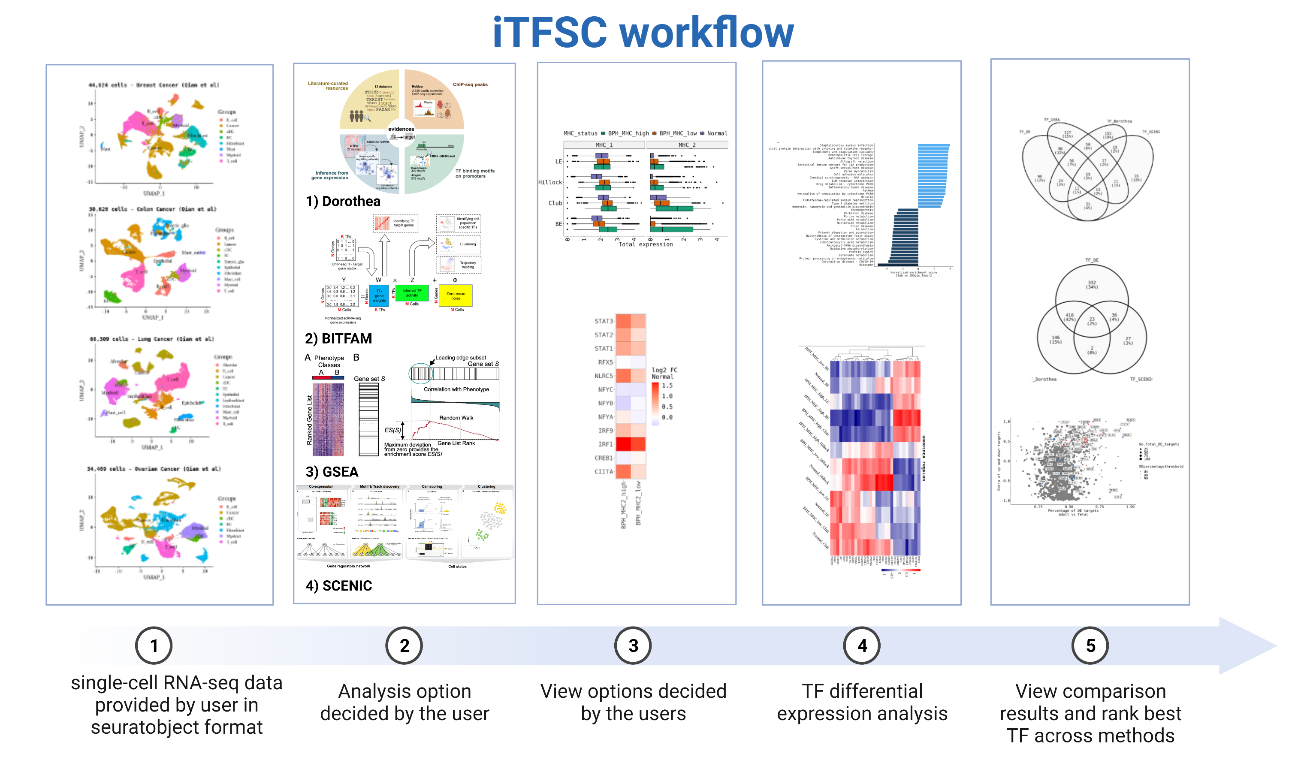
|  |
| --- |
| BIOINF 545 |
| integrated transcription factor analysis for single-cell data (iTFSC) |
| *An R package for robust transcription factor evaluation in single-cell RNA-seq data* |

|  |
| --- |
| Gondal, Mahnoor  2-8-2023 |

****

Contents

[**Software requirement specifications** 1](#_Toc126777718)

[**1.** **Introduction** 1](#_Toc126777719)

[1.1. Existing tools 1](#_Toc126777720)

[1.2. Goals for iTFSC 1](#_Toc126777721)

[1. Overall description 1](#_Toc126777722)

[1.1. Input 1](#_Toc126777723)

[1.2. Output 1](#_Toc126777724)

[1.2.1. Statistics 1](#_Toc126777725)

[1.2.2. Visualization 1](#_Toc126777726)

[1.3. Overall workflow design 1](#_Toc126777727)

[1.3.1. package flow 1](#_Toc126777728)

[1.3.2. Expected module list: 2](#_Toc126777729)

[1.3.3. Case studies for testing 2](#_Toc126777730)

[2. Requirement/dependencies 3](#_Toc126777731)

[3. Timeline 3](#_Toc126777732)

[4. Expected complications 3](#_Toc126777733)

# **Software requirement specifications**

## **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

## Overall description

## Input

To run this package user needs to provide a Seurat object and ident which they want to group by. In the comparison function, the user would also need to indicate the comparisons they want to be between.

## Output

## Statistics

Depending on the function the user employs there are different outputs. The rough idea for each function output is described below in section 4.1.1:

## Visualization

The visualization will be in the form of heatmap, boxplots, scatterplot, and venn diagrams for the comparison functions, depending on what the user is trying to visualize.

## Overall workflow design

## package flow

* Run existing function for TF
* Do DE
* Do GSEA
* Do comparison

## Expected module list:

* **Module 1**: Running the packages independently
  + Function 1: run SCENIC
    - Scope: run SCENIC with a user-defined Ident list
  + Function 2: run Dorothea
    - Scope: run Dorothea with a user-defined Ident list
  + Function 3: run BITFAM
    - Scope: run BITFAM with user-defined Ident list
  + Function 4: run GSEA
    - Scope: run GSEA (via piano) with user-defined Ident list
* **Module 2**: Running comparison analysis within each tool’s output
  + Function 5: run DE between tools
    - Scope: aggregate the activity scores and use limma for DE
  + Function 6: run GSEA between tools
    - Scope: Based on function 5 output run simple GSEA
* **Module 3**: Running comparison methods to rank robust TFs
  + Function 7: extract robust TFs
    - Scope: Use the output from the tools that the user has run and combine the output (preferably using a venn diagram)

(Each function will have its own quality checks)

## 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

## Requirement/dependencies

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

* library(SCENIC)
* library(BITFAM)
* library(Dorothea)
* library(piano)

## Timeline

**A picture containing text

Description automatically generated**

## Expected complications

I think for this project the main complications would be to figure out the interdependencies between the methods and how best to visualize and share data objects across functions.