Gene Expression Data Analysis and Visualization in RStudio, Libraries:

```
library(edgeR)
biocLite (GEOquery)
library(GEOquery)
library(Biobase)
library(plyr)
library(gplots)
library(limma)
library(impute)
library(fpc)
library(gplots)
library(Biobase)
library(annotate)
library(multtest)
library(MASS)
library(lda)
library(EMV)
Library(grepl)
library(gdata)
library(matrixStats)
library(plyr) library(class)
library(kernlab)

GSE_data <-
read.delim("~/Documents/FinalHW/GSE174443_gene_counts_matrix_with_Blanks.txt",
row.names=1,header=T)
dim(GSE_data)
[1] 27744    24

#Reorder the Columns in the File

GSEreorder <- GSE_data[, order(names(GSE_data))]
dim(GSEreorder)
[1] 27744    24

colnames(GSEreorder)
[1] "Control_011"     "Control_013"     "Control_018"     "Control_022"     "Control_024"
[6] "Control_027"     "Control_031"     "Sarcoidosis_001" "Sarcoidosis_005" "Sarcoidosis_007"
[11] "Sarcoidosis_014" "Sarcoidosis_017" "Sarcoidosis_020" "Sarcoidosis_023" "Sarcoidosis_026"
[16] "Sarcoidosis_029" "Sarcoidosis_032" "TB_006"          "TB_008"          "TB_012"
[21] "TB_015"          "TB_021"          "TB_025"          "TB_030"

View(GSEreorder)
```

Filter

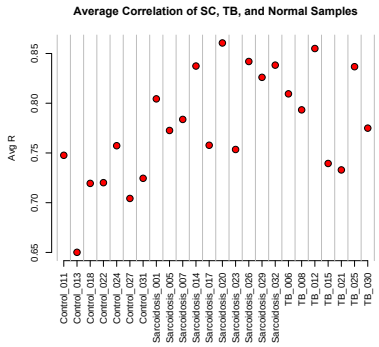
	Control_011	Control_013	Control_018	Control_022
A1BG	65.550236	32.1730124	100.7923407	48.8314546
A1CF	2.621931	11.2054612	2.6515764	1.7414190
A2M	3589.100443	852.3659730	5701.4038200	1481.4245740
A2ML1	25.419226	45.0940567	75.1124064	44.2517309
A2MP1	15.219934	1.1367500	NA	18.7487699
A3GALT2	NA	NA	NA	NA

g 1 to 6 of 27,744 entries, 24 total columns

1. First, we test for **outlier samples** and provide visual proof.

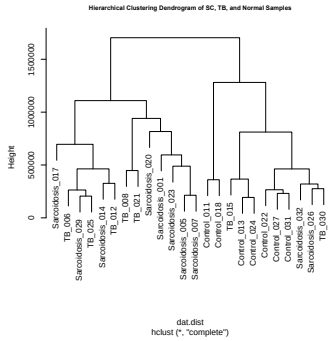
```
#Average Correlation Plot

dat.cor <- cor(GSEreorder, use="pairwise.complete.obs")
dat.cor
dat.avg <- apply(dat.cor,1,mean)
```

```
par(oma=c(6,0.1,0.1,0.1))
plot(c(1,length(dat.avg)),range(dat.avg),type="n",xlab="",ylab="Avg
R",main="Average Correlation of SC, TB, and Normal Samples",axes=F)
points(dat.avg,bg="red",col=1,pch=21,cex=1.5)
axis(1,at=c(1:length(dat.avg)),labels=dimnames(GSEreorder)[[2]],las=2,cex=0.5)
axis(2)
abline(v=seq(0.5,25,1),col="grey")
```



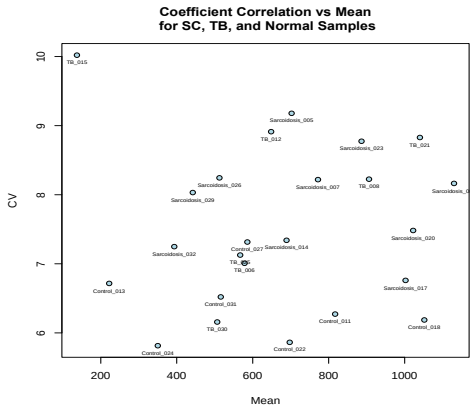
#Hierarchical Clustering Dendrogram

```
dat.t <- t(GSEreorder)
dat.dist <- dist(dat.t, method="euclidean")
dat.clust <- hclust(dat.dist,method="single")
plot(hclust(dat.dist), labels=dimnames(GSEreorder)[[2]],main="Hierarchical
Clustering Dendrogram of SC, TB, and Normal Samples",cex.main=0.75)
```



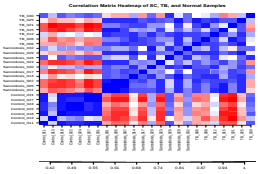
#CV vs. Mean Plot

```
dat.mean <- apply(GSEreorder,2,mean, na.rm=TRUE)
dat.sd <- apply(GSEreorder,2,sd, na.rm=TRUE)
dat.cv <- dat.sd/dat.mean

plot(dat.mean,dat.cv,main="Coefficient Correlation vs Mean\n for SC, TB, and
Normal Samples",xlab="Mean",ylab="CV",col="blue",cex=1.5,type="n")
points(dat.mean,dat.cv,bg="lightblue",col=1,pch=21)
text(dat.mean,dat.cv,label=dimnames(GSEreorder)[[2]],pos=1,cex=0.5)
```



#Correlation Plot (Heatmap)

```
dat.cor <- cor(GSEreorder, use="pairwise.complete.obs")
dat.cor

layout(matrix(c(1,1,1,1,1,1,1,1,2,2), 5, 2, byrow = TRUE))
par(oma=c(2,2,1,1))
cx<- rev(colorpanel(25,"blue", "white", "red"))
leg <- seq(min(dat.cor,na.rm=T),max(dat.cor,na.rm=T),length=10)

image(dat.cor,main="Correlation Matrix Heatmap of SC, TB, and Normal Samples",
axes=F, col=cx)
axis(1,at=seq(0,1,length=ncol(dat.cor)),label=dimnames(dat.cor)[[2]],
cex.axis=0.9, las=2)
axis(2,at=seq(0,1,length=ncol(dat.cor)),label=dimnames(dat.cor)[[2]],
cex.axis=0.9, las=2)

image(as.matrix(leg),col=cx, axes=F)
tmp <- round(leg,2)
axis(1,at=seq(0,1,length=length(leg)),labels=tmp,cex.axis=1)
```



see presentation

Remove these outliers.

```
dim(GSEreorder)
[1] 27744    24

colnames(GSEreorder)
[1] "Control_011"     "Control_013"     "Control_018"     "Control_022"     "Control_024"
[6] "Control_027"     "Control_031"     "Sarcoidosis_001" "Sarcoidosis_005" "Sarcoidosis_007"
[11] "Sarcoidosis_014" "Sarcoidosis_017" "Sarcoidosis_020" "Sarcoidosis_023" "Sarcoidosis_026"
[16] "Sarcoidosis_029" "Sarcoidosis_032" "TB_006"          "TB_008"          "TB_012"
[21] "TB_015"          "TB_021"          "TB_025"          "TB_030"
```

Sample Control_013 is the outlier sample.

```
#Remove Outlier Control_13

datout <- GSEreorder[, !(colnames(GSEreorder) %in% c("Control_013"))]
dim(datout)
[1] 27744    23

colnames(datout)
[1] "Control_011"     "Control_018"     "Control_022"     "Control_024"     "Control_027"
[6] "Control_031"     "Sarcoidosis_001" "Sarcoidosis_005" "Sarcoidosis_007" "Sarcoidosis_014"
[11] "Sarcoidosis_017" "Sarcoidosis_020" "Sarcoidosis_023" "Sarcoidosis_026" "Sarcoidosis_029"
[16] "Sarcoidosis_032" "TB_006"          "TB_008"          "TB_012"          "TB_015"
[21] "TB_021"          "TB_025"          "TB_030"
```

Log2 transform the data and filter out transcripts that have **low expression** values or missing.

```
#Log2 Transform the Data
datout.log <- log2(datout)
dim(datout.log)
[1] 27744    23

#Remove Missing Values (NAs, Blanks, 0) from the log2 transformed data
datout.log.No.NA <- na.omit(datout.log)
dim(datout.log.No.NA)
[1] 10331    23

Output: Dataset is now composed of 23 samples and 10331 transcripts.
```

Method of feature selection with a statistical test. For two conditions comparisons, a two-sample test is used, for three group comparison conditions - ANOVA tests is used.

```
#ANOVA (Analysis of Variance, p-value reported)

aov.all.genes <- function(x,s1,s2,s3) {
 x1 <- as.numeric(x[s1])
 x2 <- as.numeric(x[s2])
      x3 <- as.numeric(x[s3])
fac <- c(rep("A",length(x1)), rep("B",length(x2)), rep("C",length(x3)))
a.dat <- data.frame(as.factor(fac),c(x1,x2,x3))
names(a.dat) <- c("factor","express")
 p.out <- summary(aov(express~factor, a.dat))[[1]][1,5]
 return(p.out)    }
aov.run <- apply(datout.log.No.NA,1,
 aov.all.genes,s1=c(1:6),s2=c(7:16),s3=c(17:23))
aov.run

#Number of Statistically Significant Transcripts at different p-value
 thresholds
sum(aov.run<.05)
[1] 1846
sum(aov.run<.01)
[1] 683
sum(aov.run<(.05/10331))
[1] 20

#Obtain the List of Transcripts with P-value < 0.05
pv <-as.data.frame(aov.run)
pvA <- subset(pv, pv<0.05)
pvA
```

```
dim(pvA)
[1] 1846

#ANOVA (F-statistic reported)
#this is an Extra Code
aov.all.genes <- function(x,s1,s2,s3) {
      x1 <- as.numeric(x[s1])
      x2 <- as.numeric(x[s2])
      x3 <- as.numeric(x[s3])
fac <- c(rep("A",length(x1)), rep("B",length(x2)), rep("C",length(x3)))
a.dat <- data.frame(as.factor(fac),c(x1,x2,x3))
names(a.dat) <- c("factor","express")
      p.outF <- summary(aov(express~factor, a.dat))[[1]][1,4]
      return(p.outF)
}
aov.runF <- apply(datout.log.No.NA,1,
aov.all.genes,s1=c(1:6),s2=c(7:16),s3=c(17:23))
aov.runF
```

Adjust for multiplicity.

```
#Multiple testing correction procedures: Holm and Bonferroni.

#Holm Multiple Testing Correction Method.
aov.run.holm <- p.adjust(aov.run, method="holm")
aov.run.holm
sum(aov.run.holm<.05)
[1] 20

pvh <-as.data.frame(aov.run.holm)
pvH <- subset(pvh, pvh<0.05)
pvH
          aov.run.holm
ALAS1     2.700519e-02
C15orf48  1.361710e-05
C7        8.518631e-03
CCL21     6.379488e-04
CHI3L1    5.786055e-06
CLEC4E    3.988156e-02
CXCL9     1.340064e-02
CYP27B1   4.673795e-03
DNAJC5B   1.330780e-02
DSP       1.337230e-03
FCGR2A    6.095653e-03
FTH1      1.301722e-02
FTH1P11   1.387408e-02
FTH1P2    3.384643e-03
FTH1P23   4.031312e-03
GBP1      7.403439e-03
MYOF      5.184531e-03
SNX10     1.877543e-03
TFEC      2.592348e-03
TLR8      2.877153e-03

#Subset Data With statistically significant Transcripts from the Holm Test.
```

```
datout.log.No.NA.sub.Holm <- datout.log.No.NA[rownames(pvH),]
dim(datout.log.No.NA.sub.Holm)
[1] 20 23

datout.log.No.NA.sub.Holm
summary(datout.log.No.NA.sub.Holm)
```

#Bonferroni Multiple Testing Correction Method.

```
aov.run.bonferroni <- p.adjust(aov.run, method="bonferroni")
aov.run.bonferroni
sum(aov.run.bonferroni<.05)
[1] 20

pvb <-as.data.frame(aov.run.bonferroni)
pvB <- subset(pvb, pvb<0.05)
pvB
         aov.run.bonferroni
ALAS1          2.705232e-02
C15orf48       1.361842e-05
C7             8.529364e-03
CCL21          6.380724e-04
CHI3L1         5.786055e-06
CLEC4E         3.995504e-02
CXCL9          1.342143e-02
CYP27B1        4.677870e-03
DNAJC5B        1.332715e-02
DSP            1.337618e-03
FCGR2A         6.102150e-03
FTH1           1.303489e-02
FTH1P11        1.389695e-02
FTH1P2         3.386938e-03
FTH1P23        4.034436e-03
GBP1           7.412049e-03
MYOF           5.189554e-03
SNX10          1.878271e-03
TFEC           2.593603e-03
TLR8           2.878825e-03
```

#Subset Data with statistically significant Transcripts from the Bonferroni Test

```
datout.log.No.NA.sub.Bonferroni <- datout.log.No.NA[rownames(pvB),]
dim(datout.log.No.NA.sub.Bonferroni)
[1] 20 23

#See the final table in the power point presentation.
```

Provide the number of genes retained with the associated score (p-value, weight, test statistic, etc.) and threshold value that is used.

1846 transcripts are identified to be significantly differently expressed with p-value < 0.05 threshold; and **683** - with p-value threshold < 0.01.

When Holm's multiple testing correction method was applied, **20** genes were found to be statistically significant at p < 0.5 threshold.

When Bonferroni's multiple testing correction method was applied, **20** genes were found to be statistically significant at p < 0.5 threshold.

#Calculate Mean Gene Expression of Transcripts for 3 Groups:

#Subset Data with statistically significant genes from Holm Test.

```
datout.log.No.NA.sub.Holm <- datout.log.No.NA[rownames(pvH),]
dim(datout.log.No.NA.sub.Holm)
[1] 20 23
```

```
View(datout.log.No.NA.sub.Holm)
```

	Control_011	Control_018	Control_022	Control_024	Control_027
ALAS1	9.488019	9.625796	8.23670240	8.492510	8.957351
C15orf48	6.206223	5.981261	6.08358463	6.098518	5.320281
C7	9.928106	10.363578	8.45453702	8.453027	10.032477
CCL21	14.030219	15.640148	14.80891844	13.605963	14.160617

```
datout.log.No.NA.sub.HolmC <- datout.log.No.NA[rownames(pvH),c(1:6)]

datout.log.No.NA.sub.HolmC.t <-t(datout.log.No.NA.sub.HolmC)
summary(datout.log.No.NA.sub.HolmC.t)
```

```
      ALAS1          C15orf48          C7              CCL21          CHI3L1     
 Min.   :8.237   Min.   :5.320   Min.   : 8.453   Min.   :13.61   Min.   :5.763  
 1st Qu.:8.495   1st Qu.:6.007   1st Qu.: 8.718   1st Qu.:13.88   1st Qu.:7.062  
 Median :8.730   Median :6.091   Median : 9.718   Median :14.10   Median :7.740  
 Mean   :8.884   Mean   :5.980   Mean   : 9.457   Mean   :14.35   Mean   :7.550  
 3rd Qu.:9.355   3rd Qu.:6.166   3rd Qu.:10.006   3rd Qu.:14.65   3rd Qu.:8.000  
 Max.   :9.626   Max.   :6.206   Max.   :10.364   Max.   :15.64   Max.   :9.133  
      CLEC4E           CXCL9           CYP27B1            DNAJC5B           DSP        
 Min.   :2.976   Min.   : 8.310   Min.   :-0.02046   Min.   :3.722   Min.   :6.733  
 1st Qu.:5.156   1st Qu.: 8.675   1st Qu.: 4.09254   1st Qu.:3.880   1st Qu.:7.757  
 Median :6.799   Median : 9.438   Median : 5.39302   Median :4.548   Median :8.068  
 Mean   :6.493   Mean   : 9.282   Mean   : 4.31897   Mean   :4.847   Mean   :8.034  
 3rd Qu.:7.722   3rd Qu.: 9.505   3rd Qu.: 5.51309   3rd Qu.:5.331   3rd Qu.:8.310  
 Max.   :9.750   Max.   :10.558   Max.   : 5.93610   Max.   :7.016   Max.   :9.291  
      FCGR2A           FTH1           FTH1P11          FTH1P2          FTH1P23     
 Min.   :5.501   Min.   :15.01   Min.   :4.942   Min.   :5.170   Min.   :3.641  
 1st Qu.:6.060   1st Qu.:15.19   1st Qu.:5.700   1st Qu.:5.310   1st Qu.:4.010  
 Median :7.014   Median :15.50   Median :6.099   Median :5.736   Median :4.675  
 Mean   :6.824   Mean   :15.57   Mean   :6.061   Mean   :5.726   Mean   :4.503  
 3rd Qu.:7.635   3rd Qu.:15.71   3rd Qu.:6.624   3rd Qu.:5.878   3rd Qu.:4.939  
 Max.   :7.829   Max.   :16.52   Max.   :6.860   Max.   :6.621   Max.   :5.211  
      GBP1            MYOF           SNX10           TFEC           TLR8       
 Min.   : 9.288   Min.   :7.732   Min.   :4.145   Min.   :6.379   Min.   :5.637  
 1st Qu.: 9.314   1st Qu.:8.053   1st Qu.:4.370   1st Qu.:6.645   1st Qu.:5.830  
 Median : 9.363   Median :8.364   Median :4.416   Median :7.084   Median :6.194  
 Mean   : 9.522   Mean   :8.346   Mean   :4.495   Mean   :7.248   Mean   :6.124  
 3rd Qu.: 9.404   3rd Qu.:8.493   3rd Qu.:4.531   3rd Qu.:7.616   3rd Qu.:6.330  
 Max.   :10.405   Max.   :9.128   Max.   :5.068   Max.   :8.648   Max.   :6.637  
```

```
datout.log.No.NA.sub.HolmS <- datout.log.No.NA[rownames(pvH),c(7:16)]

datout.log.No.NA.sub.HolmS.t <-t(datout.log.No.NA.sub.HolmS)
summary(datout.log.No.NA.sub.HolmS.t)
```

```
      ALAS1          C15orf48          C7              CCL21          CHI3L1     
```

```
 Min.   :10.14   Min.   :10.39   Min.   : 6.877   Min.   :11.01   Min.   :12.57
 1st Qu.:11.14   1st Qu.:10.56   1st Qu.: 7.950   1st Qu.:12.29   1st Qu.:14.06
 Median :11.58   Median :11.97   Median : 8.297   Median :12.71   Median :14.38
 Mean   :11.50   Mean   :11.63   Mean   : 8.425   Mean   :12.55   Mean   :14.18
 3rd Qu.:12.05   3rd Qu.:12.42   3rd Qu.: 8.916   3rd Qu.:13.11   3rd Qu.:14.55
 Max.   :12.43   Max.   :12.75   Max.   :10.192   Max.   :13.80   Max.   :14.83
       CLEC4E          CXCL9          CYP27B1          DNAJC5B           DSP
 Min.   :10.22   Min.   : 9.963   Min.   : 7.888   Min.   : 7.573   Min.   :4.740
 1st Qu.:10.76   1st Qu.:12.439   1st Qu.: 9.163   1st Qu.: 8.860   1st Qu.:5.243
 Median :11.16   Median :12.511   Median :10.416   Median : 9.843   Median :5.834
 Mean   :11.19   Mean   :12.540   Mean   :10.188   Mean   : 9.501   Mean   :5.798
 3rd Qu.:11.64   3rd Qu.:13.078   3rd Qu.:11.290   3rd Qu.:10.207   3rd Qu.:6.470
 Max.   :12.46   Max.   :13.825   Max.   :11.584   Max.   :10.265   Max.   :6.668
       FCGR2A           FTH1          FTH1P11          FTH1P2         FTH1P23
 Min.   : 8.859   Min.   :17.38   Min.   : 8.005   Min.   :7.911   Min.   :6.758
 1st Qu.:10.341   1st Qu.:18.21   1st Qu.: 8.754   1st Qu.:8.279   1st Qu.:7.287
 Median :10.685   Median :18.95   Median : 9.142   Median :8.524   Median :7.920
 Mean   :10.437   Mean   :18.77   Mean   : 9.192   Mean   :8.650   Mean   :7.843
 3rd Qu.:10.975   3rd Qu.:19.38   3rd Qu.: 9.617   3rd Qu.:9.102   3rd Qu.:8.428
 Max.   :11.413   Max.   :19.86   Max.   :10.643   Max.   :9.343   Max.   :8.854
         GBP1            MYOF           SNX10           TFEC            TLR8
 Min.   :10.43   Min.   :10.11   Min.   :6.481   Min.   : 9.027   Min.   : 8.483
 1st Qu.:11.52   1st Qu.:10.95   1st Qu.:7.181   1st Qu.: 9.554   1st Qu.: 9.239
 Median :11.83   Median :11.66   Median :7.875   Median :10.112   Median : 9.617
 Mean   :11.91   Mean   :11.67   Mean   :7.685   Mean   : 9.960   Mean   : 9.649
 3rd Qu.:12.34   3rd Qu.:12.48   3rd Qu.:8.338   3rd Qu.:10.367   3rd Qu.: 9.879
 Max.   :12.98   Max.   :13.17   Max.   :8.449   Max.   :10.582   Max.   :11.508
```

```
datout.log.No.NA.sub.HolmT <- datout.log.No.NA[rownames(pvH),c(17:23)]

datout.log.No.NA.sub.HolmT.t <-t(datout.log.No.NA.sub.HolmT)
summary(datout.log.No.NA.sub.HolmT.t)
```

```
       ALAS1          C15orf48            C7           CCL21          CHI3L1
 Min.   : 9.351   Min.   : 9.542   Min.   :3.250   Min.   : 7.554   Min.   :10.97
 1st Qu.:10.792   1st Qu.:11.279   1st Qu.:4.258   1st Qu.: 8.519   1st Qu.:12.34
 Median :11.037   Median :12.294   Median :5.946   Median : 9.427   Median :13.24
 Mean   :10.945   Mean   :11.990   Mean   :5.273   Mean   : 9.278   Mean   :13.06
 3rd Qu.:11.408   3rd Qu.:12.760   3rd Qu.:6.282   3rd Qu.: 9.620   3rd Qu.:13.56
 Max.   :11.825   Max.   :14.015   Max.   :6.637   Max.   :11.690   Max.   :15.40
       CLEC4E          CXCL9          CYP27B1          DNAJC5B           DSP
 Min.   : 9.28   Min.   :11.77   Min.   : 8.495   Min.   :6.119   Min.   :2.700
 1st Qu.:10.72   1st Qu.:12.83   1st Qu.:10.000   1st Qu.:7.078   1st Qu.:3.807
 Median :11.06   Median :13.64   Median :10.849   Median :8.257   Median :4.228
 Mean   :10.94   Mean   :13.46   Mean   :10.494   Mean   :8.176   Mean   :4.137
 3rd Qu.:11.33   3rd Qu.:13.88   3rd Qu.:11.031   3rd Qu.:9.487   3rd Qu.:4.397
 Max.   :12.17   Max.   :15.36   Max.   :12.055   Max.   :9.723   Max.   :5.622
       FCGR2A           FTH1          FTH1P11          FTH1P2         FTH1P23
 Min.   : 8.146   Min.   :16.54   Min.   : 7.054   Min.   :6.653   Min.   :5.769
 1st Qu.:10.154   1st Qu.:17.73   1st Qu.: 8.456   1st Qu.:7.560   1st Qu.:6.701
 Median :10.297   Median :18.41   Median : 8.644   Median :8.588   Median :7.590
 Mean   :10.214   Mean   :18.21   Mean   : 8.635   Mean   :8.212   Mean   :7.223
 3rd Qu.:10.789   3rd Qu.:18.89   3rd Qu.: 8.986   3rd Qu.:8.909   3rd Qu.:7.762
 Max.   :11.165   Max.   :19.32   Max.   : 9.862   Max.   :9.306   Max.   :8.276
         GBP1            MYOF           SNX10           TFEC            TLR8
 Min.   :11.26   Min.   :10.52   Min.   :6.106   Min.   : 8.942   Min.   : 6.998
 1st Qu.:12.37   1st Qu.:11.23   1st Qu.:7.324   1st Qu.: 9.455   1st Qu.: 8.688
 Median :13.35   Median :11.72   Median :7.619   Median :10.055   Median : 9.080
 Mean   :13.06   Mean   :11.74   Mean   :7.778   Mean   : 9.924   Mean   : 8.934
 3rd Qu.:14.00   3rd Qu.:12.29   3rd Qu.:8.383   3rd Qu.:10.312   3rd Qu.: 9.530
 Max.   :14.03   Max.   :12.91   Max.   :9.306   Max.   :10.937   Max.   :10.021
```

Plot the scores of those genes retained in a histogram.

```
#Histogram for Retained Genes (Holm Test)
pvh <-as.data.frame(aov.run.holm)
pvH <- subset(pvh, pvh<0.05)
pvH
```

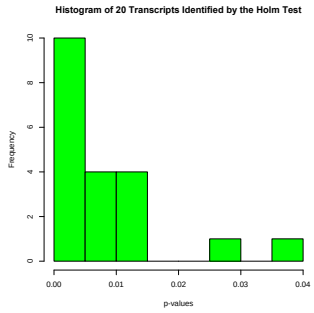
#Remove Row Names from the Subset of Genes Identified by Holm's Test

```
dat.f.sub.H <- pvH[rownames(pvH),]
```

dat.f.sub.H double [20] 2.70e-02 1.36e-05 8.52e-03 6.38e-04 5.79e-06 3.99e

```
[1] 2.700519e-02 1.361710e-05 8.518631e-03 6.379488e-04 5.786055e-06 3.988156e-02 1.340064e-02
[8] 4.673795e-03 1.330780e-02 1.337230e-03 6.095653e-03 1.301722e-02 1.387408e-02 3.384643e-03
[15] 4.031312e-03 7.403439e-03 5.184531e-03 1.877543e-03 2.592348e-03 2.877153e-03
```

```
par(oma=c(1,1,1,1))
hist(dat.f.sub.H, main="Histogram of 20 Transcripts Identified by the Holm Test",xlab="p-values",col="green")
```

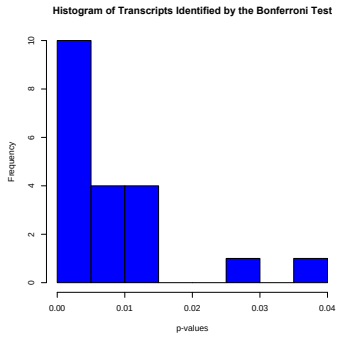


#Histogram for Retained Genes (from Bonferroni Test)

```
pvb <-as.data.frame(aov.run.bonferroni)
pvB <- subset(pvb, pvb<0.05)
pvB
```

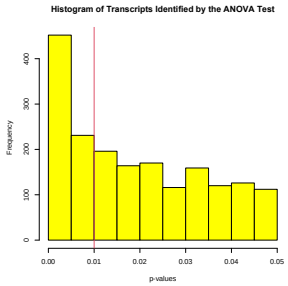
#Remove Row Names from the Subset of Genes Identified by Bonferroni's Test

```
dat.f.sub.B <- pvB[rownames(pvB),]

par(oma=c(1,1,1,1))
hist(dat.f.sub.B, main="Histogram of Transcripts Identified by the Bonferroni Test",xlab="p-values",col="blue")
```



#Histogram for ANOVA p-values for All Expressed Genes (for comparison)

```
dat.f.sub.A <- pvA[rownames(pvA),]

hist(dat.f.sub.A, main="Histogram of Transcripts Identified by the ANOVA Test",xlab="p-values",col="yellow")
abline(v=.01, col=2, lwd=2)
```



Subset data by the genes identified above. Perform **clustering** and **dimensionality** reduction methods to visualize the samples in two-dimensional space (xy scatter plot, dendrogram).

#Obtain the List of Transcripts with P-value < 0.05 from ANOVA on Log2-Transformed Data

```
pv<-as.data.frame(aov.run)
pvA <- subset(pv, pv<0.05)
pvA
```

#Number of Statistically Significant Transcripts

```
sum(aov.run<.05)
[1] 1846
```

#Subset Data with Genes That Have P-value < 0.05 from ANOVA, 23 samples.

```
datout.log.No.NA.subA <- datout.log.No.NA[rownames(pvA),].
dim(datout.log.No.NA.subA)
[1] 1846   23
```

```
View(datout.log.No.NA.subA)
```

	Control_011	Control_018	Control_022	Control_024	Control_027	Control_031	Sarcoidosis_001
A2M	11.809407	12.477101	10.532769	10.010271	10.831902	10.799529	13.457847
A2ML1	4.667848	6.230979	5.467662	5.152338	4.297682	4.778384	3.751900
AAK1	8.004322	8.488579	7.396583	8.267713	6.999145	7.433171	8.295180
ABAT	6.870629	7.502427	7.919306	6.249471	7.365509	4.901415	6.790291
ABCA8	7.692852	9.330889	8.275876	7.321939	6.809839	6.620989	5.300421
ABCB11	8.438146	6.147477	8.582570	6.700805	8.300537	7.533668	7.978083

#Calculate kmeans PCA on the samples and retain the first two component vectors, k=3.

```
dat.pca <- prcomp(t(datout.log.No.NA.subA))
dat.loadingsA <- dat.pca$x[,1:2]
```

```
View(dat.loadingsA)
```

	PC1	PC2
Control_011	-45.324519	24.2088463
Control_018	-62.797067	23.7031687
Control_022	-51.532867	36.6557683
Control_024	-8.588321	41.4221814
Control_027	-29.671270	39.5423217
Control_031	-21.411361	36.9352978
Sarcoidosis_001	-28.381400	-35.8772522
Sarcoidosis_005	-2.119743	-19.7447686

```
View(clA)
```

clA	list [9] (S3: kmeans)	List of length 9
cluster	integer [23]	2 2 2 2 2 2 ...
centers	double [3 x 2]	-2.38 -36.55 62.58 -17.79 33.74 7.19
totss	double [1]	51806.08
withinss	double [3]	6256 2352 8290
tot.withinss	double [1]	16898.24
betweenss	double [1]	34907.84
size	integer [3]	13 6 4
iter	integer [1]	2

#Plot K-means Clustering (from ANOVA set)

```
groupsA <- factor(substr(names(datout.log.No.NA.subA),1,1))
groupsA
```

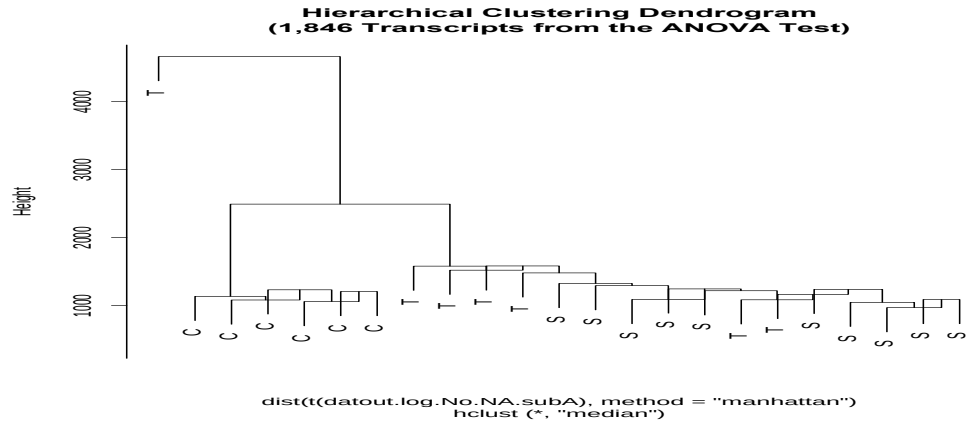
```
 [1] C C C C C C S S S S S S S S S S T T T T T T T
Levels: C S T
```

```
par(oma=c(1,1,1,1))

plot(dat.loadingsA, col=clA$cluster,cex=1, main="PCA Scatter Plot of K-
means Classification\n(1,846 Transcripts from the ANOVA Test)", xlab='PC1',
ylab='PC2')
points(clA$centers, col =1:3, pch ="*",cex=2.5)
text(dat.loadingsA,labels=groupsA, pos=4)
```

*, centroid. PC, Principal Component.

#Plot a Hierarchical Clustering Dendrogram (from ANOVA set)

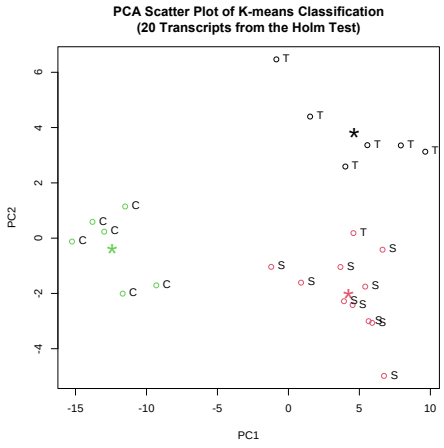
```
datout.log.No.NA.subA <- datout.log.No.NA[rownames(pvA),]
dim(datout.log.No.NA.subA)
[1] 1846   23
groupsA <- factor(substr(names(datout.log.No.NA.subA),1,1))

par(oma=c(1,1,1,1))

hcA <- hclust(dist(t(datout.log.No.NA.subA), method="manhattan"), "median")
plot(hcA, main="Hierarchical Clustering Dendrogram\n(1,846 Transcripts from
the ANOVA Test)", labels=groupsA)
```



 # Multiple Testing Corrections, MTC (Holm and Bonferroni Tests on Log2-transformed data)

#Obtain the List of Transcripts with P-values < 0.05 from Holm Test

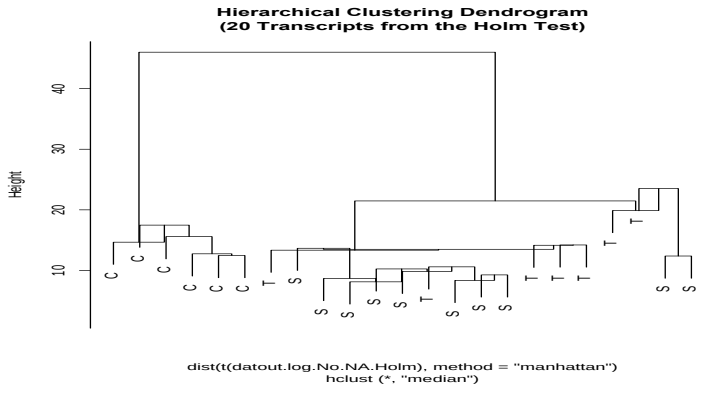
```
aov.run.holm <- p.adjust(aov.run, method="holm")
aov.run.holm

pvh <-as.data.frame(aov.run.holm)
pvH <- subset(pvh, pvh<0.05)
pvH

#Number of Statistically Significant Transcripts from Holm Test
sum(aov.run.holm<.05)
[1] 20

#Subset Data with Genes That Have P-value < 0.05 from Holm Test

datout.log.No.NA.sub.Holm <- datout.log.No.NA[rownames(pvH),]
dim(datout.log.No.NA.sub.Holm)
[1] 20 23

#Calculate PCA on the samples and retain the first two component vectors, k=3

dat.pca.Holm <- prcomp(t(datout.log.No.NA.sub.Holm))
dat.loadings.Holm <- dat.pca.Holm$x[,1:2]

clH <- kmeans(dat.loadings.Holm, centers=3, iter.max=20)

#Plot K-means PCA Scatter Plot from Holm Test.

groupsH <- factor(substr(names(datout.log.No.NA.sub.Holm),1,1))

par(oma=c(1,1,1,1))

plot(dat.loadings.Holm, col = clH$cluster,cex=1, main="PCA Scatter Plot of K-means
Classification\n(20 Transcripts from the Holm Test)", xlab='PC1', ylab='PC2')
points(clH$centers, col = 1:3, pch = "*",cex=2.5)
text(dat.loadings.Holm,labels=groupsH, pos=4)
```



*, centroid

```
#Plot a Hierarchical Clustering Dendrogram from Holm Test.

datout.log.No.NA.Holm <- datout.log.No.NA[rownames(pvH),]
dim(datout.log.No.NA.Holm)
[1] 20 23

groupsH <- factor(substr(names(datout.log.No.NA.sub.Holm),1,1))
```

```
hch <- hclust(dist(t(datout.log.No.NA.Holm), method="manhattan"), "median")
plot(hch, main="Hierarchical Clustering Dendrogram\n(20 Transcripts from
the Holm Test)", labels=groupsH)
```



(not shown in presentation)

Using linear projections of the original data (i.e. cluster centroids, latent variables), classify the samples into their respective classes.

```
#Subset Data with Genes that Have P-value < 0.05 from ANOVA.
datout.log.No.NA.subA <- datout.log.No.NA[rownames(pvA),]
dim(datout.log.No.NA.subA)
[1] 1846   23

#Rename rows creating new column and add new header

#Create factor groups with 3 levels
    groupsA <- factor(substr(names(datout.log.No.NA.subA),1,1))
              > groupsA
               [1] C C C C C C S S S S S S S S S S T T T T T T T
              Levels: C S T

    d.groups <- data.frame(groupsA, t(datout.log.No.NA.subA))

    dim(d.groups)
    [1]   23 1847

    View(d.groups)
```

	groupsA	A2M	A2ML1	AAK1	ABAT	ABCA8
Control_011	C	11.809407	4.667848	8.004322	6.870629	7.692852
Control_018	C	12.477101	6.230979	8.488579	7.502427	9.330889
Control_022	C	10.532769	5.467662	7.396583	7.919306	8.275876
Control_024	C	10.010271	5.152338	8.267713	6.249471	7.321939
Control_027	C	10.831902	4.297682	6.999145	7.365509	6.809839
Control_031	C	10.799529	4.778384	7.433171	4.901415	6.620989
Sarcoidosis_001	S	13.457847	3.751900	8.295180	6.790291	5.300421

```
#Make first row names from column 1.
     rownames(d.groups) <- d.groups[,1]

    rownames(d.groups)<-NULL
    View(d.groups)
```

	groupsA	A2M	A2ML1	AAK1	ABAT
1	C	11.809407	4.667848	8.004322	6.870629
2	C	12.477101	6.230979	8.488579	7.502427
3	C	10.532769	5.467662	7.396583	7.919306
4	C	10.010271	5.152338	8.267713	6.249471
5	C	10.831902	4.297682	6.999145	7.365509
6	C	10.799529	4.778384	7.433171	4.901415
7	S	13.457847	3.751900	8.295180	6.790291

 # Visualization and Comparisons of Train-Test Data via Classification Methods

#Linear Discriminant Analysis of original dataset.

```
> dim(datout)
[1] 27744    23
```

```
groupsall <- factor(substr(names(datout),1,1))
d.groups.all <- data.frame(groupsall,t(datout))
```

```
> groupsall
 [1] C C C C C C S S S S S S S S S S T T T T T T T
Levels: C S T
```

```
View(d.groups.all)
```

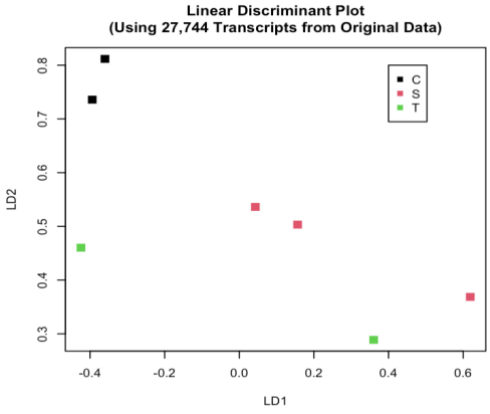
	groupsall	A1BG	A1CF	A2M	A2ML1
Control_011	C	65.55024	2.621931	3589.1004	25.419226
Control_018	C	100.79234	2.651576	5701.4038	75.112406
Control_022	C	48.83145	1.741419	1481.4246	44.251731
Control_024	C	29.59242	1.732562	1031.3160	35.563810
Control_027	C	30.98524	NA	1822.7509	19.666680
Control_031	C	14.54406	NA	1782.3058	27.443337

ving 1 to 6 of 23 entries, 27745 total columns

```
trainall <- d.groups.all[c(1:4,7:13,17:21),]
testall <- d.groups.all[c(5:6,14:16,22:23),]
output <- testall[,1]
testall <- testall[,-1]
  lda.train.all <- lda(groupsall ~ ., data=trainall)
  lda.test.all <- predict(lda.train.all, testall)
```

```
table(lda.test.all$class, output)
  output
    C S T
  C 0 0 0
  S 2 2 2
  T 0 1 0
```

```
 par(oma=c(1,1,1,1))
plot(lda.test1$x,col=as.numeric(labs1),cex=1.5,pch=15,main="Linear Discriminant
Plot\n(Using 27,744 Transcripts from Original Data)", ylab="LD2", xlab="LD1")
legend(0.4,0.8,pch=15,col=
unique(as.numeric(output)),unique(as.character(output)))
```



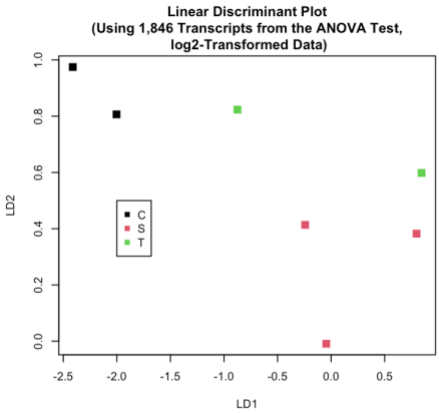
Conclusion: 5 test samples got misclassified.

#Linear Discriminant Analysis on ANOVA (1846 transcripts).

```
train <- d.groups[c(1:4,7:13,17:21),]
test <- d.groups[c(5:6,14:16,22:23),]
labs <- test[,1]
test <- test[,-1]
lda.trainA <- lda(groupsA ~ ., data=train)
lda.testA <- predict(lda.trainA,test)
```

```
table(lda.testA$class, labs)
labs
  C S T
C 2 0 0
S 0 2 1
T 0 1 1
```

```
par(oma=c(1,1,1,1))
plot(lda.testA$x,col=as.numeric(labs),cex=1.5,pch=15,main="Linear
Discriminant Plot\n(Using 1,846 Transcripts from the ANOVA Test,\n log2-
Transformed Data)", ylab="LD2", xlab="LD1")
legend(-2.0,0.5,pch=15,col=
unique(as.numeric(labs)),unique(as.character(labs)))
```



Conclusion: 2 samples from test set were classified incorrectly.

#Subset Data with Genes from the Holm Test(20).

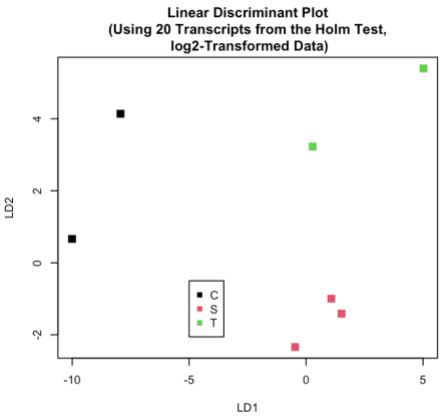
```
datout.log.No.NA.Holm <- datout.log.No.NA[rownames(pvH),]

dim(datout.log.No.NA.Holm)
[1] 20 23

View(datout.log.No.NA.Holm)
```

	Control_011	Control_018	Control_022	Control_024
ALAS1	9.488019	9.625796	8.236702	8.492510
C15orf48	6.206223	5.981261	6.083585	6.098518
C7	9.928106	10.363578	8.454537	8.453027
CCL21	14.030219	15.640148	14.808918	13.605963
CHI3L1	7.552528	9.133276	6.898893	7.927586
CXCL9	9.749784	9.975636	8.138897	7.458444
CYP27B1	9.511228	9.352033	8.310270	9.481552
DNAJC28	5.036008	5.550901	−0.020456	5.309268
	7.015804	5.397410	3.851662	5.133182

```
groupsH <- factor(substr(names(datout.log.No.NA.Holm),1,1))
d.groupsH <- data.frame(groupsH,t(datout.log.No.NA.Holm))
View(d.groupsH)
```

	groups	ALAS1	C15orf48	C7	CCL21	CHI3L1
Control_011	C	9.488019	6.206223	9.928106	14.030219	7.552528
Control_018	C	9.625796	5.981261	10.363578	15.640148	9.133276
Control_022	C	8.236702	6.083585	8.454537	14.808918	6.898893
Control_024	C	8.492510	6.098518	8.453027	13.605963	7.927586
Control_027	C	8.957351	5.320281	10.032477	14.160617	5.763376
Control_031	C	8.503534	6.188508	9.508168	13.825192	8.024509
Sarcoidosis_001	S	12.278727	12.485991	10.191503	13.397618	14.182792
Sarcoidosis_005	S	12.132968	10.415717	8.274248	12.707869	14.237976
Sarcoidosis_007	S	11.459151	10.998071	8.968351	11.010182	14.024740

```
train <- d.groupsH[c(1:4,7:13,17:21),]
test <- d.groupsH[c(5:6,14:16,22:23),]
labs <- test[,1]
test <- test[,-1]
lda.train <- lda(groups ~ ., data=train)
lda.test <- predict(lda.train,test)

table(lda.test$class,labs)
  labs
    Co Sa TB
  Co  2  0  0
  Sa  0  3  0
  TB  0  0  2

 par(oma=c(1,1,1,1))
plot(lda.test$x,col=as.numeric(labs),cex=1.5,pch=15,main="Linear Discriminant
Plot\n(Using 20 Transcripts from the Holm Test,\n log2-Transformed Data)",
ylab="LD2", xlab="LD1")
legend(-5.0,-0.5,pch=15,col= unique(as.numeric(labs)), unique(as.character(labs)))
```



Test sample set was classified correctly.

```
#Support vector machine algorithm (SVM)
library(kernlab)
```

```
datout.log.No.NA.Holm <- datout.log.No.NA[rownames(pvH),]
View(datout.log.No.NA.Holm)
```

	Control_011	Control_018	Control_022	Control_024
ALAS1	9.488019	9.625796	8.23670240	8.49251
C15orf48	6.206223	5.981261	6.08358463	6.09851
C7	9.928106	10.363578	8.45453702	8.45302
CCL21	14.030219	15.640148	14.80891844	13.60596
CHI3L1	7.552528	9.133276	6.89889277	7.92758
CLEC4E	9.749784	2.975636	6.13889759	7.45844

```
  groupsH <- factor(substr(names(datout.log.No.NA.Holm),1,1))
  d.groupsH <- data.frame(groupsH,t(datout.log.No.NA.Holm))

  View(d.groupsH)
```

	groupsH	ALAS1	C15orf48	C7
Control_011	C	9.488019	6.206223	9.928106
Control_018	C	9.625796	5.981261	10.363578
Control_022	C	8.236702	6.083585	8.454537
Control_024	C	8.492510	6.098518	8.453027
Control_027	C	8.957351	5.320281	10.032477
Control_031	C	8.503534	6.188508	9.508168

owing 1 to 6 of 23 entries, 21 total columns

```
dat.f.sub.H <- pvH[rownames(pvH),]
```

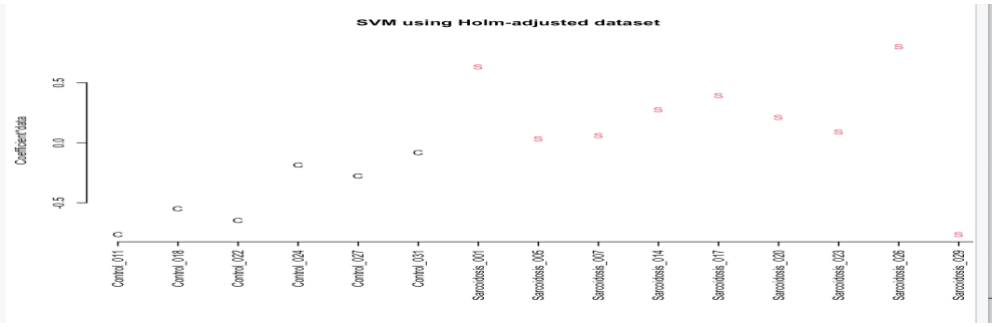
```
> View(dat.f.sub.H)
> dat.f.sub.H
 [1] 2.700519e-02 1.361710e-05 8.518631e-03 6.379488e-04
 [5] 5.786055e-06 3.988156e-02 1.340064e-02 4.673795e-03
 [9] 1.330780e-02 1.337230e-03 6.095653e-03 1.301722e-02
[13] 1.387408e-02 3.384643e-03 4.031312e-03 7.403439e-03
[17] 5.184531e-03 1.877543e-03 2.592348e-03 2.877153e-03
```

```
svp <- ksvm(t(datout.log.No.NA.Holm), d.groupsH$groupsH, type="C-svc")
svp
```

```
> svp
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.115666823818562

Number of Support Vectors : 21

Objective Function Value : -2.4094 -2.5351 -5.8004
Training error : 0.043478
```

```
fit <- fitted(svp)

   par(oma=c(0.2,0.1,0.2,0.1), xpd=NA)
plot(svp@coef[[1]], type="n", ylab="Coefficient*data", xlab="", axes=F,
main="SVM using Holm-adjusted dataset")
text(c(1:23), svp@coef[[1]], fit, col=as.numeric(factor(fit)))
axis(2)
axis(1, at=c(1:23),labels=dimnames(d.groupsH)[[1]], cex.axis=1.0, las=3,
cex=0.3)

table(d.groupsH$groupsH, fit)
> table(d.groupsH$groupsH, fit)
   fit
    C  S  T
```

```
  C  6  0  0
  S  0 10  0
  T  0  1  6
```



one test sample got misclassified.

#Select Differentially Expressed Transcripts from Filtered DATA (10,331 Significantly Expressed Genes)

 # Calculate the Fold Change Between the Groups (data are already on a $\log_2$ scale).

```
foldCvsSC <- apply((datout.log.No.NA[,c(1:6)]),1,mean) - apply((datout.log.No.NA
[,c(7:16)]),1,mean)
summary(foldCvsSC)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.6335 -0.1797  0.2553  0.1954  0.6534  7.2298

foldCvsTB <- apply((datout.log.No.NA[,c(1:6)]),1,mean) - apply((datout.log.No.NA
[,c(17:23)]),1,mean)
summary(foldCvsTB)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.3287  0.2447  0.7672  0.7375  1.2672  6.4738


foldSCvsTB <- apply((datout.log.No.NA[,c(7:16)]),1,mean) - apply((datout.log.No.NA
[,c(17:23)]),1,mean)
summary(foldSCvsTB)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.1012  0.1959  0.5261  0.5421  0.8616  5.8939
```

#Groups Control vs Sarcoidosis, 10,331 genes.
#Student's t-test (Group C vs SC)

```
      t.test.all.genes <- function(x,s1,s2) {
      x1 <- x[s1]
      x2 <- x[s2]
      x1 <- as.numeric(x1)
      x2 <- as.numeric(x2)
      t.out <- t.test(x1,x2, alternative="two.sided", var.equal=T)
      out <- as.numeric(t.out$p.value)
      return(out) }

t.test.runCvsSC <- apply(datout.log.No.NA,1,t.test.all.genes,s1=c(1:6),s2=c(7:16))
```

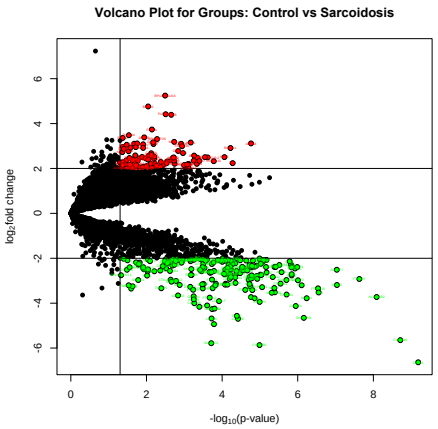
#Transform pvs

```
      p.transCvsSC <- -1 * log10(t.test.runCvsSC)
```

#Volcano Plot for Groups C vs SC
#p-value=0.05

```
      x.line <- -log10(.05)
```

```
#Fold change cut off =2
      y.line <- log2(4)

plot(range(p.transCvsSC),range(foldCvsSC),type="n", xlab=expression(paste("-
",log[10],"(p-value)")), ylab=expression(paste(log[2],"fold change")),
main="Volcano Plot for Groups: Control vs Sarcoidosis")
points(p.transCvsSC,foldCvsSC,col="black",pch=16)
points(p.transCvsSC[(p.transCvsSC>x.line&foldCvsSC>y.line)],
foldCvsSC[(p.transCvsSC>x.line&foldCvsSC>y.line)],col=1,pch=21,bg='red')
points(p.transCvsSC[(p.transCvsSC >x.line&foldCvsSC<(-1*y.line))],
foldCvsSC[(p.transCvsSC>x.line&foldCvsSC< (-1*y.line))], col=1, pch=21,
bg='green')
abline(v=x.line)
abline(h=y.line)
abline(h=(-1*y.line))

text(p.transCvsSC [(p.transCvsSC >x.line&foldCvsSC>y.line)],
foldCvsSC[(p.transCvsSC >x.line&foldCvsSC>y.line)],
col='red',label=dimnames(datout.log.No.NA)[[1]][(p.transCvsSC
>x.line&foldCvsSC>y.line)], cex=0.3)

text(p.transCvsSC [(p.transCvsSC >x.line&foldCvsSC<(-1*y.line))],
foldCvsSC[(p.transCvsSC >x.line&foldCvsSC<(-1*y.line))],
col='green',label=dimnames(datout.log.No.NA)[[1]][(p.transCvsSC
>x.line&foldCvsSC<(-1*y.line))], cex=0.3)
```




```
probesCvsSC <- dimnames(datout.log.No.NA[(t.test.runCvsSC <(.05/10331) &
abs(foldCvsSC)>log2(4)),])[[1]]

summary(probesCvsSC)
Length     Class      Mode
       31 character character

cbind(t.test.runCvsSC[probesCvsSC], foldCvsSC[probesCvsSC])
                 [,1]      [,2]
ADAMDEC1 1.995667e-06 -2.970779
ALAS1    2.682151e-06 -2.620558
ATP6V1B2 3.064140e-06 -2.299304
C15orf48 1.943733e-09 -5.648423
CAPG     1.455742e-06 -2.765726
CHI3L1   6.504628e-10 -6.633524
CLEC7A   1.131246e-06 -4.127716
CTSB     1.510108e-06 -2.285517
```

```
CTSZ     1.640562e-06 -2.136761
DNAJC5B  6.914977e-07 -4.653297
FCER1G   1.184003e-06 -2.456877
FCGR2A   1.972683e-06 -3.613768
FTH1     7.874777e-07 -3.202373
FTH1P10  4.590626e-06 -2.838834
FTH1P11  2.851928e-06 -3.131324
FTH1P2   2.349724e-08 -2.923592
FTH1P21  4.048606e-06 -3.514155
FTH1P23  2.908397e-07 -3.339674
HLA-DQB2 8.188807e-09 -3.725602
IFI27    1.390034e-06 -2.375212
LPCAT2   9.902832e-07 -2.979450
MYOF     4.162601e-06 -3.323409
NUPR1    5.814264e-07 -3.780972
PGD      4.101183e-06 -2.389976
SCPEP1   1.031069e-06 -2.390965
SLAMF7   9.278260e-08 -2.510986
SNX10    9.197458e-08 -3.190226
TFEC     1.447131e-06 -2.711974
TFRC     1.458967e-06 -2.405256
TLR8     2.794675e-07 -3.524701
TYROBP   3.803220e-06 -2.383775
```

#Group SC vs TB (10,331 Significantly Expressed transcripts).

#Student's t-test (Group C vs TB)

```
t.test.all.genes <- function(x,s1,s2) {
      x1 <- x[s1]
      x2 <- x[s2]
      x1 <- as.numeric(x1)
      x2 <- as.numeric(x2)
      t.out <- t.test(x1,x2, alternative="two.sided", var.equal=T)
      out <- as.numeric(t.out$p.value)
      return(out) }

t.test.runSCvsTB <-
apply(datout.log.No.NA,1,t.test.all.genes,s1=c(7:16),s2=c(17:23))
```

#Transform pvs

```
p.transSCvsTB <- -1 * log10(t.test.runSCvsTB)
```

#Volcano Plot for Groups SC vs TB

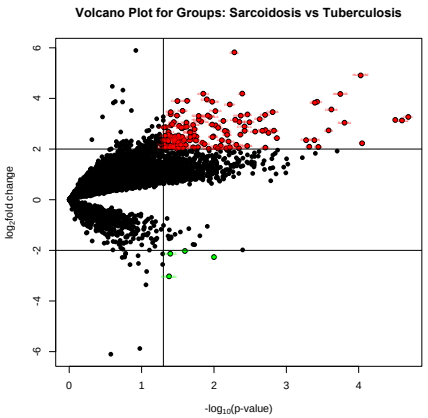
#p-value=0.05

```
      x.line <- -log10(.05)
```

#fold change cut off =2

```
      y.line <- log2(4)

plot(range(p.transSCvsTB),range(foldSCvsTB),type="n", xlab=expression(paste("-
",log[10],"(p-value)")), ylab=expression(paste(log[2],"fold change")),
main="Volcano Plot for Groups: Sarcoidosis vs Tuberculosis")
points(p.transSCvsTB,foldSCvsTB,col="black",pch=16)
points(p.transSCvsTB[(p.transSCvsTB>x.line&foldSCvsTB>y.line)],
foldSCvsTB[(p.transSCvsTB>x.line&foldSCvsTB>y.line)],col=1,pch=21,bg='red')
points(p.transSCvsTB[(p.transSCvsTB >x.line&foldSCvsTB<(-1*y.line))],
foldSCvsTB[(p.transSCvsTB>x.line&foldSCvsTB< (-1*y.line))], col=1, pch=21,
bg='green')
abline(v=x.line)
abline(h=y.line)
```

```
abline(h=(-1*y.line))

text(p.transSCvsTB [(p.transSCvsTB >x.line&foldSCvsTB>y.line)],
foldSCvsTB[(p.transSCvsTB >x.line&foldSCvsTB>y.line)],
col='red',label=dimnames(datout.log.No.NA)[[1]][(p.transSCvsTB
>x.line&foldSCvsTB>y.line)], cex=0.3)

text(p.transSCvsTB [(p.transSCvsTB >x.line&foldSCvsTB<(-1*y.line))],
foldSCvsTB[(p.transSCvsTB >x.line&foldSCvsTB<(-1*y.line))],
col='green',label=dimnames(datout.log.No.NA)[[1]][(p.transSCvsTB
>x.line&foldSCvsTB<(-1*y.line))], cex=0.3)
```



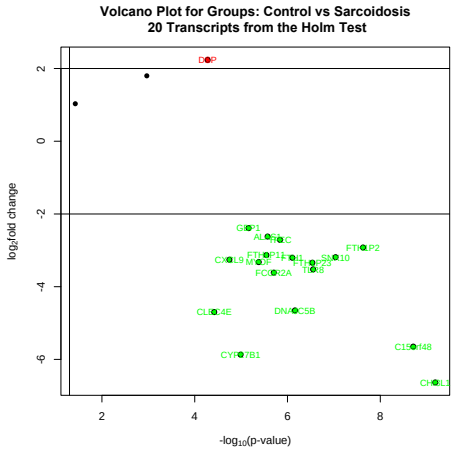

```
probesSCvsTB <- dimnames(datout.log.No.NA[(t.test.runSCvsTB <(.05/10331) &
abs(foldSCvsTB)>log2(4)),])[[1]]

summary(probesSCvsTB)
Length     Class      Mode
         0 character character

cbind(t.test.runSCvsTB[probesSCvsTB], foldSCvsTB[probesSCvsTB])
   [,1] [,2]
(you should get 69-genes column)
```

#Identify DEG in Data Subset from the Holm Test (20 transcripts).
 # Calculate the Fold Change Between the Groups.

```
#Subset Data With Genes from the Holm Test
datout.log.No.NA.Holm <- datout.log.No.NA[rownames(pvH),]
datout.log.No.NA.Holm
dim(datout.log.No.NA.Holm)

foldCvsSC <- apply((datout.log.No.NA.Holm[,c(1:6)]),1,mean) -
apply((datout.log.No.NA.Holm[,c(7:16)]),1,mean)

summary(foldCvsSC)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -6.634  -3.874  -3.230  -2.983  -2.689   2.236


foldCvsTB <- apply((datout.log.No.NA.Holm [,c(1:6)]),1,mean) -
apply((datout.log.No.NA.Holm [,c(17:23)]),1,mean)

summary(foldCvsTB)
```

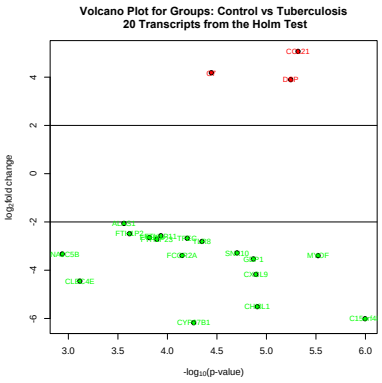
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -6.175  -3.693  -3.046  -2.404  -2.552   5.067
```

```
foldSCvsTB <- apply((datout.log.No.NA.Holm [,c(7:16)]),1,mean) -
apply((datout.log.No.NA.Holm [,c(17:23)]),1,mean)

summary(foldSCvsTB)


#Group C vs SC (from Holm Test)
#Student's t-test (Group C vs SC) and Volcano Plot

t.test.all.genes <- function(x,s1,s2) {
      x1 <- x[s1]
      x2 <- x[s2]
      x1 <- as.numeric(x1)
      x2 <- as.numeric(x2)
      t.out <- t.test(x1,x2, alternative="two.sided", var.equal=T)
      out <- as.numeric(t.out$p.value)
      return(out) }

t.test.runCvsSC <-
apply(datout.log.No.NA.Holm,1,t.test.all.genes,s1=c(1:6),s2=c(7:16))

#Transform pvs
p.transCvsSC <- -1 * log10(t.test.runCvsSC)

#Volcano Plot for Groups C vs SC
#p-value=0.05
x.line <- -log10(.05)

#log fold change comparaison >2, C vs SC
y.line <- log2(4)

plot(range(p.transCvsSC),range(foldCvsSC),type="n", xlab=expression(paste("-
",log[10],"(p-value)")), ylab=expression(paste(log[2],"fold change")),
main="Volcano Plot for Groups: Control vs Sarcoidosis\n 20 Transcripts from the
Holm Test")
points(p.transCvsSC,foldCvsSC, col="black",pch=16)
points(p.transCvsSC[(p.transCvsSC>x.line&foldCvsSC>y.line)],
foldCvsSC[(p.transCvsSC>x.line&foldCvsSC>y.line)],col=1,pch=21,bg='red')
points(p.transCvsSC[(p.transCvsSC >x.line&foldCvsSC<(-1*y.line))],
foldCvsSC[(p.transCvsSC>x.line&foldCvsSC< (-1*y.line))], col=1, pch=21,
bg='green')
abline(v=x.line)
abline(h=y.line)
abline(h=(-1*y.line))

text(p.transCvsSC [(p.transCvsSC >x.line&foldCvsSC>y.line)],
foldCvsSC[(p.transCvsSC >x.line&foldCvsSC>y.line)],
col='red',label=dimnames(datout.log.No.NA.Holm)[[1]][(p.transCvsSC
>x.line&foldCvsSC>y.line)], cex=0.8)

text(p.transCvsSC [(p.transCvsSC >x.line&foldCvsSC<(-1*y.line))],
foldCvsSC[(p.transCvsSC >x.line&foldCvsSC<(-1*y.line))],
col='green',label=dimnames(datout.log.No.NA.Holm)[[1]][(p.transCvsSC
>x.line&foldCvsSC<(-1*y.line))], cex=0.8)
```




```
probesCvsSC <- dimnames(datout.log.No.NA.Holm [(t.test.runCvsSC <(.05/20) &
abs(foldCvsSC)>log2(4)),])[[1]]

summary(probesCvsSC)
Length     Class      Mode
       18 character character

cbind(t.test.runCvsSC[probesCvsSC], foldCvsSC[probesCvsSC])
                [,1]      [,2]
ALAS1    2.682151e-06 -2.620558
C15orf48 1.943733e-09 -5.648423
CHI3L1   6.504628e-10 -6.633524
CLEC4E   3.799928e-05 -4.698049
CXCL9    1.784240e-05 -3.257872
CYP27B1  1.026112e-05 -5.868966
DNAJC5B  6.914977e-07 -4.653297
DSP      5.272044e-05  2.236074
FCGR2A   1.972683e-06 -3.613768
FTH1     7.874777e-07 -3.202373
FTH1P11  2.851928e-06 -3.131324
FTH1P2   2.349724e-08 -2.923592
FTH1P23  2.908397e-07 -3.339674
GBP1     6.899795e-06 -2.389508
MYOF     4.162601e-06 -3.323409
SNX10    9.197458e-08 -3.190226
TFEC     1.447131e-06 -2.711974
TLR8     2.794675e-07 -3.524701

## DEG Group C vs TB (from Holm Test)
#Student's t-test (Group C vs TB) and Volcano Plot

t.test.all.genes <- function(x,s1,s2) {
      x1 <- x[s1]
      x2 <- x[s2]
      x1 <- as.numeric(x1)
      x2 <- as.numeric(x2)
      t.out <- t.test(x1,x2, alternative="two.sided", var.equal=T)
      out <- as.numeric(t.out$p.value)
      return(out) }

t.test.runCvsTB <-
apply(datout.log.No.NA.Holm,1,t.test.all.genes,s1=c(1:6),s2=c(17:23))

#Transform pvs
```

```
p.transCvsTB <- -1 * log10(t.test.runCvsTB)

#Volcano Plot for Groups C vs TB
#p-value=0.05
x.line <- -log10(.05)

#fold change cut off =2
y.line <- log2(4)

plot(range(p.transCvsTB),range(foldCvsTB),type="n", xlab=expression(paste("-
",log[10],"(p-value)")), ylab=expression(paste(log[2],"fold change")),
main="Volcano Plot for Groups: Control vs Tuberculosis\n 20 Transcripts from the
Holm Test")
points(p.transCvsTB,foldCvsTB, col="black",pch=16)
points(p.transCvsTB[(p.transCvsTB>x.line&foldCvsTB>y.line)],
foldCvsTB[(p.transCvsTB>x.line&foldCvsTB>y.line)],col=1,pch=21,bg='red')
points(p.transCvsTB[(p.transCvsTB >x.line&foldCvsTB<(-1*y.line))],
foldCvsTB[(p.transCvsTB>x.line&foldCvsTB< (-1*y.line))], col=1, pch=21,
bg='green')
abline(v=x.line)
abline(h=y.line)
abline(h=(-1*y.line))

text(p.transCvsTB [(p.transCvsTB >x.line&foldCvsTB>y.line)],
foldCvsTB[(p.transCvsTB >x.line&foldCvsTB>y.line)],
col='red',label=dimnames(datout.log.No.NA.Holm)[[1]][(p.transCvsTB
>x.line&foldCvsTB>y.line)], cex=0.8)

text(p.transCvsTB [(p.transCvsTB >x.line&foldCvsTB<(-1*y.line))],
foldCvsTB[(p.transCvsTB >x.line&foldCvsTB<(-1*y.line))],
col='green',label=dimnames(datout.log.No.NA.Holm)[[1]][(p.transCvsTB
>x.line&foldCvsTB<(-1*y.line))], cex=0.8)
```




```
probesCvsTB <- dimnames(datout.log.No.NA.Holm[(t.test.runCvsTB <(.05/20) &
abs(foldCvsTB)>log2(4)),])[[1]]
summary(probesCvsTB)
Length     Class      Mode
       20 character character

cbind(t.test.runCvsTB[probesCvsTB], foldCvsTB[probesCvsTB])
ALAS1    2.747376e-04 -2.060521
C15orf48 1.009456e-06 -6.010223
C7       3.599599e-05  4.183338
CCL21    4.800004e-06  5.066724
```

```
CHI3L1   1.237486e-05 -5.509020
CLEC4E   7.666812e-04 -4.450430
CXCL9    1.276823e-05 -4.173201
CYP27B1  5.428212e-05 -6.175469
DNAJC5B  1.153625e-03 -3.328349
DSP      5.690382e-06  3.897188
FCGR2A   7.118997e-05 -3.390011
FTH1     1.521722e-04 -2.648333
FTH1P11  1.164838e-04 -2.573746
FTH1P2   2.415613e-04 -2.486179
FTH1P23  1.277927e-04 -2.719947
GBP1     1.352982e-05 -3.532579
MYOF     3.009921e-06 -3.395993
SNX10    1.980435e-05 -3.282646
TFEC     6.314291e-05 -2.675928
TLR8     4.474801e-05 -2.809198
```

##Select Differentially Expressed Transcripts from Data Subset = 1,846 transcripts from the ANOVA Test.

 # Calculate the Fold Change Between the Groups.

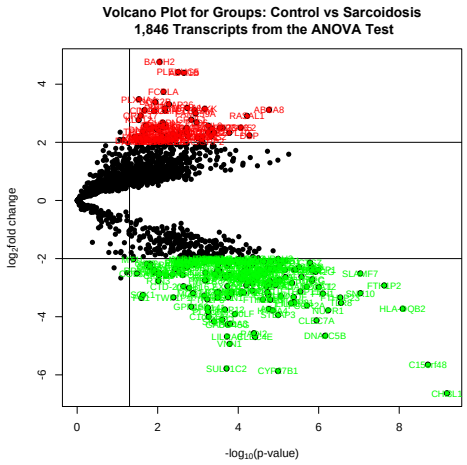
#Subset Data with Genes that Have P-value < 0.05 from ANOVA

```
datout.log.No.NA.subA <- datout.log.No.NA[rownames(pvA),]
dim(datout.log.No.NA.subA)
[1] 1846   23

foldCvsSCa <- apply((datout.log.No.NA.subA [,c(1:6)]),1,mean) -
apply((datout.log.No.NA.subA [,c(7:16)]),1,mean)

summary(foldCvsSCa)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.6335 -0.5814  0.5541  0.1300  0.9862  4.7626

foldCvsTBa <- apply((datout.log.No.NA.subA [,c(1:6)]),1,mean) -
apply((datout.log.No.NA.subA [,c(17:23)]),1,mean)

summary(foldCvsTBa)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.1755  0.7366  1.4060  1.0515  1.9422  6.4738

foldSCvsTBa <- apply((datout.log.No.NA.subA [,c(7:16)]),1,mean) -
apply((datout.log.No.NA.subA [,c(17:23)]),1,mean)

summary(foldSCvsTBa)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.0265  0.5519  0.8633  0.9216  1.2527  5.8200
```

#Group comparison C vs SC (1,846 transcripts)
#Student's t-test (Group C vs SC) and Volcano Plot.

```
t.test.all.genes <- function(x,s1,s2) {
      x1 <- x[s1]
      x2 <- x[s2]
      x1 <- as.numeric(x1)
      x2 <- as.numeric(x2)
      t.out <- t.test(x1,x2, alternative="two.sided", var.equal=T)
      out <- as.numeric(t.out$p.value)
```

```
      return(out) }

t.test.runCvsSCa <-
apply(datout.log.No.NA.subA,1,t.test.all.genes,s1=c(1:6),s2=c(7:16))

#Transform pvs
p.transCvsSCa <- -1 * log10(t.test.runCvsSCa)

#Volcano Plot for Groups C vs SC
#p-value=0.05
x.line <- -log10(.05)

#fold change=2
y.line <- log2(4)

plot(range(p.transCvsSCa),range(foldCvsSCa),type="n", xlab=expression(paste("-
",log[10],"(p-value)")), ylab=expression(paste(log[2],"fold change")),
main="Volcano Plot for Groups: Control vs Sarcoidosis\n 1,846 Transcripts from the
ANOVA Test")
points(p.transCvsSCa,foldCvsSCa, col="black",pch=16)
points(p.transCvsSCa[(p.transCvsSCa>x.line&foldCvsSCa>y.line)],
foldCvsSCa[(p.transCvsSCa>x.line&foldCvsSCa>y.line)],col=1,pch=21,bg='red')
points(p.transCvsSCa[(p.transCvsSCa >x.line&foldCvsSCa<(-1*y.line))],
foldCvsSCa[(p.transCvsSCa>x.line&foldCvsSCa< (-1*y.line))], col=1, pch=21,
bg='green')
abline(v=x.line)
abline(h=y.line)
abline(h=(-1*y.line))

text(p.transCvsSCa [(p.transCvsSCa >x.line&foldCvsSCa>y.line)],
foldCvsSCa[(p.transCvsSCa >x.line&foldCvsSCa>y.line)],
col='red',label=dimnames(datout.log.No.NA.subA)[[1]][(p.transCvsSCa
>x.line&foldCvsSCa>y.line)], cex=0.8)

text(p.transCvsSCa [(p.transCvsSCa >x.line&foldCvsSCa<(-1*y.line))],
foldCvsSCa[(p.transCvsSCa >x.line&foldCvsSCa<(-1*y.line))],
col='green',label=dimnames(datout.log.No.NA.subA)[[1]][(p.transCvsSCa
>x.line&foldCvsSCa<(-1*y.line))], cex=0.8)
```



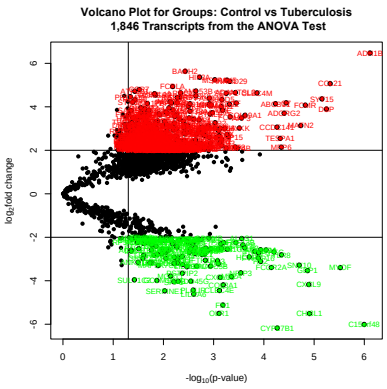

```
probesCvsSCa <- dimnames(datout.log.No.NA.subA [(t.test.runCvsSCa <(.05/1846) &
abs(foldCvsSCa)>log2(4)),])[[1]]
```

```
summary(probesCvsSCa)
Length     Class      Mode
       68 character character

cbind(t.test.runCvsSCa[probesCvsSCa], foldCvsSCa[probesCvsSCa])
                  [,1]      [,2]
ABCA8     1.718201e-05  3.116552
ACO1      1.877782e-05 -2.304259
ADAMDEC1  1.995667e-06 -2.970779
AK4       1.728302e-05 -3.730563
ALAS1     2.682151e-06 -2.620558
ATP6V1B2  3.064140e-06 -2.299304
C15orf48  1.943733e-09 -5.648423
C1QB      1.243881e-05 -2.333185
CAPG      1.455742e-06 -2.765726
CD68      9.937159e-06 -2.074532
CHI3L1    6.504628e-10 -6.633524
CLEC7A    1.131246e-06 -4.127716
CTSB      1.510108e-06 -2.285517
CTSD      2.306033e-05 -2.102741
CTSZ      1.640562e-06 -2.136761
CXCL16    1.874852e-05 -2.836073
CXCL9     1.784240e-05 -3.257872
CYP27B1   1.026112e-05 -5.868966
DNAJC5B   6.914977e-07 -4.653297
DOCK4     9.785807e-06 -2.731029
DRAM1     2.416273e-05 -2.498420
FCER1G    1.184003e-06 -2.456877
FCGR2A    1.972683e-06 -3.613768
FTH1      7.874777e-07 -3.202373
FTH1P10   4.590626e-06 -2.838834
FTH1P11   2.851928e-06 -3.131324
FTH1P16   2.443195e-05 -3.408279
FTH1P2    2.349724e-08 -2.923592
FTH1P21   4.048606e-06 -3.514155
FTH1P23   2.908397e-07 -3.339674
GBP1      6.899795e-06 -2.389508
GNS       6.694482e-06 -2.072059
HEXB      1.371907e-05 -2.374707
HLA-DQB2  8.188807e-09 -3.725602
HTRA4     1.381300e-05 -3.784227
IFI27     1.390034e-06 -2.375212
ITGB5     1.617719e-05 -2.877775
LILRB3    4.857033e-06 -3.536005
LPCAT2    9.902832e-07 -2.979450
LYZ       1.965911e-05 -2.962790
MCTP1     1.657141e-05 -2.403003
MMP14     2.158014e-05 -2.544764
MOSPD1    8.364913e-06 -2.155091
MREG      1.215367e-05 -2.790176
MYOF      4.162601e-06 -3.323409
NCF2      9.174493e-06 -2.748483
NR1H3     1.021327e-05 -2.011003
NRIP3     1.588936e-05 -3.417022
NUPR1     5.814264e-07 -3.780972
PGD       4.101183e-06 -2.389976
PLA2G7    6.685064e-06 -2.631595
PLAU      1.884722e-05 -3.149676
```

```
PLBD1    8.779794e-06 -3.288518
PLXDC2   1.702212e-05 -2.899976
PTAFR    1.656208e-05 -2.761481
SCPEP1   1.031069e-06 -2.390965
SIRPA    1.620093e-05 -2.340678
SLAMF7   9.278260e-08 -2.510986
SLC1A3   1.180835e-05 -2.916421
SNX10    9.197458e-08 -3.190226
SOD2     9.488609e-06 -2.248187
STEAP3   1.018094e-05 -3.944244
TFEC     1.447131e-06 -2.711974
TFRC     1.458967e-06 -2.405256
TLR8     2.794675e-07 -3.524701
TYROBP   3.803220e-06 -2.383775
UBE2D1   7.980097e-06 -2.075839
WARS     1.529053e-05 -2.104042
```

```
#Group comparison C vs TB (1,846 transcripts).
#Student's t-test (Group C vs TB)

t.test.all.genes <- function(x,s1,s2) {
      x1 <- x[s1]
      x2 <- x[s2]
      x1 <- as.numeric(x1)
      x2 <- as.numeric(x2)
      t.out <- t.test(x1,x2, alternative="two.sided", var.equal=T)
      out <- as.numeric(t.out$p.value)
      return(out) }

t.test.runCvsTBa <-
apply(datout.log.No.NA.subA,1,t.test.all.genes,s1=c(1:6),s2=c(17:23))

#Transform pvs
p.transCvsTBa <- -1 * log10(t.test.runCvsTBa)

#Volcano Plot for Groups C vs SC
#p-value=0.05
x.line <- -log10(.05)

#fold change=2
y.line <- log2(4)

plot(range(p.transCvsTBa),range(foldCvsTBa),type="n", xlab=expression(paste("-
",log[10],"(p-value)")), ylab=expression(paste(log[2],"fold change")),
main="Volcano Plot for Groups: Control vs Tuberculosis\n 1,846 Transcripts from
the ANOVA Test")
points(p.transCvsTBa,foldCvsTBa, col="black",pch=16)
points(p.transCvsTBa[(p.transCvsTBa>x.line&foldCvsTBa>y.line)],
foldCvsTBa[(p.transCvsTBa>x.line&foldCvsTBa>y.line)],col=1,pch=21,bg='red')
points(p.transCvsTBa[(p.transCvsTBa >x.line&foldCvsTBa<(-1*y.line))],
foldCvsTBa[(p.transCvsTBa>x.line&foldCvsTBa< (-1*y.line))], col=1, pch=21,
bg='green')
abline(v=x.line)
abline(h=y.line)
abline(h=(-1*y.line))

text(p.transCvsTBa [(p.transCvsTBa >x.line&foldCvsTBa>y.line)],
foldCvsTBa[(p.transCvsTBa >x.line&foldCvsTBa>y.line)],
```

```
col='red',label=dimnames(datout.log.No.NA.subA)[[1]][(p.transCvsTBa
>x.line&foldCvsTBa>y.line)], cex=0.8)

text(p.transCvsTBa [(p.transCvsTBa >x.line&foldCvsTBa<(-1*y.line))],
foldCvsTBa[(p.transCvsTBa >x.line&foldCvsTBa<(-1*y.line))],
col='green',label=dimnames(datout.log.No.NA.subA)[[1]][(p.transCvsTBa
>x.line&foldCvsTBa<(-1*y.line))], cex=0.8)
```




```
probesCvsTBa <- dimnames(datout.log.No.NA.subA [(t.test.runCvsTBa <(.05/1846) &
abs(foldCvsTBa)>log2(4)),])[[1]]
summary(probesCvsTBa)
   Length     Class      Mode
       12 character character

cbind(t.test.runCvsTBa[probesCvsTBa], foldCvsTBa[probesCvsTBa])
                 [,1]      [,2]
ADH1B     7.697494e-07  6.473826
C15orf48  1.009456e-06 -6.010223
CCL21     4.800004e-06  5.066724
CHI3L1    1.237486e-05 -5.509020
CXCL9     1.276823e-05 -4.173201
DSP       5.690382e-06  3.897188
FCMR      1.498026e-05  4.077411
GBP1      1.352982e-05 -3.532579
MATN2     1.839001e-05  3.136921
MYOF      3.009921e-06 -3.395993
SNX10     1.980435e-05 -3.282646
SYT15     6.986989e-06  4.363464
```

Using the top 5 discriminant gene names (positive and negative direction) and functional information (NCBI's DAVID associated pathways, GO terms) for top 10 genes.

Genes	CvsSC pvalue	CvsSC FC	CvsTB pvalue	CvsTB FC	GO, biological process	Cellular Pathway	name
CHI3L1	6.50E-10	-6.633524	1.24E-05	-5.509	inflammatory response	regulation of neutrophil chemotaxis	chitinase 3 like 1
C15orf48	1.94E-09	-5.648423			oxidation-reduction	ion transport	chromosome 15 open reading frame 48
HLA-DQB2	8.19E-09	-3.725602			innate immune response	antigen processing and presentation	Major Histocompatibility Complex, Class II, DQ Beta 2
CLEC7A	1.13E-06	-4.127716			innate immune response	pattern recognition receptor signaling	C-type lectin domain family 7 member A
NUPR1	5.81E-07	-3.780972			inflammatory response	regulation of transcription by RNA polymerase II	Nuclear protein 1, short sequence motif:Nuclear localization signal
DSP	5.27E-05	2.236074			Glycolysis / Gluconeogenesis	ethanol oxidation	Alcohol dehydrogenase 1B, chain
CCL21	5.27E-05	2.236074	4.80E-06	5.06672	immune response	Chemokine signaling	C-C motif chemokine ligand 21
ABCA8	1.72E-05	3.116552			xenobiotic transport	ABC transporter	ATP binding cassette subfamily A member 8
C7			3.60E-05	4.18334	innate immune response	complement activation signaling	complement C7
MATN2	1.84E-05	3.136921			calcium ion binding	cell migration	matrilin 2

See presentation.