Formulating a Gene Signature for Diagnosis of Autoimmune and Infectious Diseases

Gupta, Riya (School: Saratoga High School)

Many infectious and autoimmune diseases are difficult to diagnose properly, and the current diagnostic tests for these diseases are either pathogen/antibody specific, or have a slow turnaround time. A drug for one disease may harm another, so I am working with the Khatri Lab at Stanford to create a gene signature to accurately diagnose these diseases. We collected and curated the blood transcriptome profiles from 23,572 patients with 7,532 genes across 224 independent cohorts from over 50 countries and sorted them into 23 major disease groups. We used a co-normalization method called COCONUT to merge the data, using the healthy samples in each dataset as a common baseline. After trying machine learning models such as Random Forest and SVM, I used MANATEE's statistical analysis to create my gene signature. MANATEE avoids overfitting to the training data by starting with a small number of genes and using a greedy algorithm to add or subtract genes to maximize the AUC. I created two signatures to differentiate autoimmune/infectious from healthy samples and to differentiate autoimmune from infectious samples. My model achieves a ROC AUC of over 0.87 on validation datasets. Because my data includes heterogeneity across countries, specific disease, and more, my gene signature can be used in almost any clinical population, including both developing and western countries, as a blood test screening tool. In the future, I would like to translate my signature into clinical practice following wet lab confirmation and expand it to more disease groups.