Article

scGEApp: a Matlab app for feature selection on single-cell RNA sequencing data

**James J. Cai 1,2,**\*

1 Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843-4458, USA

2 Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843-4458, USA;

**\*** Correspondence: [jcai@tamu.edu](mailto:jcai@tamu.edu) (J.J.C.)

Received: date; Accepted: date; Published: date

**Abstract:** The recent development of single-cell technologies, especially single-cell RNA se-quencing (scRNA-seq), provides an unprecedented level of resolution to the cell type heterogenei-ty. It also enables the study of gene expression variability across individual cells within a homoge-nous cell population. Feature selection algorithms have been used to select biologically meaning-ful 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 (rdrop), 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 rdrop 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.

**Keywords:** *Populus*, heat shock transcription factors, gene duplication, gene expression, alternative splicing, single nucleotide polymorphism, protein structure, co-expression network

1. Introduction

Single cell technologies, especially single-cell RNA sequencing (scRNA-seq), have revolutionized the way biologists and geneti-cists 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.

https://www-ncbi-nlm-nih-gov.ezproxy.library.tamu.edu/pmc/articles/PMC6158772/.

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 [[1](#_ENREF_1)]. 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 [[2](#_ENREF_2)]. 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. Materials and Methods

*2.1. Characteristics genes*

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 [[1](#_ENREF_1)] 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

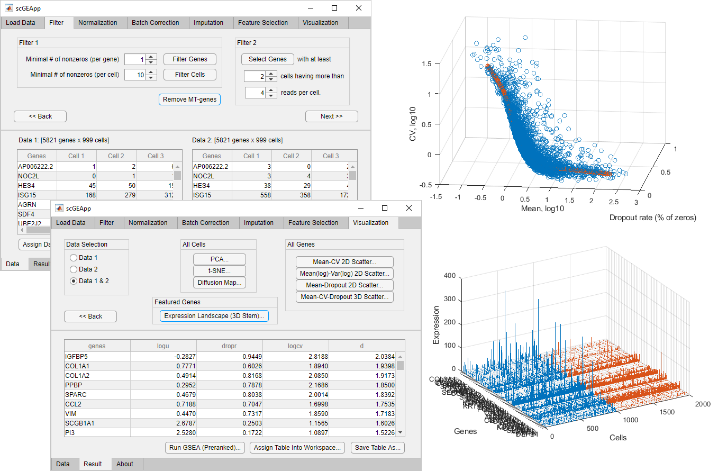
*3.1. Genome-wide duplication of Hsf genes in Populus*

Batch effects in scRNA-seq data can have severe effects on downstream analysis if not properly accounted for. Many methods such as ComBat [[3](#_ENREF_3)] and RUVwere designed to correct bulk expression data but can be applied to scRNA-seq data.. In the simplest possible case—technical replicates that were otherwise homogeneous—ComBat corrected the data and preserved the underlying biological properties (Supplementary Table 2). On biological replicates with greater batch-to-batch variability, such as two independent cultures of the same cell type, ComBat again performed well, because of its regularization for low sample numbers. A study in which ComBat was used on complex tissue data reported similar results27. Batch effect [[4](#_ENREF_4)]. [[5](#_ENREF_5)]

For complex tissue data, CCA7 and MNN8 provide generalized, nonlinear modeling approaches to align similar populations. In contrast to ComBat, both methods are independent of variations in population density7,8. Although CCA and MNN did not outperform linear methods in the small-scale examples we tested, they have potential in future large-scale data integration.

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 (rdrop) 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 rdrop) 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)[[6](#_ENREF_6)] and GSM3204305 (CCR10-) vs GSM3204304 (CCR10+ epithelial cells)[[7](#_ENREF_7)]. 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 [[8](#_ENREF_8)] 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 [[6](#_ENREF_6)] and lung airway epithelial cells [[7](#_ENREF_7)]. 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.



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

(**A**) \_\_\_\_\_\_\_\_\_. (**B**) \_\_\_\_\_\_. (**C**) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

*3.2. \_\_\_\_\_\_\_\_\_\_\_\_\_*

To \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

**Table 1. Main functions included in scGEApp and algorithms and usages of these functions.**

|  |  |  |  |
| --- | --- | --- | --- |
| Function Class | Function Syntax | Description | Reference |
| Normalization |  |  |  |
| Library size | sc\_norm(X,'type','libsiz') |  |  |
| DESeq | sc\_norm(X,'type','deseq') |  |  |
| Imputation |  |  |  |
| MAGIC | X=run\_magic(X) |  | [[9](#_ENREF_9)] |
| mcImpute | X=run\_mcimpute(X) |  | [[10](#_ENREF_10)] |
| Feature selection |  |  |  |
| HVG | T=sc\_hvg(X,genelist) | Highly variable genes | [[1](#_ENREF_1)] |
| VEG | T=sc\_veg(X,genelist) | Variably expressed genes | [[11](#_ENREF_11)] |
| Cell clustering |  |  |  |
| SC3 | C=run\_sc3(X) |  | [[12](#_ENREF_12)] |
| SIMLR | C=run\_simlr(X) | SIMLR enhances clustering performance while improving the visualization and interpretability of single-cell sequencing data | [[13](#_ENREF_13)] |
| Network construction |  |  |  |
| GENIE3 | G=run\_genie3(X) |  | [[14](#_ENREF_14)] |
|  |  |  |  |
| Batch Correction |  |  |  |
| ComBat | run\_combat |  | [[15](#_ENREF_15)] |
| HCP | run\_hcp | Hidden Covariates with Prior | [[16](#_ENREF_16)] |

4. Discussion

The broad applicability of a toolbox critical to \_\_ is the standardization of the input and output interface of methods so that users can effortlessly execute the methods of interest and compare different outs.

5. Conclusions

In this study \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

**Supplementary Materials:** The following are available online at [www.mdpi.com/xxx/s1](http://www.mdpi.com/xxx/s1),

**Figure S1.** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

**Table S1.** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

**Author Contributions:** J.J.C. conceived the project, developed the software, collected and analyzed the data, and wrote the manuscript. The author thanks Jianhua Huang, Yan Zhong and Guanxun Li for helpful discussion and inspiration during the development of this software tool.

**Funding:** This study was funded by the NIH R21AI126219 to J.J.C.

**Conflicts of Interest:** The authors declare no conflict of interest.

References

1. Brennecke, P.; Anders, S.; Kim, J.K.; Kolodziejczyk, A.A.; Zhang, X.; Proserpio, V.; Baying, B.; Benes, V.; Teichmann, S.A.; Marioni, J.C., et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **2013**, *10*, 1093-1095, doi:10.1038/nmeth.2645.

2. Yip, S.H.; Sham, P.C.; Wang, J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform* **2018**, 10.1093/bib/bby011, doi:10.1093/bib/bby011.

3. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118-127, doi:10.1093/biostatistics/kxj037.

4. Buttner, M.; Miao, Z.; Wolf, F.A.; Teichmann, S.A.; Theis, F.J. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* **2019**, *16*, 43-49, doi:10.1038/s41592-018-0254-1.

5. Liu, Q.; Herring, C.A.; Sheng, Q.; Ping, J.; Simmons, A.J.; Chen, B.; Banerjee, A.; Li, W.; Gu, G.; Coffey, R.J., et al. Quantitative assessment of cell population diversity in single-cell landscapes. *PLoS biology* **2018**, *16*, e2006687, doi:10.1371/journal.pbio.2006687.

6. Hagai, T.; Chen, X.; Miragaia, R.J.; Rostom, R.; Gomes, T.; Kunowska, N.; Henriksson, J.; Park, J.E.; Proserpio, V.; Donati, G., et al. Gene expression variability across cells and species shapes innate immunity. *Nature* **2018**, *563*, 197-202, doi:10.1038/s41586-018-0657-2.

7. Habiel, D.M.; Espindola, M.S.; Jones, I.C.; Coelho, A.L.; Stripp, B.; Hogaboam, C.M. CCR10+ epithelial cells from idiopathic pulmonary fibrosis lungs drive remodeling. *JCI Insight* **2018**, *3*, 2379-3708, doi:10.1172/jci.insight.122211.

8. Zyla, J.; Marczyk, M.; Weiner, J.; Polanska, J. Ranking metrics in gene set enrichment analysis: do they matter? *BMC Bioinformatics* **2017**, *18*, 256, doi:10.1186/s12859-017-1674-0.

9. van Dijk, D.; Sharma, R.; Nainys, J.; Yim, K.; Kathail, P.; Carr, A.J.; Burdziak, C.; Moon, K.R.; Chaffer, C.L.; Pattabiraman, D., et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **2018**, *174*, 716-729 e727, doi:10.1016/j.cell.2018.05.061.

10. Mongia, A.; Sengupta, D.; Majumdar, A. McImpute: Matrix Completion Based Imputation for Single Cell RNA-seq Data. *Frontiers in genetics* **2019**, *10*, 9, doi:10.3389/fgene.2019.00009.

11. Chen, H.I.; Jin, Y.; Huang, Y.; Chen, Y. Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC genomics* **2016**, *17 Suppl 7*, 508, doi:10.1186/s12864-016-2897-6.

12. Kiselev, V.Y.; Kirschner, K.; Schaub, M.T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K.N.; Reik, W.; Barahona, M.; Green, A.R., et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **2017**, *14*, 483-486, doi:10.1038/nmeth.4236.

13. Wang, B.; Zhu, J.; Pierson, E.; Ramazzotti, D.; Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* **2017**, *14*, 414-416, doi:10.1038/nmeth.4207.

14. Huynh-Thu, V.A.; Irrthum, A.; Wehenkel, L.; Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PloS one* **2010**, *5*, doi:10.1371/journal.pone.0012776.

15. Stein, C.K.; Qu, P.; Epstein, J.; Buros, A.; Rosenthal, A.; Crowley, J.; Morgan, G.; Barlogie, B. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics* **2015**, *16*, 63, doi:10.1186/s12859-015-0478-3.

16. Mostafavi, S.; Battle, A.; Zhu, X.; Urban, A.E.; Levinson, D.; Montgomery, S.B.; Koller, D. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PloS one* **2013**, *8*, e68141, doi:10.1371/journal.pone.0068141.

|  |  |
| --- | --- |
| copyRight | © 2019 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). |