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 feature selection on single-cell RNA sequencing data

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: 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. It also enables the study of gene expression variability across individual cells within a homogenous cell population. Feature selection algorithms have been used to select biologically meaningful genes while controlling for sampling noise. An easy-to-use application for feature selection on scRNA-seq data requires integration of functions for data filtering, normalization, visualization, and enrichment analyses. Graphic user interfaces (GUIs) are desired for such an application.

Results: We used native Matlab and App Designer to develop scGEApp for feature selection on single-cell gene expression data. We specifically designed a new feature selection algorithm based on the 3D spline fitting of expression mean (μ), coefficient of variance (CV), and dropout rate (r_{drop}), making scGEApp a unique tool for feature selection on scRNA-seq data. Our method can be applied to single-sample or two-sample scRNA-seq data, identify feature genes, e.g., those with unexpectedly high CV for given μ and r_{drop} of those genes, or genes with the most feature changes. Users can operate scGEApp through GUIs to use the full spectrum of functions including normalization, batch effect correction, imputation, visualization, feature selection, and downstream analyses with GSEA and GOrilla.

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

Contact: jcai@tamu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 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). 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. 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., x_{ij} is the expression of gene i in cell j (for $i = 1, \dots, n$ and $j = 1, \dots, m$), 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 proba-

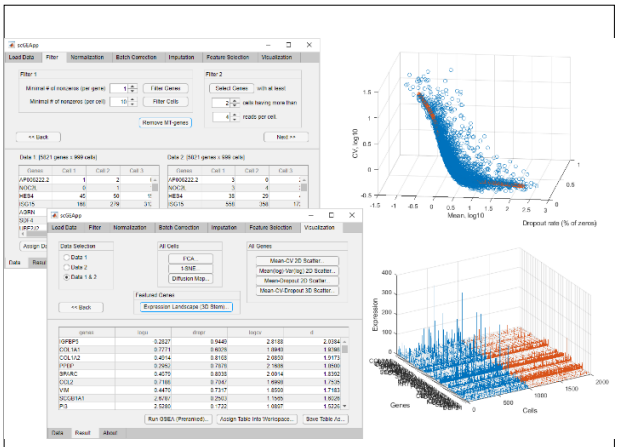

Fig. 1. Screenshot of an execution of scGEApp.

bilistic 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.

2 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: one uses the method of (Brennecke, et al., 2013) to identify HVGs and the other uses our 3D spline curve-based method to identify highly deviated genes (HDGs).

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.

3 Results

Here we introduce a new non-parametric feature selection method using only summary statistics computed from given scRNA-seq data. Our method is based on the 3D spline fit curve in a space defined by expression mean (μ), CV, and the dropout rate (r_{drop}) of genes. It can be applied to a single sample to identify HDGs, i.e., genes with the cross-cell expression feature (involving μ, CV, and r_{drop}) deviated from the 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*, is computed for each gene. We have tested our method using 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). After feature selection, genes can be ordered by their *dd* values and the ranked genes can be analyzed using downstream programs, e.g., GSEA Preranked (Zyla, et al., 2017) and GOrilla. With both data sets, enrichment tests for the ranked genes produced highly relevant results, showing the function of the tissues from which samples are derived, i.e., 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, 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

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.