Exercise 2016-09-26

Hubert Rehrauer

September 23, 2016

1 Introduction

A study examines a liver disease. In affected patients the liver does not work properly. Part of the liver shows acute signs other parts are affected but do not show severe symptoms. The study searches for the genes that have changed expression due to the disease. The goal is to understand at a molecular level the causes and consequences of the disease.

The study has used Affymetrix Whole Genome Microarrays of the type HG-U133-Plus2 to measure the gene expression. In total 5 sick patients were examined and from each patient a sample from acutely and moderately tissue was measured. As a reference normal tissue from 6 healthy patients were examined.

In this exercise we run some exploratory analysis of the data set. We want to check if there are outliers and/or systematic biases.

2 Loading the data

The phenotype information is in the file `SampleAnnotation.txt` that we read into a data frame,

```
>   anno = read.table("SampleAnnotation.txt", as.is=TRUE, sep="\t", quote="",
+                     row.names=1, header=TRUE)
```

that holds the sample name, tissue type, patient ID, and associated file:

```
>   anno
```

```
           TissueType PatientId          File
norm-02          norm       P02    norm-02.CEL
norm-05          norm       P05    norm-05.CEL
norm-07          norm       P07    norm-07.CEL
norm-09          norm       P09    norm-09.CEL
norm-10          norm       P10    norm-10.CEL
norm-11          norm       P11    norm-11.CEL
sick-04          sick       P04    sick-04.CEL
sick-12          sick       P12    sick-12.CEL
sick-13          sick       P13    sick-13.CEL
sick-14          sick       P14    sick-14.CEL
sick-15          sick       P15    sick-15.CEL
acute-04        acute       P04   acute-04.CEL
acute-04-a      acute       P04 acute-04-a.CEL
acute-12        acute       P12   acute-12.CEL
acute-13        acute       P13   acute-13.CEL
acute-14        acute       P14   acute-14.CEL
acute-15        acute       P15   acute-15.CEL
```

For labeling and coloring the subsequent plots we define the variables:

```
>   samples = rownames(anno)
> colors = rainbow(nrow(anno))
```

and generate boolean indices with which we can access the normal, sick and acute samples only:

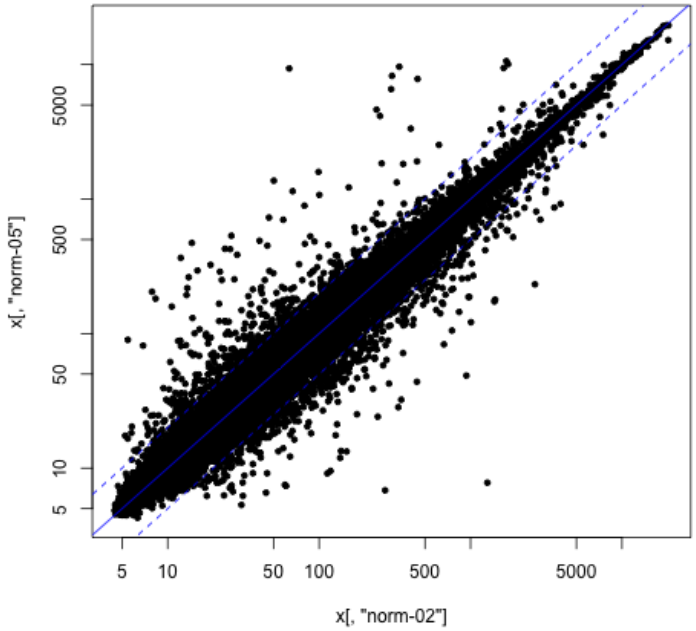
```
>   isNorm = anno$TissueType == "norm"
> isSick = anno$TissueType == "sick"
> isAcute = anno$TissueType == "acute"
```

Now we load the expression data

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

We visually compare the expression signals from sample 1 and 2 by plotting them with the standard plot command:

```
> plot(x[ , "norm-02"], x[, "norm-05"], log="xy", pch=20)
> abline(0, 1, col="blue")
> abline(log10(2), 1, col="blue", lty=2);
> abline(-log10(2), 1, col="blue", lty=2);
```



The solid blue line gives the first diagonal and the dashed lines give the boundaries for 2-fold up- or down-regulation.

3 Checking the distribution of the intensities

A basic assumption for the subsequent analysis is, that the intensity distributions of the different arrays are similar. Do create boxplots that summarize the value distribution for each sample, and use the function plotDensities in the package limma.

4 Checking the consistency of the replicates

Do check the consistency of the replicates by computing sample correlation.

In order to check whether the replicates correlate well, we compute the correlation matrix on the logarithmic values and print it

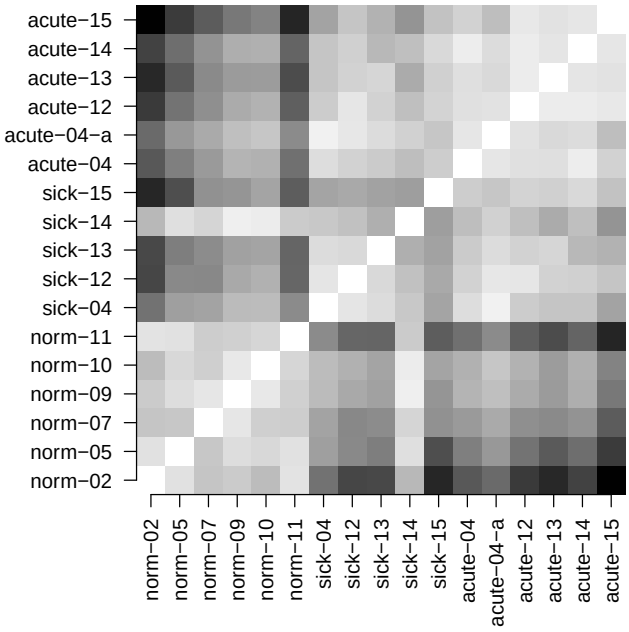
```
> corrMatrix = cor(x)
> signif(corrMatrix, digits=3)
```

	norm-02	norm-05	norm-07	norm-09	norm-10	norm-11	sick-04	sick-12	sick-13	sick-14
norm-02	1.000	0.980	0.962	0.965	0.956	0.982	0.907	0.878	0.879	0.953
norm-05	0.980	1.000	0.963	0.977	0.974	0.980	0.937	0.923	0.915	0.979
norm-07	0.962	0.963	1.000	0.984	0.968	0.967	0.940	0.922	0.925	0.972
norm-09	0.965	0.977	0.984	1.000	0.985	0.969	0.955	0.943	0.938	0.989
norm-10	0.956	0.974	0.968	0.985	1.000	0.973	0.956	0.949	0.940	0.987
norm-11	0.982	0.980	0.967	0.969	0.973	1.000	0.924	0.899	0.898	0.966

```
sick-04      0.907   0.937   0.940   0.955   0.956   0.924   1.000   0.983   0.977   0.964
sick-12      0.878   0.923   0.922   0.943   0.949   0.899   0.983   1.000   0.975   0.959
sick-13      0.879   0.915   0.925   0.938   0.940   0.898   0.977   0.975   1.000   0.948
sick-14      0.953   0.979   0.972   0.989   0.987   0.966   0.964   0.959   0.948   1.000
sick-15      0.857   0.883   0.928   0.931   0.940   0.893   0.940   0.944   0.939   0.936
acute-04     0.890   0.915   0.934   0.951   0.949   0.906   0.977   0.970   0.966   0.957
acute-04-a   0.901   0.933   0.944   0.958   0.963   0.924   0.991   0.984   0.977   0.970
acute-12     0.870   0.908   0.926   0.945   0.950   0.895   0.966   0.983   0.970   0.958
acute-13     0.858   0.891   0.923   0.934   0.935   0.882   0.962   0.970   0.973   0.945
acute-14     0.876   0.904   0.929   0.947   0.948   0.898   0.962   0.969   0.953   0.958
acute-15     0.833   0.871   0.892   0.911   0.919   0.857   0.940   0.962   0.949   0.930
           acute-15
norm-02       0.833
norm-05       0.871
norm-07       0.892
norm-09       0.911
norm-10       0.919
norm-11       0.857
sick-04       0.940
sick-12       0.962
sick-13       0.949
sick-14       0.930
sick-15       0.960
acute-04      0.970
acute-04-a    0.957
acute-12      0.985
acute-13      0.981
acute-14      0.983
acute-15      1.000
```

We visualize the matrix as an image:

```
>   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)
```



From the correlation we can see that

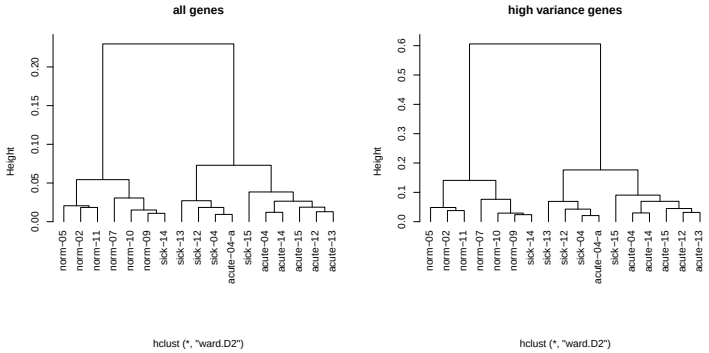
- normals show high correlation among each other
- the normal are very different from both sick and acute
- the sick and acute are rather similar
- `sick-14` rather looks like a normal sample
- `sick-15` has overall low correlation but rather looks like an "acute"
- `acute-04-a` which is a technical 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. In order to compute the similarities of the samples one can use all genes or only a subset of the genes. When using all genes where is the risk that the absent genes drive the clustering of the samples. This is because in many studies the absent genes make up the majority of the measured genes. But those genes have a low intensity that is strongly influenced by the background signal measured on each chip and not by real gene expression

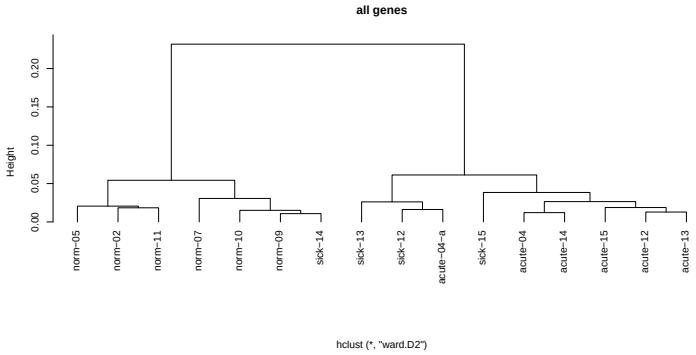
```
>   x.sd = apply(x, 1, sd, na.rm=TRUE)
> ord = order(x.sd, decreasing=TRUE)
> highVarGenes = ord[1:500]
```

```
> 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="")
```



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

```
>   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="")
```



6 Apply quantile normalization

Use the function `limma::normalizeQuantiles` to normalize the data and plot again the histograms.

7 Sample Representation in Principal Component Space

Use the functions `cmdscale` and `prcomp` to create a plot that represents the sample distances in a reduced space.