Integrative Cellular Biology
& Bioinformatics

CHAIR FOR
CLINICAL
BIOINFORMATICS
SAARLAND UNIVERSITY

Project 1

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: 26.11.2021 23:59

You are allowed to work in groups of two people.

A good result in the projects will give you a bonus on your final grade.

Tutorials:

- If you have any Problems, please contact Omar (omla00001@stud.uni-saarland.de) or visit the tutorials.
- During the project, there will be tutorials on some of the tasks.

Publishing of the project	25.10.2021	
Tutorial	25.10.2021	Introduction, Seurat and Preprocessing
	01.11.2021	-- Holiday: No Tutorial --
Tutorial	08.11.2021	Dimensionality Reduction, Clustering, Cell type annotation
Tutorial	15.11.2021	Trajectory Analysis
Tutorial	22.11.2021	DEG analysis and pathways
Deadline project	26.11.2021	

Submission:

- You will have to submit **one** tar.gz file that includes
  - The code
  - 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 the named packages

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

Instruction to set up the system

Install conda

When conda is installed, run:

```
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")
```

Test if you have installed all necessary libraries:

```
library(dplyr)
library(Seurat)
library(patchwork)
library(DoubletFinder)
library(SingleR)
library(velocyto.R)
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.

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.

Task 1: Preprocessing: (10P)

Task 1.1: Loading the Data

Load the expression matrices from the dataset
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.

Hint: Load the data, construct a seurat object and familiarize yourself with it. Where is the information such as expression matrix and meta-data stored? How can you access and manipulate the data? You can find a useful list of commands here.

Task 1.2: Pre-processing

Perform the following pre-processing steps and merge all four samples into one dataset:

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

Do not use the Data-Integration Function for Merging as it will be topic in the later projects. Instead use the Merge-Function of Seurat.

Attention: These pre-processing steps should not necessarily be executed in the given order.
Questions:

1. Which steps do you perform before and after merging and why?
2. Name the parameters that have been used for filtering and argument how and why you have used them.
3. If you think a Batch-correction may be necessary, name the parameters and explain why a correction for this parameter may be necessary. Also show the plot you used to check it.

Task 1.3: Add Meta-data

Label each sample with the corresponding metadata from Error! Reference source not found.**.**

Hint: 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)*

Task 2: Dimensionality Reduction: (5P)

Task 2.1: Dimensionality Reduction

Perform a dimensionality reduction using PCA with tSNE and using PCA with UMAP and plot it in the 2-dimensional space.

Questions:

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

Task 2.2: Clustering

Do a clustering of the PCA+UMAP data and display a 2-dimensional plot of the result. Keep the results of the dimensionality reduction and clustering for the next tasks.

Task 3: Cell type Annotation: (10 P)

Task 3.1: Manual Annotation

Task 3.1.1: Do a differential expression analysis for the cell type annotation and determine the genes that are differentially expressed between the clusters.

Task 3. 1. 2: Use a database such as CellMarker to determine the cell type for each cluster. Cell types that may occur in this dataset are:

Hematopoietic Stem Cells (HSC)
Lymphoid-primed multipotent progenitors (LMPP)
Granulocyte-Monocyte progenitors (GMP)
Common Lymphoid Progenitor (CLP1)
Granulocyte-Monocyte progenitors/ Neutrophils (GMP/Neut)
Monocyte Cells such as CD14+ Monocytes (CD14) or CD16+ Monocytes (CD16)
B Cells (B)
pre B-cell progenitors (Pre B)
T cells such as CD8+ T Cells (CD8), CD4+ T Cells (CD4), Natural Killer Cells (NK)

Name each cluster with a unique identifier including the cell type abbreviations given in the brackets and cluster.

Hint: You should end up with 7-15 clusters. If you have too few clusters, you can either go back to the clustering in task 2 and repeat the clustering with adjusted parameters or use sub-clustering (clustering on a subset of the data) to get to the necessary number of clusters. If you have too many clusters, you can always merge clusters with cells from the same cell-type after the annotation.

Task 3. 1. 3: Plot the result of the cell type-annotation in a UMAP plot.

Task 3. 1. 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.

Task 3.2: Automatic Annotation

Use SingleR, a tool for the automatic cell-type annotation, to determine an alternative Annotation. Use the “HumanPrimaryCellAtlasData” from the celldex package as reference.

Plot the results of the automatic annotation in a UMAP plots.

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 4: Trajectory Analysis (5P)

Task 4.1: 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.

Question:

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

Task 4.2: Select root-nodes manually

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.

Questions:

1. Why is the selection of the root-nodes important for the algorithm?
2. Which points are a good choice for root nodes of the analysis and why?

Task 4.3: Select root-nodes automatically

Also try to automatically choose the root nodes.
Questions:

3. Did it improve the results? Explain why.
4. Choose one path in the trajectory and explain which cells are located on this path (I.e., the biological meaning)

Task 5: Differential Expression Analysis (5P)

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 NK cells
CD4 T cells vs CD14 Monocytes
Hematopoietic Stem Cells vs Common Lymphoid Progenitor

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

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.

Task 6.2: Pathway analysis on groups

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

Questions:

- Are the DEGs used for the pathway analysis based on single-cell or bulk methods. Explain your answer in 2-3 short sentences.

Task 6.3: Biological interpretation

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

Task 7: Cell-Cell Communication (5P)

Task 7.1: Cell-cell communication

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 8: Summary (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.

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