Exercise 2

Franca Eichenberger

02.10.2016

1 Introduction

> In a particular liver disease, parts of the liver are acutely affected while other parts show no severe signs at all. In this study we analyze the gene expression levels of 5 patients with this liver disease and try to understand the molecular causes and consequences of it. The gene expression from moderately and acute affected tissue of these 5 patients has been measured. As a reference normal tissue of six healthy patients was included.

In this exercise first exploratory data anylsis is undertaken in order to identify outliers or systematic biases.

2 Loading the data

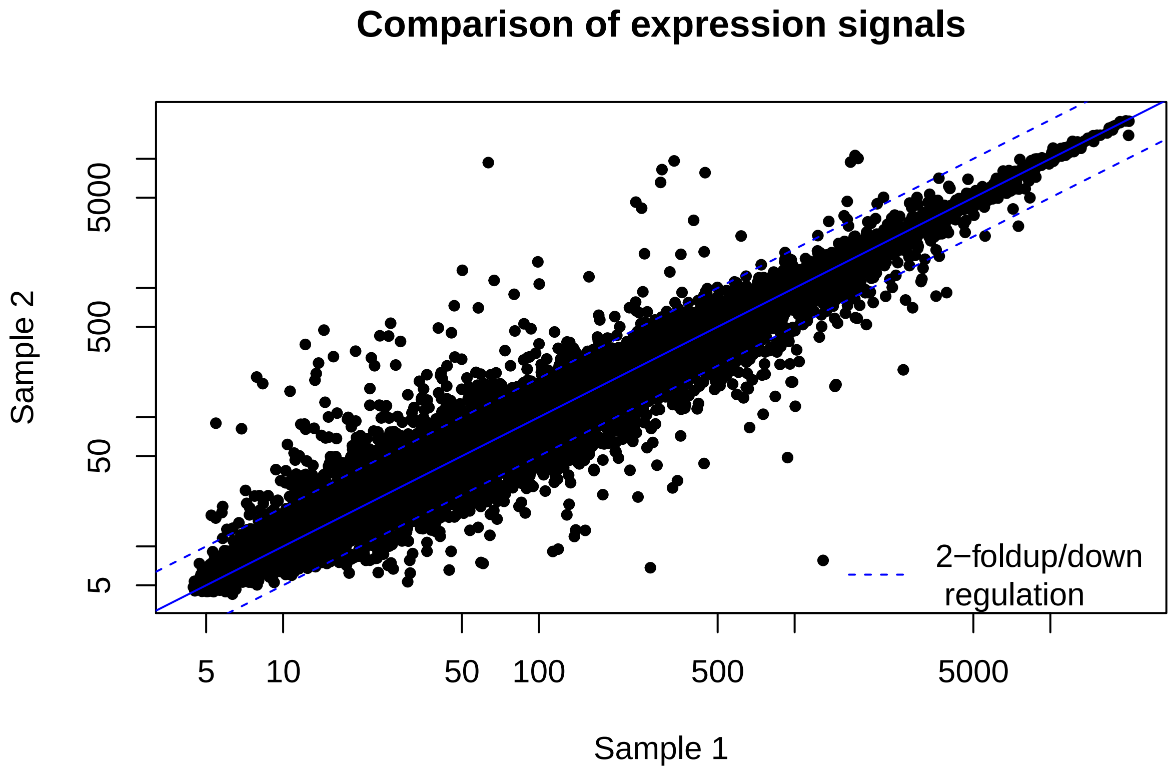
Pheontype information data

```
anno = read.table("SampleAnnotation.txt", as.is=TRUE, sep="\t", quote="",
                    row.names=1, header=TRUE)
samples = rownames(anno)
colors = rainbow(nrow(anno))
isNorm = anno$TissueType == "norm"
isSick = anno$TissueType == "sick"
isAcute = anno$TissueType == "acute"
```

Expression data

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

Now we compare the expression signal from sample 1 and 2



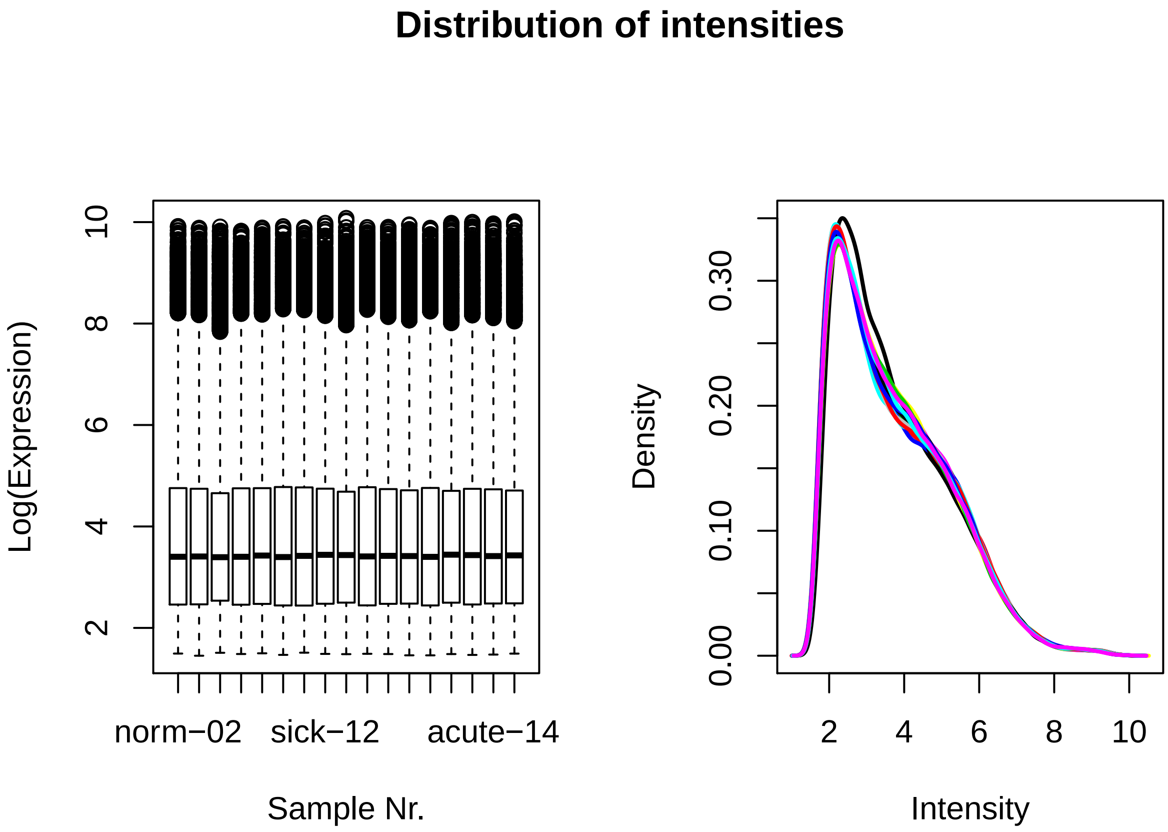
3 Checking the distribution of intensities

A basic assumption for the analysis is, that the intensity distribution of the different arrays are similar. It is therefore important to check whether these assumption is met.

```
source("http://bioconductor.org/biocLite.R")
biocLite("limma")
```

```
##
## The downloaded binary packages are in
##  /var/folders/mg/xypky2dj3cg27y2z9k4hyv9h0000gn/T//RtmpZzeWXr/downloaded_packages
```

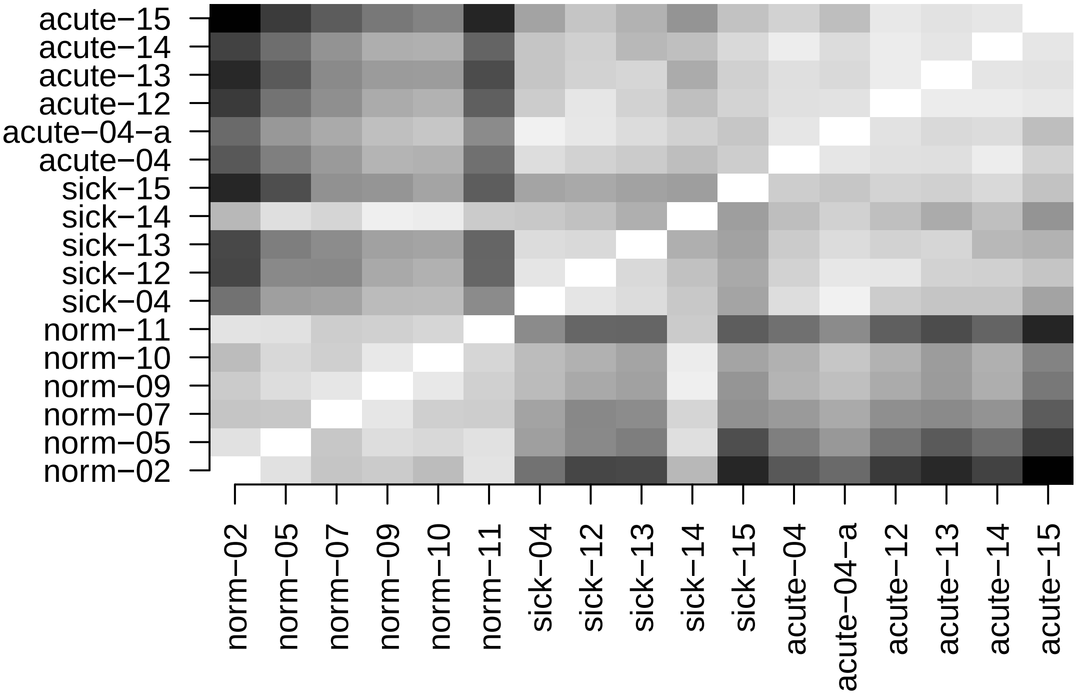
```
require("limma")
par(mfrow=c(1,2), oma=c(0,0,1,0))
boxplot(log(x),ylab="Log(Expression)",xlab="Sample Nr.")
plotDensities(log(x),legend=F)
title("Distribution of intensities", cex=1, outer=T)
```



4 Checking the consistency of the replicates

The consistency of the replicates is checked by computing sample correlation matrices.

```r
corrMatrix = cor(x)
par(mar=c(8,8,2,2))
grayScale <- gray((1:256)/256)
image(corrMatrix, col=grayScale,  axes=FALSE)
axis(1, at=seq(from=0, to=1, length.out=length(samples)), labels=samples, las=2)
axis(2, at=seq(from=0, to=1, length.out=length(samples)), labels=samples, las=2)
```



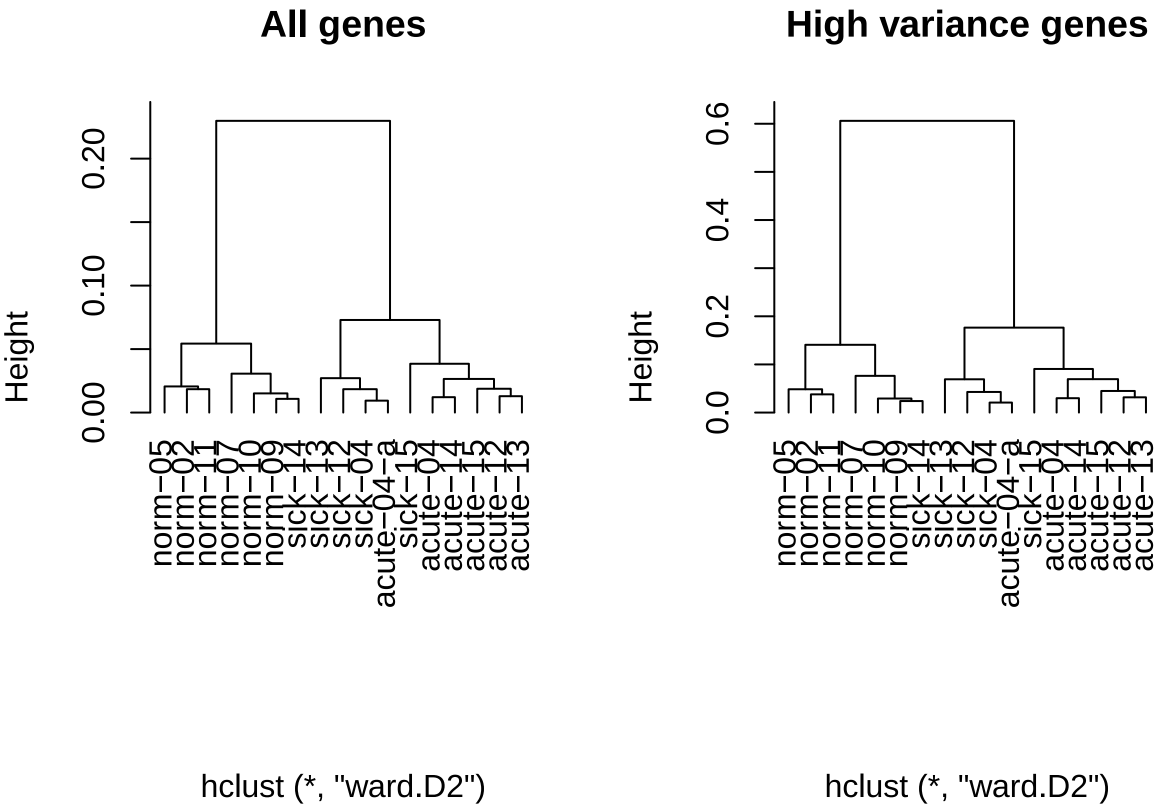
1. From the correlation we can see that:

- normals show high correlation among each other
- the normal are very different from both sick and acute
- the sick and acute are rather similar
- sick-14 rather looks like a normal sample
- sick-15 has overall low correlation but rather looks like an ???acute???
- acute-04-a which is a technical replicate of acute-04 is more similar to sick-04 than to acute-04.

5 Sample Clustering

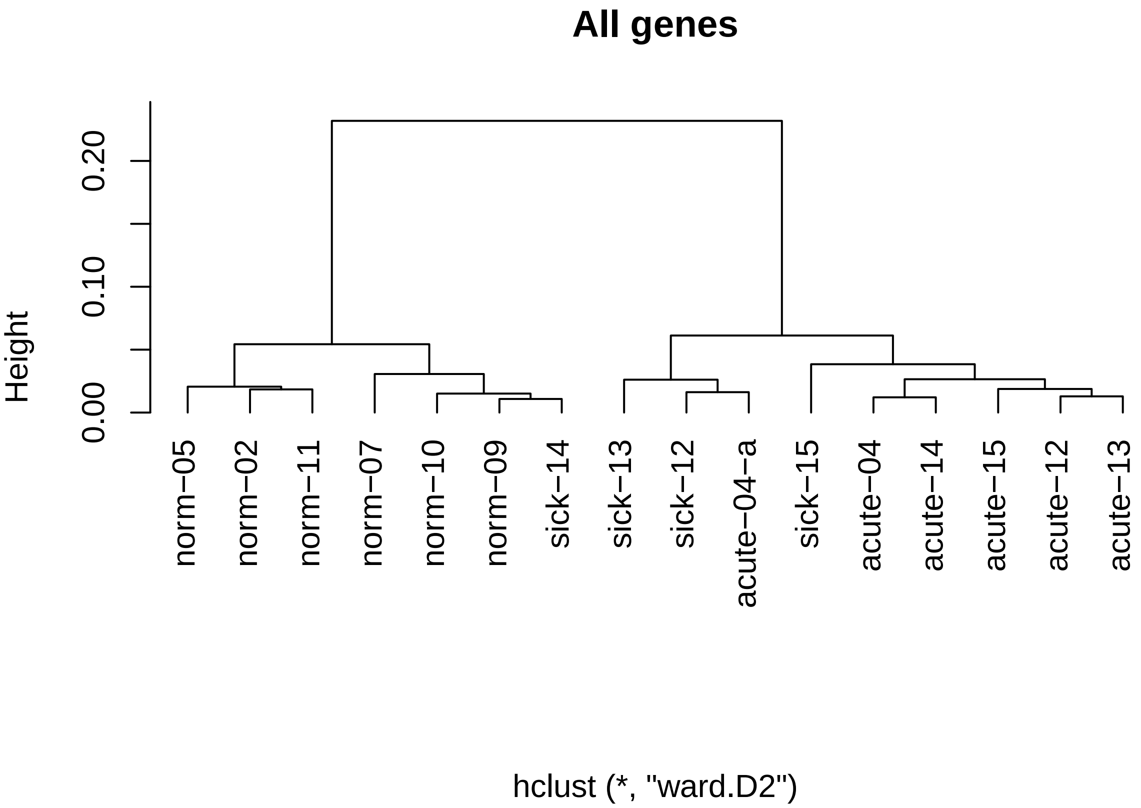
The sample clustering shows the similarities of the expression patterns of the samples in a tree.

```
x.sd = apply(x, 1, sd, na.rm=TRUE)
ord = order(x.sd, decreasing=TRUE)
highVarGenes = ord[1:500]
par(mfrow=c(1,2));
d = as.dist(1-cor(x));
c=hclust(d, method="ward.D2");
plot(c, hang=-0.1, main="All genes", xlab="")
d = as.dist(1-cor(x[highVarGenes, ]));
c=hclust(d, method="ward.D2");
plot(c, hang=-0.1, main="High variance genes", xlab="")
```



Excluding sick-04:

```
sub = x[ , samples != "sick-04"]
d = as.dist(1-cor(sub));
c=hclust(d, method="ward.D2");
plot(c, hang=-0.1, main="All genes", xlab="")
```

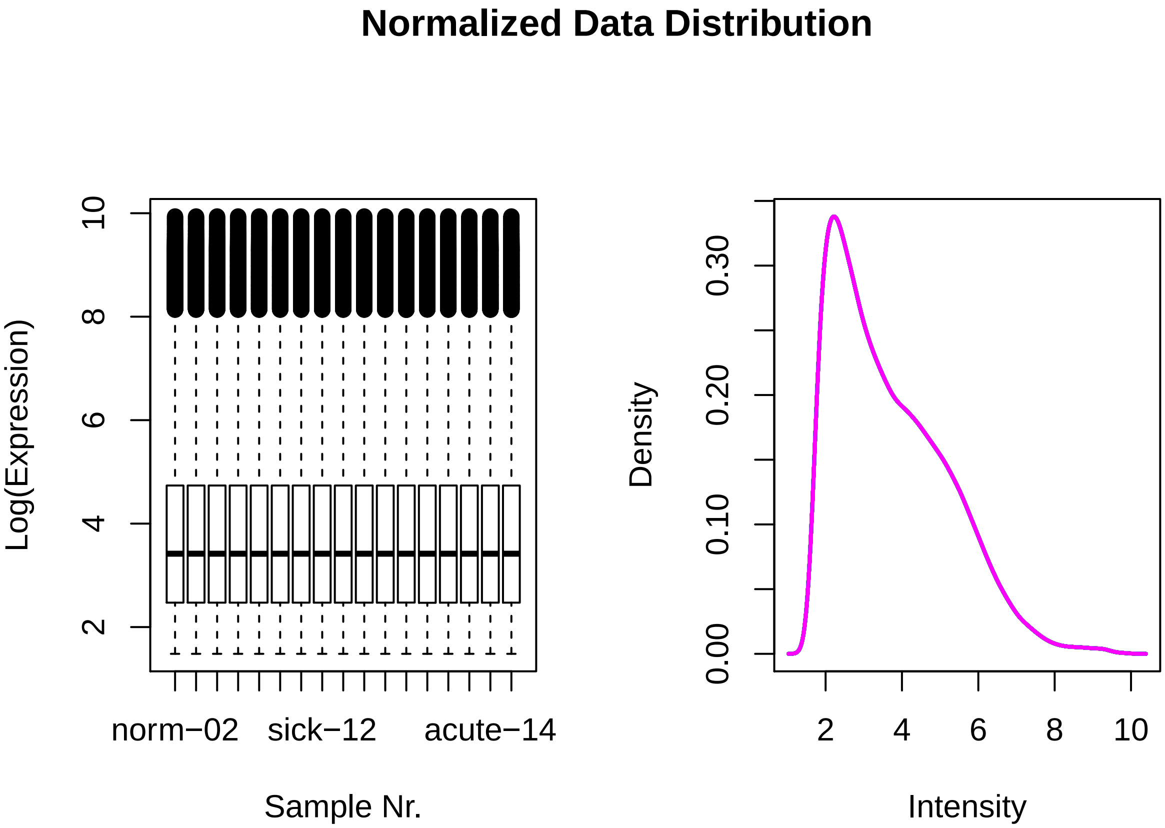


If we run the clustering without sample sick-04, the acute-04 does no longer cluster in the branch with the other sick samples.

6 Apply quantile normalization

Normalization is needed whenever the data does not fullfill the assumptions required by a certain analysis.

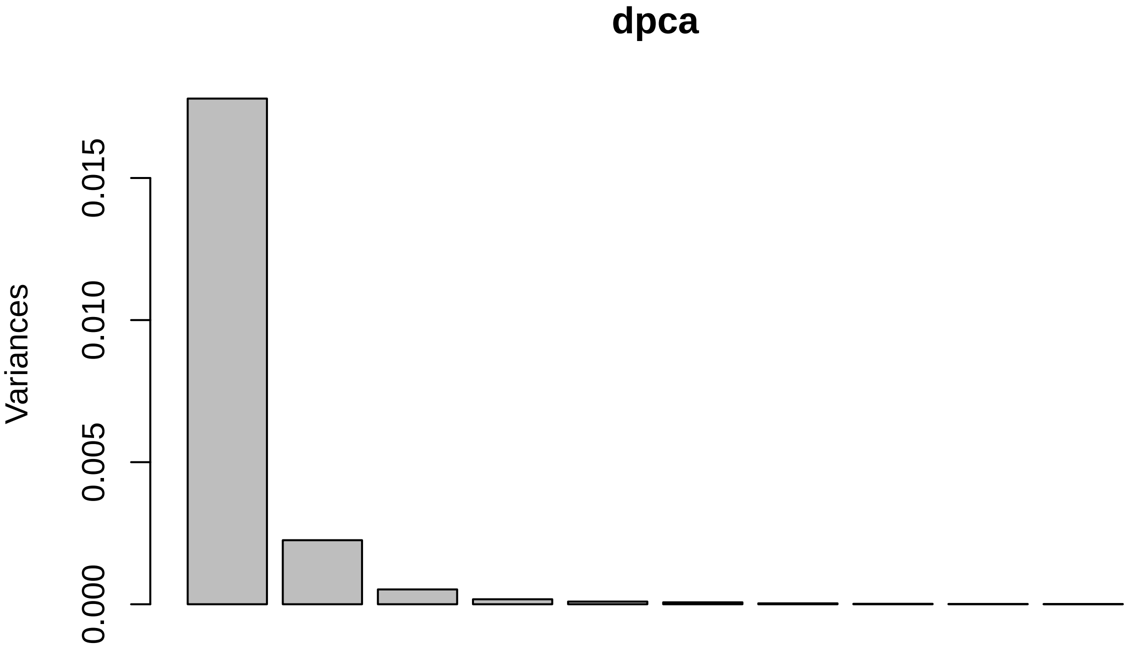
```
par(mfrow=c(1,2),oma=c(0,0,1,0))
boxplot(log(normalizeQuantiles(x)),ylab="Log(Expression)",xlab="Sample Nr.")
plotDensities(log(normalizeQuantiles(x)),legend=FALSE)
title("Normalized Data Distribution",cex=1,outer=T)
```



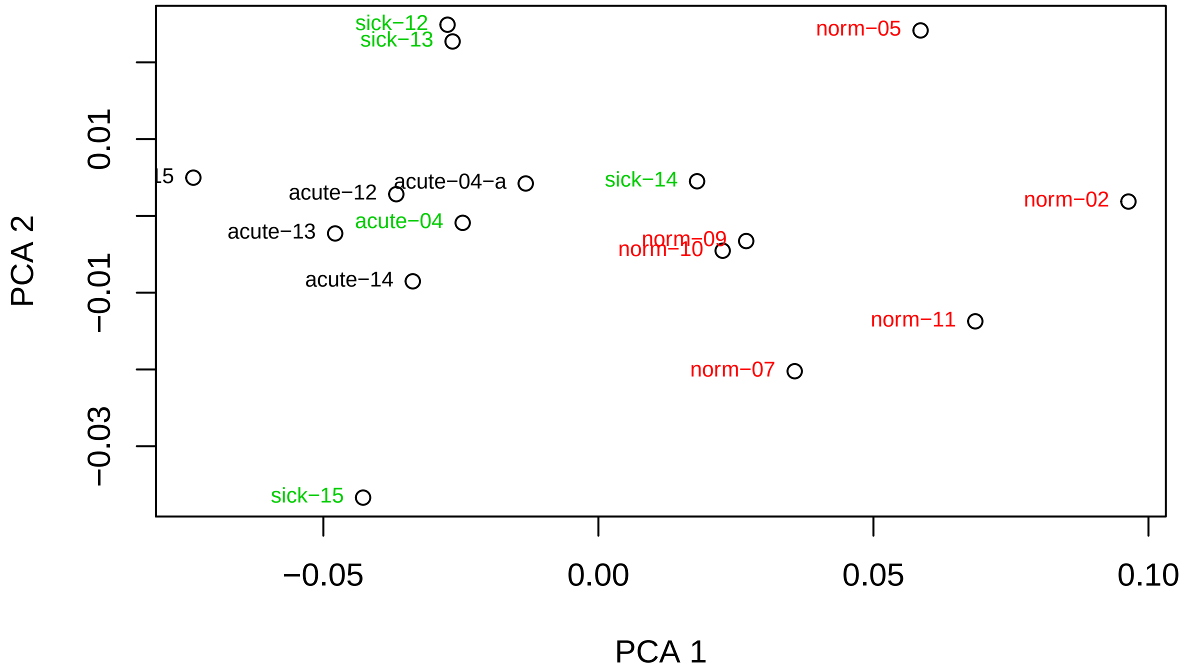
##7

Sample Representation in Principal Component Space We create a plot that represents the sample distances in a reduced space.

```
dpca<-prcomp(d)
plot(dpca)
```



```
pca<-cmdscale(d,k=2)
plot(pca,xlab="PCA 1",ylab="PCA 2")
text(pca,rownames(pca),cex=0.7,pos=2,col=as.numeric(as.factor(anno$TissueType)))
```


sick−12
sick−13
norm−05
acute−12
acute−04−a
sick−14
norm−02
acute−13
acute−04
norm−09
norm−10
acute−14
norm−11
norm−07
sick−15
PCA 1
PCA 2