**Biomedical Engineering 580.483/683**

**Annotate a Genome**

**Spring 2023 (3 credits)**

**Final Project**

This year, we will work in groups on the final project. Groups can be either 2 or 3 students. Make sure that your group contains at least one person who is a strong programmer and one person who has experience with machine learning.

There are two parts to the final project:

1. Do a mini-GWAS using 23andMe genotype files from OpenSNP.   
  
 You willl be using OpenSNP <https://opensnp.org/>

Pull down the Data menu from the main page and select Phenotypes

Here are the phenotypes with the largest number of users.

![Graphical user interface, table

Description automatically generated]()

There are 500+ phenotypes listed on OpenSNP. Browse through this list.

You will choose a phenotype, download the relevant genotype files, and do a GWAS.

Use the methods we applied in the Lecture 8 Homework.

Do you find any significant SNPs (before or after multiple testing correction)?

Show your results on a Manhattan plot.  
  
 Some tips:

Select a phenotype with >500 users.

If the phenotype is not continuous, figure out a way to binarize the phenotype

You will need the following files to do your analysis:

https://www.opencravat.org/data/phenotypes\_202304020100.csv

CSV file in which phenotypes are the columns and rows give each user's id and each phenotype characteristic.

To get the 23andMe genotype files, write a script to download these files for the users you will be working with. You can wget these files from opencravat.org using the file names in genotype\_files.txt (on Piazza) and this command:

wget "https://opencravat.org/data/open-snp/FILENAME"

Example

wget "https://opencravat.org/data/open-snp/ user10026\_file8287\_yearofbirth\_unknown\_sex\_unknown.23andme.txt"

You can convert 23andMe format to plink format with this command:

plink --23file 23andme\_raw\_data.txt --make-bed --out mydata

2. Develop a machine learning prediction model (of your choice) for your phenotype using selected SNPs as features. To assess the model's performance, you will split it into a training and test set, then use the training set to discover SNPs (as you did above). You will use the test set to evaluate performance.

Possible machine learning models: neural network with one or two hidden layers, random forest, naive bayes, logistic regression, support vector machine.

Requirements for the final project.

1. Prepare a powerpoint presentation (approximately 20 minutes long) describing your project and present it to the class.

2. Write a short paper (5-10 pages). Paper should be of professional quality, including citations.

Presentations will be on 4/26/23

Paper and powerpoint slide submission is due on 5/5/23

Submit short paper and your powerpoint slides to annotateagenome@gmail.com

We have posted on Piazza examples of an excellent paper and presentation.