*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Application Notes

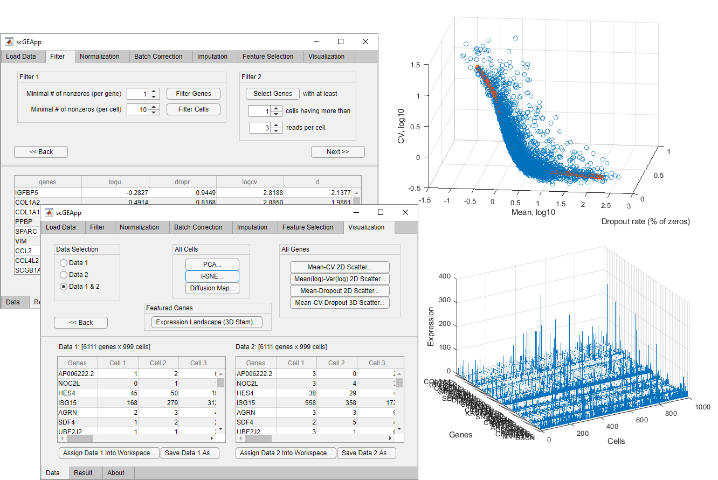
|  |
| --- |
| Application Notes  scGEApp: a Matlab app for single-cell gene expression analysis  James J. Cai1,2,\*  1Department of Veterinary Integrative Biosciences, 2Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843-4458, USA.  \*To whom correspondence 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 cellular stochasticity, and select biologically meaningful genes or cells for downstream analysis. An easy-to-use application for feature selection with scRNA-seq data requires a number of functional modules, including those for data filter, normalization, visualization, and ideally gene ontology enrichment analysis or other downstream analyses. Graphic user interfaces (GUI) 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 dropout rate, making scGEApp a powerful tool for extracting features from scRNA-seq data of single sample and comparing features between two samples. Users can operate scGEApp through GUI to access a variety of functions for data normalization, batch effect control, imputation, visualization, and feature selection using our algorithm or other algorithms for identifying highly variable genes. scGEApp provides built-in functionality of GSEA (Preranked) analysis for featured genes.  **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 can be classified into two major categories: (1) identification of cell types in a sample, and (2) characterize variability across individual cells of the same cell type. The latter application has gained a growing attention because the variability across individual cells may indicate a diversity of hidden functional capacities within the ensemble of identical cells, and the change of this functional diversity may facilitate collective behavior in tissue function and normal development (Bahar Halpern, et al., 2015; Chang, et al., 2008) (Richard, et al., 2016)(Mohammed, et al., 2017).

Nevertheless, interpreting scRNA-seq data remains challenging. The data is confounded by nuisance factors such as variation in capture efficiency and sequencing depth, amplification bias, batch effects, and transcriptional noise.

Feature selection is statistical procedure that can be used to select highly variable genes (HVGs). Features selection is a statistical procedure to remove less relevant genes, which are subject to technical noise, and preserve biological meaningful genes.

One is to treat each data point (cell x gene) as a random variable, i.e., xij is the expression of gene I in cell j (for i=1,…,n and j = 1,…,n), 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. The other focuses on only one of these tasks, often without explicit probabilistic modeling. Many algorithms of scRNA-seq data analysis are based on parametric models of gene expression in single cells. These algorithms assume that the expression level of a gene in a cell follows a statistical distribution that describes the probability of the observed value in the sample. While such a distribution can be inferred from given data and takes into account of the influence of technical noise, the model is often insufficient to capture variance in the real data. Also, even though for a given model, parameters defining the distribution can be re-evaluated, the process has to be repeated for each of different data sets.

Here we introduce a non-parametric feature selection method, which is only based on the summary statistics computed from given data.

We compared seven HVG analysis methods from six different packages, BASiCS, Brennecke, scLVM [6], scran [7], scVEGs and Seurat [9]. In summary, if the same data set is analyzed by two different tools, the resulting list of HVGs can be different. This can lead to conflicting conclusions if downstream analyses are performed using these differing lists of results (Yip, et al., 2018).

Currently gene expression data are being produced at a phenomenal rate. The general objective is to try to gain a better understanding of the functions of cellular tissues. In particular, one specific goal is to relate gene expression to cancer diagnosis, prognosis and treatment. However, a key obstacle is that the availability of analysis tools or lack thereof, impedes the use of the data, making it difficult for cancer researchers to perform analysis efficiently and effectively.

# Methods

We developed scGEApp using Matlab v9.5 (R2018b). Functions are completely written in native Matlab and GUIs are created with App Designer. In the development of our new feature selection algorithm, for each gene, we considered three summary statistics of expression: mean, CV, and dropout rate. Mean and CV are computed without removing zeros across cells, and the dropout rate is the fraction of cells with zero expression. Every gene is characterized by using these three feature variables, which are inter-related to each other by themselves. A 3-dimenal scatter plot can be generated to visualize the distribution of genes in the space defined by the three variables. We used real droplet-based scRNA-seq data (BioProject: PRJNA508890) to show that data points (genes) form an ‘S’-shaped manifold in the space (Fig. 1). To fit the curve, we used SPLINEFIT function (by Jonas Lundgren, https://www.mathworks.com/matlabcentral/fileexchange/13812-splinefit), which solves the problem of unwanted oscillations of curve from noisy data. We compute the shortest distance between a data point to the spline curve as the feature of the gene.

We developed wn fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

# Results

The Gene Expression Analyzer (GEA) presented in this demo is designed and developed to provide better data mining and analysis support for gene expression data. It makes three key contributions: Amongst all data mining paradigms that have been proposed and studied, clustering is the most widely adopted paradigm for analyzing genomic data [1]. Most of the studies regard cluster analysis as a one-step process, which have the assumption that one only needs to apply a clustering algorithm once to the data in order to get the desired outcome. However, real data mining and cluster involves repeated manipulation of the data, as well as the results of previous manipulations. GEA attempts to model this reality by providing a set of operations for cancer/biology researchers to conduct cluster analysis more effectively [2]. A key objective for performing cluster analysis on gene expression data is to identify candidate genes for further analysis, including clinical studies in a more traditional laboratory setting. But traditional clustering techniques do not provide too much help in identifying such genes. In contrast, the proposed GEA provides various facilities to accomplish this task. One example is the integrated support for finding Fascicles. There are two general types of data essential to effective gene expression analysis. The first type is the actual gene expression data, produced by the Microarray technique, or the SAGE. The second type is auxiliary data, such as for mapping tags to genes (UNIGENE), for mapping genes to proteins.

**Fig. 1.** **Screenshot of an execution of scGEApp.** This example has only example has only example has only example has only two.

In this demo, we will show several case studies conducted using the SAGE data set.

• Identify a list of candidate genes by contrasting libraries in the fascicle and the normal brain libraries. For some genes, the expression levels are much lower in the cancerous libraries than in the normal libraries or vice versa.

• The GEA can identify the genes that behave differently between cancerous brain tissues inside and outside of the fascicle. For some genes, cancerous libraries in the fascicle show a much lower level of expression than cancerous libraries outside. Identification of this kind of genes may lead to the discovery of different sub-types of brain cancer.

• The GEA is also able to identify a list of genes that for multiple tissue types always have higher expression levels in the cancerous libraries than in the normal ones.

• Based on libraries from different tissue types, finding dissimilarities may provide information of a unique set of genes that have great differences between cancerous and normal libraries in one tissue type, but not in the other.

These results may provide information on the different types of cancer possibly caused by different sets of genes.

In sum, the GEA provides a rather expressive algebraic environment for supporting multi-step cluster analysis on gene expression data. The algebraic framework ensures closure in that the output of some operation can form the input of other operations. In this way, with a creative combination of operations, a cancer/biology researcher can identify a very small number of genes that have a very high potential to be "'interesting" targets for further clinical analysis.

The purpose of our method is to identify genes whose expression levels vary across

single cells within a single population of cells. These cells are supposedly similar or, at

least, they are not a priori known to come from two distinct cell populations. This

scenario is significantly different from the more common experimental setup of finding

genes that are differentially expressed between two or more discrete groups of cells.

In the latter setting, one calculates the difference between the average expression

of a gene in one group and the average in the other group and wishes to show that this

difference is large compared to the variability within a group. Such a comparison can be

performed with any method suitable for differential expression analysis of ordinary, nonsingle-cell,

RNA-Seq data, e.g., edgeR21 or DESeq19 or even simply a t test. Such a test

will account for the total within-group variability, which comprises both the contributions

from technical noise and from biological cell-to-cell variability, and there is no need to

assess the extent to which either part contributes to the total variability. Consequently,

such a test does not need technical noise estimates derived from technical spike-in data.

In our setting, however, we seek to find genes that are variable within a single

population of cells. In other words, the biological variability, which was part of the

nuisance parameter in the two-group comparison setting, now becomes the parameter of

interest. Hence, distinguishing biological noise from technical noise is critical, and only

in this situation does it become necessary to resort to spike-in data to characterize the

strength of technical noise..

**2.3 Biological relevance**

Zyla, J., Marczyk, M., Weiner, J., & Polanska, J. (2017). Ranking metrics in gene set enrichment analysis: do they matter?. BMC Bioinformatics, 18(1), 256

Acknowledgements

Jianhua Huang.

Funding

This work has been supported by the …..

*Conflict of Interest:* none declared.

References

Alexandrescu,A. (2001) Modern C++ Design: Generic Programming and Design Patterens Applied. Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.

Alexandrescu,A. (2001) *Modern C++ Design: Generic Programming and Design Patterens Applied.* Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.

Alexandrescu,A. (2001) *Modern C++ Design: Generic Programming and Design Patterens Applied.* Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.

Yoo,M.S. *et al*. (2003) Oxidative stress regulated genes in nigral dopaminergic neurnol cell: correlation with the known pathology in Parkinson’s disease. *Brain Res. Mol. Brain Res.*, **110**(Suppl. 1), 76–84.

Crenshaw, B.,III, and Jones, W.B.,Jr (2003) The future of clinical cancer management: one tumor, one chip. *Bioinformatics*, doi:10.1093/bioinformatics/btn000.

Auhtor,A.B. *et al.* (2000) Chapter title. In Smith, A.C. (ed.), *Book Title*, 2nd edn. Publisher, Location, Vol. 1, pp. ???–???.

Bardet, G. (1920) Sur un syndrome d’obesite infantile avec polydactylie et retinite pigmentaire (contribution a l’etude des formes cliniques de l’obesite hypophysaire). PhD Thesis, name of institution, Paris, France.

<https://academic.oup.com/bioinformatics/article/26/9/1262/199010?searchresult=1>

<https://academic.oup.com/bioinformatics/article/26/3/432/214374?searchresult=1>

<https://academic.oup.com/bioinformatics/article/30/1/137/238151?searchresult=1>

<https://academic.oup.com/bioinformatics/article/28/19/2546/289556?searchresult=1>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258739/>

Bahar Halpern, K.*, et al.* Bursty gene expression in the intact mammalian liver. *Mol Cell* 2015;58(1):147-156.

Chang, H.H.*, et al.* Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* 2008;453(7194):544-547.

Mohammed, H.*, et al.* Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. *Cell Rep* 2017;20(5):1215-1228.

Richard, A.*, et al.* Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLoS biology* 2016;14(12):e1002585.

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.