Phenome-wide association study of Cystic Fibrosis Modifier Genes

Faizan Khalid Mohsin

October 25, 2018

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.

Methods: A PheWAS study was performed for the following modifier genes and their respective SNP's: 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 data registry, which has over 500,000 participants, to perform the PheWAS study using the ICD10 codes that were converted into phenotypes. A logistic regression model was used for finding associations between phenotypes and the modifier genes with phenotypes modeled as binary outcome variables 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 performed. The model was adjusted for covariates age, age-square and sex. Since, there were 1511 phenotypes, 1511 logistic regression were performed for each gene and a bonferroni correction was applied to determine statistically significant associations giving a corrected p-value of $3.309\text{x}10^{-5}$. Further, separate analysis was performed for males and females for the gene SLC6A14 as it is found on X chromosome.

Findings: For the adjusted analysis for SCL9A3 with SNP rs57221529, we was found it to be statistically associated with having Esophagitis, GERD and related diseases (OR = 1.064, S.E. = 0.013, p-value = $1.79\text{x}10^{-6}$). Further, for each allele count (0, 1, 2) the number of cases were 7% to 8% of the controls with the minimum number of cases being 784 for allele count 2. With a total of 19,687 cases and 243,236 controls. For males the gene SLC6A14 at SNP rs3788766 was statistically associated with having Urinary obstruction (OR = 1.68, S.E. = 0.127, p-value = $4.24\text{x}10^{-5}$). Further, for each allele count (0 or 2 as males cannot have one allele as this gene is on the X chromosome) there were only a total of 64 cases and 117,334 controls, with 37 cases for allele count 2 vs. 27 for allele count 0. For females no statistically significant association was found. Lastly, for the modifier gene SLC26A9 at SNP rs4077469 no association with any of the phenotypes was found. Very similar results were attained for the unadjusted analysis.

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 a 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 is possibly a sporadic relationship and not a real association. Hence, it is possible that there is an association between the gene SCL9A3 near the location SNP rs57221529 with having Esophagitis, GERD and related diseases, since we already know that CF affects the digestive system.

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 lung, sweat glands, vas deferens, and pancreas (Cohn et al. 1998).

2.2 Modifier Genes

Now, there have been identified common variation in several genes (genes that have nothing to do with CFTR gene) that contribute to CF disease severity. These genes are known as modifier genes and the majority of these genes are transporters. However, it is unknown wheather variation in these genes impact phenotypes (all physical and observable characteristics) in individuals who do not have CF, that is, individuals without two mutations in the CF causal gene. Hence, the primary goal of this papar is to see the effect of variation in modifier genes in the general public. We do this by carrying out a Phenome-wide association study (PheWAS).

We do this using 500,000 individuals from the UKBiobank who have been genotyped genome-wide and have detailed, comprehensive phenotypic data, by carrying out a Phenome-wide association study (PheWAS).

2.3 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. Due to the large number of associations tested, a Bonferroni correction is applied to identify statistically significant association.

Ultimately, we will have an improved understanding of the phenotype associated with normal variation in these genes of interest; genes which, with a background of CFTR mutations, can cause severe disease. Understanding of 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.

Statistical methods/analyses to be employed and level of familiarity needed for these methods: Statistical methods implemented will be multiple logistic and linear regression, principal component analysis, exploratory data analysis.

3 Material and Methods

3.1 Data

3.1.1 UK Biobank data

The UK Biobank is a large-scale, population-based, prospective cohort that enrolled over 500,000 participants between the ages 40 and 69 years. Externsive amounts of data is collected for each recruited participant using a varied 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. Their national health records have also been linked with the baseline and genotypic data obtaining extensive healthcare information on the participants, allowing for longitudinal follow-up. Genotypic and phenotypic data used in this study were

obtained from UK Biobank under an approved data request application (application ID: 10775) (Sudlow et al. 2015) (X. Li et al. 2018)

3.1.2 Genotypic data

For the purpose of this study we are interested in three modifier genes SCL26A9, SLC6A14 and SCL9A3. In particular, we are interested in the variation at particular SNP's in each gene. However, the SNP's we are interested in were not present in the UK Biobank. Hence, instead of imputing the SNP's, we decided to use substitute SNP's that have a high correlation with the SNP's of interest and were present in Uk Biobank. Since imputing SNP's has uncertainty, we thought that it would be more reliable to find SNP's with high correlation present in the Uk Biobank and use those in the analysis (see Table 1).

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

Note that the chosen SNP only serve as a proxy for the location in the gene and is a representative of the region in the gene.

Downloading the biobank data http://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=accessing_data_guide HES Data https://biobank.ctsu.ox.ac.uk/showcase/refer.cgi?id=2406 HES ICD data uk biobank http://biobank.ctsu.ox.ac.uk/crystal/docs/UsingUKBData.pdf Info on ICD https://biobankengine.stanford.edu/faq Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics ICD10 https://biobank.ctsu.ox.ac.uk/crystal/coding.cgi?id=19&nl=1 Main ICD10 http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=41202 ICD10 Show ICD codes http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202 More data coding: https://biobank.ctsu.ox.ac.uk/crystal/coding.cgi?id=19&nl=1 Secondary ICD10 http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=41204 Convert Ukbiobank to ICD codes https://www.rdocumentation.org/packages/ukbtools/versions/0.11.0 Ukbiobank ukbtools Try getting icd codes from here. https://rdrr.io/cran/ukbtools/man/ukb_icd_code_meaning.html https:// cloud.r-project.org/web/packages/ukbtools/vignettes/explore-ukb-data.html This may work: ukb_df(fileset, path = ".", n_threads = "dt", data.pos = 2) Link: https://rdrr.io/cran/ukbtools/man/ukb_df.html Paper on this: https://www.biorxiv.org/content/biorxiv/early/2017/06/30/158113.full.pdf Parent Category http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=2022

3.1.3 Phenotypic data

3.1.3.1 Phenotyping and mapping ICD-10 or ICD-9 to phecode

obtained using the ICD-10 codes from the UK Biobank registry.

We analysed two phenotypic data sets in the UK Biobank using the phecode schema (see online supplementary text for phenotyping and mapping process) (Denny et al. 2013).

We focused on phenotypes in relation to diagnostic disease outcomes (either an individual had or had not a disease outcome). The coding for clinical diagnoses in these data sets followed the WHO's International Classification of Diseases (ICD) coding systems, but used different ICD versions (ICD-10 or ICD-9) according to the date of record. We included only ICD-10 codes, primarily for simplicity and also because most of the hospitals in the UK transitioned to ICD-10 codes, leaving very few people in the early period of the study recruitment period with ICD-9 codes (X. Li et al. 2018).

The ICD-10 codes were converted to phecodes using the phecode schema and code provided by the supervisor, in (Denny et al. 2013).

We included only ICD-10 codes to define the case and control groups,

https://ard.bmj.com/content/77/7/1039

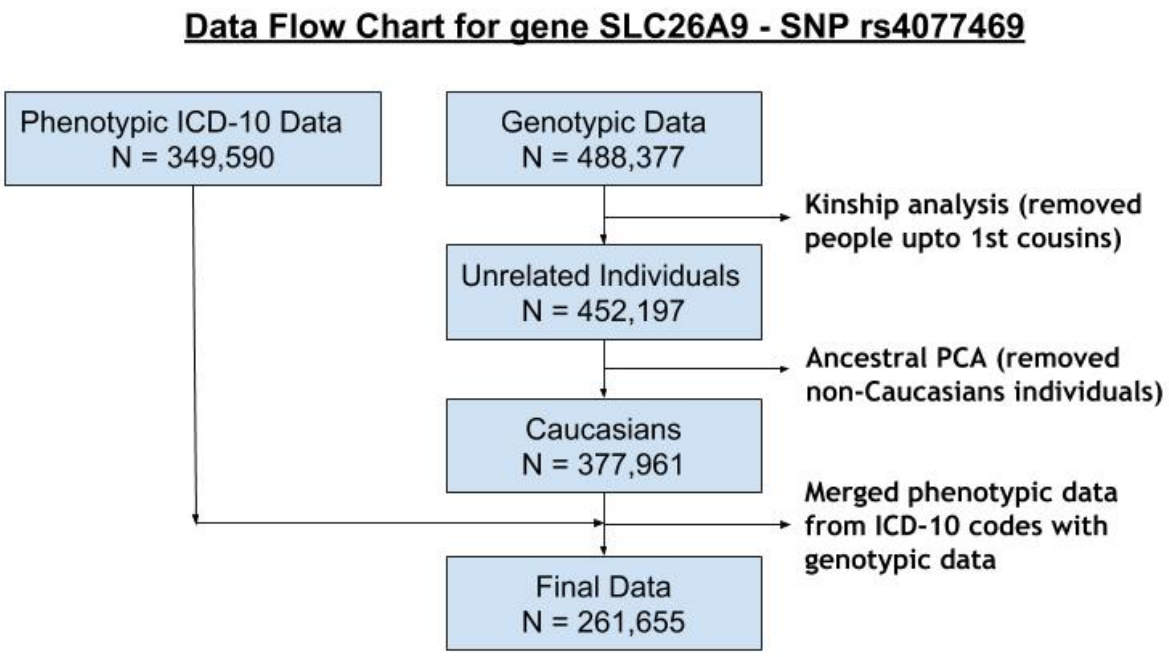

Figure 1: Data Flow Chart for gene SLC26A9 - SNP rs4077468

Similarly, the final data sets for the genes SCL9A3 and SLC6A14 were 262,923 and 117,398 respectively.

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

Further, as can be seen, all and any missing data was deleted. For example, many people with ICD-10 codes in the UK biobank data did not have genotypic data for different genes and 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 simply means they did not visit a hospital before we received the UK Biobank data.

3.1.4 Study Design

3.1.5 Primary Endpoints

3.1.6 Secondary Endpoints

Is the study vetted by a research Ethics committee. Does it have proper consent. What is the effect size. What is your population.Is analysis generalizable. unit of analysis. genes, patient, How did you handle the missing data. How much cleaning was involved in the papers.

Write a line a two about the quality of the data. The sample was not based on power calculations but rather the priori available data. Talk about the assumptions of the models. Talk about

If the models or methods are on the novel side provide more The methods section should have a flow diagram showing how many you started with and how many did you end with. conhort flow chart. Patients you started with and the patients you analysed.

3.2 Statisal Methods

3.2.1 Hardy-Weinburg Equilibrium

In the paper (Blackman et al. 2013) the risk allele at SNP rs4077468 was A and G was the other allele.

From our results, for both phenotypes or traits, or the disease Uterine leiomyoma (phecode 218.1) and disease Benign neoplasm of uterus (phecode 218), the allele T for SNP rs4077469 was found to be protective, (OR = 0.9222197 and 0.9254563, respectively).

Hardy-Weinburg Equilibrium, reason, there are certain assumptions. THe assumptions are that if have a gene pool of alleles for SNP Table1_SCL9A3_rs17497684_C_Dist. Alleles C and A, Assume: Equal mixing of individuals in this gene pool, then we should have a ditribution that follows HWE: $p**2 + 2pq + q**2 = 1$.

Assumption for equal mixing. Random mating No migration Basically, want to look into this is for quality control. If the SNP is faulty, just within in the controls we would not see HWE. If we have a twice as risk for the phenotype. additive model has highest power when tested genotype - paper. M

3.2.2 PheWAS Study:

R "pheWAS" package details.

A logistic regression was done with the number of recessive allele as the predictor, which was modeled as an additive model (Carroll, Bastarache, and Denny 2014).

$$logit(p\_{phenotype\_i}) = SLC26A9 + age + age^2 + sex$$

where

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

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

Note that SLC26A9 is equal to the number of T alleles at SNP rs4077469.

Plan of action:

Also need to do the phewas with rs4077469 for icd9 codes.

The model used for calculating the p-values is a univariate logistic regression without any covariates. The predictor variable is the number of T's the person has at the SNP. The outcome variable is if the person had the icd10 condition or not. Cases for a particular icd10 condition, are all the people with the condition, and the controls are all the other people. No exclusion criteria was implemented (for each icd10 condition, everyone was either a case or a control).

Follow-up plan: 1. Add gender covariate, as this is a gender specific disease. 2. Add age and ancestral PCA covariates. 3. Is this a spurious association? 4. Recheck and make sure the ICD10 mapping to the phecodes is done properly. 5. Try to understand what the "count" variable in the pheWAS package for the phenotypic data input is. The phenotypic data has three variables "ID", "icd9"/"phecodes" and "count". 6. Get another SNP which we know causes a disease and then do a PheWAS to see if we get the known association. 7. Need to remove people with CF. 8. Make sure the model for association in the PheWAS package is properly understood. 9. Reperform this PheWAS using the original SNP rs4077468 (imputed) instead of the substitute SNP: rs4077469 (r = 1). 10. Instead of using ICD10, use ICD9 codes and see if get the same result. 11. Once a good, validated model is established, repeat this PheWAS for the other SNP's: rs3788766 and SNP: rs57221529.

3.2.3 UK Biobank Data Validation.

Performing negative and positive controls to check if data is fine. Perform PheWAS study on a SNP i.e. rs3135388, and this should give a statistically significant result for the phenotype multiple sclerosis. Perform similar PheWAS study that of known SNP and pheno-types with know associations. Make sure that all the results are as expected. rs3135388 would be an example of a postive control. cite source establishing this relationship.

4 Results

4.1 Modifier Gene SLC9A3

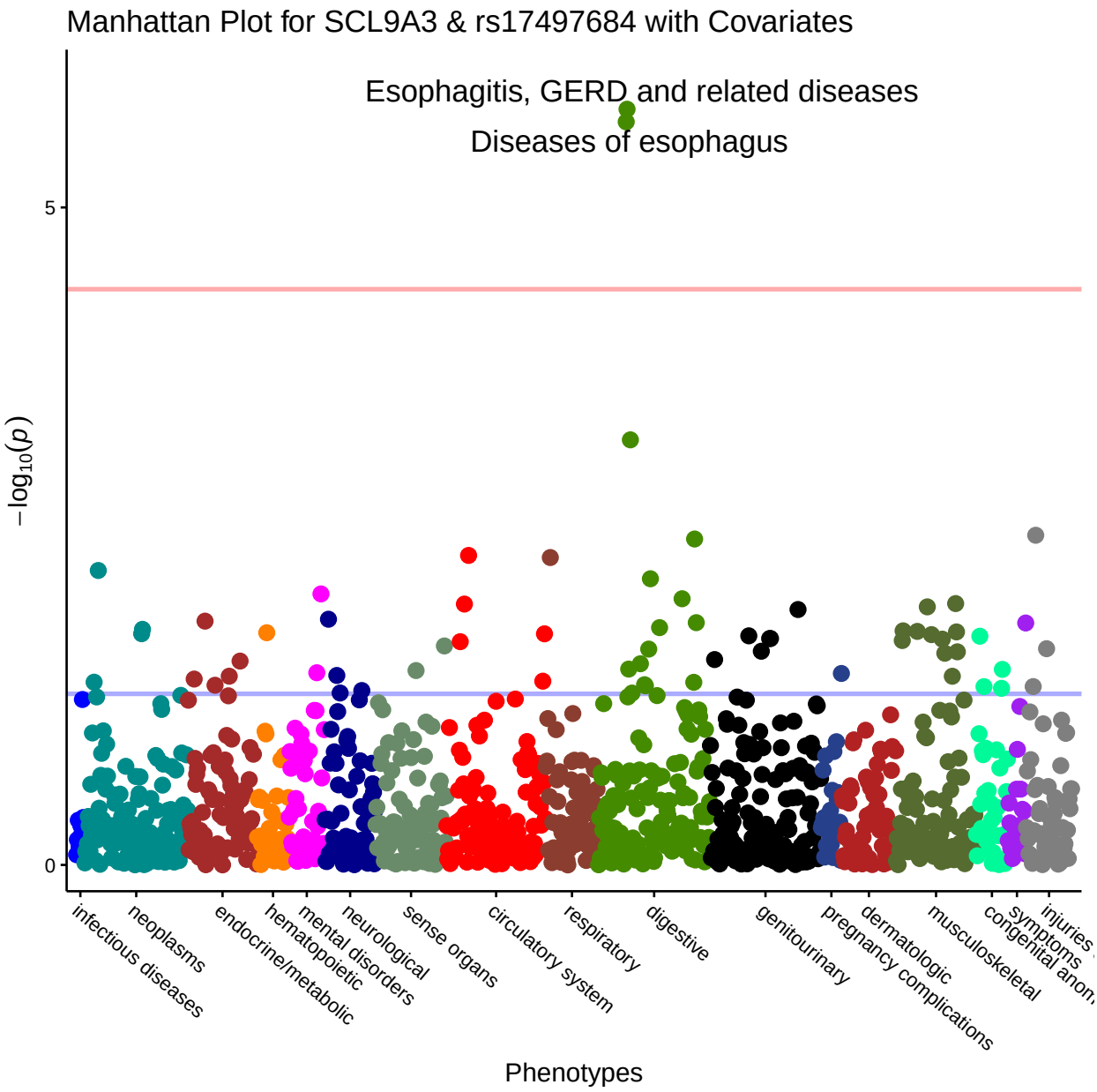

Figure 2: PheWAS Manhatten Plot with covariates.

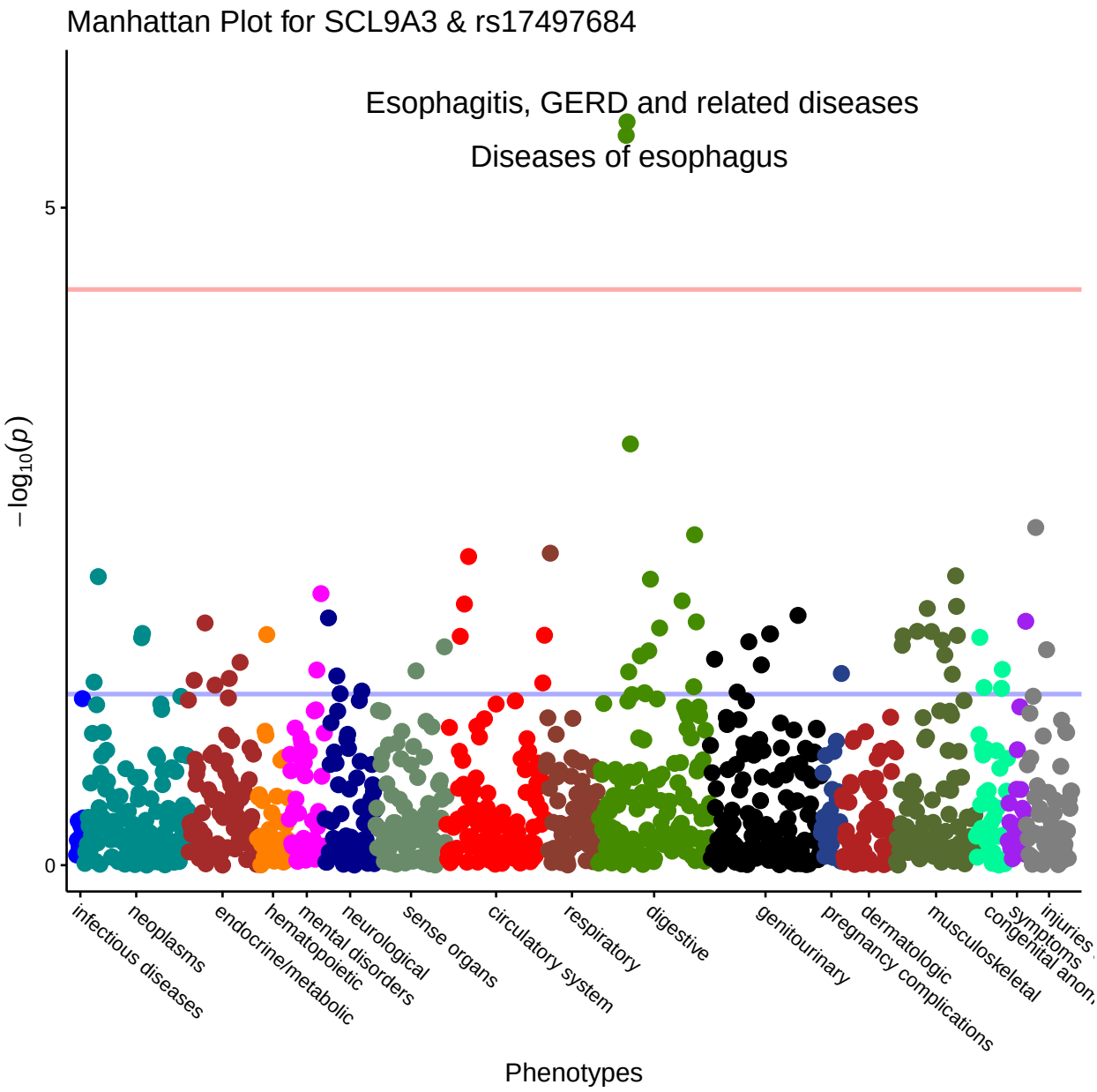

Figure 3: PheWAS without covariates.

4.2 Modifier Gene SLC6A14 for Males

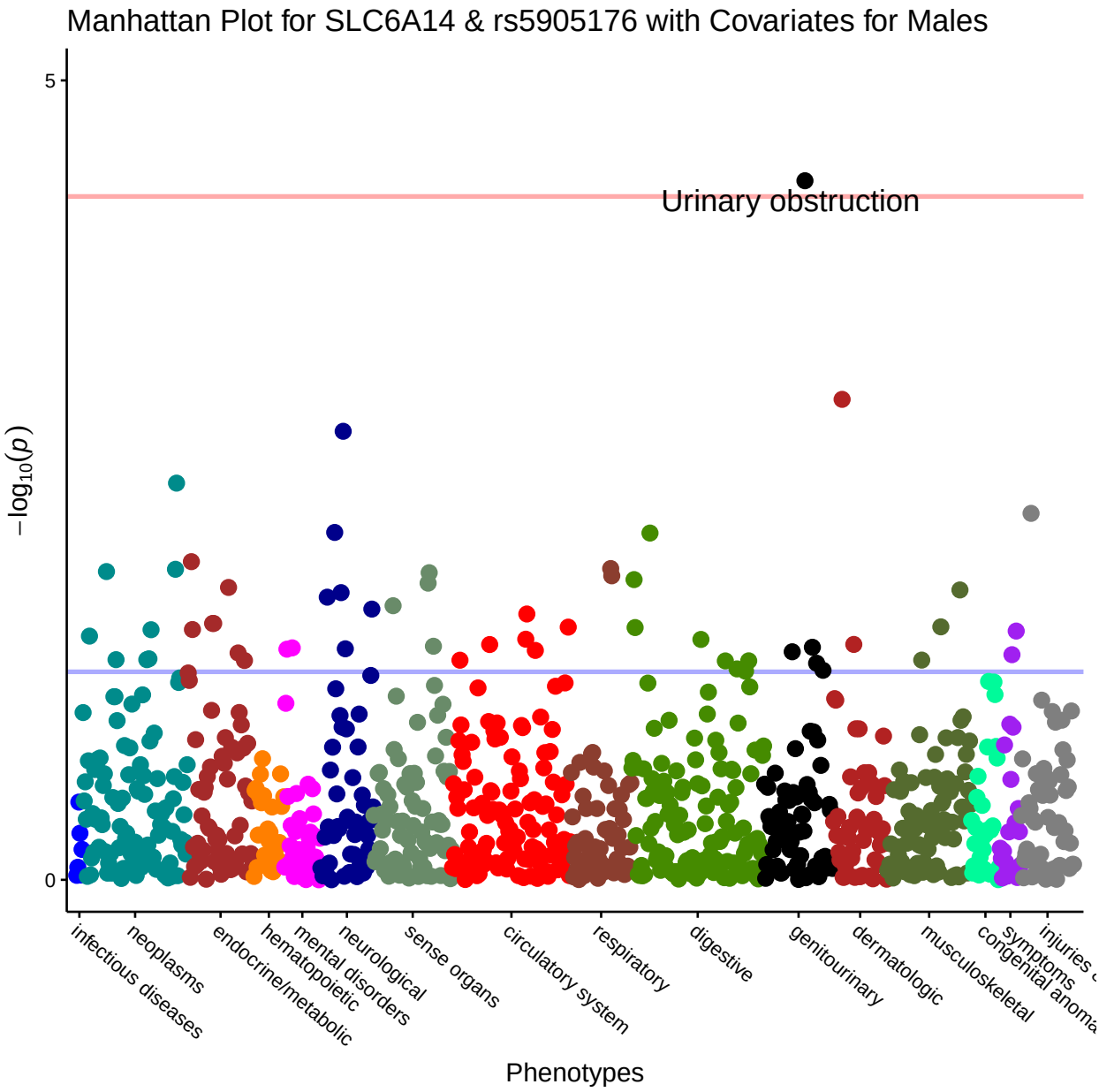

Figure 4: PheWAS with covariates.

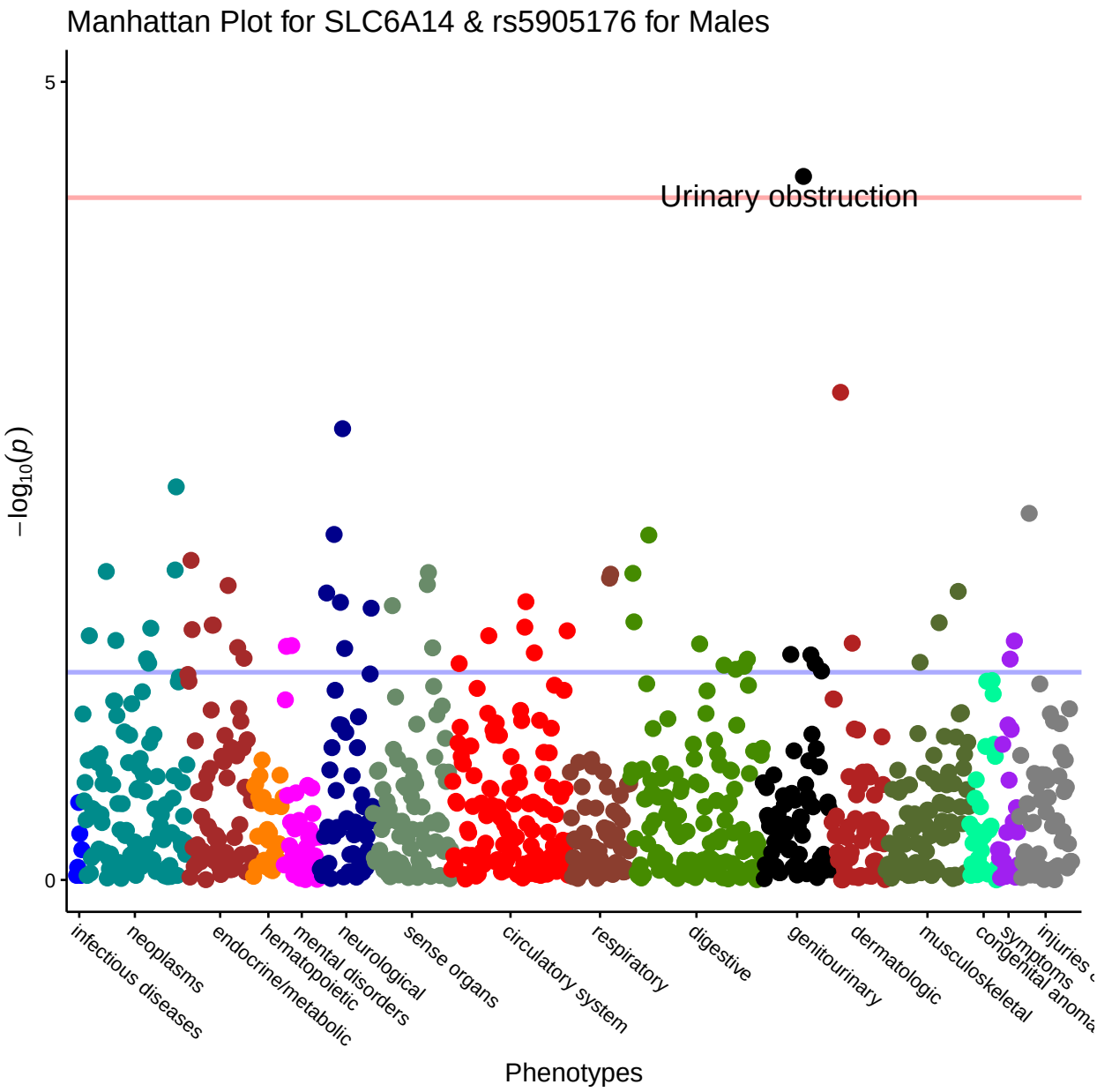

Figure 5: PheWAS without covariates.

4.3 Modifier Gene SLC6A14 for Females.

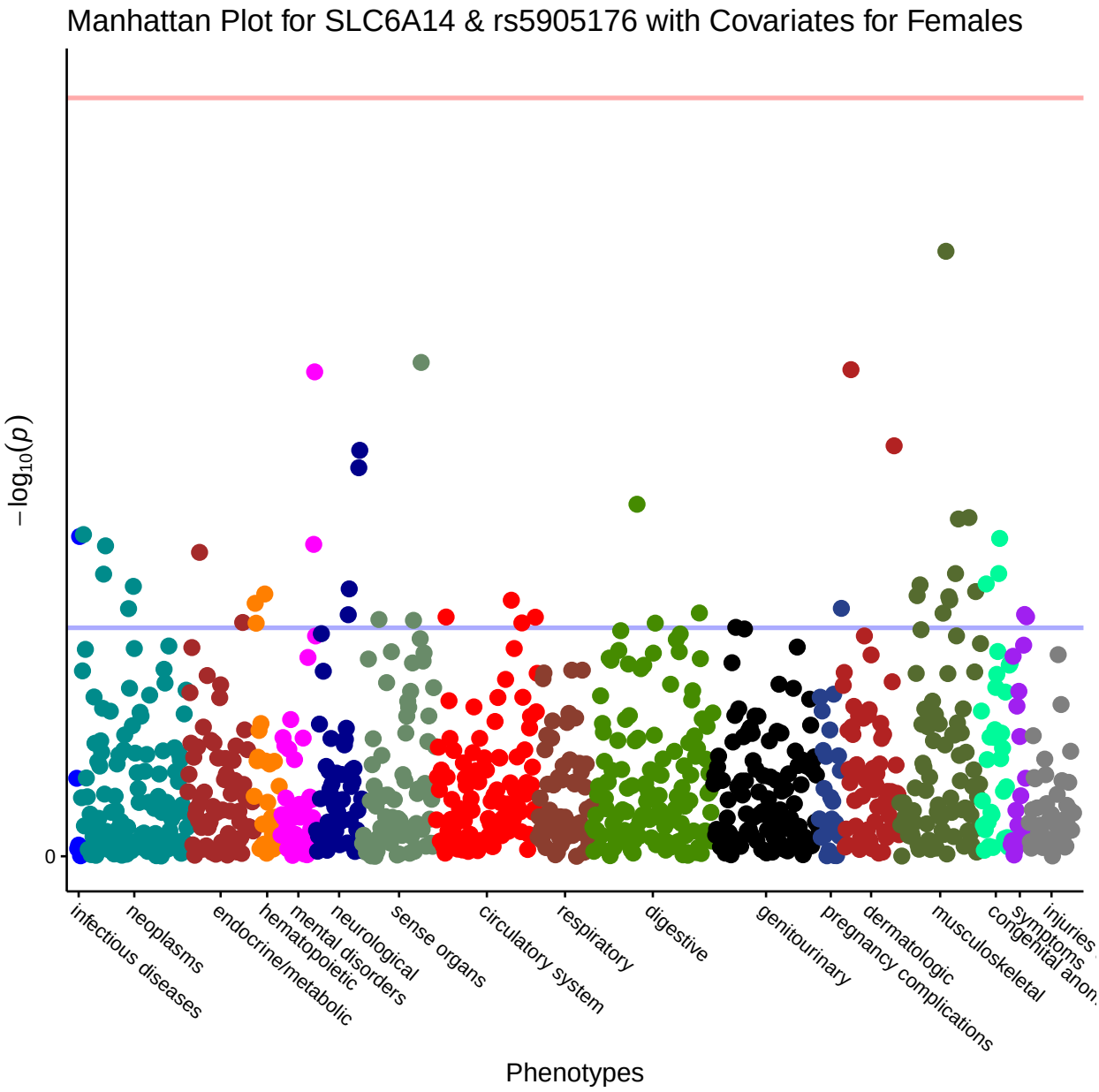

Figure 6: PheWAS with covariates.

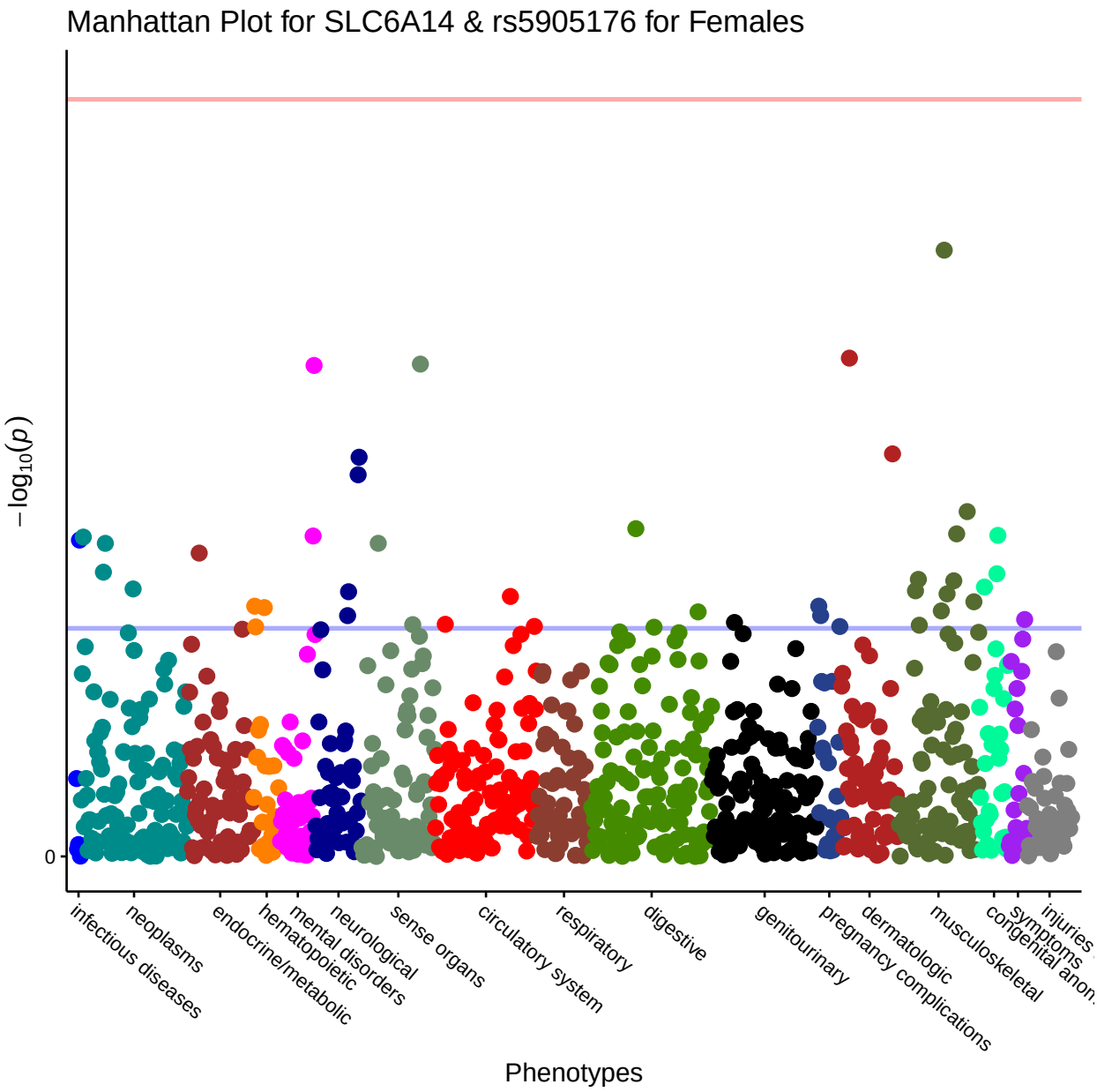

Figure 7: PheWAS without covariates.

4.4 Modifier Gene SLC26A9

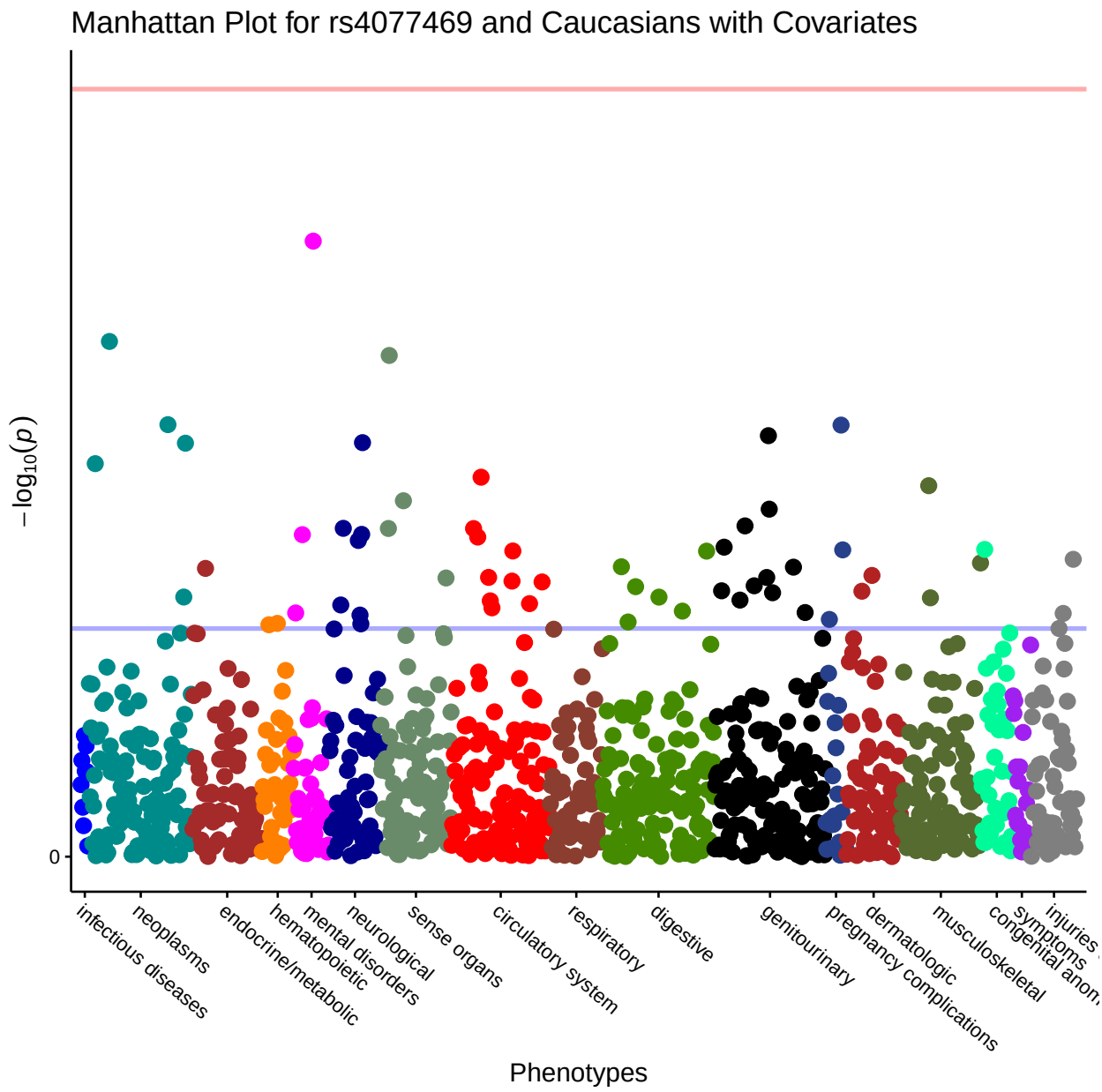

Figure 8: PheWAS with covariates after removing non-Caucasians and related people.

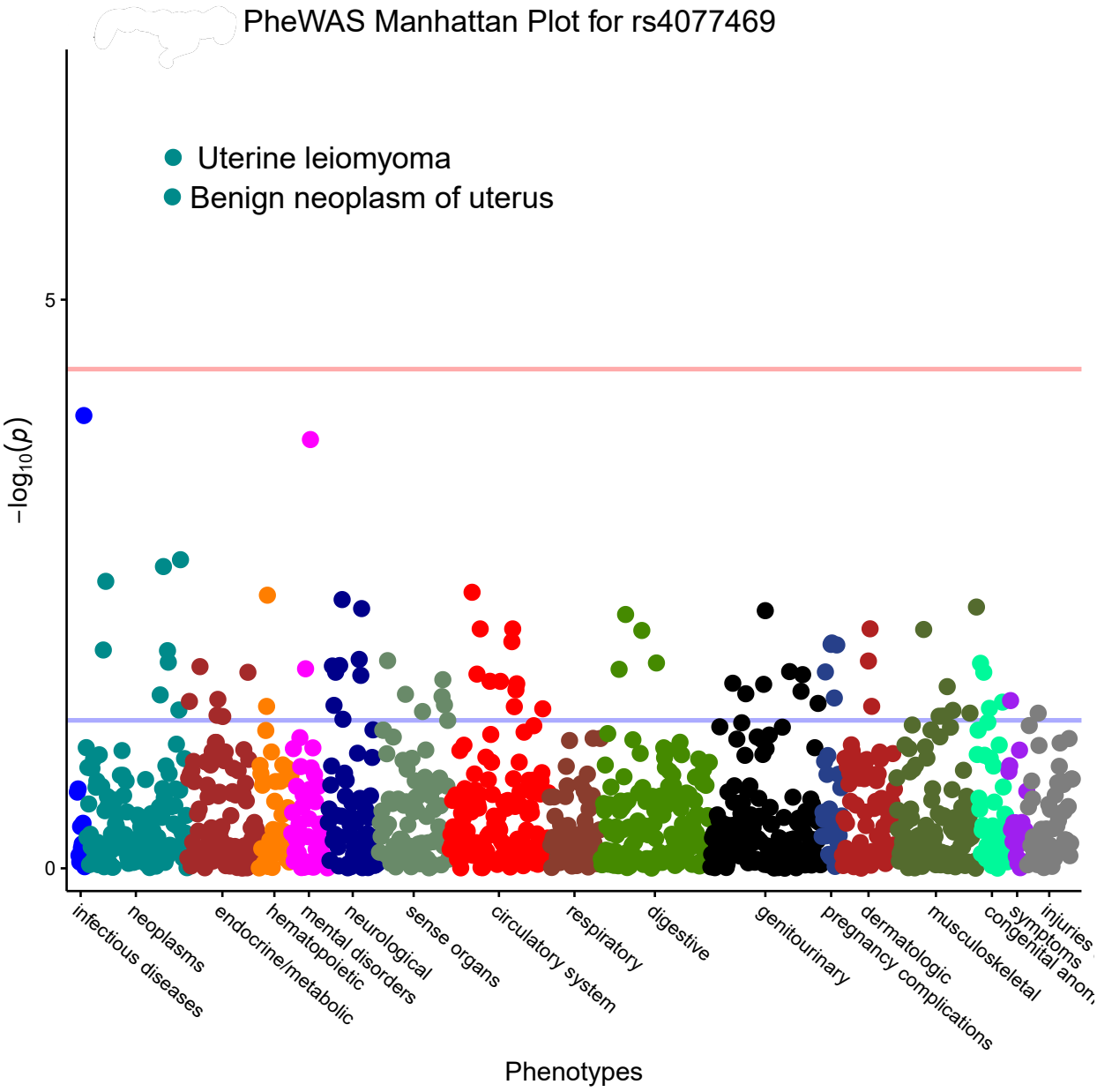

Figure 9: PheWAS without covariates, and without removing non-Caucasians and related people.

Table 2: Top 10 phenotypes associated with gene SLC26A9 and SNP rs4077469 by p-value

phecode	description	group	beta	SE	OR	p	n_total	n_cases	n_controls	HWE_p	allele_freq	bonferroni
300.12	Agorophobia, social phobia, and panic disorder	mental disorders	-0.5061935	0.1402922	0.6027857	0.0003084	261655	126	261529	0.9894264	0.3937666	FALSE
159.40	Malignant neoplasm of retroperitoneum and peritoneum	neoplasms	-0.4027249	0.1238978	0.6684959	0.0011522	261655	155	261500	0.9894264	0.3937666	FALSE
362.80	Retinal hemorrhage/ischemia	sense organs	-1.0686904	0.3341903	0.3434580	0.0013846	261655	30	261625	0.9894264	0.3937666	FALSE
208.00	Benign neoplasm of colon	neoplasms	0.0350976	0.0119958	1.0357208	0.0034356	261655	15656	245999	0.9894264	0.3937666	FALSE
653.00	Problems associated with amniotic cavity and membranes	pregnancy complications	0.4034860	0.1379881	1.4970342	0.0034549	261655	105	261550	0.9894264	0.3937666	FALSE
609.20	Abnormal spermatozoa	genitourinary	-0.4133914	0.1435154	0.6614034	0.0039709	261655	116	261539	0.9894264	0.3937666	FALSE
348.90	Other conditions of brain, NOS	neurological	0.4338051	0.1521318	1.5431181	0.0043512	261655	87	261568	0.9894264	0.3937666	FALSE
225.20	Benign neoplasm of spinal cord, meninges	neoplasms	-0.7004571	0.2458440	0.4963584	0.0043830	261655	45	261610	0.9894264	0.3937666	FALSE
149.10	Cancer of oropharynx	neoplasms	-0.2812049	0.1017883	0.7548736	0.0057335	261655	220	261435	0.9894264	0.3937666	FALSE
425.12	Other hypertrophic cardiomyopathy	circulatory system	-0.5457999	0.2018495	0.5793782	0.0068512	261655	62	261593	0.9894264	0.3937666	FALSE

Most important findings should be presented in your tables and figures.

report everything you did. If you did 10 analysis then report all of them.

Tables should be self contained. Legend of tables and figures should spell out all accronyms.

fastidious.

For graphs or tables, never assume audience can see colour. Lines should have different paterns.

5 Discussion

State upfront all the limits of the study.

Did not use the ICD-9 codes in the UK Biobank registry. Hence, could have introduced bias in the data by not including people with disease codes in the early stage of the study. 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. For example, many people with ICD-10 codes in the UK biobank data did not have genotypic data for different genes and 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 simply means they did not visit a hospital before we received the UK Biobank data.

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.

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.

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.

7 Additional information

7.1 Competing financial interests

The authors declare no competing interests.

8 Appendix

In total, the main part of the report, should be roughly 10-20 pages. This does not include the Appendix. Is the reference included.

Below is the script for running the entire PheWAS for a particular modifier gene.

```
####
#  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"), stringsAsFac

# 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")
```

```r
# 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))
```

```r
#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, rs1749
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.

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()
```