Phenome-wide Association Study of Cystic Fibrosis Modifier Genes

Student's Name
Supervised by Supervisor's Name

Date

Contents

1 Abstract

Background: Cystic fibrosis (CF) is the most common fatal genetic disease affecting Canadian children and young adults. At present, there is no cure. Further, non-CF genes have been identified that affect the severity of the symptoms of CF. These genes are called modifier genes. Our goal is to study the effects of three such modifier genes in the general public, i.e. people without CF.

Methods: A PheWAS study was performed for the following modifier genes and their respective SNPs: SCL26A9 and rs4077468 on chromosome 1, SLC6A14 and rs3788766 on Chromosome X and lastly, SCL9A3 and rs57221529 on Chromosome 5. We used the UK Biobank (UKBB) data registry, which has about 500,000 participants, to perform the PheWAS study using phenotypes converted from the ICD-10 codes in the UKBB. A multiple logistic regression was used for finding associations between phenotypes and the modifier genes with phenotypes modeled as binary outcomes and the minor allele count as the predictor (e.g. minor allele T count for rs4077469 C/T, being either 0, 1 or 2). The allele count was modeled as an additive model. Adjusted and unadjusted analysis were both conducted The model was adjusted for covariates age, age-squared and sex. Since, there were 1511 phenotypes, 1511 logistic regressions were performed for each gene/SNP and a Bonferroni correction was applied to determine statistically significant associations giving a corrected significance threshold of 3.309x10^{-5}. Further, separate analysis was performed for males and females for the gene SLC6A14 as it is found on the X chromosome. Lastly, related people were removed from the data, and our analysis was further restricted to only individuals identified as Caucasians through ancestral PCA performed in the UKBB.

Findings: For SCL9A3 and SNP rs57221529, we found it to be significantly associated with having Esophagitis, GERD and related diseases (OR = 1.064, S.E. = 0.013, p-value = 1.79x10^{-6}). Further, for each allele count (0, 1, 2) the number of cases were 12403, 6500 and 784 the controls were 157647, 76284 and 9305, respectively. For males, the gene SLC6A14 for SNP rs3788766 was significantly associated with having urinary obstruction (OR = 1.68, S.E. = 0.127, p-value = 4.24x10^{-5}). Further, for each allele count (0 or 2 as males cannot have 1 allele as this gene is on the X chromosome) there were only a total of 64 cases and 117,334 controls, with 27 cases for allele count 0 vs. 37 for allele count 1. For females no statistically significant association was found. Lastly, for the modifier gene SLC26A9 for SNP rs4077469, no association with any of the phenotypes was found.

Interpretation: For the gene SCL9A3 and SNP rs57221529, C is the risk allele with OR = 1.064 which means that with every additional C allele the odds of having Esophagitis, GERD and related diseases increases by 6.4%. Further, since there are a significant amount of cases and controls for each allele count this suggests that this is an actual association and not a sporadic one. For males even though the gene SLC6A14 was found to be statistically significantly associated with urinary obstruction, there were only 64 cases in total versus 117334 control suggesting this can possibly be a sporadic relationship and not a real association. Hence, we would conclude that our data suggests that among Caucasians there is an association between the gene SCL9A3 and the SNP rs57221529 with having Esophagitis, GERD and related diseases.

key words: Cystic fibrosis, UK Biobank, modifier genes, SCL26A9, SLC6A14, SCL9A3

2 Introduction

2.1 Cystic Fibrosis

Cystic fibrosis (CF) is the most common fatal genetic disease affecting Canadian children and young adults. It is caused by mutations of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. It causes various effects on the body, but mainly affects the digestive system and lungs such as dysfunction of the lungs, sweat glands, vas deferens, and pancreas (Cohn et al. 1998). In 2008, the average life span for people with CF who lived to adulthood was about 37, mainly due to diseases of the lung (Cohen-Cymberknoh, Shoseyov, and Kerem 2011) and is a disease that is primarily found in people with Caucasian ancestry. To get CF the individuals need to inherent a faulty CF gene from each parent. Hence, CF is a Mendelian disease.

2.2 Modifier Genes

There have been identified common variation in several genes (genes that themselves have nothing to do with CFTR gene) that contribute to CF disease severity. These genes are known as CF modifier genes and the majority of these genes are transporters. The effects of these modifier genes are well studied for people with CF. However, it is unknown whether variation in these genes impact phenotypes (i.e. all physical and observable characteristics) of individuals who do not have CF, that is, individuals without two mutations in the CF causal gene.

2.3 Single-Nucleotide Polymorphism (SNP)

Single-nucleotide polymorphisms (SNPs) are a common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. A SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA as illustrated below in Figure 1. Moreover, at a specific base position in the human genome, the C nucleotide may appear in most individuals, but in a minority of individuals, the position is occupied by an A. This means that there is a SNP at this specific position, and the two possible nucleotide variations – C or A – are said to be alleles for this position (Uppu, Krishna, and Gopalan 2018). Please see Figure 1 below for an illustration.

Some SNPs have been found to be responsible for the differing susceptibility of diseases in individuals (e.g. sickle-cell anemia, and cystic fibrosis result from SNPs). The severity of illness and the way the body responds to treatments are also manifestations of genetic variations.

The primary goal of this study is to see the effect of genetic variation in SNPs of interest in the general public. We do this by carrying out a Phenome-wide association study (PheWAS).

2.4 PheWAS

A PheWAS correlates the genetic variants of interest with every possible phenotype measured to characterize the clinical impact across the body system of these genes. For example, in a PheWAS study, the SNP rs3135388 in the gene HLA-DRA was found to be associated with multiple sclerosis (R. J. Carroll, Bastarache, and Denny 2014). Due to the large number of associations tested, a Bonferroni correction is applied to identify statistically significant associations.

In our study we perform a PheWAS using 500,000 individuals from the UK Biobank who have been genotyped genome-wide and have detailed, comprehensive phenotypic data.

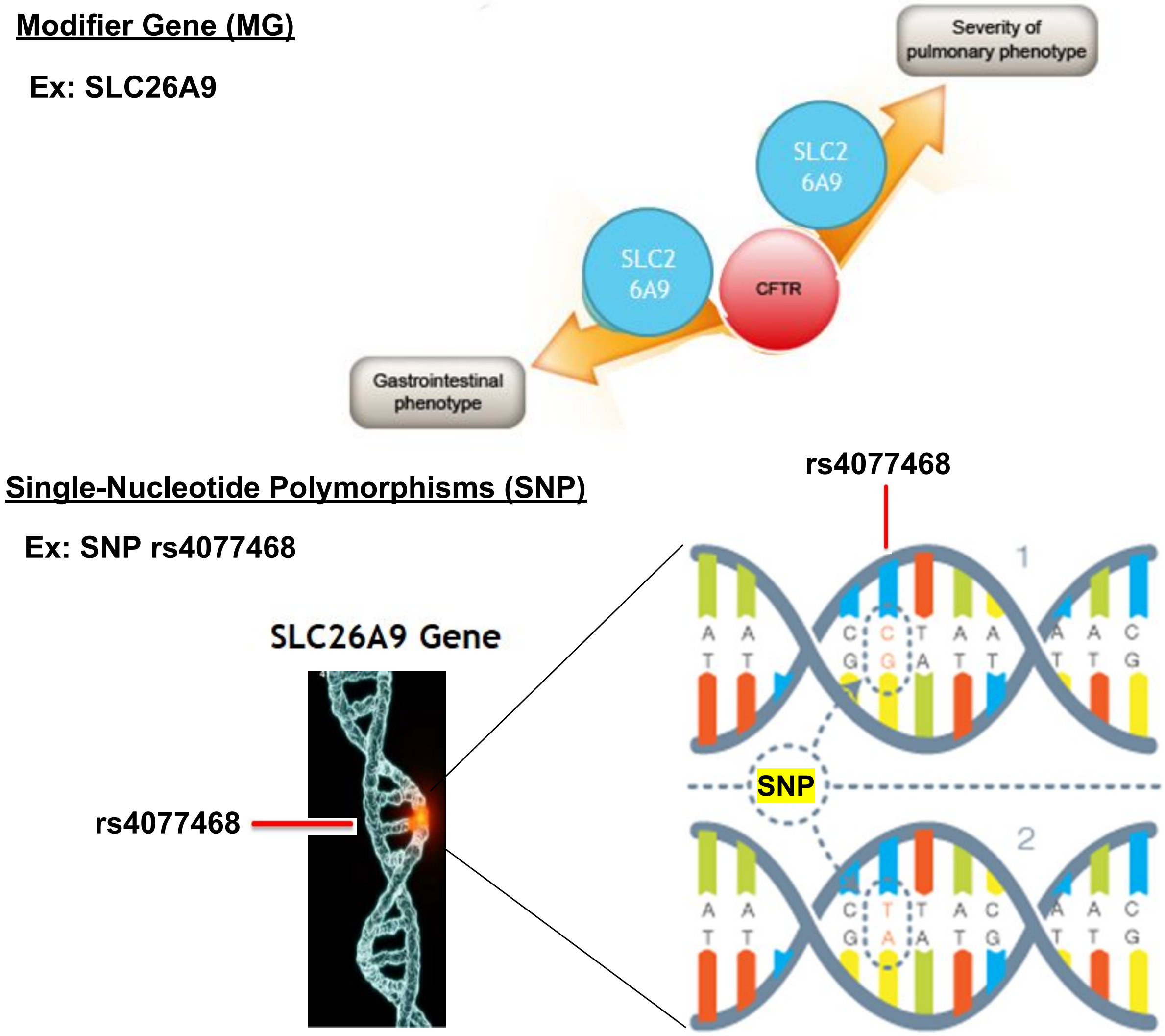

Figure 1: Illustrating modifier genes and single-nucleotide polymorphisms

2.5 International Classification of Diseases (ICD)

The International Classification of Diseases (ICD) “is the international standard diagnostic tool for epidemiology, health management and clinical purposes” and is stored in electronic health records (EHRs) from when patients visit hospitals and clinics. It was established and is maintained by the World Health Organization (WHO) to track morbidity and mortality statistics across the world. This allows researchers to “analyze hundreds of human diseases, drug responses, and many observable clinical traits” in a standardized method (Wei et al. 2017). However, ICD codes are not organized meaningfully with proper disease groupings, also referred to as chapters, for the purposes of a PheWAS. For instance, malignant neoplasm of breast (ICD9 code: 174.9) and personal history of malignant neoplasm of breast (ICD9 code: V10.3) both are in different chapters (Wei et al. 2017). Hence, enters the phecode system.

2.6 Phecodes

The phecode system was built upon the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) for phenome-wide association studies (Wu et al. 2018). It was created for proper hierarchical grouping of phenotypes, and for high throughput.

Phecodes are grouped into seventeen phenotypic groupings:

- Neoplasms
- Genitourinary
- Dermatologic
- Infectious diseases
- Endocrine/metabolic
- Neurological
- Musculoskeletal
- Congenital anomalies
- Sense organs
- Injuries & poisonings
- Circulatory system
- Digestive
- Respiratory
- Mental disorders
- Pregnancy complications
- Symptoms
- Hematopoietic

For a PheWAS, an important thing to keep in mind is that when performing statistical analysis and clinical studies using phenotypic data, one not only needs to define cases but also the controls. That is, rather than simply using all the people who do not have the phenotype serving as controls, exclusions need to be applied based on the phenotype being analyzed. For example, studies of type 2 diabetes often exclude subjects with any other type of diabetes from the control group (Wei et al. 2017).

However, ICD codes do not have any "ready-made approach to automatically exclude populations of patients that are "possible" cases or have similar or potentially overlapping disease states (e.g., type 1 diabetes and secondary diabetes mellitus)" (Wei et al. 2017). Hence, this is another reason ICD codes are translated into phecodes for performing PheWAS.

In general, phecodes have been successfully used in a number of PheWAS to replicate hundreds of known genetic associations and discovered new ones.

2.7 Study Goal Summary

Ultimately, by performing PheWAS for general population, we will have an improved understanding of the phenotype associated with normal variation in these genes of interest in the general population; genes which, with a background of CFTR mutations, can cause severe disease. Understanding the impact of these variants in a normal CFTR background may also suggest milder CF-related phenotypes not previously appreciated, as well as alternative uses for therapeutics that are designed to target these genes.

3 Material and Methods

3.1 Data

3.1.1 UK Biobank

The UK Biobank (UKBB) is a large-scale, population-based, prospective cohort that enrolled over 500,000 participants between the ages 40 and 69 years across the UK; primarily from England, Scotland and Wales. All participants were recruited on a volunteer bases. Initial enrollment took place over four years from 2006. Extensive amounts of data was collected, and is continued to be collected, for each recruited participant using a variety of methods such as questionnaires, physical measures, sample assays, accelerometry and multimodal imaging among others. Hence, for each recruited participants there is a wide range of baseline information. Blood samples are also collected for biochemical tests and genotyping. Individual's national health records have also been linked with their baseline and genotypic data obtaining extensive healthcare information on the participants, allowing for longitudinal follow-up (Sudlow et al. 2015). Genotypic and phenotypic data used in this study were obtained from UKBB under an approved data request application.

3.1.2 Genotypic data

For the purpose of this study we are interested in three modifier genes: SCL26A9, SLC6A14 and SCL9A3 and their SNPs rs4077468, rs57221529 and rs3788766, respectively. However, the SNPs we are interested in were not present in the UKBB. Hence, instead of imputing the SNPs, we decided to use substitute SNPs that are highly correlated with the original SNPs present in UKBB. We chose to take this route because we found SNPs that had a correlation between 77%-100% (see Table 1) and we thought this would be more reliable than imputing the SNPs. Detailed information on the genes and SNPs of interest is presented in Table 1 below.

Table 1: Genes with the original and substitute SNP's and their correlation.

Gene	SNP of Interest	Chromosome	Substitute SNP	Correlation
SCL26A9	rs4077468	Chromosome 1	rs4077469	r = 1
SCL9A3	rs57221529	Chromosome 5	rs17497684	r = 0.821
SLC6A14	rs3788766	Chromosome X	rs5905176	r = 0.770

Further, it is important to mention that the chosen SNPs only serve as a representative of the region in the gene. If an association is found between a SNP and a phenotype, that would mean there is some association between that region in the gene and the phenotype.

3.1.3 Phenotypic Data and Mapping ICD Codes to Phecodes

For our particular data set and the variables we had requested from UKBB, the phenotypic data had almost 14,000 variables. Among the variables, one variable was the individuals' ICD-9 codes and another was the individuals' ICD-10 codes.

For our primary analysis, we focused only on phenotypic data obtained from the ICD-10 codes in the UKBB, i.e. on phenotypes in relation to diagnostic disease outcomes (either an individual had or did not have a disease). We will only be using ICD-10 codes because most of the hospitals in the UK transitioned to ICD-10 codes, leaving very few people in the early recruitment period of the study with ICD-9 codes (X. Li et al. 2018) and also because of the purposes of simplicity.

Even though the phecode system was built upon the ICD-9 revision, studies have been done to validated

ICD-10 mapping to phecode for the UKBB data for the purposes of performing PheWAS (Wu et al. 2018). Therefore, if ICD-10 codes are properly and reasonably mapped to phecodes then they should be able to properly perform PheWAS with very reliable results. In our study the mapping of ICD-10 codes to phecodes was done by the Strug Lab.

Lastly, as mentioned before, for a PheWAS, one would want to define both cases and controls for a phenotype so that one can implement an exclusion scheme when defining controls for the analysis for particular phenotypes (e.g. when analyzing the phenotype of type 2 diabetes excluding all people with any other type of diabetes from the control group). However, when conducting the PheWAS, the software we used (R "PheWAS" package, see the Software Section for details) only had the capability of implementing the exclusion scheme for phecodes mapped from ICD-9 codes, not ICD-10. Hence, in our study no exclusion scheme was implemented and controls were defined as everyone who did not have the phenotype.

3.1.4 Data Cleaning and Preperation for Analysis

A lot of time and effort was spent on data cleaning and quality control before performing any analysis. For brevity, only the overall data processing method is mentioned here, along with a few important points.

For each SNP, the genotypic data was obtained from the UKBB data set, which also included data on the person's sex. The person's sex was also obtained directly from their genotypic data using the X and Y chromosomes. This was done using PLINK. These two were compared to make sure there were no misidentifications. For each SNP, all individuals' sexes matched.

The data cleaning process was the same for each SNP: After obtaining the genotypic data for the SNP of interest, all the individuals who were related up to 2$^{\text{nd}}$ degree of relatedness (up to siblings) were removed. Further, since CF primarily affects people with Caucasian ancestry, the data was further restrained to only individuals confirmed to be Caucasians based on the ancestral principal component analysis performed by the UKBB. Lastly, this data set was merged with the phenotypic data obtained from the ICD-10 codes. This is illustrated for gene SLC26A9 and SNP rs4077469 in Figure 2 below.

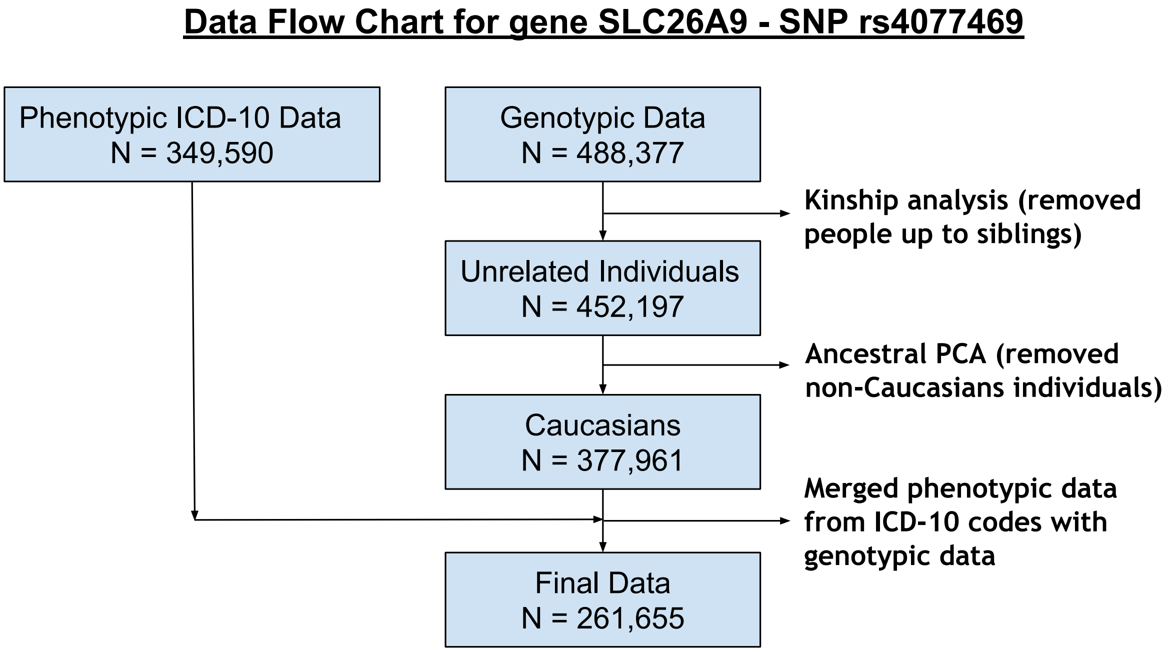
Data Flow Chart for gene SLC26A9 - SNP rs4077469
Phenotypic ICD-10 Data
N = 349,590
Genotypic Data
N = 488,377
Kinship analysis (removed people up to siblings)
Unrelated Individuals
N = 452,197
Ancestral PCA (removed non-Caucasians individuals)
Caucasians
N = 377,961
Merged phenotypic data from ICD-10 codes with genotypic data
Final Data
N = 261,655

Figure 2: Data Flow Chart for gene SLC26A9 - SNP rs4077468

Using the above data cleaning process generated a final data set for the genes ***SCL9A3***, ***SCL26A9*** and ***SLC6A14*** with **262,923**, **261,655** and **117,398** individuals, respectively.

Further, as can be surmised, all and any missing data was deleted. For example, many people with ICD-10 codes in the UKBB data did not have genotypic data for different genes/SNPs. Those individuals were simply deleted. Similarly, there were individuals with genotypic data but no ICD-10 phenotypic data. These were also deleted. This in fact may be a bit troublesome since people without any ICD-10 data may simply mean that they did not visit the hospital before we received the UKBB data.

As can be seen the sample size of the data sets were not based on any power calculation but on the available data sets.

3.2 Statistical Methodology

We performed a PheWAS using the ICD-10 phenotypic data for the three modifier genes of interest. Both, an unadjusted and adjusted analysis with covariates age, age-squared and sex, were performed.

3.2.1 PheWAS Study:

A PheWAS study tests the association between genetic variants of interest with every phenotype measured. They are cross-sectional studies.

The outcome variable is the the person's phenotype. The predictor variable is the allele variants for the SNP (e.g. AA, AT, TT for rs4077469). For binary type ICD-10 phenotypes, the cases were defined as all the people with the particular ICD-10 code, and controls were defined as everyone else. As mentioned before and to reiterate, no exclusion criteria was implemented.

Since the outcome variables are binary, a multiple logistic regression was employed as the underlying model for testing association between a SNP and the phenotypes.

Defining the Logistic Regression Model:

First, we define the outcome variable $Y_j^{phenotype_i}$. It is one if the jth individual has phenotype$_i$.

$$Y_j^{phenotype_i} = \begin{cases} 1 & phenotype_i = 1 \\ 0 & phenotype_i = 0 \end{cases}$$

where:

i = 1, . . . , 1511; is the ith phenotype, and

j = 1, . . . , N; is the jth individuals in the data set

and

$$E(Y_j^{phenotype_i}) = p_{\_phenotype_i}$$

where

$p_{\_phenotype_i}$ is the probability of having phenotype i.

The logistic regression equation for the adjusted analysis is then:

$$logit(p_{\_phenotype_i}) = SLC26A9 + age + age^2 + sex$$

where

$$sex = \begin{cases} 1 & female \\ 0 & male \end{cases}$$

Now there are two ways of modeling predictor variable SLC26A9.

Additive Model

In the additive model, SLC26A9 is equal to the number of T alleles at SNP rs4077469. Our assumption in this model is that whatever the effect of having one T allele is, it is doubled when two T alleles are present. The biological reasoning is that the SNPs we are studying are introns (noncoding sections of DNA) and the role they play is that they control the amount of protein that the RNA produces. If one T allele produces X amount of protein then two T alleles will produce 2X.

$$SLC26A9 = \begin{cases} 0 & rs4077469 = AA \\ 1 & rs4077469 = AT \\ 2 & rs4077469 = TT \end{cases}$$

This is the default model implemented in the PheWAS package in R (R. J. Carroll, Bastarache, and Denny 2014).

Genotypic Model

In the genotypic model the alleles of the SNP are treated as categorical variables with the SNP with two non-risk allele coded as the reference category. For example, for gene SLC26A9 and SNP rs4077469, AA would be the reference category.

Assumptions of Logistic Regression Model

The assumption of the model that the observations must be independent, was ensured by removing people that were related up to siblings. And, although not required for logistic regression, the sample sizes are large enough, (the smallest data set has over 170,000 individuals), to ensure normality of the covariates, as can be seen from Tables 3, 5, and 7. Also logistic regression should have little or no multicollinearity, this assumption may be an issue for us since we have the age and age-squared term in the model. Logistic regression also assumes the independent variables are linearly related to the log odds, which we found no way to verify using the PheWAS package and did not find time to verify independently, as of the time this report.

Lastly, logistic regression requires a large sample size. This is because maximum likelihood (ML) estimates are less powerful than ordinary least squares (OLS). For OLS it is recommended to have 5 cases per independent variable, where ML is recommended to have at least 10 to 30. This requirement is also taken care of in our study because of the large sample sizes.

3.2.2 Primary Analysis

Our primary analysis is to perform a PheWAS for the SNPs rs4077469 (SCL26A9), rs5905176 (SLC6A14) and rs17497684 (SCL9A3), and find which phenotypes are associated with them, using the additive model. We use the additive model in our primary analysis due to the model's smaller degrees of freedom (fewer parameters to estimate), hence higher power, compared to the genotypic model which has more parameters to estimate.

3.2.3 Secondary Analysis

The secondary analysis for this study is to perform a similar PheWAS as above, except using the genotypic model instead of the additive model. And also to include the interaction terms between sex and three allele categories.

At the time of this report, the secondary analysis was not performed due time constraints.

3.2.4 Tertiary Analysis

So far we have only discussed phenotypes that are binary variables obtained from the ICD-10 codes. However, the UKBB has collected extensive data on myriad other phenotypes. The tertiary analysis is to use curated phenotypes from the UKBB, mainly those of lung functionality such as the FEV_1/FVC ratio, and find an association between these and our SNPs of interest. These curated phenotypes are mostly continuous variables and we would be performing multiple linear regression adjusted for covariates using the genotypic model. The PheWAS package in R can handle contunious outcome variables.

As of the day this report was written, the tertiary analysis has not been performed due to time constraints.

3.2.5 Hardy-Weinburg Equilibrium

Hardy-Weinburg Equilibrium (HWE) assumes that if there is equal mixing of individuals in this gene pool, then we should have a distribution that follows HWE: $p^2 + 2pq + q^2 = 1$. Where, for example, p is the allele frequence of A and q is the allele frequency of T for the SNP rs4077469 (SLC26A9).

We look at the HWE to make sure that the individuals selected in case and control groups provide unbiased allele frequency estimates of the true underlying distribution in affected and unaffected members of the population of interest. If not, association findings will merely reflect biases resulting from the study design (Clarke et al. 2011).

We also look at HWE for the SNP for quality control. If the SNP is faulty, then within the controls, we would have deviance from the HWE. Therefore, a statistically significant p-value for the HWE is evidence that the SNP may be faulty or the individuals selected in case and control groups do not provide ubniased allele frequencies.

3.2.6 Data and Methodology Validation

We would want to confirm that our data, data cleaning process, converting ICD-10 to phecodes process and PheWAS analysis are all correctly done. To do this we would replicate results from other PheWAS studies using our data and methodology. For example, multiple sclerosis and the SNP rs3135388 were found to be statistically significant associated with OR of 2.56 and P-value of 1.4×10^{-7} (R. J. Carroll, Bastarache, and Denny 2014). We will call such a confirmatory result a positive control study and would want to perform several such positive control studies.

However, due to time constraints, at the time of this report, we were not been able to perform any positive control study.

3.3 Software

All statistical analysis was performed using the “R” software (R Core Team 2017). The phenome-wide association study was performed using the “PheWAS” package in R (R. J. Carroll, Bastarache, and Denny 2014). The PheWAS package is able to handle phenotypes that are either binary or continuous (R. J. Carroll, Bastarache, and Denny 2014). Descriptive statistics and tables were created using the “dplyr” package (Wickham et al. 2018) and the “tableone” package (Yoshida and Bohn. 2018). Sample R code for all the analysis is provided in the Appendix.

High performance computing was used to run the R code using the Hospital for Sick Children’s High Performance Facility (HPF). Bash scripts were written in Linux to run the R code. A sample Bash script is provided in the Appendix.

All genotypic data cleaning and preparation was done using PLINK prior to loading the genotypic data into R.

4 Results

4.1 Modifier Gene SLC9A3

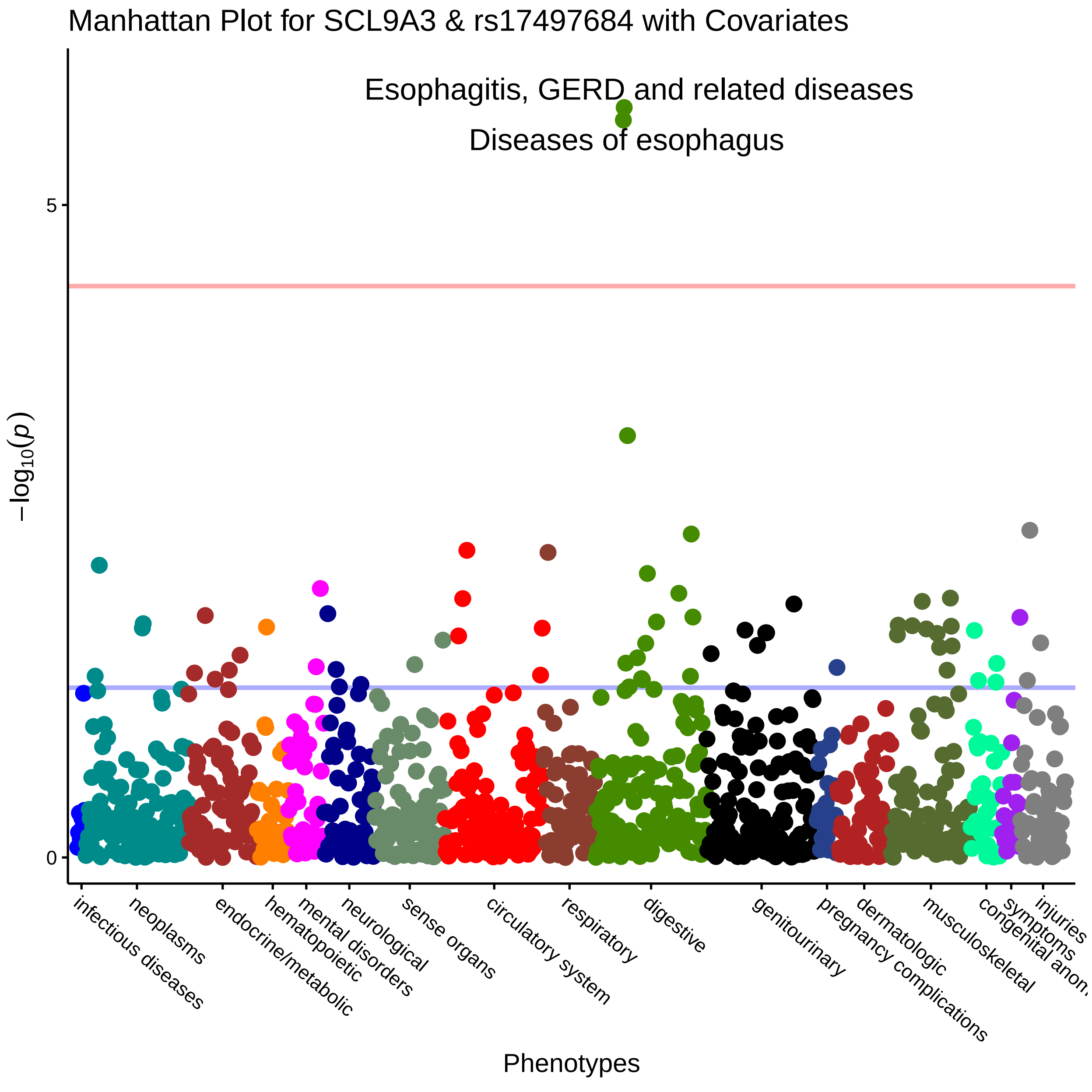

Figure 3: PheWAS Manhatten Plot with covariates. The line close to 5 is the Bonferroni corrected significance level (red line). The other line is the 5% significance level (blue line).

Table 2: Top 10 phenotypes associated with gene SCL9A3 and SNP rs17497684 by p-value. The risk allele is C for the first phenotype.

phecode	description	group	OR	p	n_cases	n_controls	HWE_p	bonferroni
530.10	Esophagitis, GERD and related diseases	digestive	1.06	0.0000018	19687	243236	0.088	TRUE
530.00	Diseases of esophagus	digestive	1.06	0.0000022	21330	241593	0.088	TRUE
530.14	Reflux esophagitis	digestive	1.07	0.0005850	7187	255736	0.088	FALSE
803.30	Fracture of clavicle or scapula	injuries & poisonings	1.23	0.0031095	576	262347	0.088	FALSE
575.00	Other biliary tract disease	digestive	0.87	0.0033213	1459	261464	0.088	FALSE
418.10	Precordial pain	circulatory system	0.91	0.0044285	3024	259899	0.088	FALSE
471.00	Nasal polyps	respiratory	1.11	0.0045946	2380	260543	0.088	FALSE
153.30	Malignant neoplasm of rectum, rectosigmoid junction, and anus	neoplasms	1.12	0.0057644	1840	261083	0.088	FALSE
550.10	Inguinal hernia	digestive	1.84	0.0066634	47	262876	0.088	FALSE
315.00	Develomental delays and disorders	mental disorders	2.05	0.0086885	30	262893	0.088	FALSE

HWEp: Hardy–Weinberg equilibrium p-value. Testing deviation from HWE.

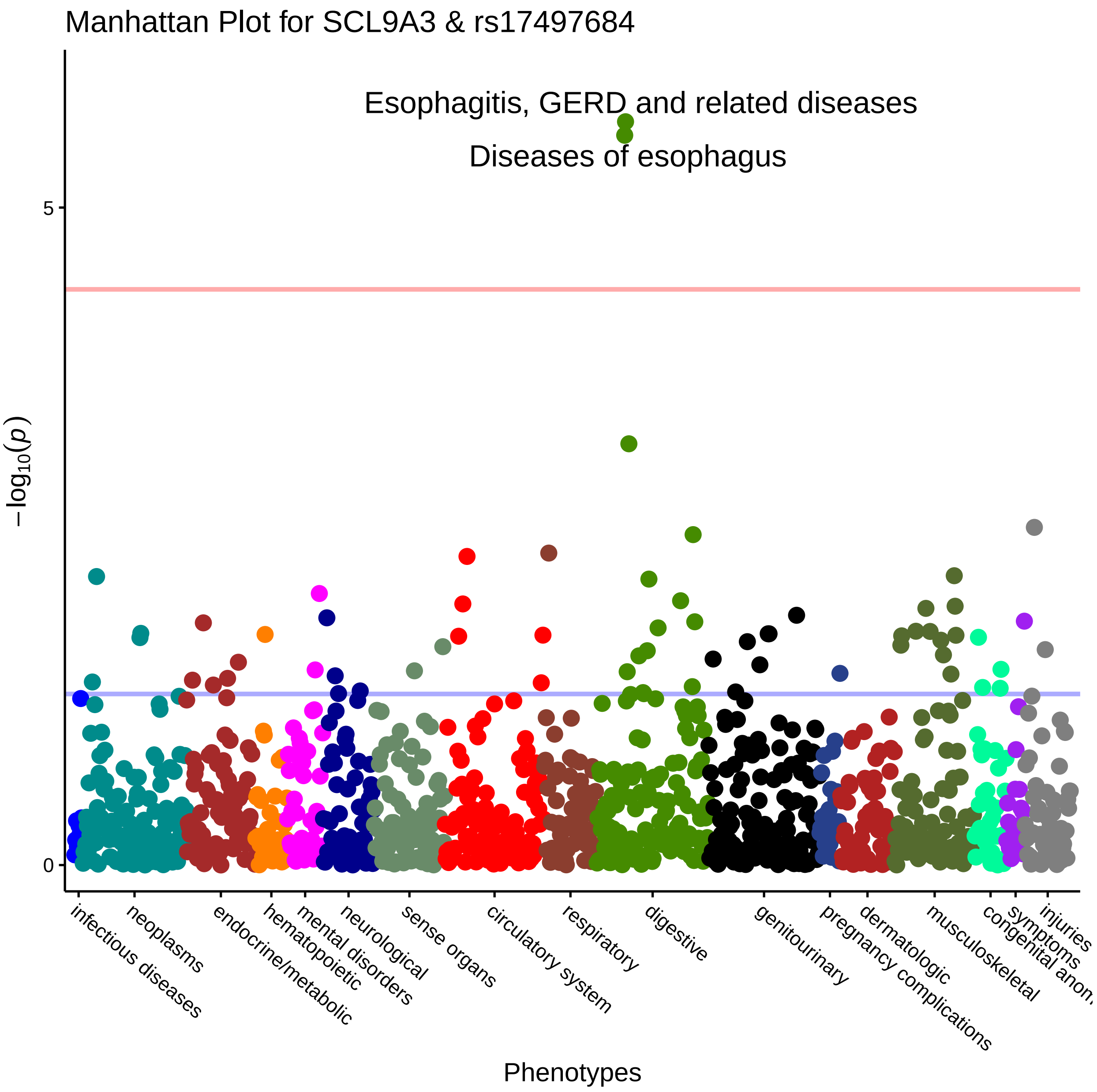

Figure 4: PheWAS without covariates. The line close to 5 is the Bonferroni corrected significance level (red line). The other line is the 5% significance level (blue line).

Table 3: Distribution of variables by allele for SCL9A3.

Variables	Allele:	AA	AC	CC
.	C Allele Count:	0	1	2
n	.	170050	82784	10089
Esophagitis, GERD and related diseases	Controlls	157647 (92.7)	76284 (92.1)	9305 (92.2)
phecode: 530.1 (%)	Case	12403 (7.3)	6500 (7.9)	784 (7.8)
Diseases of esophagus	Controlls	156593 (92.1)	75759 (91.5)	9241 (91.6)
phecode: 530 (%)	Cases	13457 (7.9)	7025 (8.5)	848 (8.4)
age (mean (SD))	.	57.85 (7.78)	57.84 (7.79)	57.70 (7.77)
age-sqrd (mean (SD))	.	3407.06 (871.43)	3406.22 (872.61)	3389.88 (869.40)
SEX (%)	male	94320 (55.5)	45716 (55.2)	5598 (55.5)
.	female	75730 (44.5)	37068 (44.8)	4491 (44.5)

We found Esophagitis, GERD and related diseases to be significantly associated with gene SCL9A3 and SNP rs17497684 with OR = 1.064, p-value = 1.79x10^{-6} for the adjusted analysis. Very similar result for the unadjusted analysis was found as can be seen by comparing Figure 3 and Figure 4. We do not present the results table for the unadjusetd analysis for the interest of brevity.

The disease of the esophagus is simply a roll-up phenotype and includes the phenotype Esophagitis, GERD and related diseases as can be seen from the phecodes in Table 2.

4.2 Modifier Gene SLC6A14 for Males

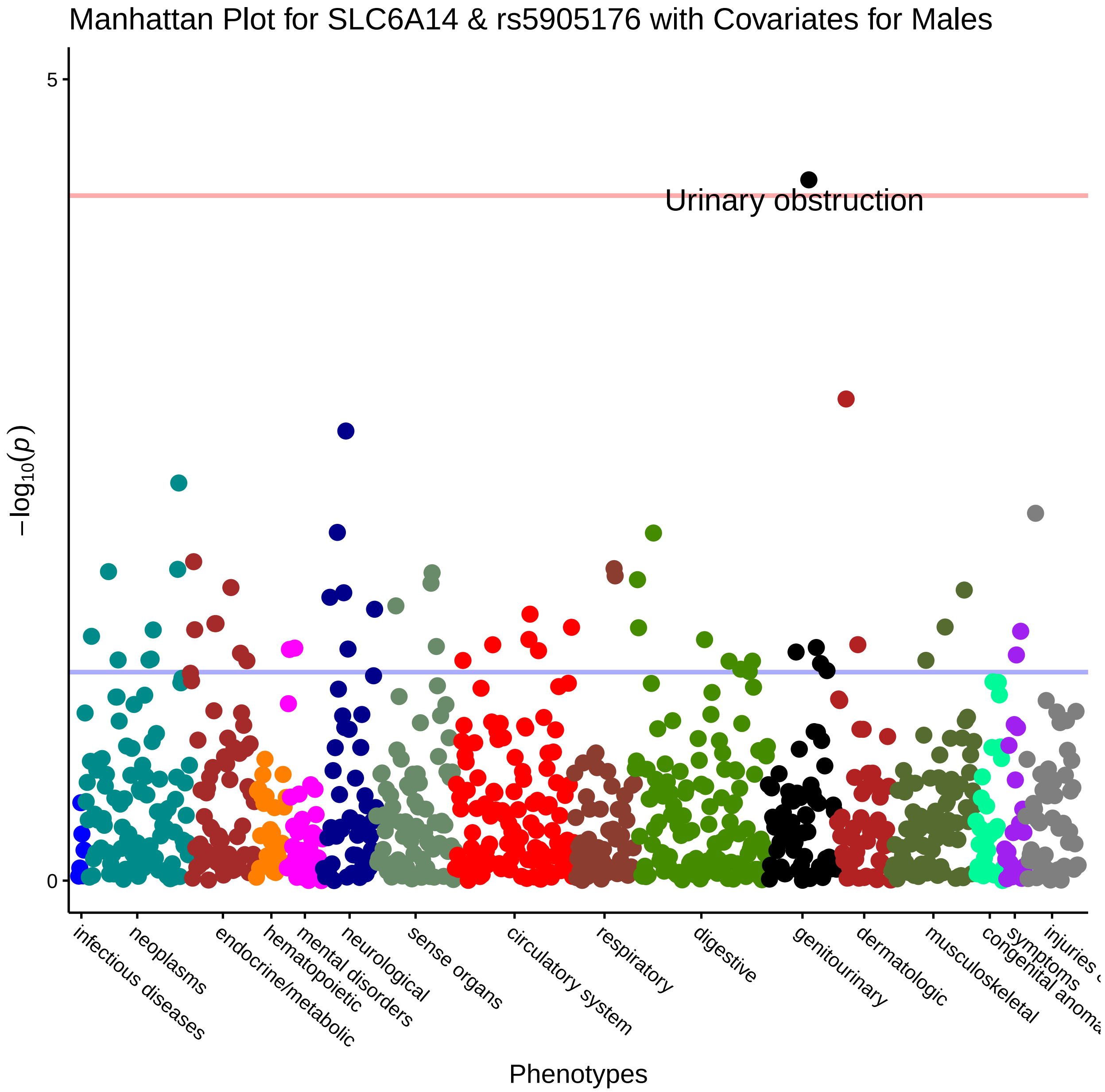

Figure 5: PheWAS with covariates. The line close to 5 is the Bonferroni corrected significance level (red line). The other line is the 5% significance level (blue line).

Table 4: Top 10 phenotypes associated with gene SLC6A14 and SNP rs5905176 by p-value. The risk allele is G for the first phenotype.

phecode	description	group	OR	p	n_cases	n_controls	HWE_p	bonferroni
599.1	Urinary obstruction	genitourinary	1.68	0.0000424	64	117334	1	TRUE
689.0	Disorder of skin and subcutaneous tissue NOS	dermatologic	1.09	0.0009880	1751	115647	1	FALSE
341.0	Other demyelinating diseases of central nervous system	neurological	0.57	0.0015652	67	117331	1	FALSE
223.0	Benign neoplasm of kidney and other urinary organs	neoplasms	0.67	0.0033017	90	117308	1	FALSE
803.1	Fracture of humerus	injuries & poisonings	0.87	0.0051026	491	116907	1	FALSE
334.0	Degenerative disease of the spinal cord	neurological	0.80	0.0067156	210	117188	1	FALSE
527.2	Sialoadenitis	digestive	1.29	0.0067768	117	117281	1	FALSE
242.0	Thyrotoxicosis with or without goiter	endocrine/metabolic	0.72	0.0102311	95	117303	1	FALSE
507.0	Pleurisy; pleural effusion	respiratory	0.91	0.0112994	888	116510	1	FALSE
222.0	Benign neoplasm of male genital organs	neoplasms	0.74	0.0114173	109	117289	1	FALSE

HWEp: Hardy–Weinberg equilibrium p-value. Testing deviation from HWE.

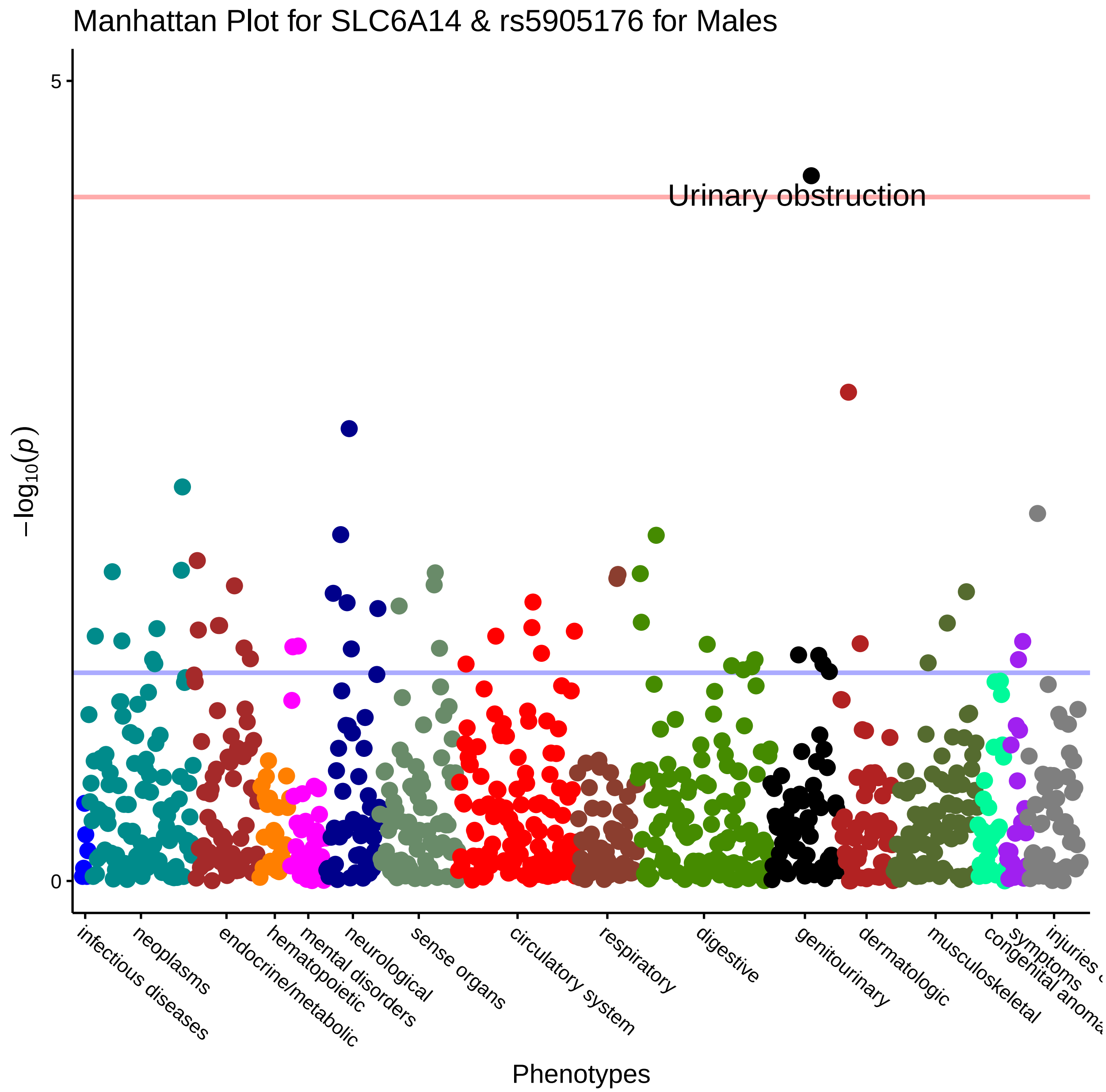

Figure 6: PheWAS without covariates. The line close to 5 is the Bonferroni corrected significance level (red line). The other line is the 5% significance level (blue line).

Table 5: Distribution of variables by phenotype Urinary obstruction for SLC6A14 for males.

.	..	Urinary obstruction	...
.	level	Controls	Cases
n	.	117334	64
rs5905176_G (%)	0	79076 (67.4)	27 (42.2)
.	2	38258 (32.6)	37 (57.8)
age (mean (SD))	.	58.44 (7.73)	62.19 (5.96)
age2 (mean (SD))	.	3475.18 (868.88)	3902.25 (706.87)

For males, we found urinary obstruction to be significantly associated with gene SLC6A14 and SNP rs5905176 with OR = 1.68, p-value = $4.24\text{x}10^{-5}$ for the adjusted analysis. Very similar result for the unadjusted analysis was found as can be seen by comparing Figure 5 and Figure 6.

4.3 Modifier Gene SLC6A14 for Females.

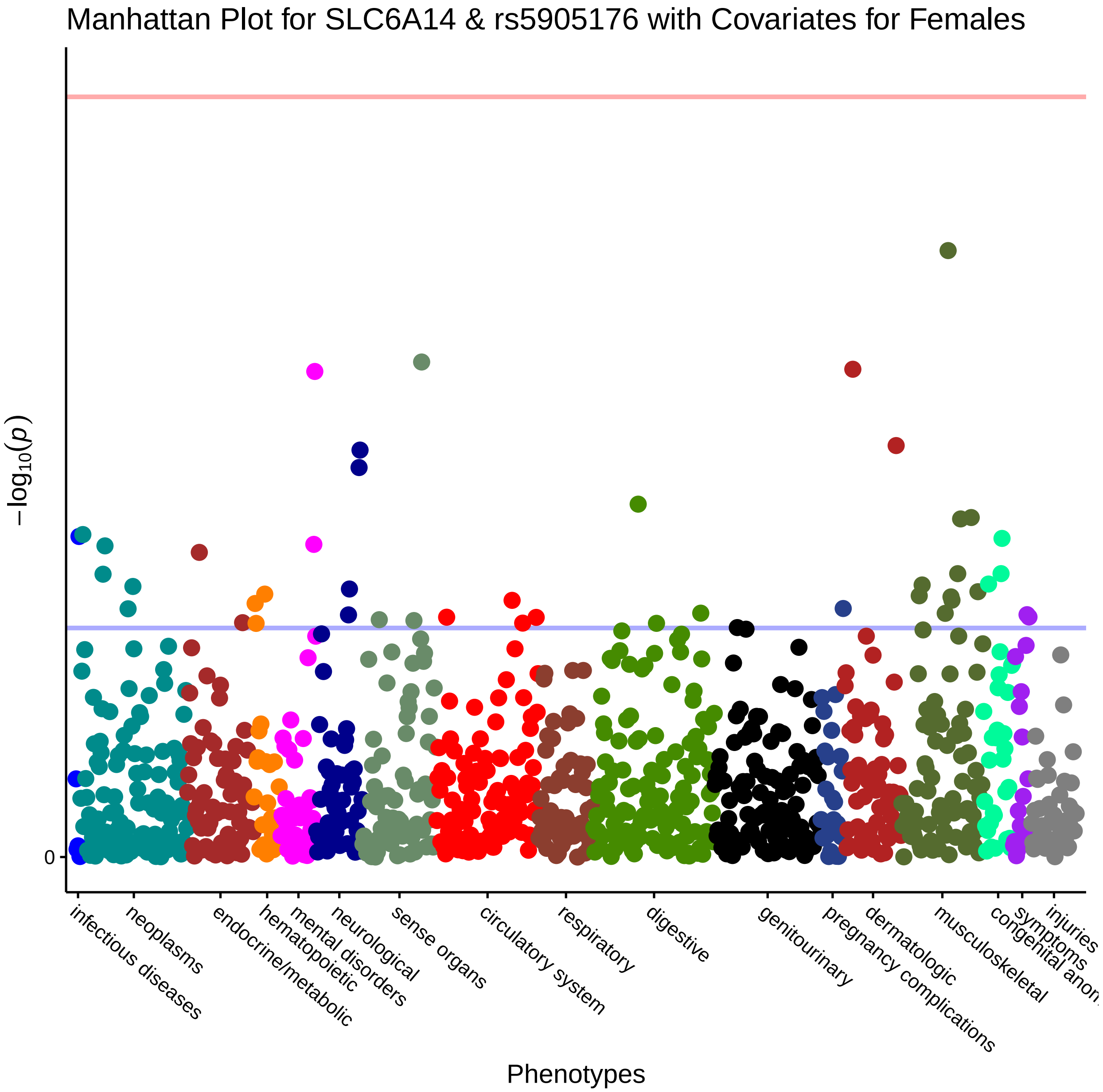

Figure 7: PheWAS with covariates. The line close to 5 is the Bonferroni corrected significance level (red line). The other line is the 5% significance level (blue line).

Table 6: Top 10 phenotypes associated with gene SLC6A14 and SNP rs5905176 by p-value. The risk allele is G for the first phenotype.

phecode	description	group	OR	p	n_cases	n_controls	HWE_p	bonferroni
733.6	Costochondritis	musculoskeletal	2.05	0.0003589	49	145456	0.924	FALSE
382.0	Otalgia	sense organs	0.56	0.0015416	89	145416	0.924	FALSE
689.0	Disorder of skin and subcutaneous tissue NOS	dermatologic	0.90	0.0016945	2243	143262	0.924	FALSE
317.1	Alcoholism	mental disorders	1.54	0.0017445	106	145399	0.924	FALSE
707.2	Chronic ulcer of leg or foot	dermatologic	1.48	0.0045980	108	145397	0.924	FALSE
358.1	Myasthenia gravis	neurological	1.81	0.0048776	45	145460	0.924	FALSE
358.0	Myoneural disorders	neurological	1.75	0.0061299	48	145457	0.924	FALSE
531.2	Gastric ulcer	digestive	1.13	0.0098949	1052	144453	0.924	FALSE
741.4	Joint effusions	musculoskeletal	0.78	0.0117856	261	145244	0.924	FALSE
738.4	Acquired spondylolisthesis	musculoskeletal	1.20	0.0120033	399	145106	0.924	FALSE

HWEp: Hardy–Weinberg equilibrium p-value. Testing deviation from HWE.

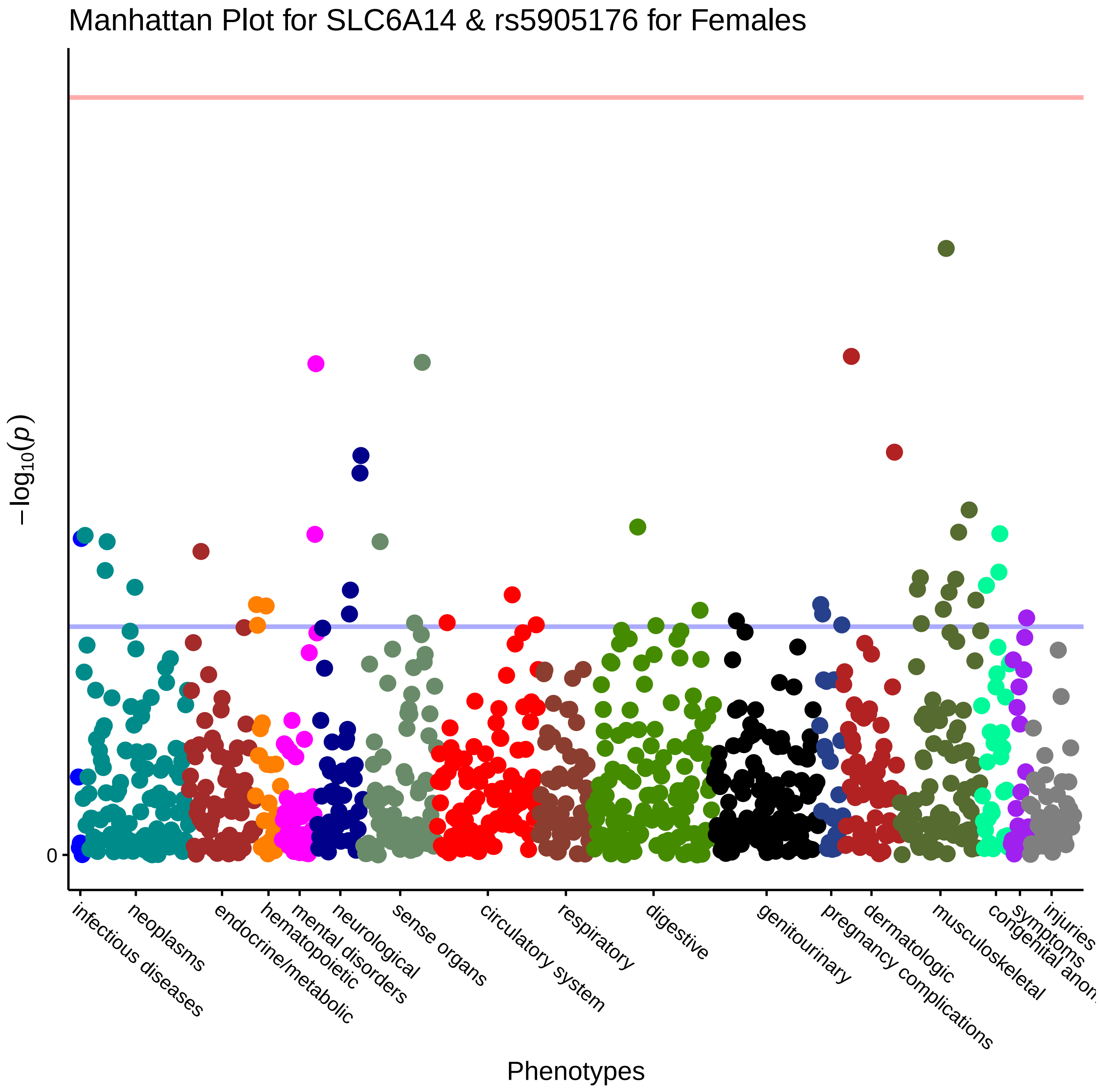

Figure 8: PheWAS without covariates. The line close to 5 is the Bonferroni corrected significance level (red line). The other line is the 5% significance level (blue line).

Table 7: Distribution of variables by phenotype Costochondritis for SLC6A14 for females.

.	..	Costochondritis	. . .
.	level	Cases	Controls
n	.	145456	49
rs5905176_G (%)	0	66142 (45.5)	12 (24.5)
.	1	63668 (43.8)	25 (51.0)
.	2	15646 (10.8)	12 (24.5)
age (mean (SD))	.	57.35 (7.80)	56.49 (8.25)
age-sqrd (mean (SD))	.	3350.21 (870.04)	3257.80 (908.41)

No statistically significant association was found for females.

4.4 Modifier Gene SLC26A9

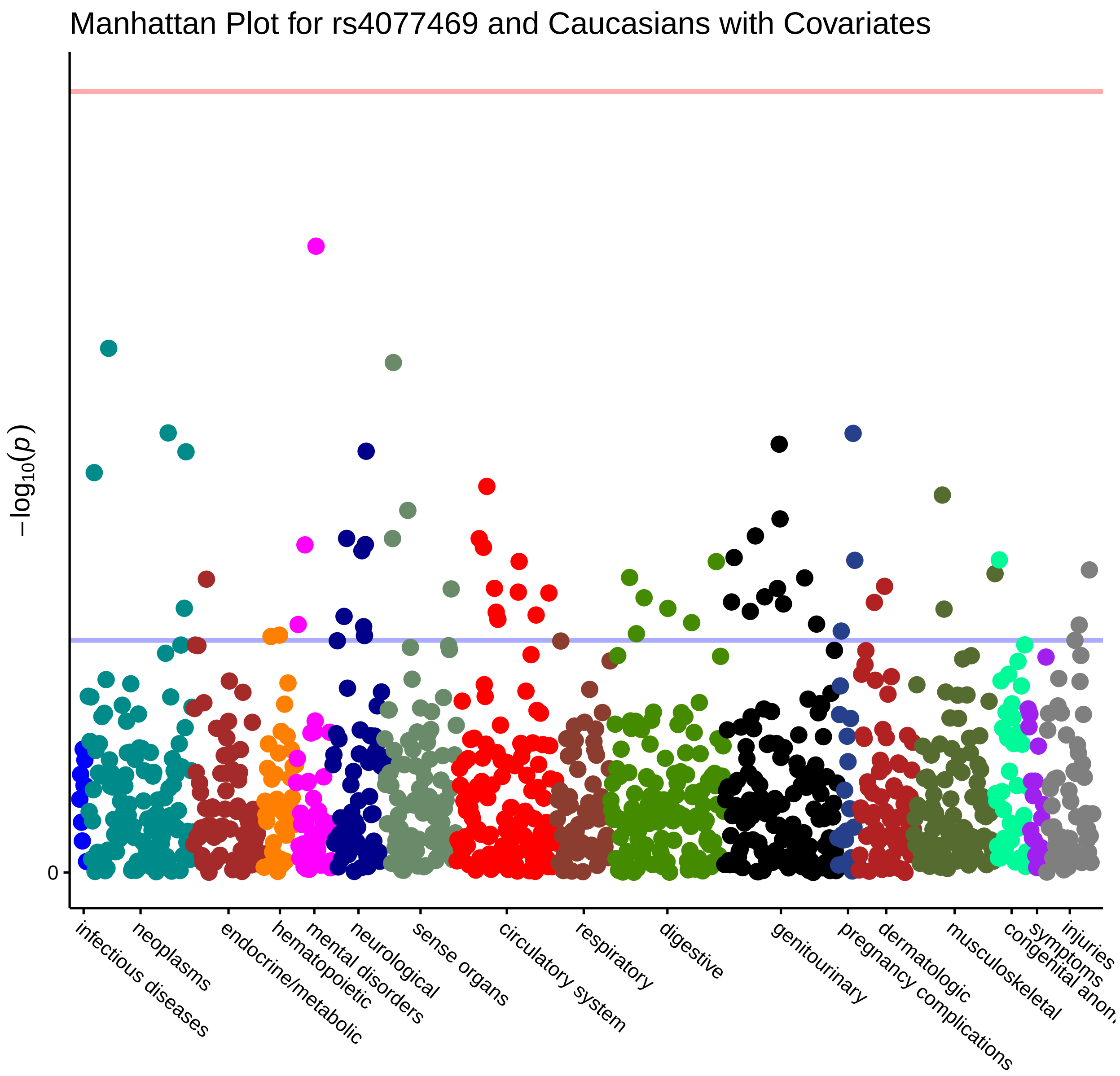

Figure 9: PheWAS with covariates after removing non-Caucasians and related people. The line close to 5 is the Bonferroni corrected significance level (red line). The other line is the 5% significance level (blue line).

Table 8: Top 10 phenotypes associated with gene SLC26A9 and SNP rs4077469 by p-value. The risk allele is C for the first phenotype.

phecode	description	group	OR	p	n_cases	n_controls	HWE_p	bonferroni
300.12	Agorophobia, social phobia, and panic disorder	mental disorders	0.603	0.0003084	126	261529	0.989	FALSE
159.40	Malignant neoplasm of retroperitoneum and peritoneum	neoplasms	0.668	0.0011522	155	261500	0.989	FALSE
362.80	Retinal hemorrhage/ischemia	sense organs	0.343	0.0013846	30	261625	0.989	FALSE
208.00	Benign neoplasm of colon	neoplasms	1.036	0.0034356	15656	245999	0.989	FALSE
653.00	Problems associated with amniotic cavity and membranes	pregnancy complications	1.497	0.0034549	105	261550	0.989	FALSE
609.20	Abnormal spermatozoa	genitourinary	0.661	0.0039709	116	261539	0.989	FALSE
348.90	Other conditions of brain, NOS	neurological	1.543	0.0043512	87	261568	0.989	FALSE
225.20	Benign neoplasm of spinal cord, meninges	neoplasms	0.496	0.0043830	45	261610	0.989	FALSE
149.10	Cancer of oropharynx	neoplasms	0.755	0.0057335	220	261435	0.989	FALSE
425.12	Other hypertrophic cardiomyopathy	circulatory system	0.579	0.0068512	62	261593	0.989	FALSE

HWEp: Hardy–Weinberg equilibrium p-value. Testing deviation from HWE.

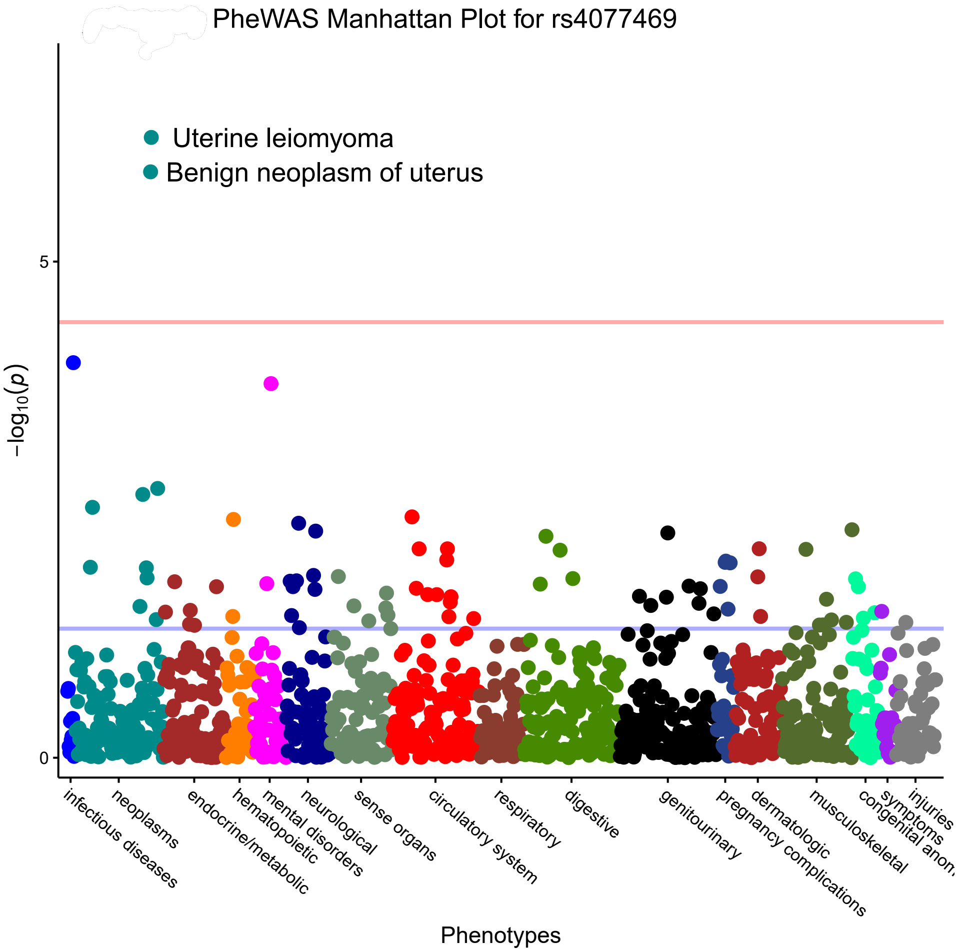

Figure 10: PheWAS without covariates, and without removing non-Caucasians and related people. The line close to 5 is the Bonferroni corrected significance (red line). The other line is the 5% significance level (blue line).

Table 9: Distribution of variables by allele for SLC26A9.

Variables	Allele:	CC	CT	TT
.	T Allele Count:	0	1	2
n	.	136712	180337	58080
age (mean (SD))	.	56.99 (7.93)	57.01 (7.92)	56.96 (7.95)
age-sqrd (mean (SD))	.	3311.05 (882.28)	3312.80 (880.44)	3307.63 (884.15)
SEX (%)	male	73585 (53.8)	96722 (53.6)	31111 (53.6)
.	female	63127 (46.2)	83615 (46.4)	26969 (46.4)

For the unadjusted analysis Uterine leiomyoma and Benign neoplasm of uterus, were found to be statistically significant (OR = 0.92, p-value = 5.61x10^{-7} and OR = 0.93, p-value = 1.25x10^{-6}, respectively). However, for the adjusted analysis no statistically significant association was found.

5 Discussion

Modifier Gene SCL9A3

For our primary analysis we found that the SNP rs57221529 for gene SCL9A3 (original SNP rs17497684, r = 0.821) was statistically significantly associated with Esophagitis, GERD and related diseases (OR = 1.064, S.E. = 0.013, p-value = 1.79x10^{-6}). Hence, since our model is an additive model each additional C allele increases the odds of having Esophagitis, GERD and related diseases by 6.4% (Table 2). Not only is this is statistcally significant but also clinically meaningful increase in odds. Further, this is not a sporadic relationship because for each allele count (0, 1, 2) the number of cases were 12403, 6500 and 784, and the controls were 157647, 76284 and 9305, respectively (Table 3). To our knowledge SCL9A3 has not been found to be associated with Esophagitis, GERD and related diseases in any previous study. The gene SCL9A3 with SNP rs57221529 has, however, been found to be significantly associated with reduced lung function (Corvol et al. 2015).

Modifier Gene SLC6A14

For males, the SNP rs3788766 for gene SLC6A14 (original SNP rs5905176) was statistically associated with having urinary obstruction (OR = 1.68, S.E. = 0.127, p-value = 4.24x10^{-5}). However, there were a total of only 64 cases of urinary obstruction, with 27 cases for allele count 0 and 37 cases for allele count 2 (note that males cannot have 1 allele count as this gene is on the X chromosome) with 79076 and 38258 controls, respectively. Since, there are so few cases of urinary obstruction it is indeed possible that this association is sporadic and purely due to chance. For logistic regression it is recommended to have 10 to 30 cases per independent variable.

Modifier Gene SLC26A9

For the unadjusted analysis we found an association with the phenotypes uterine leiomyoma and benign neoplasm of uterus (OR = 0.92, p-value = 5.61x10^{-7} and OR = 0.93, p-value = 1.25x10^{-6}, respectively) with risk allele C. These diseases are gender specific and only occur in females, and after adjusting for covariates were no longer significant. However, in the study by S. Blackman et al. (2013) the C allele was also found to be the risk allele for CF-related diabetes with hazard ratio [HR] 1.38, p-value = 3.6x10^{-8} in their adjusted and unadjusted analysis, see their Table 2 (S. Blackman et al. 2013). Hence, one further analysis to do would be to perform the same analysis adjusted to covariates but for each sex separately. This will increase the power and allow us to see if uterine leiomyoma and benign neoplasm of uterus are associated with SLC26A9 and SNP rs4077469. However, the CF modifier gene SLC26A9 and SNP rs4077469 have never been found to be associated with these kinds of gender specific diseases and we would not expect to find a statistically significant association for the separate analyses (At the time of this report, due to time constraints, we have not been able to perform this subanalysis yet). For the adjusted analysis no statistically significant association with any of the phenotypes was found.

Bonferroni Correction

As three different SNPs were tested, the proper Bonferroni corrected p-value threshold for significance should not be 5%/1511, but in fact be (5%/1511)/3 = 1.1x10^{-5}. With this threshold only the association between gene SCL9A3/SNP rs57221529 (original SNP rs17497684, r = 0.821) and Esophagitis, GERD and related diseases, is statistically significantly (p-value = 1.79x10^{-6}).

Hardy–Weinberg Equilibrium

A fundamental assumption of the case-control study is that the individuals selected in case and control groups provide unbiased allele frequencies of the true underlying distribution in affected and unaffected members of the population. If not, association findings will merely reflect biases resulting from the study design (Clarke et al. 2011). Hence, since all the SNPs had HWE p-values that were not statistically significant, association findings reflect true finding and not biases from the study design.

Limitations

There were several limitations in our study. Firstly, we did not use the ICD-9 codes in the UKBB registry. Therefore, could have introduced bias by not including these people as cases or controls in the analysis. These people were primarily in the beginning of the recruitment period of the study.

Further, as can be seen, all and any missing data was deleted. Thus, for the gene SLC26A9, we went from 377,961 to 263,607 individuals when we merged the final genotypic data with the phenotypic data. Hence, this may be introducing bias in our analysis if the missing data is not missing completely at random. And it is quite possible that a substantial number of people with missing data were people who only had ICD-9 codes in the UKBB data, therefore not missing at random.

Moreover, no exclusion scheme was implementing when we defined our controls. This will cause our analysis to have more controls, increasing the chance of committing a Type I error, because people who could not have a particular phenotype due to the exclusion are included as controls when they should not. For instance, if a person's phenotypic sex is female, then when analyzing if the person has disease of the uterine, all males should be excluded as they will artificially inflate the number of controls. This is indeed what could have transpired in the Results Section 4.4 when we performed the unadjusted analysis for the gene SLC26A9 and SNP rs4077469 (see Figure 10) and found uterine leiomyoma and benign neoplasm of uterus to be statistically significant. Hence, it is recommended to implement the exclusion scheme for controls when performing PheWAS. This is an important limitation of our study.

Also, the phecode system was built upon the ICD-9 codes. However, we used ICD-10 codes to obtain the phecodes. In the study by Wu et al. (2018), they assessed the development and validation of mappings for ICD-10/ICD-10-CM to phecode for the UKBB. Of the unique codes observed in the UKBB (ICD-10) cohort, >90% were mapped to phecodes. They observed 70-75% reproducibility for chronic diseases and <10% for an acute disease (Wu et al. 2018). To further assess the validity of the ICD-10 mapping to phecodes, they performed a PheWAS with lipoprotein(a), rs10455872, using the ICD-9-CM and ICD-10-CM maps, replicating two genotype-phenotype associations with very similar effect sizes: coronary atherosclerosis (ICD-9-CM: $P < .001$, OR = 1.60 vs. ICD-10-CM: $P < .001$, OR = 1.60) and with chronic ischemic heart disease (ICD-9-CM: $P < .001$, OR = 1.5 vs. ICD-10-CM: $P < .001$, OR = 1.47) (Wu et al. 2018). Therefore, if ICD-10 codes are properly mapped to phecodes then the results of the PheWAS should be reliable.

Lastly, several phenotypes and phecodes are related to one another, for example, the phenotype of diseases of esophagus (phecode 530.00) is a roll-up of Esophagitis, GERD and related diseases (phecode 530.10). Hence, not all the phenotypes are independent and our Bonferroni correction is overly conservative.

Conclusion

In summary, the results of this study show that among Caucasians, the gene SCL9A3 and the region near the SNP rs17497684 is associated with the disease Esophagitis, GERD and related diseases, and each additional C allele increases the odds of having the diseases by 6.4%. Since CF is a disease mainly affecting Caucasians, and our analysis was restricted to Caucasians, our results are generalizable to people with Caucasian ancestry. With this PheWAS we may have found a phenotype associated with gene SCL9A3 in the non-CF population.

6 References

Blackman, Scott, Clayton Commander, Christopher Watson, Kristin M Arcara, Lisa Strug, Jaclyn R Stonebraker, Fred A Wright, et al. 2013. “Genetic Modifiers of Cystic Fibrosis–Related Diabetes.” *Diabetes* 62 (May). doi:10.2337/db13-0510.

Carroll, Robert J, Lisa Bastarache, and Joshua C Denny. 2014. “R PheWAS: Data Analysis and Plotting Tools for Phenome-Wide Association Studies in the R Environment.” *Bioinformatics* 30 (16). Oxford University Press: 2375–6.

Clarke, Geraldine M, Carl A Anderson, Fredrik H Pettersson, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. 2011. “Basic Statistical Analysis in Genetic Case-Control Studies.” *Nature Protocols* 6 (2). Nature Publishing Group: 121.

Cohen-Cymberknoh, Malena, David Shoseyov, and Eitan Kerem. 2011. “Managing Cystic Fibrosis: Strategies That Increase Life Expectancy and Improve Quality of Life.” *American Journal of Respiratory and Critical Care Medicine* 183 (11). American Thoracic Society: 1463–71.

Cohn, Jonathan A, Kenneth J Friedman, Peadar G Noone, Michael R Knowles, Lawrence M Silverman, and Paul S Jowell. 1998. “Relation Between Mutations of the Cystic Fibrosis Gene and Idiopathic Pancreatitis.” *New England Journal of Medicine* 339 (10). Mass Medical Soc: 653–58.

Corvol, Harriet, Scott M Blackman, Pierre-Yves Boëlle, Paul J Gallins, Rhonda G Pace, Jaclyn R Stonebraker, Frank J Accurso, et al. 2015. “Genome-Wide Association Meta-Analysis Identifies Five Modifier Loci of Lung Disease Severity in Cystic Fibrosis.” *Nature Communications* 6. Nature Publishing Group: 8382.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.” *PLoS Medicine* 12 (3). Public Library of Science: e1001779.

Uppu, Suneetha, Aneesh Krishna, and Raj P Gopalan. 2018. “A Review on Methods for Detecting Snp Interactions in High-Dimensional Genomic Data.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15 (2). IEEE: 599–612.

Wei, Wei-Qi, Lisa A Bastarache, Robert J Carroll, Joy E Marlo, Travis J Osterman, Eric R Gamazon, Nancy J Cox, Dan M Roden, and Joshua C Denny. 2017. “Evaluating Phecodes, Clinical Classification Software, and Icd-9-Cm Codes for Phenome-Wide Association Studies in the Electronic Health Record.” *PloS One* 12 (7). Public Library of Science: e0175508.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2018. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wu, Patrick, Aliya Gifford, Xiangrui Meng, Xue Li, Harry Campbell, Tim Varley, Juan Zhao, et al. 2018. “Developing and Evaluating Mappings of Icd-10 and Icd-10-Cm Codes to Phecodes.” *BioRxiv.* Cold Spring Harbor Laboratory, 462077.

Yoshida, Kazuki, and Justin Bohn. 2018. *Tableone: Create ’Table 1’ to Describe Baseline Characteristics.* https://CRAN.R-project.org/package=tableone.

7 Additional information

7.1 Competing financial interests

The authors declare no competing interests.

8 Appendix

Below is the Bash file to run the R script in HPF. Note that the below code is commented out.

```
# #!/bin/bash -x
#
# #PBS -l mem=65g
# #PBS -l vmem=65g
# #PBS -l walltime=3:59:00
# #PBS -l nodes=1:ppn=1
# #PBS -o /hpf/largeprojects/cfcentre/strug/faizan/job_output/
# #PBS -e /hpf/largeprojects/cfcentre/strug/faizan/job_output/
# #PBS -d /hpf/largeprojects/cfcentre/strug/faizan/
# #PBS -N phewasSCL9A3_Ch5_March08
#
# hostname
# date
#
# echo "Working dir is ${PBS_O_WORKDIR}"
# cd $PBS_O_WORKDIR
#
#
# module load R/3.4.0
# Rscript phewas1script2v7_SCL9A3_Ch5.R > phewas1script2v7_SCL9A3_Ch5.Rout
#
#
#
# echo "Done"
#
# date
```

Below is the R code for running the entire PheWAS for the modifier gene SCL9A3.

The R code for the modifier gene SCL26A9 is almost identical to the below code.

```
####
#  R script for running phewas.
#
# Can simply change the name of the gene and the SNP in the script below
# and the adjusted and unadjusted analysis with the Table 1 of the covariates and
# the distribution of the alleles against the top two phenotypes will be produced.
#
# Things to change for new Gene and SNP (not an exhaustive list)
# ukb_chr5_rs17497684.raw
# SCL9A3
# rs17497684
# rs17497684_C
# Ch5
```

```
#####

require(data.table)
require(PheWAS)
require(dplyr)
require(tableone)

# Loading the dataset of unrealted people
unrelated = fread("09-kinship-degree2-unrelatedunrelated.txt", col.names = c("fid", "id"),
                  stringsAsFactors=F, header=F, na.strings=c(""," ","NA"))

# Cleaning the data.
print(str(unrelated))
unrelated$id = as.numeric(unrelated$id)
unrelated1 = unrelated  %>% filter(!is.na(id)) %>% filter(id>0) #Should not have neg or NA's as id's
print(dim(unrelated1))
print(str(unrelated1))

gdata0 = fread("ukb_chr5_rs17497684.raw", stringsAsFactors=F, header=T, na.strings=c(""," ","NA"))
gdata0[1:10,]
print(dim(gdata0))
gdata1 = rename(gdata0, gender = SEX, id = IID)
gdata1[1:10,]
print(str(gdata1))

# Creating the independent data set for the genotypic data.
gdata2 = inner_join(unrelated1, gdata1, by ="id")
dim(gdata2)
gdata2[1:10,]
print(str(gdata2))
summary(gdata2)

# Reading in genetic ethnic group data ( caucasian = T/F)
ethnic_data0 = fread("ukb24727_22006_genetic_ethnic_groups.tab",
                    col.names = c("id", "caucasian"), stringsAsFactors=F,
                    header=T, na.strings=c(""," ","NA"))
dim(ethnic_data0)
ethnic_data = filter(ethnic_data0, !is.na(caucasian))
dim(ethnic_data)
print(str(ethnic_data))
summary(ethnic_data)

gdata3 = inner_join(gdata2, ethnic_data, by = "id")
print(dim(gdata3))
summary(gdata3)
str(gdata3)
genotypes = gdata3 %>% select( id, rs17497684_C)
genotypes[1:10,]

covariates_age_gender = fread("pheno_age_sex.tab",  stringsAsFactors=F,
                              col.names = c("id", "age", "SEX"),
                              header=T, na.strings=c(""," ","NA"))
print(dim(covariates_age_gender))
```

```
covariates1 = mutate(covariates_age_gender, age2 = age**2)
covariates = semi_join(covariates1, genotypes, by = "id")
print(dim(covariates))
summary(covariates)
print(str(covariates))
print(covariates[1:10,])

phedata = fread("icd10_data_with_phecodes2.txt", stringsAsFactors=F, header=T,
                na.strings=c(""," ","NA"))
head(phedata)
print(str(phedata))
summary(phedata)
dim(phedata)

phedata0 = phedata %>% rename(phecode = code) %>% select(id, phecode, count)
head(phedata0)
str(phedata0)
print(dim(phedata0))
summary(phedata0)

phenotypes=createPhewasTable(phedata0, min.code.count = 1, add.exclusions = F, translate = F)
# Recommended to turn off exlucusions: add.exclusions = F
phenotypes[1:10, 1:10]
print(dim(phenotypes))
str(phenotypes[1:10, 1:10])

# Do not use the below one. Use "phenotypes" data.
phenotypes1 = semi_join(phenotypes, genotypes, by = "id")
print(dim(phenotypes1))
print(phenotypes1[1:10, 1:10])

genotypes_1 = semi_join(genotypes, phenotypes,  by = "id")
print(dim(genotypes_1))
print(genotypes_1[1:10,])

# Missing id's
phenotypes_missing_id = anti_join(phenotypes, genotypes, by = "id")
print(dim(phenotypes_missing_id))
print(phenotypes_missing_id[1:10, 1:10])

genotypes_missing_id = anti_join(genotypes, phenotypes, by = "id")
print(dim(genotypes_missing_id))
print(phenotypes_missing_id[1:10, 1:10])

missing_ids_from_gdata1 = anti_join(phenotypes, gdata1, by = "id")
print(dim(missing_ids_from_gdata1))
print(missing_ids_from_gdata1[1:10, 1:10])

# Also, do not use this genotypes1. Use "genotypes".
genotypes1 = semi_join(genotypes, phenotypes1, by = "id")
summary(genotypes1)
str(genotypes1)
dim(genotypes1)
```

```
print(sum(is.na(genotypes1$rs17497684_C)))

print(dim(genotypes))
print(dim(phenotypes))
print(dim(covariates))

#Run the unadjusted PheWAS (unvariable analysis)
results_uni=phewas(phenotypes, genotypes
               , cores=1
               , significance.threshold=c("bonferroni")
               )

#Plot the results
pdf("phewasplot_uni_SCL9A3_rs17497684_C_caucasians.pdf")
phewasManhattan(results_uni, annotate.angle=0,
                title="Manhattan Plot for SCL9A3 & rs17497684")
dev.off()

#Add PheWAS descriptions
results_uni_d=addPhecodeInfo(results_uni)
#List the significant results
results_uni_d[results_uni_d$bonferroni&!is.na(results_uni_d$p),]
#List the top 10 results
results_uni_d[order(results_uni_d$p)[1:10],]

# Save the top 10 results
r = results_uni_d[order(results_uni_d$p)[1:10],]
write.csv(r, "Results_uni_Top10_SCL9A3_rs17497684_C_final_data.csv")

# Save the entire PheWAS Study results
write.table(results_uni_d, "phewasresults_uni_SCL9A3_rs17497684_C_final_data.tsv"
            ,quote = F, row.names = F
            , col.names = T, sep = "\t")


# Save "results" to plot the PheWAS Study later again if need be.
write.table(results_uni, "phewasresults_uni_SCL9A3_rs17497684_C_data_forplotting.tsv"
            ,quote = F, row.names = F
            , col.names = T, sep = "\t")

#Run the adjusted PheWAS with covariates
results=phewas(phenotypes, genotypes
               , covariates = covariates
               , cores=1
               , significance.threshold=c("bonferroni")
               )

#Plot the results
pdf("phewasplot_SCL9A3_rs17497684_C_covariates_caucasians.pdf")
phewasManhattan(results, annotate.angle=0,
                title="Manhattan Plot for SCL9A3 & rs17497684 with Covariates")
dev.off()
```

```
#Add PheWAS descriptions
results_d=addPhecodeInfo(results)
#List the significant results
results_d[results_d$bonferroni&!is.na(results_d$p),]
#List the top 10 results
results_d[order(results_d$p)[1:10],]

# Save the top 10 results
r = results_d[order(results_d$p)[1:10],]
write.csv(r, "Results_Top10_SCL9A3_rs17497684_C_covariates_final_data.csv")

# Save the entire PheWAS Study results
write.table(results_d, "phewasresults_SCL9A3_rs17497684_C_covariates_final_data.tsv"
            ,quote = F, row.names = F
            , col.names = T, sep = "\t")

# Save "results" to plot the PheWAS Study later again if need be.
write.table(results, "phewasresults_SCL9A3_rs17497684_C_data_forplotting.tsv"
            ,quote = F, row.names = F
            , col.names = T, sep = "\t")

# Store the first two phecodes.
phecode1 = as.character(r$phecode[1])
phecode2 = as.character(r$phecode[2])

#gdata3 = rename(gdata3, gender = SEX)
rs17497684_C_data0 = inner_join(gdata3, phenotypes,  by="id") %>% filter(!is.na(rs17497684_C))
rs17497684_C_data = left_join(rs17497684_C_data0, covariates, by="id")
print(dim(rs17497684_C_data))

#str(rs17497684_C_data[,1:10])
#head(rs17497684_C_data$"218.1")
# write.table(rs17497684_C_data, "data_rs17497684_C_phecodes_covariates_final_data.tsv"
#             ,quote = F, row.names = F
#             , col.names = T, sep = "\t")

#rs17497684_C_data_table1 = select(rs17497684_C_data, '218.1', '218', age, age2, gender, SEX, rs1749768
# Select the data variables for Creating Table1, including the top two phecode results.
rs17497684_C_data_table1 = select(rs17497684_C_data, phecode1, phecode2, age, age2,
                                  gender, SEX, rs17497684_C)
rs17497684_C_data_table1[1:10,]
print(summary(rs17497684_C_data_table1))
print(str(rs17497684_C_data_table1))

# Distribution of alleles
print(length(gdata3$rs17497684_C))
print(sum(is.na(gdata3$rs17497684_C)))
print(table(gdata3$rs17497684_C))

print(length(rs17497684_C_data_table1$rs17497684_C))
print(sum(is.na(rs17497684_C_data_table1$rs17497684_C)))
print(table(rs17497684_C_data_table1$rs17497684_C))
```

```
# Create and save table 1's

variables_names = names(rs17497684_C_data_table1)
print(variables_names)

factor_names = c(phecode1,
                 phecode2,
                 "gender",
                 "SEX",
                 "rs17497684_C")
print(factor_names)

table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs17497684_C_data_table1
                        #, strata = "SEX"
                        )

table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SCL9A3_rs17497684_C.csv")

table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs17497684_C_data_table1
                        , strata = "SEX")
table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SCL9A3_rs17497684_C_gender.csv")


table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs17497684_C_data_table1
                        , strata = "rs17497684_C")
table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SCL9A3_rs17497684_C_bySNP.csv")

table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs17497684_C_data_table1
                        , strata = c(phecode1, "SEX"))
table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SCL9A3_rs17497684_C_gender_phecode1.csv")


table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs17497684_C_data_table1
                        , strata = c(phecode2, "SEX"))
table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SCL9A3_rs17497684_C_gender_phecode2.csv")


# Create input for Flow Chart.
```

```r
sink('SCL9A3_rs17497684_C_Ch5_flow_chart.txt')

print("dim(gdata0)")
print(dim(gdata0))

print("dim(unrelated)")
print(dim(unrelated))

print("dim(unrelated1)")
print(dim(unrelated1))

print("gdata2 = inner_join(unrelated1, gdata1)")
print("dim(gdata2)")
print(dim(gdata2))

print("dim(ethnic_data0)")
print(dim(ethnic_data0))

print("ethnic_data = filter(ethnic_data0, !is.na(caucasian))")
print("dim(ethnic_data)")
print(dim(ethnic_data))

print("gdata3 = inner_join(gdata2, ethnic_data)")
print(dim(gdata3))
print(dim(gdata3))

print("dim(covariates_age_gender)")
print(dim(covariates_age_gender))

print("covariates = semi_join(covariates1, genotypes)")
print("dim(covariates)")
print(dim(covariates))

print("dim(phenotypes)")
print(dim(phenotypes))

pritn("genotypes_1 = semi_join(genotypes, phenotypes)")
print("dim(genotypes_1)")
print(dim(genotypes_1))

print("sum(is.na(genotypes1$rs17497684_C))")
print(sum(is.na(genotypes1$rs17497684_C)))

sink()
```

Below is the R code for running the entire PheWAS for the modifier gene SLC6A14 for males.

Note that the R code for gene SLC6A14 for females is extremely similar to this.

```r
# R script for running phewas for genes in the X-Chromosome.
# This is for Males
#
# Things to change for new Gene and SNP (not an exhaustive list)
# ukb_chrX_rs5905176.raw
# SLC6A14
# rs5905176
# rs5905176_G
# ChX


require(data.table)
require(PheWAS)
require(dplyr)
require(tableone)


unrelated = fread("09-kinship-degree2-unrelatedunrelated.txt", col.names = c("fid", "id"),
                  stringsAsFactors=F, header=F, na.strings=c(""," ","NA"))
#unrelated_tb_removed = fread("09-kinship-degree2-unrelatedunrelated_toberemoved.txt", col.names = c("f
#454029+ 36100
#488377- 36100

print(str(unrelated))
unrelated$id = as.numeric(unrelated$id)
unrelated1 = unrelated  %>% filter(!is.na(id)) %>% filter(id > 0 )
print(dim(unrelated1))
print(str(unrelated1))

gdata0 = fread("ukb_chrX_rs5905176.raw", stringsAsFactors=F, header=T, na.strings=c(""," ","NA"))
gdata0[1:10,]
print(dim(gdata0))
gdata1 = rename(gdata0, gender = SEX, id = IID)
gdata1[1:10,]
print(str(gdata1))

# Creating the independent data set for the genotypic data.
gdata2 = inner_join(unrelated1, gdata1, by ="id")
dim(gdata2)
gdata2[1:10,]
print(str(gdata2))
summary(gdata2)


# Reading in genetic ethnic group data ( caucasian = T/F)
ethnic_data0 = fread("ukb24727_22006_genetic_ethnic_groups.tab",
                    col.names = c("id", "caucasian"), stringsAsFactors=F,
                    header=T, na.strings=c(""," ","NA"))
dim(ethnic_data0)
ethnic_data = filter(ethnic_data0, !is.na(caucasian))
dim(ethnic_data)
print(str(ethnic_data))
summary(ethnic_data)


gdata3 = inner_join(gdata2, ethnic_data, by = "id")
print(dim(gdata3))
summary(gdata3)
```

```r
str(gdata3)
genotypes = gdata3 %>% select( id, rs5905176_G)
genotypes[1:10,]


covariates_age_gender = fread("pheno_age_sex.tab",  stringsAsFactors=F,
                              col.names = c("id", "age", "SEX"),
                              header=T, na.strings=c(""," ","NA"))
print(dim(covariates_age_gender))
covariates1 = mutate(covariates_age_gender, age2 = age**2)
covariates = semi_join(covariates1, genotypes, by = "id")
print(dim(covariates))
summary(covariates)
print(str(covariates))
print(covariates[1:10,])

males = filter(covariates, SEX == 1)
print(dim(males))
print(str(males))

covariates = males %>% select(-SEX)
print(dim(covariates))
print(str(covariates))

genotypes = semi_join(genotypes, males, by = "id")

phedata = fread("icd10_data_with_phecodes2.txt", stringsAsFactors=F, header=T,
                na.strings=c(""," ","NA"))
head(phedata)
print(str(phedata))
summary(phedata)
dim(phedata)

phedata0 = phedata %>% rename(phecode = code) %>% select(id, phecode, count)
head(phedata0)
str(phedata0)
print(dim(phedata0))
summary(phedata0)

phenotypes=createPhewasTable(phedata0, min.code.count = 1, add.exclusions = F, translate = F)
# Recommended to turn off exlucusions: add.exclusions = F
phenotypes[1:10, 1:10]
print(dim(phenotypes))
str(phenotypes[1:10, 1:10])

phenotypes = semi_join(phenotypes, males, by = "id")

# Do not use the below one. Use "phenotypes" data.
phenotypes1 = semi_join(phenotypes, genotypes, by = "id")
print(dim(phenotypes1))
print(phenotypes1[1:10, 1:10])

genotypes_1 = semi_join(genotypes, phenotypes,  by = "id")
```

```
print(dim(genotypes_1))
print(genotypes_1[1:10,])


# Missing id's
phenotypes_missing_id = anti_join(phenotypes, genotypes, by = "id")
print(dim(phenotypes_missing_id))
print(phenotypes_missing_id[1:10, 1:10])

genotypes_missing_id = anti_join(genotypes, phenotypes, by = "id")
print(dim(genotypes_missing_id))
print(phenotypes_missing_id[1:10, 1:10])

missing_ids_from_gdata1 = anti_join(phenotypes, gdata1, by = "id")
print(dim(missing_ids_from_gdata1))
print(missing_ids_from_gdata1[1:10, 1:10])




# Also, do not use this genotypes1. Use "genotypes".
genotypes1 = semi_join(genotypes, phenotypes1, by = "id")
summary(genotypes1)
str(genotypes1)
dim(genotypes1)
print(sum(is.na(genotypes1$rs5905176_G)))

print(dim(genotypes))
print(dim(phenotypes))
print(dim(covariates))


#Run the unadjusted PheWAS (unvariable analysis)

results_uni=phewas(phenotypes, genotypes
               , cores=1
               , significance.threshold=c("bonferroni")
               )


#Plot the results
pdf("phewasplot_uni_SLC6A14_rs5905176_G_males.pdf")
phewasManhattan(results_uni, annotate.angle=0,
                title="Manhattan Plot for SLC6A14 & rs5905176 for Males")
dev.off()

#Add PheWAS descriptions
results_uni_d=addPhecodeInfo(results_uni)
#List the significant results
results_uni_d[results_uni_d$bonferroni&!is.na(results_uni_d$p),]
#List the top 10 results
results_uni_d[order(results_uni_d$p)[1:10],]
```

```r
# Save the top 10 results
r = results_uni_d[order(results_uni_d$p)[1:10],]
write.csv(r, "Results_uni_Top10_SLC6A14_rs5905176_G_males.csv")

# Save the entire PheWAS Study results
write.table(results_uni_d, "phewasresults_uni_SLC6A14_rs5905176_G_males.tsv"
            ,quote = F, row.names = F
            , col.names = T, sep = "\t")


# Save "results" to plot the PheWAS Study later again if need be.
write.table(results_uni, "phewasresults_uni_SLC6A14_rs5905176_G_data_forplotting_males.tsv"
            ,quote = F, row.names = F
            , col.names = T, sep = "\t")




#Run the adjusted PheWAS with covariates
results=phewas(phenotypes, genotypes
               , covariates = covariates
               , cores=1
               , significance.threshold=c("bonferroni")
               )


#Plot the results
pdf("phewasplot_SLC6A14_rs5905176_G_covariates_males.pdf")
phewasManhattan(results, annotate.angle=0,
                title="Manhattan Plot for SLC6A14 & rs5905176 with Covariates for Males")
dev.off()

#Add PheWAS descriptions
results_d=addPhecodeInfo(results)
#List the significant results
results_d[results_d$bonferroni&!is.na(results_d$p),]
#List the top 10 results
results_d[order(results_d$p)[1:10],]

# Save the top 10 results
r = results_d[order(results_d$p)[1:10],]
write.csv(r, "Results_Top10_SLC6A14_rs5905176_G_covariates_males.csv")

# Save the entire PheWAS Study results
write.table(results_d, "phewasresults_SLC6A14_rs5905176_G_covariates_males.tsv"
            ,quote = F, row.names = F
            , col.names = T, sep = "\t")


# Save "results" to plot the PheWAS Study later again if need be.
write.table(results, "phewasresults_SLC6A14_rs5905176_G_data_forplotting_males.tsv"
            ,quote = F, row.names = F
            , col.names = T, sep = "\t")
```

```r
# Store the first two phecodes.
phecode1 = as.character(r$phecode[1])
phecode2 = as.character(r$phecode[2])

#gdata3 = rename(gdata3, gender = SEX)
rs5905176_G_data0 = inner_join(gdata3, phenotypes,  by="id") %>% filter(!is.na(rs5905176_G))
rs5905176_G_data = inner_join(rs5905176_G_data0, covariates, by="id")
print(dim(rs5905176_G_data))

#str(rs5905176_G_data[,1:10])
#head(rs5905176_G_data$"218.1")
# write.table(rs5905176_G_data, "data_rs5905176_G_phecodes_covariates_final_data.tsv"
#             ,quote = F, row.names = F
#             , col.names = T, sep = "\t")

#rs5905176_G_data_table1 = select(rs5905176_G_data, '218.1', '218', age, age2, gender, SEX, rs5905176_G
# Select the data variables for Creating Table1, including the top two phecode results.
rs5905176_G_data_table1 = select(rs5905176_G_data, phecode1, phecode2, age, age2, rs5905176_G)
rs5905176_G_data_table1[1:10,]
print(summary(rs5905176_G_data_table1))
print(str(rs5905176_G_data_table1))


# Distribution of alleles
print(length(gdata3$rs5905176_G))
print(sum(is.na(gdata3$rs5905176_G)))
print(table(gdata3$rs5905176_G))

print(length(rs5905176_G_data_table1$rs5905176_G))
print(sum(is.na(rs5905176_G_data_table1$rs5905176_G)))
print(table(rs5905176_G_data_table1$rs5905176_G))


# Create and save table 1's

variables_names = names(rs5905176_G_data_table1)
print(variables_names)

factor_names = c(phecode1,
                 phecode2,
                 "rs5905176_G")
print(factor_names)


table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs5905176_G_data_table1
                        #, strata = "SEX"
                        )

table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SLC6A14_rs5905176_G.csv")
```

```
# table1 = CreateTableOne(vars = variables_names
#                         , factorVars = factor_names
#                         , data = rs5905176_G_data_table1
#                         , strata = "gender")
# table_1 = print(table1, showAllLevels = T)
# write.csv(table_1, file = "Table1_final_SLC6A14_rs5905176_G_gender.csv")


table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs5905176_G_data_table1
                        , strata = "rs5905176_G")
table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SLC6A14_rs5905176_G_bySNP.csv")

table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs5905176_G_data_table1
                        , strata = c(phecode1))
table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SLC6A14_rs5905176_G_phecode1_males.csv")


table1 = CreateTableOne(vars = variables_names
                        , factorVars = factor_names
                        , data = rs5905176_G_data_table1
                        , strata = c(phecode2))
table_1 = print(table1, showAllLevels = T)
write.csv(table_1, file = "Table1_final_SLC6A14_rs5905176_G_phecode2_males.csv")


# Create input for Flow Chart.

sink('SLC6A14_rs5905176_G_ChX_males_flow_chart.txt')

print("dim(gdata0)")
print(dim(gdata0))

print("dim(unrelated)")
print(dim(unrelated))

print("dim(unrelated1)")
print(dim(unrelated1))

print("gdata2 = inner_join(unrelated1, gdata1)")
print("dim(gdata2)")
print(dim(gdata2))

print("dim(ethnic_data0)")
print(dim(ethnic_data0))

print("ethnic_data = filter(ethnic_data0, !is.na(caucasian))")
print("dim(ethnic_data)")
```

```r
print(dim(ethnic_data))

print("gdata3 = inner_join(gdata2, ethnic_data)")
print(dim(gdata3))
print(dim(gdata3))

print("dim(covariates_age_gender)")
print(dim(covariates_age_gender))

print("dim(males)")
print(dim(males))

print("covariates = semi_join(covariates1, genotypes)")
print("dim(covariates)")
print(dim(covariates))

print("dim(phenotypes)")
print(dim(phenotypes))

pritn("genotypes_1 = semi_join(genotypes, phenotypes)")
print("dim(genotypes_1)")
print(dim(genotypes_1))

print("sum(is.na(genotypes1$rs5905176_G))")
print(sum(is.na(genotypes1$rs5905176_G)))

sink()
```

END OF REPORT