Bioinformatics 525 – Module 3 – Homework 4
Due 4/20/2017, 5 pm

Work through the exercises below, and provide your answers in either paper or electronic form to Lauren Jepsen (ljepsen@umich.edu) by the deadline.

Read the accompanying singlecell_dat.csv file into R. This file contains single-cell RNA-seq data for about 39,000 transcripts on each of ~95 cells each in S phase, G1 phase, or G2/M phase. The data are originally from Buettner et al., Nat. Biotechnol. 33:155-160, 2015; I have normalized the data and removed outliers to make it more reasonable to work with. In the loaded data, each row is a single cell, and each column except for the last has normalized counts for expression. The last column instead labels each cell as being in G1, G2/M, or S phase.

First, apply a PCA-based dimensionality reduction to the data set (since it has already been normalized, you should NOT use the scale.=TRUE argument specified in the lab assignment). Note that you need to make sure to omit the last column from the PCA and similar procedures. You can find out the number of rows and columns in a data frame with dim(dataframe).

1. Does the PCA appear to give meaningful dimensionality reduction? How many components are required to capture 75% of the variance in the data?

Plot the data set along the first two components, with the points colored by the label column of the data frame. (Hint: use levels(as.factor(dataframe$label) to see the ordering of the set labels; they will be colored black, red, and green, in order, if you plot like we did in the lab. Alternatively, you can use ggplot or a different method that automatically labels the colors.)

2. Include the plot in your solution, along with a legend identifying the growth phase corresponding to each color. Would it be possible to separate the populations using only these dimensions?

For the sake of computational efficiency, at this point we will restrict ourselves to working on a feature set containing only the first 150 principal components from your PCA, instead of the complete feature list. If you named your pca object my.pca, you could extract the coordinates of each cell in this space as
new.coords = my.pca$x[,1:150]

Don't forget to turn this into a data frame and add the labels back in.

Based on your preliminary look at the data, apply an unsupervised clustering method of your choice to the data set.

3. How many clusters does your method divide the data into? What fraction of the cells appear to be reasonably classified (you may want to support your answer by giving a table of the number of cells from each class that were assigned to each cluster)? Justify your choice of which method to use and any parameters that you had to pick.

Now, apply a supervised learning method of your choice to classify the cells. Be sure to appropriately evaluate the possibility of overtraining.

4. Describe the model that you used and why you chose it. What fraction of the cells are correctly classified within your training set? What fraction would you expect to be correctly classified if the same model were applied to new data not used in training? Explain how you reached both answers.