# SMRT- and Illumina-based RNA-seq Analyses Unveil Triterpenoid Saponin Biosynthesis and Related Key Genes in *Platycodon Grandiflorus* (Jacq.) A. DC.

**Hanwen Yu**
Anhui University of Traditional Chinese Medicine
**Mengli Liu**
Anhui University of Traditional Chinese Medicine
**Minzhen Yin**
Anhui University of Traditional Chinese Medicine
**Tingyu Shan**
Anhui University of Traditional Chinese Medicine
**Huasheng Peng**
Anhui University of Traditional Chinese Medicine
**Jutao Wang**
Anhui University of Traditional Chinese Medicine
**Xiangwei Chang**
Anhui University of Traditional Chinese Medicine
**Daiyin Peng**
Anhui University of Traditional Chinese Medicine
**Liangping Zha**
Anhui University of Traditional Chinese Medicine
**Shuangying Gui** ( guishy0520@126.com )
Anhui University of Traditional Chinese Medicine

**Research Article**

**Keywords:** Platycodon grandiflorus, Full-length transcriptome, SMRT sequencing, RNA-Seq, Triterpenoid saponin biosynthesis

**Posted Date:** January 6th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-139609/v1

# Abstract

**Background:** *Platycodon grandiflorus*, a traditional Chinese medicine, contains considerable triterpene saponins with broad pharmacological activities. To date, information on the molecular mechanism of triterpenoid saponin biosynthesis in *P. grandiflorus* is limited. Here, single-molecule real-time (SMRT) and next-generation sequencing technologies were combined to comprehensively analyse the transcriptome and unveil triterpenoid saponin biosynthesis in *P. grandiflorus.*

**Results:** We quantified four saponin monomers in *P. grandiflorus*, and found that the total content of the four saponins was the highest in the roots and the lowest in the stems and leaves. A total of 173,354 non-redundant transcripts generated from the PacBio platform were successfully annotated to seven functional databases, among which 1,765 transcripts were aligned to the "metabolism of terpenoids and polyketides" pathway in the KEGG database. Three full-length transcripts of β-amyrin synthase (β-AS), the key synthase of the β-amyrin, were identified. Furthermore, a total of 132,610 clean reads of BGISEQ sequences were utilised to explore key genes related to the triterpenoid saponin biosynthetic pathway in *P. grandiflorus*, and 96 differentially expressed genes (DEGs) involved were selected as candidates. Notably, 9 of the 96 DEGs showed the highest expression in the roots, which were considered key genes for synthesising triterpenoid saponins in *P. grandiflorus.* Furthermore, 3,469 genes encoding transcription factors (TFs) were identified and classified into 57 TF families, including MYB, bHLH, mTERF, and AP2-EREBP. The expression levels of genes were verified by quantitative real-time PCR.

**Conclusions:** Our reliable transcriptome data provide valuable information on the related biosynthesis pathway and may provide new insights into the molecular mechanisms of triterpenoid saponin biosynthesis in *P. grandiflorus*.


# Background

*Platycodon grandiflorus* is among the most widely used traditional Chinese medicines possessing various biological activities. The primary chemical constituents of *P. grandiflorus* are triterpenoid saponins, predominantly platycodin D and E and polygalacin D. Platycodin D is a marker compound for quality evaluation and standardisation [1]. Numerous studies have reported the pharmacological efficacy of *P. grandiflorus*, such as anti-inflammatory [2, 3], antitumor [4, 5], and antiobesity [6–8] effects, phlegm and cough elimination [9], and protection against hepatotoxicity [10–12].

The synthesis of triterpenoid saponins has been extensively investigated and is divided into three main phases: mevalonate (MVA) and 2-methyl-Derythritol 4-phosphate (MEP) route constitute the inception phase, the terpenoid backbone biosynthesis phase, and the modification phase [13–15]. The MVA pathway begins with the conversion of acetyl-CoA to acetoacetyl-CoA catalysed by acetoacetyl-CoA thiolase (AACT), and then mevalonate is synthesised after successive catalysis by hydroxymethylglutaryl-CoA synthase (HMGS) and hydroxymethylglutaryl-CoA reductase (HMGR). Isopentenyl diphosphate (IPP) is the final product of the MVA pathway generated by enzymatic transformations of mevalonate kinase (MK), phosphomevalonate kinase (MVK), and diphosphomevalonate decarboxylase (MVD). Additionally, the MEP pathway begins with D-glyceraldehyde-3-phosphate (G3P) and ultimately produces dimethylallyl-PP (DMAPP). DMAPP is synthesised sequentially via a series of enzymes, including 1-deoxy-D-xylulose-5-phosphate synthase (DXS), 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR), 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (MCT), 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (CMK), 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS), 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (HDS), and 46-hydroxy-3-methyl-but-2-en-1-yl diphosphate reductase (HDR). IPP and DMAPP are the precursors of terpenoids, in addition to interconverting into each other, under the catalysis of isopentenyl-diphosphate delta-isomerase (IDI) [16–18]. IPP and DMAPP transform into squalene, the direct triterpenoid precursor, through geranyl diphosphate synthase (GPPS), farnesyl diphosphate synthase (FPPS), and squalene synthetase (SQS) catalysis. Subsequently, 2,3-

oxidosqualene is synthesised from squalene via squalene epoxidase (SQE) and further cyclised to β-amyrin under β-amyrin synthase (β-AS), followed by the formation of oleanolic acid under β-amyrin 28-oxidase. In conclusion, oleanolic acid is catalysed by cytochrome P450 (CYP450) and glycosyltransferases (GTs) to form various triterpenoid saponins [19].

Increasing investigations have focused on the transcriptome and genome levels with the understanding of plant secondary metabolic pathways. Concomitantly, next-generation sequencing (NGS) technology with high efficacy and low cost is widely used to investigate putative genes and construct genome and transcriptome resources [20–21]. RNA-seq has been applied diffusely to identify gene function related to secondary metabolites and determine differential gene expression. Based on short-read RNA-seq, studies on the *P. grandiflorus* transcriptome have been conducted [22]. Nevertheless, because of the short sequence reads of NGS, incompletely assembled transcripts may inevitably occur, as well as base mis-assembly and incorrect annotation. Fortunately, with the process of sequencing technology, single-molecule real-time (SMRT) technology using the PacBio Iso-Seq protocol can compensate for the lack of NGS and has become a popular third-generation sequencing platform [23]. However, SMRT technology still has a high error rate, rectified through short-read data [24, 25]. Based on this, it provides an efficient combined approach to obtain full-length transcript sequences of diversified plants [26–28].

Here, the NGS and SMRT sequencing technologies were combined to characterise the comprehensive transcriptome of the roots, stems, and leaves (three duplicates) of *P. grandifloras.* Moreover, we also determined the contents of four saponin monomers to confirm the main accumulation organs of triterpenoid saponins. Comprehensive analysis was conducted to obtain key genes related to triterpenoid saponin synthesis in *P. grandiflorus*. The reliability of the transcriptome data was verified using real-time quantitative PCR (qRT-PCR). The experiment may provide valuable information on the biosynthesis of triterpenoid saponins in *P. grandiflorus*.

# Results

### Determination of contents of saponin monomers

The raw dried extracts obtained from the roots, stems, and leaves of *P. grandiflorus* were measured using high performance liquid chromatography with evaporative light scattering detection (HPLC-ELSD). The contents of four saponin monomers in three organs with three duplicates were expressed as the mean with standard deviation (Fig. 1). The contents of platycodin D, platycodin D3, and polygalacin D were higher in the root than in aerial parts. In addition, the sum of mean saponin contents in the root was also higher than that in the stem and leaf.

### P. grandiflorus transcriptome analysis via RNA-Seq and PacBio Iso-Seq

To obtain a comprehensive transcriptome profile of *P. grandiflorus*, BGISEQ-500 and PacBio Sequel platforms were combined. Nine RNA samples from different organs (stem, leaf, and root, each in triplicate) were sequenced separately using a BGISEQ-500 HiSeq platform, with total raw reads in three tissues produced (145.49 M, 143.74 M, and 140.15 M from stems, leaves, and roots, respectively). After data filtering, a total of 132,610 clean reads were obtained, and the number of total clean reads from stems, leaves, and roots was 131.57 M, 130.47 M, 127.65 M, respectively (Fig. 2a, Table S1). To obtain comprehensive coverage of the *P. grandiflorus* transcriptome, a full-length transcriptome was generated using full-length cDNAs from pooled poly (A) RNA of three organs normalised and subjected to SMRT sequencing using the PacBio Sequel platform. Altogether, 717,560 polymerase reads were generated, and 23,173,163 subreads (32.64 Gb) were obtained from one SMRT cell after filtering, with a mean length of 1,408.61 bp and N50 length of 1,992 bp. A total of 685,102 reads of insert (ROIs) were obtained after data processing, with a mean length of 2,219 bp, mean read quality of 0.95, and 28 passes. Of these, 547,395 ROIs containing two primers and poly (A) tails

were classified as full-length non-chimeric (FLNC) reads with a mean length of 1,341 bp (Fig. S1a). Furthermore, FLNC reads were clustered and corrected using the Interactive Clustering and Error Correction (ICE) algorithm and Quiver program. As a result, 180,602 full-length consensus isoforms were obtained, with a mean length of 1,449 bp and quality of 0.98 (Fig. S1b), of which, the high-quality isoforms were further combined, and redundant sequences were removed using CD-Hit. A total of 173,354 non-redundant transcripts were identified (Fig. 2b), with lengths ranging from 199 bp to 15,360 bp, N50 of 1,924 bp, N90 of 818 bp, and a GC content of 42.49% (Table 1).

Table 1
Statistics of non-redundant transcripts

| Total number | Total length (bp) | N50 (bp) | N90 (bp) | Max length (bp) | Min length (bp) | Sequence GC (%) |
|---|---|---|---|---|---|---|
| 173,354 | 252,336,203 | 1924 | 818 | 15,360 | 199 | 42.49% |

**Functional annotation of full-length** P. grandiflorus **transcriptome**

For a comprehensive annotation of the *P. grandiflorus* transcriptome, 173,354 non-redundant transcripts were functionally annotated against the Nr, Nt, KEGG, SwissProt, Pfam, GO, and KOG databases using Blast and Blast2GO. Among them, 146,868 transcripts (84.72%) were successfully matched to at least one of the seven databases with an achievement ratio ranging from 22.62 to 84.72% (Fig. 2c, Table 2). A total of 139,018 transcripts of *P. grandiflorus* were analysed to be homologues by aligning sequences to the Nr database, of which the most significant homology species was *Cynara cardunculus* var. scolymus (11.64%), followed by *Helianthus annuus* (11.38%), *Vitis vinifera* (9.87%), and *Daucus carota* subsp. *sativus* (5.48%) Furthermore, the remaining 61.63% sequences were mapped to other plants (Fig. 2d).

To explore the main biological processes in *P. grandiflorus*, 111,280 transcripts were mapped to the KEGG database and classified into five categories, including cellular processes, environmental information processing, genetic information processing, metabolism, and organismal systems (Fig. 3a, Table S2), where "transport and catabolism," "signal transduction," "translation," "global and overview maps," and "environmental adaptation" were the most abundant subcategories, respectively. Notably, 1,765 transcripts of *P. grandiflorus* were enriched in "metabolism of terpenoids and polyketides"; specifically, 493 transcripts were associated with "terpenoid backbone biosynthesis" (ko00900), followed by 268 transcripts and 106 transcripts involved in "sesquiterpenoid and triterpenoid biosynthesis" (ko00909) and "diterpenoid biosynthesis" (ko00904) pathways, respectively.

In addition, the KOG annotation demonstrated that 109,686 transcripts were assigned to 25 functional clusters, and "general function prediction only" (24,091 transcripts) was the largest category, followed by "signal transduction mechanisms" (16,154 transcripts) and "posttranslational modification, protein turnover, chaperones" (13,603 transcripts) (Fig. 3b, Table S3).

Based on the Nr annotation results, 77,618 transcripts were assigned to the GO database and classified into three functional categories, biological process, molecular function, and cellular component, using the Blast2GO program (Fig. 3c, Table S4). The cellular component was the largest cluster comprising 164,035 transcripts, whose major terms were "cell" and "cell part," followed by biological process containing 111,975 transcripts and molecular function with 91,034 transcripts. Under the biological process and molecular function categories, "metabolic process" and "catalytic activity" were the most abundant terms, respectively, indicating that the study might provide insight into the genes involved in secondary metabolite synthetic pathways.

Table 2
Functional annotation results from seven public databases

| Databases | Nr | Nt | Swiss-prot | KEGG | KOG | Pfam | GO | Intersection | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Number | 139,018 | 119,342 | 108,337 | 111,280 | 109,686 | 91,636 | 77,618 | 39,248 | 146,868 |
| Percentage (%) | 80.19 | 68.84 | 62.49 | 64.19 | 63.27 | 52.86 | 44.77 | 22.64 | 84.72 |

### Identification of differentially expressed genes

To explore the variation in gene abundance and expression profiles among the stem, leaf, and root (leaf vs stem, stem vs root, and leaf vs root), clean reads from RNA-Seq were aligned to reference transcripts, and the expression level calculated using RSEM. There were 50,957 differentially expressed genes (DEGs) derived from the leaf vs stem analysis, including 34,124 that were upregulated and 16,833 downregulated. Root and stem tissue comparison resulted in 37,213 DEGs, where 14,091 were upregulated, and 23,122 downregulated. Additionally, 57,203 DEGs were identified in the leaf vs root comparison, of which 35,109 were upregulated, and 22,094 were downregulated (Fig. 4a). Moreover, among these DEGs, 14,029 common DEGs were identified in three tissues that might be vital for metabolism among the three organs. Comparing DEGs in diverse groups resulted in 8,233 specific DEGs in leaf vs root, followed by 5,700 in leaf vs stem and 3,009 in stem vs root, indicating that the difference between leaf and root was the greatest (Fig. 4b). For a thorough exploration of differential transcripts, KEGG and GO enrichment analyses were performed using the Phyper function. In KEGG analysis, 15,894 and 23,334 DEGs were identified in stem vs root and leaf vs root comparisons and mapped to 137 and 138 pathways, respectively, where the most enriched pathway for either comparison was "phenylpropanoid biosynthesis" (Fig. 4c, d).

There were 60,906 DEGs in the leaf vs stem comparison, mainly associated with "plant-pathogen interaction" (Fig. S2a). Moreover, 137 and 121 DEGs were identified and separately matched to "terpenoid backbone biosynthesis" and "sesquiterpenoid and triterpenoid biosynthesis", respectively, in stems compared with roots. There were 219 and 154 DEGs annotated to the above pathways in the leaf vs root comparison. The most expansive GO term among the different comparison groups was "cellular component" (Fig. S2b, c, d).

### Identification of transcription factors

Transcription factors (TFs) play vital roles in the regulation of secondary metabolic processes. A total of 3,469 genes encoding TFs were identified through alignment with the Plant TF database and classified into 57 TF families, including 743 upregulated and 1,675 downregulated genes in the leaf vs root comparison and 878 upregulated and 1,168 downregulated genes in the stem vs root comparison (Table 3 and Table S5). A total of 313 genes encoded the mTERF family, accounting for the largest proportion, followed by 281, 262, and 259 genes encoding the MYB, bHLH, and AP2-EREBP families. We further annotated TFs with the KEGG functional database and obtained eight FHA TFs involved in terpenoid and polyketide metabolism; 14 zf-HD TFs, 5 MYB TFs, 2 RWP-RK TFs, and 1 C2C2-CO-like TF participated in the biosynthesis of other secondary metabolites. Among them, 4 zf-HD TFs, 2 RWP-RK TFs, and 1 MYB TF showed the highest expression in the roots, which are thought to be closely related to triterpene synthesis (Fig. S3, Table S6).

Table 3
Classification and number of TF families identified in the DEG database

| TF Family | Number of genes | Number of upregulated genes | | Number of downregulated genes | |
|---|---|---|---|---|---|
| | | Leaf vs Root | Stem vs Root | Leaf vs Root | Stem vs Root |
| mTERF | 313 | 94 | 27 | 67 | 54 |
| MYB | 281 | 57 | 129 | 87 | 98 |
| bHLH | 262 | 65 | 134 | 74 | 100 |
| AP2-EREBP | 259 | 16 | 174 | 71 | 120 |
| GRAS | 205 | 24 | 112 | 65 | 70 |
| C3H | 165 | 16 | 109 | 16 | 81 |
| WRKY | 131 | 22 | 68 | 56 | 29 |
| NAC | 111 | 9 | 64 | 1 | 32 |
| Trihelix | 103 | 21 | 58 | 15 | 51 |
| C2H2 | 98 | 30 | 38 | 23 | 41 |
| G2-like | 91 | 14 | 52 | 19 | 27 |
| SBP | 84 | 21 | 27 | 23 | 16 |
| C2C2-GATA | 79 | 19 | 36 | 18 | 34 |
| TCP | 78 | 52 | 13 | 24 | 23 |
| ARF | 69 | 2 | 53 | 12 | 37 |
| FAR1 | 68 | 8 | 27 | 5 | 28 |
| Tify | 65 | 15 | 36 | 28 | 19 |
| C2C2-Dof | 61 | 3 | 36 | 7 | 30 |
| HSF | 60 | 11 | 33 | 14 | 21 |
| Other | 886 | 244 | 449 | 253 | 257 |
| Total number | 3469 | 743 | 1675 | 878 | 1168 |

**Putative genes involved in triterpenoid saponin biosynthesis**

Triterpenoid saponins are the primary curative components of *P. grandiflorus.* As previously described, many known enzymes participate in the synthesis of triterpenoid saponins, such as AACT, HMGS, DXS, and SQS. Although the biosynthetic pathway of platycodins has been widely studied, exhaustive information on the genes encoding relevant key enzymes is still limited. Platycodins are oleanane-type triterpenoid saponins synthesised from oleanolic acid through several modifications by CYP450 (cytochrome P450 monooxygenase) and GTs. Subsequent modification pathways catalytically produce various platycodins (Fig. 5a).

The synthesis of platycodins predominantly comprised "terpenoid backbone biosynthesis" (map00900) and "sesquiterpenoid and triterpenoid biosynthesis" (map00909) pathways based on KEGG enrichment analysis (Fig. S4).

Divergent gene expression has a vital influence on platycodin biosynthesis. After screening of FPKM > 1, a total of 113 genes encoding 18 key enzymes, 96 of which were further identified as DEGs, were considered as putative genes related to triterpenoid saponin synthesis in *P. grandiflorus* (Table 4 and Table S7).

DEGs distributed in MVA and MEP upstream pathways were greater than those in downstream pathways. In our transcriptome dataset, 22 DEGs encoding six enzymes relevant to the MVA pathway were identified, including nine *PgAACT* genes, eight *PgHMGR* genes, one *PgHMGR* gene, three *PgMK* genes, and one *PgMDC* gene. Moreover, 34 DEGs encoding seven enzymes involved in the MEP pathway were confirmed, including nine *PgDXS* genes, one *PgDXR* gene, two *PgMCT* genes, two *PgCMK* genes, seven *PgMDS* genes, six *PgHDS* genes, and seven *PgHDR* genes. The MEP pathway demonstrated that nearly all DEGs showed the highest expression levels in leaves and the least in roots. In comparison with the MEP pathway, highly expressed DEGs were concentrated in the root and stem in the MVA pathway. Research has suggested that the biosynthesis of triterpenoids and sesquiterpenoids occurs via the MVA pathway, whereas monoterpenoids and diterpenoids are involved in the MEP pathway. Additionally, based on our transcriptome data, only one transcript for genes encoding *PgHMGS* and *PgMDC* identified as DEGs was obtained, and the expression level was the highest in the stem, followed by the root or leaf. For DEGs of *PgHMGR*, the expression levels of most transcripts in the stem and leaf were higher than those in the root, yet *PgHMGR1* and *PgHMGR3* showed their highest expression in the roots. As genes encoding key enzymes of the MVA pathway, *PgAACT* and *PgMK* displayed higher expression in stems or roots than in leaves. Among them, *PgAACT2*, *PgAACT3*, and *PgMK1* all had the highest expression levels in the roots, especially for *PgAACT3*, which exhibited much higher expression than other *PgAACTs*, presumably vital for platycodin biosynthesis.

Most of the 13 *PgFPPS* DEGs were expressed at higher levels in leaves, except for *PgFPPS9* and *PgFPPS13* whose expression was the highest in the root and stem, respectively. IDI is a crucial enzyme that catalyses the conversion between the common precursors IPP and DMAPP. There were 11 *PgIDIs* found to be DEGs among the three tissues, and most were highly expressed in the stem and root; notably, *PgIDI6* and *PgIDI7* displayed the highest expression in the roots. Pertaining to “sesquiterpenoid and triterpenoid biosynthesis”, 14 DEGs were found encoding *PgSQS*, *PgSQE*, and *Pgβ-AS*. Of the eight candidate genes of *PgSQS*, only *PgSQS1* was identified as the DEG with the highest expression in the stems.

Additionally, differential transcripts of *PgSQE* and *Pgβ-AS* had similar expression trends, with the highest expression in the leaves. β-AS is a significant modification enzyme for 2,3-oxidized squalene for oleanolic acid formation, and *Pgβ-AS1* was expressed at the highest level in the roots. Statistically, there were nine genes with the highest expression in the root, which were considered key genes for triterpenoid saponin biosynthesis in *P. grandiflorus* (Fig. 5b).

Table 4
Number of genes encoding key enzymes (FPKM > 1)

| Enzyme Abbreviation | EC number | Gene number |
|---|---|---|
| AACT | 2.3.1.9 | 14 |
| HMGS | 2.3.3.10 | 5 |
| HMGR | 1.1.1.34 | 5 |
| MK | 2.7.1.36 | 3 |
| PMK | 2.7.4.2 | 1 |
| MDC | 4.1.1.33 | 4 |
| DXS | 2.2.1.7 | 5 |
| DXR | 1.1.1.267 | 1 |
| MCT | 2.7.7.60 | 2 |
| CMK | 2.7.1.148 | 2 |
| MCS | 4.6.1.12 | 8 |
| HDS | 1.17.7.1, 1.17.7.3 | 7 |
| HDR | 1.17.7.2, 1.17.7.4, 1.17.1.4 | 6 |
| IDI | 5.3.3.2 | 13 |
| FPPS | 2.5.1.1, 2.5.1.10 | 20 |
| SQS | 2.5.1.21 | 8 |
| SQE | 1.14.14.17 | 8 |
| β-AS | 5.4.99.39 | 1 |

## Characterisation of the β-amyrin synthase

β-AS, belonging to the oxidosqualene cyclase (OSC) family, plays a vital role in regulating oleanane-type triterpenoid saponin biosynthesis. β-AS catalyses the conversion of 2,3-oxidosqualene into β-amyrin, the precursor of oleic acid. To date, five types of OSCs have been isolated and identified from ginseng, including β-AS, dammarenediol synthase (DS), cycloartenol synthase (CAS), lupeol synthase (LUS), and lanosterol synthase (LS). From the screening, three full-length genes encoding putative β-AS proteins (*isoform_10598, isoform_71380*, and *isoform 102853*) in *P. grandiflorus* were identified and characterised using the transcript library.

Domain analysis revealed that three β-AS isoforms contained one catalytic acid site and a conserved squalene cyclase domain belonging to the ISOPREN-C2-like superfamily. The ISOPREN-C2-like superfamily contains class II terpene cyclases, including squalene cyclase and 2,3-oxidosqualene cyclase (OSC) (Fig. S5). Ultimately, we selected two complete sequences encoding β-AS from ginseng (*OSCPNY1* and *OSCPNY2*) to perform homology analysis with three isoforms. The comparison of five amino acid sequences indicated that their consistency was 74.04%, and three conserved regions (motif QW, DCTAE, and MWCYCR) belonging to β-AS were also identified (Fig. 6d) [29–30]. The complete alignment result is shown in Fig. S6. The *isoform_10598* has a mutation site, "L," in the DCTAE motif, and

*isoform_71380* has two mutation sites, "L" and "F," in the MWCYCR motif. Three isoforms were further selected to create 3D constructure models based on human OSC complexed with lanosterol (PDB ID: 1w6k.1.A) (Fig. 6a, b, c) [31].

The phylogenetic tree was divided into four branches, LUS, LS, CAS, and β-AS, where the typical genes were *MtLUS*, *VrLS*, *CrCAS and GaβAS* from *Medicago truncatula*, *Vigna radiata* var. *radiata*, *C. reinhardtii*, and *G. arboreum*, respectively (Fig. 7, Table S8). The results revealed that three isoforms were homologous to β-AS from other plants, consistent with our previous structure analysis and indicating that the three isoforms were β-AS genes from *P. grandiflorus*. Additionally, *isoform_71380* clustered together with β-AS from *G. arboreum*, and *isoform_102853* clustered together with β-AS from *Panax ginseng*. These two sub-branches and *isoform_10598* were further clustered.

### qRT-PCR validation of DEGs from the RNA-seq analysis

To validate the reliability of transcriptome analysis data, 20 DEGs related to triterpenoid saponin biosynthesis were verified using qRT-PCR with β-actin as the reference; all primers used are listed in Table S9. The results of RNA-seq and qRT-PCR revealed a high-ranked consistency, indicating that the RNA-seq data are dependable and accurate (Fig. 8). More importantly, *PgAACT2*, *PgHMGR1*, and *Pgβ-AS1* displayed higher expression in roots than the tested genes.

## Discussion

As an important medicinal plant, *P. grandiflorus* is extensively used worldwide for its therapeutic effects. Many studies have focused on the biological effects of *P. grandiflorus*, especially of the inherent triterpenoid saponins. The contents of four saponin monomers, namely, platycoside E, platycodin D, platycodin $D_3$, and polygalacin D, have been determined previously, with the contents of platycodin D, platycodin $D_3$, and polygalacin D being the highest in the roots. Notably, platycoside E, platycodin D, and platycodin $D_3$ have the same nuclear structure as platycodigenin, and polygalacin D is enzymatically modified from polygalacic acid.

Platycoside E and platycodin $D_3$ are precursors of platycodin D with the conversion occurring under β-D-glucosidase hydrolysis, and the three components above belong to one synthetic pathway. Studies have recently focused on the genome level [32]; Previously, research on genes involved in platycodin biosynthesis was carried out based on RNA-seq.

Here, SMRT sequencing and RNA-seq of three tissues (root, stem, and leaf) were combined to generate a more comprehensive transcriptome of *P. XXXrandifloras.* A total of 685,102 CCS reads were obtained from PacBio ISO-seq, yielding 173,354 non-redundant transcripts (N50 = 2,517 bp) after correction using RNA-sEq. The BUSCO database was used to evaluate transcript quality, compared with conserved genes, indicating the integrity of transcriptome assembly. The combination of RNA-Seq and ISO-seq could provide a thorough understanding of triterpenoid saponin synthesis in *P. grandiflorus* at the molecular level.

In this study, clean reads obtained from RNA-seq were further identified based on a Poisson distribution to generate DEGs. According to the KEGG annotation results, the differential genes were classified into various biological pathways. A total of 477 genes were mapped to the "metabolism of terpenoids and polyketides" pathway. We further screened the DEGs related to triterpenoid saponin biosynthesis and obtained 96 DEGs encoding 18 related key enzymes. Among all the DEGs, 34 were found to be related to the MEP pathway and 13 to the MVA pathway. By further analysing these genes, we obtained nine DEGs encoding AACT, HMGR, MK, IDI, FPPS, and β-AS expressed the highest in the roots. Four of these nine genes in roots were verified using qRT-PCR, where *PgAACT2*, *PgHMGR1*, and *Pgβ-AS1* also showed the highest expression in the roots. Previous chemical experiments have confirmed that triterpenoid saponins of *P. grandiflorus* primarily accumulate in the root. Therefore, we speculate that the nine genes expressed the highest in the roots are involved in triterpenoid saponin biosynthesis in *P. grandiflorus*. The genes identified above are candidate

genes for triterpenoid saponin biosynthesis in *P. grandiflorus*, and the mechanism of their action in triterpenoid saponin biosynthesis regulation warrants further investigation.

Previous research also analysed the expression of one AACT and β-AS genes from *P. grandiflorus* in different organs, including the roots, young stems, leaves, and flowers, indicating that the unigene encoding β-AS was expressed at a much higher level in the leaves than in the roots, young stems, and flowers. In contrast, the unigene encoding β-AS showed the highest expression in the roots, consistent with our study [22]. Furthermore, tissue-specific expression profiles in different *P. grandiflorus* tissues (leaf, root, stem, seed, petal, pistil, sepal, and stamen) were also revealed, and of the 24 *bAS* genes, four genes showed significantly higher expression in the roots than in other tissues [33]. β-amyrin is the oleanane-type backbone and the precursor of oleanolic acid, with synthesis controlled by β-AS [34, 35]. One *EsBAS* has been isolated and cloned from the leaves of *E. senticosus*, and *EsBAS* was functionally characterised via heterologous expression in yeast and tobacco [36]. In addition, one *PlgOSC1* gene encoding β-AS from *P. grandiflorus* was cloned via RACE-PCR and then successfully expressed in heterologous yeast cells [37]. Methylome data of β-AS in *P. grandiflorus* have also been presented in control and methyl jasmonate (MJ) treatments. The relative dominance of hypo-CG-DMCs in the MJ treatment was detected only at 48 h with *bAS* genes; the hypomethylation of *bAS* genes may affect platycoside biosynthesis [33]. We screened out three full-length sequences from the differential genes encoding *Pgβ-AS*. The length of the three full-length sequences was approximately 2,200 bp. Three transcripts were confirmed to encode β-AS after bioinformatics and phylogenetic analyses.

TFs, as sequence-specific DNA-binding proteins, have a drastic influence on plant metabolism and regulation [38]. In our study, a total of 3,469 genes were classified into 57 TF families, with the largest proportion of the mTERF family (313 genes). mTERFs are key regulators of organellar gene expression in mitochondria as well as chloroplasts and implicated in all organellar gene expression steps ranging from transcription modulation to tRNA maturation, hence translation [39]. Previous studies have demonstrated that *VvMYB5b* TF overexpression in tomato can upregulate terpenoid metabolism [40]. In our study, 281 MYB TFs were identified, five of which participate in the biosynthesis of other secondary metabolites, after classification using the KEGG database. More notably, one of the five MYB TFs had the highest expression in roots, which is speculated to play an essential role in triterpenoid saponin synthesis in *P. grandiflorus*. Likewise, functional studies of some other TF families have also been conducted. Research has shown that overexpression of two bHLH TFs in *Medicago truncatula* hairy roots led to increased transcription levels of known triterpenoid saponin biosynthesis genes and dramatically increased the accumulation of triterpene saponins [41]. The overexpression of AP2/ERF gene clusters in tobacco and hairy roots activated nicotine and terpenoid indole alkaloid pathway genes [42]. In our study, 262 candidate TFs were identified as belonging to the bHLH family. Through further differential expression analysis, 74 and 100 bHLH TFs were downregulated in the leaf vs root and stem vs root comparisons, respectively. The 259 AP2-EREBP family TFs were also determined, and 71 AP2-EREBP TFs were upregulated in roots compared with stems, and 120 TFs were upregulated in roots compared with leaves. These TFs upregulated in the root may play an essential role in triterpenoid saponin biosynthesis.

# Conclusion

Transcriptome analysis of *P. grandiflorus* was performed using PacBio Iso-Seq combined with RNA-SEq. A total of 173,354 full-length transcripts and 3,469 TFs were obtained, and 173,354 full-length transcripts were successfully annotated to seven databases to obtain an improved annotation. Three full-length genes were confirmed to encode β-AS, providing a basis for subsequent functional research. Among the 3,469 TFs, 281, 262, and 259 TFs of the MYB, bHLH, and AP2-EREBP families, respectively, were analysed. Furthermore, we screened DEGs involved in the triterpenoid saponin biosynthetic pathway based on RNA-seq, and nine DEGs with the highest expression in the root were identified, encoding key enzymes, including *PgAACT*, *PgHMGR*, *PgMK*, *PgIDI*, *PgFPPS*, and *Pgβ-AS*. This article reports on the

integrated transcriptome sequencing analysis of *P. grandiflorus*; moreover, the sequencing results were demonstrated to be reliable and accurate using qRT-PCR. The obtained transcript information may provide a varied and well-grounded candidate pool to study functional genes involved in secondary metabolite biosynthesis.

# Methods

### Plant materials and total RNA extraction

*P. grandiflorus* (Jacq.) A. DC. was collected from Tongcheng city (Anhui province, China) and separated into leaves, stems, roots with three replicates per organ. The samples were snap-frozen in liquid nitrogen after a quick rinse with sterile water and stored at -80 °C for subsequent sequencing and saponin analysis. According to the manufacturer's instructions, total RNA isolation was performed using the ethanol precipitation protocol and CTAB-PBIOZOL. The concentrations of extracted RNA, $OD_{260/280}$ and $OD_{260/230}$ ratios, 28S/18S ratio, and RNA integrity number (RIN) were determined using a Nanodrop micro-spectrophotometer and Agilent 2100 bioanalyser based on the Agilent RNA 6000 nano Reagents Part 1 kit (Agilent Technologies, Santa Clara, CA, USA).

### Analysis of contents of saponin monomers

Dried stem, leaf, and root samples (with three replicates per organ) were ground into a powder and sifted using an 80 holes per inch sieve. Nine weighed powder samples (1.0 g) were extracted with 5 mL of 70% methanol using an ultrasonic method for 1.5 h (80 W, 40 Hz). After ultrasonic treatment, the samples were filtered through a 0.45-μm syringe filter for subsequent HPLC-ELSD analysis [43]. Finally, four saponin monomers (platycoside E, platycodin D, platycodin $D_3$, and polygalacin D) were chromatographically separated on an Agilent Eclipse XDS-C18 (4.6 mm × 250 mm, 5 μm) column (Agilent Technologies) at 35 °C; the mobile phase was acetonitrile-water at a flow rate of 1.0 mL/min. ELSD was used as the detector with a gas flow rate of 1.6 L/min, a drift tube temperature of 85 °C, and a nebulisation temperature of 50 °C.

### BGISEQ-500 RNA-Seq library construction and sequencing

For sequencing, mRNA with Poly (A) of three organs (three replicates) was enriched from total RNA using oligo (dT) magnetic beads [44–46]. The DNA probe was digested using DNase I after hybridisation with rRNA to obtain the purified RNA. RNA was then converted into short fragments using a fragmentation buffer. First-strand cDNA was synthesised using random N6 primers, followed by second-strand cDNA synthesis. The ends of double cDNA were repaired, 5' ends were phosphorylated, and 3' ends formed cohesive ends with A-Tailing. Then, cDNA was ligated to the sequencing adapters. The ligation products were amplified using PCR to build a cDNA library and sequenced on the BGISEQ-500 platform (Beijing Genomics Institute, Shenzhen, China). After RNA-seq, raw reads containing sequencing adapters with more than 5% unknown base or more than 20% low-quality base were removed using SOAPnuke software (version 1.5.2) to obtain clean reads [47]. Furthermore, clean reads were aligned to the PacBio reference sequence, taking the comparison ratio and the distribution of the reference sequence as conditions for further analysis.

### PacBio Iso-seq library construction and sequencing

According to the PacBio Sequencing protocol, a Clontech UMI base PCR cDNA Synthesis Kit (BGI-Shenzhen) was used to synthesise first-strand cDNA. The CDS Primer was first annealed to the polyA + tail of transcripts, followed by full-length first-strand synthesis with Reverse Transcriptase. Subsequently, double-stranded cDNA was produced by large-scale PCR. The cDNAs were measured using a Qubit HS (Life Technologies, Carlsbad, CA, USA) and an Agilent 2100 Bioanalyzer (Agilent DNA 12000 Reagents; Agilent Technologies) and then used for sequencing using a PacBio Sequel sequencer (BGI-Shenzhen) with a Sequel Sequencing Kit 2.1 and a Sequel SMRT Cell 1M v2 Tray.

### Iso-Seq data processing

The raw data from the PacBio Sequel (Pacific Biosciences, Menlo Park, CA, USA) were processed by the SMRT analysis suite (version 2.3.0) [48]. ROIs were recognised from sub-read data, and a circular consensus sequence (CCS) was generated according to the condition that the minimum predicted consensus accuracy was 0.75. Then, 3' and 5' primers of CCS were detected and sequence-oriented by internal scripting. After CCSs shorter than 300 bp in length were filtered, the rest were classified as full-length (FL) and non-full-length sequences depending on whether the 3' and 5' ends and poly (A) tail were simultaneously observed. The FL reads were corrected with RNA-seq reads using the LSC software [24]. Subsequently, the FL reads were clustered to generate *de novo* consensus isoforms via the iterative clustering for error correction algorithm and then polished using the Quiver quality-aware algorithm [49]. Two strategies were employed to improve the accuracy of consensus transcripts. First, the Quiver algorithm was used for isoform rectification to distinguish between high-quality (the expected Quiver accuracy ≥ 0.95) and low-quality isoforms. Second, the *de novo* consensus isoforms of high quality were merged to remove redundancy using CD-HIT obtaining final unique isoform sequences [50].

### Functional annotation

To derive more integral annotation information, the final unique full-length isoforms were mapped to seven functional databases, including NCBI non-redundant nucleotide sequence (NT, ftp://ftp.ncbi.nlm.nih.gov/blast/db), NCBI non-redundant protein sequences (NR, ftp://ftp.ncbi.nlm.nih.gov/blast/db), Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg), Clusters of euKaryotic Orthologous Groups (KOG, http://www.ncbi.nlm.nih.gov/KOG), SwissProt (http://ftp.ebi.ac.uk/pub/databases/swissprot), Gene Ontology (GO, http://geneontology. org), and Pfam (http://pfam.xfam.org). Nr, KOG, KEGG, and SwissProt annotations for isoforms were obtained using Blastx and Diamond software [51, 52]. Isoforms were annotated in the NT database using the Blastn software. GO annotation and classification were performed in the Blast2GO program based on the Nr annotation results [53].

### Analysis of differentially expressed genes

Quantitative analysis of gene expression levels in different organs was performed using RSEM software (version 1.2.8). Briefly, clean reads obtained from RNA-seq were mapped onto reference full-length transcripts using the Bowtie2 software (version 2.2.5) [54]. Subsequently, the expression level of each sample was calculated using RSEM software (version 1.2.12), and the read counts were normalised using fragment per kilobase of transcript per million fragments mapped (FPKM) [55]. DEGseq2 (version 1.4.5) was used to screen differentially expressed genes (DEGs) with a Q value of ≤ 0.001 [56, 57]. DEGs were mapped to GO and KEGG databases to obtain annotated information by Phyper based on a hypergeometric test for further enrichment and classification analyses. The *P*-values were corrected to Q-values with a threshold Q value of ≤ 0.05 using a Bonferroni correction.

### Identification of β-amyrin synthase (β-AS) genes

The full-length sequences encoding β-AS were confirmed after alignment of their amino acid sequences to the NCBI BLAST database (https://blast.ncbi.nlm.nih.gov/Blast.cgi), and their conserved domains were identified using NCBI Online tools (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). The β-AS genes were further selected to create the 3D constructure models using SWISS-MODEL (https://swissmodel.expasy.org/) and PyMOL software. To further determine the β-AS genes, a phylogenetic tree was constructed in MEGA7.0 software using the neighbour-joining method. A bootstrap of 1000 replications was also used to test the confidence levels, and the scale bar represented a 0.2 amino acid substitution per site.

### Quantitative real-time PCR validation

To verify the expression of genes from combined sequencing platforms, 20 DEGs were arbitrarily selected for qRT-PCR validation. Reverse transcription of isolated total RNA was processed using PrimeScript™ II 1st Strand cDNA Synthesis Kit (Waryong, China). A qRT-PCR analysis was performed in 96-well plates on an Agilent Mx3000P system (Agilent Technologies) using a TB Green® Premix Ex Taq™ II kit. The reaction conditions were as follows: 95 °C for 30 s, 40 cycles of 95 °C for 15 s, and 60 °C for 30 s. The melting curve was generated by heating the amplicon from 60 °C to 72 °C. The relative expression of each selected transcript was normalised to the expression of the internal reference gene β-actin and calculated via the $2^{-\Delta\Delta Ct}$ method [58]. Primers for all transcripts were designed using Primer software (version 5.0).

## Abbreviations

SMRT: single-molecule real-time; NGS: next-generation sequencing; DEGs: differentially expressed genes; HPLC-ELSD: High performance liquid chromatography-Evaporative Light Scattering Detector; TF: Transcription factors; FLNC: full-length non-chimeric; ICE: Interactive Clustering and Error Correction; NT: NCBI non-redundant nucleotide sequence; NR: NCBI non-redundant protein sequences; KEGG: Kyoto Encyclopedia of Genes and Genomes; KOG: Clusters of euKaryotic Orthologous Groups; GO: Gene Ontology; AACT: acetyl-CoA C-acetyltransferase; HMGS: hydroxymethylglutaryl-CoA synthase; HMG-CoA: Hydroxymethylglutaryl-CoA; HMGR: hydroxymethylglutaryl-CoA reductase; MVA: mevalonate; MK: mevalonate kinase; MVAP: (R)-Mevalonic acid 5-phosphate; PMK: phosphomevalonate kinase; MVAPP: (R)-5-Diphosphomevalonate; MDC: diphosphomevalonate decarboxylase; IPP: Isopentenyl diphosphate; G3P: D-glyceraldehyde 3-phosphate; DXS: 1-deoxy-D-xylulose-5-phosphate synthase; DXP: 1-deoxy-D-xylulose 5-phosphate; DXR: 1-deoxy-D-xylulose-5-phosphate reductoisomerase; MEP: 2-C-Methyl-D-erythritol 4-phosphate; MCT: 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase; CME: 4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol; CMK: 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; PCME: 2-Phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol; MCS: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; MECP: 2-C-Methyl-D-erythritol 2,4-cyclodiphosphate; HDS: (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase; HMBDP: 1-Hydroxy-2-methyl-2-butenyl 4-diphosphate; HDR: 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase; DMAPP: Dimethylallyl diphosphate; GPP: Geranyl diphosphate; FPPS: farnesyl diphosphate synthase; FPP: Farnesyl diphosphate; SQS: farnesyl-diphosphate farnesyltransferase; SQE: squalene monooxygenase; β-AS: β-amyrin synthase; β-A28O: β-amyrin 28-oxidase; MJ: methyl jasmonate; qRT-PCR: real-time quantitative PCR; RACE: rapid-amplification of cDNA ends.

## Declarations


### Acknowledgements

We would like to thank the BGI Institute for assistance with experiments.


### Authors' contributions

All authors read and approved the manuscript. H.Y., M.L., L.Z. and S.G. initiated and designed the research, H.Y., M.Y., T.S. and H.P. performed the experiments, H.Y., J.W., X.C. and D.P. analyzed the data and wrote the manuscript, and L.Z. and S.G. revised and edited the manuscript and provided advice on the experiments.


### Funding

This work was supported by National Natural Science Foundation of China (82073957, 81703633, 81773853), the Anhui Provincial Natural Science Foundation (1808085QH290),the Special Fundfor Guiding Local Science and

Technology Development, awarded by the Central Government of Anhui Province (YDZX20183400004233), the Key Project at the Central Government Level: The Ability Establishment of Sustainable Use for ValuableChinese Medicine Resources (2060302), the CAMS Innovation Fund for Medical Sciences (2019-I2M-5-065), and the Major scientific and technological projects in Anhui Province (18030801128). These funding bodies took part in the design of the study and collection, analysis, and interpretation of data, and the writing of the manuscript, as well as in the open access payment.

### Availability of data and materials

The Sequencing dataset(s) supporting the conclusions of this article is (are) available in the NCBI Sequence Read Archive (SRA) repository, accession number PRJNA688328 under the following link: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA688 328.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## References

1. Aarthy T, Mulani FA, Pandreka A, et al. Tracing the biosynthetic origin of limonoids and their functional groups through stable isotope labeling and inhibition in neem tree (Azadirachta indica) cell suspension. BMC Plant Biol. 2018;18(1):230.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–410.
3. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
4. Bao E, Xie F, Song C, Song D. FLAS: fast and high-throughput algorithm for PacBio long-read self-correction. Bioinformatics. 2019;35(20):3953–3960.
5. Beijing: Chemical Industry Press. The Pharmacopoeia Committee of China. Pharmacopoeia of the People's Republic of China. 2020;pp289.
6. Benny J, Pisciotta A, Caruso T, Martinelli F. Identification of key genes and its chromosome regions linked to drought responses in leaves across different crops through meta-analysis of RNA-Seq data. BMC Plant Biol. 2019;19(1):194.
7. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
8. Buchwald W, Szulc M, Baraniak J, Derebecka N, Kania-Dobrowolska M, Piasecka A, Bogacz A, Karasiewicz M, Bartkowiak-Wieczorek J, Kujawski R, Gryszczyńska A, Kachlicki P, Dreger M, Ożarowski M, Krajewska-Patan A, et al. The Effect of Different Water Extracts from Platycodon grandiflorum on Selected Factors Associated with Pathogenesis of Chronic Bronchitis in Rats. Molecules. 2020;25(21):5020.

9. Buhaescu I, Izzedine H. Mevalonate pathway: a review of clinical and therapeutical implications. Clin Biochem. 2007;40(9-10):575–584.
10. Chen X, Li J, Wang X, Zhong L, Tang Y, Zhou X, Liu Y, Zhan R, Zheng H, Chen W, Chen L. Full-length transcriptome sequencing and methyl jasmonate-induced expression profile analysis of genes related to patchoulol biosynthesis and regulation in Pogostemon cablin. BMC Plant Biol. 2019;19(1):266.
11. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, Li Y, Ye J, Yu C, Li Z, Zhang X, Wang J, Yang H, Fang L, Chen Q. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience. 2018;7(1):1–6.
12. Chesnut RM, Temkin N, Dikmen S, Rondina C, Videtta W, Petroni G, Lujan S, Alanis V, Falcao A, de la Fuenta G, Gonzalez L, Jibaja M, Lavarden A, Sandi F, Mérida R, Romero R, Pridgeon J, Barber J, Machamer J, Chaddock K. A Method of Managing Severe Traumatic Brain Injury in the Absence of Intracranial Pressure Monitoring: The Imaging and Clinical Examination Protocol. J Neurotrauma. 2018;35(1):54–63.
13. Choi D, Ward BL, Bostock RM. Differential induction and suppression of potato 3-hydroxy-3-methylglutaryl coenzyme A reductase genes in response to Phytophthora infestans and to its elicitor arachidonic acid. Plant Cell. 1992;4(10):1333–1344.
14. Choi YH, Yoo DS, Cha MR, Choi CW, Kim YS, Choi SU, Lee KR, Ryu SY. Antiproliferative effects of saponins from the roots of Platycodon grandiflorum on cultured human tumor cells. J Nat Prod. 2010;73(11):1863–1867.
15. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–3676.
16. Débora VE, Kijne JW, Memelink J. Transcription factors controlling plant secondary metabolism: what regulates the regulators?. Phytochemistry. 2002, 61(2):107–114.
17. Fu CL, Liu Y, Leng J, Zhang J, He YF, Chen C, Wang Z, Li W. Platycodin D protects acetaminophen-induced hepatotoxicity by inhibiting hepatocyte MAPK pathway and apoptosis in C57BL/6J mice. Biomed Pharmacother. 2018;107:867–877.
18. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–3152.
19. Hu Q, Pan R, Wang L, Peng B, Tang J, Liu X. Platycodon grandiflorum induces apoptosis in SKOV3 human ovarian cancer cells through mitochondrial-dependent pathway. Am J Chin Med. 2010;38(2):373–386.
20. Hwang KA, Hwang YJ, Im PR, Hwang HJ, Song J, Kim YJ. Platycodon grandiflorum Extract Reduces High-Fat Diet-Induced Obesity Through Regulation of Adipogenesis and Lipogenesis Pathways in Mice. J Med Food. 2019;22(10):993–999.
21. Ito R, Masukawa Y, Hoshino T. Purification, kinetics, inhibitors and CD for recombinant β-amyrin synthase from Euphorbia tirucalli L and functional analysis of the dcta motif, which is highly conserved among oxidosqualene cyclases. FEBS J. 2013;280(5):1267–1280.
22. Jang KJ, Kim HK, Han MH, Oh YN, Yoon HM, Chung YH, Kim GY, Hwang HJ, Kim BW, Choi YH. Anti-inflammatory effects of saponins derived from the roots of Platycodon grandiflorus in lipopolysaccharide stimulated BV2 microglial cells. Int J Mol Med. 2013;31(6):1357–1366.
23. Jin J, Panicker D, Wang Q, Kim MJ, Liu J, Yin JL, Wong L, Jang IC, Chua NH, Sarojam R. Next generation sequencing unravels the biosynthetic ability of spearmint (Mentha spicata) peltate glandular trichomes through comparative transcriptomics. BMC Plant Biol. 2014;14:292.
24. Jo HJ, Han JY, Hwang HS, Choi YE. β-Amyrin synthase (EsBAS) and β-amyrin 28-oxidase (CYP716A244) in oleanane-type triterpene saponin biosynthesis in Eleutherococcus senticosus. Phytochemistry. 2017;135:53–63.

25. Kamps R, Brandão RD, Bosch BJ, Paulussen AD, Xanthoulea S, Blok MJ, Romano A. Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. Int J Mol Sci. 2017;18(2):308.
26. Ke W, Bonilla-Rosso G, Engel P, Wang P, Chen F, Hu X. Suppression of High-Fat Diet-Induced Obesity by Platycodon Grandiflorus in Mice Is Linked to Changes in the Gut Microbiota. J Nutr. 2020;150(9):2364–2374.
27. Kim J, Kang SH, Park SG, Yang TJ, Lee Y, Kim OT, Chung O, Lee J, Choi JP, Kwon SJ, Lee K, Ahn BO, Lee DJ, Yoo SI, Shin IG, Um Y, Lee DY, Kim GS, Hong CP, Bhak J, Kim CK. Whole-genome, transcriptome, and methylome analyses provide insights into the evolution of platycoside biosynthesis in Platycodon grandiflorus, a medicinal plant. Hortic Res. 2020;7:112.
28. Kim TW, Lee HK, Song IB, Lim JH, Cho ES, Son HY, Jung JY, Yun HI. Platycodin D attenuates bile duct ligation-induced hepatic injury and fibrosis in mice. Food Chem Toxicol. 2013;51:364–369.
29. Kim YJ, Choi JY, Ryu R, Lee J, Cho SJ, Kwon EY, Lee MK, Liu KH, Rina Y, Sung MK, Choi MS. Platycodon grandiflorus Root Extract Attenuates Body Fat Mass, Hepatic Steatosis and Insulin Resistance through the Interplay between the Liver and Adipose Tissue. Nutrients. 2016;8(9):532.
30. Kushiro T, Shibuya M, Masuda K, Ebizuka Y. Mutational studies on triterpene synthases: engineering lupeol synthase into β-amyrin synthase. J Am Chem Soc. 2000, 122(29):6816–6824.
31. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–359.
32. Lee S, Han EH, Lim MK, Lee SH, Yu HJ, Lim YH, Kang S. Fermented Platycodon grandiflorum Extracts Relieve Airway Inflammation and Cough Reflex Sensitivity In Vivo. J Med Food 2020;23(10):1060–1069.
33. Leng J, Wang Z, Fu CL, Zhang J, Ren S, Hu JN, Jiang S, Wang YP, Chen C, Li W. NF-κB and AMPK/PI3K/Akt signaling pathways are involved in the protective effects of Platycodon grandiflorum saponins against acetaminophen-induced acute hepatotoxicity in mice. Phytother Res. 2018;32(11):2235–2246.
34. Li B, Dewey CN. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
35. Li Q, Li Y, Song J, Xu H, Xu J, Zhu Y, Li X, Gao H, Dong L, Qian J, Sun C, Chen S. High-accuracy de novo assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. New Phytol. 2014;204(4):1041–1049.
36. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25(4):402–408.
37. Liu D, Chen L, Chen C, An X, Zhang Y, Wang Y, Li Q. Full-length transcriptome analysis of Phytolacca americana and its congener P. icosandra and gene expression normalization in three Phytolaccaceae species. BMC Plant Biol. 2020;20(1):396.
38. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
39. Ma CH, Gao ZJ, Zhang JJ, Zhang W, Shao JH, Hai MR, Chen JW, Yang SC, Zhang GH. Candidate Genes Involved in the Biosynthesis of Triterpenoid Saponins in Platycodon grandiflorum Identified by Transcriptome Analysis. Front Plant Sci. 2016;7:673.
40. Mahjoub A, Hernould M, Joubès J, Decendit A, Mars M, Barrieu F, Hamdi S, Delrot S. Overexpression of a grapevine R2R3-MYB factor in tomato affects vegetative development, flower morphology and flavonoid and terpenoid metabolism. Plant Physiol Biochem. 2009;47(7):551–561.
41. Mertens J, Pollier J, Vanden Bossche R, Lopez-Vidriero I, Franco-Zorrilla JM, Goossens A. The bHLH Transcription Factors TSAR1 and TSAR2 Regulate Triterpene Saponin Biosynthesis in Medicago truncatula. Plant Physiol. 2016;170(1):194–210.

42. Niu Y, Luo H, Sun C, Yang TJ, Dong L, Huang L, Chen S. Expression profiling of the triterpene saponin biosynthesis genes FPS, SS, SE, and DS in the medicinal plant Panax notoginseng. Gene. 2014;533(1):295–303.
43. Oh J, Shin Y, Ha IJ, Lee MY, Lee SG, Kang BC, Kyeong D, Kim D. Transcriptome Profiling of Two Ornamental and Medicinal Papaver Herbs. Int J Mol Sci. 2018;19(10):3192.
44. Patterson J, Carpenter EJ, Zhu Z, An D, Liang X, Geng C, Drmanac R, Wong GK. Impact of sequencing depth and technology on de novo RNA-Seq assembly. BMC Genomics. 2019;20(1):604.
45. Paul P, Singh SK, Patra B, Liu X, Pattanaik S, Yuan L. Mutually Regulated AP2/ERF Gene Clusters Modulate Biosynthesis of Specialized Metabolites in Plants. Plant Physiol. 2020;182(2):840–856.
46. Peng, Y., Zhang, F., Tao, H., Wang, W., Sun, L., Chen, W., & Wang, C. Simultaneous determination of multiple platycosides with a single reference standard in Platycodi Radix by high-performance liquid chromatography coupled with evaporative light scattering detection. J Sep Sci. 2015;38(21):3712–3719.
47. Pütter KM, van Deenen N, Unland K, Prüfer D, Schulze Gronover C. Isoprenoid biosynthesis in dandelion latex is enhanced by the overexpression of three key enzymes involved in the mevalonate pathway. BMC Plant Biol. 2017;17(1):88.
48. Rao J, Peng L, Liang X, Jiang H, Geng C, Zhao X, Liu X, Fan G, Chen F, Mu F. Performance of copy number variants detection based on whole-genome sequencing by DNBSEQ platforms. BMC Bioinformatics. 2020;21(1):518.
49. Shan C, Wang C, Zhang S, Shi Y, Ma K, Yang Q, Wu J. Transcriptome analysis of Clinopodium gracile (Benth.) Matsum and identification of genes related to Triterpenoid Saponin biosynthesis. BMC Genomics. 2020;21(1):49.
50. Sun R, Liu S, Tang ZZ, Zheng TR, Wang T, Chen H, Li CL, Wu Q. β-Amyrin synthase from Conyza blinii expressed in Saccharomyces cerevisiae. FEBS Open Bio. 2017;7(10):1575–1585.
51. Thoma R, Schulz-Gasch T, D'Arcy B, Benz J, Aebi J, Dehmlow H, Hennig M, Stihle M, Ruf A. Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. Nature. 2004;432(7013):118–122.
52. Um Y, Jin ML, Lee DY, Kim CK, Hong CP, Lee Y, Kim OT. Functional characterization of the β-amyrin synthase gene involved in platycoside biosynthesis in Platycodon grandiflorum. Hortic Environ Biotechnol. 2017;58(6):613–619.
53. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018;34(9):666–681.
54. Wobbe L. The molecular function of plant mTERFs as key regulators of organellar gene expression. Plant Cell Physiol. 2020;pcaa132.
55. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.
56. Zautner AE, Goldschmidt AM, Thürmer A, Schuldes J, Bader O, Lugert R, Groß U, Stingl K, Salinas G, Lingner T. SMRT sequencing of the Campylobacter coli BfR-CA-9557 genome sequence reveals unique methylation motifs. BMC Genomics. 2015;16:1088.
57. Zhang LL, Huang MY, Yang Y, Huang MQ, Shi JJ, Zou L, Lu JJ. Bioactive platycodins from Platycodonis Radix: Phytochemistry, pharmacological activities, toxicology and pharmacokinetics. Food Chem. 2020;327:127029.
58. Zhang X, Yu Y, Jiang S, Yu H, Xiang Y, Liu D, Qu Y, Cui X, Ge F. Oleanane-Type Saponins Biosynthesis in Panax notoginseng via Transformation of β-Amyrin Synthase Gene from Panax japonicus. J Agric Food Chem. 2019;67(7):1982–1989.

# Figures

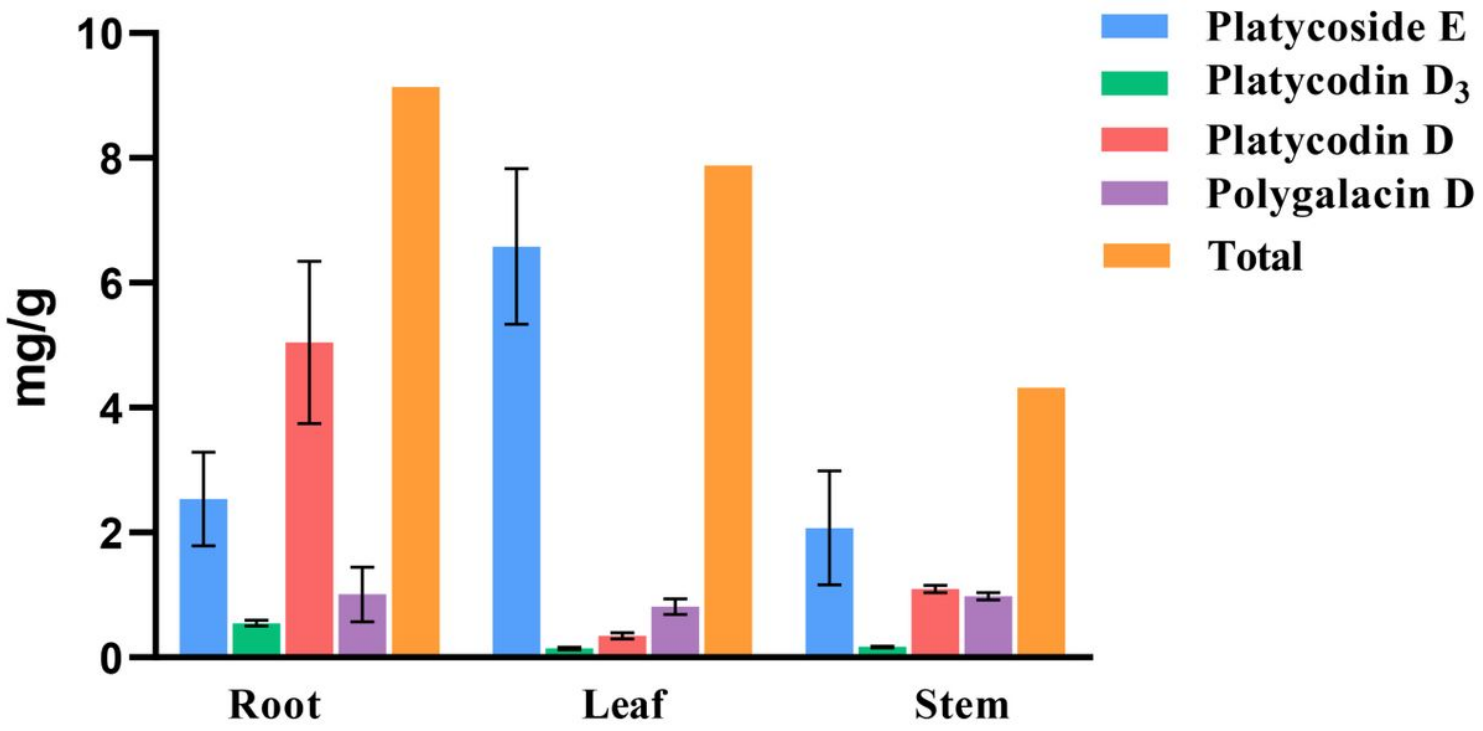
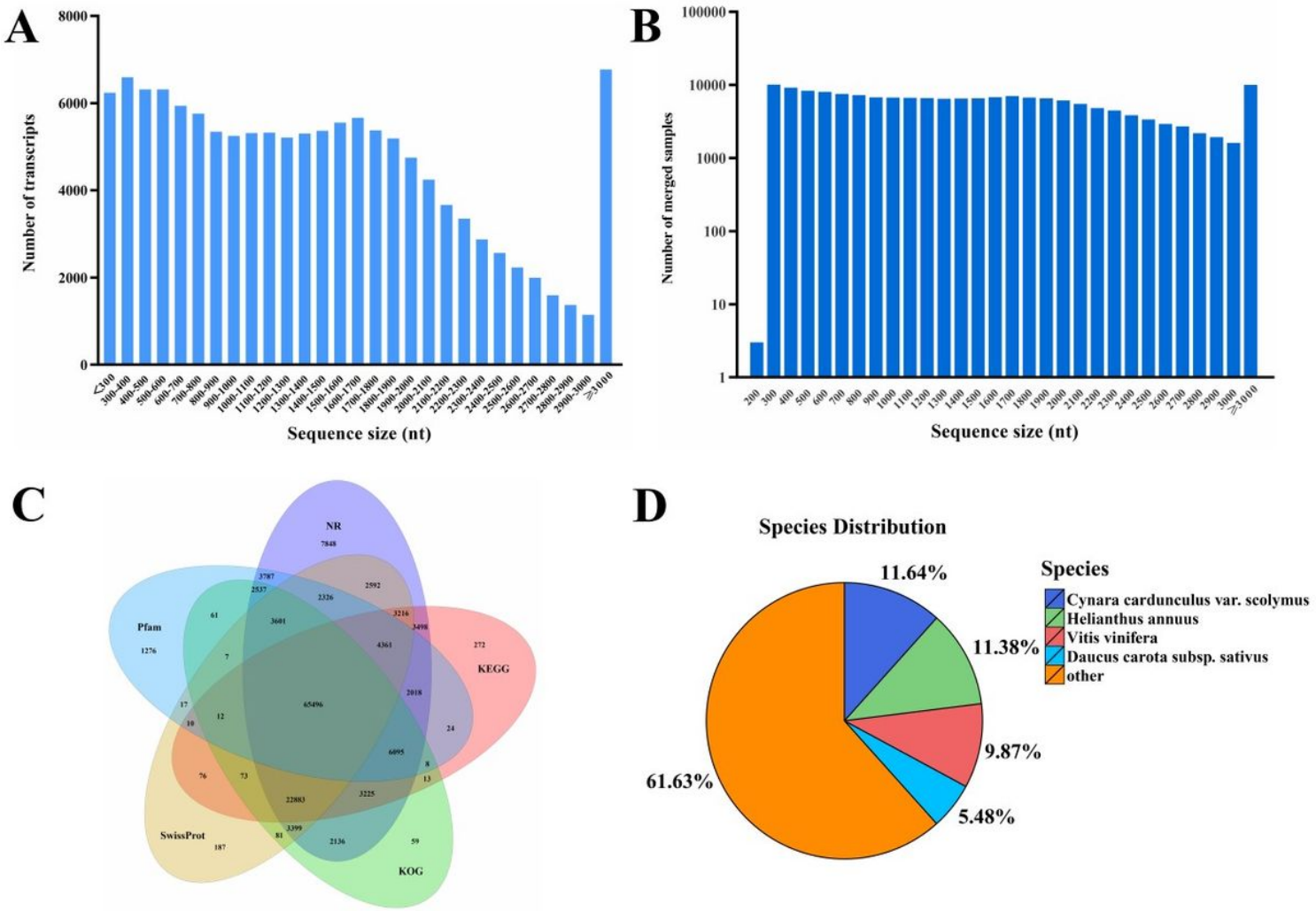


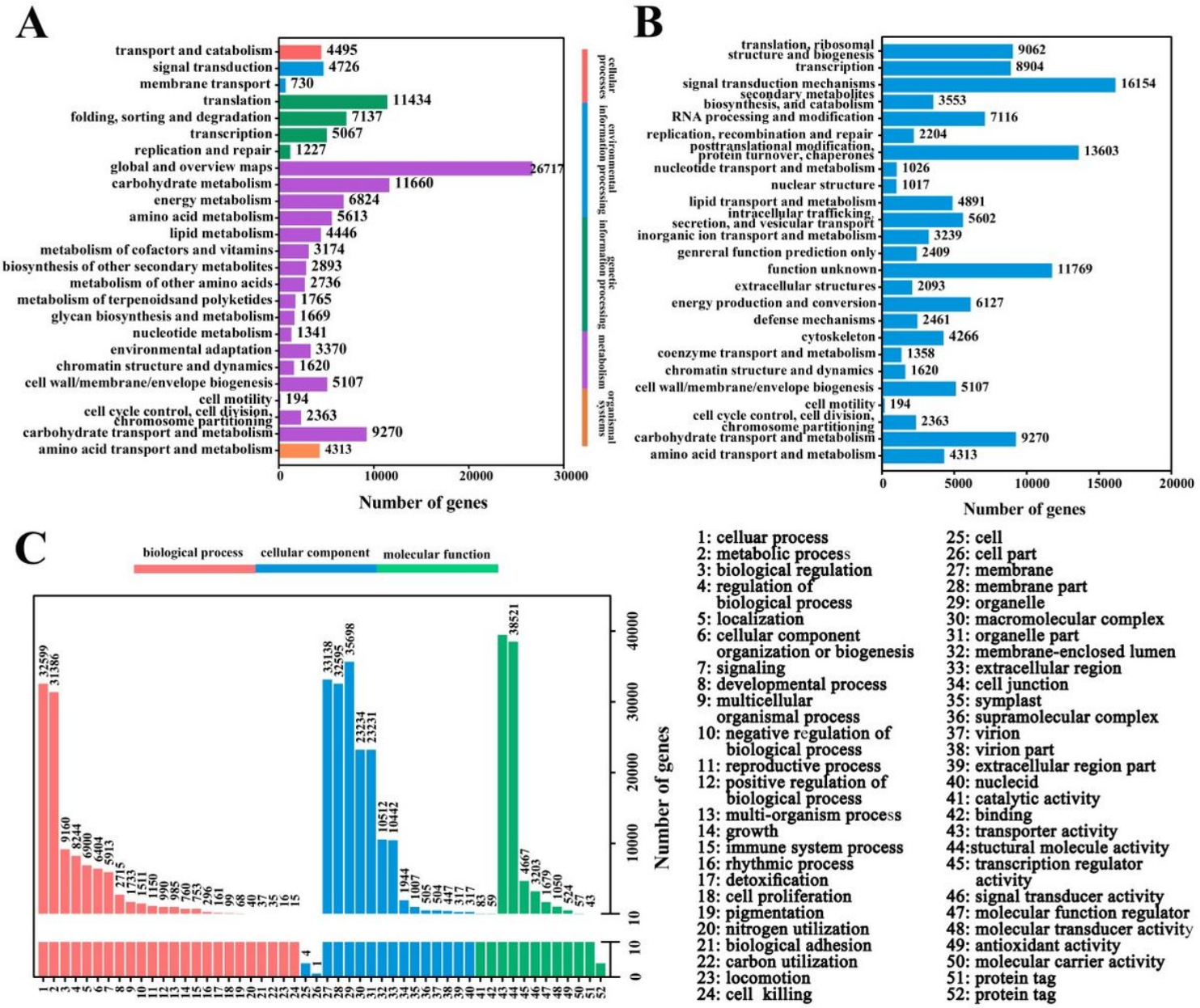

Figure 1

Contents of saponin monomers in different tissues of P. grandiflorus plants



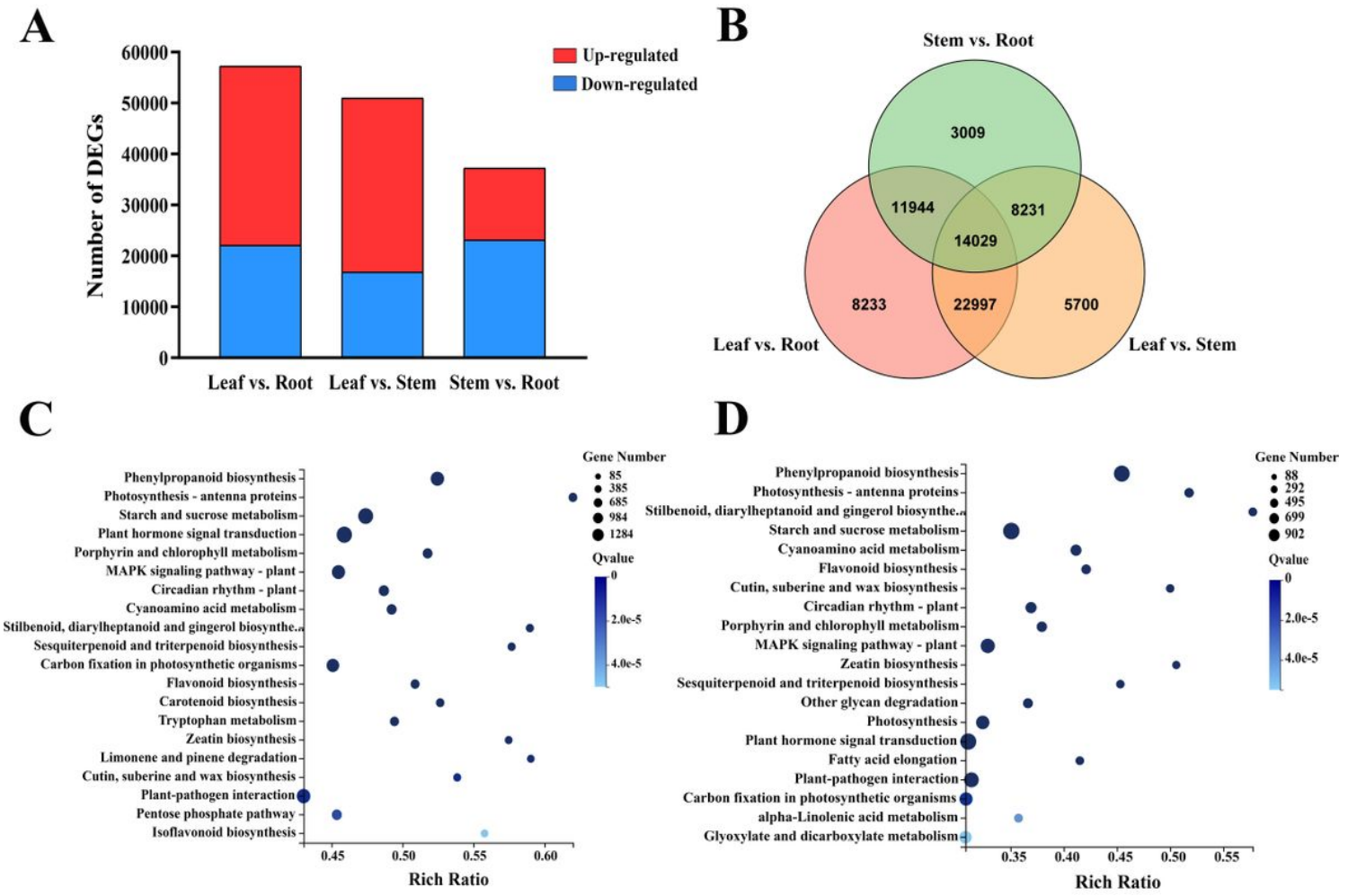
## Figure 2

Summary of RNA-Seq and Iso-Seq. a. Number and length distribution of transcripts from RNA-seq. b. Number and length distribution of non-redundant transcripts from Iso-seq. c. Functional annotation results. d. Species distribution of homologous sequences against Nr database




## Figure 3

Functional annotations for P. grandiflorus. a. KEGG pathway classification. b. KOG functional classification. c. GO functional classification




## Figure 4

Analysis of DEGs. a. Upregulated and downregulated DEGs in different tissues. b. Venn diagram of DEGs in diverse comparison groups. c. KEGG enrichment analysis of DEGs in stem vs root. d. KEGG enrichment analysis of DEGs in leaf vs root

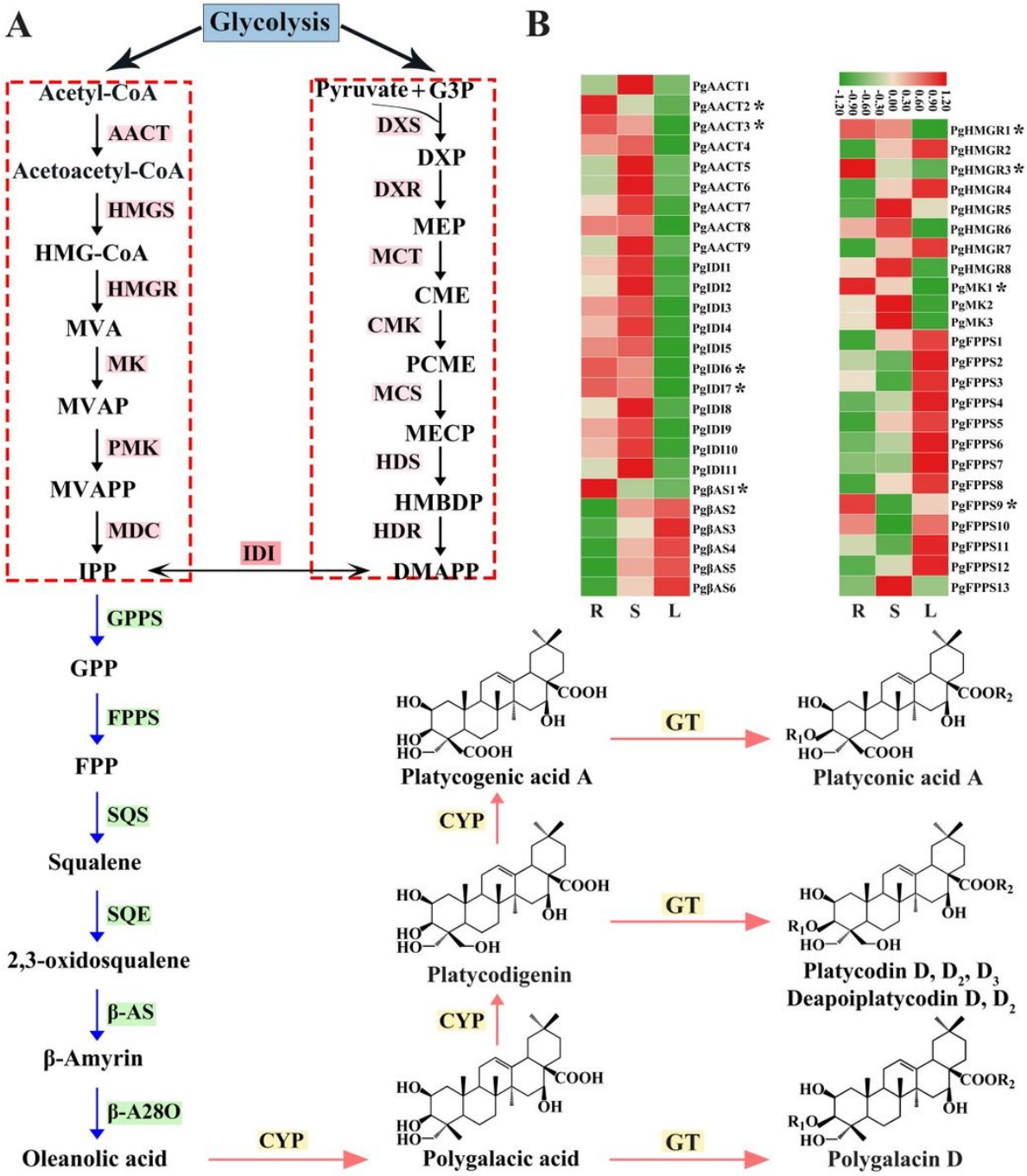



Figure 5

Pathway for triterpenoid saponin biosynthesis and relative expression patterns in P. grandiflorus. a. Pathway map of triterpenoid saponin biosynthesis. b. Expression levels of DEGs encoding key enzymes are shown using a heatmap. Asterisks (*) indicates the nine DEGs with the highest expression in the root

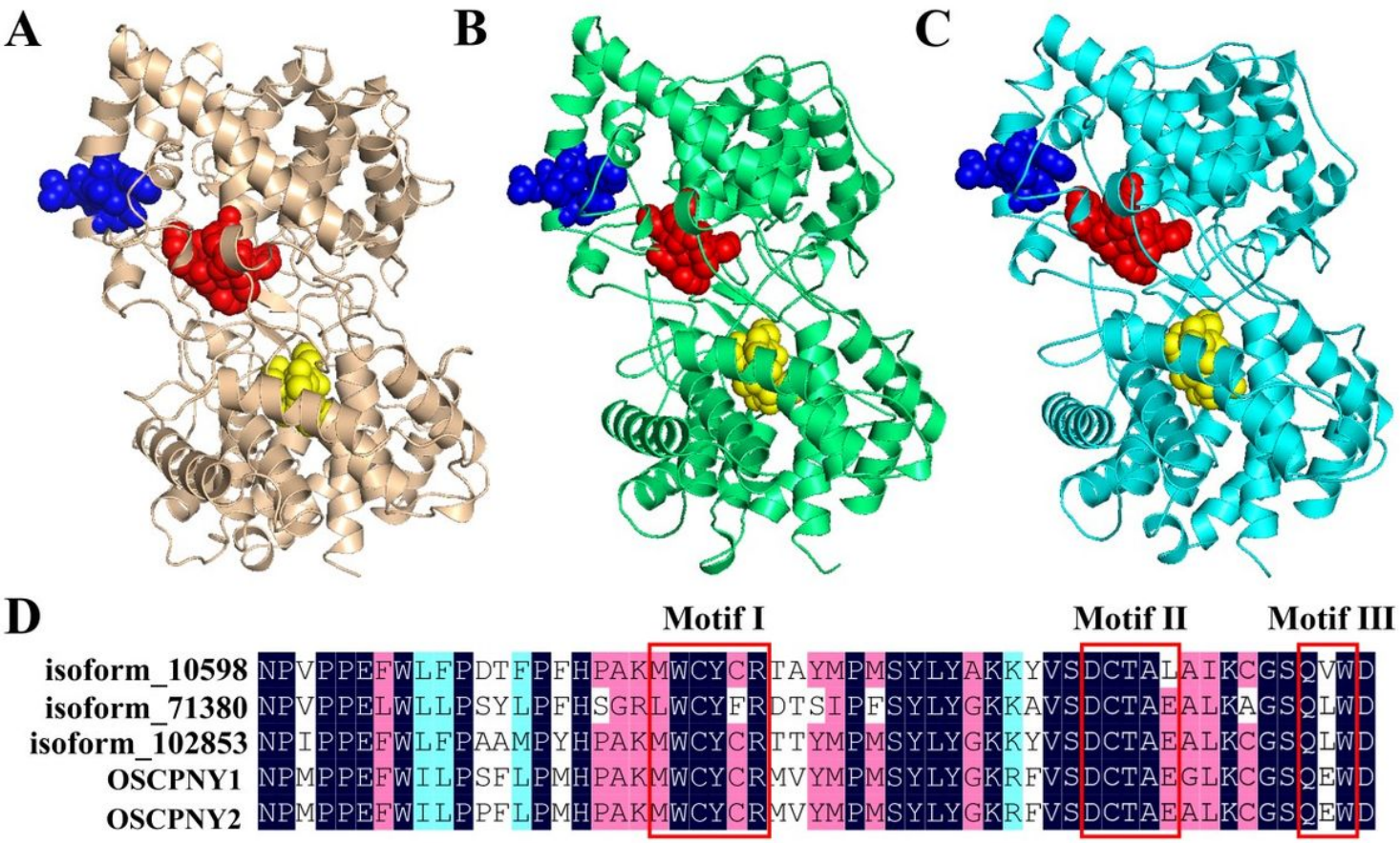



Figure 6

Spatial structure models and domain analysis of three β-AS genes. a–c Spatial structure models of isoform_10598, isoform_71380, and isoform 102853; template: 1w6k.1.A; sequence identity: 39.66, 38.50, and 39.38%, respectively). Conserved regions, motif , motif , and motif , are separately depicted as spheres in red, yellow, and blue. d. Multiple sequence alignment (dark blue indicates identical amino acids, red and light sky blue indicate similar amino acids, and the identity comparison was performed using DNAMAN software)

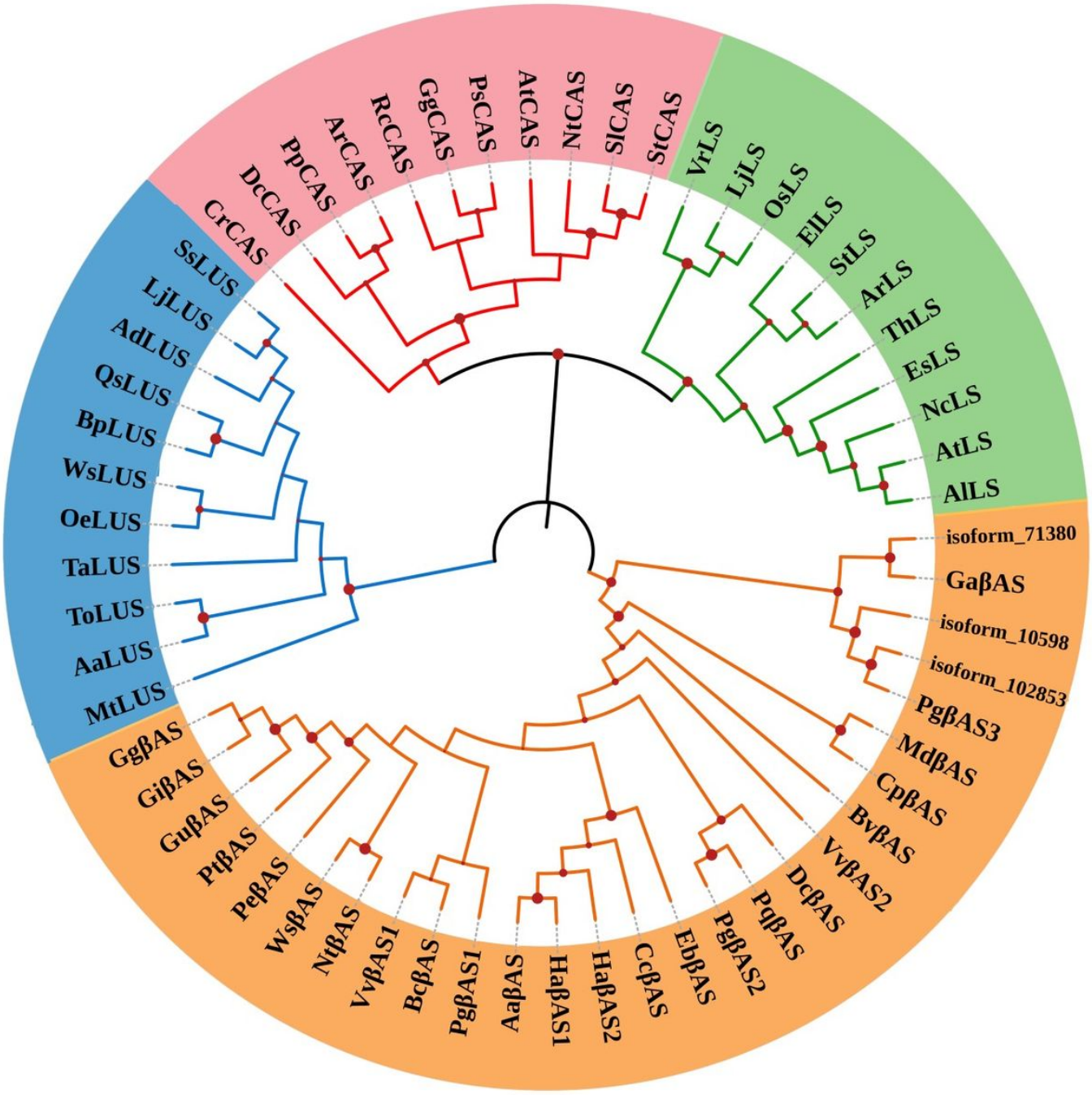



**Figure 7**

Phylogenetic tree showing the three isoforms and other oxidosqualene cyclases. The purple pentagram black indicates the three β-AS isoforms in P. grandiflorus. The neighbour-joining phylogenetic trees were constructed using the bootstrap method in MEGA 7.0, and the number of bootstrap replications was 1000

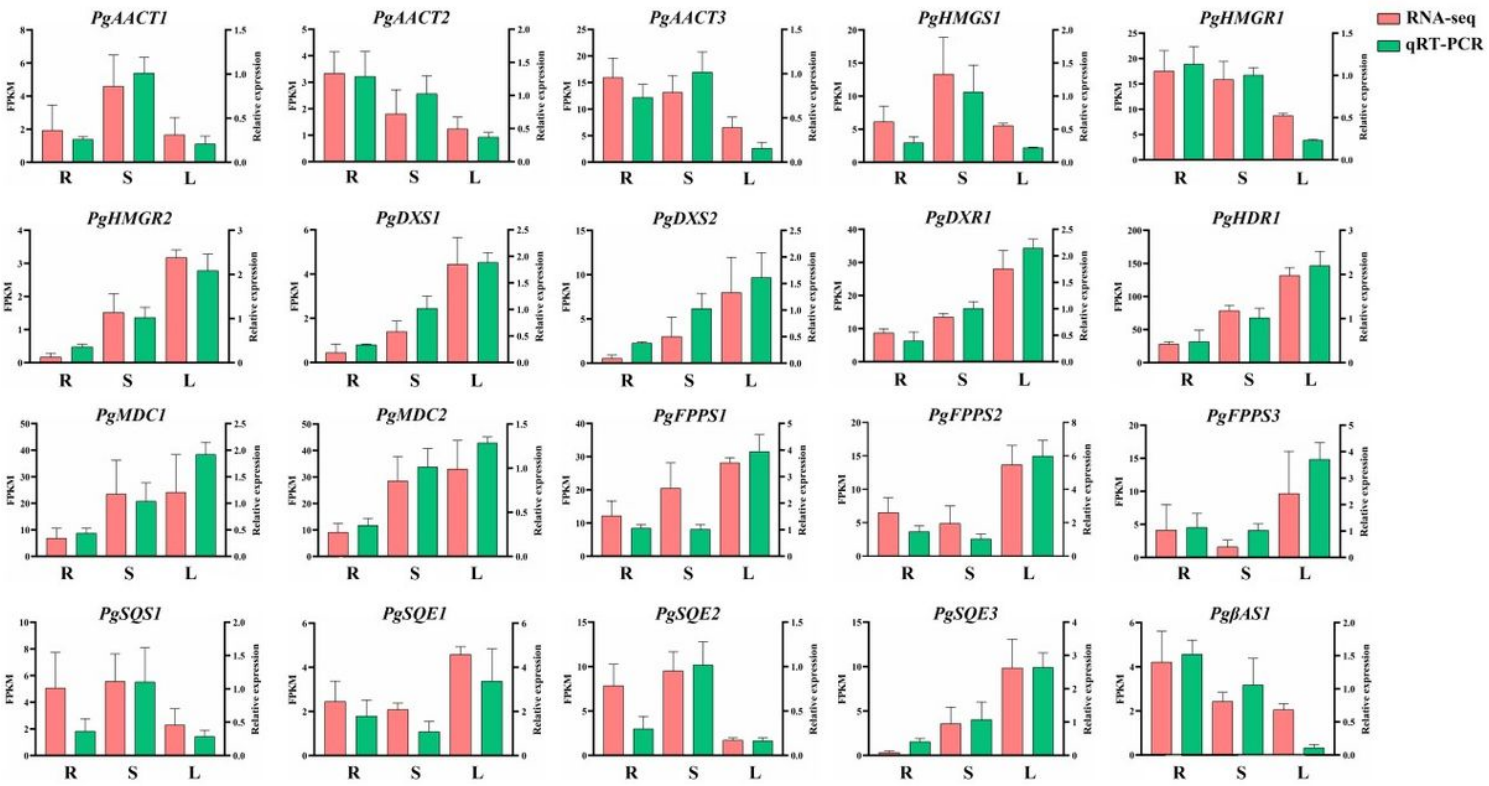



**Figure 8**

qRT-PCR validation of selected genes. The red bars represent the FPKM values of genes from RNA-seq, and green bars represent the relative expression determined by qRT-PCR. The error bars indicate the standard errors from three biological replicates

## Supplementary Files

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

- Additionalfile1.docx
- Additionalfile2.docx