Cell type classification using single cell RNAseq data from the human mammary gland.

Amanda Paulson
Biomedical Sciences Graduate Program
University of California, San Francisco
San Francisco, California
amanda.paulson@ucsf.edu

June 18, 2017

Abstract

Classical tools for identifying the most important genes in large datasets generally focus on genes that display the highest variability across samples. However, this could highlight uninteresting variations due to cell cycle or batch effects. Additionally, ignoring less variable genes could miss subtle but important variations such as master transcription factors that only increase or decrease 1-2 fold but initiate signaling cascades that result in differentiation to various cell types. The human mammary gland consists of two main types of epithelial cells: luminal epithelial cells which are responsible for producing milk and myoepithelial cells which are contractile and responsible for expelling milk. I used a dataset consisting of 22 luminal epithelial cells and 64 myoepithelial cells as a practice case for developing a machine learning algorithm. My main questions are whether or not a machine learning algorithm can classify the cells, and whether I can develop a list of biologically interesting candidate genes that the algorithm uses to classify the two cell types.

1 Introduction

The human mammary gland consists of two main types of epithelial cells: luminal epithelial cells which are responsible for producing milk and myoepithelial cells which are contractile and responsible for expelling milk. It is hypothesized that cancers arise in luminal cells while the presence of myoepithelial cells is protective against tumors. Understanding the basic transcriptomic differences between these two cell types can help us understand how tumors arise in one cell type over another as well as how the cells are related to each other in a developmental lineage hierarchy[2][5].

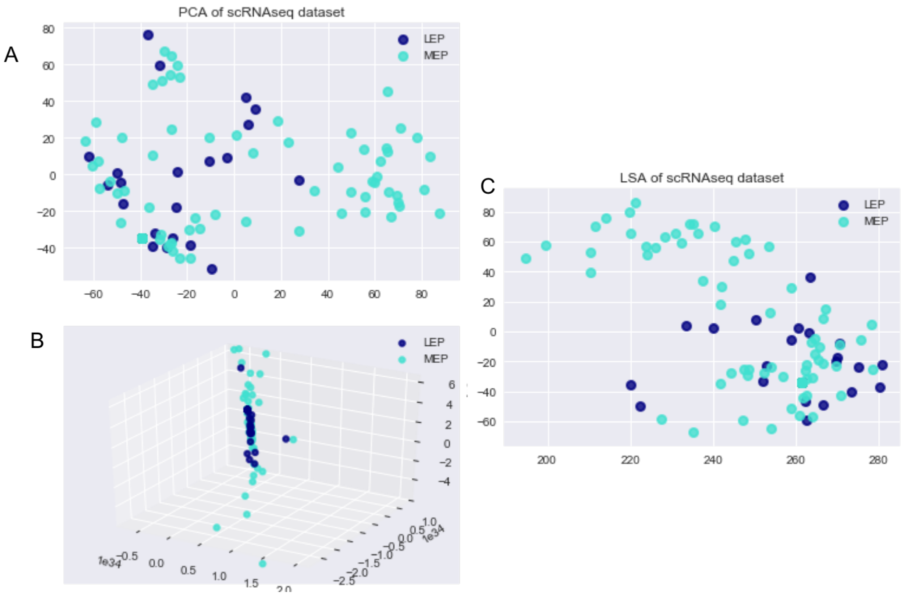

Figure 1: A: 2D PCA plot. B: 3D tSNE plot. C: 2D tSVD plot.

2 Preproccessing

Single cell RNAseq data was acquired using a plate based method in my lab. Preproccessing was done in R according to simpleSingleCell's recommended workflow[4]. I filtered out cells with library sizes or number of genes that were 3 median absolute deviations below the median (indicating poor sample preparation). I filtered out genes with a mean expression value across all cells of ¡0.4. This will select for genes that are sufficiently expressed in low numbers of cells, or that are expressed at low levels but in many cells. After filtering I was left with 86 cells and 12,799 genes.

3 Visualization of data

3.1 Dimensionality reduction techniques

In order to visualize this super high dimensional data, I tried PCA, tSNE and truncated SVD. None of the methods successfully separated LEPs and MEPs. Figure 3.1 shows 2D PCA, 3D tSNE and 2D tSVD (LSA) respectively.

3.2 Further analysis with tSVD

After visualizing data, I investigated how many components the data could be collapsed down to while still preserving all the variation. Using the tSVD, I found that the data

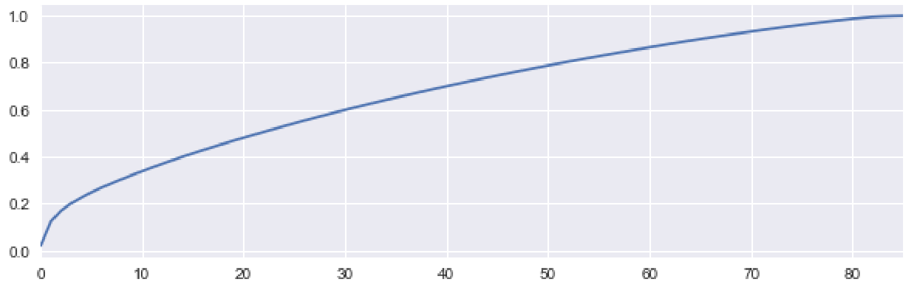

Figure 2: Variance preserved by first 86 components = 100.00%. Variance preserved by first 50 components = 78.88%

Table 1: Metrics for various machine learning algorithms

Algorithm	Precision	Recall	F1-Score
Decision Tree	0.8	0.82	0.8
Log Reg	0.8	0.82	0.8
SVM w/ RFE	0.69	0.73	0.7
SVM	0.69	0.73	0.7
SVM w/ F.S.	0.6	0.77	0.67
GridSearch SVM	0.6	0.77	0.67
kNN w/ CV	0.6	0.77	0.67
LogRegCV	0.6	0.77	0.67
Nave Bayes	0.59	0.73	0.65
LinearSVM	0.59	0.73	0.65
kNN	0.6	0.55	0.57

could be collapsed to 86 dimensions while still preserving 100 percent of the variation. Figure 2 shows a graph of variance preserved versus number of components. However, these components still do not show great separation between classes, as shown in Figure 3.

4 Machine Learning Algorithms

I tried a variety of machine learning algorithms and cross validation schemes with this data. According to the literature, a support vector machine should be the best for this data [3][1]. However, I was unable to get good classification with any of the algorithms I tried. I think this is because my data was unbalanced and I had very few samples. Cross validation optimized based on accuracy, which always resulted in 100% correct classification of MEPs but 0% classification of LEPs. Table 1 shows the metrics for various algorithms. Figure 4 shows the confusion matrix for the best scoring algorithm, logistic regression.

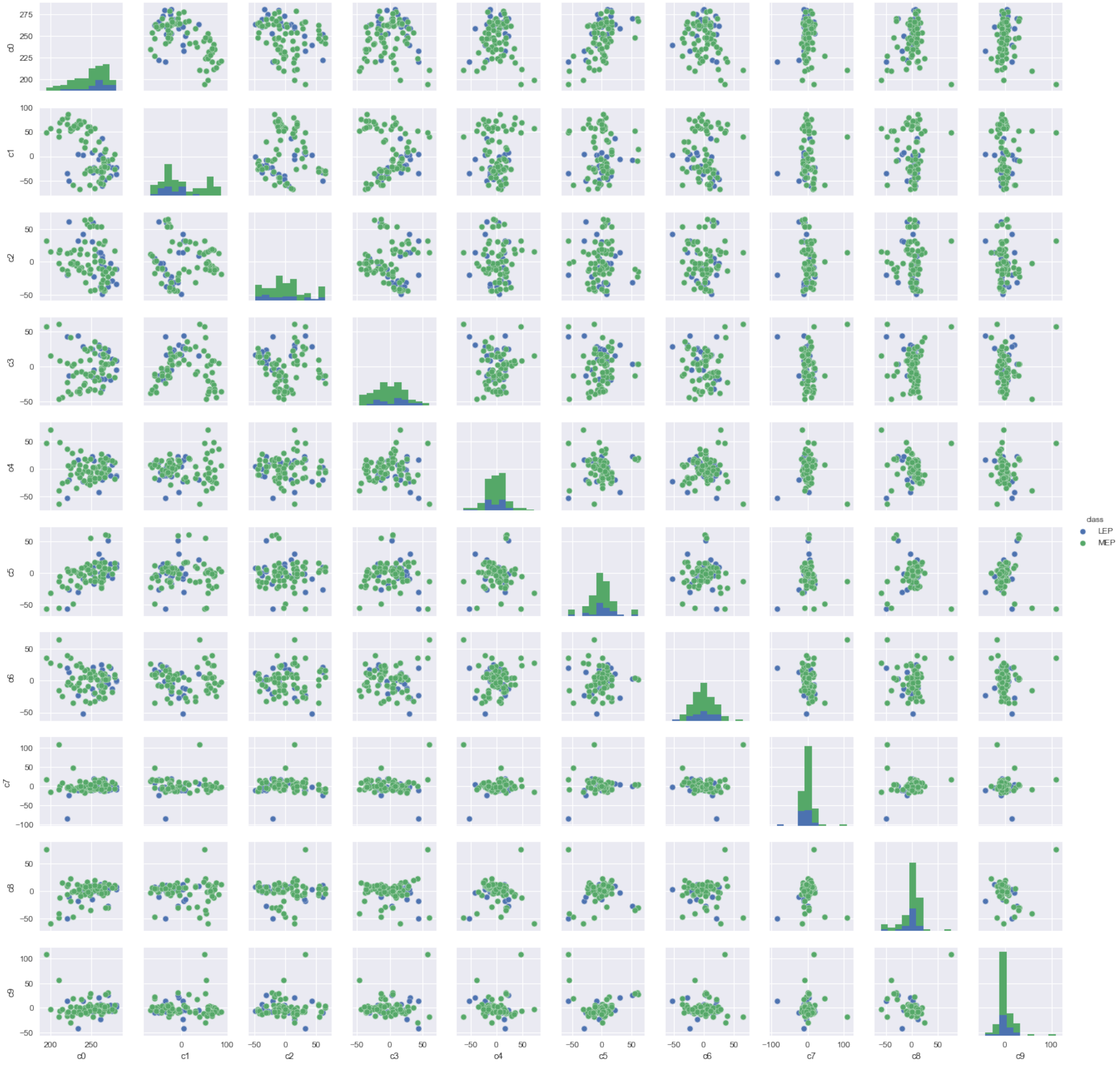

Figure 3: Pairwise plots of first ten tSVD components

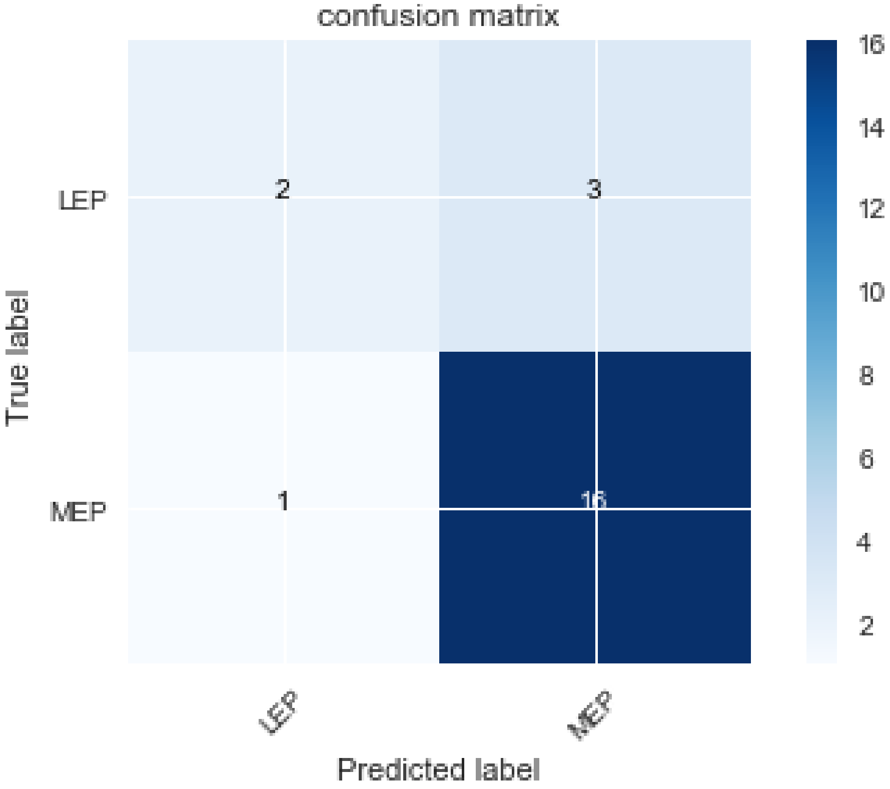

Figure 4: Confusion matrix of logistic regression algorithm

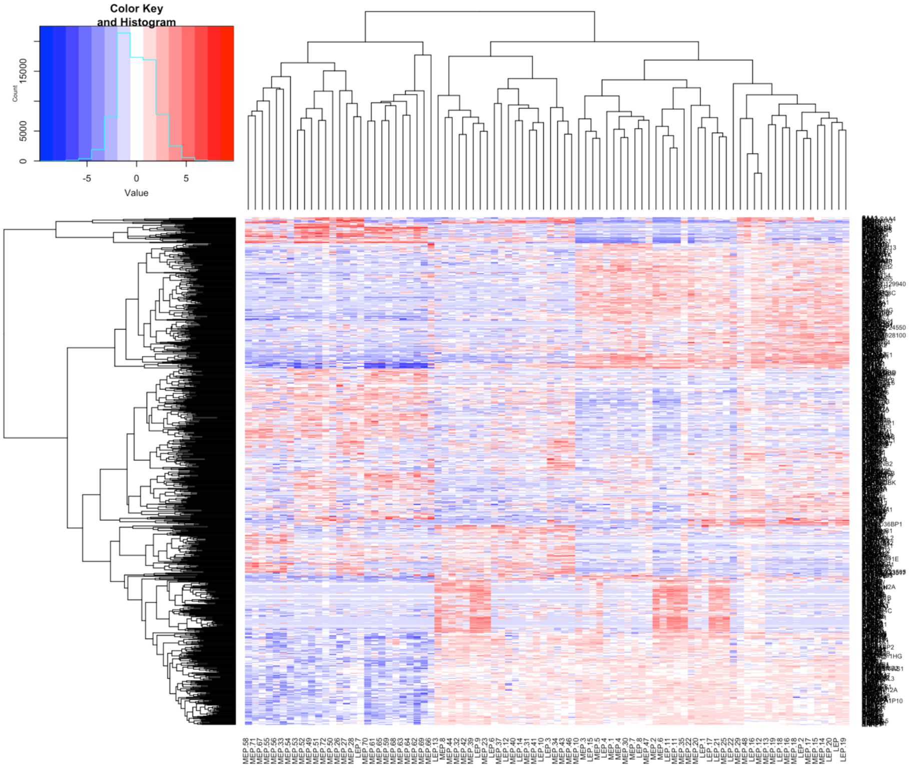

Figure 5: Hierarchical clustering of cells by 953 highly variable genes

5 Conclusion and Future Directions

Although I was not able to get good classification with these algorithms, there are still more things I could try. Adding additional samples would help immensely. Barring that, I could explore this data in other ways. For example, hierarchical clustering on the highly variable genes of these samples shows clear subpopulations (Figure 5). I could examine these subpopulations for key markers and decide whether my labels are accurate or begin to understand why these cells do not separate well.

References

[1] "Choosing the right estimator — scikit-learn 0.18.1 documentation". In: ().

[2] Thorarinn Gudjonsson et al. "Myoepithelial Cells: Their Origin and Function in Breast Morphogenesis and Neoplasia". In: *Journal of mammary gland biology and neoplasia* 10.3 (July 2005), pp. 261–272.

[3] Yongli Hu et al. "A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data". In: *BMC Genomics* 17.13 (Dec. 2016), p. 1025.

[4] Aaron T L Lun, Davis J McCarthy, and John C Marioni. "A step-by-step workflow for low-level analysis of single-cell RNA-seq data". In: *F1000Research* 5 (Aug. 2016), p. 2122.

[5] P M Siegel and W J Muller. "Transcription factor regulatory networks in mammary epithelial development and tumorigenesis". In: *Oncogene* 29.19 (May 2010), pp. 2753–2759.