An Example of Differential miRNA Expression Analysis Using R and Bioconductor Packages

Ricardo Gonzalo and Alex Sànchez
Unitat d'Estadística i Bioinformàtica
Vall d'Hebron Institut de Recerca (VHIR)

August 7, 2017

1 Introduction and Study Organization

This document aims to explain how to analyze data from miRNA 4.1 arrays from Affymetrix manufacturer.

The analysis has been performed following the usual "pipeline" for microarray data analysis. Roughly speaking each of the items below corresponds to one section in the analysis report:

1. Quality control: Do we have good data? Should any slides be rejected?
2. Data preprocess: Summarization, filtering and normalization.
3. Selection of genes differentially expressed for each set of conditions.

The statistic analysis has been performed using the statistical language "R"(R version 3.4.1 (2017-06-30), Copyright (C) 2015 The R Foundation for Statistical Computing), and the libraries developed for the microarray analysis in the Bioconductor Project (www.bioconductor.org). More details about the methods used in these analysis could be found in .

First of all it is necessary to set the folders where the data is and where the results will be saved:

```
mainDir <-getwd()
workingDir <- mainDir
dataDir <-file.path(mainDir, "data")
```

```
resultsDir <- file.path(workingDir, "results")
imagesDir<-file.path(mainDir,"images")
```

We have to load the necessary pacakges for this analysis:

```
library(xtable)
library(Biobase)
library(oligo)
library(arrayQualityMetrics)
library(devtools)
library(ggplot2)
library(ggrepel)
library(pd.mirna.4.1)
```

It is important to remark that nowadays (at least on 07/08/2017) no package (platform design) for miRNA_4.1 arrays exists on Bioconductor web page. Therefore, it is necessary to build your own package following the instructions kindly posted at https://support.bioconductor.org/p/96882/. Briefly, it could something like:

```
pgfFile <- "miRNA-4_1-st-v1.pgf"
clfFile <- "miRNA-4_1.clf"
csvAnnoFile=csvAnno

pkg <- new("AffyMiRNAPDInfoPkgSeed",
             version="0.1",
             author="RGS", email="ricardo.gonzalo@vhir.org",
             biocViews="AnnotationData",
             genomebuild="NCBI Build 35, May 2004",
             pgfFile=pgfFile, clfFile=clfFile)
makePdInfoPackage(pkg, destDir=".")

library(devtools)
build("pd.mirna.4.1")

install("pd.mirna.4.1")
```

2 Data for the analysis

Data for the analysis has been obtained from https://www.thermofisher.com/order/catalog/product/902410, and has been summarized in the `targets file`:

```
targets <-read.table(file.path(dataDir,"targets.csv"), header = TRUE, sep = ";")
x.big<-xtable(targets,caption="Targets file showing samples and covariates")
print(x.big,tabular.environment=longtable,floating=FALSE,size="small")
```

	FileName	Group	ShortName	Color
1	Human-Brain-AM7962-130ng_rep1_(miRNA-4_1-Array-Plate).CEL	Brain	Brain1	blue
2	Human-Brain-AM7962-130ng_rep2_(miRNA-4_1-Array-Plate).CEL	Brain	Brain2	blue
3	Human-Brain-AM7962-130ng_rep3_(miRNA-4_1-Array-Plate).CEL	Brain	Brain3	blue
4	Human-Brain-AM7962-130ng_rep4_(miRNA-4_1-Array-Plate).CEL	Brain	Brain4	blue
5	Human-Lung-AM7968-130ng_rep1_(miRNA-4_1-Array-Plate).CEL	Lung	Lung1	green
6	Human-Lung-AM7968-130ng_rep2_(miRNA-4_1-Array-Plate).CEL	Lung	Lung2	green
7	Human-Lung-AM7968-130ng_rep3_(miRNA-4_1-Array-Plate).CEL	Lung	Lung3	green
8	Human-Lung-AM7968-130ng_rep4_(miRNA-4_1-Array-Plate).CEL	Lung	Lung4	green

Table 1: Targets file showing samples and covariates

3 Results

3.1 Quality Control

Different types of quality controls (QC) have been performed in this analysis. It is important to do this QC before and after array normalization to check if all the arrays are suitable for normalization process, and check if normalized data is appropriated for differential expression analysis. Besides, a comprehensive report and some more figures of quality control, are provided for the raw data `(QCDir.raw/index.html)` and for the normalized data `(QCDir.norm/index.html)` to help the user to understand whether a particular array can be considered as an outlier.

First of all it is necessary to read the CEL files of the experiment:

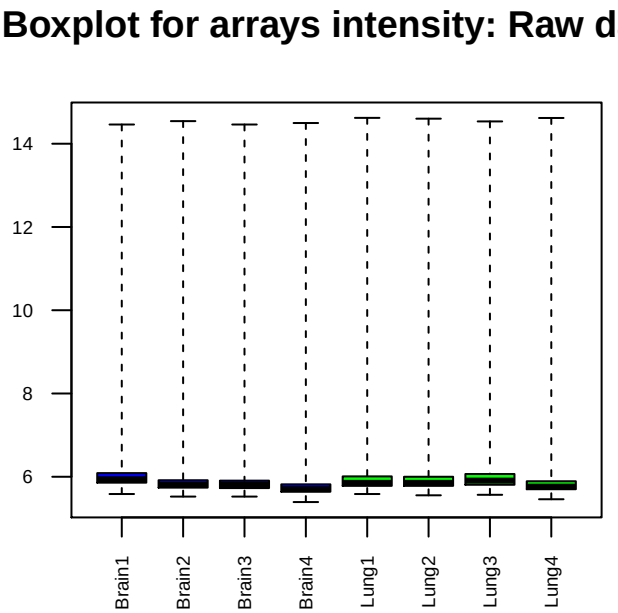
```
celFiles<-list.celfiles(dataDir,full.names=TRUE)
rawData<-read.celfiles(celFiles)
```

Some settings for the plots are useful:

```
colores <- as.character(targets$Color)
grupos <- targets$Group
sampleNames <-targets$ShortName
```

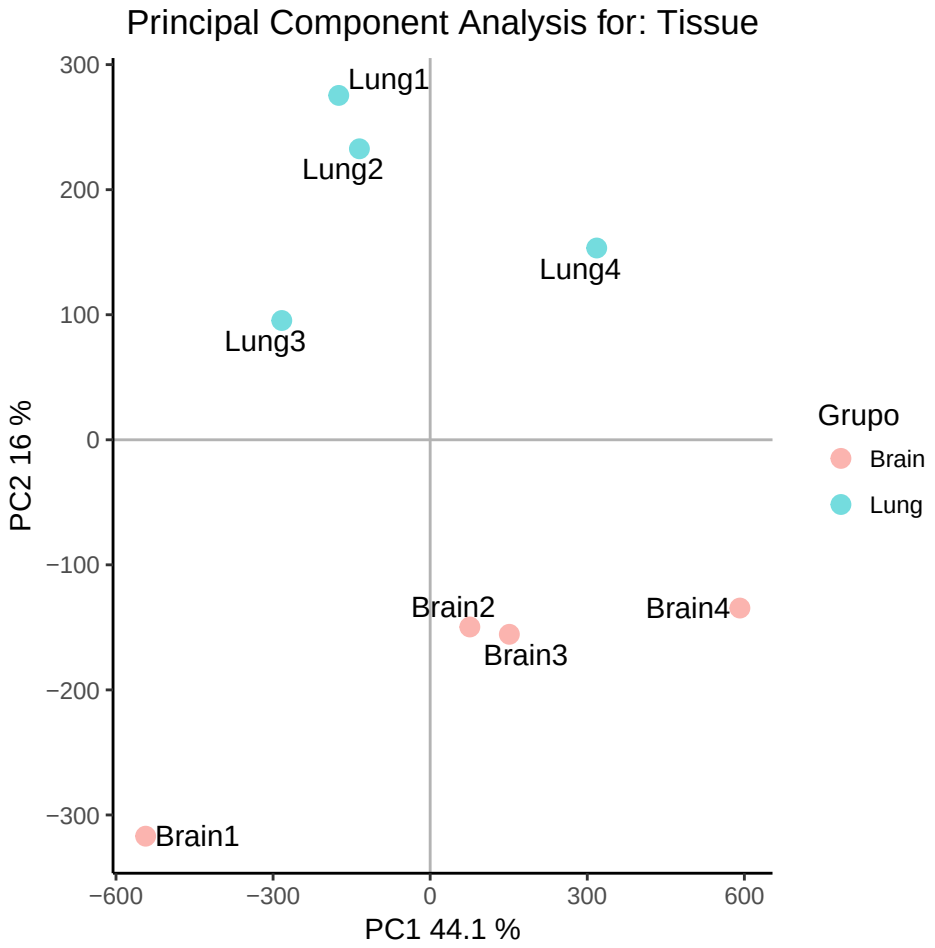
Next figure shows a boxplot representing summaries of the signal intensity distributions of the arrays. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate a problem.

```
boxplot(rawData, cex.axis=0.6, col=colores,  las=2, names=sampleNames, main="Boxplot for arrays in
```



Next figure shows a scatterplot of the arrays along the first two principal components. Principal components Analysis (PCA) is a dimension reduction technique that may be used to represent the values of an expression matrix in two (or three) dimensions. The plot is constructed in such a way that "similar" arrays should appear together in the plot, so that if a sample appears near others that are not considered similar it can be suspected the presence of some kind of technical problems such as batch effects, mislabelling of samples, etc.

```
source("https://raw.githubusercontent.com/uebvhir/UEB_PCA/master/UEB_PCA.R")
plotPCA2(exprs(rawData),labels =sampleNames, factor=grupos,title="Tissue",scale = TRUE )
```



Once all the analyses have been performed the different outlier–detection approaches are compared in order to decide whether or not an array should be removed. Usually only arrays that are called outliers by more than one criteria are considered to rule out, although this depends on every specific study.

3.2 Normalization

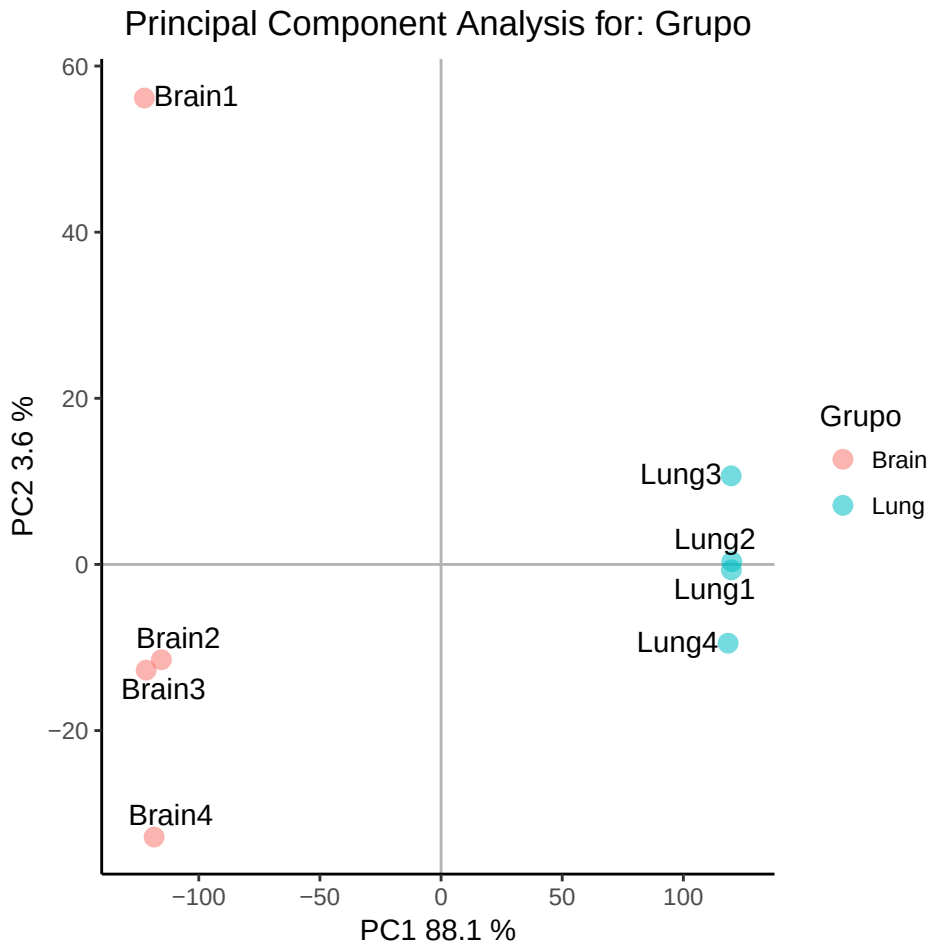
In order to make the data comparable as well as to remove technical biases the arrays have been preprocessed using the RMA method (Irizarry, 2003 and Gentleman, 2005).

```
eset_rma <- rma(rawData)

## Background correcting
## Normalizing
## Calculating Expression
```

The PCA is performed again with the normalized data.

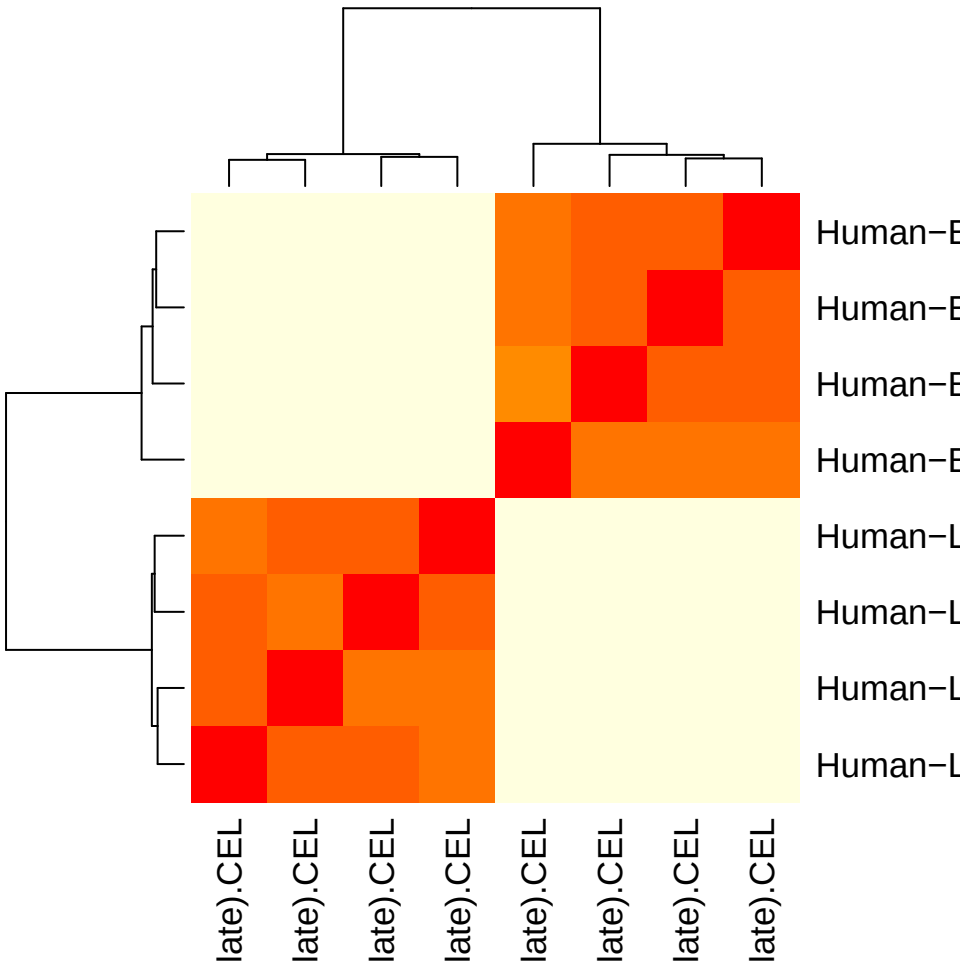
```
plotPCA2(exprs(eset_rma),labels =sampleNames, factor=grupos,title="Grupo",scale = FALSE )
```



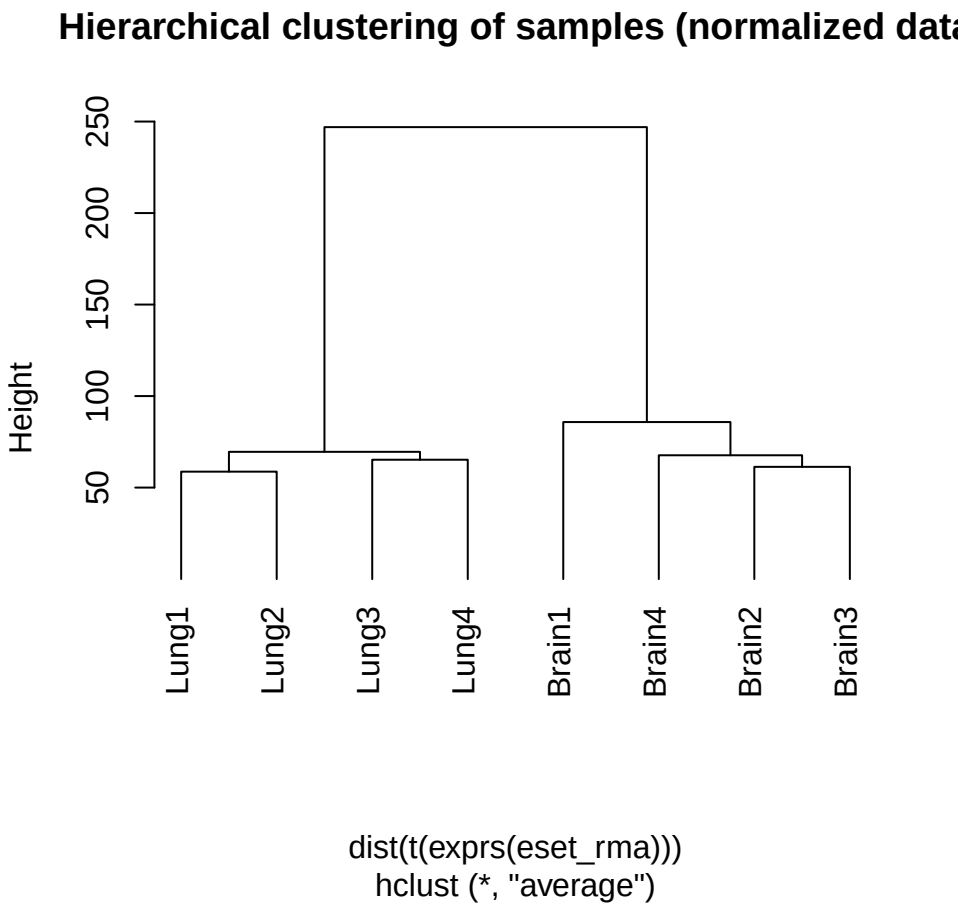
Some other Quality checks can be performed:

```
manDist <-  dist(t(exprs(eset_rma)))
clust.euclid.average <- hclust(dist(t(exprs(eset_rma))),method="average")

heatmap (as.matrix(manDist),  col=heat.colors(16))
```




```
plot(clust.euclid.average, labels=sampleNames,
     main="Hierarchical clustering of samples (normalized data",  hang=-1)
```




```
arrayQualityMetrics(eset_rma, outdir = file.path(resultsDir, "QCDir.Norm"),
                    force=TRUE)
```

It could be useful to save the normalized data in a csv file:

```
write.csv2(exprs(eset_rma), file.path(resultsDir, "Normalized.csv"),row.names = TRUE)
```

3.3 Data filtering

3.4 Filtering of the Data

Usually, in order to increase statistical power and reduce unnecessary noise it is necesary to remove some genes. In this array (`GeneChip miRNA 4.1 array`), a lot of species, apart from *Homo sapiens*, are included. This study is focused on human miRNAs, therefore, the rest of the species have to be removed from the data.

```
#Anotacions file could be downloaded from
#Thermofisher web page (https://www.thermofisher.com/order/catalog/product/902410)
anotacions <- read.csv(file.path(dataDir,"miRNA-4_1-st-v1.annotations.20160922.mod.csv"),
                       sep=";",header=TRUE)
#add Probe set name as rownames
rownames(anotacions) <- anotacions[,2]
anotacions <- anotacions[,-2]
dim(anotacions)
```

```
## [1] 36353    13
```

```
#select the human miRNA
Hanotacions <- anotacions[which(anotacions$Species.Scientific.Name == "Homo sapiens"),]
dim(Hanotacions)
```

```
## [1] 6631   13
```

```
#save the human anotations in a file
write.csv2(Hanotacions,file.path(resultsDir,"Anotations.miRNA.human.csv"),sep=";")

#filter eset_rma to have only the human miRNA
head(exprs(eset_rma))
```

```
##             Human-Brain-AM7962-130ng_rep1_(miRNA-4_1-Array-Plate).CEL
## 14q0_st                                                     3.0217828
## 14qI-1_st                                                   0.9532970
## 14qI-1_x_st                                                 0.9394577
## 14qI-2_st                                                   0.5937052
## 14qI-3_x_st                                                 0.7011243
## 14qI-4_st                                                   4.0174040
##             Human-Brain-AM7962-130ng_rep2_(miRNA-4_1-Array-Plate).CEL
## 14q0_st                                                     2.5985842
## 14qI-1_st                                                   0.9811157
## 14qI-1_x_st                                                 0.9640493
## 14qI-2_st                                                   0.5513905
## 14qI-3_x_st                                                 0.7995114
## 14qI-4_st                                                   2.9052008
##             Human-Brain-AM7962-130ng_rep3_(miRNA-4_1-Array-Plate).CEL
## 14q0_st                                                     3.0735790
## 14qI-1_st                                                   0.7361616
## 14qI-1_x_st                                                 0.7267089
## 14qI-2_st                                                   0.6295069
## 14qI-3_x_st                                                 1.1363285
```

```
## 14qI-4_st                                                    5.5950927
##             Human-Brain-AM7962-130ng_rep4_(miRNA-4_1-Array-Plate).CEL
## 14q0_st                                                      3.0309105
## 14qI-1_st                                                    0.7718235
## 14qI-1_x_st                                                  0.8392236
## 14qI-2_st                                                    0.8132312
## 14qI-3_x_st                                                  0.7076853
## 14qI-4_st                                                    5.0504709
##             Human-Lung-AM7968-130ng_rep1_(miRNA-4_1-Array-Plate).CEL
## 14q0_st                                                      1.7926686
## 14qI-1_st                                                    0.9065172
## 14qI-1_x_st                                                  0.9479105
## 14qI-2_st                                                    0.5139178
## 14qI-3_x_st                                                  0.5156264
## 14qI-4_st                                                    0.5467712
##             Human-Lung-AM7968-130ng_rep2_(miRNA-4_1-Array-Plate).CEL
## 14q0_st                                                      1.0004942
## 14qI-1_st                                                    0.8997445
## 14qI-1_x_st                                                  0.8997445
## 14qI-2_st                                                    0.8206331
## 14qI-3_x_st                                                  0.9447606
## 14qI-4_st                                                    2.5318527
##             Human-Lung-AM7968-130ng_rep3_(miRNA-4_1-Array-Plate).CEL
## 14q0_st                                                      2.0346517
## 14qI-1_st                                                    1.1688263
## 14qI-1_x_st                                                  0.8757832
## 14qI-2_st                                                    0.5377510
## 14qI-3_x_st                                                  0.6293537
## 14qI-4_st                                                    1.4628896
##             Human-Lung-AM7968-130ng_rep4_(miRNA-4_1-Array-Plate).CEL
## 14q0_st                                                      1.6883420
## 14qI-1_st                                                    0.8132312
## 14qI-1_x_st                                                  0.8132312
## 14qI-2_st                                                    0.6987774
## 14qI-3_x_st                                                  0.7382063
## 14qI-4_st                                                    2.1825145
```

```
eset <- data.frame(exprs(eset_rma))
data <- merge(eset, Hanotacions, by=0)
head(data)
```

```
##     Row.names Human.Brain.AM7962.130ng_rep1_.miRNA.4_1.Array.Plate..CEL
```

```
## 1     14q0_st                                                    3.0217828
## 2   14qI-1_st                                                    0.9532970
## 3 14qI-1_x_st                                                    0.9394577
## 4   14qI-2_st                                                    0.5937052
## 5 14qI-3_x_st                                                    0.7011243
## 6   14qI-4_st                                                    4.0174040
##   Human.Brain.AM7962.130ng_rep2_.miRNA.4_1.Array.Plate..CEL
## 1                                                 2.5985842
## 2                                                 0.9811157
## 3                                                 0.9640493
## 4                                                 0.5513905
## 5                                                 0.7995114
## 6                                                 2.9052008
##   Human.Brain.AM7962.130ng_rep3_.miRNA.4_1.Array.Plate..CEL
## 1                                                 3.0735790
## 2                                                 0.7361616
## 3                                                 0.7267089
## 4                                                 0.6295069
## 5                                                 1.1363285
## 6                                                 5.5950927
##   Human.Brain.AM7962.130ng_rep4_.miRNA.4_1.Array.Plate..CEL
## 1                                                 3.0309105
## 2                                                 0.7718235
## 3                                                 0.8392236
## 4                                                 0.8132312
## 5                                                 0.7076853
## 6                                                 5.0504709
##   Human.Lung.AM7968.130ng_rep1_.miRNA.4_1.Array.Plate..CEL
## 1                                                1.7926686
## 2                                                0.9065172
## 3                                                0.9479105
## 4                                                0.5139178
## 5                                                0.5156264
## 6                                                0.5467712
##   Human.Lung.AM7968.130ng_rep2_.miRNA.4_1.Array.Plate..CEL
## 1                                                1.0004942
## 2                                                0.8997445
## 3                                                0.8997445
## 4                                                0.8206331
## 5                                                0.9447606
```

```
## 6                                                    2.5318527
##   Human.Lung.AM7968.130ng_rep3_.miRNA.4_1.Array.Plate..CEL
## 1                                                    2.0346517
## 2                                                    1.1688263
## 3                                                    0.8757832
## 4                                                    0.5377510
## 5                                                    0.6293537
## 6                                                    1.4628896
##   Human.Lung.AM7968.130ng_rep4_.miRNA.4_1.Array.Plate..CEL Probe.Set.ID
## 1                                                    1.6883420     20532563
## 2                                                    0.8132312     20532564
## 3                                                    0.8132312     20532565
## 4                                                    0.6987774     20532566
## 5                                                    0.7382063     20532567
## 6                                                    2.1825145     20532568
##   Accession Transcript.ID.Array.Design. Sequence.Type
## 1      14q0                        14q0         CDBox
## 2    14qI-1                      14qI-1         CDBox
## 3    14qI-1                      14qI-1         CDBox
## 4    14qI-2                      14qI-2         CDBox
## 5    14qI-3                      14qI-3         CDBox
## 6    14qI-4                      14qI-4         CDBox
##   Species.Scientific.Name                    Alignments Sequence.Length
## 1            Homo sapiens chr14:101364257-101364333 (+)              77
## 2            Homo sapiens chr14:101391158-101391227 (+)              70
## 3            Homo sapiens chr14:101391158-101391227 (+)              70
## 4            Homo sapiens chr14:101393679-101393749 (+)              71
## 5            Homo sapiens chr14:101396256-101396326 (+)              71
## 6            Homo sapiens chr14:101402828-101402901 (+)              74
##                                                                  Sequence
## 1 TGGACCAATGATGAGACAGTGTTTATGAACAAAAGATCATGATTAATCCAGTTCTGCACAAAACACTGAGGTCCATT
## 2        AAAGTGAGTGATGAATAGTTCTGTGGCATATGAATCATTAATTTTGATTAAACCCTAAACTCTGAAGTCC
## 3        AAAGTGAGTGATGAATAGTTCTGTGGCATATGAATCATTAATTTTGATTAAACCCTAAACTCTGAAGTCC
## 4       ATAGCCAATCATTAGTATTCTGAGCTGTAGGAATCAAAGATTTTGATTAGATTCTGTAACTCAGAGGTTTA
## 5       TAGACCAATGATGAGTATTCTGGGGTGTCTGAATCAATGATTTTGATTAAACCCTGTAACTCTGAGGTCCA
## 6    TGGACCAATGATGAGTACCATGGGGTATCTGAAACAGGATTTTTGATTAAACCCATATGCAATTCTGAGGTCCA
##   Genome.Context Clustered.miRNAs.within.10kb Target.Genes GeneChip.Array
## 1            ---                          ---          ---      miRNA-4_1
## 2            ---                          ---          ---      miRNA-4_1
## 3            ---                          ---          ---      miRNA-4_1
```

```
## 4            ---                              ---          ---      miRNA-4_1
## 5            ---                              ---          ---      miRNA-4_1
## 6            ---                              ---          ---      miRNA-4_1
##   Sequence.Source
## 1      snoRNABase
## 2      snoRNABase
## 3      snoRNABase
## 4      snoRNABase
## 5      snoRNABase
## 6      snoRNABase

dim(data)

## [1] 6631   22

rownames(data) <- data[,1]
colnames(data)

##  [1] "Row.names"
##  [2] "Human.Brain.AM7962.130ng_rep1_.miRNA.4_1.Array.Plate..CEL"
##  [3] "Human.Brain.AM7962.130ng_rep2_.miRNA.4_1.Array.Plate..CEL"
##  [4] "Human.Brain.AM7962.130ng_rep3_.miRNA.4_1.Array.Plate..CEL"
##  [5] "Human.Brain.AM7962.130ng_rep4_.miRNA.4_1.Array.Plate..CEL"
##  [6] "Human.Lung.AM7968.130ng_rep1_.miRNA.4_1.Array.Plate..CEL"
##  [7] "Human.Lung.AM7968.130ng_rep2_.miRNA.4_1.Array.Plate..CEL"
##  [8] "Human.Lung.AM7968.130ng_rep3_.miRNA.4_1.Array.Plate..CEL"
##  [9] "Human.Lung.AM7968.130ng_rep4_.miRNA.4_1.Array.Plate..CEL"
## [10] "Probe.Set.ID"
## [11] "Accession"
## [12] "Transcript.ID.Array.Design."
## [13] "Sequence.Type"
## [14] "Species.Scientific.Name"
## [15] "Alignments"
## [16] "Sequence.Length"
## [17] "Sequence"
## [18] "Genome.Context"
## [19] "Clustered.miRNAs.within.10kb"
## [20] "Target.Genes"
## [21] "GeneChip.Array"
## [22] "Sequence.Source"

data <- data[,-c(1,10:22)]
```

After removing miRNA from other species, the data to be included in the analysis is a list of **6.631 probes** (before removing the other species the length of the data was 36.353 probes).

3.5 Selection of Differentially Expressed Genes

The goal of the study is to find differentially expressed miRNAs between the `Brain` and `Lung` tissues.

To achieve this main objective specific comparison has been performed:

- Effect of *Tissue* condition:
    1. Brain vs Lung = Brain - Lung

The analysis to select differentially expressed genes has been based on adjusting a linear model with empirical bayes moderation of the variance. This is a technique similar to ANOVA specifically developed for microarray data analysis by Gordon K. Smyth in 2004 [**?**].
Each comparison yields a list of genes sorted from most to least differentially expressed genes which is called generically a top table. The resulting top tables are presented in an csv file. Besides for each comparisons the corresponding Volcano Plot is provided in the `results` folder.
First we make the contrast matrix:

```
require(limma)
grupo <- as.factor(targets$Group)
design <- model.matrix( ~0 + grupo)
dim(design)

## [1] 8 2

colnames(design) <- c( "Brain",  "Lung")
ContrastMatrix <- makeContrasts(BrainvsLung = Brain - Lung, levels=design)
print(ContrastMatrix)

##        Contrasts
## Levels  BrainvsLung
##   Brain           1
##   Lung           -1
```

Second, fit the model:

```
fit <- lmFit(data, design)
fit.main <- contrasts.fit(fit, ContrastMatrix)
fit.main <- eBayes(fit.main)
```

Third, obtain the tobTable:

```
topTab <- topTable (fit.main, number=nrow(fit.main), coef="BrainvsLung", adjust="fdr")
write.csv2(topTab,file.path(resultsDir,"topTabBrainvsLung.csv"),sep=";")
```

The results of the number of differentially expressed genes in each comparisons at different thresholds are shown in next table:

```
source("https://raw.githubusercontent.com/uebvhir/UEB_NumGenesChanged/master/UEB_NumGenesChangedFu
BrainvsLung <- genesSelectable(topTab,0.01,0.05,0.25,0.01,0.05)
numGenesChanged <- cbind.data.frame(BrainvsLung)
write.csv2(numGenesChanged,file.path(resultsDir,"numGenesChanged.csv"),sep=";")
```

```
x.big2 <- xtable(numGenesChanged,caption="Number of DEG for different thresholds",label="tab:numge
print(x.big2, tabular.environment = longtable, floating = FALSE,size="small")
```

	BrainvsLung
upReg_B	196
downReg_B	382
upRegAdj0.01	230
downRegAdj0.01	516
upRegAdj0.05	311
downRegAdj0.05	747
upRegAdj0.25	636
downRegAdj0.25	1194
upRegP0.01	324
downRegP0.01	785
upRegP0.05	541
downRegP0.05	1116

Table 2: Number of DEG for different thresholds

p-values adjustment If one wishes to have a statistically grounded criteria, the selection of the differentially expressed genes should be based on adjusted p-values (less than 0.01) or B statistic (greater than 0). If these criteria yield too few genes, the table **??** indicates how many genes will yield a less restrictive criteria such as calling those differentially expressed genes with, for instance, adjusted p-values less than 0.25 or unadjusted p-values smaller than 0.05.

In next table the 10 more differentially expressed genes for the comparison can be found.

```
csvtable <- xtable(topTab[1:10,], caption="10 genes more differentially expressed")
print(csvtable, tabular.environment=longtable, floating=FALSE, size="normal",
      include.rownames=TRUE)
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
MIMAT0000422_st	12.80	6.82	177.40	0.00	0.00	24.57
MIMAT0004675_st	9.02	5.30	113.46	0.00	0.00	22.88
HBII-52-32_x_st	8.27	5.00	107.75	0.00	0.00	22.63
MIMAT0000738_st	6.09	3.35	83.20	0.00	0.00	21.18
HBII-52-23_x_st	5.13	3.05	71.98	0.00	0.00	20.24
MIMAT0004605_st	6.24	4.30	68.15	0.00	0.00	19.86
MIMAT0004954_st	5.17	3.31	67.64	0.00	0.00	19.81
MIMAT0000442_st	8.35	5.10	66.84	0.00	0.00	19.73
HBII-52-22_x_st	7.29	5.00	66.83	0.00	0.00	19.73
MIMAT0004978_st	6.95	4.68	56.76	0.00	0.00	18.54

Table 3: 10 genes more differentially expressed

A Volcano plot could be easy performed:

```
volcanoplot(fit.main, coef = 1, highlight = 10)
```