Segmentation of Somatic Mutation: Clustering

1. **Objective**: The goal is to **<u>segment different types of Somatic Germline Mutations</u>** in human genes associated with inherited and acquired diseases. It will be a one-stop-shop comprehensive collection of mutation data(Segments) for easy discovery in the era of personalized medicine. As part of this project I would like to find:

   a. Somatic Germline segmentation

   b. Acquired Disease with maximum number of Somatic Mutation

   c. Inherited Disease with maximum number of Somatic Mutation

   <u>Outcome</u>: Comprehensive Somatic Mutation Database an invaluable resource for all scientists.

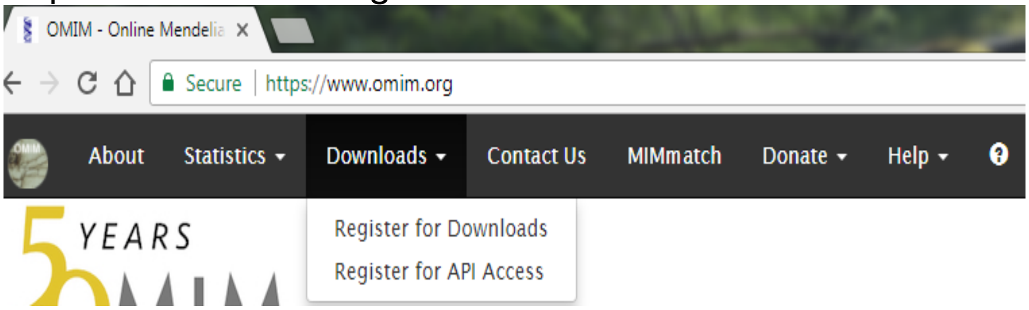
2. The Client

The client for this project is Georgetown University (www.gwu.edu) and the Bioinformatics Group. The purpose is to find an ML model which can be used to correctly cluster somatic mutation.

3. Data source and Credits

1. http://www.uniprot.org/Download Reviewed text file
   http://www.uniprot.org/downloads

2. https://www.omim.org/



3. dbsnp

http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/
Download snp147.txt.gz

4. Clinvar data ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/ Download Variant_summary.txt.gzUpdate 9/22: After discussing with Curator Navelke I have decided to keep assembly Grch38. Delete NCBI136 and Blanks as there is no mapped snp.

5. GWAS Data https://www.ebi.ac.uk/gwas/

6. Cbio portal

7. COSMIC

   http://cancer.sanger.ac.uk/cosmic/download COSMIC Mutation Data

Download

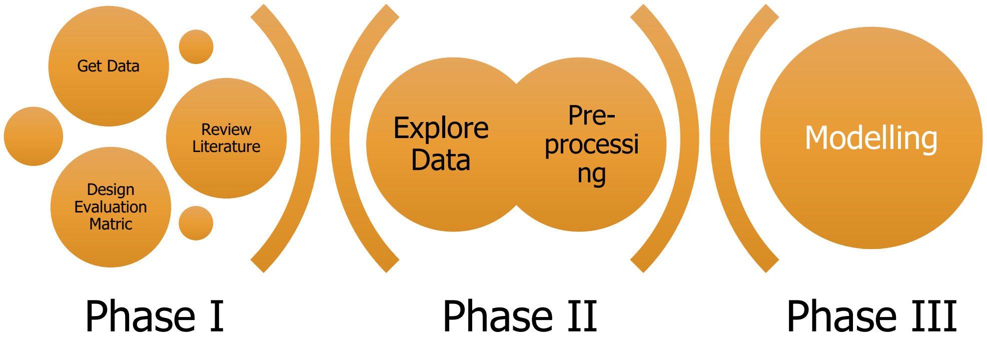
Download a full copy of the GWAS Catalog in spreadsheet format and current and older versions of GWAS diagram in SVG format.

A tab separated table of all COSMIC coding point mutations from targeted and genome wide screens from the current release. 'CosmicMutantExport.tsv.gz'

4. Solution Approach

I plan to use PCA and NMF to help with dimension reduction and segmentation of Somatic Mutation.

The solution approach is specifically designed to address large dataset with biodiversity and quality issues like redundancy, missing, wrong label etc. The solution is sub-divided into three phases as listed below.



a) **Data Assembly - Phase I**: This phase of the project is designed to gather and do basic cleanup like join, merge, add or update attributes.

b) **Explore and Preprocessing – Phase II**: This phase of the project is designed to validate and explore the dataset for all the problems listed in the "Problem" section of this proposal.

c) **Modelling and Evaluation Phase III**: In this phase of the project I will be exploring various machine learning algorithms to find the best model to cluster.

5. Project Deliverables

a) Project deliverables are listed below.

1. An analysis report (a .pdf document) on:
   a. Showing the different segments of Somatic Mutation
   b. Identify top 5 disease clusters for somatic mutation.
   c. Labeling and creating a database of gene mutation clusters.
   d. ML model
2. All project artifacts like – IPython Notebook with code, description and charts.
3. Project Presentation (.pptx)