732A51 Bioinformatics

LAB 4 Bioinformatics

Hugo Morvan
William Wiik

STIMA
Department of Computer and Information Science
Linköpings universitet

2024-12-05

Contents

```r
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install()
BiocManager::install("GEOquery")
library(GEOquery)

#ERROR: dependencies 'affy', 'genefilter', 'gcrma' are not available for package 'simpleaffy'
BiocManager::install('affy')
BiocManager::install('affyPLM')
BiocManager::install('genefilter')
BiocManager::install('gcrma')
install.packages('./simpleaffy_2.50.0.tar.gz', type='source') #Needs Linux ?
library(simpleaffy)

library(RColorBrewer)

library(limma)

BiocManager::install('hgu133plus2.db')
library(hgu133plus2.db)

library(annotate)

library(ggplot2)
```

1 Question 1

Run all the R code and reproduce the graphics. Go carefully through the R code and explain in your words what each step does. HINT Recall what a design/model matrix is from linear regression.

```r
# Important note: before knitting, delete the data folder
library(GEOquery)
x = getGEOSuppFiles("GSE20986")
x
```

```
##                                                                    size
## /home/h/Documents/LiU/Bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 56360960
##                                                                   isdir mode
## /home/h/Documents/LiU/Bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar FALSE  664
##                                                                           mtime
## /home/h/Documents/LiU/Bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 2024-12-05 09:29:07
##                                                                           ctime
## /home/h/Documents/LiU/Bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 2024-12-05 09:29:07
##                                                                           atime
## /home/h/Documents/LiU/Bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 2024-12-05 09:28:14
##                                                                    uid  gid
## /home/h/Documents/LiU/Bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 1000 1000
```

```
##                                                                    uname
## /home/h/Documents/LiU/Bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar     h
##                                                                   grname
## /home/h/Documents/LiU/Bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar     h
```

```
untar("GSE20986/GSE20986_RAW.tar", exdir = "data")
cels = list.files("data/", pattern = "[gz]")
sapply(paste("data", cels, sep = "/"), gunzip)
```

```
## data/GSM524662.CEL.gz data/GSM524663.CEL.gz data/GSM524664.CEL.gz
##              13555726              13555055              13555639
## data/GSM524665.CEL.gz data/GSM524666.CEL.gz data/GSM524667.CEL.gz
##              13560122              13555663              13557614
## data/GSM524668.CEL.gz data/GSM524669.CEL.gz data/GSM524670.CEL.gz
##              13556090              13560054              13555971
## data/GSM524671.CEL.gz data/GSM524672.CEL.gz data/GSM524673.CEL.gz
##              13554926              13555042              13555290
```

```
phenodata = matrix(rep(list.files("data"), 2), ncol =2)
class(phenodata)
```

```
## [1] "matrix" "array"
```

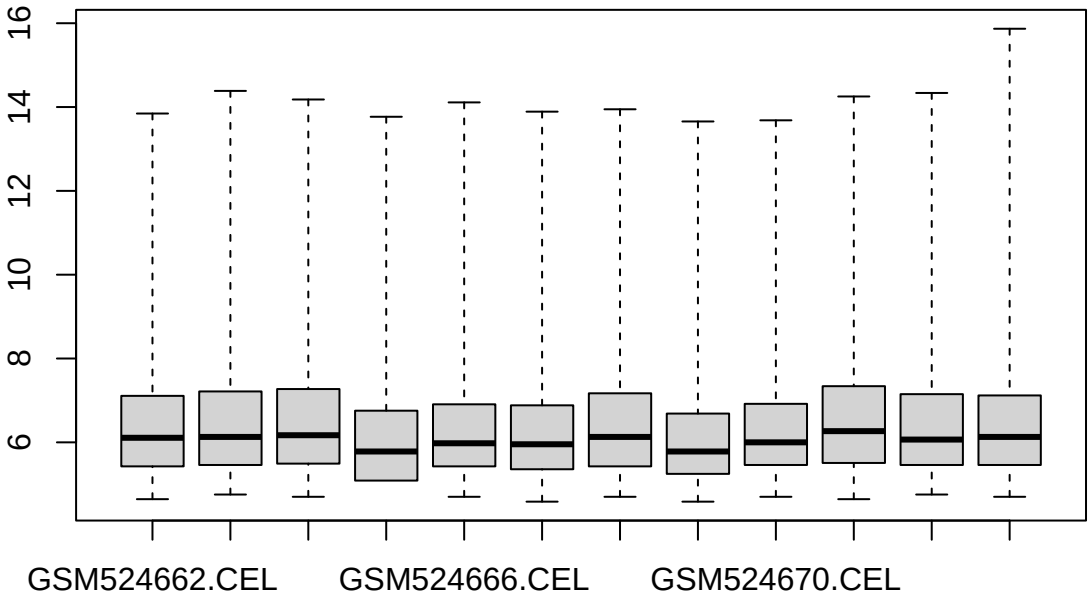
```
phenodata <- as.data.frame(phenodata)
colnames(phenodata) <- c("Name", "FileName")
phenodata$Targets <- c("iris",
                       "retina",
                       "retina",
                       "iris",
                       "retina",
                       "iris",
                       "choroid",
                       "choroid",
                       "choroid",
                       "huvec",
                       "huvec",
                       "huvec")
write.table(phenodata, "data/phenodata.txt", quote = F, sep = "\t", row.names = F)
```

```
library(simpleaffy)
celfiles <- read.affy(covdesc = "phenodata.txt", path = "data")
boxplot(celfiles)
```

```
## Warning: replacing previous import 'AnnotationDbi::tail' by 'utils::tail' when
## loading 'hgu133plus2cdf'
```

```
## Warning: replacing previous import 'AnnotationDbi::head' by 'utils::head' when
## loading 'hgu133plus2cdf'
```

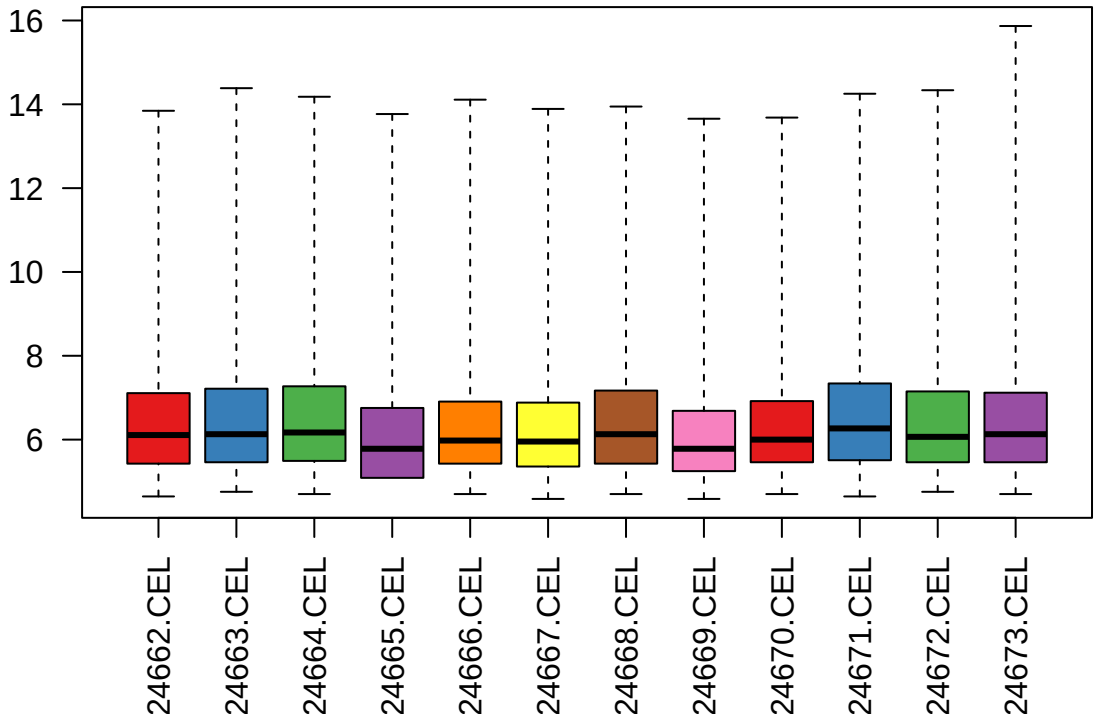
##



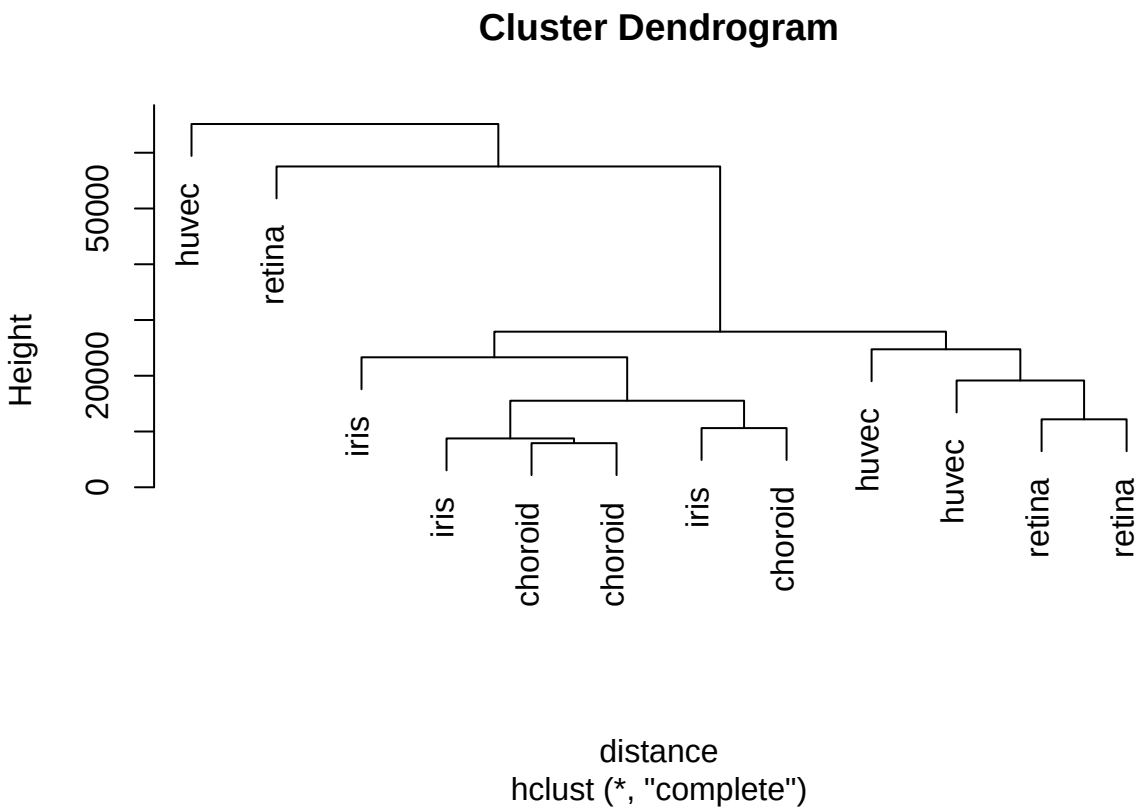

```
library(RColorBrewer)
cols = brewer.pal(8, "Set1")
eset <- exprs(celfiles)
samples <- celfiles$Targets
colnames(eset)
```

```
##  [1] "GSM524662.CEL" "GSM524663.CEL" "GSM524664.CEL" "GSM524665.CEL"
##  [5] "GSM524666.CEL" "GSM524667.CEL" "GSM524668.CEL" "GSM524669.CEL"
##  [9] "GSM524670.CEL" "GSM524671.CEL" "GSM524672.CEL" "GSM524673.CEL"
```

```
colnames(eset) <- samples
boxplot(celfiles, col = cols, las = 2)
```




```
distance <- dist(t(eset), method = "maximum")
clusters <- hclust(distance)
plot(clusters)
```




```
require(simpleaffy)
require(affyPLM)
```

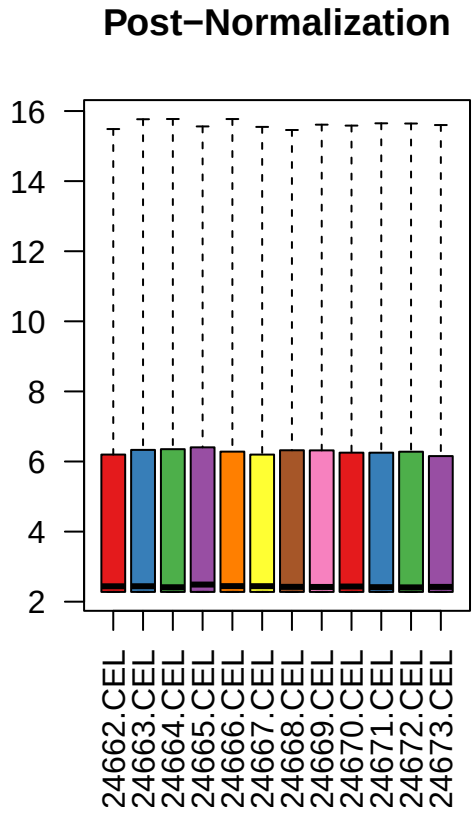
```
## Loading required package: affyPLM
```

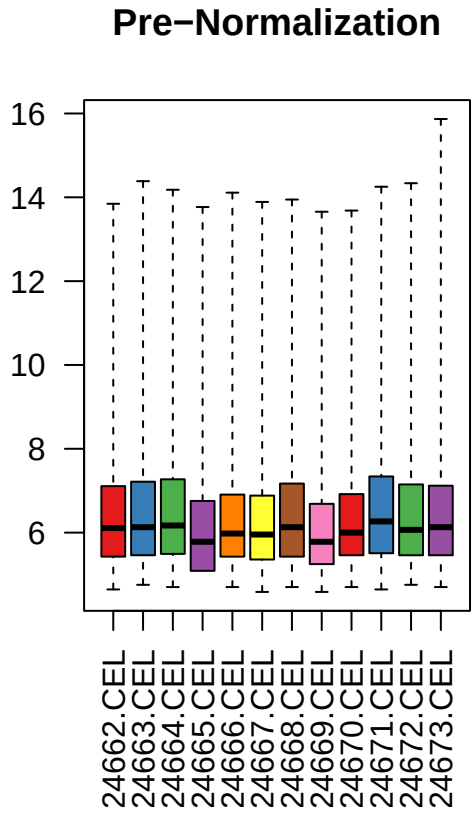
```
## Loading required package: preprocessCore
```

```
celfiles.gcrma = gcrma(celfiles)
```

```
## Adjusting for optical effect.....Done.
## Computing affinities.Done.
## Adjusting for non-specific binding.....Done.
## Normalizing
## Calculating Expression
```

```
par(mfrow=c(1,2))
boxplot(celfiles.gcrma, col = cols, las = 2, main = "Post-Normalization");
boxplot(celfiles, col = cols, las = 2, main = "Pre-Normalization")
```






```
#dev.off()
library(limma)
```

```
phenodata
```

```
##              Name      FileName Targets
## 1  GSM524662.CEL GSM524662.CEL    iris
## 2  GSM524663.CEL GSM524663.CEL  retina
## 3  GSM524664.CEL GSM524664.CEL  retina
## 4  GSM524665.CEL GSM524665.CEL    iris
## 5  GSM524666.CEL GSM524666.CEL  retina
## 6  GSM524667.CEL GSM524667.CEL    iris
## 7  GSM524668.CEL GSM524668.CEL choroid
## 8  GSM524669.CEL GSM524669.CEL choroid
## 9  GSM524670.CEL GSM524670.CEL choroid
## 10 GSM524671.CEL GSM524671.CEL   huvec
## 11 GSM524672.CEL GSM524672.CEL   huvec
## 12 GSM524673.CEL GSM524673.CEL   huvec
```

```
samples <- as.factor(samples)
design <- model.matrix(~0+samples)
colnames(design)
```

```
## [1] "sampleschoroid" "sampleshuvec"   "samplesiris"    "samplesretina"
```

```
colnames(design) <- c("choroid", "huvec", "iris", "retina")
design
```

```
##    choroid huvec iris retina
## 1        0     0    1      0
## 2        0     0    0      1
## 3        0     0    0      1
## 4        0     0    1      0
## 5        0     0    0      1
## 6        0     0    1      0
## 7        1     0    0      0
## 8        1     0    0      0
## 9        1     0    0      0
## 10       0     1    0      0
## 11       0     1    0      0
## 12       0     1    0      0
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$samples
## [1] "contr.treatment"
```

```
contrast.matrix = makeContrasts(
              huvec_choroid = huvec - choroid,
              huvec_retina = huvec - retina,
              huvec_iris <- huvec - iris,
              levels = design)

fit = lmFit(celfiles.gcrma, design)
huvec_fit <- contrasts.fit(fit, contrast.matrix)
huvec_ebay <- eBayes(huvec_fit)
```

```
library(hgu133plus2.db)

library(annotate)
```

```
probenames.list <- rownames(topTable(huvec_ebay, number = 100000))
getsymbols <- getSYMBOL(probenames.list, "hgu133plus2")
results <- topTable(huvec_ebay, number = 100000, coef = "huvec_choroid")
results <- cbind(results, getsymbols)

summary(results)
```

```
##      logFC              AveExpr             t                 P.Value
##  Min.   :-9.19178   Min.   : 2.279   Min.   :-39.77095   Min.   :0.0000
##  1st Qu.:-0.05972   1st Qu.: 2.281   1st Qu.: -0.70703   1st Qu.:0.1522
##  Median : 0.00000   Median : 2.480   Median :  0.00000   Median :0.5080
##  Mean   :-0.02355   Mean   : 4.375   Mean   :  0.07445   Mean   :0.5345
##  3rd Qu.: 0.03970   3rd Qu.: 6.241   3rd Qu.:  0.67369   3rd Qu.:1.0000
##  Max.   : 8.66974   Max.   :15.542   Max.   :295.37719   Max.   :1.0000
##    adj.P.Val            B           getsymbols
##  Min.   :0.0000   Min.   :-7.711   Length:54675
##  1st Qu.:0.6036   1st Qu.:-7.711   Class :character
##  Median :1.0000   Median :-7.452   Mode  :character
##  Mean   :0.7436   Mean   :-6.583
##  3rd Qu.:1.0000   3rd Qu.:-6.498
##  Max.   :1.0000   Max.   :21.296
```

```
results$threshold <- "1"
a <- subset(results, adj.P.Val < 0.05 & logFC > 5)
results[rownames(a), "threshold"] <- "2"
b <- subset(results, adj.P.Val < 0.05 & logFC < -5)
results[rownames(b), "threshold"] <- "3"
table(results$threshold)
```

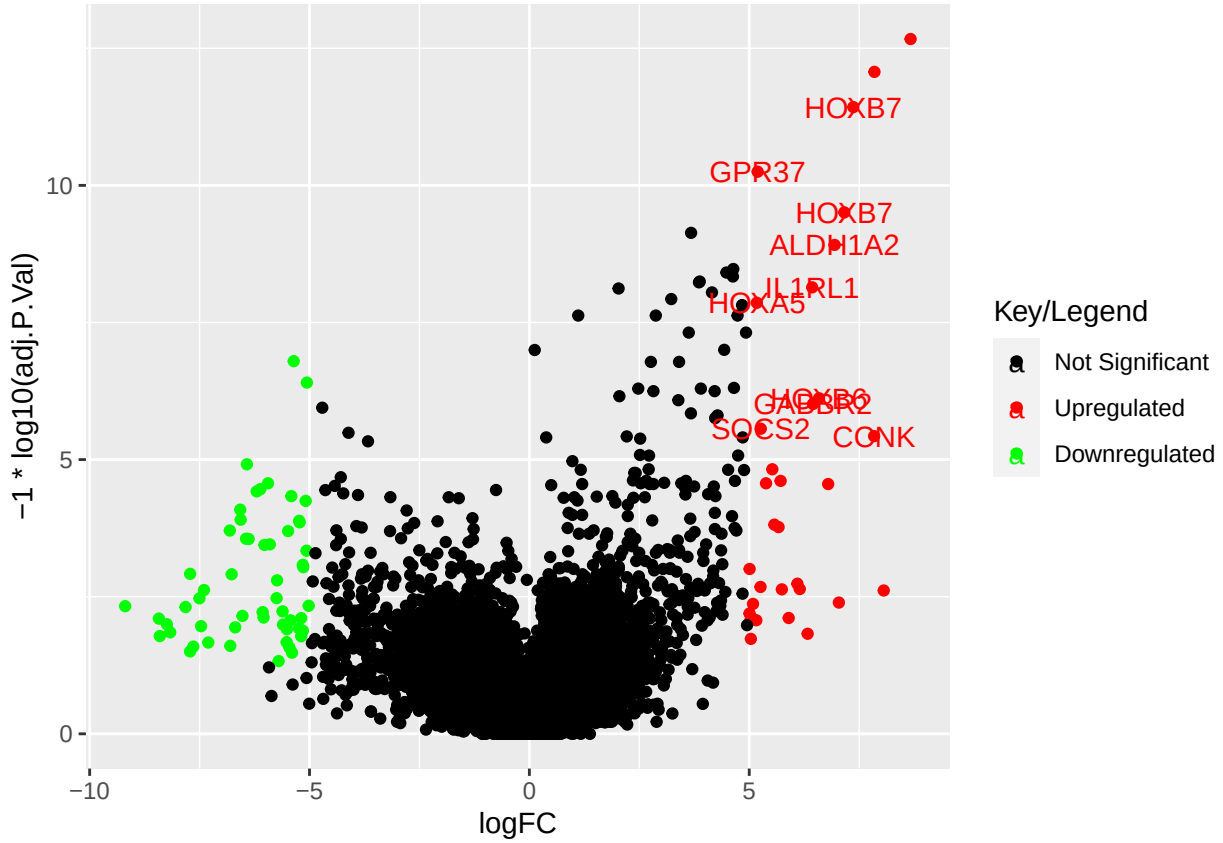
```
##
##     1     2     3
## 54587    33    55
```

```
library(ggplot2)
volcano <- ggplot(data = results,
                  aes(x = logFC, y = -1*log10(adj.P.Val),
                      colour = threshold,
                      label = getsymbols))

volcano <- volcano +
  geom_point() +
  scale_color_manual(values = c("black", "red", "green"),
                     labels = c("Not Significant", "Upregulated", "Downregulated"),
                     name = "Key/Legend")

volcano +
  geom_text(data = subset(results, logFC > 5 & -1*log10(adj.P.Val) > 5), aes(x = logFC, y = -1*log10(adj
```

```
## Warning: Removed 2 rows containing missing values (`geom_text()`).
```



2 Question 2

In the presented analysis, there are no plots of raw paired data. In the section where the contrasts are defined find the three contrasts. Present the variables versus each other original, log–scaled and MA–plot for each considered pair both before and after normalization. A cluster analysis is performed on the page but not reported. Present plots and also draw heatmaps.

3 Question 3

The volcano plot is only for huvec versus choroid. Provide volcano plots for the other pairs. Indicate significantly differentially expressed genes. Explain how they are found.

4 Question 4

Try to find more information on the genes that are reported to be significantly differentially expressed. The place to start off is https://www.ncbi.nlm.nih.gov/gene/, remember that the data is from the species human.

Try to look also for other databases where (some) information on the genes may be found. Try to follow on some of the provided links. Report in your own words on what you find. Report all the Gene Ontology (GO) terms associated with each gene. Are any of the GO terms common between genes? If so do the common GO terms seem to be related to anything particular? Try to present GO analysis in an informative manner, if possible visualize.