Application number/Title: 26664 - Novel machine-learning framework for improved inference in GWAS

Applicant PI: Prof Michal Linial

Applicant institution: Hebrew University of Jerusalem, Department of Biological Chemistry, Edmond J. Safra Campus, Jerusalem, 91904, Israel

Funding: ELIXIR

Keywords provided by the Applicant PI to describe the research project:

GWAS, machine learning, methodology, complex diseases

Application Lay Summary:

1a: Finding associations between genotypes to phenotypes (GWAS) is crucial for future precise personalized medicine. Current approaches are very limited due to the small number of available samples relative to the huge amount of studied genetic variants (e.g. SNPs). We want to develop a new GWAS framework making use of machine-learning approaches that would allow much stronger statistical inference. An improved statistical framework would allow finding many new associations and have a positive impact in bringing these discoveries to clinics. The success of this method is strongly dependent on abundance of genetic data coupled with rich phenotypic information.

1b: The low discovery rate of GWAS undermines the research community's efforts at bringing personalized medicine. An improved GWAS methodology is thus a burning need that should be at the top of the public interest as a health-related research. Success of our project will not only directly find new genetic associations, but will also allow future studies making better use of genotype-phenotype databases like the UK Biobank. Our project has the potential of improving the understanding of complex diseases inflicting a substantial fraction of the population (such as diabetes type II, asthma, and cardiovascular, autoimmune and neurodegenerative diseases)

1c: We will process tens of thousands of UK Biobank's high-quality genetic samples, each comprising of close to 800,000 informative SNPs, in order to predict the damage on each of the ~20,000 human genes for each of the studied

individuals. Combining these assessments with the detailed phenotypic data available in the UK Biobank, we will use statistical methods to uncover the associations between these phonotypes to specific genes. We will use machine-learning algorithms in order to assign the proper effect size of each gene.

1d: Any individual with genetic information (DNA sequencing or SNP-array data) will improve the quality of our model and overall effectiveness of our framework. The key for success in finding significant associations despite the unavoidable variation in the population is having a very large dataset of individuals. According to our estimates, we will need at least tens of thousands of samples. A rich repertoire of phenotypes is crucial (to serve both as predicted and predicting variables). Therefore, we will need the entire set of individuals with genomic data coupled with all of their phenotypic attributes.