2024년 2월
석사학위 논문

Genomic association analysis of sleep-related disorders using large scale genome and questionnaire matching

조선대학교 대학원

글로벌바이오융합학과

윤 다 빈

Genomic association analysis of sleep-related disorders using large scale genome and questionnaire matching

대규모 유전체-설문 매칭을 이용한
수면장애 연관 유전체분석

2024년 2월 23일

조선대학교 대학원

글로벌바이오융합학과

윤 다 빈

Genomic association analysis of sleep-related disorders using large scale genome and questionnaire matching

지도교수 김 정 수

이 논문을 이학 석사학위 신청 논문으로 제출함

2023년 10월

조선대학교 대학원

글로벌바이오융합학과

윤 다 빈

윤다빈의 석사학위논문을 인준함

위원장 김 석 준 (인)

위　원 김 규 민 (인)

위　원 김 정 수 (인)

2023년 12월

조선대학교 대학원

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Genomic association analysis of sleep-related disorders using large scale genome and questionnaire matching

Dabin Yoon

Advisor: Prof. Jungsoo Gim, Ph.D.

Department of Integrative Biological Sciences

Graduate School of Chosun University

Obstructive sleep apnea (OSA) is a complex disease where genetic variants play a critical role as one of considerable risk factors. The previous genetic studies, however, have analyzed samples of limited ethnicity with underdiagnosis of the OSA, leading to misclassification of the disease due to incorrect case definition and underpower to identification of responsible loci. Here, we investigate a genetic burden of the OSA defined in a continuous scale using STOP-BANG questionnaire from a large Korean sample.

We performed a genome-wide association study on 25,712 Korean subjects, a genetically homogeneous population selected from the Korean Genome and Epidemiology Study (n=72,291). To measure the severity of OSA, we used a standard measure calculated from the questionnaire called STOP-BANG which is scaled from 0 to 8. We replicated our findings using the FinnGen GWAS summary statistics for sleep apnea. For further validation of the identified variants, we evaluated the polygenic risk score (PRS) for sleep-related risk assessment on the independent dataset which were not used in the discovery.

We identified 9 genome-wide significant loci (45 variants). Of these, 8 were not previously reported in association with OSA nor STOP-BANG, while 2 variants corresponding to 1 locus (MME), associated with Cerebellar Ataxia were replicated in the nominal significance level from the FinnGen study. We further observed the validity of our finding from the PRS for sleep-related traits evaluated with the independent dataset.

Our study uncovered multiple genetic loci associated with OSA. We identified 1 locus at the MME gene associated with sleep apnea, and 8 loci as candidates about OSA in East Asian.

I. INTRODUCTION

I-1. Background of obstructive sleep apnea (OSA)

Obstructive sleep apnea (OSA) is sleep disorder characterized by repetitive respiratory arrest and intermittent decrease in oxygen saturation during sleep due to anatomically narrow airways or obstruction of the upper airway during sleep [1]. Although the cause of OSA is not yet well known, several epidemiological studies report that it occurs at a high frequency mainly in men [2], and studies are mainly been conducted on anatomical risk factors such as neck circumference and obesity that can cause airway obstruction.

The prevalence of OSA defined at an apnea-hypopnea index (AHI) ≥5 was a mean of 22% (range, 9–37%) in men and 17% (range, 4–50%) in women in eleven published epidemiological studies published between 1993 and 2013 [2]. The prevalence of OSA increases with age and overweight. In particular, a prevalence in the elderly population is strikingly high; at ≥5 events/h AHI, this was 88% in men aged 65–69 y and 90% in men aged 60–85 y [3].

Recently, large-scale inter-ethnic studies have shown differences in prevalence and severity by race [4, 5] and differences in major risk factors for developing diseases before and after elderly population (anatomical risk factors are the main risk factor before elder age, and non-anatomical risk factors are the main risk factor after elder age) play significant role [2, 6-8]. Above all, OSA patients are at increased risk of comorbidities such as coronary heart disease (CHD) and type 2 diabetes (T2D) due to experience oxygen deficiency during repeated sleep, and systemic inflammation and abnormal activation of the sympathetic nervous system, so the cause and treatment of OSA are important [9].

I-2. Genetics of OSA

Recently reported some genetic association studies on OSA and its associated phenotypes have revealed the previously overlooked heritability of OSA: The heritability of OSA-related phenotypes is estimated to be up to 25–40% in familial study [10], around 40% of the variance in AHI has been shown to be explained by genetic factors [11], and high heritability was also confirmed in twin study [12]. In addition to confirming the high heritability of OSA and OSA-related phenotypes, attempts are being made to discover related genes through GWAS.

Several associated genetic variants were identified in GWAS using the FinnGen cohort, and four loci were verified and two new loci were additionally discovered in the multi-ethnic cohort (MVP) GWAS performed by Tamar Sofer et al [13]. Huajun Xu et al found two genes (PACRG, SLC52A3) from 20,590 Chinese people and confirmed the function of one gene (SLC52A3) using a mouse model [14], as a result, attempts are underway to identify genetic associations and causative genes in the development of OSA

I-3. Limitation of previous OSA genetic studies and purpose of this study

OSA is a complex disease and is a polygenic disease, so many GWAS studies are needed. However, large-scale genomic research is difficult due to overlooking the risk of OSA, diagnostic inaccuracy due to underdiagnosis, and genetic heterogeneity of the disease. Although OSA is a common disease, it is known that more than 85% of patients with clinically severe OSA have never been diagnosed, and there is a limitation of underdiagnosis, with many individuals suffering from OSA being misclassified as controls [15]. A report estimated that approximately 20% of US adults thought to have OSA, about 90% have not been diagnosed with OSA [16], and a study on the imbalance between the awareness rate and the estimated prevalence of OSA [17] suggest a need for genomic research to overcome the limitation of OSA underdiagnosis.

This study performed GWAS using the STOP-BANG questionnaire, which is one of the

highly reliable tools for determining OSA severity, and genomic data from a large-scale Korean cohort of 72,219 individuals, and matched GWAS was performed using clinical data from an additional 523 individuals.

II. METHODS AND MATERIALS

II-1. Dataset preparation

II-1.1. Cohort information and participants

In this study, we used the Korean genome and epidemiology study (KoGES) cohort, a population-based longitudinal design survey cohort, which was followed every two years from 2001 to 2018. This sample consisted of 72,219 men and women over 40 years of age, and repeated follow-up surveys were conducted every two years. All participants included genomic and epidemiological information, and epidemiological information consisted of questionnaire items including demographic information, disease history, family history, and lifestyle. In addition, it includes examination items such as blood pressure, body composition analysis, and clinical tests [18].

The clinical data used in the matched GWAS are obtained from Chonnam National University Hospital (CNUH), and consist of 403 participants diagnosed to be the OSA case group. These samples are patients over the age of 20 with a history of outpatient visits at Chonnam National University with an AHI of 5 or more on the polysomnography (PSG) or 3 or more on the STOP-BANG questionnaire.

II-1.2. Phenotype and covariates

To measure the severity of OSA, STOP-BANG, one of the OSA screening tools, was used. PSG is generally used as the gold standard for diagnosing the presence and severity of OSA, but it has limitations in diagnosing OSA due to its high cost, relatively difficult access, and long inspection time. Therefore, STOP-BANG, a relatively simple and accurate screening tool, has been widely used [19, 20]. This questionnaire consisted of four yes/no demographic questions (BMI, Age, Neck, Gender) and four clinical attribute measurements (Snoring, Tired, Observed, Pressure), with a “yes” answer for each question. It is a tool that can evaluate the severity of

OSA from 0 to 8 by giving 1 point for each answer [21].

In this study, we constructed the STOP-BANG questionnaire items using variables from the epidemiology and screening variables of KoGES for primary GWAS. Among the eight items, there was no measured value for neck circumference (N), so the neck circumference was predicted using data from 6,837 Korean men and women downloaded from Size Korea (release version 8) Human Body Survey (https://sizekorea.kr/human-info/meas-report?measDegree=8). The prediction model using an artificial neural network predicted neck circumference using standard waist and hip circumference values, and gender, age, weight, and height were adjusted as covariates (mean square error (MSE)=0.002). Among the 72,219 participants of KoGES cohorts used in this study, 25,712 participants who could calculate all eight items of the STOP-BANG questionnaire were used as a discovery set to discover genetic variants associated with OSA. And the 46,575 participants who could not calculate all eight items of the STOP-BANG questionnaire were used as a validation set for GWAS validation.

For the OSA phenotype of the matched GWAS for secondary GWAS, 25,712 participants, the discovery set of the KoGES cohort in which STOP-BANG was calculated, were divided into a low-risk OSA group with a STOP-BANG questionnaire score of 0–2; an intermediate risk OSA group with a score of 3–4; and a high-risk OSA group with a score of 5–8 [19]. Among the discovery sets of KoGES, 9,613 participants, the low risk OSA group, were used as the control group, and 403 participants with CNUH diagnosed as OSA were used as the case group.

II-1.3. Quality control and imputation of genotype data

Genotyping was performed using Korean Biobank Array and imputed using 1,000 Genome Project phase 3 V5 data as a reference panel. Quality control (QC) of genomic data was performed by excluding the following conditions: samples with less than 99% of valid genotypes; markers that presented a minor allele frequency lower than 0.05; valid genotypes that were present in less than 99% of samples; and markers outside the Hardy-Weinberg

Equilibrium. After QC process, association analysis evaluated 8,011,979 variants and 25,712 individuals. Principal component analysis was used to determine genetic background.

The genotype data used in matched GWAS is the genome data of the KoGES cohort and CNUH, and after imputation, 8,341,383 genetic mutations and 9,997 samples were used for analysis through the same quality control process as above.

II-2. Statistical analysis

II-2.1. Genome-wide association study (GWAS) and replication

In this study, three GWAS were performed to identify genetic variants associated with OSA. For quantitative analysis for the STOP-BANG score (0 to 8) in the discovery set (n=25,712), linear regression analysis was performed using PLINK v1.90b6.21, and age, gender, and BMI were used as covariates. We used the conventional P value threshold of 5×10^{-8} as the genome-wide significance threshold and genome-wide suggestive threshold of 1×10^{-5}.

Typically, the calculated STOPBANG is classified as 0 to 2 as low risk OSA, 3 to 4 as intermediate risk OSA, and 5 to 8 as high risk OSA. Therefore, after dividing the participants in the discovery set into three groups, to determine the results according to OSA risk, 12,862 low risk OSA and high risk OSA participants were selected and a logistic regression analysis was performed adjusting age, gender, and BMI as covariates. Additionally, obesity is considered a major risk factor for the development and progression of OSA, and the prevalence of OSA in obese or severely obese patients is nearly twice that of normal-weight adults. [22]. Therefore, BMI-stratified GWAS was performed to identify OSA-associated variants according to obesity level. Using the WHO-Asian BMI classification criteria, some of the 12,862 participants in the low risk and high risk groups were classified into two groups, the normal group (BMI range 18.5–22.9, n=4,361) and obese group (BMI range 25–, n=4,950). Logistic regression analysis was performed adjusting age and gender as covariates.

We used publicly available GWAS results from Finnish Caucasian individuals in FinnGen version.7 to replicate the variants found in samples of participants of the KoGES cohort. GWAS summary statistics for “sleep apnoea”, OSA-related phenotype, were used, and data were downloaded from the FinnGen portal (https://risteys.finregistry.fi/endpoints/G6 SLEEPAPNO).

In addition, the matching GWAS performed using the KoGES cohort and CNUH used logistic regression, with gender, age, and BMI corrected as covariate.

II-2.2. Polygenic risk score (PRS)

For validation of significant variants identified from continuous GWAS, polygenic risk score (PRS) was evaluated and compared. PRS analysis was performed using PRSice-2.

The first validation confirmed the PRS distribution of OSA-related traits using the validation set. The OSA-related traits used were whether or not you feel refreshed after sleep (0=always, 1=most of the time, 2=some of the time, 3=not at all), whether you have insomnia (1=no, 2=yes), and daytime sleepiness using Epworth Sleepiness Scale (ESS) (0–10=normal, 11–14=mild, 15–17= moderate, 18–24=severe). After calculating the PRS of the validation set, the bottom 10% group of PRS, and the top 10% group of PRS were compared using the chi-square test.

The second validation evaluated the PRS of the two groups, OSA low risk group and OSA high risk group, using the discovery set and validation set, and the t-test was used to compare PRS between the two groups. The OSA low risk group used 9,613 individuals with a STOP-BANG score of 0 to 2 from the discovery set, and the OSA high risk group used 426 individuals with a STOP-BANG score of 5 to 8 from a validation set independent of the discovery set. Because the variables for constructing the STOP-BANG questionnaire did not completely exist in the validation set, the STOP-BANG score for all participants could not be completely calculated from 0 to 8. Therefore, in order to satisfy the minimum conditions for classifying high risk of OSA, participants with at least 5 variables that could constitute items were selected, and participants with a calculated STOP-BANG score of 5 or more were classified as high risk. OSA participants were used.

II-2.3. Functional validation with public data

Gene expression was confirmed in the GTEX (Genotype-Tissue Expression) portal to confirm the expression of genes where significant mutations found from GWAS are located.

II-2.4. Propensity score matching (PSM)

KoGES cohort and CNUH data were used for matched GWAS using clinical information. Since all participants in CNUH are OSA cases, propensity score matching was used to select new control targets with similar propensity among participants in the KoGES cohort. R package "MatchIt" was used for matching, and the well-known risk factors of OSA, gender, age, BMI, and neck circumference, were used as matching features. Case and control were matched 1:1 or 1:2, and nearest (N), nearest with discard option (ND), and optimal (O) were used as the matching method. Therefore, various matching groups were compared by combining matching ratios and matching methods, and the matching group with optimal matching performance was used for analysis. Therefore, various matching groups were compared by combining matching ratios and matching methods, and the matching group with optimal matching performance was used for analysis. The matching group is indicated as N1-1, for example, for a 1:1 matching group using the nearest method. To determine the best matching method among the various matching methods used, the covariate distribution between the case and the control was evaluated using the balance measures (standardized mean difference, variance ratio, and eCDF mean (mean of differences in epidemiological curative distribution functions)).

III. RESULTS

III-1. Study characteristics

After selecting participants for whom a complete STOP-BANG score could be calculated and going through imputation and quality control, 25,712 individuals and 8,011,979 genetic variants were used in GWAS. According to the risk group of OSA classified by STOP-BANG (9,613 low risk samples, 12,854 intermediate risk samples, and 3,245 high risk samples), the participants tended to be male (3.2% in low risk samples vs. 56.4% in intermediate samples vs. 95.4% in high samples), older (52.0 ± 7.5 vs. 56.3 ± 8.1 vs. 57.0 ± 7.0 years old), and more obese in the risk group (23.8 ± 2.9 vs. 24.7 ± 3.0 vs. 25.8 ± 2.9 kg/m^2). The demographic characteristics of the participants used in the analysis are shown in Table 1.

Table 2 shows the demographic characteristics of CNUH participants (n=403) and low-risk OSA group participants (n=9,613) of the KoGES cohort used in Matched GWAS. The difference in gender distribution between the case and the control can be identified (78.2% of male in case, 96.8% of female in control), and the case has a higher BMI than the control (27.4 ± 4.4 in case, 23.8 ± 2.9 in control). The neck circumference is also higher (11.7% in case, 1.2% in control) in the case than in the control (more than 17 inches for men and more than 16 inches for women). The genetic analysis of OSA was performed according to the genetic analysis flow described in Figure 1.

Table 1. Sample description of KoGES

	OSA Risk Group		
	Low Risk (N=9,613)	**Intermediate Risk (N=12,854)**	**High Risk (N=3,245)**
Sex			
Male	307 (3.2%)	7,252 (56.4%)	3,096 (95.4%)
Female	9,306 (96.8%)	5,602 (43.6%)	149 (4.6%)
Age	52.0 ± 7.5	56.3 ± 8.1	57.0 ± 7.0
BMI	23.8 ± 2.9	24.7 ± 2.9	25.8 ± 2.9
STOPBANG			
1	2,587 (26.9%)	0 (0.0%)	0 (0.0%)
2	7,026 (73.1%)	0 (0.0%)	0 (0.0%)
3	0 (0.0%)	7,794 (60.6%)	0 (0.0%)
4	0 (0.0%)	5,060 (39.4%)	0 (0.0%)
5	0 (0.0%)	0 (0.0%)	2,495 (76.9%)
6	0 (0.0%)	0 (0.0%)	677 (20.9%)
7	0 (0.0%)	0 (0.0%)	73 (2.2%)
8	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 2. Sample description of matched subjects

	CNUH (N=403)	KoGES (N=9,613)
Sex		
Male	315 (78.2%)	307 (3.2%)
Female	88 (21.8%)	9,306 (96.8%)
Age	52.3 ± 14.2	52.0 ± 7.5
BMI	27.4 ± 4.4	23.8 ± 2.9
Neck		
Yes	47 (11.7%)	115 (1.2%)
No	356 (88.3%)	9,498 (98.8%)
AHI	35.5 ± 23.7	-
STOPBANG		
0	3 (0.8%)	0 (0.0%)
1	21 (5.4%)	2,587 (26.9%)
2	40 (10.3%)	7,026 (73.1%)
3	101 (26.1%)	0 (0.0%)
4	95 (24.5%)	0 (0.0%)
5	64 (16.5%)	0 (0.0%)
6	47 (12.1%)	0 (0.0%)
7	15 (3.9%)	0 (0.0%)
8	1 (0.3%)	0 (0.0%)

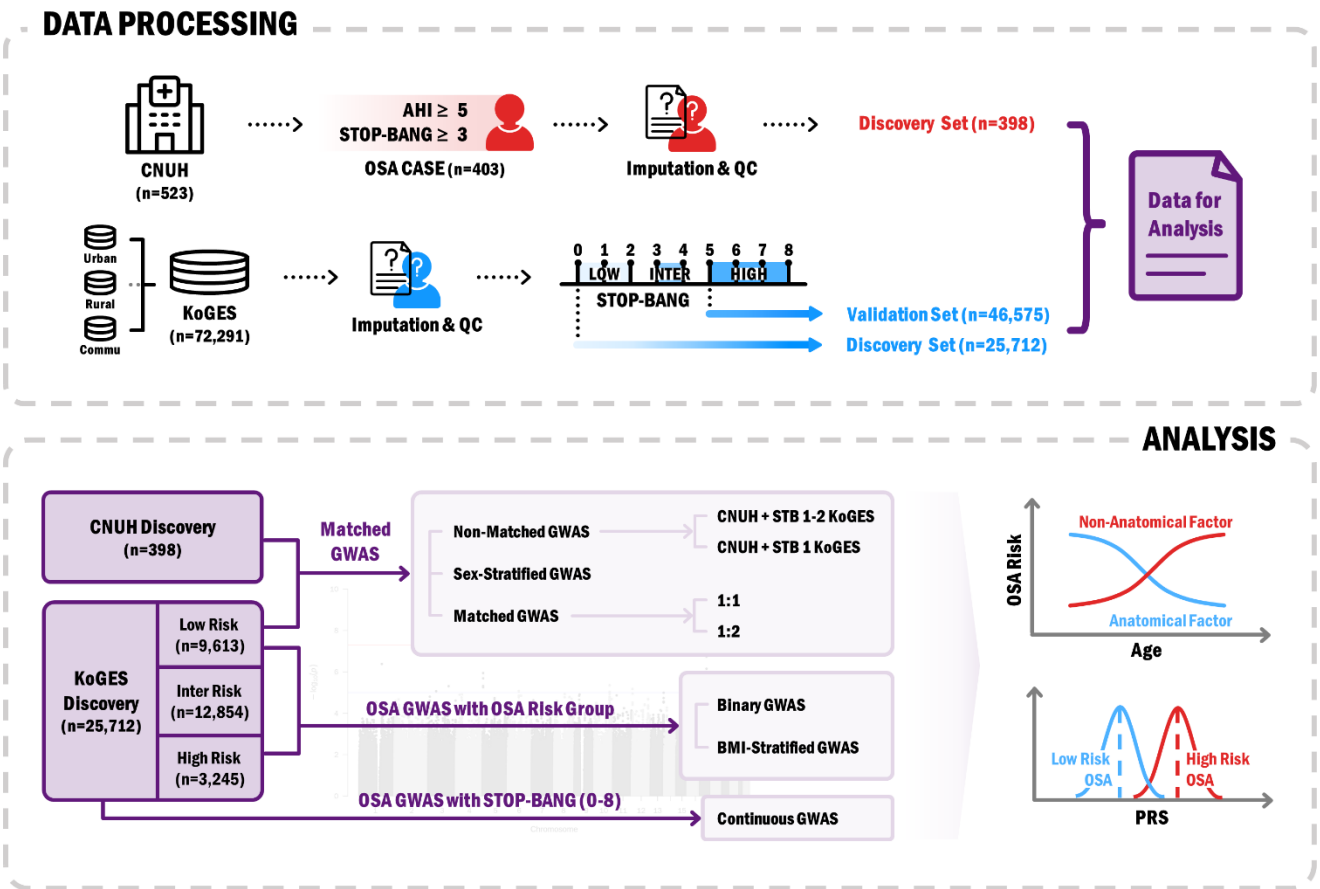

DATA PROCESSING
CNUH
(n=523)
AHI ≥ 5
STOP-BANG ≥ 3
OSA CASE (n=403)
Imputation & QC
Discovery Set (n=398)
Urban
Rural
Commu
KoGES
(n=72,291)
Imputation & QC
0 1 2 3 4 5 6 7 8
LOW
INTER
HIGH
STOP-BANG
Validation Set (n=46,575)
Discovery Set (n=25,712)
Data for Analysis
ANALYSIS
CNUH Discovery
(n=398)
KoGES Discovery
(n=25,712)
Low Risk
(n=9,613)
Inter Risk
(n=12,854)
High Risk
(n=3,245)
Matched GWAS
Non-Matched GWAS
CNUH + STB 1-2 KoGES
CNUH + STB 1 KoGES
Sex-Stratified GWAS
Matched GWAS
1:1
1:2
OSA GWAS with OSA Risk Group
Binary GWAS
BMI-Stratified GWAS
OSA GWAS with STOP-BANG (0-8)
Continuous GWAS
OSA Risk
Non-Anatomical Factor
Anatomical Factor
Age
Low Risk OSA
High Risk OSA
PRS

Figure 1. Overall workflow. The top box (DATA PROCESSING) shows the pre-processing process of the two datasets used in this analysis. The first dataset, CNUH, is a clinical dataset composed of patients diagnosed with OSA (AHI ≥ 5 or STOP-BANG score ≥ 3) at Chonnam National University Hospital. The second dataset is KoGES. After the two datasets went through the imputation and QC process, the CNUH dataset was used as a discovery set, and the KoGES dataset was used as a discovery set and validation set based on STOP-BANG. The bottom box (ANALYSIS) is the analysis flow using the dataset defined in the top box process. The CNUH discovery set and low risk of KoGES discover set were used for matched GWAS, and GWAS according to OSA risk group and GWAS according to STOP-BANG were performed using the KoGES discovery set. After GWAS, validation of significant variants from continuous GWAS were performed by comparing PRS between risk groups, and additional analysis was conducted on the effect of risk factors on OSA.

III-2. Primary GWAS and replication

III-2.1. Continuous GWAS

Figure 2 panel A visualizes continuous GWAS results with adjusted gender, age, and BMI and provides a plot that shows the annotations of the nearest genes together for each associated loci. Panel B provides a QQ-plot, and a regional association plot for locus of interest among GWAS results is provided in Panel C and D. 9 loci (45 variations), which are genome-wide significant loci, were identified in continuous GWAS, and summary statistics of the lead variants of each association are provided in Table 3. 7 of these loci show the same direction of association in KoGES and FinnGen, with one association (two variants; rs4680142 (P-value=4.39e-09, BETA=-0.048) and rs1915714 (P-value=1.07e-08, BETA=-0.04629) located around MME genes replicated in FinnGen (nominal signal level).

III-2.2. Binary GWAS

Figure 3 panel A visualizes binary GWAS (n=12,862) results with adjusted gender, age, and BMI, and panel B provides a QQ-plot. As a result of the binary GWAS, 1 genome-wide significant variants and 26 genome-wide suggestive variants were identified, and their summary statistics are provided in Table 4. Identified 1 significant variant, rs1358310000 (P-value=2.90e-08, OR=3.365), is located in the CERS6 gene.

III-2.3. BMI-stratified GWAS

Figure 4 panel A is miami plot of BMI-stratified GWAS. The top part of the miami plot provides the analysis results of the normal group (n=4,361), and the bottom provides the analysis results of the obese group (n=4,950). The QQ-plot of each analysis is provided in panels B and C. As a result of the analysis of the normal group, 1 genome-wide significant variant and 31 genome-wide suggestive variants were identified. And as a result of the analysis of the obese group, 1 genome-wide significant variant and 120 genome-wide suggestive

variants were identified, and their summary statistics are provided in Table 5. Each significant variant, rs397769528 (P-value=1.96e-08, OR=18.97) and rs1207477605 (P-value=2.09, OR=2.123) are located in the OTOF and FARP2 gene.

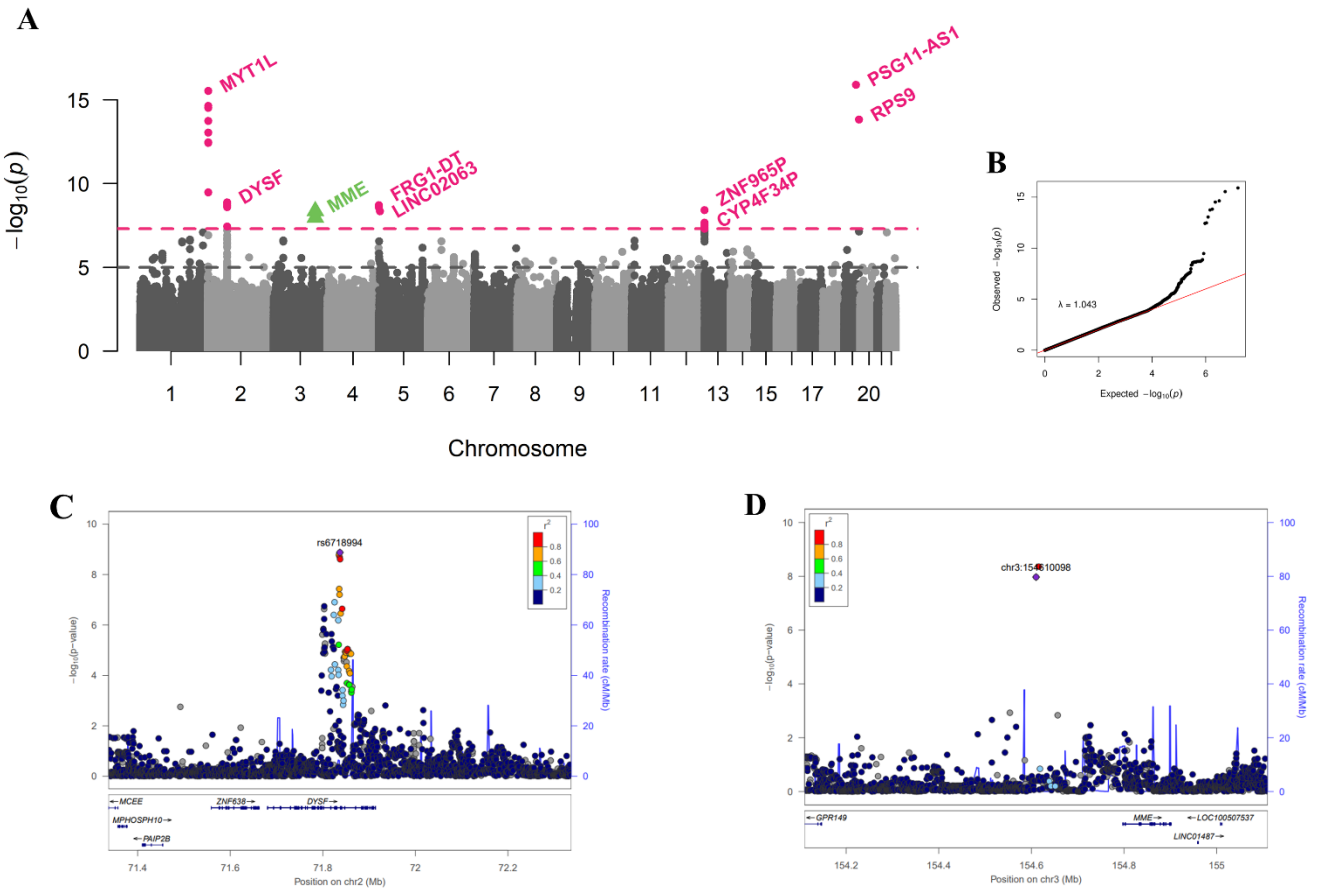

A
MYT1L
DYSF
MME
FRG1-DT
LINC02063
ZNF965P
CYP4F34P
PSG11-AS1
RPS9
−log10(p)
Chromosome
B
λ = 1.043
Observed −log10(p)
Expected −log10(p)
C
rs6718994
r²
Recombination rate (cM/Mb)
Position on chr2 (Mb)
MCEE
MPHOSPH10
PMP2B
ZNF638
DYSF
D
chr3:154610098
GPR149
MME
LOC100507537
LINC01487
Position on chr3 (Mb)

Figure 2. Result of continuous GWAS. A is the manhattan plot of continuous GWAS. The x-axis represents the chromosome location and y represents –log10 of the P-value. The pink dotted line represents the genome-wide significance level, and the gray dotted line represents the genome-wide suggestive level. Each dot represents each variant. Among them, pink dots represent variants at a significant level, and light green triangles represent replicated variants among significant variants. B is QQ-plot. C and D are regional plots for chromosomes 2 and 3, respectively. The x-axis indicates the chromosome location and the y-axis indicates the P-value of –log10. Also shown are recombination rates and genes in the regions. SNP color indicates the strength of LD (r2) with the index SNP.

Table 3. A summary of significant variants and replication in continuous GWAS

								KoGES Discovery				FinnGen Validation			
CHR	**BP**	**Lead SNP**	**Ref**	**Alt**	**A1**	**MAF**	**Nearest Gene**	***P* value**	**BETA**	**SE**	**95% CI**	***P* value**	**BETA**	**SE**	**ALT_FRQ**
19	43164753	rs57122208	T	G	G	0.3118	PSG11-AS1	1.25E-16	0.06428	0.00776	0.05-0.08	0.850779	-0.00191344	0.0101711	0.175429
2	2515763	rs11690231	C	T	T	0.03086	MYT1L	2.88E-16	0.171	0.0209	0.13-0.21	0.307179	0.00807683	0.00790948	0.379519
19	54225766	rs56128527	A	G	G	0.09844	RPS9	1.51E-14	0.0941	0.01223	0.07-0.12	0.985304	0.00171599	0.0931602	0.00175963
2	71609762	rs6718994	T	C	T	0.489	DYSF	1.34E-09	0.04413	0.007277	0.03-0.06	0.301237	-0.00809593	0.00783136	0.593762
4	189757692	rs5020494	A	T	T	0.1892	FRG1-DT	1.97E-09	-0.05564	0.00927	-0.03	-	-	-	-
	189757691	rs1222053389	TAG			-		-	-	-	-	0.282787	0.00835389	0.00777773	0.573385
3	**154898021**	**rs4680142**	**A**	**G**	**G**	**0.2799**	**MME**	**4.39E-09**	**-0.04769**	**0.008123**	**-0.03**	**0.0225554**	**-0.0198304**	**0.00869417**	**0.278278**
5	4222774	rs12656983	C	T	T	0.104	LINC02063	4.57E-09	-0.06974	0.01189	-0.04	0.505267	-0.0238592	0.0358125	0.0123341
3	**154892309**	**rs1915714**	**C**	**A**	**A**	**0.2841**	**MME**	**1.07E-08**	**-0.04629**	**0.008091**	**-0.03**	**0.00965784**	**-0.0214259**	**0.00827944**	**0.333072**
13	18739027	rs149980675	G	T	T	0.05215	ZNF965P	2.14E-08	-0.09137	0.01631	-0.06	0.824333	0.00428805	0.0193177	0.0424147
13	18745045	rs370968193	T	G	G	0.05072	CYP4F34P	2.82E-08	-0.0917	0.01651	-0.06	0.497077	-0.0174525	0.0256997	0.0236974

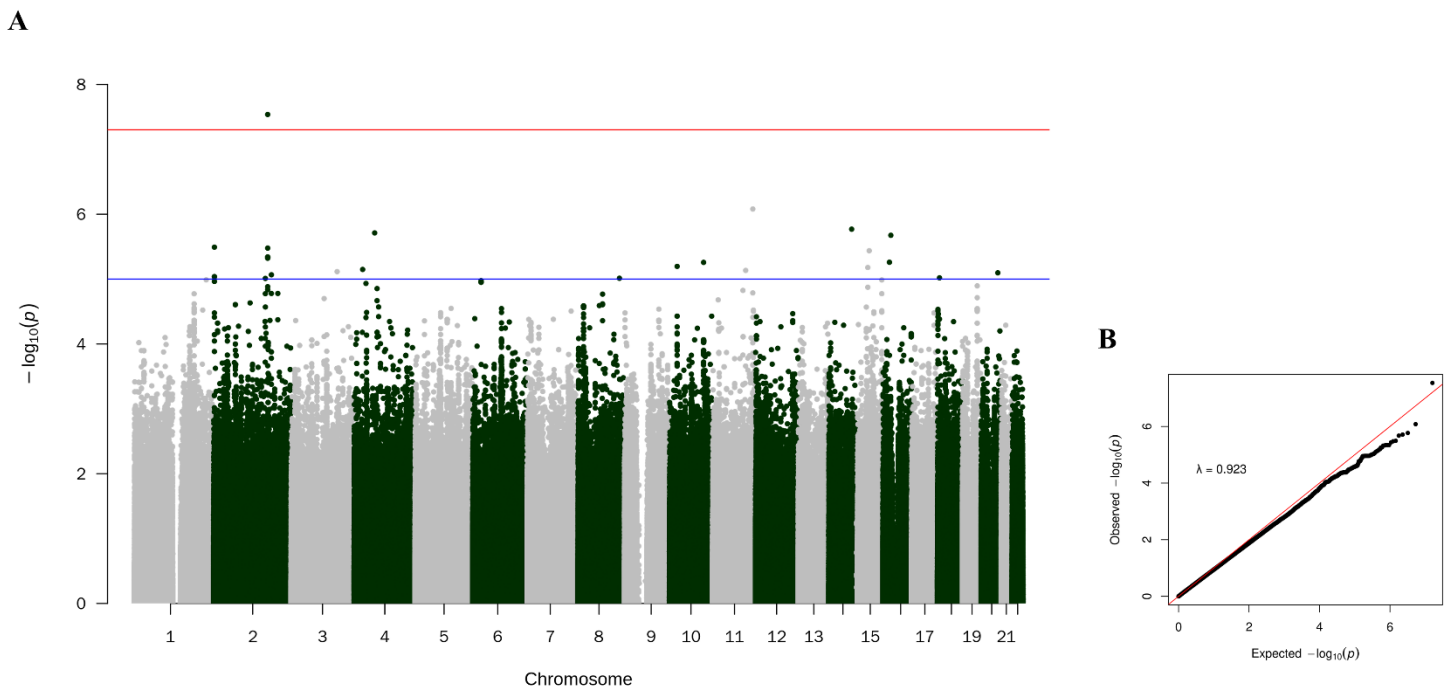


Figure 3. Result of binary GWAS. A is manhattan plot. The x-axis represents the chromosome location and y represents –log10 of the P-value. The red line represents the genome-wide significance level, and the blue line represents the genome-wide suggestive level. Each dot represents each variant. Among them, the green dot indicates a significant level of variant. B is QQ-plot.

Table 4. A summary of lead variants of binary GWAS

CHR	BP	rsID	A1	A2	MAF	OR	STAT	*P* value	Overlapped.Gene	Type	Annotation
2	169619418	rs1358310000	CT	C	0.03207	3.365	5.547	2.90E-08	CERS6	protein_coding	intronic
11	126279469	rs2298476	A	G	0.01905	3.809	4.928	8.32E-07	ST3GAL4	protein_coding	3downstream,intronic,non-coding intronic
14	91789117	rs138497065	T	C	0.01559	4.04	4.786	1.70E-06	CCDC88C	protein_coding	intronic
4	64212994	rs76302397	T	C	0.01061	5.09	4.759	1.94E-06	None	None	None
16	25038823	rs76790640	A	G	0.01213	4.262	4.742	2.11E-06	None	None	None
2	2520359	rs13023303	T	G	0.0206	3.257	4.656	3.22E-06	None	None	None
2	169650371	rs3845727	C	T	0.0325	2.684	4.649	3.33E-06	NOSTRIN	protein_coding	intronic
15	59601136	rs183892273	T	C	0.01042	4.798	4.63	3.66E-06	MYO1E	protein_coding	non-coding intronic,intronic
2	169646622	rs3845726	G	C	0.03234	2.66	4.584	4.57E-06	NOSTRIN	protein_coding	intronic
2	169645524	rs3856426	A	G	0.03234	2.66	4.584	4.57E-06	NOSTRIN	protein_coding	intronic

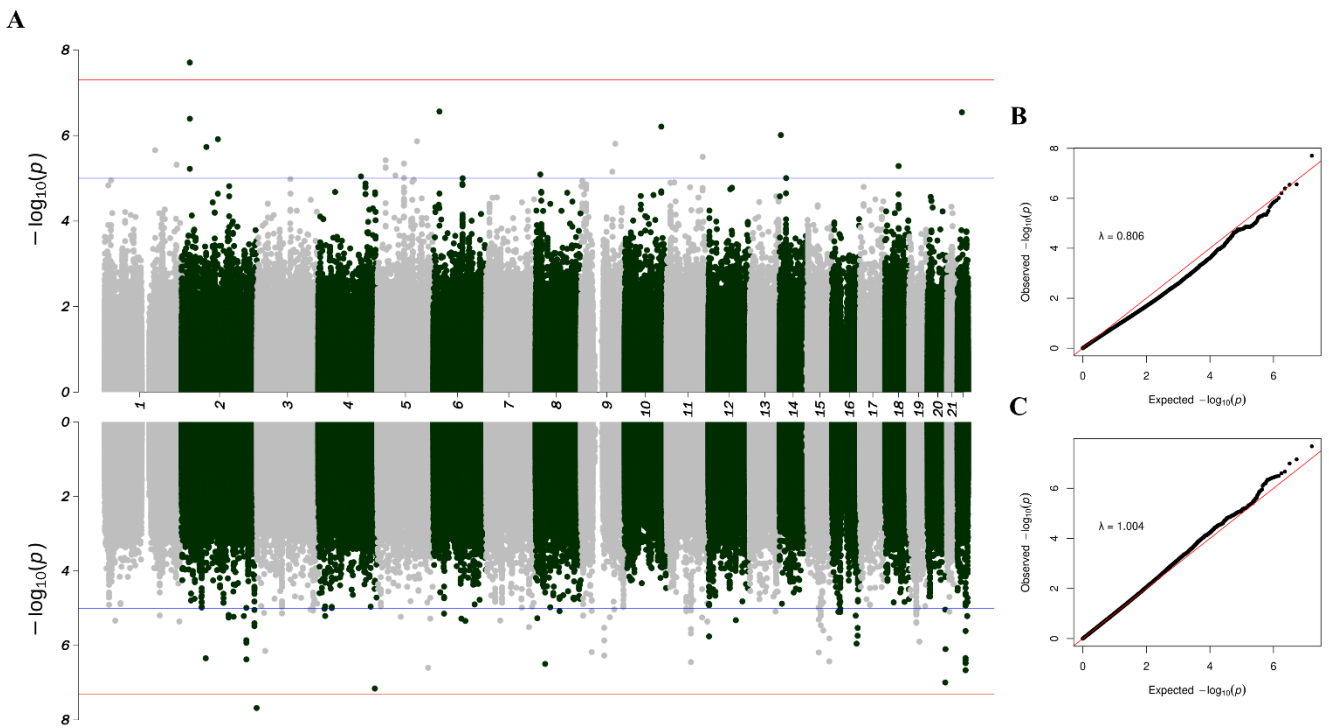


Figure 4. Result of BMI-stratified GWAS. A is miami plot. The x-axis represents the chromosome location and y represents –log10 of the P-value. The red line represents the genome-wide significance level, and the blue line represents the genome-wide suggestive level. Each dot represents each variant. Among them, the green dot indicates a significant level of variant. The top plot shows the results using the normal group, and the bottom plot shows the results using the obese group. B is the QQ-plot of the analysis results using the normal group, and C is the QQ-plot of the results using the obese group.

Table 5. A summary of lead variants of BMI-stratified GWAS

CHR	Variation.ID	BP	A1	A2	MAF	OR	STAT	*P* value	Overlapped.Gene	Type	Annotation
Normal											
2	**rs397769528**	**26738286**	**G**	**A**	**0.01364**	**18.97**	**5.616**	**1.96E-08**	**OTOF**	**protein_coding**	**intronic**
6	rs142288353	19899293	G	A	0.01261	15.66	5.14	2.74E-07	None	None	None
22	-	31078665	A	AG	0.01353	11.99	5.133	2.86E-07	-	-	-
2	rs3739079	26863388	A	G	0.02041	12	5.067	4.04E-07	CIB4	protein_coding	intronic
10	-	118772897	C	CT	0.009402	22.01	4.985	6.19E-07	-	-	-
14	rs17102928	24794905	T	C	0.01227	12	4.897	9.75E-07	ADCY4	protein_coding	non-coding intronic,3downstream,intronic
2	rs117470902	117356952	A	G	0.02075	8.969	4.853	1.22E-06	None	None	None
5	rs187490799	128280793	A	T	0.02052	7.843	4.831	1.36E-06	SLC27A6	protein_coding	intronic
9	rs138531605	111245549	A	T	0.01307	18.14	4.803	1.56E-06	None	None	None
Obese											
2	**rs1207477605**	**242413732**	**C**	**CAAAAAAAAAAAAAAAAAA A**	**0.2896**	**2.123**	**5.604**	**2.09E-08**	**FARP2**	**protein_coding**	**intronic,non-coding intronic**
4	rs79673198	182815088	G	A	0.01202	7.337	5.392	6.97E-08	AC108142.1	antisense	non-coding intronic
20	-	57458243	GT	G	0.0202	4.975	5.324	1.01E-07	-	-	-
22	rs149015834	42949368	T	C	0.03778	3.99	5.185	2.16E-07	None	None	None
5	rs117704152	163940014	A	G	0.01465	6.596	5.157	2.52E-07	CTC-340A15.2	antisense	non-coding intronic
8	rs1007499568	31161840	TAC	T	0.01222	6.924	5.112	3.19E-07	RP11-566H8.3	lincR-	non-coding intronic
22	rs5751310	43009492	A	C	0.04475	3.7	5.103	3.34E-07	POLDIP3	protein_coding	intronic,non-coding intronic
11	rs74869738	79146597	A	G	0.0197	5.314	5.091	3.56E-07	TENM4	protein_coding	intronic,non-coding intronic

III-3. PRS for validation

III-3.1. PRS with OSA-related traits

To validate the 45 significant variants identified from continuous GWAS, a first validation was performed using them to calculate PRS in a validation set independent of the discovery set. As a result of comparing the distribution of OSA-related traits in the bottom 10% group of PRS and the top 10% group of PRS, no significant differences between the two groups according to PRS distribution were found in the three OSA-related traits; For feeling refreshed after sleep, P-value=0.1836, for insomnia, P-value=0.6335, and for daytime sleepiness, P-value=0.8126.

III-3.2. PRS with STOP-BANG

An additional second validation of significant variation in continuous GWAS was performed in the discovery set and validation set. To calculate the PRS, the same 45 variants as in the first validation were used, and the PRS comparison of the OSA low risk group and the high risk group is shown in Figure 5. By confirming that the average PRS was significantly higher in high risk OSA subjects compared to low risk OSA subjects (P-value = 6.03×10^{-5}), it was possible to validate that there was a clear difference between the two groups.

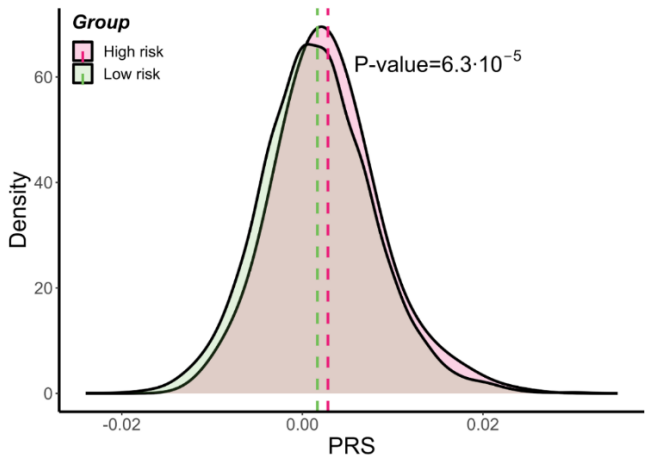


Figure 5. PRS comparison by OSA risk group. This is the PRS distribution calculated using 45 significant variants among the results of continuous GWAS. The pink distribution represents the PRS distribution of the high risk OSA group, and the light green distribution represents the PRS distribution of the low risk OSA group.

III-4. Secondary GWAS with an additional clinical sample

III-4.1. Non-matched GWAS

Two GWAS results using CNUH and KoGES without PSM are provided in Figure 6 and 7. Both results were adjusted for sex, age, and BMI. Figure 6 is the GWAS results of 398 cases and 9,599 controls with low risk OSA with a STOP-BANG score of 1 to 2, and Figure 7 is the GWAS results of 398 cases and 2,586 controls with a lower risk STOP-BANG score of 1. 3 loci (8 variants) were found in the first GWAS, which is genome-wide significant loci, but no significant loci were found in the second GWAS. The summary statistics of their lead variants are provided in Table 6. The regional plot for the 3 loci found in the first GWAS is provided in Figure 8.

III-4.2. Matched GWAS using PSM

Figure 9 show the results of the degree to which matched CNUH and KoGES are balanced according to various methods of PSM using sex, age, and BMI as matching variables. The plot of Figure 9 show balance measurement according to the matching model, with standard mean difference provided to panel B, variance ratio provided to panel D, and eCDF provided to panel F. The x-axis represents the model according to the matching method, and the y-axis represents the value of each performance. Each color represents a covariate performance value used in the model according to the matching method. It shows better balance values at N1-1 and ND1-2 compared to the balance measure of the baseline model. Therefore, matched GWAS adjusting sex, age, and BMI was performed using two datasets. The results are provided in Figure 10, panel A and B show GWAS results using N1-1 dataset, and panel C and D show GWAS results using ND1-2 dataset. 6 significant variants were identified in the GWAS of N1-1 dataset and ND1-2 dataset, respectively, and summary statistics of each of these variants are provided in Table 7.

Matching results according to various methods of PSM using sex, age, BMI, and neck as matching variables are provided in Figure 11, and similarly, they perform best in N1-1 and

ND1-2 models. Figure 12 and Table 8 show the results of GWAS adjusting sex, age, and BMI using datasets matched from the two models.

As a result of GWAS between two models using sex, age, and BMI as matching variables and two models using sex, age, BMI, and neck as matching variables, it was confirmed that some of the variants found in non-matched GWAS without PSM were found as significant variants.

III-4.3. Sex-stratified GWAS

According to Table 2, the similar number of male in the two dataset, CNUH and KoGES, was confirmed. Therefore, sex-stratified GWAS was performed using only additional 611 male (CNUH=304, KoGES=307). As a result, no significant variant was found, and 34 suggestive variants were identified, and their summary statistics are provided in Table 9 and visualized in Figure 13.

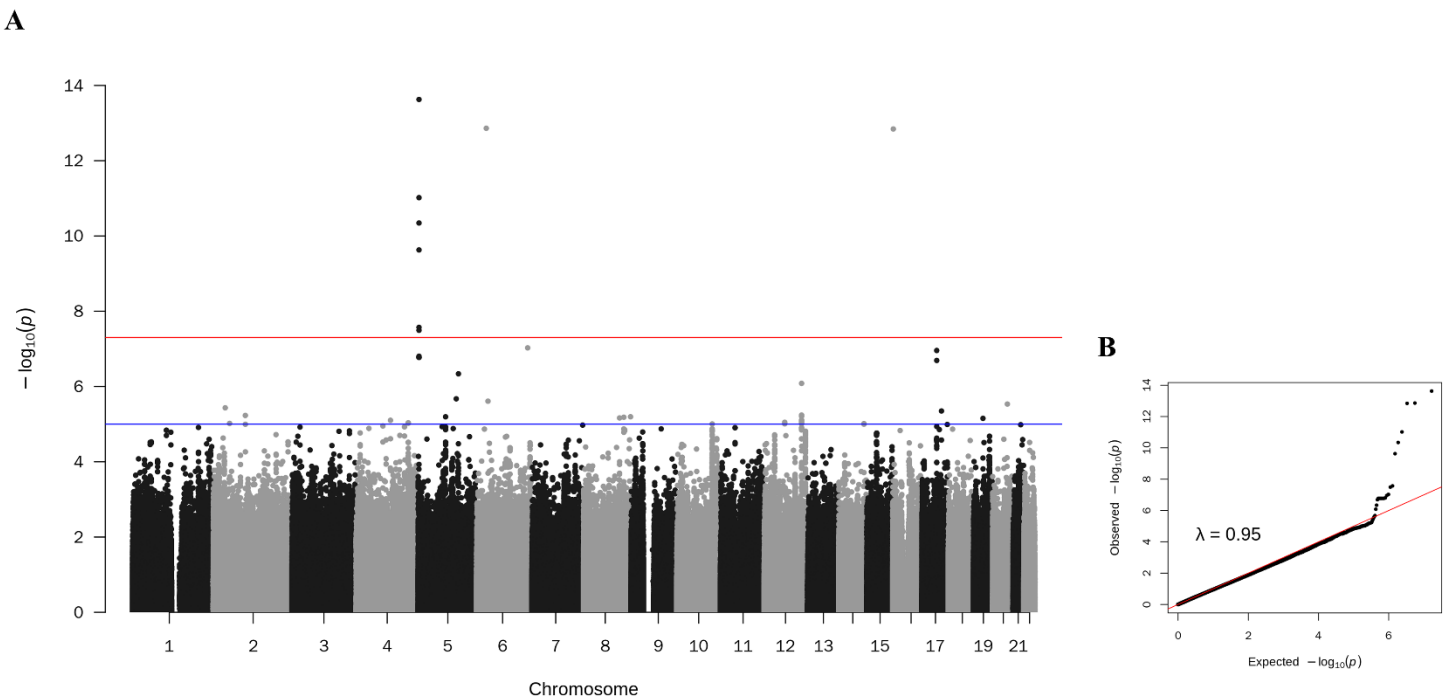


Figure 6. Result of non-matched GWAS using all participants in CNUH and participants in KoGES with 1 to 2 of STB. A is manhattan plot. The x-axis represents the chromosome location and y represents –log10 of the P-value. The red line represents the genome-wide significance level, and the blue line represents the genome-wide suggestive level. Each dot represents each variant. Among them, black dots represent significant levels of variant. B is QQ-plot.

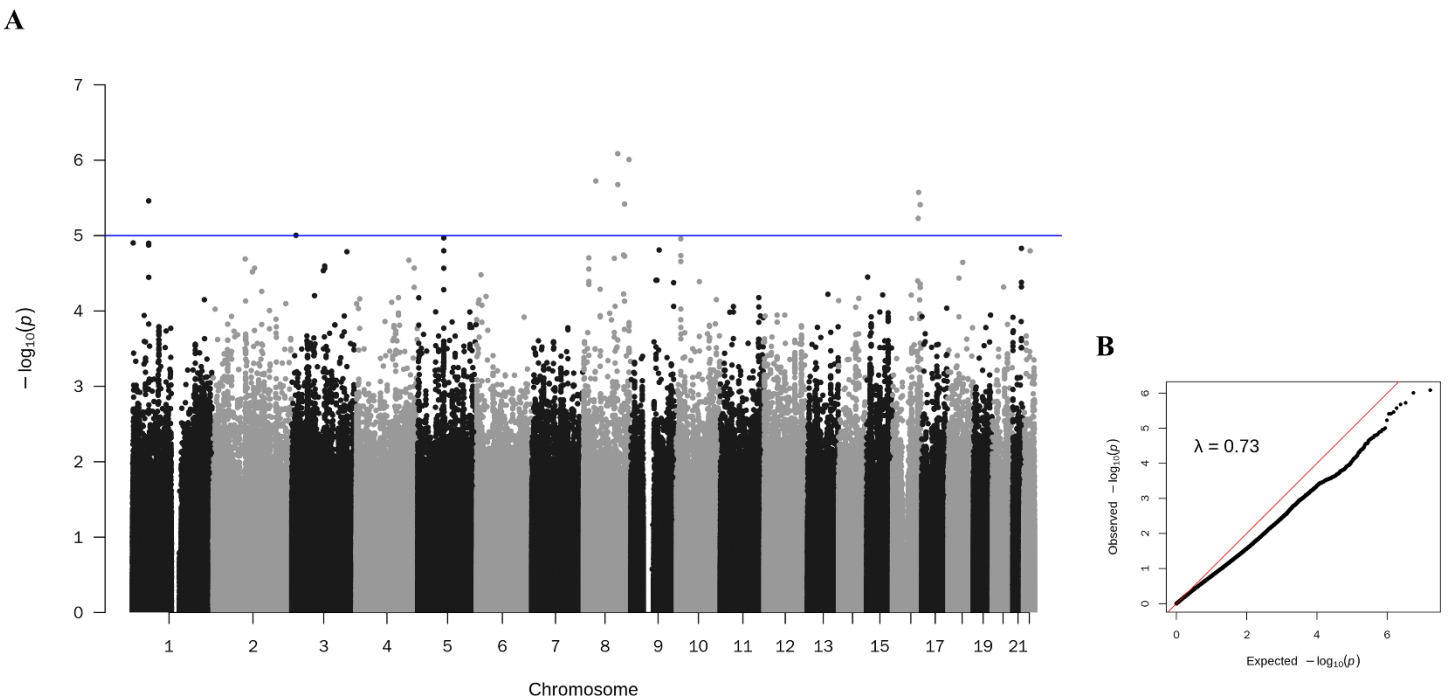


Figure 7. Result of non-matched GWAS using all participants in CNUH and participants in KoGES with 1 of STB. A is manhattan plot. The x-axis represents the chromosome location and y represents –log10 of the P-value. The blue line represents the genome-wide suggestive level. Each dot represents each variant. B is QQ-plot.

Table 6. A summary of significant or lead variants of non-matched GWAS

rsID	CHR	BP	A1/A2	MAF	OR	SE	95% CI	*P* value	Type	Nearest Gene	SNP Type	R^2
All CNUH + All low risk KoGES with STB 1-2												
rs112877582	5	3002638	A/G	0.06857	3.428	0.1614	2.509-4.723	2.36E-14	None	RP11-35O7.1 / CTD-2029E14.1	Genotyped	-
rs9261027	6	29966000	C/A	0.07792	3.114	0.1533	2.32-4.231	1.38E-13	None	MCCD1P2 / ZNRD1-AS1	Imputed	0.858379
rs200305090	16	1871961	G/T	0.02851	4.738	0.21	3.153-7.181	1.43E-13	protein_coding	HAGH	Genotyped	-
rs6873065	5	3003801	C/G	0.08292	2.884	0.1554	2.136-3.928	9.59E-12	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.89416
rs867569627	5	3000078	CA/C	0.07202	2.906	0.1619	2.124-4.006	4.52E-11	None	RP11-35O7.1 / CTD-2029E14.2	Imputed	0.86349
rs1355923525	5	3003489	G/GTCCT	0.1006	2.505	0.1448	1.895-3.344	2.35E-10	None	RP11-35O7.1 / CTD-2029E14.3	Imputed	0.85959
rs80298604	5	2994576	T/C	0.03911	3.121	0.2046	2.095-4.672	2.69E-08	None	RP11-35O7.1 / CTD-2029E14.4	Imputed	0.83503
rs7721620	5	2992739	G/C	0.03976	3.091	0.204	2.078-4.622	3.21E-08	None	RP11-35O7.1 / CTD-2029E14.5	Imputed	0.84305
All CNUH + Low risk KoGES with STB 1												
rs187965074	8	106433906	C/T	0.01508	10.37	0.4745	4.094-26.29	8.20E-07	protein_coding	ZFPM2	Imputed	0.95739
-	8	141021598	CT/C	0.05379	4.611	0.3122	2.501-8.503	9.81E-07	-	-	Imputed	0.93379
rs1333527519	8	38703957	T/TA	0.02044	8.133	0.4398	3.435-19.26	1.89E-06	protein_coding, antisense	TACC1, RP11-723D22.3	Imputed	0.90851
rs183549291	8	106454943	G/C	0.01491	9.988	0.4852	3.859-25.85	2.10E-06	protein_coding	ZFPM2	Imputed	0.96095
rs149964305	16	80678322	G/T	0.01927	9.178	0.4722	3.637-23.16	2.67E-06	protein_coding	CDYL2	Imputed	0.92047
rs7550594	1	50891093	T/C	0.02162	7.357	0.43	3.167-17.09	3.46E-06	None	DMRTA2 / RP5-850O15.4	Imputed	0.9964
rs138704473	8	127305212	G/A	0.01776	7.898	0.4472	3.287-18.97	3.82E-06	None	RFPL4AP5 / RP11-65D17.1	Imputed	0.87658
rs115100491	16	85542521	C/G	0.0191	7.54	0.4375	3.198-17.77	3.89E-06	None	AC092377.1 / RP11-118F19.1	Imputed	0.89517
rs150723989	16	79302835	G/T	0.009048	11.4	0.5374	3.978-32.69	5.91E-06	None	RNA5SP431 / RP11-467I17.1	Imputed	0.99676
rs76954666	3	12916429	T/C	0.02061	6.026	0.4064	2.717-13.37	9.91E-06	None	CAND2 / RP11-767C1.1	Imputed	0.83641

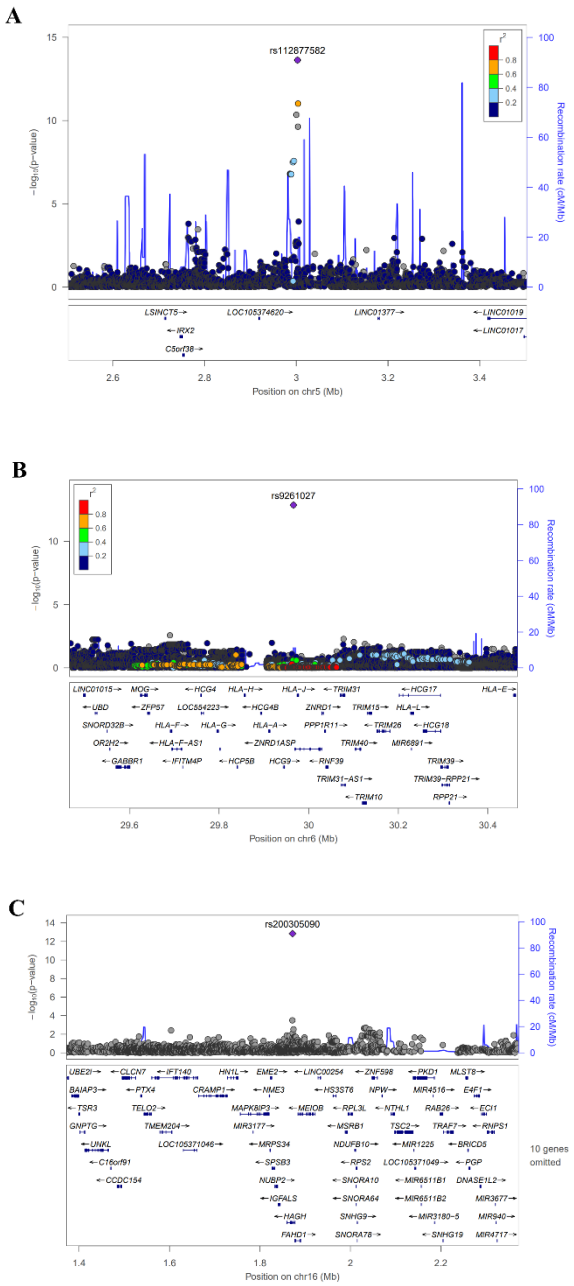
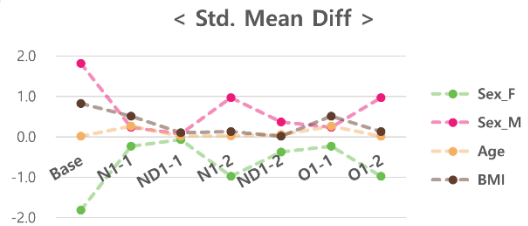


Figure 8. Regional plot of non-matched GWAS. Each panel is a regional plot for chromosomes 5, 6, and 16. The x-axis indicates the chromosome location and the y-axis indicates the P-value of –log10. Also shown are recombination rates and genes in the regions. SNP color indicates the strength of LD (r2) with the index SNP.

A

	Std. MeanDiff						
	Base	N1-1	ND1-1	N1-2	ND1-2	O1-1	O1-2
Case	403	403	279	403	279	403	403
Ctrl	9,613	403	279	806	589	403	806
Sex_F	-1.8147	-0.2342	-0.0694	-0.9730	-0.3731	-0.2342	-0.9730
Sex_M	1.8147	0.2342	0.0694	0.9730	0.3731	0.2342	0.9730
Age	0.0175	0.2614	0.0126	0.0232	0.0582	0.2614	0.0181
BMI	0.8233	0.5139	0.1049	0.1273	0.0133	0.5139	0.1313

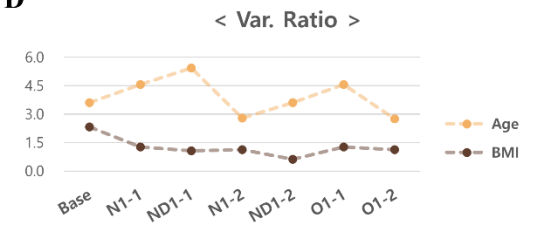
B



C

	Var.Ratio						
	Base	N1-1	ND1-1	N1-2	ND1-2	O1-1	O1-2
Case	403	403	279	403	279	403	403
Ctrl	9,613	403	279	806	589	403	806
Sex_F	.	.	.	.	.	.	.
Sex_M	.	.	.	.	.	.	.
Age	3.6182	4.5655	5.4335	2.8030	3.6140	4.5655	2.7628
BMI	2.3393	1.2638	1.0745	1.1341	0.6143	1.2638	1.1313

D



E

	eCDF Mean						
	Base	N1-1	ND1-1	N1-2	ND1-2	O1-1	O1-2
Case	403	403	279	403	279	403	403
Ctrl	9,613	403	279	806	589	403	806
Sex_F	0.7497	0.0968	0.0287	0.4020	0.1541	0.0968	0.4020
Sex_M	0.7497	0.0968	0.0287	0.4020	0.1541	0.0968	0.4020
Age	0.0920	0.1310	0.1300	0.0876	0.2437	0.1310	0.0865
BMI	0.2113	0.1335	0.0345	0.0497	0.1541	0.1335	0.0501

F

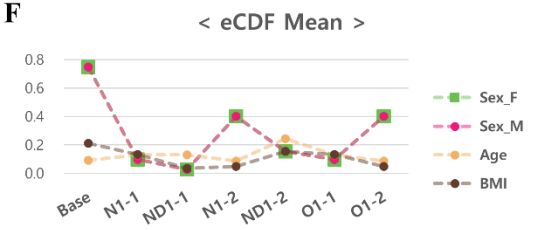


Figure 9. PSM balance performance using sex, age, and BMI as matching variables. The following are matching results using sex, age, and BMI as matching variables. A, C, and E are tables showing matching results according to the matching method after PSM, and B, D, and F are graphs visualizing the tables. A and B are Std. Mean Diff (standardized mean difference), C and D are Var. Ratio (variance ratio), and E and F are the results for eCDF Mean. The x-axis of the graph represents the different methods used for matching, and the y-axis represents the value of matching measure (Std. Mean Diff, Var. Ratio, and eCDF Mean). Each color coding shows the value of matching measure of the matching variable used for PSM.

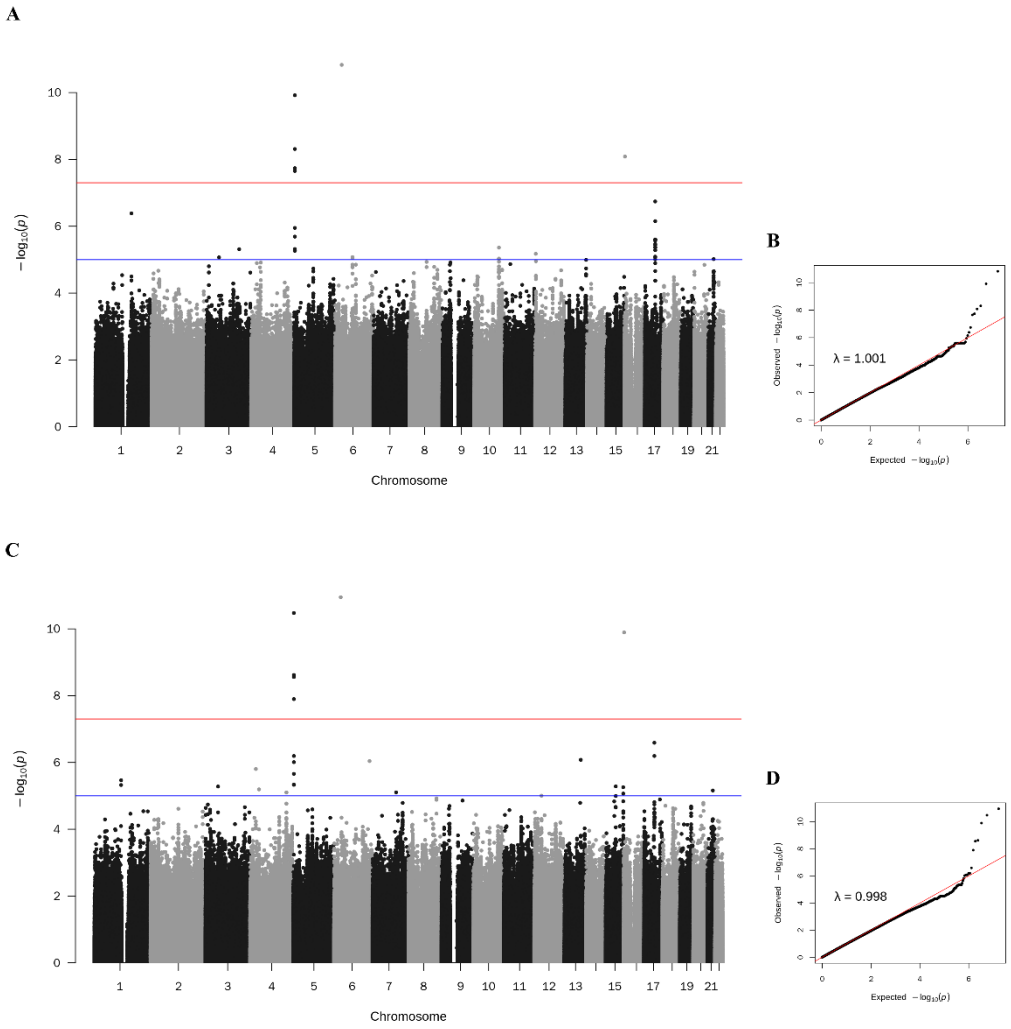


Figure 10. Result of matched GWAS using sex, age, and BMI as matching variables. A and B visualize the 1:1 matching results, and C and D visualize the 1:2 matching results. A and C are manhattan plots, where the x-axis represents the chromosome location and y represents –log10 of the P-value. The red line represents the genome-wide significance level, and the blue line represents the genome-wide suggestive level. Each dot represents a variant. Among them, black dots represent significant levels of variation. B and D are QQ-plots.

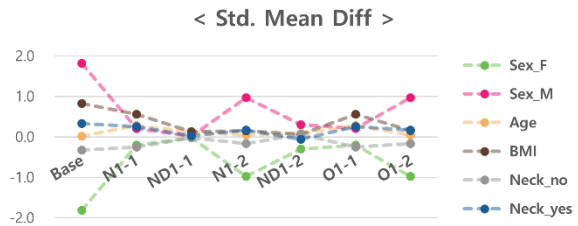
Table 7. A summary of significant variants of matched GWAS using sex, age, and BMI as matching variables

rsID	CHR	BP	A1/A2	MAF	OR	SE	95% CI	*P* value	Type	Nearest Gene	SNP Type	R^2
1:1 matching using nearest neighbor matching												
rs9261027	6	29966000	C/A	0.07477	2.973	0.1907	2.493-5.265	1.49E-11	None	MCCD1P2 / ZNRD1-AS1	Imputed	0.858379
rs112877582	5	3002638	A/G	0.06596	2.973	0.1958	2.403-5.177	1.20E-10	None	RP11-35O7.1 / CTD-2029E14.1	Genotyped	-
rs6873065	5	3003801	C/G	0.07958	2.973	0.1839	2.045-4.204	4.90E-09	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.89416
rs200305090	16	1871961	G/T	0.02605	2.973	0.2587	2.675-7.374	8.20E-09	protein_coding	HAGH	Genotyped	-
rs867569627	5	3000078	CA/C	0.07019	2.973	0.1936	2.034-4.344	1.83E-08	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.86349
rs1355923525	5	3003489	G/GTCCT	0.09929	2.973	0.1706	1.859-3.628	2.22E-08	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.85959
1:2 matching using nearest neighbor matching with discard												
rs9261027	6	29966000	A/C	0.07477	3.253	0.1737	2.314-4.572	1.13E-11	None	MCCD1P2 / ZNRD1-AS1	Imputed	0.858379
rs112877582	5	3002638	G/A	0.06596	3.297	0.1799	2.317-4.691	3.36E-11	None	RP11-35O7.1 / CTD-2029E14.1	Genotyped	-
rs200305090	16	1871961	T/G	0.02605	4.575	0.2365	2.878-7.274	1.28E-10	protein_coding	HAGH	Genotyped	-
rs6873065	5	3003801	G/C	0.07958	2.779	0.1713	1.986-3.887	2.41E-09	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.89416
rs867569627	5	3000078	C/CA	0.07019	2.9	0.1791	2.042-4.12	2.77E-09	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.86349
rs1355923525	5	3003489	GTCCT/G	0.09929	2.461	0.1583	1.805-3.357	1.27E-08	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.85959

A

	Std. MeanDiff						
	Base	N1-1	ND1-1	N1-2	ND1-2	O1-1	O1-2
Case	403	403	269	403	269	403	403
Ctrl	9,613	403	269	806	538	403	806
Sex_F	-1.8147	-0.2042	-0.0180	-0.9730	-0.3014	-0.2042	-0.9730
Sex_M	1.8147	0.2042	0.0180	0.9730	0.3014	0.2042	0.9730
Age	0.0175	0.2849	0.1250	0.0453	0.0833	0.2835	0.0446
BMI	0.8233	0.5586	0.1348	0.1540	0.0625	0.5593	0.1543
Neck_no	-0.3261	-0.2551	-0.0232	-0.1662	0.0637	-0.2551	-0.1662
Neck_yes	0.3261	0.2551	0.0232	0.1662	-0.0637	0.2551	0.1662

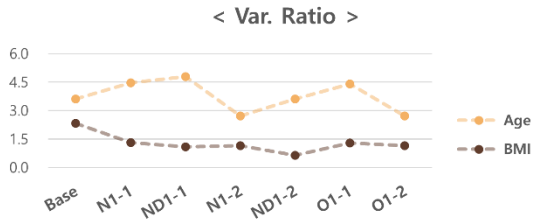
B



C

	Var.Ratio						
	Base	N1-1	ND1-1	N1-2	ND1-2	O1-1	O1-2
Case	403	403	269	403	269	403	403
Ctrl	9,613	403	269	806	538	403	806
Sex_F	.	.	.	.	.	.	.
Sex_M	.	.	.	.	.	.	.
Age	3.6182	4.4646	4.7922	2.7272	3.6103	4.4036	2.7153
BMI	2.3393	1.3075	1.0988	1.1457	0.6406	1.3068	1.1450
Neck_no	.	.	.	.	.	.	.
Neck_yes	.	.	.	.	.	.	.

D



E

	eCDF Mean						
	Base	N1-1	ND1-1	N1-2	ND1-2	O1-1	O1-2
Case	403	403	269	403	269	403	403
Ctrl	9,613	403	269	806	538	403	806
Sex_F	0.7497	0.0844	0.0074	0.4020	0.1245	0.0844	0.4020
Sex_M	0.7497	0.0844	0.0079	0.4020	0.1245	0.0844	0.4020
Age	0.0920	0.4144	0.1279	0.0899	0.1137	0.1330	0.0897
BMI	0.2113	0.3350	0.0395	0.0459	0.0646	0.1449	0.0460
Neck_no	0.1047	0.0819	0.0074	0.0533	0.0204	0.0819	0.0533
Neck_yes	0.1047	0.0819	0.0074	0.0533	0.0204	0.0819	0.0533

F

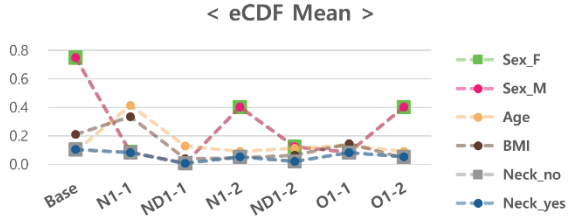


Figure 11. PSM balance performance using sex, age, BMI, and neck as matching variables. The following are matching results using sex, age, BMI, and neck circumference as matching variables. A, C, and E are tables showing matching results according to the matching method used for PSM, and B, D, and F are graphs visualizing the tables. A and B are Std. Mean Diff (standardized mean difference), C and D are Var. Ratio (variance ratio), and E and F are the results for eCDF Mean. The x-axis of the graph represents the different methods used for matching, and the y-axis represents the value of matching measure (Std. Mean Diff, Var. Ratio, and eCDF Mean). Each color coding shows the value of matching measure of the matching variable used for PSM.

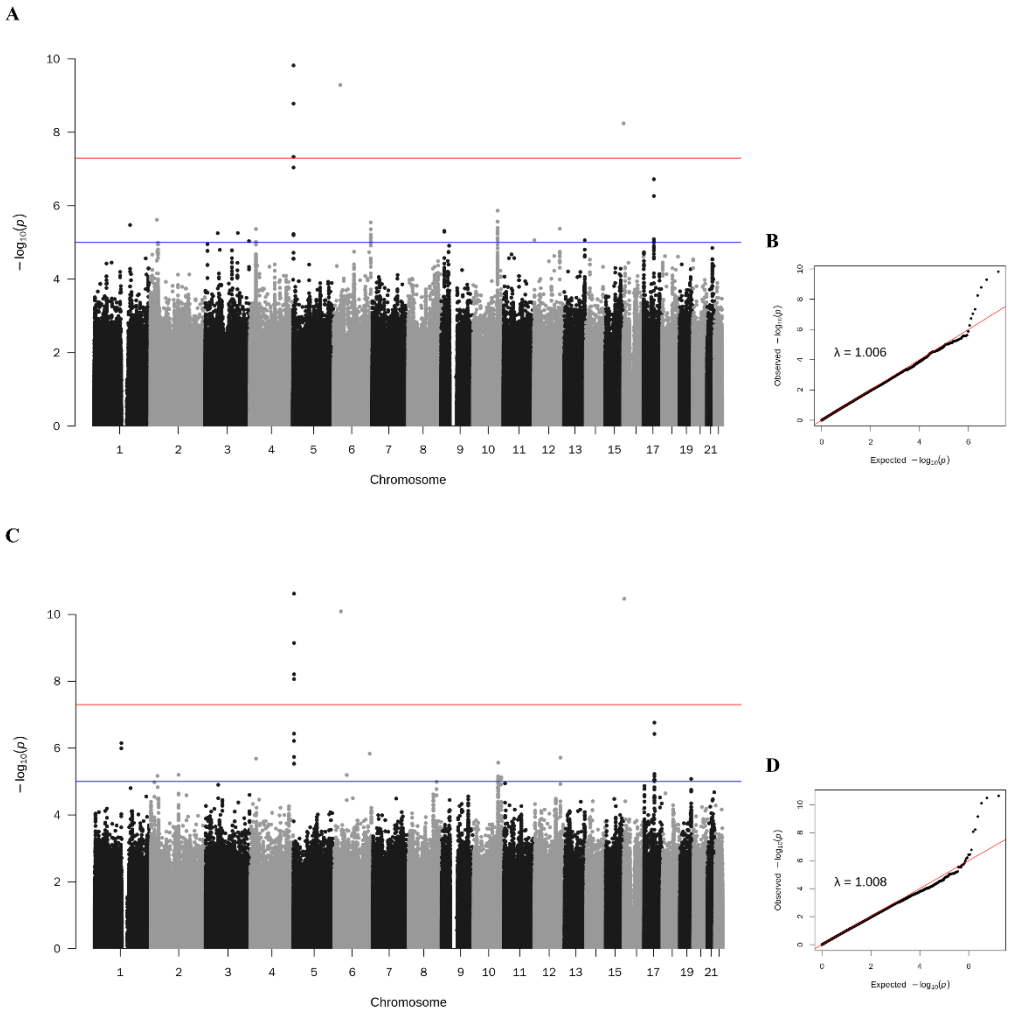


Figure 12. Result of matched GWAS using sex, age, BMI and neck as matching variables. A and B visualize the 1:1 matching results, and C and D visualize the 1:2 matching results. A and C are manhattan plots, where the x-axis represents the chromosome location and y represents –log10 of the P-value. The red line represents the genome-wide significance level, and the blue line represents the genome-wide suggestive level. Each dot represents a variant. Among them, black dots represent significant levels of variation. B and D are QQ-plots.

Table 8. A summary of significant variants of matched GWAS using sex, age, BMI, and neck as matching variables

rsID	CHR	BP	A1/A2	MAF	OR	SE	95% CI	*P* value	Type	Nearest Gene	SNP Type	R^2
1:1 matching using nearest neighbor matching												
rs112877582	5	3002638	A/G	0.06596	3.529	0.1969	2.399-5.191	1.51E-10	None	RP11-35O7.1 / CTD-2029E14.1	Genotyped	-
rs9261027	6	29966000	C/A	0.07477	3.125	0.1833	2.182-4.476	5.12E-10	None	MCCD1P2 / ZNRD1-AS1	Imputed	0.858379
rs6873065	5	3003801	C/G	0.07958	3.093	0.1873	2.142-4.465	1.66E-09	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.89416
rs200305090	16	1871961	G/T	0.02605	4.6	0.2619	2.753-7.686	5.70E-09	protein_coding	HAGH	Genotyped	-
rs867569627	5	3000078	CA/C	0.07019	2.874	0.1932	1.968-4.198	4.63E-08	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.86349
1:2 matching using nearest neighbor matching with discard												
rs112877582	5	3002638	A/G	0.06596	3.428	0.1844	2.388-4.921	2.39E-11	None	RP11-35O7.1 / CTD-2029E14.1	Genotyped	-
rs200305090	16	1871961	G/T	0.02605	5.224	0.2494	3.204-8.518	3.41E-11	protein_coding	HAGH	Genotyped	-
rs9261027	6	29966000	C/A	0.07477	3.085	0.1733	2.196-4.333	8.11E-11	None	MCCD1P2 / ZNRD1-AS1	Imputed	0.858379
rs6873065	5	3003801	C/G	0.07958	2.961	0.1762	2.097-4.182	7.16E-10	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.89416
rs1355923525	5	3003489	G/GTCCT	0.09929	2.546	0.1608	1.858-3.489	6.18E-09	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.85959
rs867569627	5	3000078	CA/C	0.07019	2.842	0.1815	1.991-4.057	8.63E-09	None	RP11-35O7.1 / CTD-2029E14.1	Imputed	0.86349

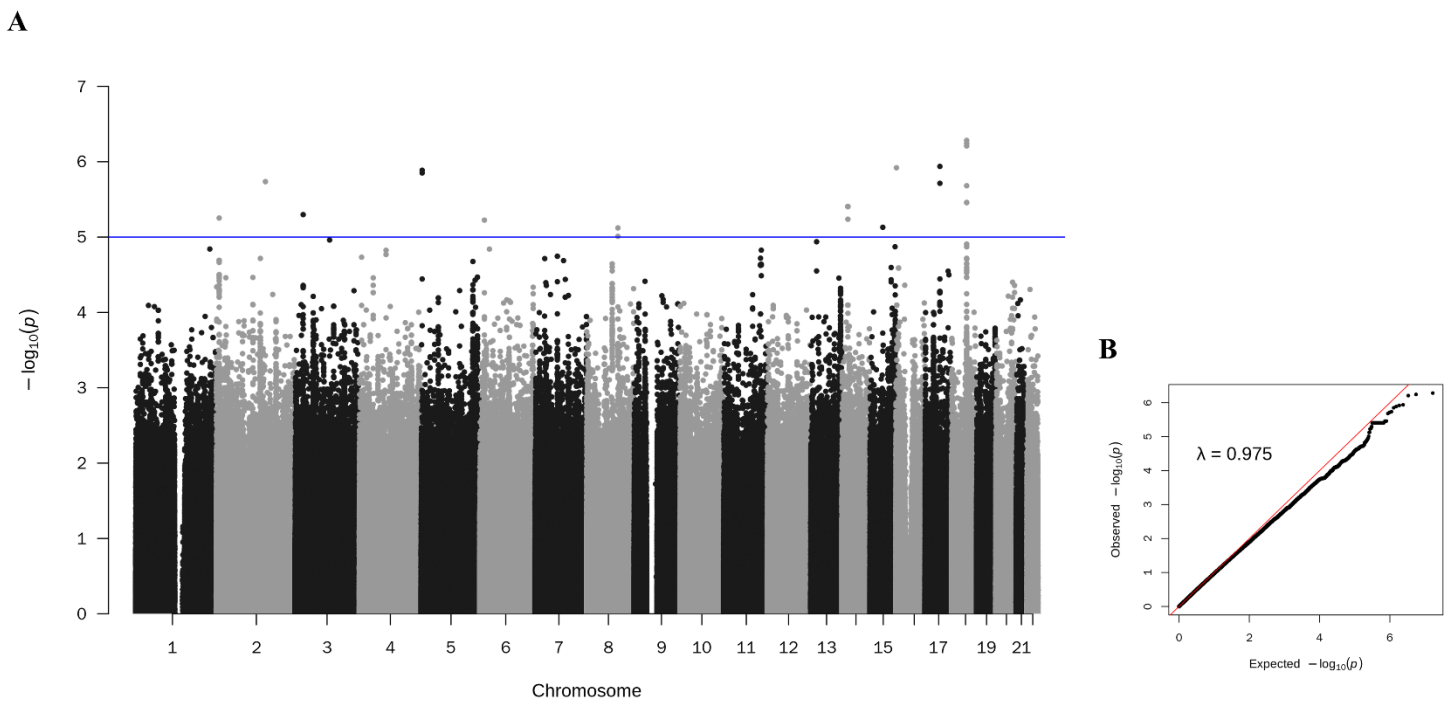


Figure 13. Result of sex-stratified GWAS. A is manhattan plot. The x-axis represents the chromosome location and y represents –log10 of the P-value. The blue line represents the genome-wide suggestive level. Each dot represents a variant. B is QQ-plot.

Table 9. A summary of lead variants of sex-stratified GWAS

rsID	CHR	BP	A1/A2	MAF	OR	SE	95% CI	*P* value	Type	Nearest Gene	SNP Type	R^2
rs11659301	18	47362880	C/T	0.2717	2.319	0.1676	1.67-3.221	5.21E-07	protein_coding	SCARNA17, MYO5B	Imputed	0.98398
rs1426122093	17	45636413	G/GA	0.2349	0.3881	0.1946	0.265-0.5683	1.15E-06	protein_coding	NPEPPS	Imputed	0.80357
rs200305090	16	1871961	G/T	0.07283	5.587	0.3543	2.79-11.19	1.20E-06	protein_coding	HAGH	Genotyped	0.81048
rs6873065	5	3003801	C/G	0.1285	3.424	0.2543	2.08-5.636	1.30E-06	lincRNA	RP11-35O7.1/CTD-2029E14.1	Imputed	0.89416
rs112877582	5	3002638	A/G	0.1154	3.563	0.2634	2.126-5.971	1.41E-06	lincRNA	RP11-35O7.1/CTD-2029E14.1	Genotyped	0.90426
rs6746068	2	151384942	T/G	0.338	0.4502	0.1673	0.3243-0.6249	1.84E-06	protein_coding	RND3	Imputed	0.8557
rs10143001	14	38999670	A/G	0.08265	3.943	0.2972	2.202-7.06	3.93E-06	lincRNA/pseudogene	RP11-96D24.1/RP11-14N4.1	Imputed	0.9819
rs10498342	14	39001036	T/C	0.08265	3.943	0.2972	2.202-7.06	3.93E-06	lincRNA/pseudogene	RP11-96D24.1/RP11-14N4.1	Genotyped	0.99757
rs12490340	3	24706758	C/T	0.0982	3.432	0.2702	2.021-5.828	5.03E-06	pseudogene/lincRNA	EIF3KP2/AC133680.1	Imputed	0.96255
rs12471031	2	8734140	C/T	0.4403	0.4762	0.1634	0.3457-0.6559	5.58E-06	lincRNA	AC011747.4/AC011747.6	Imputed	0.90196
rs13437378	6	14507185	C/T	0.1506	0.3912	0.2073	0.2606-0.5873	5.95E-06	lincRNA	RP3-448I9.1/RP11-330A16.1	Imputed	0.91942
rs201595233	15	62110500	A/AT	0.02537	9.815	0.5096	3.615-26.65	7.41E-06	miRNA/antisense	AC018618.1/RP11-16B9.1	Imputed	0.98569
rs10109270	8	96410843	C/T	0.3617	0.4629	0.172	0.3304-0.6485	7.55E-06	lincRNA	KB-1047C11.2	Imputed	0.81622

III-5. Genetic (non-anatomical) effect of the identified loci over age

The results of an analysis performed to determine the influence of anatomical and non-anatomical factors affecting OSA according to age are visualized in Figure 14. The analysis results using the KoGES cohort data are shown in panel A. BMI and neck circumference were used as anatomical factors, and 45 significant variants found from continuous GWAS were used as non-anatomical factors. As a result of the analysis, it was confirmed that the influence of non-anatomical factors gradually increases compared to BMI as one ages from the young to the elderly, but the anatomical factor, neck circumference, still has a high influence. The analysis results using CNUH dataset are visualized in panels B and C. As a result of checking the explanatory power of AHI and STOP-BANG using the PRS of the anatomical factor, BMI, and the non-anatomical factor, 45 significant variants from continuous GWAS, it was confirmed that the influence of the non-anatomical factor clearly increases with age.

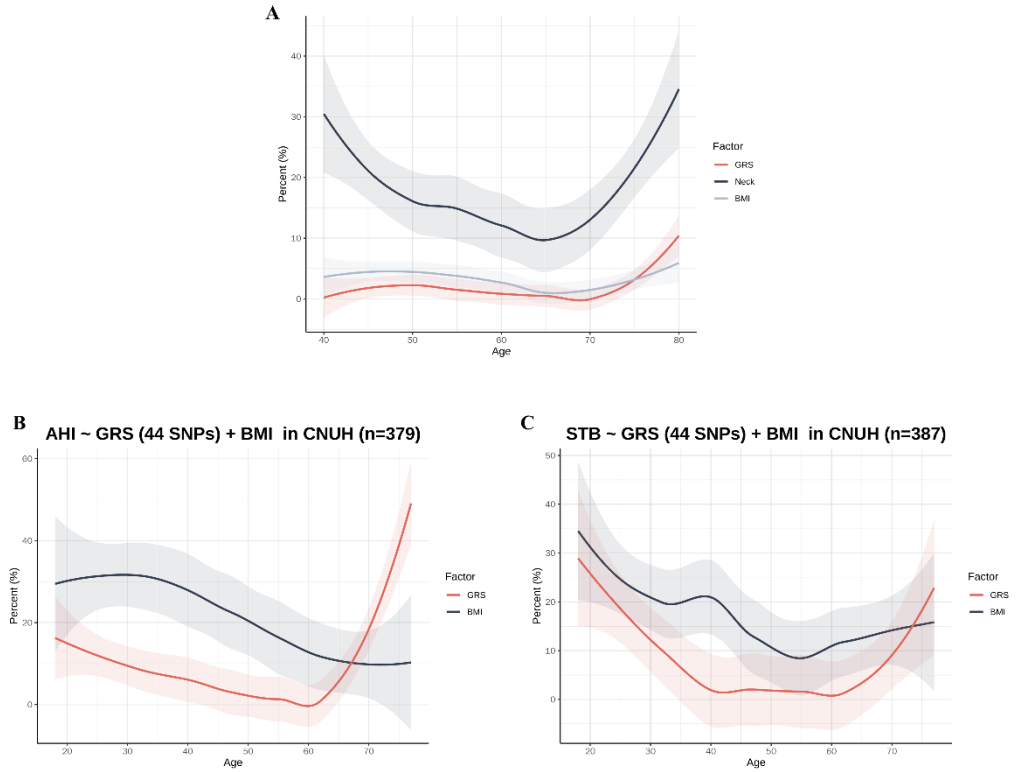


Figure 14. Effect of OSA risk factors over age. This figure shows the effect of anatomical and non-anatomical factors on OSA over age. Neck circumference (Neck) and BMI were used as anatomical factors, and PRS was used as a non-anatomical factor. A is the result using the KoGES dataset, and B and C are the results using the CNUH dataset.

IV. DISCUSSION

We performed OSA GWAS on a dataset of n=25,712 KoGES participants. Furthermore, matched GWAS of CNUH and KoGES was performed using a total of 10,016 individuals using additional clinical data. We were able to see genetic evidence for OSA by discovering and cloning genetic loci associated with OSA, and identified loci that had not yet been reported. We were able to confirm the genetic role in OSA, by confirming that PRS based on continuous GWAS is associated with OSA even in independent datasets. We also confirmed the OSA explanatory power of each anatomical and non-academic factor over age. In particular, among the genetic variants associated with OSA discovered from continuous GWAS, we focused on rs4680142 and rs1915714, located around the MME gene replicated from FinnGen GWAS, and rs6718994, located around the DYSG gene, which was another significant variant although not replicated. A disease related to the MME gene is spinocerebellar ataxia, which is one of the cerebellar disorders and is a disease that shows gradual incoordination of gait and movement due to degeneration of the cerebellum as well as the brainstem and spinal cord. Expression of this gene was confirmed in the brain, especially the basal ganglia, and previous studies have reported that the basal ganglia and brainstem system can contribute to awake-sleep state regulation through involvement of the cerebellum [23]. The DYSF gene is associated with diseases such as muscular dystrophy, Miyoshi muscular dystrophy, and limb griddle muscular dystrophies. These diseases are muscle-related diseases, starting with weakness of the distal muscles and causing anatomical weakness of the pharyngeal muscles, and have been reported to be disease with a high risk of OSA [24-26]. Based on function of different patterns associated with OSA by each gene, which located the identified variants and previous study that OSA risk varies with age, it was possible to hypothesize that anatomical and non-anatomical factors would apply differently depending on age. As a result of analyzing this using our data, we were able to actually confirm the trend in Figures 14.

Our study has several strengths. First, the sample used in this study was obtained from a large-scale single ethnic group. Genetic study of OSA has mainly been conducted in other ethnic groups, and as far as we know, this study used the largest amount of data among OSA association studies in Asian ethnic groups. Therefore, by using a large sample size in genetic research on OSA, which has polygenic characteristics, statistical power was improved and meaningful results were obtained. Second, by assessing OSA using STOP-BANG, we were able to overcome the time and cost limitations of PSG and use a large number of participants. Additionally, it was possible to overcome the problem of underdiagnosis of OSA due to dichotomous diagnosis. Third, more accurate results could be obtained by using a case diagnosed using hospital-based PSG as well as STOP-BANG. However, this study has several limitations. First, since only the Korean population was used, caution should be exercised in interpreting the identified variants in other races. Second, the sample size of the elderly is insufficient to test the hypothesis.

In conclusion, we identified genetic candidates associated with OSA using large-scale Korean data. Because it is widely known that overall AHI is likely to reflect heterogeneous phenotypes, STOP-BANG allowed us to evaluate more specific OSA phenotypes. Additionally, it was possible to propose a hypothesis about the progression of OSA from the identified candidate genes

V. 초록

대규모 유전체-설문 매칭을 이용한
수면장애 연관 유전체분석

윤 다 빈

지도교수 : 김 정 수

글로벌바이오융합학과

조선대학교 대학원

폐쇄성 수면무호흡증은 여러 위험요인이 작용하지만 그 중 유전 변이가 중요한 역할을 하는 복합질병이다. 그러나 이전 유전 연구들은 폐쇄성 수면무호흡증에 대한 낮은 인지로 인해 질병군이 대조군으로 오분류됨으로써 발생하는 과소진단 문제로 표본 크기의 한계를 겪게 되었고 유의미한 유전자좌의 식별에 어려움을 겪었다. 따라서 본 연구는 대규모 한국인 표본으로부터 STOP-BANG 설문지를 사용하여 연속적 규모로 폐쇄성 수면무호흡증을 정의하고 폐쇄성 수면무호흡증에 대한 전장유전체연관성분석를 수행하였다.

40세 이상의 한국인 성인 남녀 72,219명으로 구성된 한국인유전체역학 조사사업으로부터 유전적으로 동질적인 25,712명을 선별하여 전장유전체연관성-분석을 수행하였다. 폐쇄성 수면무호흡증의 중증도를 측정하기 위해, 대표적인 폐쇄성 수면무호흡증 선별도구 중 하나인 STOP-BANG 설문지를 사용하여 폐쇄성 수면무호흡증의 중증도를 0점부터 8점까지 정의하였다. 전장유전체연관성분석

결과의 복제를 위해 서양인 자료인 FinnGen을 사용한 수면 무호흡증 전장유전체-연관성분석 요약 통계를 사용하였으며, 발굴된 유전변이에 대한 추가 검증을 위해 전장유전체연관성분석에 사용되지 않은 독립적인 검증 데이터세트를 사용하여 폐쇄성 수면무호흡증 관련 변수에 대한 다유전자 위험점수를 평가하였다.

전장유전체연관성분석 결과, 9개의 유의한 유전자좌(45개의 유전변이)를 식별하였다. 이 중 8개의 유전자좌는 폐쇄성 수면무호흡증 또는 STOP-BANG과 관련하여 이전에 보고되지 않았지만, 2개의 유전변이가 위치하는 1개 유전자좌(MME)는 FinnGen 연구의 일반적인 유의 수준(nominal significance level)에서 복제가 확인되었다. 발견된 45개의 유전변이에 대한 검증을 위해 전장유전체연관성분석에 사용된 발견 데이터세트와 독립적인 검증 데이터세트를 사용하여 수면 관련 특성과 폐쇄성 수면무호흡증에 대한 유전변이의 다유전자 위험점수를 비교하였고, 그의 타당성을 추가로 확인할 수 있었다.

본 연구는 폐쇄성 수면무호흡증과 관련된 여러 유전자좌를 발견하였다. 결과적으로 폐쇄성 수면무호흡증과 관련된 MME 유전자의 1개 유전자좌와 동아시아인 특이적인 후보로 8개의 유전자좌를 확인하였다.

VI. REFERENCES

1. Young, T., J. Skatrud, and P.E. Peppard, *Risk factors for obstructive sleep apnea in adults.* Jama, 2004. **291**(16): p. 2013-2016.
2. Franklin, K.A. and E. Lindberg, *Obstructive sleep apnea is a common disorder in the population—a review on the epidemiology of sleep apnea.* Journal of thoracic disease, 2015. **7**(8): p. 1311.
3. Senaratna, C.V., et al., *Prevalence of obstructive sleep apnea in the general population: a systematic review.* Sleep medicine reviews, 2017. **34**: p. 70-81.
4. Villaneuva, A.T.C., et al., *Ethnicity and obstructive sleep apnoea.* Sleep medicine reviews, 2005. **9**(6): p. 419-436.
5. Benjafield, A.V., et al., *Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis.* The Lancet Respiratory Medicine, 2019. **7**(8): p. 687-698.
6. Bixler, E.O., et al., *Prevalence of sleep-disordered breathing in women: effects of gender.* American journal of respiratory and critical care medicine, 2001. **163**(3): p. 608-613.
7. Quintana-Gallego, E., et al., *Gender differences in obstructive sleep apnea syndrome: a clinical study of 1166 patients.* Respiratory medicine, 2004. **98**(10): p. 984-989.
8. Resta, O., et al., *Gender, age and menopause effects on the prevalence and the characteristics of obstructive sleep apnea in obesity.* European journal of clinical investigation, 2003. **33**(12): p. 1084-1089.
9. Strausz, S., et al., *Genetic analysis of obstructive sleep apnoea discovers a strong association with cardiometabolic health.* European Respiratory Journal, 2021. **57**(5).
10. Redline, S., et al., *The familial aggregation of obstructive sleep apnea.* American journal of respiratory and critical care medicine, 1995. **151**(3): p. 682-687.
11. Mukherjee, S., R. Saxena, and L.J. Palmer, *The genetics of obstructive sleep apnoea.* Respirology, 2018. **23**(1): p. 18-27.
12. Szily, M., et al., *Genetic influences on the onset of obstructive sleep apnoea and daytime sleepiness: a twin study.* Respiratory research, 2019. **20**(1): p. 1-6.
13. Sofer, T., et al., *Genome-wide association study of obstructive sleep apnoea in the Million Veteran Program uncovers genetic heterogeneity by sex.* EBioMedicine, 2023. **90**.
14. Xu, H., et al., *Genome-wide association study of obstructive sleep apnea and objective sleep-related traits identifies novel risk loci in Han Chinese individuals.* American Journal of Respiratory and Critical Care Medicine, 2022. **206**(12): p. 1534-1545.
15. Motamedi, K.K., A.C. McClary, and R.G. Amedee, *Obstructive sleep apnea: a growing problem.* Ochsner Journal, 2009. **9**(3): p. 149-153.
16. Finkel, K.J., et al., *Prevalence of undiagnosed obstructive sleep apnea among adult surgical patients in an academic medical center.* Sleep medicine, 2009. **10**(7): p. 753-758.
17. Kapur, V., et al., *Underdiagnosis of sleep apnea syndrome in US communities.* Sleep and Breathing, 2002. **6**(02): p. 049-054.
18. Kim, Y., B.-G. Han, and K. Group, *Cohort profile: the Korean genome and epidemiology study (KoGES) consortium.* International journal of epidemiology, 2017. **46**(2): p. e20-e20.
19. Chiu, H.-Y., et al., *Diagnostic accuracy of the Berlin questionnaire, STOP-BANG, STOP, and Epworth sleepiness scale in detecting obstructive sleep apnea: a bivariate meta-analysis.* Sleep medicine reviews, 2017. **36**: p. 57-70.
20. Pivetta, B., et al., *Use and performance of the STOP-Bang questionnaire for obstructive sleep apnea screening across geographic regions: a systematic review and meta-analysis.*

JAMA network open, 2021. **4**(3): p. e211009-e211009.
21. Chung, F., H.R. Abdullah, and P. Liao, *STOP-Bang questionnaire: a practical approach to screen for obstructive sleep apnea.* Chest, 2016. **149**(3): p. 631-638.
22. Romero-Corral, A., et al., *Interactions between obesity and obstructive sleep apnea: implications for treatment.* Chest, 2010. **137**(3): p. 711-719.
23. Takakusaki, K., et al., *Role of basal ganglia–brainstem pathways in the control of motor behaviors.* Neuroscience research, 2004. **50**(2): p. 137-151.
24. Halani, P., et al., *793 Obstructive Sleep Apnea and Hypoventilation in Patients with Muscular Dystrophies: Analysis of 42 Polysomnogram Studies.* Sleep, 2021. **44**: p. A309.
25. Hoque, R., *Sleep-disordered breathing in Duchenne muscular dystrophy: an assessment of the literature.* Journal of Clinical Sleep Medicine, 2016. **12**(6): p. 905-911.
26. Merrison, A.F., *Myopathies in the adult patient.* Medicine, 2020. **48**(9): p. 619-624.