RESEARCH ARTICLE

Transcriptome Sequencing of *Codonopsis pilosula* and Identification of Candidate Genes Involved in Polysaccharide Biosynthesis

Jian Ping Gao¹*, Dong Wang¹‡, Ling Ya Cao¹‡, Hai Feng Sun²

¹ College of Pharmacy, Shanxi Medical University, Taiyuan, Shanxi, China, ² College of Chemistry & Chemical Engineering, Shanxi University, Taiyuan, Shanxi, China

‡ These authors contributed equally to this work.

* jpgao123@sina.com

Abstract

Background

*Codonopsis pilosula* (Franch.) Nannf. is one of the most widely used medicinal plants. Although chemical and pharmacological studies have shown that codonopsis polysaccharides (CPPs) are bioactive compounds and that their composition is variable, their biosynthetic pathways remain largely unknown. Next-generation sequencing is an efficient and high-throughput technique that allows the identification of candidate genes involved in secondary metabolism.

Principal Findings

To identify the components involved in CPP biosynthesis, a transcriptome library, prepared using root and other tissues, was assembled with the help of Illumina sequencing. A total of 9.2 Gb of clean nucleotides was obtained comprising 91,175,044 clean reads, 102,125 contigs, and 45,511 unigenes. After aligning the sequences to the public protein databases, 76.1% of the unigenes were annotated. Among these annotated unigenes, 26,189 were assigned to Gene Ontology categories, 11,415 to Clusters of Orthologous Groups, and 18,848 to Kyoto Encyclopedia of Genes and Genomes pathways. Analysis of abundance of transcripts in the library showed that genes, including those encoding metallothionein, aquaporin, and cysteine protease that are related to stress responses, were in the top list. Among genes involved in the biosynthesis of CPP, those responsible for the synthesis of UDP-L-arabinose and UDP-xylose were highly expressed.

Significance

To our knowledge, this is the first study to provide a public transcriptome dataset prepared from *C. pilosula* and an outline of the biosynthetic pathway of polysaccharides in a
medicinal plant. Identified candidate genes involved in CPP biosynthesis provide understanding of the biosynthesis and regulation of CPP at the molecular level.

**Introduction**

Codonopsis Radix (CR, known as “Dangshen” in Chinese) is a famous traditional Chinese herbal medicine that has been widely used in Chinese medicine for replenishing qi (vital energy) deficiency, strengthening the immune system, improving poor gastrointestinal function, curing gastric ulcer and appetite, etc [1]. In Chinese Pharmacopoeia, CR is prescribed as dried roots of *Codonopsis pilosula* (Franch.) Nannf., *C. pilosula* Nannf. var. *modesta* (Nannf.) L. D. Shen and *C. tangshen* Oliv. of the family Campanulaceae [2]. Among these source plants, *C. pilosula* (Franch.) Nannf. is widely distributed in China and the raw herb from the species accounts for the largest part of the market [3]. At present, commercial CR is predominantly produced in Shanxi, Shaanxi, Gansu, and Sichuan. Among these different resources, CRs from Lingchuan, Shanxi taste sweet and have been considered to be of good quality for hundreds of years [4]. Chemical and pharmacological studies of CR have shown that codonopsis polysaccharides (CPPs) are major contributors to the sweet taste and are capable of modulating immune system functions, preventing tumor growth, improving memory, curing diabetes, and increasing hemoglobin [5–10]. These documents also reveal that the components’ bioactivities are related to their structure and composition. Because the content and composition of CPPs varies according to the geographic region, method used for cultivation, and the age of the plant, it is important to understand the biosynthesis, metabolism, and regulation of CPPs [11].

Using simple capillary electrophoresis, Yang et al. demonstrated that CPP is an acidic heteropolysaccharide composed of arabinose, glucose, rhamnose, galactose, mannose, glucuronic acid, and galacturonic acid in the molar contents of 48.1, 103.5, 16.1, 48.5, 7.5, 4.2, and 119.1 μmol, respectively [12]. Structural analysis of CPPS(3), a water-soluble polysaccharide, has shown that it is composed of galactose, arabinose, and rhamnose in a molar ratio of 1.13:1.12:1 [13]. There is remarkable diversity in the composition and structure of CPPs [7, 14, 15]. However, the mechanism by which CPPs are produced and accumulated in *C. pilosula* is not known. Additionally, pathways responsible for polysaccharide biosynthesis in other medicinal plants are also poorly understood [16, 17]. Use of efficient high throughput techniques may allow the identification of genes involved in the biosynthesis of polysaccharides in medicinal plants.

Transcriptome sequencing using next-generation sequencing (NGS) technology provides much information within a short time and with enormous depth and coverage [18]. To date, a large volume of sequences relevant to the biosynthesis of a number of natural products has been identified using this technique [19–24]. However, these efforts focused primarily on the biosynthesis of flavonoids, unsaturated fatty acids, triterpenoid sapogenins, and saponins. In the current study, a combined transcriptome of roots, stems, leaves, and flowers of *C. pilosula* grown in Lingchuan was developed with the help of NGS. The sequences were then analyzed to understand the biosynthesis of secondary products, particularly to identify candidate genes involved in CPP biosynthesis. Finally, to verify the quality of the dataset as well as to characterize the biosynthesis of CPP, identified classical transcripts with respect to different branches of CPPs biosynthesis were cloned and their abundance quantified by real time PCR. Our results provide insight into the biosynthesis of polysaccharides in *C. pilosula* and provide new avenues for sustainable production of bioactive natural products such as CPPs.
Materials and Methods

Ethics statement

The study was carried out on a private land and Guo Feng Zhao should be contacted for future permissions. The location is not protected in any way, and the field studies did not involve endangered or protected species.

Plant material

Two-year-old *C. pilosula* planted in a field, where good agricultural practices were implemented, was harvested and used as plant material in 2012 and 2013. The field was located in Lingchuan, Shanxi, China (1,522 m altitude, 35°47'N, 113°24'E).

Previous studies have shown that the level of CPPs in root tissue of *C. pilosula* increases steadily during the growing season [25, 26]. At flowering and boll-forming stage (Fig. 1B), the content of CPPs was the highest in roots, and the content in flower buds was relatively higher than in other aerial parts (S1 Fig.). Thus, to comprehensively generate the *C. pilosula* transcriptome, flower buds, stems, leaves, as well as roots at this stage (Fig. 1B) were collected in the midafternoon of July 18, 2012. The other root materials were sampled in the midafternoon of June 18, August 8, and September 5 in 2012. At these sampling points, the plants were at the following growth stage of the year: tendril-pulling stage, flowering stage, seed maturity stage. Morphological characterization of the plants at each of these stages was presented in Fig. 1. After collection, the specimens were first washed with tap water, dried using filter paper, and then quickly cut into small pieces on ice before being snap-frozen in liquid nitrogen. The samples were stored at -70°C for library preparation and RNA-sequencing (RNA-seq).

For real time PCR analysis, leaves, stems, flower buds, and roots were collected in the midafternoon of July 27, 2013, at the flowering and boll-forming stage of plants. They were sampled, washed, and stored using the same method as that described above.

Library preparation and sequencing (mRNA-seq)

Total RNA was isolated from each sample using a TRIzon kit (CWBIO, Beijing, China) following the manufacturer’s instructions. DNA contamination was removed using a Qiagen RNase free DNase I (Qiagen, Hilden, Germany). The integrity of isolated RNA was verified by agarose gel electrophoresis. RNA was quantified using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., Wilmington, USA). Equal quantities of high-quality RNA from each sample were pooled and used for cDNA synthesis.

The cDNA library was constructed using a mRNA-Seq Sample Preparation Kit (Illumina Inc., San Diego, USA) according to the manufacturer’s recommendations. Briefly, mRNA was purified from the total RNA by using mRNA-Seq Sample Preparation Kit. The purified mRNA was then fragmented into small pieces by using fragmentation buffer and the first strand cDNA was synthesized from fragmented mRNA by using random hexamers as primers. After the synthesis of the second strand cDNA using DNA polymerase I (Invitrogen, Carlsbad, CA, USA) and RNase H (New England Biolabs Inc., Ipswich, MA, USA), a 3'-adenine was added to the terminus for further ligation of Solexa adapters. Ligation products were separated on a 2% TAE agarose gel. A specific region (200 ± 25 bp) of the gel was excised and the cDNA contained in the excised gel was purified using a gel extraction kit (Qiagen). These fragments were used for 12 rounds of PCR amplification with primers complementary to the previously ligated adaptors. The final mRNA-seq libraries were constructed after the products were purified using a QIAquick PCR kit (Qiagen). The mRNA-seq libraries were sequenced on an Illumina Genome HiSeq 2000 platform (BioMarker Technologies Co., Ltd, Beijing, China).
De novo assembly and functional annotation

Clean reads were obtained by removing the adapter and low quality sequences as well as sequences that were less than 60 bases in length. De novo assembly was carried out using Trinity [27]. The K-mer value was set at 25 while other parameters were set at default levels. Redundancy removal was further performed with sequence clustering software TGICL to acquire non-redundant unigenes [28]. The unigenes obtained after gene family clustering were composed of clusters (corresponding prefix was CL), which contained several unigenes and singletons (the prefix was Unigene). Similarity between unigenes within one cluster was higher than 70%.

The unigenes thus obtained were aligned (online) to the following databases: GenBank non-redundant (NR, http://www.ncbi.nlm.nih.gov), Swiss-Prot (http://web.expasy.org/docs/relnotes/relstat.html), COG (http://www.ncbi.nlm.nih.gov), GO (http://www.geneontology.org/), and KEGG (http://www.genome.jp/kegg/) [29–31]. A cut-off value of E < 10^{-5} was set.

Fig 1. Plant material. Morphological characters of the aerial parts of 2-year-old plants at representative growth stages. (A) Tendril-pulling stage, stem elongated; (B) Flowering and boll-forming stage, stem extended further and the number of stalks and leaves went up sharply, flower buds become visible in appearance until petals extend; (C) Flowering stage, the number of flower buds increased and flowers open; (D) Seed maturity stage, seeds begin to be mature, and capsule hemispheric are at base, conical turns toward apex.

doi:10.1371/journal.pone.0117342.g001
for all BLASTx searches. Based on BLASTx hits against the NR database, BLAST2GO program was carried out to obtain the GO annotations [32]. WEGO was used to obtain GO classifications according to molecular function, biological process, and cellular component [33].

**Prediction of the protein-coding region**

For predicting the protein-coding region, the unigenes were first aligned (E-value <10^{-5}) to protein databases in the priority order of NR, Swiss-Prot, KEGG, and COG. The unigenes aligned to a database of higher priority were no longer aligned to a lower priority database. Proteins in the blast results with the highest ranks were selected and the coding sequences (CDSs) as well as the amino acid sequences were determined using standard codons as references. Unigenes that did not align to any available database were scanned using ESTScan, which produced nucleotide and amino acid sequences of the predicted coding region [34].

**Analysis of the genes involved in the biosynthesis of CPPs**

The analysis of genes involved in the biosynthesis of CPPs depended on the KEGG pathway annotation dataset. Candidate genes involved in CPPs biosynthesis were identified in KEGG database based on the metabolic pathways of starch, glucose, amino sugar, and nucleotide sugar. Neighbor joining (NJ) phylogenetic analysis of UGPase (uridine diphosphate glucose pyrophosphorylase) was performed with MEGA5[35]. All potential transcripts annotated as GTs (glycosyltransferases) [EC: 2.4.x.y] were extracted with the help of Perl script based on the Swiss-Prot annotation and were classified by BLASTx searches against local Arabidopsis thaliana GTs database using a threshold of 1.0^{-5}, which encompasses all genes encoding GTs in A. thaliana (http://www.cazy.org/e1.html).

**Real-time PCR**

Nine genes potentially involved in the biosynthesis of CPPs were selected for further analysis. Specific primers were designed using Primer Premier 3.0 [36]. The details regarding the genes and primers are presented in Table 1. For expression analysis, total RNA was extracted from the leaf, stem, flower bud, and root tissues at the same growth stage in 2013 using the protocol described earlier. The RNA was treated with DNase I to remove contaminating DNA (TaKaRa, Dalian, China). Following this, total RNA was reverse transcribed into cDNA by using a Prime-Script RT Master Mix (TaKaRa). All primer pairs were examined using standard real-time PCR and Premix Ex Taq (TaKaRa), and the presence of a single amplification product of the expected size for each gene was verified by electrophoresis on a 1.5% agarose gel followed by ethidium bromide staining.

Real-time PCR was performed using a SYBR Premix ExTaq Kit (TaKaRa) on an Applied Biosystems 7300 Real-Time PCR system (Applied Biosystems, Foster City, CA, USA). The total volume of the reaction mixture was 20 μL, and it contained 2 μL cDNA, 10 μL SYBR Premix Ex Taq, 0.4 μL ROX Reference Dye II, and each primer at a concentration of 4 pmol. PCR conditions were as follows: 95°C for 30 s, followed by 40 cycles of 95°C for 5 s and 60°C for 31 s. The relative expression levels of genes were calculated by 2^{-ΔΔCt} method [37, 38]. Biological samples from three independent experiments were analyzed and each reaction was run in triplicates. Raw data on the relative abundance of each transcript were expressed as mean ± standard deviation (SD).
Results

Illumina sequencing, de novo assembly, and assessment of assembly program

At the flowering and boll-forming stage, reproduction and vegetative growth concur in the species, and flower buds become gradually visible and the petals become extended (Fig. 1B). At this stage, concentrations of CPPs in the roots and flower buds are rather higher than in the other aerial parts. To cover the C. pilosula transcriptome comprehensively, total RNA was isolated from the four tissues collected at this stage as well as from the root tissue collected at other growth stages. Equal amounts of total RNA from these samples were collected and pooled. The mRNA was purified, fragmented, and reverse-transcribed into cDNA. The cDNA was subjected to Illumina HiSeq 2000 paired-end sequencing. After the removal of adaptor sequences, ambiguous reads, and low quality reads, 91,175,044 clean reads 101 bp in length, were obtained. All HiSeq 2000 sequencing data were combined and assembled using Trinity, which produced 102,125 contigs with an N50 of 999 bp (i.e., 50% of the assembled bases were incorporated into contigs of 999 bp or longer). Then, TGICL was used to reduce the redundancy. Finally, 45,511 unigenes were generated, which included 10,563 clusters and 34,948 singletons (Table 2). The size distribution of the unigenes is shown in Fig. 2. The size distribution was homogeneous and 25.3% of unigenes were longer than 1000 bp. The longest unigene recovered in this work was Unigene CL5433.Contig2, the length of which was 9,150 bp and it encoded alpha, alpha-trehalose phosphate synthase (according to Swiss-Port annotation result). All reads were deposited in the National Center for Biotechnology Information (NCBI) and can be accessed in the Short Read Archive (SRA) Sequence Database under project accession number: SRP038022.

Functional annotation and analysis of the predominant transcripts

All C. pilosula unigenes generated by HiSeq 2000 sequencing were aligned to public protein databases (NR, COG, Swiss-Prot, and KEGG) and nucleotide database (Nt) by BLASTx and BLASTn, respectively (with cut-off E-values of 10^-5). In total, 34,616 (76.1%) unigenes were aligned to homologous sequences in public databases (Table 3). To understand the expression

Table 1. Codonopsis pilosula candidate genes and primers used in real-time PCR.

| Gene ID     | Candidate gene | Upstream primer (5'-3') | Tm(°C) | Downstream Primer (5'-3') | Tm(°C) | Length(bp) |
|-------------|----------------|-------------------------|--------|---------------------------|--------|------------|
| Unigene12384| GAPDH          | tgctctgtgaatcatacttc    | 58.1   | ctaaactgtctgctttcc        | 61.9   | 164        |
| Unigene186  | manA           | ttctgctgtataattccttcc   | 56.9   | acatatgtggtatgcttt        | 57.4   | 162        |
| Unigene11590| manB           | aacgcctccaggtacaaac     | 59.4   | gcactcccaagcagacgga       | 60.9   | 86         |
| Unigene16117| UGPase         | ttactctccagagacgca      | 58.6   | tctgatgctgatgtgaccc       | 60.1   | 192        |
| Unigene7955 | RHM            | ctaactctgtctgatggtt     | 53.2   | gaatcccaatcagaccc         | 55.4   | 111        |
| Unigene12223| UER            | ccaactttggttaactcttt    | 58.3   | tattcctgcggctcttt         | 59.8   | 131        |
| CL5269.Contig2| UGDH         | gtgcttataaggggagcgaa    | 61.6   | gagggtttcataatggtagacat   | 63.2   | 193        |
| Unigene9227 | UXE            | tgtggcagcaagaaagggg     | 62.6   | cggacgagaaagaaatca        | 60.6   | 97         |
| Unigene15345| UGlcAE         | cagcggtttatcttgcag      | 55.7   | ctcctctttaccgcctga        | 57.3   | 95         |
| Unigene14947| AXS            | gataaagggcatgacgata     | 55.7   | aagagcgcttaagaggtt        | 58.8   | 146        |

Abbreviations: GAPDH—glyceraldehyde-3-phosphate dehydrogenase, manA—mannose-6-phosphate isomerase, manB—phosphomannomutase, UGPase—uridine diphosphate glucose pyrophosphorylase, RHM—UDP-glucose 4,6-dehydratase, UER—3,5-epimerase/4-reductase, UGDH—UDP-glucose 6-dehydrogenase, UXE—UDP-arabinose 4-epimerase, UGlcAE—UDP-glucuronate 4-epimerase, AXS—UDP-apiose/xylose synthase.

doi:10.1371/journal.pone.0117342.t001
profile of the assembled sequences, we aligned the high-quality clean reads to the assembled C. pilosula transcripts with the help of Bowtie program, allowing one mismatch. Following the alignments, the count of raw reads for each unigene was normalized to fragments per kilobase of exon model per million mapped fragments (FPKM) [39]. We also mapped the reads to the assembled transcriptome; 92.51% of the reads could be mapped to the C. pilosula library. The data on the mapped efficiency are presented in S1 Table.

NR annotation
Approximately 74.2% (33,748) of the unigenes were annotated with reference to the NR database. Based on the NR annotations, 30.3% of the annotated sequences had the highest homology (E-value < $10^{-100}$) and 15% had moderate homology (10$^{-100}$ < E-value < $10^{-60}$) (Fig. 3A). With respect to similarity distribution, 24.9% of the hits had higher similarity than 80%, and 46.8% had moderate similarities of approximately 60–80% (Fig. 3B). With respect to species, 38.4% of the unigenes had top matches to the sequences from Vitis vinifera, 15.8% to the sequences from Solanum lycopersicum, and 9.6% to those from Prunus persica (Fig. 3C).

GO categories
Among the 33,748 most significant BLASTx hits against the NR database, 26,189 unigenes could be assigned to one or more terms, which were categorized into 55 functional groups as shown in Fig. 4. Among the unigenes that were categorized in “the biological process,” 16,286 were related to the “cellular process” and represented the largest proportion, followed by “metabolic process” (15,732), and “single-organism process” (11,743). Within the cellular components, the assignments were mostly enriched in the “cell part” (19,471) and “cell” (19,472). The GO terms for molecular function were mainly grouped into “catalytic activity” (13,164) and “binding” (12,307).

COG annotation
Assignment of COG was used to further evaluate the completeness of C. pilosula transcriptome library and the reliability of the annotation process. Overall, 11,415 (25.1%) unigenes were
assigned to 25 COG categories (Fig. 5). Among these groups, unigenes belonging to the “general function prediction” occupied the largest part (3,694, 32.6%), followed by “transcription” (2,047, 17.9%), “replication, recombination, and repair” (1,857, 13.8%). Additionally, 1,354 (11.7%) unigenes were assigned to “carbohydrate transport and metabolism,” and 647 (5.7%) unigenes to “signal transduction mechanisms.”

### Table 3. Summary statistics of functional annotations for *Codonopsis pilosula* unigenes in public databases.

| Annotated databases | Number of Unigenes | Percentage (%) |
|---------------------|--------------------|----------------|
| NR                  | 33,748             | 74.2           |
| Nt                  | 27,319             | 60.0           |
| GO                  | 26,189             | 57.5           |
| COG                 | 11,415             | 25.1           |
| KEGG                | 18,848             | 41.4           |
| Swiss-Prot          | 21,265             | 46.7           |
| Total               | 34,616             | 76.1           |

doi:10.1371/journal.pone.0117342.t003
were assigned to "secondary metabolites biosynthesis, transport, and catabolism." Unigenes assigned to "extracellular structures" and "nuclear structure" constituted the smallest category.

**KEGG pathway mapping**

The unigenes annotated to the KEGG database were mapped to further understand the biological processes in *C. pilosula*. Totally, 18,848 unigenes were assigned to 128 KEGG pathways (see S2 Table). Among these, 4,191 (22.24%) were mapped to metabolic pathways and 2,076 (11.01%) were mapped to biosynthesis of secondary metabolites (Fig. 6). Moderate numbers of unigenes were mapped to the following pathways: "plant-pathogen interaction" (1,096, 5.81%), "plant hormone signal transduction" (986, 5.23%), "RNA transport" (747, 3.96%), "spliceosome" (601, 3.19%), "starch and sucrose metabolism" (582, 3.09%), "glycerophospholipid metabolism" (506, 2.68%), and "endocytosis" (501, 2.66%). Only four unigenes were distributed to "betalain biosynthesis" and two unigenes were distributed to "Caffeine metabolism" in the library.

**Analysis of highly expressed transcripts**

The value of FPKM represents the abundance of a certain unigene in a certain gene pool, and is often used to indicate the level of gene expression in various biological samples. The top 20...
relatively highly abundant genes in the dataset are listed in Table 4. Transcripts encoding chlorophyll a-b binding protein 40, ribulose-1,5-bisphosphate carboxylase small subunit, and chloroplast photosystem II 10 kDa polypeptide were abundant in the transcriptome library, indicating that the photosynthetic process was active in the tissues investigated. Metallothionein is responsible for the detoxification of heavy metals, metal ion transport, maintenance of dynamic balance, ion metabolism, and reactive oxygen removal [40]. Two transcripts (Unigene8032_Codonopsis and Unigene12118_Codonopsis) encoding metallothionein were also found abundantly in the tissues. Additionally, unigenes encoding aquaporin, cysteine protease, chloroplastic thiamine thiazole synthase, and dormancy-associated protein were also abundant in the cDNA library. The remaining unigenes with higher FPKM values could not be annotated successfully, indicating their unidentified roles in *C. pilosula*.

**Predicted coding sequences**

Based on three public protein databases, 36,146 CDSs were obtained, which accounted for 79.42% of the total number of unigenes. Among these, 33,804 CDSs were predicted by BLASTx and 2,342 by ESTScan, indicating relatively higher identities of the predicted CDSs to those available in the public databases. The length distribution of CDSs is shown in Fig. 7. A total of 27,562 CDSs were between 200 bp and 1000 bp in length, 68.25% of which was longer than 500 bp, thus constituting a relatively higher proportion of the predicted CDSs. The number of CDSs longer than 2000 bp was 1,563, which was approximately 18.21% of the CDSs longer than 1000 bp, indicating a relatively higher ratio of CDSs with larger size. The CDSs predicted by ESTScan tool were primarily 200 bp to 1000 bp in length.

**Functional annotation and candidate genes involved in CPP biosynthesis**

To understand the biosynthesis of CPPs and to identify the genes involved, the transcripts related to starch and sucrose metabolism (ko00500) as well as amino sugar and nucleotide sugar
Starch and sucrose metabolism

Among 128 metabolic pathways detected in *C. pilosula*, candidate genes related to starch and sucrose metabolic pathways were ranked seventh as far as the number of involved genes was concerned (S3 Table). To understand the biosynthesis of CPPs, the unigenes involved in the biosynthesis of starch and other related polymers as well as those related to carbon assimilation in the leaf tissues were annotated. The key enzymes involved in these metabolic processes were identified and annotated based on KEGG (see Table 5 and S3 Table).
In plants, UGPase catalyzes the formation of uridine diphosphate glucose (UDP-Glu), a key precursor of NDP-sugars formed from 1-P-glucose (Glu-1-P) \[45, 46\]. In *Astragalus membranaceus*, the enzyme activity is positively correlated with the concentration of Astragalus polysaccharides, indicating that UGPase plays an important role in polysaccharide biosynthesis \[16\]. In *Arabidopsis thaliana*, UGPase is a rate-limiting factor and contributes primarily to UDP-Glu metabolism in the vegetative phase \[45\]. In *C. pilosula* library, Unigene16117_Codonopsis (KJ470627) was identified as a homologue of UGPase. This 2087-bp unigene contained a complete open reading frame (ORF) and showed the highest amino acid identity (88%) to UGPase from *V. vinifera* (Accession number: XP_002282276) (S2 Fig.). In *A. membranaceus*, granule-bound starch synthase (GBSS) is a key enzyme required for the synthesis of polysaccharides \[17\]. A 2308-bp Unigene7794_Codonopsis (KJ470628) was also identified as a homologue of GBSS and contained an intact ORF encoding 622 amino acids.

**Amino sugar and nucleotide sugar metabolism**

Previous studies indicated that CPPs are formed by the polymerization of NDP-sugars such as UDP-glucose (UDP-Glu), UDP-glucuronic acid (UDP-GluA), UDP-arabinose (UDP-Ara), UDP-rhamnose (UDP-Rha), GDP-mannose (GDP-Man), UDP-galactose (UDP-Gla), UDP-xylose (UDP-Xyl), and UDP-galacturonic acid (UDP-GlaA) \[12–15\]. Among these NDP-
sugars, UDP-Glu and GDP-Man are primarily derived from glucose-6-phosphate (Glu-6-P) and fructose-6-phosphate (Fru-6-P), whereas others are derived from UDP-Glu and GDP-Man through the actions of NDP-sugar interconversion enzymes (NSEs) such as isomerase, decarboxylase, and dehydrogenase \[47\]. Six subclasses of NSEs were identified in \textit{C. pilosula}, including 4-epimerase, 3,5-epimerase, 3,5-epimer-4-reductase, 4,6-dehydrogenase, and 6-dehydrogenase and decarboxylase.

Thirty-seven transcripts encoding NSEs were identified in \textit{C. pilosula} (Table 6, S4 Table). A comparison of FPKM values revealed that in the transcriptome library, the transcript of \textit{UGl-cAE} (UDP-glucuronate 4-epimerase) was the most abundant, followed by those of \textit{UGDH} (UDP-glucose 6-dehydrogenase) and \textit{AXS} (UDP-apiose/xylose synthase). The enzyme encoded by \textit{UGl-cAE} is responsible for the production of UDP-GalA, using UDP-GluA as a precursor. The enzyme encoded by \textit{UGDH} converts UDP-Glu into UDP-GluA, thus taking part in the biosynthesis of glycosaminoglycans. AXS converts UDP-GluA to UDP-Xyl. Relatively higher abundance of such transcripts suggested that the biological process of conversion from UDP-GluA to UDP-Xyl should be very active.

### CPP biosynthetic pathway

Based on the annotation results of the abovementioned pathway and documents published \[41\], biosynthetic pathway for the formation of CPPs from sucrose was cautiously constructed and was shown in Fig. 8. In this scheme, synthesis of CPPs is divided into three main steps. The first step is the formation of UDP-Glu and GDP-Man from Glu-6-P and Fru-6-P, which are then converted in the second step into other NDP-sugars. Finally, the active monosaccharide units, NDP-sugars, are added to the sugar residues of various polysaccharides and glycoconjugates by the action of various GTs \[48\].

### Table 4. Predominant unigenes in \textit{Codonopsis pilosula} transcriptome library.

| Gene ID          | FPKM      | NR-ID       | NR-annotation                                      |
|------------------|-----------|-------------|----------------------------------------------------|
| CL3472.Contig1   | 41524.464 | ref|XP_003614394.1| Hypothetical protein                               |
| Unigene9157_Codonopsis | 33994.62 | ref|XP_003614382.1| Hypothetical protein                               |
| CL595.Contig1_Codonopsis | 6054.2321 | ref|XP_003614391.1| Tar1p                                              |
| CL655.Contig2_Codonopsis | 4182.2431 | ref|XP_002285299.1| RD21a-like cysteine proteinase                      |
| Unigene8032_Codonopsis | 3955.4621 | dbj|BAD18924.1 | Metallothionein 2                                  |
| Unigene14155_Codonopsis | 3759.3688 | ref|XP_004170751.1| Uncharacterized protein                            |
| Unigene17719_Codonopsis | 2440.7739 | ref|XP_002510636.1| Phosphoprotein ECPP44                              |
| CL40.Contig5_Codonopsis | 2340.6985 | sp|P27495.1|CB24_TOBAC Chlorophyll a-b binding protein 40     |
| Unigene9176_Codonopsis | 2271.6946 | gb|AAA33866.1 | Ribulose-1,5-bisphosphate carboxylase small subunit|
| CL2635.Contig3_Codonopsis | 1994.5632 | gb|ACB38232.1 | Aquaporin                                          |
| Unigene12118_Codonopsis | 1818.9061 | gb|AAO92264.1 | Metallothionein-like protein                       |
| Unigene3690_Codonopsis | 1786.4363 | gb|ADB93062.1 | Chloroplast photosystem II 10 kDa polypeptide     |
| CL655.Contig1_Codonopsis | 1759.3512 | gb|AAW34315.1 | Cysteine protease gp2b                             |
| Unigene3692_Codonopsis | 1680.0371 | gb|EMJ23383.1 | Hypothetical protein                               |
| Unigene3701_Codonopsis | 1610.1937 | ref|XP_002301827.1| Predicted protein                                  |
| CL655.Contig3_Codonopsis | 1584.7255 | sp|P60994.1|ERV_B_TABDI Ervatamin-B                             |
| Unigene15471_Codonopsis | 1543.9714 | sp|Q023787.1|THI4_CITSI Thiamine thiazole synthase, chloroplastic|
| Unigene12117_Codonopsis | 1407.0655 | ref|XP_002513211.1| Lipid binding protein                             |
| Unigene1975_Codonopsis | 1370.3016 | gb|AAW02792.1 | Dormancy-associated protein                        |
In *C. pilosula* library, 723 transcripts encoding GTs were identified based on the annotations in *Arabidopsis thaliana*, including 591 GTs and 132 GTNCs (referenced for glycosyltransferases with no classification) (see S5 Table). Among these GTs, the transcripts of family 1 glycosyltransferase (GT1s), often referred to as UDP glycosyltransferase (UGTs), were the most abundant, indicating that GT1s might be encoded by comparably large multigene families. The transcripts of GT2, GT8, GT4, and GT48 were more abundant than that of others, and GT31, GT20, GT41, GT5, and GT90 showed moderate abundance, whereas the remaining had only limited number of transcripts (less than ten in number). An imbalance in the number of transcripts was also observed in annotated UGTs. Since these genes have specific functions in the biosynthesis of CPPs, the differing abundances are indicative of a preference for the synthesis of specific NDP-sugars.

In addition to polysaccharides, UGTs are also responsible for the transfer of sugar moieties from UDP sugars to other plant natural products, such as phenolic compounds, cyanins,
steroids, flavonoids, and terpenes [49]. In general, approximately 48% of UGTs contain a common motif known as plant secondary product glycosyltransferase (PSPG) box [50]. As key sugar donors, NDP-sugars are primarily attached to PSPG motif of UGTs. Therefore, the PSPG motif is very important for UGTs. This motif contains 44 conserved amino acids. The amino acids at C-terminus are glutamine (Q) and histidine (H), which are highly conserved in glycosyltransferase and galactosyltransferase respectively [50]. An analysis of the amino acid sequences of PSPG in the annotated UGTs of *C. pilosula* showed that the terminal amino acid residue of all annotated UGTs was Q except in unigene2366, whose terminal amino acid residue was H, indicating an obvious imbalance in the transfer of NDP-sugars.

Amino acid sequences after multiple sequence alignments reflect their PSPG structures and their functions. Therefore, amino acid sequence analysis of this motif will provide insights into their functional diversities in CPPs biosynthesis. Conservative analysis on the amino acid sequences of fifty UGTs possessing a typical PSPG sequence was performed using WebLogo

| Enzyme name                                      | EC            | Number of Unigenes |
|------------------------------------------------|---------------|--------------------|
| Glucose-6-phosphate isomerase                    | 5.3.1.9       | 2                  |
| Phosphoglomutase                                | 5.4.2.2       | 2                  |
| Uridine diphosphate glucose pyrophosphorylase    | 2.7.7.9       | 1                  |
| Cellulose synthase                              | 2.4.1.12      | 55                 |
| Sucrose synthase                                | 2.4.1.13      | 15                 |
| Sucrose-phosphate synthase                      | 2.4.1.14      | 16                 |
| Glucose-1-phosphate adenylyltransferase          | 2.7.7.27      | 8                  |
| Starch synthase                                 | 2.4.1.21      | 13                 |
| Starch branching enzyme                         | 2.4.1.18      | 13                 |
| Granule-bound starch synthase                   | 2.4.1.242     | 1                  |
| Invertase                                       | 3.2.1.26      | 24                 |
| Hexokinase                                      | 2.7.1.1       | 7                  |
| Fructokinase                                    | 2.7.1.4       | 9                  |

Table 6. Number of unigenes annotated as NDP-sugar interconversion enzymes (NSEs).

| Enzyme code | Enzyme name                                      | Abbreviation | Number | FPKM    |
|-------------|--------------------------------------------------|--------------|--------|---------|
| 5.3.1.8     | mannose-6-phosphate isomerase                    | manA         | 2      | 19.3816 |
| 5.3.1.9     | glucose-6-phosphate isomerase                     | GPI          | 2      | 158.1845|
| 5.4.2.8     | Phosphomannomutase                               | manB         | 5      | 111.3943|
| 2.7.7.13    | mannose-1-phosphate guanylyltransferase          | GMPP         | 7      | 86.7738 |
| 5.4.2.2     | Phosphoglomutase                                 | pgm          | 2      | 143.3946|
| 2.7.7.9     | Uridine diphosphate glucose pyrophosphorylase     | UGPase       | 1      | 242.5288|
| 4.2.1.76    | UDP-glucose 4,6-dehydratase                       | RHIM         | 4      | 222.6735|
| 5.1.3.-1.1.1.1- | 3,5-epimerase/4-reductase                      | UER          | 1      | 110.7206|
| 5.1.3.6     | UDP-glucuronate 4-epimerase                       | UGlcAE       | 4      | 488.2218|
|             | UDP-apiose/xylose synthase                       | AXS          | 1      | 343.0033|
| 5.1.3.5     | UDP-arabinose 4-epimerase                        | UXE          | 1      | 48.15   |
| 5.1.3.2     | UDP-glucose 4-epimerase                           | UGE          | 2      | 125.0329|
| 5.1.3.18    | GDP-mannose 3,5-epimerase                         | GME          | 1      | 18.1922 |
| 1.1.1.22    | UDP-glucose 6-dehydrogenase                      | UGDH         | 4      | 400.1352|
The font size represented the conservative properties of amino acid residues, the bigger the font the more conservative amino acid residue. It indicates that the first and twenty-second tryptophan (W), fourth and forty-fourth glutamine (Q), tenth and nineteenth histidine (H), twenty-first glycine (G), twenty-fourth serine (S), twenty-seventh glutamic acid (E), and thirty-fourth and thirty-ninth proline (P) residues at N-terminus are more conserved than other amino acids in the motif.

Studies on the structure-activity relationship of UGT have demonstrated that the nucleotide sugar donor primarily interacts with the C-terminal domains and that the acceptor mainly binds to the N-terminal domains [50]. Since diverse natural products act as acceptors of UGTs, further studies are needed to understand whether specific UGTs are involved in CPPs biosynthesis.

Fig. 8. Proposed pathways for polysaccharide biosynthesis in *Codonopsis pilosula*. Activated monosaccharide units were marked in red. Words in bold are key intermediates. The first step of CPPs biosynthesis was shown with blue arrows. Conversion of monosaccharides were marked with red arrows. The key enzymes were highlighted in purple, and the corresponding FPKM values were indicated in brackets. Abbreviations: SS: sucrose synthase; HK: fructokinase/hexokinase; UGPase: uridine diphosphate glucose pyrophosphorylase; RHM: UDP-glucose 6-dehydrogenase; UGE: UDP-glucose 4-epimerase; UGlcAE: UDP-glucuronate 4-epimerase; AXS: UDP-apiose/xylose synthase; UXE: UDP-arabinose 4-epimerase; UER: 3, 5-epimerase/4-reductase; PGM: phosphoglucomutase; GPI: glucose-6-phosphate isomerase; manA: mannose-6-phosphate isomerase; manB: phosphomannomutase; GMPP: mannose-1-phosphate guanylyltransferase; GTs: glycosyltransferases.

doi:10.1371/journal.pone.0117342.g008
Real-time PCR analysis

Analysis of the abundance of transcripts of enzymes related to CPP biosynthesis provides evidence at two levels. On one hand, the detection of PCR products verifies if the sequences obtained by high throughput screening are reliable and accurate. On the other hand, it provides insights into their functions in different branches of CPPs biosynthesis. Based on the preliminary results of the selection of reference genes in *C. pilosula* (see S6 Table), GAPDH was used as internal reference in the present study. Nine candidate unigenes were selected based on the alignment results from online BLAST for verification. Among these candidates, manA (Unigene186) had the lowest identity to its homolog in the dataset (66% to 67%). Relatively lower identity of UXE (Unigene 9227) to its homolog was also obtained, but the value was larger than 79% (*V. vinifera*, XP_002264946). For following candidates: manB (Unigene11590), UGPase (Unigene 16117), RHM (Unigene 7955), UER (Unigene 12223), and UGlcAE (Unigene 15345), their identities to the corresponding homolog ranged from 81% to 93%. UGDH (CL5269.Contig 2) and AXS (Unigene 14947) shared rather higher identities (equal or larger than 90%) to their homologs available in the public dataset. The detailed results of the alignment were listed in the S7 Table. The analysis primarily confirmed that the selected potential unigenes are rather conservative and should be responsible for the synthesis of CPPs.

Among these candidate unigenes, the ones encoding UXE (UDP-arabinose 4-epimerase), UER (3,5-epimerase/4-reductase), manA (mannose-6-phosphate isomerase), and manB (phosphomannomutase) should function downstream of the predicted pathway of CPPs biosynthesis (see Fig. 8), whereas the ones encoding UGDH (UDP-glucose 6-dehydrogenase), RHM (UDP-glucose 4,6-dehydratase), and UGPase should act upstream. The predicted pathway also indicated that the unigenes responsible for UGlcAE (UDP-glucuronate 4-epimerase) and AXS (UDP-apiose/xylose synthase) should possess the same precursor and work at two branch points. Real-time PCR was carried out using equally mixed total RNA extracted from all tissues collected at flowering and boll-forming stage as a template for transcription, and PCR products were detected by 1.5% agarose gel electrophoresis. The result was presented in S3 Fig. All unigenes investigated were successfully amplified, confirming the reliability and high accuracy of the used deep sequencing technique.

The results of the relative expression of the nine candidate genes in the four tissues are presented in Fig. 10. Among these, the expression of ManA was relatively stable and there was no
significant difference among the tissues. A similar expression pattern of \textit{UGPase} was also observed. Relatively lower expressions in leaves and relatively higher expressions in stem and root tissues were observed for \textit{manB}, \textit{RHM}, \textit{UER}, \textit{UGDH}, \textit{UGlceAE}, and \textit{AXS}, indicating that the expression of these candidate genes were tissue-dependent. Comparatively, the expression of \textit{UXE} in roots was much higher than the expression in other tissues, indicating an important role the gene might play in the biosynthesis of CPPs in the species.

Besides real-time PCR, another dataset on the abundance of the transcripts of these unigenes was obtained by RNA-seq (S8 Table). A comparison between the two datasets was performed and was presented in Fig. 11. In the figure, the data on relative abundance of the transcripts obtained by real-time PCR were presented as a column graph, while the FPKM value of each transcript was indicated with a broken line. It revealed that there was some disagreement in the two datasets of each unigene even in the same tissue excluding \textit{UGPase}. For example, the data from the two datasets on \textit{manA} in roots, and \textit{RHM} and \textit{UGlceAE} in stems were not in accordance, indicating bias in the datasets obtained by the two methods. In general, similar patterns of gene expression in the same tissue and at the same growth stage should be observed by both techniques. In the present work, two sets of samples were collected from
plants at the same growth stage but not in the same year (2012 and 2013). Inconsistent expression patterns might be caused by different environmental conditions or stress factors. The comparison between the two datasets also demonstrated that the expression of \textit{UGPase} is relatively stable, suggesting conservation and vital role of the enzyme in the process of CPPs biosynthesis. Due to the higher sensitivity of Illumina sequencing compared to real-time PCR, further investigation is needed to understand their functions at a molecular level [51].

**Discussion**

To our knowledge, this is the first \textit{de novo} sequencing of samples from \textit{C. pilosula} using Illumina HiSeq 2000 platform. The sequencing produced 91,175,044 clean reads and 45,511 unigenes after assembly and made it possible to understand the biosynthesis of CPPs at molecular level. Compared to datasets published on medicinal plants, the number of unigenes obtained from \textit{C. pilosula} was lower than that obtained from \textit{Polygonum cuspidatum} (86,418), \textit{Dimocarpus longan} (68,925), and \textit{Hypericum perforatum} (59,184), but was higher than that from \textit{Siraitia grosvenorii} (43,891) [21, 24, 52, 53]. However, the average length of unigenes obtained here was 728 bp, which was longer than that from other medicinal plants, including \textit{P. cuspidatum}.
(365 bp), S. grosvenorii (668 bp), D. longan Lour. (448 bp), and H. perforatum (422 bp). A relatively higher annotation rate (76.1%) was also observed in the present work. Notably, the number of unigenes longer than 1000-bp was 8,584, accounting for 23.75% of all unigenes, indicating higher quality of the transcriptome library and the reliability of Illumina Hiseq-2000 technology for global screening of functional genes in non-model plants, especially medicinal plants.

Using bioinformatics tools, we identified unigenes involved in CPPs biosynthesis. Annotation results suggested that sucrose is used as the primary starting material for CPPs biosynthesis. Further analysis showed relatively higher abundance of CPPs biosynthesis-related transcripts in the root and stem tissues, which might be related to the functions of the studied tissues and might be beneficial to the synthesis and accumulation of polysaccharide, the main active component in the plant. The real-time PCR analysis showed that the transcripts of genes related to the conversion of UDP-glucose to UDP-L-arabinose (branch pathway) was more abundant than those of the others, suggesting that this pathway functions as a key branch of CPPs biosynthesis at the flowering and boll-forming stage of C. pilosula. However, a previous study on C. pilosula collected during traditional harvest season showed that the content of arabinose was not the highest and was substantially lower than that of glucose and galacturonic acid [12].

Consistent with a previous study, which found that xylose was an important unit in CPPs biosynthesis and that its molar ratio (15.41) was higher than that of galactose (2.87), glucose (1.92), and rhamnose (12) [54], our results showed higher abundance of AXS transcript in the C. pilosula library. Further studies are needed to identify the diverse molecular entities used for the biosynthesis of CPPs.

UGPase is a key enzyme that functions upstream of the biosynthesis pathway of Astragalus polysaccharides, and is responsible for the conversion of glucose-1-phosphate to UDP-Glu. Although there was a positive correlation between the activity of UGPase and the polysaccharide content in A. membranaceus, higher abundance of the corresponding transcript was not observed in our study [16]. Alignment on UGPase homolog from representative species in plant kingdom showed that Unigene16117 was more similar to UGPase from V. vinifera (Accession number: XM_002282240) than to UGPase from A. membranaceus (AF281081) both at nucleotide and amino acid levels. Phylogenetic analysis showed that Unigene16117 and UGPase from Vitis vinifera are clustered in the same subclass and that the homologues from C. pilosula and A. membranaceus are located relatively far from each other (see S2 Fig.). Therefore, the low abundance of Unigene16117 may be related to the relatively low content of CPPs in the raw materials investigated or may have been caused by divergent function of the protein. However, consistent abundance pattern of the transcript in all tissues examined across two growing seasons confirm the conservation and importance of the enzyme in the biosynthesis of CPPs. Further studies are needed to understand the relationships between enzyme activity, transcript abundance, and the content of CPPs in C. pilosula.

UDP-glucose dehydrogenase (UGDH) catalyzes the oxidation of UDP-glucose (UDP-Glc) to UDP-glucuronate (UDP-GlcA), a key sugar nucleotide involved in the biosynthesis of plant cell wall polysaccharides. Analysis on endogenous expression of a UGDH in Eucalyptus grandis (GenBank accession number: EF179384) showed that it was mainly expressed in roots, stem and bark of 6-month-old saplings, with lower expression level in leaves [55]. In C. pilosula, a relatively higher expression of UGDH-like gene was observed in stem and root tissues, demonstrating similar expression pattern to its homologue from E. grandis. Both works carried out in two distantly related plants also indicated the conservation of UGDH homology’s function in plant kingdom. L-Rhamnose (Rha) is an important constituent of CPPs, which is synthesized
by the enzyme encoded by AtRHM1 in Arabidopsis [56]. Expression studies with AtRHM1 promoter-GUS fusion gene showed that AtRHM1 is expressed almost ubiquitously, with stronger expression in roots and cotyledons of young seedlings and inflorescences. The most abundant transcript of RAH1-like unigene was detected in roots, which was consistent with AtRHM1 in Arabidopsis. Thus, the studies on UGDH and RHM in Arabidopsis confirmed that their homology in C. pilosula should act in a similar way in the biosynthesis of CPPs, and that they should be functional genes involved in the biosynthesis of CPPs. Although UGPase, UGDH, and RHM function in parallel in separated branches of biosynthesis pathway of CPPs, the expression patterns of the latter two were more similar to each other, possessing relatively higher expression in stems and roots, indicating that the transcription of them might be regulated in a similar molecular mechanism. More work should be needed to understand the mechanism and regulation of biosynthesis and accumulation of CPPs at molecular level.

To summarize, an EST collection was created from a transcriptome library prepared from different parts of C. pilosula. The results demonstrated the usefulness of this methodology for the identification of metabolic pathways related to polysaccharide biosynthesis in medicinal plants. Our results provide understanding of the biosynthesis of CPPs at the molecular level in C. pilosula inferred by real-time PCR analysis of candidate unigenes functioning at key separate branches of biosynthesis pathway of CPPs. These findings will allow for an efficient and sustainable production of CPPs and related bioactive natural products.

Supporting Information

S1 Fig. Quantification of polysaccharides in different tissues at the flowering and boll-forming stage in Codonopsis pilosula.

S2 Fig. Neighbor joining (NJ) phylogenetic analysis of UGPase homologs in plants. The phylogenetic analysis was performed using MEGA 5 software. The tree was derived from nine UGPase homologs according to their amino acid sequences.

S3 Fig. Validation of candidate unigenes involved in the biosynthesis of CPP by real-time PCR. Lanes 1 through 9 represent the amplified fragments generated via PCR. Lane 1: manA (162 bp); lane 2: manB (86 bp); lane 3: UGPase (192 bp); lane 4: RHM (111 bp); lane 5: UER (131 bp); lane 6: UGDH (193 bp); lane 7: UXE (97 bp); lane 8: UGlcAE (95 bp); lane 9: AXS (146 bp); lane 10: 2kb plus molecular weight marker.

S1 Table. Statistical analysis of the efficiency of mapped reads.

S2 Table. List of unigene numbers from Codonopsis pilosula assigned to KEGG reference pathway.

S3 Table. Functional annotations of unigenes involved in the biosynthesis of starch and sucrose in Codonopsis pilosula.

S4 Table. Functional annotations of unigenes annotated as NDP-sugar interconversion enzymes (NSEs) in Codonopsis pilosula.
S5 Table. GT classification of *Codonopsis pilosula* against *Arabidopsis thaliana* GT categories.

(XLS)

S6 Table. Primary selection of reliable internal control for gene expression analysis in different tissues in *Codonopsis pilosula*.

(XLS)

S7 Table. Identity analysis of candidate unigenes related to CPP biosynthesis at encoded amino acid level.

(DOC)

S8 Table. FPKM values of the candidate unigenes involved in the biosynthesis of CPPs in four tissues.

(XLSX)

Acknowledgments

The authors are grateful to Prof. Qiu-fen Cao and Dr. Chang-biao Wang from Agri-Biotechnology Research Center of Shanxi Province (Taiyuan, Shanxi) for their assistance with experimental design and bioinformatics analysis, to Prof. Yuan-huai Han and Dr. Si-yu Hou from Shanxi Agricultural University (Jinzhou, Shanxi) for their constructive advice to this manuscript, and to Guo Feng Zhao from Lingchuan Agricultural Comprehensive Development field (Lingchuan, Shanxi) for his assistance in sample collection.

Author Contributions

Conceived and designed the experiments: JPG. Performed the experiments: DW LYC. Analyzed the data: JPG DW LYC. Contributed reagents/materials/analysis tools: JPG DW LYC. Wrote the paper: JPG HFS.

References

1. Nanjing University of Chinese Medicine (2006) Dictionary of Chinese traditional medicine, 2nd ed. Shanghai: Shanghai Scientific and Technical Publishers. 3875 p.
2. Chinese Pharmacopoeia Commission (ed) (2010) Pharmacopoeia of the People's Republic of China. Vol I, China Medical Science Press, Beijing, pp. 264–265.
3. Zhang XD, Gao JP, Cao LY, Zhen X (2013) Survey of Codonopsis germplasm resource and production status. Chinese Archives of Traditional Chinese Medicine 31: 496–498.
4. Yang FD, Li CY (2007) Textual study on Codonopsis Radix documented in ancient Chinese medicinal literatures. Chinese Jouranl of Information on Traditional Chinese Medicine 14: 100–101.
5. Sun YX (2009) Immunological adjuvant effect of a water-soluble polysaccharide, CPP, from the roots of *Codonopsis pilosula* on the immune responses to ovalbumin in mice. Chem Biodivers 6: 890–896. doi: 10.1002/cbdv.200800154 PMID: 19551790
6. Xu C, Liu Y, Yuan G, Guan M (2012) The contribution of side chains to antitumor activity of a polysaccharide from *Codonopsis pilosula*. Int J Biol Macromol 50: 891–894. doi: 10.1016/j.ijbiomac.2012.01.013 PMID: 22285989
7. He K, Li X, Chen X, Ye X, Huang J, et al. (2011) Evaluation of antidiabetic potential of selected traditional Chinese medicines in STZ-induced diabetic mice. J Ethnopharmacol 137: 1135–1142. doi: 10.1016/j.jep.2011.07.033 PMID: 21798327
8. Tsai KH, Lee NH, Chen GY, Hu WS, Tsai CY, et al. (2013) Dung-shen (*Codonopsis pilosula*) attenuated the cardiac-impaired insulin-like growth factor II receptor pathway on myocardial cells. Food Chem 138: 1856–1867. doi: 10.1016/j.foodchem.2012.11.056 PMID: 23411518
9. Zhang XD, Liu L, Tong X (2003) Comparison of Radix Codonopsis and Radix Astragalion central nervous system. Zhong Cao Yao 34: 822–823.
31. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the
biology. The Gene Ontology Consortium. Nat Genet 25: 252–258. doi:10.1016/j.ygeno.2004.02.005 PMID:
15015075
30. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Karyutin B, et al. (2003) The COG database: an up-
dated version includes eukaryotes. BMC Bioinformatics 4: 41. PMID:12969510
29. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unifica-
tion of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. PMID:10802651
28. Kanehisa M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. (2000)KEGG: Kyoto Encyclopedia of Genes
and Genomes. Nucleic Acids Res 28: 27–30. doi:10.1093/nar/28.1.27 PMID:10648165
27. Nan HJ, Qin XM, Wu B, Li Y, Zhang LZ, et al. (2008). Ultrasonic extraction and content determination of
polysaccharide in Codonopsis. Journal of Shanxi Medical University 39: 641–643. doi:10.1016/j.
fitote.2009.08.011 PMID:19686809
26. Ye G, Li C, Huang CG, Li ZX, Wang XL, et al. (2005) Chemical structure of fructosan from Codonopsis
pilosula. Zhongguo Zhong Yao Za Zhi 30: 1338–1340. PMID:16323542
25. Zhang YJ, Zhang LX, Yang JF, Liang ZY (2010) Structure analysis of water-soluble polysaccharide
CPPS3 isolated from Codonopsis pilosula. Fitoterapia 81: 157–161. doi:10.1016/j.fitote.2009.08.011
PMID:19686809
24. Tang Q, Ma X, Mo C, Wilson IW, Song C, et al. (2011) An efficient approach to finding Siraitia grosve-
norii triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. BMC Genomics 12:
343. doi:10.1186/1471-2164-12-343 PMID:21729270
23. Lee EK, Jin YW, Park JH, Yoo YM, Hong SM, et al. (2010) Cultured cambial meristematic cells as a
source of plant natural products. Nat Biotechnol 28: 1213–1217. doi:10.1038/nbt.1693 PMID:
20972422
22. Peng JS, Zhao SJ, Wu XJ, Liu D, Hu ZB, et al. (2000) cDNA cloning and structural analysis of Granule-
bound Starch Synthase gene of hairy roots of Astragalus membranaceus. J Integr Plant Biol 42: 940–945.
21. Xiao M, Zhang Y, Chen X, Lee EJ, Barber CJ, et al. (2013) Transcriptome analysis based on next-gen-
eration sequencing of non-model plants producing specialized metabolites of biotechnological interest.
J Biotechnol 166: 122–134. doi:10.1016/j.jbiotec.2013.04.004 PMID:23602801
20. Wenping H, Yuan Z, Jie S, Lijun Z, Zhezhi W (2011) De novo transcriptome sequencing in Salvia mili-
torrhiza to identify genes involved in the biosynthesis of active ingredients. Genomics 98: 272–279.
doi:10.1016/j.ygeno.2011.03.012 PMID:21473906
19. Nan HJ, Qin XM, Wu B, Li Y, Zhang LZ, et al. (2008). Ultrasonic extraction and content determination of
polysaccharide in Codonopsis. Journal of Shanxi Medical University 39: 641–643. doi:10.1016/j.
fitote.2009.08.011 PMID:19686809
18. Zhang XJ, Zhu CC, Hu L, Lai X, Mo J (2003) Pharmacological action of polysaccharides from Radix
Codonopsis on immune function and hematopoeisis in mice. Traditional Chinese Drug Research and
Clinical Pharmacology 14: 174–176.
17. Ye G, Li C, Huang CG, Li ZX, Wang XL, et al. (2005) Chemical structure of fructosan from Codonopsis
pilosula. Zhongguo Zhong Yao Za Zhi 30: 1338–1340. PMID:16323542
16. Wu XJ, Liu D, Hu ZB (2000) Study on relationship between uridine diphosphate glucose pyrophosphor-
ylase activity and polysaccharide content in hairy root of Astragalus membranaceus. Acta Universitatis
Traditionis Medicalis Sinensia Pharmacologiaeque Shangai 14: 53–55.
15. Ye G, Li C, Huang CG, Li ZX, Wang XL, et al. (2005) Chemical structure of fructosan from Codonopsis
pilosula. Zhongguo Zhong Yao Za Zhi 30: 1338–1340. PMID:16323542
14. Ye G, Li C, Huang CG, Li ZX, Wang XL, et al. (2005) Chemical structure of fructosan from Codonopsis
pilosula. Zhongguo Zhong Yao Za Zhi 30: 1338–1340. PMID:16323542
13. Zhang YJ, Zhang LX, Yang JF, Liang ZY (2010) Structure analysis of water-soluble polysaccharide
CPPS3 isolated from Codonopsis pilosula. Fitoterapia 81: 157–161. doi:10.1016/j.fitote.2009.08.011
PMID:19686809
12. Zhang YJ, Zhang LX, Yang JF, Liang ZY (2010) Structure analysis of water-soluble polysaccharide
CPPS3 isolated from Codonopsis pilosula. Fitoterapia 81: 157–161. doi:10.1016/j.fitote.2009.08.011
PMID:19686809
11. Zhang YJ, Zhang LX, Yang JF, Liang ZY (2010) Structure analysis of water-soluble polysaccharide
CPPS3 isolated from Codonopsis pilosula. Fitoterapia 81: 157–161. doi:10.1016/j.fitote.2009.08.011
PMID:19686809
10. Zhang YJ, Zhang LX, Yang JF, Liang ZY (2010) Structure analysis of water-soluble polysaccharide
CPPS3 isolated from Codonopsis pilosula. Fitoterapia 81: 157–161. doi:10.1016/j.fitote.2009.08.011
32. Conesa A, Gótz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676. PMID: 16081474

33. Ye J, Fang L, Zheng H, Zhang Y, Chen J, et al. (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34: W293–W297. PMID: 16845012

34. Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol: 138–148. PMID: 10786296

35. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121 PMID: 21546353

36. Rozen S, Skaletsky H (1999) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365–386.

37. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2−ΔΔCT method. Methods 25: 402–408. PMID: 11846609

38. Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative CT method. Nat Protoc 3: 1101–1108. PMID: 18546601

39. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcripomes by RNA-Seq. Nat Methods 5: 621–628. doi: 10.1038/nmeth.1226 PMID: 18516045

40. Quan XQ, Zhang HT, Shan L, Bi YP (2006) Advances in plant metallothionein and its heavy metal detoxification mechanisms. Yi Chuan 28: 375–382. PMID: 16551609

41. Gille S, Cheng K, Skinner ME, Liepman AH, Wilkerson CG, et al. (2011) Deep sequencing of voodooly (Amorphophallus konjac): an approach to identify relevant genes involved in the synthesis of the hemicellulose glucomannan. Planta 234: 515–526. doi: 10.1007/s00425-011-1422-z PMID: 21538106

42. Nag A, Karpinets TV, Chang CH, Bar-Peled M (2012) Enhancing a Pathway-Genome Database (PGDB) to capture subcellular localization of metabolites and enzymes: the nucleotide-sugar biosynthetic pathways of Populus trichocarpa. Database (Oxford) doi: 10.1093/database/bas013

43. Pauly M, Gille S, Liu LF, Mansoori N, de Souza A, et al. (2013) Hemicellulose biosynthesis. Planta 238: 627–642. doi: 10.1007/s00425-013-1921-1 PMID: 23801299

44. Wang Y, Alonso AP, Wilkerson CG, Keegstra K (2012) Deep EST profiling of developing fenugreek endosperm to investigate galactomannan biosynthesis and its regulation. Plant Mol Biol 79: 243–258. doi: 10.1007/s11103-011-9909-y PMID: 22927750

45. Park JI, Ishimizu T, Suwabe K, Sudo K, Masuko H, et al. (2010) UDP-glucose pyrophosphorylase is rate limiting in vegetative and reproductive phases in Arabidopsis thaliana. Plant Cell Physiol 51: 981–996. doi: 10.1093/pcp/pcq057 PMID: 20435647

46. Zhang Q, Hrmova M, Shirley NJ, Lahnstein J, Fincher GB. (2006) UDP-glucose pyrophosphorylase is rate limiting in vegetative and reproductive phases in Arabidopsis thaliana. Plant Cell Physiol 51: 981–996. doi: 10.1093/pcp/pcq057 PMID: 20435647

47. Yin Y, Huang J, Gu X, Bar-Peled M, Xu Y (2011) Evolution of plant nucleotide-sugar interconversion enzymes. PLoS One 6: e27995. doi: 10.1371/journal.pone.0027995 PMID: 22126650

48. Gomez-Casati DF, Martin M, Busi MV (2013) Polysaccharide-synthesizing glycosyltransferases and carbohydrate binding modules: the case of starch synthase III. Protein Pept Lett 20: 856–865. doi: 10.1080/10295016.2013.766201

49. Wang X (2009) Structure, mechanism and engineering of plant natural product glycosyltransferases. FEBS Lett 583: 3303–3309. doi: 10.1016/j.febslet.2009.09.042 PMID: 19796637

50. Gachon CM, Langlois-Meurinne M, Saiendran P (2005) Plant secondary metabolism glycosyltransferases: the emerging functional analysis. Trends Plant Sci 10: 542–549. PMID: 16214386

51. Qiu YX, Su YC, Guo JL, Wu QB, Xu LP (2014) A global view of transcriptome dynamics during Sporisorium scitamineum challenge in sugarcane by RNA-Seq. PLoS One 9: e106476. doi: 10.1371/journal.pone.0106476

52. He M, Wang Y, Hua WP, Zhang Y, Wang ZZ (2012) De novo sequencing of Hypericum perforatum transcriptome to identify potential genes involved in the biosynthesis of active metabolites. PLoS ONE 7: e42081. doi: 10.1371/journal.pone.0042081

53. Lai Z, Lin Y (2013) Analysis of the global transcriptome of longan (Dimocarpus longan Lour.) embryogenic callus using Illumina paired-end sequencing. BMC Genomics 14:561. doi: 10.1186/1471-2164-14-561 PMID: 23957614
54. Zhang YJ, Liang ZY, Zhao W, Huo X, Zhang LX, et al. (2005) Separation, purification and compositional analysis of water soluble polysaccharide CPPS3 from Codonopsis pilosula. Chinese Pharmaceutical Journal 40: 1107–1109.

55. Labate MT, Bertolo AL, do Nascimento DD, Gutmanis G, de Andrade A, et al. (2010) Cloning and endogenous expression of a Eucalyptus grandis UDP-glucose dehydrogenase cDNA. Genet Mol Biol 33: 686–695. doi: 10.1590/S1415-47572010005000078 PMID: 21637578

56. Wang J, Ji Q, Jiang L, Shen S, Fan Y, et al. (2009) Overexpression of a cytosol-localized rhamnose biosynthesis protein encoded by Arabidopsis RHM1 gene increases rhamnose content in cell wall. Plant Physiol Biochem 47: 86–93. doi: 10.1016/j.plaphy.2008.10.011 PMID: 19056285