Genome-wide association study of genetic susceptibility and clinical features related to obsessive-compulsive disorder in Korean population

Chun Il Park

Department of Medicine

The Graduate School, Yonsei University

Genome-wide association study of genetic susceptibility and clinical features related to obsessive-compulsive disorder in Korean population

Chun Il Park

Department of Medicine

The Graduate School, Yonsei University

Genome-wide association study of genetic susceptibility and clinical features related to obsessive-compulsive disorder in Korean population

Directed by Professor Se Joo Kim

The Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Chun Il Park

June 2021

This certifies that the Doctoral Dissertation
of Chun Il Park is approved.

Thesis Supervisor: Se Joo Kim

Thesis Committee Member #1: Chan Hyung Kim

Thesis Committee Member #2: Kyung Sue Hong

Thesis Committee Member #3: Kyung A Lee

Thesis Committee Member #4: Sangwoo Kim

The Graduate School
Yonsei University

June 2021

ACKNOWLEDGEMENTS

My sincere thanks to everyone who supported me during the writing of this doctoral thesis.

Above all, I would like to express my deepest gratitude to my supervisor, Prof. Se Joo Kim, who patiently guided me through the research period. I was able to complete this dissertation because he supported me in all aspects of this research. Thank you for your support.

I would also like to thank Prof. Chan Hyung Kim, chair of this thesis committee, and all of the thesis committee members, Prof. Kyung Sue Hong, Prof. Kyung A Lee, and Prof. Sangwoo Kim. I learned a lot about research and attitude of a scientist from your valuable comments on the thesis.

I would like to thank Prof. Kee Namkoong for giving me important advice so that I could walk on the path I am now. I thank Prof. Jee In Kang for always supporting and guiding me. I want to be of help to you, but I always end up getting help from you. I would also like to thank Prof. Sang-Hyuk Lee, Prof. Tai Kiu Choi, Prof. Kang Soo Lee, and Prof. Minji Bang for their consideration and support in completing my thesis.

All of this would not have been possible without the support of my family. I am so grateful to my parents and sister who always believe in me. Finally, I would like to thank my beloved wife, Jeongjin, my lovely Sihwan, and my lovely Sia. I am sorry for not being able to spend much time with you and thank you for understanding everything.

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Genome-wide association study of genetic susceptibility and clinical features related to obsessive-compulsive disorder in Korean population

Chun Il Park

Department of Medicine

The Graduate School, Yonsei University

(Directed by Professor Se Joo Kim)

Purpose: Obsessive–compulsive disorder (OCD) is a severe and disabling clinical condition characterized by intrusive thoughts and repetitive behaviors. A growing body of studies has investigated the genetic etiology of OCD, and genome-wide association study (GWAS) has emerged as a powerful tool for investigating the genetic architecture of common diseases. However, only few studies have conducted GWAS on OCD among Asian populations. This study aimed to identify common variants associated with OCD using genome-wide analysis in the Korean population. In addition, further analyses, including polygenic risk score (PRS) analysis, were performed.

Methods: We collected two genome-wide variant datasets produced by the Korean Biobank Array Chip. Dataset #1 included 429 patients with OCD and 2400 healthy controls. In addition, dataset #2 of independent samples, including 115 patients with OCD and 228 healthy controls, was gathered for the replication study. We performed a logistic regression analysis for case–control associations. Further, gene-based analysis was performed using the web-based platforms FUMA and MAGMA. PRS analyses

were performed using PRSice-2 in two steps: (1) we used the previous GWAS results of the Psychiatric Genomics Consortium as a discovery set and dataset #1 as the target sample, and (2) we used the summary statistics from dataset #1 as a discovery set and calculated the PRS in dataset #2. Finally, the association between genetic predisposition and clinical characteristics and onset age of OCD was investigated.

Results: There were no significant genome-wide markers in the multiple logistic regression analysis. Further analysis with gene annotation revealed that upregulated gene expression was associated with genetic differences in OCD in the frontal areas of the brain. PRS was calculated using a discovery set from a previously reported GWAS in other populations, and the threshold of P = 0.4288 showed the best model fit. When using the GWAS results of the present study as a discovery set, the R2 value was 0.018 with a P-value threshold of 0.0249 in independent Korean samples. Regarding the genetic association with onset age of OCD, rs372803 in the *CIT* gene showed genome-wide significance, although it was not replicated in independent samples.

Conclusions: Although we could not identify the genome-wide significant loci associated with OCD susceptibility, we found a genetic effect on the onset age of OCD. Further studies with larger sample sizes are needed to explore the candidate SNPs and related genes.

Key words: obsessive–compulsive disorder, genome-wide association study, onset age, polygenic risk score

Genome-wide association study of genetic susceptibility and clinical features related to obsessive-compulsive disorder in Korean population

Chun Il Park

Department of Medicine

The Graduate School, Yonsei University

(Directed by Professor Se Joo Kim)

Ⅰ. INTRODUCTION

An obsession is a recurrent, unwanted and intrusive thought, feeling, urge, or image. An individual with an obsession attempts to suppress or neutralize the obsession to relieve anxiety. This attempt to suppress the obsession is called compulsion, which is a conscious, standardized, and recurrent behavior, such as counting, checking, or washing. Obsession and compulsion are the main symptoms and bases of obsessive–compulsive disorder (OCD). OCD is a severe and disabling clinical condition, and the lifetime prevalence is estimated at 2–3% in the general population [1]. Several studies suggested that susceptibility to OCD has a close genetic relationship. In a family study, the lifetime prevalence of OCD was higher in the first-degree relatives of OCD probands than in controls (10.3% vs. 1.9%, respectively) [2] and the onset age of obsessive–compulsive symptoms was related to familiarity in OCD patients [3]. Another family and twin study with large and multigenerational samples suggested that familial risk for OCD was largely attributable to additive genetic factors and that there was no significant effect of a shared environment [4].

Considering the evidence of genetic etiology in OCD, a large number of molecular genetic studies have been conducted to identify candidate genes that influence the development of OCD. The majority of candidate gene association studies have focused on genetic variants related to serotonin, dopamine, and glutamate pathways based on the current knowledge of OCD neurobiology. Dysregulation of neurotransmitter systems may alter the function of the cortico-striato-thalamo-cortical circuits [5]. Numerous candidate gene studies on OCD have been published. A meta-analysis study that reviewed 113 genetic association studies concluded that OCD was associated with serotonin-related polymorphisms (5-HTTLPR and HTR2A) and with polymorphisms involved in catecholamine modulation in males (COMT and MAOA) [6]. However, polymorphisms in two dopamine system-related genes (*DAT1* and *DRD3*) and in a glutamate-related gene (*SLC1A1*) showed no significant correlation with OCD development [6]. In addition, among other candidate genes, significant genetic causes of OCD were found in bradykinin (BDKRB2) [7], trophic factors (BDNF; only in the male subgroup) [8], immunologic factors (TNFA) [9], and glutamate-related gene (GRIK2) [10]. The findings of previous studies mentioned above suggest that multiple genes play a role in the development of OCD. However, because the candidate gene approach is based on a theoretical background, there are limitations in elucidating the genetic predisposition of OCD [11].

Genome-wide association study (GWAS) has emerged as a powerful tool for investigating the genetic architecture of common diseases [12] while overcoming the limitations of candidate gene studies. To date, only two OCD GWASs have been

reported [13,14]. The first GWAS [14] analyzed 469,410 autosomal and 9,657 X-chromosomal single nucleotide polymorphisms (SNPs) using the Illumina Human610-Quadv1_B SNP arrays among European, South African Afrikaner, and Ashkenazi Jewish ethnicities. In the case–control analysis, two SNPs that showed the lowest p-values were located in DLGAP1 ($p = 2.49 \times 10^{-6}$ − and $p = 3.44 \times 10^{-6}$ −), which is known to be a member of the neuronal postsynaptic density complex. The second GWAS [13] examined 549,123 autosomal markers in a United States population using Illumina HumanOmniExpress bead chips. In the study above, a marker on chromosome 9 near the gene encoding PTPRD protein showed the smallest p-value ($p = 4.13 \times 10^{-7}$). The presynaptic PTPRD promotes the differentiation of glutamatergic synapses [15,16] and plays a role in regulating GABAergic synapse development [17]. Although the results of previous GWASs on OCD did not reach genome-wide significance ($p < 5 \times 10^{-8}$), a possible gene was presented for further study to reveal the genetic susceptibility of OCD. However, studies that conduct GWASs in OCD among Asian populations are lacking. Considering the difference in allelic distribution according to ethnicity, it is valuable to explore the genetic architecture of OCD in the Korean population.

This study aimed to find OCD-related variants in the Korean population using the Korea Biobank Array. In addition, we investigated the association between genetic data and other phenotypic characteristics, including onset age of the OCD group. Further, we conducted additional analysis with significant markers to increase the explanatory power of the genetic effect using polygenic risk score (PRS).

Ⅱ. MATERIALS AND METHODS

1. Subjects

We collected two datasets of patients with OCD and healthy controls.

Dataset #1 comprised a total of 429 unrelated individuals with OCD from the outpatient department of psychiatry at Severance Hospital, Yonsei University Health System and 2400 healthy individuals from the National Biobank of Korea, the Center for Disease Control and Prevention, Republic of Korea. These control data were provided after quality control according to the institution's policy. The genotype data of healthy controls were approved for use in this study by the National Biobank of Korea. This dataset was used for genome-wide analysis between cases and controls, and PRS analysis.

For dataset #2, we recruited additional Korean subjects including 115 OCD patients and 228 healthy controls. Patients with OCD were recruited from the outpatient department of psychiatry at Severance Hospital, similar to dataset #1, and healthy controls were recruited using posters and online advertisements. The dataset #2 was used for replication study of GWAS results and as a target set for PRS analysis.

All patients with OCD were diagnosed using the Korean version of the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) Axis I disorders by a trained psychiatrist. The exclusion criteria were as follows: age < 19 or > 65 years, a lifetime history of psychotic symptoms, history of substance abuse or dependence in the preceding 6 months, or severe organic or neurologic disorders. In addition, subjects with a comorbid DSM-IV Axis I disorder other than OCD were excluded. Ethnicity was

ascertained through self-reports and the administrator's observations; only those subjects who were identified as ethnically Korean were enrolled.

2. Genotyping and quality control

Blood samples of patients with OCD (datasets #1 and #2) and healthy controls (dataset #2) were collected in 3 ml EDTA tubes and stored in a freezer until they were genotyped using the Korea Biobank Array kit. The Korea Biobank Array was designed by the Center for Genome Science at the Korea National Institute of Health (4845-301, 3000-3031). Samples that revealed the following thresholds were excluded: sex inconsistency, markers with a high missing rate (>5%), individuals with a high missing rate (>5%), a minor allele frequency < 0.01, and a significant deviation from the Hardy–Weinberg equilibrium ($p < 0.0001$). In addition, markers that were related to one another in linkage disequilibrium (LD $R^2 \geq 0.5$) were excluded.

3. SNP-based analysis

We performed multiple logistic regression analysis in an additive model using PLINK version 1.9 [18], controlling for age and sex. We used the FUMA web platform [19] version 1.3.6a (https://fuma.ctglab.nl/) for visualization of the GWAS results, for example using Manhattan plot.

4. Gene-based analysis

The summary statistics of the SNP-based analysis were used for genome-

wide gene association analysis using MAGMA [20] version 1.08 in the FUMA platform. In the FUMA, gene annotation function is provided based on Ensembl genes (build 92) including 19,436 protein-coding genes, 9249 non-coding RNAs, and other genes [19]. SNPs in the GWAS summary statistics were mapped to gene by the location of SNPs on the protein coding genes, and the gene-based P-value was computed. The Bonferroni correction was used for adjusting multiple comparison.

5. Gene-set analysis and GTEx gene expression

The P-value of the gene-based analysis was used for 4728 curated gene sets and 6166 GO terms acquired from MsigDB v5.2 in FUMA [19]. This analysis was computed with the default MAGMA setting, and FDR was used to adjust for multiple testing. To investigate the gene expression in specific tissues, FUMA utilizes the information on normalized gene expression of 54 tissue types from GTEx. FUMA provides a total of 22,146 mapped genes to entrez ID. This information is used to find differentially expressed gene (DEG) sets that are over- or under-expressed in a given tissue types against all other tissues.

6. PRS analysis

To determine the genetic risk measured by PRS, we performed PRS analysis in two steps. First, we received summary statistics from the Psychiatric Genomics Consortium GWAS for OCD from the meta-analysis by the International Obsessive Compulsive Disorder Foundation Genetics Collaborative and OCD Collaborative Genetics Association Studies [13,14,21]. We used the results of 2,688 cases and 7,037

controls as a discovery sample, and PRS was computed for each individual in dataset #1 of the present study for each selected SNP using PRSice-2 tool version 2.3.3 [22]. We determined the threshold of P-value of the discovery set showing the highest effect size. Second, based on the summary statistics of GWAS results of the present study (with dataset #1 as the discovery set), we investigated the best PRS model and the P-value threshold in the additional Korean samples (with dataset #2 as the target set).

8. Genetic correlation with onset age of OCD

We conducted further analyses to investigate the genetic effect on the onset age of OCD. In the case group of OCD in dataset #1, a total of 350 subjects had information on onset age. Using Plink 1.9, linear regression analysis was performed for risk loci associated with onset age of OCD, controlling for age and sex. GWAS results were visualized with the Manhattan plot using the FUMA platform. Genome-wide significant markers related to the onset age of OCD were analyzed in independent samples (dataset #2) for replication.

Ⅲ. RESULTS

1. Subjects and quality control

Dataset #1: After quality control, 420 patients with OCD were included in the analysis. Because the data of healthy controls were provided after quality control by the National Biobank of Korea, 2400 controls were analyzed. Regarding sex composition, there were 174 and 246 female and male patients, respectively, and 960 female and 1440 male healthy controls (chi-square = 0.303, p = 0.582). The mean age was significantly different between cases (30.02 years) and controls (52.57 years; t = -40.09, p < 0.001). Therefore, we performed logistic or linear regression analysis adjusting for age as a covariate. A total of 800,998 autosomal SNPs were identified, and after quality control, 428,570 SNPs remained for analysis.

Dataset #2: Six patients with OCD were excluded based on low call rates. In the control group, samples of eight individuals were excluded: seven individuals with low call rates, and one individual who was a sibling of another control sample. Finally, 109 cases and 220 controls were used for further analyses. After quality control, 565,433 SNPs were included in the analysis.

2. SNP-based association

A quantile-quantile (QQ) plot of the observed versus expected log (P) values under the null hypothesis is presented in Figure 1. The QQ plot for the genetic association of OCD status revealed no evidence of inflation of the test statistics.

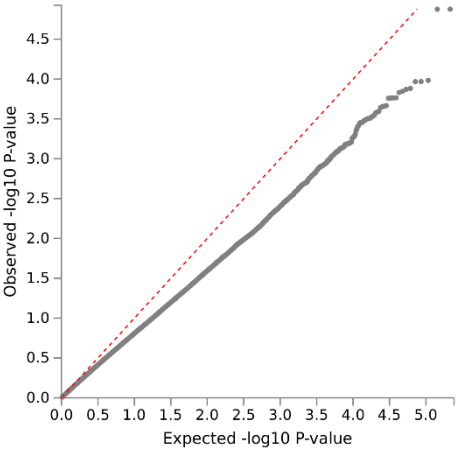


Figure 1. Quantile-quantile (QQ) plots of observed versus expected -log (*P*) statistics in SNP-based analysis

No SNPs were found to be more than the genome-wide significance threshold ($P < 5 \times 10^{-8}$) in the case–control analysis (Figure 2, Table 1). The SNP rs186251548 located on chromosome 20 within LOC284788 gene showed the lowest P-value (1.326×10^{-5}) (Table 1). Thirty-six SNPs had p-values lower than 0.0005, and the number of SNPs ($P < 0.01$) was 1288.

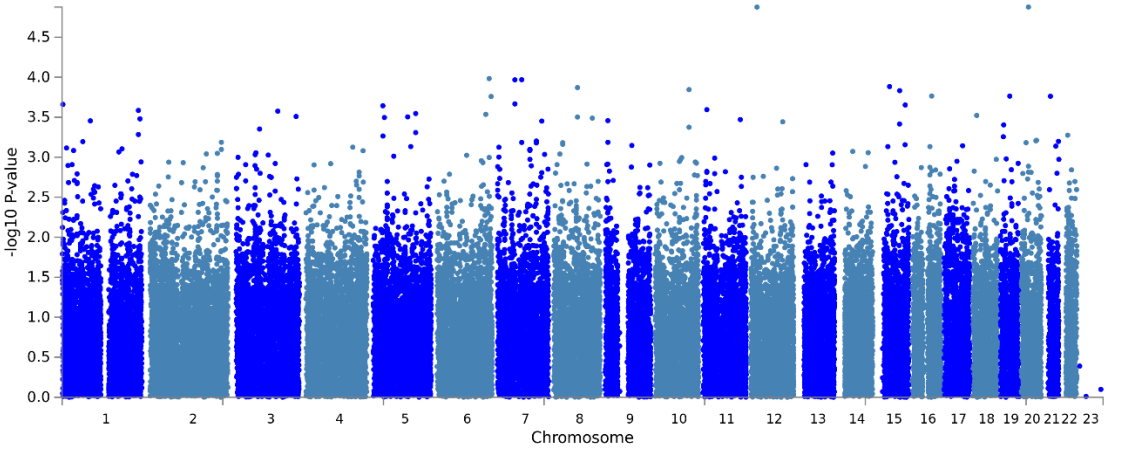


Figure 2. Manhattan plots of all genotyped single-nucleotide polymorphisms (SNPs) for case-control samples

Table 1. Genomic risk loci of OCD at p-value = 5×10^{-4}

Chr	Position	SNP	Genes	Region	Risk allele	Allele frequency in cases	Allele frequency in controls	OR	P-value
20	22510390	rs186251548	LOC284788,LINC00261	upstream,downstream	C	0.02262	0.01398	5.91	0.00001326
12	19851124	rs186176805	AEBP2,LOC100506393	intron,downstream,upstream	T	0.02738	0.02004	4.774	0.00001329
6	161882221	rs9346856	PARK2	intron	C	0.05	0.03096	3.298	0.0001039
7	75422442	rs62475544	CCL26,CCL24	upstream,downstream	T	0.2088	0.1819	1.878	0.0001074
7	54225506	rs17646638	HPVC1,LINC01446	upstream,downstream	A	0.03452	0.02255	3.561	0.0001078
15	41768900	rs138345663	RTF1	intron	C	0.02381	0.01317	5.038	0.0001309
8	74081273	rs77727575	SBSPON,LOC100130301	downstream,upstream	G	0.2064	0.1769	1.863	0.0001348
10	105509477	rs183599888	SH3PXD2A,SH3PXD2A-AS1	intron	C	0.02148	0.01087	5.696	0.0001429
15	73263455	rs76512277	NEO1,ADPGK-AS1	upstream,downstream	G	0.02143	0.01258	5.209	0.0001472
16	60137498	rs147470091	LOC101927580,LOC729159	upstream,downstream	C	0.03341	0.02029	3.621	0.0001718
19	29386942	rs149513234	LINC00906,LOC100420587	intron,upstream	T	0.02024	0.01104	5.476	0.0001723
21	21777160	rs151027610	LINC00320,LOC101927797	downstream,upstream	A	0.01667	0.01087	5.229	0.000173
6	168182043	rs34463918	C6orf123,LOC441178	upstream,downstream	G	0.3947	0.3382	1.679	0.0001741
1	2704698	rs114858168	TTC34	intron	T	0.03214	0.02107	3.68	0.0002189
15	90507390	rs79222095	ZNF710,C15orf38-AP3S2	upstream,downstream	T	0.03571	0.03111	2.984	0.0002224

5	30081449	rs187645562	CDH6,LOC101929681	downstream,upstream	C	0.02262	0.01341	4.3	0.0002275
11	12109374	rs11022162	MICAL2,DKK3	intron,upstream	T	0.4119	0.4527	0.6143	0.0002544
1	237502215	rs2808226	RYR2	intron	C	0.05357	0.03112	3.028	0.00026
3	130236369	rs75831106	COL6A5,COL6A6	downstream,upstream	C	0.04405	0.03172	2.986	0.0002657
5	132200057	rs138136756	GDF9	missense,intron,UTR-5	A	0.06563	0.04042	2.456	0.0002844
18	11881804	rs75501567	GNAL	UTR-3	C	0.06548	0.03783	2.669	0.0003006
3	187285995	rs151224154	SST,RTP4	downstream	C	0.02262	0.01462	4.69	0.0003097
5	107214953	rs77248417	FBXL17	intron,UTR-3	C	0.02381	0.01091	4.915	0.0003131
8	74096655	rs16938586	SBSPON,LOC100130301	downstream,upstream	G	0.1202	0.1024	2.059	0.0003145
5	34749416	rs9283739	RAI14	intron	G	0.1158	0.09078	1.989	0.000319
8	120171334	rs148348901	COLEC10,MAL2	downstream,upstream	T	0.03095	0.02838	3.574	0.0003246
1	242492973	rs4658476	PLD5	intron	T	0.02381	0.01209	4.191	0.0003317
11	116384580	rs12361273	LOC101929011,BUD13,LINC00900	upstream,downstream	T	0.2086	0.1834	1.756	0.0003386
9	8115293	rs989724	PTPRD,TMEM261	downstream,upstream	G	0.1464	0.1167	1.946	0.000349
1	88087820	rs12144799	LINC01364,PKN2-AS1	upstream,downstream	T	0.02976	0.01836	4.076	0.0003507
7	137932270	rs10954609	TRIM24,MIR4468	upstream,downstream	A	0.1167	0.1515	0.4681	0.0003535
12	99920334	rs141177655	ANKS1B	intron	G	0.02738	0.01485	4.215	0.0003611
15	72903605	rs7496834	GOLGA6B,MIR630	intron,upstream,downstream	T	0.1238	0.08038	2.141	0.0003845
19	10397238	rs281437	ICAM1	UTR-3,exon	T	0.1074	0.08708	2.134	0.0003949

10	105504654	rs55672765	SH3PXD2A	intron	T	0.09167	0.06943	2.226	0.0004216
5	132208861	rs765760	LEAP2,UQCRQ	UTR-5,downstream,upstream	T	0.1512	0.125	1.828	0.0004922

Notes: OCD, obsessive–compulsive disorder; Chr, chromosome; SNP, single nucleotide polymorphism; OR, odds ratio

From the GWAS results of the present study, we explored the characteristics of functional consequences on genes among the candidate SNPs (P-value threshold as 0.01 and r^2 threshold as 0.6). Functional consequences of the annotated SNPs (n = 917) on genes are presented in Figure 3; most of the SNPs were intergenic (38.3%) and intronic (41.5%). The proportion of SNPs located in the exon was 4% (n = 63; Figure 3).

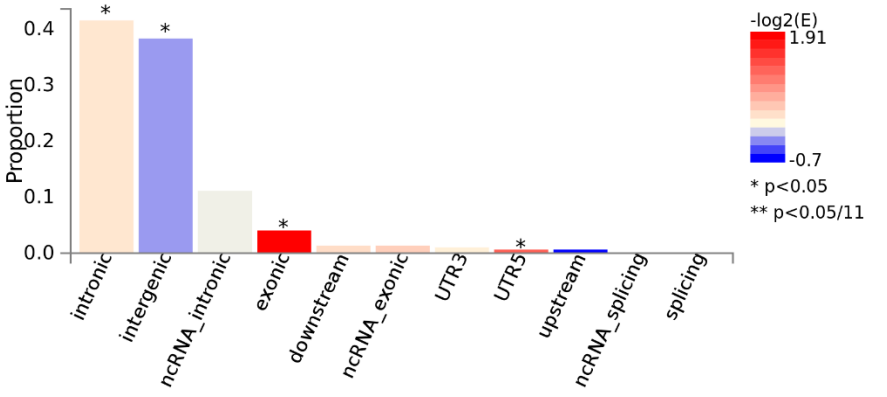


Figure 3. Functional consequences of SNPs on genes

3. Gene-based association

The QQ plot of gene-based analysis is presented in Figure 4.

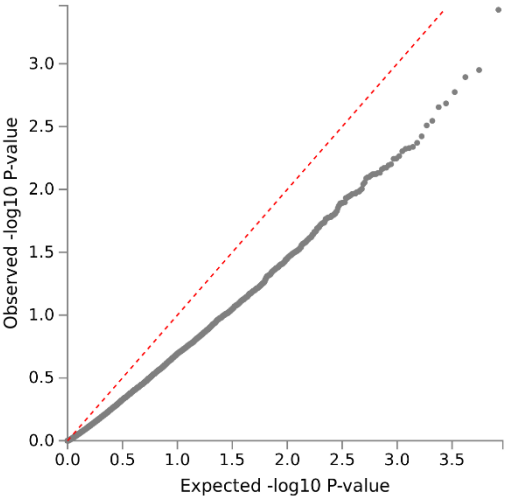


Figure 4. Quantile–quantile (QQ) plots of observed versus expected -log (*P*) statistics in gene-based analysis

Input SNPs were mapped to 16718 protein-coding genes; therefore, genome-wide significance was defined at $p = 0.05/16718 = 2.991 \times 10^{-6}$. No statistically significant genes were identified (Figure 5).

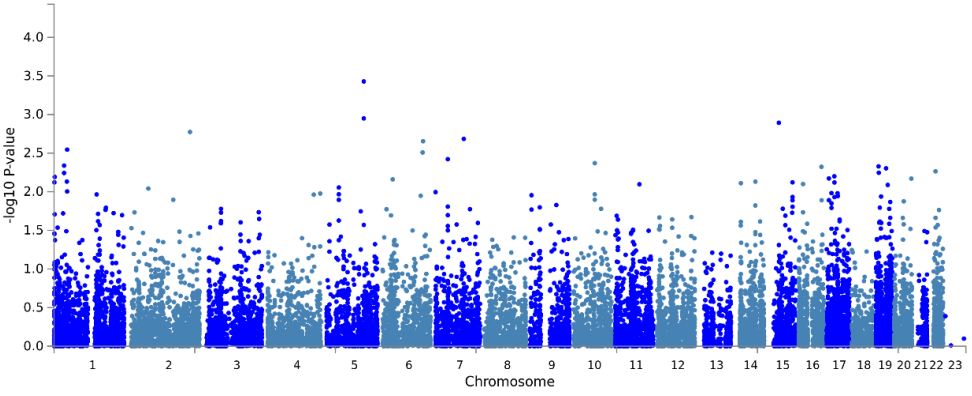
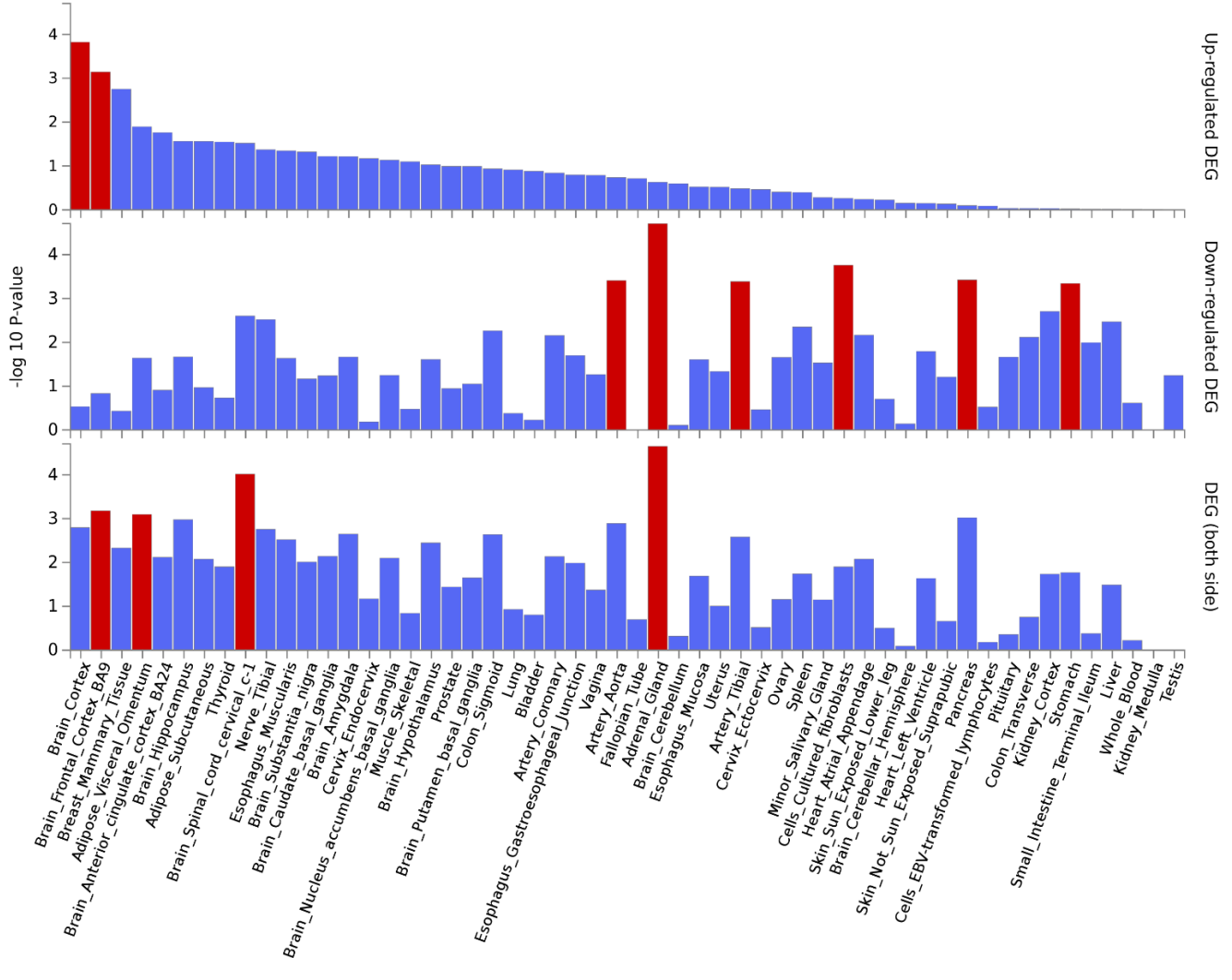


Figure 5. Manhattan plots of genes-based association for case–control samples

4. Tissue and cell types associated with annotated gene

In gene-set analysis, we used 917 lead SNPs that selected by P-value<0.01, there was no statistically significant gene set after Bonferroni correction.

Gene-based tissue expression was computed according to the GTEx v8 54 tissue types (Figure 6). From our GWAS results, a total of 3365 genes were input with unique entrez IDs for implementing differentially tissue expression. The expression of annotated genes by case–control comparison was increased in both the brain cortex and brain frontal cortex_BA9. The significance of differential gene expression in the brain frontal cortex_BA9 remained with the combined results of upregulation and downregulation.



Notes: DEG, differentially expressed genes. Red bars: significant difference

Figure 6. Gene-based expression analysis results by general tissue types

5. PRS analysis

[Step 1] Using the summary statistics of the previous meta-analysis [13,14,21] as the discovery set, PRS of the subjects in the present study was calculated (dataset #1). We conducted PRS analysis with matching SNP information by rsID, and among 428,570 SNPs in dataset #1, 8 SNPs were mismatched. When the P-value threshold was set at 0.4288 within the discovery set, the best effect size ($R^2 = 0.0044$) was shown with genotype information of 44,528 SNPs (P-value = 0.008) (Figure 7).

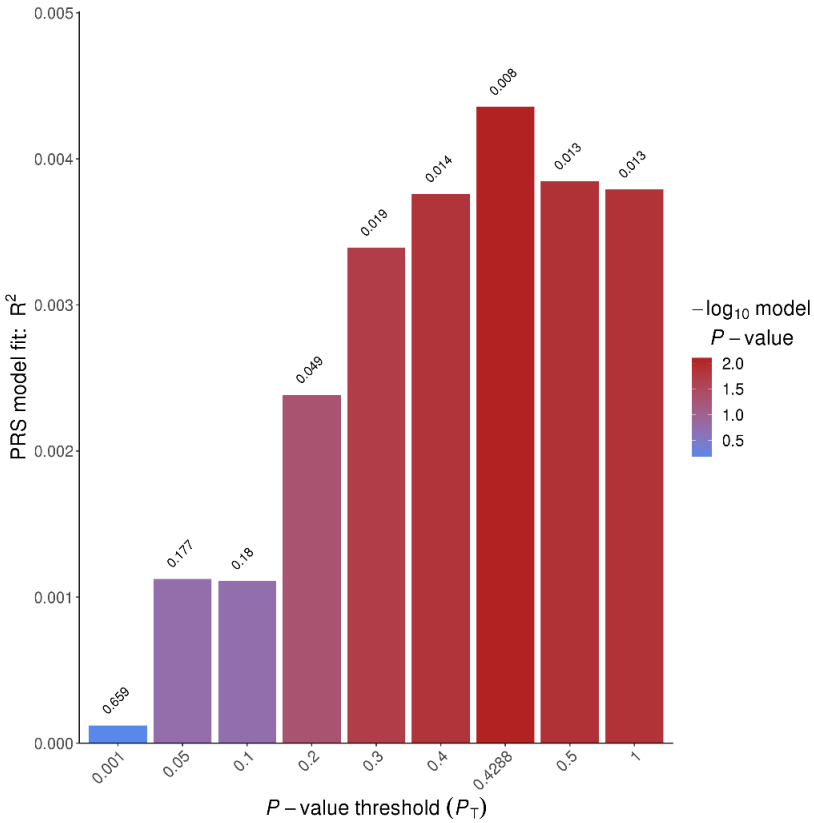


Figure 7. Effect size of the PRS model according to the P-value threshold when using previous GWAS results of Psychiatric Genomics Consortium

[Step 2] When the discovery data were set as GWAS results of the present study consisting of 420 cases and 2400 cotnrols (dataset #1), we computed additional PRS in independent target samples of 115 cases and 228 controls (dataset #2). The best effect size was $R^2 = 0.018$ (P-value = 0.041), with a P-value threshold of 0.0249 (number of SNPs: 2791) (Figure 8).

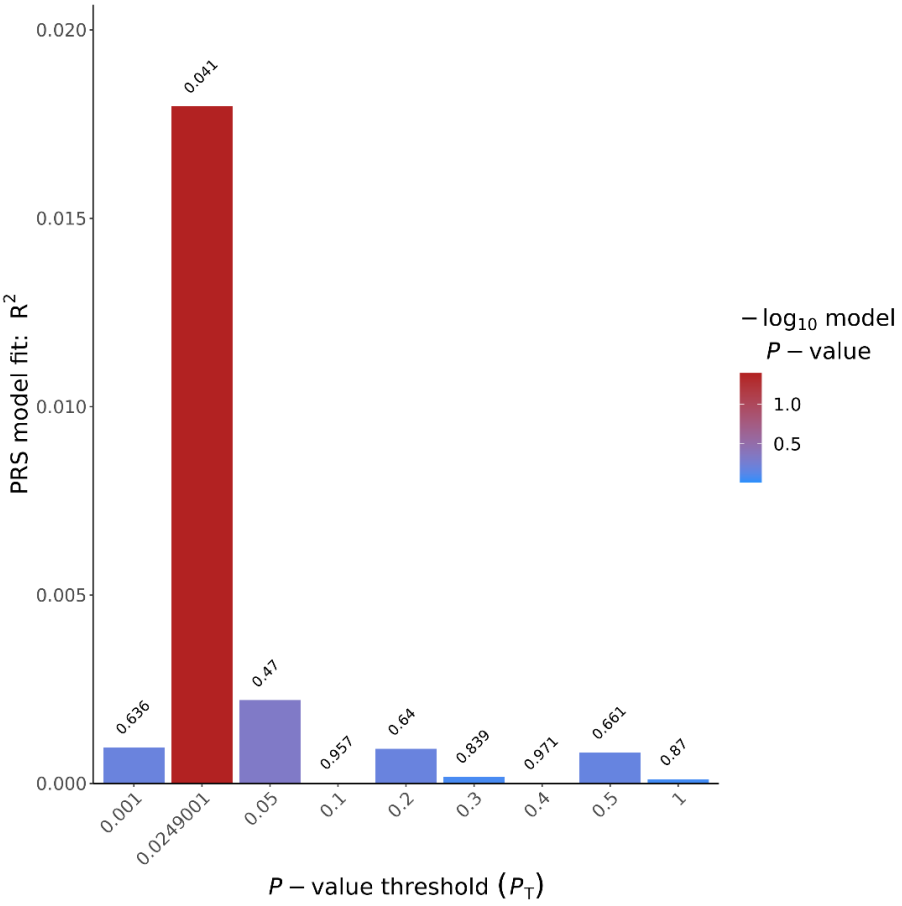


Figure 8. Effect size of the PRS model according to the P-value threshold when using GWAS results of the present study

6. Genome-wide association for onset age of OCD

We conducted further analysis in patients with OCD for genetic association with onset age. One SNP, rs372803, in the citron gene (*CIT*), showed genome-wide significance (P-value = 3.35×10^{-9}). (Figure 9)

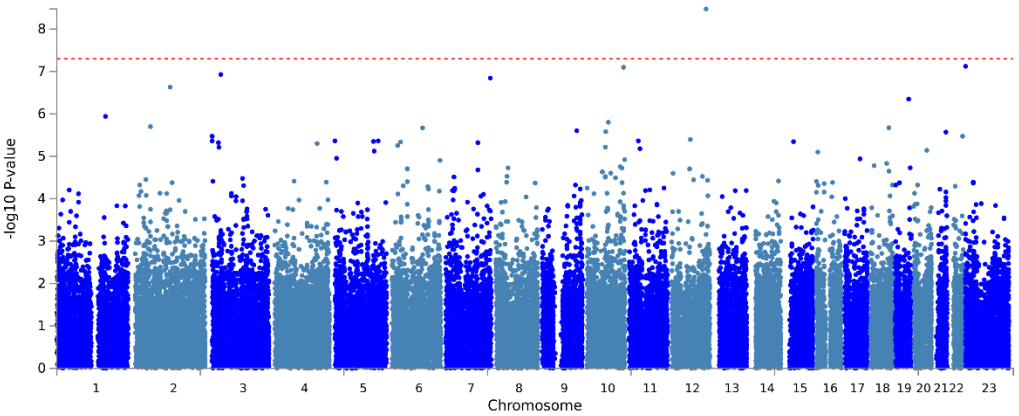


Figure 9. Manhattan plots of all genotyped single-nucleotide polymorphisms (SNPs) for onset age of OCD

Replication analysis with additional samples of 87 patients with OCD in dataset #2 was performed to validate the genetic effect of rs372803 on the onset age of OCD. There were 76 samples with major allele homozygote (GG), 11 samples with heterozygote (GT), and no minor allele homozygotes. The linear regression model with the enter method revealed that the genotype of rs372803 was not significant in explaining the onset age of OCD in replication samples (P-value = 0.897).

IV. DISCUSSION

We investigated 420 patients with OCD and 2400 healthy controls in the Korean population for 428,570 autosomal SNPs using the Korea Biobank Array. To our knowledge, this is the first GWAS on Asian patients with OCD. This study failed to detect genome-wide significant loci through SNP-based analysis or gene-based analysis. Further analysis investigating the genetic vulnerability of OCD using PRS method was performed, and we found that our PRS model explained 1.8% of variance of OCD status. Regarding the genetic influence on the clinical characteristics of OCD, we found a genome-wide significant marker, rs372803, in the *CIT* gene that is related to the onset age of OCD. However, we failed to replicate this positive result in the independent samples.

Genetic factors that affect OCD occurrence have been previously explored by GWAS and commercial SNP chips [13,14]. However, commercial SNP chips developed by non-Korean companies are not suitable for analyzing genetic traits in the Korean population because they are designed for a Western population whose genetic basis differs from that of an Asian population. Therefore, it is important to identify genetic SNPs specific to the Korean population using an appropriate analysis tool. In this regard, the Korea Biobank Array was developed under the Korea Biobank Array project by the Korea National Institute of Health of the Korea Centers for Disease Control and Prevention in 2014. This is a low-cost customized chip optimized for genetic research on diseases and complex traits in the Korean population; thus, utilization of the Korea Biobank Array is an appropriate method for genetic analysis in the Korean population.

Except for some rare diseases with well-known genetic predispositions, the purpose of GWAS is to identify genetic markers with odd ratios between 1.0 and 1.5, which correspond to the common disease–common variant hypothesis for most psychiatric diseases [23]. In addition, mental illness is a complex disease that is caused by a combination of genetic and environmental causes, and it is difficult for a single genetic mutation to have a very high odds ratio for disease onset. A previous GWAS on OCD with 1,406 cases (2,895 of which were of European, Afrikaner, and Jewish ancestry) failed to detect genome-wide significant markers, and the authors suggested that additional samples were needed to reach a sufficient sample size [13]. Recently, a GWAS using the Padua Inventory Revised Abbreviated, which measures obsessive–compulsive symptoms, analyzed 6931 subjects as part of the Netherland National Twin Registry [24]. In this study, authors identified genome-wide significant SNPs in the myocyte enhancer factor 2B neighbor gene in region 9p13 on chromosome 19 [24]. However, this positive finding may not have been valid because the self-report questionnaire may not be diagnostically valid. In a study on a major depressive disorder, the authors argued that genome-wide significant markers can be discovered when the sample size is more than 50,000 [25] therefore, the relatively small sample size of the present study (420 cases) could have resulted in a negative finding.

Other previous studies of GWAS in OCD attempted to find genetic architectures related to severity of symptom [26], sex difference [27], or relationships with other psychiatric disorders [28,29]. A recent study of 401 patients with OCD utilized the Yale-Brown Obsessive Compulsive Scale (Y-BOCS) for measuring the symptom severity and conducted linear regression with genome-wide genotype data

using the Infinium PsychArray-24 BeadChip kit (Illumina) [26]. Although this study could not find novel markers associated with OCD symptom severity with genome-wide significance, the authors suggested that the most associated SNP (rs7578149, P-value=1.89E-6) located between RPS16P2 and LAPTM4A was reported as an associated marker for MDD in female [30]. In regards to sex difference in genetics in OCD, a study using the dataset of OCD Psychiatry Genomics Consortium [13,14] presented results that there was no evidence for sex-dependent genetic effect on OCD both in autosomal chromosome and in X chromosome suggesting genetic homogeneity between the sexes in OCD [27]. There were two studies that investigated the shared genetic vulnerability between OCD and other psychiatric disorders; a study of OCD and anorexia nervosa (AN) using genotype data of 3495 AN patients, 2688 OCD patients, and 18,013 healthy controls, suggested a significant genetic relationship between AN and OCD [29]. However, the other study which was a meta-analysis of OCD and attention deficit hyperactivity disorder (ADHD) with samples of 2998 OCD and 5415 ADHD cases showed little evidence for shared genetic architectures between OCD and ADHD [28].

We also performed further gene-based analyses among possible OCD-related GWAS variants (P-value < 0.01). The general tissue expression data of GTEx v8 showed increased gene expression in the frontal area of the brain from our annotated gene list. The mechanism by which the upregulated gene expression level in the frontal lobe affects the susceptibility of OCD is uncertain, and some findings of brain imaging studies seem to be linked with our results. Several studies reported abnormally high activities throughout the frontal cortex, especially the orbitofrontal cortex, using functional magnetic resonance imaging, positron emission tomography,

or single photon emission computed tomography [31,32]. Therefore, the increased activity of the frontal area in OCD might be related to the upregulated gene expression. Further research is needed to elucidate the mechanisms underlying the influence of gene expression on the functional activity in the frontal lobe.

Most psychiatric disorders occur in a polygenic manner; [33,34] therefore, several studies investigated the genetic influences of thousands of SNPs with small effect sizes. PRS has emerged as an alternative method for calculating the degree of genetic risk in individuals. Some studies reported ethnicity-specific genetic risk loci and suggested potential population heterogeneity [35,36], while on the other hand, other studies suggested the trans-ethnic consistency in GWAS study results [37,38]. Therefore, some recent studies have attempted to investigate the genetic risk by comparing results of ethnic and trans-ethnic studies [39]. Here, we found that the PRS explains only 0.4% of the variance in OCD status when using the previous GWAS results of a large meta-analysis as a discovery set. This very small explanatory power might be due to differences in the genetic architecture according to ethnicity or small sample size. The samples in the discovery set were of European, South African Afrikaner, and Ashkenazi Jewish ancestry, and east-Asian participants were not enrolled [14]. We performed further PRS analysis with the discovery data set as GWAS results of the present study, we performed additional PRS analysis using independent cases and control samples of Korean ethnicity. The best PRS model could explain 1.8% of the variance in susceptibility of OCD, which showed better performance than that of the PRS model using the discovery set of European samples. However, compared with the findings of a previous study that suggested an effect size from 0.011 to 0.031 of PRS of OCD according to the P-value threshold [40], our

results showed relatively low explanatory power. A larger sample size of GWAS not only leads to a greater possibility of finding statistically significant loci but also increases the variance explained by PRS [41]. Although the PRS is an alternative method for compensating a limitation of conventional GWAS, such as missing heritability [42], our small sample size may encompass low variance explained by PRS for OCD status.

Although there was no significant marker in the association between patients with OCD and healthy controls, we found a novel SNP rs372803 with genome-wide significance associated with onset age. Positioned on chromosome 12, rs372803 is an intron variant of the *CIT* gene. Previous studies suggested that the CIT is concentrated at the post-synapse of γ-aminobutyric acid and glutamatergic neurons and modulates Rho signaling and glutamate receptor signaling [43,44]. In some animal studies, the CIT gene variant was related to the developing CNS [45,46]. Furthermore, the *CIT* gene polymorphism was reported to be related to schizophrenia [47] and bipolar disorder [48]. Considering early-onset OCD can be associated with neurodevelopmental problems, the *CIT* gene might play a role in the neurodevelopmental process, consequently influencing the onset age of OCD. A previous candidate gene study [49] targeted dopamine-related (dopamine transporter gene, DAT; dopamine receptor type 4, DRD4; catechol-O-methyltransferase gene, COMT) and serotonin related gene (5-HT receptor types 2A, 5-HT2A; 1Dβ, 5-$HT_{1D\beta}$). Among 180 Caucasian samples, there was a significant difference in allelic distribution of DRD4 SNP between early- and late-onset OCD [49]. Relatively recent study with 74 adolescent OCD patients, compared the very early-onset (ages under

10 years) and early-onset groups [50]. In this study, among 44 SNPs of 5 candidate genes of serotonin system, two SNPs (rs6311, rs4942587) of the HTR2A were significantly related to the very-early onset OCD [50]. However, these two studies [49,50] are limited by the candidate gene approach, and the strength our study is that it was the first to carry out a genome-wide analysis to examine the genetic influence on the onset age of OCD.

Our study has several limitations. First, the number of subjects was too small to find the genome-wide significance (420 cases and 2400 controls). However, the present study was the first GWAS of OCD in East Asian population with well-defined OCD phenotype. Larger studies with consortium are needed to find novel significant SNPs. Second, the ethnicity of samples was limited to Koreans, and therefore, the results of this study could not be generalized for another ethnicity. This homogeneity of ethnicity could be useful for further analysis with international genetic consortium conducting trans-ethnic genome-wide meta-analysis. Third, the onset age of OCD was defined by retrospective recall, and the possibility of recall bias cannot be excluded.

Ⅴ. CONCLUSION

We have conducted GWAS in Korean population to investigating the association between genetic variation and susceptibility of OCD. The present study could not discover the novel markers with genome-wide significance, however, gene-based results showed a possibility that effect of genetic risk on the pathophysiology of OCD. In addition, we found the novel candidate marker that related to the age of onset in OCD. Further studies with a large sample size to reveal the risk loci in OCD.

REFERENCES

1 Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 2005; **62**: 593-602.

2 Pauls DL, Alsobrook II JP, Goodman W, Rasmussen S, Leckman JF A family study of obsessive-compulsive disorder. *The American Journal of Psychiatry* 1995; **152**: 76.

3 Nestadt G, Samuels J, Riddle M, Bienvenu OJ, Liang K-Y, LaBuda M, et al A family study of obsessive-compulsive disorder. *Archives of general psychiatry* 2000; **57**: 358-63.

4 Mataix-Cols D, Boman M, Monzani B, Rück C, Serlachius E, Långström N, et al Population-based, multigenerational family clustering study of obsessive-compulsive disorder. *JAMA psychiatry* 2013; **70**: 709-17.

5 Saxena S, Rauch SL Functional neuroimaging and the neuroanatomy of obsessive-compulsive disorder. *Psychiatric Clinics of North America* 2000; **23**: 563-86.

6 Taylor S Molecular genetics of obsessive–compulsive disorder: a comprehensive meta-analysis of genetic association studies. *Molecular psychiatry* 2013; **18**: 799.

7 Gratacòs M, Costas J, de Cid R, Bayés M, González JR, Baca-García E, et al Identification of new putative susceptibility genes for several psychiatric disorders by association analysis of regulatory and non-synonymous SNPs of

306 genes involved in neurotransmission and neurodevelopment. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2009; **150**: 808-16.

8 Hemmings SM, Kinnear CJ, Van Der Merwe L, Lochner C, Corfield VA, Moolman-Smook JC, et al Investigating the role of the brain-derived neurotrophic factor (BDNF) val66met variant in obsessive-compulsive disorder (OCD). *The World Journal of Biological Psychiatry* 2008; **9**: 126-34.

9 Hounie AG, Cappi C, Cordeiro Q, Sampaio AS, Moraes I, do Rosário MC, et al TNF-alpha polymorphisms are associated with obsessive-compulsive disorder. *Neuroscience letters* 2008; **442**: 86-90.

10 Sampaio AS, Fagerness J, Crane J, Leboyer M, Delorme R, Pauls DL, et al Association between polymorphisms in GRIK2 gene and obsessive-compulsive disorder: a family-based study. *CNS neuroscience & therapeutics* 2011; **17**: 141-7.

11 Hirschhorn JN, Daly MJ Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**: 95-108.

12 Bush WS, Moore JH Genome-wide association studies. *PLoS computational biology* 2012; **8**: e1002822.

13 Mattheisen M, Samuels JF, Wang Y, Greenberg BD, Fyer AJ, McCracken JT, et al Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. *Molecular psychiatry* 2015; **20**: 337.

14 Stewart SE, Yu D, Scharf JM, Neale BM, Fagerness JA, Mathews CA, et al Genome-wide association study of obsessive-compulsive disorder. *Mol*

Psychiatry 2013; **18**: 788-98.

15 Kwon SK, Woo J, Kim SY, Kim H, Kim E Trans-synaptic adhesions between netrin-G ligand-3 (NGL-3) and receptor tyrosine phosphatases LAR, protein-tyrosine phosphatase delta (PTPdelta), and PTPsigma via specific domains regulate excitatory synapse formation. *J Biol Chem* 2010; **285**: 13966-78.

16 Takahashi H, Arstikaitis P, Prasad T, Bartlett TE, Wang YT, Murphy TH, et al Postsynaptic TrkC and presynaptic PTPsigma function as a bidirectional excitatory synaptic organizing complex. *Neuron* 2011; **69**: 287-303.

17 Takahashi H, Katayama K, Sohya K, Miyamoto H, Prasad T, Matsumoto Y, et al Selective control of inhibitory synapse development by Slitrk3-PTPdelta trans-synaptic interaction. *Nat Neurosci* 2012; **15**: 389-98, s1-2.

18 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 2007; **81**: 559-75.

19 Watanabe K, Taskesen E, Van Bochoven A, Posthuma D Functional mapping and annotation of genetic associations with FUMA. *Nature communications* 2017; **8**: 1-11.

20 de Leeuw CA, Mooij JM, Heskes T, Posthuma D MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015; **11**: e1004219.

21 Arnold PD, Askland KD, Barlassina C, Bellodi L, Bienvenu O, Black D, et al Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Molecular psychiatry* 2018; **23**: 1181-.

22 Choi SW, Mak TS-H, O'Reilly PF Tutorial: a guide to performing polygenic

risk score analyses. *Nature Protocols* 2020; **15**: 2759-72.

23 Hodge SE, Greenberg DA How can we explain very low odds ratios in GWAS? I. Polygenic models. *Human Heredity* 2016; **81**: 173-80.

24 Den Braber A, Zilhao N, Fedko I, Hottenga J, Pool R, Smit D, et al Obsessive–compulsive symptoms in a large population-based twin-family sample are predicted by clinically based polygenic scores and by genome-wide SNPs. *Translational psychiatry* 2017; **6**: e731.

25 Nishino J, Ochi H, Kochi Y, Tsunoda T, Matsui S Sample size for successful genome-wide association study of major depressive disorder. *Frontiers in genetics* 2018; **9**: 227.

26 Alemany-Navarro M, Cruz R, Real E, Segalàs C, Bertolín S, Baenas I, et al Exploring genetic variants in obsessive compulsive disorder severity: A GWAS approach. *Journal of affective disorders* 2020; **267**: 23-32.

27 Khramtsova EA, Heldman R, Derks EM, Yu D, Consortium TSOCDWGotPG, Davis LK, et al Sex differences in the genetic architecture of obsessive–compulsive disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2019; **180**: 351-64.

28 Ritter ML, Guo W, Samuels JF, Wang Y, Nestadt PS, Krasnow J, et al Genome Wide Association Study (GWAS) between Attention Deficit Hyperactivity Disorder (ADHD) and Obsessive Compulsive Disorder (OCD). *Frontiers in molecular neuroscience* 2017; **10**: 83.

29 Yilmaz Z, Halvorsen M, Bryois J, Yu D, Thornton LM, Zerwas S, et al Examination of the shared genetic basis of anorexia nervosa and obsessive–compulsive disorder. *Molecular psychiatry* 2020; **25**: 2036-46.

30 Aragam N, Wang K-S, Pan Y Genome-wide association analysis of gender differences in major depressive disorder in the Netherlands NESDA and NTR population-based samples. *Journal of affective disorders* 2011; **133**: 516-21.

31 Pauls DL, Abramovitch A, Rauch SL, Geller DA Obsessive–compulsive disorder: an integrative genetic and neurobiological perspective. *Nature Reviews Neuroscience* 2014; **15**: 410-24.

32 Nakao T, Okada K, Kanba S Neurobiological model of obsessive–compulsive disorder: Evidence from recent neuropsychological and neuroimaging findings. *Psychiatry and clinical neurosciences* 2014; **68**: 587-605.

33 Gandal MJ, Geschwind DH Polygenicity in Psychiatry—Like It or Not, We Have to Understand It. *Biological Psychiatry* 2021; **89**: 2-4.

34 Sullivan PF, Geschwind DH Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell* 2019; **177**: 162-83.

35 Simino J, Wang Z, Bressler J, Chouraki V, Yang Q, Younkin SG, et al Whole exome sequence-based association analyses of plasma amyloid-β in African and European Americans; the Atherosclerosis Risk in Communities-Neurocognitive Study. *PloS one* 2017; **12**: e0180046.

36 Li X, Su X, Liu J, Li H, Li M, Li W, et al Transcriptome-wide association study identifies new susceptibility genes and pathways for depression. *Translational psychiatry* 2021; **11**: 1-13.

37 Marigorta UM, Navarro A High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS genetics* 2013; **9**: e1003566.

38 Strawbridge RJ, Ward J, Lyall LM, Tunbridge EM, Cullen B, Graham N, et

al Genetics of self-reported risk-taking behaviour, trans-ethnic consistency and relevance to brain gene expression. *Translational psychiatry* 2018; **8**: 1-11.

39 Shigemizu D, Mitsumori R, Akiyama S, Miyashita A, Morizono T, Higaki S, et al Ethnic and trans-ethnic genome-wide association studies identify new loci influencing Japanese Alzheimer's disease risk. *Translational psychiatry* 2021; **11**: 1-10.

40 Bey K, Weinhold L, Grützmann R, Heinzel S, Kaufmann C, Klawohn J, et al The polygenic risk for obsessive-compulsive disorder is associated with the personality trait harm avoidance. *Acta Psychiatrica Scandinavica* 2020; **142**: 326-36.

41 Wray NR, Lin T, Austin J, McGrath JJ, Hickie IB, Murray GK, et al From basic science to clinical application of polygenic risk scores: a primer. *JAMA psychiatry* 2021; **78**: 101-9.

42 Uricchio LH Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. *Human genetics* 2020; **139**: 5-21.

43 Furuyashiki T, Fujisawa K, Fujita A, Madaule P, Uchino S, Mishina M, et al Citron, a Rho-target, interacts with PSD-95/SAP-90 at glutamatergic synapses in the thalamus. *Journal of Neuroscience* 1999; **19**: 109-18.

44 Zhang W, Benson DL Targeting and clustering citron to synapses. *Molecular and Cellular Neuroscience* 2006; **31**: 26-36.

45 Di Cunto F, Imarisio S, Hirsch E, Broccoli V, Bulfone A, Migheli A, et al Defective neurogenesis in citron kinase knockout mice by altered cytokinesis and massive apoptosis. *Neuron* 2000; **28**: 115-27.

46 Sarkisian MR, Li W, Di Cunto F, D'Mello SR, LoTurco JJ Citron-Kinase, A Protein Essential to Cytokinesis in Neuronal Progenitors, Is Deleted in the FlatheadMutant Rat. *Journal of Neuroscience* 2002; **22**: RC217-RC.

47 Nicodemus KK, Callicott JH, Higier RG, Luna A, Nixon DC, Lipska BK, et al Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. *Human genetics* 2010; **127**: 441-52.

48 Lyons-Warren A, Chang J, Balkissoon R, Kamiya A, Garant M, Nurnberger J, et al Evidence of association between bipolar disorder and Citron on chromosome 12q24. *Molecular psychiatry* 2005; **10**: 807-9.

49 Hemmings SaM, Kinnear CJ, Lochner C, Niehaus DJ, Knowles JA, Moolman-Smook JC, et al Early-versus late-onset obsessive–compulsive disorder: investigating genetic and clinical correlates. *Psychiatry Research* 2004; **128**: 175-82.

50 Plana MT, Torres T, Rodríguez N, Boloc D, Gassó P, Moreno E, et al Genetic variability in the serotoninergic system and age of onset in anorexia nervosa and obsessive-compulsive disorder. *Psychiatry research* 2019; **271**: 554-8.

ABSTRACT (IN KOREAN)

한국인 강박장애의 전장 유전체 연관성 분석

<지도교수 김 세 주>

연세대학교 대학원 의학과

박 천 일

강박 장애는 강박 사고 및 강박 행동을 특징으로 하며 일상생활에 장애를 일으키는 임상적 질환으로 많은 선행 연구에서 강박 장애의 유전 적 원인을 조사해 왔다. 가설에 기반하지 않고 유전체 전체에 걸쳐 원인 유전자를 탐색하는 전장 유전체 연관성 분석 (GWAS, genome-wide association study) 이 대두되었으나 강박 장애에서의 연구는 매우 부족하다. 기존 두 개의 강박 장애 연구에서 유의한 마커를 발굴하지 못하였으며, 현재까지 아시아인을 대상으로 한 강박 장애 GWAS 연구는 전무한 실정이다. 따라서 본 연구에서는 한국인 강박 장애 환자의 전장 유전체 연관성 분석을 통하여 원인 마커를 규명하고자 하였다. 강박 장애 환자 429 명을 모집하여 채취한 전혈을 한국인칩 분석을 시행하여 전장 유전체 정보를 획득하였다. 대조군은 질병관리청 산하 국립보건연구원 바이오뱅크에서 2400명의 한국인칩 데이터를 분양 받았다. 환자-대조군 연관성에 대하여 로지스틱 회귀 분석을 수행하였으며 발병 연령 등의

임상 정보와의 유전 연관성을 추가 분석 시행하였다. 분석 결과 환자-대조군 분석에서는 통계적으로 유의한 마커는 확인되지 않았다. 유전자 기반 분석을 추가로 실시한 결과 유전적 특성과 뇌 전두엽 영역에서의 유전자 발현 증가가 연관되어 있음을 확인할 수 있었다. 본 연구의 환자-대조군 분석 결과를 바탕으로 다중유전자위험점수를 계산한 결과 추가 환자-대조군을 대상으로 하였을 때 1.8%의 설명력을 갖는 것으로 나타났다. 마지막으로 강박 장애 환자군 만을 대상으로 한 추가 분석에서는 발병 연령과 통계적으로 유의한 연관성을 보이는 한 개의 변이를 확인할 수 있었으나, 독립적인 환자-대조군에서 결과가 재현되지 않았다. 강박 장애 감수성과 관련된 유전 변이를 확인하기 위하여 컨소시엄 등을 통한 더 많은 대상자를 대상으로 한 연구가 필요하다.

핵심되는 말: 강박장애, 전장유전체연관성분석, 발병연령, 다중유전자위험점수