EMT Multi-Omics

Deconvolution of proteomics data using scRNA-Seq

Ahmed Youssef

October 29, 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. While single-cell proteomics methods are still in their infancy, 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 begging the question of whether it can be used for making inferences at the protein level. **In this report, I 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.

Approach

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 factors, such as alternative splicing and post-translational modifications. Leveraging the timepoints present in this dataset, which conveniently show shifts in cell type abundances across time, we can instead look for changes in cell-type proportions and corresponding changes in bulk-level protein abundance as suggestive of relationships between specific cell types and specific proteins. This information can then potentially be used to estimate the contribution of individual cell types to the bulk proteomics measurements.

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

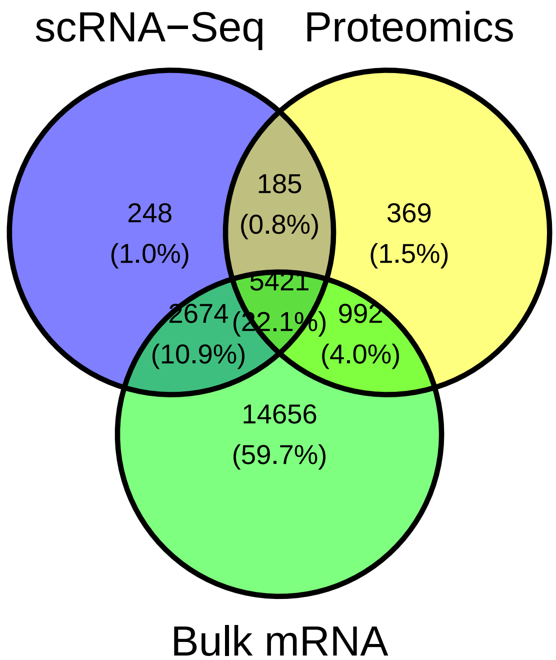
The bulk mRNA data comes from a microarray experiment. Summary of the dataset follows:

- 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.

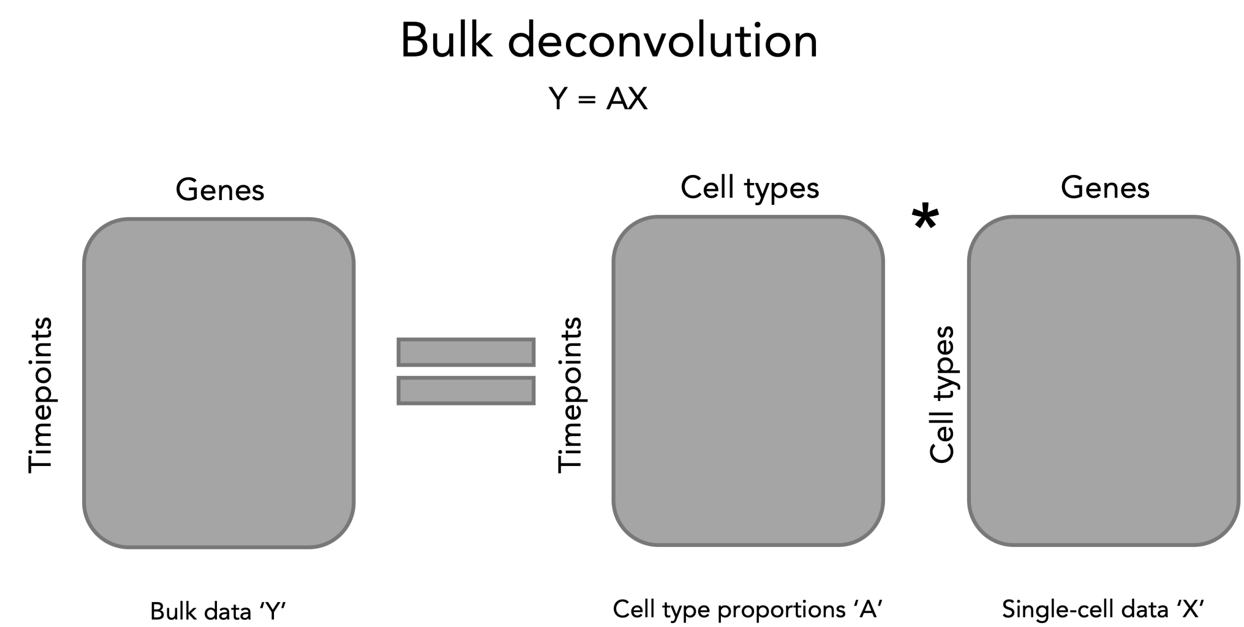
Protein overlap

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.

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 values between 10^{-10} and 1 for each number of clusters. For each value of λ, we sum the errors in estimating each gene's single-cell profile as a measure of the accuracy of the predicted single-cell profiles. The λ that leads to the minimal error is selected as the optimal value.

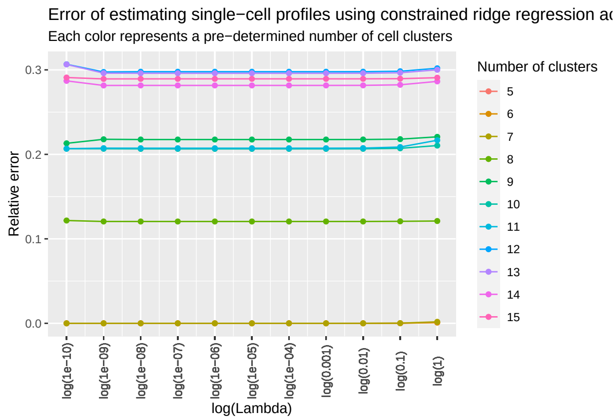
To summarize, the following three matrices represent the key variables in our model:

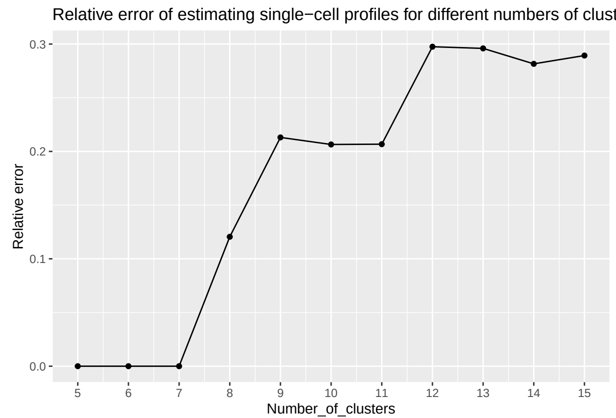
- Matrix A of dimensions $timepoints * clusters$. (cell type counts in each timepoint)
- Matrix X of dimensions $clusters * genes$. (cluster-averaged single-cell RNA data)
- Matrix Y of dimensions $timepoints * genes$. (bulk/pseudo-bulk RNA data)

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 number of clusters 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.

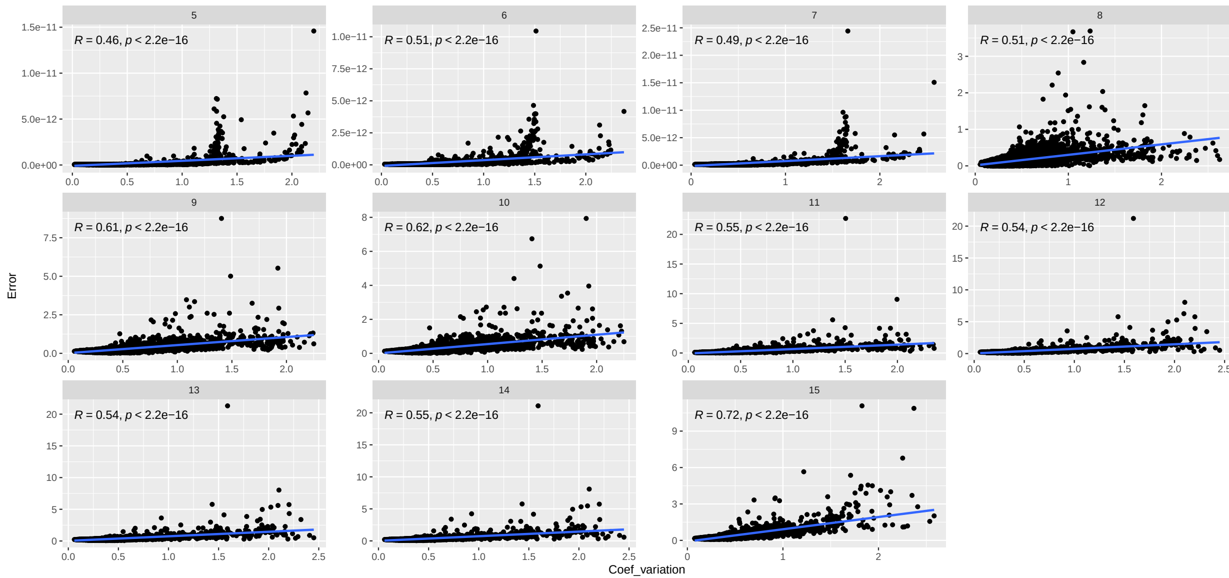
To explore the relationship between properties of the genes and their corresponding predicted values, for each gene we compute the following measures:

- ***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.

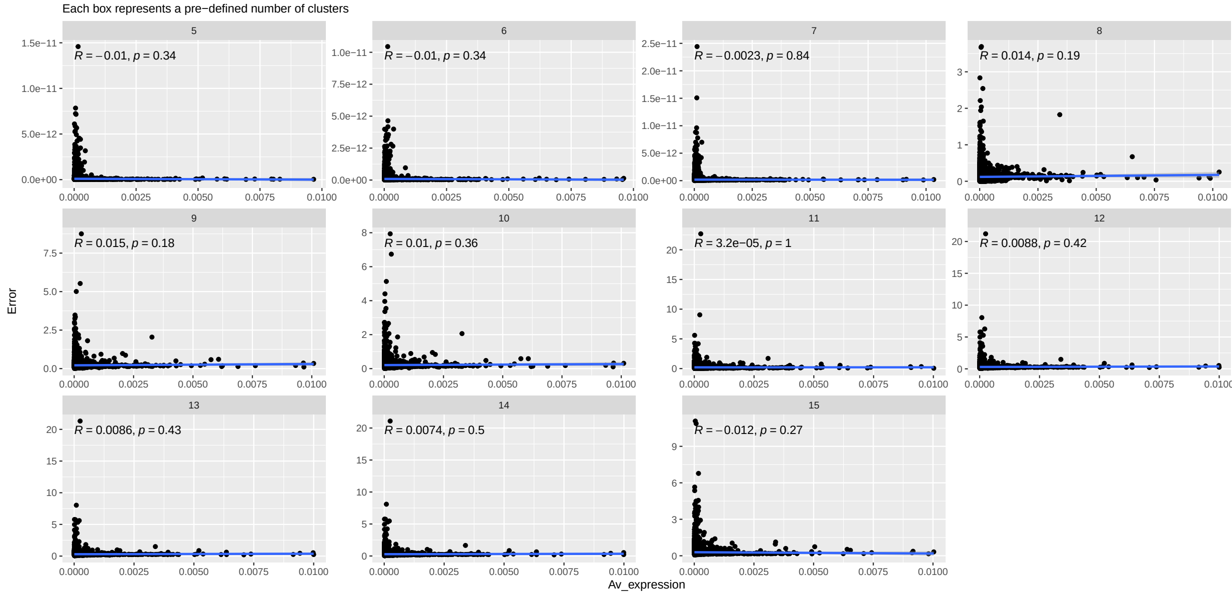
Relationship between a gene's cluster specificity and prediction errors

Each box represents a pre−defined number of clusters



Relationship between a gene's average expression and prediction errors

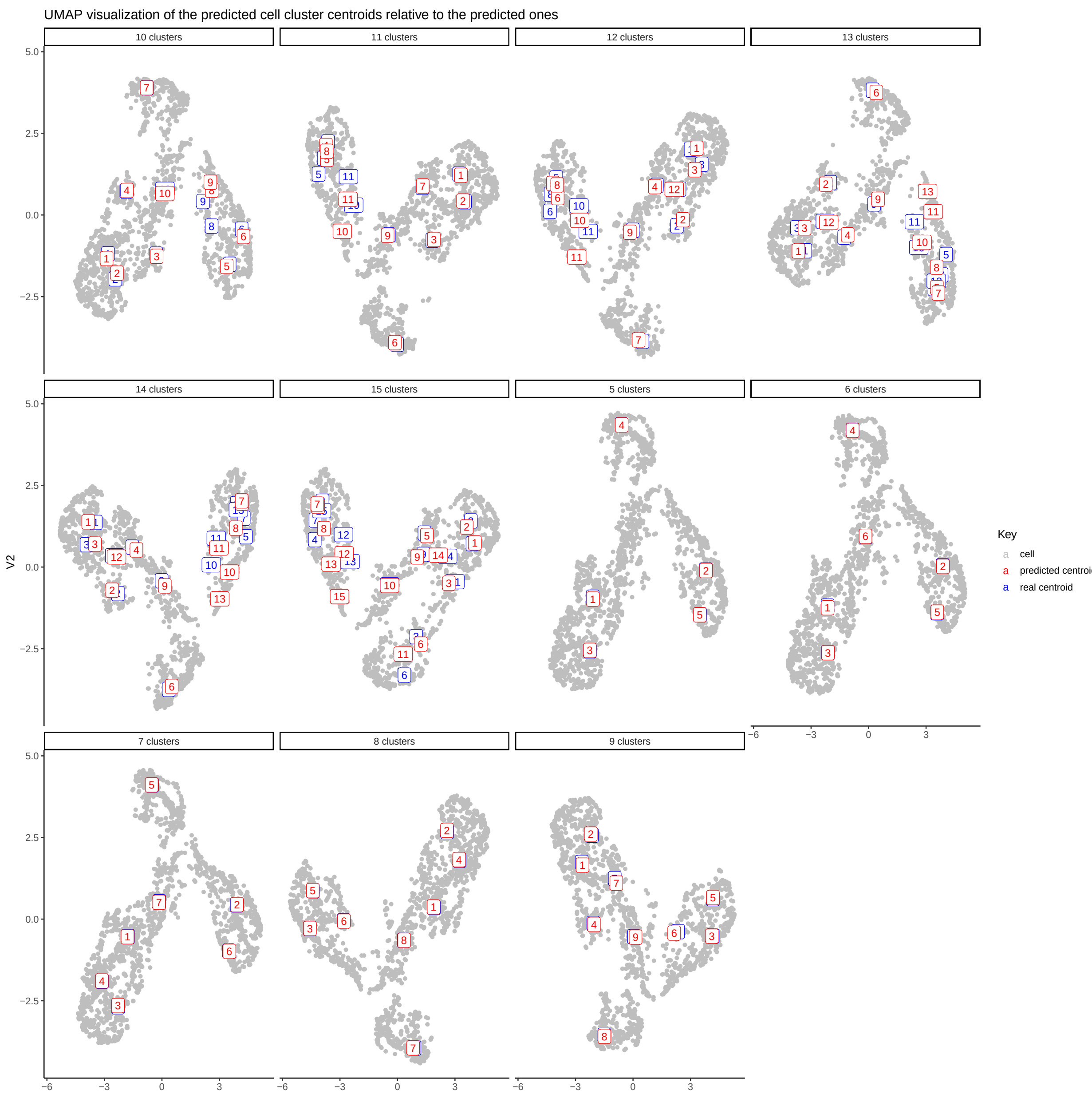
Each box represents a pre−defined number of clusters



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 ($Y = AX$).

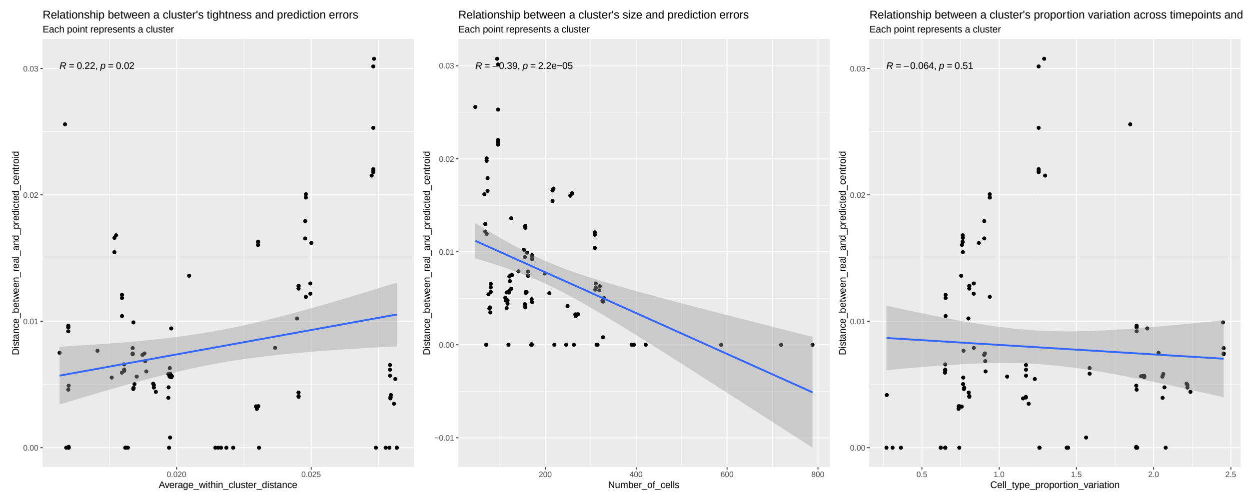
Are our predicted cluster centroids close to the real ones? The cluster centroid is each cluster's average gene expression profile. Visualizing the two sets of centroids (real & predicted) side-by-side on a *UMAP* will help us answer this question. For each number of clusters between 5 and 15, we cluster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

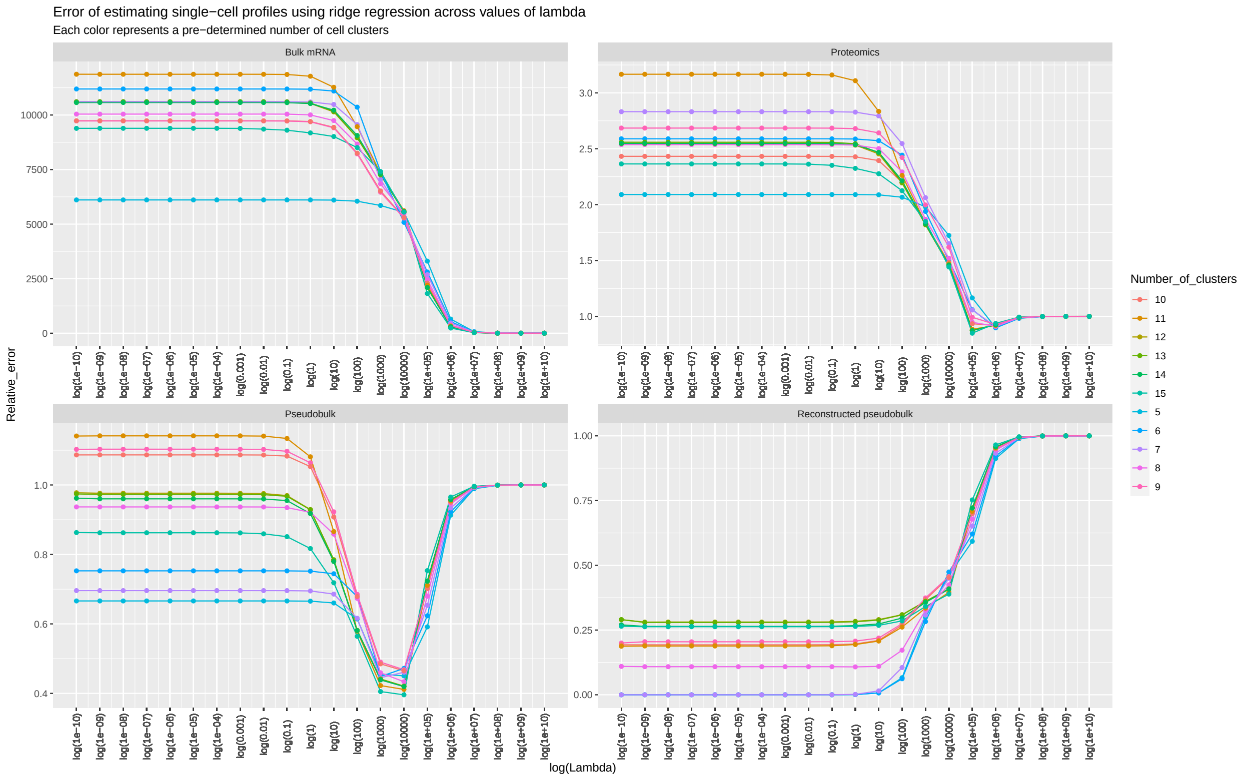


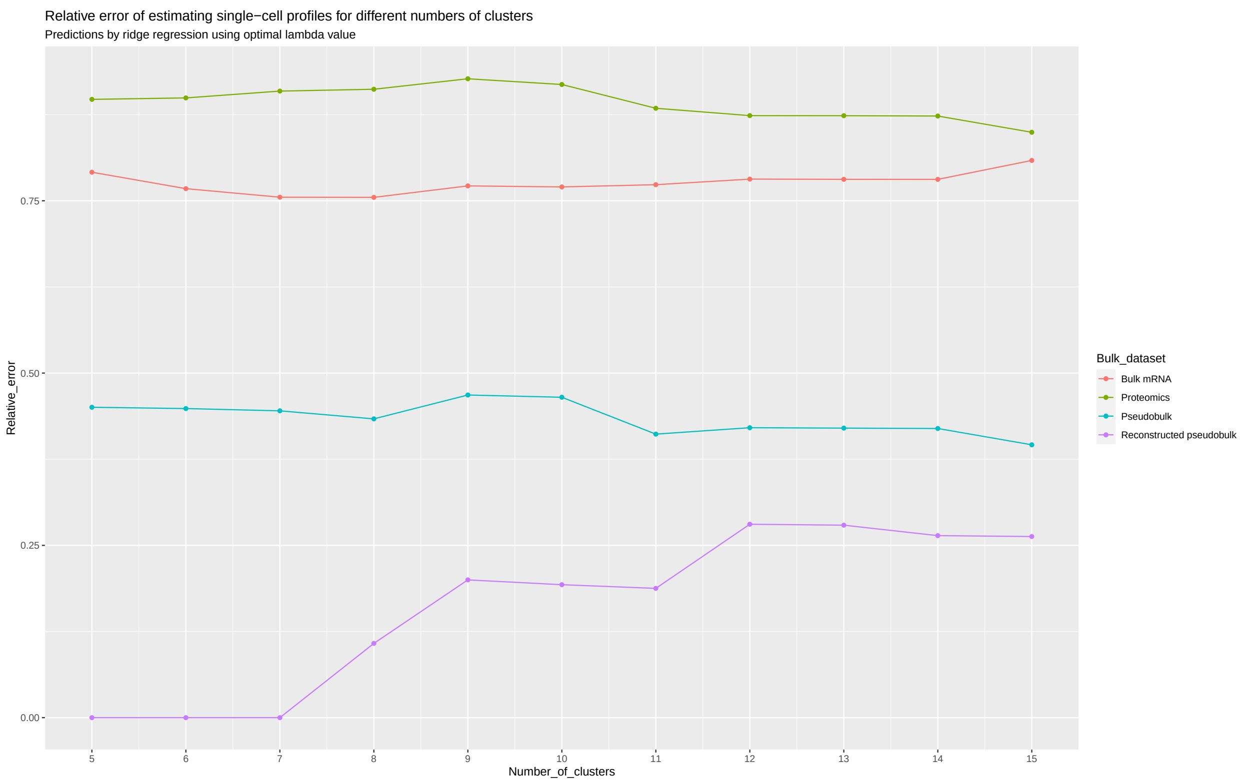
Algorithm Results

The previous sections in this report optimized the deconvolution algorithm by using only the single-cell RNA-Seq data. This section applies the resultant deconvolution algorithm to the following 4 bulk datasets:

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 each timepoint, i.e. $Y = AX$.
3) ***Bulk mRNA***: Bulk mRNA data from a microarray experiment.
4) ***Proteomics***: Bulk proteomics data from a mass spectrometry experiment.

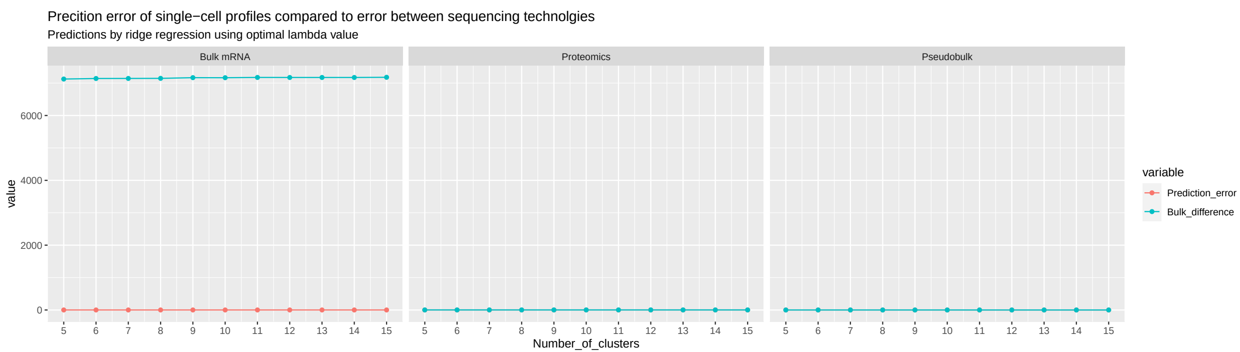
For each of the above datasets, we apply the deconvolution algorithm and compute the prediction error as the relative error between our predicted cell-cluster matrix X' and the one obtained from the Seurat clustering of the scRNA data X. Since the identified genes differ across the three technologies, we retain only the ~5,000 genes common to all datasets for this comparison.


Error of estimating single−cell profiles using ridge regression across values of lambda
Each color represents a pre−determined number of cell clusters
Bulk mRNA
Proteomics
Pseudobulk
Reconstructed pseudobulk
Relative_error
log(Lambda)
Number_of_clusters
10
11
12
13
14
15
5
6
7
8
9


Relative error of estimating single−cell profiles for different numbers of clusters
Predictions by ridge regression using optimal lambda value
Relative_error
Number_of_clusters
Bulk_dataset
Bulk mRNA
Proteomics
Pseudobulk
Reconstructed pseudobulk

Accounting for differences in experimental technologies

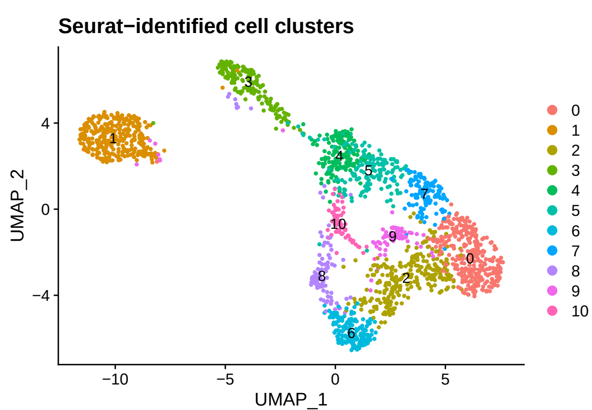
The errors computed in the previous section compared the predicted single-cell profiles to those derived from scRNA-Seq without accounting for the fact that the bulk mRNA and proteomics data were obtained from different technologies, namely microarray and mass spectrometry, which introduce their own technical noise. Indeed, even the pseudobulk that was reconstructed from the scRNA-Seq data after clustering differs from the actual summed-up measurements from the raw data. To account for these differences, we find the relative error between each of the bulk datasets and the "reconstructed" pseudobulk that was used to optimize the algorithm. By plotting that difference on the same plot as the prediction errors, we can place the prediction errors within the context of the underlying experimental differences. For example, in the case of proteomics it would correspond to adjusting the prediction error based on the discrepancy between the protein and RNA levels.

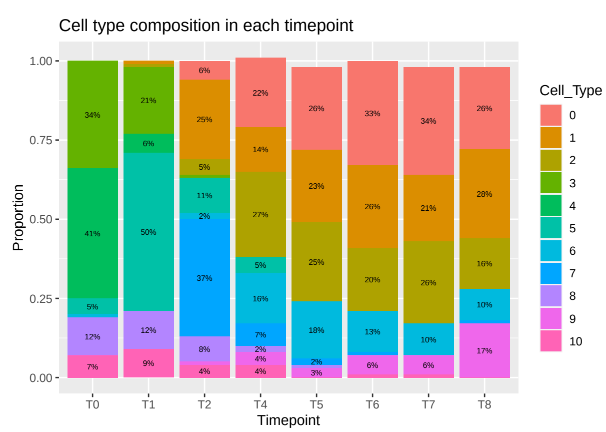


Deconvolution of proteomics data

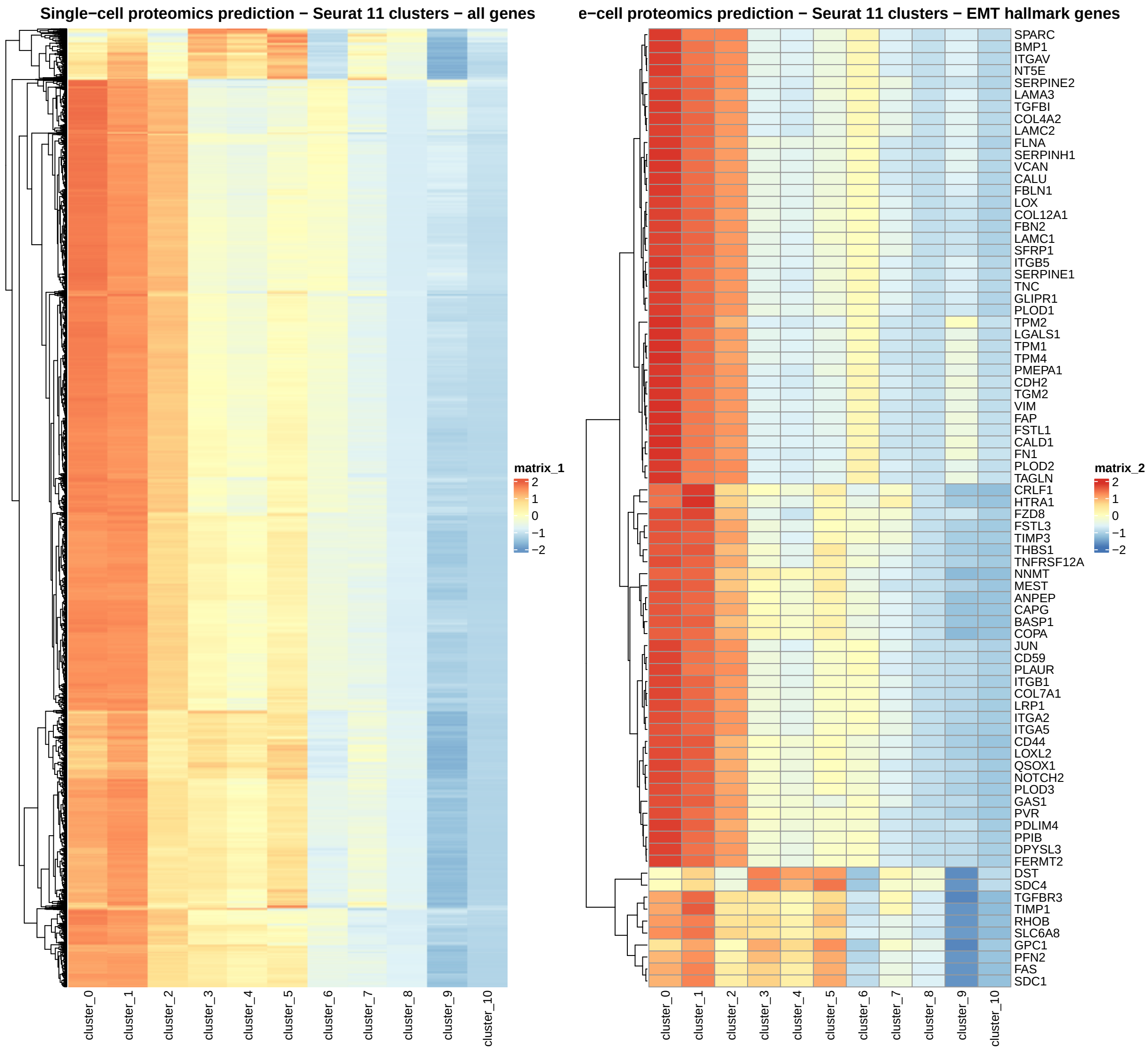
In this section of the report, we focus on deconvoluting the bulk proteomics data with 11 Seurat-defined clusters. We furthermore focus on a set of 80 hallmark genes associated with EMT from the MSigDB database.

Seurat clustering results



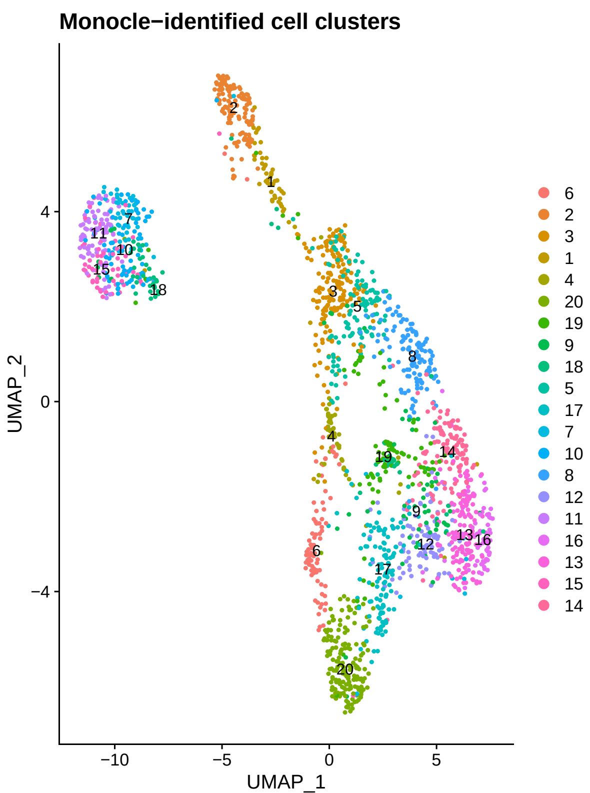


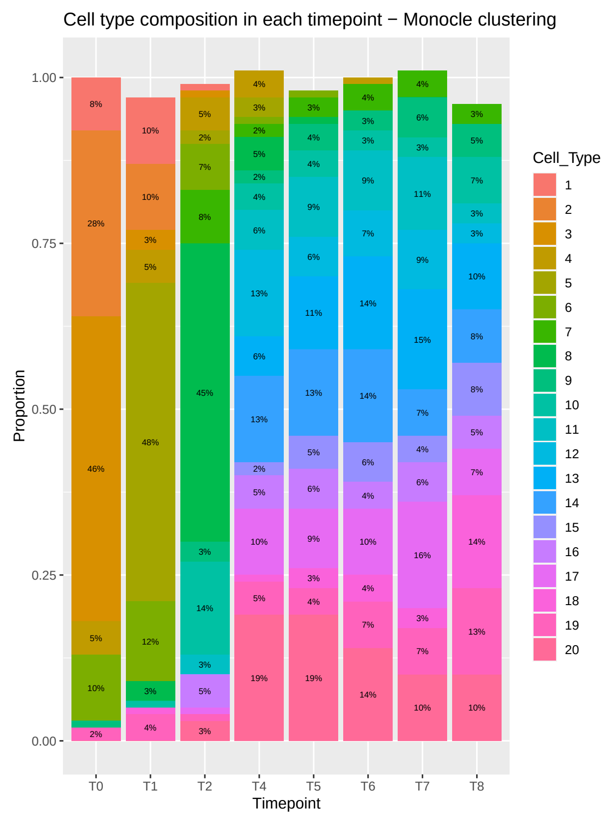
Single-cell proteomics predictions

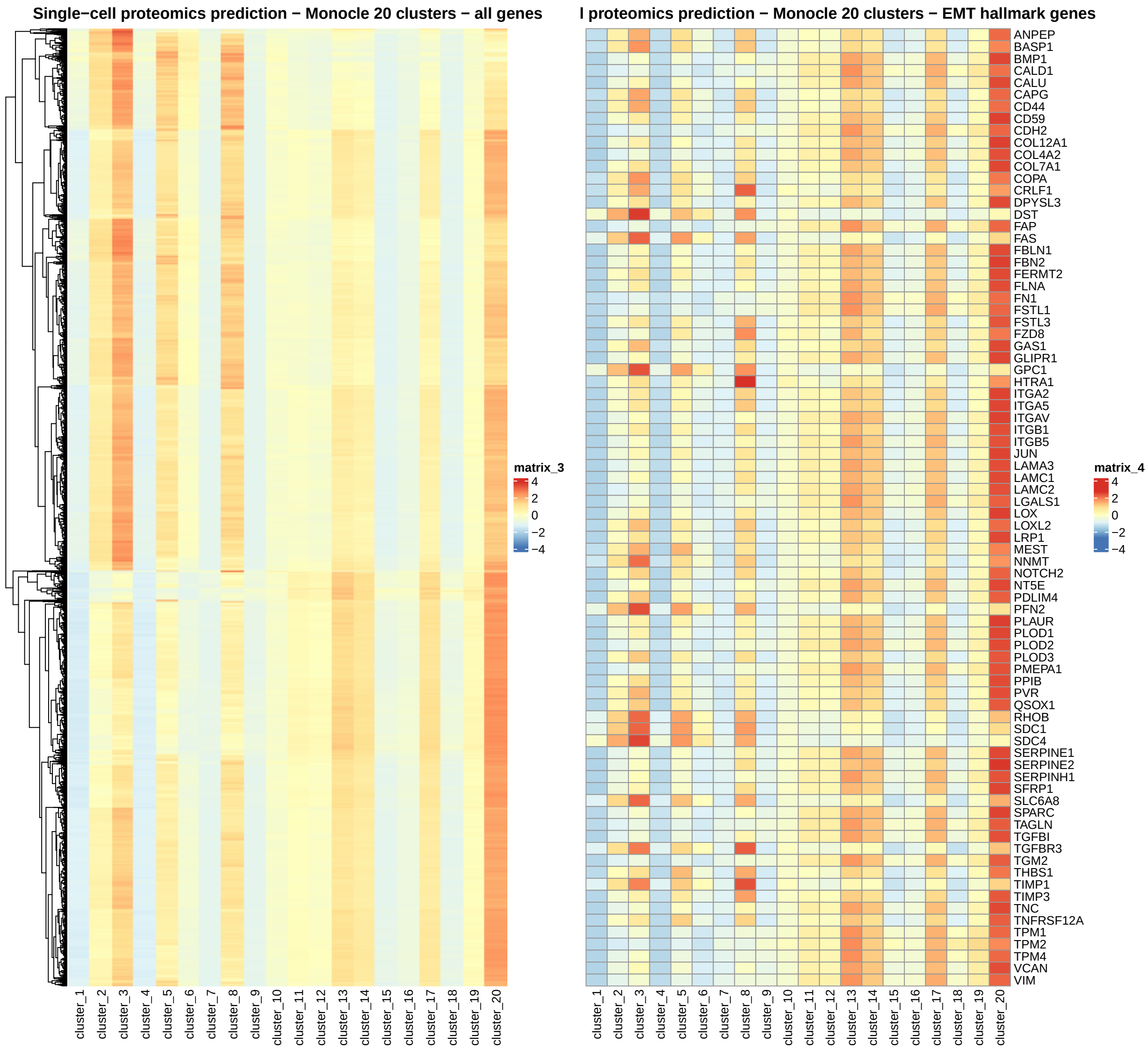

Single−cell proteomics prediction − Seurat 11 clusters − all genes
matrix_1
2
1
0
−1
−2
cluster_0
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
cluster_6
cluster_7
cluster_8
cluster_9
cluster_10
e−cell proteomics prediction − Seurat 11 clusters − EMT hallmark genes
matrix_2
SPARC
BMP1
ITGAV
NT5E
SERPINE2
LAMA3
TGFBI
COL4A2
LAMC2
FLNA
SERPINH1
VCAN
CALU
FBLN1
LOX
COL12A1
FBN2
LAMC1
SFRP1
ITGB5
SERPINE1
TNC
GLIPR1
PLOD1
TPM2
LGALS1
TPM1
TPM4
PMEPA1
CDH2
TGM2
VIM
FAP
FSTL1
CALD1
FN1
PLOD2
TAGLN
CRLF1
HTRA1
FZD8
FSTL3
TIMP3
THBS1
TNFRSF12A
NNMT
MEST
ANPEP
CAPG
BASP1
COPA
JUN
CD59
PLAUR
ITGB1
COL7A1
LRP1
ITGA2
ITGA5
CD44
LOXL2
QSOX1
NOTCH2
PLOD3
GAS1
PVR
PDLIM4
PPIB
DPYSL3
FERMT2
DST
SDC4
TGFBR3
TIMP1
RHOB
SLC6A8
GPC1
PFN2
FAS
SDC1

Alternative clustering

Here, we examine an alternative clustering of the data, which clustered the scRNA data into 20 clusters using the Monocle3 algorithm. Note: the regression parameter λ chosen was the same one used for the Seurat clustering in the previous section.





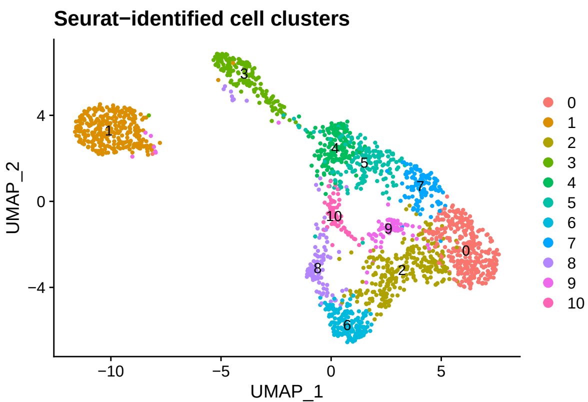

Single−cell proteomics prediction − Monocle 20 clusters − all genes
l proteomics prediction − Monocle 20 clusters − EMT hallmark genes
matrix_3
4
2
0
−2
−4
matrix_4
4
2
0
−2
−4
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
cluster_6
cluster_7
cluster_8
cluster_9
cluster_10
cluster_11
cluster_12
cluster_13
cluster_14
cluster_15
cluster_16
cluster_17
cluster_18
cluster_19
cluster_20
ANPEP
BASP1
BMP1
CALD1
CALU
CAPG
CD44
CD59
CDH2
COL12A1
COL4A2
COL7A1
COPA
CRLF1
DPYSL3
DST
FAP
FAS
FBLN1
FBN2
FERMT2
FLNA
FN1
FSTL1
FSTL3
FZD8
GAS1
GLIPR1
GPC1
HTRA1
ITGA2
ITGA5
ITGAV
ITGB1
ITGB5
JUN
LAMA3
LAMC1
LAMC2
LGALS1
LOX
LOXL2
LRP1
MEST
NNMT
NOTCH2
NT5E
PDLIM4
PFN2
PLAUR
PLOD1
PLOD2
PLOD3
PMEPA1
PPIB
PVR
QSOX1
RHOB
SDC1
SDC4
SERPINE1
SERPINE2
SERPINH1
SFRP1
SLC6A8
SPARC
TAGLN
TGFBI
TGFBR3
TGM2
THBS1
TIMP1
TIMP3
TNC
TNFRSF12A
TPM1
TPM2
TPM4
VCAN
VIM

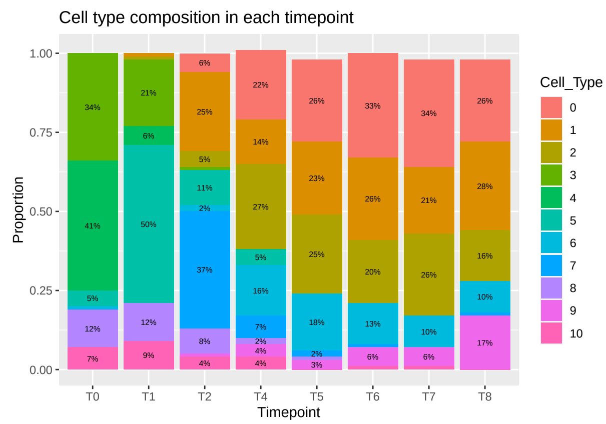
Deconvolution of proteomics data - tuning lambda to individual genes

In this section of the report, we once again focus on deconvoluting the bulk proteomics data with 11 Seurat-defined clusters, with a focus on a set of 80 hallmark genes associated with EMT from the MSigDB database. This time, we fine-tune the regression parameter λ to individual genes.

For each gene, we test the prediction accuracy of 21 values of λ (ranging from $1e^{-10}$ to $1e^{10}$) for deconvoluting pseudobulk RNA data to the single-cell level, with the scRNA data being the benchmark value. The value of λ that minimized the error to the scRNA profiles is retained as the optimal value.

Seurat clustering results





Single-cell proteomics predictions