PCP
PLANT & CELL PHYSIOLOGY

De Novo Assembly of Expressed Transcripts and Global Analysis of the *Phalaenopsis aphrodite* Transcriptome

Chun-lin Su[1,4], Ya-Ting Chao[2,4], Yao-Chien Alex Chang[3,4], Wan-Chieh Chen[1], Chun-Yi Chen[1], Ann-Ying Lee[1], Kee Tuan Hwa[1] and Ming-Che Shih[1,*]

[1]Agricultural Biotechnology Research Center, Academia Sinica, Taipei, 11529, Taiwan
[2]Department of Computer Science and Engineering, Yuan Ze University, Chungli, Taoyuan, 32003, Taiwan
[3]Department of Horticulture, National Taiwan University, Taipei, 10617, Taiwan
[4]These authors contributed equally to this work
*Corresponding author: E-mail, mcshih@gate.sinica.edu.tw; Fax, +886-2-26515693
(Received April 23, 2011; Accepted July 12, 2011)

Being one of the largest families in the angiosperms, Orchidaceae display a great biodiversity resulting from adaptation to diverse habitats. Genomic information on orchids is rather limited, despite their unique and interesting biological features, thus impeding advanced molecular research. Here we report a strategy to integrate sequence outputs of the moth orchid, *Phalaenopsis aphrodite*, from two high-throughput sequencing platform technologies, Roche 454 and Illumina/Solexa, in order to maximize assembly efficiency. Tissues collected for cDNA library preparation included a wide range of vegetative and reproductive tissues. We also designed an effective workflow for annotation and functional analysis. After assembly and trimming processes, 233,823 unique sequences were obtained. Among them, 42,590 contigs averaging 875 bp in length were annotated to protein-coding genes, of which 7,263 coding genes were found to be nearly full length. The sequence accuracy of the assembled contigs was validated to be as high as 99.9%. Genes with tissue-specific expression were also categorized by profiling analysis with RNA-Seq. Gene products targeted to specific subcellular localizations were identified by their annotations. We concluded that, with proper assembly to combine outputs of next-generation sequencing platforms, transcriptome information can be enriched in gene discovery, functional annotation and expression profiling of a non-model organism.

Keywords: Assembly • Database • Expression profiling • *Phalaenopsis* • Transcriptome.

Abbreviations: BLAST, basic local alignment search tool; CAM, crassulacean acid metabolism; DAS, days after sowing; DSN, duplex-specific nuclease; ER, endoplasmic reticulum; EST, expressed sequence tag; FPKM, fragments per kilobase of exon per million fragments mapped; GFP, green fluorescent protein; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; NGS, next-generation sequencing; NLS, nuclear localization signal; SNP, single nucleotide polymorphism; TKL, tyrosine kinase like.

The nucleotide sequences of raw reads and contigs from this work were deposited in GenBank (NCBI accession number: SRA030409; contigs: JI626343–JI831113).

Introduction

Being the largest family of angiosperms with >25,000 species (Pridgeon et al. 2005), the Orchidaceae contain members with many unique physiological characteristics that are absent in more popularly studied model plants, such as rice and Arabidopsis. The great biodiversity of orchids is displayed not only in terms of speciation but also in their vast geographical distribution, resulting in variations in morphology, physiology, habitat and interaction with the ecosystem. Great efforts have been made to accumulate valuable genomic information for model organisms rapidly. Orchids, like many other non-model organisms, have received less attention from researchers by comparison. An effective workflow to generate and accumulate genomic data quickly is critical to facilitate in-depth research of non-model organisms.

Orchids have some conserved characteristics that distinguish them from other plants, such as floral zygomorphy (Rudall and Bateman 2002), a fused stamen and pistil (Rudall and Bateman 2002), aggregation of pollen into pollinia (Dressler 1990), and mycotrophy at an early stage of life (Rasmussen 2002). Orchid pollination often involves complex mechanisms, which had been described as early as Darwinian times (Darwin 1862) and have been researched extensively ever since (Nilsson 1992, van der Cingel 2007, Micheneau et al. 2009). Palaeontological evidence of orchid–insect interaction was found in amber fossil dating back 15–20 million years (Ramirez et al. 2007). Based on the photosynthetic patterns,

Plant Cell Physiol. 52(9): 1501–1514 (2011) doi:10.1093/pcp/pcr097, available online at www.pcp.oxfordjournals.org

orchids can be broadly divided into C_3 and crassulacean acid metabolism (CAM) plants. Closely related orchid taxa may have different photosynthetic patterns: thick-leaved *Oncidium* species carry out CAM photosynthesis, while the photosynthetic pattern of thin-leaved *Oncidium* species is C_3 (Silvera et al. 2010). It is of particular interest that some orchids such as *Phalaenopsis* shift their photosynthetic pattern between growth stages or when facing environmental changes (Guo and Lee 2006, Ping et al. 2010). The prevalence of CAM photosynthesis with epiphytism at low altitude under the canopy of a rainforest indicated a functional relationship between traits under correlated evolutionary divergence (Silvera et al. 2009). Research in functional genomics and the genome structure of orchids may shed light on the molecular regulation or evolutionary track of the photosynthetic pattern shift.

In recent years, orchids have gained commercial importance worldwide as ornamentals. With such agroeconomic significance, studying the reproductive behavior of orchids is as important as investigating the vegetative physiology described above. The reproduction of orchids is unique due to the absence of endosperm in the mature seeds, with embryo development arrested at the globular stage (Arditti 1992). Characteristics of embryo maturation only become evident in the protocorm, a unique structure that is developed after germination and found only in orchids (Vinogradova and Andronova 2002). Symbiosis with mycorrhiza is required for germination of the poorly differentiated seeds in nature (Burgeff 1959). This dependence on fungus can be eliminated through the addition of sugars to artificial culture medium (Knudson 1922). Tissue culture and seed germination in an artificial environment became available through years of development aimed at facilitating orchid breeding and propagation.

With recent advances in sequencing technologies, genome-scale sequencing projects including de novo transcriptome analysis of many emerging model organisms (Shin et al. 2008, Meyer et al. 2009, Wakaguri et al. 2009) and reference mapping of expressed transcripts (Nagalakshmi et al. 2008) were launched. High-throughput technologies, e.g. RNA-Seq, were also applied to functional research such as gene expression profiling of rice anther (Huang et al. 2009) and overall transcriptome analysis with the combination of tiling array in *Drosophila* (Graveley et al. 2011). However, very few functional genomic studies in orchids have been conducted so far. Compared with model plants, orchids have large genome sizes and longer life cycles, thus making them more challenging subjects in genetic and genomic works. Previously, several sequencing efforts have been made to provide a collection of orchid genes on a genomic scale. In one study, conventional capillary sequencing was employed to identify expressed sequence tags (ESTs) of *Phalaenopsis equestris*, resulting in a relatively low output of 3,688 clustered unigenes (Tsai et al. 2006). Very recently, results from multiple sequencing techniques were integrated to generate 8,501 contigs and 76,116 singletons for *Phalaenopsis* spp. (Fu et al. 2011). Nevertheless, the number of orchid genes in current databases is still not enough to support further functional genomics studies when compared with other plant species.

Because *Phalaenopsis aphrodite* (moth orchid) is an important parent in commercial breeding programs for desirable traits such as large and white flowers, it was selected as our model orchid for functional genomic research. It is native to Taiwan and epiphytic with CAM photosynthesis. To investigate expressed genes involved in many biological processes of orchids, we constructed cDNA libraries from various organs of *P. aphrodite*, including root and shoot tissues, germinating seeds and flowers, at different developmental stages. An efficient workflow was developed using data generated from different high-throughput sequencing platforms to maximize EST output in order to increase the number and length of assembled contigs and to reduce sequencing bias.

Next-generation sequencing (NGS) platforms, such as Roche's 454 GS FLX, Illumina/Solexa Genome Analyzer and Applied Biosystems' SOLiD, have been widely used in recent years for high-throughput sequencing of many organisms (Wall et al. 2009, Metzker 2010). In general, Roche's sequencing technology produces long reads and is advantageous for assembly of sequences into longer contigs; however, the number of reads generated in each run is lower than that of other platforms and not enough to reach deep coverage for low-abundant genes. This technology also has shortcomings in decoding homopolymeric nucleotide tracts correctly. Additional problems identified in our study were sequencing bias, the number of uncalled bases and a software glitch that produced duplicated contigs. The Solexa technology provides a high number of reads for deeper coverage, which is beneficial for gene discovery. However, its short read length limits de novo contig assembly efficiency. Discussions on sequencing bias of high-throughput technologies have taken place in several publications describing bias such as base substitution caused by the amplification procedure of Solexa (Dohm et al. 2008, Hillier et al. 2008) or IN/DELs in homopolymeric repeat regions caused by the chemistry used in Roche 454 (Quinlan et al. 2008). Proper integration of data to minimize sequencing bias by either technology will assist in reducing sequence mistakes.

Assembled contigs from sequencing reads were annotated and archived into an orchid-specialized database, Orchidstra (http://orchidstra.abrc.sinica.edu.tw), which is publicly available. Potential applications of this database are not limited to *Phalaenopsis* since the establishment of an orchid genome database will provide invaluable references for future works in other orchid genera.

Results

Generation of transcriptome information

Phalaenopsis aphrodite cDNA samples were prepared from vegetative tissues (mixture of root, leaf and stem), reproductive tissues (young inflorescence as reproductive library 1; flower buds and open blossom as reproductive library 2) and

germinating seeds [0–30 days after sowing (DAS) as seed library 1; 40–75 DAS as seed library 2, and >75 DAS as seed library 3]. Due to expression redundancy, vegetative cDNA library was normalized by the duplex-specific nuclease (DSN) method to facilitate novel gene discovery (Zhulidov et al. 2004). The cDNA samples were sequenced using the 454 GS FLX Titanium and the Illumina/Solexa Genome Analyzer II platforms. Pooling of the 454 sequencing runs resulted in 3,302,528 qualified reads with an average length of 307 bases after adaptor trimming, i.e. about 1 Gb of sequence data in total (**Supplementary Table S1**). Illumina/Solexa GAII reads were generated in 76 or 120 bp paired-end format. In total, 28.9 Gb of adaptor-filtered and Q20-trimmed sequence were obtained with this platform. Altogether, nearly 30 Gb of raw sequence were produced. After the process of data trimming, 99.64% of Roche 454 and 72.64% of Solexa sequence results were preserved (**Supplementary Table S1**).

We have optimized assembly procedures through a series of tests, the results of which are summarized in **Fig. 1A**. Our procedure consist of three steps: primary assembly of 454 reads, reference mapping of Solexa reads and de novo assembly of non-mapped Solexa reads (**Fig. 1A**; **Supplementary Table S3**). In the first step, reads from all libraries were pooled; assembly of all 454 reads was performed with the Newbler Assembler software. After removing contigs <200 bp, 34,563 contigs were obtained from this primary assembly with an average size of 1,194 bp. Additionally, 85,144 singletons >200 bp with an average size of 298 bp were obtained and included in the unique sequences for later mapping.

In the second step, the 454 contigs and singletons obtained (119,707 reference sequences altogether) were used as reference frames, and Illumina/Solexa GAII reads were aligned to the reference frames using CLC Genomic Workbench. About 92% of Solexa reads were mapped to the reference contigs and singletons in this step. When inconsistencies or rare variants were encountered, Solexa reads provided deeper coverage to adjust 454 reference sequences. There are two types of conflict, uncalled bases (N) and miscalled bases. About 97% of bases uncalled by 454 sequencing (1,350 out of a total of 1,393 Ns) were identified after mapping of Solexa reads. We found that the remaining 43 uncorrected Ns were not covered by Solexa reads in the mapping process. There were also 427,790

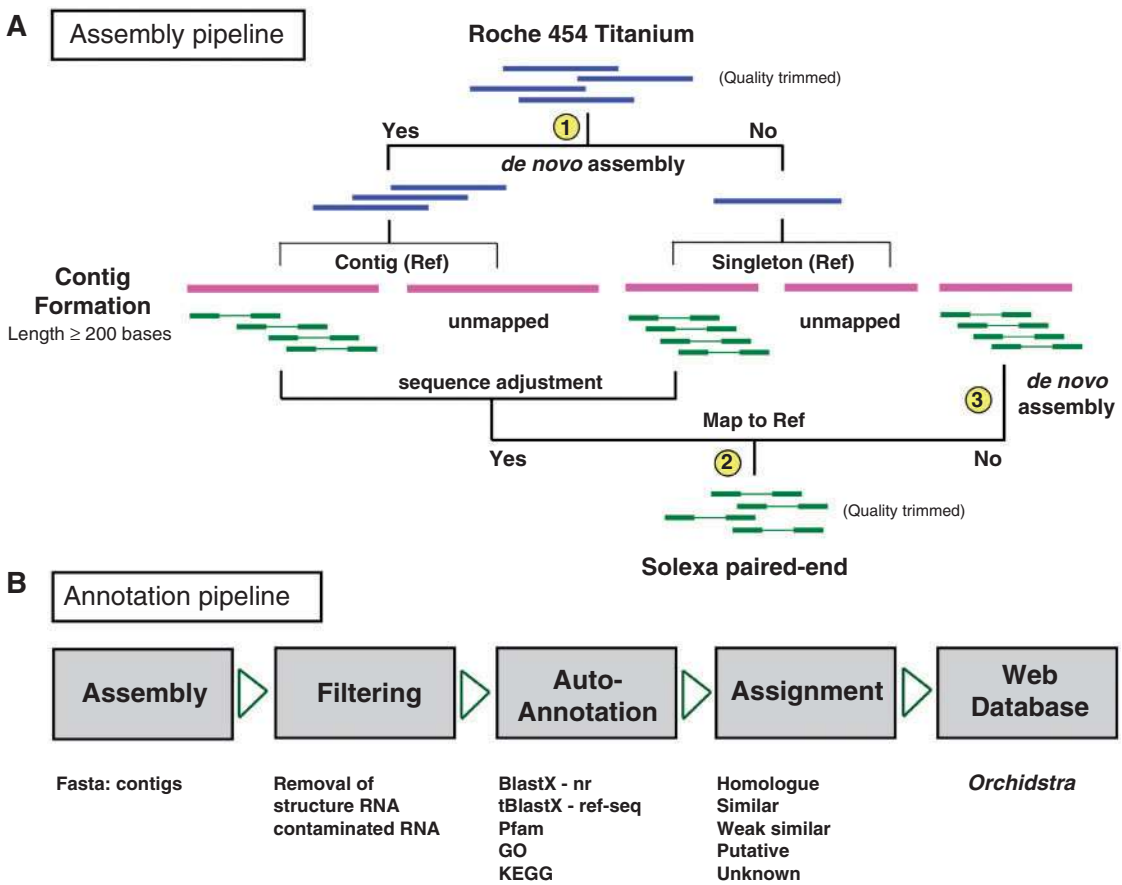

Fig. 1 Assembly scheme and annotation pipeline. (A) The strategy used for contig assembly is illustrated. Step 1: Roche 454 reads were assembled into Ref (reference sequence). Step 2: Solexa reads were mapped to the reference sequence and were used to adjust sequencing bias. Step 3: the remaining Solexa reads were de novo assembled into contigs. (B) Bioinformatic pipeline for annotation. For details, see the Materials and Methods and Results.

inconsistent bases adjusted by Solexa reads; bases were mainly present in homopolymeric nucleotide tracts or were polymorphisms with minor allele frequencies.

In the third step, Solexa reads that could not be aligned to 454 references were separately assembled de novo. The third assembly resulted in 126,535 transcripts with an average size of 334 bp per contig. Altogether, 93.6% of Roche 454 and 97.5% of Solexa quality data were assembled into unique sequences containing either contigs or singletons that are longer than 200 bp (**Supplementary Table S3**). In total, 246,212 unique sequences (prefixed with PACT plus a serial number) with a size larger than 200 bp were assembled and were considered as *Phalaenopsis*-expressed transcripts for further processing within the annotation pipeline. Nucleotide sequences of raw reads and contigs were deposited in the GenBank database.

Annotation

The annotation pipeline is shown in **Fig. 1B**. Contaminated and structure RNA sequences in the assembled 246,212 unique sequences were removed by comparison against a customized database comprising virus, rRNA, general vector and chloroplast sequences using a blastn cut-off E-value of 1e-30. A software bug was identified in the Roche 454 Newbler Assembler program that generated duplicate contigs. After blast and comparison, 6,683 duplicate sequences were found and removed from the transcript pool. Altogether, 12,389 contigs were eliminated which consisted of about 5% of the total transcripts (**Supplementary Table S4**). Among them, transcripts originating from exogenous contaminations including virus, bacteria and yeast make up <1% of the assembled contigs, indicating the healthiness of the greenhouse plants. After removal of contaminating sequences, 233,823 unique contigs were preserved for further annotation.

The remaining unique transcript sequences were then annotated according to sequence comparison results from BlastX analysis and functional annotation, using Gene Ontology (GO; http://www.geneontology.org/), Pfam (http://pfam.sanger.ac.uk/) and Kyoto Encyclopedia of Genes and Genomes KEGG; http://www.genome.jp/kegg/) databases. All 233,823 unique contigs were compared against the nr protein sequence database using standalone Blastx, tBlastX and customized Perl scripts. The homology search against nr resulted in 42,590 annotated genes (18.2% of the total unique contigs) with at least one Blastx hit at the E-value cut-off of 1e-10. **Table 1** shows that 88.7% of the 42,590 annotated genes were assembled by sequences derived from both Roche 454 and Solexa platforms. De novo assembly of the remaining Solexa reads contributed to 11.1% of the contigs, whereas contigs identified solely by Roche 454 reads were only 0.2%. The size distribution of the annotated contigs is shown in **Fig. 2**. Size distribution at N50 equals 1,485 bp in length and average contig size is 875 bp. Among the annotated coding sequences, 7,263 contigs (17% of the annotated genes) are predicted to be near

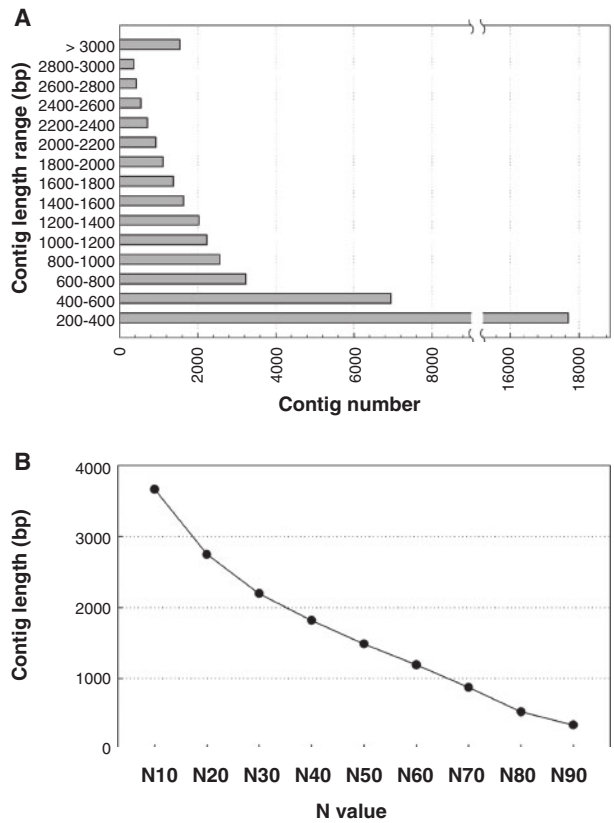

Fig. 2 Length distribution of the 42,590 annotated contigs. (A) Length distribution in base pairs. The average contig length is 875 bp. (B) Contig size distribution plot, where N50 is equal to 1,485 bases.

Table 1 Data source of assembled contigs

Source of assembly	Annotated genes				Non-annotated genes			
	No. of contigs	Average length (bp)	No. of bases	Base ratio (%)	No. of contigs	Average length (bp)	No. of bases	Base ratio (%)
454 contig with Solexa mapped	31,321	1,055.82	33.07 Mb	88.7	77,364	336.77	26.05 Mb	43.7
454 only	253	323.02	81.73 kb	0.2	820	267.05	218.98 kb	0.4
Solexa de novo assembly	11,016	375.35	4.13 Mb	11.1	113,049	295.03	33.35 Mb	55.9
Total	42,590	875.47 p	37.28 Mb	100.0	191,233	311.79 bp	59.62 Mb	100.0

A total of 42,590 annotated and 191,233 non-annotated contigs were traced back to three data sources. No. of bases indicates the sum of nucleotide bases of all the contigs, and base ratio is the number of bases from each step as a fraction of total bases.

full length, as they covered >90% of amino acid sequences of the corresponding homologs. Arbitrary nomenclature was applied to give names to assembled contigs with various degrees of similarity identified in current databases according to E-value and identity to the blasted genes (**Table 2**). There were 24,131 Pfam domains and 16,744 GO identities identified in the annotated genes, while 3,221 genes were confirmed by the EC number provided by KEGG.

Overall GO distribution is shown in **Fig. 3**. High counts of ESTs in metabolic processes were identified, such as genes involved in the biosynthesis of macromolecules and in nitrogen, protein and nucleic acid metabolism. Gene retrieval was performed to identify important genes in three functional groups: transcription factors, kinases and transporters. Using the Plant Transcription Factor Database categories, 1,275 *P. aphrodite* transcription factors were identified (**Fig. 4**), of which zinc finger proteins (41%) were the most abundant type followed by MYB (17%) and NAC (8%). A similarly high ratio of zinc finger proteins in the transcription factor category is present in Arabidopsis (38%) and rice (25%). There were 946 orchid transporters (including channels and carriers) identified in this analysis, whereas 644 Arabidopsis transporters and 949 rice transporters were reported. The numbers of amino acid transporters are close among the three species. A total of 739 kinases were found in the *Phalaenopsis* annotation using the classification system of the RKD database. This enzyme class is under-represented when compared with the number of kinases found in rice (1,411) and Arabidopsis (1,552), probably due to tissue distribution for library preparation (**Fig. 4**). However, this may also result from species variations, which clearly exist between Arabidopsis and rice. For example, there is a much higher number of TKLs (tyrosine kinase like) in rice when compared with Arabidopsis, i.e. 1,127 vs. 372 (Dardick et al. 2007). However, in analogy to rice and Arabidopsis kinases, the TKL type is still the most abundant kinase in orchid (504 TKLs). In spite of the massive amounts of sequence data acquired, the investigation of the transcriptome in this study may not be sufficient for a complete list of expressed transcripts. More efforts to collect transcriptome data by sequencing more libraries obtained from orchid plants grown under various treatments or at different growth stages may further enrich the database.

Construction of the database

In order to facilitate orchid functional genomics research, we created Orchidstra, a publicly available resource that combines *Phalaenopsis* orchid transcriptome data with relevant annotations. Orchidstra (http://orchidstra.abrc.sinica.edu.tw) is a web-based interface for annotated transcriptomic and genomic data of orchids generated from sequencing technology. Currently available organisms in Orchidstra include *P. aphrodite*, *P. bellina* and *P. equestris*. Among them, *P. aphrodite* transcripts were generated by our sequencing efforts. The interface of Orchidstra provides access to information on sequence, accession number, top hit alignments of blastx against the nr database, GO mapping results, EC number, KEGG pathway map, KEGG Orthology and Pfam domain alignments for each transcript.

Expression profiling and gene category analysis

Solexa reads from individual library were submitted to expression analysis applying the FPKM (fragments per kilobase of exon per million fragments mapped) value as the expression index (**Fig. 5A**). Genes differentially expressed among the three different tissue types (vegetative, reproductive and seed tissues) were analyzed in a Venn diagram (**Fig. 5B**). The majority of expressed genes appear in all three tissue types (24,248 out of 42,590 genes, or 57 %). Seed tissue has the greatest number of unique genes (3,690), while reproductive tissue has 2,608 unique genes and vegetative tissue has 762 unique genes. GO analysis of uniquely expressed transcripts is shown in **Supplementary Fig. S2**.

Validation of sequence accuracy

Forty-eight *P. aphrodite* contigs were randomly chosen from the Orchidstra transcriptome database for PCR amplification and analysis in order to validate both the accuracy of the assembly and base assignment. Gene identities and primers used in this study are listed in **Supplementary Table S5**. All of the target genes were successfully amplified by PCR with the predicted amplicon size (1,217 bp on average). Overall sequence accuracy from assembly is as high as 99.9% when comparing contig sequences derived from next-generation sequencers

Table 2 Summary of gene annotation

Tentative annotation	Blast cut-off	No. of genes
Homolog to accession/definition of protein	Identity ≥90% and E-value ≤1e-40	848
Similar to accession/definition of protein	Identity <90% and E-value ≤1e-40 or 1e-40 <E-value ≤1e-30)	20,285
Weakly similar to accession/definition of protein	1e-30 <E-value ≤1e-20	6,607
Putative protein to accession/definition of protein	1e-20 <E-value ≤1e-10	14,921
Total annotated genes		42,590
Total unknown transcripts	E-value >1e-10	191,233

The definition of automatic annotation was assigned to assembled contigs based on the E-value and sequence identity from the BlastX result. A total of 233,823 assembled contigs were classified into 42,590 annotated genes and 191,233 non-annotated (unknown) genes.

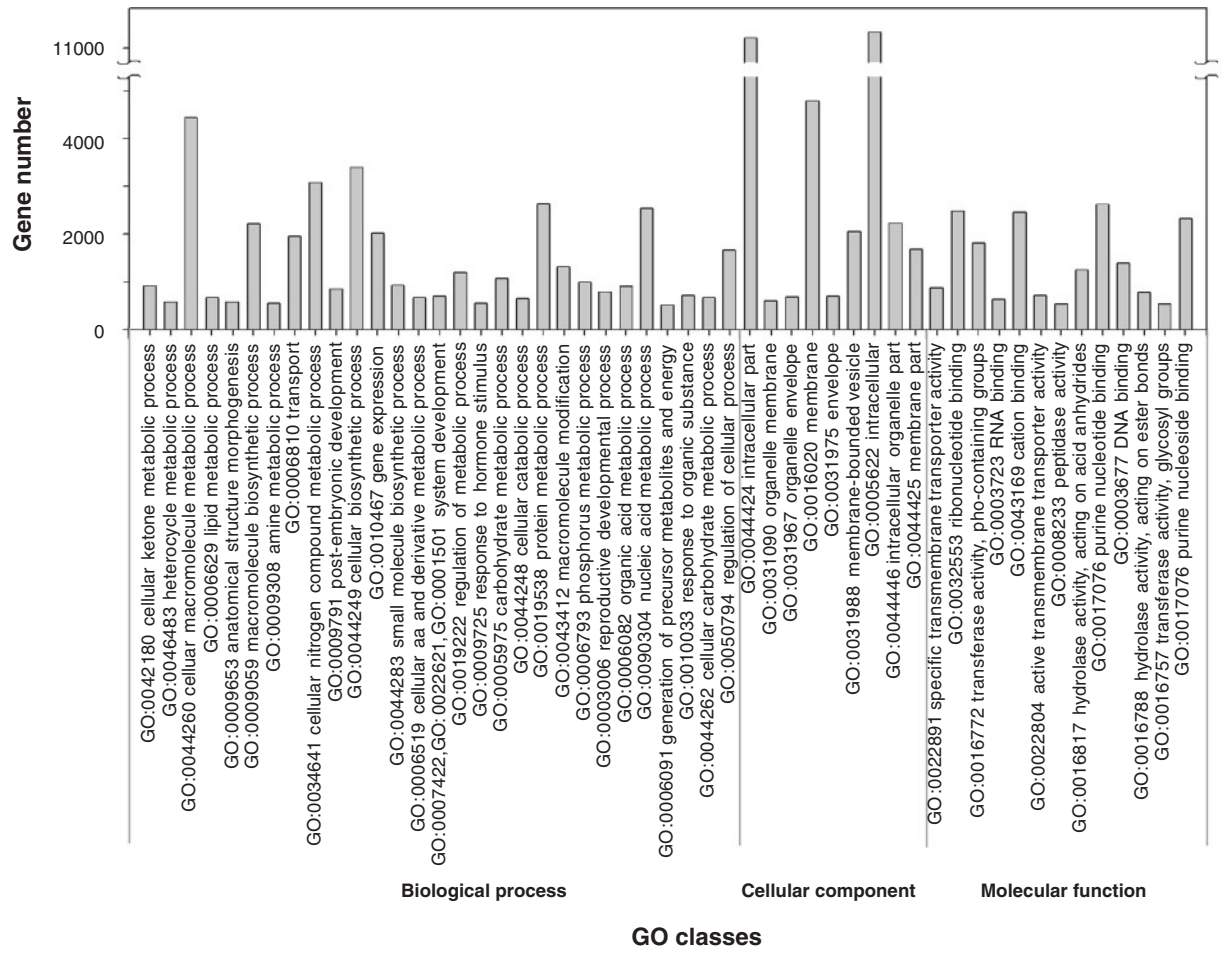

Fig. 3 Gene Ontology analysis of *Phalaenopsis*-annotated genes. Categories pertaining to biological process, cellular component and molecular function were defined by Gene Ontology classification.

with those of the PCR amplicons obtained through a conventional capillary sequencer, AB3730 (**Supplementary Table S5**).

Subcellular localization

In order to validate the annotation functionally in terms of cellular components, constructs of selective *Phalaenopsis* genes fused with soluble modified green fluorescent protein (smGFP; Davis and Vierstra 1998) were bombarded into *Phalaenopsis* Sogo Yukidian 'V3' white petals or sepals to examine their subcellular localizations. Based on GO annotation and the literature (Donaldson et al. 1992, Mao et al. 2005, Osterrieder et al. 2010), the following *Phalaenopsis* genes were chosen for this approach: 60S ribosomal protein (PATC141573) and an smD2 protein of small nuclear ribonucleoprotein (snRNP; PATC144393) genes for nuclear localization, a cyclophilin gene (PATC140899/PATC155502) for the cytoplasm, an actin gene (PATC141888) for the cytoskeleton, a GTPase (PATC138866) and ceramide synthase (PATC140948) for the endoplasmic reticulum (ER) and an ADP-ribosylation factor (PATC142447) for the Golgi apparatus (**Supplementary Table S6**). Green fluorescence indicating expression and targeting to cellular compartments was observed using a confocal microscope, indicating successful expression of the target genes, and it was also found co-localized with images of red fluorescence derived from expression of the corresponding Arabidopsis subcellular markers carrying mCherry standard constructs localized to compartments including the Golgi, ER (Nelson et al. 2007) and a nuclear construct with a nuclear localization signal (NLS) peptide (Lee et al. 2008) in the co-bombardment experiment (**Fig. 6**). Genes selected from the Orchidstra database were proven to be localized to the correct cellular compartments according to annotation and may be useful as cellular markers for future cell biology studies.

Discussion

Large number of *P. aphrodite* expressed genes were identified and analyzed with a high-throughput sequencing approach described in this study. In addition, a transcriptome database was constructed which is accessible to the public. Gene

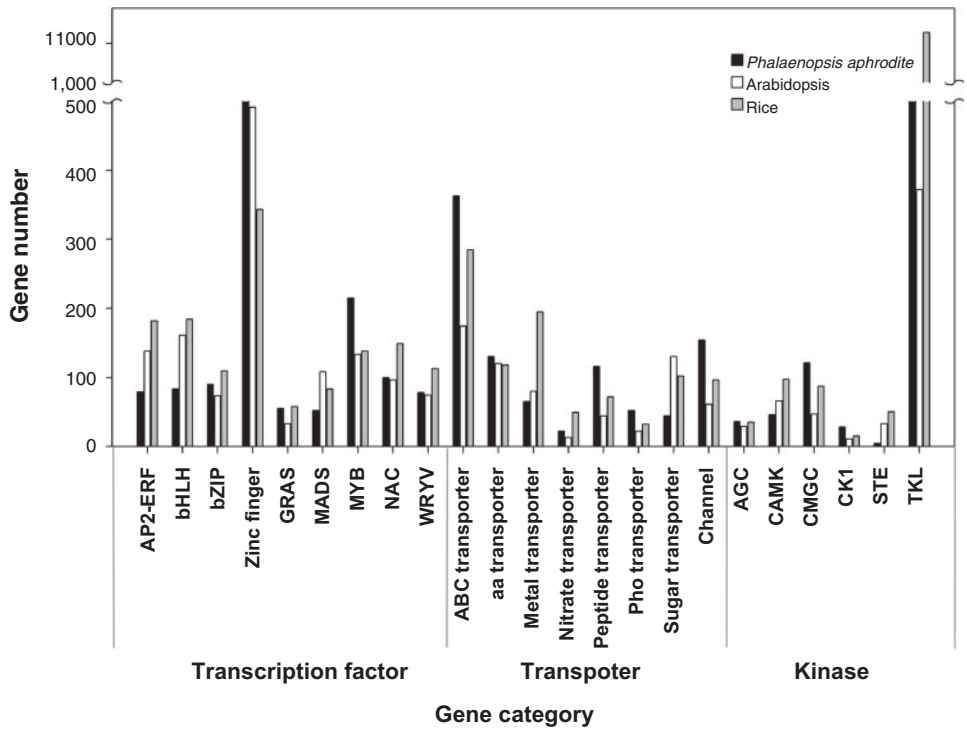

Fig. 4 Gene discovery by a high-throughput EST approach. Genes encoding transcription factors, transporters and kinases were identified through combination search of annotation, Pfam and GO. Numbers of genes in subcategories identified in the orchid database as well as those in the Arabidopsis and rice databases are shown.

expression profiling analysis was performed to differentiate genes with tissue-specific expression. These transcripts can serve as molecular and cellular markers in breeding programs and future genome research.

Assembly efficiency

Several types of non-commercial assembly software such as Velvet from EMBL-EBI and SOAPdenovo from BGI were evaluated for assembling reads generated in this study. Neither of them is suitable for assembly with input sequencing data that diverges greatly in numbers and lengths of reads, thus illustrating the complexity of and difficulty in integrating such data. Fu et al. (2011) attempted to combine sequencing output from Sanger sequencing as well as Roche 454 and Solexa high-throughput sequencing by consensus annealing of contigs generated from each individual platform. From our observation, the annealing of assembled contigs from different data sources purely based on the consensus is not reliable and is inefficient in terms of data usage (**Supplementary Table S2**). Neither is it reliable to mix long and short sequences or to mix few and rich data for assembly simply because algorithm and parameter adjustment is very different with distinct data sets. A strategy to optimize the assembly protocol and a logical working procedure is necessary in consolidating data from multiple platforms to improve sequencing efficiency for transcriptomic studies of a novel species.

The optimized assembly workflow described in **Fig. 1A** not only increased the number of unique contigs but also led to better base usage efficiency. A high base usage rate of this working pipeline was achieved, as 93.6% of Roche 454 trimmed sequence and 97.5% of Solexa quality sequence assembled into unique contigs (**Fig. 1A**). A total of 246,212 unique contigs and singletons with a size longer than 200 bp were obtained from this assembly method before annotation.

In conclusion, this assembly process has at least four benefits to satsify the general interest of researchers engaged in transcriptomic research. The first is the ability to obtain long contigs with the Roche 454 data assembly. The second is the ability to correct sequencing errors present in low-fold coverage 454 reads using the high-depth coverage sequencing data obtained with the Solexa platform. Thirdly, de novo assembly of the remaining unmapped Solexa data enriched novel gene discovery. Finally, the frequency of the mapped Solexa reads can be used for gene expression profiling.

Annotation

Considering error inheritance and insufficient information for a gene that could play multiple roles in various functions, we first divided annotation outcome into five categories according to criteria of identity and E-value from blastX and tblastX results: homolog, similar, weakly similar, putative and unknown (**Table 2**). The first four categories with an E-value cutoff at

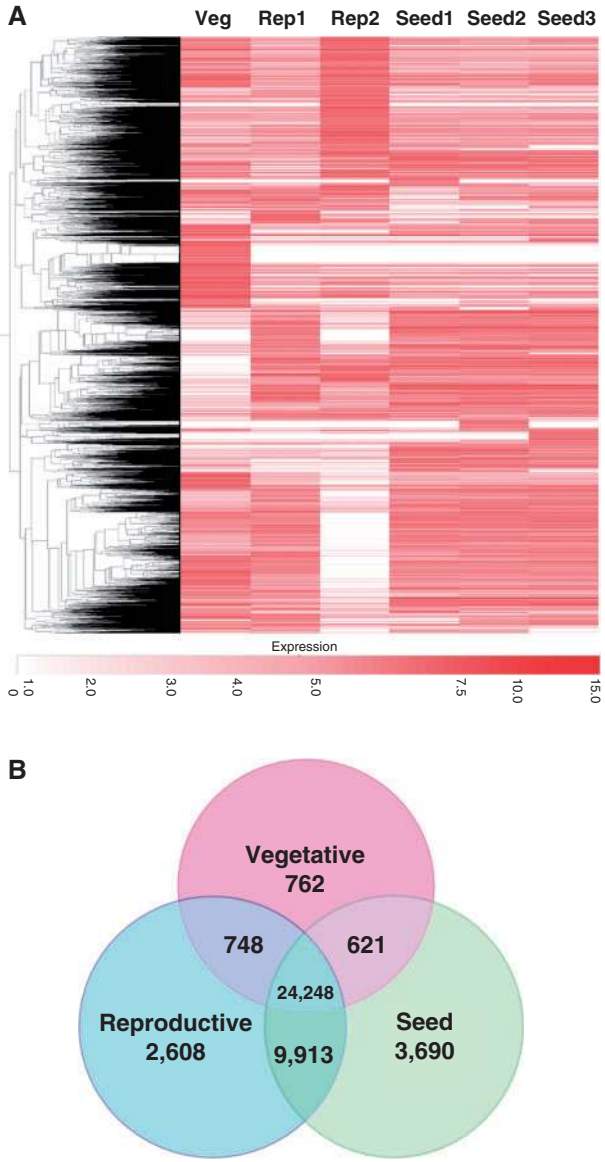

Fig. 5 Differential gene expression between tissues. (A) Gene cluster analysis. Expression levels were determined by FPKM values calculated from Solexa read counts. Veg, vegetative library; Rep1 and 2, reproductive library 1 and 2; Seed 1–3, seed library 1, 2 and 3. (B) The numbers of genes differentially expressed or shared among vegetative, reproductive and seed libraries are shown.

10E-10 (42,590 unique contigs) were defined as protein-coding genes, while unknown transcripts were defined as non-annotated (191,233 contigs). To provide additional information, Pfam, GO and KEGG identities were also assigned to the genes when applicable.

Of the top-hit species list, 5,288 *Phalaenopsis* genes were found to be similar to genes of common grapevine (*Vitis vinifera*), while 2,647 and 1,394 genes were similar to genes of rice and *Ricinus communis*, respectively (**Supplementary Fig. S3**). Given that many plant species have a much higher number of UniGenes or ESTs in current databases than grapevine, this similarity should not be coincidental; however, at present, the biological sense of this correlation remains unclear.

Apart from those genes that were found to be similar to protein-coding genes in current databases, there are 191,233 unique contigs in our database that lack similarity to known genes (E-value >1e-10) and do not possess an open reading frame long enough for comparison. Non-annotated genes are not uncommon since a significant proportion of non-annotated UniGenes exists in current databases of rice and Arabidopsis (data not shown). These unknown contigs may contain fragmented RNA including untranslated regions (UTRs), introns, long non-coding RNAs (ncRNAs), microRNA precursors and other types of transcripts (including contaminated transcripts from unidentified sources).

Functional category and differential expression

RNA-Seq taking advantage of the large number of reads generated from high-throughput sequencing technology has been proved to be a powerful tool in analysis of gene expression profiling (Wang et al. 2009). When filtered through criteria requiring the expression index (Log_2FPKM value) to be >5, gene clustering analysis revealed 8,804 differentially expressed genes (**Fig. 5A**); many of them are proteins with unknown functions. A unique glycine-rich RNA-binding protein (PATC230728), glutaredoxin (PATC153260), thioredoxin (PATC143280), glutathione transferase (PATC148647) and peroxidase (PATC156909) genes were highly expressed in all tissues, indicating an active defense mechanism and stress response. A few other common genes, including those encoding ubiquitins, proteases, histone H2B, Chl *a*/*b*-binding protein and ribosomal proteins, were among the highly expressed genes generally found in all tissues examined. A cyclophilin homolog (PATC155502) and an immunophilin (PATC148202), both immunosuppressants in humans, were also highly expressed in all tissues. Cyclophilin and immunophilin are enzymes with peptidyl-prolyl isomerase activity involved in protein folding. An orchid cyclophilin homolog had been reported in a Vanda orchid hybrid and in a *Phalaenopsis* hybrid with expression correlated to a peloric mutant phenotype (Chen et al. 2005).

Unique genes identified in three types of tissues (vegetative; reproductive 1 and 2 combined; seed 1, 2 and 3 combined) were analyzed by their GO distribution (**Supplementary Fig. S2**). Many of the differentially expressed genes were annotated as unknown function, exhibiting a vast research potential. The seed library contains the most unique genes, presumably due to the high diversity of genes needed for the architecture of early embryo formation. The novel genes identified in seed libraries present promising and intriguing projects for future functional research. Cupin, LEA (late embryogenesis abundant), metallothionein, heat shock 70 kDa protein, several dehydrogenases and some other genes with unknown function are among the most highly expressed seed-specific genes. LLA-115 (a tapetum-specific protein), a pollen-specific protein

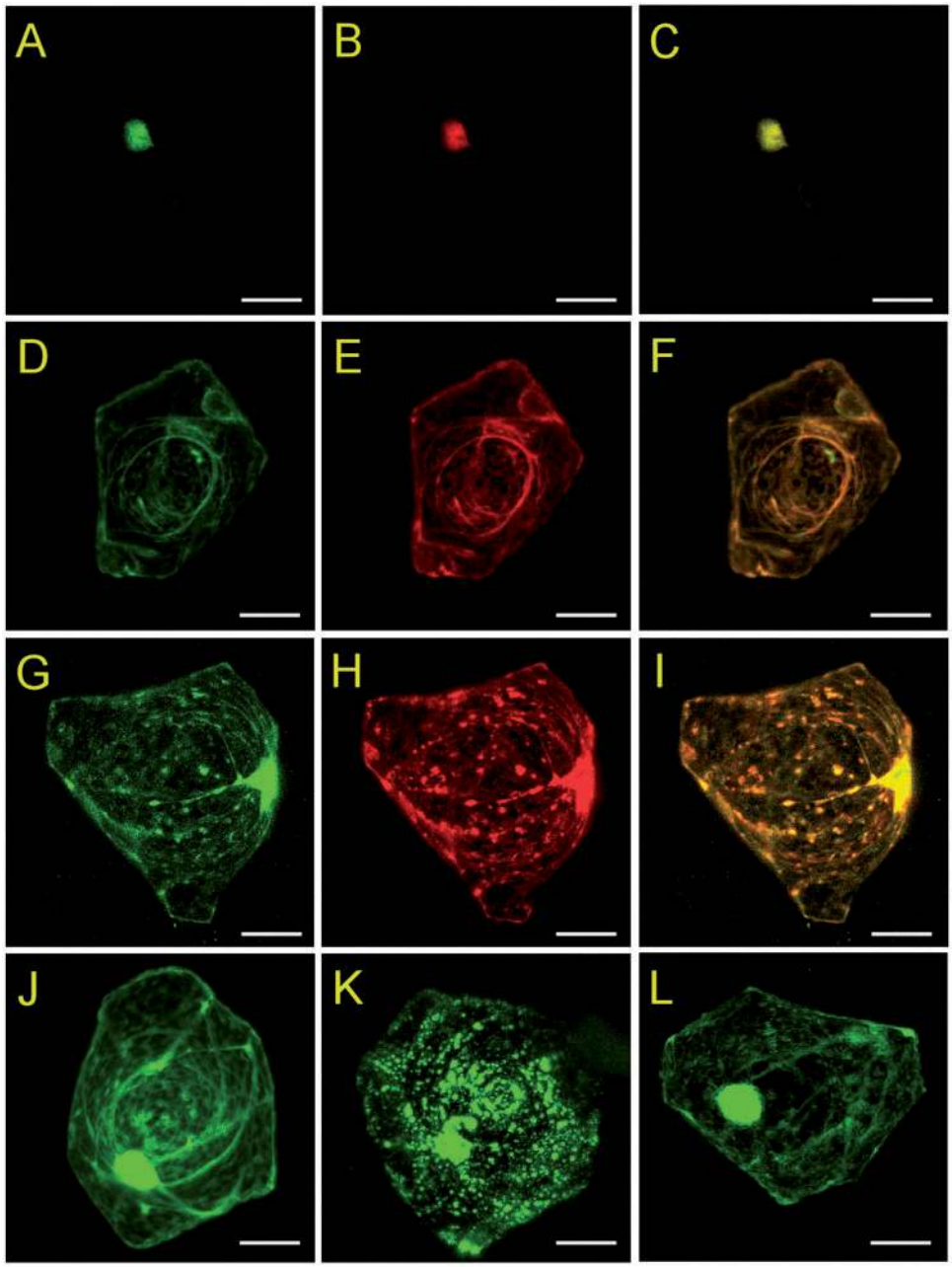

Fig. 6 Subcellular localization of *Phalaenopsis aphrodite* genes. Selective *Phalaenopsis* genes were fused with GFP and expressed in petal cells by particle bombardment. Confocal images derived from co-expression of *Phalaenopsis* constructs and Arabidopsis marker constructs were arranged in the following order: A–C, nuclear localization; D–F, ER localization; G–I, Golgi localization. *Phalaenopsis* constructs with GFP: A, PATC144393; D, PATC140948; and G, PATC142447 are in the green channel, and Arabidopsis marker constructs with mCherry: B, ER-rk CD3-959; E, G-rb CD3-968; and H, E3170 are in the red channel. C, F and I are merged images taken in the green and red channels from A and B, D and E and G and H, respectively. Green fluorescence images of the cytoplasm construct, J, cyclophilin (PATC140899/PATC155502); and the cytoskeleton construct, K, actin (PATC141888) were taken without co-bombardment with Arabidopsis marker genes. A fluorscence image of free GFP with vector-only control, 326-GFP, is shown in L. Scale bars indicate 20 μm.

homolog, a MADS box transcription factor, and many others were identified as reproductive-specific expressed genes. These tissue-specific expressed genes present opportunities to investigate genes involved in unique molecular and physiological functions in orchids further.

Validation of sequence accuracy and functional annotation of the assembled contigs

When amplicons of randomly selected contigs were submitted to capillary sequencing and the sequences confirmed in both

directions, the overall accuracy of the assembled contigs was estimated to be as high as 99.9%. This result assured us that our assembly procedure is fairly reliable. This high accuracy is due to the deep coverage and sequence adjustment from Solexa mapping. Among the 48 genes used for validation, six were found to have single nucleotide polymorphisms (SNPs) detected by Sanger sequencing (**Supplementary Table S5**); however, additional SNPs with lower minor allele frequency can be identified by Solexa consensus comparison (data not shown).

Particle bombardment (You et al. 2003) with GFP fusion constructs to transiently express genes of interest in petals of a commercially popular *Phalaenopsis* hybrid is newly developed in this study for functional confirmation of the annotation process. Subcellular localization of *Phalaenopsis* contig–GFP fusion constructs was observed by fluorescence co-localization with Arabidopsis standard constructs (with mCherry) in cellular compartments such as nuclei, Golgi and ER in orchid petal cells, indicating a similar mechanism of protein targeting between the two species (**Fig. 6**). Additionally, genes targeting cytoplasm and cytoskeleton localizations were also observed. The result not only demonstrates correct functional annotation of a selection of *Phalaenopsis* genes but also provides subcellular markers for future cell biology studies. Lack of a reliable protocol for stable transformation and the long life cycle of orchids are hurdles for a transgenic approach in functional studies of orchid genes in general. Transient expression in white petal cells of a commercial *Phalaenopsis* hybrid by particle bombardment provides an alternative way to investigate gene functions.

Perspectives of an orchid database

Orchid genomic information including genome structure, transcriptome sequences and markers such as physical, genomic and cytogenetic markers are fairly limited in the current databases. Sequence information of expressed genes of *P. aphrodite* obtained in our study was deposited in the Orchidstra database. Orchidstra is a web database designated for orchid genomic and transcriptomic information. Orchidstra is a combined word from orchid and orchestra, denoting the harmonious interplay of a collection of genes to bring about the beauty of orchids. It is not simply a sequence archive but serves as a value-added database with functional annotations and gene expression profiling data, among other features. For example, an orchid specialized microarray was designed based on *Phalaenopsis* transcriptomic sequences assembled in this work, and the performance test is in progress. The relative expression level derived from microarray analysis is included in the annotation to support functional studies (data not shown, see Orchidstra database). In the future, we will keep on enriching genomic information of more orchid species with continuous efforts in data integration with functional research.

Conclusion

To initiate genomic research for a non-model organism, an EST approach for transcriptome analysis has many advantages and potential applications. An effective workflow to integrate data from advanced sequencing technologies was described. Significant features of this working procedure include pooling of libraries from various tissues for a comprehensive source of expressed genes, steps to combine reads from two platforms to enhance assembly performance, a streamlined annotation process, generation of gene expression profiles using sequencing result as tags, and construction of a web database. The entire workflow minimizes time and labor for informatic process and maximizes output in both novel gene discovery and sequence information. Furthermore, it can easily be adapted to most research subjects on demand.

Materials and Methods

Plant materials

Taiwan's native moth orchid, *P. aphrodite* Rchb.f., was collected from its original mountain habitat in Dawu, Taitung county, and kindly provided by Dr. Tsai-Mu Shen, National Chyayi University, Chyayi county, Taiwan. *Phalaenopsis* Sogo Yukidian 'V3' (a popular tetraploid commercial *Phalaenopsis* hybrid) used in the particle bombardment experiment was purchased from commercial nurseries.

Mature plants were maintained in growth chambers at 22–27°C under a 12 h light/dark cycle with regular irrigation and fertilization. Seeds from hand-pollinated capsules (120 d after pollination) were germinated on 1/4 Murashige and Skoog medium including Gamborg B5 vitamins (Duchefa Biochemie), pH 5.6, with 1% tryptone, 2% sucrose and 0.85% agar under the same growth conditions as for mature plants.

Three libraries were built from collections of various orchid tissues. Leaf, stem and root from mature *P. aphrodite* plants were pooled for the vegetative library. A reproductive library was built from young inflorescence (emerging inflorescence about 10–15 cm in length) and mature stalk (reproductive 1 library), and flower buds, full blossoms and senescing flowers (reproductive 2 library). A seed library was built from a collection of germinating seeds divided into three phases: seed 1 library, protocorm formation (0–30 DAS) when the globular shape of the protocorm with absorbing hairs was formed; seed 2 library, protocorm development (40–75 DAS) with pseudo-rhizome and cleavage forming; and seed 3 library, seedling formation (75–100 DAS) with the first leaf and sometimes the root emerging from the protocorm.

Orchid RNA extraction

Orchid total RNA was isolated as described previously with minor modification (Yu and Goh 2000). Plant tissues were frozen and ground in liquid nitrogen. RNA was extracted by

vortexing with 5–10 vols. (w/v) of extraction buffer (2% hexadecyltrimethylammonium bromide, 1% polyvinylpyrrolidone 40, 100 mM Tris–HCl pH 7.5, 20 mM EDTA, 2 M NaCl, 2% 2-mercaptoethanol) pre-warmed at 65°C. The homogenate was incubated at 65°C for 15 min with frequent mixing. The homogenate was centrifuged at 3,000 × *g* for 10 min at room temperature. The supernatant was extracted twice with an equal volume of chloroform : isoamyl alcohol (24 : 1, v/v) and centrifuged at 8,000 × *g* for 15 min. One-third volume of 8 M lithium chloride was added to the aqueous phase to precipitate total RNA at −20°C overnight. The RNA pellet was harvested by centrifugation at 12,000 × *g* for 30 min at 4°C, washed with ice-cold 75% ethanol, and resuspended in RNase-free water. RNA purity and concentration were determined by Nanodrop 2000 (Thermo Scientific) or Qubit (Invitrogen) measurement. The quality of RNA was evaluated on an Agilent 2100 Bioanalyzer (Agilent).

Solexa cDNA library preparation

The Illumina/Solexa Genome Analyzer IIx system was used in this massively parallel sequencing approach. Six libraries were prepared for Solexa paired-end sequencing. The vegetative library with pooled RNAs from root (1 cm from the root tip), stem and leaf was constructed by a SMART cDNA Synthesis kit (Clontech) and normalized via DSN (Evrogen) (Zhulidov et al. 2004) treatment followed by pair-end DNA Sample Preparation Kits (Illumina) which included DNA shearing by sonication, adaptor ligation and DNA enrichment. The young inflorescence (reproductive 1) and flower (buds and full blossom; reproductive 2) libraries were prepared using the mRNA-Seq kit (Illumina). Three seed libraries including protocorm formation, protocorm development and seedling formation (seed 1, seed 2 and seed 3) were constructed using the Multiplex sample preparation kit (Illumina). Both kits were used in all steps of library construction, i.e. mRNA isolation, RNA fragmentation, cDNA synthesis, adaptor ligation and DNA enrichment, starting with 10 μg of total RNA. All procedures followed protocols provided by the manufacturer.

Roche 454 cDNA library preparation

Equal amounts of mRNAs were pooled from the indicated tissues to construct three cDNA libraries for Roche 454 sequencing. The tissue samples included vegetative tissues (leaf, stem and root), reproductive tissues (reproductive 1 and 2) and germinating seeds (seed 1, 2 and 3). mRNA was purified from total RNA using the PolyATract mRNA Isolation kit (Promega). Each cDNA library was constructed using the cDNA Rapid Library Preparation Kit (454 Life Sciences, Roche), starting from 200 ng of mRNA. All steps including RNA fragmentation, cDNA synthesis, adaptor ligation and product quantification followed protocols provided by the manufacturer. The resulting cDNA libraries were run on the Roche 454 GS FLX Titanium system.

Sequence processing and assembly

Processing of sequence data for assembly was conducted using Newbler (GS De Novo Assembler) (Brandford) for assembling Roche 454 reads, CLC Genomics Workbench (CLC bio) for Solexa mapping and customized Perl scripts for data streaming. Sequenced bases with a quality lower than Q20 or equivalent measurement were discarded. All sequences assembled from Roche 454 reads were taken as reference frames for mapping of Illumina GAII reads (**Fig. 1A**). Illumina GAII reads were produced in paired-end formats (76 or 120 bp) from six cDNA libraries: vegetative, reproductive 1, reproductive 2, seed 1, seed 2 and seed 3. Solexa reads were assembled onto reference frames; reads that could not be aligned to 454 references were de novo assembled. All assembled contigs from procedures described above were submitted to the annotation pipeline. Sequence data were deposited in GenBank (NCBI accession Nos. SRA030409; contigs JI626343–JI831113).

Annotation pipeline

As shown in **Fig. 1B**, the assembled contigs were first filtered to remove non-coding RNA and contaminating sequences such as virus (CymMV, accession No. NC_001812; and ORSV, accession No. NC_001728), *Escherichia coli*, *Phalaenopsis* chloroplast (accession No. NC_007499), rRNA (data source: SILVA rRNA database project, http://www.arb-silva.de/, release version: SILVA 102) (Pruesse et al. 2007), and other non-coding sequences in the Rfam database (Rfam 10.0, January 2010 release, ftp://ftp.sanger.ac.uk/pub/databases/Rfam/10.0/) such as tRNA and snRNA. Duplicate contig sequences derived from Newbler error were also removed. The remaining contigs were submitted to BlastX against the NCBI nr database (the nr database was downloaded from ftp://ftp.ncbi.nih.gov/blast/db; last updated in August 2010) followed by tBlastx search against the RefSeq RNA database. The first four categories: homolog, similar, weakly similar and putative, are considered protein-coding genes. When the E-value became larger than 1e-10, the genes were defined as unknown transcripts.

Only those sequences that met our default criteria for annotation were scanned with Pfam 24.0. In addition, those sequences were blasted again with an E-value cut-off at 1e-30 and the output was transformed into XML format for running Blast2Go analysis (Gotz et al. 2008). Default parameter values were used in the annotation configuration, except for the E-value Hit filter (1e-30 instead of 1e-6). After the Blast2GO mapping process, EC numbers from the KEGG pathway (http://www.genome.jp/kegg/) and GO terms were generated. When an EC number was obtained, the contig is mapped to the relative KEGG pathway and KEGG Orthology (K number). Through the pipeline, gene names or products, protein domains, GO terms and EC numbers were assigned to *P. aphrodite* transcripts based on the similarity to functionally characterized proteins and/or functional domains

ESTs from other *Phalaenopsis* species such as *P. equestris* and *P. bellina* are present in the NCBI database. They were

downloaded using the search term 'Phalaenopsis[organism] AND gbdiv_EST[PROP]'. Annotation of these ESTs was conducted using the same procedure of our annotation pipeline.

Database construction

Nucleotide fasta sequences, blast results, annotations, GO results, KEGG results and Pfam results were stored in a normalized MySQL database. To access these genomic resources in a user-friendly way, we built a customized database, Orchidstra. The website of this database runs with the Apache Web server in a Linux environment. PHP and JavaScript scripts were used to create the user interface coupled with MySQL, a relational database management system. The URL address of the Orchidstra genome database is http://orchidstra.abrc.sinica.edu.tw. Current features include resource browsing and online tools such as searching, blasting and linking to corresponding Pfam, GO and KEGG.

Analysis of gene expression profile and functional category analysis

Solexa reads were mapped to final assembled contigs from the original six libraries. The SAM files were generated separately. Fragment counts that are properly normalized can be used as a measure of relative abundance of transcripts, so we used the Cufflinks software (Mortazavi et al. 2008, Trapnell et al. 2010) and customized scripts to calculate the average FPKM value of each mapped contig. The FPKM value was transformed by $\log_2$ as an expression index and loaded in GeneSpring GX7.3 software (Agilent) for further analyses. A filter of $\log_2$ value <5 in all tissues was set to remove low-abundant expression transcripts.

Transcription factors of *Phalaenopsis* orchid were searched and organized based on sequence homology in PlantTFDB (http://planttfdb.cbi.pku.edu.cn/, Center for Bioinformatics, Peking University, China) (Gao et al. 2006) as well as Pfam annotation. Transcription factors of rice (*Oryza sativa* subsp. *japonica*) were also identified from the same database and those of Arabidopsis were obtained from The Arabidopsis Information Resource (TAIR: http://www.arabidopsis.org/). Transporters of Arabidopsis were searched in the TAIR database; an exception is the ABC family that was taken from a previous publication (Verrier et al. 2008). Transporters of rice were identified from the gene list in TransportDB (http://www.membranetransport.org/). Kinases of rice and Arabidopsis were identified and categorized by searching the RKD database (http://phylomics.ucdavis.edu/kinase/index.shtml, University of California, Berkeley, CA, USA) (Dardick et al. 2007). Orchid kinases were found in our own database using the same criteria as in RKD.

Validation of assembled contigs

A total of 48 contigs with a size larger than 900 bp was randomly selected from the *Phalaenopsis* transcriptome database for sequence validation. Primers were designed to generate an amplicon size >900 bp for each contig (**Supplementary Table S5**). Each amplicon was generated by regular PCR using cDNA derived from mRNA of vegetative tissues as template. DNA sequencing of amplicons was conducted using a capillary DNA sequencer (DNA Analyzer 3730, Applied Biosystems) using both forward and reverse primers that were used in PCR.

Subcellular localization

Genes with annotations of potential subcellular targeting were selected from the Orchidstra database by either GO analysis or the literature (Donaldson et al. 1992, Jin et al. 2004, Mao et al. 2005, Osterrieder et al. 2010). Gene identities, subcellular compartments and PCR primers are listed in **Supplementary Table S6**. PCR-amplified cDNAs were cloned into an smGFP vector (326-GFP), accession No. U70495 (Davis and Vierstra 1998). All amplified constructs were confirmed by DNA sequencing. One of the marker genes, cyclophilin, was assembled into two contigs (PATC140899/PATC155502) with a six base gap, one contig mapped to the N-terminus and the other to the C-terminus of the protein. PCR primers were then designed to amplify the entire coding region and successfully filled the gap to bring the two contigs into one. The amplified cyclophilin full-length cDNA with the GFP fusion construct was expressed and displayed fluorescence in the cytoplasm.

Standard markers were obtained with mCherry constructs of ER-rk CD3-959 for the ER, G-rb CD3-968 for the Golgi (Nelson et al. 2007) and an NLS signal peptide fusion (E3170, kindly provided by Dr. Gelvin) for nuclei (Lee et al. 2008). Coating to gold particles and bombardment were conducted following the manufacturer's manual of Biolistic® PDS-1000/He (Bio-Rad) with a slight modification with regard to vacuum pressure. Gold particles and macrocarriers were purchased from InBIO GOLD, Australia. White petals and sepals of a commercial hybrid, *Phalaenopsis* Sogo Yukidian 'V3', were used as target explants. Petals and sepals were excised and placed on top of 1% water agar for bombardment. Optimal configuration parameters for bombardment were tested and optimized with a rupture disc [900 p.s.i. and vacuum pressure (27 mmHg)]. The bombarded explants were incubated in the dark for 24 h and images of cells with fluorescence were taken by confocal microscopy (LSM 510 META NLO DuoScan, Carl Zeiss).

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by Academia Sinica [Development Program of Industrialization for Agricultural Biotechnology (http://dpiab.sinica.edu.tw/index_en.php) grant No. 098S0311].

Acknowledgments

The authors wish to thank Dr. Tsai-Mu Shen of the National Chiayi University, Chiayi, Taiwan, for providing the native collection of *Phalaenopsis aphrodite*. The authors greatly appreciate technical support from Academia core facilities and personnel, including Dr. Mei-Yeh Lu for the Solexa and Roche 454 sequencing technique, Miss Shu-Chen Shen for assistance in image acquisition with the confocal microscope at the Scientific Instrument Center of Academia Sinica, and the Transgenic Plant Laboratory for assistance in particle bombardment.

References

Arditti, J. (1992) *In* Fundamentals of Orchid Biology. John Wiley and Sons, New York.

Burgeff, H. (1959) Mycorrhiza of orchids. *In* The Orchids. A Scientific Survey. Edited by Withner, C.L. pp. 361–395. Ronald Press Co., New York.

Chen, Y.H., Tsai, Y.J., Huang, J.Z. and Chen, F.C. (2005) Transcription analysis of peloric mutants of Phalaenopsis orchids derived from tissue culture. *Cell Res.* 15: 639–657.

Dardick, C., Chen, J., Richter, T., Ouyang, S. and Ronald, P. (2007) The rice kinase database. A phylogenomic database for the rice kinome. *Plant Physiol.* 143: 579–586.

Darwin, C. (1862) *In* On the Various Contrivances by which British and Foreign Orchids are Fertilized by Insects. John Murray, London.

Davis, S.J. and Vierstra, R.D. (1998) Soluble, highly fluorescent variants of green fluorescent protein (GFP) for use in higher plants. *Plant Mol. Biol.* 36: 521–528.

Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36: e105.

Donaldson, J.G., Cassel, D., Kahn, R.A. and Klausner, R.D. (1992) ADP-ribosylation factor, a small GTP-binding protein, is required for binding of the coatomer protein beta-COP to Golgi membranes. *Proc. Natl Acad. Sci. USA* 89: 6408–6412.

Dressler, R.L. (1990) *In* The Orchids, Natural History and Classification Harvard University Press, Cambridge, MA.

Fu, C.H., Chen, Y.W., Hsiao, Y.Y., Pan, Z.J., Liu, Z.J., Huang, Y.M. et al. (2011) OrchidBase: A collection of sequences of transcriptome derived from orchids. *Plant Cell Physiol.* 52: 238–243.

Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W., Zheng, W. et al. (2006) DRTF: a database of rice transcription factors. *Bioinformatics* 22: 1286–1287.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J. et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36: 3420–3435.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L. et al. (2011) The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471: 473–479.

Guo, W.J. and Lee, N. (2006) Effect of leaf and plant age, and day/night temperature on net CO2 uptake in Phalaenopsis amabilis var. formosa. *J. Amer. Soc. Hort. Sci,* 131: 320–326.

Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D. et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nat. Methods* 5: 183–188.

Huang, M.D., Wei, F.J., Wu, C.C., Hsing, Y.I. and Huang, A.H. (2009) Analyses of advanced rice anther transcriptomes reveal global tapetum secretory functions and potential proteins for lipid exine formation. *Plant Physiol.* 149: 694–707.

Jin, A., Itahana, K., O'Keefe, K. and Zhang, Y. (2004) Inhibition of HDM2 and activation of p53 by ribosomal protein L23. *Mol. Cell. Biol.* 24: 7669–7680.

Knudson, L. (1922) Nonsymbiotic germination of orchid seeds. *Bot. Gaz.* 73: 1–25.

Lee, L.Y., Fang, M.J., Kuang, L.Y. and Gelvin, S.B. (2008) Vectors for multi-color bimolecular fluorescence complementation to investigate protein–protein interactions in living plant cells. *Plant Methods* 4: 24.

Mao, T., Jin, L., Li, H., Liu, B. and Yuan, M. (2005) Two microtubule-associated proteins of the Arabidopsis MAP65 family function differently on microtubules. *Plant Physiol.* 138: 654–662.

Metzker, M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11: 31–46.

Meyer, E., Aglyamova, G.V., Wang, S., Buchanan-Carter, J., Abrego, D., Colbourne, J.K. et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* 10: 219.

Micheneau, C., Johnson, S.D. and Fay, M.F. (2009) Orchid pollination: from Darwin to the present day. *Bot. J. Linn. Soc.* 1611–1619.

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

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344–1349.

Nelson, B.K., Cai, X. and Nebenfuhr, A. (2007) A multicolored set of in vivo organelle markers for co-localization studies in Arabidopsis and other plants. *Plant J.* 51: 1126–1136.

Nilsson, L.A. (1992) Orchid pollination biology. *Trends Ecol. Evol.* 7: 255–259.

Osterrieder, A., Hummel, E., Carvalho, C.M. and Hawes, C. (2010) Golgi membrane dynamics after induction of a dominant-negative mutant Sar1 GTPase in tobacco. *J. Exp. Bot.* 61: 405–422.

Ping, C.Y., Lee, Y.I., Lin, T.S., Yang, W.J. and Lee, G.C. (2010) Crassulacean acid metabolism in Phalaenopsis aphrodite var. formosa during different developmental stages. *Acta Hort.* 878: 71–77.

Pridgeon, A.M., Cribb, P.J., Chase, M.W. and Rasmussen, F.N., eds. (2005) Genera Orchidacearum: Epidendroideae (Part One). Oxford University Press, Oxford.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J. et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35: 7188–7196.

Quinlan, A.R., Stewart, D.A., Stromberg, M.P. and Marth, G.T. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods* 5: 179–181.

Ramirez, S.R., Gravendeel, B., Singer, R.B., Marshall, C.R. and Pierce, N.E. (2007) Dating the origin of the Orchidaceae from a fossil orchid with its pollinator. *Nature* 448: 1042–1045.

Rasmussen, H.N. (2002) Recent developments in the study of orchid mycorrhiza. *Plant Soil* 244: 149–163.

Rudall, P.J. and Bateman, R.M. (2002) Roles of synorganization, zygomorphy and heterotopy in floral evolution: the gynostemium and labellum of orchids and other lilioid monocots. *Biol. Rev.* 77: 403–441.

Shin, H., Hirst, M., Bainbridge, M.N., Magrini, V., Mardis, E., Moerman, D.G. et al. (2008) Transcriptome analysis for Caenorhabditis elegans based on novel expressed sequence tags. *BMC Biol.* 6: 30.

Silvera, K., Neubig, K.M., Whitten, W.M., Williams, N.H., Winter, K. and Cushman, J.C. (2010) Evolution along the crassulacean acid metabolism continuum. *Funct. Plant Biol.* 37: 995–1010.

Silvera, K., Santiago, L.S., Cushman, J.C. and Winter, K. (2009) Crassulacean acid metabolism and epiphytism linked to adaptive radiations in the Orchidaceae. *Plant Physiol.* 149: 1838–1847.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J. et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28: 511–515.

Tsai, W.C., Hsiao, Y.Y., Lee, S.H., Tung, C.W., Wang, D.P., Wang, H.C. et al. (2006) Expression analysis of the ESTs derived from the flower buds of Phalaenopsis equestris. *Plant Sci.* 170: 426–432.

van der Cingel, N.A. (2007) Pollination of orchids by Lepidoptera: outcrossing by long distance transport. *In* Orchid Biology: Reviews and Perspectives. Edited by Cameron, K.M., Arditti, J. and Kull, T. pp. 201–260. The New York Botanical Garden Press, New York.

Verrier, P.J., Bird, D., Burla, B., Dassa, E., Forestier, C., Geisler, M. et al. (2008) Plant ABC proteins—a unified nomenclature and updated inventory. *Trends Plant Sci.* 13: 151–159.

Vinogradova, T. and Andronova, E.V. (2002) Development of orchid seeds and seedlings. *In* Orchid Biology: Reviews and Perspectives. Edited by Kull, T. and Arditti, J. pp. 167–234. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Wakaguri, H., Suzuki, Y., Katayama, T., Kawashima, S., Kibukawa, E., Hiranuka, K. et al. (2009) Full-Malaria/Parasites and Full-Arthropods: databases of full-length cDNAs of parasites and arthropods, update 2009. *Nucleic Acids Res.* 37: D520–D525.

Wall, P.K., Leebens-Mack, J., Chanderbali, A.S., Barakat, A., Wolcott, E., Liang, H. et al. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* 10: 347.

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

You, S.J., Liau, C.H., Huang, H.E., Feng, T.Y., Prasad, V., Hsiao, H.H. et al. (2003) Sweet pepper ferredoxin-like protein (pflp) gene as a novel selection marker for orchid transformation. *Planta* 217: 60–65.

Yu, H. and Goh, C.J. (2000) Identification and characterization of three orchid MADS-box genes of the AP1/AGL9 subfamily during floral transition. *Plant Physiol.* 123: 1325–1336.

Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., Vagner, L.L., Khaspekov, G.L., Kozhemyako, V.B. et al. (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* 32: e37.