OMass Bioinformatics Internship

Coding challenge

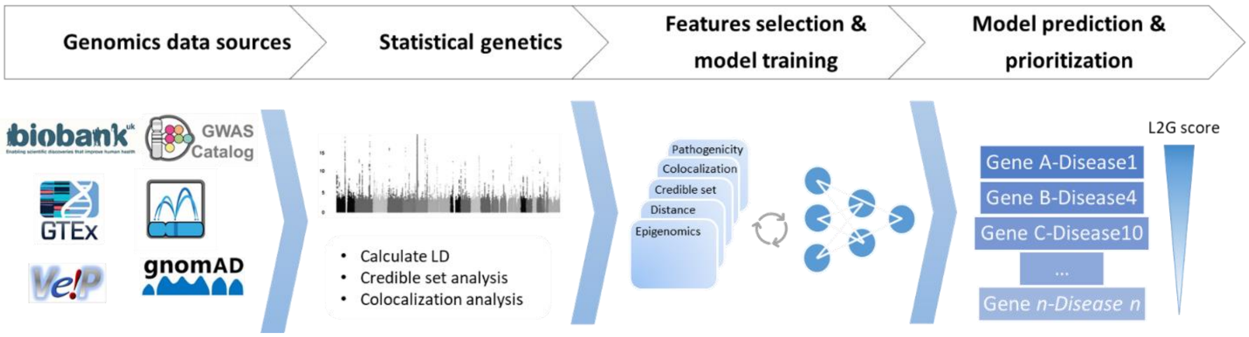
Duration 4 hours

This coding assignment is designed based on the advertise project to exercise the various programming skills to answer some relevant questions. Please follow the instructions on how to report your results.

Through the assignment you will be using the data file “**gene_disease_opt.csv**” which is a gene- disease association table programmatically collected from Open Target Genetics Platform.

Briefly, Open Target Genetics Platform uses a ML algorithm to combine various GWAS and functional genomics features for prioritizing potentially causal variant/genes for diseases.

To put it simple, in the data table, each **gene** is associated with a **disease** through a specific **variant** (coming from a GWAS) and each association has been scored by an ML analysis pipeline to make it possible to rank each of the gene-disease associations. The figure below summarizes how this table is generated.



you can read more about the Open Target Genetics Platform here

The coding challenge:

1- *Create a github repository (if you don’t have a github account, first you need to sign up for github) with your name and the following folder structure and clone it to your working environment.*

```
.
../code/
../output/
../graphics/
```

2- *Write a bash script that does the followings. (Save your bash script into yourfistname_lastename.sh and push it to your github repository).*

a) From the input file “gene_disease_opt.csv”, select the “target.approvedSymbol”,”disease.name”, “variantRsId”, “score” columns, then filter records for the disease name “rheumatoid arthritis”, and sort the data based on “score” column in descending order. Write the results to “/output/ rheumatoid_arthritis.csv”. (Hint- you can use | to pipe the output of bash functions that are needed to perform the filtering and sorting tasks).

b) Print the number of the records in the file that belong to the GPCR family (Hint: use the “gene_family” column). Please use the following line in the STDOUT.
“Number of the GPCRs:”

c) Finds all the records for the gene “SLC15A4” and replaces all the NA with 0 in “beta” column and saves it to the new file as “/output/ SLC15A4.csv”

In the second part, you will load the “gene_disease_opt.csv” file into your preferred programming environment (Python/R) and follow the instructions to generate some summary stats and graphics (The instructions assume you use R but it is up to you if you prefer to use Python).

d) Load the following libraries into your R environment: dplyr, ggplot2, tidyverse, (if you don’t have these libraries in your R environment, you are required to install them first).

e) Load the “gene_disease_opt.csv”

f) Group the associations based on “gene_family” and use “0.4” as cut off on “score” column to create a new logical variable called “selected”. Then, find top 5 associated diseases that are above the cut off in each gene family and save it as a “selected_associations.csv” file in the /output/ folder. Plot a stacked bar plot of the associations per gene family using “selected” column.

g) Find how many associations are based on studies published after 2020, then assess if any of the records carry missense variants (Hint. Use “publicationYear” and “variantFunctionalConsequence.label” columns).

h) Plot the distribution of all association scores (all data table) and color them based on *gene family* and save it into /graphics/score_dst.pdf

i) Write an R function (Python method) that takes the data as input and returns the sorted list of variant ids and their corresponding score as an indexed matrix.

j) Create a network plot of all gene-disease associations and map the scores to the edges. Color the nodes upon if they are “gene” or “disease”. Save the Network plot in /graphics/gene_disease_ntw.pdf

Please push all the files and codes of the assignment to the repository and share the link with amir.feizi@omass.com .

//End