

Group work 2 - part a

The file [raw_trascr_count.txt](#) is a tab delimited file that you can download from moodle.

It contains raw count data of human bone marrow samples (leukemia and control samples). Leukemia patients are divided in 2 groups: asymptomatic (group 1) and symptomatic (group 2).

The file [raw_trascr_count_annot.txt](#) provides additional information about the different genes, including their length.


You are asked to submit an R script which implement 3 functions described in the next slides.

NOTE:

- the name of the function and of the name of the input parameters is MANDATORY. Do not change it!
- The R script must not contain any other code but the 3 functions

MvAplot

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise



The first function:


```
MvAplot <- function(exprData,pdffilename) {  
  BODY OF THE FUNCTION  
}
```

1. Take as input
 - **exprData**: a numeric data matrix that is supposed to have the same format of the raw transcript data provided in stem (i.e. genes'IDs on rows and subjects'IDs on columns)
 - **pdffilename** : the name of the .pdf file where to save the plots
2. Generates a .pdf file with MvA plots of each sample vs. sample 1

TMMnorm

The second function:

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise



```
TMMnorm <- function(exprData, annot, Atrim = c(0,8), Mtrim=0.02) {  
  BODY OF THE FUNCTION  
}
```

1. Takes as input

- **exprData**: a numeric data matrix that is supposed to have the same format of the raw transcript data provided in stem (i.e. genes'IDs on rows and subjects'IDs on columns)
- **annot**: a data frame that is supposed to have the same format of the annotation data provided in stem
- **Atrim**: a vector of 2 elements indicating the lower and upper thresholds to trim the most extreme values of A (where A is the average in log2 scale)
- **Mtrim**: a number between 0 and 1 indicating the fraction of observations to be deleted from each end (positive and negative values) of the sorted vector M before calculating the mean (where Ms are the log ratios defined as in the MvA plot)

TMMnorm

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise

TMMnorm <- function(exprData, annot, Atrim = c(0,8), Mtrim=0.02) {
 BODY OF THE FUNCTION
}

2. The function TMMnorm

- Scale the data by their sequencing depth and multiply by 10^6
- Calculates the scaling factors SF (with respect to sample 1) by trimming the most extreme values of A and taking the trimmed means of M values (suggestion: use the R function mean)
- Normalizes the data by their scaling factors SF with respect to sample 1
- Scale the genes (in the original scale, not in log scale) by their length and multiply by 10^3
- Returns a list of 2 elements: i) the vector of the scaling factors; ii) the normalized matrix (in the original scale, not in log scale)

Suggestion: Generates a .pdf file with MvA plots before and after the normalization to check your work (use the function MvAplot)

Group work 2 - part b

You are asked to perform the differential expression (DE) analysis using edgeR.
In particular you have to:


1. Implement an appropriate design matrix to inform edgeR about the group labels
2. Calculate the p-values (not corrected for multiple testing) between all patients (group 1 and 2) and controls.
3. Calculate the expected number of false positives (FP), false negatives (FN) and the FDR in correspondence to the given choice of alpha. **Suggestion: be careful about estimates... negative numbers are not meaningful.**
4. Calculate q-values using Benjamini-Hockberg procedure.

You are not allowed to use the FDR obtained using edgeR. You have to compute it!

DEbyEdgeR

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise

The third function:



```
DEbyEdgeR <- function(rawdat, groups, alpha = 0.05) {  
  BODY OF THE FUNCTION  
}
```


1. Takes as input

- **rawdat**: a numeric data matrix that is supposed to have the same format of the raw transcript data provided in stem (i.e. genes'IDs on rows and subjects'IDs on columns)
- **groups**: a vector of labels corresponding to the rawdat columns labels
- **alpha**: the significance level (default 0.05)

DEbyEdgeR

The third function:

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise



```
DEbyEdgeR <- function(rawdat, groups, alpha = 0.05) {  
  BODY OF THE FUNCTION  
}
```

2. Gives as output a list of two elements

- A vector of 4 elements with: i) the number of selected genes in correspondence to the significance level α ; ii) the corresponding estimate of the expected number of false positives (FP) considering $G_0 = 0.8 * G$ (G is the total number of analyzed genes); iii) the corresponding estimate of the expected number of false negatives (FN); the corresponding FDR
- A matrix with p values and q values (in this case obtained with the Benjamini Hockberg procedure) of each gene in the columns. The matrix must have row names corresponding to the gene names

Suggestion: EdgeR takes raw data in input (not the normalized ones).

I will test your solution

Only one student from each group has to submit the code on behalf of his/her mates as representative of the group.

I will test your code with different inputs... this is not a programming course so **a valid submission means that your code must be submitted fully working**... I am not going to correct coding and programming bugs...

Therefore, before submitting it, try running your code placing the input files in a directory different from the working directory and cleaning the workspace before running your functions.

To assign a grade I will evaluate if the MvA plots, the normalization and the DE tests perform correctly and give appropriate and reasonable results