Deconvolution of bulk proteomics using single-cell RNA-Seq

Ahmed Youssef [1,4], Indranil Paul [2,4], Mark Crovella [3], Andrew Emili [2,4]

[1]Bioinformatics Program, Boston University, Boston, MA, [2]Department of Biochemistry, Boston University School of Medicine, Boston, MA,
[3]Department of Computer Science, Boston University, Boston, MA, [4] Center for Network Systems Biology, Boston University School of Medicine, Boston, MA

INTRODUCTION

- Proteins perform the majority of essential biological processes governing cellular functions, yet the proteome remains largely unexplored at the resolution of single cells.

- Existing multi-omics studies that combine RNA-Seq and proteomics data are confounded by weak RNA-protein correlations.

- We developed a novel deconvolution algorithm that combines single-cell RNA-seq with standard proteomics data to model the proteome at the single-cell level.

- Our approach does not rely on direct correlations between RNA and protein levels and provides a genralizable framework for bridging the gap between bulk and single-cell omics layers.

METHODS

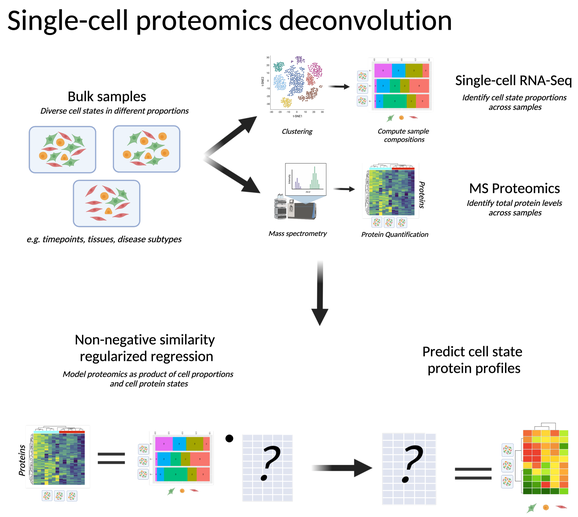
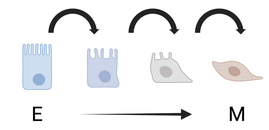


Fig. 1) Overview of algorithm for using single-cell RNA-Seq data to deconvolute bulk proteomics data. Paired proteomics and scRNA-Seq datasets are generated using established techniques, or extracted from public sources, for a set of samples examining a dynamic biological system. The algorithm uses scRNA-Seq data to identify the cell populations in the data and their relative proportions within each sample. These proportions are then used to deconvolute the bulk proteomics measurements to uncover the contribution of each cell population to each protein's observed bulk measurement using a regression-based framework.

RESULTS

Case study: EMT Multi-omics Data



Multi-omics data generated across 8 timepoints for cells undergoing EMT:

- Proteomics: 6,967 proteins quantified

- scRNA-Seq: 9,785 genes across 1,913 cells

Classifying cell state composition using scRNA

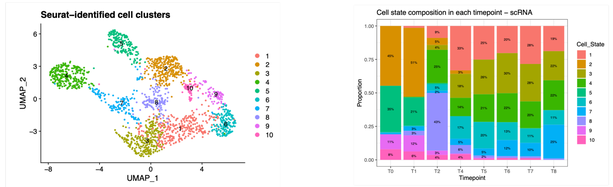
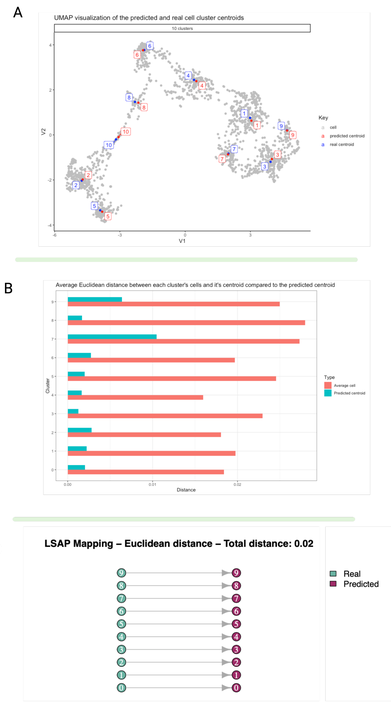


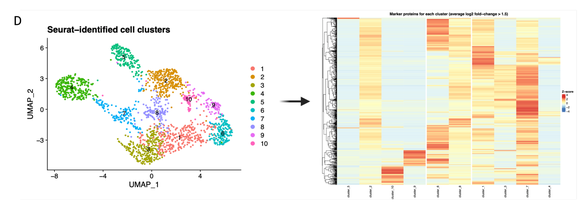
Fig. 2) Top: Overview of experiment. Cells in the scRNA-Seq dataset were pooled across timepoints and clustered using the Seurat R package. Left: UMAP plot visualizing the 10 distinct cell states identified using Seurat. Right: Bar plot visualizing proportions of cell states across the timepoints.

Right: Benchmarking the algorithm on pseudobulk data created by summing the scRNA data. Using the pseudobulk allows for the mathematical validation of the algorithm since the predictions can be evaluated against the true scRNA profiles. A) UMAP visualizing the overlap between the real and predicted centroids for each cell cluster. B) Bar plots showing that our predicted centroids fall within the correct intra-cluster distances from the ground truth data. C) Using linear sum assignment problem to quantify the distance between the real and predicted centroids.

Benchmarking the algorithm



Deconvoluting proteomics using scRNA



Recovery of canonical EMT markers

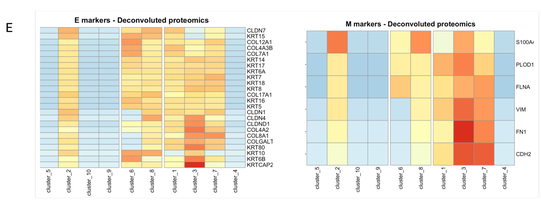
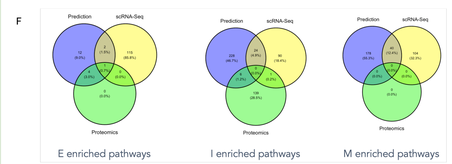


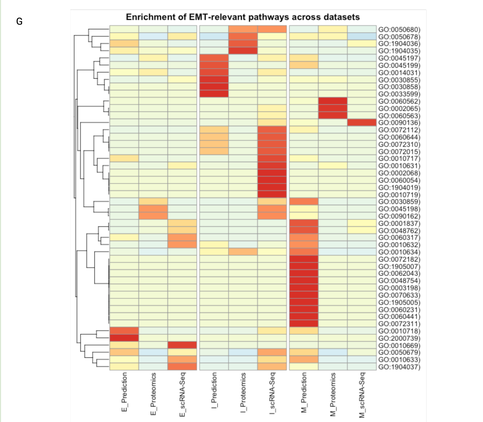
Fig. 3) Top: Results of applying our algorithm to the EMT proteomics data. D) Our algorithm identified the relative protein levels, i.e. proteome, for each of the scRNA-defined clusters. E) Heatmaps showing the predicted protein expression levels of canonical epithelial (E) and mesenchymal (M) markers.

Right: Comparing information gleaned from our predicted single-cell proteomes to that from the raw datasets. F) Overlap of enriched pathways among differential markers of each EMT stage. G) Relative enrichment scores of key EMT-related pathways across datasets

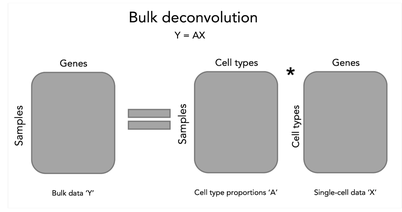
Prediction of novel enriched pathways



Differential enrichment of key pathways across datasets



ALGORITHM



- *Premise*: Model bulk data as product of cell type expression profiles and cell type proportions
- *Method*: Use regression-based approach to predict single-cell matrix *X* given bulk omics matrix *Y* and mixture matrix *A*
- *Output*: Map bulk data onto the cell types defined in the mixture matrix

CONCLUSION

- We applied our algorithm to cell differentiation datasets and demonstrate its ability to accurately reconstruct single-cell profiles from bulk-level measurements at both the proteome and transcriptome levels.

- Our approach is able to successfully cross the protein-RNA divide by using scRNA-Seq in combination with bulk proteomics to distinguish both canonical and novel protein markers.

- Our method offers researchers the abilty to study the proteome at the single-cell level by integrating proteomics into the single-cell analysis toolkit.

FUTURE WORK

- Extend framework to predict cell state-specific protein-protein interactions

- Application to large-scale single-cell and proteomics reference databases including the Human Cell Atlas

- Develop R package and interactive visualization tool

Scan the QR code to reach GitHub for source code, more details, and updates: