Available online at www.scholarsresearchlibrary.com

Scholars Research Library

J. Nat. Prod. Plant Resour., 2022, 12 (3): 1-14

(http://scholarsresearchlibrary.com/archive.html)

Transcriptome Analysis and Identification of Genes Involved in Terpenoid Biosynthesis of *Eurycoma longifolia Jackroot*

Norlia Basherudin[1]*, Nor Hasnida Hassan[1], Nur Nabilah Alias[1], Norwati Muhammad[1], Mohd Faizal Abu Bakar[2], Mohd Noor Mat Isa[2]

[1]Forestry Biotechnology Division, Forest Research Institute Malaysia (FRIM), Kepong, Selangor , 52109

[2]Malaysia Genome Institute, Jalan Bangi, Kajang, Selangor, 43000

ABSTRACT

Eurycoma longifolia has been widely used for various traditional medicinal purposes in Malaysia. Among the major compounds produced in E. longifolia is quassinoids. However little is known about the biosynthesis pathway leading to these compounds. As a starting point, genes involved in terpenoids biosynthesis pathway, which is the primary pathway leading to quassinoid were analysed using Illumina high-throughput RNA-sequencing technology. The transcriptome profiles were generated from roots of the mature 10-year-old and the young 1-year-old E. longifolia. BLAST against the Nr database of 60,753 non-redundant unigenes obtained indicated that only 34,673 (57%) showed homology to known proteins and was highly similar to citrus. Most assignments of gene ontology based on the biological process category were to "metabolic process". KEGG analysis showed that key enzymes in major secondary metabolite pathways were present in the transcriptome and the highest were assigned to "phenylpropanoid biosynthesis" and "terpenoid backbone biosynthesis". Differentially expressed gene analysis predicted 154 unigenes were potentially related to quassinoid biosynthesis. They were up-regulated in 10-year-old roots and were either involved in terpenoids backbone biosynthesis, encoded for transcription factors (WRKY and AP2/ERF genes) or encoded for cytochrome P450.

Keywords: *Ali, Quassinoids, Medicinal plant, High-Throughput sequencing*

INTRODUCTION

*Eurycoma longifolia*Jack or locally known as Tongkat Ali which belongs to the family of *Simaroubaceae* is a popular medicinal plant in Malaysia. It is indigenous to South-East Asian countries such as Malaysia, Indonesia and Vietnam. Extracts of this plant, particularly from the roots, have been used in traditional medicines for their antimalarial [1-3], antiulcer [4], cytotoxic [5] and aphrodisiac effects [6]. The extract is widely used in pharmaceutical industries and various chemical compounds including canthin-6-one alkaloids, quassinoids, eurycomaoside, squalene derivatives, biphenylneoligans and eurycomalactone have been isolated from the root [7]. Among these isolates, quassinoids account for a major portion of the root phytochemicals [8,9].

Quassinoids are compounds found exclusively in plants under the *Simarouboidaea* subfamily of *Simaroubaceae*. Quassinoids are considered principal active constituents, responsible for bitter taste and have potential pharmacological and therapeutic properties [10,11]. Phytochemical studies on the roots of *E. longifolia* have isolated and identified nearly seventy quassinoids related to triterpenes. Study on *E. longifolia* extracts rich in quassinoids exhibit biological activities of anti-malaria, anti-tumor and aphrodisiac [12-14]. Quassinoids are categorized according to their basic skeleton into five distinct groups, which are C-18, C-19, C-20, C-22 and C-25 [15]. However little is known about the biosynthesis pathway leading to quassinoids in *E. longifolia*, which hinders progress in determining the underlying mechanism. Thus for the start it is important to identify genes involved in terpenoids pathways, which is the primary pathway leading to quassinoid biosynthesis.

Terpenoids or isoprenoids are the largest family of secondary metabolites. All terpenoids are formed through the condensation of the central metabolic intermediates of terpenoid metabolism, isopentenyl and dimethylallyl diphosphate (IDP and DMADP). Two distinct biochemical pathways involve in the synthesis are 2C-methyl-D-erythritol-4-phosphate (MEP) and the Mevalonic Acid (MVA). Both MVA and MEP pathways contribute terpene skeletons leading to functionalized and biologically active of many

secondary metabolites. Generally, the MEP pathway supplies C_5 prenyl diphosphate for the synthesis of C_{10} monoterpenes, C_{20} diterpenes and C_{40} tetraterpenes while the MVA pathways provides precursors for the synthesis of C_{15} sesquiterpenes, C_{27-29} sterols, C_{30} triterpenes and their saponin derivatives [16]. Quassinoids are believed to be formed by oxidative degradation of triterpene derivatives [17] from the proto-triterpene, apo-euphol or apo-tirucallol [18].

Transcriptomes analysis using high-throughput sequencing technologies is an effective approach to facilitate gene discovery. This technology is increasingly being used to analyse numbers of plant species especially those related to biosynthesis of secondary metabolites [19,20]. Availability of software such as Blast2GO [21], which associates individual sequences and related expression information with biological function, facilitates gene discovery research. The technologies would benefit non model species such as *E. longifolia*, which has very limited transcriptome and genomic data in public databases. Here we report the transcriptome analysis of mature 10-year-old and young 1-year-old roots of *E. longifolia* using Illumina high-throughput RNA-sequencing technology to discover putative genes involved in terpenoid biosynthesis, a major pathway leading to quassinoids formation.

MATERIALS AND METHODS

Plant materials

Roots of a ten-year-old *E. longifolia* tree was collected from a trial planting plot at Bukit Hari, Forest Research Institute Malaysia (FRIM), Selangor while roots of a one-year-old *E. longifolia* tree were obtained from the Kepong Botanical Garden nursery, FRIM. Harvested roots were immediately put in liquid nitrogen until further processing. For longer storage, the roots were kept at -80°C.

RNA isolation and integrity analysis

RNA isolation from the roots of *E. longifolia* was carried out using the RNeasy Plant Mini Kit (Qiagen) according to the protocol outlined by the supplier. Samples were ground into powder in a pre-chilled DEPC-treated mortar in the presence of liquid nitrogen. A QIAshredder spin column was used to filter the cell-debris and RNeasy spin column was used to trap all the RNA. The RNA was then eluted with 50 µl of RNase-free water and kept at 4°C or -80°C for longer storage. High quality RNA for next generation DNA sequencing is very important to ensure a reliable result. Agilent RNA 6000 Nano Chips run on a Agilent 2100 bioanalyzer (Agilent Technologies) were used to assess the RNA quality and graded it as RNA Integrity Number (RIN), a numbering system of 1 to 10, with 1 being the most degraded profile and 10 being the most intact. Sample preparation and apparatus set-up to determine RNA quality was carried out according to the manual provided by the supplier.

Illumina sequencing

The total RNA with absorbance 260/280 nm ratio of ~2.0 and RIN number more than 8.0 was chosen for Illumina sequencing. Each of the RNA samples was used to generate a paired end cDNA library of 100 bp sequencing reads. Generation of cDNA libraries using mRNA-Seq assay and sequencing on Illumina HiSeq 2000 were outsourced to Beijing Genome Institute (BGI), China.

Transcriptome sequencing data analysis

The raw sequence data was quality checked and trimmed. All the quality reads were assembled *de novo* using the SOAP *denovo* program [22], which involved three processes. In the first process the clean reads were overlapped with each other to form contigs. The process was carried out until the contigs were no longer extended. Second, the contigs were joined together to form scaffolds based on paired-end information and finally, the paired-end reads were reused to fill the scaffold gaps to obtain unigenes with the fewest Ns that could not be extended on either end. The sequences were defined as unigenes. The unigenes from the two libraries were combined to generate non-redundant unigenes. Unigene sequences were aligned by Blast X (an E-value<$1.0e^{-5}$ was used as the cut-off) to public protein databases NCBI's nonredundant protein (Nr) database, Swiss-Prot protein database and Kyoto Encyclopedia of Genes and Genomes (KEGG). Blast2GO software, an automated tool for the assignment of Gene Ontology (GO) terms was used to functionally categorize the BLAST matches and construct pie charts using standard graph configurations. The software was also used to map unique sequences with Enzyme Commission (EC) number to KEGG biochemical pathways according to the EC distribution in the pathways database. Unigenes encoded for Transcription Factor (TF) and Cytochrome P450 were also selected for analysis. ORF finder and MOTIF search software were used to search for the open reading frame and protein sequence motif, respectively.

Quality of coverage and assembled sequence validation

Coverage of the assembled unigenes were evaluated by realigning the sequencing reads of the 1-and 10-year-old roots to the reference transcriptome. The quality of assembled sequences was validated using cDNA of eight randomly selected unigenes. Specific primer pairs for each unigene were designed and synthesised. PCR was carried out using cDNA synthesized from root RNA. The amplified fragments were eluted from the gel and cloned in PCR 2.1 vector (Thermofisher Scientific). Plasmids were isolated, sequenced via Sanger sequencing technology and subjected to BLAST analysis.

Identification of differentially expressed genes

Reads Per Kilobase Per Million Mapped Reads (RPKM) was used to normalize the mapped reads [23]. The ρ values and $\log_2$ of each gene were calculated and the differentially expressed genes (DEGs) were identified at thresholds of $|\log_2|>2$ and FDR ≤

0.001. The RPKM method is able to eliminate the influence of different gene length and sequencing level on the calculation of gene expression. Therefore the calculated gene expression can be directly used to compare differences in gene expression between samples.

RESULTS AND DISCUSSION

Sequencing output and de novo transcriptome assembly

Illumina high-throughput-sequencing of the *E. longifolia* transcriptome data went through stringent quality assessment and data filtering, only reads with Q20 bases (those with a base quality greater than 20) were defined as high quality reads. Clean paired-end reads, contigs, scaffolds and unigenes generated were listed in (Table 1).

Table 1. A summary of the transcriptome sequencing and assembly results of *E. longifolia* root

Sample		Total Number	Total Length (nt)	Mean Length (nt)	N50 (bp)
Clean reads	1 year-old root	4,13,57,870	3,72,22,08,300		
Contigs	10 year-old root	3,88,57,074	3,49,71,36,660		
	1 year-old root	6,16,318	8,09,97,931	131	90
	10 year-old root	1,11,340	3,06,02,922	275	429
Scaffolds	1 year-old root	1,12,414	3,80,01,762	338	449
	10 year-old root	50,942	2,60,87,967	512	803
Unigenes	1 year-old root	44,857	2,52,14,755	419	510
	10 year-old root	81,907	3,42,92,972	562	843
	All-unigenes	60,753	3,82,00,050	629	921

The quality reads of the two roots were combined for non-redundant sequencing as the reference transcriptome of *E. longifolia*. A total of 60,753 All-unigenes were obtained from the reference transcriptome with an average length of 629 bp and N50 of 921 bp. Of these unigenes, 52.6% were in the range of 300 bp-400 bp, 30.6% were longer than 500 bp and the remaining 16.8% were longer than 1000 bp (Figure 1).

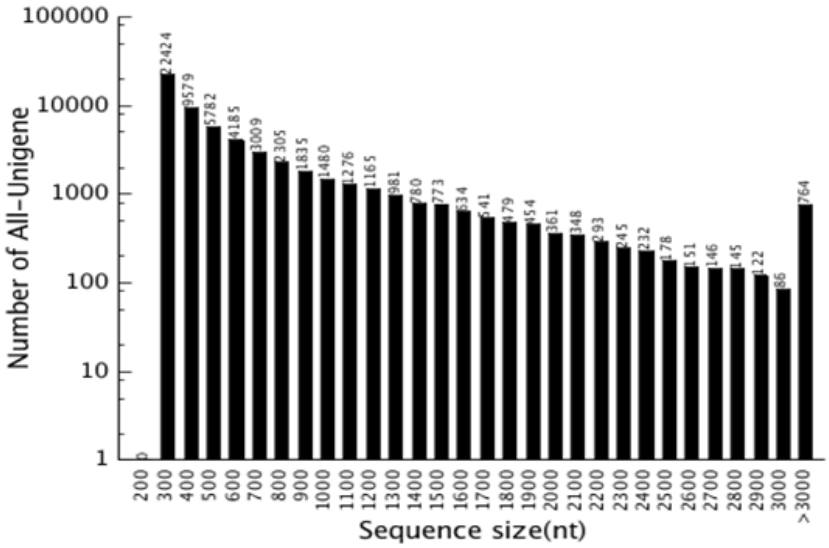

Figure 1. Size distribution of unigenes of the *Eurycoma longifolia* reference transcriptome

Mapping of the short reads back to the 60,753 All-unigenesrevealed that the read numbers mapped with per unigene (from 11-100) comprised the largest distribution, followed by 1-10 and 101-200 (Figure 2). A total of 39,206 (about 74.03%) unigeneswere realigned with more than 10 reads; 20677 (approximately 39.04%) and 6015 (about 11.3%) unigenes were remapped by more than 100 and 1,000 reads, respectively; whereas only 71 (0.13%) unigenes were remapped with more than 8,000 reads. The sequencing depth of the assembled unigenes for 1- and 10-year-old rootsaveraged 42.18 folds and 42.92 folds, respectively. These results suggested that the assembled unigenes were well overlapped by the sequencing reads. Amplification and resequencing of eight randomly selected unigenes also indicated that sequencing and assembly of the unigenes were carried out appropriately, as all unigenes were equal or more than 97% similar at nucleotide level with their corresponding amplicon sequence (Table 2).

Table 2. Similarity percentage of amplicons to their assembled unigenes and sequences of primer used to amplify the amplicons

Unigene	Primer	Amplicon size (bp)	Identity (%)
Unigene10491_All	**forward**: 5'-TTTGTCAGTCCAGAGGTTTCCGT-3'	400	99
	reverse:5'-TTTATTTTCCAGCTGTAGTTAAGGAGA-3'		
Unigene114_All	**forward**: 5'-TCCTGCAGCAGAGTCCCCAC-3'	1264	99
	reverse:5'-CCCGGGGCTTGCCATTTGTT-3'		
Unigene17680_All	**forward**: 5'-TCGATCGACCACAGTGTTGTCTCA-3'	507	99
	reverse:5'-ACACGGATGGGAAGCGGAGG-3'		
Unigene23968_All	**forward**: 5'-GAGGACGCTGTAGCACTACC-3'	907	99
	reverse:5'-CCTACGTCATCCACCACACC-3'		
Unigene3976_All	**forward** : 5'-GCAACTTGGCTGTGCGGGAG-3'	1032	99
	reverse :5-CGCCCATACCAGGTGCCAAA-3'		
Unigene49971_All	**forward** : 5'-TCTCTTTGTTTTCGACTGCTTAGGC-3'	301	97
	reverse :5'-TTGCCTCTAACCTTGGTAGGGCT-3'		
Unigene52285_All	**forward** : 5'-TCTTTCCTGTCGCAGCAATGGC-3'	324	99
	reverse :5'-CAGCCATGGCCTGATCCCGA-3'		
Unigene60390_All	**forward** : 5'-ATTGCGCGATCCCTCCATAA-3'	510	99
	reverse:5'-TCCTTTATACAATGTCTTGGCCAGA-3'		

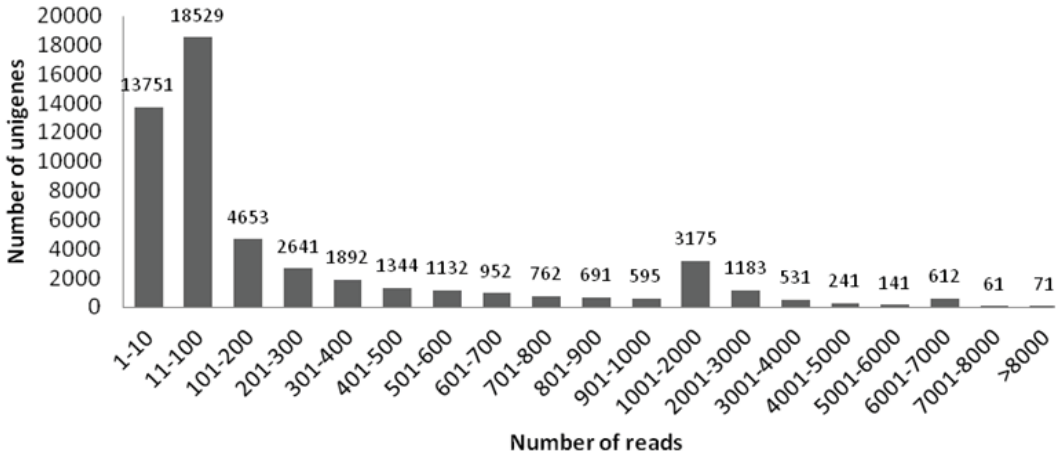

Figure 2. Assessment of assembled unigene quality by the distribution of unique mapped reads within the assembled unigenes for the 10-year-old roots

Similarity search and functional annotation

Among 60,753 All-unigenes, 34,673 (57%) showed homology to known proteins in the Nr database. E-value distribution of the top hits with significant homology (less than 1.0 e^{-49}) was detected for only 34% of the unigenes andonly 12% had similarity greater than 80% (Figure 3). Only 24,167 unigenes (40%) had matches to known protein sequences in the Swiss-Prot database. Within the matched unigenes only 24% had significant homology less than 1.0e^{-49} and 6% had similarity equal or greater than 80%. The mapped sequences of *E. longifolia* had a high similarity with sequences of *Citrus sinensis*, *Citrus clementina*, Citrus unshiu, Vitis vinifera, Theobroma cacao and *Hevea brasilliensis* (Figure 4). The low percentage of matched unigenes, e-value and low similarity to the protein databases might be due to limited publicly available genomic information for the genus *Eurycoma.* Therefore transcriptome data generated from this study may allow for the identification of novel genes of *E. longifolia.*

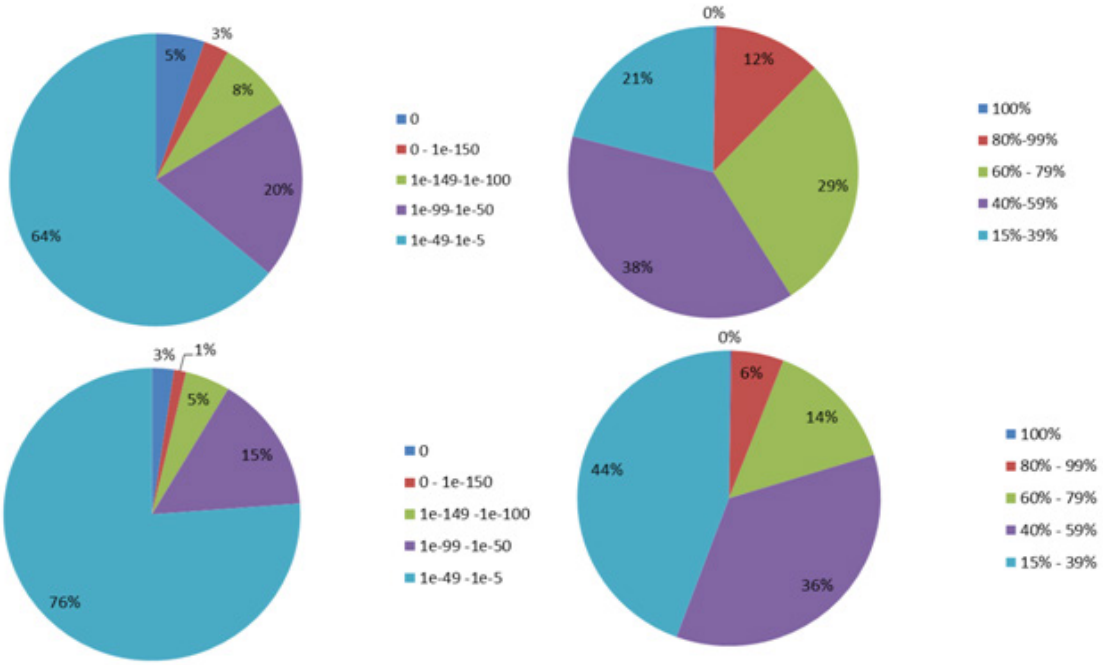

Figure 3. Unigene homology searches against Nr and Swiss Prot databases: proportional frequency of the E-value distribution for (A) Nr and (C) Swiss-Prot, and the proportional frequency of the sequence similarity distribution for (B) Nr and (D) Swiss-Prot

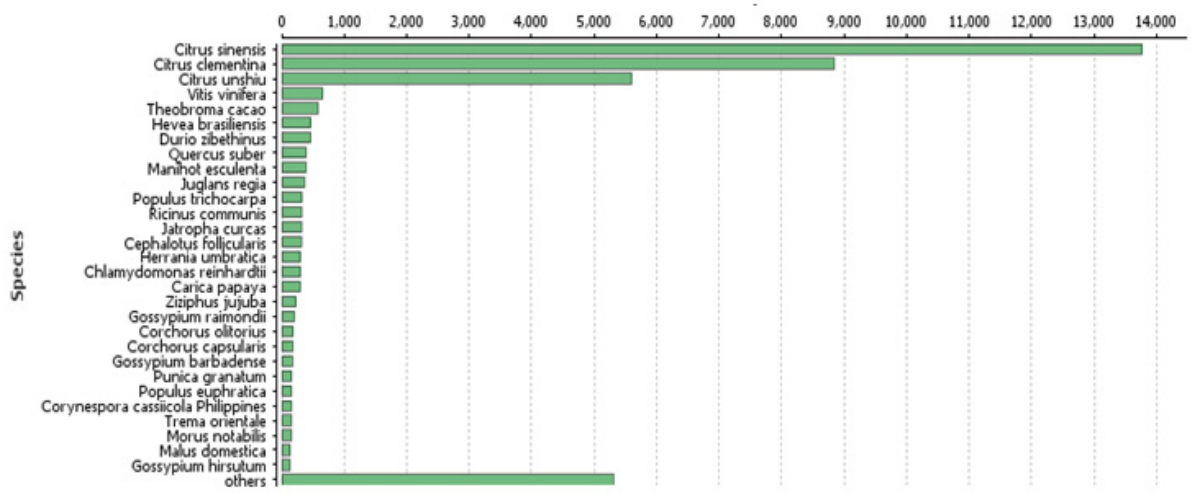

Figure 4. Species distribution of the top BLAST hits against the NR database for the assembled unigene (E-value ≤ 1.0×10^{-5})

To analyse putative protein function, the Cluster of Orthologous Group of proteins (COG) database was used for annotation. Based on transcribed amino acid sequence and function similarity, a total of 11,540 unigenes (19%) were annotated by COG and distributed into 25 categories including proteins with "unknown function" and "general function prediction only" (Figure 5). The cluster for "general function prediction only" (3,381 unigenes) represented the largest group, followed by "transcription" (1,672 unigenes), and "replication, recombination and repair" (1,605 unigenes). Between 1,000-1,500 unigenes were assigned to "carbohydrate transport and metabolism"; "translation, ribosomal structure and biogenesis";"post-translation modification, protein turnover, chaperones";"signal transduction mechanisms" and "function unknown". "Amino acid transport and metabolism"; "cell wall/membrane/envelope biogenesis"; "energy production and conversion"; "cell cycle control, cell division and chromosome partitioning"; inorganic ion transport and metabolism"; "secondary metabolite biosynthesis, transport and catabolism" and "lipid transport and metabolism" possessed 500-900 unigenes. Remaining groups possessed less than 500 unigenes.

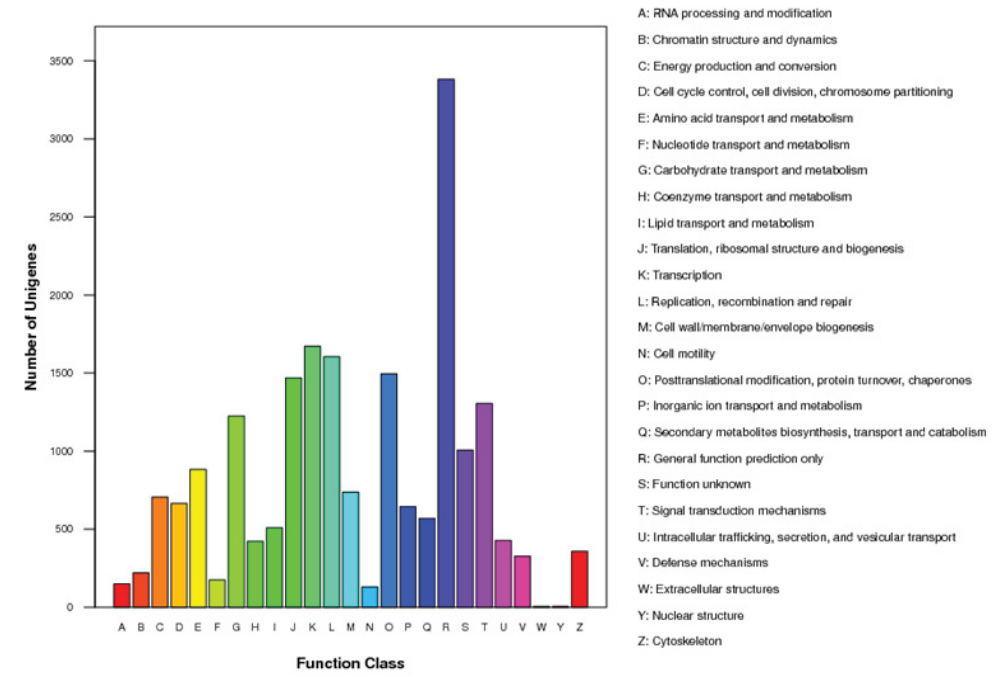

Figure 5. Histogram of the COG (clusters of orthologous groups) functional classification of all the *Eurycoma longifolia unigenes*. Total 11, 540 unigenes were annotated and grouped into 25 molecular families

GO classification revealed that 14,702 unigenes (24%) were categorized into three main categories: biological process, cellular component and molecular function; and 44 functional sub-categories (Figure 6). Biological process represented the majority of the functional terms, followed by cellular component and molecular functions. Within the molecular function category, catalytic activity (6,922, 47%) and binding activity (6,538, 44.5%) were the two most dominant groups. Within the biological process category, metabolic process (6,008, 40.8%) and cellular process (5,548, 37.7%) were the most represented terms. In the cellular component category, cell (9580, 65%) and cell part (8937, 60.7%) were dominant. High numbers of unigenes in groups under the three main categories have also been observed in other medicinal plant root transcriptome studies, e.g., *Polygonum cuspidatum* [24] and *Hedera helix L* [25]. Annotation results of the reference transcriptome data indicated that the root of *E. longifolia* possesses extensive metabolic activity. This was supported by other studies that reported *E. longifolia* contains various chemical compounds such as alkaloids, quassinoids [2], triterpenes [26], squalene derivatives [27] and biphenylneoligans [5].

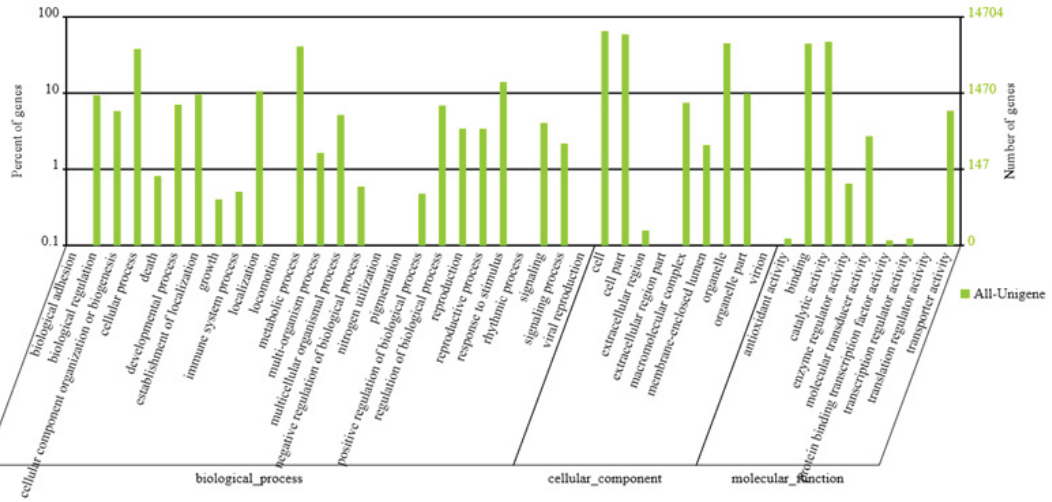

Figure 6. Gene Ontology (GO) classification of the annotated unigenes. The unigenes were classified into three main categories: biological processes, cellular components and molecular functions.

Among the 14,702 unigenes, a total of 6,020 unigenes were assigned to all three categories and 4,230 unigenes were assigned to two categories. Biological process and molecular function categories were shared by 2296 unigenes, while 1188 unigenes shared by molecular function and cellular component and 746 unigenes were shared by cellular component and biological process. Only 366 unigenes were unique to biological process, 2314 unigenes to molecular function and 1772 unigenes to cellular component (Figure 7).

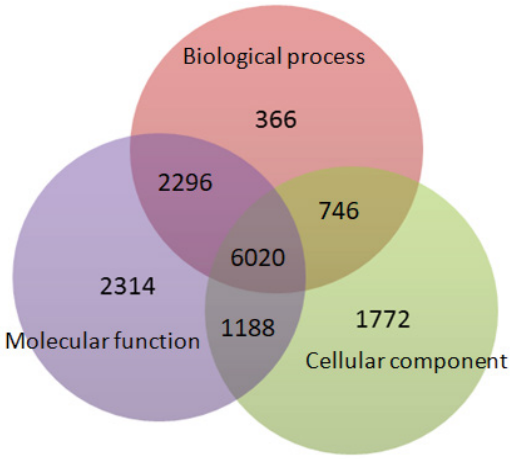

Figure 7. Venn diagram of shared unigenes among the GO classifications of biological process, molecular function and cellular component

KEGG biochemical pathways analysis

To further understand the active biochemical pathways of *E. longifolia*, all assembled unigenes were annotated using the KEGG pathway database. In total, 13,641 unigenes were annotated to 150 pathways based on the fact that one or more unigenes may be mapped to a single Enzyme Commission (EC) number. The top five pathways with the highest number of unigenes were 'purine metabolism' (2,141 unigenes, 15.7%), 'thiamine metabolism' (1446 unigenes, 10.6%), 'biosynthesis of antibiotics' (851 unigenes, 6.2%) and 'starch and sucrose metabolism' (466 unigenes, 3.4%).

Several secondary metabolite pathways have been identified in *E. longifolia* root. Among the pathways were the 'phenylpropanoid biosynthesis' (275 unigenes), 'terpenoid backbone biosynthesis' (85 unigenes), 'isoquinoline alkaloid biosynthesis' (58 unigenes) and 'ubiquinone and other terpenoid-quinone biosynthesis' (41 unigenes). The broad coverage of these genes offers opportunities to examine the biosynthesis process of secondary metabolites in *E. longifolia*. 'Terpenoid backbone biosynthesis' and 'sesquiterpenoid and triterpenoid biosynthesis', which are the primary pathway leading to quassinoid biosynthesis, were further analysed in our study.

Differential expression gene analysis

Differential expression gene was carried out to identify genes that were up-regulated in 10-and 1 year-old-root. It is believed that the active compounds in *E. longifolia* correlated with the root maturity. Generally maturity of the plant takes up to more than 10 years, however, for commercial uses roots are harvested after 4 years of cultivation [7]. Comparison of gene expression of the 1-and 10-year-old root samples revealed 15,994 differentially expressed genes, of which 6,063 and 9,931 were up-regulated in the 1-and 10-year-old roots, respectively (Figure 8). Of the 15,994 transcripts, approximately 10,141 (62.8%) had significant BLASTX hits while another 5,953 (37.2%) bore no similarity with any protein accession.

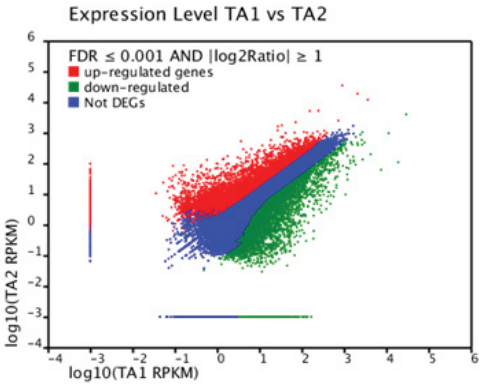

Figure 8. Identification of the differentially expressed genes between 10-(TA1) and 1-year-old (TA2) *E. longifolia* roots. The red, green and blue dots represent transcripts more prevalent in 1-, 10-year-old roots, and both root samples, respectively.

Categorization of the differentially expressed genes based on biological process, cellular component and molecular function is given in Figure 9. Most of the 10-year-old root up-regulated genes in the biological process category were involved in 'gene expression', 'response to stimulus' and 'transport'. For the 1 year-old root, the up-regulated genes were largely involved in 'response to stimulus', 'transport' and 'RNA metabolic process'. The cellular component category showed that the up-regulated genes of both roots mostly took place in the 'cytoplasmic part' and 'integral component of the membrane'. Categorization based on molecular function revealed that genes from both roots were most up-regulated in 'hydrolase activity'. The second highest category for the 10-year-old root was 'nucleic acid binding', while for the 1 year-old root it was 'protein binding'.

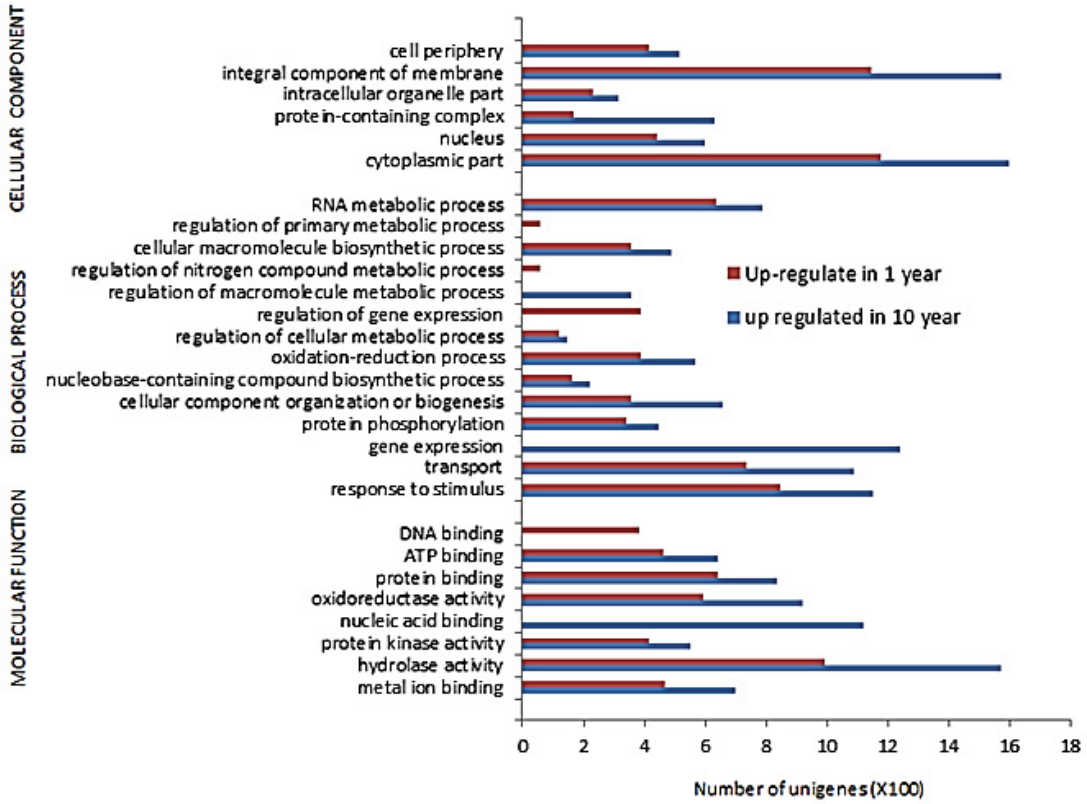

Figure 9. Gene Ontology (GO) classification of up-regulated unigenes in 1-and 10-year-old roots. The unigenes were classified into three main functions: biological process, cellular component and molecular function

Candidate genes involved in terpenoid biosynthesis pathways

Major compounds isolated and characterized from *E. longifolia*, particularly from the roots are quassinoids and alkaloids [9]. However the pathway leading to synthesis of the compound remains unknown. In order to predict the genes involved in the synthesis of quassinoids in *E. longifolia* the annotated enzymes that participated in the terpenoid backbone biosynthesis and sesquiterpenoid and triterpenoid biosynthesis pathways were analysed. Most of genes involved in terpenoid biosynthesis pathways were expressed in the transcriptome of *E. longifolia* root (Figure 10) and a total of 85unigenes potentially involve in the pathway were identified (Table 3). Of the 85 unigenes involved in both pathways, 14 were differentially expressed and of these, 12 and 1 unigenes were up-regulated in the 10-and 1-year-old root, respectively. Among the up-regulated unigenes were unigenes that bear similarity to farnesyl diphosphate synthase and squalene synthase. Overexpression of squalene synthase has been reported to increase triterpene and phytosterol accumulation in *Eleutherococcus senticosus* [28] and *Panax ginseng* [29]. Overexpression of farnesyl diphosphate synthase in *Panax ginseng* hairy root has also been reported to increase ginsenoside content by approximately 2.4-fold [30]. Further study on these two genes may yield more information on pathways by which compounds are synthesized in *E. longifolia.*

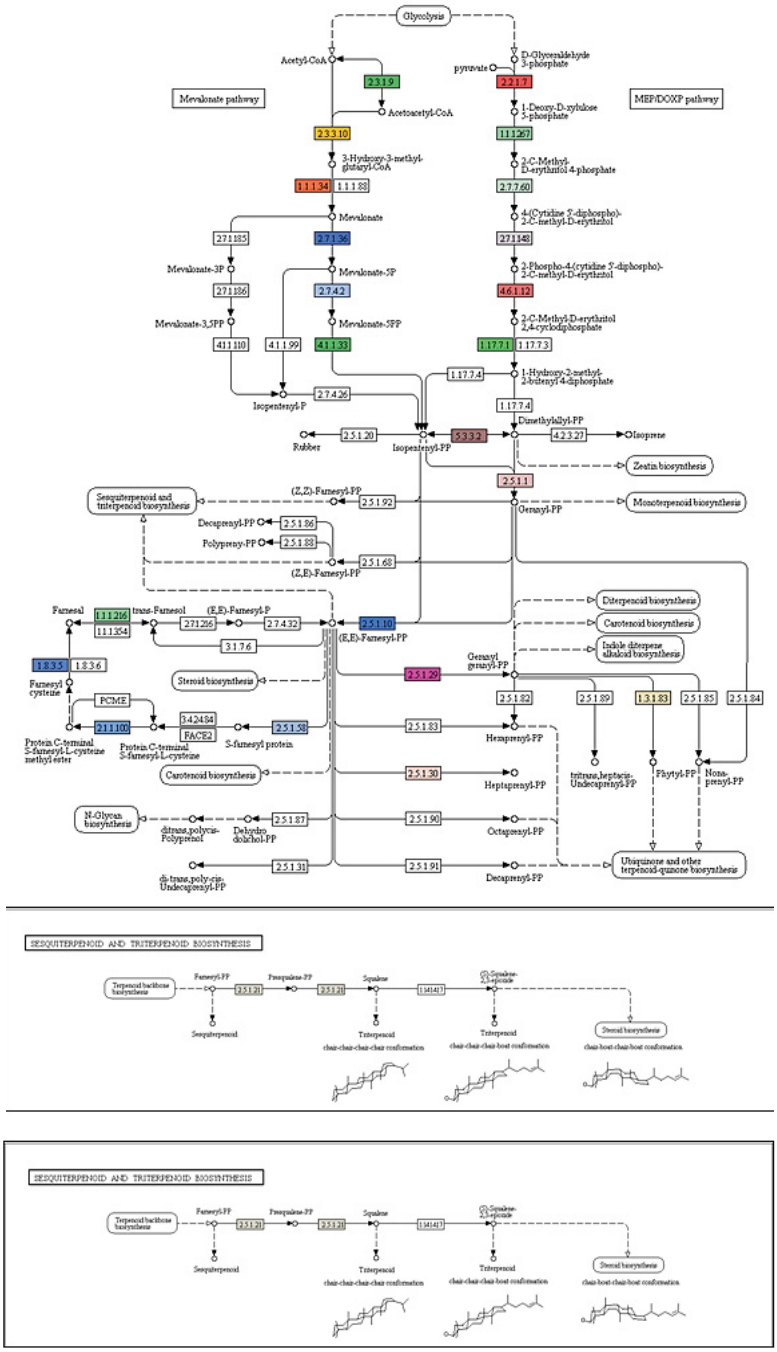

Figure 10. *Eurycoma longifolia* transcriptome encoded enzymes (highlighted) involved in terpenoid backbone biosynthesis and sesquiterpenoid and triterpenoid biosynthesis

Table 3. Unigenes involved in the 'terpenoid backbone biosynthesis' and' sesquiterpenoid and triterpenoid biosynthesis' pathways in *E. longifolia*

Enzymes	EC number	Unigene number	10-year-old up-regulated	1-year-old up-regulated
Terpenoid backbone biosynthesis pathway				
Step 1				
MEP				
1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	10	2	-

1-deoxy-D-xylulose-5-phosphate reductoisomerase	1.1.1.267	5	2	-
2-C-methyl-D-erythritol4-phosphate cytidyl transferase	2.7.7.60	2	-	-
4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	2.7.1.148	1	-	-
2-C-methyl-D-erythritol2,4-cyclodiphosphate synthase	4.6.1.12	1	-	-
4-hydroxy-3-methylbut-2-enyldiphosphate synthase	1.17.7.1	3	1	-
MAV				
Acetyl CoA acetyltransferase	2.3.1.9	2	1	-
Hydroxymethyl glutaryl CoA synthase	2.3.3.10	3	-	-
3-hydroxy-3-methylglutaryl-coenzymeA reductase	1.1.1.34	5	-	-
Mevalonate kinase	2.7.1.36	2	1	-
Phosphomevalonate kinase	2.7.4.2	2	-	-
Mevalonate diphosphosphate decarboxylase	4.1.1.33	1	-	-
Step 2				
Isopentenyl-diphosphate Delta-isomerase	5.3.3.2	2	-	-
Geranyl-diphosphate synthase	2.5.1.1	8	-	-
Farnesyl diphosphate synthase	2.5.1.10	10	2	-
NADP+-dependent farnesol dehydrogenase	1.1.1.216	1	-	-
Prenylcysteine oxidase/Farnesylcysteine lyase	1.8.3.5	1	-	-
Protein farnesyltransferase subunit beta	2.5.1.58	3	-	-
Protein-S-isoprenylcysteine O-methyltransferase	2.1.1.100	1	-	-
Geranylgeranyl diphosphate synthase	2.5.1.29	12	-	-
Heptaprenyl diphosphate synthase	2.5.1.30	1	-	-
Diphosphate reductase	1.3.1.83	4	1	2
Sesquiterpenoid and triterpenoid biosynthesis				
Squalene synthase	2.5.1.21	5	2	-

Based on BlastX search, 967 unigenes of *E. longifolia* were annotated as putative TFs. Among these TFs, 112, 60, 54 52, 52 and 50 unigenes were annotated to the MYB, WRKY, bHLH, AP2/ERF, GRAS and bZIP families, respectively (Table 4). Transcription factors (TFs) are sequence-specific DNA-binding proteins that interact with the regulatory regions of the target genes and modulate the expression initiation [31]. Several families of TFs have been identified as regulators of plant secondary metabolism. Among TFs frequently reported to regulate plant secondary metabolism biosynthesis are the families of WRKY, AP2/ERF, MYB and BHLH [32,33].

Table 4. Putative transcription factors encoding unigenes in *E. longifolia*

TF family	unigenes	10 year-old root	1 year-old root	Non- differential
MYB	112	55	46	11
WRKY	60	41	17	2
BHLH	54	27	25	2
AP2/ERF	52	36	6	10
GRAS	52	24	17	11
bZIP	51	31	16	4
AP2	32	21	9	2
C2H2L	31	16	13	12
ARF	21	5	16	-
Jumonji	18	12	3	3

Among the annotated TFs in *E. longifolia* root, 482 showed a higher expression level in the 10-year-old root, 371 in the 1 year-old root and 114 were equally expressed in both roots. Within the up-regulated TFs of the 10-year-old root, the highest three ranking families were MYB, WRKY and AP2/ERF. Obvious differential expression of WRKY and AP2/ERF gene families were observed between both roots, with 41 WRKY genes up-regulated in the 10-year-old old and only 17 in the 1-year-old root; and 36 AP2/ERF genes enriched in the 10-year-old root but only 6 in the 1-year-old root. These two TFs might be involved in regulation of secondary metabolic pathways of *E. longifolia*. The WRKY family has been reported as a positive regulator of artemisinin biosynthesis in *Artemisia annua* by activating the sesquiterpene synthase gen promoter [34]. In *Withania somnifera*, WRKY regulates the accumulation of triterpenes with anolide by binding to squalene synthase [35]. In *Catharanthus roseus*, the AP2/ERF protein members were reported to regulate terpenoid indole alkaloid metabolism by binding to the promoter of the strictosidine synthase (STR) gene and activating its expression [36].

BlastX searchof *E. longifolia* transcriptome also identified 142 unigenes that bear similarity to cytochrome P450s. The unigenes exhibit the three conserved consensus sequence of cytochrome P450 signature. The first is the heme motif signature (Phe-X-X-Gly-X-Arg-X-Cys-X-Gly), a motif 10 amino acids long that facilitates the binding of heme iron to the cytochrome P450 enzymes. The second motif is Glu-X-X-Arg and the third is Arg/Gly-Gly-X-Asp/Glu-Thr-Thr/Ser (Figure 11). Cytochrome P450 proteins are the largest family of plant proteins and catalyze most of the oxidation steps in the biosynthesis of plant secondary metabolites [37]. Some of the cytochrome P450s have been reported to play critical roles in the modification of triterpenes including hydroxyl, ketone, aldehyde, carboxyl or introduction of epoxy groups [38,39].

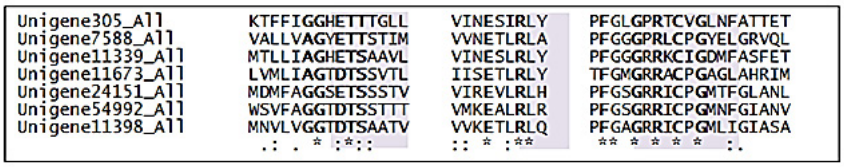

Figure 11. Seven representative unigenes that bear similarity to cytochrome P450 shows the three conserved consensus sequence of cytochrome P450: a motif of Arg/Gly-Gly-X-Asp/Glu-Thr-Thr/Ser, Glu-X-X-Arg and Phe-X-X_Gly_X-Arg_X-Cys_X-Gly

Differential expression analysis using RPKM revealed that of 142 unigenes encoded for the cytochrome P450 enzyme, 65 unigenes were up-regulated (Table 5) and 28 unigenes were down-regulated in 10-year-old roots while another 46 unigenes showed no differrential expression in both roots. Better understanding the function and involvement of cytochrome P450 genes in secondary metabolite production in *E. longifolia* could be useful in metabolite engineering studies in enhancing the production of potential secondary metabolites [40-42].

Table 5. Unigene homologs to cytochrome P450 genes and up-regulated in 10-year-old roots

Gene ID	Gene length	RPKM	RPKM	p value	FDR
		10-year-old	1-year-old		
Unigene305_All	2083	84.8765	19.3012	0	0
Unigene54999_All	1862	9.7235	0	9.30E-109	2.82E-107
Unigene54997_All	1855	21.9669	1.4486	4.25E-175	2.11E-173
Unigene39048_All	1791	12.2469	0.4039	1.38E-109	4.22E-108
Unigene54992_All	1772	11.9781	0.2916	1.50E-110	4.60E-109
Unigene11015_All	1724	40.0607	9.2023	2.50E-158	1.13E-156
Unigene37136_All	1691	13.698	0.8557	1.11E-101	3.15E-100
Unigene11673_All	1596	13.3582	3.9826	5.17E-39	6.08E-38
Unigene54963_All	1562	39.5246	0.397	0	0
Unigene291_All	1545	10.9537	5.1844	6.43E-16	3.82E-15
Unigene11368_All	1515	70.799	14.1215	5.55E-274	4.78E-272
Unigene11231_All	1327	8.1221	4.05	1.25E-09	5.32E-09
Unigene11788_All	1257	153.1362	14.6767	0	0
Unigene5006_All	1193	96.9614	36.4292	2.24E-150	9.51E-149
Unigene37961_All	1117	4.9092	0	1.97E-33	2.04E-32
Unigene6665_All	1105	271.6544	9.1194	0	0
Unigene2506_All	1052	8.3131	3.2912	6.50E-12	3.18E-11
Unigene794_All	1047	46.7755	15.6461	4.45E-76	9.50E-75
Unigene11398_All	1044	28.6168	6.5833	2.02E-69	3.93E-68
Unigene54788_All	1041	37.963	0.4964	4.61E-217	2.94E-215
Unigene32168_All	1032	7.2374	1.0516	7.94E-25	6.52E-24
Unigene38187_All	885	11.4308	1.1094	1.58E-39	1.88E-38
Unigene11644_All	821	50.0359	18.7572	9.24E-55	1.43E-53
Unigene10680_All	794	6.0727	2.4732	6.79E-07	2.32E-06
Unigene12099_All	776	145.3496	55.9386	2.48E-141	9.93E-140
Unigene2582_All	731	9.3122	0.7776	1.86E-28	1.70E-27
Unigene54285_All	678	6.6237	1.9817	2.61E-09	1.09E-08
Unigene14544_All	655	52.5406	13.0178	1.11E-74	2.32E-73
Unigene54004_All	598	40.0786	0.9506	4.86E-125	1.71E-123

Unigene26944_All	590	21.7933	5.6056	5.82E-28	5.25E-27
Unigene18797_All	558	45.4931	2.2227	6.75E-118	2.22E-116
Unigene53813_All	557	2.5461	0	3.59E-09	1.48E-08
Unigene53633_All	527	55.4349	1.5689	9.81E-149	4.11E-147
Unigene53596_All	520	10.4544	1.1925	1.42E-20	1.01E-19
Unigene18841_All	517	117.586	19.4912	6.89E-177	3.47E-175
Unigene35385_All	511	6.7532	0.809	2.90E-13	1.53E-12
Unigene53415_All	497	37.3802	1.1437	4.50E-94	1.19E-92
Unigene53405_All	496	4.0982	0	7.73E-13	3.98E-12
Unigene6104_All	492	71.0043	25.1031	6.77E-51	9.92E-50
Unigene53260_All	477	3.0722	0.2167	2.57E-07	9.13E-07
Unigene23711_All	467	89.483	4.6476	1.73E-190	9.43E-189
Unigene53144_All	465	3.2531	0.6668	4.66E-05	1.31E-04
Unigene27253_All	458	4.4382	0.4513	8.65E-09	3.47E-08
Unigene55222_All	449	127.7085	47.6484	2.82E-76	6.02E-75
Unigene26504_All	444	46.4204	10.1258	1.53E-50	2.22E-49
Unigene37048_All	432	3.8299	0	1.40E-10	6.31E-10
Unigene52549_All	417	2.2672	0	2.38E-06	7.70E-06
Unigene18778_All	399	147.9772	39.3727	1.80E-118	5.96E-117
Unigene32553_All	398	3.682	0	1.88E-09	7.90E-09
Unigene14582_All	396	10.5049	3.5234	9.44E-08	3.48E-07
Unigene31349_All	387	13.6808	1.4688	1.48E-20	1.05E-19
Unigene27237_All	381	3.9704	0	9.80E-10	4.21E-09
Unigene51765_All	372	5.083	0	5.42E-12	2.67E-11
Unigene10228_All	334	74.1635	13.6154	1.67E-68	3.21E-67
Unigene55165_All	321	151.2414	45.3982	5.34E-86	1.28E-84
Unigene55309_All	318	156.0872	32.176	1.07E-124	3.75E-123
Unigene55211_All	295	60.2519	15.2402	1.20E-38	1.40E-37
Unigene55315_All	291	105.2658	22.553	2.29E-75	4.81E-74
Unigene58779_All	268	38.6291	12.7263	1.20E-17	7.65E-17
Unigene46694_All	261	5.6147	1.188	7.61E-05	2.07E-04
Unigene46697_All	261	9.4182	0.396	7.88E-13	4.05E-12
Unigene45374_All	247	13.9711	0	2.66E-21	1.94E-20
Unigene44461_All	239	16.8123	0	1.09E-24	8.92E-24
Unigene43287_All	227	36.235	2.9595	2.41E-34	2.55E-33
Unigene25473_All	221	103.3142	26.6567	2.80E-48	3.92E-47

CONCLUSION

De novo assembly of transcriptomes has been widely used to identify the biosynthetic and regulatory genes in medicinal plants such as *Cassia obtusifolia*, *Salvia miltiorrhiza* and *Panax quinquefolius*. This paper reported a comprehensive study on de novo assembly of transcriptome data of *E. longifolia*, one of the most important medicinal plants in Malaysia. Illumina Hi-Seq 2000 sequencing technology together with bioinformatics analysis generated 60,753 non-redundant unigenes from the roots of 1-and 10-year-old trees. This transcriptome dataset will serve as a public information platform for gene expression and functional genomics investigations of the species. Homology searches showed that only 57% unigenes were homologs to known proteins and most of the genes were similar to the *Citrus* genus. Detailed analysis of the transcriptome dataset showed that roots of *E. longifolia* are the site of extensive metabolic activity.

KEGG analysis revealed that most genes that encode for key enzymes in major secondary metabolites were present in the transcriptome. The broad coverage of secondary metabolic genes offers opportunities to examine the biosynthesis process of secondary metabolites in *E. longifolia*. Based on differential expression analysis, 154 unigenes that were up-regulated in the 10-year-old roots were potentially related to quassinoid biosynthesis. Twelve of the unigenes were involved in terpenoid backbone biosynthesis and sesquiterpenoid and triterpenoid biosynthesis pathways; 41 and 36 were similar to WRKY and AP2/ERF TFs, respectively; and 65 were similar to the cytochrome P450 enzyme. Analysis of the transcriptome of *E. longifolia* roots at 1 and 10 years of age has provided valuable resources for gene discovery, which will accelerate progress in molecular biology research for

this species. These results demonstrate that the Illumina paired-end sequencing is a reliable and cost-effective approach for gene discovery in non-model plants such as *E. longifolia*.

ACKNOWLEDGEMENT

We thank the staff of the Genetic Laboratory, FRIM for their assistance, and staff of Etnobotany and Kepong Botanical Garden, FRIM for providing *E. longifolia* root samples. This work was supported by FRIM and the 11th Malaysian Development Fund.

REFERENCES

1. Chan, K.L.,et al., Plants as sources of antimalarial drugs. Planta Medica, 1986. 52(02): p. 105-107.
2. Kardono , L.B.,et al., Cytotoxic and antimalarial constituents of the roots of Eurycoma longifolia. Journal of Natural Products, 1991. 54(5): p. 1360-1367.
3. Ang , H. H., et al., In vitro antimalaria activity of quassinoids from Eurycoma longifolia against Malaysian chloroquine-resistant Plasmodium falciparum isolates. Planta Medica, 1995. 61(3): p. 177-178.
4. Tada , H.,et al., New antiulcer quassinoids from Eurycoma longifolia. European Journal of Medicinal Chemistry, 1991. 26: p. 345-349.
5. Morita, H.,et al., New quassinoids from the roots of Eurycoma longifolia. Chemistry Letters, 1990. 19(5): p. 749-752.
6. Low, B. S.,et al., Standardized quassinoid-rich Eurycoma longifolia extract improved spermatogenesis and fertility in male rats via the hypothalamic-pituitary-gonadal axis. Journal of Ethnopharmacology, 2013. 145(3): p. 706-714.
7. Bhat, R., Karim, A. A., Tongkat Ali (Eurycoma longifolia Jack): a review on its ethnobotany and pharmacological importance. Fitoterapia, 2010. 81(4): p. 669-679.
8. Al-Salahi ,OS.,et al., Anti-angiogenic quassinoid-rich fraction from Eurycoma longifolia modulates endothelial cell function. Microvascular Research, 2013. 90: p. 30-39.
9. Rehman, S. U., Choe, K, Yoo, H. H., Review on a traditional herbal medicine, Eurycoma longifolia Jack (Tongkat Ali): its traditional uses, chemistry, evidence-based pharmacology and toxicology. Molecules, 2016. 21(3): p. 331-362.
10. Tung, NH.,et al., Quassinoids from the root of Eurycoma longifolia and their antiproliferative activity on human cancer cell lines. Pharmacognosy Magazine 2017. 13(51): p. 459-462.
11. Miyake , K.,et al.,Quassinoids from Eurycoma longifolia. Journal of Natural Products, 2009. 72(12): p. 2135-2140.
12. Chan, K.L.,et al., Antiplasmodial studies of Eurycoma longifolia Jack using the lactate dehydrogenase assay of Plasmodium falciparum. Journal of Ethnopharmacology, 2004. 92(2-3): p. 223-227.
13. Kuo ,P.C.,et al.Cytotoxic and antimalarial constituents from the roots of Eurycoma longifolia. Bio-organic Medical Chemistry, 2004. 12(3): p.537-544.
14. Abd Razak ,M.F., Aidoo, K.E., Candlish ,A.G., Mutagenic and cytotoxic properties of three herbal plants from Southeast Asia. Tropical Biomedicine, 2007. 24(2): p. 49-59.
15. Chakraborty, D., Pal, A., Quassinoids: Chemistry and novel detection techniques. Hand book of natural products: Phytochemistry, botany and metabolism of alkaloids, phenolics and terpenes. Springer-Verlag, 2013. 1(1): p. 3345-3366.
16. Bergman, M. E., Davis, B., Phillips, M. A., Medically useful plant terpenoids: Biosynthesis, occurrence and mechanism of action. Molecules, 2019. 24(21): p. 3961-3984.
17. Houel, E., et al., Quassinoids: anticancer and antimalaria activities. In: Ramawat KG, Merillon JM. Natural Products, 2013. 6: p. 161.
18. Guo, Z., et al., Biologically active quassinoids and their chemistry: potential leads for drug design. Current medicinal chemistry, 2005. 12(2):173-90.
19. Wu, Z. J., et al., De novo assembly and transcriptome characterization: novel insights into catechins biosynthesis in Camellia sinensis. BMC Plant Biology, 2014. 14(1) : p. 1-6.
20. Lulin, H., et al., The first illumine-based de novo transcriptome and analysis of Safflower flowers. PloS ONE, 2012. 7(6): p. e38653.
21. Conesa, A., et al., Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 2005. 21(18): p. 3674-3676.
22. Li, R., et al., De novo assembly of human genomes with massively parallel short read sequencing. Genome Research, 2010. 20(2): p. 265-272.

23. Hao, D., et al., De novo characterization of the root transcriptome of a traditional Chinese medicinal plant Polygonum cuspidatum. China Life Science, 2012. 55(5): p. 452-466.
24. Sun, H., et al., De novo leaf and root transcriptome analysis to identify putative genes involved in triterpenoid saponins biosynthesis in Hedera helix L. PlosOne, 2017. 12(8): p. e0182243.
25. Itokawa, H., et al.,C18 and C19 quassinoids from Eurycoma longifolia. Journal of Natural Products, 1993. 56(10): p. 1766-1771.
26. Morita, H., et al.,Squalene derivatives from Eurycoma longifolia. Phytochemistry, 1993. 34(3): p. 765-771.
27. Seo, J. W., et al., Overexpression of squalene synthase in Eluetherococcus senticosus increase phytosterol and triterpene accumulation. Phytochemistry, 2005. 66(8): p. 867-877.
28. Lee, M. H., et al., Enhanced triterpene and phytosterol biosynthesis in Panax ginseng overexpressing squalene synthase gene. Plant Cell Physiology,2004. 45(8): p. 976-984.
29. Kim, O. T., et al., Molecular characterization of ginseng farnesyl diphosphate synthase gene and its up-regulation by methyl jasmonate. Biology of Plant, 2010; 54(1): p.47-53.
30. Vom-Endt, D., Kijne, J. W., Memelink, J., Transcription factor controlling plant secondary metabolism: what regulates the regulators. Phytochemistry, 2002. 61(2): p. 107-114.
31. Misra, R. C., et al., Methyl jasmonate-elicited transcriptional responses and pentacyclic triterpene biosynthesis in sweet basil. Plant Physiology, 2014. 164(2): p. 1028-1044.
32. Yang, C. Q., et al., Transcriptional regulation of plant secondary metabolism. International Journal Plant Biology, 2012. 54(10) : p. 703-712.
33. Ma, D., et al., Isolation and characterization of AaWRKY1, an Artemisia annua transcription factor that regulates the amorpha-4,11-diene synthase gene, a key gene of artemisinin biosynthesis. Plant and Cell Physiology, 2009. 50(12): p. 2146-2161.
34. Singh, A. K., et al. A WRKY transcription factor from Withania somnifera regulates triterpenoid withanolide accumulation and biotic stress tolerance through modulation of phytosterol and defense pathways. New Phytology, 2017. 215(3): p. 1115-1131.
35. Van der Fits, L., Memelink, J.,The jasmonate-inducible AP2/ERF-domain transcription factor ORCA3 activates gene expression via interaction with a jasmonate-responsive promoter element. Plant Journal, 2001. 25(1): p. 43-53.
36. Coon, M. J., Cytochrome P450: nature's most versatile biological catalyst. Annual Review Pharmacological Toxicology, 2005. 45: p. 1-25.
37. Ghosh, S., Triterpene structural diversification by plant cytochrome P450 enzymes. Front Plant Sciences, 2017. 9(8): p. 1886.
38. Thimmappa, R., et al.,Triterpene biosynthesis in plants. Annual Review Plant Biology, 2014. 65: p. 225-257.
39. Liu, Z., De novo assembly and analysis of Cassia obtusifolia seed transcriptome to identify genes involved in the biosynthesis of active metabolites. Bioscience Biotechnology and Biochemistry, 2014. 78(5): p. 791-799.
40. Wenping, H.,et al., De novo transcriptome sequencing in Salvia miltiorrhiza to identify genes involved in the biosynthesis of active ingredients. Genomics, 2011. 98(4): p. 272-279.
41. Sun, C., et al., De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics, 2010.11: p. 262.