*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 to cell type heterogeneity and gene expression variability across individual cells. Analyzing scRNA-seq data is challenging due to sparsity and high dimensionality of the data.  **Results:** We developed scGEAToolbox—a Matlab toolbox for scRNA-seq data analysis, includign 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 developed 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 graphic user interfaces (GUIs) generated with App Designer, making it an easy-to-use application for quick data filtering, normalization, visualization, and 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. |

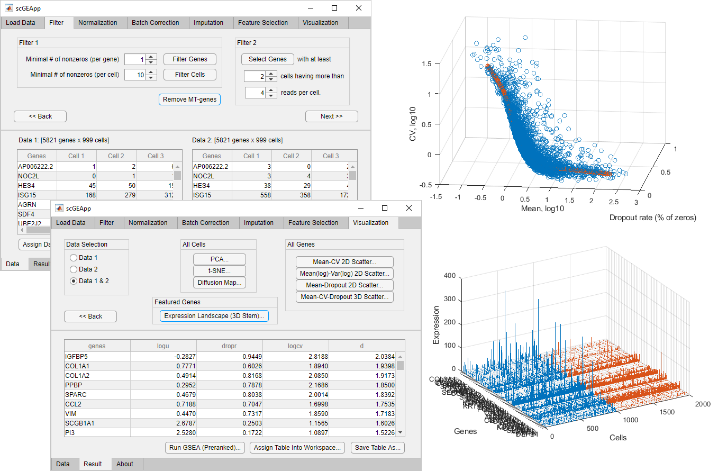
# Introduction

Single-cell technologies, especially single-cell RNA sequencing (scRNA-seq), have revolutionized the way biologists and geneticists study cell type heterogeneity and gene expression variability. Analyzing scRNA-seq data, however, is a challenging task. This is due to that scRNA-seq data sets tend to be sparse. The sparsity is rooted from the limitation in the sensitivity of single-cell assay system. The data sets are also naturally high-dimensional and often confounded by nuisance technical effects. The analyses of scRNA-seq data involves, in general, filtering, normalization, feature selection, cell clustering, marker gene identification, cell type identification, pseudotime and trajectory analysis, and single-cell regulatory network construction. When multiple data sets are included, batch effect correction and manifold alignment between cells from different data sets are often required. For every aspect of these scRNA-seq data analyses, there is a plethora collection of software tools that have already been developed to fulfill the task. The majority of these tools are developed in computer language other than Matlab, including 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 contains many toolboxes, such as statistics, bioinformatics, optimization, and image processing, but dedicated toolboxes for comprehensive analyses of scRNA-seq data are still missing. Given that scRNA-seq data is increasing exponentially over time, we believe a new Matlab toolbox for scRNA-seq data analysis is highly desired.

# 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 scGEAToolbox functions take two variables: X and genelist, as the input of scRNA-seq data. X is a matrix of dimension *n*×*m*, where *n* denotes the total number of genes and *m* denotes the total number of cells; genelist is a *n*×1 string array that contains the name of *n* genes. Main categories of functions of scGEAToolbox include file I/O, 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 (Anders and Huber, 2010), 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 scGEApp (**Fig. 1**, upper right). 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 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. Under the main panel is the panel for viewing data matrices and result tables. Data and results in tables can be exported into the workspace as variables or saved into external files.

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

# Results

We implemented multiple Matlab functions for each category of analyses involved in the scRNA-seq data processing. For example, sc\_hvg and sc\_veg implement methods of (Brennecke, et al., 2013) and (Chen, et al., 2016) for HVG detection; sc\_sc3 implements SC3 for cell clustering; sc\_pcnet implements the pcNet method for network construction; and sc\_tscan implements TSCAN for trajectory analysis. We developed functions that perform tasks commonly shared in many analytical tools. For example, scGEAToolbox contains a function that uses different methods to compute the cell-to-cell similarity matrix and another function that different methods to estimate the number of clusters. We believe these “modular” functions can be utilized in the development of new algorithms.

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 (rdrop) of genes. It can be applied to identify feature genes, i.e., those with cell-to-cell expression variability in µ, CV, and rdrop deviated from the majority of other genes (**Fig. 1**, upper right). In addition to its own implementations, scGEAToolbox incorporates many existing analytical tools such as Combat, HCP, MAGIC, mcImpute, fitSNE, PHATE, UMAP, GENIE3, SCode, Monocle 2, simlr, SinNLRR, soptsc, and so on. All these tools can be through wrapper function in Matlab directly and most of these wrapper functions, just like the other functions, take X and genelist as input variables. This design results in great convenience in using such diverse sources of functions.

Many functions of scGEAToolbox can be accessed through the main GUI of scGEApp and are organized under each tab by their categories. For example, functions for selecting cells and genes by the number of mapped reads are under Filter; functions for normalization by using library size and by using the method of DESeq are under Normalization. The Feature Selection tab panel contains two functions for HVG selection: one uses the method of (Brennecke, et al., 2013) and the other uses the method of (Chen, et al., 2016). Truncated data sets derived from real 10x Genomics scRNA-seq data are included in scGEAToolbox as example data in a subfolder. 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. It makes two key contributions: (1) implementing and incorporating a large number of functions, and (2) defining an easy-to-use GUI for a number of commonly used methods in scRNA-seq data analysis. We anticipate that these two 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 partially supported by the Texas A&M University T3 grant and NIH grant R21AI126219.

*Conflict of Interest:* none declared.

References

Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome Biol* 2010;11(10):R106.

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.

Habiel, D.M.*, et al.* CCR10+ epithelial cells from idiopathic pulmonary fibrosis lungs drive remodeling. *JCI Insight* 2018;3(16).

Hagai, T.*, et al.* Gene expression variability across cells and species shapes innate immunity. *Nature* 2018;563(7730):197-202.

Yip, S.H., Sham, P.C. and Wang, J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform* 2018.

Zyla, J.*, et al.* Ranking metrics in gene set enrichment analysis: do they matter? *BMC Bioinformatics* 2017;18(1):256.