*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

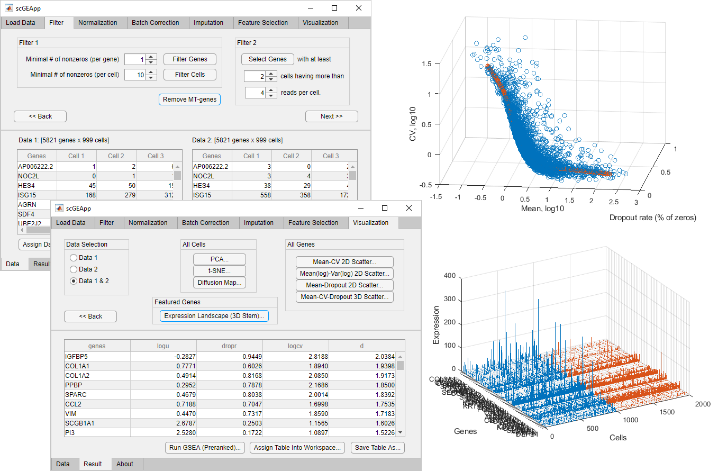
Application Notes

|  |
| --- |
| Application Notes  scGEApp: a Matlab app for single-cell gene expression 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:** The recent development of single cell technologies, especially single-cell RNA sequencing (scRNA-seq), provides an unprecedented level of resolution to the cell type heterogeneity and also facilitates the study of transcriptome variation across individual cells of the same cell type. Feature selection algorithms have been developed to control for sampling noise and select biologically meaningful genes for downstream analyses. An easy-to-use application for feature selection with scRNA-seq data also requires support functions for data filtering, normalization, visualization, and functional enrichment analysis. Graphic user interfaces (GUIs) are also desired for such an application.  **Results:** We used native Matlab and App Designer to develop scGEApp for the analysis of single-cell gene expression data. We specifically designed a new algorithm for feature selection based on 3D spline fitting of expression mean, coefficient of variance, and the dropout rate, making scGEApp a powerful tool for feature selection with scRNA-seq data. Using real data, we show that our new method can be applied single-sample data to identify highly deviated genes; it can also be used to data from two samples to identify differentially deviated genes. Users can operate scGEApp through GUIs to access all functions for data normalization, batch effect control, imputation, visualization, and feature selection, as well as downstream analyses with GSEA Preranked and GOrilla.  **Availability:** <https://github.com/jamesjcai/scGEApp>  **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 gene expression. 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). Characterizing cell-to-cell variation in gene expression remains challenging because scRNA-seq data is often confounded by nuisance factors. Feature selection is the statistic process of selecting a subset of relevant features, variables, predictors, for use in model construction.

In the scRNA-seq analysis, feature selection can be to control for nuisance factors of technical noise and select biologically meaningful genes, e.g., highly variable genes (HVGs) driving 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. In nonparametric settings, statistics are inferred without explicit probabilistic modeling. 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 that feature selection is an important step to identify genes contribute to cell heterogeneity, relating gene expression to cellular functions of the tissue, effective feature selection methods and easy-to-use software tools are highly desired.

**Fig. 1.** **Screenshot of an execution of scGEApp.**

# Methods

We developed scGEApp using Matlab v9.5 (R2018b). Functions in scGEApp are written in native Matlab and the app GUIs are created with App Designer. The main panel of the current version of scGEApp includes seven tabs, namely Load Data, Filter, Normalization, Batch Correction, Imputation, Feature Selection, and Visualization, which are ordered following the workflow of 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: HVG selection using the method of (Brennecke, et al., 2013) and HDG selection using our 3D spline curve-based method.

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 largest *d* are called highly deviated genes (HDGs). The source code of scGEApp is provided free for academic use. Stand-alone application of scGEApp is provided for platforms without Matlab installed.

# Results

Here we introduce a non-parametric feature selection method, which is only based on the summary statistics computed from given scRNA-seq data. Our method is based on the 3D spline fit curve in a space defined by mean, CV, and the dropout rate of genes. Our method can be applied to a single sample to identify HDGs, whose expression feature across cells is deviated from majority of other genes. Our method can also be applied to two samples from comparative analysis. In the two-sample setting, the deviation from the spline curve, *d*, is computed for each gene for the two samples independently. Then, the difference in the deviation, *dd*, can be computed for each gene. We have tested our method through scGEApp GUIs with two comparative scRNA-seq data sets: E-MTAB-5988 (unstimulated) vs E-MTAB-5989 (stimulated dermal fibroblasts)(Hagai, et al., 2018) and GSM3204305 (CCR10-) vs GSM3204304 (CCR10+ epithelial cells)(Habiel, et al., 2018). Genes then can be ordered by using the value of their *dd*, and subject to two downstream analyses: GSEA Preranked (Zyla, et al., 2017) and GOrilla. In both cases, enrichment tests for genes ranked with *dd* produced results highly relevant to functions of the tissues from which samples are derived: primary dermal fibroblasts (Hagai, et al., 2018) and lung airway epithelial cells (Habiel, et al., 2018). 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.

Acknowledgements

The author thanks Jianhua Huang, Yang 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 seed 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.

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.