의학박사 학위논문

파킨슨병의 전장유전체상관분석: 한국인 특이적, 성별 특이적 분석

Ethnicity-specific and sex-specific genome wide association study on Parkinson's disease

울 산 대 학 교 대 학 원

의 학 과

박 계 원

파킨슨병의 전장유전체상관분석: 한국인 특이적, 성별 특이적 분석

지 도 교 수　　정 선 주

이 논문을 의학박사 학위 논문으로 제출함

2022년 2월

울 산 대 학 교　대 학 원

의　학　과

박 계 원

박계원의 의학박사학위 논문을 인준함

심사위원장 김 재 승 인

심사위원 고 성 범 인

심사위원 김 성 윤 인

심사위원 전 상 용 인

심사위원 정 선 주 인

울 산 대 학 교 대 학 원

2022년 2월

감사의 글

끝이 보이지 않는 터널 같던 박사 논문이 이렇게 완성되어 갑니다. 해변을 가득 메운 지식의 모래사장에 그래도 한 톨의 모래를 뿌렸구나 스스로를 자랑해봅니다. 재작년 시작되어 아직도 그 위세를 떨치고 있는 코로나19 바이러스의 영향으로 박사과정에 많은 우여곡절이 있었습니다. 많은 분들의 도움이 없었다면 무사히 논문을 발표할 수 없었을 것입니다.

먼저 파킨슨병 유전체 연구를 시작하여 이 박사 학위 논문이 나오기까지, 그리고 신경과 의사로 걸음마를 시작하였던 전공의 시절부터 이상운동질환분야 전문의가 되기까지 한결같이 든든한 멘토가 되어 주신 지도교수님, 정선주 교수님께 깊이 감사드립니다. 교수님께서는 연구와 임상 뿐 아니라 인생의 고민에 대해서도 늘 날카로운 통찰력으로 혜안을 제시하여 주시었습니다. 논문을 심사하는 과정에서 꼼꼼히 분석하고 아낌없는 조언을 주신 심사위원장 김재승 교수님, 심사위원 김성윤, 전상용 교수님께 깊은 감사의 인사를 올립니다. 엄중한 시국에 대면 회의의 어려움에도 불구하고, 외부 심사위원으로 좋은 말씀 들려주신 고성범 교수님께도 깊이 감사드립니다. 분석에 많은 도움을 주시고 잦은 질의로 폐를 끼쳤던 신은순 박사님과 DNA LINK 일원 여러분께도 항상 감사드립니다.

물심양면 서로를 도우며 함께 공부하며 밤을 지샌 동반자 김홍범과 사랑하는 아들 태민이에게 이 논문을 바칩니다. 태민이에게 저의 빈자리를 부족함없이 채워 주시는 부모님, 사랑합니다. 여러분이 없었다면 이 논문을 완성하지 못하였을 것입니다. 이제 막 대한민국의 의사-과학자로 발걸음을 뗀 저의 앞으로의 연구가 파킨슨병과 이상운동질환 환자들에게 실체적 도움을 줄 수 있기를, 이 논문이 그 초석이 되기를 기도하며 글을 마칩니다.

Summary

Parkinson's disease (PD) is one of the most common neuro-degenerative disorder worldwide. Genome-wide association studies (GWAS) on PD have discovered 90 genome-wide significant risk variants associated with PD. However, majority of the PD GWAS has focused on Europeans. Furthermore, despite several sex-specific clinical difference of PD, little attention has been paid upon the genetic difference between sex. Our study aimed to identify the genome-wide risk variants of sporadic PD in Korean-specific and sex-specific manner.

A total of 1,050 PD cases and 5,000 controls with Korean ethnicity were included. For the case-control analyses, we used logistic additive model adjusted for the age at the onset of Parkinsonism, or age at sample, and sex. We also performed candidate gene analysis with previously known 59 genes associated with sporadic PD with a more relaxed quality control standard. We applied same statistical model to sex-specific analysis, between 554 female cases and 2,610 female controls, and between 496 male cases and 2,390 male controls.

In the case-control analysis with age at onset of Parkinsonism and sex as the covariate, 492,970 SNPs passed the marker quality control (QC). Of them, 8 SNPs surpassed Bonferroni-corrected genome-wide significance ($P < 1.01 \times 10^{-7}$). These variants included 4 SNPs from *SNCA* locus (rs3796661, rs356203, rs11931074, and rs12640100), and 3 SNPs from *PARK16* locus (rs708726, rs947211, and rs708723). Candidate gene analysis further identified rs34778348 (G2385R) at the *LRRK2* locus associated with PD ($P = 4.77 \times 10^{-13}$). In the case-control analysis with age at sample and sex as the covariate, 493,000 SNPs passed marker QC. The 4 SNPs from the SNCA (rs3796661, rs356203, rs11931074, and rs12640100) surpassed Bonferroni-corrected genome-wide significance. However, variants at the *GBA* or *MAPT* locus, which are strongly associated with European PD, did not show significant association in Korean ethnicity. In female-only analysis, 486,510 SNPs passed the marker QC. The 5 SNPs

surpassing genome-wide significance threshold under Bonferroni correction ($P < 1.03 \times 10^{-7}$) included rs34778348 in *LRRK2* locus, which was a missense SNP ($P = 1.25 \times 10^{-9}$), and the other 4 in the *SNCA* locus (rs3796661, rs126401100, rs356203, and rs11931074). In male-only analysis, 488,631 SNPs passed the marker QC. None of the SNPs surpassed genome-wide significance threshold under Bonferroni correction ($P < 1.02 \times 10^{-7}$). However, the most significant signal was the rs708726 in the *PARK16* locus (P= 8.23 × 10-^{6}). Only rs3796661 in the *SNCA* locus showed *P* value under 10^{-4}, reflecting its small effect on male than on female patients.

In this first Korean ethnicity-specific GWAS on the susceptibility of PD, we found the SNPs in the *SNCA* and *PARK16* strongly associated with Korean PD. We did not find any association with variants in the *GBA* or *MAPT* gene, which are one of the strongest signals associated with European PD. We also observed strong association of variants in the SNCA locus in female PD patients but not in male patients, while variants of the *PARK16* were the strongest association with male PD but not in female PD. These findings suggest the different genetic contribution to sporadic PD across ethnicity and gender.

차 례

Table and Figure Contents

Introduction

Parkinson's disease (PD) is one of the most common neuro-degenerative disorder worldwide. PD is caused by the neuronal loss of the substantia nigra pars compacta of the midbrain, with its pathological hallmark being the presence of Lewy bodies and abnormal accumulation of the protein alpha-synuclein in the brain.[1] The disease is categorized as a movement disorder, its cardinal motor features being resting tremor, rigidity, bradykinesia, stooped posture, and postural instability.

For the past two decades, genome-wide association studies (GWAS) have shed lights to the genetic background of various sporadic common disease traits.[2] Previous GWAS on PD has also made a great progress,[3] discovering about 90 genome-wide significant risk signals in the most recent meta-analysis.[4] The studies confirmed the effect of genes responsible for familial PD, such as *SNCA*, and *LRRK2*, in sporadic PD as well. These signals reflect the important roles of genes related to abnormal pathogenesis of alpha-synuclein and genes related to the function in autophagy and lysosome, suggesting its role in the disease susceptibility and possible utilization for drug development.[3]

However, the majority of previous GWAS on sporadic PD focused primarily on European population.[4] A certain risk variant which is common in a population may be uncommon or not effective in other population. Thus, adopting the appropriate genotyping platform and performing accurate population stratification strategy is crucial for the proper conduct of a GWAS. Several recent PD GWAS focusing on populations outside Western world discovered a discrepancy in the genetic contribution to PD across ethnicities.[5] Korea is one of the most rapidly aging society, hence the prevalence of PD which is an age-related degenerative disorder is rapidly rising in the country.[6] Despite such backgrounds, no studies have previously performed GWAS on sporadic

PD specific to Korean population. Furthermore, there are only a couple of GWAS on PD in East Asian population which is the largest worldwide.

On the other hand, there are difference in the clinical characteristics in PD according to sex.[7] Possible reasons proposed for such sex-specific clinical differences of PD includes environmental factors, such as pesticide use, hormonal factors, such as the neuroprotective effect of estrogen, and unveiled genetic factors.[8] Despite the striking sex-specific clinical difference of PD, little attention has been paid to the genetic difference between male and female patients with PD. Only recently, a PD GWAS was conducted in a sex-specific manner in European population, showing no sex-specific differences.[9] Thus, further replication studies in such sex-specific manner, especially outside the western world, are warranted.

For such backgrounds, our study primarily aimed to identify the genetic variants associated with PD by applying customed GWAS chip, including ethnicity-specific genetic variants, focusing on a genetic isolate, the Koreans. We secondly aimed to determine the genomic risk variants for PD in sex-specific manner, by applying GWAS separately on male and female population.

Methods

Subjects

We recruited patients with PD in Asan Medical Center, Seoul, South Korea from January 2011 to April 2016. A total of 1,070 cases were enrolled. included ethnically Korean patients who were diagnosed as sporadic PD by movement disorder specialists according to the United Kingdom Parkinson's Disease Brain Bank Criteria.[10] Baseline demographics including age at sample, age at the onset of PD, sex, and family history of Parkinsonism were collected. We defined the age at onset as at the time when one of the motor cardinal symptoms (resting tremor, rigidity, bradykinesia, stooped posture, or postural instability) was noted by the patient or close caregiver. Exclusion criteria were those who were ethnically not Korean; genetically confirmed hereditary Parkinsonism; and those with signs of atypical Parkinsonism. This included cerebellar signs, Parkinsonism not-responsive to levodopa, supranuclear gaze palsy, early severe autonomic dysfunction, early severe dementia with disturbances of memory, language, and praxis, and otherwise-unexplained pyramidal signs. Informed consent was obtained in every patient for locally approved protocols. For controls, we obtained the samples of 5,000 age and sex matched healthy controls from the Korea Biobank Project.

Genotyping

All patients underwent peripheral blood sampling for DNA extraction. All samples were genotyped on Korean Chip obtained from the Korean Chip Consortium, designed by Center for Genome Science, Korea National Institute of Health.[11] Genotyping chip is a tiny chip that contains up to millions of genetic variations used for genomic research. Choosing an appropriate genotyping platform for the specific population being studied is essential for

GWAS.[12] However, preexisting genotyping platforms were mostly developed for European population. Its genomic coverage of common SNPs for Asians was poorer compared with that for Europeans. The Korean Chip was developed by the Korean Chip Project of Korean National Institute of Health to standardize the genotypic platform optimal for Koreans. It is an SNP microarray chip containing 833,000 specific genetic variations of Korean people. Its genomic coverage is more than 85% for SNP of minor allele frequency (MAF) > 5%. It contains 585,300 tagging variants covering 95% of SNPs with MAF>5% and 73% for SNPs with MAF 1-5%,[13] 148,800 genetic variation in protein structural change (nonsynonymous), 44,500 genetic variants that causes protein dysfunction predicted to be damaging, and 54,400 other functional variations.

All samples were assayed on Affymetrix Axiom® 2.0 Reagent Kit. A total of 200 ng from each extracted and purified genomic DNA were prepared for each sample. Manual target preparation for the assay was processed according to the manufacturer's protocol. *DNA amplification:* A volume of 20 ul of each genomic DNA at a concentration of 5ng/ul or 10ng/ul for each sample were placed on the plate. Denaturation Master Mix was added to each well and were incubated for 10 minutes. Neutralization solution was added to each sample. Then, the Amplification Master Mix was added to each well and were incubated in an oven set 37°C for 24 hours. *Fragmentation and precipitation:* After 24 hours, the amplification reaction was stopped by transferring the amplification plate from the 37°C oven to another oven set at 65°C, then back to the 37°C oven. Fragmentation Master Mix was added to the wells and were incubated, and then stop solution was added to each reaction. Precipitation Master Mix was added subsequently and was incubated in -20°C freezer overnight. *Drying, resuspension, and quality control:* The Precipitation Plate was thawed, and was centrifuged for 40 minutes at 4°C. Following centrifugation, the liquid from the Precipitation Plate was emptied and were dried in an oven for 20 minutes at 37°C. Resuspension and Hybridization Master Mix was prepared and were added to the dry pellet. At this stage, the Resuspension Plate underwent

quality control under the recommended manufacturer's protocol. *Denaturation and hybridization:* The fragmented genomic DNA segments were end-labeled with biotinylated nucleotides. After ligation, the arrays were imaged on the GeneTitanMC Instrument and the image was analyzed using Genotyping Console™ Software.

Quality control

Because GWAS analyzes up to millions of SNP between cases and controls, genotyping errors should be diligently sought out to avoid spurious associations. Thus, a number of quality control (QC) steps should be applied both on per-sample and per-SNP basis. Our study applied seven QC steps to filter out low-quality samples and low-quality SNPs. All QC steps were done by the PLINK software version 1.90. First, markers with low call rate which have high rates of genotype missingness over 5% for cases or controls were removed. Second, samples with discrepancy between the reported sex of the individuals in the dataset and their sex based on X chromosome zygosity was removed for possible sample mix-ups. Third, individuals with high or low heterozygosity rates were removed for possible sample contamination or inbreeding. Individuals who deviated ±3 standard deviation from the samples' heterozygosity rate mean were removed. Fourth, SNPs with low minor allele frequency (MAF) were removed. SNPs with low MAF < 1% for cases or controls were eliminated. Rare SNPs lack power to detect SNP-trait associations and are also more prone to genotyping errors, thus should be removed. Fifth, SNPs which deviate from Hardy-Weinberg equilibrium with P-value < 10^{-4} were excluded. The Hardy-Weinberg equilibrium is a theorem that allele frequencies in a population will remain constant from generation to generation, in the absence of other evolutionary influences. Deviations from Hardy-Weinberg equilibrium may be a sign of non-random mating, inbreeding, or genotyping errors. Thus, SNPs which deviate from the Hardy-Weinberg equilibrium should be excluded for GWAS. Sixth, individuals with cryptic

relatedness were excluded. Cryptic relatedness in the sample can interfere with the association analysis. We excluded the subjects with pi-hat above 0.2 (i.e., second degree relatives). Lastly, we performed cluster QC for the SNPs with P-value under 10^{-4} (Figure 1). For each marker, the genotype (AA, Aa, or aa) was estimated by the sample signals. The genotype AA, Aa, and aa were colored as red, purple, and blue. The cluster QC was performed for both the case and control separately. If the genotypes were clearly clustered into the three colors, QC was passed. If there are genotyping errors, the three genotypes become not clearly separated like in Figure 1B. All markers with $P < 10^{-4}$ were visually inspected and those with such features were excluded.

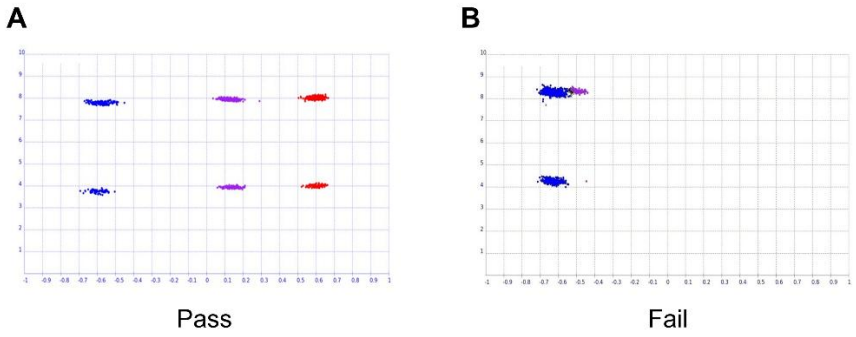

Figure 1. An example of the cluster quality control.

In general, the three genotypes, denoted as the blue, purple, and red dots, are clearly clustered (A). If the three genotypes were not clearly separated like in (B), the markers were excluded.

Statistical analysis

We performed primary analysis between cases and controls by multiple logistic additive models. Two models were analyzed, one adjusted for the age at sample and sex, and the other for age at onset of PD for cases (and age at sample for controls) and sex. We performed a separate candidate gene analysis with the same QC procedure but without visual inspection

for cluster QC, focusing on the SNPs in the genes that were identified from previous GWAS on PD. The genes included 59 genes around the 41 PD loci identified in the largest European GWAS at the time of analysis.[14] The list of genes is as follows: *GBA, SYT11, ITPKB, SLC45A3, NUCKS1, RAB7L1, SLC41A1, PM20D1, SIPA1L2, ASH1L, STK39, SCN3A, IL1R2, MCCC1, MISCH, STAB1, ITIH3, DNAH1, NCKIPSD, BAP1, PHF87, TLR9, SNCA, TMEM175, CAMK2D, ANK2, AM47E, BST1, ELOVL7, HLA-DQB1, ZNF184, GPNMB, FGF20, C8orf58, SORBS3, CTSB, BIN3, PDLIM2, SH3GL2, INPP5F, FAM171A1, DLG2, LRRK2, CCDC62, GCH1, GALC, TMEM229B, LRP10, VPS13C, TOX3, COQ7, MAPT, PSMC3IP, ATP6V0A1, TUBG2, RIT2, TMPRSS9,* and *DDRGK1*.

Secondary analyses, the sex-specific GWAS, were performed in male population and female population separately, using the same model and covariates as the primary analysis. PLINK software (version 1.90) was used for the association analysis. Quantile-quantile plots and Manhattan plots were plotted using the R software (version 3.5.2, R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/). Regional association plots were generated using the LocusZoom software (version 0.4.8). Power calculations were performed using Quanto software (version 1.2.4.). Regional association plots were generated and were visually inspected in all SNPs with $P < 10^{-4}$. Conservative Bonferroni corrections were applied to correct multiple testing, because of the large number of tests performed.

Results

Case-Control GWAS: Demographics and clinical characteristics

A total of 1,070 patients were initially recruited in the study. Sample of 20 patients among the 1,070 were excluded for low sample quality. The remaining 1,050 patients were included as the case group for the GWAS. The baseline characteristics of these patients are shown in Table 1. Their mean age at sample was 64.0 ± 9.7 (mean ± standard deviation) years, ranging from 31 to 89 years. Their age at onset of Parkinsonism symptom was 58.7 ± 10.2 years, ranging from 28 to 87 years. Among them, 554 were female patients. Their mean disease duration at study enroll was 5.3 ± 4.4 years. Their mean education years was 8.6 ± 4.4 years. Mini-mental status examination (MMSE) score, which is a 30-point questionnaire which is used to screen cognitive impairment. Their average MMSE score was 26.0 ± 3.5, ranging from 10 to 30. Power calculation showed that our sample had 80% power to detect variants with odds ratio (OR) of 1.25 with an allele frequency of 10% (Table 2).

Table 1. Baseline demographics of the case-control analyses

Characteristics	Cases (n = 1,050)	Control (n = 5,000)	P value
Female	554 (53)	2610 (52)	0.740
Age at sample, years	64.0 ± 9.7	64.0 ± 10.0	1.000
Age at onset of Parkinsonism symptoms, years	58.7 ± 10.2	-	
Disease duration at sample, years	5.3 ± 4.4	-	
Education years, years	8.6 ± 6.0	-	
MMSE	26.0 ± 3.5	-	
Disease duration from PD onset to MMSE, years	5.2 ± 4.1	-	

Data are presented as mean ± standard deviation or number of patients (%).

Abbreviations: PD, Parkinson's disease; MMSE, mini-mental status examination.

Table 2. Power calculation of the case-control analyses

Minor Allele Frequency	Least OR to satisfy statistical power of 80%
0.05	1.34
0.10	1.25
0.15	1.20
0.20	1.18
0.25	1.18
0.30	1.17
0.35	1.17
0.40	1.16
0.45	1.16
0.50	1.16

Abbreviation: OR, odd's ratio.

Case-Control GWAS: Age at onset as the covariate

In the case-control analysis which used age at onset of Parkinsonism and sex as the two covariates, 492,970 SNPs passed the marker QC. The quantile-quantile plot of the case-control analysis did not show significant inflation from the diagonal identity (Figure 2). The diagonal identity depicts the null distribution that no SNP is associated with the trait. Deviation from the null distribution suggest that there is a bias in population stratification or the uneven distribution of genotyped regions. Thus, the quantile-quantile plot in the current analysis suggests homogeneous population structure of our sample.

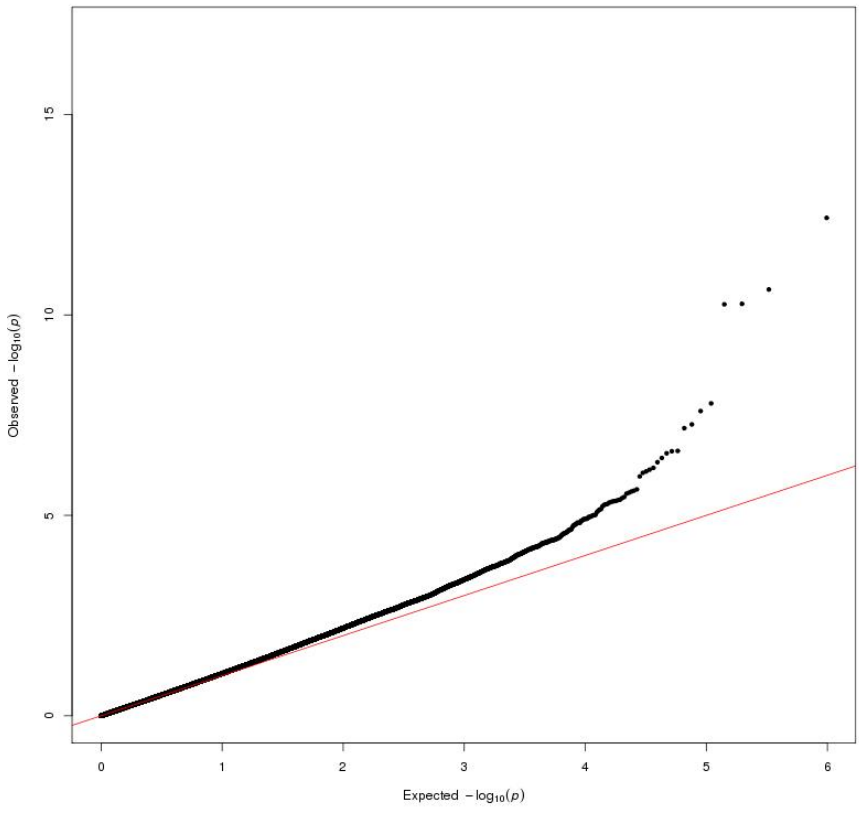

Figure 2. Quantile-Quantile plot of the case-control GWAS of age at onset and sex as the covariates

Of the 492,970 SNPs which passed marker QC on this analyses, 8 SNPs surpassed the strict Bonferroni-corrected genome-wide significance, the threshold being 1.01×10^{-7} (0.05 / 492,970) (Table 3, Figure 3). Four SNPs within or near the *SNCA* locus were the most significant among the 8 SNPs. The most significant was rs3796661 ($P = 3.79 \times 10^{-13}$), followed by rs356203 ($P = 2.32 \times 10^{-11}$), rs11931074 with ($P = 5.29 \times 10^{-11}$), and rs12640100 ($P = 5.45 \times 10^{-11}$). Two SNPs in the *SLC41A1* gene, which is located within the *PARK16* locus, were also genome-wide significant, including rs708726 ($P = 1.61 \times 10^{-18}$), and rs947211 ($P = 2.50 \times 10^{-8}$). The other two SNPs were rs2451713 of *ZNF322/GUSBP2* gene ($P = 5.39 \times 10^{-8}$), and rs708723 of *RAB29* gene which is located within the *PARK16* locus ($P = 6.69 \times 10^{-8}$) (Table 4).

Nine SNPs surpassed the *P* threshold of 1.0×10^{-6}, but not to the extent of Bonferroni correction (Table 5). These included 3 SNPs around *POM121L2* gene (rs61736085, rs9295732, and 6919033), 3 additional SNPs from *SNCA* gene (rs17016196, rs356204, and rs2736990), one from *PM20D1* gene (rs954206), and one from the *NUCKS1/SLC45A3* gene (rs12748933). The latter two SNPs are located in the *PARK16* locus Twenty-six more SNPs surpassed the *P* threshold of 1.0×10^{-5}, and these SNPs are listed in Table 6.

Table 3. The number of SNPs according to the level of genome-wide significance in the case-control analysis of age at onset and sex as the covariate

P value	Number of SNPs	Cumulative Number
Bonferroni Pass ($< 1.01 \times 10^{-7}$)	8	8
$1.01 \times 10^{-7} \leq P < 1.0 \times 10^{-6}$	9	17
$1.0 \times 10^{-6} \leq P < 1.0 \times 10^{-5}$	26	43
$1.0 \times 10^{-5} \leq P < 1.0 \times 10^{-4}$	139	182
$1.0 \times 10^{-4} \leq P < 1.0 \times 10^{-3}$	771	953
$1.0 \times 10^{-3} \leq P < 0.01$	6,165	7,118
$0.01 \leq P < 0.05$	22,096	29,214
≥ 0.05	463,756	492,970

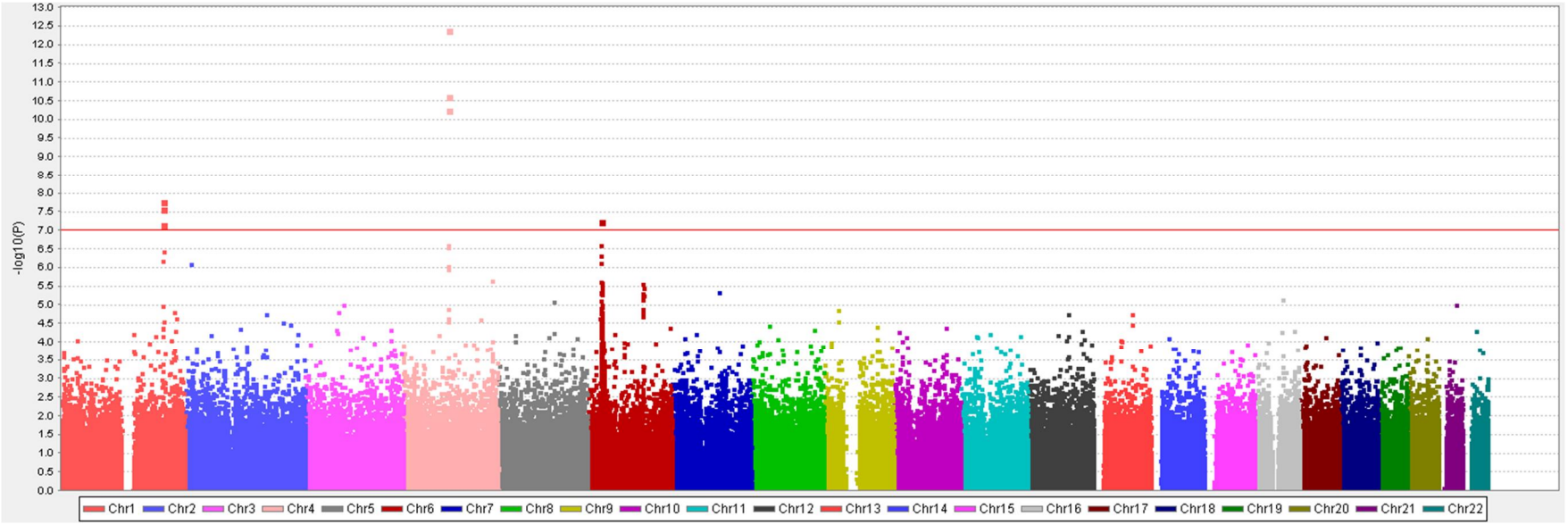

Figure 3. Manhattan plot the case-control GWAS of age at onset and sex as the covariates.

The red line indicates the Bonferroni threshold.

Table 4. Genomic variants with genome-wide significance after Bonferroni correction of the case-control GWAS of age at onset and sex as the covariates

Chr	SNP number	Gene	Region	Allele (minor/ major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	P
4	rs3796661	*SNCA*	intron	C/T	0.69 (0.62–0.76)	0.37/0.46	3.79×10^{-13}
4	rs356203	*SNCA*	intron	T/C	0.71 (0.64–0.79)	0.39/0.46	2.32×10^{-11}
4	rs11931074	*SNCA, GPRIN3*	intron, downstream, upstream	G/T	0.72 (0.65–0.79)	0.39/0.47	5.29×10^{-11}
4	rs12640100	*SNCA, GPRIN3*	intron, downstream, upstream	G/A	0.72 (0.65–0.79)	0.39/0.47	5.45×10^{-11}
1	rs708726	*SLC41A1*	intron	T/G	0.75 (0.68–0.83)	0.43/0.50	1.61×10^{-8}
1	rs947211	*SLC41A1, RAB29*	downstream, upstream	A/G	0.75 (0.68-0.83)	0.43/0.50	2.50×10^{-8}
6	rs2451713	*ZNF322, LOC101929855, GUSBP2*	upstream, downstream	C/G	1.88 (1.50–2.36)	0.05/0.04	5.39×10^{-8}
1	rs708723	*RAB29*	UTR-3	C/T	0.76 (0.69–0.84)	0.90/0.33	6.69×10^{-8}

Table 5. Genomic variants with *P* value between 1.01×10^{-7} and 1.0×10^{-6} of the case-control GWAS of age at onset and sex as the covariates

Chr	SNP number	Gene	Region	Allele (minor/ major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	P
6	rs61736085	*POM121L2*	missense, intron	A/G	1.77 (1.42–2.19)	0.06/0.04	2.47×10^{-7}
4	rs17016196	*SNCA*	intron	C/T	1.31 (1.18–1.45)	0.39/0.33	2.51×10^{-7}
4	rs356204	*SNCA*	intron	C/T	0.77 (0.69–0.85)	0.35/0.41	2.83×10^{-7}
1	rs954206	*PM20D1*	intron	C/T	0.77 (0.7–0.85)	0.45/0.51	3.70×10^{-7}
6	rs9295732	*PRSS16, POM121L2*	downstream, upstream	C/T	1.74 (1.4–2.15)	0.06/0.04	4.71×10^{-7}
1	rs12748961	*NUCKS1, SLC45A3*	downstream, upstream	T/C	0.78 (0.71–0.86)	0.44/0.50	6.53×10^{-7}
6	rs6919033	*HIST1H2AH*	UTR-5	T/C	1.77 (1.41–2.21)	0.06/0.03	7.26×10^{-7}
2	rs7563844	*LOC100506474*	intron	G/A	1.28 (1.16–1.42)	0.54/0.49	8.05×10^{-7}
4	rs2736990	*SNCA*	intron	A/G	0.77 (0.7–0.86)	0.35/0.41	8.71×10^{-7}

Table 6. Genomic variants with *P* value between 1.01×10^{-6} and 1.0×10^{-5} of the case-control GWAS of age at onset of Parkinsonism and sex as the covariates.

Chr	SNP number	Gene	Region	Allele (minor/ major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	P
4	rs2119787	*SNCA*	intron	G/A	0.77 (0.69–0.85)	0.3/0.36	1.07×10^{-6}
4	rs73865898	*LINC01098, LINC00290*	upstream,downstream	T/G	1.28 (1.16–1.42)	0.39/0.34	2.25×10^{-6}
6	rs16894368	*MIR3143, PRSS16*	upstream,downstream	A/G	1.73 (1.38–2.17)	0.06/0.03	2.42×10^{-6}
6	rs16894986	*MAS1L, LINC01015*	intron,upstream	G/C	1.38 (1.21–1.58)	0.17/0.13	2.54×10^{-6}
6	rs9374291	*FYN*	intron	C/T	0.76 (0.68–0.85)	0.24/0.28	2.72×10^{-6}
6	rs16894996	*LINC01015, MAS1L*	upstream,downstream	A/G	1.38 (1.21–1.58)	0.17/0.13	2.88×10^{-6}
6	rs74717803	*MARCKS, LINC01268*	downstream	A/C	1.40 (1.21–1.61)	0.15/0.12	3.48×10^{-6}
6	rs1362077	*OR2H1, MAS1L*	downstream,upstream	T/C	1.37 (1.20–1.57)	0.17/0.13	3.73×10^{-6}
6	rs61730668	*MAS1L*	missense	A/G	1.37 (1.20–1.57)	0.17/0.13	4.08×10^{-6}
6	rs7759855	*PGBD1, ZSCAN31*	downstream	G/A	1.44 (1.23–1.69)	0.12/0.09	4.16×10^{-6}
6	rs1362076	*OR2H1, MAS1L*	downstream,upstream	T/G	1.37 (1.20–1.57)	0.17/0.13	4.37×10^{-6}
7	rs11763156	*DYNC1I1*	intron	G/C	2.04 (1.50–2.76)	0.03/0.02	4.42×10^{-6}
6	rs9467704	*HIST1H4I, BTN3A2, HIST1H4H*	upstream,downstream	T/C	1.65 (1.33–2.04)	0.06/0.04	4.53×10^{-6}
6	rs7752077	*FYN*	intron	T/C	0.76 (0.68–0.86)	0.24/0.28	4.68×10^{-6}
6	rs72500814	*BTN3A1, BTN2A2*	upstream,downstream	T/G	1.45 (1.24–1.71)	0.11/0.08	4.87×10^{-6}
6	rs2523943	*HCG9*	exon	G/T	1.30 (1.16–1.45)	0.26/0.22	5.25×10^{-6}
6	rs12664800	*FLJ34503, LINC01268*	upstream	A/G	1.49 (1.26–1.77)	0.1/0.07	5.30×10^{-6}
6	rs6919306	*FYN*	intron,exon	T/C	0.77 (0.68–0.86)	0.24/0.28	5.51×10^{-6}

6	rs9481192	*FYN*	intron	T/C	0.77 (0.68–0.86)	0.24/0.28	5.95×10^{-6}
6	rs1409839	*FYN*	intron	C/T	0.77 (0.69–0.86)	0.24/0.28	6.98×10^{-6}
16	rs117332104	*MT4, BBS2*	upstream	C/T	2.38 (1.63–3.48)	0.02/0.01	7.16×10^{-6}
6	rs149494377	*HIST1H2AI, HIST1H2AK*	missense	T/G	1.43 (1.22–1.68)	0.12/0.09	7.66×10^{-6}
5	rs79724263	*APC*	intron	G/T	1.86 (1.41–2.43)	0.04/0.02	8.31×10^{-6}
21	rs142918131	*LOC101928107*	intron	C/T	2.16 (1.53–3.03)	0.03/0.01	9.67×10^{-6}
3	rs117894153	*ROBO2*	intron	G/A	1.60 (1.30–1.97)	0.06/0.04	9.83×10^{-6}
6	rs259942	*ZNRD1-AS1*	intron	T/C	1.28 (1.15–1.44)	0.27/0.23	9.97×10^{-6}

Associatations of SNPs with PD near the most significant SNP, rs3796661, within the *SNCA* gene is depicted in figure 4. The index SNP rs3796661 is shown in the purple filled diamond diamond. The locus contained several SNPs in LD with the leading SNP rs3796661, including 3 SNPs with $r^2 > 0.8$, 1 with $0.6 < r^2 < 0.8$, and 5 others with $0.4 < r^2 < 0.6$. The *PARK16* locus with the second leading genome-wide significant variants under Bonferroni correction around the SNP rs708726 is depicted in Figure 5. A number of SNPs that are in LD with the leading SNP rs708726 is observable within the *PARK16* locus.

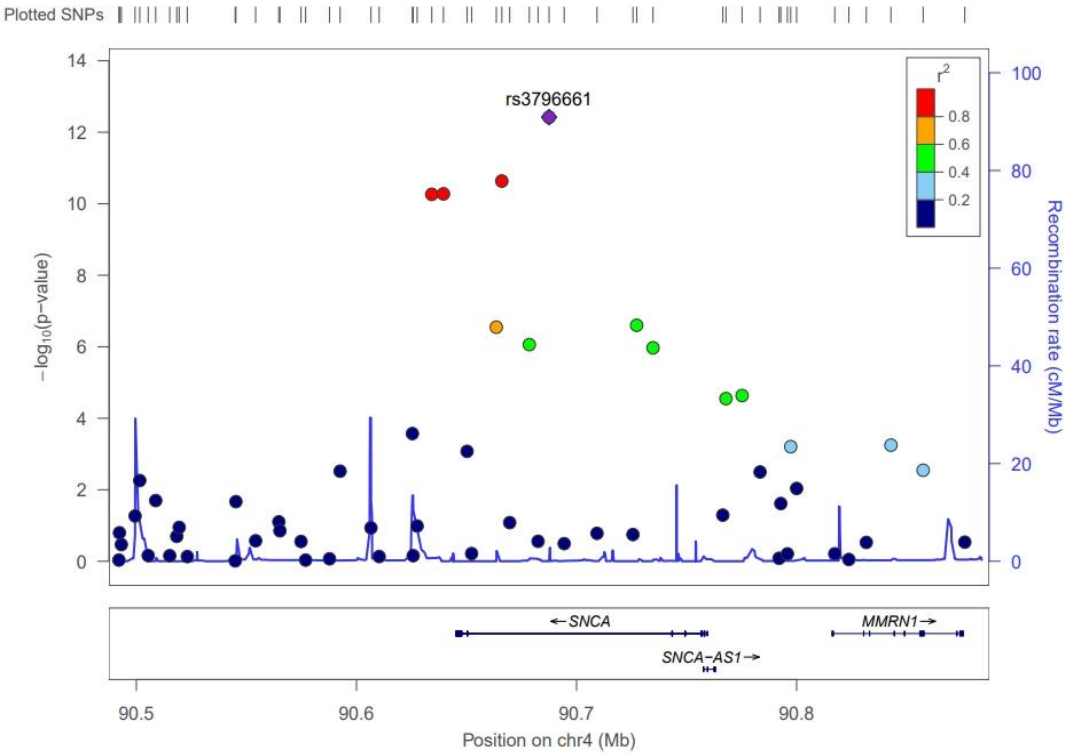

Figure 4. Regional association plot around rs3796661 in the case-control GWAS of age at onset and sex as the covariates.

The top purple diamond is the leading SNP, rs3796661. The other SNPs are colored according to their degree of linkage disequilibrium (r^2) with rs3796661.

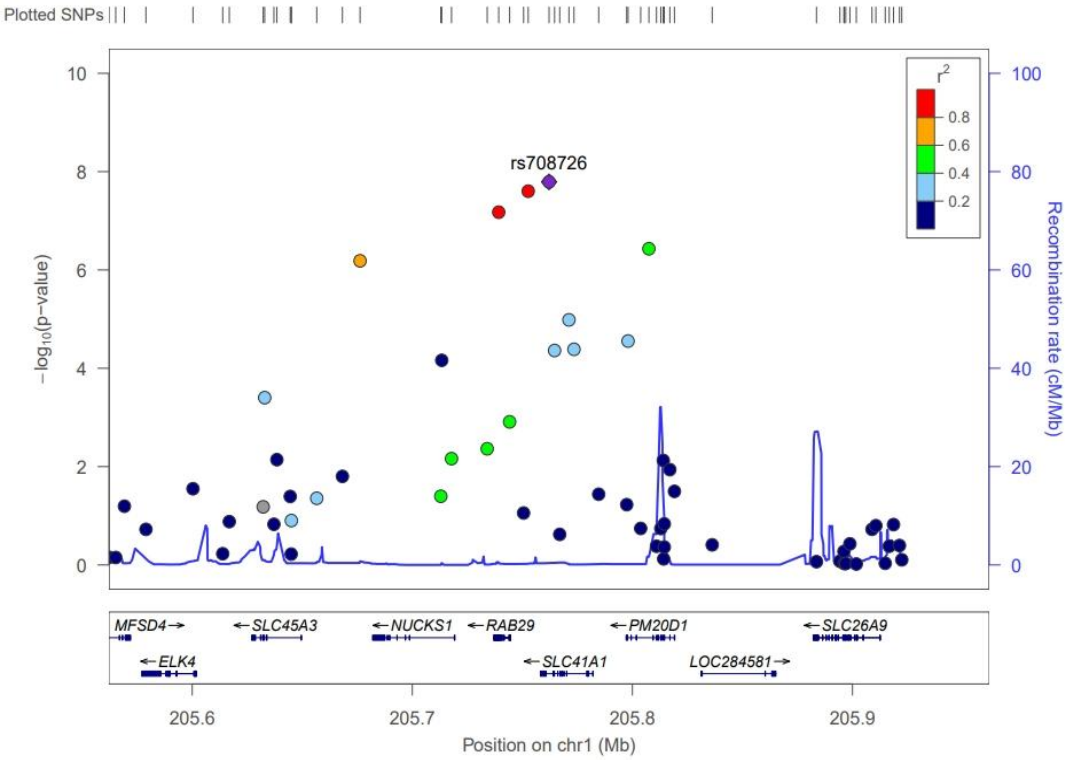

Figure 5 Regional association plot around SNP rs708726 in the case-control GWAS of age at onset and sex as the covariates.

The top purple diamond is the leading SNP, rs708726. The other SNPs are colored according to their degree of linkage disequilibrium (r^2) with the leading SNP.

Candidate gene analysis

In previous GWAS on the susceptibility of PD, there have been 41 PD loci identified associated with PD at the time of 2017. We further examined 1260 SNPs located in 59 genes located near these previously identified loci. Twenty-four SNPs surpassed Bonferroni correction adjusted significance ($P < 3.89 \times 10^{-5}$ (0.05 / 1,260)). The SNPs that did not appear priorly in the case-control analysis are shown in Table 7. Notably, rs34778348 at the *LRRK2*, which is a well-recognized causal gene for familial PD and risk gene for sporadic PD in Caucasians, showed a very strong association ($P = 4.77 \times 10^{-13}$). This variant was an exonal missense variant (G2385R). However, there was no further SNPs with statistical significance in LD with the variant found. There was also rs3754413 at *ITPKB* locus ($P = 1.05 \times 10^{-5}$), which was also a missense variant, but without nearby significant SNPs in LD. Rs823128 at *NUCKS1* gene located in the *PARK16* locus ($P = 1.53 \times 10^{-5}$) was also significant.

Table 7. SNPs that surpassed Bonferroni correction in the candidate gene analysis additional to the original genome-wide analysis

Chr	SNP number	Gene	Region	Allele (minor/ major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	P
12	rs34778348	*LRRK2*	missense, exon	A/G	2.56 (1.99–3.31)	0.05/0.02	4.77×10^{-13}
11	rs1022308	*DLG2*	intron	C/T	1.28 (1.14–1.43)	0.31/0.25	1.05×10^{-5}
1	rs3754413	*ITPKB*	missense	T/C	1.42 (1.22–1.68	0.12/0.09	1.73×10^{-5}
1	rs823085	*PM20D1*	intron	G/A	0.76 (0.67–0.86)	0.17/0.21	2.79×10^{-5}

Case-Control GWAS: Age at sample as the covariate

We performed another case-control analysis using age at sampling of DNA and sex as the covariates. A total of 493,000 SNPs passed the marker QC in this analysis. The quantile-quantile plot of the case-control analysis of age at sample and sex as covariates also did not show significant inflation from the null hypothesis (Figure 6).

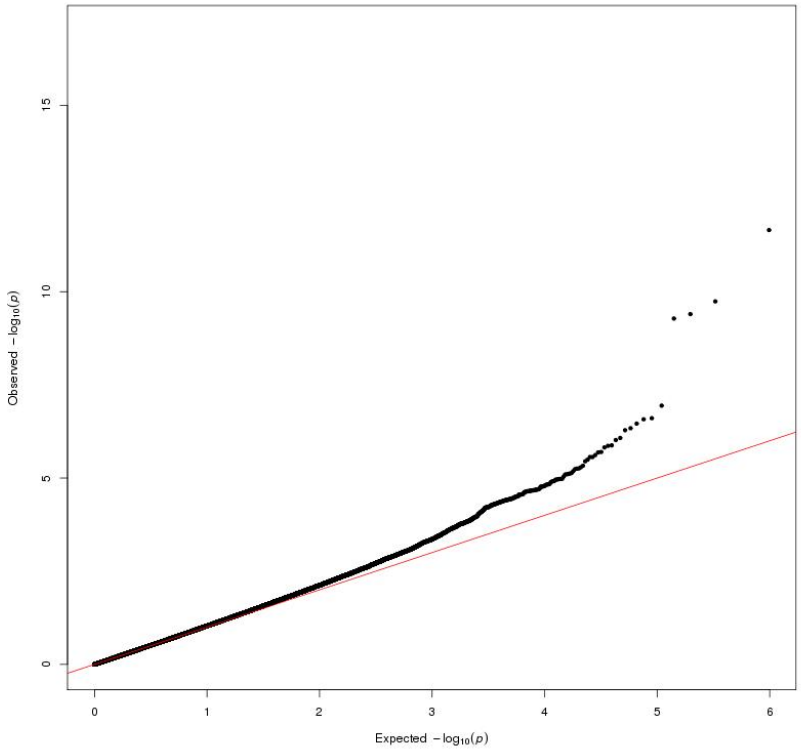

Figure 6. Quantile-Quantile plot of the case-control GWAS of age at sample and sex as the covariates.

Of the 493,000 SNPs which passed marker QC on this analysis, only 4 SNPs surpassed genome-wide significance threshold under Bonferroni correction ($P < 1.01 \times 10^{-7}$ (0.05/493,000) (Table 8, Figure 7). All 4 SNPs were at the *SNCA* locus, the most significant being rs3796661 ($P = 2.21 \times 10^{-12}$) followed by rs356203 ($P = 1.82 \times 10^{-10}$) rs126401100 ($P = 3.99 \times 10^{-10}$), and rs11931074 ($P = 3.99 \times 10^{-10}$) (Table 9). These SNPs were also the most significant ones in the prior analysis of age at onset as the covariate. Eight SNPs surpassed the P threshold of 1.0×10^{-6}, but not to the extent of Bonferroni correction. These included 3 SNPs from chromosome 6 around *POM121L2* locus (rs2451713, rs61736085, and rs9295732), 2 SNPs from *SLC41A1* gene (rs708726 and rs947211), and 2 more from *SNCA* locus (rs17016196 and rs356204). Twenty-two more SNPs surpassed the P threshold of 1.0×10^{-5}. These SNPs under P threshold of 1.0×10^{-5} but not surpassing Bonferroni correction are summarized in Table 10. The regional association plot showing the associations of SNPs at the *SNCA* and *SLC41A1* locus are shown in figure 9 and 10 respectively, the most significant SNPs within the locus as the top purple diamond.

Table 8. The number of SNPs according to the level of genome-wide significance in the case-control analysis of age at sample and sex as the covariate

***P* value**	**Number of SNPs**	**Cumulative Number**
Bonferroni Pass ($< 1.01 \times 10^{-7}$)	4	4
$1.01 \times 10^{-7} \leq P < 1.0 \times 10^{-6}$	8	12
$1.0 \times 10^{-6} \leq P < 1.0 \times 10^{-5}$	22	34
$1.0 \times 10^{-5} \leq P < 1.0 \times 10^{-4}$	160	194
$1.0 \times 10^{-4} \leq P < 1.0 \times 10^{-3}$	714	908
$1.0 \times 10^{-3} \leq P < 0.01$	5,424	6,332
$0.01 \leq P < 0.05$	21,202	27,534
≥ 0.05	465,466	493,000

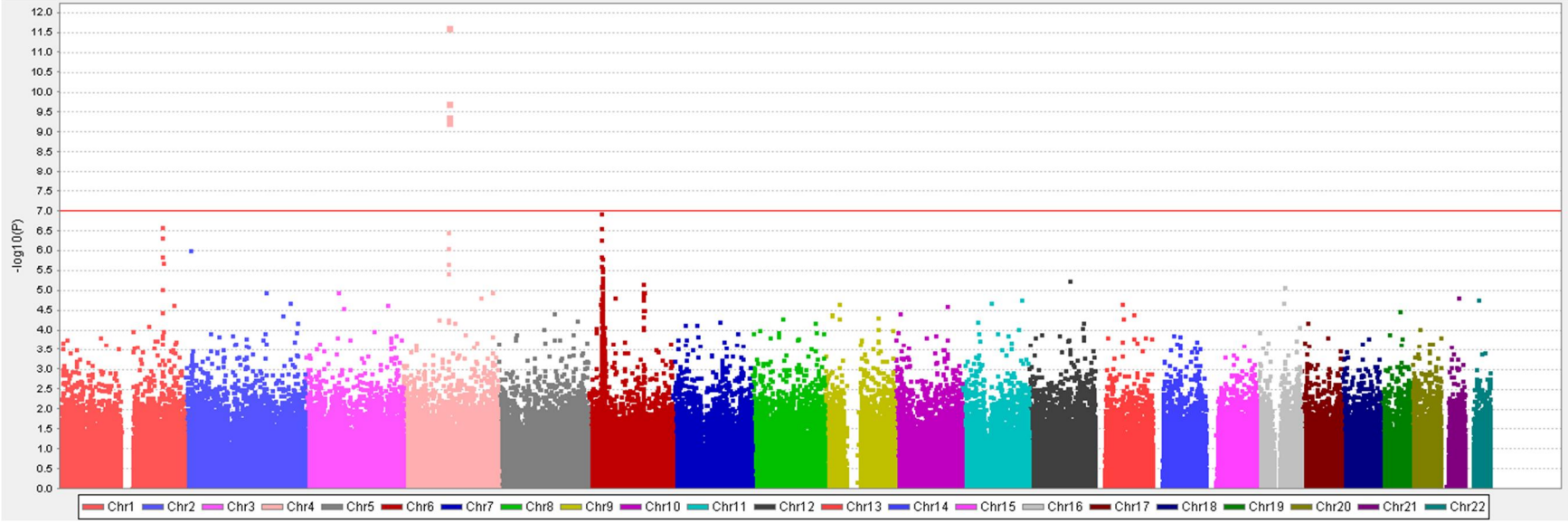

Figure 7. Manhattan plot the case-control GWAS of age at sample and sex as the covariates.

The red line indicates the Bonferroni threshold.

Table 9. Genomic variants with genome-wide significance after Bonferroni correction of the case-control GWAS of age at sample and sex as the covariates

Chr	SNP number	Gene	Region	Allele (minor/ major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	P
4	rs3796661	*SNCA*	intron	C/T	0.67 (0.60–0.75)	0.37/0.46	2.21×10^{-12}
4	rs356203	*SNCA*	intron	T/C	0.78 (0.63–0.78)	0.39/0.46	1.82×10^{-10}
4	rs12640100	*SNCA, GPRIN3*	intron, downstream, upstream	G/T	0.71 (0.63–0.79)	0.39/0.47	3.99×10^{-10}
4	rs11931074	*SNCA, GPRIN3*	intron, downstream, upstream	G/A	0.71 (0.64–0.79)	0.39/0.47	5.24×10^{-10}

Table 10. Genomic variants with significance between 1.01×10^{-7} and 1.0×10^{-5} of the case-control GWAS of age at sample and sex as the covariates

Chr	SNP number	Gene	Region	Allele (minor/ major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	P
6	rs2451713	*ZNF322, LOC101929855, GUSBP2*	upstream, downstream	C/G	1.99 (1.54–2.56)	0.06/0.03	1.14×10^{-7}
1	rs708726	*SLC41A1*	intron	T/G	0.75 (0.67–0.84)	0.43/0.50	2.48×10^{-7}
6	rs61736085	*POM121L2*	missense, intron	A/G	1.88 (1.48–2.4)	0.06/0.04	2.65×10^{-7}
4	rs17016196	*SNCA*	intron	C/T	1.34 (1.20–1.50)	0.39/0.33	3.46×10^{-7}
1	rs947211	*SLC41A1, RAB29*	downstream, upstream	A/G	0.76 (0.68–0.84)	0.43/0.50	4.60×10^{-7}
6	rs9295732	*PRSS16, POM121L2*	downstream, upstream	C/T	1.84 (1.45–2.34)	0.06/0.04	5.21×10^{-7}
4	rs356204	*SNCA*	intron	C/T	0.76 (0.68–0.85)	0.35/0.41	8.39×10^{-7}
2	rs7563844	*LOC100506474*	intron	G/A	1.31 (1.18–1.46)	0.54/0.49	9.53×10^{-7}
6	rs6919033	*HIST1H2AH*	UTR-5	T/C	1.85 (1.44–2.38)	0.06/0.03	1.33×10^{-6}
1	rs708723	*RAB29*	UTR-3	C/T	0.77 (0.69–0.85)	0.43/0.49	1.37×10^{-6}
6	rs2523943	*HCG9*	exon	G/T	1.36 (1.20–1.53)	0.26/0.22	1.51×10^{-6}
1	rs954206	*PM20D1*	intron	C/T	0.77 (0.69–0.86)	0.45/0.51	2.00×10^{-6}
4	rs2736990	*SNCA*	intron	A/G	0.76 (0.68–0.85)	0.35/0.41	2.05×10^{-6}
6	rs16894368	*MIR3143, PRSS16*	upstream, downstream	A/G	1.84 (1.43–2.37)	0.06/0.03	2.42×10^{-6}
6	rs259942	*ZNRD1-AS1*	intron	T/C	1.34 (1.18–1.51)	0.27/0.23	2.73×10^{-6}
6	rs16894996	*LINC01015, MAS1L*	upstream, downstream	A/G	1.43 (1.23–1.65)	0.17/0.13	2.73×10^{-6}
6	rs7759855	*PGBD1, ZSCAN31*	downstream	G/A	1.50 (1.27–1.78)	0.12/0.09	3.21×10^{-6}

4	rs2119787	*SNCA*	intron	G/A	0.76 (0.68–0.85)	0.3/0.36	3.56×10^{-6}
6	rs16894986	*MAS1L, LINC01015*	intron, upstream	G/C	1.41 (1.22–1.63)	0.17/0.13	4.68×10^{-6}
6	rs1362077	*OR2H1, AS1L*	downstream, upstream	T/C	1.41 (1.22–1.63)	0.17/0.13	5.13×10^{-6}
6	rs61730668	*MAS1L*	missense	A/G	1.41 (1.21–1.63)	0.17/0.13	5.54×10^{-6}
12	rs138619186	*TMTC2*	intron	G/A	2.20 (1.57–3.09)	0.03/0.02	5.59×10^{-6}
6	rs1362076	*OR2H1, AS1L*	downstream, upstream	T/G	1.41 (1.21–1.63)	0.17/0.13	5.77×10^{-6}
6	rs9374291	*FYN*	intron	C/T	0.75 (0.66–0.85)	0.24/0.28	6.55×10^{-6}
6	rs149494377	*HIST1H2AI, HIST1H2AK*	missense	T/G	1.49 (1.25–1.77)	0.12/0.09	7.36×10^{-6}
6	rs1150722	*PGBD1*	intron	G/A	1.48 (1.25–1.76)	0.12/0.09	7.58×10^{-6}
6	rs12665039	*HLA-A, HCG9*	intron, downstream, upstream	C/T	1.33 (1.18–1.51)	0.25/0.21	7.83×10^{-6}
6	rs72500814	*BTN3A1, BTN2A2*	upstream, downstream	T/G	1.49 (1.25–1.78)	0.11/0.08	7.87×10^{-6}
16	rs117332104	*MT4, BBS2*	upstream	C/T	2.62 (1.71–3.99)	0.02/0.01	8.15×10^{-6}
1	rs12748961	*NUCKS1, SLC45A3*	downstream, upstream	T/C	0.78 (0.70–0.87)	0.44/0.50	9.19×10^{-6}

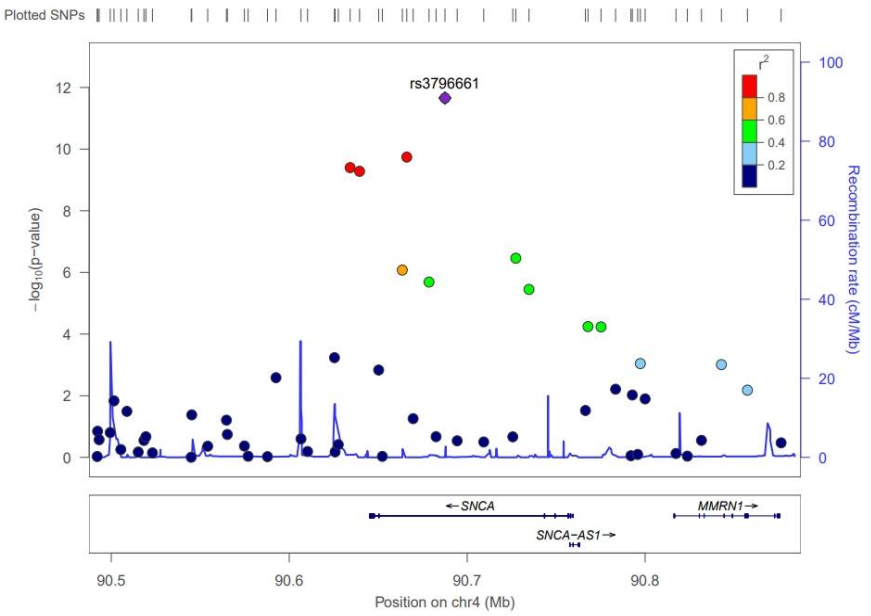

Figure 8. Regional association plot around rs3796661 in the case-control GWAS of age at onset and sex as the covariates.

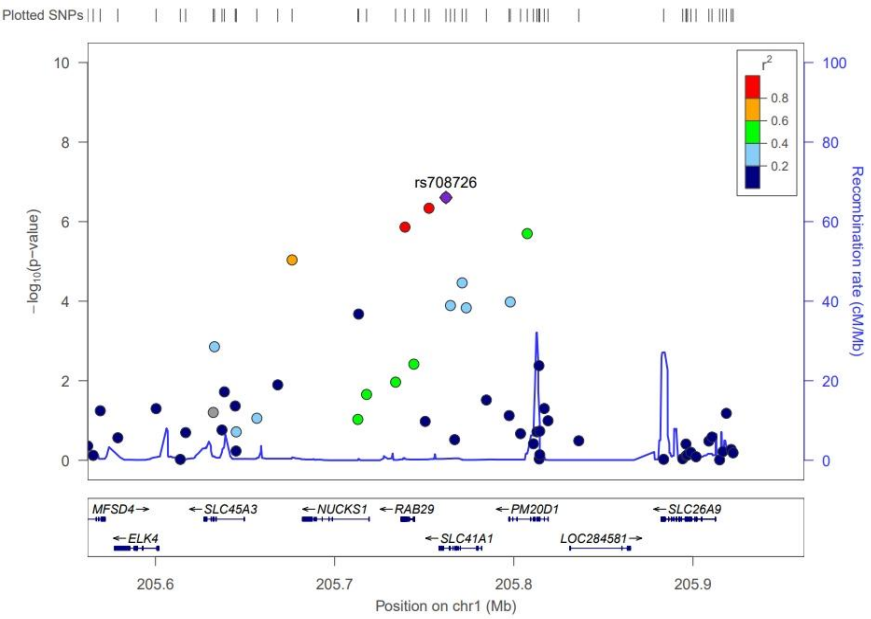

Figure 9. Regional association plot around rs708726 in the case-control GWAS of age at onset and sex as the covariates.

Sex-specific GWAS: Female-only analysis

To determine the sex-specific difference in the genetic contribution to sporadic PD, we performed female-only and male-only case-control analysis. In female-only analysis, 554 female PD cases and 2,610 controls were included. The demographics of female patients are shown in Table 11. Power calculation showed that the female-only analysis had 80% power to detect variants with OR of 1.34 with an allele frequency of 10% (Table 12). In the female-only analysis, 486,510 SNPs passed the marker QC. The quantile-quantile plot did not show inflation from the null hypothesis (Figure 10).

Table 11. Demographics of the female cases and controls

Characteristics	Female Cases (n = 554)	Female Control (n = 2,610)	P value
Age at sample, years	64.0 ± 9.2	64.0 ± 9.0	1.000
Age at onset of Parkinsonism symptoms, years	58.6 ± 9.8	-	-
Disease duration at sample, years	5.4 ± 4.8	-	-
Education years, years	7.2 ± 5.5	-	-
MMSE	25.5 ± 3.7	-	-
Disease duration from PD onset to MMSE, years	5.4 ± 4.4	-	-

Table 12. Power calculation of the female-only analysis

Minor Allele Frequency	Least OR to satisfy statistical power of 80%
0.05	1.47
0.10	1.34
0.15	1.28
0.20	1.26
0.25	1.24
0.30	1.23
0.35	1.22
0.40	1.21
0.45	1.21
0.50	1.21

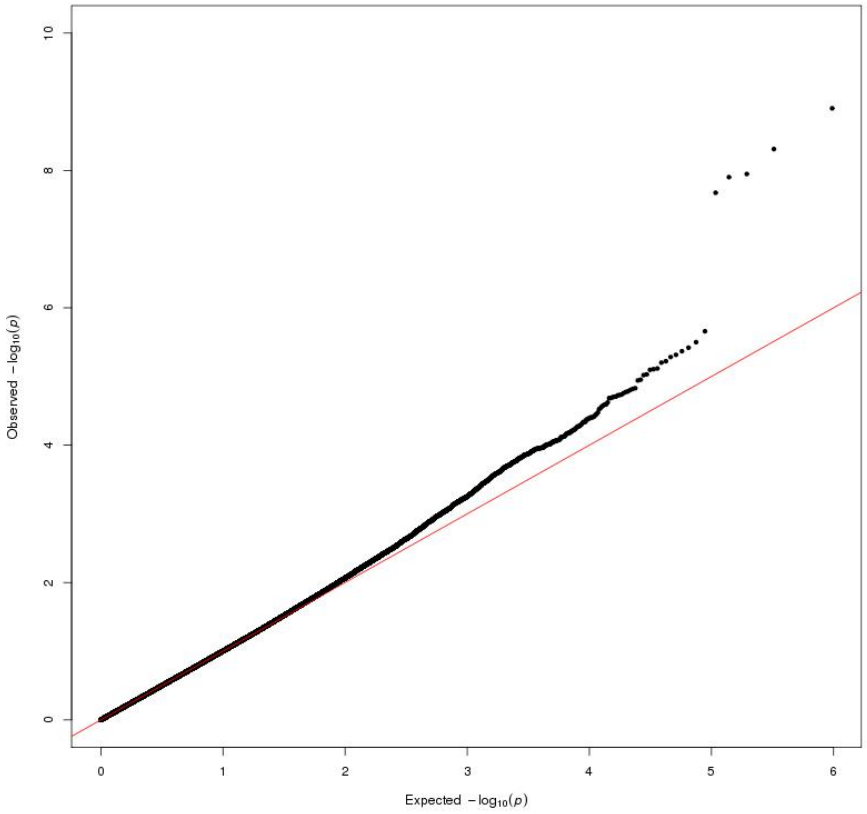

Figure 10. Quantile-Quantile Plot of the female-only analysis.

Of the 486,510 SNPs which passed marker QC on female-only analysis, 5 SNPs surpassed genome-wide significance threshold under Bonferroni correction ($P < 1.03 \times 10^{-7}$ (0.05 / 486,510) (Table 13, Figure 11). The most significant SNP was rs34778348 in *LRRK2* locus, which was a missense SNP placed in exon ($P = 1.25 \times 10^{-9}$) (Table 14). The other 4 SNPs which surpassed Bonferroni threshold were all in the *SNCA* locus, the most significant being rs3796661 ($P = 4.89 \times 10^{-9}$), followed by rs126401100 ($P = 1.83 \times 10^{-8}$), rs356203 ($P = 1.25 \times 10^{-8}$), and rs11931074 ($P = 2.11 \times 10^{-8}$). No other SNPs surpassed the *P* threshold of 1.0×10^{-6}, but 13 SNPs surpassed the *P* threshold of 1.0×10^{-5}. These SNPs under *P* threshold of 1.0×10^{-5} but not surpassing Bonferroni correction are summarized in Table 10. None of the variants in the *SLC41A1* gene had significance under $P < 1.0 \times 10^{-5}$ in this female-only analysis.

The regional association plot showing the associations of SNPs at the *SNCA* locus are shown in figure 12. The regional association plot of the rs34778348 in *LRRK2* locus was not created, because there was nearby significant SNPs in LD with the variant.

Table 13. The number of SNPs according to the level of genome-wide significance in the female-only analysis

***P* value**	**Number of SNPs**	**Cumulative Number**
Bonferroni Pass ($< 1.01 \times 10^{-7}$)	5	5
$1.01 \times 10^{-7} \leq P < 1.0 \times 10^{-6}$	0	5
$1.0 \times 10^{-6} \leq P < 1.0 \times 10^{-5}$	13	18
$1.0 \times 10^{-5} \leq P < 1.0 \times 10^{-4}$	92	110
$1.0 \times 10^{-4} \leq P < 1.0 \times 10^{-3}$	790	800
$1.0 \times 10^{-3} \leq P < 0.01$	4,783	5,583
$0.01 \leq P < 0.05$	19,180	24,763
≥ 0.05	461,747	486,510

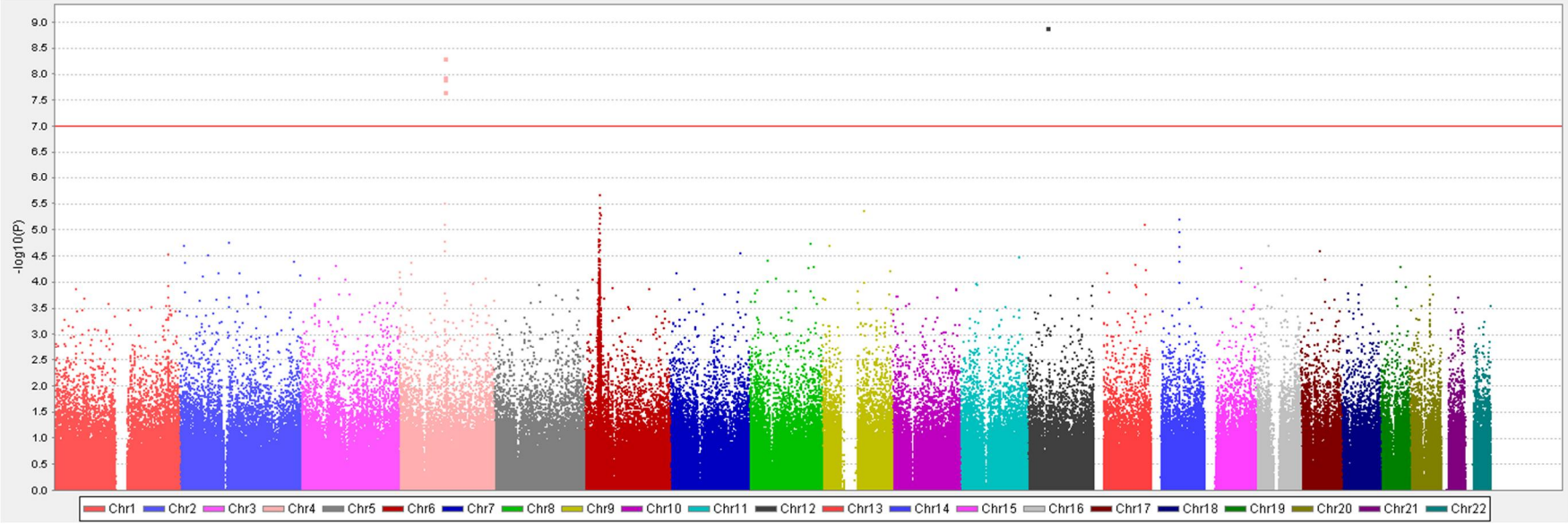

Figure 11. Manhattan plot of female-only analysis.

The red line indicates the Bonferroni threshold.

Table 14. Genomic variants with genome-wide significance after Bonferroni correction of the female-only GWAS

Chr	SNP number	Gene	Region	Allele (minor/ major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	*P*
12	rs34778348	*LRRK2*	missense,exon	A/G	3.53 (2.35–5.29)	0.05/0.02	1.25×10^{-9}
4	rs3796661	*SNCA*	intron	C/T	0.63 (0.54–0.74)	0.37/0.46	4.89×10^{-9}
4	rs12640100	*SNCA,GPRIN3*	intron,downstream,upstream	G/A	0.64 (0.55–0.75)	0.38/0.47	1.13×10^{-8}
4	rs356203	*SNCA*	intron	T/C	0.64 (0.55–0.75)	0.38/0.47	1.25×10^{-8}
4	rs11931074	*SNCA,GPRIN3*	intron,downstream,upstream	G/T	0.65 (0.56–0.75)	0.38/0.47	2.11×10^{-8}

Table 15. Genomic variants with P value under 1.0×10^{-5} of the female-only GWAS

Chr	SNP number	Gene	Region	Allele (minor/major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	*P*
6	rs2523943	*HCG9*	exon	C/G	1.99 (1.54–2.56)	0.06/0.03	2.20×10^{-6}
4	rs17016196	*SNCA*	intron	T/G	0.75 (0.67–0.84)	0.43/0.50	3.16×10^{-6}
6	rs12665039	*HLA-A,HCG9*	intron,downstream,upstream	A/G	1.88 (1.48–2.4)	0.06/0.04	3.81×10^{-6}
9	rs1889065	*LOC101927477*	upstream,downstream	C/T	1.34 (1.20–1.50)	0.39/0.33	4.29×10^{-6}
6	rs2523961	*HLA-A,HCG9*	exon,downstream,upstream	A/G	0.76 (0.68–0.84)	0.43/0.50	4.84×10^{-6}
6	rs241429	*TAP2*	intron	C/T	1.84 (1.45–2.34)	0.06/0.04	5.21×10^{-6}
6	rs2256902	*HCG9,HLA-A*	upstream,downstream	C/T	0.76 (0.68–0.85)	0.35/0.41	6.00×10^{-6}
14	rs58737648	*DLGAP5*	intron	G/A	1.31 (1.18–1.46)	0.54/0.49	6.27×10^{-6}
6	rs76748682	*HLA-A,HCG9*	downstream,upstream	T/C	1.85 (1.44–2.38)	0.06/0.03	7.65×10^{-6}
13	rs9554699	*PCCA*	intron	C/T	0.77 (0.69–0.85)	0.43/0.49	7.83×10^{-6}
4	rs356204	*SNCA*	intron	G/T	1.36 (1.20–1.53)	0.26/0.22	8.00×10^{-6}
6	rs16894986	*MAS1L,LINC01015*	intron,upstream	C/T	0.77 (0.69–0.86)	0.45/0.51	9.34×10^{-6}
6	rs16894996	*LINC01015,MAS1L*	upstream,downstream	A/G	0.76 (0.68–0.85)	0.35/0.41	9.55×10^{-6}

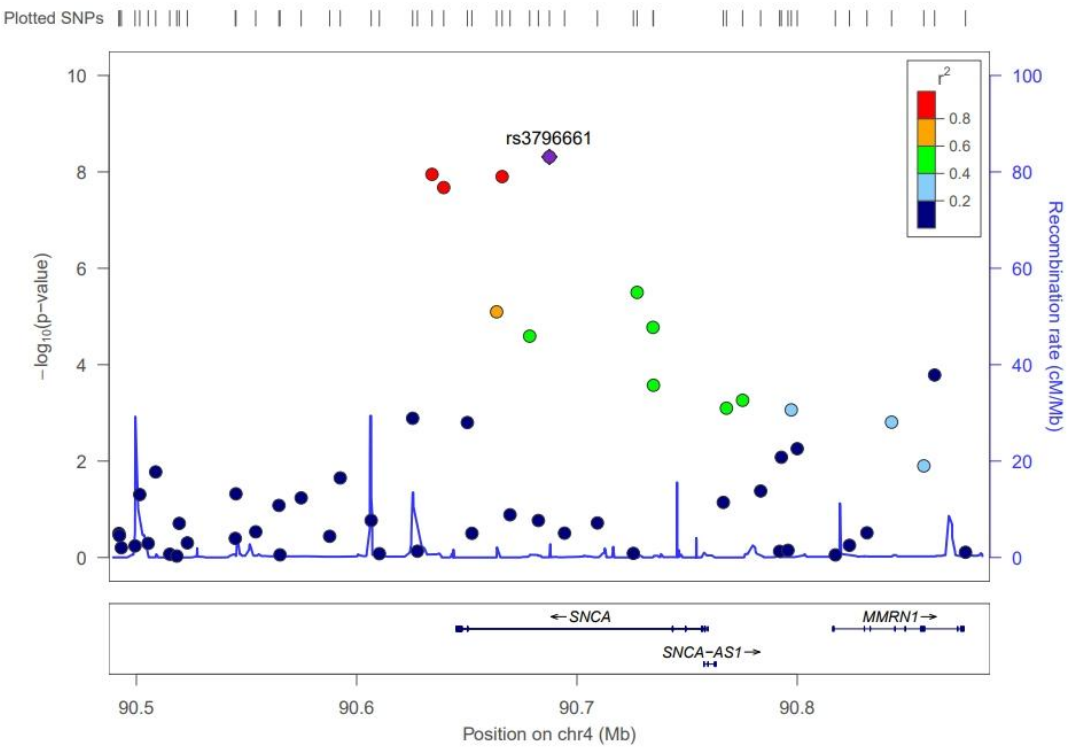

Figure 12. Regional association plot around the rs379661 of the female only analysis.

Sex-specific GWAS: Male-only analysis

Male-only analysis was performed as the same manner in female-only analysis. A total of 496 male PD cases and 2,390 male controls were included in the analysis. The demographics of the male subjects are shown in Table 16. There was no significant difference in age at samples, as it was matched during the subject selection. Power calculation of the male-only analysis had 80% power to detect variants with OR of 1.36 with an allele frequency of 10% (Table 17). A total of 488,631 SNPs passed the marker QC. The quantile-quantile plot did not show significant inflation from the null hypothesis (Figure 14).

In the genome-wide analysis, none of the SNPs surpassed genome-wide significance threshold under Bonferroni correction ($P < 1.02 \times 10^{-7}$ (0.05 / 488,631) (Table 18, Figure 15). The most significant SNP was rs708726 in the *SLC41A1* gene ($P = 8.23 \times 10^{-6}$) (Table 19, Figure 16). Meanwhile, the most significant SNP within the SNCA locus was rs3796661 with *P* value of 5.25×10^{-5}, and there was no additional SNP within the *SNCA* gene with *P* value under 10^{-4}, indicating its small effect on male patients than on female patients.

Table 16. Demographics of the male subjects

Characteristics	Male Cases (n = 496)	Male Control (n = 2,390)	P value
Age at sample, years	64.0 ± 10.2	64.0 ± 9.0	1.000
Age at onset of Parkinsonism symptoms, years	58.8 ± 10.6	-	-
Disease duration at sample, years	5.1 ± 4.0	-	-
Education years, years	7.2 ± 5.5	-	-
MMSE	26.5 ± 3.0	-	-
Disease duration from PD onset to MMSE, years	5.1 ± 3.9	-	-

Table 17. Power calculation of the male-only analysis

Minor Allele Frequency	Least OR to satisfy statistical power of 80%
0.05	1.50
0.10	1.36
0.15	1.30
0.20	1.27
0.25	1.26
0.30	1.24
0.35	1.24
0.40	1.23
0.45	1.23
0.50	1.23

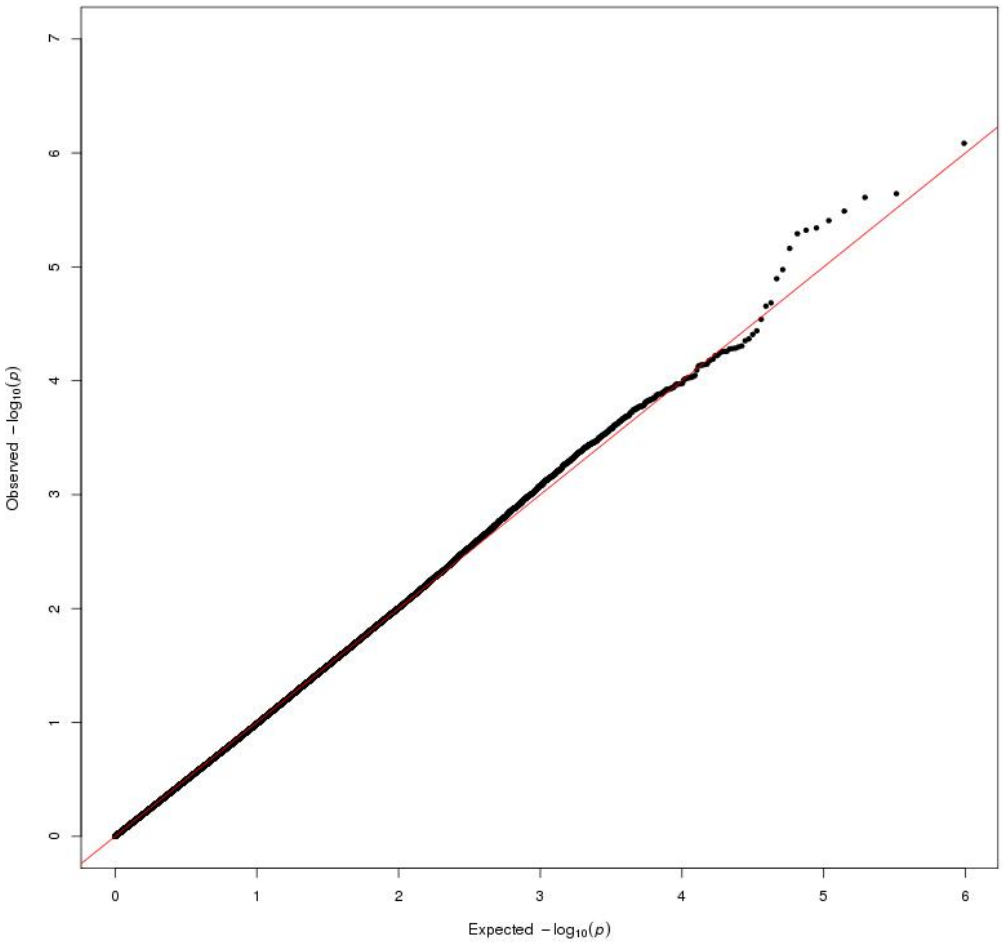

Figure 13. Quantile-Quantile Plot of the male-only analysis.

Table 18. The number of SNPs according to the level of genome-wide significance in the male-only analysis

P value	Number of SNPs	Cumulative Number
Bonferroni Pass ($< 1.01 \times 10^{-7}$)	0	0
$1.01 \times 10^{-7} \leq P < 1.0 \times 10^{-6}$	0	0
$1.0 \times 10^{-6} \leq P < 1.0 \times 10^{-5}$	1	1
$1.0 \times 10^{-5} \leq P < 1.0 \times 10^{-4}$	8	9
$1.0 \times 10^{-4} \leq P < 1.0 \times 10^{-3}$	39	48
$1.0 \times 10^{-3} \leq P < 0.01$	525	573
$0.01 \leq P < 0.05$	4,363	4,936
≥ 0.05	19,204	24,140

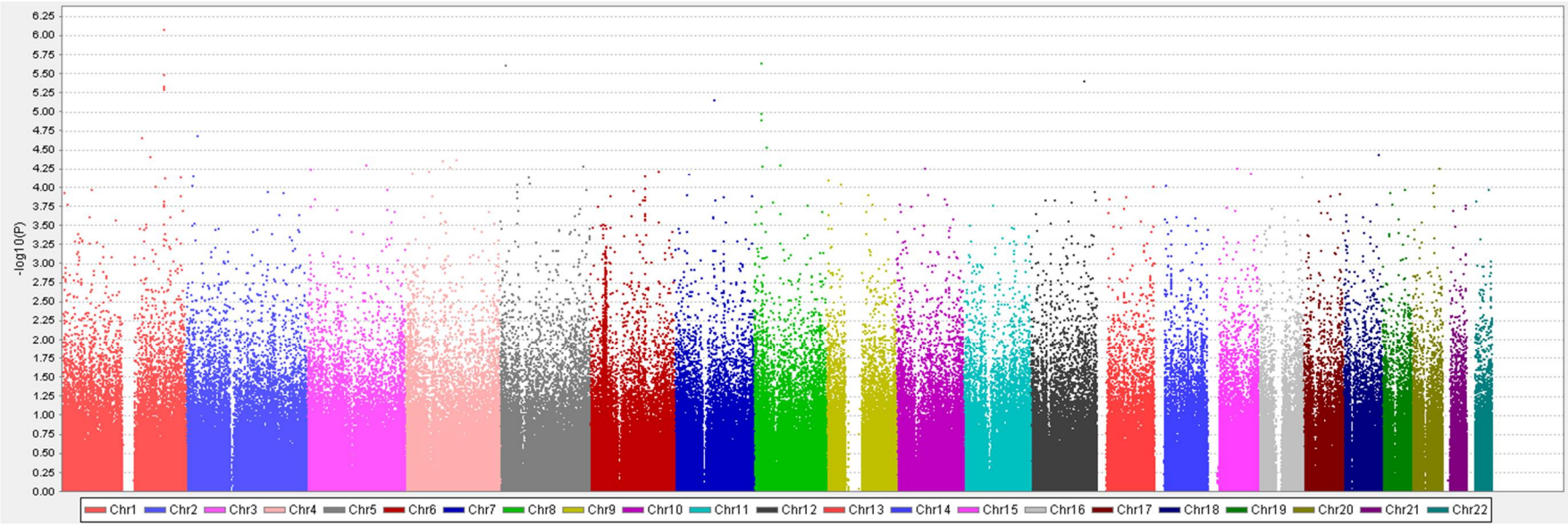

Figure 14. Manhattan plot of male-only analysis.

Table 19. Genomic variants with P value under 1.0×10^{-5} of the male-only GWAS

Chr	SNP number	Gene	Region	Allele (minor/major)	OR (95% CI)	Minor Allele Frequency (Case/Control)	P
1	rs708726	*SLC41A1*	intron	T/G	0.67 (0.57–0.79)	0.40/0.49	8.23×10^{-6}
8	NA	*FGF20,MSR1*	intron,downstream,upstream	G/T	1.52 (1.28–1.81)	0.32/0.25	2.28×10^{-6}
5	rs139422381	*ANKRD33B,LOC101929412*	downstream,upstream	A/G	4.02 (2.25–7.16)	0.03/0.01	2.46×10^{-6}
1	rs947211	*SLC41A1,RAB29*	downstream,upstream	A/G	0.69 (0.59–0.8)	0.41/0.50	3.24×10^{-6}
12	rs10746109	*WSCD2,LOC728739*	upstream	A/G	1.46 (1.24–1.71)	0.51/0.44	3.93×10^{-6}
1	rs708723	*RAB29*	UTR-3	C/T	0.69 (0.59–0.81)	0.40/0.49	4.56×10^{-6}
1	rs1775145	*SLC41A1,RAB29*	downstream,upstream	C/A	0.69 (0.59–0.81)	0.42/0.50	4.77×10^{-6}
1	rs12748961	*NUCKS1,SLC45A3*	downstream,upstream	T/C	0.69 (0.59–0.81)	0.42/0.50	5.12×10^{-6}
7	rs1949132	*GNAI1,LOC101927269*	intron,downstream	C/T	1.78 (1.39–2.29)	0.12/0.09	6.90×10^{-6}

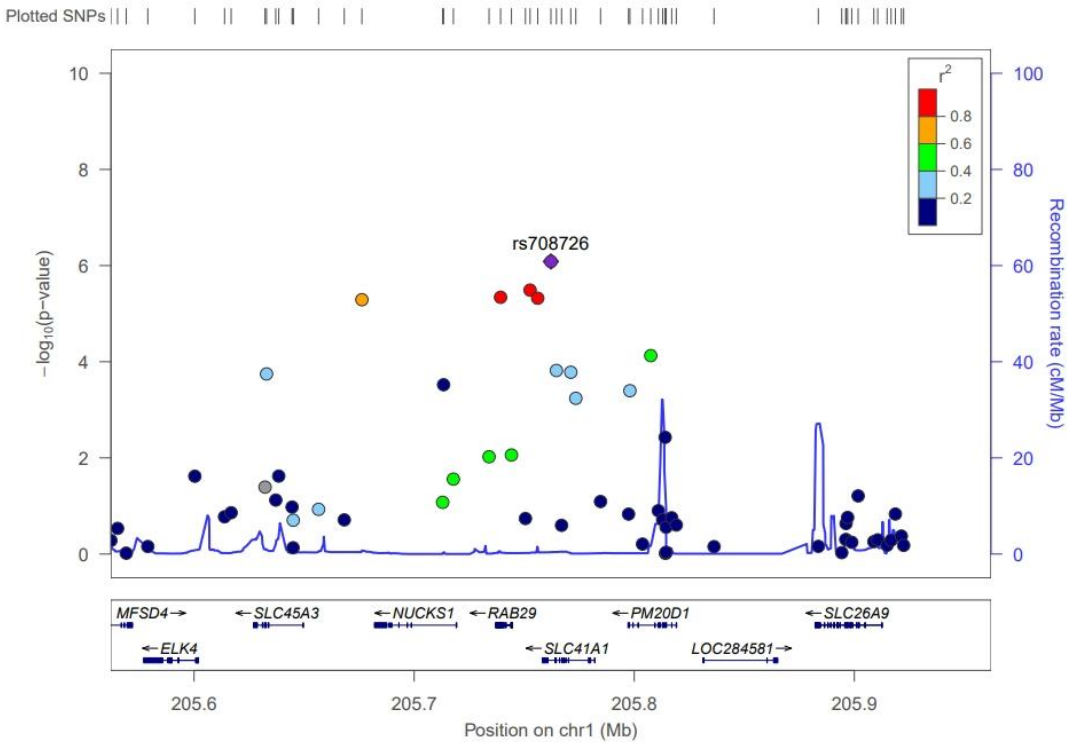

Figure 15. Regional association plot around the rs708726 in the male-only analysis.

Discussion

In this study, we investigated the ethnicity-specific, and sex-specific genomic variants associated with PD. In the case-control analysis, we observed strong genome-wide significant signals passing Bonferroni correction at the *SNCA* and *SLC41A1* loci. In female-only analysis, rs3796661 in the *SNCA* locus and rs34778348 in the *LRRK2* locus showed genome-wide significant associated with PD. In contrast, rs708726 from *SLC41A1* locus was the most significant association though not surpassing genome-wide significance under Bonferroni correction in male patients. There was no signal from the *SNCA* gene with *P* value under 10^{-5} in male patients, reflecting the sex-specific effect of genetic variants of *SNCA*.

Variants within the *SNCA* gene, the most representative one being rs3796661, were the most significant and most consistently associated variant in the various analyses performed in this current study. Alpha-synuclein is the major component of Lewy bodies, which is the pathological hallmark of PD.[15] It is a neuronal protein that physiologically regulates synaptic vesicle trafficking and the subsequent neurotransmitter release.[16] The mechanism how the α-synuclein results in dopaminergic neurodegeneration in PD remains to be established, but it is related to the modulation of dopamine homeostasis in synapses and its binding on dopamine transporter.[16] *SNCA* is the gene encoding for α-synuclein. The genetic link between PD and *SNCA* was first suggested in 1997 by Polymeropoulos et al., in a report of G209A mutation in the *SNCA* gene resulting in familial PD in large Italian kindred. The triplication of the *SNCA* locus in a separate family with PD with autosomal-dominant inheritance was subsequently reported, shedding light to the crucial role of *SNCA* on the occurrence of PD. The association of common variants of *SNCA* with PD was identified from the earliest large-scale PD GWAS. The subsequent meta-analyses consistently showed a very strong effect of common variants of *SNCA* on the risk of sporadic PD. Such strong effect was also replicated in one of the largest GWAS in East Asian. Our top SNP being in the *SNCA* locus demonstrates the universal strong

effect of variants of SNCA on the risk of PD across ethnicity.

Despite the small size of our study resulting in less power to detect genome-wide significance, multiple variants within the *PARK16* locus, including the SNP rs708726 in the *SLC41A1* gene, showed genome-wide significance in the case-control analyses. Minor alleles within the region were associated with reduced risk for PD. Variant in the *PARK16* locus was one of the first reported loci associated with PD in the earliest GWAS in Japan and was consistently replicated in the following European studies. The *PARK16* locus spans in five genes, including *SLC45A3, NUCKS1, RAB29/RAB7L1, SLC41A1* and *PM20D1*.[17] Among these regions, *RAB29* has gained the most highlight for its possible role in PD. *RAB29* is closely related to *LRRK2*.[18] The protein Rab29 is the master regulator of LRRK2 protein, controlling its activation, localization, and phosphorylation.[19] *SLC41A1 i*s a Mg^{2+} transporter, that may role in magnesium homeostasis.[20] The precise mechanism of how the variants within this region effect susceptibility to PD is still vague. In the large European GWAS, *PARK16* was a consistently replicated risk region but was not in the top rank.[4, 14] In contrast, *PARK16* was always the one of the most significant risk regions in Asians, including Japanese and Han Chinese.[17] Koreans are genetically closely related to these two populations. Our result adds evidence to the stronger effect of the *PARK16* locus particularly in East Asian population.

In this study, the strong association of rs34778348 with PD in *LRRK2* was another notable finding of the case-control analyses. The variant was found significant only in the candidate gene analyses in our study, but not in the original case-control GWAS because the variant failed to pass the cluster QC. Thus, a careful interpretation of the association is needed. The SNP rs34637584, known as the *LRRK2* G2019S variant, is one of the most well-known variants that is strongly associated with PD risk in Caucasian and Jewish population. In perfect contrast to this, rs34778348, which is the *LRRK2* G2385R variant, is only found in Asian populations. This variant was found to be a genetic risk factor for sporadic PD in Chinese,

Japanese, and in Korean.[21-23] Among these, the previous study on Korean population included only a small number of subjects. Our study is the replication of the result in a large size.[21] The *LRRK2* G2019S variant results in the kinase overactivity. Consequently, downregulating the *LRRK2* function with kinase inhibitor has been suggested as a potential therapeutic target of PD.[24] It has been proposed that the *LRRK2* G2385R results in partial loss-of-function of the kinase activity in vitro,[25] in opposite to the G2019S counterpart. The discrepancy in *LRRK2* variants and subsequent protein dysfunction between Caucasian and Asian poses great importance, warranting different therapeutic approach according to the ethnicity.

We could not observe any evidence for association across other previously known PD risk loci from European GWAS, such as *GBA* locus and *MAPT* locus. Signals within the two loci were the strongest signal along with those in the SNCA locus in previous large European GWAS.[4, 14] However, signals within these two loci did not show any meaningful significance in our case-control analysis. One possible explanation is the small power of our study. Otherwise, SNPs within the loci could have been excluded during QC. *MAPT* is genetically highly homogenous in East Asian populations with only H1 haplotype, whereas European heritage has both H1 and H2 haplotype. However, multiple variants exist even in H1 haplotype, reflecting greater diversity of *MAPT* than explained only by H1 and H2 clades.[26] Thus, the lack of the association of *MAPT* or *GBA* in this Korean specific analysis suggest the difference in the susceptibility to PD by the variants within these genes. The two loci were not replicated other Asian PD GWAS, including Japanese and Han-Chinese, supporting the idea.

There have been several clinical features of PD with sex differences. For example, PD is 1.5 times more prevalent in men than women worldwide. Men shows earlier age at onset of disease than women.[8] It was hypothesized that the sex-difference in PD is determined by the sex-hormonal influence on nigrostriatal dopaminergic system. However, little has been investigated about the genetics of PD in sex-specific manner. In our analysis, rs34778348 of

LRRK2 locus and 4 SNPs of *SNCA* locus showed genome-wide significance in female, but the significance was not replicated in male. The most significant SNPs in male were those in the *PARK16* locus, though not surpassing Bonferroni correction. However, a recent investigation on autosomal genetic sex differences in PD found no significant genetic differences between male or female PD cases.[9] It is notable that PD is more prevalent in men worldwide but is more prevalent in women in Asian populations.[6] The discrepancy between the European and Korean sex-specific PD GWAS may implicate such ethnicity-specific difference in sex ratio of PD. On the other hand, in another study upon the transcriptional profile of PD patients, SNCA were exclusively down-regulated in male PD patients under stringent statistical significance level, though not in more relaxed level. Genes of the *PARK16* were not analyzed in that study.[27] Although their results were not in perfect line with our study, it implicates the sex-specific transcriptional profile of nigral dopaminergic neurons in PD. Thus, our study may implicate the not only the gender-genetic interaction for PD, but also the gender-ethnic-genetic interaction for the development of PD. However, genetically determined sex difference in PD is an under-investigated area. Further investigation focusing on the genetic sex difference of PD, including X-chromosome wide association study, should be encouraged.

In our case-control analyses, we adopted two ways of adjusting the age of patients; age at onset of Parkinsonism and age at sample. The overall results were similar, but the SNPs of *PARK16* locus surpassed the Bonferroni corrected genome-wide significance threshold in the analysis adjusting age at onset of Parkinsonism, but barely failed in the analysis adjusting age at sample. Such subtle difference could have resulted from the small sample size of the current study, compared to other large size GWAS recruiting hundreds of thousands of subjects. Because PD is a long-standing neurodegenerative disorder with various age at onset and progression rate, it would be ideal to adjust the age at a certain common time-point. The most universally accessible would be the age at onset of disease. However, there are several hurdles fort this. First, the self-report of the patients' age at onset of Parkinsonism symptom may be

inaccurate and is difficult to collect. Also, matching the healthy controls' age at the age at onset of Parkinsonism may result in recruiting more younger controls, thus encompass bigger chance of including pre-symptomatic potential PD as healthy controls. For such reasons, previous PD GWAS has mostly adopted simply the age at sample as the adjustment covariates.[4, 14, 28] There are such strengths and weaknesses in adjusting either age at Parkinsonism or age at sample for PD GWAS, and we performed both methods to supplement each other.

There are several limitations in our study. First, our study lacks a separate replication analysis. GWAS is typically composed of 2-stage design, composed to the initial discovery study and a replication study in independent samples to sort out the true-positive associations from the many possible false-positive associations. However, there have been about 5 major meta-analysis of PD GWAS since the start of GWAS era.[4, 5, 28, 29] In a way, our study itself could be a version of replication study in Korean population. Second, our study lacks functional validation of the discovered variants. GWAS identifies a genomic location associated with the disease but does not identify in what biological consequences the variant causes the disease. Methods for functional validation includes SNP enrichment, single-cell technologies, such as transcriptomics and chromatin accessibility, colocalization of disease variants with quantitative trait loci (QTL), and dissecting the effects of the variants with genome-editing technologies.[30] However, the primary aim of our study was to demonstrate the difference in the genetic contribution between Korean population and the others, rather than to discover a new variants associated with PD. As we did not identify any new variants from new loci associated with PD, functional validation was not essential. Still, whether the variants identified in our study affect the susceptibility to PD differently between male and female needs further functional validations. These techniques may include sex-specific expression QTL, sex-specific transcriptome, etc.

Conclusion

In this first Korean ethnicity-specific GWAS on the susceptibility of PD, we found the variants in the *SNCA* and *PARK16* are strongly associated with Korean PD. However, we could not identify any association with variants in the *GBA* or *MAPT* gene, which are the strongest signals associated with European PD. We also observed strong association of variants in the *SNCA* locus in female PD patients but not in male patients. On the other hand, the strongest variants associated with male PD was in the *PARK16* locus. These findings suggest the disproportionate gradient in the genetic contribution for the susceptibility of PD across ethnicity and gender. Further functional validations specific to ethnicity and sex should be encouraged.

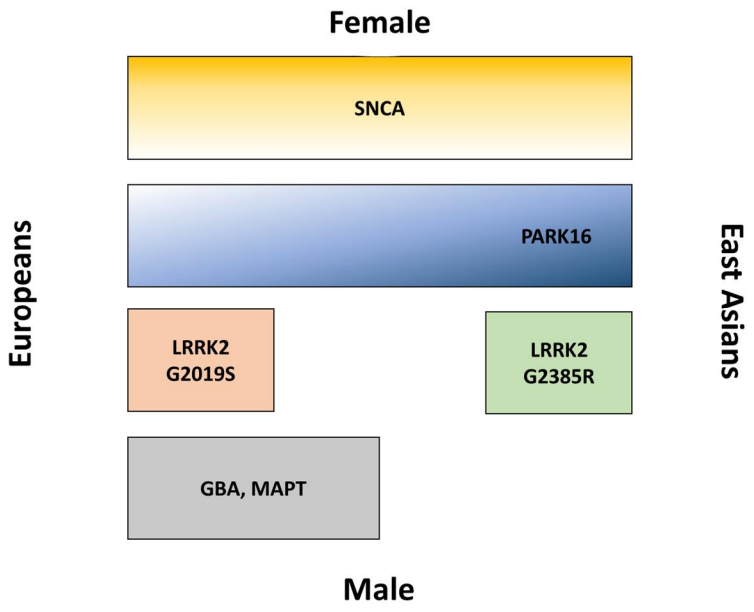

Figure 16. Graphical summary of the study.

The figure depicts the disproportionate gradient in the genetic contribution of known genetic loci for the susceptibility of PD across ethnicity and gender.

Reference

1. Poewe W, Seppi K, Tanner CM, et al. Parkinson disease. Nature Reviews Disease Primers 2017;3:17013.

2. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics 2019;20:467-484.

3. Day JO, Mullin S. The Genetics of Parkinson's Disease and Implications for Clinical Practice. Genes (Basel) 2021;12.

4. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. The Lancet Neurology 2019;18:1091-1102.

5. Foo JN, Tan LC, Irwan ID, et al. Genome-wide association study of Parkinson's disease in East Asians. Hum Mol Genet 2017;26:226-232.

6. Park J-H, Kim D-H, Kwon D-Y, et al. Trends in the incidence and prevalence of Parkinson’s disease in Korea: a nationwide, population-based study. BMC Geriatrics 2019;19:320.

7. Klein C, König IR. Exploring Uncharted Territory: Genetically Determined Sex Differences in Parkinson's Disease. Annals of Neurology 2021;90:15-18.

8. Gillies GE, Pienaar IS, Vohra S, Qamhawi Z. Sex differences in Parkinson's disease. Front Neuroendocrinol 2014;35:370-384.

9. Blauwendraat C, Iwaki H, Makarious MB, et al. Investigation of Autosomal Genetic Sex Differences in Parkinson's Disease. Annals of Neurology 2021;90:35-42.

10. Gibb W, Lees A. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. Journal of Neurology, Neurosurgery & Psychiatry 1988;51:745-752.

11. Moon S, Kim YJ, Han S, et al. The Korea Biobank Array: Design and Identification of Coding Variants Associated with Blood Biochemical Traits. Sci Rep 2019;9:1382.

12. Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008;299:1335-1344.

13. Korean Disease Control and Prevention Agency. Available at: http://www.kdca.go.kr/contents.es?mid=a30327000000. Accessed Nov 18, 2021.

14. Chang D, Nalls MA, Hallgrimsdottir IB, et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. Nat Genet 2017;49:1511-1516.

15. Siddiqui IJ, Pervaiz N, Abbasi AA. The Parkinson Disease gene SNCA: Evolutionary and structural insights with pathological implication. Sci Rep 2016;6:24475-24475.

16. Wong YC, Krainc D. α-synuclein toxicity in neurodegeneration: mechanism and therapeutic strategies. Nat Med 2017;23:1-13.

17. Tucci A, Nalls MA, Houlden H, et al. Genetic variability at the PARK16 locus. European Journal of Human Genetics 2010;18:1356-1359.

18. Gan-Or Z, Dion PA, Rouleau GA. Genetic perspective on the role of the autophagy-lysosome pathway in Parkinson disease. Autophagy 2015;11:1443-1457.

19. MacLeod DA, Rhinn H, Kuwahara T, et al. RAB7L1 interacts with LRRK2 to modify intraneuronal protein sorting and Parkinson's disease risk. Neuron 2013;77:425-439.

20. Arjona FJ, Latta F, Mohammed SG, et al. SLC41A1 is essential for magnesium homeostasis in vivo. Pflügers Archiv - European Journal of Physiology 2019;471:845-860.

21. Kim JM, Lee JY, Kim HJ, et al. The LRRK2 G2385R variant is a risk factor for sporadic Parkinson's disease in the Korean population. Parkinsonism Relat Disord 2010;16:85-88.

22. Tan EK, Peng R, Teo YY, et al. Multiple LRRK2 variants modulate risk of Parkinson disease: a Chinese multicenter study. Hum Mutat 2010;31:561-568.

23. Zabetian CP, Yamamoto M, Lopez AN, et al. LRRK2 mutations and risk variants in Japanese patients with Parkinson's disease. Mov Disord 2009;24:1034-1041.

24. Taymans JM, Greggio E. LRRK2 Kinase Inhibition as a Therapeutic Strategy for Parkinson's Disease, Where Do We Stand? Curr Neuropharmacol 2016;14:214-225.

25. Carrion MDP, Marsicano S, Daniele F, et al. The LRRK2 G2385R variant is a partial loss-of-function mutation that affects synaptic vesicle trafficking through altered protein interactions. Sci Rep 2017;7:5377.

26. Pittman AM, Myers AJ, Abou-Sleiman P, et al. Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration. Journal of Medical Genetics 2005;42:837-846.

27. Simunovic F, Yi M, Wang Y, Stephens R, Sonntag KC. Evidence for Gender-Specific Transcriptional Profiles of Nigral Dopamine Neurons in Parkinson Disease. PLOS ONE 2010;5:e8856.

28. Simon-Sanchez J, Schulte C, Bras JM, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet 2009;41:1308-1312.

29. Blauwendraat C, Nalls MA, Singleton AB. The genetic architecture of Parkinson's disease. The Lancet Neurology 2020;19:170-178.

30. Lichou F, Trynka G. Functional studies of GWAS variants are gaining momentum. Nat Commun 2020;11:6283.

국문요약

파킨슨병은 전세계적으로 가장 흔한 신경퇴행성질환 중 하나이다. 파킨슨병에 대한 전장유전체상관분석은 지난 십여년 동안 파킨슨병의 발생과 관련된 90여개의 위험 변이를 규명하였다. 하지만 과거의 파킨슨병에 대한 전장유전체상관분석은 절대 다수가 유럽인을 대상으로 한 연구였다. 또한, 파킨슨병이 성별에 따라 다른 임상적 특징을 보임에도 불구하고, 성별에 따른 파킨슨병의 유전적 차이는 주목받지 못한 분야였다. 본 연구는 파킨슨병에 대해 한국인 특이적, 성별 특이적인 전장유전체상관분석을 시행하였다.

총 1,050명의 한국인 파킨슨병 환자와 5,000명의 건강한 한국인을 대조군으로 비교하였다. 전체 환자-대조군 분석은 파킨슨병 발병 나이와 성별을 교란변수로, 또한 혈액채취 당시의 나이와 성별을 교란변수로 각각 분석을 시행하였으며, 로지스틱 첨가(additive) 모델을 사용하였다. 또한 기존에 파킨슨병과 연관된 것으로 알려진 유전자들로 구성된 후보 유전자 분석(candidate gene analysis)을 더 완화된 정도관리(quality control) 기준을 적용하여 진행하였다. 그리고 같은 모델을 이용하여 554명의 여자 환자와 2,610명의 여자 대조군, 그리고 496명의 남자 환자와 2,390명의 남자 환자로 성별 특이적 분석을 시행하였다.

파킨슨병 발병 나이와 성별을 교란 변수로 시행한 환자-대조군 분석에서는 총 492,970개의 단일염기다형성(single nucleotide polymorphism, SNP)이 마커 정도관리를 통과하였다. 그 중 8개의 SNP이 본페로니 교정을 거친 유의성 기준을 통과하였다 (P 값 < 1.01×10^{-7}). 여기에는 *SNCA* 유전자좌에 위치한 rs3796661, rs356203, rs11931074, rs12640100의 4개 SNP과 and *PARK16* 유전자좌에 위치한 rs708726, rs947211, rs708723의 3개 SNP이 포함되었다. 후보 유전자 분석에서는 추가적으로 *LRRK2* 유전자좌의 rs34778348 (G2385R) 변이가 한국인 파킨슨병과 연관이 있는 것으로 나타났다 (P = 4.77×10^{-13}). 혈액채취 당시의 나이와 성별을 교란변수로 적용한 환자-대조군 분석에서는 493,000개의 SNP이 마커 정도관리를 통과하였다. 이 중 *SNCA* 유전자좌에 위

치한 4개의 SNP이 본페로니 교정을 거친 유의성 기준을 통과하였다. 반면, 유럽인 파킨슨병 환자들과 연관성이 강하게 나타났던 *GBA−SYT11*나 *MAPT* 유전자좌에 위치한 변이들은 한국인에서는 연관성을 보이지 않았다. 여성만을 대상으로 한 분석에서는 486,510개의 SNP이 정도관리를 통과하였다. 총 5개의 SNP이 본페로니 교정을 거친 유의성을 통과하였는데, 이 중에는 *LRRK2* 유전자좌의 rs34778348가 포함되어 있었으며 이는 엑손에 위치한 점돌연변이에 해당한다. 나머지 4개는 SNCA 유전자좌에 위치한 SNP들이었다. 남성만을 대상으로 한 분석에서는 본페로니 교정 하의 유의성을 통과할 정도로 강한 연관성을 보이는 SNP는 없었다. 하지만 가장 유의했던 변이는 *PARK16* 유전자좌에 위치한 rs708726이었다 ($P = 8.23 \times 10^{-6}$). *SNCA* 유전자좌의 변이 중에서는 오직 rs3796661 만이 0.0001보다 작은 P 값을 보여, 이 변이에 의한 효과가 여성보다 남성에서 더 작음을 반영하였다.

본 연구는 한국인을 대상으로 파킨슨병 발병의 취약도를 전장유전체분석을 한 첫 연구로, *SNCA*와 *PARK16* 유전자좌의 SNP들이 한국인 파킨슨병과 연관이 깊음을 밝혔다. 유럽인 파킨슨병에서 강한 연관성을 보이는 *GBA−SYT11*나 *MAPT* 유전자좌의 SNP들은 한국인 파킨슨병과는 연관성을 보이지 않았다. 또한 우리는 *SNCA* 유전자좌의 SNP들이 남성보다 여성 환자에서 강한 연관을 보이는 반면 *PARK16* 유전자좌의 SNP들은 여성보다 남성 환자에서 더 강한 연관성을 보이는 점을 확인하였다. 이러한 결과들은 특발성 파킨슨병 발병에 있어 인종과 성별에 따라 그 유전적 기여도가 다름을 의미한다.