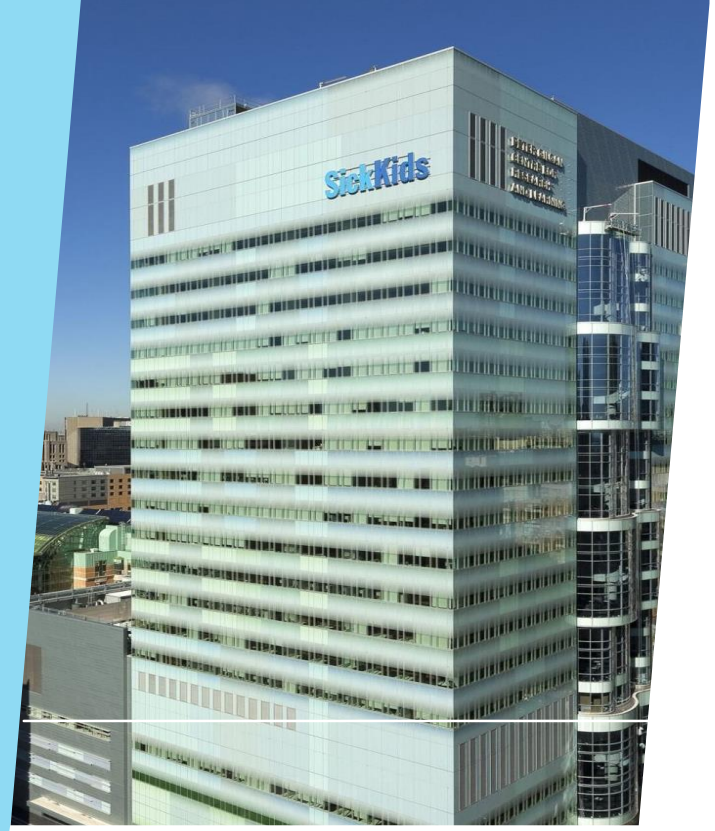
# Phenome-wide Association Study of Cystic Fibrosis Modifier Genes

Supervisor: Dr. Lisa Strug

Faizan Khalid Mohsin




# The Hospital for Sick Children

SickKids®
RESEARCH INSTITUTE

## Strug Lab

# Overview

- Background: Cystic Fibrosis and Modifier Genes
- Research Question
- What is a PheWAS?
- UK Biobank Database
- Statistical Methodology and Results
- Limitations and Challenges
- Future Work

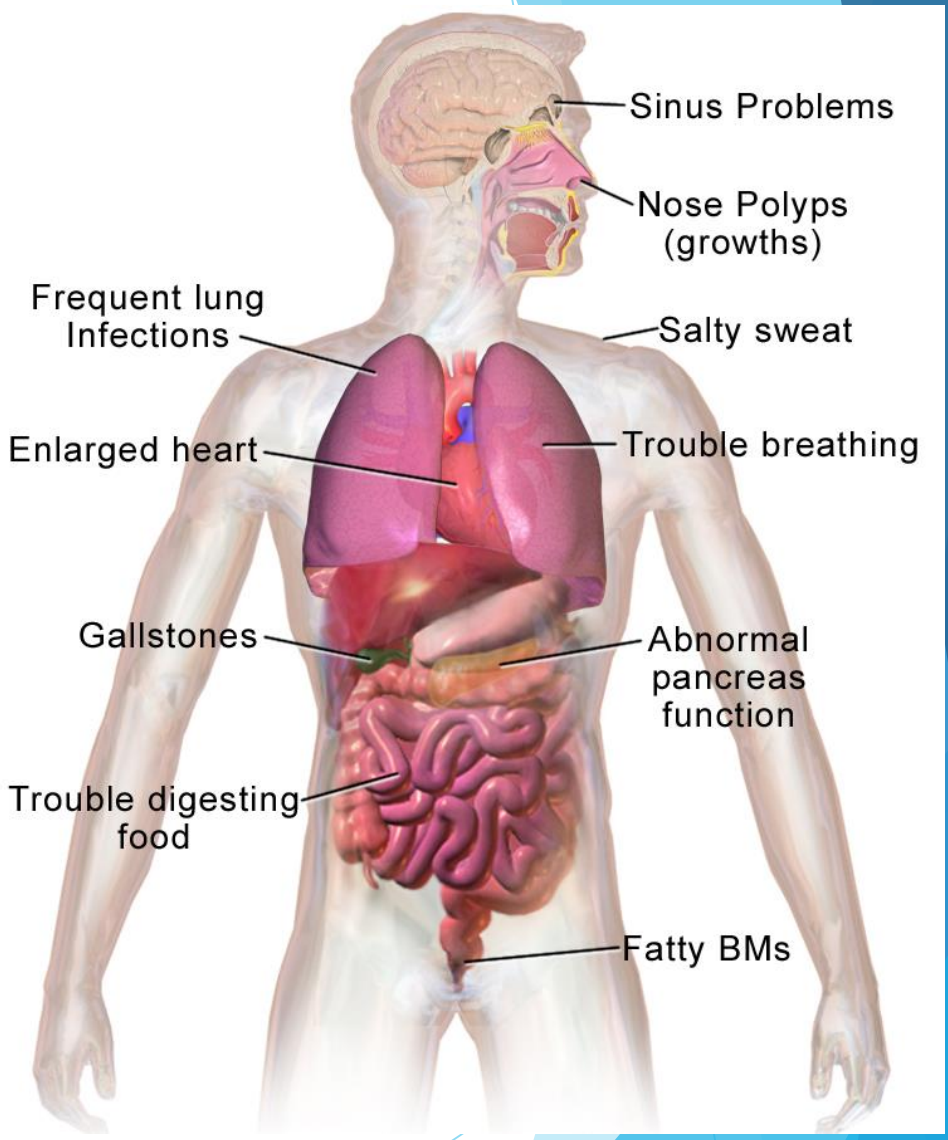
# BACKGROUND

## Cystic Fibrosis

- Cystic fibrosis (CF) is the most common fatal genetic disease affecting Canadian children and young adults. At present, there is no cure.
- Commonly suffer from lung disease.

## Phenotype

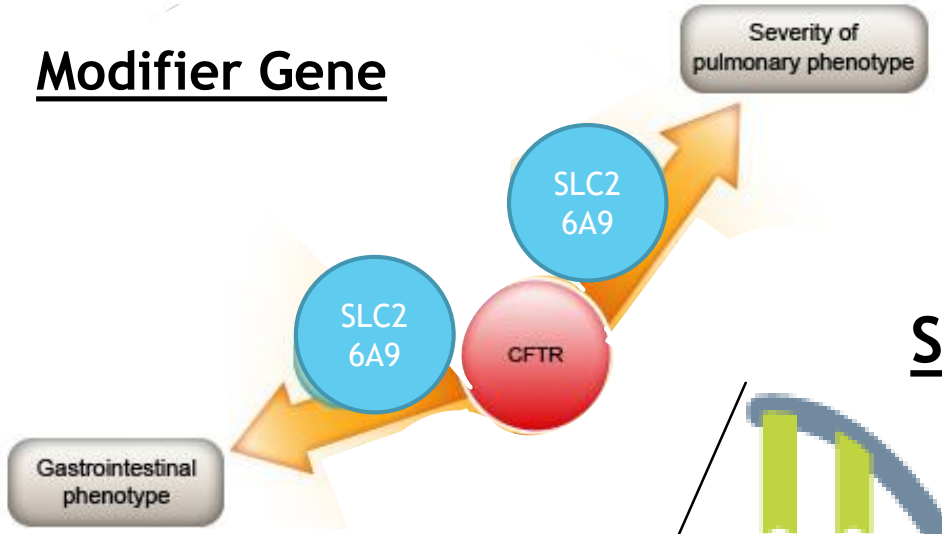
- All physical and observable traits.
- E.g. Height, hair color, white blood cell count, and diseases you may have (diabetes, cystic fibrosis, etc.).



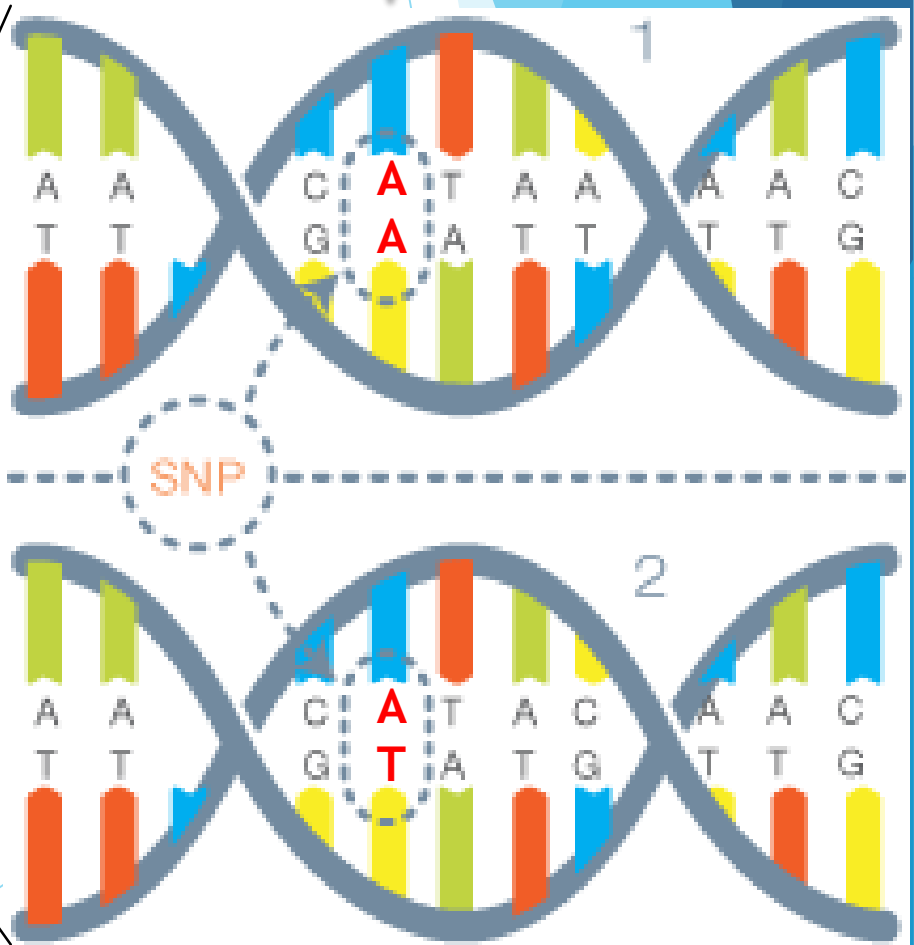

**Typically**: genotype (G) + environment (E) → phenotype (P)

# Modifier Genes

- **Cystic Fibrosis (CF) Genetic modifiers** are SNPs that affect the severity of the disease.
- Modifier gene: SLC26A9
- Affects lung function for people with CF.
- SNP: RS4077468
- SNP Variation:
  - AA
  - AT
  - TT







# Research Question

## In the general public what is the impact of variation in the three modifier genes on a person's phenotypes.

- We will answer this question through a Phenome-wide Association Study (PheWAS).

The 3 modifier genes of interest:

1) SLC26A9 (Chromosome 1 - SNP rs4077468 Substitute: rs4077469; r = 1)
2) SLC6A14 (Chromosome X – SNP rs3788766 Substitute: rs5905176; r = 0.770)
3) SLC9A3 (Chromosome 5 – SNP rs57221529 Substitute: rs17497684; r = 0.821)

# UK Biobank

## Cohort

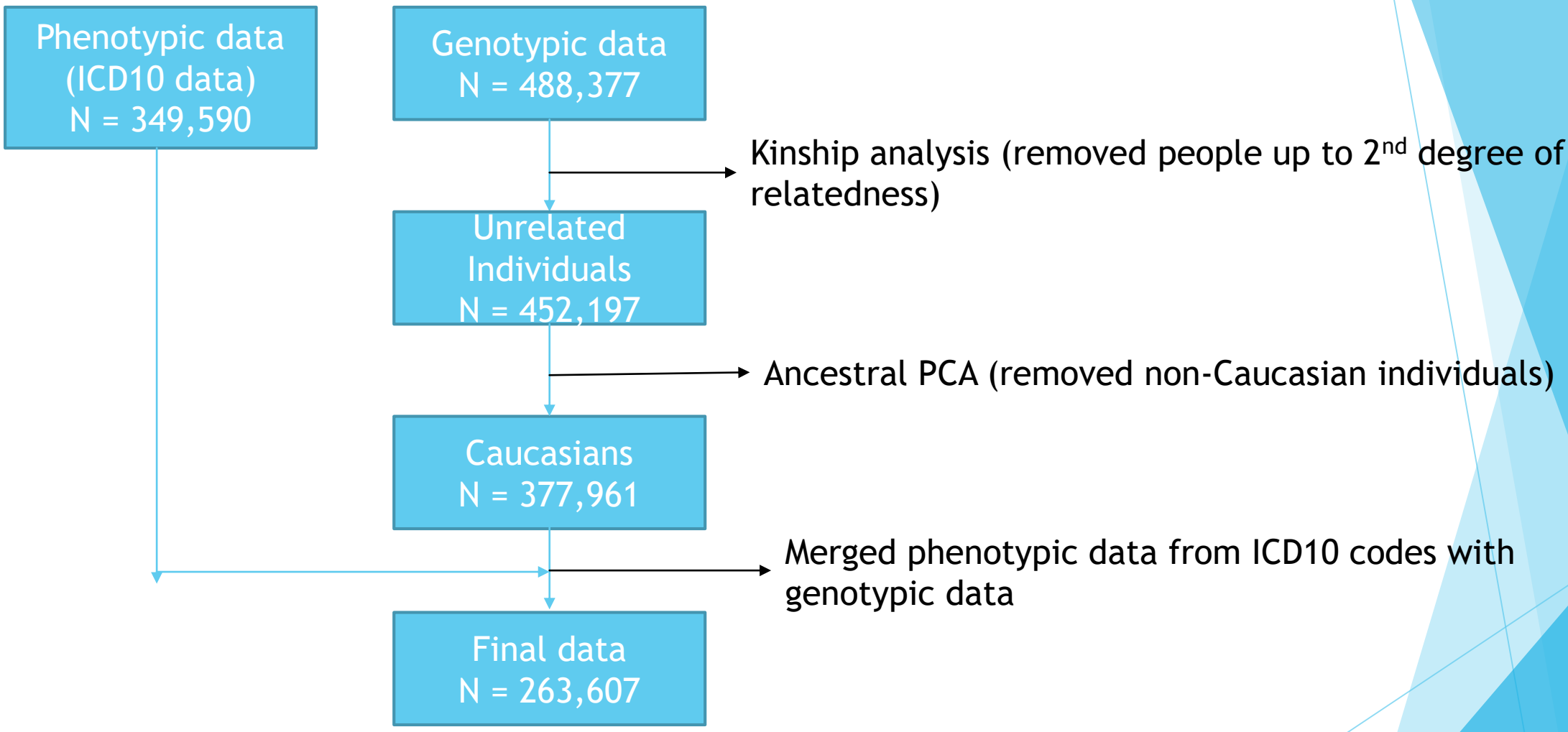
- Approximately 500,000 people aged between 40-69 years in 2006-2010 from across the country (Mainly England, Scotland and Wales).
- Initial enrollment took place over four years from 2006.
- Baseline data collected via questionnaires, physical measures, sample assays, accelerometry and multimodal imaging among others.
- Individual's national health records have also been linked with their baseline and genotypic data.
- 1511 phenotypes obtained from ICD10 codes (30GB).
- Genotype data (100GB). Micro arrays: between 500,000 to 1 million SNPs per person.

Significant time spent data cleaning (formatting, merging, etc.).

# 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).
- Purpose:
  - It was created for proper hierarchical grouping of phenotypes.
  - High throughput.
  - For ensuring exclusion of mutually exclusive phenotypes. For example, excluding subjects with any other type of diabetes from the control group when studying type 2 diabetes (Wei et al. 2017).
- In general, phecodes have been successfully used in a number of PheWAS to replicate hundreds of known genetic associations and discovered new ones.

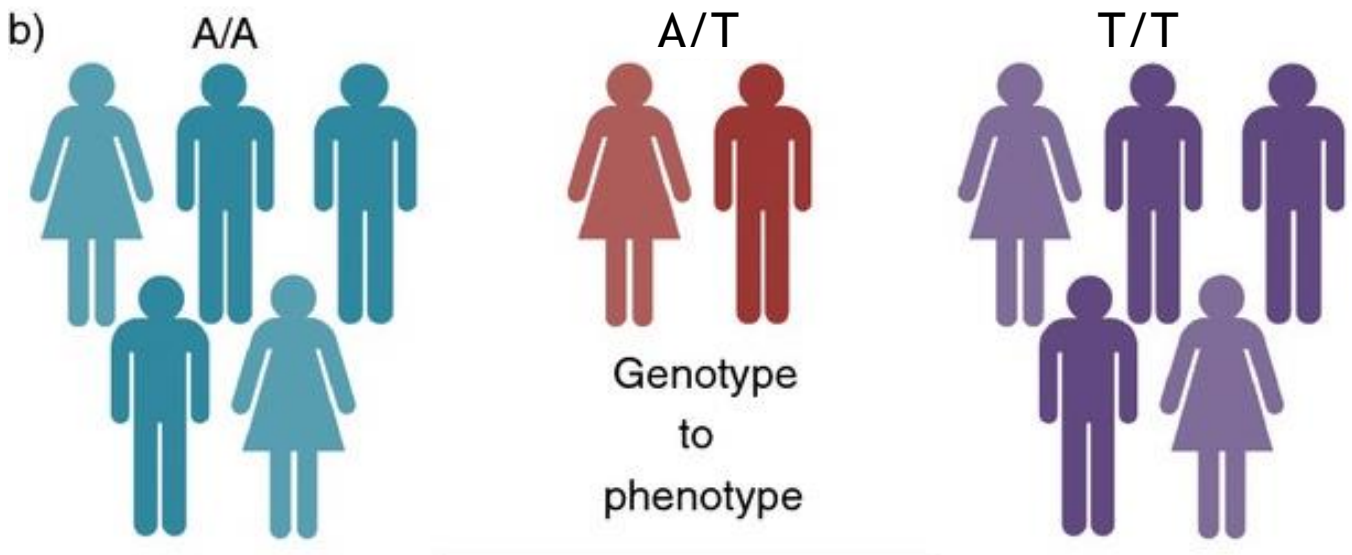
# Data Flow Chart for SLC26A9 - SNP rs4077468)




Final data set for ***SCL9A3***, ***SCL26A9*** and ***SLC6A14*** had individuals **262,923**, **261,655** and **117,398**, respectively.

# What is a PheWAS?

- PheWAS: Phenome-Wide Association Study
- Tests the association between genetic variants of interest with every phenotype measured.
- They are cross-sectional studies.
- The outcome variable is the person's phenotype. The predictor variable is the allele variation of the SNP.




Association: genotype (G) → all phenotypes (P's)

# Phenome-Wide Association Study (PheWAS)

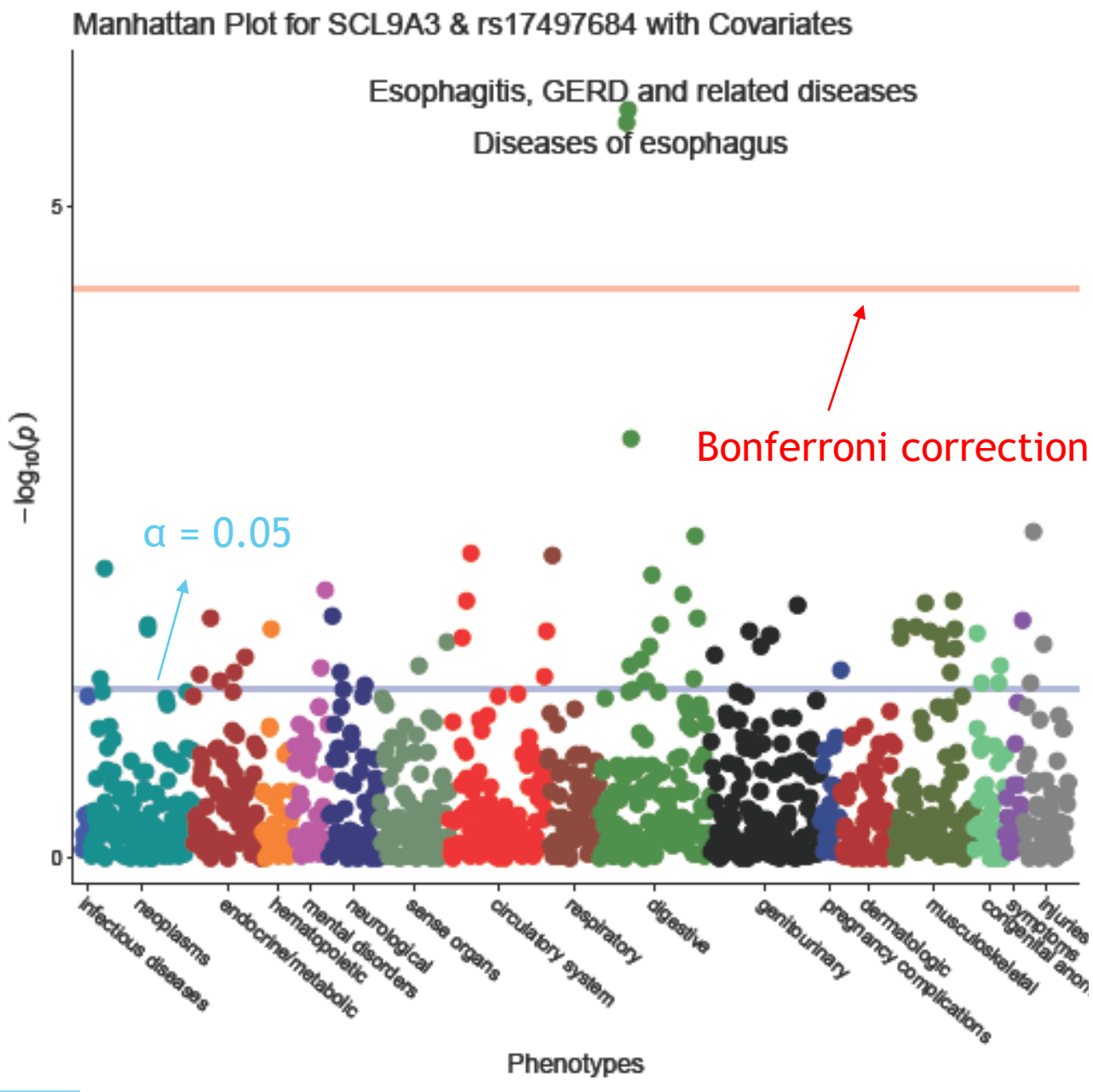
- **Statistical Method: Additive Model for performing PheWAS:**
  - Logit(Phenotype_i) = SLC26A9 + covariates  i=1, …,1511

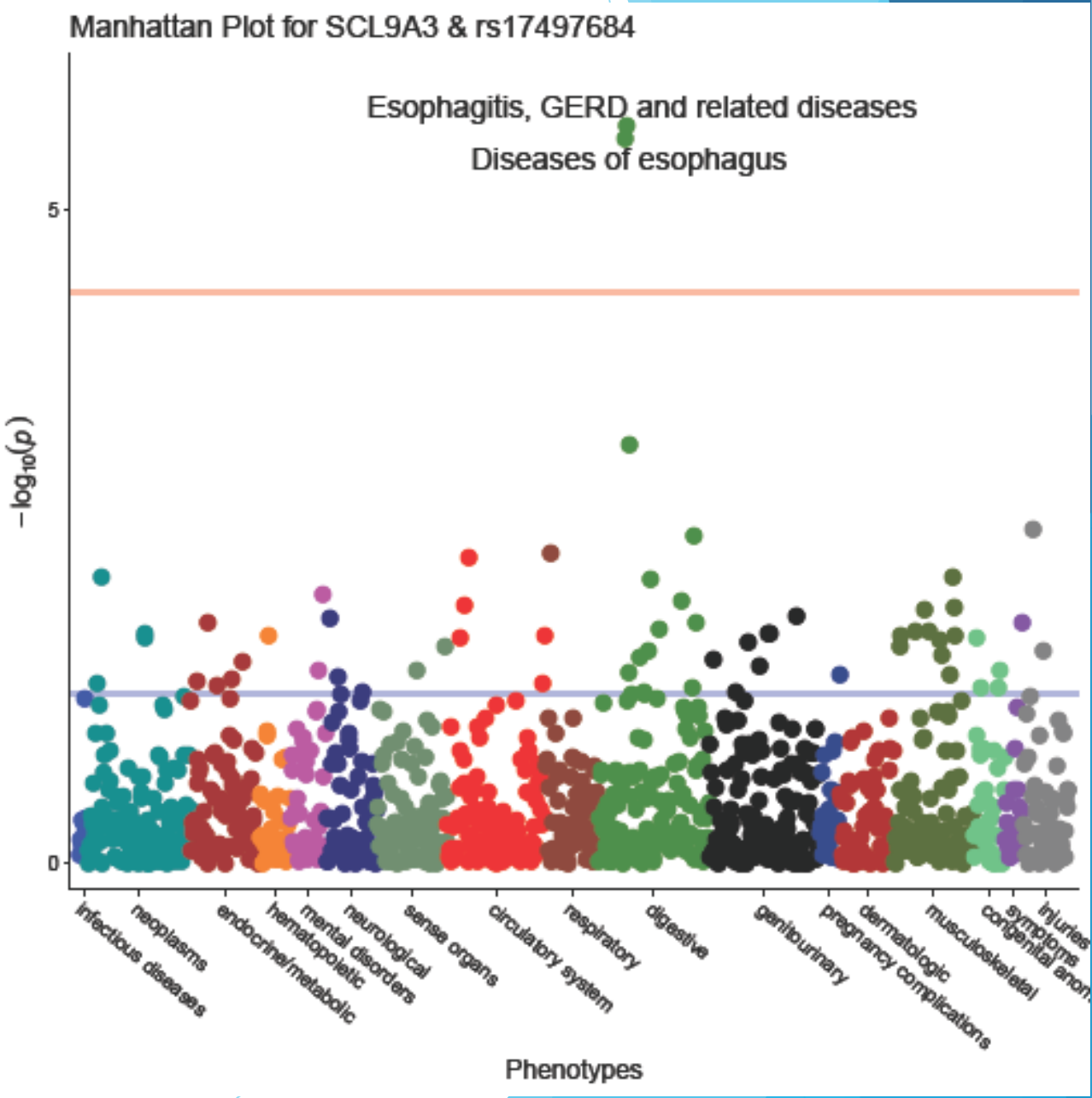
$$\text{Phenotype\_i} = \begin{cases} 0 & \text{if do not have phenotype i} \\ 1 & \text{if have phenotype i} \end{cases}$$

$$\text{SLC26A9} = \begin{cases} 0 & \text{if RS4077468\_AA} \\ 1 & \text{if RS4077468\_AT} \\ 2 & \text{if RS4077468\_TT} \end{cases}$$

  - Perform adjusted and unadjusted logistic regression.
  - Adjusted for covariates: Age, age-squared and sex.
- **Software:**
  - R ("PheWAS/PheWAS" package from github) and PLINK.
  - Linux environment for high performance computing.

# SLC9A3 (Chromosome 5 – SNP rs57221529 Substitute: rs17497684; r = 0.821)



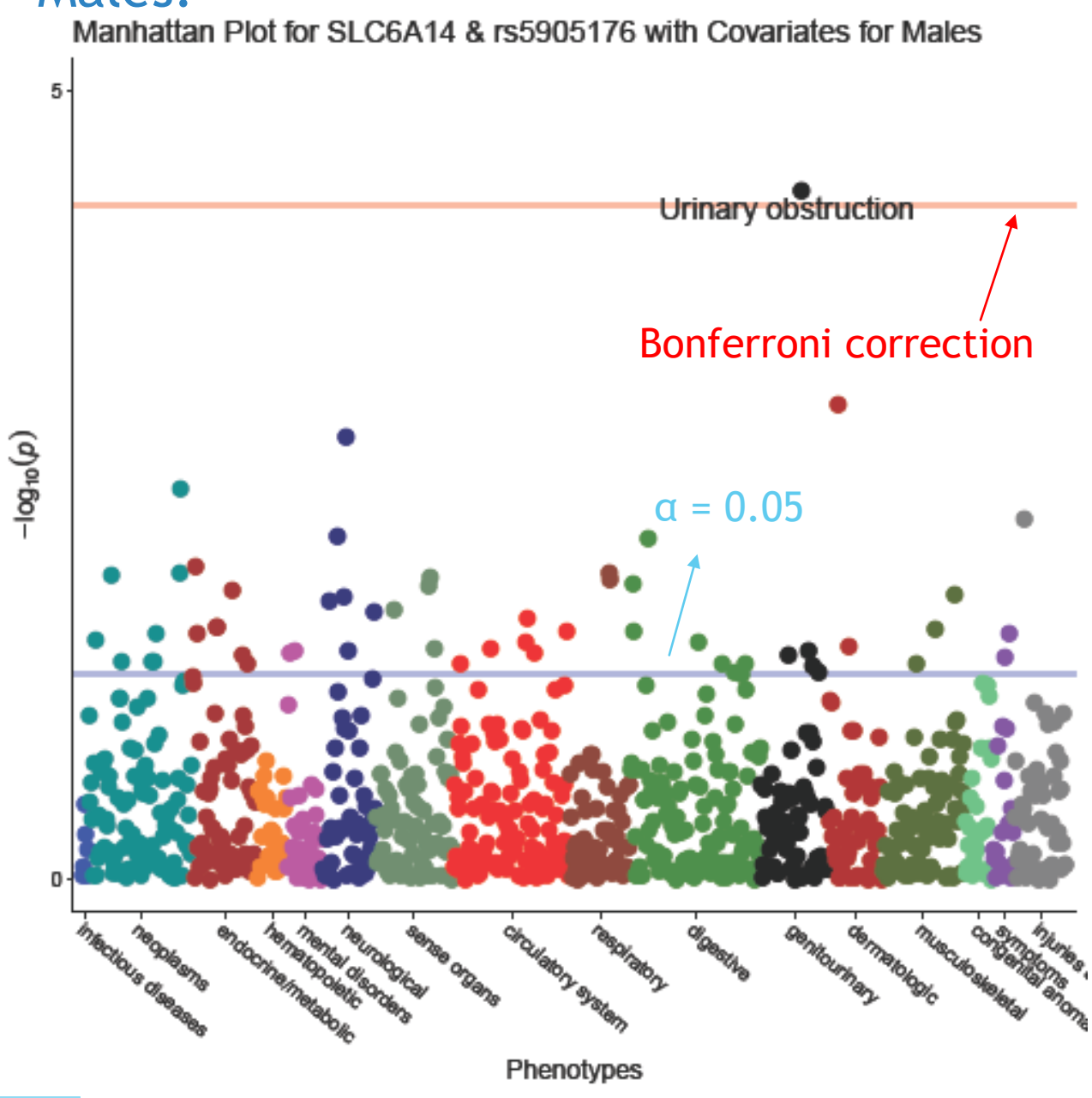
# SLC9A3 (Chromosome 5 – SNP rs57221529 Substitute: rs17497684; r = 0.821)

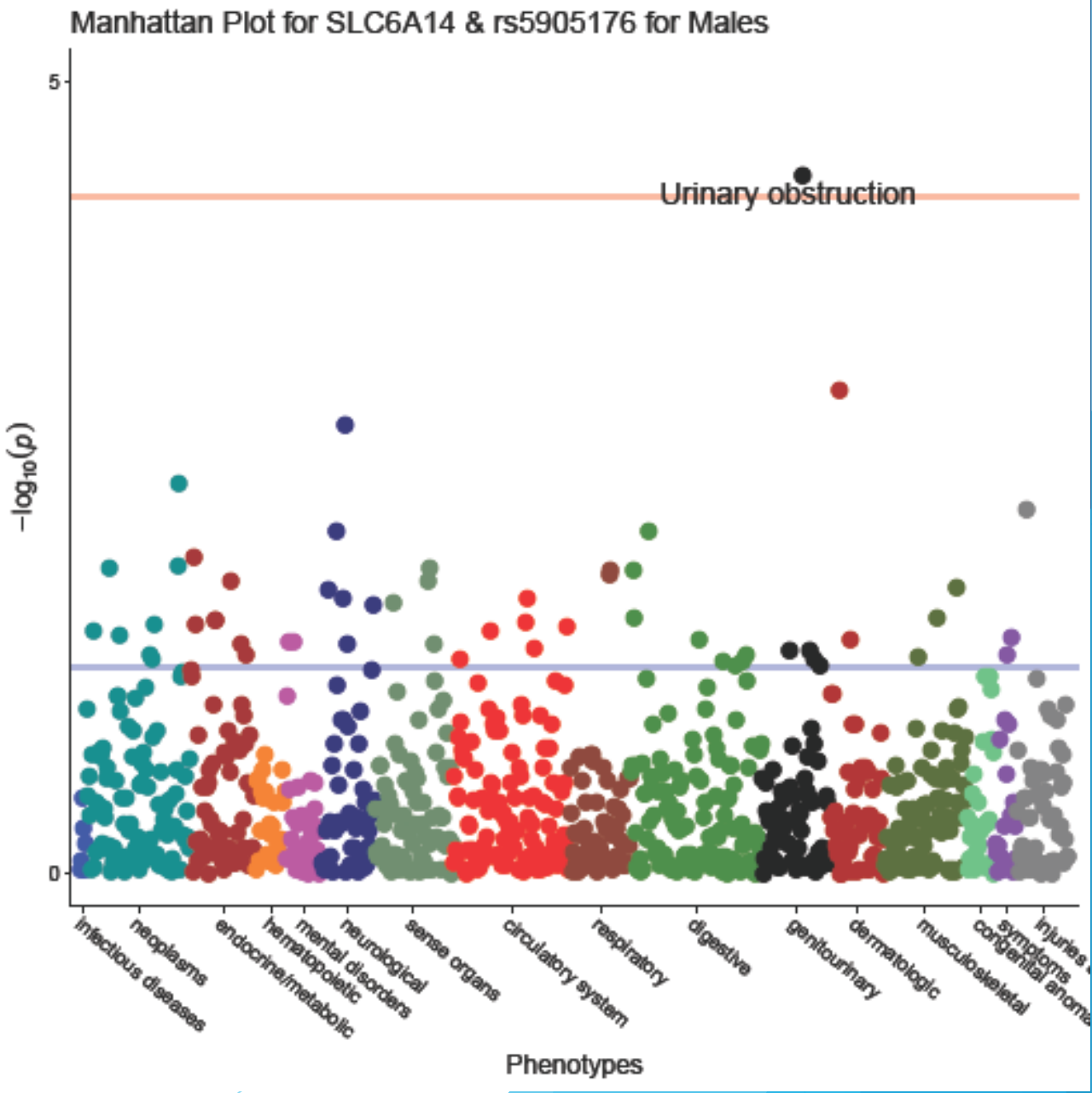
**Esophagitis, GERD and related diseases**

OR = 1.064
P-value = 1.79E-06
Cases = 19,687
Controls = 243,236

| | | C Allele Count | | |
|---|---|---|---|---|
| | level | **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) |
| **age2 (mean (SD))** | | 3407.06 (871.43) | 3406.22 (872.61) | 3389.88 (869.40) |
| **SEX (%)** | 0 | 94320 ( 55.5) | 45716 ( 55.2) | 5598 ( 55.5) |
| | 1 | 75730 ( 44.5) | 37068 ( 44.8) | 4491 ( 44.5) |

# SLC6A14 (Chromosome X – SNP rs3788766 Substitute: rs5905176; r = 0.770) for Males.



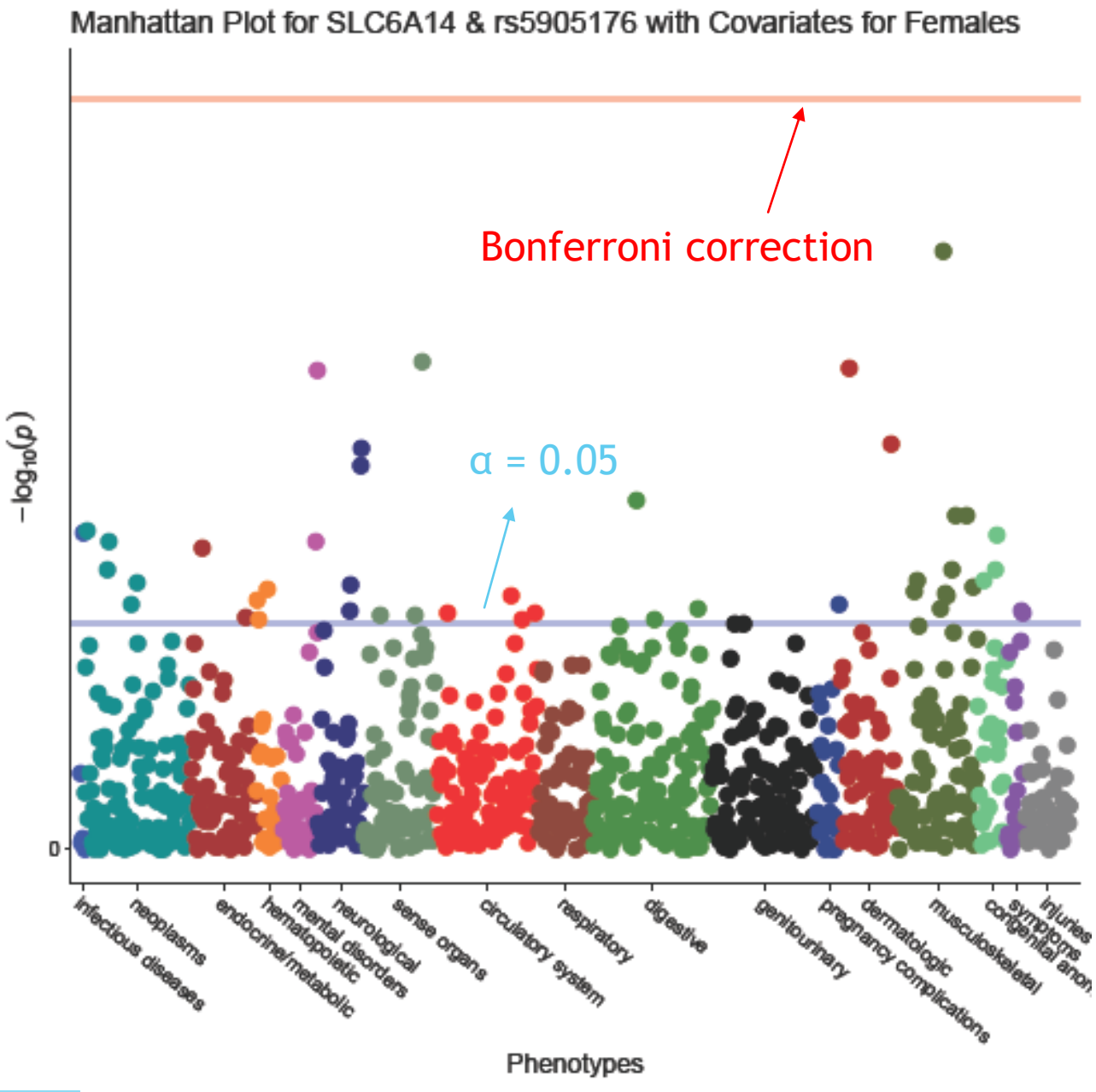
# SLC6A14 (Chromosome X – SNP rs3788766 Substitute: rs5905176; r = 0.770) for Males.

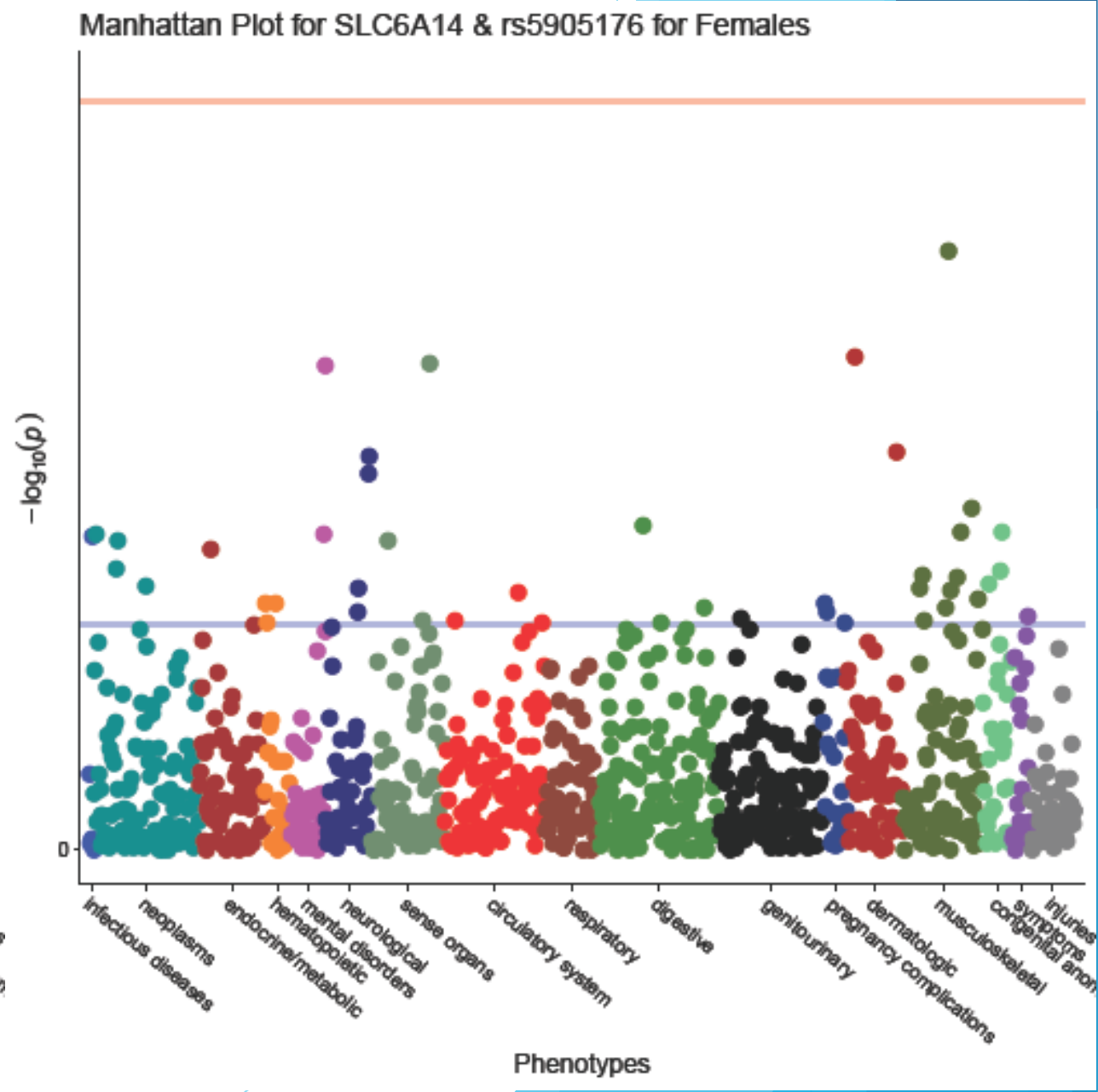
**Esophagitis, GERD and related diseases**

OR = 1.68
P-value = 4.24E-05
Cases = 64
Controls = 117,334

| Phecode | | | Urinary obstruction | | |
|---|---|---|---|---|---|
| | | level | Controls | Cases | p |
| n | | | 117334 | 64 | |
| rs5905176_G (%) | AA | 0 | 79076 ( 67.4) | 27 ( 42.2) | <0.001 |
| | GG | 2 | 38258 ( 32.6) | 37 ( 57.8) | |
| age (mean (SD)) | | | 58.44 (7.73) | 62.19 (5.96) | <0.001 |
| age2 (mean (SD)) | | | 3475.18 (868.88) | 3902.25 (706.87) | <0.001 |

# SLC6A14 (Chromosome X – SNP rs3788766 Substitute: rs5905176; r = 0.770) for Females.



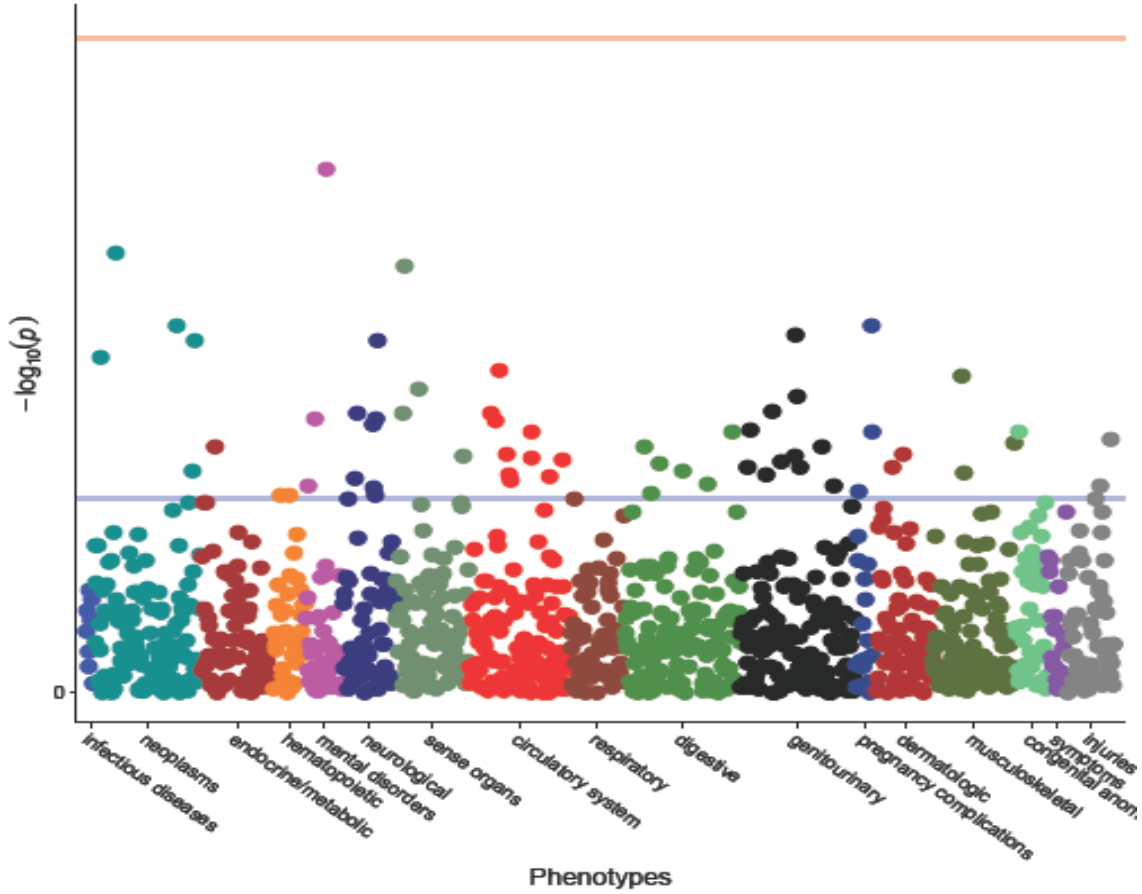



# SLC6A14 (Chromosome X – SNP rs3788766 Substitute: rs5905176; r = 0.770) for Females.

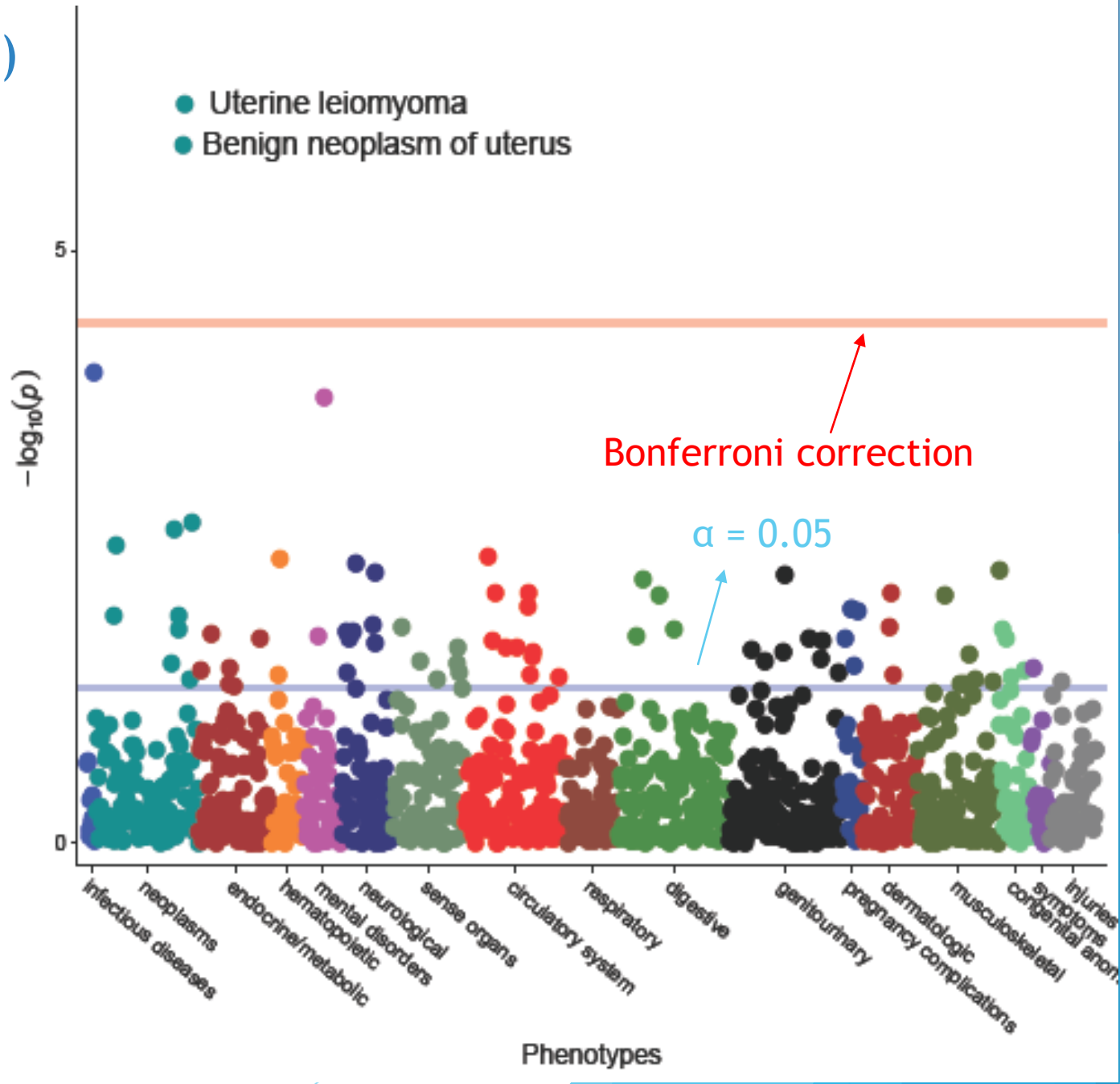
| | | G Allele Count | | |
|---|---|---|---|---|
| | level | **AA** | **AG** | **GG** |
| | G Allele Count | **0** | **1** | **2** |
| **n** | | 66154 | 63693 | 15658 |
| **Urinary obstruction** | Controls | 66142 (100.0) | 63668 (100.0) | 15646 ( 99.9) |
| phecode: 733.6 (%) | Cases | 12 ( 0.0) | 25 ( 0.0) | 12 ( 0.1) |
| **age (mean (SD))** | | 57.43 (7.77) | 57.31 (7.83) | 57.21 (7.80) |
| **age2 (mean (SD))** | | 3358.36 (867.71) | 3345.77 (872.86) | 3333.50 (868.15) |

# SLC26A9 (Chromosome 1 - SNP rs4077468 Substitute: rs4077469; r = 1)

Manhattan Plot for rs4077469 with Covariates




Manhattan Plot for rs4077469 without Covariates



# Some Challenges and Limitations.

- Multiple testing.
  - 1511 phenotypes and three SNP's
- Missing data (phenotype & genotype). All and any missing data was deleted. Most likely data not MCAR; people with ICD9 codes were deleted. Mostly people in the beginning of the recruitment period.
- No exclusion scheme when defining controls based on ICD codes.
- Converting ICD10 codes to phenotypic codes (phecode system built was on ICD9 codes) Validated by Wu et al (2018) on the UKBB data.
- Several phecodes and phynotypes are related, hence Bonferroni correction is conservative.
- Relatedness, kinship analysis.
- Computing the ancestral PCA.
- Implementing all of this in a HPF.

# Future Work

- Instead of using additive model to perform the PheWAS, use a genotypic model (treat allele count as categorical variable).
- Included interaction term between allele count and sex.
- Use curated phenotypic data
  - Lung function: $FEV_1/FVC$ ratio
- Validating the UKBB data, our data cleaning and conversion of ICD-10 to phecodes process and our PheWAS analysis, by replicating previously published PheWAS studies.

# Conclusion

- Results suggest that there maybe an association between gene SLC9A3 near SNP rs57221529, and having Esophagitis, GERD and related diseases.
- Every additional C allele increases the odds by about 6.4% of having the related diseases in an individual.
- Results generalizable to people with Caucasian ancestry.
- With this PheWAS we may have found a phenotype associated with gene SCL9A3 in the non-CF population
- However, further research work is required.

# Questions?

# Additional Slides

# Phenome-Wide Association Study (PheWAS)

- **Statistical Method: "*Genotypic*" Model for performing PheWAS:**
  - Logit(Phenotype_i) = intercept + I(RS4077468_AT) + I(RS4077468_TT) + covariates

$$i=1, \ldots,1511 , \quad SLC26A9 = \begin{cases} 0 & \text{if RS4077468\_AA (reference)} \\ 1 & \text{if RS4077468\_AT} \\ 2 & \text{if RS4077468\_TT} \end{cases}$$

  - Perform adjusted and unadjusted logistic regression.
  - Adjusted for covariates: Age, age-squared and sex.

- **Software:**
  - R ("PheWAS/PheWAS" package from github) and PLINK.
  - Linux environment for high performance computing.

# PheWAS Manhattan Plot

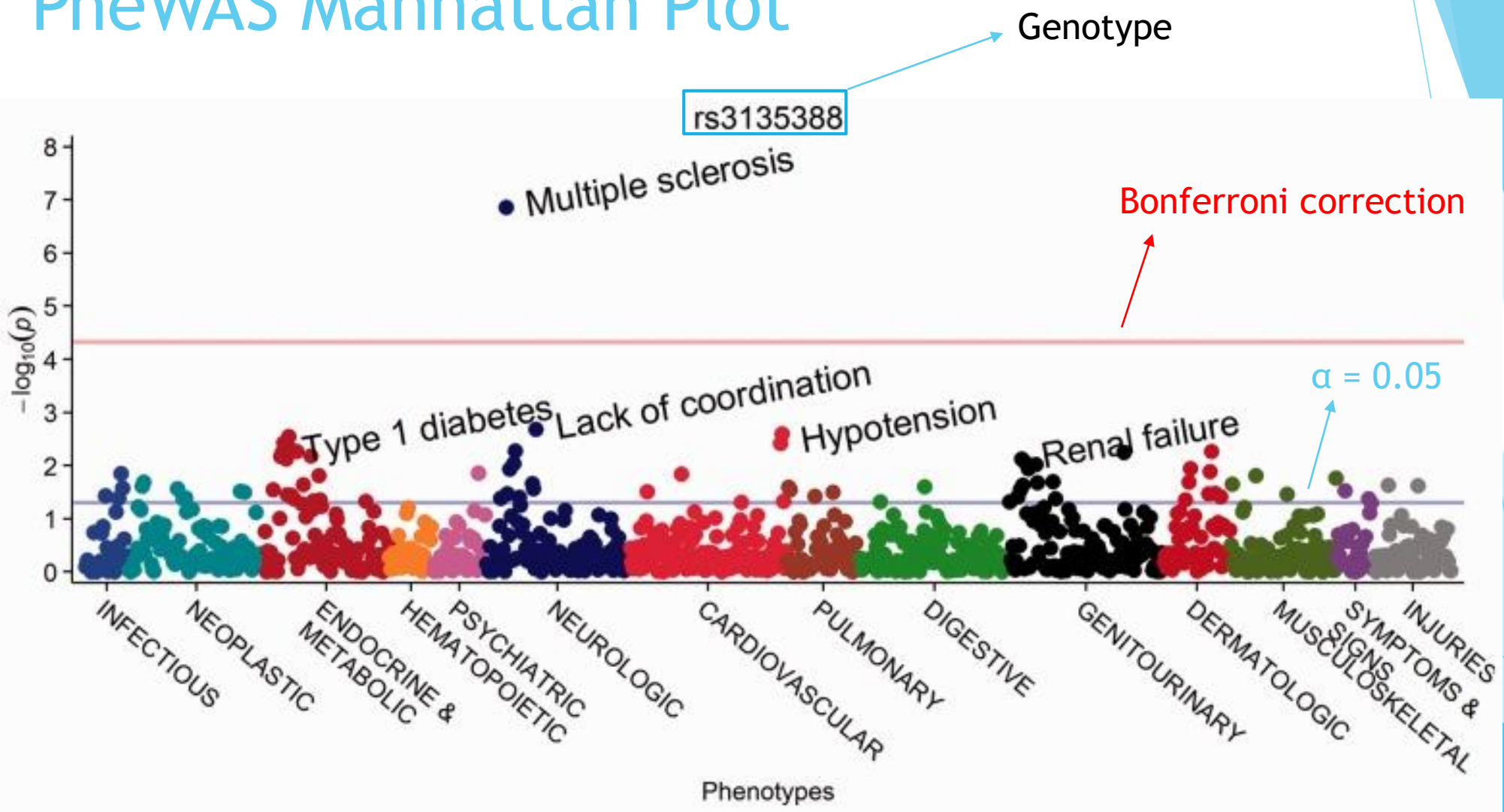



Figure 5. PheWAS Manhattan plot for rs3135388, with phenotypes ordered by PheWAS code. (Carrol et al. 2014)

# Detailed Results for Multivariable Analysis.

# SLC9A3 (Chromosome 5 – SNP rs57221529 Substitute: rs17497684; r = 0.821)

| phecode | description | group | snp | beta | SE | OR | p | type | n_total | n_cases | n_controls | HWE_p | allele_freq | n_no_snp | bonferroni |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 530.1 | Esophagitis, GERD and related diseases | digestive | rs17497684_C | 0.062196 | 0.013023 | 1.064171 | 1.79E-06 | logistic | 262923 | 19687 | 243236 | 0.087651 | 0.195803 | 684 | TRUE |
| 530 | Diseases of esophagus | digestive | rs17497684_C | 0.059461 | 0.012569 | 1.061265 | 2.23E-06 | logistic | 262923 | 21330 | 241593 | 0.087651 | 0.195803 | 684 | TRUE |

| | | C Allele Count | | |
|---|---|---|---|---|
| | level | 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) |
| **age2 (mean (SD))** | | 3407.06 (871.43) | 3406.22 (872.61) | 3389.88 (869.40) |
| **SEX (%)** | 0 | 94320 ( 55.5) | 45716 ( 55.2) | 5598 ( 55.5) |
| | 1 | 75730 ( 44.5) | 37068 ( 44.8) | 4491 ( 44.5) |

# SLC6A14 (Chromosome X – SNP rs3788766 Substitute: rs5905176; r = 0.770) for Males.

| phecode | description | group | snp | beta | SE | OR | p | type | n_total | n_cases | n_controls | HWE_p | allele_freq | n_no_snp | bonferroni |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 599.1 | Urinary obstruction | genitourinary | rs5905176_G | 0.518311 | 0.126602 | 1.679189 | 4.24E-05 | logistic | 117398 | 64 | 117334 | 1 | 0.326198 | 194 | TRUE |

| Phecode | | Urinary obstruction | | |
|---|---|---|---|---|
| | level | Controls | Cases | p |
| **n** | | 117334 | 64 | |
| **rs5905176_G (%)** | 0 | 79076 ( 67.4) | 27 ( 42.2) | <0.001 |
| | 2 | 38258 ( 32.6) | 37 ( 57.8) | |
| **Urinary obstruction** | | 117334 (100.0) | 0 ( 0.0) | <0.001 |
| phecode: 599.1 (%) | TRUE | 0 ( 0.0) | 64 (100.0) | |
| **Disorder of skin and subcutaneous tissue NO** | FALSE | 115583 ( 98.5) | 64 (100.0) | 0.639 |
| phecode: 689 (%) | TRUE | 1751 ( 1.5) | 0 ( 0.0) | |
| **age (mean (SD))** | | 58.44 (7.73) | 62.19 (5.96) | <0.001 |
| **age2 (mean (SD))** | | 3475.18 (868.88) | 3902.25 (706.87) | <0.001 |

# Diseases

- **Gastroesophageal reflux disease**, or **GERD**, is a digestive disorder that affects the lower esophageal sphincter (LES), the ring of muscle between the esophagus and stomach. Many people, including pregnant women, suffer from heartburn or acid indigestion caused by **GERD**.
- **Esophagitis** (uh-sof-uh-JIE-tis) is inflammation that may damage tissues of the esophagus, the muscular tube that delivers food from your mouth to your stomach.**Esophagitis** can cause painful, difficult swallowing and chest pain.

# UK Biobank (Variables)

- Baseline characteristics
- Blood count
- Blood count processing
- Blood pressure
- Blood sample collection
- Body size measures
- Bone-densitometry of heel
- Bread/pasta/rice yesterday
- Breathing
- Cancer register
- Cancer screening
- Cannabis use
- Cereal yesterday
- Chest pain
- Claudication and peripheral artery disease
- Consent
- Consent timings and usage
- Death register
- Depression
- Diet
- Diet by 24-hour recall

**Baseline characteristics**

| Field ID | Field title |
|---|---|
| 21022 | Age at recruitment |
| 52 | Month of birth |
| 31 | Sex |
| 189 | Townsend deprivation ind |
| 34 | Year of birth |

- Blood count
- Blood count processing
- Blood pressure
- Blood sample collection
- Body size measures
- Bone-densitometry of heel
- Bread/pasta/rice yesterday
- Breathing
- Cancer register

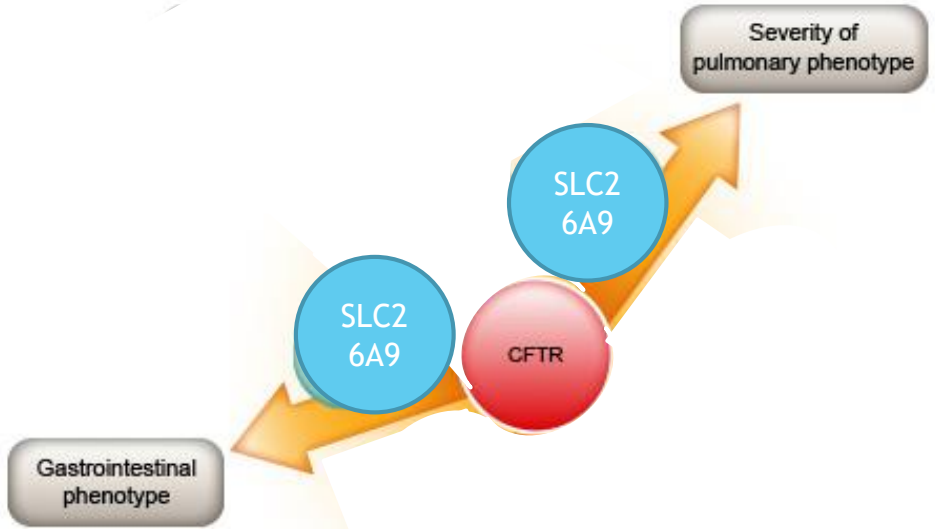

Severity of pulmonary phenotype
SLC2 6A9
SLC2 6A9
CFTR
Gastrointestinal phenotype

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 |