Computational Method of Analyzing Genetic and Epigenetic Information for Predicting Mutations in Colorectal and Pancreatic Ductal Adenocarcinoma Cancers

Long, Hannah (School: Magoffin County High School)

This study investigates the carcinogenesis of cancer drivers by leveraging epigenetics, genomics, known biomarkers, and phenotypic information creating an algorithm identifying tumor suppressors (TSG) and oncogenes (OG), focusing on all driver mutations, including rare drivers such as KMT2D. KMT2D is a rare epigenetic regulator that mediates histone 3 lysine 4-trimethylation by forming a multi-protein complex. Studies have linked a connection between epigenetic alterations and tumorigenesis of multiple types of cancers (Lyu et al, 2020). Due to low background mutability rates, rare drivers are hard to distinguish from "typical "drivers. The framework of the algorithm is similar to another computational approach using DORGE (Lyu et al.), which took advantage of known epigenetic, genomic, and phenotypic information to create an applicable algorithm that identified both ordinary and rare drivers. After researching, the decision to include additional features to the dataset arose: CIMP (CpG Island Methylation Phenotype) and MSI (Microsatellite Instability). The dataset created for this work included multiple known drivers and neutral genes to validate results. Using the scikit-learn software package in Python enabled the comparison of several known algorithms to identify which garnered the best accuracy. The methodology included creating a LASSO, RIDGE, Elastic Net, Random Forest Model, and Support Vector Machine. The results consist of successful comparisons of algorithms that could aid research studies in genomics, artificial intelligence (AI), epigenetics, cancer driver database creations, pharmaceuticals, therapy options, and many more.