

Project No.	Project Title	
2021-01-186	Application of machine learning on gene-gene interaction	
Academic Advisor		Co-Advisor
Dr. Jonathan Rosenblatt		Dr. Nadav Rappoport
Team Members		
Yarden Hochenberg		
hochyard@post.bgu.ac.il		

Abstract

Background: the variance explained by genetic variants as identified in genome-wide genetic association studies is typically small compared to family-based heritability estimates. Explanations of this "missing heritability" have been mainly genetic, such as genetic heterogeneity and complex genetic mechanisms. Our goal is to create a machine learning model based on large and wide genomic data, in order to identify complex relationship between SNPs and phenotypes.

Methods: our method involves building a baseline model using Polygenic Risk Score based on Genome Wide Associate Study (GWAS) statistics. We will implement that using Python as an infrastructure for further use of gene interactions for phenotype prediction. Comparison of different libraries in Python that allow the construction of a regression model in order to optimize computational running time.

Results: we conclude that solve function in Numpy.linalg package is the fastest way for construction of a regression model to identify significant SNPs. The running time of using Numpy library is about 5 times faster than other tested Python libraries like Scikit-learn and Statsmodel.

Conclusion: the model built with Numpy library can later in the study enable us to incorporate gene-gene interactions in an accessible way.

Keywords: genome-wide association study (GWAS), single nucleotide polymorphism (SNP), missing heritability, computational complexity.