Case Studies in Bioinformatics

Lucas Anchieri

1/7/2019

Introduction

This module covers several studies that analysed gene expression patterns in relation to breast cancer. Their goal is to describe gene expression signatures (GES) in different subtypes of breast cancer in order to find potential targets for future therapy techniques.

In class, we reproduced the analysis done in (Parker et al. 2009), which covered a set of genes called PAM50 and studied expression patterns of these genes in different subtypes of breast cancer. They proceeded by plotting the gene expression scores on a heatmap. The scores of a given gene for each sample was represented in a cell by a color on a spectrum from blue (downregulated) to white (no change) to red (upregulated). The resulting heatmap linked samples (x axis) to genes (y axis) with a color. This makes it possible to discern patterns of up- or downregulation between several genes.

The next step was to regroup the samples into known subtypes. Replotting the heatmap according to the subtypes allowed them to discern patterns of expression between genes more easily. The genes could then be regrouped into gene expression signatures for further analysis.

For this assignment, the goal is to use what we learned while working on the PAM50 gene set to reproduce a study by (Gatza et al. 2014). This study went one step further by plotting a heatmap of several GES (each containing several genes) in relation to breast cancer subtypes. We were given a set of gene expression data for Uterine Corpus Endometrial Carcinoma (UCEC) and 3 subtypes to analyse in relation to the GES used in Gatza's study.

For readability, all the heatmaps are also provided in separate files in the "output" folder in the zip file joined to this report.

Method and Results

The Data

We'll start by loading the ucec gene expression data and take a look at it.

```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
a <- load("data/ucec_ge.RData")
ge <- ucec_ge
a <- load("data/ucec_annotation.RData") #samples classed within 3 subtypes
a <- load("data/signatures_gatza.RData") #genes regrouped in 49 signatures

dim(ucec_ge)
```

```
## [1] 20502   381
```

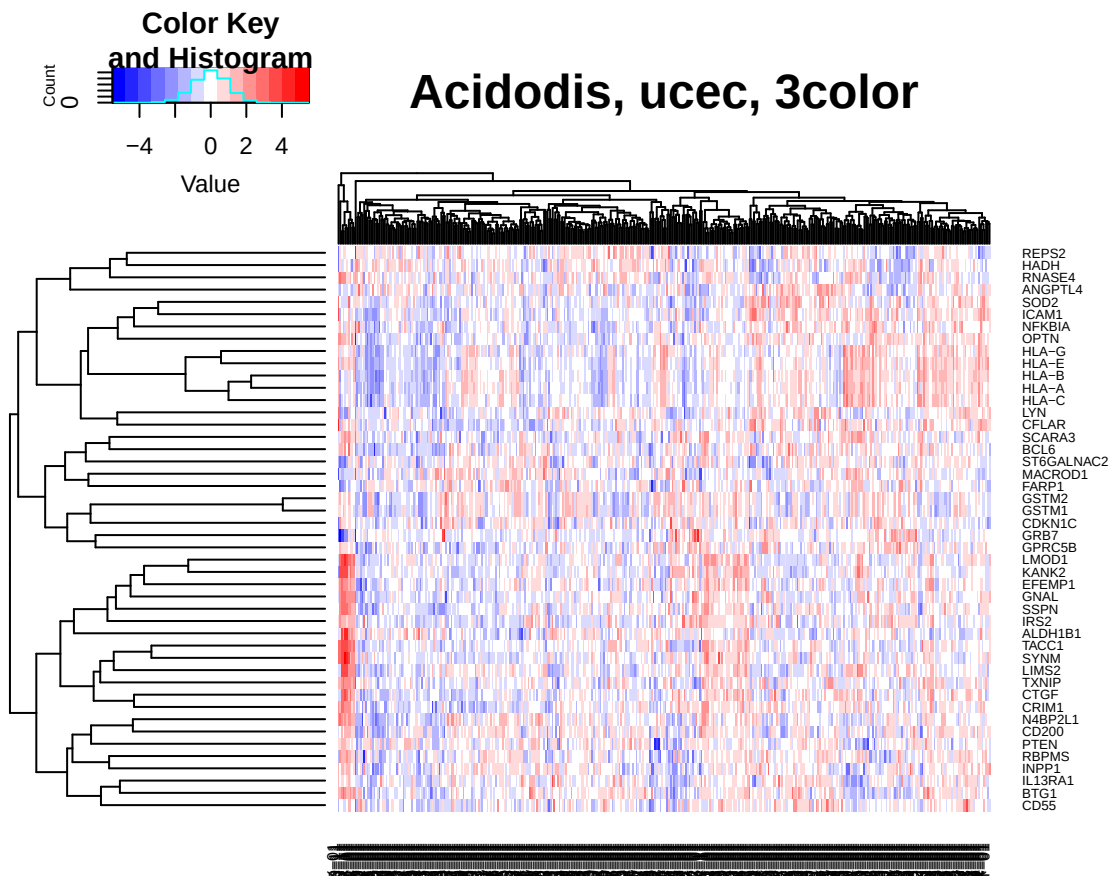
```
ge[1:4,1:4]
```

```
##       TCGA-A5-A0G1-01 TCGA-A5-A0G2-01 TCGA-A5-A0G3-01 TCGA-A5-A0G5-01
## A1BG          68.5764         14.2999        134.5286        656.2865
## A1CF           0.0000          0.0000          0.0000          0.0000
## A2BP1          0.0000          0.0000          0.0000          0.0000
## A2LD1         96.8229        107.7100         62.2714        167.0424
```

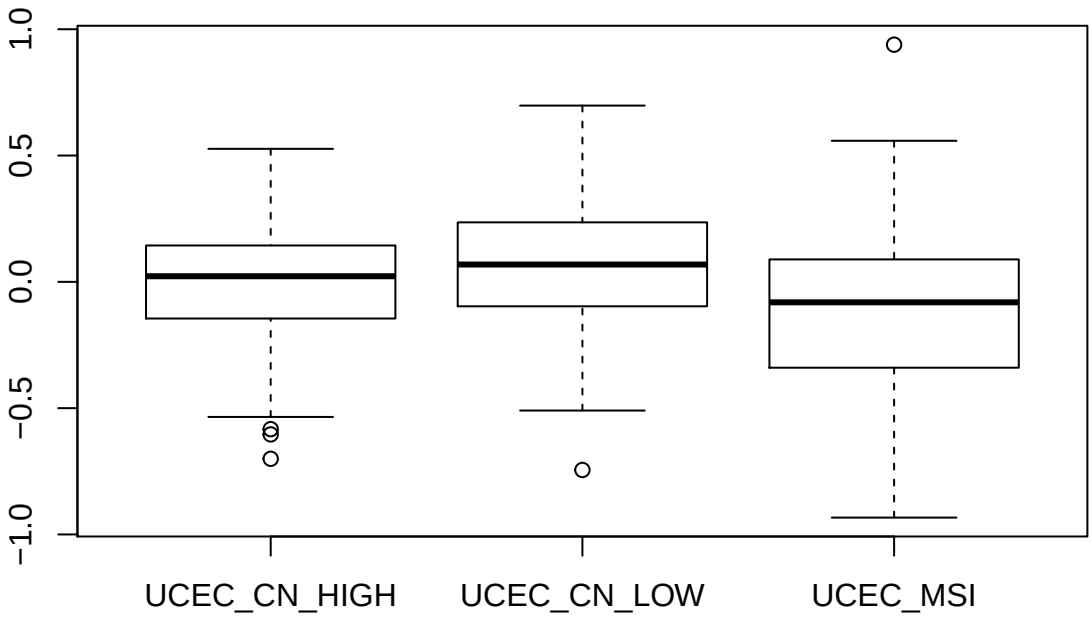
The data set comprises of 20502 genes with expression scores for 381 samples.

Before analysing all the GES, will take a look at and analyse one of them. We'll take ACIDOSIS, as it is the first one in the data set.

```
#getting the names of the genes in the GES
signatures_acidosis <- unlist(signatures_gatza["ACIDOSIS"])

#keeping ony those in the gene expression matrix
acidosis_ge <- ge[signatures_acidosis,]

#apply a log to all expression values so that the data will be readable
ac_ge_log <- log2(acidosis_ge+1)

#normalizing the data
ac_ge_st <- t(apply(ac_ge_log,1,function(x) (x - mean(x))/sd(x)))

#computing the clusterings as done during class
gene_clust = hclust(dist(ac_ge_st, method='euclidean'), method = 'complete')
sample_clust = hclust(dist(t(ac_ge_st), method='euclidean'), method = 'complete')

#plotting the heatmap
heatmap.2(ac_ge_st, trace = "none",
          Rowv = as.dendrogram(gene_clust),
          Colv = as.dendrogram(sample_clust),
          col=colorRampPalette(c('blue', 'white', 'red')),
          main = "Acidodis, ucec, 3color",
          margins = c(2,8),
          cexRow = 0.5)
```




```r
ac_ge_st_mn <- apply(ac_ge_st,2,mean) #calculating the GES score

#keeping the subtype annotation only for the 381 samples studied
subtyping <- droplevels(as.factor(ucec_annotation[names(ucec_annotation) %in% names(ac_ge_st_mn)]))

#keeping the subtype annotation only for the 381 samples studied
ac_ge_st_mn <- ac_ge_st_mn[names(ac_ge_st_mn) %in% names(ucec_annotation)]

boxplot(ac_ge_st_mn~subtyping) #boxplot of the gene expression score for each subtype
```



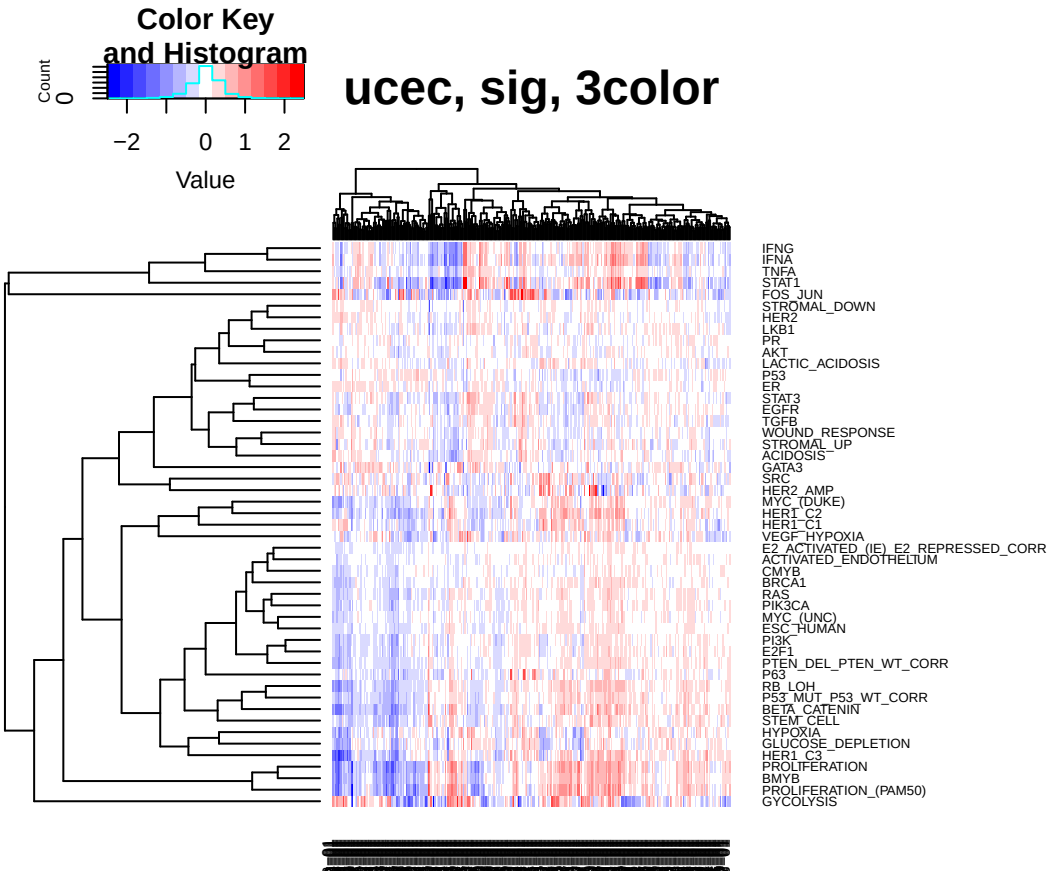
GES scores

We will now calculate the GES scores as we did for ACIDOSIS for all the signatures and store the data in a matrix.

```
ges_tot <- matrix(0L, nrow = 49, ncol = 304) #creating the matrix

rownames(ges_tot) <- names(signatures_gatza) #adding the names of the signatures

#applying what we did for ACIDOSIS to every signature with a loop
for (id in names(signatures_gatza)) {
  signatures_id <- unlist(signatures_gatza[[id]])

  signatures_id <- signatures_id[signatures_id %in% rownames(ge)]

  id_ge <- ge[signatures_id,]

  id_ge_log <- log2(id_ge+1)

  id_ge_st <- t(apply(id_ge_log,1,function(x) (x - mean(x))/sd(x)))

  id_ge_st_mn <- apply(id_ge_st,2,mean)

  id_ge_st_mn <- id_ge_st_mn[names(id_ge_st_mn) %in% names(ucec_annotation)]

  ges_tot[id,] <- id_ge_st_mn #adding the GES score to the matrix

}

colnames(ges_tot) <- names(ac_ge_st_mn) #adding the sample names
```

We can now generate a first heatmap of our GES scores.

```
gene_clust = hclust(dist(ges_tot, method='euclidean'), method = 'complete')
sample_clust = hclust(dist(t(ges_tot), method='euclidean'), method = 'complete')

heatmap.2(ges_tot, trace = "none",
          Rowv = as.dendrogram(gene_clust),
          Colv = as.dendrogram(sample_clust),
          col=colorRampPalette(c('blue', 'white', 'red')),
          main = "ucec, sig, 3color",
          margins = c(2,16),
          cexRow = 0.5)
```



We can already see that there are some patterns there. What we need to do now is to sort the samples by subtype.

```
subtypes <- sort(levels(subtyping)) #isolating the names of the subtypes

subtype_color_scale = c('red', 'green', 'blue') #setting up the color scale
names(subtype_color_scale) = subtypes

table(subtyping)
```

```
## subtyping
## UCEC_CN_HIGH  UCEC_CN_LOW     UCEC_MSI
##           83          118          103
```

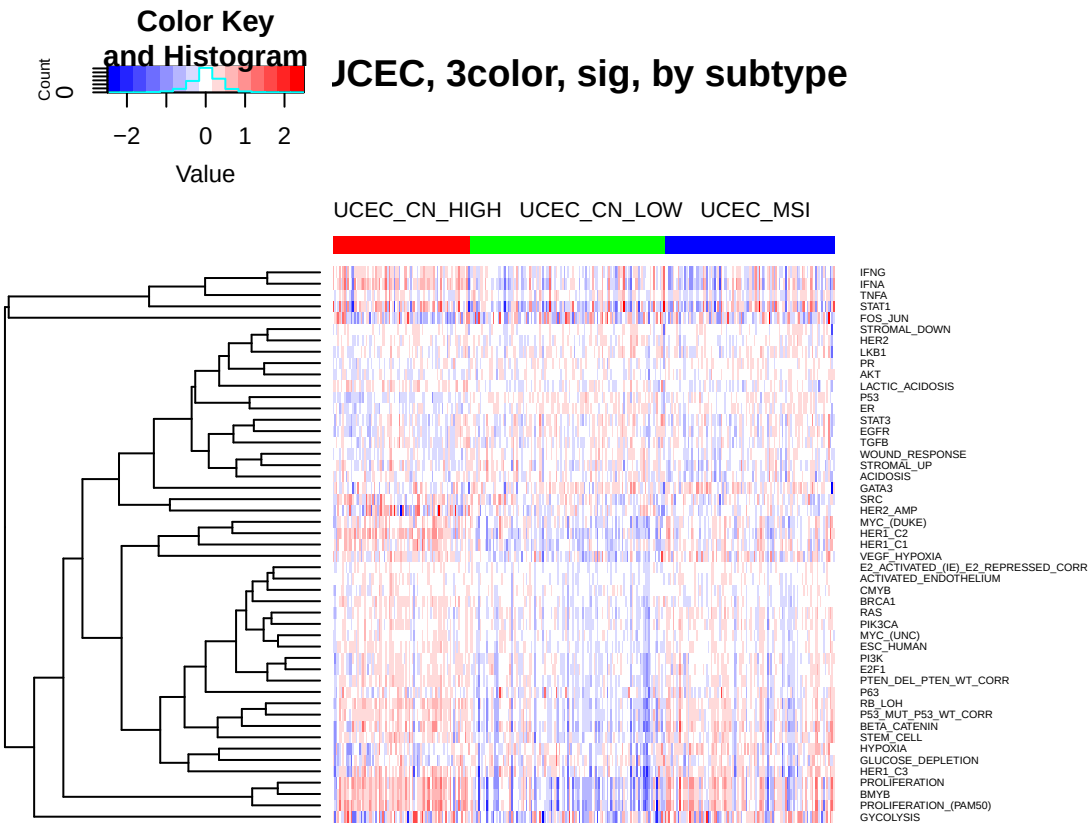
```
#testing e^whether the names in subtyping correspond those in ges_tot
names(subtyping) == colnames(ges_tot)
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [85] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [99] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [113] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [127] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [141] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [155] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [169] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [183] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [197] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [211] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [225] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [239] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [253] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [267] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [281] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [295] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
subtype_colors = subtype_color_scale[subtyping]
names(subtype_colors) = names(subtyping)
print(subtype_color_scale)
```

```
## UCEC_CN_HIGH  UCEC_CN_LOW     UCEC_MSI
##        "red"      "green"       "blue"
```

```
heatmap.2(ges_tot[, order(subtyping)],
          trace='none',
          col=colorRampPalette(c('blue', 'white', 'red')),
          ColSideColors=subtype_colors[order(subtyping)],
          Colv = FALSE,
          dendrogram = 'row',
          Rowv = as.dendrogram(gene_clust),
          labCol = paste(unlist(names(subtype_color_scale)), collapse = "   "),
          cexCol = 1.0,
          cexRow = 0.5,
          srtCol = 0,
          margins = c(2,16),
          adjCol = c(0,-41),
          main = "UCEC, 3color, sig, by subtype")
```



We can see that there is a correlation in the scores of some signatures, like PROLIFERATION, PROLIFERATION_(PAM50), BMYB, HER1_C1, HER1_C2, RB_LOH, P53_MUT_P53_WT_CORR, BETA_CATENIN, STEM_CELL, IFNG, IFNA, etc.

Statistical Analysis

We will now test the results to see whether the observed correlations are significative or not. Again, we'll first start by doing it for one signature step by step, then apply it to the whole data set with a loop.

For One Signature

We could use a t-test to test the significance of the difference of the signature's gene expression score between one subtype and the other, but the test isn't strong enough and might give us p values that are too low. We will therefore use teh Tukey's Honest Significance test, which is more reliable in this case.

```
pairwise.t.test(ges_tot[1,], subtyping[names(ges_tot[1,])]) #t-test for comparison
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ges_tot[1, ] and subtyping[names(ges_tot[1, ])]
##
##             UCEC_CN_HIGH UCEC_CN_LOW
## UCEC_CN_LOW 0.16878      -
## UCEC_MSI    0.02742      0.00012
##
## P value adjustment method: holm
```

```r
al = aov(ges_tot[1,] ~ subtyping[names(ges_tot[1,])]) #arguments for tukey's test
TukeyHSD(al)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ges_tot[1, ] ~ subtyping[names(ges_tot[1, ])])
##
## $`subtyping[names(ges_tot[1, ])]`
##                                 diff         lwr          upr     p adj
## UCEC_CN_LOW-UCEC_CN_HIGH  0.05804112 -0.04106187  0.157144107 0.3529622
## UCEC_MSI-UCEC_CN_HIGH    -0.10741249 -0.20945182 -0.005373164 0.0364233
## UCEC_MSI-UCEC_CN_LOW     -0.16545361 -0.25873721 -0.072170011 0.0001142
```

Here, we see that the difference in gene expression score for ACIDOSIS between the CN_HIGH and CN_LOW subtypes isn't significant ($p = 0.35$). However, the differences between MSI and CN_HIGH ($p < 0.05$) as well as between MSI and CN_LOW ($p< 0.01$) are both significant.

For All Signatures

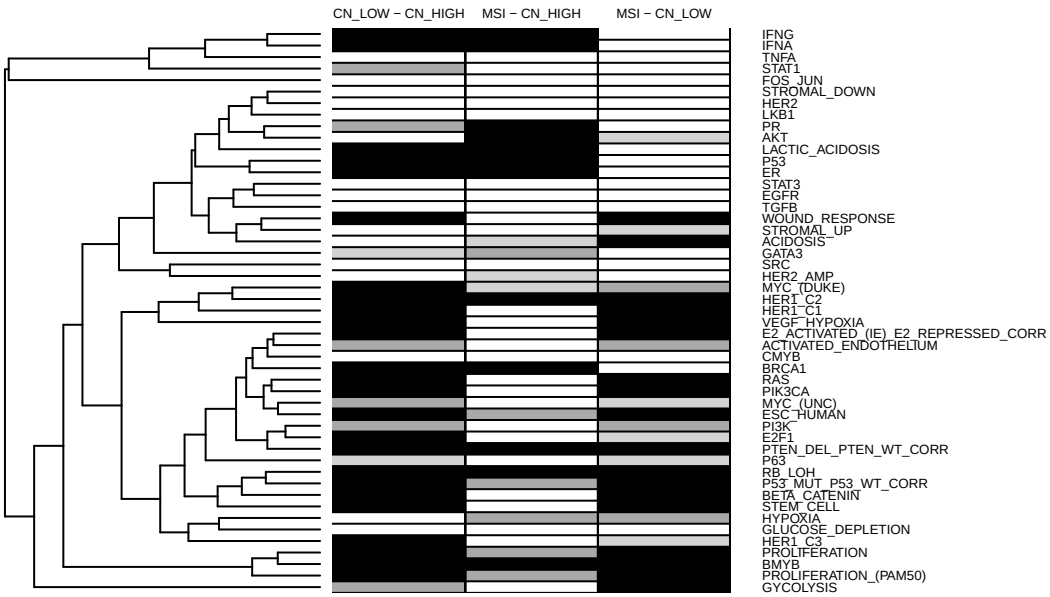
We can now apply this to all the other signatures with a loop and store the p values in a new matrix.

```r
p.difference <- matrix(0L, nrow = 49, ncol = 3) #creating the matrix

rownames(p.difference) <- row.names(ges_tot) #naming its rows after the signatures

for (signature in row.names(ges_tot)){
  al = aov(ges_tot[signature,] ~ subtyping[names(ges_tot[signature,])])
  #doing the test
  tukey <- TukeyHSD(al)
  #storing the result in a data frame for accessibility
  tukey.result <- data.frame(TukeyHSD(al)$subtyping)
  #extracting the p values
  p.values <- tukey.result$p.adj
  #adding the p values to the matrix
  p.difference[signature,] <- p.values
}

#naming the columns after the subtype comparisons
colnames(p.difference) <- c("CN_LOW - CN_HIGH", "MSI - CN_HIGH", "MSI - CN_LOW")
```

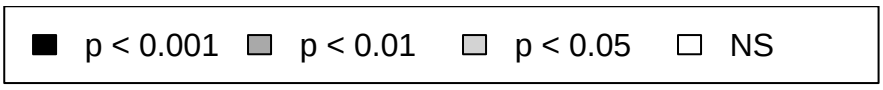
Now that this is done, we can display the p values in a heatmap.

```r
heatmap.2(p.difference,
          trace='none',
          Colv = FALSE,
          dendrogram = 'row',
          Rowv = as.dendrogram(gene_clust),
          breaks = c(0.00,0.001,0.01,0.05,1),
          col=colorRampPalette(c('black', 'dark grey', "light grey", "white")),
          colsep=1:ncol(p.difference),
          rowsep=1:nrow(p.difference),
          sepcolor="black",
          sepwidth=c(0.001,0.001),
```

```
          cexCol = 0.5,
          cexRow = 0.5,
          srtCol = 0,
          margins = c(2,16),
          adjCol = c(0.5,-63),
          key = FALSE,
          main = "UCEC, p values")
```




```
#plotting the legend
plot.new()
legend("top",
       legend = c("p < 0.001", "p < 0.01", "p < 0.05", "NS"),
       fill = c("black", "dark grey", "light grey", "white"),
       border = "black",
       ncol = 4)
```



The p values inform us on the significance of the difference in expression between one subtype and the other and will help us confirm whether some signatures truly are up- or downregulated in certain subtypes.

Discussion

Looking at both the heatmap with subtypes and the statistics allows us to determine which gene expression signatures are upregulated in a certain subtype by testing wheter the difference of expression with both other subtypes are significative ($p < 0.05$).

Some GES that are upregulated in the CN_HIGH subtype : * IFNG * IFNA * HER1_C2 * RBH_LOH *

P53_MUT_P53_WT_CORR * PROLIFERATION * PROLIFERATION_(PAM50) * BMYB

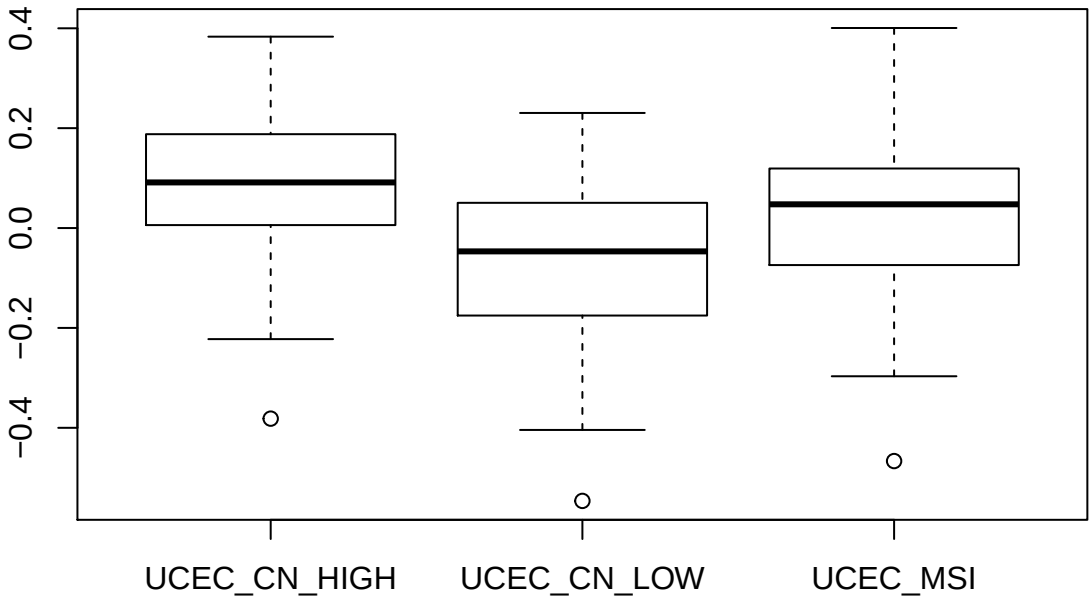
Several of these subtypes (all but IFNG and IFNA actually), along with others (HER1_C1, VEGF_HYPOXIA, BETA_CATENIN, STEM_CELL, etc), also seem to be downregulated in the CN_LOW subtype.

Others are upregulated in the MSI subtype : * PROLIFERATION * BMYB * PROLIFERATION_(PAM50) * RB_LOH * P53_MUT_P53_WT_CORR

However, many of these up- and downregulations seem low, although significant.

We can plot the GES score distribution between subtypes :

```
ges_mean <- apply(ges_tot,2,mean) #calculating the mean score for each sample
boxplot(ges_mean~subtyping) #plotting them in function of the subtypes
```



We can see that although there seems to be a slight tendency for upregulation in CN_HIGH tumors and downregulation in CN_LOW tumor, there is a lot of overlap between the boxplots and we cannot conclude that there is any general signifcant tendency in any subtype.

Let's take a look at one of the signatures. BMYB, for example, plays a role in cell proliferation. This explains why it can be upregulated in the CN_HIGH subtype, where there is a high level of transcription and the cell cycle is deregulated (Kandoth, Schultz, et al. 2013). It is probable that most other signatures are linked to cell proliferation or cell cycle.

This kind of analyses are useful to detect patterns between GES and subtypes and therefore help us predict which symptoms a certain type of cancer can cause, or which treatment is more appropriate, etc. They therefore play a key role in research against cancer.

Bibliography

- Gatza, M. L. et al. (2014) An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. Nature Genetics, 46(10), 1051–1059
- Kandoth C, Schultz N, et al. (2013) Integrated genomic characterization of endometrial carcinoma. Nature. 497(7447):67-73.
- Parker, J. S. et al. (2009) Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology, 27(8), 1160–1167.