Michael G. Hirezi
L2
LSU Health Sciences Center, New Orleans, LA

Chindo Hicks, PhD, Mentor, Bioinformatics and Genetics Program, Department of Genetics, LSUHSC

A Machine Learning Approach for Biomarker Discovery using Hepatocellular Carcinoma Transcriptome Data

Advances in high-throughput microarray and next-generation sequencing technologies have revolutionized genomic research on cancer. Large multi-center and multinational projects such as The Cancer Genome Atlas and the International Cancer Genome Consortium have conducted large scale sequencing of cancer genomes for the discovery of driver mutations, clinically actionable biomarkers and targets for the development of novel therapeutics. Molecular markers discovered from these studies are poised to improve clinical outcomes via precision oncology. However, while opportunities for understanding cancer biology and improving human health using genomics data have become unprecedented, the major bottleneck in genomic research with the aim of improving health is no longer data generation—the computational challenges around data analysis, display and integration are now the rate limiting factor. Of the possible answers to addressing this critical unmet need, Machine Learning (ML) algorithms and artificial intelligence offer the best prospects. The objective of this project is to employ ML algorithms for the discovery of prospective clinically actionable biomarkers in Liver Cancer. To achieve this objective, we used publicly available RNA-sequence (RNA-Seq) data from the National Cancer Institutes "The Cancer Genome Atlas". The project serves to characterize a RNA-seq bioinformatics pipeline capable of describing the unique genomic attributes observed within a machine learning analysis. Further questions are explored by harnessing the transcriptome data to obtain robust foundational knowledge about the structure and function of the liver cancer genome and about the genetic underpinnings of the disease. In summary, the work should be able to at least serve as a demonstration of the potential usefulness of machine learning methods in investigating the high volume of genomics data (and subsequently trends) that are inherent to next-generation sequencing data. Furthermore, the results of the work are considered within the context of Oncology, a promising field for the implementation of clinical diagnostics fueled by the management of omics data to advance precision oncology.