Comparative transcriptomics reveals candidate transcription factors involved in costunolide biosynthesis in medicinal plant-Saussurea lappa

Vasundhara Thakur, Savita Bains, Shivalika Pathania, Shailesh Sharma, Ravneet Kaur, Kashmir Singh

PII: S0141-8130(19)40217-1

DOI: https://doi.org/10.1016/j.ijbiomac.2020.01.312

Reference: BIOMAC 14648

To appear in: *International Journal of Biological Macromolecules*

Received date: 11 December 2019

Revised date: 28 January 2020

Accepted date: 28 January 2020

Please cite this article as: V. Thakur, S. Bains, S. Pathania, et al., Comparative transcriptomics reveals candidate transcription factors involved in costunolide biosynthesis in medicinal plant-Saussurea lappa, *International Journal of Biological Macromolecules*(2020), https://doi.org/10.1016/j.ijbiomac.2020.01.312

Comparative transcriptomics reveals candidate transcription factors involved in costunolide biosynthesis in medicinal plant- *Saussurea lappa*

Vasundhara Thakur[1#], Savita Bains[1#], Shivalika Pathania, Shailesh Sharma[2], Ravneet Kaur[1], Kashmir Singh[1*]

[1]Department of Biotechnology, Panjab University, Chandigarh, India-160014.

[2]National Institute of Animal Biotechnology (NIAB), Hyderabad, Telangana, India - 500049

# Both Authors contributed equally for this work.

*Corresponding Author: Dr. Kashmir Singh, Associate Professor, Email: kashmirbio@pu.ac.in, kashmir123@gmail.com, Tel: +91-172-2534085

E-mail ID

VT: tkr.vasu93@gmail.com

SB: savita.bains18@gmail.com

SS: haitoshailesh@gmail.com

RK: ravk14@gmail.com

ABSTRACT

Costunolides, an important sesquiterpene lactone (STL) isolated from *Saussurea lappa*, are the major pharmaceutical ingredient of various drug formulations. Identification of the genes and transcriptional regulation of costunolide biosynthesis pathway in *S. lappa* will propose alternatives for engineering enhanced metabolite biosynthesis in plant. Here, we aimed to unravel the transcription factors (TFs) regulating the costunolide biosynthesis. Comparative transcriptome analysis of root and leaf tissues and transcripts were annotated using various *in silico* tools. Putative transcription factors were identified using PlantTFDB and TF- gene co-expression network was generated followed by clustering using module based analysis to observe their coordinated behaviour. The module 1 was found to be significant based on its enrichment with major pathway genes. Further, promoter cloning determined the *cis* acting elements in *costunolide synthase* (*SlCOS1*) gene which catalyses the final key step of costunolide biosynthesis. Bioinformatics tools were employed to predict the cis regulatory elements, leading to the identification of MYB family of TFs as an interacting partner of *SlCOS1* gene. The present study is the pioneer attempt for TF prediction and elucidation of their regulatory role in costunolide synthesis. This will help in future metabolic engineering of the pharmaceutically important STLs and their yield improvement.

Keywords

Saussurea lappa, Transcriptome, Sesquiterpene lactones, Costunolide, Co-expression, Transcription factors

1. INTRODUCTION

Sesquiterpene lactones (STLs) are a major category of biologically active constituents most prevalent among Asteraceae family with around 5000 reported structures. They are a diverse subclass of naturally occurring sesquiterpenoids comprising of lactone group at its C15 backbone which contributes significantly to their sturdy role in various biological activities [1]. The plethora of STLs reported in plants such as *Artemisia annua, Lactuca sativa, Cichorium intybus, Helianthus annuus etc.,* attributes to their allelopathic signalling. STLs are vital components of drugs such as artimisinin, anti-migraine and anti-helminths [2–4]. They also help in relieving pain, cough and insomnia. These compounds are reported to possess remarkable anti-cancerous properties against different carcinomas like blood, breast, prostate, ovarian and liver cancer [5]. Further, costunolide, the simplest STL, has remarkable properties against bladder cancer cell line. It also promotes the loss of viability and apoptosis which leads to subsequent elevated levels of ROS and alterations in membrane potential [6]. Thus, it gained attention of the researcher because of its importance as pharmaceutical component of various drug formulations [7].

S. lappa has tremendous pharmaceutical potential accredited to liberal occurrence of STLs in it. Its dried roots and the essential oil has great economic importance in international drug market [8]. The roots are copious source of STLs such as dehydrocostus lactone and costunolide. Costunolide, first extracted from roots of *S. lappa,* holds promising medicinal value including anti-diabetic, anti-carcinogenic, anti-viral and anti-fungal activities [9]. Its biogenesis occurs mainly through mevalonate pathway, yielding isopentenyl diphosphate (IPP) or dimethylallyl diphosphate (DMAPP), further converted to farnesyl diphosphate (FPP) which is the precursor for sesquiterpenes biosynthesis. It is later acted upon by a series of cyclase, hydroxylases and cytochrome P450 enzymes to yield costunolide[10]. Previous studies shows nine major genes

involved in the synthesis of costunolide, including hydroxymethyl glutaryl-CoA (HMGS), hydroxymethylglutaryl-CoA reductase (HMGR), mevalonate kinase (MK), phosphomevalonate kinase (PMK), diphosphomevalonate decarboxylase (DPD), farnesyl diphosphate synthase (FDP), germacrene A synthase (GAS), germacrene A oxidase (GAO) and costunolide synthase (COS) [10,11]. Costunolide synthase (COS) is a cytochrome P450 mono-oxygenase that catalyze the final and crucial step of lactone ring formation in costunolide skeleton [12] . The yielded costunolide is an important precursor of certain other STLs, such as eudesmanolide, germacranolide and guaianolide, which have been reported to possess immense medicinal properties[13]. Despite its vast economic importance, not many data sources are available to explore the STL biosynthesis mechanism in *S. lappa*. In our previous study we have reported in depth annotation and spatial relative expression of the genes encoding for costunolide synthesis [11]. Here, we focused on an elucidated regulation of *COS* gene that holds a pivotal role in STL biosynthesis, identifying potential candidates through comparative transcriptomic studies and co-expression analysis.

At present, the transcriptional regulation mediated by transcription factors (TFs) in costunolide synthesis has not been reported. Since, TFs are key regulators of numerous biological processes including stress responses and developmental processes in plants, it demands the unravelling of TFs networks regulating genes involved in sesquiterpenoid biosynthesis, that bind to *cis*-acting elements found in the upstream promoter regions of specific genes, thus activating or repressing transcription response [14].

Plant TFs are associated with chief cellular pathways including metabolic processes, plant growth and development and various abiotic and biotic stress responses [15,16]. TFs expression depends on the induction of the target genes, which further involves the identification of the

promoters or regulatory regions [17]. Various bioinformatics tools and algorithms have been created that make predictions on the basis of identification of core promoter elements and others on the basis of their sequence composition [18,19]. Hu *et al* presented a genome wide identification of TFs and associated TFBSs (transcription factor binding sites) by performing characteristic domain analysis in *Nannochloropsis* [20]. Suttipanta et al reported the isolation of *G10H* promoter sequence by DNA walking in geraniol 10-hydroxylase (G10H), a crucial enzyme in TIA biosynthesis in *Catharanthus roseus.* The analysis and characterization of identified promoter was performed resulting in detection of unique binding regions for various TFs [21]. Different studies have reported the role of several TFs such as certain MYB family members along with bHLH factors are involved in flavonoid pathway regulation [22]. Ma *et al* suggested the involvement of a WRKY TF in biosynthesis of a sesquiterpene lactone, artemisinin, in *Artemisia annua* [23]. Further, the co-expression studies will allow the mining of regulatory networks interplaying in many biological and molecular processes. It also aids in annotation of novel genes in plants using various bioinformatics tools on the basis of their identical gene expression patterns [24].

In the previous studies it is well documented that plants have an elaborative regulatory mechanism comprising of positive and negative regulators of gene expression which further markedly influence the development of plant and their defence mechanism against various stress conditions [25,26]. However, studies related to involvement of TFs in *S. lappa* are not done yet. To generate a more comprehensive insight into the gene expression and its regulation in *S. lappa*, a transcriptome assembly was generated. The present study was initiated in order to discover the interlinked active costunolide pathway and associated TFs regulating the processes. Promoter cloning followed by the *in silico* analysis suggested MYB TFs capable of regulating *COS* gene

expression in sesquiterpenoid biosynthesis. The systematic analysis will help in further investigation of the biological pathways related to important secondary metabolites isolated from *S. lappa* in order to enhance the properties and yield of the plant.

2. MATERIAL AND METHODS

2.1 Plant sample collection and RNA isolation

The plant samples were obtained from Kullu, HP, India (31.9579° N, 77.1095° E). These plants were maintained in plant tissue culture facility of the Department of Biotechnology, Panjab University, Chandigarh (30.7333° N, 76.7794° E). The leaves and root tissues were snap frozen in liquid nitrogen and preserved for further use at -80°C. The RNA isolation was performed using young leaves and root tissues [27]. The RNA samples were purified and treated on column DNase I with miRNAeasy mini kit (Qiagen, Germany) in order to remove DNA contamination. The RNA purity and concentration was assessed by taking the absorbance (A260/A280) using BioSpectrometer® (Eppendorf, USA). Thus, obtained DNA free RNA sample were quantified and assessed for purity and were then sent for transcriptome sequencing to AgriGenome Labs Pvt. Ltd., Kochi, India.

2.2 Library construction and *de novo* transcriptome assembly

The cDNA sequencing libraries were generated using TruSeq^{TM} cDNA synthesis kit for illumina following the manufacturer's protocol. The libraries were sequenced on Illumina HiSeq 2500 platform providing paired end reads. Six transcriptome libraries were generated from leaf and root tissues (3 biological replicates each). The raw reads retrieved from the sequencer were carried forward for low quality reads filtration (Q<30) using sickle v1.3.3 and adapter sequences were removed using Cutadapt v1.8. The trimmed raw reads (wherever necessary) were further

refined by the contamination removal i.e. non-coding RNAs and mitochondrial RNAs elimination using Bowtie2 (version 2.2.6) and in-house built Perl scripts. The cleaned reads were then assembled into transcripts using Trinity program. The over-lapping high quality transcripts were further clustered to generate unigenes using UniGene module of Trinity.

2.3 *In-silico* annotation and similarity analysis

The assembled transcripts were submitted to Basic local alignment search tool (BLAST) for similarity search against various databases such as non-redundant (NR) protein database, Protein Family (Pfam), InterPro, UniProt and PROSITE. Further, Gene ontology (GO) enrichment was performed on resulting unigenes and biological, molecular and cellular categories were allotted to them using Blast2GO software. The KO IDs were assigned to the transcripts and were mapped to respective metabolic pathways using KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway mapping scheme. All the assembled unigenes were anticipated against the plant transcription factor database (PlantTFDB 4.0) (http://planttfdb.cbi.pku.edu.cn/download.php) to identify the transcription factors (TFs) represented in the transcriptome using BLASTX with an E-value cut-off of 1e-05. The transcription factors unigenes were classified into various plant TF families on the basis of conserved domains predicted in the above analysis.

2.4 Differential genes expression (DGE) analysis

The raw reads from root and leaf transcriptome data were used to calculate the transcript level abundance using RSEM (RNA-Seq by Expectation-Maximization) software from Trinity v-2.4.0 package [28]. The reads (for each tissue in triplicates) were aligned to the assembled transcriptome and read count profile was acquired using Bowtie-2, followed by transcript abundance estimation. A gene-specific count matrix was generated across the samples and

normalized expression values matrix were generated as FPKM (fragments per kilobase transcript length per million mapped reads) values for downstream analysis. The differentially expressed genes (DEGs) were identified with the help of EdgeR of Trinity package that performs TMM (trimmed mean of M-values) normalization scaling any sample differences. To acquire the differentially expressed genes, a cut off *P*-value of 0.01 and two-fold change in the expression values was used. This was done for transcripts encoding for putative TFs and costunolide pathway genes. To generate heat maps for visualizing comparative expression in the two tissues, Hierarchical Clustering Explorer 3.5 (http://www.cs.umd.edu/hcil/hce/) was used. The GO enrichment analysis was carried out for DEGs using Blast2GO software.

2.5 Network Construction and threshold selection

TF-gene regulatory network analysis was performed to determine the functionality of putative TFs on the basis of similarity pattern of expression across interacting partners under different tissue-specific conditions for costunolide pathway. The expression values (in the form of FPKM (fragments per kilobase of transcript per million fragments mapped)) obtained in earlier steps for all TFs and gene involved in costunolide pathway (HMGS, HMGR, MK, PMK, DPD1, DPD2, FDP1, FDP2, GAS1, GAS2, GAO, COS1, and COS2), available for two conditions (root and leaf), were used to calculate Pearson correlation coefficient (PCC) [29] using PERL script [30]. Here, DPD1 and DPD2, FDP1, and FDP2, GAS1 and GAS2, as well as COS1 and COS2 were identified as the isoforms of costunolide pathway genes DPD, FDP, GAS, and COS, respectively. While calculating the PCC, mean of expression values were determined by sample size which is different for both TFs and pathway genes.

PCC was computed based on the following equation:

$$PCC = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{(x - \bar{x})^2}\sqrt{(y - \bar{y})^2}}$$

Where x and $\bar{x}$ as well as y and $\bar{y}$ represents expression data and the corresponding mean of pathway genes and TFs, respectively. Mean of expression profile data is determined by sample size which is different for both pathway genes and TFs.

In order to construct TF-gene network, biologically significant PCC threshold was determined by comparing various topological parameters of network such as number of nodes, edges, and network density (ND) against different PCC cutoffs [31]. ND was calculated as follows:

$$ND = \frac{2m}{n(n-1)}$$

Where m and $\frac{n(n-1)}{2}$ specify the number of observed edges and possible links of nodes, respectively.

To further reduce the rate of false positive, "t-test" was carried out to select statistically significant interacting pairs with *p* value ≤0.05 [32] followed by multiple testing correction using false discovery rate (FDR) [33] strategy using "stats" library of R statistical package (http://www.r-project.org/). The resultant co-expressed TF-genes pairs obtained during successive filtering steps were shortlisted at defined PCC threshold and a weighted co-expression network was constructed and visualized by Cytoscape v 3.7.0 (http://www.cytoscape.org/). Correlation value for each TF-gene pair was used as edge weights to construct TF-gene network, where weights constitute the strength of correlation or co-expression.

To decipher the biological nature of weighted TF-gene network, random network with same number of nodes and edges was constructed using "igraph" library in R package [34], and both networks were compared for their topological parameters to illustrate their biological behaviour.

2.6 Network partitioning and Enrichment Analysis

In order to reduce the complex analysis of whole network, fragmentization of the TF-gene network was performed to obtain highly interconnected but comparatively smaller and functional gene modules using Markov cluster (MCL) algorithm [35]. This systematic network clustering algorithm works on the basis of "random walks" which states that starting with a node randomly travel to connected nodes will make to stay within a cluster than travel between, i.e. the flow is easier within dense regions than across sparse boundaries. The clusterMaker2 [36] plugin of Cytoscape was used to perform clustering through MCL algorithm at default inflation value of 2.5. Enrichment analysis of all modules was carried out with singular enrichment analysis (SEA) of agriGO web-based tool (http://bioinfo.cau.edu.cn/agriGO/index.php) [37], and significantly enriched GO terms were determined in all comparative conditions by comparing it against background reference. Bonferroni correction [38] was carried out with hypergeometric test [39] for the selection of statistically significant terms. Database for Annotation, Visualization and Integrated Discovery (DAVID) v 6.7 [40], a web-based program, was used to carry out KEGG pathway analysis of significant modules. Expression profile data of three replicates for each condition (root and leaf) was used to create heat maps using "gplots" library of R package.

2.7 Isolation and identification of promoter regions in *costunolide synthase* (*SlCOS*)

The crucial key step of costunolide biosynthetic pathway involving lactonization is catalyzed by costunolide synthase, thus our study focused on its transcriptional regulation. The young leaves were used to isolate total genomic DNA with modified version of cetyl trimethyl ammonium bromide (CTAB) method [41] . The concentration and quality was analysed on 0.8 % agarose gel and BioSpectrometer® (Eppendorf, USA). Genome walking method based on PCR was performed using GenomeWalker Universal Kit (TAKARA) to isolate the promoter regions of

SlCOS gene. Genome walking libraries were constructed following the manufacturer's manual. The genomic DNA was digested using different restriction enzymes (EcoRV, DraI, PvuII and StuI) in separate reactions and the products ligated to Genome Walker adaptors. Two rounds of nested PCR amplifications were carried out using obtained products as the templates. Adaptor specific primer, AP1 (provided in the kit) and gene specific primer, GSP1 (designed according to the manufacturer's protocol) were used to carry out the primary PCR reaction. This was followed by secondary reaction involving AP2 (nested adaptor primer) and GSP2 (nested gene primer) yielding nested PCR product. This was cloned in pGEM-T Easy vector (Promega, USA) and the promoter region was sequenced by AgriGenome Labs. The sequenced region was extracted and used to analyze the *cis*–acting elements found in the promoter using PlantCARE [42] and PLACE.

2.8 Identification of associated TF family genes

The analysis revealed the presence of various binding sites for MYB transcription factor, suggesting its involvement in transcriptional regulation of *COS* gene expression. The putative MYB family sequences were subjected to ESTScan to recognize the coding regions in the assembled transcripts. The hidden Markov model profile for DNA-binding domain of MYB TFs was obtained from PFAM database (pfam00249) and used to predict putative MYB TFs with significant hits against the HMM profile. DGE analysis was performed and the acquired read count of MYB TFs in root and leaf was normalized by determining their FPKM values. HCE3.5 (Hierarchical Clustering Explorer) was used to conduct hierarchical clustering analysis.

2.9 Sequence alignment and phylogenetic analysis of SlMYB

To study the evolutionary relationship among the putative MYB TFs co-expressing with *SlCOS1* gene, multiple sequence alignment was carried out using Clustal omega [43] and MUSCLE (MUltiple Sequence Comparison by Log- Expectation). BLASTP search for MYB proteins was carried out against the NR database to obtain their corresponding high similarity proteins in other species. For each MYB protein sequence in *S. lappa*, top hits were extracted and after removing the redundant sequences, a phylogenetic tree was generated with the help of MEGA7 using 1000 bootstrap value and neighbour-joining method. A total of 65 sequences were used to perform phylogenetic analysis of MYB family TFs.

2.10 Protein structure and physiochemical analysis

Analysis of conserved domains in the predicted MYB family proteins was carried out using Pfam database (https://pfam.xfam.org/), SMART (http://smart.embl-heidelberg.de/), InterProScan (https://www.ebi.ac.uk/interpro/search/sequence-search) and Conserved Domains Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). For the identification of conserved motifs among the MYB proteins, MEME suite (http://meme-suite.org/tools/meme) was employed with default parameters [44].Various physical and chemical characteristics of the identified MYBs, such as isoelectric point (pI) (theoretical), molecular weight, amino acid composition and instability index, were computed using ProtParam tool (https://web.expasy.org/protparam/) [45]. Also, the sub-cellular localization of the proteins was predicted with the help of ProtComp v. 9.0 server (http://www.softberry.com).

3. RESULTS

3.1 Illumina based sequencing and *de novo* assembly

The number of raw reads obtained after sequencing of root samples (three biological replicates) were 145, 88 and 91million with mean read length of 100bp whereas after read cleaning, 134, 82 and 83 million high-quality reads were obtained, respectively. Similarly in leaf tissue (three biological replicates) the raw reads generated were 210, 130 and 62 million (mean read length 100bp). When the reads were put through the quality filtering, the number reduced to 195, 120 and 40 million, respectively (Table S1). The assembled clean reads using Trinity assembler (default parameters) resulted in 388,941 transcripts, further clustered into 162,539 unigenes based on sequence similarity, in order to reduce the redundancy and remove partial transcripts. The length of transcripts were more than 200bp (Fig. 1A) and mean GC content of 39.36%.

3.2 Functional annotation of the transcripts

The unigenes were annotated against NR and UniProt database using BLASTX program. 70,947 transcripts could find significant hit in UniProt database whereas the remaining 91,592 transcripts found match in nr database. Overall, we found that 73,532 transcripts were found to have at least one hit in UniProt or NR protein database. Only the transcripts showing similarity score >=40% and e-value <=10^{-5} were further retained for annotation that could establish the homology. The top BLASTX hits for each transcript were studied and *Cynara cardunculus* was found to be the most similar organism (Fig. S1). A sum of 27,121 (16.69%), 28640 (17.62%) and 20086 (12.35%) unigenes were annotated against PFAM, InterPro, PROSITE database, respectively (Fig. 1B). 89,008 transcripts had no significant BLAST hit which may be due to the less information available pertaining to them. The BLASTX analysis produced around 36.5% of the transcripts which were able to recognize a homolog with minimum E-value confidence of 1e-5 (Fig. 1C), thus indicating significant conservation level of the proteins. A similarity score of

more than 60% was observed among 85% of the transcripts identifying homologs with at protein level with the existing proteins (Fig. 1D).

3.3 Identification and classification of transcription factors

TFs mediate gene expression and cellular mechanism, regulated by recognizing the sequence specific *cis*-regulatory elements enriched in the target gene promoter region. Several metabolic processes in plants are coordinated by transcription factors involved for controlled gene expression. The assembled 162,539 unigenes were annotated against known TF families for homology search, identified a total of 30,070 putative TF transcripts in the data and classified into 58 families with reference to the PlantTFDB (Fig. S2). *Cynara cardunculus* was one of the top blast hit species. It represented 18.5% of *S. lappa* transcripts with an average of 1756 bp and a range of sequence lengths observed from 201 to 21,992 bp. The most abundant families were MYB followed by bHLH, ERF, NAC, WRKY and C2H2 (Fig. 2A). Many families including NAC, bHLH, MYB, WRKY, ERF and C2H2 have been reportedly involved in biosynthesis of secondary metabolites [46,47]. MYB family proteins are associated with secondary metabolism like flavonoid and carotenoid biosynthetic pathway, developmental processes and biotic & abiotic stresses [26,48]. bHLH family proteins are also involved in flavonoid and anthocyanin biosynthesis, along with abiotic and biotic stress [49].

3.4 GO enrichment and KEGG pathway mapping

To functionally classify the unigenes, a total of 2,781 GO terms were assigned to biological process (BP), 1,935 to molecular function (MF) and 716 to cellular component (CC). GO terms were assigned to 59.24% and grouped them into BP, MF and CC with 14, 24 and 16 functional

subcategories, respectively (Fig. S3). Molecular function category being the most abundantly represented about 21,825 transcripts (46%), followed by biological processes being given to 12,523 transcripts and cellular component category with 11,527 transcripts. The transcription factor unigenes were also annotated by GO enrichment and the maximum are depicted in Fig. 3A. The KEGG distribution of all annotated transcripts majorly predicts the functional involvement of transcripts in metabolism (Fig. S4). The transcripts annotated by KEGG for various TFs were mostly the part of metabolic pathways, signalling pathways and biosynthesis of secondary metabolism (Fig. 3B).

The transcripts encoding for mevalonate as well as non-mevalonate pathway were assigned KO IDs and were mapped to terpenoid backbone. The transcripts catalysing various steps towards generation of IPP and DMAPP through mevalonate pathway include acetyl-CoA C-acetyltransferase, hydroxymethyl glutaryl-CoA synthase, hydroxymethylglutaryl-CoA reductase, mevalonate kinase, phosphomevalonate kinase, diphosphomevalonate decarboxylase, isopentenyl-diphosphate Delta-isomerase and farnesyl diphosphate synthase. The relative expression of costunolide pathway genes identified is depicted in Fig. S5.

Similarly 5 KO IDs were assigned to transcripts playing role in sesquiterpenoid and triterpenoid biosynthesis, for examples squalene monooxygenase, (-)-germacrene A synthase , germacrene A oxidase, and costunolide synthase could be mapped to KEGG pathway scheme.

3.5 Tissue-specific differential expression analysis

FPKM method was used to numerate the expression level of the transcripts of root and leaf tissues from three biological replicates. The number of transcripts having FPKM >=1.0 in case of root and leaf tissue was determined to be 31,965 and 28,980, respectively. The differential

FPKM distribution in root and leaf tissue is provided Table S2. After DGE analysis, it was found that 2,801 transcripts are down-regulated, 3,160 transcripts are up-regulated and 156,578 transcripts are not significant in root versus leaf differential gene expression analysis.

Transcription factors regulate gene expression by binding to the upstream promoter region of a gene. Thus, the regulation of genes in various pathways is reportedly modulated by the expression of TFs in plants. Around 6.5% of TFs transcripts i.e. 1,973 sequences were observed to be differentially expressed (Fig. 4). These DEGs were found to be distributed to 53 TF families, and distribution of top 10 TF families is depicted in Fig. 2B. The TF families like bHLH, ERF, NAC, MYB_related and WRKY were found to be highly up-regulated in leaf tissue as compared to the root (Fig. 5). From the resulting data, a total of 976 unigenes were differentially expressed in leaf as well as root tissue while 241 and 756 unigenes were noted to be differentially expressing only in root and only in leaf tissue, respectively. Among these, 30 and 39 DEGs from leaf and root tissue only, respectively, were found to be specific TF transcripts regulating certain signalling and metabolic pathways (Table S3). Nine transcripts encoding NAC, bHLH, MYB, G2-like and E2F/DP TFs regulating the secondary metabolites biosynthesis were observed to be specifically expressed in leaf tissue.

The DEGs were annotated by GO analysis, classifying them into the three categories i.e. molecular function (1,436 sequences), biological process (716 sequences) and cellular component (840 sequences) (Fig. S6). Most of the transcripts were found in ATP binding (329 sequences), DNA binding (209 sequences), structural constituent of ribosome (205 sequences) and nucleic acid binding (96 sequences) subcategories enriched in molecular function category.

3.6 Network Topology Based Threshold Selection

Since simple annotation was unable to determine the tissue-specific functionality of putative TFs in costunolide biosynthesis, thus their congruity against pathway genes was calculated in terms of PCC for three replicates of each root and leaf tissue (Table S4). Various studies [50,51] have shown that the number of samples used in this study is sufficient to carry out network-based functional annotation. The PCC computed in earlier step was used to create a correlation matrix of expression profile data from which weighted co-expression network was constructed [29]. A total of 388,713 TF-gene pairs were obtained for 13 pathway genes and 30,070 TFs, and pairs with positive correlation were considered to largely understand the up-regulation of genes associated with costunolide biosynthesis under control of these TFs. Network Density (ND) represents the fraction of possible connections which appears among the nodes of a network that further highlights its scarcity or density. ND has also been used for selecting a statistically significant threshold since biologically relevant modules are known to be predicted at this cut-off [31,52]. To determine the PCC threshold, change in number of edges, nodes, and network density were investigated as a function of different PCC cut-off values. As the cut-off value escalated, edge and node number, as well as the actual number and all possible edges were declined. However, decreasing rate of edges became slower than that of all possible edges and nodes at significantly high cut-off values, which assists to an increase in ND. ND was at its minimum at 0.80 cut-off (Fig. 6A), but contrarily PCC threshold was adjusted at 0.75 to preserve the significant number of connections in network construction and thereby, leads to minimal loss of biological information. TF-gene pairs at and above this threshold (0.75) were evaluated to obtain biologically significant modules. These TF-gene connections (56,801) were further subjected to statistical evaluation, and were found to be significant (p-value ≤ 0.05) at and above

the threshold of 0.75 (Table S5). This assessment criterion ensures minimum number of false positives that helps to strengthen TF-gene pair regularity information.

3.7 Construction of Weighted Network and Module Detection

Weighted co-expression network was created to interpret system-level understanding of costunolide biosynthesis under the regulation of putative TFs. Such weighted networks often employs calculated values of PCC as attribute that quantifies robustness of connections among nodes to determine biologically relevant modules in various organisms [31,52]. The TF-gene co-expression network was generated, and was constituted of 24,623 nodes (including unique 13 genes and 24,610 TFs) as well as 56,801 edges representing genes and connections between them, respectively. All putative 24,610 TFs were grouped to 58 TF families, based on the BLAST search against PlanTFDB, being MYB (MYB & MYB-related), bHLH, ERF, NAC, and WRKY at top 5 positions in terms of their count. Scale-free behaviour of networks are often determined by comparing the biological networks against random network on the basis of distinct topological properties [53]. Therefore, a weighted random network was created with same number of nodes, as of actual TF-gene co-expression network, by generating random connection among edges for 10,000 iterations. The degree distribution (DD) of TF-gene network was highly skewed (Fig. 6B) as compared to bell-shaped pattern of distribution in random network that represents the similar average degree for majority of nodes in random graphs and thereby, satisfies scale-free nature of biological networks [54]. DD of TF-gene network also fitted to the power law (Fig. S7) since there was no apex around average degree, and on the contrary graph rather sloped sharply downwards with increase in degree indicating the presence of large number of nodes with very few connections oppressed by some immensely connected ones. As compared to random network (0.0014), negative value of assortativity (-0.81) also

augmented the scale-free behaviour of TF-gene network [55]. This affirmed that TF-gene network had scale-free topology and robust against random perturbations, as such networks are most likely to hit a node with only few neighbours and disrupt only a small portion [56]**.**

Module-based analysis was carried out to determine the TFs that may work in an integrative manner to estimate the effect of a group of genes having common features to control specific pathway. Also, identification of gene clusters is a better choice to reveal the coordinated behaviour of various TF families and genes since each module constitutes appropriate interpreting unit that occupies certain set of TFs shared by specific genes to understand the complex biological system [57]. Graph clustering algorithms can be used to generate sub-networks based on integrity of network connections using top-down approach (or non-targeted approach) [31]. Network clustering has been carried out for module detection through the active implementation of MCL algorithm in various studies [30,52]. This clustering algorithm divided the large network into 6 sub-networks to simplify the analysis of weighted TF-gene co-expression network at default inflation value, and typically assigned by their number henceforth. These six modules (1–6) were large in size with 8,077, 6,493, 4,761, 2,540, 2,123, and 629 nodes including 4 (*GAS1*, *GAO*, *COS1*, *DPD1*), 3 (*DPD2*, *PMK*, *HMGS*), 2 (*HMGR*, *FDP1*), 2 (*FDP2*, *GAS2*), 1 (*COS2*), and 1 (*MK*) genes from costunolide biosynthesis pathway, respectively. Since, we are focusing on the synthesis of costunolide which is reported to occur in roots [12], expression profiles for each module were generated and visualized in terms of heat maps. Module 1 (Fig. 7), 4 (Fig. S8A), and 5 (Fig. S8B) as well as Module 2 (Fig. S9A), 3 (Fig. S9B), and 6 (Fig. S9C) were found to be highly expressed in roots and leaves, respectively. Additionally, out of all the modules obtained to be expressed in roots, module 1 was considered to be significant since all the key genes (*GAS*, *GAO* and *COS*) known to be involved in

costunolide synthesis [12] were present together in this module, and considered for further understanding of regulatory mechanisms in costunolide biosynthesis.

3.8 Enrichment Analysis Significant Module 1

Module 1 was observed to be significant on the basis of tissue-specific expression in roots and inclusion of key genes involved in costunolide biosynthesis. Important pathway genes (*DPD2*, *GAS1*, *GAO*, and *COS1*) were found to be integrated to multiple TF families and visualized using cytoscape (Fig. 8), with maximum number of MYB-group (MYB-related (1,879) and MYB (1,123)) (Fig. 2A). Enrichment analysis for given module 1 was carried out (Fig. S10), and was found to be associated with following significant BP terms like response to jasmonic acid stimulus, response to abscisic acid stimulus, cellular response to salicylic acid stimulus, salicylic acid mediated signalling pathway, hormone-mediated signalling pathway, response to cold, response to salt, aromatic compound biosynthetic process, response to blue light, oxygen and reactive oxygen species metabolic process, secondary metabolic process, flavonoid metabolic process, phenylpropanoid metabolic process, post-translational protein modification, and regulation of transcription, DNA dependent. Terms associated with hormone-mediated signaling has already been reported to be involved in secondary metabolite production of various chemical classes [58,59]. Similarly, terms like response to jasmonic acid stimulus and salicylic acid mediated signalling pathway has already been known for its involvement in costunolide synthesis [60,61], which further represents its significance in costunolide biosynthetic pathway. Additionally, the GO term ‘oxygen and reactive oxygen species metabolic process’ presented that costunolide synthesis is integrated to high reactive oxygen species (ROS) production.

Similarly, the significant terms under MF category (Fig. S11) were observed to associate with transcription factor activity, sequence-specific DNA binding, oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor, hydrolase activity, and phosphotransferase activity, alcohol group as acceptor. Significant presence of these terms further complemented the involvement of module 1 in costunolide synthesis in *S. lappa* as *COS* gene is already reported for NADPH dependent synthesis of costunolide in different plants [60]. Significant terms like cytosol, cytosolic ribosome, ribosome, nucleus and many more (Fig. S12) were obtained under CC category. Since the location of the key genes (*GAS*, *GAO*, and *COS*) is also known to be reported in cytosol and endoplasmic reticulum [62], it further compliment the involvement of module 1 in costunolide synthesis. Pathway analysis was also performed to determine significant pathways (Biosynthesis of antibiotics, Carbon metabolism, Glyoxylate and dicarboxylate metabolism, Lysine degradation) (Fig. 9), which were found to be associated with synthesis of isoprenoids as well as mevalonate pathway (https://www.genome.jp/kegg/). Interestingly, these results revealed that this module may constitute the group of genes and integrated TFs which are importantly involved in systematic regulation of costunolide biosynthesis in *S. lappa*.

3.9 Promoter isolation of *COS* gene and *in silico* analysis

Costunolide synthase (COS) hydroxylates germacrene A acid from the previous step in costunolide biosynthesis to yield costunolide by the cyclization of hydroxyl group at C6 and carboxylic group at C12 positions [13]. A genome walker library was made for the promoter cloning of *SlCOS1* gene and isolated a sequence of 1,417 bp upstream of the gene. The UTR region was identified by comparing against the transcriptomic data and predicted the transcription start site (TSS) 30 bp upstream of the start codon. PlantCARE resulted in

identification of 32 upstream sites and PLACE gave around 25 distinct cis-acting regulatory sites. The TATA-box was found 115 bp upstream of the UTR region and 22 CAAT sequences also located upstream of the TATA box.

Predicted cis regulated elements (Table 1) mediated the gene expression of the costunolide biosynthesis pathway genes. From the distal region, CCAAT-box is found which is a stress responsive element and binding site for MYBHv1. MSA-like site located at -271 bp is another cis acting element indulged in cell cycle regulation. Another CGTCA-motif is found acting as a cis regulatory element for MeJA-responsiveness. Certain cis acting regulatory elements like ACE, G-box and GATA motif are also present in the upstream region of the gene. Important MYB cis binding elements including BOXLCOREDCPAL, EECCRCAH1, MYB1AT, MYB2CONSENSUSAT, MYBCORE, MYBGAHV and MYBPLANT were found which proposes the involvement of MYB proteins in the regulation of *SlCOS1* gene. Other motifs identified using PLACE were corresponding to WRKY and Dof binding regions. MYB TFs generally have their role in various biological processes like phenylpropanoid metabolism, plant defence, biotic and abiotic stresses, hormone responses etc and it was predicted to be the potential TF regulating *COS* gene.

3.10 Phylogenetic analysis of putative MYB proteins

Putative MYB transcripts from the transcriptome data were translated and their ESTs were obtained. These were around 2,988 in number, out of which 28 were expressed only in root while 73 were specific to the leaf tissue. These were searched against the publically available EST sequences of MYB domains in Pfam database with help of HMM search. This resulted in identification of 241 unique transcript sequences showing match with MYB domains. The

heatmap between 241 MYB transcripts and *SlCOS1* gene revealed three clusters on basis of PCC, with 138 MYB transcripts appearing in cluster 1 along with *SlCOS1* gene (Fig. 10).

Amongst these, the 18 putative MYB transcripts co-expressing with *SlCOS1* were identified and subsequently analysed with the help of a phylogenetic tree, grouping together 65 total sequences into homologous pairs (Fig. 11). Most of the sequences shared homology with *Arabidopsis*, *Cynara, Vitis, Helianthus* and *Lactuca*, suggesting the presence of a common ancestor. However, some MYBs couldn't relate homology to the sequences of other plants possibly reasoning to their divergence in ancestry.

The eighteen members were clustered into 2 classes i.e. R1 and R2R3 family of MYB proteins with eight and ten members, respectively; further classifying the R2R3-MYBs into previously elucidated subgroups (SG) in *Arabidopsis* [63]. The defined subgroups included SG6, SG8, SG13.b, SG17.a, SG20, SG22, SG23, SG31 and SG37. Out of 10 R2R3 MYBs, 5 MYB transcripts (TRINITY_DN83927_c2_g1, TRINITY_DN82635_c1_g5, TRINITY_DN81434_c0_g1, TRINITY_DN86897_c2_g2 and TRINITY_DN89446_c1_g1) shared sequence 94%, 96%, 95.5%, 92% and 91% identity with *Cynara* and *Vitis* R2R3-MYB proteins regulating certain biosynthetic pathways. R2R3-type is the largest class of MYB family and generally involved in metabolic synthesis pathways. TRINITY_DN81434_c0_g1 showed close homology with MYB90 like and MYB113 like proteins which are reported in anthocyanin biosynthesis [64]. TRINITY_DN86897_c2_g2 was found to be related to MYB124 and MYB88, reported to regulate cold-responsive expression of genes in apple [22]. Another transcript (TRINITY_DN82208_c3_g5) was clustered with AtMYB7 transcription factor, which is also R2R3 type of MYB and reported their involvement in phenylpropanoid biosynthesis gene expression and negative regulation of flavonol biosynthesis [65].

In addition, 5 transcripts (TRINITY_DN88143_c2_g2, TRINITY_DN84554_c0_g7, TRINITY_DN86599_c2_g6, TRINITY_DN83137_c2_g1 and TRINITY_DN83393_c0_g1) showed homology with CcrdMYB proteins i.e. MYB59-like, MYB73-like, MYBS3, DIVARICATA-like and SRM1 like transcription factors, respectively. These are responsible for controlling stress related factors and plant growth & development [66–68]. Amongst these, the first two were found to be a part of R2R3 family while the last three were R1 type MYB TFs. TRINITY_DN88143_c2_g2, shows 49% and 50% sequence identity with *Arabidopsis* protein AtMYB48 and AtMYB50, suggesting that TRINITY_DN88143_c2_g2 might have been acquired from extensively allocated R2R3 *MYB* genes paralog. Further, the expression analysis resulted in identification of 12 out of the 18 genes, four from R1 and eight from R2R3 family, follow similar gene expression in root and leaf tissue.

3.11 Protein structure and physiochemical properties

The domain architecture was analyzed and revealed the presence of MYB-like DNA-binding domains (pfam00249 and pfam13921) in 13 sequences and one pfam00249 domain in three sequences (TRINITY_DN90031_c1_g1, TRINITY_DN86559_c2_g6 and TRINITY_DN90224_c8_g1), while no MYB specific domains detected in two sequences (TRINITY_DN86237_c0_g2 and TRINITY_DN82539_c0_g3). These identified domains i.e. pfam00249 of ~49 aa residues and pfam13921 of ~64 aa residues are the signature domains of MYB TF family proteins and found in all the other related species too [69]. Fig. 12A represents the homology analysis among the 16 MYB TFs with conserved domains, grouping together the two families i.e. R1 an R2R3 type MYBs.

The identified MYB TFs were observed to possess an average length of 363 amino acid residues and average molecular weight of 40.6 kDa. The length of the MYBs varied from 187 to 953 amino acid residues, with the molecular weight ranging from 21 kDa to 105 kDa. An average pI (physical index) of 7.25 was observed, with minimum and maximum values being 4.64 and 9.54 (Table S6). The MYBs showed mixed nature i.e. both acidic and basic TFs based on the pI predictions. The instability index ranged from 38.78 to 59.43 i.e. minimum and maximum values with an average of 49.68. Most of the proteins were observed to be unstable according to instability index, stating their instability during *in vitro* conditions.

Sub-cellular localization studies using ProtComp software revealed that among 10 R2R3 MYB TFs, seven were localized in the nucleus while rest of the three were extracellular proteins. Amongst eight R1 family proteins, four were nucleus localized and were characterized from BLAST analysis, while other four were uncharacterized extracellular proteins (Table S6).

Conserved motifs were further identified in the 16 MYBs with the help of MEME suite, obtaining a total of 15 motifs widely distributed among the sequences (Fig. 12B). The signature MYB type HTH DNA binding sequences included motifs 1-4 ascertaining the significance of identified proteins as MYB TFs. Motif 2 was found to be occurring among almost all the sequences while rest of the motifs were exclusively present in combinations in R1 and R2R3 classes of MYB proteins. Motif 1 was present in all the R2R3 proteins along with other motifs 3, 5, 6, 10 and 11 scattered amongst them. In case of R1 family of MYB proteins, motif 4, 7, 8, 9, 11, 14 and15 were found distributed among them. The distribution of motifs was observed in relevance to their occurrence among evolutionary related MYBs, stating them to be significant.

4. DISCUSSION

Costunolide, the major sesquiterpene lactone in the roots of *S. lappa*, is a potential anti-carcinogenic component as reported to repress cell proliferation caused by carcinogenic induction, inhibits tumor invasion, metastasis and induce apoptosis in the target tissues [70]. The volatile oil obtained from the roots (richly constituting STLs) of *S. lappa* is reported to possess potential therapeutic and anti-breast cancer activities [71]. Despite its well established pharmacological properties, the mechanism regulating transcriptional patterns in costunolide biosynthetic pathway has not yet been elucidated.

In this study, assembled transcriptome was investigated to review the gene expression and transcriptional regulation of costunolide biosynthetic pathway in *S. lappa*. This will contribute unravelling the in depth embedded regulatory mechanism occurring in the biosynthesis of medicinally essential metabolites in the plant leaf and root tissues. In our previous report, the genes regulating the sesquiterpenoid pathway were characterized, followed by their relative expression studies [11]. The present work has reported the comparative transcriptome analysis of *S. lappa* leaf and root tissue, further identifying and annotating various transcripts encoding costunolide pathway genes and associated TFs, along with the active co-expression networks underplaying. In non-model plant species, the unavailability of genomic or transcriptomic information is the major hindrance in molecular studies. Thus, de novo library preparation presents the alternative to it, for instance similar studies are reported from many plant species such as *Chlorophytum borivilianum, Isodon amethysoides, Dioscorea zingiberensis* [72–74]. Also, supporting the well established fact that secondary metabolites are manufactured in leaves and later stored in root tissues, similar pattern was obtained as anticipated in *S. lappa*.

In the present dataset, 162,539 unique transcripts were obtained with average GC content of 39.36% and length more than 200bp long, which is indicative of good library preparation. In an attempt to elucidate sesquiterpenoid biosynthesis in chicory by Testone [1], relatively less number of unigenes were obtained in comparison to our dataset. Henceforth, our work provides detailed and in-depth transcriptomic information. A study in *C. winterianus* incorporated comparative differential expression studies in root and leaf tissue in order to explore the genes regulating terpene biosynthesis pathway and its regulation [75].

TFs play pivotal position in mediating various biological processes occurring in plants including development, growth, response to stresses and so on [76]. A total of 30,070 TFs were annotated which represents 58 TF families comprising bHLH, WRKY, MYB, NAC, C2H2 and ERF, representing them as the predominant families in regulation of plant processes. In other studies conducted on *Isodon amethystoides,* only 2% TFs were located which belongs to 55 families [73] in comparison to 18.5% TF identification done in our study, which infers that more comprehensive understanding of the regulatory role played by them in sesquiterpenoid pathway can be deciphered. bHLH TF family proteins also have reported involvement in the regulation of diterpenoid synthesis is previous studies, however, none of the studies have been reported in *S. lappa* [49,77]. The TFs such as AP2, MYB, bZIP and Zinc finger play a crucial role in abiotic stress in *B. nivea* [78]. Additionally, Nguyen studied the involvement of TFs like NAC, WRKY and ARF families in fruit ripening in bilberry [79]. In cotton, GaWRKY is reported to possess regulatory role in biosynthesis of sesquiterpene, similar to putative WRKY transcritps annotated (Table S7) in our plant [80].

The TF transcripts were assigned with KEGG pathways, revealing that many transcripts were involved in several secondary metabolites biosynthesis. The one with no match were considered

to be short sequences of proteins lacking domains or untranslated regions. In *T. ammi*, tissue specific differential expression of transcripts encoding for various TFs such as WRKY, MYB, bZIP, GATA was studied, wherein these TFs were reported to be highly expressed in aerial tissues [81]. The annotation results are consistent with our findings in *S. lappa* with bHLH, C2H2, WRKY and B3 as highly upregulated in leaf tissue. Likewise, the up-regulation of MYB studied in carrot is reportedly linked to biosynthesis of anthocyanin [82].

Differential gene expression analysis was performed to annotate the TFs transcripts. This resulted in identification for certain TF families expressing highly in leaf tissues s compared to the roots. Since various metabolic pathways precursors are synthesized in the leaves, it can be stated that most TFs regulating these pathways are thus active in the leaves [83]. Highly expressing TF transcripts included bHLH and MYB families which are reported to be acting as a complex along with WD40 proteins regulating flavonoid synthesis in many studies [84]. This suggest a coordinated role of MYB and bHLH proteins which is also reported in earlier studies for STLs biosynthesis [85]. While the expression of bHLH family TFs varies extremely in roots and leaves, the MYB and MYB-related family proteins show considerably moderate changes in the expression patterns in the two tissues, supporting our prediction on MYBs regulating costunolide pathway genes in roots.

Network analysis carried out in costunolide pathway genes directed us towards the correlated TF-gene partners and predicted the probable TF family of proteins regulating various steps in the costunolide biosynthesis. Various studies have used the gene co-expression analysis by the means of network studies [86,87] and identified clusters of genes correlated functionally. A module based approach was opted to study the correlating costunolide pathway gene and TF pairs. Six functionally enriched modules with significant interactions were obtained with tissue

specific expression patterns, supporting the facts regarding tissue specific synthesis of secondary metabolite. The DGE analysis also revealed that the genes encoding for costunolide biosynthesis enzymes were predominantly expressed in root as compared to the leaf. This implies that the oxidation and hydroxylation steps occurring later in the costunolide biogenesis might occur in the root tissue, similar to a previous report in *Tanacetum cinerariifolium* [88]. Amongst them, module 1 was chosen for our analysis since it comprised of key genes regulating STL biosynthesis, revealing putative TFs interacting along. The genes involved in mevalonate pathway were mostly co-expressed with MYB_related, bHLH and ERF TF family proteins. *SlHMGR* gene had MYB-related and bHLH TFs among its co-expressing partners, supported by Ginzberg suggesting that mevalonate pathway genes such as HMGR are controlled by GAME9 TF at some level. Co-expression analysis also revealed an APETALA2/Ethylene Response Factor (GAME9) regulated the production of alkaloids cooperatively with a MYC TF in *Catharanthus roseus* and tobacco [89]. *SlGAS1* gene was found to be co-expressed with WRKY and bHLH TF family proteins while *GAS2* gene was found co-expressed with NAC, G2-like and bHLH TF family along with up regulated expression in leaf tissue. In *Cichorium endivia,* it was reported that *GAS* gene showed positive correlation with WRKY and MYB TFs [1]. The co-expression data shows *DPD* having interacting MYB-related TFs, supported by a study revealing several MYB binding sites in *DPD* gene promoter region in *Salvia miltiorrhiza* and its expression regulation by maize C1 transcription factor (a R2R3 –MYB TF) [90].

The transcriptional control of several secondary metabolite pathway genes suggests that TFs holds an equally essential part in the regulation of terpenoid biosynthesis pathway. The costunolide biosynthesis related transcriptional regulation has not been reported in the literature yet. Costunolide synthase is involved in the final key step of the pathway, thus studying its

transcriptional regulation can provide certain privileges in understanding the costunolide synthesis. However, the TFs responsible for terpenoid build-up in plants are currently being identified [91].

The basic understanding of regulatory elements is provided by the upstream promoter region of a gene by reviewing their interaction with specific TFs. Thus it reveals well regulated expression pattern. A previous study reported that a number of transcription factors can be responsible for controlling a single gene and these may be part of same family of TFs sharing the common DNA binding domain [92]. Promoter analysis was performed in Geraniol 10-hydroxylase (G10H) gene in TIA biosynthesis which regulate the first committed step and observed binding sites for various TFs in *C. roseus* [21]. Similarly, Xu *et al* performed the analysis of transcription factor regulating delta (+) cadinene synthase in cotton. They found the putative motif for binding of WRKY TFs in delta (+) cadinene synthase gene and gradually identified GaWRKY1 having significant interacting reports with the promoter region of delta (+) cadinene synthase gene and further suggested that GaWRKY1 might be responsible for regulating terpenoid biosynthesis pathway in cotton [80]. The promoter analysis of *SlCOS1* gene in our study reported the presence of predominating MYB binding domains (CCWACC, CANNTG and WAACCA). This suggested the involvement of MYB transcription family proteins in regulation of *SlCOS1* in costunolide biosynthesis pathway. Moreover, the promoter region of *CcrdCOS* gene (which shows 93% identity to *SlCOS1*) was analysed to reveal various MYB-TF binding elements present upstream the gene. This is in relevance to a recent study by Testone *et al* where MYB and MYB related TFs were deduced correlated with *GAS/GAO/COS* genes transcription patterns, backing up our results positively [1]. Earlier reports suggest positive correlation in the expression of MYB TFs and genes of STL pathway including *GAS* and *GAO*. Transcription studies also

reveal the existence of strong correlation among the three genes (*GAS, GAO, COS*) as well as with the amount of STL in the plant [85]. This can predict MYBs as positively regulating transcription factor in costunolide synthesis. Intriguingly, when promoter analysis was performed for *Cynara COS* (95% homologus to *SlCOS1)* it also divulged various *cis* elements related to MYB binding present in its promoter. Henceforth, it further supports the predictions regarding MYB family of TFs to be involved in transcriptional regulation process of *SlCOS1*.

MYB TF family has been reported as a regulatory protein in various secondary metabolism pathways such as phenylpropanoid, proanthocyanidin and flavonoids [48,93]. Anthocyanin pathway related MYB TFs usually possess certain key motifs, thus showing an association with bHLH family proteins. As reported in a study, a MYB family transcription factor in carrot, *DcMYB6*, had correlated expression pattern with the production of anthocyanin and regulated its biosynthesis in purple carrots[94]. In addition, Kodama studied the regulation of anthocyanin biosynthesis by the complex consisting of MYB-bHLH-WD repeats [82]. Widely reported are the R2R3-MYB family proteins which play essential role in cold acclimation, petal and cell morphogenesis, epidermal differentiation, drought stress and plant secondary metabolism [26,95]. Certain flavonoid biosynthetic pathways are regulated by the transcriptional activation activity of MYB-bHLH-WD40 (MBW) family proteins [96]. The phylogenetic analysis revealed 18 potential DEGs where 10 shows homology with R2R3 type MYB proteins that vastly regulate secondary metabolism in plants, along with majorly conserved motif regions among 16 proteins for the two groups of MYB family TFs. The R1 and R2R3 family MYBs were depicted in two clusters along with significant homology with other Asteraceae family MYBs suggesting their divergence from a common ancestor with prominent variations in amino acid composition. Also, the subgroups defined on the basis of phylogeny reveal four MYBs clustered along secondary

metabolite biosynthesis proteins. Liu *et al* (2019) reported various MYB, bHLH and WD TFs that regulate anthocyanin content in pear including MYB20-like, MYB108-like, MYB1 like and MYB90 like TFs which are also reported in our analysis [97]. The results along with the data reported in previous studies suggest that MYB candidates might play an important role in regulating key costunolide pathway gene in *S. lappa* thus regulating an essential biosynthetic pathway for STLs. Although the role of MYB TFs in costunolide biosynthesis regulation is still to be characterized, the existing information and datasets presented in the study provided a major framework for future analyses.

5. Conclusion

The comparative transcriptomic study between root and leaf tissue was performed in order to predict the TFs regulating the costunolide biosynthesis pathway. Thus, the present study encompasses the thorough understanding of relative expression of costunolide biosynthetic genes and TFs associated with their regulation. Our earlier study contributed towards the transcriptomic repertoire of *S. lappa*. Henceforth, the combinatorial data of transcriptome sequencing, expression profiling and co-expression network analysis will enable the heterologous expression of the pathway and to increase the concentration of active metabolites. Further molecular studies and transcriptional validations will improve the understanding of basic regulatory mechanisms, opening new perspectives for exploring the costunolide production in the plant. This will enable us to unravel underplaying molecular networks and interacting transcription factors that can simultaneously manipulate the pathway genes and help developing molecular markers, contributing to an improved yield of essential oils from this plant of enriched medicinal value.

DECLARATIONS

FUNDING: Not applicable

AUTHOR CONTRIBUTIONS: KS and RK; conceived the idea, designed the experiments, analyzed the results, finalized the manuscript. VT and SB; performed all the experiments, analyzed the data, compiled the results and wrote the manuscript. SS and SP; analysis and compilation of transcriptome data and networks analysis.

AVAILABILITY OF DATA AND MATERIAL: The SRA data associated with this study has been submitted to NCBI SRA database under ID SUB5375309.

CONFLICT OF INTEREST: All the authors declare that there is no conflict of interest.

CONSENT FOR PUBLICATION: Not Applicable

ETHICS APPROVAL AND CONSENT TO PARTICIPATE: Not Applicable

ACKNOWLEDGEMENTS: VT is thankful to University Grant Commission (UGC), India for providing fellowship. RK is thankful to Department of Science and Technology (DST), India for providing fellowship under WOS-A scheme.

REFERENCES

[1] G. Testone, G. Mele, E. di Giacomo, G.C. Tenore, M. Gonnella, C. Nicolodi, G. Frugis, M.A. Iannelli, G. Arnesi, A. Schiappa, T. Biancari, D. Giannino, Transcriptome driven characterization of curly- and smooth-leafed endives reveals molecular differences in the sesquiterpenoid pathway, Horticulture Research. 6 (2019). https://doi.org/10.1038/s41438-018-0066-6.

[2] U. Amara, Z. ur R. Mashwani, A. Khan, S. Laraib, R. Wali, U. Sarwar, Q.T. Ain, S. Shakeel, Rahimullah, Sohail, Conservation Status and Therapeutic Potential of <i>Saussurea lappa</i>: An Overview, American Journal of Plant Sciences. 08 (2017) 602–614. https://doi.org/10.4236/ajps.2017.83041.

[3] N. Ikezawa, J.C. Göpfert, D.T. Nguyen, S.-U. Kim, P.E. O'Maille, O. Spring, D.-K. Ro, Lettuce Costunolide Synthase (*CYP71BL2*) and Its Homolog (*CYP71BL1*) from Sunflower Catalyze Distinct

Regio- and Stereoselective Hydroxylations in Sesquiterpene Lactone Metabolism, Journal of Biological Chemistry. 286 (2011) 21601–21611. https://doi.org/10.1074/jbc.M110.216804.

[4] J.C. Göpfert, G. MacNevin, D.-K. Ro, O. Spring, Identification, functional characterization and developmental regulation of sesquiterpene synthases from sunflower capitate glandular trichomes, BMC Plant Biology. 9 (2009) 86. https://doi.org/10.1186/1471-2229-9-86.

[5] T. Julianti, Y. Hata, S. Zimmermann, M. Kaiser, M. Hamburger, M. Adams, Antitrypanosomal sesquiterpene lactones from Saussurea costus, Fitoterapia. 82 (2011) 955–959. https://doi.org/10.1016/j.fitote.2011.05.010.

[6] M.M. Pandey, S. Rastogi, A.K.S. Rawat, Evaluation of Pharmacognostical Characters and Comparative Morphoanatomical Study of Saussurea costus (Falc.) Lipchitz and Arctium lappa L. Roots, Natural Product Sciences. 13 (2007) 7.

[7] R.N. Rao, S.S. Raju, K.S. Babu, P.R. Vadaparthi, OF Saussurea lappa AND ITS HERBAL FORMULATIONS, (2013) 9.

[8] C.P. Kuniyal, Y.S. Rawat, S.S. Oinam, J.C. Kuniyal, S.C.R. Vishvakarma, Kuth (Saussurea lappa) cultivation in the cold desert environment of the Lahaul valley, northwestern Himalaya, India: arising threats and need to revive socio-economic values, Biodiversity and Conservation. 14 (2005) 1035–1045. https://doi.org/10.1007/s10531-004-4365-x.

[9] J.Y. Cho, K.U. Baik, J.H. Jung, M.H. Park, In vitro anti-inflammatory effects of cynaropicrin, a sesquiterpene lactone, from Saussurea lappa, European Journal of Pharmacology. 398 (2000) 399–407. https://doi.org/10.1016/S0014-2999(00)00337-X.

[10] J.-W. de Kraker, M.C.R. Franssen, A. de Groot, W.A. König, H.J. Bouwmeester, (+)-Germacrene A Biosynthesis: The Committed Step in the Biosynthesis of Bitter Sesquiterpene Lactones in Chicory, Plant Physiology. 117 (1998) 1381–1392. https://doi.org/10.1104/pp.117.4.1381.

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

[12] J.-W. de Kraker, Biosynthesis of Costunolide, Dihydrocostunolide, and Leucodin. Demonstration of Cytochrome P450-Catalyzed Formation of the Lactone Ring Present in Sesquiterpene Lactones of Chicory, PLANT PHYSIOLOGY. 129 (2002) 257–268. https://doi.org/10.1104/pp.010957.

[13] Q. Liu, M. Majdi, K. Cankar, M. Goedbloed, T. Charnikhova, F.W.A. Verstappen, R.C.H. de Vos, J. Beekwilder, S. van der Krol, H.J. Bouwmeester, Reconstitution of the Costunolide Biosynthetic Pathway in Yeast and Nicotiana benthamiana, PLoS ONE. 6 (2011) e23255. https://doi.org/10.1371/journal.pone.0023255.

[14] B.P. Berman, Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, M.B. Eisen, Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome, Proceedings of the National Academy of Sciences. 99 (2002) 757–762. https://doi.org/10.1073/pnas.231608898.

[15] R. Shankar, A. Bhattacharjee, M. Jain, Transcriptome analysis in different rice cultivars provides novel insights into desiccation and salinity stress responses, Scientific Reports. 6 (2016). https://doi.org/10.1038/srep23719.

[16] S. Okay, E. Derelli, T. Unver, Transcriptome-wide identification of bread wheat WRKY transcription factors in response to drought stress, Molecular Genetics and Genomics. 289 (2014) 765–781. https://doi.org/10.1007/s00438-014-0849-x.

[17] G.S. Duraisamy, A.K. Mishra, T. Kocabek, J. Matoušek, Identification and characterization of promoters and cis-regulatory elements of genes involved in secondary metabolites production in hop (Humulus lupulus. L), Computational Biology and Chemistry. 64 (2016) 346–352. https://doi.org/10.1016/j.compbiolchem.2016.07.010.

[18] C.M. Hernandez-Garcia, J.J. Finer, Identification and validation of promoters and cis-acting regulatory elements, Plant Science. 217–218 (2014) 109–119. https://doi.org/10.1016/j.plantsci.2013.12.007.
[19] L.D. Ward, H.J. Bussemaker, Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences, Bioinformatics. 24 (2008) i165–i171. https://doi.org/10.1093/bioinformatics/btn154.
[20] J. Hu, D. Wang, J. Li, G. Jing, K. Ning, J. Xu, Genome-wide identification of transcription factors and transcription-factor binding sites in oleaginous microalgae Nannochloropsis, Scientific Reports. 4 (2015). https://doi.org/10.1038/srep05454.
[21] N. Suttipanta, S. Pattanaik, S. Gunjan, C.H. Xie, J. Littleton, L. Yuan, Promoter analysis of the Catharanthus roseus geraniol 10-hydroxylase gene involved in terpenoid indole alkaloid biosynthesis, Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression. 1769 (2007) 139–148. https://doi.org/10.1016/j.bbaexp.2007.01.006.
[22] D. Vom Endt, Transcription factors controlling plant secondary metabolism: what regulates the regulators?, Phytochemistry. 61 (2002) 107–114. https://doi.org/10.1016/S0031-9422(02)00185-1.
[23] D. Ma, G. Pu, C. Lei, L. Ma, H. Wang, Y. Guo, J. Chen, Z. Du, H. Wang, G. Li, H. Ye, B. Liu, Isolation and Characterization of AaWRKY1, an Artemisia annua Transcription Factor that Regulates the Amorpha-4,11-diene Synthase Gene, a Key Gene of Artemisinin Biosynthesis, Plant and Cell Physiology. 50 (2009) 2146–2161. https://doi.org/10.1093/pcp/pcp149.
[24] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, J.P. de Magalhães, Gene co-expression analysis for functional classification and gene–disease predictions, Briefings in Bioinformatics. (2017) bbw139. https://doi.org/10.1093/bib/bbw139.
[25] Y. Jiang, M.K. Deyholos, Functional characterization of Arabidopsis NaCl-inducible WRKY25 and WRKY33 transcription factors in abiotic stresses, Plant Molecular Biology. 69 (2009) 91–105. https://doi.org/10.1007/s11103-008-9408-3.
[26] N.W. Albert, A.H. Thrimawithana, T.K. McGhie, W.A. Clayton, S.C. Deroles, K.E. Schwinn, J.L. Bowman, B.R. Jordan, K.M. Davies, Genetic analysis of the liverwort *Marchantia polymorpha* reveals that R2R3MYB activation of flavonoid production in response to abiotic stress is an ancient character in land plants, New Phytologist. 218 (2018) 554–566. https://doi.org/10.1111/nph.15002.
[27] S. Ghawana, A. Paul, H. Kumar, A. Kumar, H. Singh, P.K. Bhardwaj, A. Rani, R.S. Singh, J. Raizada, K. Singh, S. Kumar, An RNA isolation system for plant tissues rich in secondary metabolites, BMC Research Notes. 4 (2011). https://doi.org/10.1186/1756-0500-4-85.
[28] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, (2011) 16.
[29] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proceedings of the National Academy of Sciences. 95 (1998) 14863–14868. https://doi.org/10.1073/pnas.95.25.14863.
[30] S. Pathania, V. Acharya, Computational Analysis of "-omics" Data to Identify Transcription Factors Regulating Secondary Metabolism in Rauvolfia serpentina, Plant Mol Biol Rep. 34 (2016) 283–302. https://doi.org/10.1007/s11105-015-0919-1.
[31] K. Aoki, Y. Ogata, D. Shibata, Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology, Plant and Cell Physiology. 48 (2007) 381–390. https://doi.org/10.1093/pcp/pcm013.
[32] A. McCluskey, A.G. Lalkhen, Statistics II: Central tendency and spread of data, Continuing Education in Anaesthesia Critical Care & Pain. 7 (2007) 127–130. https://doi.org/10.1093/bjaceaccp/mkm020.

[33] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society: Series B (Methodological). 57 (1995) 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
[34] G. Csardi, T. Nepusz, The igraph software package for complex network research, (n.d.) 9.
[35] S. Van Dongen, Graph Clustering Via a Discrete Uncoupling Process, SIAM J. Matrix Anal. & Appl. 30 (2008) 121–141. https://doi.org/10.1137/040608635.
[36] J.H. Morris, L. Apeltsin, A.M. Newman, J. Baumbach, T. Wittkop, G. Su, G.D. Bader, T.E. Ferrin, clusterMaker: a multi-algorithm clustering plugin for Cytoscape, BMC Bioinformatics. 12 (2011) 436. https://doi.org/10.1186/1471-2105-12-436.
[37] Z. Du, X. Zhou, Y. Ling, Z. Zhang, Z. Su, agriGO: a GO analysis toolkit for the agricultural community, Nucleic Acids Research. 38 (2010) W64–W70. https://doi.org/10.1093/nar/gkq310.
[38] R.J. Simes, An improved Bonferroni procedure for multiple tests of significance, (n.d.) 4.
[39] M. Shimizu, R. Fujimoto, H. Ying, Z. Pu, Y. Ebe, T. Kawanabe, N. Saeki, J.M. Taylor, M. Kaji, E.S. Dennis, K. Okazaki, Identification of candidate genes for fusarium yellows resistance in Chinese cabbage by differential expression analysis, Plant Mol Biol. 85 (2014) 247–257. https://doi.org/10.1007/s11103-014-0182-0.
[40] D.W. Huang, B.T. Sherman, X. Zheng, J. Yang, T. Imamichi, R. Stephens, R.A. Lempicki, Extracting Biological Meaning from Large Gene Lists with DAVID, Current Protocols in Bioinformatics. 27 (2009) 13.11.1-13.11.13. https://doi.org/10.1002/0471250953.bi1311s27.
[41] S. Porebski, L.G. Bailey, B.R. Baum, Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components, Plant Molecular Biology Reporter. 15 (1997) 8–15. https://doi.org/10.1007/BF02772108.
[42] M. Lescot, PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences, Nucleic Acids Research. 30 (2002) 325–327. https://doi.org/10.1093/nar/30.1.325.
[43] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol Syst Biol. 7 (2011) 539. https://doi.org/10.1038/msb.2011.75.
[44] T.L. Bailey, FITTING A MIXTURE MODEL BY EXPECTATION MAXIMIZATION TO DISCOVER MOTIFS IN BIOPOLYMERS, (n.d.) 33.
[45] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, A. Bairoch, Protein Identification and Analysis Tools on the ExPASy Server, in: J.M. Walker (Ed.), The Proteomics Protocols Handbook, Humana Press, Totowa, NJ, 2005: pp. 571–607. https://doi.org/10.1385/1-59259-890-0:571.
[46] J. Zhang, S. Zhang, H. Li, H. Du, H. Huang, Y. Li, Y. Hu, H. Liu, Y. Liu, G. Yu, Y. Huang, Identification of Transcription Factors ZmMYB111 and ZmMYB148 Involved in Phenylpropanoid Metabolism, Frontiers in Plant Science. 7 (2016). https://doi.org/10.3389/fpls.2016.00148.
[47] F. Zhang, X. Fu, Z. Lv, X. Lu, Q. Shen, L. Zhang, M. Zhu, G. Wang, X. Sun, Z. Liao, K. Tang, A Basic Leucine Zipper Transcription Factor, AabZIP1, Connects Abscisic Acid Signaling with Artemisinin Biosynthesis in Artemisia annua, Molecular Plant. 8 (2015) 163–175. https://doi.org/10.1016/j.molp.2014.12.004.
[48] S. Czemmel, R. Stracke, B. Weisshaar, N. Cordon, N.N. Harris, A.R. Walker, S.P. Robinson, J. Bogs, The Grapevine R2R3-MYB Transcription Factor VvMYBF1 Regulates Flavonol Synthesis in Developing Grape Berries, PLANT PHYSIOLOGY. 151 (2009) 1513–1530. https://doi.org/10.1104/pp.109.142059.
[49] A. Van Moerkercke, P. Steensma, F. Schweizer, J. Pollier, I. Gariboldi, R. Payne, R. Vanden Bossche, K. Miettinen, J. Espoz, P.C. Purnama, F. Kellner, T. Seppänen-Laakso, S.E. O'Connor, H. Rischer, J.

Memelink, A. Goossens, The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in *Catharanthus roseus*, Proceedings of the National Academy of Sciences. 112 (2015) 8130–8135. https://doi.org/10.1073/pnas.1504951112.

[50] C. Schluttenhofer, S. Pattanaik, B. Patra, L. Yuan, Analyses of Catharanthus roseus and Arabidopsis thaliana WRKY transcription factors reveal involvement in jasmonate signaling, BMC Genomics. 15 (2014) 502. https://doi.org/10.1186/1471-2164-15-502.

[51] A. Paul, A. Jha, S. Bhardwaj, S. Singh, R. Shankar, S. Kumar, RNA-seq-mediated transcriptome analysis of actively growing and winter dormant shoots identifies non-deciduous habit of evergreen tree tea during winters, Sci Rep. 4 (2015) 5932. https://doi.org/10.1038/srep05932.

[52] L. Mao, J.L. Van Hemert, S. Dash, J.A. Dickerson, Arabidopsis gene co-expression network and its functional modules, BMC Bioinformatics. 10 (2009) 346. https://doi.org/10.1186/1471-2105-10-346.

[53] U. Alon, An Introduction to Systems Biology, (n.d.) 45.

[54] R. Albert, Scale-free networks in cell biology, Journal of Cell Science. 118 (2005) 4947–4957. https://doi.org/10.1242/jcs.02714.

[55] S. Johnson, J.J. Torres, J. Marro, M.A. Muñoz, Entropic Origin of Disassortativity in Complex Networks, Phys. Rev. Lett. 104 (2010) 108702. https://doi.org/10.1103/PhysRevLett.104.108702.

[56] F. Emmert-Streib, M. Dehmer, ROBUSTNESS IN SCALE-FREE NETWORKS: COMPARING DIRECTED AND UNDIRECTED NETWORKS, Int. J. Mod. Phys. C. 19 (2008) 717–726. https://doi.org/10.1142/S0129183108012510.

[57] Z. Wang, A. Maity, C.K. Hsiao, D. Voora, R. Kaddurah-Daouk, J.-Y. Tzeng, Module-Based Association Analysis for Omics Data with Network Structure, PLoS ONE. 10 (2015) e0122309. https://doi.org/10.1371/journal.pone.0122309.

[58] H. Wang, C. Ma, Z. Li, L. Ma, H. Wang, H. Ye, G. Xu, B. Liu, Effects of exogenous methyl jasmonate on artemisinin biosynthesis and secondary metabolites in Artemisia annua L., Industrial Crops and Products. 31 (2010) 214–218. https://doi.org/10.1016/j.indcrop.2009.10.008.

[59] V.. Bulgakov, G.. Tchernoded, N.. Mischenko, M.. Khodakovskaya, V.. Glazunov, S.. Radchenko, E.. Zvereva, S.. Fedoreyev, Y.. Zhuravlev, Effect of salicylic acid, methyl jasmonate, ethephon and cantharidin on anthraquinone production by Rubia cordifolia callus cultures transformed with the rolB and rolC genes, Journal of Biotechnology. 97 (2002) 213–221. https://doi.org/10.1016/S0168-1656(02)00067-6.

[60] R.F. Benevenuto, T. Seldal, S.J. Hegland, C. Rodriguez-Saona, J. Kawash, J. Polashock, Transcriptional profiling of methyl jasmonate-induced defense responses in bilberry (Vaccinium myrtillus L.), BMC Plant Biol. 19 (2019) 70. https://doi.org/10.1186/s12870-019-1650-0.

[61] M. Majdi, M.R. Abdollahi, A. Maroufi, Parthenolide accumulation and expression of genes related to parthenolide biosynthesis affected by exogenous application of methyl jasmonate and salicylic acid in Tanacetum parthenium, Plant Cell Rep. 34 (2015) 1909–1918. https://doi.org/10.1007/s00299-015-1837-2.

[62] S.A. Petropoulos, I.C.F.R. Ferreira, L. Barros, eds., Phytochemicals in Vegetables: A Valuable Source of Bioactive Compounds, BENTHAM SCIENCE PUBLISHERS, 2018. https://doi.org/10.2174/9781681087399118010l.

[63] R. Stracke, M. Werber, B. Weisshaar, The R2R3-MYB gene family in Arabidopsis thaliana, Current Opinion in Plant Biology. 4 (2001) 447–456. https://doi.org/10.1016/S1369-5266(00)00199-0.

[64] J.A. Bac-Molenaar, E.F. Fradin, J.A. Rienstra, D. Vreugdenhil, J.J.B. Keurentjes, GWA Mapping of Anthocyanin Accumulation Reveals Balancing Selection of MYB90 in Arabidopsis thaliana, PLOS ONE. 10 (2015) e0143212. https://doi.org/10.1371/journal.pone.0143212.

[65] S. Fornalé, E. Lopez, J.E. Salazar-Henao, P. Fernández-Nohales, J. Rigau, D. Caparros-Ruiz, AtMYB7, a New Player in the Regulation of UV-Sunscreens in Arabidopsis thaliana, Plant and Cell Physiology. 55 (2014) 507–516. https://doi.org/10.1093/pcp/pct187.

[66] C.F. Su, Y.C. Wang, T.H. Hsieh, C.A. Lu, T.H. Tseng, S.M. Yu, A Novel MYBS3-Dependent Pathway Confers Cold Tolerance in Rice, PLANT PHYSIOLOGY. 153 (2010) 145–158. https://doi.org/10.1104/pp.110.153015.

[67] T. Wang, T. Tohge, A. Ivakov, B. Mueller-Roeber, A.R. Fernie, M. Mutwil, J.H.M. Schippers, S. Persson, Salt-Related MYB1 Coordinates Abscisic Acid Biosynthesis and Signaling during Salt Stress in Arabidopsis, Plant Physiology. 169 (2015) 1027–1041. https://doi.org/10.1104/pp.15.00962.

[68] Ao Gao, Jingbo Zhang, Wenheng Zhang, Evolution of RAD- and DIV-Like Genes in Plants, International Journal of Molecular Sciences. 18 (2017) 1961. https://doi.org/10.3390/ijms18091961.

[69] C. Dubos, R. Stracke, E. Grotewold, B. Weisshaar, C. Martin, L. Lepiniec, MYB transcription factors in Arabidopsis, Trends in Plant Science. 15 (2010) 573–581. https://doi.org/10.1016/j.tplants.2010.06.005.

[70] S.G. Ko, H.-P. Kim, D.-H. Jin, H.-S. Bae, S.H. Kim, C.-H. Park, J.W. Lee, Saussurea lappa induces G2-growth arrest and apoptosis in AGS gastric cancer cells, Cancer Letters. 220 (2005) 11–19. https://doi.org/10.1016/j.canlet.2004.06.026.

[71] J. Qiu, F. Gao, G. Shen, C. Li, X. Han, Q. Zhao, D. Zhao, X. Hua, Y. Pang, Metabolic Engineering of the Phenylpropanoid Pathway Enhances the Antioxidant Capacity of Saussurea involucrata, PLoS ONE. 8 (2013) e70665. https://doi.org/10.1371/journal.pone.0070665.

[72] S. Kalra, B.L. Puniya, D. Kulshreshtha, S. Kumar, J. Kaur, S. Ramachandran, K. Singh, De Novo Transcriptome Sequencing Reveals Important Molecular Networks and Metabolic Pathways of the Plant, Chlorophytum borivilianum, PLoS ONE. 8 (2013) e83336. https://doi.org/10.1371/journal.pone.0083336.

[73] F. Zhao, M. Sun, W. Zhang, C. Jiang, J. Teng, W. Sheng, M. Li, A. Zhang, Y. Duan, J. Xue, Comparative transcriptome analysis of roots, stems and leaves of Isodon amethystoides reveals candidate genes involved in Wangzaozins biosynthesis, BMC Plant Biology. 18 (2018). https://doi.org/10.1186/s12870-018-1505-0.

[74] J. Li, Q. Liang, C. Li, M. Liu, Y. Zhang, Comparative Transcriptome Analysis Identifies Putative Genes Involved in Dioscin Biosynthesis in Dioscorea zingiberensis, Molecules. 23 (2018) 454. https://doi.org/10.3390/molecules23020454.

[75] K. Devi, S.K. Mishra, J. Sahu, D. Panda, M.K. Modi, P. Sen, Genome wide transcriptome profiling reveals differential gene expression in secondary metabolite pathway of Cymbopogon winterianus, Scientific Reports. 6 (2016). https://doi.org/10.1038/srep21026.

[76] Q. Shen, L. Zhang, Z. Liao, S. Wang, T. Yan, P. Shi, M. Liu, X. Fu, Q. Pan, Y. Wang, Z. Lv, X. Lu, F. Zhang, W. Jiang, Y. Ma, M. Chen, X. Hao, L. Li, Y. Tang, G. Lv, Y. Zhou, X. Sun, P.E. Brodelius, J.K.C. Rose, K. Tang, The Genome of Artemisia annua Provides Insight into the Evolution of Asteraceae Family and Artemisinin Biosynthesis, Molecular Plant. 11 (2018) 776–788. https://doi.org/10.1016/j.molp.2018.03.015.

[77] J. Mertens, J. Pollier, R. Vanden Bossche, I. Lopez-Vidriero, J.M. Franco-Zorrilla, A. Goossens, The bHLH Transcription Factors TSAR1 and TSAR2 Regulate Triterpene Saponin Biosynthesis in *Medicago truncatula*, Plant Physiology. 170 (2016) 194–210. https://doi.org/10.1104/pp.15.01645.

[78] X. An, J. Chen, J. Zhang, Y. Liao, L. Dai, B. Wang, L. Liu, D. Peng, Transcriptome Profiling and Identification of Transcription Factors in Ramie (Boehmeria nivea L. Gaud) in Response to PEG Treatment, Using Illumina Paired-End Sequencing Technology, International Journal of Molecular Sciences. 16 (2015) 3493–3511. https://doi.org/10.3390/ijms16023493.

[79] N. Nguyen, M. Suokas, K. Karppinen, J. Vuosku, L. Jaakola, H. Häggman, Recognition of candidate transcription factors related to bilberry fruit ripening by de novo transcriptome and qRT-PCR analyses, Scientific Reports. 8 (2018). https://doi.org/10.1038/s41598-018-28158-7.

[80] Y.-H. Xu, Characterization of GaWRKY1, a Cotton Transcription Factor That Regulates the Sesquiterpene Synthase Gene (+)- -Cadinene Synthase-A, PLANT PHYSIOLOGY. 135 (2004) 507–515. https://doi.org/10.1104/pp.104.038612.

[81] M. Soltani Howyzeh, S.A. Sadat Noori, V. Shariati J., M. Amiripour, Comparative transcriptome analysis to identify putative genes involved in thymol biosynthesis pathway in medicinal plant Trachyspermum ammi L., Scientific Reports. 8 (2018). https://doi.org/10.1038/s41598-018-31618-9.

[82] M. Kodama, H. Brinch-Pedersen, S. Sharma, I.B. Holme, B. Joernsgaard, T. Dzhanfezova, D.B. Amby, F.G. Vieira, S. Liu, M.T.P. Gilbert, Identification of transcription factor genes involved in anthocyanin biosynthesis in carrot (Daucus carota L.) using RNA-Seq, BMC Genomics. 19 (2018). https://doi.org/10.1186/s12864-018-5135-6.

[83] C.-Q. Yang, X. Fang, X.-M. Wu, Y.-B. Mao, L.-J. Wang, X.-Y. Chen, Transcriptional Regulation of Plant Secondary Metabolism [F], Journal of Integrative Plant Biology. 54 (2012) 703–712. https://doi.org/10.1111/j.1744-7909.2012.01161.x.

[84] L. Zhao, L. Gao, H. Wang, X. Chen, Y. Wang, H. Yang, C. Wei, X. Wan, T. Xia, The R2R3-MYB, bHLH, WD40, and related transcription factors in flavonoid biosynthesis, Functional & Integrative Genomics. 13 (2013) 75–98. https://doi.org/10.1007/s10142-012-0301-4.

[85] G. Testone, G. Mele, E. Di Giacomo, M. Gonnella, M. Renna, G.C. Tenore, C. Nicolodi, G. Frugis, M.A. Iannelli, G. Arnesi, A. Schiappa, D. Giannino, Insights into the Sesquiterpenoid Pathway by Metabolic Profiling and De novo Transcriptome Assembly of Stem-Chicory (Cichorium intybus Cultigroup "Catalogna"), Frontiers in Plant Science. 7 (2016). https://doi.org/10.3389/fpls.2016.01676.

[86] J. Ruan, A.K. Dean, W. Zhang, A general co-expression network-based approach to gene expression analysis: comparison and applications, BMC Syst Biol. 4 (2010) 8. https://doi.org/10.1186/1752-0509-4-8.

[87] A. Fukushima, T. Nishizawa, M. Hayakumo, S. Hikosaka, K. Saito, E. Goto, M. Kusano, Exploring Tomato Gene Functions Based on Coexpression Modules Using Graph Clustering and Differential Coexpression Approaches, Plant Physiol. 158 (2012) 1487–1502. https://doi.org/10.1104/pp.111.188367.

[88] A.M. Ramirez, N. Saillard, T. Yang, M.C.R. Franssen, H.J. Bouwmeester, M.A. Jongsma, Biosynthesis of Sesquiterpene Lactones in Pyrethrum (Tanacetum cinerariifolium), PLoS ONE. 8 (2013) e65030. https://doi.org/10.1371/journal.pone.0065030.

[89] P.D. Cárdenas, P.D. Sonawane, J. Pollier, R. Vanden Bossche, V. Dewangan, E. Weithorn, L. Tal, S. Meir, I. Rogachev, S. Malitsky, A.P. Giri, A. Goossens, S. Burdman, A. Aharoni, GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway, Nature Communications. 7 (2016). https://doi.org/10.1038/ncomms10654.

[90] S. Zhao, J. Zhang, R. Tan, L. Yang, X. Zheng, Enhancing diterpenoid concentration in *Salvia miltiorrhiza* hairy roots through pathway engineering with maize C1 transcription factor, Journal of Experimental Botany. 66 (2015) 7211–7226. https://doi.org/10.1093/jxb/erv418.

[91] P. Broun, Y. Liu, E. Queen, Y. Schwarz, M.L. Abenes, M. Leibman, Importance of transcription factors in the regulation of plant secondary metabolism and their relevance to the control of terpenoid accumulation, Phytochemistry Reviews. 5 (2006) 27–38. https://doi.org/10.1007/s11101-006-9000-x.

[92] J.O. Borevitz, Y. Xia, J. Blount, R.A. Dixon, C. Lamb, Activation Tagging Identifies a Conserved MYB Regulator of Phenylpropanoid Biosynthesis, (n.d.) 12.

[93] R. Stracke, H. Ishihara, G. Huep, A. Barsch, F. Mehrtens, K. Niehaus, B. Weisshaar, Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling: TFs of A. thaliana R2R3-MYB subgroup 7, The Plant Journal. 50 (2007) 660–677. https://doi.org/10.1111/j.1365-313X.2007.03078.x.

[94] Z.-S. Xu, K. Feng, F. Que, F. Wang, A.-S. Xiong, A MYB transcription factor, DcMYB6, is involved in regulating anthocyanin biosynthesis in purple carrot taproots, Scientific Reports. 7 (2017). https://doi.org/10.1038/srep45324.

[95] Z. Gu, J. Zhu, Q. Hao, Y.-W. Yuan, Y.-W. Duan, S. Men, Q. Wang, Q. Hou, Z.-A. Liu, Q. Shu, L. Wang, A Novel R2R3-MYB Transcription Factor Contributes to Petal Blotch Formation by Regulating Organ-Specific Expression of *PsCHS* in Tree Peony (*Paeonia suffruticosa*), Plant and Cell Physiology. 60 (2019) 599–611. https://doi.org/10.1093/pcp/pcy232.

[96] W.-L. Wang, Y.-X. Wang, H. Li, Z.-W. Liu, X. Cui, J. Zhuang, Two MYB transcription factors (CsMYB2 and CsMYB26) are involved in flavonoid biosynthesis in tea plant [Camellia sinensis (L.) O. Kuntze], BMC Plant Biology. 18 (2018). https://doi.org/10.1186/s12870-018-1502-3.

[97] B. Liu, L. Wang, S. Wang, W. Li, D. Liu, X. Guo, B. Qu, Transcriptomic analysis of bagging-treated 'Pingguo' pear shows that MYB4-like1, MYB4-like2, MYB1R1 and WDR involved in anthocyanin biosynthesis are up-regulated in fruit peels in response to light, Scientia Horticulturae. 244 (2019) 428–434. https://doi.org/10.1016/j.scienta.2018.09.040.

TABLES

Table 1: Putative cis acting elements found in *SlCOS1* upstream region using PLACE tool.

Site Name	Signal Sequence	No. of cis-site	Binding Site	Function
GT1CONSENSUS	GRWAAW	11	S000198	Light responsive element
SURECOREATSULTR11	GAGAC	1	S000499	Auxin response factor (ARF) binding sequence
POLASIG2	AATTAAA	2	S000081	Poly A signal found in rice alpha-amylase
ARR1AT	NGATT	17	S000454	ARR1-binding element
CAATBOX1	CAAT	10	S000028	CAAT promoter consensus sequence
IBOXCORE	GATAA	3	S000199	Light responsive element
GATABOX	GATA	15	S000039	Light responsive element
MYB2CONSENSUSAT	YAACKG	3	S000409	MYB recognition site found in the promoters of the dehydration-responsive gene in Arabidopsis
MYBCORE	CNGTTR	3	S000176	MYB recognition site found in Arabidopsis
MYBCOREATCYCB1	AACGG	2	S000502	Myb core found in the promoter of Arabidopsis thaliana cyclin
POLASIG1	AATAAA	4	S000080	PolyA signal found in legA gene of pea, rice alpha-amylase
CCAATBOX1	CCAAT	2	S000030	CCAAT box found in the promoter of heat shock protein
LTRECOREATCOR15	CCGAC	1	S000153	Core of low temperature responsive element (LTRE)
MYBST1	GGATA	2	S000180	MYB recognition site
MYB1AT	WAACCA	2	S000408	MYB recognition site, dehydration-responsive
EECCRCAH1	GANTTNC	1	S000494	MYB recognition site
POLASIG3	AATAAT	2	S000088	PolyA signal-Consensus sequence for plant polyadenylation signal
MYBGAHV	TAACAAA	1	S000181	MYB recognition site

FIGURE LEGENDS.

Fig 1. Transcriptome analysis and annotation of *S. lappa* transcripts. (A) Length distribution of the assembled transcripts, **(B)** Venn diagram representing the annotation of transcripts according to various databases along with identified DEGs, **(C)** E value distribution of transcripts, and **(D)** BLASTX similarity score distribution.

Fig. 2. Classification and distribution of TFs into different TF families when compared to PlantTFDB in *S. lappa*. The top10 TF families for **(A)** complete set of putative TFs and **(B)** differentially expressed TFs are depicted in pie chart.

Fig. 3. Functional annotation (A) GO classification of putative TF transcripts in *S. lappa*, the most represented were the terms DNA integration (2411) of biological process category, Integral component of membrane (5190) from cellular components and nucleic acid binding (1617) of molecular function category and **(B)** Transcription factor transcripts classified into different families annotated with major KEGG pathways.

Fig. 4. Heat map of expression profiles of differentially expressed TFs from different (leaf and root) tissues.

Fig. 5. Differential expression analysis of TF families in *S. lappa* in leaf tissue as compared to root (control) tissue. Differences in top20 differentially expressed TF families has been represented.

Fig. 6. Topological analysis of co-expression network (A) Network density-based PCC threshold selection and **(B)** comparison of degree distribution of TF-gene network against random network.

Fig. 7. Heat map of expression profiles of all genes present in significant module 1. Genes in module 1 were compared for their differential expression for both leaf and root tissue, and was found to be up-regulated in root tissue.

Fig. 8. Depiction of the network of module 1 representing the differentially expressed TFs integrated to key pathway genes (involved in costunolide synthesis).

Fig. 9. Pie chart representing the percentage count of significant KEGG pathways of module 1.

Fig. 10. Differential expression profile of 241 putative MYB transcripts along with SlCOS1 in root and leaf tissues. The color scale on the top right corner represents log2 transformed expression values. The transcripts are classified into 3 clusters (left to right) depicted by red and blue mosaics patterns while our transcripts of interest fell in first cluster along with SlCOS1.

Fig. 11. Phylogenetic relationship among MYB proteins in *S. lappa* and other related plant species. Neighbour joining method was used to construct the phylogenetic tree using MEGA 7. The putative amino acid sequences of MYB were used and a bootstrap value of 1000 replicates.

The transcripts marked with red, purple, turquoise, orange, blue, yellow and grey represent sequences from *S. lappa*, *A. thaliana*, *C. cardunculus*, *V. vinifera*, *H. annus*, *G. max*, and *L. sativa*, respectively. The highlighted purple and green regions show R2R3 type and R1 subclass of MYB proteins, respectively along with their putative functions.

Fig. 12. Phylogenetic relationship and analysis of conserved motifs in MYB proteins of *S. lappa*. **(A)** The phylogenetic tree representing evolutionary relationship among 16 putative MYB proteins was constructed using Neighbor-joining method and 1000 bootstrap replicates, where the green and blue labeled clusters represent R2R3 and R1 families of MYB proteins in *S. lappa*. **(B)** The motif composition of the MYB proteins was identified using MEME tool and represented in a schematic diagram. A total of 15 motifs were elucidated and represented as colored boxes.

AUTHOR STATEMENT

KS and RK; conceived the idea, designed the experiments, analyzed the results, finalized the manuscript. VT and SB; performed all the experiments, analyzed the data, compiled the results and wrote the manuscript. SS and SP; analysis and compilation of transcriptome data and networks analysis.