Lab 4 - Gr. 14 - Bioinformatics (732A93)

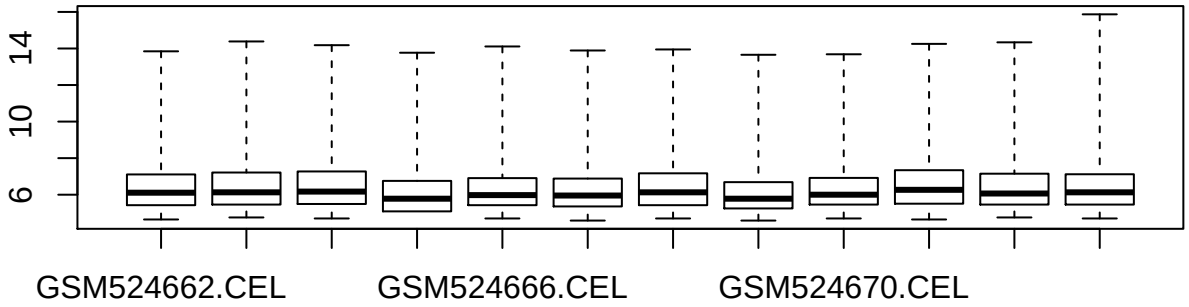
Julius Kittler (julki092), Stefano Toffol (steto820), Saewon Jun (saeju204), Maximilian Pfundstein (maxpf364)

Provided source code we have to explain.

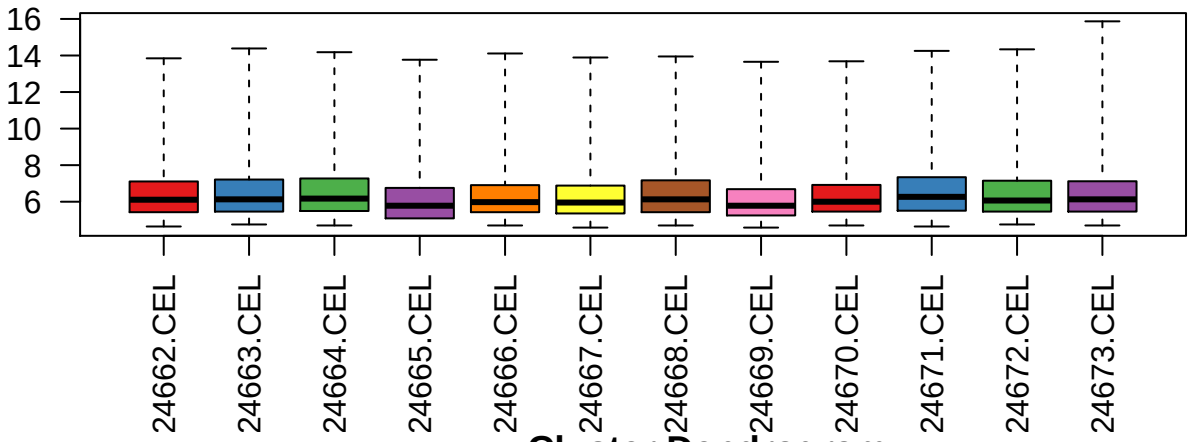
```
##                                                                    size
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 56360960
##                                                                   isdir
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar FALSE
##                                                                    mode
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar  644
##                                                                           mtime
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 2018-12-03 14:35:00
##                                                                           ctime
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 2018-12-03 14:35:00
##                                                                           atime
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 2018-12-03 14:34:52
##                                                                   uid gid
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar 501  20
##                                                                     uname
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar flennic
##                                                                   grname
## /Users/flennic/git/bioinformatics/Lab4/GSE20986/GSE20986_RAW.tar  staff
```

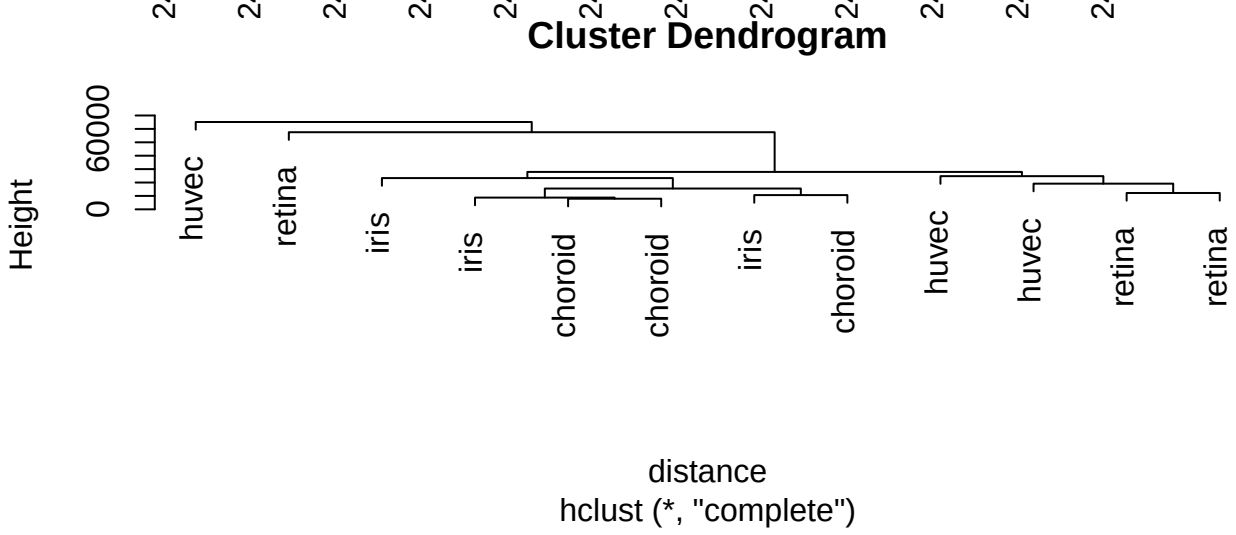
```
## data/GSM524662.CEL.gz data/GSM524663.CEL.gz data/GSM524664.CEL.gz
##              13555726              13555055              13555639
## data/GSM524665.CEL.gz data/GSM524666.CEL.gz data/GSM524667.CEL.gz
##              13560122              13555663              13557614
## data/GSM524668.CEL.gz 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
```

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

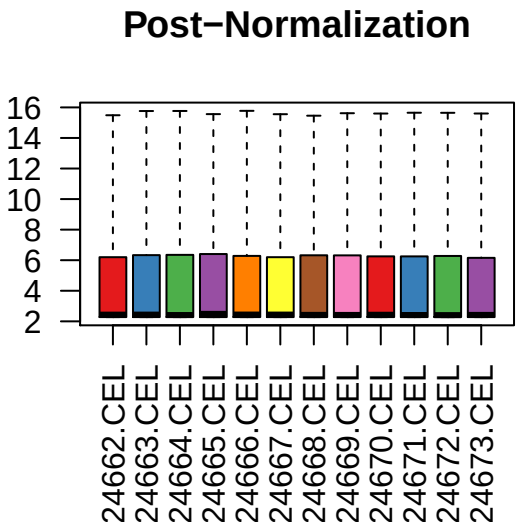


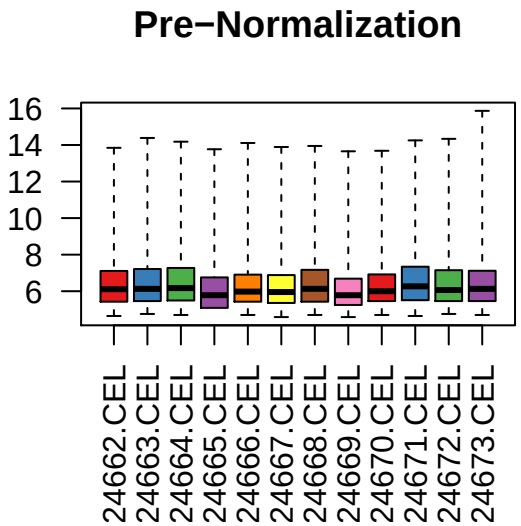

```
##  [1] "GSM524662.CEL" "GSM524663.CEL" "GSM524664.CEL" "GSM524665.CEL"
##  [5] "GSM524666.CEL" "GSM524667.CEL" "GSM524668.CEL" "GSM524669.CEL"
##  [9] "GSM524670.CEL" "GSM524671.CEL" "GSM524672.CEL" "GSM524673.CEL"
```






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






```
## null device
```

```
##           1
```

```
##             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
```

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

```
##    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"
```

```
##      logFC             AveExpr             t                 P.Value
##  Min.   :-9.19111   Min.   : 2.279   Min.   :-39.77473   Min.   :0.0000
##  1st Qu.:-0.05967   1st Qu.: 2.281   1st Qu.: -0.70649   1st Qu.:0.1523
##  Median : 0.00000   Median : 2.480   Median :  0.00000   Median :0.5079
##  Mean   :-0.02353   Mean   : 4.375   Mean   :  0.07441   Mean   :0.5346
##  3rd Qu.: 0.03986   3rd Qu.: 6.241   3rd Qu.:  0.67455   3rd Qu.:1.0000
##  Max.   : 8.67086   Max.   :15.541   Max.   :296.84201   Max.   :1.0000
##
##    adj.P.Val            B             getsymbols
##  Min.   :0.0000   Min.   :-7.710   YME1L1  :   22
##  1st Qu.:0.6036   1st Qu.:-7.710   HFE     :   15
##  Median :1.0000   Median :-7.451   CFLAR   :   14
##  Mean   :0.7436   Mean   :-6.582   NRP2    :   14
##  3rd Qu.:1.0000   3rd Qu.:-6.498   ARHGEF12:   13
##  Max.   :1.0000   Max.   :21.290   (Other) :41857
##                                    NA's    :12740
```

```
##
##     1     2     3
```

```
## 54587    33    55
```

Assignment 1

Task:

- Run all the R code and reproduce the graphics.
- Go carefully through the R code and explain in your words what each step does.

Assignment 2

Task:

- 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 report. Present plots and also draw heatmaps.

Assignment 3

Task:

- Provide volcano plots for the other pairs.
- Indicate significantly differentially expressed genes.
- Explain how they are found.

Assignment 4

Task:

- Try to find more information on the genes that are reported to be significantly differentially expressed. 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.

Appendix

```
knitr::opts_chunk$set(fig.width = 7, fig.height = 3, echo = FALSE,
                      warning = FALSE, message = FALSE)

# All privided Links:
# R Bio: Untangling Genomes
# https://www.bioconductor.org/help/course-materials/ 2015/Uruguay2015/
# Step by Step HUVEC and OVE
# https://www.bioconductor.org/help/course-materials/2015/Uruguay2015/ day3-gene.expression.html
# Cell Data:
# ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE20nnn/GSE20986/suppl/
# Description of Cell Data:
```

```r
# https://www.ncbi.nlm.nih.gov/geo/query/acc. cgi?acc=GSE20986
# Additional
#  https://www.bioconductor.org/help/course-materials/ 2015/Uruguay2015/day5-data_analysis.html
# Explanations Graphic
# https://www.bioconductor.org/help/course-materials/2015/Uruguay2015/V6-RNASeq.html

library(ggplot2)

# Use this if BiocManager is not installed
#if (!requireNamespace("BiocManager", quietly = TRUE))
#    install.packages("BiocManager")
#library("BiocManager")

# BiocManager packages
# BiocManager::install("GEOquery", version = "3.8")
# BiocManager::install("simpleaffy", version = "3.8")
# BiocManager::install("RColorBrewer", version = "3.8")
# BiocManager::install("affyPLM", version = "3.8")
# BiocManager::install("limma", version = "3.8")
# BiocManager::install("annotate", version = "3.8")
# BiocManager::install("hgu133plus2.db", version = "3.8")
library(GEOquery)
library(simpleaffy)
library(RColorBrewer)
library(affyPLM)
library(limma)
library(hgu133plus2.db)
library(annotate)


# Get the Data
x = getGEOSuppFiles("GSE20986")
x

# Untar and Unzip
# DONT ADD THE FILE TO GIT IT'S 60MB!
untar("GSE20986/GSE20986_RAW.tar", exdir = "data")
cels = list.files("data/", pattern = "[gz]")
sapply(paste("data", cels, sep = "/"), gunzip)

# Creating phenodata
phenodata = matrix(rep(list.files("data"), 2), ncol =2)
class(phenodata)
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)
celfiles <- read.affy(covdesc = "phenodata.txt", path = "data")
boxplot(celfiles)

cols = brewer.pal(8, "Set1")
eset <- exprs(celfiles)
samples <- celfiles$Targets
colnames(eset)
colnames(eset) <- samples
boxplot(celfiles, col = cols, las = 2)
distance <- dist(t(eset), method = "maximum")
clusters <- hclust(distance)
plot(clusters)
celfiles.gcrma = gcrma(celfiles)

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()

distance <- dist(t(exprs(celfiles.gcrma)), method = "maximum")
clusters <- hclust(distance)
plot(clusters)

phenodata

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

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

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)

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)
```

```r
summary(results)

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)

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.P.Val), colour = threshold,
                label = getsymbols)  )

# ------------------------------------------------------------------------------
# Question 1
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# Question 2
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# Question 3
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# Question 4
# ------------------------------------------------------------------------------
```