

# Comparison of Gene Expression in Epithelial Cells of Phenotypic Normal Smokers vs. Normal Non-smokers

Raymond Li

Final Project

December 16, 2009

# Introduction

- Chose to analyze dataset GSE5056.
- Dataset is publicly available from Gene Expression Omnibus (GEO).
- Dataset is part of a study aimed at understanding how cigarette smoking modifies neuroendocrine cells.
- Lung cancer tumors exhibit neuroendocrine properties, and chronic smokers have increased numbers of neuroendocrine cells.
- GSE5056 was originally submitted on June 12, 2006.
- GSE5056 is part of a superseries (GSE5060).

# Background of Dataset

- Contains 44 HuGeneFL GeneChip human samples.
- MAS5 used to analyze the microarray data.
- Each sample is from the airway epithelium of phenotypically normal smokers and non-smokers with large airways.
- 18 of the 44 samples are from non-smokers.
- 26 of the 44 samples are from phenotypically normal smokers.
- 7129 genes are included in the dataset.

# Background of Study

- Of the 11 genes considered to be neuroendocrine cell-specific, only ubiquitin C-terminal hydrolase L1 (UCHL1, a member of the ubiquitin proteasome pathway) was consistently up-regulated in smokers compared to non-smokers.
- Up-regulation of UCHL1 at the protein level was observed with immunohistochemistry of bronchial biopsies of smokers compared to non-smokers.
- While UCHL1 expression was present only in neuroendocrine cells of the airway epithelium in non-smokers, UCHL1 expression was also expressed in ciliated epithelial cells in smokers.
- UCHL1 is involved in the degradation of unwanted, misfolded or damaged proteins within the cell and is over-expressed in >50% of lung cancers, its over-expression in chronic smokers may represent an early event in the complex transformation from normal epithelium to overt malignancy.

# Loading the Dataset

- Before loading the dataset into R, the annotations at the top of the file must be filtered out.
- Annotations are in this format:

```
!Series_title "Airway epithelium, large airways, phenotypically normal smokers vs  
non-smokers, MAS5 (HuGeneFL)"  
!Series_geo_accession "GSE5056"  
!Series_status "Public on Nov 21 2006"  
!Series_submission_date "Jun 12 2006"  
!Series_last_update_date "Nov 21 2006"  
!Series_pubmed_id "17108109"
```

- One easy way to deal with the annotations is to manually remove them and put them in a separate file.
- With the annotations in a separate file, the code snippets used throughout the semester can be used to load the data and to load the annotations.

# Processing Outliers

3 approaches are used to identify outliers (plots can be seen in the next few slides):

1. Correlation plot

Samples 3, 4, 5, 6, 25, 26, and 28 could be possible outliers.

2. Cluster dendrogram

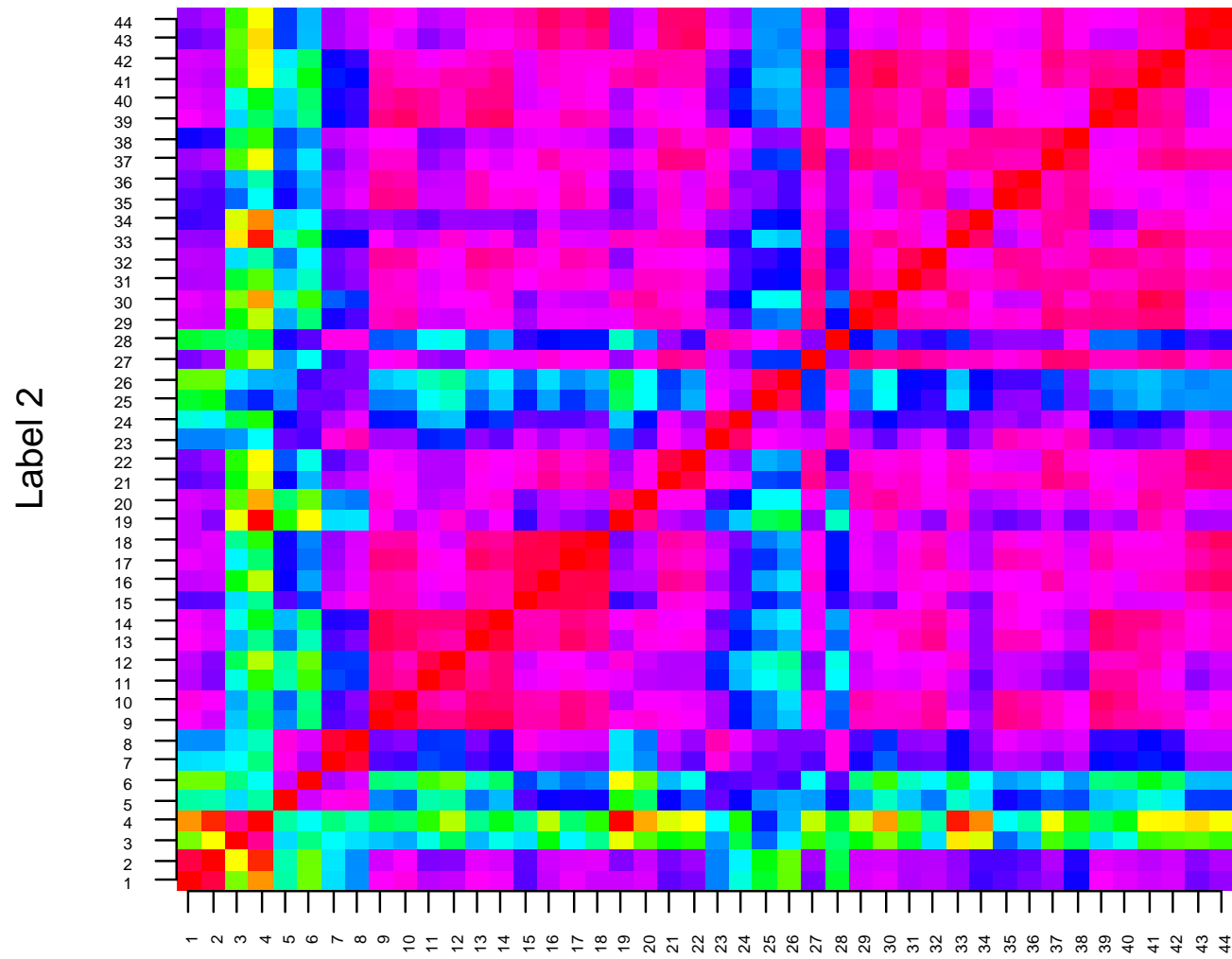
Samples 3, 4, 5, and 6 could be possible outliers.

3. CV versus Mean plot

Samples 4, possibly 3, 5, 6, and 25 could be possible outliers.

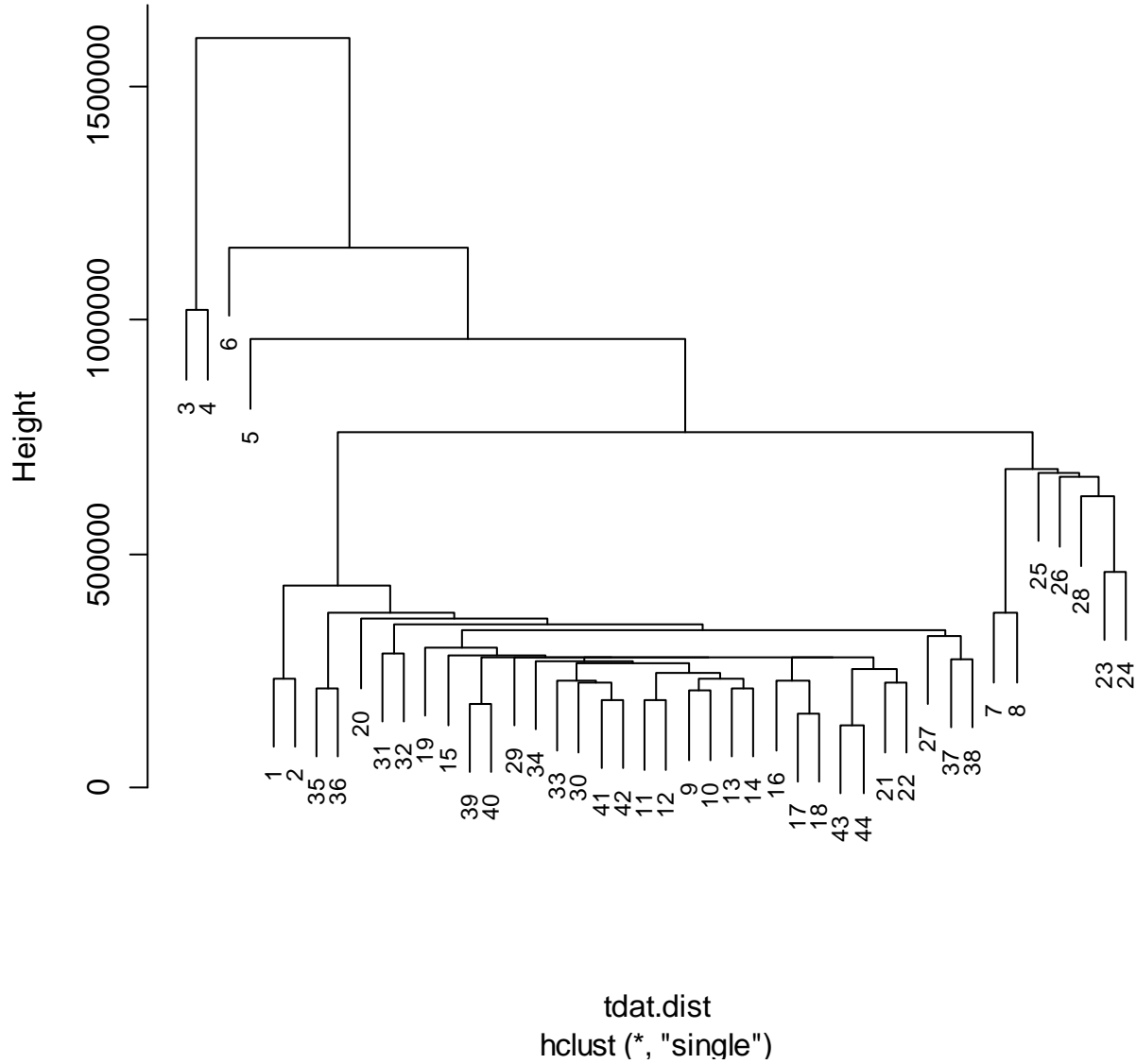
- Despite possible evidence pointing to potential outliers, they are not very strong outliers.
- Additionally, the stronger outlier candidates (samples 3, 4, 5, and 6) tend to cluster together and share similar characteristics (they are all non-smokers).
- Although they tend to show up on our radar using these 3 approaches, they exhibit characteristics that indicate they are not outliers and should be included in our dataset.

# Correlation plot for all genes

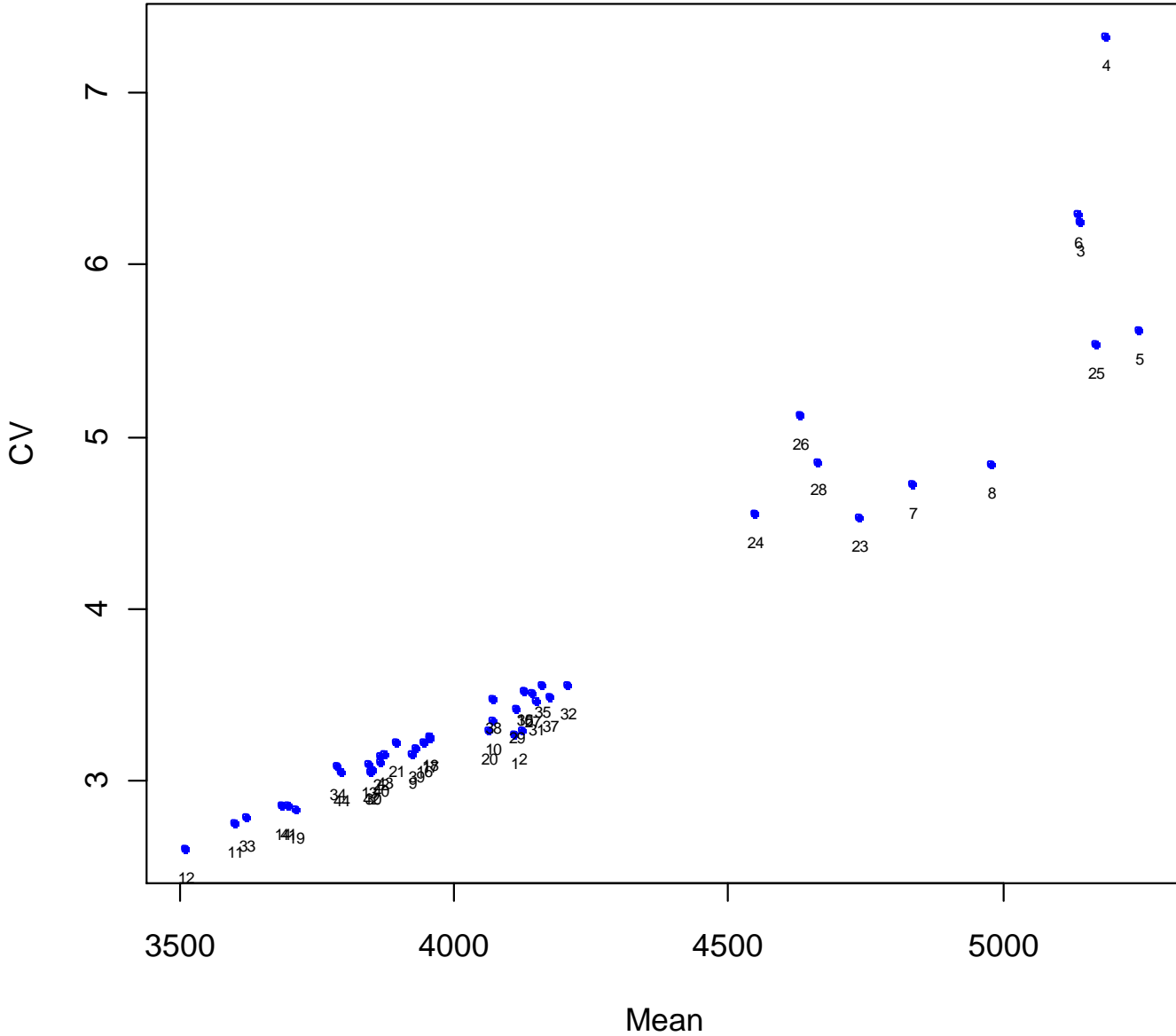


Label 1

# Cluster Dendrogram



# Sample CV vs. Mean

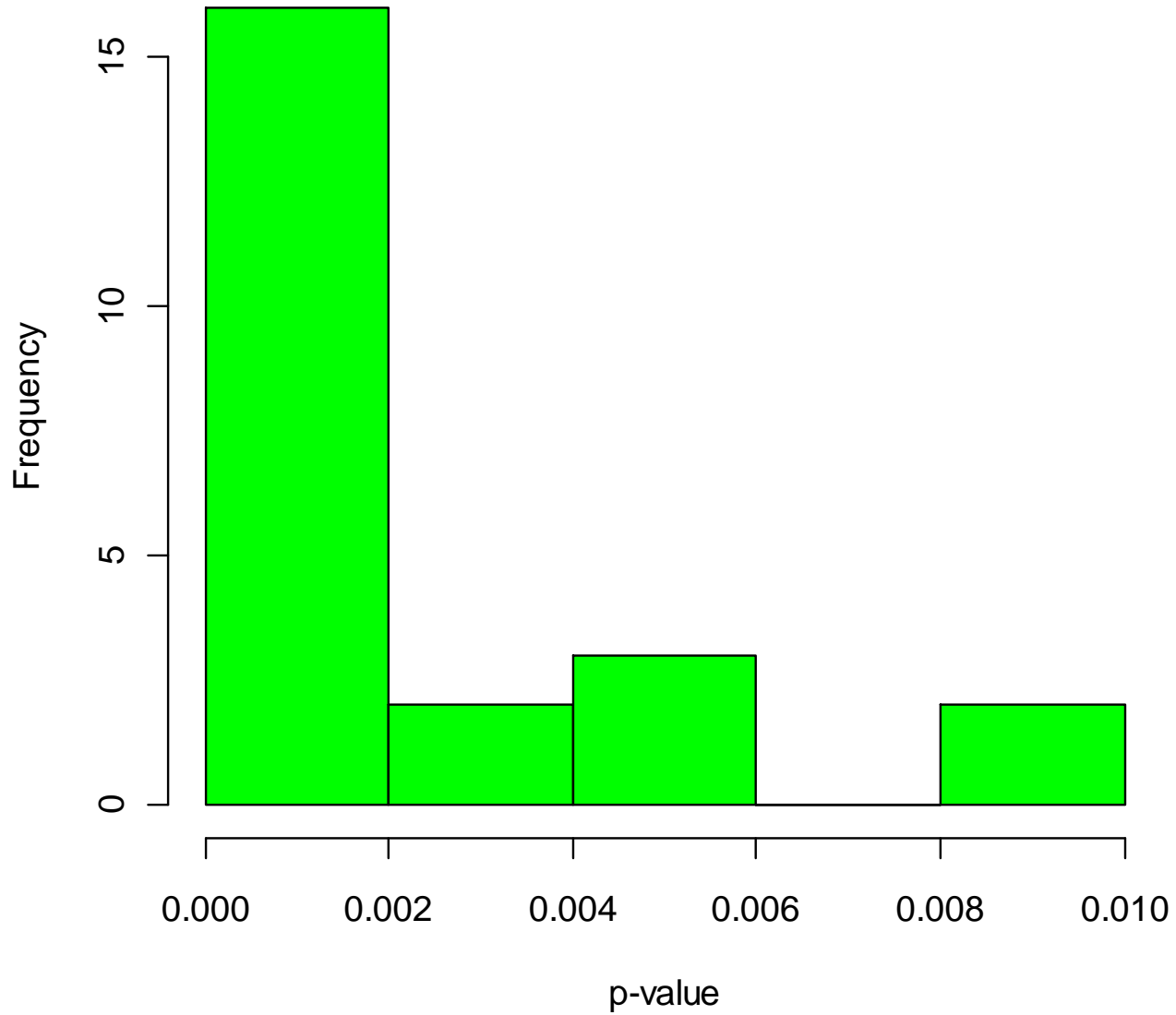


# Gene Filtering

2 approaches are used to filter out low expression genes:

1. Remove genes with an average expression value  $< 50$ .
  2. Remove genes with a coefficient of variation in the lowest quantile (5%).
- These 2 approaches leave us with 6708 out of the original 7129 genes.
  - Of the remaining 6708 genes, a Student's t-test and the Holm's method of multiple test correction is applied to further focus the list of useful genes.
  - Using a p-value threshold of  $< 0.01$ , this method of feature selection leaves 23 remaining genes (a histogram of their p-values is shown on the next slide).

### Histogram of p-values for Remaining Genes

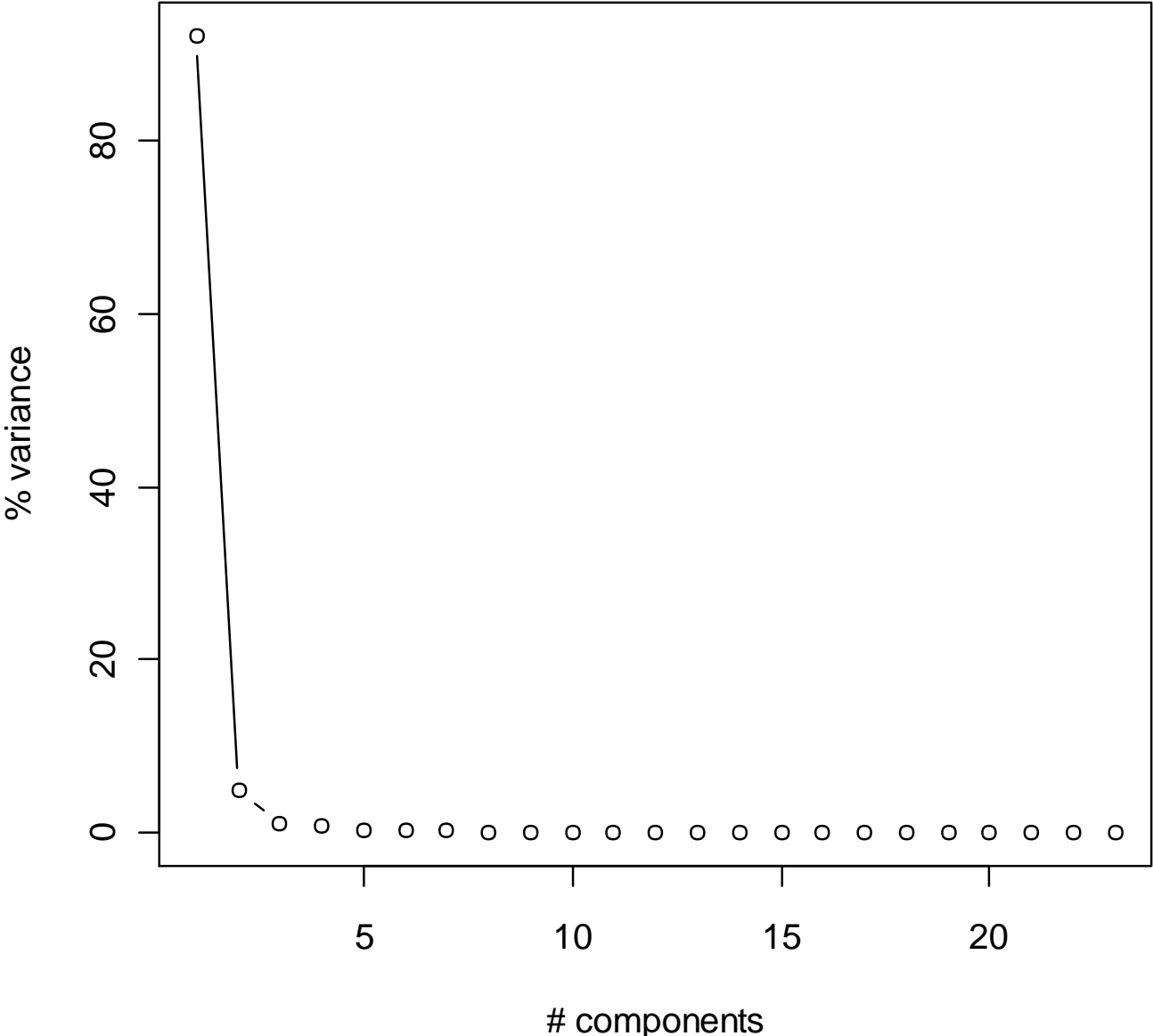


# Data Visualization

- PCA is used for dimensionality reduction.
- In the next slide, the first 2 components from the PCA analysis are plotted in 2-dimensional space.
- The non-smoker samples (red N's) are clearly clustered away from the smoker samples (blue S's).
- Although the smoker samples do not form as clear a cluster as the non-smoker samples, they are easily distinguishable from the non-smoker samples.
- Following the PCA plot is a Scree plot of all the components.
- It shows the first 2 components account for around 95% of the variance in the filtered dataset.



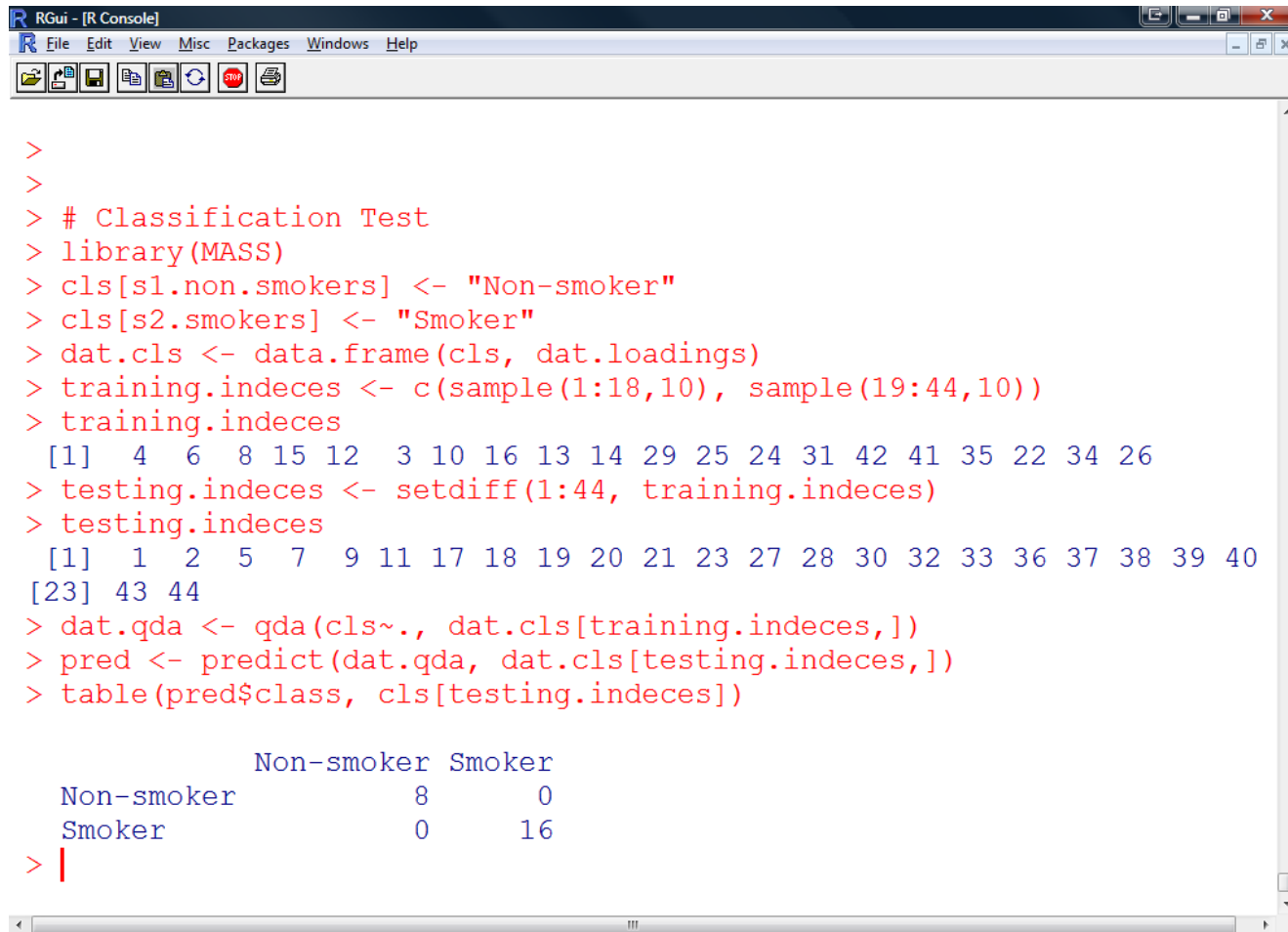
# Scree plot



# Data Classification

- Using quadratic discriminant analysis (QDA), the previously identified PCA components are used to classify the samples into their respective classes.
- The PCA plot a few slides back seems to indicate the existence of a linear discriminant.
- However, LDA produces occasional errors in classification even with as few as 10 random runs.
- Based on 10 random runs, QDA classifies without errors.
- Out of the 44 samples, 10 non-smokers and 10 smokers are chosen at random for the training set.
- The testing set consists of the remaining 24 samples.

# One Sample Run

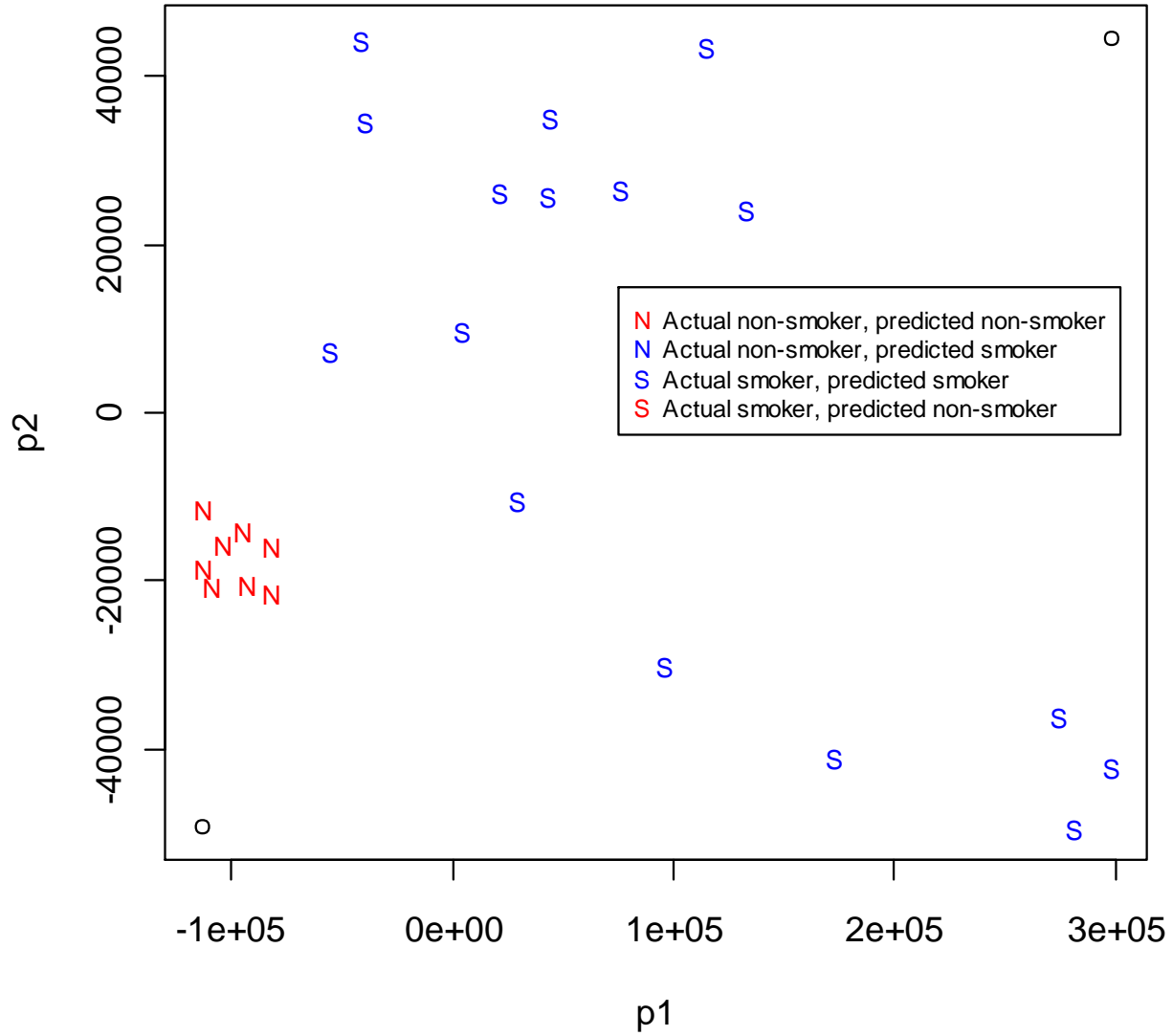


```
RGui - [R Console]
File Edit View Misc Packages Windows Help

>
>
> # Classification Test
> library(MASS)
> cls[s1.non.smokers] <- "Non-smoker"
> cls[s2.smokers] <- "Smoker"
> dat.cls <- data.frame(cls, dat.loadings)
> training.indeces <- c(sample(1:18,10), sample(19:44,10))
> training.indeces
[1] 4 6 8 15 12 3 10 16 13 14 29 25 24 31 42 41 35 22 34 26
> testing.indeces <- setdiff(1:44, training.indeces)
> testing.indeces
[1] 1 2 5 7 9 11 17 18 19 20 21 23 27 28 30 32 33 36 37 38 39 40
[23] 43 44
> dat.qda <- qda(cls~., dat.cls[training.indeces,])
> pred <- predict(dat.qda, dat.cls[testing.indeces,])
> table(pred$class, cls[testing.indeces])

      Non-smoker Smoker
Non-smoker      8      0
Smoker          0     16
> |
```

# PCA plot of Filtered GSE5056 Data



# Gene Identification

- Fold change is calculated for the 23 selected genes, and the 10 most discriminant genes are chosen based on this calculated value.
- These 10 genes consist of the 5 genes with the greatest positive fold change and the 5 genes with the greatest negative fold change.
- The 10 genes are listed below:  
M74542\_at, X68314\_at, J03934\_s\_at, X76342\_at, U05861\_at,  
U17077\_at, L42563\_at, M69203\_s\_at, K02765\_at, U22028\_at.
- The gene information is included in the file  
final\_genes\_raymond\_li.txt.

# Gene Breakdown

- Alcohol detoxification/response (2 genes): M74542\_at  
X76342\_at
- Inflammatory response (3 genes): X68314\_at  
M69203\_s\_at  
K02765\_at
- Oxidative stress response (2 genes): X68314\_at  
J03934\_s\_at
- Cholesterol homeostasis (2 genes): U05861\_at  
U17077\_at  
U22028\_at
- Bile acid binding (1 gene): U05861\_at
- Organophosphorus response (1 gene): U05861\_at
- Immune response (1 gene): M69203\_s\_at
- Toxin response (1 gene): J03934\_s\_at

# Cigarette Smoking

- Ethanol and a number of other alcohols are found in cigarettes. Inflammatory response and oxidative stress are very highly correlated with cigarette smoking.
- HDL (good cholesterol) is lowered by cigarette smoking.
- Bile salt concentration is increased by cigarette smoking.
- Organophosphorus is found in pesticide which is used on the tobacco found in cigarettes.
- Immune response is suppressed by smoking cigarettes.
- Toxin response is triggered by smoking cigarettes.

# Conclusions

- The process of removing outliers, filtering out low expression genes, and feature selection seem to zero in on those genes that are highly correlated with smokers (rather than non-smokers).
- The classification of a subset of the data based on dimensionality reduction and classification methods also proved effective.
- Despite a seemingly successful pipeline, a key issue is the use of the same dataset to both train and test the data.
- Using a distinct dataset would be a more robust test of the classification model based on the selected components.
- Additionally, the lack of samples (only 44 in total) is a clear weakness.
- For future experimentation, another feature selection method (t-test, Wilcoxon, Welch's, etc.) may yield different (maybe better) results.

# References

1. [http://whyquit.com/whyquit/A Tobacco Additives.html](http://whyquit.com/whyquit/A_Tobacco_Additives.html)
2. <http://respiratory-research.com/content/7/1/132>
3. <http://www.ncbi.nlm.nih.gov/pubmed/209795>
4. <http://www.ncbi.nlm.nih.gov/pubmed/3956939>
5. <http://tih.sagepub.com/cgi/content/abstract/22/9/399>
6. <http://jpet.aspetjournals.org/content/293/1/166.abstract>
7. [http://www.med.upenn.edu/ceet/documents user/BlairNCI.doc](http://www.med.upenn.edu/ceet/documents_user/BlairNCI.doc)
8. <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5056>