**Project Title:**

"**Reproducible Pipeline for Immune Microenvironment Profiling in Pediatric Solid Tumors Using CRDC APIs and Local HPC** "

Abstract:

Understanding the tumor immune microenvironment (TIME) in pediatric cancers is critical for identifying biomarkers of immunotherapy response. While large public datasets such as TARGET, Kids First, and CCDI provide valuable genomic and clinical data, workflows for accessing and analyzing these resources are often inconsistent and not easily reproduced across systems.

This project aims to develop a reproducible workflow for immune profiling of pediatric solid tumors by combining CRDC API-based cohort selection with local high-performance computing (HPC) analysis on NIH’s Biowulf cluster. We will use the GDC API to identify tumor samples based on diagnosis, data type (e.g., RNA-seq), and clinical features. API-generated manifests will be used to select and download consistent expression and clinical data for local analysis.

We will perform immune profiling using tools such as Seurat, Harmony, and CellTypist to characterize immune cell types and states in tumor samples. These profiles will be compared with clinical features such as stage or treatment response. The full process—from cohort selection to analysis—will be documented using plain-text scripts or notebooks (e.g., R Markdown or Snakemake) that record all commands and parameters to ensure reproducibility.

This project will demonstrate a hybrid workflow that links cloud-based data discovery with local analysis, offering a practical model for reproducible and portable research methods. By testing this approach, we aim to provide useful examples and lessons that others can apply to reuse existing data in future cancer studies. The results will also help guide the development of the Immuno-Oncology Data Commons (IODC), an NIH effort to improve access to standardized immuno-oncology data and tools.