Categorization of Kidney Cancer Using Machine Learning Based on lncRNA Expression

Bodily, Lily (School: Caddo Parish Magnet High School)

Kidney cancer treatment depends on several factors, such as cancer subtype and stage. It is extremely important that the doctor knows what subtype of kidney cancer their patient has in order to most effectively treat them. Current diagnostic methods use biopsies, procedures where tumor samples are viewed under a microscope, to determine subtype. However, molecular analysis offers greater precision and versatility in a clinical setting. One potential avenue is the use of long non-coding RNAs (lncRNAs) as biomarkers. This subtype of RNA is >200 nucleotides in length and does not contain information to make a protein. They have a diverse array of functions, particularly in regulating the gene expression of protein-coding RNAs. In addition, they are cancer-type specific: about 60% of aberrantly expressed lncRNAs are only abnormal in one type of cancer. By measuring levels of certain lncRNAs, it is possible to ascertain the specific type of cancer. I hypothesized that these molecules can also be used to differentiate between subtypes of the same cancer (kidney). To explore this idea, I performed a Principal Component Analysis on lncRNA data obtained from TANRIC, a RNA-seq database hosted by the MD Anderson Cancer Center. This technique simplified the data into dimensions more easily used to train a machine learning algorithm. I trained a logistic regression model on two of the principal components found in the PCA. When used to distinguish different forms of kidney cancer, the model achieved an accuracy of 92%. By combining the disease specificity of lncRNAs with the predictive power of machine learning algorithms, we can increase the accuracy, efficiency, and ease of cancer diagnoses.