Progress in Neuropsychopharmacology & Biological Psychiatry 81 (2018) 50–54

Contents lists available at ScienceDirect

Progress in Neuropsychopharmacology & Biological Psychiatry

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

Integrating genome-wide association study and expression quantitative trait locus study identifies multiple genes and gene sets associated with schizophrenia

Yan Zhao[a,1], Awen He[a,1], Feng Zhu[b], Miao Ding[a], Jingcan Hao[a], Qianrui Fan[a], Ping Li[a], Li Liu[a], Yanan Du[a], Xiao Liang[a], Xiong Guo[a], Feng Zhang[a,*], Xiancang Ma[b,**]

[a] Key Laboratory of Trace Elements and Endemic Diseases of National Health and Family Planning Commission, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, PR China

[b] Department of Psychiatry, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, PR China

ABSTRACT

Schizophrenia is a serious mental disease with high heritability. To better understand the genetic basis of schizophrenia, we conducted a large scale integrative analysis of genome-wide association study (GWAS) and expression quantitative trait loci (eQTLs) data. GWAS summary data was derived from a published GWAS of schizophrenia, containing 9394 schizophrenia patients and 12,462 healthy controls. The eQTLs dataset was obtained from an eQTLs meta-analysis of 5311 subjects, containing 923,021 cis-eQTLs for 14,329 genes and 4732 trans-eQTLs for 2612 genes. Genome-wide single gene expression association analysis was conducted by SMR software. The SMR analysis results were further subjected to gene set enrichment analysis (GSEA) to identify schizophrenia associated gene sets. SMR detected 49 genes significantly associated with schizophrenia. The top five significant genes were CRELD2 (p value = 1.65×10^{-11}), DIP2B (p value = 3.94×10^{-11}), ZDHHC18 (p value = 1.52×10^{-10}) and ZDHHC5 (p value = 7.45×10^{-10}), C11ORF75 (p value = 3.70×10^{-9}). GSEA identified 80 gene sets with p values < 0.01. The top five significant gene sets were COWLING_MYCN_TARGETS (p value < 0.001) and CHR16P11 (p value < 0.001), ACTACCT_MIR196A_MIR196B (p value = 0.002), CELLULAR_COMPONENT_DISASSEMBLY (p value = 0.002) and GRAESSMANN_RESPONSE_TO_MC_AND_DOXORUBICIN_DN (p value = 0.002). Our results provide useful information for revealing the genetic basis of schizophrenia.

1. Introduction

Schizophrenia is a complex mental disease characterized by heterogeneous display of positive symptoms, negative symptoms, and cognitive dysfunction. Schizophrenia generally occurs in late adolescence and early adulthood, and affects ~1% of the population worldwide (Owen et al., 2005). Schizophrenia can be a disabling disorder, which often results in functional decline, difficulties navigating the social field and low quality of life. The etiopathogenesis of schizophrenia remains unclear now.

It is well known that genetic factors play an important role in the pathogenesis of schizophrenia. Twin studies showed that the heritability of schizophrenia is approximately 80% (Sullivan et al., 2003). Extensive efforts have been paid to uncover the genetic basis of schizophrenia and multiple susceptibility genes have been identified. For instance, the largest GWAS to date discovered 108 genome-wide significant loci associated with schizophrenia (Ripke et al., 2014). Another study found that common variants in MKL1 gene are involved in the development of schizophrenia (Luo et al., 2015). Significant association between the haplotypes of COMT gene and schizophrenia has been reported (Shifman et al., 2002). Nevertheless, the genetic risks of schizophrenia explained by the reported loci are limited (Zhu et al., 2016b), suggesting the existence of undiscovered susceptibility genes.

Genome-wide association studies (GWAS) are wildly used in susceptibility gene mapping of complex diseases and traits. However, due to the strict statistical significant threshold, GWAS are likely to miss

* Correspondence to: Feng Zhang, Key Laboratory of Trace Elements and Endemic Diseases of National Health and Family Planning Commission, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an 710061, PR China.

** Correspondence to: Xiancang Ma, Department of Psychiatry, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, PR China.

E-mail addresses: fzhxjtu@mail.xjtu.edu.cn (F. Zhang), maxiancang@xjtu.edu.cn (X. Ma).

[1] The two authors contributed equally to this work.

http://dx.doi.org/10.1016/j.pnpbp.2017.10.003
Received 26 July 2017; Received in revised form 25 September 2017; Accepted 4 October 2017
Available online 09 October 2017
0278-5846/

susceptibility gene with moderate or weak genetic effects (Wang et al., 2010). Furthermore, molecular networks and cellular pathways are generally involved in the development of complex diseases (Schadt, 2009). The significant genetic loci identified by GWAS are usually functionally independent, and provide limited information for revealing the mechanism of complex diseases. Inspired by gene set enrichment analysis of gene expression profiles, researchers proposed GWAS-based GSEA (Wang et al., 2007). By integrating GWAS summary data and prior functional information of biological pathways, GWAS-based GSEA is capable of identifying diseases associated biological pathways. PAS has been successfully applied in the genetic studies of multiple complex diseases (Deng, 2010; Zhang et al., 2010b).

Expression quantitative traits loci (eQTLs) are genetic loci, the polymorphisms of which are associated transcript abundances. A lot of eQTLs have been discovered by now (Hernandez et al., 2012; Koopmann et al., 2014). Recently, summary data–based Mendelian randomization (SMR) approach was proposed (Zhu et al., 2016b), Through integrating summary-level data from GWAS and eQTLs studies, SMR is capable of identifying the genes, the expression levels of which are associated with complex diseases due to pleiotropy. SMR showed good performance in identifying new candidate genes for complex diseases (Zhu et al., 2016b).

In this study, we conducted a genome-wide SMR single gene expression association analysis followed by GWAS-based GSEA for schizophrenia. SMR was first applied to a large scale GWAS summary data of schizophrenia for screening genes associated with schizophrenia. Utilizing SMR single gene analysis results, GSEA was then carried out to identify the gene sets associated with schizophrenia.

2. Materials and methods

2.1. GWAS summary dataset of schizophrenia

A large-scale GWAS summary dataset of schizophrenia was applied here (Consortium, 2011). Briefly, this GWAS dataset contained 21,856 individuals (including 9394 cases and 12,462 controls) from 17 separate GWAS of European ancestry. Genotyping was conducted using commercial platforms, such as Affymetrix 5.0 array, Affymetrix 6.0 array and Illumina 550 K array. Genotype imputation was conducted by Beagle 3.0.4 using CEU + TSI HapMap3 data as reference panel (Altshuler et al., 2010). Genome-wide association analysis was performed using standard logistic regression. Detailed descriptions of study subjects, genotyping, imputation, and quality control can be found in the published study (Consortium, 2011).

2.2. Genome-wide eQTLs dataset

A large scale eQTLs dataset established by Westra et al. was applied here (Westra et al., 2013). In brief, this eQTLs dataset was derived from a meta-analysis of 5311 samples from peripheral blood. Gene expression levels were evaluated using Illumina gene expression arrays. Genotype imputation was conducted using HapMap2 data as reference panel. The identified eQTLs were further validated in 5 independent datasets. 923,021 cis-eQTLs and 4732 trans-eQTLs were identified at false discovery rate < 0.05.

2.3. SMR single gene analysis

The GWAS summary data of schizophrenia and eQTL data were analyzed by SMR software to identify the genes, the expression levels of which were associated with the risks of schizophrenia. SMR resembled a Mendelian randomization (MR) analysis, in which genetic variants were used as instrumental variables to evaluate the effects of gene expression on the variations of disease phenotypes. SMR analysis showed good power to evaluate the impact of gene expression variation on complex diseases by integrating GWAS summary data and eQTLs annotation information (Zhu et al., 2016a). SMR executable program for Linux system were downloaded from SMR website (http://cnsgenomics.com/software/smr/index.html). SMR analysis was carried out using the default parameters recommended by the developers. In this study, 2776 genes with both GWAS summary data and eQTLs data were analyzed by SMR. A p value was calculated for each gene by SMR. The genes with SMR p values $< 1.80 \times 10^{-5}$ (0.05/2776) after Bonferroni correction were identified as significant genes.

2.4. Gene set enrichment analysis

The single gene analysis results of SMR were further subjected to GSEA to identify the biological gene sets enriched in the significant genes identified by SMR (Wang et al., 2007). This approach has been widely used for identifying diseases related biological pathways (Liu et al., 2010; Zhang et al., 2010a). The biological gene sets analyzed by this study were driven from the GSEA Molecular Signatures Database (Subramanian et al., 2005). The latest gene set annotation database (msigdb.v5.1, contained 9623 gene sets) was downloaded from the GSEA website (http://software.broadinstitute.org/gsea/msigdb/index.jsp). 5000 permutations were conducted to calculate the empirical p value of each gene set.

3. Results

For SMR single gene expression association analysis, a total of 2776 genes were analyzed in this study. After strict Bonferroni correction, SMR detected 49 genes significantly associated with schizophrenia (Table 1). The top five significant genes were CRELD2 (p value $= 1.65 \times 10^{-11}$), DIP2B (p value $= 3.94 \times 10^{-11}$), ZDHHC18 (p value $= 1.52 \times 10^{-10}$), and ZDHHC5 (p value $= 7.45 \times 10^{-10}$), C11ORF75 (p value $= 3.70 \times 10^{-9}$). The single gene analysis results of SMR were further subjected to GSEA to identify the biological gene sets enriched in the significant genes identified by SMR. GSEA detected 80 gene sets with p values < 0.01 (Supplementary Table S1). The top five significant gene sets were COWLING_MYCN_TARGETS (p value $<$ 0.001) and CHR16P11 (p value < 0.001), ACTACC-T_MIR196A_MIR196B (p value $= 0.002$), CELLULAR_COMPONENT_DISASSEMBLY (p value $= 0.002$) and GRAESSMANN_RESPONSE_TO_MC_AND_DOXORUBICIN_DN (p value $= 0.002$). Table 2 summarizes the 42 gene sets with p value < 0.005.

4. Discussion

GWAS are widely used for susceptibility gene mapping of complex diseases. However, it is common that the susceptibility loci identified by GWAS individually have minor effects, and even cumulatively explain only a modest fraction of the genetic predisposition (Harrison, 2015). A large part of identified genetic loci lies within non-coding sequences, which also complicates their functional evaluation (Maurano et al., 2012). To better illuminate the genetic basis of schizophrenia and take full advantage of published GWAS data, we conducted a large scale integrative analysis of GWAS and eQTLs datasets for schizophrenia. We identified a group of schizophrenia associated genes and gene sets, providing useful information for revealing the biological mechanism of schizophrenia.

The most significant gene identified by this study is CRELD2, which has been reported to be associated with schizophrenia (De et al., 2012). CRELD2 is an endoplasmic reticulum(ER) stress-inducible gene (Oh-Hashi et al., 2009). ER is essential in synthesizing, folding, and assembling proteins in eukaryotic cells. Many pathophysiological conditions can results in the accumulation of unfolded proteins in ER, which will trigger stress responses and disrupt ER function (Kim et al., 2008). Previous studies have found that ER plays important roles in neurodegenerative diseases, such as Parkinson's disease and Alzheimer's disease (Lindholm et al., 2006). Our study results support that CRELD2