**Introduction**

The UK Biobank (<https://www.ukbiobank.ac.uk/>) is a great resource of data being collected on over 500,000 volunteer participants on a wide variety of illnesses, environmental exposures, genetic data, population characteristics, biological samples and imaging (<http://biobank.ctsu.ox.ac.uk/crystal/label.cgi>).

Ben Neale’s group has taken the UK Biobank resource with the genetic dataset, and performed a genome-wide association analysis (GWAS) on each phenotype. A GWAS consists of a regression with the phenotype set as the dependent variable, and the genotype per genomic region (aka. SNP or single nucleotide polymorphism) as the independent variable. Since a test is done on millions of SNPs, the multiple test burden requires that the p-value be lower than 5x10-8 in a GWAS analysis for significance. A significant result *tags* a genomic region for a relationship with the particular phenotype tested.

Often, SNPs nearby in a genomic region will be correlated due to linkage disequilibrium (LD), and one particular region may have more than one SNP associated. This presents a particular “pattern” of association of the SNPs tested in a plot of the -log10P-value versus genomic region, such that a particular association region will follow a particular pattern depending on what SNPs and to what extent those SNPs contribute to the phenotype being tested.

**Method**

While there are thousands of phenotypes that have been tested (<https://docs.google.com/spreadsheets/d/1kvPoupSzsSFBNSztMzl04xMoSC3Kcx3CrjVf4yBmESU/edit?ts=5b5f17db#gid=178908679>), we will focus on the following list of lung function phenotypes:

3062 Forced vital capacity (FVC)

3063 Forced expiratory volume in 1-second (FEV1)

3064 Peak expiratory flow (PEF)

20150 FEV1, best measure

20153 FEV1, predicted

20154 FEV, predicted percentage

20002\_1115 interstitial lung disease

22127 Doctor diagnosed asthma

22128 Doctor diagnosed emphysema

22129 Doctor diagnosed chronic bronchitis

22130 Doctor diagnosed COPD (chronic obstructive pulmonary disease)

22133 Doctor diagnosed sarcoidosis

22134 Doctor diagnosed bronchiectasis

22135 Doctor diagnosed idiopathic pulmonary fibrosis

22137 Doctor diagnosed tuberculosis

22502 Cough on most days

22504 Bring up phlegm/sputum/mucus on most days

Step 1 – Download the desired list of phenotypes using Python

These phenotypes can be searched and downloaded from <https://docs.google.com/spreadsheets/d/1kvPoupSzsSFBNSztMzl04xMoSC3Kcx3CrjVf4yBmESU/edit?ts=5b5f17db#gid=178908679>. Each phenotype analysis consists of about 14 million rows of summary statistics, each row summarizing the result per SNP in a tab-separated format.

Step 2 – Build SQL database into a giant SQL table with all the desired phenotypes

These results will be loaded from each tsv file and pushed to a SQL database in chunked phases.

The RefSeq database will also be required in order to draw the genes present in a particular genomic region that’s queried under the results. This will be obtained as a tsv file from the UCSC Table browser and pushed to a SQL database (<https://genome.ucsc.edu/cgi-bin/hgTables>). (While I could simply query the UCSC tables via MySQL (http://genomewiki.ucsc.edu/index.php/Programmatic\_access\_to\_the\_Genome\_Browser), this has proven to be slow in the past, but may explore this a bit more).

Step 3 – Flask

Build the RESTful API using Flask to deliver the query in JSON format. The query will consist of a specific genomic region and phenotype. The desired phenotype will be selectable from a dropdown.

Store the p-values and genomic region desired.

Step 4 – Website

The website will be simple, with a dropdown to select a particular phenotype and a textbox to enter the genomic region.

Step 5

*Time-permitting:*

* allow the selection of more than one phenotype – draw them as overlaid lines with a moving window selecting the minP per window and placing the x-value basepair coordinate in the correct place of the minP value within a window (ie., not just the middle of the window)
* allow uploading own GWAS datasets with a maximum window of about 2 Mbp to plot as dots, and selected phenotypes as lines overlaid on the same plot