EMT Multi-Omics

Deconvolution of proteomics data using scRNA-Seq

Ahmed Youssef

December 09, 2021

Introduction

Biological research has become increasingly defined by the generation of large-scale datasets in need of specialized computational analysis to unravel the complexities underlying living systems. 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, through their networks of interactions, carry out most of the vital biological processes governing cellular functions, yet remain largely unexplored at the resolution of single cells, representing crucial gaps in our knowledge of cell biology. The current pandemic has exposed our need for a better understanding of the contextual nature of cellular functions in order to develop effective targeted therapies. While there remain challenging technical hurdles to overcome for experimentally measuring proteins on a systems-wide scale at the level of individual cells, single-cell RNA sequencing (scRNA-Seq) has emerged in recent years as a powerful technology for defining cell states on a large scale, enabling breakthroughs in many areas of cell biology research, and raising the question of whether this type of data can be used for making inferences at the protein level. **In this report, we explore the deconvolution of bulk proteomics data to the single-cell level using scRNA-Seq data.**

Experiment summary

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, microarray, and proteomics datasets generated in this study.

Rationale

A regression-based approach to this problem is suggested here which is primarily concerned with the lack of strong correlation between a gene's RNA and protein levels, which is the main challenge with analyses aiming to integrate these two data modalities. Multiple biological and technical factors, such as alternative splicing and post-translational modifications, drive the weak correlation between the RNA and protein levels of the same set of genes, thus preventing us from using the gene expression values derived from scRNA-Seq to make inferences on the corresponding protein levels. Our approach aims to bridge this gap by utilizing the ability of scRNA-Seq to delineate functionally-distinct cell populations from heterogeneous mixtures, allowing us

to distinguish the identity and proportions of the different cells populations making up the bulk sample for which we have the proteomics data measurements, as well as how these proportions change under different experimental contexts such as health and disease. We cannot reliably use these proportions to directly estimate the contribution of each population to each protein's abundance at the bulk-level, however, due to variations in the relative abundance of each protein across the different cell populations. Leveraging the shifts in the bulk sample composition, as quantified by scRNA-Seq, we can instead use a regression model to combine the changes in cell population proportions with the corresponding fluctuations in bulk-level protein abundance as suggestive of relationships between specific cell populations and bulk protein levels. This information can then be used to estimate the protein levels within each cell population. Essentially, this approach aims to leverage scRNA data to break down bulk proteome measurements into cell population-level proteomes.

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 bulk proteomics 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.

Data summary - Bulk mRNA

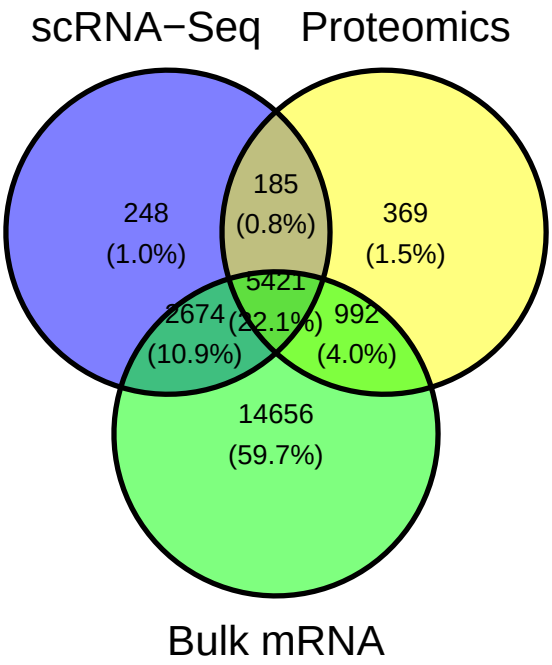
The bulk mRNA data comes from a microarray experiment. Summary of the dataset follows:

- 23,743 genes
- 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.

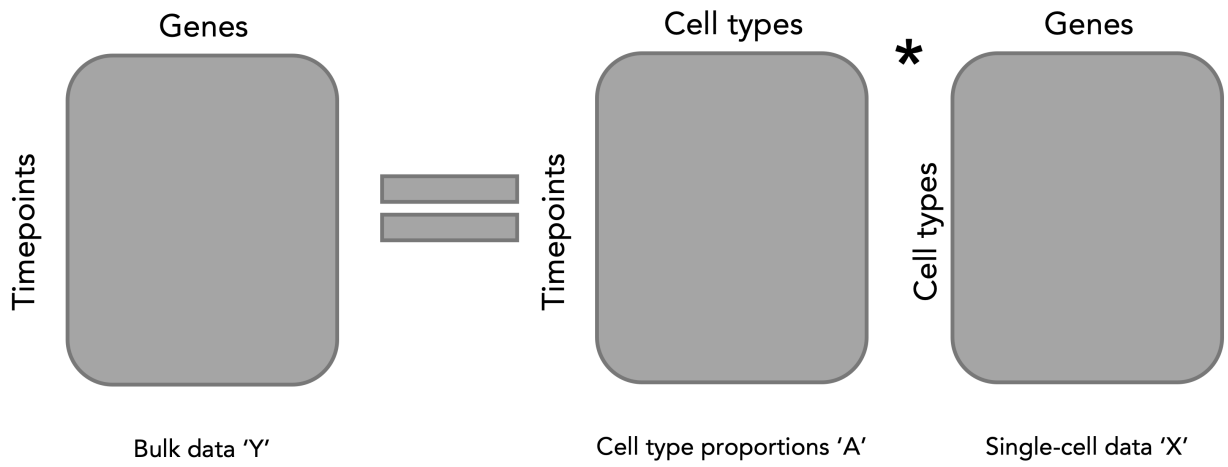
Protein overlap

The Venn diagram below shows the overlap of the identified proteins in the datasets.



Approach

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.



The five main steps underlying our approach are outlined 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. We tested our approach on different pre-defined numbers of clusters in our analysis to examine the effect of varying the number of cell populations.

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 using ridge regression***: Re-create the single-cell data from the pseudo-bulk data Y and the timepoint-specific cell cluster counts A based on the formula $Y = AX$ by using the formula $X' = (A^T A)^{-1}(A^T Y)$, which is essentially the pseudo-inverse of A multiplied by the pseudo-bulk Y. To achieve this, we solve the non-negative constrained equation $\hat{X} = min_{x>=0}(-2Y^T AX + X^T A^T AX)$ after adding a ridge penalty λ to the diagonal of the matrix $A^T A$. The R function *solve.QP* from the *quadprog* package is used to solve this equation one gene at a time to estimate the expression profile at the cell-type level. To decide on the optimal value for the parameter λ, we test a range of 21 values between 1^{-10} to 1^{10} for each gene and each number of clusters. For each value of λ, we compute 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 that leads to the minimal error is selected as the optimal value for the corresponding gene.

To summarize, the following three matrices represent the key variables in our model:

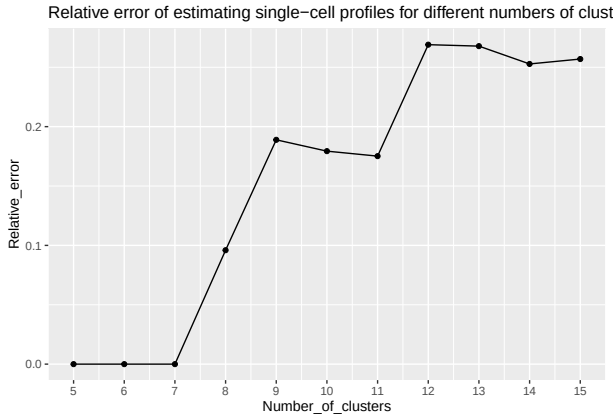
- Matrix A of dimensions $timepoints * clusters$. (cell type counts in each timepoint)
- Matrix X of dimensions $clusters * genes$. (cluster-averaged single-cell RNA data)

- Matrix Y of dimensions $timepoints * genes$. (bulk data)

We then attempt to re-create the single-cell matrix X' data by computing $Y = AX'$.

Method Development and Optimization

The key input parameter for the deconvolution algorithm is the regularization parameter λ for solving the ridge regression problem. After selecting the optimal λ for each gene we report the errors as relative RMAD (relative mean absolute deviation) using the formula $|X - X'|/X$, where $|X|$ is the absolute value of the difference. This error is computed for each gene and the final reported score is the average RMAD value, in other words: *on average, how different is a gene's predicted values compared to the true ones?* The resultant relative RMAD values are shown below.



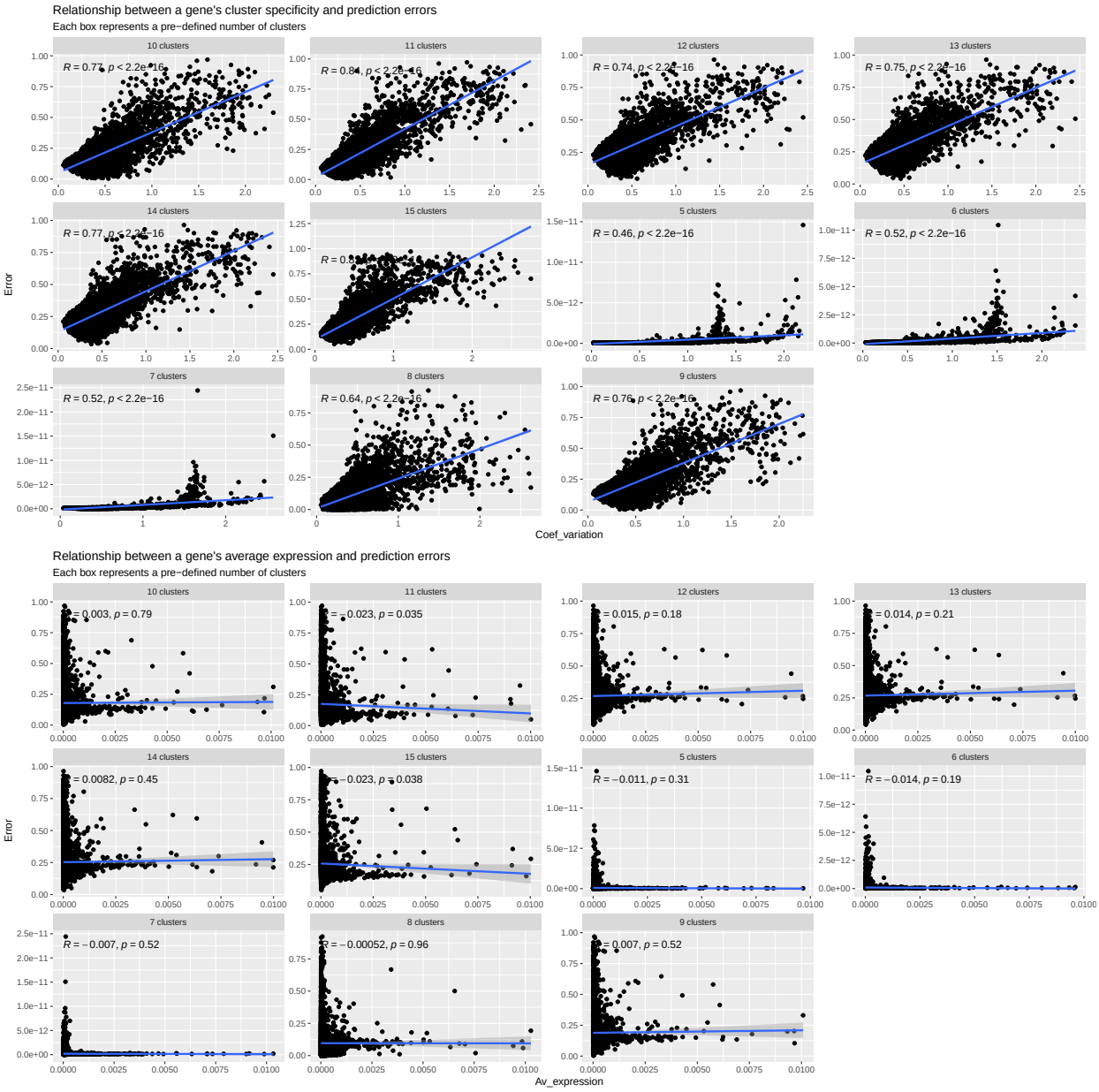
Distribution of errors across genes

This section explores whether the single-cell predictions might be more accurate for a subset of the genes, such as differential markers, by comparing the distributions of prediction errors for individual genes as they relate to cell-type expression specificity.

To explore the relationship between properties of the genes and their corresponding predicted values, for each gene we compute the following measures:

- ***Expression specificity:*** This metric looks at the relative specificity of a given gene's expression to the clusters. The coefficient of variation of the gene's cluster-specific expression values is computed as the standard deviation of the per-cluster expression values divided by the mean.
- ***Average expression:*** This metric is concerned with the relative abundance of each gene's transcript, and is simply computed as the mean expression of the gene across clusters.

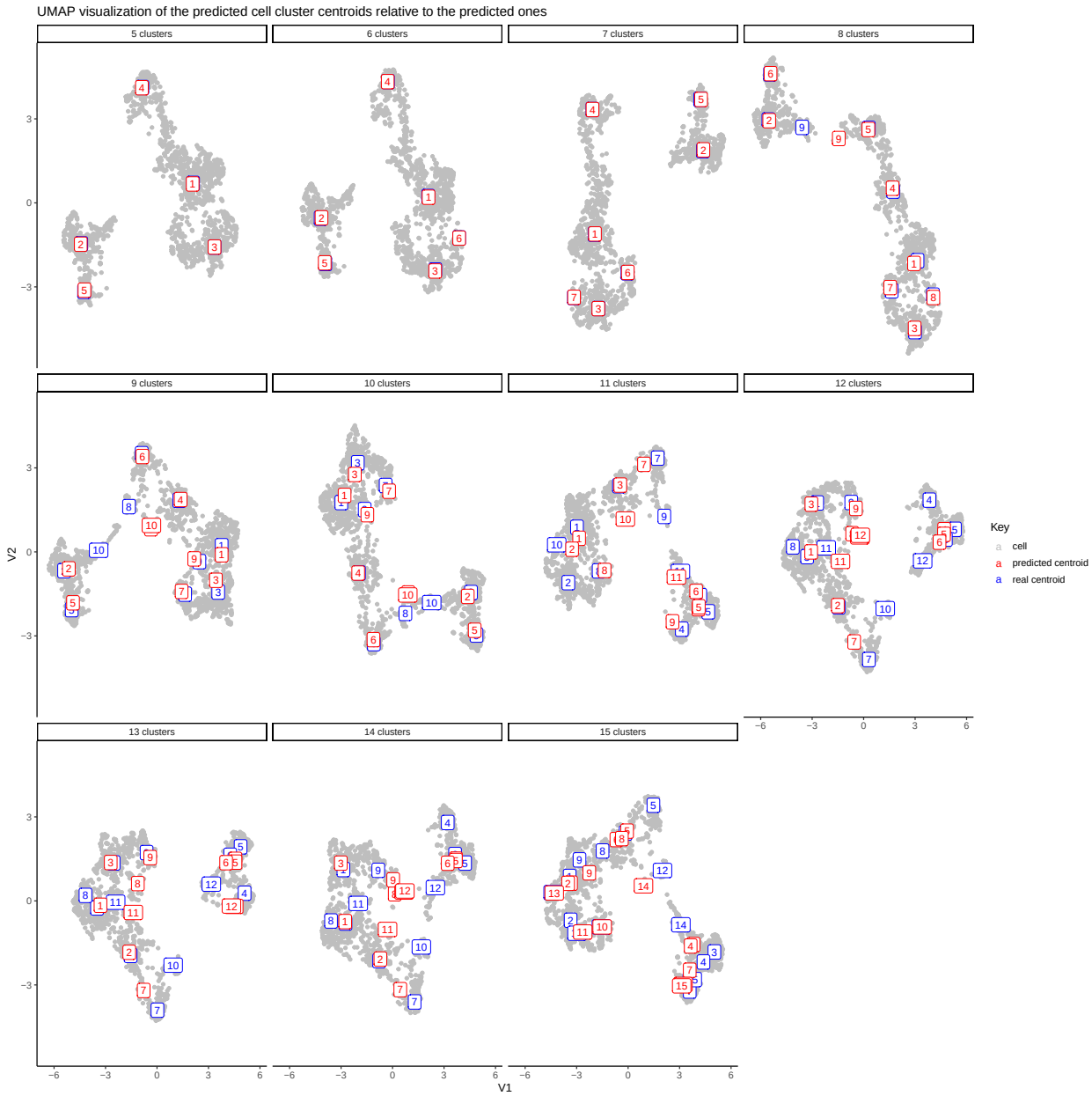
Each of the above measures are then compared to the error in predicting the gene's expression value using regression.


Relationship between a gene's cluster specificity and prediction errors
Each box represents a pre−defined number of clusters
10 clusters
R = 0.77, p < 2.2e−16
11 clusters
R = 0.84, p < 2.2e−16
12 clusters
R = 0.74, p < 2.2e−16
13 clusters
R = 0.75, p < 2.2e−16
14 clusters
R = 0.77, p < 2.2e−16
15 clusters
5 clusters
R = 0.46, p < 2.2e−16
6 clusters
R = 0.52, p < 2.2e−16
7 clusters
R = 0.52, p < 2.2e−16
8 clusters
R = 0.64, p < 2.2e−16
9 clusters
R = 0.76, p < 2.2e−16
Error
Coef_variation
Relationship between a gene's average expression and prediction errors
Each box represents a pre−defined number of clusters
10 clusters
= 0.003, p = 0.79
11 clusters
= −0.023, p = 0.035
12 clusters
= 0.015, p = 0.18
13 clusters
= 0.014, p = 0.21
14 clusters
= 0.0082, p = 0.45
15 clusters
= −0.023, p = 0.038
5 clusters
R = −0.011, p = 0.31
6 clusters
R = −0.014, p = 0.19
7 clusters
R = −0.007, p = 0.52
8 clusters
R = −0.00052, p = 0.96
9 clusters
= 0.007, p = 0.52
Error
Av_expression

Evaluation of single-cell predictions

The below sections examine the accuracy of the single-cell predictions made using ridge regression in the previous section. Note: the bulk matrix that is deconvoluted in this section is the reconstructed pseudo-bulk matrix ($Y = AX$).

Are our predicted cluster centroids close to the real ones? The cluster centroid is each cluster's average gene expression profile. Visualizing the two sets of centroids (real & predicted) side-by-side on a *UMAP* will help us answer this question. For each number of clusters between 5 and 15, we cluster the data using *Seurat* and construct the matrix X which contains the average gene expression by cluster. We then use the ridge regression approach defined in the previous section to construct the matrix X'. Finally, we perform a *UMAP* projection using the R package uwot on the original gene expression matrix but with the addition of the real and predicted cluster centroids, i.e. the rows of X and X'. The *UMAP*s below show the projections for each number of clusters.

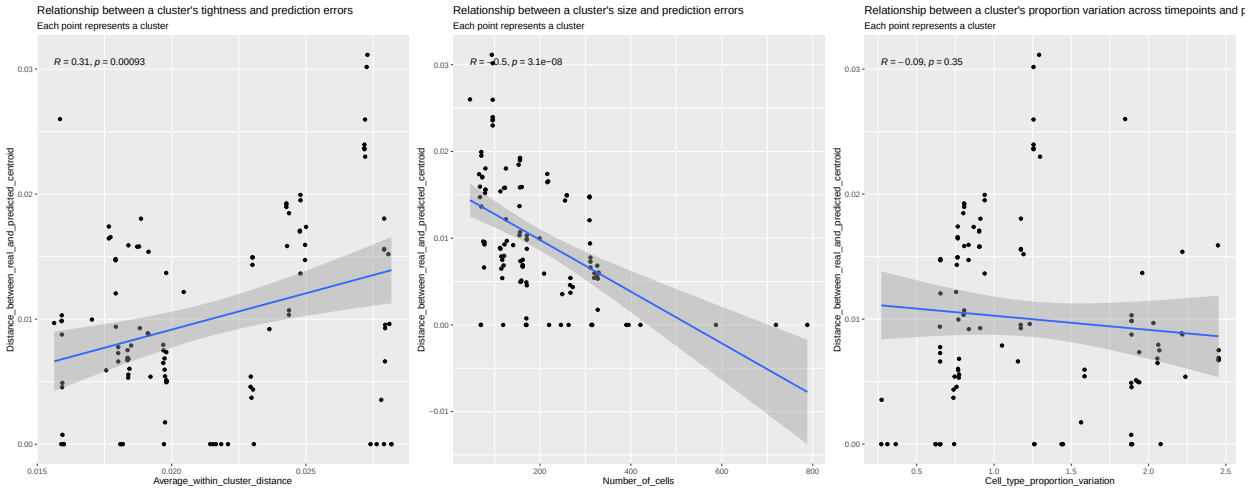


How similar is the average cell to it's real centroid compared to the predicted one? To determine if our predictions are reasonable approximations of the real data, we also compare the predictions to the real data by computing the Euclidean distance between individual cells' measurements and their real cluster average as opposed to the distance between our predictions and the same cluster average.

More specifically, for a given number of clusters, we cluster the data using *Seurat*, followed by iterating over each of the ~1,900 cells in the original single-cell expression matrix and computing the Euclidean distance between it's profile and that of its cluster's centroid. Next, for each cluster we compute the average distance of it's cells to the centroid, as well as the distance between our predicted centroid and the corresponding cluster centroid. By comparing these two distances side-by-side, we can determine whether our predicted centroids fall within the correct intra-cluster range.

Which clusters are easier to predict? The accuracy of recapturing the cell cluster profiles varied by cluster. In this section, we are interested in examining the mathematical properties of the cell clusters derived

from the scRNA-Seq data that influence the quality of our predictions. We first compute the correlation between how 'tightly-knit' a cluster is, i.e. average within-cluster distance to the centroid, and the error in predicting the profiles. We also correlate the prediction error with the number of cells in each cluster and the variation of each cell type's proportion across timepoints. The error in this section is taken as the Euclidean distance between the cluster's predicted centroid and the actual centroid.



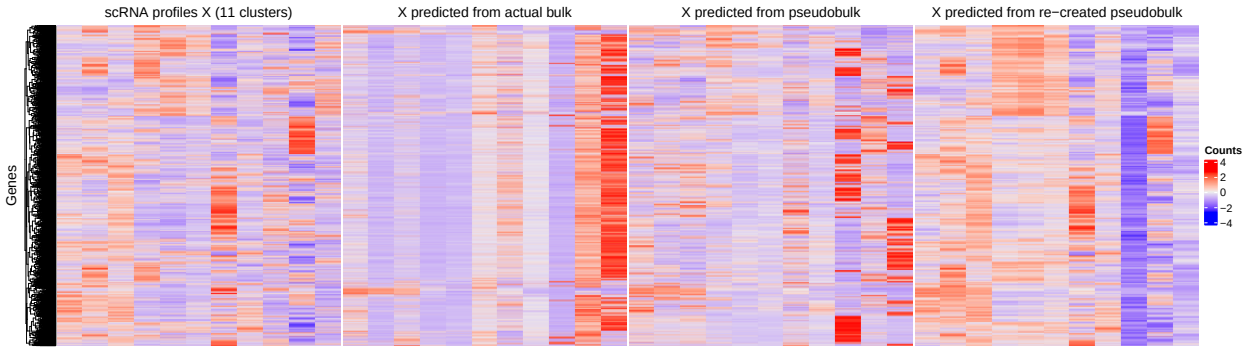
Visualizing predictions and real values side-by-side

The below heatmaps visualize the real scRNA-defined cluster profiles with the predicted ones from deconvoluting the different RNA datasets described below.

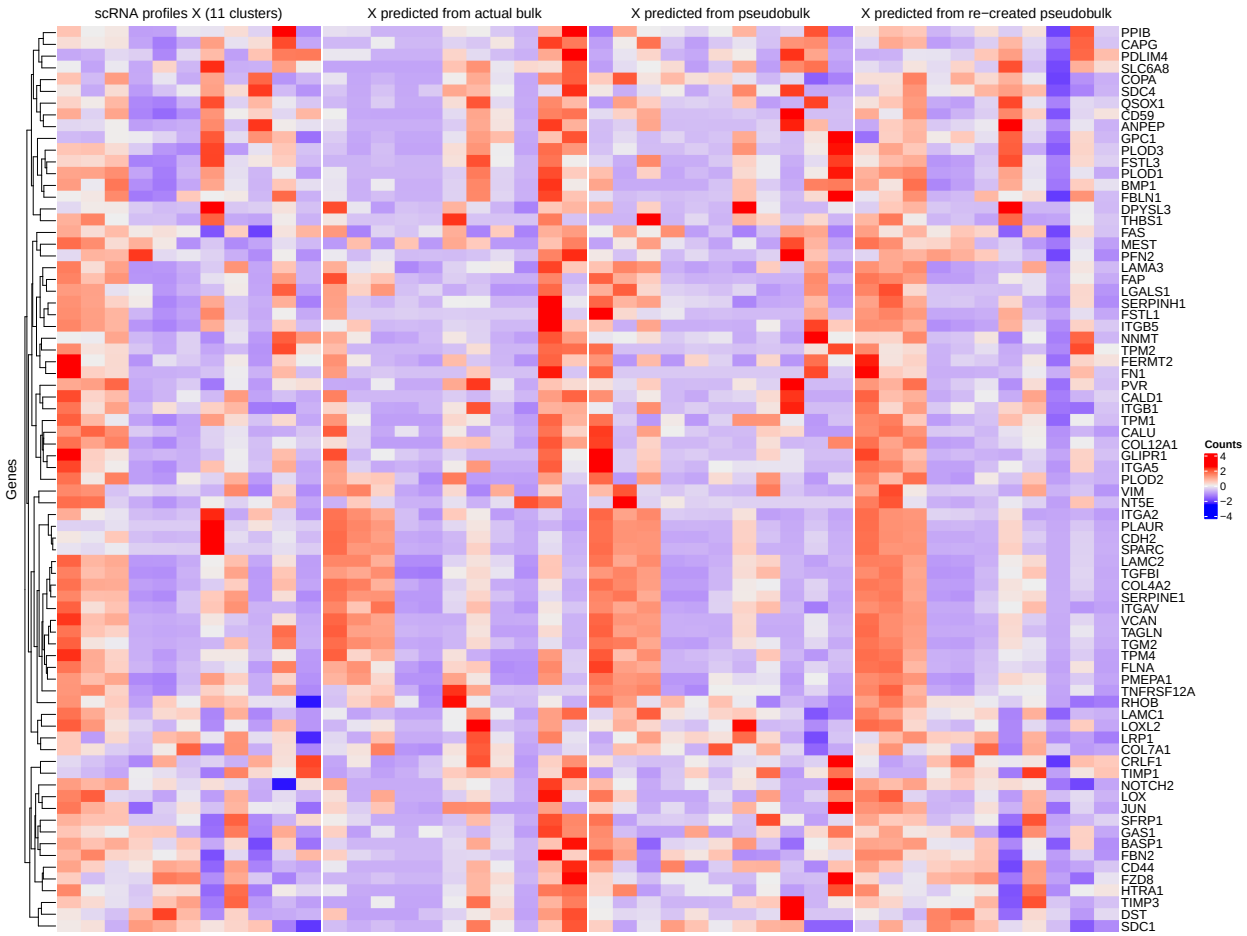
1) ***Pseudo-bulk***: A pseudo-bulk RNA dataset is created for each timepoint by summing the gene counts of all cells within the timepoint.
2) ***Reconstructed Pseudo-bulk***: The reconstructed pseudo-bulk is created by matrix multiplication of the average cell cluster expression of each gene and the number of cells in each cluster in each timepoint, i.e. $Y = AX$.
3) ***Bulk mRNA***: Bulk mRNA data from a microarray experiment.

The λ values used for the regression were those deemed to be optimal based on the reconstructed pseudobulk dataset. We furthermore focus on a set of 80 hallmark genes associated with EMT from the MSigDB database.

All overlapping genes (scaled across clusters)



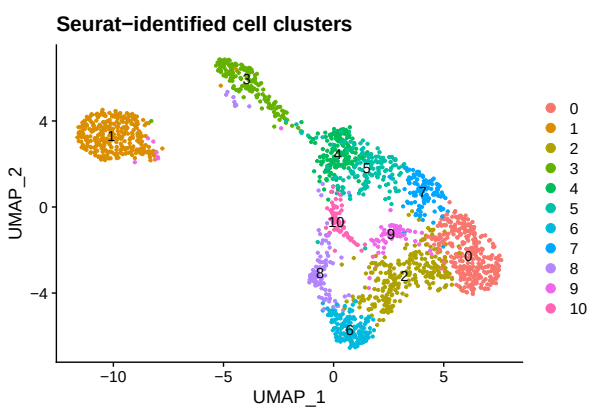
EMT hallmark genes (scaled across clusters)

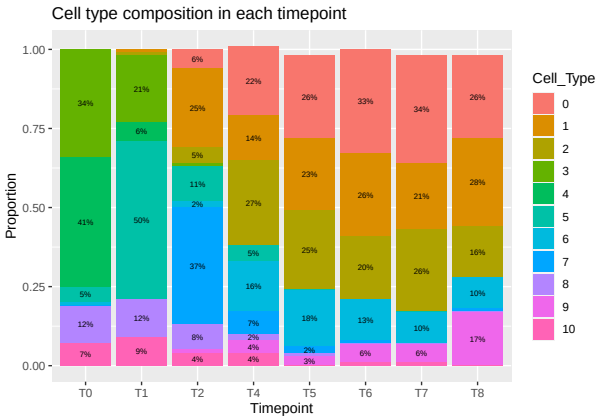


Deconvolution of proteomics data

In this section of the report, we focus on deconvoluting the bulk proteomics data with 11 Seurat-defined clusters. We furthermore focus on a set of 80 hallmark genes associated with EMT from the MSigDB database.

Seurat clustering results





Predicted cell cluster profiles