Tittle:

AI-driven analysis for exploring differences in Tumor Mutation Burden in Acute Lymphoid Leukemia in distinct ethnic groups

Authors

Maria G. Jimenez[1], Jonathon E. Mohl[1,2,3]

[1]Bioinformatics Program, [2]Department of Mathematical Sciences, [3]Border Biomedical Research Center, The University of Texas at El Paso, El Paso, TX

Abstract

Tumor Mutation Burden (TMB) is a key biomarker in cancer research, providing insights into genomic instability and immunotherapy efficacy. While extensively studied across cancer types, its variation across ethnic groups remains little explored, particularly in Acute Lymphoblastic Leukemia (ALL), a malignancy caused by uncontrolled proliferation of immature B or T lymphocyte precursors, where genomic alterations influence disease progression, prognosis, and treatment response. Understanding these differences may inform personalized medicine and improve outcomes. This AI-driven study uses Machine Learning (ML) and Deep Learning (DL) to evaluate whether TMB and mutational profiles vary across ethnic groups of ALL patients and to identify group-specific and shared frequently mutated genes. Clinical demographic patient data (TARGET-ALL-P2) were obtained from the GDC portal and preprocessed using OncoMiner and other Python-based pipelines, focusing on coding sequence (CDS) genes. TMB values were calculated as the number of mutations per total bases and normalized for cross-sample comparison. To address high-dimensional data, Principal Component Analysis (PCA) was applied to preserve informative components while minimizing noise. ML models and DL architectures were trained and evaluated using cross-validation and standard accuracy metrics. By combining predictive modeling with gene-level analysis, this study aims to uncover ethnic-specific differences and shared mutational patterns in ALL, providing insights for more equitable precision medicine strategies. Limitations include cohort size and representation, sequencing variability, and potential constraints of predictive models in capturing subtle or complex genomic features. Future work will extend analyses to larger cohorts, stratify by clinical subtypes, and integrate multi-omics data to elucidate biological mechanisms underlying observed differences.