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

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Assignment 1

For this exercise, you have been provided several gene expression datasets generated by single-cell mRNA sequencing. While we will discuss this technology in greater detail later in the lecture, it is not necessary to understand its technological details here. You find an overview of single-cell mRNA sequencing here.

You will be required to analyze the data and generate the indicated output. You will be graded both on the correctness of your analysis and a detailed explanation of how the analysis was performed. For some tasks, additional data has been provided and will be necessary to successfully complete the exercise.

Task 1 – Identifying and annotating cells (60 points)

In the lecture, we discussed the occurrence of various cell types and states in tissues of multicellular organisms. In this task, we are going to investigate the cell state composition of immune cells in the context of chronic infections. Specifically, you are expected to identify the cell type and organ of origin of each cell based on the following paper:

Ioana Sandu et. al. Landscape of Exhausted Virus-Specific CD8 T Cells in Chronic LCMV Infection, Cell Reports, Volume 32, Issue 8, 2020, 108078, ISSN 2211-1247, https://doi.org/10.1016/j.celrep.2020.108078.

Expression data has been provided in the Matrix Market File Format. Cell-level annotations have been provided as tab-delimited files. For each task use the skeleton code provided to output the results of your analysis in the required format.

A. For this subtask, we will work with a partially analyzed dataset: (45 points)

gene expression data – *expression_data_1.mtx*
annotations file – *expression_data_1_metadata.tsv.*

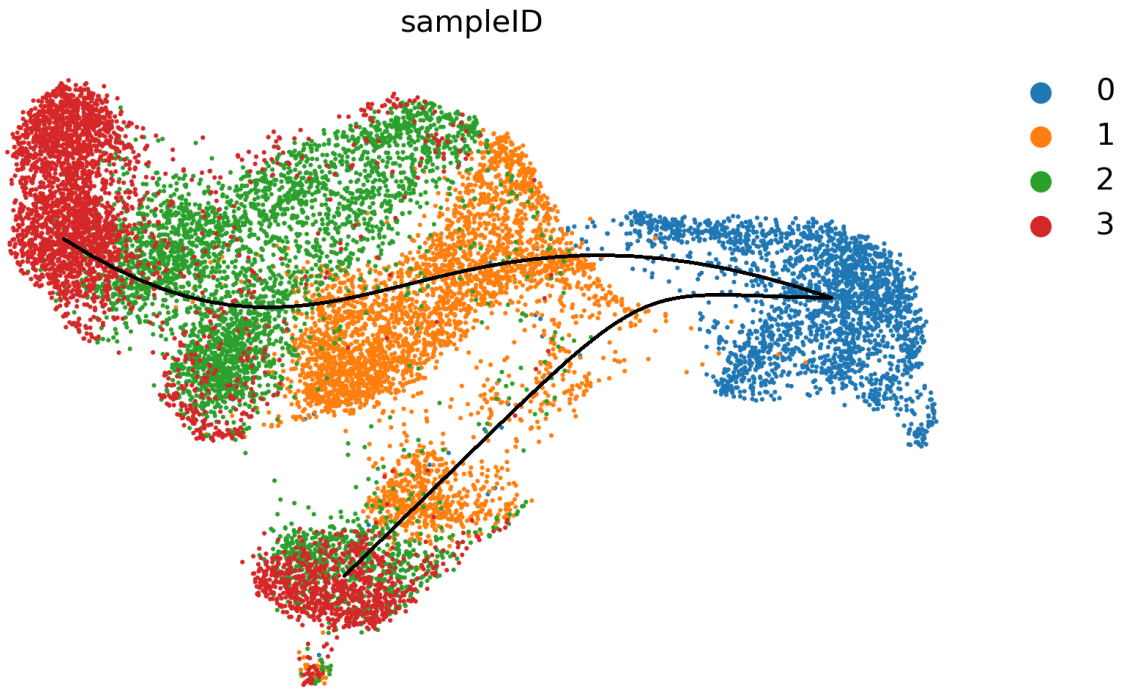
The cells in this dataset have already been grouped by organ (*new_sampleID*) and cell type (*new_pheno*); however, the identity of each group is still unknown. Based on your reading of the paper, identify the cell types and the organs. Give a detailed account of your analysis and indicate the content of the paper that was instrumental in generating your results. To facilitate automatic evaluation, please use the cell type names (E, Exh, I, M, and P) and organ names (blood, BM, liver, LN, lung, spleen) as specified in the paper.

B. Next, we are going to analyze data of a heterogeneous group of cells, i.e., the cells have not been grouped. The task is to annotate each cell with its cell type (E, Exh, I, M, or P). For this subtask, use gene expression data – *expression_data_2.mtx.* (15 points)

Task 2 – Analyzing time series scRNAseq data (40 points)

The lecture introduced dynamic single-cell processes. Here we are going to investigate the differentiation of T cells in the context of chronic infection. The data was generated in a time series experiment where each cell belongs to one of four samples (*sampleID*) that form a temporal sequence. The cells in the dataset have a common origin but evolve towards two distinct terminal states, thus forming a branching topology of two trajectories. In addition, cells have also been clustered using the Louvain algorithm (*louvain*).

A. You have been provided 4000 sequences of cells that represent the temporal evolution of cell states (*tex_sampling.npy*), each sequence consisting of 50 cells. Use the data provided (*expression_data_3_metadata.txt*) to identify the sequence of Louvain clusters that compose either trajectory. The figure below illustrates the two trajectories and annotations described above. Each dot in this figure is a two-dimensional projection of its gene expression profile generated using UMAP. (20 points)



B. Next, we will investigate which genes are relevant to the temporal evolution of cell states. We will do this by means of a correlation analysis. Use the *expression_data_3.mtx* to investigate the correlation (spearman correlation) of each gene's expression with sample ID.

Plot your results as a volcano plot where the x-axis is the correlation and the y-axis is the negative of log-transformed p-values. Also, report the top ten correlating genes (with respect to absolute correlation). (20 points)