Title

Workflow development for GDC cancer data of AML and ALL samples

Authors

Olamide Adefioye[1] and Jonathon E. Mohl[1,2]

[1]Bioinformatics Program and [2]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Acute Myeloid Leukemia (AML) and Acute Lymphoid Leukemia (ALL) are two distinct hematological malignancies characterized by uncontrolled proliferation of undifferentiated myeloid and lymphoid cells, respectively. While phenotypic and cytogenetic methods are traditionally employed for classification and prognosis, molecular profiling offers deeper insights into the disease's biology, potentially unveiling novel therapeutic targets. This study sought to elucidate the differential gene expression patterns between different groups of AML and ALL patients, utilizing RNA sequencing (RNA-seq) data. The Genomic Data Commons (GDC) is an initiative of the National Cancer Institute (NCI) that provides a unified platform for sharing genomic, transcriptomic, and clinical data among cancer researchers. Our workflow approach was centered on downloaded sequencing data processed by the program "Spliced Transcripts Alignment to a Reference (STAR)" aligner obtained from the GDC repository. Python was used for the preliminary data cleaning, followed by differential expression analysis in R using DESeq2. This workflow is important because it can also be used to compare subgroups between AML or ALL or other types of cancer datasets from the GDC.