*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 given that the data is sparse and high-dimensional.  **Results:** We developed scGEAToolbox—a Matlab toolbox for scRNA-seq data analysis, that contains a comprehensive set of functions for data normalization, feature selection, batch correction, imputation, cell clustering, trajectory analysis, and network inference. While most of functions are implemented by using native Matlab code, a number of wrapper functions are developed to allow Matlab users to call the “third-party” tools, which are not necessarily developed in Matlab. Furthermore, scGEAToolbox is also equipped with graphic user interfaces 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 types and gene expression variability. Analyzing scRNA-seq data, however, is a challenging task. The reasons include that scRNA-seq data sets tend to be sparse due to the limitation of sensitivity of single-cell assay system; the data is naturally high-dimensional and is often confounded by nuisance technical effects. The analyses of scRNA-seq data in general contain 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 alignment between cells from different data sets are often required. For every aspect of these scRNA-seq data analyses, there is plethora collection of software tools that have been developed to fulfil the task. However, 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. Given that scRNA-seq data s increasing exponentially over time, we believe new Matlab toolboxes dedicated to scRNA-seq data analysis are 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 takes two variables: X and genelist, as the input 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. Main categories of functions include: file input and output, data normalization, gene and cell filtration, detection of highly variable genes (HVGs), batch effect correction, cell clustering, data visualization, trajectory analysis, and regulatory network construction. For each of functional categories, at least two algorithms were implemented or incorporated in scGEAToolbox. For example, norm\_libsize and norm\_deseq are two functions for normalization using library size and using the method of DESeq (Anders and Huber, 2010). Furthermore, scGEAToolbox also provides an unified entry function called sc\_norm, with which the two normalization functions can be accessed using sc\_norm(X, 'type', 'libsize') and sc\_norm(X, 'type', 'deseq'), respectively. The functionSignatures.json file has been edited to specify the function usage. The main GUI application in scGEAToolbox is scGEApp, which is a Matlab App with a main panel of seven tabs, namely Load Data, Filter, Normalization, Batch Correction, Imputation, Feature Selection, and Visualization. These tabs are ordered following a general workflow of scRNA-seq data, i.e., data acquiring, processing, and information extraction. Moving between tabs can be done by clicking the tab name or clicking ‘Next’ and ‘Back’ buttons on each tab panel. Under the main panel is the panel for viewing data matrices and the result table. Data and results in tables can be exported into the workspace as variables or saved into external files. Most functions of scGEApp can be accessed through the main GUI 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).

# Results

We implemented Matlab functions for each category of analyses commonly required in 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 implments the pcNet method for network construction; and sc\_tscan implments 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 visualization of genes’ summary statistics was introduced. The method is based on the 3D spline fit curve in a space defined by expression mean (µ), CV, and the dropout rate (rdrop) of genes. It can be applied to identify HDGs, i.e., genes with the cross-cell expression feature (involving µ, CV, and rdrop) deviated from the majority of other genes (Fig. 1, upper right). In addition to its own implementations, scGEAToolbox incorporates a plethora of existing analytical tools, namely Combat, HCP, MAGIC, mcImpute, fitSNE, PHATE, UMAP, Gorilla, Enrichr, GSEA Preranked, GENIE3, SCode, Monocle 2, simlr, SinNLRR, soptsc, and so on.

A truncated data set derived from GSM3204305 and GSM3204304 is provided as example data in one of the subfolders of scGEApp to allow users to identify genes play a role in tissue structural remolding in idiopathic pulmonary fibrosis lungs.

In summary, scGEApp is designed and developed to provide better data analysis support for scRNA-seq data. It makes two key contributions: (1) introducing a non-parametric, 3D spline-based feature selection method, 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 scGEApp a useful tool for researchers to conduct feature selection analysis with scRNA-seq data more effectively.

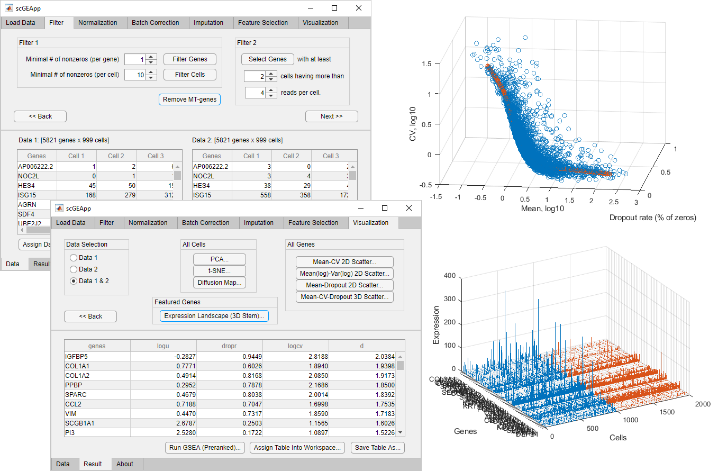
In the development of our feature selection method, we considered three summary statistics of scRNA-seq expression for each gene: mean, CV, and dropout rate. Mean and CV are computed across cells without removing zeros, and the dropout rate is computed as the fraction of cells with zero expression for the given gene. Every gene is characterized by these three variables and has its own unique position in the 3D space defined by the three variables. We used real droplet-based scRNA-seq data (BioProject: PRJNA508890) to show the distribution of genes in such a 3D space: data points (genes) form an ‘S’-shaped manifold (Fig. 1). To fit the curve, we used function SPLINEFIT (by Jonas Lundgren). This function handles noisy data and removing unwanted oscillations in the spline curve from noisy data. We compute, *d*, the shortest distance from each data point to the spline curve, and use it as the feature of the gene. Genes with large *d* are called highly deviated genes (HDGs). The source code of scGEApp is provided free for academic use, and stand-alone applications of scGEApp are provided for all major platforms with or without Matlab installed.

Applications of scRNA-seq include (1) identification of cell types in a sample and (2) characterization of variability across individual cells of the same cell type. The latter application has gained a growing attention because, within an ensemble of identical cells, cell-to-cell variation often indicates a diversity of hidden functional capacities that facilitate collective behavior in tissue function and normal development, and the change of this functional diversity may be associated with disease development (Habiel, et al., 2018; Hagai, et al., 2018). Nevertheless, characterizing cell-to-cell variation in gene expression remains challenging because scRNA-seq data is often confounded by nuisance technical effects.

Feature selection is the statistic process of selecting a subset of relevant features, variables, or predictors for use in model construction.

During the last two decades, feature selection techniques have become an active and fruitful research field in machine learning [[[1]](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0005), [[2]](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0010), [[3]](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0015), [[4]](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0020)], pattern recognition [[5](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0025),[6](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0030)], and bioinformatics [[[7]](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0035), [[8]](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0040), [[9]](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0045)]. Feature selection, a.k.a. Variable selection or gene selection (in bioinformatics), is the process of selecting a subset of relevant features for model construction or interpretation of results. It improves model predictive accuracy and reduces model complexity by eliminating irrelevant and redundant features and provides a better understanding of the underlying processes [[10](https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/#bb0050)].

In the scRNA-seq analysis, feature selection can be used to control for nuisance factors of technical noise and select biologically meaningful genes, e.g., highly variable genes (HVGs) that drive the heterogeneity across cells in a population (Brennecke, et al., 2013). Feature selection algorithm can be parametric or nonparametric. In parametric modeling, each data point is treated as a random variable, i.e., is the expression of gene *i* in cell *j* (for and ), and fit a parametric statistical model to this variable. Once these models have been fit to the data, they can then be used for various downstream tasks such as normalization, imputation, and clustering. On the other hand, in nonparametric settings, such probabilistic modeling is not explicitly employed. Although a number of parametric and nonparametric algorithms and tools have been developed for feature selection with scRNA-seq data, different methods capture different aspects of gene features. A comparative study of seven HVG analysis methods from six different packages showed that, even with the same data set, different tools produce different resulting lists of HVGs (Yip, et al., 2018). Given feature selection is an important step to identify genes contribute to cell heterogeneity, effective feature selection algorithms and easy-to-use software tools are highly desired

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

Acknowledgements

The author thanks Jianhua Huang, Yan Zhong and Guanxiong 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.