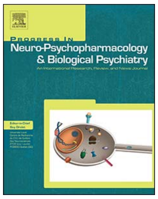
Contents lists available at ScienceDirect

Progress in Neuropsychopharmacology & Biological Psychiatry

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



Integrating genome-wide association study and expression quantitative trait loci data identifies multiple genes and gene set associated with neuroticism

Qianrui Fan[1], Wenyu Wang[1], Jingcan Hao, Awen He, Yan Wen, Xiong Guo, Cuiyan Wu, Yujie Ning, Xi Wang, Sen Wang, Feng Zhang*

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

ARTICLE INFO

Keywords:
Neuroticism
Genome-wide association study
Expression quantitative trait loci
Gene set enrichment analysis

ABSTRACT

Neuroticism is a fundamental personality trait with significant genetic determinant. To identify novel susceptibility genes for neuroticism, we conducted an integrative analysis of genomic and transcriptomic data of genome wide association study (GWAS) and expression quantitative trait locus (eQTL) study. GWAS summary data was driven from published studies of neuroticism, totally involving 170,906 subjects. eQTL dataset containing 927,753 eQTLs were obtained from an eQTL meta-analysis of 5311 samples. Integrative analysis of GWAS and eQTL data was conducted by summary data–based Mendelian randomization (SMR) analysis software. To identify neuroticism associated gene sets, the SMR analysis results were further subjected to gene set enrichment analysis (GSEA). The gene set annotation dataset (containing 13,311 annotated gene sets) of GSEA Molecular Signatures Database was used. SMR single gene analysis identified 6 significant genes for neuroticism, including *MSRA* (p value = 2.27×10^{-10}), *MGC57346* (p value = 6.92×10^{-7}), *BLK* (p value = 1.01×10^{-6}), *XKR6* (p value = 1.11×10^{-6}), *C17ORF69* (p value = 1.12×10^{-6}) and *KIAA1267* (p value = 4.00×10^{-6}). Gene set enrichment analysis observed significant association for Chr8p23 gene set (false discovery rate = 0.033). Our results provide novel clues for the genetic mechanism studies of neuroticism.

1. Introduction

Health is not only the absence of infirmity and diseases, but also a state of completely physical, mental and social well-being (Constitution of the World Health Organization, 1946). As one of the five important traits of personality, neuroticism is characterized as being vulnerable to negative emotions, like anxiety and fear (Ormel et al., 2013). Besides its relation to crisis events, neuroticism also impairs physical health (Ormel et al., 2013), such as reversible heart failure (Christensen et al., 2016), sleep disorder (Huang et al., 2016) and depression (Enns and Cox, 1997). Neuroticism accounts for a substantial proportion of current and lifetime comorbidity (Ormel et al., 2013).

It was estimated that genetic factors contributed > 30% of the variations of neuroticism (Nivard et al., 2015). Multiple studies have been conducted to uncover the genetic basis of neuroticism and several susceptibility genes have been identified. For instance, Criado et al. (2014) found that CHRNA3 and CHRNA5 genes were involved in the development of neuroticism in Mexican American young adults. Another cohort-based study observed that BDNF gene interacted with recent stressful life events may lead to neuroticism (Lehto et al., 2016). Okbay et al. (2016) reported 11 neuroticism-associated genetic loci by large GWAS (n = 170,911). Most recently, variants on chromosome 8p23.1 and in L3MBTL2 were detected to be significantly associated with neuroticism (Lo et al., 2016). However, the genetic risk of neuroticism explained by the reported candidate loci was limited (Zhu et al., 2016), suggesting the existence of undiscovered susceptibility genes for neuroticism.

In spite of its great power, GWAS generally focuses on the most significant genetic loci, which are limited and mostly functionally independent. Although individual genes can participate in multiple cellular processes, identifying several disease-associated genes are usually inadequate for revealing the pathogenesis of complex diseases. It has been demonstrated that GWAS has limited power to detect the causal loci with moderate or weak genetic effects due to the strict statistical significant threshold (Marchetti-Bowick et al., 2016). Inspired by the gene set enrichment analysis of microarray data, pathway-

* Corresponding author at: Key Laboratory of Trace Elements and Endemic Diseases of National Health and Family Planning Commission, School of Public Health, Health Science Center, Xi'an Jiaotong University, Yan Ta West Road 76, Xi'an 710061, PR China.

E-mail address: fzhxjtu@mail.xjtu.edu.cn (F. Zhang).

[1] These two authors contribute equally to this manuscript.

http://dx.doi.org/10.1016/j.pnpbp.2017.05.017
Received 9 October 2016; Received in revised form 17 May 2017; Accepted 19 May 2017
Available online 25 May 2017

based association study (PAS) using GWAS summary data was proposed (Wang et al., 2007), and successfully applied in the genetic studies of complex diseases (Liu et al., 2010; Zhang et al., 2010). By integrating the results of GWAS and prior functional information of biological pathways, PAS has the potential to provide additional clues for pathogenetic studies of complex diseases.

Expression quantitative trait loci (eQTLs) are genomic loci which can regulate gene expression level. Through genome-wide detecting associations between gene transcript abundance and genomic polymorphisms, a lot of eQTLs have been identified in human genome (Fu et al., 2012; Hernandez et al., 2012; Koopmann et al., 2014; Zou et al., 2012). Recently, summary data–based Mendelian randomization (SMR) analysis was proposed. By using the statistic model of Mendelian Randomization, SMR is able to identify the genes whose expression levels are associated with a complex trait (Zhu et al., 2016). Using published summary data of GWAS and eQTL study, SMR identified a group of novel genes associated with human complex diseases, showing good performance for susceptibility gene mapping (Zhu et al., 2016). SMR provides a novel tool to prioritize genes underlying GWAS hits for follow-up functional studies. However, to the best of our knowledge, SMR cannot be directly applied for genome-wide pathway association analysis.

In this study, we conducted a genome-wide single gene association analysis and gene sets enrichment analysis, integrating GWAS and eQTL study data. SMR was first applied to a large scale GWAS summary data of neuroticism for screening novel genes, the expression levels of which were associated with neuroticism. Furthermore, the SMR single gene analysis results were subjected to PAS to identify neuroticism associated gene sets. The novelty of this research is combining SMR with biological pathway enrichment analysis, which may help to detect novel susceptibility gene sets for neuroticism.

2. Material and methods

2.1. GWAS summary datasets

A recent large-scale genome-wide meta-analysis of neuroticism was used here (Okbay et al., 2016). Briefly, genome-wide summary data was collected from two GWAS of neuroticism, including a new GWAS in UK Biobank (UKB) cohort (n = 107,245) and a published GWAS meta-analysis of 30 cohorts conducted by the Genetics of Personality Consortium (GPC, n = 63,661) (de Moor et al., 2015; Sudlow et al., 2015). Neuroticism was diagnosed according to the respondent's score on a 12-item version of the Eysenck Personality Inventory Neuroticism (Eysenck and Eysenck, 1975). Genotyping was conducted using commercial platforms, such as Affymetrix Axiom array, Affymetrix 6.0 array and Illumina 550 K array. Genotype imputation was conducted by IMPUTE, Beagle or MACH software, using various reference populations, including 1000 Genomes Project, UK10K haplotype reference panel and HapMap reference panel (CEU + TSI). A sample-size-weighted fixed-effects meta-analysis of the GWAS summaries of UKB and GPC data was conducted (Okbay et al., 2016). Detailed information of cohorts, genotyping, imputation, meta-analysis and quality control approaches can be found in the published studies (de Moor et al., 2015; Okbay et al., 2016; Sudlow et al., 2015).

2.2. SMR single gene analysis

The genome-wide meta-analysis statistics of neuroticism were input into SMR for detecting association between gene expression levels and neuroticism. SMR executable files (version 0.66) for Linux system were downloaded from SMR website (http://cnsgenomics.com/software/smr/index.html). The input GWAS summary data and eQTL data files were prepared according users' document of SMR. SMR analysis was conducted using the default parameters recommended by SMR developers. The principle of SMR-based single gene analysis resembles a Mendelian randomization that regards genetic variants as instrumental variable to evaluate the effect of gene expression on traits (Zhu et al., 2016). SMR pooled both GWAS summary data and eQTL information together to evaluate the causal effects of gene expression variation on target diseases. The principle of SMR analysis is to use a genetic variant as an instrumental variable to estimate and test for the causative effect of an exposure variable. From a Mendelian randomization (MR) analysis perspective, if we denote z as a genetic variant (for example, a SNP), x as the expression level of a gene and y as the target trait, then the two-step least-squares estimate of the effect of x on y from a MR analysis is $\widehat{b_{xy}} = \widehat{b_{zy}}/\widehat{b_{zx}}$, where $\widehat{b_{zy}}$ and $\widehat{b_{zx}}$ are the least-squares estimates of y and x on z, respectively. $\widehat{b_{zy}}$ represents the effect of the genetic variant on the target trait and $\widehat{b_{zx}}$ represents the effect of the genetic variant on gene expression level. However, such estimate requires genotype, gene expression and phenotype to be measured on the same sample. These data are usually unavailable in practice. In SMR, we can use the summary statistics from GWAS and eQTL studies to estimate the effect of a SNP on gene expression ($\widehat{b_{zx}}$) and disease phenotype ($\widehat{b_{zy}}$), respectively. $\widehat{b_{zx}}$ is derived from the eQTL studies, define by $\widehat{b_{zx}} = z_{zx}S_{zx}$, where $\mathrm{S_{zx}} = 1/\sqrt{2\mathrm{p}(1 - \mathrm{p})(\mathrm{n} + \mathrm{z}_{\mathrm{zx}}^2)}$. z_{zx} is the z statistic in eQTL studies. p is the allele frequency and n is the sample size. $\widehat{b_{zy}}$ can be collected from GWAS summary data. Within each gene, the most significant eQTL SNP from the eQTL study and the same SNP from the GWAS are used by SMR to detect association between the gene and target phenotype. A more detailed explanation of SMR algorithms was presented in Supplementary note. Further description of SMR approach can be found in the published paper (Zhu et al., 2016).

Preferred for the two-stage design and large sample sizes, the eQTL dataset established by Westra et al. (2013) was applied here. Briefly, this eQTL dataset was first driven from a meta-analysis of 5311 samples from peripheral blood. Gene expression levels were assessed by Illumina gene expression arrays. SNP genotype data was imputed against HapMap 2 reference panels. The identified eQTLs were further validated in another independent sample of 2775 individuals which were got from 5 independent datasets of 4 cohorts, including data obtained from lymphoblastoid cells (HapMap 3, n = 608), B cells and monocytes (Oxford, n = 282 and 283, respectively) and whole peripheral blood (KORA F4, n = 740 and BSGS, n = 862). 923,021 cis-eQTL for 14,329 gene expression probes and 4732 trans-eQTL for 2612 gene expression probes were identified at false discovery rate (FDR) < 0.05. For this study, 4674 genes with both GWAS summary and eQTL data were analyzed. A p value was calculated by SMR for each gene. Significant genes were identified at SMR p values $< 1.07 \times 10^{-5}$ (0.05 / 4674) after Bonferroni correction.

2.3. Gene set enrichment analysis

The SMR gene-level results of neuroticism were further analyzed using the gene set enrichment analysis (GSEA) approach developed by Wang et al. (2007). GSEA was first proposed by Subramanian et al. (2005) for gene expression profile analysis. GSEA algorithm was modified by Wang et al. for GWAS-based gene set association analysis (Wang et al., 2007). GSEA is a competitive method that tests whether a gene set is associated with the target trait by comparing the genetic effects of genes in the set with genetic effects of genes not in the set. In this study, the analyzing gene sets were defined according to the GSEA Molecular Signatures Database (Subramanian et al., 2005). The latest gene set annotation database (msigdb.v5.1) was downloaded from the GSEA Molecular Signatures Database website (http://software.broadinstitute.org/gsea/msigdb/index.jsp). It contained a total of 13,311 annotated gene sets. During the gene set association analysis, 5000 permutations were conducted to calculate the p value and FDR of each gene set (Wang et al., 2007). Significant gene sets were identified at FDR < 0.05.