Deconvolution of proteomics data using scRNA-Seq

Multi-omics tracking of cells undergoing EMT

Ahmed Youssef

May 16, 2022

Introduction

The fundamental unit of all living organisms is the cell, and recent technological advances have granted us unprecedented opportunities to study life at this principal level. Proteins perform the majority of vital biological processes governing cellular functions, yet the proteome remains largely unexplored at the resolution of single cells, representing crucial gaps in our understanding of cellular complexity. Here, we present a novel deconvolution algorithm that combines single-cell RNA-sequencing (scRNA-Seq) with bulk proteomics to model the global proteome at the single-cell level. Our approach leverages cell profile similarities to overcome the weak correlation between RNA-Seq and proteomics that confounds existing deconvolution strategies. We apply 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. Furthermore, we show that our algorithm is able to successfully cross the protein-RNA divide by using scRNA-Seq in combination with bulk proteomics to distinguish established canonical markers. Our method provides a generalizable computational framework for charting the relationship between bulk and single-cell molecular layers, and offers researchers the ability to study the proteome at the single-cell level using established bulk-proteomics workflows. This work also lays the foundation for transferring cell-state information between RNA and protein modalities, integrating the under-served layer of proteomics into the single-cell analysis toolkit to enhance the prioritization of cell populations for targeted therapeutics.

Case Study - EMT Multi-Omics Experiment

Epithelial-to-mesenchymal transition (EMT) is a biological process in which epithelial cells gradually lose their adhesion and transition into mesenchymal cells. As one of the hallmarks of cancer progression, it is one of the long-standing interests of the biomedical research community. Towards profiling this process, protein and RNA samples were extracted from cells at 8 different timepoints during EMT and multiple layers of omics data were generated. These omics layers include proteomics, transcriptomics, phosphoproteomics, secretome, exosome among others. A pre-print with more details on the experiment and generated data can be found on bioRxiv here (Paul et al, 2021). This report is interested in the scRNA-Seq and proteomics datasets generated in this study.

Rationale

Bulk proteomics data gives a view of the aggregated protein abundance from all cell types within a sequenced sample. Using single-cell data, derived from the same samples, we can investigate the sample heterogeneity by estimating proportions of cell types within the bulk sample. We cannot reliably use these proportions to directly estimate the contribution of each population to each gene/protein's expression at the bulk-level

however, since there is low correlation between RNA and protein levels of the same genes due to multiple biological and technical factors, such as alternative splicing and post-translational modifications. Leveraging the timepoints present in this dataset, which conveniently show shifts in cell state proportions across time, we can instead combine changes in cell state proportions with the corresponding changes in bulk-level protein abundance as suggestive of relationships between specific cell states and protein levels. This information can then be used to estimate the contribution of individual cell states to the bulk proteomics measurements.

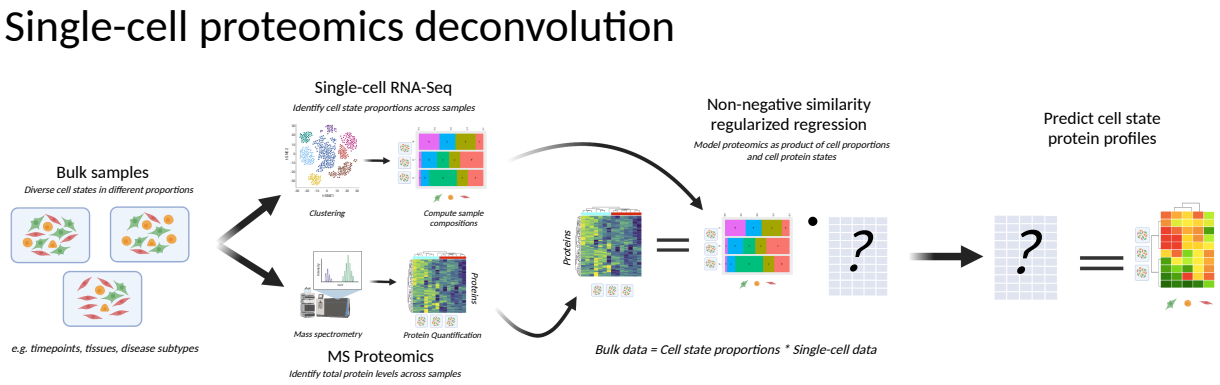


Figure 1: Simplified schematic of single-cell deconvolution model

Methods

Background

In our problem, we have $AX = Y$, where

- A is timepoints $\times$ clusters
- X is cluster centroids, meaning clusters $\times$ genes
- Y are bulk timepoint data, meaning timepoints $\times$ genes

We seek to deconvolve Y to find X:

$$\hat{X} = \min_X \|AX - Y\|_F^2$$

In general we expect to have less timepoints than clusters. Hence this problem is highly underdetermined, and we need additional information to guide us to a good solution.

Strategy

Our problem involves matrices $AX = Y$, but now let's use work one gene at a time (one column of X and the corresponding column of Y).

Nonnegativity

The first constraint we add is nonnegativity. This leads to the nonnegative least squares problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\geq 0} \|A\mathbf{x} - \mathbf{y}\|^2.$$

Algebraically this is the same as solving the quadratic program:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\geq 0} -2\mathbf{y}^T A\mathbf{x} + \mathbf{x}^T A^T A\mathbf{x}.$$

ℓ_2 Regularization

The next constraint we add is ℓ_2 regularization (ridge regression):

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\geq 0} \|A\mathbf{x} - \mathbf{y}\|^2 + \lambda_{\mathbf{x}}\|\mathbf{x}\|^2$$

We use the notation $\lambda_{\mathbf{x}}$ to indicate that the proper amount of regularization depends on the properties of $\mathbf{x}$ (which we will exploit).

This is equivalent to the quadratic program:

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\min_{\mathbf{x}\geq 0} -2\mathbf{y}^T A\mathbf{x} + \mathbf{x}^T A^T A\mathbf{x} + \lambda_{\mathbf{x}}\mathbf{x}^T\mathbf{x} \\ &= \arg\min_{\mathbf{x}\geq 0} -2\mathbf{y}^T A\mathbf{x} + \mathbf{x}^T(A^T A + \lambda_{\mathbf{x}} I)\mathbf{x}\end{aligned}$$

Similarity Regularization

Next we make the following observation: the process used to construct A may have access to additional information. For example, if we use scRNA clusters to define the cell types leading to the cluster mixtures in A, we can make use of the cluster RNA expression profiles.

We can use any similarity or dissimilarity matrix. For concreteness, assume we have matrix C which consists of cluster centroids; ie, it has clusters on the rows and genes on the columns, and entries in C correspond to average gene expression in the cluster. (Note that if the original scRNA data has missing values, this must be considered in forming cluster centroids).

Perform standard normalization on the rows of C, so that rows have zero mean and unit norm, to obtain $\tilde{C}$. Then form the correlation matrix of clusters,

$$M = \tilde{C}\tilde{C}^T$$

Now for a given gene $\mathbf{x}$, we seek to minimize $\mathbf{x}^T M^{-1}\mathbf{x}$. This will tend to make entries in $\mathbf{x}$ close if the corresponding clusters have high similarity.

So to include cluster similarity as a form of regularization, we minimize:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\geq 0} \|A\mathbf{x} - \mathbf{y}\|^2 + \lambda_{\mathbf{x}}\|\mathbf{x}\|^2 + \beta\mathbf{x}^T M^{-1}\mathbf{x}$$

which is equivalent to the quadratic program:

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\min_{\mathbf{x}\geq 0} -2\mathbf{y}^T A\mathbf{x} + \mathbf{x}^T(A^T A + \lambda_{\mathbf{x}} I)\mathbf{x} + \beta\mathbf{x}^T M^{-1}\mathbf{x} \\ &= \arg\min_{\mathbf{x}\geq 0} -2\mathbf{y}^T A\mathbf{x} + \mathbf{x}^T(A^T A + \lambda_{\mathbf{x}} I + \beta M^{-1})\mathbf{x}\end{aligned}$$

Implementation

The R function `solve.QP` can solve any of the above quadratic programs.

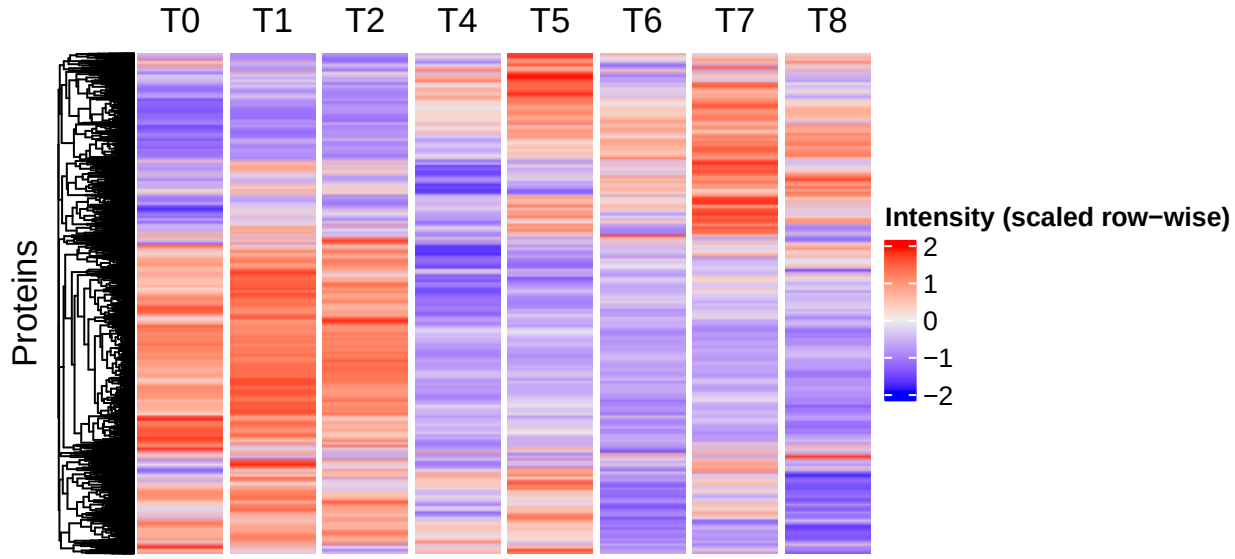
So, again working one gene at a time, we apply `solve.QP` solve the constrained minimization above to estimate the expression profile at the cell type level.

Data summary - Proteomics

The bulk proteomics data was generated in the Emili Lab using standard mass-spectrometry. Summary of the dataset follows:

- 6,967 proteins
- 10 different timepoints
- Three replicates

The average intensity across replicates was computed for each protein in each timepoint. Timepoints 3 and 9 were removed since they are not present in the scRNA data.



Data summary - scRNA-Seq

The scRNA-Seq data was generated in the Emili Lab using standard mass-spectrometry. Summary of the dataset follows:

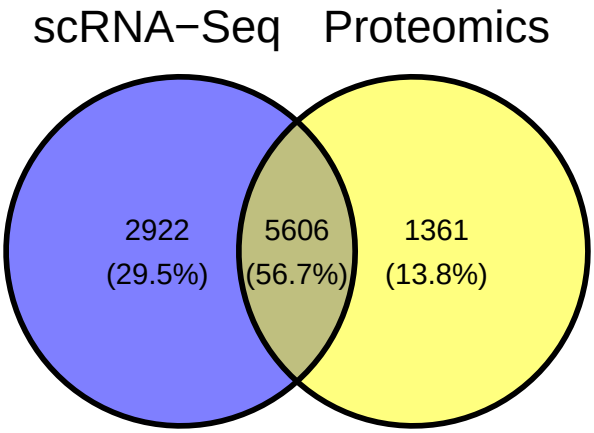
- 9,785 genes

- 1,913 cells (~200 cells per timepoint)
- 8 different timepoints

Prior to this summation, genes with zero variance as well as those with non-zero counts in less than 5% of all cells were removed. This removed 17 genes (0.2% of all genes). The data was also normalized such that each cell sums to 1.

Protein overlap

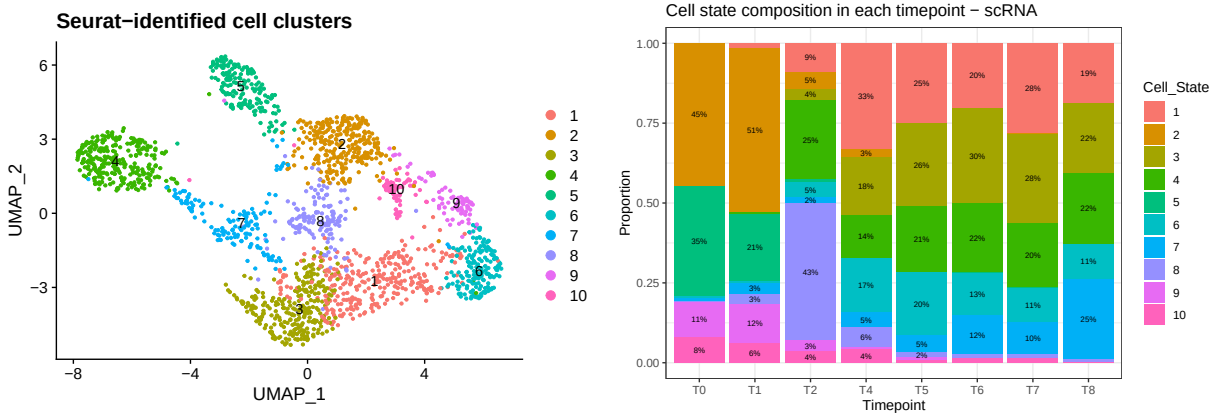
The venn diagram below shows the overlap of the identified proteins in the datasets. Only the genes that overlap between the two datasets are retained for downstream analysis.



Clustering scRNA data

The scRNA data is clustered in an unsupervised manner based on similarity of gene expression profiles using *Seurat* into 10 clusters. All the 420 cells across the 8 timepoints were pooled together for this clustering. The plot on the right shows the proportion of each cell cluster in each timepoint.

Note: this approach assumes the single-cell data accurately captures the sample heterogeneity. In practice, biased cell sampling upstream could lead to an inaccurate view of sample heterogeneity here.



Validating the approach using pseudobulk data

Prior to making inferences from the proteomics data, we first investigate the ability to recover the scRNA data from the bulk data at the RNA-level where we have the true single-cell profiles to compare against. The underlying principle of our model is that the bulk data is the summation of the single-cell data, which can be represented using the simple formula $Bulk = Number\_of\_cells * Single\_cell\_expression$, for which we will use the notation $Y = AX$ throughout this report. The figure below shows a graphical representation of this model.

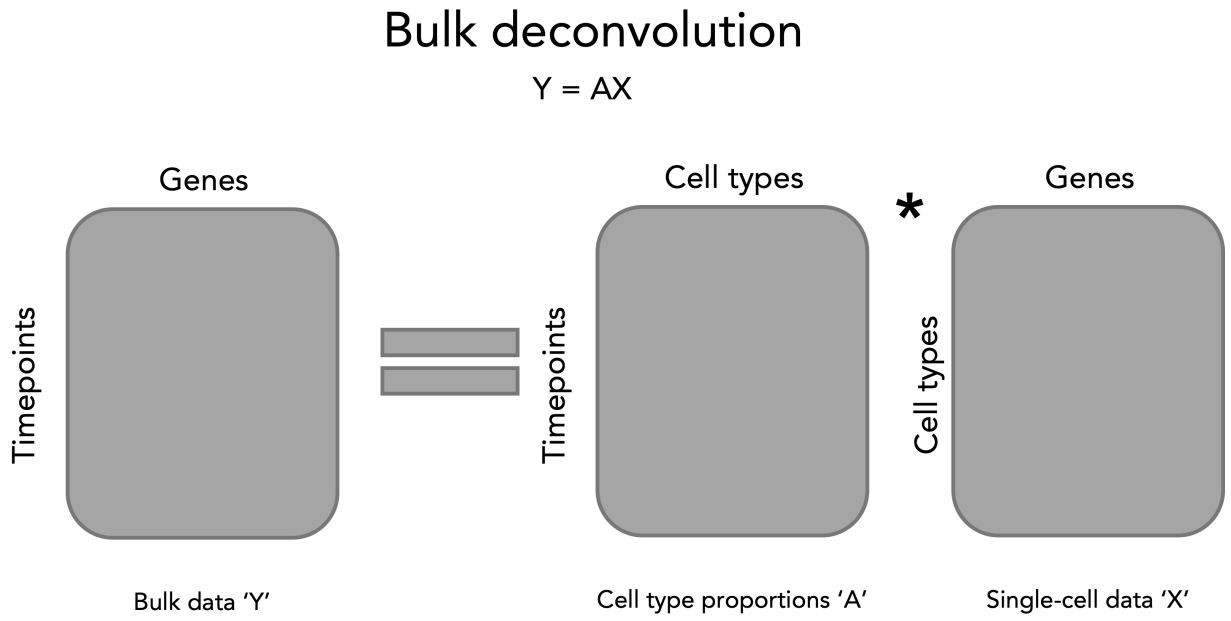


Figure 2: Schematic of single-cell deconvolution model

The process to to test our method on pseudobulk RNA data is detailed below:

1) ***Clustering***: The cell states in our dataset are identified in an unsupervised manner based on similarity of gene expression profiles. All cells from all timepoints are pooled together for this analysis. For data pre-processing, we remove the genes with low expression counts, retaining genes with a minimum of 3 counts in at least 3 cells. This removed 1,240 genes (13% of all genes). On average, each cell expressed ~3,600 genes after processing. Seurat is then used to cluster the cells with their default workflow based on the 2,000 most variable genes.

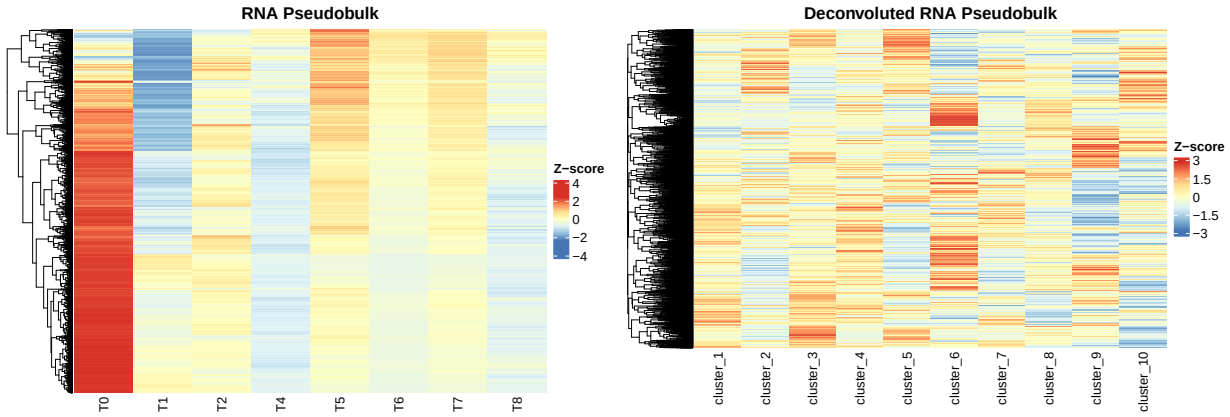
2) ***Construct cell type proportions matrix A***: The timepoint * cluster mixing matrix A is constructed by counting the numbers of cell from each cluster in each timepoint.

3) ***Construct cell cluster matrix X***: The cluster * gene matrix X is constructed by averaging the gene expression of each cluster.

4) ***Create pseudo-bulk matrix Y***: Construct timepoint * gene pseudobulk matrix Y using the formula $Y = AX$.

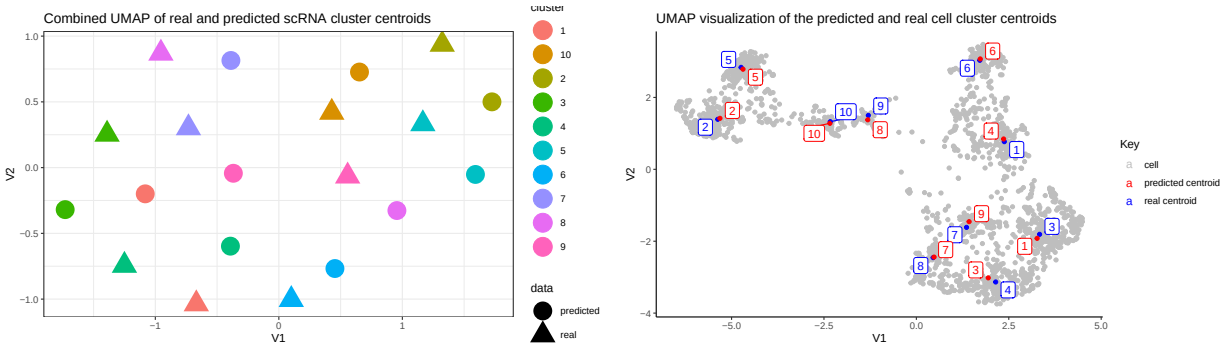
5) ***Predicting single-cell profiles***: Re-create the single-cell data from the pseudo-bulk data Y using the method detailed in the methods section. To decide on the optimal value for the parameters λ and β, we tested a range of 21 values between 1^{-10} to 1^{10} for each gene and each number of clusters. For each value of λ and β, we computed the relative error in estimating each gene's single-cell profile as a measure of the accuracy of the predicted single-cell profile. The λ and β pairing that lead to the minimal error was selected as the optimal value.

Deconvolution results



Comparing predicted cluster centroids to real ones

The below UMAP plots include the centroid of each cluster in each dataset. The centroid is computed as the average abundance of each protein in each cluster's cells.

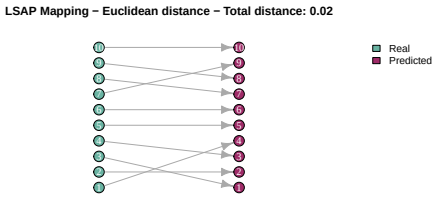


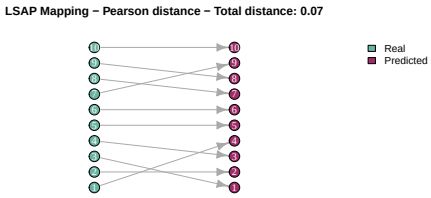
Mapping predicted clusters to real clusters - LSAP

In this section, we map the cluster centroids by solving the linear sum assignment problem (*LSAP*) using the Hungarian method as implemented in the `solve_LSAP()` function in the *clue* R package. In summary, the LSAP algorithm expects similarities between cluster centroids as the entries in the input matrix. The idea of the matching is that given two sets A and B, we find the matching that maximizes $sum(similarity(a_i, b_j))$ where a_i is matched with b_j. *Each member of the set is matched with exactly one member of the other set.*

The algorithm is run twice:

1) Using the Euclidean distance as the similarity metric for the cluster centroids, where the algorithm will minimize the sum of assigned euclidean distances.
2) Using the Pearson correlation as the similarity metric for the cluster centroids. The correlations are converted to distances to become non-negative, and the algorithm will minimize the sum of assigned distances


LSAP Mapping − Euclidean distance − Total distance: 0.02
Real
Predicted


LSAP Mapping − Pearson distance − Total distance: 0.07
Real
Predicted

Testing method with single-cell proteomics data

To aid as an external benchmark of our method, the Nikolai Slavov group at Northeastern University shared with us a MS-derived single-cell proteomics dataset from cells undergoing EMT. Their experimental setup consists of three timepoints (days 0, 3, and 9), and the data has 1,827 genes x 420 cells, out of which 1,064 genes, including 27 of the EMT hallmark genes, were also detected in our own data. For all downstream analysis, only the set of ~1,000 genes that overlap between the single-cell proteomics, bulk proteomics, and scRNA-Seq datasets are retained. Notably, 65% of the original single-cell proteomics matrix consists of missing values.

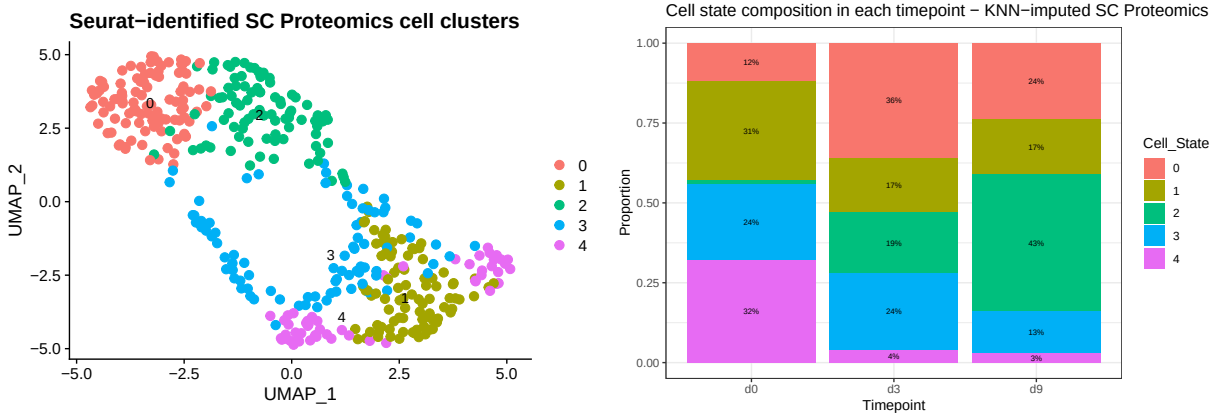
Data Summary:

- 1,776 proteins
- 420 cells
- 205 proteins detected in every single cell
- 605 proteins detected in each cell on average (Range: 528 - 702)
- Each protein is present in 145 cells on average with a median of 49 cells (IQR 11 - 315 cells)

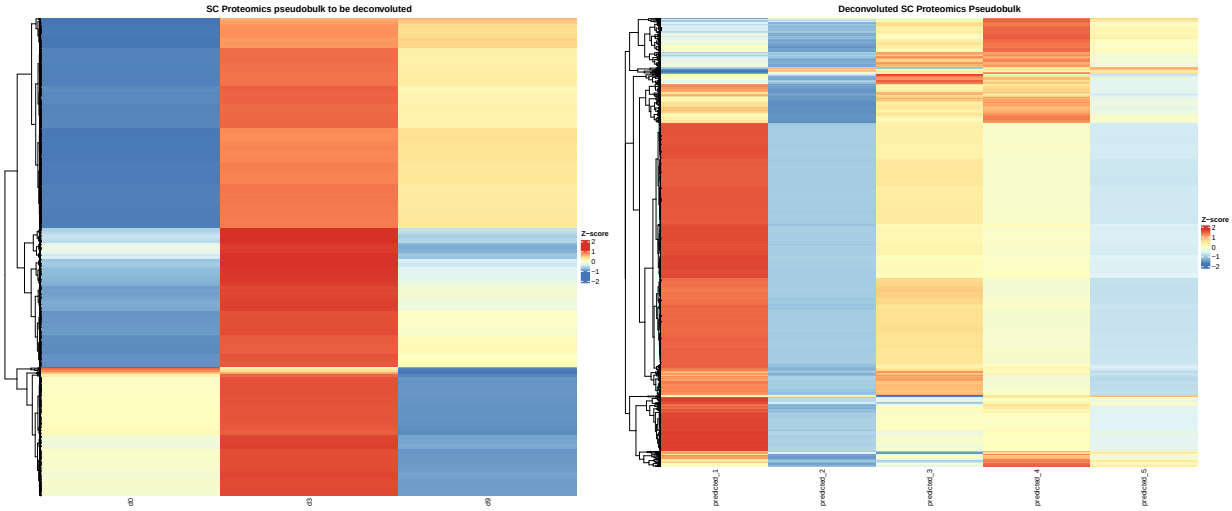
KNN Imputation

The Specht et al publication proposed a strategy to impute the missing values by k-nearest neighbor imputation (k = 3) using Euclidean distance as a similarity measure between the cells. Briefly, for a given missing value, the expression of that gene in that cell is taken as the mean of it's expression in the 3 most similar cells for which it's expression is present. The key parameter for this KNN algorithm is the number of neighbors K. We select the K based on a five-fold cross-validation within the training dataset and computing the resultant mean absolute prediction error on the held out non-missing values. The optimal value for K was shown to be 10.

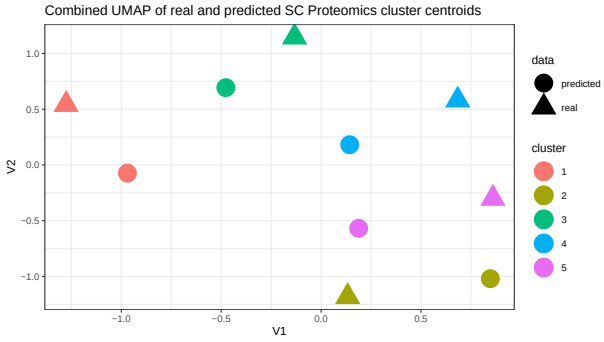
Clustering

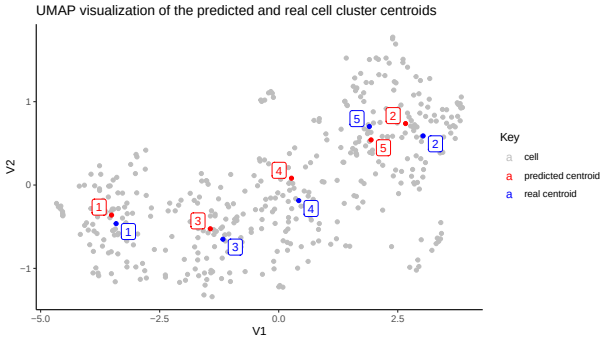


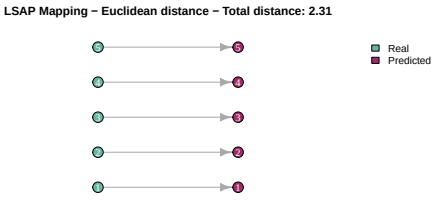
Deconvolution results

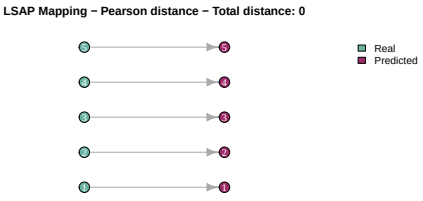


Comparing predicted and real centroids







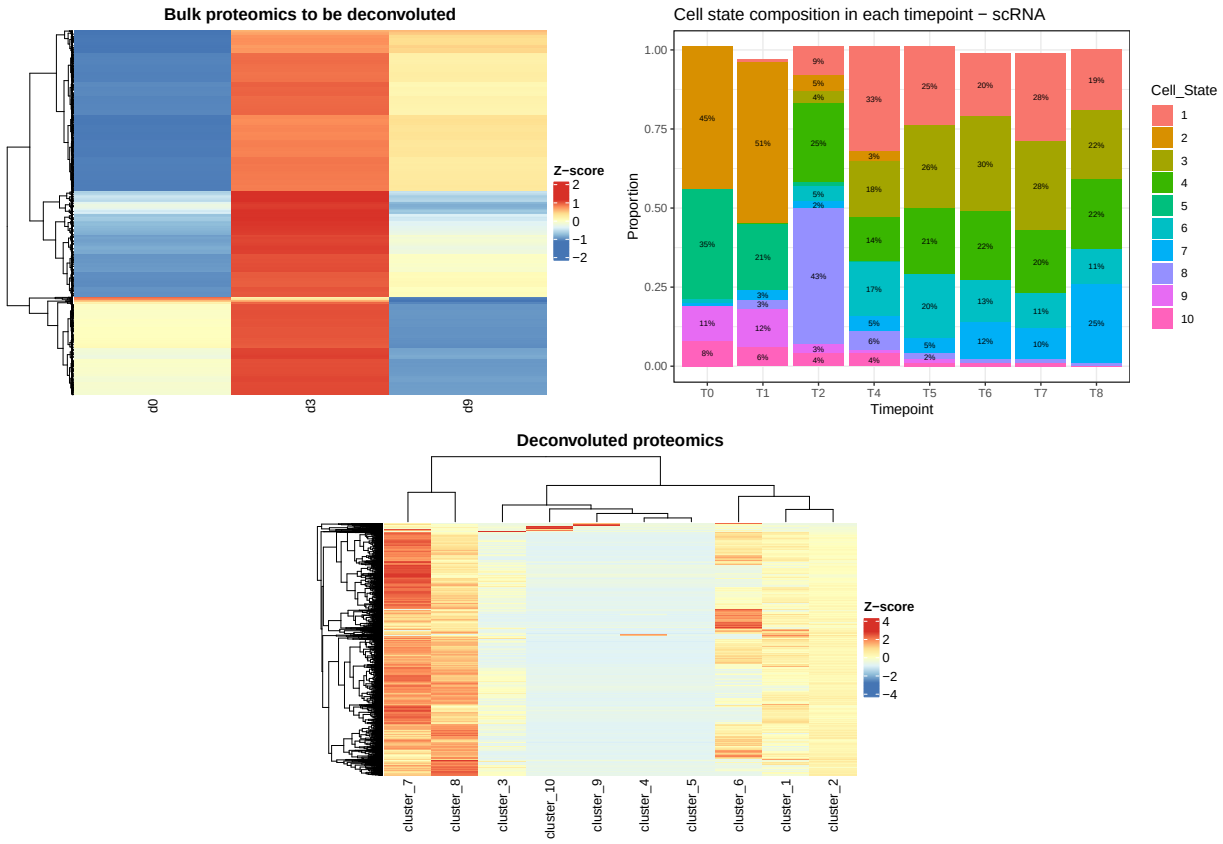


Deconvolution of bulk proteomics using scRNA-defined cell clusters

Now that we've validated our algorithm using the pseudobulk data, we applied our novel deconvolution algorithm to the bulk proteomics data with the scRNA-defined cell proportions matrix defined in the previous sections, setting the parameters λ and β to 1^{-7} and 1^{-4} respectively based on tuning results obtained from deconvoluting pseudobulk data.

Deconvolution results

Note: The deconvoluted proteomics heatmap below was first normalized by dividing each cluster's values by the number of cells in that cluster, followed by cross-cluster scaling.

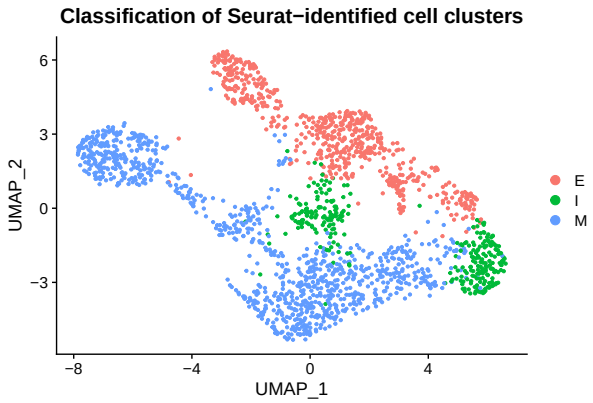


Classification of cell clusters

Each of the 10 clusters identified in the previous step are classified into one of three categories based on their change in proportions across time: Epithelial (E), Mesenchymal (M), or Intermediate (I). The clusters with their maximum count in the first 2 timepoints were labeled *E*, and those in the last 3 timepoints were labeled *M*, with the rest being *I*.

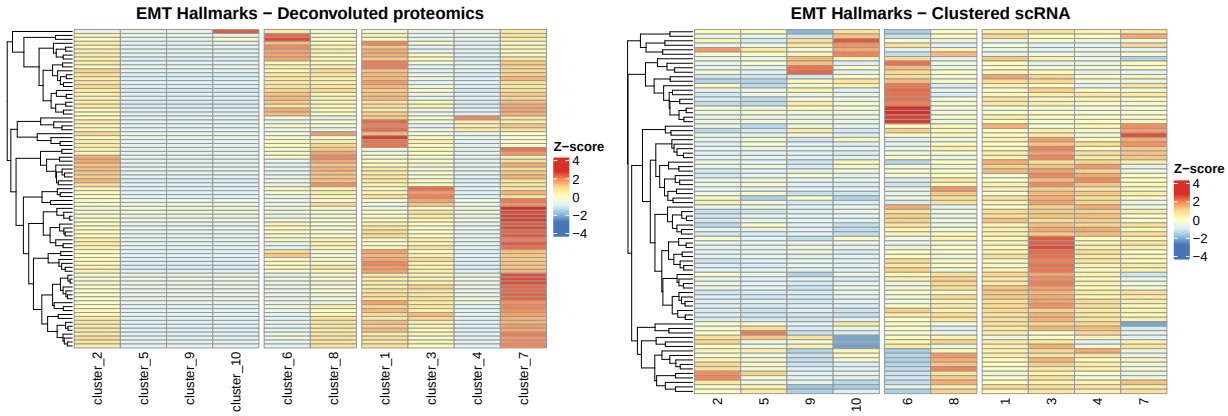
The following classifications were assigned:

- Epithelial clusters: 2, 5, 9, 10
- Intermediate clusters: 6, 8
- Mesenchymal clusters: 1, 3, 4, 7



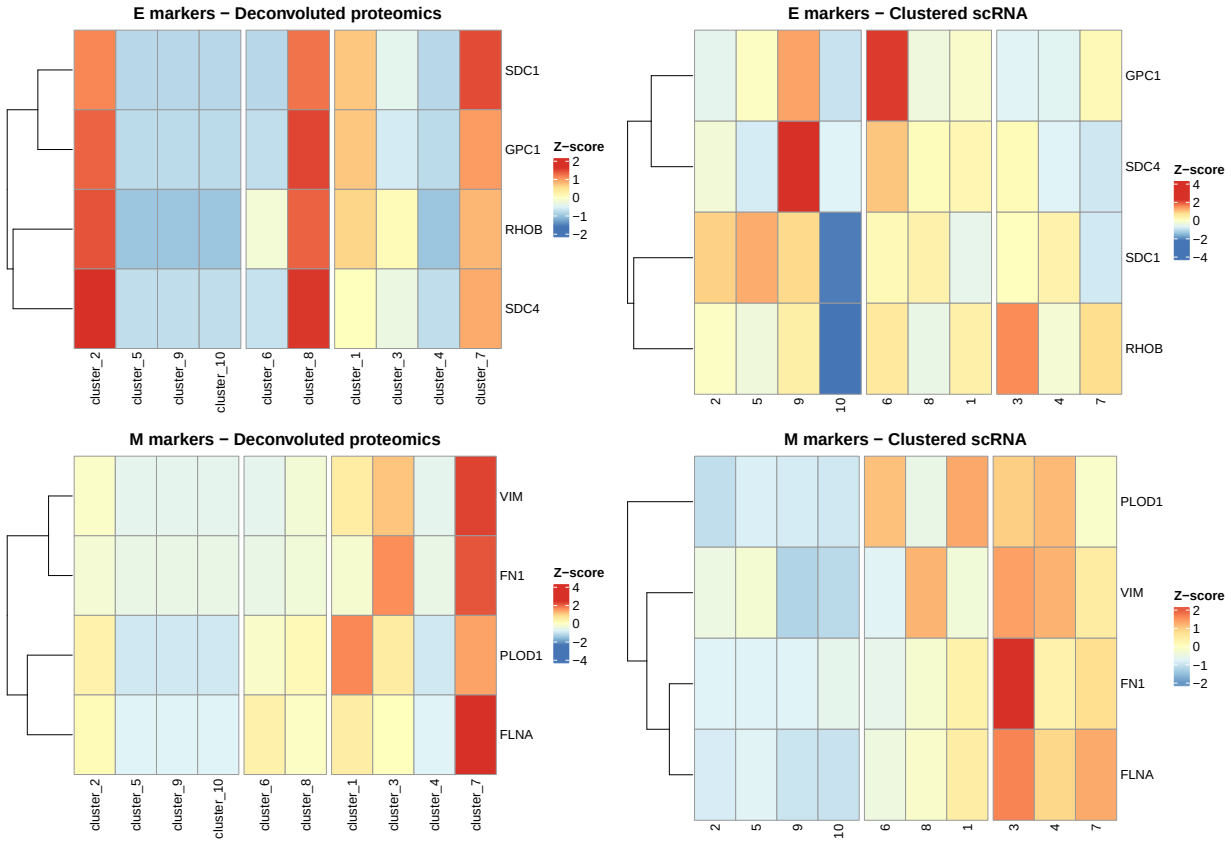
EMT hallmark genes

We focus on a set of 81 hallmark genes associated with EMT from the MSigDB database. The heatmaps are divided into 3 sections: E, I, and M. Note: The deconvoluted proteomics below was first normalized by dividing each cluster's values by the number of cells in that cluster, followed by cross-cluster scaling.



Focused set of known E/M markers

The below heatmaps show the expression patterns for a smaller subset of 4 known E markers and 4 known M markers in both the original single-cell RNA and the deconvoluted single-cell proteomics data. Note: no normalization was applied to the data.

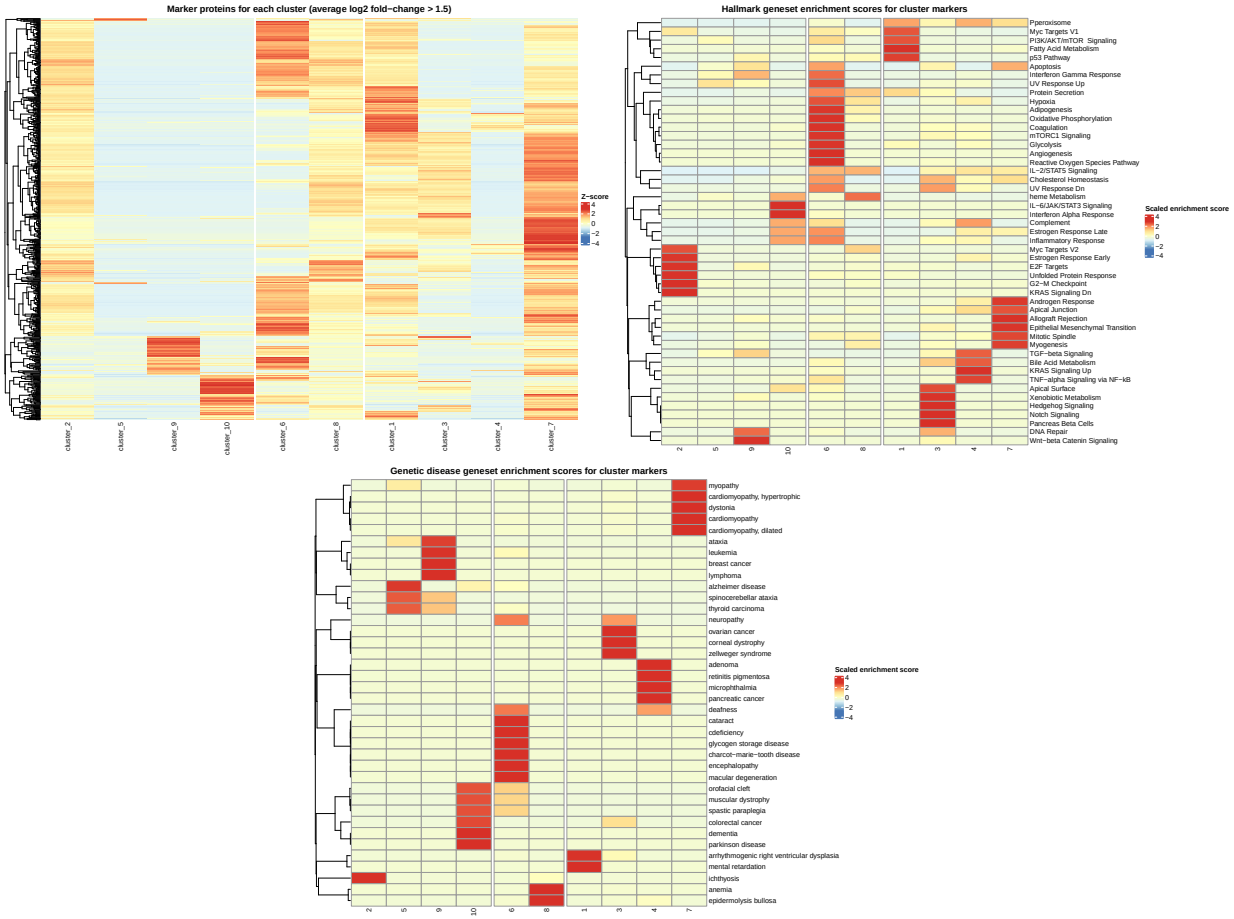


Pathway enrichment across predicted cluster centroids

For each cluster, we extract the proteins with a log2 fold-change of at least 1.5 between the cluster and all the other ones, as the set of markers for the cluster in the predicted single-cell proteomics data. The number of markers per cluster varied from 3 (cluster 2) to 198 (cluster 3). We then use the *enrichR* tool to compute the enrichment score for each cluster's markers across the set of MSigDB hallmark gene sets.

Table 1: Number of differential proteins for each cluster (average log2 fold-change > 1.5)

Cluster	Markers	Classification
1	15	M
2	3	E
3	198	M
4	84	M
5	44	E
6	185	I
7	19	M
8	14	I
9	64	E
10	121	E


Marker proteins for each cluster (average log2 fold−change > 1.5)
Z−score
4
2
0
−2
−4
cluster_2
cluster_5
cluster_9
cluster_10
cluster_6
cluster_8
cluster_1
cluster_3
cluster_4
cluster_7
Hallmark geneset enrichment scores for cluster markers
Peroxisome
Myc Targets V1
PI3K/AKT/mTOR Signaling
Fatty Acid Metabolism
p53 Pathway
Apoptosis
Interferon Gamma Response
UV Response Up
Protein Secretion
Hypoxia
Adipogenesis
Oxidative Phosphorylation
Coagulation
mTORC1 Signaling
Glycolysis
Angiogenesis
Reactive Oxygen Species Pathway
IL−2/STAT5 Signaling
Cholesterol Homeostasis
UV Response Dn
Heme Metabolism
IL−6/JAK/STAT3 Signaling
Interferon Alpha Response
Complement
Estrogen Response Late
Inflammatory Response
Myc Targets V2
Estrogen Response Early
E2F Targets
Unfolded Protein Response
G2−M Checkpoint
KRAS Signaling Dn
Androgen Response
Apical Junction
Allograft Rejection
Epithelial Mesenchymal Transition
Mitotic Spindle
Myogenesis
TGF−beta Signaling
Bile Acid Metabolism
KRAS Signaling Up
TNF−alpha Signaling via NF−kB
Apical Surface
Xenobiotic Metabolism
Hedgehog Signaling
Notch Signaling
Pancreas Beta Cells
DNA Repair
Wnt−beta Catenin Signaling
Scaled enrichment score
2
0
−2
2
5
9
10
6
8
1
3
4
7
Genetic disease geneset enrichment scores for cluster markers
myopathy
cardiomyopathy, hypertrophic
dystonia
cardiomyopathy
cardiomyopathy, dilated
ataxia
leukemia
breast cancer
lymphoma
alzheimer disease
spinocerebellar ataxia
thyroid carcinoma
neuropathy
ovarian cancer
corneal dystrophy
zellweger syndrome
adenoma
retinitis pigmentosa
microphthalmia
pancreatic cancer
deafness
cataract
cdeficiency
glycogen storage disease
charcot−marie−tooth disease
encephalopathy
macular degeneration
orofacial cleft
muscular dystrophy
spastic paraplegia
colorectal cancer
dementia
parkinson disease
arrhythmogenic right ventricular dysplasia
mental retardation
ichthyosis
anemia
epidermolysis bullosa
Scaled enrichment score
4
2
0
−2
−4
2
5
9
10
6
8
1
3
4
7