Predicting Disease Status from Gene Expression Data

Nick Treloar, Jiamin Zou, Yalei Zhao

Motivation

In modern biomedical research, large-scale gene expression data have made it possible to study how expression patterns relate to human disease in a meaningful way. At the same time, these datasets are high-dimensional and computationally challenging, which motivates our interest in exploring gene–disease relationships using real-world genomic data.

Data

We will use the GenoTEX dataset from Kaggle https://www.kaggle.com/datasets/haoyangliu14/genotex-llm-agent-benchmark-for-genomic-analysis/data. In total, GenoTEX includes 1384 gene–trait analysis problems derived from 911 cohorts, with an average of 167 samples per cohort. This dataset is organized by trait: each trait directory contains 1–11 cohort datasets, and each cohort dataset includes clinical data with a binary trait-status variable (0 = control, 1 = disease) and high-dimensional gene-expression data with approximately 18,530 normalized features.

Question of Interest

We are interested in identifying which genes are most informative for predicting a given disease across multiple cohorts, and whether these outcomes align with what has been reported in the biomedical literature.

Computational Challenges

The gene-expression data are high-dimensional, with more genes than samples, hence, feature selection and model stability is one challenge. We also need to fit models across multiple cohorts, each with different sample sizes and gene distributions. These challenges require efficient regularization and computational strategies to avoid overfitting and to handle the scale of the data.

Proposed Approach

We plan to fit random forest models to predict disease status for its outstanding performance in handling nonlinear effects and providing variable-importance measures to identify influential genes. We will run the models on the cluster using batch job submission to avoid substantial runtime for large feature sets.

After developing a model for each trait of interest, we will apply the same pipeline to data from other traits to see whether the model generalizes. If the approach transfers well, it can be expanded to all available cohorts. This helps us assess the stability of our model and examine whether gene–disease patterns are trait-specific or shared across related conditions.

Expected Outcomes

We expect to identify a set of genes that are strongly predictive of disease status within each trait. Some of these genes may match known causal or disease-associated genes reported in the literature, while others may appear as potential biomarkers without established causal roles. By comparing results across traits, we anticipate seeing both trait-specific gene signals and broader patterns that may be shared across related conditions.