RNA- Seq analysis of *Taverniera cuneifolia* (Roth) Ali.

Talibali Momin (talib.a.m-botanyphd@msubaroda.ac.in)
The Maharaja Sayajirao University of Baroda
Dr. Apurvasinh Puvar
Gujarat Biotechnology Research Centre
Dr. Harshvardhan Zala
Sardarkrushinagar Dantiwada Agricultural University
Dr. Garima Ayachit
Gujarat Biotechnology Research Centre
Dr. Madhvi Joshi
Gujarat Biotechnology Research Centre
Dr. Padamnabhi Nagar (padamnabhi.nagar-botany@msubaroda.ac.in)
The Maharaja Sayajirao University of Baroda.

Research Article

Keywords: Taverniera cuneifolia, De novo assembly, Transcriptome, Licorice, Glycyrrhizin, Sesquiterpenoid pathway

Posted Date: March 29th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2731450/v1

Abstract

Aims: To unravel the potential of *Taverniera cuneifolia*in transcriptomics

Methods: RNA Seq experiment was conducted using NGS to understand the key metabolites and genes from *Taverniera cuneifolia*(Roth) Ali.

Key results: 7.29 Gb of raw data and generated 5,59,91,233 raw reads. The high-quality reads were *de novo* assembled by Trinity assembler followed through CD-HIT resulted into 35,590 Unigene with an average of 419 bp. The unigenes analyzed using BLAST2GO resulted in 27884 (78.35%) transcripts with blast hits, 22510 (63.25%) transcripts with mapping and 21066 (59.19%) transcripts with annotation. Functional annotation was carried out using NCBI non-redundant and Uniprot databases resulted in the identification of 21066 (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 179 metabolic enriched pathways were identified which included pathways like Sesquiterpenoid and triterpenoid pathway which were involved in synthesis of important secondary metabolite Glycyrrhizin biosynthesis. Transcription factors were predicted by comparison with Plant Transcription Factor Database and 1557 transcripts belonging to 85 transcription factor families were identified.

Conclusions: This transcriptome analysis provided an important resource for future genomic studies in *Taverniera cuneifolia*., besides identifying functional secondary metabolites genes of *Taverniera cuneifolia* with Pharmacological potential for future functional genomics and metabolomic engineering of secondary metabolites form these plants towards diversified industrial applications.

Implications: The study confirms that presence of glycyrrhizin producing genes with six key genes that are backbone of sesquiterpenoid biosynthesis pathway in *Taverniera cuneifolia*.

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 (Y. 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 (Pal & Shukla, 2003) and Barda dungar ni Vanaspati ane upyog (Indraji, 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.

Introduction

Glycyrrhiza glabra, commonly known as Liquorice, has been employed in traditional formulations since India is rich in potential medicinal herbs. Licorice (*Glycyrrhiza glabra*) root has been employed as a primary component in over 1200 traditional Chinese herbal remedies. This plant has numerous important applications in industries such as food, herbal, cosmetics, nutraceuticals, and so on (Pastorino et al., 2018). Owing to market demand, it is imported from foreign nations and is not available locally of higher quality. *Taverniera cuneifolia* has been regarded as a powerful replacement for licorice in India. Glycyrrhizin is a valuable bioactive sesquiterpenoid found in this plant.

The fabaceae family, which has over 800 genera and 20,000 species, is the third-largest family of flowering plants, and includes *Taverniera cuneifolia*. Mimosaceae, Papilionaceae, and Caesalpiniacea are the three principal subfamilies. Mendel's discovery (1866), which laid the groundwork for contemporary plant genetics, employed the pea (*Pisum sativum* L.) as a model organism. The basal chromosomal count, ploidy level, and reproductive biology of the lineages varies substantially from one another. The Trifolieae and Loteae tribes of the Galegoid clade's *Medicago truncatula* and *Lotus japonicus*, respectively, were chosen as the model systems for investigating the genomes and biology of legumes. The soybean is the most commonly cultivated and economically significant legume, and it has undergone extensive research. The soybean genome has been available since 2010. Since 2014, the genome of the common bean (*Phaseolus vulgaris*), the most popular grain legume, has been made public. Since then, a great deal more legumes have been sequenced (Smýkal et al., 2020).

According to Charak Samita, written during the Nigantu era, *Taverniera cuneifolia* is a significant traditional medicinal herb used in India. Because that it has the same sweet flavour as *Glycyrrhiza glabra* (commercial licorice), it is frequently referred to as Indian licorice (Zore et al., 2008). There are sixteen different species in the genus *Taverniera* (*Www.Worldfloraonline.Org/Taxon/Wfo-4000037624*, 2018). With several clinical and commercial applications, including food, herbal medicine, cosmetics, etc., licorice is a significant component of traditional Chinese medicine (Y. Liu et al., 2015). The tribal people of Jamnagar in Western India's Saurashtra and Gujarat use *Taverniera cuneifolia*, also known as Jethimad, as a substitute for licorice; the plant itself is thought to be *Glycyrrhiza glabra* (P.S.Nagar, 2005).There have been numerous reports of the plants medicinal uses, including as an expectorant, blood purifier, anti-inflammatory, wound healer, anti-ulcer, and therapy for spleen tumours (Indraji, 1910; Manglorkar, 2013; P.S.Nagar, 2005).

Taverniera cuneifolia has demonstrated the existence of alkaloids, flavonoids, tannins, proteins, reducing sugar, and saponins at the biochemical level. The presence of oil revealed polyunsaturated fatty acids, monounsaturated fatty acids, and saturated fatty acids in the seeds of *Taverniera cuneifolia* (P & H, 2016). There have been relatively few attempts to define *Taverniera cuneifolia* at the molecular level, and there has been very little evaluation of this plant on a phytochemical basis. The plant has eight chromosomes (Perveen & Khatoon, 1989). On NCBI, there is not much genetic data about this species. RNA polymerase beta subunit, ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit, ribosomal protein L32, photosystem 1 assembly protein Ycf4, cytochrome b6/f complex subunit VIII, D1

protein, photosystem 2 protein M, MaturaseK, phosphogluconate dehydrogenase, UDP-sulfoquinovose synthase, and matura (P. L. Liu et al., 2017).

The current study concentrated on Glycyrrhizin, the most valuable secondary metabolite, as well as other significant secondary metabolites. This experiment offers detailed descriptions of this plant. Based on the aforementioned information, root transcriptome sequencing has been used to try and identify the genes of different metabolic pathways in *Taverniera cuneifolia*. The research will provide a scientific understanding of *Taverniera cuneifolia*'s molecular network.

Materials And Methods

Plant material and RNA isolation

The plant *Taverniera cuneifolia* was taken from its native environment close to the Lakhpat area in Kutch, Gujarat, India (23.7887 N, 68.79580 E). For longer-term storage, the plant tissue, or roots, were washed with water rather than ethanol and then placed in RNA Later Solution (Qiagen). The temperature was then changed in the fridge to -20°C. Following the manufacturer's instructions, total RNA was extracted from the plant's root tissues using the RNeasy Plant Mini Kit (Qiagen). With the help of formaldehyde agarose gel electrophoresis, the RNA's integrity was evaluated. Using a Qiaxpert (Qiagen), Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA, USA), and Qiaxcel capillary electrophoresis, total RNA was measured (Qiagen). The sample's RNA integrity number (RIN) was more than about 7.0.

cDNA library preparation and Sequencing

A RiboMinus RNA plant kit for RNA- Seq was used to do ribosomal RNA depletion (Life Technologies, C.A). Ion total RNA-Seq kit v2 (Life Technologies, C.A.) was used to create mRNA fragmentation and a cDNA library, which was then further purified with Ampure XP beads (Beckman coulter, Brea, CA, USA). Dynabeads MyOne Streptavidin C1 was used to enrich the library on ion sphere particles while following the established procedures for ion proton sequencing. The sequence read archive (SRA) NCBI database has received the raw transcriptome data and assigned it the accession number SRR5626167. The DDBJ/EMBL/GenBank accession number for this Transcriptome Shotgun Assembly project is GJAF00000000.

RNA-Seq data processing and *De novo* assembly

Data from sequencing was gathered, and then using FastQCv.0.11.5, the raw reads were subjected to strict filtering rules to remove reads with adaptors, reads with ambiguous bases, and reads of low quality. For the transcriptome assembly, high quality (HQ) reads—reads with each base having a Phred score of at least 20—were taken into account. Using a minimum contig length of 200 bases and a k-mer size of 25 bp, the "Trinity" assembler (Grabherr et al., 2011) was used to do primary assembly by combining the HQ reads. The Inchworm algorithm constructed contigs with a minimum count of 2 k-mers, and a minimum

of 5 reads were used to join two contigs. Using the CD-HIT EST (v4.6.1) tool, a secondary assembly was performed to group contigs coming from the same gene or protein (W. Li & Godzik, 2006). 80% identical homologous contigs were clustered to produce full-length transcripts. We used Bowtie to map the assembled transcriptome onto the processed reads in order to calculate the proportion of reads that were mapped to the assembled transcriptome. The finished file was then entered into a functional annotation and categorization programme.

Functional annotation of transcripts and classification

The non-redundant (nr) database was searched for contigs in the functional characterization of assembled sequences (https://www.ncbi.nlm.nih.gov/). Blast2GO was used to perform further annotation after an e-value cut-off of 1E-5 (Conesa et al., 2005). The classification of the predicted coding sequences' functions was done using a Gene Ontology (GO) analysis. Biological process (BP), Molecular function (MF), and Cellular component were the three main domains that the GO used to categories the functionally annotated coding sequences (CC). using KEGG, the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2017).

Validation of gene for sesquiterpenoid biosynthesis pathway transcripts using qRT-PCR

Using Primer3 online programme, the primers were created from roots sequences associated with the sesquiterpenoid biosynthesis pathway and differentially expressed transcripts and genes (in fasta format) (Rozen & Skaletsky, 2000). Using the Prime ScriptTMfirst strand cDNA synthesis kit from TAKARA Bio Corporation, the genotypes from total RNA were created using 2 g of total RNA in a 20 µl reaction volume. Using the QuantStudioTM 5 system Real Time PCR with SYBR Green detection chemicals, PCR reactions were carried out in 96-well plates. As endogenous reference genes for normalisation, GAPDH and Actin were employed. The QuantStudioTM Real-Time PCR detection system is compatible with the QuantStudioTM 5 system software 1.5.1, a ThermoFisher Scientific exclusive. Using the equation described in, the relative expression of these transcripts was determined (Saussoy et al., 2004).

Identification of transcription factors families

By evaluating InterProScan domain patterns in protein sequences with high coverage and sensitivity using the PlantTFcat analysis programme (http://plantgrn.noble.org/PlantTFcat/), transcription factors (TFs) were discovered utilizing genome-scale protein and nucleic acid sequences (Dai et al., 2013).

SSR prediction

Using the MISA tool (Microsatellite; http://pgrc.ipk-gatersleben.de/misa/misa.html). simple sequence repeats (SSRs) were found. We looked for SSRs with sizes ranging from one mononucleotide to six. For SSR searches, the minimum repetition count was set at 10 for mononucleotides, 6 for dinucleotides, and 5 for trinucleotides to hexanucleotides. With a compound microsatellite, the maximum number of bases that can separate two SSRs is 100, meaning that 100 bases were chosen as the minimum distance between two adjacent SSR markers.

Results

Ion torrent Sequencing and *De novo* assembly

Ion Total RNA-Seq kit v2 (Life Technologies, C.A.) was used to convert the whole RNA from two root samples with RIN values more than 7.0 into a cDNA library, which was then further purified using Ampure XP beads (Beckman coulter, Brea, CA, USA). Ion spherical particles were added to the collection using MyOne C1 DynabeadsTM. Ion proton sequencing standard methods generated a total of 7.29 gigabytes of raw data (Table 1). For the RNA sequencing, *Taverniera cuneifolia's* high-quality root were utilized, and a total of 55,991,233 reads containing 7,286,727,421 bases were produced. The Fast QC tool was used to do a quality check on the raw reads, and the average base quality was higher than Q20. The 36,896 reads obtained from the de novo transcriptome assembly were combined with 35,590 unique high-quality reads using CD-HIT at 85% sequence similarity and a N50 value of 441 bp to create the final assembly. For *Taverniera cuneifolia*, an average GC content of 43.25% was obtained. The following table lists the statistics for transcriptome sequencing and assembly produced by the Trinity assembler (Table 2).

Functional annotation of transcripts

The Nr Protein database, KEGG, UniProt, and other databases were used to perform functional annotation on a total of 35,590 transcripts (contigs) assembled using Trinity. Transcripts were given GO terms (Supplementary Fig. S2). Based on the statistics of species-specific distribution, all transcripts were tested for similarity to known organisms. It can be deduced that the transcript had the largest blast hits with *Medicago truncatula* (18,734, 52.63%), *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). Using Blast2GO, the functionally annotated transcripts of *Taverniera cuneifolia* were divided into three primary categories: biological processes, cellular components, and molecular function gene ontology (27,884, 78.34%). (Table S1). The biological processes were the most prevalent among them, including 44,395 (40.8%) sequences, followed by various molecular functions with 25,025 (27.5%) sequences, and cellular components with 21,508 (23.6%) sequences (Figure 2, 3, 4). The annotated transcripts were put via KEGG's (Kyoto Encyclopedia of Genes and Genomes) pathway, which connected the transcripts to enzymes that could be located in a variety of pathways that were made available by KEGG. The class of enzymes with the most annotated transcripts is hydrolases, followed by transferases and oxidoreductases (Figure 5).

Gene ontology classification

Biological Processes (44,395[49%]), Molecular Function (25,025[27%]), and Cellular Component (21,508[24%]) category were used to further annotate the contigs using assembled 27,884 transcripts GO keywords (Table S1).

The most prevalent GO word subset was biological processes. The term "Biological process," which was represented by the most contigs out of the 44,395 biological processes, was followed by "Metabolic process" and "Cellular process" (Figure 2).

A total of 25,025 transcripts were annotated to molecular function, with "Molecular function" being the category with the highest number of transcripts, followed by "Catalytic activity" and "Binding," respectively (Figure 3).

Moreover, Cellular Component was annotated to a total of 21,508 transcripts, ranking as the highest match, followed by "Cell" and "Cell portion," 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 179 different pathways were annotated and among them, Metabolic pathways (665), Biosynthesis of secondary metabolites (335), Ribosome (101) 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). Of the numerous chemicals that plants create, terpenoids (isoprenoids) are the largest and most diversified class. Moreover, terpenoids' ecological significance has drawn more focus in the creation of techniques for long-term pest management and abiotic stress defense.

Terpenoids' backbone biosynthesis is encoded by the Kegg ID ko00900. It contains twenty-three enzymes that are expressed in all the color enzymes boxes were present in *Taverniera cuneifolia.* Details for the enzymes are in Table 5. Among them enzyme acetyl-CoA C-acetyltransferase [EC:2.3.1.9] showed minimum expression compare to 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase [EC:1.17.7.4] which showed high expression in *Taverniera cuneifolia* (Table 5, Supplementary Fig. S3).

Sesquiterpenoid and triterpenoid biosynthesis (ko00909), it contains six enzymes that are expressed among them squalene monooxygenase [EC:1.14.14.17] showed the maximum expression while beta-amyrin 24-hydroxylase [EC:1.14.14.134] showed minimum expression in *Taverniera cuneifolia* (Table 5, Supplementary Fig. S4). 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), it contains eight enzymes that are expressed among them beta-carotene 3-hydroxylase [EC:1.14.15.24] showed maximum expression while abscisic-aldehyde oxidase [EC:1.2.3.14] showed the minimum expression in *Taverniera cuneifolia* (Table 5, Supplementary Fig. S5).

Ubiquinone and other terpenoid-quinone biosynthesis (ko00130), it contains eight enzymes among them low expression was observed on 4-coumarate--CoA ligase [EC:6.2.1.12] enzymes in *Taverniera cuneifolia* (Table 5, Supplementary Fig. S6).

Zeatin biosynthesis (ko00908), it contains four enzymes among them tRNA dimethylallyltransferase [EC:2.5.1.75] showed maximu expression while cytokinin dehydrogenase [EC:1.5.99.12] showed the minimum expression in *Taverniera cunefolia*

(Table 5, Supplementary Fig. S7).

Flavonoid biosynthesis (ko00941), it contains nine enzymes among them flavonol synthase [EC:1.14.20.6] showed the maximum expression while anthocyanidin synthase [EC:1.14.20.4] showed minimum expression in *Taverniera cuneifolia* (Table 5, Supplementary Fig. S8).

Candidate genes involved in biosynthesis pathways

Glycyrrhizin is a triterpenoid saponin found in the roots of licorice plants (Glycyrrhiza species) that is widely used in the food, pharmaceutical, and cosmetics industries. The production of glycyrrhizin in plants is controlled by several genes, and understanding the regulation of these genes is important for improving glycyrrhizin production in licorice plants and other related species. Five genes have been found among the 35,591 transcripts that have been annotated using several databases to play a significant part in the biosynthetic process of the formation of glycyrrhizin from *Taverniera cuneifolia* (Table S3 and Supplementary material 1).

One of the key genes involved in glycyrrhizin production is the cytochrome P450 enzyme CYP72A154. This gene encodes an enzyme that catalyzes the conversion of β-amyrin to glycyrrhetinic acid, which is the precursor of glycyrrhizin. Several studies have demonstrated that overexpression of CYP72A154 can increase glycyrrhizin production in licorice plants (Seki et al., 2011). Another important gene in glycyrrhizin production is the 2,3-oxidosqualene cyclase (OSC) gene. OSC catalyzes the cyclization of 2,3-oxidosqualene, which is the precursor of all triterpenoids, including glycyrrhizin. Several OSC genes have been identified in licorice plants, including GgOSC1 and GgOSC2 (Seki, 2015). These genes are expressed specifically in the roots of licorice plants and are important for glycyrrhizin biosynthesis. In addition to these genes, several transcription factors (TFs) have also been implicated in the regulation of glycyrrhizin

production. For example, the MYB TF GmMYB58 has been shown to activate the expression of CYP72A154 and increase glycyrrhizin production in soybean hairy roots(Z. Chen et al., 2021). Similarly, the bHLH TFs GtbHLH1 and GtbHLH2 have been shown to regulate the expression of OSC genes and increase glycyrrhizin production in licorice plants (Tamura et al., 2018).

Overall, understanding the regulation of glycyrrhizin production genes in plants is crucial for improving glycyrrhizin production in licorice plants and other related species. By identifying and manipulating these genes and their regulators, researchers can potentially increase glycyrrhizin production in plants and provide a sustainable source of this valuable compound for various industries.

There were 5414 transcripts annotated with that of hypothetical proteins from the database based on the sequence similarity search. Among them 1745 transcript had percent identity for more than 90. These proteins were reported from most of the fabaceae family Plants (Supplementary material 2).

There were 153 transcripts/gene annotated with Cytochrome P450 protein family in plants. Among them 65 transcripts had percent identity for more than 90 (Supplementary material 3).

Relative expression analysis of genes involved in sesquiterpenoid biosynthesis.

Seven significant transcripts associated with sesquiterpenoid biosynthesis pathways were chosen for RT-qPCR in order to validate the relative expression levels from the transcript abundance estimation. The transcripts' primer sequences were created as shown in (Table S4). Using the equation described in, the relative expression of these transcripts was determined (Saussoy et al., 2004). Actin and GAPDH were used as the references in the CT calculation. In Supplemental, the transcripts' expression has been presented (Table S5).

Discussion

Secondary metabolites play a crucial function in giving plants a defensive mechanism against stressors, and they are crucial for many economically significant businesses, such as the pharmaceutical industry (Pang et al., 2021). There are no known molecular details about this particular plant. High-throughput sequencing data produced by the most recent developments in omics technologies have allowed us to predict genes, enzymes, and complex pathways (Metzker, 2010). De novo of many medicinally important plants such as *Saussurea lappa* (Bains et al., 2019)*, Vigna radiate* L (H. Chen et al., 2015), *Glycyrrhiza glabra* (Chin et al., 2007), Pigeon pea *Cajanus cajan* L. Millspaugh (Murray et al., 2011), *Dracocephalum tanguticum* (H. Li et al., 2017) etc. have reported the transcripts involved in active metabolite production using NGS technology.

One of the most cutting-edge techniques for identifying the genes expressed in various metabolic, growth, developmental, stress-response, cell signaling, etc. pathways is transcriptome analysis. This has aided in identifying and dividing up the various roles played by secondary metabolic compounds. Licorice roots include a well-known secondary metabolite called glycyrrhizin, which shares the same properties as

Taverniera cuneifolia roots and has many of the same purposes as those listed above. The unique transcripts that are first reported from this plant to be engaged in the pathways of primary and secondary metabolism have been revealed by a complete transcriptome analysis of the root of *Taverniera cuneifolia* (Sharma et al., 2020).

Using the blastx programme, *Taverniera cuneifolia's* de novo assembled transcripts were mapped to a non-redundant protein database. *Medicago truncatula* [(18,734) 52.6%], *Cicer arietinum* [(16,044) 45%], *Glycine max* [(15,991) 44.9%], and other members of the same family Fabaceae order fabales, had the highest degree of similarity among the 35,590 transcripts annotated to the database.

Proteins are one of the most important biomolecules found in living organisms. They perform a wide range of functions such as catalysis, transport, storage, and signalling. However, there are still many hypothetical proteins whose functions are unknown. Hypothetical proteins in Plant genomes have been found to contain a large number of genes encoding for hypothetical proteins, which have unknown functions. For example, in Arabidopsis thaliana, a model plant species, approximately 20-30% of the genome encodes for hypothetical proteins (J. Li & Wang, 2000). Hypothetical proteins in plants represent a vast area of unexplored biology. While their functions remain largely unknown, there are several studies that suggest potential roles in plant development, photosynthesis, and growth. Further research is needed to fully understand the functions of these proteins and their potential applications in agriculture and biotechnology.

Cytochrome P450 (CYP450) enzymes are a group of heme-containing monooxygenases that play an important role in various metabolic processes in plants. These enzymes are involved in the synthesis of various plant secondary metabolites, such as alkaloids, flavonoids, and terpenoids. Functions of CYP450 Enzymes in Plant Metabolism CYP450 enzymes are involved in a wide range of metabolic processes in plants, including the biosynthesis of plant secondary metabolites, hormone signalling, and detoxification of xenobiotics. One of the primary functions of CYP450 enzymes in plants is the synthesis of plant secondary metabolites. These compounds are important for plant growth and development, as well as for their defense against various biotic and abiotic stresses (Pandian et al., 2020).

One of the best-studied examples of CYP450 enzymes in plant secondary metabolism is the biosynthesis of the alkaloid nicotine in tobacco. Nicotine biosynthesis is catalyzed by a series of CYP450 enzymes, including CYP82E4, CYP82E5, and CYP82E10. These enzymes are involved in the conversion of nicotine precursors, such as nornicotine and anatabine, into nicotine(Moghbel, 2016). CYP450 enzymes are also involved in the biosynthesis of flavonoids, which are important plant secondary metabolites that play a role in plant pigmentation and protection against UV radiation. For example, CYP75B1 and CYP75B2 are involved in the biosynthesis of the flavonoid anthocyanin in Arabidopsis thaliana (Casati, 2012). In *Taverniera cuneifolia* there were 153 gene that annotated for Cytochrome P450 enzymes and sixty five transcripts were having more than 90 % identity to different plants *Cicer arietinum*, *Vigna anguiculata*, *Lotus japonicus* etc which were mostly the fabaceae family plants. In addition to their roles in secondary metabolism, CYP450 enzymes are also involved in hormone signalling in plants. For example, CYP90B1

is involved in the biosynthesis of the plant hormone brassinosteroid. In *Taverniera cuneifolia* we found 8 transcripts that annotated for the brassinosteroid. while CYP83B1 is involved in the biosynthesis of the hormone indole-3-acetic acid (IAA), *Taverniera cuneifolia* had 3 transcripts that annotated with IAA.

Main metabolism-related gene of *Taverniera cuneifolia.*

Licorice roots generate the triterpenoid-saponin glycyrrhizin. The cytosolic melvonic acid route is used to create 2,3-oxidosqualene, which is then converted to β-amyrin by β-amyrin synthase (bAS). Finally, as seen in (Supplementary Fig. S9), β-amyrin proceeds through a two-step oxidation at the C-30 position followed by glycosylation reactions at the C-3 hydroxyl group to create glycyrrhizin Seki et al., 2008, 2011). *Glycyrrhiza glabra* can be replaced by *Taverniera cuneifolia*, often known as Indian Licorice, because it shares many of the same characteristics with this plant. Triterpenoids, flavonoids, polysaccharides, and other compounds that have been discovered for the first time from this plant are among the many distinct compounds found in this plant that can be employed in future study.

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 *Taverniera cuneifolia* 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 four gene identified from *Taverniera cuneifolia* 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 11-oxo-beta-amyrin 30-oxidase/CYP72A to form Glycyrrhetinic acid. There were nine genes in *Taverniera cuneifolia* that annotated for CYP72A and one gene that annotated for CYP93E1. 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 thirty two different UDP-glycuronosyl genes which have been identified from *Taverniera cuneifolia* that led to last reaction given in table (Table S3).

We have associated *Taverniera cuneifolia* with the reference pathway proposed by Seki et al. 2008, 2011 since *Taverniera cuneifolia* has not yet undergone extensive study, and as a result, there are no papers that provide specific information about the enzymes involved in the Glycyrrhizin pathway. Due to the fact that the known pathway for glycyrrhizin has not been properly investigated. We have extensively worked upon the proteins which we have opted from our data of *Taverniera cuneifolia*.

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).

Declarations

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.

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

Funding

There was no funding provided for the Research project.

Competing Interests:

“The authors have no relevant financial or non-financial interests to disclose.”

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 TM finalized the manuscript.

References

Bains, S., Thakur, V., Kaur, J., Singh, K., & Kaur, R. (2019). Elucidating genes involved in sesquiterpenoid and flavonoid biosynthetic pathways in Saussurea lappa by de novo leaf transcriptome analysis. *Genomics*, *111*(6), 1474–1482. https://doi.org/10.1016/j.ygeno.2018.09.022

Casati, P. (2012). *Flavonoids : biosynthesis , biological functions , and biotechnological applications*. *3*(September), 1–15. https://doi.org/10.3389/fpls.2012.00222

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

Chen, Z., Fang, X., Yuan, X., Zhang, Y., Li, H., Zhou, Y., & Cui, X. (2021). Overexpression of transcription factor gmtga15 enhances drought tolerance in transgenic soybean hairy roots and arabidopsis plants. *Agronomy*, *11*(1). https://doi.org/10.3390/agronomy11010170

Chin, Y. W., Jung, H. A., Liu, Y., Su, B. N., Castoro, J. A., Keller, W. J., Pereira, M. A., & 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

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *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

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Indraji, J. (1910). Botany : a complete and comprehensive account of the flora of Barda Mountain (Kathiawad) / by Jayakrishna Indraji. In *Botany : a complete and comprehensive account of the flora of Barda Mountain (Kathiawad) / by Jayakrishna Indraji.* Bombay : Printed at the “Gujarati” Printing Press. https://doi.org/10.5962/bhl.title.10897

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

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*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-00078-y

Li, J., & Wang, D. (2000). Cloning and in vitro expression of the cDNA encoding a putative nucleoside transporter from Arabidopsis thaliana. *Plant Science*, *157*(1), 23–32. https://doi.org/10.1016/S0168-9452(00)00261-2

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

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), 1–19. https://doi.org/10.1371/journal.pone.0170596

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

Manglorkar, P. S. (2013). *Bioprospecting the potential of Taverniera cuneifolia (Roth) Ali.*

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

Moghbel, N. (2016). *Investigations into the pituri plant: nicotine content, nicotine conversion to nornicotine, nicotine release and cytotoxicity of Australian native Nicotiana spp*. 135.

Murray, M. G., Thompson, W. F., Singh, B. P., Gupta, D. K., Singh, S., Dogra, V., Gaikwad, K., Sharma, T. R., Raje, R. S., Bandhopadhya, T. K., Datta, S., Singh, M. N., Bashasab, F., Kulwal, P., Wanjari, K., Varshney, R. K., Cook, D. R., Singh, N. K., Winga, R., … Cook, D. (2011). Develpment of genic-SSR markers by deep transcriptome sequencing in pigeonpea. *Nucleic Acids Research*, *8*(19), 4321–4326. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/8.19.4321

P.S.Nagar. (2005). *Floristic biodiversity of barda hills and its surroundings* (P.S.Nagar (ed.); 1st ed.). SCIENTIFIC PUBLISHERS (INDIA).

P, M., & H, J. (2016). Chemical Composition and Characteristics of Taverniera Cuneifolia (Roth) Ali Seed Oil. *Journal of Pharmacognosy & Natural Products*, *1*(1), 1–3. https://doi.org/10.4172/2472-0992.1000101

Pal, S. K., & Shukla, Y. (2003). Herbal medicine: Current status and the future. *Asian Pacific Journal of Cancer Prevention*, *4*(4), 281–288.

Pandian, B. A., Sathishraj, R., Djanaguiraman, M., Prasad, P. V. V., & Jugulam, M. (2020). Role of cytochrome P450 enzymes in plant stress response. *Antioxidants*, *9*(5), 1–15. https://doi.org/10.3390/antiox9050454

Pang, Z., Chen, J., Wang, T., Gao, C., Li, Z., Guo, L., Xu, J., & Cheng, Y. (2021). Linking Plant Secondary Metabolites and Plant Microbiomes: A Review. *Frontiers in Plant Science*, *12*(March). https://doi.org/10.3389/fpls.2021.621276

Pastorino, G., Cornara, L., Soares, S., Rodrigues, F., & Oliveira, M. B. P. P. (2018). Liquorice (Glycyrrhiza glabra): A phytochemical and pharmacological review. In *Phytotherapy Research* (Vol. 32, Issue 12, pp. 2323–2339). John Wiley and Sons Ltd. https://doi.org/10.1002/ptr.6178

Perveen, S., & Khatoon, S. (1989). Chromosome numbers in Papilionaceae from Pakistan. In *Pakistan Journal of Botany* (Vol. 21, pp. 247–251).

Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for General Users and for Biologist Programmers In: Krawetz S, Misener S (eds) Methods and protocols: methods in molecular biology. *Methods in Molecular Biology*, *132*(1), 365–386. http://dx.doi.org/10.1385/1-59259-192-2:365

Saussoy, P., Vaerman, J. L., Straetmans, N., Deneys, V., Cornu, G., Ferrant, A., & Latinne, D. (2004). Differentiation of acute myeloid leukemia from B- and T-lineage acute lymphoid leukemias by real-time quantitative reverse transcription-PCR of lineage marker mRNAs. *Clinical Chemistry*, *50*(7), 1165–1173. https://doi.org/10.1373/clinchem.2003.030114

Seki, H. (2015). *Title : P450s and UGTs : key players in the structural diversity of triterpenoid saponins Running title : P450s and UGTs in saponin biosynthesis Title : P450s and UGTs : key players in the structural diversity of triterpenoid saponins Running title : P450*. *2*.

Seki, H., Sawai, S., Ohyama, K., Mizutani, M., Ohnishi, T., Sudo, H., Fukushima, E. O., Akashi, T., Aoki, T., Saito, K., & Muranaka, T. (2011). Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. *Plant Cell*, *23*(11), 4112–4123. https://doi.org/10.1105/tpc.110.082685

Sharma, P., Kumar, S., Beriwal, S., Sharma, P., Bhairappanavar, S. B., Verma, R. J., & Das, J. (2020). Comparative transcriptome profiling and co-expression network analysis reveals functionally coordinated genes associated with metabolic processes of Andrographis paniculata. *Plant Gene*, *23*, 100234. https://doi.org/10.1016/j.plgene.2020.100234

Smýkal, P., von Wettberg, E. J. B., & McPhee, K. (2020). Legume genetics and biology: From mendel's pea to legume genomics. *International Journal of Molecular Sciences*, *21*(9). https://doi.org/10.3390/ijms21093336

Tamura, K., Yoshida, K., Hiraoka, Y., Sakaguchi, D., Chikugo, A., Mochida, K., Kojoma, M., Mitsuda, N., Saito, K., Muranaka, T., & Seki, H. (2018). The Basic Helix-Loop-Helix Transcription Factor GubHLH3 Positively Regulates Soyasaponin Biosynthetic Genes in Glycyrrhiza uralensis. *Plant and Cell Physiology*, *59*(4), 778–791. https://doi.org/10.1093/pcp/pcy046

www.worldfloraonline.org/taxon/wfo-4000037624. (2018). http://www.worldfloraonline.org/taxon/wfo-4000037624

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

Tables

Tables 1-5 are not available with this version.

Supplementary Material

Supplementary Figures and Tables are not available with this version.

Figures

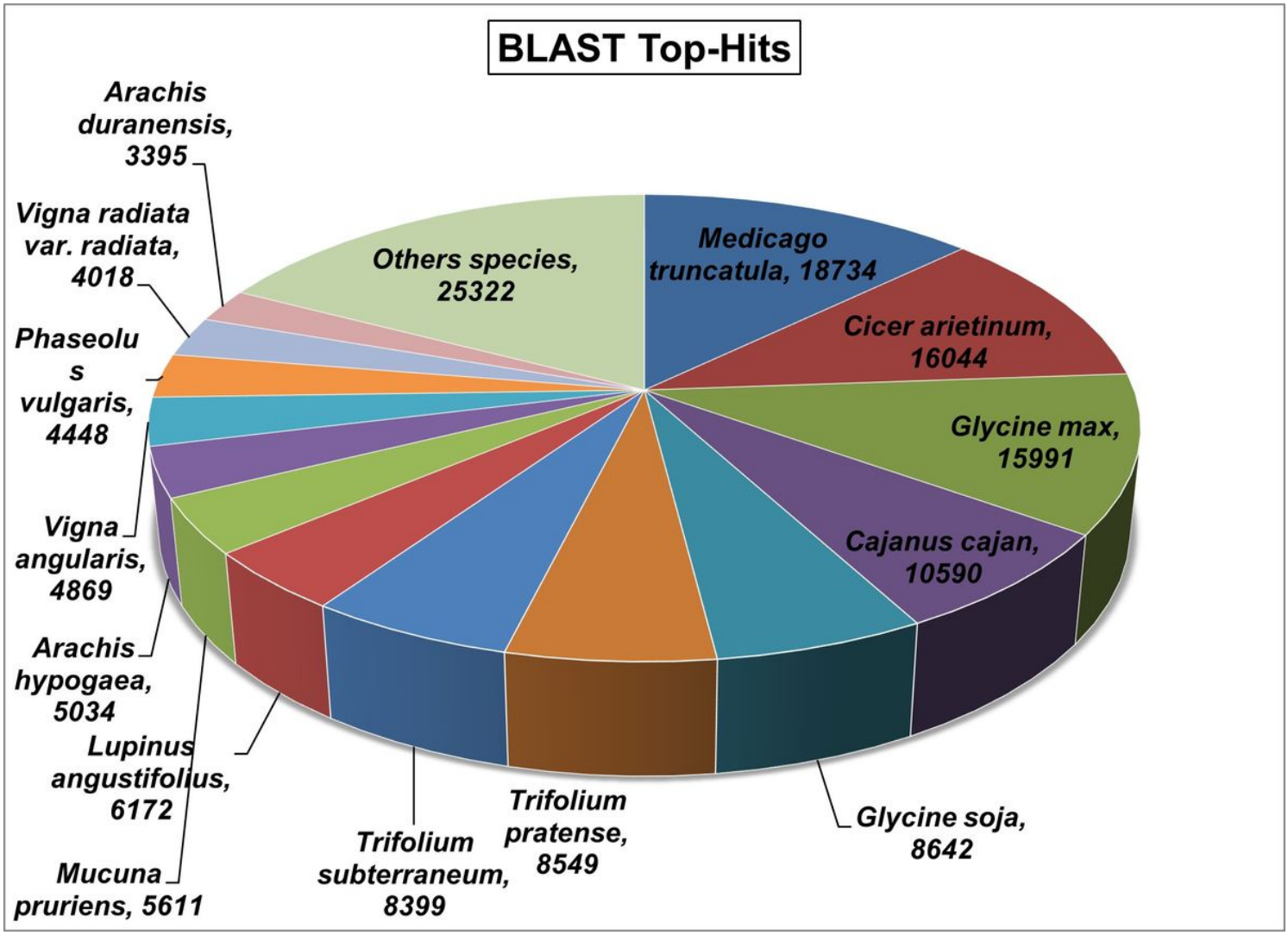

Figure 1

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

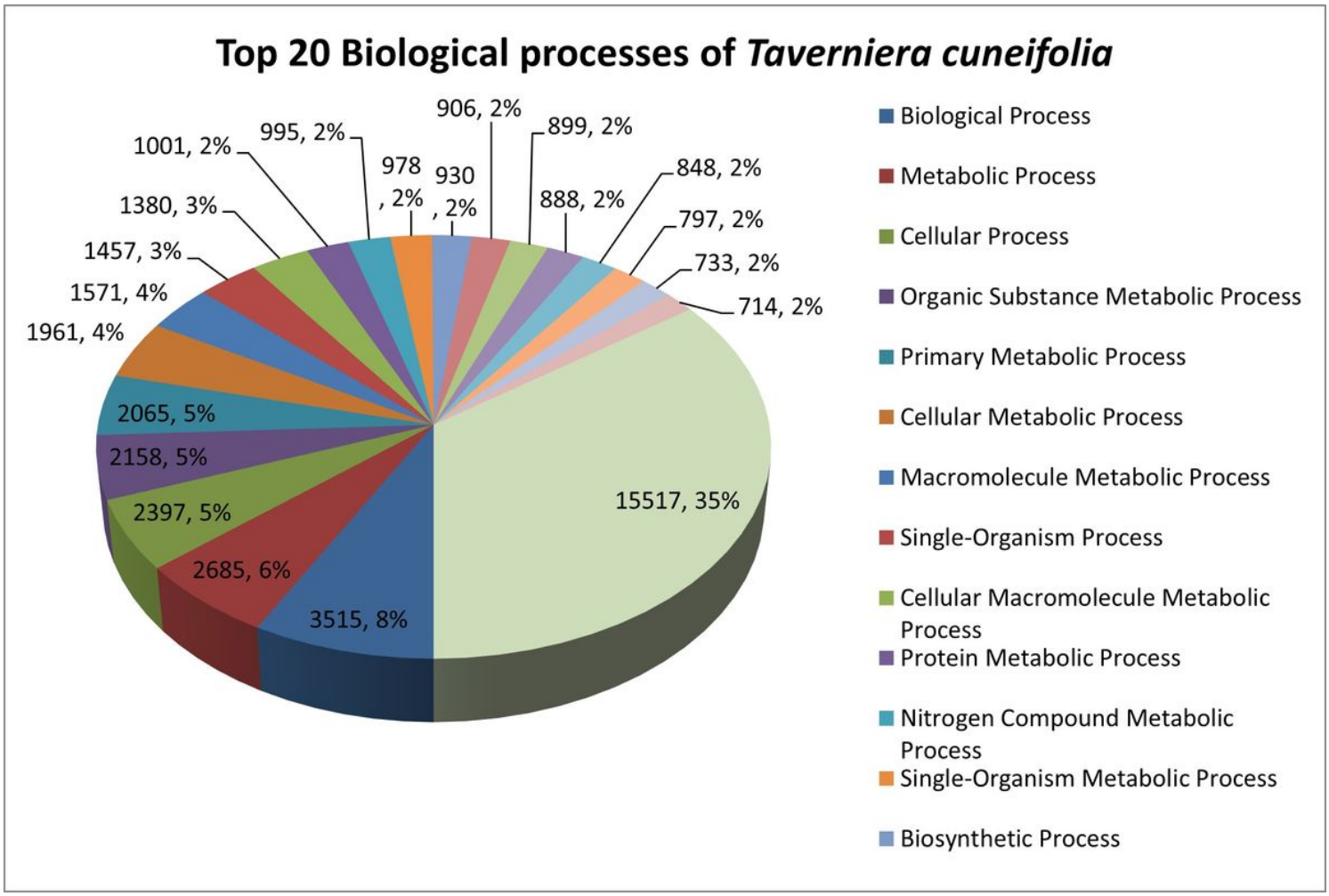

Figure 2

Biological processes gene ontology of *Taverniera cuneifolia* transcripts.

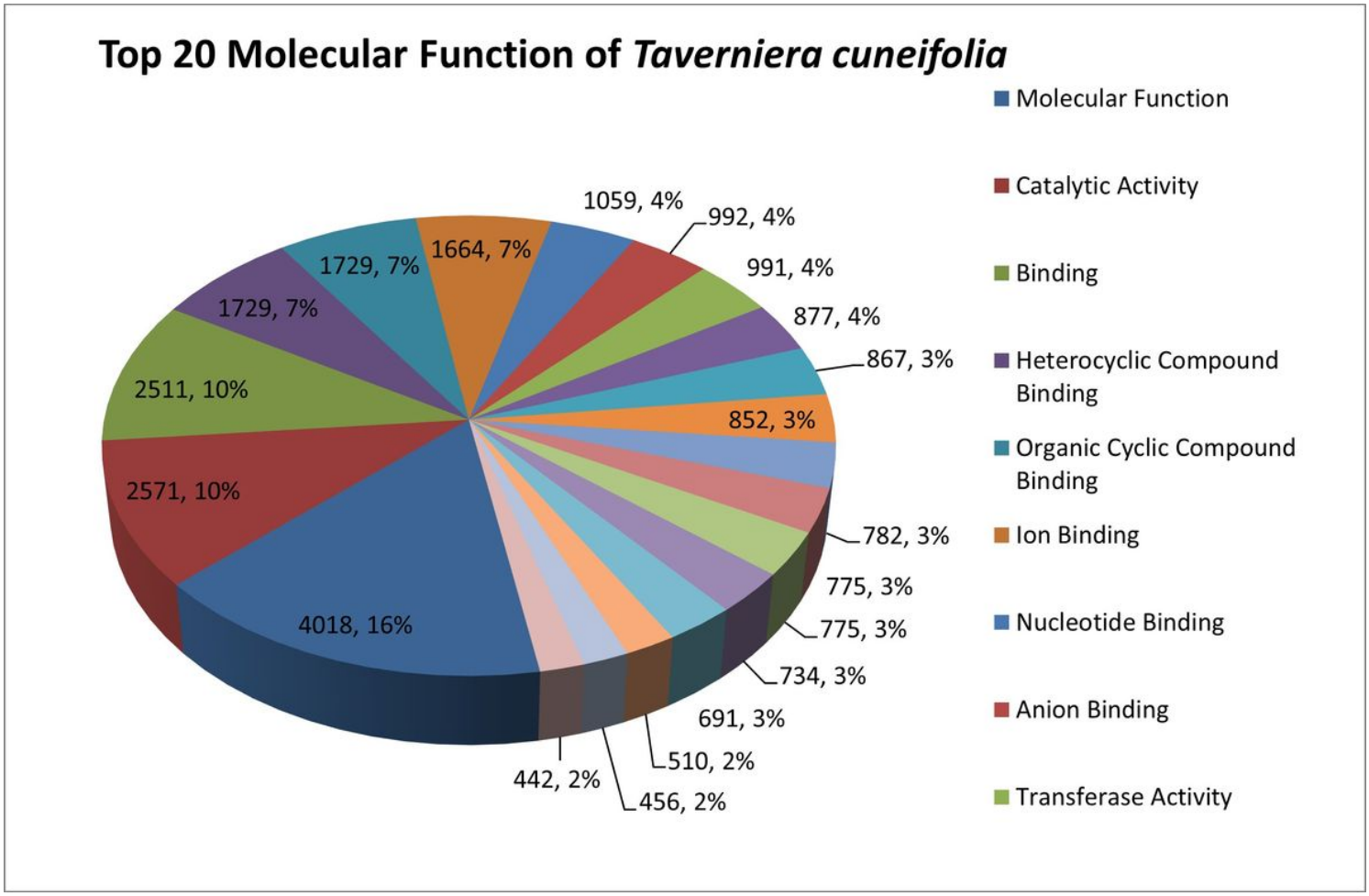

Figure 3

Molecular functions gene ontology of *Taverniera cuneifolia* transcripts.

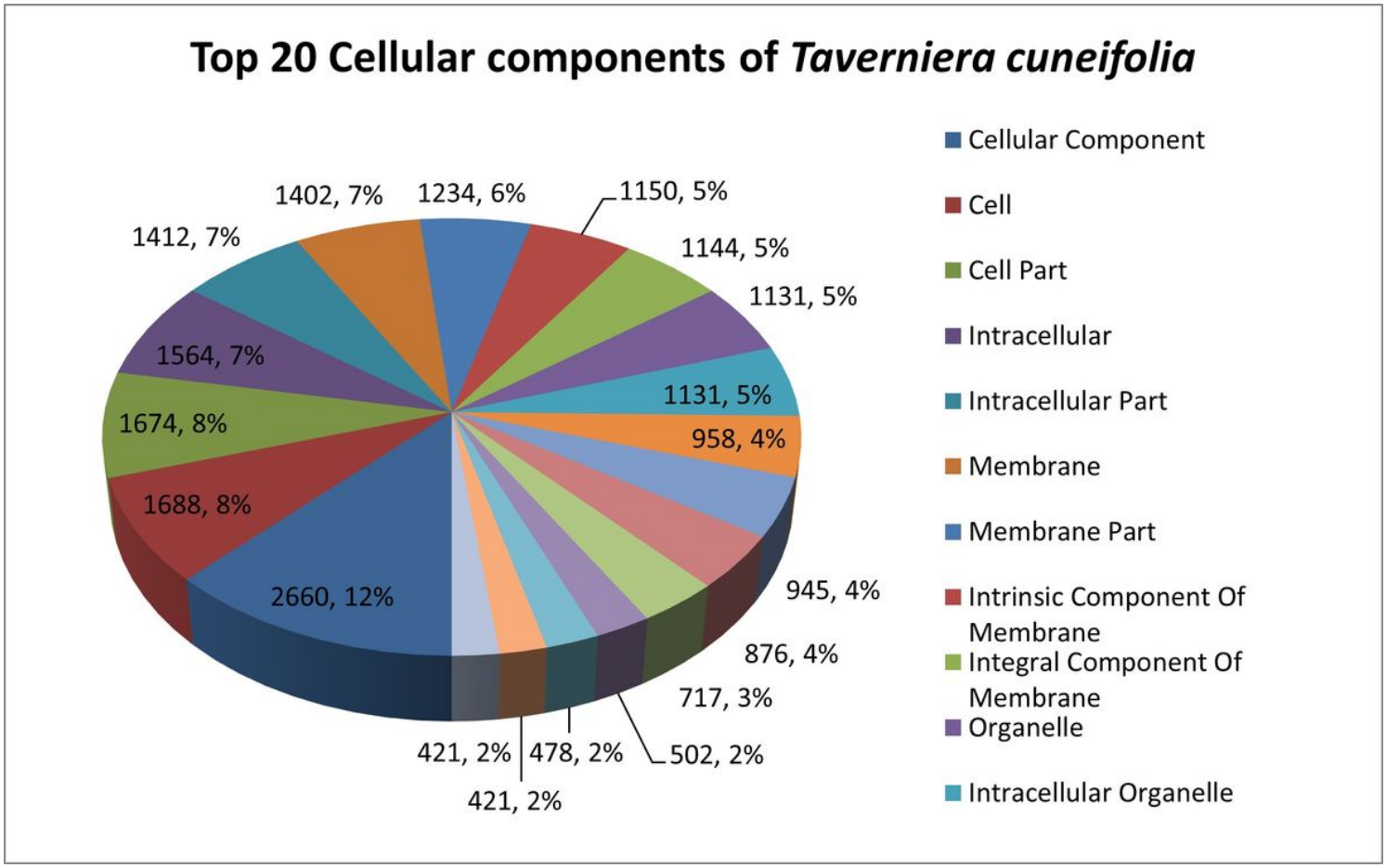

Figure 4

Cellular components gene ontology of *Taverniera cuneifolia* transcripts.

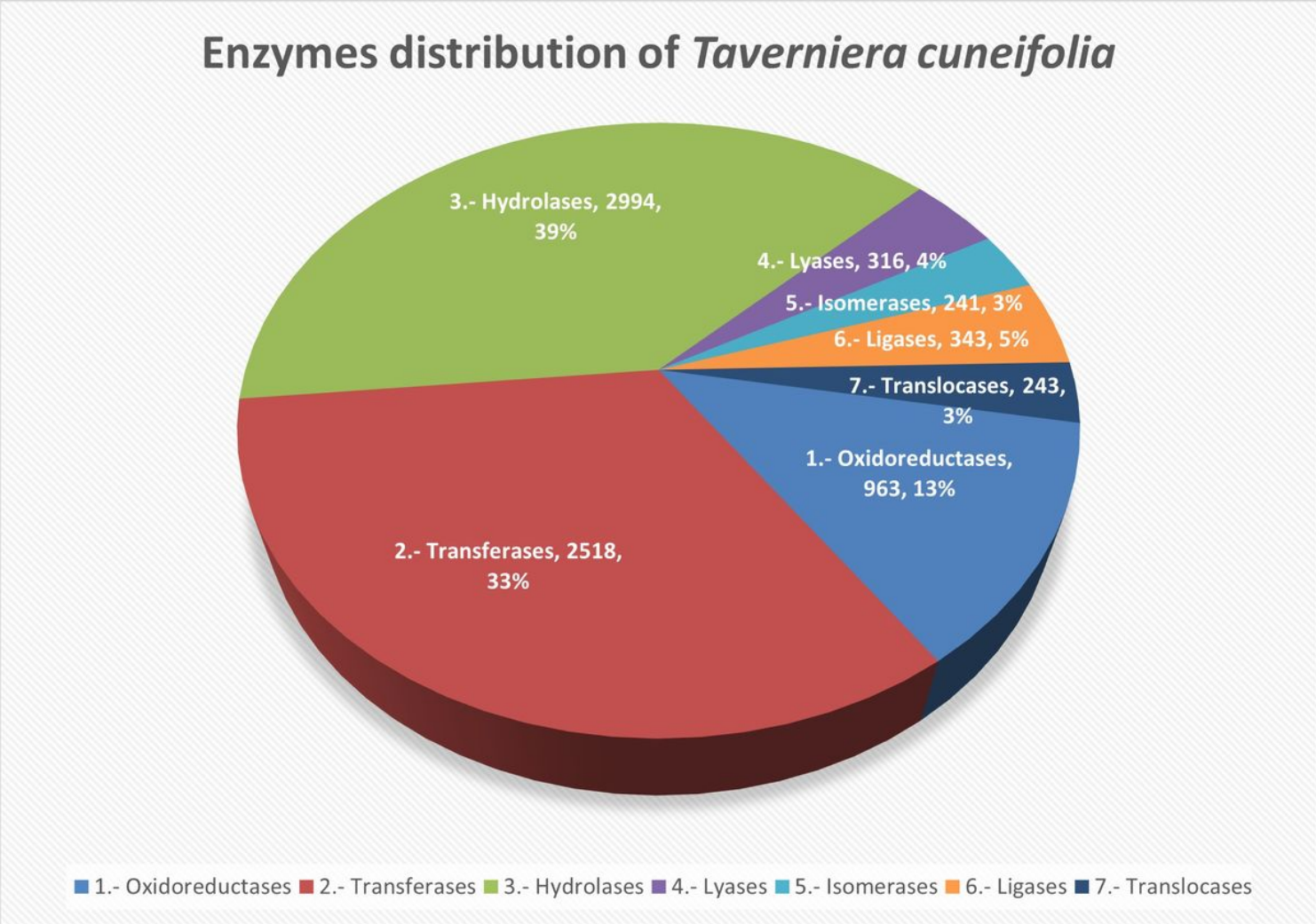

Figure 5

Enzyme classification of *Taverniera cuneifolia* transcripts based on KEGG pathway.

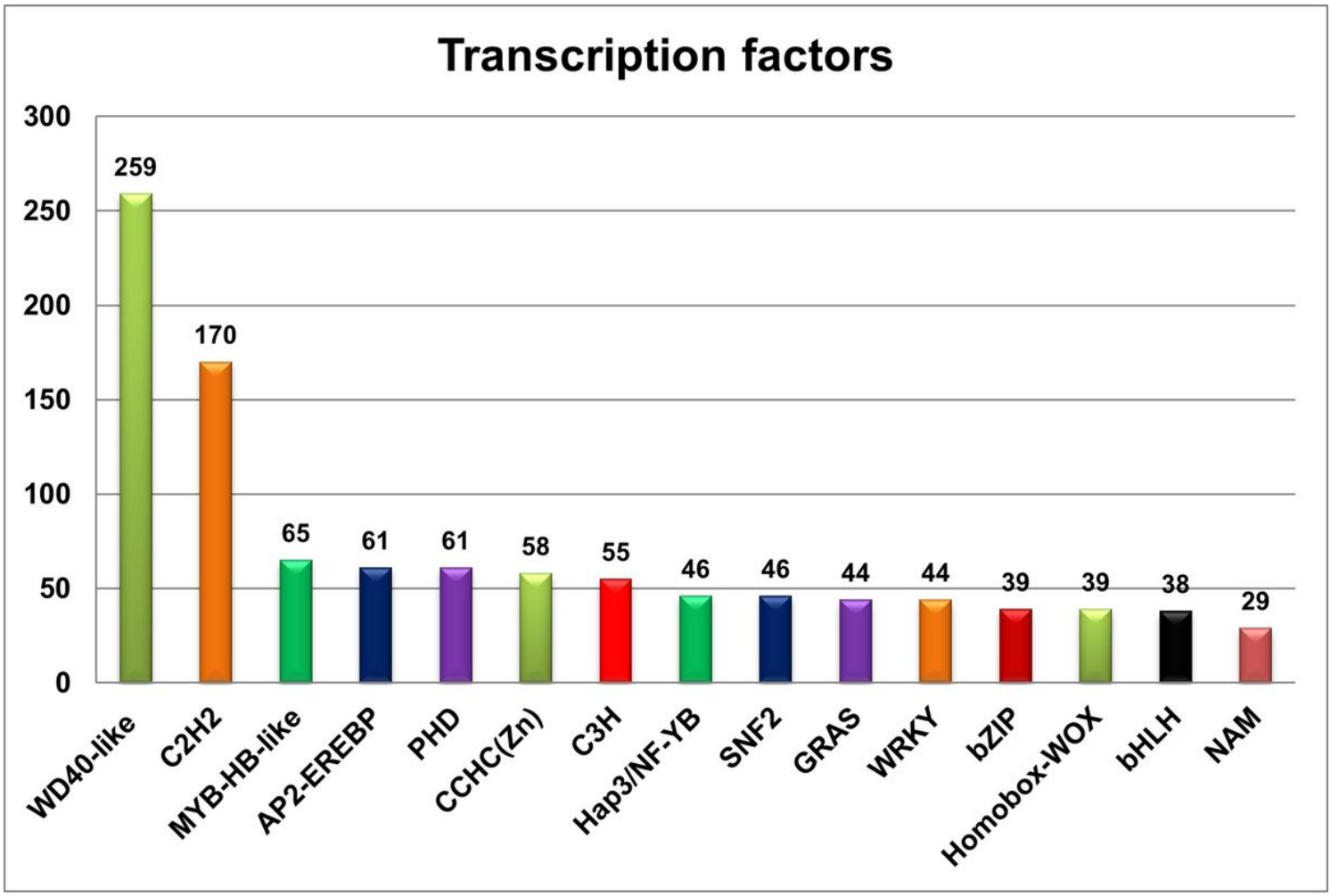

Figure 6

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

Supplementary Files

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

- supplementarymaterial1GeneofGlycyrrhizin.xlsx
- Supplementarymaterial2HypotheticalproteinspresentinTC.xlsx
- Supplementarymaterial3CytochromeP450genes.xlsx
- Supplementarymaterial4RTPCR.xlsx