*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 gene expression technologies, and especially single cell transcriptomics, have revolutionized the way biologists and clinicians investigate organs and organisms, allowing an unprecedented level of resolution to the description of cell demographics in both healthy and diseased states.. The microarray technology allows the high-throughput quantification of the mRNA level of thousands of genes under dozens of conditions, generating a wealth of data which must be analyzed using some form of computational means. A popular framework for such analysis is Matlab, a powerful computing language for which many functions have been written. However, although complex topics like neural networks or principal component analysis are freely available in Matlab, functions to perform more basic tasks like data normalization or hierarchical clustering in an efficient manner are not. The MatArray toolbox aims at filling this gap by offering efficient implementations of the most needed functions for microarray analysis. The functions in the toolbox are command-line only, since it is geared toward seasoned Matlab users.  Fluorescent and luminescent reporter gene systems in combination with automated microplate readers allow real-time monitoring of gene expression on the population level at high precision and sampling density. This generates large amounts of data for the analysis of which computer tools are missing to date.  **Results:** We have developed WellReader, a MATLAB program for the analysis of fluorescent and luminescent reporter gene data. WellReader allows the user to load the output files of microplate readers, remove outliers, correct for background effects and smooth and fit the data. Moreover, it computes biologically relevant quantities from the measured signals, notably promoter activities and protein concentrations, and compares the resulting expression profiles of different genes under different conditions.  **Availability:** <https://github.com/jamesjcai/scGEApp>  **Contact:** jcai@tamu.edu  **Supplementary information:** Supplementary data are available at *Bioinformatics* online. |

# Introduction

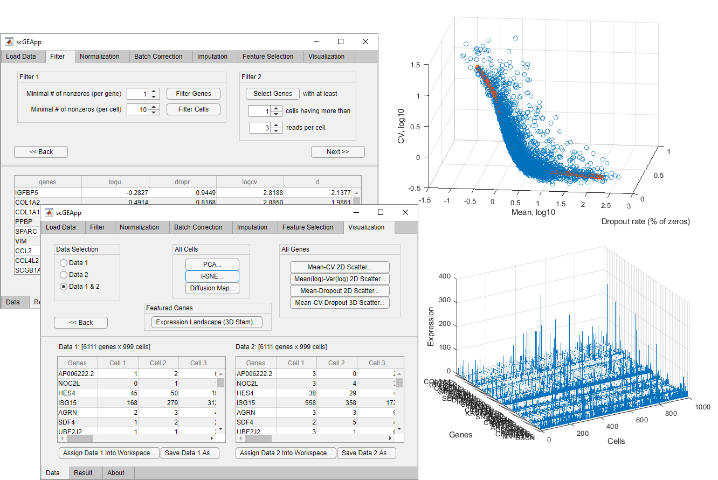
Up to 2 pages; this is approx. 1,300 words or 1,000 words plus one figure): Applications Notes are short descriptions of novel software or new algorithm implementations, databases and network services (web servers, and interfaces). Software or data must be freely available to non-commercial users. Availability and Implementation must be clearly stated in the article. Authors must also ensure that the software is available for a full TWO YEARS following publication. Web services must not require mandatory registration by the user. Additional Supplementary data can be published online-only by the journal. This supplementary material should be referred to in the abstract of the Application Note. If describing software, the software should run under nearly all conditions on a wide range of machines. Web servers should not be browser specific. Application Notes must not describe trivial utilities, nor involve significant investment of time for the user to install. The name of the application should be included in the title.

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

MATLAB Version: 9.5.0.944444 (R2018b) 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. 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. 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. 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/>