Prof. Dr. Manfred Claassen
Matthias Bruhns & Jan Schleicher
Machine Learning for
Single Cell Biology 2024

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Assignment 3

In the previous assignments, we explored the concept of cell type as a pattern of co-expression of genes. In this assignment, we will focus on identifying genes that are differentially expressed between cell types.

The expression of individual genes that are known to be specifically expressed in cell types is often used to identify them. These genes are referred to as markers.

However, to fully understand the functional differences between often closely related cell types, it is important to identify sets of genes that are differentially expressed between cell types.

Task 1 Choosing cell type markers. (30 P)

For this task, you have been provided with the following expression dataset and cell type annotations. The dataset was normalized such that each cell had the same total expression. Subsequently, each entry of the normalized expression matrix was transformed like so: log(count+1).

Expression data: *expression_data_1.txt*

Cell type labels: *celltype_data_1.txt*

A. Plot a heatmap of mean gene expression per cell type of the following genes: (10 P)

CCR7, CD14, IL7R, S100A4, MS4A1, CD8A, FCGR3A, MS4A7, FCGR3A, NKG7, FCER1A, PPBP

B. Based on the heatmap (Task 1A) identify one gene per cell type that could serve as a cell type marker. Explain your reasoning. (10 P)

C. Project the expression data into 2D using UMAP and plot each marker on the projection. (10 P)

Task 2 Differential gene expression. (25 P)

For this task, use the expression data and cell type labels from Task 1.

A. For each cell type and each gene perform differential expression testing using the Mann-Whitney U test. Use the provided python function. (10 P)

B. Based on the list of differentially expressed genes per cell type, what biological functions can you infer for each cell type? Briefly explain your reasoning/methodology. (15 P)

Task 3 Identifying differential gene sets using factor analysis. (45 P)

For this task, use the expression data and cell type labels from Task 1.

Use the code provided to perform factor analysis with 3 factors.

A. Plot the factor activity for each factor on the 2D UMAP generated in Task 1C. What can you infer about the relationship between the cell types and factors based on these plots? (10 P)

B. Extract the gene loadings on each factor and perform gene set enrichment analysis. What inference can you make about the cell types based on this analysis? (35 P)