Multi-Omics, Big Data, and How Machine Learning is Revealing New Biological Insights in Pediatric Brain Cancers

Shehbeel Arif[1,2,3]

[1] Drexel University College of Medicine, Philadelphia, PA
[2] Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA
[3] Center for Data Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA

Introduction: The advent of “omics” sequencing technology has revolutionized our understanding of human genetics, but also has created the paradoxical situation in which many scientists are now metaphorically “drowning” in vast amounts of biological data. As we find ourselves submerged in this genomic deluge, there is an urgent need for innovative strategies and advanced computational approaches to navigate, interpret, and extract meaningful insights from the ocean of omics information. This brief case study demonstrates how Machine Learning can be used to integrate vast amounts of multi-omics data to unveil shared biology between histologically-disparate cancers and opportunities for common drug targets.

Methods: A comprehensive, unsupervised machine learning investigation was performed using sample-matched WGS, RNA-sequencing, miRNA profiling, proteomics, and phospho-proteomics data of 238 pediatric brain tumors (consisting of seven different histologies) from the Open Pediatric Brain Tumor Atlas.

Results: Through unsupervised clustering, we identified four broad groups of brain cancers that share similar biological pathways and miRNA signatures irrespective of histology. The most aggressive group (consisting of Medulloblastoma, ATRT, High-Grade Astrocytic Tumors, and some Ependymomas) exhibited the shortest patient survival and displayed a marked increase in the expression of miR-17/92 cluster family of oncogenic miRNAs ($p<0.001$). Further exploration of the upstream regulatory mechanisms of miR-17/92 identified E2F1/2/3 significantly upregulated ($p<0.01$). Utilizing phospho-proteomics data, we found hyperphosphorylation of RB1, the upstream regulator of E2F, at three sites, providing additional evidence of RB1-E2F regulation upstream of MIR17HG. RB1 hyperphosphorylation and subsequent complete inactivation are mediated by the CDK2/Cyclin E complex, which is inhibited by the tumor suppressor p21. Interestingly, p21 (*CDKN1A*) gene expression was entirely depleted in the aggressive brain tumors group.

Conclusion: Overall, this study highlights a use case of advanced computational methods, like Machine Learning, to derive biologically-meaningful information from large amounts of multi-omics data. In our case, we demonstrate how shared biological pathways can be extracted from histologically-different tumors and novel regulatory mechanisms, such as that of miR-17/92 cluster, can be uncovered.

Keywords: Machine Learning, Big Data, Multi-omics, Pediatric brain cancers

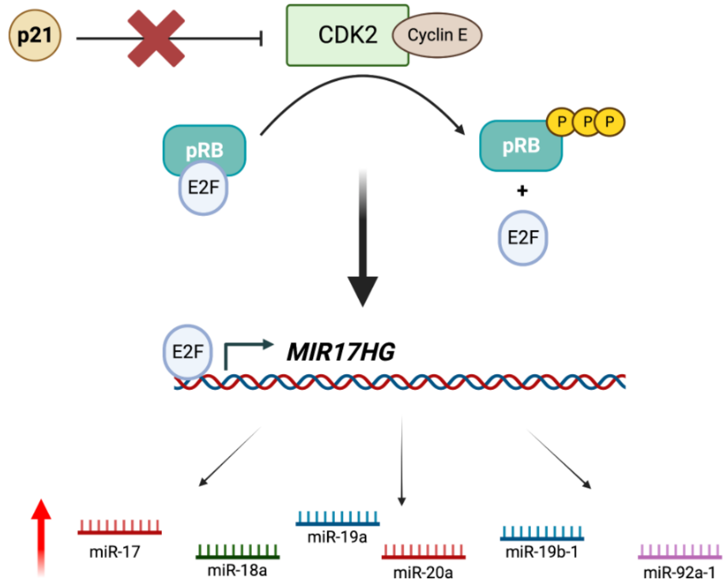

Figure 1. Novel pathway uncovered using Machine Learning and advanced computational analyses