Gene 525 (2013) 107–115

Contents lists available at SciVerse ScienceDirect

Gene

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

# Pathway analysis of a genome-wide association study in schizophrenia

Young Ho Lee *, Jae-Hoon Kim, Gwan Gyu Song

*Division of Rheumatology, Department of Internal Medicine, Korea University College of Medicine, Seoul, Republic of Korea*

ARTICLE INFO

*Article history:*
Accepted 1 April 2013
Available online 1 May 2013

*Keywords:*
Schizophrenia
Genome-wide association study
Pathway-based analysis

ABSTRACT

*Objective:* The aim of this study was to identify the candidate single nucleotide polymorphisms (SNPs) and candidate mechanisms that contribute to schizophrenia susceptibility and to generate a SNP to gene to pathway hypothesis using an analytical pathway-based approach.
*Methods:* We used schizophrenia GWAS data of the genotypes of 660,259 SNPs in 1378 controls and 1351 cases of European descent after quality control filtering. ICSNPathway (Identify candidate Causal SNPs and Pathways) analysis was applied to the schizophrenia GWAS dataset. The first stage involved the pre-selection of candidate SNPs by linkage disequilibrium analysis and the functional SNP annotation of the most significant SNPs found. The second stage involved the annotation of biological mechanisms for the pre-selected candidate SNPs using improved-gene set enrichment analysis.
*Results:* ICSNPathway analysis identified fifteen candidate SNPs, ten candidate pathways, and nine hypothetical biological mechanisms. The most strongly associated potential pathways were as follows. First, rs1644731 and rs1644730 to RDH8 to estrogen biosynthetic process ($p < 0.001$, FDR $< 0.001$). The genes involved in this pathway are RDH8 and HSD3B1 ($p < 0.05$). All-trans-retinol dehydrogenase (RDH8) is a visual cycle enzyme that reduces all-trans-retinal to all-trans-retinol in the presence of NADPH. The chemical reactions and pathways involved result in the formation of estrogens, which are C18 steroid hormones that can stimulate the development of female sexual characteristics. Second, rs1146031 to ACVR1 to mesoderm formation and activin binding ($p < 0.001$, FDR $= 0.032, 0.034$). Two of 15 candidate genes are known genes associated with schizophrenia: KCNQ2 and APOL2. One of the 10 candidate pathways, estrogen biosynthetic process, is known to be associated with schizophrenia ($p < 0.001$, FDR $< 0.001$). However, 13 of candidate genes (*RDH8*, *ACVR1*, *PSMD9*, *KCNAB1*, *SLC17A3*, *ARCN1*, *COG7*, *STAB2*, *LRPAP1*, *STAB1*, *CXCL16*, *COL4A4*, *EXOSC3*) and 9 of candidate pathways were novel.
*Conclusion:* By applying ICSNPathway analysis to schizophrenia GWAS data, we identified candidate SNPs, genes like KCNQ2 and APOL2 and pathways involving the estrogen biosynthetic process may contribute to schizophrenia susceptibility. Further analyses are needed to validate the results of this analysis.


## 1. Introduction

Schizophrenia is a mental disorder characterized by a breakdown of thought processes and poor emotional responsiveness. It most commonly manifests itself as auditory hallucinations, paranoid or bizarre delusions, or disorganized speech and thinking, and it is accompanied by significant social and occupational dysfunction. The onset of symptoms typically occurs in young adulthood, and the disease has a global lifetime prevalence of about 0.3–0.7% (van Os and Kapur, 2009). Although the etiology of schizophrenia is not fully understood, it has been suggested to involve suitable interactions between a susceptible genetic background and environmental factors. Many molecular candidates have been proposed to be involved, including specific copy number variations, NOTCH4, and histone protein loci (Tiwari et al., 2010).

Genome-wide association studies (GWAS) provide a powerful means of screening susceptibility genes of complex diseases (Harley et al., 2008). The number of GWAS conducted is growing rapidly and this growth has resulted in the discovery and replication of new disease genes (Manolio, 2010). However, although large-scale GWAS have been carried out on schizophrenia, and on other complex diseases, much of the genetic component of schizophrenia remains unexplained.

Research based on GWAS datasets offers powerful opportunities (Hom et al., 2008; Johnson and O'Donnell, 2009). It is well known that genes do not work in isolation. Instead, complex molecular networks and cellular pathways are often involved in disease susceptibility (Schadt, 2009). Individual genetic variants make small risk contributions but may interact with each other to cause diseases, such as, schizophrenia. However, genetic signals have only been examined at the single marker level in GWAS and suggested biological mechanisms lack support (Manolio, 2010). The key challenges posed during the GWAS data interpretation are the identification of single nucleotide polymorphisms (SNPs) and the provision of evidence and of hypotheses

* Corresponding author at: Division of Rheumatology, Department of Internal Medicine, Korea University Anam Hospital, Korea University College of Medicine, 126-1, Anam-dong 5-ga, Seongbuk-gu, Seoul 136-705, Korea. Tel.: +82 2 920 5645; fax: +82 2 922 5974.
*E-mail address:* lyhcgh@korea.ac.kr (Y.H. Lee).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.04.014

regarding the mechanism by which they affect traits. Furthermore, using new methods to study existing GWAS data may provide additional biological insight and highlight new candidate genes (Wang et al., 2010). Pathway-based approaches have been developed and improved, and one of these pathway-based approaches, the ICSNPathway (Identify candidate Causal SNPs and Pathways) approach was developed to identify biologically plausible candidate SNPs and their corresponding candidate pathways from GWAS data by integrating linkage disequilibrium (LD) analysis, functional SNP annotation, and pathway-based analysis (PBA) (Zhang et al., 2011). Thus, the application of ICSNPathway analysis may improve GWAS data interpretations from variants to biological mechanisms because it identifies candidate SNPs and their corresponding candidate pathways.

In the present study, we applied ICSNPathway analysis to a schizophrenia GWAS dataset to identify biologically plausible candidate SNPs and mechanisms that contribute to schizophrenia susceptibility, and to generate SNP to gene to pathway hypotheses.

## 2. Methods

### 2.1. Study populations

We used a schizophrenia GWAS dataset extracted from NCBI dbGap (http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap) (phs000021.v3.p2) (O'Donovan et al., 2008). Genome-wide genotyping data has been produced with the Affymetrix 6.0 platform and with the Birdseed calling algorithm. This analysis includes only the subset of subjects including 1378 controls and 1351 cases with European ancestry. This consent group includes a subset of schizophrenia cases and all controls (which overlap with Bipolar study controls in Bipolar and this consent group includes a subset of schizophrenia cases. Inclusion criteria were as follows: 1) All subjects must give signed, informed consent, 2) probands must have a consensus best-estimate DSM-IV (Diagnostic and Statistical Manual of Mental Disorders) diagnosis of schizophrenia or of schizoaffective disorder with at least six months' duration of the "A" criteria for schizophrenia, 3) subjects must be over 18 years of age at interview, male or female, and 4) the informant should have known the subject for at least two years, be familiar with the psychiatric history, and have at least 1 h of contact per week with the proband (close family members preferred). Exclusion criteria were as follows: 1) Unable to give informed consent to all aspects of the study, 2) unable to speak and be interviewed in English, 3) psychosis is deemed secondary to substance use by the consensus diagnostic procedure because psychotic symptoms are limited to periods of likely intoxication or withdrawal, or there are persistent symptoms which are likely to be related to substance use, 4) the psychotic disorder is deemed secondary to a neurological disorder such as epilepsy based on the nature and timing of symptoms. For example, non-specific, non-focal EEG abnormalities are common in schizophrenia, but subjects with psychosis that emerged in the context of temporal lobe epilepsy would be excluded, and 5) subjects with severe mental retardation (MR). Subjects with mild MR (IQ is greater than or equal to 55 or based on clinical and educational history) will be included, if schizophrenia symptoms and history can be clearly established. Data was filtered using criteria: 1) sample heterozygosity range from 0.26 to 0.285; 2) sample call rate greater than 97%; 3) SNP MAF >0.01; 4) SNP call rate >0.95, and HWE p-value >0.000001; 4) plate effect p-value test above $10^{-8}$ for a single plate and/or $10^{-4}$ for two plates; and 5) less than 2 discrepancies between SNP calls on duplicated samples. These thresholds resulted in 1378 controls and 1351 cases genotyped over 729,454 SNPs used in the association analysis. Genome-wide association scan has been pre-computed by NCBI. Genome wide association using a chi-square test was carried out using the "–assoc" option in the program PLINK (allelic association). p-Values and allelic odds ratios are reported. We filtered the dataset to remove individuals with a $p < 1 \times 10^{-4}$ for Hardy–Weinberg violation and a call rate of <98%, to reduce the impact of genotyping errors. The 660,259 SNPs passed quality control filters. No one was omitted from the analysis following the quality control procedure, thus in total 1378 controls and 1351 cases were included in this analysis.

### 2.2. Identification of candidate SNPs and pathways from the schizophrenia GWAS

ICSNPathway analysis was applied to the schizophrenia GWAS data (Zhang et al., 2011). ICSNPathway analysis involves two stages (Zhang et al., 2011). The first stage involves the pre-selection of candidate SNPs by LD analysis and functional SNP annotation based on most significant SNPs, whereas the second stage involves the annotation of biological mechanisms to pre-selected candidate SNPs using a PBA algorithm named *i*-GSEA (improved-gene set enrichment analysis). A full list of GWAS SNP p-values was entered into ICSNPathway analysis. One concept utilized in ICSNPathway analysis is LD analysis, which searches SNPs in LD with most significant SNPs of a GWAS dataset to identify more possible candidate SNPs based on the extended dataset, which includes HapMap data (International HapMap C et al., 2010). The other involves the use of functional SNPs. ICSNPathway analysis pre-selects candidate SNPs based on functional SNPs, which are important for understanding the underlying genetics of human health. Functional SNPs are defined as SNPs that may alter protein or gene expressions or the role of a protein in context of a pathway. Functional SNPs include deleterious and non-deleterious non-synonymous SNPs, SNPs leading the gain or loss of a stop codon, SNPs resulting in a frame shift, SNPs in essential splice sites, and SNPs in regulatory regions. The ICSNPathway server applies a PBA algorithm, called improved-gene set enrichment analysis (*i*-GSEA), to the full list of GWAS SNP p-values to detect pathways associated with traits. Briefly, (1) each SNP is mapped to its nearest gene according to the SNP and gene localization in the Ensembl 61 database (http://www.wnsembl.org/biomart/martview), and the maximum $t = -\log(\text{p-value})$ values of SNPs mapped to genes are assigned to represent those genes. Then all genes are ranked by decreasing their representation value *t*. (2) For each pathway *S*, *ES* (enrichment score, i.e. a Kolmogorov–Smirnov like running-sum statistics with weight [a]), which measures the tendency that genes of a pathway are located at the top of the ranked gene list, is calculated. (3) ES is converted to a *SPES* (significant proportion based ES) by multiplying it with $m_1/m_2$, where $m_1$ is the proportion of significant genes (defined as genes mapped with at least one of the top 5% most significant SNPs of all SNPs in the GWAS) for pathways *S*, and $m_2$ is the proportion of significant genes for all genes in the GWAS. (4) SNP label permutation and normalization are employed to generate the distribution of *SPES* and to correct gene variation (the bias caused by different genes with different numbers of mapped SNPs) and pathway variation (the bias due to different pathways consistent of different number of genes). (5) Based on all the distribution of *SPES* generated by permutation, a nominal p-value is calculated and a false discovery rate (FDR) is computed for multiple testing correction. By " most significant SNPs", we mean SNPs with p-values below a certain cutoff, which can be specify to extract the most significant SNPs from GWAS SNP p-values. ICSNPathway analysis identified significant pathways from the original GWAS dataset when input a cutoff p-value of $<5 \times 10^{-2}$, and thus, this cutoff was used throughout the study.

The first condition applied for ICSNPathway analysis was 'within gene,' which meant that only p-values of SNPs located within genes were utilized in the PBA algorithm. The second was a false discovery rate (FDR) cutoff (0.05) for multiple testing corrections. Control of the FDR is preferred for large-scale testing. The FDR, defined as the expected proportion of false positives among all significant tests, allows researchers to identify a set of "candidate positive", which a high proportion is likely to be true. The FDR, a permutation-based approach for multiple comparisons problem, was used for identification of statistically significant genes. 'Within gene' was selected as a