Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Notes

Application Notes

scGEAToolbox: a Matlab toolbox for single-cell RNA sequencing data analysis

James J. Cai[1,2,*]

[1]Department of Veterinary Integrative Biosciences, [2]Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843-4458, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Single-cell RNA sequencing (scRNA-seq) technology has revolutionized the way research is done in biomedical sciences. It provides an unprecedented level of resolution across individual cells for studying cell heterogeneity and gene expression variability. Analyzing scRNA-seq data is challenging though, due to the sparsity and high dimensionality of the data.

Results: We developed scGEAToolbox—a Matlab toolbox for scRNA-seq data analysis, including a comprehensive set of functions for data normalization, feature selection, batch correction, imputation, cell clustering, trajectory inference, and network construction. While most of the functions are implemented in native Matlab language, wrapper functions are also provided to allow Matlab users to call a large number of the "third-party" tools, which are not necessarily developed in Matlab. Furthermore, scGEAToolbox is equipped with sophisticated graphic user interfaces (GUIs) generated with App Designer, making it an easy-to-use application for quick data filtering, normalization, visualization, as well as downstream functional enrichment analyses.

Availability: https://github.com/jamesjcai/scGEAToolbox

Contact: jcai@tamu.edu

1 Introduction

Single-cell technologies, especially single-cell RNA sequencing (scRNA-seq), have revolutionized the way biologists and geneticists study cell heterogeneity and gene expression variability. Analyzing scRNA-seq data, however, is a challenging task due to the sparsity and dimensionality of the data. The sparsity is rooted from the limitation in the sensitivity of single-cell assay system; the high dimensionality is a characteristic property of scRNA-seq data. The data sets are also confounded by nuisance technical effects. The analyses of scRNA-seq data involve, in general, data filtering, normalization, feature selection, cell clustering, marker gene identification, cell type identification, pseudotime and trajectory analysis, regulatory network construction, and so on. When multiple data sets are compared, batch effect correction is often required. For every aspect of these analyses, there has been a plethora collection of software tools to fulfill the task. The majority of these tools are developed in computer languages other than Matlab, such as R and python.

Matlab is a scientific programming language and provides strong mathematical and numerical support for the implementation of advanced algorithms. Its basic data element is the matrix; mathematical operations that work on arrays or matrices are built-in to the Matlab environment. Matlab comes with many toolboxes, such as statistics, bioinformatics, optimization, and image processing. Nevertheless, a dedicated Matlab toolbox for comprehensive analyses of scRNA-seq data is still missing. Given scRNA-seq data is increasing exponentially over time, we believe a new Matlab toolbox for scRNA-seq data analysis is highly desired.

2 Methods

We developed scGEAToolbox using Matlab v9.5 (R2018b). Functions in scGEAToolbox are written in native Matlab, and the app GUIs are created with App Designer. Most functions take two variables: `X` and `genelist`, as inputs of scRNA-seq data. `X` is a matrix of dimension $n \times m$, where n denotes the total number of genes and m denotes the total number of cells; `genelist` is an $n \times 1$ string array holding the names of n genes. Main

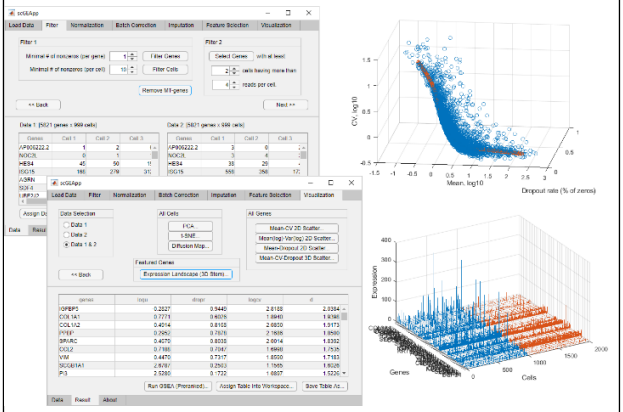

Fig. 1. Screenshots of an execution of scGEApp, part of scGEAToolbox GUIs.

categories of functions of scGEAToolbox include: file input and output, data normalization, gene and cell filtration, detection of highly variable genes (HVGs), batch effect correction, cell clustering, dimensionality reduction, data visualization, trajectory analysis, and network construction. For each of these functional categories, at least two algorithms were implemented. For example, for data normalization, `norm_libsize`, and `norm_deseq` are two functions that normalize `X` using library size and the method of DESeq, respectively. Furthermore, an "entry" function called `sc_norm` was developed. The two normalization functions can be accessed using `sc_norm(X,'type','libsize')` and `sc_norm(X,'type', 'deseq')`. Accordingly, the functionSignatures.json file was edited to specify the usage of all entry functions. The main GUI application in scGEAToolbox is called scGEApp (**Fig. 1**). It contains a main panel with seven tabs, namely *Load Data*, *Filter*, *Normalization*, *Batch Correction*, *Imputation*, *Feature Selection*, and *Visualization*. These tabs are ordered according to the order of the general workflow of scRNA-seq data. Moving between tabs can be done by clicking the tab or clicking 'Next' and 'Back' buttons on each tab panel. Command-line functions can be executed and applied to the loaded data by clicking buttons under each tab. For example, functions for selecting cells by library size and genes by the number of mapped reads are under *Filter*; functions for HVG selection are under *Feature Selection*; functions for t-SNE and PHATE are under *Visualization*. Under the main panel is a panel for viewing data matrices and result tables, where data and results can be exported into the workspace as variables or saved into external files.

3 Results

For each of the categories of scRNA-seq data analysis methods, several algorithms were either implemented in Matlab or incorporated through wrapper functions in scGEAToolbox. For example, `sc_hvg` and `sc_veg` contain implementations of two methods of HVG detection: one is the method of (Brennecke, et al., 2013) and the other is the method of (Chen, et al., 2016); `sc_sc3` contains the implementation of SC3 (Kiselev, et al., 2017) for consensus clustering; `sc_pcnet` contains the implementation of the dna/PCnet method (Gill, et al., 2010) for principal component network inference; and `sc_tscan` implements TSCAN (Ji and Ji, 2016) for trajectory analysis. Some computational tasks are shared by many tools. In this case, we developed "modular" functions that perform these common tasks, e.g., a function that uses different methods to compute the cell-to-cell similarity matrix and a function that uses different methods to estimate the number of clusters. These modular functions can be utilized in the process of new algorithm development. In scGEAToolbox, a new function for the visualization of genes' summary statistics was introduced. The method is based on the 3-D spline fit curve in a space defined by expression mean (μ), coefficient of variation (CV), and the dropout rate (r_{drop}) of genes. It can be applied to identify feature genes, i.e., those with cell-to-cell expression variability in μ, CV, and r_{drop} deviated from the majority of other genes (**Fig. 1**, upper right).

To expand its functionality, scGEAToolbox incorporates many existing Matlab-based packages such as ComBat, HCP (Hidden Covariates with Prior), MAGIC, McImpute, SIMLR, SinNLRR, SoptSC, bigSCale, DensityClust, PHATE, scDiffMap and GENIE3. All these tools can be accessed through corresponding wrapper functions, e.g., `run_magic`, `run_simlr` and `run_genie3`. Like other functions, most of these wrapper functions take `X` and `genelist` as inputs. Furthermore, scGEAToolbox also includes wrapper functions for selected R packages, e.g., UMAP, SCODE and Monocle, which facilitates the use of diverse sources of functions and tools developed in R.

As many functions can be run through the GUI of scGEApp without using the command line, scGEAToolbox is a useful training tool for beginners. Example data sets are given in a subfolder; four demonstration script files are provided. The source code of scGEAToolbox is provided free for academic use. When needed, stand-alone applications of scGEApp can be built for all major platforms with or without Matlab installed.

In summary, scGEAToolbox is designed and developed to provide better data analysis support for scRNA-seq data using the Matlab environment. It makes two key contributions: (1) implementing and incorporating a large number of high-level analytical functions, and (2) defining an easy-to-use GUI for commonly used methods in scRNA-seq data analysis. We anticipate that these key features will make scGEAToolbox a useful tool for researchers to conduct analysis with scRNA-seq data more effectively and develop new algorithms more efficiently.

Acknowledgements

The author thanks Jianhua Huang, Yan Zhong and Guanxun Li for helpful discussion and inspiration during the development of this software tool.

Funding

This work has been supported by the Texas A&M University T3 grant and NIH grant R21AI126219.

Conflict of Interest: none declared.

References

Brennecke, P., *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013;10(11):1093-1095.

Chen, H.I., *et al.* Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics* 2016;17 Suppl 7:508.

Gill, R., Datta, S. and Datta, S. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics* 2010;11:95.

Ji, Z. and Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* 2016;44(13):e117.

Kiselev, V.Y., *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;14(5):483-486.