Enhancement of the Single Cell Toolkit (SCTK)

Adam Lewis[1,2], David Jenkins[3] , Dr. Evan Johnson[3], Dr. Joshua Campbell[3]
Clark University[1], Boston University Bioinformatics BRITE REU Program, Summer 2018[2] ,
Boston University Bioinformatics Program[3]

Abstract

Bulk RNA sequencing, or bulk RNA-seq, is a well-explored sequencing pipeline that has many academic and medical uses, but does not provide adequate information about expression in specific cells. Single Cell RNA sequencing (scRNA-seq) is a revolutionary process that allows for specific cell type identification, and to detect specific alterations in expression. Since the process is relatively new, there is not a single standard technique or pipeline in either data acquisition or downstream analysis. To address this problem, the single cell toolkit was created to consolidate many downstream analyses into one easy-to-use R application. The toolkit has a variety of functions that filter the data and provide visualizations such as basic heatmaps, boxplots, and scatterplots, as well as providing dimension reducing visualizations such as principal component analysis (PCA) and t distributed stochastic neighbor embedding (tSNE), which try to show the greatest possible variance of the data in two dimensions.

While the toolkit was fully functional, the toolkit was not fully tested or completed. This project's goal is to use a dataset to find flaws in the toolkit as well as to address those flaws. We then tested the toolkit with the ischemic spleen sensitivity dataset from the Human Cell Atlas, a dataset of 2000 cells sequenced using 10x sequencing. The toolkit was used to determine what types of cells were contained within the sample. Using the filter and dimensional reduction tools, several types of immune cells were identified, and a weakness of the toolkit was found.

Background

Historically, high throughput transcriptomics was limited to bulk RNA sequencing. While this was satisfactory among fields such as comparative transcriptomics, where the comparison of tissues between species means an average expression is enough, in many other areas of research in which bulk RNA seq is not acceptable data. Certain tissues contain multiple subtypes of cells, and an average does not properly explain the function of each. Any method of single cell RNA sequencing was low throughput, and therefore not very useful to examining large scale transcriptomics. More recently, however, scientific advances have allowed for high throughput single cell RNA sequencing, known as scRNA-seq.

Methods

First we did quality assurance testing for existing functions in the toolkit. This was done by using a several test datasets and testing different function settings to make sure all worked properly. The use of several datasets guarantees generality of the program. Next, to ensure that the program works on a variety of datasets, and to ensure that the program has broad utility, we used an unused dataset to test the program's functionality, followed by a dimensional reduction analysis to test practicality.

Results

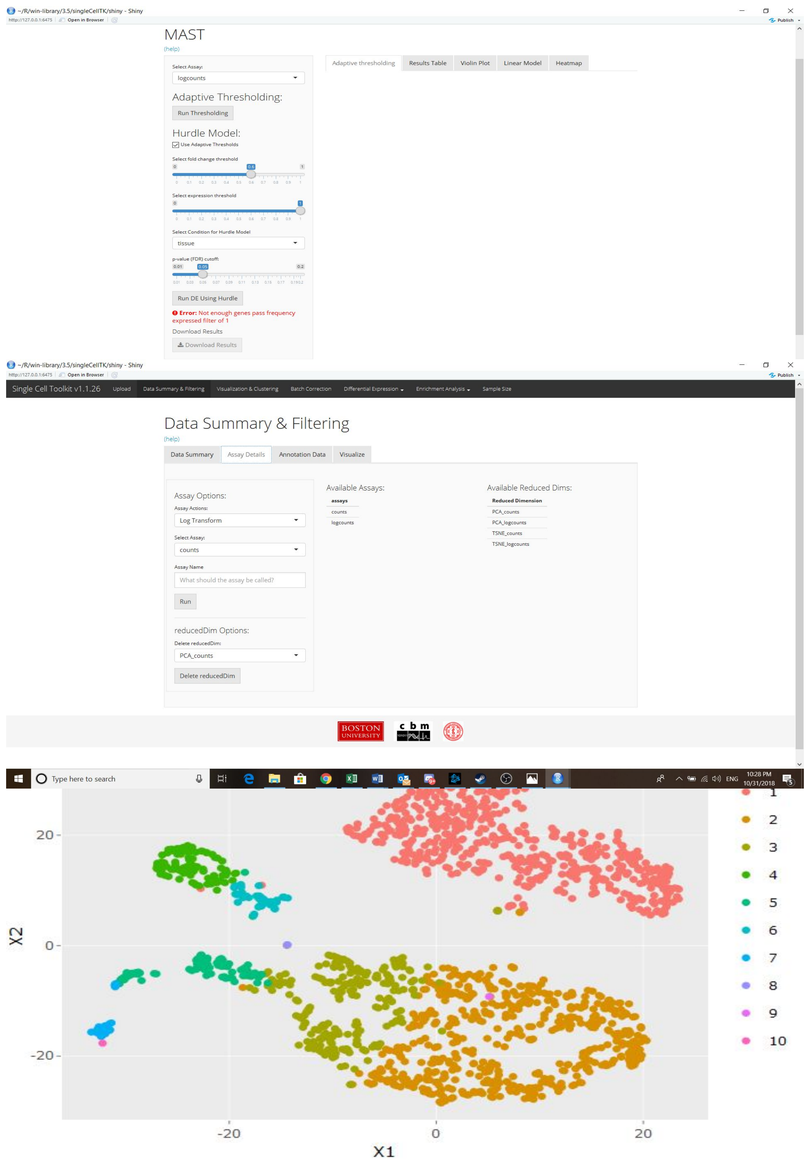

Figure 1. The updated error message in the MAST function. Previously, an unintuitive message was displayed.

Figure 2. The updated assay details page. This page was added in order to make manipulating assays to work with the program.

Figure 3.The tSNE plot of the ischemic dataset. Each color is a different cluster of cells.

Conclusions

The toolkit had several bugs that required fixing. During MAST analysis, if all genes were filtered out, the program would give an unintuitive error. This was corrected, as shown in Fig 1. There was another error in which assay names were required to follow a convention, despite no convention being standard practice. We added a function that allowed the user to rename functions, as that solved the problem as well as allowed the program to have more utility, as seen in Fig 2. After, we moved on to the ischemic spleen data set, a dataset made public by the Human Cell Atlas. The program was able to work with the data, meaning that the program has general usability. Therefore, we continued with dimensional reduction analysis. As seen in Fig 3., ten clusters were recognized using tSNE.

Future Research

In order to make the toolkit's dimensionality reduction feature easier to use, we are in the process of adding a new function that allows users to view most variable genes via a heatmap. Once a user knows the most variable genes, it becomes easier to identify cell types and subtypes using the tSNE and PCA functions. Initially, standard deviation will be used to measure variability, however other functions will be added during future development.

References

Tung, Po-Yuan, et al. "Batch effects and the effective design of single-cell gene expression studies." *Scientific reports* 7 (2017): 39921.

https://preview.data.humancellatlas.org/

https://hemberg-lab.github.io/scRNA.seq.course/cleaning-the-expression-matrix.html#dealing-with-confounders

Acknowledgements

This work was funded, in part, by NSF grant DBI-1559829, awarded to the Boston University Bioinformatics BRITE REU program, and NIH grant U01CA220413.