Kim Prechtl

Final Project Proposal

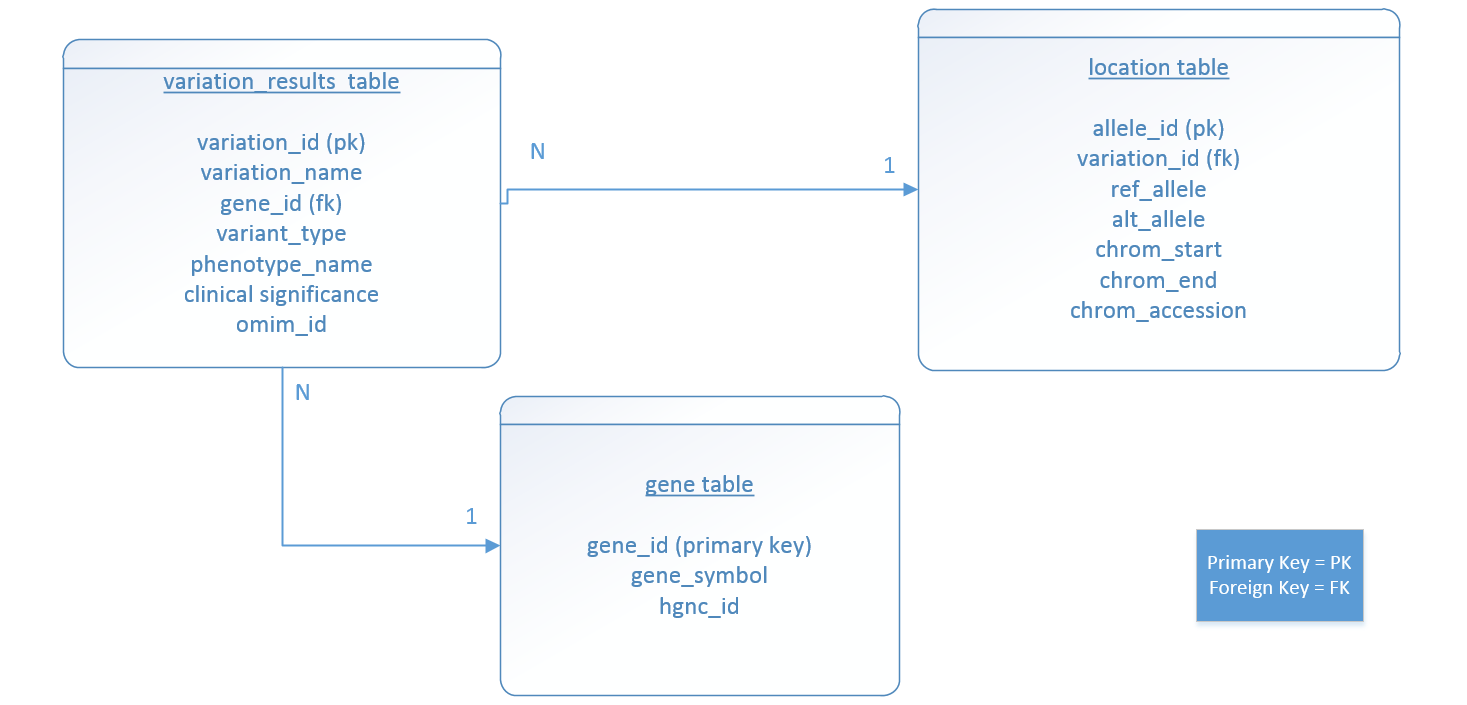
BACKGROUND:

*Question: What are the genomic variations are associated with a specific disease phenotype?*

When studying a specific disease, it is important for researchers to study different variations such as Single Nucleotide Polymorphisms (SNPs) and mutations to determine the origin of the disease or how to make new therapeutic targets. There is an online website called NCBI ClinVar that contains variations found in different diseases, but this website has a lot of extra information and can be difficult to navigate. The aim of this project is to create a database from the information found on NCBI ClinVar and then allow the user to search for their specific disease of interest through a user-friendly website. Once they enter in the disease of interest, they will be able to see all of the different SNPs and variations associated with that disease. It will also provide them with statistics such as the number of SNPs or variations found in this disease as well as how many of each different type.

PROJECT PLAN:

First, I will download a data file called a “variant summary table” from NCBI ClinVar. This text file that contains the SNP name, the gene it is associated with, the condition it is associated with, the clinical significance, the review status, Allele ID, type, name, geneID, geneSymbol, chromosomal location and clinical significance. I will parse the file using a python script. The text file has many columns separated by tabs. The python script will loop through each line of the file and pull the information that I am interested in into one location. With this data, I will make a database schema like the one I created below.



I will use MySQL commands to insert this information into the new database. Once the database is created, I will begin to create the user interface website. This website will give users a search box where they enter in the disease of interest. This search box will have an autofill function that will start to populate diseases as they start typing. This autocomplete box will be generated from a cgi script that runs a SQL query in the new database for the column that contains the disease name and then outputs the information into a JSON format (key: variant\_id, value: phenotypic\_name).

SELECT vr.variant\_id  
, vr.phenotype\_name

FROM variation\_results vr

WHERE vr.phenotype\_name like %

The % will be populated from the what the user enters in the text box. When they enter in the disease of interest and click “search”, another CGI script will run a SQL query to populate the results for that disease.

SELECT vr.variation\_name

, vr.variant\_type  
, vr.phenotype

, vr.significance  
, g.gene\_symbol  
, g.hgnc\_id  
, l.ref\_allele  
, l.alt\_allele  
, l.chrom\_accession

, l.chrom\_start  
, l.chrom\_end

, l.omim\_id

FROM snp\_variation\_results vr  
LEFT JOIN snp\_location l on vr.allele\_id = l.allele\_id  
LEFT JOIN snp\_gene g on g.gene\_id = vr.gene\_id

WHERE vr.phenotype like %

The results from this query will then be populated into a table that is defined on an HTML file. This HTML file will be generated using AJAX so that the page does not have to refresh. This table will be kept simple and only show the variation name, variant type, phenotype name, clinical significance and gene symbol. However, I will make the first variation name in each column row into a link so if you click it, it will take you to a page that contains more information about that variant in a list format including the OMIM number, the reference allele, the alternate allele and the chromosome accession number, start location and stop location. Statistics will be populated at the top of the original page, showing how many SNPs were found, what percentage are missense vs frameshifts, etc.