Databases and ontologies

ZooDDD: a cross-species database for digital differential display analysis

Yun-Ching Chen*1, Chung-Der Hsiao*2, Wen-Dar Lin1, Chung-Ming Hu1, Pung-Pung Hwang#2 and Jan-Ming Ho#1

1Institute of Information Science, Academia Sinica, NanKang 115, Taipei, Taiwan, ROC. 2Institute of Cellular and Organismic Biology, Academia Sinica, NanKang 115, Taipei, Taiwan, ROC.

ABSTRACT
SUMMARY: In this paper, we combined EST information from the UniGene database and orthologous relationships from the Ensembl database to construct a ZooDDD database. The primary function of ZooDDD is to mine evolutionary conserved, highly expressed, tissue-specific orthologues in model animals. The candidate genes of interest derived from the ZooDDD database will provide biologists with a good step for comparing the expression, functions, and evolution of animal genomes.

Contact: cyc@iis.sinica.edu.tw

1 INTRODUCTION

Although the human genome was decoded in April 2003, the biological functions of most human genes are still unclear (Sauer et al., 2005). In order to fully explore human gene function, model animals like mouse, chicken, Xenopus, zebrafish, drosophila, and nematodes have been successfully utilized to perform reversed (by gene knockout or knockdown) and forward (by mutant screening) genetic assays on human orthologues (Cullen and Arndt, 2005; Drummond, 2005; Michno et al., 2005). However, during vertebrate evolution, several rounds of genome duplication and rearrangement have occurred, leading to the expression and function of some orthologues being shuffled by paralogues (Prince and Pickett, 2002) or partitioned by co-orthologues (Rastogi and Liberles, 2005). These complex genomic evolutionary processes have increased genetic diversity and possibly benefited diverse vertebrate species in dealing with specific developmental and physiological requirements. Taking this into consideration, it is worthwhile comparing the expression patterns of orthologues and paralogues between model animals and humans in order to gain insights into the possible functions of their human counterparts.

NCBI UniGene is a database which groups expressed sequence tags (ESTs) into clusters, each of which represents a unique gene (Pontius et al., 2003). Because ESTs grouped in UniGene clusters are tagged with information related to tissue origin and/or developmental stage, NCBI has developed a web-based tool of digital differential display (DDD) to mine tissue- and stage-specific gene transcripts. However, the current version of DDD only enables users to search differentially expressed gene transcripts in a single species. For this study, we have constructed a ZooDDD database by merging the EST information from UniGene and the orthologous relationships from the Ensembl database. The establishment of the ZooDDD database makes it possible to mine evolutionarily converged orthologues with tissue- and stage-specific expression patterns and will provide scientists with an excellent step for comparing the functions, regulation, and evolution of animal genomes.

2 METHODS

Taking advantage of the rich bioinformatics of ESTs and genomes in model animals, we selected nine vertebrates to construct the ZooDDD database: human (Homo sapiens), mouse (Mus musculus), rat (Rattus norvegicus), cattle (Bos taurus), dog (Canis familiaris), chicken (Gallus gallus), frog (Xenopus tropicalis), zebrafish (Danio rerio), and tunicate (Ciona intestinalis). Basically, different cDNA libraries generated from the same tissues were combined and selected for further studies when the total number of merged ESTs exceeded 10,000. Next, we broke down ESTs grouped in particular UniGene clusters according to their tissue origins. The relative expression levels of the ESTs in each UniGene cluster were further normalized to the level of transcripts per million. To further narrow down the relative number of interesting UniGene clusters, we also calculated the tissue- or stage-specificity of each UniGene cluster among different cDNA libraries. Finally, we mapped the UniGene clusters to the Ensembl transcript using BLAST and utilized the orthologous relationship built up by Ensembl to perform cross-species comparisons. Below, we use two UniGene clusters of Homo sapiens Hs.436385 (Solute carrier family 22) and Dr.13231 (Zgc:64076) as examples to describe the key steps in constructing the ZooDDD database.

(1) Breaking down ESTs grouped in UniGene according to their tissue origins. For each animal species, ESTs were downloaded from the NCBI UniGene database and merged into single cDNA libraries based on whether they were generated from the same tissue. To enhance the comparison quality, merged cDNA libraries with fewer than 10,000 ESTs were discarded and not used for further analysis. Next, we broke down ESTs grouped in particular UniGene clusters according to their tissue origin. Taking zebrafish Dr.13231 as an example, there were 34 ESTs derived from the kidney library and two ESTs from the liver library. For human UniGene Hs.436385, there were 43 ESTs from the kidney library and few ESTs from other libraries.

(2) Normalizing the expression levels and calculating the specificities of the UniGene clusters. Because the numbers of ESTs generated from various tissues differed, there needs to be a normalization process for making further comparisons more reasonable. We followed the concept of transcripts per million (TPM) which was introduced by the NCBI to normalize gene expression levels. Initially, we normalized the numbers of ESTs generated from different tissues to one million. Next, the TPM value of UniGene clusters in particular tissues was calculated as:
For example, there were 13,568 and 48,447 ESTs in the zebrafish liver and kidney cDNA libraries. After normalization, the relative expression levels of Dr.13231 in the liver and kidney were determined to be 147 ($= (2/13,568 \times 1,000,000)$) and 701, while the relative expression level of Hs.436385 in the kidney was 237. In addition, we also defined the expression specificity of UniGene clusters in particular tissue as

$$\text{Specificity} = \frac{\text{TPM of the tissue in the UniGene}}{\text{Sum of TPMs in the UniGene}} \times 100\%$$

For example, the kidney-specific expression scores of Dr.13231 and Hs.436385 were 82% ($= 701/(701 + 147) \times 100\%$) and 89%, respectively.

(3) Mapping UniGene clusters to Ensembl transcripts using the BLAST program

Basically, four types of prediction methods in terms of unique best reciprocal hit (MRBH), multiple best reciprocal hit (MBRH), reciprocal best hit based on synteny information (RHS), and derived from whole-genome alignment (DWGA) are employed by the Ensembl database to predict orthologous relationships among diverse vertebrate species. Initially, we defined the representative ESTs from each UniGene cluster as query sequences and the Ensembl gene transcripts as the subject sequences. Then representative ESTs were compared using BLASTN (Altschul et al., 1990) against the Ensembl database with the parameter of the E-value < $e^{-100}$ to find the best matching gene transcript. Results demonstrated that 65% of human UniGenes and 56% of zebrafish UniGenes could be mapped to Ensembl transcripts. For example, Hs.436385 was mapped to ENST00000345033, and Dr.13231 was mapped to ENSDART00000048890. By utilizing the built-in orthologous relationship predicted in the Ensembl database, we learned that human ENST00000345033 and zebrafish ENSDART00000048890 have a UBRH-orthologous relationship. Therefore, UniGene clusters of Hs.436385 and Dr.13231 are defined as being highly expressed and kidney-specific orthologous genes by the ZooDDD database.

### 3 DATA PRESENTATION

The program used to construct the ZooDDD database was implemented in PERL script v5.8. The expressions of the UniGene clusters as well as the orthologous relationships are stored in the MySQL database. The ZooDDD website is very user-friendly and was designed for both novices and experts. New users can choose an interesting species to perform single-species DDD analyses as an entry point (Fig. 1a), and ZooDDD will produce a table to summarize the UniGene clusters which fit different specificity settings (Fig. 1b). By clicking appropriate specificity settings, the system returns a list of matches arranged in descending order according to their expression levels (Fig. 1c). For cross-species ZooDDD analysis, users can further choose another species to find orthologues and set parameters to constrain gene expression patterns. For practiced users, one can bypass the previous steps and directly proceed to the cross-species comparison step with precise queries by setting the TPM and specificity. The results returned from ZooDDD include UniGene ID, NCBI descriptions, and related Ensembl proteins (Fig. 1d). ZooDDD also provides descriptions of gene ontology (GO), InterPro domains, Ensembl descriptions, as well as Online Mendelian Inheritance in Man (OMIM) information with which users can perform data mining. The outcomes listed in the results table are also hyperlinked to their original URL site for further investigation. In addition, ZooDDD also uses a JAVA-based Gene Ontology Browsing Utility (GOBU) (Lin et al., 2006) for users to browse related GO terms of the matching UniGene clusters. Finally, ZooDDD automatically and periodically updates ESTs and orthologous relationships from the UniGene and Ensembl databases.

### ACKNOWLEDGEMENTS

We thank Drs. Ming-Jing Hwang and Wen-Chang Lin at the Institute of Biomedical Sciences, Academia Sinica for fruitful discussions, and NSC 93-2317-B-001-008 for support.

### REFERENCES


