Application number/Title: 43661 - Predicting heritable phenotypic traits and conditions from genomic data using cutting-edge deep learning methods

Applicant PI: Dr Gregory Kucherov

Applicant institution: Skolkovo Institute of Science and Technology, Moscow, Russian Federation

Keywords provided by the Applicant PI to describe the research project:
deep learning, genome-wide association studies, genotype-to-phenotype prediction, machine learning

Application Lay Summary:

The sequencing of the human genome and the subsequent advent of large-scale genotyping and high-throughput sequencing promised to usher in the era of truly personalized medicine, where an individual's genetic disease risk could be accurately evaluated and medical treatments tailored to his or her genotype could be better defined. Successes in this area, however, are mostly limited to cases where disease status or therapy choice depend on one or few genetic variants, such as in monogenic diseases or certain types of cancer. In the general case, the problem of mapping genetic data to actual phenotypic traits, such as disease risk, is far from being solved; it is particularly challenging for traits and conditions that arise from interactions among many genes. To help to bring about the fulfillment of the promise of personal genomics and truly individualized medicine, we plan to tackle the genotype-to-phenotype problem using complex state-of-the-art machine learning techniques, such as gradient-boosted decision trees and deep neural networks. These methods have already demonstrated their success in many difficult medical tasks such as automated recognition of diseases from tissue images (e.g., for diabetic retinopathy or tumor detection) or from EEG/ECG signals (e.g. for cardiac arrhythmia detection), and should significantly improve our understanding of the relationship between phenotypes and genomic data. The power of these methods comes from having very large data sets for training, and UK BioBank is a unique source of genotype and phenotype data at the scale necessary for contemporary machine learning to fulfil its potential. We plan to train our models to predict phenotypes of different kinds (height, BMI, presence or absence of medical conditions, such as bipolar disorder, certain autoimmune diseases, and others) with a range of heritabilities

and genotype-environment interactions; the diversity of phenotypic data available at the UK BioBank will make this exploration possible. When training the models, we plan to consider environmental and behavioral features, also available in the BioBank, in addition to the genomic information. Furthermore, we plan to improve "traditional" interpretable models using insights obtained from applying ML techniques. Our work should contribute to the advancement of the general area of genotype-to-phenotype prediction and, in particular, provide more accurate models for disease prediction from genotyping data. We expect our project to complete in three years.