Noah M. Jorgenson

Bioinformatics CS6643

Lab 10

I performed my analysis for all the genes by leaving ‘genes’ unspecified in most of the analysis. I could have, in retrospect, maybe just looked at the IBC or non-IBC grouped genes which may have reduced my set a bit further. At first though, I was running into an out of memory error when attempting to use the entire stromaData variable in my analysis, since there are so many pairwise distances this resulted in a huge vector of results which took up nearly 3 GB of memory (yikes). I chose to only look at the upper 95 percentile of the stromaData though by utilizing a mask generated by the genevarfilter function and this resulted in a much more reasonable mask, bringing the intial stromaData set of about 18 thousand values to about 900 values.

After I eliminated the times variable and genes (since genes are a part of the stromaData set by default), I was able to run the analysis for principal components and the various types of clustering. The generated graphs were cool but I won’t lie, a majority of the information presented was over my head. I think if I went back through I’d be a bit more thoughtful in how I filter the stromaData (i.e. how I generate the mask), and try to find more meaningful results. The K-means clustering functions and Clustergram, for example, had pretty crowded graphs, which was most likely due to me having not selected a small enough percentile or meaningful subset of the data.