Whole-transcriptome sequencing of *Pinellia ternata* using the Illumina platform

X. Huang, Y. Jing, D.J. Liu, B.Y. Yang, H. Chen and M. Li

Ministry of Education Key Laboratory of Standardization of Chinese Herbal Medicine, College of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, China

Corresponding author: M. Li
E-mail: 028limin@163.com

Received November 17, 2015
Accepted February 12, 2015
Published July 15, 2016
DOI http://dx.doi.org/10.4238/gmr.15028062

**ABSTRACT.** Pinelliae rhizoma is the dried tuber of *Pinellia ternata* (Thunb.) Breit., and has been used for thousands of years as a traditional Chinese medicine. However, its genomic background is little known. With the development of high-throughput genomic sequencing, it is now easy and cheap to obtain genomic information. In this study, 193,032,910 high-quality clean reads were generated using the Illumina Hiseq 2000 platform. A total of 53,544 unigenes were identified from the contigs assembled. Functional annotation analysis annotated 37,318, 27,697, 23,043, 22,869, 23,328, and 27,415 unigenes. KEGG analysis revealed that five pathways (169 genes) were associated with alkaloid synthesis, 201 unigenes were related to fatty acid biosynthesis (ko00061), and 133 unigenes were involved in the biosynthesis of unsaturated fatty acids (ko01040). In addition, 6703 simple sequence repeats were designed based on the unigene sequences for screening germplasm resources in the future. These data are a valuable resource for genomic studies on *Pinellia* plants.

**Key words:** *Pinellia ternata*; Next-generation sequencing; Organic acids; Alkaloid
INTRODUCTION

Pinelliae rhizoma is the dried tuber of *Pinellia ternata* (Thunb.) Breit. (Figure 1). It has been widely used in Chinese medicine for drying dampness and resolving phlegm, suppressing Qi to stop vomiting, relieving stuffiness, dissipating nodulations, and relieving swelling (Chinese Pharmacopoeia Commission, 2015) by influencing the respiratory, digestive, cardiovascular, and nervous systems. It is anticarcinogenic, anti-inflammatory, and antibacterial (Li et al., 2004). *P. ternata* is widely distributed in China, except in Neimenggu, Jilin, Heilongjiang, Xinjiang, Qinghai, and Tibet Provinces (Li et al., 2009). It is mainly cultivated in Gansu, Shanxi, Shandong, Guizhou, and Yunnan Provinces (Huadong and Faming, 2012). Because of its wide distribution in China, it is difficult to evaluate its quality.

Figure 1. *Pinellia* medicinal materials.

Pinelliae rhizoma contains alkaloids, sterols, organic acids, amino acids, inorganic elements, and nucleosides (Han et al., 2006). Organic acids are regarded as important for evaluating the quality of *P. ternata* (Chinese Pharmacopoeia Commission, 2015), and alkaloids exhibit significant biological and pharmacological activity. Many factors affect the quality of *P. ternata*, including the place of origin, timing of the harvest, germplasm, light, and processing methods.

In recent years, researchers have focused on identifying the genetic diversity of *P. ternata*. The genetic control of heterosis for guanosine and organic acid content has been reported (Luo et al., 2013), and 12 microsatellite markers have been identified that could provide a useful tool for studying the genetic structure of *Pinellia* species (Chen et al., 2008). Inter-simple sequence repeat (ISSR) and sequence-related amplified polymorphism (SRAP) markers have been used to analyze genetic relationships among the *Pinellia* in China (Liu et al., 2012), and random amplified polymorphic DNA (RAPD) and polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) have been employed to authenticate
Pinellia ternata transcriptome

Chinese and Korean herbal medicines (Chung et al., 2002). A total of 54 microRNAs (miRNAs) belonging to 23 miRNA families have been identified (Xu et al., 2012); however, no studies have investigated associations between genes and phenotypic quality in P. ternata, or its DNA and protein sequences. Indeed, little is known about what the genes encode in the species. As of October 2015, only 387 nucleotide sequences, 353 proteins, and 131 genes from P. ternata have been deposited in the GenBank database of National Center for Biotechnology Information (NCBI). Transcriptome analysis is a useful and cost-effective method for discovering new genes and providing information on gene expression and regulation (Mugulies et al., 2005; Huse et al., 2007; Novaes et al., 2008). In this study, we surveyed the poly (A) + transcriptome of P. ternata using next-generation sequencing technology. We obtained 17.33 Gb of raw reads, and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation revealed five pathways (169 genes) that were associated with alkaloid synthesis. A total of 201 unigenes were related to fatty acid biosynthesis (ko00061) and 133 unigenes were involved in the biosynthesis of unsaturated fatty acids (ko01200). These data can be used to study alkaloid, saturated fatty acid, and unsaturated fatty acid synthesis at the gene level in P. ternata, and reveal associations between genes and chemical components. They will also provide research ideas on virus infection, the death of improved varieties, the low coefficient of natural reproduction, the cultivation of low technology, and counterfeit P. ternata plants.

MATERIAL AND METHODS

Plant materials and RNA isolation

P. ternata plants were planted in traditional Chinese medicine-producing areas of the town of Hongjiang (Peng xi County, Suining City, Sichuan Province, China; 30.69°N, 105.45°E) (Figure 2). We selected three strains as one biological replication with three replications (each sample is about 0.5 g) in September 2014, when the plants were growing. After selection, the specimens were washed with pure water, dried using filter paper, and small pieces were cut on ice and immediately frozen in liquid nitrogen. All of the samples were stored at -80°C until analysis.

Figure 2. Pinellia ternata plant.
Total RNA was isolated from each sample using TRIzol (Invitrogen) following the manufacturer instructions. RNase-free DNase I (New England BioLabs) for 30 min at 37°C was used to expunge residual DNA from the total RNA. Equal amounts of RNA from each sample were prepared for mRNA-seq library construction.

**RNA-seq library construction for Hiseq 2000 sequencing**

Total RNA was purified using Oligo(dT) beads (Dynabeads® mRNA Purification Kit, Invitrogen), then poly (A)-containing mRNA was fragmented into 200-250 bp with Fragment Buffer (Ambion). First-strand cDNA was generated using an N6 primer, First-Strand Master Mix, and Super Script II Reverse Transcriptase (Invitrogen). The reaction conditions included 25°C for 10 min, 42°C for 30 min, and 70°C for 15 min, with a 4°C hold. Second-Strand Master Mix was then added to synthesize the second-strand cDNA (16°C for 2 h), the cDNA was purified using a QIAquick® PCR Purification Kit (Qiagen), End Repair Mix was added, and it was incubated at 20°C for 30 min. It was then purified using A-Tailing Mix and incubated at 37°C for 30 min. The purified Adenylate 3'-Ends DNA, Adapter and Ligation Mix was combined, and the ligation reaction was incubated at 20°C for 20 min. A 2% agarose gel was then run in order to select a narrow 300-350-bp size range, and the gel was purified using a QIAquick® Gel Extraction Kit (Qiagen). Several rounds of PCR amplification using PCR Primer Cocktail and PCR Master Mix were performed to enrich the cDNA fragments. The PCR products were then purified with AMPure XP Beads (Agencourt). The final library was quantitated in two ways: average molecule length was determined using an Agilent 2100 Bioanalyzer (Agilent DNA 1000 Reagents), and the library was quantified by PCR (TaqMan® Probe). The quantified libraries were amplified on cBot to generate a cluster on a flow cell (TruSeq PE Cluster Kit v3-cBot-HS, Illumina). Finally, the amplified flow cell was paired-end sequenced on the HiSeq 2000 System (TruSeq SBS KIT-HS V3, Illumina). A read length of 90 was the most common sequencing strategy.

**Sequence data assessment and assembly**

Clean reads were obtained after removing adaptors and low-quality reads (the proportion of reads with a quality value of 10 or less was over 20%). The CycleQ20 value was used to evaluate the clean reads. The *de novo* transcriptome assembly of the short clean reads was conducted using Trinity (release-20130225, http://trinityrnaseq.sourceforge.net/), which recovers full-length transcripts across a broad range of expression levels with a sensitivity similar to methods that rely on genome alignments (Grabherr et al., 2011). The overlap settings used for the assembly were 25 bp and 70% similarity.

**Unigene annotation**

The assembled unigenes were compared against the NCBI non-redundant protein (Nr) and Swiss-Prot databases using BLASTx (version 2.2.14), with an E-value lower than 10⁻⁵. Gene names were assigned to each assembled sequence based on the best BLAST hit (highest score). Blast2GO (Conesa et al., 2005) was then used to obtain a Gene Ontology (GO) annotation of the unigenes. For each unigene within a GO annotation, the WEGO software (Ye et al., 2006) was used to obtain the GO functional classification for the unigenes, and to
ascertain the distribution of gene functions of the species at the macro-level.

Unigenes were aligned to the Clusters of Orthologous Groups (COGs) database to predict and classify possible functions (Tatusov et al., 2000). KEGG pathways were assigned to the assembled unigenes using the online KEGG Automatic Annotation Server (http://www.genome.jp/kegg/kaas/).

Expressed sequence tag-simple sequence repeat (EST-SSR) detection

By using the unigenes as references, SSRs were detected using the MicroSAtellite (MISA) program (Temnykh et al., 2001). Perfect di-, tri-, tetra-, penta-, and hexanucleotide motifs were adjusted to identify the parameters with a minimum of six, five, five, four, and four repeats, respectively. Among the submitted unigenes, we obtained a number of sequences containing SSRs, sequence IDs, SSR motifs, repeats (di-, tri-, tetra-, penta-, and hexanucleotide repeat units), repeat lengths, SSR starts, and SSR ends (Yu et al., 2004; Varshney et al., 2005a,b).

RESULTS

P. ternata transcriptome sequencing and de novo assembly

Three parallel samples of P. ternata were collected and sequenced on the Illumina Hiseq 2000. After removing the adaptors and low-quality reads (Q20, 20), 193,032,910 high-quality clean reads that were 75 bp long were obtained (Table 1). All of the high-quality reads were assembled de novo using Trinity (Grabherr et al., 2011), which produced 189,150, 72,615, and 52,419 contigs from P1, P2, and P3, respectively, with an N50 of 214, 472, and 452 bp, respectively.

<table>
<thead>
<tr>
<th>Transcript length</th>
<th>Total number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-500</td>
<td>33,929</td>
<td>63.37</td>
</tr>
<tr>
<td>500-1000</td>
<td>10,744</td>
<td>20.07</td>
</tr>
<tr>
<td>1000-1500</td>
<td>4616</td>
<td>8.62</td>
</tr>
<tr>
<td>1500-2000</td>
<td>2201</td>
<td>4.11</td>
</tr>
<tr>
<td>2000</td>
<td>2054</td>
<td>3.84</td>
</tr>
<tr>
<td>Total number</td>
<td>53,544</td>
<td></td>
</tr>
<tr>
<td>Total length</td>
<td>32,716,910</td>
<td></td>
</tr>
<tr>
<td>N50 length</td>
<td>907</td>
<td></td>
</tr>
<tr>
<td>Mean length</td>
<td>611</td>
<td></td>
</tr>
</tbody>
</table>

The size distributions of the contigs are presented in Figure 3A-C. The contigs were then assembled into unigenes, the size distributions of which are presented in Figure 3D-F. Although the unigene distributions closely matched the contig distributions with the majority being short sequences, we obtained more long unigenes than long contigs (Figure 3A-C). We obtained 53,544 unigenes with an average length of 611 bp and an N50 of 907 bp (Figure 3G). Of these, 19,615 (36.63%) were longer than 500 bp, 8871 (16.57%) were longer than 1000 bp, and 4255 (7.95%) were longer than 1500 bp. Although the majority of the unigenes were smaller than 500 bp, 2054 (3.84%) were greater than 2000 bp in length. The longest unigene was CL823.Contig2. All that was 18,172 bp long, and according to the COG annotation encodes outer membrane proteins during cell wall/membrane/envelope biogenesis.
In order to analyze the coding DNA sequences (CDSs) and gene annotations, all of the assembled unigenes were first aligned in BLASTx (E-value, $10^{-5}$) to protein databases in the order Nr, Swiss-Prot, KEGG, and COG, i.e., we first aligned the unigenes to Nr, then Swiss-Prot, then KEGG, and finally to COG (Natale et al., 2000). We obtained 42,687 significant BLAST hits (79.72% of all unigenes). The size distributions of the CDSs and predicted proteins are presented in Figure 3H and I. Assembled unigenes that could not be annotated by any database were predicted by ESTScan, and were then translated into peptide sequences (Iseli et al., 1999); 3884 unigenes were analyzed by ESTScan. The size distributions of the ESTs and proteins are presented in Figure 3J and K.

Figure 3. Overview of *Pinellia ternata* transcriptome assembly. **A**–**C**, size distributions of the contigs; **D**–**G**, size distributions of the unigenes; **H**, size distributions of the coding DNA sequences; **I**, size distributions of the proteins; **J** and **K**, size distributions of the expressed sequence tags and proteins.
Functional annotation

Nr annotation

Almost 70.70% (37,318) of the unigenes were annotated using the Nr database. The highest homology (E-value < 10^{-100}) of the unigenes was 16.1% and the moderate homology (10^{-100} < E-value < 10^{-60}) was 14.1% (Figure 4A). A total of 38.1% of the sequences had a higher similarity than 80%, 27.6% had moderate similarities of approximately 60-80%, and 22.8% had a similarity between 40-60% (Figure 4B). For the species distribution of aligned hits, 18.6% of the hits had top matches to sequences from Japanese rice, 15.3% to sequences from *Vitis vinifera*, and 8.2% to sequences from *Cucumis sativus* (Figure 4C). Therefore, the assembled transcripts were high quality.

Figure 4. Characteristics of a homology search of *Pinellia ternata* unigenes. A. E-value distribution of the top BLASTx hits against the non-redundant (Nr) protein database for each unigene; B. similarity distribution of unigene BLASTx hits against the Nr database; C. number of unigenes matching the top seven species using BLASTx in the Nr database.
COG classification

A total of 23,328 unigenes were annotated to 25 functional categories using COG (Figure 5A). “General function prediction only” was the largest COG category (22.81%), followed by “translation, ribosomal structure, and biogenesis” (17.15%), “transcription” (16.92%), “replication, recombination, and repair” (11.71%), “posttranslational modification, protein turnover, chaperones” (11.15%), “signal transduction mechanisms” (11.01%), “amino acid transport and metabolism” (10.76%), “carbohydrate transport and metabolism” (10.70%), “lipid transport and metabolism” (5.04%), and “secondary metabolites biosynthesis, transport, and catabolism” (4.69%). The smallest groups were “nuclear structures” (10 unigenes) and “extracellular structures” (nine unigenes) (Figure 5A).

Figure 5. Clusters of Orthologous Groups (COGs) functional classification and Gene Ontology (GO) classification of the Pinellia ternata transcriptome.

GO classification

A total of 27,415 unigenes were assigned 5043 GO-term annotations using BLAST2GO, and the terms were summarized into the three main GO categories and 56 subcategories (functional groups) (Figure 5B). In the “biological processes” category, a high percentage of genes were involved in “cellular processes” (62.30%) and “metabolic processes” (59.74%), followed by “single-organism processes” (39.09%) and “response to stimulus” (28.18%). For the “cellular components” category, the dominant terms were “cells” (77.93%), “cell parts” (77.92%), and “organelles” (65.87%); “virion” and “virion parts” contained only 17 genes. In the “molecular function” category, the main terms were “binding” (48.87%) and “catalytic activity” (50.30%), followed by “transporter activity” (5.84%) and “structural molecule activity” (4.06%) (Figure 5B).

KEGG pathway mapping

We found 128 KEGG pathways with 22,869 unigenes (the main pathways were “metabolic pathways” (6577, 28.76%) and “biosynthesis of secondary metabolites” (2872, 12.56%). Moderate numbers of unigenes were mapped to the following pathways: “RNA transport” (1707, 7.46%), “mRNA surveillance” (1253, 5.48%), “endocytosis” (1155, 5.05%), “ribosome” (1090, 4.77%), “glycerophospholipid metabolism” (1060, 4.64%), “spliceosome” (1038, 4.54%), “ether lipid metabolism” (932, 4.08%), “plant hormone signal transduction” (864, 3.78%), and “plant-pathogen interactions” (854, 3.73%). Only five unigenes were assigned to “betalain biosynthesis”, and one unigene was assigned to “D-Arginine and D-ornithine metabolism”.

SSR discovery

The MISA analysis identified 6703 putative microsatellites from 5525 unigenes (10.32%). Of these, seven types of SSR were discovered and 926 sequences contained more than one SSR. The most abundant was the trinucleotide repeat motif, which accounted for 39.82%, followed by dinucleotide (30.87%), mononucleotide (26.21%), tetranucleotide (1.22%), pentanucleotide (0.78%), and hexanucleotide repeats (1.10%) (Table 2).

<table>
<thead>
<tr>
<th>Search item</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of sequences examined</td>
<td>53,544</td>
</tr>
<tr>
<td>Total number of SSRs identified</td>
<td>6703</td>
</tr>
<tr>
<td>Number of SSRs containing sequences</td>
<td>5525</td>
</tr>
<tr>
<td>Number of SSRs containing more than one SSR</td>
<td>926</td>
</tr>
<tr>
<td>Number of SSRs present in compound formation</td>
<td>520</td>
</tr>
<tr>
<td>Mononucleotide</td>
<td>1757</td>
</tr>
<tr>
<td>Dinucleotide repeats</td>
<td>2069</td>
</tr>
<tr>
<td>Trinucleotide</td>
<td>2669</td>
</tr>
<tr>
<td>Tetranucleotide</td>
<td>82</td>
</tr>
<tr>
<td>Pentanucleotide</td>
<td>52</td>
</tr>
<tr>
<td>Hexanucleotide</td>
<td>74</td>
</tr>
</tbody>
</table>

Table 2. Summary of simple sequence repeat (SSR) types in Pinellia ternata.
DISCUSSION

Carles et al. (2005) used a silicon-based DNA microarray to identify toxic Chinese medicines associated with *P. ternata*, and found that many have a similar genetic heritage. By directly sequencing *P. ternata* and its adulterant *P. pedatisecta*, Schott’s nuclear genome sequence provides a molecular basis for the original identification of *P. ternata*. Lin et al. (2006) used special DNA primers by PCR to identify different types of *P. ternata* and its adulterants, and found that two specific sites could be identified by the restriction endonucleases *Bam*HI and *Nco*I in the opening DNA sequence. RAPD and PCR-RFLP have also been used to authenticate Chinese and Korean herbal medicines (Chung et al., 2002). In the present study, we obtained a large amount of genetic information using transcription, and a great deal of SSR information (6703 SSRs). It is now possible to develop cDNA SSR (cSSR) tags for breeding by transcriptome sequencing (Wang et al., 2010), and EST-SSR markers from transcriptome sequencing data (Li et al., 2012; Luo et al., 2015). Using the *P. ternata* transcriptome sequencing data, cSSRs can be developed that can be used in genetic mapping, molecular marker-assisted selection, variety identification, etc. Studies of *P. ternata* genes and their components have mainly focused on *Pinellia* proteins, which are involved in clotting, fertility, tumor resistance, and insect resistance. Yao et al. (2003b) first reported the *P. ternata* lectin gene and coded the amino acid sequence. They imported *Pinellia* protein genes into tobacco, which were successfully expressed in response to attack by homopteran pests such as aphids (Yao et al., 2003a). SJ-10 was cloned from the *P. ternata* agglutinin (PTA) gene and inoculated into rice plants, which then exhibited insecticidal activity against the white-backed planthopper, *Sogota furcifera* Horvath (Zhang et al., 2011). In our study, 3884 assembled unigenes that could not be aligned to any database were predicted by ESTScan and then translated into peptide sequences. Using KEGG, 2417 unigenes were mapped to the “amino acid metabolism” pathway and 536 to the “metabolism of other amino acids” pathway. There were 13 pathways in the “amino acid metabolism” pathway and most unigenes were in the “glycine, serine, and threonine metabolism” pathway (293 unigenes, 12.12%) followed by the “cysteine and methionine metabolism” pathway (278 unigenes, 11.50%); only 73 unigenes were aligned to the “lysine biosynthesis” pathway. Regarding the “metabolism of other amino acids” pathway, 186 unigenes (34.70%) were mapped to the “glutathione metabolism” pathway and only one unigene was mapped to the “D-Arginine and D-ornithine metabolism” pathway. These data will provide a reference for research on *Pinellia* proteins and their relationships with genes.

*P. ternata* genes involved in alkaloids and organic acids have been little studied. Liu et al. (2010a,b) investigated the potential use of protocorm-like bodies (PLBs) for studying alkaloid metabolism in *P. ternata* by investigating the induction and *in vitro* alkaloid yield of calluses and PLBs. Organic acids are one of the active ingredients of *Pinelliae rhizoma*, which contains a large amount of unsaturated fatty acids (49.15%) that include linoleic acid (37.10%); the saturated fatty acids present include palmitic acid (15.16%) (Kewei et al., 2002). In the KEGG analysis, 201 unigenes were related to fatty acid biosynthesis (ko00061) and 133 were involved in the biosynthesis of unsaturated fatty acids (ko00061). These genes will provide a reference for research into total organic acid biosynthesis in *P. ternata*.

Only 1093 unigenes (4.69%) were related to secondary metabolite biosynthesis, transport, and catabolism in the COG classification. Twelve pathways were found in the KEGG classification (9.38% of the total number of pathways). In these pathways, five unigenes were
involved in betalain biosynthesis (ko00965) and nine were related to caffeine metabolism (ko00232). Eleven unigenes were involved in indole alkaloid biosynthesis (ko00901), 66 were involved in isoquinoline alkaloid biosynthesis (ko00950), and 78 were involved in tropane, piperidine, and pyridine alkaloid biosynthesis (ko00960). All of these alkaloid metabolism-related genes are potential candidate genes for alkaloid synthesis in *P. ternata*.

There are few data available regarding *P. ternata* genome sequences. Researchers have focused on the complete chloroplast genome sequence of *P. ternata*, which has the potential to advance population and phylogenetic studies of this plant (Han et al., 2015). Suppression subtractive hybridization has been used to clone transcripts that are upregulated during heat stress in *P. ternata* leaves (Lu et al., 2013). Fifteen cDNA fragments have been obtained from the small tubers of *P. ternata*, which probably play an important role during tuber development (Yueqin, 2008).

De novo transcriptome sequencing is the transcriptome profiling of organisms or species based on high-throughput technologies without the aid of a reference genome, and includes library construction, sequencing, and bioinformatic analysis. To our knowledge, this is the first de novo sequencing of three parallel samples from *P. ternata* using an Illumina Hiseq 2000. The sequencing produced 193,032,910 clean reads and 53,544 unigenes after assembly. The number of unigenes obtained from *P. ternata* was lower than that obtained from *Carthamus tinctorius* (120,778) (Lulin et al., 2012), *Polygonum cuspidatum* (86,418) (Hao et al., 2012), and *Salvia miltiorrhiza* (64,139) (Yang et al., 2013). However, the number of unigenes obtained from *P. ternata* was higher than that from *Codonopsis pilosula* (45,511) (Gao et al., 2015), and *Isatis indigotica* (36,367) (Chen et al., 2013). The average unigene length (611 bp) was slightly greater than that of other medicinal plants, including *C. tinctorius* (446 bp) (Lulin et al., 2012), *S. miltiorrhiza* (413 bp) (Yang et al., 2013), and *P. cuspidatum* (365 bp) (Hao et al., 2012). The unigene annotation rate observed in the present study (79.72%) was relatively higher. It is noteworthy that 8871 unigenes were longer than 1000 bp and accounted for 16.57% of all the unigenes, indicating that the Illumina Hiseq 2000 can provide a high-quality transcriptome library and reliable global screening of functional genes in non-model plants, particularly medicinal plants.

According to the GO classification, the unigenes were related to biological processes (75.62%), cellular components (81.03%), and molecular functions (76.19%). A total of 16,378 unigenes were related to metabolic processes, including secondary metabolite biosynthesis, transport, and catabolism. The active ingredients in *P. ternata* may be related to secondary metabolite biosynthesis. There were also high numbers of unigenes that were linked to “response to stimulus” and “signaling”, demonstrating that *P. ternata* can rapidly adapt to changes in the natural environment.

In this study, we obtained 6703 putative microsatellites that can be used to investigate the population-level diversity of molecular markers in future studies. For designing primers and developing nuclear gene markers, unigenes related to low-copy nuclear genes can be used, which is important for studying plant phylogenies. The functions of 3089 (5.77%) unigenes were unknown; therefore, future studies are needed to elucidate these unigenes. To our knowledge, this is the first time that the Illumina Hiseq 2000 platform has been used for three parallel samples of *P. ternata*. It produced about 53,544 unigenes, 69.70% of which were aligned to the Nr database. A total of 1093 unigenes (4.69%) were aligned to secondary metabolite biosynthesis, transport, and catabolism (12 pathways), and 6703 SSRs were identified that will promote the development of the marker. These data provide insights into
the genetic diversity of *P. ternata*, and provide a better understanding of alkaloid and organic acid biosynthesis in *P. ternata* at the genetic level. In conclusion, these data will provide a new resource for genomic studies on the main chemical components of *P. ternata*.

**Conflicts of interest**

The authors declare no conflict of interest.

**ACKNOWLEDGMENTS**

The authors would like to thank the Beijing Genomics Institute (Shenzhen) for assistance with raw data processing and the related bioinformatic analysis. We thank the anonymous referees and editors for carefully reviewing our manuscript and providing constructive comments. Research supported by the National Natural Science Fund Project (#81173492) and the Department of Sichuan Province “Twelfth Five Year” Crop Breeding Research (#2011NZ0098-12).

**REFERENCES**


**Pinellia ternata transcriptome**


Yao JH, Zhao XY, Liao ZH, Lin J, et al. (2003b). Cloning and molecular characterization of a novel lectin gene from...