BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 9 2010, pages 1273–1275
doi:10.1093/bioinformatics/btq109

Databases and ontologies

Advance Access publication March 11, 2010

TiSGeD: a database for tissue-specific genes

Sheng-Jian Xiao[1], Chi Zhang[1], Quan Zou[1] and Zhi-Liang Ji[1,2,*]

[1]Key Laboratory for Cell Biology and Tumor Cell Engineering, the Ministry of Education of China, School of Life Sciences and [2]The Key Laboratory for Chemical Biology of Fujian Province, School of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, Fujian, P. R. China

Associate Editor: Jonathan Wren

ABSTRACT

Summary: The tissue-specific genes are a group of genes whose function and expression are preferred in one or several tissues/cell types. Identification of these genes helps better understanding of tissue–gene relationship, etiology and discovery of novel tissue-specific drug targets. In this study, a statistical method is introduced to detect tissue-specific genes from more than 123 125 gene expression profiles over 107 human tissues, 67 mouse tissues and 30 rat tissues. As a result, a novel subject-specialized repository, namely the tissue-specific genes database (TiSGeD), is developed to represent the analyzed results. Auxiliary information of tissue-specific genes was also collected from biomedical literatures.

Availability: http://bioinf.xmu.edu.cn/databases/TiSGeD/index.html

Contact: appo@bioinf.xmu.edu.cn; zhiliang.ji@gmail.com

Received on December 3, 2009; revised on February 18, 2010; accepted on March 7, 2010

1 INTRODUCTION

The spatial and temporal variation in gene expression carries crucial information of what the gene does (Bassett *et al.*, 1999). Thanks to the emergence of high-throughput technologies like microarray, the expressions of thousands of genes in multiple tissues can be now monitored simultaneously. Databases were thus constructed for storing, representing and retrieving the information. Upon these databases, a series of effective analysis tools were developed as well to explore and interpret the massive amounts of data in the context of existing knowledge. These include some sophisticated algorithms that are able to 'align' or 'cluster' the gene expression profiles so as to identify those non-trivial patterns such as correlated expression, differential expression and specific expression (Barrett *et al.*, 2007; Parkinson *et al.*, 2009; Vaquerizas *et al.*, 2005; Wang *et al.*, 2006). Some even mapped the gene expression profiles onto the regulatory, metabolic and cellular pathways (Liu *et al.*, 2008; Mlecnik *et al.*, 2005; Slonim, 2002; Van Steensel, 2005). Nowadays, there exist a lot of databases containing tissue-specific expression information. Some databases like GEO (Barrett *et al.*, 2007) and ArrayExpress (Parkinson *et al.*, 2009) generally deposit microarray datasets produced by various experiments. Some databases like TiGER (Liu *et al.*, 2008), BODYMAP (Ogasawara *et al.*, 2006), BioGPS (Su *et al.*, 2002) and TissueDistributionDBs, however, mainly collect datasets of tissue-specific expressions. Many of these databases enable users to visualize and statistically analyze gene expression profiles. Some further provide special tools to facilitate the detection of non-trivial patterns (Liu *et al.*, 2008). In this study, we introduce a novel subject-specialized database, namely the tissue-specific genes database (TiSGeD), which particularly provides information of tissue-specific genes derived from literature and data mining of gene expression profiles.

*To whom correspondence should be addressed.

2 THE DATA

The microarray datasets and their latest gene chip annotation files were derived from public microarray repositories GNF BioGPS, NCBI GEO and EBI ArrayExpress. In current version of database, eight high-quality datasets were chosen with criteria of: (i) proper normalization of profiles by MAS or GCRMA methods, (ii) large tissue scales, and (iii) enough 'fluctuation' of gene expressions over multiple tissues. These criteria ensure the datasets reliable, useful and representative. Removing those incomplete or defective profiles, the TiSGeD release 1.0 currently compiled more than 123 125 distinct gene expression profiles over 107 human tissues, 67 mouse tissues and 30 rat tissues. In addition to microarray data, tissue-specific genes were also manually collected by keyword search of NCBI biomedical literature database PubMed.

3 THE METHOD

Strictly, the tissue-specific gene is defined as gene whose function and expression is restricted to a particular tissue or cell type. However, in many cases, the specificity concept is broadened to tissue selectivity that gene expression is enriched in one or several tissues/cell types (Shuang *et al.*, 2006). In TiSGeD, we presented genes of both strictly specific and highly selective in tissues.

Briefly, the tissue-specific gene is detected by solving a linear algebra problem of scalar projection in this study. First, transform each expression profile of gene x into a vector X_p:

$$X_p = (x_1, x_2, x_3, \ldots, x_i, \ldots, x_n) \quad (1)$$

where n is the number of tissues in the profile and x_i is the gene expression level in tissue i. For each element in the profile X_p, the expression x_i can also be represented by a vector X_i in high-dimension tissue spaces:

$$X_i = (0, 0, 0, \ldots, x_i, \ldots, 0) \quad (2)$$

Then, the tissue specificity of gene x in tissue i is determined by calculating the ratio of vector X_i's scalar projection in the direction of vector X_p (i.e. $||X||$) against the length of X_p (i.e. $||X_p||$):

$$||X|| = |(X_i \bullet X_p)/||X_p||\,|; \quad (3)$$

$$\text{Specificity measure (SPM)} = ||\mathrm{X}||/||\mathrm{X}_p||. \quad (4)$$

Theoretically, the SPM ranges from 0 to 1.0. A value close to 1.0 indicates that element x_i is the major contributor to the length of profile X_p in high-dimension tissue spaces; in biological term, high tissue specificity.

In practice, user can rely on the SPM value to quantitatively estimate the tissue specificity of a gene in a profile regardless of its absolute expression level. The larger the SPM value, the higher the tissue specificity is. However, in the cases of profiles with 'spiked' expression patterns (which gene expressions are highly selective in several similar tissues), a large SPM value may not be achieved. As a feasible solution, the original gene expression profiles are reformatted by merging several similar subtissues into an 'integrative' tissue (individual expressions are summed up as the expression of the representative) according to the organ hierarchy tree adopted in this study. For example, intestine represents for small intestine, large intestine, etc. As a result, the 'spiked' expression pattern can then be detected by SPM analysis as an enriched expression.

4 DATABASE DESCRIPTION

The TiSGeD database is now accessible at http://bioinf.xmu.edu.cn/databases/TiSGeD/index.html. It was curated on Red Hat Linux release 9 operating system. The data were managed by the RDBMS Oracle 10 g. Interactive user interfaces and search engines were coded by PHP or JavaScript. Two methods were developed for rapid access of the TiSGeD database. They are briefly described as follows:

A standard search method is designed to retrieve information via keyword query forms. User is required to type partial or full keywords in the respective text fields of either gene symbol or tissue name. Wild-card characters like '*, &, ?' are not supported. At the same time, a positive threshold SPM value, e.g. 0.9 (highly tissue specific), is required to initiate the search. Once a keyword and a valid SPM value are submitted, a list of gene symbols or tissue names that meet the query criteria will be responded in alphabet order, respectively. Clicking on the gene symbol or tissue name will finally lead to the detailed information page of tissue-specific gene. The literature records are always given regardless of the threshold SPM value. For the user who has no specific searching intention, a foolproof quick search method is available for flexible access of database by just providing a keyword or keyword combination.

TiSGeD also offers an alternative browse method for direct retrieval of tissue-specific genes. User is allowed to query the tissue-specific genes by species and tissues/cell types. Data of three species (human, mouse and rat) are now available for browsing. Every tissue in these three species was non-redundantly assigned into branches of the physiological organ hierarchy trees and arranged in alphabet order. All cell types were put under an isolated branch of 'Cells'. The branches of the trees are expandable up to maximum five levels. Via the tissue hierarchy trees, user can easily browse information by tissues. Before that, a threshold SPM value is required in advance. Clicking on a tissue name, genes that specifically/selectively expressed in this tissue will be listed in descending order of their SPM values.

The detailed information page is presented in sections of Gene Information, Microarray/Literature Evidences and Homologenes. The Gene Information section contains brief information of gene including gene symbol, gene description and UniGene ID. Cross-links to the Gene Ontology database and the UniGene database via gene symbol or UniGene ID are also provided. In the Microarray/Literature Evidences section, records of the tissue-specific expressions are listed one by one in descending order of SPM values. Each record contains the information of probeset_ID, SPM value, species, dataset and tissue while available. The respective gene expression profiles are further analyzed and illustrated in charts. The literature evidences are also presented in this section while available, including the contents of experiment method, condition, material, PubMed ID, etc. In the last section of Homologenes, the TiSGeD lists the homologous genes in human, mouse or rat while available for quick access and comparison of their expressions in tissue.

Currently, total of 6782 distinct tissue specifically expressed genes (SPM > 0.9) were identified via SPM analysis of microarray data, including 2032 human genes, 4229 mouse genes and 521 rat genes. Identification of these tissue-specific genes facilitates our better understanding of gene function. For example, the myosin light chain 3 (Myl3) gene was reported to function in cardiac muscle contraction and may play a role in cardiomyopathy. This gene is detected consistently as highly heart-specific gene in human, mouse and rat (SPM = 1.00, 0.99 and 0.97, respectively) in TiSGeD. Villin-1 is the Ca^{2+}-regulated actin-binding protein and major component of microvilli of intestinal epithelial cells and kidney proximal tubule cells. In TiSGeD, gene of Villin-1 (Vil1) is found highly specific in small intestine (much more than in large intestine and kidney) of mouse, selective in liver (especially in HepG2 cell lines) and kidney of human and non-selective in tissues of rat. The different tissue specificity of Vil1 in animal models suggests that this gene may play various physiological roles.

5 DATABASE COMPARISON AND UPDATE

According to our survey, most microarray databases, e.g. GEO, ArrayExpress Altas and BioGPS, do not provide similar information as TiSGeD does. Comparing to the TiGER which used EST data for assessment of the tissue-specific expression, the TiSGeD mainly adopted microarray data on a much broader tissue scale. Worthy of mention, TiSGeD is the only one among the above databases that provides auxiliary literature information.

TiSGeD is scheduled to be updated regularly. More high-quality microarray datasets will be included into the database continually. The literature data, however, will be updated yearly.

6 DISCUSSION AND CONCLUSION

In this study, we presented a novel database of tissue-specific genes as a useful resource for the communities of functional genomics, system biology, etiology and drug discovery. A statistical parameter SPM is introduced as a sensitive indicator in quantitative estimation of gene spatial expression patterns. This simple idea is also promising on the identification of time-specific/selective genes during development when high-quality microarray datasets are available in the future. However, the SPM value does not reflect the real expression level of a gene since all profiles are transformed and projected into a linear region of 0–1.0. This weakness may be made up by introducing another statistical

parameter to indicate the relative expression level of a gene against all genes in a dataset. Besides, sometimes the assignment of tissue-specific genes is inconsistent. This is mainly caused by the instability of microarray technology in detecting genes at a low level. The different tissue scales between experiments are also responsible for the inconsistence. Therefore, introduction of a scoring system to evaluate the detection of tissue-specific genes is also desired in the future. Furthermore, SAGE is another high-throughput technology that is popularly used in accurate measurement of gene expressions. It may serve as a promising source in detection of the tissue (time) specific genes if the SAGE experiments are demonstrated in relatively large number of tissue (time) samples and in a simultaneous manner.

ACKNOWLEDGEMENTS

The support from the National Natural Science Foundation of China (NSFC#30873159 to Z.-L.J.), the Program for New Century Excellent Talents (NCET) of MOE (to Z.-L.J.) and the Science Planning Program of Fujian Province (2009J1010) are gratefully acknowledged.

Conflict of Interest: none declared.

REFERENCES

Barrett,T. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

Bassett,D.E. *et al.* (1999) Gene expression informatics–it's all in your mine. *Nat. Genet.*, **21**, 51–55.

Liu,X. *et al.* (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.

Mlecnik,B. *et al.* (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.*, **33**, W633–W637.

Ogasawara,O. *et al.* (2006) BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. *Nucleic Acids Res.*, **34**, D628–D631.

Parkinson,H. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.

Shuang,L. *et al.* (2006) Detecting and profiling tissue-selective genes. *Physiol. Genomics*, **26**, 158–162.

Slonim,D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.*, **32**, 502–508.

Su,A.I. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.

Van Steensel,B. (2005) Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat. Genet.*, **37**, S18–S24.

Vaquerizas,J.M. *et al*. (2005) GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res.*, **33**, W616–W620.

Wang,Y.P. *et al.* (2006) GEPS: the Gene Expression Pattern Scanner. *Nucleic Acids Res.*, **34**, W492–W497.