Genetics and Molecular Research

De novo assembly and characterization of *Gleditsia sinensis* transcriptome and subsequent gene identification and SSR mining

S. Han[1], Z. Wu[1], X. Wang[1], K. Huang[1], Y. Jin[1], W. Yang[1] and H. Shi[1,2]

[1]Hubei Key Laboratory of Genetic Regulation and Integrative Biology, School of Life Sciences, Central China Normal University, Wuhan, China
[2]Department of Chemistry and Biochemistry, Texas Tech University, Lubbock, TX, USA

Corresponding authors: W. Yang / H. Shi
E-mail: yangwn@mail.ccnu.edu.cn / huazhong.shi@ttu.edu

Genet. Mol. Res. 15 (1): gmr.15017740
Received September 28, 2015
Accepted December 22, 2015
Published January 26, 2016
DOI http://dx.doi.org/10.4238/gmr.15017740

ABSTRACT. *Gleditsia sinensis* is a Chinese native deciduous tree with a high economic and medicinal value. However, there is limited knowledge on the molecular processes responsible for the medical properties of this species owing to lack of bioinformatic resources such as available whole-genome sequences. In the present study, RNA sequencing data were used to analyze the transcriptome of *G. sinensis*, and a series of bioinformatic tools was used to explore the main genes involved in important molecular processes. A total of 75.57 million paired-end reads, with a length of 101 bp, were acquired from *G. sinensis*. Using the assembly tool Trinity, 233,751 transcripts were discovered. Among these, 85,795 were identified as unique transcripts and 59,326 unique transcripts were found to contain coding regions. Gene ontology analysis identified 27,637 unique transcripts that were clustered into 56 functional groups. Genes involved in flavonoid and terpenoid backbone biosynthesis and those encoding transcription

factors were further analyzed. Sequence analysis revealed four putative *G. sinensis* chalcone isomerase genes (*GsCHI*) encoding the enzymes for flavonoid biosynthesis. *GsCHI1* was found to be phylogenetically related to the chalcone isomerase of the family Leguminosae, and its transcript levels in different tissues were higher than those of *GsCHI2*, *GsCHI3*, and *GsCHI4*. Furthermore, 15,014 simple sequence repeat (SSR) markers were discovered in the transcript library, and 5170 primers were generated for the SSR loci. The genetic and genomic information presented in this study will be helpful for future studies on gene discovery and molecular processes in *G. sinensis*.

Key words: Chalcone isomerase; *Gleditsia sinensis*; Transcriptome assembly; Unique transcripts; Gene identification; SSR mining

INTRODUCTION

Gleditsia sinensis Lam. is a leguminous tree native to China. It has high economic, medicinal, and ecological value, and it has been used as a detergent in China for thousands of years (Tilstone et al., 1998; Lan et al., 2004; Li et al., 2014). In Chinese traditional medicine, the thorns of *G. sinensis* are used for the treatment of inflammatory diseases, including swelling, suppuration, carbuncle, and skin diseases (Ahn, 2003; Ha et al., 2008; Seo et al., 2015). *G. sinensis* can survive in arid regions and is tolerant to extreme temperature conditions. Therefore, this species is useful for landscaping purposes and reforestation (Tilstone et al., 1998; Lan et al., 2004).

In the last twenty years, several antimicrobial compounds have been identified from natural resources, including plants because of their superior structural diversity, unique bioactivity, and environmental compatibility as compared to compounds derived from organic synthesis (Tang et al., 2010). In fact, a variety of antimicrobial compounds including flavonoids, phenols, saponins, and alkaloids, have been isolated from different leguminous plants. Three flavonoids isolated from *G. sinensis* (dihydrokaempferol, quercetin, and 3,30,50,5,7-pentahydroflavanone) were shown to have significant antibacterial activities (Zhou et al., 2007). Thus, *G. sinensis* could be an important medicinal plant for modern natural medicine. However, studies on *G. sinensis* have been very limited, and little is known about the main active ingredients and their biosynthesis in this species.

RNA sequencing (RNA-Seq) is a powerful technology for genome-wide analysis of RNA transcripts, which can be used for all organisms including those with and without sequenced genomes. In relation to other technologies, such as microarray, which is dependent on available whole-genome sequences, RNA-Seq can be used to assemble short reads for gene annotation and gene expression analysis in organisms without the use of reference genomes (Xu et al., 2012). In addition, RNA-Seq can also be used for analysis of mRNA splicing, discovery of single nucleotide polymorphisms, and mining of expressed sequence tag-simple sequence repeat (EST-SSR), among other methods (Torres et al., 2008). Zhu et al. (2014) analyzed the transcriptome of *G. sinensis* using RNA-Seq; however, the assembled sequences are not available. In the present study, we used the existing RNA-Seq data of *G. sinensis* and carried out transcriptome assembly, gene annotation, and EST-SSR mining. Our results provide a reference for future research on *G. sinensis* such as studies on gene cloning or on the biosynthesis and regulation of active pharmaceutical compounds.

MATERIAL AND METHODS

RNA-Seq data processing and sequence assembly

G. sinensis RNA-Seq data (accession No. SRX365131) was generated by Zhu et al. (2014). The methodology used for tissue sampling, total RNA isolation, and Illumina sequencing is explained in Zhu et al. (2014). The raw RNA-Seq reads of *G. sinensis* were processed using three tools: 1) fastq-dump (from the NCBI SRA toolkit), used to convert data for downstream analysis from the sequence read archive format to the FASTQ format (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software); 2) FastQC (v. 0.10.1), used to assess the quality of the reads (http://www.nipgr.res.in/ngsqctoolkit.html); and 3) Trimmomatic (v. 0.32), a flexible trimmer for Illumina sequence data that was used to filter the high-quality sequence data for further processing and assembly (http://www.usadellab.org/cms/index.php?page=trimmomatic).

The clean reads were then *de novo* assembled with Trinity (Trinityrnaseq _r20131110) (Grabherr et al., 2011), and the longest transcripts among the assembled transcripts were selected. TransDecoder (http://sourceforge.jp/projects/sfnet_transdecoder/downloads/OLDER/TransDecoder_r20131117.tar.gz) was used to identify coding regions within unique transcript sequences. The deduced protein sequences were acquired using default parameters, with exception of the -m parameter which was set to 50.

Gene annotation with the BLASTp algorithm

In order to annotate the protein sequences, these were subjected to BLASTp analysis (E-value cut-off of 1 x 10^{-3}) against several databases such as the NCBI non-redundant protein database (NR) (http://www.ncbi.nlm.nih.gov), SWISS-PROT (http://www.expasy.ch/sprot), eukaryotic orthologous groups (KOG), and clusters of orthologous groups (COG) (Kanehisa et al., 2008). The functional classification of the unique transcripts was analyzed using Blast2GO (http://blast2go.com/webstart/blast2go1000.jnlp) based on gene ontology (GO) terms (Conesa et al., 2005; Gene Ontology Consortium, 2008) and plotted using WEGO (Ye et al., 2006). To obtain the various metabolic pathways, sequences and corresponding enzymes were extracted from Blast2GO and mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathway database (Kanehisa and Goto 2000). To further enrich the pathway annotation and identify the BRITE functional hierarchies, sequences were submitted to the KEGG automatic annotation server and the single-directional best hit information method was selected (Moriya et al., 2007).

Gene validation and expression analysis

Fruits, stems, leaves, and rachises of *G. sinensis* were collected from trees cultivated in the Wuhan University campus (Hubei, China) (30.539636°N, 114.358518°E). Total RNAs from these tissues were extracted as described by Jaakola et al. (2001). *G. sinensis* mRNA was reverse-transcribed into double-stranded cDNA using a PrimeScript RT reagent kit with gDNA Eraser (Perfect Real Time, TaKaRa, China) and a modified oligo(dT)18 primer. The expression level of the chalcone isomerase gene in *G. sinensis* (*GsCHI*) relative to the actin gene was determined in different tissues using quantitative polymerase chain reaction (qPCR). The qPCR was performed with the SYBR Premix DimerEraser kit (TaKaRa, China), following manufacturer protocol, and

using a Bio-Rad C1000 system (Bio-Rad, Hercules, CA, USA). Three biological replicates were taken of each sample and each reaction was performed in triplicate. The cDNA sequences of *GsCHI* genes were determined by Sanger sequencing.

Sequences of GsCHI homologous proteins were obtained from NCBI, and aligned using Clustal X (Larkin et al., 2007). A phylogenetic tree based on the conserved amino acid sequences was generated by the Neighbor-Joining method with Poisson model using the Mega 5.0 program (http://www.megasoftware.net/). Bootstrap values were derived from 1000 replicates (Tamura et al., 2011).

SSR mining and primer design

Microsatellite mining was conducted by using the MIcroSAtellite (MISA, http://pgrc.ipk-gate-rsleben.de/misa/) software. Default parameters (unit_size-min_repeats: 1-10, 2-6, 3-5, 4-5, 5-5, 6-5) were used. The primer pair for SSR was designed by Primer3 (v. 2.3.6, http://primer3.sourceforge.net) considering default settings and a PCR product length ranging from 100 to 250 bp (Thiel et al., 2003).

RESULTS

De novo assembly and data analysis

Illumina sequencing generated a total of 75,570,488 paired-end raw reads with 7,632,619,288 bases. Pre-processing resulted in 70,928,068 clean reads and 6,944,576,794 bases. All bases had quality scores >Q30 (**Figure S1**), indicating that sample size and read quality were appropriate for further analysis (Xia et al., 2011; Long et al., 2014; Wu et al., 2015).

The assembly of these clean reads generated 233,751 transcripts, of which 85,795 were unique transcripts. Unique transcripts had an average length of 1648 bp and a N50 value of 2556 bp (GenBank accession No. GCKC01000000). The length of the assembled transcripts ranged from 201 bp to 17,276 bp, among which approximately 41% were in the range of 201-1000 bp, and 80,507 transcripts (34%) were longer than 2.1 kb (Figure 1).

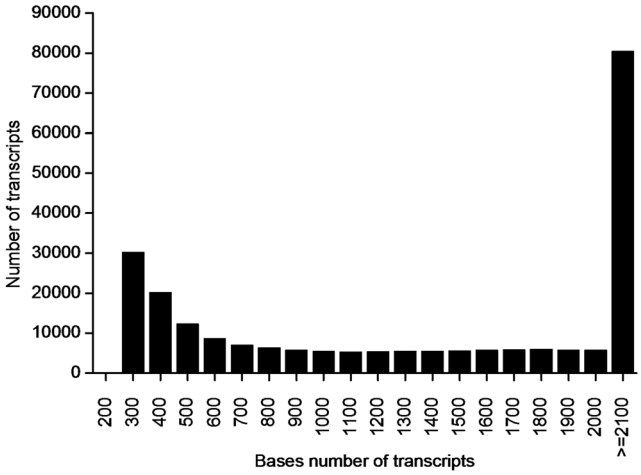

Figure 1. Number of transcripts in relation to transcript length.

Annotation and analysis of unique transcripts

From the 85,795 unique transcripts, 59,326 transcripts with coding regions were identified and translated into protein sequences using TransDecoder. A total of 48,343 protein sequences were annotated (Table 1). Among the 59,326 unique transcripts, 40,024 (67.46%) were matched to the protein database NCBI NR with an E-value ≤1 x 10^{-3}, 27,866 (46.97%) were aligned to SWISS-PROT, 33,691 (56.79%) showed significant homology in the KOG database (**Figure S2**), and 17,474 (29.45%) were annotated by COG database (**Figure S3**). All the annotation details are shown in **Table S1**. A total of 10,983 unique transcripts could not be annotated by any of the public databases.

Table 1. Annotations of *G. sinensis* assembled transcripts against public databases.

Database	59,326 transcripts with predicted coding regions	
	Annotated (N)	Percentage (%)
NR	40,024	67.46
KOG	33,691	56.79
COG	17,474	29.45
SWISS-PROT	27,866	46.97
KEGG	16,583	27.95
GO	27,636	46.58
Total	48,343	81.49

GO

Among the 59,326 unique transcripts with coding regions, 27,636 (46.58%) fell into at least one functional group. Figure 2 summarizes the number of the deduced proteins in each of the 56 GO functional groups, which were divided into three main categories: biological process (17,892 genes), cellular component (12,909 genes), and molecular function (19,404 genes). The most abundant proteins were found in the categories cellular processes (49.1%) and metabolic processes (48.1%) of the biological process group, in the cell (46.0%) and cell part (45.5%) of the cellular component group, and in the binding protein (44.2%) and catalytic activity (45.7%) categories of the molecular function group (Figure 2).

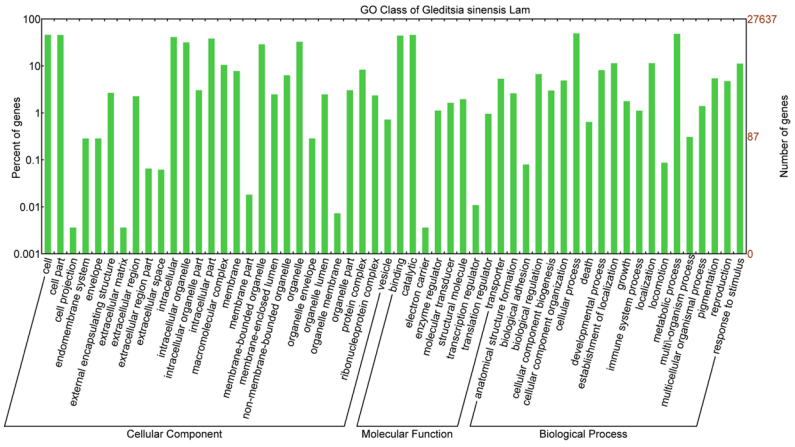

Figure 2. GO annotation results of the transcriptome of *Gleditsia sinensis*.

Protein sequences within *G. sinensis* functional categories were used to identify top-hit species using BLASTp (E-value cutoff of 1 x 10^{-3}). Results suggest that *G. sinensis* is a close relative to the leguminous plants *Glycine max* (28.97%), *Phaseolus vulgaris* (11.07%), *Cicer arietinum* (9.81%), and *Medicago truncatula* (5.42%) (Figure 3). This result is consistent with the general knowledge about the evolutionary history of *G. sinensis* in the plant kingdom. Interestingly, the BLAST result also showed that 11.43% of the deduced proteins of *G. sinensis* are similar to the proteins from *Rhizophagus irregularis*, a fungus that forms symbiotic relationships with plant roots and contributes to the phosphorus cycle (Figure 3).

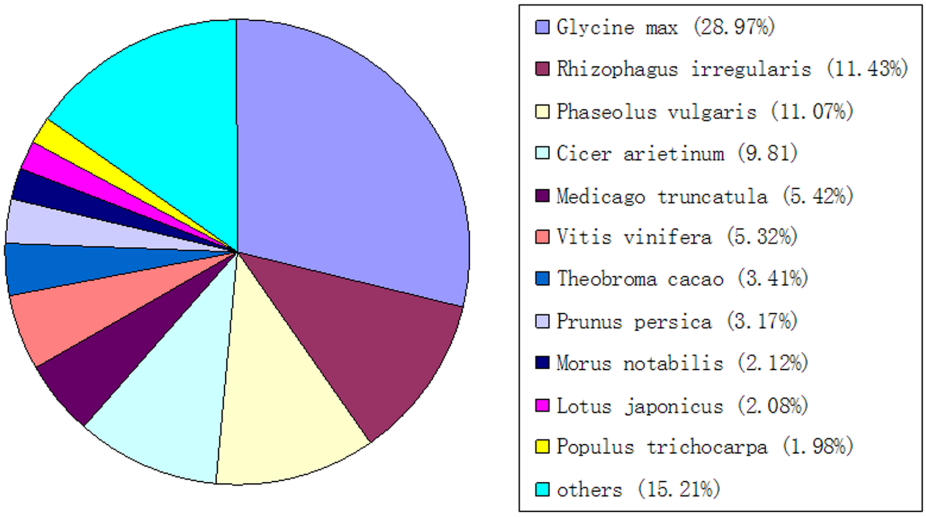

Figure 3. BLAST top-hit species distribution based on the transcripts of *Gleditsia sinensis*.

Secondary metabolic pathway analysis

To identify the enzymes and the involved secondary metabolic pathways, unique protein sequences were annotated with Enzyme Commission (EC) numbers from BLASTP alignments against the KEGG database (E-value ≤1 x 10^{-3}), and the assigned EC numbers were subsequently mapped to the defined metabolic pathways. As a result, 921 unique open reading frames (ORFs) involved in secondary metabolic pathways, possibly contributing to the biosynthesis of medicinal ingredients, were identified in *G. sinensis* (Figure 4). Among these, 144 may be involved in the biosynthesis of alkaloids, 2 of brassinosteroids, 33 of caffeine, 19 of carotenoids, 76 of diterpenoids, 24 of flavones and flavonols, 35 of limonene and pinene, 112 of flavonoids including isoflavonoids, 287 of phenylpropanoids, 156 of terpenoids, and 33 of zeatin. An example of 82 unique ORFs classified into 12 groups based on key enzymes involved in the biosynthesis of flavonoids is found in **Table S2**. The 12 groups of key enzymes [phenylalanine ammonia lyase, cinnamic acid 4-hydroxylase, 4-coumarate-CoA ligase, chalcone synthase, chalcone isomerase (CHI), flavanone 3-hydroxylase, flavonoid 3'-hydroxylase, flavonol synthase, NADPH-dihydromyricetin reductase, anthocyanidin synthase, anthocyanidin reductase, and leucoanthocyanidin reductase] encompassed almost all enzymes in the flavonoid biosynthetic pathway (**Table S2**) (Blount et al.,

2000; Lacampagne et al., 2010; Yao et al., 2012; Dong and Shang 2013; Han et al., 2014; Morita et al., 2014; Schwinn et al., 2014). In addition, 82 unique transcripts encoded proteins thought to be involved in terpenoid backbone biosynthesis. These 82 proteins were classified into 24 classes of key enzymes, which covered almost all processes involved in the metabolic pathways of terpenoid backbone biosynthesis (**Table S3**). Results show that the assembled sequences are relatively complete. Similar analysis was also done for cytochrome P450 superfamily, MYB, and bHLH transcription factors (**Table S4**).

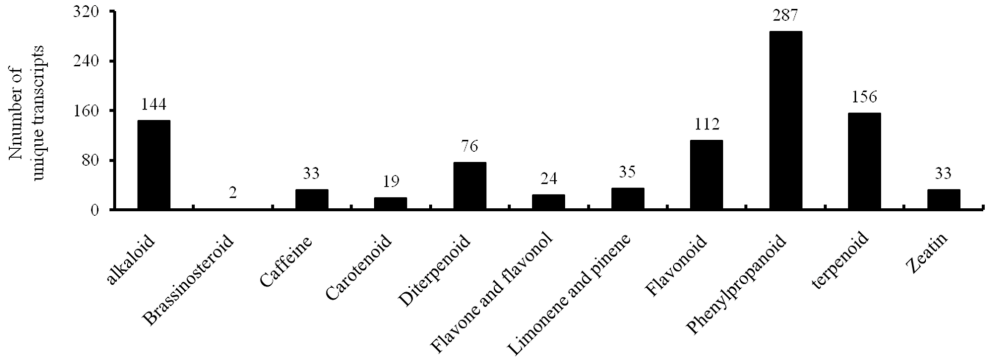

Figure 4. Unique transcripts from *Gleditsia sinensis* related to secondary metabolic pathways.

Verification and phylogenetic analysis of chalcone isomerase CHI

CHI (EC 5.5.1.6) is a key enzyme for flavonoid biosynthesis that catalyzes the stereospecific isomerization of chalcones into their corresponding (2S)-flavanones (Jez et al., 2000). Sequence analysis suggested four unique ORFs encoding chalcone isomerase in *G. sinensis*. In order to verify the accuracy of the RNA-Seq assembly results and understand the role of this key enzyme in the flavonoid biosynthesis, the four unique ORFs of *CHI* (Gs48559_c0_seq1, Gs42647_c0_seq1, Gs19260_c0_seq1, and Gs13596_c1_seq1) were confirmed by Sanger sequencing. For convenience, the related sequences were named as *GsCHI1*, *GsCHI2*, *GsCHI3*, and *GsCHI4*, respectively (**Table S5**). The four cloned sequences showed 99% (896/906), 100% (627/627), 100% (832/832), and 100% (633/633) identity to the assembled sequences (**Figure S4**). Therefore, Gs48559_c0_seq1, Gs42647_c0_seq1, Gs19260_c0_seq1, and Gs13596_c1_seq1 identified from the RNA-Seq data analysis were indeed from *G. sinensis*.

Phylogenetic analysis provides the evolutionary relationship between genes of different species (Durbin et al., 2000). The genetic relationship among the *CHI* genes of different plant species, including *GsCHI1*, is shown in Figure 5. Phylogenetic analysis revealed that the *GsCHI1* gene was closely related to *CHI* genes from leguminous plants (**Figure S5**). In addition, qPCR showed that *GsCHI* genes were expressed in rachises, leaves, fruits, and young stems, and that *GsCHI1* exhibited much higher expression levels than other *GsCHI* genes (Figure 6).

SSR mining in *G. sinensis* transcriptome

SSR markers supply useful information for molecular breeding in plants. In this study,

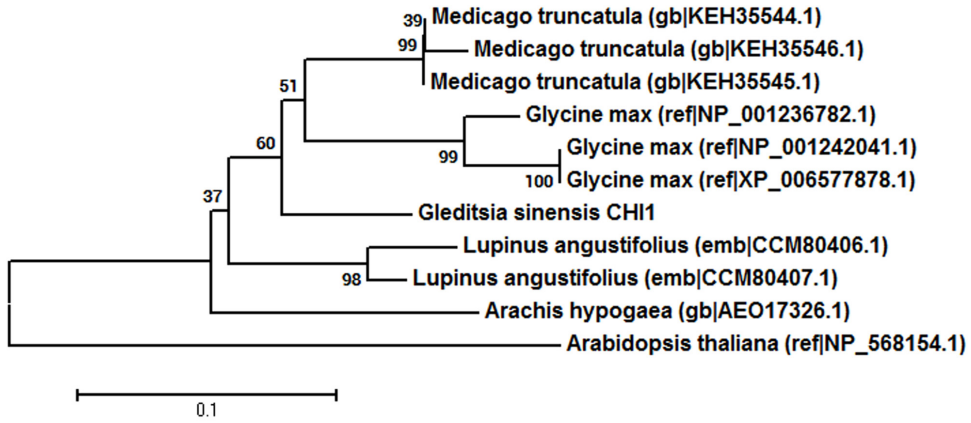

Figure 5. Phylogenetic tree based on the deduced amino acid sequences of various *CHI* open reading frames (ORFs). Amino acid sequences were generated by the Neighbor-Joining method with Poisson model. Bootstrap values were derived from 1000 replicates.

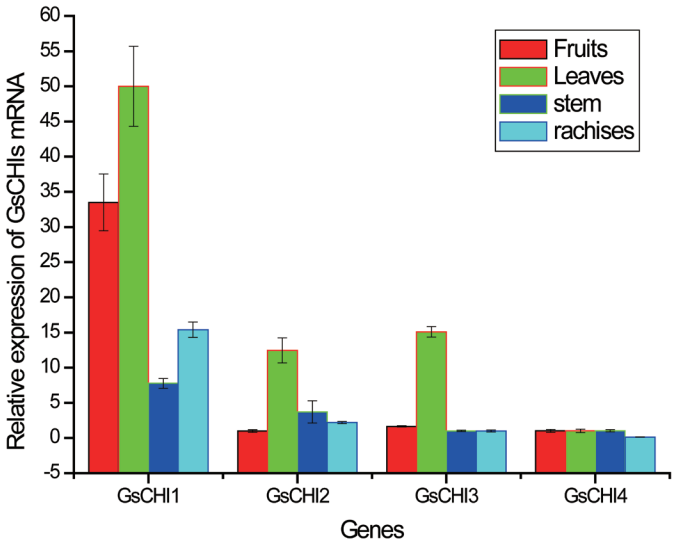

Figure 6. Expression of the GsCHI1 gene in different tissues of *Gleditsia sinensis*.

15,788 SSR markers were identified in 13,122 unique transcript sequences, of which 2169 sequences had more than one SSR marker. Among all the identified SSR markers, 15,014 were shown to be mono-, di-, tri-, tetra-, penta-, or hexa-nucleotide repeats. Among these, dinucleotide repeats were most abundant (63.02%), followed by trinucleotide repeats (34.42%), while other types were very low (Figure 7A). The most abundant motif was AG/CT, followed by AT/AT, AC/GT, AAG/CTT, ATC/ATG, AAT/ATT, AGC/CTG, AGG/CCT, AAC/GTT, ACC/GGT, and ACG/CGT, respectively (Figure 7B). Among all the detected SSR markers, 5170 primer pairs were designed by Primer3 (**Table S6**).

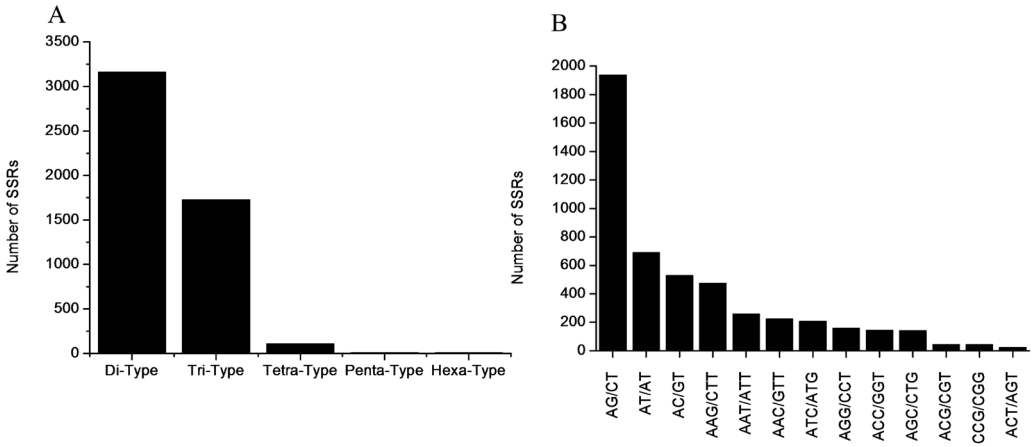

Figure 7. Characterization of SSR markers. **A.** Frequency of the different nucleotide repeat types; **B.** frequency of the classified repeat motifs.

DISCUSSION

G. sinensis is a species of deciduous tree belonging to the family Leguminosae, which has high economic and medicinal value. *G. sinensis* can be used in medical, cosmetic, and health products, as well as in natural raw materials used in cleaning products (Lian and Zhang, 2013). Guar gum, an important vegetable gum present in the seeds of *G. sinensis*, can be used to make food more appetizing (Jian et al., 2013), whereas the thorns of *G. sinensis* have an economic and medicinal value because they contain flavonoid glycosides, amino acids, and phenols (Yi et al., 2012). Thus, *G. sinensis* could be used as natural medicine and as an environment-friendly industrial material.

Traditional Chinese medicine has been practiced for thousands of years, yet the active ingredients in most medicinal plants are still largely unknown. In recent years, studies focusing on the genetic diversity of Chinese herbs and the production of secondary metabolites have been considered important for a better effectiveness of the treatment (Zheng et al., 2015). However, only a small number of reports suggested that secondary metabolites such as alkaloids, brassinosteroids, caffeine, and flavonoids are the active ingredients in medicinal plants (Zhang et al., 1999; Zhou et al., 2007). In this study, 921 unique transcripts were considered to be involved in secondary metabolic pathways important for the production of active ingredients in *G. sinensis* (Figure 4). The transcripts encoding putative enzymes for flavonoid and terpenoid biosynthesis covered almost all the pathways, suggesting that the transcriptome assembled in this study represents genes expressed in the whole genome of *G. sinensis*. The analysis of secondary metabolic pathways is useful for studying the genes involved in the production of active ingredients and their expression patterns. In this study, CHI, an important enzyme in the flavonoid biosynthesis pathway, was studied based on the assembled transcripts and their annotation (following Morita et al., 2014). Four putative *CHI* genes (*GsCHI1*, *GsCHI2*, *GsCHI3*, and *GsCHI4*) were cloned and their expression levels were determined in different tissues. The expression level of *GsCHI1* was high in fruits and leaves. The fruits of *G. sinensis* have been used for a long time in traditional Chinese medicine in the treatment of various diseases. Triterpenoid saponins present in the fruit of *G. sinensis* have been considered potential antitumor agents (Lu et al., 2014), but little

is known about the enzymes and biochemical pathways involved in saponin biosynthesis in this species. The present study provided candidate genes for key enzymes involved in the terpenoid backbone biosynthesis pathway that are, therefore, helpful for the identification of key enzymes involved in the biosynthesis of triterpenoid saponins. Results support the use of RNA-Seq in the discovery and cloning of novel genes in species without a known whole-genome sequence. Using this strategy, transcription factors such as MYB and bHLH that are known to be involved in the regulation of metabolite synthesis (Broun et al., 2006; Mizutani and Sato, 2011; Patra et al., 2013), have also been identified (**Table S4**). These results will help elucidate the mechanisms controlling or regulating the metabolic pathways in *G. sinensis*.

Overall, 233,751 transcripts and 85,795 unique transcripts were assembled from the RNA-Seq data. Among them, 41.22% were in the size range of 201-1000 bp and 80,507 transcripts were longer than 2.1 kb. The average transcript length and the calculated N50 value suggest that the results in this study provide additional information to the previous reports by Zhu et al. (2014). Among the 85,795 unique transcripts, 59,326 transcripts with coding regions were deduced to protein sequences and 81.49% (48,343) of the deduced protein sequences were annotated based on a similarity search against public databases. A total of 10,983 unique transcripts could not be annotated. These transcripts might encode proteins that are specific to *G. sinensis* or other related plant species, and should be included in a future study. A total of 15,788 SSR markers were identified in 13,122 unique transcripts and 5170 primer pairs were designed based on the sequence library. These results are important for future genetic studies in *G. sinensis* (Morgante et al., 2002; Zhang et al., 2015). Overall, the results of this study, especially the assembled transcript sequences, are useful for future molecular studies of this important medicinal plant.

To conclude, this study provided 233,751 transcripts to NCBI, among them 59,326 unique transcripts were annotated to contain coding regions and 48,343 unique transcripts had at least one hit with a sequence in public databases. In addition, this study provides the most comprehensive transcriptome assembly for *G. sinensis*. A total of 921 unique ORFs involved in secondary metabolic pathways, possibly contributing to the biosynthesis of medicinal ingredients, were identified. Furthermore, four *GsCHI* genes were cloned for the first time. Additionally, a large number of SSR markers were identified and 5170 primer pairs were designed. This work will be useful for future studies on genetics, genomics, molecular evolution, and biotechnological improvement in *G. sinensis.*

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research partially supported by the National Natural Science Foundation of China (grant #31270316 to W. Yang and #31328004 to H. Shi) and by the Excellent Doctoral Dissertation Cultivation Grant from the Central China Normal University (grant #2013YBZD24) to S. Han.

REFERENCES

Ahn DK (2003). Illustrated book of Korean medicinal herbs. Kyohaksa, Seoul.

Blount JW, Korth KL, Masoud SA, Rasmussen S, et al. (2000). Altering expression of cinnamic acid 4-hydroxylase in transgenic plants provides evidence for a feedback loop at the entry point into the phenylpropanoid pathway. *Plant Physiol.* 122:

107-116. http://dx.doi.org/10.1104/pp.122.1.107

Broun P, Liu Y, Queen E, Schwarz Y, et al. (2006). Importance of transcription factors in the regulation of plant secondary metabolism and their relevance to the control of terpenoid accumulation. *Phytochem. Rev.* 5: 27-38. http://dx.doi.org/10.1007/s11101-006-9000-x

Conesa A, Götz S, García-Gómez JM, Terol J, et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676. http://dx.doi.org/10.1093/bioinformatics/bti610

Dong CJ and Shang QM (2013). Genome-wide characterization of phenylalanine ammonia-lyase gene family in watermelon (*Citrullus lanatus*). *Planta* 238: 35-49. http://dx.doi.org/10.1007/s00425-013-1869-1

Durbin ML, McCaig B and Clegg MT (2000). Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Mol. Biol.* 42: 79-92. http://dx.doi.org/10.1023/A:1006375904820

Gene Ontology Consortium (2008). The gene ontology project in 2008. *Nucleic Acids Res.* 36: D440-D444.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29: 644-652. http://dx.doi.org/10.1038/nbt.1883

Ha HH, Park SY, Ko WS and Kim Y (2008). Gleditsia sinensis thorns inhibit the production of NO through NF-kappaB suppression in LPS-stimulated macrophages. *J. Ethnopharmacol.* 118: 429-434. http://dx.doi.org/10.1016/j.jep.2008.05.004

Han Y, Zhao W, Wang Z, Zhu J, et al. (2014). Molecular evolution and sequence divergence of plant chalcone synthase and chalcone synthase-Like genes. *Genetica* 142: 215-225. http://dx.doi.org/10.1007/s10709-014-9768-3

Jaakola L, Pirttilä AM, Halonen M and Hohtola A (2001). Isolation of high quality RNA from bilberry (*Vaccinium myrtillus* L.) fruit. *Mol. Biotechnol.* 19: 201-203. http://dx.doi.org/10.1385/MB:19:2:201

Jez JM, Bowman ME, Dixon RA and Noel JP (2000). Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase. *Nat. Struct. Biol.* 7: 786-791. http://dx.doi.org/10.1038/79025

Jian HL, Zhu LW, Zhang WM, Sun DF, et al. (2013). Enzymatic production and characterization of manno-oligosaccharides from *Gleditsia sinensis* galactomannan gum. *Int. J. Biol. Macromol.* 55: 282-288. http://dx.doi.org/10.1016/j.ijbiomac.2013.01.025

Kanehisa M and Goto S (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27-30. http://dx.doi.org/10.1093/nar/28.1.27

Kanehisa M, Araki M, Goto S, Hattori M, et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36: D480-D484. http://dx.doi.org/10.1093/nar/gkm882

Lacampagne S, Gagné S and Gény L (2010). Involvement of abscisic acid in controlling the proanthocyanidin biosynthesis pathway in grape skin: new elements regarding the regulation of tannin composition and leucoanthocyanidin reductase (LAR) and anthocyanidin reductase (ANR) activities and expression. *J. Plant Growth Regul.* 29: 81-90. http://dx.doi.org/10.1007/s00344-009-9115-6

Lan YP, Zhou LD, Li SY, Cao QC, et al. (2004). Advances in research of *Gleditsia* and its prospect of industrializational development. *World Forestry Res.* 6: 003.

Larkin MA, Blackshields G, Brown NP, Chenna R, et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948. http://dx.doi.org/10.1093/bioinformatics/btm404

Li JJ, Wang J, He JB, Ren ML, et al. (2014). The standard operating procedure (SOP) of *Gleditsia sinensis* planting technologies. *Hans J. Agric. Sci.* 4: 151-159.

Lian XY and Zhang Z (2013). Quantitive analysis of *gleditsia* saponins in the fruits of *Gleditsia sinensis* Lam. by high performance liquid chromatography. *J. Pharm. Biomed. Anal.* 75: 41-46. http://dx.doi.org/10.1016/j.jpba.2012.11.007

Long Y, Zhang J, Tian X, Wu S, et al. (2014). *De novo* assembly of the desert tree *Haloxylon ammodendron* (C. A. Mey.) based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. *BMC Genomics* 15: 1111. http://dx.doi.org/10.1186/1471-2164-15-1111

Lu D, Xia Y, Tong B, Zhang C, et al. (2014). *In vitro* anti-angiogenesis effects and active constituents of the saponin fraction from *Gleditsia sinensis*. *Integr. Cancer Ther.* 13: 446-457. http://dx.doi.org/10.1177/1534735412442377

Mizutani M and Sato F (2011). Unusual P450 reactions in plant secondary metabolism. *Arch. Biochem. Biophys.* 507: 194-203. http://dx.doi.org/10.1016/j.abb.2010.09.026

Morgante M, Hanafey M and Powell W (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30: 194-200. http://dx.doi.org/10.1038/ng822

Morita Y, Takagi K, Fukuchi-Mizutani M, Ishiguro K, et al. (2014). A chalcone isomerase-like protein enhances flavonoid production and flower pigmentation. *Plant J.* 78: 294-304. http://dx.doi.org/10.1111/tpj.12469

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, et al. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35: W182-5. http://dx.doi.org/10.1093/nar/gkm321

Patra B, Schluttenhofer C, Wu Y, Pattanaik S, et al. (2013). Transcriptional regulation of secondary metabolite biosynthesis in plants. *Biochim. Biophys. Acta* 1829: 1236-1247. http://dx.doi.org/10.1016/j.bbagrm.2013.09.006

Schwinn K, Miosic S, Davies K, Thill J, et al. (2014). The B-ring hydroxylation pattern of anthocyanins can be determined through activity of the flavonoid 3'-hydroxylase on leucoanthocyanidins. *Planta* 240: 1003-1010. http://dx.doi.org/10.1007/s00425-014-2166-3

Seo CS, Lim HS, Ha H, Jin SE, et al. (2015). Quantitative analysis and anti-inflammatory effects of *Gleditsia sinensis* thorns in RAW 264.7 macrophages and HaCaT keratinocytes. *Mol. Med. Rep.* 12: 4773-4781.

Tamura K, Peterson D, Peterson N, Stecher G, et al. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28: 2731-2739. http://dx.doi.org/10.1093/molbev/msr121

Tang J, Meng X, Liu H, Zhao J, et al. (2010). Antimicrobial activity of sphingolipids isolated from the stems of cucumber (*Cucumis sativus* L.). *Molecules* 15: 9288-9297. http://dx.doi.org/10.3390/molecules15129288

Thiel T, Michalek W, Varshney RK and Graner A (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 106: 411-422.

Tilstone GH, Pasiecznik NM, Harris PJC and Wainwright SJ (1998). The growth of multipurpose tree species in the Almeria province of Spain and its relationship to native plant communities. *Int. Tree Crops J.* 9: 247-259. http://dx.doi.org/10.1080/01435698.1998.9752982

Torres TT, Metta M, Ottenwälder B and Schlötterer C (2008). Gene expression profiling by massively parallel sequencing. *Genome Res.* 18: 172-177. http://dx.doi.org/10.1101/gr.6984908

Wu L, Wen C, Qin Y, Yin H, et al. (2015). Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. *BMC Microbiol.* 15: 125. http://dx.doi.org/10.1186/s12866-015-0450-4

Xia Z, Xu H, Zhai J, Li D, et al. (2011). RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*. *Plant Mol. Biol.* 77: 299-308. http://dx.doi.org/10.1007/s11103-011-9811-z

Xu DL, Long H, Liang JJ, Zhang J, et al. (2012). De novo assembly and characterization of the root transcriptome of *Aegilops variabilis* during an interaction with the cereal cyst nematode. *BMC Genomics* 13: 133. http://dx.doi.org/10.1186/1471-2164-13-133

Yao X, Shang E, Zhou G, Tang Y, et al. (2012). Comparative characterization of total flavonol glycosides and terpene lactones at different ages, from different cultivation sources and genders of *Ginkgo biloba* leaves. *Int. J. Mol. Sci.* 13: 10305-10315. http://dx.doi.org/10.3390/ijms130810305

Ye J, Fang L, Zheng H, Zhang Y, et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34: W293-7. http://dx.doi.org/10.1093/nar/gkl031

Yi JM, Park JS, Oh SM, Lee J, et al. (2012). Ethanol extract of *Gleditsia sinensis* thorn suppresses angiogenesis *in vitro* and *in vivo*. *BMC Complement. Altern. Med.* 12: 243. http://dx.doi.org/10.1186/1472-6882-12-243

Zhang C, Wang G, Hou L, Ji Z, et al. (2015). De novo assembly and characterization of the skeletal muscle transcriptome of sheep using Illumina paired-end sequencing. *Biotechnol. Lett.* 37: 1747-1756. http://dx.doi.org/10.1007/s10529-015-1854-9

Zhang Z, Koike K, Jia Z, Nikaido T, et al. (1999). Four new triterpenoidal saponins acylated with one monoterpenic acid from *Gleditsia sinensis*. *J. Nat. Prod.* 62: 740-745. http://dx.doi.org/10.1021/np980441k

Zheng WH, Zhuo Y, Liang L, Ding WY, et al. (2015). Conservation and population genetic diversity of *Curcuma wenyujin* (Zingiberaceae), a multifunctional medicinal herb. *Genet. Mol. Res.* 14: 10422-10432. http://dx.doi.org/10.4238/2015.September.8.3

Zhou L, Li D, Wang J, Liu Y, et al. (2007). Antibacterial phenolic compounds from the spines of *Gleditsia sinensis* Lam. *Nat. Prod. Res.* 21: 283-291. http://dx.doi.org/10.1080/14786410701192637

Zhu L, Zhang Y, Guo W and Wang Q (2014). *Gleditsia sinensis*: transcriptome sequencing, construction, and application of its protein-protein interaction network. *BioMed Res. Int.* 2014: 404578. http://dx.doi.org/10.1155/2014/404578

Supplementary material

Table S1. The information of the blast data against the public databases.

Table S2. Putative unique transcripts involved in flavonoid biosynthesis. BLAST analysis against the KEGG database (based on Lin X and Saito K et al. [1] [2] and with a few modifications). aEnzyme code.

Table S3. Candidate genes encoding key enzymes that participate in terpenoid backbone biosynthesis. aEnzyme code.

Table S4. Putative unique transcripts encoding cytochrome P450 superfamily, MYB, and bHLH transcription factor.

Table S5. Primers used for Q-PCR in the experiment of *Gleditsia sinensis*.

Table S6. Primer pairs for SSRs.

www.geneticsmr.com/year2016/vol15-1/pdf/gmr7740_supplementary.pdf