

PAC2

Carles M. Bosch Herrera

28/5/2020

Índice

1. Entorno de trabajo y lectura de ficheros
2. Elección de datos de cada grupo de targets
3. Extracción aleatoria de muestras de cada grupo
4. Coincidencias con el archivo counts.csv
5. Packages
6. Lectura de datos y filtrado y eliminación de genes con contaje bajo
7. Counts to DGEList object
8. Quality Control
9. Diagrama de barras de los library sizes
10. Diagrama de cajas
11. Multidimensional scaling plots
12. Hierarchical clustering with heatmaps
13. Normalization for "composition bias"
14. Differential expression with limma-voom
15. Create the design matrix
16. Testing for differential expression
17. Annotation and saving the results
18. Volcano plot
19. Heatmap
20. Referencias

El objetivo de esta PEC es ilustrar el proceso de análisis de datos de ultrasecuenciación mediante la realización de un estudio, tal como se llevaria a cabo en una situación real. La PEC se basa en los datos suministrados de un estudio del que se debe extraer una muestra aleatoria con el fin de garantizar que cada conjunto de datos sea distinto.

1. Entorno de trabajo y lectura de ficheros

Creamos el entorno de trabajo y leemos el fichero targets.csv

```
setwd("C:/Users/CarlesM/Desktop/pac2")  
getwd()
```

```
## [1] "C:/Users/CarlesM/Desktop/pac2"
```

```
targets<- read.csv2(file.path("./data", "targets.csv"), head=T, sep=",")  
head(targets,5)
```

```
##      Experiment SRA_Sample      Sample_Name Grupo_analisis body_site
## 1  SRX567480   SRS626942 GTEX-111CU-0226-SM-5GZXC          1  Thyroid
## 2  SRX615964   SRS644174 GTEX-111FC-1026-SM-5GZX1          1  Thyroid
## 3  SRX563960   SRS625636 GTEX-111VG-0526-SM-5N9BW          3  Thyroid
## 4  SRX564185   SRS625665 GTEX-111YS-0726-SM-5GZY8          1  Thyroid
## 5  SRX559141   SRS624025 GTEX-1122O-0226-SM-5N9DA          1  Thyroid
##      molecular_data_type      sex Group ShortName
## 1 Allele-Specific Expression  male   NIT 111CU_NIT
## 2              RNA Seq (NGS)  male   NIT 111FC_NIT
## 3              RNA Seq (NGS)  male   ELI 111VG_ELI
## 4 Allele-Specific Expression  male   NIT 111YS_NIT
## 5              RNA Seq (NGS) female   NIT 1122O_NIT
```

2. Elección de datos de cada grupo de targets

Escogemos los datos de cada grupo del archivo targets

```
datos_NIT<-targets[targets$Group=="NIT",]
head(datos_NIT,5)
```

```
##      Experiment SRA_Sample      Sample_Name Grupo_analisis body_site
## 1  SRX567480   SRS626942 GTEX-111CU-0226-SM-5GZXC          1  Thyroid
## 2  SRX615964   SRS644174 GTEX-111FC-1026-SM-5GZX1          1  Thyroid
## 4  SRX564185   SRS625665 GTEX-111YS-0726-SM-5GZY8          1  Thyroid
## 5  SRX559141   SRS624025 GTEX-1122O-0226-SM-5N9DA          1  Thyroid
## 6  SRX561718   SRS625313 GTEX-1128S-0126-SM-5H12S          1  Thyroid
##      molecular_data_type      sex Group ShortName
## 1 Allele-Specific Expression  male   NIT 111CU_NIT
## 2              RNA Seq (NGS)  male   NIT 111FC_NIT
## 4 Allele-Specific Expression  male   NIT 111YS_NIT
## 5              RNA Seq (NGS) female   NIT 1122O_NIT
## 6 Allele-Specific Expression female   NIT 1128S_NIT
```

```
datos_ELI<-targets[targets$Group=="ELI",]
head(datos_ELI,5)
```

```
##      Experiment SRA_Sample      Sample_Name Grupo_analisis body_site
## 3  SRX563960   SRS625636 GTEX-111VG-0526-SM-5N9BW          3  Thyroid
## 29 SRX628009   SRS648152 GTEX-11NV4-0626-SM-5N9BR          3  Thyroid
## 40 SRX619829   SRS644736 GTEX-11XUK-0226-SM-5EQLW          3  Thyroid
## 100 SRX582762   SRS631169 GTEX-13NZ9-1126-SM-5MR37          3  Thyroid
## 119 SRX601511   SRS638114 GTEX-13QJC-0826-SM-5RQKC          3  Thyroid
##      molecular_data_type      sex Group ShortName
## 3              RNA Seq (NGS)  male   ELI 111VG_ELI
## 29              RNA Seq (NGS)  male   ELI 11NV4_ELI
## 40              RNA Seq (NGS) female   ELI 11XUK_ELI
## 100              RNA Seq (NGS)  male   ELI 13NZ9_ELI
## 119 Allele-Specific Expression female   ELI 13QJC_ELI
```

```
datos_SFI<-targets[targets$Group=="SFI",]
head(datos_SFI,5)
```

```
##      Experiment SRA_Sample      Sample_Name Grupo_analisis body_site
## 9      SRX557750  SRS623875  GTEX-117YW-0126-SM-5EGGN          2  Thyroid
## 14     SRX578169  SRS629611  GTEX-11DXY-0426-SM-5H12R          2  Thyroid
## 21     SRX619524  SRS644703  GTEX-11EQ8-0826-SM-5N9FG          2  Thyroid
## 22     SRX558144  SRS623916  GTEX-11EQ9-0626-SM-5A5K1          2  Thyroid
## 23     SRX567902  SRS627040  GTEX-11GS4-0826-SM-5986J          2  Thyroid
##      molecular_data_type  sex Group ShortName
## 9              RNA Seq (NGS) male   SFI 117YW_SFI
## 14              RNA Seq (NGS) male   SFI 11DXY_SFI
## 21 Allele-Specific Expression male   SFI 11EQ8_SFI
## 22              RNA Seq (NGS) male   SFI 11EQ9_SFI
## 23              RNA Seq (NGS) male   SFI 11GS4_SFI
```

3. Extracción aleatoria de muestras de cada grupo

Extraemos las 10 muestras aleatoriamente de cada grupo

```
muestra.NIT = datos_NIT[sample(nrow(datos_NIT),10) , ]
muestra.SFI = datos_SFI[sample(nrow(datos_SFI),10) , ]
muestra.ELI = datos_ELI[sample(nrow(datos_ELI),10) , ]
```

```
muestra.NIT
```

```
##      Experiment SRA_Sample      Sample_Name Grupo_analisis body_site
## 187  SRX203936  SRS374928  GTEX-REY6-0526-SM-2TF5M          1  Thyroid
## 277  SRX615716  SRS644145  GTEX-ZV7C-0126-SM-57WDE          1  Thyroid
## 230  SRX415902  SRS525220  GTEX-XGQ4-0426-SM-4AT4I          1  Thyroid
## 240  SRX564506  SRS625717  GTEX-Y3IK-0526-SM-4WWE3          1  Thyroid
## 71   SRX604905  SRS639191  GTEX-132NY-1026-SM-5P9IY          1  Thyroid
## 125  SRX607930  SRS639555  GTEX-13W46-0926-SM-5LU3T          1  Thyroid
## 234  SRX408795  SRS524547  GTEX-XUZC-0126-SM-4BOO6          1  Thyroid
## 27   SRX598448  SRS637651  GTEX-11NSD-0126-SM-5987F          1  Thyroid
## 217  SRX261799  SRS408833  GTEX-WHSB-1626-SM-3LK6J          1  Thyroid
## 180  SRX203980  SRS374948  GTEX-QV44-0826-SM-2S1RG          1  Thyroid
##      molecular_data_type  sex Group ShortName
## 187 Allele-Specific Expression  male   NIT REY6-_NIT
## 277              RNA Seq (NGS)  male   NIT ZV7C-_NIT
## 230              RNA Seq (NGS)  male   NIT XGQ4-_NIT
## 240 Allele-Specific Expression female  NIT Y3IK-_NIT
## 71                RNA Seq (NGS)  male   NIT 132NY_NIT
## 125               RNA Seq (NGS)  male   NIT 13W46_NIT
## 234               RNA Seq (NGS) female  NIT XUZC-_NIT
## 27                RNA Seq (NGS)  male   NIT 11NSD_NIT
## 217 Allele-Specific Expression  male   NIT WHSB-_NIT
## 180 Allele-Specific Expression  male   NIT QV44-_NIT
```

```
muestra.SFI
```

##	Experiment	SRA_Sample	Sample_Name	Grupo_analisis	body_site
## 69	SRX572541	SRS627900	GTEX-131YS-0726-SM-5P9G9	2	Thyroid
## 58	SRX601535	SRS638117	GTEX-12ZZX-1226-SM-5EGHS	2	Thyroid
## 59	SRX567279	SRS626844	GTEX-12ZZY-0826-SM-5EQMT	2	Thyroid
## 67	SRX561789	SRS625323	GTEX-131XG-0226-SM-5IFG1	2	Thyroid
## 22	SRX558144	SRS623916	GTEX-11EQ9-0626-SM-5A5K1	2	Thyroid
## 188	SRX203696	SRS374817	GTEX-RM2N-0526-SM-2TF4N	2	Thyroid
## 79	SRX597654	SRS637303	GTEX-139UW-0126-SM-5KM1B	2	Thyroid
## 252	SRX630577	SRS648438	GTEX-YFCO-0326-SM-4W1ZP	2	Thyroid
## 21	SRX619524	SRS644703	GTEX-11EQ8-0826-SM-5N9FG	2	Thyroid
## 44	SRX576955	SRS629456	GTEX-1211K-0726-SM-5FQUW	2	Thyroid
##	molecular_data_type	sex	Group	ShortName	
## 69	Allele-Specific Expression	female	SFI	131YS_SFI	
## 58	RNA Seq (NGS)	female	SFI	12ZZX_SFI	
## 59	RNA Seq (NGS)	male	SFI	12ZZY_SFI	
## 67	RNA Seq (NGS)	female	SFI	131XG_SFI	
## 22	RNA Seq (NGS)	male	SFI	11EQ9_SFI	
## 188	RNA Seq (NGS)	male	SFI	RM2N-_SFI	
## 79	Allele-Specific Expression	male	SFI	139UW_SFI	
## 252	Allele-Specific Expression	male	SFI	YFCO-_SFI	
## 21	Allele-Specific Expression	male	SFI	11EQ8_SFI	
## 44	Allele-Specific Expression	female	SFI	1211K_SFI	

muestra.ELI

##	Experiment	SRA_Sample	Sample_Name	Grupo_analisis	body_site
## 211	SRX222429	SRS389623	GTEX-TMMY-0826-SM-33HB9	3	Thyroid
## 147	SRX607358	SRS639491	GTEX-14AS3-0226-SM-5Q5B6	3	Thyroid
## 149	SRX568916	SRS627158	GTEX-14BMU-0226-SM-5S2QA	3	Thyroid
## 29	SRX628009	SRS648152	GTEX-11NV4-0626-SM-5N9BR	3	Thyroid
## 167	SRX199272	SRS333099	GTEX-PLZ4-1226-SM-2I5FE	3	Thyroid
## 146	SRX575932	SRS629299	GTEX-14ABY-0926-SM-5Q5DY	3	Thyroid
## 290	SRX568364	SRS627095	GTEX-ZYY3-1926-SM-5GZXS	3	Thyroid
## 100	SRX582762	SRS631169	GTEX-13NZ9-1126-SM-5MR37	3	Thyroid
## 253	SRX583148	SRS631283	GTEX-YJ89-0726-SM-5P9F7	3	Thyroid
## 251	SRX615373	SRS644099	GTEX-YFC4-2626-SM-5P9FQ	3	Thyroid
##	molecular_data_type	sex	Group	ShortName	
## 211	Allele-Specific Expression	female	ELI	TMMY-_ELI	
## 147	RNA Seq (NGS)	female	ELI	14AS3_ELI	
## 149	Allele-Specific Expression	female	ELI	14BMU_ELI	
## 29	RNA Seq (NGS)	male	ELI	11NV4_ELI	
## 167	RNA Seq (NGS)	female	ELI	PLZ4-_ELI	
## 146	Allele-Specific Expression	male	ELI	14ABY_ELI	
## 290	Allele-Specific Expression	female	ELI	ZYY3-_ELI	
## 100	RNA Seq (NGS)	male	ELI	13NZ9_ELI	
## 253	RNA Seq (NGS)	male	ELI	YJ89-_ELI	
## 251	Allele-Specific Expression	female	ELI	YFC4-_ELI	

4. Coincidencias con el archivo counts.csv

Seleccionamos las columnas de counts que coincidan con la columna Sample-Name de los 30 targets y leemos el archivo resultante scounts. La elección de las columnas se ha hecho usando Excel

```
scounts<- read.csv2(file.path("./data", "selectcounts.csv"), head=T, sep=";")
str(scounts)
```

```
## 'data.frame':    56202 obs. of  31 variables:
##  $ X                      : Factor w/ 56202 levels "ENSG000000000003.10",...: 26352
28704 39144 36095 53325 37828 16363 36388 33329 36152 ...
##  $ GTEX.111VG.0526.SM.5N9BW: int   1 474 1 0 1 1 0 3 7 427 ...
##  $ GTEX.11EM3.0126.SM.5985K: int   2 669 2 1 1 1 0 3 20 791 ...
##  $ GTEX.11EMC.0226.SM.5EGLP: int   5 786 0 0 0 1 0 10 8 553 ...
##  $ GTEX.11NSD.0126.SM.5987F: int   0 408 1 0 0 0 2 11 19 800 ...
##  $ GTEX.11NV4.0626.SM.5N9BR: int   3 1301 1 0 0 1 0 5 7 1132 ...
##  $ GTEX.11O72.2326.SM.5BC7H: int   0 633 2 1 0 1 1 14 11 1075 ...
##  $ GTEX.12WSG.0226.SM.5EGIF: int   3 369 1 3 1 2 2 3 10 235 ...
##  $ GTEX.139UW.0126.SM.5KM1B: int   2 430 0 0 0 0 0 9 9 679 ...
##  $ GTEX.13NZ9.1126.SM.5MR37: int   0 1002 1 0 0 1 0 15 19 602 ...
##  $ GTEX.13O1R.0826.SM.5J2MB: int   3 460 0 1 2 0 1 7 12 279 ...
##  $ GTEX.13OVG.0226.SM.5LU93: int   4 719 2 1 2 2 1 6 14 1064 ...
##  $ GTEX.13QJC.0826.SM.5RQKC: int   0 825 1 0 0 1 1 10 21 853 ...
##  $ GTEX.13U4I.0526.SM.5LU59: int   2 636 0 0 0 0 0 8 13 606 ...
##  $ GTEX.14ABY.0926.SM.5Q5DY: int   1 775 2 0 0 0 1 10 2 580 ...
##  $ GTEX.14AS3.0226.SM.5Q5B6: int   0 834 1 1 0 0 0 6 6 445 ...
##  $ GTEX.14BMU.0226.SM.5S2QA: int   2 423 0 0 2 1 0 18 6 325 ...
##  $ GTEX.PWN1.2626.SM.2I3FH  : int   5 297 0 0 1 2 0 0 453 229 ...
##  $ GTEX.S7SE.0726.SM.2XCD7 : int   4 422 0 1 1 2 1 4 12 247 ...
##  $ GTEX.T5JW.1226.SM.3GACY  : int   1 541 2 0 0 0 1 1 9 1468 ...
##  $ GTEX.WYVS.0326.SM.3NM9V  : int   6 820 0 1 0 4 5 12 18 973 ...
##  $ GTEX.XBED.0126.SM.47JY7  : int   3 766 3 4 0 4 1 10 11 374 ...
##  $ GTEX.XMK1.0626.SM.4B65A  : int   9 568 1 1 1 0 1 5 14 738 ...
##  $ GTEX.Y5V6.0526.SM.4VBRV  : int   3 482 3 2 2 2 2 2 27 681 ...
##  $ GTEX.YEC4.0626.SM.5CVLU  : int   1 365 1 1 0 1 1 1 20 359 ...
##  $ GTEX.YFC4.2626.SM.5P9FQ  : int   1 1472 1 0 0 1 2 38 24 2020 ...
##  $ GTEX.YJ89.0726.SM.5P9F7  : int   4 1325 1 0 2 1 2 4 8 853 ...
##  $ GTEX.Z9EW.0226.SM.5CVM7  : int   3 450 2 2 0 1 0 2 10 352 ...
##  $ GTEX.ZLV1.0126.SM.4WWBZ  : int   2 689 2 4 0 2 0 18 9 809 ...
##  $ GTEX.ZYVF.1126.SM.5E458  : int   2 838 1 4 1 1 0 0 21 1212 ...
##  $ GTEX.ZYY3.1926.SM.5GZXS  : int   6 1003 1 2 0 1 4 8 12 960 ...
```

5. Packages

Ahora tendremos que cargar los paquetes que necesitaremos

```
library(edgeR)
```

```
## Loading required package: limma
```

```
library(limma)
```

```
library(Glimma)
```

```
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 3.6.3
```

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

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

```
library(org.Mm.eg.db)
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##  
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':  
##  
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,  
##     clusterExport, clusterMap, parApply, parCapply, parLapply,  
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following object is masked from 'package:limma':  
##  
##     plotMA
```

```
## The following objects are masked from 'package:stats':  
##  
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':  
##  
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,  
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,  
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,  
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,  
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,  
##     union, unique, unsplit, which, which.max, which.min
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase)", and for packages 'citation("pkgname)".
```

```
## Loading required package: IRanges
```

```
## Loading required package: S4Vectors
```

```
## Warning: package 'S4Vectors' was built under R version 3.6.3
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:gplots':
##
##     space
```

```
## The following object is masked from 'package:base':
##
##     expand.grid
```

```
##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:grDevices':
##
##     windows
```

```
##
```

```
library(org.Hs.eg.db)
```

```
##
```

```
library(RColorBrewer)
library(DESeq2)
```

```
## Loading required package: GenomicRanges
```

```
## Loading required package: GenomeInfoDb
```

```
## Warning: package 'GenomeInfoDb' was built under R version 3.6.3
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: DelayedArray
```

```
## Warning: package 'DelayedArray' was built under R version 3.6.3
```

```
## Loading required package: matrixStats
```

```
## Warning: package 'matrixStats' was built under R version 3.6.3
```

```
##  
## Attaching package: 'matrixStats'
```

```
## The following objects are masked from 'package:Biobase':  
##  
##      anyMissing, rowMedians
```

```
## Loading required package: BiocParallel
```

```
##  
## Attaching package: 'DelayedArray'
```

```
## The following objects are masked from 'package:matrixStats':  
##  
##      colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges
```

```
## The following objects are masked from 'package:base':  
##  
##      aperm, apply, rowsum
```

6. Lectura de datos y filtrado y eliminacion de genes con contajes bajos

Los genes con recuentos muy bajos en todas las bibliotecas proporcionan poca evidencia en la expresión diferencial e interfieren con algunas de las aproximaciones estadísticas que se utilizan más adelante dentro del pipeline del análisis.

Asimismo añaden “ruido” en el ajuste por múltiple testing mediante FDR, reduciendo “potencia estadística” en la detección de genes expresados diferencialmente (como ya hemos discutido en debates anteriores).

Estos genes deben filtrarse antes de un análisis posterior.

Hay diferentes maneras de filtrar genes poco expresados. En este caso optamos por retener los genes si se expresan en un conteo por millón (CPM) por encima de 0.5 en al menos dos muestras.

Utilizaremos la función `cpm` del package `edgeR` para generar los valores de CPM y luego filtrar. Hay que tener presente que al convertir a CPM estamos normalizando segun el “Sequencing depth” de cada muestra.

Nota: Sequencing depth es comúnmente un término usado para la secuenciación del genoma o del exoma y significa el número de lecturas que cubren cada posición.

```
rownames(scounts)<-scounts[,1]
scounts<-scounts[,-(1)]
library(edgeR)
dgeList_counts<-DGEList(scounts)
counts_cpm<-cpm(dgeList_counts,log=TRUE)
head(counts_cpm)
```

##	GTEX.111VG.0526.SM.5N9BW	GTEX.11EM3.0126.SM.5985K
##	ENSG00000223972.4	-4.162194 -3.890311
##	ENSG00000227232.4	3.188591 3.367278
##	ENSG00000243485.2	-4.162194 -3.890311
##	ENSG00000237613.2	-4.768198 -4.263485
##	ENSG00000268020.2	-4.162194 -4.263485
##	ENSG00000240361.1	-4.162194 -4.263485
##	GTEX.11EMC.0226.SM.5EGLP	GTEX.11NSD.0126.SM.5987F
##	ENSG00000223972.4	-3.149473 -4.768198
##	ENSG00000227232.4	3.583036 2.973675
##	ENSG00000243485.2	-4.768198 -4.162034
##	ENSG00000237613.2	-4.768198 -4.768198
##	ENSG00000268020.2	-4.768198 -4.768198
##	ENSG00000240361.1	-4.268205 -4.768198
##	GTEX.11NV4.0626.SM.5N9BR	GTEX.11O72.2326.SM.5BC7H
##	ENSG00000223972.4	-3.404865 -4.768198
##	ENSG00000227232.4	4.647684 3.214006
##	ENSG00000243485.2	-4.160086 -3.923594
##	ENSG00000237613.2	-4.768198 -4.284952
##	ENSG00000268020.2	-4.768198 -4.768198
##	ENSG00000240361.1	-4.160086 -4.284952
##	GTEX.12WSG.0226.SM.5EGIF	GTEX.139UW.0126.SM.5KM1B
##	ENSG00000223972.4	-3.128109 -3.676954
##	ENSG00000227232.4	3.261766 3.162977
##	ENSG00000243485.2	-3.997905 -4.768198
##	ENSG00000237613.2	-3.128109 -4.768198
##	ENSG00000268020.2	-3.997905 -4.768198
##	ENSG00000240361.1	-3.498427 -4.768198
##	GTEX.13NZ9.1126.SM.5MR37	GTEX.13O1R.0826.SM.5J2MB
##	ENSG00000223972.4	-4.768198 -3.538156
##	ENSG00000227232.4	4.030623 2.927702
##	ENSG00000243485.2	-4.238656 -4.768198
##	ENSG00000237613.2	-4.768198 -4.233559
##	ENSG00000268020.2	-4.768198 -3.844358
##	ENSG00000240361.1	-4.238656 -4.768198
##	GTEX.13OVG.0226.SM.5LU93	GTEX.13QJC.0826.SM.5RQKC
##	ENSG00000223972.4	-3.039814 -4.768198
##	ENSG00000227232.4	3.935232 4.081484
##	ENSG00000243485.2	-3.659318 -4.128516
##	ENSG00000237613.2	-4.109743 -4.768198
##	ENSG00000268020.2	-3.659318 -4.768198
##	ENSG00000240361.1	-3.659318 -4.128516
##	GTEX.13U4I.0526.SM.5LU59	GTEX.14ABY.0926.SM.5Q5DY
##	ENSG00000223972.4	-3.630268 -4.261134
##	ENSG00000227232.4	3.812274 3.586691
##	ENSG00000243485.2	-4.768198 -3.886682
##	ENSG00000237613.2	-4.768198 -4.768198
##	ENSG00000268020.2	-4.768198 -4.768198
##	ENSG00000240361.1	-4.768198 -4.768198
##	GTEX.14AS3.0226.SM.5Q5B6	GTEX.14BMU.0226.SM.5S2QA
##	ENSG00000223972.4	-4.768198 -3.614303
##	ENSG00000227232.4	4.313859 3.254818
##	ENSG00000243485.2	-4.046922 -4.768198
##	ENSG00000237613.2	-4.046922 -4.768198
##	ENSG00000268020.2	-4.768198 -3.614303
##	ENSG00000240361.1	-4.768198 -4.078837
##	GTEX.PWN1.2626.SM.2I3FH	GTEX.S7SE.0726.SM.2XCD7

##	ENSG00000223972.4	-2.620847	-2.922711
##	ENSG00000227232.4	2.909522	3.333199
##	ENSG00000243485.2	-4.768198	-4.768198
##	ENSG00000237613.2	-4.768198	-4.047100
##	ENSG00000268020.2	-4.014570	-4.047100
##	ENSG00000240361.1	-3.522061	-3.568528
##	GTEX.T5JW.1226.SM.3GACY	GTEX.WYVS.0326.SM.3NM9V	
##	ENSG00000223972.4	-4.101799	-3.121589
##	ENSG00000227232.4	3.547515	3.422774
##	ENSG00000243485.2	-3.647705	-4.768198
##	ENSG00000237613.2	-4.768198	-4.329733
##	ENSG00000268020.2	-4.768198	-4.768198
##	ENSG00000240361.1	-4.768198	-3.492806
##	GTEX.XBED.0126.SM.47JY7	GTEX.XMK1.0626.SM.4B65A	
##	ENSG00000223972.4	-3.316634	-2.275165
##	ENSG00000227232.4	4.026265	3.427442
##	ENSG00000243485.2	-3.316634	-4.169451
##	ENSG00000237613.2	-3.039889	-4.169451
##	ENSG00000268020.2	-4.768198	-4.169451
##	ENSG00000240361.1	-3.039889	-4.768198
##	GTEX.Y5V6.0526.SM.4VBRV	GTEX.YEC4.0626.SM.5CVLU	
##	ENSG00000223972.4	-3.637457	-4.147308
##	ENSG00000227232.4	2.817869	2.856089
##	ENSG00000243485.2	-3.637457	-4.147308
##	ENSG00000237613.2	-3.925709	-4.147308
##	ENSG00000268020.2	-3.925709	-4.768198
##	ENSG00000240361.1	-3.925709	-4.147308
##	GTEX.YFC4.2626.SM.5P9FQ	GTEX.YJ89.0726.SM.5P9F7	
##	ENSG00000223972.4	-4.350836	-3.461805
##	ENSG00000227232.4	4.182595	4.165509
##	ENSG00000243485.2	-4.350836	-4.315807
##	ENSG00000237613.2	-4.768198	-4.768198
##	ENSG00000268020.2	-4.768198	-3.971922
##	ENSG00000240361.1	-4.350836	-4.315807
##	GTEX.Z9EW.0226.SM.5CVM7	GTEX.ZLV1.0126.SM.4WWBZ	
##	ENSG00000223972.4	-3.310837	-3.780305
##	ENSG00000227232.4	3.270219	3.640098
##	ENSG00000243485.2	-3.654472	-3.780305
##	ENSG00000237613.2	-3.654472	-3.199401
##	ENSG00000268020.2	-4.768198	-4.768198
##	ENSG00000240361.1	-4.106431	-3.780305
##	GTEX.ZYVF.1126.SM.5E458	GTEX.ZYY3.1926.SM.5GZXS	
##	ENSG00000223972.4	-3.657650	-2.649670
##	ENSG00000227232.4	4.158799	4.360455
##	ENSG00000243485.2	-4.108604	-4.129351
##	ENSG00000237613.2	-3.037644	-3.688108
##	ENSG00000268020.2	-4.108604	-4.768198
##	ENSG00000240361.1	-4.108604	-4.129351

```
# Which values in myCPM are greater than 0.5?
thresh <- counts_cpm > 0.5
# This produces a logical matrix with TRUEs and FALSEs
head(thresh)
```

##	GTEX.111VG.0526.SM.5N9BW	GTEX.11EM3.0126.SM.5985K
## ENSG00000223972.4	FALSE	FALSE
## ENSG00000227232.4	TRUE	TRUE
## ENSG00000243485.2	FALSE	FALSE
## ENSG00000237613.2	FALSE	FALSE
## ENSG00000268020.2	FALSE	FALSE
## ENSG00000240361.1	FALSE	FALSE
##	GTEX.11EMC.0226.SM.5EGLP	GTEX.11NSD.0126.SM.5987F
## ENSG00000223972.4	FALSE	FALSE
## ENSG00000227232.4	TRUE	TRUE
## ENSG00000243485.2	FALSE	FALSE
## ENSG00000237613.2	FALSE	FALSE
## ENSG00000268020.2	FALSE	FALSE
## ENSG00000240361.1	FALSE	FALSE
##	GTEX.11NV4.0626.SM.5N9BR	GTEX.11O72.2326.SM.5BC7H
## ENSG00000223972.4	FALSE	FALSE
## ENSG00000227232.4	TRUE	TRUE
## ENSG00000243485.2	FALSE	FALSE
## ENSG00000237613.2	FALSE	FALSE
## ENSG00000268020.2	FALSE	FALSE
## ENSG00000240361.1	FALSE	FALSE
##	GTEX.12WSG.0226.SM.5EGIF	GTEX.139UW.0126.SM.5KM1B
## ENSG00000223972.4	FALSE	FALSE
## ENSG00000227232.4	TRUE	TRUE
## ENSG00000243485.2	FALSE	FALSE
## ENSG00000237613.2	FALSE	FALSE
## ENSG00000268020.2	FALSE	FALSE
## ENSG00000240361.1	FALSE	FALSE
##	GTEX.13NZ9.1126.SM.5MR37	GTEX.13O1R.0826.SM.5J2MB
## ENSG00000223972.4	FALSE	FALSE
## ENSG00000227232.4	TRUE	TRUE
## ENSG00000243485.2	FALSE	FALSE
## ENSG00000237613.2	FALSE	FALSE
## ENSG00000268020.2	FALSE	FALSE
## ENSG00000240361.1	FALSE	FALSE
##	GTEX.13OVG.0226.SM.5LU93	GTEX.13QJC.0826.SM.5RQKC
## ENSG00000223972.4	FALSE	FALSE
## ENSG00000227232.4	TRUE	TRUE
## ENSG00000243485.2	FALSE	FALSE
## ENSG00000237613.2	FALSE	FALSE
## ENSG00000268020.2	FALSE	FALSE
## ENSG00000240361.1	FALSE	FALSE
##	GTEX.13U4I.0526.SM.5LU59	GTEX.14ABY.0926.SM.5Q5DY
## ENSG00000223972.4	FALSE	FALSE
## ENSG00000227232.4	TRUE	TRUE
## ENSG00000243485.2	FALSE	FALSE
## ENSG00000237613.2	FALSE	FALSE
## ENSG00000268020.2	FALSE	FALSE
## ENSG00000240361.1	FALSE	FALSE
##	GTEX.14AS3.0226.SM.5Q5B6	GTEX.14BMU.0226.SM.5S2QA
## ENSG00000223972.4	FALSE	FALSE
## ENSG00000227232.4	TRUE	TRUE
## ENSG00000243485.2	FALSE	FALSE
## ENSG00000237613.2	FALSE	FALSE
## ENSG00000268020.2	FALSE	FALSE
## ENSG00000240361.1	FALSE	FALSE
##	GTEX.PWN1.2626.SM.2I3FH	GTEX.S7SE.0726.SM.2XCD7

##	ENSG00000223972.4	FALSE	FALSE
##	ENSG00000227232.4	TRUE	TRUE
##	ENSG00000243485.2	FALSE	FALSE
##	ENSG00000237613.2	FALSE	FALSE
##	ENSG00000268020.2	FALSE	FALSE
##	ENSG00000240361.1	FALSE	FALSE
##	GTEX.T5JW.1226.SM.3GACY	GTEX.WYVS.0326.SM.3NM9V	
##	ENSG00000223972.4	FALSE	FALSE
##	ENSG00000227232.4	TRUE	TRUE
##	ENSG00000243485.2	FALSE	FALSE
##	ENSG00000237613.2	FALSE	FALSE
##	ENSG00000268020.2	FALSE	FALSE
##	ENSG00000240361.1	FALSE	FALSE
##	GTEX.XBED.0126.SM.47JY7	GTEX.XMK1.0626.SM.4B65A	
##	ENSG00000223972.4	FALSE	FALSE
##	ENSG00000227232.4	TRUE	TRUE
##	ENSG00000243485.2	FALSE	FALSE
##	ENSG00000237613.2	FALSE	FALSE
##	ENSG00000268020.2	FALSE	FALSE
##	ENSG00000240361.1	FALSE	FALSE
##	GTEX.Y5V6.0526.SM.4VBRV	GTEX.YEC4.0626.SM.5CVLU	
##	ENSG00000223972.4	FALSE	FALSE
##	ENSG00000227232.4	TRUE	TRUE
##	ENSG00000243485.2	FALSE	FALSE
##	ENSG00000237613.2	FALSE	FALSE
##	ENSG00000268020.2	FALSE	FALSE
##	ENSG00000240361.1	FALSE	FALSE
##	GTEX.YFC4.2626.SM.5P9FQ	GTEX.YJ89.0726.SM.5P9F7	
##	ENSG00000223972.4	FALSE	FALSE
##	ENSG00000227232.4	TRUE	TRUE
##	ENSG00000243485.2	FALSE	FALSE
##	ENSG00000237613.2	FALSE	FALSE
##	ENSG00000268020.2	FALSE	FALSE
##	ENSG00000240361.1	FALSE	FALSE
##	GTEX.Z9EW.0226.SM.5CVM7	GTEX.ZLV1.0126.SM.4WWBZ	
##	ENSG00000223972.4	FALSE	FALSE
##	ENSG00000227232.4	TRUE	TRUE
##	ENSG00000243485.2	FALSE	FALSE
##	ENSG00000237613.2	FALSE	FALSE
##	ENSG00000268020.2	FALSE	FALSE
##	ENSG00000240361.1	FALSE	FALSE
##	GTEX.ZYVF.1126.SM.5E458	GTEX.ZYY3.1926.SM.5GZXS	
##	ENSG00000223972.4	FALSE	FALSE
##	ENSG00000227232.4	TRUE	TRUE
##	ENSG00000243485.2	FALSE	FALSE
##	ENSG00000237613.2	FALSE	FALSE
##	ENSG00000268020.2	FALSE	FALSE
##	ENSG00000240361.1	FALSE	FALSE

```
# Summary of how many TRUEs there are in each row
# There are 13142 genes that have TRUEs in all 30 samples.
table(rowSums(thresh))
```

```
##
##      0      1      2      3      4      5      6      7      8      9     10     11     12
## 37003  825   439   332   294   244   211   174   164   180   140   154   125
##      13     14     15     16     17     18     19     20     21     22     23     24     25
##    130    139    136    142    125    124    121    113    138    108    117    148    161
##      26     27     28     29     30
##    174    206    256    437 13142
```

```
# we would like to keep genes that have at least 2 TRUES in each row of thresh
keep <- rowSums(thresh) >= 2
# Subset the rows of countdata to keep the more highly expressed genes
counts.keep <- scounts[keep,]
summary(keep)
```

```
##      Mode      FALSE      TRUE
## logical    37828    18374
```

```
dim(counts.keep)
```

```
## [1] 18374    30
```

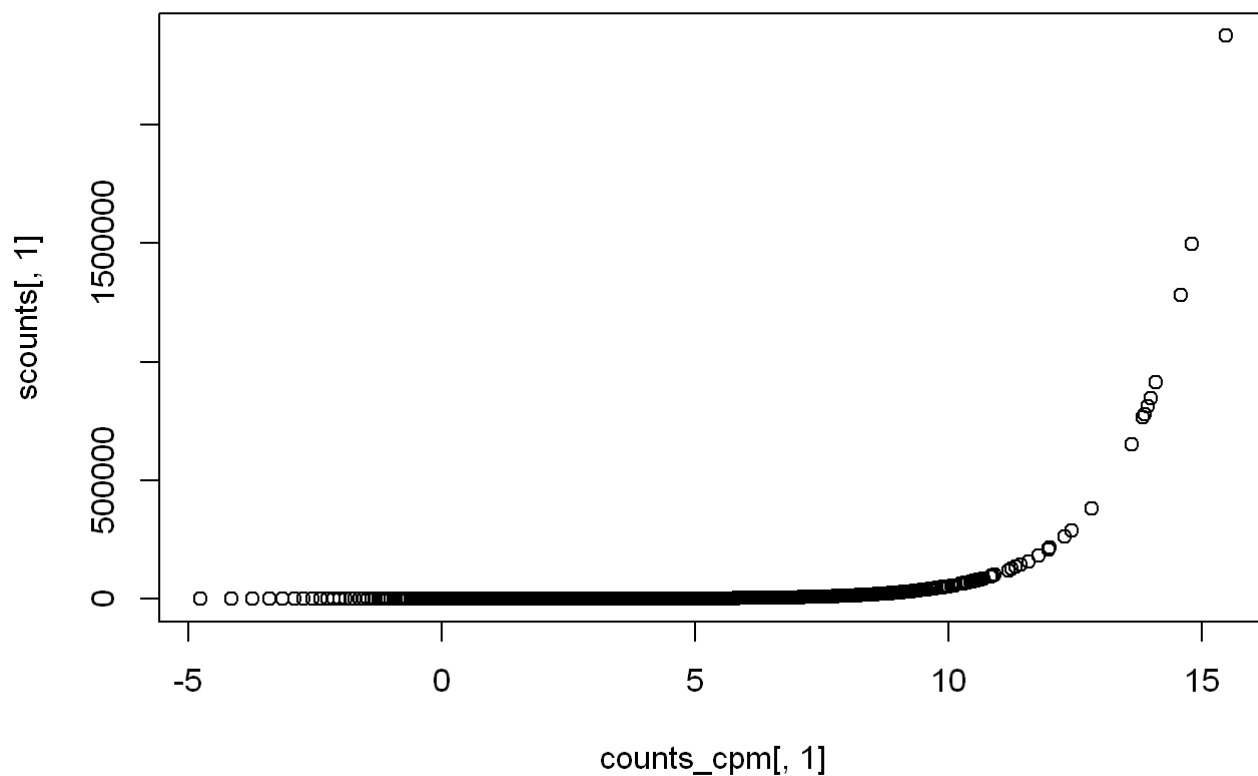
En este caso, se usa un CPM de 0.5 ya que corresponde a un “recuento por gen” de 10-15 segun los “library size” de este conjunto de datos.

Asimismo se utiliza la condición de que la la expresión sea en 2 o más “libraries” ya que en este caso cada situación experimental contiene dos replicas y ello nos “asegura” que “analizaremos” genes que como mínimo se expresen en un grupo.

Como regla general, se puede elegir un buen umbral identificando el CPM que corresponde a un recuento de 10.

Se debe filtrar a partir de el objeto CPM en lugar de filtrar los counting data (recuentos directamente), ya que este último no tiene en cuenta las diferencias en los tamaños de biblioteca (library sizes) entre las muestras.

```
# Let's have a look and see whether our threshold of 0.5 does indeed correspond to a
count of about 10-15
# We will look at the first sample
plot(counts_cpm[,1],scounts[,1])
```



7. Counts to DGEList object

A continuación crearemos un objeto DGEList. Este es un objeto utilizado por edgeR para almacenar datos de recuento

```
y <- DGEList(counts.keep)
# have a look at y
y
```

```

## An object of class "DGEList"
## $counts
##
## GTEX.111VG.0526.SM.5N9BW GTEX.11EM3.0126.SM.5985K
## ENSG00000227232.4 474 669
## ENSG00000237683.5 427 791
## ENSG00000241860.2 92 68
## ENSG00000228463.4 9 9
## ENSG00000225972.1 87 155
##
## GTEX.11EMC.0226.SM.5EGLP GTEX.11NSD.0126.SM.5987F
## ENSG00000227232.4 786 408
## ENSG00000237683.5 553 800
## ENSG00000241860.2 107 29
## ENSG00000228463.4 57 51
## ENSG00000225972.1 82 135
##
## GTEX.11NV4.0626.SM.5N9BR GTEX.11O72.2326.SM.5BC7H
## ENSG00000227232.4 1301 633
## ENSG00000237683.5 1132 1075
## ENSG00000241860.2 46 82
## ENSG00000228463.4 92 16
## ENSG00000225972.1 23 57
##
## GTEX.12WSG.0226.SM.5EGIF GTEX.139UW.0126.SM.5KM1B
## ENSG00000227232.4 369 430
## ENSG00000237683.5 235 679
## ENSG00000241860.2 21 95
## ENSG00000228463.4 6 26
## ENSG00000225972.1 101 54
##
## GTEX.13NZ9.1126.SM.5MR37 GTEX.13O1R.0826.SM.5J2MB
## ENSG00000227232.4 1002 460
## ENSG00000237683.5 602 279
## ENSG00000241860.2 96 41
## ENSG00000228463.4 39 16
## ENSG00000225972.1 4678 99
##
## GTEX.13OVG.0226.SM.5LU93 GTEX.13QJC.0826.SM.5RQKC
## ENSG00000227232.4 719 825
## ENSG00000237683.5 1064 853
## ENSG00000241860.2 91 94
## ENSG00000228463.4 41 98
## ENSG00000225972.1 64 55
##
## GTEX.13U4I.0526.SM.5LU59 GTEX.14ABY.0926.SM.5Q5DY
## ENSG00000227232.4 636 775
## ENSG00000237683.5 606 580
## ENSG00000241860.2 29 69
## ENSG00000228463.4 91 5
## ENSG00000225972.1 18 37
##
## GTEX.14AS3.0226.SM.5Q5B6 GTEX.14BMU.0226.SM.5S2QA
## ENSG00000227232.4 834 423
## ENSG00000237683.5 445 325
## ENSG00000241860.2 40 41
## ENSG00000228463.4 66 42
## ENSG00000225972.1 33 25
##
## GTEX.PWN1.2626.SM.2I3FH GTEX.S7SE.0726.SM.2XCD7
## ENSG00000227232.4 297 422
## ENSG00000237683.5 229 247
## ENSG00000241860.2 50 73
## ENSG00000228463.4 53 56
## ENSG00000225972.1 192 77
##
## GTEX.T5JW.1226.SM.3GACY GTEX.WYVS.0326.SM.3NM9V

```



```

## ENSG00000227232.4          541          820
## ENSG00000237683.5          1468         973
## ENSG00000241860.2           85         101
## ENSG00000228463.4           35          66
## ENSG00000225972.1           50          49
##           GTEX.XBED.0126.SM.47JY7 GTEX.XMK1.0626.SM.4B65A
## ENSG00000227232.4          766         568
## ENSG00000237683.5          374         738
## ENSG00000241860.2           71          67
## ENSG00000228463.4           28          52
## ENSG00000225972.1           74          81
##           GTEX.Y5V6.0526.SM.4VBRV GTEX.YEC4.0626.SM.5CVLU
## ENSG00000227232.4          482         365
## ENSG00000237683.5          681         359
## ENSG00000241860.2           63          61
## ENSG00000228463.4           51          15
## ENSG00000225972.1          110          82
##           GTEX.YFC4.2626.SM.5P9FQ GTEX.YJ89.0726.SM.5P9F7
## ENSG00000227232.4          1472        1325
## ENSG00000237683.5          2020         853
## ENSG00000241860.2           196          94
## ENSG00000228463.4           52          44
## ENSG00000225972.1           54          50
##           GTEX.Z9EW.0226.SM.5CVM7 GTEX.ZLV1.0126.SM.4WWBZ
## ENSG00000227232.4          450         689
## ENSG00000237683.5          352         809
## ENSG00000241860.2           43          82
## ENSG00000228463.4            9          61
## ENSG00000225972.1          116          37
##           GTEX.ZYVF.1126.SM.5E458 GTEX.ZYY3.1926.SM.5GZXS
## ENSG00000227232.4          838        1003
## ENSG00000237683.5          1212         960
## ENSG00000241860.2            89          59
## ENSG00000228463.4            21          26
## ENSG00000225972.1            51          66
## 18369 more rows ...
##
## $samples
##           group lib.size norm.factors
## GTEX.111VG.0526.SM.5N9BW      1 52085501      1
## GTEX.11EM3.0126.SM.5985K      1 64954617      1
## GTEX.11EMC.0226.SM.5EGLP      1 65673287      1
## GTEX.11NSD.0126.SM.5987F      1 52084492      1
## GTEX.11NV4.0626.SM.5N9BR      1 51837308      1
## 25 more rows ...

```

```

# See what slots are stored in y
names(y)

```

```
## [1] "counts" "samples"
```

```

# Library size information is stored in the samples slot
y$samples

```

```
##              group lib.size norm.factors
## GTEX.111VG.0526.SM.5N9BW      1 52085501      1
## GTEX.11EM3.0126.SM.5985K      1 64954617      1
## GTEX.11EMC.0226.SM.5EGLP      1 65673287      1
## GTEX.11NSD.0126.SM.5987F      1 52084492      1
## GTEX.11NV4.0626.SM.5N9BR      1 51837308      1
## GTEX.11O72.2326.SM.5BC7H      1 68388840      1
## GTEX.12WSG.0226.SM.5EGIF      1 38526920      1
## GTEX.139UW.0126.SM.5KM1B      1 48121274      1
## GTEX.13NZ9.1126.SM.5MR37      1 61301417      1
## GTEX.13O1R.0826.SM.5J2MB      1 60635354      1
## GTEX.13OVG.0226.SM.5LU93      1 47021443      1
## GTEX.13QJC.0826.SM.5RQKC      1 48725791      1
## GTEX.13U4I.0526.SM.5LU59      1 45294109      1
## GTEX.14ABY.0926.SM.5Q5DY      1 64593320      1
## GTEX.14AS3.0226.SM.5Q5B6      1 41908407      1
## GTEX.14BMU.0226.SM.5S2QA      1 44406229      1
## GTEX.PWN1.2626.SM.2I3FH      1 39644072      1
## GTEX.S7SE.0726.SM.2XCD7      1 41928929      1
## GTEX.T5JW.1226.SM.3GACY      1 46317315      1
## GTEX.WYVS.0326.SM.3NM9V      1 76564730      1
## GTEX.XBED.0126.SM.47JY7      1 47027068      1
## GTEX.XMK1.0626.SM.4B65A      1 52878007      1
## GTEX.Y5V6.0526.SM.4VBRV      1 68606086      1
## GTEX.YEC4.0626.SM.5CVLU      1 50583432      1
## GTEX.YFC4.2626.SM.5P9FQ      1 80995956      1
## GTEX.YJ89.0726.SM.5P9F7      1 73817346      1
## GTEX.Z9EW.0226.SM.5CVM7      1 46741531      1
## GTEX.ZLV1.0126.SM.4WWBZ      1 55313069      1
## GTEX.ZYVF.1126.SM.5E458      1 46918338      1
## GTEX.ZYY3.1926.SM.5GZXS      1 48818257      1
```

8. Quality control

Ahora que hemos eliminado los genes de baja expresión y hemos almacenado nuestros conteos en un objeto DGEList, vamos a llevar a cabo algunos gráficos que nos permitan realizar un pequeño informe de los mismos (Quality control).

Library sizes and distribution plots

Primero, podemos verificar cuántas lecturas tenemos para cada muestra en el objeto creado (counting data)

```
y$samples$lib.size
```

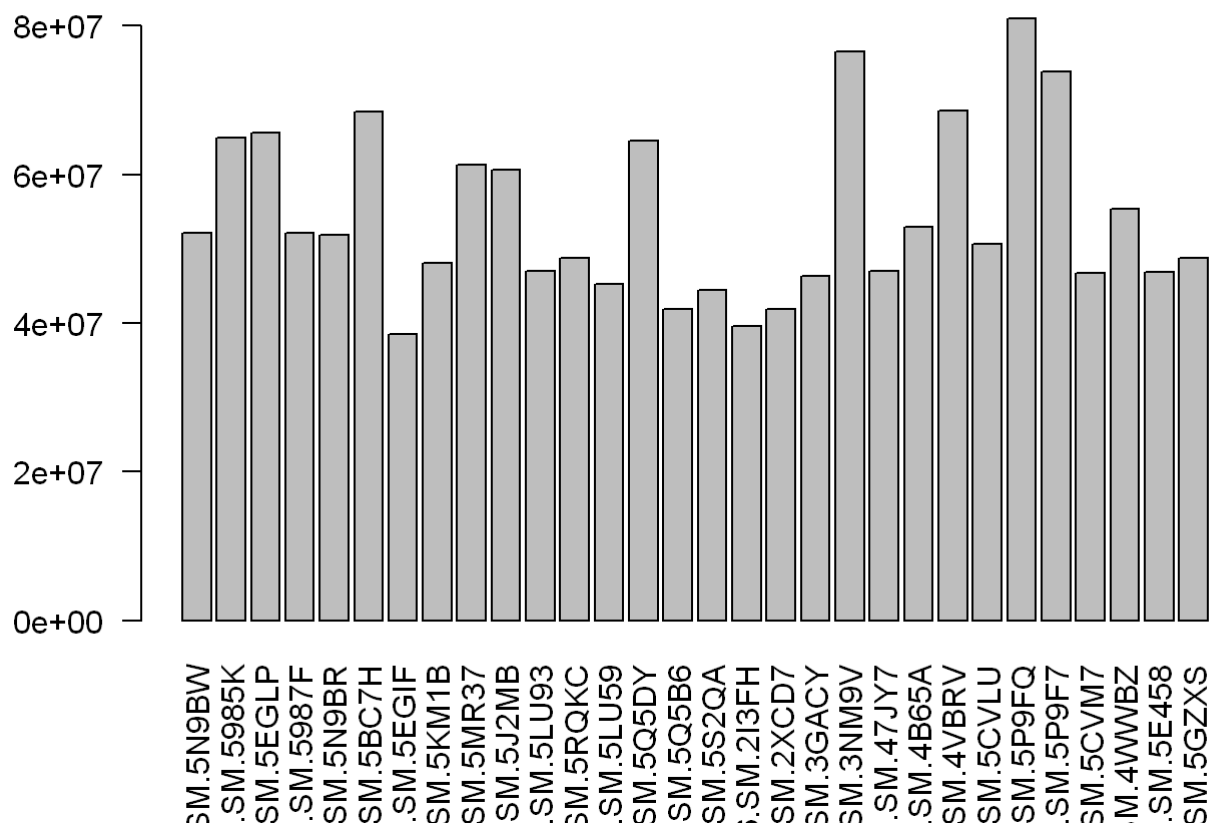
```
## [1] 52085501 64954617 65673287 52084492 51837308 68388840 38526920 48121274
## [9] 61301417 60635354 47021443 48725791 45294109 64593320 41908407 44406229
## [17] 39644072 41928929 46317315 76564730 47027068 52878007 68606086 50583432
## [25] 80995956 73817346 46741531 55313069 46918338 48818257
```

9. Diagrama de barras de los library sizes

También podemos plotear a partir de un diagrama de barras de los “library sizes” para ver si hay discrepancias importantes entre las muestras

```
# The names argument tells the barplot to use the sample names on the x-axis
# The las argument rotates the axis names
barplot(y$samples$lib.size, names=colnames(y), las=2)
# Add a title to the plot
title("Barplot of library sizes")
```

Barplot of library sizes



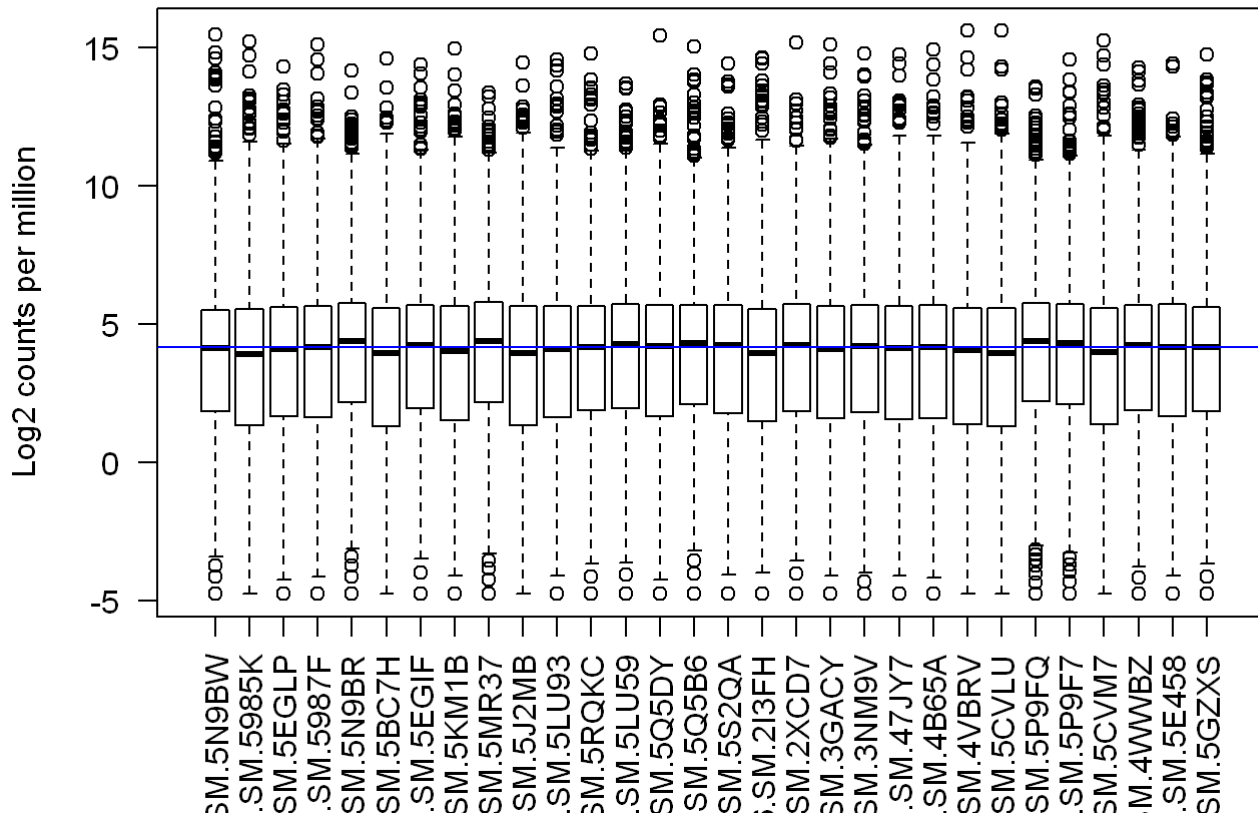
Los “ounting data” (datos de recuento) no se distribuyen segun una Distribución Normal, por lo que si queremos examinar las distribuciones de los recuentos sin procesar, utilizaremos Boxplots para verificar la distribución de los recuentos de lectura en escala log2.

Podemos usar la función cpm para obtener recuentos de log2 por millón, corregidos por los library sizes (tamaños de biblioteca). La función cpm también incorpora una pequeña “modificación” para evitar el problema asociado al logaritmo de valores de cero.

10, Diagrama de cajas

```
# Get log2 counts per million
logcounts <- cpm(y, log=TRUE)
# Check distributions of samples using boxplots
boxplot(logcounts, xlab="", ylab="Log2 counts per million", las=2)
# Let's add a blue horizontal line that corresponds to the median logCPM
abline(h=median(logcounts), col="blue")
title("Boxplots of logCPMs (unnormalised)")
```

Boxplots of logCPMs (unnormalised)



De los boxplots, vemos que, en general, las distribuciones del counting data no son idénticas, pero tampoco son muy diferentes.

Si una muestra está realmente muy por encima o por debajo de la línea horizontal azul, es posible que tengamos que investigar más esa muestra.

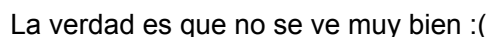
11. Multidimensional scaling plots

Uno de los gráficos más importante en el Quality control es el MDS. Un MDSplot es un gráfico, que nos permite “visualizar” variabilidad en los datos. Si su experimento está bien “controlado” y funcionó bien, lo que esperamos ver es que las principales fuentes de variación en los datos sean los tratamientos / grupos que nos interesan.

También nos puede ayudar en “la visualización de valores atípicos. Podemos usar la función plotMDS para crear el diagrama de MDS.

Diagrama de MDS

```
plotMDS(y)
```



Podemos complementar la visualización de los datos con la función `heatmap.2` que nos permitiría obtener la representación del cluster jerárquico de las muestras, en concreto, en este ejemplo, se grafica (a partir del método `average`) la matriz de distancias euclídeas del `logCPM` (objeto `logcounts`) para los 500 genes más variables. El diagrama del heatmap se representará en el último apartado.

```
## [1] "ENSG00000229807.5" "ENSG00000110680.8" "ENSG00000012817.11"
## [4] "ENSG00000114374.8" "ENSG00000131002.7" "ENSG00000129824.11"
```

```
# Subset logcounts matrix  
highly_variable_lcpm <- logcounts[select_var,]  
dim(highly_variable_lcpm)
```

```
## [1] 500 30
```

```
head(highly_variable_lcpm)
```

##	GTEX.111VG.0526.SM.5N9BW	GTEX.11EM3.0126.SM.5985K
## ENSG00000229807.5	-2.718176	9.3032207
## ENSG00000110680.8	-1.622558	-0.5821209
## ENSG00000012817.11	7.263680	-1.8206317
## ENSG00000114374.8	6.060782	-2.1745202
## ENSG00000131002.7	6.654994	-2.7904720
## ENSG00000129824.11	7.462084	-0.8761533
##	GTEX.11EMC.0226.SM.5EGLP	GTEX.11NSD.0126.SM.5987F
## ENSG00000229807.5	8.8746914	-2.128050
## ENSG00000110680.8	7.7763875	-2.718155
## ENSG00000012817.11	-2.8022915	7.275198
## ENSG00000114374.8	-2.1877475	6.507863
## ENSG00000131002.7	-2.6566599	6.091193
## ENSG00000129824.11	-0.8909509	7.677126
##	GTEX.11NV4.0626.SM.5N9BR	GTEX.11O72.2326.SM.5BC7H
## ENSG00000229807.5	-2.712949	-1.602891
## ENSG00000110680.8	-2.712949	3.397195
## ENSG00000012817.11	7.871551	6.678998
## ENSG00000114374.8	6.485204	6.132382
## ENSG00000131002.7	7.409647	5.157281
## ENSG00000129824.11	7.852049	7.846139
##	GTEX.12WSG.0226.SM.5EGIF	GTEX.139UW.0126.SM.5KM1B
## ENSG00000229807.5	9.4114330	-2.300322
## ENSG00000110680.8	-1.8869714	-1.805201
## ENSG00000012817.11	-3.9947683	6.975253
## ENSG00000114374.8	-3.4951997	6.810346
## ENSG00000131002.7	-3.1248326	6.436111
## ENSG00000129824.11	-0.6000913	7.553627
##	GTEX.13NZ9.1126.SM.5MR37	GTEX.13O1R.0826.SM.5J2MB
## ENSG00000229807.5	-3.848986	-1.384193
## ENSG00000110680.8	-2.445467	9.692514
## ENSG00000012817.11	7.977681	6.677879
## ENSG00000114374.8	6.397054	6.536891
## ENSG00000131002.7	6.683575	6.145760
## ENSG00000129824.11	7.784346	7.762029
##	GTEX.13OVG.0226.SM.5LU93	GTEX.13QJC.0826.SM.5RQKC
## ENSG00000229807.5	-2.003235	9.434566
## ENSG00000110680.8	-3.036955	-1.299790
## ENSG00000012817.11	7.366544	-3.072626
## ENSG00000114374.8	6.597902	-4.125467
## ENSG00000131002.7	6.512768	-3.683767
## ENSG00000129824.11	7.557474	-1.820885
##	GTEX.13U4I.0526.SM.5LU59	GTEX.14ABY.0926.SM.5Q5DY
## ENSG00000229807.5	9.6226138	-1.1867559
## ENSG00000110680.8	0.3231476	0.7197045
## ENSG00000012817.11	-2.9990437	6.5049404
## ENSG00000114374.8	-4.0868357	6.3479110
## ENSG00000131002.7	-3.2792143	5.8086103
## ENSG00000129824.11	-1.3583875	7.9679308
##	GTEX.14AS3.0226.SM.5Q5B6	GTEX.14BMU.0226.SM.5S2QA
## ENSG00000229807.5	9.6504170	9.805710
## ENSG00000110680.8	-4.7652806	-3.611581
## ENSG00000012817.11	-1.4311584	-2.978820
## ENSG00000114374.8	-1.9912251	-4.765281
## ENSG00000131002.7	-3.5650022	-3.611581
## ENSG00000129824.11	-0.6056577	-2.978820
##	GTEX.PWN1.2626.SM.2I3FH	GTEX.S7SE.0726.SM.2XCD7

##	ENSG000000229807.5	9.7351668	-3.2066175
##	ENSG000000110680.8	0.7668791	-0.7727064
##	ENSG000000012817.11	-1.2683373	6.9801434
##	ENSG000000114374.8	-2.8607245	6.3623148
##	ENSG000000131002.7	-3.5191958	6.1859211
##	ENSG000000129824.11	-1.7907803	8.1136462
##	GTEX.T5JW.1226.SM.3GACY	GTEX.WYVS.0326.SM.3NM9V	
##	ENSG000000229807.5	8.9684011	10.0114465
##	ENSG000000110680.8	6.2705314	2.4060716
##	ENSG000000012817.11	0.2134548	-2.5788043
##	ENSG000000114374.8	-1.0936478	-1.5684495
##	ENSG000000131002.7	-0.4164854	-2.3695848
##	ENSG000000129824.11	-0.2968670	-0.7726863
##	GTEX.XBED.0126.SM.47JY7	GTEX.XMK1.0626.SM.4B65A	
##	ENSG000000229807.5	-1.776240	-2.272477
##	ENSG000000110680.8	10.001647	9.879521
##	ENSG000000012817.11	7.177538	6.686635
##	ENSG000000114374.8	6.682904	6.318754
##	ENSG000000131002.7	6.279340	6.166787
##	ENSG000000129824.11	7.791057	7.714708
##	GTEX.Y5V6.0526.SM.4VBRV	GTEX.YEC4.0626.SM.5CVLU	
##	ENSG000000229807.5	-2.848890	-0.967942
##	ENSG000000110680.8	10.194599	1.721901
##	ENSG000000012817.11	6.711817	7.010032
##	ENSG000000114374.8	6.037534	6.076172
##	ENSG000000131002.7	5.757531	5.865838
##	ENSG000000129824.11	7.404991	7.569556
##	GTEX.YFC4.2626.SM.5P9FQ	GTEX.YJ89.0726.SM.5P9F7	
##	ENSG000000229807.5	9.374923	-1.413327
##	ENSG000000110680.8	-3.536899	-3.082495
##	ENSG000000012817.11	-2.757425	7.297902
##	ENSG000000114374.8	-4.765281	6.496223
##	ENSG000000131002.7	-4.765281	6.638566
##	ENSG000000129824.11	-2.641747	7.947008
##	GTEX.Z9EW.0226.SM.5CVM7	GTEX.ZLV1.0126.SM.4WWBZ	
##	ENSG000000229807.5	-0.1640095	9.459529
##	ENSG000000110680.8	9.9851052	-2.085349
##	ENSG000000012817.11	7.2177655	-4.765281
##	ENSG000000114374.8	6.2123057	-2.975218
##	ENSG000000131002.7	5.9585426	-3.777364
##	ENSG000000129824.11	7.5016355	-3.196451
##	GTEX.ZYVF.1126.SM.5E458	GTEX.ZYY3.1926.SM.5GZXS	
##	ENSG000000229807.5	9.848223161	8.5286664
##	ENSG000000110680.8	-2.270345829	3.7565253
##	ENSG000000012817.11	-0.983729159	-1.1634167
##	ENSG000000114374.8	-3.654745620	-2.4726482
##	ENSG000000131002.7	-3.311691623	-2.4726482
##	ENSG000000129824.11	-0.005949852	-0.7615973

13. Normalization for “composition bias”

El proceso de normalización denominado TMM se realiza para eliminar los sesgos de composición (bias composition) entre las bibliotecas.

Este método genera un conjunto de factores de normalización, donde el producto de estos factores y los tamaños de la biblioteca (library sizes) definen el tamaño efectivo de la biblioteca (effective library size).

La función `calcNormFactors` calcula los factores de normalización entre bibliotecas.

```
# Apply normalisation to DGEList object  
y <- calcNormFactors(y)  
head(y)
```

```
## An object of class "DGEList"
## $counts
##
##          GTEX.111VG.0526.SM.5N9BW GTEX.11EM3.0126.SM.5985K
## ENSG00000227232.4                474                669
## ENSG00000237683.5                427                791
## ENSG00000241860.2                 92                 68
## ENSG00000228463.4                 9                  9
## ENSG00000225972.1                 87                155
## ENSG00000225630.1            31646            12906
##
##          GTEX.11EMC.0226.SM.5EGLP GTEX.11NSD.0126.SM.5987F
## ENSG00000227232.4                786                408
## ENSG00000237683.5                553                800
## ENSG00000241860.2                107                 29
## ENSG00000228463.4                 57                 51
## ENSG00000225972.1                 82                135
## ENSG00000225630.1            9595            9332
##
##          GTEX.11NV4.0626.SM.5N9BR GTEX.11O72.2326.SM.5BC7H
## ENSG00000227232.4                1301                633
## ENSG00000237683.5                1132               1075
## ENSG00000241860.2                 46                 82
## ENSG00000228463.4                 92                 16
## ENSG00000225972.1                 23                 57
## ENSG00000225630.1            8718            19638
##
##          GTEX.12WSG.0226.SM.5EGIF GTEX.139UW.0126.SM.5KM1B
## ENSG00000227232.4                369                430
## ENSG00000237683.5                235                679
## ENSG00000241860.2                 21                 95
## ENSG00000228463.4                  6                 26
## ENSG00000225972.1                101                 54
## ENSG00000225630.1            8997            7455
##
##          GTEX.13NZ9.1126.SM.5MR37 GTEX.13O1R.0826.SM.5J2MB
## ENSG00000227232.4                1002                460
## ENSG00000237683.5                602                279
## ENSG00000241860.2                 96                 41
## ENSG00000228463.4                 39                 16
## ENSG00000225972.1            4678                 99
## ENSG00000225630.1            8450            13275
##
##          GTEX.13OVG.0226.SM.5LU93 GTEX.13QJC.0826.SM.5RQKC
## ENSG00000227232.4                719                825
## ENSG00000237683.5                1064               853
## ENSG00000241860.2                 91                 94
## ENSG00000228463.4                 41                 98
## ENSG00000225972.1                 64                 55
## ENSG00000225630.1            14271            18090
##
##          GTEX.13U4I.0526.SM.5LU59 GTEX.14ABY.0926.SM.5Q5DY
## ENSG00000227232.4                636                775
## ENSG00000237683.5                606                580
## ENSG00000241860.2                 29                 69
## ENSG00000228463.4                 91                  5
## ENSG00000225972.1                 18                 37
## ENSG00000225630.1            7910            10398
##
##          GTEX.14AS3.0226.SM.5Q5B6 GTEX.14BMU.0226.SM.5S2QA
## ENSG00000227232.4                834                423
## ENSG00000237683.5                445                325
## ENSG00000241860.2                 40                 41
## ENSG00000228463.4                 66                 42
## ENSG00000225972.1                 33                 25
```

##	ENSG00000225630.1	9934	7614
##	GTEX.PWN1.2626.SM.2I3FH	GTEX.S7SE.0726.SM.2XCD7	
##	ENSG00000227232.4	297	422
##	ENSG00000237683.5	229	247
##	ENSG00000241860.2	50	73
##	ENSG00000228463.4	53	56
##	ENSG00000225972.1	192	77
##	ENSG00000225630.1	11276	9076
##	GTEX.T5JW.1226.SM.3GACY	GTEX.WYVS.0326.SM.3NM9V	
##	ENSG00000227232.4	541	820
##	ENSG00000237683.5	1468	973
##	ENSG00000241860.2	85	101
##	ENSG00000228463.4	35	66
##	ENSG00000225972.1	50	49
##	ENSG00000225630.1	12188	14976
##	GTEX.XBED.0126.SM.47JY7	GTEX.XMK1.0626.SM.4B65A	
##	ENSG00000227232.4	766	568
##	ENSG00000237683.5	374	738
##	ENSG00000241860.2	71	67
##	ENSG00000228463.4	28	52
##	ENSG00000225972.1	74	81
##	ENSG00000225630.1	10542	34546
##	GTEX.Y5V6.0526.SM.4VBRV	GTEX.YEC4.0626.SM.5CVLU	
##	ENSG00000227232.4	482	365
##	ENSG00000237683.5	681	359
##	ENSG00000241860.2	63	61
##	ENSG00000228463.4	51	15
##	ENSG00000225972.1	110	82
##	ENSG00000225630.1	14469	36579
##	GTEX.YFC4.2626.SM.5P9FQ	GTEX.YJ89.0726.SM.5P9F7	
##	ENSG00000227232.4	1472	1325
##	ENSG00000237683.5	2020	853
##	ENSG00000241860.2	196	94
##	ENSG00000228463.4	52	44
##	ENSG00000225972.1	54	50
##	ENSG00000225630.1	12782	27107
##	GTEX.Z9EW.0226.SM.5CVM7	GTEX.ZLV1.0126.SM.4WWBZ	
##	ENSG00000227232.4	450	689
##	ENSG00000237683.5	352	809
##	ENSG00000241860.2	43	82
##	ENSG00000228463.4	9	61
##	ENSG00000225972.1	116	37
##	ENSG00000225630.1	51193	6712
##	GTEX.ZYVF.1126.SM.5E458	GTEX.ZYY3.1926.SM.5GZXS	
##	ENSG00000227232.4	838	1003
##	ENSG00000237683.5	1212	960
##	ENSG00000241860.2	89	59
##	ENSG00000228463.4	21	26
##	ENSG00000225972.1	51	66
##	ENSG00000225630.1	6015	48449
##			
##	\$samples		
##	group lib.size norm.factors		
##	GTEX.111VG.0526.SM.5N9BW	1 52085501	0.9417468
##	GTEX.11EM3.0126.SM.5985K	1 64954617	0.8841413
##	GTEX.11EMC.0226.SM.5EGLP	1 65673287	0.9728161
##	GTEX.11NSD.0126.SM.5987F	1 52084492	1.0252789

```
## GTEX.11NV4.0626.SM.5N9BR      1 51837308      1.1079924
## 25 more rows ...
```

Esta linea “actualizará” los factores de normalización en el objeto DGEList (sus valores predeterminados son 1).

```
y$samples
```

```
##              group lib.size norm.factors
## GTEX.111VG.0526.SM.5N9BW      1 52085501      0.9417468
## GTEX.11EM3.0126.SM.5985K      1 64954617      0.8841413
## GTEX.11EMC.0226.SM.5EGLP      1 65673287      0.9728161
## GTEX.11NSD.0126.SM.5987F      1 52084492      1.0252789
## GTEX.11NV4.0626.SM.5N9BR      1 51837308      1.1079924
## GTEX.11O72.2326.SM.5BC7H      1 68388840      0.9028173
## GTEX.12WSG.0226.SM.5EGIF      1 38526920      1.0468427
## GTEX.139UW.0126.SM.5KM1B      1 48121274      0.9521874
## GTEX.13NZ9.1126.SM.5MR37      1 61301417      1.1459429
## GTEX.13O1R.0826.SM.5J2MB      1 60635354      0.9216591
## GTEX.13OVG.0226.SM.5LU93      1 47021443      0.9875460
## GTEX.13QJC.0826.SM.5RQKC      1 48725791      0.9969542
## GTEX.13U4I.0526.SM.5LU59      1 45294109      1.0698486
## GTEX.14ABY.0926.SM.5Q5DY      1 64593320      1.0162815
## GTEX.14AS3.0226.SM.5Q5B6      1 41908407      1.0875705
## GTEX.14BMU.0226.SM.5S2QA      1 44406229      1.0165402
## GTEX.PWN1.2626.SM.2I3FH      1 39644072      0.9110875
## GTEX.S7SE.0726.SM.2XCD7      1 41928929      1.0824931
## GTEX.T5JW.1226.SM.3GACY      1 46317315      0.9887829
## GTEX.WYVS.0326.SM.3NM9V      1 76564730      1.0296372
## GTEX.XBED.0126.SM.47JY7      1 47027068      0.9998168
## GTEX.XMK1.0626.SM.4B65A      1 52878007      1.0106043
## GTEX.Y5V6.0526.SM.4VBRV      1 68606086      0.9260320
## GTEX.YEC4.0626.SM.5CVLU      1 50583432      0.9171709
## GTEX.YFC4.2626.SM.5P9FQ      1 80995956      1.1087606
## GTEX.YJ89.0726.SM.5P9F7      1 73817346      1.1032385
## GTEX.Z9EW.0226.SM.5CVM7      1 46741531      0.9145234
## GTEX.ZLV1.0126.SM.4WWBZ      1 55313069      0.9860888
## GTEX.ZYVF.1126.SM.5E458      1 46918338      1.0280662
## GTEX.ZYY3.1926.SM.5GZXS      1 48818257      0.9886884
```

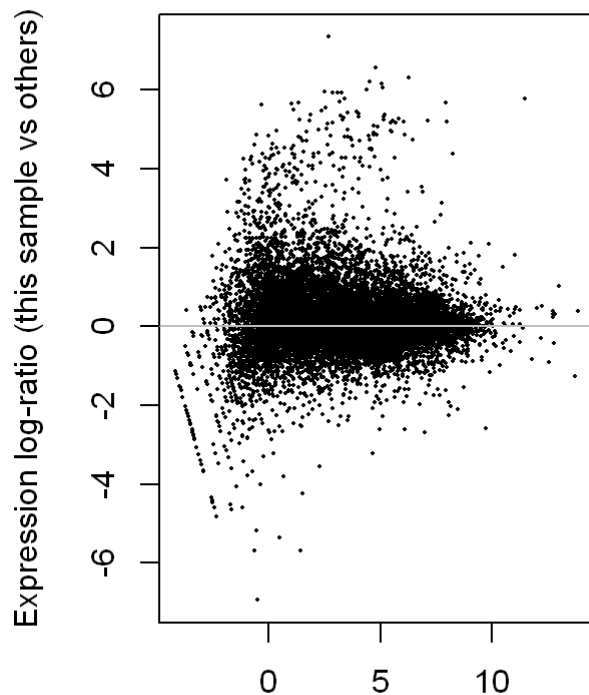
Un factor de normalización por debajo de uno indica que el tamaño de la biblioteca se reducirá, ya que “hay más sesgo de composición” (composition bias) en esa biblioteca en relación con las otras bibliotecas.

Es decir estamos re-escalando los recuentos “incrementandolos” en esa muestra. Por el contrario, un factor por encima de uno es equivalente a “reescalar a la baja” los recuentos.

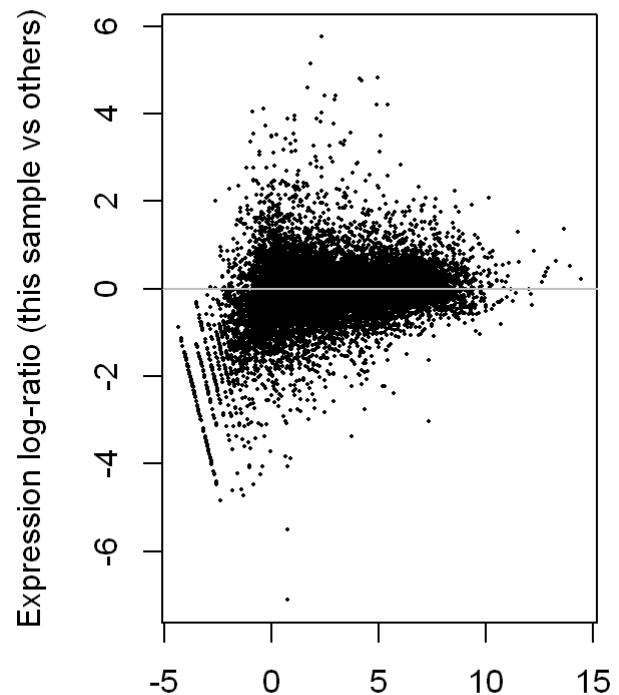
Si graficamos la diferencias medias usando la función plotMD para estas muestras, deberíamos poder ver el problema de sesgo de composición (bias composition).

Utilizaremos los logcounts, “normalizados por el tamaño de la biblioteca” (library size)“, pero no para el sesgo de composición (bias composition)

```
par(mfrow=c(1,2))
plotMD(logcounts,column = 7)
abline(h=0,col="grey")
plotMD(logcounts,column = 11)
abline(h=0,col="grey")
```

GTEX.12WSG.0226.SM.5EGIF

Average log-expression

GTEX.13OVG.0226.SM.5LU93

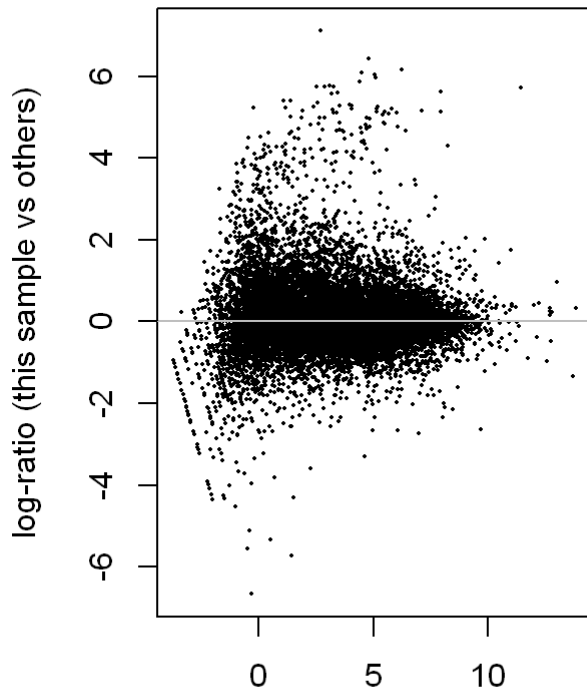
Average log-expression

Los gráficos de “diferencia de medias” muestran la expresión promedio (media: eje x) contra los cambios log-fold (diferencia: eje y).

Debido a que nuestro objeto DGEList contiene los factores de normalización, si rehacemos estos gráficos usando y(el objeto y), deberíamos ver que el problema de sesgo de composición (bias composition) ha sido resuelto.

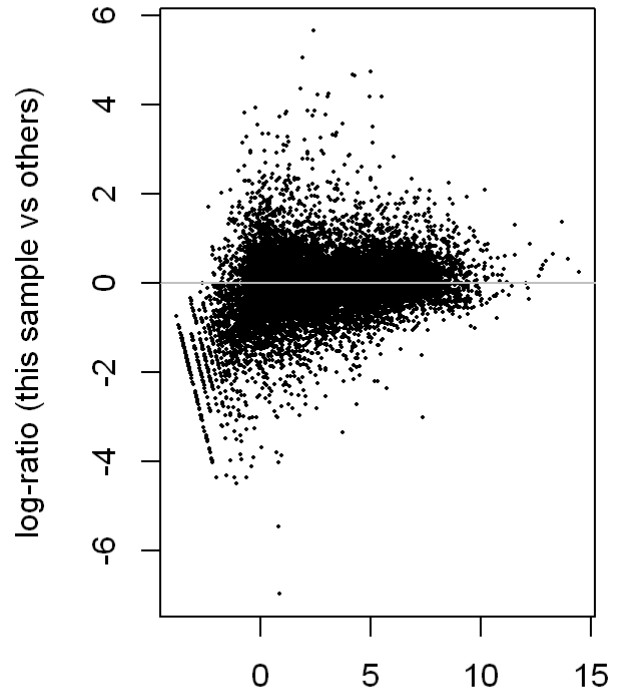
```
par(mfrow=c(1,2))
plotMD(y,column = 7)
abline(h=0,col="grey")
plotMD(y,column = 11)
abline(h=0,col="grey")
```

GTEX.12WSG.0226.SM.5EGIF



Average log CPM (this sample and others)

GTEX.13OVG.0226.SM.5LU93



Average log CPM (this sample and others)

14. Differential expression with limma-voom

Hay una serie de paquetes para analizar datos de RNA-Seq. El paquete limma (Ritchie et al., 2015) (desde la versión 3.16.0) ofrece la función voom, que transforma los recuentos de lectura en logCPM teniendo en cuenta la relación de la media y varianza de los datos (Law et al., 2014).

Después de aplicar voom, los usuarios pueden aplicar un modelo lineal a los datos transformados por voom para identificar genes expresados diferencialmente, utilizando comandos estándar de limma.

Leemos los targets seleccionados y los guardamos en una variable targets. Después determinamos los factores y niveles que tenemos. Vemos que nos salen los 3 niveles: SFI, NIT y ELI

```
targets<- read.csv2(file.path("../data", "targets.csv"), head=T, sep=";")
head(targets,5)
```

```
##      Experiment SRA_Sample      Sample_Name Grupo_analisis body_site
## 1  SRX223301   SRS389914  GTEX-T5JW-1226-SM-3GACY           2  Thyroid
## 2  SRX605452   SRS639261  GTEX-12WSG-0226-SM-5EGIF           2  Thyroid
## 3  SRX559198   SRS624031  GTEX-11072-2326-SM-5BC7H           2  Thyroid
## 4  SRX614680   SRS644012  GTEX-1301R-0826-SM-5J2MB           2  Thyroid
## 5  SRX597594   SRS637292  GTEX-ZLV1-0126-SM-4WWBZ           2  Thyroid
##      molecular_data_type      sex Group ShortName
## 1 Allele-Specific Expression female   SFI T5JW-_SFI
## 2              RNA Seq (NGS) female   SFI 12WSG_SFI
## 3              RNA Seq (NGS)   male   SFI 11072_SFI
## 4              RNA Seq (NGS)   male   SFI 1301R_SFI
## 5 Allele-Specific Expression female   SFI ZLV1-_SFI
```

```
group<-factor(stargets$Group)
group
```

```
## [1] SFI SFI SFI SFI SFI SFI SFI SFI SFI SFI SFI NIT NIT NIT NIT NIT NIT NIT NIT NIT
## [20] NIT ELI ELI ELI ELI ELI ELI ELI ELI ELI ELI ELI
## Levels: ELI NIT SFI
```

Create the design matrix

Primero, necesitamos crear una matriz de diseño para los grupos (lo teneis como material de consulta la guía del usuario de limma para obtener más información sobre las matrices de diseño y ya fue trabajado en la primera parte del curso).

Hay muchas formas diferentes de configurar la matriz de diseño, y estan supeditadas a las comparaciones que se “quieren testar”. En este análisis, supongamos que queremos testar las diferencias de estado (status) en los diferentes tipos por separado.

Por ejemplo, queremos saber qué genes se expresan diferencialmente

Anteriormente “hemos codificado como variable grupo”, que lleva implicito “cell type and status”.

Codificar de esta manera nos permite ser flexibles al especificar las comparaciones que nos interesan

```
# Look at group variable again
group
```

```
## [1] SFI SFI SFI SFI SFI SFI SFI SFI SFI SFI SFI NIT NIT NIT NIT NIT NIT NIT NIT NIT
## [20] NIT ELI ELI ELI ELI ELI ELI ELI ELI ELI ELI ELI
## Levels: ELI NIT SFI
```

```
# Specify a design matrix without an intercept term
design <- model.matrix(~ 0 + group)
design
```

```
##      groupELI groupNIT groupSFI
## 1          0          0          1
## 2          0          0          1
## 3          0          0          1
## 4          0          0          1
## 5          0          0          1
## 6          0          0          1
## 7          0          0          1
## 8          0          0          1
## 9          0          0          1
## 10         0          0          1
## 11         0          1          0
## 12         0          1          0
## 13         0          1          0
## 14         0          1          0
## 15         0          1          0
## 16         0          1          0
## 17         0          1          0
## 18         0          1          0
## 19         0          1          0
## 20         0          1          0
## 21         1          0          0
## 22         1          0          0
## 23         1          0          0
## 24         1          0          0
## 25         1          0          0
## 26         1          0          0
## 27         1          0          0
## 28         1          0          0
## 29         1          0          0
## 30         1          0          0
## attr(,"assign")
## [1] 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

```
## Make the column names of the design matrix a bit nicer
colnames(design) <- levels(group)
design
```



```
##      ELI NIT SFI
## 1      0   0   1
## 2      0   0   1
## 3      0   0   1
## 4      0   0   1
## 5      0   0   1
## 6      0   0   1
## 7      0   0   1
## 8      0   0   1
## 9      0   0   1
## 10     0   0   1
## 11     0   1   0
## 12     0   1   0
## 13     0   1   0
## 14     0   1   0
## 15     0   1   0
## 16     0   1   0
## 17     0   1   0
## 18     0   1   0
## 19     0   1   0
## 20     0   1   0
## 21     1   0   0
## 22     1   0   0
## 23     1   0   0
## 24     1   0   0
## 25     1   0   0
## 26     1   0   0
## 27     1   0   0
## 28     1   0   0
## 29     1   0   0
## 30     1   0   0
## attr(,"assign")
## [1] 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

Cada columna de la matriz de diseño nos remite a las muestras que corresponden a cada grupo

voom estima la tendencia de la varianza respecto a la media en el counting data, para luego asignar un peso a cada observación en función de la predicción de la varianza (según el modelo que nos da la tendencia). Los pesos se usan luego en el proceso de modelado lineal para ajustar la heterocedasticidad.

Así pues voom ajustará automáticamente los tamaños de biblioteca (library size) utilizando norm.factors ya calculados.

La transformación de voom usa la matriz de diseño de experimento y produce un objeto EList.

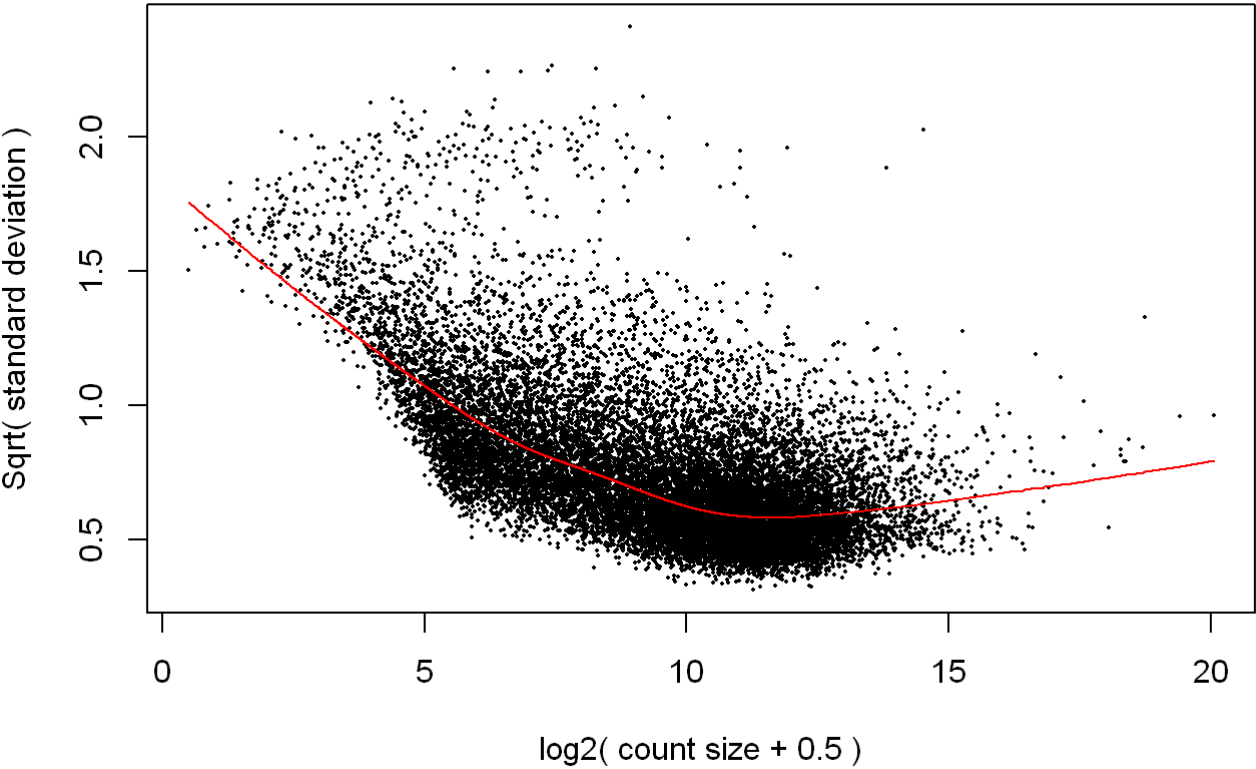
Podemos agregar plot = TRUE para generar un gráfico de la tendencia de media-varianza.

Este diagrama es importante ya que nos “informa” de si hay algún gen con “alta variabilidad” en nuestros datos, y sobretodo porque nos indica si hemos filtrado los recuentos bajos adecuadamente.

Los recuentos log2 normalizados que nos aporta voom se pueden encontrar en v\$E.

```
par(mfrow=c(1,1))
v <- voom(y,design,plot = TRUE)
```

voom: Mean-variance trend



▼

```
## An object of class "EList"
## $targets
##               group lib.size norm.factors
## GTEX.111VG.0526.SM.5N9BW      1 49051354      0.9417468
## GTEX.11EM3.0126.SM.5985K      1 57429059      0.8841413
## GTEX.11EMC.0226.SM.5EGLP      1 63888030      0.9728161
## GTEX.11NSD.0126.SM.5987F      1 53401129      1.0252789
## GTEX.11NV4.0626.SM.5N9BR      1 57435341      1.1079924
## 25 more rows ...
##
## $E
##               GTEX.111VG.0526.SM.5N9BW GTEX.11EM3.0126.SM.5985K
## ENSG00000227232.4                3.2740432                3.543231
## ENSG00000237683.5                3.1235595                3.784737
## ENSG00000241860.2                0.9151604                0.254323
## ENSG00000228463.4               -2.3682936               -2.595782
## ENSG00000225972.1                0.8349900                1.437062
##               GTEX.11EMC.0226.SM.5EGLP GTEX.11NSD.0126.SM.5987F
## ENSG00000227232.4                3.6218292                2.93539391
## ENSG00000237683.5                3.1149657                3.90595924
## ENSG00000241860.2                0.7507191               -0.85615531
## ENSG00000228463.4               -0.1519837               -0.05229783
## ENSG00000225972.1                0.3688484                1.34335069
##               GTEX.11NV4.0626.SM.5N9BR GTEX.11O72.2326.SM.5BC7H
## ENSG00000227232.4                4.5020927                3.3590057
## ENSG00000237683.5                4.3014285                4.1225967
## ENSG00000241860.2               -0.3047080                0.4181272
## ENSG00000228463.4                0.6875146               -1.9038009
## ENSG00000225972.1               -1.2892780               -0.1027050
##               GTEX.12WSG.0226.SM.5EGIF GTEX.139UW.0126.SM.5KM1B
## ENSG00000227232.4                3.1955909                3.2319490
## ENSG00000237683.5                2.5457436                3.8904094
## ENSG00000241860.2               -0.9075749                1.0595084
## ENSG00000228463.4               -2.6333999               -0.7899999
## ENSG00000225972.1                1.3314963                0.2502639
##               GTEX.13NZ9.1126.SM.5MR37 GTEX.13O1R.0826.SM.5J2MB
## ENSG00000227232.4                3.8350028                3.0426648
## ENSG00000237683.5                3.1004337                2.3223120
## ENSG00000241860.2                0.4580733               -0.4293531
## ENSG00000228463.4               -0.8306030               -1.7599984
## ENSG00000225972.1                6.0574466                0.8322321
##               GTEX.13OVG.0226.SM.5LU93 GTEX.13QJC.0826.SM.5RQKC
## ENSG00000227232.4                3.9536841                4.0869116
## ENSG00000237683.5                4.5187934                4.1350345
## ENSG00000241860.2                0.9785330                0.9600296
## ENSG00000228463.4               -0.1621274                1.0198390
## ENSG00000225972.1                0.4740604                0.1922030
##               GTEX.13U4I.0526.SM.5LU59 GTEX.14ABY.0926.SM.5Q5DY
## ENSG00000227232.4                3.7153585                3.56236978
## ENSG00000237683.5                3.6457056                3.14453906
## ENSG00000241860.2               -0.7160152                0.08232788
## ENSG00000228463.4                0.9170416               -3.57718158
## ENSG00000225972.1               -1.3892049               -0.80779450
##               GTEX.14AS3.0226.SM.5Q5B6 GTEX.14BMU.0226.SM.5S2QA
## ENSG00000227232.4                4.1944915                3.22986068
## ENSG00000237683.5                3.2890049                2.85015626
## ENSG00000241860.2               -0.1704267               -0.12131805
```

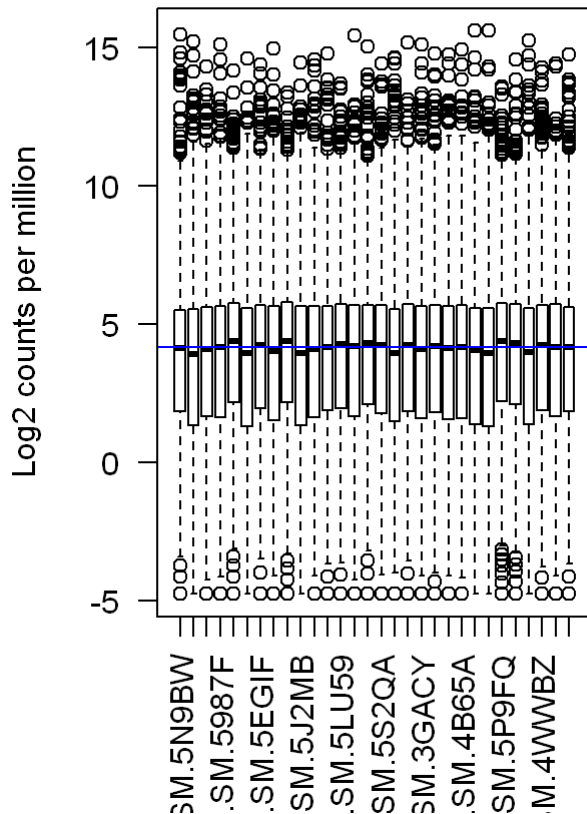
```
## ENSG00000228463.4          0.5450057          -0.08696654
## ENSG00000225972.1          -0.4441876          -0.82393214
##                               GTEX.PWN1.2626.SM.2I3FH GTEX.S7SE.0726.SM.2XCD7
## ENSG00000227232.4          3.0420511          3.2185756
## ENSG00000237683.5          2.6676556          2.4470528
## ENSG00000241860.2          0.4835167          0.6954404
## ENSG00000228463.4          0.5667723          0.3159470
## ENSG00000225972.1          2.4140199          0.7718925
##                               GTEX.T5JW.1226.SM.3GACY GTEX.WYVS.0326.SM.3NM9V
## ENSG00000227232.4          3.5636121          3.3796154
## ENSG00000237683.5          5.0029221          3.6262930
## ENSG00000241860.2          0.9006470          0.3645917
## ENSG00000228463.4          -0.3674584          -0.2454617
## ENSG00000225972.1          0.1410060          -0.6713875
##                               GTEX.XBED.0126.SM.47JY7 GTEX.XMK1.0626.SM.4B65A
## ENSG00000227232.4          4.0269869          3.41120241
## ENSG00000237683.5          2.9936668          3.78863998
## ENSG00000241860.2          0.6047162          0.33700147
## ENSG00000228463.4          -0.7222651          -0.02556861
## ENSG00000225972.1          0.6640134          0.60891402
##                               GTEX.Y5V6.0526.SM.4VBRV GTEX.YEC4.0626.SM.5CVLU
## ENSG00000227232.4          2.9249865268          2.9778721
## ENSG00000237683.5          3.4231712414          2.9539925
## ENSG00000241860.2          -0.0007139186          0.4066590
## ENSG00000228463.4          -0.3028980782          -1.5816591
## ENSG00000225972.1          0.7985039540          0.8304668
##                               GTEX.YFC4.2626.SM.5P9FQ GTEX.YJ89.0726.SM.5P9F7
## ENSG00000227232.4          4.0353260          4.0246882
## ENSG00000237683.5          4.4917708          3.3896146
## ENSG00000241860.2          1.1296596          0.2146097
## ENSG00000228463.4          -0.7744804          -0.8718993
## ENSG00000225972.1          -0.7205416          -0.6894212
##                               GTEX.Z9EW.0226.SM.5CVM7 GTEX.ZLV1.0126.SM.4WWBZ
## ENSG00000227232.4          3.39765822          3.6600687
## ENSG00000237683.5          3.04375438          3.8915493
## ENSG00000241860.2          0.02521842          0.5969842
## ENSG00000228463.4          -2.16979756          0.1731765
## ENSG00000225972.1          1.44646107          -0.5405193
##                               GTEX.ZYVF.1126.SM.5E458 GTEX.ZYY3.1926.SM.5GZXS
## ENSG00000227232.4          4.11965373          4.3778882
## ENSG00000237683.5          4.65175578          4.3147051
## ENSG00000241860.2          0.89180253          0.3018811
## ENSG00000228463.4          -1.16574849          -0.8650162
## ENSG00000225972.1          0.09448728          0.4623458
## 18369 more rows ...
##
## $weights
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## [1,] 4.5030196 4.9672876 5.305353 4.7482229 4.967624 5.1961140 3.9941982
## [2,] 4.4399822 4.8979258 5.233247 4.6808738 4.898256 5.1227426 3.9400419
## [3,] 1.1203512 1.2607530 1.363865 1.1944368 1.260855 1.3303272 0.9653701
## [4,] 0.5522524 0.6176995 0.667198 0.5864319 0.617748 0.6508351 0.4819840
## [5,] 1.8087636 1.9948351 2.124401 1.9084084 1.994965 2.0829684 1.5866789
##           [,8]      [,9]     [,10]     [,11]     [,12]     [,13]     [,14]
## [1,] 4.3171825 5.6181665 4.8844544 4.6240495 4.7564756 4.7491173 5.717847
## [2,] 4.2570748 5.5410157 4.8164980 4.5022091 4.6302729 4.6231577 5.573086
## [3,] 1.0638239 1.4604738 1.2357549 1.1996647 1.2406274 1.2383870 1.542690
## [4,] 0.5264915 0.7152142 0.6057789 0.9451511 0.9783536 0.9765018 1.228841
```

```
## [5,] 1.7297023 2.2389950 1.9628821 0.9931146 1.0274919 1.0255752 1.288662
##      [,15]      [,16]      [,17]      [,18]      [,19]      [,20]      [,21]
## [1,] 4.5706526 4.5433151 3.9634813 4.5587553 4.5843372 6.325885 4.8684665
## [2,] 4.4505626 4.4241039 3.8631540 4.4390547 4.4637991 6.186931 4.8656209
## [3,] 1.1828085 1.1741993 0.9918591 1.1790600 1.1871233 1.747588 1.2779043
## [4,] 0.9318373 0.9250376 0.7816860 0.9288767 0.9352453 1.405188 0.6561726
## [5,] 0.9789747 0.9717428 0.8205094 0.9758258 0.9825997 1.472476 1.1794696
##      [,22]      [,23]      [,24]      [,25]      [,26]      [,27]      [,28]
## [1,] 5.2699987 5.8394604 4.8286008 6.981922 6.6647792 4.587306 5.3350296
## [2,] 5.2670029 5.8363192 4.8257843 6.978751 6.6617867 4.584592 5.3319875
## [3,] 1.4033024 1.5866558 1.2653449 1.986180 1.8696105 1.190606 1.4241754
## [4,] 0.7206152 0.8198366 0.6498234 1.066296 0.9900518 0.612590 0.7316222
## [5,] 1.2973029 1.4713290 1.1675956 1.858008 1.7448538 1.098021 1.3171905
##      [,29]      [,30]
## [1,] 4.9457510 4.9477069
## [2,] 4.9428488 4.9448032
## [3,] 1.3023259 1.3029452
## [4,] 0.6684714 0.6687823
## [5,] 1.2025686 1.2031545
## 18369 more rows ...
##
## $design
##      ELI NIT SFI
## 1      0      0      1
## 2      0      0      1
## 3      0      0      1
## 4      0      0      1
## 5      0      0      1
## 25 more rows ...
```

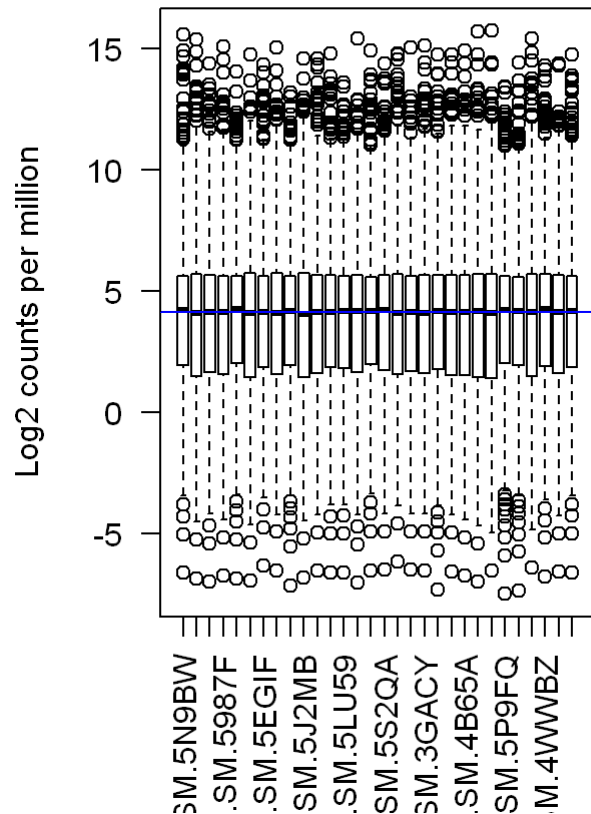
Ahora podemos comparar los boxplot despues antes y despues de normalizar. Los valores de expresión en v\$E ya son valores en escala logarítmica log2.

```
par(mfrow=c(1,2))
boxplot(logcounts, xlab="", ylab="Log2 counts per million",las=2,main="Non normalised
logCPM")
## Let's add a blue horizontal line that corresponds to the median logCPM
abline(h=median(logcounts),col="blue")
boxplot(v$E, xlab="", ylab="Log2 counts per million",las=2,main="Voom transformed log
CPM")
## Let's add a blue horizontal line that corresponds to the median logCPM
abline(h=median(v$E),col="blue")
```

Non normalised logCPM



Voom transformed logCPM



15. Testing for differential expression

Ahora que tenemos los datos obtenidos a partir de la función voom, podemos usar limma para obtener la expresión diferencial. Primero ajustamos un modelo lineal para cada gen usando la función lmFit en limma. lmFit necesita el objeto voom y la matriz de diseño que ya hemos especificado, que se encuentra dentro del objeto generado por voom

```
# Fit the linear model
fit <- lmFit(v)
names(fit)
```

```
## [1] "coefficients"      "stdev.unscaled"    "sigma"              "df.residual"
## [5] "cov.coefficients"  "pivot"              "rank"               "Amean"
## [9] "method"            "design"
```

Hay una serie de elementos dentro del objeto fit la mayoría de los cuales, son prácticamente idénticos a los vistos cuando aplicamos dicha función en la primera parte del microarray data analysis.

Dado que estamos interesados en obtener genes diferencialmente expresados entre los grupos, debemos especificar qué comparaciones queremos probar.

Las comparaciones se pueden especificar utilizando la función makeContrasts.

Aquí, estamos interesados en saber qué genes se expresan diferencialmente entre los distintos grupos

Los nombres de los grupos deben coincidir exactamente con los nombres de columna de la matriz de diseño.

```
#cont.matrix <- makeContrasts(SFI-NIT,SFI-ELI,NIT-ELI,levels=design)
cont.matrix <- makeContrasts(SFIvsNIT=SFI-NIT,SFIvsELI=SFI-ELI,NITvsELI=NIT-ELI,levels=design)
cont.matrix
```

```
##           Contrasts
## Levels SFIvsNIT SFIvsELI NITvsELI
##      ELI         0        -1        -1
##      NIT        -1         0         1
##      SFI         1         1         0
```

Las siguientes líneas se corresponden con las ya “presentadas y llevadas a cabo” y que se encuentran dentro del material de la primera parte de la asignatura.

```
fit.cont <- contrasts.fit(fit, cont.matrix)
fit.cont <- eBayes(fit.cont)
```

```
summa.fit <- decideTests(fit.cont)
summary(summa.fit)
```

```
##           SFIvsNIT SFIvsELI NITvsELI
## Down           0         0         0
## NotSig      18374      18374      18374
## Up            0         0         0
```

```
fit.cont
```

```
## An object of class "MArrayLM"
## $coefficients
##           Contrasts
##           SFIvsNIT  SFIvsELI  NITvsELI
## ENSG00000227232.4 -0.1231904 -0.2343738 -0.1111834
## ENSG00000237683.5 -0.1107212 -0.2871451 -0.1764239
## ENSG00000241860.2 -0.1806831 -0.3124049 -0.1317218
## ENSG00000228463.4 -0.9789662 -0.4027514  0.5762148
## ENSG00000225972.1  1.2761546  0.9660592 -0.3100954
## 18369 more rows ...
##
## $stdev.unscaled
##           Contrasts
##           SFIvsNIT  SFIvsELI  NITvsELI
## ENSG00000227232.4 0.2031788 0.1975970 0.1977131
## ENSG00000237683.5 0.2052181 0.1983561 0.1991050
## ENSG00000241860.2 0.4005397 0.3873543 0.3837840
## ENSG00000228463.4 0.5154630 0.5458204 0.4811301
## ENSG00000225972.1 0.3825753 0.3540027 0.4108064
## 18369 more rows ...
##
## $sigma
## [1] 1.008990 1.557569 0.629783 1.023444 1.817747
## 18369 more elements ...
##
## $df.residual
## [1] 27 27 27 27 27
## 18369 more elements ...
##
## $cov.coefficients
##           Contrasts
## Contrasts  SFIvsNIT SFIvsELI NITvsELI
## SFIvsNIT      0.2      0.1      -0.1
## SFIvsELI      0.1      0.2       0.1
## NITvsELI     -0.1      0.1       0.2
##
## $rank
## [1] 3
##
## $Amean
## ENSG00000227232.4 ENSG00000237683.5 ENSG00000241860.2 ENSG00000228463.4
##           3.5814555           3.5493974           0.3107755           -0.7259797
## ENSG00000225972.1
##           0.4625031
## 18369 more elements ...
##
## $method
## [1] "ls"
##
## $design
##   ELI NIT SFI
## 1   0   0   1
## 2   0   0   1
## 3   0   0   1
## 4   0   0   1
## 5   0   0   1
## 25 more rows ...
```



```
##
## $contrasts
##      Contrasts
## Levels SFivsNIT SFivsELI NITvsELI
##      ELI      0      -1      -1
##      NIT     -1       0       1
##      SFI      1       1       0
##
## $df.prior
## [1] 3.146253
##
## $s2.prior
## [1] 0.7620268
##
## $var.prior
## [1] 0.01312290 0.01312290 0.04351578
##
## $proportion
## [1] 0.01
##
## $s2.post
## [1] 0.9913403 2.2523567 0.4347621 1.0176503 3.0388867
## 18369 more elements ...
##
## $t
##      Contrasts
##      SFivsNIT  SFivsELI  NITvsELI
## ENSG00000227232.4 -0.6089577 -1.1912893 -0.5647978
## ENSG00000237683.5 -0.3594982 -0.9645777 -0.5904139
## ENSG00000241860.2 -0.6841416 -1.2231610 -0.5205292
## ENSG00000228463.4 -1.8826560 -0.7314556  1.1871966
## ENSG00000225972.1  1.9135029  1.5654524 -0.4330129
## 18369 more rows ...
##
## $df.total
## [1] 30.14625 30.14625 30.14625 30.14625 30.14625
## 18369 more elements ...
##
## $p.value
##      Contrasts
##      SFivsNIT  SFivsELI  NITvsELI
## ENSG00000227232.4 0.54711446 0.2428352 0.5763903
## ENSG00000237683.5 0.72172852 0.3424308 0.5593136
## ENSG00000241860.2 0.49911159 0.2307447 0.6064971
## ENSG00000228463.4 0.06942983 0.4701489 0.2444210
## ENSG00000225972.1 0.06522173 0.1279150 0.6680878
## 18369 more rows ...
##
## $lods
##      Contrasts
##      SFivsNIT  SFivsELI  NITvsELI
## ENSG00000227232.4 -4.687422 -4.562866 -4.883086
## ENSG00000237683.5 -4.714943 -4.621983 -4.872106
## ENSG00000241860.2 -4.616419 -4.577709 -4.692876
## ENSG00000228463.4 -4.541923 -4.605211 -4.570792
## ENSG00000225972.1 -4.498652 -4.533537 -4.690077
## 18369 more rows ...
##
```

```
## $F
## [1] 0.7029646 0.4693886 0.7450031 1.8294749 2.0990762
## 18369 more elements ...
##
## $F.p.value
## [1] 0.5030504 0.6298785 0.4832714 0.1778725 0.1401262
## 18369 more elements ...
```

```
toptable_SFIVsELI<-topTable(fit.cont,coef="SFIVsELI",sort.by="p")
toptable_SFIVsNIT<-topTable(fit.cont,coef="SFIVsNIT",sort.by="p")
toptable_NITvsELI<-topTable(fit.cont,coef="NITvsELI",sort.by="p")
```

```
# View(toptable_SFIVsELI)
# View(toptable_SFIVsNIT)
# View(toptable_NITvsELI)
```

16. Annotation and saving the results

```
library(org.Hs.eg.db)
columns(org.Hs.eg.db)
```

```
## [1] "ACCNUM"      "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
## [6] "ENTREZID"    "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
## [11] "GO"          "GOALL"        "IPI"          "MAP"          "OMIM"
## [16] "ONTOLOGY"    "ONTOLOGYALL"  "PATH"         "PFAM"         "PMID"
## [21] "PROSITE"     "REFSEQ"       "SYMBOL"       "UCSCKG"       "UNIGENE"
## [26] "UNIPROT"
```

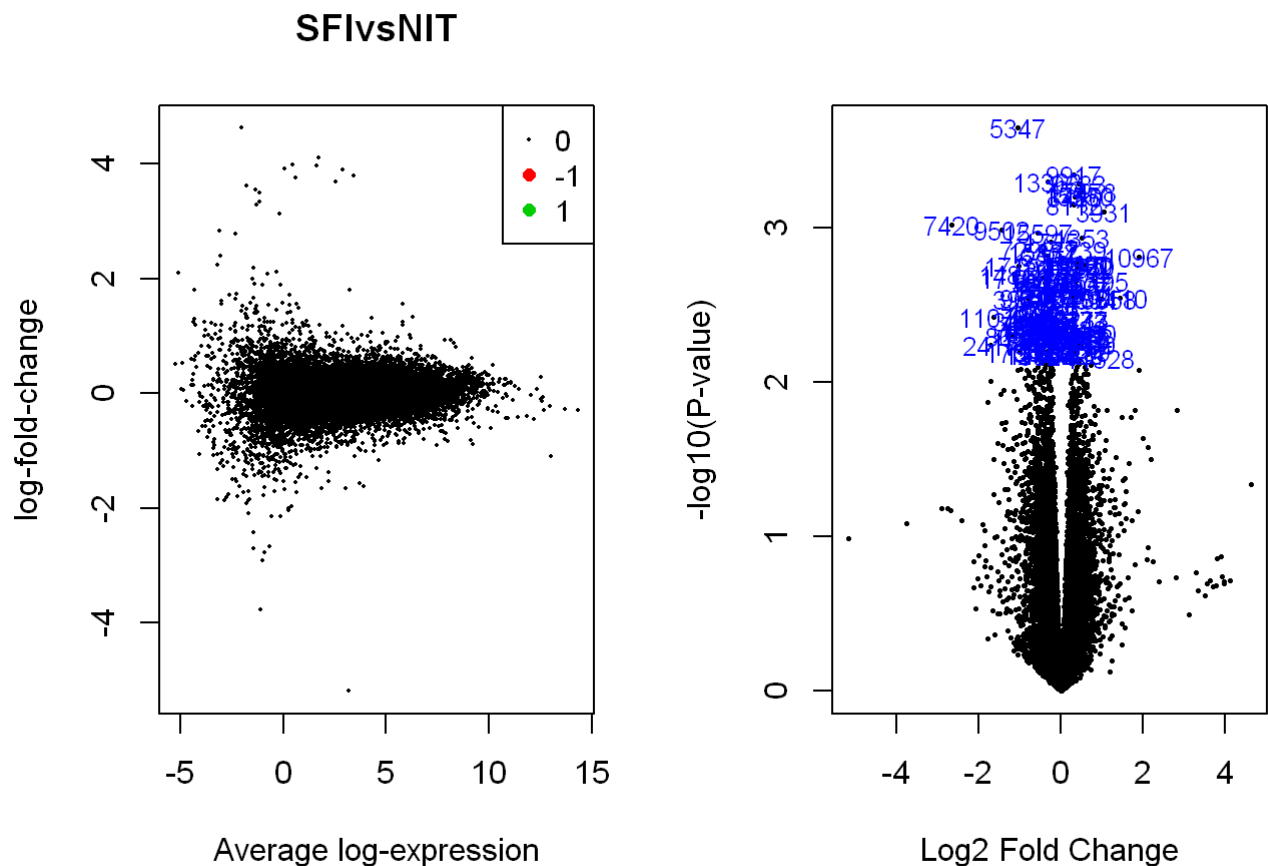
No he aconseguit crear les anotacions !

```
# ann <- select(org.Hs.eg.db,keys=rownames(fit.cont),columns=c("ENTREZID","SYMBOL","GENENAME"))
```

17. Volcano Plot

```
# We want to highlight the significant genes. We can get this from decideTests.
par(mfrow=c(1,2))
plotMD(fit.cont,coef=1,status=summa.fit[, "SFIVsNIT"], values = c(-1, 1))

# For the volcano plot we have to specify how many of the top genes to highlight.
# We can also specify that we want to plot the gene symbol for the highlighted genes.
# let's highlight the top 100 most DE genes
volcanoplot(fit.cont,coef=1,highlight=100,names=fit.cont$genes$SYMBOL)
```



Hay una función llamada `treat` en el paquete `limma` (McCarthy y Smyth 2009) que a partir del objeto `fit.conty` de de un “log fold change (`logFC`)” determinado por el usuario como “`threshold`” permite “recalcular the moderate t-statistics and p-values”. Este procedimiento es mucho más “preciso” “en el control de falsos positivos” que “listar” los p-valores y descartar a continuación genes con `logFC` pequeños.

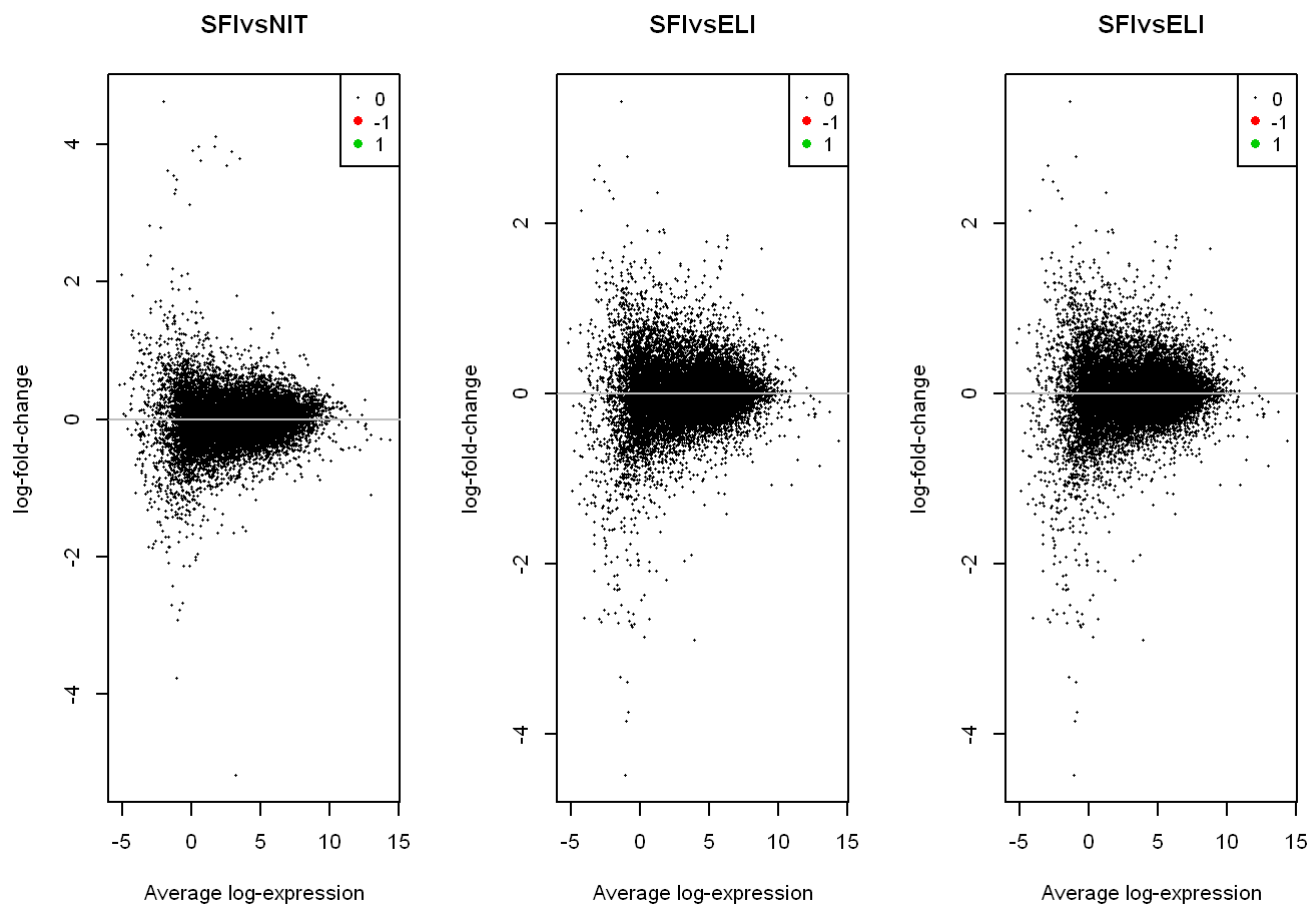
```
# This is easy to do after our analysis, we just give the treat function the fit.cont
object and specify our cut-off.
fit.treat <- treat(fit.cont,lfc=1)
res.treat <- decideTests(fit.treat)
summary(res.treat)
```

##	SFIvsNIT	SFIvsELI	NITvsELI
## Down	0	0	0
## NotSig	18374	18374	18374
## Up	0	0	0

```
topTable(fit.treat,coef=1,sort.by="p")
```

##		logFC	AveExpr	t	P.Value	adj.P.Val
##	ENSG00000170356.8	-2.667071	-0.6691073	-2.285105	0.01476582	1
##	ENSG00000225851.1	2.828226	-3.0957945	1.661617	0.05425436	1
##	ENSG00000170523.3	1.898064	-0.1426545	1.645973	0.05508328	1
##	ENSG00000171195.6	4.624876	-2.0323915	1.630333	0.06515299	1
##	ENSG00000160951.3	1.906968	-1.3702753	1.341059	0.09504230	1
##	ENSG00000025423.7	-1.651809	3.0560536	-1.237182	0.11279568	1
##	ENSG00000175535.6	-3.758024	-1.0910852	-1.319796	0.11344780	1
##	ENSG00000253288.1	2.123276	-0.3166287	1.235337	0.11399672	1
##	ENSG00000261600.1	-1.707952	-0.4107199	-1.230485	0.11404171	1
##	ENSG00000162078.7	2.190884	-1.4227673	1.226245	0.11609204	1

```
# Notice that much fewer genes are highlighted in the MAplot
par(mfrow=c(1,3))
plotMD(fit.treat,coef=1,status=res.treat[, "SFivsNIT"], values=c(-1,1))
abline(h=0,col="grey")
plotMD(fit.treat,coef=2,status=res.treat[, "SFivsELI"], values=c(-1,1))
abline(h=0,col="grey")
plotMD(fit.treat,coef=2,status=res.treat[, "NITvsELI"], values=c(-1,1))
abline(h=0,col="grey")
```

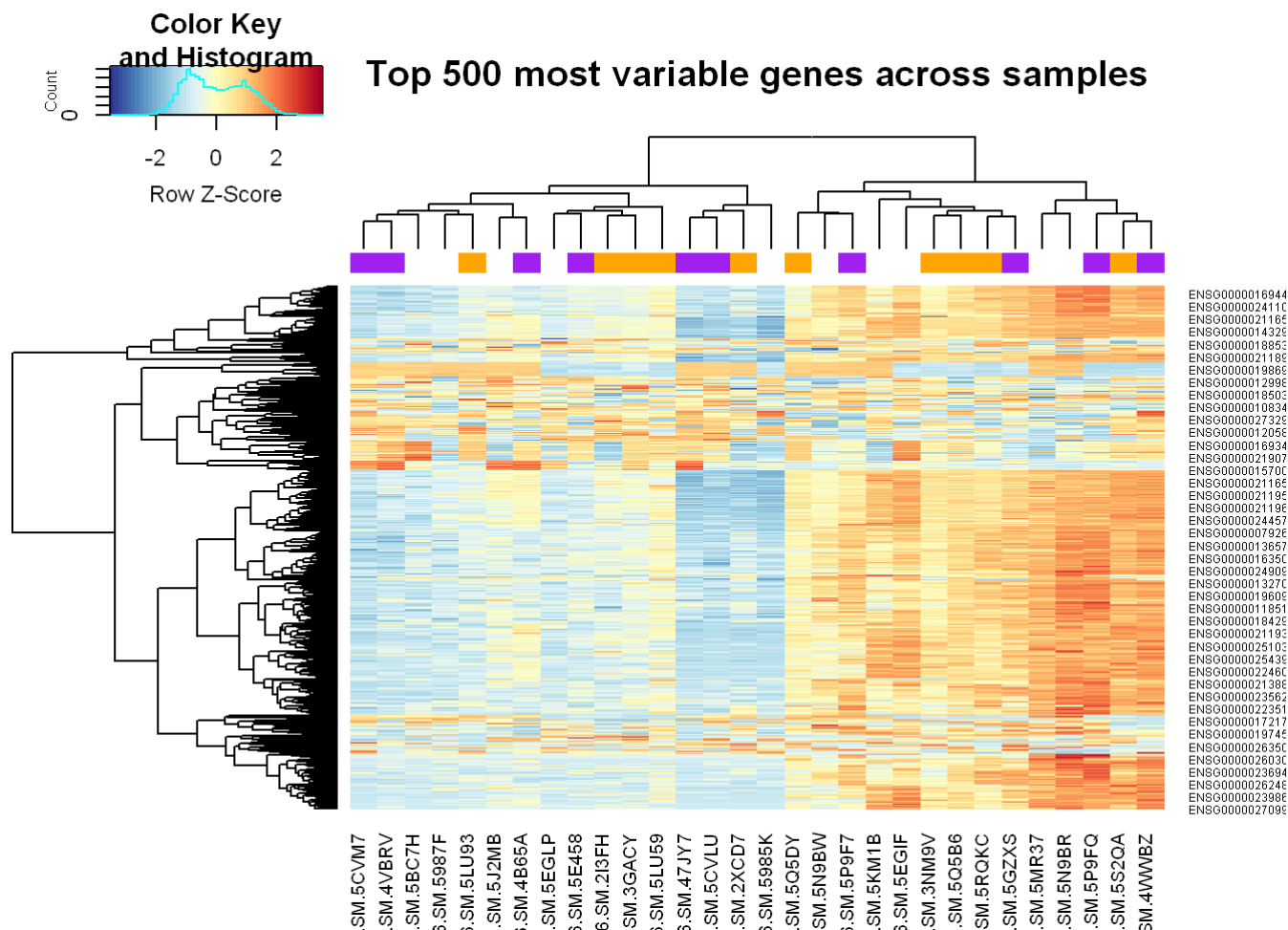


18. HeatMap

Finalmente dibujaremos el HeatMap que no se representó en el apartado 12 y quedaba pendiente.

```
## Get some nicer colours
mypalette <- brewer.pal(11,"RdYlBu")
morecols <- colorRampPalette(mypalette)
# Set up colour vector for celltype variable
col.cell <- c("purple","orange")[stargets$Group]

# Plot the heatmap
par(mfrow=c(1,1))
heatmap.2(highly_variable_lcpm,col=rev(morecols(50)),trace="none", main="Top 500 most
variable genes across samples",ColSideColors=col.cell,scale="row")
```



19. Referencias:

www.google.com

RNAseqTutorialUOCv2.html

[Statistical analysis of RNA-seq data.pdf](#)

[IntroToAnnotationPackages.pdf](#)

ENLACE A GitHub:

<https://github.com/cmbosch/PAC2> (<https://github.com/cmbosch/PAC2>)