Determination of cell types in Chicken Utricle

CS229 Machine Learning Project Final Report

Yanli Wang, December 11th 2015, Stanford University

INTRODUCTION

Hair cells are the final sensory cells in the cochlea that transfer the mechanical energy of vibration of air to electrochemical energy that fire the auditory nerves. Mammalian hair cells cannot regenerate after damage, for instance after exposure to loud sound or ototoxic antibiotics such as aminoglycoside, which is still an absolute essential for treatments of many fatal diseases. In contrast to mammalian hair cells, chicken utricle **hair cells (HC)** can regenerate from the **supporting cells (SC)** underneath. Furthermore, there are two visually different regions in chicken utricle - **striola (S)** and **extrastriola (ES)** regions as shown in Fig. 1, containing both HCs and SCs. The hypothesis is that type I and type II HCs are in S and ES regions respectively, where type I HCs are especially regenerative. The motivation of current work is to study the gene expression of chicken hair cells to potentially help with regeneration of human hair cells. On top of distinguishing HC and SC from gene expressions, it is also within the area of interest of current work to identify the differences between type I and type II hair cells. The first aim of the project is to identify the cell types in chicken utricle from **220** gene expression as positive real numbers from **192** cells. The genes of interest are selected based on relevance to known HCs and SCs specific genes in human, and potential relevance to regeneration. This is an ongoing research in the field, and the data is obtained from Dr. Scheibinger and Prof. Heller at Stanford Medicine School Department of Otolaryngology Head and Neck Surgery.

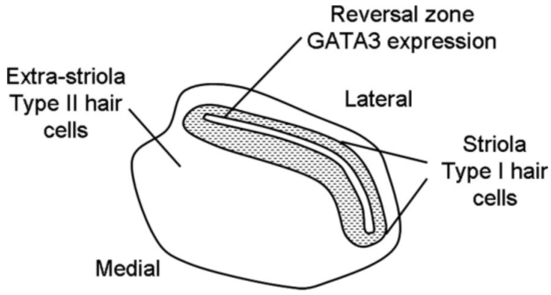

Figure 1: A schematic chicken utricle.

METHOD and RESULTS

The data is processed so that zero expressing genes across all cells and unsuccessful cells with no gene expression detected are deleted, leading to **160 observations** (cells) with **192 variables** (genes). In this case, dimension reduction and variable selection are essential to the problem. From the raw data, HC/SC type is unknown, however, each cell is known to be from either S or ES region from the dissection process. Moreover, there might be other cell types in chicken utricle. To learn about the general cell types, hierarchical clustering is performed first. All the analysis done in this work is written in R.

Hierarchical Clustering

To learn some general information about the cell types, without assuming the most prominent categories of the cells, hierarchical clustering with complete linkage is performed. The result is shown in Fig. 2. The color coding for Fig. 2 (a) is according to the gene marker - TMC1 (transmembrane-channel-like 1), which is known to express more prominently in HCs than in SCs in human cochlea, whereas, the color coding for Fig. 2 (b) is according to the S/ES regions of the cells.

From the dendrogram, it is believed that the cells can be divided mainly into three categories. From the color coding in (a), we can conclude with some confidence that the branch 1 on the left of the dendrogram are the HCs, and branch 2 in the middle are the SCs. It is hypothesized that the third branch on the right having a smaller population are the transitional cells from SC to HC. It can also be seen that S/ES difference is secondary comparing to the HC/SC difference as no pattern of S/ES is observed corresponding to the clustering. Thus, cells are categorized as HCs, SCs, and transitional cells according to the hierarchical clustering for further analysis through the rest of the text.

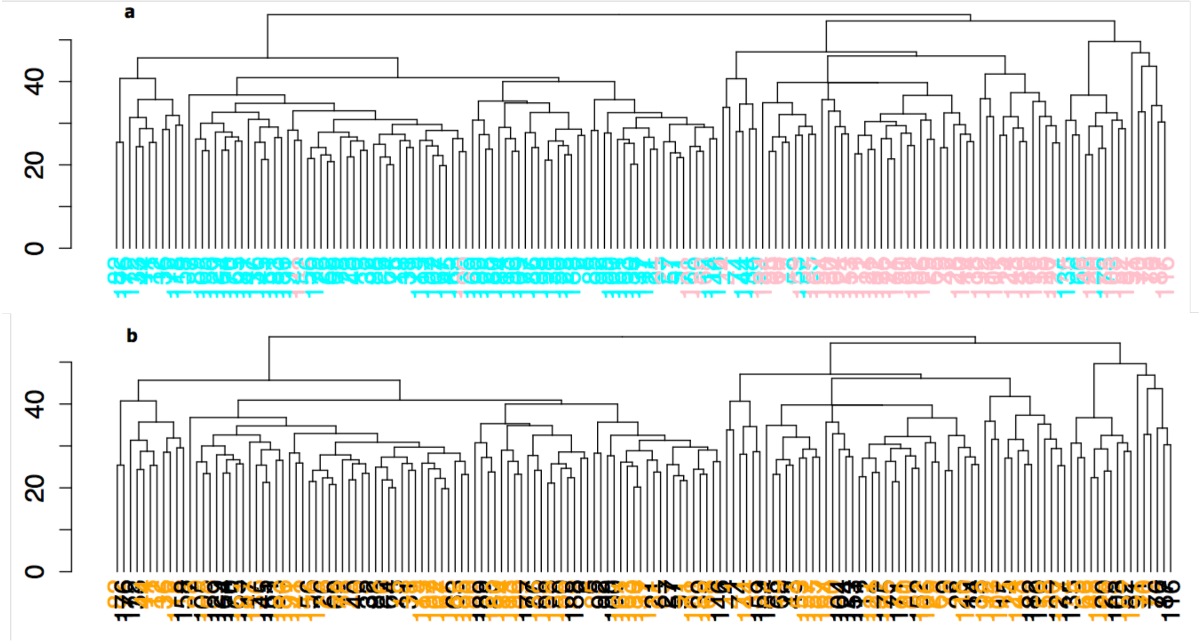

Figure 2: Dendrogram of all cells. **a**. The color coding is one of the known gene expression TMC1 (transmembrane-channel-like-1 gene) primarily found in human HCs. Cyan - with TMC1 expression, likely to be HCs; Pink - no TMC1 expression, likely to be other types than HCs. **b**. Same dendrogram with the color coding according to the regions where the cells are from. Orange - striola (S) region; Black - extrastriola (ES) region.

Principle Component Analysis

As a first step, the principle component analysis (PCA) is performed. The percentage of variance captured by the first 10 principle components (PC) are computed via bootstrap with $k = 1000$. The ranking of the principal components is by their eigenvalues. The histogram of the percentage of variance is plotted in Figure 3.

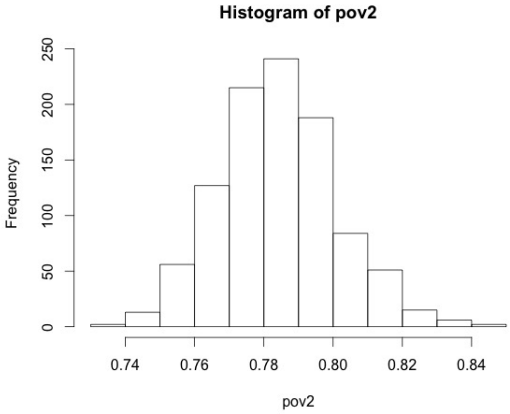

Figure 3: Histogram of percentage of variance captured by the first 10 principle components.

It can be seen that about 80% of the variance is captured by the first 10 PCs. The first 3 PC scores of the data are plotted in 2D with color-coding of HC/SC shown in Figure 4. Prominent clusters can be seen from PC1 vs. PC2 plots, where PC3 even shows the clear differences between transitional cells vs. the rest. The same plots are also color coded (not shown) according to S/ES region categories, however, no clustering according to regions has been observed even up to 10 PCs.

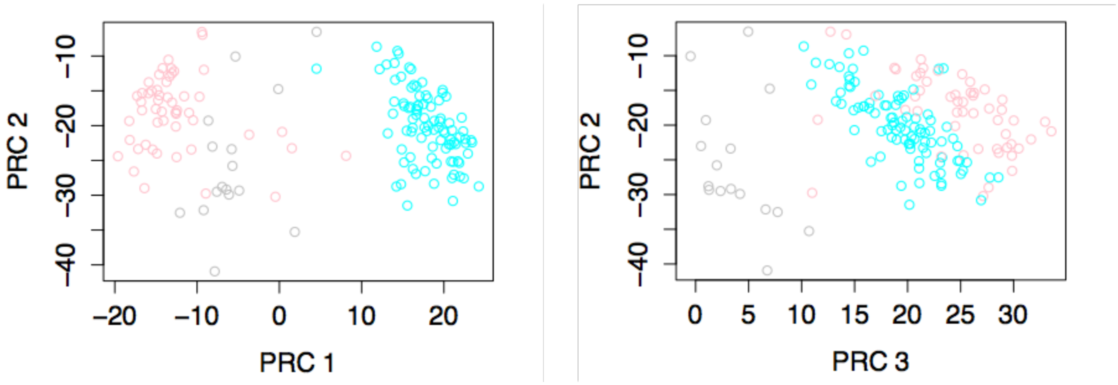

Figure 4: Scattering plot of PC 2 vs. PC 1 and PC 2 vs. PC 3, color coded with S (orange) and ES (black) region. The ranking of the principal components is by their eigenvalues.

Since HC vs. SC are the most prominent differences in the entire population, it is possible that the differences between S/ES regions can be amplified if looking only at HC or SC population. Therefore, the PC Analysis is applied to HC and SC population separately, and part of the results are shown in Fig. 5. In Fig. 5 (a), only HC population is used for PCA, and PC2 vs. PC1, PC2 vs. PC3 are plotted according to S/ES categories. Fig. 5 (b) is similar to (a) but only SC population is used. No pattern is observed for either population even up to 10 PCs.

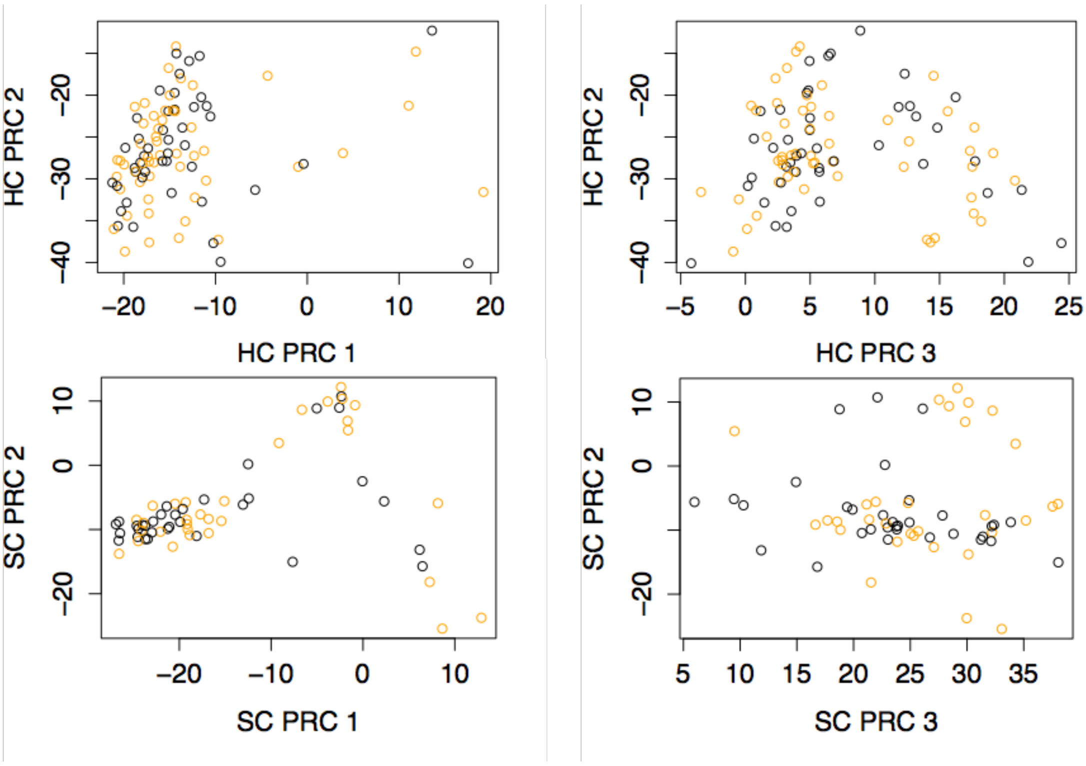

Figure 5: Scattering plot of PC 2 vs. PC 1 and PC 2 vs. PC 3, color coded with HCs (cyan), SCs (pink), and transitional cells (grey). The ranking of the principal components is by their eigenvalues.

Support Vector Machine with PCA

As the first attempt of dimensionality reduction, PC scores are used for support vector machine (SVM) algorithm. The intuition of choosing SVM is from separability of data observed in Fig. 4. Therefore, HC/SC for entire population, S/ES in HC population and SC population separately are classified using SVM. Due to small size of the data, cross-validation is performed in a slightly differently way. For 100 runs, 75% of observations are randomly chosen as training set, and the rest are used as testing set. The PCA is performed on the training set only. n eigenvectors with n largest eigenvalues are chosen using SVM. The mean and standard deviation of the 100 testing errors are reported for each n. The result for distinguishing HC/SC is very successful with only 2 PCs included as shown in Fig. 6 (a). However as seen in (b), the result for S/ES is not very successful even performed only on HC population and SC population. The SVM with polynomial kernel is also shown in grey in Fig. 6 (b). Radial kernel has also been tried, however, none of polynomial kernel or radial kernel is shown improvements to the results.

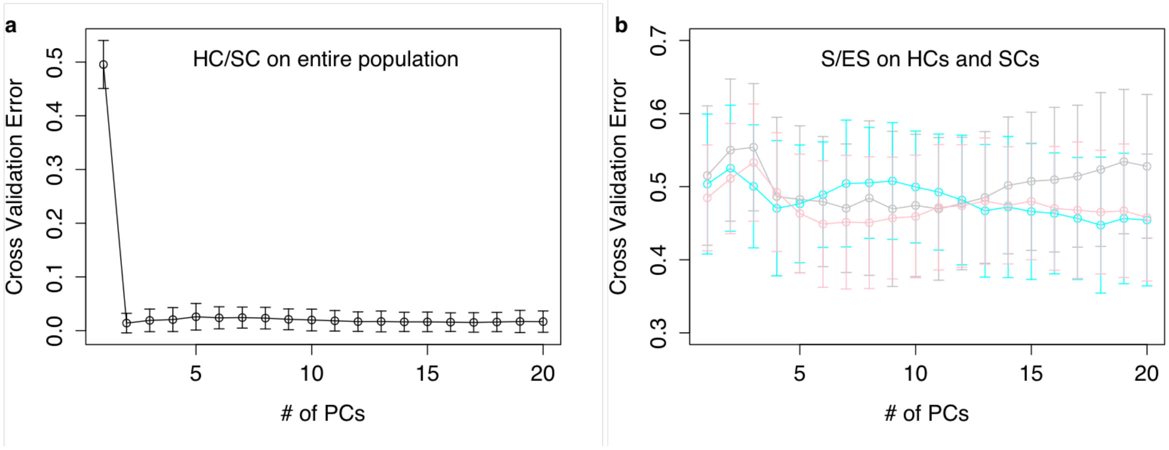

Figure 6: Cross validation error for SVM performed using PCs. **a** HC/SC classification on entire population with linear kernel. **b** S/ES classification on HC population with linear kernel (cyan); S/ES classification on SC population with linear kernel (pink); S/ES classification on HC population with polynomial kernel of degree 2 (grey).

Support Vector Machine using Forward Stepwise Selection

The down side of using principal components is loss of information in particular genes. Alternatively, we could look at the most prominent genes contributing to the PCs. However, the SVM on distinguishing S/ES region was not successful. Because the selection of the PCs is based on the percentage of variation explained, it can be thought that the major differences observed in these genes among the cells is not the regions. To identify the S/ES region, other dimension reduction or variable selection technique is needed. It will be interesting to try to identify the cell types explained by these genes, for example the four branches within HCs and three branches within SCs from the dendrogram in Fig. 2. However, for the purpose of current work, to identify the S/ES region, forward stepwise selection is performed using SVM.

The data is randomly divided into four folds. The variable with lowest testing error from SVM is selected in each iteration up to a total of 30 variables selected. Due to the small size of the observations, a small number of fold - four - is chosen so that the testing set is not too small for generating diverse testing error rates for picking the best variable. Since the number of folds is small, this process is repeated for 25 times, so that there are 100 trials for each number of variables used in SVM. The mean and standard deviation of the testing errors for each number of variables selected is shown in Fig. 7.

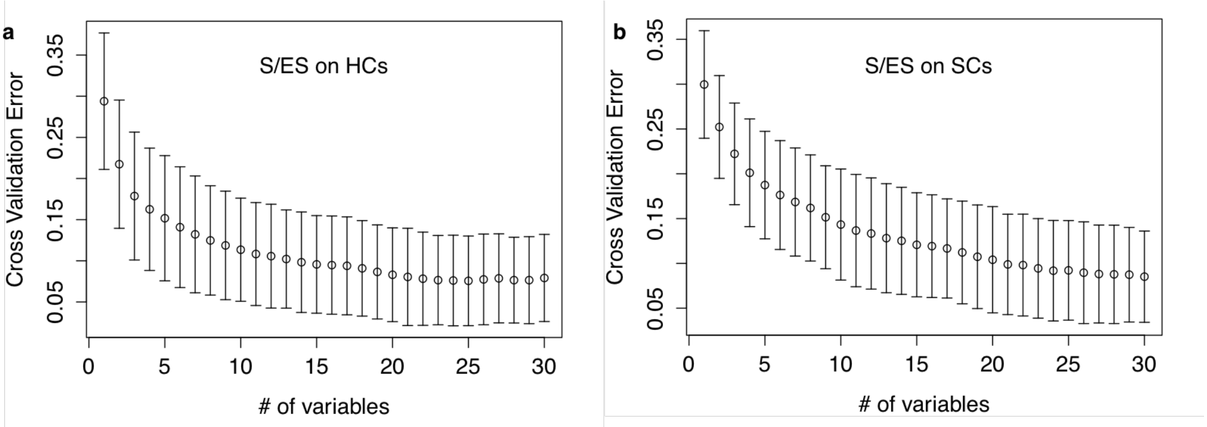

Figure 7: (a) Cross-validation error of SVM using forward step selection on HC population, and (b) cross-validation error of the same method on SC popultion.

Correlations of Genes to Cell Types

To study the relationship of genes and the cell types, the correlation of the gene expression to HC/SC and to S/ES are plotted in Fig. 8.

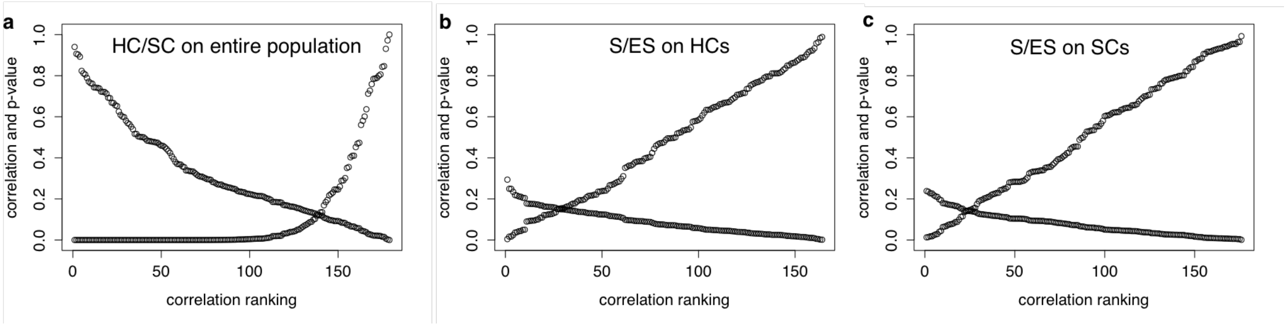

Figure 8: (a) The correlation of each gene to HC/SC category on entire population. (b) The correlation of each gene to S/ES category on HC population. (c) The correlation of each gene to S/ES on SC population.

It can be seen that there is no single strongly correlated gene expression to the S/ES category in either HC or SC population. The SVM algorithm using most correlated genes is also tested. The results are not satisfactory as might be expected.

CONCLUSION and FUTURE WORK

The first layer difference in cell types explored by these gene expressions is the category of HCs and SCs in chicken utricle. This can be seen from the hierarchical clustering in Fig. 2 and the principal component analysis in Fig. 4. The difference between cells from S and ES region is not as prominent as we originally hypothesized. There is no gene in these set of genes that is highly correlated to the S/ES classification. However, the method of forward stepwise selection is shown to be successful in determining the region of the cell.

It is important to see which genes are frequently selected by the forward stepwise selection method. The union and intersection of these sets from HC population and SC population should lead the focus of studies of genes in chicken utricle. The nature of these selected genes may provide information on the regenerative capability. For the purpose of studying the cell types in chicken utricle, it is important to study the third branch of the hierarchical clustering, hypothesized as transitional cells. The study of the differences between these cells and the supporting cells may shine light on regenerative mechanism. Moreover, It will be interesting to try to identify the cell types, for example the four branches within HCs and three branches within SCs from the dendrogram in Fig. 2. Further more, it is important to study the differences in gene expression between human HCs and chicken HCs.