Plant Omics Journal POJ

POJ 7(5):328-336 (2014) ISSN:1836-3644

Transcriptome analysis of differential expression genes from petals and lips of *Phalaenopsis amabilis* to identify genes associated with floral development

Yan-Xiao Li, Xiao-Ming Song, Zhen Wang, Shan-Wu Lv, Chang-Wei Zhang*

State Key Laboratory of Crop Genetics and Germplasm Enhancement/Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in East China, Ministry of Agriculture, Nanjing Agricultural University, Nanjing 210095, China

∗Corresponding author: changweizh@njau.edu.cn

Abstract

Phalaenopsis amabilis belongs to the genus *Phalaenopsis*, which is member of Orchidaceae, one of the largest families of flowering plants. Orchidaceae includes species with greatly diversified and specialized floral morphology. But few genomic resources are available for these non-model plants. In order to understand the genetic mechanisms underlying floral patterning, we investigated differential expression genes from petals and lips of this species. Two libraries were prepared from petals and lips and then sequenced using short reads sequencing technology (Illumina) to identify the differential expression genes. The total reads were 8,889,080 and 16,224,038 for these two libraries, respectively. The open reading frame of unigenes was predicted using Getorf software. The MISA software was used to identify SSR markers for the unigenes that were larger than 1 Kb. Sequencing data between samples compared with the Unigene database, using soapsnp build consensus sequence, and then analyzed between samples to get homozygous SNP loci. By comparing the transcripts from petals and lips, we finally obtained 2,389 differentially expressed genes. These genes were significantly enriched in 101 KEGG pathways and 55 GO terms. The transcriptome analysis provided a comprehensive understanding of the complexity of floral development and organ identity. The results let us know more details about the relationship between MADS gene family and floral morphology. This information broadens our understanding of the mechanisms of floral patterning and contributes to molecular and genetic research by enriching the *Phalaenopsis* database.

Keywords: *Phalaenopsis amabilis*, Transcriptome, sequencing, differential expression genes, MADS-genes.
Abbreviations: MADS_genes; GO_Gene ontology.

Introduction

The family of Orchidaceae is one of the largest and most widespread families of flowering plants, with contains more than 250,000 species (Leitch et al., 2009). They show a wide diversity of epiphytic and terrestrial growth forms and have successfully colonized almost every habitat on earth (Cozzolino and Widmer, 2005; Hsu et al., 2011). Phalaenopsis, a very important economically flowering genus, is member of Orchidaceae. *Phalaenopsis amabilis* is a plant of the orchid genus *Phalaenopsis* and native to Taiwan. It has become an important commodity in the international floral trade, and they are among the top-traded blooming potted plants worldwide (Blanchard and Runkle, 2006). Hybrids of this genus are of great economic value as house and garden plants as well as cut flowers. *Phaleonopsis* plays an important role in the development of novel species and it is frequently used to cross with other hybrids. The well-known *P. amabilis* is one of the most important ancestor species of *Phalaenopsis* hybrids (Semiarti et al., 2007).

Among the angiosperms, orchids are unique in their floral patterning, particularly in floral structures and organ identity. The beautiful orchid flower is bilaterally symmetrical. With fascinating complexity, the orchid flower includes three outer tepals (sepals) in the first floral whorl, two lateral inner tepals (petals) and a highly modified median inner tepal (lip or labellum) in the second whorl (Rudall and Bateman, 2002; Mondragón-Palomino and Theißen, 2008; Aceto and Gaudio, 2011). Compared with the tepals, the lip is usually decorated with calli, spurs and glands, and exhibits a distinctive shape (Mondragón-Palomino and Theißen, 2009). It also shares a different color pattern from that of the tepals. The deep analysis of differentially expressed genes from lip and petals will provide a comprehensive understanding of the transcriptome complexity of floral structures and organ identity.

Basic mechanisms of flower development have been elucidated primarily in model plants including Antirrhinum, Arabidopsis and Petunia (Su et al., 2011). The molecular studies on orchids groups are scarce, compared to other model plants (Tsai et al., 2004; Xu et al., 2006; Hsiao et al., 2011). With extreme diversity and specialization of floral morphology, a deep analysis of genes involved in the flower development of this family has a very important significance. The genome projects and sequencing efforts are already under development to provide a collection of Orchid genes on a genomic scale. Pyrosequencing to develop ESTs for Phalaenopsis was launched to generate 8,233 contigs and 34,630 singletons (Hsiao et al., 2011). Multiple sequencing techniques were integrated to generate 8,501 contigs and 76,116 singletons for Phalaenopsis spp. (Su et al., 2011; Fu et al., 2011). De novo transcriptome analysis of gene-related information associated with vegetative and reproductive growth of *C. sinense* was performed. The Illumina sequencing generated 54,248,006 high quality reads that were assembled into 83,580 unigenes with an average sequence length of 612 base pairs, including 13,315 clusters and 70,265 singletons (Zhang et al., 2013). These contributions, together with recent knowledge on floral developmental control genes in orchids, enable an improved understanding of orchid evolution.

Being part of the complex network of regulatory genes driving the formation of flower organs, the MADS-box gene family is among one of the most studied gene families (Schwarz-Sommer et al., 1990). Aceto and Gaudio, (2011) proposed that the diversification of the orchid perianth was a consequence of duplication events and changes in the regulatory regions of the MADS-box genes, followed by sub- and neo-functionalization. This specific developmental-genetic code is termed the "orchid code" (Aceto and Gaudio, 2011).

With recent advances in sequencing technologies, genome-scale sequencing projects such as de novo transcriptome analysis, reference mapping of expressed transcripts, and high-throughput technologies were launched in many emerging model organisms. In our research, two sequencing libraries prepared from petals and lips were sequenced using short reads sequencing technology (Illumina) to investigate differential expression genes (DEGs) in different flower organs. By deep sequencing analysis, 2,389 differentially expressed genes were obtained. Further functional annotation of DEGs provided a comprehensive understanding of the transcriptome complexity of floral structures and organ identity. The results let us know more details about the relationship between MADS gene family and floral morphology. This information broadens our understanding of the mechanisms of floral patterning and contributes to molecular and genetic research by enriching the Phalaenopsis database.

Result

Illumina sequencing and sequence assembly

Two cDNA libraries were generated with mRNA from two lateral inner tepals (petals) and a highly modified median inner tepal and then sequenced by short reads sequencing technology (Illumina). We identified the lateral inner tepals (petals) as Amabilis_Stem-red, and the lip as Amabilis_red. After cleaning and quality checks, 8,889,080 and 16,224,038 clean reads were generated from petals and lip cDNA libraries, respectively (Table 1). De novo assembly was carried out by Trinity software. The sequence reads were finally assembled into 60,933 non-redundant unigenes. Overviews of the assembly results were shown in Table 2. The assembly produced a substantial number of large unigenes: 12,046 (19.77%) unigenes were >1,000 bp in length and 21,224 unigenes were >500 bp, although most contigs were between 200 and 300 bp in length. All unigenes were longer than 200 bp. Mean length of final unigenes and N50 were 663.89 bp and 1168 bp, respectively. The unigene length distributions were shown (Supplementary File 1, Fig S1 and Fig S2).

Prediction of ORF, SSR markers and SNP

The open reading frame of unigenes was predicted using Getorf software. In general, when there are multiple open reading frame sequences, the longest reading frame of the coding sequence as the region of this sequence will be identified (Supplementary file 2, Table S1). Then the SSR markers were identified for the unigenes that were larger than 1Kb. A set of 12,046 unigenes sequences searched for SSRs. The total size of examined sequences was 22264058 bp and 3706 SSRs were identified. A 2974 unigene sequences with a repeat motif length ranging from one to six nucleotides was obtained using SSIT. In total, 1987 Mono-nucleotide SSRs, 856 di-nucleotide SSRs and 834 Tri-nucleotide SSRs were identified. In total, 570 sequences contained more than 1 SSR, and 259 SSRs presented in compound formation (Table 3; Supplementary File 2, Table S2). Analysis of these SSR motifs revealed that the proportion of SSR unit sizes was not evenly distributed.

Sequencing data between samples compared with the Unigene database (soap), using soapsnp build consensus sequence, and then analyzed between samples to get homozygous SNP loci. Finally after screening, SNPs with scoring 30 or more by soapsnp, and depth between 10x to 100x were obtained. Within the high coverage dataset, 15,149 putative SNPs from 1510 unigenes sequences recognized that may be used for population genetic analyses of *Phalaenopsis* (Supplementary File 2, Table S3).

Expression analysis and identification of differently expressed genes

Comparing sequencing data of different samples in the database with the unigene gene sequences, gene expression abundance analysis was done according to the ratio of the number on the difference reads for different aspects. The results could be used to evaluate the quality of the assembly and sequencing results. Sample statistics of the efficiency ratio were shown in Table 4.

Gene expression level was calculated using RPKM (Reads per Kb per million reads). The RPKM method can eliminate the amount of genetic differences in the length and calculate the gene expression. The calculated amount of gene expression can be directly used to compare differences in gene expression between samples. The IDEG6 analysis can find differentially expressed genes according to expression abundance of gene expression among the different samples. Comparing the transcriptomes from petals and lip, the IDEG6 analysis predicted 2,389 differentially expressed genes (false discovery rate≤0.001 and |log2Ratio|≥1) including 1,256 up-regulated unigenes and 1,133 unigenes down-regulated in lips (Fig 1). Hierarchical cluster analysis was done for screening of differentially expressed genes to cluster genes which had the same or similar conducted expression (Fig 2; Supplementary File 1, Fig S3).

Functional classifications of DEGs

Functions of 2,389 DEGs were annotated based on sequence similarities to sequences in the seven public databases (NT, NR, SwissProt, TrEMBL, KEGG, COG and GO, Supplementary File 2, Table S4). The statistics of annotation results were showed in Table 5.

To determine the possible functions of DEGs tagged, we used the Gene Ontology (GO) cataloging system for plants (Fig 3). The functions of identified genes covered three main categories (cellular components, molecular functions and biological processes) and distributed into 55 categories, including the most dominant such as, cell, cell part, catalytic activity, organelle, cellular binding, metabolic processes, and response to stimulus.

Table 1. Summary of transcriptome sequencing and assembly results.

Sample ID	Reads number	Base number	GC%	N%	Q20%	CycleQ20%
Amablilis-red	8,889,080	1,767,091,925	46.53	0.07	94.61	100.00
Amabilis stem-red	16,224,038	3,276,861,889	45.08	0.00	99.84	100.00

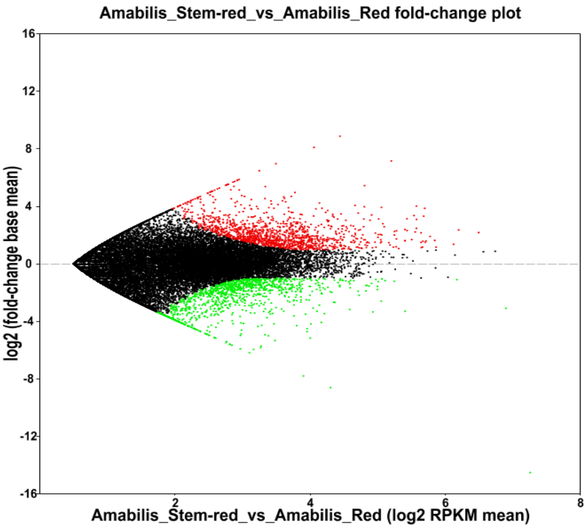

Fig 1. Identification of DEGs between tepals and lips. DEGs were determined using a threshold of FDR≤0.001 and |log2Ratio|≥1. Red spots represent up-regulated DEGs and green spots indicate down-regulated DEGs. Those shown in black are unigenes that did not show obvious changes.

Table 2. Statistics of the assembly with the Trinity method.

	Total number (percentage)		
Length range	Amabilis_Red unigene	Amabilis_Stem-red unigene	All unigenes
200-300	11,699(31.46%)	10,482(30.81%)	24,046(39.46%)
300-500	9,189(24.71%)	7,554(22.20%)	15,663(25.71%)
500-1000	7,801(20.98%)	6,038(17.75%)	9,178(15.06%)
1000-2000	6,317(16.99%)	7,014(20.62%)	8,266(13.57%)
2000+	2,183(5.87%)	2,932(8.62%)	3,780(6.20%)
Total number	37,189	34,020	60,933
Total length	26,762,841	27,626,110	40,453,025
N50 length	1,143	1,398	1,168
Mean length	719.64	812.35	663.89

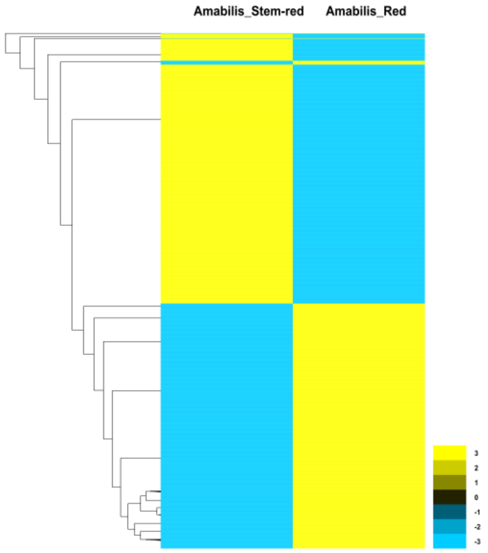

Fig 2. Hierarchical cluster analysis of differentially expressed genes. Expression differences are shown in different colors. Yellow means high expression and blue means low expression. Legends must be more informative, explaining the critical point of each Fig.

The mostly represented biological processes were "response to salt stress" (1.07%), "response to cadmium ion" (1.04%), "response to cold" (0.75%). Among the molecular functions, the most represented were "protein binding" (8.83%), "ATP binding" (3.94%), "DNA binding" (2.68%). In the category of Cellular Component, the most enriched groups were "nucleus" (7.58%), "plasma membrane" (6.79%), "cytosol" (5.81%), "mitochondrion" (4.93%), "chloroplast" (4.07%).

To further evaluate the integrity of our transcriptome library and the effectiveness of our annotation process, unigene sequences were subjected to Clusters of Orthologous Groups (COG) classification. Among the 25 COG categories, the cluster for "general function prediction only" (259DEGs,) was the largest group, followed by "transcription (124, DEGs)", "Posttranslational modification, protein turnover, chaperones" (122, DEGs), "replication, recombination and repair" (104, DEGs), "transcription, ribosomal structure and biogenesis" (99, DEG),"signal transduction mechanisms" (87, DEGs), and "carbohydrate transport and metabolism" (77, DEGs). There were no corresponding genes to the categories "nuclear structure" and "extracellular structures" (Fig 4).

In our study, 2,389 annotated sequences were mapped to the reference canonical pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG). In total, 734 DEGs were assigned to 101 KEGG pathways (Table 7; Supplementary File 2, Table S5). The pathways with most representation by the unigenes were "Ribosome" (37 DEGs), "Ubiquitin mediated proteolysis" (32 DEGs), and "Oxidative phosphorylation" (29 DEGs), "Spliceosome" (24 DEGs), "RNA transport" (23DEGs), "Protein processing in endoplasmic reticulum" (22 DEGs), "Plant hormone signal transduction "(22 DEGs), "Plant-pathogen interaction" (21 DEGs).

Discussion

Sequence annotation and DEGs analysis

In addition to the ecological significance, *P. amabilis* has been considered as an economically important floriculture industry worldwide. However, little is known about the mechanisms responsible for floral development and genomic information. The aims of this project were to identify the genes controlling floral structures and organ identity and generate a large amount of transcriptome data that would facilitate more detailed studies for *Phalaenopsis.* The availability of transcriptome data for *Phalaenopsis* would meet the initial information required for functional studies of this species and its relatives. In this study, RNA-seq was performed using Illumina sequencing, which generated total reads of 8,889,080 and 16,224,038, respectively. By comparing the transcriptomes from petals and lip, we finally obtained 2,389 differentially expressed genes. All the DEGs unigenes were used for BLASTX and annotation against protein databases like nr, SwissPort, COG, KEGG and GO. These genes were significantly enriched in 101 KEGG pathways and 55 GO terms.

To better understand the information related to flowering, we analyzed the different expression genes between tepals and lips. Among the 2,397DEGs, 1,257 DEGs were up-regulated while 1,134 were down-regulated in lips. In comparison, 36 DEGs only expressed in lips, and 159 DEGs only expressed in tepals. As these genes showed specific expression in the two phases, they were likely to involve in floral development and floral structure. Some genes have no significant similarities to any other protein, indicating that the short size has a negative effect on successful annotation. However, among the angiosperms, the orchids are unique in their floral patterning, particularly in floral structures and organ identity. This suggests that these genes may perform specific roles in orchids and be quite divergent from those of other plant species. These could be new genes that related to the structure of flower.

Floral meristem identity genes

The acronym MADS box is derived from the initials of four loci, MCMI of *Saccharomyces cerevisiae*, AG of *Arabidopsis thaliana*, DEF of *Antirrhinum majus* and SRF of *Homo sapiens*, all of which contain the MADS-box domain, a conserved 56-amino-acid DNA-binding domain (Schwarz-Sommer et al., 1990). The MADS box genes termed the "orchid code" are part of the complex network of regulatory genes driving the formation of flower organs. In our transcriptome database, we identified the MADS-box transcription factor Suppressor of Overexpression of CO 1. Eight *SOC1* genes were identified in our transcriptome database (Table 6). Three *SOC1* showed up-regulation in lips, while five showed down-regulation. The expression of *SOC1* is regulated by the light pathway, autonomous pathway, vernalization and gibberellin pathway. The flowering integration genes accept the signal from the genetic pathway, and then induce floral meristem identity (FMI) genes for flowering as a whole (Komeda, 2004). The floral meristem is initiated by a set of *FMI* genes that include *LFY*, *AP1*, *CAL*, *AP2*, and Unusual Floral Organs (UFO) (Yanofsky et al., 1995). The *OMADS1* gene belongs to the AP1/AGL9 group (Hsu et al., 2003). Previous studies believed that the expression pattern of *OMADS1* in the mature flower is restricted to the lip and carpel (Ma et al., 1991; Mena et al., 1995), and represent a class of MADS-box genes which are similar to that of the carpel-specific MADS-box genes in regulating floral initiation and ovary development in orchids. However, in our study, *OMADS1* gene was observed both in lips and petals and their expression levels showed great differences. It was strongly expressed in lips, and at very lower levels in petals. The expression in petals might be an error, as the short size might have a negative effect on successful annotation. AP2 showed different expression levels between lips and petals in our transcriptome database. These information of flowering integration genes and *FMI* genes would facilitate more detailed studies on the mechanism of floral differentiation for Orchidaceae.

In *Phalaenopsis equestris*, the four class B genes, PeMADS2-5, are AP3/DEF-like paralogs that are expressed during developmental stages ranging from early to late inflorescence (Tsai et al., 2004). Their organ-specific expression pattern demonstrates an absence of functional redundancy (Aceto and Gaudio, 2011). In fact, PeMADS2 was strongly expressed in the inner tepals, and at a lower level in lips. The PeMADS5 was expressed both in the inner tepals and lips. All These results were in agreement with the study of Aceto and Gaudio (2011). But in their research, PeMADS4 was only expressed in the lips and the column. In our research, the PeMADS4 was observed both in lips and the inner tepals, but they showed a great differential expression level in the two organs. The expression pattern of

Table 3. Statistics of the SSR markers with Getorf software.

Total number of sequences examined	12046
Total size of examined sequences (bp)	22264058
Total number of identified SSRs	3706
Number of SSR containing sequences	2974
Number of sequences containing more than 1 SSR	570
Number of SSRs present in compound formation	259

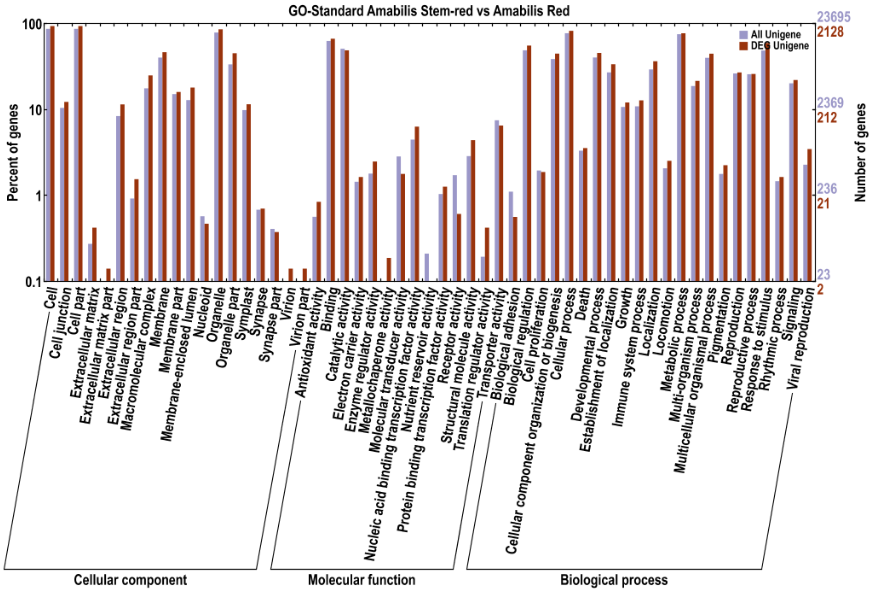

Fig 3. Functional classifications of GO terms DEGs. Unigenes with best BLAST hits were aligned to GO database. All unigenes were grouped into three main GO categories and 55 sub-categories. Right Y-axis represents number of genes in a category. Left Y-axis indicates percentage of a specific category of genes in each main category.

Table 4. Sample statistics of the efficiency ratio.

ID	Total Reads	Mapped Reads	Perfect Mapped Reads	Perfect Mapped Reads/ Total Reads
Amabilis_Red	8889080	6774730	4997244	73.76%
Amabilis_Stem-red	16224038	13800639	12016202	87.06%

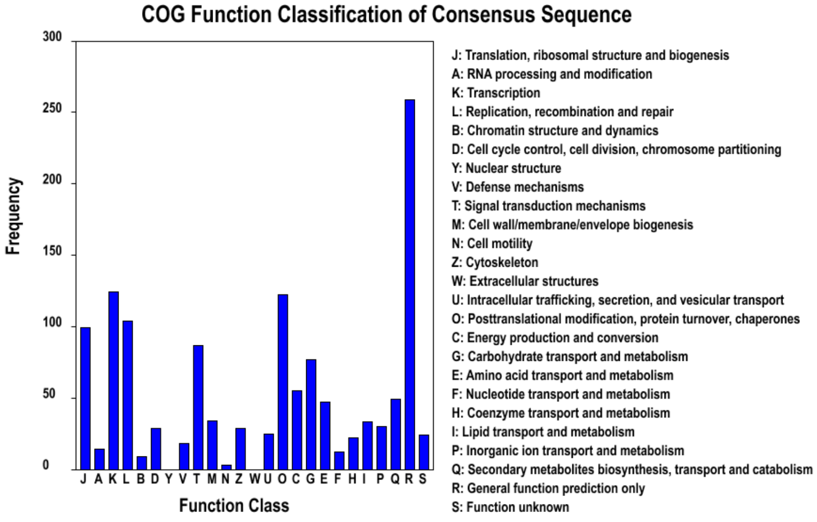

Fig 4. Functional classifications of COG terms DEGs.

Table 5. Statistics of annotation results for Phalaenopsis unigenes.

Type	Nr	Nt	SwissProt	TrEMBL	GO	KEGG	COG
Amabilis_Stem-red_vs_Amabilis_Red	2,356	1,895	2,018	2,363	2,129	678	940

Table 6. DEGs that share homology with flowering time genes

ID	Amabilis_Stem-red	Amabilis_Red	log2 (A/B)	Nt annotation
Amabilis_Stm-red_Unigene_BMK.31063	268	18	3.896	Oncidium Gower Ramsey MADS box transcription factor 1 (MADS1) mRNA, complete cds
Amabilis_Yellow_Unigene_BMK.24397	44	6	2.8748	Phalaenopsis equestris MADS box transcription factor (MADS5) gene, complete cds
Amabilis_Red_Unigene_BMK.3487	49	7	2.807	Phalaenopsis equestris MADS box transcription factor (MADS4) mRNA, complete cds
Amabilis_Stem-red_Unigene_BMK.18578	152	24	2.663	Cymbidium ensifolium SEP-like MADS-box protein mRNA, complete cds
Amabilis_Yellow_Unigene_BMK.20731	80	319	-1.995	Phalaenopsis equestris MADS box transcription factor (MADS2) mRNA, complete cds
Amabilis_Yellow_Unigene_BMK.28767	20	2	3.322	Phalaenopsis equestris MADS box protein (SOC1) gene, complete cds
Amabilis_Stem-red_Unigene_BMK.24227	1557	312	2.319	Phalaenopsis equestris MADS box protein (SOC1) gene, complete cds
Amabilis_Stem-red_Unigene_BMK.24246	830	259	1.680	Phalaenopsis equestris MADS box protein (SOC1) gene, complete cds
Amabilis_Red_Unigene_BMK.29711	32	101	-1.658	Phalaenopsis equestris MADS box protein (SOC1) gene, complete cds
Amabilis_Red_Unigene_BMK.29618	63	258	-2.034	Phalaenopsis equestris MADS box protein (SOC1) gene, complete cds
Amabilis_Red_Unigene_BMK.29514	3	58	-4.273	Phalaenopsis equestris MADS box protein (SOC1) gene, complete cds
Amabilis_Yellow_Unigene_BMK.17719	0	11	-13.425	Phalaenopsis equestris MADS box protein (SOC1) gene, complete cds
Amabilis_Red_Unigene_BMK.29714	0	23	-14.489	Phalaenopsis equestris MADS box protein (SOC1) gene, complete cds
Amabilis_Red_Unigene_BMK.9167	32	85	-1.409	AGL6[Cymbidium goeringii] Cymbidium ensifolium SEP-like MADS-box protein mRNA, complete cds

Table 7. DEGs associated with flower color.

ID	Amabilis_Stem-red	Amabilis_Red	log2 (A/B)	Nt annotation
Amabilis_Stem-red_Unigene_BMK.24974	30	0	14.872	Onobrychis viciifolia chalcone synthase (CHS) mRNA, complete cds
Amabilis_Yellow_Unigene_BMK.29304	209	595	-1.509	Phalaenopsis hybrid cultivar chalcone synthase (CHS5) gene, complete cds
Amabilis_Red_Unigene_BMK.27970	126	572	-2.183	Phalaenopsis equestris UFGT3 (UFGT3) gene, complete cds

these AP3/DEF-like genes reveals that PeMADS2, PeMADS4 and PeMADS5 are involved in specifying the development of the outer tepals, lip and inner tepals, respectively. Our research somehow supports the theory of Aceto and Gaudio, (2011) in transcription aspects, except the fact that PeMADS4 was observed in the inner tepals.

The AGAMOUS LIKE 6 lineage of MIKC-type MADS-box transcription factors is rooted in a superclade with both SEPALLATA-like genes and APETALA1/FRUIT-FUL-like genes (Tsai et al., 2004; Theissen and Melzer, 2007). So far, researches suggested that AGL6 plays a redundant role in establishing the flower and its organs (Viaene et al., 2010). In our study, AGL6 was among the 2389 DEGs. It showed a lower expression level in lips.

DEGs associated with flower color

Flower colour as an important trait, is mainly determined by anthocyanins. The pathway of anthocyanin biosynthesis is usually divided into two sections, the early and the late sections (Deroles, 2009; Niu et al., 2011). Chalcone synthase (CHS) is among the early sections formations, and UDPGlucose: flavonoid 3-O- glucosyltranferase (UFGT) is among the late section formations. In our DEGs transcriptome database, chalcone synthase (CHS) was only found in lips while CHS5 could be observed both in lips and petals (Table 7). As the color of lips is darker than petals, the more expression of CHS may contribute to this phenomenon. It has been reported UFGT played a predominant and positive-regulated role in the anthocyanin accumulation in litchi. However, in our database, UFGT3 was down-regulated in lips.

Recently, studies indicate that expression of biosynthetic genes in anthocyanin accumulation is regulated by MYB transcription factor in the fruit of grapes (Kobayashi et al., 2002), apples (Takos et al., 2006; Espley et al., 2007), mangosteen (Palapol et al., 2009), Chinese bayberries (Palapol et al., 2009) and red pear (Zhang et al., 2011). We generated Myb family transcription factor from both libraries. Most of them were down-regulated in lips (Supplementary File 2, Table S5). All the results suggest more researches should be done on the metabolism of anthocyanin biosynthesis.

Materials and Methods

Plant material

The Red Phalaenopsis Dtps. Jiuhbao Red Rose (Taisuco Firebird × King Shiangs Rose) was used as plant materials. It was grown in greenhouses at Nanjing Agricultural University under a 14 h photoperiod at 25°C in the daytime and 20°C at night. When the flower was fully bloomed, we collected 0.1g samples from lip and petals, respectively. Three replicates were performed for three independent plants.

cDNA library construction and sequencing

Total RNA was extracted with EASYspin plant RNA rapid extraction kit according to the manufacturer's protocol (Yuan Ping Hao Bio, Beijing, China). Beads with Oligo (dT) were used to enrich the eukaryotic mRNA. Added fragmentation buffer (Ambion, Austin, TX,USA) was used to break mRNA into short fragments ken into short segments, using random hexamers (random hexamer primers) synthesis of the first-strand cDNA, then buffer, dNTPs, RNaseH and DNA polymerase I (Invitrogen, Grand Island, NY, USA) synthesis were added to make the second strand cDNA. Short fragments were purified with a QIAquick PCR extraction kit (Qiagen, Hilden, Germany) and resolved with EB buffer for end-repair. The poly (A) was added and linked to sequencing adapters. Agarose gel electrophoresis and PCR amplification were used to select suitable fragments. Two cDNA libraries were constructed and sequenced on the Illumina HiSeq^{TM}2000 platform.

De novo assembly and predict ORF, SSR, SNP

After filtering the raw reads, De novo assembly of the transcriptome was carried out with a short reads assembling program–Trinity (Grabherr et al., 2011). Trinity connects the contigs and obtains sequences defined as unigenes.

The open reading frame of unigenes was predicted using Getorf software (Mortazavi et al., 2008). In general, for a sequence with multiple open reading frame, the longest reading frame of the coding sequence as the region of sequence are identified. The MISA software (Kanehisa et al., 2008) was used to identify SSR markers for the unigenes that were larger than 1Kb. Sequencing data between samples comptared with the Unigene database, using Soapsnp (Conesa et al., 2005), which builds consensus sequence, and then analyzes samples to get homozygous SNP loci.

Expression analysis and identification of differentially expressed genes (DEGs)

The sequencing reads were compared with the Unigene database to obtain the expression abundance information, using RPKM (Mortazavi et al., 2008) to reflect the value of the expression of the corresponding Unigenes abundance. For each gene analysis, the statistical information including length, depth, reads coverage (coverage x), RPKM (expression abundance), the total compared number of reads, Unique (unique) compared to the number of reads, multi (multi-position) compared to the number of reads. The differentially expressed genes were found based on expression abundance of gene among the different samples. Comparing the transcriptomes from petals and lip, the IDEG6 (Romualdi et al., 2003) analysis predicted 2,389 differentially expressed genes including 1256 up-regulated unigenes and 1133 unigenes down-regulated in lips.

Functional categorization

The generated unigenes were used for BLASTX (Altschul et al., 1997) and annotation against protein databases, including non-redundant (nr), NT, SwissPort, TrEMBL, GO (Ashburner et al., 2000), COG (Tatusov et al., 2000) and KEGG (Kanehisa et al., 2004). GO (http://www.geneontology.org) has three ontologies: molecular function, cellular component and biological process. To get general gene ontology (GO) annotations for all unigenes were aligned to three public databases (NR, Swiss-Prot and KEGG) by BLASTX with E-value < =1e-5. The GO annotations for the top blast hits were retrieved with Blast2GO program. GO functional classification was performed by WEGO website tool. The GO database can be applicable to all species, capable of limiting and description of genes or proteins. COG database is based on bacteria, algae, eukaryotic phylogenetic relationships. The COG database can accomplish the classification of orthologous gene product. KEGG is a major public pathway-related database that is able to analyze a gene product during a metabolic process and related gene function in cellular processes. With the help of the KEGG database, we can further study genes' biological complex behaviors. By KEGG annotation we can obtain pathway annotation for unigenes. The KEGG pathways annotation was performed using BlastAll software against the KEGG database.

Conclusion

Although the molecular functions of *Phalaenopsis* genes and the associated floral genetic pathways remained unknown, the combination of RNA-seq and DGE analysis based on Illumina sequencing technology provided comprehensive information on gene expression, which could facilitate further investigations of the detailed floral development mechanisms of this culturally important orchid. The candidate genes on floral meristem identity genes, such as MADS-box genes, were identified by this approach, which could let us know more details about the relationship between MADS gene family and floral morphology. This data could be used as a tool to investigate the flowering pathway and various other biological pathways in *Phalaenopsis*.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (31101574), Youth Science and Technology Innovation Fund of NJAU (KJ2011008), and Agricultural Science and Technology Support Program of Zhenjiang (NY2011015).

References

Aceto S, Gaudio L (2011) The MADS and the beauty: genes involved in the development of orchid flowers. Curr Genomics. 12:342.

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

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT (2000) Gene Ontology: tool for the unification of biology. Nat Genet. 25:25-29.

Blanchard MG, Runkle ES (2006) Temperature during the day, but not during the night, controls flowering of Phalaenopsis orchids. J Expe Bot. 57,4043-4049.

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

Cozzolino S, Widmer A (2005) Orchid diversity: an evolutionary consequence of deception? Trends Ecol Evol. 20:487-494.

Deroles S (2009) Anthocyanin biosynthesis in plant cell cultures: A potential source of natural colourants. In: Anthocyanins, Springer New York. 108-167.

Espley RV, Hellens RP, Putterill J, Stevenson DE, Kutty-Amma S, Allan AC (2007) Red colouration in apple fruit is due to the activity of the MYB transcription factor, MdMYB10. Plant J. 49:414-427.

Fu CH, Chen YW, Hsiao YY, Pan ZJ, Liu ZJ, Huang YM, Tsai WC, Chen HH (2011) OrchidBase: a collection of sequences of the transcriptome derived from orchids. Plant Cell Physiol. 52(2):238-243.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29:644-652.

Hsiao YY, Chen YW, Huang SC, Pan ZJ, Fu CH, Chen WH, Tsai WC, Chen HH (2011) Gene discovery using next-generation pyrosequencing to develop ESTs for Phalaenopsis orchids. BMC Genomics. 12,360.

Hsu CC, Chung YL, Chen TC, Lee YL, Kuo YT, Tsai WC, Hsiao YY, Chen YW, Wu WL, Chen HH (2011) An overview of the Phalaenopsis orchid genome through BAC end sequence analysis. BMC Plant Biol. 11:3.

Hsu HF, Huang CH, Chou LT, Yang CH (2003) Ectopic expression of an orchid (Oncidium Gower Ramsey) AGL6-like gene promotes flowering by activating flowering time genes in Arabidopsis thaliana. Plant Cell Physiol. 44:783-794.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res. 32:D277-D280.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36:D480-D484.

Kobayashi S, Ishimaru M, Hiraoka K, Honda C (2002) Myb-related genes of the Kyoho grape (Vitis labruscana) regulate anthocyanin biosynthesis. Planta 215:924-933.

Komeda Y (2004). Genetic regulation of time to flower in Arabidopsis thaliana. Annu Rev Plant Biol. 55:521-535.

Leitch I, Kahandawala I, Suda J, Hanson L, Ingrouille MJ, Chase M, Fay M (2009) Genome size diversity in orchids: consequences and evolution. Ann Bot-London. 104:469-481.

Ma H, Yanofsky MF, Meyerowitz EM (1991) AGL1-AGL6, an Arabidopsis gene family with similarity to floral homeotic and transcription factor genes. Gene Dev 5:484-495.

Mena M, Mandel MA, Lerner DR, Yanofsky MF, Schmidt RJ (1995) A characterization of the MADS-box gene family in maize. Plant J. 8:845-854.

Mondragón-Palomino M, Theißen G (2008) MADS about the evolution of orchid flowers. Trends Plant Sci. 13:51-59.

Mondragón-Palomino M, Theißen G (2009) Why are orchid flowers so diverse? Reduction of evolutionary constraints by paralogues of class B floral homeotic genes. Ann Bot-London. 104:583-594.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 5:621-628.

Niu SS, Xu CJ, Zhang WS, Zhang B, Li X, Lin-Wang K, Ferguson IB, Allan AC, Chen KS (2010) Coordinated regulation of anthocyanin biosynthesis in Chinese bayberry (Myrica rubra) fruit by a R2R3 MYB transcription factor. Planta. 231:887-899.

Palapol Y, Ketsa S, Lin-Wang K, Ferguson IB, Allan AC (2009) A MYB transcription factor regulates anthocyanin biosynthesis in mangosteen (*Garcinia mangostana* L.) fruit during ripening. Planta. 229:1323-1334.

Romualdi C, Bortoluzzi S, d'Alessi F, Danieli GA (2003) IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. Physiol Genomics. 12:159-162.

Rudall PJ, Bateman RM (2002) Roles of synorganisation, zygomorphy and heterotopy in floral evolution: the gynostemium and labellum of orchids and other lilioid monocots. Biol Revi Camb Philos. 77:403-441.

Schwarz-Sommer Z, Huijser P, Nacken W, Saedler H, Sommer H (1990) Genetic control of flower development by homeotic genes in *Antirrhinum majus*. Science. 250:931-936.

Semiarti E, Indrianto A, Purwantoro A, Isminingsih S, Suseno N, Ishikawa T, Yoshioka Y, Machida Y, Machida C (2007) Agrobacterium-mediated transformation of the wild orchid species Phalaenopsis amabilis. Plant Biotechnol. 24,265-272.

Su Cl, Chao YT, Chang YCA, Chen WC, Chen CY, Lee AY, Hwa KT, Shih MC (2011) De novo assembly of expressed transcripts and global analysis of the *Phalaenopsis aphrodite* transcriptome. Plant Cell Physiol. 52:1501-1514.

Takos AM, Jaffé FW, Jacob SR, Bogs J, Robinson SP, Walker AR (2006) Light-induced expression of a MYB gene regulates anthocyanin biosynthesis in red apples. Plant Physiol. 142:1216-1232.

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33-36.

Theissen G, Melzer R (2007) Molecular mechanisms underlying origin and diversification of the angiosperm flower. Ann Bot-London. 100:603-619.

Tsai WC, Kuoh CS, Chuang MH, Chen WH, Chen HH (2004) Four DEF-like MADS box genes displayed distinct floral morphogenetic roles in Phalaenopsis orchid. Plant Cell Physiol. 45:831-844.

Viaene T, Vekemans D, Becker A, Melzer S, Geuten K (2010) Expression divergence of the AGL6 MADS domain transcription factor lineage after a core eudicot duplication suggests functional diversification. BMC Plant Biol. 10: 148.

Xu Y, Teo LL, Zhou J, Kumar PP, Yu H (2006) Floral organ identity genes in the orchid *Dendrobium crumenatum*. Plant J. 46,54-68.

Yanofsky MF. (1995). Floral meristems to floral organs: genes controlling early events in Arabidopsis flower development. Ann Rev Plant Biol. 46:167-188.

Zhang J, Wu K, Zeng S, da Silva JAT, Zhao X, Tian CE, Xia H, Duan J (2013) Transcriptome analysis of Cymbidium sinense and its application to the identification of genes associated with floral development. BMC Genomics. 14(1):279.

Zhang X, Allan AC, Yi Q, Chen L, Li K, Shu Q, Su J (2011) Differential gene expression analysis of Yunnan red pear, *Pyrus pyrifolia*, during fruit skin coloration. Plant Mol Biol Rep. 29:305-314.