Predicting Molecular Phenotypes with Single Cell RNA Sequencing: An Assessment of Unsupervised Machine Learning Models

Dunca, Anastasia (School: West High School)

The National Cancer Institute reported 9.5 million cancer-related deaths in 2018. A challenge in improving treatment is resistance in genetically unstable cells. The purpose of this study is to evaluate unsupervised machine learning on classification of treatment-resistant phenotypes in heterogeneous tumors through single cell RNA sequencing (scRNAseq) data analysis with a unique pipeline and innovative evaluation metrics. scRNAseq quantifies mRNA in cells and characterizes cell phenotypes. Two scRNAseq datasets were analyzed: cells of different cell phases (S, G1, G2/M) and tumor/non-tumor cells of different molecular subtypes. Accurately identifying these cells is vital because irregular cell cycles may fail to respond to treatment, and tumor cell subtypes may have resistant phenotypes. The pipeline consists of data filtering, dimensionality reduction with Principal Component Analysis, projection with Uniform Manifold Approximation and Projection, clustering with nine methods (Ward, BIRCH, Gaussian Mixture, DBSCAN, Spectral, Affinity Propagation, Agglomerative, Mean Shift, K-Means), and evaluation. Six models divided G2 v. S cells; Spectral, Ward, and K-Means ranked highest with ~60% accuracy. Seven models divided tumor v. non-tumor cells; K-Means, Ward, and BIRCH ranked highest with ~80% accuracy. Because Ward and K-Means ranked high in both tasks, analysis was optimized by using these models; three subtypes in the identified tumor cell population were discovered, verifying the efficacy of unsupervised analysis in tumor heterogeneity research. In clinical settings where there is no standard scRNAseq analysis protocol, this pipeline can be used to generate clusters that hold key information about the tumor microenvironment, directly affecting the success of treatment.