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

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

Contents

[**vignettes** 2](#_Toc131972096)

[**1.** **Introduction** 2](#_Toc131972097)

[1.1. Existing tools 2](#_Toc131972098)

[1.2. Goals for iTFSC 2](#_Toc131972099)

[1. Goal of tutorial 2](#_Toc131972100)

[2. Case studies for testing 2](#_Toc131972101)

[3. Steps-by-step 3](#_Toc131972102)

[4. Applications of the tool 4](#_Toc131972103)

## **vignettes**

## **Introduction**

### Existing tools

Nowadays, with the advent of high throughput single-cell sequencing, there are many tools/methods to employ single-cell data to quantify transcription factor expression. These are very widely employed in the scientific community, however, each of these methods employs different methods to quantify transcription factor “activity” and none of them look at the differential expression of these activities across groups such as across normal and healthy tissue or across cell types. integrated transcription factor analysis for single-cell data (iTFSC) is an R package that helps users run some of the well-known existing single-cell transcription factor activity computing tools based on their output, iTFSC evaluates the most robust TFs list and allows users to perform differential expression analysis.

### Goals for iTFSC

The main overarching goals for iTFSC, are, therefore, to do the following.

* + - Run existing tools
    - Perform differential expression (DE) analysis between groups using different tool outputs
    - Perform gene set enrichment analysis (GSEA) between groups using different tool outputs
    - Establish a robust list of TFs validated across methods

## Goal of tutorial

The primary goal of this tutorial is to guide the user to:

* Load the sample case study
* Downsample the data
* Run BITFAM
* Run Dorothea
* Run SCENIC
* Generate combined results

## Case studies for testing

For package validation, the following datasets will be employed from Qian et al [PMID: 32561858]. The data used would be in the form of a seuratobject RDS file. For computing purposes, we will first downsample the data to 1000 cells. We will also place relevant checks to evaluate the quality of the dataset by looking at total UMI (unique molecular identifier) and features present.

* **Case study 1**:
  + Breast cancer scRNA-seq data [44,024 cells]
* **Case study** **2**:
  + Colorectal scRNA-seq data [30,626 cells]
* **Case study 3**:
  + Lung cancer scRNA-seq data [66,309 cells]
* **Case study 4**:
  + 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

## Steps-by-step

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()

This combines results from the three methods to generate a common and more robust list of TFs.

## Applications of the tool

Since single-cell data is very noisy. This tool will generate a robust TF which are common to all three methods. Such a list of TF dependent on celltype or disease condition, depending on what the user is interested in, will be very beneficial because it would allow a very accurate assessment of the regulators in cell types which can then be used in regression modeling and allow predictors models to be build downstream.