International Research Journal of Biotechnology
Vol.11 (4) pp. 1-4, Oct, 2021
Available online @http://www.interesjournals.org/IRJOB

Research Article

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

Talib Ali Momin*[1], Apurva Punvar[2], Harshvardhan Zala[3], Ayachit Garima[2], Madhvi Joshi[2], Padamnabhi S. Nagar[1]

[1]Department of Botany, Faculty of Science, The Maharaja Sayajirao University of Baroda -390002.
[2]Department of Science and Technology, Gujarat Biotechnology Research Center (GBRC), 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
E-mail: talib429gmail.com

Abstract

Background: India is rich in many potential medicinal plants, *Glycyrrhiza glabra* popularly known as Liquorice have been used in traditional formulation. Licorice (*Glycyrrhiza glabra* roots) 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. 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. Since there are no molecular studies on this plant the present experiment focuses on in-depth transcriptome analysis in *Taverniera cuneifolia*.

Results: Transcriptomic analysis of *Taverniera cuneifolia* roots resulted in a total of ~7.29 Gb of raw data, generated from Ion Torrent platform (PROTON). The pre-processing of raw reads 55,991,233 was carried out using FastQC and FastXTool kit (Andrews, 2010). The pre-processing of raw reads resulted in high-quality reads, assembled into 36,896 contigs by Trinity assembler, finally assembled using CD-HIT (85% sequence similarity). Assembled transcripts were functionally annotated using Blast X (Nr) databases. A total of 279 metabolic enriched pathway 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.

Conclusion: Transcriptome analysis of *T.cuneifolia* provides the first time information about the gene and enzymes involved in Glycyrrhizin biosynthesis and other secondary metabolites. The transcriptome data will help in development of Molecular markers based on the EST.

Key words: *Taverniera cuneifolia*, Transcriptomic, Glycyrrhizin, Ion torrent, Sesquiterpenoid pathway, Cytochrome P450, Next-generation sequencing.

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.

Keywords: *Salinity tolerance, SSR primers, Rice (Oryza sativa), Polymorphism information content, Cluster analysis.*

INTRODUCTION

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). *Taverniera cuneifolia* (family Fabaceae) 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). Licorice is used as important traditional Chinese medicine with many clinical and industrial applications like Food, Herbal medicine, cosmetics etc. *T.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. At the biochemical level, *T*. *cuneifolia* has shown the presence of alkaloids, flavonoids, Tannins, proteins, Reducing sugar and Saponins. The presence of oil content in the seeds of T. cuneifolia showed polyunsaturated fatty acids, monounsaturated fatty acids and saturated fatty acids (Manglorkar, 2016).

T.cuneifolia has been assessed on phytochemical basis; there are no attempts to characterize this phytochemical basis at on molecular level. Based on the above references attempts have been made to identify the genes of various metabolic pathways in *T.cuneifolia* through Root transcriptome sequencing. The study will give scientific insight into the molecular network of *Taverniera cuneifolia*. The coding regions will assist in identifying the proteins involved in Glycyrrhizin biosynthesis and other pathways. SSR markers were identified and transcription factors from the data can help in further research in future.

MATERIALS AND METHODS

Plant material and RNA isolation

Taverniera cuneifolia plant was collected from Kutch, Gujarat, India (23.7887° N, 68.79580° E) on 16/12/2016 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 Score (RIS) was higher than approx. 7.0 for the sample.

cDNA library preparation and ion torrent 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 my one C1 dynabeads 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.

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 MG, Haas BJ 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 W., and Godzik, 2006).

Functional annotation of transcripts and classification

Functional characterization of assembled sequences was done by performing BlastX (Altschul et al.,1990) 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). GO (Gene Ontology) study was used to classify the functions of the predicted coding sequences. The gene ontology 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. The identification of the transcription factor was done by Blastx using Plant TFDB (http://planttfdb.cbi.pku.edu.cn).

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). (Beier et al., 2017).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 2 SSRs in a compound microsatellite is 100 i.e. the minimum distance between two adjacent SSR markers was set 100 bases.

RESULTS AND DISCUSSION

The total RNA of two root samples along with RIN value more than 7, 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 my one 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 T.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 contig length of 419.45 bp was obtained. The statistics of transcriptome sequencing and assembly generated by Trinity assembler as given (Table 2).

Functional annotation

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 unigenes.(Figures 1 and 2). 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 *Cicer arietinum* (6488) followed by *Medicago truncatula* (3498) and *Trifolium subterraneum* (2119). A total of 2103, 1843, 1754, 1644 contigs were found to be similar to Glycine max, *Cajanus cajan, Trifolium pratense, Mucuna pruriens*, respectively. (Table 3 and supplementary Figure 3).The functionally annotated transcripts (27,884) of *Taverniera cuneifolia* were classified using Blast2GO into three main domains: Biological processes gene ontology, Cellular component gene ontology and Molecular function gene ontology (Table 4 and supplementary Figure 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 (Supplementary Figures 5 and 6).

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 Molecular Function (26,382[38%]), Biological Processes (25,890[37%]) and Cellular Component (17,099[25%]) category (Supplementary Figures 7 and 8). The Molecular functions were the most abundant component of GO terms. Among the 26,382 Molecular functions, the maximum number of contigs i.e. represented “Nucleotide Binding,” followed by “Hydrolase activity” and “Transferase activity”.

In addition, Biological Processes a total of 25,890 transcripts were associated with the “Biosynthetic process” as the highest match followed by “cellular protein modification process” and “nucleo base-containing compound metabolic process” respectively.

A total of 17,099 transcripts were associated with the

Table 1. Summary of the Ion torrent sequencing data generated for two 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. Results based on combined assembly of two root transcriptome.

Sr. No.	Characteristics	Values
1	Total assembled contigs/transcript	35,590
2	GC %	43.25
3	Contig N50 (bp)	441
4	Median Contig length (bp)	322
5	Average Contig length (bp)	419.45
6	Total assembled bases	14,928,144

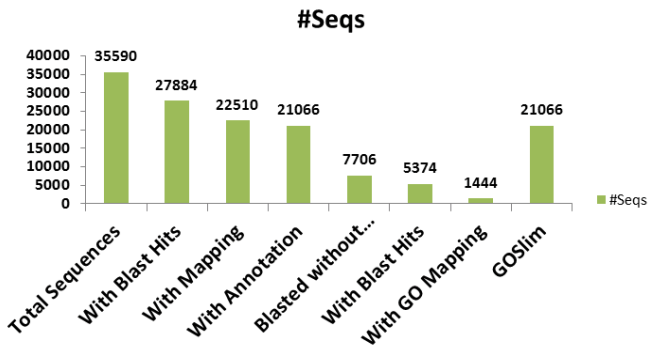

Figure 1. Data ditribution of *Taverniera cuneifolia* subject to functional annotation with Blast.

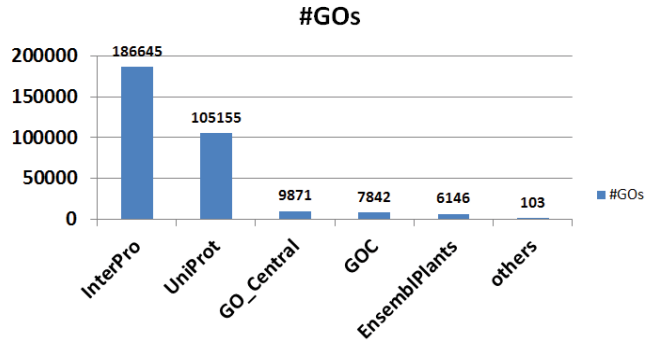

Figure 2. Annotation of transcripts to different database sources.

Table 3. Distribution of top hit species by BLAST.

Species	BLAST Top-Hits
Cicer arietinum	6488
Medicago truncatula	3498
Trifolium subterraneum	2119
Glycine max	2103
Cajanus cajan	1843
Trifolium pratense	1754
Mucuna pruriens	1644
Glycine soja	1344
Lupinus angustifolius	1023
Phaseolus vulgaris	925
Arachis duranensis	638
Vigna angularis	562
Arachis hypogaea	516
Lotus japonicus	426
Vigna radiata var. radiata	393
Others	4381

cellular component and a relatively large no of the transcript was associated with “Membrane” followed by “Nucleus” and “Cytoplasm”, respectively.

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 279 pathways were annotated and among them, metabolic pathways (100), Biosynthesis of secondary metabolites (46), biosynthesis of antibiotics (24) showed the maximum hit with the database (Table 5).

Candidate genes involved in biosynthesis pathway

There were 11 unigenes specific that matched with Glycyrrhiza species which were associated with the Glycyrrhizin biosynthesis from this plant (Table 6). There were 4912 unigenes hypothetical protein predicted from this plant, of which 30 unigenes that had a hit length above 400 were noted (Table 7). 94 unigenes that predicted Cytochrome P450 family protein from this plant, of which 17 unigenes with a hit length above 150 were noted (Table 8).

DISCUSSION

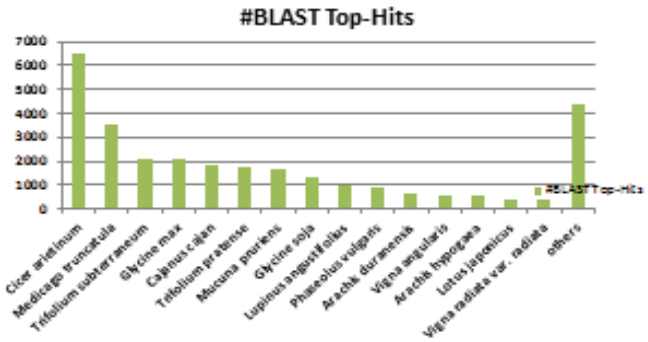

Figure 3. Numbers of transcripts showing similarity with different plant species.

Table 4. GO sequence distribution of biological processes, cellular components and molecular functions using (Blast2Go).

GO term	Process	No. of Transcripts
Biological processes (25,890)	Biosynthetic process	3019
	Cellular protein modification process	2751
	Nucleobase-containing compound metabolic process	2513
	Transport	2385
	Cellular process	2188
	Cellular component organization	1917
	Metabolic process	1288
	Carbohydrate metabolic process	1172
	Catabolic process	1107
	Protein metabolic process	886
	Signal transduction	850
	Response to stress	820
	Response to chemical	767
	Lipid metabolic process	743
	Translation	615
	DNA metabolic process	592
	Generation of precursor metabolites and energy	493
	Response to endogenous stimulus	344
	Cell cycle	284
	Response to abiotic stimulus	224
	Response to external stimulus	221
	Multicellular organism development	197
	Cellular homeostasis	179
	Response to biotic stimulus	169
	Reproduction	166

Cellular components (17,099)	Membrane	5950
	Nucleus	2427
	Cytoplasm	1665
	Plasma membrane	828
	Cytosol	731
	Chloroplast	725
	Mitochondrion	662
	Golgi apparatus	617
	Ribosome	502
	Endoplasmic reticulum	413
	Nucleoplasm	329
	Cytoskeleton	299
	Intracellular	284
	Cell	278
	Vacuole	248
	Extracellular region	210
	Endosome	191
	Thylakoid	171
	Nucleolus	148
	Cell wall	143
	Nuclear envelope	104
	Peroxisome	71
	Cellular component	51
	Plastid	39
	Lysosome	13
Molecular functions (26,382)	Nucleotide binding	4355
	Hydrolase activity	3607
	Transferase activity	2849
	Catalytic activity	2655
	Binding	2344
	Kinase activity	2002
	Protein binding	1515
	DNA binding	1438
	RNA binding	1063
	Transporter activity	1021
	Nucleic acid binding	766
	Structural molecule activity	623
	DNA-binding transcription factor activity	356
	Translation factor activity, RNA binding	351
	Carbohydrate binding	246
	Enzyme regulator activity	237
	Lipid binding	197
	Nuclease activity	194
	Motor activity	166
	Transcription regulator activity	137
	Signaling receptor activity	119
	Chromatin binding	98
	Signaling receptor binding	22
	Molecular function	18
	Translation regulator activity	3
	Oxygen binding	3

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 plants as such.

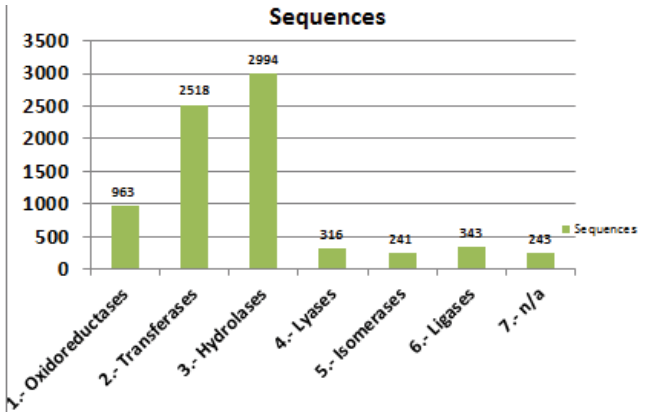

Figure 4. Transcripts linked with enzymes found in KEGG Pathways using Blast2Go.

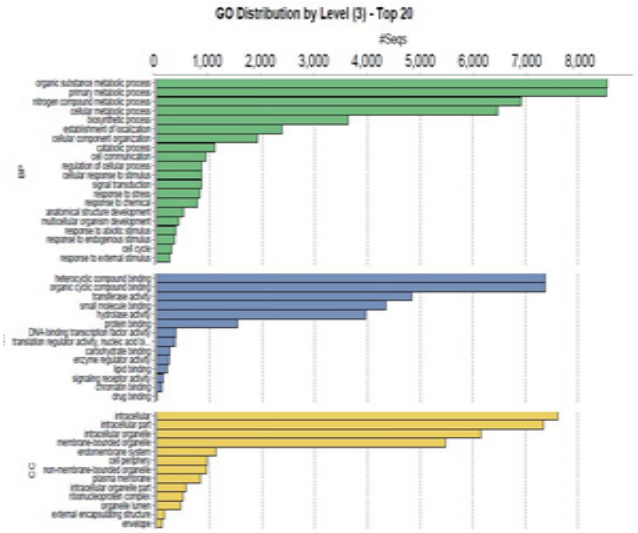

Figure 5. Distribution of transcripts based on Gene Ontology (GO); categorized into cellular component, molecular function and biological process.

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*

RESULTS OF MICROSATELLITE SEARCH

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

Unit size	Number of SSRs
Mono-nucleotide	832
Di-nucleotide	597
Tri-nucleotide	1291
Tetra-nucleotide	153
Penta-nucleotide	33
Hexa-nucleotide	6

Figure 5. Distribution of transcripts based on Gene Ontology (GO); categorized into cellular component, molecular function and biological process.

RESULTS OF MICROSATELLITE SEARCH

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

Unit size	Number of SSRs
Mono-nucleotide	832
Di-nucleotide	597
Tri-nucleotide	1291
Tetra-nucleotide	153
Penta-nucleotide	33
Hexa-nucleotide	6

Figure 6. Results of the microsatellite for *Taverniera cuneifolia.*

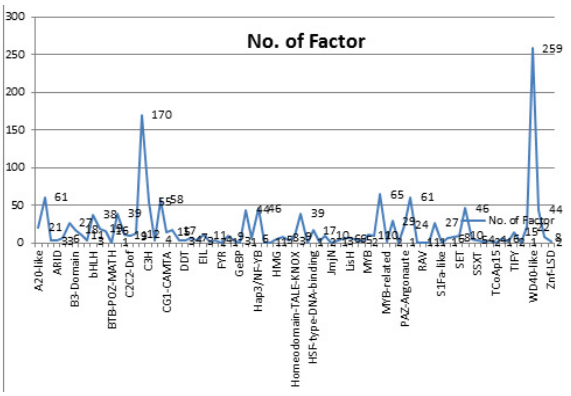

Figure 7. Transcription factors for *Taverniera cuneifolia*.

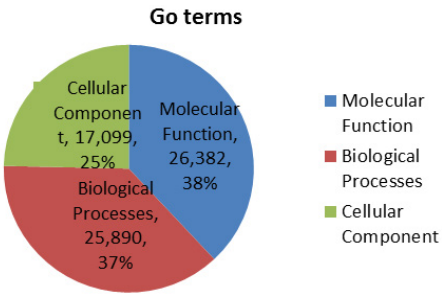

Figure 8. Number of Go terms identified in molecular function, biological processes and cellular components categories**.**

Table 5. Distribution of transcripts to biological pathways using KEGG.

Pathway	No. of enzymes
Metabolic pathways	100
Biosynthesis of secondary metabolites	46
Biosynthesis of antibiotics	24
Microbial metabolism in diverse environments	22
Plant hormone signal transduction	15
Spliceosome	14
Ribosome	13
Endocytosis	13
Carbon metabolism	13
RNA transport	12
Thermogenesis	12
Plant-pathogen interaction	12
MAPK signaling pathway - plant	11
Protein processing in endoplasmic reticulum	10
Ubiquitin mediated proteolysis	10
RNA degradation	10
Biosynthesis of amino acids	10
Amino sugar and nucleotide sugar metabolism	10
Oxidative phosphorylation	9
PI3K-Akt signaling pathway	8
mRNA surveillance pathway	8
FoxO signaling pathway	7
Pyruvate metabolism	7
Glycerolipid metabolism	7
Phagosome	7
Pyrimidine metabolism	5

Table 6. Transcripts/genes that matched with Glycyrrhiza genus from Nr data base in *Taverniera cuneifolia*.

Sr. No.	Transcript ID	Best hit Transcripts associated with Glycyrrhizin biosynthesis pathway in Nr database
1	TRINITY_DN11206_c0_g1_i1	gi|295822111|gb|ADG36709.1| squalene synthase 1
2	TRINITY_DN11206_c0_g1_i2	gi|295822111|gb|ADG36709.1| squalene synthase 1
3	TRINITY_DN20252_c0_g1_i1	gi|403399720|sp|B5BSX1.1|BAMO_GLYUR RecName: Full=Beta-amyrin 11-oxidase; AltName: Full=Cytochrome P450 88D6
4	TRINITY_DN7116_c0_g1_i1	gi|133917969|emb|CAJ77655.1| squalene synthase 2
5	TRINITY_DN11613_c0_g1_i1	gi|550600133|sp|H1A988.1|C7254_GLYUR RecName: Full=11-oxo-beta-amyrin 30-oxidase; AltName: Full=Cytochrome P450 72A154
6	TRINITY_DN11652_c0_g1_i2	gi|838228579|gb|AKM97308.1| cytochrome P450 88D6
7	TRINITY_DN11652_c0_g1_i3	gi|403399720|sp|B5BSX1.1|BAMO_GLYUR RecName: Full=Beta-amyrin 11-oxidase; AltName: Full=Cytochrome P450 88D6
8	TRINITY_DN9998_c0_g1_i1	gi|133917206|emb|CAJ77652.1| squalene synthase 1
9	TRINITY_DN9998_c0_g1_i2	gi|133917206|emb|CAJ77652.1| squalene synthase 1
10	TRINITY_DN9998_c0_g1_i3	gi|253993202|gb|ACT52826.1| squalene synthase 1
11	TRINITY_DN11489_c375_g1_i1	gi|403399720|sp|B5BSX1.1|BAMO_GLYUR RecName: Full=Beta-amyrin 11-oxidase; AltName: Full=Cytochrome P450 88D6

Table 7. 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

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 (P. Sharma, S. Kumar, S. Beriwal, et al, 2019).

The *de novo* assembled transcripts of *T.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 *Cicer arietinum* (18.2%), *Medicago truncatula* (9.8%), *Trifolium*

Table 8. Transcripts/genes that showed the Cytochrome P450 family protein in *Taverniera cuneifolia* with hit length above 150, (total over all 94 Cytochrome P450).

Sr. No.	Transcript ID	Transcripts associated with Glycyrrhizin biosynthesis pathway
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

subterraneum (6%) 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 synthesise glycyrrhizin.as shown in Figure 9 (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 6 different transcripts that we have found in our plants that are associated for the production of squalene given in (Table 6) and then by oxidation by squalene epoxidase (SQE) to 2,3 – oxidosqualene or

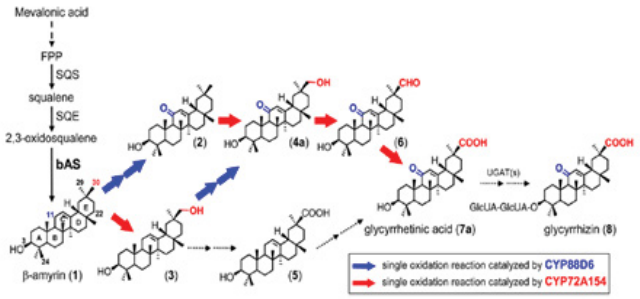

Figure 9. Proposed Glycyrrhizin biosynthesis pathway in liquorice roots by seki, et al 2011.

cyclization catalyzed by bAS i.e β- Amyrin *CYP88D6* given in (Table 6) and one gene catalyzed by *CYP72A154* (11-oxo-β-amyrin) which show a single oxidation reaction a dotted arrow signify undefined oxidation and glycosylation steps in the pathways.

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 2011 - Garg et al 2013). As there have been no proper investigation for the pathway of glycyrrhizin known till today.

We have extensively worked upon the proteins which we have opted from our datas of *Taverniera cuneifolia*. Approx. 4912 gene have been isolated that showed different proteins reported firstly from this plant among them the details have been provided in (Tabel 7) (we have approx. shown only those hypothetical protein whose hit length is above 400) (Maroufi et al 2016 - Rasool et al 2016) In our studies we

Table 9. Top 15 Transcription factors families detection from *Taverniera cuneifolia* root transcriptome.

Sr no.	Transcription factors Family	No of factors from 1557 unigenes
1	WD40-like	259
2	C2H2	170
3	MYB-HB-like	65
4	AP2-EREBP	61
5	PHD	61
6	CCHC(Zn)	58
7	C3H	55
8	Hap3/NF-YB	46
9	SNF2	46
10	GRAS	44
11	WRKY	44
12	bZIP	39
13	Homobox-WOX	39
14	bHLH	38
15	NAM	29
16	Others approx. (70)	505

Table 10. Distribution of SSRs.

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
Single nucleotide	832
Di-nucleotide	597
Tri-nucleotide	1291
Tetra-nucleotide	153
Penta-nucleotide	33
Hexa-nucleotide	6

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 (Li 2014 - Villa-Ruano 2015). Tri-nucleotide represented the maximum numbers of SSRs (1291), followed by Mono-nucleotide (832) and then Di-nucleotide (597). 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 (Tables 9 and 10).

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.

AUTHOR'S CONTRIBUTION

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

REFERENCES

Andrews S(2010). FastQC: A quality control tool for high throughput sequence data.

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.

Kamboj VP (2000). Herbal Medicine. Current Science. 78: 35-39.

Thakar J (1910). Herbal Examination and Use of Bardadungar, Kathiawar, Gujarati Press Publishers, Mumbai

Roskov Y, Bisby A, Zarucchi J, Schrire B, White R(2006). ILDIS World Database of Legumes: Draft checklist, version 10 [published June 2006, but CD shows November 2005 date. ILDIS, Reading, UK.

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

Zore G, WinstonU, Surwase B, Meshram N, Sangle V, Kulkarni S, Karuppayil M(2008). Chemoprofile and bioactivities of *Taverniera cuneifolia* (Roth) Arn: A wild relative and possible substitute of *Glycyrrhiza glabra* L. Phytomdicine. 15(4): 292–300.

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

Mangalorkar(2016). 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.

Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, Regev A(2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7): 644-652.

Haas B, Delcher A, Mount S, Wortman J, Jr R, Hannick L, White O(2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31(19): 5654–5666.

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.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ(1990). Basic local alignment search tool. J Mol Biol. 215: 403–410.

Conesa A, Gotz S, García-gómez J, Terol J, Talón M, Genómica D, Valencia U(2005). Blast2GO : A universal tool for annotation , visualization and analysis in functional genomics research. 21(18): 3674–3676.

Kanehisa M, Goto S (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids Res 28(1): 27–30.

Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M(2007). KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35: W182--W185.

Dai X, Sinharoy S, Udvardi M, Zhao P(2013). PlantTFcat: An online plant transcription factor and transcriptional regulator categorization and analysis tool. BMC Bioinformatics. 14(1).

Beier S, Thiel T, Münch T, Scholz U, Mascher M(2017). MISA-web: A web server for microsatellite prediction. Bioinformatics. 33(16): 2583–2585.

Pagare, Saurabh, Bhatia M, Tripathi N, Pagare, Sonal, Bansal Y(2015). Secondary metabolites of plants and their role: Overview. Curr Trends Biotechnol Pharm. 9:293 –304.

Metzker M(2010). Sequencing technologies: The next generation. Nature Reviews Genetics 11(1): 31–46.

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.

Chen H, Wang L, Wang S, Liu C, Blair M, Cheng X(2015). Transcriptome sequencing of mung bean (*Vigna radiate* L.) genes and the identification of EST-SSR markers. PLoS ONE. 10(4).

Chin Y, JungH, Liu Y, Su B, Castoro J, Keller W, Kinghorn A(2007). Anti-oxidant constituents of the roots and stolons of licorice (*Glycyrrhiza glabra*). J Agri Food Chem. 55(12): 4691–4697.

Li B, Fillmore N, Bai Y, Collins M, Thomson J, Stewart R, Dewey C(2014). Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. Genome Biology. 1–21.

Dutta S, Kumawat G, Singh B, GuptaD, Singh S, Dogra V, Singh N(2011). Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea.

Sharma P, Kumar S, S. Beriwal, Shruti, Sharma, Priyanka, Bhairappanavar, Shivarudrappa B, Verma, Ramtej J, Das, Jayashankar(2020). Comparative transcriptome profiling and co-expression network analysis reveals functionally coordinated genes associated with metabolic processes of *Andrographis paniculata*. Plant Gene.

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.

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. Prep Biochem Biotechnol. 34(3): 209–214.

Amit G, Daniel M(2014). Development of quality standards of *Taverniera cuneifolia* (Roth) arn. root - A substitute drug for liquorice. Int J Pharmacog Phytochem Res. 6(2): 255–259.

Garg R, Jain M(2013). RNA-Seq for transcriptome analysis in non-model plants. Methods Mol Biol. 1069: 43-58.

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).

Chomczynski P, Sacchi N(1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 162(1): 156–159.

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 J. 89(2): 181–194.

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.

Nadiya F, Anjali N, Thomas J, Gangaprasad A, Sabu K(2017). Transcriptome profiling of *Elettaria cardamomum* (L.) Maton (small cardamom). Genomics Data. 11: 102–103.

Li J, Dai X, Zhuang Z, Zhao P(2016). LegumeIP 2.0—A platform for the study of gene function and genome evolution in legumes. Nucleic Acids Res. 44: D1189–D1194.

Anvar SY, Khachatryan L, Vermaat M, Galen M, Van, Pulyakhina I, Ariyurek Y, Laros J(2014). Determining the quality and complexity of next-generation sequencing data without a reference genome. Genome Biol. 15(12): 555.

Chirumbolo S(2016). Commentary: The antiviral and antimicrobial activities of licorice, a widely-used Chinese herb. Front Microbiol. 1–3.

Ramilowski J, Sawai S, Seki H, Mochida K, Yoshida T, Sakurai T, Daub C(2013). *Glycyrrhiza uralensis* transcriptome landscape and study of phytochemicals. Plant Cell Physiol. 54(5): 697–710.

Rasool S, Mohamed R(2016). Plant cytochrome P450s: Nomenclature and involvement in natural product biosynthesis. Protoplasma.

Li B, Fillmore N, Bai Y, Collins M, Thomson J, Stewart R, Dewey C(2014). Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. Genome Biology. 1–21.

Assefa M, Dagne E(2010). Phytochemical Investigation on Roots of *Taverniera Abyssinica. (*Dingetegna)

Li Y, Luo H, Sun C, Song J, Sun Y, Wu Q, Chen S(2010). EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. BMC Genomics. 11(268).

Sudo H, Seki H, Sakurai N, Suzuki H, Shibata D, Toyoda A, Saito K(2009). Expressed sequence tags from rhizomes of *Glycyrrhiza uralensis*. Plant Biotechnol. 26(1): 105–107.

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 .) 106: 411–422.

Gore R, GaikwadS (2015). Checklist of Fabaceae Lindley in Balaghat Ranges of Maharashtra, India. Biodivers Data J. 3: 4541.

Varshney R, Graner A, Sorrells M(2005). Genic microsatellite markers in plants: Features and applications. 23(1).

Varshney R, Song C, Saxena R, Azam S, Yu S, Sharpe A, hang G(2013). Draft genome sequence of chickpea (*Cicer arietinum)* provides a resource for trait improvement. Nat Biotechnol. 31(3): 240–246.

Villa-Ruano N, Pacheco-Hernández Y, Lozoya-Gloria E, Castro-Juárez C, Mosso-Gonzalez C, Ramirez-Garcia S(2015). Cytochrome P450 from Plants: Platforms for valuable phytopharmaceuticals. Trop J Pharmaceut Res. 14(4):731-742.