Discovering Biomarkers for Breast Cancer Subtypes With Gene Expression Data Using Machine Learning

Lin, Daniel (School: Fairview High School)

Breast cancer is the most commonly diagnosed cancer in the world and accounts for the fifth most cancer deaths globally. Breast cancers are generally categorized into four main subtypes, namely basal-like, HER2+, luminal A, and luminal B, which are differentiated by their gene expression signatures. The identification of breast tumor subtypes with genomic information is critical to determining treatment options for patients. Elastic net regression, a regularized regression model that combines the penalties of both lasso and ridge regressions, is very useful for feature selection when there are groups of highly correlated independent variables in the data, like gene expression data. Here I employed elastic net regression on breast cancer microarray data to identify signature genes for the breast cancer subtypes. The selected signature genes were verified using publicly available tools. Functional analysis of selected signature genes revealed that hormone therapy is not suitable for basal-like breast cancer, as suggested by the identified steroid-related terms. The results also suggest that immunotherapy is ineffective for luminal A and luminal B breast cancers. This study identified important and novel breast cancer biomarkers for different subtypes and established a framework for identifying disease-associated genes. These biomarkers shed light on breast cancer diagnostics and treatments and the framework established in this study can also be used on other diseases.

Awards Won:

Arizona State University: Arizona State University ISEF Scholarship (valued at up to $52,000 each)