Transcriptome analysis of polycyclic sesquiterpene biosynthesis in Leonurus sibiricus L.

Xiaoguang Yan
Tianjin University
Weiguo Li
Tianjin University
Dongmei Liang
Tianjin University
Lei Zhang
Tianjin University
Xiaoyu Qin
Tianjin University
Qinggele Caiyin
Tianjin University
Guangrong Zhao
Tianjin University
Zhijun Zhang
Tianjin Research Institute of Forestry and Pomology
Jian Liu
Wuqing District Center for Disease Control and Prevention
Meiqing Sun
Wuqing Distrist Center for Disease Control and Prevention
Jianjun Qiao (jianjunq@tju.edu.cn)
School of Chemical Engineering and Technology, Tianjin University, China

Research article

Keywords: Leonurus sibiricus, Sesquiterpene biosynthesis, Polycyclic sesquiterpene synthases, Transcriptome

Posted Date: July 4th, 2019

DOI: https://doi.org/10.21203/rs.2.10684/v1

Abstract

Abstract: Background: Leonurus sibiricus L. is a kind of medicinal herb widely used in China possessing a range of pharmacological activities. Polycyclic sesquiterpenes, one of the major bioactive components of L. sibiricus mainly existed in the root tissues, show the antibacterial, anti-inflammatory, antioxidant, and antiproliferative properties. However, for lacking of genome or transcriptome information, there is no report about terpene synthases (TPSs) genes of L. sibiricus. Results: In this research, six cDNA libraries were prepared from roots and leaves of L. sibiricus using the BGISEQ-500 platform. A total of 83,244 unigenes were finally obtained with an average length of 1,025 bp, of which 50,356 unigenes (60.49%) have annotations compared with BLAST searching results. 68 differentially expressed genes (DEGs) were putative biosynthetic genes involved in sesquiterpene metabolism among which five unigenes were speculated to be related to ylangene, copaene, gurjunene, selinene and cadinol biosynthesis in L. sibiricus. Conclusions: This research is the first report of transcriptome analysis of L. sibiricus, and unigenes probably involved in the biosynthesis of polycyclic sesquiterpenes in L. sibiricus. Furthermore, this database will supply important clues to explore other biological characteristics genetically in L. sibiricus. Keywords: Leonurus sibiricus, Sesquiterpene biosynthesis, Polycyclic sesquiterpene synthases, Transcriptome

Background

Leonurus sibiricus L. (*Lamiaceae*) is a kind of medicinal herb mainly used in China possessing a wide range of pharmacological activities [1, 2]. Essential oils of *L. sibiricus* containingvaluable secondary metabolites have shown antimicrobial, anti-inflammatory, antioxidant, and antiproliferative activities [3, 4]. Polycyclic sesquiterpenes paly vital roles among these beneficial secondary metabolites. For instance, the essential oil mainly containing α-copaene of *Casearia genu* (a kind of medicinal plant used in South America) had a special antifungal properties [5]. The essential oil containing α-copaene and β-ylangene extracted from *Syzygium aromaticum* seed showed inhibitory activity for *Escherichia coli*, *Pseudomonas aeruginosa* and *Staphylococcus aureus* [6]. β-selinene rich essential oils from *Callicarpa macrophylla* exhibited prominent antioxidant activity *in vitro* [7].

In *L. sibiricus*, just like any other plants, the basal skeleton for the biosynthesis of various sesquiterpenoids is the simple isoprenoids C5-unit isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP) [8, 9] (Fig. 1). IPP and DMAPP are synthesized through either mevalonic acid (MVA) pathway in all eukaryotic cells or the 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway in bacteria, some prokaryotes and plastids in plants [10]. The substrates farnesyl diphosphate (FPP) is formed by link of one DMAPP and two IPPs. Sesquiterpenes is further cycled and modified from FPP [11, 12]. Chemical diversity of polycyclic sesquiterpenes in *L. sibiricus* results from a different rearrangement of carbocation intermediates in the reactions catalyzed by sesquiterpene synthases (STSs). Besides, valuable sesquiterpenes differentially distribute in various tissues of *L. sibiricus*. The constituents of α-copaene, α-ylangene, β-selinene, α-cadinol and γ-gurjunene in normal roots of *L. sibiricus* were higher than that in other tissues [13]. Therefore, the research of profile of STSs in different tissues of *L. sibiricus* may help us understand in depth into sesquiterpene secondary metabolites of *L. sibiricus*. However, for lacking of genome or transcriptome information, there is no report about STS genes of *L. sibiricus.*

Owing to the developing sequencing technology, RNA sequencing (RNA-seq) become a sensible way for the genomic and transcriptome data analysis especially in non-model species [14]. Transcriptome analysis also assist

to excavate unique genes of key enzymes of secondary metabolites, which possibly include ingenious STSs. In order to exploit STS genes in *L. sibiricus*, functional genes involved in sesquiterpene biosynthesis in *L. sibiricus* were screened using *de novo* transcriptome sequencing. Five candidate genes of STSs were identified which could be used as an important resource to investigate the unique polycyclic sesquiterpenes in *L. sibiricus*. Furthermore, this database will supply important clues to explore biological characteristics genetically in *L. sibiricus*.

Results

Chemical composition of essential oils from *L. sibiricus* roots and leaves

The qualitative and quantitative determination of chemical composition of essential oils from *L. sibiricus* roots and leaves were performed using GC-MS. α-cadinol, β-selinene and α-copaene are the major sesqiuterpenes of essential oils from *L. sibiricus* roots and the concentrations of total polycyclic sesquiterpenes of essential oils from *L. sibiricus* roots are higher than that in leaves (Table 1).

L. sibiricus transcriptome analysis using BGISEQ-500

L. sibiricus was sequenced to generate a gene expression profile by using BGISEQ-500. Six cDNA libraries prepared from three leaf samples (L1, L2 and L3) and three root samples (R1, R2 and R3) were sequenced. Total of 107.60, 109.07, 109.26, 109.46, 109.15 and 109.38 Mb clean reads were obtained after eliminated the low-quality reads (Table 2). Subsequently, 63,927, 55,113, 51,261, 33,902, 30,833 and 36,609 unigenes were acquired by Trinity with mean lengths of 1,006, 994, 915, 936, 886 and 1,048 bp respectively (Table 3).

After *de novo* assembly, 83,244 unigenes were finally obtained with an average length of 1,025 bp (Table 3). Among these, 30,466 unigenes have lengths longer than 1 kb (>1000 bp) and 12,434 unigenes have lengths range between 1,000 bp and 1,500 bp. Summary of the sequencing and assembly results is shown in Table 3 and the length distribution of all unigenes is shown in Fig 2.

Functional annotation of *L. sibiricus* transcriptome

After assembly, unigenes functional annotation was performed with seven functional database (NR, NT, SwissProt, KEGG, KOG, InterPro and GO), and a total of 45,106, 36,184, 33,369, 34,498, 35,806, 40,197 and 23,790 unigenes were aligned, respectively. There were 50,356 unigenes (60.49% of all unigenes) annotated within at least one functional database (Table 4).

GO analysis is an international standard system of gene function classification, and the mainly terms including “Biological process”, “Cellular component” and “Molecular function”. Among the GO terms, cellular process was identified as the most annotation in “Biological process”, followed by metabolic process. 18 groups were involved in “Cellular component”, in which cell and cell part occupied the mainly categories. In addition, 13 groups existing in “Molecular function” and high percentage of unigenes were catalytic activity and binding (Fig 3).

Apart from GO functional classification, Kyoto Encyclopedia of Genes and Genomes (KEGG) was also used to explore the pathways of 34,498 unigenes. The majority of annotated categories were classified into “Metabolism” related pathways, in which metabolism of terpenoids and polyketides occupied the obvious part. Additionally, carbohydrate metabolism, amino acid metabolism and lipid metabolism were the top three pathways with most abundant transcripts. These pathways provide a valuable reference for investigating specific processes, functions and pathways of *L. sibiricus* transcriptome (Fig 4).

As to KOG functional classification, 35,806 unigenes were classified into 25 functional classifications (Fig 5). “General function prediction only” (8,499) was the most dominant term, followed by “Signal transduction mechanisms” (4,606) and “Posttranslational modification, protein transduction mechanisms” (3,493). Notably, 1,509 unigenes in the “secondary metabolites biosynthesis transport and catabolism” category may play vital roles in the biosynthesis of terpenes.

Overall, the results from GO, KEGG and KOG functional annotation and classification of unigenes, allowed us to obtain a comprehensive functional characterisation for the transcripts to further study.

DEGs in the root Vs leaf of *L. sibiricus*

DEGs annotated in Terpenoid biosynthesis of KEGG pathway was calculated by FPKM. DEGs of the six transcriptome libraries were used to dig the unigenes with significant differences in expression and were used to comment for GO classification and KEGG pathway analysis.

A total of 43,005 DEGs were obtained including 40,312 up-regulated and 2,693 down-regulated unigenes in root Vs leaf. Furthermore, 3,608, 1,953 and 5,700 unigenes expressed uniquely in R1, R2 and R3, respectively, and 35,510 unigenes were expressed in all three libraries, but at different levels. 5,363, 3,595 and 2,417 unigenes expressed uniquely in L1, L2, and L3, respectively, and 51,398 unigenes were expressed in all three libraries, but at different levels (Fig 6). A total of 12,559 DEGs could be annotated in KEGG based on sequence homologies in root Vs leaf, and 438 DEGs were related to metabolism of terpenoids and polyketides. A total of 16,171 DEGs could be annotated in GO based on sequence homologies in root Vs leaf, among which 5,898 DEGs were related to metabolic process. Furthermore, the number of up-regulated genes was much more than the number of down-regulated genes involved in metabolic process.

DEGs involved in terpenoid backbone biosynthesis in *L. sibiricus*

Based on the KEGG pathway annotation of metabolism of terpenoids and polyketides, a total of 64 DEGs out of 113 contigs/unigenes correlated with the terpene backbone biosynthesis as shown in Table 5.

Most of the DEGs were up regulated in root Vs leaf and only parts of DEGs related to HMGR, MK and IPPI were down regulated. Besides, 19 up-regulated unigenes were related to MVA pathway and 24 up-regulated unigenes were related to MEP pathway or DXP pathway. These contigs/unigenes might participate in the biosynthesis of FPP that is the building block of terpenoids (Table 5).

DEGs involved in sesquiterpene biosynthesis in *L. sibiricus*

According to the KEGG pathway annotation, there are 23 DEGs correlated with sesquiterpene biosynthesis in the root Vs leaf. Briefly, four up regulated unigenes are related to farnesene synthesis, five up regulated unigenes are related to nerolidol synthesis, 11 up regulated unigenes and one down regulated unigene are related to (-)-germacrene D synthesis, three up regulated unigenes are related to (+)-valencene synthesis, one up regulated unigene and three down regulated unigenes are related to vetispiradiene synthesis, three up regulated unigenes are related to 7-epi-α-selinene synthesis and one up regulated unigene is related to δ-cadinene synthesis.

Validation and expression analysis of key genes

In order to confirm the accuracy of the BGISEQ-500 sequencing and FPKM calculated results, we selected 15 unigenes and used qRT-PCR to determine their relative expression level in the root and leaf tissues of *L. sibiricus*. Most of them were putative sesquiterpene biosynthetic genes containing five lower-expressed unigenes (Unigene40848_All, CL4840.Contig1_All, CL8702.Contig2_All, Unigene9512_All and CL8997.Contig2_All), five higher-expressed unigenes (CL1461.Contig1_All, CL1461.Contig2_All, Unigene23382_All, CL6561.Contig1_All and Unigene19443_All) and five unchanged unigenes (CL9720.Contig2_All, CL3333.Contig3_All, CL3811.Contig3_All, CL1537.Contig4_All and CL9747.Contig2_All) calculated by FPKM. Quantification was performed using standard dilution curves for each studied gene fragment and the data were normalized for the quantity of housekeeping β-actin. The qRT-PCR and FPKM results were shown in Fig 7, and the expression levels are similar.

Disscussion

BGISEQ-500 sequencing and sequence annotation

In recent years, a number of novel components have been isolated and purified from plants and have a variety of biological activities [15, 16], many of which can be used as lead compounds for drug development [17, 18]. Sesquiterpene is a kind of the most abundant substance existing in the plant essential oils which often have a strong odor and may protect themselves from animals, insects and parasites [19, 20]. Due to this, these essential oils can be used in insecticide application or fragrance developments. *L. sibiricusi* is known as a rich source for sesquiterpense with unique aromadendrane skeletons such as ylangene, globulol and selinene [13]. The aims of this study is to carry out transcriptome of *L. sibiricus* with reference genome-free that would facilitate more detailed studies on various bioactive polycyclic sesquiterpenes biosynthesis. RNA-seqs were carried out using BGISEQ-500 sequencing and a total of 83,244 unigenes were obtained. 50,356 (60.49%) unigenes were provided significant BLAST results. This is the first time to report transcriptome research of *L. sibiricus*,offering sufficient references to study on sesquiterpene biosynthesis of Lamiaceae plants.

Terpenoid backbone biosynthetic genes and their differential expression patterns in *L. sibiricus*

Unique polycyclic sesquiterpenes in *L. sibiricus* like ylangene, copaene, gurjunene, selinene and cadinol were shown to be distributed mainly in root as shown in Table 1 and other literuture [13]. The terpene metabolic pathways related genes expressed higher in root may encode some enzymes responsible for unique sesquiterpene synthesis. Therefore, choosing the root and the leaf for comparative transcriptome analysis will more realistically dissect the genes responsible for the biosynthesis of ylangene, copaene, gurjunene, selinene and cadinol which have special antimicrobial or antioxidant properties.

Definitely, the basal skeleton for the biosynthesis of almost the 80,000 terpenoids is the simple isoprenoids C5-unit IPP and its isomer DMAPP [8, 9] which are synthesized through either MVA pathway or MEP pathway (Fig. 1). In the map of terpenoid backbone biosynthesis, 64 DEGs out of 113 contigs/unigenes were involved in MVA pathway and MEP pathway and listed in Table 4. The up-regulated genes involved in MVA pathway and MEP pathway may be used as the candidate genes to rebuild engineering microorganisms for the production of different isoprenoids and enhance their yields. Engineering microorganisms such as *Saccharomyces cerevisiae* and *Escherichia coli* were usually chose as the microbial host to produce various sesquiterpenes by introducing heterogenous genes to improve the availability of the precursor FPP. High titers of artemisinic acid have been obtained by the engineering of *S. cerevisiae* using enzymes from *A. annua* to engineer MVA pathway [21]. The same method was also used to achieve high yields of lycopene through expressing four exogenous genes to synthesize IPP and DMAPP from mevalonate in *E. coli* [22].

Sesquiterpene biosynthetic genes and their differential expression in *L. sibiricus*

Based on the reaction mechanism, TPSs can be classified into two groups: class I and class II. Class I TPSs ionize an isoprenoid diphosphate substrate to yield an allylic cation and inorganic pyrophosphate by assist of metal cluster whereas class II TPSs relies on a general acid (an aspartic acid side chain) to protonate the terminal carbon−carbon double bond of an isoprenoid substrate to yield a tertiary carbocation [14]. In depth, seven clades, a, b, c, d, e/f, g and h types are recognized according to TPSs amino acid sequence relativity [23, 24]. Most of plant STSs belong to TPS-a subfamily. Many researches have paid attention to the digging of STSs both in theory and in practice [25, 26], and the novel STSs truly improve the application of natural products [27]. However, even a lot of known terpenes and their derivatives have not been applied to manufacturing due to difficult isolation and purification active molecules from limited plant materials as well as the lack of accurate STSs genes.

According to the phylogenetic analysis with known TPSs (Fig 8), most of predicted STSs of *L. sibiricus* fell into TPS-a subfamily, and others fell into TPS-b and TPS-g subfamilies. Unigene4075_All, Unigene24876_All, Unigene7177_All, CL6561.Contig1_All, CL6561.Contig2_All, CL9601.Contig1_All, CL1461.Contig1_All, CL1461.Contig2_All, CL1626.Contig1_All and CL1626.Contig3_All fell into TPS-a subfamily. Sequence alignment of these TPS-a subfamily genes with 5-epi-aristolochene synthase (TEAS) from *Nicotiana tabacum* showed that the first STSs aspartate-rich, metal binding motif ‘DDXXD’ was conserved. This motif is responsible for catalyze the cleavage of the diphosphate group from the FPP allmost exist in all plant STSs. In contrast, the second metal bingding motif ‘(N,D)DXX(S,T)XXXE’ is not as conserved as the first one (Fig 9). Some known plant STSs only contain the ‘DXXDD’ motif as well [10]. Even more, these genes are the preferred STSs related to ylangene, copaene, gurjunene, selinene and cadinol which are all synthesized through germacrenyl cation intermediate from FPP [28-31]. Unigene4075_All and Unigene24876_All were presumed to be germacrene D synthase.

CL6561.Contig1_All was predicted to be δ-Cadinene synthase. CL6561.Contig2_All and CL9601.Contig1_All were supposed to be 7-epi-α-Selinene synthase. Further more, all of these genes were up regulated in root Vs leaf.

Conclusions

L. sibiricus transcriptome analysis was carried out using BGISEQ-500 for the first time so as to exploit the polycyclic sesquiterpene synthase genes in specific root tissue. 83,244 unigenes have generated from *L. sibiricus* transcriptome of which 50,356 (60.49% of all unigenes) unigenes provided significant BLAST results.

The candidate terpene synthetase genes and other detail data analysis delineates preliminary information for the putative sesquiterpene biosynthesis in *L. sibiricus* which may serve as targets to study polycyclic sesquiterpenes. Furthermore, this transcriptome data may provide precise clues to trace other natural products biosynthesis processes in *L. sibiricus*.

Methods

Plant materials and RNA isolation

L. sibiricus was collected in March 2018 from Shantou City, Guangdong Province, China and authenticated by Prof. Shihong Luo, Kunming Institute of Botany, China. A voucher specimen has been deposited at our laboratory. Leaf and root samples (L1, L2, L3, R1, R2 and R3) were harvested for RNA extraction according to the manufacturer's protocol using an RNA plant Plus Reagent (Tiangen, Beijing, China). The RNA integrity number of each sample was evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies Co. Ltd., Santa Clara, CA, USA).

Isolation and analysis of essential oils from *L. sibiricus*

The essential oils of root and leaf (20 g of each sample) *L. sibiricus* were obtained by hydrodistillation and their composition were analyzed by GC-MS. Gas chromatography was performed on a 30 m×0.25 mm×0.25um DB-5MS column (thermo). Helium was used as the carrier gas at a flow rate of 1 mL/min. Oven temperature started at 70℃ and held for 2 min, then raised at10℃/min to 300℃. The MS source temperature was 200°C. Split injection at 280℃ (split ratio: 1:10). Mass spectra were compared to the 2013 NIST library and the literature. The percentages were calculated from response peak area.

cDNA library production and BGISEQ-500 sequencing

The isolation of poly (A) + mRNA, reverse transcription of double-strand cDNA and cDNA library production were carried out as previously described [32]. Each cDNA library was sequenced in a single lane of the BGISEQ-500 system according to the operational instructions at the Beijing Genomics Institute (BGI-Shenzhen, China). The number of reads formed per sample was 10–11 Gb to get wide range of transcripts for de novo assembly.

De novo assembly and functional annotation analysis of BGISEQ-500 sequencing

In order to get perfect clean read data for *de novo* assembly, the raw reads from BGISEQ-500 were filtered and trimmed mainly using software Trinity (v2.0.6) [33], and software TGICL (v2.0.6) [34] as represented in our previous article [32].

To describe the functional annotation of the unigenes, a BLASTx search was performed with an E-value of 10^{-5} against protein databases, including NR (non-redundant protein database), SwissPort, KOG (euKaryotic Orthologous Group database), and KEGG (Kyoto Encyclopedia of Genes and Genomes protein database). Besides, a BLASTn search was also performed against NT (NCBI non-redundant nucleotide sequence database). With NR annotation, the Blast2GO [35] and InterProScan5 programwas used to obtain the GO (Gene ontology) and InterPro annotation of unigenes, respectively. GO classification was then performed using WEGO software [36] to illustrate the distribution of gene functions including Biological Process, Cellular Component and Molecular Function.

Differentially expressed unigene analysis

After assembly, clean reads were mapped to unigenes using Bowtie2 (v2.2.5) [37], and then gene expression level was calculated with RSEM (v1.1.12) [38]. To compare the difference of gene expression among different samples, the FPKM (Fragments per kilobase per transcript per million mapped reads) method was used [39]. The DEseq2 (Fold Change >= 2.00 and Adjusted Pvalue <= 0.05) and PossionDis (Fold Change >= 2.00 and FDR <= 0.001) were proposed to identify DEGs, and the P-value and FDR (false discovery rate) for each gene were calculated. DEGs were required to have thresholds of “log2 ratio>= 1” and “FDR< 0.001” [40]. Next, GO and KEGG analysis were again performed on the DEGs.

Phylogenetic analysis

TPSs sequence were aligned using ClustalW. A phylogenetic tree was built by the software MEGA version 6 [41], employing the neighbor-joining algorithm.

qRT-PCR analysis

The expression levels of this transcriptome were detected by qRT-PCR. The sequences of the specific primer sets are listed in Additional file 1. The β-actin gene (CL5763.Contig1_All) was used as an endogenous control. RNA extraction and reverse transcription were conducted according to the manufacturer’s instructions (Tiangen, Beijing, China). The expression levels were examined with Bio-Rad CFX96 real-time PCR system (Bio-Rad, CA, USA) with a SYBR Green-based PCR assay. The final volume for each reaction was 20 μL with the following components: 2 μL diluted cDNA template (1 mg/mL), 10 μL SYBR Green Mix (Bio-Rad, CA, USA), 0.4 μL forward primer (10 μM), 0.4 μL reverse primer (10 μM) and 7.2 μL ddH_2O. The reaction was conducted under the following profiles: 95℃ for 3 min, followed by 40 cycles of denaturation at 95℃ for 10 s and 60℃ for 30 s. The melting curve was obtained by heating the amplicon from 65℃ to 95℃ at increments of 0.5℃ per 5 s. Each qRT-PCR analysis was performed

with three biological replicates. The relative quantification of gene expression was computed using the $2^{-\Delta\Delta Ct}$ method.

Abbreviations

TPSs: Terpene synthases; DEGs: Differentially expressed genes; STSs: Sesquiterpene synthases MVA: Mevalonate; MEP: Methylerythritol phosphate; AACT: Acetyl-CoA C-acetyltransferase ; HMGS: Hydroxymethylglutaryl-CoA synthase; HMGR: Hydroxymethylglutaryl-CoA reductase; MK: Mevalonate kinase; PMK: Phosphomevalonate kinase; MCD: Mevalonate diphosphate decarboxylase; IPPI: Isopentenylpyrophosphate isomerase; DXPS: 1-deoxy-D-xylulose 5-phosphate synthase; DXR: 2-C-methylerythritol 4-phosphate reductase; MCT: 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase; CMK: 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MDS: 2-C-methyl-D-erythritol 2:4-cyclodiphosphate synthase; HDS: 4-hydroxy-3-methylbut-2-enyl diphosphate synthase; HDR: 1-hydroxy-2-methyl-butenyl 4-diphosphate reductase; DMAPP: Dimethylallyl pyrophosphate; IPP Isopentenyl diphosphate; FPPS: Farnesyl pyrophosphate synthase; NR: NCBI non-redundant protein database; KOG: Clusters of euKaryotic Orthologous Group database; KEGG: Kyoto Encyclopedia of Genes and Genomes protein databases; NT: NCBI non-redundant nucleotide sequence database; GO: Gene ontology; FPKM: Fragments per kilobase per transcript per million mapped reads; qRT-PCR: Quantitative Real-Time PCR.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The raw RNA-Seq reads will be available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database and will be public after the reception of this manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by Creative Research Groups of China (21621004) and National Key R&D Program of China (2017YFD0201400).

Authors’ contributions

XY, WL, DL, JQ, QC, GZ and conceived and designed the experiment; XY, ZZ, LZ, XQ, MS and JL collected plant samples, extracted RNA for sequencing and analyzed RNA-Seq data; XY, WL, and JL analyzed sesquiterpene in *L. sibiricus* by GC-MS; XY, DL, and WL wrote and modified the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful for Prof. Shihong Luo for his collection of leaf and root samples of *L. sibiricus*.

References

1. Sitarek P, Skala E, Wysokinska H, Wielanek M, Szemraj J, Toma M et al.The Effect of Leonurus sibiricus Plant Extracts on Stimulating Repair and Protective Activity against Oxidative DNA Damage in CHO Cells and Content of Phenolic Compounds. *Oxid Med Cell Longev* 2016, 2016:5738193.

2. Shin HY, Kim SH, Kang SM, Chang IJ, Kim SY, Jeon H et al.Anti-inflammatory activity of Motherwort (Leonurus sibiricus L.). *Immunopharmacol Immunotoxicol* 2009, 31(2):209-213.

3. Asadollahi MA, Maury J, Schalk M, Clark A, Nielsen J.Enhancement of farnesyl diphosphate pool as direct precursor of sesquiterpenes through metabolic engineering of the mevalonate pathway in Saccharomyces cerevisiae. *Biotechnol Bioeng* 2010, 106(1):86-96.

4. Gershenzon J, Dudareva N.The function of terpene natural products in the natural world. *Nat Chem Biol* 2007, 3(7):408-414.

5. Pereira FG, Marquete R, Domingos LT, Rocha MEN, Ferreira-Pereira A, Mansur E et al.Antifungal activities of the essential oil and its fractions rich in sesquiterpenes from leaves of Casearia sylvestris Sw. *An Acad Bras Cienc* 2017, 89(4):2817-2824.

6. Ajiboye TO, Mohammed AO, Bello SA, Yusuf, II, Ibitoye OB, Muritala HF et al.Antibacterial activity of Syzygium aromaticum seed: Studies on oxidative stress biomarkers and membrane permeability. *Microb Pathog* 2016, 95:208-215.

7. Chandra M, Prakash O, Kumar R, Bachheti RK, Bhushan B, Kumar M et al.β-Selinene-Rich Essential Oils from the Parts of Callicarpa macrophylla and Their Antioxidant and Pharmacological Activities. *Medicines* 2017, 4(3):52.

8. Sacchettini JC, Poulter CD.Creating isoprenoid diversity. *Science* 1997, 277(5333):1788-1789.

9. Christianson DW.Chemistry. Roots of biosynthetic diversity. *Science* 2007, 316(5821):60-61.

10. Chen F, Tholl D, Bohlmann J, Pichersky E.The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J* 2011, 66(1):212-229.

11. Tholl D.Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Curr Opin Plant Biol* 2006, 9(3):297-304.

12. Fraga BM.Natural sesquiterpenoids. *Nat Prod Rep* 2013, 30(9):1226-1264.

13. Sitarek P, Rijo P, Garcia C, Skala E, Kalemba D, Bialas AJ et al.Antibacterial, Anti-Inflammatory, Antioxidant, and Antiproliferative Properties of Essential Oils from Hairy and Normal Roots of Leonurus sibiricus L. and Their Chemical Composition. *Oxid Med Cell Longev* 2017, 2017:7384061.

14. Gao Y, Honzatko RB, Peters RJ.Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Nat Prod Rep* 2012, 29(10):1153-1175.

15. Takemura M, Maoka T, Misawa N.Carotenoid analysis of a liverwort Marchantia polymorpha and functional identification of its lycopene beta- and epsilon-cyclase genes. *Plant Cell Physiol* 2014, 55(1):194-200.

16. Wang S, Li RJ, Zhu RX, Hu XY, Guo YX, Zhou JC et al.Notolutesins A-J, dolabrane-type diterpenoids from the Chinese liverwort Notoscyphus lutescens. *J Nat Prod* 2014, 77(9):2081-2087.

17. Nagashima F, Asakawa Y.Terpenoids and bibenzyls from three Argentine liverworts. *Molecules* 2011, 16(12):10471-10478.

18. Yu HN, Wang L, Sun B, Gao S, Cheng AX, Lou HX.Functional characterization of a chalcone synthase from the liverwort Plagiochasma appendiculatum. *Plant Cell Rep* 2015, 34(2):233-245.

19. Ramirez M, Kamiya N, Popich S, Asakawa Y, Bardon A.Insecticidal constituents from the argentine liverwort Plagiochila bursata. *Chem Biodivers* 2010, 7(7):1855-1861.

20. Ren F, Mao H, Liang J, Liu J, Shu K, Wang Q.Functional characterization of ZmTPS7 reveals a maize tau-cadinol synthase involved in stress response. *Planta* 2016, 244(5):1065-1074.

21. Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM et al.Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 2006, 440(7086):940-943.

22. Yoon SH, Lee YM, Kim JE, Lee SH, Lee JH, Kim JY et al.Enhanced lycopene production in Escherichia coli engineered to synthesize isopentenyl diphosphate and dimethylallyl diphosphate from mevalonate. *Biotechnol Bioeng* 2006, 94(6):1025-1032.

23. Yamada Y, Cane DE, Ikeda H.Diversity and analysis of bacterial terpene synthases. *Methods Enzymol* 2012, 515:123-162.

24. Silva GN, Rezende LC, Emery FS, Gosmann G, Gnoatto SC.Natural and Semi synthetic Antimalarial Compounds: Emphasis on the Terpene Class. *Mini Rev Med Chem* 2015, 15(10):809-836.

25. Bian G, Han Y, Hou A, Yuan Y, Liu X, Deng Z et al.Releasing the potential power of terpene synthases by a robust precursor supply platform. *Metab Eng* 2017, 42:1-8.

26. Stajich JE, Wilke SK, Ahren D, Au CH, Birren BW, Borodovsky M et al.Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus). *Proc Natl*

Acad Sci U S A 2010, 107(26):11889-11894.

27. Dai Z, Liu Y, Zhang X, Shi M, Wang B, Wang D et al.Metabolic engineering of Saccharomyces cerevisiae for production of ginsenosides. *Metabolic Engineering* 2013, 20(5):146-156.

28. Lee S, Chappell J.Biochemical and genomic characterization of terpene synthases in Magnolia grandiflora. *Plant Physiol* 2008, 147(3):1017-1033.

29. Kumar S, Kempinski C, Zhuang X, Norris A, Mafu S, Zi J et al.Molecular Diversity of Terpene Synthases in the Liverwort Marchantia polymorpha. *The Plant Cell* 2016:tpc.00062.02016.

30. Zhou H, Yang YL, Zeng J, Zhang L, Ding ZH, Zeng Y.Identification and Characterization of a δ-Cadinol Synthase Potentially Involved in the Formation of Boreovibrins in Boreostereum vibrans of Basidiomycota. *Natural Products & Bioprospecting* 2016, 6(3):167-171.

31. Baer P, Rabe P, Fischer K, Citron CA, Klapschinski TA, Groll M et al.Induced-fit mechanism in class I terpene cyclases. *Angew Chem Int Ed Engl* 2014, 53(29):7652-7656.

32. Li W, Xu R, Yan X, Liang D, Zhang L, Qin X et al.De novo leaf and root transcriptome analysis to explore biosynthetic pathway of Celangulin V in Celastrus angulatus maxim. *BMC Genomics* 2019, 20(1):7.

33. Grabherr MG, Haas BJ, Moran Y, Levin JZ, Thompson DA, Ido A et al.Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 2011, 29(7):644.

34. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S et al.TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 2003, 19(5):651-652.

35. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ et al.High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 2008, 36(10):3420-3435.

36. Young MD, Wakefield MJ, Smyth GK, Oshlack A.Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010, 11(2):R14.

37. Langmead B, Salzberg SL.Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012, 9(4):357-359.

38. Bo L, Dewey CN.RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics* 2011.

39. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008, 5(7):621-628.

40. Audic S, Claverie JM.The significance of digital gene expression profiles. *Genome Res* 1997, 7(10):986-995.

41. Moniodis J, Jones CG, Barbour EL, Plummer JA, Ghisalberti EL, Bohlmann J.The transcriptome of sesquiterpenoid biosynthesis in heartwood xylem of Western Australian sandalwood (Santalum spicatum). *Phytochemistry* 2015, 113:79-86.

Tables

Table 1 Main sesquiterpenes of essential oils from roots and leaves of *L. sibiricus*

Constituent	Sesquiterpene composition of essential oil (%)	
	roots	leaves
Nerolidol	0.5	0.4
α-ylangene	1.7	1.1
α-copaene	9.7	-
Farnesene	-	0.4
γ-gurjunene	0.5	-
Selina-4,11-diene	1.8	0.3
β-selinene	12.7	0.9
β-eudesmol	5.6	1.4
α-cadinol	21	1.3

Table 2 Summary of data output quality of various libraries

Sample	Total Raw Reads(M)	Total Clean Reads(M)	Clean Reads Q20(%)	Clean Reads Q30(%)
L1	110.02	107.6	98.33	91.57
L2	112.06	109.7	98.27	91.78
L3	112.06	109.26	97.32	89.29
R1	112.06	109.46	97.63	90.2
R2	112.06	109.15	98.43	91.88
R3	112.06	109.38	98.5	92.1

Q20: The percentage of bases with a Phred value > 20

Q30: The percentage of bases with a Phred value > 30

Table 3 Summary of assembly results of *L. sibiricus*

Sample	Total Number	Total Length	Mean Length	N50	GC(%)
L1	63927	64360490	1006	1715	40.81
L2	55113	54812204	994	1663	41.69
L3	51261	46949093	915	1575	41.71
R1	33902	31736339	936	1455	42.75
R2	30833	27346958	886	1371	42.88
R3	36609	38374245	1048	1618	42.17
All-Unigene	83244	85329981	1025	1779	41.1

N50: a weighted median statistic that 50% of the Total Length is contained in Transcripts great than or equal to this value

Table 4 Summary of functional annotations of *L. sibiricus*

Values	Total	Nr	Nt	Swissprot	KEGG	KOG	Interpro	GO	Overall
Number	83,244	45,106	36,184	33,369	34,498	35,806	40,197	23,790	50,356
Percentage	100%	54.19%	43.47%	40.09%	41.44%	43.01%	48.29%	28.58%	60.49%

Overall: the number of transcripts which be annotated with at least one functional database.

Table 5 Discovery and expression of unigenes involved in sesquiterpene biosynthesis in *L. sibiricus*

Enzymes name	Abbreviation	EC number	Different regulated unigenes
Acetyl-CoA C-acetyltransferase	AACT	EC 2.3.1.9	Unigene18540_All , Unigene8183_All, Unigene29150_All CL2370.Contig1_All, CL2370.Contig2_All
Hydroxymethylglutaryl-CoA synthase,	HMGS	EC 2.3.3.10	Unigene32765_All, Unigene14772_All
Hydroxymethylglutaryl-CoA reductase	HMGR	EC 1.1.1.34	CL695.Contig2_All, CL4778.Contig1_All, Unigene14400_All CL4778.Contig5_All, CL4778.Contig4_All, CL4778.Contig2_All Unigene17226_Al
Mevalonate kinase	MK	EC 2.7.1.36	CL3052.Contig2_All
Phosphomevalonate kinase	PMK	EC 2.7.4.2	Unigene13112_All, CL6082.Contig5_All, Unigene5627_All CL6082.Contig2_All
1-deoxy-D-xylulose-5-phosphate synthase	DXP	EC 2.2.1.7	CL1664.Contig3_All, CL2703.Contig1_All, Unigene15550_All CL2221.Contig6_All, CL2221.Contig4_All, CL2703.Contig2_Al CL2221.Contig3_All , CL2221.Contig2_All, CL2221.Contig5_All CL2221.Contig7_All, Unigene3358_All
1-deoxy-D-xylulose-5-phosphate reductoisomerase	DXR	EC 1.1.1.267	CL1224.Contig2_All , CL1224.Contig3_All
	MCT	EC 2.7.7.60	CL1219.Contig3_All, CL1219.Contig2_All, CL1219.Contig1_All
	CMK	EC 2.7.1.148	Unigene14314_All
	MDS	EC 4.6.1.12	Unigene9342_All
	HDS	EC 1.17.7.1	CL1328.Contig1_All, CL1328.Contig2_All
	HDR	EC 1.17.7.4	CL5134.Contig1_All, CL5134.Contig2_All, CL5134.Contig4_All
isopentenyl-diphosphate δ-isomerase	IPPI	EC 5.3.3.2	CL6011.Contig3_All, Unigene24940_All
Farnesyl diphosphate synthase	FPP	EC 2.5.1.10	Unigene18210_All, Unigene23632_All, Unigene10235_All Unigene18237_All, Unigene23426_All, Unigene9949_All Unigene44417_All, CL6467.Contig1_All, Unigene1578_All Unigene1044_All, Unigene9123_All, CL6467.Contig2_All CL6467.Contig3_All, CL6467.Contig4_All, Unigene24404_All Unigene17435_All, Unigene18630_All, CL4800.Contig3_All CL4800.Contig2_All
Farnesene	_	EC 4.2.3.46	CL1626.Contig3_All, CL1626.Contig1_All, CL1626.Contig2_All
Nerolidol	_	EC 4.2.3.48	CL7235.Contig3_All, CL7235.Contig4_All, CL7235.Contig2_All

			CL7235.Contig1_All, Unigene23069_All
Germacrene D synthase	_	EC 4.2.3.75	CL9601.Contig1_All, CL6561.Contig2_All, CL9601.Contig2_All Unigene24876_All, Unigene4075_All, CL2126.Contig1_All Unigene8124_All, CL6561.Contig3_All, CL2126.Contig2_All CL2126.Contig4_All, CL2126.Contig3_All, CL6561.Contig1_All
Valencene synthase	_	EC 4.2.3.73	CL9601.Contig1_All, CL6561.Contig2_All, CL6561.Contig3_All
Vetispiradiene	_	EC 4.2.3.21	Unigene7177_All
7-epi-α-Selinene	_	EC 4.2.3.86	CL9601.Contig1_All, CL6561.Contig2_All, CL6561.Contig3_All
δ-Cadinene	_	EC 4.2.3.13	CL6561.Contig1_All

Figures

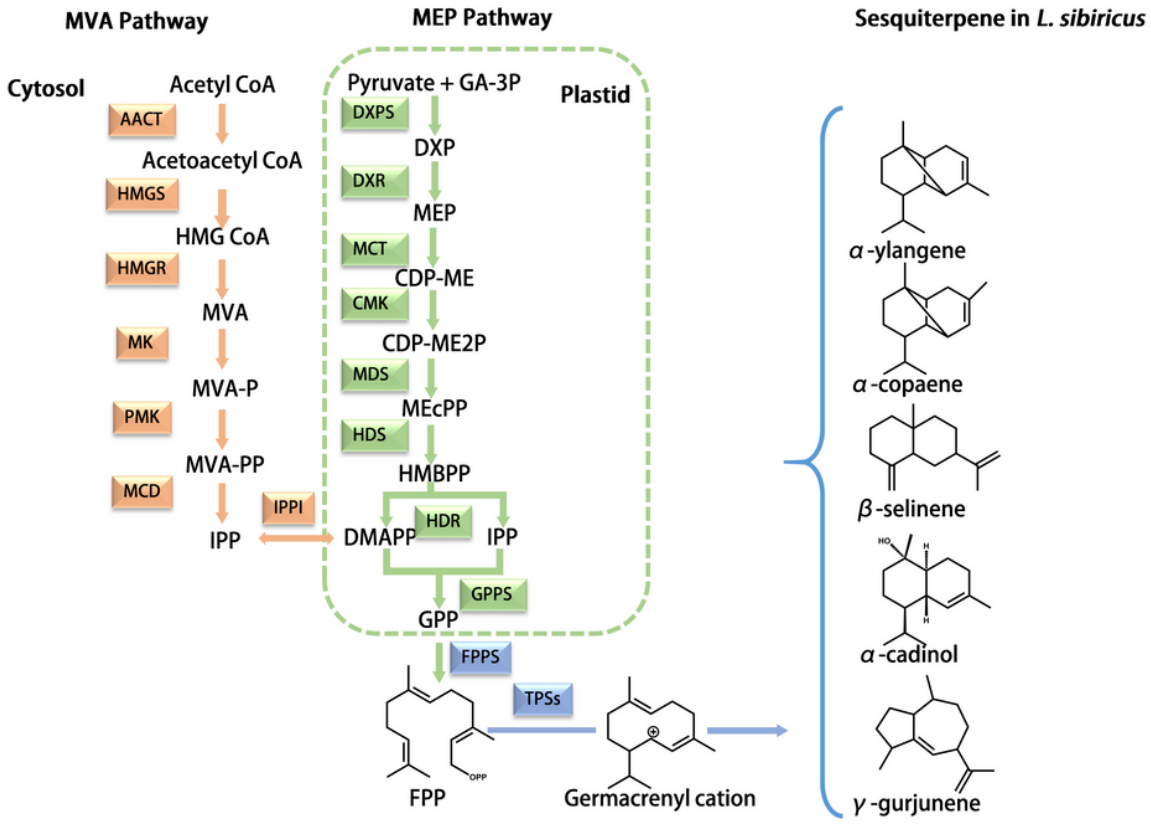

Figure 1

Metabolic pathway for polycyclic sesquiterpene biosynthesis of L. sibiricus Abbreviations: MVA, Mevalonate; MEP, methylerythritol phosphate; AACT, acetyl-CoA C-acetyltransferase ; HMGS, Hydroxymethylglutaryl-CoA synthase;

HMGR, hydroxymethylglutaryl-CoA reductase; MK, mevalonate kinase; PMK, phosphomevalonate kinase; MCD, mevalonate diphosphate decarboxylase; IPPI, isopentenylpyrophosphate isomerase; DXPS, 1-deoxy-D-xylulose 5-phosphate synthase; DXR, 2-C-methylerythritol 4-phosphate reductase; MCT, 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase; CMK, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MDS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS, 4-hydroxy-3-methylbut-2-enyl diphosphate synthase; HDR, 1-hydroxy-2-methyl-butenyl 4-diphosphate reductase; DMAPP, dimethylallyl pyrophosphate; IPP Isopentenyl diphosphate; FPPS, farnesyl pyrophosphate synthase; TPSs, sesquiterpene synthases

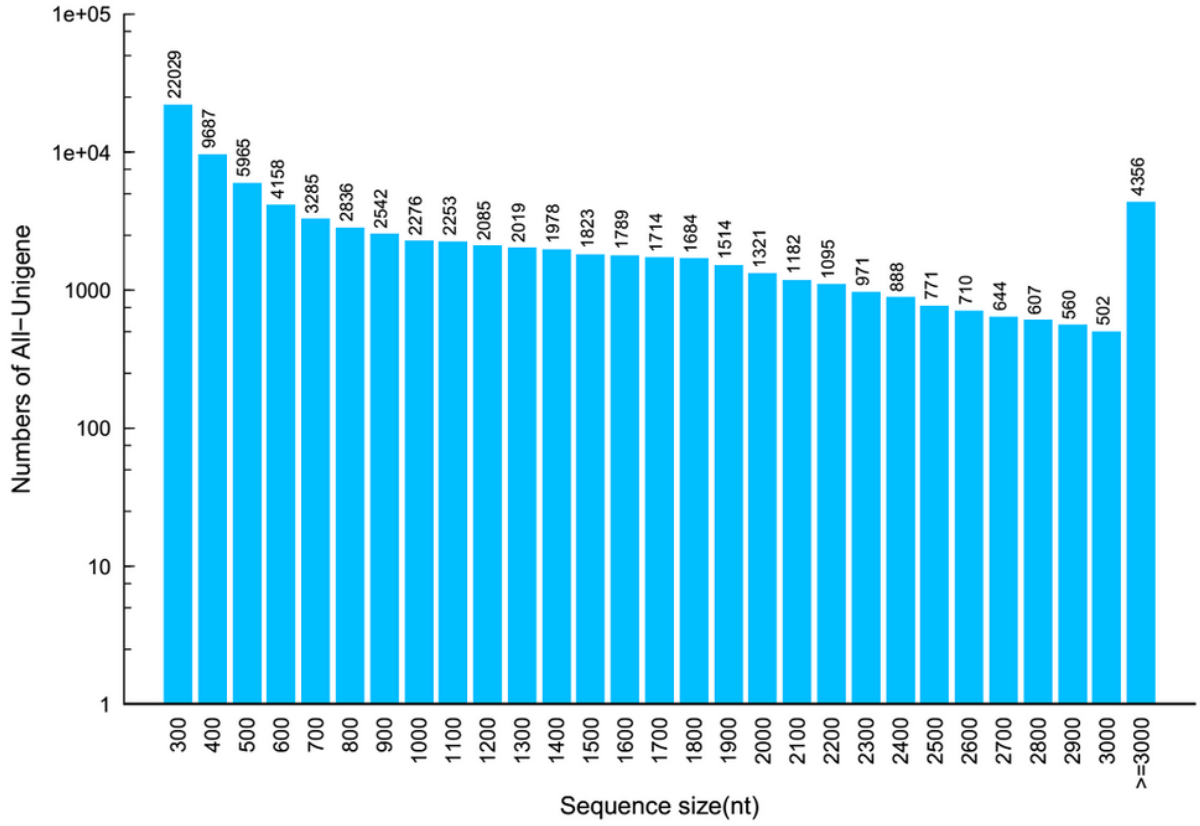

Figure 2

Distribution size of de novo assembled unigenes for L. sibiricus. A total of 83,244 unigenes sizes were calculated for L. sibiricus

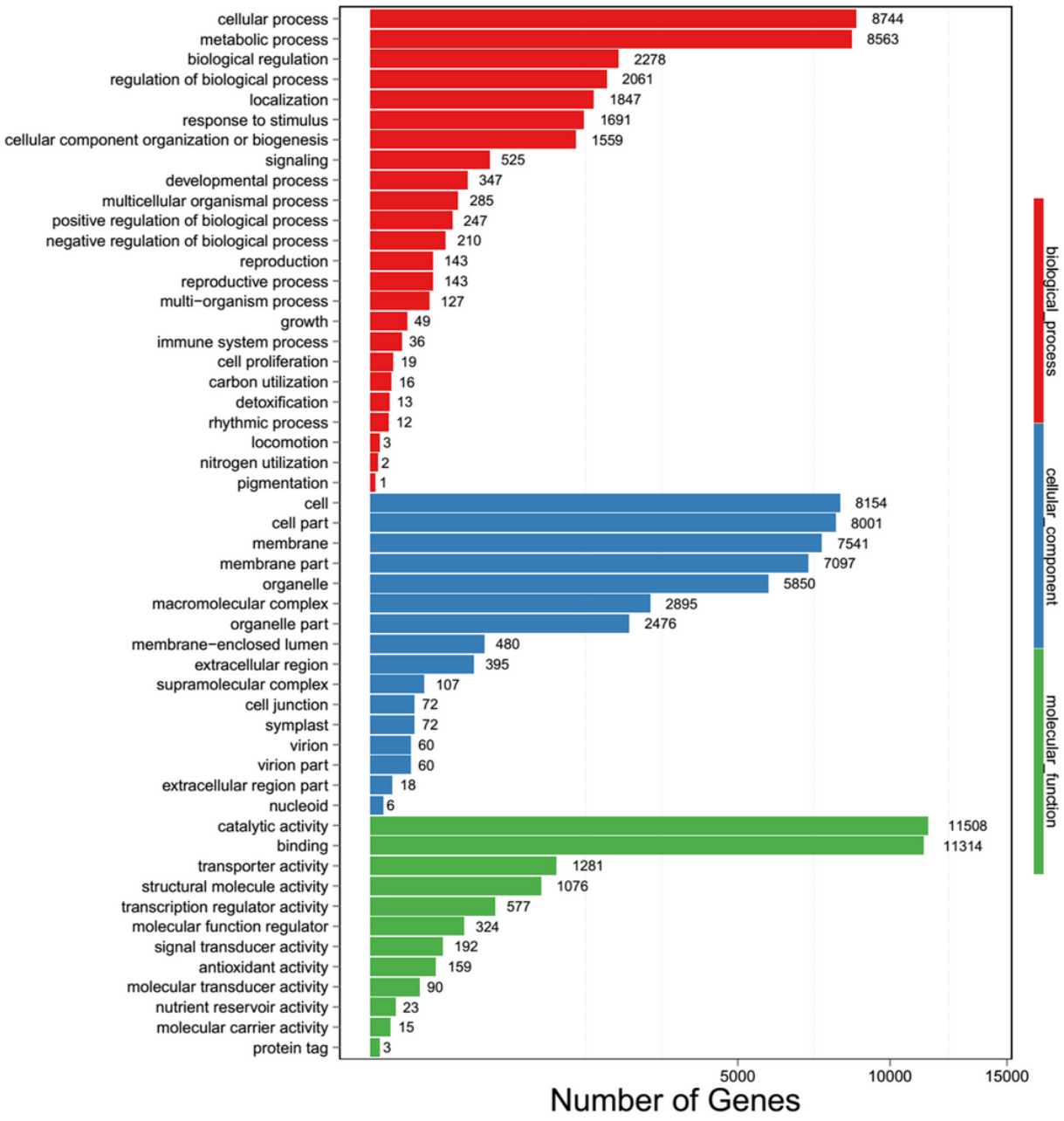

Figure 3

GO categories of L. sibiricus. The results are summarized in mainly three categories: biological process, cellular component and molecular function

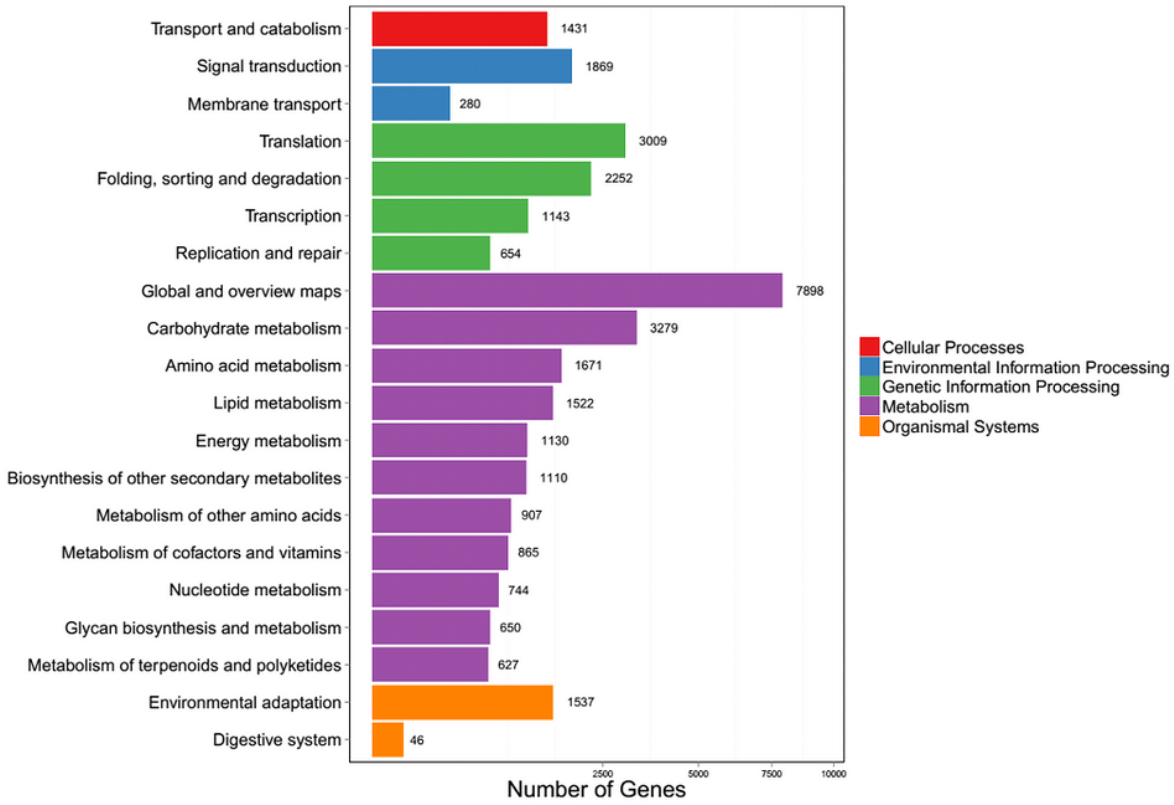

Figure 4

KEGG function classification of L. sibiricus

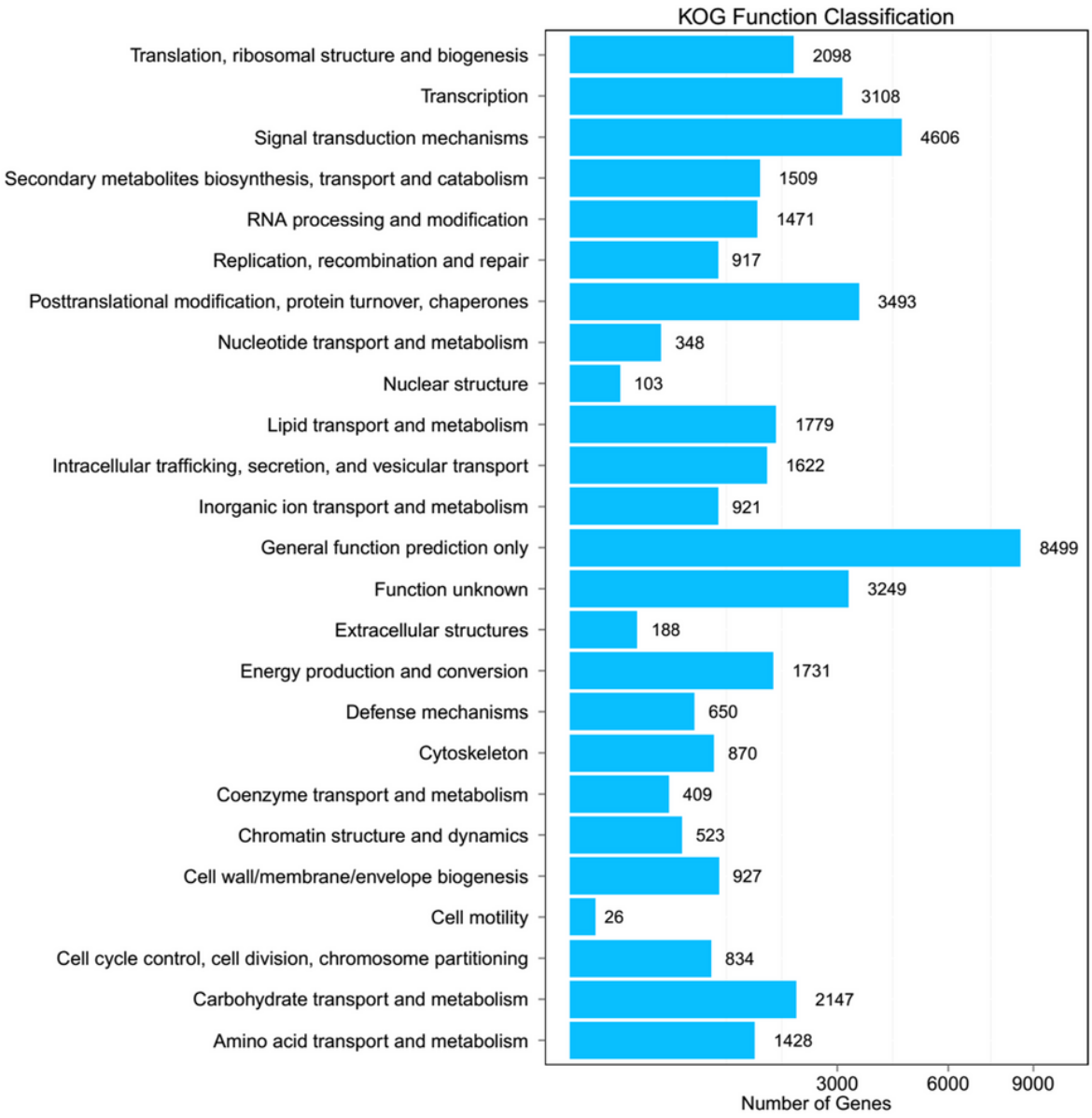

Figure 5

KOG function classification of L. sibiricus. A total of 35,806 unigenes were classified into 25 functional categories according to their predicted gene products using the COG database

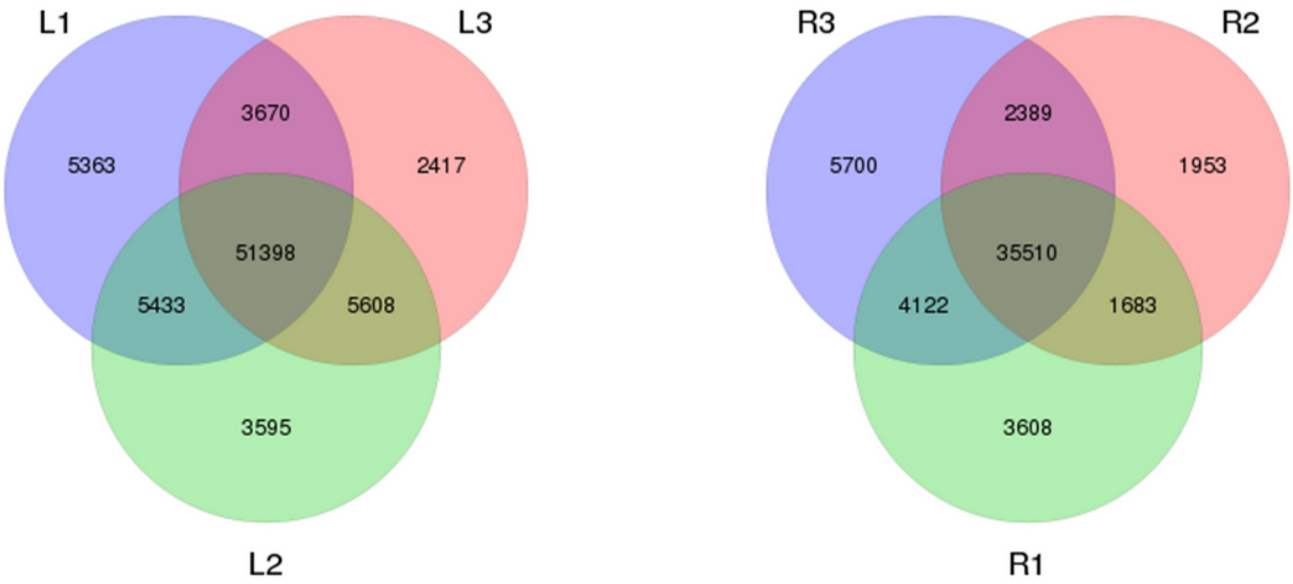

Figure 6

Venn diagram of the unigenes at root and leaf

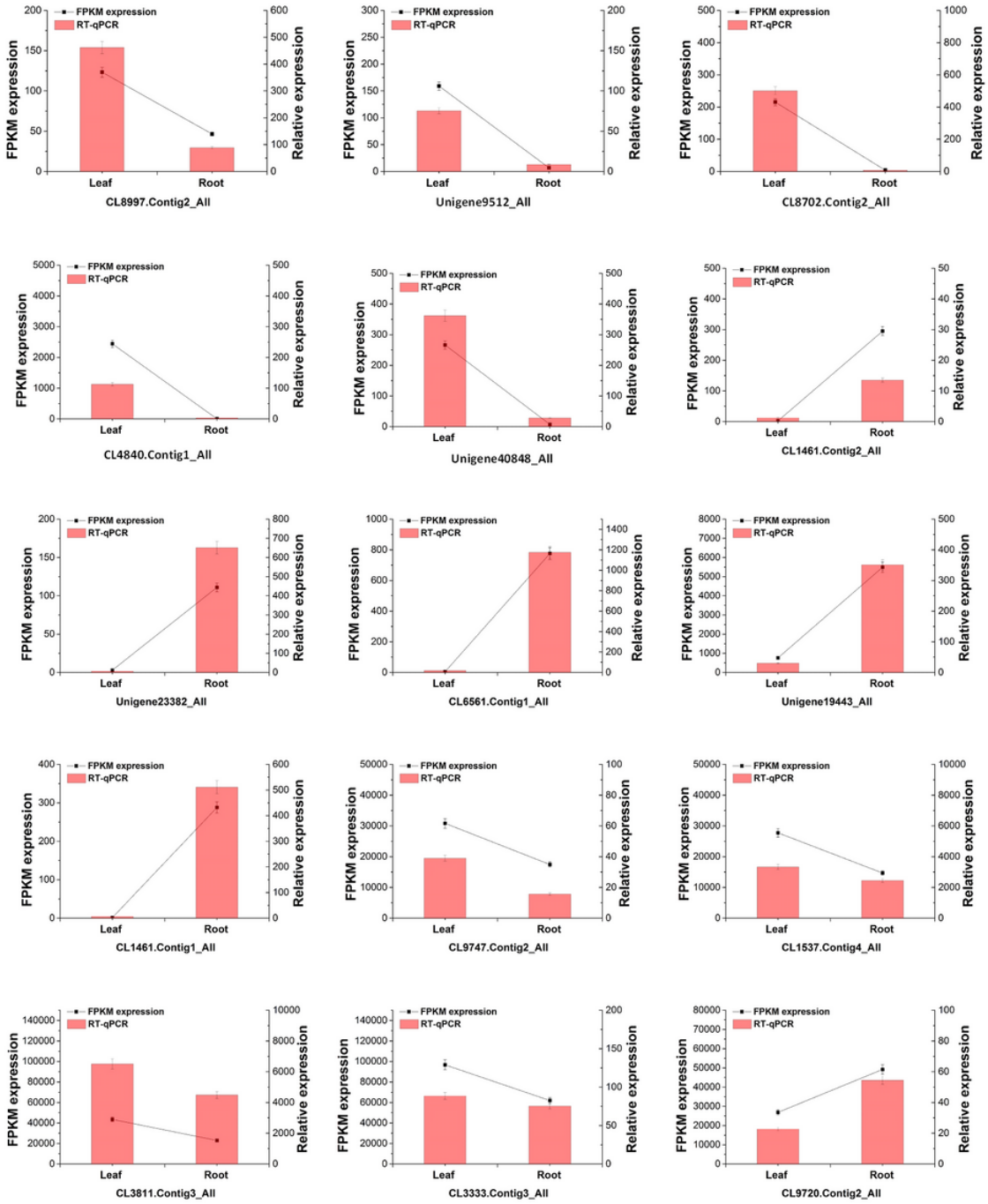

Figure 7

qRT-PCR validation of selected 15 DEGs at root and leaf. The relative expression level of each selected gene was determined by 2-ΔΔCT. Each bar represents the mean ± STD of triplicate assays. Values with different letters indicate significant differences at $P < 0.05$ according to Duncan's multiple range tests

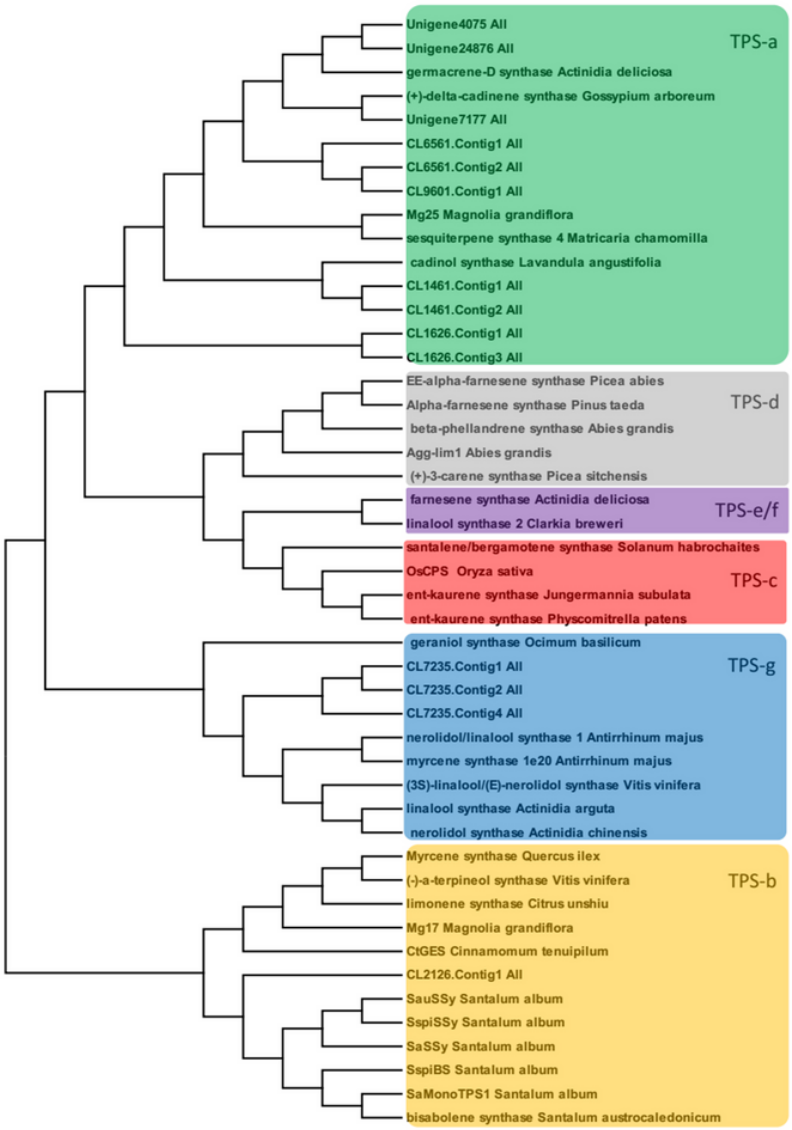

Figure 8

Phylogenetic analysis of sesquiterpene synthases of L. sibiricus. Accession numbers of proteins used in phylogenetic analysis are listed in Additional file 2

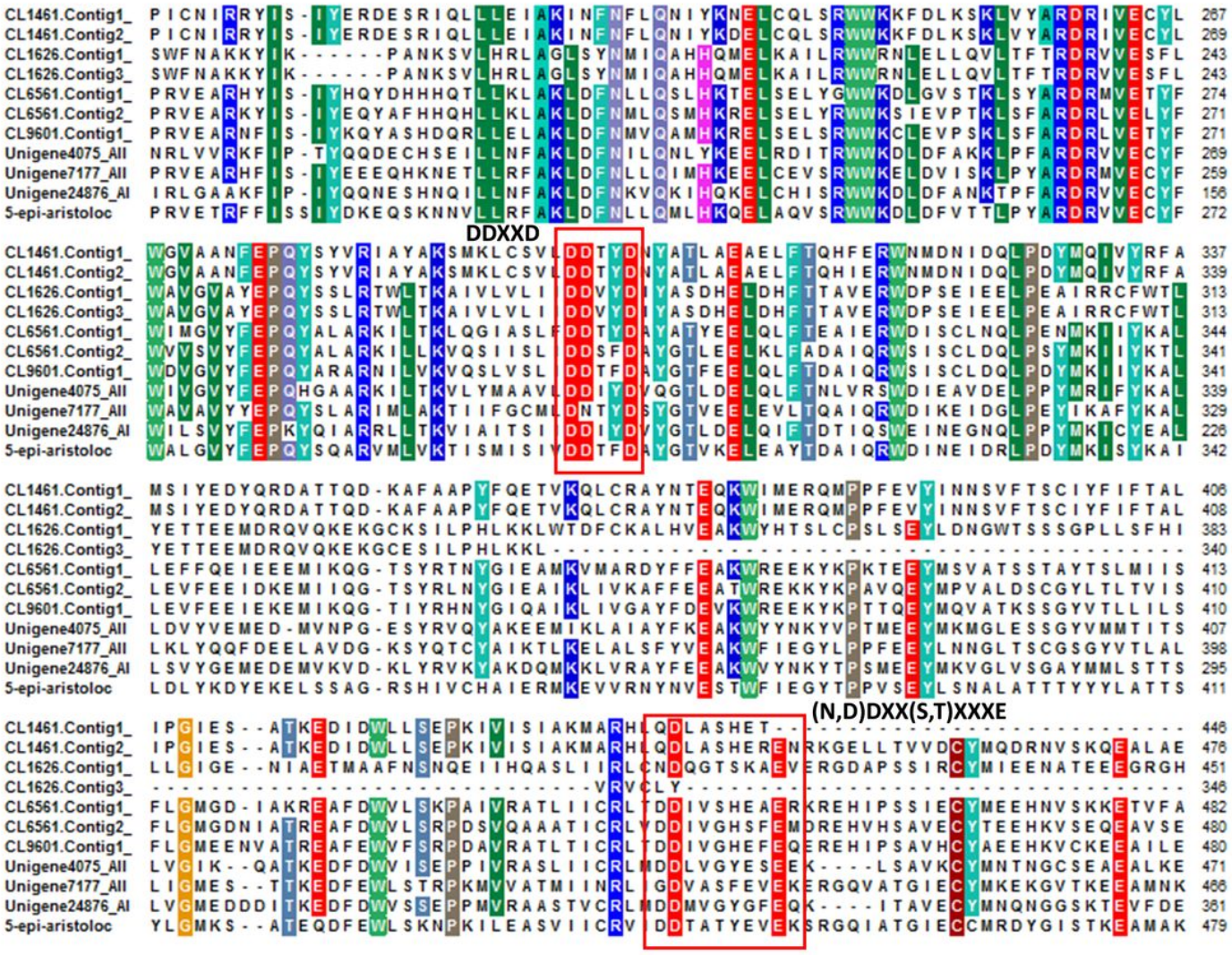

Figure 9

Sequence alignment of putative TPSs of L. sibiricus genes with TEAS (PDB: 5EAS_A). Aspartate-rich metal binding motif are highlighted in red boxes.

Supplementary Files

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

- Additionalfile1.xlsx
- Additionalfile2.xlsx