Exercise 2

Hubert Rehrauer, modified by Cyril Statzer

25 9 2017

Exploratory Data Analysis

Do an exploratory data analysis of a matrix of expression values. Load the data and display:

```
#install.packages("limma")
library(limma)
#install.packages("pheatmap")
library(pheatmap)
```

Data Import

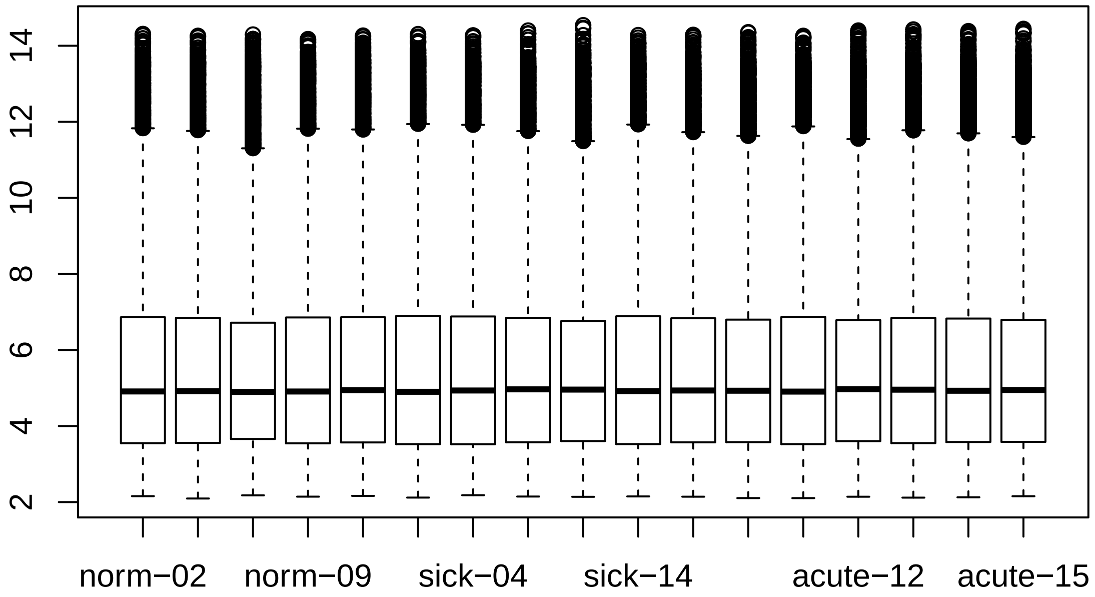
```
anno = read.table("SampleAnnotation.txt", as.is=TRUE, sep="\t", quote="",
                  row.names=1, header=TRUE)
x = read.table("expressiondata.txt", as.is=TRUE, sep="\t", quote="", row.names=1, header=TRUE, check.nam
x = log2(as.matrix(x))
```

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"
```

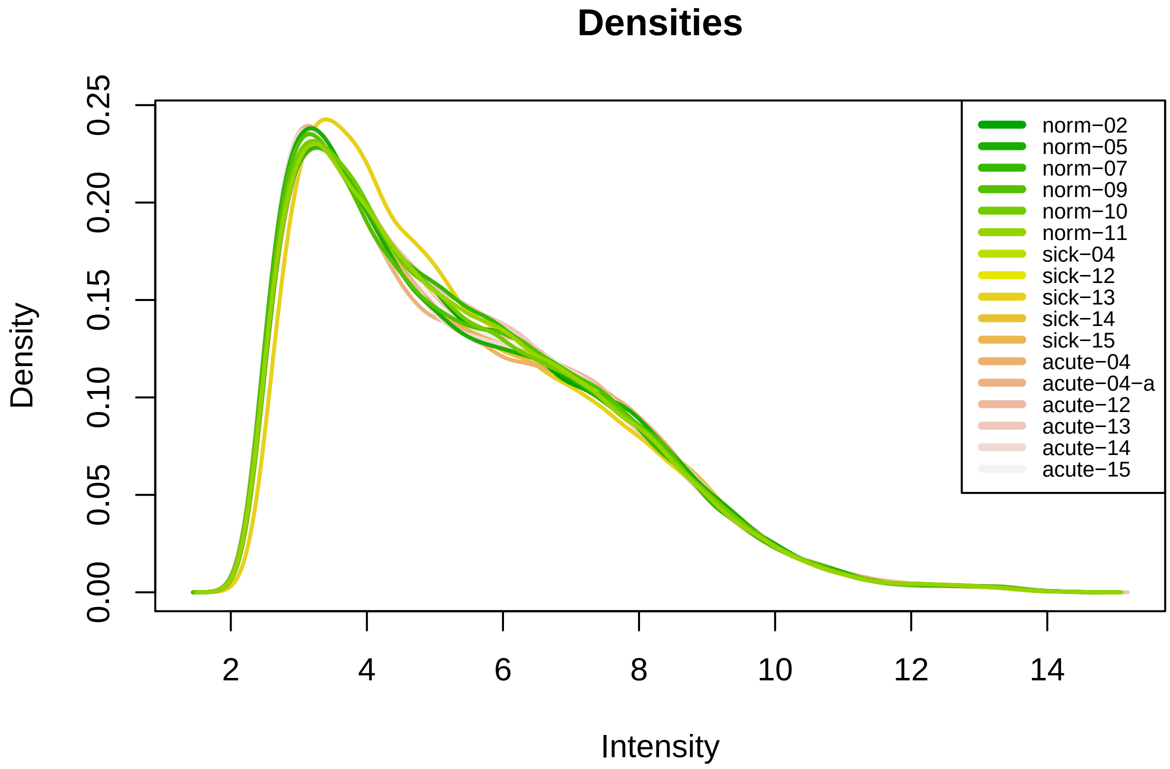
Distribution analysis *boxplot*

```
boxplot(x,use.cols=1)
```



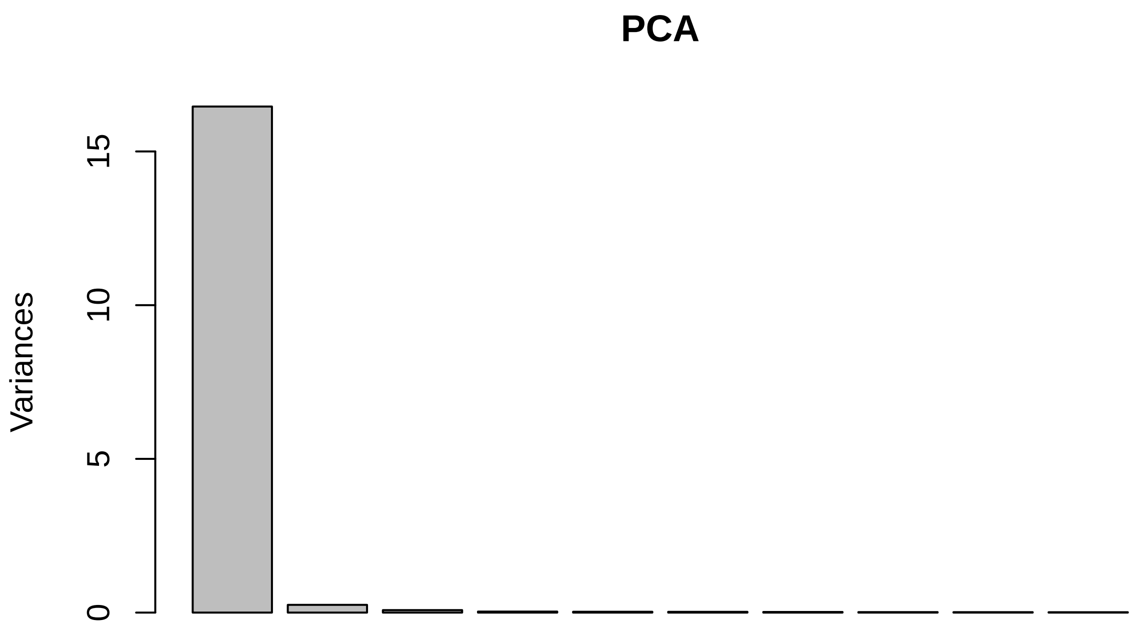
density

```r
limma::plotDensities(x,main = "Densities",legend=F,col = terrain.colors(nrow(anno)))
legend("topright",legend = colnames(x),cex = 0.7,col=terrain.colors(nrow(anno)),lty = 1, lwd = 4, y.inte
```



Principle component analysis

```r
pca <- prcomp(x, center = T, scale. = T)
plot(pca, main = "PCA")
```



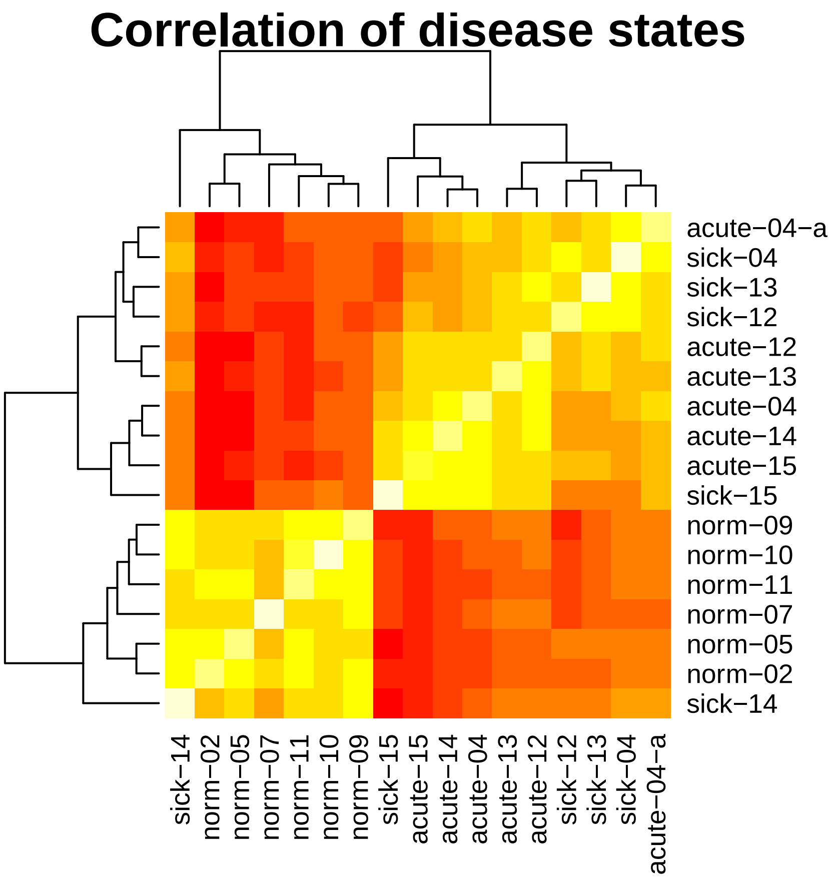
```r
#install.packages("ggbiplot")
#library(ggbiplot)
#g <- ggbiplot(pca)
```

Build correlation matrix from expression matrix (corr(x)) Normalization is performed using the min-max methond

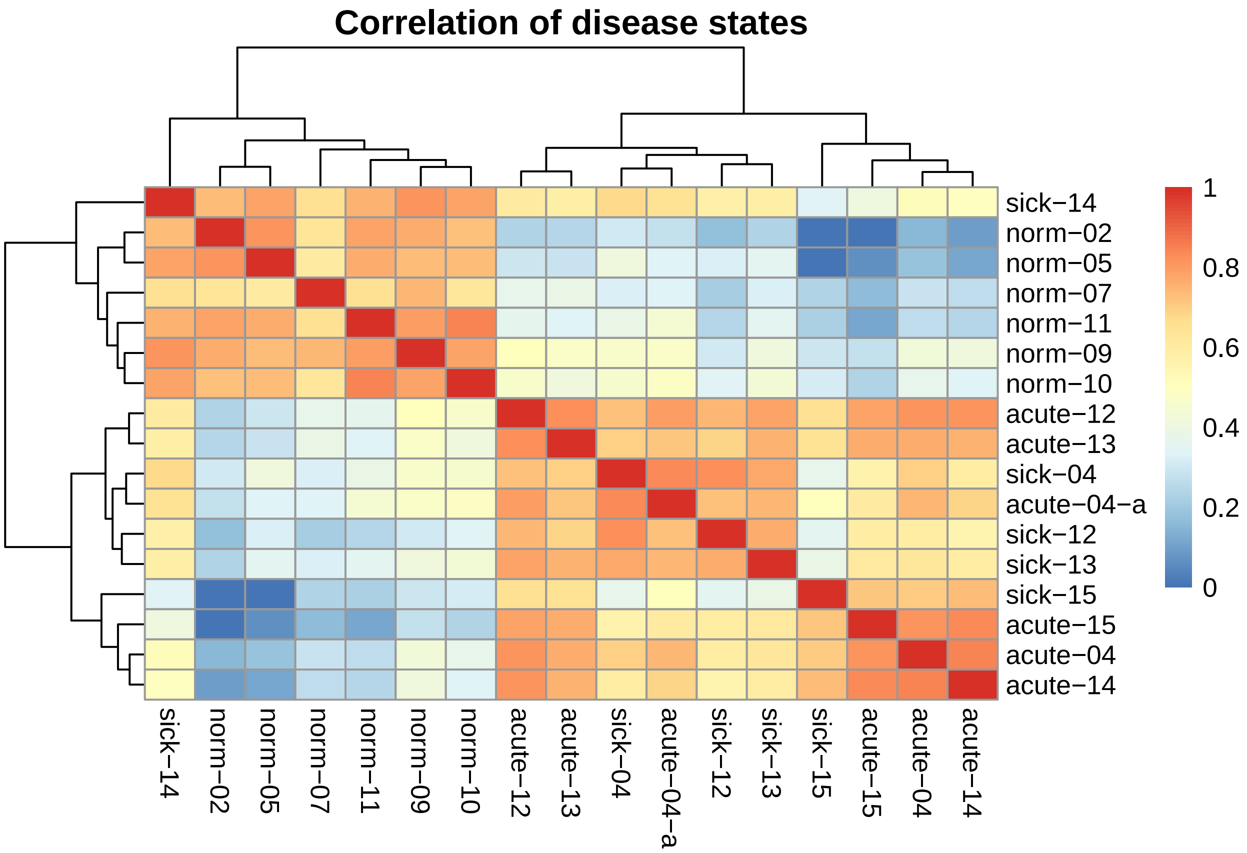
```r
corr <- cor(x)
corr <- (corr - min(corr))/(max(corr) - min(corr))
#corr <- normalizeQuantiles(corr)
```

Generate a heatmap to analyze clustering of samples Both the heatmap and pheatmap functions were used. The general clustering is very good, indicating a strong difference between normal and other samples (acute and disease). Within the acute and disease groups the clustering is very weak and they cannot be separated.

```r
heatmap(corr,main = "Correlation of disease states")
```



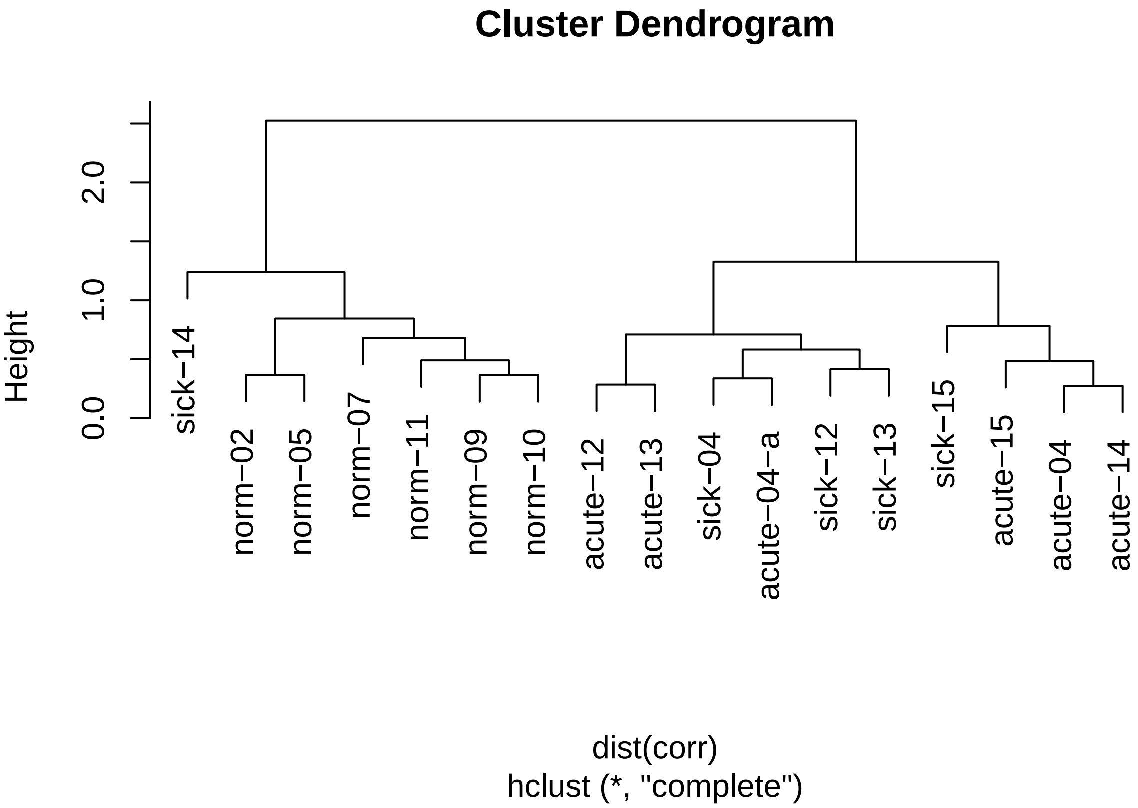
```
pheatmap(corr,main = "Correlation of disease states")
```



Similarly to the heatmap here the clustering between patients is visualized in a separate dendrogram. The sick-14 sample is in the other cluster. All other acute and sick samples cluster within the same cluster.

- clustering: *hclust*

```
hc <- hclust(dist(corr))
plot(hc)
```


Cluster Dendrogram
Height
0.0
1.0
2.0
sick−14
norm−02
norm−05
norm−07
norm−11
norm−09
norm−10
acute−12
acute−13
sick−04
acute−04−a
sick−12
sick−13
sick−15
acute−15
acute−04
acute−14
dist(corr)
hclust (*, "complete")