SC1: A web-based single cell RNA-seq analysis pipeline

Marmar Moussa  
Computer Science and Engineering  
University of Connecticut  
CT, USA  
marmar.moussa@uconn.edu

Ion I. Mǎndoiu  
Computer Science and Engineering  
University of Connecticut  
CT, USA  
ion@engr.uconn.edu

Abstract—Single cell RNA-Seq (scRNA-Seq) is critical for studying cellular heterogeneity and development of tissues and tumors. There are currently only few tools that allow researchers to analyze scRNA-Seq data without considerable coding efforts. Here, we present an interactive web-based pipeline for single-cell RNA-Seq data analysis that allows researchers to analyze data in an efficient work-flow. The tool also implements a novel and unique method of selecting informative genes before clustering and differential expression analysis.

Index Terms—Single cell RNA-Seq, bioinformatics pipeline

I. INTRODUCTION

There are currently only few packages (mostly in R) for scRNA-Seq data analysis. Most of them however require considerable programming knowledge and are not easy to use by biologists. Recently, web-based scRNA-Seq analysis pipelines such as Granatum [1] have begun to offer a more user-friendly interface, however there is no widely adopted analysis workflow yet. Here, we present a web-based scRNA-Seq data analysis tool publicly accessible at https://sc1.engr.uconn.edu/. The tool includes several quality control (QC) options, a novel method for gene selection based on Term-Frequency Inverse-Document-Frequency (TF-IDF) scores, followed by cell clustering and visualization tools as well as Differential Expression (DE) analysis and gene enrichment steps.

II. SC1 WORKFLOW

The SC1 workflow is implemented in the R programming language, with an interactive web-based front-end built using the Shiny framework [2]. To facilitate interactive data exploration, time consuming computations such as computing PCA and t-SNE projections of the data following basic QC are performed in a preprocessing step. Preprocessed data is saved in the tool’s .scDat file format that can be uploaded for interactive analysis.

a) Quality Control: Before further analyses, SC1 allows users to perform several QC checks, where poor quality cells and outliers can be excluded from subsequent analysis. The tool includes widely used criteria like relative library size, number of detected genes, fraction of reads mapping to mitochondrial genes, ribosomal protein genes, or synthetic spike-in.

b) Informative Gene Selection and Clustering: SC1 implements a novel method of selecting informative genes based on the average TF-IDF scores [3], where we fit a mixture of gaussian models to the distribution of TF-IDF gene averages and select the genes assigned to the mixture components with highest weight. The user can also control the minimum number of genes to use for clustering. Genes with highest average TF-IDF scores differentiate best between heterogeneous cell populations (see Fig. 1). Clustering can be performed using Ward’s hierarchical agglomerative clustering algorithm using the top average TF-IDF genes as features.

c) Differential Expression and Gene Enrichment Analyses: DE analysis is done by performing one vs. the rest t-tests for each identified cluster. DE analysis is followed by functional enrichment analysis performed using the ‘gProfileR’ R package [4] with results visualized as word clouds.

d) Data Visualization: The tool includes multiple interactive visualization options like interactive 3D t-SNE plots with the ability to visualize genes individually, in pairs, or in groups. The Cell View and Gene-Pairs View visualizations options help researchers explore co-expressions of any given pair of genes. Heatmaps are used to visualize results of clustering and DE steps.

III. FUTURE WORK

Additional workflow steps are currently under development, including integrating the LSImpute tool [5] for imputing missing data, additional clustering methods, as well as other differential expression analysis methods.

REFERENCES