Genomics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Analysis of the transcriptome of *Marsdenia tenacissima* discovers putative polyoxypregnane glycoside biosynthetic genes and genetic markers

Kaiyan Zheng [a], Guanghui Zhang [b], Nihao Jiang [b], Shengchao Yang [b], Chao Li [a], Zhengui Meng [b], Qiaosheng Guo [a,*], Guangqiang Long [b,**]

[a] Institute of Chinese Medicinal Materials, Nanjing Agricultural University, Nanjing 210095, Jiangsu, People's Republic of China
[b] Yunnan Research Center on Good Agricultural Practice for Dominant Chinese Medicinal Materials, Yunnan Agricultural University, Kunming 650201, Yunnan, People's Republic of China

ARTICLE INFO

Article history:
Received 18 March 2014
Accepted 25 July 2014
Available online xxxx

Keywords:
Marsdenia tenacissima
Transcriptome
Polyoxypregnane glycosides
Biosynthesis

ABSTRACT

Marsdenia tenacissima is a well-known anti-cancer medicinal plant used in traditional Chinese medicine due to bioactive constituents of polyoxypregnane glycosides, such as tenacissosides, marsdenosides and tenacigenosides. Genomic information regarding this plant is very limited, and rare information is available about the biosynthesis of polyoxypregnane glycosides. To facilitate the basic understanding about the polyoxypregnane glycoside biosynthetic pathways, de novo assembling was performed to generate a total of 73,336 contigs and 65,796 unigenes, which represent the first transcriptome of this species. These included 27 unigenes that were involved in steroid biosynthesis and could be related to pregnane backbone biosynthesis. The expression patterns of six unigenes involved in polyoxypregnane biosynthesis were analyzed in leaf and stem tissues by quantitative real time PCR (qRT-PCR) to explore their putative function. Furthermore, a total of 15,295 simple sequence repeats (SSRs) were identified from 11,911 unigenes, of which di-nucleotide motifs were the most abundant.

1. Introduction

Marsdenia tenacissima (Roxb.) Wight et Arn. is a perennial climber belonging to the Asclepiadaceae family, which is widely distributed in tropical to subtropical areas in Asia, mainly in the Guizhou and Yunnan Provinces of China. The dried stems of *M. tenacissima*, known as "Tong-guang-teng" or "Tong-guang-san", are used in Chinese folk medicine for the treatment of asthma, cancer, tracheitis, tonsillitis, pharyngitis, cystitis, and pneumonia [1,2]. Clinical studies have shown that the aqueous extractions of *M. tenacissima* are beneficial for treating patients with various cancers [3–5]. Polyoxypregnane glycosides are the major bioactive constituents in the stem of *M. tenacissima* [6]. More than 40 polyoxypregnane glycosides have been isolated from *M. tenacissima*, mainly tenacissosides [7], marsdenosides [8–11] and tenacigenosides [12–14], and all of which have aglycones derived from tenacigenin B. Two other polyoxypregnane glycosides with aglycones of sarcogein and drevogenin P were also detected from *M. tenacissima* [15] (Fig. 1). The main biosynthetic pathway of phytosterol has been studied extensively and is well understood [16–20], but the biosynthesis of steroidal derivatives as secondary metabolites is still largely unknown, especially pregnane and their glycosides.

Pregnane glycosides are C-21 steroidal compounds conjugated with sugars [21]. In plants, pregnane derivatives are intermediates in cardenolide glycoside biosynthesis where cholesterol is a direct precursor [22,23]. The biosynthesis of cardenolide glycosides has been sufficiently elucidated [24], and most of the enzymes and genes involved in this pathway are well characterized [25–33]; however, there are some genes whose functions are still not clear, such as cholesterol monooxygenase (side chain-cleaving enzyme), Δ5-Δ4-ketosteroid isomerase and pregnane 14β-hydroxylase [24]. Comparing the molecular structures with cardenolide glycosides [24], we explored the putative biosynthetic pathway of polyoxypregnane glycosides in *M. tenacissima* (Fig. 1). Clearly, pregnanes must be modified by hydroxylation, acylation and glycosylation at C-atoms in its backbone for the formation of polyoxypregnane glycosides. Currently, only enzymes that catalyze those modifications at C-atoms in the side chain of sterols have been identified and characterized, including modifications at C-21 [32], C-22 [34–37] and C-24 [38–41]; however, little is known about the molecular mechanism of the modification of C-atoms in the backbone of sterols in plants (Fig. 1).

Moreover, cholesterol and other phytosterols, such as campesterol and sitosterol, are biosynthesized via cycloartenol and catalyzed by cycloartenol synthase (*CAS*) in higher plants (cycloartenol pathway), contributing to membrane sterol biosynthesis. New evidence has suggested that another route (the lanosterol pathway) catalyzed by lanosterol synthase (*LAS*) might contribute to the biosynthesis of not only phytosterols but also steroids as secondary metabolites [42].

* Corresponding author. Fax: +86 25 84395980.
** Corresponding author.
E-mail addresses: gqs@njau.edu.cn (Q. Guo), ynaulong2316@163.com (G. Long).

http://dx.doi.org/10.1016/j.ygeno.2014.07.013
0888-7543/

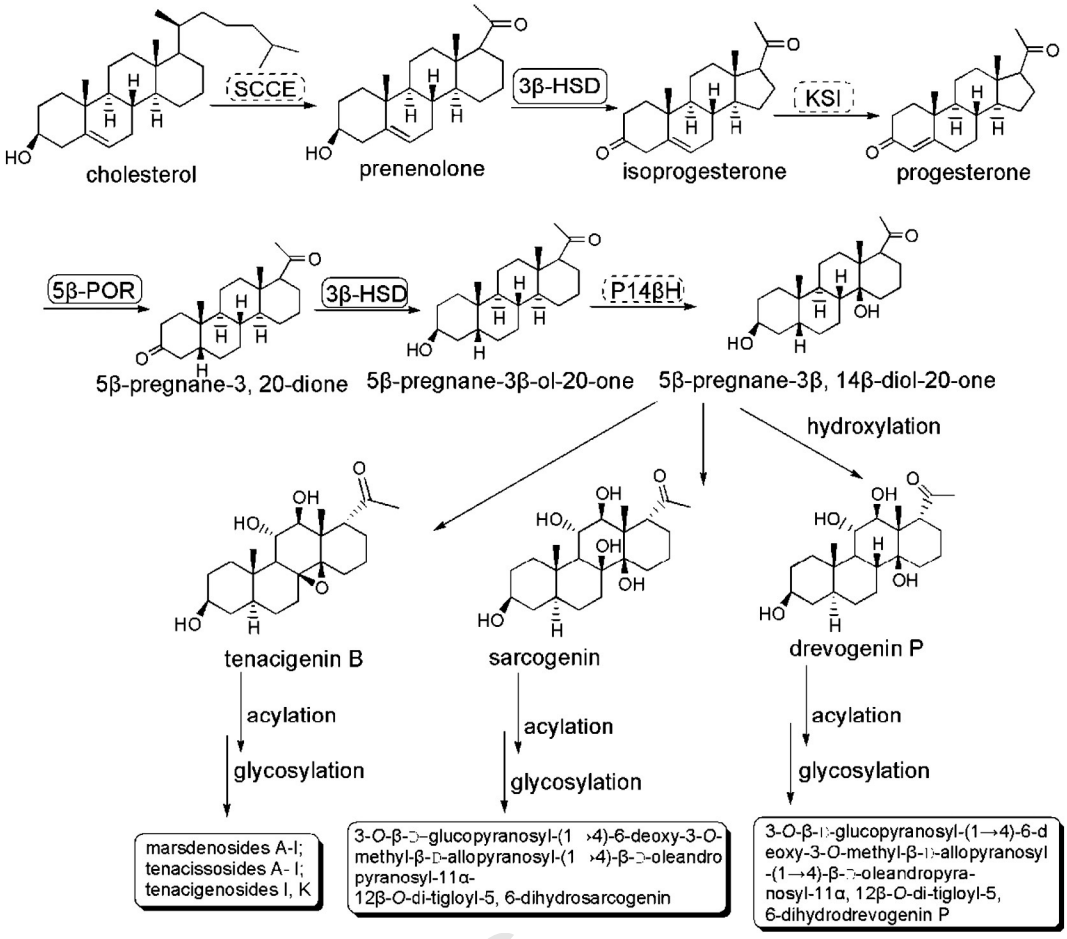

Fig. 1. The putative biosynthetic pathway of polyoxypregnane glycosides in *M. tenacissima*. Enzymes found in this study are surrounded by boxes; enzymes which are not found or hypothetical are surrounded by dashed boxes. Enzymes involved in the pathways are: SCCE, cholesterol monooxygenase (side-chain-cleaving enzyme); *3β-HSD*, 3β-hydroxysteroid dehydrogenase; KSI, Δ5-Δ4-ketosteroid isomerase; *5β-POR*, progesterone 5β-reductase; P14βH, pregnane 14β-hydroxylase (hypothetical). This putative biosynthetic pathway is modified according to Kreis and Müller-Uri [24].

Although the biosynthetic steps involved in the conversion of lanosterol to cholesterol have been postulated [43], most enzymes and their genes have not been identified or characterized.

RNA-seq has been widely used for de novo transcriptome sequencing in many medicinal plants. The objective of the present study is to analyze the transcriptome of *M. tenacissima* using Illumina paired-end sequencing technology on a HiSeq 2000 platform to discover candidate genes that encode enzymes involved in polyoxypregnane glycoside biosynthesis. Based on RNA-seq, many simple sequence repeat (SSR) markers were found, which will facilitate marker-assisted breeding of this plant.

2. Results and discussion

2.1. Illumina sequencing and de novo assembly

To obtain a comprehensive *M. tenacissima* transcriptome, cDNA libraries were generated from an equal mixture of RNA extracted from fresh leaves or stems and were paired-end sequenced using an Illumina HiSeq 2000 platform. After quality assessment and data cleaning, 63,175,764 high-quality reads were generated, comprising a total length of 6,317,576,400 nucleotides. Among these clean reads, 95.15% of reads had Q20 bases (base quality more than 20) and 46.83% GC-content. Based on high-quality reads, we obtained 73,336 contigs with lengths ranging from 201 bp to 15,808 bp with an average of 1123 bp; 43.77% of contigs were longer than 1000 bp (Fig. 2). After all the clean reads were assembled using the Trinity assembling program, de novo assembly yielded 65,796 unigenes with an average of 1087 bp, and 27,347 unigenes (41.56%) were longer than 1000 bp (Fig. 2). The sequences of all unigenes are shown in the NCBI SRA database. Of the

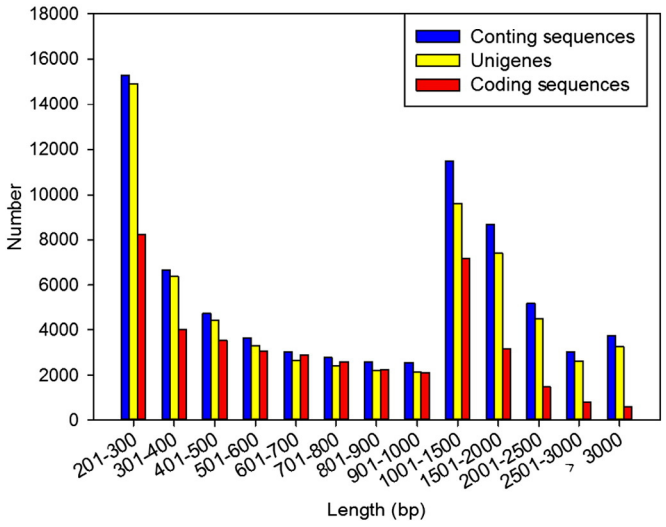

Fig. 2. Overview of the *M. tenacissima* transcriptome assembly and the length distribution of the CDS. (Blue) Length distribution of contig sequences. (Yellow) Length distribution of unigenes. (Red) Length distribution of the coding sequence (CDS). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

total number of unigenes identified, 63.66% (41,883) were identified as having a CDS. The size distribution showed that 68.40% of CDS (28,650) ranged from 201 to 1000 bp and that the percentage of CDS longer than 1000 bp was 31.60% (13,233) (Fig. 2). The high-quality reads produced in this study have been deposited in the NCBI SRA database (accession number: SRA140234).

2.2. Functional annotation

To obtain complete annotations, all of the assembled unigenes were similarity searched against four public protein databases (NCBI Nr, SwissProt, KEGG and COG). A total of 42,258 unigenes (64.23%) were annotated in the public databases Nr, SwissProt, KEGG and COG. Of these, 8299 unigenes were annotated as common in the four public databases. Unigenes that were annotated as unique in public databases are as follows: 8596 unigenes in the Nr database, 27 unigenes in the SwissProt database, 2 unigenes in the COG database and 45 unigenes in the KEGG database.

The results showed that approximately 75% of unigenes over 1000 bp in length had BLAST matches against the Nr database, whereas only 60% of unigenes with lengths shorter than 1000 bp generated BLAST matches. The same tendency was also observed in BLAST results against the SwissProt database. The statistical analysis of the E-value features that were distributed in the Nr databases revealed that 36.16% of the mapped unigenes showed significant homology (E-value < 10^{-50}) and 18.78% showed high similarity (E-value < 10^{-100}) to the available plant sequences. In contrast, the E-value and similarity distributions of the SwissProt database were 50.48% and 10.30%, respectively. Furthermore, identified unigenes were compared to sequences from *Arabidopsis thaliana*, *Glycine max*, *Medicago truncatula*, and *Vitis vinifera*. These results revealed that the transcriptome sequences of *M. tenacissima* showed high similarity to *V. vinifera* (34.23%), *A. thaliana* (12.14%), *G. max* (11.29%), and *M. truncatula* (8.86%).

2.3. Gene ontology classification

A total of 16,642 unigenes were characterized using gene ontology (GO) analysis based on Nr annotation, including biological process, cellular component, and molecular function. Under the biological process category, metabolic process (8427, 49.56%), cellular process (7687, 46.19%), and response to stimulus (3297, 19.81%) were prominently represented. In the cellular component group, unique sequences related to cell (11,279, 67.77%), cell part (11,279, 67.77%), organelle (8418, 50.56%), and organelle part (2331, 14.01%) were well-represented categories. For the molecular function category, catalytic activity (8228, 49.44%) and binding (8606, 48.47%) represented the majority of unique sequences (Fig. 3). These GO annotations provide comprehensive information on specific biological processes, molecular functions, and cellular structures of *M. tenacissima* transcripts and may lead to the identification of novel genes involved in secondary metabolite synthesis pathways (Additional file 1).

2.4. Functional classification by COG

All unigenes were also subjected to a search against the COG database for functional prediction and classification. In total, 17,587 unigenes were assigned to one or more of the 25 COG classification categories. Among the 25 COG categories, the largest cluster was predicted as general function (5749, 32.69%), followed by replication, recombination and repair (2752, 15.65%), transcription (2722, 15.48%), signal transduction mechanisms (2185, 12.42%), posttranslational modification, protein turnover and chaperones (2047, 11.64%), carbohydrate transport and metabolism (1886, 10.72%), translation, ribosomal structure and biogenesis (1760, 10.01%), amino acid transport and metabolism (1434, 8.15%), and function unknown (1416, 8.05%). The remaining categories account for only a small proportion of COG classifications.

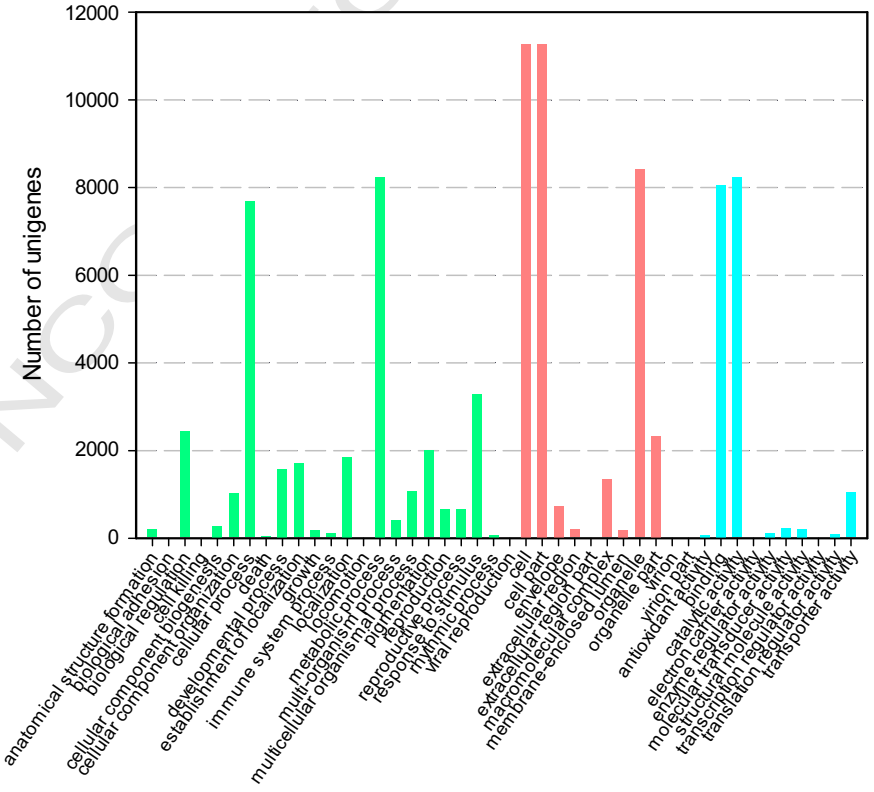

Fig. 3. Gene ontology classification of the *M. tenacissima* transcriptome. 16,642 unigenes with BLASTX matches against the plant NR database were classified into three main GO categories (biological process, cellular component, molecular function). The scale on the y-axis indicates the number of unigenes in the same category.

Please cite this article as: K. Zheng, et al., Genomics (2014), http://dx.doi.org/10.1016/j.ygeno.2014.07.013

2.5. Functional classification by KEGG

To identify active biological pathways in *M. tenacissima*, a total of 16,778 unigenes had significant matches in the KEGG database with corresponding enzyme commission (EC) numbers from BLASTX alignments and were assigned to 124 KEGG pathways (Additional file 2). Metabolic pathways had the largest number of unigenes (3478, 20.73%) followed by biosynthesis of secondary metabolites (1555, 9.27%), carbohydrate, starch and sucrose metabolism (335, 2.00%), nucleotide and purine metabolism (334, 1.99%) and translation and RNA transport (319, 1.90%). Among them, approximately 6814 unigenes were assigned to metabolic pathways, followed by carbohydrate metabolism (2007, 11.96%), amino acid metabolism (1320, 7.87%), lipid metabolism (827, 4.39%), energy metabolism (590, 3.52%), and nucleotide metabolism (569, 3.39%). Furthermore, it is worth noting that 827 unigenes were assigned to lipid biosynthetic pathways, the most represented categories of which were glycerophospholipid metabolism (196, 1.17%), fatty acid metabolism (104, 0.62%), glycerolipid metabolism (92, 0.55%), linoleic acid metabolism (67, 0.40%), and steroid biosynthesis (27, 0.16%), which could be related to pregnane backbone biosynthesis (Fig. 4). In addition to metabolism pathways, genes corresponding to genetic information processing (3359) and cellular processes (561) were highly represented categories. There were 27 unigenes (0.16), associated with steroid biosynthesis.

2.6. Candidate gene encoding enzymes involved in pregnane backbone biosynthesis

Like cardenolides, pregnanes are steroids and supposed to be derived from the mevalonate pathway via triterpenoid and phytosterol intermediates. Previous studies have shown that heterologous expression of the *A. thaliana HMGCR* (3-hydroxy-3-methylglutaryl-CoA reductase) gene in *Digitalis minor* could increase the content of cardenolide and phytosterol [31]. Based on the KEGG pathway annotation, we found that all of the gene encoding enzymes involved in the mevalonate pathway and farnesyl diphosphate biosynthesis in this study, including *ACAT* (acetyl-CoA acetyltransferase), *HMGCS* (hydroxymethylglutaryl-CoA synthase), *HMGCR*, *MVK* (mevalonate kinase), *PMVK* (phosphomevalonate kinase), *MVD* (mevalonate pyrophosphate decarboxylase), *IDI* (isopentenyl diphosphate isomerase), *FPS* (farnesyl diphosphate synthase), *SQS* (squalene synthase), and *SQLE* (squalene epoxidase) (Table 1; Additional file 3). This result might help us further understand polyoxypregnane glycoside biosynthetic mechanisms and increase their level of accumulation by overexpressing these genes in *M. tenacissima*.

Cholesterol is the direct precursor for pregnane biosynthesis, which comes from the lanosterol pathway [42]. Most gene encoding enzymes involved in cholestenol and pregnane backbone biosynthesis were found in this study, including *LAS* (lanosterol synthase), *14-SDM*

Table 1
Transcripts involved in pregnane derivatives biosynthesis in *Marsdenia tenacissima*.

Gene name	EC number	Unigene numbers
Mevalonate pathway and farnesyl diphosphate biosynthesis		
ACAT, acetyl-CoA acetyltransferase	2.3.1.9	17
HMGCS, hydroxymethylglutaryl-CoA synthase	2.3.3.10	1
HMGCR, 3-hydroxy-3-methylglutaryl-CoA reductase	1.1.1.34/1.1.1.88	3
MVK, mevalonate kinase	2.7.1.36	1
PMVK, phosphomevalonate kinase	2.7.4.2	3
MVD, mevalonate pyrophosphate decarboxylase	4.1.1.33	2
IDI, isopentenyl diphosphate isomerase	5.3.3.2	1
FPS, farnesyl diphosphate synthase	2.5.1.1/2.5.1.10	5
SQS, squalene synthase	2.5.1.21	2
SQLE, squalene epoxidase	1.14.13.132/1.14.99.7	4
Cholestenol biosynthesis		
LAS, lanosterol synthase	5.4.99.7	1
14-SDM, sterol 14α-demethylase (CYP51)	1.14.13.70	3
14SR, Δ14-sterol reductase	1.3.1.70	3
4-MSO, C4-methylsterol oxidase	1.14.13.72	3
EBP, cholestenol Δ-isomerase	5.3.3.5	1
DHCR24, Δ24-sterol reductase	1.3.1.72	1
SC5DL, sterol C5 desaturase/lathosterol oxidase	1.14.21.6	2
DHCR7, 7-dehydrocholesterol reductase	1.3.1.21	2
Pregnane derivatives biosynthesis		
3β-HSD, 3β-hydroxysteroid dehydrogenase	1.1.1.145	8
KSI, Δ5-Δ4-ketosteroid isomerase (delta 5-delta 4-steroid isomerase)	5.3.3.1	4
5β-POR, progesterone 5β-reductase	1.3.1.3	4
SOAT, sterol *O*-acyltransferase = Acyl-CoA cholesterol acyltransferase = Acyl-CoA cholesterin acyltransferase	2.3.1.26	14
Sterol 3-*O*-glucosyltransferase	2.4.1.173	40

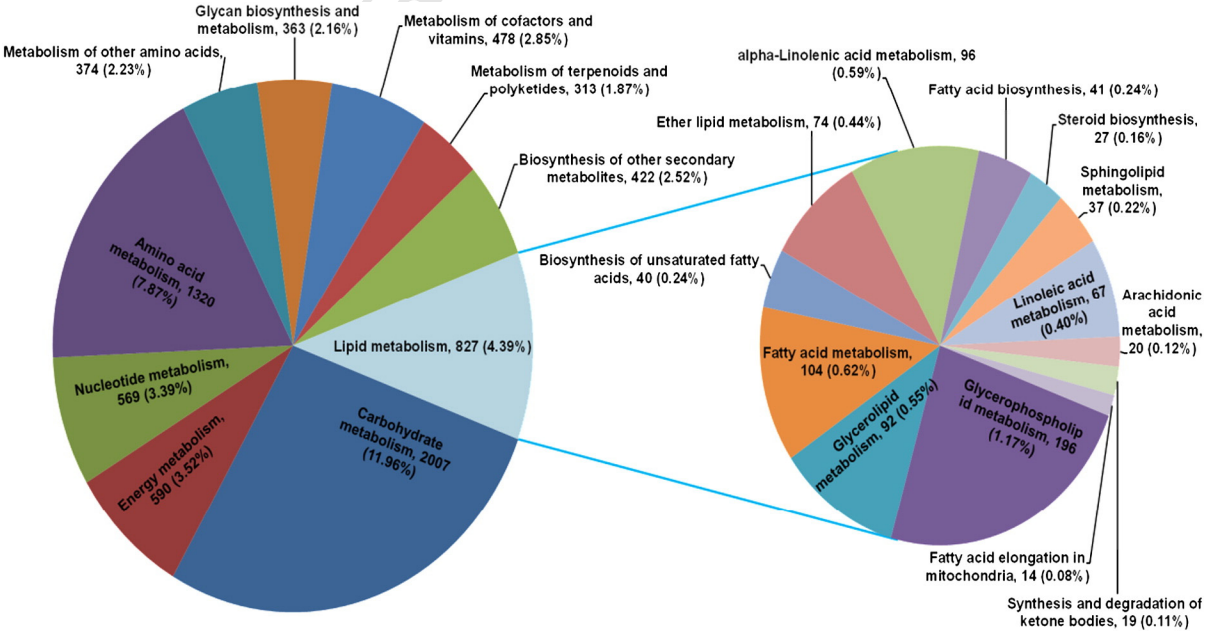

Fig. 4. Pathway assignment based on the KEGG. Classification based on metabolism categories.

(sterol 14α-demethylase), *14SR* (Δ14-sterol reductase), *4-MSO* (C4-methylsterol oxidase), *EBP* (cholestenol Δ-isomerase), *DHCR24* (Δ24-sterol reductase), *SC5DL* (sterol C5 desaturase/lathosterol oxidase), *DHCR7* (7-dehydrocholesterol reductase), *3β-HSD* (3β-hydroxysteroid dehydrogenase), and *5β-POR* (progesterone 5β-reductase), suggesting that the biosynthesis of polyoxypregnane glycosides in *M. tenacissima* might be similar to cardenolides biosynthesis in *Digitalis*, both of which share similar early enzymatic steps in their biosynthetic pathway. The functions of these genes, and the relationship between their expression levels and polyoxypregnane glycoside accumulation, will be studied in the future.

Some gene encoding enzymes involved in cholesterol and pregnane backbone biosynthesis were not found in this study (Table 1) because they are only isolated and characterized in animals or microorganisms, such as NSDHL (sterol-4α-carboxylate 3-dehydrogenase, decarboxylating), 3-KSR (3-keto-steroid reductase), SCCE (cholesterol monooxygenase, side chain-cleaving enzyme), KSI (Δ5-Δ4-ketosteroid isomerase), and P14βH (pregnane 14β-hydroxylase). To search for these genes in *M. tenacissima* transcriptome, we compared all unigenes with *Stenotrophomonas maltophilia* NSDHL (P-001970132.1), *Homo sapiens* 3-KSR (NP-057455.1), *Rattus norvegicus* SCCE (AAA40989.1), *Comamonas testosteroni* KSI (AAA25871.1) and *Oryctolagus cuniculus* cholesterol-7α-hydroxylase gene (AAA74382.1). The most similar unigenes only had 23–31% identity to genes identified in the other species mentioned (data not shown), suggesting that these genes have little similarity with those in animals or microorganisms and cannot be cloned by homology-based cloning methods.

The mitochondrial CYP-dependent side chain cleaving enzyme (SCCE) catalyzing the reaction converts sterols into pregnenolone [23]. No evidence of such a P450 (*CYP11A* in animals) has yet been found in plants [44]; therefore, more attention was given to possible interaction partners, such as acyl-CoA-binding protein (*ACBP*) and peripheral-type benzodiazepine receptor (*PBR*) [45,46]. In the mitochondrial envelope, *ACBPs* bind to *PBR* and stimulate the transport of cholesterol into the mitochondria [47]. Unigenes annotated to *ACBP* and *PBR* were also found in *M. tenacissima* transcriptome in this study (Table 1), which will help us elucidate their function in polyoxypregnane biosynthesis.

2.7. Candidate gene encoding enzymes that catalyze pregnane modifications

For the synthesis of different polyoxypregnane glycosides in *M. tenacissima*, the pregnane backbone must be modified by hydroxylation, acylation and glycosylation, catalyzed by hydroxylases, acyltransferases and glucosyltransferases, respectively. The main aglycone of polyoxypregnane glycosides in *M. tenacissima* is tenacigenin B, which has five hydroxyl groups at 3-, 8-, 11-, 12-, 14-C in the backbone of pregnane, respectively (Fig. 1). There must be some pregnane hydroxylases (that belong to cytochrome P450, CYP) that catalyze these hydroxylation reactions. Some steroid hydroxylases have been found that hydroxylate different C-atoms, such as cholesterol-7α-hydroxylase (*CYP7A1*), steroid 17α-hydroxylase (*CYP17*) and *CYP90B1* [48,49,37]. Though homologs of those genes do not exist in the *M. tenacissima* transcriptome, we did find 208 unigenes annotated to the CYP family (Additional file 4), which will help us to identify pregnane hydroxylase in *M. tenacissima*. Moreover, 14 and 40 unigenes that were annotated as sterol *O*-acyltransferase and sterol 3-*O*-glucosyltransferase, respectively, were also found in this study (Table 1), some of which were distantly related to genes from other plant species (Fig. 5, Additional file 5), indicating that these unigenes might encode enzymes that catalyze acylations and glycosylations in *M. tenacissima*.

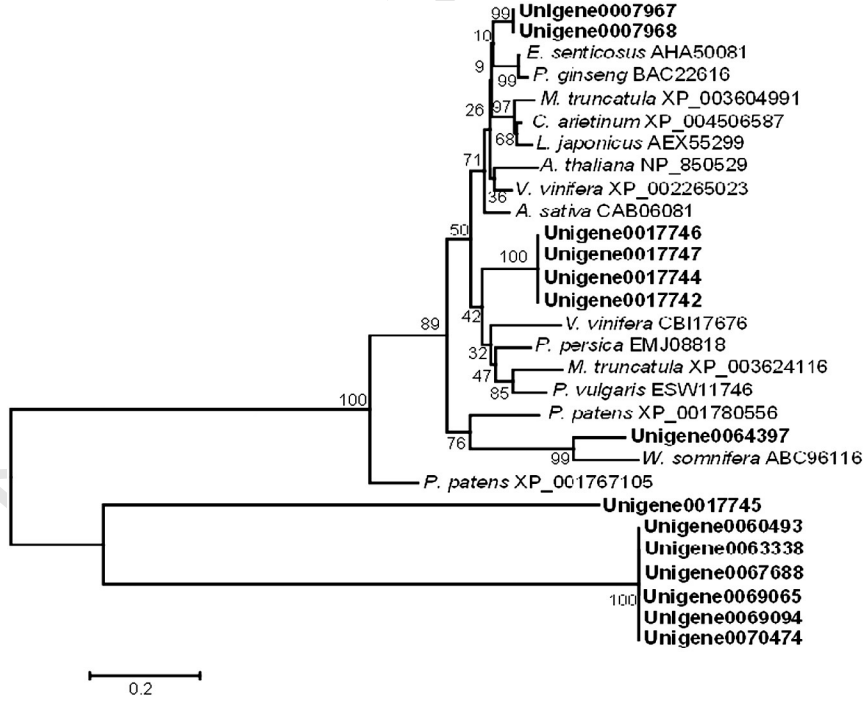

Fig. 5. Phylogenetic analysis of sterol 3-*O*-glucosyltransferase genes from *M. tenacissima* (bold letters) and characterized sterol 3-*O*-glucosyltransferase genes from other plants. Phylogenetic tree constructed based on the deduced amino acid sequences. Amino acid sequences were aligned using the ClustalW program, and evolutionary distances were computed using MEGA5.10 with the Poisson correction method. Bootstrap values obtained after 1000 replications are indicated on the branches. Bar = 0.2 amino acid substitutions/site. Protein sequences were retrieved from NCBI GenBank using the following accession numbers (source organism and proposed function, if any, are given in parentheses): AHA50081 (*Eleutherococcus senticosus*, sterol 3-*O*-glucosyltransferase); BAC22616 (*Panax ginseng*, sterol 3-*O*-glucosyltransferase); XP_003604991 (*M. truncatula*, sterol 3β-glucosyltransferase); XP_004506587 (*Cicer arietinum*, sterol 3β-glucosyltransferase-like); AEX55299 (*Lotus japonicas*, sterol glucosyltransferase 1); NP_850529 (*A. thaliana*, sterol 3β-glucosyltransferase); XP_002265023 (*V. vinifera*, sterol 3β-glucosyltransferase-like); CAB06081 (*Avena sativa*, sterol glucosyltransferase); CBI17676 (*V. vinifera*, UDP-glucuronosyltransferase); EMJ08818 (*Prunus persica*, UDP-glucuronosyltransferase); XP_003624116 (*M. truncatula*, sterol 3β-glucosyltransferase); ESW11746 (*Phaseolus vulgaris*, UDP-glucuronosyltransferase); XP_001780556 (*Physcomitrella patens*, UDP-glucuronosyltransferase); ABC96116 (*Withania somnifera*, sterol glucosyltransferase); XP_001767105 (*P. patens*).

Please cite this article as: K. Zheng, et al., Genomics (2014), http://dx.doi.org/10.1016/j.ygeno.2014.07.013

2.8. Expression patterns of five unigenes related to polyoxypregnane glycoside biosynthesis

The candidate genes *SQS*, *SQLE*, *CAS*, *4-MSO*, *3β-HSD* and *5β-POR* were selected for further analysis, and their expression patterns in leaves and stems were analyzed by qRT-PCR. The expression patterns of these genes are shown in Fig. 6. Among them, the gene expression levels of *CAS*, *SQS*, *SQLE*, *4-MSO* and *3β-HSD* were higher in leaves than in stems; conversely, the expression level of the *5β-POR* gene was 36.50% higher in stems than in leaves. Higher expression levels in leaves of *CAS*, *SQS*, *SQLE*, *4-MSO* and *3β-HSD* genes indicate that leaves are the main organs for synthesizing the precursors of polyoxypregnane, and higher expression levels of the downstream enzyme *5β-POR* in stems suggest that polyoxypregnane is modified and stored in stems, which is the major medicinal part of *M. tenacissima* and contains high contents of polyoxypregnane glycosides. The analysis of the expression patterns of these genes in leaves and stems will be helpful to further understand the mechanism of polyoxypregnane glycoside biosynthesis.

2.9. EST-SSR discovery: distribution and frequencies

To develop new molecular markers, using MISA software, all of the 15,296 microsatellites were detected in 11,911 unigenes. Of all the SSR-containing unigenes, 2573 sequences contained more than 1 SSR, and 989 SSRs were present in compound form. On average, we found 2.14 SSR per 10 Kb in this study. Microsatellites included 8217 (53.72%) dinucleotide motifs, 5094 (33.31%) trinucleotide motifs, 1394 (9.11%) tetranucleotide motifs, 313 (2.05%) pentanucleotide motifs and 277 (1.81%) hexanucleotide motifs. The length of SSRs was also analyzed; the majority were between 18 bp to 27 bp. SSRs with six tandem repeats (4386, 28.68%) were the most common, followed by five tandem repeats (3273, 21.40%), seven tandem repeats (2563, 16.76%), and four tandem repeats (1526, 10.21%) (Table 2). The information of SSRs derived from all unigenes is shown in Additional file 6. The most abundant repeat type was AT/AT (26.34%), followed by AG/CT (21.02%), AAG/CTT (8.49%), and AAT/ATT (6.43%). Based on those SSRs, 27,189 primer pairs were successfully designed using Primer 3 (Additional file 6). The unique sequence-derived markers generated in this study represent a valuable genetic resource for SSR mining and future applications in research and molecular marker-assistant breeding in this plant.

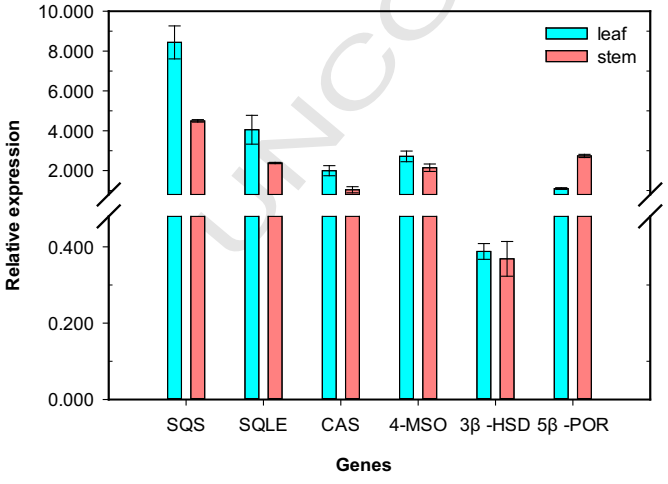

Fig. 6. Expression patterns of six genes related to the biosynthesis of polyoxypregnane glycosides in leaves and stems. Bars represent the mean (±SD). *SQS*: squalene synthase, *SQLE*: squalene epoxidase, *CAS*: cycloartenol synthase, *4-MSO*: C4-methylsterol oxidase, *3β-HSD*: 3β-hydroxysteroid dehydrogenase, and *5β-POR*: progesterone 5β-reductase.

3. Conclusion

Based on the analysis of the *M. tenacissima* transcriptome, valuable gene candidates for the biosynthesis of polyoxypregnane glycosides were identified and will likely facilitate functional studies aiming to produce larger quantities of this compound for cancer treatment. These data not only enrich genomic resources for the species but also benefit research on genetics, functional genomics, and gene expression.

4. Materials and methods

4.1. Plant material and RNA extraction

One-year-old *M. tenacissima* seedlings were grown in the experimental station of the Yunnan Agricultural University, Kunming, China [latitude: 25°7′ 60″ N, longitude: 102°45′ 10″ E, altitude: 1895 m]. Samples were collected from fresh leaves and stems, which were immediately frozen in liquid nitrogen and stored at −80 °C until further processing. Total RNA was extracted using the TRIzol Kit (Promega, USA), and RNA quality was measured using Agilent's Bioanalyzer and agarose gel electrophoresis. To obtain complete gene expression information, equal amounts of total RNA from leaves and stems were pooled together for cDNA preparation.

4.2. cDNA library construction and sequencing

For constructing an mRNA library, poly (A) RNA was purified from 20 mg total RNA using Sera-mag Magnetic Oligo (dT) Beads (Illumina). Then, the mRNA was fragmented using a fragmentation buffer. The mRNA fragments were transcribed into first-strand cDNA using random hexamer primers. Second-strand cDNA was synthesized using DNA polymerase I and RNase H. The cDNA fragments were purified and enriched with PCR for end reparation and the addition of poly (A) and were connected with sequencing adaptors. After resolution by agarose gel electrophoresis, suitable fragments were selected for PCR amplification. Lastly, the cDNA fragments were sequenced using Illumina HiSeq 2000 at Gene Denovo Corporation (Guangzhou, China).

4.3. Illumina read processing and assembly

The raw reads obtained from the sequencing machine were preprocessed by trimming adaptors and discarding low-quality reads (reads containing more than 50% bases with Q-value ≤ 20). The remaining high-quality sequences were then used for de novo transcriptome assembly using the short reads assembling program Trinity. The assembly was performed using the default parameters.

4.4. Functional annotation and predicted CDS

For functional annotations, the generated unigenes were compared with a series of public databases, such as the non-redundant protein database (Nr, http://www.ncbi.nlm.nih.gov/) and the Swiss-Prot database (http://www.expasy.ch/sprot), using BLASTx (E-value < 10^{-5}) and BLAST (E-value < 10^{-10}), respectively. The unigenes were also aligned to the Cluster of Orthologous Groups (COG) of protein database (http://www.ncbi.nlm.nih.gov/COG/) and Kyoto Encyclopedia of Genes and Genomes database (KEGG, http://www.genome.jp/kegg) [50] using BLASTx with an E-value < 10^{-10}. Through the comparison against the KEGG database, we can further study the complex biological behaviors of genes and obtain pathway annotation for unigenes. A Perl script was used to retrieve KO (KEGG ontology) information from the BLAST results to establish pathway associations between unigenes and KEGG. The gene ontology (GO) (http://www.geneontology.org) [51] database annotates genes as belonging to one of three functional categories: biological process, molecular function, or cellular component. The functional categories of these unigenes were further identified

Please cite this article as: K. Zheng, et al., Genomics (2014), http://dx.doi.org/10.1016/j.ygeno.2014.07.013

Table 2
Distribution of identified SSRs using the MISA software.

Motif	Repeat numbers									Total	%
	4	5	6	7	8	9	10	11	12		
Di-	0	0	2884	1905	1393	1117	647	249	22	8217	53.72
Tri-	0	2888	1465	658	83	0	0	0	0	5094	33.31
Tetra-	1009	348	37	0	0	0	0	0	0	1394	9.11
Penta-	276	37	0	0	0	0	0	0	0	313	2.05
Hexa-	277	0	0	0	0	0	0	0	0	277	1.81
Total	1526	3273	4386	2563	1476	1117	647	249	22		
%	10.21	21.40	28.68	16.76	9.65	7.30	4.23	1.63	0.14		

using the GO Database, and GO trees were generated using the WEGO tool (http://wego.genomics.org.cn/cgibin/wego/index.pl) [52].

The CDSs (coding DNA sequences) of all unigenes were predicted by using BLASTX and ESTScan. First, we performed BLASTx alignment (E-value < 10^{-5}) between unigenes and protein databases such as Nr, SwissProt, KEGG and COG. The best alignment results were used to determine the sequence direction of unigenes. Unigenes with sequences that produced matches in only one database were not searched further. When a unigene would not align to any database, ESTScan was used to predict coding regions and determine sequence direction.

4.5. Real-time PCR analysis

To assay the expression levels of mRNA of putative key genes in the stems and leaves of *M. tenacissima*, qRT-PCR was performed using an Applied Biosystems 7500 Fast Real-Time PCR system with three replicates using FSQ-301 (Toyobo, Japan). Total RNA was treated with 4× DN Master Mix (with gDNA remover added) at 37 °C for 5 min to remove DNA. The reverse transcription reaction was performed using the 5× RT Master Mix II according to the manufacturer's instructions. For quantitative RT-PCR, reactions (20 μL) consisted of 2 μL of first-strand cDNA, 0.4 μM primers, 10 μL of SYBR® Premix Ex Taq™ (2×) (Fermentas), and 7.6 μL of ddH_2O. PCR cycling conditions were as follows: 95 °C for 2 min followed by 40 cycles of 95 °C for 15 s and then 60 °C for 40 s. A melting curve was performed from 65 °C to 95 °C to check the specificity of the amplified product. Primer sequences are listed in Additional file 7. Based on the transcriptome sequencing and annotation, almost all of the putative unigenes involved in the polyoxypregnane glycosides biosynthetic pathway were identified. To further analyze the expression patterns of these genes in leaf and stem tissues, five essential genes were selected for verification by qRT-PCR: squalene synthase (*SQS*), squalene epoxidase (*SQLE*), cycloartenol synthase (*CAS*), C4-methylsterol oxidase (*4-MSO*), 3β-hydroxysteroid dehydrogenase (*3β-HSD*) and progesterone 5β-reductase (*5β-POR*). The glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) was used as an internal control gene. The relative expression levels of the selected genes were normalized to GAPDH and calculated using the $2^{-\Delta\Delta}$CT method [53].

4.6. EST-SSR detection and primer design

The potential SSR markers with motifs ranging from di- to hexa-nucleotides were detected among the 65,796 unigenes by using the MISA tool (http://pgrc.ipk-gatersleben.de/misa/). The minimum of repeat units were set as follows: six for di-, five for tri-, and four for tetra-, penta- and hexa-nucleotides. The maximum interruption distance between two SSRs was specified as 100 bases. The primers for the identified SSR loci were designed using Primer 3 (http://primer3.ut.ee/). Among all the designed primers, GC content ranged between 40% and 60%, and the expected PCR product sizes ranged from 100 to 280 bp.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.07.013.

Acknowledgments

This work was funded by the Nanjing Sanhome Pharmaceutical and by Yunan Base Construction Project of Technical Industry Actions for TCM Modernization.

References

[1] Jiangsu New Medicinal College, Zhongyao Dacidian (Encyclopedia of Chinese Materia Medica), Shanghai Science and Technology Press, Shanghai, 1977. 1976.
[2] Chinese Pharmacopoeia Commission, Chinese Pharmacopoeia, Chinese Science and Technology Press, Beijing 100061, China, 2010. 277–278.
[3] Z. Huang, H. Lin, Y. Wang, Z. Cao, W. Lin, Q. Chen, Studies on the anti-angiogenic effect of *Marsdenia tenacissima* extract *in vitro* and *in vivo*, Oncol. Lett. 5 (2013) 917–922.
[4] Z. Huang, Y. Wang, J. Chen, R. Wang, Q. Chen, Effect of Xiaoaiping injection on advanced hepatocellular carcinoma in patients, J. Tradit. Chin. Med. 33 (2013) 34–38.
[5] W. Li, Y. Yang, Z. Ouyang, Q. Zhang, L. Wang, F. Tao, Y. Shu, Y. Gu, Q. Xu, Y. Sun, Xiaoai-ping, a TCM injection, enhances the antigrowth effects of cisplatin on lewis lung cancer cells through promoting the infiltration and function of CD8(+) T lymphocytes, Evid. Based Complement. Alternat. Med. (2013) 879512.
[6] J. Deng, F. Shen, D. Chen, Quantitation of seven polyoxypregnane glycosides in *Marsdenia tenacissima* using reversed-phase high-performance liquid chromatography-evaporative light-scattering detection, J. Chromatogr. A 116 (2006) 83–88.
[7] S. Miyakawa, K. Yamaura, K. Hayashi, K. Kaneko, H. Mitsuhashi, Five glycosides from the Chinese drug "tong-guang-san": the stems of *Marsdenia tenacissima*, Phytochemistry 25 (1986) 2861–2865.
[8] S.X. Qiu, S.Q. Luo, L.Z. Lin, G.A. Cordell, Further polyoxypregnanes from *Marsdenia tenacissima*, Phytochemistry 41 (1996) 1385–1388.
[9] J. Deng, Z. Liao, D. Chen, Marsdenosides A–H, polyoxypregnane glycosides from *Marsdenia tenacissima*, Phytochemistry 66 (2005) 1040–1051.
[10] Z.H. Xia, W.X. Xing, S.L. Mao, A.N. Lao, J. Uzawa, S. Yoshida, Y. Fujimoto, Pregnane glycosides from the stems of *Marsdenia tenacissima*, J. Asian Nat. Prod. Res. 6 (2004) 79–85.
[11] Z.H. Xia, S.L. Mao, A.N. Lao, J. Uzawa, S. Yoshida, Y. Fujimoto, Five new pregnane glycosides from the stems of *Marsdenia tenacissima*, J. Asian Nat. Prod. Res. 13 (2011) 477–485.
[12] X.L. Wang, Q.F. Li, K.B. Yu, S.L. Peng, Y. Zhou, L.S. Ding, Four new pregnane glycosides from the stems of *Marsdenia tenacissima*, Helv. Chim. Acta 89 (2006) 2738–2744.
[13] X.L. Wang, S.L. Peng, L.S. Ding, Further polyoxypregnane glycosides from *Marsdenia tenacissima*, J. Asian Nat. Prod. Res. 12 (2010) 654–661.
[14] M. Yang, W.L. Wang, H. Wu, G. Zhu, X.L. Wang, Pregnane glycosides from stems of *Marsdenia tenacissima*, Chin. Tradit. Herb. Drugs 42 (2011) 1473–1476.
[15] J. Chen, X. Li, C. Sun, Y. Pan, U.P. Schlunegger, Identification of polyoxypregnane glycosides from the stems of *Marsdenia tenacissima* by high-performance liquid chromatography/tandem mass spectrometry, Talanta 77 (2008) 152–159.
[16] P. Benveniste, Biosynthesis and accumulation of sterols, Annu. Rev. Plant Biol. 55 (2004) 429–457.
[17] Y. Sun, H. Luo, Y. Li, C. Sun, J. Song, Y. Niu, Y. Zhu, L. Dong, A. Lv, E. Tramontano, S. Chen, Pyrosequencing of the *Camptotheca acuminata* transcriptome reveals putative genes involved in camptothecin biosynthesis and transport, BMC Genomics 12 (2011) 533.
[18] X. Guo, Y. Li, C. Li, H. Luo, L. Wang, J. Qian, X. Luo, L. Xiang, J. Song, C. Sun, H. Xu, H. Yao, S. Chen, Analysis of the *Dendrobium officinale* transcriptome reveals putative alkaloid biosynthetic genes and genetic markers, Gene 527 (2013) 131–138.
[19] C. Li, Y. Zhu, X. Guo, C. Sun, H. Luo, J. Song, Y. Li, L. Wang, J. Qian, S. Chen, Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng* C. A. Meyer, BMC Genomics 14 (2013) 245.
[20] L. Xiang, Y. Li, Y. Zhu, H. Luo, C. Li, X. Xu, C. Sun, J. Song, L. Shi, L. He, W. Sun, S. Chen, Transcriptome analysis of the *Ophiocordyceps sinensis* fruiting body reveals putative genes involved in fruiting body development and cordycepin biosynthesis, Genomics 103 (2014) 154–159.
[21] N. Panda, S. Banerjee, N.B. Mandal, N.P. Sahu, Pregnane glycosides, Nat. Prod. Commun. 1 (2006) 665–695.
[22] H.H. Sauer, R.D. Bennett, E. Heftmann, Biosynthesis of pregnane derivatives in *Strophanthus kombe*, Phytochemistry 8 (1969) 69–77.

[23] P. Lindemann, M. Luckner, Biosynthesis of pregnane derivatives in somatic embryos of *Digitalis lanata*, Phytochemistry 46 (1997) 507–513.
[24] W. Kreis, F. Müller-Uri, Cardenolide aglycone formation in *Digitalis*, Isoprenoid Synthesis in Plants and Microorganisms, Springer, New York, 2013, pp. 425–438.
[25] D.E. Gärtner, W. Keilholz, H.U. Seitz, Purification, characterization and partial peptide microsequencing of progesterone 5β-reductase from shoot cultures of *Digitalis purpurea*, Eur. J. Biochem. 225 (1994) 1125–1132.
[26] I. Gavidia, P. Pérez-Bermúdez, H.U. Seitz, Cloning and expression of two novel aldo-keto reductases from *Digitalis purpurea* leaves, Eur. J. Biochem. 269 (2002) 2842–2850.
[27] Y. Kazeto, S. Ijiri, H. Matsubara, S. Adachi, K. Yamauchi, Molecular cloning and characterization of 3β-hydroxysteroid dehydrogenase/Delta5-Delta4 isomerase cDNAs from Japanese eel ovary, J. Steroid Biochem. Mol. Biol. 85 (2003) 49–56.
[28] L. Roca-Pérez, R. Boluda, I. Gavidia, P. Pérez-Bermúdez, Seasonal cardenolide production and *Dop5βr* gene expression in natural populations of *Digitalis obscura*, Phytochemistry 65 (2004) 1869–1878.
[29] V. Herl, G. Fischer, R. Bötsch, F. Müller-Uri, W. Kreis, Molecular cloning and expression of progesterone 5β-reductase (5beta-POR) from *Isoplexis canariensis*, Planta Med. 72 (2006) 1163–1165.
[30] V. Herl, G. Fischer, F. Müller-Uri, W. Kreis, Molecular cloning and heterologous expression of progesterone 5β-reductase from *Digitalis lanata* Ehrh, Phytochemistry 67 (2006) 225–231.
[31] E. Sales, J. Muñoz-Bertomeu, I. Arrillaga, J. Segura, Enhancement of cardenolide and phytosterol levels by expression of an N-terminally truncated 3-hydroxy-3-methylglutaryl CoA reductase in transgenic *Digitalis minor*, Planta Med. 73 (2007) 605–610.
[32] S.P. Kuate, R.M. Pádua, W.F. Eisenbeiss, W. Kreis, Purification and characterization of malonyl-coenzyme A: 21-hydroxypregnane 21-*O*-malonyltransferase (*Dp*21MaT) from leaves of *Digitalis purpurea* L. Phytochemistry 69 (2008) 619–626.
[33] P. Pérez-Bermúdez, A.A. García, I. Tuñón, I. Gavidia, *Digitalis purpurea P5βR2*, encoding steroid 5β-reductase, is a novel defense-related gene involved in cardenolide biosynthesis, New Phytol. 185 (2010) 687–700.
[34] S. Fujita, T. Ohnishi, B. Watanabe, T. Yokota, S. Takatsuto, S. Fujioka, S. Yoshida, K. Sakata, M. Mizutani, Arabidopsis CYP90B1 catalyses the early C_{22} hydroxylation of C_{27}, C_{28} and C_{29} sterols, Plant J. 45 (2006) 765–774.
[35] T. Ohnishi, B. Watanabe, K. Sakata, M. Mizutani, CYP724B2 and CYP90B3 function in the early C-22 hydroxylation steps of brassinosteroid biosynthetic pathway in tomato, Biosci. Biotechnol. Biochem. 70 (2006) 2071–2080.
[36] T. Morikawa, M. Mizutani, D. Ohta, Cytochrome P450 subfamily CYP710A genes encode sterol C-22 desaturase in plants, Biochem. Soc. Trans. 34 (2006) 1202–1205.
[37] T. Morikawa, H. Saga, H. Hashizume, D. Ohta, CYP710A genes encoding sterol C22-desaturase in *Physcomitrella patens* as molecular evidence for the evolutionary conservation of a sterol biosynthetic pathway in plants, Planta 229 (2009) 1311–1322.
[38] A.C. Diener, H. Li, W. Zhou, W.J. Whoriskey, W.D. Nes, G.R. Fink, *Sterol methyltransferase 1* controls the level of cholesterol in plants, Plant Cell 12 (2000) 853–870.
[39] F. Sitbon, L. Jonsson, Sterol composition and growth of transgenic tobacco plants expressing type-1 and type-2 sterol methyltransferases, Planta 212 (2001) 568–572.
[40] A.K. Neelakandan, Z. Song, J. Wang, M.H. Richards, X. Wu, B. Valliyodan, H.T. Nguyen, W.D. Nes, Cloning, functional expression and phylogenetic analysis of plant sterol 24 C-methyltransferases involved in sitosterol biosynthesis, Phytochemistry 70 (2009) 1982–1998.
[41] F. Carland, S. Fujioka, T. Nelson, The sterol methyltransferases SMT1, SMT2, and SMT3 influence *Arabidopsis* development through nonbrassinosteroid products, Plant Physiol. 153 (2010) 741–756.
[42] K. Ohyama, M. Suzuki, J. Kikuchi, K. Saito, T. Muranaka, Dual biosynthetic pathways to phytosterol via cycloartenol and lanosterol in *Arabidopsis*, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 725–730.
[43] W.D. Nes, Biosynthesis of cholesterol and other sterols, Chem. Rev. 111 (2011) Q7 6423–6451.
[44] T. Ohnishi, T. Yokota, M. Mizutani, Insights into the function and evolution of P450s in plant steroid metabolism, Phytochemistry 70 (2009) 1918–1929.
[45] M. Metzner, K.P.Ruecknagel, J. Knudsen, G. Kuellertz, F. Mueller-Uri, B. Diettrich, Isolation and characterization of two acyl-CoA-binding proteins from proembryogenic masses of *Digitalis lanata* Ehrh, Planta 210 (2000) 683–685.
[46] V. Papadopoulos, M. Baraldi, T.R. Guilarte, T.B. Knudsen, J.J. Lacapère, P. Lindemann, M.D. Norenberg, D. Nutt, A. Weizman, M.R. Zhang, M. Gavish, Translocator protein (18 kDa): new nomenclature for the peripheral-type benzodiazepine receptor based on its structure and molecular function, Trends Pharmacol. Sci. 27 (2006) 402–409.
[47] V. Papadopoulos, H. Amri, N. Boujrad, C. Cascio, M. Culty, M. Garnier, M. Hardwick, H. Li, B. Vidic, A.S. Brown, J.L. Reversa, J.M. Bernassau, K. Drieu, Peripheral benzodiazepine receptor in cholesterol transport and steroidogenesis, Steroids 62 (1997) 21–28.
[48] J.C. Cohen, J.J. Cali, D.F. Jelinek, M. Mehrabian, R.S. Sparkes, A.J. Lusis, D.W. Russell, H. H. Hobbs, Cloning of the human cholesterol 7 alpha-hydroxylase gene (CYP7) and localization to chromosome 8q11–q12, Genomics 14 (1992) 153–161.
[49] T.A. Sushko, A.A. Gilep, A.V. Yantsevich, S.A. Usanov, Role of microsomal steroid hydroxylases in Δ^7-steroid biosynthesis, Biochemistry 78 (2013) 282–289.
[50] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, From genomics to chemical genomics: new developments in KEGG, Nucleic Acids Res. 34 (Database issue) (2006) 354–357.
[51] M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G.M. Rubin, J.A. Blake, C. Bult, M. Dolan, H. Drabkin, J.T. Eppig, D.P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J.M. Cherry, K. R. Christie, M.C. Costanzo, S.S. Dwight, S. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, R.S. Nash, A. Sethuraman, C.L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S.Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E.M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, Gene Ontology Consortium, The gene ontology (GO) database and informatics resource, Nucleic Acids Res. 32 (Database issue) (2004) 258–261.
[52] J. Ye, L. Fang, H. Zheng, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, J. Wang, WEGO: a web tool for plotting GO annotations, Nucleic Acids Res. 34 (Web Server issue) (2006) 293–297.
[53] Kenneth J. Livak, Thomas D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method, Methods 25 (2001) 402–408.

Please cite this article as: K. Zheng, et al., Genomics (2014), http://dx.doi.org/10.1016/j.ygeno.2014.07.013