BMC Genomics

RESEARCH ARTICLE Open Access

EST analysis reveals putative genes involved in glycyrrhizin biosynthesis

Ying Li[1], Hong-Mei Luo[1], Chao Sun[1], Jing-Yuan Song[1], Yong-Zhen Sun[1], Qiong Wu[1], Ning Wang[3], Hui Yao[1], André Steinmetz*[3] and Shi-Lin Chen*[1,2]

Abstract

Background: *Glycyrrhiza uralensis* is one of the most popular medicinal plants in the world and is also widely used in the flavoring of food and tobacco. Due to limited genomic and transcriptomic data, the biosynthetic pathway of glycyrrhizin, the major bioactive compound in *G. uralensis*, is currently unclear. Identification of candidate genes involved in the glycyrrhizin biosynthetic pathway will significantly contribute to the understanding of the biosynthetic and medicinal chemistry of this compound.

Results: We used the 454 GS FLX platform and Titanium regents to produce a substantial expressed sequence tag (EST) dataset from the vegetative organs of *G. uralensis*. A total of 59,219 ESTs with an average read length of 409 bp were generated. 454 ESTs were combined with the 50,666 *G. uralensis* ESTs in GenBank. The combined ESTs were assembled into 27,229 unique sequences (11,694 contigs and 15,535 singletons). A total of 20,437 unique gene elements representing approximately 10,000 independent transcripts were annotated using BLAST searches (e-value ≤ 1e-5) against the SwissProt, KEGG, TAIR, Nr and Nt databases. The assembled sequences were annotated with gene names and Gene Ontology (GO) terms. With respect to the genes related to glycyrrhizin metabolism, genes encoding 16 enzymes of the 18 total steps of the glycyrrhizin skeleton synthesis pathway were found. To identify novel genes that encode cytochrome P450 enzymes and glycosyltransferases, which are related to glycyrrhizin metabolism, a total of 125 and 172 unigenes were found to be homologous to cytochrome P450s and glycosyltransferases, respectively. The cytochrome P450 candidate genes were classified into 32 CYP families, while the glycosyltransferase candidate genes were classified into 45 categories by GO analysis. Finally, 3 cytochrome P450 enzymes and 6 glycosyltransferases were selected as the candidates most likely to be involved in glycyrrhizin biosynthesis through an organ-specific expression pattern analysis based on real-time PCR.

Conclusions: Using the 454 GS FLX platform and Titanium reagents, our study provides a high-quality EST database for *G. uralensis*. Based on the EST analysis, novel candidate genes related to the secondary metabolite pathway of glycyrrhizin, including novel genes encoding cytochrome P450s and glycosyltransferases, were found. With the assistance of organ-specific expression pattern analysis, 3 unigenes encoding cytochrome P450s and 6 unigenes encoding glycosyltransferases were selected as the candidates most likely to be involved in glycyrrhizin biosynthesis.

Background

The sequencing and analysis of expressed sequence tags (ESTs) has been a primary tool for the discovery of novel genes in plants, especially in non-model plants for which full genome sequences are not currently available [1]. EST sequencing represents a rapid and cost-effective method for analyzing the transcribed regions of genomes. EST analysis is also a powerful tool for the discovery of genes involved in plant secondary metabolism. The 454 GS FLX sequencing technology has made EST-based resources more readily accessible for non-model organism transcriptomes [2,3]. Our experimental focus for this study was *Glycyrrhiza uralensis* Fisch. ex DC., which is one of the most ancient medicinal herbs and has been used as a Chinese herbal medicine to treat infectious diseases for over 3,000 years [4]. This herb has been extensively stud-

* Correspondence: andre.steinmetz@crp-sante.lu, slchen@implad.ac.cn
[3] Centre de Recherche Public-Santé, Luxembourg, L-1526 Luxembourg
[1] Institute of Medicinal Plant Development (IMPLAD), Chinese Academy of Medical Sciences & Peking Union Medical College, No.151, Malianwa North Road, HaiDian District, Beijing 100193, China
Full list of author information is available at the end of the article

ied and is widely used as a flavoring agent, medicament and tobacco additive. Many of the biological activities of the bioactive constituents of *G. uralensis* have been investigated, including the protection against hepatotoxicity [5,6], anti-ulcer effects [7], anti-inflammatory [8] and anti-tumor promoting activities [9]. This herb also exhibits antiviral activity against various DNA and RNA viruses, including herpes simplex virus [10], HIV [11,12] and severe acute respiratory syndrome (SARS)-associated coronavirus [13].

These biological activities of *G. uralensis* have been primarily attributed to two of its components, flavonoids and saponins. Our research interests primarily concern glycyrrhizin, an oleanane-type triterpene saponin and a well-known natural sweetener that is fifty times sweeter than sugar [14]. Although the various chemical and pharmacological properties of glycyrrhizin in *G. uralensis* have been extensively studied, the biosynthetic pathway of this compound remains poorly understood. Two functional genes encoding squalene synthase (SQS) have been isolated from *G. uralensis* [15]. Two cytochrome P450 genes have also been isolated from *G. uralensis* based on the traditional EST sequencing method [16]. *CYP88D6*, a cytochrome P450 monooxygenase, was characterized by *in vitro* enzymatic activity assays and was shown to catalyze the oxidation of β-amyrin at C-11 to produce 11-oxo-β-amyrin, a possible biosynthetic intermediate in the glycyrrhizin biosynthetic pathway [16]. Another cytochrome P450 from *G. uralensis*, *CYP93E3*, possesses β-amyrin 24-hydroxylase activity in *in vitro* enzymatic activity assays [16]. A functional β-amyrin synthase gene (bAS) has been isolated from *G. glabra* [17,18]. Thus far, only one glycosyltransferase in the *Glycyrrhiza* genus, the isoflavonoid glucosyltransferase in *G. echinata*, has been identified [19]. However, no progress has been made in the identification of the genes involved in the glycosylation of glycyrrhetinic acid to produce glycyrrhizin. Transcriptome sequencing would provide a foundation for detailed studies of gene expression and genetic connectivity with respect to plant secondary metabolism.

In our study, we constructed a cDNA library using the vegetative organs of five-year-old wild *G. uralensis* cultivated from the city of Yanchi in the Ningxia province of China, one of the most famous areas for the production of wild *G. uralensis*. The library was sequenced using the 454 GS FLX platform and Titanium reagents. There are currently 50,666 *G. uralensis* ESTs in the GenBank (GuEST) dbEST database, which were determined using conventional sequencing techniques [20]. In our study, we increased this collection with an additional 59,219 ESTs generated from 454 GS FLX Titanium sequencing. Bioinformatic analyses indicated that almost all of the genes involved in the biosynthesis of the glycyrrhizin skeleton were within the combined EST database, except for mevalonate kinase (EC 2.7.1.36) and DXP synthase (EC 2.2.1.7). Additionally, a pool of candidate genes for cytochrome P450s and glycosyltransferases was established, containing 125 and 172 unigenes, respectively. Finally, using an organ-specific expression pattern analysis, a few unigenes were selected as the candidates most likely to be responsible for glycyrrhizin skeleton modifications. The method described here is a cost-effective technology for the identification of novel genes in non-model organisms that serve as medicinal plants. Furthermore, the ESTs and unigenes described in our study constitute an important resource for future studies of the molecular genetics and functional genomics of *G. uralensis*.

Results and Discussion

454 sequencing and assembly

A *G. uralensis* cDNA library was constructed from a pool of mRNA isolated from the vegetative organs of *G. uralensis* and was sequenced using the 454 GS FLX platform and Titanium regents. After the initial quality filtering, this one-eighth sequencing run produced 59,219 high-quality reads (HQ reads) with an average length of 409 bp (mode, 498; range, 10-653) and a total length of 24.2 Mb. Of these HQ reads, 57,997 exceeded our minimal quality standards (SMART primer filtering; length threshold of 50 bp) and were used in the assembly. The large portion of reads used after trimming and filtering indicates that most of the sequences were used in the assembly and that the quality of the data was considerably high. An additional 50,666 *G. uralensis* ESTs were downloaded from GenBank (GuEST) and combined with our 454 derived ESTs. A summary of these two EST datasets is given in Table 1. Figure 1 illustrates the sequence length distributions of the ESTs derived from 454 sequencing and GenBank. Assembly of the trimmed, size-selected ESTs produced 11,694 contigs with a mean length of 571 bp and a range of 85 - 3,812 bp, as well as an additional 15,535 singletons, for a total of 27,229 unigenes (contigs and singletons) (Table 2). The length distribution of the contigs is shown in Figure 2. The assembly produced a substantial number of large contigs (10,868 contigs were > 200 bp in length). Approximately 16% of the unigenes (4,337 out of 27,229) were composed of the 454 ESTs and GuESTs data (Figure 3). The 454 sequencing identified 16,130 novel unigenes, bringing the total number of different *G. uralensis* unigenes in the databases to 27,229. This number should cover the vast majority of genes from this species, including those expressed at low levels.

For homology searches against known genes, unigenes longer than 200 bp are widely accepted as valid sources, making them sufficient for the effective assignment of functional annotations [21]. Most of the unigene

Table 1: Summary of *G. uralensis* ESTs derived from 454 sequencing and GenBank

	454 EST	GuEST
HQ reads[a]	59,219	50,666
total bases of HQ reads[a]	24,231,155 bp	24,211,504 bp
average HQ read length[a]	409 ± 120 bp	478 ± 198 bp
average quality	32	
reads used in assembly (after trimming and filtering)	57,997	50,451
bases used in assembly	23,526,020 bp	24,099,889 bp
	(97.1%)	(99.5%)

[a]HQ indicates high-quality

sequences were longer than 200 bp and were valid in this study (85%; 23,158/27,229).

Functional annotation

Our annotation method was based on sequence homology searches and the annotations that accompanied them. Its aim was to capture the most informative and complete annotation possible. Additional file 1 shows the hit numbers and percentages relative to those of the major public databases. These annotation statistics show all the unigenes annotated by the BLAST [22] search against the public protein and nucleotide databases (SwissProt [23], KEGG [24], TAIR [25], Nr [26] and Nt [27]) where the e-value threshold was set at 1e-5. Of the 27,229 unigenes, 20,437 unigenes had at least one hit within these databases. The remaining unigenes (24.9%) that were not annotated likely comprise *G. uralensis*-specific genes, as well as genes with homologs in other species whose corresponding biological functions have not yet been investigated.

Table 2: Summary of G. uralensis EST assembly

Contigs	
reads assembled as contigs	92,221
bases assembled as contigs	39,171,687 bp
number of contigs	11,694
total length of contigs	6,673,089 bp
average length of contigs	571 ± 368 bp
range of contig lengths	85 - 3,812 bp
contigs above 200 bp	10,868
N50 contig size	816 bp
depth of contigs	5.84
Singletons	
number of singletons	15,535
total length of singletons	5,687,305 bp
average length of singletons	366 ± 163 bp
range of singleton lengths	50 - 853 bp
singletons above 200 bp	12,290
Unigenes[a]	
number of unigenes	27,229
total length of unigenes	12,360,394 bp
unigenes above 200 bp	23,158

[a]unigenes includes contigs and singletons

Gene Ontology classification

As a result of the completed genomic sequencing of the plant *Arabidopsis thaliana*, the currently available expressed sequences have been invaluable in defining the correct components of the gene structure in this species [28]. To obtain an overview of the gene functions of the ESTs from *G. uralensis*, the annotated unigenes were categorized according to Gene Ontology (GO) on the basis of the TAIR GO slim provided by TAIR [29]. The *G. uralensis* unigenes were compared with the *Arabidopsis* proteome [25] using BlastX [22]. We used the annotations of each unigene to assign it to the GO categories for Molecular Function, Biological Process and Cellular Component (Figure 4). The GO analysis showed that the functions of the identified genes are involved various biological processes. A large number of hydrolases, kinases and transferases were annotated, which suggests that our study may allow for the identification of novel genes involved in the secondary metabolite synthesis pathways.

Putative genes related to the biosynthesis of glycyrrhizin

In this study, our primary goal was to identify genes involved in the glycyrrhizin biosynthetic pathway (Additional file 2 and Table 3) [16,30]. The biosynthesis of glycyrrhizin involves the synthesis of dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP), the biochemically active isoprene units of all terpenoids [30]. This step is followed by the synthesis of the triterpene skeleton, also known as β-amyrin [30], and then by a series of oxidative reactions and glucuronyla-

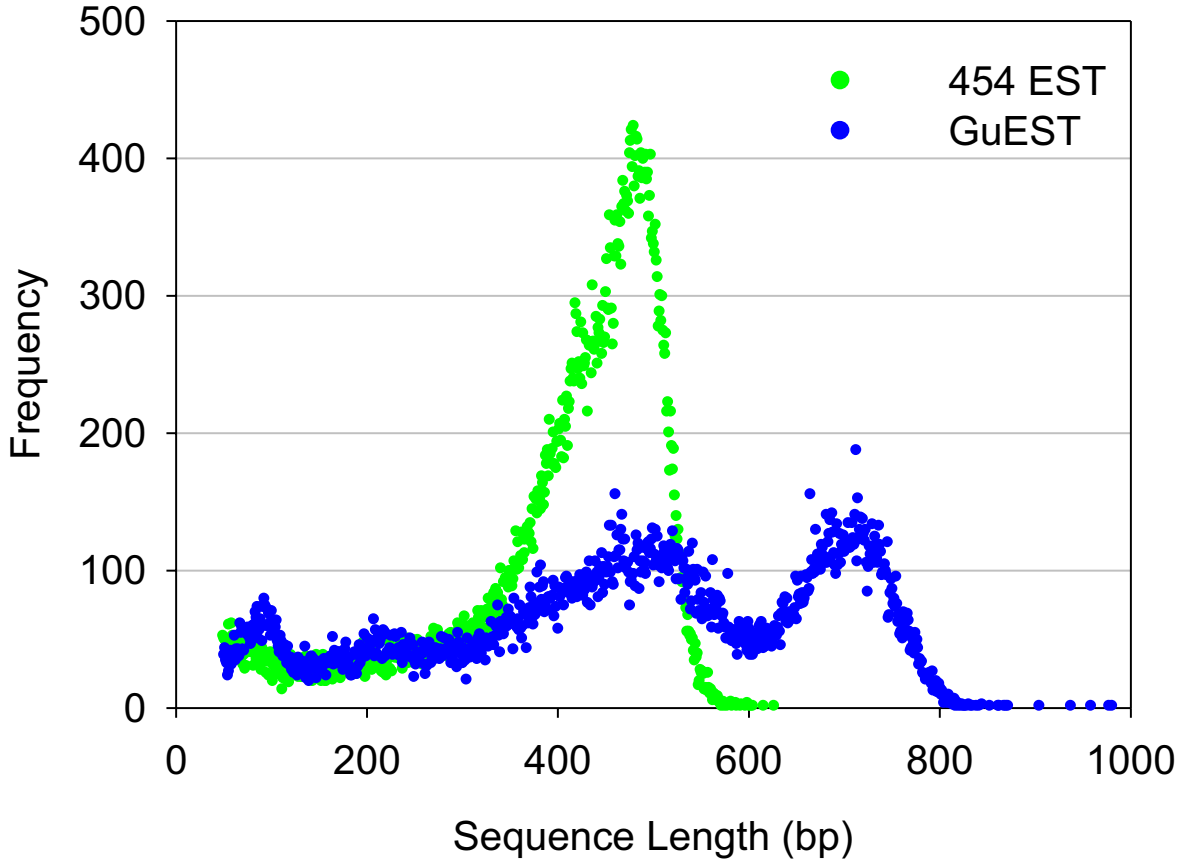

Figure 1 Sequence length distribution of *G. uralensis* 454 ESTs and GenBank ESTs.

tions, which produce glycyrrhizin. The precise order of the intermediate products is still unknown [16].

In the early stage of active isoprene unit formation, plants have the ability to produce DMAPP and IPP using two pathways, the mevalonate pathway (MVA pathway) and the methylerythritol phosphate pathway (MEP pathway). In plants, these two pathways appear to be separate; enzymes of the MVA pathway are found in the cytosol, whereas enzymes of the MEP pathway are localized in plastids. Triterpenoids are known to be formed by the MVA pathway because they are cytosolic products. However, there are examples where the two pathways can act cooperatively to create a molecule [30]. No progress has been made toward determining the precise source of isoprene units in glycyrrhizin biosynthesis. Using a BLAST [22] search against the SwissProt [23] and KEGG [24] databases, we found the genes encoding all of the enzymes from both of these two pathways in the EST database, except for mevalonate kinase (EC 2.7.1.36), which is located in the MVA pathway, and DXP synthase (EC 2.2.1.7), which is located in the MEP pathway. In this study, we found all of the putative genes encoding the enzymes involved in the triterpene skeleton β-amyrin synthesis step: farnesyl diphosphate synthase (FPS), squalene synthase (SQS), squalene monooxygenase and β-amyrin synthase (bAS). The enzymes involved in the biosyntheses of the isoprene unit and the triterpene skeleton are listed in Table 3. A list of putative unigenes involved in the glycyrrhizin biosynthetic pathway is shown in Additional file 3.

Cytochrome P450 and glycosyltransferase

Glycyrrhizin is derived from the triterpene β-amyrin, which is an initial from product of thethe cyclization of 2, 3-oxidosqualene. The subsequent steps in glycyrrhizin biosynthesis include a series of oxidative and glycosyl transfer reactions. We have little knowledge of the later steps in the glycyrrhizin biosynthetic pathway, which include multiple oxidation and glycosylation steps that are catalyzed by enzymes from the cytochrome P450 (CYP) and glycosyltransferase superfamilies, respectively.

Cytochrome P450 is a very large and diverse superfamily of hemoproteins that are found in all higher organisms [31,32]. Plant P450s catalyze many different reactions involved in the biosynthesis of secondary metabolites, including terpenoids [33]. Some members of the CYP88

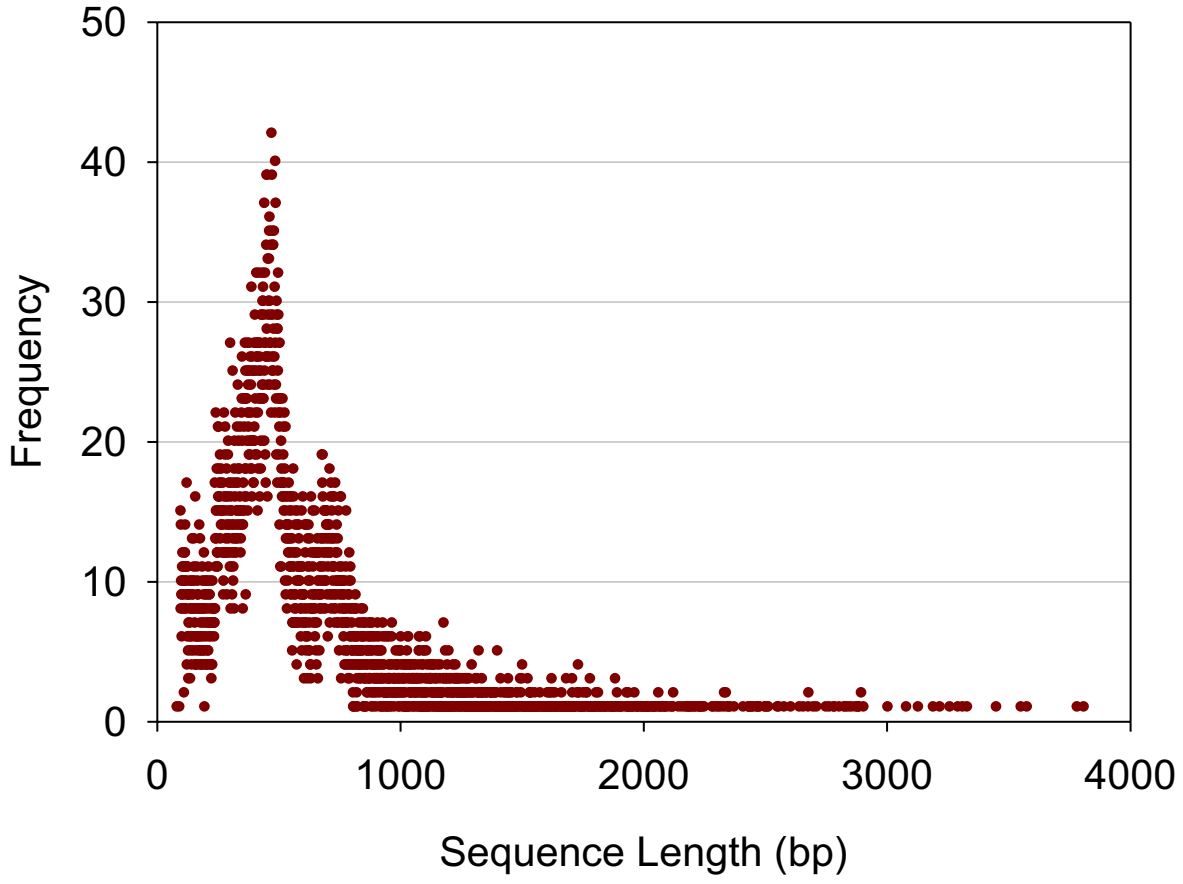

Figure 2 Contig length distribution.

and CYP93 families have been shown to act on β-amyrin or related triterpene substrates with unique reaction specificities [16,34-36]. Thus far, all known cytochrome P450s that act on triterpenes and sterols have been classified into two clans: the CYP71 clan and the CYP85 clan, which includes CYP93 and CYP88, respectively [16,34,37-39]. Only two *CYP* genes of *G. uralensis* have been identified [16]. An organ-specific transcript profiling approach was used in other studies to identify *CYP88D6*, which catalyzes the oxidation of β-amyrin at C-11 to produce 11-oxo-β-amyrin in the glycyrrhizin biosynthetic pathway. The expression profile of *CYP88D6* was consistent with the organ-specific accumulation pattern of glycyrrhizin [16]; a higher level of expression was seen in the root than in the stem and leaf. By mining the EST database, we found 125 unigenes (500 ESTs) annotated as putative *CYP* genes, which were further classified into 32 CYP families and 47 subfamilies (Additional file 4). To narrow down the candidate cytochrome P450s, these unigenes were further screened according to their classification. In the candidate P450 dataset, two unigenes (23 ESTs) were annotated as CYP88, while six unigenes (30 ESTs) were annotated as CYP93. In total, 29 contigs annotated as cytochrome P450 and belonging to the CYP71 and CYP85 clans (Additional file 5), were chosen for organ-specific expression pattern assays (Figure 5A). Contig01314 (No. 2 in Figure 5A) was exactly the same as the *CYP88D6* gene according to the BLAST annotation results. An additional 3 unigenes had a similar organ-specific expression pattern as *CYP88D6*, including contig06734 (No. 19 in Figure 5A), contig07137 (No. 20 in Figure 5A) and contig07899 (No. 23 in Figure 5A). However, additional experiments are needed to determine which of these unigenes participate in glycyrrhizin biosynthesis.

Glycosyltransferases, a ubiquitous family of enzymes, catalyze reactions involving the transfer of a nucleotide-activated sugar moiety onto another molecule [40]. These enzymes are encoded by a large multigene family; approximately 120 secondary metabolism glycosyltransferase genes have been identified in *Arabidopsis*. The conjugation of a sugar moiety to a substrate is called glycosylation, which is a process that contributes to the synthesis of glucidic polymers, glycoproteins and glycolipids.

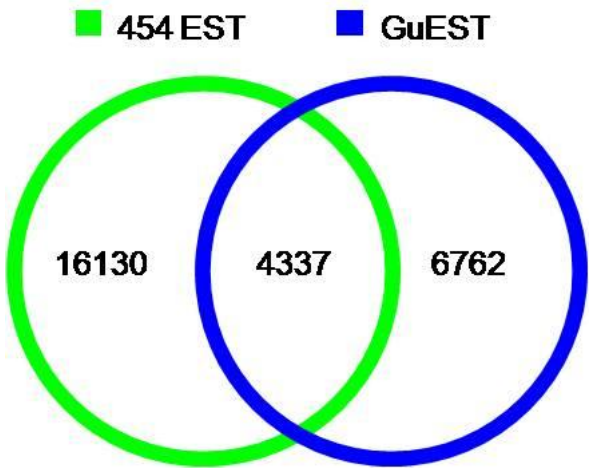

Figure 3 Comparison of the 454 ESTs and GuESTs. The overlapping section corresponds to unigenes composed of ESTs in the two datasets.

Glycosyltransferases often use specific substrates in the glycosylation reaction and are relevant for the synthesis of secondary metabolites. No genes encoding relevant glycosyltransferases have been identified in *Glycyrrhiza*.

Using BLAST searches, approximately 172 unigenes (1205 ESTs) in our study showed sequence similarities to glycosyltransferase (EC: 2.4.-.-) in the KEGG database. According to the GO category analysis, these unigenes were classified into 45 categories (Additional file 6). Among these categories, 27 unigenes (83 ESTs) encoded for UDP-glucosyltransferases, which is obviously involved in the biosynthesis of secondary metabolites. We also pick up unigenes annotated as glucuronosyltransferases because glycyrrhizin is composed of aglycone glycyrrhetic acid and two glucuronic acid units. We found 11 unigenes (33 ESTs) that encoded glucuronosyltransferases, and it is possible that these are involved in the last steps of glycyrrhizin biosynthesis. From these two categories, 17 contigs were chosen (Additional file 5) for organ-specific expression pattern analysis by real-time PCR (Figure 5B). The expression patterns of 6 glycosyltransferase unigenes were similar to that of *CYP88D6*. These glycosyltransferases included contig01209 (No. 3 in Figure 5B), contig03646 (No. 6 in Figure 5B), contig05219 (No. 11 in Figure 5B), contig09428 (No. 13 in Figure 5B), contig09463 (No. 14 in Figure 5B) and contig09686 (No. 15 in Figure 5B). These contigs were regarded as candidate glycosyltransferases that encode the enzymes responsible for glycyrrhizin biosynthesis and will be the subject of further study. We did not select singletons that were annotated as cytochrome P450s or glycosyltransferases for the organ-specific expression pattern analysis because of the high content of glycyrrhizin in the *Glycyrrhiza* plant (3 - 5% in the root) [30]. On the other hand, 22 ESTs were annotated as CYP88D6 (contig01314, No. 2 in Figure 5A), which is a known cytochrome P450 gene in the glycyrrhizin biosynthetic pathway. The lists of candidate unigenes for cytochrome P450s and glycosyltransferases are found in Additional files 7 and 8, respectively.

Conclusions

Our study established a high-quality EST database for *G. uralensis* using 454 GS FLX Titanium sequencing technology. With this work, we initiated a large-scale investigation of the transcriptome of *G. uralensis* in terms of functional genomics, molecular biology and biochemistry. A large number of novel candidate genes involved in glycyrrhizin biosynthesis, including cytochrome P450s and glycosyltransferases, were identified in our EST dataset. The information from these ESTs represents a significant contribution toward the exploration of the molecular mechanisms of glycyrrhizin biosynthesis. More importantly, a few candidate genes encoding the enzymes responsible for glycyrrhizin skeleton modifications were obtained by screening functional annotations and by organ-specific expression pattern analyses.

Methods

Plant materials

G. uralensis material was collected from a five-year-old, field-grown *G. uralensis* plant growing in Ningxia, China. Previous research has shown that wild *G. uralensis* contains much more glycyrrhizin than cultivated plants [41,42]. One possible reason for this difference is that under cultivated plant conditions, *G. uralensis* grows more vigorously and has a more active primary metabolism, while glycyrrhizin accumulation results from secondary metabolism. Wild *G. uralensis* primarily grows in dry areas where lean soil inhibits vegetative growth and thus favors the synthesis and accumulation of glycyrrhizin. Additional studies have shown that the glycyrrhizin content of the *G. uralensis* plant is related to its growth period [42,43]. In contrast to the chemical methods that are mainly used to investigate glycyrrhizin content and accumulation, our transcriptome sequencing method is designed to only reveal genes that are expressed during sampling. Therefore, to investigate the secondary metabolites of glycyrrhizin, plants should be sampled during the glycyrrhizin biosynthetic period. Since this period has not been well studied thus far, we decided to use five-year-old wild *G. uralensis*.

RNA extraction and cDNA library synthesis

A mixture of approximately 1 g of roots, 1 g of stems and 1 g of leaves was ground to a fine powder under liquid nitrogen. Total RNA purification was performed with the Universal Plant Total RNA Rapid Extraction kit (Bioteke,

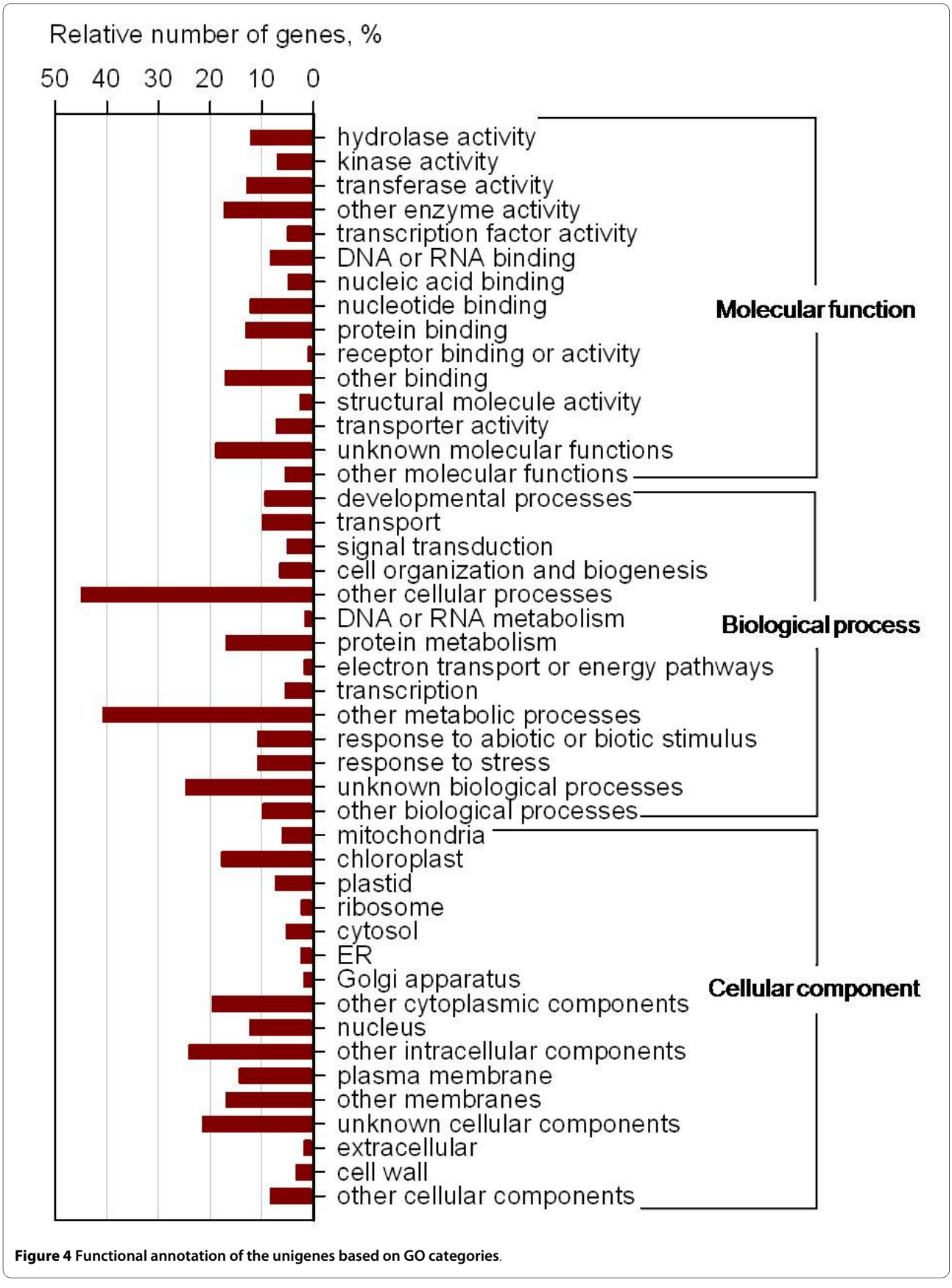

Figure 4 Functional annotation of the unigenes based on GO categories.

Table 3: Number of putative unigenes and ESTs involved in glycyrrhizin skeleton biosynthesis[a]

ECb	Enzyme name	Number of Unigenes	Number of ESTs	Number of 454 ESTs	Number of GuESTs
2.3.1.9	acetyl-CoA acetyltransferase	3	30	28	2
2.3.3.10	HMG-CoA synthase	2	11	5	6
1.1.1.34	HMG-CoA reductase	2	17	16	1
2.7.1.36	mevalonate kinase	0	0	0	0
2.7.4.2	phosphomevalonate kinase	2	5	4	1
4.1.1.33	mevalonate-5-diphosphate decarboxylase	1	1	1	0
2.2.1.7	DXP synthase	0	0	0	0
1.1.1.267	DXP reductoisomerase	3	6	4	2
2.7.7.60	MEP cytidylyltransferase	2	2	2	0
2.7.1.148	CDP-ME kinase	1	1	1	0
4.6.1.12	MECDP synthase	1	6	4	2
1.17.7.1	4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	2	25	25	0
1.17.1.2	4-hydroxy-3-methylbut-2-enyl-diphosphate reductase	1	24	20	4
5.3.3.2	isopentenyl-PP isomerase	1	63	31	32
2.5.1.10	farnesyl diphosphate synthase	1	1	1	0
2.5.1.21	squalene synthase	1	1	1	0
1.14.99.7	squalene monooxygenase	4	23	23	0
5.4.99.-	β-amyrin synthase	1	1	1	0

[a]BLAST against the SwissProt and KEGG databases; [b]Enzyme code

China). The poly(A) RNA was then separated from the total RNA using the Oligotex® mRNA kit (Qiagen). The quality and purity of the poly(A) RNA was analyzed with 1.0% agarose gels and a GE GeneQuant 100 Spectrophotometer. Ethidium bromide-stained mRNA appeared as a smear that ranged from 500 to 2,000 bp, and the ratio between the absorbance values at 260 and 280 nm was 2.03. Subsequently, 1 μg of mRNA was using the SMART™ PCR cDNA Synthesis kit (Clontech). The cDNA was amplified using the PCR Advantage II polymerase (Clontech) and the following thermal profile: 1 min at 95°C, and then 12 cycles of 95°C for 15 sec, 65°C for 30 sec and 68°C for 6 min. The amplified cDNA product was purified using the PureLink™ PCR Purification kit (Invitrogen) to remove fragments of less than 300 bp. Finally, approximately 5 μg of the resulting cDNA was used to construct a 454 library.

454 library preparation and sequencing

Sequencing was based on the 454 GS FLX platform and Titanium reagents. Preparation of the 454 library was performed according to the supplier's instructions. In summary, approximately 5 μg of amplified cDNA was nebulized and selected for length, which ranged from 300 to 800 bp. The FLX specific adapters, Adapter A (GCCTCCCTCGCGCCATCAG) and Adapter B (GCCTTGCCAGCCCGCTCAG), were added to each fragmented cDNA, resulting in Adapter A-DNA fragment-Adapter B constructs. The DNA fragments were then denatured to generate single-stranded DNA, which was then amplified by emulsion PCR for sequencing. The sequencing of the libraries was performed on a 454 GS FLX platform (454 Life Sciences, Roche).

Sequence assembly

All processing and analyses of the sequencing data was performed with the GS-FLX Software v2.0.01 (454 Life Sciences, Roche). Using a series of normalization, correction and quality-filtering algorithms, the 454 sequencing data were processed to screen and filter for weak signals and low-quality reads, and to trim the read ends for low-quality and 454 adaptor sequences. The resulting 59,219 HQ reads were then submitted to the Short Read Archive at NCBI and assigned the accession number SRX011915. The HQ reads were combined with the 50,666 *G. uralensis* ESTs in GenBank and filtered, clustered and assembled into transcript contigs using GS *De Novo* Assembler Software, which is an application of the GS FLX Software. The filtering step included the masking of SMART PCR primer sequences (Clontech) and the removing of reads

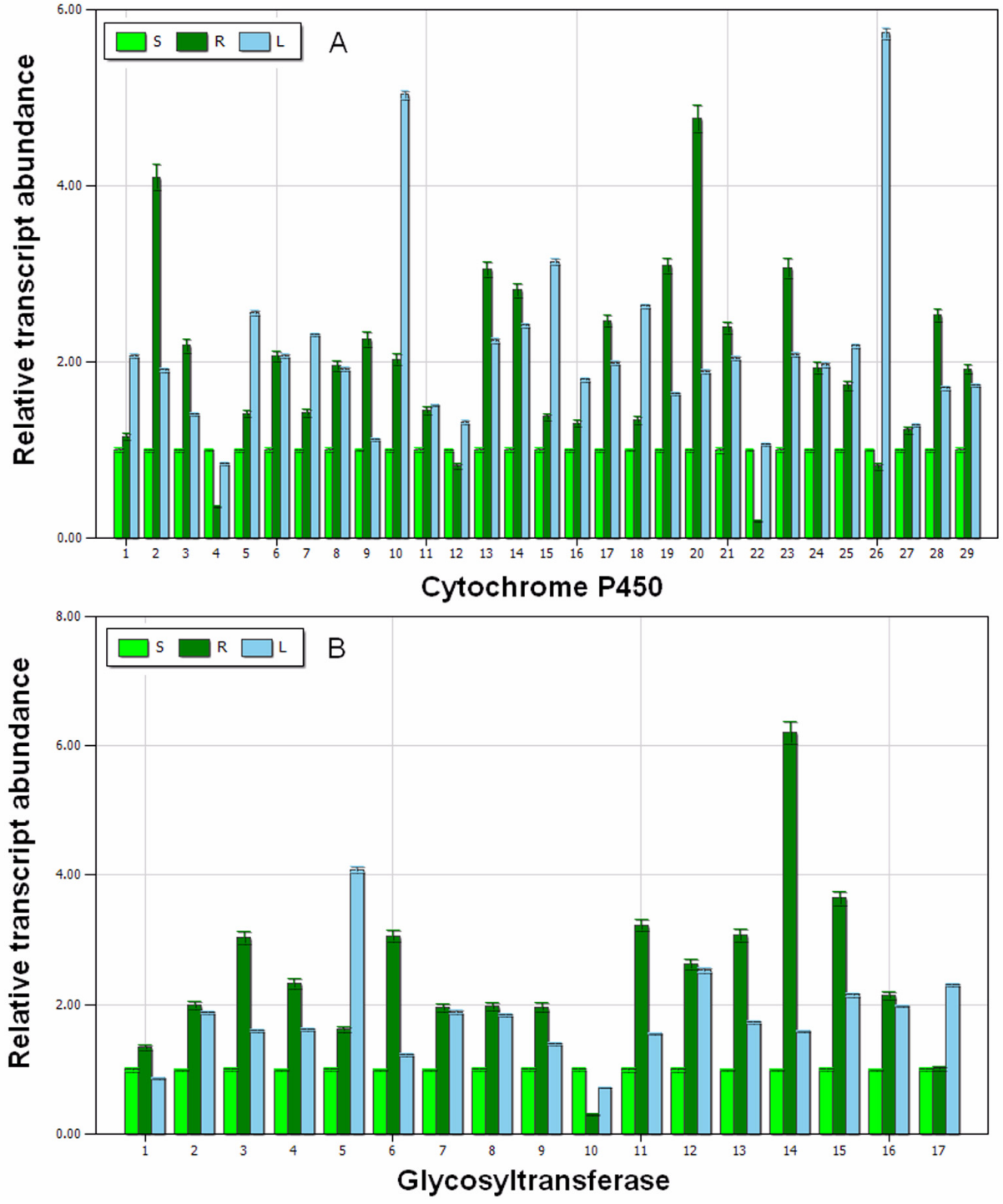

Figure 5 Real-time PCR analysis of cytochrome P450s (A) and glycosyltransferases (B) in different plant organs. R represents the root, S represents the stem, and L represents the leaf. The corresponding cytochrome P450 and glycosyltransferases contigs represented by the numbers are listed in Additional file 5. A) The gene expression of *cytochrome P450*s in different organs. B) The gene expression of *glycosyltransferases* in different organs.

that were shorter than 50 bases. The assembly was conducted using the default parameters. Reads that did not fit into a contig were defined as singletons. The resulting singletons and contigs (unigenes) represented the *G. uralensis* candidate gene set.

Functional annotation and GO classification

The unigenes were annotated by a BLAST (version Blast 2.2.17) [22] search against a series of protein and nucleotide databases, including the curated protein database of Uniprot/SwissProt (released on 06/19/2009) [23], the

KEGG GENE database (version KEGG 50) [24], the *Arabidopsis thaliana* proteome databases (version TAIR9) [25] and the NCBI non-redundant protein (Nr) [26] and nucleotide (Nt) [27] databases (released on 06/23/2009). The unigenes were compared against these databases with a significance threshold of e-value ≤ le-5. To maximize computational speed, the search was limited to the first five significant hits for each query. The definition line of the top BLAST hit was used as a description of the putative function of the queried unigene. Customized Perl scripts were used to parse the BLAST outputs.

The Gene Ontology annotations were assigned based on similarity to the *A. thaliana* proteomic sequences (TAIR9) [25,29]. This database was chosen because it has been extensively annotated in GO terms. Each of the unigenes was assigned a GO term based on the top BLAST hit for that query. The transcripts were classified into 45 GO categories under the major categories of Cellular Component, Molecular Function and Biological Process.

Gene discovery and classification for glycyrrhizin biosynthesis

To evaluate the completeness of our transcriptome library and the effectiveness of our annotation procedure, we searched the annotated sequences for genes involved in the glycyrrhizin metabolic pathway. These simple text searches were based on standard gene names or synonyms.

Real-time PCR

The mRNA levels of selected *cytochrome P450s* and *glycosyltransferases* genes in different *G. uralensis* organ types were analyzed by RT-PCR. Reverse transcription was performed with DNase I-treated total RNA of *G. uralensis* roots, stems and leaves using the PrimeScript™ 1st Strand cDNA Synthesis Kit (TaKaRa, Dalian, China). The quantitative reaction was performed on an IQ5 Multicolor Real-Time PCR Detection System (Bio-Rad, USA) using SYBR Premix Ex Taq™ (TaKaRa, Dalian, China). PCR amplification was performed under the following conditions: 2 min at 50°C and 30 sec at 95°C, and then 40 cycles of 95°C for 15 sec and 62°C for 1 min. The gene expression of *cytochrome P450s* and *glycosyltransferases* was normalized against an internal reference gene, *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*), which was found in our EST library. All primers used in this study are listed in Additional file 5.

List of abbreviations

cDNA: complementary DNA; EST: expressed sequence tag; bp: base pairs; GuEST: *G. uralensis* ESTs derived from GenBank; BLAST: Basic Local Alignment Search Tool; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; NCBI: National Center for Biotechnology Information; Nr: non-redundant protein data bank; Nt: Entrez nucleotide database; TAIR: The Arabidopsis Information Resource.

Additional material

Additional file 1 Percentages of EST having hits in major public databases. Word document containing the hit numbers and percentages relative to those of the major public databases, including SwissProt, KEGG, TAIR, Nr and Nt.

Additional file 2 The putative glycyrrhizin biosynthetic pathway. Word document containing the putative glycyrrhizin biosynthetic pathway of *G. uralensis*.

Additional file 3 Gene discovery for glycyrrhizin skeleton synthesis. Excel document containing the annotations of putative genes corresponding to the glycyrrhizin skeleton synthesis.

Additional file 4 Classification of the candidate P450 genes. Word document containing the classification of the candidate P450 genes by CYP families.

Additional file 5 Primers used in this study. Excel document containing the primers used in this study.

Additional file 6 Classification of the candidate glycosyltransferase genes. Word document containing the classification of the candidate glycosyltransferase genes according to the GO category.

Additional file 7 Cytochrome P450 gene discovery. Excel document containing the annotations of putative genes annotated as cytochrome P450.

Additional file 8 Glycosyltransferase gene discovery. Excel document containing the annotations of putative genes annotated as glycosyltransferase.

Authors' contributions

YL contributed to the tissue sample collection, RNA extraction, cDNA library construction, bioinformatic analysis and writing of the manuscript. HML helped with the RNA extraction, construction of the cDNA library and writing of the manuscript. CS helped to prepare the first draft of the manuscript and discussed the results. JYS initiated the EST project and helped with the construction of the cDNA library. YZS performed the real-time PCR and the corresponding data analysis. QW aided in the RNA extraction. NW contributed to revisions of the manuscript. HY helped with the RNA extraction and the construction of the cDNA library. AS contributed to the discussion of the gene candidates for the biosynthesis of secondary metabolites and the revisions of the manuscript. This work was conducted in the laboratory of SLC, who initiated the 454 sequencing project and contributed to the evaluation and discussion of the results, as well as to the revisions of the manuscript. All authors contributed to the content of the manuscript, and have read and approved the final version.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (30772735). We thank Prof. Jun Wang, Mr. Qing Zhu and Mr. Ping Zhang (School of Biological Science, Ningxia University, China) for their kind help with the *G. uralensis* sampling in the Ningxia province. We are grateful to Mr. Hai-Bo Sun and Mr. Tao Feng (MininGene Biotechnology Co. Ltd, Beijing, China) for their help with the bioinformatic analysis. We are also grateful to laboratory assistant Yunyun Niu for her help with the experiments.

Author Details

[1]Institute of Medicinal Plant Development (IMPLAD), Chinese Academy of Medical Sciences & Peking Union Medical College, No.151, Malianwa North Road, HaiDian District, Beijing 100193, China, [2]Hubei College of Traditional Chinese Medicine, No. 1, Huangjia Lake West Road, Hongshan District, Wuhan City, Hubei Province 430065, China and [3]Centre de Recherche Public-Santé, Luxembourg, L-1526 Luxembourg

Received: 8 November 2009 Accepted: 28 April 2010
Published: 28 April 2010

References

1. Parkinson J: *Expressed Sequence Tags (ESTs) Generation and Analysis* Humana Press; 2009.
2. Brautigam A, Shrestha RP, Whitten D, Wilkerson CG, Carr KM, Froehlich JE, Weber AP: **Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in non-model species: comparison of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes.** *J Biotechnol* 2008, **136(1-2):**44-53.
3. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17(7):**1636-1647.
4. Zhang HC, Liu JM, Lu HY, Gao SL: **Enhanced flavonoid production in hairy root cultures of Glycyrrhiza uralensis Fisch by combining the over-expression of chalcone isomerase gene with the elicitation treatment.** *Plant Cell Rep* 2009, **28(8):**1205-1213.
5. Shibata S: **A drug over the millennia: pharmacognosy, chemistry, and pharmacology of licorice.** *Yakugaku Zasshi* 2000, **120(10):**849-862.
6. van Rossum TG, Vulto AG, de Man RA, Brouwer JT, Schalm SW: **Review article: glycyrrhizin as a potential treatment for chronic hepatitis C.** *Aliment Pharmacol Ther* 1998, **12(3):**199-205.
7. He JX, Akao T, Nishino T, Tani T: **The influence of commonly prescribed synthetic drugs for peptic ulcer on the pharmacokinetic fate of glycyrrhizin from Shaoyao-Gancao-tang.** *Biol Pharm Bull* 2001, **24(12):**1395-1399.
8. Matsui S, Matsumoto H, Sonoda Y, Ando K, Aizu-Yokota E, Sato T, Kasahara T: **Glycyrrhizin and related compounds down-regulate production of inflammatory chemokines IL-8 and eotaxin 1 in a human lung fibroblast cell line.** *Int Immunopharmacol* 2004, **4(13):**1633-1644.
9. Li W, Asada Y, Yoshikawa T: **Flavonoid constituents from Glycyrrhiza glabra hairy root cultures.** *Phytochemistry* 2000, **55(5):**447-456.
10. Pompei R, Flore O, Marccialis MA, Pani A, Loddo B: **Glycyrrhizic acid inhibits virus growth and inactivates virus particles.** *Nature* 1979, **281(5733):**689-690.
11. Ito M, Nakashima H, Baba M, Pauwels R, De Clercq E, Shigeta S, Yamamoto N: **Inhibitory effect of glycyrrhizin on the in vitro infectivity and cytopathic activity of the human immunodeficiency virus [HIV (HTLV-III/LAV)].** *Antiviral Res* 1987, **7(3):**127-137.
12. Ito M, Sato A, Hirabayashi K, Tanabe F, Shigeta S, Baba M, De Clercq E, Nakashima H, Yamamoto N: **Mechanism of inhibitory effect of glycyrrhizin on replication of human immunodeficiency virus (HIV).** *Antiviral Res* 1988, **10(6):**289-298.
13. Cinatl J, Morgenstern B, Bauer G, Chandra P, Rabenau H, Doerr HW: **Glycyrrhizin, an active component of liquorice roots, and replication of SARS-associated coronavirus.** *Lancet* 2003, **361(9374):**2045-2046.
14. Hanrahan C: *Gale Encyclopedia of Alternative Medicine: Licorice [book on CD-ROM]* Farmington Hills, Thomson Gale; 2001.
15. Lu H-Y, Liu J-M, Zhang H-C, Yin T, Gao S-L: **Ri-mediated Transformation of Glycyrrhiza uralensis with a Squalene Synthase Gene (GuSQS1) for Production of Glycyrrhizin.** *Plant Mol Biol Rep* 2008, **26:**1-11.
16. Seki H, Ohyama K, Sawai S, Mizutani M, Ohnishi T, Sudo H, Akashi T, Aoki T, Saito K, Muranaka T: **Licorice beta-amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin.** *Proc Natl Acad Sci USA* 2008, **105(37):**14204-14209.
17. Hayashi H, Hirota A, Hiraoka N, Ikeshiro Y: **Molecular cloning and characterization of two cDNAs for Glycyrrhiza glabra squalene synthase.** *Biol Pharm Bull* 1999, **22(9):**947-950.
18. Hayashi H, Huang P, Kirakosyan A, Inoue K, Hiraoka N, Ikeshiro Y, Kushiro T, Shibuya M, Ebizuka Y: **Cloning and characterization of a cDNA encoding beta-amyrin synthase involved in glycyrrhizin and soyasaponin biosyntheses in licorice.** *Biol Pharm Bull* 2001, **24(8):**912-916.
19. Nagashima S, Inagaki R, Kubo A, Hirotani M, Yoshikawa T: **cDNA cloning and expression of isoflavonoid-specific glucosyltransferase from Glycyrrhiza echinata cell-suspension cultures.** *Planta* 2004, **218(3):**456-459.
20. Sudo H, Seki H, Sakurai N, Suzuki H, Shibata D, Toyoda A, Totoki Y, Sakaki Y, Iida O, Shibata T, *et al.*: **Expressed sequence tags from rhizomes of Glycyrrhiza uralensis.** *Plant Biotechnology* 2009, **26(1):**105-107.
21. Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM: **Shedding light on an extremophile lifestyle through transcriptomics.** *New Phytol* 2009, **183(3):**764-775.
22. **Basic Local Alignment Search Tool** [ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.17/]
23. **The UniProt-SwissProt Database** [http://www.uniprot.org/downloads]
24. **The KEGG Database** [ftp://ftp.genome.jp/pub/kegg/release/archive/kegg/50/]
25. **The TAIR Database** [ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR9_blastsets/]
26. **NCBI Nr Database** [ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz]
27. **NCBI Nt Database** [ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nt.gz]
28. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, *et al.*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Res* 2003, **31(19):**5654-5666.
29. Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G, *et al.*: **Functional annotation of the Arabidopsis genome using controlled vocabularies.** *Plant Physiol* 2004, **135(2):**745-755.
30. Dewick PM: *Medicinal Natural Products: A Biosynthetic Approach* Wiley; 2009.
31. Sigel A, Sigel H, Sigel RKO: *The Ubiquitous Roles of Cytochrome P450 Proteins: Metal Ions in Life Sciences* Wiley; 2007.
32. Danielson PB: **The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans.** *Curr Drug Metab* 2002, **3(6):**561-597.
33. Kim GT, Tsukaya H: **Regulation of the biosynthesis of plant hormones by cytochrome P450s.** *J Plant Res* 2002, **115(3):**169-177.
34. Shibuya M, Hoshino M, Katsube Y, Hayashi H, Kushiro T, Ebizuka Y: **Identification of beta-amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay.** *Febs J* 2006, **273(5):**948-959.
35. Winkler RG, Helentjaris T: **The maize Dwarf3 gene encodes a cytochrome P450-mediated early step in Gibberellin biosynthesis.** *Plant Cell* 1995, **7(8):**1307-1317.
36. Davidson SE, Elliott RC, Helliwell CA, Poole AT, Reid JB: **The pea gene NA encodes ent-kaurenoic acid oxidase.** *Plant Physiol* 2003, **131(1):**335-344.
37. Fujita S, Ohnishi T, Watanabe B, Yokota T, Takatsuto S, Fujioka S, Yoshida S, Sakata K, Mizutani M: **Arabidopsis CYP90B1 catalyses the early C-22 hydroxylation of C27, C28 and C29 sterols.** *Plant J* 2006, **45(5):**765-774.
38. Shimada Y, Fujioka S, Miyauchi N, Kushiro M, Takatsuto S, Nomura T, Yokota T, Kamiya Y, Bishop GJ, Yoshida S: **Brassinosteroid-6-oxidases from Arabidopsis and tomato catalyze multiple C-6 oxidations in brassinosteroid biosynthesis.** *Plant Physiol* 2001, **126(2):**770-779.
39. Bishop GJ, Nomura T, Yokota T, Harrison K, Noguchi T, Fujioka S, Takatsuto S, Jones JD, Kamiya Y: **The tomato DWARF enzyme catalyses C-6 oxidation in brassinosteroid biosynthesis.** *Proc Natl Acad Sci USA* 1999, **96(4):**1761-1766.
40. Gachon CM, Langlois-Meurinne M, Saindrenan P: **Plant secondary metabolism glycosyltransferases: the emerging functional analysis.** *Trends Plant Sci* 2005, **10(11):**542-549.
41. Yan YH, Duan TX, Wang WQ: **Studies on the HPLC fingerprint of Radix Glycyrrhizae.** *Chin J Nat Med* 2006, **14(12):**116-120.
42. Fu YJ: *GANCAO: The Chinese Licorice* Beijing/New York, Science Press; 2004.
43. Sun L, Yu JG, Li DY, Luo XZ, Zhao CJ, Yang SL: **Comparison of the content of glycyrrhizin and liquiritin of wild and cultivated root of Glycyrrhiza uralensis Fisch.** *Journal of Chinese Medicinal Materials* 2001, **24(8):**550-552.

doi: 10.1186/1471-2164-11-268
Cite this article as: Li *et al.*, EST analysis reveals putative genes involved in glycyrrhizin biosynthesis *BMC Genomics* 2010, **11**:268