Journal of Psychiatric Research 82 (2016) 23–29

Contents lists available at ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/psychires

A combined analysis of genome-wide expression profiling of bipolar disorder in human prefrontal cortex

Jinglu Wang [a, b], Susu Qu [a, b], Weixiao Wang [a, b], Liyuan Guo [a], Kunlin Zhang [a], Suhua Chang [a, *], Jing Wang [a, **]

[a] The Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 30 May 2016
Received in revised form
12 July 2016
Accepted 15 July 2016

Keywords:
Bipolar disorder
Gene expression
Prefrontal cortex
Mega-analysis
Differentially expressed genes

ABSTRACT

Numbers of gene expression profiling studies of bipolar disorder have been published. Besides different array chips and tissues, variety of the data processes in different cohorts aggravated the inconsistency of results of these genome-wide gene expression profiling studies. By searching the gene expression databases, we obtained six data sets for prefrontal cortex (PFC) of bipolar disorder with raw data and combinable platforms. We used standardized pre-processing and quality control procedures to analyze each data set separately and then combined them into a large gene expression matrix with 101 bipolar disorder subjects and 106 controls. A standard linear mixed-effects model was used to calculate the differentially expressed genes (DEGs). Multiple levels of sensitivity analyses and cross validation with genetic data were conducted. Functional and network analyses were carried out on basis of the DEGs. In the result, we identified 198 unique differentially expressed genes in the PFC of bipolar disorder and control. Among them, 115 DEGs were robust to at least three leave-one-out tests or different pre-processing methods; 51 DEGs were validated with genetic association signals. Pathway enrichment analysis showed these DEGs were related with regulation of neurological system, cell death and apoptosis, and several basic binding processes. Protein-protein interaction network further identified one key hub gene. We have contributed the most comprehensive integrated analysis of bipolar disorder expression profiling studies in PFC to date. The DEGs, especially those with multiple validations, may denote a common signature of bipolar disorder and contribute to the pathogenesis of disease.

1. Introduction

Bipolar disorder (BD) is an episodic recurrent pathological mood disturbance characterized by the cycles between bouts of mania and depression (Goodwin et al., 2007). Studies of families and twins showed bipolar disorder had estimated heritability ranging from 80% to 85% (Barnett and Smoller, 2009). Many studies have indicated that BD is a polygenic disease influenced by many genes with small effect (Baum et al., 2008). But, the specific pathogenesis of BD is still not well understood. Microarray technology provides a powerful tool for studying the gene dysfunction contributed to complex disorders (Bunney et al., 2003). Study of the differences in gene expression between patients and controls can not only help understand the relationship between genes and disease, but also provide evidences from the aspect of biological function for the existing genetic results. So far, there are numbers of gene expression profiling studies of BD being published (Chen et al., 2013, Clelland et al., 2013, de Baumont et al., 2015, Harris et al., 2008, Iwamoto et al., 2005, Iwamoto et al., 2004, Lanz et al., 2015, Matigian et al., 2007, Reinhart et al., 2015, Ryan et al., 2006, Yang et al., 2009). Besides different array chips and tissues, variety of the data processes in different cohorts aggravated the inconsistency of results of these genome-wide gene expression profiling studies.

Results of functional magnetic resonance imaging (fMRI) studies have demonstrated that BD is associated with multiple brain areas

* Corresponding author. Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China.
** Corresponding author. Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China.
E-mail addresses: changsh@psych.ac.cn (S. Chang), wangjing@psych.ac.cn (J. Wang).

http://dx.doi.org/10.1016/j.jpsychires.2016.07.016
0022-3956/

(Delvecchio et al., 2013, Hall et al., 2010, Passarotti et al., 2010, Pavuluri et al., 2008, Townsend and Altshuler, 2012, Yoshimura et al., 2014). Prefrontal cortex (PFC) is an important brain region to affect human cognition, thinking, perception and emotion, which are inseparable with mental illness. Recent years, many groups have attempted to identify changes of gene expression in the brains of BD, often focusing on the PFC (Lanz et al., 2015; Reinhart et al., 2015; Ryan et al., 2006). These studies have suggested several altered molecular processes and loci of BD, but the heterogeneity of their findings cannot be neglected. Generally, the sample sizes of these studies are usually small, in addition to the modest gene expression changes in brain tissue, which leads to a reduction in power when identifying differentially expressed genes (DEGs). Consequently, it is necessary to conduct an integrated analysis of the studies from the same area or tissue using a unified data process. After compiling the studies together, the sample size will be enlarged, which can enhance the credibility of the obtained DEGs. Furthermore, the loci identified from genetic studies are statistically meaningful. Integration of the genetic data and expression data could facilitate the understanding of the biological function of the genetic marker as Zhao et al. did in (Zhao et al., 2015). Hence, the integrative analysis for the genome-wide expression profile studies in PFC and the genetic data can be a good way to integrate the results of multiple levels (genetics, gene expression and brain function) to have a better explanation of the pathogenesis of BD.

In this study, we conducted a combined analysis for the genome-wide expression data sets of bipolar disorder in prefrontal cortex by integrating six data sets for the Brodmann area 9, 10, 11 and 46. After using consistent data pre-processing and quality control, the combined gene expression data included 101 cases and 106 controls. A mixed-effects model was used to control several factors, such as age, brain pH, post-mortem interval (PMI), gender and batch effect. Totally, we identified 198 differential expressed genes. Further validations were conducted by using leave-one-out test, different pre-processing methods, comparison with the results from the original gene expression analysis, and cross-analysis with genetic data. Our results confirmed some previously reported expression changes in BD in addition to identifying potential novel genes related to BD. Furthermore, pathway enrichment analysis and protein-protein interaction for the DEGs were conducted to explore the function and interaction of the differentially expressed genes.

2. Methods and materials

2.1. Data pre-processing and quality control

We searched bipolar disorder related gene expression data from Gene Expression Omnibus (GEO) (Barrett et al., 2011), ArrayExpress (Parkinson et al., 2007), and the Stanley Medical Research Institute online genomics database (SMRIDB) (Higgs et al., 2006). In GEO and ArrayExpress, we searched with keyword *bipolar disorder*, and filtered the result by setting organism as *Homo sapiens*, array type as *Expression profiling by array* in GEO or *transcription profiling by array* in ArrayExpress. For the data sets of bipolar disorder in SMRIDB, some of them represented repeated runs of the same samples from the data sets in GEO or ArrayExpress, so we only selected three data sets after removing the duplicate individuals. Data sets with samples from prefrontal cortex (BA9, 10, 11 or 46) and consisted of single-channel intensity data generated from Affymetrix Human Genome U133A Array or U133 Plus 2.0 Array were included for analysis. Besides, the included data sets should be with available information on covariates including age, brain pH, post-mortem interval (PMI), gender and batch effect, and the raw data. After the selection of data sets, six separate data sets were included in our study. Each data set was composed of a cohort of healthy control subjects and a cohort of bipolar disorder subjects, as diagnosed and reported in their respective studies (Table 1).

In order to consistently handle all datasets and eliminate bias introduced by different algorithms used in the original studies, we reprocessed each data set individually using the same pre-processing method Robust Multi-array Average (RMA) since it has been shown to be a high performer on gold standard data sets (Bolstad et al., 2003, Irizarry et al., 2003). Outlier samples from each dataset were identified as those showing $r < 0.8$ with all the other samples by using an inter-sample correlation analysis for the vector of probe expression values of each sample by pair-wise Pearson correlations. For probe sets, only those on the HGU133A chip were used. Finally, two samples (one control and one BD patient) were removed, and the merged gene expression value matrix contained 207 samples and 22,277 probe sets. Batch information was obtained using the 'scan date' stored in the CEL files; chips run on different days were considered different batches (Mistry et al., 2013). Sample characteristics for the subjects were collected and are summarized in Table S1.

2.2. Statistical modeling

Gene expression value for each probe set was modeled using a standard linear mixed-effects model, in which, disease, age, brain pH, post-mortem interval (PMI) were used as fixed effects, while gender and batch effect were used as random effects. For each probe set, t-statistic for the disease effect was extracted from the model to show the expression is up-regulated or down-regulated. Statistical significance was calculated using the likelihood ratio test by comparing this model with the null model, which included all of the six factors in the original model except disease. The resulting *P*-values were further adjusted for multiple testing using the p.adjust function in R by setting method as Benjamini-Hochberg (FDR). The probe sets with FDR <0.1 were considered as significant result.

2.3. Validation for the differentially expressed genes

The significant probe sets were then annotated with gene symbols and EntrezIDs by using R package "annotate". To test the robustness of the findings, we used a jackknife procedure to do "leave-one-out" test, in which, one of the six data sets was removed at once and the remaining five were analyzed using the same DEGs analysis procedure, then compared the two results to find the overlapped results. Two other pre-processing methods gcRMA and MAS 5.0 were also used to evaluate the robustness of our significant probe sets. In addition, the DEGs were compared with the gene lists obtained from the original analysis result for each data set. Both GSE5388 and GSE5389 were published in Ryan et al. (Ryan et al., 2006), but there was no reported result for data set GSE53987. For the unpublished SMRI data sets, the gene lists were extracted from SMRI online genomics database.

Validations with the genetic data were conducted by three ways. One is directly check if the DEGs were reported to be associated with BD in at least one study in the BD genetic database BDgene (Chang et al., 2013). Secondly, we calculated gene-level *P*-value for the bipolar disorder genome-wide association study data from the Psychiatric Genomics Consortium (PGC) (Sklar et al., 2011) using HYST in KGG (Li et al., 2011) by setting the default parameters, then multiple correction for the gene *P*-value was calculated using the p.adjust function in R by setting method as Benjamini-Hochberg (FDR). Thirdly, we checked if the susceptible SNPs of BD and their LD-proxies regulated the expression of these DEGs by integrating