Análisis de datos Ómnicos (M0 - 157) - PEC 1

MYC Transactome Mapped by global Array-based Nuclear Run - on (ANRO - Affymetrix)

Esther Martí

8/4/2020

Contents

```
## Loading required package: printr
```

```
## Registered S3 method overwritten by 'printr':
##   method                from
##   knit_print.data.frame rmarkdown
```

1 Antes de empezar.

Antes de empezar a trabajar se tienen que instalar los siguientes paquetes.

Primero instalamos el paquete basico bioconductor.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install()
```

y seguidamente todos los paquetes que vamos a necesitar

```
install.packages("knitcitations")
install.packages("knitr")
install.packages("colorspace")
install.packages("gplots")
install.packages("ggplot2")
install.packages("ggrepel")
install.packages("htmlTable")
install.packages("prettydoc")
install.packages("devtools")
install.packages("BiocManager")
BiocManager::install("oligo")
BiocManager::install("arrayQualityMetrics")
BiocManager::install("pvca")
# NOT NEEDED UNTIL ANALYSES ARE PERFORMED
BiocManager::install("limma")
BiocManager::install("org.Hs.eg.db")
BiocManager::install("genefilter")
BiocManager::install("pd.huex.1.0.st.v2")
BiocManager::install("annotate")
BiocManager::install("org.Mm.eg.db")
BiocManager::install("ReactomePA")
BiocManager::install("reactome.db")
```

Seguidamente he creado los siguientes directorios.

Data : Donde almacenaré todos los ficheros de entrada, recogidos de la base de datos www.ncbi.nlm.nih.gov/geo/ **figures** : Directorio donde almacenaré todos los gráficos que vaya obteniendo. **results** : Dónde almacenaré todos los ficheros de resultados que vaya almacenando.

2 Abstract (Resumen)

El estudio describe un método para medir la transcripción naciente de genes nucleares con un ensayo Nuclear Run-On (ANRO), basado en Array utilizando plataformas comerciales en microarrays. Las mediciones ANRO en un modelo de celulas B P493-6 humano que expresa c-Myc inducible se realizaron después de 48 horas con o sin inducción del gen MYC, Las muestras se preparaon a partir de ARN Nuclear y ARN Total.

Todo el trabajo está subido en el repositorio con la siguiente url, https://github.com/EstheMar04/Marti_Esther_ADO_PEC1

3 Objetivo del Estudio

La expresión génica se compara a nivel global después de 48 horas con y sin tetraciclina, medido tanto para el ARNO como para el ARN total.

4 Materiales y Métodos

4.1 Naturaleza de los datos, tipo de experimento, tipo de microarrays utilizados

Los datos estan identificados con el número de accesion: **GSE17239** , el tratamiento que utiliza es comparar el nivel global con y sin tetraciclina. Los datos se encuentran en ficheros CELL. Para poder importarlos he preparado un fichero csv, creando 4 grupos dependiendo si se trabaja con ANRO o con RNA Total y para ambos el tratamientto utilizado con o sin tetracilina. Los grupos son NRO_NoTet (ANRO sin tetraciclina), NRO_Tet (ANRO con tetraciclina), Total_Tet (RNA Total con tetraciclina) y Total_Tet (RNA Total sin tetraciclina).

Table 1: Contiene los datos de los ficheros utilizados

ï..NomFile	Group	Tipo.RNA	Treatment	ShortName
GSM431552	NRO_NoTet	NRO	No Tet	U7 NRO
GSM431553	NRO_NoTet	NRO	No Tet	U6 NRO
GSM431554	NRO_NoTet	NRO	No Tet	U5 NRO
GSM431555	NRO_NoTet	NRO	No Tet	U4 NRO
GSM431556	NRO_Tet	NRO	Tet maintained	T7 NRO
GSM431557	NRO_Tet	NRO	Tet maintained	T6 NRO
GSM431558	NRO_Tet	NRO	Tet maintained	T5 NRO
GSM431559	NRO_Tet	NRO	Tet maintained	T4 NRO
GSM431560	Total_Tet	Total	Tet maintained	T4 Total
GSM431561	Total_Tet	Total	Tet maintained	T5 Total
GSM431562	Total_Tet	Total	Tet maintained	T6 Total
GSM431563	Total_Tet	Total	Tet maintained	T7 Total
GSM431564	Total_NoTet	Total	No Tet	U4 Total
GSM431565	Total_NoTet	Total	No Tet	U5 Total
GSM431566	Total_NoTet	Total	No Tet	U6 Total
GSM431567	Total_NoTet	Total	No Tet	U7 Total

La plataforma de **Microarrays Affimetrix** Exon.

El siguiente paso es leer los ficheros CELL, para ello leemos primero la lista de ficheros que tenemos en el directorio **Data** y los alamacenamos en la variable **CelFiles** , seguidamente creamos una variable **my.targets**

en la cual cruzamos los datos que tenemos en el fichero targets.csv. Finalmente podemos cruzar la información de ambos ficheros en la variable **RowData** y le cambiamos el nombre de las columnas por nuestro nombre corto del fichero.

5 Métodos utilizados en el análisis

5.1 Control de calidad

El siguiente paso que debemos realizar, es comprobar si los datos tienen suficiente calidad para la normalización. Si no fuera así ocurriria que se introduciria mucho ruido en el análisis, procando no poder resolver el proceso.

Uso el paquete **ArrayQualityMetrics** que realiza diferentes enfoques de calidad, como diagrama de caja de la intensidad de los datos y Análisis de componentes principales (PCA), entre otros.

Se puede obtener un análisis má completo de los datos utilizando funciones especificas diseñadas para dicho análisis. Mostramos en un grafico el resultado de este análisis, segun los grupos que tenemos montados.

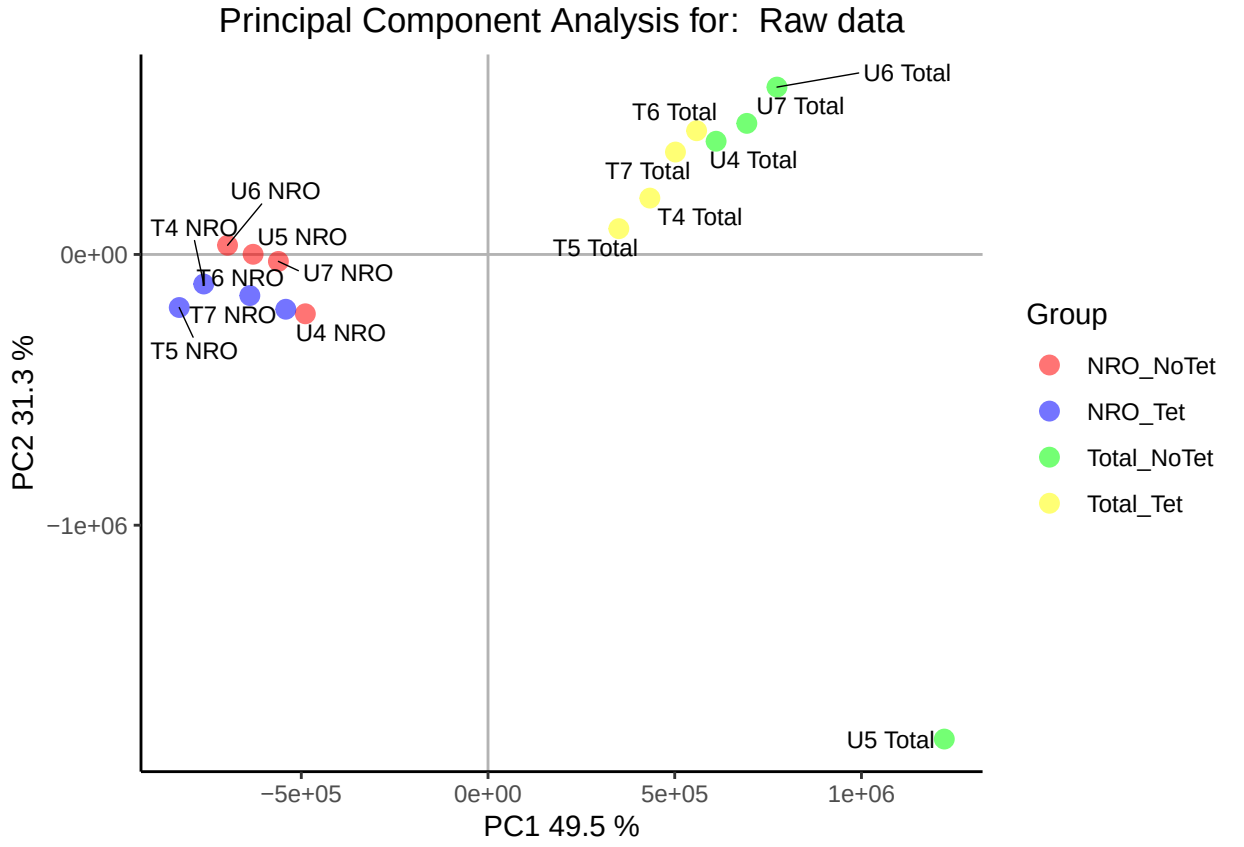


Figure 1: Visualización de los dos principales componentes de RawData

En este grafico de observamos que: * Las etiquetas de cada uno de los puntos de la gráfica son los nombres cortos que indicamos en nuestro fichero de targets. * Las características de cada tipo de muestras, es la columna de grupo, también de nuestro fichero targets * Finalmente los colores tenemos 4, uno para cada uno de nuestros grupos.

En este gráfico vemos que obtenemos una PCA de 49,5% que es el total de la variabilidad de las muestras, dependiendo si se trabaja con RNA total o con ANRO. tenemos en la parte izquierda del grafico las muestras que trabajan con ANRO y en la parte derecha las que trabajan con RNA Total.

De la misma manera podemos obtener un gráfico para ver las intensidades de las muestras utilizando la función boxplot.

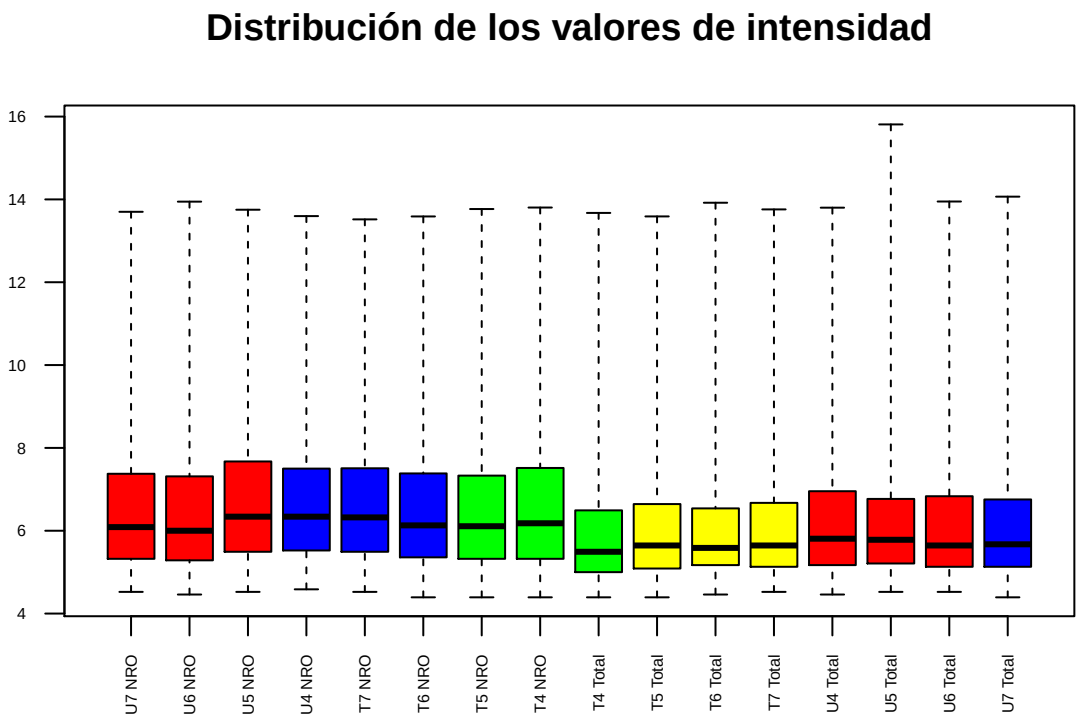


Figure 2: Boxplot for arrays intensities (Raw Data)

5.2 Normalización de lo datos

Antes de iniciar el analisis debemos de conseguir que las matrices sean comparables entre ellas, para ello debemos conseguir reducir o eliminar la variabilidad de las muestras que no se deba a razones biológicas. Para ello debemos normalizar los datos.

```
## Background correcting
## Normalizing
## Calculating Expression
```

5.3 Control de calidad después de la normalización

Después de normalizar los datos debemos volver a comprobar los datos, para ver si hemos conseguido cambiar la variabilidad de las muestras. Lo hacemos de la misma manera que lo hemos hecho anteriormente.

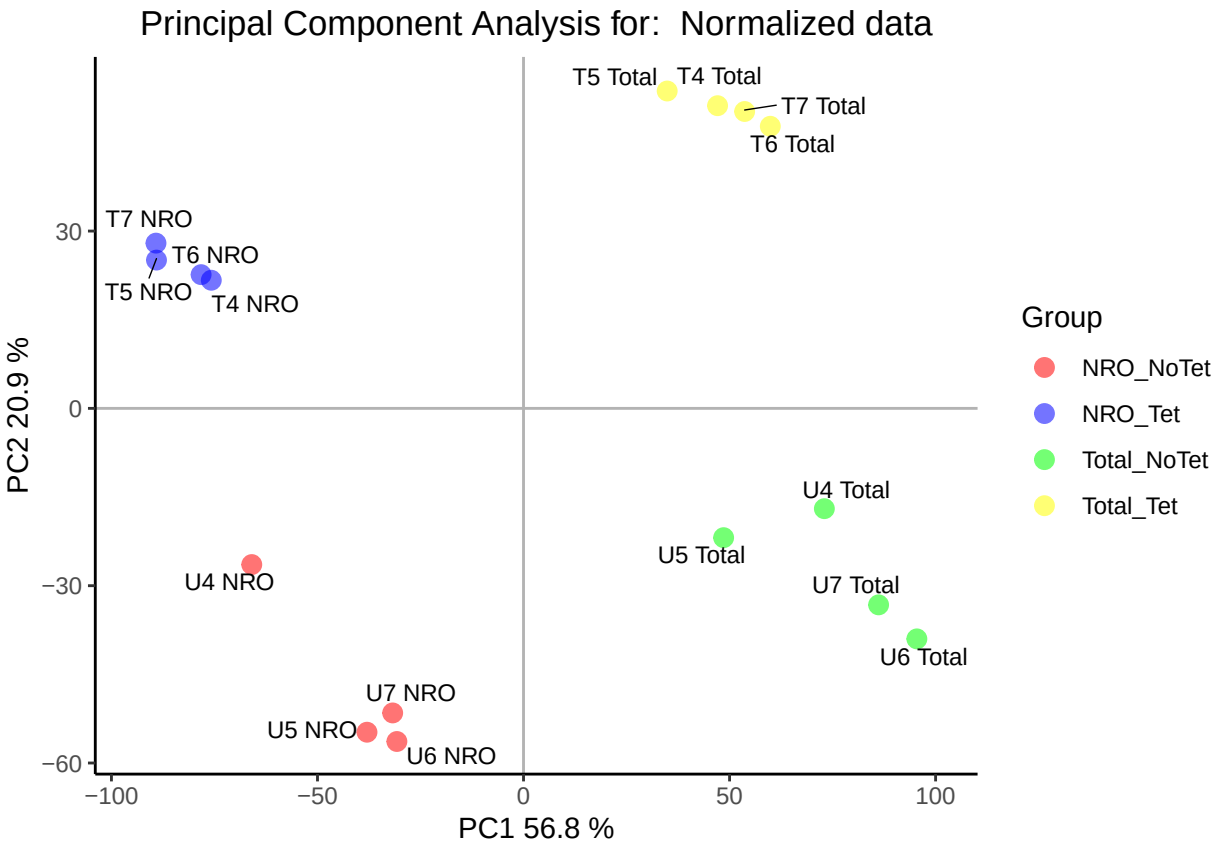


Figure 3: Visualización de los dos principales componente de los datos normalizados

Mostramos las muestras después de normalizar en un grafico el resultado de este análisis, según los grupos que en los que noe hemos basado.

Una vez normalizados los datos la variabilidad a aumentado a 56.9% y visualmente ya vemos que tenemos a la derecha las muestras de ARN Total y a la izquierda las de ANRO, pero ahora vemos qtanto en ANRO como en ARN total tenemos en la parte positiva las muestras tratadas con tetraciclina y en la parte negativa las muestras que no han sido tratadas con tetraciclina.

Finalmente mostramos los datos normalizados con boxplot.

```
boxplot(eset_rma, cex.axis=0.5, las=2,  which="all",
        col = c(rep("red", 3), rep("blue", 3), rep("green", 3), rep("yellow", 3)),
        main="Boxplot for arrays intensity: Normalized Data")
```

```
## Warning in .local(x, ...): Argument 'which' ignored (not meaningful for
## ExpressionSet)
```

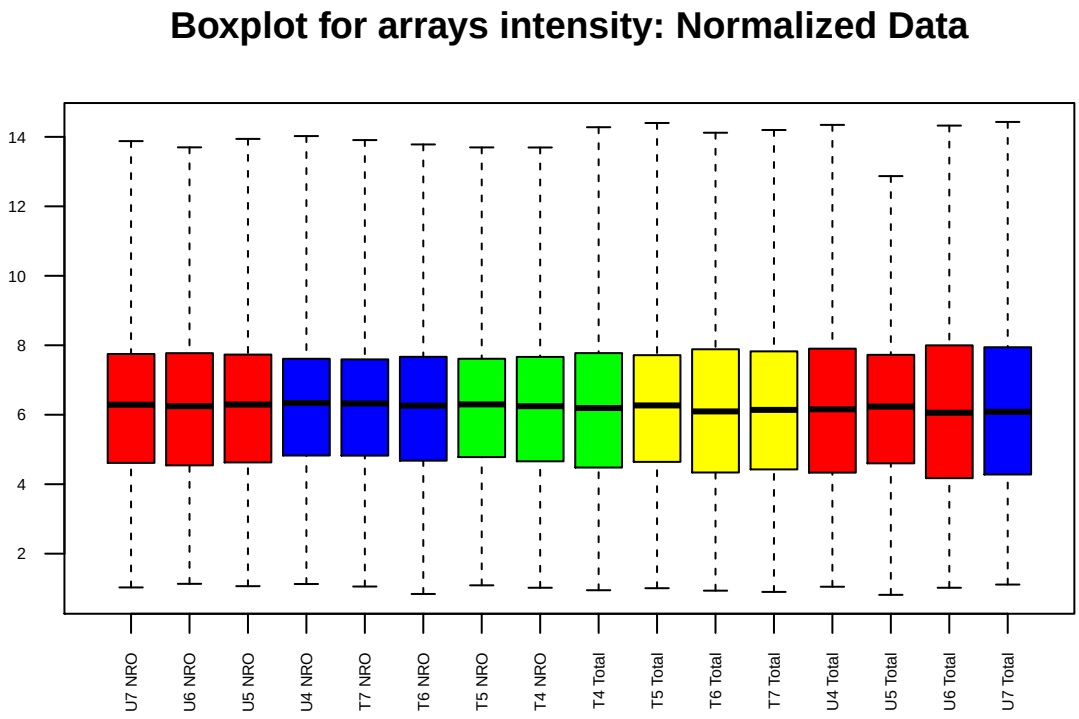


Figure 4: Distribution of intensities for normalized data

5.4 Detección de Lotes

Los resultados que obtengamos de los microarrays tienen pequeñas diferencias según el lote de los reactivos, los técnicos que realicen el test e incluso la fecha en la que se hace el experimento. El error acumulativo introducido por estas variaciones experimentales dependientes del tiempo y el lugar se denomina “efectos

por lotes". Existen enfoques para identificar y eliminar los efectos por lote como el análisis de variables sustitutas, el análisis de componentes de variación principal y de combate (PVCA).

```
#load the library
library(pvca)
pData(eset_rma) <- targets
#select the threshold
pct_threshold <- 0.6
#select the factors to analyze
batch.factors <- c("Tipo.RNA", "Treatment")
#run the analysis
pvcaObj <- pvcaBatchAssess (eset_rma, batch.factors, pct_threshold)
```

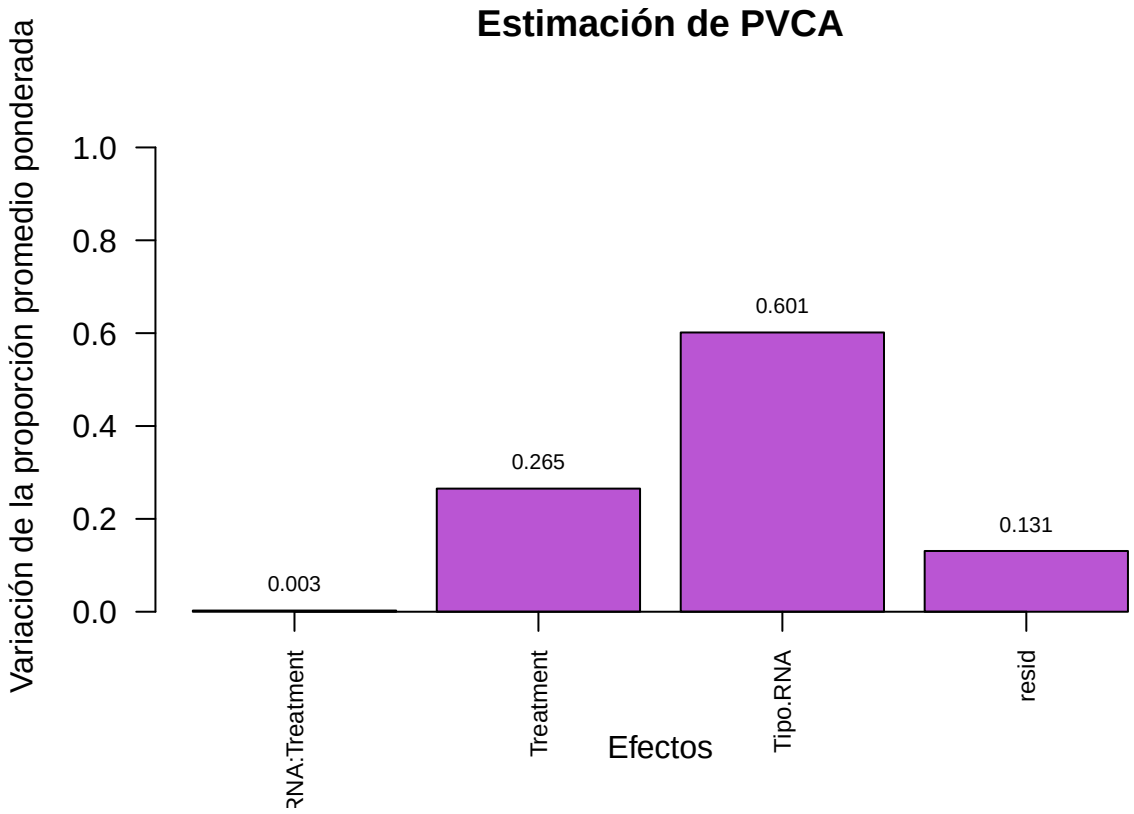


Figure 5: Relativa a la importancia para los distintos factores con los que trabajamos Tipo.RNA y Tratamiento e interacción afectando a la expresión del Gen

5.5 Detección de la mayoría de los genes variables

En el siguiente gráfico mostramos las desviaciones estándar de todos los genes ordenados de menor a mayor. Los genes más variables son aquellos con una desviación estándar suoerior a 90%-95% de todas las desviaciones estándar. Es decir, tal y como vemos en el gráfico son todas aquellas con valores superiores a 20000.

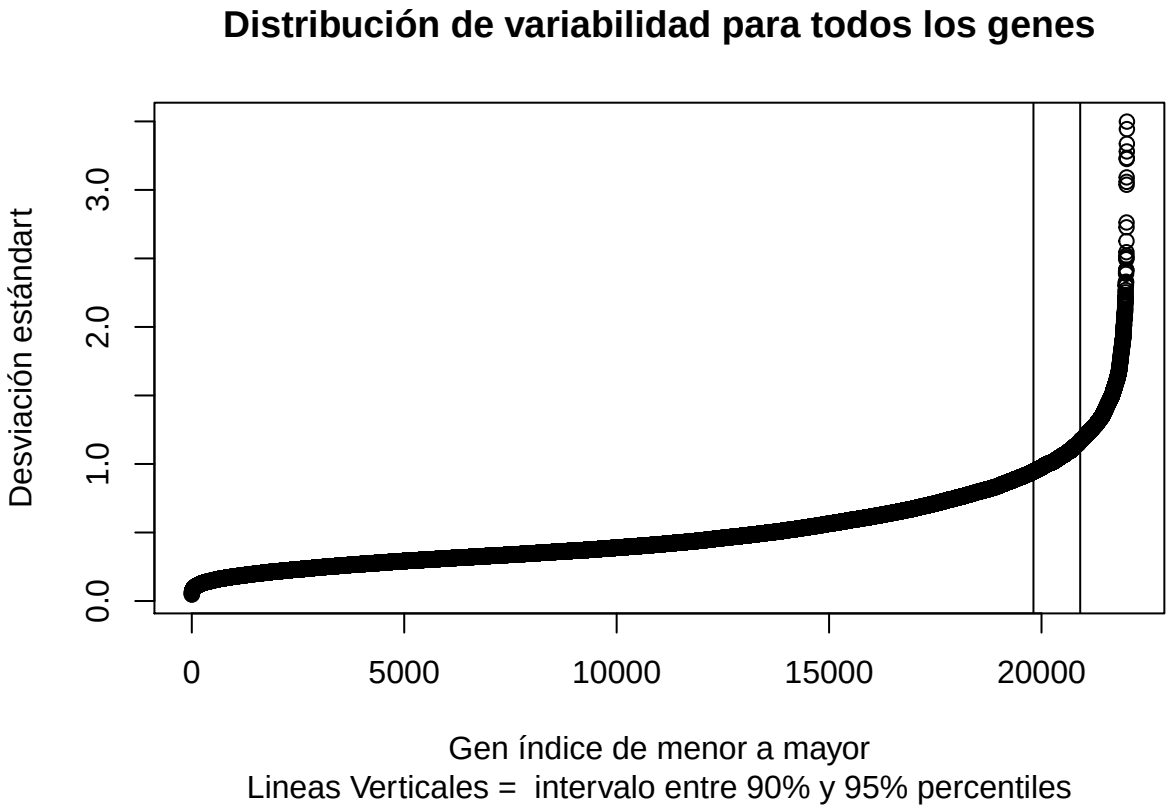


Figure 6: Los valores de las desviaciones estándar abarcan todas las muestras para todos los genes ordenados de menor a mayor

5.6 Filtraje no específico

Este es recomendable para eliminar el ruido de fondo y limitar los ajustes posteriores a los necesarios. Los principales son tres:

- Eliminación de spots marcados como erróneos mediante flags
- Eliminación de spots con señales muy bajas debido a problemas en el spoting.
- **Eliminación de genes que no presenten una variación significativa en su señal** , es decir filtraje por variabilidad. Este es el que vamos a utilzar a continuación debido a que nos permite reducir el filtraje al mínimo.

Para hacer este filtraje he buscado en la información de GSE17239 y este trabaja con [HuEx-1_0-st] Affymetrix Human Exon 1.0 ST Array [transcript (gene) version], debido a que no conseguia encontrar que base de datos existia para el paquete pd.huex.1.0.st.v2, encontré en Internet la siguiente información:

"The annotation packages for the Gene and Exon ST arrays in the current Bioconductor release (versions 8.3.0) are based on the na34 annotation files distributed by Affymetrix. It appears that there were some problems with their annotation pipeline. We asked Affy about the mouse Exon 1.0 ST arrays, and the response was:_

"There was a large issue with the pipeline used for NA34, which is why the annotation numbers are so low. It had significant issues and Affymetrix was well aware of them. We took a great deal of time and generated a new pipeline for NA35 that we hope will resolve a great many issues people were experiencing with NA34."_

The na35 data were released 15 April, which was after we released BioC 3.1. They have since (as of 1 June) released na35.1 versions for some arrays, which hopefully will be the last release from them. I am currently re-building all of the annotation packages for these arrays, and will hopefully have them available for download by the end of this week/early next week.

Para ello usamos la función en R **nsFilter**, que nos devuelve un report con los resultados filtrados.

```
## $numDupsRemoved
## [1] 169
##
## $numLowVar
## [1] 10984
##
## $numRemoved.ENTREZID
## [1] 7197
```

Después de filtrar hay 3661 genes a la izquierda.Estos datos los hemos almacenado en la variable *eset_filtered*

5.7 Guardamos los ficheros de normalización y de filtro

Empezamos a guardar resultados en el directorio results que hemos creado al principio. Guardamos los datos en ficheros csv de los datos normalizados, los datos filtrados y los datos nuevamente normalizados después de filtrarlos.

```
write.csv(exprs(eset_rma), file="./results/normalized.Data.csv")
write.csv(exprs(eset_filtered), file="./results/normalized.Filtered.Data.csv")
save(eset_rma, eset_filtered, file="./results/normalized.Data.Rda")
```

5.8 Diseño de la Matriz

A continuación muestro una Matriz basándome en los grupos que he introducido en el fichero targets. El estudio se basa en cuatro niveles ANRO/RNA Total combinado con tratado con tetracilcina o no tratados.

```
##    NRO_NoTet NRO_Tet Total_Tet Total_NoTet
## 1          1       0         0           0
## 2          1       0         0           0
## 3          1       0         0           0
## 4          1       0         0           0
## 5          0       1         0           0
## 6          0       1         0           0
## 7          0       1         0           0
## 8          0       1         0           0
## 9          0       0         0           1
## 10         0       0         0           1
## 11         0       0         0           1
## 12         0       0         0           1
## 13         0       0         1           0
## 14         0       0         1           0
## 15         0       0         1           0
## 16         0       0         1           0
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$Group
## [1] "contr.treatment"
```

5.9 Definir comparación con contrastes

He hecho una comparación de las muestras, dependiendo de la preparación de las muestras, tal y como he indicado al principio, se prepararon a partir de ARN Nuclear (NRO) y ARN Total (Total). Estas muestras se trataron con (Tet) o sin tetraciclina (noTet).

Así pues he realizado las siguientes tres comparaciones, * NROvsTotal.Tet -> ARN Nuclear versus ARN Total que han sido tratados con Tetraciclina * NROvsTotal.NoTet -> ARN Nuclear versus ARN Total que no han sido tratados con Tetraciclina * INT -> una comparativa de las dos anteriores.

A continuación muestro una tabla con el resultado obtenido

```
cont.matrix <- makeContrasts (NROvsTotal.Tet = NRO_Tet-Total_Tet,
                              NROvsTotal.NoTet = NRO_NoTet-Total_NoTet,
                              INT = (NRO_Tet-Total_Tet) - (NRO_NoTet-Total_NoTet),
                              levels=designMat)
print(cont.matrix)
```

```
##              Contrasts
## Levels        NROvsTotal.Tet NROvsTotal.NoTet INT
##   NRO_NoTet                0                1  -1
##   NRO_Tet                  1                0   1
##   Total_Tet               -1                0  -1
##   Total_NoTet              0               -1   1
```

5.10 Estimación del modelo y selección de genes

Una vez definida la matriza y las comparaciones, se puede pasar a estimar el modelo y los contrastes y realizar las pruebas de significación las cuales nos van a ayudar a tomar la decisión.

Dentro del paquete lima, tenemos los modelos empíricos de Bayes para combinar una estimación de variabilidad basada en la matriz completa con estimaciones individuales basadas en cada valor individual proporcionadas por estimaciones de error mejoradas.

El análisis proporciona las estadísticas de pruebas habituales para ordenar los genes expresados diferencialmente de mayor a menor, segun p-valor. Ajustamos estos valores de p-valor para tener un control sobre los falsos positivos.

Almacenamos los valores en una clase con el nombre fit.main.

```
library(limma)
fit<-lmFit(eset_filtered, designMat)
fit.main<-contrasts.fit(fit, cont.matrix)
fit.main<-eBayes(fit.main)
class(fit.main)
```

```
## [1] "MArrayLM"
## attr(,"package")
## [1] "limma"
```

5.11 Obtención de listas de genes expresados diferencialmente

Podemos obtener una vista para las primeras lineas de cada tabla.

Para la primera comparación NROvsTotal: Genes que cambian su expresión de los distintos RNA según si son tratados con tetraciclina:

	logFC	AveExpr	t	P.Value	adj.P.Val	B
3334125	-4.361007	9.229873	-37.08962	0	0	28.67555
3146433	-5.256716	9.445767	-33.71231	0	0	27.31271
3293280	-4.512346	8.149439	-32.24019	0	0	26.66444
3565303	-3.353582	7.810610	-28.48601	0	0	24.83646
3903361	-3.407773	9.389357	-27.69724	0	0	24.41647
3292413	-4.271833	6.461712	-27.16445	0	0	24.12495

Explicación de cada columna:

- La Primera columna de cada tabla contiene el ID de la fabrica de Affimetrix para cada probeset. El siguiente paso se corresponde a cada ID, este proceso es llamado **annotation**.
- `logFC`: Diferencia entre grupos.

- `AveExpr`: Promedio de todos los genes en la comparación.
- `t` : Estadística t moderada.
- `P.Value`: p-valor.

- `adj.P.Val`: p-valor ajustado

- `B`: Estadístiaca B

Para la segunda compraración (NROvsTotal): Genes que cambian su expresión de los distintos RNA según si son tratados sin tetraciclina, con las mismas columnas que en la tabla anterior:

	logFC	AveExpr	t	P.Value	adj.P.Val	B
3674199	4.126386	8.730105	34.02862	0	0	27.38117
3334125	-3.701858	9.229873	-31.48367	0	0	26.26547
3830993	-3.994580	8.407288	-29.16947	0	0	25.15078
3333622	-3.152174	9.417919	-26.73805	0	0	23.86087
2326448	-3.490911	8.856791	-26.38518	0	0	23.66243
3872983	-2.705796	10.058237	-26.12678	0	0	23.51519

Finalmente para la última comparación, con las mismas columnas mostramos los Genes que difieren entre ambas comparaciones anteriores

	logFC	AveExpr	t	P.Value	adj.P.Val	B
3933817	-4.616043	8.566163	-29.32409	0	0	24.97735
3674199	-4.986847	8.730105	-29.07941	0	0	24.86062
3742727	-4.248686	7.401830	-27.43246	0	0	24.04115
3677752	-5.035445	9.427868	-27.29102	0	0	23.96791
3401099	-5.989461	8.588911	-26.18988	0	0	23.38130
3715489	-6.071859	7.276322	-25.91267	0	0	23.22887

5.12 Gene Annotation

Una vez tenemos las tables con información adicional las características han sido selecionadas. Este proceso en llamado anotación y esencialmente lo que hace es buscar información para asociar identificadores que aparecen en la tabla superior.

Finalmente almacenamos el resultado en el fichero topAnnotated_INT.csv.

Mostramos una tabla con los datos obtenidos.

```
##             logFC  AveExpr         t      P.Value
## 3334125 -4.361007 9.229873 -37.08962 7.356271e-17
## 3146433 -5.256716 9.445767 -33.71231 3.293024e-16
## 3293280 -4.512346 8.149439 -32.24019 6.629420e-16
## 3565303 -3.353581 7.810610 -28.48601 4.591201e-15
## 3903361 -3.407773 9.389357 -27.69724 7.113203e-15
```

5.13 Visualizando la expresión diferencial

Para poder mostrar la visualización de expresión diferencial tenemos el plot de volcan, que nos muestra la cantidad de genes que contiene.

Mostramos la tabla y almacenamos la grafica en el directorio figures.

```
library(huex10sttranscriptcluster.db)
geneSymbols <- select(huex10sttranscriptcluster.db, rownames(fit.main), c("SYMBOL"))
SYMBOLS<- geneSymbols$SYMBOL
volcanoplot(fit.main, coef=1, highlight=4, names=SYMBOLS,
```

```
          main=paste("Genes expresados diferencialmente", colnames(cont.matrix)[1], sep="\n"))
abline(v=c(-2.5,2.5))
```

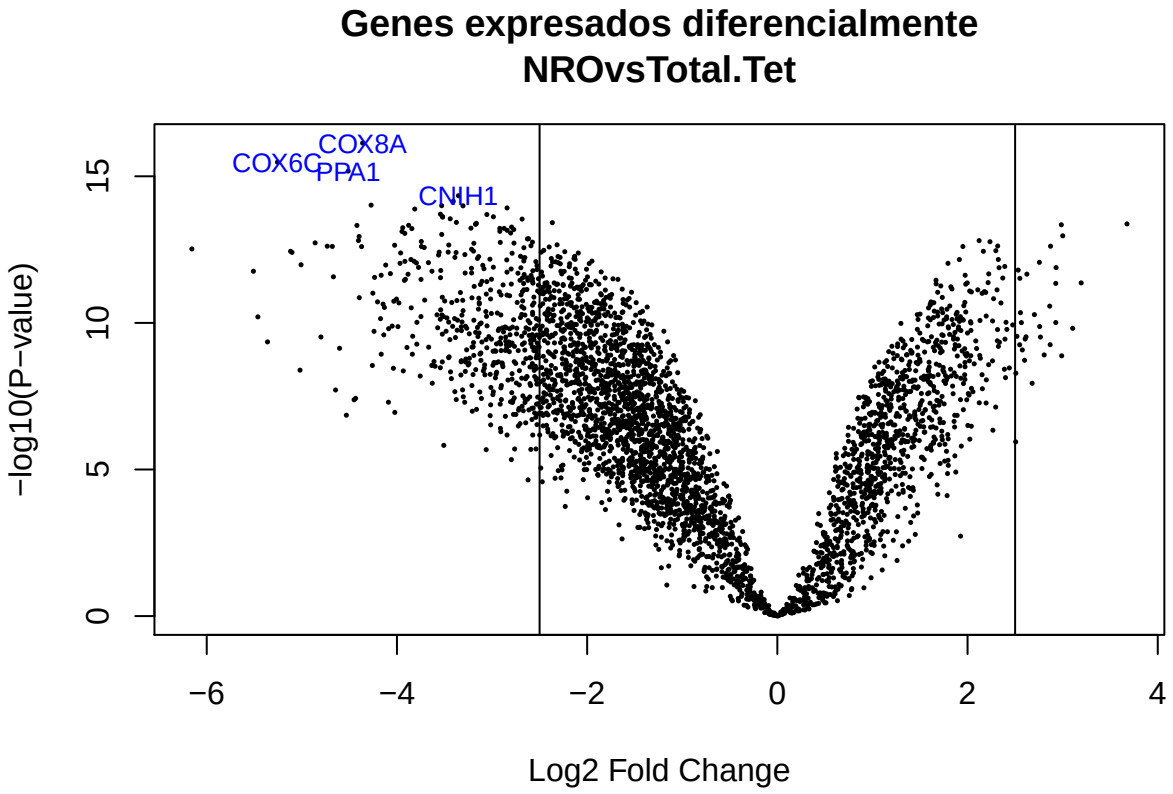


Figure 7: Plot Volcan para mostrar la comparación entr ANRO y ARN total tratados con tetraciclina y sin

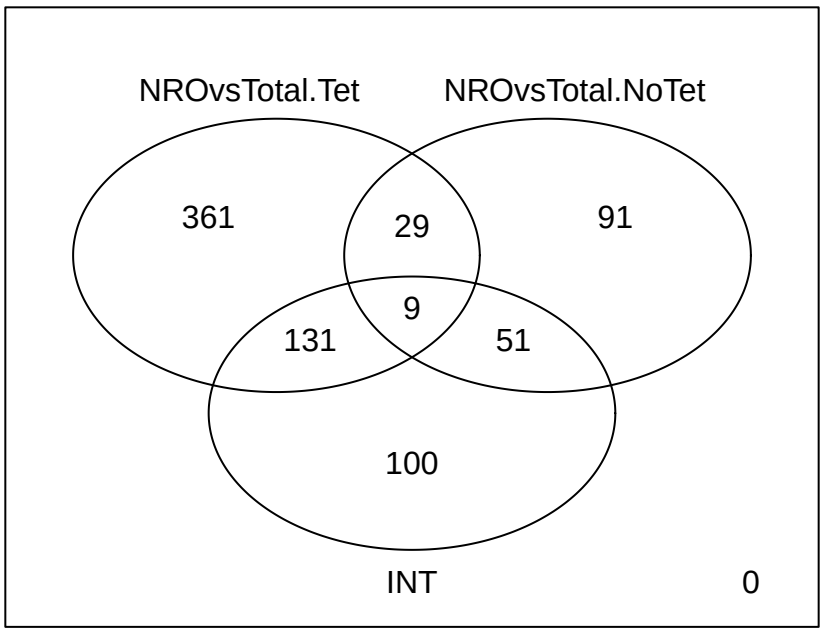
5.14 Múltiples comparaciones

Al seleccionar los genes en varias comparaciones, hay que conocer que genes se han seleccionado en cada comparación . En algunas ocasiones, los genes biológicamente relevantes serán aquellos que se seleccionan en uno de ellos pero no en otros. sin embargo en otras ocasiones, su interés radicará en los genes que se seleccionan en todas las comparaciones. Dentro del paquete lima tenemos la función `decideTests` y `vennDiagram` con los cuales podemos recontar los genes.

La primera función nos muestra los datos agrupados en columnas y la segunda función nos muestra una grafica en la que podemos ver como estan agrupados los genes.

```
##        NROvsTotal.Tet NROvsTotal.NoTet  INT
## Down              434              125  211
## NotSig           3131             3481 3370
## Up                 96               55   80
```

```
vennDiagram (res.selected[,1:3], cex=1)
title("Genes más comunes de las tres comparaciones\n Genes seleccionados con FDR < 0.0000000001 y logFC
```



Mapa de calor

```
probesInHeatmap <- rownames(res.selected)
HMdata <- exprs(eset_filtered)[rownames(exprs(eset_filtered)) %in% probesInHeatmap,]

geneSymbols <- select(huex10sttranscriptcluster.db, rownames(HMdata), c("SYMBOL"))
SYMBOLS<- geneSymbols$SYMBOL
rownames(HMdata) <- SYMBOLS
write.csv(HMdata, file = file.path("./results/data4Heatmap.csv"))
```

```
my_palette <- colorRampPalette(c("blue", "red"))(n = 299)
library(gplots)

heatmap.2(HMdata,
          Rowv = FALSE,
          Colv = FALSE,
          main = "Genes expresados diferencialmente \n FDR < 0.000000001, logFC >=1",
          scale = "row",
          col = my_palette,
          sepcolor = "white",
          sepwidth = c(0.05,0.05),
          cexRow = 1,
          cexCol = 1,
          key = TRUE,
          keysize = 1.5,
          density.info = "histogram",
```

```
          ColSideColors = c(rep("red",4),rep("blue",4), rep("green",4), rep("yellow",4)),
          tracecol = NULL,
          dendrogram = "none",
          srtCol = 30)
```

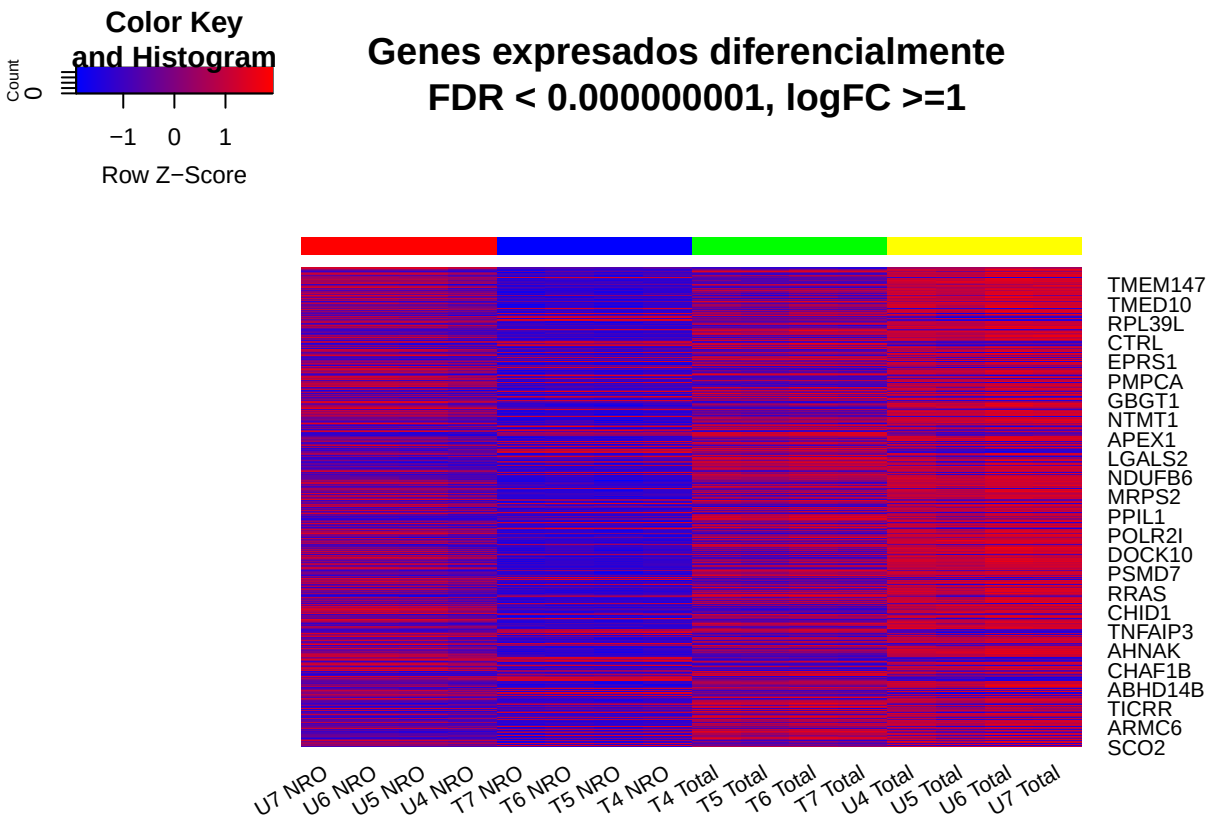


Figure 8: Heatmap for expression data without any grouping

6 Significado Biologico de los resultados

Finalmente una vez tenemos todas las comparaciones entre los genes, estas deben ser interpretadas.

Como primer paso, preparamos la lista de listas de genes que se analizarán:

```
listOfTables <- list(NROvsTotal.Tet = topTab_NROvsTotal.Tet,
                     NROvsTotal.NoTet  = topTab_NROvsTotal.NoTet,
                     INT = topTab_INT)
listOfSelected <- list()
for (i in 1:length(listOfTables)){
  topTab <- listOfTables[[i]]
  whichGenes<-topTab["adj.P.Val"]<0.15
  selectedIDs <- rownames(topTab)[whichGenes]
  EntrezIDs<- select(huex10sttranscriptcluster.db, selectedIDs, c("ENTREZID"))
  EntrezIDs <- EntrezIDs$ENTREZID
```

```
  listOfSelected[[i]] <- EntrezIDs
  names(listOfSelected)[i] <- names(listOfTables)[i]
}
sapply(listOfSelected, length)
```

```
##   NROvsTotal.Tet NROvsTotal.NoTet              INT
##             3471             3274             3034
```

Cómo segunda opción definimos nuestro universo como todos los genes que tienen al menos una anotación en la ontología genetica,

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


mapped_genes2GO <- mappedkeys(org.Hs.egGO)
mapped_genes2KEGG <- mappedkeys(org.Hs.egPATH)
mapped_genes <- union(mapped_genes2GO , mapped_genes2KEGG)
```

A las dos primeras listas se les aplica el análisis de significación biológica.

```
## ####################################
## Comparación:  NROvsTotal.Tet
##                          ID                           Description GeneRatio
## R-HSA-69620     R-HSA-69620               Cell Cycle Checkpoints  155/2262
## R-HSA-69306     R-HSA-69306                      DNA Replication   82/2262
## R-HSA-72766     R-HSA-72766                          Translation  142/2262
## R-HSA-69239     R-HSA-69239                    Synthesis of DNA   78/2262
## R-HSA-5368287 R-HSA-5368287            Mitochondrial translation   66/2262
## R-HSA-5368286 R-HSA-5368286 Mitochondrial translation initiation   62/2262
##                  BgRatio       pvalue     p.adjust       qvalue
## R-HSA-69620   290/10616 3.664889e-34 4.881632e-31 3.244391e-31
## R-HSA-69306   127/10616 4.770839e-26 2.308457e-23 1.534228e-23
## R-HSA-72766   291/10616 5.199228e-26 2.308457e-23 1.534228e-23
## R-HSA-69239   119/10616 1.786973e-25 5.950621e-23 3.954854e-23
## R-HSA-5368287  93/10616 8.985793e-25 2.393815e-22 1.590958e-22
## R-HSA-5368286  87/10616 1.666433e-23 3.170984e-21 2.107474e-21
##
## R-HSA-69620   BIRC5/NUP37/PSMB5/H2BC14/CCNB2/CDC6/CDC20/PSMB2/RANGAP1/PSMC3/MAPRE1/PSMD9/AURKB/CDK1/C
## R-HSA-69306
## R-HSA-72766                                PPA1/RPS15A/EIF3K/MRPL51/MRPS35/MRPS7/AIMP2/EIF3H/MRPL58/MRI
## R-HSA-69239
## R-HSA-5368287
## R-HSA-5368286
##               Count
## R-HSA-69620     155
## R-HSA-69306      82
## R-HSA-72766     142
## R-HSA-69239      78
## R-HSA-5368287    66
## R-HSA-5368286    62
## ####################################
## Comparación:  NROvsTotal.NoTet
```

```
##                       ID                   Description GeneRatio   BgRatio
## R-HSA-69620 R-HSA-69620         Cell Cycle Checkpoints  132/2126 290/10616
## R-HSA-69306 R-HSA-69306                DNA Replication   73/2126 127/10616
## R-HSA-69239 R-HSA-69239               Synthesis of DNA   70/2126 119/10616
## R-HSA-69481 R-HSA-69481               G2/M Checkpoints   84/2126 167/10616
## R-HSA-69002 R-HSA-69002 DNA Replication Pre-Initiation   54/2126  85/10616
## R-HSA-69206 R-HSA-69206                G1/S Transition   71/2126 131/10616
##                     pvalue     p.adjust       qvalue
## R-HSA-69620 2.361923e-23 3.150805e-20 2.163024e-20
## R-HSA-69306 1.091169e-20 4.975591e-18 3.415738e-18
## R-HSA-69239 1.118949e-20 4.975591e-18 3.415738e-18
## R-HSA-69481 1.349497e-18 4.500572e-16 3.089638e-16
## R-HSA-69002 2.193442e-18 5.852104e-16 4.017463e-16
## R-HSA-69206 3.252011e-18 7.230304e-16 4.963596e-16
##
## R-HSA-69620 PSMB10/H4C2/ANAPC16/CENPT/RANGAP1/PLK1/PSME1/INCENP/H4C1/H4C3/MCM10/CDC45/NUP37/YWHAB/H4
## R-HSA-69306
## R-HSA-69239
## R-HSA-69481
## R-HSA-69002
## R-HSA-69206
##             Count
## R-HSA-69620   132
## R-HSA-69306    73
## R-HSA-69239    70
## R-HSA-69481    84
## R-HSA-69002    54
## R-HSA-69206    71
```

```
cnetplot(enrich.result, categorySize = "geneNum", schowCategory = 15,
       vertex.label.cex = 0.75)
```

Table 5: Primeras filas y columnas para results Reactome en comparación NROvsTtotal.Tet.csv

	Description	GeneRatio	BgRatio	pvalue	p.adjust
R-HSA-69620	Cell Cycle Checkpoints	155/2262	290/10616	3.66488885172562e-34	4.88163195049852e-31
R-HSA-69306	DNA Replication	82/2262	127/10616	4.77083866967684e-26	2.30845742184186e-23
R-HSA-72766	Translation	142/2262	291/10616	5.19922842757176e-26	2.30845742184186e-23
R-HSA-69239	Synthesis of DNA	78/2262	119/10616	1.7869732255601e-25	5.95062084111514e-23

Los Resultados que hemos obtenido son:

- un fichero csv con el resumen a partir de la función enrichPathway asociado al mapeo de los genes que hemos escogido.
- un grafico con el mejor resultado.
- un gráfico con la red de toda la información de los datos enriquecidos relacionados con los genes escogidos.

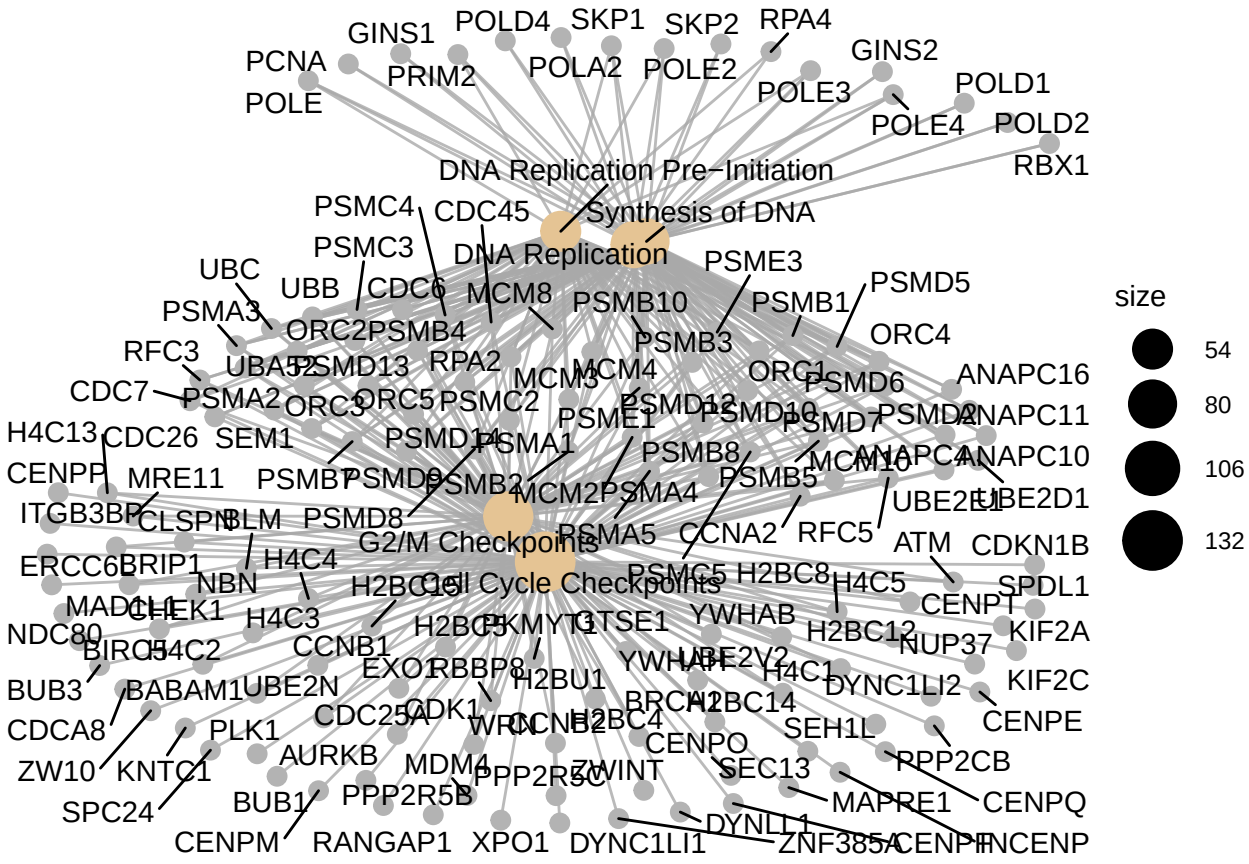


Figure 9: Red obtenida del análisis de enriquecimiento de Reactome en la lista obtenida de la comparación entre ARN nuclear y ARN total según se trate con tetraciclina o no

7 Resultados

A continuación mostramos un resumen con todos los ficheros y los resultados que hemos obtenido

Table 6: Lista de Ficjeros generada en los análisis

Lista de ficheros
data4Heatmap.csv
normalized.Data.csv
normalized.Data.Rda
normalized.Filtered.Data.csv
QCDir.Norm
ReactomePA.Results.NROvsTotal.NoTet.csv
ReactomePA.Results.NROvsTotal.Tet.csv
ReactomePABarplot.NROvsTotal.NoTet.pdf
ReactomePABarplot.NROvsTotal.Tet.pdf
ReactomePAcnetplot.NROvsTotal.NoTet.pdf
ReactomePAcnetplot.NROvsTotal.Tet.pdf
topAnnotated_INT.csv
topTab_NROvsTotal.NoTet.csv
topTab_NROvsTotal.Tet.csv

8 Apendice

8.1 Comentarios de codigo R

A continuación mostramos el código R que he utilizado para obtener el Significado Biológico de los resultados.

Primero creamos una lista con todas las comparaciones que hemos obtenido anteriormente de ARN nuclear y ARN total tratado o no tratado con tetraciclina, añadimos también la comparación entre las dos anteriores en la variable INT.

listOfTables <- list(NROvsTotal.Tet = topTab_NROvsTotal.Tet, NROvsTotal.NoTet = topTab_NROvsTotal.NoTet, INT = topTab_INT) listOfSelected <- list()

Seguidamente creamos un bucle para el conjunto de tablas que acabamos de crear en el que seleccionamos todos los genes con un p-valor ajustado < 0.15. Este resultado lo añadimos a una lista llamada listOfSelected.

for (i in 1:length(listOfTables)){

topTab <- listOfTables[[i]]

whichGenes<-topTab["adj.P.Val"]<0.15 selectedIDs <- rownames(topTab)[whichGenes]

EntrezIDs<- select(huex10sttranscriptcluster.db, selectedIDs, c("ENTREZID")) EntrezIDs <- EntrezIDs$ENTREZID listOfSelected[[i]] <- EntrezIDs names(listOfSelected)[i] <- names(listOfTables)[i] } sapply(listOfSelected, length)

Una vez tenemos la lista anterior creamos nuestro universo, para ello usamos la libreria org.Hs.eg.db, porque estamos trabajando con genes de homosapiens (human). Mapeamos egGO y egPATH y realizamos la unión de ambos, para temer todos los datos en una variable y poderlo comparar con los datos de nuestro universo.

library(org.Hs.eg.db)

mapped_genes2GO <- mappedkeys(org.Hs.egGO) mapped_genes2KEGG <- mappedkeys(org.Hs.egPATH) mapped_genes <- union(mapped_genes2GO , mapped_genes2KEGG)

A través de la libreria **ReactomePA**, comparamos la lista de los genes que hemos obtenido en el mapeo anterior con nuestro universo.

enrich.result <- enrichPathway(gene = genesIn, pvalueCutoff = 0.05, readable = T, pAdjustMethod = "BH", organism = "human", universe = universe)

Una vez tenemos los resultados lo grabamos en ficheros csv. Hay que tener en cuenta que solo trabajamos con las 2 comparaciones primeras, no trabajamos con la comparación que tenemos entre las dos primeras comparaciones.

if (length(rownames(enrich.result@result)) != 0) { write.csv(as.data.frame(enrich.result), file =paste0("./results/","ReactomePA.Results.",comparison,".csv"), row.names = FALSE)

9 Referencias

[1] _Paso 2: Crear un repositorio de GitHub_.
urlhttps://docs.aws.amazon.com/es_es/codedeploy/latest/userguide/tutorials-github-create-github-repos
2019.
##
[2] C. C. Fan J. "MYC Transactome Mapped by Global Array-based Nuclear
Run-On (ANRO - Affimetrix)". In: _Bioinformatics_ (2020).
##
[3] J. W. McDonals. _News: Rebuilding annotation packages for
Affymetrix Gene/Exon ST Arrays_.
urlhttps://support.bioconductor.org/p/68341/. 2016.
##
[4] G. Yu. _enrichPathway From ReactomePA v1.16.2_.
urlhttps://www.rdocumentation.org/packages/ReactomePA/versions/1.16.2/topics/enrichPathway.
2019.