Homework Assignment 1

Group 3: *Ezgi Karaesmen, Mary Spencer, Jiefei Wang, Shruti Dhiren Shah*

installing necessary libraries

```r
#install.packages('NMF')
library(knitr)
library(GEOquery)
library(NMF)
library(gplots)
library(vioplot)
library(matrixStats)
```

Problem 1

The data was downloaded and the designated row was extracted. For Group 3, our assigned row of the gse19439 expression data matrix was 26058. The following code was used to extract this row of information for further analysis.

```r
#clean the workspace
rm(list=ls())

#set the working directory,please set it to the data file path
setwd('D:/course material/STA525 Statistics of bioinformation/homework/hw1')

#loading previously downloaded data for quick access
load(file="HW1 data.RData")
#extracting assigned row from expression data matrix
group_data=X[26058,]

#determining if our feature corresponds to a gene
probe_name=rownames(X)[26058]
probe_data_feature=pData(featureData(gse19439))
index=match(probe_name,probe_data_feature[,"ID"])
gene_name=probe_data_feature[index,]
gene=gene_name$Symbol

probe_name
gene
```

It was determined that row 26059, probe ILMN_1798827, corresponds to the SRBD1 gene. Little is known about this particular gene other than it is a protein coding gene that assists in RNA binding, but some common aliases include *S1 RNA Binding Domain 1* and *S1 RNA-Binding Domain-Containing Protein 1*. SRBD1 is located on the p arm of chromosome 2 (2p21) and contains 995 amino acids with a final molecular mass of 111,776 Da. Additionally, alternative splicing produces 2 isoforms of this gene. There is evidence to suggest that SRBD1 is linked to disease suceptibility, especially glaucoma.

1. http://www.genecards.org/cgi-bin/carddisp.pl?gene=SRBD1
2. http://www.uniprot.org/uniprot/Q8N5C6#Q8N5C6-1
3. http://www.ncbi.nlm.nih.gov/pubmed/20363506
4. http://www.ncbi.nlm.nih.gov/pubmed/24040232

Problem 2

Our assigned data for row 2058 was plotted as a function of the factors CON, LTB, and PTB. These factor assignments correspond to the control group, latent TB group, and active TB group, respectively.

We first compared the factors with a box plot.

```
probe_data_pheno=pData(phenoData(gse19439))[,1]
probe_data_type=as.factor(substring(probe_data_pheno,1,3))
boxplot(group_data~probe_data_type)
```

While the control and latent stage groups have comparable results, subjects with active TB appear to have a substantially increased expression of this gene.

The second method of comparison was to make both a violin and bean plot. While comparable plots, the bean plot allows the possible multimodal nature of the data to be visible.

```
vioplot(group_data[as.numeric(probe_data_type)==1],
        group_data[as.numeric(probe_data_type)==2],
        group_data[as.numeric(probe_data_type)==3],names=c("CON", "LTB", "PTB"),col="tomato1")
beanplot::beanplot(group_data[as.numeric(probe_data_type)==1],
                  group_data[as.numeric(probe_data_type)==2],
                   group_data[as.numeric(probe_data_type)==3],names=c("CON", "LTB", "PTB"), col=3)
```

The bean plot illustrates the the violin plot oversimplified the kernel density for each factor. Espcially for the active TB group, there were multiple nodes hidden with the violin plot.

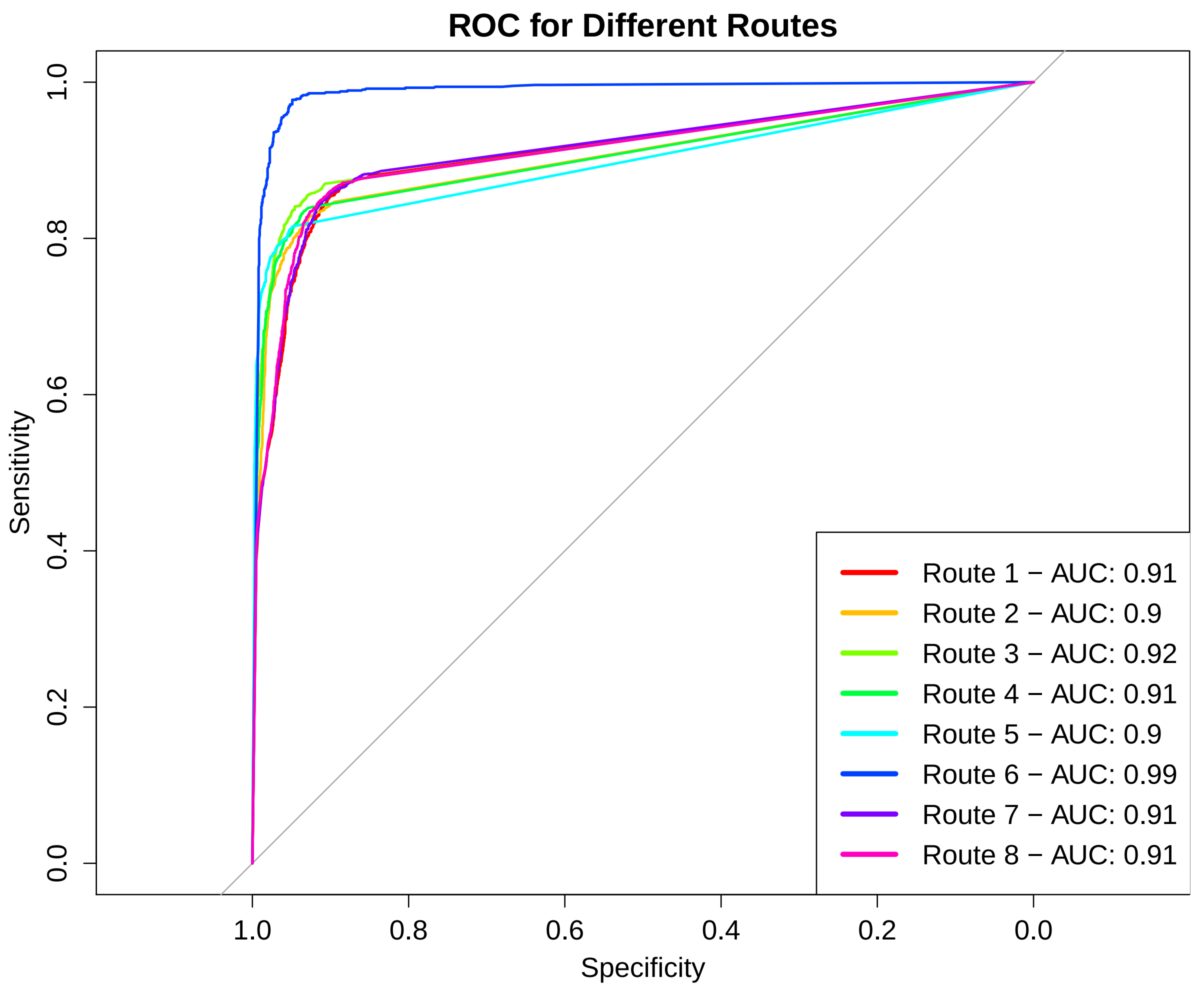
Problem 3

```
#Find the best probes that have the smallest P-value
best20=order(myPvals)[1:20]
best20.X= X[best20,]
colnames(best20.X)=probe_data_type

heatmap.2(best20.X, trace = "none")
heatmap.2(best20.X, trace = "none", scale = "row")


Info = pData(phenoData(gse19439))[,1:2]
Info[,1] <- probe_data_type
group = Info[1]
heatmap2 = aheatmap(X[best20,], annCol=group)
```

All the figures suggest that the features are able to distinguish the PTB patients the best, though the separation is not perfect. The CON and LTB samples are less distinguishable.

Problem 4

For Problem 4, a control microarray dataset as mentioned in **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset** paper by Halfon et al


ROC for Different Routes
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Route 1 − AUC: 0.91
Route 2 − AUC: 0.9
Route 3 − AUC: 0.92
Route 4 − AUC: 0.91
Route 5 − AUC: 0.9
Route 6 − AUC: 0.99
Route 7 − AUC: 0.91
Route 8 − AUC: 0.91