**Visualization of single cell RNA-sequencing data from tens of thousands of cells in real time (Group 7 Report)**

Group Lead: Grace

Group Members: Celia, JC, Ogan, Ed, Shing, Ben, Erin

**Abstract**

Grace’s Version

Characterizing the transcriptome of individual cells is fundamental to understanding complex biological systems such as the nervous system, the immune system and cancer. Recent developments in single cell RNA sequencing (scRNA-seq) technologies have enabled profiling of thousands of cells in a single sample. However, existing analysis tools are designed for datasets with much lower cell numbers, and to be used by people with a programming background, limiting their adoption. There is a strong need for an intuitive user interface to facilitate real-time data visualization and exploration by biologists to accelerate the discovery cycle of scRNA-seq analysis.

We have developed a visualization application, SINGLE CELL VIS, that can be used for scRNA-seq data from thousands of cells in real time. It takes the gene-barcode matrix and the clustering result as input. Users can visualize the cells using t-distributed stochastic neighbour embedding (t-SNE) plots. They can explore the expression pattern of specific genes. In addition, they can investigate the identity of cell clusters by examining genes that are specific to a cluster.

The application uses the plotly tool with Shiny to enable interaction over a web browser (currently tested on Firefox and Chrome). The source code is open on https://github.com/hackseq/2016\_project\_7. All plots can be saved and tables exported. We have tested single cell datasets with thousands of cells, but the application can be extended to support the interactive analysis of tens of thousands of cells.

**Scientific Abstract**

Single-cell RNA sequencing technologies have become a valuable tool in measuring the heterogeneity and similarity within cell populations. The visualization of single-cell data a crucial to fully understanding the molecular heterogeneity of a cell population, and allows the effective query and fractionation of data. t-distributed stochastic neighbour embedding (t-SNE) is a machine learning algorithm that allows for dimension reduction in big data (van der Maaten, 2008). In our application OUR CATCHY NAME, we examine the difference between the RNA-sequencing profiles of single cells and allow for visualization of single-cell RNA seq.

The interactive web application is written in R with the R packages shiny, readr, pdlyr, Matrix, plotly, magritter, vioplot, and DT. The application is hosted on the shiny server, allowing for interactive data analysis. Upon opening the application, the user is directed to a t-SNE representation of their data. This is linked to a bar graph od the number of cells in each cluster, identified as cell types within the t-SNE by expression of specific marker genes for each cell type.

Additionally, the user is able to analyze the specific expression by searching for expression of a gene of interest. The data is displayed as a heatmap for each gene, customizable by colour and percentile-gene-expression maximums and minimums. Histograms of specific gene expression data are available by cellular subset. The user is also able to custom-select regions of the t-SNE plot to compare gene expression of selected areas. The user is also able to select pre-defined data clusters, if desired. Gene expression values are available for download in .tsv format, and the data for the top 10 differentally-expressed genes between the selected groups can be visualized by histogram. The relative cluster components of each groups are displayed by bar graph. In this application, all produced graphics can be exported in .png format from Chrome or Firefox.

**Introduction**

Single-cell RNA sequencing technologies have become an interesting method of measuring the heterogeneity and similarity within cell populations. The advent of this technology has allowed for the quantification of gene expression between similar and dissimilar cell populations. These heterogeneities can be used to analyze de novo cell type discovery by genomic cell markers. The visualization of single-cell data a crucial to fully understanding the molecular heterogeneity of a cell population, and allows the effective query and fractionation of data.

t-distributed stochastic neighbour embedding (t-SNE) is a machine learning algorithm that allows for dimension reduction in big data (van der Maaten, 2008). t-SNE works by assigning each datapoint to a location on a 2 or 3-dimesnional map, allowing the observer to visualize how different two data points are from each other. In this case, we examine the difference between the RNA-sequencing profiles of single cells.

With our application, allow for easy visualization of gene expression of single cells, and comparison between expression profiles of selected cells

**Methods**

Normalization of data: by logged total transcript counts to account for cell size

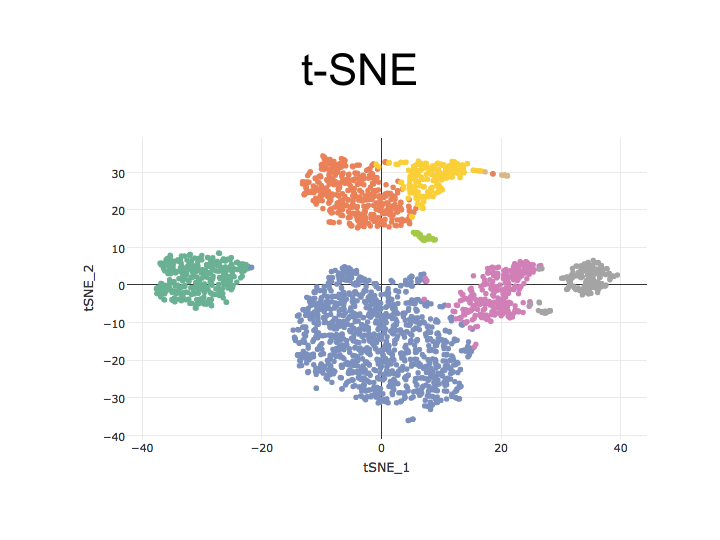
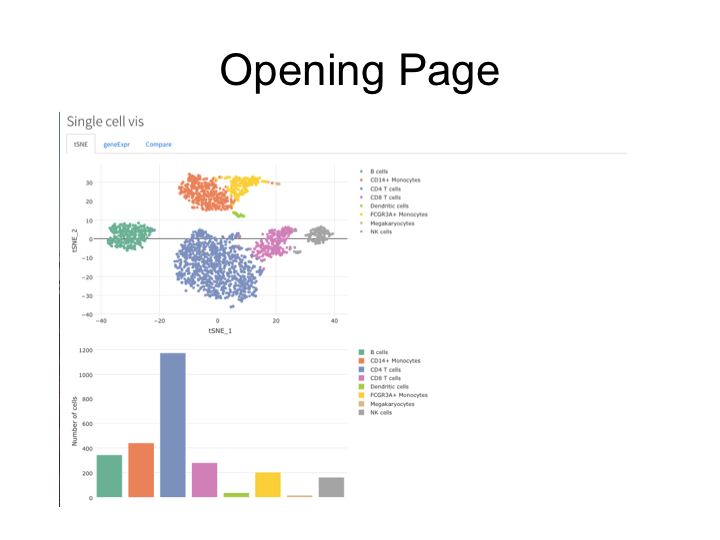
* number of zeros in the data is still an issue

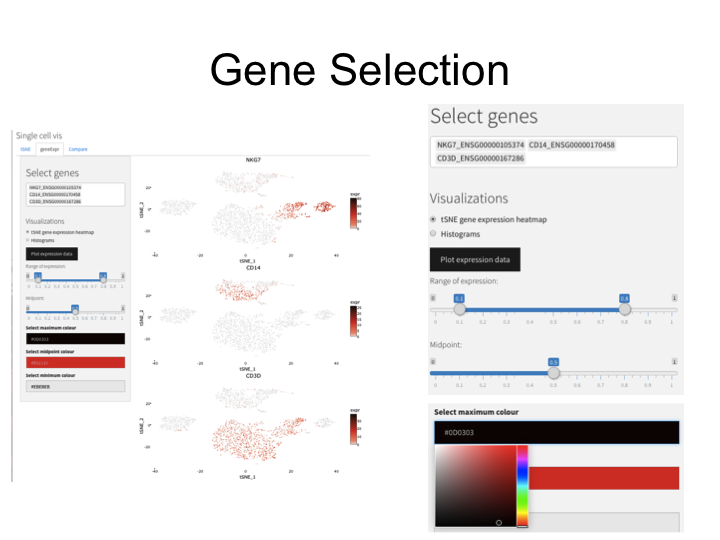
**Interface**

Upon opening the application, the user is directed to a t-SNE representation of their data. This is linked to a bar graph od the number of cells in each cluster, identified as cell types within the t-SNE by expression of specific marker genes for each cell type.

The user also able to analyze the expression of a specific gene

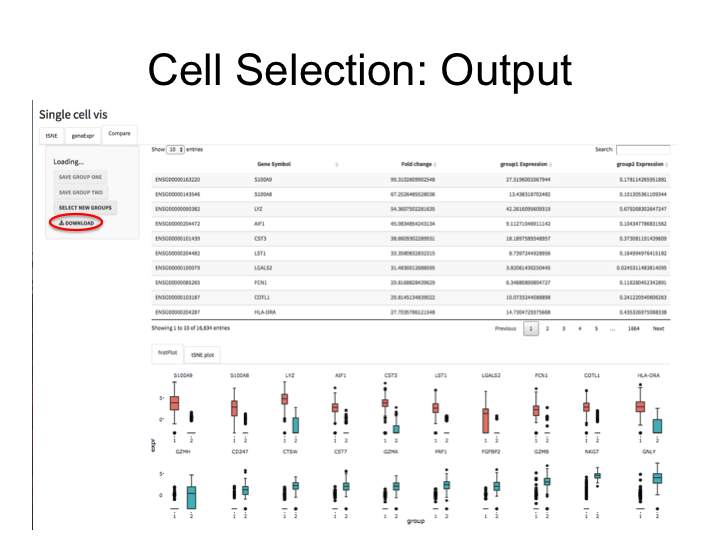
**Discussion and Conclusion**

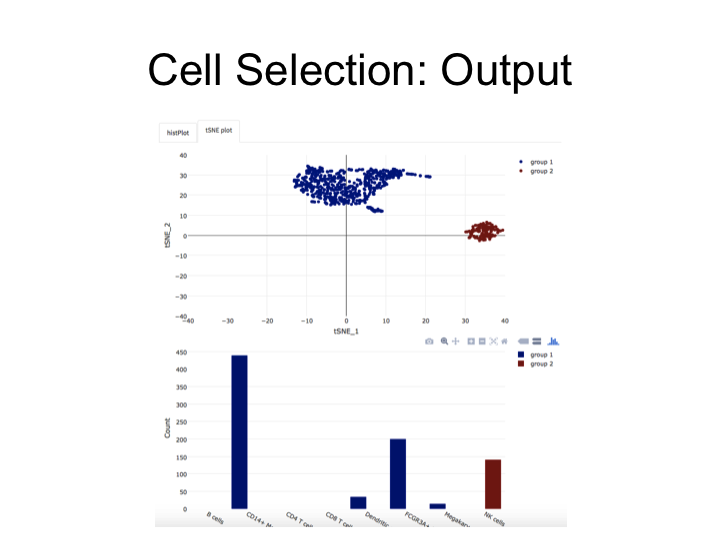
****

****

****

****

****

****

**References**

van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). ["Visualizing High-Dimensional Data Using t-SNE"](http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf) (PDF). *Journal of Machine Learning Research*. 9: 2579–2605.