Journal of Animal and Veterinary Advances 10 (5): 659-662, 2011
ISSN: 1680-5593

PGED: A *Porcine* Gene Expression Data Repository Based on Affymetrix Genechip

[1, 2]Haiyan Wang, [1]Jianhua Cao and [1]Shuhong Zhao
[1]Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education,
Key Laboratory of Swine Genetics and Breeding of Ministry of Agriculture,
HuaZhong Agricultural University, 430070 Wuhan, P.R. China
[2]Department of Computer Science, HuaZhong Agricultural University, 430070 Wuhan, P.R. China

Abstract: The *porcine* Gene Expression Database (PGED) is an open-source web data repository specially for *porcine* gene expression data. It archives *porcine* gene expression data in different experiments and associated functional annotation. A user-friendly web interface allows users to browse, retrieve, analyze and visualize data which is intended to facilitate to identify genes of interest and explore their expression profiles.

Key words: Swine, database, gene expression, functional annotation, *porcine* gene, China

INTRODUCTION

Expression datasets derived from high-throughput experiments can be utilized to find a set of genes differentially expressed in a particular condition and generate new hypotheses or inferences of gene function. Microarray and other high-throughput methodologies have generated large amount of transcriptome data in the last decade. These data is often obtained from biological samples under a variety of experimental conditions which adds more dimensions of difficulties for data management and processing.

In compliance with the MIAME standards (Brazma *et al*., 2001) large amounts of gene expression data to publish are required to deposit in a public repositories such as the Gene Expression Omnibus (GEO) (Barrett *et al*., 2009) or ArrayExpress (Parkinson *et al*., 2009). The databases generally contain massive volumes of experimental and analytical data produced by various experiments. Users need to download the relevant datasets and re-analyse them in order to find differentially expressed genes under particular conditions which makes it difficult to effectively look for interested genes. The GEO supports queries for gene expression profiles however, it does not provide queries for genes specific to a particular biological condition. Moreover, it does not allow straightforward visualization of expression levels. ArrayExpress provides retrieval of condition-specific gene expression patterns but it lacks relevant probesets information and cross-platform annotation functions.

Other databases such as ANEXdb (Couture *et al*., 2009) or BioGPS (Wu *et al*., 2009) have housed large gene expression datasets from the public data resources, re-analyzed and displayed them through various interfaces. Most of these databases are specific to particular biological domains. ANEXdb is a porcine-specific expression database that houses microarray expression and EST annotation data. It supports retrievals of Affymetrix-based expression data however, it also does not provide visualization of expression data. BioGPS is a comprehensive database integrating disparate gene annotation resources based on microarray chips. It primarily focuses on data for human, mouse and rat genes however, it does not contain the porcine data.

Pig (*Sus scrofa*) is an important model organism for health science researches for its many similarities with humans in physiology, anatomy and size. Transcriptomics studies in the pig have greatly contributed to formulate hypotheses about the role of genes.

Here, introduce the *porcine* Gene Expression Database (PGED), a unique *porcine* gene expression database based on a seamless fusion of genetic and genomic resources. The PGED is designed to archive *porcine* gene expression data under various biological conditions including disease states and developmental stages.

The gene expression data derived from the use of the Affymetrix GeneChip. The Affymetrix GeneChip® Porcine Genome Array (https://www.affymetrix.com/products_

Corresponding Author: Shuhong Zhao, Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Key Laboratory of Swine Genetics and Breeding of Ministry of Agriculture,
HuaZhong Agricultural University, 430070 Wuhan, P.R. China

services/arrays/specific/porcine.affx) which contains over 23 K probesets that interrogate transcripts representing approximately 20, 201 *S. scrofa* genes and has become a popular tool for systematic study of the *Sus scrofa* transcriptome biology.

To maximize the use of publicly-available Affymetrix GeneChip data, the PGED also incorporates associated functional annotation for gene expression data from public databases such as NCBI and DAVID (Dennis *et al*., 2003). The PGED provides a range of web-based browse, query, analysis and visualization functions for identifying genes of interests and exploring their expression profiles.

Implementation and architecture: The PGED is developed on a Linux server (Fedora Core 12 distribution) with Apache as web server. The open source MySQL server 5.0 is used as the backend DBMS. Managements of the data interface are implemented with phpMyAdmin and the MySQL client.

The PGED is built following a three-tiered software architecture (Fig. 1). The architecture consists of a data abstraction layer, a business layer and a presentation layer. The data abstraction layer, implemented with PHP and Perl, contains data processing modules for web session, program process and communications with the database server and the Blast server. The Blast server is built using the NCBI BLAST toolkit wwwBLAST (ftp://ftp.ncbi.nih.gov/blast/executables/LATEST/). The business layer is also programmed by PHP and Perl for retrieval, computation and parsing data from other resources. The presentation layer is responsible for receiving user requests and rendering web pages as well as visualizing gene expression data. The presentation layer is implemented with PHP and Java Scripts.

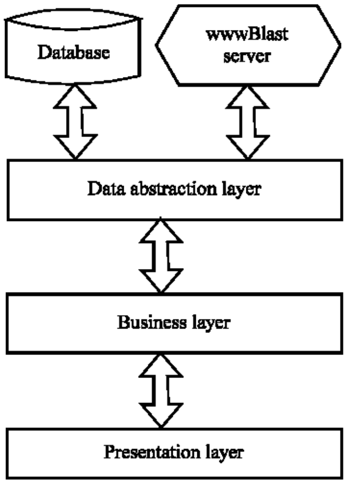

Fig. 1: The software architecture of the PGED

Database content: The *porcine* gene expression data stored in the PGED are based on the Affymetrix GeneChip data structure. The PGED also integrates associated annotation of individual transcripts to help elucidate their functions. Figure 2 shows an overview of all data types comprised by the PGED web server. The information in the web server is compiled into different database tables and the tables are organized through common primary/foreign key pairs to facilitate interpreted joint query of multiple datasets.

The PGED currently hosts expression data from 3 experiments represented by 27 GeneChips including 12 GeneChips hybridized with Erhualian and Large White placentas on day 75 and 90 of gestation (Zhou *et al*., 2009), 6 GeneChips detected gene expression in porcine spleen under *Haemophilus parasuis* infection (Chen *et al*., 2009) and 9 GeneChips detected gene expression in porcine muscle on 33, 65 days of gestation and adult (Huang *et al*., 2008).

All raw chip data in CEL format were converted to gene signal files by MAS 5.0 (microarray analysis system 5.0, Affymetrix). The data were normalized between slides using the quantile normalization method (Bolstad *et al*., 2003) in an R routine. The PGED displays the expression values, standard deviations and p-values for each probeset, linking each gene to each experimental condition.

The probeset data obtained from Affymetrix contains probeset ID and their details, consensus sequences,

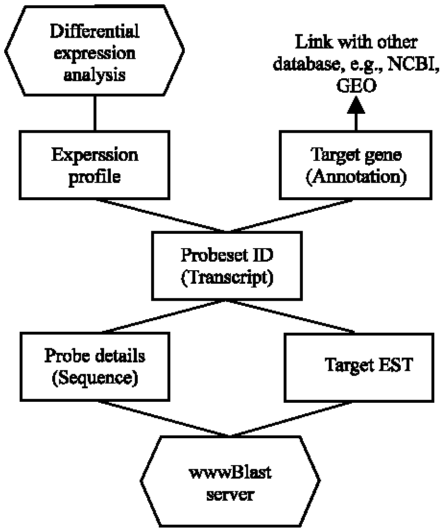

Fig. 2: The data types in the PGED. Ractangles indicate different data types with corresponding search functions. Pentagons show the major analytical functions provided by the PGED

target sequences and their initial annotation. Since the gene functional annotation is critical for differential expression analysis, we have integrated annotations of target gene from NCBI and DAVID.

Database functions: The PGED displays the complete set of target sequence in one table. Users can browse the associated properties of each sequence manually. Moreover, the PGED provides a flexible interface to query, analyze and visualize data described as follows:

Search by ID or functional annotation: Users may enter one or multiple gene identifiers (Affymetrix ID, gene symbol, entrez ID and so on) to view associated features such as annotations and expression changes. Functional annotations with natural language such as Gene Ontology terms can also be used this way. Selecting a gene out of the search result list will show a gene annotation report.

Search by expression levels: The PGED offers expression profiles queries for a particular biological condition. For example, users can query all genes up-regulated (p value<0.05 and FC>2) in porcine spleen under *Haemophilus parasuis* infection. The search returns 412 results including probeset ID, gene names with corresponding expression data.

Moreover, the PGED supports differential expression analysis with a flexible fold-change option. The fold-changes with a custom threshold value could intuitively reflect different degrees of either up-regulation or down-regulation. Users can effectively retrieve interesting genes with the fold-changes option.

Differential expression visualization: Gene expression values of each probeset in various biological conditions are visualized using histograms to compare the expression difference. The thumbnail plots provide a direct link to show information about the expression profiles of the selected gene in the associated experimental condition.

Sequence alignment: To facilitate to identify the *porcine* GeneChip potential homologous sequences related to genes of other model organisms such as human and mouse, the PGED offers sequence alignment via Blast. Users could carry out a Blast search against the interesting target sequence to find the corresponding homologs.

CONCLUSION

The PGED creates a single location to host gene expression data and associated annotation in differential expression experiments. These data has systematically normalized and stored for effective comparative and integrative analyses across datasets from different studies. The PGED will be useful to easier and faster analysis of the large amounts of data generated through high-throughput expression experiments. Moreover, it will improve the power of investigators to identify their interesting gene from different biological angles and aid efforts to interpret the porcine genome through functional genomics.

The PGED will be under consistent development to extend the scope of data and analysis functions. In the near future, the PGED will expand to cover available Affymetrix GeneChip data in the pig and provide a comprehensive resource for porcine functional genomics.

ACKNOWLEDGEMENTS

This research was financially supported by the National High Science and Technology Foundation of China 863 (2006AA10Z195 and 2008AA101008), the NSFC-Guangdong Key Project (U0631005), the Key Project of National Basic Research and Developmental Plan (2006CB102105) of China and the creative team project of education Ministry of China (IRT0831). We thanks Dr. Zhiliang Hu and Pro. Lingling Chen for valuable suggestions.

REFERENCES

Barrett, T., D.B. Troup, S.E. Wilhite, P. Ledoux and D. Rudnev *et al*., 2009. NCBI GEO: Archive for high-throughput functional genomic data. Nucleic Acids Res., 37: 885-890.

Bolstad, B.M., R.A. Irizarry, M. Astrand and T.P. Speed, 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19: 185-193.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock and P. Spellman, 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet., 29: 365-371.

Chen, H., C. Li, M. Fang, M. Zhu and X. Li *et al*., 2009. Understanding Haemophilus parasuis infection in porcine spleen through a transcriptomics approach. BMC Genomics, 10: 64-64.

Couture, O., K. Callenberg, N. Koul, S. Pandit and R. Younes, 2009. ANEXdb: An integrated animal ANnotation and microarray EXpression database. Mamm. Genome, 20: 768-777.

Dennis, G., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane and R.A. Lempicki, 2003. DAVID: Database for annotation, visualization and integrated discovery. Genome Biol., 4: R60-R60.

Huang T.H., M.J. Zhu, X.Y. Li and S.H. Zhao, 2008. Discovery of porcine microRNAs and profiling from skeletal muscle tissues during development. PLoS. One, 3: e3225-e3225.

Parkinson, H., M. Kapushesky, N. Kolesnikov, G. Rustici and M. Shojatalab *et al*., 2009. ArrayExpress update-from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res., 37: D868-D872.

Wu, C., C. Orozco, J. Boyer, M. Leglise and J. Goodale *et al*., 2009. BioGPS: An extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol., 10: R130-R130.

Zhou. Q.Y., M.D. Fang, T.H. Huang, C.C. Li, M. Yu and S.H. Zhao, 2009. Detection of differentially expressed genes between Erhualian and large white placentas on day 75 and 90 of gestation. BMC Genomics, 10: 337-337.