Phenome-Wide Association Study to Determine the Effects of Cystic Fibrosis Modifier Genes in the UK Biobank Population.

Faizan Khalid Mohsin[1], M.Sc; Lisa Strug[1,2], Ph.D; Naim Panjwani[2], M.Sc; and Zeynep Baskurt[2], Ph.D

[1] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON; [2] Hospital for Sick Children, Toronto, ON

ABSTRACT

Question: What is the impact of having the gene variants that increase severity of Cystic Fibrosis disease in people who do not have Cystic Fibrosis?

Findings: In UKBiobank population

1. People with **allele C** at SNP **rs17497684** of the gene **SLC9A3** have **6.4%** higher probability of developing **Esophagitis, GERD and related disease.**

2. **Males** with **allele G** at SNP **rs5905176** of the gene **SLC6A14** were associated with having **68%** higher probability of developing **Urinary Obstruction.**

Method: Sample size after QC steps ~**264,000** unrelated individuals.
Used **UKBiobank** data (~500,000 individuals).
PheWAS method used to find associations.

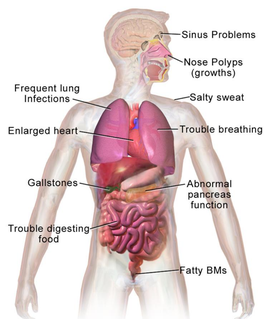
1. INTRODUCTION

1.1. CYSTIC FIBROSIS

Cystic fibrosis (CF) is the most common fatal genetic disease affecting Canadian children and young adults.

At present, there is no cure.

Fig 1. Issues associated with CF



1.2. MODIFIER GENES

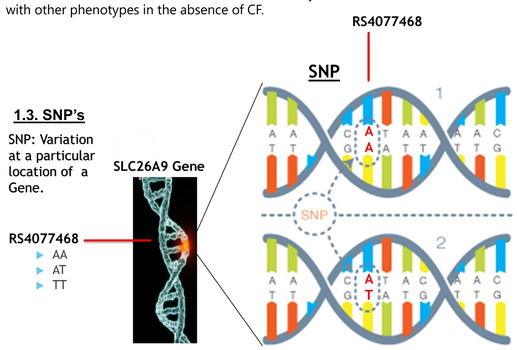
The 3 modifier genes of interest and the SNP's (location of genetic variation) of interest*:

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

*These modifier genes were identified through Sun et al (2012) to be associated with meconium ileum in CF. Here we want to know if they are associated with other phenotypes in the absence of CF.

1.3. SNP's

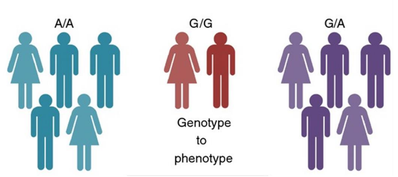
SNP: Variation at a particular location of a Gene.



2. METHODOLOGY

2.1 STATISTICAL METHOD & SOFTWARE

What is a PheWAS?
Phenome-Wide Association Study



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

(Variation in the gene) (Different Physical Traits such as colour blindness or genetic diseases)

- The data had 1511 phenotypes.
- ICD10 data codes from participants' electronic health recode data was mapped to phenotypes using the PheCode system codes. The phecode mapping was done through collaboration.

Statistical Method: Additive Model for performing PheWAS:

- Logit(P(phenotype_i = 1)) = SLC26A9_i + covariates_i i=1, …, 264,000 individual

$$\text{Phenotype\_i} = \begin{cases} 0 & \text{if do not have phenotype i} \\ 1 & \text{if have phenotype i} \end{cases} \qquad \text{SLC26A9} = \begin{cases} 0 & \text{if RS4077468\_AA} \\ 1 & \text{if RS4077468\_AT} \\ 2 & \text{if RS4077468\_TT} \end{cases}$$

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

Software:

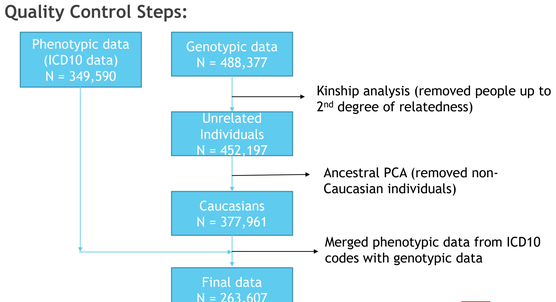
- R ("PheWAS" package from github) and PLINK.
- Linux environment for high performance computing.

2.2 DATA & QC STEPS

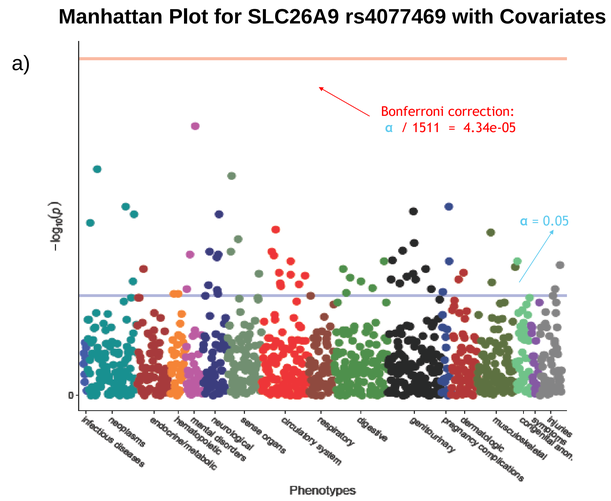
Data:

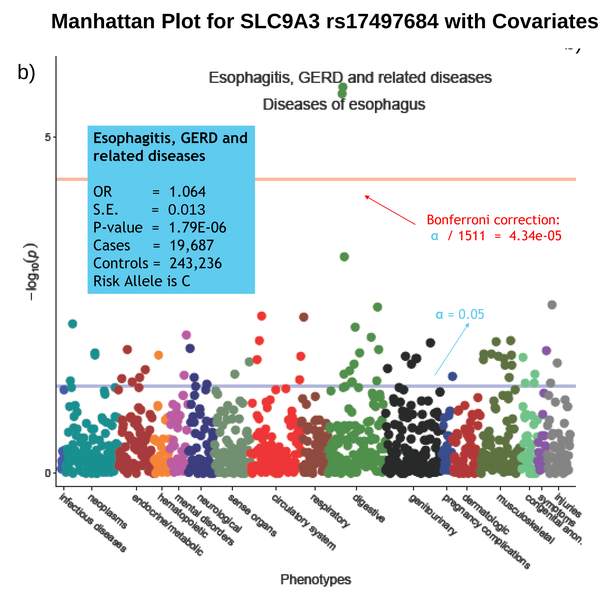
1) Approximately 500,000 people aged between 40-69 years in 2006-2010 from across the country (Mainly England, Scotland and Wales).
2) All participants volunteered to provide their genetic data: Genotype data (100GB): between 500,000 to 1 million SNPs per person.
3) Individual's national health records have also been linked with their baseline and genotypic data.
4) The SNP's of interest were not genotyped for a high number of individuals (hence were missing). Therefore, instead of imputing them we chose to substitute them with SNP's that were genotyped and had high pearson correlation (r > 0.7) as identified in section 1.2.
5) Final data set for ***SLC9A3***, ***SLC26A9*** and ***SLC6A14*** had individuals n= **262,923**, **261,655** & **117,398**, respectively.

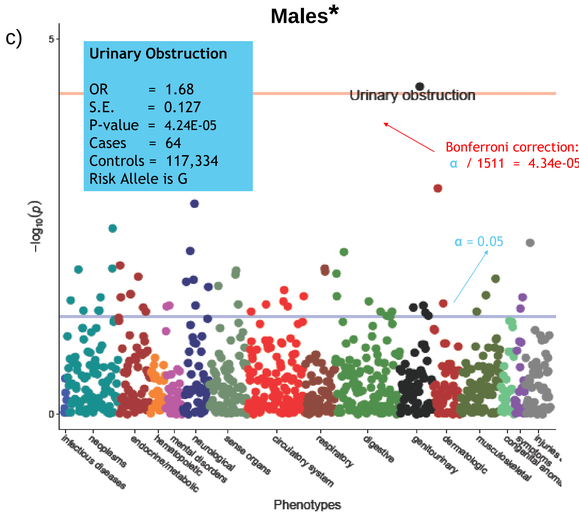
Quality Control Steps:

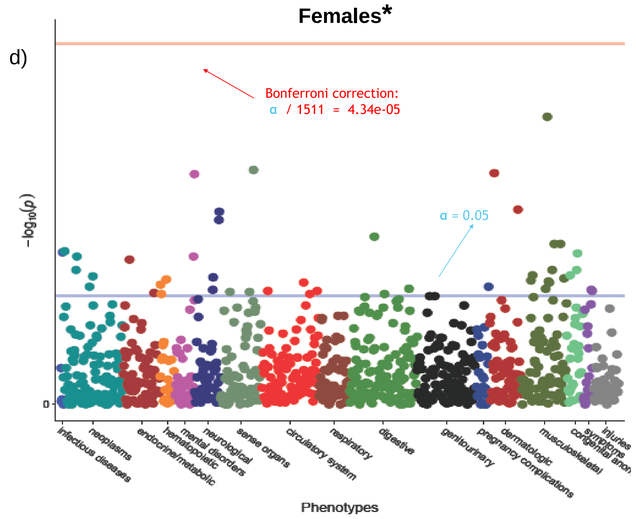


3. RESULTS









4. CONCLUSION

* The analysis for the SLC6A14 rs5905176 SNP was carried out separately for males and females because it is found on the X chromosome.

Results:

- Results suggest there to be an association between SNP rs17497684 of gene SLC9A3 with 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 in non-CF populations.

Future Work:

- Use curated phenotypes. E.g., Lung function: FEV1/FVC ratio
- Include interaction term between allele count and sex.
- Instead of using additive model use a genotypic model (treat allele count as categorical variable).
- Adjust our analysis for the case control imbalance.

5. ACKNOWLEDGEMENT: **This research has been conducted using the UK Biobank Resource under Application Number 40946.**