The production of *Phalaenopsis* is continuously increasing because of its wide popularity in China, the United States, and many European countries (Su and Hsu 2003). Several commercial features, such as brightness of the color, multiple color forms, and different color patterns were enhanced with the improvement of *Phalaenopsis* cultivation. In recent years, the global consumption of *Phalaenopsis* has gradually increased (Griesbach et al. 2002) and the ornamental and commercial values of *Phalaenopsis* are mainly determined by the quality and variety of flower colors (Yang et al. 2014).

Previous studies on *Phalaenopsis* flower color focused on the breeding of rare or colorful varieties and the mechanism of color formation (Chugh et al. 2009; Wang et al. 2018). *Phalaenopsis* is an ideal plant for studying flower coloring mechanisms because of the color diversities. Tanaka et al. first reported that the diversity of anthocyanins was the primary factor of its colourful flowers (Tanaka et al. 2008). Anthocyanin originates from malonyl-coenzyme A and coumaryl-coenzyme A were catalysed by a series of enzymes in the cytoplasm, which is subsequently modified.
by glycosylation, methylation, and acylation. Through transporters or transporter vesicles, anthocyanin enters vacuoles and accumulates to create various colors from white to purple (Tanaka et al. 2008; Zhao and Dixon 2010; Gomez et al. 2011). The most related enzymes in the anthocyanin synthesis pathway were encoded by genes that exist in multiple gene families (Pourcel et al. 2010). Four key enzyme genes, including chalcone synthase (CHS), dihydro flavonoid reductase (DFR), flavanone 3', 5'-hydroxylase (F3’5’H), and anthocyanin synthetase (ANS) were highly associated with anthocyanin synthesis in Phalaenopsis. Su et al. (Su and Hsu 2003) isolated the flavonoid 3', 5'-hydroxylase gene from the cytochrome P450 family and transferred it into petals of Phalaenopsis ssaflower by gene marksmanship, which resulted in the petal color changing from red to purple within 48 h. These genes had high homology with color-related genes and also were identified in other ornamental plants (Zhong et al. 2013). In addition, Tatsuzawa et al. isolated cyanidin 3, 7, 3'-triglucoside acylated by acid from five red-purple Phalaenopsis varieties (Tatsuzawa et al. 1997) and identified several anthocyanin components such as delphinidin, peonidin, petunidin, malvidin, and pelargonidin (Ling and Subramanian 2007).

Gene expression related to flower color formation can be analysed through RNA-seq sequencing technology, which has been extensively used for decoding differential genes due to its low cost and high throughput. Chen et al. (Chen et al. 2015) conducted RNA-seq sequencing of flowers and leaves from Osmanthus serrulatus Rehd. and obtained 2,602 differentially expressed genes (DEGs), 33 of them were involved in carotenoid biosynthesis in the metabolic pathway. Zhou et al. (Zhou et al. 2017) reported that the expression of carotenoid biosynthesis genes (PSY, CrtZ, and BCH) and flavonoid biosynthesis genes (CHS, F3H, FLS, and ANS) resulted in the golden yellow petals of Camellia nitidissima. Wu et al. (Wu et al. 2016) identified 127 unigenes related to color synthesis and confirmed that the candidate gene UA3GT in the flavonoid metabolism pathway caused color formation of blue petals in Nymphaea ‘King of Siam.’

The flower color formation mechanism of Phalaenopsis is an important research topic, and the majority of current research has focused on genetic engineering technology, genetic analysis, and breeding selections in Phalaenopsis. Gao et al. (Gao et al. 2016) have investigated the regulation mechanism of flower color by using RNA-seq at the transcriptome level and revealed that the anthocyanin synthesis pathway was responsible for differences between red and yellow flowers in Phalaenopsis. Recently, RNA-seq has been used to analyse a large number of single nucleotide polymorphisms (SNPs) associated with plant biological and agronomic traits (Zhao et al. 2019). However, the relationship between transcription-level SNPs and DEGs of flower color traits still need to be further studied in Phalaenopsis. Therefore, we collected two contrasting flower colors (white and purple) and analysed DEGs and SNPs at the transcriptome level. Moreover, the contribution of SNPs related to DEGs in flower color traits was decoded. This study provides insight into the mechanism of flower color formation and enhances the understanding of the theoretical basis for the breeding improvement of Phalaenopsis flower color and cultivation of new varieties.

## Materials and methods

### Sample collection and extraction

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

### Construction of cDNA library and sequencing

The stranded RNA-seq was used to build the library, and eukaryotic mRNA was enriched using magnetic beads with oligo (dT). The fragmentation buffer solution was added to break the mRNA into short fragments, and mRNA was used as a template to synthesise first-stranded cDNA by random hexamers. Then, the second strand cDNA was synthesised with the buffer solution, dNTPs, and DNA polymerase I, and then purified with AMPure XP beads. The purified

---

![Fig. 1](image_url)

The experimental materials used in this study (A: purple petals of Phalaenopsis; B: white petals of Phalaenopsis)
second-strand cDNA was optimised, and the sequencing adapter was added. Finally, the cDNA library was amplified by PCR, and library quality was assessed using the Agilent Bioanalyzer 2100 system. The cDNA library products were sequenced using paired-end sequencing technology with read lengths of 150 bp on an Illumina HiSeq 2500 instrument platform.

**Quality control and function annotation**

The quality of raw data from RNA-seq was controlled by fastQC v0.11.7. The reads with adapters and low-quality (Qphred ≤ 20 bases account for more than 50% of the whole read) were removed by Trimmomatic v0.36, and the clean reads were used for subsequent analysis. Hisat2 v2.1.0 was used to align against the reference genome (www.ncbi.nlm.nih.gov/genome/11403) (Cai et al. 2015), and the aligned sequences were assembled into potential transcripts by StringTie v1.3.6. Secondly, sequences less than 150 bp were filtered and redundant transcript sequences were removed by cd-hit v4.8.1. Finally, featurecounts v2.0.0 software was used for quantitative analysis of the follow-up transcript (Kim et al. 2015; Liao et al. 2014; Fu et al. 2012; Pertea et al. 2016). The transcripts 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 (Huerta-Cepas et al. 2018; Buchfink et al. 2015).

**Differential expression analysis**

To confirm the reliability of sample selection and reduplication, we carried out a correlation analysis of the gene expression among the six samples. The correlation between paired samples from the six datasets was calculated using the fragments per kilobase of transcript per million (FPKM) values of transcripts. The Pearson’s correlation coefficient ($R^2$) was set greater than 0.9 as the evaluation standard for correlation analysis between biological replication samples. Transcripts abundances were calculated, and normalized, and DEGs were identified by DESeq2 package of R. A $P$-value $< 0.01$ was set as cut-off for DEGs, and DEGs with $|\log_{2}\text{fold change}| > 1.5$ were considered up-regulated or down-regulated genes. The foldchange was the ratio of the expression value of purple flower samples to that of white flower samples.

**Enrichment analysis of differential genes**

The DEGs were compared with the analysis results of eggNOG v2.0.0 software, and the annotation information of DEGs was obtained from GO databases. GO enrichment analysis was implemented by using the clusterProfiler package of R (Yu et al. 2012), and GO terms with corrected $P$-value less than 0.05 were considered as the significant enrichment. Furthermore, the enrichment of DEGs in KEGG pathways were analysed by KOBAS v3.0 with FDR ≤ 0.05 (Xie et al. 2011).

**qRT-PCR validation of flower color-related genes**

Quantitative real-time PCR (qRT-PCR) was performed according to the method of Zheng et al. (Zheng et al. 2020). Total RNA was isolated by using an RNAprep Pure Plant Kit (Tiangen) according to the manufacturer’s instructions. The cDNA was synthesised using M-MLV reverse transcriptase (Promega) and was diluted 10 times, and 2 μL cDNA was used for subsequent PCR. qRT-PCR was performed using SYBR® Premix Ex Taq™ II (Takara) according to the manufacturer’s instructions on a 7300 Real-time PCR System (Applied Biosystems), and the relative expression of each gene was calculated according to the $2^{-\Delta\Delta CT}$ method. The glyceraldehyde-3-phosphate dehydrogenase gene was used as an endogenous reference for real-time PCR, and analyses were performed with three technical and three biological replicates. The primers used in this study were listed in Table S1.

**SNP identification and relationship with DEGs**

Unique mapped data from RNA-seq was used to identify single nucleotide polymorphisms (SNPs). SNPs from transcripts were identified using GATK4 v1.8.1 with default parameters (Lopez-Maestre et al. 2016). To improve the accuracy of SNPs, SNPs located in the gene coding region and read coverage of SNP was greater than 2 (Figure S1).

Two evaluations were conducted to further study the relationship between SNPs and DEGs: (1) Gene level: The mutation frequency of the gene was calculated by the formula: Mutation frequency = $\frac{\text{Number of SNPs}}{\text{Length of gene}}$. All genes were divided into two groups: DEGs and non-DEGs, and the mutation frequencies of these two groups were compared and tested using t-test. (2) Base level: Genes were divided into four categories according to the mutated nucleotide bases: A, T, C, and G. Then, each type of gene was divided into two groups: DEGs and non-DEGs, and the mutation frequencies of these two groups were also compared and tested.
Results

Transcripts mapped 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 the Illumina HiSeq platform, generating 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) were treated; adapters, impurities, and low-quality reads were removed, and 329,491,250 clean reads were retained. Samples P2 and W2 had the highest (58,673,380) and lowest (48,567,124) amounts of clean data, respectively (Table 1). Moreover, the GC content of each sample was approximately 45%, and the nucleotide bases with Phred (Phred = -10log10(e)) values greater than 20 and greater than 30 accounted for 97% and 93% of the total nucleotide bases, respectively (Table 1). These results demonstrated that the quality of clean data was adequate for downstream analyses.

A total of 30,134 transcript sequences with a total nucleotide base number of 323,511,212 bp were sequenced. The maximum and minimum lengths of the sequences were 264,352 bp and 152 bp, respectively (Table 2). The average and N50 lengths of the mapped sequences were 10,735.75 bp and 33,917 bp, respectively (Table 2). Moreover, transcripts with the sequences lengths distributed from 1,000 bp to 3,000 bp accounted for 53.37% (Figure S2). The Nr, Swiss-Prot, GO, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Clusters of Orthologous Groups of proteins (COG) databases were used to obtain the functional annotation of the amino acid sequences of the transcripts. The number of annotations obtained in Nr, Swiss-Prot, GO, KEGG, and COG were 21,143 (90.7%), 15,062 (64.6%), 10,827 (46.4%), 9,854 (42.3%), and 18,954 (82.7%), respectively (Table S2).

COG database distribution

To further annotate homologous transcripts, 18,954 transcripts from the COG database were further analysed. 12,436 (61.84%) transcript sequences were grouped into three categories: (1) information storage and processing (3,947), (2) cellular processes and signalling (4,432), and (3) metabolism (4,057), They accounted for 19.63%, 22.04%, and 20.17% of all sequences, respectively (Table 3 and Figure S3). In addition, 32.41% of the sequences were functionally unknown.

Identification of DEGs

The three biological replications from purple and white flowers had a high degree of similarity (Fig. 2A, R^2 > 0.95). A total of 1,175 DEGs were detected between purple and white samples; 718 genes were up-regulated, and 457 genes were down-regulated (Fig. 2B and C, Table S3). To illustrate the differential expression of identified genes in different colors of *Phalaenopsis*, the genes with normalised FPKM values from six samples were compared and clustered in the heatmap (Figure S4). DEGs were annotated using the Nr, Swiss-Prot, GO, KEGG, and COG databases, and the number of annotated genes was 1,163 (98.9%), 898 (76.4%), 474 (40.3%), 431 (36.7%), and 882 (75.1%), respectively. Of them, 277 genes were simultaneously annotated in all five databases (Fig. 2D and Table S4).

GO Enrichment and KEGG pathway analysis of DEGs

To further annotate the function of DEGs, GO enrichment analysis was performed on the DEGs. 1,247 genes were involved in biological processes, 134 genes related to cell components, and 314 genes involved in molecular functions (Fig. 3A and Table S5). Biological processes were significantly enriched in response to organonitrogen compounds, chitin, and wounding and the secondary metabolite
biosynthetic process (Fig. 3A). The enrichment of cell components included apoplasts, peroxisomes, microbodies, plastoglobuli, and peroxisomal membranes (Fig. 3A). The significant enrichment of the molecular function was primarily involved in calcium ion binding, CoA-ligase activity, ligase activity, forming carbon-sulphur bonds, etc. (Fig. 3A).

KEGG pathway analysis indicated that DEGs involved in 76 signals pathways, including biosynthesis of secondary metabolites, metabolic pathways, phenylpropanoid biosynthesis, flavonoid biosynthesis, and carotenoid biosynthesis. Among these pathways, the biosynthesis of the secondary metabolite was the most significant (Fig. 3B and Table S6). A total of four DEGs (F3’H, C4H, CCoAOMT, and UA3’5’GT) were identified and annotated in the flavonoid biosynthesis metabolic pathway (Figure S5), which had been reported to play an important role in flower color formation (Patra et al. 2013). Among them, the F3’H gene was up-regulated and C4H, CCoAOMT, and UA3’5’GT were all down-regulated (Table 4 and Table S7). In addition, 10 DEGs were identified and annotated in the phenylpropanoid biosynthesis of the KEGG pathway (Figure S6). The expression of CCoAOMT, C4H, PAL, 4CL, CCR, CALDH, and bglx were down-regulated, except for CAD, SGTase, and E1.11.1.7, (Table 4 and Table S7). Phenylpropanoid biosynthesis was also an important pathway in flower color formation reported in previous investigation (Patra et al. 2013).

Furthermore, 27 DEGs were enriched in the DNA-binding transcription factor activity in GO enrichment. 9 genes of them were from MYB family, of which six genes were up-regulated and three genes were down-regulated (Figure S7A and Table S8). 18 genes were from WRKY family, of which 16 genes were up-regulated and 2 genes were down-regulated (Figure S7B and Table S8). MYB and WRKY gene families were reported to regulate gene expression during flavonoid and anthocyanin biosynthesis (Lloyd et al. 2017).

### The qRT-PCR validation of DEGs

Seven color-related genes (CCoAOMT, C4H, 4CL, CCR, CALDH, E1.11.1.7, and PAL) were cross-validated using RNA-seq data and qRT-PCR experiments (Fig. 4A and Table S9). Although expression levels of qRT-PCR and the fold changes of RNA-seq in the transcriptomes were different, the regulation trends of all validated genes were the same (Fig. 4B). Among them, CALDH was up-regulated, and all other genes were down-regulated. The down-regulated DEGs C4H and PAL were highly expressed in purple petals, while the expression of CALDH was higher in white petals than that in purple petals. The above results demonstrated that the qRT-PCR results were consistent with the transcriptome sequencing data from RNA-seq. These DEGs may played an important role in color change formation regulation of Phalaenopsis petals.

### SNP identification and relationship with DEGs

A total of 123,168 SNPs were identified in transcripts (Table S10), and the transition (G-A and C-T) frequencies were higher than those in transversion sites (Fig. 5A). The C->T, G->A, A->G, and T>C ratio of transition sites were 13.86%, 13.77% and 12.01%, 11.75%, respectively (Table S10 and Fig. 5C), which were transversion sites (C->A: 7.10%, G->T: 7.10%, G->C: 6.07%, C->G: 5.96%, A->T: 0%...
that in non-DEGs (0.0019) (Tables S12 and S13). 95 SNPs were identified in 12 key genes, and the average mutation frequency was 0.0026 (Table S14), which was also higher than that in non-DEGs.

To further explain whether SNP mutations were associated with flower color, the mutation location of genes was identified (Table S11), and the mutation frequency of DEGs

5.82%, T->A: 5.80%, T->G: 5.43% and A->C: 5.33%) (Table S10). From the level of gene expression, the average expression of genes with SNP transversion was higher than that with transformation (Fig. 5B).

Furthermore, a total of 7,675 and 115,493 SNPs were identified in DEGs and non-DEGs, respectively. The average mutation frequency of DEGs (0.0035) was higher than that in non-DEGs (0.0019) (Tables S12 and S13). 95 SNPs were identified in 12 key genes, and the average mutation frequency was 0.0026 (Table S14), which was also higher than that in non-DEGs.

To further explain whether SNP mutations were associated with flower color, the mutation location of genes was identified (Table S11), and the mutation frequency of DEGs
Table 4 The annotation pathway of identified transcripts of Phalaenopsis

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

and non-DEGs from the transcript sequences were compared. The results showed that the mutation frequency of SNPs between DEGs and non-DEGs was extremely significant (P < 0.01) (Fig. 5D). In the DEGs group, the mutation frequency of SNPs in up- and down-regulated genes was highly significant (Fig. 5E), which was beneficial for pigmentation based on the foldchange value. The significance of 12 key genes and differential genes was compared, and the P value was 0.055 (Fig. 5F). Furthermore, genes of each group were divided into four categories: A, C, T, and G at the base level. The mutation frequencies in the four categories were compared between the two groups (DEGs and non-DEGs). The SNP mutations in A, C, T, and G categories between DEGs and non-DEGs were all significant (A: \( P = 1.4e^{-11} \), C: \( P = 6.4e^{-10} \), T: \( P = 2.7e^{-4} \), G: \( P = 1.9e^{-12} \)) (Figure S8A, B, C, and D). The above results illustrated that the SNP mutations were strongly associated with color formation in Phalaenopsis.

Discussion

Transcriptome refers to the transcription of all genes at a specific developmental stage or physiological state, is necessary to explain functional elements of the genome and the underlying mechanisms of biological growth and diseases. (Yang and Kim 2015) In this study, the color formation mechanism of purple and white petal in Phalaenopsis was explored at the transcriptional level by using RNA-seq. 30,134 transcripts with annotated information were identified from RNA-seq data, and 1,175 of them were DEGs. Among them, 718 DEGs were up-regulated, and 457 DEGs were down-regulated. Moreover, this study first reports the correlation between SNP and gene expression in color formation of Phalaenopsis, these findings revealed the flower color formation mechanism and provided a new insight into the correlation between genetic variation and gene expression at the transcription level.

Previous studies had found that SNPs at the DNA level were related to flower color in chrysanthemum and cabbage, which were used as markers to assist breeding (Chong et al. 2016; Zhang et al. 2020). In Phalaenopsis hybrids, DNA allele diversity of SNPs had been used as an assistant marker to associate and predict the color of flowers (Sudarsono et al. 2017). But the detailed relationship between SNPs and phenotypes at the transcriptional level is rarely reported. RNA-seq, a powerful technique for gene expression at the transcriptome level, can be used not only for differential expression gene identification, but also for genetic variation analysis (Chang et al. 2015). In this study, 123,168 SNP sites were identified through RNA-seq, and mutation frequencies of SNPs associated with DEGs in genes were first investigated at the nucleotide base level. Results revealed 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 in flower color in Phalaenopsis.

Flavonoid biosynthesis and phenylpropanoid biosynthesis are two metabolic pathways related to flower color formation (Patra et al. 2013). In this study, we identified 12 genes (C4H, CCoAOMT, F3’H, UA3’5’GT, PAL, 4CL, CCR, CAD, CALDH, bgx, SGThirse and E1.11.1.7) that related flower color synthesis pathways. qRT-PCR was used to verify the expression pattern of these DEGs, and seven color-related genes (CCoAOMT, C4H, 4CL, CCR, CALDH, E1.11.1.7, and PAL) were experimentally confirmed. The results demonstrated that qRT-PCR expression levels of these genes were consistent with the change trend.
of RNA-seq in the transcriptomes. These genes had been investigated and reported in previous studies (Reinprecht et al. 2017; Shaipulah et al. 2016; Guo and Qiu 2013; Sui et al. 2011; Yamazaki et al. 1999; Holcroft and Kader 1999; Tao et al. 2010; Bai et al. 2019; Sun et al. 2015; Oren-Shamir 2009). C4H has been reported not only as the key enzyme involved in the second step of flavonoid synthesis, but also as the first oxidoreductase of cytochrome P450 in the phenylpropane biosynthesis pathway, which catalysed a specific hydroxylation reaction and generated coumaric acid, which was a precursor of flavonoids (Reinprecht et al. 2017). CCoAOMT contributes to the formation of phenylpropylene, and downregulation of its expression leads to the activation of anthocyanin biosynthesis (Shaipulah et al. 2016). F3’H was a microsomal cytochrome P450-dependent monoxygenase which is a critical enzyme in the generation of the hydroxylate B-ring in flavonoids and played an important role in flower coloring (Guo and Qiu 2013). Yang et al. (Yang et al. 2013) reported that the upregulation and inhibition of the F3’H gene were closely related to anthocyanin accumulation in the torenia hybrid, and the overexpression of F3’H increased the accumulation of cyanidin and resulted in red sunflower petals. Anthocyanin 3’-, 5’- O-glucosyltransferase activities, catalysed the formation of the first stable anthocyanin by UDP-glucose: anthocyanidin 3-O-glucosyltransferase (Sui et al. 2011) and modified anthocyanins to be more stable molecules complex and produced purple color by UDP-glucose: anthocyanin 5-O-glucosyltransferase (Yamazaki et al. 1999). Phenylalanine ammonia lyase (PAL) catalysed the first step of the phenylpropanoid pathway in plants, and coniferyl-aldehyde dehydrogenase (CALDH) was a gene involved in this pathway, which was related to the synthesis of anthocyanins (Tao et al. 2010; Holcroft and Kader 1999; Bai et al. 2019). 4CL was a key enzyme that functions early in the phenylpropane pathway. The proprotein 4CL converted 4-coumaric acid and other cinnamic acids to the corresponding CoA thiol esters, which can be used for many secondary metabolites, such as flavonoids, isoflavones, and lignin (Sun et al. 2015). SGTase was used as a multiple phenylpropanoid glucosylation enzyme, which exhibited significant activity against flavonoids (Taguchi et al. 2000). β-Glucosidase (bglx) is an anthocyanidase that hydrolyses anthocyanidins into anthocyanin, (Oren-Shamir 2009) while E1.11.1.7 was a peroxidase that lead to the degradation of anthocyanin (Luo et al. 2019). MYB transcription factor is the most important transcription regulator in plant metabolic network, which controls
secondary metabolism, development, signal transduction. Anthocyanin is a secondary metabolite modified by flavonoids and a subclass of flavonoids. The biosynthesis of anthocyanin and other phenylpropanoids was also regulated by MYB (Yan et al. 2021). In this study, 9 MYB family genes were obtained, 3 of them genes were up-regulated and 6 genes were down-regulated, which indicated that MYB not only promoted but also inhibited anthocyanin accumulation to prevent anthocyanin excess in Phalaenopsis.

WRKY family plays an important role in regulating physiological metabolism and abiotic stress responses. For example, PbWRKY75 promote the expression of PbMYB10b, activate the anthocyanin biosynthesis genes DFR and UFGT, and promote the accumulation of anthocyanin. The interaction of McWRKY71 and McMYB12 directly activated McANR and participated in the regulation of proanthocyanidin biosynthesis (Wang et al. 2023). In this study, 18 genes from WRKY family were identified, and 16 of them were up-regulated, indicating that they played a positive role in the flower color synthesis of Phalaenopsis. These studies further confirmed the rationality and reliability of identified genes in Phalaenopsis from RNA-seq in this study.

In summary, this study provides a reference and verification basis for the selection and identification of flower color regulatory genes and the discovery of regulatory pathways. Furthermore, the relationship between SNP mutation and DEGs related to flower color of Phalaenopsis was explained at the RNA level. These results provide new insights to further study gene regulation and expression in genetic variants and DEGs from RNA-seq data in other species.

**Contributions**

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

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s13258-023-01422-5.

**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).
Data Availability All sequencing data can be obtained in the NCBI project PRJNA757191 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA757191). The analysis data is included in the manuscript and the supplemented materials.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests There were no competing interests.

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.

References

Bai L, Chen Q, Jiang L, Lin Y, Ye Y, Liu P, Wang X, Tang H (2019) Comparative transcriptome analysis uncovers the regulatory functions of long noncoding RNAs in fruit development and color changes of Fragaria pennanthylla. Hortic Res 6(1):42. https://doi.org/10.1038/s41438-019-0128-4

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12(1):59–60. https://doi.org/10.1038/nmeth.3176

Cai J, Liu X, Vanneste K, Proost T, Tsai W-C, Liu L-J, He Y, Xu Q, Bian C, Zheng Z, Sun F, Liu W, Hsiao Y-Y, Pan Z-J, Hsu C-C, Yang Y-P, Hsu Y-C, Chuang Y-C, Dievart A, Dufayard J-F, Xu X, Wang J-Y, Wang J, Xiao X-J, Zhao X-M, Du R, Zhang G-Q, Wang M, Su Y-Y, Xie G-C, Liu G-H, Li L-Q, Huang L-Q, Luo Y-B, Chen H-H, Van de Peer Y, Liu Z-J (2015) The genome sequence of the orchid Phalaenopsis equestris. Nat Genet 47(1):65–72. https://doi.org/10.1038/ng.3149

Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X (2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol 16:30. https://doi.org/10.1186/s13059-015-0396-2

Chen L, Li L, Dai Y, Wang X, Duan Y, Yang G (2015) De novo transcriptome analysis of Osmanthus serrulatus Rehd. Flowers and leaves by Illumina sequencing. Biochem Syst Ecol 61:531–540. https://doi.org/10.1016/j.bse.2015.07.027

Chong X, Zhang F, Wu Y, Yang X, Zhao N, Wang H, Guan Z, Fang W, Chen F (2016) A SNP-Enabled Assessment of Genetic Diversity, Evolutionary Relationships and the identification of candidate genes in Chrysanthemum. Genome Biol Evol 8(12):3661–3671. https://doi.org/10.1093/gbe/evw270

Chugh S, Guha S, Rao IU (2009) Micropropagation of orchids: a review on the potential of different explants. Sci Hort 122(4):507–520. https://doi.org/10.1016/j.scienta.2009.07.016

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gao LW, Jiang DH, Yang YX, Li YX, Sun GS, Ma ZH, Zhang CW (2016) De novo sequencing and comparative analysis of two Phalaenopsis orchid tissue-specific transcriptomes. Russ J Plant Physiol 63(3):391–400. https://doi.org/10.1134/s1021443716020072

Gomez C, Conejero G, Torregrosa L, Cheynier V, Ageorges A (2011) In vivo grapevine anthocyanin transport involves vesicle-mediated trafficking and the contribution of ANTHOMATE transporters and GST. Plant J 67(6):960–970

Griesbach RJ, Janick J, Whippley A (2002) Development of Phalaenopsis orchids for the mass-market. In: Trends in New Crops & New Uses Fifth National Symposium

Guo Y, Qiu L-J (2013) 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 126(6):1445–1455. https://doi.org/10.1007/s00122-013-2063-3

Holcroft DM, Kader AA (1999) Carbon dioxide–induced changes in color and anthocyanin synthesis of stored strawberry fruit. HortScience 34(7):1244–1248

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen Lars J, von Mering C, Bork P (2018) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47(D1):D309–D314. https://doi.org/10.1093/nar/gky1085

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357–360.

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30(7):923–930. https://doi.org/10.1093/bioinformatics/btu656

Ling LF, Subramaniam S (2007) Biochemical analyses of Phalaenopsis violacea Orchids. Asian J Biochem 2(4):237–246

Lloyd A, Brockman A, Aguirre L, Campbell A, Bean A, Cantero A, Gonzalez A (2017) Advances in the MYB–bHLH–WD repeat (MBW) Pigment Regulatory Model: Addition of a WRKY factor and co-option of an anthocyanin MYB for Betalain Regulation. Plant Cell Physiol 58(9):1431–1441. https://doi.org/10.1093/pcbd/pxz075

Lopez-Maestre H, Brinza L, Marchet C, Kielbassa J, Bastien S, Boutigny M, Monnin D, Filali AE, Cararetto CM, Vieira C, Picard F, Kremer N, Vavre F, Sagot M-F, Lacroix V (2016) SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. Nucleic Acids Res 44(19):e148–e148. https://doi.org/10.1093/nar/gkw655

Luo H, Li W, Zhang X, Deng S, Xu Q, Hou T, Pang X, Zhang Z, Zhang X (2019) In planta high levels of hydrolysable tannins inhibit peroxidase mediated anthocyanin degradation and maintain abaxially red leaves of Excoecaria Cochinchinensis. BMC Plant Biol 19(1):315. https://doi.org/10.1186/s12870-019-1903-y

Oren-Shamir M (2009) Does anthocyanin degradation play a significant role in determining pigment concentration in plants? Plant Sci 177(4):310–316. https://doi.org/10.1016/j.plantsci.2009.06.015

Patra B, Schlutenhofer C, Wu Y, Pattanaik S, Yuan L (2013) Transcriptional regulation of secondary metabolite biosynthesis in plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 1829(11):1236–1247.

Pourcel L, Irani NG, Lu Y, Riedl K, Schwartz S, Grotewold E (2010) The formation of Anthocyanic Vacuolar Inclusions in Arabidopsis thaliana and Implications for the sequestration of anthocyanin functions of long noncoding RNAs in fruit development and transcriptional, national, and international guidelines and legislation.
pigments. Mol Plant 3(1):78–90. https://doi.org/10.1093/mp/ssp071

Reinprecht Y, Perry GE, Peter Pauls K (2017) A Comparison of Phenylpropanoid Pathway Gene Families in Common Bean. Focus on P450 and C4H Genes. In: The Common Bean Genome. Compendium of Plant Genomes. pp 219–261. https://doi.org/10.1007/978-3-319-63526-2_11

Shaipulah NF, Muhlemann JK, Woodward BD, Van Moerkercke A, Verdgon JC, Ramirez AA, Haring MA, Dudareva N, Schuurink RC (2016) CCoAOMT Down-Regulation activates anthocyanin biosynthesis in Petunia. Plant Physiol 170(2):717–731. https://doi.org/10.1104/pp.15.01646

Su Y, Hsu B-D (2003) Cloning and expression of a putative cytochrome P450 gene that influences the colour of Phalaenopsis flowers. Biotechnol Lett 25(22):1933–1939. https://doi.org/10.1023/b:bile.0000039899.19657.53

Sudarsono, Haristianita MD, Handini AS, Sukma D (2017) Molecular marker development based on diversity of genes associated with pigment biosynthetic pathways to support breeding for novel colors in Phalaenopsis. Acta Hort 116705–312. https://doi.org/10.17660/ActaHortic.2017.1167.44

Sui X, Gao X, Ao M, Wang Q, Yang D, Wang M, Fu Y, Wang L (2011) cDNA cloning and characterization of UDP-glucose: anthocyanidin 3-O-glucosyltransferase in Freesia hybrida. Plant Cell Rep 30(7):1209–1218

Sun H, Guo K, Feng S, Zou W, Li Y, Fan C, Peng L (2015) Positive selection drives adaptive diversification of the 4-coumarate: CoA ligase (4CL) gene in angiosperms. Ecol Evol 5(16):3413–3420. https://doi.org/10.1002/ece3.1613

Taguchi G, Imura H, Maeda Y, Kodaira R, Hayashida N, Shimo-saka M, Okazaki M (2000) Purification and characterization of UDP-glucose: hydroxycoumarin 7-O-glucosyltransferase, with broad substrate specificity from tobacco cultured cells. Plant Sci 157(1):105–112. https://doi.org/10.1016/s0168-9452(00)00270-3

Tanaka Y, Sasaki N, Ohmiya A (2008) Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. The Plant Journal: for cell and molecular biology 54(4):733–749. https://doi.org/10.1111/j.1365-313X.2008.04347.x

Tao J, Cao C, Zhao D, Zhou C, Liang G (2010) Molecular analysis and expression of phenylalanine ammonia-lyase from poonsettia (Euphorbia pulcherrima willd). Afr J Biotechnol 10(2):126–135

Tatsuzawa F, Saito K, Nemoto T, Harada H, Honda T (1997) Acylated cyanidin glycosides in the red-purple flowers of Phalaenopsis. Phytochemistry 45(1):173–177

Wang LM, Zhang J, Dong XY, Fu ZZ, Jiang H, Zhang HC (2018) Identification and functional analysis of anthocyanin biosynthesis genes in Phalaenopsis hybrids. Biol Plant 62(1):45–54

Wang N, Song G, Zhang F, Shu X, Cheng G, Zhuang W, Wang T, Li Y, Wang Z (2023) Characterization of the WRKY Gene Family related to Anthocyanin Biosynthesis and the Regulation mechanism under Drought stress and Methyl Jasmonate Treatment in Lycoris radiata. Int J Mol Sci 24(3). https://doi.org/10.3390/ijms24032423

Wu Q, Wu J, Li SS, Zhang HJ, Feng CY, Yin DD, Wu RY, Wang LS (2016) Transcriptome sequencing and metabolite analysis for revealing the blue flower formation in waterlily. BMC Genomics 17(1):897. https://doi.org/10.1186/s12864-016-3226-9

Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39:W316–322 (Web Server issue). https://doi.org/10.1093/nar/gkr483

Yamazaki M, Gong Z, Fukushima-Mizutani M, Fukui Y, Tanaka Y, Kusumi T, Saito K (1999) Molecular cloning and biochemical characterization of a Novel anthocyanin 5-O-Glucosyltransferase by mRNA Differential Display for Plant Forms regarding antho-cyanin. J Biol Chem 274(11):7405–7411

Yan H, Pei X, Zhang H, Li X, Zhang X, Zhao M, Chiang VL, Seder-offr RR, Zhao X (2021) MYB-Mediated regulation of Anthocy-anin Biosynthesis. Int J Mol Sci 22(6). https://doi.org/10.3390/ijms22061303

Yang IS, Kim S (2015) Analysis of whole transcriptome sequencing data: Workflow and Software. Genomics Inf 13(4):119–125

Yang Y, Sun F, Zhang C (2013) Construction of full-length cDNA Library and the cDNA cloning of F3’H in Phalaenopsis. Acta Bot Boreali-Occidentalia Sinica 33(9):1731–1731

Yang Y, Wang J, Ma Z, Sun G, Zhang C (2014) De novo sequencing and comparative transcriptome analysis of white petals and red labelia in Phalaenopsis for discovery of genes related to flower color and floral differentiation. Acta Soc Bot Pol 83(3):191–199. https://doi.org/10.5586/asbp.2014.023

Yu G, Wang LG, Han Y, He QY (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 16(5):284–287. https://doi.org/10.1089/omi.2011.0118

Zhang N, Chen L, Ma S, Wang R, He Q, Tian M, Zhang L (2020) Fine mapping and candidate gene analysis of the white flower gene brfw in cabbage (Brassica rapa L). Sci Rep 10(1):6080–6080. https://doi.org/10.1038/s41598-020-63165-7

Zhao J, Dixon R, US Department of Agriculture (2010). 2015. Floriculture crops.14 summary. Washington: US Department of Agriculture. The ‘ins’ and ‘outs’ of flavonoid transport Trends Plant Sci 15 (2):72–80. https://doi.org/10.13271/j.mpb.016.004530

Zhao Y, Wang K, Wang W-l, Yin T-t, Dong W-q, Xu C-j (2019) A high-throughput SNP discovery strategy for RNA-seq data. BMC Genomics 20 (1)

Zheng X, Liu C, Qiao L, Zhao J, Han R, Wang X, Ge C, Zhang W, Zhang S, Qiao L, Zheng J, Hao C (2020) The MYB transcription factor TaPHR3-A1 is involved in phosphate signaling and governs yield-related traits in bread wheat. J Exp Bot 71(19):5808–5822. https://doi.org/10.1093/jxb/eraa355

Zhong HQ, Huang ML, Jian-Shen WU, Fang RH (2013) Cloning and expression of key enzyme genes involved in Phalaenopsis Anthocy-anins synthesis. Fujian Journal of Agricultural Sciences 36(5):284–287. https://doi.org/10.1186/s12864-016-3226-9

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.