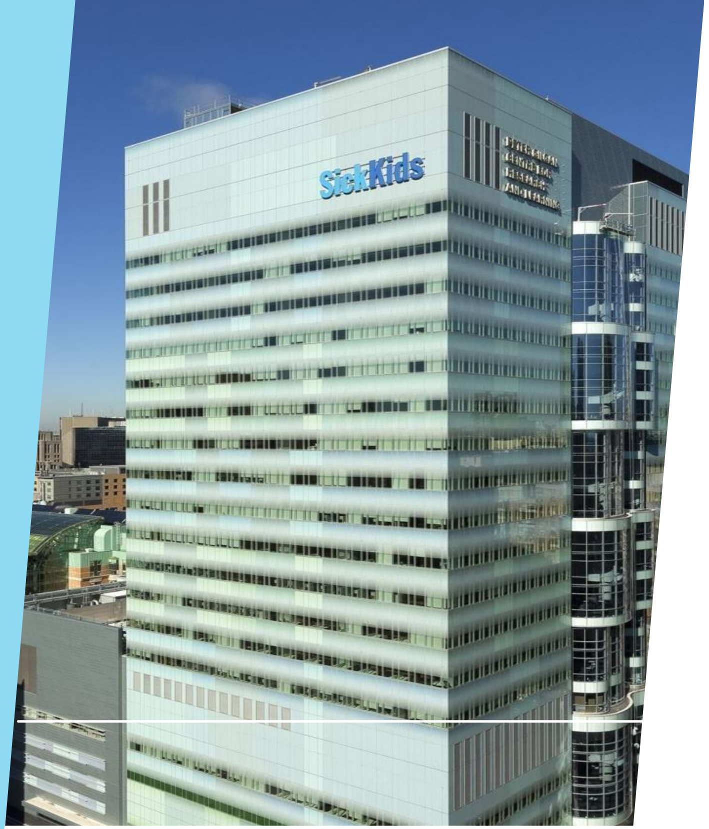
Phenome-wide Association Study of Cystic Fibrosis Modifier Genes

Supervisor: Dr. Lisa Strug

Faizan Khalid Mohsin



The Hospital for Sick Children

Strug Lab

SickKids®

RESEARCH
INSTITUTE

Overview

- Background: Cystic Fibrosis and Modifier Genes
- Research Question
- What is a PheWAS?
- PheWAS Methods
- Database: UK Biobank
- Limitations and Challenges

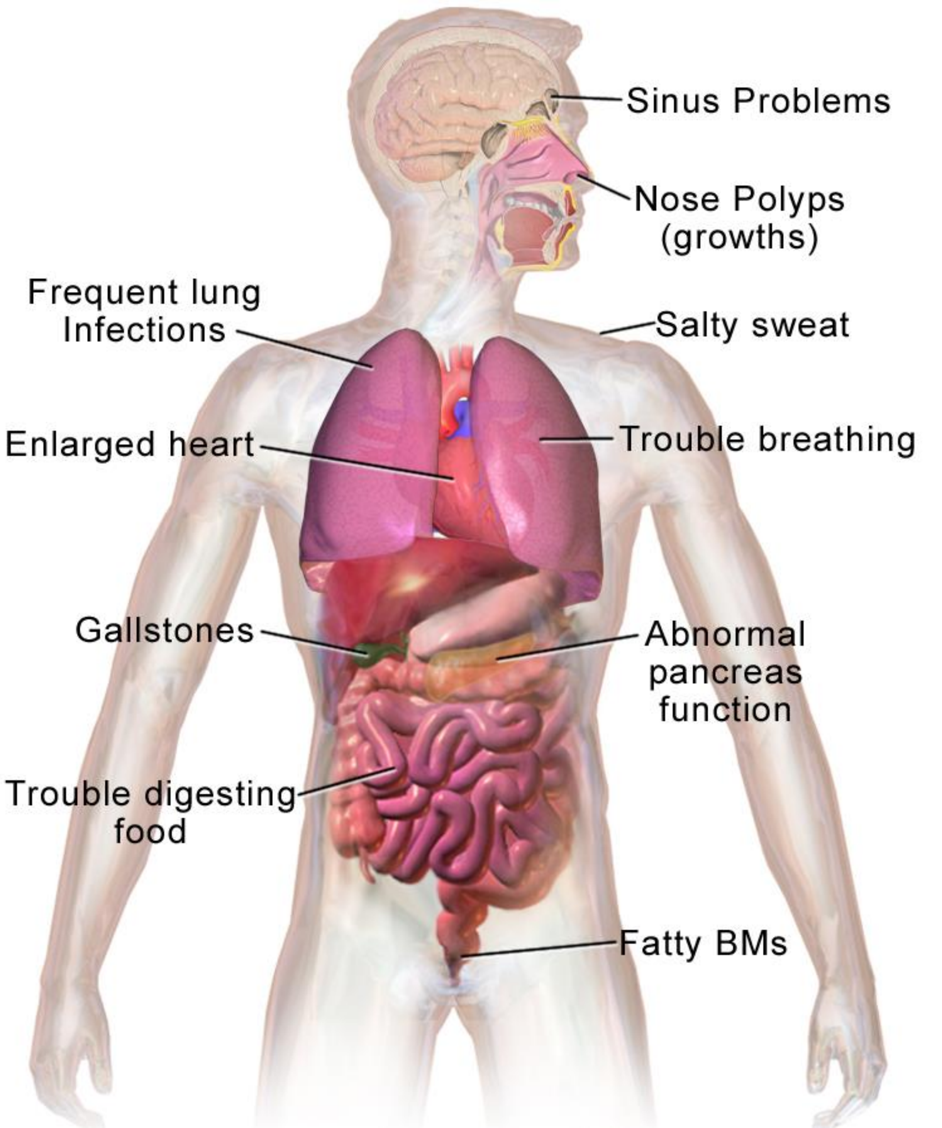
BACKGROUND

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.).

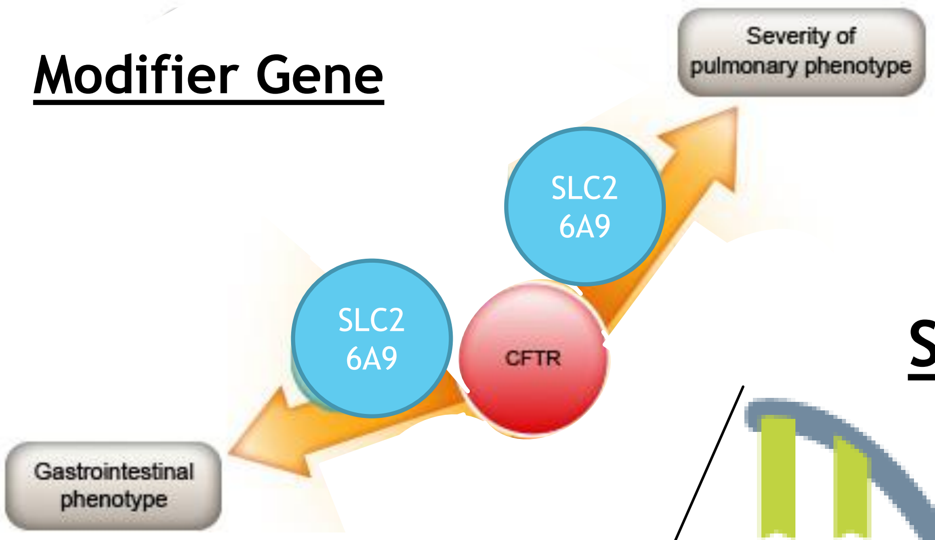


<u>**Typically**</u>: genotype (G) + environment (E) → phenotype (P)

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

Modifier Gene

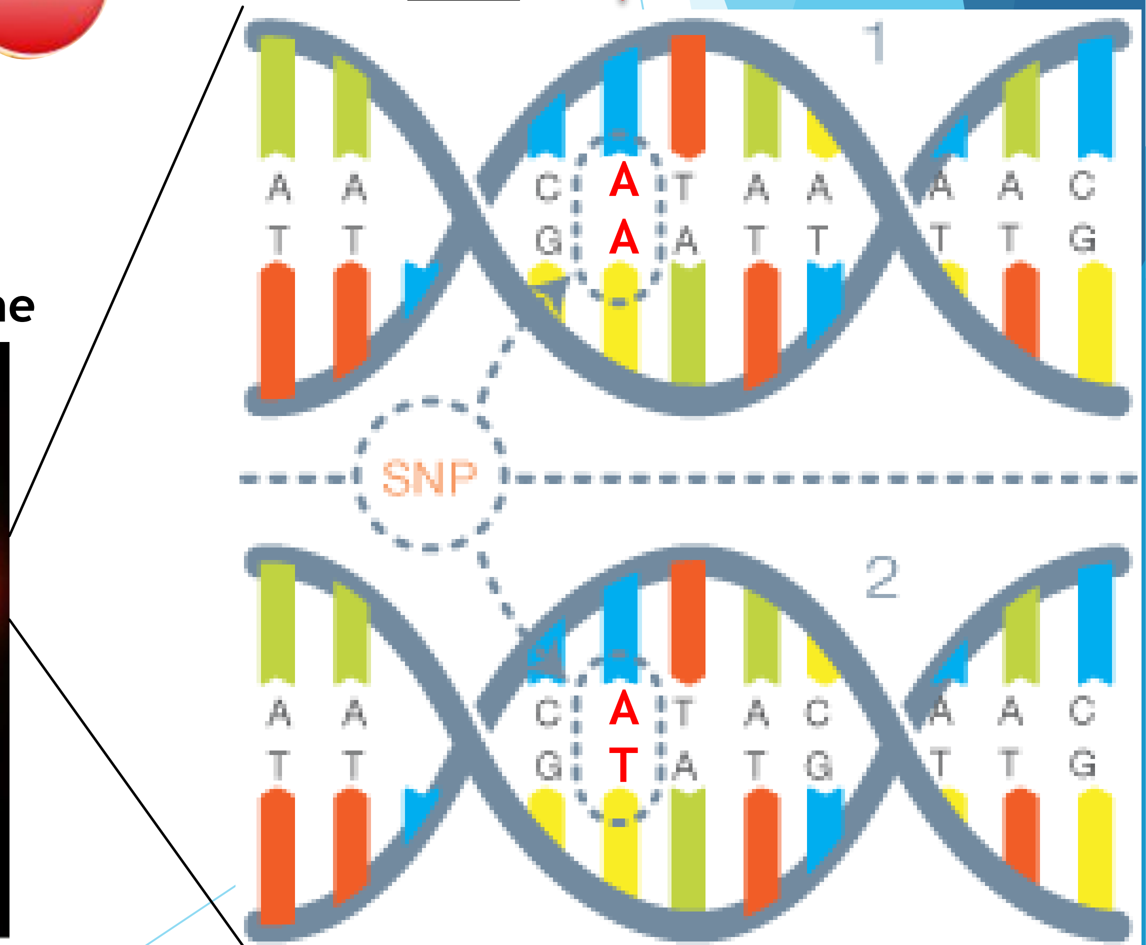


SLC26A9 Gene

RS4077468

RS4077468

SNP



Research Question

If do not have cystic fibrosis 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:

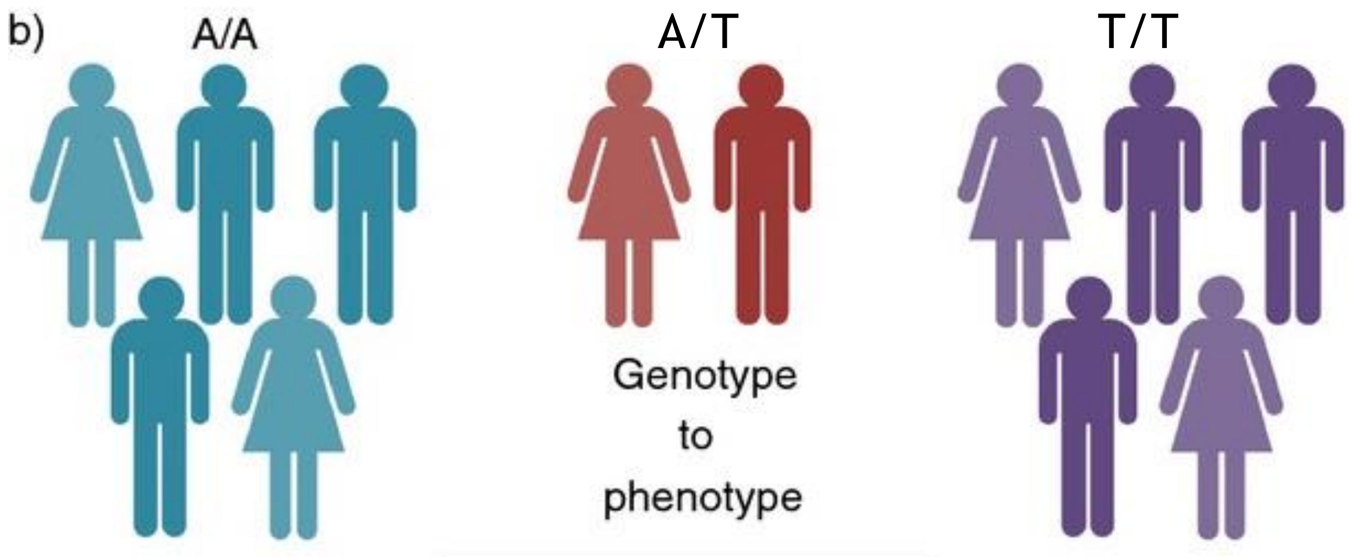
1) SLC26A9 (E.g. look at SNP RS4077468 with variation AA, AT or TT)

2) SLC6A14

3) SLC9A3

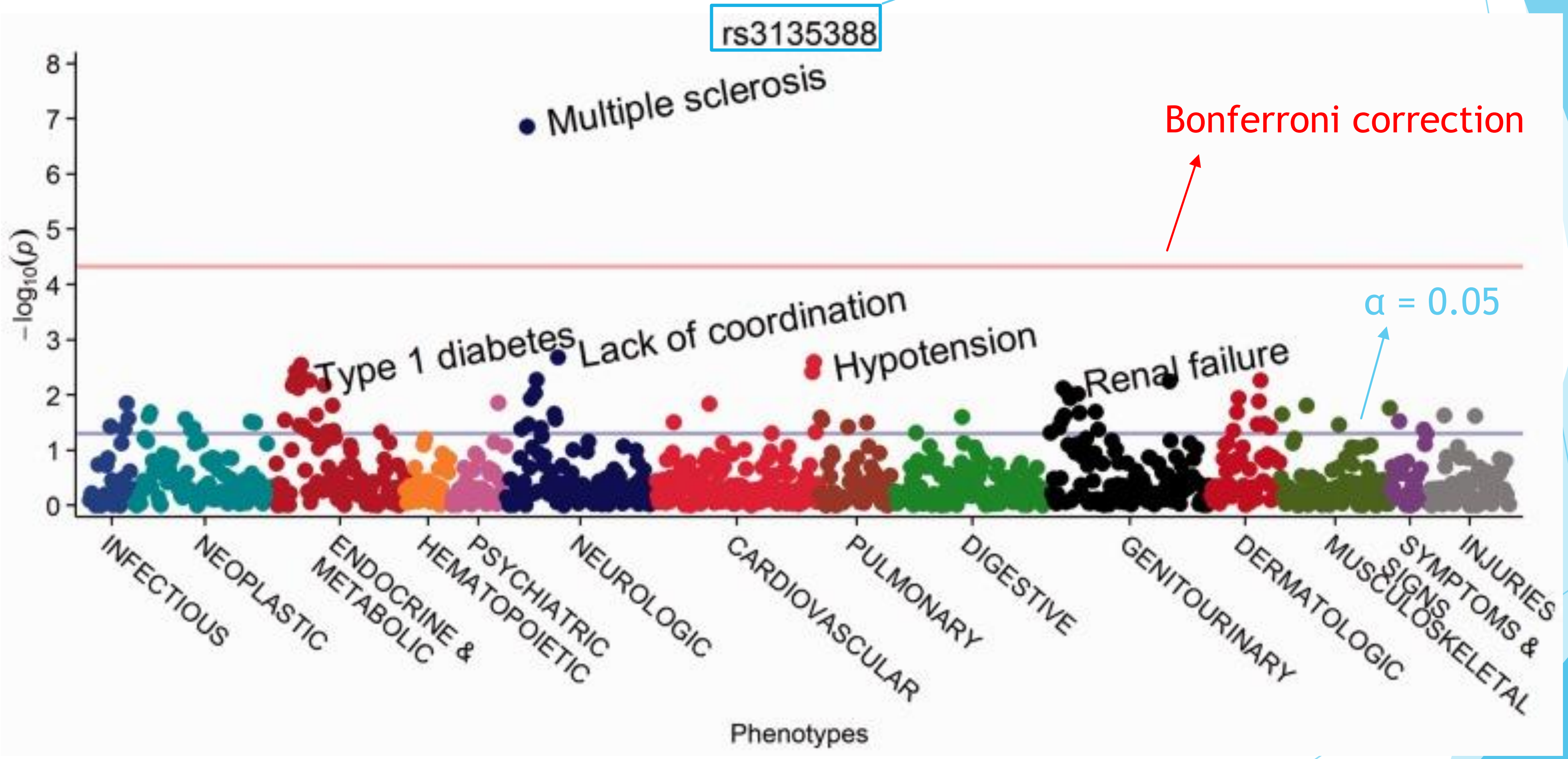
What is a PheWAS?

- PheWAS: Phenome-Wide Association Study
- Tests the association between genetic variants of interest with every phenotype measured.



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

PheWAS Manhattan Plot



Genotype

Bonferroni correction

α = 0.05

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

Phenome-Wide Association Study (PheWAS)

- **Statistical Methods for performing PheWAS**
  - Phenotype_i = covariates + SLC26A9 i=1, …, 2000

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

  - Covariates: Age, gender and ancestral PCA.
  - Perform Linear, Logistic Regression, etc. – depending on variable type of the phenotype

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

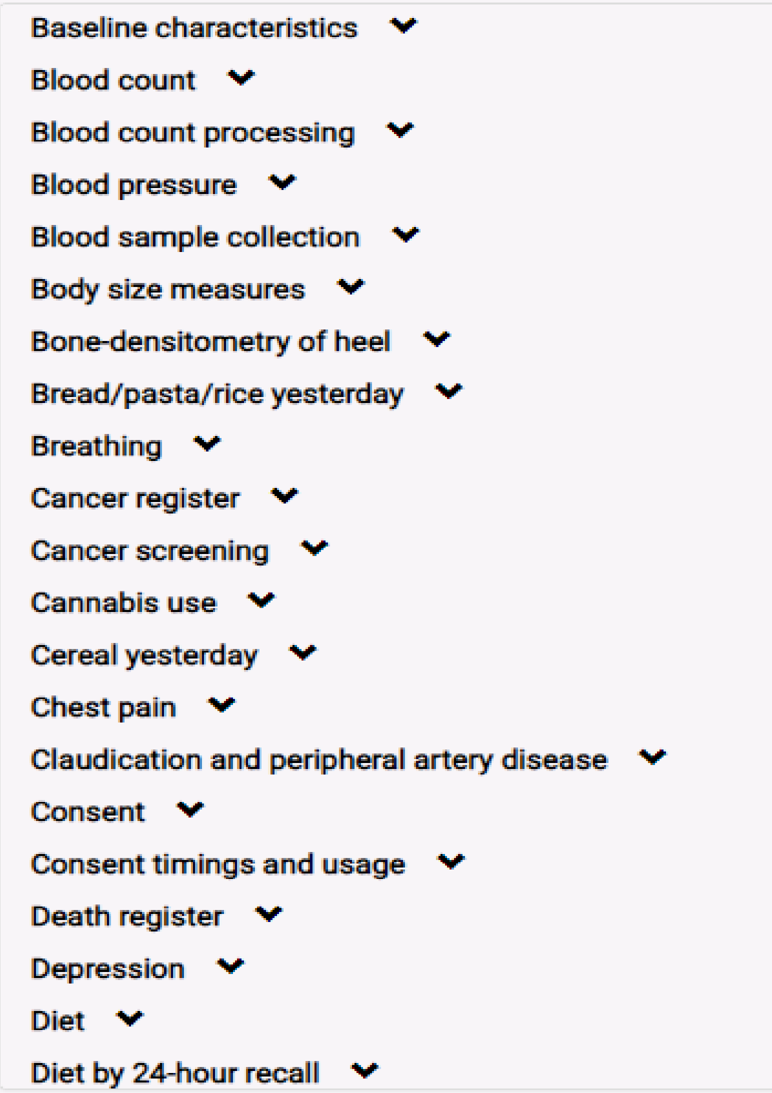
UK Biobank

Cohort

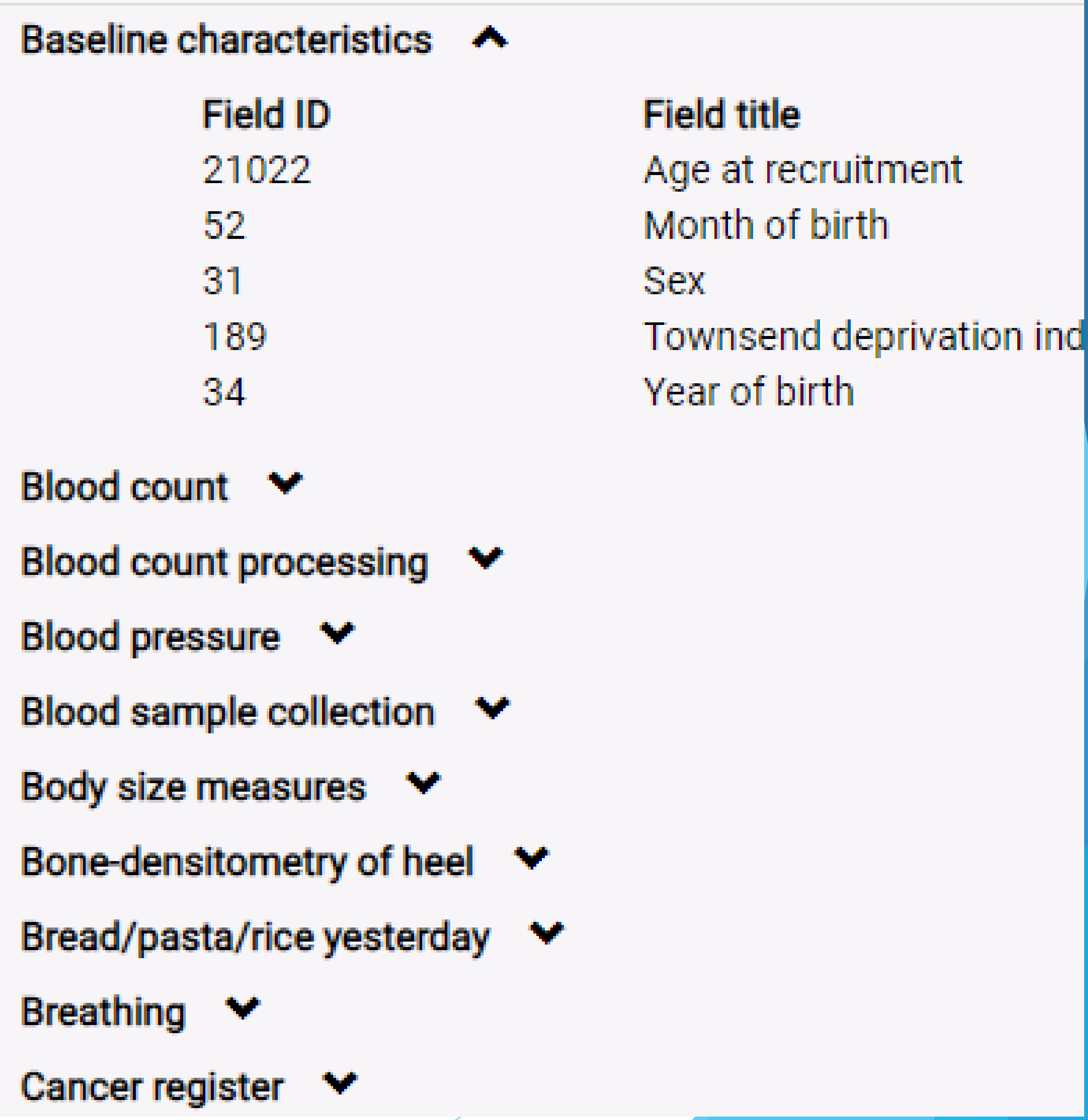
- 500,000 people aged between 40-69 years in 2006-2010 from across the country (UK).
- Volunteers recruited from England, Scotland and Wales.
- Approximately 2000 phenotypes (30GB).
- Genotype data (100GB).
  Micro arrays: between 500,000 to 1 million SNPs per person.

Anticipate spending a lot of time data wrangling.

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

Some Challenges and Limitations.

- Multiple testing.
  - 2000 phenotypes
  - Experimental wide α of 0·05
  - Bonferroni correction: *P* < 2·5E-5
- Missing data (phenotype & genotype).
- Choosing covariates.
- Relatedness, kinship checking.
- Computing the ancestral PCA.
- Implementing all of this in a HPF.
- External validation and generalizability of results.

Questions?