Predicting Immune Cell Admixture in Bulk Proteomics Data Using ProteoMixture

Joshua P Schaaf[1]; Tamara Abulez[1]; Pang-ning Teng[1]; Brian L. Hood[1]; Kelly A. Conrads[1]; Litzi J. Tracy[1]; Allison L. Hunt[2]; Matthew D. Wilkerson[3]; Kathleen M. Darcy[1]; Neil Phippen[1]; Christopher M. Tarney[1]; Larry Maxwell[1]; Thomas P. Conrads[2]; Nicholas W. Bateman[1]
[1]Gynecologic Cancer Center of Excellence and the Women's Health Integrated Research Center, Annandale, VA; [2]Women's Health Integrated Research Center, Women's Service Line, Inova Health System, Annandale, VA; [3]The American Genome Center, Department of Anatomy Physiology and Genetics, Uniformed Services University of the Health Sciences, Bethesda, MD, United States, Bethesda, MD

Introduction

Immune cell admixture within the tumor microenvironment is correlated with cancer prognosis. Current tools for predicting immune cell subpopulations in bulk expression data, such as xCell and CIBERSORT, are developed and optimized using transcript-level measurements. We recently developed the ProteoMixture tool to support prediction of tumor, stroma, and immune cell admixture in bulk tissue proteomics data. Here we describe developmental efforts to refine ProteoMixture immune scores, enabling prediction of relative immune cell sub-population signatures within proteomics data obtained from bulk tissue specimens. We prioritize optimized protein signatures from a pan-cancer cohort of >1000 patient tumors exhibiting diverse immune cell admixtures predicted by transcript-level data and explore signature performance in bulk tissue proteome data from independent patient cohorts.

Methods

Global proteome and transcriptome data from ten cancer types (BRCA, ccRCC, COAD, GBM, HNSCC, LSCC, LUAD, OV, PDAC, and UCEC) harmonized from the Baylor College of Medicine were downloaded using the Python cptac package. Duplicate proteins were consolidated by calculating median abundance, samples were median centered, and proteins with missing data were excluded. xCell was run in R using the GSVA and xCell libraries. Proteome signatures were generated in Python by selecting proteins with high Spearman correlations to cellular subpopulation scores predicted using xCell, keeping protein-level features that optimally predict cellular subpopulations of interest. Protein signatures were then validated in bulk tissue proteome data for independent ovarian (APOLLO-OV, ZhangAW-HGSOC, PMID: 38480868) and lung (APOLLO-LUAD, PMID: 36384096) cancer cohorts.

Preliminary Data

Proteome-informed geneset scores from bulk tissue proteomic data significantly correlated with xCell transcript scores for diverse cellular subpopulations including immune cell subtypes (Benjamini-Hochberg (BH) Spearman $p < 0.05$ and exhibited significantly higher (paired Wilcoxon $p < 0.001$) correlation coefficients ($\rho = 0.79 \pm 0.08$) compared to transcript-based genesets assessed on proteomic data alone ($\rho = 0.45 \pm 0.27$) across all CPTAC samples. Proteome-informed genesets (22±28) relative to transcript genesets (155±93) were significantly smaller (paired Wilcoxon $p < 0.001$). Despite smaller genesets, proteome-informed geneset accuracy did not correlate with geneset size in CPTAC nor independent validation cohorts

(Spearman $p > 0.05$). Comparison of proteome-informed and xCell transcript genesets revealed significantly higher gene-wise RNA to protein spearman correlations (GRPC) between overlapping candidates than protein-level candidates alone across all tumor samples in CPTAC and APOLLO-LUAD (Mann-Whitney $p < 0.001$). Geneset correlation trends in CPTAC OV were replicated in the APOLLO OV validation cohort, with strong geneset performance trends showing high correlation between these datasets ($\rho = 0.80$, $p < 0.001$). A total of 35 proteome-informed immune cell subtype genesets were significantly more predictive of different cellular subpopulations within bulk tissue proteomics data across cancer types than xCell genesets alone (BH adjusted MWU $p < 0.05$). Both CD8+ T cell and CD4+ T cell proteome-informed geneset enrichment scores in the ZhangAW-HGSOC samples were positively correlated with CD8 and CD4 T cell populations quantified by immunohistochemistry ($\rho = 0.73$ and $\rho = 0.60$, $p < 0.001$).

Novel Aspect

The ProteoMixture software tool has been refined to accurately predict immune cell subpopulations from bulk proteomics data.