A Multi-Omics Predictive Pipeline for the Discovery of Cancer Driver Genes

Beesam, Shriya

Beesam, Adhya

Cancer is among the most debilitating diseases in the world with its pervasiveness originating from its complex genetic makeup. Although cancer research is at the forefront of medical concern, the specific genes that catalyze tumorigenesis are still not well understood. There is currently no accurate method for identifying the causal “driver genes” in cancer. The lack of a specific understanding in cancer’s genetic landscape has led to deficiencies in fields such as drug target and biomarker identification and consequently an inability to effectively treat certain cancers. Therefore, a novel pipeline utilizing machine learning was developed for cancer driver gene discovery. The pipeline consists of three steps: (1) Identifying cancer-specific differentially expressed genes (DEGs) in mRNA, methylation, and somatic mutation data, (2) Developing a machine learning framework for feature selection, and (3) Predicting probable candidate driver genes. The pipeline was applied to a case study of kidney renal adenocarcinoma (KIRC) with 3982 DEGs identified in mRNA, 263 DEGs in methylation, and 15 DEGs in somatic mutation levels. The multi-omics framework utilized feature selection to identify the most significant genes from both DEG data and the genes' subsequent protein-protein interaction (PPI) data, and was validated using survival curves and literature-based evaluation tools. It identified many previously validated driver genes as well as novel candidate drivers, and provided an efficient pipeline to propel advancement in cancer research from diagnosis to treatment. Ultimately, this pipeline has the potential to revolutionize oncology, and has many far-reaching applications for the future.