|  |
| --- |
| BIOINF 545 |
| Real dataset tutorial |
| *iTFSC - An R package for robust transcription factor evaluation in single-cell RNA-seq data* |

|  |
| --- |
| Gondal, Mahnoor  4-9-2023 |

Contents

[**Real dataset tutorial** 2](#_Toc131973088)

[**1.** **Introduction** 2](#_Toc131973089)

[1.1. Goal of the tutorial 2](#_Toc131973090)

[1.2. Information on the dataset 2](#_Toc131973091)

[1. The pdf version of the notebook can be found here: 2](#_Toc131973092)

[2. Step-by-step real dataset tutorial 2](#_Toc131973093)

[3. Summary of the results 4](#_Toc131973094)

[4. Conclusion on the results 4](#_Toc131973095)

## **Real dataset tutorial**

## **Introduction**

### Goal of the tutorial

In this tutorial we will run an analysis on a lung cancer dataset adopted by Qian et al [PMID: 32561858].

### Information on the dataset

This dataset has been widely cited and employed in many downstream studies. It contains four different types of cancers so users can have more flexibility to select the cancer type that they are interested in. In the tutorial I am interested in looking at the TFs which are activated in cancer cell types. Each dataset has cancer cells, therefore, it is an ideal dataset to investigate cancer agnostic transcription factor activity. At the end of the analysis, I am expecting a set of TFs which are shown to be up in cancer cells and is generated from three different methods.

Breast cancer scRNA-seq data [44,024 cells]

Colorectal scRNA-seq data [30,626 cells]

Lung cancer scRNA-seq data [66,309 cells]

Ovarian cancer scRNA-seq data [34,469 cells]

The RDS file for these datasets can be downloaded from here:

https://drive.google.com/drive/folders/1WL0TxDAQpPGzmGy8gltT-x-ezSw6Ndh1?ths=true

Ideally processed data with raw counts This file will be provided by the users

## The pdf version of the notebook can be found here:

<https://github.com/MahnoorNGondal/iTFSC/blob/main/design_document/real_dataset_tutorialNotebook.pdf>

## Step-by-step real dataset tutorial

In order to run this package, you will need to install the following dependencies:

library(Seurat)

library(SeuratDisk)

library(SCENIC)

library(BITFAM)

library(dorothea)

library(piano)

library(ggplot2)

library(dplyr)

library(tidyr)

library(AUCell)

library(RcisTarget)

library(GENIE3)

library(base)

library(tibble)

library(ComplexHeatmap)

library(ggVennDiagram)

Once the dependencies have been installed, the user can then run the following function (in order) to generate a robust list of transcription factors for any condition or cell type that they are interested in.

Step 1: Reading the scRNA-seq data

Users can download this RDS file on your local machine using this url: https://drive.google.com/drive/folders/1WL0TxDAQpPGzmGy8gltT-x-ezSw6Ndh1?ths=true

# Code example

*Qian\_merged <- readRDS("/mctp/share/users/gondal/DC\_Jenny/03\_output/Lung\_Breast\_Colon\_Qian/version\_01\_02\_06\_23/Qian\_merged.RDS")*

Step 2: Select the tissue (using existing case study)

Function name:

Tissue\_selection()

While running the analysis on sample data, users can choose which case study they are interested in. We have provided 4 case studies for this package.

Users can choose to employ the case studies of their choice or use their own data

In this function, users can choose which case study they want to work on

The case studies include: lung cancer 2 -> breast cancer 3 -> ovarian cancer 4 -> colon cancer

# Code example

A screenshot of a computer

Description automatically generated with medium confidence

Step 3: Performing data QC

Function names:

Normalization\_check()

Rawcount\_check()

This part of the code will check if the data is in the write format.

Step 4: Downsample seurat object

Function name:

Down\_sample()

The user will also need to provide the name of ident they want to downsample from and the number of cells per cluster in that ident.

Step 5a-c: Running individual methods

Function name:

Runnin\_BITFAM()

Runnin\_Dorothea()

Runnin\_SCENIC()

Each of the above function provides a customized script for running individual method for transcription factor assessments.

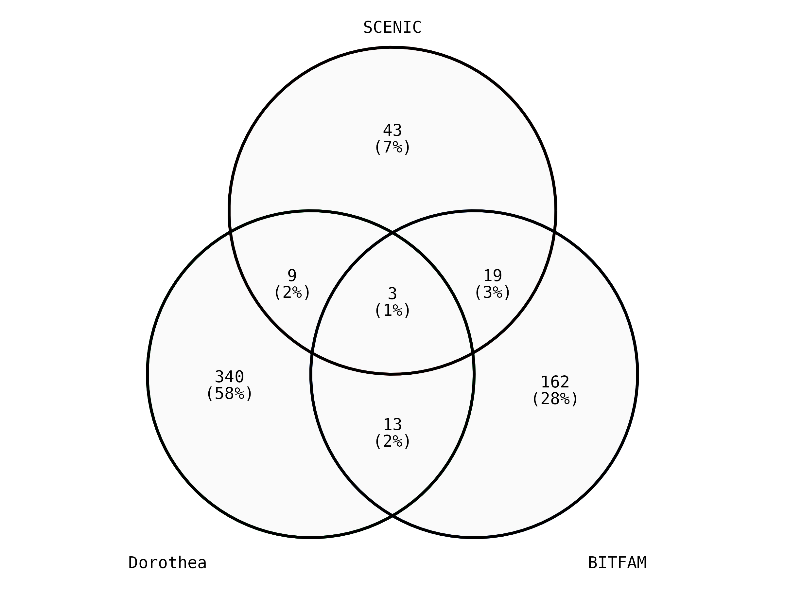
Step 6: Combining results from the three methods

Function name:

Combining\_res()

## Summary of the results

After running the code we were able to generate the following venn diagram successfully.



## Conclusion on the results

The three common TFs as been in the venn diagram above are: "YY1" "NR3C1" "BPTF". This is very interesting because YY1 has been shown to modulates lung cancer progression [PMID: 28972861] and BPTF has been involved in tumor progression [PMID: 26418899]. There are not many cases suggesting NR3C1’s involvement in cancer but this might be a novel finding in our analysis.