STAT115 Homework 3

(your name)

2017-02-16

Part I– Single Cell RNA-Seq

For this exercise, we will be analyzing a single cell RNA-Seq dataset of mouse brain (Cortex, hippocampus, and subventricular zone) from the 10X Genomics platform. The full dataset consists of nearly 1.3M single cells, but for this assignment, we'll consider a random subset of these cells. A full description of the data is available here.

1. Describe the composition of the raw dataset (i.e. number of genes, number of samples, and dropout rate). After filtering against weakly detected cells and lowly expressed genes using reasonable parameters, how do these summary statistics change?

2. What proportion of the counts from your filtered dataset map to mitochondrial genes? Compare these values to other mitochondrial read distributions in the PBMC dataset shown in lab and in the Seurat vignette. If you determine that mitochondrial reads represent a source of unwanted variation in the filtered data, use techniques discussed in lab to remove this unwanted source of variation.

3. Perform linear dimensionality reduction (PCA) on the filtered dataset. Provide summary plots, statistics, and tables to show A) how many PCs are statistically significant, B) which genes contribute to which principle components, and C) how much variability is explained in these top components. Compare the variability in the top PCs to other scRNA-Seq datasets.

3a. [Graduate students] Determine which PCs are heavily weighted by cell cycle genes. Provide plots and other quantitative arguments to support your argument. Assuming that cell cycle is a source of unwanted variation in the data, how could you correct for it?

4. Perform a non-linear dimensionality reduction (tSNE) using the principle components as features. Visualize the cells and their corresponding tSNE coordinates and comment on the number of cell clusters that become apparent from the visualization. Are the number of clusters that form robust when rerunning tSNE?

5. Using the principle components as features, perform a clustering algorithm of your choice (either supervised or unsupervised) to uncover potential subpopulations in this data. How many cells become assigned to each group? Visualize these clusters on the tSNE graph.

6. Using differential expression analyses between clusters, identify putative biomarkers for each cell subpopulation. Visualize the gene expression values of these potential markers on your tSNE coordinates. Comment on any cluster heterogeneity or rare subpopulation characteristics based on these gene expression values.

6a. [Graduate students] Based on the data-driven characteristics of your cell clusters, provide a putative biological annotation (e.g. hippocampal cells) to the identified populations. This paper may serve as a good resource as well as the Allen Brain Atlas.

7. Seurat is one of many analysis packages for scRNA-Seq. As many of these frameworks are very young, what feedback do you have to improve the user experience of single cell analyses?

Submission

Please submit your solution directly on the canvas website. Please provide your code (.Rmd) and a pdf file for your final write-up. Please pay attention to the clarity and cleanness of your homework. Page numbers and figure or table numbers are highly recommended for easier reference.

The teaching fellows will grade your homework and give the grades with feedback through canvas within one week after the due date. Some of the questions might not have a unique or optimal solution. TFs will grade those according to your creativity and effort on exploration, especially in the graduate-level questions.