GMR
The Original by FUNPEC-RP

PhalDB:A comprehensive database for molecular mining of the *Phalaenopsis* genome, transcriptome and miRNome

C.Y. Lee[1], K.K. Viswanath[1], J.Z. Huang[1], C.P. Lee[2], C.P. Lin[2,4], T.C. Cheng[1], B. C. Chang[2,3], S.W. Chin[1], and F.C. Chen[1]

[1] Department of Plant Industry, National Pingtung University of Science and Technology, Pingtung, Taiwan
[2] Yourgene Bioscience, Shu-Lin District, New Taipei City, Taiwan
[3] Faculty of Veterinary Science, The University of Melbourne, Parkville, Victoria, Australia
[4] Department of Biotechnology, School of Health Technology, Ming Chuan University, Gui Shan District, Taoyuan, Taiwan

Corresponding author: F.C. Chen
E-mail: fchen@mail.npust.edu.tw

Genet. Mol. Res. 17 (4): gmr18051
Received June 06, 2018
Accepted July 26, 2018
Published October 05, 2018
DOI http://dx.doi.org/10.4238/gmr18051

ABSTRACT. The orchid family Orchidaceae is one of the largest families of angiosperms; it includes approximately 35,000 species, displays a wide range of unique flower shapes and is a valued ornamental crop. Among the various species of orchids, the *Phalaenopsis* genus is the most widely commercialized among blooming potted plants. Our aim was to provide in-depth sequence information for the various parts of flowers, different time treatments of the protocorm-like bodies and temperature treated shortened stems for flowering pathway regulation, in addition to adding to the available data on normal and peloric mutants. We also obtained transcriptome data from shortened stems under cool temperature for spike induction. The deep sequenced, assembled and annotated information was integrated and built into a database that we named PhalDB. Existing databases, such as OrchidBase, OrchidBase 2.0, Orchidstra, and Orchidstra 2.0 mainly contain information on expressed sequence tags, unigenes, and microRNA only from floral organs of orchid species. These databases do not

include information about somaclonal variations and cool temperature treatment, which are important for commercial variety development. PhalDB provides sequence information from 24 samples and covers the above-mentioned tissues or conditions so that comprehensive gene data related to flower development, somaclonal variation and some horticulture traits are available for searching. A user can access DNA level information and miRNA structure, etc. It also provides an opportunity to explore mRNA level information and interactions between genes and miRNA. PhalDB is equipped with a BLAST tool to perform similarity searches among the various gene sequences.

Key words: *Phalaenopsis*; PhalDB; Next-generation sequencing; DNA; mRNA; miRNA

INTRODUCTION

Orchidaceae, the orchid family classified in class Liliopsida, order Asparagales, with five subfamilies, is one of the largest and most evolved families of the angiosperms (Cozzolino et al., 2005 and Gorniak et al., 2010). It comprises 880 genera and around 17,000-35,000 species representing about 10% of flowering families or almost 30% of the monocotyledons (Dressler 1993, Hsiao et al., 2011b and da Silva et al., 2014). It has various unique and attractive features, such as shape, color and fragrances which makes it special among all flowering plants; consequently, the demand in the floricultural industry has increased tremendously in the last few decades (Khentry et al., 2006). Other unique physiological characteristic features include large and complex polyploid genomes, long life span, slow growth, zygomorphy, specialized pollination, symbiosis with mycorrhizae, crassulacean acid metabolism (CAM) and epiphytism (Gravendeel et al., 2004, Leitch et al., 2009 and Su et al., 2011). Orchids also have numerous distinctive reproductive strategies, such as a delicate composite structure shaped by the union of the style and at least part of the pistil (gynostemium or column) and pollinia and labellum or lip (Rudall et al., 2002). This reproductive strategy of the orchid flower has great specificity for insect or other pollination vectors. Ovule development does not occur upon pollination and female meiosis only takes place about 10 weeks later to produce ovules and female gametophytes; fertilization occurs at this point (Nadeau et al., 1996). Orchids have established exotic mechanisms for radial adaptation and signify a very progressive and terminal line of floral evolution (Chen et al., 2012). Similar to other flowering plants, the progress of orchid flower development involves floral transition and subsequent initiation and formation of floral organs.

Despite their importance among angiosperms families, the genetic investigation on orchids has been comparatively limited. The large size of the genome of the orchid family signifies great evolutionary diversity compared with other plant models (Leitch et al., 2009). In recent times, next-generation sequencing technology has transformed our earlier understanding of plant genomes by producing large amounts of information, at a high pace and a continuous drop in per-base cost (Martin and Wang, 2011 and Jain

2012). The studies of the transcriptome profiling of the species with functional complexity are analyzed with novel methods, such as massively parallel sequencing technologies, digital gene expression, which offers distinctive advantages (Wang et al., 2009). The development of various methods in bioinformatics, particularly in the *de novo* assembly has sped up the study of genomics of various living organisms. In order to obtain massive sequential data, RNA-Seq, an alternative tool for gene expression profiling is extensively used for transcriptome profiling, gene discovery and molecular marker development (Hsiao et al., 2011a).

The horticulturally significant orchid genera include *Cymbidium*, *Cattleya*, *Dendrobium*, *Oncidium* and *Phalaenopsis*. The *Phalaenopsis* (moth orchid) genus consists of about 66 species distributed throughout tropical Asia, and these are prevalent ornamental plants because of their stylish appearance, the ease of production, low maintenance and prolonged flower longevity (Christenson 2001). Potted *Phalaenopsis* is one of the best outstanding orchids with more than 32,000 hybrids bred and is of great economic significance yielding more than 500 million US dollars annually for the floral industry (Lin et al., 2016). *Phalaenopsis* species have the same chromosome number (2n = 2x = 38), with chromosome sizes ranging from 1.5 to 3.5 µm and a predicted genome size of nearly 1.5 gigabases (Arends 1970 and Lin et al., 2001). *Phalaenopsis* plants generally contain three types of perianth organs with zygomorphic structure; an outer whorl with three sepals, an inner whorl with two lateral inner petals, and a labellum. The labellum normally has a distinct morphology and offers a landing platform for pollinators. The sepals and petals are usually called tepals, since they are very similar in appearance. In addition to the tepals, a highly evolved gynostemium is present in the center of the flower.

Recently the genome sequence of *Phalaenopsis* species has been sequenced and cDNA libraries, expressed sequence tags (ESTs) from cDNA-AFLP and cDNA-RAPD have been created to study gene expression (Hsu et al., 2008 and Hsiao et al., 2011a). Based on ESTs sequence information, the transcript expression patterns of floral scent products in *P. equestris* and *P. bellina* have been compared (Hsiao et al., 2006). The sequenced genome of *P. equestris* comprises 29,431 predicted protein-coding genes and provides information about the evolution of CAM photosynthesis, flowering MADS-box B-,C/D-class genes and interactions between type I MADS-box genes (Cai et al., 2015). Recently, a genomic study was performed on *Phalaenopsis* orchid tissue using Illumina sequencing technology to understand the biosynthesis of anthocyanin in two different colors (red and yellow) (Gao et al., 2016). Most of the differentially expressed genes in yellow color were associated to the biosynthesis of anthocyanins. In our previous study, we examined the regulation of flower organ development in normal type and peloric mutant *Phalaenopsis* with obtained RNA-Seq (Huang et al., 2015). After *de novo* assembly, a total of 43,552 contigs were obtained. The transcript profile was analyzed and the labellum development was proposed to be regulated by the MADS-box genes, including *PhAGL6a*, *PhAGL6b*, and *PhMADS4*. In addition to this, we generated draft genomes for winter flowering *Phalaenopsis* Brother Spring Dancer ‘KHM190’ cultivar and summer flowering species *Phalaenopsis pulcherrima* ‘B8802’ (Huang et al., 2016). It revealed alternative splicing of *PhAGL6b* that led to differential labellum development, and we also proposed gibberellin pathway regulation of the flowering expression.

Recently, differentially expressed genes were analyzed in *Phalaenopsis* using Illumina high-throughput technology for explant browning (Xu et al., 2015).

Till now, various databases have been constructed from *Phalaenopsis* orchid genomes. OrchidBase was constructed from 11 *Phalaenopsis* orchid cDNA libraries for management and analysis of EST data. Later, OrchidBase was upgraded to OrchidBase 2.0 with 1,562,071 newly added unigenes from 10 orchid species (Tsai et al., 2013). Orchidstra, and its upgraded version Orchidstra 2.0 were constructed for orchid functional genomics studies (Su et al., 2013 and Chao et al., 2017). The Orchidstra database was developed with the sequence information of five orchid species and one commercial hybrid *P. aphrodite* subsp. *formosana*. Orchidstra 2.0 has been constructed with a new database system to store protein-coding genes and non coding transcripts, covering 18 orchid species. Three databases were presented for 11 tissues of *P. equestris* from the RNA-Seq raw reads, assembled unigenes and predicted coding sequences (Niu et al., 2016).

In the floriculture industry, some common problems affecting the sale of *Phalaenopsis* cultivars occur that lead to lowered pot plant quality, including spiking traits, peloric flower mutation due to tissue culture or of a genetic nature, regeneration events from micropropagation, and factors controlling flowering during summer or spring seasons. Large-scale production of genomic information about *Phalaenopsis* gives clues to the hidden genome-scale expression. Currently, very limited information on *Phalaenopsis* genome, RNA and small RNA sequences is available. In this study, we chose to focus on the genome, transcriptome, and miRNA of the various tissues, such as peloric flowers, protocorm-like bodies (PLBs) proliferation of *Phalaenopsis* Brother Spring Dancer ‘KHM190’, and transcriptomes from shortened stems of *P. aphrodite* subsp. *formosana* under cool and warm temperature treatments at different durations. The information is stored in a user-friendly database named PhalDB. Compared with existing orchid databases PhalDB provides unique, in-depth sequential information about various parts of the peloric flower, genes up-and down-regulated at different period of PLBs proliferation, and cool temperature induced gene expression leading to spiking and variegated leaf mutants. The main reason for building PhalDB is to share integrated information about the DNA, mRNA, and miRNA from *Phalaenopsis* with researchers in order to simplify and speed studies about the functional, interacting and regulatory mechanisms of genes in orchids, identification of various *Phalaenopsis* traits and provide information about reproductive as well as vegetative phases for orchid production.

MATERIAL AND METHODS

Plant material collection and DNA isolation

The *Phalaenopsis* Brother Spring Dancer ‘KHM190’ with normal and peloric mutant types was obtained from I-Hsin Biotechnology (Chiayi, Taiwan). *Phalaenopsis aphrodite* subsp. *formosana* plants (Huang et al., 2015) were used for spike induction under cool temperature. These plants were grown under natural daylight and well-controlled temperature ranging from 27°C to 30°C in a fan-and-pad greenhouse of National Pingtung University of Science and Technology (Pingtung, Taiwan). For spike

induction, mature plants were forced at 22/18°C (day/night). Flowering plants were preserved at 26/20°C (day/night) temperature in a cooling greenhouse.

Different plant tissues, such as flowers and PLBs, were collected from the *Phalaenopsis* Brother Spring Dancer 'KHM190' (Huang et al., 2015). Three distinct floral parts, including sepal, petal, and labellum, were collected from normal and peloric mutant type plants (Figure 1). Also, PLBs produced at different time periods, cool and warm temperature treated shortened stems of *Phalaenopsis aphrodite* subsp. *formosana* (Huang et al., 2015), hybrids between *Phalaenopsis aphrodite* subsp. *formosana* and *Phalaenopsis* Dalyan, harlequin flower (Figure 2) and non-harlequin *Phalaenopsis* Pingtung Black Venus flower tissues were collected for sequencing. A total of 24 tissue samples were collected for database construction. The complete list of the plant tissues, the various parts of the flower, different time treatments of the PLBs and temperature treated shortened stems and leaves, is given in Table 1. Orchid virus (CymMV and ORSV) affected plants were examined by use of RT-PCR and excluded from the transcriptome experiment to avoid background interference. Genomic DNA was extracted from *Phalaenopsis* PLB (B190N-OX) using a standard CTAB method and this extracted DNA used as a control.

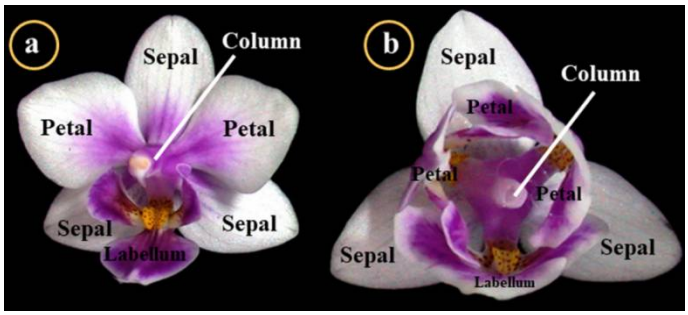

Figure 1. Flowers of the wild-type and peloric mutant of *Phalaenopsis* Brother Spring Dancer 'KHM190'. (a) Wild-type and (b) Peloric flower.

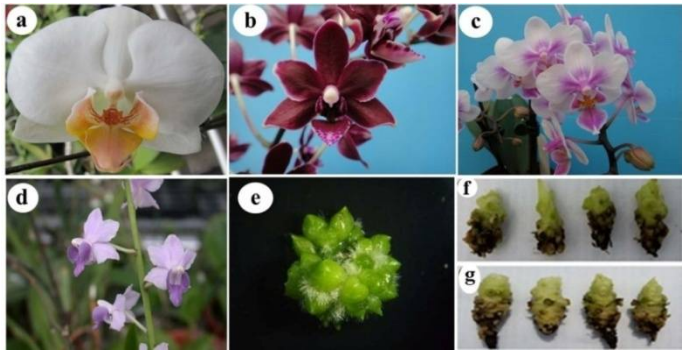

Figure 2. *Phalaenopsis* sample tissues for sequencing. (a) *Phalaenopsis* Dalyan (b) Harlequin *Phalaenopsis* Pingtung Black Venus 'NPU1332' (c) *Phalaenopsis* Brother Spring Dancer 'KHM190' (d) *P. pulcherrima* 'B8802' (e) Protocorm-like bodies (PLBs) (f) *P. aphrodite* with shoot tip tissues from shortened stems: constant high-temperature treatment at 30/27°C (day/night) (g) low-temperature treatment at 22/18°C (day/night).

Table 1. Plant sample for DNA, RNA and miRNA isolation and sequencing

Sample ID	Tissue part and conditions
100-004-D	*Phalaenopsis* Dalyan
100-004-5	*Phalaenopsis* Dalyan
98095-B	Harlequin *Phalaenopsis* Pingtung Black Venus
98095-NB	Non harlequin *Phalaenopsis* Pingtung Black Venus
B1774-NL	Normal labellum
B1774-PL	Peloric mutant labellum
B1774-NP	Normal petal
B1774-PP	Peloric petal
B1774-NS	Normal sepal
B1774-PS	Phal.Sogo Pure 'Sogo F1774'
B190-0	PLBs treated for 0 hours
B190-6	PLBs treated for 6 hours
B190-24	PLBs treated for 24 hours
B190-48	PLBs treated for 48 hours
B190N-1W	PLBs treated for 1 week
B190N-2W	PLBs treated for 2 weeks
B190N-OX	PLB for control
BH	Shorten stem at constant hightemperature (30/27 °C; day/night)
BL1	Shorten stem at cool temperature/ 1 week (22/18 °C; day/night)
BL2	Shorten stem at cool temperature/ 2 weeks (22/18 °C; day/night)
BL3	Shorten stem at cool temperature/ 3 weeks (22/18 °C; day/night)
BL4	Shorten stem at cool temperature/ 4 weeks (22/18 °C; day/night)
BN	Normal leaves
BW	Variegated leaves

DNA was isolated from control sample B190N-OX (PLB)
BH, BL1~BL4, transcriptomes from *P. aphrodite* subsp. *formosana*

Small RNA samples for sequencing: B1774-NL, B1774-PL, B1774-NP, B1774-PP, B1774-NS, B1774-PS, B190-0, B190-6, B190-24, B190-48, B190-48, B190N-1W, B190N-2W, B190N-OX,BH, BL1, BL2, BL3, BL4, BN, BW.

20 small RNA samples

B1774-NL	B1774-PL	B1774-NP	B1774-PP	B1774-NS	B1774-PS
B190-0	B190-6	B190-24	B190-48	B190N-1W	B190N-2W
B190N-OX	BH	BL1	BL2	BL3	BL4
BN	BW				

24 RNA samples

100-004-D	100-004-5	98095-B	98095-NB	B1774-NL	B1774-PL
B1774-NP	B1774-PP	B1774-NS	B1774-PS	B190-0	B190-6
B190-24	B190-48	B190N-1W	B190N-2W	B190N-OX	BH
BL1	BL2	BL3	BL4	BN	BW

Sequencing, quality control, and draft genome construction

In our previous study, we examined sequencing and quality checking of *Phalaenopsis*, and construction of draft genome (Huang et al., 2016). Here explained briefly, to sequence the *Phalaenopsis* genome from extracted DNA, we applied next-generation sequencing technologies using the Illumina HiSeq 2000 platform. Illumina short-insert paired-end (insert size: 250 bp) and large-insert mate-pair (3, 5, and 8 kb) libraries were prepared following the manufacturer's instructions. In total, we generated approximately 300.5 Gb of sequences, and 278.89 Gb were retained for assembly after performing quality trimming using CLC Genomic Workbench 5.5 (http://www.clcbio.com/) to filter out low-quality reads.

Estimation of genome size through *K*-mer analysis

Genome size was estimated with an accepted technique based on the *K*-mer distribution. We used high-quality reads (215.9 Gb) from short-insert size libraries (250 bp) to obtain accurate estimates. The occurrence of each *K*-mer was estimated from the genomic reads, and the *K*-mer frequencies followed a Poisson distribution for a deep-sequenced genome. Thus the genome size G is calculated as G = Knum / Kdepth, where Knum is the total number of *K*-mers, and Kdepth is the highest peak detected. K was set to 17 in our project based on our empirical analysis. In this work, K was 17; Knum was 175,493,961,632; and Kdepth was 50. We, therefore, estimated the *Phalaenopsis* 'KHM190' genome size to be 3.45 Gb.

De novo assembly and dataset annotation

In our previous report, we explained about sequence assembly and dataset annotation (Huang et al., 2016). Here we performed whole-genome assembly using Velvet (https://www.ebi.ac.uk/~zerbino/velvet/) (Zerbino et al., 2008) with *K*-mer 63. In order to fill the gaps inside the constructed scaffolds, we used Gapcloser to reduce the N ratio in the final assembly (http://soap.genomics.org.cn/down/GapCloser_release_2011.tar.gz). The total contig length of the assembly reached 2.39 Gb (69.28% of 3.45 Gb), with an N50 length of 1.49 kb (longest, 50.94 kb), and the genome assembly was 3.1 Gb, with a scaffold N50 length of 100.94 kb (longest, 1.4 Mb). Scaffolds with lengths greater than 10 kb accounted for more than 74.8% of the assembly. The assembled genome of *Phalaenopsis* was validated using 8,188 Sanger-derived ESTs for *Phalaenopsis* downloaded from NCBI and was aligned with the assembly using BLAST with the default parameters. As a result, 7,701 genes (95% identity and over 50% coverage) were matched to the *de novo Phalaenopsis* assembly. The validation procedure gave 6,928 genes in the *Phalaenopsis* genome assembled with stringent parameters (95% identity and 90% coverage). Thus, the draft sequences represent a considerable portion of the *Phalaenopsis* genome, with high quality and good coverage.

Recognition of differentially expressed transcripts

The expression profile of raw transcripts was examined using gapped alignment mode of the program Bowtie 2 (Langmead and Salzberg, 2012). First, the trimmed reads to raw transcript sequence were mapped; after alignment, raw transcript expression was quantified with the eXpress 1.3.0 software (Roberts et al., 2013). The value of reading counts from eXpress would be the input of DESeq.

RNA isolation and cDNA synthesis

We isolated total RNA from sepal, petal, and labellum of the flower, different time treated PLBs during proliferation, and cool and warm temperature treated shortened stems. A total of 24 orchid samples were collected and stored at -80°C for further analysis (Table 1). Total RNA was extracted from the frozen orchid tissues by the

TriSolution method (GeneMark, Taipei). To eliminate contaminating DNA, RNA samples were treated with RNase-free DNase I (Promega, Taipei). After assessment of RNA quality, samples with a quality indicator (RQI) > 8 were used for mRNA purification and cDNA synthesis at Yourgene Bioscience (New Taipei City, Taiwan).The cDNA library was constructed according to the manufacturer's protocol using a Truseq RNA sample prep kit (Illumina, San Diego, CA,USA). An aliquot of 5 μg mRNA was directly fragmented after the oligo-dT purification step. Next, first-strand cDNA was synthesized with mRNA as the template by priming with random hexamer primers and then second-strand cDNA was synthesized. After purification, cDNA underwent end repair, A-tailing, Illumina adaptor ligation, size-selection of the range 320–420 bp (approximately insert size 200–300 bp) and finally for enriched the cDNA PCR amplification for 15 cycles was run. The products were loaded onto flow cell channels at 12 pM for pair-end 100 bp×2 sequencing with the Illumina HiSeq2000 platform (Yourgene Bioscience, New Taipei City, Taiwan).

RNA-Seq mapping and transcript reconstruction

To annotate transcriptionally active regions of the *Phalaenopsis* genome, RNAs from 24 different tissues were sequenced using Illumina transcriptome sequencing technology. The 100-bp paired ends of the samples were pooled, and each sample dataset was aligned against the library-based repeat-masked assembly of *Phalaenopsis* using Bowtie 2 (Langmead and Salzberg, 2012) and TopHat (v2.0.8b) (Trapnell et al., 2009) with the default settings and the previously determined mean inner distance between mate pairs. We utilized TopHat to identify exon-intron splicing junctions and refine the alignment of the RNA-Seq reads to the genome. Cufflinks software (v2.1.1) (Trapnell et al., 2012 and Pollier et al., 2013) was then employed to define a final set of predicted genes. Using the RNA-Seq approach, we predicted 54,659 gene loci and 76,370 spliced transcripts in the assembly.

miRNA analysis

miRNA library development and sequencing

To develop the miRNA library, we used 10 μg of total RNA obtained from *Phalaenopsis* floral organs (sepal, petal, and labellum), shortened stems and PLBs as the initial input for library construction. Following 15% polyacrylamide denaturing gel electrophoresis, the small RNA fragments with lengths in the range of 16–32 nt were isolated from the gel and purified. Next, a miRNA library was prepared with the Illumina TruSeq Small RNA Sample Prep Kit. Finally, the miRNA library was sequenced by using the Illumina HiSeq 2000 platform at Yourgene Bioscience in Taiwan.

miRNA gene prediction

To delete low-quality reads, adapters, and contamination from the raw sequencing data we used Perl scripts. After filtration of the raw data, the clean reads were aligned against plant repeat databases (http://rfam.sanger.ac.uk/ and http://plantrepeats.plantbiology.msu.edu/) using Bowtie 2 software (Langmead and Salzberg, 2012) to dispose of numerous non-coding RNAs. The residual unique filtered sequences were then compared with known mature and precursor miRNAs (pre-miRNAs) from other plant species deposited in miRBase 19 (Kozomara et al., 2013)(http://www.mirbase.org/) using Bowtie software to search the conserved miRNAs. We used miRDeep2 (Friedländer et al., 2011) and INFERNAL (v 1.1) software (Nawrocki et al., 2013) to predict miRNA precursor sequences from the sequenced small RNAs. Later the putative target sites of the miRNA candidates were identified by aligning the miRNA sequences with the assembled ESTs of *Phalaenopsis* using Bowtie software.

RESULTS AND DISCUSSION

Database contents

PhalDB development and architecture

The database was developed using dancer, a light-weight Perl web framework and by Bootstrap with a front-end component library and MySQL relational database on a Linux platform. The database we constructed is named PhalDB because it contains a uniting database system, a Windows application that performs genome, transcriptome and miRNA data analysis of *Phalaenopsis* orchids. The comprehensive library descriptions together with a number of sequences can be openly queried from the website of the database. All identified nucleotide sequences in fasta format, blast results, annotations were stored in a user-friendly database. PhalDB is constructed to store and search the vast amount of DNA, mRNA and miRNA sequence information consistent annotation in every library built distinctly from numerous cDNA libraries through Solexa Illumina sequencers and explored through the web application. The complete flowchart of the database development is illustrated in Figure 3. The exploration of deep sequencing results of DNA, mRNA, and miRNA for *Phalaenopsis* Brother Spring Dancer ‘KHM190’ and other samples provides information about the genome, transcriptome and putative target genes for this species. The entire database contents were classified into four different search categories, including DNA, mRNA, miRNA by BLAST search and mRNA by gene annotation (Figure 4).

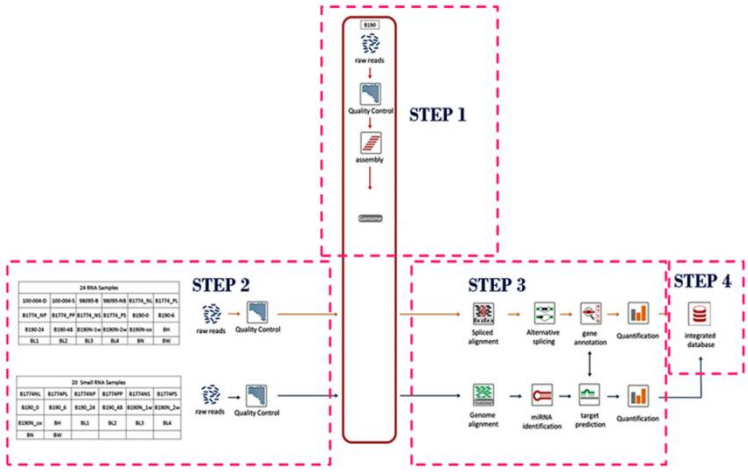

Figure 3. The schematic pipeline of the DNA, mRNA, and miRNA assembly of PhalDB construction. PhalDB is a comprehensive database for browsing and retrieving resources about functional genomics and associated data of *Phalaenopsis* accessions.

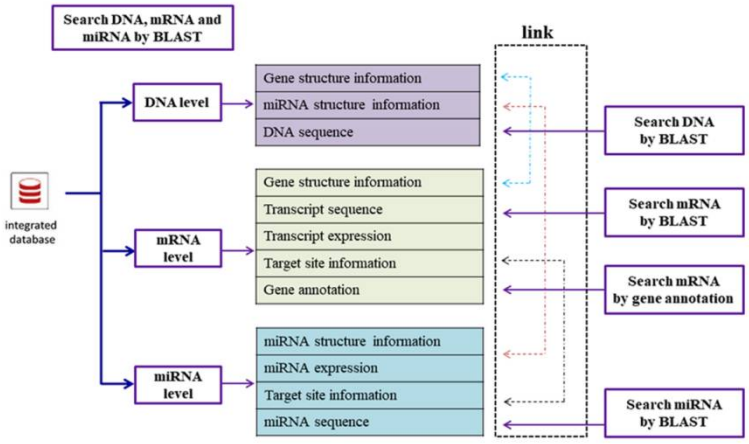

Figure 4. The schematic data relationship of PhalDB. It provides the information about DNA, mRNA and miRNA with regard to flower development, leaf variegation, PLBs proliferation and spike induction of *Phalaenopsis* orchids.

Raw read processing and *de novo* assembly (Figure 3, step 1 and step 2)

The draft genome was produced by solely using the Illumina HiSeq 2000 platform performed at Yourgene Bioscience (Step 1). A total of 149,151 scaffolds were obtained after assembly of sequences. The cDNA library for transcriptome sequencing was constructed using Illumina HiSeq 2000 platform. An RNA-Seq has been applied for sequencing of sepal, petal and lip tissues of both the wild-type and peloric mutant (Table 2); the Illumina Genome Analyzer II system was used to generate 100 bp raw pair-end reads. To avoid a negative impact on analyses, adapter and low-quality sequences were

pre-processed from the raw read pairs using the Cross Match software. After pre-processing, Velvet, a short-read assembly method was used with the default setting for assembling of processed sequences into consensus sequences (Table 3). After assembly, 752,203 transcripts were produced, from these assembled transcripts after choosing the highest confidence score of a locus (with at least 50% of overlapping bases and at least 95% identity); finally, 43,552 contigs were obtained using CLC genomic workbench (step 2).

Table 2. Transcriptome analysis of four organs of *Phalaenopsis* via RNA-Seq

Organ (Gb)	Usable data (bp)	Transcripts (bp)	Average length (bp)	Maximum length	Total size of transcripts
Shorten stems	35.2	56,609	914	19,384	51,769,072
Floral organs	32.5	40,192	1,081	17,075	43,464,697
Leaves	11.9	43,719	504	4,720	22,054,235
PLB	9.9	61,736	653	5,042	40,324,120

Shorten stems: Constant high temperature (BH) and a cool temperature (BL) (1 to 4 weeks)
Floral organs: sepal, petal and lip tissues of both the 'KHM190' wild-type and peloric mutant
Leaves: *Phalaenopsis aphrodite* wild-type leaf and *Phalaenopsis aphrodite* mutant with leaf intervein chlorosis.
Protocorm-like body: *Phalaenopsis* Brother Spring Dancer 'KHM190' PLB after cutting and growth on induction medium for 0 weeks and Primary *Phalaenopsis* 'KHM190' PLB after cutting and growth on induction medium for 2 weeks.

Table 3. Summary of *Phalaenopsis* floral-organ transcriptome assembly

Total number of transcripts	NS	PS	NP	PP	NL	PL
Raw data						
Total no. of reads	39,281,522	63,691,838	43,899,068	51,106,016	81,707,498	67,469,926
Total nucleotides (nt)	3,967,433,722	6,432,875,638	4,433,805,868	5,161,707,616	8,252,457,298	6,814,462,526
High-quality reads	36,697,424	58,805,774	41,084,182	48,504,500	78,458,762	65,091,470

Functional annotation (Figure 3, step 3)

Structural RNAs including 562 rRNAs, 655 tRNAs, 290 and snoRNAs and 263 snRNAs were identified and removed from mRNA with Rfam (http://rfam.sanger.ac. uk/). We obtained 6,976,375 unique small RNA (sRNA) tags from 92,811,417 sRNA raw reads ranging from 18 to 27 bp, among which the 24 nt category was the most abundant type of small RNA (34.59%). After assembly, mRNA and miRNA sequences with high quality were used for functional annotation. We predicted 41,153 genes with an average length of 1,014 bp; the putative functional annotation of identified genes was executed using BLASTX search against the NCBI nr database. After annotation procedures, explanation of the best BLAST hit was allocated to every protein-coding mRNA that met the specified threshold. A total of 650 miRNAs dispersed in 188 families were recognized, compared with the NCBI nr database and 1,644 miRNA-targeted genes were predicted. The annotation of miRNAs from *Phalaenopsis* Brother Spring Dancer 'KHM190' and *P. aphrodite* subsp. *formosana* (BL1-BL4) through the deep sequencing results provides precursors and putative target genes for this species.

Information visualization (step 4)

PhalDB makes available data on the processed sequence status, functional annotation and for curating the biological meaning assigned to all DNA, mRNA and miRNA sequences.

System requirements

Browser compatibility of PhalDB was implemented to support most of the most widely-used browsers, including Chrome, Safari, Opera and Mozilla Firefox. For best view and functionality of PhalDB, we recommend an upgraded version of these browsers.

Information retrieval system

The PhalDB is a useful, web-accessible interactive database. It offers information about DNA, mRNA, and miRNA sequences (Figure 5).

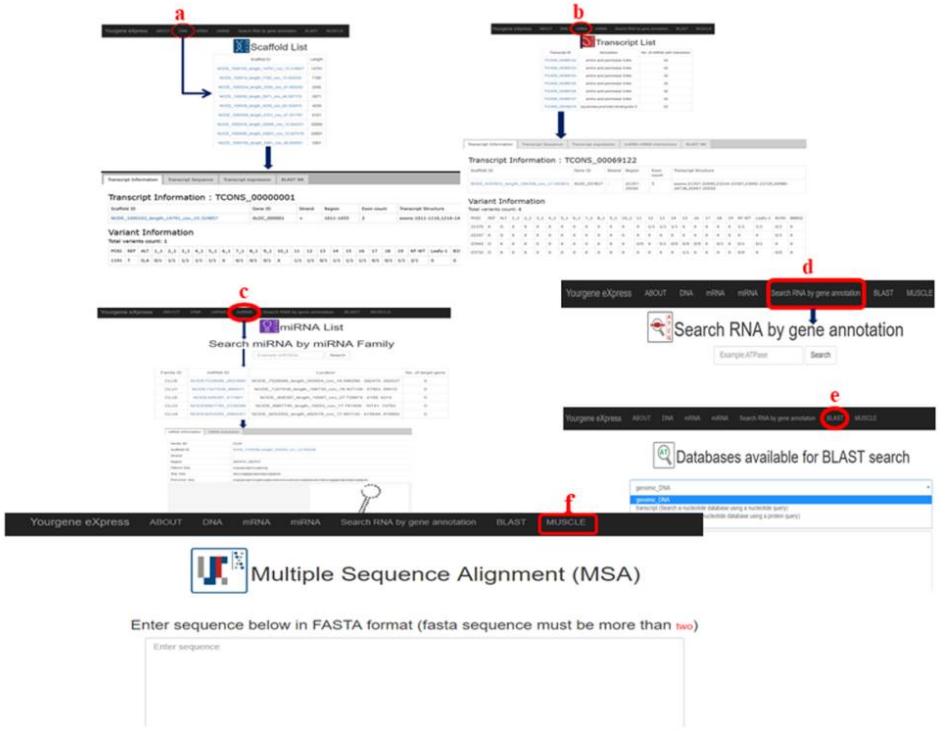

Figure 5. Architecture and snapshots of the PhalDB web interface. (a) Sequence reports, search DNA page, (b) Sequence reports, search transcripts, (c) Sequence reports, search miRNA, (d) Search RNA by gene annotation, (e) Database for searching of genomic DNA, a nucleotide search using a nucleotide query, translated nucleotide using a protein query and miRNA, and (f) a MUSCLE tab for the Multiple Sequence Alignment for nucleotide or protein sequences.

I. Sequence reports, search DNA: The database pages enable listing of the DNA sequences. This page simply lists the scaffold of all analyzed DNA data sets. Each scaffold consists of a length of DNA with scaffold ID numbers. Users may click on a scaffold list that shows gene ID, transcript ID, region and exon count of DNA. Users click on transcript ID that gives variant and sequence information, and the expression level of the transcript. Nucleotide BLAST shows similarities and identified sequences of various plants along with their accession numbers.

II. Sequence reports, search mRNA: This tab contains the list of the transcript, scaffold and gene IDs, variant and transcript information, transcript sequence and expression. This page provides information about the interaction between miRNA-mRNA. Nucleotide BLAST shows similarities and identified sequences of various plants along with their accession numbers.

III. Sequence reports, search miRNA: This tab comprises the list of miRNA list with their family ID, miRNA ID, and location. Click on miRNA ID or search tool to provide information and miRNA expression.

IV. Sequence similarity search: In order to perform the sequence similarity searches PhalDB is also equipped with a number of tools, such as BLASTN, TBLASTN, and TBLASTX. The Blast output gives data about the DNA, RNA and contig names, and sorts by e-value and sequence ID.

CONCLUSIONS

In the last decade, high-throughput technologies have been developed to extensively study the genome, transcriptome, small RNAs and their regulatory mechanisms in many organisms. The Orchidaceae is one of the most species-rich plant families. Among the various orchid species, *Phalaenopsis* is the top-traded ornamental potted plants in the world. We obtained sequencing information from flowers, PLBs of *Phalaenopsis* Brother Spring Dancer ‘KHM190’, shortened stems of *P. aphrodite* subsp. *formosana* and other accessions with different flowering behaviors by using high-throughput technologies. In order to share our data results of DNA, mRNA, and miRNA, we combined the information from each step of our exploration into a web-based database and named it PhalDB. The present database PhalDB is developed by Dancer, a light-weight Perl web framework and by Bootstrap with front-end component library and MySQL relational database on a Linux platform. We enriched these data with essential information that may enlighten and support further research on various aspects of orchid biology. Compared to other existing orchid databases, PhalDB contains significant sequence information about the genome, transcriptome and miRNA of various parts of a flower, different time treatments of the PLBs and temperature treated shortened stems and other sample accessions with a friendly interface to query or download data. The data provided in PhalDB should help improve knowledge on orchid flowering regulation, flower development, micropropagation through PLB proliferation, and novel breeding strategies. In the future, we will expand our database by integrating more data with small RNAs, mutations and SNPs for further analysis of orchids.

Database weblink: http://www.yourgenebio.com/cylee2014/index
Username: Viswanath
Password: 612812

ACKNOWLEDGMENTS

This work was supported by grants from the Council of Agriculture, Agriculture and Food Agency (grant numbers 105AS-9.1.1-FD-Z1, 106AS-8.6.3-FD-Z1, 107AS-7.6.3-FD-Z2) and the Ministry of Science and Technology (grant numbers MOST 105-2321-B-020-003, MOST 106-2321-B-020-002).

CONFLICT OF INTEREST

The authors have declared that they have no conflict of interest.

REFERENCES

Arends J (1970). Cytological observations on genome homology in eight interspecific hybrids of Phalaenopsis. *Genetica*. 41:88-100. https://doi.org/10.1007/BF00958896

Cai J, Liu X, Vanneste K, Proost S, et al. (2015). The genome sequence of the orchid Phalaenopsis equestris. *Nat. Genet*. 47:65-72. doi:10.1038/ng.3149

Chao Y-T, Yen S-H, Yeh J-H, Chen W-C, et al. (2017). Orchidstra 2.0—a transcriptomics resource for the orchid family. *Plant Cell Physiol*. 58:e9-e9. https://doi.org/10.1093/pcp/pcw220

Chen Y-Y, Lee P-F, Hsiao Y-Y, Wu W-L, et al. (2012). C-and D-class MADS-box genes from Phalaenopsis equestris (Orchidaceae) display functions in gynostemium and ovule development. *Plant Cell Physiol.* 53:1053-1067. https://doi.org/10.1093/pcp/pcs048

Christenson EA (2001). *Phalaenopsis*: a monograph. Portland Or.: Timber Press. 330p, ISBN 881924946

Cozzolino S and Widmer A (2005). Orchid diversity: an evolutionary consequence of deception?. *Trend Ecol Evol.* 20:487-494. doi:10.1016/j.tree.2005.06.004

da Silva JAT, Aceto S, Liu W, Yu H, et al. (2014). Genetic control of flower development, color and senescence of Dendrobium orchids. *Sci. Hortic*. 175:74-86. https://doi.org/10.1016/j.scienta.2014.05.008

Dressler RL (1993). Phylogeny and classification of the orchid family. Cambridge University Press.

Friedländer MR, Mackowiak SD, Li N, Chen W, et al. (2011). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* 40:37-52. https://doi.org/10.1093/nar/gkr688

Gao L, Jiang D, Yang Y, Li Y, et al. (2016). *De novo* sequencing and comparative analysis of two *Phalaenopsis* orchid tissue-specific transcriptomes. *Russ. J. Plant Physiol.* 63:391-400. https://doi.org/10.1134/S1021443716020072

Gorniak M, Paun O, Chase MW (2010). Phylogenetic relationships within Orchidaceae based on a low-copy nuclear coding gene, Xdh: Congruence with organellar and nuclear ribosomal DNA results. *Mol Phylogenet Evol.* 56:784-795. doi:10.1016/j.ympev.2010.03.003

Gravendeel B, Smithson A, Slik FJ, Schuiteman A (2004). Epiphytism and pollinator specialization: drivers for orchid diversity? *Philos. Trans. R. Soc. Lond. B Biol. Sci*. 359:1523-1535. DOI: 10.1098/rstb.2004.1529

Hsiao Y-Y, Tsai W-C, Kuoh C-S, Huang T-H, et al. (2006). Comparison of transcripts in *Phalaenopsis bellina* and *Phalaenopsis equestris* (Orchidaceae) flowers to deduce monoterpene biosynthesis pathway. *BMC Plant Biol.* 6:14. https://doi.org/10.1186/1471-2229-6-14

Hsiao YY, Chen YW, Huang SC, Pan ZJ, et al. (2011a). Gene discovery using next-generation pyrosequencing to develop ESTs for *Phalaenopsis orchids*. *BMC Genomics.* 12:360. doi:10.1186/1471-2164-12-360

Hsiao YY, Pan ZJ, Hsu CC, Yang YP, et al. (2011b). Research on orchid biology and biotechnology. *Plant Cell Physiol.* 52:1467-1486. doi:10.1093/pcp/pcr100

Hsu T-W, Tsai W-C, Wang D-P, Lin S, et al. (2008). Differential gene expression analysis by cDNA-AFLP between flower buds of *Phalaenopsis* Hsiang Fei cv. HF and its somaclonal variant. *Plant Sci*. 175:415-422. https://doi.org/10.1016/j.plantsci.2008.06.010

Huang J-Z, Lin C-P, Cheng T-C, Chang BC-H, et al. (2015). A *de novo* floral transcriptome reveals clues into *Phalaenopsis* orchid flower development. *PLoS One.* 10:e0123474. doi.org/10.1371/journal.pone.0123474

Huang J-Z, Lin C-P, Cheng T-C, Huang Y-W, et al. (2016). The genome and transcriptome of *Phalaenopsis* yield insights into floral organ development and flowering regulation. *PeerJ*. 4:e2017. https://doi.org/10.7717/peerj.2017

Jain M (2012). Next-generation sequencing technologies for gene expression profiling in plants. *Brief Funct Genomics*. 11:63-70. doi:10.1093/bfgp/elr038

Khentry Y, Paradornuwat A, Tantiwiwat S, Phansiri S, et al. (2006). Incidence of Cymbidium mosaic virus and Odontoglossum ringspot virus in Dendrobium spp. in Thailand. *Crop Prot.* 25:926-932. https://doi.org/10.1016/j.cropro.2005.12.002

Kozomara A and Griffiths-Jones S (2013). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 42:D68-D73. https://doi.org/10.1093/nar/gkt1181

Langmead B and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 9:357-359. doi:10.1038/nmeth.1923

Leitch I, Kahandawala I, Suda J, Hanson L, et al. (2009). Genome size diversity in orchids: consequences and evolution. *Ann Bot*. 104:469-481. https://doi.org/10.1093/aob/mcp003

Lin B-Y, Chang C-D, Huang L, Liu Y-C, et al. (2016). The mitochondrial DNA markers for distinguishing *Phalaenopsis* species and revealing maternal phylogeny. *Biol. Plant*. 60:68-78. https://doi.org/10.1007/s1053

Lin S, Lee H-C, Chen W-H, Chen C-C, et al. (2001). Nuclear DNA contents of *Phalaenopsis* sp. and *Doritis pulcherrima*. *J. Am. Soc. Hortic. Sci*. 126:195-199

Martin JA and Wang Z (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12:671-682. doi:10.1038/nrg3068

Nadeau JA, Zhang XS, Li J, O'Neill SD (1996). Ovule development: identification of stage-specific and tissue-specific cDNAs. *Plant Cell*. 8:213-239. https://doi.org/10.1105/tpc.8.2.213

Nawrocki EP and Eddy SR (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 29:2933-2935. doi:10.1093/bioinformatics/btt509

Niu S-C, Xu Q, Zhang G-Q, Zhang Y-Q, Tsai W-C, et al. (2016). *De novo* transcriptome assembly databases for the butterfly orchid *Phalaenopsis equestris*. *Scientific Data.* 3:160083. doi:10.1038/sdata.2016.83

Pollier J, Rombauts S, Goossens A (2013). Analysis of RNA-Seq data with TopHat and Cufflinks for genome-wide expression analysis of jasmonate-treated plants and plant cultures. *Methods Mol. Biol*. 1011: 305-15. doi: 10.1007/978-1-62703-414-2_24

Roberts A and Pachter L (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods*. 10:71-73. doi:10.1038/nmeth.2251

Rudall PJ and Bateman RM (2002). Roles of synorganisation, zygomorphy and heterotopy in floral evolution: the gynostemium and labellum of orchids and other lilioid monocots. *Biol. Rev*. 77:403-441. https://doi.org/10.1017/S1464793102005936

Su C-l, Chao Y-T, Alex Chang Y-C, Chen W-C, et al. (2011). *De novo* assembly of expressed transcripts and global analysis of the *Phalaenopsis aphrodite* transcriptome. *Plant Cell Physiol*. 52:1501-1514. https://doi.org/10.1093/pcp/pcr097

Su C-l, Chao Y-T, Yen S-H, Chen C-Y, et al. (2013). Orchidstra: an integrated orchid functional genomics database. *Plant Cell Physiol.* 54:e11-e11. https://doi.org/10.1093/pcp/pct004

Trapnell C, Pachter L, Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 25:1105-1111. https://doi.org/10.1093/bioinformatics/btp120

Trapnell C, Roberts A, Goff L, Pertea G, et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc*. 7:562-578. doi:10.1038/nprot.2012.016

Tsai W-C, Fu C-H, Hsiao Y-Y, Huang Y-M, et al. (2013). OrchidBase 2.0: comprehensive collection of Orchidaceae floral transcriptomes. *Plant Cell Physiol.* 54:e7-e7. https://doi.org/10.1093/pcp/pcs187

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*.10:57-63. doi:10.1038/nrg2484

Xu C, Zeng B, Huang J, Huang W, et al. (2015). Genome-wide transcriptome and expression profile analysis of *Phalaenopsis* during explant browning. *PLoS One.* 10:e0123356. https://doi.org/10.1371/journal.pone.0123356

Zerbino DR and Birney E (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18:821-829. doi:10.1101/gr.074492.107