Bioinformatics, 35(1), 2019, 172–174
doi: 10.1093/bioinformatics/bty576
Advance Access Publication Date: 7 July 2018
Applications Note

OXFORD

Databases and ontologies

BrainEXP: a database featuring with spatiotemporal expression variations and co-expression organizations in human brains

Chuan Jiao[1], Pengpeng Yan[2,3], Cuihua Xia[1], Zhaoming Shen[2], Zexi Tan[2], Yanyan Tan[2], Kangli Wang[1], Yi Jiang[1], Lingling Huang[1], Rujia Dai[1], Yu Wei[1], Yan Xia[1], Qingtuan Meng[1], Yanmei Ouyang[2], Liu Yi[1], Fangyuan Duan[1], Jiacheng Dai[1], Shunan Zhao[1], Chunyu Liu[1,4,*] and Chao Chen[1,5,*]

[1]Center for Medical Genetics, School of Life Science, Central South University, Changsha 410012, China, [2]Hunan Qingpeng International Information Technology, Changsha 410000, China, [3]Qingres Limited, London EC1A 4EN, UK, [4]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY 13201, USA and [5]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, Hunan 410012, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 1, 2017; revised on June 14, 2018; editorial decision on July 3, 2018; accepted on July 5, 2018

Abstract

Summary: Gene expression changes over the lifespan and varies among different tissues or cell types. Gene co-expression also changes by sex, age, different tissues or cell types. However, gene expression under the normal state and gene co-expression in the human brain has not been fully defined and quantified. Here we present a database named Brain EXPression Database (BrainEXP) which provides spatiotemporal expression of individual genes and co-expression in normal human brains. BrainEXP consists of 4567 samples from 2863 healthy individuals gathered from existing public databases and our own data, in either microarray or RNA-Seq library types. We mainly provide two analysis results based on the large dataset: (i) basic gene expression across specific brain regions, age ranges and sexes; (ii) co-expression analysis from different platforms.

Availability and implementation: http://www.brainexp.org/

Contact: liuch@upstate.edu or chenchao@sklmg.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Comprehensive knowledge about the pattern of gene expression in normal human brain ('normal' means no neuropsychiatric disorders) is fundamental to understand the molecular function and physiology of human brain. Spatiotemporal analyses of human brain have identified many genes differentially expressed among brain regions (Johnson *et al.*, 2009; Kang *et al.*, 2011; Miller *et al.*, 2014), age stages (Johnson *et al.*, 2009; Kang *et al.*, 2011; Miller *et al.*, 2014) and sexes (Kang, *et al.*, 2011). Quantification of the spatiotemporal expression in human brain is essential for understanding the neurodevelopment, gender differences and neuropsychiatric disorders.

Gene co-expression, which may reflect regulatory relationships among genes, was also implicated in neurodevelopment (Bakken *et al.*, 2016; Johnson *et al.*, 2009; Miller *et al.*, 2014) and the etiology of multiple brain disorders (Bakken *et al.*, 2016; Tebbenkamp *et al.*, 2014). Weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) is a popular method to construct the gene co-expression network for identifying sets of

genes with correlated patterns among samples. Genes with high topological overlap matrix named modules can be defined by this method through the dynamic tree cut algorithm. Modules of coordinated expression of genes and co-regulation relationships can be identified.

Here, we developed a novel database called Brain EXPression (BrainEXP), which provides the normal brain expression levels and co-expression analysis results for reference. Compared with other databases of the same type (Colantuoni *et al.*, 2011; Higgs *et al.*, 2006; Miller *et al.*, 2014), BrainEXP provides a more in-depth analysis based on the largest sample-sizes. Combined data from several published databases and our own data were strictly analyzed by consistent workflow. Thus, large sample-sizes and abundant brain regions or age stages are included in BrainEXP. BrainEXP is the first database that supplies gene co-expression information based on WGCNA in the human brain data.

2 Materials and methods

Currently, BrainEXP contains 4567 normal human brain samples of 2863 normal individuals from both our own data and existing public databases, including the Gene Expression Omnibus database (Barrett *et al.*, 2013), ArrayExpress (Kolesnikov *et al.*, 2015), Genotype-Tissue Expression project (Consortium, 2013), Brain Cloud (Colantuoni *et al.*, 2011), Brainspan (Miller *et al.*, 2014), Stanley Medical Research Institute online genomics database (SMRIDB) (Higgs *et al.*, 2006) and BrainGVEx (Psych *et al.*, 2015). Datasets with fewer than 15 samples were not included. The samples' ages range from embryonic stages to late adulthood, involving 56 brain regions. The raw expression data were measured by eight platforms, two RNA-Seq and six microarray platforms. Data from different platforms were analyzed separately with stringent quality control and a consistent pre-processing pipeline (see Supplementary Material). *ComBat* was used to correct the batch effect within and among datasets (Johnson *et al.*, 2007; Leek *et al.*, 2012).

The database was designed using relational tables and was implemented by Microsoft's SQL Server Management Studio. Its website was built using C# on Windows Server.

3 Results

BrainEXP consists of two major parts: spatiotemporal expression variations and WGCNA co-expression networks described as follows (also see Fig. 1):

3.1 Spatiotemporal expression variations

BrainEXP provides three types of charts: The Differential expression analysis, the *Gene expression in different brain regions and sexes* shown by boxplot and the *Gene expression in different brain regions and ages* demonstrated by scatterplot.

3.2 WGCNA co-expression networks

We applied WGCNA to get the matrix of gene co-expression values. The results are displayed through four different modes. The first is the *Co-expression gene network*. Gene nodes connect if the corresponding genes significantly co-expressed across samples at each platform. We choose top 10 co-expressed genes ordered by the adjacency (the higher the absolute adjacency value, the stronger the connection will be) (see Supplementary Material). Second, the *Co-expression pattern* shows the other gene expression levels with the similar pattern. Next, the *Correlation of co-expression genes* gives

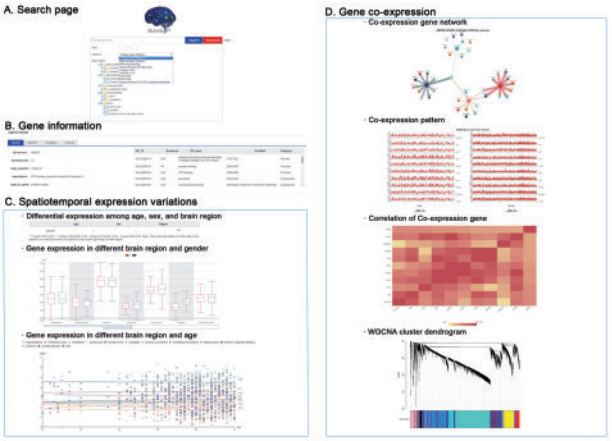

Fig. 1. The query and search results page. (**A**) The Search page. (**B**) The gene information interface. (**C**) The Spatiotemporal expression variations interface. (**D**) The Gene co-expression results interface

the correlation coefficients of these 11 genes (top 10 genes plus the query) displayed by the heat map. Finally, the *WGCNA cluster dendrogram* will tell you in which module the query gene is detected and how many genes are in the same module.

3.3 The advantages of combined datasets

The increase of sample size can improve the statistical power, overcome the influence of outliers and represent population better. We test whether a larger dataset can produce more information in the differential expression analysis and co-expression analysis. In differential expression analysis, since different datasets have different age ranges and brain regions, which may affect the results, we analyzed sex-related genes only. The test dataset chosen from our database has the same brain region (Hippocampus), platform (HG-U133P) and age range (22∼68 years old). This combined dataset contains 30 (12 and 18) samples. The results showed that 21 sex-related genes were detected in 30-sample dataset, while zero was detected in either 12- or 18-sample datasets. So, the combined dataset identified more sex-related genes.

In the WGCNA analysis, larger sample data resulted in more robust and refined results (Langfelder and Horvath, 2008). In summary, the combined data are proven valuable in detecting differential expression and co-expression.

3.4 Future update

The data will be updated as new human brain expression datasets are available. The datasets involving different cell types will be included in the future. New functions such as the gene co-expression in particular brain regions, age stages and sex will be added.

Acknowledgements

We sincerely thank Chicago Biomedical Consortium for its supports (to C. Liu). All the data contributors are sincerely appreciated for data submitted in the GEO and other databases. We are grateful to Qingres | Journal of Psychiatry and Brain Science for hosting our webserver.

Funding

This work was supported by National Natural Science Foundation of China grants 81401114, 31571312, the National Key Plan for Scientific Research and Development of China (2016YFC1306000), Innovation-Driven Project

of Central South University (No. 2015CXS034, 2018CX033) (to C. Chen) and NIH grants 1 U01 MH103340-01, 1R01ES024988 (to C. Liu).

Conflict of Interest: none declared.

References

Bakken,T.E. *et al*. (2016) A comprehensive transcriptional map of primate brain development. *Nature*, **535**, 367–375.

Barrett,T. *et al*. (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*., **41**, D991–D995.

Colantuoni,C. *et al*. (2011) Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*, **478**, 519–523.

Consortium,G.T. (2013) The genotype-tissue expression (GTEx) project. *Nat Genet*., **45**, 580–585.

Higgs,B.W. *et al*. (2006) An online database for brain disease research. *BMC Genomics*, **7**, 70.

Johnson,M.B. *et al*. (2009) Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron*, **62**, 494–509.

Johnson,W.E. *et al*. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Kang,H.J. *et al*. (2011) Spatio-temporal transcriptome of the human brain. *Nature*, **478**, 483–489.

Kolesnikov,N. *et al*. (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res*., **43**, D1113–D1116.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Leek,J.T. *et al*. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Miller,J.A. *et al*. (2014) Transcriptional landscape of the prenatal human brain. *Nature*, **508**, 199–206.

Psych,E.C. *et al*. (2015) The PsychENCODE project. *Nat Neurosci*, **18**, 1707–1712.

Tebbenkamp,A.T. *et al*. (2014) The developmental transcriptome of the human brain: implications for neurodevelopmental disorders. *Curr. Opin. Neurol*., **27**, 149–156.