Full-length transcriptome profiling of *Gentiana straminea* Maxim provides new insights into iridoid biosynthesis pathway (#112216)

First submission

Guidance from your Editor

Please submit by **9 Mar 2025** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files

Download and review all files from the materials page.

14 Figure file(s)
10 Table file(s)
1 Raw data file(s)
1 Other file(s)

Custom checks

DNA data checks

- Have you checked the authors data deposition statement?
- Can you access the deposited data?
- Has the data been deposited correctly?
- Is the deposition information noted in the manuscript?

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- **Impact and novelty is not assessed.** Meaningful replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Full-length transcriptome profiling of *Gentiana straminea* Maxim provides new insights into iridoid biosynthesis pathway

Lina Yang [1], **Tao He** [Corresp., 2, 3], **Le Wang** [Corresp., 3], **Xiaochun Ning** [4], **Shuai Wang** [3]

[1] College of Agriculture and Animal Husbandry, Qinghai University, Xi' ning, Qinghai, China

[2] School of Ecol-Environmental Engineering, Qinghai university, Xi' ning, Qinghai, China

[3] State Key Laboratory of Plateau Ecology and Agriculture, Qinghai university, Xi' ning, Qinghai, China

[4] Xining Center of Natural Resources Comprehensive Survey, China Geological Survey, Xi' ning, Qinghai, China

Corresponding Authors: Tao He, Le Wang
Email address: hetaoxn@aliyun.com, wangleqhu@163.com

Gentiana straminea Maxim is a traditional Chinese medicinal plant renowned for its rich array of bioactive compounds, particularly iridoid glycosides. These compounds are recognized as the main components that exert therapeutic effects against rheumatism,osteoarthritis,hepatitis, gastritis , and cholecystitis . The study of secondary metabolites in *G.straminea* become an exciting area of research, however the genetic factor sunderlying the production and diversification of secondary metabolites in *G.straminea* are still poorly understood, especially the pathway of iridoid biosynthesis. In the present study, full-length transcriptome-based Illumina sequencing was performed to identify genes that differentially expressed in five *G.straminea* tissues, and proteins catalyzing iridoid biosynthesis was characterized. After sequence clustering and redundancy removal, a total of 32,776 isoforms were identified in PacBio sequencing, with an average length of 2589.14bp, an N50 value of 2767bp, and a GC content of 41.43% . Results of illumina sequencing unveiled that a total of 31,330 genes were found in common in all the five tissues. KEGG enrichment analysis revealed that the DEGs were mainly enriched interms related to biosynthesis of secondary metabolites, metabolic pathways, MAPK signaling pathway, etc. A total of 117 isoforms encoding 19 key enzymes related to the iridoid synthesis pathway were identified, including two geranyl diphosphate synthases (GPPS) and four geranylgeranyl diphosphate synthases (GGPPS). A phylogenetic analysis further classified plant G(G)PPSs into three distinct branches. The profiling of tissue-specific expression of key genes involved iniridoid synthesis revealed that the RT-qPCR results demonstrated the consistent trend with the FPKM values of in the root, stem, leaf, flower, ovary, non-embryoniccalli (NEC) and embryoniccalli (EC) . Among them, *AACT, IDI, ISPH, and GCPE* had the highest expression levels in leaves, while *DXS* and *GPPS*

had the highest expression level sinstems. This work provides the first transcriptomic analysis of *G.straminea* , which will be a valuable resource for mechanisms of bioactive medicinal compound formation and molecular and genomic studies of the species.

Full-length transcriptome profiling of *Gentiana straminea* Maxim provides new insights into iridoid biosynthesis pathway

Lina Yang[1], Tao He[2,3], Le Wang[3], Xiaochun Ning[4], Shuai Wang[3]

[1] College of Agriculture and Animal Husbandry, Qinghai University, Xi'ning, Qinghai, China

[2] School of Ecol-Environmental Engineering, Qinghai University, Xi'ning, Qinghai, China China

[3] State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xi'ning, Qinghai,

[4] Xining Center of Natural Resources Comprehensive Survey, China Geological Survey, Xi'ning, Qinghai, China

Corresponding Author:

Tao He

Street Address, Xi'ning, Qinghai, 810016, China

Email address: hetaoxn@aliyun.com

Le Wang

Street Address, Xi'ning, Qinghai, 810016, China

Email address: wangleqhu@163.com

Abstract

Gentiana straminea Maxim is a traditional Chinese medicinal plant renowned for its rich array of bioactive compounds, particularly iridoid glycosides. These compounds are recognized as the main components that exert therapeutic effects against rheumatism, osteoarthritis, hepatitis, gastritis, and cholecystitis. The study of secondary metabolites in *G. straminea* become an exciting area of research, however the genetic factors underlying the production and diversification of secondary metabolites in *G. straminea* are still poorly understood, especially the pathway of iridoid biosynthesis. In the present study, full-length transcriptome-based Illumina sequencing was performed to identify genes that differentially expressed in five *G. straminea* tissues, and proteins catalyzing iridoid biosynthesis was characterized. After sequence clustering and redundancy removal, a total of 32,776 isoforms were identified in PacBio sequencing, with an average length of 2589.14 bp, an N50 value of 2767 bp, and a GC content of 41.43%. Results of illumina sequencing unveiled that a total of 31,330 genes were found in common in all the five tissues. KEGG enrichment analysis revealed that the DEGs were mainly enriched in terms related to biosynthesis of secondary metabolites, metabolic pathways, MAPK signaling pathway, etc. A total of 117 isoforms encoding 19 key enzymes related to the iridoid synthesis pathway were identified, including two geranyl diphosphate synthases (GPPS) and four geranylgeranyl diphosphate synthases (GGPPS). A phylogenetic analysis further classified plant G(G)PPSs into three distinct branches. The profiling of tissue-specific expression of key genes involved in iridoid synthesis revealed that the RT-qPCR results demonstrated the consistent trend with the FPKM values of in the root, stem, leaf, flower, ovary, non-embryonic calli (NEC) and embryonic calli (EC). Among them, *AACT, IDI, ISPH, and GCPE* had the highest expression levels in leaves, while *DXS and GPPS* had the highest expression levels in stems. This work provides the first transcriptomic analysis of *G. straminea*, which will be a valuable resource for mechanisms of bioactive medicinal compound formation and molecular and genomic studies of the species.

Introduction

Gentiana straminea Maxim, which is a member of the Gentianaceae family and termed “Mahuajiao” in Chinese, is used in traditional Chinese medicine (*Ye et al. 2021*). It is distributed mainly in Qinghai, Xizang and Sichuan, as well as other regions. *G. straminea* usually grows in

alpine meadows, forests, and grassland at altitude of 2000 ~ 4950 m (*Jia et al. 2012*). Previous studies have indicated that iridoids from the roots of *G. straminea* have therapeutic effects against rheumatism, osteoarthritis, hepatitis, gastritis, and cholecystitis (*Zhou et al. 2016*). The main medicinal effects are associated with gentiopicroside, loganic acid, sweroside and swertiamarin, which are all iridoids compounds (*Wei et al. 2012; Wu et al. 2016*). Gentiopicroside, considered the main active compound (*Wu et al. 2016*), is predominantly synthesized from iridoids that originate from terpenoid biosynthesis. The annotated metabolites and identified enzymes suggest that the biosynthesis of iridoids is similar to the synthesis of vincristine in *Catharanthus roseus* (*Oudin et al. 2007*).

As reported in early findings, iridoids are an oxygenated monoterpene compounds that are composed of two isopentane units, and their synthesis pathway comprises three stages. The first stage involves synthesis of the precursors isopentyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), which can be formed via the mevalonate pathway (MVA) and the methylerythritol phosphate pathway (MEP). MVA occurs mainly in the cytoplasm, whereas MEP occurs mainly in plastids (*Zhan et al. 2023*). The second stage involves the formation of the iridoid skeleton, during which IPP and DMAPP are catalytically condensed by geranyl diphosphate synthases (GPPS) to produce GPP, while geranylgeranyl diphosphate (GGPP) is generated through a catalytic process mediated by geranylgeranyl diphosphate synthases (GGPPS) (*Eva et al. 2013*). Then, GPP and GGPP are used as the raw materials for the synthesis of different terpenoids (monoterpenes, diterpenes, triterpenes, etc.) through different metabolic pathways (*Sun et al. 2012*). The third stage is the synthesis of iridoids from GPP. GPP is converted to geraniol and by geranyl diphosphate diphosphatase (GES)-mediated catalysis and hydrolysis (*Oudin et al. 2007*). The structure of geraniol is then modified via glycosylation, hydroxylation, methylation, isomerisation and other reactions to form iridoid (*Zhao and Wang 2020*). Because of their abundant pharmacological properties, iridoids have become a research hotspot in relevant studies. However, the biosynthesis pathway of iridoid in *G. straminea* is still unclear. As a result, analyzing the biosynthesis mechanism has become crucial for effectively increasing the levels of key medicinal components in *G. straminea*. It is necessary to acquire the relevant sequences of the target genes involved in iridoid biosynthesis. As sequencing technology developed, transcriptome sequencing has gradually been applied to identify transcripts, discover new genes, and determine which genes are expressed in plants.

The transcriptome consists of all the RNA transcripts of a species and reflects the functions of different cells and tissues during a particular period. Advancements in high-throughput RNA sequencing technologies currently enable the analysis of genes that regulate the synthesis of secondary metabolites in non-model species. This approach can uncover new genes, potential metabolic pathways, and associated genetic regulatory mechanisms (*Ozsolak and milos 2012*). The third-generation single-molecule real-time sequencing (SMRT) enables sequencing of transcripts up to 10kb without a reference genome, however, it is limited by high cost per base, high error rates, and low throughput (*Rhoads and Au, 2015*). Second-generation sequencing produces short read lengths, but provides high sequencing accuracy. Due to limitations imposed by its read length and assembly algorithms, second-generation sequencing cannot accurately obtain the complete sequence of transcripts, particularly for different transcripts with high homology. Consequently, the integration of second- and third-generation sequencing techniques allows for the acquisition of high-quality sequencing results with low error rate. Full-length transcriptome-based Illumina sequencing has been applied in research involving *Coptis deltoidei* (*Zhong et al. 2020*), *Ranunculus japonicus* (*Xu et al. 2023*), *Fritillaria hupehensis* (*Guo et al. 2021*), *Angelica sinensis* (*Gao et al. 20*21), *Torreya grandis*(*Lou et al. 2019*), and *Salvia miltiorrhiza* (*Xu et al. 2016*). In the present study, a combined sequencing strategy was utilized to identify differentially expressed genes (DEGs) in the roots, stems, leaves, flowers, and ovaries of *G. straminea*. Furthermore, genes related to iridoid biosynthesis were characterized.

Materials & Methods

Preparation and collection of samples for transcriptome sequencing and qPCR analysis

Samples of *G. straminea* individuals were collected in August 2023 during the flowering stage from Yushu, Maqin County, Qinghai Province, China (N34°38′380, E100°23′546, altitude 4200 m). Fresh tissues from five types—root, stem, leaves, flower, and ovary—were collected, washed with sterilized water, wrapped in foil, and then preserved in liquid nitrogen. Tissues of these five types were further utilized for library construction and SMRT sequencing. Additionally, for the quantification of gene expression via qPCR, two additional tissue types—non-embryonic calli (NEC) and embryonic calli (EC)—were included. The generation of EC and NEC tissues was carried out following the protocol outlined by *He Tao et al* [21], utilizing leaves as explants. Three biological replicates were collected for all samples.

RNA extraction and SMRT sequencing

Samples of *G. straminea* were used for total RNA extraction on ice, according to the manufacturer's protocol, with TRIzol reagent (Life Technologies, Karlsbad, California/USA). An Agilent 2100 Bioanalyzer and agarose gel electrophoresis were used to determine the integrity of the total RNA. A Nanodrop microspectrophotometer (Waltham, MA, USA; Thermo Fisher) was used to check the purity and concentration of the RNA. The Clontech SMARTer PCR cDNA Synthesis Kit was used to reverse transcribe the oligo (dT) magnetic bead-enriched mRNA to cDNA. PCR cycle optimization was used to determine the optimal number of amplification cycles for downstream large-scale PCRs. Double-stranded cDNAs were generated with the optimized cycle number. Additionally, size selection at >5 kb and equal mixing without size selection of cDNA were performed with the BluePippin^{TM} Size Selection System. The next step in the construction of the SMRTbell library was carried out by large-scale PCR. The sequencing primer was matched with the SMRTbell template by annealing, and then linked to the polymerase. Sequencing was performed on the PacBio Sequel II platform at Gene Denovo Biotechnology Co.

The raw sequencing reads from the cDNA library were classified via the Pacific Biosciences Iso-Seq pipeline, with high-quality CCSs first extracted. Transcript integration was assessed according to whether the CCS reads contained all 5' primers, the 3' primer and the poly-A sequences. Full length sequences (FLs) were those that contained all three sequences. After the removal of primers, barcodes and poly A tails, full-length nonchimeric (FLNC) reads were obtained. Reads less than 50 bp in length were discarded. The entire isoform was generated by clustering the FLNC reads. Minimap2 was used for similar FLNC reads, which were then clustered hierarchically to obtain a consistent sequence. The consistent sequence was then further corrected via the Quiver algorithm. The high-quality isoforms (prediction accuracy ≥ 0.99) were used for subsequent analysis.

Library construction and Illumina sequencing

Total RNA was enriched by Oligo(dT) beads to form mRNA, then was fragmented into short fragments. With random primers, fragments was transcribed into cDNA, then syntheisized the second-strand cDNA with DNA polymerase I, Rnase H, dNTP and buffer. the obtained cDNA was purified with QiaQuick PCR extraction kit (Qiagen, Venlo, The Netherlands), end repaired, poly(A) added, and ligated to Illumina sequencing adapters. The ligated products were

screened by agarose gel electrophoresis, amplified by PCR, and sequenced by Gente Denovo Biotechnology Co. (Guangzhou, China) using Illumina HiSeq^{TM} 4000. High quality clean reads were obtained by fastp (Version 0.18.0), with removing adapters, containing more than 10% of unknown nucleotides (N) and low quality reads.

Isoform expression and differential expression analysis

Using the full-length transcriptome as the reference, the clean and high quality reads were mapped using RSEM (version 1.2.8) to determine the isoform expression in five different tissues of *G. straminea*. The results were expressed in terms of fragments per kilobase per million mapped fragments (FPKM). Differential analysis of gene expression in different tissues was performed by using DESeq2 software, and genes with false discovery rate (FDR) parameter below 0.05 and absolute fold change $\geqslant 2$ were considered as differentially expressed genes**.**

Functional annotation, structure analysis

The sequences of the isoforms were checked against the non-redundant protein (Nr) database of the NCBI (http://www.ncbi.nlm.nih.gov), the COG/KOG database (http://www.ncbi.nlm.nih. gov/COG), the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome. jp/kegg), and the Swiss-Prot protein database (http://www.expasy.ch/sprot) via the BLASTx program (http://www.ncbi.nlm.nih.gov/BLAST/), with an E value threshold of $1e^{-5}$, to assess the similarity of the sequences to those of genes from other species. Gene Ontology (GO) annotation was analyzed using isoforms from the Nr annotation results by Blast2GO software. The top 20 scoring isoforms and no fewer than 33 high-scoring segment pair hits (HSPs) were selected for the Blast2GO analysis. Isoforms were functionally classified using WEGO software. TFs were predicted via hmmscan by aligning the protein coding sequences to the Plant TFdb (http://planttfdb.cbi.pku. edu.cn/). The sequence annotated to iridoids biosynthesis pathway was submitted to string database (https://cn.string-db.org/) for protein interaction analysis.

Identification and bioinformatic analysis of G(G)PPSs in *G. straminea*

For identification of GsG(G)PPSs, local BLAST search was performed using GGPPSs from *Arabidopsis thaliana* (*Beck et al. 2013*) or *Chimonanthus praecox* (*Kamran et al. 2020*) as queries. A threshold of e-value < 10^{-10} was applied for preliminary screening. After searching, sequences of putative GsG(G)PPSs were further subjected to CDD (https://www.ncbi.nlm.nih.gov/cdd/) and InterPro

(https://www.ebi.ac.uk/interpro/result/InterProScan/) for domain confirmation (*Paysan- Lafosse, 2022*). Prediction and analysis of the physicochemical properties of the GsG(G)PPS amino acid sequences were performed via ExPASy (https://web. expasy.org /protparam/) (*Artimo et al. 2012*). Sequences were submitted to SignalP4.1 server for prediction of the signal peptide (https://services.healthtech. dtu.dk/services/SignalP-4.1/) (*Thomas et al, 2011*). Subcellular localization were determined via WoLF PSORT (https://wolfpsort.hgc.jp/), transmembrane structure were predicted by HMHMM2.0 (https://services.healthtech.dtu.dk/services/TMHMM-2.0/). In addition, the annotated sequence information was submitted to the MEME website (http://meme -suit.org), using 6-100 residues as the optimal motif size to search for 10 conserved motifs and predicted the conserved protein motifs in the sequence (*Bailey et al, 2015*). Similar to the GsGGPPS SSU, GsGGPPS, and GsGPPS amino acid sequences were downloaded from NCBI BLAST (https://blast.ncbi.nlm. nih.gov/Blast.cgi). Protein sequence alignment was performed via DNAMAN. Protein structure prediction was performed via SWISS-MODEL (https://swissmodel.expasy. org/).

Phylogentic analysis of G(G)PPS gene family

The GsG(G)PPSs obtained, and G(G)PPSs from other species were incorporated for phylogenetic analysis. G(G)PPSs in *Arabidopsis thaliana* and *Nicotiana tabacum* genomes were identified through BLAST online by using the ensemble database (https://asia.ensembl.org/Multi/ Tools/Blast). G(G)PPS homologues from other species available in the NCBI database were included for phylogenetic analyses of the G(G)PPS family. Details on all the G(G)PPSs used for phylogenetic analysis were listed in supplemental Table S1. Phylogenetic inference of G(G)PPSs was conducted using the neighbor-joining method in MEGA 11.0 software, with a bootstrap test of 1000 replicates (*Tamura et al, 2021*). The refinement of the evolutionary tree was completed using the online software Evoview (https://www.evolgenius.info/evolview/#/).

Expression analysis of key enzymes by real - time quantitative PCR

First-strand cDNA synthesis was performed using a cDNA reverse transcription kit (PrimeScriptTMII 1st Strand cDNA Synthesis Kit), following the protocol provided. Primers for RT-qPCR were designed using the OligoArchitect online sever and synthesized by Sangon Biotech (Shanghai) Co., Ltd. The primers sequence shown in supplementary Table S2, qPCR was performed using TB Green Premix Ex Taq with a 20μL reaction system, which included

10μL of TB Green Premix, 0.8μL each of forward and reverse primers (10 μM), 2μL of cDNA, 6.4μL of ddH_2O. The reaction procedure consisted of the following steps: pre-denaturation at 94°C for 5 minutes, denaturation at 94°C for 30 seconds, annealing at 53°C for 30 seconds, extension at 72°C for 30 seconds, followed by 40 cycles. The GAPDH gene was utilized as the internal reference for relative expression analysis. The quantification of gene expressions was conducted using three biological replicates. Relative expression was calculated using the Ct ($2^{-\Delta\Delta Ct}$) method, following the approach described by Livak and Schmittgen (*Livak and Schmittgen 2001*), with root expression serving as the control. The significance analysis of difference tissue was conducted with means of gene expression by the Duncan test at 5%.

Results

Transcriptome sequencing of *G. straminea*

Both SMRT and Illumine sequencing were performed for the root, stem, leaf, flower and ovary tissues of *G. straminea.* The average amount of raw data generated was 6.5 GB for second-generation sequencing per sample (Table 1). For PacBio sequencing, a total of 62.47 GB of raw data was obtained. A total of 23,318,162 subreads were generated from third-generation sequencing. After self-correction and merging, 499,496 circular consensus sequence (CCS) were formed, with an average CCS read length of 2789 bp, and the number and length distributions of the CCS reads and passes are shown in Fig. S1(a)-(b). The full-length nonchimeric sequences with high-precision CCS reads were identified, and similar FLNC reads were clustered hierarchically to obtain consensus sequences (Fig. S1(c)). A total of 41,785 high-quality isoforms (HQs) and 140 low-quality isoforms (LQs) were obtained after further correction. After removing redundant sequences, the total length of the isoforms was 84,861,577bp, 32,776 isoforms were obtained, and the lengths ranged from 165 to 10169bp, with an average length of 2589.14 bp, an N50 of 2767 bp, and a GC content of 41.43%. The number and length distribution of the isoforms were shown in Fig. S1(d).

Functional annotation of the full-length transcriptome of *G. straminea*

The HQ unigenes were annotated via four functional annotation databases NR, Swiss-Prot, KEGG and KOG. A total of 31,434 (95.9%) unigenes were successfully annotated, while 1342 were unannotated. The highest number of unigenes (31,235; 97.47%) were annotated to Nr database, followed by the KEGG database and the Swiss-Prot database, in which 30,990

(94.55%) and 27,622 (84.27%) unigenes, respectively, were annotated. The lowest number of unigenes were annotated in the KOG database (22,753; 69.42%). Summary, 21,742 common unigenes (66.34%) were annotated in all four databases (Fig. 1a). These findings were compared with those for 414 species annotated in the Nr database (top ten shown in Fig. 1b). The species with the most annotated sequence information was *Coffea arabica,* with 8,701 (27.86%) unigenes, followed by *Coffea eugenioides, Coffea canephora,* and *Olea europaea*, with 5,318 (17.03%), 3,512 (11.24%), and 1,051 (3.36%) unigenes, respectively.

The KOG analysis identified 22,753 unigenes, which could be classified into 25 categories (Fig. 2). The largest number of annotated genes were associated with general function prediction only 4,733 genes (20.80%); followed by 4,146 genes (18.22%) annotated to signal transduction mechanisms; 2,859 genes (12.57%) annotated to posttranslational modifications, protein turnover, and chaperones; 1,617 genes (7.11%) annotated to carbohydrate transport and metabolism; and 1,533 genes (6.74%) annotated to RNA processing and modification. The lowest number was observed for cell motility (37; 0.16%). In addition, 1,210 (5.30%) genes with unknown functions were identified.

The unigenes annotated by GO function analysis were associated with 51 GO terms, which were grouped into 3 categories: cellular component, molecular function, and biological process (Fig. S2). The top three GO enriched terms in the biological process category were cellular process, metabolic process, and response to stimulus, with 21,052, 18,417 and 7,559 genes, respectively. The top three enriched GO terms in the molecular function category were binding (18,864), catalytic activity (16,549), and transporter activity (3,175). In the cellular component category, the cellular anatomical entity (16,552) and protein-containing complex (6946) terms were highly enriched.

In the KEGG database, 30,990 unigenes of *G. straminea* were annotated and divided into 5 categories and 19 subclasses (shown in Table S3). In the KEGG pathway analysis, 9,485 genes were annotated. The greatest number of genes (4,647; 48.99%) were annotated to metabolism pathways, followed by secondary metabolite biosynthesis (2,494; 26.29%), carbon metabolism (826; 8.71%), and biosynthesis of amino acids (635; 6.69%) (Table S4). Genes annotated to secondary metabolite biosynthesis pathways may be related to the synthesis of the medicinal components of *G. straminea*. In addition to carbon metabolism, the biosynthesis of amino acids and other metabolic pathways may be related to cellular osmotic regulation and the oxidative

stress response. These annotated genes provide important sequence information for investigating the biosynthetic mechanism of the metabolites of *G. straminea*.

In this study, 708 annotated genes were found to participate in 20 standard KEGG secondary metabolism pathways in the transcriptome of *G. straminea* (Table S5), of which 121 genes were annotated to the terpenoid backbone biosynthesis pathway and 67 genes were enriched in terpenoids (monoterpenoid, diterpenoid, sesquiterpenoid and triterpenoid biosynthesis) pathways. Second, there were 84 genes involved in phenylpropanoid biosynthesis, and 41 genes were involved in flavonoids, isoflavonoid, flavone and flavonol biosynthesis. In addition, 89 genes related to the synthesis of various alkaloids (indole, isoquinoline, tropane, piperidine, and pyridine alkaloid biosynthesis) were annotated, as shown in Table S5.

Predicting TFs

According to the assembly results, 1,151 genes were annotated to TFs, distributed in 51 TF families. The largest number of genes belonged to the GRAS family, with 128 genes (11.12%), followed by the ARF, C3H, bHLH, and WRKY families, with 92, 70, 66 and 66 genes, respectively. The least common families were the NF-YB (1), M-type (1), Whirly (1), AP2 (1), and YABBY (1) families. The ten TF families with the greatest number of genes in *G. straminea* were shown in Fig. S3.

DEGs analysis

A total of 32,470 genes were detected, and venn diagram analysis revealed that 31,330 genes were commonly found in the five tissues (Fig.3b). In the comparisons of root and stem, root and leaf, root and flower, root and ovary, a total of 9809, 10503, 13195, 9699 DGEs were identified, respectively, of which, 6594, 5762, 6572, 5727 DGEs were up-regulated and 3260, 4741, 6623, 3972 DGEs were down-regulated, respectively. In the contrast between leaf and stem, leaf and flower, leaf and ovary, in sum of 6980, 10475, 10006 DEGs were separately detected, and 4030, 4707, 5002 DEGs were up-reaulated and 2950, 5768, 5004 DEGs were down-regulated, individually. A total of 8855, 7021 DEGs were expressioned in the comparison group of stem vs flower, stem vs ovary, reseparately. Comparing with ovary, in the tissue of flower, 3456 DEGs were up-regulated, and 2218 DEGs were down-regulated, these was shown in Fig.3a.

In the comparisons of five different tissue, DGEs genes annotated were mainly enriched in the metabolic pathways, biosynthesis of secondary metabolites. Secondly, DGEs genes were

enriched in the pentose and glucuronate interconversions in the group of root-vs-flower and root-vs-ovaries; carbon metabolism in leaves-vs-stem and flowers-vs-ovaries; amino sugar and nucleotide sugar metabolism in roots-vs-stem, separately. (Fig.4).

Analysis of iridoid biosynthesis genes in *G. straminea*

Iridoids compound, which are common secondary metabolite components found in various medicinal plants, are the main components of *G. straminea* and have significant biological activity. By combining these results with previous research results (*Ni et al. 2019; Liu et al. 2017*), we identified a putative pathway for iridoid biosynthesis and the isoforms involved (Fig. 5). Our results revealed that 117 isoforms encoded 19 key enzymes (Table S6). According to previous reports, MVD has been identified as an important enzyme in the MVA pathway of iridoid synthesis, HDR is an important enzyme in the MEP pathway, G(G)PPS is the key enzyme for the conversion of IPP and DMAPP to GPP or GGPP, and plays an important role in the formation of geraniol. The expression levels of these key enzymes isoforms in five tissues were shown with heatmap (Fig. 5), of which, *GCPE, STR, ISPE,DXR, ISPH,7-DLS* showed relatively high expression in leaves, other genes showed different expression patterns in different tissues. Protein-protein interaction (PPI) network analysis was performed on the enzymes annotated in the iridoids biosynthesis pathway. The PPI network contained 19 nodes and 114 edges (Fig. 6), average node degree was 9.91, average local clustering coefficient was 0.615, PPI enrichment p-value $< 1.0e^{-16}$.

Bioinformatics analysis of G(G)PPS

Our results revealed that ten isoforms had GPPS/GGPPS annotations, six of which had open reading frames (ORFs). Four genes were annotated as GGPPS, while two genes were annotated as GPPS, among the four GGPPSs, three were categorized as GGPPS small subunits (GGPPS SSU), and one was classified as a typical GGPPS. The prediction results revealed that the amino acid length of G(G)PPS (SSU) ranged from 342 to 424 aa, with corresponding molecular weights in range of 37.47 ~ 46.45 kDa and pI values ranging from 5.81 to 6.48 (Table 1). Four of GsGGPPS (SSU) possessed nagative GRAVY values ranging from -0.050 to -0.187, Indicating that these proteins have hydrophilicity. Two of GsGPPS had a positive GRAVY value (0.049, 0.041), suggesting hydrophobicity of them. Six of GsG(G)PPS (SSU) were identified no signal peptide. Four of GsGGPPS (SSU) were localized in the chloroplast, Two of GsGPPS were predicated to be mitochondrion. No transmembrane structures were detected in all the G(G)PPS

proteins on the basis of the TMHMM2.0 predictions (Table 2). Pfam protein structural domain prediction revealed a distinctive polyprenyl-synt domain shared by all the G(G)PPS proteins (Fig. S4a b).

G(G)PPS usually contains two highly conserved aspartic acid-rich regions-FRAM and SARM with the sequences of $DD(XX)_{1\text{-}2}D$ (D is aspartic acid, and X refers to any amino acid). The first conserved region FRAM (DDXXXXD) is consistent with the binding site of the substrate dimethylallyl diphosphate (DMAPP), and the second conserved region SARM (DDXXD) corresponds to the binding site of the substrate isopentenyl diphosphate (IPP). which affects the catalytic activity of G(G)PPS. Some G(G)GPPS proteins also have the characteristic sequence CXXXC (C is cysteine, and X refers to any hydrophobic amino acid) of them structural domain, which is essential for the interaction of G(G)PPS proteins with other proteins (*Beck et al. 2013*). Sequence alignment results revealed that the GsGGPPS SSU sequences were similar to those of PjGGPPS SSU, AeGGPPS SSU, CaGGPPS SSU, and SiGGPPS SSU2, with identity values of 81.74%, 81.55%, 81.49% and 81.61%, respectively, according to DNAMAN (Table S7). The identities of the GsGGPPS sequences were similar to those of CrGGPPS, CaGPPS, CeGGPPS and GjGGPPS, with values of 74.06%, 72.29%, 72.04% and 71.28%, respectively (Table S7). The GsGPPS sequences were similar to those of CeSPPS, CaSPPS, CrGPPS1, CrGPPS2, GsyFPPS, SiSPPS and NaSPPS, with identities of 91.84%, 91.76%, 92%, 91.84%, 91.53%, 90.68%, and 90.75%, respectively (Table S7). GsGGPPS SSU1~GsGGPPS SSU3 were enriched with one FARM ($DD(XX)_2D$) and two CXXXC regions. The GsGGPPS subunit underwent a change in the second aspartic acid enrichment motif, from D to E, i.e, DDXXE (Fig. S4a). GsGGPPS was enriched with one FARM region ($DD(XX)_2D$), one SARM region each (DDXXD), and one CXXXC region (Fig. S4a). GsGPPS was enriched with two SARM regions (DDXXD) (Fig. S4b).

The analysis of the conserved motifs in GsG(G)PPS revealed that all of them contained conserved motif 1, 2, and 4, with the exception that both GsGGPPS SSU and GsGPPS also included conserved motif 7. GsGGPPS SSU contained additional conserved motifs 3, 5, and 6, while GsGPPS included conserved motifs 8, 9, and 10 (Fig. 7). The difference in motif composition may influence the function of GsG(G)PPSs, leading to changes in catalytic activity, protein subcellular localization, and other aspects.. The results of the study revealed that the protein tertiary structure of GsGGPPS SSU1 was highly similar to that of the *Mucuna pruriens*

(velvet bean) template (A0A371F419), with a GMQE value of 0.85; GsGGPPS was similar to that of *Handroanthus impetiginosus* GGPPS (A0A2G9GV50), with a GMQE value of 0.81; and GsGPPS1 and GsGPPS2 were similar to those from *C. roseus* GPPSs (B2MV87), with a GMQE value of 0.79. GsG(G)PPS mainly contained α-helices, and random coils in the tertiary structure (Fig. S5).

Phylogentic analysis of G(G)PPSs

Phylogenetic analysis revealed that the G(G)PPSs identified can be categorized into three distinct branches. Among them, GsGGPPS, together with large subunits of GGPPS (GGPPS LSUs) and GGPPSs from other species, clustered into group 1. GsGGPPS SSU1 to GsGGPPS SSU3, along with the small subunits of GGPPS (GGPPS SSUs) from other species, were grouped into the second branch (group 2). GsGPPSs formed the third branch, together with GPPS, SPPS, and FPPS from various other species (group 3) (Fig. 8). GsG(G)PPSs were categorized into three distinct groups based on their sequence and functional divergence.

Expression analysis by real-time quantitative PCR

To further identify the candidate genes involved in the iridoid synthesis pathway, *AACT, DXS, IDI, MVD, ISPH, GCPE*, and *GPPS* were selected for RT-qPCR analysis. As shown in Fig. 9, the seven genes demonstrated the trend of increased-decreased and increased again in seven different tissue from RT-qPCR results. Among them, *AACT, IDI, ISPH, and GCPE* had the highest expression levels in leaves, while *DXS and GPPS* had the highest expression levels in stems. The relative expression quantitation demonstrated the consistent trend with the FPKM values of *DXS, IDI, ISPH, GCPE, and GPPS* genes in different tissues. The expression levels of *DXS, IDI, MVD, ISPH, and GPPS* in NEC were generally higher than those in EC tissues. Except for *MVD*, the expression levels of other six genes showed significant differences in the seven tissues. Differential expression of these genes may result in varying iridoid content in different tissues.

Discussion

As a medicinal plant, *G. straminea* contains various iridoids compound, as the main active substances, its mainly synthesized through terpenoids. Abundant transcripts annotated to the synthesis of secondary metabolites, especially terpenoid backbone biosynthesis accounted for 121 genes. Compared with the results obtained for *G. straminea* via Illumina NGS, and *Gentiana*

waltonii and *Gentiana robusta* via the Illumina Hiseq X Ten platform (*Ni et al. 2019*), we obtained more annotation information, which could enrich the gene library of *G. straminea*, making it more extensive and complete.

TFs can regulate gene expression by recognizing specific DNA sequences in gene promoters, which is important for understanding gene expression regulatory mechanisms (*Jose et al. 2016*). In plants, the GRAS, bHLH and WRKY families are common TF families, which are related to hormone metabolism and secondary metabolism. Based on the annotated results, most of the TFs distributed in the GRAS, ARF, C3H, bHLH, WRKY and FAR1 families. The SmDELLA1 protein of the GRAS gene family in *Salvia miltiorrhiza* was found to be a positive regulatory factor in total phenolic acid and flavonoid biosynthesis (*Li et al. 2024*). In addition, DELLA also participates in the regulation of jasmonic acid (JA) signaling and cell wall formation (*Hou et al. 2010; Wang et al. 2021*). bHLH was constitute the second largest class of TFs in angiosperms, they are ubiquitous in various eukaryotes participates in plant epidermal differentiation, environmental stress response and secondary metabolism regulation, and are a key regulators of anthocyanin biosynthesis in a variety of plants (*Jaakola et al. 2013*). Previous studies have shown MYB can regulate the terpenoid alkaloids produced (*Zhao et al. 2013*), GmbHLH can positived regulated the biosynthesis of loganic acid (*Fu et al. 2024*). WRKYs are unique to plants, and the highly conserved N-terminal domains can specifically integrate into the promoter region of target genes, and then activate the expression of downstream genes (*Brand et al. 2013*). For example, AaWRKY1 isolated from *Artemisia carvifolia* can integrate into the cis-acting W-box element in the promoter region of ADS, promoting artemisinin biosynthesis via the activation of the expression of the key enzyme sesquiterpene synthase (*Ma et al. 2009*).

Iridoids are present in traditional medicinal plants and regulate various diseases in the human body. The synthesis of iridoids has been reported in *C. roseus* (*Oudin et al. 2007*), *Gentiana rigescens* (*Zhang et al. 2015*), *Valeriana jatamansi* (*Zhao and Wang. 2020*), *Swertia mussotii* (*Liu et al. 2017*) and *Rehmannia glutinos*a (*Sun et al. 2012*). In our study, 117 isoforms involved in 19 key enzymes were annotated, which contained different stages of iridoid synthesis. In *S. mussotii*, 24 enzyme categories associated with 39 transcripts were identified (*Liu et al. 2023*), in *Gentiana lhassica,* 171 unigenes were annotated as encoding 27 key enzymes (*Heng et al. 2021*), and in *V. jatamansi* Jones, 24 unigenes were identified and classified into 24 enzyme categories associated with three metabolic pathways leading to iridoid biosynthesis

(*Zhao and Wang et al. 2020*). In *Panax ginseng* (*Kim et al. 2014*) and *Ganoderma lucidum* (*Shi et al. 2012*), overexpressed MVD could significantly increase the accumulation of terpenoids in plants. The overexpression of HDR gene in *Artemisia annua* (*Ma et al. 2017*) and *Ginkgo biloba* (*Kim et al. 2021*) could significantly increase the terpenoids content. In the present study, the seven genes of *AACT, DXS, IDI, MVD, ISPH,GCPE and GPPS* demonstrated the same trend between RT-qPCR results and FPKM values in seven different tissue, and genes associated with iridoid synthesis were most abundant in stem and leaf tissues. Avanish Rai *et al*. compared the expression of *GPPS* across different tissues (root, stem, leaf, flower, silique) of *C. roseus* and discovered that GPPS exhibited the highest expression in the flower, followed by the stem (*Rai et al. 2013*). Zhou *et al* (*Zhou et al. 2016*) found that GPPS exhibited higher expression levels in the flowers comparing to root. The genes related to the iridoid synthesis pathway exhibit differential expression in various tissues of different species.

Some researchers have reported that IPP and DMAPP form GPP under the catalytic action of GPPS for monoterpene synthesis, whereas under the catalytic effect action of GGPPS, they form GGPP for diterpene synthesis, triterpene synthesis, etc. (*Tholl et al. 2004; Liang et al.2002*). Since both of them act on the same substrate, some scholars have hypothesized that the IPP flow direction determines the different products (*Tholl et al. 2004*). In the third stage, geraniol is formed via the action of GES, and then 10-hydroxygeraniol is formed via the catalytic action of G10H (*Liang et al. 2002*). The genes encoding G10H in *C. roseus* (*Krithika et al. 2015*). and *S. mussotii* (*Wang et al. 2010*) have been cloned. Although the G10H gene was not annotated in our results, cytochrome P450 reductase (CPR, POR, EC1.6.2.4) was annotated; this enzyme is the partner of G10H, in the catalytic production of 10-hydroxygeraniol from geraniol. Some scholars have reported that cytochrome P450 monooxygenases (P450s) which constitute one of the major families of enzymes can catalyze the conversion of geraniol to loganic acid (*Wang et al. 2010; Collu et al. 2001*). For example, CYP76B6 from *C. roseus* (*Hofer et al. 2013*), and CYP76B10 from *S. mussotii* (*Wang et al. 2010*) are considered to have the same catalytic activity for production of 10-hydroxygeraniol. The catalytic activity of most cytochrome P450s in eukaryotes depends on their partner in the reduction process, cytochrome P450 reductase (CPR, POR, EC1.6.2.4). This gene expression profile was similar to that of G10H, and the genes presented similar kinetics to jasmonic acid induction (*Hofer et al. 2013*). It is possible that G10H is a member of the cytochrome P450 monooxygenase family, and almost

all plant CYP450s relay on the electron cytochrome P450 reductase provided by the oxidation reduction partner NADPH cytochrome (*Durst and Nelson 1995*). Peng *et al* reported that geraniol was converted to 10-hydroxygeraniol under the catalytic action of cytochrome P450 reductase and G10H in *R. glutinosa* (*Sun et al. 2012*). Therefore, the POR annotated in this study may catalyze geraniol formation.

GPP synthase catalyzed the conversion of DMAPP and IPP to GPP, and it is a member of the short chain prenyltransferase family. Both FPPS and GGPPS belong to this group, they play a regulatory role in IPP flux (*Durst and Nelson 1995*). Our result revealed that the amino acid sizes, molecular weights and isoelectric points of GsG(G)PPS annotated in this study were essentially similar to those reported for GGPPS in *S. miltiorrhiza* (*Li et al. 2024*), *Liriodendron tulipifera* (*Zhang et al. 2021*) and wintersweet flower (*Kamran et al. 2020*). The characteristic conserved motif of GsGGPPS SSU1~GsGGPPS SSU3 was consistent with that of CpGPPS.SSU2 and CpGPPS.SSU1 reported in wintersweet flower (*Kamran et al. 2020*). The GsGGPPS was consistent with the LtuGGPPS2 reported in the *Liriodendron tulipifera* (*Zhang et al. 2021*) and the CpGPPS reported in wintersweet flower plants (*Kamran et al. 2020*). The characteristic conserved motif of GsGPPS was similar to other characteristics of homologous GPPSs (*Kamran et al. 2020*).

G(G)PPS was shown to exist in both homologous and heterologous forms in the plant material (*Chen et al. 2015*), heterodimeric G(G)PPS contained one LSU and one SSU, and the LSU of the heterodimeric GPPS showed 50%-75% sequence similarity to that of GGPPS and possessed isopentenyl transferase activity, which catalyzes the production of mainly GGPP, as well as a small amount of GPP and FPP (*Tholl et al. 2004; Kamran et al. 2020*). However, the heterodimeric GPPS SSU shares little sequence similarity to with GGPPS, only 22%-38%, lacks the $DD(XX)_{1\text{-}2}D$ motif, and shows no isoprenyl transferase activity (*Tholl et al. 2004*). Five full-length GPPS and GGPPS genes were successfully annotated in the wintersweet flower transcriptome, these genes were classified into three branches by phylogenetic analysis, namely the SSU representing the heterodimeric GPPS and the homodimeric GPPS and GGPPS (*Kamran et al. 2020*).

It has been shown that the LSU of GGPPS can combine with the inactive SSU of GPPS to form a heterodimer, and then catalyzed the synthesis of monoterpene precursor substances. For example, homologous and heterologous GPP synthetases have been identified in *C. roseus*, and

classified as the LSU of CrGPPS, the SSU of heterologous CrGPPS, and homologous CrGPPS, the LSU of CrGPPS is bifunctional in the formation of GPP and GGPP, whereas the inactive SSU of CrGPPS can integrate with CrGPPS LSU, increasing enzyme activity, and result in the production of only GPP (Rai et al. 2013). It was hypothesized that the inactive SSU of the heterodimeric CrGPPS interacting with the bifunctional G(G)PPS redirected metabolic flux towards, and thus acting as an important regulator of monoterpene indole alkaloid biosynthesis (*Zhang et al. 2021*). It has been shown that the synthesis of monoterpenes in flowers is dependent on the heterodimeric rather than the homodimeric G(G)PPS in Arabidopsis (*Orlova et al. 2010*). In addition, It has been reported that GGPPS is involved in heterodimer formation and promotes monoterpene synthesis in *Antirrhinum majus* (*Tholl et al. 2004*) and *C. roseus* (*Zhang et al. 2021*). In tobacco, Overexpression of AmSSU increased the activity of total GPPS enzymes in leaves and flowers and promoted monoterpene formation (*Orlova et al. 2010*). On the basis of the above analysis, both homodimeric and heterodimeric G(G)PPS are clearly related to the formation of monoterpenes in different plant species, and the LSU of heterodimeric G(G)PPS may promote monoterpene formation either by binding to the SSU or by acting as a homodimer to regulate the flow of IPPs, leading to the formation of different products. However, the reason for this phenomenon in *G. straminea* is still unknown, and further studies of related genes in the future will provide new insights into this process.

Conclusions

Based on the third-sequencing through PacBio, and second-sequencing with the Illumina HiSeq™ 4000, the full-length transcriptome and the differentially expressed in five *G. straminea* tissues was performed in this study. A total of 32,776 full-length transcripts of high-quality without redundancy were obtained, and 31,434 isoforms were annotated in the NR, KEGG, KOG and Swiss-Prot databases. Illumina sequencing revealed 31,330 genes common expressed in five tissues of *G. straminea.* DEGs were mainly enriched in biosynthesis of secondary metabolites, metabolic pathways, MAPK signaling pathway, etc from the result of KEGG enrichment. In summary, 708 genes were classified into 20 KEGG secondary metabolism pathways in the transcriptome of *G. straminea*. All genes involved in the biosynthesis of iridoids were screened, and a total of 117 isoforms were annotated into the iridoid synthesis pathway, resulting in the identification of key genes encoding 19 enzymes. RT-qPCR results shown that *AACT, IDI, ISPH*, and *GCPE* had the highest expression levels in leaves, while *DXS* and *GPPS* had the

highest expression levels in stems; *DXS, IDI, MVD, ISPH*, and *GPPS* exhibited the highest expression levels in NEC than in EC, RT-qPCR results shown similar trend with the expression abundance in seven tissue. The polyprenyl-synt domain was highly conserved in both the identified GsGGPPSs and GsGPPSs. Through phylogenetic analysis, the GsG(G)PPSs annotated in this study can be classified into three branches. These new results provide valuable information for further research on functional gene development, and active ingredient accumulation patterns in *G. straminea*.

Acknowledgements

This research was supported by the Youth Fund projects, Qinghai University (No. 2020-DNY-7) and Central Leading Local Science and Technology Development Fund Project (2024ZY030).

Author contribution

Sample collection: Xiaochun Ning, Lina Yang. Conceived and designed the experiments: Lina Yang, Tao He, Le Wang. Performed bioinformatic analysis: Le Wang, Lina Yang, Shuai Wang. Wrote the paper: Lina Yang, Tao He, Le Wang.

Data available statement

The raw sequence data reported in this paper have been deposited in Genome Sequence Archive, China National Center for Bioinformation, the accession number is CRA017932 and CRA019968 that are publicly accessible at (https://download.cncb.ac.cn/gsa4/CRA017932 and https://download.cncb.ac.cn/gsa4/CRA019968/)

Conflict of interest

The authors declare that they have no conflicts of interest.

Ethical approval

This article does not contain any studies with animals or human participants performed by any of the authors.

References

542 Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G. 2012. ExPASy: SIB bioinformatics
543 resource portal. *Nucleic Acids Research*.40 (Web Server issue): W597-W603. doi:
544 10.1093/nar/gks400.
545 Beck G, Coman D, Herren E. guila Ruiz-Sola M, Rodríguez-Concepción M, GruissemW,
546 Vranová E. 2013. Characterization of the GGPP synthase gene family in *Arabidopsis thaliana*.
547 *Plant Molecular Biology* 82:393–416. https://doi.org/10.1007 /s11103-013 -0070-z.
548 Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Research* 43
549 (W1):W39-49. https://doi:10.1093/nar/gkv416
550 Brand L H, Fischer N M, Harter K, Kohlbacher O and Wanke D. 2013. Elucidating the
551 evolutionary conserved DNA-binding specificities of WRKY transcription factors by molecular
552 dynamics and in vitro binding assays. *Nucleic Acids Research* 41(21):9764-9778. https://doi.org/
553 10.1093/nar/gkt732
554 Collu G, Unver N, Peltenburg-Looman AM, Heijden RVD, Verpoorte R, Memelink J. 2001.
555 Geraniol 10-hydroxylase, a cytochrome P450 enzyme involved in terpenoid indole alkaloid
556 biosynthesis. *FEBS Letters* 508:215-220. https://doi.org/10.1016/ s0014-5793 (01)03045-9
557 Chen QW, Fan DJ and Wang GD. 2015. Heteromeric geranyl (geranyl) diphosphate synthase is
558 involved in monoterpene biosynthesis in Arabidosis flowers. *Molecular Plant* 8(9):1434-1437.
559 https://doi.org/10.1016/j.molp.2015.05.001
560 Durst F and Nelson D R. 1995. Diversity and evolution of plant P450 and P450-reductases.*Drug*
561 *Metabol Drug Interact* 12(3):4. https://doi.org/10.1515/DMDI. 1995.12.3-4.189
562 Eva Vranová, Coman D and Gruissem W. 2013. Network Analysis of the MVA and MEP
563 Pathways for Isoprenoid Synthesis. *Annual Review Plant Biology* 64(1):665-670.
564 https://doi.org/10.1146/ annurev-arplant-050312-120116
565 Fu HH , Wang YM , Mi FK, Wang L, Yang Y, Wang F, Yue ZG, He YH. 2024. Transcriptome
566 and metabolome analysis reveals mechanism of light intensity modulating iridoid biosynthesis in
567 *Gentiana macrophylla* Pall. *BMC Plant Biology* 24(1).526. DOI:10.1186/s12870-024-05217-y.
568 Guo KY, Chen J, Niu YQ and Lin XM. 2021. Full-Length Transcriptome Sequencing Provides
569 Insights into Flavonoid Biosynthesis in *Fritillaria hupehensis. Life* 11(4):287. https://doi.org/
570 10.3390/life11040287

571 Gao X, Guo FX, Chen Y, Bai G, Liu YX, Jin JQ and Wang Q. 2021. Full-length transcriptome
572 analysis provides new insights into the early bolting occurrence in medicinal *Angelica sinensis.*
573 *Scientific Reports* 11(1):13000. https://doi.org/10.1038 /S41598-021-92494-4
574 He T, Yang LN, Zhao ZG. 2011. Embryogenesis of *Gentiana straminea* and assessment of
575 genetic stability of regenerated plants using inter simple sequence repeat (ISSR) marker. *African*
576 *Journal of Biotechnology* 10(39):7604–7610.
577 Hou XL, Yun L, Lee C, Xia KF, Yan YY and Yu H. 2010. DELLAs Modulate Jasmonate
578 Signaling via Competitive Binding to JAZs. *Developmental Cell* 19(6): 884-894.
579 https://doi.org/10.1016/ j.devcel.2010.10.024
580 Heng K, Zhao ZL, Ni LH, Li WT, Zhao SJ and Liu TH. 2021. Transcriptome analysis and
581 validation of key genes involved in biosynthesis of iridoids in *Gentiana lhassica. China Journal*
582 *Chinese Materia Medica* 46(18):4704-4711. https://doi.org/10.19540 /j.cnki. cjcmm.20210610.
583 101
584 Hofer R, Lemeng D, Andre F and Ginglinger J F, Lugan R, Gavira C, Grec S, Lang G, Memelink
585 J, and Krol SVD. 2013. Geraniol hydroxylase and hydroxygeraniol oxidase activities of the
586 CYP76 family of cytochrome P450 enzymes and potential for engineering the early steps of the
587 (seco)iridoid pathway. *Metabolic Engineering* 20: 221-232. https://doi.org/10.1016/j.
588 ymben.2013.08.001
589 Jia N, Li, YW, Wu Y, Xi MM, Hur GM, Zhang XX, Cui J, Sun WJ and Wen AD. 2012.
590 Comparison of the anti-inflammatory and analgesic effects of *Gentiana macrophylla* Pall. and
591 *Gentiana straminea* Maxim, and identification of their active constituents. *Journal of*
592 *Ethnopharmacology* 144(3):638-645. https://doi.org /10.1016/j.jep.2012. 10.004
593 Jose M, Franco Z and Roberto S. 2016. Identification of plant transcription factor target
594 sequences, Gene Regulatory Mechanisms. *Biochimica et Biophysica Acta-biomembranes*
595 1860(1):21-30. https://doi.org/ 10.1016/j.bbagrm.2016.05.001
596 Kamran HM, Hussain SB, Shang JZ, Xiang L and Chen LQ. 2020. Identification and Molecular
597 Characterization of Geranyl Diphosphate Synthase (GPPS) Genes in Wintersweet Flower. *Plants*
598 9: 666. https://doi.org/10.3390/plants9050666
599 Kim Y K, Kim Y B, Uddin M R, Lee S H, Kim S U and Park S U. 2014. Enhanced triterpene
600 accumulation in *Panax ginseng* Hairy roots overexpressing mevalonate-5-pyrophosphate

601 decarboxylase and farnesyl pyrophosphate synthase. *ACS Synthetic Biology* 3(10):773-779.
602 https://doi.org/ 10.1021/sb400194g
603 Kim YB, Kim SM, Sathasivam R, Kim YK and Kim SU. 2021. Overexpression of *Ginkgo biloba*
604 Hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase 2 gene (GbHDR2) in *Nicotiana*
605 *tabacum* cv. Xanthi. *3. Biotech* 11(7): 337. https://doi.org/ 10.1007 /s13205-021-02887-5
606 Krithika R, Srivastava PL, Rani B, Kolet SP, Chopade M, Soniya M, Thulasiram HV. 2015.
607 Characterization of 10-Hydroxygeraniol Dehydrogenase from *Catharanthus roseus* Reveals
608 Cascaded Enzymatic Activity in Iridoid Biosynthesis. *Scientific Reports* 5:8258.
609 https://doi.org/10.1038 /srep08258
610 Lou HQ, Ding HQ, Wu MZ, Zhang JS, Zhang FC, Chen WC, Yang Y, Suo JW, Yu WW, Xu CM
611 and Song LL. 2019. Full-Length Transcriptome Analysis of the Genes Involved in Tocopherol
612 Biosynthesis in *Torreya grandis*. *Journal of Agricultural and Food Chemistry* 67(7): 1877-1888.
613 https://doi.org/10.1021/ acs.jafc.8b06138
614 Livak KJ and Schmittgen TD. 2001. Analysis of Relative Gene Expression Data Using Real Time
615 Quantitive PCR and the $2^{-\triangle\triangle CT}$ Method. *Methods* 25:402-408. https://doi. org/10.1006
616 /meth.2001.1262
617 Liu Y, Wang Y, Guo FX, Zhan L, Mohr T, Cheng P, Huo NX, Gu RH, Pei DN, Sun JQ, Tang L,
618 Long CL, Huang LQ and Gu YQ. 2017. Deep sequencing and transcriptome analyses to identify
619 genes involved in secoiridoid biosynthesis in the Tibetan medicinal plant *Swertia mussotii,*
620 *Scientific Reports* 7:43108. https://doi.org /10.1038/srep43108
621 Li YY, Pang QY, Li B, Fu YC, Guo MY, Zhang CJ, Tian Q, Hu SY, Niu JF and Wang SQ. 2024.
622 Characteristics of CXE family of *Salvia miltiorrhiza* and identification of interactions between
623 SmGID1s and SmDELLAs. *Plant Physiology and Biochemistry* 206: 108140.
624 https://doi.org/10.1016/ j.plaphy.2023.108140.
625 Liang PH, Ko TP, Wang AH. 2002. Structure, mechanism and function of prenyltransferases.
626 *European journal of Biochemistry* 269: 3339-3354.https://doi.org/ 10.1046 /j.1432-
627 1033.2002.03014.x
628 Ma DM, Pu GB, Lei CY, Ma LQ, Wang HH, Guo YW, Chen JL, Du ZG, Wang H, Li GF, Ye HC
629 and Liu BY. 2009. Isolation and characterization of AaWRKY1, an *Artemisia annua* transcription
630 factor that regulates the amorpha-4,11-diene synthase gene, a key gene of artemisinin
631 biosynthesis. *Plant and Cell Physiology* 50(12): 2146-2161. https://doi.org/10. 1093/ pcp/pcp149

Ma DM, Li G, Zhu Y, and Xie DY. 2017. Overexpression and suppression of *Artemisia annua* 4-hydroxy-3-methylbut-2-enyldiphosphate reductase 1 gene (AaHDR1) differentially regulate artemisinin and terpenoid biosynthesis. Frontiers in Plant Science 8: 77. https://doi.org/10. 3389/fpls.2017. 00077

Ni LH, Zhao ZL, Wu JR, GAAWE Dorje and Ma M. 2019. Analysis of transcriptomes to explore genes contributing to iridoid biosynthesis in *Gentiana waltonii* and *Gentiana robusta* (Gentianaceae). *Acta Pharmaceutica Sinica B* 54(5): 944-953

Oudin A, Courtois M, Rideau M and Clastre M. 2007. The iridoid pathway in *Catharanthus roseus* alkaloid biosynthesis. *Phytochemistry Reviews* 6(2):259-276. https://doi.org/10. 1007/s11101-006- 9054-9

Ozsolak F and Milos PM. 2012. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* 12: 87-98. https://doi.org/10.1038/nrg 2934

Orlova I, Nagegowda DA, Kish CM, Gutensohn AM and Haeda AH. 2010. The small subunit of snapdragon geranyl diphosphate synthase modifies the chain length specificity of tobacco geranylgeranyl diphosphate synthase in planta. *Plant* 21(12):4002-4017. https://doi.org/10. 1105/ TPC. 109.071282

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunić I, Marchler-Bauer A, Mi H, Natale DA, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A. 2022. InterPro in 2022. *Nucleic Acids Research*. https://doi: 10.1093/nar/gkac993

Rhoads A and Au KF. 2015. PacBio sequencing and its applications. Genomics, *Proteomics. Bioinformatics* 3(5): 278-289. https://doi.org/10.1016/j.gpb. 2015.08.002

Rai A, Smita SS, Singh AK and Shanker K and Nagegowda DA. 2013. Heteromeric and Homomeric Geranyl Diphosphate Synthases from *Catharanthus roseus* and Their Role in Monoterpene Indole Alkaloid Biosynthesis. *Molecular Plant* 6(5):1531-1549. https://doi.org /10.1093/ mp/sst058

Sun P, Song SH, Zhou LL, Zhang B, Qi JJ and Li XE. 2012. Transcriptome Analysis Reveals Putative Genes Involved in Iridoid Biosynthesis in *Rehmannia glutinosa. International Journal of Molecular Sciences* 13(10):13748-13763. https://doi.org/ 10.3390/ijms131013748

662 Sun L, Luo HT, Bu DC, Zhao GG, Yu KT, Zhang CH, Liu YM, Chen RS and Zhao Y. 2013.
663 Utilizing sequence intrinsic composition to classify proteincoding and long non-coding
664 transcripts, *Nucleic Acids Research* 41(17):166. https://doi.org /10.1093/nar /gkt646
665 Shi L, Qin L, Xu YJ, Ren A, Fang X, Mu DS, Tan Q and Zhao MW. 2012. Molecular cloning,
666 characterization, and function analysis of a mevalonate pyrophosphate decarboxylase gene from
667 *Ganoderma lucidum*. *Molecular Biology Reports* 39(5): 6149-6159. https://doi.org/10.
668 1007/s11033-011-1431-9
669 Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne & Henrik Nielsen. 2011.
670 discriminating signal peptides from transmembrane regions. *Nature Methods*. 8:785-786.
671 Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis
672 Version 11. *Molecular Biology and Evolution* 38 (7):3022-3027. https://doi: 10.1093/molbev
673 /msab120
674 Tholl D, Kish CM, Orlova I, Sherman D, Gershenzon J, Pichersky E and Dudareva N. 2004.
675 Formation of monoterpenes in *Antirrhinum majus* and *Clarkia breweri* flowers involves
676 heterodimeric geranyl diphosphate synthases. *Plant Cell* 16(4):977-992. https://doi.org/10.
677 1105/tpc.020156
678 Wei SH, Zhang PC, Feng XZ, Kodama H, Yu CY and Chen G. 2012. Qualitative and quantitative
679 determination of ten iridoids and secoiridoids in *Gentiana straminea* Maxim. by LC-UV-ESI-
680 MS. *Journal of Natural Medicines* 66:102-108. https://doi.org /10.1007/s11418-011-0560-8
681 Wu JR, Zhao ZL, Wu LH and Wang ZT. 2016. Authentication of *Gentiana straminea* Maxim and
682 its substitutes based on chemical profiling of iridoids using liquid chromatography with mass
683 spectrometry. *Biomedical Chromatography* 30:2061–2066. https://doi.org/10.1002/bmc.3763
684 Wang Y.i, Yu WT, Ran LF, Chen Z, Wang CN, Dou Y, Qin YY, Suo QW, Li YH, Zeng JJ, Liang
685 AM, Dai YL, Wu YP, Qu XF and Xiao YH. 2021. DELLA-NAC Interactions Mediate GA
686 Signaling to Promote Secondary Cell Wall Formation in Cotton Stem. *Frontier in Plant Science*
687 12(12):655127. https://doi.org/10.3389/ fpls.2021.655127
688 Wang J, Liu Y, Cai Y, Zhang F, Xia G, and Xiang F. 2010. Cloning and functional analysis of
689 geraniol 10-hydroxylase, a cytochrome P450 from *Swertia mussotii* Franch. *Bioscience*
690 *Biotechnology and Biochemistry* 74(8):1583. https://doi.org/10.1271 /bbb.100175

691 Xu J, Shan TY, Zhang JJ, Zhong XX and Tao YJ. 2023. Full-length transcriptome analysis
692 provides insights into flavonoid biosynthesis in *Ranunculus japonicus. Physiologia*
693 *Plantarum*175(4): 1-13. https://doi.org/10.1111/ppl.13965
694 Xu ZC, Luo HM, Ji AJ, Zhang X, Song JY and Chen SL. 2016. Global identification of the full-
695 length transcripts and alternative splicing related to phenolic acid biosynthetic genes in *Salvia*
696 *miltiorrhiza. Frontier in Plant Science* 7:100-110. https://doi.org/10.3389/fpls.2016. 00100
697 Ye HG, Li CY, Ye WC, Zeng FY, Liu FF, Liu YY, Wang FG, Ye YS, Fu L, Li JR. 2021.
698 Medicinal Angiosperms of Gentianaceae. *Common Chinese Materia Medica*. DOI:10.1007/978-
699 981-16-5900-3_8.
700 Yue J, Tan YQ, Wei RJ, Wang X, Mubeen S, Chen CN, Cao S, Wang CJ, Chen P. 2024.
701 Genome-wide identification of bHLH transcription factors in Kenaf (*Hibiscus cannabinus* L.) and
702 gene function analysis of HcbHLH88. *Physiology Molecular Biology of Plants.*
703 https://doi.org/10.1007/ s12298- 024-01504-y
704 Zhou DW, Gao S, Wang H, Lei TX, Shen JW, Gao J, Chen SL, Yin J and Liu JQ. 2016. De novo
705 sequencing transcriptome of endemic *Gentiana straminea* (Gentianaceae) to identify genes
706 involved in the biosynthesis of active ingredients. *Gene* 575:160-170. https://doi.org/10.1016/
707 j.gene.2015. 08.055
708 Zhan T, Li F, Lan J, Lin LH, Yang ZR, Xie CZ, Wang HB, Zheng XS. 2023. Functional
709 characterization of four mono-terpene synthases (TPSs) provided insight into the biosynthesis of
710 volatile monoterpenes in the medicinal herb *Blumea balsamifera*. *Physiology and Molecular*
711 *Biology of Plants* 29:459 – 469. https://doi. org/10.1007/s12298-023-01306-8
712 Zhao S and Wang CS. 2020. Deep sequencing and transcriptome analyses to identify genes
713 involved in iridoid biosynthesis in the medicinal plant *Valeriana jatamansi* Jones. *Not. Bot. Horti*
714 *Agrobot. Cluj-Napoca* 48(1):189-199. https://doi.org /10.15835 /nbha 48111759
715 Zhong FR, Huang L, Qi LM, Ma YT and Yan ZY. 2020. Full-length transcriptome analysis of
716 *Coptis deltoidea* and identification of putative genes involved in benzylisoquinoline alkaloids
717 biosynthesis based on combined sequencing platforms. *Plant Molecular Biology* 102(5):477–499.
718 https://doi.org/10.1007/s11103-019-00959-y
719 Zhao L, Gao LP, Wang HX, Chen XT, Wang YS, Yang H, Wei CL, Wan XC and Xia T. 2013.
720 The R2R3-MYB, bHLH, WD40, and related transcription factors in flavonoid biosynthesis, *Funct*
721 *Integr Genomic* 13(1):75–98. https://doi.org/10.1007/s10142-012-0301-4

722 Zhang XD, Allan A, Li CH, Wang YZ, Yao QY. 2015. De Novo Assembly and Characterization
723 of the Transcriptome of the Chinese Medicinal Herb, *Gentiana rigescens, International Journal of*
724 *Molecular Sciences* 16:11550-11573. https://doi. org/10.3390 /ijms160511550
725 Zhang CG, Liu HH, Zong YX, Tu ZH and Li HG. 2021. Isolation, expression, and functional
726 analysis of the geranylgeranyl pyrophosphate synthase (GGPPS) gene from *Liriodendron*
727 *tulipifera. Plant Physiol and Bioch* 166:700-711. https://doi.org/10.1016 /j. plaphy.2021.06.052
728
729
730

Figure 1

Figure. 1 Venn diagram and species distribution

(a) Venn diagram showing the number of unigenes annotated to four databases; (b) The top ten species distribution annotated in theN r database

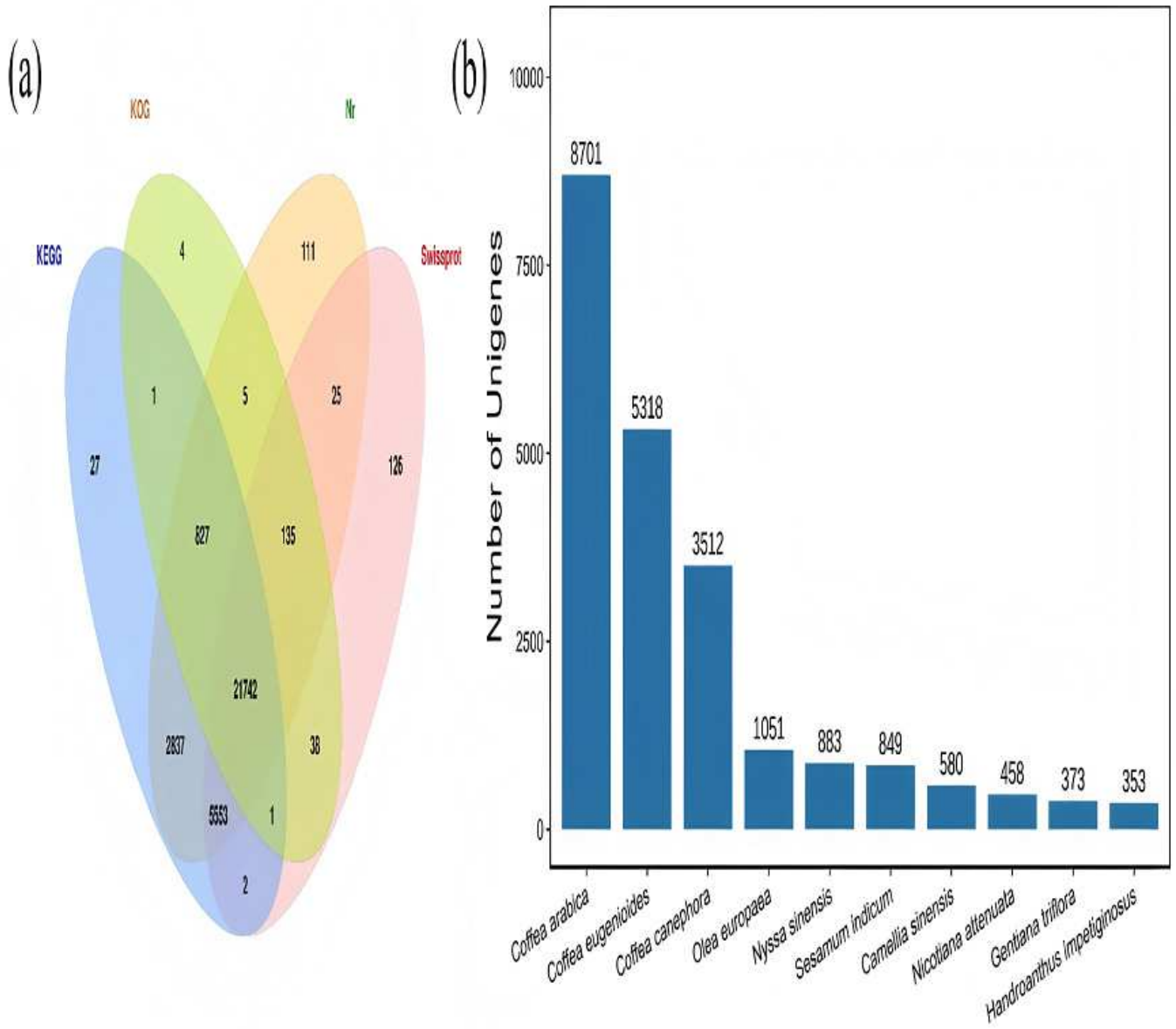

(a)
KOG
Nr
KEGG
Swissprot
4
111
1
5
25
27
126
827
135
21742
2837
38
5553
1
2
(b)
Number of Unigenes
10000
7500
5000
2500
0
8701
5318
3512
1051
883
849
580
458
373
353
Coffea arabica
Coffea eugenioides
Coffea canephora
Olea europaea
Nyssa sinensis
Sesamum indicum
Camellia sinensis
Nicotiana attenuata
Gentiana triflora
Handroanthus impetiginosus

Figure 2

Figure . 2 K OG function classification

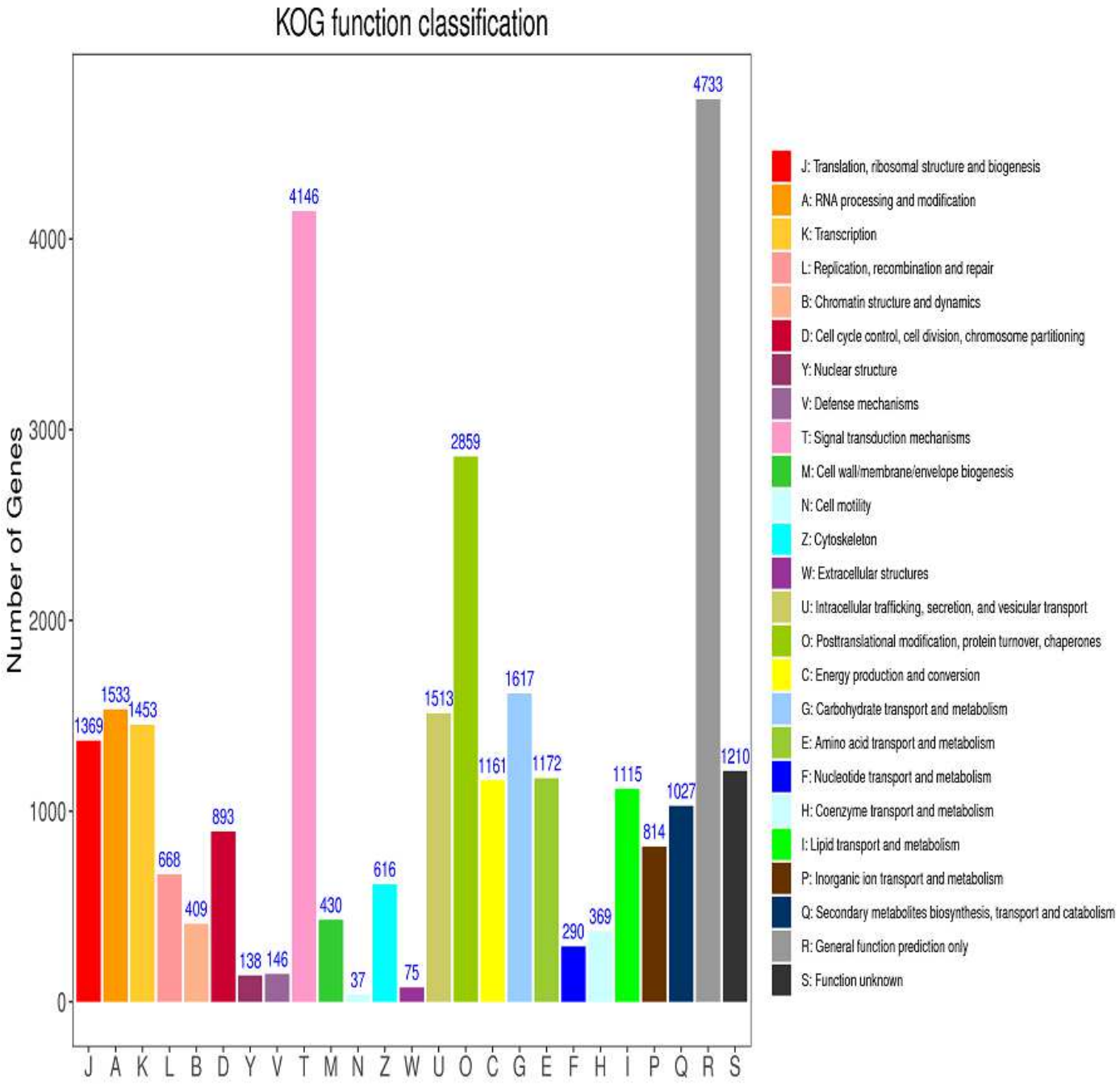


Figure 3

Fig ure . 3 Distribution of the number of DEGs gene expression in different group and Venn diagram of DEGs genes in different tissue

(a) Up-regulated and down-regulated number distribution of DEGs gene expression in different group (b) Venn diagram of DEGs genes in different tissue

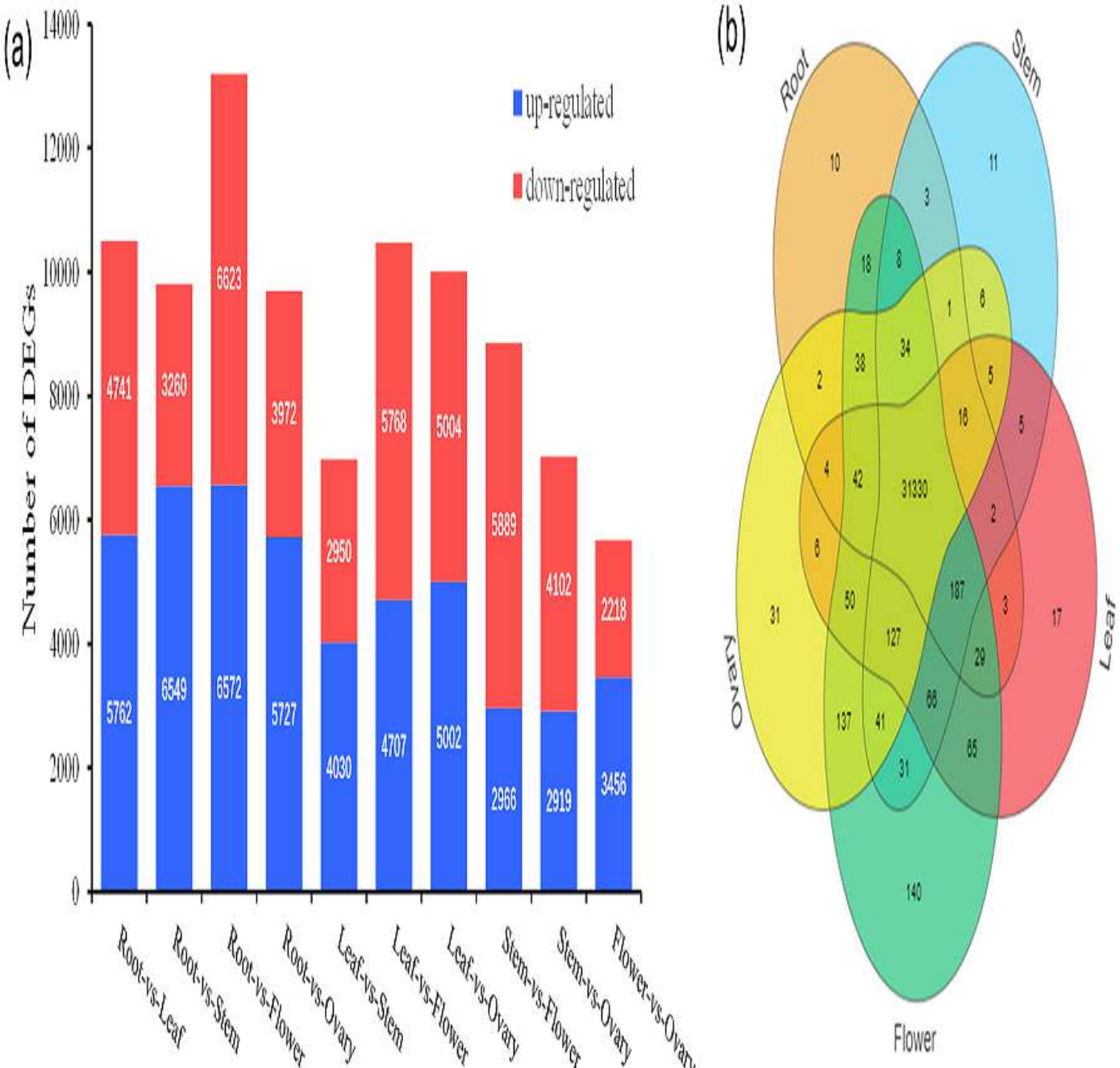

(a)
Number of DEGs
up-regulated
down-regulated
14000
12000
10000
8000
6000
4000
2000
0
4741
5762
3260
6549
6623
6572
3972
5727
2950
4030
5768
4707
5004
5002
5889
2966
4102
2919
2218
3456
Root-vs-Leaf
Root-vs-Stem
Root-vs-Flower
Root-vs-Ovary
Leaf-vs-Stem
Leaf-vs-Flower
Leaf-vs-Ovary
Stem-vs-Flower
Stem-vs-Ovary
Flower-vs-Ovary
(b)
Root
Stem
Ovary
Leaf
Flower
10
11
3
18
8
1
6
2
38
34
5
4
42
31330
16
5
2
8
50
187
3
17
31
127
29
66
137
41
65
31
140

Figure 4

Figure.4 KEGG pathway enrichment of DEGs in different groups

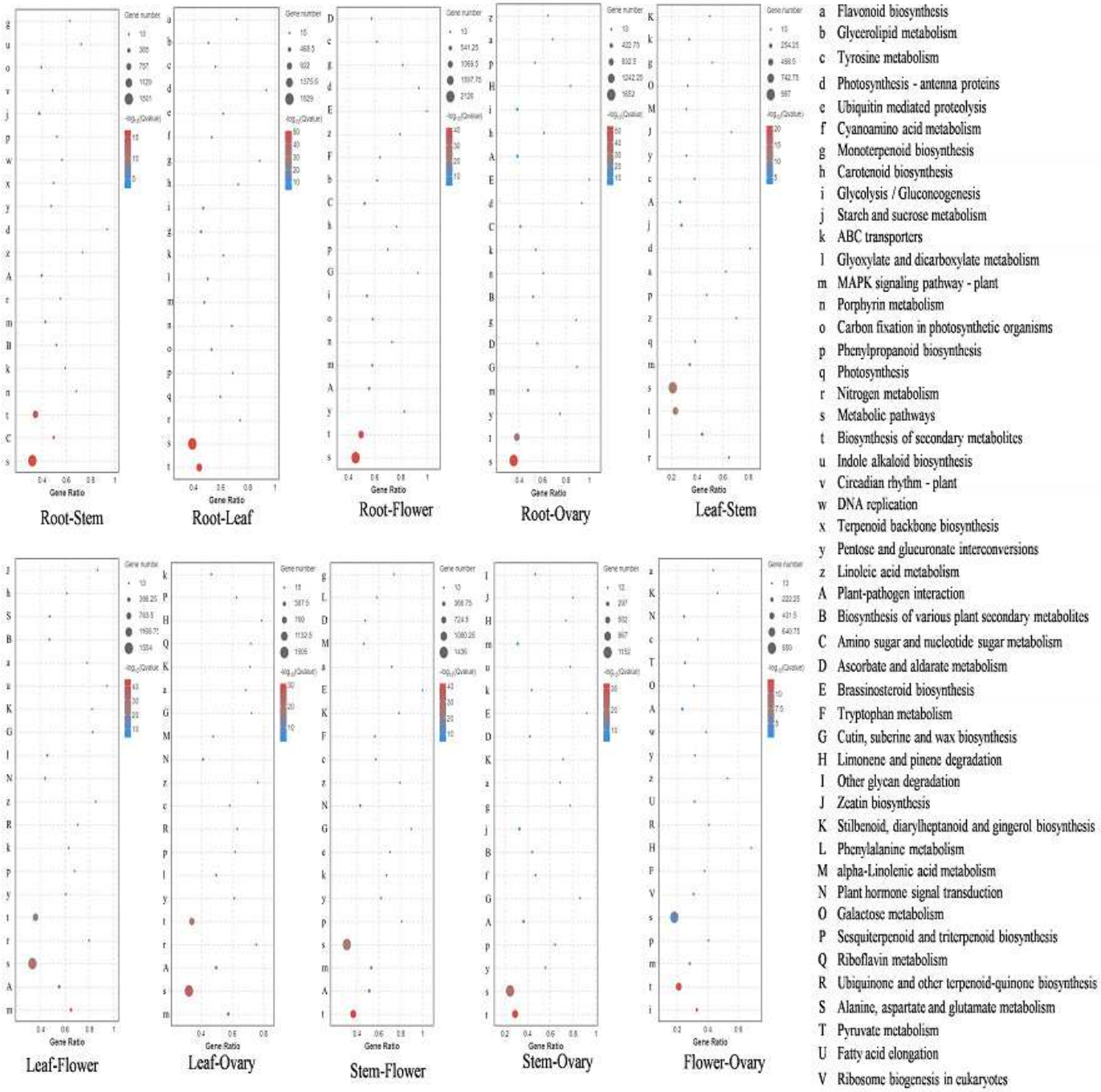


Figure 5

Fig ure . 5 Putative pathways and heatmap of isoforms related to iridoids biosynthesi s in different tissue

Note: Enzymes labelled in red are annotated in *G.staminea* , while those labelled in black are unannotated, The number of isoforms in *G.staminea* is indicated by the red number on the bracket. H eatmap was drawn based on the FPKM values of gene expression levels in different tissues ; R, root; S, stem; L, leaf; F, flower; O, ovary; NEC, non-embryonic calli; EC, embryonic calli. AACT: Acetyl-CoAC-acetyltransferase; HMGCS: Hydroxymethylglutaryl-CoAsynthase; HMGCR: Hydroxymethylglutaryl-CoAreductase(NADPH); MVK: Mevalonatekinase; PMK:Phosphomevalonatekinase; MVD: Diphosphomevalonatedecarboxylase; IDI: Isopentenyl-diphosphatedelta-isomerase; DXS: 1-Deoxy-D-xylulose-5-phosphatesynthase; DXR: 1-Deoxy-D-xylulose-5-phosphatereductoisomerase; ISPD: 2-C-methyl-D-erythritol4-phosphatecytidylyltransferase; ISPE: 4-Diphosphocytidyl-2-C-methyl-D-erythritolkinase; ISPF: 2-C-methyl-D-erythritol2,4-cyclodiphosphatesynthase; GCPE: (E)-4-Hydroxy-3-methylbut-2-enyl-diphosphatesynthase; ISPH: 4-Hydroxy-3-methylbut-2-en1yldiphosphatereductase; GPPS : Geranyldiphosphatesynthase ; GGPPS: Geranylgeranyldiphosphatesynthase ; GES: Geranyldiphosphatediphosphatase; POR: CytochromeP450reductase; G10H: Geraniol 10 -hydroxylase; 10-HG O : 10 -Hydroxygeranioloxidoreductase; ISY1: Iridoidsynthase; 7-DLS: 7-Deoxyloganeticacidsynthase; 7-DLGT: 7-Deoxyloganeticacidglucosyltransferase; 7-DLH: 7-Deoxyloganicacidhydroxylase; LAMT: LoganicacidO-methyltransferase; SLS: Secologaninsynthase; STR: Strictosidinesynthase .

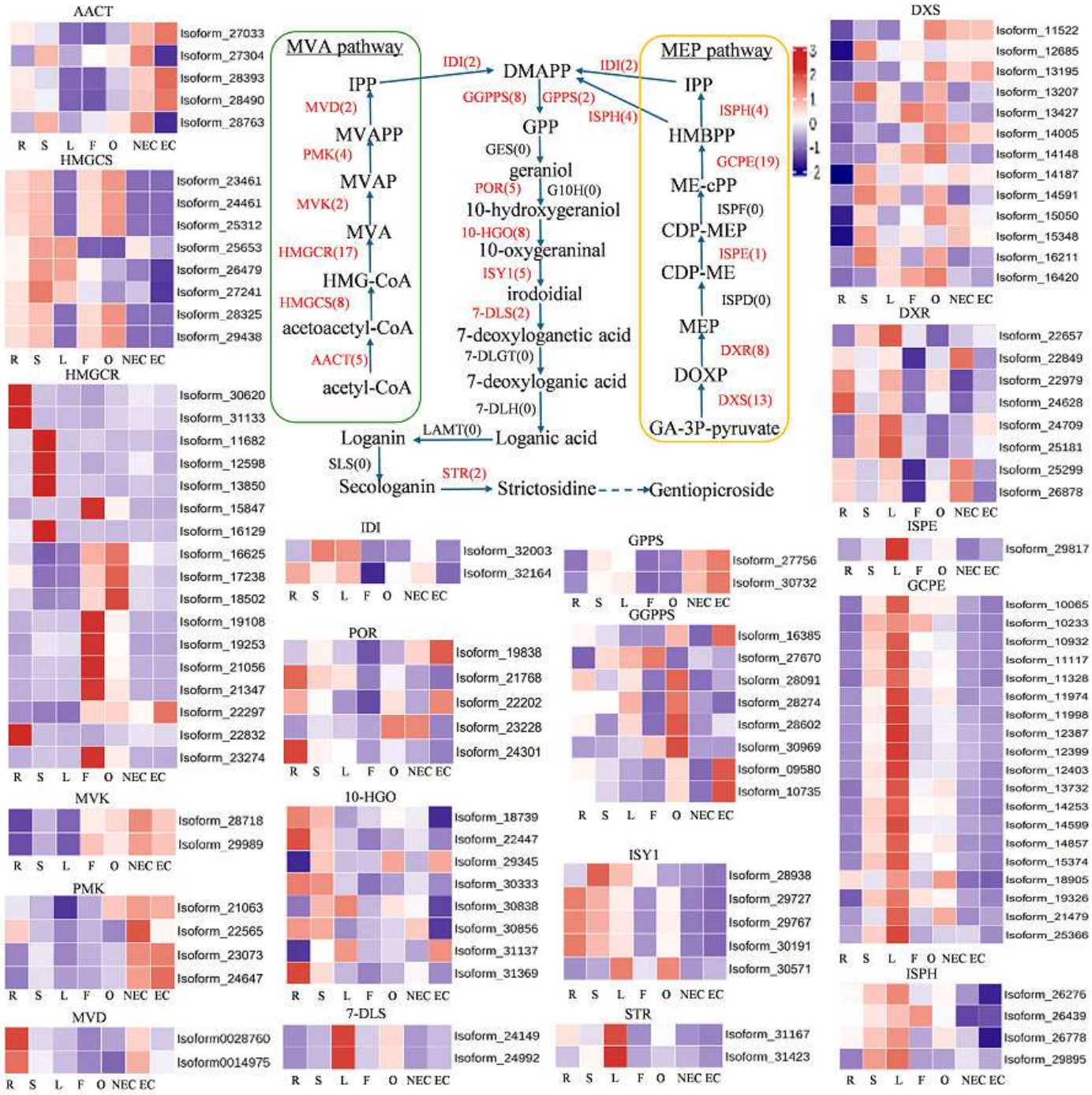

MVA pathway
MEP pathway
IPP
DMAPP
MVAPP
MVAP
MVA
HMG-CoA
acetoacetyl-CoA
acetyl-CoA
MVD(2)
PMK(4)
MVK(2)
HMGCR(17)
HMGCS(8)
AACT(5)
IDI(2)
GGPPS(8)
GPPS(2)
ISPH(4)
GPP
GES(0)
geraniol
POR(5)
G10H(0)
10-hydroxygeraniol
10-HGO(8)
10-oxygeranianl
ISY1(5)
irodoidial
7-DLS(2)
7-deoxyloganetic acid
7-DLGT(0)
7-deoxyloganic acid
7-DLH(0)
Loganic acid
LAMT(0)
Loganin
SLS(0)
Secologanin
STR(2)
Strictosidine
Gentiopicroside
HMBPP
ME-cPP
CDP-MEP
CDP-ME
MEP
DOXP
GA-3P-pyruvate
GCPE(19)
ISPF(0)
ISPE(1)
ISPD(0)
DXR(8)
DXS(13)
AACT
HMGCS
HMGCR
MVK
PMK
MVD
IDI
POR
10-HGO
7-DLS
GPPS
GGPPS
ISY1
STR
DXS
DXR
ISPE
GCPE
ISPH
R S L F O NEC EC

Figure 6

Figure.6 The interaction network with key enzymes annotated to the iridoids biosynthesis pathway of *G.straminea*

Line thickness indicates the strength of data support

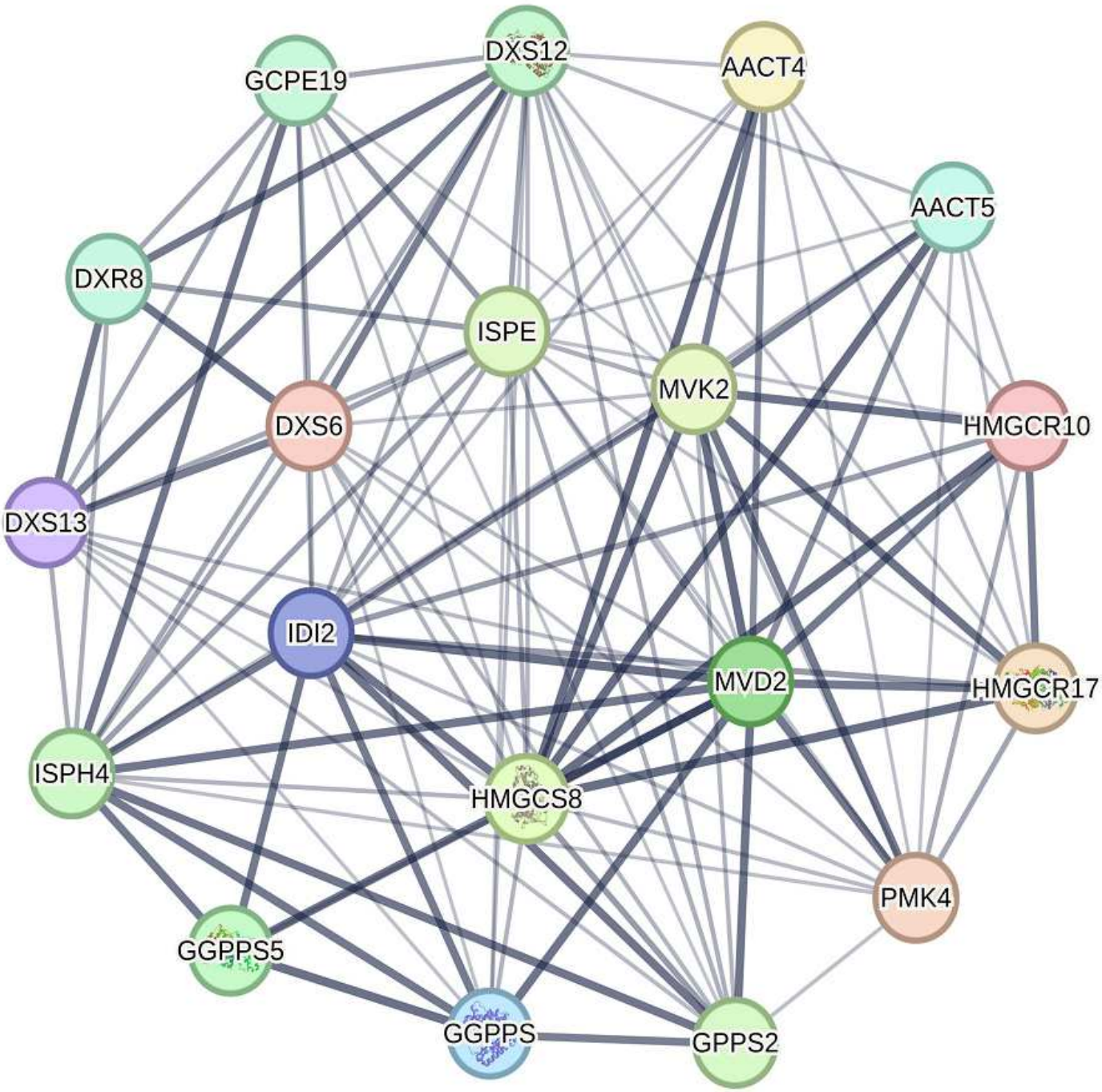
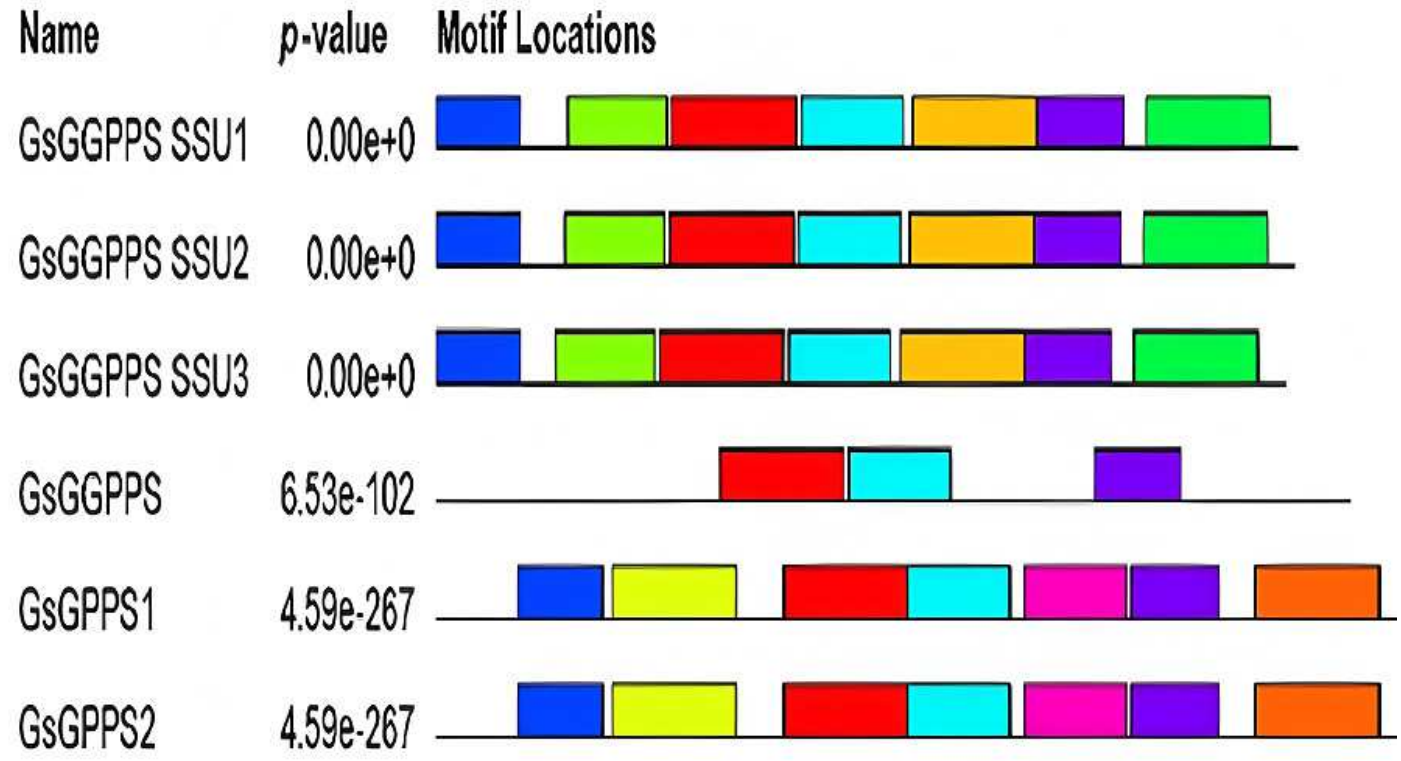

DXS12
GCPE19
AACT4
AACT5
DXR8
ISPE
MVK2
DXS6
HMGCR10
DXS13
IDI2
MVD2
HMGCR17
ISPH4
HMGCS8
PMK4
GGPPS5
GGPPS
GPPS2

Figure 7

Fig ure. 7 Conserved motif analysis from G(G)PPS of *G.straminea*



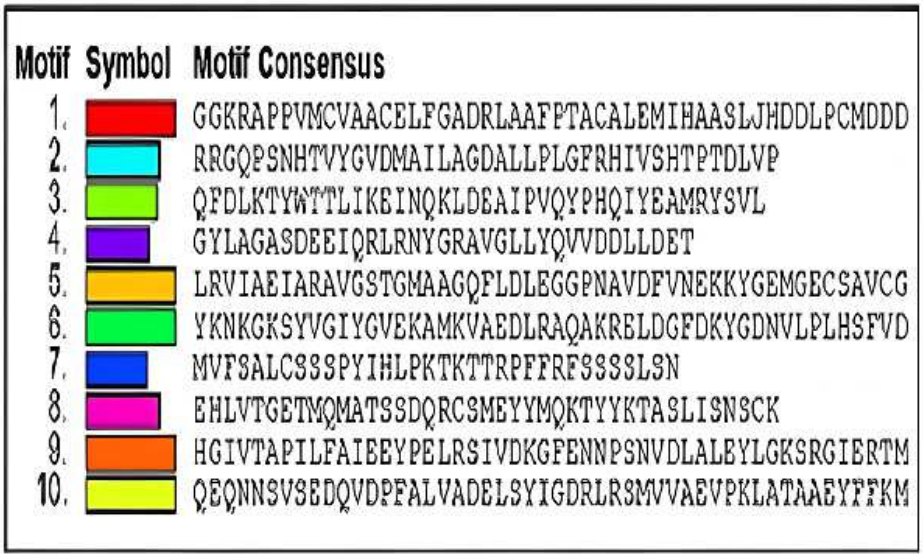


Figure 8

Fig ure . 8 Phylogenetic tree of G(G)PPS gene family in different species

The abbreviations and sequence ID of G(G)PPS gene family are shown in the table S1. Red triangles, red circles and red stars indicate d the proteins annotated in this study

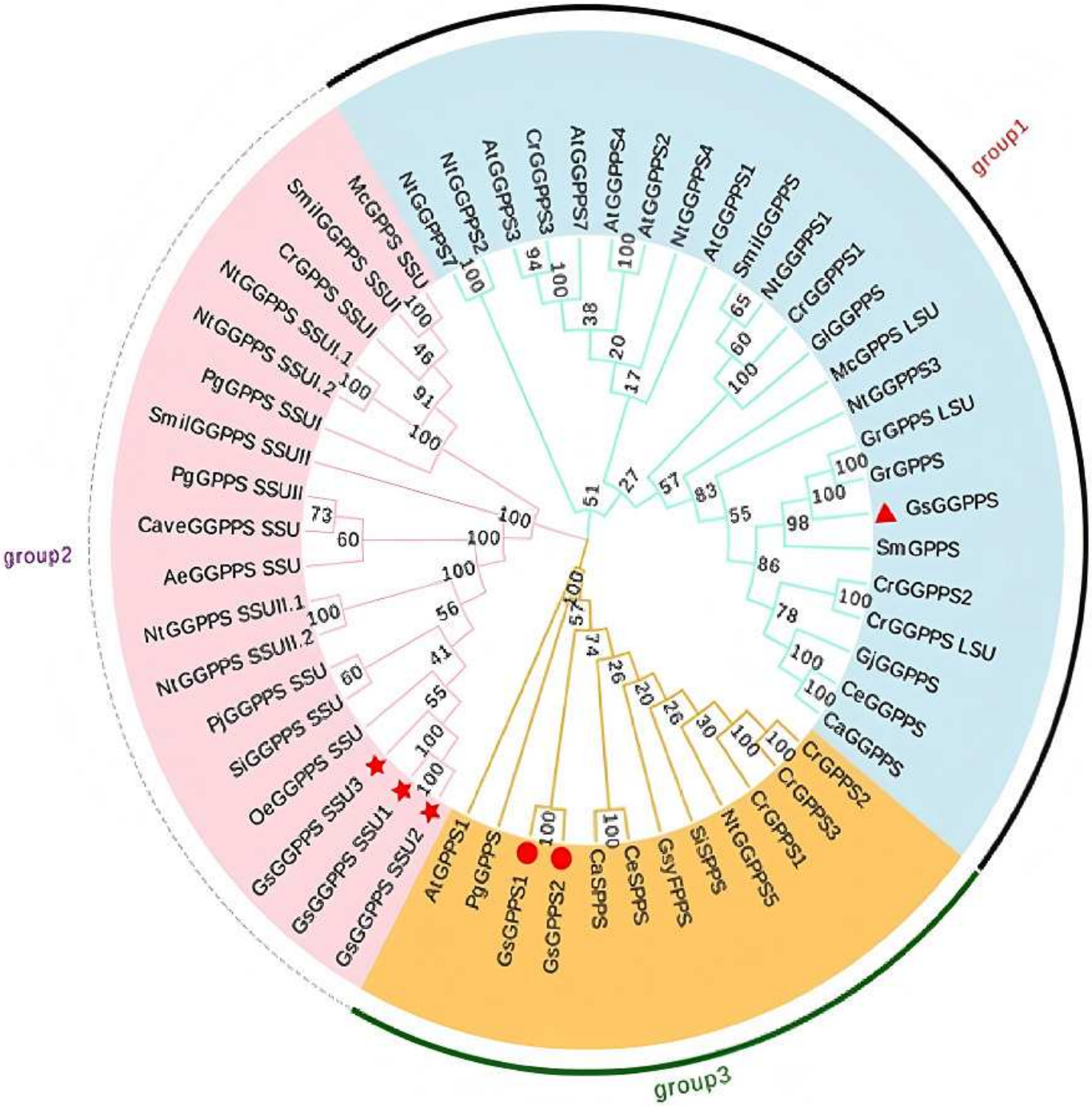

group1
group2
group3
AtGGPPS1
AtGGPPS2
AtGGPPS4
NtGGPPS1
SmIIGGPPS
CrGGPPS1
GlGGPPS
McGPPS LSU
NtGGPPS3
GrGPPS LSU
GrGPPS
GsGGPPS
SmGPPS
CrGGPPS2
CrGGPPS LSU
GjGGPPS
CeGGPPS
CaGGPPS
CrGPPS2
CrGPPS3
CrGPPS1
NtGGPPS5
SlSPPS
GsyFPPS
CaSPPS
CaSPPS
GsGPPS2
GsGPPS1
PgGPPS
AtGPPS1
GsGGPPS SSU2
GsGGPPS SSU1
GsGGPPS SSU3
OeGGPPS SSU
SlGGPPS SSU
PjGGPPS SSU
NtGGPPS SSUII.2
NtGGPPS SSUII.1
AeGGPPS SSU
CaveGGPPS SSU
PgGPPS SSUII
SmIIGGPPS SSUII
PgGPPS SSUI
NtGGPPS SSUI.2
NtGGPPS SSUI.1
CrGPPS SSUI
SmIIGGPPS SSUI
McGPPS SSU
NtGGPPS7
NtGGPPS2
AtGGPPS3
CrGGPPS3
AtGPPS7
AtGGPPS4
AtGGPPS2
NtGGPPS4
AtGGPPS1
SmIIGGPPS
NtGGPPS1
CrGGPPS1
GlGGPPS
McGPPS LSU
NtGGPPS3
GrGPPS LSU
GrGPPS
GsGGPPS
SmGPPS
CrGGPPS2
CrGGPPS LSU
GjGGPPS
CeGGPPS
CaGGPPS
CrGPPS2
CrGPPS3
CrGPPS1
NtGGPPS5
SlSPPS
GsyFPPS
CaSPPS
CaSPPS
GsGPPS2
GsGPPS1
PgGPPS
AtGPPS1
GsGGPPS SSU2
GsGGPPS SSU1
GsGGPPS SSU3
OeGGPPS SSU
SlGGPPS SSU
PjGGPPS SSU
NtGGPPS SSUII.2
NtGGPPS SSUII.1
AeGGPPS SSU
CaveGGPPS SSU
PgGPPS SSUII
SmIIGGPPS SSUII
PgGPPS SSUI
NtGGPPS SSUI.2
NtGGPPS SSUI.1
CrGPPS SSUI
SmIIGGPPS SSUI
McGPPS SSU
NtGGPPS7
NtGGPPS2
AtGGPPS3
CrGGPPS3
AtGPPS7
AtGGPPS4
AtGGPPS2
NtGGPPS4
AtGGPPS1
SmIIGGPPS
NtGGPPS1
CrGGPPS1
GlGGPPS
McGPPS LSU
NtGGPPS3
GrGPPS LSU
GrGPPS
GsGGPPS
SmGPPS
CrGGPPS2
CrGGPPS LSU
GjGGPPS
CeGGPPS
CaGGPPS
CrGPPS2
CrGPPS3
CrGPPS1
NtGGPPS5
SlSPPS
GsyFPPS
CaSPPS
CaSPPS
GsGPPS2
GsGPPS1
PgGPPS
AtGPPS1
GsGGPPS SSU2
GsGGPPS SSU1
GsGGPPS SSU3
OeGGPPS SSU
SlGGPPS SSU
PjGGPPS SSU
NtGGPPS SSUII.2
NtGGPPS SSUII.1
AeGGPPS SSU
CaveGGPPS SSU
PgGPPS SSUII
SmIIGGPPS SSUII
PgGPPS SSUI
NtGGPPS SSUI.2
NtGGPPS SSUI.1
CrGPPS SSUI
SmIIGGPPS SSUI
McGPPS SSU
NtGGPPS7

Figure 9

Fig ure . 9 Tissue-specific expression abundance of key genes involved in iridoids ynthesis .

Note: Bar chart indicated the relative expression levels of genes, line chart indicated the FPKM values of genes. NEC indicated non-embryonic calli , EC indicated embryonic calli . Bars represent standard deviation; Different lowercase letters indicating significant differences at the 0.05 level of probability according to Duncan ’ s multiple-rangetest

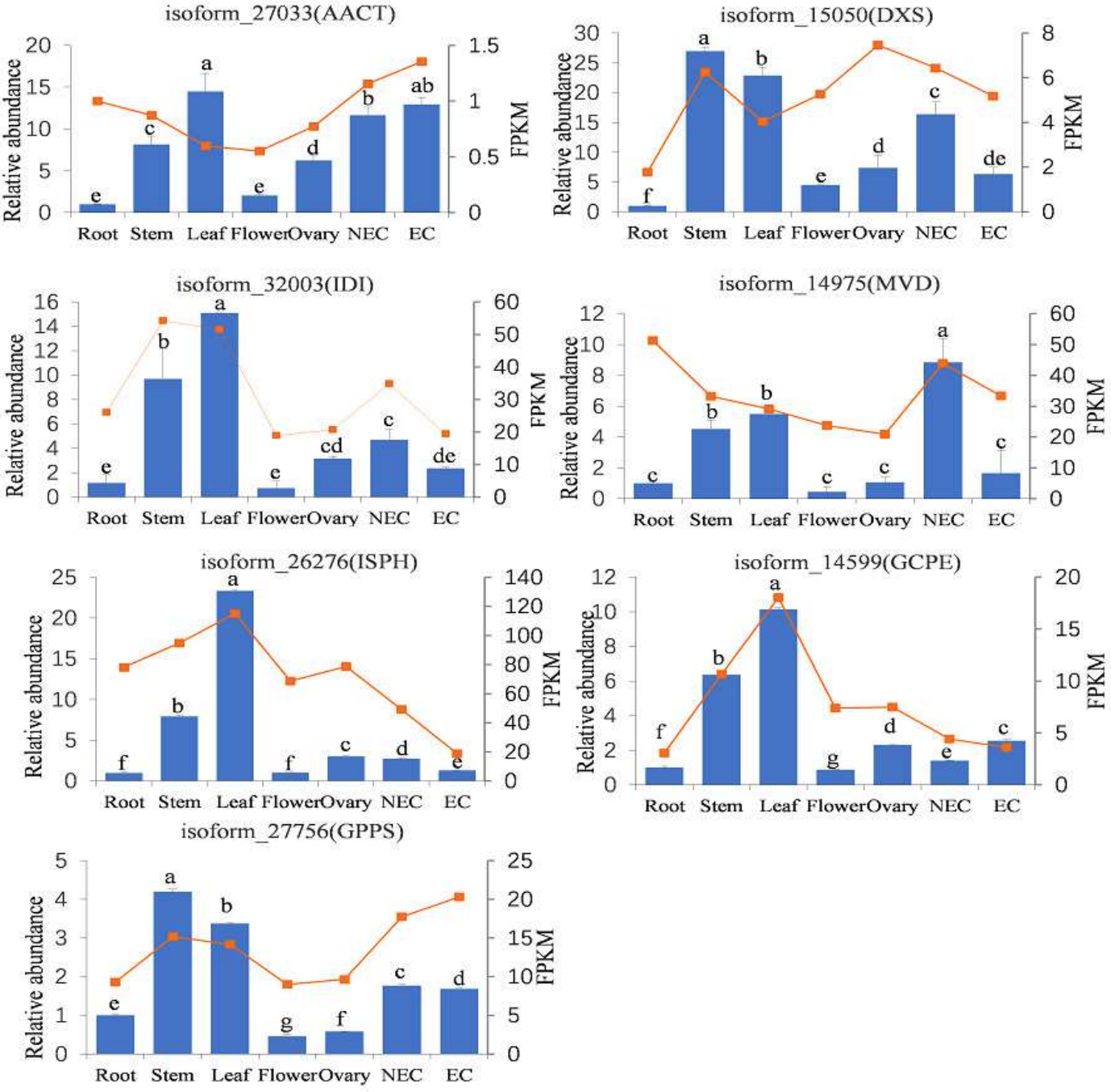

isoform_27033(AACT)
Relative abundance
FPKM
Root Stem Leaf Flower Ovary NEC EC
isoform_15050(DXS)
Relative abundance
FPKM
Root Stem Leaf Flower Ovary NEC EC
isoform_32003(IDI)
Relative abundance
FPKM
Root Stem Leaf Flower Ovary NEC EC
isoform_14975(MVD)
Relative abundance
FPKM
Root Stem Leaf Flower Ovary NEC EC
isoform_26276(ISPH)
Relative abundance
FPKM
Root Stem Leaf Flower Ovary NEC EC
isoform_14599(GCPE)
Relative abundance
FPKM
Root Stem Leaf Flower Ovary NEC EC
isoform_27756(GPPS)
Relative abundance
FPKM
Root Stem Leaf Flower Ovary NEC EC

Table 1(on next page)

Table 1 Comparison with reference gene sequence Pure reads obtained in second generation sequencing

Note:R indicated tissue of root;S indicated stem, L indicated leaf, F indicated flower, O indicated ovary, the numbers after letters indicated three biological replicates

Table 1 Comparison with reference gene sequence Pure reads obtained in second generation sequencing

Sample	CleanData(GB)	Total_Mapped(%)	Unique_Mapped(%)
R-1	6.574	76.63	18.14
R-2	6.944	76.75	18.04
R-3	6.449	76.95	18.08
S-1	6.712	73.75	18.86
S-2	5.728	73.86	18.9
S-3	5.835	73.33	19.04
L-1	6.398	76.13	20.3
L-2	5.922	76.74	19.34
L-3	6.232	76.95	19.39
F-1	6.507	71.36	18.76
F-2	6.837	70.92	18.59
F-3	6.406	71.18	18.65
O-1	7.384	72.48	19.04
O-2	6.594	71.91	18.92
O-3	7.001	72.3	18.89

Note: R indicated tissue of root; S indicated stem, L indicated leaf, F indicated flower, O indicated ovary, the numbers after letters indicated three biological replicates.

Table 2(on next page)

Table 2 Physicochemical , s tructural propert ies and subcellular localization of G s G(G)PPS

MW: molecular weight; pI: isoelectricpoint; SP : Signal peptide cleavage site; SL: Subcellular localization; GRAVY, grand ave rage of hydropathicity ; TS: T ransmembrane structures , o: indicates that the proteinis predicted to be out side the membrane

Table 2 Physicochemical, structural properties and subcellular localization of GsG(G)PPS

Isoform number	Gene name	length (aa)	MW(kD)	pI	SP	SL	GRAVY	TS
Isoform0028091	GsGGPPS SSU1	347	37.90001	5.81	NO	chloroplast	-0.187	o
Isoform0028274	GsGGPPS SSU2	346	37.81293	5.81	NO	chloroplast	-0.185	o
Isoform0027670	GsGGPPS SSU3	342	37.474.66	5.81	NO	chloroplast	-0.180	o
Isoform0030969	GsGGPPS	368	40.01208	6.28	NO	chloroplast	-0.050	o
Isoform0027756	GsGPPS1	424	46.45244	6.48	NO	mitochondrion	0.049	o
Isoform0030732	GsGPPS2	424	46.39234	6.48	NO	mitochondrion	0.041	o

MW: molecular weight; pI: isoelectric point; SP: Signal peptide cleavage site; SL: Subcellular localization; GRAVY, grand average of hydropathicity; TS: Transmembrane structures, o: indicates that the protein is predicted to be outside the membrane.