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

**A B S T R A C T**

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*.

© 2017 Pontificia Universidad Católica de Valparaíso. Production and hosting by Elsevier B.V. All rights reserved. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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-terpinol 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].
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 N2 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 of Genes and Genomes (KEGG), and the transcriptome of *C. longepaniculatum* leaves.

Table 1

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total raw reads</th>
<th>Total clean reads</th>
<th>Total clean nucleotides (nt)</th>
<th>Error (%)</th>
<th>Q20 (%)</th>
<th>GC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf</td>
<td>58,564,542</td>
<td>56,857,306</td>
<td>8,331,942,311</td>
<td>0.0109</td>
<td>98.02</td>
<td>48.88</td>
</tr>
</tbody>
</table>

* 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

<table>
<thead>
<tr>
<th>Type</th>
<th>Unigene</th>
<th>Transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sequence no.</td>
<td>82,061</td>
<td>105,028</td>
</tr>
<tr>
<td>GC (%)</td>
<td>42.87</td>
<td>42.71</td>
</tr>
<tr>
<td>Largest (bp)</td>
<td>61,022</td>
<td>61,022</td>
</tr>
<tr>
<td>Smallest (bp)</td>
<td>201</td>
<td>201</td>
</tr>
<tr>
<td>Average length (bp)</td>
<td>897.43</td>
<td>999.38</td>
</tr>
<tr>
<td>N50</td>
<td>1387</td>
<td>1523</td>
</tr>
</tbody>
</table>

* 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.

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.
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.

![Statistical information on annotated unigenes in five databases.](image)

### Table 3

<table>
<thead>
<tr>
<th>Database</th>
<th>Annotated unigene</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfam</td>
<td>21,499</td>
<td>49.87%</td>
</tr>
<tr>
<td>String</td>
<td>16,104</td>
<td>37.36%</td>
</tr>
<tr>
<td>KEGG</td>
<td>19,260</td>
<td>44.68%</td>
</tr>
<tr>
<td>Swissprot</td>
<td>24,642</td>
<td>57.17%</td>
</tr>
<tr>
<td>NR</td>
<td>42,809</td>
<td>99.31%</td>
</tr>
</tbody>
</table>

### 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
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 vinifera* (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. 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.
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...
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,
pentanucleotides, and hexanucleotides. 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 monoterprenoid 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, 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 plastidial 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 monoterprenoid 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*.
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


