Contents lists available at ScienceDirect

Electronic Journal of Biotechnology

Research article

Transcriptome analysis of *Cinnamomum longepaniculatum* by high-throughput sequencing

Kuan Yan [a,b], Qin Wei [b], Ruizhang Feng [b], Wanhai Zhou [b], Fang Chen [a,*]

[a] *College of Life Science, Sichuan University, Chengdu 610044, China*
[b] *College of Life Science and Food Engineering, Yibin University, Yibin 644000, China*

ARTICLE INFO

Article history:
Received 8 January 2017
Accepted 23 May 2017
Available online 30 May 2017

Keywords:
Breeding
Crop
Gene annotation
Gene Ontology
Pathway
Simple sequence repeats (SSR)
Transcription
Transcriptomics
Unigenes
Volatile terpenoids

ABSTRACT

Background: Cinnamomum longepaniculatum is an important commercial crop and the main source of volatile terpenoids. The biosynthesis of key bioactive metabolites of *C. longepaniculatum* is not well understood because of the lack of available genomic and transcriptomic information. To address this issue, we performed transcriptome sequencing of *C. longepaniculatum* leaves to identify factors involved in terpenoid metabolite biosynthesis.

Results: Transcriptome sequencing of *C. longepaniculatum* leaves generated over 56 million raw reads. The transcriptome was assembled using the Trinity software and yielded 82,061 unigenes with an average length of 879.43 bp and N50 value of 1387 bp. Furthermore, Benchmarking Universal Single-Copy Orthologs analysis indicated that our assembly is 91% complete. The unigenes were used to query the nonredundant database depending on sequence similarity; 42,809 unigenes were homologous to known genes in different species, with an annotation rate of 42.87%. The transcript abundance and Gene Ontology analyses revealed that numerous unigenes were associated with metabolism, while others were annotated in functional categories including transcription, signal transduction, and secondary metabolism. The Kyoto Encyclopedia of Genes and Genomes pathway analysis showed that 19,260 unigenes were involved in 385 metabolic pathways, with 233 unigenes found to be involved in terpenoid metabolism. Moreover, 23,463 simple sequence repeats were identified using the microsatellite identification tool.

Conclusion: This is the first detailed transcriptome analysis of *C. longepaniculatum*. The findings provide insights into the molecular basis of terpenoid biosynthesis and a reference for future studies on the genetics and breeding of *C. longepaniculatum*.

1. Introduction

Cinnamomum longepaniculatum (Gamble) N. Chao is an important industrial crop in China. Essential oils can be extracted from its roots, stem, leaves, and seeds; the main constituents of leaf essential oil are terpenoids (>85%) [1], including 1,8-cineole, α-terpilenol and γ-terpinene [2,3,4,5]. These essential oils have antioxidant, anti-inflammatory, and antimicrobial properties [6]; however, their yield is low because of the variable content of secondary metabolites and the effects of other factors such as environmental conditions. Substantial efforts have been made to improve the yield of *C. longepaniculatum* essential oil using new technologies.

Next-generation sequencing is a powerful tool for de novo transcriptome assembly and annotation owing to its low operational cost and available computational resources; it has revolutionized the fields of phytochemistry and natural medicine. It is especially useful for studies of nonmodel plants for which no genomic information is available. For instance, gene expression profiles of important industrial or medicinal plants at a specific developmental stage or from tissues can be rapidly obtained by high-throughput transcriptome sequencing [7,8,9,10]. This can provide insight into metabolic processes and thus a basis for developing strategies to increase the biosynthesis of desired metabolites [11,12,13,14,15,16,17,18].

[*] Corresponding author.
E-mail address: fangchenscu@163.com (F. Chen).
Peer review under responsibility of Pontificia Universidad Católica de Valparaíso.

http://dx.doi.org/10.1016/j.ejbt.2017.05.006

Table 1
Sequencing statistics.

Sample	Total raw reads	Total clean reads	Total clean nucleotides (nt)	Error (%)[a]	Q20 (%)[c]	GC (%)[b]
Leaf	58,564,542	56,857,306	8,331,942,311	0.0109	98.02	48.88

[a] Base error rate.
[b] Sum of G and C bases as a percentage of the total number of bases.
[c] Percentage of bases with Phred value >20 as a percentage of the total number of bases.

Table 2
Assembly quality.

Type	Unigene	Transcripts
Total sequence no.	82,061	105,028
GC (%)[a]	42.87	42.71
Largest (bp)	61,022	61,022
Smallest (bp)	201	201
Average length (bp)	897.43	999.38
N50[b]	1387	1523

[a] Sum of G and C bases as a percentage of the total number of bases.
[b] Assembled transcripts were sorted from the largest to the smallest according to length; N50 is the transcript length when the length of the transcript is half of the total length.

Transcriptome sequencing and analysis has been performed for various medicinal plants [19,20,21,22]; however, fragrant plants such as *C. longepaniculatum* have not been examined in detail, which has hindered their genetic improvement and industrial development. To this end, the present study used high-throughput sequencing to analyze the transcriptome of *C. longepaniculatum* leaves and identify genes involved in terpenoid biosynthesis.

2. Material and methods

2.1. Plant materials, RNA isolation, and cDNA library construction

C. longepaniculatum (Gamble) N. Chao leaves were collected from the Red Rock mountain in Yibin, Sichuan Province, China (located at 27°50′ N; 105°20′ E). Those with similar vigor were quickly cleaned with sterile water, dried, and stored in a 50-ml centrifuge tube. The tube was then flash frozen in liquid N_2 at -80°C. Total RNA was extracted using TRIzol reagent, and the concentration and purity were evaluated using a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). RNA integrity was evaluated by agarose gel electrophoresis, and the RNA integrity number was determined using a Bioanalyzer2100 (Agilent Technologies, Santa Clara, CA, USA). mRNA was enriched from total RNA using oligo (dT) primers and then fragmented and used as a template for the synthesis of first-strand cDNA using random primers. A cDNA library was constructed using the Sure Select Strand-specific RNA library kit (Agilent Technologies).

2.2. Illumina sequencing and raw data processing and assembly

The library was sequenced using Illumina Hiseq 4000; mass analysis of total raw reads was then performed. Bases at the sequenced end (3′) with low masses (<20) and low-quality reads with mass < 10 were removed. Reads with "N" bases > 10% were removed, and adapter sequences or sequences with an average length of <20 bp were excluded. The transcriptome was assembled de novo using the Trinity software [23] from paired or unpaired clean reads (forward and reverse). Briefly, reads with overlapping segments were assembled into contiguous sequences. The contigs were then assembled into unigenes by paired-end assembly and gap filling [24].

2.3. Functional annotation and single sequence repeat locus identification

Unigene sequences were compared (E value < 1e-5) against Non-redundant (NR), Search Tool for the Retrieval of Interacting Genes (String), Swissprot, Protein Families (Pfam), Kyoto Encyclopedia

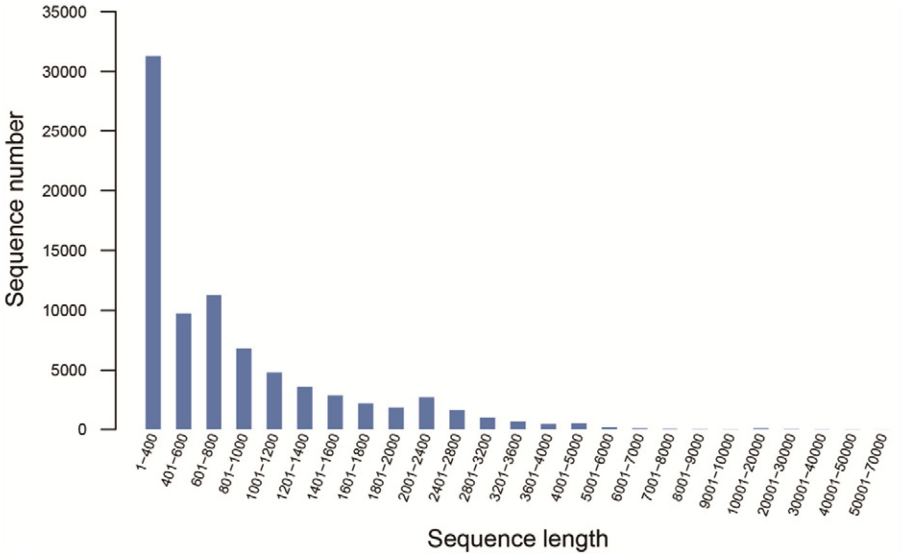

Fig. 1. Length distribution of assembled unigenes. The abscissa and ordinate show the length range of assembled unigenes and the corresponding number of unigenes, respectively.

Table 3
Annotation of unigenes against five databases.

Database	Annotated unigene	Percentage
Pfam	21,499	49.87%
String	16,104	37.36%
KEGG	19,260	44.68%
Swissprot	24,642	57.17%
NR	42,809	99.31%

of Genes and Genomes (KEGG), and Clusters of Orthologous Groups of Proteins (COG) databases using BLASTx. Each unigene was annotated according to information from NR against the Gene Ontology (GO) database using Blast2GO, and GO functional classification analysis was performed. Poly-A at the 3′ end and Poly-T at the 5′ end of unigenes were first trimmed by est_trimmer.pl (http://pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl, options: "-tr5 = T,5,50 -tr3 = A,5,50"). Then MISA was employed to identify single sequence repeats (SSRs) in the assembled transcriptome of *C. longepaniculatum*; the repeat thresholds for mono-, di-, tri-, tetra-, penta-, and hexanucleotide motifs were a minimum of 10, 6, 5, 5, 5, and 5, respectively.

3. Results

3.1. Illumina sequencing and de novo assembly

A total of 58,564,542 raw read fragments were obtained by high-throughput sequencing. Over 56 million clean reads were obtained after optimization, with Q20 > 98%, error rate = 0.0109%, and GC content of 48.88% (Table 1). The sequencing quality was good, making the subsequent analyses reliable. All raw sequence data have been deposited in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/) under the accession number SRR5388904.

De novo assembly was performed with the Trinity software, yielding 105,028 contigs with an average length of 999.38 bp. These were paired-end assembled into 82,061 unigenes with an average length of 879.43 bp and N50 of 1387 bp (Table 2). The length distribution of assembled unigenes is shown in Fig. 1. Unigenes with a length of <400 bp comprised the largest fraction (approximately 38.15%), followed by those ranging in size from 601 to 800 bp (13.73%).

For quantitative assessment of the assembly and completeness of annotations, we used Benchmarking Universal Single-Copy Orthologs (BUSCO) [25], which is based on evolutionarily informed expectations of gene content, with default settings. Compared to the 1440 single-copy orthologs for the embryophyta lineage, our assembly was 91% complete (926 complete single-copy and 385 complete duplicated BUSCO), while 3.3% of contigs were fragmented (48 BUSCOs) and 5.6% were missing (81 BUSCOs). These results indicated that transcriptome

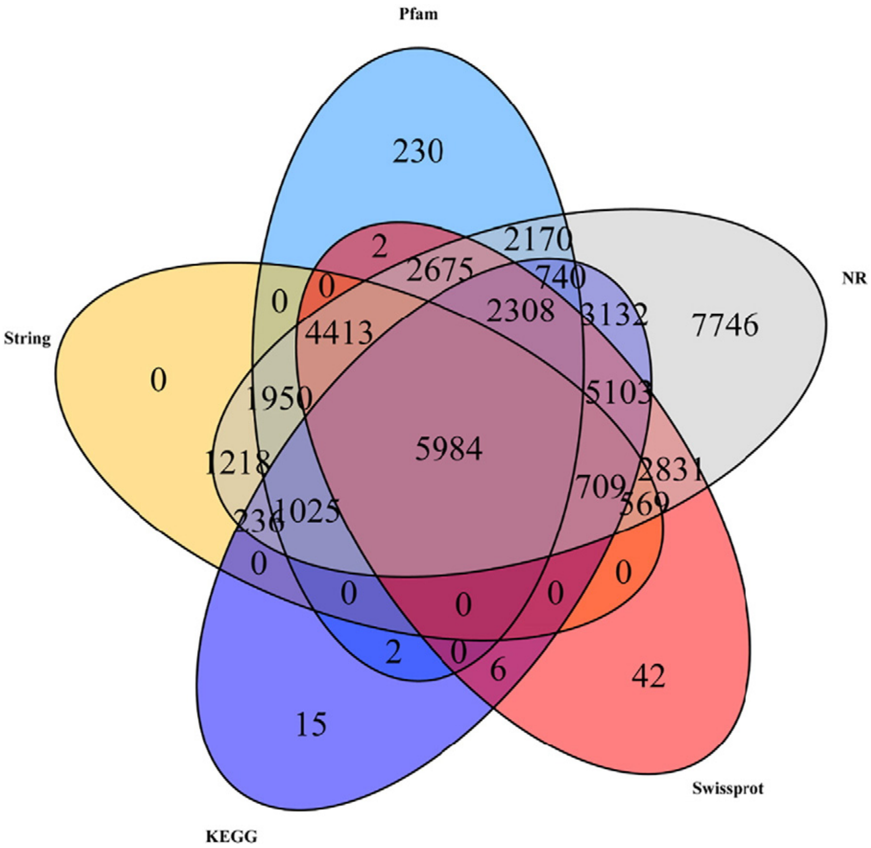

Fig. 2. Statistical information on annotated unigenes in five databases. Circles of different colors represent the number of unigenes annotated against a database and the areas of intersection represent the number of genes simultaneously annotated in the corresponding libraries.

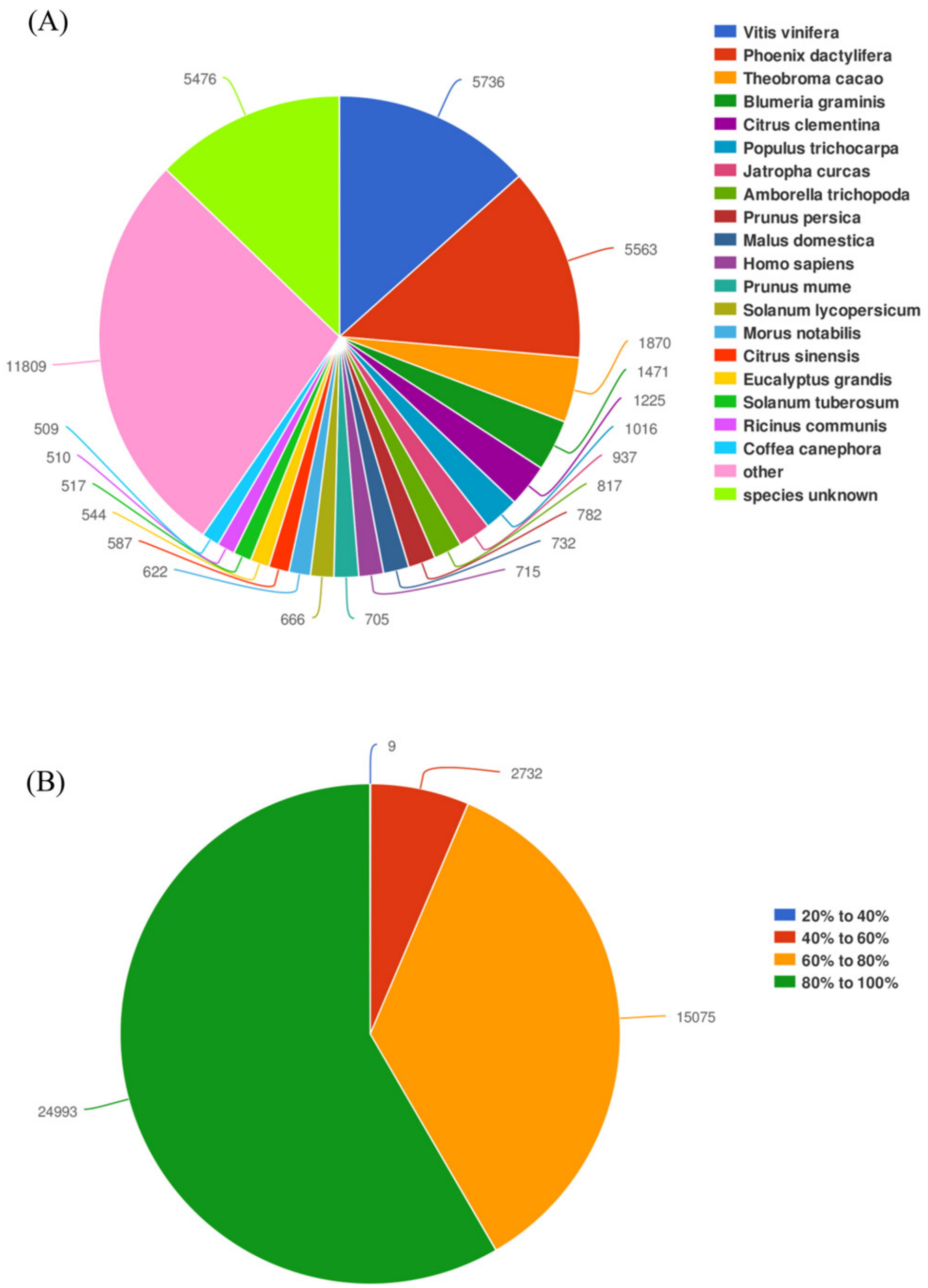

Fig. 3. Statistics of NR annotation results. (A) Sequence of a species from the sample relative to near-source species (each sector represents a species). A larger fan-shaped area reflects a greater number of sequences aligned to the species. (B) Similarity refers to the ratio of similar amino acids to total amino acids (E value < 1e-5). Each sector represents a similarity interval; a larger fan-shaped area reflects a greater number of genes with a similarity degree within this interval.

assembly was useful for further transcriptomic analyses of *C. longepaniculatum*.

3.2. Unigene function annotation and classification

3.2.1. Unigene sequence similarity analysis

The assembled unigenes were annotated on the basis of sequence similarity against NR, String, Swissprot, Pfam, and KEGG databases using BLASTx. A total of 43,106 unigenes were annotated against the NR database, with 42,809 (99.31%) showing significant homology; 16,104 unigenes (19.62%) were annotated against the String database (Table 3). Annotation information in other databases was compared against the NR database to determine species similarity and functions of homologous sequences. A total of 5984 unigenes were annotated in all five databases (Fig. 2).

The species distribution analysis based on BLASTx results showed that approximately 70% of unigenes with a BLAST hit shared high similarity with sequences from *Vitis vinitera* (13.40%), *Phoenix dactylifera* (12.99%), and *Theobroma cacao* (4.37%) (Fig. 3). Moreover, 24,993 unigenes showed 80%–100% homology and 15,075 unigenes showed 60%–80% homology. Some unigenes could not be annotated against the above-mentioned databases because of the absence of a

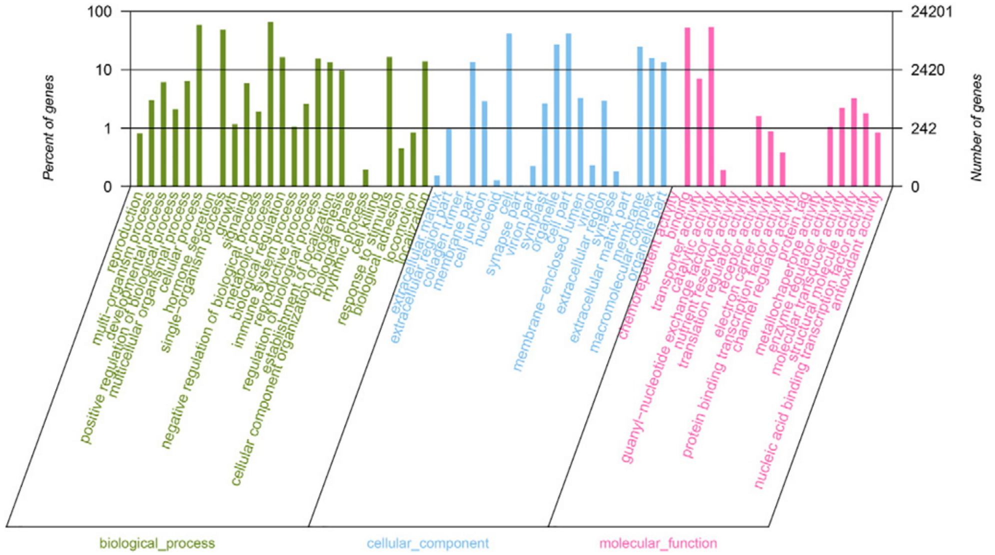

Fig. 4. Statistics of GO annotation results. The abscissa and left and right ordinates show the secondary GO classification terms, the percentage of total isogenes contained in the secondary classification, and the number of isogenes of the secondary classification in the alignment, respectively. Green, blue, and red represent biological process, cellular component, and molecular function, respectively.

C. longepaniculatum genome and expressed sequence tag and protein sequence information.

We classified genes on the basis of GO annotation as biological process, molecular function, and cellular component. The unigenes were classified into the following functional groups: metabolic process (n = 15,929), cellular process (n = 14,250), catalytic activity (n = 13,124), binding (n = 12,735), and cell part (n = 10,090). There were few unigenes in the protein tag, translation regulator activity, and chemorepellent activity groups (Fig. 4). These results reveal that many unigenes participated in metabolic processes involving binding and catalytic activities.

3.2.2. COG annotation and classification

Reference unigenes were aligned with sequences in the COG database. A total of 16,104 unigenes were clustered into 25 categories depending on sequence homology. The top category was "General function prediction only" (n = 1116), followed by "Signal transduction mechanisms" (n = 1028) and "Post-translational modification, protein turnover, chaperones" (n = 928). In contrast, there was only one unigene in the "Nuclear structure" category and none in the "Extracellular structures" category, and 427 unigenes were in the "Function unknown" category (Fig. 5).

3.2.3. KEGG pathway annotation

The assembled unigenes were assigned to biochemical pathways described in KEGG; a pathway-based analysis provided information on

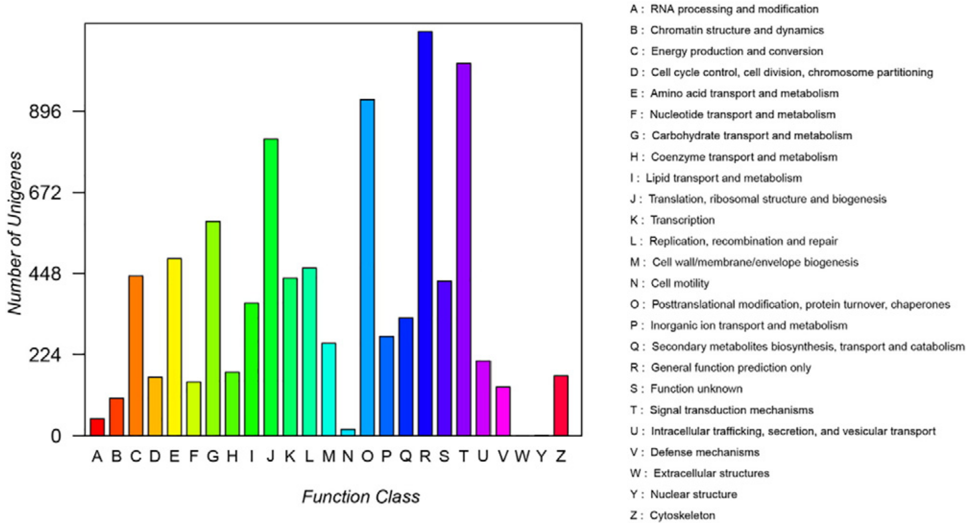

Fig. 5. Statistics of COG annotation results. Columns of each color represent the functional classification of a COG (in uppercase letters A to Z; see right side for the specific meaning of the label); the height of the column indicates the number of unigenes with that function.

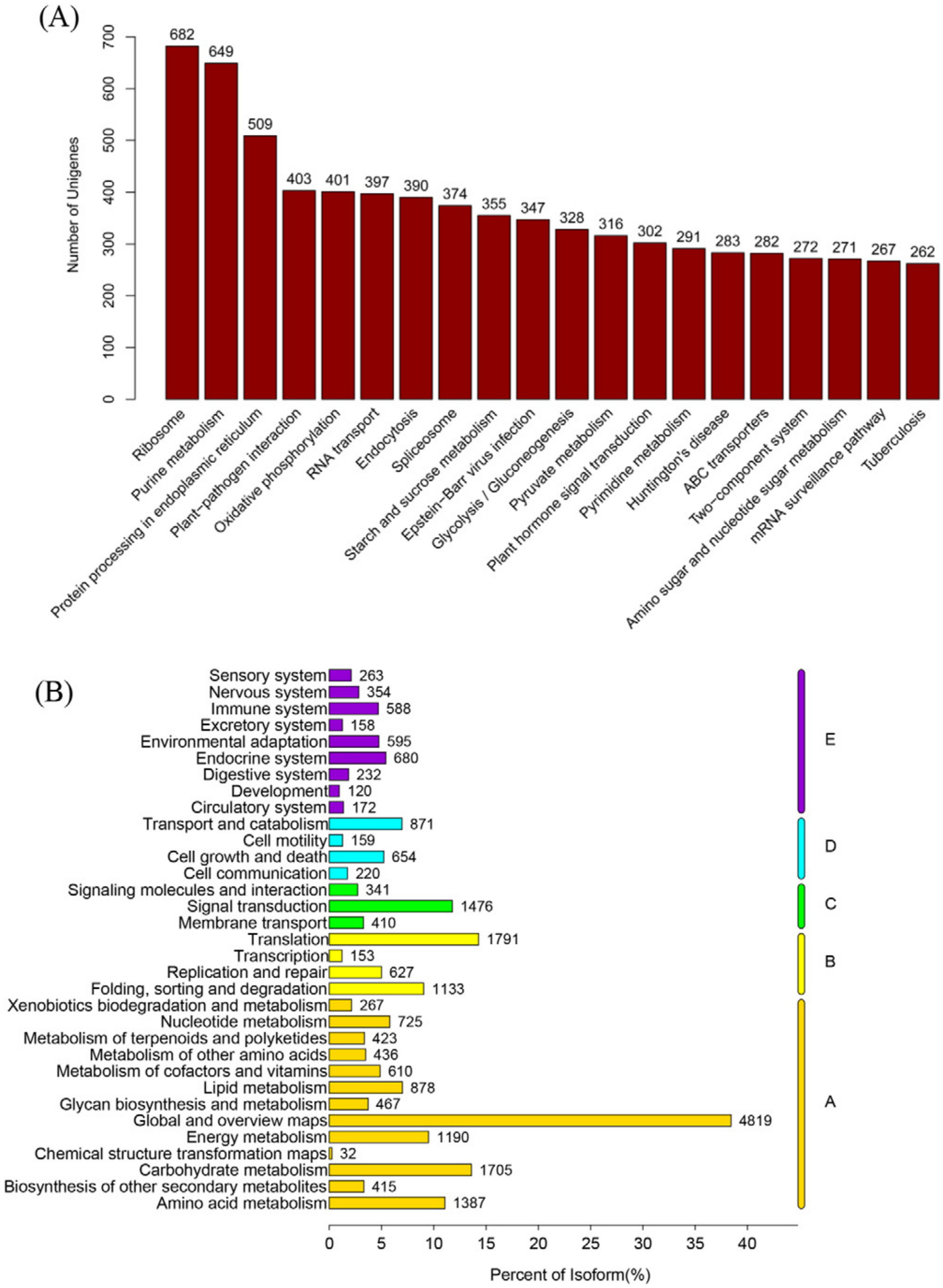

Fig. 6. Statistics of KEGG pathway annotation results. (A) From left to right in accordance with the number of unigenes included in order from high to low, taller columns reflect more active biological pathways in the measured sample. (B) The ordinate and abscissa show the name of the KEGG metabolic pathway and the ratio of the number of unigenes in the pathway to the total number of annotated unigenes. Unigenes are divided into five branches according to the KEGG metabolic pathway: A, Metabolism; B, Genetic information processing; C, Environmental information processing; D, Cellular processes; and E, Organismal systems.

the biological functions of the identified genes [26]. The KEGG analysis showed that 19,260 unigenes were associated with 385 metabolic pathways. The top annotated pathway was "Ribosome" (n = 682), followed by "Purine metabolism" (n = 649) and "Protein processing in endoplasmic reticulum" (n = 509) (Fig. 6A). The KEGG metabolic pathways were divided into five categories: (A) Metabolism; (B) Genetic information processing; (C) Environmental information processing; (D) Cellular processes; and (E) Organismal systems (Fig. 6B). Of these, 1750 unigenes were assigned to "Carbohydrate metabolism" (14%), 1387 to "Amino acid metabolism" (11%), and 1190 to "Energy metabolism" (9%). Notably, 423 unigenes were assigned to "Metabolism of terpenoids and polyketides," which may be important for *C. longepaniculatum* terpenoid synthesis and were therefore isolated for future studies.

3.2.4. Identification of C. longepaniculatum SSRs based on the de novo assembled transcriptome

SSRs are ubiquitously distributed throughout eukaryote genomes and have high variability and a large number of repeats [27]. We detected 23,463 SSRs, of which mononucleotides comprised the largest fraction (12,124, 51.67%) followed by dinucleotides (6918, 29.48%). A total of 269 SSRs were detected by tetranucleotides,

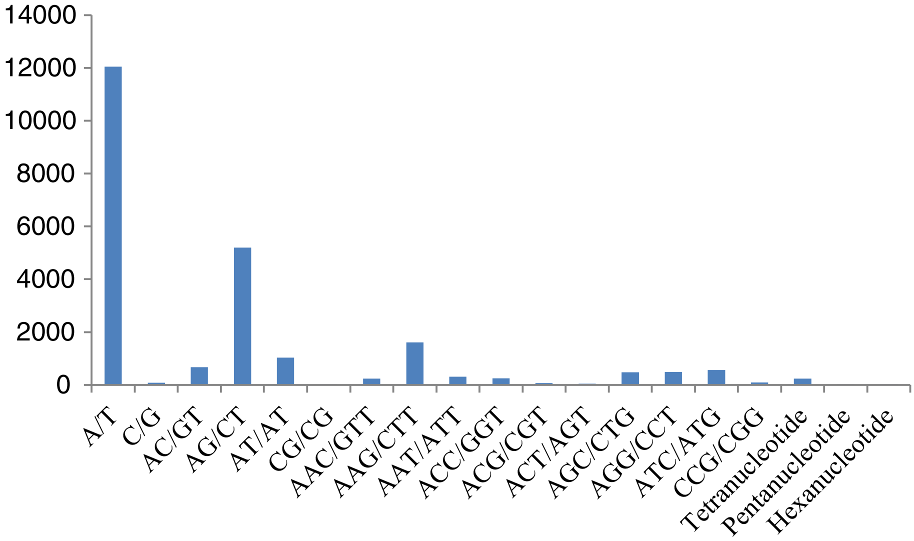

Fig. 7. Statistical analysis of SSR type. The abscissa represents the SSR type and the ordinate represents the number of SSRs.

pentanucleotides, and hexaonucleotides. A/T (12,046) represented the largest fraction of mononucleotides, whereas AG/CT (5202) and AAG/CTT (1614) were the most abundant dinucleotides and trinucleotides, respectively. Thus, A/T, AG/CT, and AAG/CTT were the major SSR markers in *C. longepaniculatum* (Fig. 7). SSRs can be used to analyze differences between the genomes of *C. longepaniculatum* and those of other plants within the same family, which can provide insight into genetic diversity and provide a set of genetic markers.

3.3. Unigenes related to terpenoid biosynthesis

The two major terpenoid biosynthesis pathways are the cytoplasmic mevalonate and the 2-C-methyl-D-erythritol-4-phosphate (MEP) pathways. Monoterpenoids, the major volatile terpenoids in *C. longepaniculatum*, are mainly generated through the latter route [28,29]. The KEGG pathway analysis revealed 109 unigenes related to biosynthesis of the terpenoid skeleton. Of these, 17 were identified for 1-deoxy-D-xylulose 5-phosphate synthase (DXS) and 12 for acetoacetyl-CoA-thiolase (AACT) (Table 4). In addition, 14 unigenes were related to monoterpenoid synthesis, including six (-)-alpha-terpineol synthase, six (3S)-linalool synthase, and three (+)-neomenthol dehydrogenase (Table 5). These results provide reliable information for further studies on the expression of unigenes related to volatile terpenoid metabolism in *C. longepaniculatum*.

4. Discussion

Next-generation sequencing is more cost-effective and practical than whole-genome sequencing and is a powerful tool in plant research [30,31]. In the present study, we performed a transcriptome analysis of *C. longepaniculatum* by high-throughput sequencing. A total of 82,061 unigenes with an average length of 879.43 bp were generated by de novo transcriptome assembly, with an N50 value of 1387 bp; the completeness of the assembly was 91% for 1440 single-copy orthologs of the embryophyta lineage. Thus, the quantity and quality of the transcriptome sequence data obtained for *C. longepaniculatum* met the requirements for reliable analyses.

Owing to a lack of genomic and transcriptomic resources, the molecular basis underlying the biosynthesis of characteristic key bioactive metabolites remains unclear. We annotated 43,106 unigenes (52.53% of the total) against NR, String, Swissprot, KEGG, and Pfam databases; 5984 unigenes were annotated in all databases simultaneously, whereas some were not identified because of limited genomic sources and biodiversity. The GO analysis showed that many of the unigenes were related to cellular and metabolic processes and were mainly involved in binding and catalytic activities. The classification of unigenes in the COG database revealed orthologous genes that allowed the prediction of biological function. We found that the transcriptome of *C. longepaniculatum* harbored many unigenes associated with transcription, signal transduction, and secondary metabolism. A KEGG pathway analysis revealed that 19,260 unigenes were involved in 385 metabolic pathways related to carbohydrate and amino acid metabolism and energy production.

The terpenoids in *C. longepaniculatum* are mostly generated by the MEP synthesis pathway. Identifying the components of this pathway and expressing them in microbial hosts can enable industrial-scale production of terpenoids [32]. Hundreds of genes in the MEP and closely related metabolic pathways have recently been identified in *Camellia sinensis* [33], *Salvia miltiorrhiza* [34], *Euphorbia fischeriana* [35], and other plants. For the taxane-producing *Taxus cuspidata*, genes encoding the seven enzymes of the plastidial MEP pathway were identified in 454 datasets [36]; additionally, *Taxus mairei* genes encoding all seven enzymes in the plastidial MEP pathway were identified in Illumina datasets [17]. In this study, several unigenes involved in terpenoid synthesis were assigned to genes encoding the eight key enzymes in the MEP pathway, including DXS (17 unigenes), DXR, CMS, MCS, HDS, IDS, and GPPS. In addition, 14 unigenes were assigned to genes encoding three enzymes involved in monoterpenoid synthesis. Transcriptome sequencing of *C. longepaniculatum* revealed 23,463 SSRs, which can be used to estimate genetic resources and used for marker-assisted selection [37].

In conclusion, our results can serve as a resource for further studies on *C. longepaniculatum* physiology and provide an important reference for synthesizing and maximizing the yield of *C. longepaniculatum*

Table 4
Unigenes related to terpenoid skeleton biosynthesis.

EC number	Name	KO name	Unigene number
2.2.1.7	DXS	K01662	17
5.3.3.2	IDI	K01823	4
1.17.7.1	HDS/ispG	K03526	6
2.5.1.87	DHDDS	K11778	3
2.3.3.10	HMGS	K01641	5
1.2.1.83	CHLP	K10960	5
2.5.1.29	GGPPS	K13789	7
2.7.1.148	CMK/ispE	K00919	1
1.1.1.216	FLDH	K15891	1
2.5.1.10	FDPS	K00787	5
2.3.1.9	AACT	K00626	12
2.7.4.2	PMK	K00938	2
4.1.1.33	MVD	K01597	3
3.4.22.-	RCE1	K08658	1
2.7.1.36	MVK	K00869	3
1.1.1.267	DXR/ispC	K00099	3
3.1.1.-	PCME	K15889	3
2.7.7.60	ispD/CMS	K00991	1
1.1.1.34	HMGCR	K00021	4
2.5.1.58	FNTB	K05954	1
2.1.1.100	ICMT	K00587	1
2.5.1.84	SPS	K05356	1
4.6.1.12	ispF/MCS	K01770	1
2.5.1.58	FNTA	K05955	1
2.5.1.29	GPPS	K14066	1
1.17.1.2	ispH/IDS	K03527	5
3.4.24.84	STE24	K06013	4
2.7.1.216	FOLK	K15892	2
2.7.7.60	ispDF	K12506	1
2.5.1.90	ispB	K02523	3
1.8.3.6	FCLY	K05906	1
2.5.1.31	uppS	K00806	1

AACT, acetoacetyl-CoA transferase; CHLP, geranylgeranyl reductase; CMK/ispE, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; DHDDS, dehydrodolichyl diphosphate synthase; DXR/ispC, 1-deoxy-D-xylulose-5-phosphate synthase reductoisomerase; DXS, 1-deoxy-D-xylulose 5-phosphate synthase; FCLY, farnesylcysteine lyase; FDPS, farnesyl diphosphate synthase; FLDH, dependent farnesol dehydrogenase; FNTA, farnesyl protein transferase alpha subunit; FNTB, farnesyl protein transferase beta subunit; FOLK, 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase; GGPPS, geranylgeranyl pyrophosphate synthase; GPPS, geranyl diphosphate synthase; HDS/ispG, 1-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate synthase; HMGCR, hydroxymethylglutaryl-CoA reductase; HMGS, hydroxymethylglutaryl-CoA synthase; ICMT, isoprenylcysteine carboxyl methyltransferase; IDI, isopentenyl diphosphate isomerase; ispB, octaprenyl-disphosphate synthase; ispD/CMS, 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase; ispDF, 4-diphosphocytidyl-2C-methyl-D-erythritol synthase/2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; ispF/MCS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; ispH/IDS, 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase; KEGG, Kyoto Encyclopedia of Genes and Genomes; MVD, mevalonate diphosphate decarboxylase; MVK, mevalonate kinase; PCME, prenylcysteine alpha-carboxyl methylesterase; PMK, phosphomevalonate kinase; RCE1, related to ubiquitin 1-conjugating enzyme 1; SPS, solanesyl diphosphate synthase; STE24, STE24 endopeptidase; uppS, ubiquitous plant pathway S.

Table 5
Unigenes related to monoterpenoid biosynthesis.

EC number	Name	KO name	Unigene number
4.2.3.111	(-)-alpha-terpineol synthase	K18108	6
1.1.1.208	(+)-neomenthol dehydrogenase	K15095	3
4.2.3.25	(3S)-linalool synthase	K15086	5

essential oils, thereby facilitating further research in functional genomics.

Conflicts of interests

The authors declare that they have no competing interests.

Financial support

This work was co-supported by the Project of Sichuan Education Department (grant no. 14TD0031); Innovative Program of Sichuan Undergraduates (grant no. 201510641016); and Key Lab of Aromatic Plant Resources Exploitation and Utilization in Sichuan Higher Education (grant nos. 2015XLZ001 and 2015XLZ004).

References

[1] Hu WJ, Gao HD, Jang XM. Analysis on constituents and contents in leaf essential oil from three chemical types of *Cinnamum camphora*. J Cent South Univ Technol 2012; 32:186–94.

[2] Li L, Zheng WL, Zhong QY. Antibacterial activity of leaf essential oil and its constituents from *Cinnamomum longepaniculatum*. Int J Clin Exp Med 2014;7:1721–7.

[3] Xu S, Zhong QY, Wei Q. Anti-hepatoma effect of safrole from *Cinnamomum longepaniculatum* leaf essential oil in vitro. Int J Clin Exp Pathol 2014;7:2265–72.

[4] Li N, Zu YG, Wang W. Antibacterial and antioxidant of celery seed essential oil. Chin Condiment 2012;37:28–30.

[5] Deng S, Liang ZY, Zhang LL. Three kinds of antibacterial activity of essential oil. J Guizhou Norm Univ 2012;28:30–2.

[6] Wei Q, Tan YY, Li Q. Effects of fungal endophytes on cell suspension culture of *Cinnamomum longepaniculatum*. Guihaia 2016;36:923–9.

[7] Alagna FD, Agosino N, Torchia L, et al. Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics 2009; 10:399. http://dx.doi.org/10.1186/1471-2164-10-399.

[8] Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, Powell WA, et al. Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. BMC Plant Biol 2009;9:51. http://dx.doi.org/10.1186/1471-2229-9-51.

[9] Dassanayake M, Haas JS, Bohnert HJ, Bohnert HJ, Cheeseman JM. Shedding light on an extremophile lifestyle through transcriptomics. New Phytol 2009;183:764–75. http://dx.doi.org/10.1111/j.1469-8137.2009.02913.x.

[10] Maher CA, Kumar-Sinha C, Cao XH, Kalyana-Sundaram S, Han B, Jing XJ, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458:97–101. http://dx.doi.org/10.1038/nature07638.

[11] Wu Q, Sun CH, Chen SL. Application of transcriptomics in the studies of medicinal plants. World Sci Technol 2010;12:457–65.

[12] Franssen SU, Shrestha RP, Brautigam A, Bornberg-Bauer E, Weber A. Comprehensive transcriptome analysis of the highly complex *Pisumsativum* genome using next generation sequencing. BMC Genomics 2011;12:227. http://dx.doi.org/10.1186/1471-2164-12-227.

[13] Der JP, Barker MS, Wickett NJ, de Pamphilis CW, Wolf PG. *De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. BMC Genomics 2011;12:99. http://dx.doi.org/10.1186/1471-2164-12-99.

[14] Li RQ, Zhu HM, Ruan J, Qian W, Fang XD, Shi ZB, et al. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res 2010;20: 265–72. http://dx.doi.org/10.1101/gr.097261.109.

[15] Zhang GJ, Guo GW, Hu XD, Zhang Y, Li QY, Li RQ, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res 2010;20:646–54. http://dx.doi.org/10.1101/gr.100677.109.

[16] Lu TT, Lu GJ, Fan DL, Zhu CR, Li W, Zhao Q, et al. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res 2010;20: 1238–49. http://dx.doi.org/10.1101/gr.106120.110.

[17] Hao DC, Ge GB, Xiao PG, Zhang YY, Yang L. The first insight into the tissue specific *Taxus* transcriptome via Illumina second generation sequencing. PLoS One 2011;6: e21220. http://dx.doi.org/10.1371/journal.pone.0021220.

[18] Conesa A, Gotz S, Garcia JM, Terol J, Talon M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2006;21:3674–6. http://dx.doi.org/10.1093/bioinformatics/bti610.

[19] Chen SL, Luo HM, Li Y, Sun Y, Wu Q, Niu Y, et al. 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Rep 2011; 30:1593. http://dx.doi.org/10.1007/s00299-011-1070-6.

[20] Luo HM, Sun C, Song JY, Wu Q, Li Y, Song JY, et al. Analysis of the transcriptome of *Panaxnoto ginseng* root uncovers putative triterpenoid saponin-biosynthetic genes and genetic markers. BMC Genomics 2011;12(Suppl. 5):S5. http://dx.doi.org/10.1186/1471-2164-12-S5-S5.

[21] Sun C, Li Y, Wu Q, Luo HM, Sun YZ, Lui E, et al. *De novo* sequencing and analysis of the American ginseng root transcriptome using a GSFLX *Titanium platform* to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 2010;11:262. http://dx.doi.org/10.1186/1471-2164-11-262.

[22] Li Y, Luo HM, Sun C, Song JY, Sun YZ, Wu Q, et al. EST analysis reveals putative genes involved in glycyrrhizin biosyntheses. BMC Genomics 2010;11:268. http://dx.doi.org/10.1186/1471-2164-11-268.

[23] Glowacka K. A review of the genetic study of the energy crop *Miscanthus*. Biomass Bioenergy 2011;35:2445–54. http://dx.doi.org/10.1016/j.biombioe.2011.01.041.

[24] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;29:644–52. http://dx.doi.org/10.1038/nbt.1883.

[25] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;6:1–2.

[26] Camacho C, Coulouris G, Avagyan V. BLAST plus: Architecture and applications. BMC Bioinformatics 2009;10:421. http://dx.doi.org/10.1186/1471-2105-10-421.

[27] Faircloth BC. Msatcommander: Detection of microsatellite repeat arrays and automated, locus-specific primer design. Mol Ecol Resour 2008;8:92–4. http://dx.doi.org/10.1111/j.1471-8286.2007.01884.x.

[28] Hampel D, Mosandl A, Wust M. Biosynthesis of mono and sesquiterpenoids in carrot roots and leaves (*Daucuscarota* L.): Metabolic cross talk of cytosolic mevalonate and plastidial methylerythritol phosphate pathways. Phytochemistry 2005;66:305–11. http://dx.doi.org/10.1016/j.phytochem.2004.12.010.

[29] Wang LJ, Fang X, Yang CQ. Biosynthesis and regulation of secondary terpenoid metabolism in plants. Sci Sin Vitae 2013;43:1030–46.

[30] Shendure J. The beginning of the end for microarrays? Nat Methods 2008;5:585–7.

[31] Wilhelm BT, Landry JR. RNA-Seq quantitative measurement of expression through massively parallel RNA sequencing. Methods 2009;48:249–57. http://dx.doi.org/10.1016/j.ymeth.2009.03.016.

[32] Ajikumar PK, Xiao WH, Tyo KE. Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. Science 2010;330:70–4. http://dx.doi.org/10.1126/science.1191652.

[33] Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, et al. Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics 2011;12:131. http://dx.doi.org/10.1186/1471-2164-12-131.

[34] Hua WP, Zhang Y, Song J, Zhao LJ, Wang ZZ. *De novo* transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. Genomics 2011;98:272–9. http://dx.doi.org/10.1016/j.ygeno.2011.03.012.

[35] Barrero RA, Chapman B, Yang Y, Moolhuijzen P, Gagnere GK, Zhang N, et al. *De novo* assembly of *Euphorbia fischeriana* root transcriptome identifies prostratin pathway related genes. BMC Genomics 2011;12:600. http://dx.doi.org/10.1186/1471-2164-12-600.

[36] Wu Q, Sun C, Luo H, Li Y, Niu YY, Sun YZ, et al. Transcriptome analysis of *Taxus cuspidata* needles based on 454 pyrosequencing. Planta Med 2011;77:394–400. http://dx.doi.org/10.1055/s-0030-1250331.

[37] Yang HB, Liu WY, Kang WH, Jahn M, Kang BC. Development of SNP markers linked to the L locus in *Capsicum* spp. by a comparative genetic analysis. Mol Breed 2009;24:433–46. http://dx.doi.org/10.1007/s11032-009-9304-9.