hw2

Cheng Li

2017/9/26

Data Import

```
anno = read.table("/Users/Hana/Desktop/STA/sta46/hw2/SampleAnnotation.txt", as.is=TRUE, sep="\t", quote=
                  row.names=1, header=TRUE)
x = read.table("/Users/Hana/Desktop/STA/sta46/hw2/expressiondata.txt", as.is=TRUE, sep="\t", quote="", r
x = as.matrix(x)
```

Define samples and colors and phenotype

```
samples = rownames(anno)
colors = rainbow(nrow(anno))
isNorm = anno$TissueType == "norm"
isSick = anno$TissueType == "sick"
isAcute = anno$TissueType == "acute"
```

Data Transformation

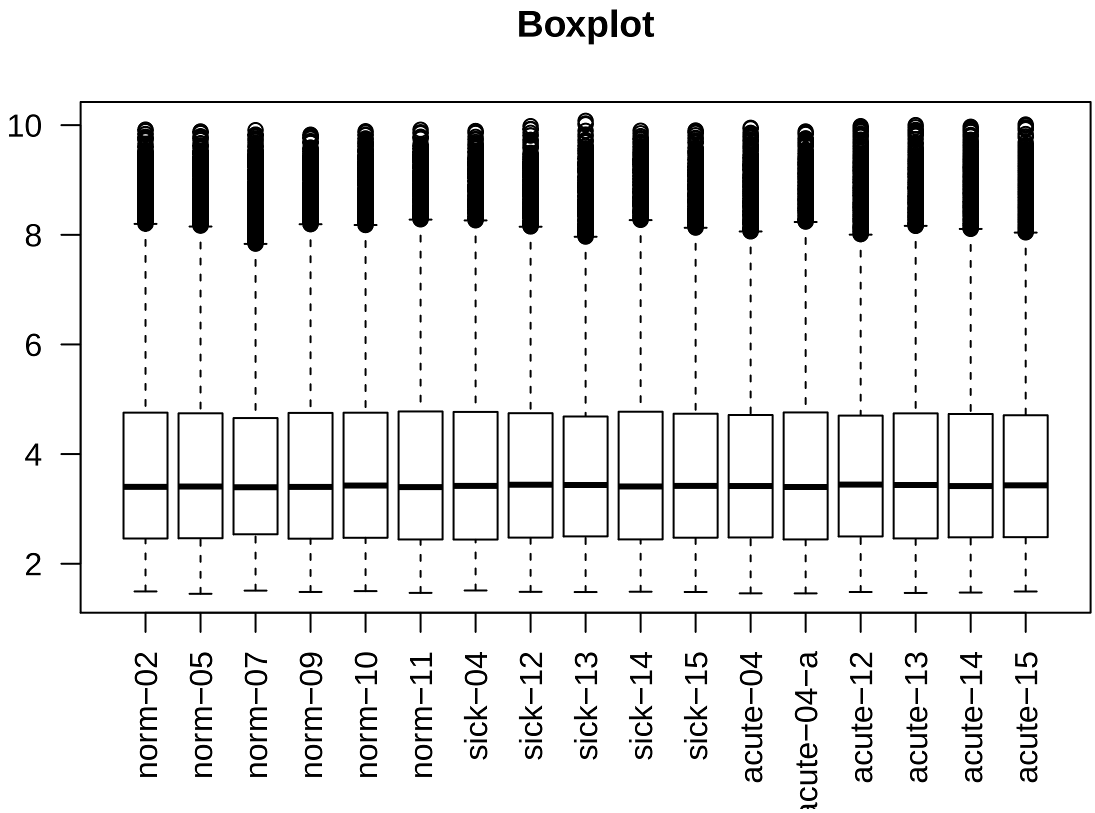
It seems to be too dispersed when we try boxplot on the original data. So we transfer data into log-form to do the further analysis.

```
data <- log(x)
```

Distributions

Boxplot

```
boxplot(data,las=2,main="Boxplot")
```

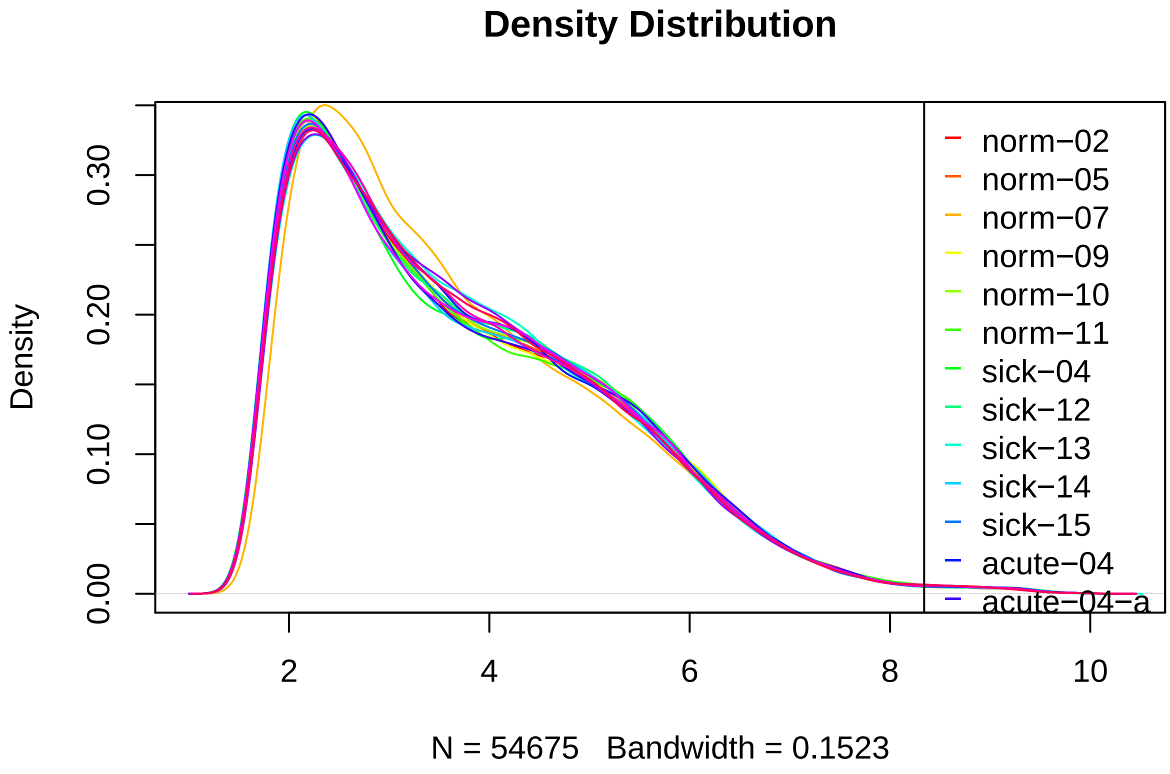


```
outliers = data[data[,1]>6|data[,2]>6|data[,3]>6|data[,4]>6|data[,5]>6|data[,6]>6|data[,7]>6|data[,8]>6
```

The plot shows us that there are some outliers of each sample, which refer to highly expressed genes.

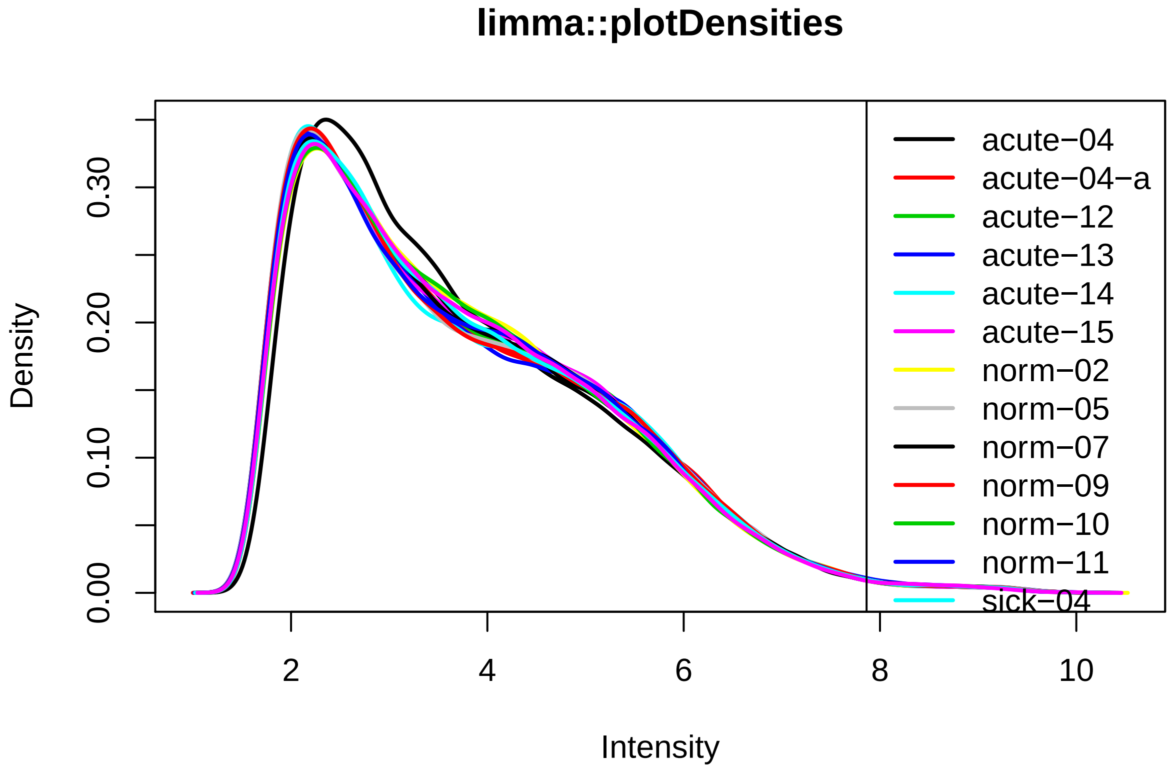
Density

```
plot(density(data[,1]),col=colors[1],main="Density Distribution")
for(i in 2:ncol(data))
{
  lines(density(data[,i]),col=colors[i])
}
legend("topright",col=colors,legend=samples,pch="-")
```



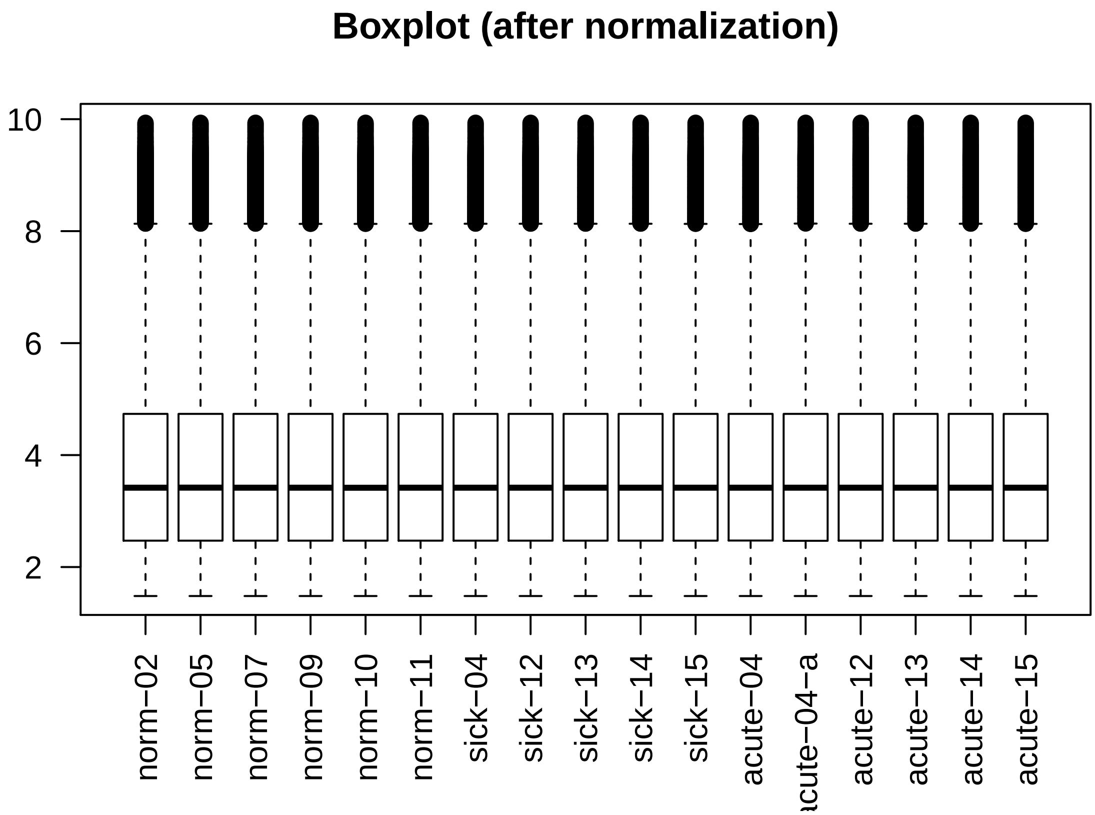
limma::plotDensities

```
library(limma)
plotDensities(data,main="limma::plotDensities",legend="topright")
```

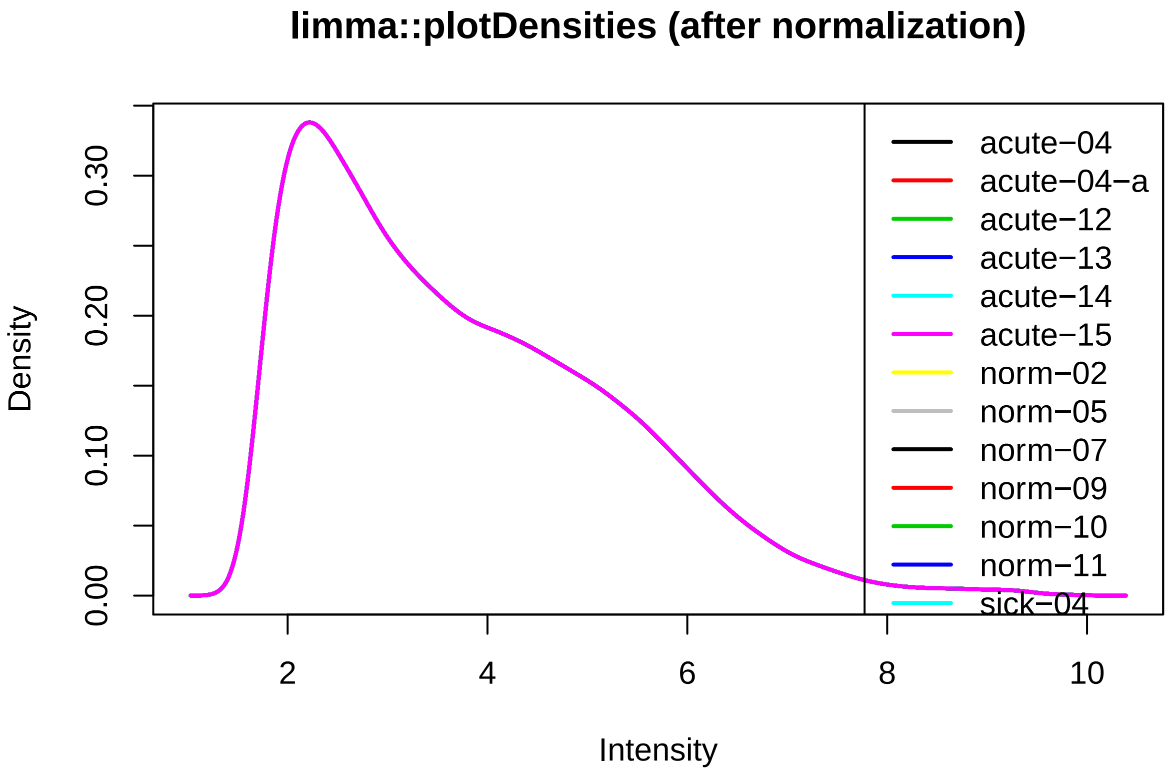


Normalization

```r
norm_data<-normalizeQuantiles(data,ties=TRUE)
boxplot(norm_data,las=2,main="Boxplot (after normalization)")
```



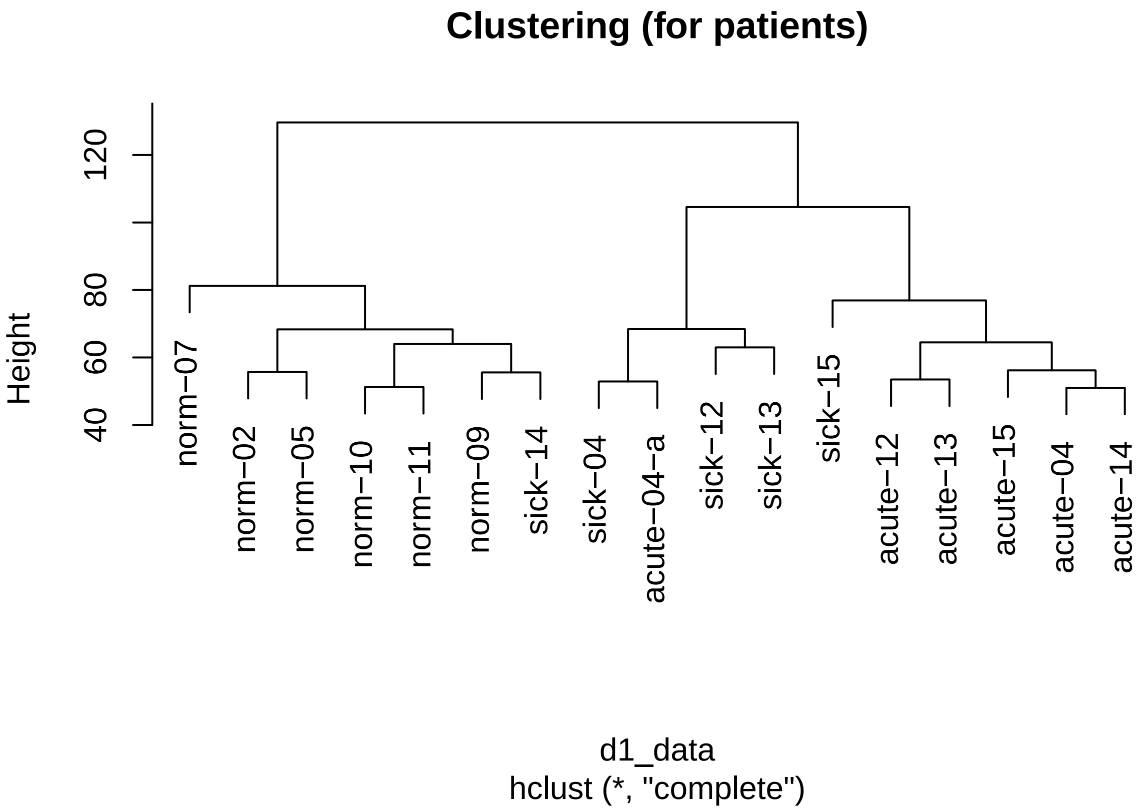
```r
plotDensities(norm_data,main="limma::plotDensities (after normalization)",legend="topright")
```



Clustering

For patients:

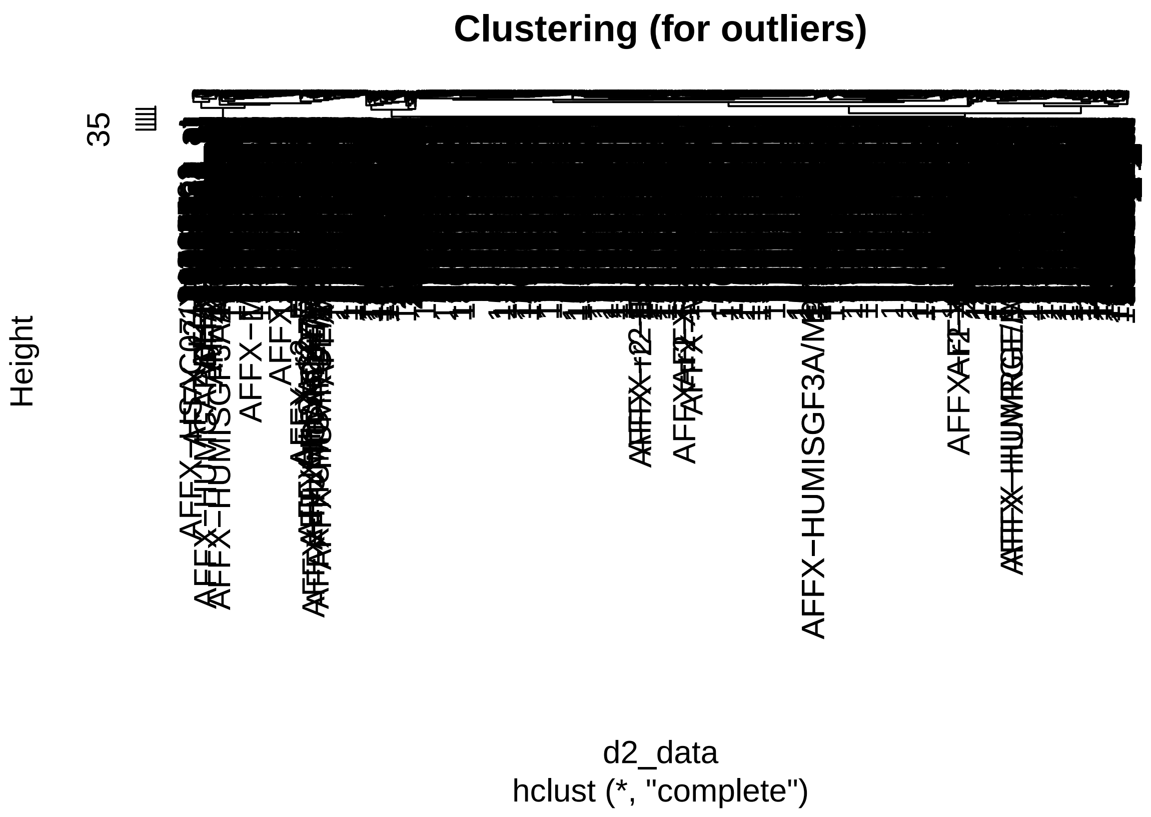
```
d1_data<-dist(t(norm_data))
clusters_p<-hclust(d1_data)
plot(clusters_p,main="Clustering (for patients)")
```



For genes:

The whole dataset is too large to perform clustering among all genes. Instead, we select thoses genes which are outliers(log-value larger than 6 according to boxplot above) in at least one sample.

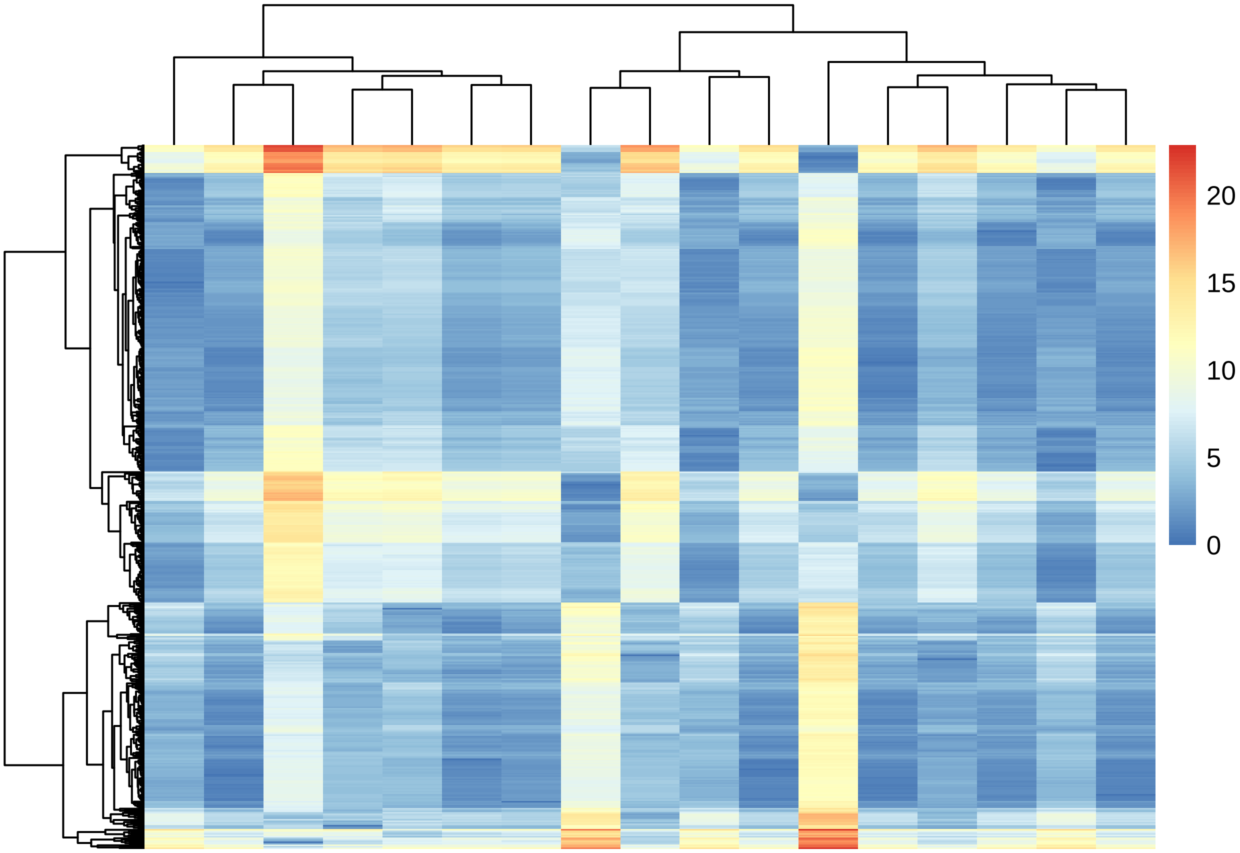
```
norm_outliers<-normalizeQuantiles(outliers,ties=TRUE)
d2_data <-dist(norm_outliers)
clusters_g<-hclust(d2_data)
plot(clusters_g,main="Clustering (for outliers)")
```



Heatmap

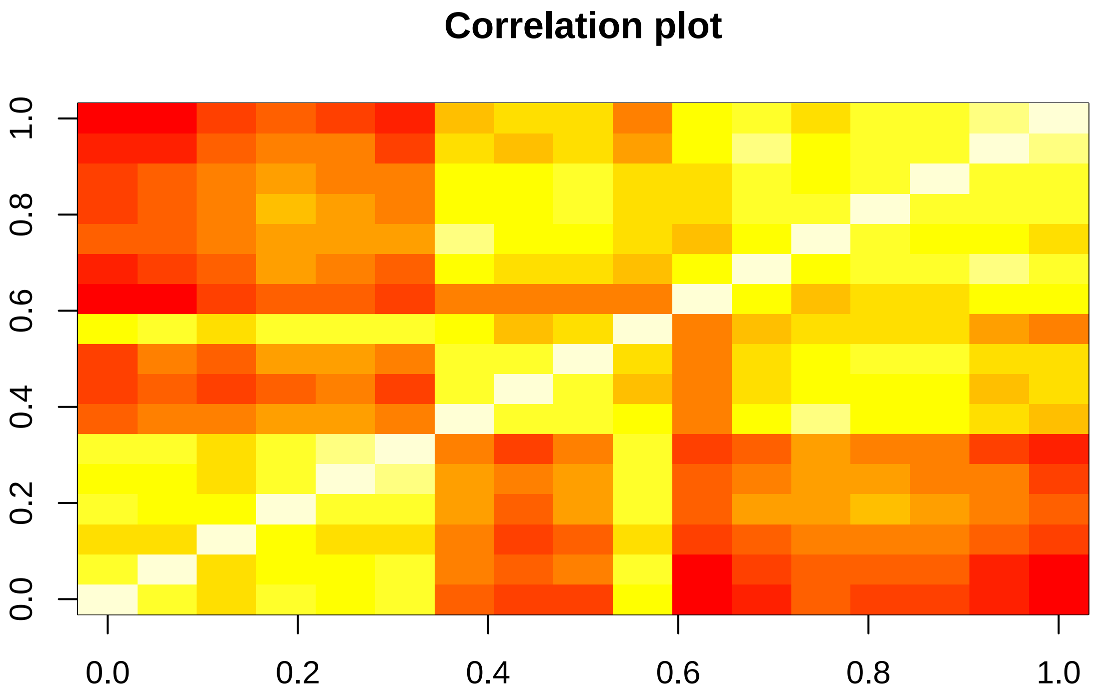
Plot the hearmap of outliers. However, the dataset is still too large to run. So I randomly select 500 genes from outliers.

```
s_outliers=outliers[sample(nrow(outliers),500),]
norm_s_outliers<-normalizeQuantiles(s_outliers,ties=TRUE)
ds_data <-dist(norm_s_outliers)
library(pheatmap)
clusters_g2<-hclust(ds_data)
pheatmap(ds_data,cluster_rows =clusters_g2,cluster_cols=clusters_p )
```



Correlation matrix

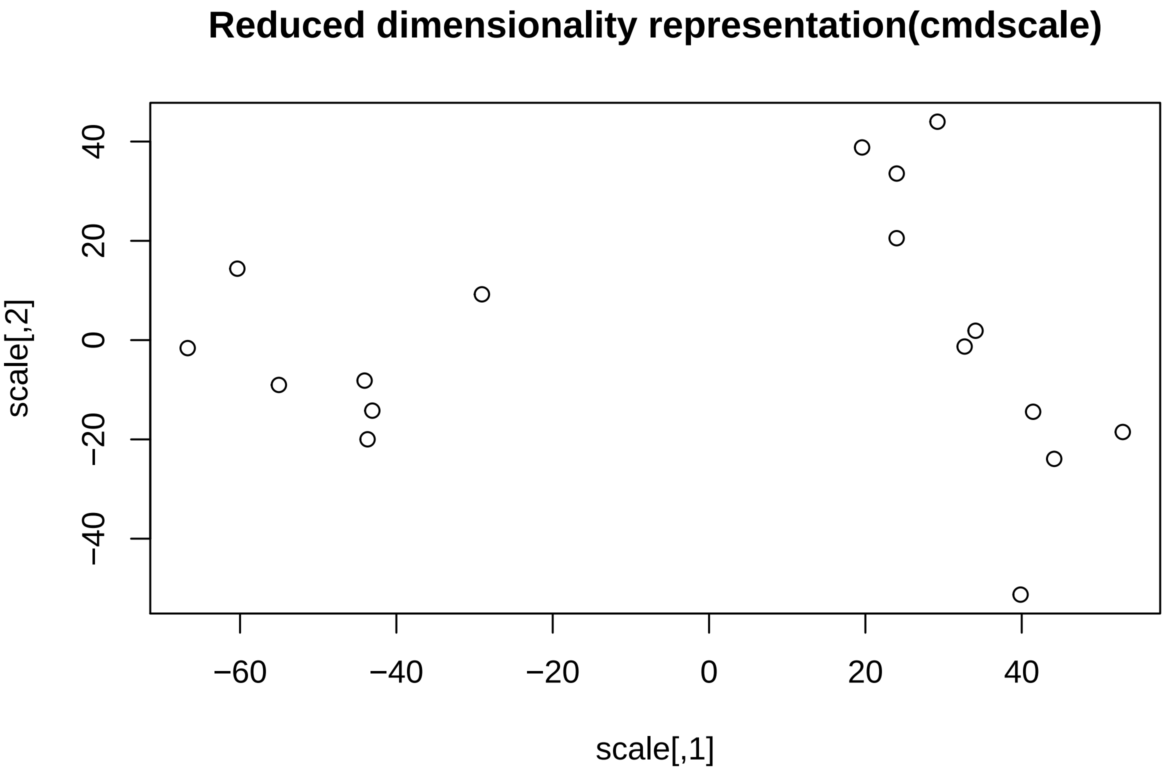
```
corr <- cor(norm_data)
image(corr,main="Correlation plot")
```



Reduced dimensionality representation

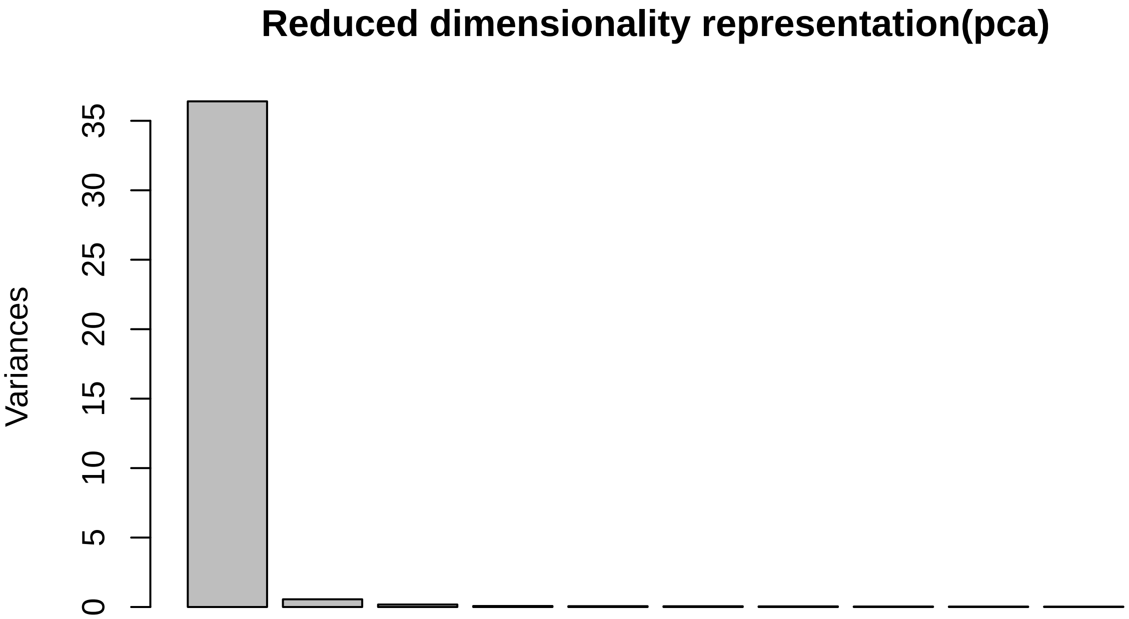
Classical Multidimensional Scaling

```
scale<-cmdscale(d1_data,k=2)
plot(scale,main="Reduced dimensionality representation(cmdscale)")
```



Principal Components Analysis

```
pca<-prcomp(norm_data)
plot(pca,main="Reduced dimensionality representation(pca)")
```


Reduced dimensionality representation(pca)
Variances