***Pak. J. Bot.*, 50(3): 1047-1059, 2018.**

TRANSCRIPTOME ANALYSIS OF *SCHISANDRA SPHENANTHERA* DISCOVERS PUTATIVE LIGNAN BIOSYNTHESIS GENES AND GENETIC MARKERS

SEN WANG[1†], ZHEN ZHANG[1†], WAN-LING SONG[2†], NI-HAO JIANG[2], FENG-XIA SHAO[1], SHENG-CHAO YANG[2], FANG-DE LV[1], GUANG-HUI ZHANG[2], CHUN-HUA MA[2*] AND JUN-WEN CHEN[2*]

[1]College of Forestry, Central South University of Forestry and Technology, Changsha, 410004, Hunan, People's Republic of China
[2]Yunnan Agricultural University National & Local Joint Engineering Research Center on Gemplasm Utilization & Innovation of Chinese Medicinal Materials in Southwest China, Kunming, 650201, Yunnan, People's Republic of China
[*]Corresponding author's email: pony0207@126.com, cjw31412@hotmail.com
[†]These authors contribute equally to this paper.

Abstract

Based on the transcriptional analysis of *Schisandra sphenanthera* Rehd. Et Wils., a total of 129,951 assembled unigenes were obtained. This article found some cytochrome P450 enzymes (CYP450), such as CYP81Q1 (piperitol/sesamin synthase, PSS), CYP719A24 and CYP71923 catalyze the formation of two methylenedioxy bridges in sesamin and podophyllotoxin biosynthesis, respectively.

The candidates of these CYP450s and *O*-methyltransferase (OMT) genes that catalyze lignan*O*-methylation have also discovered from the transcriptome database of *S. sphenanthera*, indicating that dibenzocyclooctadiene-type lignan might be biosynthesized in the similar pathways of sesamin and podophyllotoxin. Moreover, cloning and functional characterization of those genes will help us to illuminate the biosynthesis mechanisms of dibenzocyclooctadiene-type lignan. Additionally, twenty sets of primers for SSR were chosen in random for validating the polymorphism and amplification. The results showed that 18 (90.00%) primer pairs could be amplified successfully and 15 (83.33%) primer pairs exhibited polymorphisms.This study represents the first report to analyze the transcriptome of *S. sphenanthera* using high-throughput RNA-seq technology. These data will enrich the genomic data and provide a solid evidence for functional genomics and molecular genetic researching in this herb.

Key words: *Schisandra sphenanthera*; Transcriptome; Lignan; Biosynthesis.

Introduction

Schisandra sphenanthera Rehd. Et Wils. is a perennial plant belonging to the Magnoliaceae family, which is widely distributed in south and southwest of China, mainly in Sichuan, Yunnan, Guizhou, Guangxi, and Hunan Provinces. The dry fruit of *S. sphenanthera*, known as Nan-Wuweizi (literally 'Southern Magnoliavine Fruit'), has been used as superior traditional drugs and functional foods for several thousand years (Wang *et al*., 2011). It has been officially listed as an important sedative and tonic agent (Liu *et al*., 2012) because of its activities in treating astringency of sweating, seminal emission, enuresis, frequent urination, diarrhea and tranquillizing of the mind (Chinese Pharmacopoeia Commission, 2010). Lignans, particularly the dibenzocyclooctadiene-type lignans, mainlydeoxyschizandrin, schisantherin A, and 2,3-dimethyl-1,4-diarylbutane-type lignan, anwulignan, are believed to be the main active principles of *S. sphenanthera* (Lu & Chen, 2009; Liu *et al*., 2012; Xia *et al*., 2014). Some of them were reported to possess valuable bioactivities of tumor suppression (Chen *et al*., 2002; Fong *et al*., 2007; Yoo *et al*., 2007), anti-platelet aggregation (Jiang *et al*., 2005) and anti-HIV (human immunodeficiency virus) effects (Chen *et al*., 1997; 2006).

Lignan is a group of phenylpropanoid dimers and can be classified into eight subgroups, most of them have oxygen at C9 (C9') except for some of those from dibenzocyclooctadiene, dibenzylbutane and furan (Whiting, 1985; Umezawa, 2003). The biosynthesis of lignans with C9 (C9') oxygen is well elucidated, which is formed by enantioselective dimerization of two coniferyl alcohol units to produce pinoresinol with the help of dirigent protein (DIR), then is converted to matairesinol catalyzed by pinoresinol/lariciresinol reductase (PLR), secoisolariciresinol dehydrogenase (SIRD) (Suzuki & Umezawa, 2007, Fig. 1). The conversion from coniferyl alcohol to matairesinol is believed to be the general lignan biosynthetic pathway and has been identified in a variety of plant species for the biosynthesis of most lignans, such as podophyllotoxin in Linum (Xia *et al*., 2000; Hano *et al*., 2006; Hemmati *et al*., 2007a, 2007b; 2010; Bayindir *et al*., 2008; Renouard *et al*., 2012) and other plants (Dinkova-Kostova *et al*., 1996; Marques *et al*., 2013). The furofuran type of lignansemamin has also been well elucidated, which is synthesized by formation of two methylenedioxy bridges from pinoresinol by means of piperitol mediated by cytochrome P450 CYP81Q1 (also called piperitol/sesamin synthase, PSS) in *Sesamum indicum* (Ono *et al*., 2006); the CYPs with similar functions of converting matairesinol into pluviatolide by catalyzing methylenedioxy bridge formation are also found in *Podophyllum* species (Marques *et al*., 2013).Even though the main biosynthetic pathway of lignans has been studied extensively, especially for furofuran and aryltetalin types of lignans, some downstream steps have only been reported using crude enzymatic assays; no genes have yet been identified or the enzymes purified to homogeneity (Molog *et al*., 2001; Federolf *et al*., 2007). Moreover, little is known about their biosynthesis, and the available genomic information and molecular markers of *S. sphenanthera* are also limited.

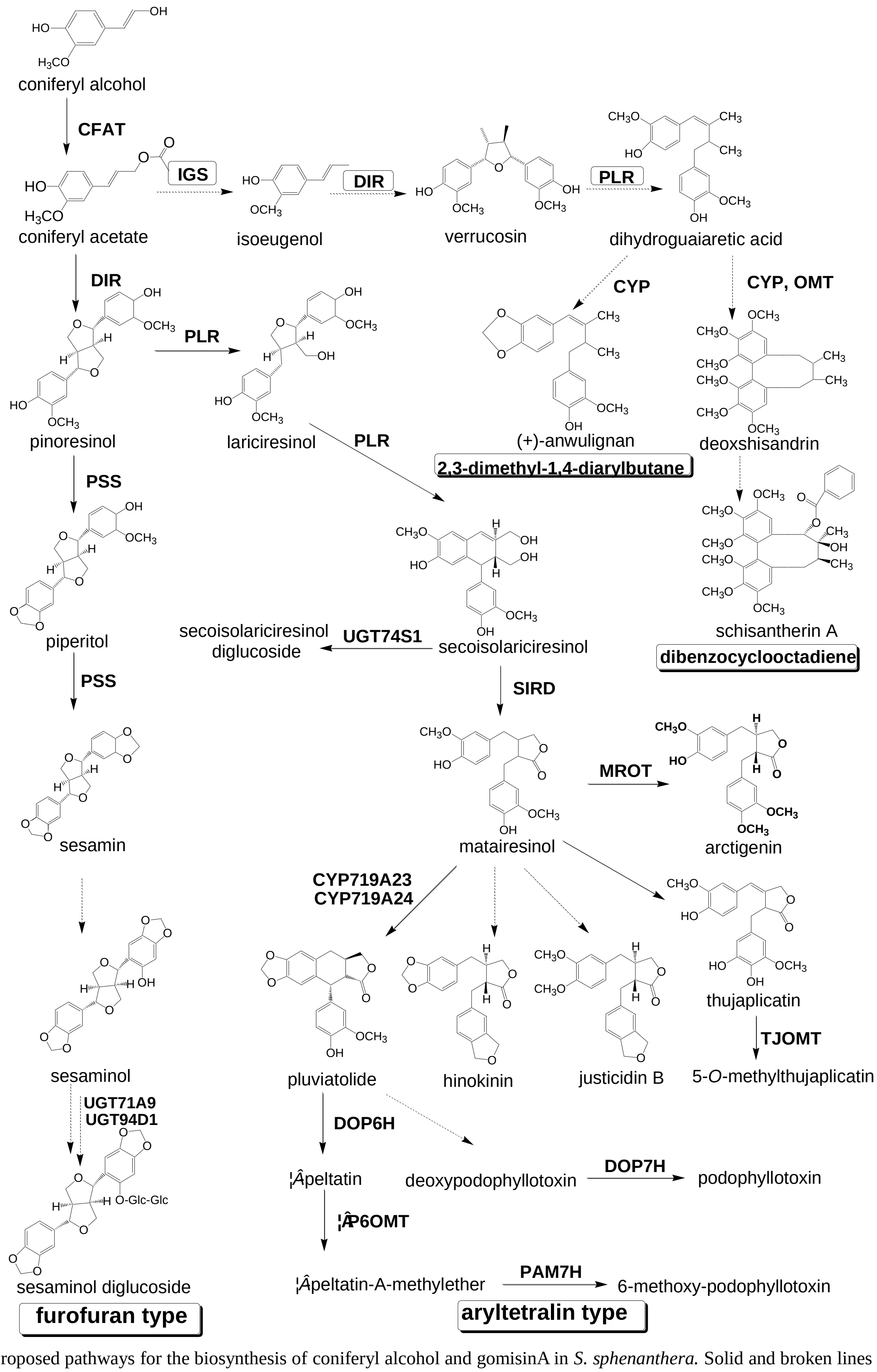


Fig. 1. Proposed pathways for the biosynthesis of coniferyl alcohol and gomisinA in *S. sphenanthera.* Solid and broken lines represent identified and unidentified enzyme-catalyzed reactions, respectively. Enzymes involved in the pathways are: CFAT, coniferylacetyltransferase; IGS, isoeugenol synthase; DIR, dirigent protein; PLR, pinoresinol/lariciresinol reductase; CYP, cytochrome P450; OMT,*O*-methyltransferase; SIRD, secoisolariciresinol dehydrogenase; DOP7H, deoxypodophyllotoxin 7-hydroxylase; MOMT, matairesinol *O*-methyltransferase; SMGT, sesaminol 2-*O*-glucosyltransferase; DOP6H, deoxypodophyllotoxin 6-hydroxylase; βP6OMT, β-peltatin 6-*O*-methyl-transferase; PAM7H, β-peltatin-A-methylether 7-hydroxylase.

Considering that *S. sphenanthera* is a high value medical plant with limited available transcriptome data, we analyzed the *de novo* transcriptome of the *S. sphenanthera* leaf, root and fruit with the Illumina HiSeq^{TM} 2000 sequencing platform, which is a more effective strategy with the characteristics of low-cost and high-output. The present study aimed to enrich the genetic information of *S. sphenanthera* with these data and to predict several functional genes involved in pathways of lignan metabolism. Meanwhile, we selected a lot of simple sequence repeats (SSRs) markers, which were developed and used in marker-assisted breeding of this medicinal plant. This is the first report to analyze the transcriptome of *S. sphenanthera*. We believe that it can be a potential strategy to discover the candidate genes of main medicinal component biosynthesis in other non-model medicinal plants.

Materials and Methods

Plant material and RNA extraction: The fresh leafs, roots, and mature fruits of *S. sphenanthera* were collected from the experimental field of Luanchuan Forestry Administration, located in Luanchuan County, Henan Province, middle of China (33° 47' 1"N, 111° 36' 58"E, alt. 750 m). All the samples were frozen immediately after collection in liquid nitrogen, and stored at −80°C before utilization. The total RNA was extracted from all samples with the Trizol Kit (Promega, USA) according to the manufacturer’s instructions, and subsequently purified by RNeasy Mini Elute Cleanup Kit (Qiagen). The RNA quality and concentration were detected by 1% agarose gel and spectrophotometer, respectively. With the purpose of getting comprehensive gene information, at least 20 μg pooled RNA from all samples were prepared for a cDNA library construction and further *de novo* sequencing.

cDNA library construction and de novo sequencing: The mRNA was accumulated from the blended RNA using oligo (dT) magnetic beads, and fractionated into 200-700 bp with fragmentation buffer. Following this, the first and second cDNAs were synthesized successively with the standard protocol and the double-stranded cDNA was purified using QiaQuick PCR purification kit. Subsequently, end repair, poly (A) tailing and ligation of adapters were performed, and the fragment sizes were then selected using agarose gel electrophoresis. Finally, these products were amplified with PCR to construct the cDNA library, which was then sequenced using an Illumina Hi Seq™ 2000.

Data filtering and de novo assembly: The image data generated from the sequencing were transformed into sequence data (raw data or raw reads) via base calling. Then, readquality filtering was conducted, and clean reads were obtained by removal of low-quality reads with Perl script, including reads containing adaptors, having more than 50% bases with Q-value≤20, and having a frequency of unknown nucleotides more than 5%. Subsequently, transcriptome*de novo* assembly was carried out using the short read assembly program: Trinity (Haas *et al*., 2013; Pertea *et al*., 2003). We utilized the Trinity at the fixed default k-mer size of 25 to connect the reads with overlap into longer fragments without N, which were termed contigs. Following this, contigs were processed using sequence clustering software TGICL to splice sequences, remove redundant, and get sequences without N that were defined as unigenes.

Functional annotation and classification: The functional annotation was conducted utilizing various bioinformatics measures. Firstly, a BLASTX with an E-value cutoff of 1.0E-5 was performed between all assembled unigenes and the protein databases following the priority order non-redundant (NR) protein database (http://www.ncbi.nlm.nih.gov/), Swiss-Prot database (http://www.expasy.ch/sprot), the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (http://www.genome.jp/kegg), and the Cluster of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG/) (Altschul *et al*., 1997; Tatusov *et al*., 1997; Kanehisa *et al*., 2006). The protein with highest sequence similarity was annotated to corresponding unigene, followed by the remaining unaligned sequences consisting of unnamed, unknown, uncharacterized and hypothetical unigenes were discarded. Following the NR database annotation, Blast2GO software (Conesa *et al*., 2005) was used to analyze the GO terms: molecular function, cellular component and biological process, and all the GO annotations were calculated. Subsequently, we adopted the WEGO software (Ye *et al*., 2006) to classify the functions of all the unigenes and a macro level of the distribution characteristics of gene functions in *S. phenanthera* was cognized ultimately. The unigene sequences were also compared with COG database to forecast and cluster conceivable functions, and the conserved domains/families of our unigenes were performed following the Pfam database (version 26.0) employing Pfam_Scan program. Finally, we detected the pathway assignments following the KEGG database based on BLASTX with an E-value less than 1.0E-5.

CDS prediction: The protein coding sequence (CDS) prediction was carried out by aligning to the four protein databases according to the priority order NR, Swiss-Prot, KEGG, COG with BLASTX (E-value<1.0E-5). The highest rank of the BLASTX alignments was selected to detect the coding sequence of corresponding unigene, and coding sequence was translated into amino acid sequence relying on standard code table. Then, the Nucleotide (from 5-prime to 3-prime ends) and amino acid sequences of the coding region were generated. If a unigene could not match with any of those databases, the ESTScan (Iseli *et al*., 1999) software was recommended to predict coding region and its nucleotide and amino acid sequences.

SSR and primer design: The MISA tool (Thiel *et al*., 2003) (http://pgrc.ipk-gatersleben.de/misa/) was used to search for the SSR markers in the unigene sequences according the parameters: di-nucleotide, tri-nucleotide and tetra-nucleotide to hexa-nucleotide motifs with a minimum of six repetitions, five repetitions and four repetitions, respectively. Mono-nucleotide repeats were omitted due to the SSR among this motif type lacked practical application and were mismatched easily. Based on the searched SSR, Primer3 (Rozen *et al*., 2000) (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) was used to design Primer pairs following the design criterion: the annealing temperature 55 °C to 64°C, PCR product size 100-500 bp, the difference in TM value between the forward primer and reverse primer was less than 5°C, primer length 18 bp to 27 bp, GC content was between 40% and 60%, no secondary structures, including hairpin structures, dimers, mismatches. Based on these design criterion, 20 primer pairs were designed and synthesized for further potential investigation of SSR primers that were generated based on the *S. sphenanthera* transcriptome sequences.

DNAextraction, filter of EST-SSR primers and PCR amplification: We selected 8 *S. sphenanthera* superior plants without pests and robust to extract DNA, which was used for PCR amplification and validation with the abovementioned 20 primer pairs including 2-6 nucleotide repeats. Total DNA isolation from tender leaves was carried out with the modified CTAB method (Porebski *et al*., 1997). Because of the difference within the SSR primers, the PCR was performed with a gradient annealing temperature with the DNA from the three random superior plants in order to optimize the annealing temperature (Zheng *et al*., 2013).PCR was performed in a reaction (20 μL) containing 1 μL template DNA (about 60 ng/μL), 1 μL of each primer (10 μM), 8.5 μL 2 × Taq PCR MasterMix and 8.5 μL ddH$_2$O. The standard protocol included an initial denaturation for 5 minat 94°C, followed by 30 cycles of denaturation at 94°C for 60 s, annealing at Tm (annealing temperature) for 50 s and extension at 72°C for 1.5 min, with a final extension of 10 min at 72°C. A 1% agarose gel electrophoresis was utilized to determine the specificity of the EST-SSR Primers. The amplification of DNA from 8 *S. sphenanthera* superior plants was carried out with the optimized SSR primer sets, and the PCR products were separated on 8% polyacrylamide gels (Jiang *et al*., 2014), which were then stained with silver nitrate as described previously (Bassam *et al*., 1991). A 500 bp Marker (Dongsheng Biotech, Guangzhou co., LTD.) was used for calculating the size of the EST-SSR amplicons.

Results and Discussion

Illumina sequencing and reads assembly: To obtain a comprehensive overview of the transcriptome profile of *S. sphenanthera*, a cDNA library was prepared from the mixture of RNA extracted from fresh leaves, roots and mature fruits, and was then paired-end sequenced using the Illumina Hi Seq TM 2000 sequencing platform. Clean reads were screened from raw reads by removing the reads of low-quality, containing adapters and uncertain-base. We obtained a total of 87, 755, 638 clean reads, with a total of 8, 775, 563, 800 nucleotides, had the Q20 percentage (sequencing error rate<1%) of 99.09% and GC content of 46.70%. All these high-quality reads were deposited in the NCBI database and could be viewed under the accessing number: SRA175342. The detailed information generated from the *de novo* sequencing and assembled was shown in Table 1. The high-quality reads were assembled into 139,831 contigs with an N50 size of 918 bp. The contigs with lengths ranging from 201 to 11,203 bp and the average length was 596 bp, in which 46,054 contigs were >500 bp, 22,459 contigs>1000 bp in length and the large proportion of contigs (93,327) with the lengths between 200 and 500 bp (Fig. 2). Further treatment of splicing sequences and removing redundancy were conducted utilizing the TGICL sequence clustering software, we generated 129,951 unigenes with an average size of 570 bp ranging from 201 to 11,203 bp. There were 39,981 unigenes with length >500 bp, 19,067 unigenes with length >1000 bp, and the majority of unigenes (89,970) were 200-500 bp in size (Fig. 2). Among the total assembled unigenes, 76,460 (59.05%) had CDS that varied in size. The CDS size ranging from 200 to 500 bp (53,324) was the most predominant ones, followed by size >500 bp (23,136) and >1000 bp (9,764) (Fig. 2). This is the first comprehensive research of *S. sphenanthera* transcriptome, these abundant and high-quality data are helpful for enriching the genetic information and facilitating the research of regulation mechanism about active constituents in *S. sphenanthera*.

Table 1. Summary of Illumina Paired-end sequencing and assembly for *S. sphenanthera*.

Database	Number	Total length (bp)
Total clean reads	87,755,638	8,775,563,800
Q20 percentage	99.09%	
GC percentage	46.70%	
Number of contigs	139,831	83,334,949
Average length of contigs (bp)	595	
Max length of contigs (bp)	11,203	
Min length of contigs (bp)	201	
Contig size N50 (bp)	918	
Number of unigenes	129,951	74,198,804
Average length of unigenes (bp)	570	
Max length of unigenes (bp)	11,203	
Min length of unigenes (bp)	201	
Unigene size N50 (bp)	851	

Functional annotation: All the 129,951 unigenes generated by the *de novo* sequencing and assembly were searched against with the four public protein databases: NR, Swiss-Prot, KEEG and COG with an E-value threshold<1.0E^{-5} and a cut off similarity value of 17%. A total of 62,174 (47.80%) unigenes were successful matched (Table 2), among them, 12,391 unigenes could be annotated in all the four public protein databases and 14,090, 1393, 71, 131 unigenes could be matched the unique protein database of NR, Swiss-Prot, KEEG and COG, respectively (Fig. 3). Meanwhile, 67,777 (52.16%) unigenes could not be annotated against any database, indicating that these unigenes may be the novel unigenes. By comparing the unigene lengths between hit and no hit unigenes against the Nr and Swiss-Prot databases, we found that longer sequences were more likely to have BLASTx homologs in protein database. Particularly, in the Nr database, 95.03% of the sequences with length longer than 1000 bp had the significant matches, and the match ratio declined to 70.13% of sequences between 500 to 1000 bp and further down to

30.35% for those length ranging from 200 to 500 bp (Fig. 4A). A similar tendency was observed in Swiss-Prot database. The proportion of the sequences with length of > 1000 bp, 500 to 1000 bp, 200 to 500 bp was 81.09%, 49.56%, 22.46%, respectively (Fig. 4B). The E-value frequency distribution of the top hits against the Nr database indicated that 36.55% of the sequences had high homologies (E-value smaller than1.0E^{-50}); 63.45% of the matched sequences with the E-value between 1.0E^{-50} and 1.0E^{-5} (Fig. 5A). The transcriptome sequences also showed higher similarity comparing with sequences from the Nr Database. 88.16% of the top BLAST hits had the similarity ranging from 17% to 80%; 13.84% of the hits with similarity value more than 80% (Fig. 5C). We also matched the sequences against the protein database of Swiss-Prot, and it was found that 29.67% of the matched sequences showed significant homologies with E-value<1.0E^{-50} and the remaining 70.33% had E-value ranging from 1.0E^{-50} to 1.0E^{-5} (Fig. 5B). The similarity distribution against Swiss-Prot was similar to that against the Nr database. The majority of the sequences had the similarities between 17% and 80%; only 11.85% of sequences showed high homologies similarity value more than 80% (Fig. 5D). In addition, results further showed that 46.27% of the unigenes were significantly homologous to the sequences of *Vitis vinifera* (14,677, 24.41%), followed by *Theobroma cacao* (8,947, 14.88%), *Oryza sativa* (3,357, 5.58%), and *Cucumis sativus* (2,919, 4.85%) (Fig. 6). This suggests that the *S. sphenanthera* genome is more closely related to *V. vinifera* genome than to other model plant genomes.

Table 2. Summary of the annotation percentage of *S. sphenanthera* as compared to public database.

Database	Number of unigenes	Annotation percentage (%)
Nr	60,129	46.27
SwissProt	46,034	35.42
COG	22,972	17.68
KEGG	19,882	15.30
All	62,174	47.80
Total unigenes	129,951	

Gene ontology (GO) classification: To functionally categorize all the unigenes annotated into the Nr database, Gene Ontology (GO) analysis was performed. A total of 27,971 unigenes were assigned to at least one of the GO category and 14,8022 functional terms were generated. These functional terms were assigned into the main three GO categories: biological process (62,847, 42.46%), cellular component (55,651, 37.59%), and molecular function (29,524, 19.95%) with 43 sub-categories (Fig. 7). Within the biological process, the top three GO terms were metabolic process (15,228, 24.23%), cellular process (14,721, 23.42%), and response to stimulus (5,728, 8.40%). Among the cellular component, cell (17,466, 31.38%), cell part (17,466, 31.38%), organelle (12,795, 22.99%) were dominant terms. For the molecular function, the majority of the GO terms were classified into binding (13,535, 45. 84%) and catalytic activity (12,260, 41.53%) (Fig. 7; Table 3). These comprehensive GO annotations could be as important reference information for the researches of *S. sphenanthera*.

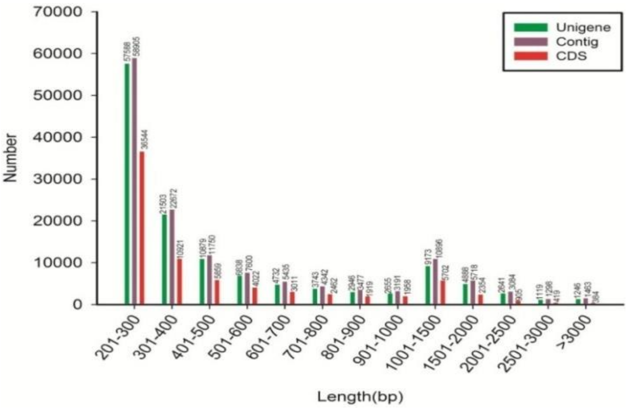


Fig. 2. Overview of the *S. sphenanthera* transcriptome assembly and the length distribution of the Unigene, Contig and CDS.

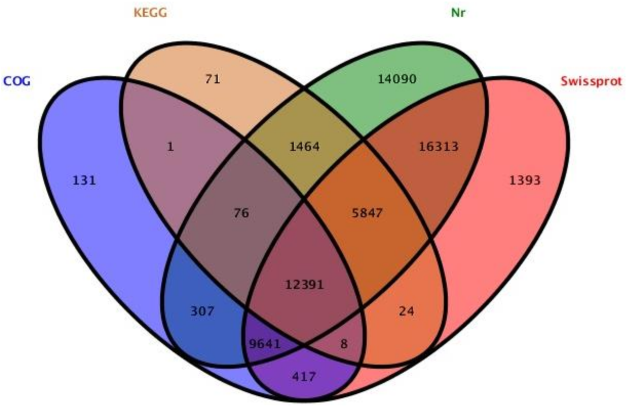


Fig. 3. Venn diagram results from diverse databases. Venn diagram of number of unigenesan notated by BLASTX with an E-value threshold of 10^{-5} against protein databases. The numbers in the circles indicate the number of unigenes annotated by single or multiple databases.

Conserved domain annotationand COG classification: In order to perfect the accuracy of our transcriptome annotations, we used conserved domains/families as an annotation standard rather than other annotations that simply based on the length and depth of sequences for further researching. In total, 52,011 unigenes were classified into 4,099 domains/ families. Within these protein domains/families, PPR repeat, leucine rich repeat, PPR repeat family, protein kinase domain, WD domain, G-beta repeat were the primary five represented domains. The top 15 Pfam domains/families of *S. sphenanthera* were listed in Fig. 8. Simultaneously, the unigenes were compared with COG for exploring theorthology genes of these assembled sequences. Based on the functional annotation, we got 22,972 (17.68%) unigenes possessing high consistency with the COG clusters. Since some individual unigenes maybe annotated into multiple COG functional categories, 40,840 functional annotations were obtained, which were further subdivided into 25 subclasses. Among these functional categories, general function prediction only (6,407, 15.69%) was the largest group, followed by translation, ribosomal structure and biogenesis (3,930, 9.62%), transcription (3,447, 8.44%), posttranslational modification, protein turnover, chaperones (3,251, 7.96%), replication, recombination and repair (3,162, 7.74%). The COG function group of RNA processing and modification (274, 0.67%), extracellular structures (13, 0.03%), nuclear structure (9, 0.02%) were the three smallest represented groups (Fig. 9).

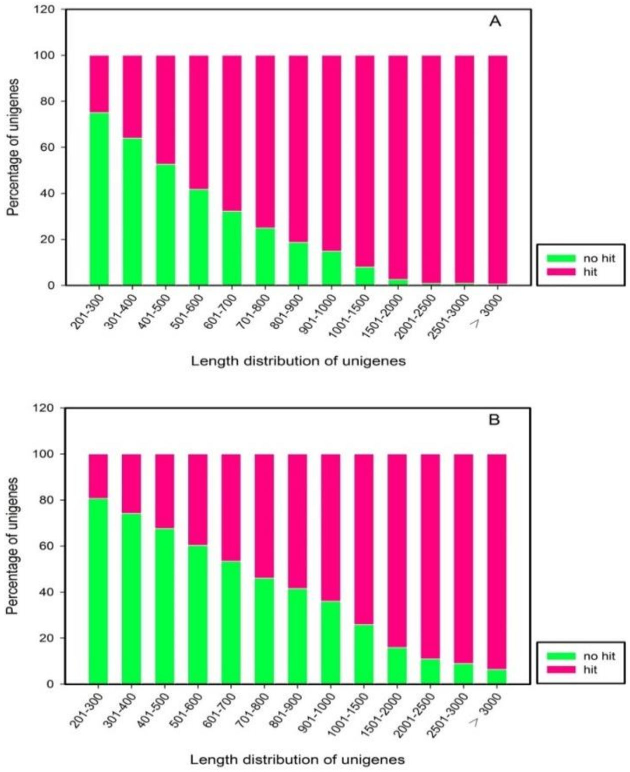


Fig. 4. Comparison of unigene length between hit and no hit unigenes. (A) Comparison of unigene length between hit and no hit unigenes in the Nr databases. (B) Comparison of unigene length between hit and no hit unigenes in the Swiss-prot database. Longer unigenes were more likely to have BLASTx homologs in protein database. In this study, more than 81% of unigenes over 1000 bp in length had BLAST matches, whereas only less than 22% of unigenes shorter than 500bp did.

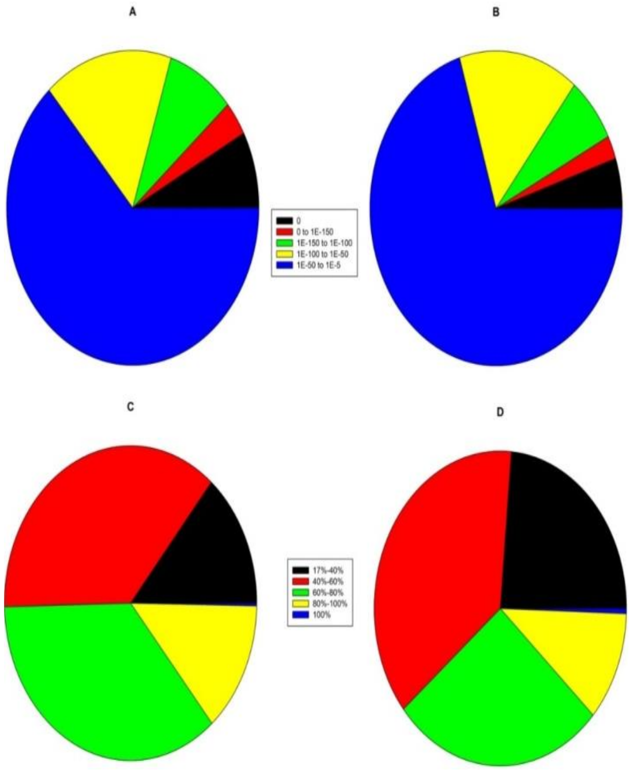


Fig. 5. Characterization of searching the assembled unigenes against NCBI Nr and Swiss-Prot protein databases. (A) E-value proportional frequency distribution of BLAST hits against the Nr database. (B) E-value proportional frequency distribution of BLAST hits against the Swiss-Prot database. (C) Similarity distribution of the top BLAST hits for the assembled unigenes with a cut off of 1E-5 in Nr database. (D) Similarity distribution of the top BLAST hits for the assembled unigenes with a cut off of 1E-5 in Swiss-Prot database.

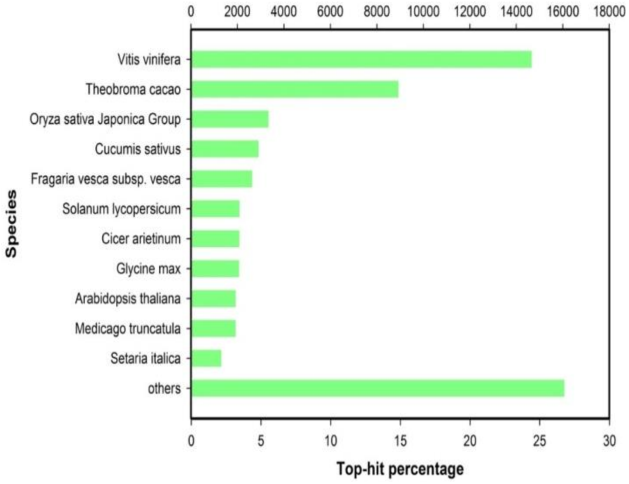


Fig. 6. Top-hit species distribution for sequences from *S. sphenanthera* submitted BLASTX against the NCBI-Nr database.

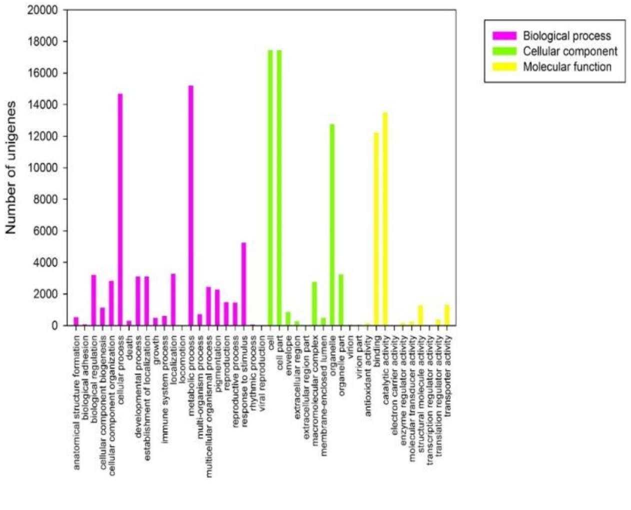


Fig. 7. Gene Ontology classification of assembled unigenes. Total 27,971 unigenes were categorized into three main categories: biological process, cellular component and molecular function.

KEGG pathway mapping: For researching the biological pathways of *S. sphenanthera*, we mapped all the unigenes against the specific pathways in KEGG database using BLASTX with an E-value threshold<1.0E $^{-10}$ and the matched unigenes were assigned into the corresponding KEGG Pathway based on the EC numbers. We found that a total of 19,882 unigenes were annotated into KEGG database, in which 9,408 unigenes could be mapped to a single Enzyme Commission (EC) number and 34,937 matches (several unigenes may assigned into multiple pathways) were classified into 275 KEGG pathways. Among them, the top five KEGG pathways groups were ribosome (1,846, 5.28%), protein processing in endoplasmic reticulum (900, 2.58%), spliceosome (801, 2.29%), RNA transport (785, 2.25%), purine metabolism (751, 2.15%).Among the metabolism pathway, the primary represented subclass was carbohydrate metabolism (2,724, 24.10%), followed by amino acid metabolism (1,822, 16.12%), energy metabolism (1,413, 12.50%), lipid metabolism (1,186, 10.49%), nucleotide metabolism(1,094, 9.68%), metabolism of cofactors and vitamins (554, 4.90%), metabolism of other amino acids (545, 4.82%), xenobiotics biodegradation and metabolism (532, 4.71%), biosynthesis of other secondary metabolites (511, 4.52%), metabolism of terpenoids and polyketides (490, 4.33%), glycan biosynthesis and metabolism (433, 3.83%) (Fig. 10A). We focused on the analysis about the information of biosynthesis of other secondary metabolites, because the lignan was the product of secondary metabolites (Kim*et al*., 2009). Within the biosynthesis of other secondary metabolites, phenylpropanoid biosynthesis had the largest number of unigenes (210, 41.10%), followed by flavonoid biosynthesis (59, 11.55%), streptomycin biosynthesis (45, 8.81%), tropane, piperidine and pyridine alkaloid biosynthesis (42, 8.22%), isoquinoline alkaloid biosynthesis (41, 8.02%), stilbenoid, diarylheptanoid and gingerol biosynthesis (38, 7.44%), novobiocin biosynthesis (25, 4.89%), caffeine metabolism (12, 2.35%), butirosin and neomycin biosynthesis (20, 3.91%), flavone and flavonol biosynthesis (9, 1.76%), penicillin and cephalosporin biosynthesis (3, 0.59%), anthocyanin biosynthesis (2, 0.39%), betalain biosynthesis (2, 0.39%), indole alkaloid biosynthesis (2, 0.39%), glucosinolate biosynthesis (1, 0.20%) (Fig. 10B). It is noteworthy that phenylpropanoid and caffeine were generally considered as a putative precursor in the biosynthesis of lignan among plants. In addition to 126 metabolism pathways, we obtained 7,404 and 3,779 unigenes corresponding to genetic information processing and cellular processes, respectively (Table 4). The KEGG pathway mapping, along with the functional annotation, GO classification and COG analysis provide abundantly useful information contributing to further researching in *S. sphenanthera*.

The distribution and frequency of SSR: In order to develop new molecular marker with the SSR generating from the *S. sphenanthera* transcriptome and evaluate the potential SSR, all the 129,951 assembled unigenes were detected using MISA. In total, 13,427 SSR were identified in 11,860 sequences, of which 1,362 sequences had more than one SSR, and 610 SSR were presented in compound form. Frequency of occurrences for SSR was 10.33%, and distribution density was on average 1/5.52 kb (Table 5). The detailed information of SSR gained from assembled unigenes could be examined. Within the SSR, di-nucleotide repeats (9,467, 70.51%) constituted the primary nucleotide type, which is similar to the previous studies about other plants (Wei *et al*., 2011; Kumpatla & Mukhopadhyay., 2005), followed by tri-nucleotide (2,974, 22.15%), tetra-nucleotide (562, 4.19%), and hexa-nucleotide (263, 1.96%), penta-nucleotide (161, 1.20%) motifs (Table 6).

We also researched the SSR frequency based on motif forms. Among the di-nucleotide repeats, AG/CT (58.22%) represented the most abundant one, followed by AC/GT (7.73%) and AT/AT (4.43%). Usually, CG motif was rarely found in plants, but in *S. sphenanthera*, 17 CG motifs were generated, which might represent an important new finding and it biological significance needs to be explored in the future. In the tri-nucleotides, AAG/CTT (7.64%) formed the largest repeat group, followed by the ACC/GGT (3.00%), ATC/ATG (2.85%) and AGG/CCT (2.77%) (Fig. 11). In addition, 14,973 primer pairs were designed using Primer premier 6.0 software. Currently, most researchers utilized the traditional methods to develop molecular markers that were high consume and low efficiency, and only 43 available expressed sequence tags of *S. sphenanthera* could be searched in the NCBI (until July 1, 2014), which severely limited the study of molecular marker-assisted selection. Thus, we believe that the identified 13,427 SSR and the designed 14,973 primer pairs in our research will play an important role in the future study on genetic diversity, marker-assisted breeding, and genetic map of *S. sphenanthera*.

Table 3. Gene Ontology classification.

Gene Ontology	Class	Total numbers
Biological process	Anatomical structure formation	555
	Biological adhesion	103
	Biological regulation	3238
	Cellular component biogenesis	1165
	Cellular component organization	2863
	Cellular process	14721
	Death	322
	Developmental process	3144
	Establishment of localization	3135
	Growth	509
	Immune system process	628
	Localization	3313
	Locomotion	3
	Metabolic process	15228
	Multi-organism process	749
	Multicellular organismal process	2471
	Pigmentation	2299
	Reproduction	1520
	Reproductive process	1489
	Response to stimulus	5278
	Rhythmic process	100
	Viral reproduction	14
Cellular component	Cell	17466
	Cell part	17466
	Envelope	903
	Extracellular region	289
	Extracellular region part	9
	Macromolecular complex	2787
	Membrane-enclosed lumen	510
	Organelle	12795
	Organelle part	3266
	Virion	80
	Virion part	80
Molecular function	Antioxidant activity	174
	Binding	12260
	Catalytic activity	13535
	Electron carrier activity	9
	Enzyme regulator activity	197
	Molecular transducer activity	261
	Structural molecule activity	1311
	Transcription regulator activity	25
	Translation regulator activity	411
	Transporter activity	1341

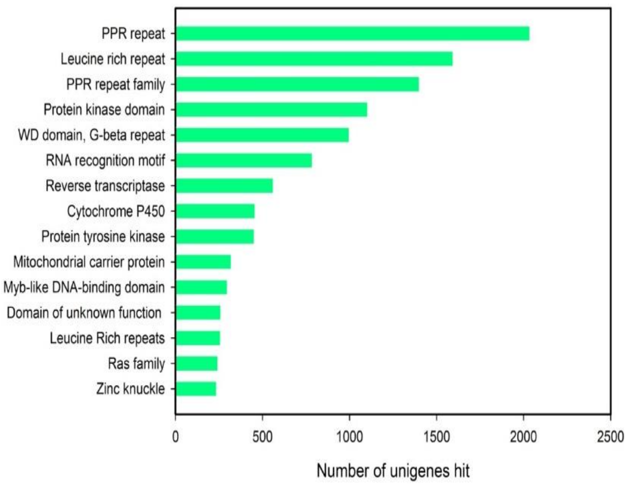


Fig. 8. Top 15 Pfam domains/families predcted in *S. sphenanthera*.

Table 4. Mapping of *S. sphenanthera* unique sequences to KEGG biochemical pathways.

KEGG categories represented	No. of uniques
Metabolism	11,304
Carbohydrate metabolism	2,724
Amino acid metabolism	1,822
Energy metabolism	1,413
Lipid metabolism	1,186
Nucleotide metabolism	1,094
Metabolism of cofactors and vitamins	554
Metabolism of other amino acids	545
Xenobiotics biodegradation and metabolism	532
Biosynthesis of other secondary metabolites	511
Metabolism of terpenoids and polyketides	490
Glycan biosynthesis and metabolism	433
Genetic Information Processing	7,404
Translation	3,630
Folding, sorting, and degradation	2,094
Transcription	1,050
Replication and repair	630
Environmental information processing	1,911
Signal transduction	1,861
Membrane transport	42
Signaling molecules and interaction	8
Cellular processes	3,779
Transport and catabolism	1,576
Cell growth and death	1,264
Cell communication	635
Cell motility	304
Organismal systems	5,077
Human diseases	5,462

The validation and evaluation of SSR markers in *S. sphenanthera*: In this study, 20 primer pairs were designed to validate and evaluate the potential of SSR markers generating from *S. Sphenanthera* transcriptome sequences (Table 7). 18primer pairs (90%) successfully yielded amplified bands. This result was consistent with previous studies from *Lpomoeabatatas*, *Arachis hypogaea* L. etc. (Liang *et al*., 2009; Wang *et al*., 2011; Cordeiro *et al*., 2001; Yu *et al*., 2004). Among the 18 successful primer pairs, we obtained 15 (83.3%) polymorphic SSR markers, for which the band sizes were very close to our expected fragments (Fig. 12). The sizes of the PCR products with the other three primer pairs were quite different from the expected sizes, which may be caused by large introns, chimeric primers, large insertions or assembly errors (Saha *et al*., 2004; Varshney *et al*., 2005). These results suggested that the SSR identified in our dataset were suitable for specific primer design and could be used as a new means of molecular markers in the future. Of the 15 polymorphic SSR markers, 12 SSR markers had the length of 100-250 bp and the main repeat units were di-nucleotide and tri-nucleotide repeats. Therefore, the sequences with length ranging from 100 bp to 250 bp and containing low-level repeat units should be selected to design SSR primers for higher polymorphism and better amplification (Dreisigacker *et al*., 2004). The majority of our transcriptome sequences had the length between 200 bp and 500 bp and the di-nucleotide repeats were the most common type, with a frequency of 70.51% (9,467), hence, we believe that more SSR primers could be designed based on the 13,427 SSR identified in our dataset in the future as tools for assessment of germplasm polymorphism, mapping of quantitative trait loci, and cloning of functional genes in *S. sphenanthera*.

Table 5. Summary of SSR searching results.

Item	Number
Total number of sequences examined	129,951
Total size of examined sequences (bp)	74,198,804
Total number of identified SSRs	13,427
Number of SSR containing sequences	11,860
Average number of SSRs per 10 kb	1.81
Number of sequences containing more than 1 SSR	1,362
Number of SSRs present in compound formation	610

Transcripts encoding enzymes involved inlignanbiosynthesis: In plants, most lignans with oxygen at C9 (C9’) are formed by enantioselective dimerization of two coniferyl alcohol units (Umezawa, 2003), all of the genes encoding enzymes involved in coniferyl alcohol (Humphreys & Chapple, 2002) were found in this Illumina dataset, including PAL (phenylalanine ammonia-lyase), C4H (cinnamate 4-hydroxylase), C3H (*p*-coumarate 3-hydroxylase) , COMT (caffeate *O*-methyltransferase), 4CL (4-coumarate:CoA ligase), CCR (cinnamoyl CoA reductase), CAD (cinnamyl alcohol dehydrogenase), CCoAOMT (caffeoyl CoA *O*-methyltransferase), CQT (hydroxycinnamoyl CoA: quinate hydroxycinnamoyl transferase) and CST (hydroxycinnamoyl CoA: shikimate hydroxycinnamoyl transferase) (Table 8). Considering that dibenzocyclooctadiene-type lignan is the major active component of *S. sphenanthera*, we focused on the unigenes that encode enzymes involving in lignan biosynthesis. First, we found 3 transcripts annotated to isoeugenol synthase (IGS) catalyzing the conversion of coniferyl acetate to isougenol, the precursor of dibenzocyclooctadiene-type lignan (Dexter *et al*., 2007) (Fig. 1). In addition, the mRNAs encoding all existing enzymes involved in lignan biosynthesis were also identified, including dirigent protein (DIR), pinoresinol/lariciresinol reductase (PLR), SIR (secoisolariciresinol dehydrogenase). These results suggested that dibenzocyclooctadiene-type lignan might be biosynthesized in the similar pathways of other lignans. Characterization of the functions of those unigenes will undoubtedly help us better understand the molecular mechanism of lignan biosynthesis in *S. sphenanthera*.

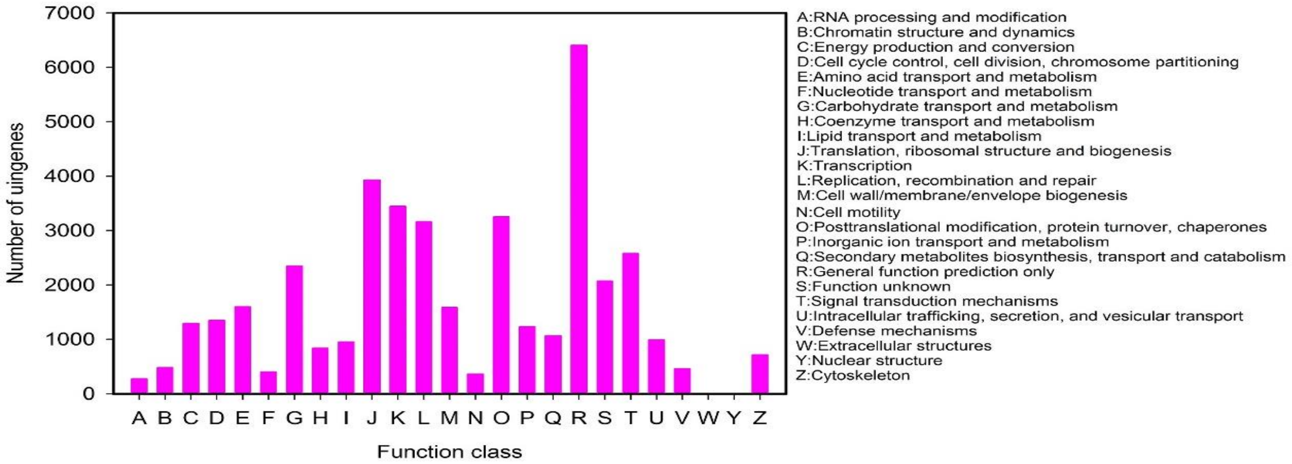


Fig. 9. COG function classification of *S. sphenanthera* unigenes.

Table 6. Distribution of identified SSRs using the MISA software.

Motif	Repeat numbers									Total	%
	4	5	6	7	8	9	10	11	≥11		
Di-	0	0	1,619	1,329	1,661	2,716	1,727	375	40	9,467	70.51
Tri-	0	1,650	792	481	46	2	0	1	2	2,974	22.15
Tetra-	427	118	15	0	1	1	0	0	0	562	4.19
Penta-	148	13	0	0	0	0	0	0	0	161	1.20
Hexa-	232	19	3	4	4	1	0	0	0	263	1.96
Total	807	1,800	2,429	1,814	1,712	2,720	1,727	376	42	13,427	100.00
%	6.01	13.41	18.09	13.51	12.75	20.26	12.86	2.80	0.31	100.00	

Table 7. *S. sphenanthera* for validation and evaluation with EST-SSRs.

Code	Source
LC1	Luanchuan, Henan, China
LC2	Luanchuan, Henan, China
LC3	Luanchuan, Henan, China
LC4	Luanchuan, Henan, China
LC5	Luanchuan, Henan, China
LC6	Luanchuan, Henan, China
LC7	Luanchuan, Henan, China
LC8	Luanchuan, Henan, China

Cytochrome P450 CYP81Q1, also called piperitol/sesamin synthase (PSS) catalyzes the formation of two methylenedioxy bridges for biosynthesis of semamin (Ono *et al*., 2006). The similar cytochrome P450 enzymes were also found in *Podophyllum hexandrum* (CYP719A23) and *P. peltatum* (CYP719A24), both of them are able to convert (—)-matairesinol into (—)-pluviatolide by catalyzing methylenedioxy bridge formation (Marques *et al*., 2013). In the Illumina dataset, 1 unigene (unigene 0023187) is very close to CYP81Q1 (with the similarity of 51.25% (Figs. 13 & 14) might catalyze the conversion of dihydroguaiaretic acid to anwulignan in *S. sphenanthera* (Fig. 1). Like other lignans, dibenzocyclooctadiene-type lignans also showed various *O*-methylation patterns, until now, only two *O*-methyltransferase (OMT) enzymes that catalyze lignin *O*-methylation have been found. One of them, named *Carthamus tinctorius* matairesinol (CtMROMT), catalyzes the methylation of matairesinol (Umezawa *et al*., 2013), another is *Anthriscus sylvestris* thujaplicatin OMT (AsTJOMT), which catalyzes regioselective methylation of thujaplicatin to produce 5-*O*-methylthujaplicatin (Ragamustari *et al*., 2013). In the Illumina dataset, only 1 unigene (unigene 0079412) is close to AsTJOMT, thus this unigene is regarded as candidate of lignan OMT in *S. sphenanthera* (Fig. 15).

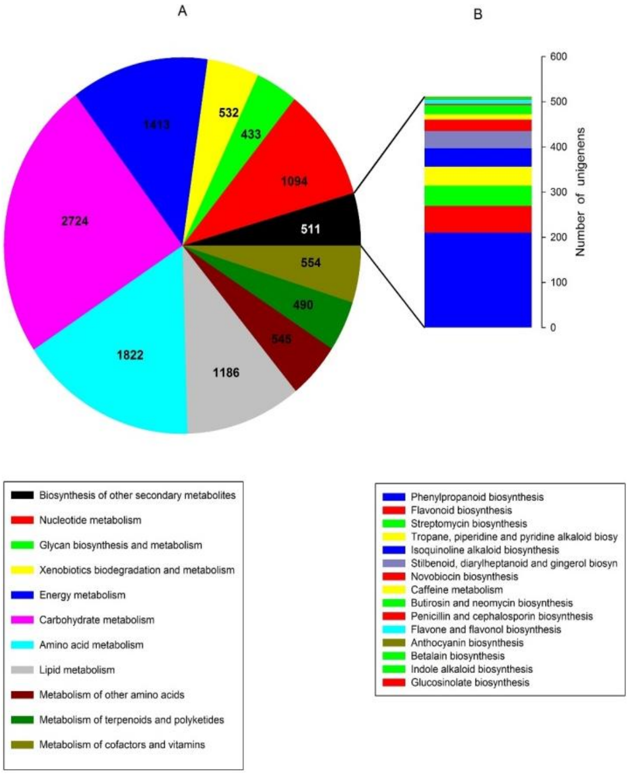


Fig. 10. Pathway assignment based on KEGG. (A) Classification based on metabolism categories; (B) Classification based on amino acid metabolism categories.

Table 8. The information of genes involved in coniferyl alcohol and lignan biosynthesis in *S. sphenanthera*.

Gene name	EC number	Unigene numbers
coniferyl alcohol biosynthesis		
PAL, phenylalanine ammonia-lyase	4.3.1.24	15
C4H, cinnamate 4-hydroxylase	1.14.13.11	5
C3H, *p*-coumarate 3-hydroxylase	1.14.13.-	3
COMT, caffeate*O*-methyltransferase	2.1.1.68	1
4CL, 4-coumarate:CoA ligase	6.2.1.12	40
CCR, cinnamoyl CoA reductase	1.2.1.44	27
CAD, cinnamyl alcohol dehydrogenase	1.1.1.195	22
CCoAOMT, caffeoyl CoA *O*-methyltransferase	2.1.1.104	8
CQT, hydroxycinnamoyl CoA: quinatehydroxycinnamoyltransferase	2.3.1.99	9
CST, hydroxycinnamoyl CoA: shikimatehydroxycinnamoyltransferase	2.3.1.133	7
Lignanbiosynthesis		
DIR, dirigent protein		1
PLR,pinoresinol/lariciresinol reductase		4
SIRD,secoisolariciresinol dehydrogenase	1.1.1.331	2
IGS, isoeugenol synthase	1.1.1.319	3

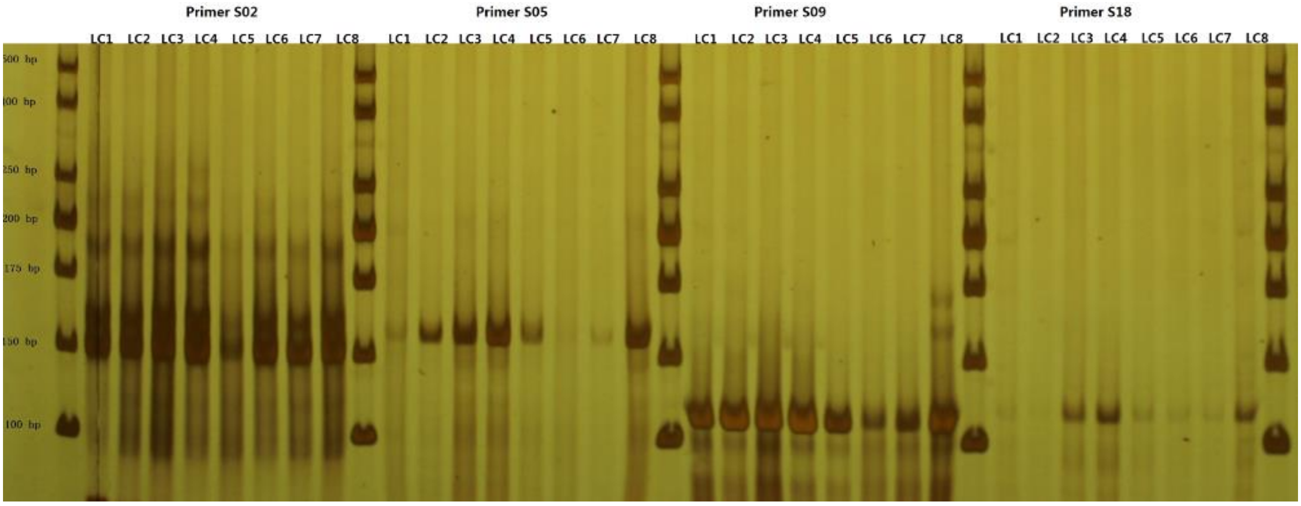
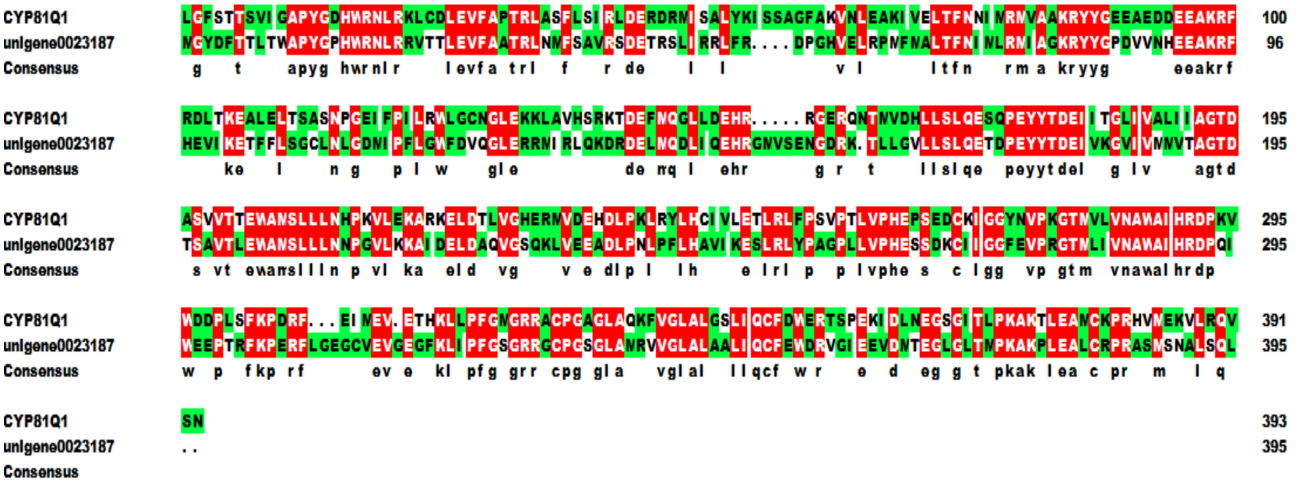

Primer S02
Primer S05
Primer S09
Primer S18
LC1 LC2 LC3 LC4 LC5 LC6 LC7 LC8
LC1 LC2 LC3 LC4 LC5 LC6 LC7 LC8
LC1 LC2 LC3 LC4 LC5 LC6 LC7 LC8
LC1 LC2 LC3 LC4 LC5 LC6 LC7 LC8
500 bp
00 bp
250 bp
200 bp
175 bp
150 bp
100 bp

Fig. 12. Amplification of primer S02, S05, S09 and S18 showed in eight clones of *S. sphenanthera*. The bands size of these SSR primers with length ranging from 100 bp to 250 bp and were very close to the length of transcriptome data.

Among all the aligned sequences.



```
CYP81Q1         LGFSTTSVIGAPYGDHWRNLRKLCDLEVFAPTRLASFLSIRLDERDRMISALYKISSAGFAKVNLEAKIVELTFNNIMRMVAAKRYYGEEAEDDEEAKRF   100
unigene0023187  MGYDFTTLTWAPYGPHWRNLRRVTTLEVFAATRLNMFSAVRSDETRSLIRRLFR....DPGHVELRPMFNALTFNIMLRMIAGKRYYGPDVVNHEEAKRF    96
Consensus        g   t    apyg hwrnlr    levfa trl  f   r de    i  l          v l      ltfn   rm a kryyg      eeakrf

CYP81Q1         RDLTKEALELTSASNPGEIFPILRWLGCNGLEKKLAVHSRKTDEFMQGLLDEHR.....RGERQNTMVDHLLSLQESQPEYYTDEIITGLIVALIIAGTD   195
unigene0023187  HEVIKETFFLSGCLNLGDMIPFLGWFDVQGLERRMIRLQKDRDELMQDLIQEHRGMVSENGDRK.TLLGVLLSLQETDPEYYTDEIVKGVIVMMVTAGTD   195
Consensus           ke   l    n g   p l w    gle          de mq l  ehr      g r  t    llslqe  peyytdei  g iv    agtd

CYP81Q1         ASVVTTEWANSLLLNHPKVLEKARKELDTLVGHERMVDEHDLPKLRYLHCIVLETLRLFPSVPTLVPHEPSEDCKIGGYNVPKGTMVLVNAWAIHRDPKV   295
unigene0023187  TSAVTLEWANSLLLNNPGVLKKAIDELDAQVGSQKLVEEADLPNLPFLHAVIKESLRLYPAGPLLVPHESSDKCIIGGFEVPRGTMLIVNAWAIHRDPQI   295
Consensus        s vt ewansllln p vl ka  eld  vg    v e dlp l  lh    e lrl p  p lvphe s  c igg  vp gtm  vnawaihrdp

CYP81Q1         WDDPLSFKPDRF...EIMEV.ETHKLLPFGMGRRACPGAGLAQKFVGLALGSLIQCFDWERTSPEKIDLNEGSGITLPKAKTLEAMCKPRHVMEKVLRQV   391
unigene0023187  WEEPTRFKPERFLGEGCVEVGEGFKLIPFGSGRRGCPCSGLAWRVVGLALAALIQCFEWDRVGIEEVDNTEGLGLTMPKAKPLEALCRPRASMSNALSQL   395
Consensus       w  p  fkp rf      ev e  kl pfg grr cpg gla   vglal  liqcf w r   e  d  eg g t pkak lea c pr  m   l q

CYP81Q1         SN                                                                                                     393
unigene0023187  ..                                                                                                     395
Consensus
```

Fig. 14. Alignment of amino acid sequences of the putative *S. sphenanther*a CYPs with cytochrome P450 [*Sesamum indicum*] (BAE48234.1). Identical amino acid residues are shaded in red blue. Light green shade indicates 50% or more identity.

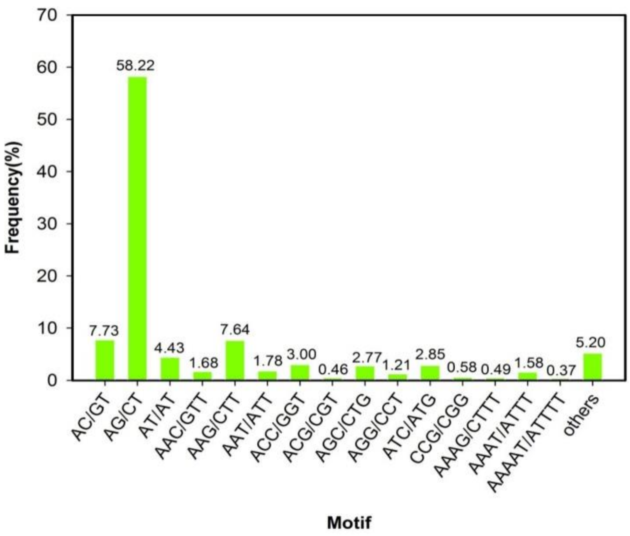


Fig. 11. Frequency distribution of SSRs based on motif types. The AG/CT di-nucleotide repeat motif was the most abundant motif detected.

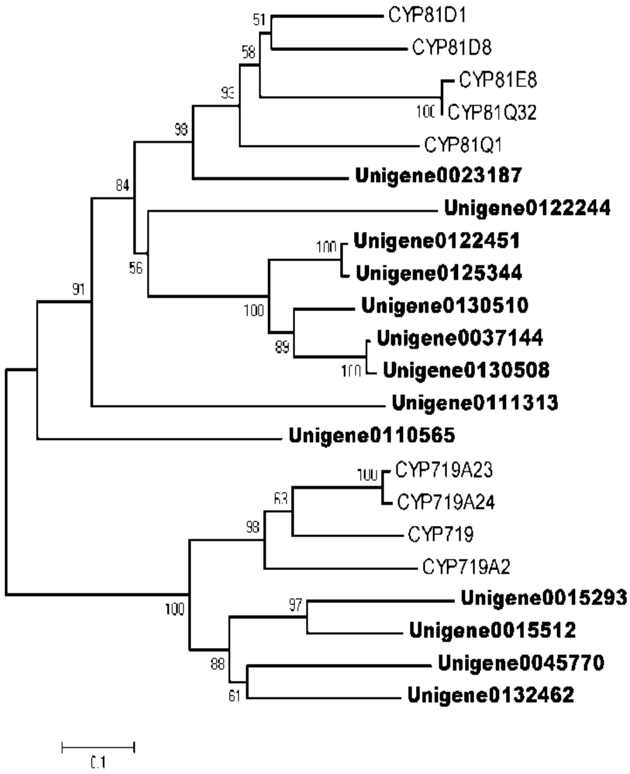


Fig. 13. Phylogenetic tree of the *S. sphenanthera* CYPs. Phylogenetic tree constructed based on the deduced amino acid sequences for the *S. sphenanthera* CYPs (bold letters) and other plant CYPs. Protein sequences were retrieved from NCBI GenBank using the following accession numbers (source organism and proposed function, if any, are given in parentheses): CYP81D1, EXB59542.1 (*Morus notabilis*); CYP81D8, XP_002866952.1 (*Arabidopsis lyrata* subsp. *lyrata*); CYP81E8, AAQ20042.1 (*Medicago truncatula*); CYP81Q32, AHK60837.1 (*Catharanthus roseus*); CYP81Q1, BAE48234.1 (*Sesamum indicum*); CYP719A23, AGC29953.1 (*Sinopodophyllum hexandrum*); CYP719A24, AGC29954.1 (*Podophyllum peltatum*); CYP719, AAU20771.1 (*Thalictrum flavum* subsp. *glaucum*, (S)-canadine synthase); CYP719A2, ACO90219.1 (*Eschscholzia californica*, stylopine synthase).

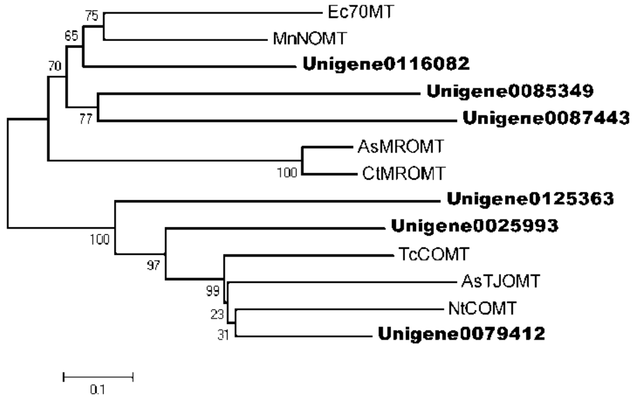


Fig. 15. Phylogenetic tree of the *S. sphenanthera* OMTs. Phylogenetic tree constructed based on the deduced amino acid sequences for the *S. sphenanthera* OMTs (bold letters) and other plant OMTs. Protein sequences were retrieved from NCBI GenBank using the following accession numbers (source organism and proposed function, if any, are given in parentheses): Ec7OMT, BAE79723.1 (*Eschscholzia californica*, reticuline-7-*O*-methyltransferase); MnNOMT, EXB55626.1 (*Morusnotabilis*, (RS)-norcoclaurine 6-*O*-methyltransferase); AsMROMT, BAO79381.1 (*Anthriscus sylvestris*, matairesinol *O*-methyltransferase); CtMROMT, BAN63362.1 (*Carthamus tinctorius*, matairesinol *O*-methyltransferase); AsTJOMT, BAO79384.1 (*A. sylvestris*, thujaplicatin *O*-methyltransferase); NtCOMT, CAA50561.1 (*Nicotiana tabacum*, catechol *O*-methyltransferase); TcCOMT, XP_007019090.1 (*Theobroma cacao*, caffeic acid 3-*O*-methyltransferase 1).

Conclusions

Using Illumina Hi Seq TM 2000 sequencing platform, a total of 129,951 assembled unigenes were generated that could be used as a valuable resource for enriching the transcriptomic and genomic data. Meanwhile, we found several functional genes involved in lignan biosynthesis and these data could provide solid support for studying the biosynthetic pathway and regulation mechanism of lignan. In addition, we identified 13,427 SSR and designed 14,973 primer pairs. Subsequently, we randomly selected 20 primer pairs for validating and evaluating the potential of SSR markers, and most of them could yield amplified bands and had higher polymorphism. This reveals that based on the transcriptome data we could develop new SSR molecular markers that will contribute to study population genetic structure, diversity analysis, linkage mapping and germplasm characterization analysis in *S. sphenanthera*. To the best of our knowledge, this is the first comprehensive research about the *S. sphenanthera* transcriptome analysis, and we believe it could serve as a valuable database for further research about this medicinal plant.

References

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*., 17: 3389-3402.

Bassam, B.J., G. Caetano-Anolles and P.M. Gresshoff. 1991. Fast and sensitive silver staining of DNA in polyacrylamide gels. *Anal. Biochem*., 196(1): 80-83.

Bayindir, U., A.W. Alfermann and E. Fuss. 2008. Hinokinin biosynthesis in *Linumcorymbulosum* Reichenb. *Plant J.*, 55(5): 810-820.

Chen, D.F., S.X. Zhang, M. Kozuka, Q.Z. Sun, J. Feng, Q. Wang, T. Mukainaka, Y. Nobukuni, H. Tokuda, H. Nishino, H.K. Wang, S.L. Morris Natschke and K.H. Lee. 2002. Interiotherins C and D, two new lignans from *Kadsura interior* and antitumor-promoting effects of related neolignans on Epstein-Barr virus activation. *J. Nat. Prod.*, 65(9): 1242-1245.

Chen, D.F., S.X. Zhang, L. Xie, J.X. Xie, K. Chen, Y. Kashiwada, B.N. Zhou, P. Wang, L.M. Cosentino and K.H. Lee. 1997. Anti-AIDS agents--XXVI. Structure-activity correlations of gomisin-G-related anti-HIV lignans from Kadsura interior and of related synthetic analogues. *Bioorg Med. Chem.*, 5(8):1715-1723.

Chen, M., N. Kilgore, K.H. Lee and D.F. Chen. 2006. Rubrisandrins A and B, lignans and related anti-HIV compounds from *Schisandra rubriflora*. *J. Nat. Prod.*, 69(12):1697-1701.

Chinese Pharmacopoeia Commission. 2010. The Pharmacopoeia of the People's Republic of China Version. Chin. Med. Sci. Technol. Press, Beijing.

Conesa, A., S. Götz, J.M. García-Gómez, J. Terol, M. Talón and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18): 3674-3676.

Cordeiro, G.M., R. Casu, C.L. McIntyre, J.M. Manners and R.J. Henry. 2001. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci*., 160(6): 1115-1123.

Dexter, R., A. Qualley, C.M. Kish, C.J. Ma,T. Koeduka, D.A. Sagegowda, N. Dudareva, E. Pichersky and D. Clark. 2007. Characterization of a petunia acetyltransferase involved in the biosynthesis of the floral volatile isoeugenol. *Plant J.*, 49: 265-275.

Dinkova-Kostova, A.T., D.R. Gang, L.B. Davin, D.L. Bedgar, A. Chu and N.G. Lewis. 1996. (+)-Pinoresinol/ (+)-lariciresinol reductase from *Forsythia intermedia*. Protein purification, cDNA cloning, heterologous expression and comparison to isoflavone reductase. *J. Biol. Chem.*, 271(46): 29473-29482.

Dreisigacker, S., P. Zhang, M.L. Warburton, M. Van Ginkelc, D. Hoisingtonc, M. Bohnb and A.E. Melchinger. 2004. SSR and Pedigree analyses of genetic diversity among CIMMYT wheat lines targeted to different Mega environments. *Crop Science*, 44(2): 381-388.

Federolf, K., A.W. Alfermann and E. Fuss. 2007. Aryltetralin-lignan formation in two different cell suspension cultures of *Linum album*: deoxypodophyllotoxin 6-hydroxylase, a key enzyme for the formation of 6-methoxypodophyllotoxin. *Phytochemistry*, 68(10): 1397-1406.

Fong, W.F., C.K. Wan, G.Y. Zhu, A. Chattopadhyay, S. Dey, Z. Zhao and X.L. Shen. 2007. Schisandrol A from *Schisandra chinensis* reverses P-glycoprotein-mediated multidrug resistance by affecting Pgp-substrate complexes. *Planta Med.*, 73(3): 212-220.

Haas, B., A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. MacManes, M. Ott, J. Orvis, N Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. LeDuc, N. Friedman and A. Regev. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8: 1494-1512.

Hano, C., I. Martin, O. Fliniaux, B. Legrand, L. Gutierrez, R.R.J. Arroo, F. Mesnard, F. Lamblin and E. Lainé. 2006. Pinoresinol-lariciresinol reductase gene expression and secoisolariciresinol diglucoside accumulation in developing flax (*Linum usitatissimum*) seeds. *Planta*, 224(6): 1291-1301.

Hemmati, S., T.J. Schmidt and E. Fuss. 2007a. (+)-Pinoresinol/(-)-lariciresinol reductase from *Linumperenne* Himmelszelt involved in the biosynthesis of justicidin B. *FEBS Lett*, 581(4):603-610.

Hemmati, S., B. Schneider, T.J. Schmidt, K. Federolf, A.W. Alfermann and E. Fuss. 2007b. Justicidin B 7-hydroxylase, a cytochrome P450 monooxygenase from cell cultures of *Linumperenne* Himmelszelt involved in the biosynthesis of diphyllin. *Phytochemistry*, 68(22-24): 2736-2743.

Hemmati, S., C.B. von Heimendahl, M. Klaes, A.W. Alfermann, T.J. Schmidt and E. Fuss. 2010. Pinoresinol-lariciresinol reductases with opposite enantio specificity determine the enantiomeric composition of lignans in the different organs of *Linum usitatissimum* L. *Planta Med*., 76(9): 928-934.

Humphreys, J.M. and C. Chapple. 2002. Rewriting the lignin roadmap. *Curr. Opin. Plant Biol.*, 5: 224-229.

Iseli, C., C.V. Jongeneel and P. Bucher. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 7: 138-148.

Jiang, N.H., G.H. Zhang, J.J. Zhang, L.P. Shu, W. Zhang, G.Q. Long, T. Liu, Z.G. Meng, J.W. Chen and S.C. Yang. 2014. Analysis of the transcriptome of erigeron breviscapus uncovers putative scutellarin and chlorogenic acids biosynthetic genes and genetic markers. *PLoS One*, 9(6): e100357.

Jiang, S.L., Y.Y. Zhang and D.F. Chen. 2005. Effects of heteroclitin D, schisanhenol and (+)-anwulignan on platelet aggregation. *Fudan Univ. J. Med. Sci.*, 32(4): 467-470, 478.

Kanehisa, M., S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucl. Acids Res.*, 34(Database issue): D354-357.

Kim, H.J., E. Ono, K. Morimoto, T.Yamagaki, A. Okazawa, A. Kobayashi and H. Satake. 2009. Metabolic engineering of lignan biosynthesis in *Forsythia* cell culture. *Plant Cell Physiol.*, 50: 2200-2209.

Kumpatla, S.P. and S. Mukhopadhyay. 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome.*, 48: 985-998.

Liang, X., X. Chen, Y. Hong, H. Liu, G. Zhou, S. Li and B. Guo. 2009. Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and Arachis wild species. *BMC Plant Biol*., 9: 35.

Liu, H., J. Zhang, X. Li, Y. Qi, Y. Peng, B. Zhang and P. Xiao. 2012. Chemical analysis of twelve lignans in the fruit of *Schisandra sphenanthera* by HPLC–PAD-MS. *Phytomedicine*, 19(13): 1234-1241.

Lu, Y. and D.F. Chen. 2009. Analysis of *Schisandra chinensis* and *Schisandra sphenanthera*. *J. Chromatogr A.*, 1216(11): 1980-1990.

Marques, J.V., K.W. Kim, C. Lee, M.A. Costa, G.D. May, J.A. Crow, L.B. Davin and N.G. Lewis. 2013. Next generation sequencing in predicting gene function in podophyllotoxin biosynthesis. *J. Biol. Chem.*, 288(1): 466-479.

Molog, G.A., U. Empt, S. Kuhlmann, U.W. Van, N. Pras, A.W. Alfermann and M. Petersen. 2001. Deoxypodophyllotoxin 6-hydroxylase, a cytochrome P450 monooxygenase from cell cultures of *Linum flavum* involved in the biosynthesis of cytotoxic lignans. *Planta*, 214(2): 288-294.

Ono, E., M. Nakai, Y. Fukui, N. Tomimori, M. Fukuchi-Mizutani, M. Saito, H. Satake, T. Tanaka, M. Katsuta, T. Umezawa and Y Tanaka. 2006. Formation of two methylenedioxy bridges by a *Sesamum* CYP81Q protein yielding a furofuranlignan, (+)-sesamin. *Proc. Natl. Acad. Sci. USA.*, 103(26): 10116-10121.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai and J. Quackenbush. 2003. TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics,* 19: 651-652.

Porebski, S., L.G. Bailey and B.R. Baum. 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.*, 15(1): 8-15.

Ragamustari, S.K., T. Nakatsubo, T. Hattori, E. Ono, Y. Kitamura, S. Suzuki, M. Yamamura and T. Umezawa. 2013. A novel *O*-methyltransferase involved in the first methylation step of yatein biosynthesis in *Anthriscus sylvestris*. *Plant Biotechnology,* 30: 375-384.

Renouard, S., C. Corbin, T. Lopez, J. Montguillon, L. Gutierrez, F. Lamblin, E. Lainé and C. Hanol. 2012. Abscisic acid regulates pinoresinol-lariciresinol reductase gene expression and secoisolariciresinol accumulation in developing flax (*Linum usitatissimum* L.) seeds. *Planta,* 235(1): 85-98.

Rozen, S. and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., 132(3): 365-386.

Saha, M.C., M.A.R. Mian, I. Eujayl, J.C. Zwonitzer, L.J. Wang and G.D. May. 2004. Tall fescue EST-SSR markers with transferability across several grass species. *Theor. Appl. Genet.,* 109(4): 783-791.

Suzuki, S. and T. Umezawa. 2007. Biosynthesis of lignans and norlignans. *J. Wood Sci.,* 53(4): 273-284.

Tatusov, R.L., E.V. Koonin and D.J. Lipman. 1997. A genomic perspective on protein families. *Science,* 278(5338): 631-637.

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

Umezawa, T., S.K. Ragamustari, T. Nakatsubo, S. Wada, L.G. Li, M. Yamamura, N. Sakakibara, T. Hattori, S. Suzuki and V.L. Chiang. 2013. A novel lignin *O*-methyltransferase catalyzing the regioselective methylation of matairesinol in *Carthamus tinctorius*. *Plant Biotechnol.,* 30: 97-109.

Umezawa, T. 2003. Diversity in lignan biosynthesis. *Phytochem. Rev.,* 2(3): 371-390.

Varshney, R.K., A. Graner and M.E. Sorrells. 2005. Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.,* 23(1): 48-55.

Wang, Z.S., H.X. Chen, W.J. Zhang, G.S. Lan and L.K. Zhang. 2011. Comparative studies on the chemical composition and antioxidant activities of *Schisandra chinensis* and *Schisandra sphenanthera* fruits. *Journal of Medicinal Plants Research,* 5(7): 1207-1216.

Wang, Z.Y., J. Li, Z.X. Luo, L.F. Huang, X.L. Chen, B.P. Fang, Y.J. Li, J.Y. Chen and X.J. Zhang. 2011. Characterization and development of EST-derived SSR markers in cultivated sweet potato (*Ipomoea batatas*). *BMC Plant Biol.,* 11: 139.

Wei, W.L., X.Q. Qi, L.H. Wang, Y.X. Zhang, W. Hua, D.H. Li, H.X. Lv and X.R. Zhang. 2011. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics.*, 12: 451.

Whiting, D.A. 1985. Lignans and neolignans. *Nat. Prod. Rep.,* 2:191-211.

Xia, Y.G., B.Y. Yang, J. Liang, Q. Yang, D. Wang and H.X. Kuang. 2014. Quantitative analysis and fingerprint profiles for quality control of Fructus Schisandrae by gas chromatography: mass spectrometry. *Scientific World Journal*, 806759.

Xia, Z.Q., M.A. Costa, J. Proctor, L.B. Davin and N.G. Lewis. 2000. Dirigent-mediated podophyllotoxin biosynthesis in *Linum flavum* and *Podophyllum peltatum*. *Phytochemistry*, 55(6): 537-549.

Ye, J., L, Fang, H.K. Zheng, Y. Zhang, J. Chen, Z.J. Zhang, J. Wang, S.T. Li, R.Q. Li, L. Bolund and J. Wang. 2006. WEGO: a web tool for plotting GO annotations, *Nucleic Acids Res*., 34(Web Server issue): W293-W297.

Yoo, H.H., M. Lee, M.W. Lee, S.Y. Lim, J. Shin and D.H. Kim. 2007. Effects of *Schisandra lignans* on P-glycoprotein-mediated drug efflux in human intestinal Caco-2. *Planta Med.,* 73(5): 444-450.

Yu, J.K., T.M. Dake, S. Singh, D. Benscher, W. Li, B. Gill and M.E. Sorrells. 2004. Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. *Genome*, 47(5): 805-818.

Zheng ,X.F., C. Pan, Y, Diao, Y.N You, C.Z. Yang and Z.L. Hu. 2013. Development of microsatellite markers by transcriptome sequencing in two species of Amorphophallus (Araceae). *BMC Genomic,* 14:490.

(Received for publication 20 March 2017)