Molecular Classification of Pediatric High-Risk Leukemias Using Expression Profiles of Multimodally Expressed Genes

Sneha Jariwala[1], Alfred Geoffrey Lyle[1,2], Jacob Pfeil[2], Lauren Sanders[1,2], Holly Beale[1,2], Ellen Kephart[2], Katrina Learned[2], Allison Cheney[1,2], and Olena M. Vaske[1,2]
[1]Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, CA
[2]University of California Santa Cruz Genomics Institute, Santa Cruz, CA, USA

Introduction: Childhood leukemia is the most common type of cancer among children and teenagers. Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) account for the majority of childhood leukemia patients. While known markers for poor prognosis include higher age, higher white blood cell count at diagnosis, and certain translocations, gene expression analysis can indicate new prognostic factors that can be targeted in therapy.

Methods: In order to determine the gene expression signatures that characterize leukemia subtypes, we used a novel unsupervised analysis model called Hydra. Hydra uses multimodal gene expression to characterize clusters within cancer cohorts. The approach uses a Dirichlet process mixture model in order to detect multimodal genes and gene expression signatures. The output of the Hydra model identifies clusters of the cancer cohort, as well as enriched pathways and genes for further investigation. Differences among these clusters can be investigated by finding enriched pathways via Gene Set Enrichment Analysis (GSEA). The enriched pathways are specific to each cluster, which can be used in conjunction with data about event free survival and overall survival, to determine how certain pathways are associated with poor outcome. This analysis used publicly available data from the National Cancer Institute (NCI) Therapeutically Applicable Research to Generate Effective Treatments (TARGET) database that was uniformly processed and made available by the Treehouse Childhood Cancer Initiative. TARGET data was used as it provides thorough and detailed metadata about patient age, sex, fusion status, event free survival time, and overall survival time.

Results: To look more into each cluster, the enriched pathways provided more insight as to what gene expression patterns characterize the cluster and differentiate it from others. First, 202 fusion negative AML and fusion negative B-cell precursor ALL samples were run through Hydra and 5 clusters were identified. Since the Hydra program is an unsupervised model that solely creates clusters based on gene expression, one could infer that the clusters should be distinctive of AML or ALL samples. Clusters had different enriched pathways, such as high mitochondrial activity, high cell proliferation, and high cell signaling. Though these are characteristics of all cancer cells, each cluster demonstrated that one pathway was most distinctive of those samples compared to others. Most clusters mainly differentiated by disease, however, one cluster with enriched heme metabolism and immunoglobulin pathways contained almost equal amounts of AML and ALL samples, suggesting that specific cohorts of AML and ALL patients have this subtype of increased inflammatory response. Another cluster containing 72 AML samples and 4 ALL samples indicated a unique subtype of ALL samples that have molecular characteristics of AML.

Discussion: Application of Hydra to the analysis of pediatric high-risk leukemias may reveal novel expression-based disease subtypes. Future work will focus on characterizing these subtypes and assessing their therapeutic significance.