Developing reliable and scalable methods for deep immune phenotyping in public Childhood Cancer RNA-Seq Data repositories

Recent advancements in RNA-Seq technologies have significantly improved our ability to analyze individual transcriptomes, allowing for comprehensive gene expression characterization. While vast amounts of RNA-Seq data are publicly available, current analyses in immunology, particularly concerning childhood cancer, are limited. Conventional gene expression analysis often overlooks crucial information like ancestry, cell type composition, HLA type, killer cell Ig-like receptor (KIR) expression, and T/B Cell Receptor (TCR/BCR) repertoires. This missing data is vital for understanding immune responses and disease susceptibility across diverse populations.

Existing RNA-Seq and immunological databases fall short. For instance, some provide gene expression counts but lack essential immunological and ancestry data. Conversely, other immunological repositories offer detailed immune repertoire information but miss key immunological phenotypes and associated ancestry. This critical gap prevents comprehensive immunological studies that consider the variability in immune responses across different populations, which is especially important for understanding and treating complex diseases like childhood cancer.

To overcome these limitations, we propose developing advanced bioinformatics tools and a comprehensive database to infer and integrate critical immune phenotypes and ancestry information directly from RNA-Seq data. Our approach will enable more accurate and thorough analysis of immune-related diseases, including childhood cancers, across diverse populations. We will leverage public RNA-Seq samples from major public repositories, encompassing individuals of various ancestries. Our methods will be rigorously benchmarked on both simulated and real data to ensure their feasibility and reliability. This extensive and ethnically diverse dataset will allow us to identify significant differences in immune responses across various populations and disease conditions, providing crucial insights into health disparities.

We will develop robust and scalable methods to infer a rich set of immunological phenotypes from RNA-Seq data. This includes a highly accurate HLA typing tool that uses a pan-genome reference to minimize ancestral bias and ensure comprehensive allele identification. We will also create a tool to infer individual Adaptive Immune Receptor Repertoires (AIRR) alleles directly from RNA-Seq data, with the potential to discover novel alleles. Our existing method for T and B cell receptor assembly will be enhanced, utilizing paired-end read information for improved precision in V(D)J recombination inferences and clonotype assembly. Furthermore, for more accurate cell type composition analysis, we will introduce a consensus-based method that synthesizes results from multiple transcriptomics-based deconvolution techniques. Each of these tools will undergo rigorous validation to ensure their effectiveness for deep immune phenotyping from RNA-Seq data.

The insights gained from these methods will be disseminated through a novel, user-friendly database. This platform will be the largest collection of individuals with detailed immunological phenotypes across diverse ethnic backgrounds and disease conditions, providing the biomedical community with the tools to tackle key questions in medical immunology and work towards reducing health disparities. Unlike existing repositories, our platfrom will offer a comprehensive set of functionalities, including options for sample normalization and meta-analysis, and will be accessible via an intuitive R package, graphical user interface, and application programming interface. We will also address ethical and security issues related to genotyping and phenotyping analyses, ensuring privacy and responsible data distribution. This initiative aims to promote access and reuse of pediatric cancer data and foster interdisciplinary collaborations.