OrchidBase: A Collection of Sequences of the Transcriptome Derived from Orchids

Chih-Hsiung Fu1, Yun-Wen Chen2, Yu-Yun Hsiao3, Zhao-Jun Pan3, Zhong-Jian Liu4,5,6, Yueh-Min Huang1,*, Wen-Chieh Tsai2,7,* and Hong-Hwa Chen2,3,7,*

1Department of Engineering Science, National Cheng Kung University, Tainan 701, Taiwan
2Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, Taiwan
3Department of Life Sciences, National Cheng Kung University, Tainan 701, Taiwan
4Shenzhen Key Laboratory for Orchid Conservation and Utilization, The National Orchid Conservation Center of China and The Orchid Conservation & Research Center of Shenzhen, Shenzhen 518114, PR China
5The Center for Biotechnology and BioMedicine, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, PR China
6College of Forestry, South China Agricultural University, Guangzhou 510642, PR China
7Orchid Research Center, National Cheng Kung University, Tainan 701, Taiwan
*Corresponding authors: Hong-Hwa Chen, E-mail, hhchen@mail.ncku.edu.tw; Fax, +886-6-235-6211; Wen-Chieh Tsai, E-mail, tsaiwc@mail.ncku.edu.tw; Fax, +886-6-274-2583; Yueh-Min Huang, E-mail: huang@mail.ncku.edu.tw; Fax +886-6-276-6549

(Received October 14, 2010; Accepted December 16, 2010)

Orchids are one of the most ecological and evolutionarily significant plants, and the Orchidaceae is one of the most abundant families of the angiosperms. Genetic databases will be useful not only for gene discovery but also for future genomic annotation. For this purpose, OrchidBase was established from 37,979,342 sequence reads collected from 11 in-house Phalaenopsis orchid cDNA libraries. Among them, 41,310 expressed sequence tags (ESTs) were obtained by using Sanger sequencing, whereas 37,908,032 reads were obtained by using next-generation sequencing (NGS) including both Roche 454 and Solexa Illumina sequencers. These reads were assembled into 8,501 contigs and 76,116 singletons, resulting in 84,617 non-redundant transcribed sequences with an average length of 459 bp. The analysis pipeline of the database is an automated system written in Perl and C#, and consists of the following components: automatic pre-processing of EST reads, assembly of raw sequences, annotation of the assembled sequences and storage of the analyzed information in SQL databases. A web application was implemented with HTML and a Microsoft .NET Framework C# program for browsing and querying the database, creating dynamic web pages on the client side, analyzing gene ontology (GO) and mapping annotated enzymes to KEGG pathways. The online resources for putative annotation can be searched either by text or by using BLAST, and the results can be explored on the website and downloaded. Consequently, the establishment of OrchidBase will provide researchers with a high-quality genetic resource for data mining and facilitate efficient experimental studies on orchid biology and biotechnology. The OrchidBase database is freely available at http://lab.fhes.tn.edu.tw/est.

Keywords: Expressed sequence tags • KEGG pathways • OrchidBase • Phalaenopsis orchids.

Abbreviations: EST, expressed sequence tag; NGS, next-generation sequencing; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; SOAP, simple object access protocol; TC, tentative consensus.

Introduction

Orchidaceae as a group represents one of the largest families of the flowering plants and the number of species may exceed 25,000 (Atwood 1986). They show a wide diversity of epiphytic and terrestrial growth forms and have successfully colonized almost every habitat on earth. Specific interaction between the orchid flower and pollinator (Cozzolino and Widmer 2005), sequential and rapid interplay between drift and natural selection (Tremblay et al. 2005), the role of obligate orchid–mycorrhizal interactions (Otero and Flanagan 2006) and epiphytism may be the factors promoting orchid species richness. In addition to their specialized pollination and ecological strategies, orchids also show several reproductive characteristics unique in the plant kingdom. These include their flowers being highly evolved with a gynostemium or column (a fusion of the male and female reproductive organs) and a highly modified petal, the labellum or lip, mature pollen grains packaged as pollinia, pollination-regulated ovary/ovule development, synchronized timing of micro- and mega-gametogenesis for effective fertilization, and the release of thousands or millions of immature embryos (seeds without endosperm) in mature pods (Yu and Goh 2001, Tsai and Chen 2006). Despite their unique developmental reproductive biology, as well as specialized pollination and ecological strategies, orchids remain under-represented in
molecular studies relative to other species-rich plant families (Peakall 2007).

Hybrids within the genus *Phalaenopsis* comprise one of the top traded blooming potted plants worldwide. Analyses of floral bud expressed sequence tags (ESTs) in public databases have been initiated for both *P. bellina* and *P. equestris* (Hsiao et al. 2006, Tsai et al. 2006). These collections provide valuable sources for direct access to genes of interest (Tsai et al. 2004, Tsai et al. 2005, Hsiao et al. 2008) and for the development of molecular markers for marker-assisted breeding programs or cultivar identification. However, to date there has been no integrated database to curate finished annotations and simultaneously provide processed ESTs for *Phalaenopsis* orchids.

Large-scale EST sequencing provides a gateway into the genome of organisms owing to the massive amount of information buried in the genome-scale expression data. It is a useful information source for investigating a wide variety of genetic characteristics of a species, such as how many genes exist in the species and how gene expression patterns differ between tissues or developmental stages. By generating and storing ornamental plant ESTs, it is possible to mine the candidate genes associated with ecologically important traits. Recently developed cDNA deep sequencing technologies, such as 454 Life Sciences (Margulies et al. 2005) and the Solexa/Illumina platform (Bennett et al. 2005), have dramatically changed the way in which the functional complexity of transcriptomes is investigated (Delseny et al. 2010). In addition, these technologies are also powerful for identification of genes, structure of transcripts, alternative splicing, non-coding RNAs and new transcription units. Both technologies applied to transcriptome analyses have confirmed that the relatively short reads produced can be effectively assembled and used for gene discovery (Alagna et al. 2009, Wang et al. 2009, Wu et al. 2010).

In this report, we present the web-based EST database, named OrchidBase, which provides integrated information of ESTs from *Phalaenopsis* orchids. The OrchidBase currently contains 84,617 non-redundant transcribed sequences (unigenes) including 8,501 contigs and 76,116 singletons. Besides the common approach of providing data stocks of ESTs, OrchidBase is also integrated comprehensive information, including information of clusters, annotations, gene ontology (GO) and assignment to metabolic pathways based on BLAST similarity searches. Information can be retrieved using text searches, a query assistant or BLAST searches.

**Implementation and architecture**

OrchidBase architecture consists of a cooperating database systems, a Windows application that performs sequences analysis, a web application created using html and the latest .NET (Microsoft .NET framework 4) software technology (asp.net/ c#) which dynamically execute multiple database queries and execute a request by simple object access protocol (SOAP) Web Services (Labarga et al. 2007). It operates under IIS 7.0 on Microsoft Windows Server 2008 R2. Perl and C# program were used to parse data and construct the database automatically. OrchidBase is designed to store and explore the vast amount ESTs, implying complex biological information. Sequence information and corresponding annotation in each library constructed separately from various cDNA libraries can be accessed and searched through the web application.

**Construction and database content**

OrchidBase is a database for Orchid EST data management and analysis. OrchidBase was established from 37,979,342 ESTs collected from 11 in-house *Phalaenopsis* orchid cDNA libraries (Table 1). The detailed library descriptions including the origins and number of sequences can be directly queried from the Release Summary at the website of the database. The web pages also contain detailed information about contigs in each library. Among them, 41,310 ESTs were obtained by using traditional Sanger sequencing, whereas 37,908,032 reads were obtained by using next-generation sequencing (NGS) including both Roche 454 and Solexa Illumina sequencers. These reads were assembled into 8,501 contigs and 76,116 singletons, resulting in 84,617 unigenes with an average length of 459 bp.

To analyze these data, we have built a pipeline using the Perl and C# programming language to construct OrchidBase for comprehensive EST data analyses for raw and processed data (Fig. 1). The pipeline is composed of a series of fully integrated systems and will automatically process, analyze and store the data in a SQL Server 2008 database management system.

The analysis pipeline infrastructure consists of the following modules.

**Pre-processing (Fig. 1, step 1 and step 2)**

We use the Sequencher V. 4.1.2 (Gene Codes Corp.) to extract high-quality regions from raw sequences generated by Sanger sequencing. After confirming EST quality with ‘Low Quality Sequences Trimming’ and having trimmed the vector sequence to obtain high-quality sequence, EST sequences were clustered and assembled into contigs and singletons to reduce inherent redundancy (step 1). The DNA sequence data were assembled using Sequencer. At least 40 bases of overlap in which there was at least 95% identity were required for the assembly.

The NGS technologies were also adapted to generate ESTs. Sequences generated by the 454 pyrosequencing (GS FLX se- quencer) were trimmed to remove adaptor and low-quality sequences. Reads < 50 bp after processing were excluded from the analysis. For assembly, GS FLX gsAssembler was used to construct contigs and singletons under the criteria of 40 bases of overlap in which there was at least 95% identity (step 1).

---

(Continued)
The other high-throughput deep transcriptome sequencing was conducted using Solexa/Illumina RNA-seq. The 75 bp raw pair-end reads were generated by the Illumina Genome Analyzer II system. Sequencing quality of no more than 1% error rate (Q20) was set as a criterion. Adaptor sequences were trimmed using the Cross_Match software. Processed sequences were then assembled into consensus sequences using SOAPdenovo (step 1) by default setting (Li et al. 2010). The respectively constructed contigs were pulled together to assemble by the CLC genomic workbench with at least 50% of overlapping bases in which there was at least 95% identity to build orchid unigenes (step 2).

Functional annotation (Fig. 1, step 3)

Functional annotation was performed using BLASTX through NCBI and our local BLAST server. The filtered ESTs and assembled EST contigs were compared with the NCBI nr database and TAIR database (Rhee et al. 2003) (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9_genome_release/TAIR9_sequences/TAIR9_pep_20090619/) using the BLASTX program, with the E-value set at <1e-7. The successful hits based on the highest scoring of BLASTX results in TAIR were mapped to the GO terms (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt) and KEGG metabolic pathways (ftp://ftp.genome.jp/pub/kegg/genes/organisms/ath/ath_tair.list) by the corresponding locus name using Perl programming. For all successful matches (as ‘the results from BLAST searches’, ‘search results against TAIR’ in Fig. 1), the top five hits and their alignment results were stored in the OrchidBase. The top hit of each BLAST result was automatically extracted and transferred into a relational database.

Information visualization and visual data mining (step 4)

OrchidBase provides high-level information on processed sequence status, functional annotation and for curating the biological meaning assigned to all EST consensus sequences.

Searching the database

The OrchidBase is a highly efficient, web-accessible relational database. It provides several tools to search raw, cleaned and assembled EST sequences, genes and GO, as well as pathway information. The database is freely accessed through a web interface, and data can be queried via three main parts.

(i) Sequence Reports, Search EST: the ‘Libraries’ page allowing users facilitates listing of the raw ESTs, contigs, singletons and non-redundant sequences. The ‘tentative consensus (TC) annotator’ page and the ‘EST Annotator’ page simply list the TC annotation and EST annotation of all analyzed data sets. Users may input and submit keywords or IDs to the server using the web interface. ESTs and annotated function data are in the relational database and results will be sent back to the users in proper formats in response to a query. Search options include simple searches by using keyword, species, accession number or sequence ID (Fig. 2a).

(ii) Functional Annotation and Analysis: in OrchidBase, we used well-annotated GO information to interpret the classification of ESTs by GO vocabularies (Ashburner et al. 2000). The ‘Gene Ontology’ page includes the information for distribution of the three categories (biological processes, cellular components and molecular functions) based on GO for all the sequences.
The 'Metabolic Pathways' page also provides access to the pathway information in KEGG (Kanehisa et al. 2004). This is useful for mapping and investigating the relationships among a whole system of annotated EST sequences and is especially valuable for those who are interested in biological pathways (Fig. 2b).

(iii) Sequence Similarity Search: BLASTN, TBLASTN and TBLASTX searches can also be performed against OrchidBase. The search output contains information about the EST and contig names, and sorts by e-value and sequence ID (Fig. 2c).

Conclusions
The OrchidBase is the first incorporated online information database of orchid EST sequences that can be freely accessed and downloaded. The orchid EST database is not only a uniquely comprehensive world-wide orchid EST database, but also the most useful one which includes sufficient information on genes that are representative of the characteristics of orchids which are of use for breeding purposes. In addition, it includes important information for studying the properties of orchid genes at the molecular level and gene function level. Moreover, it is a tool for information retrieval, visualization and management. Consequently, the development of OrchidBase will provide a significant contribution to biologists for data mining and can lead to efficient experimental studies.

Availability
The OrchidBase database is freely available at http://lab.fhes.tn.edu.tw/est. All questions, comments and requests should be sent by e-mail to tsaiwc@mail.ncku.edu.tw.
Funding
This work was supported by the National Science Council, Taiwan [grants NSC96-2317-B-006-003 and NSC97-2317-B-006-002 to H.H.C. and W.C.T.].

References


Fig. 2 Snapshots of the OrchidBase web interface. (a) Sequence Reports, Search EST page; (b) Functional Annotation and Analysis page. An example of the tree menu listing all the metabolic pathways and showing the pathway schema where EST sequences highlighted in red were mapped; (c) Sequence Similarity Search page.


