De novo transcriptome of *Taverniera cuneifolia* (Roth) Ali.

Talibali Momin[1], Apurva Punvar[2], Harshvardhan Zala[3], Garima Ayachit [2], Madhvi Joshi[2], Padamnabhi Nagar[1].

[1]Department of Botany, Faculty of Science, The Maharaja Sayajirao University of Baroda -390002.
[2]Gujarat Biotechnology Research Center (GBRC), Department of Science and Technology, Govt. of Gujarat. Gandhinagar 382 011.
[3] Department of Genetics and Plant Breeding, C. P. College of Agriculture, Sardarkrushinagar Dantiwada Agricultural University, Sardarkrushinagar - 385 506, Gujarat - INDIA

*Corresponding author:

Padamnabhi Nagar:drnagar@gmail.com
Talibali Momin: talib429gmail.com
Department of Botany,
Faculty of Science,
The Maharaja Sayajirao University of Baroda -390002.

ABSTRACT

Taverniera cuneifolia has been described as a potent substitute of Licorice in India. It has been used as an expectorant, anti-inflammatory, anti-ulcer, wound healing, blood purifier etc. Glycyrrhizin is one of the most useful bioactive sesquiterpenoid present in this plant. The present study aim to carry out transcriptome analysis in root tissue of *Taverniera cuneifolia* to identify specific functional genes involved in the biosynthesis of secondary metabolites. The root transcriptome sequencing of *Taverniera cuneifolia* resulted in a total of ~7.29 Gb of raw data and generated 55,991,233 raw reads. The high quality reads were *de novo* assembled by Trinity assembler followed through CD-HIT resulted into 35,590 “Unigene” transcripts with an average size of 419 bp. The unigenes were analyzed using BLAST2GO resulted in 27,884 (78.35%) transcript with blast hits, 22,510 (63.25%) transcript with mapping and 21,066 (59.19%) transcript with annotation. Functional annotation was carried out using NCBI’s non-redundant and Uniprot databases resulted in the identification of 21,066 (59.19%) annotated transcripts and GO assigned to 24751 (69.54%) transcripts. The gene ontology result shows maximum sequences match with Biological Processes (48%), Molecular Function (27%) and Cellular components (23%). A total of 289 metabolic enriched pathways were identified, which included pathways like Sesquiterpenoid and triterpenoid pathway which were involved in synthesis of secondary metabolite Glycyrrhizin biosynthesis. The enzymes, squalene monooxygenase, farnesyl-diphosphate farnesyltransferase, beta amyrin synthase, beta-amyrin 24-hydroxylase, were identified by functional annotation of transcriptome data. There were several other pathways like terpenoid backbone biosynthesis, steroid biosynthesis, Carotenoid biosynthesis, Flavonoids biosynthesis etc. which have been reported first time from this plant. Transcription factors were predicted by comparison with Plant Transcription Factor Database, and 1557 trancripts belonging to 85 trancription factor families were identified. This transcriptome analysis provided an important resource for future genomic studies in *Taverniera cuneifolia*, therefore representing basis in further investigation of the plant.

De novo transcriptome analysis of *Taverniera cuneifolia* (Roth) Ali.

ABSTRACT

Taverniera cuneifolia has been described as a potent substitute of Licorice in India. It has been used as an expectorant, anti-inflammatory, anti-ulcer, wound healing, blood purifier etc. Glycyrrhizin is one of the most useful bioactive sesquiterpenoid present in this plant. The present study Root transcriptome sequencing of *Taverniera cuneifolia* resulted in a total of ~7.29 Gb of raw data and generated 55,991,233 raw reads. The high quality reads were *de novo* assembled by Trinity assembler followed through CD-HIT resulted into 35,590 contigs transcripts with an average size of 419 bp. Functional annotation was carried out using NCBI's non-redundant and Uniprot databases resulted in the identification of 21,066 annotated transcripts and GO assigned to 24,751 transcripts. The gene ontology result shows maximum sequences match with Biological Processes (48%), Molecular Function (27%) and Cellular components (23%). A total of 289 metabolic enriched pathways were identified, which included pathways like Sesquiterpenoid and triterpenoid pathway which were involved in synthesis of secondary metabolite Glycyrrhizin biosynthesis. The enzymes, squalene monooxygenase, farnesyl-diphosphate farnesyltransferase, beta amyrin synthase, beta-amyrin 24-hydroxylase, were identified by functional annotation of transcriptome data. There were several other pathways like terpenoid backbone biosynthesis, steroid biosynthesis, Carotenoid biosynthesis, Flavonoids biosynthesis etc. which have been reported first time from this plant. Transcription factors were predicted by comparison with Plant Transcription Factor Database, and 1557 trancripts belonging to 85 trancription factor families were identified. This transcriptome analysis provided an important resource for future genomic studies in *Taverniera cuneifolia*, therefore representing basis in further investigation of the plant.

Significance

Licorice (*Glycyrrhiza glabra* roots) is used as traditional Chinese herbal medicines in majority of formulations. Licorice is also used in Industries like food, herbal and cosmetics etc. due to its high demand in the market it is imported from foreign countries and is not available locally of superior quality (Liu et al., 2015). In India, *Taverniera cuneifolia* has been described as a potent substitute of Licorice, it has been quoted in ancient books like Charak Samhita during the Nigandu period (Kamboj, 2000) and Barda dungar ni Vanaspati ane upyog (Thaker 1910). It has been used as an expectorant, anti-inflammatory, anti-ulcer,

wound healing, blood purifier etc. Transcriptomic studies will assist in understanding the basic molecular structure, function and organization of information within the genome of *Taverniera cuniefolia*. This study will help us to identify the key metabolites their expressions and genes responsible for their production.

Key words: *Taverniera cuneifolia*, *De novo assembly*, Transcriptome, Licorice, Glycyrrhizin, Sesquiterpenoid pathway.

Bioproject ID: 388043

This Transcriptome Shotgun Assembly project has been deposited at DDBJ/ENA/GenBank under the accession GJAF00000000. The version described in this paper is the first version, GJAF01000000.

Sequences Accession numbers: SRR5626167

1. Introduction

India is rich in many potential medicinal plants, *Glycyrrhiza glabra* popularly known as Liquorice has been used in the traditional formulation. A licorice (*Glycyrrhiza glabra*) root has been used in more than 1200 formulations in traditional Chinese herbal medicines as major formulations. There are many essential uses of this plant in industries like food, herbal, cosmetics, nutraceuticals etc. (Pastorino et al., 2018). Due to its high demand in the market, it is imported from foreign countries and not available locally of superior quality. In India, *Taverniera cuneifolia* has been described as a potent substitute for Licorice. Glycyrrhizin is one of the most useful bioactive sesquiterpenoid present in this plant.

Taverniera cuneifolia belong to fabaceae family, the third largest family of flowering plants, with over 800 genera and 20,000 species. The three major subfamilies include Mimosaceae, Papilionaceae and Caesalpiniacea. The pea (*Pisum sativum* L.) was the model organism used in Mendel's discovery (1866) and is the foundation of modern plant genetics. The phylogenetic differ greatly in their genome size, base chromosome number, ploidy level and reproductive biology. Two legume species in the Galegoid clade, *Medicago truncatula* and *Lotus japonicus*, from Trifolieae and Loteae tribe respectively, were selected as model system of studying legume genomics and biology. There are many other legumes that have been studies like the soybeans, the most widely grown and economically important legume whose genome has been available since 2010.The common bean (*Phaseolus vulgaris*) the most widely grown grain legume whose genome is available since 2014. Many more legumes have been sequenced since (Smýkal, P. et al., 2020).

Taverniera cuneifolia is an important traditional medicinal plant of India as mention in Charak Samita in Nigantu period. It is often referred to as Indian licorice having the same sweet taste as of *Glycyrrhiza glabra* (commercial Licorice) (Zore, 2008). The genus *Taverniera* has sixteen different species (Roskov et al., 2006). It is endemic to North-east Africa and South-west Asian countries (Naik, 1998). Licorice is used as important traditional Chinese medicine with many clinical and industrial applications like Food, Herbal medicine, cosmetics etc. (Liu et al. 2015). *Taverniera cuneifolia* locally known as Jethimad is used by the tribal's of Barda Hills of Jamnagar in Western India (Saurashtra, Gujarat) as a substitute for Licorice or in other words, the Plant itself is considered to be *Glycyrrhiza glabra* (Nagar, 2005). Many pharmacological benefits of the plants have been reported earlier like expectorant, blood purification, anti-inflammatory, wound healing, anti-ulcer and used in treating spleen tumors (Thaker, Manglorkar and Nagar, 2013).

At the Biochemical level, *Taverniera cuneifolia* has shown the presence of alkaloids, flavonoids, tannins, proteins, reducing sugar and saponins. The presence of oil content in the seeds of *Taverniera cuneifolia* showed polyunsaturated fatty acids, monounsaturated fatty acids and saturated fatty acids (Manglorkar, 2016). *Taverniera cuneifolia* has been assessed very less on phytochemical basis there are only few attempts to characterize this plant at molecular level. *Taverniera cuneifolia* has eight numbers of chromosomes (Perveen and Khatoon, 1989). There is limited information on genetic for this plant on NCBI. Fifteen proteins have been reported from this plants which includes ribosomal protein L32, maturase, photosystem 1 assembly protein Ycf4, cytochrome b6/f complex subunit VIII, D1 protein, photosystem 2 protein M, MaturaseK, ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit, Triosephosphate translocator, Phosphogluconate dehydrogenase, UDP-sulfoquinovose synthase, RNA polymerase beta subunit (Liu et al., 2017).

The current investigation was focused on the most valuable secondary metabolite, Glycyrrhizin and other important secondary metabolites. This experiment provides the in-depth characterizations of this plant. Based on the above facts attempts have been made to identify the genes of various metabolic pathways in *Taverniera cuneifolia* through root transcriptome sequencing. The study will give scientific insight into the molecular network of *Taverniera cuneifolia*.

Materials and Methods

Plant material and RNA isolation

Taverniera cuneifolia plant was collected from Kutch, Gujarat, India (23.7887 N, 68.79580 E) from its natural habitat near the area of Lakhpat. The tissue of the plant, i.e., roots were cleaned with water than with ethanol and stored in RNA later solution (Qiagen) for longer-term storage. It was then shifted to -20°C in the refrigerator. The total RNA was isolated from the root tissues of the Plant using the RNeasy Plant Mini Kit (Qiagen) following the manufacturer's instructions. The integrity of the RNA was assessed by formaldehyde agarose gel electrophoresis. Total RNA was quantified by using a Qiaxpert (Qiagen), Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA, USA) and Qiaxcel capillary electrophoresis (Qiagen). RNA integrity number (RIN) was higher than approx. 7.0 for the sample.

cDNA library preparation and Sequencing

Ribosomal RNA depletion was carried out using a RiboMinus RNA plant kit for RNA- Seq (Life Technologies, C.A). mRNA fragmentation and cDNA library was constructed using an Ion total RNA-Seq kit v2 (Life Technologies, C.A), further purified using AMpure XP beads (Beckman coulter, Brea, CA, USA). The library was enriched on Ion sphere particles using Dynabeads MyOne Streptavidin C1 using standard protocols for the Ion Proton sequencing. The raw transcriptome data have been deposited in the sequence read archive (SRA) NCBI database with the accession number SRR5626167. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GJAF00000000.

RNA-Seq data processing and *De novo* assembly

Quality control of raw sequence reads was filtered to obtain the high-quality clean reads using bioinformatics tools such as FASTQCv.0.11.5 using a minimum quality threshold Q20 (Andrews, 2010). The clean reads were subjected to de novo assembly using the Trinity v2.4.0 (Grabherr et al., 2011) software to recover full-length transcripts. The redundancy of

Trinity generated contigs were clustered for removing duplicate reads with 85% identity using CD-HIT v4.6.1 (Li and Godzik, 2006).

Functional annotation of transcripts and classification

Functional characterization of assembled sequences was done by performing BlastX of contigs against the non-redundant (nr) database, (https://www.ncbi.nlm.nih.gov/) using an e-value cut-off of 1E-5 followed by further annotation was carried out using Blast2GO (Conesa and Gotz, 2005). Gene Ontology (GO) study was used to classify the functions of the predicted coding sequences. The GO classified the functionally annotated coding sequences into three main domains: Biological process (BP), Molecular function (MF) and Cellular component (CC). Using the Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa and Goto, 2000) pathway maps were determined. Further, KEGG Automated Annotation Server (KAAS) was used for pathway mapping in addition to Blast2GO (Moriya et al., 2007) for assignment and mapping of the coding DNA sequence (CDS) to the biological pathways. KAAS provides functional annotation of genes by BLAST comparison against the manually curated KEGG genes database.

Identification of transcription factors families

Transcription factors (TFs) were identified using genome-scale protein and nucleic acid sequences by analyzing InterProScan domain patterns in protein sequences with high coverage and sensitivity using PlantTFcat analysis tool (http://plantgrn.noble.org/PlantTFcat/) tool (Dai et al., 2013).

SSR prediction

Simple sequence repeats (SSRs) were identified using the MISA tool (Microsatellite; http://pgrc.ipk-gatersleben.de/misa/misa.html). We searched for SSRs ranging from mono to hexanucleotide in size. The minimum repeats number 10 for mononucleotide, 6 for Dinucleotide and 5 for trinucleotide to hexanucleotide was set for SSR search. The maximal number of bases interrupting two SSRs in a compound microsatellite is 100 i.e. the minimum distance between two adjacent SSR markers was set 100 bases.

Results and Discussion

Transcriptome Sequencing and *De novo* assembly

The total RNA of two root samples along with RIN value more than 7.0, converted to cDNA library using Ion Total RNA-Seq kit v2 (Life Technologies, C.A), further purified using Ampure XP beads (Beckman coulter, Brea, CA, USA). The library was enriched on Ion sphere particles using Myone C1 Dynabeads. A total of 7.29 gb of raw data was generated using standard protocols for the Ion proton sequencing (Table 1). The good quality roots of *Taverniera cuneifolia* were used for the RNA sequencing, and a total of 55,991,233 reads containing 7,286,727,421 bases were generated. The raw reads were subjected to quality check by FastQC tool and the average base quality was above Q20. De novo transcriptome assembly resulted in 36,896 reads assembled and the final assembly of 35,590 unique high-quality reads was prepared using CD-HIT at 85% sequence similarity, with N50 value of 441 bp. The average GC content of 43% and average contigs length of 419.45 bp was obtained for *Taverniera cuneifolia*. The statistics of transcriptome sequencing and assembly generated by Trinity assembler as given (Table 2).

Functional annotation of transcripts

A total of 35,590 transcripts (contigs) assembled by Trinity were subjected to functional annotation using different databases like the Nr Protein database, KEGG, UniProt, etc. GO terms were assigned to transcripts (Supplementary Fig. S2). All transcripts were screened for similarity to a known organism based on the data of species-specific distribution, and it can be concluded that the transcript showed the highest blast hits with *Medicago truncatula* (18,734 , 52.63%) followed by *Cicer arietinum* (16,044, 45.08%) and *Glycine max* (15,991, 44.93%). A total of 10590 (29.75%), 8642 (24.28%), 8549 (24.02%), 8399 (23.59%) contigs were found to be similar to *Cajanus cajan*, *Glycine soja*, *Trifolium pratense*, *Trifolium subterraneum*, respectively (Figure 1). The functionally annotated transcripts (27,884, 78.34%) of *Taverniera cuneifolia* were classified using Blast2GO into three main domains; Biological processes, Cellular component and Molecular function gene ontology (Table S1). Among them the most abundant were the Biological processes consisting of 44,395(48.8%) sequences followed by different Molecular Function consisting of 25,025 (27.5%) sequences and last the cellular components consist of 21,508 (23.6%) sequences (Figure 2, 3, 4). The annotated transcripts were subjected to the Kyoto encyclopedia genes and genomes (KEGG) pathway wherein the transcripts were linked to enzymes found in a large number of pathways available in KEGG. The maximum number of annotated transcripts assigned to hydrolases, followed by transferases and oxidoreductases class of enzymes (Figure 5).

Gene ontology classification

The contigs were further annotated by Blast2Go software with assembled 27,884 transcripts GO terms and divided into three broad categories as Biological Processes (44,395[49%]), Molecular Function (25,025[27%]) and Cellular Component (21,508[24%]) category (Table S1).

The Biological Processes were the most abundant component of GO terms. Among the 44,395 Biological Processes, the maximum number of contigs i.e. represented “Biological process,” followed by “Metabolic process” and “Cellular process” (Figure 2).

A total of 25,025 transcripts were associated with the Molecular function and a relatively large no of the transcript was associated with “Molecular function” followed by “Catalytic activity” and “Binding”, respectively ”(Figure 3).

In addition, Cellular Component a total of 21,508 transcripts were associated with the “Cellular component” as the highest match followed by “Cell” and “Cell part” respectively”(Figure 4).

Pathway Annotation by KEGG

Kyoto Encyclopedia of Genes and Genomes (KEGG) serves as knowledge source to perform functional annotation of the genes. The KEGG represents various biochemical pathways for the genes associated with it. Approximately 289 pathways were annotated and among them, Metabolic pathways (102), Biosynthesis of secondary metabolites (55), Microbial metabolism in diverse environment (22) showed the maximum hit with the database. Some of the important pathways from this plant are discussed below which have been reported with the gene and ko-id. (Table S2).

Terpenoids (isoprenoids) represent the largest and most diverse class of chemicals among the myriad compounds produced by plants. Moreover, the ecological importance of terpenoids has gained increased attention to develop strategies for sustainable pest control and abiotic stress protection. The gene that has shown in this plant includes **Terpenoids**

backbone biosynthesis (ko00900) (Supplementary Fig. S3). which includes three gene, ko:K03526 gcpE; (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase [EC:1.17.7.1 1.17.7.3], ko:K05356 SPS; all-trans-nonaprenyl-diphosphate synthase [EC:2.5.1.84 2.5.1.85], ko:K15889 PCME; prenylcysteine alpha-carboxyl methylesterase [EC:3.1.1.-]. Monoterpenoid biosynthesis having two gene ko: K21373 UGT8; 7-deoxyloganetic acid glucosyltransferase [EC:2.4.1.323], ko:K21374 UGT85A23_24; 7-deoxyloganetin glucosyltransferase [EC:2.4.1.324] and Diterpenoid biosynthesis (ko00904) includes ko:K05282 GA20ox; gibberellin-44 dioxygenase [EC:1.14.11.12].

Sesquiterpenoid and triterpenoid biosynthesis (ko00909) (Supplementary Fig. S4). which includes three gene namely ko:K00801 FDFT1; farnesyl-diphosphate farnesyltransferase [EC:2.5.1.21], ko:K15813 LUP4; beta-amyrin synthase [EC:5.4.99.39], ko:K20658 PSM; alpha/beta-amyrin synthase [EC:5.4.99.40 5.4.99.39]. This are the gene on further reactions like oxidation and reductions leads to the production of Glycyrrhizin that is important secondary metabolites as mention above.

Carotenoid biosynthesis (ko00906) (Supplementary Fig. S5) includes ko:K09842 AAO3; abscisic-aldehyde oxidase [EC:1.2.3.14], ko:K09843 CYP707A; (+)-abscisic acid 8'-hydroxylase [EC:1.14.14.137], ko:K14595 AOG; abscisate beta-glucosyltransferase [EC:2.4.1.263].

Ubiquinone and other terpenoid-quinone biosynthesis (ko00130) (Supplementary Fig. S6) include ko:K03809 wrbA; NAD(P)H dehydrogenase (quinone) [EC:1.6.5.2].

Zeatin biosynthesis (ko00908) (Supplementary Fig. S7) includes ko:K00791 miaA; tRNA dimethylallyltransferase [EC:2.5.1.75], ko:K13496 UGT73C; UDP-glucosyltransferase 73C [EC:2.4.1.-].

Flavonoid biosynthesis (ko00941) (Supplementary Fig. S8) inludes ko:K13065 E2.3.1.133; shikimate O-hydroxycinnamoyltransferase [EC:2.3.1.133].

Candidate genes involved in biosynthesis pathways

Among the 35,591 transcripts that have been annotated using different database, we have identified six gene that play important role in the biosynthesis pathway of Glycyrrhizin production from *Taverniera cuneifolia* (Table S3). Each six different gene includes in formation of Glycyrrhizin.

There were 4912 unigenes hypothetical protein predicted from this plant, of which 30 unigenes that had a hit length above 400 were noted (Table S4). 94 unigenes that predicted Cytochrome P450 family protein from this plant, of which 17 unigenes with a hit length above 150 were noted (Table S5).

Discussion

Secondary metabolites have key role in providing the defense mechanism to plants against stresses and these metabolites have very important role in many economic important like industries, pharma sector etc (Pagare et al., 2015). There has been no molecular data recorded for this plant as such. The new advancement in the field of omics technologies has led to high-throughtput sequencing data which lead us to prediction of genes, enzymes, complex pathways. (Metzker,2010). De novo of many medicinally important plants such as *Saussurea lappa* (Bains, S et al, 2018), *Vigna radiate* L (Chen, H et al, 2015), *Glycyrrhiza glabra* (Chin,Y et al, 2007), pigeonpea *Cajanus cajan* (L .) Millspaugh (Dutta, S. et al, 2011), *Dracocephalum tanguticum* (Li, H., Fu, Y., Sun, H., Zhang, Y., & Lan, X., 2017) etc. have reported the trancripts involved in active metabolite production using NGS technology.

Transcriptome analysis has proved to be one of the advanced methods for the identification of gene expressing in different pathways of metabolism, growth, development, response towards stress, cell signaling etc. This has help in classifying and categorization different role in secondary metabolic compound. Glycyrrhizin, a well-known secondary metabolite that is found in roots of Licorice has same property that is been found in the roots *Taverniera cuneifolia* which has many uses as described above. A whole transcriptome analysis of root of *Taverniera cuneifolia* has opened the unique transcripts which are reported first time from this plant to be involved in the pathways of primary and secondary metabolism (Sharma, Kumar, Beriwal, et al, 2019).

The de novo assembled transcripts of *Taverniera cuneifolia* were mapped to non-redundant protein database using blastx tool. A total of 35,590 transcripts annotated to the database showed the maximum similarity with *Medicago truncatula* [(18,734) 52.6 %] followed by *Cicer arietinum* [(16,044) 45%] and *Glycine max* [(15,991) 44.9%] and so on, which belong to same family Fabaceae order fabales.

Main metabolism-related gene of *Taverniera cuneifolia.*

Glycyrrhizin is triterpenoid-saponin produced in Licorice roots. It is synthesized via the cytosolic melvonic acid pathway for the production of 2,3-oxidosqualene, which is then cyclized to β-amyrin by β-amyrin synthase (bAS). Then, β-amyrin undergoes a two-step oxidation at the C-30 position followed by glycosylation reactions at the C-3 hydroxyl group to synthesize glycyrrhizin as shown in (Supplementary Fig. S9)(Seki et al 2008, 2011). *Taverniera cuneifolia* also known as Indian Licorice can be used as substitute of *Glycyrrhiza glabra* as it has same features that of this plant. This plant contains varieties of different compound that can be used in future research like triterpenoids, flavonoids, polysaccharides etc, which have been reported first time from this plant. Among them Glycyrrhizin is a primary focus compound that has many economic importance use in different fields. In our experiment we have compared the enzymes and genes for the production of Glycyrrhizin with proposed pathway for biosynthesis of Glycyrrhizin by (Seki et al, 2011), In which Glycyrrhizin is produce by a series of chemical reaction i.e. oxidation of different compound associated with Melvanoic Acid pathway. In this particular pathway there are series of chemical reaction by which Farnesyl diphosphate (FPP) molecule catalyzed by squalene synthase (SQS) originating Squalene. There are fifteen different transcripts that we have found in our plants that are associated for the production of squalene and then by oxidation by squalene epoxidase (SQE) to 2, 3 – oxidosqualene to form β- Amyrin. There are five gene identified from our plant that catalyzed by bAS i.e β- Amyrin synthase to form β- Amyrin. Further β- Amyrin goes into various oxidation reaction with the help of Beta-amyrin 11-oxidase /CYP88D6 and 11-oxo-beta-amyrin 30-oxidase/CYP72A154 to form Glycyrrhetinic acid. The last step includes conversion of glycyrrhetinic acid to glycyrrhizin which includes glycosylation steps in which different enzymes related to UDP-glycuronosyl transferases family are included. There were 32 different UDP-glycuronosyl genes which have been identified from our plant that led to last reaction given in table (Table S3).

At this point *Taverniera cuneifolia* have not been intensively studied and there as such no any reports that showed the details about the enzymes associated in the Glycyrrhizin pathway we have associated with reference pathway proposed by (seki et al 2008, 2011). As there has been no proper investigation for the pathway for glycyrrhizin known till today.

We have extensively worked upon the proteins which we have opted from our data of *Taverniera cuneifolia*. Approx. 4912 genes have been isolated that showed different proteins reported firstly from this plant among them the details have been provided in (Table S4) (we have approx. shown only those hypothetical proteins whose hit length is above 400 bases). In our studies we also found that there were more than 90 transcripts that showed the function related to Cytochrome P450 family protein. This protein has an immense ability to synthesis many new molecules required in the system to function and cope up with.

Identification of SSR markers and Transcription factors

The potential SSR from mono to hexanucleotide were predicted using MISA Perl script. A total of 35,590 unigene sequences were examined and 2912 SSR were obtained. It was found that only 2454 number of sequences were containing SSRs. Further, only 365 sequences contained >1 SSR marker and 265 were present in compound form. Tri-nucleotide represented the maximum numbers of SSRs (1291), followed by Mono-nucleotide (832) and then Di-nucleotide (597) (Table 3). The analysis of the transcripts revealed 1557 unique transcripts belonging to 85 transcription factor families. Among the identified unigenes, the highest of them represented the WD40 family followed by C2H2, MYB-HB, AP2-EREBP, PHD etc. the top 15 have been shown in the table.(Figure 6).

Acknowledgments:

We are grateful to GBRC (Gujarat Biotechnology Research Centre) for providing the platform for performing the experiment. All the facilities were provided by GBRC including Computational Analysis. The Department of Botany, Faculty of Science, The Maharaja Sayajirao University Baroda for the all the supports for this work.

Author’s contribution:

All authors have contributed to various aspects of this work. PN and MJ conceived the idea and designed the experiments. TM and HZ performed the experiment. TM, HZ, GA and AP analyzed the data. TM analyzed the results and wrote the manuscript. PSN, HZ and MJ finalized the manuscript.

References

A, W. L., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, 22(13), 1658-1659. https://doi.org/10.1093/bioinformatics/btl158.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of molecular biology 215, 403-410.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bains, S., Thakur, V., Kaur, J., Singh, K., & Kaur, R. (2018). Genomics Elucidating genes involved in sesquiterpenoid and flavonoid biosynthetic pathways in *Saussurea lappa* by de novo leaf transcriptome analysis. Genomics, 0-1. https://doi.org/10.1016/j.ygeno.2018.09.022.

Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. Bioinformatics (Oxford, England), 33(16), 2583–2585. https://doi.org/10.1093/bioinformatics/btx198.

Chen, H., Wang, L., Wang, S., Liu, C., Blair, M. W., & Cheng, X. (2015). Transcriptome sequencing of mung bean (*Vigna radiate* L.) genes and the identification of EST-SSR markers. PLoS ONE, 10(4). https://doi.org/10.1371/journal.pone.0120273

Chin, Y. W., Jung, H. A., Liu, Y., Su, B. N., Castoro, J. A., Keller, W. J., … Kinghorn, A. D. (2007). Anti-oxidant constituents of the roots and stolons of licorice (*Glycyrrhiza glabra*). Journal of Agricultural and Food Chemistry, 55(12), 4691–4697. https://doi.org/10.1021/jf0703553.

Chirumbolo, S. (2016). Commentary: The antiviral and antimicrobial activities of licorice, a widely-used Chinese herb, 7(April), 1–3. https://doi.org/10.1002/ptr.2295

Chomczynski, P., & Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Analytical Biochemistry, 162(1), 156–159. https://doi.org/10.1016/0003-2697(87)90021-2

Conesa, A., Götz, S., García-gómez, J. M., Terol, J., Talón, M., Genómica, D., … Valencia, U. P. De. (2005). Blast2GO : a universal tool for annotation , visualization and analysis in functional genomics research, 21(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610.

Dai, X., Sinharoy, S., Udvardi, M., & Zhao, P. X. (2013). PlantTFcat: An online plant transcription factor and transcriptional regulator categorization and analysis tool. BMC Bioinformatics, 14(1). https://doi.org/10.1186/1471-2105-14-321.

Dutta, S., Kumawat, G., Singh, B. P., Gupta, D. K., Singh, S., Dogra, V., Singh, N. K. (2011). Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L .) Millspaugh]. https://doi.org/10.1186/1471-2229-11-17

Garg, R., & Jain, M. (2013). RNA-Seq for transcriptome analysis in non-model plants. Methods in Molecular Biology. https://doi.org/10.1007/978-1-62703-613-9_4

Ghawana, S., Paul, A., Kumar, H., Kumar, A., Singh, H., Bhardwaj, P. Kumar, S. (2011). An RNA isolation system for plant tissues rich in secondary metabolites. BMC Research Notes, 4(1), 85. https://doi.org/10.1186/1756-0500-4-85

Gohil Amit, N., & Daniel, M. (2014). Development of quality standards of *Taverniera cuneifolia* (Roth) arn. root - A substitute drug for liquorice. International Journal of Pharmacognosy and Phytochemical Research, 6(2), 255–259.

Gore, R., & Gaikwad, S. (2015). Checklist of Fabaceae Lindley in Balaghat Ranges of Maharashtra, India. Biodiversity Data Journal, 3, e4541. https://doi.org/10.3897/BDJ.3.e4541

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I. Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome, 29(7). https://doi.org/10.1038/nbt.1883

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Jr, R. K. S., Hannick, L. I., White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, 31(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

J. Thaker, Kathiyawadna Bardadungarni jadibuti teni pariksha ane upyog, Gujarati Press Publishers, Mumbai (1910).

Kamboj VP (2000). Herbal Medicine. Current Science, 78, 35-9.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 28, 27–30.

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., & Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data, 1–21. https://doi.org/10.1186/s13059-014-0553-5

Li, H., Fu, Y., Sun, H., Zhang, Y., & Lan, X. (2017). Transcriptomic analyses reveal biosynthetic genes related to rosmarinic acid in *Dracocephalum tanguticum*. Scientific Reports, (January), 1–10. https://doi.org/10.1038/s41598-017-00078.

Li, J., Dai, X., Zhuang, Z., & Zhao, P. X. (2016). LegumeIP 2.0—a platform for the study of gene function and genome evolution in legumes. Nucleic Acids Research, 44(D1), D1189–D1194. https://doi.org/10.1093/nar/gkv1237

Li, Y., Luo, H.-M., Sun, C., Song, J.-Y., Sun, Y.-Z., Wu, Q., Chen, S.-L. (2010). EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. BMC Genomics, 11(268), 268. https://doi.org/10.1186/1471-2164-11-268.

Liao, Z., Chen, M., Guo, L., Gong, Y., Tang, F., Sun, X., & Tang, K. (2004). Rapid isolation of high-quality total RNA from taxus and ginkgo. Preparative Biochemistry & Biotechnology, 34(3), 209–214. https://doi.org/10.1081/PB-200026790

Liu, P. L., Wen, J., Duan, L., Arslan, E., Ertuğrul, K., & Chang, Z. Y. (2017). Hedysarum L.(Fabaceae: Hedysareae) is not monophyletic–evidence from phylogenetic analyses based on five nuclear and five plastid sequences. *PLoS One*, *12*(1), e0170596.

Liu, Y., Zhang, P., Song, M., Hou, J., Qing, M., Wang, W., & Liu, C. (2015). Transcriptome analysis and development of SSR molecular markers in *Glycyrrhiza uralensis* fisch. PLoS ONE, 10(11), 1–12. https://doi.org/10.1371/journal.pone.0143017

Mangalorkar, Bioprospecting the potential of *Taverniera cuneifolia* Roth Ali. Ph.D Thesis in Department of Botany, Faculty of Science, The Maharaja Sayajirao University of Baroda. Gujarat, India (2016).

Maroufi, A. (2016). Selection of reference genes for real-time quantitative PCR analysis of gene expression in *Glycyrrhiza glabra* under drought stress. Biologia Plantarum, 60(4). https://doi.org/10.1007/s10535-016-0601-y

Metzker, M. L. (2010). Sequencing technologies - the next generation. Nature Reviews. Genetics, 11(1), 31–46. https://doi.org/10.1038/nrg2626

Mochida, K., Sakurai, T., Seki, H., Yoshida, T., Takahagi, K., Sawai, S., Saito, K. (2017). Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume. Plant Journal, 89(2), 181–194. https://doi.org/10.1111/tpj.13385

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research 35, W182--W185.

Nadiya, F., Anjali, N., Thomas, J., Gangaprasad, A., & Sabu, K. K. (2017). Transcriptome profiling of *Elettaria cardamomum* (L.) Maton (small cardamom). Genomics Data, 11, 102–103. https://doi.org/10.1016/j.gdata.2016.12.013.

Naik, V.N., 1998. Flora of Marathwada (Ranunculaceae to Convolvulaceae). Amrut Prakashan, Aurangabad, India.

P. Sharma, S. Kumar, S. Beriwal, et al., Comparative transcriptome profiling and co-expression network analysis reveals functionally coordinated genes associated with metabolic processes of *Andrographis paniculata*, Plant Gene (2019). https://doi.org/10.1016/j.plgene.2020.100234

P.S.Nagar, Floristic Biodiversity of Barda Hills and its Surroundings, Scientific Publishers, Jodhpur, India (2005).

Pagare, Saurabh, Bhatia, M., Tripathi, N., Pagare, Sonal, Bansal, Y.K., 2015. Secondary metabolites of plants and their role: Overview. Current Trends Biotechnology Pharm 9, 293 –304.

Pastorino, G, Cornara, L, Soares, S, Rodrigues, F, Oliveira, MBPP (2018). Liquorice (*Glycyrrhiza glabra*): A phytochemical and pharmacological review. Phytotherapy Research. 2018; 32: 2323– 2339. https://doi.org/10.1002/ptr.6178.

Perveen, Shaista. & Khatoon, Surayya. (1989). Chromosome numbers in Papilionaceae from Pakistan. Pakistan J. Bot, 21, 247-251.

Ramilowski, J. A., Sawai, S., Seki, H., Mochida, K., Yoshida, T., Sakurai, T., Daub, C. O. (2013). *Glycyrrhiza uralensis* transcriptome landscape and study of phytochemicals. Plant and Cell Physiology, 54(5), 697–710. https://doi.org/10.1093/pcp/pct057.

Rasool, S., & Mohamed, R. (2016). Plant cytochrome P450s: nomenclature and involvement in natural product biosynthesis. Protoplasma. https://doi.org/10.1007/s00709-015-0884-4.

Roskov Y.R., Bisby F.A., Zarucchi J.L., Schrire B.D. & White R.J. (eds.) ILDIS World Database of Legumes: draft checklist, version 10 [published June 2006, but CD shows November 2005 date]. ILDIS, Reading, UK, 2006 [CD-Rom: ISBN 0 7049 1248 1] (also available here at https://ildis.org/LegumeWeb10.01.shtml).

School of graduate studies faculty of science departement of chemistry Bioassay Guided Phytochemical Investigation on Roots of *Taverniera Abyssinica* (Dingetegna) By : Mekuriaw Assefa Advisor : Ermias Dagne (Professor) July , (2010).

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., Muranaka, T. (2008). Licorice -amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin. Proceedings of the National Academy of Sciences, 105(37), 14204–14209. https://doi.org/10.1073/pnas.0803876105

Seki, H., Sawai, S., Ohyama, K., Mizutani, M., Ohnishi, T., Sudo, H., Muranaka, T. (2011). Triterpene Functional Genomics in Licorice for Identification of CYP72A154 Involved in the Biosynthesis of Glycyrrhizin. The Plant Cell, 23(11), 4112–4123. https://doi.org/10.1105/tpc.110.082685.

Smýkal, P., von Wettberg, E. J., & McPhee, K. (2020). Legume genetics and biology: from Mendel's pea to legume genomics.

Stadler, M., Dagne, E., Anke, H., 1994. Nematicidal activity of two phytoalexins form *Taverniera abyssynica*. Planta Med. 60 (6), 550-552.

Sudo, H., Seki, H., Sakurai, N., Suzuki, H., Shibata, D., Toyoda, A., Saito, K. (2009). Expressed sequence tags from rhizomes of *Glycyrrhiza uralensis*. Plant Biotechnology, 26(1), 105–107. https://doi.org/10.5511/plantbiotechnology.26.105

Thiel, T., Michalek, W., Varshney, K., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L .), 411–422. https://doi.org/10.1007/s00122-002-1031-0

V.N.Naik, Flora of Marathawada (Ranunculaceae to convolvulaceae), Amrut prakashan, Aurangabad, India (1998).

Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genic microsatellite markers in plants : features and applications, 23(1). https://doi.org/10.1016/j.tibtech.2004.11.005

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., … Zhang, G. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nature Biotechnology, 31(3), 240–246. https://doi.org/10.1038/nbt.2491

Villa-Ruano, N., Pacheco-Hernández, Y., Lozoya-Gloria, E., Castro-Juárez, C. J., Mosso-Gonzalez, C., & Ramirez-Garcia, S. A. (2015). Cytochrome P450 from Plants: Platforms for valuable phytopharmaceuticals. Tropical Journal of Pharmaceutical Research. https://doi.org/10.4314/tjpr.v14i4.24

Wolf, J. B. W. (2013). Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial, 559–572. https://doi.org/10.1111/1755-0998.12109

Yang, R., Yuan, B., Ma, Y., Wang, L., Liu, C., & Liu, Y. (2015). HMGR, SQS, β-AS, and Cytochrome P450 Monooxygenase Genes in *Glycyrrhiza uralensis*. Chinese Herbal Medicines, 7(4), 290–295. https://doi.org/10.1016/s1674-6384(15)60054-5

Zhang, C., Zhang, B., Vincent, M. S., Zhao, S., & Quantification, G. (2016). Bioinformatics Tools for RNA-seq Gene and Isoform Quantification Next Generation Sequencing & Applications, 3(3). https://doi.org/10.4172/2469-9853.1000140

Zhang, Y., Zhang, X., Wang, Y.-H., & Shen, S.-K. (2017). De Novo Assembly of Transcriptome and Development of Novel EST-SSR Markers in *Rhododendron rex* Lévl. through Illumina Sequencing. Frontiers in Plant Science, 8(September), 1–12. https://doi.org/10.3389/fpls.2017.01664

Zore, G. B., Winston, U. B., Surwase, B. S., Meshram, N. S., Sangle, V. D., Kulkarni, S. S., & Mohan Karuppayil, S. (2008). Chemoprofile and bioactivities of *Taverniera cuneifolia* (Roth) Arn.: A wild relative and possible substitute of *Glycyrrhiza glabra* L. Phytomedicine, 15(4), 292–300. https://doi.org/10.1016/j.phymed.2007.01.006.

Captions for Tables:

Table 1: Summary of sequencing data generated for root sample of *Taverniera cuneifolia*.

Table 2: Results based on combined assembly of *Taverniera cuneifolia* root transcriptome.

Table 3: Identification of Simple Sequence Repeats (SSRs) from *Taverniera cuneifolia* root transcriptome.

Table 4: Candidate "Unigenes" encoding enzymes involved in the Sesquiterpenoid and Triterpenoid biosynthesis, Flavonoid biosynthesis, Terpenoid backbone biosynthesis, Carotenoid biosynthesis, Monoterpenoid biosynthesis and Zeatin biosynthesis identified from *Taverniera cuneifolia* Transcriptome.

Table 1: Summary of sequencing data generated for root sample of *Taverniera cuneifolia*.

Sr. No.	Features	Raw data	
		Sample run 1	Sample run 2
1	Total reads	26,652,853	29,338,380
2	Total nucleotides (bp)	3,604,710,778	3,682,016,643
3	Mean read length (bp)	135 bp	126 bp

Table 2: Result based on combined assembly of *Taverniera cuneifolia* root transcriptome.

Sr. No.	Characteristics	Values
1	Total assembled contigs/transcript	35,590
2	GC %	43.25
3	Contig N_{50} (bp)	441
4	Median Contig length (bp)	322
5	Average Contig length (bp)	419.45
6	Total assembled bases	14,928,144

Table 3: Identification of Simple Sequence Repeats (SSRs) from *Taverniera cuneifolia* root transcriptome.

SSR statistics	Count
Total number of sequences examined	35,590
Total size of examined sequences (bp)	1,49,28,144
Total number of identified SSRs	2,912
Number of SSR containing sequences	2,454
Number of sequences containing more than 1 SSR	365
Number of SSRs present in compound formation	265
Mono-nucleotide	832
Di-nucleotide	597
Tri-nucleotide	1291
Tetra-nucleotide	153
Penta-nucleotide	33
Hexa-nucleotide	6

Table 4: Candidate "Unigenes" encoding enzymes involved in the Sesquiterpenoid and Triterpenoid biosynthesis, Flavonoid biosynthesis, Terpenoid backbone biosynthesis, Carotenoid biosynthesis, Monoterpenoid biosynthesis and Zeatin biosynthesis identified from *Taverniera cuneifolia* Transcriptome.

Pathway	Name	Description	KO no.	EC no.
Sesquiterpenoid and Triterpenoid biosynthesis	FDFT1	farnesyl-diphosphate farnesyltransferase	ko:K00801	2.5.1.21
	LUP4	beta-amyrin synthase	ko:K15813	5.4.99.39
	PSM	alpha/beta-amyrin synthase	ko:K20658	5.4.99.40

				5.4.99.39
Flavanoid biosynthesis		shikimate O-hydroxycinnamoyltransferase	ko:K13065	2.3.1.133
Terpenoid backbone biosynthesis	gcpE	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	ko:K03526	1.17.7.1 1.17.7.3
	SPS	all-trans-nonaprenyl-diphosphate synthase	ko:K05356	2.5.1.84 2.5.1.85
	PCME	prenylcysteine alpha-carboxyl methylesterase	ko:K15889	3.1.1.-]
Carotenoid biosynthesis	AAO3	abscisic-aldehyde oxidase	ko:K09842	1.2.3.14
	CYP707A	(+)-abscisic acid 8'-hydroxylase	ko:K09843	1.14.14.137
	AOG	abscisate beta-glucosyltransferase	ko:K14595	2.4.1.263
Monoterpenoid biosynthesis	UGT8	7-deoxyloganetic acid glucosyltransferase	ko:K21373	2.4.1.323
	UGT85A23_24	7-deoxyloganetin glucosyltransferase	ko:K21374	2.4.1.324
Zeatin biosynthesis	miaA	tRNA dimethylallyltransferase	ko:K00791	2.5.1.75
	UGT73C	UDP-glucosyltransferase 73C	ko:K13496	2.4.1.-

24

Captions for Figures:

Figure 1: Species distribution of the top BLAST hits of *Taverniera cuneifolia* transcripts in Nr database

Figure 2: Biological processes gene ontology of *Taverniera cuneifolia* transcripts

Figure 3: Molecular functions gene ontology of *Taverniera cuneifolia* transcripts

Figure 4: Cellular components gene ontology of *Taverniera cuneifolia* transcripts

Figure 5: Enzyme classification of *Taverniera cuneifolia* transcripts based on KEGG pathway

Figure 6: Top 15 Transcription factors families detection from *Taverniera cuneifolia* root transcriptome.

Figure 1: Species distribution of the top BLAST hits of *Taverniera cuneifolia* transcripts in Nr database.

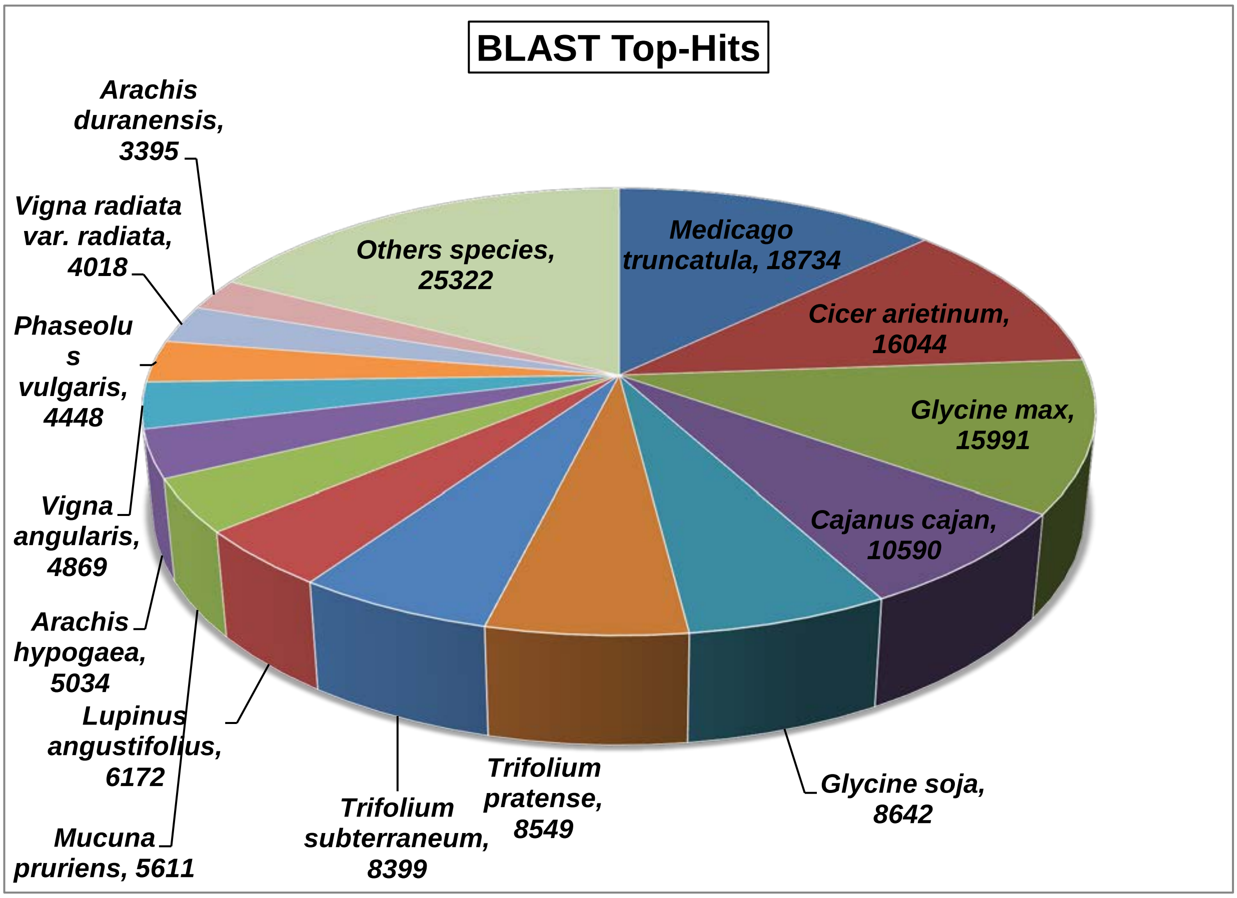


Figure 2: Biological processes gene ontology of *Taverniera cuneifolia* transcripts.

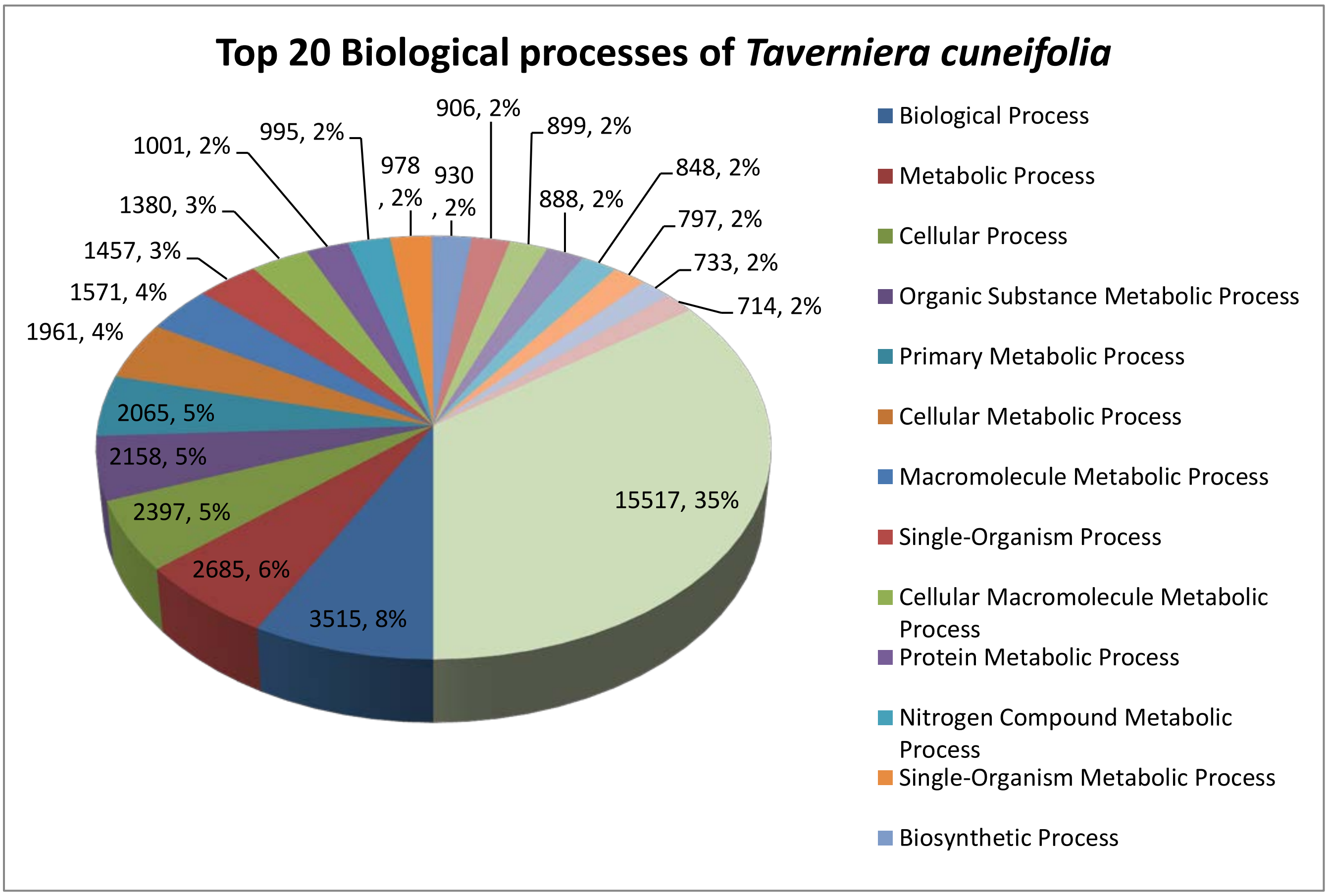


Figure 3: Molecular functions gene ontology of *Taverniera cuneifolia* transcripts.

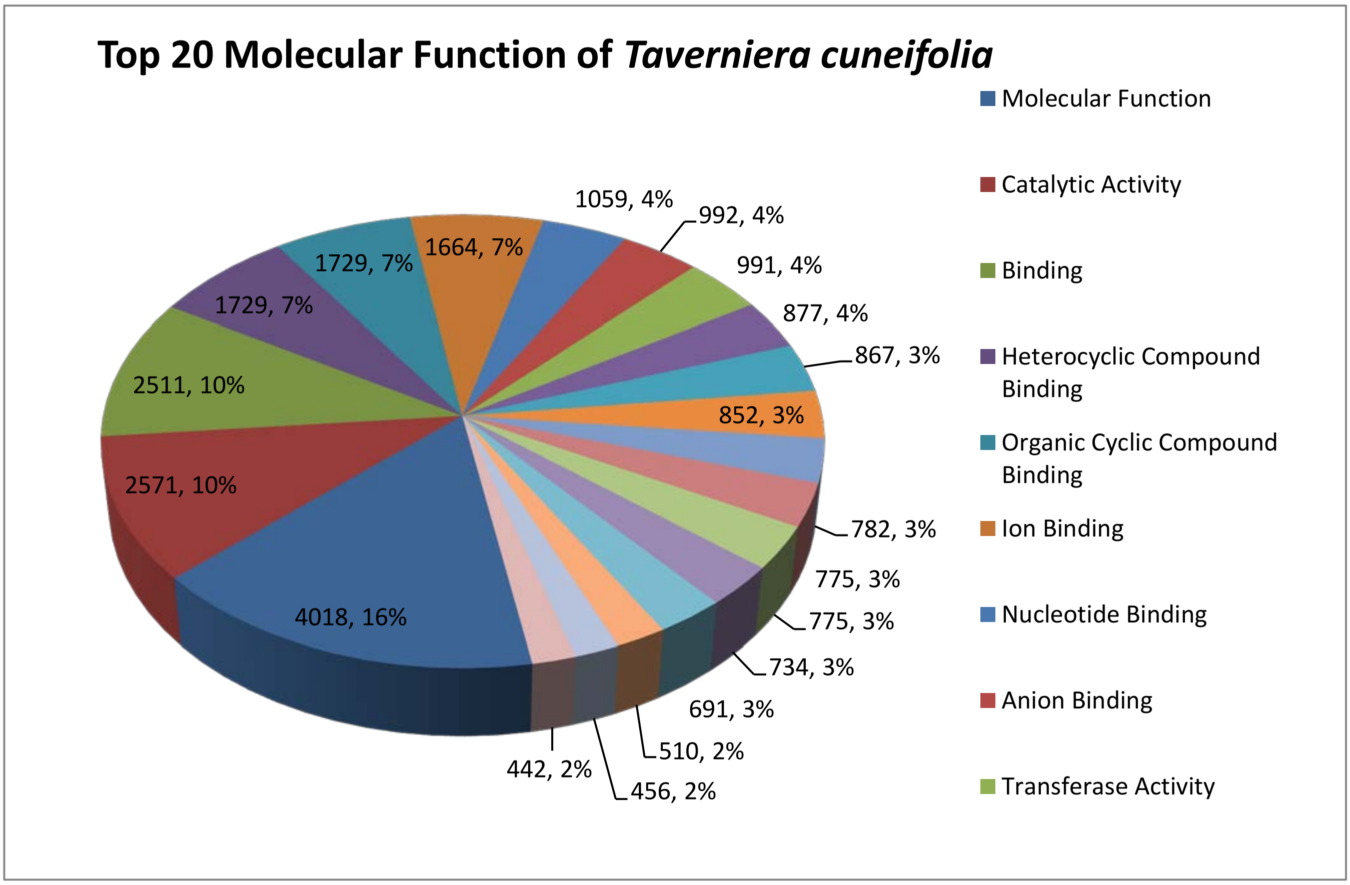
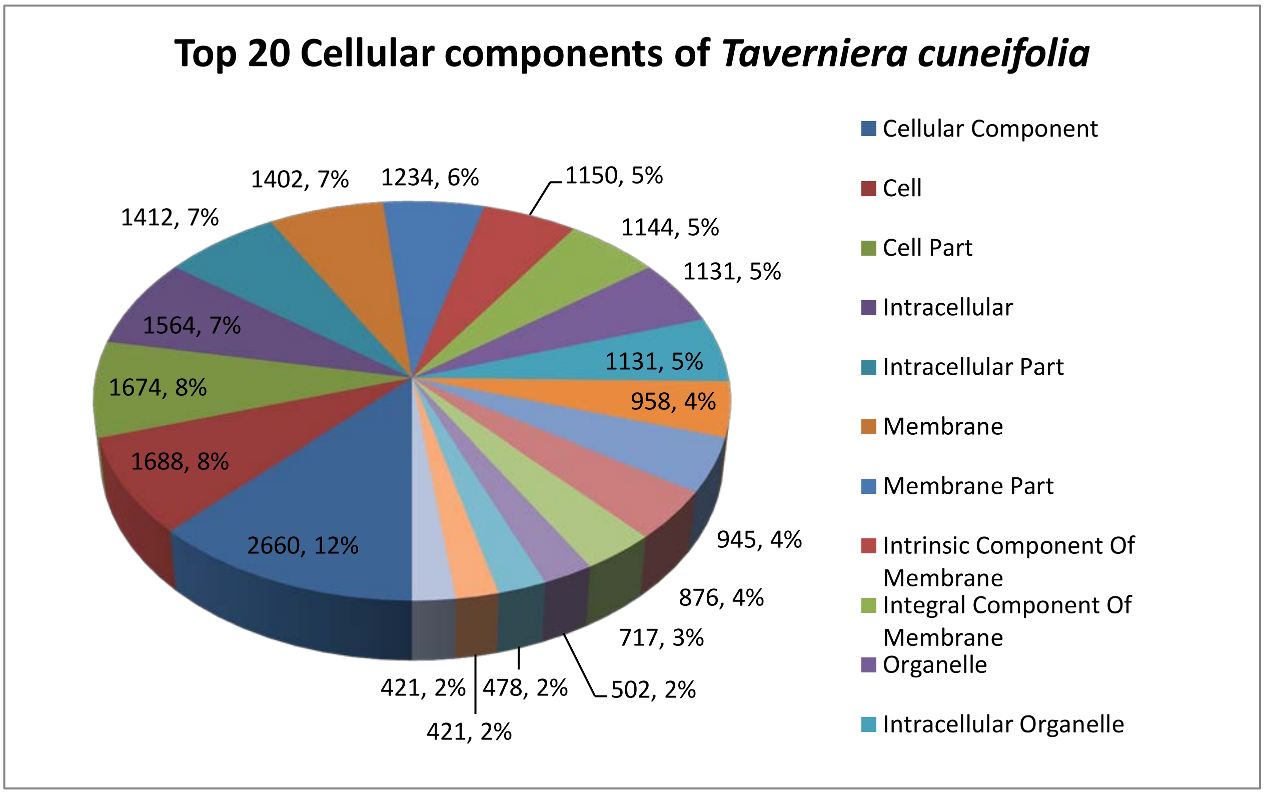


Figure 4: Cellular components gene ontology of *Taverniera cuneifolia* transcripts.



43 Figure 5: Enzyme classification of *Taverniera cuneifolia* transcripts based on KEGG pathway.

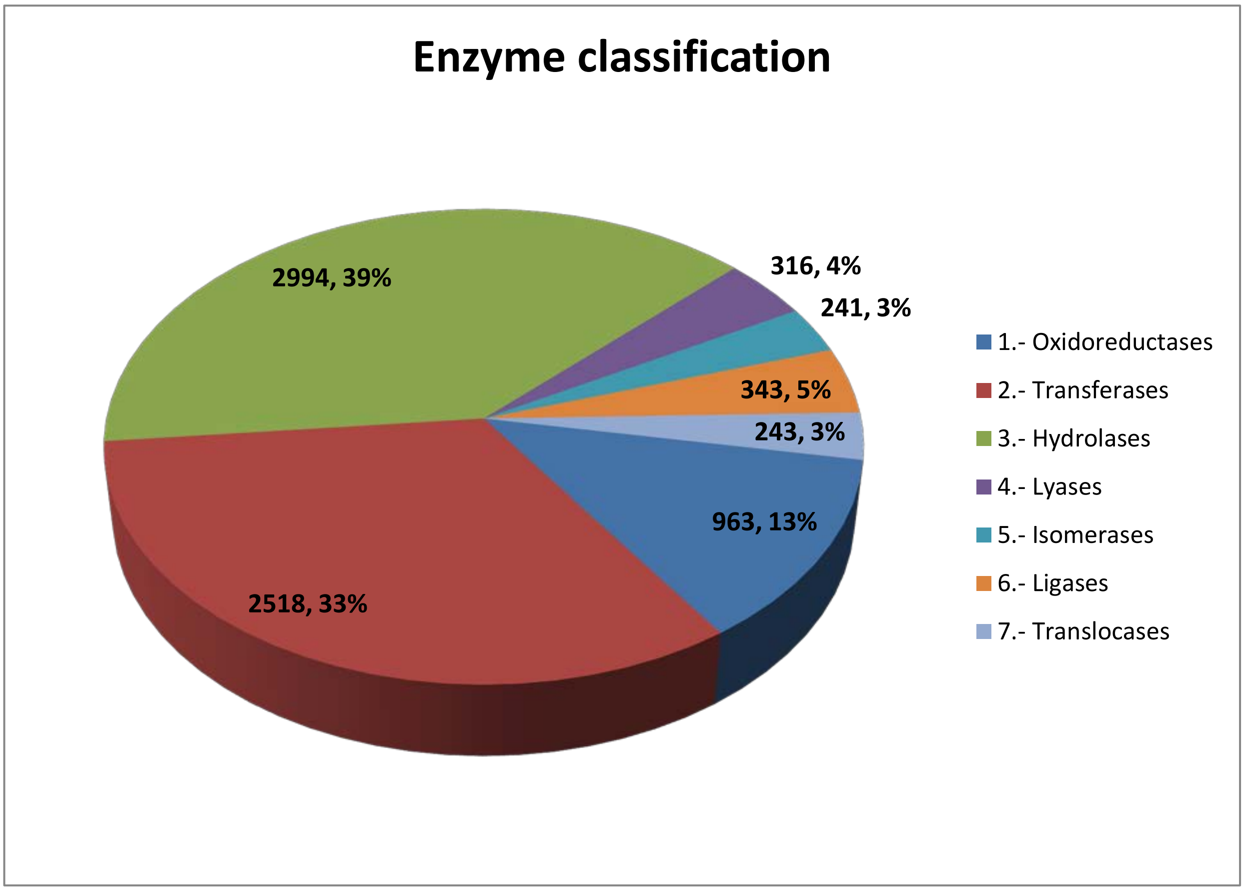
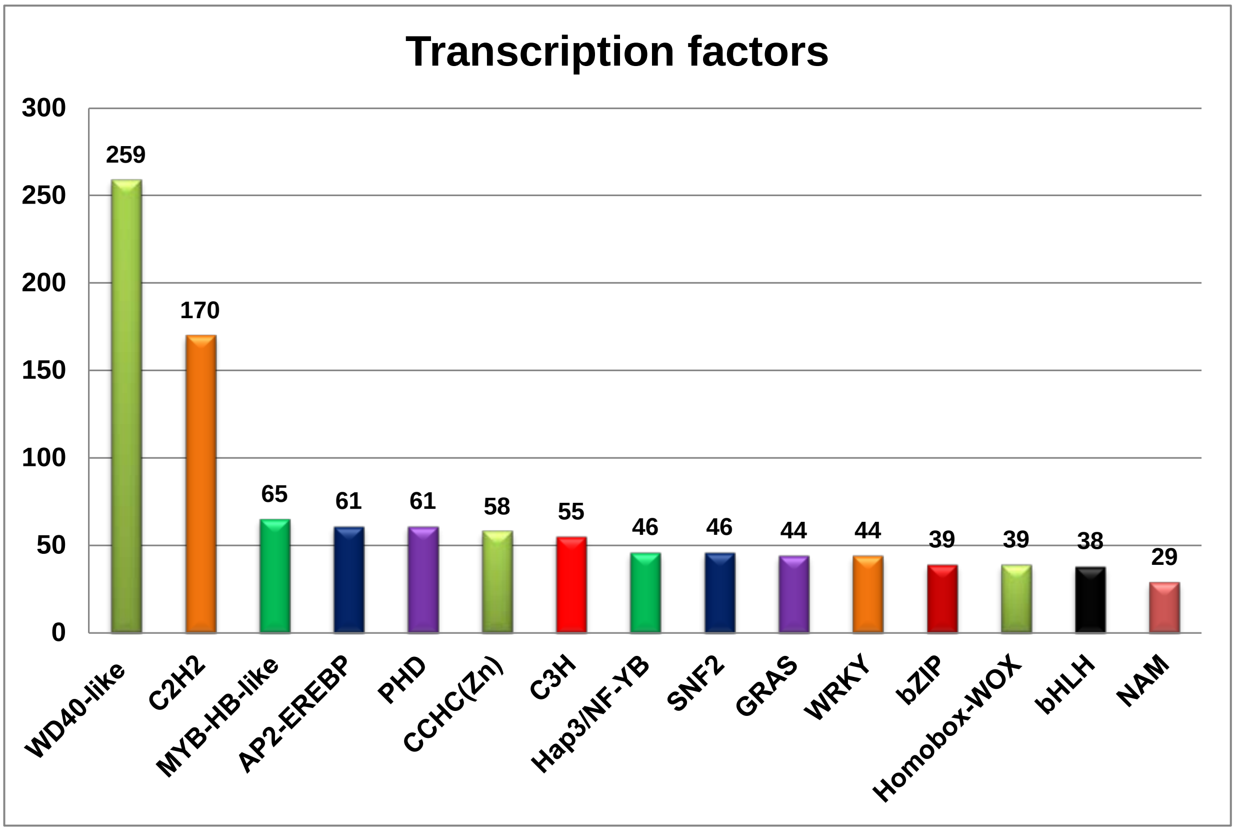


Figure 6: Top 15 Transcription factors families detection from *Taverniera cuneifolia* root transcriptome.



Supplementary Tables:

Table S1: GO sequence distribution of biological processes, molecular functions and cellular components.

GO term	Process	No. of Transcripts
Biological processes (44,395)	Biological process	**3515**
	Metabolic process	**2685**
	Cellular process	**2397**
	Organic substance metabolic process	**2158**
	Primary metabolic process	**2065**
	Cellular metabolic process	**1961**
	Macromolecule metabolic process	**1571**
	Single-organism process	**1457**
	Cellular macromolecule metabolic process	**1380**
	Protein metabolic process	**1001**
	Nitrogen compound metabolic process	**995**
	Single-organism metabolic process	**978**
	Biosynthetic process	**930**
	Organic substance biosynthetic process	**906**
	Cellular nitrogen compound metabolic process	**899**
	Cellular biosynthetic process	**888**
	Cellular protein metabolic process	**848**
	Single-organism cellular process	**797**
	Organic cyclic compound metabolic process	**733**
	Heterocycle metabolic process	**714**
	Cellular aromatic compound metabolic process	**712**
	Macromolecule biosynthetic process	**673**
	Nucleobase-containing compound metabolic process	**659**
	Cellular macromolecule biosynthetic process	**650**
	Cellular nitrogen compound biosynthetic process	**637**
	Phosphorus metabolic process	**632**
	Gene expression	**631**
	Phosphate-containing compound metabolic process	**629**
	Biological regulation	**614**
	Macromolecule modification	**589**
	Protein modification process	**576**
	Cellular protein modification process	**576**
	Regulation of biological process	**555**
	Localization	**533**
	Nucleic acid metabolic process	**531**
	Establishment of localization	**530**
	Transport	**529**
	Regulation of cellular process	**513**
	Organonitrogen compound metabolic process	**493**

	Oxidation-reduction process	**483**
	Phosphorylation	**478**
	RNA metabolic process	**459**
	Organic cyclic compound biosynthetic process	**437**
	Response to stimulus	**433**
	Heterocycle biosynthetic process	**421**
	Aromatic compound biosynthetic process	**416**
	Small molecule metabolic process	**386**
	Nucleobase-containing compound biosynthetic process	**383**
	Organonitrogen compound biosynthetic process	**359**
Cellular components (21,508)	Cellular component	**2660**
	Cell	**1688**
	Cell part	**1674**
	Intracellular	**1564**
	Intracellular part	**1412**
	Membrane	**1402**
	Membrane part	**1234**
	Intrinsic component of membrane	**1150**
	Integral component of membrane	**1144**
	Organelle	**1131**
	Intracellular organelle	**1131**
	Membrane-bounded organelle	**958**
	Intracellular membrane-bounded organelle	**945**
	Cytoplasm	**876**
	Cytoplasmic part	**717**
	Macromolecular complex	**502**
	Nucleus	**478**
	Organelle part	**421**
	Intracellular organelle part	**421**
Molecular functions (25,025)	Molecular function	**4018**
	Catalytic activity	**2571**
	Binding	**2511**
	Heterocyclic compound binding	**1729**
	Organic cyclic compound binding	**1729**
	Ion binding	**1664**
	Nucleotide binding	**1059**
	Transferase activity	**991**
	Ribonucleoside binding	**877**
	Purine ribonucleoside binding	**867**
	Hydrolase activity	**852**
	Adenyl ribonucleotide binding	**775**
	Oxidoreductase activity	**456**
	Anion binding	**992**
	Cation binding	**782**

	ATP binding	**734**
	Nucleic acid binding	**691**
	Transferase activity, transferring phosphorus-containing groups	**510**
	Metal ion binding	**775**
	Kinase activity	**442**

Table S2: Distribution of transcripts to biological pathways using KEGG specific to plants along with KO-ID.

KO-ID	**KEGGS Pathways Distribution**	**Transcripts no.**
ko01100	Metabolic pathways	102
ko01110	Biosynthesis of secondary metabolites	55
ko01120	Microbial metabolism in diverse environments	22
ko04075	Plant hormone signal transduction	15
ko03040	Spliceosome	14
ko01200	Carbon metabolism	13
ko03010	Ribosome	13
ko04144	Endocytosis	13
ko03013	RNA transport	12
ko04714	Thermogenesis	12
ko04626	Plant-pathogen interaction	12
ko04016	MAPK signaling pathway - plant	11
ko04120	Ubiquitin mediated proteolysis	10
ko04141	Protein processing in endoplasmic reticulum	10
ko01230	Biosynthesis of amino acids	10
ko00520	Amino sugar and nucleotide sugar metabolism	10
ko03018	RNA degradation	10
ko00190	Oxidative phosphorylation	9
ko03015	mRNA surveillance pathway	8
ko00500	Starch and sucrose metabolism	7
ko01240	Biosynthesis of cofactors	7
ko04146	Peroxisome	7
ko00620	Pyruvate metabolism	7
ko00010	Glycolysis / Gluconeogenesis	6
ko03050	Proteasome	5
ko00052	Galactose metabolism	5
ko04810	Regulation of actin cytoskeleton	5
ko00051	Fructose and mannose metabolism	5
ko03008	Ribosome biogenesis in eukaryotes	5
ko00270	Cysteine and methionine metabolism	5
ko00240	Pyrimidine metabolism	5

ko04142	Lysosome	5
ko00920	Sulfur metabolism	4
ko00410	beta-Alanine metabolism	4
ko03420	Nucleotide excision repair	4
ko00230	Purine metabolism	4
ko00983	Drug metabolism - other enzymes	4
ko04110	Cell cycle	4
ko00020	Citrate cycle	4
ko00970	Aminoacyl-tRNA biosynthesis	4
ko04072	Phospholipase D signaling pathway	4
ko00250	Alanine, aspartate and glutamate metabolism	4
ko00720	Carbon fixation pathways in prokaryotes	4
ko00940	Phenylpropanoid biosynthesis	4
ko00564	Glycerophospholipid metabolism	4
ko00710	Carbon fixation in photosynthetic organisms	4
ko00480	Glutathione metabolism	4
ko01212	Fatty acid metabolism	4
ko00906	Carotenoid biosynthesis	3
ko00030	Pentose phosphate pathway	3
ko00071	Fatty acid degradation	3
ko00280	Valine, leucine and isoleucine degradation	3
ko00330	Arginine and proline metabolism	3
ko04712	Circadian rhythm - plant	3
ko00040	Pentose and glucuronate interconversions	3
ko01210	2-Oxocarboxylic acid metabolism	3
ko03020	RNA polymerase	3
ko00909	Sesquiterpenoid and triterpenoid biosynthesis	3
ko00900	Terpenoid backbone biosynthesis	3
ko00460	Cyanoamino acid metabolism	2
ko03410	Base excision repair	2
ko00902	Monoterpenoid biosynthesis	2
ko00340	Histidine metabolism	2
ko00908	Zeatin biosynthesis	2
ko00980	Metabolism of xenobiotics by cytochrome P450	2
ko03440	Homologous recombination	2
ko00061	Fatty acid biosynthesis	2
ko00630	Glyoxylate and dicarboxylate metabolism	2
ko02010	ABC transporters	2
ko01040	Biosynthesis of unsaturated fatty acids	2
ko04978	Mineral absorption	2
ko04024	cAMP signaling pathway	2

ko00592	alpha-Linolenic acid metabolism	2
ko00290	Valine, leucine and isoleucine biosynthesis	2
ko00260	Glycine, serine and threonine metabolism	2
ko00625	Chloroalkane and chloroalkene degradation	2
ko00510	N-Glycan biosynthesis	2
ko04310	Wnt signaling pathway	2
ko04020	Calcium signaling pathway	2
ko00310	Lysine degradation	2
ko03060	Protein export	2
ko04922	Glucagon signaling pathway	2
ko00640	Propanoate metabolism	2
ko00941	Flavonoid biosynthesis	1
ko03430	Mismatch repair	1
ko00220	Arginine biosynthesis	1
ko02020	Two-component system	1
ko00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	1
ko00513	Various types of N-glycan biosynthesis	1
ko00903	Limonene and pinene degradation	1
ko00966	Glucosinolate biosynthesis	1
ko00730	Thiamine metabolism	1
ko00400	Phenylalanine, tyrosine and tryptophan biosynthesis	1
ko04122	Sulfur relay system	1
ko00100	Steroid biosynthesis	1
ko00380	Tryptophan metabolism	1
ko00982	Drug metabolism - cytochrome P450	1
ko00790	Folate biosynthesis	1
ko03030	DNA replication	1
ko04979	Cholesterol metabolism	1
ko00904	Diterpenoid biosynthesis	1
ko03022	Basal transcription factors	1
ko00130	Ubiquinone and other terpenoid-quinone biosynthesis	1

54 Table S3: Transcripts/genes that is associated with *Glycyrrhizin* production in *Taverniera*
55 *cuneifolia* from Nr database.

Sr. No.	**Transcript ID**	**Best hit Transcripts associated with Glycyrrhizin biosynthesis pathway from Nr database**
GENE 1 Squalene synthase/ epoxidase/monooxygenase		
1	TRINITY_DN11206_c0_g1_i1	ADG36709.1| squalene synthase 1
2	TRINITY_DN11206_c0_g1_i2	ADG36709.1| squalene synthase 1
3	TRINITY_DN25523_c0_g1_i1	XP_007041440.1 squalene monooxygenase
4	TRINITY_DN7116_c0_g1_i1	ADG36706.1squalene synthase 1
5	TRINITY_DN9998_c0_g1_i1	ADG36711.1squalene synthase 1
6	TRINITY_DN9998_c0_g1_i2	CAJ77652.1squalene synthase 1
7	TRINITY_DN9998_c0_g1_i3	ADG36699.1squalene synthase 1
8	TRINITY_DN5897_c0_g1_i1	AHY94896.1squalene epoxidase
9	TRINITY_DN10863_c0_g1_i1	AKO83630.1squalene epoxidase
10	TRINITY_DN14273_c0_g1_i2	AHY94896.1squalene epoxidase
11	TRINITY_DN14273_c1_g1_i1	KEH39980.1squalene monooxygenase
12	TRINITY_DN3414_c0_g1_i1	APA19297.1squalene synthase
13	TRINITY_DN10934_c0_g1_i2	XP_004498941.1 squalene monooxygenase-like
14	TRINITY_DN10934_c0_g1_i3	AKO83630.1squalene epoxidase
15	TRINITY_DN10934_c0_g1_i4	AKO83630.1squalene epoxidase
GENE 2 Beta-amyrin synthase		
1	TRINITY_DN11239_c0_g1_i0031	AAO33578.1beta-amyrin synthase
2	TRINITY_DN11239_c0_g1_i2	NP_001236591.2beta-amyrin synthase
3	TRINITY_DN28103_c0_g1_i1	XP_018838319.1 beta-amyrin synthase
4	TRINITY_DN11371_c0_g1_i1	NP_001236591.2 beta-amyrin synthase
5	TRINITY_DN11371_c0_g1_i2	AHI17180.1 beta-amyrin synthase
GENE 3 Beta-amyrin 11-oxidase /CYP88D6		
1	TRINITY_DN20252_c0_g1_i1	B5BSX1.1 Full=Beta-amyrin 11-oxidase; AltName: Full=Cytochrome P450 88D6
2	TRINITY_DN11652_c0_g1_i3	B5BSX1.1 Full=Beta-amyrin 11-oxidase; AltName: Full=Cytochrome P450 88D6
3	TRINITY_DN11652_c0_g1_i4	AQQ13664.1 beta-amyrin 11-oxidase
4	TRINITY_DN11652_c0_g1_i6	XP_004510262.1 beta-amyrin 11-oxidase-like
GENE 4 11-oxo-beta-amyrin 30-oxidase/CYP72A154		
1	TRINITY_DN5998_c0_g1_i1	XP_004488667.1 11-oxo-beta-amyrin 30-oxidase-like
2	TRINITY_DN11613_c0_g1_i1	H1A988.1 Full=11-oxo-beta-amyrin 30-oxidase; AltName: Full=Cytochrome P450 72A154
3	TRINITY_DN11613_c0_g1_i2	XP_004511068.1 11-oxo-beta-amyrin 30-oxidase-like
4	TRINITY_DN11613_c0_g1_i7	XP_004511068.1 11-oxo-beta-amyrin 30-oxidase-like
5	TRINITY_DN11613_c0_g1_i9	XP_004511068.1 11-oxo-beta-amyrin 30-oxidase-like
6	TRINITY_DN9161_c0_g1_i1	RHN74756.1putative 11-oxo-beta-amyrin 30-oxidase
7	TRINITY_DN10411_c0_g1_i2	XP_004511068.1 11-oxo-beta-amyrin 30-oxidase-like
8	TRINITY_DN10411_c0_g1_i3	XP_004511068.1 11-oxo-beta-amyrin 30-oxidase-like
9	TRINITY_DN10411_c0_g1_i4	XP_004511068.1 11-oxo-beta-amyrin 30-oxidase-like
10	TRINITY_DN5730_c0_g1_i1	XP_004488667.1 11-oxo-beta-amyrin 30-oxidase-like
GENE 5 Beta-amyrin 24-hydroxylase /CYP93E7		
1	TRINITY_DN11492_c0_g1_i1	AIN25419.1beta-amyrin 24-hydroxylase CYP93E7
GENE 6 UDP-glycosyltransferase family protein		
1	TRINITY_DN25514_c0_g1_i1	RHN51110.1putative UDP-glucuronosyl/UDP-glucosyltransferase
2	TRINITY_DN11708_c3_g1_i1	XP_022880903.1UDP-glycosyltransferase 73B5-like
3	TRINITY_DN1272_c0_g1_i1	AMQ26133.1UDP-glycosyltransferase 3

4	TRINITY_DN14507_c0_g1_i1	KEH43353.1 UDP-glycosyltransferase family protein
5	TRINITY_DN7469_c0_g1_i1	XP_013451680.1UDP-glycosyltransferase 1
6	TRINITY_DN30090_c0_g1_i1	XP_019428832.1 UDP-glycosyltransferase 73C6-like
7	TRINITY_DN17733_c0_g1_i1	XP_003600815.1UDP-glycosyltransferase 76B1 isoform X1
8	TRINITY_DN18323_c0_g1_i1	XP_012568016.1 UDP-glucose:glycoprotein glucosyltransferase
9	TRINITY_DN1735_c0_g1_i1	XP_004489724.1 UDP-glycosyltransferase 74E1
10	TRINITY_DN1735_c0_g1_i2	XP_004489724.1 UDP-glycosyltransferase 74E1
11	TRINITY_DN19316_c0_g1_i1	XP_020228882.1UDP-glycosyltransferase 87A1-like
12	TRINITY_DN10035_c0_g1_i1	RDX79205.1UDP-glycosyltransferase 71K2
13	TRINITY_DN6951_c0_g1_i1	XP_004490590.1 UDP-glycosyltransferase 71D1-like
14	TRINITY_DN12074_c0_g1_i1	KEH43353.1 UDP-glycosyltransferase family protein
15	TRINITY_DN27452_c0_g1_i1	AES66918.2UDP-glucosyltransferase family protein
16	TRINITY_DN16244_c0_g1_i1	XP_012568016.1 UDP-glucose:glycoprotein glucosyltransferase
17	TRINITY_DN21948_c0_g1_i1	PNY11551.1UDP-glycosyltransferase-like protein
18	TRINITY_DN29071_c0_g1_i1	RDX76823.1UDP-glycosyltransferase 72B1
19	TRINITY_DN14490_c0_g1_i1	XP_014489827.1UDP-glycosyltransferase 87A1
20	TRINITY_DN14236_c0_g1_i1	XP_012568460.1 UDP-glycosyltransferase 87A1-like
21	TRINITY_DN14292_c0_g1_i1	PNY09424.1UDP-glycosyltransferase 76F1-like protein
22	TRINITY_DN23324_c0_g1_i1	PNY15296.1UDP-glycosyltransferase 87A1-like protein
23	TRINITY_DN28174_c0_g1_i1	XP_013451922.1UDP-glycosyltransferase 74G1
24	TRINITY_DN29171_c0_g1_i1	XP_004490654.1 UDP-glycosyltransferase 87A1
25	TRINITY_DN16300_c0_g1_i1	KHN41573.1UDP-glycosyltransferase 83A1
26	TRINITY_DN14033_c0_g1_i1	XP_003544901.1UDP-glycosyltransferase 87A1
27	TRINITY_DN22391_c0_g1_i1	XP_003599976.1putative UDP-glucose glucosyltransferase
28	TRINITY_DN9589_c0_g1_i1	XP_013451680.1UDP-glycosyltransferase 1
29	TRINITY_DN9589_c1_g1_i1	XP_014515686.1UDP-glycosyltransferase 1-like
30	TRINITY_DN20336_c0_g1_i1	XP_004503216.1 UDP-glycosyltransferase 76F1-like isoform X1
31	TRINITY_DN17416_c0_g1_i1	XP_012568016.1 UDP-glucose:glycoprotein glucosyltransferase
32	TRINITY_DN29577_c0_g1_i1	XP_020240106.1UDP-glycosyltransferase 71K1 isoform X1

Table S4: Transcripts/genes that showed the Hypothetical protein in *Taverniera cuneifolia* with hit length above 400, (total over all 4912 hypothetical protien).

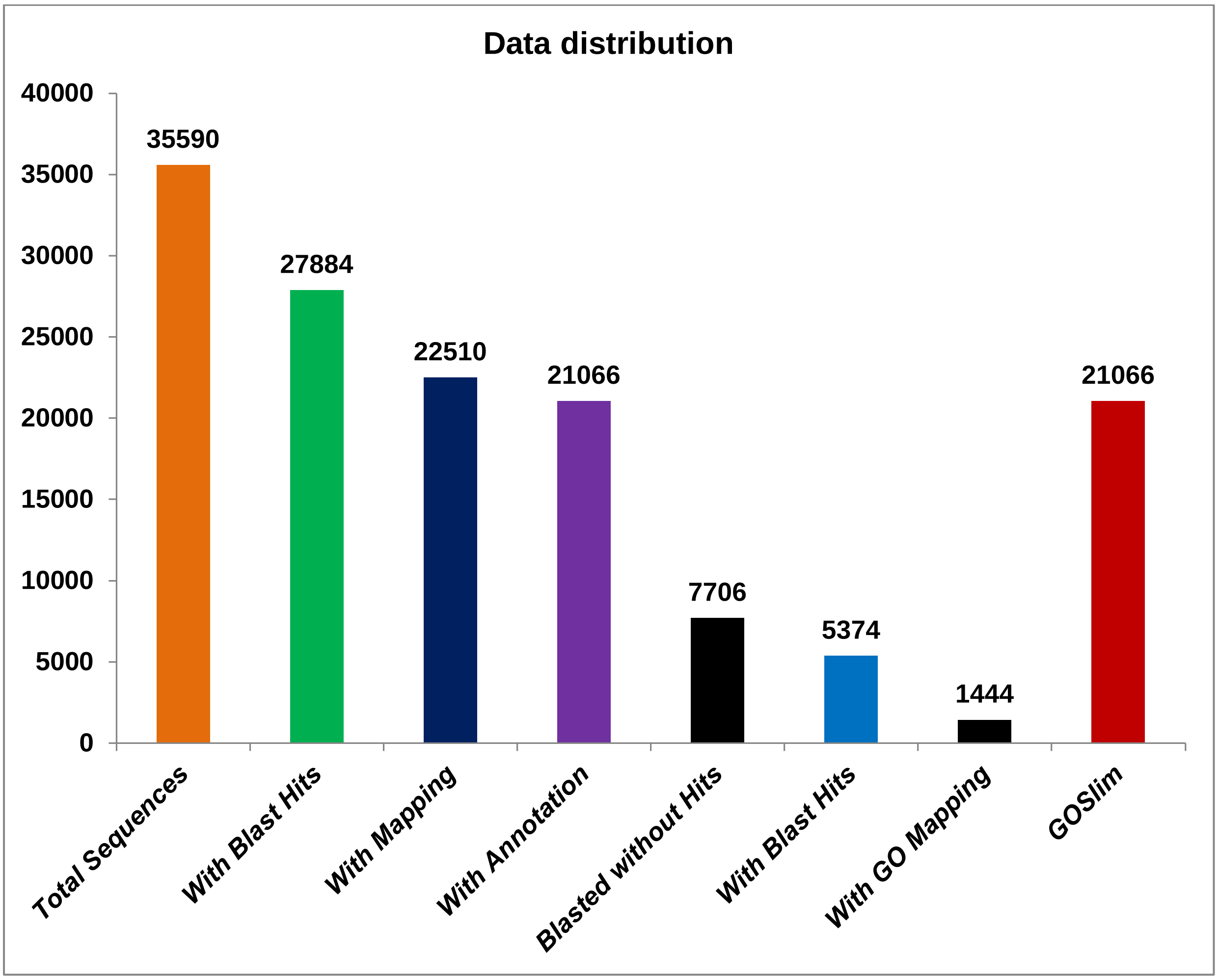
Sr. No.	Transcript ID	Transcripts associated with Glycyrrhizin biosynthesis pathway
1	TRINITY_DN11292_c0_g1_i3	gi|593801532|ref|XP_007163803.1| hypothetical protein PHAVU_001G265500g
2	TRINITY_DN11286_c0_g1_i4	gi|593795660|ref|XP_007160868.1| hypothetical protein PHAVU_001G023500g
3	TRINITY_DN10387_c0_g1_i1	gi|965601928|dbj|BAT89106.1| hypothetical protein VIGAN_05279900
4	TRINITY_DN10679_c0_g1_i1	gi|920709256|gb|KOM51253.1| hypothetical protein LR48_Vigan08g208000
5	TRINITY_DN11770_c6_g1_i1	gi|920715088|gb|KOM55176.1| hypothetical protein LR48_Vigan10g106800
6	TRINITY_DN11705_c1_g1_i3	gi|593797882|ref|XP_007161979.1| hypothetical protein PHAVU_001G113800g
7	TRINITY_DN11740_c0_g1_i2	gi|147782060|emb|CAN61004.1| hypothetical protein VITISV_015023
8	TRINITY_DN6658_c0_g1_i1	gi|920703423|gb|KOM46648.1| hypothetical protein LR48_Vigan07g035200
9	TRINITY_DN6650_c0_g1_i1	gi|965663984|dbj|BAT79693.1| hypothetical protein VIGAN_02261400
10	TRINITY_DN11563_c0_g1_i4	gi|593701389|ref|XP_007151112.1| hypothetical protein PHAVU_004G018900g
11	TRINITY_DN11531_c0_g1_i1	gi|593700643|ref|XP_007150760.1| hypothetical protein PHAVU_005G178600g
12	TRINITY_DN11540_c0_g1_i2	gi|147781743|emb|CAN61179.1| hypothetical protein VITISV_032292
13	TRINITY_DN11053_c1_g1_i2	gi|357441957|ref|XP_003591256.1| hypothetical protein MTR_1g084990
14	TRINITY_DN11043_c0_g1_i2	gi|763758066|gb|KJB25397.1| hypothetical protein B456_004G189700
15	TRINITY_DN11043_c0_g1_i4	gi|763758066|gb|KJB25397.1| hypothetical protein B456_004G189700
16	TRINITY_DN11647_c0_g1_i2	gi|965604026|dbj|BAT91203.1| hypothetical protein VIGAN_06251600
17	TRINITY_DN11647_c0_g1_i5	gi|965604026|dbj|BAT91203.1| hypothetical protein VIGAN_06251600
18	TRINITY_DN11626_c1_g1_i1	gi|593612647|ref|XP_007142864.1| hypothetical protein PHAVU_007G023200g
19	TRINITY_DN11665_c3_g1_i2	gi|947109915|gb|KRH58241.1| hypothetical protein GLYMA_05G114900
20	TRINITY_DN11468_c0_g1_i2	gi|593799252|ref|XP_007162664.1| hypothetical protein PHAVU_001G169900g
21	TRINITY_DN11472_c0_g2_i1	gi|920703664|gb|KOM46889.1| hypothetical protein LR48_Vigan07g059300
22	TRINITY_DN11487_c0_g1_i3	gi|947099253|gb|KRH47745.1| hypothetical protein GLYMA_07G047800
23	TRINITY_DN11430_c1_g2_i1	gi|593704437|ref|XP_007152592.1| hypothetical protein PHAVU_004G142900g

24	TRINITY_DN11430_c1_g2_i4	gi|593704437|ref|XP_007152592.1| hypothetical protein PHAVU_004G142900g
25	TRINITY_DN10796_c0_g1_i3	gi|965661959|dbj|BAT77668.1| hypothetical protein VIGAN_02025800
26	TRINITY_DN4344_c0_g1_i1	gi|593694898|ref|XP_007147954.1| hypothetical protein PHAVU_006G168300g
27	TRINITY_DN11312_c0_g1_i2	gi|922399741|ref|XP_013467009.1| hypothetical protein MTR_1g041275
28	TRINITY_DN11307_c0_g1_i4	gi|920679711|gb|KOM26600.1| hypothetical protein LR48_Vigan303s002200
29	TRINITY_DN10957_c0_g1_i3	gi|920681762|gb|KOM28542.1| hypothetical protein LR48_Vigan549s009700
30	TRINITY_DN11114_c0_g1_i2	gi|357466213|ref|XP_003603391.1| hypothetical protein MTR_3g107090

Table S5: Transcripts/genes that showed the Cytochrome P450 family protein in *Taverniera cuneifolia*

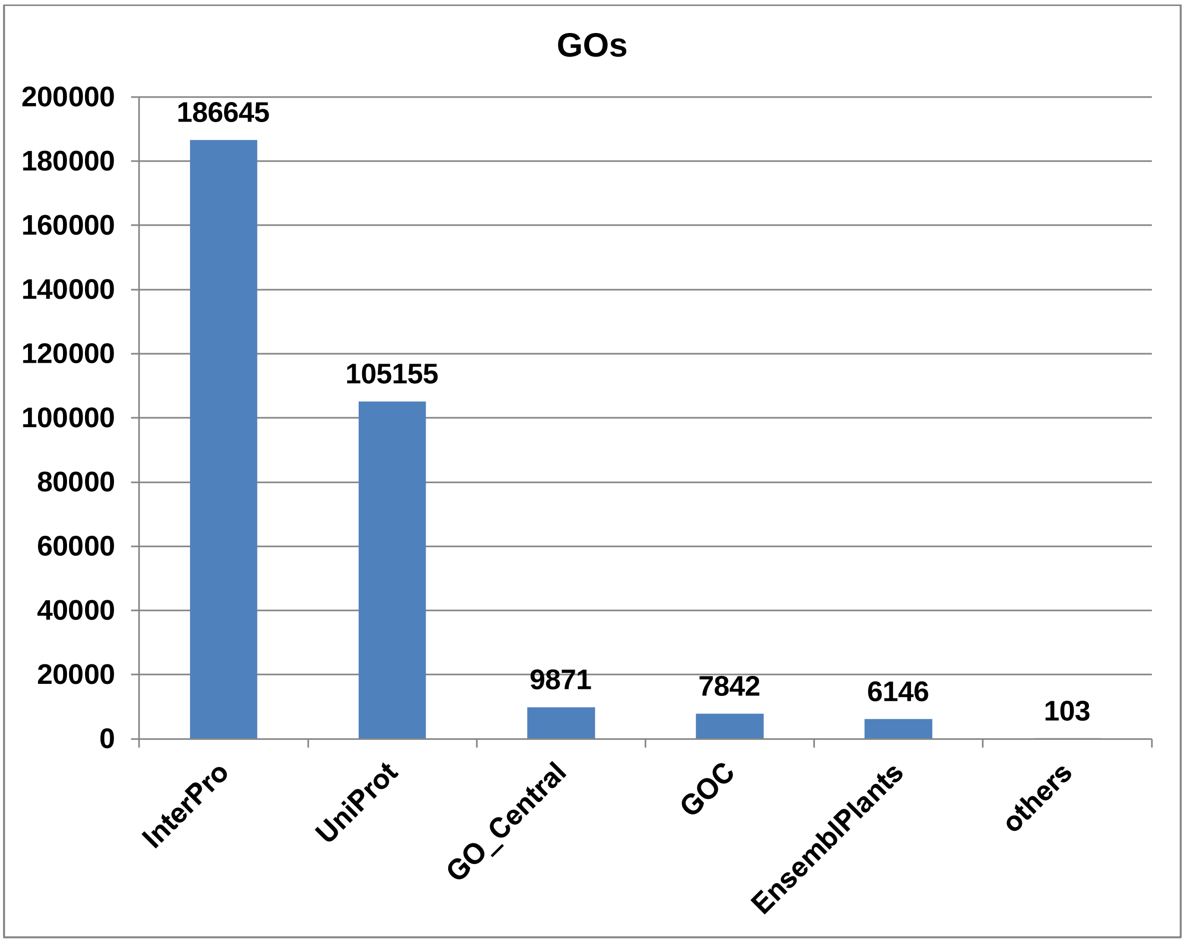
Sr. No.	Transcript ID	Transcripts associated with Cytochrome P450 family protein
1	TRINITY_DN10399_c0_g1_i2	gi|356515730|ref|XP_003526551.1| PREDICTED: NADPH--cytochrome P450 reductase
2	TRINITY_DN11569_c1_g1_i1	gi|357514033|ref|XP_003627305.1| cytochrome P450 family monooxygenase
3	TRINITY_DN11569_c1_g1_i3	gi|357514033|ref|XP_003627305.1| cytochrome P450 family monooxygenase
4	TRINITY_DN1922_c0_g1_i2	gi|502156756|ref|XP_004510631.1| PREDICTED: cytochrome P450 78A3
5	TRINITY_DN11093_c0_g1_i2	gi|922394449|ref|XP_013465628.1| cytochrome P450 family Ent-kaurenoic acid oxidase
6	TRINITY_DN11010_c1_g1_i1	gi|357470373|ref|XP_003605471.1| cytochrome P450 family monooxygenase
7	TRINITY_DN11652_c0_g1_i4	gi|838228579|gb|AKM97308.1| cytochrome P450 88D6
8	TRINITY_DN8146_c0_g1_i1	gi|371940464|dbj|BAL45206.1| cytochrome P450 monooxygenase
9	TRINITY_DN9931_c0_g1_i2	gi|502150242|ref|XP_004507858.1| PREDICTED: NADPH--cytochrome P450 reductase
10	TRINITY_DN9161_c0_g1_i1	gi|922392052|ref|XP_013464437.1| cytochrome P450 family protein
11	TRINITY_DN3071_c0_g1_i1	gi|922399435|ref|XP_013466867.1| cytochrome P450 family protein
12	TRINITY_DN5499_c0_g1_i1	gi|922394449|ref|XP_013465628.1| cytochrome P450 family Ent-kaurenoic acid oxidase
13	TRINITY_DN2780_c0_g1_i1	gi|502161259|ref|XP_004512097.1| PREDICTED: cytochrome P450 84A1
14	TRINITY_DN2767_c0_g1_i1	gi|356569428|ref|XP_003552903.1| PREDICTED: cytochrome P450 714C2-like
15	TRINITY_DN10453_c0_g1_i1	gi|356540462|ref|XP_003538708.1| PREDICTED: cytochrome P450 87A3-like
16	TRINITY_DN10993_c0_g1_i1	gi|922380457|ref|XP_013460458.1| cytochrome P450 family protein
17	TRINITY_DN11158_c1_g1_i4	gi|922327835|ref|XP_013443310.1| cytochrome P450 family 71 protein

Supplementary Figures:



Supplementary Fig. S1: Data distribution of *Taverniera cuneifolia* transcripts subjected to functional annotation.

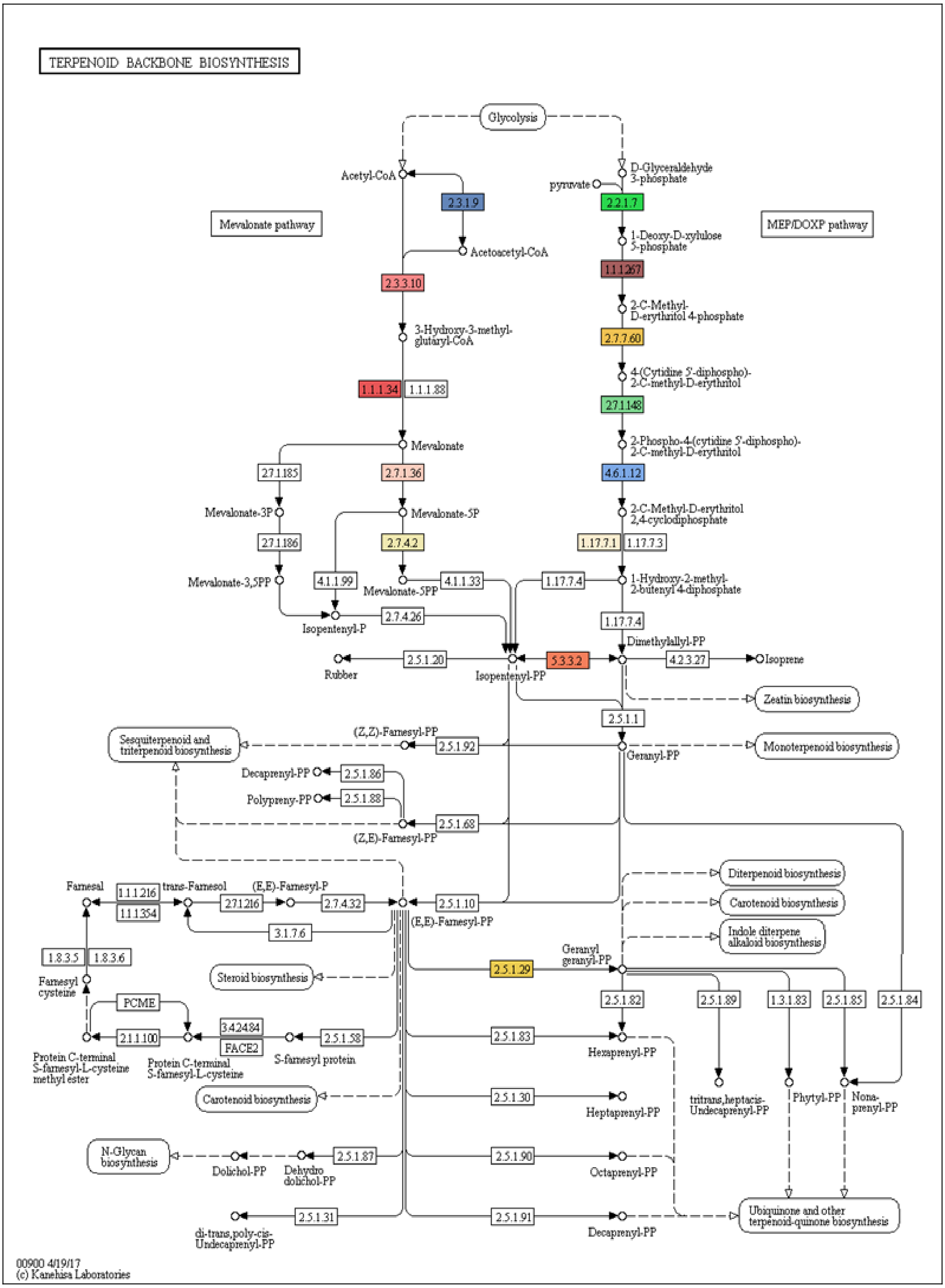
72



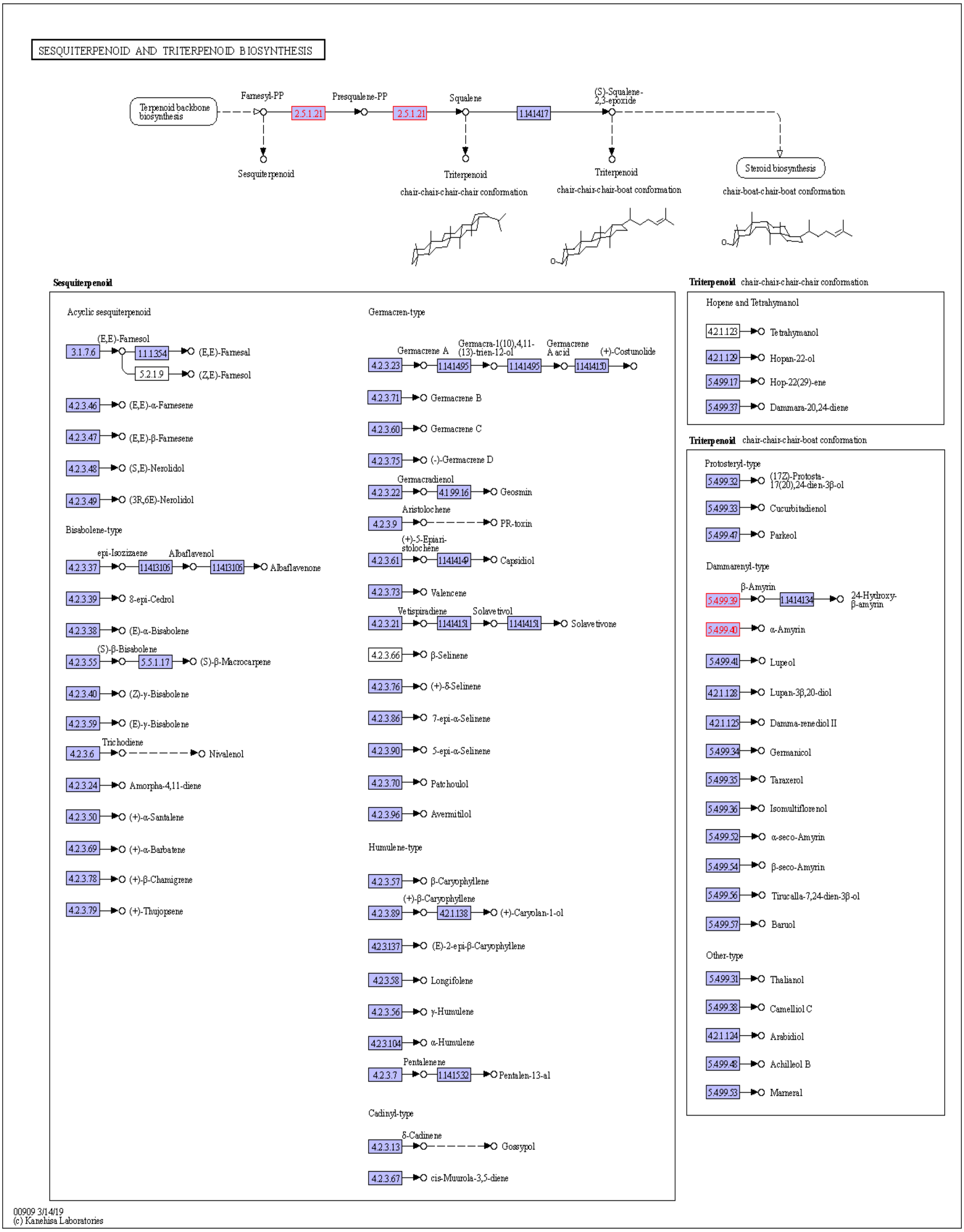
73

74 **Supplementary Fig. S2**: Annotation of *Taverniera cuneifolia* transcripts to different database sources.

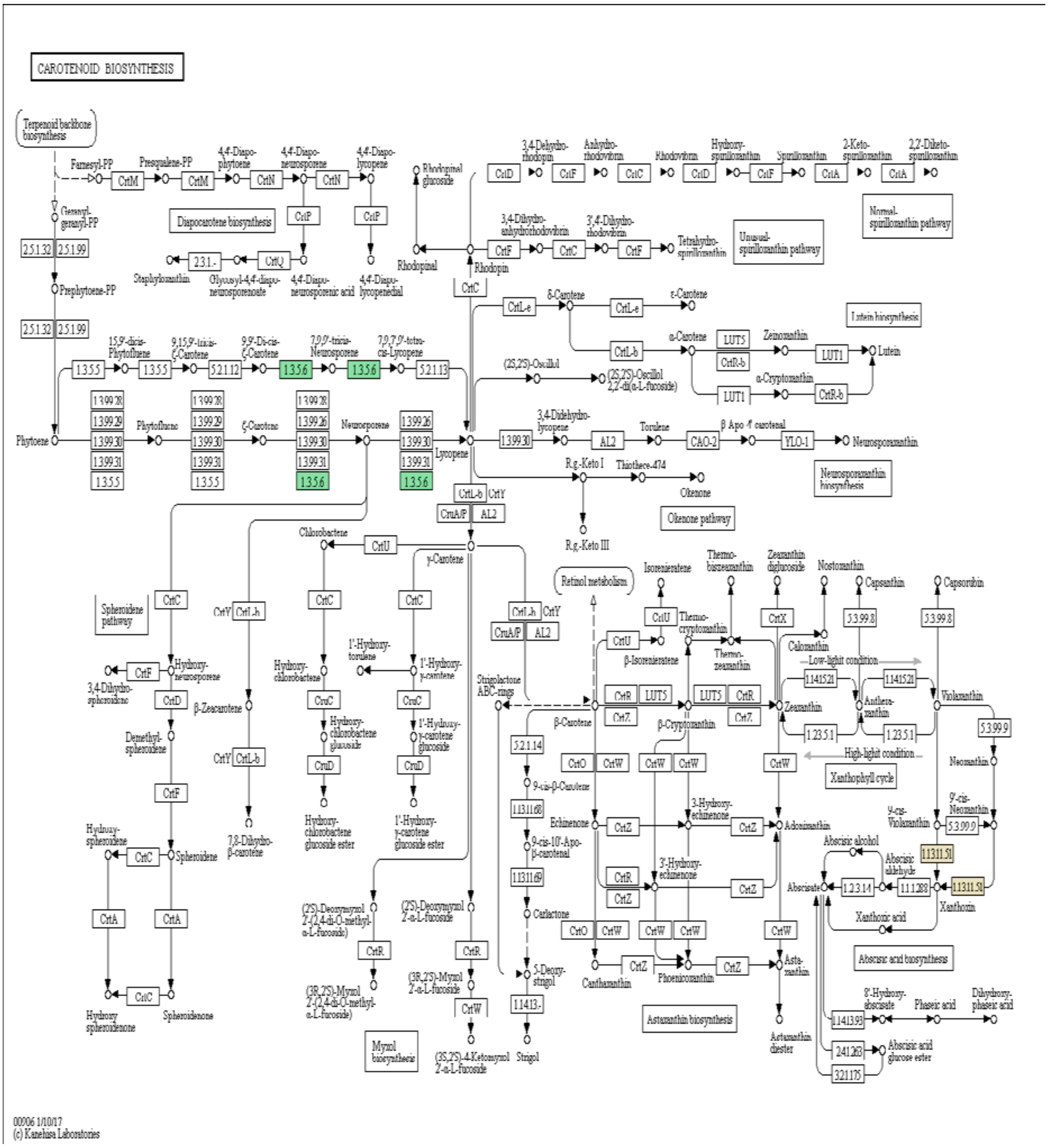
75



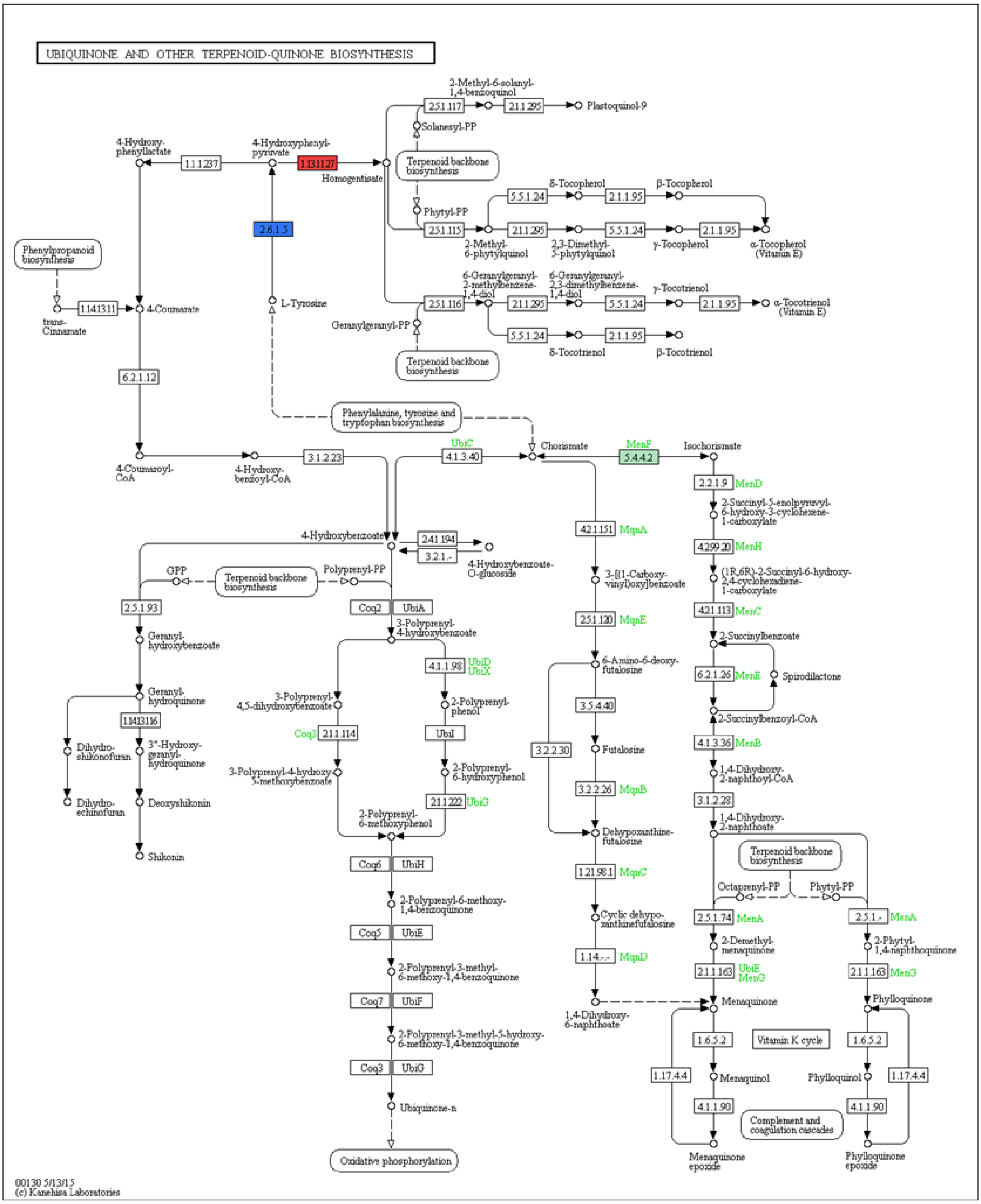
Supplementary Fig. S3: Terpenoids backbone biosynthesis pathways (Ko00900), color boxes are the gene found from *Taverniera cuneifolia* sequences.



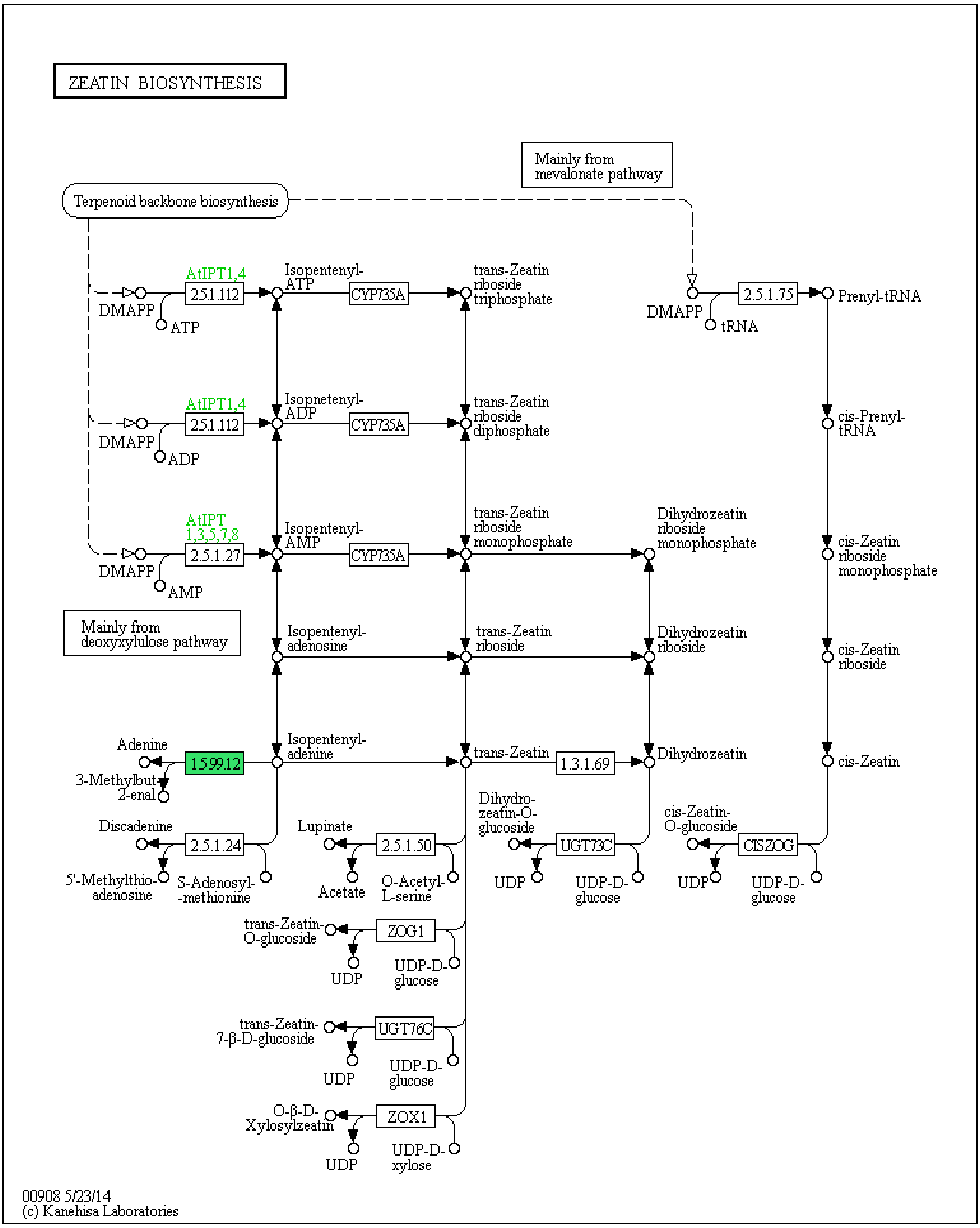
Supplementary Fig. S4: Sesquiterpenoid and triterpenoid biosynthesis pathway (ko00909) highlighted boxes are the gene found in *Taverniera cuneifolia* sequences.



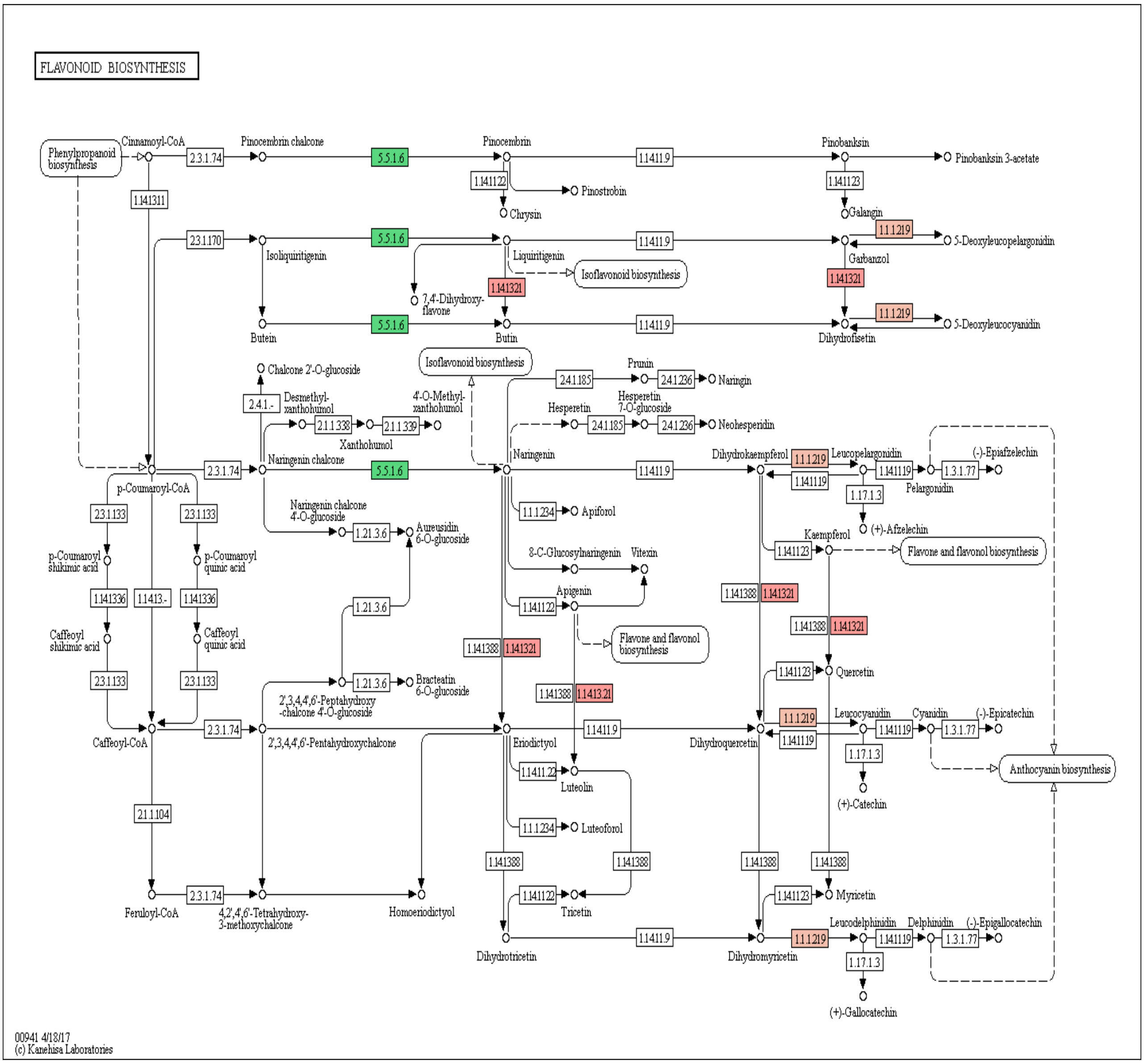
Supplementary Fig. S5: Carotenoid biosynthesis pathway (ko00906) , color boxes are the gene found in *Taverniera cuneifolia* sequences.



Supplementary Fig. S6: Ubiquinone and other terpenoid-quinone biosynthesis pathway (ko00130), color boxes are the gene found in *Taverniera cuneifolia* sequences.



Supplementary Fig. S7: Zeatin biosynthesis pathway (ko00908), color box are the gene found in *Taverniera cuneifolia* sequences.



Supplementary Fig. S8: Flavonoid biosynthesis pathway (ko00941), color boxes are the gene found in *Taverniera cuneifolia* sequences.

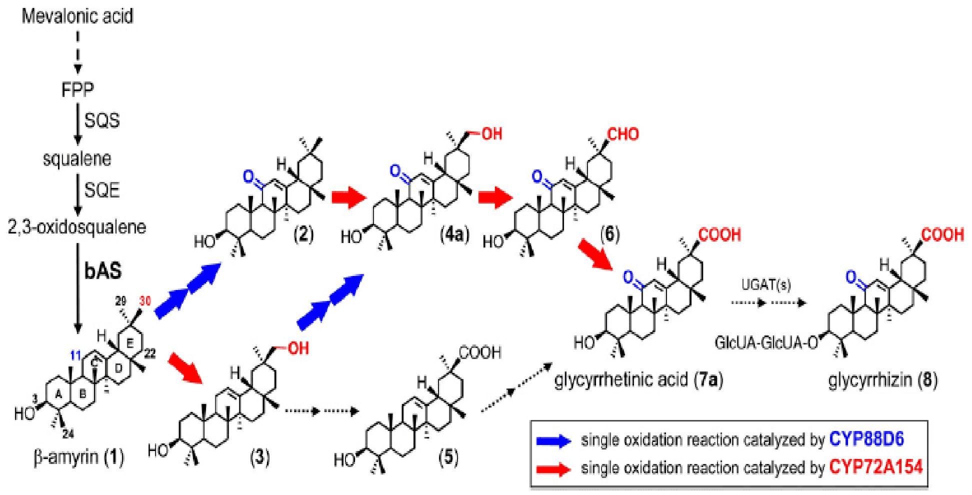


Figure 1. Proposed Pathway for Biosynthesis of Glycyrrhizin.

The structures of possible biosynthetic intermediates between β-amyrin (**1**) and glycyrrhizin (**8**) are shown: (**2**), 11-oxo-β-amyrin; (**3**), 30-hydroxy-β-amyrin; (**4a**), 30-hydroxy-11-oxo-β-amyrin; (**5**), 11-deoxoglycyrrhetinic acid; (**6**), glycyrrhetaldehyde; and (**7a**), glycyrrhetinic acid. Solid black arrows indicate a dimerization reaction of two farnesyl diphosphate (FPP) molecules catalyzed by squalene synthase (SQS) originating squalene, oxidation by squalene epoxidase (SQE) to 2,3-oxidosqualene, or cyclization catalyzed by bAS. A dashed arrow between mevalonic acid and farnesyl diphosphate indicates multiple enzyme reactions. The blue arrow indicates a single oxidation reaction catalyzed by the CYP88D6 enzyme (Seki et al., 2008); the red arrow indicates a single oxidation reaction catalyzed by the CYP72A154 enzyme, as described herein; the dotted arrows signify undefined oxidation and glycosylation steps. UGATs, UDP-glucuronosyl transferases.

Supplementary Fig. S9: Proposed Glycyrrhizin biosynthesis pathway in Liquorice roots by seki et al 2011