PEC1: Análisis de Datos Omicos

Natalia Díaz González

15/4/2020

Contents

1. Introducción y Objetivos

Tras la finalización de las primeras unidades, procedemos a poner en práctica y unificar los conocimientos adquiridos mediante esta PEC cuyo objetivo es realizar un análisis de datos de Microarrays. Para ello:

- Se partirá de un problema y unos datos públicos y se reanalizarán siguiendo las pautas presentadas en los materiales y discutidas en los dos primeros debates.
- Una vez obtenidos los resultados se procederá a redactar un informe con la estructura tradicional de un informe científico-técnico.

La presentacion de esta práctica se realizará en formato PDF mediante el uso de Markdown.

Los datos y el código para el análisis se pueden localizar mediante el siguiente repositorio de github:

www.github.com/ASPteaching/Omics_Data_Analysis-Case_Study_1-Microarrays.

1.1. Palabras Clave

Microarrays, Bioconductor, R, Genes expresados Diferencialmente.

1.2. Objetivos

El objetivo de esta PEC es ilustrar el proceso de análisis de microarrays mediante la realización de un estudio, de principio a fin, tal como se llevará a cabo en una situación real. Este estudio se seleccionará de la Base de Datos "Gene Expression Omnibus (GEO)" a la que podemos acceder mediante la url:

http://www.ncbi.nlm.nih.gov/geo/browse/?view=series.

Para la selección del mismo, he seguido las pautas marcadas en el enunciado de la PEC:

- El estudio ideal debería tener pocas muestras (10-30).
- 2-3 comparaciones.
- Que se haya realizado con microarrays de marca Afymetrix, es decir, que utilice archivos .CEL

Tras revisar varios casos de estudio me he decantado por la selección del experimento **GSE67883**, relacionado con el análisis de la desregularización de hormonas de crecimiento pulmonar después de la exposición al humo del tabaco intrauterino.

La exposición prenatal al humo del tabaco es un factor de riesgo significativo para el desarrollo de enfermedades de las vías respiratorias. Además, la alta prevalencia de mujeres embarazadas que fuman requiere el establecimiento de estrategias para la protección pulmonar de la descendencia. El objetivo del estudio del artículo consiste en comprender el mecanismo molecular de cómo la exposición prenatal al humo afecta al desarrollo pulmonar fetal. Para ello, se utiliza un modelo de ratón que recapitula los hallazgos clínicos de niños expuestos prenatalmente, donde ratones embarazadas fueron expuestos al humo hasta la cesárea o parto espontáneo, y se monitoreó el desarrollo del peso de la descendencia y la función pulmonar. Con los resultados del mismo se pretende demostrar que los ratones expuestos prenatalmente muestran un retraso del crecimiento intrauterino y postnatal, y deterioro de la función pulmonar.

Sin embargo, a diferencia del objetivo del artículo, el objetivo de nuestro análisis es encontrar genes diferencialmente expresados entre aquellos ratones que se tomaron como control y aquellos que fueron expuestos al humo del tabaco (madre expuesta vs control) teniendo en cuenta el sexo de la descendencia (macho vs hembra).

2. Materiales y Métodos

2.1. Métodos

Para llevar a cabo el análisis de Microarrays debemos proceder de forma ordenada y siguiendo el método científico. Este análisis puede ser fácilmente visualizado como un proceso que empieza por una pregunta

biológica y concluye con una interpretación de los resultados de los análisis que, de alguna forma nos acerque un poco a la respuesta de la pregunta inicial.

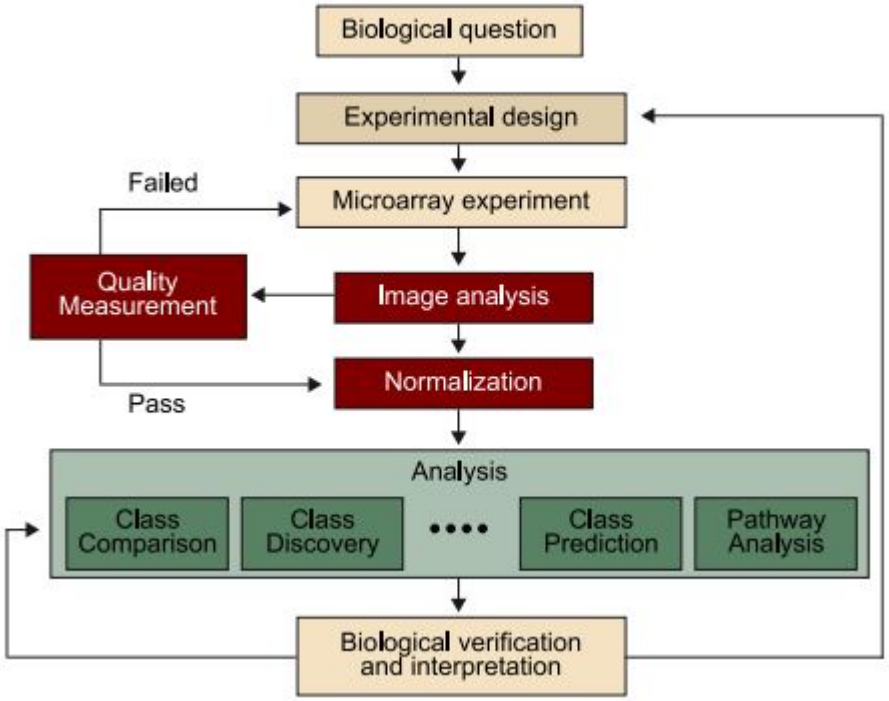

Figure 1: Imagen_Analisis_Microarrays

Para realizar un análisis completo de Microarrays debemos seguir los siguientes pasos:

1. **Lectura de los Datos**
2. **Preprocesado**:
   - Exploración y visualización
   - Control de Calidad
   - Normalización
   - Filtraje
3. **Selección de genes expresados diferencialmente**:
   - Análisis de Modelos Lineales: matriz de Diseño y de Contrastes
   - Selección de genes expresados diferencialmente
4. **Post-Procesado**:
   - Comparaciones
   - Anotaciones
   - Análisis de significación biólogica

Para realizar el **preprocesado** de los datos crudos obtenidos de los archivos CEL hemos empleado el método de RMA mediante el cual se efecturán los siguientes pasos:

–> Corrección de fondo –> Normalización para hacer los valores de los arrays comparables –> Resumen de las diversas sondas asociadas a cada grupo de sondas para dar un único valor.

Una vez obtenidos los datos normalizados, procedemos a la realización de un **filtraje** a través del cual eliminaremos aquellos genes que apenas varían entre condiciones o que deseamos quitar.

Para efectuar la **selección de genes diferencialmente expresados** debemos realizar una serie de pruebas, generalmente en términos de genes, para comparar la expresión de genes entre grupos. En nuestro caso, se aplicará la aproximación presentada por Smyth basada en la utilización del modelo lineal general, es decir, se trata de ajustar un modelo lineal a cada gen para detectar diferencias de expresión.

A fin de controlar el porcentaje de falsos positivos que puedan resultar del alto número de contrastes realizados simultáneamente, los p-valores se ajustan de forma que tengamos control sobre la tasa de falsos positivos utilizando el método de Benjamini y Hochberg.

Una vez identificados los genes expresados diferencialmente, trataremos de agrupar los mismos para localizar posibles patrones comunes entre las condiones experimentales. Para ello, nos apoyaremos, como se hará a lo largo de todo el análisis, de gráficas que nos ayuden a visualizar dichas agrupaciones: diagrama de Venn, mapas de calor, dendogramas...

En el proceso de **"anotación"** buscaremos información para asociar los identificadores de los genes, generalmente correspondientes a sondas o transcripciones que dependen del tipo de matriz, con nombres más familiares como el Símbolo del gen, el Identificador del gen Entrez o la descripción del gen.

Para finalizar el análisis, se realizará un **Análisis de Significacion Biológica** que busca establecer si, dada una lista de genes seleccionados por ser diferencial expresada entre dos condiciones, las funciones, procesos biológicos o vías moleculares que los caracterizan, aparecen en esta lista con más frecuencia que entre el resto de los genes analizados.

2.2. Herramientas Bioinformáticas

Para la realización del análisis de microarrays hemos empleado las facilidades que ofrecen R Studio y las librerías de Bioconductor. Para el correcto funcionamiento del análisis debemos aseguramos que disponermos de las versiones compatibles en ambos casos. En el caso que nos compete, la versión de R Studio es la 3.6.2 y la versión de Bioconductor corresponde a la 3.10.

2.2.1. Instalación de Paquetes en R

Para poder realizar el análisis, debemos asegurarnos que disponemos de todos los paquetes necesarios que descargaremos del repositorio CRAN en el caso de paquetes estándar o Bioconductor para paquetes Bioconductor. La instalación se llevará a cabo por medio del siguiente código:

```
> #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("pd.mogene.2.1.st")
> #BiocManager::install("arrayQualityMetrics")
> #BiocManager::install("pvca")
> # El siguiente código se empleará una vez se haya realizado el análisis
> #BiocManager::install("limma")
> #BiocManager::install("genefilter")
> #BiocManager::install("mogene21sttranscriptcluster.db")
> #BiocManager::install("annotate")
```

```
> #BiocManager::install("org.Mm.eg.db")
> #BiocManager::install("ReactomePA")
> #BiocManager::install("reactome.db")
```

3. Datos

La obtención de los datos se localiza en la base de datos Omnibus de expresión genética (GEO). El conjunto de datos seleccionado se identifica con el número de acceso: **GSE67883** y se puede localizar toda la información del mismo mediante la siguiente url:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67883.

Además, podemos acceder al artículo completo de Dehmel titulado *"Intrauterine smoke exposure deregulates lung function, pulmonary transcriptomes, and in particular insulin-like growth factor (IGF)-1 in a sex-specific manner"* a través del siguiente link:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5953988/

Los microarrays utilizados para este experimento fueron del tipo Mouse Gene 2.1 ST Array de Affymetrix. Se realizaron análisis de microarrays de expresión de ARNm en tejido pulmonar fetal (E18.5) derivado de madres expuestas al humo o de madres no tratadas/control.

3.1. Directorios y opciones de trabajo

Para facilitar nuestro estudio trabajaremos en el directorio ("C:/Users/Natalia/Master/Datos_omicos_2/PEC1") que asignaremos a la variable workingDir. Copiaremos los datos en un subdirectorio del anterior denominado data, que se almacenará en la variable dataDir y los resultados los guardaremos en un directorio "results" que asignaremos a la variable resultsDir.

```
> setwd("C:/Users/Natalia/Master/Datos_omicos_2/PEC1")
> workingDir <-getwd()
> dir.create("data")
> dir.create("results")
> dataDir <-file.path(workingDir, "data")
> resultsDir <- file.path(workingDir, "results")
> setwd(workingDir)
> options(width=80)
> options(digits=5)
```

3.2. Preparación de los datos para el Análisis

Para poder comenzar con el análisis de nuestros datos se requieren dos tipos de archivos:

- .CEL que obtendremos de la página donde hemos seleccionado el estudio. En nuestro caso el experimento elegido contiene 21 muestras .CEL comprimidas que podemos descargar en la base de datos de **GEO**
- Creación de un archivo csv denominado "targets" que contiene la información sobre los grupos y las covariables. La creación de este archivo se realiza mediante la información obtenida del estudio, ayudándonos de los datos analizados en **GEO2R**. Mi archivo targets ha quedado como sigue:

```
> targets <- read.csv2("./data/targets.csv", header = TRUE, sep = ";")
> knitr::kable(
+   targets, booktabs = TRUE,
+   caption = 'Content of the targets file used for the current analysis')
```

Table 1: Content of the targets file used for the current analysis

FileName	ShortName	Titulo	Grupo	Sexo
GSM1657523_Smoke_M121.CEL	M121_Smoke	Smoke_fetal_lung_rep1 [mRNA]	Smoke_Male	Macho
GSM1657524_Smoke_M191.CEL	M191_Smoke	Smoke_fetal_lung_rep2 [mRNA]	Smoke_Male	Macho
GSM1657525_Smoke_M205.CEL	M205_Smoke	Smoke_fetal_lung_rep3 [mRNA]	Smoke_Male	Macho
GSM1657526_Smoke_M207.CEL	M207_Smoke	Smoke_fetal_lung_rep4 [mRNA]	Smoke_Male	Macho
GSM1657527_Smoke_M443.CEL	M443_Smoke	Smoke_fetal_lung_rep5 [mRNA]	Smoke_Male	Macho
GSM1657528_Smoke_M447.CEL	M447_Smoke	Smoke_fetal_lung_rep6 [mRNA]	Smoke_Male	Macho
GSM1657529_Smoke_F195.CEL	F195_Smoke	Smoke_fetal_lung_rep7 [mRNA]	Smoke_Female	Hembra
GSM1657530_Smoke_F197.CEL	F197_Smoke	Smoke_fetal_lung_rep8 [mRNA]	Smoke_Female	Hembra
GSM1657531_Smoke_F201.CEL	F201_Smoke	Smoke_fetal_lung_rep9 [mRNA]	Smoke_Female	Hembra
GSM1657532_Smoke_F202.CEL	F202_Smoke	Smoke_fetal_lung_rep10 [mRNA]	Smoke_Female	Hembra
GSM1657533_Smoke_F444.CEL	F444_Smoke	Smoke_fetal_lung_rep11 [mRNA]	Smoke_Female	Hembra
GSM1657534_Smoke_F445.CEL	F445_Smoke	Smoke_fetal_lung_rep12 [mRNA]	Smoke_Female	Hembra
GSM1657535_Ctrl_M141.CEL	M141_Control	Control_fetal_lung_rep1 [mRNA]	Control_Male	Macho
GSM1657536_Ctrl_M187.CEL	M187_Control	Control_fetal_lung_rep2 [mRNA]	Control_Male	Macho
GSM1657537_Ctrl_M307.CEL	M307_Control	Control_fetal_lung_rep3 [mRNA]	Control_Male	Macho
GSM1657538_Ctrl_M386.CEL	M386_Control	Control_fetal_lung_rep4 [mRNA]	Control_Male	Macho
GSM1657539_Ctrl_M421.CEL	M421_Control	Control_fetal_lung_rep5 [mRNA]	Control_Male	Macho
GSM1657540_Ctrl_F181.CEL	F181_Control	Control_fetal_lung_rep6 [mRNA]	Control_Female	Hembra
GSM1657541_Ctrl_F301.CEL	F301_Control	Control_fetal_lung_rep7 [mRNA]	Control_Female	Hembra
GSM1657542_Ctrl_F381.CEL	F381_Control	Control_fetal_lung_rep8 [mRNA]	Control_Female	Hembra
GSM1657543_Ctrl_F427.CEL	F427_Control	Control_fetal_lung_rep9 [mRNA]	Control_Female	Hembra

Tanto los archivos CEL como targets deben encontrarse en el mismo directorio de trabajo, es decir en el subdirectorio data que hemos comentado en el apartado de directorios.

3.3. Lectura Archivos CEL

Para poder leer los archivos .CEL se requiere el uso de la libería oligo de Bioconductor y almacenaremos los datos CEL en una variable que llamaremos rawData:

```
> library(oligo)
> celFiles <- list.celfiles("./data", full.names = TRUE)
> library(Biobase)
> my.targets <-read.AnnotatedDataFrame(file.path("./data","targets.csv"),
+                                      header = TRUE, row.names = 1,
+                                      sep=";")
> rawData <- read.celfiles(celFiles, phenoData = my.targets)
```

Tras la lectura, procederemos a asociar la información almacenada en los archivos CEL con el archivo targets:

```
> my.targets@data$ShortName->rownames(pData(rawData))
> colnames(rawData) <-rownames(pData(rawData))
>
> head(rawData)
```

```
GeneFeatureSet (storageMode: lockedEnvironment)
assayData: 1 features, 21 samples
  element names: exprs
protocolData
  rowNames: M121_Smoke M191_Smoke ... F427_Control (21 total)
  varLabels: exprs dates
```

```
  varMetadata: labelDescription channel
phenoData
  rowNames: M121_Smoke M191_Smoke ... F427_Control (21 total)
  varLabels: ShortName Titulo Grupo Sexo
  varMetadata: labelDescription channel
featureData: none
experimentData: use 'experimentData(object)'
Annotation: pd.mogene.2.1.st
```

```
> print(pData(rawData))
```

```
                ShortName                          Titulo          Grupo   Sexo
M121_Smoke     M121_Smoke   Smoke_fetal_lung_rep1 [mRNA]     Smoke_Male  Macho
M191_Smoke     M191_Smoke   Smoke_fetal_lung_rep2 [mRNA]     Smoke_Male  Macho
M205_Smoke     M205_Smoke   Smoke_fetal_lung_rep3 [mRNA]     Smoke_Male  Macho
M207_Smoke     M207_Smoke   Smoke_fetal_lung_rep4 [mRNA]     Smoke_Male  Macho
M443_Smoke     M443_Smoke   Smoke_fetal_lung_rep5 [mRNA]     Smoke_Male  Macho
M447_Smoke     M447_Smoke   Smoke_fetal_lung_rep6 [mRNA]     Smoke_Male  Macho
F195_Smoke     F195_Smoke   Smoke_fetal_lung_rep7 [mRNA]   Smoke_Female Hembra
F197_Smoke     F197_Smoke   Smoke_fetal_lung_rep8 [mRNA]   Smoke_Female Hembra
F201_Smoke     F201_Smoke   Smoke_fetal_lung_rep9 [mRNA]   Smoke_Female Hembra
F202_Smoke     F202_Smoke  Smoke_fetal_lung_rep10 [mRNA]   Smoke_Female Hembra
F444_Smoke     F444_Smoke  Smoke_fetal_lung_rep11 [mRNA]   Smoke_Female Hembra
F445_Smoke     F445_Smoke  Smoke_fetal_lung_rep12 [mRNA]   Smoke_Female Hembra
M141_Control M141_Control Control_fetal_lung_rep1 [mRNA]   Control_Male  Macho
M187_Control M187_Control Control_fetal_lung_rep2 [mRNA]   Control_Male  Macho
M307_Control M307_Control Control_fetal_lung_rep3 [mRNA]   Control_Male  Macho
M386_Control M386_Control Control_fetal_lung_rep4 [mRNA]   Control_Male  Macho
M421_Control M421_Control Control_fetal_lung_rep5 [mRNA]   Control_Male  Macho
F181_Control F181_Control Control_fetal_lung_rep6 [mRNA] Control_Female Hembra
F301_Control F301_Control Control_fetal_lung_rep7 [mRNA] Control_Female Hembra
F381_Control F381_Control Control_fetal_lung_rep8 [mRNA] Control_Female Hembra
F427_Control F427_Control Control_fetal_lung_rep9 [mRNA] Control_Female Hembra
```

4. Preprocesado: Exploración, Control de Calidad y Normalización

El preprocesado de los datos comprende diferentes fases:

- Realización de algunos gráficos con los datos en crudo para hacerse una idea del experimento.
- Realizar un control de calidad
- Normalizar y resumir las expresiones
- Filtrado para eliminar aquellos genes que hemos detectado en las fases anteriores que no se expresan o se expresan de diferente manera que el resto de los grupos.

4.1. Exploración y Visualización

Para llevar a cabo una primera exploración de la distribución de las señales de los datos originales sin normalizar emplearemos un **histograma**. Además, nos permitirá hacernos una idea de si la distribución de los arrays son similares en forma y posición.

```
> col = c(rep("red", 3), rep("blue", 3), rep("green", 3), rep("yellow", 3))
> grupos <- pData(rawData)$Grupo
> Sexo <- pData(rawData)$Sexo
> numMuestras <- nrow(pData(rawData))
> NombreMuestras <- paste( pData(rawData)$ShortName, Sexo, sep=".")
```

```
> hist(rawData, cex.axis=0.5, las=2, which="all",
+ col = c(rep("red", 3), rep("blue", 3), rep("green", 3),
+ rep("yellow", 3)),
+ main="Distribucion Arrays")
> legend (x="topright", legend=NombreMuestras , col=col, lty=1:numMuestras)
```

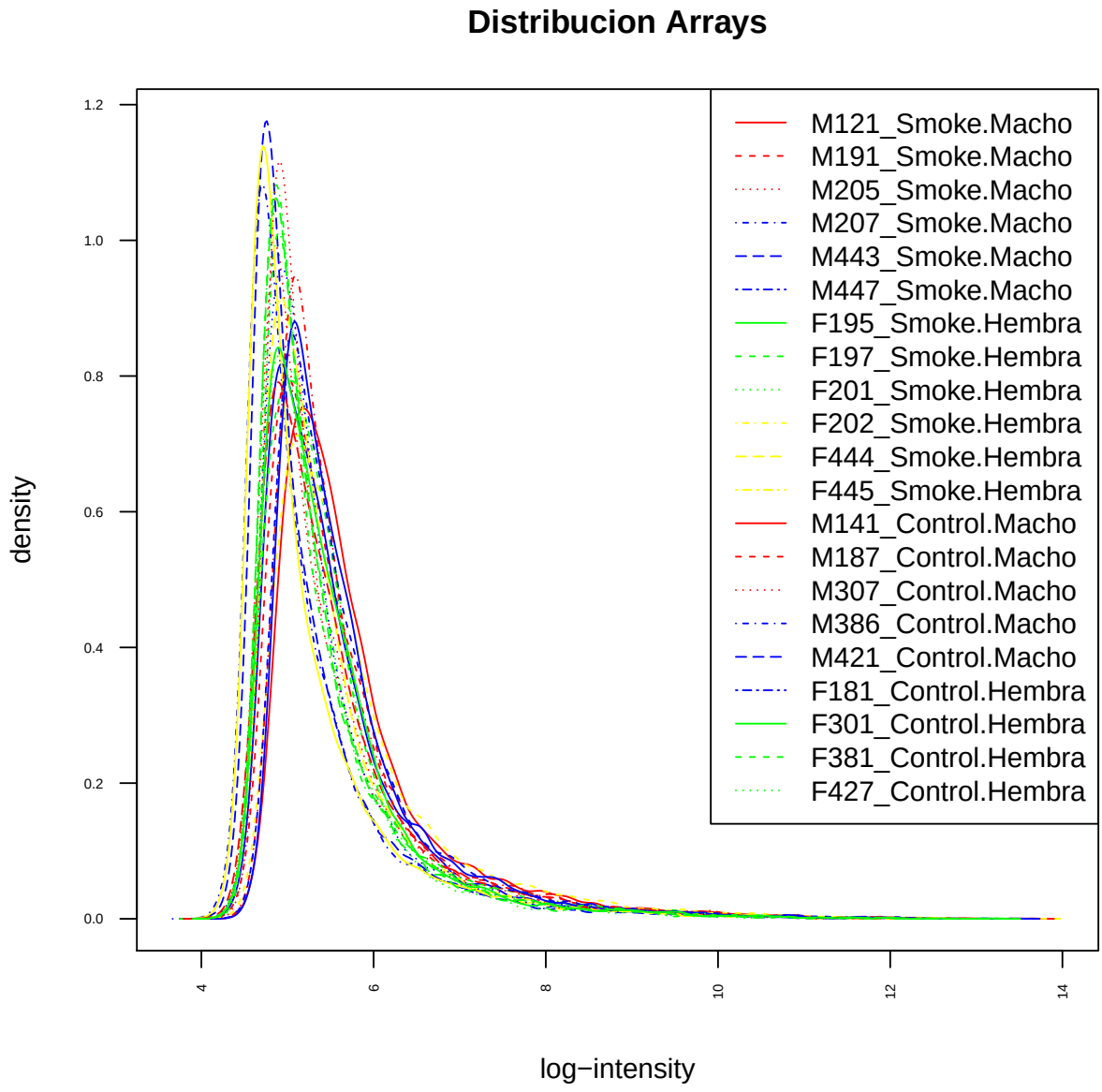

Figure 2: Histograma para para conocer la distribución de los arrays (Raw Data)

Atendiendo al histograma podemos decir que la distribución de los arrays es bastante similar en forma y posición y, por lo tanto esto nos sugiere que no parece existir problema con los datos.

4.2. Control de Calidad de los datos

Mediante el paquete ArrayQualityMetrics se realizan diferentes enfoques de calidad, como diagrama de caja de la intensidad de los datos y Análisis de componentes principales (PCA)...

```
> library(arrayQualityMetrics)
> arrayQualityMetrics(rawData, outdir = "./results/rawData_quality", force = T)
```

En la carpeta rawData_quality se generan una serie de archivos, entre los cuales existe uno denominado index que abre una página web desde donde podremos acceder a un resumen del análisis realizado.

Otra forma de poder visualizar los datos es a través de los **diagramas de caja** basados en los distintos cuantiles de los valores– que nos pueden dar una idea de la distribución de las intensidades.

```
> boxplot(rawData, cex.axis=0.5, las=2,  which="all",
+          col = c(rep("red", 3), rep("blue", 3), rep("green", 3), rep("yellow", 3)),
+          main="Distribution of raw intensity values")
```

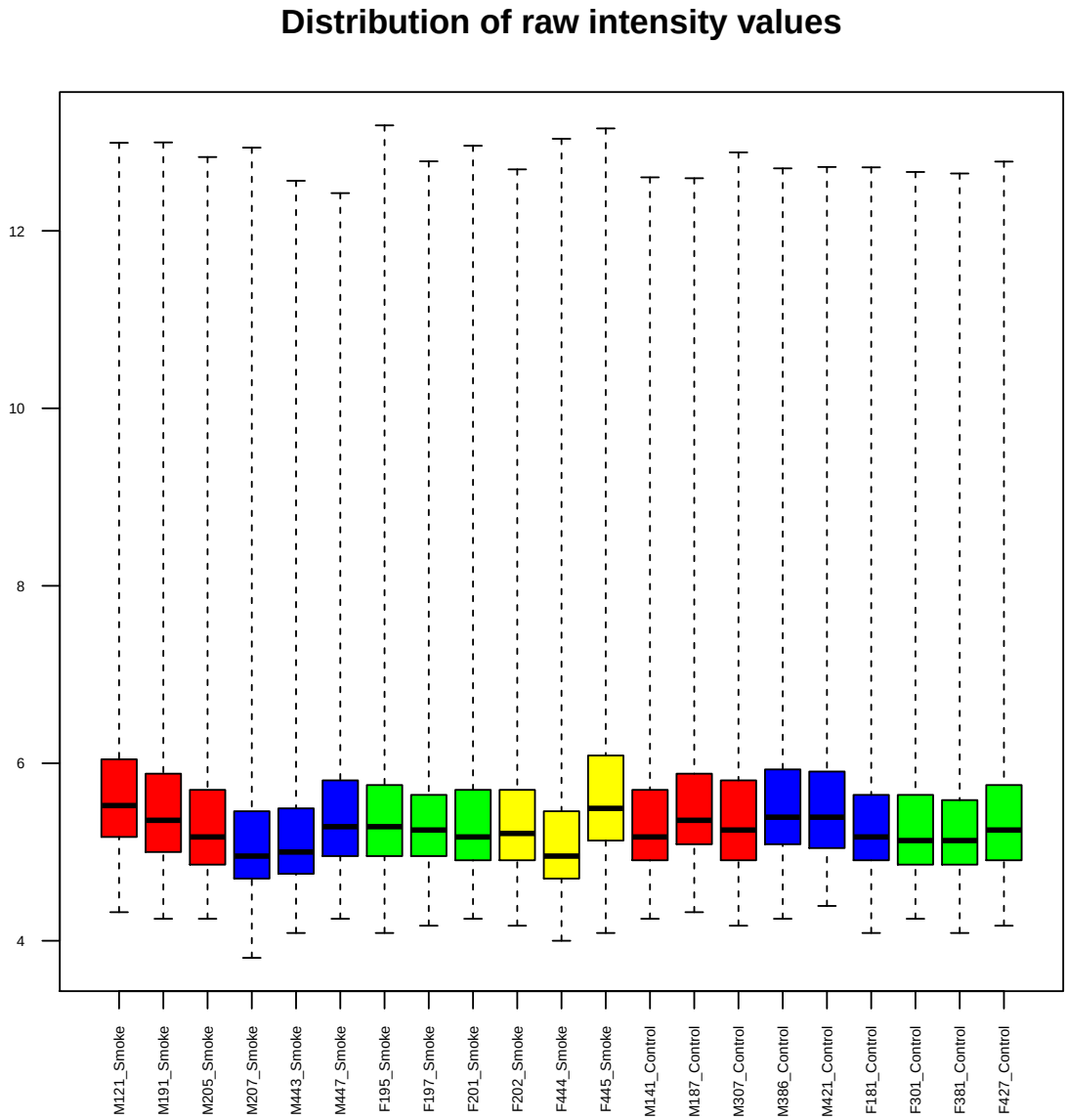

Figure 3: Boxplot para intensidades de de los arrays (Raw Data)

Las distribuciones de los datos son bastante parecidas por lo que sugiere que es conveniente normalizar pero no se aprecian arrays problemáticos, salvo el F445 que parece salirse de la distribución del resto de arrays.

También podemos realizar un análisis visual y detallado empleando las funciones ggplot que reforzarán lo conseguido con los análisis anteriores.

```
> library(ggplot2)
> library(ggrepel)
> plotPCA3 <- function (datos, labels, factor, title, scale,colores, size = 1.5, glineas = 0.25) {
+   data <- prcomp(t(datos),scale=scale)
+   # ajustes de la gráfica
```

```
+   dataDf <- data.frame(data$x)
+   Group <- factor
+   loads <- round(data$sdev^2/sum(data$sdev^2)*100,1)
+   # gráfica principal
+   p1 <- ggplot(dataDf,aes(x=PC1, y=PC2)) +
+     theme_classic() +
+     geom_hline(yintercept = 0, color = "gray70") +
+     geom_vline(xintercept = 0, color = "gray70") +
+     geom_point(aes(color = Group), alpha = 0.55, size = 3) +
+     coord_cartesian(xlim = c(min(data$x[,1])-5,max(data$x[,1])+5)) +
+     scale_fill_discrete(name = "Grupo")
+   # evitar superposición etiquetas
+   p1 + geom_text_repel(aes(y = PC2 + 0.25, label = labels),segment.size = 0.25, size = size) +
+     labs(x = c(paste("PC1",loads[1],"%")),y=c(paste("PC2",loads[2],"%"))) +
+     ggtitle(paste("Análisis de Componentes Principales: ",title,sep=" "))+
+     theme(plot.title = element_text(hjust = 0.5)) +
+     scale_color_manual(values=colores)
+   }
```

El **gráfico de componentes principales** nos muestra la varianza en los datos representando el porcentaje explicado para cada componente, en los cuales tenemos una medida de la importancia de los grupos que se puedan visualizar. Si la suma de los porecentajes es alta, por ejemplo superior al 50% las conclusiones obtenidas seran mas fiables que con valores bajos, por ejemplo inferiores al 30% de varianza explicada.

```
> plotPCA3(exprs(rawData), labels = targets$ShortName, factor = targets$Grupo,
+          title="Raw data", scale = FALSE, size = 3,
+          colores = c("red", "blue", "green", "yellow"))
```

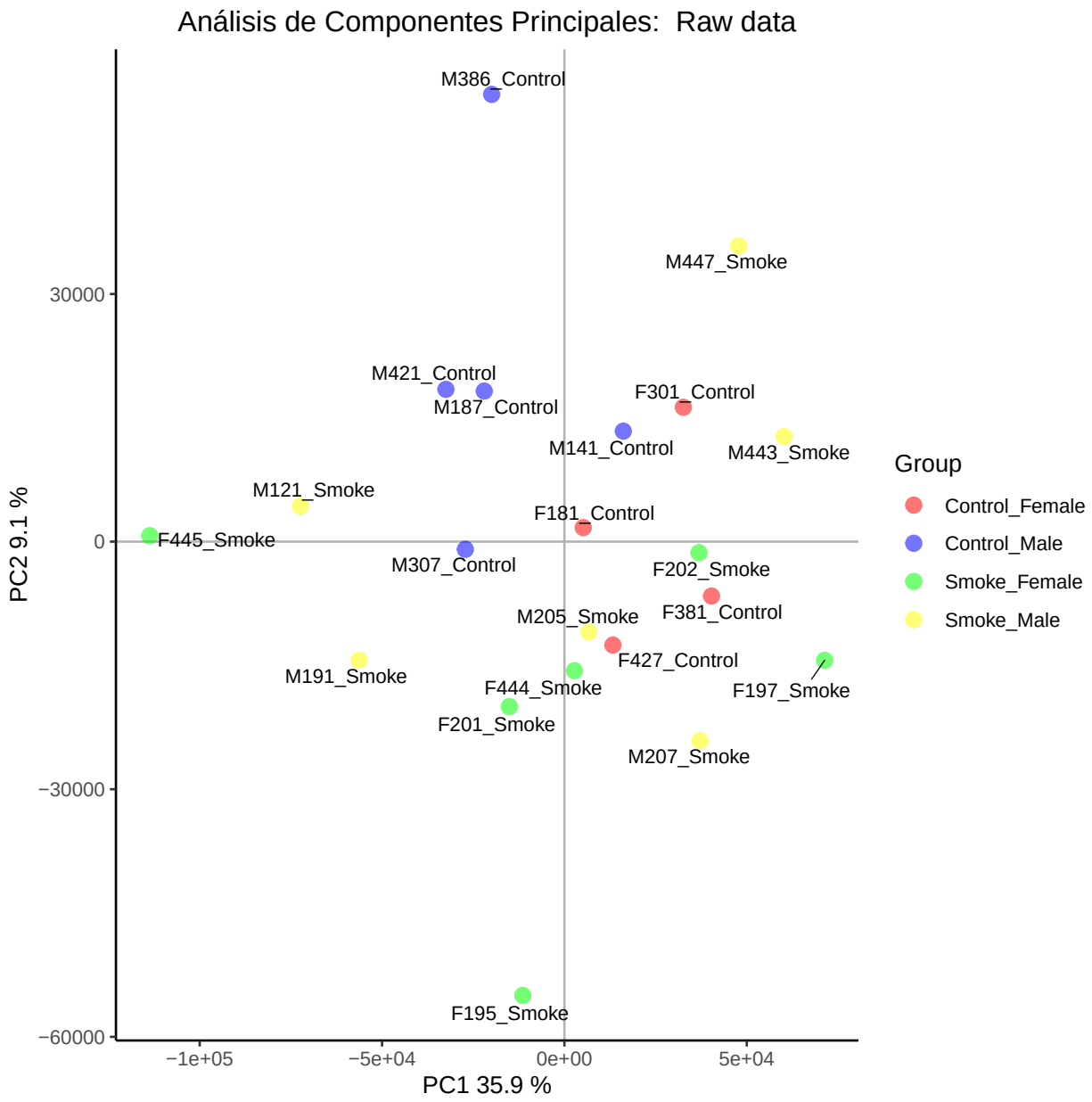

Figure 4: Visualizacion de los dos primeros componentes de raw data

Tal y como podemos observar en la gráfica, el primer componente de la PCA representa el 35.9% de la variabilidad total de las muestras y el segundo componente un 29,1%. El porcentaje de variabilidad explicado por las dos primeras componentes nos indica que no son muy explicativas.

Además, este gráfico servir para detectar si las muestras se agrupan de forma "natural", es decir, con otras muestras provenientes del mismo grupo o si no hay correspondencia clara entre grupos experimentales En nuestro caso de estudio se observa que los datos aparecen separados.

Para completar los resultados obtenidos mediante la gráfica de Componentes principales podemos emplear un cluster jerárquico seguido de un **dendograma** que nos puede ayudar a hacernos una idea de si las muestras se agrupan por condiciones experimentales

```
> Distancia <- dist(t(exprs(rawData)))
> heatmap (as.matrix(Distancia),  col=heat.colors(16))
```

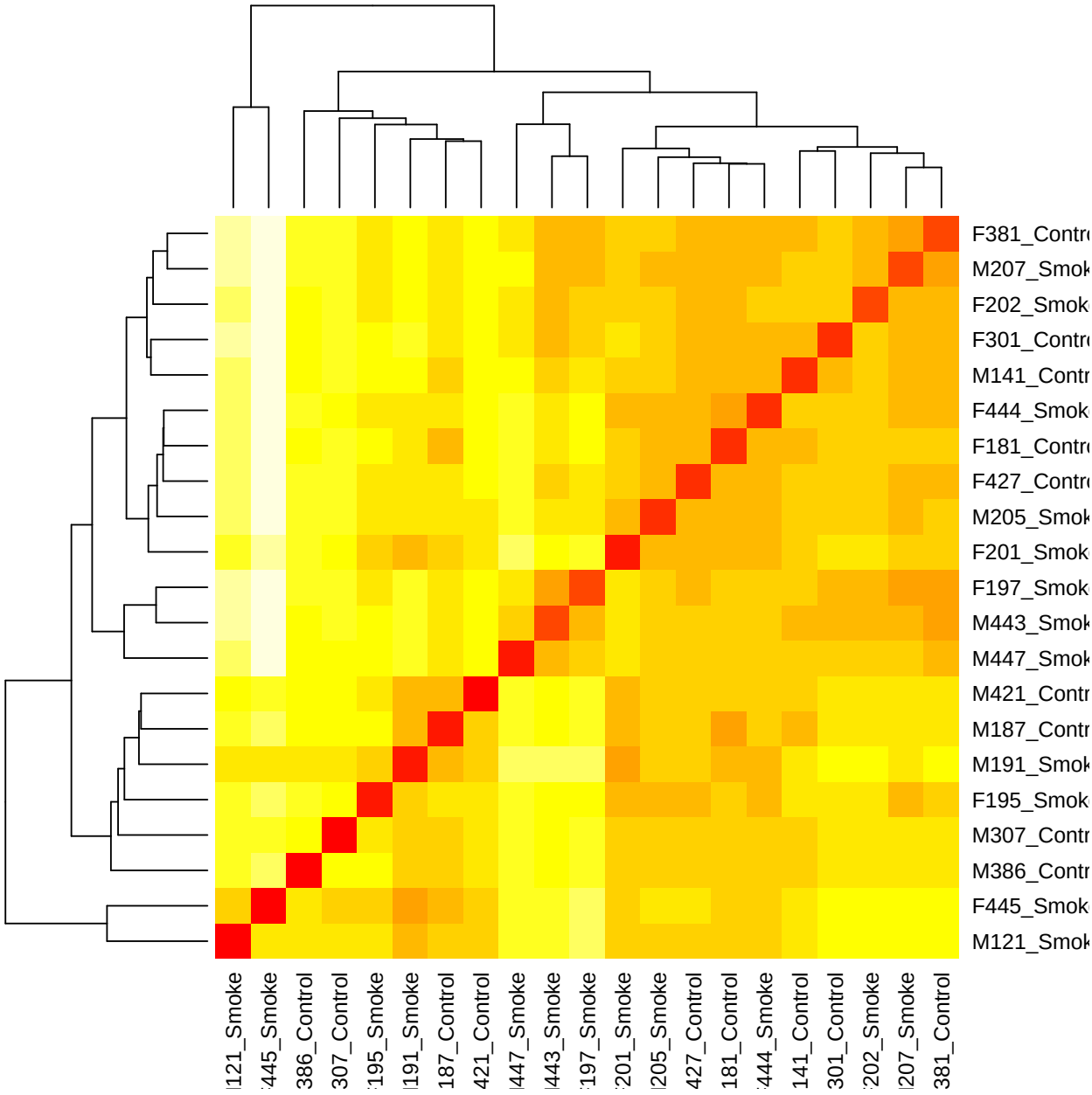

Figure 5: Mapa de Calor para visualizar la agrupación por condiciones experimentales de raw data

```
> Distancia <- dist(t(exprs(rawData)))
> den <- hclust (Distancia, "average")
> plot (den)
```

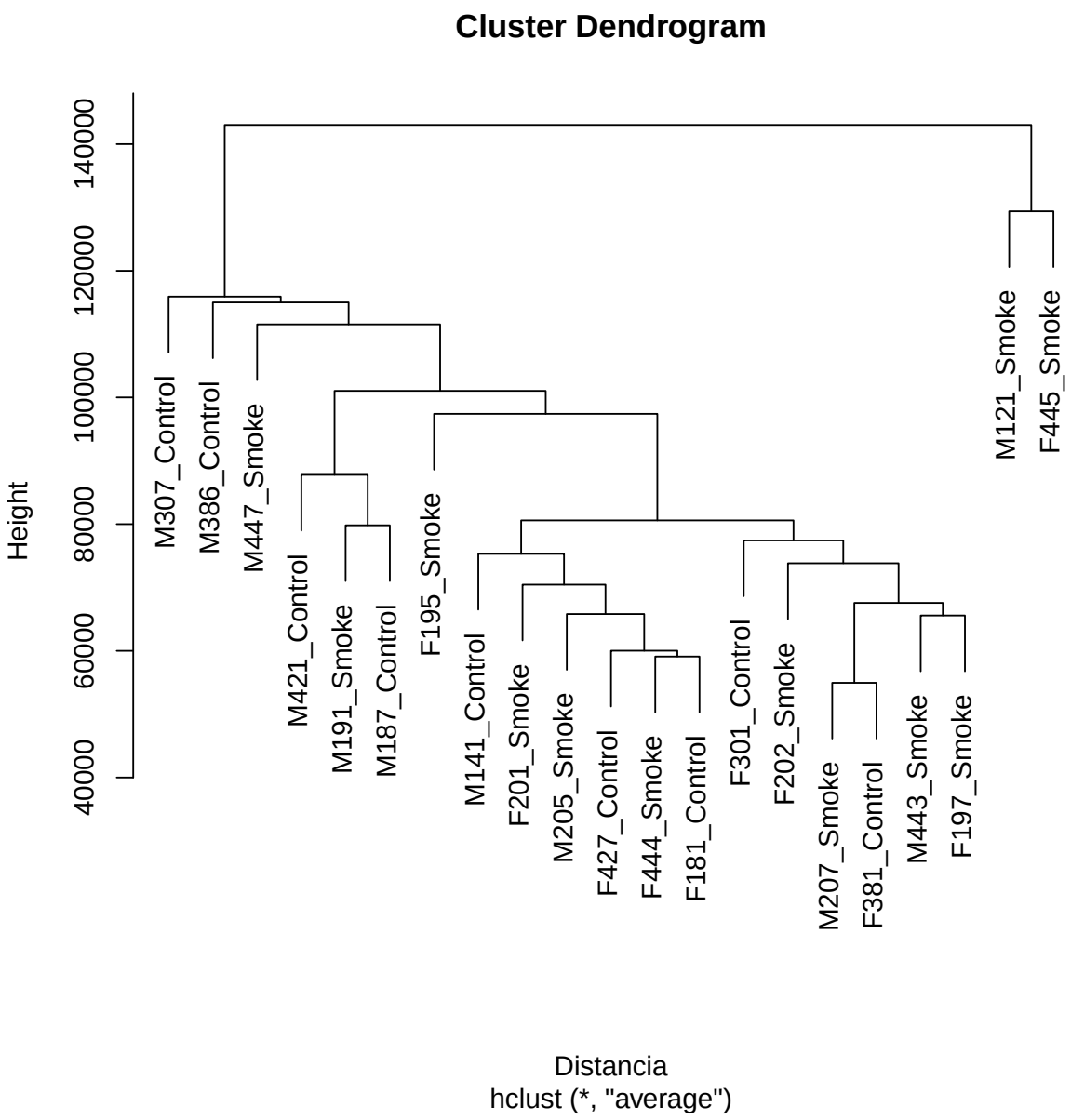

Figure 6: Dendograma para visualizar la agrupación por condiciones experimentales de raw data

En el dendograma se detectan dos muestras de un macho y una hembra (M121 Y F445) que son susceptibles a no cumplir el control de calidad.

4.3. Normalización del Dato

Con el proceso de normalización vamos intentar que las matrices sean comparables entre ellas y tratar de reducir o eliminar la variabilidad en las muestras que no se deba a razones biológicas. Realizare el procesado mediante RMA mediante el cual se llevarán a cabo los siguientes pasos:

- Corrección de fondo
- Normalización para hacer los valores de los arrays comparables

- Resumen de las diversas sondas asociadas a cada grupo de sondas para dar un único valor.

```
> eset_rma <- rma(rawData)
```

```
Background correcting
Normalizing
Calculating Expression
```

4.4. Control de Calidad de los Datos Normalizados

Realizaremos de nuevo un control de Calidad de estos Datos Normalizados tal y como lo hemos realizado anteriormente. Volveremos a generar un diagrama de cajas para estudiar la distribución de las intensidades tras la normalización.

```
> 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 de intensidades de los Arrays: Datos Normalizados")
```

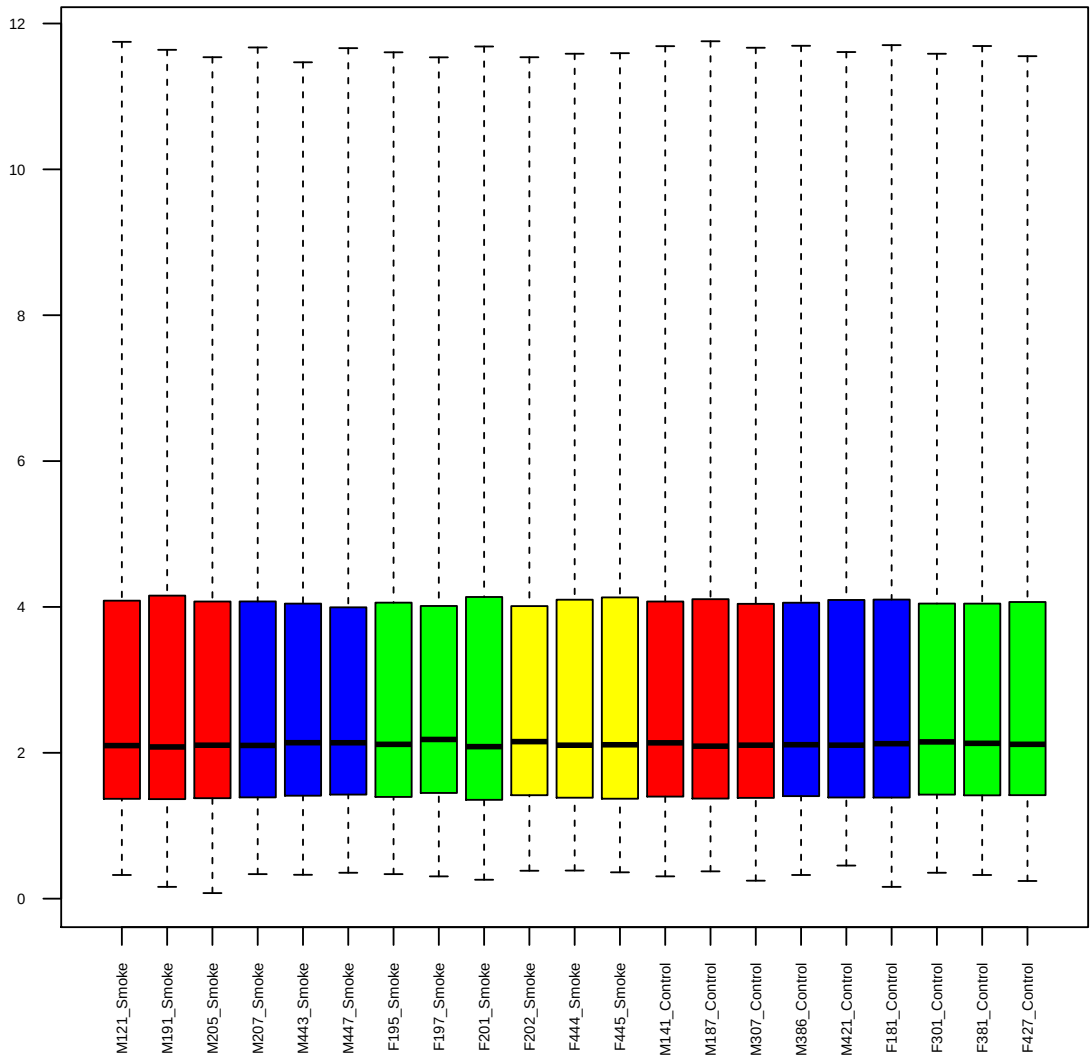

Figure 7: Distribución de las intensidades de los datos normalizados

Después de normalizar se observa como los valores se encuentran en una escala comparable y se descarta la ausencia de Microarrays problemáticos.

Ejecutamos de nuevo el paquete *ArrayQualityMetrics* y guardaremos el resultado en una nueva carpeta QCDir.Norm donde podremos acceder a un nuevo archivo index.htlm que contiene el resumen del análisis realizado.

```
> arrayQualityMetrics(eset_rma, outdir = file.path("./results", "QCDir.Norm"), force=TRUE)
```

Igual que en el apartado anterior, generamos un gráfico de componentes principales.

```
> plotPCA3(exprs(eset_rma), labels = targets$ShortName, factor = targets$Grupo,
+          title="Datos Normalizados", scale = FALSE, size = 3,
+          colores = c("red", "blue", "green", "yellow"))
```

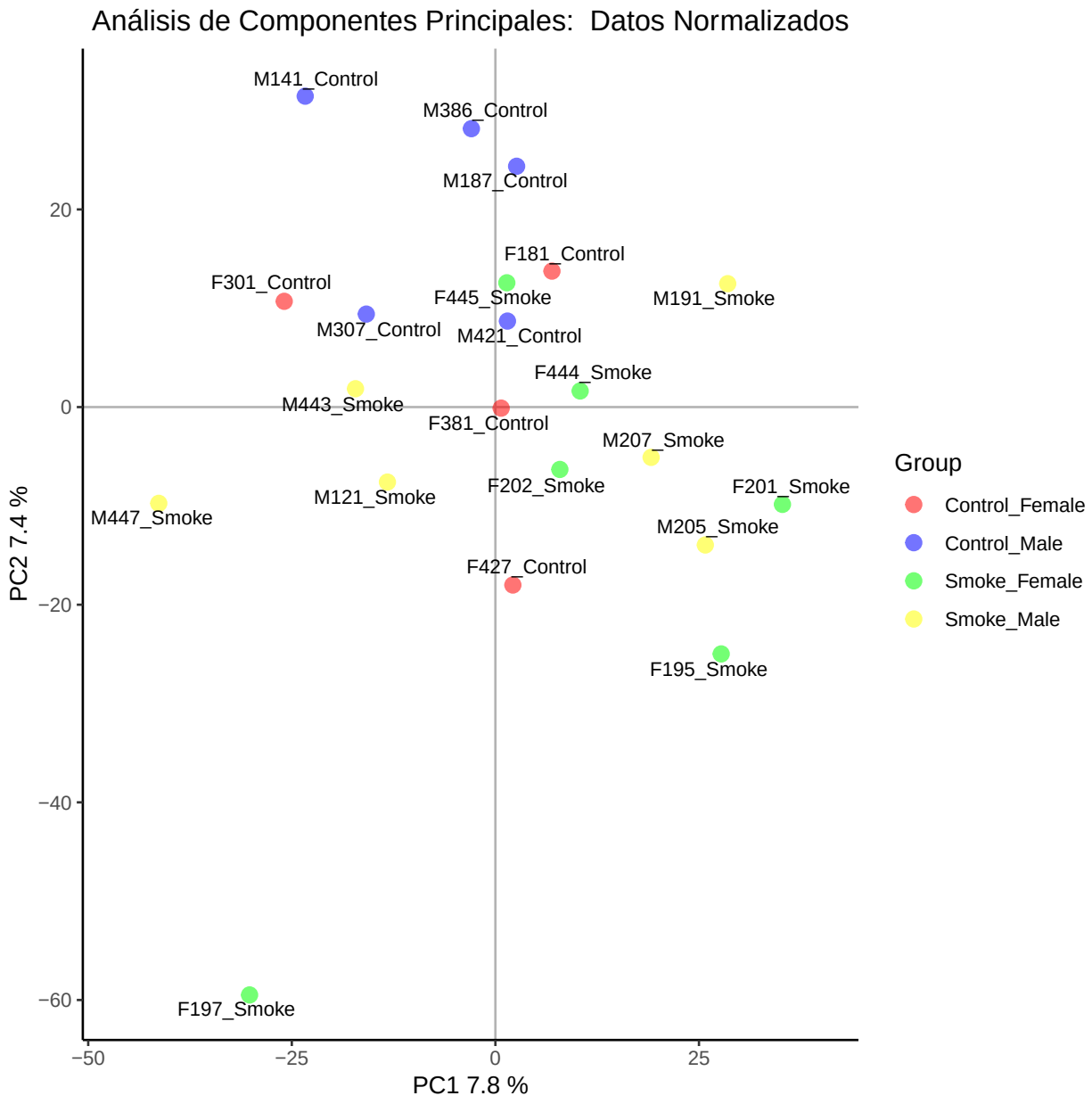

Figure 8: Visualización de los dos componentes principales para Datos Normalizados

Tras la normalización, el primer componente de la PCA representa el 7,8% de la variabilidad total de las muestras y el segundo componente un 7,4%. Se ha producido una disminución en el porcentaje de la variabilidad explicada con respecto a la PCA realizada en datos sin procesar.

4.5. Detectando una mayor variabilidad de genes

La selección de genes expresados diferencialmente se ve afectada por la cantidad de genes sobre los que hacemos el estudio. Cuanto mayor sea el número, mayor será el ajuste necesario del p-valor, lo que nos llevará a eliminar más genes.

Si un gen se expresa de manera diferencial, se espera que haya una cierta diferencia entre los grupos y, por lo tanto, la varianza general del gen será mayor que la de aquellos que no tienen expresión diferencial.

Trazar la variabilidad general de todos los genes, en nuestro caso de los *41.345 genes*, es útil para decidir qué porcentaje de genes muestra una variabilidad que puede atribuirse a otras causas que no sean la variación aleatoria.

```
> numero_genes <- nrow(eset_rma)
> numero_genes
```

```
Features
   41345
```

```
> sds <- apply (exprs(eset_rma), 1, sd)
> sds0<- sort(sds)
> plot(1:length(sds0), sds0, main="Distribución de variabilidad para todos los genes",
+      sub="La lineas verticales representan los percentiles 90% y 95%",
+      xlab="Índice de genes (de menor a mayor variabilidad)", ylab="Desviación Estándar")
> abline(v=length(sds)*c(0.9,0.95))
```

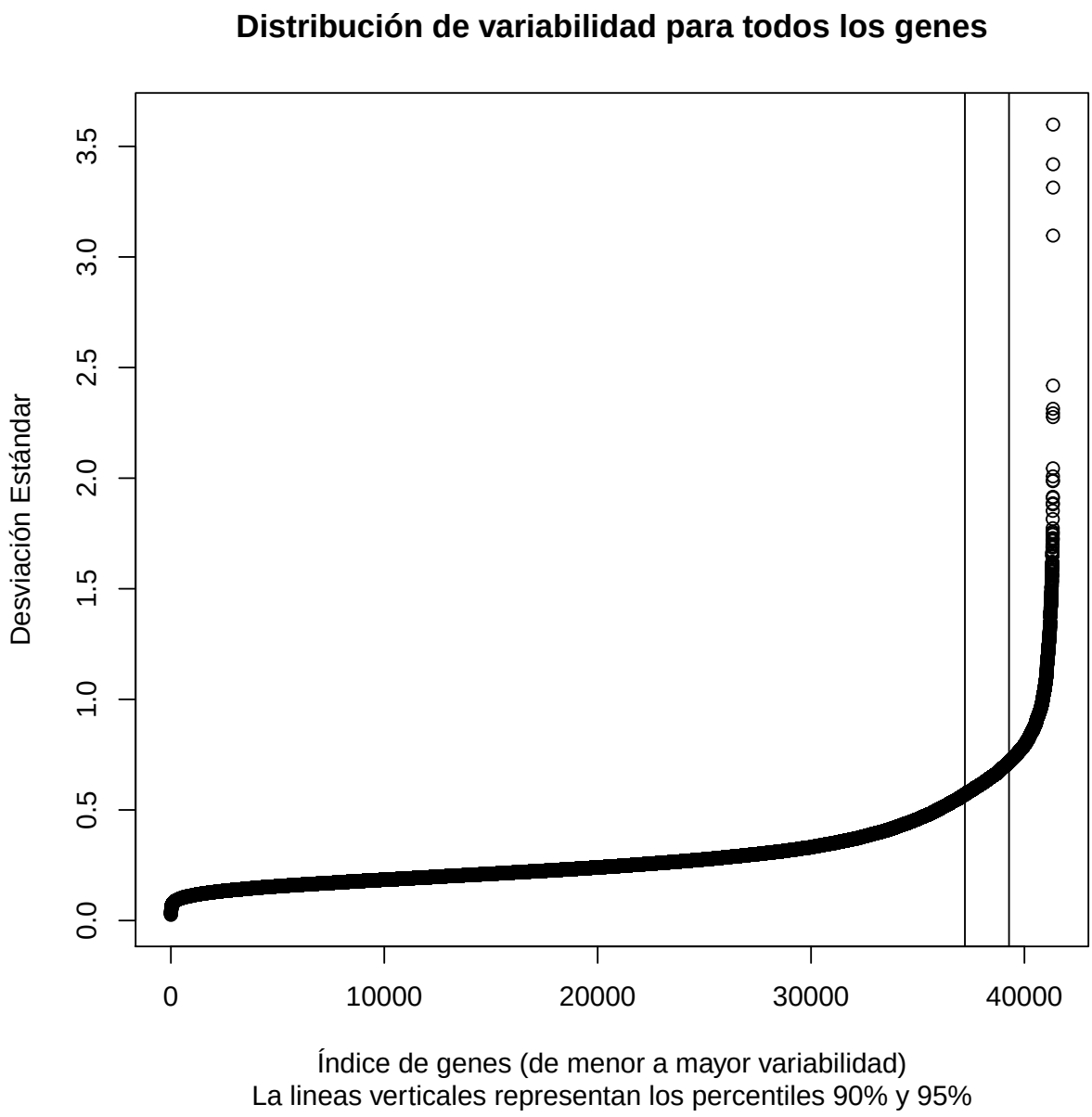

Figure 9: Valores de las desviaciones estándar que abarcan todas las muestras de los genes ordenados de menor a mayor

El gráfico representa las desviaciones estándar de todos los genes ordenados de menor a mayor valor, donde se muestra que los genes más variables son aquellos con una desviación estándar superior al 90-95% de todas las desviaciones estándar.

4.5. Filtraje

Mediante el filtraje eliminaremos los genes que apenas varían entre condiciones o que deseamos quitar. Para ello, emplearemos el paquete nsFilter localizando el paquete de **anotaciones** correspondiente.

En el caso del estudio que nos compete corresponde a Affymetrix mogene21 annotation data (chip mo-

gene21sttranscriptcluster).

```
> library(genefilter)
> library(mogene21sttranscriptcluster.db)
> annotation(eset_rma) <- "mogene21sttranscriptcluster.db"
> filtered <- nsFilter(eset_rma,
+                      require.entrez = TRUE, remove.dupEntrez = TRUE,
+                      var.filter=TRUE, var.func=IQR, var.cutoff=0.75,
+                      filterByQuantile=TRUE, feature.exclude = "^AFFX")
```

Podemos obtener los valores filtrados y un informe de los resultados del filtrado de la siguiente manera:

```
> print(filtered$filter.log)
```

```
$numDupsRemoved
[1] 671

$numLowVar
[1] 17973

$numRemoved.ENTREZID
[1] 16710
```

```
> eset_filtered <-filtered$eset
```

```
> num <- nrow(eset_filtered)
> num
```

```
Features
    5991
```

Después del filtraje nos quedan *5.991 genes* disponibles para analizar.

4.6. Guardado de los Datos Normalizados y Filtrados:

Guardamos los datos filtrados normalizados en un archivo excel por si requerimos usarlos posteriormente.

```
> 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. Selección de Genes Diferencialmente Expresados

La selección de genes expresados diferencialmente consiste en hacer algún tipo de prueba, generalmente en términos de genes, para comparar la expresión de genes entre grupos.

En este ejemplo se aplicará la aproximación presentada por Smyth basada en la utilización del modelo lineal general, combinada con un método para obtener una estimación mejorada de la varianza.

5.1. Análisis Basado en modelos lineales

5.1.1. Matriz de Diseño

El primer paso para el análisis basado en modelos lineales es crear la **matriz de diseño**. Básicamente es una tabla que describe la asignación de cada muestra a un grupo o condición experimental. Tiene tantas filas como muestras y tantas columnas como grupos (si solo se considera un factor). Cada fila contiene un 1 en la columna del grupo al que pertenece la muestra y un 0 en los demás.

```
> if (!exists("eset_filtered")) load (file="./results/normalized.Data.Rda")
```

```
> library(limma)
> designMat<- model.matrix(~0+Grupo, pData(eset_filtered))
> colnames(designMat) <- c("Control_Female", "Control_Male", "Smoke_Female", "Smoke_Male")
> print(designMat)
```

```
             Control_Female Control_Male Smoke_Female Smoke_Male
M121_Smoke                0            0            0          1
M191_Smoke                0            0            0          1
M205_Smoke                0            0            0          1
M207_Smoke                0            0            0          1
M443_Smoke                0            0            0          1
M447_Smoke                0            0            0          1
F195_Smoke                0            0            1          0
F197_Smoke                0            0            1          0
F201_Smoke                0            0            1          0
F202_Smoke                0            0            1          0
F444_Smoke                0            0            1          0
F445_Smoke                0            0            1          0
M141_Control              0            1            0          0
M187_Control              0            1            0          0
M307_Control              0            1            0          0
M386_Control              0            1            0          0
M421_Control              0            1            0          0
F181_Control              1            0            0          0
F301_Control              1            0            0          0
F381_Control              1            0            0          0
F427_Control              1            0            0          0
attr(,"assign")
[1] 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$Grupo
[1] "contr.treatment"
```

5.1.2. Matriz de Contraste

La matriz de contrastes se usa para describir las comparaciones entre grupos. Consiste en tantas columnas como comparaciones y tantas filas como grupos. Una comparación entre grupos, llamada "contraste", está representada por un "1" y un "-1" en las filas de grupos para comparar y ceros en el resto.

Nuestra **Matriz de Contraste** comparará:

- el efecto de exponer a una madre embarazada al humo de tabaco vs una madre sin dicha exposicion que tendrá descendientes machos.

  `"Control_Male" vs "Smoke_Male"`

- el efecto de exponer a una madre embarazada al humo de tabaco vs una madre sin dicha exposicion que tendrá descendientes hembras.

  `"Control_Female" vs "Smoke_Female"`

- el efecto de exponer a una madre embarazada al humo de tabaco vs una madre sin dicha exposicion independientemente del sexo de su descendencia.

  `"Control" vs "Smoke"`

```
> cont.matrix <- makeContrasts (CM_SM = Control_Male - Smoke_Male,
+                               CF_SF = Control_Female - Smoke_Female,
+                               Int = (Control_Male - Control_Female) - (Smoke_Female - Smoke_Male),
+                               levels=designMat)
> print(cont.matrix)
```

```
               Contrasts
Levels          CM_SM CF_SF Int
  Control_Female     0     1  -1
  Control_Male       1     0   1
  Smoke_Female       0    -1  -1
  Smoke_Male        -1     0   1
```

5.2. Estimación del Modelo y Seleccion de Genes

Una vez definida la matriz de diseño y los contrastes, podemos pasar a estimar el modelo, los contrastes y realizar las pruebas de significación que nos indiquen, para cada gen y cada comparación, si puede considerarse diferencialmente expresado.

El método implementado en *limma* amplía el análisis tradicional utilizando modelos de Bayes empíricos para combinar la información de toda la matriz de datos y de cada gen individual y obtener estimaciones de error mejoradas. El análisis proporciona los estadísticos de test habituales como Fold-change t-moderados o p-valores ajustados, que se utilizan para ordenar los genes de más a menos diferencialmente expresados.

A fin de controlar el porcentaje de falsos positivos que puedan resultar del alto número de contrastes realizados simultáneamente, los p-valores se ajustan de forma que tengamos control sobre la tasa de falsos positivos utilizando el método de Benjamini y Hochberg.

La función *topTable* genera para cada contraste una lista de genes ordenados de más a menos diferencialmente expresados.

```
> 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"
```

5.3. Obtención de la lista de Genes Expresados Diferencialmente

El paquete `limma` implementa la función`topTable` que contiene, para un contraste dado, una lista de genes ordenados desde el p-valor más pequeño al más grande que se puede considerar como más o menos expresado diferencialmente.

- Para la comparación 1 (Control_Male vs Smoke_Male): Genes que cambian su expresión por el efecto de exponer a una madre embarazada al humo de tabaco vs una madre sin dicha exposicion con descendientes machos.

```
> topTab_CM_SM <- topTable (fit.main, number=nrow(fit.main), coef="CM_SM", adjust="fdr")
> head(topTab_CM_SM)
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
17500996	0.71768	4.5121	4.8048	0.00005	0.30413	0.00164
17451867	-0.75676	4.3872	-4.3212	0.00019	0.55835	-0.68505

	logFC	AveExpr	t	P.Value	adj.P.Val	B
17457994	-0.67448	1.2777	-4.1033	0.00033	0.61239	-0.99986
17366728	0.76929	1.3359	3.9510	0.00050	0.61239	-1.22102
17437129	0.61978	3.5840	3.8250	0.00070	0.61239	-1.40433
17428477	0.87623	4.5068	3.8243	0.00070	0.61239	-1.40538

- Para la comparación 2 (Control_Female vs Smoke_Female): Genes que cambian su expresión por el efecto de exponer a una madre embarazada al humo de tabaco vs una madre sin dicha exposicion con descendientes hembras.

```
> topTab_CF_SF <- topTable (fit.main, number=nrow(fit.main), coef="CF_SF", adjust="fdr")
> head(topTab_CF_SF)
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
17374406	-0.75619	4.0192	-4.6650	0.00007	0.35461	-0.06199
17351053	0.70257	4.1238	4.4903	0.00012	0.35461	-0.31936
17480922	-0.82166	1.6666	-3.9313	0.00053	0.52058	-1.15606
17515315	0.75682	5.3939	3.8872	0.00059	0.52058	-1.22244
17528586	-0.85346	6.1029	-3.8463	0.00066	0.52058	-1.28416
17396334	0.76909	2.7897	3.8169	0.00071	0.52058	-1.32840

- Para la comparación 3 (Control vs Smoke): Genes que cambian su expresión por el efecto de exponer a una madre embarazada al humo de tabaco vs una madre sin dicha exposicion independientemente de su descendencia.

```
> topTab_Int <- topTable (fit.main, number=nrow(fit.main), coef="Int", adjust="fdr")
> head(topTab_Int)
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
17550478	-14.0313	5.9197	-47.6523	0	0.00000	37.5153
17546287	7.7014	3.2205	33.0376	0	0.00000	34.1430
17546834	7.2362	3.4021	28.3027	0	0.00000	32.2298
17546797	5.5485	2.7999	23.1495	0	0.00000	29.3102
17546316	5.4719	3.1343	15.8897	0	0.00000	22.7553
17366926	-2.0012	2.0231	-6.1002	0	0.00141	5.2264

La primera columna de cada tabla superior contiene la identificación del fabricante (Affymetrix) para cada conjunto de sondas. Mediante el proceso denominado **anotación** se puede generar la correspondencia de cada gen a cada ID de Affymetrix.

5.4. Visualización de Genes Significativamente Diferenciados

Una forma de visualizar los resultados es mediante un **volcano plot**, que representa en abscisas los cambios de expresión en escala logarítmica ("efecto biológico"), y en ordenadas el "menos logaritmo" del p-valor o alternativamente el estadístico.

```
> volcanoplot(fit.main, coef="CM_SM", style = "B-statistic", highlight=10,
+             main=paste("Genes Expresados Diferencialmente", colnames(cont.matrix)[1], sep="\n"))
>   abline(v=c(-1,1))
```

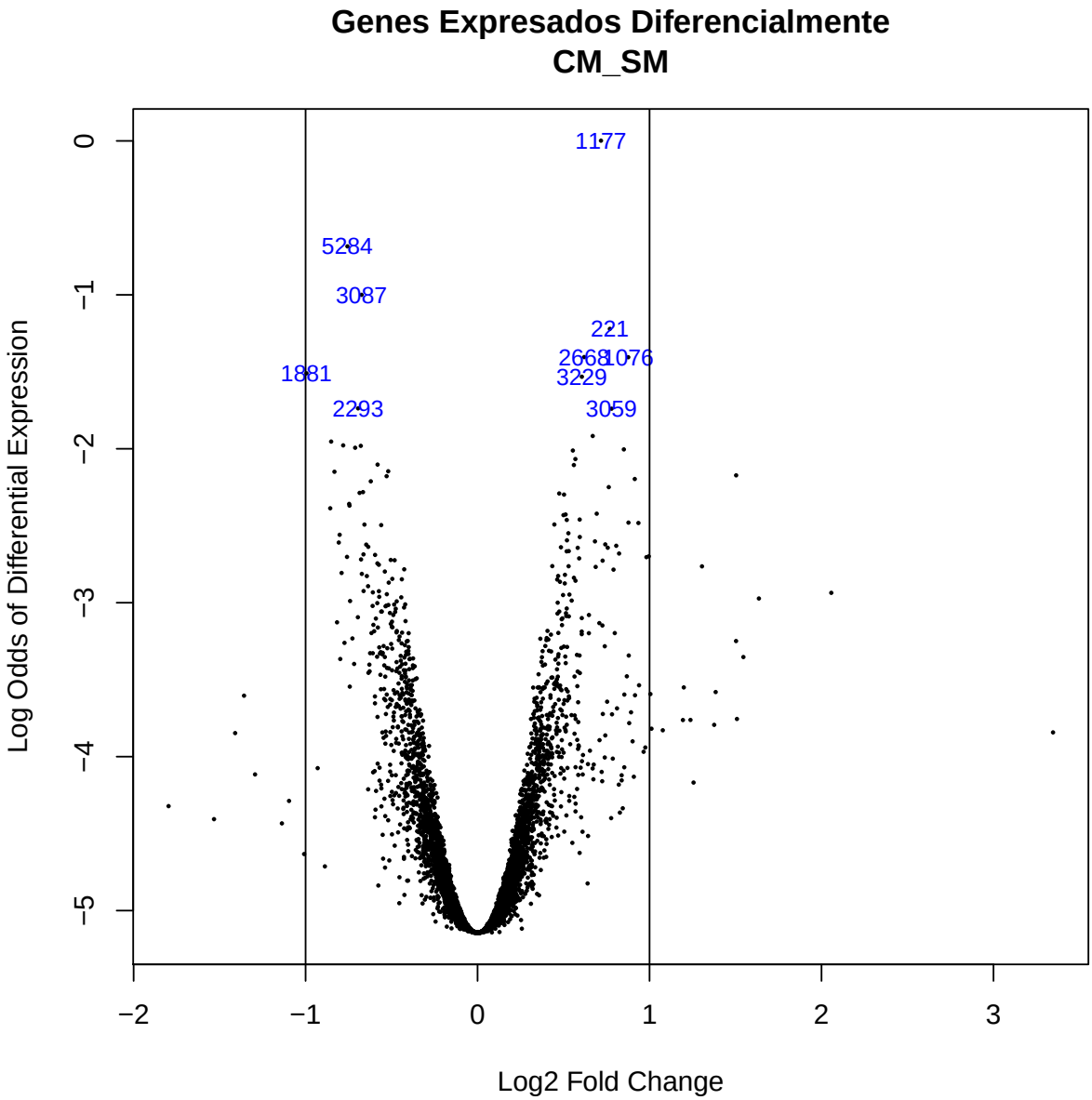

Figure 10: Volcano Plot que nos permite visualizar los genes expresados diferencialmente entre las madres control y las madres expuestas al humo del tabaco con descendencia masculina

```
> volcanoplot(fit.main, coef="CF_SF", style = "B-statistic", highlight=10,
+             main=paste("Genes Expresados Diferencialmente", colnames(cont.matrix)[2], sep="\n"))
>   abline(v=c(-1,1))
```

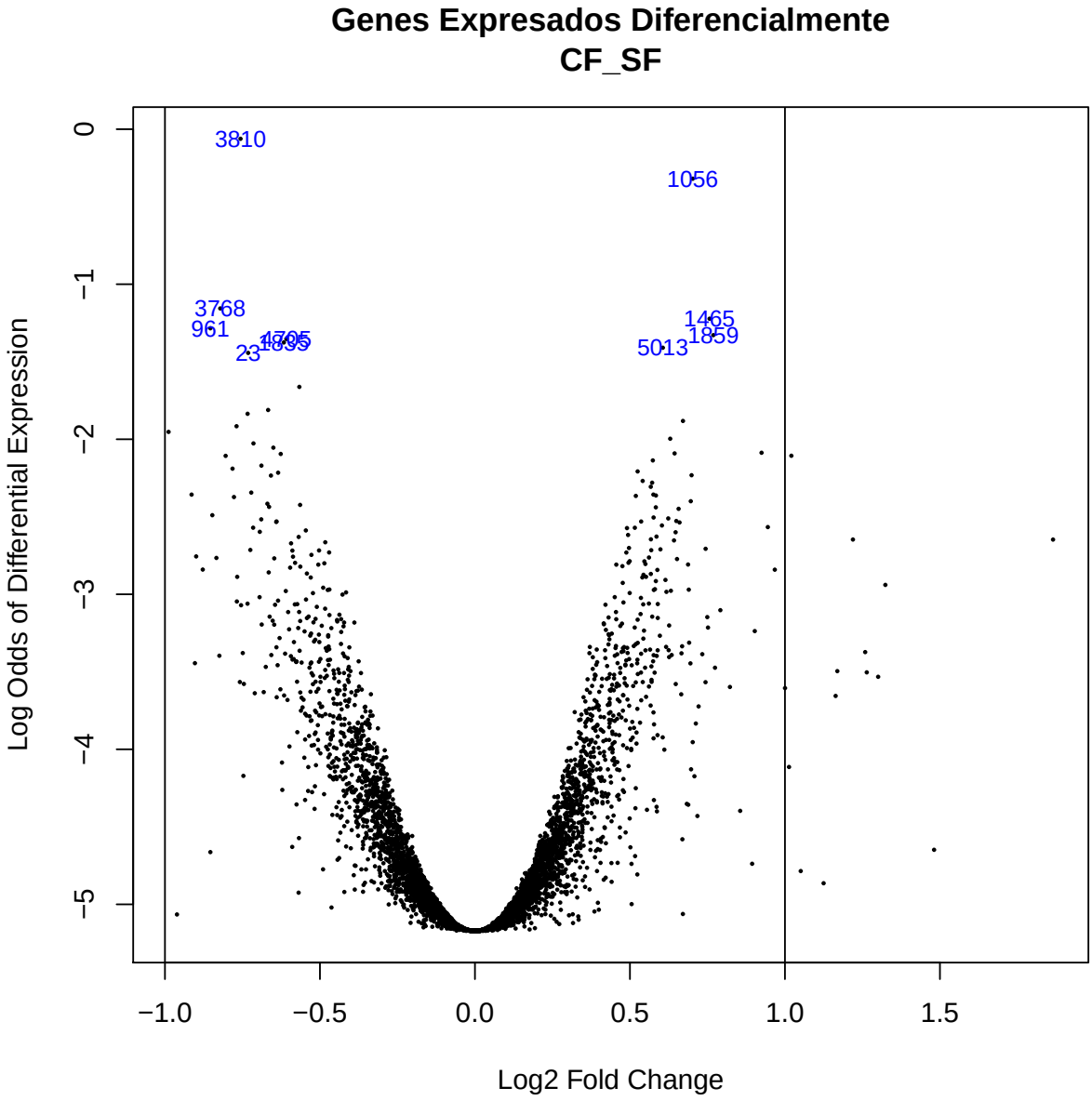

Figure 11: Volcano Plot que nos permite visualizar los genes expresados diferencialmente entre las madres control y las madres expuestas al humo del tabaco con descendencia femenina

```
> volcanoplot(fit.main, coef="Int", style = "B-statistic", highlight=10,
+             main=paste("Genes Expresados Diferencialmente", colnames(cont.matrix)[3], sep="\n"))
>   abline(v=c(-1,1))
```

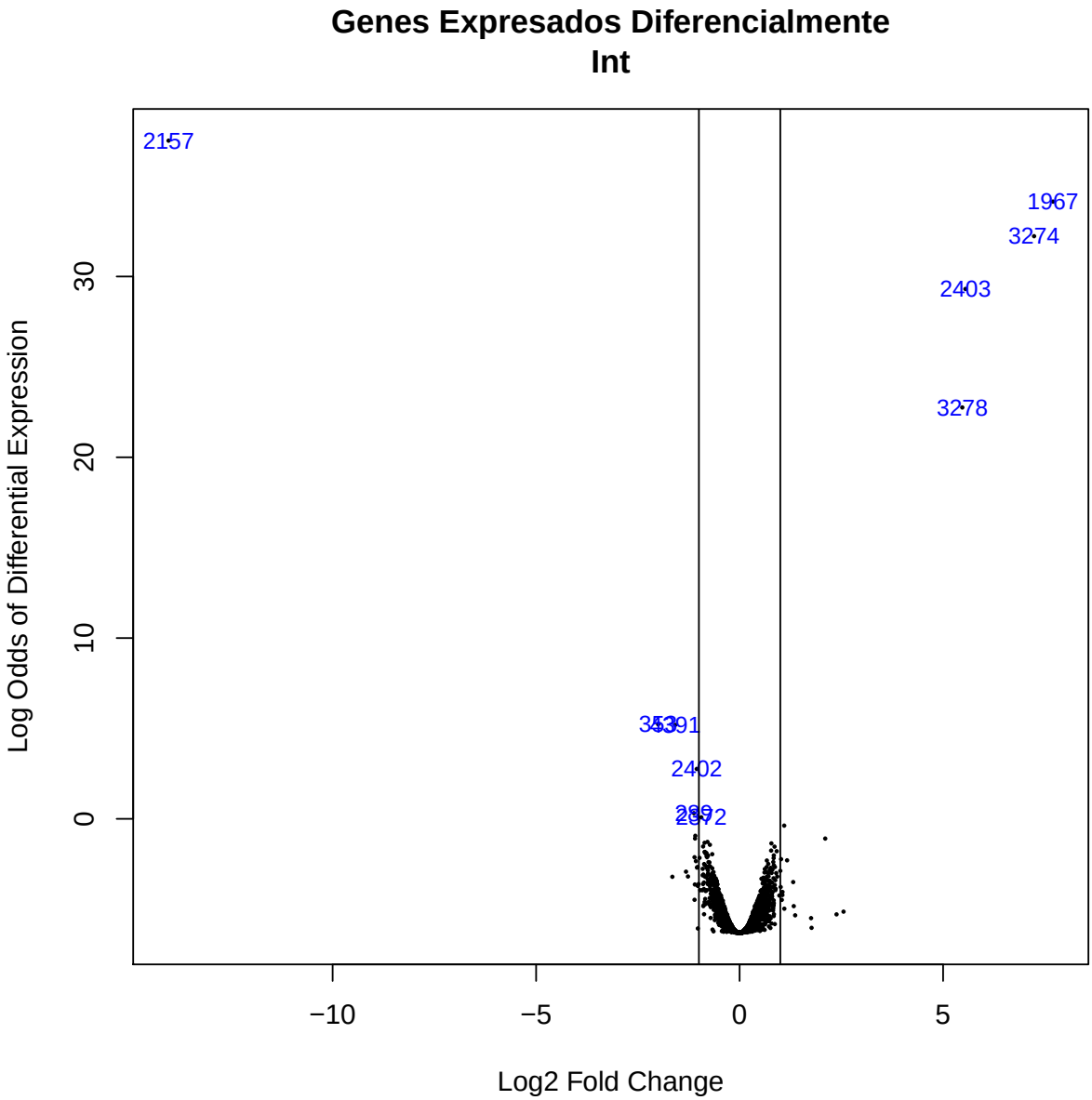

Figure 12: Volcano Plot que nos permite visualizar los genes expresados diferencialmente entre las madres control y las madres expuestas al humo del tabaco independientemente de su descendencia

Los genes cuyo log odds es superior a 0 y cuyo log2 fold change es, en valor absoluto, superior a 1, son candidatos a estar diferencialmente expresados.

Si consideraramos genes expresados diferencialmente aquellos que tienen un p-valor ajustado inferior a 0.05 o 0.01, el numero de genes diferencialmente expresado en cada caso seria:

```
Numero de genes con un p-valor inferior a 0.05:

Control_Macho vs Smoke_Macho: 0

Control_Hembra vs Smoke_Hembra:  0
```

```
Control vs Smoke:  8


Numero de genes con un p-valor inferior a 0.01:

Control_Macho vs Smoke_Macho:  0

Control_Hembra vs Smoke_Hembra:  0

Control vs Smoke:  7
```

Por lo que vemos que solo se aprecían genes diferencialmente expresados en la comparativa entre los organismos control y los organismos expuestos al humo de tabaco independientemente de su descendencia. Los 8 genes más expresados en esta comparativa son los siguientes:

```
> Int <- head(topTab_Int[1:8,1:6])
> show(Int)
```

```
           logFC AveExpr        t    P.Value  adj.P.Val       B
17550478 -14.0314  5.9197 -47.6523 1.0231e-27 6.1292e-24 37.5153
17546287   7.7014  3.2205  33.0375 1.8114e-23 5.4261e-20 34.1430
17546834   7.2361  3.4021  28.3027 1.0781e-21 2.1529e-18 32.2298
17546797   5.5485  2.7999  23.1495 2.0492e-19 3.0692e-16 29.3101
17546316   5.4719  3.1343  15.8897 2.8167e-15 3.3750e-12 22.7553
17366926  -2.0012  2.0231  -6.1002 1.5863e-06 1.4080e-03  5.2264
```

5.5. Comparaciones Multiples

Cuando se realizan varias comparaciones a la vez puede resultar importante ver qué genes cambian simultáneamente en más de una comparación. La función *decidetests* permite seleccionarán los genes que cambian en una o más condiciones.

El resultado del análisis es una tabla, que llamaremos res y que para cada gen y cada comparación contiene un 1 (si el gen esta sobreexpresado o up en esta condición), un 0 (si no hay cambio significativo) o un -1 (si esta down regulado).

Para resumir dicho análisis podemos contar qué filas tienen como mínimo una celda distinta de cero:

```
> library(limma)
> res<-decideTests(fit.main, method="separate", adjust.method="fdr", p.value=0.05, lfc=1)
```

```
> sum.res.rows<-apply(abs(res),1,sum)
> res.selected<-res[sum.res.rows!=0,]
> print(summary(res))
```

```
       CM_SM CF_SF  Int
Down       0     0    4
NotSig  5991  5991 5983
Up         0     0    4
```

El **diagrama de Venn** permite visualizar cuántos de estos genes son compartidos por una o más selecciones sin diferenciar entre genes up o down regulados.

```
> vennDiagram (res.selected[,1:3], cex=0.9)
> title("Genes en común entre las tres comparaciones \n Genes seleccionados con FDR <0.1 y logFC> 1")
```

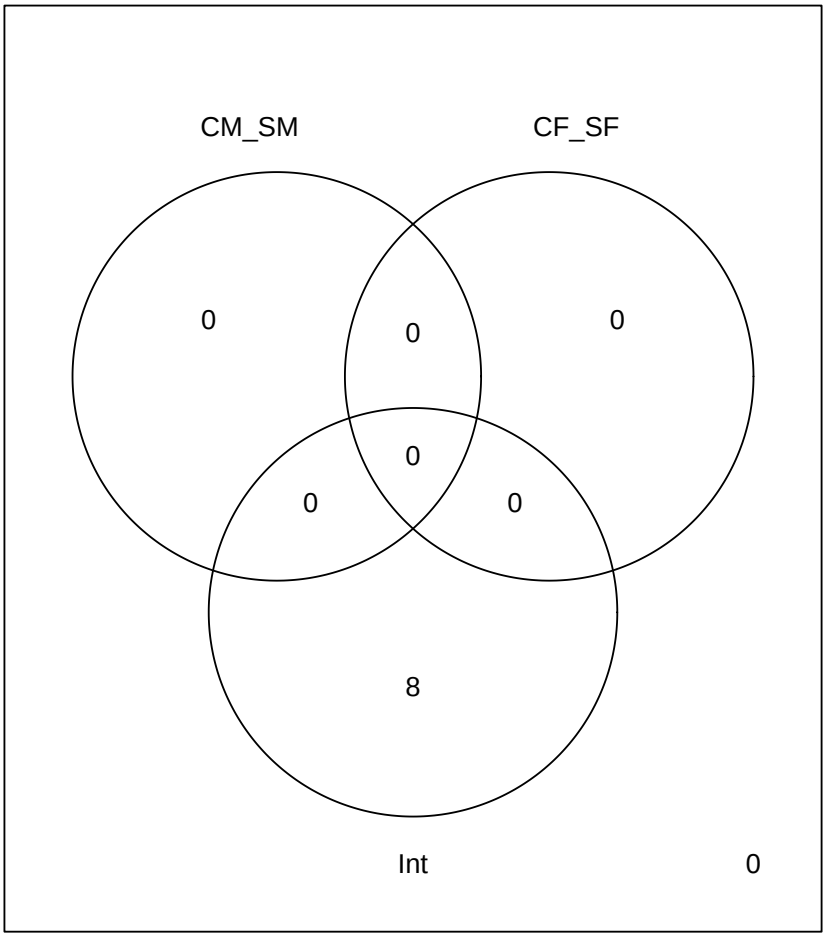

Figure 13: Diagrama de Venn que muestra los genes en común entre las tres comparaciones realizadas

Mediante la tabla anterior podemos deducir que la diferenciación de genes se establece entre organismos expuestos al humo de tabaco vs los organismos denominados control independientemente del sexo de su descendencia.

5.6. Anotación de Genes

En el proceso de *"anotación"* buscaremos información para asociar los identificadores que aparecen en las tablas superiores, generalmente correspondientes a sondas o transcripciones que dependen del tipo de matriz, con nombres más familiares como el Símbolo del gen, el Identificador del gen Entrez o la descripción del gen.

Por simplicidad, como disponemos de tres tablas, se prepara y utiliza una función que anota una tabla

superior con un paquete dado.

```
> require(mogene21sttranscriptcluster.db)
```

```
> annotatedTopTable <- function(topTab, anotPackage)
+ {
+   topTab <- cbind(PROBEID=rownames(topTab), topTab)
+   myProbes <- rownames(topTab)
+   thePackage <- eval(parse(text = anotPackage))
+   geneAnots <- select(thePackage, myProbes, c("SYMBOL", "ENTREZID", "GENENAME"))
+   annotatedTopTab<- merge(x=geneAnots, y=topTab, by.x="PROBEID", by.y="PROBEID")
+ return(annotatedTopTab)
+ }
```

```
> topAnnotated_CM_SM <- annotatedTopTable(topTab_CM_SM,
+ anotPackage="mogene21sttranscriptcluster.db")
> topAnnotated_CF_SF <- annotatedTopTable(topTab_CF_SF,
+ anotPackage="mogene21sttranscriptcluster.db")
> topAnnotated_Int <- annotatedTopTable(topTab_Int,
+ anotPackage="mogene21sttranscriptcluster.db")
> write.csv(topAnnotated_CM_SM, file="./results/topAnnotated_CM_SM.csv")
> write.csv(topAnnotated_CF_SF, file="./results/topAnnotated_CF_SF.csv")
> write.csv(topAnnotated_Int, file="./results/topAnnotated_Int.csv")
```

La anotación hace que las tablas sean más comprensibles. La tabla generada muestra las anotaciones agregadas a los resultados "topTable" para la comparación "Int" (solo se muestran las primeras cuatro columnas).

```
   PROBEID SYMBOL ENTREZID                                    GENENAME
1 17211000   Rrs1    59014          ribosome biogenesis regulator 1
2 17211004 Adhfe1    76187 alcohol dehydrogenase, iron containing, 1
3 17211198  Sulf1   240725                                sulfatase 1
4 17211347 Tfap2b    21419           transcription factor AP-2 beta
5 17211441 Mir30a   387225                               microRNA 30a
```

Para disponer de una información más completa que la obtenida en la tabla anterior, podemos emplear el paquete de Bioconductor *annaffy* que genera una tabla de anotaciones con enlaces a las Bases de Datos para cada anotación seleccionada.

```
> require(annaffy)
> genesSelected <- rownames(res.selected)
> at <- aafTableAnn(genesSelected, "mogene21sttranscriptcluster.db")
> saveHTML (at, file.path(resultsDir, "anotations.html"), "Anotaciones para Genes Seleccionados")
```

5.7. Visualizacion de Genes Expresados Diferencialmente

Tras la generación de la tabla de asociación, volvemos a generar el *volcano plot*, solo para la comparativa de los individuos Control vs los individuos expuestos al Humo independiente del sexo de su descendencia, para observar aquellos genes expresados diferencialmente ya identificados mediante el proceso de Anotación.

```
> library(mogene21sttranscriptcluster.db)
> geneSymbols <- select(mogene21sttranscriptcluster.db, rownames(fit.main), c("SYMBOL"))
> SYMBOLS<- geneSymbols$SYMBOL
```

```
> volcanoplot(fit.main, coef="Int", highlight=8, names=SYMBOLS,
+             main=paste("Genes Expresados Diferencialmente", colnames(cont.matrix)[3], sep="\n"))
>   abline(v=c(-1,1))
```

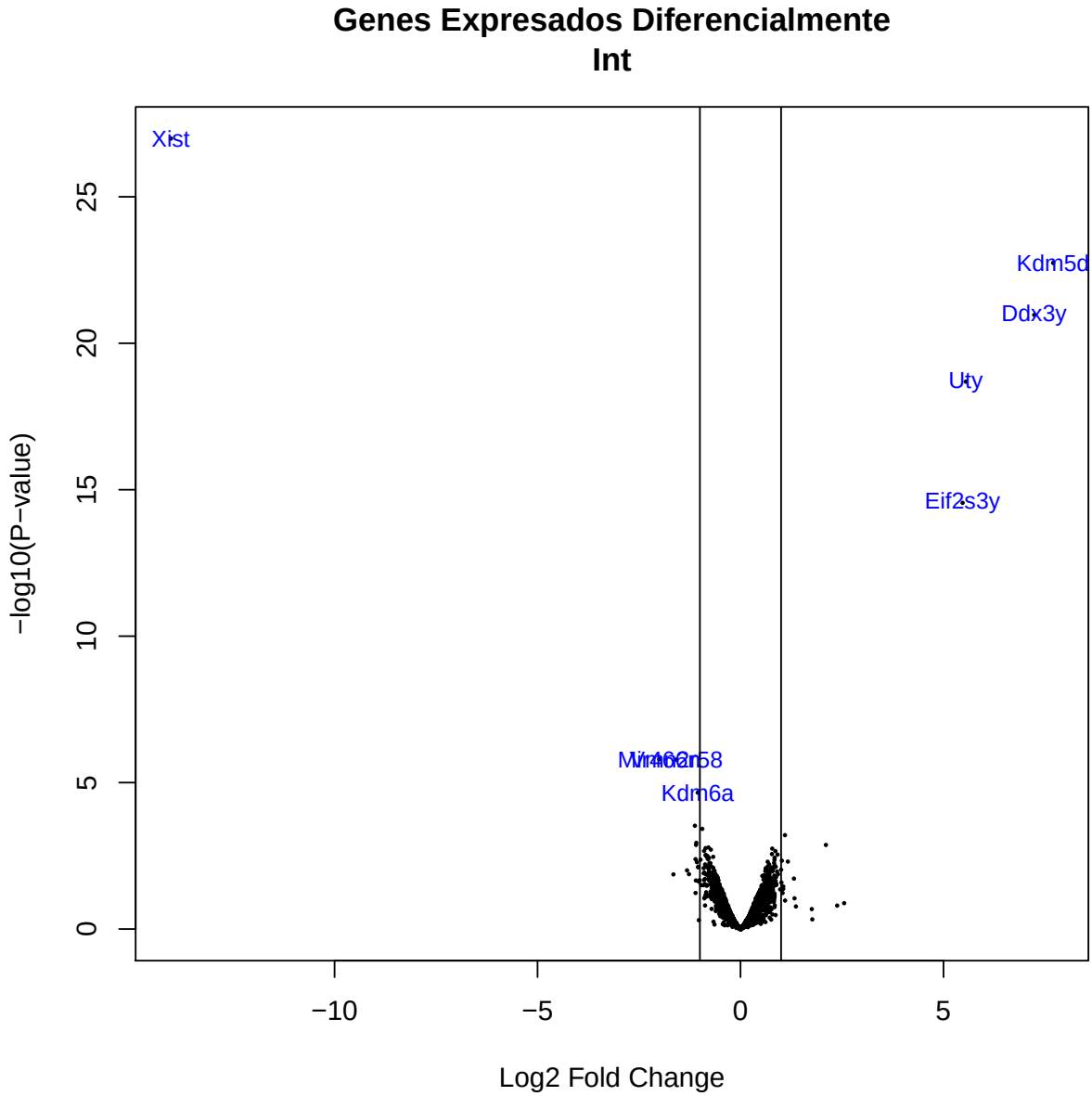

Figure 14: Volcano plot para la comparación entre Control y Smoke independientemente de su descendencia donde solo se muestran los nombres de los 8 genes principales de la tabla superior

5.8. Mapas de Calor

Los genes que han sido seleccionados como diferencialmente expresados pueden visualizarse usando un mapa de calor. Estas gráficas usan paletas de colores para resaltar valores distintos: expresiones significativamente diferenciales positivas (regulación ascendente) o negativas (regulación descendente).

Los **mapas de calor** se pueden usar para visualizar los valores de expresión de genes expresados diferencialmente sin un orden específico, pero generalmente se prefiere trazarlos haciendo un agrupamiento jerárquico en genes (filas) o columnas (muestras) para encontrar grupos de genes con patrones comunes de variación que eventualmente puede asociarse a los diferentes grupos que se comparan.

Puede haber una discusión sobre qué genes seleccionar para hacer un mapa de calor. Una opción común es seleccionar los genes que se han seleccionado en los pasos anteriores, es decir, los genes que se han denominado diferencialmente expresados en al menos una comparación.

```
> probesInHeatmap <- rownames(res.selected)
> HMdata <- exprs(eset_filtered)[rownames(exprs(eset_filtered)) %in% probesInHeatmap,]
>
> geneSymbols <- select(mogene21sttranscriptcluster.db, rownames(HMdata), c("SYMBOL"))
> SYMBOLS<- geneSymbols$SYMBOL
> rownames(HMdata) <- SYMBOLS
> write.csv(HMdata, file = file.path("./results/data4Heatmap.csv"))
```

Con los datos seleccionados se puede generar un mapa de calor con o sin agrupamiento de genes y / o muestras.

- El mapa de calor producido para todos los genes seleccionados con los mismos criterios descritos anteriormente (FDR <0.1 y logFC> 1) donde no se realiza la agrupación de genes y muestras.

```
> color.map <- function(grupo) {
+   if (grupo=="A"){
+     c<- "yellow"
+   }else{
+     if (grupo=="B"){
+       c<- "red"
+     }else{
+       c<- "blue"
+    }
+   }
+ return(c)}
> grupColors <- unlist(lapply(pData(eset_filtered)$Grupo, color.map))
> library(gplots)
> heatmap.2(HMdata,
+           Rowv = FALSE,
+           Colv = FALSE,
+           main = "Genes Diferencialmente Expresados \n FDR < 0,1, logFC >=1",
+           scale = "row",
+           col = bluered(75),
+           sepcolor = "white",
+           sepwidth = c(0.05,0.05),
+           cexRow = 0.5,
+           cexCol = 0.9,
+           key = TRUE,
+           keysize = 1.5,
+           density.info = "histogram",
+           ColSideColors = grupColors,
+           tracecol = NULL,
+           dendrogram = "none",
+           srtCol = 30)
```

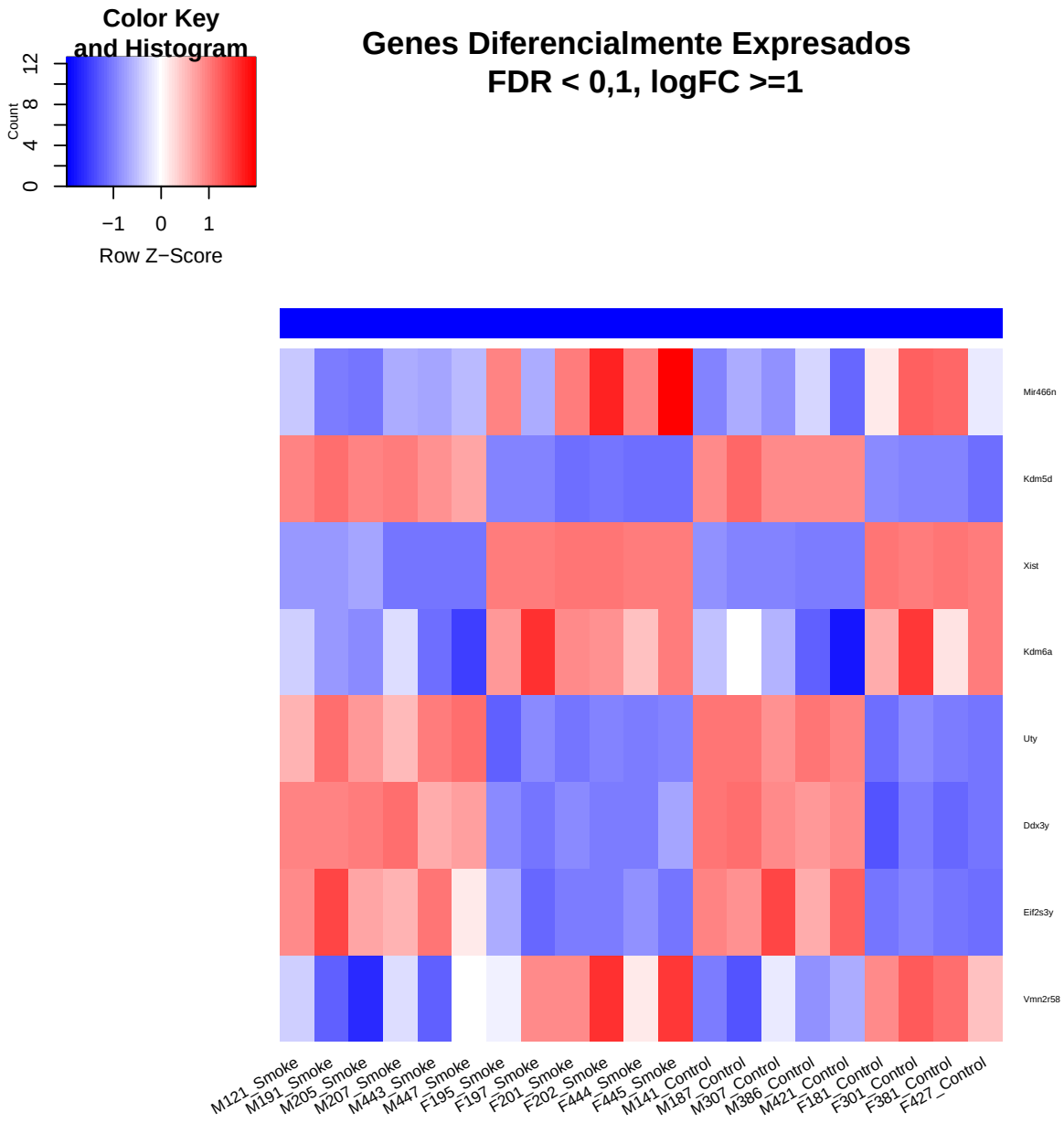

Figure 15: Mapa de Calor para datos de expresión sin ninguna agrupación

- El mapa de calor producido para todos los genes seleccionados con los mismos criterios descritos anteriormente (FDR <0.1 y logFC> 1) donde los genes y las muestras se ven obligados a agruparse por fila y columna de forma similar.

```
> heatmap.2(HMdata,
+           Rowv = TRUE,
+           Colv = TRUE,
+           dendrogram = "both",
+           main = "Differentially expressed genes \n FDR < 0,1, logFC >=1",
+           scale = "row",
+           col = bluered(75),
```

```
+            sepcolor = "white",
+            sepwidth = c(0.05,0.05),
+            cexRow = 0.5,
+            cexCol = 0.9,
+            key = TRUE,
+            keysize = 1.5,
+            density.info = "histogram",
+            ColSideColors = grupColors,
+            tracecol = NULL,
+            srtCol = 30)
```

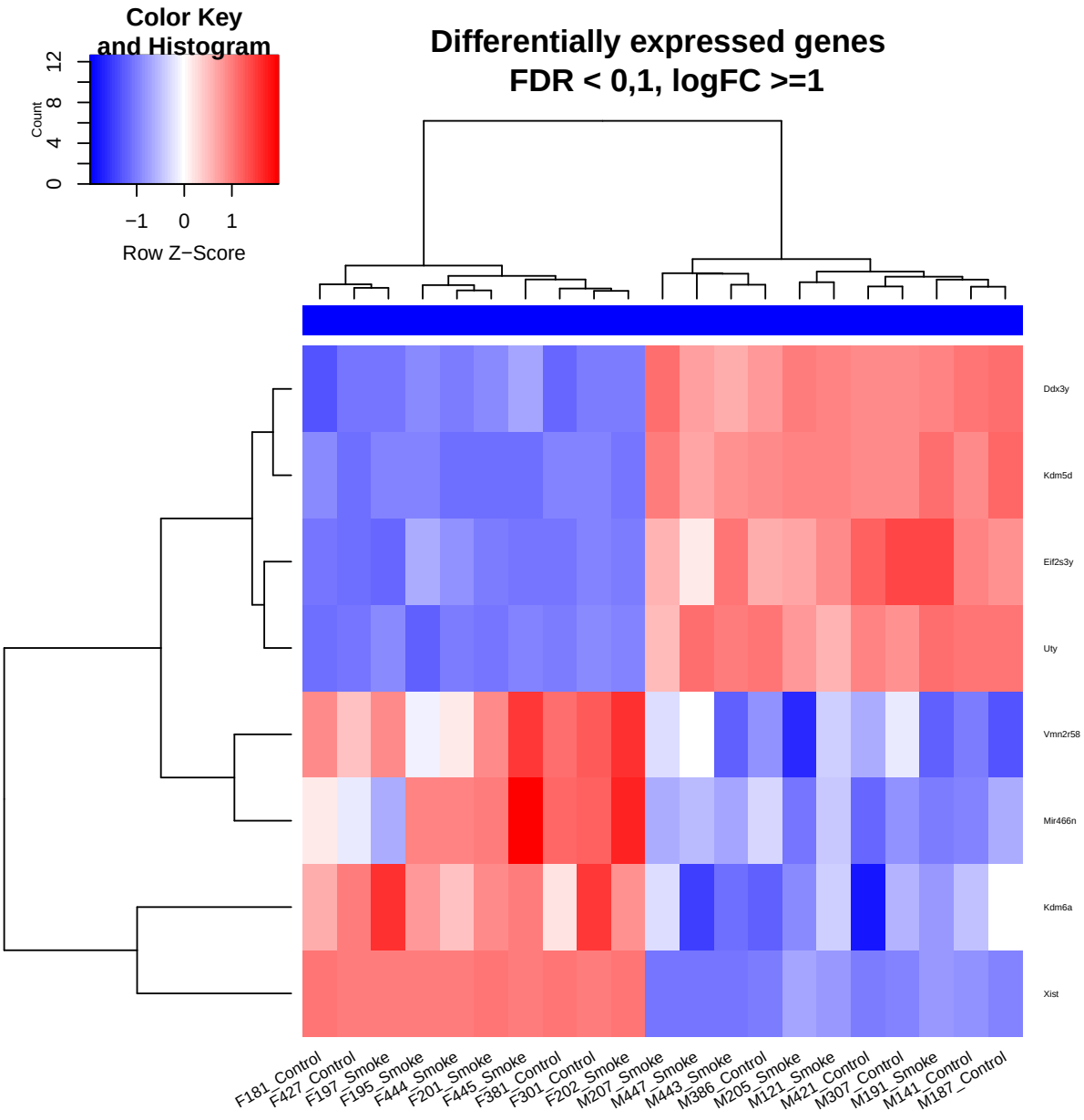

Figure 16: Mapa de calor que agrupa genes (filas) y muestras (columnas) por su similitud

Como puede verse aquellos que presentan una descendencia de machos tienen perfiles de expresión mas parecidos entre ellos, mientras que los que presentan descendencia de hembra presentan más parecidos entre ellos.

6. Análisis de la Significación Biólogica

El Análisis de Significación Biológica busca establecer si, dada una lista de genes seleccionados por ser diferencial expresada entre dos condiciones, las funciones, procesos biológicos o vías moleculares que los caracterizan aparecen en esta lista con más frecuencia que entre el resto de los genes analizados.

En este estudio usaremos el análisis de enriquecimiento básico mediante el paquete de bioconductores *ReactomePA*. El análisis se realiza en la base de datos de anotaciones **ReactomePA**.

Los análisis de este tipo necesitan un número mínimo de genes para ser confiables, preferiblemente unos pocos cientos, por lo que es común realizar una selección menos restrictiva que con los pasos anteriores.

Por ejemplo, una opción es incluir todos los genes con un límite de FDR no estricto, como FDR <0.15 sin filtrado o emplear p-valores sin ajustar.

```
> listOfTables <- list(CM_SM = topTab_CM_SM,
+                      CF_SF = topTab_CF_SF,
+                      INT = topTab_Int)
> listOfSelected <- list()
> for (i in 1:length(listOfTables)){
+   # generaremos toptable
+   topTab <- listOfTables[[i]]
+   # selección de genes para incluir en el análisis
+   whichGenes<-topTab["adj.P.Val"]<0.15
+   selectedIDs <- rownames(topTab)[whichGenes]
+   # convertimos el ID en identificador de Entrez
+   EntrezIDs<- select(mogene21sttranscriptcluster.db, selectedIDs, c("ENTREZID"))
+   EntrezIDs <- EntrezIDs$ENTREZID
+   listOfSelected[[i]] <- EntrezIDs
+   names(listOfSelected)[i] <- names(listOfTables)[i]
+ }
> sapply(listOfSelected, length)
```

```
CM_SM CF_SF   INT
    0     0     8
```

Lamentablemente, en el estudio que he seleccionado no existen genes expresados diferencialmente en 2 de las 3 comparaciones establecidas si tomamos los valores del p-valor ajustado. Solo disponemos de 8 genes expresados diferencialmente de la comparación Int, y este tipo de análisis solo trabaja bien con varios centenares de genes.

Por lo que en el caso de mi estudio no tiene mucho sentido realizar un Análisis de Significación Biológica teniendo en cuenta este aspecto. Por este motivo, basaré mi esrudio en los p-valores sin ajustar.

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

```
> listOfTables <- list(CM_SM = topTab_CM_SM,
+                      CF_SF = topTab_CF_SF,
+                      INT = topTab_Int)
> listOfSelected <- list()
> for (i in 1:length(listOfTables)){
+   # Crearemos el toptab
+   topTab <- listOfTables[[i]]
```

```
+   # seleccionaremos los genes a incluir en el análisis teniendo en cuenta el p-valor sin ajustar
+   whichGenes<-topTab["P.Value"]<0.15
+   selectedIDs <- rownames(topTab)[whichGenes]
+   # convertimos el ID en identificador de Entrez
+   EntrezIDs<- select(mogene21sttranscriptcluster.db, selectedIDs, c("ENTREZID"))
+   EntrezIDs <- EntrezIDs$ENTREZID
+   listOfSelected[[i]] <- EntrezIDs
+   names(listOfSelected)[i] <- names(listOfTables)[i]
+ }
> sapply(listOfSelected, length)
```

```
CM_SM CF_SF   INT
 1243  1234   779
```

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

El análisis de significación biológica se aplicará solo a las dos primeras listas de las cuales podemos comparar la exposición al humo del tabaco en madres embarazadas vs organismos control dependiendo del sexo de su descendencia.

```
> library(ReactomePA)
>
> listOfData <- listOfSelected[1:2]
> comparisonsNames <- names(listOfData)
> universe <- mapped_genes
>
> for (i in 1:length(listOfData)){
+   genesIn <- listOfData[[i]]
+   comparison <- comparisonsNames[i]
+   enrich.result <- enrichPathway(gene = genesIn,
+                                  pvalueCutoff = 0.05,
+                                  readable = T,
+                                  pAdjustMethod = "BH",
+                                  organism = "mouse",
+                                  universe = universe)
+
+   cat("##################################")
+   cat("\nComparison: ", comparison,"\n")
+   print(head(enrich.result))
+
+   if (length(rownames(enrich.result@result)) != 0) {
+   write.csv(as.data.frame(enrich.result),
+             file =paste0("./results/","ReactomePA.Results.",comparison,".csv"),
+             row.names = FALSE)
+
+   pdf(file=paste0("./results/","ReactomePABarplot.",comparison,".pdf"))
+     print(barplot(enrich.result, showCategory = 15, font.size = 4,
+             title = paste0("Reactome Pathway Analysis for ", comparison,". Barplot")))
+   dev.off()
+
+   pdf(file = paste0("./results/","ReactomePAcnetplot.",comparison,".pdf"))
+     print(cnetplot(enrich.result, categorySize = "geneNum", schowCategory = 15,
+          vertex.label.cex = 0.75))
```

```
+   dev.off()
+   }
+ }

###################################
Comparison:  CM_SM
                        ID
R-MMU-69278     R-MMU-69278
R-MMU-69620     R-MMU-69620
R-MMU-2500257 R-MMU-2500257
R-MMU-68877     R-MMU-68877
R-MMU-141424   R-MMU-141424
R-MMU-141444   R-MMU-141444
                                                                                Description
R-MMU-69278                                                                Cell Cycle, Mitotic
R-MMU-69620                                                              Cell Cycle Checkpoints
R-MMU-2500257                                                 Resolution of Sister Chromatid Cohesion
R-MMU-68877                                                                 Mitotic Prometaphase
R-MMU-141424                                          Amplification of signal from the kinetochores
R-MMU-141444  Amplification  of signal from unattached  kinetochores via a MAD2  inhibitory signal
              GeneRatio  BgRatio     pvalue   p.adjust     qvalue
R-MMU-69278      85/524 499/8772 1.5537e-19 1.2429e-16 1.0810e-16
R-MMU-69620      58/524 281/8772 1.9859e-17 7.9436e-15 6.9088e-15
R-MMU-2500257    32/524 118/8772 1.6193e-13 4.3180e-11 3.7556e-11
R-MMU-68877      41/524 189/8772 2.2646e-13 4.5293e-11 3.9393e-11
R-MMU-141424     27/524  91/8772 1.2655e-12 1.6873e-10 1.4675e-10
R-MMU-141444     27/524  91/8772 1.2655e-12 1.6873e-10 1.4675e-10

R-MMU-69278   Rfc5/Ccnb2/Mcm6/Bub1b/Cdc25c/Cdc6/Tubb2a/Cdk1/Ncapd2/Mastl/Anapc15/Kif18a/Fbxo5/Ccna2/Cenp
R-MMU-69620
R-MMU-2500257
R-MMU-68877
R-MMU-141424
R-MMU-141444
              Count
R-MMU-69278      85
R-MMU-69620      58
R-MMU-2500257    32
R-MMU-68877      41
R-MMU-141424     27
R-MMU-141444     27
###################################
Comparison:  CF_SF
                        ID
R-MMU-69278     R-MMU-69278
R-MMU-2500257 R-MMU-2500257
R-MMU-68877     R-MMU-68877
R-MMU-141424   R-MMU-141424
R-MMU-141444   R-MMU-141444
R-MMU-5663220 R-MMU-5663220
                                                                                Description
R-MMU-69278                                                                Cell Cycle, Mitotic
R-MMU-2500257                                                 Resolution of Sister Chromatid Cohesion
```

```
R-MMU-68877                                                                  Mitotic Prometaphase
R-MMU-141424                                          Amplification of signal from the kinetochores
R-MMU-141444  Amplification  of signal from unattached  kinetochores via a MAD2  inhibitory signal
R-MMU-5663220                                                        RHO GTPases Activate Formins
              GeneRatio  BgRatio     pvalue   p.adjust     qvalue
R-MMU-69278      89/494 499/8772 1.0309e-23 8.1437e-21 7.3028e-21
R-MMU-2500257    39/494 118/8772 1.8093e-20 7.1467e-18 6.4087e-18
R-MMU-68877      48/494 189/8772 1.6976e-19 4.4704e-17 4.0087e-17
R-MMU-141424     32/494  91/8772 6.5651e-18 1.0373e-15 9.3017e-16
R-MMU-141444     32/494  91/8772 6.5651e-18 1.0373e-15 9.3017e-16
R-MMU-5663220    37/494 132/8772 8.6487e-17 1.1387e-14 1.0212e-14

R-MMU-69278   Ccnb2/Pold3/Mcm2/Ndc80/Cdca5/Cenpu/Cenpa/B9d2/Pole/Ercc6l/Ccna2/Tubb4a/Spc25/Kif2c/Bub1b/
R-MMU-2500257
R-MMU-68877
R-MMU-141424
R-MMU-141444
R-MMU-5663220
              Count
R-MMU-69278      89
R-MMU-2500257    39
R-MMU-68877      48
R-MMU-141424     32
R-MMU-141444     32
R-MMU-5663220    37
```

Los resultados obtenidos en el análisis de importancia biológica son:

- un *archivo .csv* con un resumen de todas las rutas enriquecidas y las estadísticas asociadas.
- *un diagrama de barras* con las mejores vías enriquecidas. La altura del gráfico de barras es el número de genes de nuestro análisis relacionados con esa vía. Además, las vías están ordenadas por significación estadística.
- una trama con una *red de las vías enriquecidas* y la relación entre los genes incluidos.

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

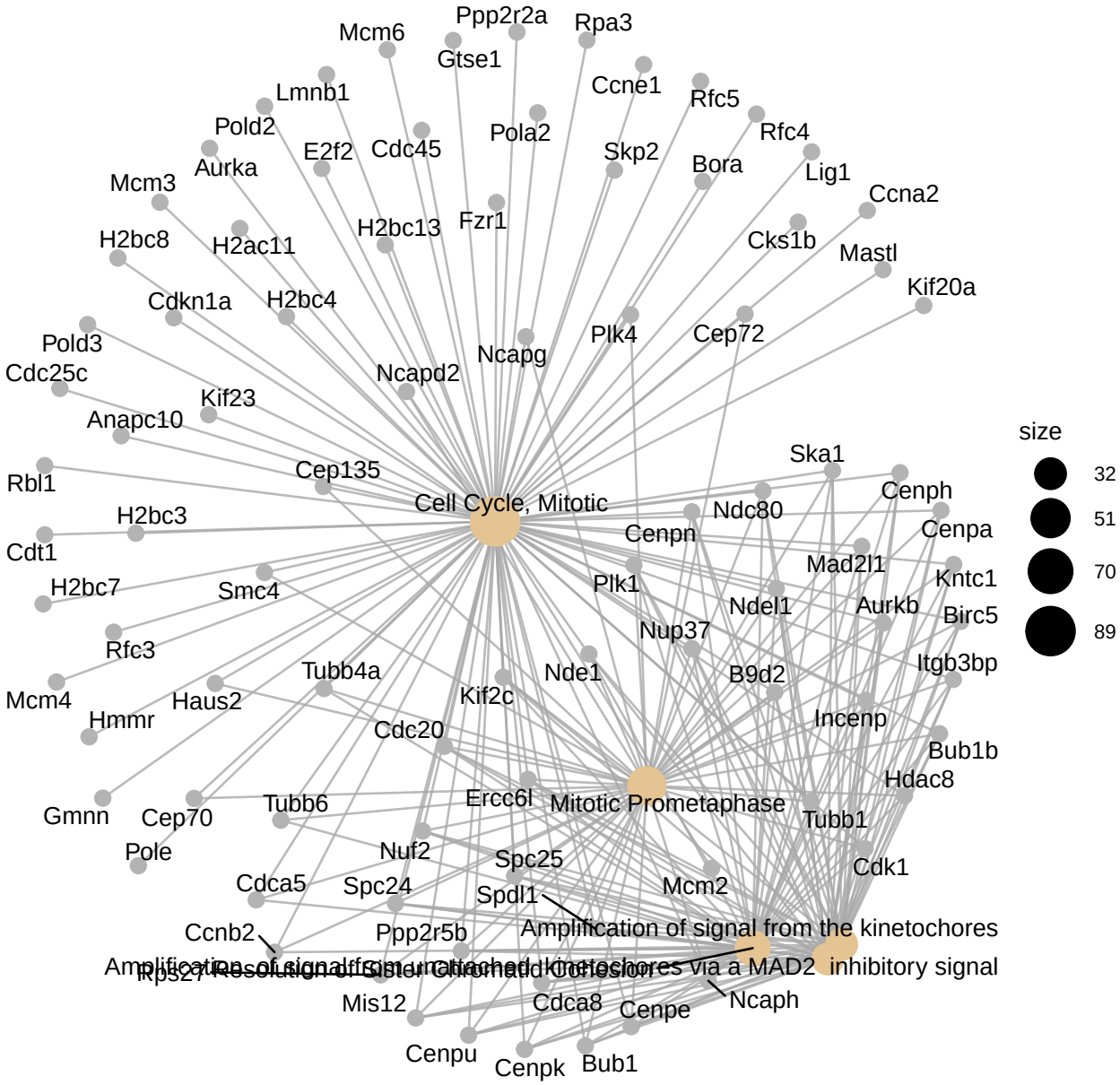

Figure 17: Red obtenida del análisis de enriquecimiento de Reactome en la lista obtenida de la comparación entre Control Male vs Smoke Male y Control Female vs Smoke Female

En nuestro estudio de comparación, tanto en el caso de *CM_SM* como en *CF_SF* se ha encontrado una vía destacable relacionada con el *Ciclo Celular Mitótico (Cell Cycle, Mitotic)* en ambos casos.

Table 5: Primeras filas y columnas para los resultados de Reactome en la Comparación CM_SM.csv

	Description	GeneRatio	BgRatio	pvalue	p.adjust
R-MMU-69278	Cell Cycle, Mitotic	85/524	499/8772	1.55368041876834e-19	1.24294433
R-MMU-69620	Cell Cycle Checkpoints	58/524	281/8772	1.98590160640613e-17	7.94360642

	Description	GeneRatio	BgRatio	pvalue	p.adjust
R-MMU-2500257	Resolution of Sister Chromatid Cohesion	32/524	118/8772	1.61926783459435e-13	4.31804755
R-MMU-68877	Mitotic Prometaphase	41/524	189/8772	2.26463742043144e-13	4.52927484

Table 6: Primeras filas y columnas para los resultados de Reactome en la Comparación CF_SF.csv

	Description	GeneRatio	BgRatio	pvalue	p.adju
R-MMU-69278	Cell Cycle, Mitotic	89/494	499/8772	1.03085354016541e-23	8.1437
R-MMU-2500257	Resolution of Sister Chromatid Cohesion	39/494	118/8772	1.80929220460399e-20	7.1467
R-MMU-68877	Mitotic Prometaphase	48/494	189/8772	1.69760992860313e-19	4.4703
R-MMU-141424	Amplification of signal from the kinetochores	32/494	91/8772	6.5650956058875e-18	1.0372

7. Resumen de Resultados y Discusión

Mediante el *preprocesado* realizado en la fase de Exploración y Control de Calidad hemos podido comprobar que, salvo en dos de las 21 muestras presentadas en el estudio donde se refleja una mínima heterogeneidad, la calidad de los datos era buena.

Los *datos normalizados y filtrados* nos han permitido detectar una serie de genes diferencialmente expresados con los que hemos procedido a realizar los estudios de significacion biologica.

En el estudio que he seleccionado, así como las comparaciones establecidas para el mismo, no existen genes expresados diferencialmente en 2 de las 3 comparaciones establecidas si tomamos los valores del p-valor ajustado. Solo disponemos de 8 genes expresados diferencialmente de la comparación Int. Por lo que en el caso de mi estudio, no tiene mucho sentido realizar un *Análisis de Significación Biológica* teniendo en cuenta este aspecto y por este motivo, fue necesario basar mi estudio en los p-valores sin ajustar.

Tras la realización de dichos análisis, se puede llegar a la *conclusión* que no existen grandes relaciones entre el sexo de la descendencia y la exposición de la madre al humo del tabaco durante el embarazo.

Los archivos generados durante el análisis han sido los siguientes:

Table 7: Listado de Ficheros generados durante el Análisis

Listado de Ficheros
anotations.html
data4Heatmap.csv
normalized.Data.csv
normalized.Data.Rda
normalized.Filtered.Data.csv
QCDir.Norm
rawData_quality
ReactomePA.Results.CF_SF.csv
ReactomePA.Results.CM_SM.csv
ReactomePABarplot.CF_SF.pdf
ReactomePABarplot.CM_SM.pdf
ReactomePAcnetplot.CF_SF.pdf
ReactomePAcnetplot.CM_SM.pdf
topAnnotated_CF_SF.csv
topAnnotated_CM_SM.csv
topAnnotated_Int.csv

7.1. Discusión

En relación al estudio realizado podemos destacar los siguientes puntos:

- El número de muestra utilizadas para el estudio es demasiado pequeña, por lo que se puede generar un mayor número de errores, más falsos negativos y problemas en la interpretación de los resultados.
- La comparaciones establecidas en el estudio no siguen los objetivos que se establecían en el artículo. Esto hace que el estudio no esté preparado para dar resultados eficientes ya que los datos no estaban destinados al tipo de análisis realizado.
- Los métodos empleados se encuentran acordes a los diferentes casos que se nos han planteado en la asignatura. Probablemente existan métodos que se ajusten mucho mejor al estudio que queremos realizar y que sabremos aplicar a medida que vayamos adquiriendo conocimientos y experiencia en el Análisis de Datos Ómicos.

8. Bibliografía

- Obtención de los datos en la Base de Datos "Gene Expression Omnibus (GEO)": (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67883)
- Artículo completo de Dehmel para el estudio titulado "Intrauterine smoke exposure deregulates lung function, pulmonary transcriptomes, and in particular insulin-like growth factor (IGF)-1 in a sex-specific manner" a través del siguiente link: (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5953988/)
- Página de Bioconductor para la localización de los paquetes necesarios para el análisis: (https://www.bioconductor.org/install/)
- Página de apoyo y soporte en la resolución de errores obtenidos durante el análisis: (https://www.biostars.org/)
- Página de anotaciones para localizar la correspondiente al estudio realizado: (http://www.bioconductor.org/packages/release/data/annotation/)
- Statistical Analysis of Microarray data (adapted for teaching purposes), Based on Gonzalo, Ricardo and Sanchez-Pla, Alex (2019)
- Genómica funcional y análisis de microarrays, Módulo 2. Análisis de datos de microarrays, M. Carme Ruíz de Villa, Alex Sánchez-Pla
- Repositorio GitHub: (https://github.com/)
- Guía para el uso de RMarkdown: (https://bookdown.org/gboccardo/manual-ED-UCH/introduccion-al-uso-de-rmarkdown-para-la-compilacion-de-resultados-de-rstudio-en-diferentes-formatos.html)