A Novel Computational Pipeline To Identify Target Genes That Define Cancer Subtypes To Improve Cancer Detection and Therapy

Parikh, Vatsal (School: Sunset High School)
Mittal, Ekansh (School: Westview High School)

Background: Cancer is the second leading cause of death globally. The heterogeneous nature of cancers with multiple subtypes makes them challenging to treat. We hypothesized that a computational approach could be implemented to identify genes and pathways specific to cancer subtypes that can serve as biomarkers and therapeutic targets. Methods and Results: We used a breast cancer dataset as a proof-of-concept for establishing a computational pipeline. We trained a logistic regression model on the dataset to predict the cancer subtype based on gene expression. Then, we applied the ExtraTreesClassifier dimensionality reduction technique to select 500 genes from the initial 8000 genes to train the model. This improved the predictive accuracy from 79% to 86%. Additionally, hyperparameter tuning was used; this increased the model accuracy from 86% to 92%. Using this model, we selected 386 genes for each subtype, from which 20 significant genes were determined using differential expression. These genes are associated with various cellular processes impacting cancer progression and are targetable. Specifically, the most aggressive subtype, triple-negative, can be targeted using a combination of IGF1R and FOXA1 inhibitors, and the Luminal A and HER2+ subtypes can be targeted using NTRK2 inhibitors. We applied our pipeline to a low-grade gliomas (LGG) dataset and found that LSD1 can be targeted in non-codel LGG. Conclusions: In summary, we established a computational workflow that can be applied to multiple cancer types to identify specific cancer subtypes and targetable pathways. In the future, we plan to apply these approaches to improve patient outcomes.