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 children and young adults of caucasion decent [@ source]. 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 rs4077468 & sub rs4077469; r = 1 - Chromosome 1, SLC6A14 rs3788766 & sub rs5905176; r = 0.770 - Chromosome X and lastly, SCL9A3 rs57221529 & sub rs17497684; r = 0.821 - Chromosome 5. We used the UK biobank data registry, which has over 500,000 participants, to perform the PheWAS study using the ICD10 codes in the registry. The 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 (e.g. minor allele T for rs4077469 C/T, being either 0, 1 or 2) as the predictor. The allelle count was modeled as an additive model. Adjusted and unadjusted analysis were both performed. The model was adjusted for covariates age and age-squared, gender. Since, there were 1511 phenotypes [@ reference] 1511 logistic regression were performed and a bonferroni correction was applied to determine statistically significant associations.

Findings: For the unadjusted analysis, gene SCL26A9 and its SNP rs4077469, the traits of having Uterine leiomyoma and Benign neoplasm of uterus, were found to be statistically significant (OR = 0.9222197, p-value = 5.614299e-07 and OR = 0.9254563, p-value = 1.247990e-06, respectively) after bonferroni correction. Moreover, these diseases are gender specific and only occur in females.

Interpretation: In the unadjusted analysis, for gene SCL26A9 and its SNP rs4077469, C is the risk allele, hence, its OR's are 1.0843403 and 1.080548, therefore, every additional C allele increases the risk of having uterine leiomyoma and benign neoplasm of uterus by 8.43% and 8.01%, respectively. (After adjusting for gender, the increased risk per allele was XX% and XXX% respectively (in females)). In the study by 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, $P = 3.6 \times 10^{-8}$ in their unadjusted analysis (Blackman et al. 2013).

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

2 Introduction

Project Description: We have identified common variation in several genes that contribute to Cystic Fibrosis (CF) disease severity. The majority of these genes are transporters, and we are unaware if variation in the genes impact phenotypes in individuals who do not have CF, that is, individuals without two mutations in the CF causal gene, CFTR. Using 500,000 individuals from the UKBiobank who have been genotyped genome-wide and have detailed, comprehensive phenotypic data, the student will carry out a Phenome-wide association study (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. 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. The student will be expected to have good work knowledge of regression methodology and ability to work with large data sets in a Unix environment. Site Location: SickKids Research Institute, 686 Bay Street Site Description: The successful candidate will be working amongst other statisticians, bioinformaticians, molecular geneticists, clinical researchers and programmers involved in the Canadian Cystic Fibrosis Gene Modifier Study. The research program is based out of The Hospital for Sick Children Research

Institute, with both wet and dry laboratory space and scientists working collaboratively to improve the lives of individuals living with Cystic Fibrosis.

clinical equipoise

3 Material and Methods

R packages: "PheWAS/PheWAS" package from the Github repository and the "PHESANT" package also from Github (Millard et al. 2017).

EHR-package

4 References

5 Methods

5.1 SNP's of interest.

SNPS we are interested in:

SNP's and the genes: SCL26A9 - Original SNP: rs4077468 - Chromosome 1. Substitute: rs4077469; r = 1 SLC6A14 - Original SNP: rs3788766 - Chromosome X. Substitute: rs5905176; r = 0.770 SCL9A3 - Original SNP: rs57221529 - Chromosome 5. Substitute: rs17497684; r = 0.821

Also need to look at SNP rs7512462 in SLC26A9. Treatment response to ivacaftor, which aims to improve CFTR-channel opening probability in patients with gating mutations, shows substantial variability in response, 28% of which can be explained by rs7512462 in SLC26A9 (P=0.0006) (L. J. Strug et al. 2016).

5.2 Statisal Methods

5.2.1 Hardy-Weinburg Equilibrium

Discussion:

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: $p\mathbf{2+\ 2pq\ +\ q}2 = 1$. 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

We are looking,

?phewas

5.2.2 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).

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.

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

5.3 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.

5.4 UK Biobank data

The UK Biobank is a large-scale, population-based, prospective cohort that enrolled over 500,000 participants aged 40–69 years. The recruited participants provided a wide range of self-reported baseline information. Blood samples were collected for biochemical tests and genotyping. Their national health records have been linked with the baseline and genotypic data 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). (X. Li et al. 2018)

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

We focused on phenotypes in relation to diagnostic disease outcomes. We analysed two phenotypic data sets (inpatient hospital episode records and cancer registry data) in the UK Biobank using the phecode schema (see online supplementary text for phenotyping and mapping process) (Denny et al. 2013). 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 both ICD-10 and ICD-9 codes to define the case and control groups. Since cancer registry data overlapped with the cancer diagnosis in inpatient hospital records, we pooled the cancer registry data into the hospital episode data as a complement to the cancer diagnosis. (X. Li et al. 2018)

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

consort flow chart.

5.6 Sample Size

5.7 Study Design

5.8 Phenotype data

5.9 Genotype data

5.10 Primary Endpoints

5.11 Secondary Endpoints

In genetics p-values are written in scientific notation.

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

The researcher does not know the thing of

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 Cite software.

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. Patients you started with and the patients you analysed.

6 Results

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.

7 Discussion

State upfront all the limits of the study.

8 Reference

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.

Denny, Joshua C, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, et al. 2013. "Systematic Comparison of Phenome-Wide Association Study of Electronic Medical Record Data and Genome-Wide Association Study Data." *Nature Biotechnology* 31 (12). Nature Publishing Group: 1102.

Li, Xue, Xiangrui Meng, Athina Spiliopoulou, Maria Timofeeva, Wei-Qi Wei, Aliya Gifford, Xia Shen, et al. 2018. "MR-Phewas: Exploring the Causal Effect of Sua Level on Multiple Disease Outcomes by Using Genetic Instruments in Uk Biobank." *Annals of the Rheumatic Diseases* 77 (7). BMJ Publishing Group Ltd: 1039–47. doi:10.1136/annrheumdis-2017-212534.

Millard, Louise AC, Neil M Davies, Tom R Gaunt, George Davey Smith, and Kate Tilling. 2017. "Software Application Profile: PHESANT: A Tool for Performing Automated Phenome Scans in Uk Biobank." *International Journal of Epidemiology*.

Strug, Lisa J, Tanja Gonska, Gengming He, Katherine Keenan, Wan Ip, Pierre-Yves Boelle, Fan Lin, et al. 2016. "Cystic Fibrosis Gene Modifier Slc26a9 Modulates Airway Response to Cftr-Directed Therapeutics." *Human Molecular Genetics* 25 (20). Oxford University Press: 4590–4600.

9 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.

Practicum 10 min presentation, 10 slides.