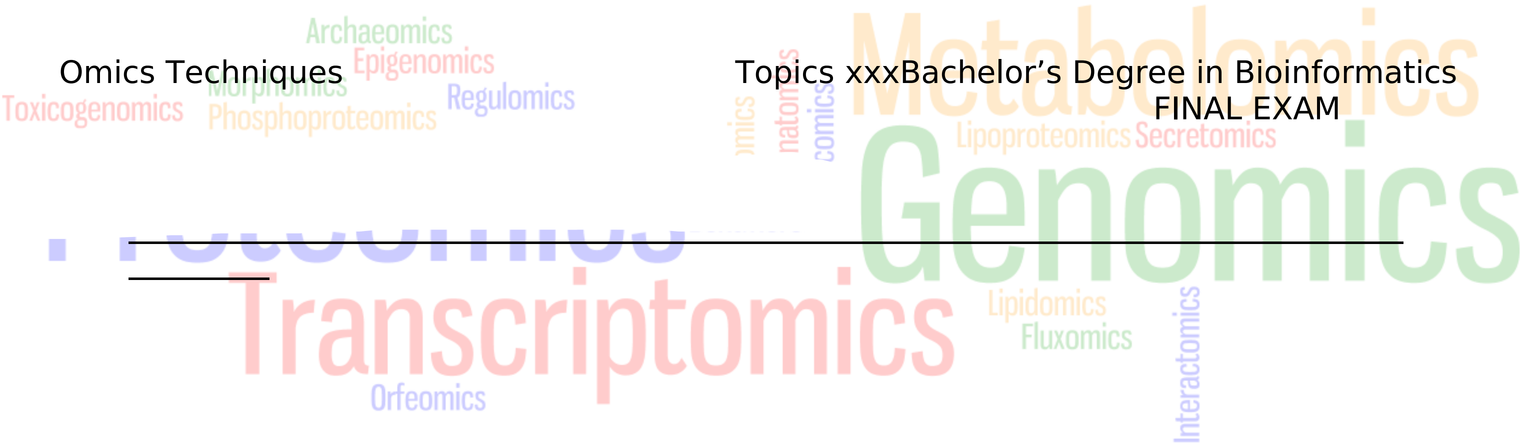


Part II. Bioconductor, Experimental design for omics studies and differential expression analysis

1. The code below illustrates a typical use of Bioconductor packages for microarray analysis. Indeed it is something you have used in class. Explain briefly (not more than 2-4 lines) what does each code chunk do

```
### chunk 1

      library(GEOquery)

      gse <- getGEO("GSE12345")

      esetFromGEO <- gse[[1]]
```

Call the GEOquery library.

Assign to the gse varialbe the dataset obtained with the getGEO function from the GEOquery library.

Assign to the esetFromGEO variable the previously obtained dataset but removing its first line, which contained the zip name of the file and can cause problems in the later analysis.

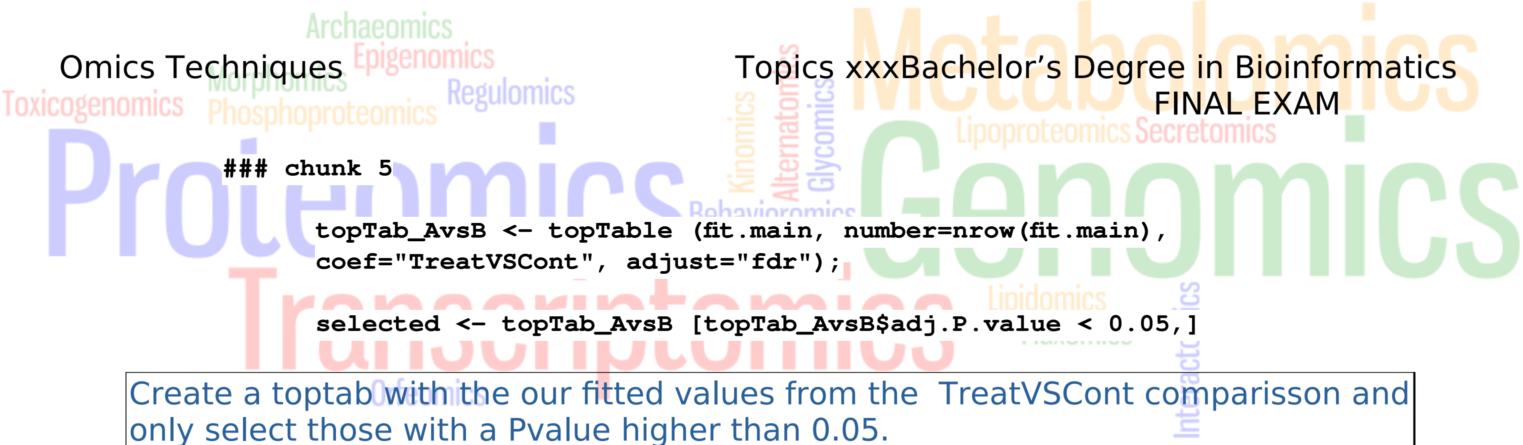
```
### chunk 3

      design_matrix <- model.matrix(~0+targets$groups)

      cont.matrix <- limma::makeContrasts(TreatVSCont, levels=design_matrix)
```

With the model.matrix function, create a design matrix taking in account our targets and the groups we have assigned.

From the limma package, call the makeConstast function and create a contrast matrix for the comparison of TreatVSCont using values from the previously created design matrix.



```
### chunk 5

        topTab_AvsB <- topTable (fit.main, number=nrow(fit.main),
        coef="TreatVSCont", adjust="fdr");

        selected <- topTab_AvsB [topTab_AvsB$adj.P.value < 0.05,]
```

Create a toptab with the our fitted values from the TreatVSCont comparisson and only select those with a Pvalue higher than 0.05.

2. A researcher has done a microarray analysis to compare gene expression between healthy donors with patients affected from endometriosis in three types of populations. She has done 12 microarrays (or RN-seq runs, the design would be the same) 6 from patients with endometriosis and 6 from healthy donors. In each group there are 3 different cell populations call them A, B, and C.

   The researcher wishes to make the following comparisons: (i) Compare healthy vs affected in each population. (ii) Compare population A vs B in affected patients.

   1. Imagine Samples are named as [H, A][A,B,D][1,2,3]. Write down an appropriate "targets" table to describe the experimental layout for this study

   2.Write down the design matrix associated with this study design. **HINT**: Do as we have done in class and combine the two main factors into a single one

   3.Build the contrast matrix needed to do the comparisons described above.

TARGETS

HA1	HA2	HA3
HB1	HB2	HB3
HC1	HC2	HC3
AA1	AA2	AA3
AB1	AB2	AB3
AC1	AC2	AC3

DESIGN MATRIX

H=0; A=1	A=0; B=1;
	C=2
		1=0; 2=1;
		3=2

0	0	0
0	0	1
0	0	2
0	1	0
0	1	1
0	1	2
0	2	0
0	2	1

0	2	2
1	0	0
1	0	1
1	0	2
1	1	0
1	1	1
1	1	2
1	2	0
1	2	1
1	2	2

CONTRAST MATRIX

(doing everything in the same population)

α1 = E(logHA1) ; α4 = E(logAA1)

α2 = E(logHB1) ; α5= E(logAB1)

α3 = E(logHC1) ; α6 = E(logAC1)

COMPARISSONS 1: Healthy vs Affected

$$\begin{pmatrix} \beta_1^1 \\ \beta_2^1 \\ \beta_3^1 \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

α1 vs α4
α2 vs α5
α3 vs α6

COMPARISSONS 2: A vs B in affected patients

λ1 = E(logAA1)

λ2= E(logAB1)

$$\begin{pmatrix} \beta_1^2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

3. Using the previously defined matrices we have fitted a linear model to the data and have obtained one top tables (one for each comparison). Table below shows the results for a few genes arbitrarily selected

```
ID             adj.P.Val      P.Value         t         B      logFC    Gene.Symbol
```

```
201282_at     0.0074911    1.281663e-05  6.773810 -4.468046  3.031999      OGDH
217232_x_at   0.0088708    5.409429e-05  5.868340 -4.477733  2.969720      HBB
210719_s_at   0.0097080    6.672291e-05  5.742075 -4.479320  3.396428      HMG20B
235361_at     0.0187080    7.161377e-05  5.699805 -4.479866  3.382755      STAMBP
207092_at     0.0339791    1.693045e-04  5.197041 -4.486967  4.011486      LEP
228499_at     0.0459791    1.903590e-04  5.130082 -4.488003  2.510947      PFKFB4
235719_at     0.0907918    2.133572e-04 -5.065256 -4.489028 -2.156150      CYP4V2
230445_at     0.0997918    2.736306e-04  4.924930 -4.491321  2.381681      BTBD17
1569386_at    0.1299791    3.180007e-04 -4.840876 -4.492746 -2.210268
215388_s_at   0.2397918    3.454539e-04 -4.794777 -4.493543 -3.641796      CFHR1
```

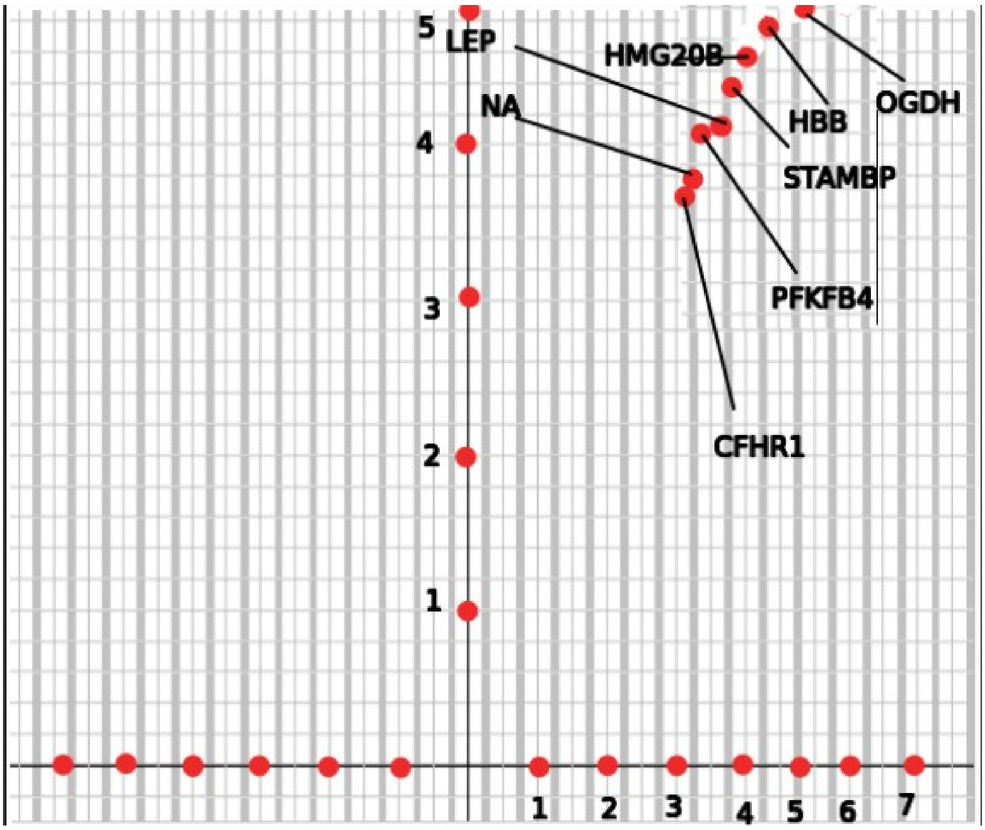
a. Which genes would you call differentially expressed? Explain the criteria used to make this.

The most differentially expressed genes are the one without name and CFHR1 because are the ones with smaller P-value.

b. Draft an approximate Volcano Plot depicting all the genes. Explain what does the Volcano plot represent and why is it interesting (**HINT: B ~ -log(P.Value)**)

Volcano plot represent the P-values fro the expression of our genes.

It is interesting because it is a very visual way for knowing with are the most expressed genes in our study.


LEP
HMG20B
NA
HBB
OGDH
STAMBP
PFKFB4
CFHR1