

# De novo full length transcriptome analysis and gene expression profiling reveal the regulation of phenylethanol glycosides biosynthesis in *Cistanche tubulosa*

**Lei Hou**

Shandong Academy of Agricultural Sciences

**Guanghui Li**

Shandong Academy of Agricultural Sciences

**Qingliang Chen**

Peking University

**JinJin Zhao**

Hebei Agricultural University

**Jiaowen Pan**

Shandong Academy of Agricultural Sciences

**Ruxia Lin**

Shandong Academy of Agricultural Sciences

**Xiujin Zhu**

Shandong Academy of Agricultural Sciences

**Pengfei Wang**

Shandong Academy of Grape

**Xingjun Wang** (✉ [xingjunw@hotmail.com](mailto:xingjunw@hotmail.com))

Shandong Academy of Agricultural Sciences

---

## Research Article

**Keywords:** *Cistanche tubulosa*, Full-length transcriptome, PacBio, RNA-seq, Phenylethanol glycosides

**Posted Date:** June 13th, 2022

**DOI:** <https://doi.org/10.21203/rs.3.rs-1374976/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Abstract

## Background

*Cistanche* as a famous tonic Chinese traditional medicine, is known as "desert ginseng". Its dried fleshy stem is a crucial tonic in Chinese traditional medicine, and the main active pharmacodynamic components are phenylethanol glycosides. *Cistanche tubulosa* is more suitable for cultivation and extracting active ingredients because it produces higher level of phenylethanol glycosides than *C. deserticola*. The molecular regulation of the phenylethanol glycosides is not clear, partially due to the lack of genomic and transcriptomic resources in *C. tubulosa*.

## Results

In this study, we performed the full-length transcriptome sequencing and gene expression profiling of *C. tubulosa* using PacBio combined with BGISEQ-500 RNA-seq technology. Totally, 237,772 unique transcripts were obtained, the transcript lengths ranging from 199 bp to 31,857 bp, and the N50 length was 1,797 bp. Among the unique transcripts, 188,135 (79.12%) transcripts were annotated. Interestingly, 1080 transcripts were annotated as 22 enzymes related to PhGs biosynthesis. We measured the content of echinacoside, acteoside and total PhGs at two development stages, and found that the content of PhGs was 46.74% of dry matter in young fleshy stem (YS1) and then decreased to 31.22% at the harvest stage (HS2). To compare with YS1, 13,631 genes were up-regulated, and 15,521 genes were down regulated in HS2. Many differentially expressed genes (DEGs) were identified to be involved in phenylpropanoid biosynthesis pathway, phenylalanine metabolism pathway and tyrosine metabolism pathway between YS1 and HS2.

## Conclusions

Based on these results, we proposed a potential model for phenylethanol biosynthesis regulation. This is the first report of transcriptome study of *C. tubulosa* which provided the foundation for understanding of PhGs biosynthesis regulation.

## Introduction

*Cistanche tubulosa* (Schenk) R. Wight belongs to the valuable Chinese herb *Cistanche*, is a perennial parasitic plant specifically parasitic on the roots of *Tamarix ramosissima*. *C. tubulosa* which is suitable for growing on sandy land in arid and semi-arid areas. *Tamarix ramosissima* is widely distributed in the saline-alkali area in eastern China, which provides the opportunity for planting *C. tubulosa* in saline-alkali land. In recent years, *C. tubulosa* has been successfully introduced into the saline areas of the Yellow River Delta in China [1]. Growing *C. tubulosa* either in the desert area or in saline-alkali land could bring great values both economically and ecologically.

Phenylethanol glycosides (PhGs), oligosaccharides and iridoids are the main active pharmacodynamic components in *C. tubulosa*. PhGs, in particular, draw the most attention of the researchers because of their promising pharmacological characteristics. Modern Pharmacological studies have shown that PhGs has outstanding biological activities such as kidney tonifying, blood and essence invigorating, intestines moistening and defecation, anti-aging, learning ability improvement, nerve and liver protection [2–6]. *C. tubulosa* is the species with the highest content of PhGs in *Cistanche*, which is also the most widely cultivated species. However, the content of PhGs is highly regulated by both environment stimuli and development progress. The processing condition after harvest also significantly affect the content of PhGs. The difference in PhGs contents of different samples are extremely significant, even for the same species. Therefore, the quality control of medicinal materials is very important and difficult to guarantee. For example, the content of echinoside in *C. deserticola* grown in Inner Mongolia is significantly higher than that grown in Gansu [7]. Saline-alkali soil environment is beneficial to the accumulation of active components in *Cistanche deserticola*. The content of PhGs in *C. deserticola* grown in saline-alkali land was higher than those grown in grassland and sandy land [8]. The content of active components in different parts of *C. deserticola* was also different in different development stages. The content of PhGs in the base part of *C. deserticola* fleshy stem was the highest, and in the top part was the lowest. The echinoside content of the base part was eight times as higher as the top part [9].

To uncover the biosynthesis and regulation mechanisms of PhGs, to stabilize and improve the content of PhGs are all hot research area in herbal medicines. However, there are only few reports on *C. deserticola* chloroplast genome sequencing and transcriptome analysis [8, 10–11]. There is even less genomic and molecular information available in *C. tubulosa*. The limited genomic and transcriptomic information of this species hinder the advance and understanding of the molecular mechanisms of PhGs biosynthesis and regulation. The third-generation single-molecule real-time sequencing platform (SMRT) can capture the full-length transcripts and avoid the assembly process [12]. SMRT sequencing has been successfully used in many plant species for high quality full-length transcripts identification and transcriptome analysis [13–16]. Full-length transcriptome analysis is becoming the powerful way for revealing the dynamics of gene expression, and understanding the molecular mechanism for complex biological processes.

In this study, the upgraded PacBio SMRT Sequel platform was employed to generate a full-length transcriptome of *C. tubulosa*. In addition, next-generation RNA-seq technology was used to investigate the gene expression dynamics of *C. tubulosa* in different developmental stages. Our study generated a high-quality transcriptome reference sequence of *C. tubulosa*, and identified the key metabolic enzyme genes for PhGs biosynthesis. These results have laid the foundation for understanding the molecular mechanism for PhGs biosynthesis, for development of molecular makers for *C. tubulosa* authenticity test.

## Materials And Methods

### Plant materials and RNA sample preparation

The fleshy stems and flower organs of *C. tubulosa* at different developmental stages were collected from an experimental field of Shandong Academy of Agricultural Sciences in Weifang City in Shandong Province of China. Dr. Lei Hou grew the plants and undertook the identification and collection of the plant material used in this study. We could grow the plants in our experimental farm whenever we need, therefore, plant specimen was not deposited in publicly available herbarium.

Four materials were collected, including three fresh succulent stem tissues and one inflorescence tissue. The young stem, stem at harvest stage, stem at flowering period, and the inflorescence. After cleaning, the tissues were cut into small pieces and immediately frozen in liquid nitrogen, and stored at -80°C until further processing. The total RNA samples were isolated using the RNeasy Plus Mini Kit (Qiagen) according to the protocol of manufacturer (<http://www.qiagen.com>). The quantity and quality of total RNA was assessed using Agilent 2100 Bioanalyzer and Fragment Analyzer Automated CE System (<http://www.agilent.com>). For each sample, three biological replications were prepared.

### **Library preparation and high-throughput sequencing**

The PacBio UMI Iso-Seq sequencing library was constructed using equally mixed RNAs from different tissues. First-strand cDNA was synthesized using UMI base PCR cDNA Synthesis Kit (BGI). After synthesis of first strand, PCR amplification was performed to generate double-strand cDNA. Then, multiple transcripts are connected end to end to generate multifold flux full-length transcriptome Sequel PacBio IsoSeq library. The problem of different size fragment preference was solved by Sequel platform. The library was subsequently sequenced using a PacBio Sequel system.

### **PacBio Iso-Seq data processing and bioinformatic analysis**

After sequenced by PacBio sequel, large number of Circular Consensus Sequencing (CCS) reads were obtained. One CCS read could be broken down to obtain multiple transcript reads. Reads of insert (ROI) was identified using SMRT analysis package version 2.3.0 (Pacific Biosciences, <https://www.pacb.com/products-and-services/analytical-software/smrt-analysis>) following the instructions of manufacture. According to whether containing 5' primers, 3' primers and poly (A) tail, ROIs were classified into full-length (FL) and non-full-length (non-FL) reads. Then, Interactive Clustering and Error Correction (ICE) algorithm was used for de novo prediction of the consensus sequences of isoforms. The Quiver algorithm was then used to correct the isoForms. Though reads classify, reads cluster and reads quiver, the high-quality full-length consistent sequence were obtained from FLs. The integrity of the transcripts was evaluated by BUSCO (Benchmarking Universal Single-Copy Orthologs).

The high-quality full-length sequences of two libraries were combined for clustering to de-redundancy and isoform expression quantification. The TransDecoder (<https://transdecoder.github.io>) was first used for recognizing the longest Open Reading Frames (ORFs), then the ORFs were further blasted with SwissProt (<http://ftp.ebi.ac.uk/pub/databases/swissprot>) and Pfam (<http://pfam.xfam.org>). Five other public databases were used to annotate the transcripts, including Nt and Nr of NCBI (<http://www.ncbi.nlm.nih.gov/>), GO (<http://geneontology.org>), KOG (<http://www.ncbi.nlm.nih.gov/KOG>)

and KEGG (<http://www.genome.jp/kegg>). To identifying the transcription factors, HMMsearch software (<http://hmmer.org>) was used for searching the plant transcription factors database (<http://plntfdb.bio.uni-potsdam.de/v3.0/>).

The coding ability of transcripts was predicted to distinguish between mRNA and lncRNA. Three Predictive softwares (CPC, txCdsPredict and CNCI) and the protein database Pfam were used to score the coding ability of the transcripts. The scoring threshold was used to distinguish lncRNA and mRNA. The threshold of CPC and CNCI was 0, greater than 0 were mRNA and those less than 0 were lncRNA, The threshold of txCdsPredict was 500, greater than 500 were mRNA and those less than 500 were lncRNA. When the transcript could be compared with Pfam, it was considered to be mRNA, otherwise lncRNA. At least three of the four judgment methods were consistent, we could confirm that the transcript was mRNA or lncRNA.

### Bioinformatics Analysis of RNA-Seq Data

For RNA-seq, YS1 and HS2 samples were used for gene expression analysis. Libraries were constructed using the method described in previous studies [17]. Briefly, the mRNA was first enriched and cleaved into short fragments, and then used to synthesize cDNAs. The cDNA fragments were purified and enriched through PCR amplification to construct cDNA library. Sequencing was conducted using BGISEQ-500 platform (BGI, China). The clean reads were obtained by removing the reads with poly N >10%, reads containing adaptor sequences, and low-quality reads through SOAPnuke (version 1.5.2) [18]. Then, all clean reads were mapped with the full-transcripts of *C. tubulosa* using Bowtie2 (version 2.2.5) software [19]. The gene expression level was calculated using RSEM, and normalized using FPKM (Fragments Per Kilobase of transcript per Million fragments mapped) method [20]. The relative gene expression level between different samples was calculated using R package DEGseq [21]. Differentially expressed genes (DEGs) were identified following the criteria of fold change  $\geq 2$  and Q-value  $\leq 0.05$ . To further understand the function of the DEGs, GO analysis was performed and enriched GO terms were identified using Blast2GO using hypergeometric test comparing with the whole transcript background [17, 22]. In addition, enriched KEGG pathways were identified by comparing the ratio of DEGs with the whole transcript background.

### Verification of RNA-seq data by qRT-PCR

qRT-PCR was used to verify the expressional levels of 10 selected genes. RNA samples were those used for high-throughput sequencing and the reverse transcription was performed using PrimeScript II 1st Strand cDNA Synthesis Kit (TaKaRa). The gene-specific primers were designed using PerlPrimer software and were listed in Additional Table S7. qRT-PCR reaction was performed on ABI7500 Real Time System (Applied Biosystems) using TB Green™ Premix Ex Taq™ II (TaKaRa). The parameters of thermal cycle were 94°C for 10 min, followed by 30 cycles of 94°C for 15 s and 60°C for 1 min in a 20 µl volume. Three biological replications were performed for each reaction with actin gene as internal reference. The relative expression level of each gene between YS1 and HS2 was calculated by  $2^{-\Delta\Delta Ct}$  method.

## Preparation and quantification of PhGs

To extract PhGs, 50 mL of 50% methanol was added into 1.0 g of *C. tubulosa* powder, and then ultrasonic treatment was conducted for 30 min (500W, 40KHz). Add 50% methanol to supplement the weight loss, shake well, stand, and collect the supernatant. The supernatant was filtered through 0.22µm microporous membrane to obtain the phenylethanol glycosides extracts.

The contents of echinacoside and acteoside were determined by HPLC using a column of Phenomenex Luna 5µm C18(A) column (4.6×250 mm, 5µm). The isocratic mobile phase consisted of mobile phase A (methanol/acetonitrile, 1:6, v/v), and mobile phase B (0.1 % formic acid). The gradient program is as followed: 10~18 %: 90~82 % at 0~8 min, 18~24 %: 82~76 % at 8~15 min, 24~26 %: 76~74 % at 15~20 min, 26~30 %: 74~70 % at 20~28 min, 30~35 %: 70~65 % at 28~35 min. The total elution time was 35 min, with the flow rate of 1 ml/min. Column temperature was kept constant at 30 °C. Ultraviolet detection wavelength was 330 nm.

The contents of total PhGs were quantitated by ultraviolet spectrophotometry. Weigh 10 mg of echinosiden and put it in a 50 ml volumetric flask and dissolved in 50% methanol to obtain the standard solution. 0.2 mL, 0.4 ml, 0.6 ml, 0.8 ml and 1.0 ml of standard solution were accurately aliquot to a 10 ml volumetric flask, diluted to scale with 50% methanol with the final volume is 10 ml, and shaken well. The absorbance was measured at 330 nm wavelength by UV-spectrophotometer, and the standard curve was done using the standard solution. The absorbance of the sample was measured at 330 nm, and the content of total PhGs was calculated according to the standard curve.

## Results

### Full-length transcriptome sequencing and de novo assembly

Two PacBio IsoSeq libraries (F01, H01) were constructed by combining equal quantity of total RNAs from compounded four types of tissues including young succulent stem, unearthed succulent stem in harvest stage, succulent stem in flowering period and the inflorescence part above ground. PacBio sequencing generated a total of 21.87 Gb dataset including 4,321,117 subreads. A total of 254,813 Circular Consensus Sequencing CCSs reads were generated with the mean read length of 5,348 bp and 6,485 bp in the two libraries. SMRT analysis identified a total of 871,064 reads of full length non-chimeric transcripts (FLs) split from CCSs. The libraries were characterized by average lengths of 1,133 and 1,154 bp, respectively. After removing the redundant sequences, a total of 237,772 unique isoforms (transcripts) were obtained. The total length of the transcripts was 323,235,817 bp, with each transcript lengths ranging from 199 bp to 31,857 bp, and a N50 of 1,797 bp. The GC content is 41.90% (Figure 1a, Supplementary Table S1). BUSCO assessment showed that the transcripts are with high quality (Supplementary Figure S1).

To further test the integrity of the protein coding transcripts, the TransDecoder software was used for identifying the candidate coding regions in isoforms, through further blasted with the SwissProt and

Pfam Database. In total, 127,241 transcripts (53.51%) showed significant homology to known proteins and contained the complete coding sequence (CDS). The total length of CDS was more than 101 Mb with N50 of 1,002 bp and 45.59% GC content. The PacBio sequencing provided large number of full-length transcripts of *C. tubulosa*, an important genomic resource for herba *Cistanche* research community.

### Functional annotation of the transcripts

To acquire the most comprehensive annotation for these transcripts of *C. tubulosa*, the sequences of the transcripts were used for aligning with seven functional databases (Nr, Nt, SwissProt, KEGG, KOG, Pfam and GO). The results showed that 174,391 (73.34%), 142,528 (59.94%), 122,202 (51.39%), 132,164 (55.58%), 130,770 (55.00%), 105,386 (44.32%), and 125,067 (52.60%) transcripts could match the Nr, Nt, SwissProt, KEGG, KOG, Pfam and GO databases, respectively (Supplementary Table S1). In total, there were 188,135 (79.12%) transcripts were annotated by several public databases. There were 29,622 transcripts were identified actively expressed with FPKM more than 0.5.

GO analysis grouped these genes into three categories: biological process, cell composition and molecular function. In terms of molecular function, catalytic activity and binding were the most abundant classes, in terms of cell, cell part, membrane and membrane part and organelle were the main classes, in terms of biological process, cellular process and metabolic process were the most abundant classes (Figure 1b). Nr annotation showed that 64,315 (36.88%), 35,698 (20.47%) and 17,491 (10.03%) transcripts could be best matched with the species of *Sesamum indicum*, *Handroanthus impetiginosus* and *Erythranthe guttata*, respectively. The other 5,947 (3.41%) transcripts were matched with *Chenopodium quinoa* and the rest of 50,940 (29.21%) transcripts had similarity with other plant species (Figure 1c).

Three softwares (CPC, txCdsPredict and CNCI) and Pfam database were used to predict the coding sequences. The results showed that the coding mRNA and the non-coding lncRNA accounted for 7,587(32.63%), and 128,218 (53.92%), respectively. The rest 31,967 (13.44%) transcripts were determined (Figure 2, Supplementary Table S2). Transcription factors (TFs) play important roles in plant growth, development, and environment adaptation. In this study, a total of 3,862 (1.62%) transcripts encoding 57 types of TFs were identified through blasting with PlnTFDB database (Supplementary Table S3). Among these TFs, MYB, C3H, bHLH, AP2-EREBP, WRKY, NAC, and mTERF accounted for 50% of the total TFs, representing the most abundant TF families. The annotation of these transcripts provides a framework for future gene identification and gene transcriptional regulation in *C. tubulosa*.

### Determination of PhGs content in succulent stem

PhGs, for example, acteoside, echinoside, salidroside and isoacteoside are known to be the primary active molecules in *Cistanche*. Previous studies indicated that the accumulation of PhGs varies in different part of the fresh succulent stem or different growth stages of *C. tubulosa*, and the young stem accumulate more PhGs than that of the old stem [23-25]. To confirm the previous results, we measured the content of echinacoside, acteoside and total PhGs in the young stem (YS1) and stem at harvest stage (HS2) (Figure

3). The results showed that the PhGs content was 46.74 % of dry matter in YS1, while it was 31.22 % in HS2. The content of echinoside and acteoside was 28.82 % and 9.17 % in YS1, while 17.26% and 7.33% in HS2, respectively.

### Identification of enzyme genes in PhGs biosynthesis

Previous study showed that *C. tubulosa* accumulates more PhGs than *C. deserticola* [26]. However, the regulation of PhGs biosynthetic pathway and key enzymes in *C. tubulosa* is unclear. Three major components of PhGs are organic acid, saccharide and phenylethanol aglycon [11]. PhGs biosynthetic pathway is related to phenylpropanoid biosynthesis pathway (PS), phenylalanine metabolism pathway (PM) and tyrosine metabolism pathway (TM), referring to the studies of *Rehmannia glutinosa* and *C. deserticola* [11, 27]. We analyzed the genes involved in these three pathways.

Using the annotated *C. tubulosa* full-length transcriptome, we identified 1080 transcripts which encoded 22 enzymes, known involved in PhGs biosynthesis (Table 1, Supplementary Table S6). In most cases, more than one unigenes were annotated as the same enzyme, and the transcripts number encoding UDP-glycosyltransferase were the most (185 transcripts) and followed by peroxidase (138 transcripts). Among these genes, the average length of copper amine oxidase, phenylalanine ammonia-lyase and cinnamyl alcohol dehydrogenase transcripts were 2212.89 bp, 1977.84 bp, and 955.84 bp, respectively.

### Differentially expressed genes (DEGs) between different developmental stages

To further identify genes that were differentially expressed in the two stages in *C. tubulosa*, we performed RNA-seq analysis to quantify transcript abundances in YS1 and HS2. Six samples were sequenced using DNBSEQ-500 platform, and on average 21.59 Mb clean reads were generated from each sample. About 87.87% of the reads were mapped to the full-length transcriptome. A total of 175,162 genes were identified expressing in these two stages. We used a stringent cutoff of FPKM  $\geq 1$  to define transcripts that were robustly expressed in specific tissue. Among these transcripts, 75,719 (43.23%) in YS1 and 75,214 (42.94%) in HS2 were detected, representing the expressed genes in a specific stage.

Between these two stages, 29,152 DEGs were identified (Supplementary Table S4). To compare with YS1, there were 13,631 and 15,521 up- and down-regulated genes in HS2. To understand the function of DEGs, GO classification and enrichment analysis were conducted. The DEGs were classified into three main GO categories: biological process, cellular component, and molecular function (Figure 4). Cellular process and metabolic process were the top two terms in the biological process. In cellular component category, DEGs were mainly distributed in terms of cellular anatomical entity and intracellular. The most abundant terms in the molecular function were catalytic activity and binding. KEGG pathway analysis classified the DEGs into 136 metabolic pathways (Supplementary Table S5). ABC transporters, DNA replication, phenylpropanoid biosynthesis, plant-pathogen interaction, plant hormone signal transduction were observed in the top 20 enriched pathways (Figure 4). In addition, phenylalanine, tyrosine and tryptophan biosynthesis pathway appeared in the top 30 enriched pathways, which were responsible for generating phenylethanol glycosides.

## DEGs involved in phenylethanol glycosides biosynthesis

Many DEGs involved in phenylpropanoid biosynthesis, phenylalanine metabolism and tyrosine metabolism pathways between YS1 and HS2 stages were identified (Table 2, Supplementary Table S6). Many of these enzymes are composed of multiple gene family members, among which the gene expression levels vary greatly, and many genes expressed in a very low level. We only selected the genes with relative high expression level (FPKM level over 10) for further analysis.

Phenylpropanoid biosynthesis is an important pathway of secondary metabolism in plants. Cinnamic acid, coumaric acid, caffeic acid and ferulic acid are intermediates of this pathway. These acids can be further converted into coumarin, chlorogenic acid, flavonoids and lignin. In phenylpropanoid metabolic pathway, most genes of catalytic enzyme families were down-regulated in HS2 stages, including phenylalanine ammonia-lyase, cinnamic acid 4-hydroxylase, and shikimate O-hydroxycinnamoyl-transferase. All four genes of Phenylalanine ammonia-lyase (*PAL*), the first key enzyme of phenylpropanoid biosynthesis pathway, were down-regulated in HS2. In some gene families, the expression level of part of the genes were elevated, while other members were decreased, for example, coumarate 3-hydroxylase, 4-coumarate-CoA ligase, caffeoylshikimate esterase and caffeic acid 3-O-methyltransferase. However, it is worth noting that in these gene families, only one of them has increased expression, while others have decreased expression. For example, there were six transcripts with relative high expression level of *C3H* gene family, among which 4 genes decreased by more than two fold in HS2 to compare with that in YS1 (*isoform\_76310*, *isoform\_10016*, *isoform\_15416*, *isoform\_16568*), but only one transcript up-regulated (*isoform\_14095*). In *CSE* gene family, there were nine transcripts, among which the expression of four genes were greatly reduced (*isoform\_18401*, *isoform\_28866*, *isoform\_25302*, *isoform\_48623*) and one transcript increased by 7.5-fold (*isoform\_27751*). Three genes showed elevated expression in HS2, including 5-O-(4-coumaroyl)-D-quinic acid 3'-monooxygenase, caffeoyl-CoA O-methyltransferase, and Cinnamoyl-CoA reductase genes. *CYP98A3* functions to catalyze p-Coumaroyl quinic acid to caffeoyl quinic acid, or p-Coumaroyl shikimic acid to caffeoyl shikimic acid, respectively. Four *CYP98A3* genes were found in *C. tubulosa*, two of them had elevated expression by 1.45-fold and 1.58-fold, respectively. *CCR* and *CCoAOMT* are located downstream of Phenylpropanoid synthesis pathway (PS) pathway and related to lignin synthesis. In addition, *CCR* is the key enzyme in lignin synthesis pathway. Two *CCR* genes were expressed highly at HS2 (*isoform\_32104*, *isoform\_91788*), while almost undetectable at YS1. These changes of the enzyme gene expression indicated that the synthesis of phenylethanoid precursors, such as caffeic acid and ferulic acid, were channeled into lignin synthesis in HS2.

Tyrosine metabolism pathway (TM) was the main pathway for synthesizing phenylethanol aglycon in *Rehmannia glutinosa* [27]. And phenylalanine metabolism pathway (PM) was speculated the pathway for synthesizing phenylethanol aglycon in *C. deserticola* [11]. Aromatic-L-amino-acid decarboxylase (*AADC*), copper amine oxidase (*CuAO*) and aryl-alcohol dehydrogenase are present in both TM and PM pathways. We found two *AADC* genes (*isoform\_74329*, *isoform\_15697*) and seven *CuAO* genes (*isoform\_75943*, *isoform\_237680*, *isoform\_4039*, *isoform\_4504*, *isoform\_3585*, *isoform\_4105*, *isoform\_5377*) were all

down-regulated in HS2. In *Rehmannia glutinosa* and *Petroselinum crispum*, tyrosine/DOPA decarboxylase catalyzes (*TyDC*) decarboxylation of tyrosine to Tyramine [27-28]. In *C. tubulosa*, the expression levels of five members of tyrosine/DOPA family were all very low, and there was no significant difference between two stages. Therefore, we suggested that *AADC* plays an important role in this decarboxylation process in *C. tubulosa*, rather than *TyDC*. The similar situation may happen with the deamination of tyrosine, because the tyrosine aminotransferase and histidinol-phosphate aminotransferase expressed lowly and unchanged between two stages.

Glycosylation is one of the final steps involved in the biosynthesis of many plant secondary metabolites. UDP-glycosyltransferase (UGT) can transfer sugar moieties from active sugar molecules (e.g. UDP-glucose) to various acceptor molecules. There were 28 *UGT* genes identified highly expressed in this study, and 13 of which were down-regulated and three were up-regulated in HS2. The shikimate O-hydroxycinnamoyltransferase was inferred involved in catalyzing the acyl-transfer from coenzyme A-activated acids to phenylethanol aglycon [27]. Using the annotated *C. tubulosa* transcriptome assembly, we identified 13 unigenes encoding shikimate O-hydroxycinnamoyltransferase, and two of which (*isoform\_31973*, *isoform\_122217*) were down-regulated in HS2.

### Verification of DEGs using qRT-PCR

To validate the RNA-seq results, 10 genes (four involved in PS pathway, three in TM pathway, two *UGT* genes and one *GuAO* gene) with different expression levels and functions were selected for qRT-PCR analysis. The qRT-PCR expressions of these genes were in accordance with the RNA-Seq results, with the correlation coefficient of 0.8665 (Fig. 5 and Supplementary Table S7).

## Discussion

The dried succulent stems of *Cistanche* were most widely used of traditional Chinese medicines. *C. deserticola* and *C. tubulosa* are the two main medicinal plants [29]. The genomic and transcriptomic resources of *Cistanche* are mainly from *C. deserticola* [8, 10–11]. In the current study, we analysing the stem full-length transcriptome of *C. tubulosa* using PacBio SMRT Sequel platform, and obtained 237,772 unique transcripts. The proportion of the annotated transcripts using the seven functional annotations software is very low, only 79.12% annotated by several public databases, while in the case of general plants, the number is about 90% [30]. It may be related to the lack of research on parasitic plants.

Due to the superior biological activity of PhGs, the synthetic pathway has been extensively studied in order to obtain PhGs rich medicinal materials. The putative PhGs biosynthesis pathway was established based on precursor feeding experiments. Feeding tyrosine and phenylalanine to cell suspension culture of *Cistanche* can increase the accumulation of acteoside, echinoside or 2'-acetyl acteoside [31–32]. Isotope labeled feeding studies in *Syringa Vulgaris* and *Olea Europaea* showed that tyrosine to 3,4-dihydroxytyrosol (DHPA) via dopamine or tyramine pathway and phenylalanine to the coffeyl moiety through phenylpropane pathway were involved in PhGs biosynthesis [33–34]. It recognized that the organic acid acyl moiety (coffeyl, feruloyl or coumaroyl) of PhGs were generated from phenylpropanoid

synthesis pathway, while the synthesis of phenylethanol moiety has several possible choices. In *Rehmannia glutinosa*, phenylethanol part was considered be started from the tyrosine precursors by the tyrosine-derived pathway [27, 35–36]. In *C. deserticola*, the synthesis of phenylethanol part is reported presumably from two pathways. One is the caffeic acid or ferulic acid pathway, which is part of the phenylpropanoid biosynthesis pathway. The other is the phenylalanine metabolism pathway, in which the phenylethanol was converted to phenylethanol aglycon [11].

Studies indicated that tyrosine could be converted to 4-hydroxyphenyl- acetaldehyde by aminotransferase and 4-hydroxyphenylacetaldehyde synthase though 4-hydroxyphenylpyruvate [27, 37]. In this 4-hydroxyphenylpyruvate pathway, Tyrosine aminotransferase (*TAT*), histidinol-phosphate aminotransferase (*hisC*), and Aspartate aminotransferase (*GOT*) are the main active enzymes. We found that both *TAT* and *hisC* genes expressed lowly and unchanged in YS1 and HS2. In *GOT* gene family, one gene was up-regulated (isoform\_14711) and one gene was down-regulated (isoform\_17258) in HS2, the change trend of expression was not obvious. Meanwhile, we didn't found 4-hydroxyphenylpyruvate decarboxylase (*4HDPC*) coding gene in the full-length transcripts of *C. tubulosa*. Therefore, we suggested that the 4-hydroxyphenylpyruvate pathway maybe not the primary one in *C. tubulosa*. Based on these results, we proposed a potential model of phenylethanol biosynthesis regulation in *C. tubulosa* (Figure. 6). In summary, the phenylethanol part is derived from two parallel pathways, the phenylalanine metabolism pathway and tyrosine metabolism pathway in *C. tubulosa*. In the phenylalanine metabolism pathway, phenylalanine is decarboxylated under the action of aromatic-L-amino-acid decarboxylase to produce phenylethylamine, and then generate phenylethylalcohol under copper amine oxidase and aryl-alcohol dehydrogenase. In the tyrosine metabolism pathway, tyrosine or its oxidation product L-DOPA were first decarboxylated to produce tyramine or dopamine, then to tyrosol or hydroxytyrosol under the action of CuAO and AADH, respectively. However, the acyltransferase corresponding gene is still uncertain, which transfers caffeoyl- group to phenylethanol aglycon in the downstream of PhGs synthesis pathway. Inferred from transcriptome data, shikimate O-hydroxycinnamoyltransferase may play the role, but the specific catalytic function requires further study.

There were significant differences in gene expression between YS1 and HS2 stages in *C. tubulosa*, and the expression levels of many genes related to synthesis of PhGs were decreased in HS2, indicating that catalytic synthesis of PhGs was weakened at HS2 development stage, which supported the rationality of harvest the succulent stem of *C. tubulosa* at this development stage in agriculture.

## Conclusions

In the present study, the full-length transcriptome and gene expression profiling of *C. tubulosa* stem in different developmental stages were achieved. Key genes for PhGs biosynthesis were identified, and a model for PhGs biosynthesis of *C. tubulosa* was proposed. The expression patterns of key genes were consistent with the accumulation of PhGs. These results laid the foundation for further studies on molecular mechanism of PhGs biosynthesis and regulation in *C. tubulosa*.

# Abbreviations

Phenylpropanoid synthesis pathway (PS) , Tyrosine metabolism pathway (TM), phenylalanine metabolism pathway (PM), Phenylethanol glycosides (PhGs), Phenylalanine ammonia-lyase (PAL), Cinnamic acid 4-hydroxylase (C4H), Coumarate 3-hydroxylase (C3H), Caffeoylshikimate esterase (CSE), 5-O-(4-coumaroyl)-D-quinic acid 3'-monooxygenase (CYP98A3), Shikimate O-hydroxycinnamoyltransferase (HCT), 4-coumarate-CoA ligase (4CL), Caffeic acid 3-O-methyltransferase (COMT), Caffeoyl-CoA O-methyltransferase (CCoAOMT), Ferulate-5-hydroxylase (F5H), Peroxidase (PER), Cinnamoyl-CoA reductase (CCR), Cinnamyl alcohol dehydrogenase (CAD), Aromatic-L-amino-acid decarboxylase (AADC), Tyrosine/DOPA decarboxylase (TyDC), Aspartate aminotransferase (GOT), Histidinol-phosphate aminotransferase (hisC), Tyrosine aminotransferase (TAT), Copper amine oxidase (CuAO) Polyphenol oxidase (PPO), aryl-alcohol dehydrogenase (AADH), UDP-glycosyltransferase (UGT), 4-hydroxyphenylpyruvate decarboxylase (4HDPC), Differentially expressed genes (DEGs), Young fleshy stem (YS1) , Harvest stage (HS2), Fragments Per Kilobase of transcript per Million fragments (FPKM).

# Declarations

The *Cistanche tubulosa* used in this study was cultivated for this study and sampling permission is not necessary. The materials were collected from plants grown in the experimental station of Shandong Academy of Agricultural Sciences.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

All the authors have signed the consent form.

## Availability of data and materials

Raw sequence reads of the RNA-seq data were available in NCBI Short Read Archive (SRA) Database (Bioproject: PRJNA786765). The assembled full length transcriptome data have been deposited in NCBI Transcriptome Shotgun Assembly (TSA) Database (TSA submission: SUB10777875).

## Competing interests

All authors read the final version of this manuscript and agreed with the journal policy. The authors declare that they have no competing interests.

## Funding

This work was financially supported by grants from the Shandong provincial crop elite variety development project (2017LZN032), Shandong province agricultural technology innovation project,

Innovative Public Service Platform of Shandong Province: Public Service Platform for Molecular Identification of Traditional Chinese Medicine (2018JGX111), Agricultural scientific and technological innovation project of Shandong Academy of Agricultural Sciences (CXGC2018E13), Taishan Scholar Foundation of Shandong Province (ts20190964).

### Authors' contributions

XJW, LH, and PFW conceived and designed the study, GHL, JJZ, JWP, RXL, and XJZ performed the experiments. QLC analyzed the data and drafted the manuscript with LH. XJW and PFW revised the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

Thanks to professor Yuhai Guo of China Agricultural University for the guidance of *Cistanche tubulosa* cultivation.

## References

1. Hou L, Wang YT, Fan ZX, Zhao SZ, Li CS, Wang XJ. Study on Feasibility Inoculating *Cistanche tubulosa* to Wild *Tamarix chinensis* in Yellow River Delta. Mod Chin Med Apr. 2018,20(4):437-40.
2. Lin WY, Yao C, Cheng J, Kao ST, Tsai FJ, Liu HP. Molecular pathways related to the longevity promotion and cognitive improvement of *Cistanche tubulosa* in *Drosophila*. Phytomedicine. 2017,26:37-44.
3. Zhang K, Ma X, He WJ, Li HX, Han SY, Jiang Y, et al. Extracts of *Cistanche deserticola* can antagonize immunosenescence and extend life span in senescence-accelerated mouse prone 8(SAM-P8) mice. Evid. Based Complement. Alternat. Med. 2014,(2):601383.
4. Peng XM, Gao L, Huo SX, Liu XM, Yan M. The mechanism of memory enhancement of acteoside (verbascoside) in the senescent mouse model induced by a combination of D-gal and AlCl<sub>3</sub>. Phytotherapy Research. 2015,29(8) :1137-44.
5. Luo HY, Cao RR, Wang LJ, Zhu LJ. Protective effect of *Cistanche* A on ethanol-induced damage in primary cultured mouse Hepatocytes. Biomed. Pharmacother. 2016,83:1071-79.
6. Wang T, Chen C, Yang M, Deng BW, Kirby GM, Zhang XY. *Cistanche tubulosa* ethanol extract mediates rat sex hormone levels by induction of testicular steroidogenic enzymes. Pharm. Biol. 2016,54 (3):481-87.
7. Wang Y, Zhang L, Du Z, X, Pei J, Huang L. Chemical diversity and prediction of potential cultivation areas of *Cistanche* Herbs. Sci. Rep. 2019;9(1) :19737.
8. Sun X, Li L, Pei J, Liu C, Huang LF. Metabolome and transcriptome profiling reveals quality variation and underlying regulation of three ecotypes for *Cistanche deserticola*. Plant Mol Biol. 2020,102(3):253-69.

9. Guo XF, Wu YD, Ni H, Jia XG. Determination on active ingredient content for different parts of *cistanche tubulosa*. Journal of Xinjiang Medical University. 2011,29(4) :52-54
10. Li X, Zhang TC, Qiao Q, Ren ZM, Zhao JY, Yonezawa T, et al. Complete Chloroplast Genome Sequence of Holoparasite *Cistanche deserticola* (*Orobanchaceae*) Reveals Gene Loss and Horizontal Gene Transfer from Its Host *Haloxylon ammodendron* (*Chenopodiaceae*). PLoS ONE. 2013,8:e58747.
11. Li Y, Wang X, Chen T, Yao F, Li C, Tang Q, Sun M, Sun G, Hu S, Yu J, Song S. RNA-Seq Based De Novo Transcriptome Assembly and Gene Discovery of *Cistanche deserticola* Fleshy Stem. PLoS One. 2015,10(5):e0125722.
12. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015,13:278-89.
13. Schaarschmidt S, Fischer A, Lawas LMF, Alam R, Septiningsih EM, Bailey-Serres J, et al. Utilizing PacBio Iso-Seq for Novel Transcript and Gene Discovery of Abiotic Stress Responses in *Oryza sativa* L. Int J Mol Sci. 2020,31,21(21):8148.
14. Li L, Liu H, Wen W, Huang C, Li X, Xiao S, et al. Full Transcriptome Analysis of Callus Suspension Culture System of *Bletilla striata*. Front Genet. 2020,11:995.
15. Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, et al. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis. Plant J. 2015,82:951-61.
16. Hoang NV, Furtado A, Mason PJ, Marquardt A, Kasirajan L, Thirugnanasambandam PP, et al. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. BMC Genomics. 2017,18:395.
17. Wang J, Zhang Q, Cui F, Hou L, Zhao S, Xia H, et al. Genome-wide analysis of gene expression provides new insights into cold responses in *Thellungiella salsuginea*. Frontiers in plant science. 2017,8:713.
18. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010,38:1767-71.
19. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012,9:357.
20. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011,12:323.
21. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010,26:136-8.
22. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005,21:3674-6.
23. Wang X, Guo Y. Study on the chemical composition variations of *Cistanche tubulosa* during the whole growth period. Journal of China Agricultural University. 2017,22(11):28-35.

24. Yang TX, Lu YX, Guo YH, Zhai ZX, Yu GJ. Studied of dry matter accumulation and echinacoside content of *Cistanche tubulosa* in Huabei plain. China journal of Chinese materia medica. 2006,31(16):1317-20.
25. Yang TX, Du YH, Liu JN, He M, Gao Q. Determination on active ingredient content of *Cistanche tubulosa* in different growth period and different parts. Lishizhen Med Mat Med Res. 2014,25:1191-3.
26. Zhu NL, Xu R, Wu HF, Ma GX, Zhu YD, Peng F, et al. Fingerprint comparative analysis of *Cistanche deserticola* Y.C. Ma and *Cistanche tubulosa* (Schrenk) Wight. Chinese Pharmaceutical Journal,. 2016,51(13):1116-19.
27. Wang F, Zhi J, Zhang Z, Wang L, Suo Y, Xie C, et al. Transcriptome Analysis of Salicylic Acid Treatment in *Rehmannia glutinosa* Hairy Roots Using RNA-seq Technique for Identification of Genes Involved in Acteoside Biosynthesis. Front Plant Sci. 2017,8:787.
28. Torrens-Spence MP, Gillaspay G, Zhao BY, Harich K, White RH, Li JY. Biochemical evaluation of a parsley tyrosine decarboxylase results in a novel 4-hydroxyphenylacetaldehyde synthase enzyme. Biochem Bioph Res Co. 2012,418(2):211-6.
29. Han JP, Song JY, Liu C, Chen J, Qian J, Zhu YJ, Shi LC, Yao H, Chen SL. Identification of *Cistanche* species (*Orobanchaceae*) based on sequences of the plastid psbA-trnH intergenic region. Yao Xue Xue Bao. 2010,45(1):126-30.
30. Zhao C, He L, Xia H, Zhou X, Geng Y, Hou L, et al. De novo full length transcriptome analysis of *Arachis glabrata* provides insights into gene expression dynamics in response to biotic and abiotic stresses. Genomics. 2021,113(3):1579-88.
31. Liu JY, Guo ZG, Zeng ZL. Improved accumulation of phenylethanoid glycosides by precursor feeding to suspension culture of *Cistanche salsa*. Biochem Eng J. 2007,33(1):88-93.
32. Hu GS, Jia JM, Kim DH. Effects of feeding tyrosine and phenylalanine on the accumulation of phenylethanoid glycosides to *Cistanche deserticola* cell suspension culture. Chin J Nat Med. 2014,12(5):367-72.
33. Ellis BE. Production of hydroxyphenylethanol glycosides in suspension cultures of *Syringa vulgaris*. Phytochemistry. 1983,22(9):1941-43.
34. Saimaru H, Orihara Y. Biosynthesis of acteoside in cultured cells of *Olea europaea*. J Nat Med-Tokyo. 2010,64(2):139-45.
35. Wang F, Yang X, Zuo X, Miao C, Zhang Z. Full-length transcriptome sequence and identification of genes involved in phenylethanol glycoside biosynthesis in *Rehmannia glutinosa*. Acta Pharmaceutica Sinica. ISSN 0513-4870, CN 11-2163/R.  
<https://kns.cnki.net/kcms/detail/11.2163.R.20211209.0131.007.html>
36. Alipieva K, Korkina L, Orhan IE, Georgiev MI. Verbascoside-a review of its occurrence (bio)synthesis and pharmacological significance. Biotechnol. Adv. 2014,32:1065-76.
37. Zhou Y, Zhu J, Shao L, Guo M. Current advances in acteoside biosynthesis pathway elucidation and biosynthesis. Fitoterapia. 2020,142:104495.

Tables

Table 1-2 are available in the Supplementary Files section.

Figures

Figure 1

Characteristics of unique isoforms of *C. tubulosa*. a: Length distribution of the transcripts, b: GO annotation of the transcripts, c: Species mapped with the transcripts.

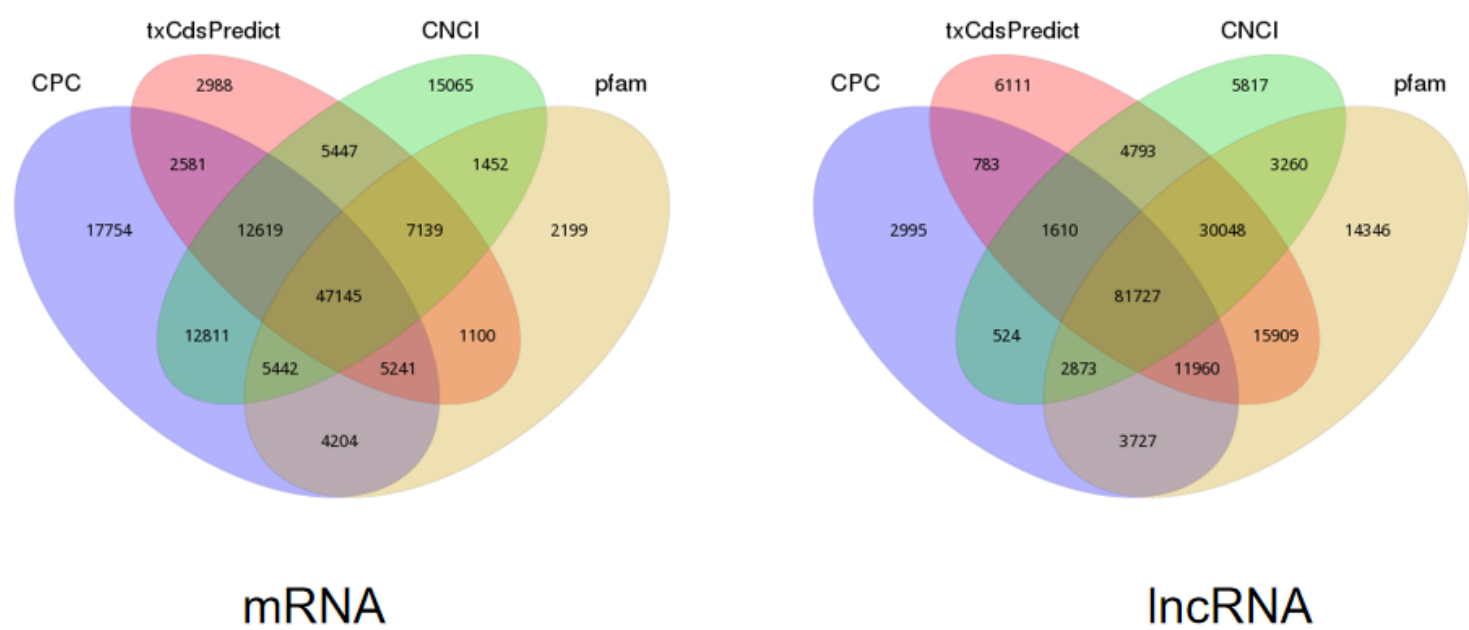


Figure 2

Venn diagram showed the prediction results of coding ability. Different colors represent different prediction methods.

Figure 3

*C. tubulosa* samples and phenylethanol glycosides content of stem at different developmental stages.

a: YS1 and HS2 stem of *C. tubulosa*, b: The contents of echinacoside, acteoside and total PhGs.

## Figure 4

GO classification and KEGG pathway enrichment analysis.

## Figure 5

Verification of DEGs by qRT-PCR. a. Transcript levels of 10 genes related to PhGs synthesis. Data are means of three replicates, and error bars represent  $\pm$ SE (n =3). b. Pearson's correlation of gene expression ratios between RNA-seq and qRT-PCR. The correlation of the fold change was analyzed by RNA-seq (x-axis) with qRT-PCR (y-axis) data.

## Figure 6

The proposed model of phenylethanol biosynthesis regulation.

# Supplementary Files

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

- [SupplementaryFigureS1.png](#)
- [SupplementaryFigureS2.png](#)
- [SupplementaryTableS1.xls](#)
- [SupplementaryTableS2.xls](#)
- [SupplementaryTableS3.xls](#)
- [SupplementaryTableS4.csv](#)
- [SupplementaryTableS5.csv](#)
- [SupplementaryTableS6.xls](#)
- [SupplementaryTableS7.xls](#)
- [Table1.xls](#)
- [Table2.xls](#)