Single-Cell Bioinformatics 2022/23

Project 1

Integrative Cellular Biology
& Bioinformatics

CHAIR FOR CLINICAL BIOINFORMATICS
SAARLAND UNIVERSITY

This project will cover the most important methods for the analysis of single cell RNA sequencing data that are covered in the lecture. This dataset consists of human cells from the bone marrow and the CD34+ Enriched Bone Marrow Cells. This dataset has been part of a study that combined scATAC-seq, CITE-seq and scRNA-seq, in this project we will analyse the scRNA-seq part.

The paper that analysed the dataset that also included these scRNA-seq data was published by Granja et al. (2019) and the original data can be found here.

Deadline: 25.11.2022 23:59

You are allowed to work in groups of two people.

To be able to attend the final exam, you must achieve at least 50% of all points in the assignments.

If you have any Problems, please contact Irem (irembgunduz@gmail.com) or visit the tutorials.

Submission:

- You will have to submit **one** tar.gz file that includes
  - The code as an R script
  - PDF file including all images and responses to the questions
- The code must be well commented and must run without an error to obtain any points
- If you use any other sources for your answers, don't forget to give the reference

Programming:

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

Introduction in Seurat

The tool that you will mainly use in this project is Seurat. It is a tool that combines many functionalities for the analysis of single cell data. For the start, there is a good documentation here for the first steps of the analysis. If you should use any other package for the analysis, it is explicitly mentioned in the tasks.

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

Instruction to set up the system

Install conda

Install all packages using the environment.yml file. If you are using a Mac, first delete the singleR line from the file.

```
conda env create -f environment.yml
```

Start the conda environment with:

```
conda activate single-cell
```

Install CellChat by **starting R** and install CellChat using devtools:

```
devtools::install_github("sqjin/CellChat")
```

If you are using a Mac, delete the line for singleR from the environment.yml file and manually install SingleR using the following commands in R:

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

Test if you have installed all necessary libraries:

```
library(dplyr)
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:
ccb-web.cs.uni-saarland.de/lecture_material/singlecell/scbi_ds1.zip

The file contains the data of four samples: BMMC_D1T1, BMMC_D1T2, CD34_D2T1 and CD34_D3T1. You will be given the expression matrix for each sample separately.

Week 1: 5 Points

Task 1: Preprocessing

Task 1.1: Loading the data (1P)

Load the expression matrices from the dataset and construct a Seurat object

You will have to load two files that include the data of Bone Marrow Mononuclear Cells (BMMC) and two files that include the data of CD34+ Enriched Bone Marrow Cells.

Task 1.2: Add Meta-data (1P)

Label each sample with the corresponding meta-data from Table 1:

Table 1: Meta Data

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

Task 1.3: Add Meta-data (3P)

For each sample report the following information:

- *How many cells are in each sample?*
- *How many genes are part of the expression-matrix?*
- *What information are now part of the meta-data of the objects?*

Hint:

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

Seurat: Guided Tutorial, Seurat Command List

Consider storing a seurat object with the so far processed data using the command *saveRDS(data, file =filename).* You can read that object with *data <- readRDS(file=filename)*

Week 2: 15 Points

Task 2: Preprocessing (10P)

Task 2.1: Pre-processing (7P)

Bring the following pre-processing steps into the correct order and perform them on your data.

- Filtering
- Doublet-Removal (DoubletFinder)
- Normalization
- Feature Selection

Which steps do you perform before and after merging and why?

Filtering:

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

Doublet-Removal with DoubletFinder:

Make sure you have correctly prepared the data for DoubletFinder. Estimate the optimal value for pK and perform the Doublet Detection.

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 the Feature Selection?

Task 2.2: Batch-Correction (3P)

A) Merge all four samples into one dataset without performing Batch-Correction.
B) 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.

Task 3: Dimensionality Reduction: (5P)

Task 3.1: Dimensionality Reduction

Perform a dimensionality reduction using PCA with UMAP and plot it in the 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.

Task 3.2: Clustering

Do a clustering of the embedded data and display a 2-dimensional plot of the result. Keep the results of the dimensionality reduction and clustering for the next 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 tutorials/documentations:

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

Consider storing a seurat object with the so far processed data using the command *saveRDS(data, file =filename).* You can read that object with *data <- readRDS(file=filename)*

Week 3: 20 Points

Task 4: Cell type Annotation: (11 P)

Task 4.1: Cells in Bone-Marrow Datasets (1P)

Make yourself familiar with the Cell-types from table 2. Draw a tree that shows the differentiation of the different cell-types.

Task 4.2: Automatic Annotation (2P)

Use SingleR, a tool for the automatic cell-type annotation, to determine a cell-type annotation.

Use the build-in reference "HumanPrimaryCellAtlasData" from the celldex package. Plot the results of the automatic annotation in a UMAP plots.

Task 4.3: Manual Annotation (8P)

Task 4.3.1: Do a differential expression analysis for the cell type annotation and determine the genes that are differentially expressed between the clusters.

Task 4.3.2: Use the markers from table 2 to determine which cell-types occur in the dataset and name the cell type for each cluster.
Name each cluster with a unique identifier including the cell type abbreviations given in the brackets and cluster.

Task 4.3.3: Plot the result of the cell type-annotation in a UMAP plot. Compare the results of the automatic annotation and the manual annotation.

Task 4.3.4: Show the gene-expression of three Marker-genes in the different clusters using a Violon-plot and in the different cells using a UMAP-Plot.

Use the results of the manual annotation for the next tasks. Therefore, you should treat all clusters from one cell-type as one cluster.

Task 5: Differential Expression Analysis (4P)

Task 5.1: 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
CD4 T cells vs CD14 Monocytes

Task 5.2: Plot Differentially expressed genes

Show a comparison of the top 5 differentially expressed gene for each comparison.

Therefore, plot the cell types on the x-axis, the genes on the y-axis and use the significance as size and the Fold-change as colour of the dots.

Task 6: Pathway Analysis (5P)

Task 6.1: 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 each cell-type separately.

Task 6.2: Pathway analysis on groups (2P)

Do a pathway analysis for GO terms for the comparison between the BMMC and the CD34 data.

You can either use Seurats *DEenrichRPlot* Function or use Genetrail for the pathway analysis

Task 6.3: 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 tutorials/documentations:

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

Table 2: Celltype Markers

	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

Week 4: 10 Points

Task 7: Trajectory Analysis (5P)

Task 7.1: Select subset

Select a group of cells that may consist of one or more clusters you think might be interesting for trajectory analysis. Use monocle 3 to perform a trajectory analysis. Plot only this group of clusters into an UMAP.

Why is this a good group to do trajectory analysis. Which other group do you think may be a good choice.

Task 7.2: 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?

Task 7.3: Select root-nodes automatically (2P)

Also try to automatically choose the root nodes.
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)

Task 8: 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 the BMMC samples
- in the 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, display the results in a circle plot for each group (CD34 and BMMC) and compare the results.

Task 9: Summary (Bonus + 5P)

Filter out the most important findings of the project and write a short summary (max. 200 words). Given these data what are findings that may be interesting to others. You can also include a short outlook including alternative methods to analyse the data or other methods to study these cells.

Hint:

You will find all methods that are necessary to solve these tasks in one of the following tutorials/documentations:

Seurat: Guided Tutorial, Seurat Command List
Monocle 3: Tutorial
CellChat: GitHub Page with Tutorials