## Project Detailed Research Plan

**Date: 11/12/2017**

**Student Name: Siddharth Sharma**

**Project Title:** What will be my eye color? Can I get Diabetes? Enhanced Machine Learning and Genomics based approach for the prediction of common traits and diseases (termed phenotypes) among humans.

**Category**: Biological Sciences and Engineering: Computational Biology and Bioinformatics

The purpose of my project is to enhance existing machine learning and genomics-based approaches for the prediction of common traits (or phenotypes) among humans.

Background

Increasingly, genomics is being used for the prediction of specific traits and diseases (phenotypes) among humans. A trait is observable characteristic that is inherited and attributed to genes. Examples of traits include height, eye color, IQ, genetic diseases (Prostate or colorectal cancer, breast cancer, Type-2 diabetes) etc. Wider availability of genomics data through multiple research projects has been a catalyst in that direction. The 1000 Genomes project <http://www.internationalgenome.org/> is an example of such research project.

Hypothesis

With the recent advances in machine learning, deep learning and big data analysis, data computation resources and data models needed for genomics data analysis are readily available. For example, tensorflow and mxnet are two widely adopted open source frameworks for deep learning. However, the prediction of traits and diseases has its own challenges in terms of statistical analysis (whether a variable is an independent variable or not), overfitting of the data models, and limited quality of data collection where research studies have data from less than 100,000 subjects. Linear Mixed Models [1] (LMM, a type of linear regression) is a common approach for Genome-wide Association Studies (GWAS) for predicting common traits using genomics. Linear mixed-effects models are an extension of linear regression models for data that is collected and summarized in groups.

Project Goal

The goal of my project is to analyze and experiment with existing LMM-based approaches for Genome-wide Association Studies (GWAS) for the prediction of common traits in humans, identify any gaps and challenges in these existing approaches, and then propose an enhanced approach on how to address these gaps and challenges using machine/deep learning and data science frameworks.

Experiment Design

I will conduct an experiment by performing data analysis and predictions using Linear Mixed Models approach for the prediction of traits on an existing genomic dataset, namely International HapMap project [4]. My goal will be to understand how these existing LMM-based approaches work for Genome-wide Association Studies for the prediction of common phenotype traits among humans. Once I understand these existing approaches, I will run the same experiment with my proposed approach and enhancements that aims to addresses any gaps and challenges in these existing approaches. My goal in this experiment will be to perform a better prediction of specific phenotype traits using my enhanced approach. I will collect experimental data on the quality of analysis and predictions using GWAS and LMM approaches, and compare the results with my proposed enhanced approaches.

Bibliography References

[1] Linear Mixed Models: <https://www.mathworks.com/help/stats/linear-mixed-effects-models.html?requestedDomain=www.mathworks.com>

# [2] The benefits of selecting phenotype-specific variants for applications of mixed models in genomics: <https://www.ncbi.nlm.nih.gov/pubmed/23657357>

[3] Potential etiologic and functional implications of genome-wide association loci for human diseases and traits <https://www.genome.gov/pages/about/od/newsandfeatures/pnasgwasonlinecatalog.pdf>

[4] <ftp://ftp.ncbi.nlm.nih.gov/hapmap/>