*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Application Notes

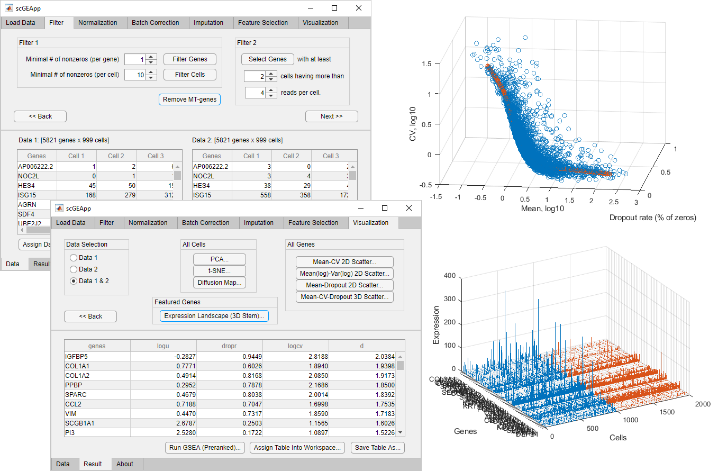
|  |
| --- |
| Application Notes  scGEAToolbox: a Matlab toolbox for single-cell RNA sequencing data analysis  James J. Cai1,2,\*  1Department of Veterinary Integrative Biosciences, 2Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843-4458, USA.  \*To whom 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:** I 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 the “third-party” tools, which are not necessarily developed in Matlab. Furthermore, scGEAToolbox is equipped with sophisticated graphical 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](mailto:jcai@tamu.edu)  **Supplementary information:** Supplementary data are available at *Bioinformatics* online. |

# 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 problem is rooted from the limitation in the sensitivity of single-cell assay system; scRNA-seq data sets are often 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 or trajectory analysis, gene regulatory network construction, and so on. When multiple data sets are compared, batch effect correction is often required to harmonize the data. 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 such as R and python; few tools, except e.g., SCell (Diaz, et al., 2016) and SCUBA (Marco, et al., 2014), are developed in Matlab, a scientific programming language and provides strong mathematical and numerical support for the implementation of advanced algorithms. The basic data element of Matlab 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. Up to date, a dedicated Matlab toolbox for comprehensive analyses of scRNA-seq data is still missing. Given scRNA-seq data is increasing exponentially over time, a new Matlab toolbox for such data analyses is highly desired.

# Methods

I 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*×*m*, where *n* denotes the number of genes and *m* the number of cells; genelist is an *n*×1 string array holding the names of the *n* genes. Main 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, dimensionality reduction, data visualization, cell clustering, trajectory analysis, and network construction. For each of these functional categories, multiple algorithms were implemented. For example, for data normalization, norm\_libsize and norm\_deseq are provided to normalize X using the method based on library size and the method of DESeq, respectively. Furthermore, an “entry” function called sc\_norm was developed to allow users to access the two normalization functions 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. 1a**). It contains the main panel with multiple tabs, namely *Load Data*, *Filter*, *Normalization*, *Batch Correction*, *Imputation*, *Feature Selection*, *Visualization*, *Clustering*, *Pseudotime*, and *Network*. On each tab panel, there are buttons for executing corresponding functions. For example, function for selecting cells by library size and selecting 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. Majority of functions in scGEAToolbox were implemented in native Matlab. These include sc\_hvg and sc\_veg, two HVG detection methods (Brennecke, et al., 2013; Chen, et al., 2016); sc\_sc3, the implementation of SC3 for consensus clustering (Kiselev, et al., 2017); sc\_pcnet, the implementation of the dna/PCnet method for principal component regression network inference (Gill, et al., 2010); and sc\_tscan, the implement of TSCAN for trajectory analysis (Ji and Ji, 2016). Some computational tasks are shared by many tools. In this case, I 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. Several new functions were introduced, including sc\_scatter3 for visualizing three summary statistics, namely, expression mean (µ), coefficient of variation (CV) and the dropout rate (rdrop) of a gene across cells. Genes are arranged in the 3-D space and a spline curve is defined to facilitate the identification of feature genes, i.e., those with cell-to-cell expression variability deviated from the majority of other genes (**Fig. 1b**). 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 such as run\_magic, run\_simlr and run\_genie3. These wrapper functions also take X and genelist as inputs. Furthermore, scGEAToolbox also includes wrapper functions for few selected R functions such as UMAP, SCODE and Monocle.

**Fig. 1. Screenshots of an execution of scGEAToolbox.** (**a**) Two example panels of the main GUI scGEApp; (**b**) A 3-D scatter plot showing genes whose position is determined by µ, CV and rdrop; (**c**) A stem plot showing expression level of 50 selected genes across 2000 cells: 1000 in one state (blue) and the other 1000 in the other state (red).

# Results

scGEAToolbox was developed to facilitate the analyses of scRNA-seq data in Matlab environment. It contains a comprehensive set of high-level functions. No other Matlab toolbox offers a nearly comparable number of functions and function categories. Nearly all functions take the same input arguments (i.e., X and genelist)—a uniform functional interface that simplifies the usage, making scGEAToolbox an ideal algorithm prototyping tool. As many functions in scGEAToolbox can be run through the GUI application without using the command line, it is a useful training tool for beginners. Visualization functions support an intuitive interpretation of the data (e.g., **Fig. 1c**). 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 provides comprehensive data analysis support for scRNA-seq data in Matlab. 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. I 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 Daniel Osorio, 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.

Diaz, A.*, et al.* SCell: integrated analysis of single-cell RNA-seq data. *Bioinformatics* 2016;32(14):2219-2220.

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.

Marco, E.*, et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci U S A* 2014;111(52):E5643-5650.