GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

De novo assembly and characterization of the root transcriptome and development of simple sequence repeat markers in *Paphiopedilum concolor*

D.M. Li, C.Y. Zhao, X.R. Liu, X.F. Liu, Y.J. Lin, J.W. Liu, H.M. Chen and F.B. Lǔ

Guangdong Key Laboratory of Ornamental Plant Germplasm Innovation and Utilization, Environmental Horticulture Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou, China

Corresponding authors: D.M. Li / F.B. Lǔ
E-mail: biology.li2008@163.com / 13660373325@163.com

Genet. Mol. Res. 14 (2): 6189-6201 (2015)
Received August 19, 2014
Accepted January 15, 2015
Published June 9, 2015
DOI http://dx.doi.org/10.4238/2015.June.9.5

ABSTRACT. *Paphiopedilum* orchids (Orchidaceae) have attracted much attention from botanists and horticulturists because of their peculiar leaves and beautiful flowers. Furthermore, the dry roots of *Paphiopedilum* plants have well-known medicinal uses. However, it is unknown how sensitive and plastic the root genes are to environmental changes or how these environmental changes regulate the biosynthesis of active ingredients. In this study, we chose *Paphiopedilum concolor* for root sequencing, as it is widely used as a parent in breeding experiments. A total of 3.77 Gb of sequence data were generated by Illumina paired-end sequencing. *De novo* assemblies yielded 72,952 contigs, 67,434 scaffolds, 64,304 unigenes with average lengths of 937, 1022, and 1047 bp, respectively. Based on Basic Local Alignment Search Tool with known protein sequences, 40,815 (63.5%) unigenes were annotated with an E-value cutoff of 1.0E-5. Among the unigenes, 24,605 were classified in the Gene Ontology database, 17,361 were assigned to Cluster of Orthologous

Groups, and 14,170 were annotated in Kyoto Encyclopedia of Genes and Genomes. Among these annotations, over 1195 unigenes related to secondary metabolic pathways, as well as 609 unigenes involved in plant hormone synthesis and signal transduction, were identified. In addition, 5322 potential simple sequence repeats (SSRs) were identified, and 4989 primer pairs for 3975 sequences containing SSRs were obtained. This study provides valuable insights into the mechanisms of genes that regulate root growth and development and provides a comprehensive resource for genes related to secondary metabolism in roots and for marker-assisted studies in *Paphiopedilum*.

Key words: Illumina paired-end sequencing; *Paphiopedilum concolor*; Simple sequence repeat markers; Root transcriptome

INTRODUCTION

Paphiopedilum orchids, commonly known as lady's slipper orchids in horticulture, belong to the family Orchidaceae (Cox et al., 1997). They are distributed mainly in tropical and subtropical regions from Asia to the islands of the Pacific Ocean. *Paphiopedilum* species usually occur in the understory in karst limestone areas below 2000 m (Cribb, 1998).

As well-known horticultural plants, *Paphiopedilum* plants have attracted much attention from botanists and horticulturists (Williams et al., 1983; Assmann and Zeiger, 1985; Cribb, 1998; Guan et al., 2011; Zhang et al., 2011). Botanists and horticulturists mainly focus their attention on peculiar leaves and beautiful flowers. For example, the *Paphiopedilum* leaf lacks chloroplasts in its guard cells (Williams et al., 1983; Assmann and Zeiger, 1985; Guan et al., 2011; Zhang et al., 2011). This lack of chloroplasts slows the induction of photosynthesis and ecophysiologically acclimatizes the leaf to low light and nutrient-poor environments with limited water (Assmann and Zeiger, 1985; Guan et al., 2011; Chang et al., 2011; Zhang et al., 2011). However, few studies have evaluated how sensitive and plastic the root genes of *Paphiopedilum* are to environmental changes with regard to evolution. Furthermore, the dry *Paphiopedilum* roots are well known for their medicinal use. They are used as a treatment to reduce inflammation and acesodyne, dysmenorrheal, colic, and cystitis (http://baike.baidu.com/view/74271.htm?fr=aladdin). To date, little is known regarding how the root genes of *Paphiopedilum* plants are involved in the biosynthesis of active ingredients.

Conventional methods for gene cloning and sequencing are not only time-consuming and expensive, but also produce only a limited amount of genetic information. Currently, next-generation sequencing technologies, such as Illumina Genome Analyzer, the Roche/454 Genome Sequencer FLX Instrument, and the ABI SOLiD System, have proven to be powerful and cost-effective tools for advanced research in many areas, including *de novo* transcriptome sequencing, gene discovery, expression profiling analysis, and molecular marker development in non-model organisms (Hsiao et al., 2011; Wang et al., 2012; Li et al., 2013). Recently, the mature flower transcriptome of *Paphiopedilum armeniacum* has been sequenced using next-generation sequencing (Tsai et al., 2013). However, the comprehensive gene expression profiles of *Paphiopedilum* roots remain unavailable. Moreover, expressed sequence tag (EST) collections can contribute to the development of molecular markers for a variety of applications in plant genetics and molecular breeding, although only a few EST-derived markers from *Paphiopedilum* have been identified and

utilized (Chung and Choi, 2012; Parveen et al., 2012). Therefore, extensive transcriptomic sequence data are needed to elucidate molecular mechanisms related to root growth and development as well as develop new molecular markers for *Paphiopedilum* orchids.

In this study, we aimed to provide a large collection of assembled and functionally annotated cDNAs from *Paphiopedilum* roots as well as to identify the genes that control root growth and development in *Paphiopedilum*. We selected one *Paphiopedilum* species, *Paphiopedilum concolor*, for root sample collection, because this species is widely used as a parent in breeding. Therefore, we carried out a comprehensive analysis of transcriptome sequences of the *P. concolor* root. *De novo* assembly of the transcriptome sequence reads generated from Illumina HiSeq™ 2000 resulted into 3.77 Gb of transcript data with 64,304 unigenes. Then, the root unigenes were functionally annotated by Basic Local Alignment Search Tool (BLAST) against public databases. Subsequently, the annotated sequences were classified into putative functional categories. Finally, we detected and characterized the EST-derived simple sequence repeat (SSR) markers from the root transcriptome of *P. concolor*. This represents the first report of publically available pyrosequencing data for roots in *Paphiopedilum* orchids. These results will provide background for functional studies of *P. concolor* and its relatives about root growth and development, but also be useful for future marker-assisted studies in *Paphiopedilum* orchids.

MATERIAL AND METHODS

Plant materials and RNA extraction

P. concolor was used for this study. Plants were grown in a greenhouse under natural light in the Environmental Horticulture Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou, China. Plants were watered and fertilized as needed. The samples consisted of 3-year-old mature root tissues, and newly grown root tissues were collected on April 18, 2012. To avoid potential expression differences among collections due to circadian rhythms, root tissues were only collected from individual plants between 9:00 and 10:00 am. The sample was collected and pooled from equal amounts of new and mature roots from five seedlings. The pooled sample was placed immediately into liquid nitrogen, and stored at -80°C until RNA was extracted.

Total RNA of the pooled root sample was isolated using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) according to manufacturer instructions. To avoid genomic DNA contamination, RNA was treated with RNase-free DNase I (TaKaRa, Dalian, China). RNA quality and quantity were analyzed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and a Nanodrop ND1000 (NanoDrop Technologies, Wilmington, DE, USA), respectively.

cDNA library construction and sequencing

The root cDNA library of *P. concolor* was prepared using an Illumina kit (Illumina, San Diego, CA, USA) following manufacturer recommendations. Briefly, the poly(A) mRNA was purified from 6.0 μg total RNA of pooled root sample using oligo(dT) magnetic beads and fragmented into short sequences using divalent cations under elevated temperatures. First- and second-strand cDNAs were synthesized from cleaved RNA fragments. After the end repair and ligation of adaptors, the products were cleaned up with a QIAquick PCR Purification Kit (Qiagen, Valencia, CA, USA) to create the final cDNA library. Finally, after validating using

an Agilent Technologies 2100 Bioanalyzer using the Agilent DNA 1000 chip kit, the root cDNA library was sequenced using Illumina HiSeq™ 2000 to obtain short sequences from both ends at Shanghai Biotechnology Corporation (SBC) in Shanghai, China.

Data filtering and *de novo* assembly

The raw reads were cleaned by removing the adapter and low-quality sequences, which included the reads with ambiguous nucleotides and more than 10% nucleotides with Q-value ≤20. The clean reads were assembled with the CLC Genomics Workbench software (CLC Bio, Denmark, http://www.clcbio.com/) using the following parameters: conflict resolution (vote), 95% similarity of 100 bp over read length and alignment mode (global, do not allow InDels). The reads were then re-assembled twice with CAP3 version 10/15/07 (Huang and Madan, 1999) using settings (threshold identity cutoff of 95% over 500 bp) and stringent parameters (threshold identity cutoff of 95% over 800 bp), respectively. Briefly, CLC first combined reads with a particular overlap to form longer fragments without N to form contigs. Next, the reads were mapped back to the contigs using CLC to construct scaffolds with the paired-end information. The program detected contigs from the same transcript as well as the distances between these contigs. Next, CLC connected the contigs between each pair of contigs using N to represent unknown bases, thus generating scaffolds. Next, the assembled scaffolds were re-assembled twice by CAP3 for gap filling. The sequences with the lowest Ns and those that could not be extended on either end were obtained. Such sequences were defined as unigenes. The unigenes were constructed for the *P. concolor* root.

Functional annotation

Functional unigenes were also aligned with sequences in the National Center for Biotechnology Information (NCBI) non-redundant (Nr) database (http://www.ncbi.nlm.nih.gov) and Swiss-Prot protein database (http://www.expasy.ch/sprot) using the BLASTx algorithm (Altschul et al., 1997) with an E-value of less than 1.0E-5. Blast2GO (Conesa et al., 2005) was used to obtain Gene Ontology (GO) annotation of the unigenes based on BLASTx top hits against the NCBI Nr database with an E-value cutoff 1.0E-5. The unigene sequences were also aligned to the Cluster of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG) to predict and classify functions with E-value cutoff of 1.0E-5. Pathway assignments were carried out based on the online Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (KAAS) (http://www.genome.jp/tools/kaas/) (Moriya et al., 2007). The Bi-directional Best Hit (BBH) method was used to obtain KEGG Orthology (KO) assignment.

SSR mining and primer design

The unique sequences of each floral tissue were used to search for potential microsatellite markers using MIcroSAtellite (MISA) (http://pgrc.ipk-gatersleben.de/misa/). The parameters were adjusted for identification of perfect di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of 5, 5, 5, 5, and 5 repeats, respectively. Based on MISA results, BatchPrimer 3 (You et al., 2008) was used to design the primer pairs with settings as following: 1) primer length of 18-28 bases (average 22 bases); 2) annealing temperature between 55° and 65°C (average 58°C) with a maximum discrepancy within 4°C between the primer pairs; and 3) PCR product size of 100-500 bp (average 300 bp).

RESULTS AND DISCUSSION

Illumina sequencing and *de novo* assembly

In this study, using Illumina HiSeq™ 2000 sequencing, root transcriptomic data for *P. concolor* was generated. After cleaning and quality checks, we obtained 37.718 million clean reads for the *P. concolor* root, encompassing 3.77 Gb of sequence data (Table 1). The raw read data of *P. concolor* roots have been submitted to the NCBI Short Read Archive under the accession No. SRR1405687.

A *de novo* assembly was performed for the root transcriptomic data of *P. concolor*. An overview of the sequencing and assembly is provided in Table 1. Based on the high-quality reads, a total of 72,952 contigs with an average length of 937 bp were assembled from the *P. concolor* root library (Table 1). Based on the paired-end information of the corresponding assembled contigs, 67,434 scaffolds were obtained with an average length of 1022 bp from the *P. concolor* root library (Table 1). After further gap filling, 64,304 unigenes were generated from the *P. concolor* root library with an average length of 1047 bp (Table 1). The length distributions of the contigs, scaffolds, and unigenes are shown in Figure 1.

Table 1. Statistics for sequencing and assembly of the *Paphiopedilum concolor* root transcriptome.

Tissue	Root
Total number of clean reads	37,718,879
Number of contigs	72,952
Average length of contigs (bp)	937
Total length of contigs (Nt)	68,362,516
N50 length of contigs	1,286
Number of scaffolds	67,434
Average length of scaffolds (bp)	1,022
Total length of scaffolds (Nt)	68,915,058
N50 length of scaffolds	1,395
Number of unigenes	64,304
Average length of unigenes (bp)	1,047
Total length of unigenes (Nt)	67,310,919
N50 length of unigenes	1,447

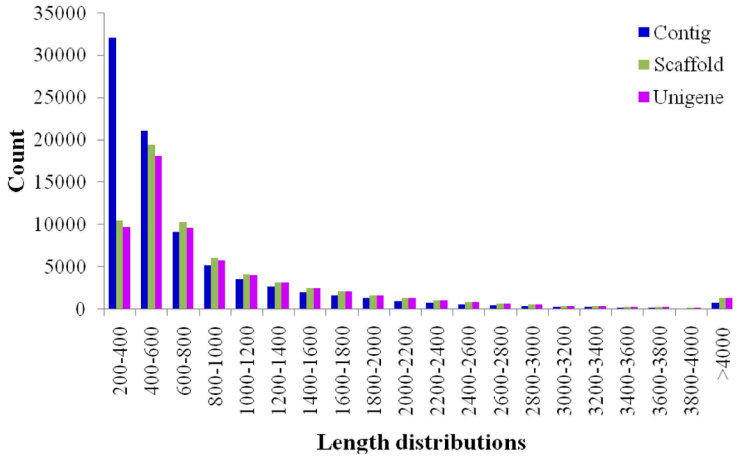

Figure 1. Overview of the sequencing and assembly of the *Paphiopedilum concolor* root transcriptome sequences.

Functional annotation of predicted proteins

All unigenes of the *P. concolor* root library that was generated by Illumina sequencing were aligned to public protein databases (Nr, Swiss-Prot, COG, and KEGG) by BLASTx (E-values < 1.0E-5). A total of 40,815 (63.5%) of all unigenes were annotated in this manner (Table 2), while the remainder (36.5%) had no significant match in any of the existing databases and require more genetic data to annotate. The Nr database annotation results are shown in **Table S1**. The E-value distribution of the top hits in the Nr database showed that 54.55% of the root library's mapped sequences have strong homology (<1.0E-50), whereas 45.45% of the root library's homologous sequences ranged between 1.0E-5-1.0E-49 (Figure 2A). The species distribution of the best match results for the root library's sequences is shown in Figure 2B. Among these, the sequences had a 21.06% match with those of *Vitis vinifera*, followed by *Nectria haematococca* mpVI 77-13-4 (7.71%), *Oryza sativa* Japonica Group (5.84%), *Populus trichocarpa* (4.65%), *Sorghum bicolor* (3.59%), *Glycine max* (3.41%), *Brachypodium distachyon* (3.22%), *Serpula lacrymans* var. lacrymans S7.3 (2.89%), *O. sativa* Indica Group (2.57%), and *Zea mays* (2.52%) (Figure 2B). The other 42.48% matched with sequences from other plant species, animals, insects, and viruses, of which a very low proportion (<2.5%) was assigned to each species. In *Vanda* Mimi Palmer, a hybrid of *Vanda* orchids, BLAST hits to non-plant organisms were also present in a floral cDNA library (Teh et al., 2011). In *Phalaenopsis* orchids, transcripts encoding triple gene block 1, triple gene block 2, triple gene block 3, and RNA-dependent RNA polymerase of *Cymbidium* mosaic virus were also found in 10 samples of different tissues by using pyrosequencing (Hsiao et al., 2011). Based on these results, it is tempting to speculate whether the root sample sequences were infected by plant viruses or if the gene is included in the *P. concolor* genome.

Table 2. Summary of annotations for the *Paphiopedilum concolor* root transcriptome.

Public protein database	Number of unigene hits	Percentage
Nr	40,397	62.8
Swiss-Prot	30,054	46.7
COG	17,361	26.9
KEGG	14,170	22.0
Total	40,815	63.5

GO assignments

The GO functional annotations of the *P. concolor* root unigenes with Nr annotations were obtained by using Blast2GO. In total, 24,605 of *P. concolor* root annotated unigenes were further classified into functional 59 GO terms (Figure 3). These GO assignments were divided into three main categories: biological process, cellular component, and molecular function. Predicted proteins that were assigned to biological process were mainly associated with metabolic process (14.08%), cellular process (12.25%), single-organism process (4.27%), biological regulation (2.24%), localization (2.17%), establishment of localization (2.15%), regulation of biological process (2.14%), and response to stimulus (1.81%). Those assigned to cellular component were mainly related to cell (7.17%), cell part (7.17%), organelle (4.69%), membrane (3.49%), membrane part (2.11%), macromolecular complex (2.11%), and organelle part (1.66%). Finally, those assigned to molecular function were mainly linked to binding

(11.71%), catalytic activity (10.55%), transporter activity (1.18%), structural molecule activity (0.85%), nucleic acid binding transcription factor activity (0.44%), electron carrier activity (0.32%), and enzyme regulator activity (0.28%).

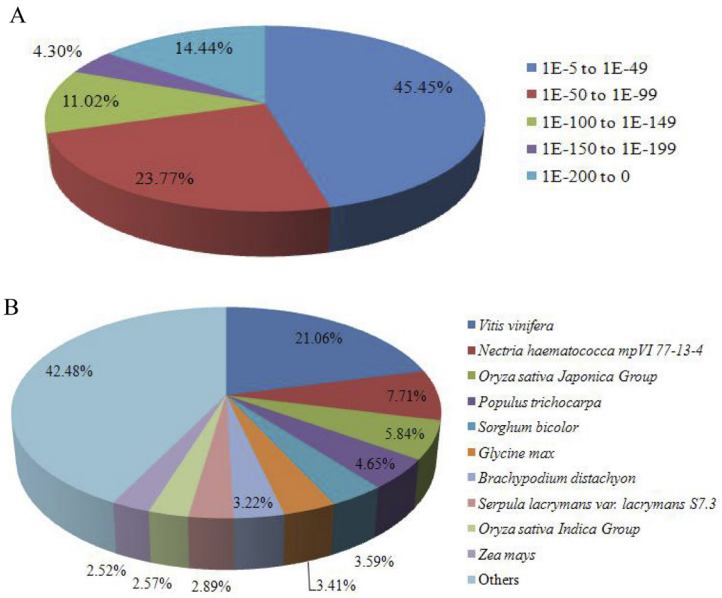

Figure 2. Characteristics of homology search of the *Paphiopedilum concolor* root Illumina sequences against the Nr database. **A.** E-value distribution of the BLAST hits for each unique sequence with a cutoff E-value of 1.0E-5. **B.** Species distribution of the BLAST results. We used the first hit of each sequence for analysis.

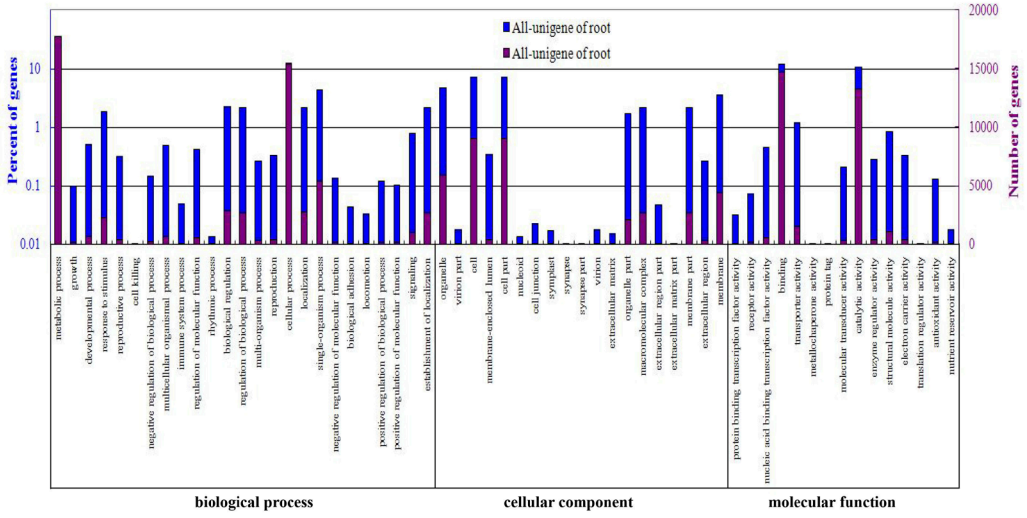

Figure 3. Gene ontology classification of *Paphiopedilum concolor*'s root unigenes by BLASTx with an E-value threshold of 10^{-5} against the Nr database.

COG classification

A total of 17,361 unigenes of the root library were assigned to the appropriate COG clusters (Figure 4). These COG classifications were grouped into 25 functional categories (Figure 4). Among the 25 COG categories, the 12 largest categories were signal transduction mechanisms (14.60%); posttranslational modification, protein turnover, and chaperones (13.27%); general function prediction only (10.58%); intracellular trafficking, secretion, and vesicular transport (6.46%); translation, ribosomal structure, and biogenesis (6.11%); RNA processing and modification (5.78%); cytoskeleton (4.89%); energy production and conversion (4.32%); transcription (4.31%); carbohydrate transport and metabolism (3.48%); lipid transport and metabolism (3.47%); and function unknown (3.29%).

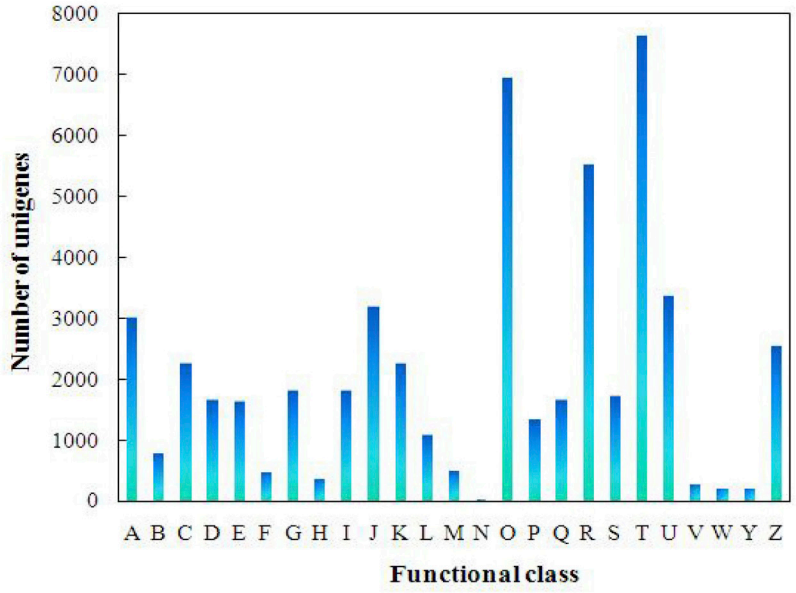

Figure 4. COG functional classification of the *Paphiopedilum concolor* root transcriptome. All putative proteins were aligned to the COG databases and can be functionally classified into 25 clusters. A = RNA processing and modification; B = chromatin structure and dynamics; C = energy production and conversion; D = cell cycle control, cell division, and chromosome partitioning; E = amino acid transport and metabolism; F = nucleotide transport and metabolism; G = carbohydrate transport and metabolism; H = coenzyme transport and metabolism; I = lipid transport and metabolism; J = translation, ribosomal structure, and biogenesis; K = transcription; L = replication, recombination, and repair; M = cell wall/membrane/envelope biogenesis; N = cell motility; O = posttranslational modification, protein turnover, and chaperones; P = inorganic ion transport and metabolism; Q = secondary metabolites biosynthesis, transport, and catabolism; R = general function prediction only; S = function unknown; T = signal transduction mechanisms; U = intracellular trafficking, secretion, and vesicular transport; V = defense mechanisms; W = extracellular structures; Y = nuclear structure; Z = cytoskeleton.

KEGG analysis

To identify the biological pathways that are active in roots of *P. concolor*, we mapped

the 40,815 annotated unigenes to the referential canonical pathways in the KEGG database. A total of 14,170 unigene sequences were assigned to 267 KEGG pathways. All of the pathways are summarized in **Table S2**. The top 12 pathways were metabolic pathways (3419 unigenes), biosynthesis of secondary metabolites (1680 unigenes), microbial metabolism in diverse environments (929 unigenes), ribosome (837 unigenes), protein processing in endoplasmic reticulum (835 unigenes), Epstein-Barr virus infection (662 unigenes), pyrimidine metabolism (579 unigenes), oxidative phosphorylation (451 unigenes), RNA transport (423 unigenes), spliceosome (376 unigenes), PI3K-Akt signaling pathway (374 unigenes), and tight junction (373 unigenes).

Identification of genes involved in plant hormone signal transduction

In this study, various genes related to the pathways of plant hormone signal transduction were found and contained the plant growth regulators, such as auxin, cytokinin, gibberellin, ethylene, abscisic acid, brassinosteroid, salicylic acid, and jasmonic acid (Table 3). Interestingly, almost all of the genes involved in the plant hormone signal transduction pathway (ko04075) had been identified, except for seven genes that remain unidentified (Figure 5). The plant hormones mediate diverse growth and developmental processes and defense responses. For example, auxin plays critical roles in root formation, apical dominance, tropic response, and senescence; brassinosteroids involve in maintaining optimal leaf angle and plant height, fruit development, and regulation of grain filling (Wang and Chong, 2010). These data indicate that plant hormones may play essential and indispensable roles in normal root growth and development of the *P. concolor* life cycle.

Table 3. Numbers of unigenes involved in the pathways and products of plant hormone signal transduction.

Pathway (ID)	Product	Mapped KO	All pathway KO	Number of unigenes
Cysteine and methionine metabolism (ko00270)	Ethylene	39	64	164
Phenylalanine metabolism (ko00360)	Salicylic acid	20	49	91
Carotenoid biosynthesis (ko00906)	Abscisic acid	15	27	33
Tryptophan metabolism (ko00380)	Auxin	21	61	94
Linoleic acid metabolism (ko00591)	Jasmonic acid	9	11	23
Diterpenoid biosynthesis (ko00904)	Gibberellin	7	33	17
Brassinosteroid biosynthesis (ko00905)	Brassinosteroid	7	8	12
Zeatin biosynthesis (ko00908)	Cytokinin	3	8	11
Indole alkaloid biosynthesis (ko00901)	Indole-acetic acid	2	10	4
Plant hormone signal transduction (ko04075)	-	36	43	160

Identification of genes related to secondary metabolic pathways

In the current study, 1195 unigenes were mapped into 34 secondary metabolic pathways (Figure 6). Moreover, we identified several genes in our root transcriptome data that encode key enzymes in the major secondary metabolites, for example, within pathways for streptomycins, terpenoids, phenylpropanoids, alkaloids, biotins, carotenoids, and flavonoids (Figure 6). These secondary metabolic genes provide more abundant information for detecting the biosynthesis process of secondary metabolites in the *P. concolor* root.

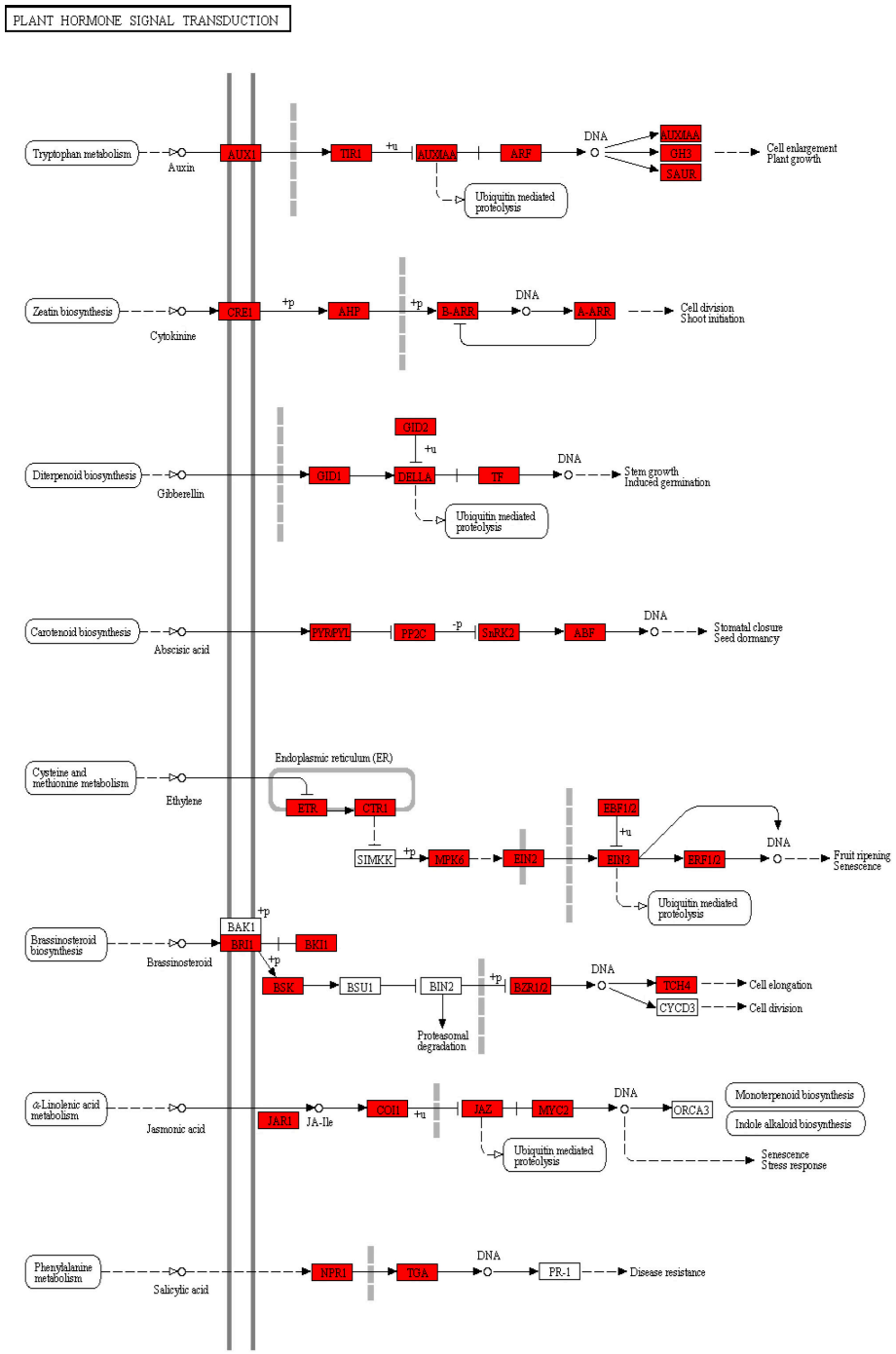

Figure 5. Unigenes predicted to be involved in the plant hormone signal transduction pathway. Red indicates genes that were identified in the analysis; white indicates genes that were not identified.

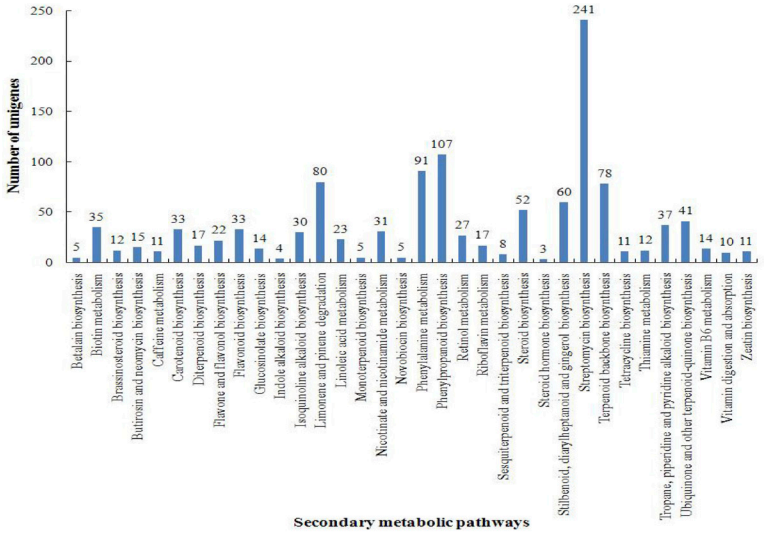

Figure 6. Unigenes from the *Paphiopedilum concolor* root transcriptome that are related to secondary metabolism.

EST-SSR marker detection and characterization

In this study, a total of 5322 potential EST-SSR markers were identified from 4878 unique sequences from the root library (Figure 7A).

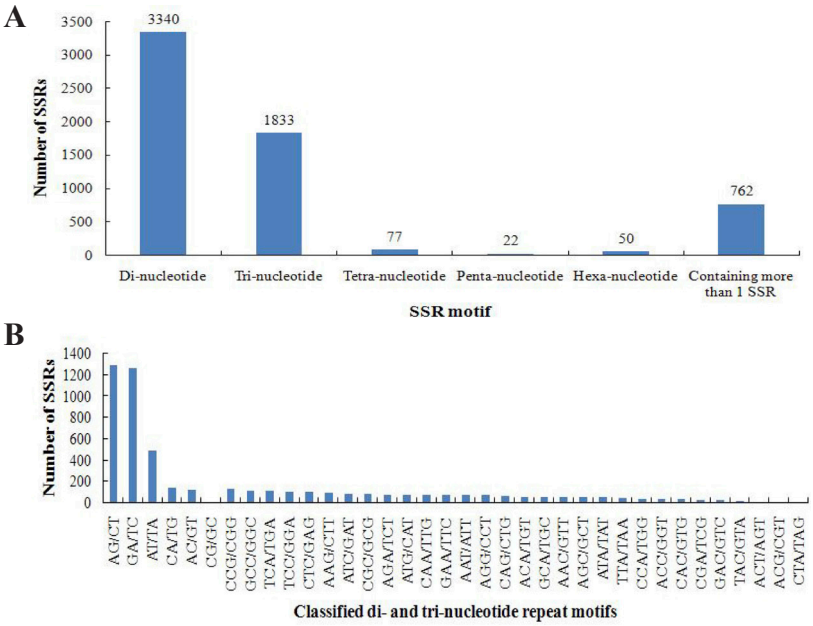

Figure 7. Characterization of SSRs that were identified in the *Paphiopedilum concolor* root transcriptome. **A.** Distribution to different repeat motifs. **B.** Distribution of classified di- and tri-nucleotide repeat motifs.

Of these, 762 sequences from the root library contained >1 EST-SSR (Figure 7A). The EST-SSRs included 3340 (62.75%) di-nucleotide SSR motifs from the root library, followed by tri-nucleotide (1833, 34.44%), tetra-nucleotide (77, 1.44%), hexa-nucleotide (50, 0.93%), and penta-nucleotide motifs (22, 0.41%) (Figure 7A). Among the di-nucleotide motifs, AG/CT (38.86%) was the most abundant repeat type, followed by GA/TC (37.99%), AT/TA (14.76%), CA/TG (4.31%), and AC/GT (3.74%), respectively (Figure 7B). Among the tri-nucleotide motifs, CCG/CGG (7.20%) was the most abundant motif, followed by GCC/GGC (6.27%), TCA/TGA (6.16%), TCC/GGA (5.56%), CTC/GAG (5.40%), and AAG/CTT (4.96%), respectively (Figure 7B). Moreover, based on the criteria described for SSRs screening, 4989 primer pairs from 3975 sequences that contain SSRs in the root library were successfully designed using BatchPrimer3 (**Table S3**).

In total, 7.58% unigene sequences possessed SSRs. This SSR frequency is consistent with the range frequencies that were reported for dicotyledonous plant species (2.65-16.82%) (Kumpatla and Mukhopadhyay, 2005). Di-nucleotide motifs were the most frequent SSR motif type. This finding is consistent with the results that were reported for sugar beets, cabbage, soybeans, sunflowers, grapes, and sesame plants (Kumpatla and Mukhopadhyay, 2005; Wei et al., 2011), whereas tri-nucleotide motifs were the most abundant SSRs in rice, wheat, barley, radish, and *Cymbidium ensifolium* (La Rota et al., 2005; Wang et al., 2012; Li et al., 2013). Among the di-nucleotide repeats, AG/CT was the most abundant motif in our data (Figure 5B). This finding is consistent with the results that were reported for other plant species (Wei et al., 2011; Wang et al., 2012). Among the tri-nucleotide motifs, the most frequent motifs in our data were CCG/CGG, whereas AAG/CTT were the most frequent motifs in other plant species, such as radish, sesame, and *C. ensifolium* (Wei et al., 2011; Wang et al., 2012; Li et al., 2013).

CONCLUSION

In this study, *P. concolor* root transcriptome was analyzed and characterized. There were a large number of assembled unigenes (64,304), and detected SSR markers (5322) were derived from the *P. concolor* root transcriptome. These results demonstrate that Illumina paired-end sequencing is a fast and cost-effective approach to novel gene discovery and molecular marker development in non-model organisms that lack a reference genome. Based on the criteria for testing SSR markers, 4989 primer pairs were obtained from 3975 SSR sequences. These EST-SSR markers and primers will enable the construction of a genetic linkage map, quantitative trait loci mapping, and marker-assisted selection breeding studies in *Paphiopedilum* orchids. Additionally, the dataset improves our understanding of the molecular mechanisms of root growth and development, plant hormone signal transduction, secondary metabolism, and other metabolic processes in *Paphiopedilum* orchids. Although it will be necessary to validate the functions of the identified genes related to secondary metabolism and plant hormone signal transduction-related genes, these results give us a starting point for future functional research on these fields in *Paphiopedilum* orchids.

ACKNOWLEDGMENTS

Research supported by the Guangzhou Municipal Science and Technology Project (#12C14071654) and the Guangdong Academy of Agricultural Sciences Fund (#201019).

Supplementary material

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Assmann SM and Zeiger E (1985). Stomatal responses to CO_2 in *Paphiopedilum* and *Phragmipedium*: role of the guard cell chloroplast. *Plant Physiol.* 77: 461-464.

Chang W, Zhang SB, Li SY and Hu H (2011). Ecophysiological significance of leaf traits in *Cypripedium* and *Paphiopedilum*. *Physiol. Plantarum* 141: 30-39.

Chung SY and Choi SH (2012). Genetic variability and relationships among interspecific hybrid cultivar and parental species of *Paphiopedilum* via ribosomal DNA sequence analysis. *Plant Syst. Evol.* 298: 1897-1907.

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

Cox AV, Pridgeon AM, Albert VA and Chase MW (1997). Phylogenetics of the slipper orchids (Cypripedioideae, Orchidaceae): nuclear rDNA ITS sequences. *Plant Syst. Evol.* 208: 197-223.

Cribb P (1998). The genus *Paphiopedilum*. 2nd edn. Natural History Publications, Borneo.

Guan ZJ, Zhang SB, Guan KY, Li SY, et al. (2011). Leaf anatomical structures of *Paphiopedilum* and *Cypripedium* and their adaptive significance. *J. Plant Res.* 124: 289-298.

Hsiao YY, Chen YW, Huang SC, Pan ZJ, et al. (2011). Gene discovery using next-generation pyrosequencing to develop ESTs for *Phalaenopsis* orchids. *BMC Genomics* 12: 360.

Huang X and Madan A (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9: 868-877.

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

La Rota M, Kantety R, Yu JK and Sorrells M (2005). Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* 6: 23.

Li X, Luo J, Yan T, Xiang L, et al. (2013). Deep sequencing-based analysis of the *Cymbidium ensifolium* floral transcriptome. *PLoS One* 8: e85480.

Moriya Y, Itoh M, Okuda S, Yoshizawa A, et al. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35: W182-W185.

Parveen I, Singh HK, Raghuvanshi S, Pradhan UC, et al. (2012). DNA barcoding of endangered Indian *Paphiopedilum* species. *Mol. Ecol. Resour.* 12: 82-90.

Teh SL, Chan WS, Abdullah JO and Namasivayam P (2011). Development of expressed sequence tag resources for *Vanda* Mimi Palmer and data mining for EST-SSR. *Mol. Biol. Rep.* 38: 3903-3909.

Tsai WC, Fu CH, Hsiao YY, Huang YM, et al. (2013). OrchidBase 2.0: comprehensive collection of Orchidaceae floral transcriptomes. *Plant Cell Physiol.* 54: e7.

Wang L and Chong K (2010). Auxin, brassinosteroids, and G-protein signaling. Integrated G Proteins Signaling in Plants (Yalovsky S, et al., eds.). *Signal. Commun. Plants.* Doi: 10.1007/978-3-642-03524-1_8: 135-154.

Wang S, Wang X, He Q, Liu X, et al. (2012). Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. *Plant Cell Rep.* 31: 1437-1447.

Wei W, Qi X, Wang L, Zhang Y, et al. (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.

Williams WE, Grivet C, Zeiger E (1983). Gas exchange in Paphiopedilum: lack of chloroplasts in guard cells correlates with low stomatal conductance. Plant Physiol. 72: 906-908.

You FM, Huo N, Gu YQ and Luo M (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* 9: 253.

Zhang SB, Guan ZJ, Chang W, Hu H, et al. (2011). Slow photosynthetic induction and low photosynthesis in *Paphiopedilum armeniacum* are related to its lack of guard cell chloroplast and peculiar stomatal anatomy. *Physiol. Plantarum* 142: 118-127.