Deconvolution of proteomics data using scRNA-Seq

Validating computationally-predicted proteomes using single-cell proteomics

Ahmed Youssef

December 16, 2021

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, through their networks of interactions, carry out most of the vital biological processes governing cellular functions, yet remain largely unexplored in the single-cell space, representing crucial gaps in our knowledge of cell biology. 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 it 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 and validate the results using single-cell proteomics 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 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 potentially be used to estimate the contribution of individual cell states to the bulk proteomics measurements.

Data summary - Proteomics

The bulk proteomics data was generated in the Emili Lab using standard tandem 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

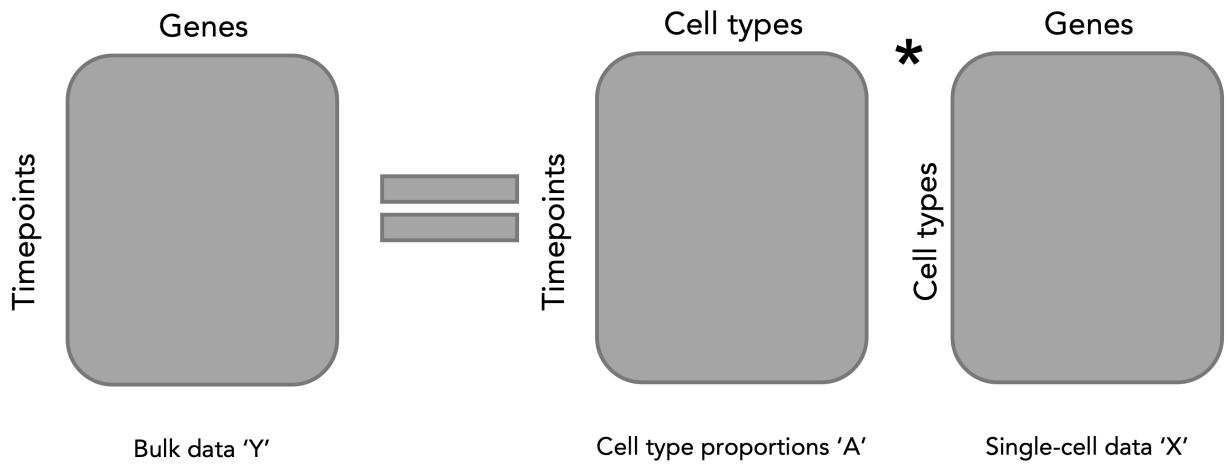
The bulk proteomics data was generated in the Emili Lab using standard mass-spectrometry. Summary of the dataset follows:

- 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.

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.

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 data)

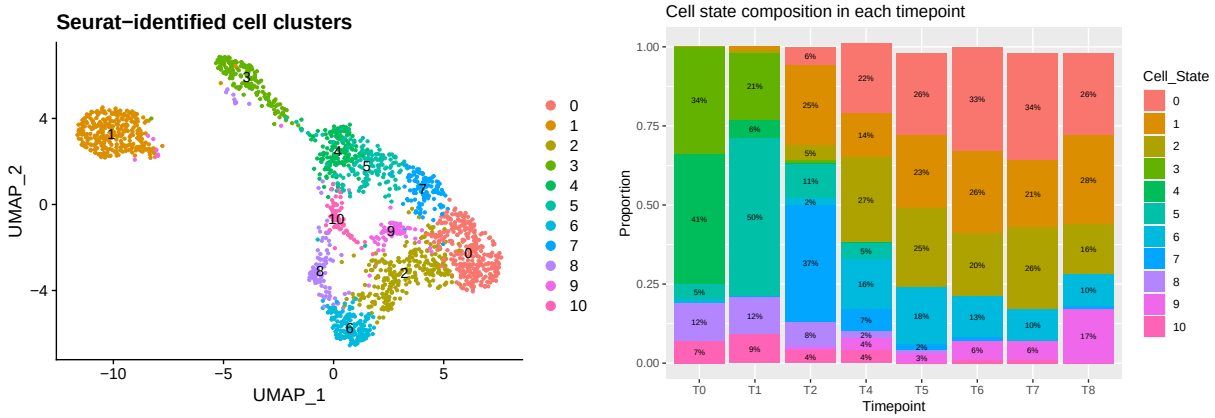
- Matrix Y of dimensions $timepoints * genes$. (bulk data)

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

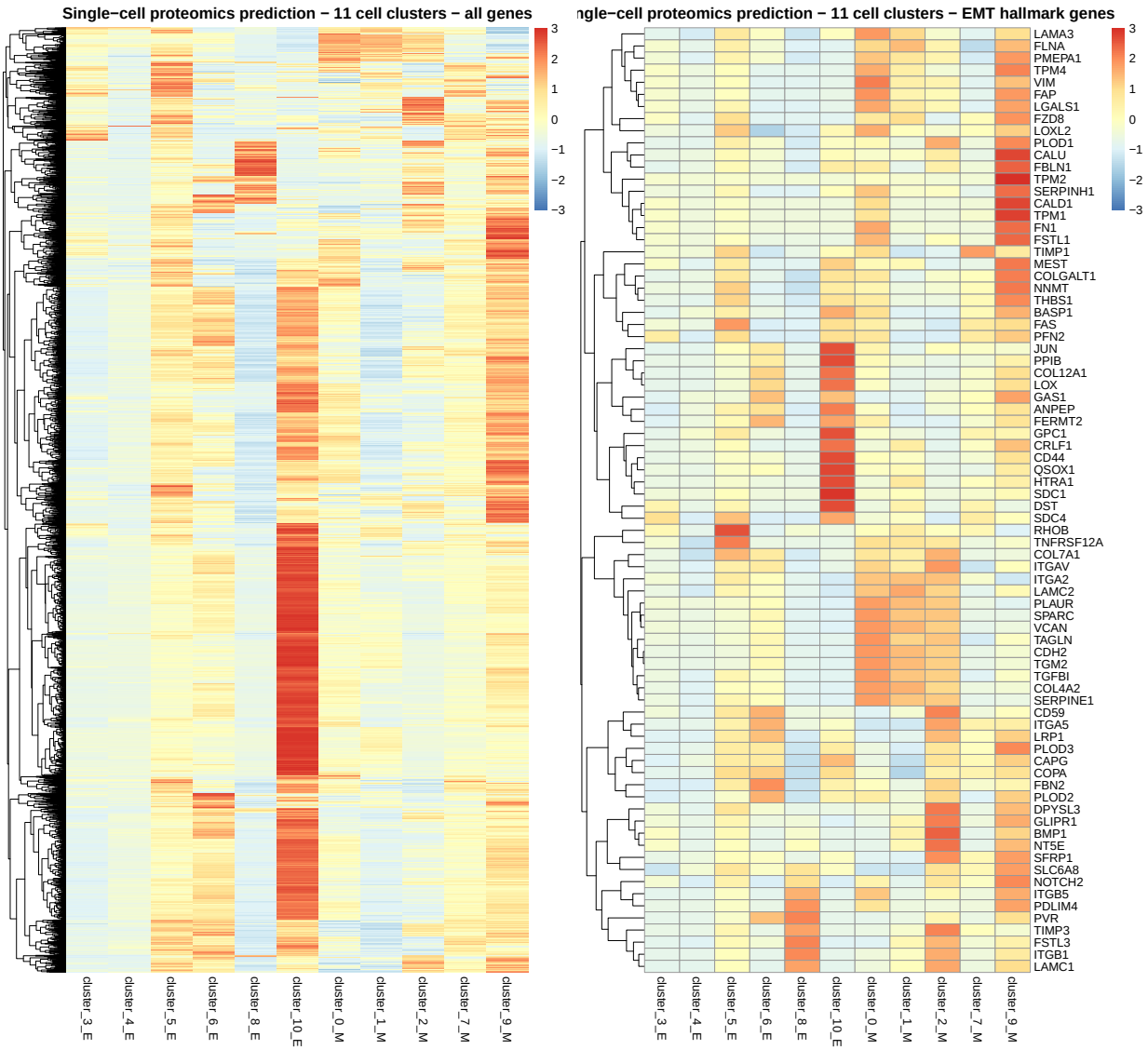
Deconvolution of bulk proteomics data using scRNA-defined cell clusters

In this section of the report, we apply the deconvolution algorithm outlined above to the bulk proteomics data with 11 Seurat-defined cell clusters from the scRNA-Seq data. We furthermore highlight a set of 80 hallmark genes associated with EMT from the MSigDB database.

Seurat clustering results



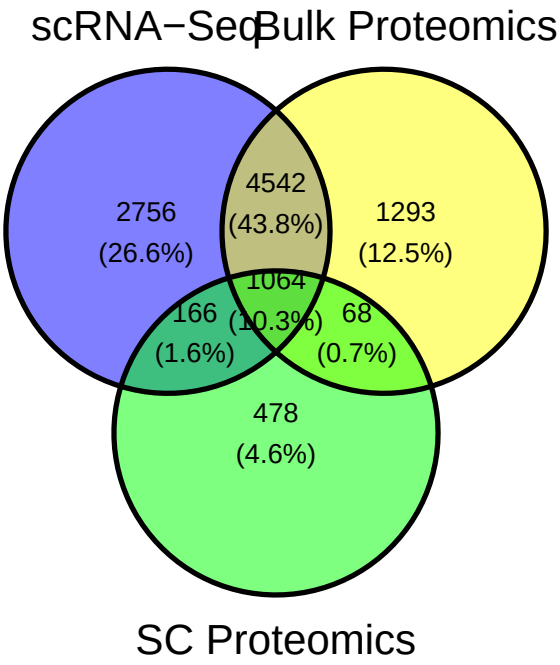
Predicted single-cell profiles



Deconvolution of bulk proteomics data using single-cell proteomics-defined cell clusters

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, 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.

Protein overlap



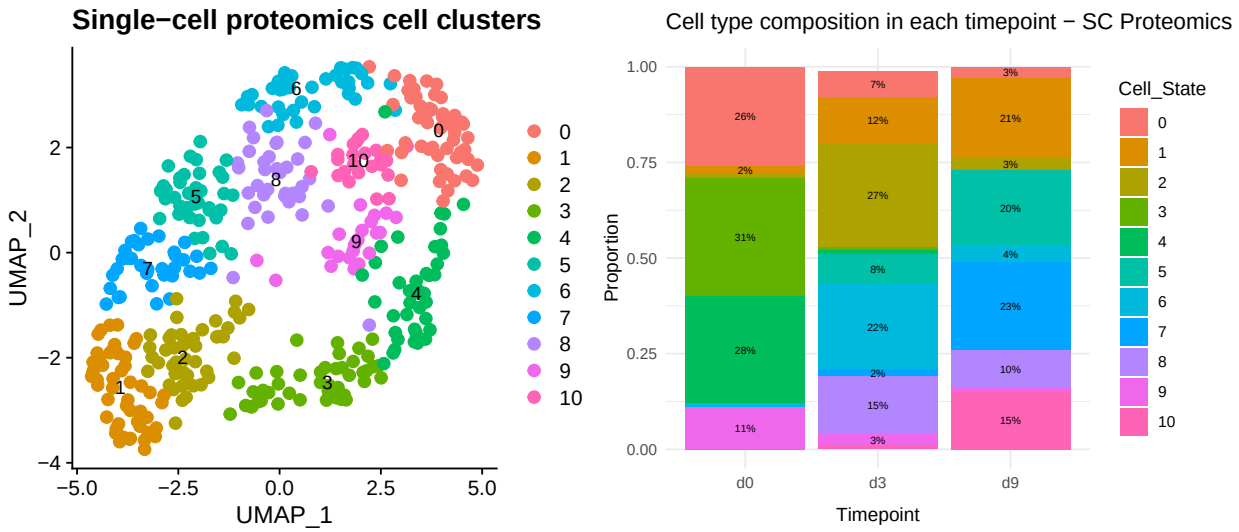
Data Pre-processing

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. 65% of the original single-cell proteomics matrix consists of missing values. 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. Similar to scRNA-Seq pre-processing strategies, the data was also normalized such that each cell sums to 1 after imputation. The final matrix consists of 1,064 proteins and 420 cells.

Identify cell clusters in single-cell proteomics data

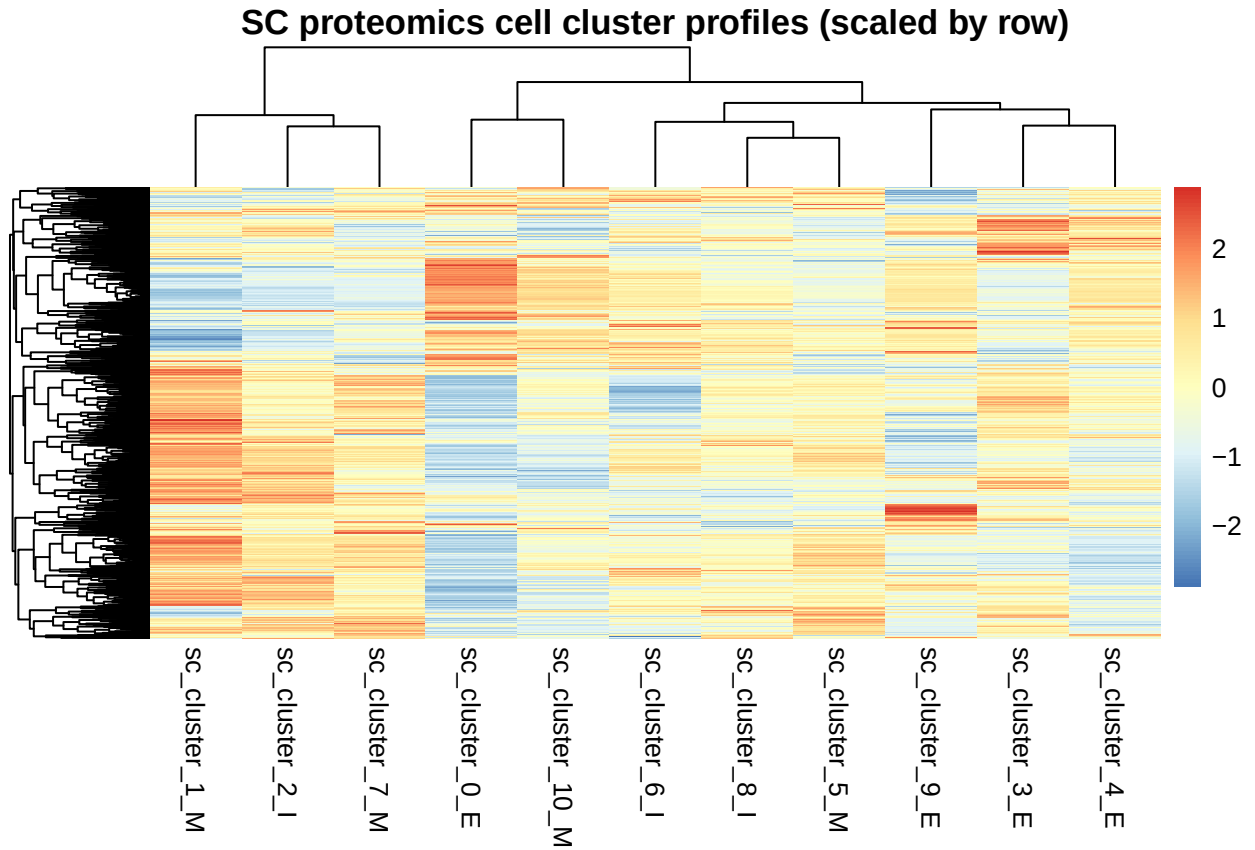
The cell states in the single-cell proteomics dataset are identified in an unsupervised manner based on similarity of protein abundance profiles. All cells from all timepoints are pooled together for this analysis. Seurat is used to cluster the cells following their standard pipeline for scRNA-Seq data. Our model defines 11 cell clusters, to be comparable to the clusters identified in the scRNA-Seq data, by using a Seurat clustering resolution of 2.6 in the proteomics data and 1.2 in the RNA data. The below UMAP plots visualize the identified cell clusters.

Note: we re-clustered the scRNA-Seq data in this section using the genes that overlapped with the single-cell proteomics data.



Cell cluster profiles

For each cluster, a cell cluster profile is constructed as the average gene expression profile of the member cells. Consequently, the cell cluster profile matrix X consists of the ~1,000 genes * 11 cell clusters. Cell clusters are labeled as epithelial (E), mesenchymal (M), or intermediate (I) based on manual inspection of the above timepoint proportion plots. The heatmap below shows the constructed cell cluster profiles scaled by row.



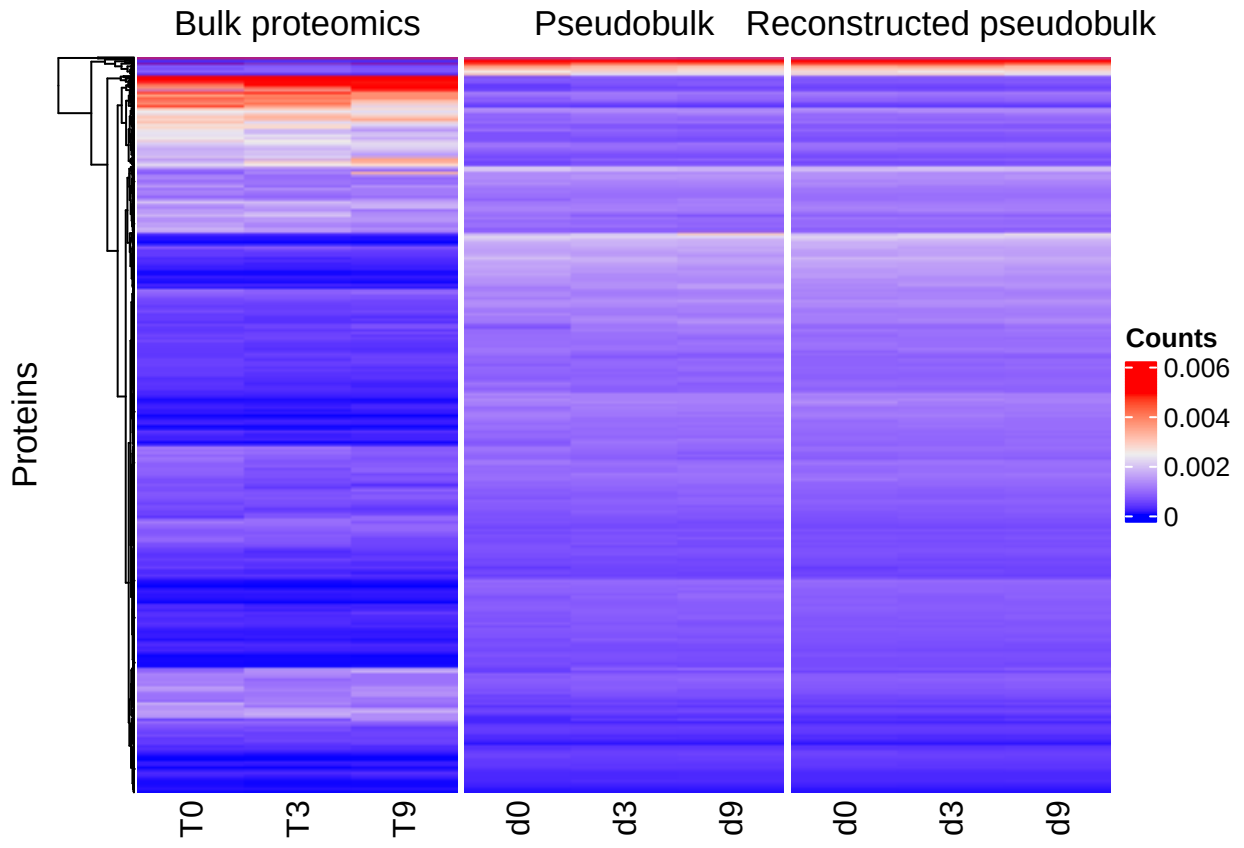
Re-constructing bulk proteomics data from single-cell proteomics data

In this section, we examine the relationship between the bulk proteomics data generated by the Emili Lab and the single-cell proteomics data generated by the Slavov Lab, both of which are dependent on mass spectrometry. The bulk proteomics data was subset to the three timepoints corresponding to days 0, 3, and 9 to correspond with the single-cell proteomics data. The three replicates of each timepoint in the bulk data were averaged to produce one profile per timepoint.

To compare the single-cell proteomics to the bulk, we create two datasets from the single-cell proteomics data:

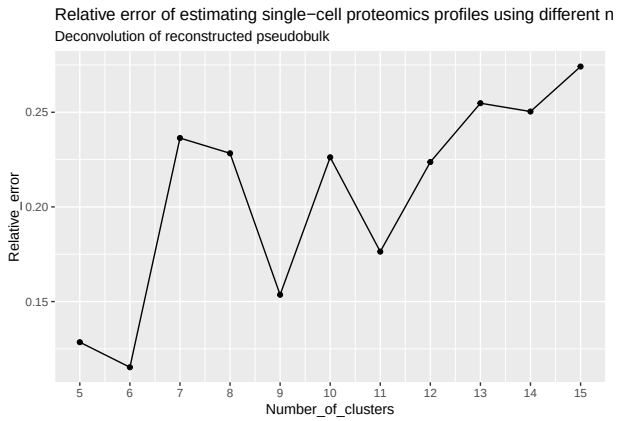
1) ***Pseudo-bulk***: A pseudo-bulk proteomics dataset is created for each timepoint by summing the protein levels of all cells within the timepoint.
2) ***Reconstructed Pseudo-bulk***: The reconstructed pseudo-bulk is created by matrix multiplication of the average cell cluster abundance of each protein by the number of cells in each cluster in each timepoint, i.e. $Y = AX$.

This gives three bulk proteomics datasets in total, including the actual bulk and the two datasets above. The three bulk datasets are shown below, with the genes maintained in the same order.



The deconvolution algorithm was applied to the reconstructed pseudobulk to examine the method's ability to re-construct the single-cell profiles. 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*

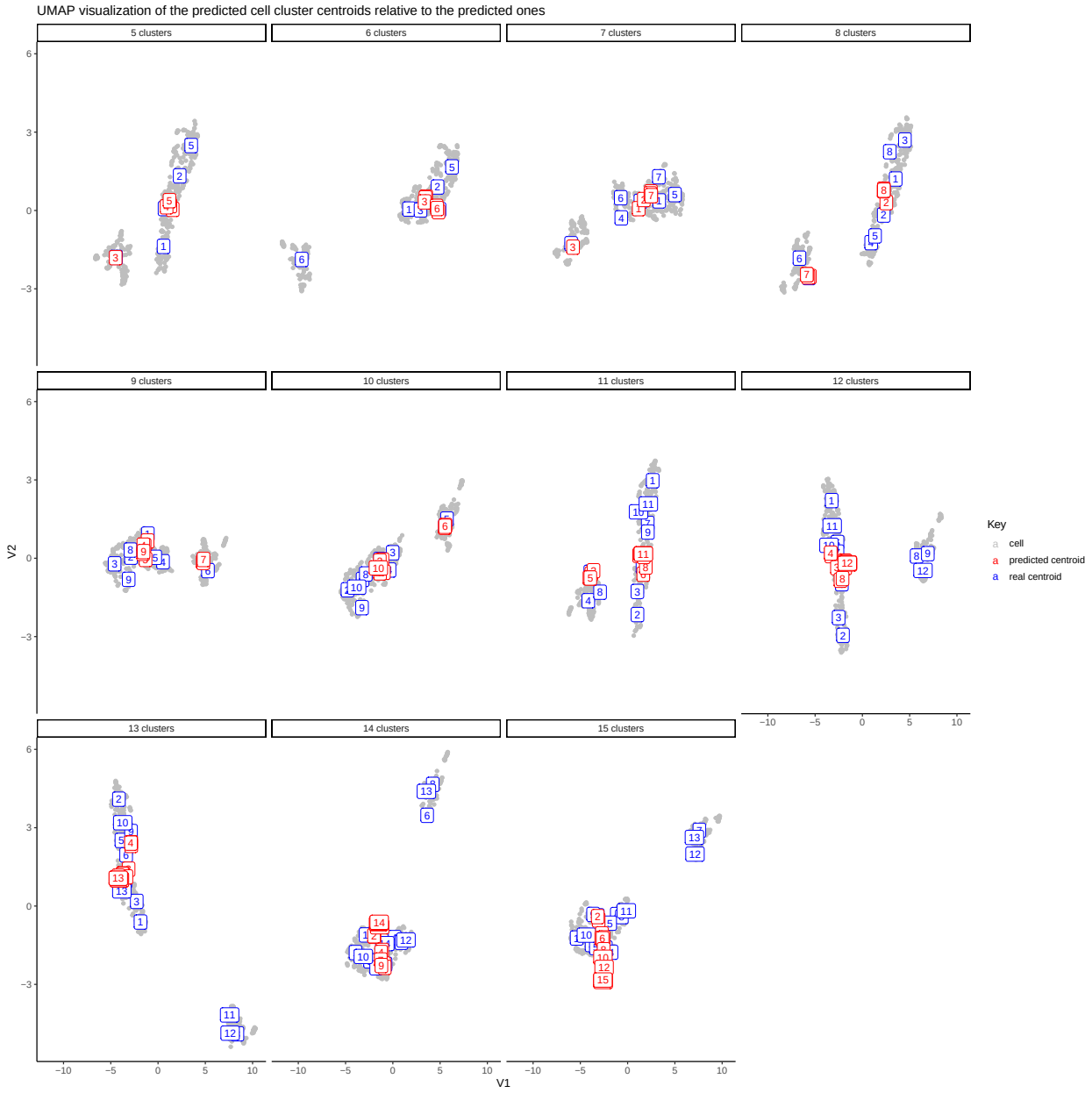
compared to the true ones? The resultant relative RMAD values are shown below for each number of clusters.



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 ($Y = AX$).

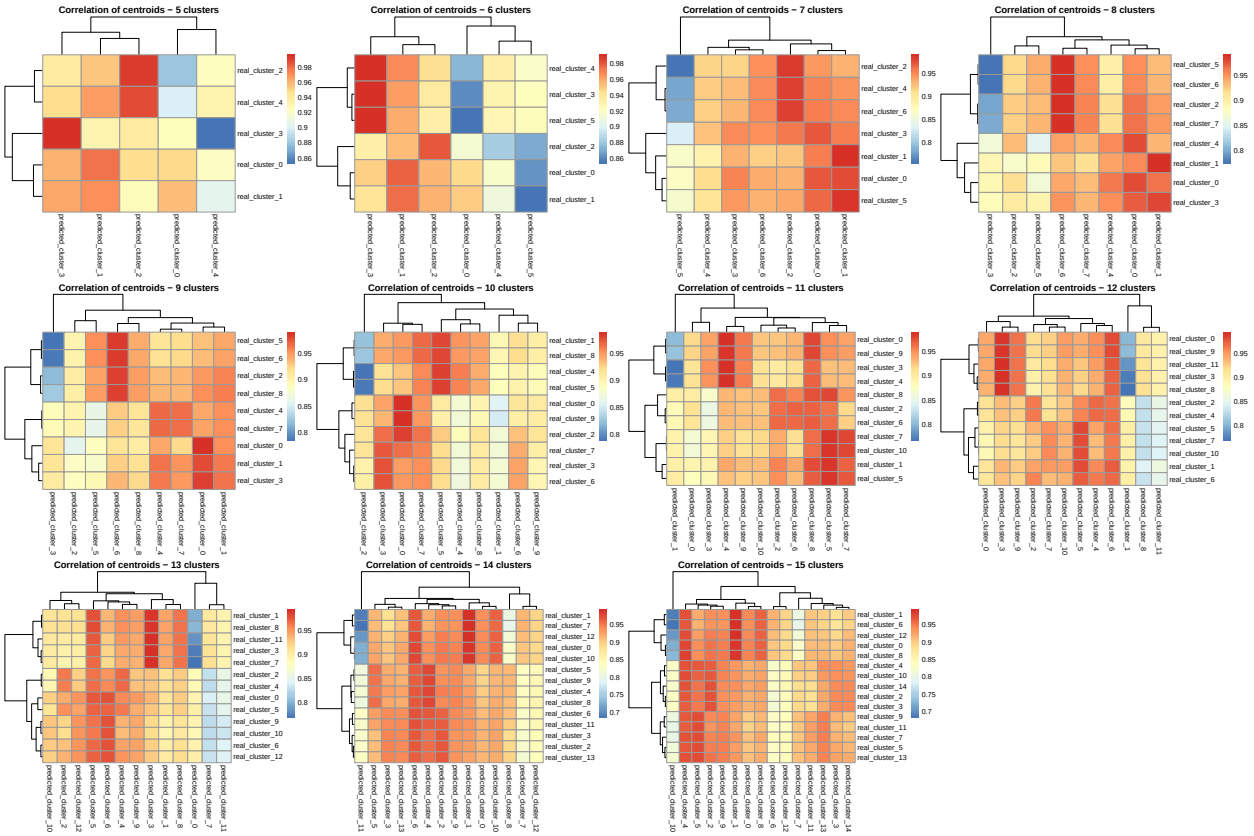
Are our predicted cluster centroids close to the real ones? The cluster centroid is each cluster's average gene expression profile. Visualizing the two sets of centroids (real & predicted) side-by-side on a *UMAP* will help us answer this question. For each number of clusters between 5 and 15, we cluster 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 single-cell proteomics 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.



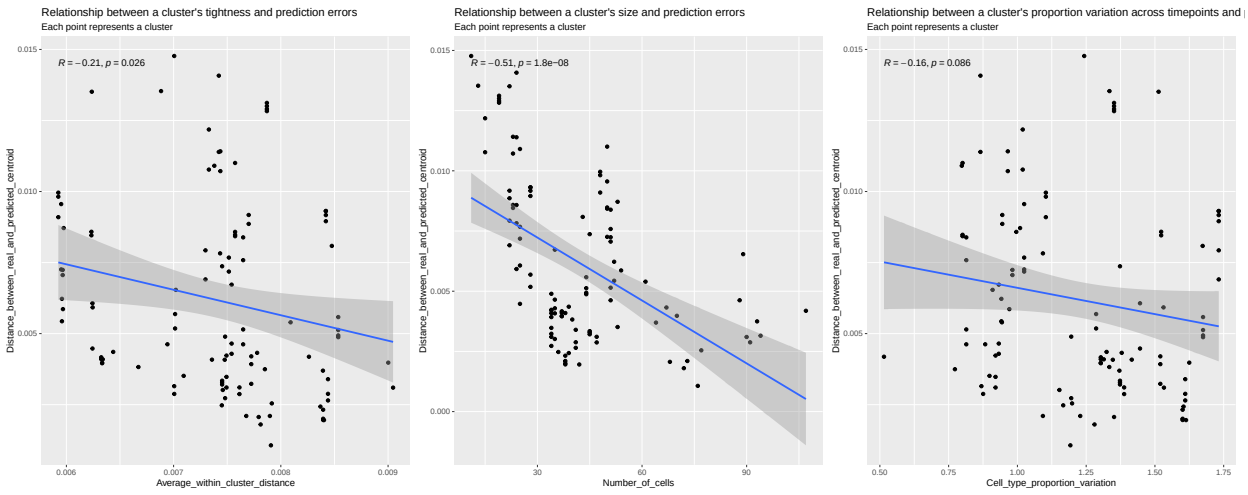
How similar is the average cell to it's real centroid compared to the predicted one? To determine if our predictions are reasonable approximations of the real data, we also compare the predictions to the real data by computing the Euclidean distance between individual cells' measurements and their real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

Correlation of cluster centroids The below plot visualizes the correlation of the cell cluster centroids predicted by our algorithm and those computed from the single-cell proteomics data.



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.

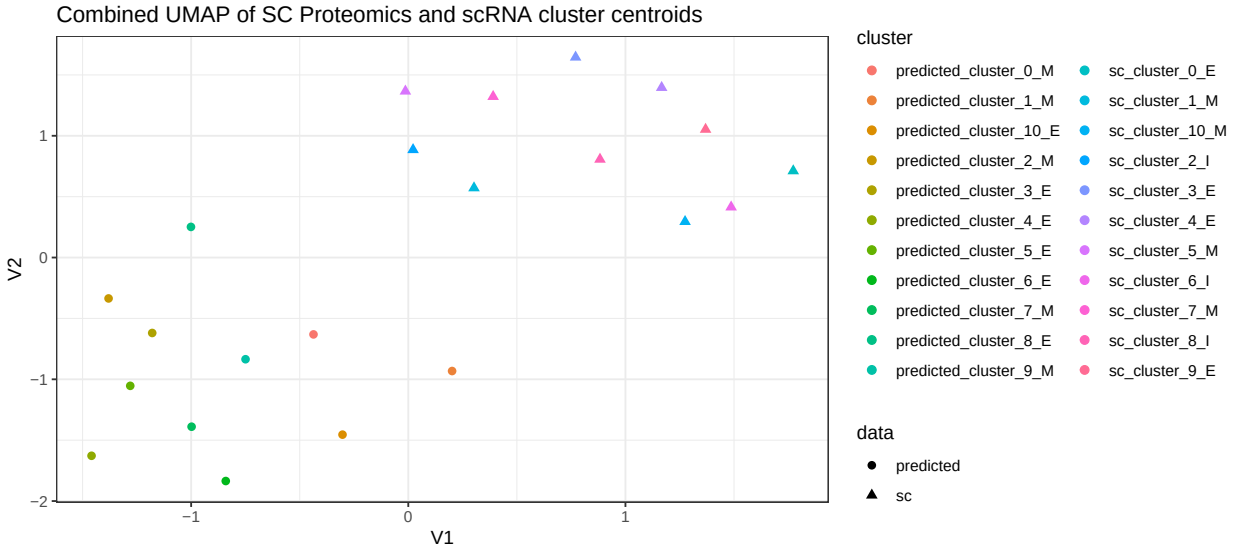


Comparing scRNA-derived predictions to single-cell proteomics data

In this section, we compare the experimentally-derived single-cell profiles to our computationally-predicted ones as an external benchmark.

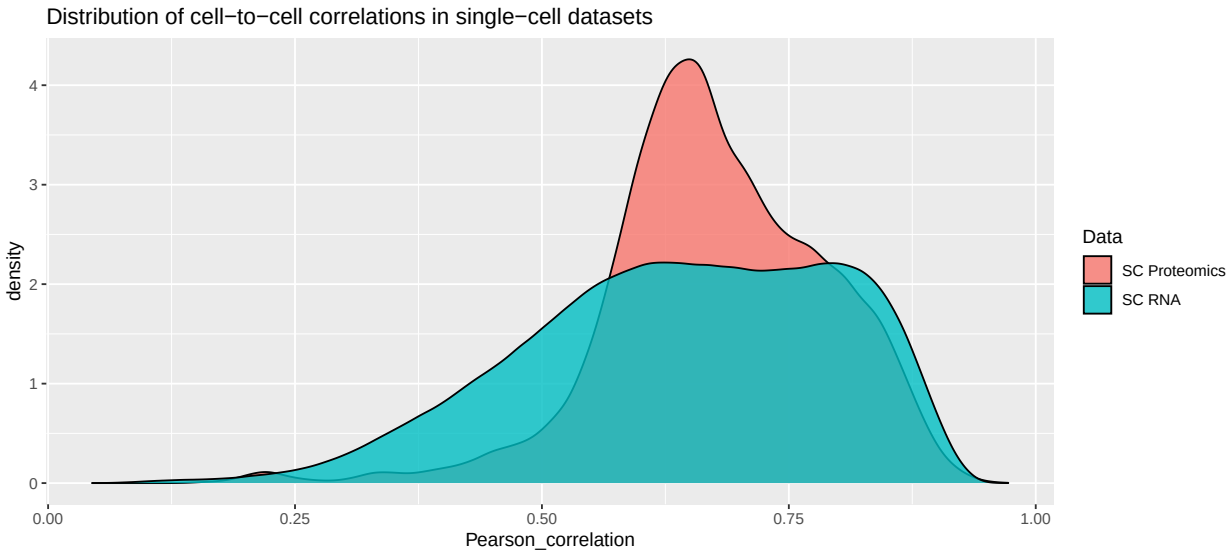
Comparing predicted and real cell clusters

The below UMAP plot includes the centroid of each cluster in each dataset. The centroid is computed as the average abundance of each protein in each cluster's cells. Prior to performing the UMAP dimensionality reduction, the values were normalized to sum to 1 in each cluster. We use the scenario with 11 cell clusters for this section.



Comparing correlation patterns in scRNA and single-cell proteomics

We examine and compare the correlation patterns in the scRNA and single-cell proteomics datasets. A correlation matrix was computed for each dataset that computed the Pearson correlation between cells based on all of the proteins (~8,500 for scRNA and ~200 for sc proteomics). The below plot visualizes the distributions of these correlation values.


Distribution of cell−to−cell correlations in single−cell datasets
density
Pearson_correlation
0
1
2
3
4
0.00
0.25
0.50
0.75
1.00
Data
SC Proteomics
SC RNA