Transcriptome analysis of flower color reveals the correlation between SNP and differential expression genes in *Phalaenopsis*

Yu Ding

Key Laboratory of Ministry of Education for Genetics and Germplasm Innovation of Tropical Special Trees and Ornamental Plants, College of Forestry, Hainan University

Ding-Hai Yang

Key Laboratory of Ministry of Education for Genetics and Germplasm Innovation of Tropical Special Trees and Ornamental Plants, College of Forestry, Hainan University

Ma-Yin Wang

Key Laboratory of Ministry of Education for Genetics and Germplasm Innovation of Tropical Special Trees and Ornamental Plants, College of Forestry, Hainan University

Dai-Cheng Hao

Hainan Boda Orchid Technology Co. Ltd.

Wei-Shi Li

Hainan Boda Orchid Technology Co. Ltd.

Peng Ling

Key Laboratory of Ministry of Education for Genetics and Germplasm Innovation of Tropical Special Trees and Ornamental Plants, College of Forestry, Hainan University

Shang-Qian Xie (✉ sqianxie@foxmail.com)

Key Laboratory of Ministry of Education for Genetics and Germplasm Innovation of Tropical Special Trees and Ornamental Plants, College of Forestry, Hainan University

Research Article

Keywords: Phalaenopsis, flower color, SNP, differential expression gene

Posted Date: March 2nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-209376/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background *Phalaenopsis* is an important ornamental plant, which occupies an important position in the world flower market and has great economic value due to its rich and diverse flower colors. In order to investigate the flower color formation of *Phalaenopsis* at transcription level, the flower color formation involved genes were identified from RNA-seq in this study.

Results White and purple petals of *Phalaenopsis* were collected in this study, and results were focused on two aspects: (1) the differential expression genes (DEGs) between white and purple flower color; and (2) association between SNP mutations and DEGs in transcriptome level. Results indicated that a total of 1,175 DEGs were identified, and the up- and down-regulation genes were 718 and 457, respectively. Gene Ontology (GO) and pathway enrichment showed that the biosynthesis of secondary metabolites pathway was key responsible for color formation and twelve crucial genes (C4H, CCoAOMT, F3'H, UA3'5'GT, PAL, 4CL, CCR, CAD, CALDH, bglx, SGTase and E1.11.17) from them involved in the regulation of flower color in *Phalaenopsis*.

Conclusion This study firstly reported that the SNP mutations strongly associated with DEGs in color formation at RNA level, and provides a new insight to further investigate the gene expression and its relationship with genetic variants from RNA-seq data in other species.

Background

Phalaenopsis is an important ornamental plant with abundant colors in petals, and lips such as white, yellow, red, purple, etc. It is also an important economical horticulture plant and has been widely sold as potted or cut flowers in the commercial trade market [1]. The production of *Phalaenopsis* is continuously increasing due to its wide popularity in China, United States and many European countries [2]. The commercial values increase with the improvement of *Phalaenopsis* cultivation, such as brightness of the color, multiple color forms and different color patterns etc. The global consumption of *Phalaenopsis* has been increasing recent years [3], and the ornamental and commercial values of *Phalaenopsis* are mainly determined by the quality and variety of the flower colors [4].

Previous studies on the *Phalaenopsis* flower color were focused on the breeding of rare or colorful varieties, and the mechanism of the color formation [5, 6]. *Phalaenopsis* is an ideal material for studying the flower coloring mechanism because of its color diversity. Tanaka et al. firstly reported that the diversity of anthocyanin was the primary cause of the colorful flowers [7]. Anthocyanin was originated from malonyl-coenzyme A and coumaryl-coenzyme A catalyzed by a series of enzymes in cytoplasm, then it was modified by different glycosylation, methylation and acylation and generated stable anthocyanin. With the assistance of transporters or transporters vesicles, anthocyanin entered the vacuoles and accumulated to emerge various colors from white to purple [1, 7, 8]. The most related enzymes in the anthocyanin synthesis pathway were encoded by genes that existed in the form of multiple gene families [9]. Among them, four key enzyme genes included Chalcone synthase (CHS),

Dihydroflavonoid reductase (DFR), Flavanone 3', 5' –hydroxylase (F3'5'H) and Anthocyanin synthetase (ANS) were highly associated with the anthocyanin synthesis in *Phalaenopsis*. Su et al. isolated flavonoid 3', 5'-hydroxylase gene from cytochrome P450 family and transferred into petal of *Phalaenopsis* safflower by gene marksmanship, which resulted petal color changed from red to purple within 48h [2]. These genes had high homology with color-related genes and also been reported in other ornamental plants [10]. In addition, Tatsuzawa et al. isolated cyanidin 3, 7, 3'- triglucoside acylated by acid from five red-purple *Phalaenopsis* varieties [11], and identified several anthocyanin components such as delphinidin, peonidin, petunidin, malvidin and pelargonidin [12].

The gene expressions related to flower color formation can be resolved by RNA-seq sequencing technology, which had been extensively used for decoding differential genes due to its low cost and high throughput of data. Chen et al. conducted RNA-seq sequencing of flowers and leaves from *Osmanthus serrulatus Rehd* and obtained 2,602 differentially expressed genes, and 33 of them involved in carotenoid biosynthesis in metabolic pathway[13]. Zhou et al. reported that the expression of carotenoid biosynthesis genes (PSY, CrtZ and BCH) and flavonoid biosynthesis genes (CHS, F3H, FLS and ANS) jointly caused the golden yellow of petals in *Camellia nitidissima* [14]. Wu et al. identified 127 unigene related to color synthesis and confirmed the candidate gene UA3GT in flavonoid metabolism pathway caused color formation of blue petals in *Nymphaea* 'King of Siam' [15].

The flower color formation mechanism of *Phalaenopsis* is an important research topic, but the most of current researches were focused on genetic engineering technology, genetic analysis and breeding selections in *Phalaenopsis*. At present, only Gao et al. investigated the regulation mechanism of flower color by using RNA-seq in transcriptome level and revealed the anthocyanin synthesis pathway was responsible for the differences between red and yellow flowers in *Phalaenopsis*[16]. Recently, RNA-seq had been used to analyze a large number of single nucleotide polymorphisms (SNPs) and associated with plant biological and agronomic traits [17]. However, there is a lack of research on the relationship between transcriptional level SNP and differential expressed genes of the flower color traits in *Phalaenopsis*. Therefore, we collected two contrasting flower colors (white and purple) in this study, and analyzed the differential expressed genes and SNP in transcriptome level. Furthermore, the contribution of SNP related to differential expression genes of flower color trait was decoded. It provides an insight on the mechanism of flower color formation, and enhances the understanding of theoretical basis for the breeding improvement of *Phalaenopsis* flower color and cultivation of new varieties.

Result

Unigene assembly and functional annotation

RNA-seq datasets of six samples from purple flowers (P1, P2 and P3) and white flowers (W1, W2 and W3) of *phalaenopsis* were sequenced on Illumina HiSeq platform, and generated 60,293,026, 48,567,124, 57,043,822, 53,586,804, 60,652,854, and 55,117,678 raw reads in P1, P2, P3, W1, W2 and W3, respectively (Table 1). The total raw data (335,261,308) was treated and removed joints, impurities and low-quality

reads, then 329,491,250 clean reads were retained. Sample P2 and W2 had the most (58,673,380) and least (48,567,124) number of clean data, respectively (Table 1). Moreover, the GC content of each sample was around 45%, and the nucleotide bases with Phred ($\text{Phred} = -10\log_{10}(e)$) values greater than 20 and 30 account for 97% and 93% of the total nucleotide bases (Table 1). These results showed that the quality of clean data obtained from sequencing was adequate enough for subsequent analysis.

Table 1
Sequencing row data of the six samples

Sample	Raw reads No.	Clean reads No.	Clean bases (G)	Q20 (%)	Q30 (%)	GC (%)
W1	60,293,026	58,137,588	8.72	97.44	93.5	45.47
W2	48,567,124	48,567,124	7.29	98.13	95.09	45.20
W3	57,043,822	57,043,822	8.56	98.11	95.03	45.39
P1	53,586,804	53,586,804	8.04	98.16	95.14	45.56
P2	60,652,854	58,673,380	8.80	97.48	93.62	44.89
P3	55,117,678	53,482,532	8.02	97.39	93.43	44.78
Total of 215,191 unigene sequences with total nucleotide base number of 54,354,436 bp were sequenced. Among them, the maximum and minimum length of the sequences were 8,698 bp and 201 bp, respectively (Table 2). And the average and N50 length of assembled sequences were 253 bp and 703 bp, respectively (Table 2). Moreover, the length of these unigene sequences was mainly distributed between 200 bp and 4,000 bp, which accounted for 98% of the sequences reads (Figure S1). Then we used Nr, Swissprot, GO, KEGG and COG databases to obtain the functional annotation of the Amino acid sequence of the unigene. The number of annotations obtained in the Nr, Swissprot, GO, KEGG, and COG databases were 21,143 (90.7%), 15,062 (64.6%), 10,827 (46.4%), 9,854 (42.3%) and 18,954 (82.7%), respectively (Table S1).						

Table 2
The statistical results of unigene sequences

Items	Assembled transcripts
Total sequence number	215,191
Total sequence base (bp)	54,354,436
GC%	43.84
Median contig length (bp)	139
Largest length (bp)	8,698
Smallest length (bp)	201
Average length (bp)	253
N50 length (bp)	703
Total Amino Acid Sequence number	23,314

COG database distribution

To further annotate homologous proteins of unigene, the 18,954 unigenes from COG database were further analyzed. About 12,436 (61.84%) unigene sequences were grouped into three categories included (1) Information Storage and Processing (3,947), (2) Cellular Processes and Signaling (4,432) and (3) Metabolism (4,057), which accounted for 19.63%, 22.04% and 20.17% of all sequences, respectively (Table 3 and Figure S2). Besides, 32.41% of the sequences were still functional unknown.

Identification of differentially expressed genes

The correlation between paired samples from six datasets was calculated by using the FPKM values of the transcripts. The three biological reduplications from purple and white flowers had a high degree of similarity (Fig. 2A, $R^2 > 0.95$). There were 1,175 differentially expressed genes (DEGs) were detected between purple and white samples, 718 genes of them were up-regulated and 457 genes were down-regulated (Fig. 2B and C, Table S2). And DEGs were annotated in Nr, Swiss-Prot, GO, KEGG, and COG databases, and the numbers were 1,163 (98.9%), 898 (76.4%), 474 (40.3%), 431 (36.7%), and 882 (75.1%), respectively (Fig. 2D and Table S3).

Table 3
COG Categories Distribution

Information Storage and Processing	
Transcription (K)	1,649
Replication, recombination and repair (L)	878
Translation, ribosomal structure and biogenesis (J)	699
RNA processing and modification (A)	555
Chromatin structure and dynamics (B)	166
Total	3,947 (19.63%)
Cellular Processes and Signaling	
Posttranslational modification, protein turnover, chaperones (O)	1,871
Signal transduction mechanisms (T)	1,393
Intracellular trafficking, secretion, and vesicular transport (U)	1,522
Cytoskeleton (Z)	188
Cell wall/membrane/envelope biogenesis (M)	175
Cell cycle control, cell division, chromosome partitioning (D)	139
Defense mechanisms (V)	137
Extracellular structures (W)	1
Nuclear structure (Y)	6
Cell motility (N)	0
Total	4,432 (22.04%)
Metabolism	
Carbohydrate transport and metabolism (G)	1,021
Energy production and conversion (C)	961
Secondary metabolites biosynthesis, transport and catabolism(Q)	526
Amino acid transport and metabolism (E)	512
Lipid transport and metabolism (I)	392
Inorganic ion transport and metabolism (P)	391
Coenzyme transport and metabolism (H)	157

Information Storage and Processing	
Nucleotide transport and metabolism (F)	97
Total	4,057 (20.17%)
Poorly Characterized	
Function unknown (S)	6,518
General function prediction only (R)	0
Total	6,518 (32.41%)

GO and KEGG pathway Enrichment analysis of DEGs

To further annotate the function of DEGs, GO enrichment analysis was performed on the DEGs. There were 1,247 genes involved in biological processes, 134 genes related with cell components and 314 genes involved in molecular functions were represented by DEGs (Fig. 3A and Table S4). The biological processes were significantly enriched in response to organonitrogen compound, in response to chitin, and in response to wounding and the secondary metabolite biosynthetic process etc (Fig. 3A). The enrichment of cell component included apoplast, peroxisome, microbody, plastoglobule and peroxisomal membrane etc (Fig. 3A). The significant enrichment of the molecular function mainly involved in calcium ion binding, CoA-ligase activity, ligase activity and forming carbon-sulfur bonds etc (Fig. 3A).

Furthermore, we used KOBAS v3.0 to analyze KEGG pathway enrichment from DEGs (Fig. 3B and Table S5). The results indicated that these differential genes were involved in 76 signaling pathways, which included Biosynthesis of secondary metabolites, Metabolic pathways, Phenylpropanoid biosynthesis, and Flavonoid biosynthesis and Carotenoid biosynthesis etc. Among them, the Biosynthesis of secondary metabolites pathway was the most significant (Fig. 3B).

Total of 4 DEGs (F3'H, C4H, CCoAOMT and UA3'5'GT) were identified and annotated in the Flavonoid biosynthesis metabolic pathway by KEGG, which had been reported to play an important role in flower color formation [18]. Among them, F3'H gene was up-regulation expression and C4H, CCoAOMT and UA3'5'GT were all down-regulation (Table 4 and Table S2). Besides, 10 DEGs were identified and annotated in the phenylpropanoid biosynthesis of KEGG pathway. And the expressions of CCoAOMT, C4H, PAL, 4CL, CCR, CALDH and Bglx were down-regulation except CAD, SGTase and E1.11.1.7 with up-regulation (Table 4 and Table S2). The phenylpropanoid biosynthesis was also an important bio-pathway in flower color formation [18].

Table 4
The annotation pathway of identified unigenes of *phalaenopsis*

Unigene id	Gene name	Definition	Pathway
LOC110024427	C4H	trans-cinnamate 4-monooxygenase	Flavonoid/ Phenylpropanoid biosynthesis
LOC110023518	CCoAOMT	caffeoyl-CoA O-methyltransferase	Flavonoid/ Phenylpropanoid biosynthesis
LOC110022396	F3'H	flavonoid 3'-monooxygenase	Flavonoid biosynthesis
LOC110030623	UA3'5'GT	anthocyanidin 5,3-O-glucosyltransferase	Flavonoid biosynthesis
LOC110031047	PAL	phenylalanine ammonia-lyase	Phenylpropanoid biosynthesis
LOC110038424 LOC110018262 LOC110026381	4CL	4-coumarate–CoA ligase	Phenylpropanoid biosynthesis
LOC110024447	CCR	cinnamoyl-CoA reductase	Phenylpropanoid biosynthesis
LOC110028720	CAD	cinnamyl-alcohol dehydrogenase	Phenylpropanoid biosynthesis
LOC110037950 LOC110019312	CALDH	coniferyl-aldehyde dehydrogenase	Phenylpropanoid biosynthesis
LOC110031840	SGTase	scopoletin glucosyltransferase	Phenylpropanoid biosynthesis
LOC110024203	bglx	beta-glucosidase	Phenylpropanoid biosynthesis
LOC110029660	E1.11.1.7	peroxidase	Phenylpropanoid biosynthesis

RNA SNP identification

Total of 207,759 SNPs were identified in transcripts (Table S6), and the transition (G-A and C-T) frequency was higher than transversion sites (Fig. 4A). The G->A, A->G, C->T, and T->C ratio of transition sites were 13.45%, 13.41% and 13.35%, 13.32%, respectively (Table S7), which was about 1.15 times of the transversion sites (T->A: 8.60%, A->T: 8.49%, A->C: 5.72%, G->T: 5.67%, C->A: 5.66%, T->G: 5.55%, G->C : 3.40% and C->G: 3.37%) (Table S7).

In addition, a total of 44,808 SNPs were identified in annotated genes (Table S8), among which 2,260 SNPs belonged to DEGs and 42,548 SNPs belonged to non-DEGs. And the average mutation frequencies of DEGs and non-DEGs were 0.0024 and 0.0029, respectively (Table S9 and Table S10). In 12 key genes,

the total number of SNPs was 20, and the average mutation frequency was 0.0015 (Table S11), which was lower than that in DEGs and non-DEGs.

To further study whether SNP mutations in transcripts were associated with flower color, the mutation location of genes were identified (Table S8) and the mutation frequency of DEGs and non-DEGs from the unigene sequences were compared. The results showed that the mutation frequency of SNPs between DEGs and non-DEGs was extremely significant ($P = 0.0021$) (Fig. 4B). In DEGs group, the mutation frequency of SNPs in up- and down-regulation genes were highly significant (Fig. 4C), which was benefit for pigmentation based on the foldchange value. Furthermore, the genes of each group were divided into four categories: A, C, T and G in base level. And the mutation frequencies in four categories were compared between the two groups (DEGs and non-DEGs). The SNPs mutations in A, C, T and G categories between DEGs and non-DEGs were all significant (A: $P = 0.0013$, C: $P = 0.0022$, T: $P = 0.039$, G: $P = 0.019$) (Figure S3A, B, C and D). All above results illustrated that the RNA SNP mutations was strongly associated with color formation in *Phalaenopsis*.

Discussion

Transcriptome refers to the transcription of all RNA at a specific developmental stage or physiological state. Understanding the transcriptome is necessary to explain the functional elements of genome and to understand the underlying mechanisms of biological growth and disease[19]. In this study, the petal color formation mechanism of *phalaenopsis* from purple flowers and white flowers were explored by RNA-seq in transcription level. We identified a total of 215,191 unigenes with annotated information, and 1,175 of them were differential expression genes. Among them, 718 DEGs were up-regulated and 457 DEGs were down-regulated. This study, first time reported the correlation between SNP and gene expression in color formation of *Phalaenopsis*, which clearly explained the flower color formation mechanism, and provided new insights for the close correlation between genetic variation and gene expression at transcription level.

Previous studies had found that the SNPs in DNA level related to flower color in chrysanthemum and cabbage, which were used as markers to assist breeding [20, 21]. In *Phalaenopsis hybrids*, DNA allele diversity of SNP had been used as an assistant marker to associate and predict the color of flowers [22]. However, it was rarely reported that the relationship between SNP and phenotype at the transcriptional level. RNA-seq, a powerful technique for gene expression at transcriptome level, can be used not only for differential expression gene identification, but also for genetic variation analysis [23]. In this study, 207,759 SNP sites were identified through the RNA-seq and the mutation frequencies of SNP associated with DEGs in genes were firstly investigated at nucleotide base level. And the results showed that the SNP mutations were significantly different between DEGs and non-DEGs, which indicated that SNPs from RNA-seq were strongly related to the change of flower color in *Phalaenopsis*.

Both GO enrichment and KEGG pathway analysis showed that DEGs were significantly enriched in biosynthesis of secondary metabolites pathway. And the Flavonoid biosynthesis and Phenylpropanoid

biosynthesis were two metabolic pathways related to the flower color formation [18]. In this study, we identified 12 genes (C4H, CCoAOMT, F3'H, UA3'5'GT, PAL, 4CL, CCR, CAD, CALDH, Bglx, SGTase and E1.11.17) that related flower color synthesis pathways. And these genes had been investigated and reported in previous studies [24–33]. C4H had been reported not only the key enzyme involved in the second step of flavonoid synthesis, but also is the first oxidoreductase of cytochrome P450 in phenylpropane biosynthesis pathway, which catalyzed a specific hydroxylation reaction and generated coumaric acid that was a precursor of flavonoids [24]. CCoAOMT contributed to the formation of Phenylpropylene, and the down-regulation of its expression would lead to the activation of biosynthesis of anthocyanins [25]. F3'H was a microsomal cytochrome P450-dependent monooxygenase which was the key enzyme to generated the hydroxylate B-ring in flavonoids that play an important role in flower coloring [26]. Yang et al. mentioned that the up-regulation and inhibition of F3'H gene were closely related to anthocyanin accumulation in *torenia hybrid*, and the over-expression of F3'H increased the accumulation of cyanidin and resulted in red sunflower petals [34]. Anthocyanin 3', 5'- O-glucosyltransferase (UA3'5'GT), a bifunctional enzyme with both anthocyanin 3'- and 5'- O-glucosyltransferase activities, catalyzed the formation of the first stable anthocyanin by UDP-glucose: anthocyanidin 3-O-glucosyltransferase [27] and modified anthocyanins to be more stable molecule complex and produced purple color by UDP-glucose: anthocyanin 5-O-glucosyltransferase [28]. Phenylalanine ammonia lyase (PAL) was the key enzyme to catalyze the first step of phenylpropanoid pathway in plants, and Coniferyl-Aldehyde Dehydrogenase (CALDH) is a gene involved in this pathway, which is related to the synthesis of anthocyanins. [29–31]. 4CL was a key enzyme that functions early in the phenylpropane pathway. And the proprotein 4CL converts 4-coumaric acid and other cinnamic acids to the corresponding CoA thiol esters, which can be used to many secondary metabolites, such as flavonoids, isoflavones, lignin, etc [32]. SGTase can be used as a multiple phenylpropanoid glucosylation enzyme, which exhibited significant activity with flavonoids [35].β-glucosidase (bglx) was an anthocyanase, which can hydrolyze anthocyanidins into anthocyanin [33]. E1.11.1.7 is a peroxidase that may lead to the degradation of anthocyanin [36]. Above previous studies further confirmed the rationality and reliability of identified genes in *phalaenopsis* from RNA-seq in this study.

In summary, this study provides a good reference and verification basis for the selection and identification of flower color regulatory genes and the discovery of regulatory pathway. Furthermore, the correlation between SNP mutation and DEGs related flower color of *Phalaenopsis* was explained in RNA level for the first time. These results would provide a new insight to further study the gene regulation and expression in genetic variants and differential expression gene from RNA-seq data in other species.

Materials And Methods

Sample collection and extraction

Purple and white flowers were collected and used from Hainan Boda Orchid Scientific Technology Company nursery in this study (Fig. 1). Three replicates of each color sample were collected from one single plant during the blooming period. These samples were put into liquid nitrogen immediately and stored at -80°C, and the total RNA of them were extracted by using the kit in Qubit 2.0 Fluorometer (Life Technologies, USA). The purity of the RNA sample was tested by Nanodrop (Nanodrop Technologies, USA). RNA concentration and RNA integrity were evaluated by using Qubit and Agilent 2100.

Construction of cDNA library and sequencing

The magnetic beads labeled with Oligo (dT) were used to enrich the mRNA of eukaryotes by A-T complementary pairing with the polyA tail of mRNA. The fragmentation buffer solution was added to break the mRNA into short fragments, and mRNA was used as a template to synthesize single-stranded cDNA by random hexamers. Then the double-stranded cDNA was synthesized with the buffer solution, dNTPs and DNA polymerase I, and then purified with AMPure XP beads. The purified double-stranded cDNA was optimized and the sequencing adapter was added. Finally the cDNA library were amplified by PCR, and library quality was assessed by Agilent Bioanalyzer 2100 system. Sequencing was carried out using an Illumina HiSeq 2500 instrument. The above work is completed by the laboratory staff of the company.

Quality control and function annotation

The quality of raw data from RNA-seq was controlled by fastQC v0.11.7. The reads with joints and low-quality (Qphred < = 20 bases account for more than 50% of the whole read) were removed, and the clean reads were used for subsequent analysis. Next, the unigenes obtained by sequencing were annotated by the eggNOG v5.0 software and the Diamond software, which aligned protein sequences into the Nr, Swiss-prot, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Cluster of Orthologous Groups of proteins (COG) databases [37, 38].

Differential expression analysis

In order to confirm the reliability of samples selection and reduplication, we carried out the correlation analysis of the gene expression among six samples. The Pearson's correlation coefficient (R^2) is set greater than 0.9 as the evaluation standard for correlation analysis between biological repeat samples. The Unigene was quantified and transcript abundances were calculated, then we used the DESeq2 package of R to normalize the abundances of gene expression and analyzed differentially expressed genes (DEGs). The P -value < 0.01 was set for DEGs, and DEGs with $|\log_2\text{foldchange}| > 1.5$ were considered as up-regulated or down-regulated genes. The foldchange was the ratio of the expression value of purple flower samples to the white flower samples.

Enrichment analysis of differential genes

The DEGs were compared with the analysis results of eggNOG v2.0.0 software and obtained the annotation information from the Gene Ontology (GO) databases of these differential genes. GO

enrichment analysis of DEGs was implemented by the clusterProfiler package of R[39], and GO terms with corrected *P*-value that less than 0.05 were considered significantly enriched by differential expressed genes. Furthermore, we used KOBAS v3.0 to analyze the statistical enrichment of DEGs in KEGG pathways under $FDR \leq 0.05$ [40].

SNP identification and relationship with differential expression genes

The single nucleotide polymorphisms (SNP) from the unigenes were identified by using GATK4 v1.8.1 with the default parameters [41]. In order to further study the relationship between SNP and differential expression genes, we conducted the evaluation in two aspects: (1) Gene level: the mutation frequency of the gene was calculated, mutation frequency was defined by the formula of

$$\text{Mutation frequency} = \frac{\text{Number of SNPs}}{\text{Length of gene}}.$$

Then, the genes were divided into two groups: DEGs and non-DEGs, and the mutation frequency of these genes between two groups were tested by T-test. (2) Base level: the genes were divided into four categories according the mutated nucleotide bases: A, T, C and G, which indicated the genes with the most mutations in base A, T, C, and G, respectively. Then each type genes were divided into two groups: DEGs and non-DEGs, and the mutation frequency of these two groups were also tested by T-test.

Declarations

Contributions

SQX and PL conceived the project and designed the experiments, DCH, WSL to provide the sample. YD, DHY and MYW collected datasets and performed the bioinformatics analysis, YD plotted figures, and YD, MYW, PL and SQX wrote the manuscript. SQX revised the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

There were no competing interests.

Availability of data and materials

All sequencing data have been deposited in <http://isodb.xieslab.org/data/download/phalaenopsis>. Analyze the relevant data generated are included in the manuscript and the supplemented materials.

Statement of Authorization of Materials

The Phalaenopsis were collected and used from Hainan Boda Orchid Scientific Technology Company nursery. The authors of this article, Dai-Cheng Hao, Wei-Shi Li, are affiliated with this company and the use of the material has been authorized by them and this study protocol comply with relevant institutional, national, and international guidelines and legislation.

Funding

This work was supported by grants from the National Natural Science Foundation of China (grant number 32060149 and 31760316), Hainan Provincial Natural Science Foundation of China (320RC500 and 2018CXTD33), Priming Scientific Research Foundation of Hainan University (grant number KYQD (ZR) 1721)

Acknowledgements

We thank Hanan Boda Orchid Scientific Technology Company Nursery for the research samples

References

1. Zhao J, Dixon R: **US Department of Agriculture. 2015. Floriculture crops.14 summary. Washington: US Department of Agriculture. *The 'ins' and 'outs' of flavonoid transport Trends Plant Sci* 2010, 15(2):72–80.**
2. Su V, Hsu B-D: **Cloning and expression of a putative cytochrome P450 gene that influences the colour of Phalaenopsis flowers. *Biotechnology letters* 2003, 25(22):1933–1939.**
3. Griesbach RJ, Janick J, Whipkey A: **Development of Phalaenopsis orchids for the mass-market. In: *Trends in New Crops & New Uses Fifth National Symposium: 2002.***
4. Yang Y, Wang J, Ma Z, Sun G, Zhang C: **De novo sequencing and comparative transcriptome analysis of white petals and red labella in Phalaenopsis for discovery of genes related to flower color and floral differentiation. *Acta Societatis Botanicorum Poloniae* 2014, 83(3):191–199.**
5. Chugh S, Guha S, Rao IU: **Micropropagation of orchids: A review on the potential of different explants. *Scientia Horticulturae* 2009, 122(4):507–520.**
6. Wang LM, Zhang J, Dong XY, Fu ZZ, Jiang H, Zhang HC: **Identification and functional analysis of anthocyanin biosynthesis genes in Phalaenopsis hybrids. *Biologia Plantarum* 2018, 62(1):45–54.**
7. Tanaka Y, Sasaki N, Ohmiya A: **Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *The Plant journal: for cell and molecular biology* 2008, 54(4):733–749.**
8. Gomez C, Conejero G, Torregrosa L, Cheynier V, Ageorges A: **In vivo grapevine anthocyanin transport involves vesicle-mediated trafficking and the contribution of ANTHOMATE transporters and GST.**

Plant Journal 2011, **67**(6):960–970.

9. Pourcel L, Irani NG, Lu Y, Riedl K, Schwartz S, Grotewold E: **The Formation of Anthocyanic Vacuolar Inclusions in *Arabidopsis thaliana* and Implications for the Sequestration of Anthocyanin Pigments.** *Molecular Plant* 2010, **3**(1):78–90.
10. Zhong HQ, Huang ML, Jian-She WU, Fang RH: **Cloning and Expression of Key Enzyme Genes Involved in *Phalaenopsis* Anthocyanins Synthesis.** *Fujian Journal of Agricultural Sciences* 2013.
11. Tatsuzawa F, Saito N, Seki H, Hara R, Honda T: **Acylated cyanidin glycosides in the red-purple flowers of *Phalaenopsis*.** *Phytochemistry* 1997, **45**(1):173–177.
12. Ling LF, Subramaniam S: **Biochemical Analyses of *Phalaenopsis violacea* Orchids.** *Asian Journal of Biochemistry* 2007, **2**(4):237–246.
13. Chen L, Li L, Dai Y, Wang X, Duan Y, Yang G: **De novo transcriptome analysis of *Osmanthus serrulatus* Rehd. flowers and leaves by Illumina sequencing.** *Biochemical Systematics and Ecology* 2015, **61**:531–540.
14. Zhou X, Li J, Zhu Y, Ni S, Chen J, Feng X, Zhang Y, Li S, Zhu H, Wen Y: **De novo Assembly of the *Camellia nitidissima* Transcriptome Reveals Key Genes of Flower Pigment Biosynthesis.** *Frontiers in Plant Science* 2017, **8**.
15. Wu Q, Wu J, Li SS, Zhang HJ, Feng CY, Yin DD, Wu RY, Wang LS: **Transcriptome sequencing and metabolite analysis for revealing the blue flower formation in waterlily.** *BMC Genomics* 2016, **17**(1):897.
16. Gao LW, Jiang DH, Yang YX, Li YX, Sun GS, Ma ZH, Zhang CW: **De novo sequencing and comparative analysis of two *Phalaenopsis* orchid tissue-specific transcriptomes.** *Russian Journal of Plant Physiology* 2016, **63**(3):391–400.
17. Zhao Y, Wang K, Wang W-l, Yin T-t, Dong W-q, Xu C-j: **A high-throughput SNP discovery strategy for RNA-seq data.** *BMC Genomics* 2019, **20**(1).
18. Patra B, Schluttenhofer C, Wu Y, Pattanaik S, Yuan L: **Transcriptional regulation of secondary metabolite biosynthesis in plants.** *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 2013, **1829**(11):1236–1247.
19. Yang IS, Kim S: **Analysis of Whole Transcriptome Sequencing Data: Workflow and Software.** *Genomics Inform* 2015, **13**(4):119–125.
20. Chong X, Zhang F, Wu Y, Yang X, Zhao N, Wang H, Guan Z, Fang W, Chen F: **A SNP-Enabled Assessment of Genetic Diversity, Evolutionary Relationships and the Identification of Candidate Genes in *Chrysanthemum*.** *Genome Biol Evol* 2016, **8**(12):3661–3671.
21. Zhang N, Chen L, Ma S, Wang R, He Q, Tian M, Zhang L: **Fine mapping and candidate gene analysis of the white flower gene *Brwf* in Chinese cabbage (*Brassica rapa* L.).** *Scientific reports* 2020, **10**(1):6080–6080.
22. Sudarsono, Haristianita MD, Handini AS, Sukma D: **Molecular marker development based on diversity of genes associated with pigment biosynthetic pathways to support breeding for novel colors in *Phalaenopsis*.** *Acta Horticulturae* 2017(1167):305–312.

23. Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X: **Bridger: a new framework for de novo transcriptome assembly using RNA-seq data.** *Genome Biol* 2015, **16**:30.
24. Reinprecht Y, Perry GE, Peter Pauls K: **A Comparison of Phenylpropanoid Pathway Gene Families in Common Bean. Focus on P450 and C4H Genes.** In: *The Common Bean Genome*. 2017: 219–261.
25. Shaipulah NF, Muhlemann JK, Woodworth BD, Van Moerkercke A, Verdonk JC, Ramirez AA, Haring MA, Dudareva N, Schuurink RC: **CCoAOMT Down-Regulation Activates Anthocyanin Biosynthesis in Petunia.** *Plant Physiol* 2016, **170**(2):717–731.
26. Guo Y, Qiu L-J: **Allele-specific marker development and selection efficiencies for both flavonoid 3'-hydroxylase and flavonoid 3',5'-hydroxylase genes in soybean subgenus soja.** *Theor Appl Genet* 2013, **126**(6):1445–1455.
27. Sui X, Gao X, Ao M, Wang Q, Yang D, Wang M, Fu Y, Wang L: **cDNA cloning and characterization of UDP-glucose: anthocyanidin 3-O-glucosyltransferase in Freesia hybrida.** *Plant Cell Reports* 2011, **30**(7):1209–1218.
28. Yamazaki M, Gong Z, Fukuchi-Mizutani M, Fukui Y, Tanaka Y, Kusumi T, Saito K: **Molecular Cloning and Biochemical Characterization of a Novel Anthocyanin 5-O-Glucosyltransferase by mRNA Differential Display for Plant Forms Regarding Anthocyanin.** *Journal of Biological Chemistry* 1999, **274**(11):7405–7411.
29. Holcroft DM, Kader AA: **Carbon dioxide-induced changes in color and anthocyanin synthesis of stored strawberry fruit.** *HortScience* 1999, **34**(7):1244–1248.
30. Tao J, Cao C, Zhao D, Zhou C, Liang G: **Molecular analysis and expression of phenylalanine ammonia-lyase from poinsettia (Euphorbia pulcherrima willd.).** *AFRICAN JOURNAL OF BIOTECHNOLOGY* 2010, **10**(2):126–135.
31. Bai L, Chen Q, Jiang L, Lin Y, Ye Y, Liu P, Wang X, Tang H: **Comparative transcriptome analysis uncovers the regulatory functions of long noncoding RNAs in fruit development and color changes of Fragaria pentaphylla.** *Horticulture Research* 2019, **6**(1):42.
32. Sun H, Guo K, Feng S, Zou W, Li Y, Fan C, Peng L: **Positive selection drives adaptive diversification of the 4-coumarate: CoA ligase (4CL) gene in angiosperms.** *Ecology and Evolution* 2015, **5**(16):3413–3420.
33. Oren-Shamir M: **Does anthocyanin degradation play a significant role in determining pigment concentration in plants?** *Plant Science* 2009, **177**(4):310–316.
34. Yang Y, Sun F, Zhang C: **Construction of Full-length cDNA Library and the cDNA Cloning of F3'H in Phalaenopsis.** *Acta Botanica Boreali-Occidentalia Sinica* 2013, **33**(9):1731–1731.
35. Taguchi G, Imura H, Maeda Y, Kodaira R, Hayashida N, Shimosaka M, Okazaki M: **Purification and characterization of UDP-glucose: hydroxycoumarin 7-O-glucosyltransferase, with broad substrate specificity from tobacco cultured cells.** *Plant Sci* 2000, **157**(1):105–112.
36. Luo H, Li W, Zhang X, Deng S, Xu Q, Hou T, Pang X, Zhang Z, Zhang X: **In planta high levels of hydrolysable tannins inhibit peroxidase mediated anthocyanin degradation and maintain abaxially red leaves of Excoecaria Cochinchinensis.** *BMC Plant Biology* 2019, **19**(1):315.

37. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen Lars J *et al*: **eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses**. *Nucleic Acids Research* 2018, **47**(D1):D309-D314.
38. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nat Methods* 2015, **12**(1):59–60.
39. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters**. *Omics* 2012, **16**(5):284–287.
40. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L: **KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W316-322.
41. Lopez-Maestre H, Brinza L, Marchet C, Kielbassa J, Bastien S, Boutigny M, Monnin D, Filali AE, Carareto CM, Vieira C *et al*: **SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence**. *Nucleic Acids Research* 2016, **44**(19):e148-e148.

Figures

Figure 1

The experimental materials used in this study (A: purple petals of Phalaenopsis; B: white petals of Phalaenopsis)

Figure 2

Differential expression genes of six samples (A: Correlation among samples; B: DEGs between purple and white samples; C: Number of up and down regulated genes; D: Functional annotation databases of DEGs)

Figure 3

Enrichment pathway of DEGs (A: GO classification; B: KEGG pathway)

Figure 4

SNP identification, comparison between DEGs and non-DEGs and comparison between down and up (A: Distribution of variation types of SNP; B: Mutation frequency comparison of DEGs and non-DEGs; C: Mutation frequency comparison of down and up)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTable.xlsx](#)
- [SupplementaryTableandFigure.docx](#)