Single Cell Bioinformatics 2024-25

Jun-Prof. Dr. Fabian Müller, Dr. Fabian Kern & Dr. Thomas Hentrich
Irem Gündüz, Midhuna Joseph Maran
Müller Lab, Kern Lab, and Schulze-Hentrich Lab @ Saarland University

Project 1- scRNAseq Worksheet

Deadline: 15.11.2024, 23:59

This project will focus on key methods for analyzing single-cell RNA sequencing (scRNA-seq) data, as covered in the lectures. The dataset comes from a project on human bone marrow cells and CD34+ enriched bone marrow cells, incorporating scATAC-seq, CITE-seq, and scRNA-seq. The dataset was originally analyzed by Granja et al. (2019).

You may work in pairs for this project. To qualify for the final exam, you must earn at least 50% of the total points across all assignments. If you have any questions, please get in touch with Irem or Midhuna via Teams, by email (irembgunduz@gmail.com, mijo00001@stud.uni-saarland.de), or by visiting the tutorials.

Submission:

1. You will have to submit one tar.gz file that includes:
   - The code as an R script
   - PDF file containing all images and responses to the questions
2. The code must be well documented and must run without an error to obtain any points,
3. If you use any additional sources for your answers, make sure to include proper references.

Programming:

1. The programming should be done in R using only the named packages

Introduction to Seurat:

In the project, you will predominantly use `Seurat`, an R package that combines many functionalities for the analysis of single-cell data. For the initial steps, you can find helpful documentation here. They will be explicitly mentioned if other packages are required for specific tasks.

Before you start programming, you should set up the system as follows:

System-setup:

1. Install Conda
2. Install all packages using the provided `environment.yml` file. [Mac users: delete `singleR` line from the .yml file before running the following command]
   ```
   conda env create -f environment.yml
   ```
3. Start the conda environment with:
   ```
   conda activate single-cell
   ```
4. Install CellChat by starting R and install CellChat using devtools
   ```
   devtools::install_github("sqjinCellChat")
   ```
   If you are using a Mac, delete the line for `singleR` from `environment.yml` file and manually install `SingleR` using the following commands in R:

```
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("SingleR")
```

5. If you have problems with Seurat's clustering functions, please try to downgrade Matrix and spatstat.core packages.

```
remove.packages(grep("spatstat", installed.packages(), value = T))
devtools::install_version("spatstat", version = "2.4.4")
install.packages("Matrix", ".", type = "source", repos = "http://R-Forge.R-p
```

6. Test if you have installed all necessary libraries:

```
suppressPackageStartupMessages({
 library(dplyr)
 library(spatstat.core)
 library(Seurat)
 library(patchwork)
 library(DoubletFinder)
 library(SingleR)
 library(enrichR)
 library(CellChat)
 library(SingleCellExperiment)
 library(SeuratWrappers)
 library(tidyverse)
 library(monocle3)
 library(celldex)
 })
```

Download the data:

You can download the dataset for this project under the following link: https://icbb-share.s3.eu-central-1.amazonaws.com/single-cell-bioinformatics/scbi_ds1.zip The file contains the data of four samples: BMMC_D1T1, BMMC_D1T2, CD34_D2T1 and CD34_D3T1, with separate expression matrices for each sample.

Week 1: (5 Points)

1 Loading the Data (1P)

Load the expression matrices from the dataset and construct a Seurat object. You will need to load two files: one containing data on Bone Marrow Mononuclear Cells (BMMC) and the other on CD34+ Enriched Bone Marrow Cells (CD34).

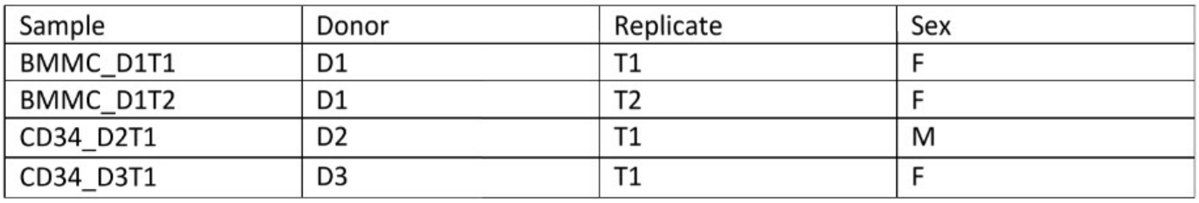
2 Create the sample sheet (1P)

Label each sample with the corresponding metadata from Table 1:

3 Add Meta-data (3P)

For each sample report the following information:

1. How many cells are in each sample?
2. How many genes are in the expression matrices?
3. What information is now part of the meta-data of the objects?



Sample	Donor	Replicate	Sex
BMMC_D1T1	D1	T1	F
BMMC_D1T2	D1	T2	F
CD34_D2T1	D2	T1	M
CD34_D3T1	D3	T1	F

Figure 1: Table 1

Hint:

You will find all methods that are necessary to solve these tasks in one of the following vignettes/documentation(s):

Seurat: Guided Tutorial, Seurat Command List

Consider saving your processed data as a Seurat object using the command:

```
saveRDS(data, file =filename)
```

You can read that object with

```
data <- readRDS(file=filename)
```

Week 2: (15 Points)

4 Preprocessing (10P)

4.1 Preprocessing (7P)

Bring the following preprocessing steps into the correct order and perform them on your data.

1. Filtering
2. Doublet removal (`DoubletFinder`)
3. Normalization
4. Feature Selection

Which steps do you perform before and after merging (task: 4.2) and why?

Filtering: Perform Filtering on the data to remove low-quality cells. *Name the parameters that have been used for filtering. Argue why you have used them and how you have chosen the cut-off parameters.*

Doublet-Removal with DoubletFinder: Estimate the optimal value for pK and perform the Doublet Detection. *Explain why we perform doublet removal.*

Normalization and Feature Selection: Perform Normalization and Feature Selection on your data. *Which Normalization method is used by the Seurat Normalization function by default? What is the purpose of Feature Selection? How are they selected?*

4.2 Batch-Correction (3P)

1. Merge all four samples into one dataset without performing batch correction.
2. Merge all four samples into one dataset using Seurat's Data Integration method as Batch Correction.

Compare the outcomes of A and B.

Is Batch-Correction necessary? If yes, name the parameters and explain (with the necessary plots) why a correction for this parameter may be necessary.

5 Dimensionality Reduction: (5P)

5.1 Dimensionality Reduction

Perform dimensionality reduction using PCA followed by UMAP, and plot the data in 2-dimensional space.

How did you choose the number of dimensions? Use a plot to explain.
Explain why we use a combination of PCA with UMAP for clustering and not only one of the methods.

5.2 Clustering

Cluster the embedded data and display the results in a two-dimensional plot. Keep the results of the dimensionality reduction and clustering for the subsequent tasks. You should end up with 7-15 clusters.

Hint:

You will find all methods that are necessary to solve these tasks in one of the following vignettes/documentation(s):

Seurat: Guided Tutorial, Seurat Command List, Data Integration
DoubletFinder: Tutorial

Week 3: (20 Points)

6 Cell Type Annotation: (10 P)

6.1 Automatic Annotation (2P)

Using `SingleR`, a tool for automatic cell type annotation, assign cell type labels to your data.
Use the built-in reference "HumanPrimaryCellAtlasData" from the `celldex` package. Plot the results of the automatic annotation as a UMAP plot.

6.2 Manual Annotation (8P)

1. Perform differential expression analysis for the cell-type annotation and determine the differentially expressed genes.
2. Use the markers from Table 2 to identify the cell types present in the dataset and assign a name to each cluster. Each cluster should be labelled with a unique identifier that includes the corresponding cell-type abbreviations in brackets.
3. Plot the result of the cell-type annotation in a UMAP plot. Compare the results of the automatic annotation and the manual annotation.
4. Show the gene expression of 3 marker genes in the different clusters using a Violin plot and in the different cells as a UMAP Plot.

You will use the results of the manual annotation for the subsequent tasks. Therefore, treat all clusters of the same cell-type as a single cluster.

6.3 Cell-Type Proportions (Bonus: Max 3P)

1. Compute the cell-type proportions for each sample. Plot the cell-type proportions as a bar plot.
2. Explain how samples vary in terms of cell-type proportions.

7 Differential Expression Analysis (4P)

7.1 Differential Expression Analysis on cell-types

Compare the following groups by performing a differential expression analysis and show the results as a volcano plot.

- B cells vs T cells
- T-cells vs Monocytes

7.1.1 Memory Formation on Cells (Bonus: Max 2P)

How do naive T cells differ from memory T cells? Do a quick search on memory formation and give a 1-2 sentence explanation.

7.2 Plot Differentially expressed genes

Show a comparison of the top 5 differentially expressed genes for each comparison. Plot the cell types on the x-axis, and the genes on the y-axis and use the significance as size and the Fold-change as colour of the dots.

8 Pathway Analysis (5P)

8.1 Differential Expression Analysis on groups (2P)

Compare the BMMC data with the CD34 data by performing a differential expression analysis independent of the cell types. Report the top 5 DEGs with p-value and Fold change.
Do the same analysis again but compare the two groups for Monocyte cells separately.

8.2 Pathway analysis on groups (2P)

Do a pathway analysis for GO terms for the comparison between the BMMC and the CD34 data.
(You can use Seurat's `DEenrichRPlot` Function for the pathway analysis)

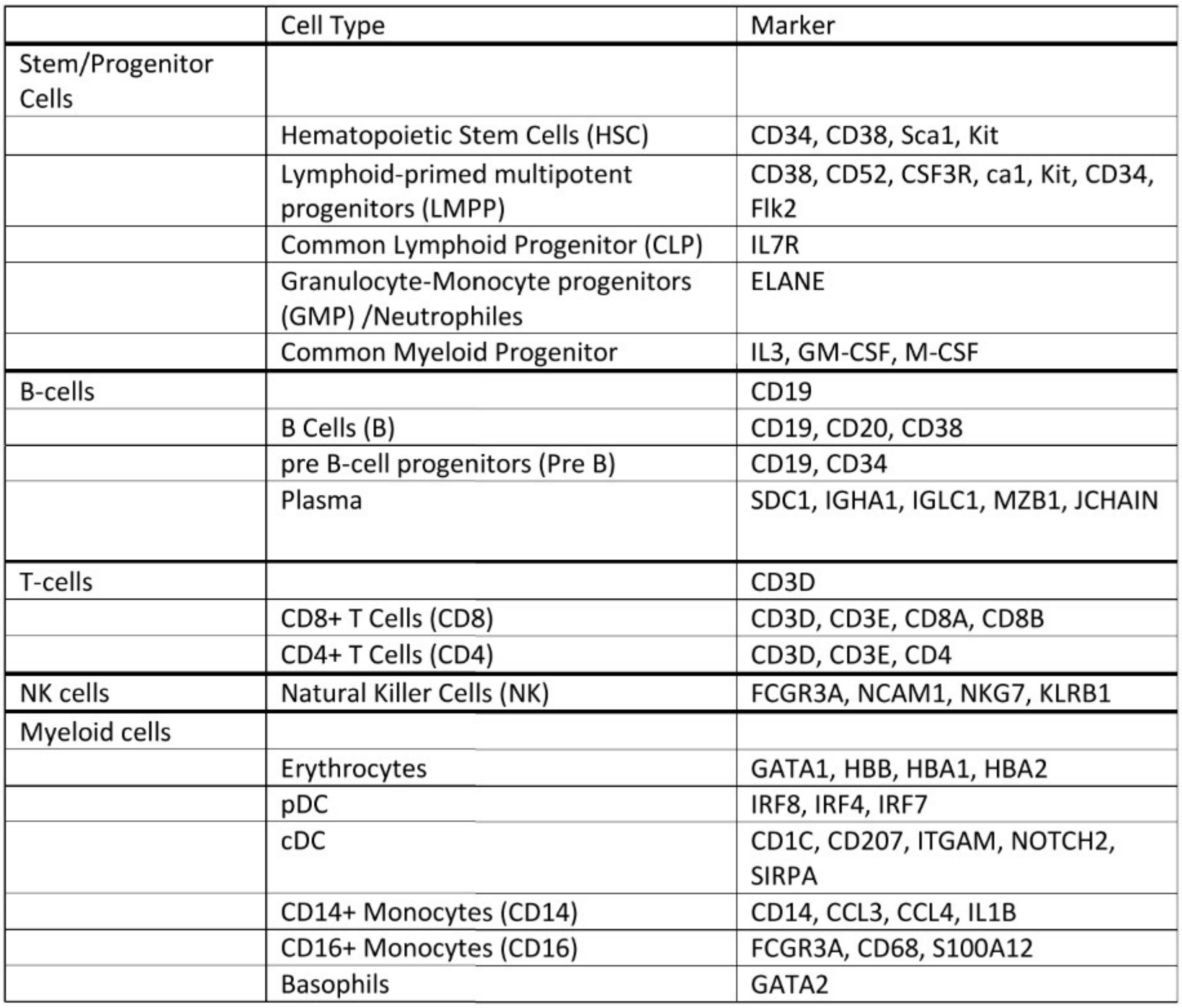
8.3 Biological interpretation (1P)

Name the pathway with the lowest p-value. Explain its biological meaning.

Hint:

You will find all methods that are necessary to solve these tasks in one of the following vignettes/documentation(s):

Seurat: Guided Tutorial, Seurat Command List, EnrichR
singleR: Tutorial



	Cell Type	Marker
Stem/Progenitor Cells		
	Hematopoietic Stem Cells (HSC)	CD34, CD38, Sca1, Kit
	Lymphoid-primed multipotent progenitors (LMPP)	CD38, CD52, CSF3R, ca1, Kit, CD34, Flk2
	Common Lymphoid Progenitor (CLP)	IL7R
	Granulocyte-Monocyte progenitors (GMP) /Neutrophiles	ELANE
	Common Myeloid Progenitor	IL3, GM-CSF, M-CSF
B-cells		CD19
	B Cells (B)	CD19, CD20, CD38
	pre B-cell progenitors (Pre B)	CD19, CD34
	Plasma	SDC1, IGHA1, IGLC1, MZB1, JCHAIN
T-cells		CD3D
	CD8+ T Cells (CD8)	CD3D, CD3E, CD8A, CD8B
	CD4+ T Cells (CD4)	CD3D, CD3E, CD4
NK cells	Natural Killer Cells (NK)	FCGR3A, NCAM1, NKG7, KLRB1
Myeloid cells		
	Erythrocytes	GATA1, HBB, HBA1, HBA2
	pDC	IRF8, IRF4, IRF7
	cDC	CD1C, CD207, ITGAM, NOTCH2, SIRPA
	CD14+ Monocytes (CD14)	CD14, CCL3, CCL4, IL1B
	CD16+ Monocytes (CD16)	FCGR3A, CD68, S100A12
	Basophils	GATA2

Figure 2: Table 2

Week 4 (10 Points)

9 Trajectory Analysis (5P)

9.1 Select subset

Select a group of cells that may consist of one or more clusters that you find interesting for trajectory analysis. Use `Monocle 3` to perform the trajectory analysis, and plot only this group of clusters in a UMAP.
Why is this a good group to do trajectory analysis? Which other group do you think may be a good choice?

9.2 Select root-nodes manually (2P)

Select root nodes manually and use `Monocle 3` to perform trajectory analysis on the data and plot the pseudo-time of the cells. Shortly explain the result you see in the plot.
Why is the selection of the root nodes important for the algorithm?
Which points are a good choice for root nodes of the analysis and why?

9.3 Select root-nodes automatically (2P)

Try to Select root nodes automatically. Did it improve the results? Explain why.
Choose one path in the trajectory and explain which cells are located on this path (I.e., the biological meaning)

10 Cell-Cell Communication (5P)

For this task, you should use only the cell types that occur both in the BMMC and the CD34 samples.

Use `CellChat` to study the cell-cell communication between the different cell types:

- In BMMC samples
- In CD34 samples

Find the signalling pathways that can be found in both groups. Show the number of interactions and the interaction strength for each group.
Choose one pathway, and display the results in a circle plot for each group (BMMC).

11 Summary (Bonus + 5P)

Write a summary of the analysis you've done (max. 200 words). You can also include a short outlook on alternative methods for analyzing the data or other approaches to analyse these cells.

Hint:
You will find all methods that are necessary to solve these tasks in one of the following vignettes/documentation(s):

Seurat: Guided Tutorial, Seurat Command List, EnrichR
Monocle 3: Tutorial
CellChat: Tutorial