Interestingly, KhDIR4 protein also possesses the three key phenylalanine residues, and exhibits almost similar three-dimensional configuration as SchDIR1 protein (Figure 12A-E). These results suggest that the role of KhDIR4 in stereoselective synthesis of (+)-pinoresinol from coniferyl alcohol. Li et al. (2017) has been reported the functional involvement of GmDIR22 (belonging to DIR-b/d subfamily) in synthesis of lignan (+)-pinoresinol by effectively directing *E*-coniferyl alcohol coupling (Li *et al.*, 2017). We have observed from our results that, among the 28 DIRs three homologs KhDIR8, KhDIR19 and KhDIR20 are in the same clade with GmDIR22, suggesting their similar functions as GmDIR22 respectively (Figure 11A). The remaining DIR proteins: KhDIR1, KhDIR2, KhDIR6, KhDIR9, KhDIR11, KhDIR12, KhDIR14, KhDIR21-26, KhDIR28 belonged to the same clade of GhDIR4 (*Gossypium hirsutum*). According to Effenberger *et al.* (2015), GhDIR4 protein plays a crucial role in conversion of two hemigossypols into one (+)-gossypol (Effenberger *et al.*, 2015). Zhao *et al.* (2020) has reported that MtDIR exhibited 40% higher expression levels in root compared to other plant parts, while we have observed similar behavior in gene expression level ranging as high as 57% in *K. heteroclita* root samples respectively (Zhao *et al.*, 2020). Thus, from these results we suspect the involvement of *KhDIR8*, *KhDIR22*, *KhDIR23* and *KhDIR28* genes in (+)-pinoresinol synthesis (Figure 11B).

Previous studies have reported that PLR protein plays a crucial role in production of lariciresinol and/or secoisolariciresinol from pinoresinol (Markulin *et al.*, 2019). However, the substrates and products involved in its production are species-dependent. For example, the substrates for LuPLR1 and TpPLR1 proteins are (-)-pinoresinol and (-)-lariciresinol which leads to the product (+)-secoisolariciresinol (Fujita *et al.*, 1999; von Heimendahl *et al.*, 2005). Similarly, the substrates for LuPLR2, TcPLR2.2, TcPLR3, TpPLR2 and PpPLR proteins are (+)-pinoresinol, (+)-lariciresinol but the product is (-)-secoisolariciresinol respectively (Fujita *et al.*, 1999; Hemmati *et al.*, 2010; Chiang *et al.*, 2019). It is still unknown that whether these two steps are catalyzed by one PLR or two different PLRs. We have observed two PLR candidate genes in *K. heteroclita* transcriptome. Further studies must be conducted to understand the function and mechanism of these PLR genes in production of lariciresinol/secoisolariciresinol.

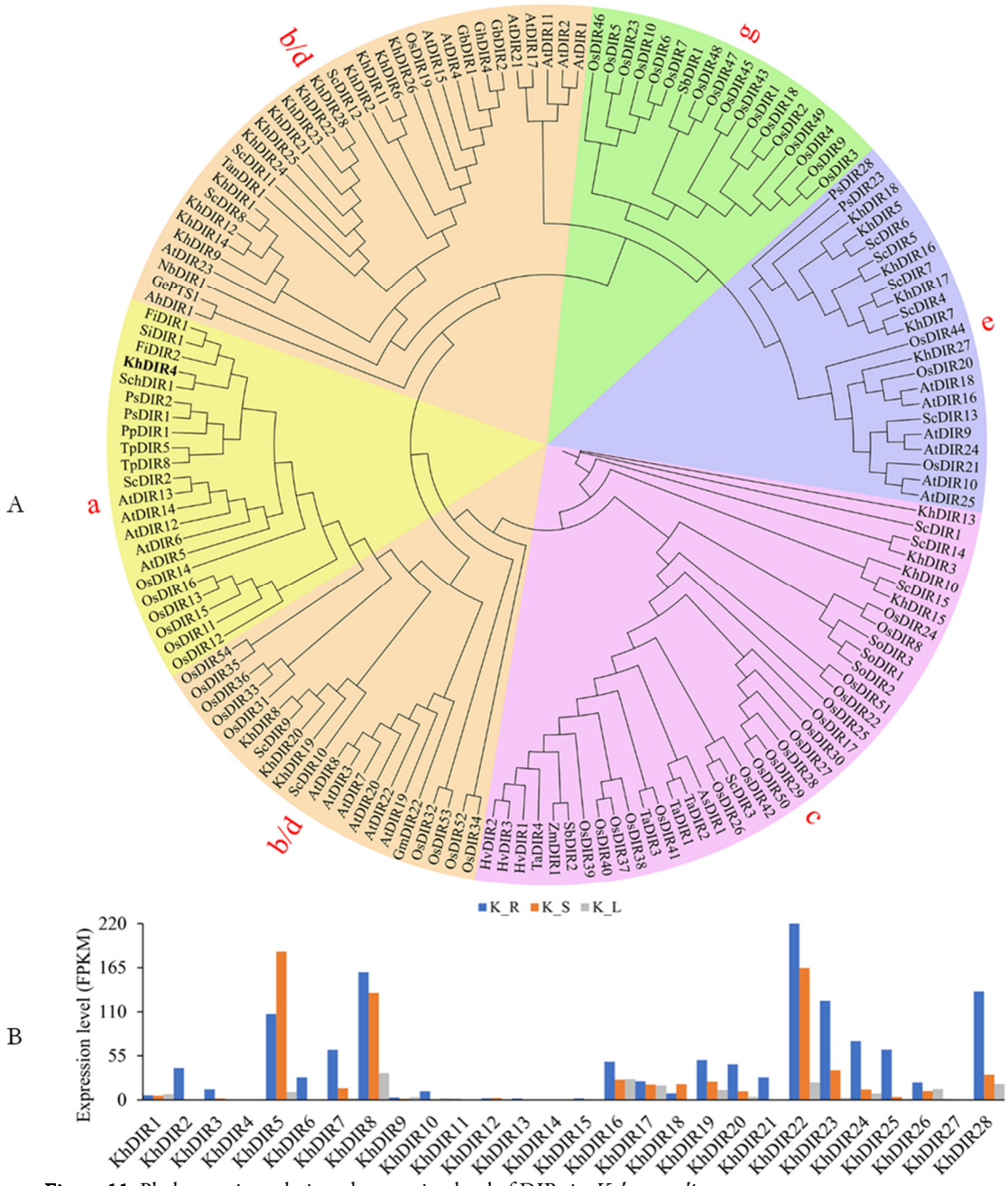

Figure 11. Phylogenetic analysis and expression level of DIRs in *K. heteroclita*
(A) Phylogenetic analysis of KhDIRs. The sequences used here are listed in Table S7. (B) Expression level of KhDIRs. Expression level for each gene is represented by the value of average RPKM in roots, stems, and leaves, separately.

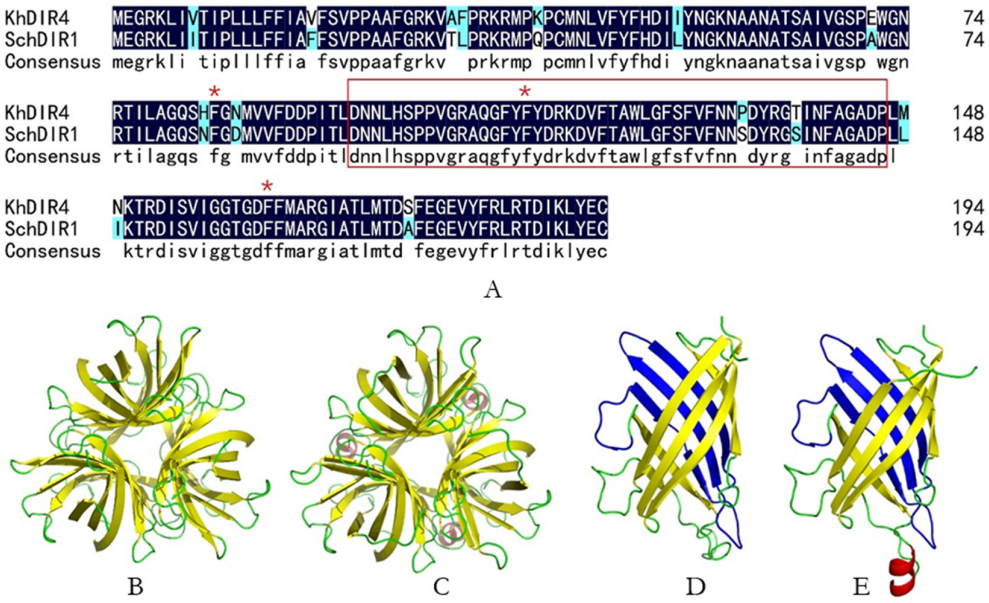

Figure 12. Alignment and three-dimensional model of KhDIR4 and SchDIR4 in *K. heteroclita*
(A) Alignment of KhDIR4 and SchDIR4. (B) Trimer model of KhDIR4. (C) Trimer model of SchDIR4. (D) Monomer model of KhDIR4, which is derived from (B). (E) Monomer model of SchDIR4, which is derived from (C).

Finally, we have observed 24 significant DEGs in *K. heteroclita* transcriptome which are coding for SDH protein. Xia *et al.* (2001) has reported that, NAD-dependent SDH protein of *Forsythia intermedia* (FiSDH) and *Podophyllum peltatum* (PpSDH) converts (-)-secoisolariciresinol into dibenzylbutyrolactone lignan (-)-matairesinol (Xia *et al.*, 2001). Results obtained from our sequence alignment studies involving KhSDH, FiSDH and PpSDH showed that they have commonly exhibited the conserved glycine-rich motif GXXGXG which was reported to be involved in binding of the pyrophosphate group of NAD^+. Also, these proteins commonly possess D47 found in SDRs which preferentially binds to NAD(H) (Figure 13A). The catalytic triad Ser153, Tyr167, and Lys171 adjacent to both NAD^+ and substrate molecules in PpSDH (Youn *et al.*, 2005) were also observed in KhSDH proteins (except the KhSDH9 protein) (Figure 13A). Results obtained from phylogenetic analysis has revealed that KhSDH17, KhSDH6, KhSDH7 proteins shared same clade with FiSDH and ShSDH2 (Figure S1). Based on these results and from these gene expression patterns we report that both KhSDH3 and KhSDH8 genes are comparatively more active during developmental stages (Figure 13B).

Biosynthetic genes of the sesquiterpenoid and triterpenoid in K. heteroclita

Naturally, sesquiterpenoid and triterpenoid are synthesized from methylerythritol phosphate (MEP) and mevalonate (MVA) pathways respectively. The genome-wide annotation of *K. heteroclite* against the KEGG database has resulted in a total of 124 genes assigned to the terpenoid backbone biosynthetic pathways, out of which 64 genes coding for 8 enzymes involved in MEP pathway and 60 genes coding for 6 enzymes involved in MVA pathway respectively (Figure 14A, Table 5). Further classifying these 124 genes has showed that 32 genes code for DXS, 26 genes code for HMGS, 15 genes code for HMGR, 14 genes code for AACT and 1 gene each for PMK, MVD, MCT and CMK respectively. The expression patterns of these genes suggest that MEP pathway is highly active in root, stem and leaf, whereas MVA pathway is highly active in root and stem respectively (Figure 14B). The genome-wide annotation of *K. heteroclite* against the KEGG database has resulted in a total of 150 genes assigned to the sesquiterpenoid biosynthetic pathway. Out of these 150 genes, 16 genes are significantly upregulated and 67 genes significantly downregulated in leaf vs. root, and 2 genes significantly upregulated and 13 genes significantly downregulated in leaf vs. stem (Figure14A, Table 5).

Further classifying these 150 genes has showed that: 121 genes code for GDS, 13 genes code for GPPS2, 6 genes code for GPPS, 4 genes code for FPPS and 1 gene each for PO, SS and VS respectively (Table 5). Among these significant DEGs, we have observed that six genes *FPPS*, *VES*, *PO*, *GDS*, *SS* and *VS* were significantly upregulated in root compared to stem and leaf samples (Figure 14B, Table 5). These results suggest that sesquiterpenoid pathway is highly active in root samples (Figure14, Table 5).

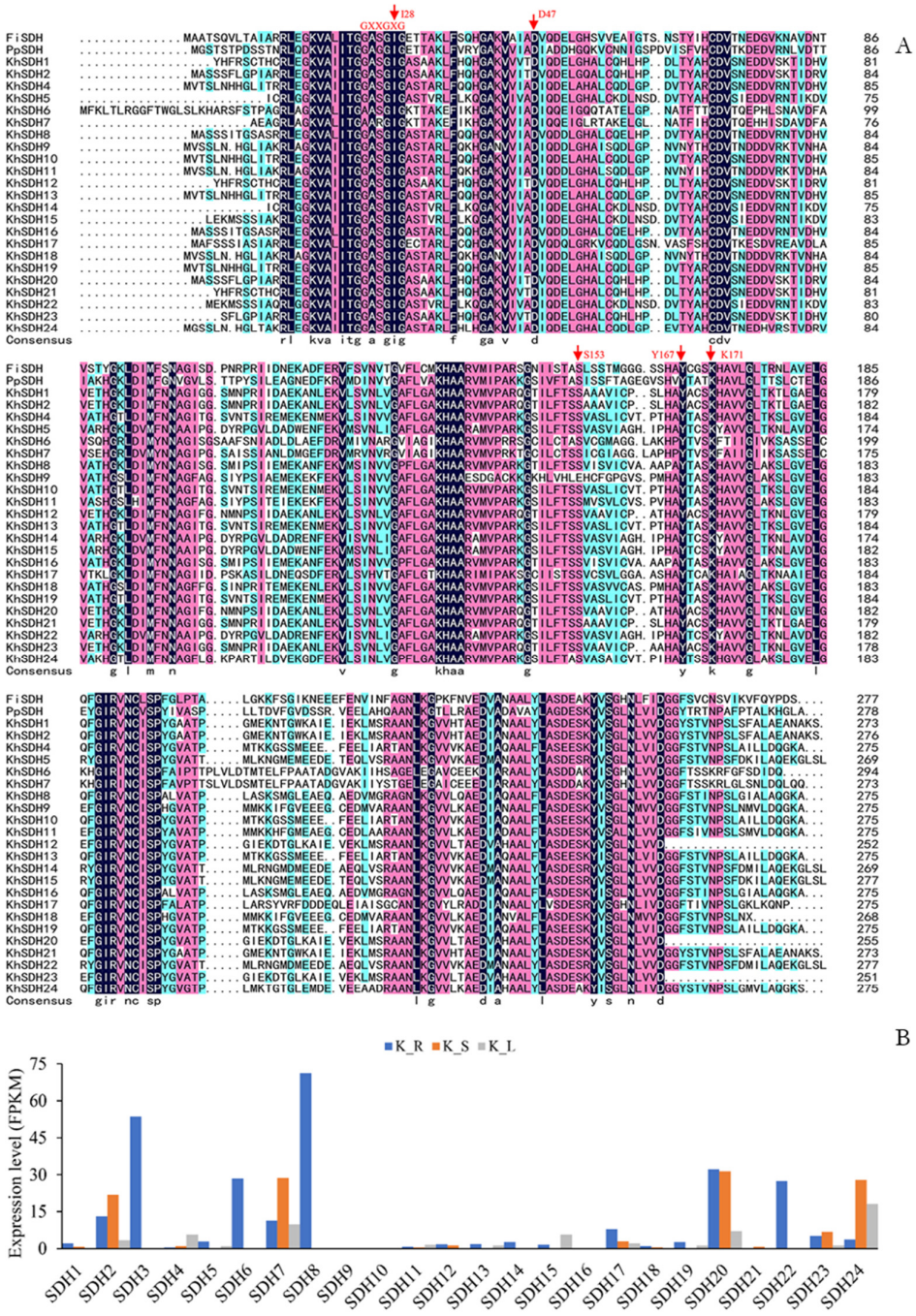

Figure 13. Alignment and expression level of KhSDHs in *K. heteroclita*
(A) Alignment of KhSDHs with FiSDH *Forsythia intermedia* and PpSDH in *Podophyllum peltatum*. The accession numbers are 24 KhSDHs (MT725696-MT725719), FiSDH(Q94KL7.1) and PpSDH(Q94KL8.1). (B) Expression level of KhSDHs.

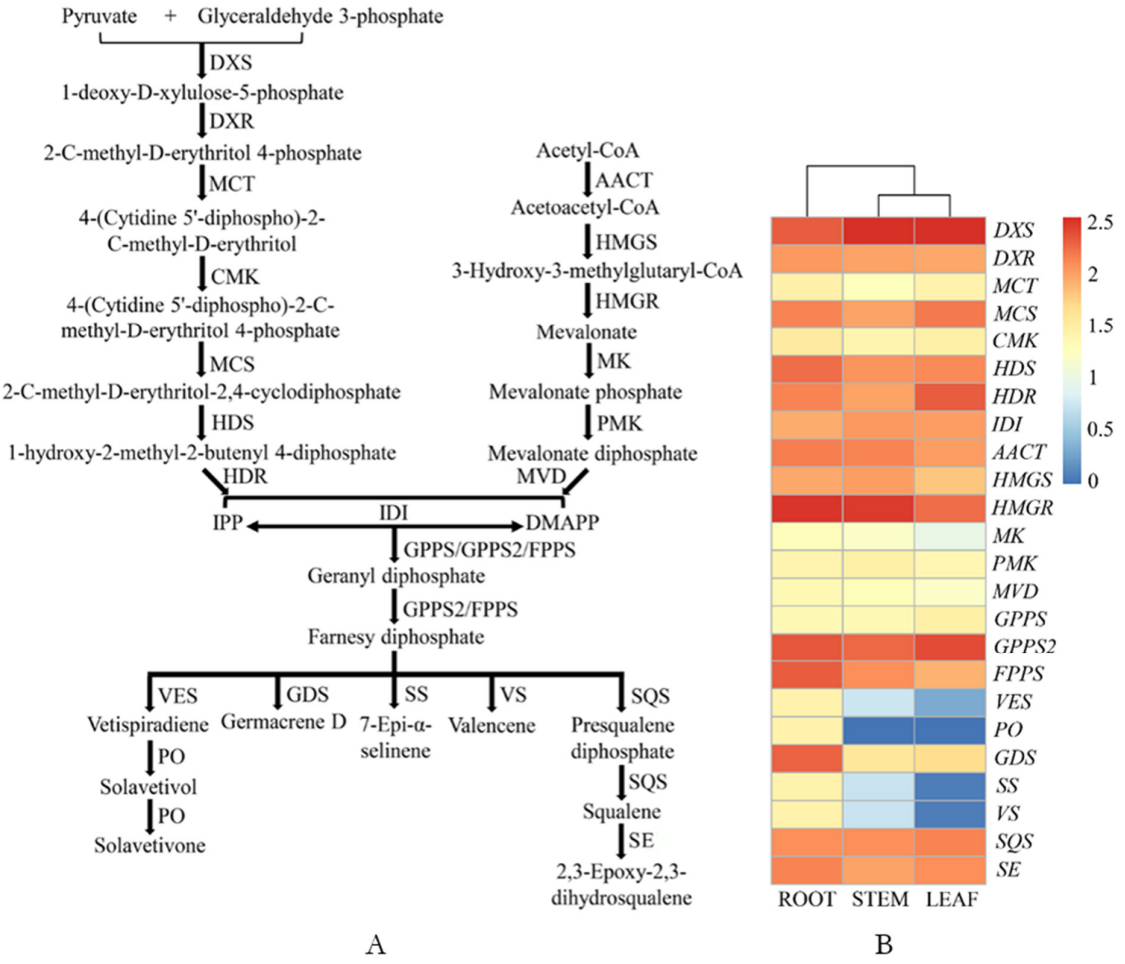

Figure 14. Expression of unigenes in the putative pathway of sesquiterpenoid and triterpenoid biosynthesis in *K. heteroclita*

(A) Proposed biosynthetic pathway of sesquiterpenoid and triterpenoid. (B) Heatmap based on the expression level of unigenes involved in sesquiterpenoid and triterpenoid biosynthesis across three tissues in *K. heteroclita*. Abbreviations: IPP, isopentenyl pyrophosphate; DMAPP, dimethylallyl pyrophosphate.

Identification of transcription factors

Transcription factors play an important role in regulation of plant secondary metabolism by activating or inhibiting the expression of functional genes involved in various biosynthesis pathways (Nag *et al.*, 2020). We have used the software HMMER3.0 to specifically search the TFs against the annotated transcriptome of *K. heteroclita*. Results obtained from the HMMER3.0 search have revealed that 4,528 genes coding for various TFs are belonging to 82 categories. Among these 4,528 genes, 297 (6.16% of total TFs) genes code for MYB, 274 (6.05%) genes code for mTERF, 247 (5.45%) genes code for bHLH, 198 (4.37%) genes code for NAC and 192 (4.24%) genes code for AP2-EREBP respectively (Figure 15, Table 6, Table S8). We have also observed the genes coding for the bZIP and WRKY TFs in *K. heteroclita* transcriptome (Table 6). In summary, the majority of the above-mentioned TFs are highly upregulated in root and stem compared with leaf samples (Table 6).

Table 5. Putative genes of MVA, MEP, sesquiterpenoid and triterpenoid biosynthesis pathways

Pathway	Gene	Gene name	EC	Number	Up (L/R)	Down (L/R)	Up (L/S)	Down (L/S)	Up (S/R)	Down(S/R)
MVA	*AACT*	Acetyl-CoA C-acetyltransferase	2.1.3.9	14	-	3	-	-	-	2
	HMGS	Hydroxymethylglutaryl-CoA synthase	2.3.3.10	26	7	1	-	-	-	-
	HMGR	Hydroxymethylglutaryl-CoA reductase	1.1.1.34	15	-	2	-	2	1	-
	MK	Mevalonate kinase	2.7.1.36	3	-	-	-	-	-	-
	PMK	Phosphomevalonate kinase	2.7.4.2	1	-	-	-	-	-	-
	MVD	Diphosphomevalonate decarboxylase	4.1.1.33	1	-	-	-	-	-	-
MEP	*DXS*	1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	32	-	-	-	2	3	3
	DXR	1-deoxy-D-xylulose-5-phosphate reductoisomerase	1.1.1.267	4	-	-	-	-	-	-
	MCT	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	2.7.7.60	1	-	-	-	-	-	-
	CMK	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	2.7.1.148	1	-	-	-	-	-	-
	MCS	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4.6.1.12	8	-	-	-	-	-	-
	HDS	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	1.17.7.1 1.17.7.3	12	-	-	-	-	-	-
	HDR	4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase	1.17.7.4	2	1	-	-	-	-	-
	IDI	Isopentenyl pyrophosphate isomerase	5.3.3.2	4	2	-	2	-	-	-
Sesquiterpenoid	*GPPS*	Geranyl diphosphate synthase	2.5.1.1	6	1	-	-	-	-	-
	GPPS2	Geranyl diphosphate synthase II	2.5.1.1; 2.5.1.10; 2.5.1.20	13	1	2	-	-	1	2
	FPPS	Farnesyl diphosphate synthase	2.5.1.1 2.5.1.10	4	-	-	-	-	-	-
	VES	Vetispiradiene synthase	4.2.3.21	3	-	2	-	1	-	1
	PO	Premnaspirodiene oxygenase	1.14.14.151	1	-	1	-	-	-	1
	GDS	Germacrene D synthase	4.2.3.75	121	14	60	2	12	8	43
	SS	7-epi-alpha-selinene synthase	4.2.3.86	1	-	1	-	-	-	1
	VS	Valencene synthase	4.2.3.73	1	-	1	-	-	-	1
Triterpenoid	*SQS*	Squalene synthase	2.5.1.21	1			-	-	-	-
	SE	Squalene monooxygenase	1.14.14.17	4	3		-	3	3	-

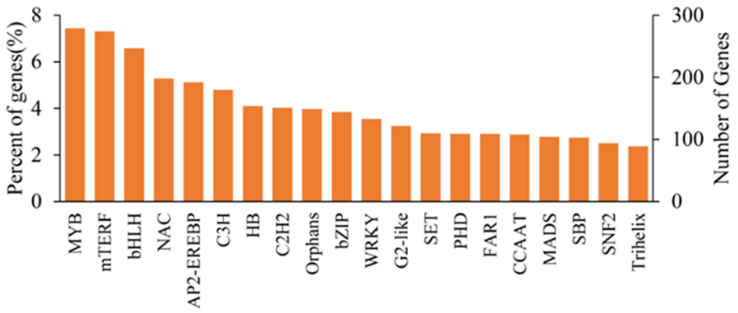

Figure 15. Top 20 TFs in *K. heteroclita*

Table 6. Summary of TF unigenes of *K. heteroclita*

TF Family	Number	Up (L/R)	Down (L/R)	Up (L/S)	Down (L/S)	Up (S/R)	Down (S/R)
MYB	279	30	49	12	29	21	18
mTERF	274	17	3	10	1	5	
bHLH	247	28	30	7	16	19	13
NAC	198	15	41	8	18	11	18
AP2-EREBP	192	20	33	4	19	22	14
bZIP	144	11	22	1	6	8	13
WRKY	133	12	32	2	13	8	10
Total	1467	133	210	44	102	94	86

Recent studies have shown that lignan and terpenoid biosynthesis are highly regulated by MYBs (Deng and Lu, 2017; Matías-Hernández *et al.*, 2017; Chiang *et al.*, 2019). In order to understand and reveal the TFs involved in *K. heteroclita's* lignan and terpenoid biosynthesis, we have performed phylogenetic analysis of all 279 MYB TFs against 37 MYB TF sequences belonging to other plant species. Based on the results obtained from our analysis we have classified MYB TFs into four categories each containing 2, 22, 92 and 163 members respectively (Figure 16). Recent study based on MYB TFs of *Taiwania cryptomerioides* has revealed that TcMYB1, TcMYB4, and TcMYB8 positively regulate lignan biosynthesis by activating the gene expressions of *TcPLR3* and *TcPLR1* respectively (Chiang *et al.*, 2019). Results obtained from our phylogenetic analysis have showed 4 *K. heteroclita* MYBs KhMYB118, KhMYB37, KhMYB203 and KhMYB36 share the same clade S6 with TcMYB1, while KhMYB249, KhMYB167, KhMYB199, KhMYB4 and KhMYB145 are in the same S7 clade with TcMYB4 and TcMYB8 (Figure 16). These results suggest that above-mentioned MYB TFs regulate lignan biosynthesis in *K. heteroclita*.

Deng and Lu (2017) reported 14 MYB TFs of *A. thaliana* (AtMYB4, AtMYB26, AtMYB32, AtMYB43, AtMYB46, AtMYB52, AtMYB54, AtMYB58, AtMYB61, AtMYB63, AtPAP1, AtMYB83, AtMYB85, and AtMYB103) which are involved in the biosynthesis of monolignols, building blocks of both lignan and lignin (Deng and Lu, 2017). The phylogenetic analysis of KhMYB against the above-mentioned *A. thaliana* MYB TFs has resulted in a total of 41 TFs exhibiting high similarity between *A. thaliana* and *K. heteroclita* MYB TFs, which implies their involvement in regulation of lignan biosynthesis. Also, recent studies have reported the regulatory effect of MYB TFs on terpenoid biosynthesis in plants. In *Artemisia annua*, *AaMYB1* can promote the artemisinin biosynthesis by activating the expression of *FDS* (*farnesyl diphosphate synthase*), *ADS*, *CYP71AV1*, *DBR2* and *ALDH1* genes (Matías-Hernández *et al.*, 2017). The phylogenetic analysis has revealed that KhMYB4 and KhMYB145 TFs share the same S7 subclade with AaMYB1, indicating their possible regulatory role in sesquiterpenoid biosynthesis (Figure 16). Recent studies based on PtMYB4 in *Pinus taeda* and VvMYB5b in *Vitis vinifera* have reported that these TFs influenced the accumulation of terpenoids and phenylpropanoids in plants (Nagegowda and Gupta, 2020). Similarly, in *Actinidia deliciosa* the genes coding for AdMYBR2, AdMYBR3, AdMYB3, AdMYB7 and AdMYB8 TFs positively regulate

carotenoid biosynthesis via transcriptional activation of lycopene beta-cyclase gene (Ampomah-Dwamena *et al.*, 2019). These results suggest that KhMYBs in clade S2, S6, S8, S9, and S10 have the potential regulatory roles in terpenoid biosynthesis (Figure 16).

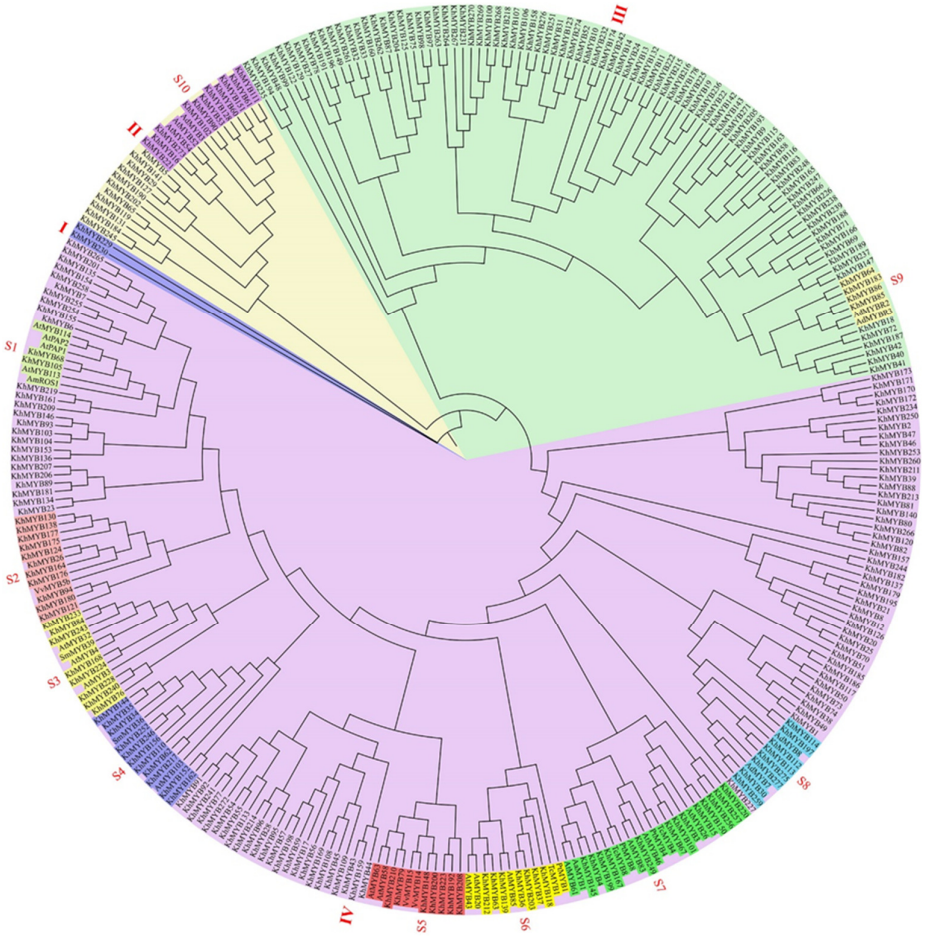

Figure 16. Phylogenetic analysis of MYBs in *K. heteroclita*
The sequences of *Arabidopsis thaliana* come from TAIR (https://www.arabidopsis.org/), while that of other plants are listed in Table S9.

The basic-helix-loop-helix (bHLH) is a class of TF which were also found to be involved in regulation of plant terpenoid biosynthesis by recruiting defined cis-regulatory elements which are part of a conserved model for jasmonate signaling pathway (Mertens *et al.*, 2016). Mertens *et al.* (2016), has reported about two candidate genes of *Medicago truncatula* coding for triterpene saponin biosynthesis activating regulators: MtTSAR1 and MtTSAR2 which play a key role in enhancing nonhemolytic and hemolytic soyasaponin biosynthesis by activating the corresponding pathway genes respectively (Mertens *et al.*, 2016). Recent study has reported that MtTSAR3 controls hemolytic saponin biosynthesis in developing seeds (Ribeiro *et al.*, 2020). Jarvis et al. (2020), has reported that *Chenopodium quinoa* genes encoding for CqTSARL1 in seed and CqTSARL2 in root also regulate the triterpenoid saponin biosynthesis (Jarvis *et al.*, 2017). Studies have also reported that *Catharanthus roseus*, a medicinal plant without triterpenoid saponin codes two bHLH TFs CrBIS1 and CrBIS2 which can regulate the biosynthesis of monoterpenoid indole alkaloid (Van Moerkercke *et al.*, 2015; Van Moerkercke *et al.*, 2016). The transcriptome of *K. heteroclita* totally codes for 247 bHLH TFs genes (Table 6), out of which 179 putative KhbHLH are obtained after removing the identical proteins which are transcribed by different genes. We have totally observed ten classical bHLH clades Ia, Ib, IIIc, IIId,

IIIe, IIIf, IVa, IVb, IVc and IVd in our phylogenetic tree (Figure 17, Mertens *et al.*, 2016). Results obtained in our study shows that KhbHLH54 and KhbHLH154 cluster with CqTSARL1, CqTSARL2, CrBIS1, CrBIS2, MtTSAR1, MtTSAR2 and MtTSAR3 belonging to IVa bHLH (Figure 17), indicating their potential functional involvement in sesquiterpenoid and triterpenoid biosynthesis in *K. heteroclita*. Study conducted on *Aquilaria sinensis* has reported that AsMYC2 promotes the agarwood sesquiterpene biosynthesis by activating the expression of *ASS1* (*Aquilaria sesquiterpene synthase 1*) gene through the jasmonate signaling pathway (Xu *et al.*, 2017). Similar study conducted on *Solanum lycopersicum*, SlMYC1 reported that it positively regulates the monoterpene biosynthesis in both leaf and stem trichomes but negatively regulate the sesquiterpene biosynthesis in stem trichomes (Xu *et al.*, 2018). However, in *Catharanthus roseus*, CrMYC2 promotes the accumulation of monoterpenoid indole alkaloids in shoot (Schweizer *et al.*, 2018). Results obtained in our study are in accordance as the protein sequences encoding for KhbHLH114, KhbHLH134, KhbHLH60 and KhbHLH114 were found to cluster with CrMYC2, SlMYC1 and AsMYC2 in the IIIe clade (Figure 17), suggesting their role in sesquiterpenoid and triterpenoid biosynthesis in *K. heteroclita*. In *A. thaliana*, clade IIId bHLH TFs AtbHLH3, AtbHLH13, AtbHLH14 and AtbHLH17 acted as transcription repressors against the transcription activators, such as MYC2 and the MYB/bHLH/WD40 complex by binding to their target sequences (Song *et al.*, 2013). Results obtained in our study shows that TFs encoding for KhbHLH60, KhbHLH163, KhbHLH162 and KhbHLH169 cluster with AtbHLH3, AtbHLH13, AtbHLH14 and AtbHLH17 in the IIId clade (Figure 17), which indicates their role as a negative regulator of the sesquiterpenoid and triterpenoid biosynthesis pathway in *K. heteroclita*.

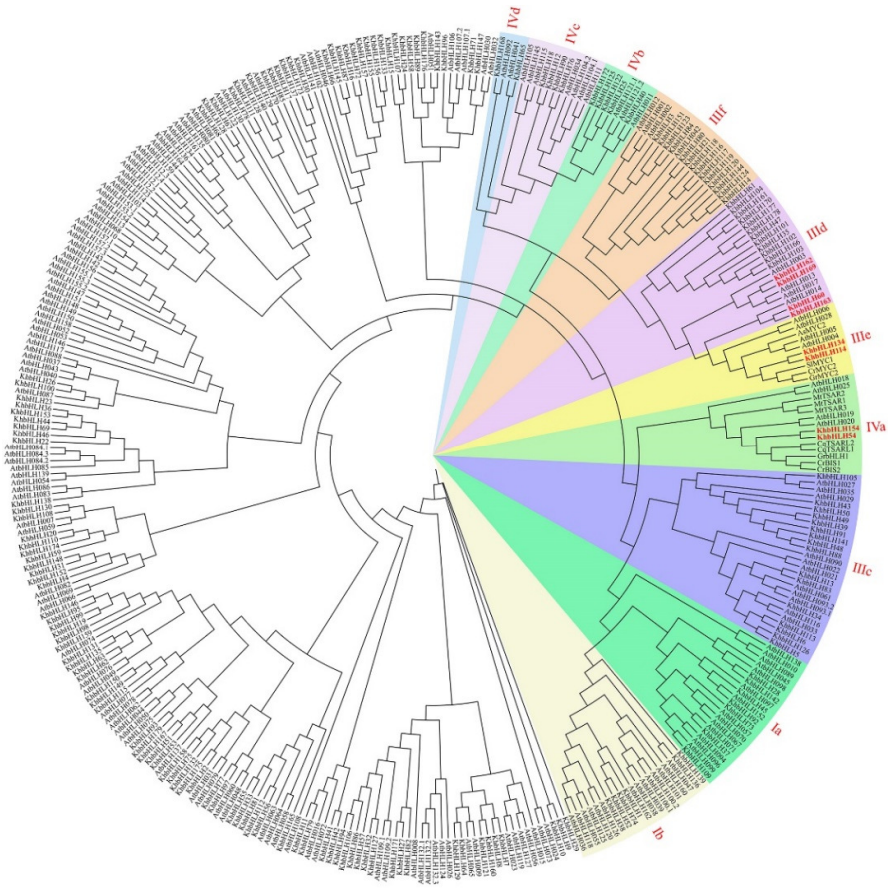

Figure 17. Phylogenetic analysis of bHLHs in *K. heteroclita*. (a) Phylogenetic analysis of KhbHLHs. The sequences of *Arabidopsis thaliana* come from TAIR (https://www.arabidopsis.org/), while that of other plants are listed in Table S10

Discussion

As of August 11th, 2020, 19,936,210 cases of CoVID-19 infection and 732,499 deaths have been confirmed worldwide (WHO, 2020). However, this epidemic has been effectively controlled in May, 2020 in China, whose anti-epidemic experience has verified the effectiveness of traditional Chinese medicine (TCM) against the new coronavirus (State Council Information Office of the PRC, 2020). Therefore, strengthening basic research on TCM will provide a solid strategic reserve for future epidemic prevention and control. *K. heteroclita* is one of the most famous TCMs widely used in Tujia and Yao Nationality in China. Despite its extensive pharmacological effects and remarkable medicinal effects, it has not been collected in the Chinese Pharmacopoeia because of lacking in-depth basic research, which also prevent its national and global use (Zhang *et al.*, 2019). At the same time, several factors are leading to the incorrect usage of *K. heteroclita* as the same species with different names and same name with different species were being found and circulated, especially *K. interior* being used for *K. heteroclita.* Thus, addressing such problems will significantly benefit the growing research on TCM. For the first time we have reported complete genome-wide transcriptome sequence of *K. hetroclita* to elucidate molecular mechanisms underlying the biosynthesis of lignan, sesquiterpenoid and triterpenoid.

Our present study reports the transcriptome of roots, stems and leaves samples obtained from *K. heteroclita* seedlings. A total of 160,248 transcripts and 98,005 genes were obtained after analysis, suggesting us that one gene may have more than one transcript. The N50 value is an important value used for comparing the *de novo* transcriptome assemblies (Sadat-Hosseini *et al.*, 2020). The average length of transcripts and genes were 927 bp and 1306 bp while their N50 were 1609 bp and 1838 bp respectively (Table 2). The average transcript length of *K. heteroclita* were found to be longer than the average transcript lengths of *Lolium multiflorum* and *Persea americana* (Ge *et al.*, 2019; Cechin *et al.*, 2020). These results indicated that *K. heteroclita* transcriptome assembly quality was good for further downstream analysis. Results obtained from the species distribution analysis of unigenes showed us three top matched species: *Nelumbo nucifera* (20.42%), *Macleaya cordata* (11.36%) and *Amborella trichopoda* (11.30%) (Figure 3). The whole genome sequences of these plant species closely related to *K. heteroclita* were not studied till date.

Earlier studies, have reported that main active components of *K. heteroclita* are lignan, sesquiterpenoid and triterpenoid (Zhang *et al.*, 2019). Lignans are widely distributed in plants and it imparts strength and disease resistance in plants. But most importantly lignans constitute the active pharmacological compounds used for wellbeing of human beings (Markulin *et al.*, 2019). Therefore, understanding the biosynthesis of lignan is gaining attention in the scientific community. In our study, we have revealed about the involvement of 191 unigenes (coding for 15 enzymes) in lignan biosynthetic pathway. Results obtained from gene expression profiling suggests that the lignan biosynthesis is more active in root, followed by stems and leaves samples respectively (Figure 10). The results obtained were in accordance with previous studies and endorses the usage of stems and roots medicinal compounds. Using the results obtained from this gene expression analysis we have deduced tentative lignan biosynthesis mechanism in *K. heteroclita* which commences via PAL, C4H, 4CL, HCT, C3'H, HCT, CCOMT, CCR, CAD, DIR, PLR and SDH as shown using red arrows in Figure 10A. It is well known that lignan is synthesized via the phenylpropanoid pathway, therefore we have specifically focused on the lignan pathway genes *KhDIR*, *KhPLR* and *KhSDH* in our analysis. Studies have reported that in *Urtica dioica* and *Schisandra chinensis*, tissue- and organ-specific distribution of lignans are dependent on the expression patterns of *DIR* and *PLR* genes (Xu *et al.*, 2019; Chen *et al.*, 2020). Results obtained from sequence alignment, phylogenetic analysis and gene expression analysis reveal that *KhDIR4* differentially expressed in roots, stems and leaves might be responsible for the conversion of coniferyl alcohol to (+)-pinoresinol. *K. heterocliata* transcriptome hosts two *KhPLR* genes, which supports the findings from previous reports, that more than one *PLR* coding genes exists in each plant (Markulin *et al.*, 2019). However, the number of *KhPLR* is far less than that of *KhDIR* and *KhSDH* genes indicating that *KhPLR* might be a rate-limiting enzyme in the lignan biosynthesis pathway of *K. heteroclita.* The relatively lower expression of *KhPLR*

gene in roots, stems and leaves samples further supports our hypothesis (Figure 10B). The expression of *PLR* gene is species- and organ-specific. In mature *Linum usitatissimum*, both *LuPLR1* and *LuPLR2* are expressed in seeds and roots, whereas only *LuPLR2* is expressed in stems and leaves (Hemmati *et al.*, 2010). Similarly, the seedlings of *U. dioica*, *UdPLR1*, *UdPLR2* and *UdPLR3* genes are highly expressed in top stems, roots and leaves, respectively (Xu *et al.*, 2019). It was also observed in the seedlings of *A. thaliana*, where *AtPrP1* is highly expressed in roots and stems whereas, *AtPrR2* is only expressed in roots (Nakatsubo *et al.*, 2008). Similar expression pattern was observed with *K. heteroclita* transcriptome, only *KhPLR1* gene is highly expressed in roots, with a lower expression in stems and leaves and *KhPLR2* gene expressions in all three tissues are found be very weak (Table S11).

Sesquiterpenoids and triterpenoids are derived from terpenoid backbone pathway which further consists of MEP and MVA pathways. Previous studies have revealed that in MEP pathway, DXS is the first and rate-limiting enzyme, and it expresses based on the feedback generated from the last metabolite. Importantly, the gene expression and activity of the first enzyme in a pathway is considered as a common regulatory mechanism (Banerjee and Sharkey, 2014). *K. heteroclita* genome codes for 32 *KhDXS* genes (Table 5) indicating that it is not a rate-limiting enzyme in *K. heteroclita* terpenoid pathway, and occurrence of so many DXS could avoid the feedback inhibition in *K. heteroclita*. Recent study conducted by Xue *et al.* (2019) have reported that in *Panax ginseng*, MCT is an rate-limiting enzyme in MEP pathway (Xue *et al.*, 2019). The transcriptome of *K. heteroclita* contains only one *KhMCT* gene and we have observed that the expression of *KhMCT* in root, stem and leaf samples were relatively low compared to other MEP pathway genes (Figure 14B, Table 5). The gene expression profile of *KhMCT* indicates that it is a rate-limiting gene in MEP pathway. Similarly, CMK could be another rate-limiting enzyme (Figure 14B, Table 5). Previous studies also suggest that plastidial IDI plays an important role in regulating the relative ratio of IPP to the DMADP, as they are the key precursors for various downstream pathways (Pankratov *et al.*, 2016). *K. heteroclita* transcriptome contains 4 *IDI* genes with two genes upregulated in leaves samples (Table 5). Although, there are no significant differences among the overall expression patterns of *IDI* genes in roots, stems and leaves (Figure 14B), suggesting that *IDI* might not be a rate-limiting gene. Similarly, *K. heteroclita* transcriptome contains two genes encoding for *KhHMGR* genes (key rate-limiting enzyme of MVA pathway) which were found to be highly up regulated in stem and root samples (Table 5). Thus, results obtained in our study shed light on a new topic of investigation focused on the relative contribution of MEP and MVA pathway, and its downstream biosynthesis of active medicinal compounds. Previous studies conducted on *Nothapodytes nimmoniana* has revealed that MEP pathway is a major route for production of camptothecin (an active medicinal compound) (Rather *et al.*, 2019). Whereas studies based on *P. ginseng* has reported that both MEP and MVA pathway contribute to the ginsenoside biosynthesis (Xue *et al.*, 2019). However, differential gene expression analysis of *K. heteroclita* leaf, stem and root samples revealed that genes encoding for MVA pathway genes such as *KhMK, KhPMK, KhMVD* are lowly expressed in comparison with MEP pathway genes (Figure 14B), suggesting that MEP pathway might play a crucial role in the production of precursors IPP and DMAPP involved in sesquiterpenoid and triterpenoid biosynthesis. Further studies must be conducted to understand and reveal the exact functional involvement of these genes.

Similarly, the FPP is a branchpoint for sesquiterpenoid and triterpenoid biosynthesis pathway. In the sesquiterpenoid pathway, FPP is cyclized into structurally diverse sesquiterpenoids by various sesquiterpene synthases, whereas in triterpenoid biosynthesis pathway FPP is dimerized into squalene by SQS (Kanehisa, 2020). Results obtained from differential gene expression analysis of *K. heteroclita* showed that genes encoding for GDS are highly expressed in root samples suggesting that the germacrene D is the main sesquiterpenoid product during this stage of development (Figure 14B). *K. heteroclita* transcriptome expresses only one *KhSQS* gene involved in the triterpenoid pathway. These results are in accordance with other species such as *Euphorbia pekinensis*, *Chlorophytum borivilianum*, *Euphorbia tirucalli*, *Taxus cuspidata*, *Lotus japonicus* and *Oryza sativa* respectively. Whereas there are more than one *SQS* gene in the genomes of *Glycyrrhiza glabra*, *Salvia miltiorrhiza* and *P. ginseng* (Rong *et al.*, 2016). In *Salvia miltiorrhiza*, the gene expression profiles of *SmSQS2*

in roots is almost twice the gene expression level of leaves samples (Rong *et al.*, 2016). Similarly, previous studies conducted on *Tripterygium wilfordii*, *Medicago sativa* and *Dryopteris fragrans*, *TwSQS*, *MsSQS* and *DfSQS1* have reported highest gene expression patterns in roots, followed by leaves and stem samples respectively (Zhang *et al.*, 2018; Gao *et al.*, 2019; Kang *et al.*, 2019). Interestingly in *K. heteroclita* transcriptome, *KhSQS* gene exhibited almost the same expression pattern in roots, stems and leaves samples (Figure 14B), suggesting the significance of *KhSQS* gene during *K. heteroclita* seedling development. Interestingly the results obtained from the phylogenetic analysis of KhSQS shows that it shares the same clade with monocots, although it belongs to eudicot, which implies its significance in the evolution from monocot to eudicot (Figure S2).

Squalene epoxidase is an important rate-limiting enzyme which is involved in conversion of squalene to 2,3-epoxy-2,3-dihydrosqualene during the biosynthesis of sterols and triterpenoids. The *T. wilfordii* genome contains a total of five *SE* genes, out of which *TwSE1-4* exhibited higher gene expression levels in the roots followed by stem samples, while *TwSE5* exhibited lower gene expression profiles in roots, stem and leaves and eventually lost its SE activity (Zhou *et al.*, 2018). Similar study conducted on *A. thaliana* reported that it contained six *SE* genes, among them *AtSQE1-3* encodes functional SEs whereas the *AtSQE4-6* gene does not code any functional genes, out of all these genes *AtSE1* is responsible for the root and seed development (Rasbery *et al.*, 2007). *K. heteroclita* transcriptome harbors four *KhSE* genes, while naturally most of the plants exhibit more than two *SE* genes (Rasbery *et al.*, 2007). *KhSE4* gene was found to be highly expressed in roots samples while *KhSE1-3* exhibited weak and lower gene expression profile (Figure S3). These results suggest us that only *KhSE4* is majorly involved in the biosynthesis of triterpenoid during the *K. heteroclita* developmental stage. Results obtained from the phylogenetic analysis has revealed that *KhSE1*-4 clusters with TwSE1-4, whereas the TwSE5 gene formed a single clade (Figure S4). These results mainly suggest that KhSE1-4 all other genes possessed the SE activity, and they are mostly organ-specific in *K. heteroclita*.

Recent studies have implemented transgenic technology for successfully enhancing the production of triterpenoid in plants. For example, over expression of *MsSQS* in alfalfa led to the accumulation of total saponins, suggesting a correlation between *MsSQS* expression level with saponins content respectively (Kang *et al.*, 2019). Similarly, in *Eleutherococcus Senticosus* gene expression of the *SQS* and *SE* genes are positively correlated with the saponin content (Wang *et al.*, 2019), *Ganoderma lingzhi*, and overexpression of *SE* doubled the production of ganoderic acid (Zhang *et al.*, 2017). Based on the occurrence of one *SQS* and the four *SE* genes and their expression profiles in *K. heteroclita* seedling samples, we believe that KhSQS and KhSE genes might exhibit great potential in increasing the active triterpenoid production. Based on the results obtained in our study we believe that *K. heteroclita* can synthesize various types of lignans, sesquiterpenoids and triterpenoids, indicating the diversity of genes and enzymes involved in their biosynthesis. For the first time our study provides preliminary information about the structural genes and possible regulatory genes involved in lignan, sesquiterpenoid and triterpenoid biosynthetic pathways in *K. heteroclita*. However, further studies must be conducted in future to figure out the downstream metabolic pathways by conducting other molecular and isotope tracing techniques.

Conclusions

Kadsura heteroclita is popularly known for its medicinal importance owing to its key ingredients used in TCM in Tujia and Yao Nationality. Although *K. heteroclita* has been used for centuries, the cellular and molecular by mechanisms involved in regulating the biosynthesis of the medicinal components, especially lignan, sesquiterpenoid and triterpenoid, is not known till date. Our present study reports genome-wide transcriptome of roots, stems and leaves samples are obtained from *K. heteroclita* seedlings. We have extensively reported about the genes involved in the lignan, sesquiterpenoid and triterpenoid biosynthetic pathways. Based on the information obtained from the gene expression profiles and sequence alignments we have reported a putative lignan biosynthetic pathway in *K. heteroclita.* Also, we have revealed that genes involved in these

biosynthetic pathways were highly upregulated in root and stem samples. We strongly believe that results obtained from *K. heteroclita* transcriptome analysis could significantly benefit future studies conducted on understanding the lignan, sesquiterpenoid and terpenoid biosynthetic pathways in *K. heteroclita.*

Authors' Contributions

Conceptualization: XZ and WQ; Investigation: XZ and CL; Methodology: XZ and CL; Formal analysis: XZ, CL, CC and TM; Funding acquisition: CL and XZ; Writing-original draft: CL, XZ and AK; and Writing-review and editing: WQ and CC. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the China Scholarship Council (grant number: 201908530057) and Start-up Fee for Doctoral Scientific Research at Xuchang University (grant number: 20201009).

Data availability: The biosamples and raw reads used in this study has been deposited in the NCBI BioProject under the accession number PRJNA631549.

Conflict of Interests

The authors declare that there are no conflicts of interest related to this article.

References

Ampomah-Dwamena C, Thrimawithana AH, Dejnoprat S, Lewis D, Espley RV, Allan AC (2019). A kiwifruit (*Actinidia deliciosa*) R2R3-MYB transcription factor modulates chlorophyll and carotenoid accumulation. New Phytologist 221(1):309-325. *https://doi.org/10.1111/nph.15362*

Banerjee A, Sharkey T (2014). Methylerythritol 4-phosphate (MEP) pathway metabolic regulation. Natural Product Reports 31(8):1043-1055. *https://doi.org/10.1039/C3NP70124G*

Cao L, Li B, Shehla N, Gong L, Jian Y, Peng C, ... Man R (2020). Triterpenoids from stems of *Kadsura heteroclita*. Fitoterapia 140:104441. *https://doi.org/10.1016/j.fitote.2019.104441*

Cao L, Shehla N, Li B, Jian Y, Peng C, Sheng W, ... Liao D (2020). Schinortriterpenoids from *Tujia* ethnomedicine Xuetong - The stems of *Kadsura heteroclita*. Phytochemistry 169:112178. *https://doi.org/10.1016/j.phytochem.2019.112178*

Cao L, Shehla N, Tasneem S, Cao M, Sheng W, Jian Y, ... Liao D (2019). New cadinane sesquiterpenes from the stems of *Kadsura heteroclita*. Molecules 24(9):1664. *https://doi.org/10.3390/molecules24091664*

Cechin J, Piasecki C, Benemann DP, Kremer FS, Galli V, Maia LC, Agostinetto D (2020). Transcriptome analysis identifies candidate target genes involved in glyphosate-resistance mechanism in *Lolium multiflorum*. Plants 9(6):685. *https://doi.org/10.3390/plants9060685*

Chen C, Liu S, Liu H, Liu H (2020). Candidate genes involved in the biosynthesis of lignan in *Schisandra chinensis* fruit based on transcriptome and metabolomes analysis. Chinese Journal of Natural Medicines 1:1-12. *https://doi.org/10.3724/SP.J.1009.2019.000000*

Chiang N, Wen C, Chu F (2019). TcMYB1, TcMYB4, and TcMYB8 participate in the regulation of lignan biosynthesis in *Taiwania cryptomerioides* Hayata. Tree Genetics & Genomes 15(5): 67. *https://doi.org/10.1007/s11295-019-1375-0*

Chiang NT, Ma LT, Lee YR, Tsao NW, Yang CK, Wang SY, Chu FH (2019). The gene expression and enzymatic activity of pinoresinol-lariciresinol reductase during wood formation in *Taiwania cryptomerioides* Hayata. Holzforschung 73(2):197-208. *https://doi.org/10.1515/hf-2018-0026*

Deng Y, Lu S (2017). Biosynthesis and regulation of phenylpropanoids in plants. Critical Reviews in Plant Sciences 36(4):257-290. *https://doi.org/10.1080/07352689.2017.1402852*

Editorial Committee of Flora of China (1996). 中国植物志 [Flora of China]. 科学出版社, 北京, 中国 30(1):238. *http://www.iplant.cn/info/Kadsura%20heteroclita?t=z*

Effenberger I, Zhang B, Li L, Wang Q, Liu Y, Klaiber I, ... Schaller A (2015). Dirigent proteins from cotton (*Gossypium* sp.) for the atropselective synthesis of gossypol. Angewandte Chemie International Edition 54(49):14660-14663. *https://doi.org/10.1002/anie.201507543*

Fan W, Hu Q, Duan T, He B, Ye Z, Meng Y (2019). 广西海风藤HPLC指纹图谱研究[HPLC fingerprint analysis of *Kadsura heteroclita* (Roxb.) Craib]. 湖北中医药大学学 21(4):46-50. *http://www.cnki.com.cn/Article/CJFDTotal-HZXX201904011.htm*

Fidan O, Zhan J (2018). Reconstitution of medicinally important plant natural products in microorganisms. In: Kermode AR (Eds). Molecular Pharming: Applications, Challenges, and Emerging Areas. John Wiley & Sons, Inc, NJ, USA, pp 383-415. *https://doi.org/10.1074/jbc.274.2.618*

Fujita M, Gang DR, Davin LB, Lewis NG (1999). Recombinant pinoresinol-lariciresinol reductases from western red cedar (*Thuja plicata*) catalyze opposite enantiospecific conversions. Journal of Biological Chemistry 274(2):618-627. *https://doi.org/10.1074/jbc.274.2.618*

Gao R, Yu D, Chen L, Wang W, Sun L, Chang Y (2019). Cloning and functional analysis of squalene synthase gene from *Dryopteris fragrans* (L.) Schott. Protein Expression and Purification 155:95-103. *https://doi.org/10.1016/j.pep.2018.07.011*

Ge Y, Cheng Z, Si X, Ma W, Tan L, Zang X, ... Zhou Z (2019). Transcriptome profiling provides insight into the genes in carotenoid biosynthesis during the mesocarp and seed developmental stages of avocado (*Persea americana*). International Journal of Molecular Sciences 20(17):4117. *https://doi.org/10.3390/ijms20174117*

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, ... Zeng Q (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29(7):644-652. *https://doi.org/10.1038/nbt.1883*

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59(3):307-321. *https://doi.org/10.1093/sysbio/syq010*

Guo H (2017). 五味子科药用植物叶绿体基因组学研究与内南五味子的分子鉴定[Study on chloroplast genome of Schisandraceae, and molecular identification of *Kadsura interior*]. 硕士论文, 北京协和医学院. *http://cdmd.cnki.com.cn/Article/CDMD-10023-1017227498.htm*

Hemmati S, von Heimendahl CB, Klaes M, Alfermann AW, Schmidt TJ, Fuss E (2010). Pinoresinol-lariciresinol reductases with opposite enantiospecificity determine the enantiomeric composition of lignans in the different organs of *Linum usitatissimum* L. Planta Medica 76(9):928-934. *https://doi.org/10.1055/s-0030-1250036*

Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, ... Tester M (2017). The genome of *Chenopodium quinoa*. Nature 542(7641):307-312. *https://doi.org/10.1038/nature21370*

Kanehisa M (2020). KEGG PATHWAY Database. Retrieved 2020 April 20 from *https://www.kegg.jp/kegg/pathway.html*.

Kang J, Zhang Q, Jiang X, Zhang T, Long R, Yang Q, Wang Z (2019). Molecular cloning and functional identification of a squalene synthase encoding gene from slfalfa (*Medicago sativa* L.). International Journal of Molecular Sciences 20(18):4499. *https://doi.org/10.3390/ijms20184499*

Kim KW, Moinuddin SG, Atwell KM, Costa MA, Davin LB, Lewis NG (2012). Opposite stereoselectivities of dirigent proteins in *Arabidopsis* and *Schizandra* species. Journal of Biological Chemistry 287(41):33957-33972. *https://doi.org/10.1074/jbc.M112.387423*

Kim MK, Jeon JH, Fujita M, Davin LB, Lewis NG (2002). The western red cedar (*Thuja plicata*) 8-8'DIRIGENT family displays diverse expression patterns and conserved monolignol coupling specificity. Plant Molecular Biology 49(2):199-214. *https://doi.org/10.1023/A:1014940930703*

Letunic I, Bork P (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research 47(W1):W256-W259. *https://doi.org/10.1093/nar/gkz239*

Li B, Dewey C (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12(1):323. *https://doi.org/10.1186/1471-2105-12-323*

Li N, Zhao M, Liu T, Dong L, Cheng Q, Wu J, ... Lu W (2017). A novel soybean dirigent gene *GmDIR22* contributes to promotion of lignan biosynthesis and enhances resistance to *Phytophthora sojae*. Frontiers in Plant Science 8:1185. *https://doi.org/10.3389/fpls.2017.01185*

Liu R, Liu Q, Li B, Liu L, Cheng D, Cai X, ... Wang W (2020). Pharmacokinetics, bioavailability, excretion, and metabolic analysis of Schisanlactone E, a bioactive ingredient from *Kadsura heteroclita* (Roxb) Craib, in rats by UHPLC-MS/MS and UHPLC-Q-Orbitrap HRMS. Journal of Pharmaceutical and Biomedical Analysis 177:112875. *https://doi.org/10.1016/j.jpba.2019.112875*

Mao X, Cai T, Olyarchuk JG, Wei L (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21(19):3787-3793. *https://doi.org/10.1093/bioinformatics/bti430*

Markulin L, Corbin C, Renouard S, Drouet S, Gutierrez L, Mateljak I, ... Lainé E (2019). Pinoresinol-lariciresinol reductases, key to the lignan synthesis in plants. Planta 249(6):1695-1714. *https://doi.org/10.1007/s00425-019-03137-y*

Matías-Hernández L, Jiang W, Yang K, Tang K, Brodelius PE, Pelaz S (2017). AaMYB1 and its orthologue AtMYB61 affect terpene metabolism and trichome development in *Artemisia annua* and *Arabidopsis thaliana*. Plant Journal 90(3):520-534. *https://doi.org/10.1111/tpj.13509*

Mertens J, Pollier J, Vanden Bossche R, Lopez-Vidriero I, Franco-Zorrilla JM, Goossens A (2016). The bHLH transcription factors TSAR1 and TSAR2 regulate triterpene saponin biosynthesis in *Medicago truncatula*. Plant Physiology 170(1):194-210. *https://doi.org/10.1104/pp.15.01645*

Mertens J, Van Moerkercke A, Vanden Bossche R, Pollier J, Vanden Bossche R, Goossens A (2016). Clade IVa basic helix-hoop-helix transcription factors form part of a conserved jasmonate signaling circuit for the regulation of bioactive plant terpenoid biosynthesis. Plant and Cell Physiology 57(12):2564-2575. *https://doi.org/10.1093/pcp/pcw168*

Nag A, Choudhary S, Masand M, Parmar R, Bhandawat A, Seth R, ... Sharma RK (2020). Spatial transcriptional dynamics of geographically separated genotypes revealed key regulators of podophyllotoxin biosynthesis in *Podophyllum hexandrum*. Industrial Crops and Products 147:112247. *https://doi.org/10.1016/j.indcrop.2020.112247*

Nagegowda DA, Gupta P (2020). Advances in the biosynthesis, regulation, and metabolic engineering of plant specialized terpenoids. Plant Science 294:110457. *https://doi.org/10.1016/j.plantsci.2020.110457*

Nakatsubo T, Mizutani M, Suzuki S, Hattori T, Umezawa T (2008). Characterization of *Arabidopsis thaliana* pinoresinol reductase, a new type of enzyme involved in lignan biosynthesis. Journal of Biological Chemistry 283(23):15550-15557. *https://doi.org/10.1074/jbc.M801131200*

Pankratov I, McQuinn R, Schwartz J, Bar E, Fei Z, Lewinsohn E, ... Hirschberg J (2016). Fruit carotenoid-deficient mutants in tomato reveal a function of the plastidial isopentenyl diphosphate isomerase (IDI1) in carotenoid biosynthesis. Plant Journal 88(1):82-94. *https://doi.org/10.1111/tpj.13232*

Paulino PR, Diego Mauricio ROP, Corrêa LGG, Rensing SA, Birgit K, Bernd MR (2010). PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Research 38:D822-D827. *https://doi.org/10.1093/nar/gkp805*

State Council Information Office of the PRC (2020). 国新办举行中医药防治新冠肺炎重要作用及有效药物发布会图文实录 [Record of a press conference held by National Information Office on the important role of Chinese medicine in the prevention and treatment of COVID-19 and the effective drugs release]. Cited 2020 May 5 *http://www.scio.gov.cn/xwfbh/xwbfbh/wqfbh/42311/42768/wz42770/Document/1675777/1675777.htm*.

Ralph SG, Jancsik S, Bohlmann J (2007). Dirigent proteins in conifer defense II: Extended gene discovery, phylogeny, and constitutive and stress-induced gene expression in spruce (*Picea* spp.). Phytochemistry 68(14): 1975-1991. *https://doi.org/10.1016/j.phytochem.2007.04.042*

Rasbery JM, Shan H, LeClair RJ, Norman M, Matsuda SP, Bartel B (2007). *Arabidopsis thaliana* squalene epoxidase 1 is essential for root and seed development. Journal of Biological Chemistry 282(23):17002-17013. *https://doi.org/10.1074/jbc.M611831200*

Rather GA, Sharma A, Jeelani SM, Misra P, Kaul V, Lattoo SK (2019). Metabolic and transcriptional analyses in response to potent inhibitors establish MEP pathway as major route for camptothecin biosynthesis in *Nothapodytes nimmoniana* (Graham) Mabb. BMC Plant Biology 19(1):301. *https://doi.org/10.1186/s12870-019-1912-x*

Ribeiro B, Lacchini E, Bicalho K, Mertens J, Arendt P, Vanden Bossche R, ... Buitink J (2020). A seed-specific regulator of triterpene saponin biosynthesis in *Medicago truncatula*. Plant Cell 32(6):2020-2042. https://doi.org/10.1105/tpc.19.00609

Rong Q, Jiang D, Chen Y, Shen Y, Yuan Q, Lin H, ... Huang L (2016). Molecular cloning and functional analysis of squalene synthase 2 (SQS2) in *Salvia miltiorrhiza* Bunge. Frontiers in Plant Science 7:1274. https://doi.org/10.3389/fpls.2016.01274

Sadat-Hosseini M, Bakhtiarizadeh MR, Boroomand N, Tohidfar M, Vahdati K (2020). Combining independent *de novo* assemblies to optimize leaf transcriptome of Persian walnut. PloS One 15(4):e0232005. https://doi.org/10.1371/journal.pone.0232005

Schweizer F, Colinas M, Pollier J, Van Moerkercke A, Vanden Bossche R, de Clercq R, Goossens A (2018). An engineered combinatorial module of transcription factors boosts production of monoterpenoid indole alkaloids in *Catharanthus roseus*. Metabolic Engineering 48:150-162. https://doi.org/10.1016/j.ymben.2018.05.016

Shehla N, Li B, Cao L, Zhao J, Jian Y, Daniyal M, ... Rahman A (2020). Xuetonglactones A-F: Highly oxidized lanostane and cycloartane triterpenoids from *Kadsura heteroclita* Roxb. Craib. Frontiers in Chemistry 7:935. https://doi.org/10.3389/fchem.2019.00935

Song M, Peng X (2019). Genome-wide identification and characterization of DIR genes in *Medicago truncatula*. Biochemical Genetics 57(4):487-506. https://doi.org/10.1007/s10528-019-09903-7

Song S, Qi T, Fan M, Zhang X, Gao H, Huang H, ... Xie D (2013). The bHLH subgroup IIId factors negatively regulate jasmonate-mediated plant defense and development. PLoS Genetics 9(7):e1003653. https://doi.org/10.1371/journal.pgen.1003653

Van Moerkercke A, Steensma P, Gariboldi I, Espoz J, Purnama PC, Schweizer F, ... Goossens A (2016). The basic helix-loop-helix transcription factor BIS2 is essential for monoterpenoid indole alkaloid production in the medicinal plant *Catharanthus roseus*. Plant Journal 88(1):3-12. https://doi.org/10.1111/tpj.13230

Van Moerkercke A, Steensma P, Schweizer F, Pollier J, Gariboldi I, Payne R, ... Goossens A (2015). The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in *Catharanthus roseus*. Proceedings of the National Academy of Sciences USA 112(26):8130-8135. https://doi.org/10.1073/pnas.1504951112

von Heimendahl CB, Schäfer KM, Eklund P, Sjöholm R, Schmidt TJ, Fuss E (2005). Pinoresinol–lariciresinol reductases with different stereospecificity from Linum album and *Linum usitatissimum*. Phytochemistry 66(11):1254-1263. https://doi.org/10.1016/j.phytochem.2005.04.026

Wang C, Zhu J, Liu M, Yang Q, Wu J, Li Z (2018). *De novo* sequencing and transcriptome assembly of *Arisaema heterophyllum* Blume and identification of genes involved in isoflavonoid biosynthesis. Scientific Reports 8(1):1-12. https://doi.org/10.1038/s41598-018-35664-1

Wang Z, Guo H, Zhang Y, Lin L, Cui M, Long Y, Xing Z (2019). DNA methylation of farnesyl pyrophosphate synthase, squalene synthase, and squalene epoxidase gene promoters and effect on the saponin content of *Eleutherococcus Senticosus*. Forests 10(12):1053. https://doi.org/10.3390/f10121053

World Health Organization (2020). Coronavirus disease (COVID-2019) situation reports. Retrieved 2020 August 11, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.

Xia Z, Costa MA, Pélissier HC, Davin LB, Lewis NG (2001). Secoisolariciresinol dehydrogenase purification, cloning, and functional expression implications for human health protection. Journal of Biological Chemistry 276(16):12614-12623. https://doi.org/10.1074/jbc.M008622200

Xu J, van Herwijnen ZO, Dräger DB, Sui C, Haring MA, Schuurink RC (2018). SlMYC1 regulates type VI glandular trichome formation and terpene biosynthesis in tomato glandular cells. Plant Cell 30(12):2988-3005. https://doi.org/10.1105/tpc.18.00571

Xu X, Guignard C, Renaut J, Hausman J, Gatti E, Predieri S, Guerriero G (2019). Insights into lignan composition and biosynthesis in stinging nettle (*Urtica dioica* L.). Molecules 24(21):3863. https://doi.org/10.3390/molecules24213863

Xu YH, Liao YC, Lv FF, Zhang Z, Sun PW, Gao ZH, ... Wei JH (2017). Transcription factor AsMYC2 controls the jasmonate-responsive expression of ASS1 regulating sesquiterpene biosynthesis in *Aquilaria sinensis* (Lour.) Gilg. Plant and Cell Physiology 58(11):1924-1933. https://doi.org/10.1093/pcp/pcx122

Xue L, He Z, Bi X, Xu W, Wei T, Wu S, Hu S (2019). Transcriptomic profiling reveals MEP pathway contributing to ginsenoside biosynthesis in *Panax ginseng*. BMC Genomics 20(1):134. https://doi.org/10.1186/s12864-019-5718-x

Youn B, Moinuddin SG, Davin LB, Lewis NG, Kang C (2005). Crystal structures of apo-form and binary/ternary complexes of *Podophyllum* secoisolariciresinol dehydrogenase, an enzyme involved in formation of health-protecting and plant defense lignans. Journal of Biological Chemistry 280(13):12917-12926. *https://doi.org/10.1074/jbc.M413266200*

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010). Method gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology 11:R14. *https://doi.org/10.1186/gb-2010-11-2-r14*

Zhang B, Liu Y, Chen M, Feng J, Ma Z, Zhang X, Zhu C (2018). Cloning, expression analysis and functional characterization of squalene synthase (sqs) from *tripterygium wilfordii*. Molecules 23(2):269. *https://doi.org/10.3390/molecules23020269*

Zhang DH, Jiang LX, Li N, Yu X, Zhao P, Li T, Xu JW (2017). Overexpression of the squalene epoxidase gene alone and in combination with the 3-hydroxy-3-methylglutaryl coenzyme A gene increases ganoderic acid production in *Ganoderma lingzhi*. Journal of Agricultural Food and Chemistry 65(23):4683-4690. *https://doi.org/10.1021/acs.jafc.7b00629*

Zhang M, Zhou J, Wei W, Hao E, Deng J, Hou X (2019). 瑶药大红钻化学成分及药理作用研究进展 [Research progress on chemical constituents and pharmacological effects of Yao medicine *Kadsura heteroclita*]. 中草药 50(14): 3493-3502. *http://www.cnki.com.cn/Article/CJFDTotal-ZCYO201914033.htm*

Zhang X, Allan A, Li C, Wang Y, Yao Q (2015). *De novo* assembly and characterization of the transcriptome of the Chinese medicinal herb, *Gentiana rigescens*. International Journal of Molecular Sciences 16(5):11550-11573. *https://doi.org/10.3390/ijms160511550*

Zhao M, Zhong Q, Tian M, Han R, Ren Y (2020). Comparative transcriptome analysis reveals differentially expressed genes associated with the development of Jerusalem artichoke tuber (*Helianthus tuberosus* L.). Industrial Crops and Products 151:112455. *https://doi.org/10.1016/j.indcrop.2020.112455*

Zhou H, Ma S, Chen B, Han Z, Yao H (2016). 鸡血藤、**滇**鸡血藤、大血藤等血藤类药材的psbA-trnH条形码分子鉴定 [Identification of *Spatholobi Caulis*, *Kadsurae Caulis* and *Sargentodoxae Caulis* using the psbA-trnH barcode]. 世界科学技术-中医药现代化 (1):40-45. *http://www.cnki.com.cn/Article/CJFDTOTAL-SJKX201601010.htm*

Zhou J, Zhang Y, Hu T, Su P, Zhang Y, Liu Y, ... Gao W (2018). Functional characterization of squalene epoxidase genes in the medicinal plant *Tripterygium wilfordii*. International Journal of Biological Macromolecules 120:203-212. *https://doi.org/10.1016/j.ijbiomac.2018.08.073*

Zhou P, Pu T, Gui C, Zhang X, Gong L (2020). Transcriptome analysis reveals biosynthesis of important bioactive constituents and mechanism of stem formation of *Dendrobium huoshanense*. Scientific Reports 10:2857. *https://doi.org/10.1038/s41598-020-59737-2*