PAC 1 - ANÁLISIS DE DATOS ÓMICOS (UOC)

Proceso de análisis de microarrays

María José Ojeda-Montes

2 de mayo, 2020

Contents

ANÁLISIS DE MICROARRAYS

Estudio de microarrays

- *Título del estudio*: **Identification of a nutrient-sensing transcriptional network in monocytes by using inbred rat models on a cafeteria diet (2016)** [1]

- *Autores*: Neus Martínez-Micaelo, Noemí González-Abuín, Ximena Terra, Ana Ardévol, Monterrat Pinent, Enrico Petretto, Jacques Behmoaras y Mayte Blay
- *Enlace al artículo*: PMID-27483348
- *Enlace a los datos*: GSE85167
- *Enlace a github*: Repositorio

El objetivo de la práctica es reproducir los resultados obtenidos por el propio artículo [1] siguiendo el protocolo ofrecido en el material de la asignatura de Análisis de Datos Ómicos [2], entendiendo el por qué de cada uno de los pasos que se realizan.

Por lo tanto, el objetivo principal de este estudio es **analizar el efecto de la dieta (STD vs cafetería) en la expresión genética de cada tipo de rata (lewis o wistar kyoto).** También se pretende **analizar si la dieta afecta en la expresión de monocitos de forma distinta en cada cepa**.

Identificación de los grupos

Se trata de una muestra de 20 animales divididos en 4 grupos cada uno de ellos con 5 réplicas por grupo. Para cada grupo se considera la combinación de dos factores con dos niveles cada uno de ellos:

- dieta (niveles: STD y CAF)
- cepa de rata (niveles: LEW y WKY)

Table 1: Identificación de los grupos para cada archivo CEL

FileName	Group	Cepa	Dieta	ShortName
GSM2259098_NML2012082901.CEL	LEW.STD	LEW	STD	LEW.STD.1
GSM2259099_NML2012082902.CEL	LEW.STD	LEW	STD	LEW.STD.2
GSM2259100_NML2012082903.CEL	LEW.STD	LEW	STD	LEW.STD.3
GSM2259101_NML2012082904.CEL	LEW.STD	LEW	STD	LEW.STD.4
GSM2259102_NML2012082905.CEL	LEW.STD	LEW	STD	LEW.STD.5
GSM2259103_NML2012082906.CEL	WKY.STD	WKY	STD	WKY.STD.1
GSM2259104_NML2012082907.CEL	WKY.STD	WKY	STD	WKY.STD.2
GSM2259105_NML2012082908.CEL	WKY.STD	WKY	STD	WKY.STD.3
GSM2259106_NML2012082909.CEL	WKY.STD	WKY	STD	WKY.STD.4
GSM2259107_NML2012082910.CEL	WKY.STD	WKY	STD	WKY.STD.5
GSM2259108_NML2012082911.CEL	LEW.CAF	LEW	CAF	LEW.CAF.1
GSM2259109_NML2012082912.CEL	LEW.CAF	LEW	CAF	LEW.CAF.2
GSM2259110_NML2012082913.CEL	LEW.CAF	LEW	CAF	LEW.CAF.3
GSM2259111_NML2012082914.CEL	LEW.CAF	LEW	CAF	LEW.CAF.4
GSM2259112_NML2012082915.CEL	LEW.CAF	LEW	CAF	LEW.CAF.5
GSM2259113_NML2012082916.CEL	WKY.CAF	WKY	CAF	WKY.CAF.1
GSM2259114_NML2012082917.CEL	WKY.CAF	WKY	CAF	WKY.CAF.2
GSM2259115_NML2012082918.CEL	WKY.CAF	WKY	CAF	WKY.CAF.3
GSM2259116_NML2012082919.CEL	WKY.CAF	WKY	CAF	WKY.CAF.4
GSM2259117_NML2012082920.CEL	WKY.CAF	WKY	CAF	WKY.CAF.5

- **Archivos CEL**

```
## GeneFeatureSet (storageMode: lockedEnvironment)
```

```
## assayData: 1 features, 20 samples
##   element names: exprs
## protocolData
##   rowNames: "LEW.STD.1" "LEW.STD.2" ... "WKY.CAF.5" (20 total)
##   varLabels: exprs dates
##   varMetadata: labelDescription channel
## phenoData
##   rowNames: "LEW.STD.1" "LEW.STD.2" ... "WKY.CAF.5" (20 total)
##   varLabels: X.Group. X.Cepa. X.Dieta. X.ShortName.
##   varMetadata: labelDescription channel
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: pd.ragene.1.0.st.v1
```

Se trata del archivo de datos con código GSE85167 que tiene 1102500 genes y 20 muestras.

Control de calidad de los datos crudos

	array	sampleNames	*1	*2	*3	X.Group.	X.Cepa.	X.Dieta.	X.ShortName.
☐	1	"LEW.STD.1"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.1"
☐	2	"LEW.STD.2"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.2"
☐	3	"LEW.STD.3"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.3"
☐	4	"LEW.STD.4"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.4"
☐	5	"LEW.STD.5"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.5"
☐	6	"WKY.STD.1"				"WKY.STD"	"WKY"	"STD"	"WKY.STD.1"
☐	7	"WKY.STD.2"				"WKY.STD"	"WKY"	"STD"	"WKY.STD.2"
☐	8	"WKY.STD.3"				"WKY.STD"	"WKY"	"STD"	"WKY.STD.3"
☐	9	"WKY.STD.4"				"WKY.STD"	"WKY"	"STD"	"WKY.STD.4"
☐	10	"WKY.STD.5"	x	x	x	"WKY.STD"	"WKY"	"STD"	"WKY.STD.5"
☐	11	"LEW.CAF.1"				"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.1"
☐	12	"LEW.CAF.2"			x	"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.2"
☐	13	"LEW.CAF.3"				"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.3"
☐	14	"LEW.CAF.4"		x		"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.4"
☐	15	"LEW.CAF.5"				"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.5"
☐	16	"WKY.CAF.1"			x	"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.1"
☐	17	"WKY.CAF.2"				"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.2"
☐	18	"WKY.CAF.3"			x	"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.3"
☐	19	"WKY.CAF.4"				"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.4"
☐	20	"WKY.CAF.5"				"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.5"

Figure 1: Detección de outliers de los datos sin procesar (extraído del archivo 'index.html' producido por el paquete 'arrayQualityMetrics')

Con el paquete `ArrayQualityMetrics` de Bioconductor analizamos la presencia de *outliers* mediante diferentes aproximaciones (ver Figura 1). Como resultado obtenemos 3 muestras que superan el *threshold* para los MA plots y una muestra que supera el *threshold* de intensidad de los datos representado en el *boxplot*. Dado que estas muestras solamente presentan un asterisco, las mantedremos sin considerarlas un problema potencial. Sin embargo, la muestra `WKY.STD.5` debería ser revisada tras la normalización y probablemente eliminada para mejorar la calidad de los datos dado que presenta asterisco para las tres metodologías analizadas.

En el gráfico de la Figura 2 se han representado las dos primeras componentes del PCA (Principal Component Analysis) que explicarían el 46% total de la variabilidad de las muestras. Observamos que las réplicas se encuentran agrupadas por los grupos determinados por la cepa de rata y la dieta administrada, con la excepción del grupo `WKY.STD` entre las cuales recordemos se encuentra el potencial *outlier*, situado arriba-derecha de la gráfica. La variabilidad de la primera componente tiene una contribución muy alta de la cepa

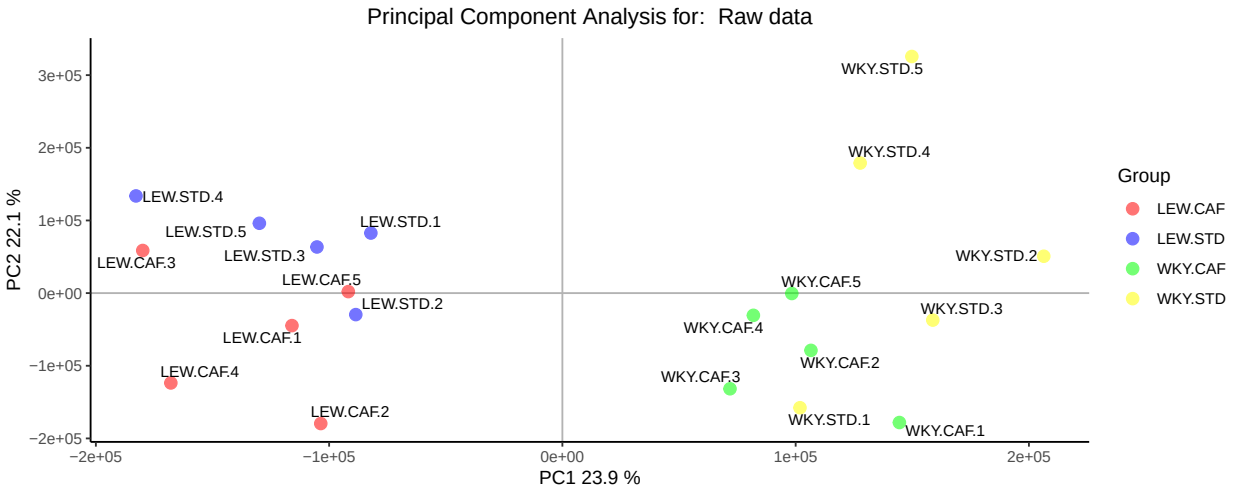

Figure 2: Visualización de las dos primeras componentes principales del PCA para los datos sin procesar

WKY, que como podemos observar se encuentran todas las muestras a la derecha de la gráfica, mientras que la cepa LEW la encontramos a la izquierda. Respecto a la contribución de la segunda componente, la contribución no está tan dividida, pero observamos que hay una tendencia a que las muestras de la dieta STD se encuentran con valores positivos más elevados que los correspondientes a la dieta CAF que tienden a encontrarse en el inferior de la gráfica.

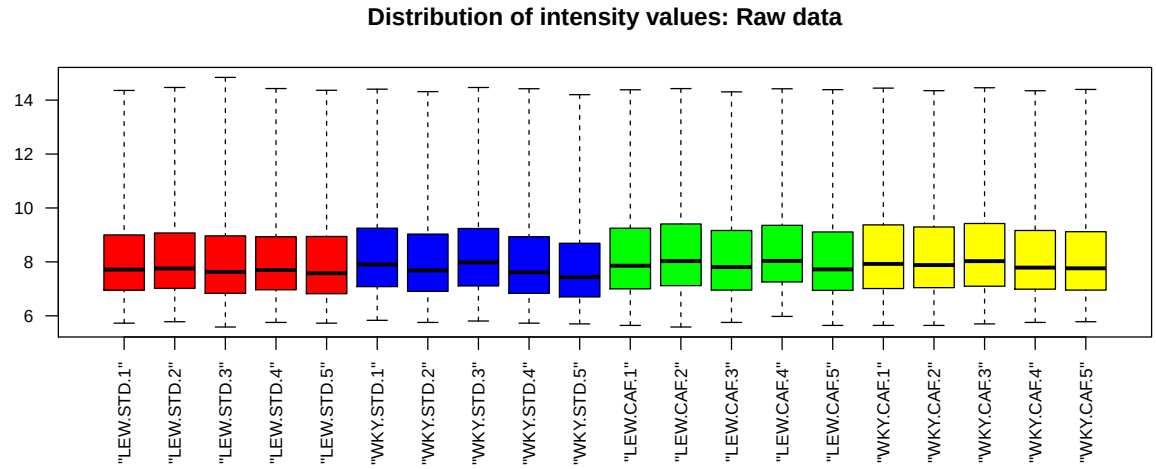

Figure 3: Diagrama de caja para intensidades de matrices de los datos sin procesar

En el diagrama de cajas se representa la distribución de la intesidad de los arrays (ver Figura 3). Podemos observar que la media es bastante regular entre las réplicas y entre los grupos. Sin embargo, observamos una mayor variabilidad en el caso de las réplicas pertenecientes a `WKY.STD` entre los cuales se encuentra el potencial *outlier*.

Observamos que la señal de todos los arrays (ver Figura 4) sigue una distribución muy parecida con los datos crudos, no hay aparentemente ninguna alteración o diferencia.

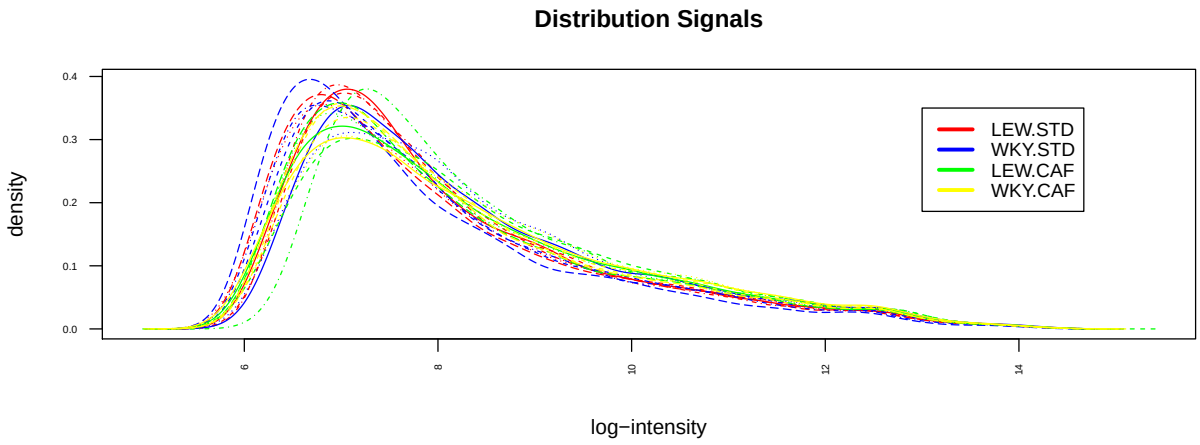

Figure 4: Histograma de la señal de los arrays correspondientes a los datos sin procesar

Normalización

Antes de inicar el análisis, realizamos la normalización de los datos para poder compararlos entre ellos y así reducir la variabilidad de las muestras no debidas a razones biológicas. Con la normalización, observaremos si reducimos el *outlier* que nos marcaba el apartado anterior o bien si debemos eliminarlo. Realizamos la normalización con la función `rma` del paquete `oligo`.

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

Tras la normalización del archivo de datos con código GSE85167, se observan 29214 genes y 20 muestras. Vemos que se ha reducido bastante el número de genes.

Control de calidad de los datos normalizados

Gracias a la normalización de los datos observamos que la muestra `WKY.STD.5` que era un potencial *outlier* se ha corregido (ver Figura 5). Aún así, se mantienen un asterisco en dos muestras de este mismo grupo que como habíamos visto presentaba mayores variaciones.

En el diagrama de dispersión de las dos primeras componentes principales realizado usando esta vez los datos normalizados (ver Figura 6), es importante destacar que ha aumentado ligeramente (*i.e.*, 49.7%) la variabilidad explicada por ambas componentes respecto al PCA realizado con los datos crudos (ver Figura 2), aunque la proporción ha variado notablemente, siendo la PC1 la que contribuye en su mayor parte. De la misma forma que en la Figura 2, las muestras quedan separadas en la PC1 de acuerdo al factor cepa de rata, situando el nivel WKY a la derecha y LEW a la izquierda. Sin embargo, resulta difícil en esta ocasión establecer una distribución diferenciada de las muestras de acuerdo al factor de la dieta, lo que provoca que los grupos de dieta no se encuentren agrupados como en la Figura 2.

El diagrama de cajas (ver Figura 7) muestra la distribución de las intesidades normalizadas de las muestras. Nuevamente, dado que todas las cajas presentan un aspecto homogéneo, podemos determinar que la normalización ha permitido corregir las pequeñas variaciones que existían, especialmente en el grupo `WKY.STD`.

	array	sampleNames	*1	*2	*3	X.Group.	X.Cepa.	X.Dieta.	X.ShortName.
☐	1	"LEW.STD.1"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.1"
☐	2	"LEW.STD.2"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.2"
☐	3	"LEW.STD.3"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.3"
☐	4	"LEW.STD.4"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.4"
☐	5	"LEW.STD.5"				"LEW.STD"	"LEW"	"STD"	"LEW.STD.5"
☐	6	"WKY.STD.1"				"WKY.STD"	"WKY"	"STD"	"WKY.STD.1"
☐	7	"WKY.STD.2"				"WKY.STD"	"WKY"	"STD"	"WKY.STD.2"
☐	8	"WKY.STD.3"		x		"WKY.STD"	"WKY"	"STD"	"WKY.STD.3"
☐	9	"WKY.STD.4"		x		"WKY.STD"	"WKY"	"STD"	"WKY.STD.4"
☐	10	"WKY.STD.5"				"WKY.STD"	"WKY"	"STD"	"WKY.STD.5"
☐	11	"LEW.CAF.1"				"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.1"
☐	12	"LEW.CAF.2"				"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.2"
☐	13	"LEW.CAF.3"				"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.3"
☐	14	"LEW.CAF.4"				"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.4"
☐	15	"LEW.CAF.5"				"LEW.CAF"	"LEW"	"CAF"	"LEW.CAF.5"
☐	16	"WKY.CAF.1"				"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.1"
☐	17	"WKY.CAF.2"				"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.2"
☐	18	"WKY.CAF.3"				"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.3"
☐	19	"WKY.CAF.4"				"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.4"
☐	20	"WKY.CAF.5"				"WKY.CAF"	"WKY"	"CAF"	"WKY.CAF.5"

Figure 5: Detección de outliers de los datos normalizados (extraído del archivo 'index.html' producido por el paquete 'arrayQualityMetrics')

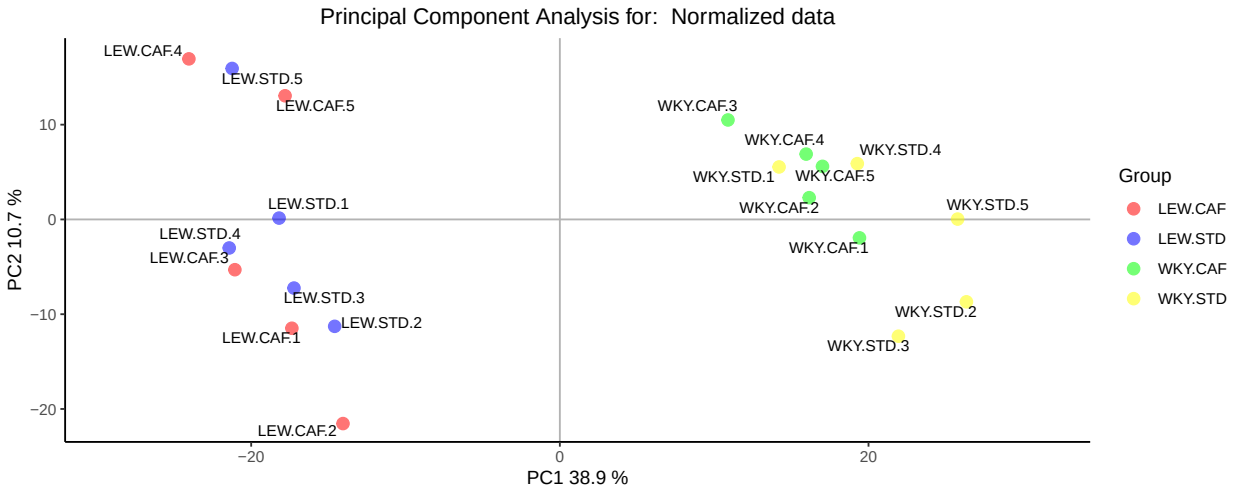

Figure 6: Visualización de las dos primeras componentes principales del PCA para los datos normalizados

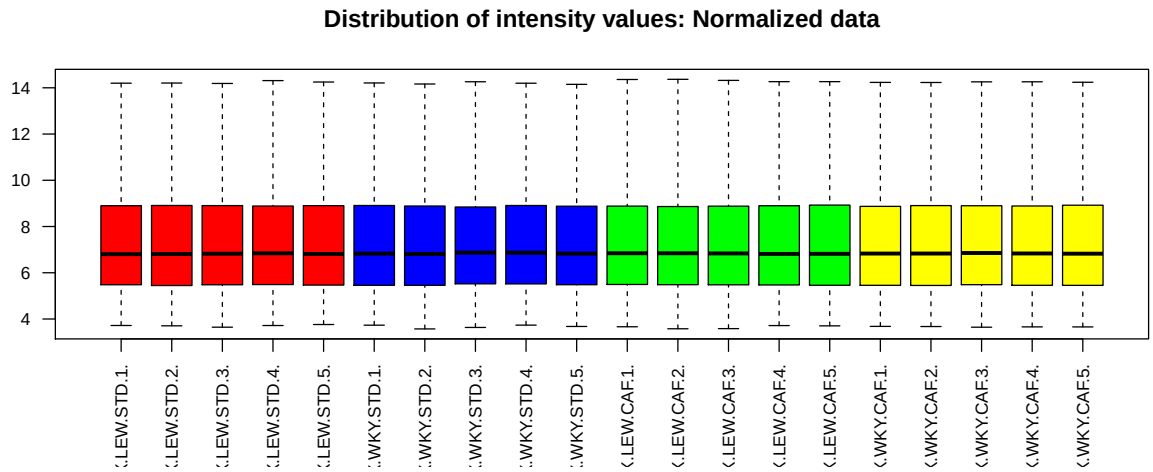

Figure 7: Diagrama de caja para intensidades de matrices de los datos normalizados

Filtraje no específico

Con el fin de analizar pequeñas variaciones en la expresión génica por causas no biológicas (*i.e.*, investigador-técnico, fecha del procesamiento, reactivos), realizamos un *Principal variation component analysis* (PVCA).

En primer lugar, comprobamos que todos los microarrays se realizaron en la misma fecha: 2012-08-30

- ***Principal variation component analysis* (PVCA)**

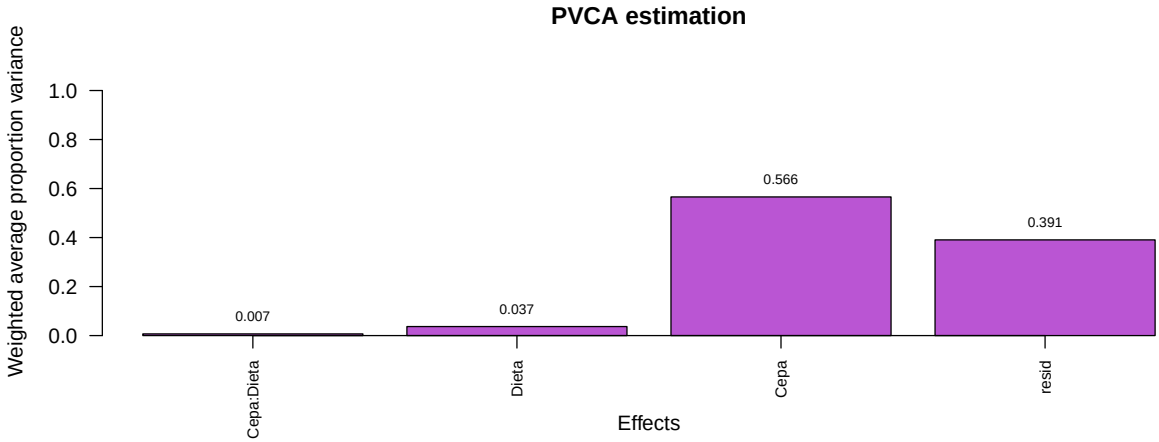

Figure 8: Importancia relativa de los diferentes factores (Cepa, Dieta e interacción) que afectan la expresión génica

El histograma (ver Figura 8) muestra una barra por cada fuente de variación del análisis (*i.e.*, cepa, dieta e interacción). Observamos que la principal fuente de variación es la condición de la cepa de rata que contribuye con un 56.6% a la variación de los datos causada por el factor lote, como se vio en las anteriores gráficas de PCA (ver Figuras 2 y 6). La dieta en cambio tiene una contribución mínima.

Identificación de genes diferencialmente expresados

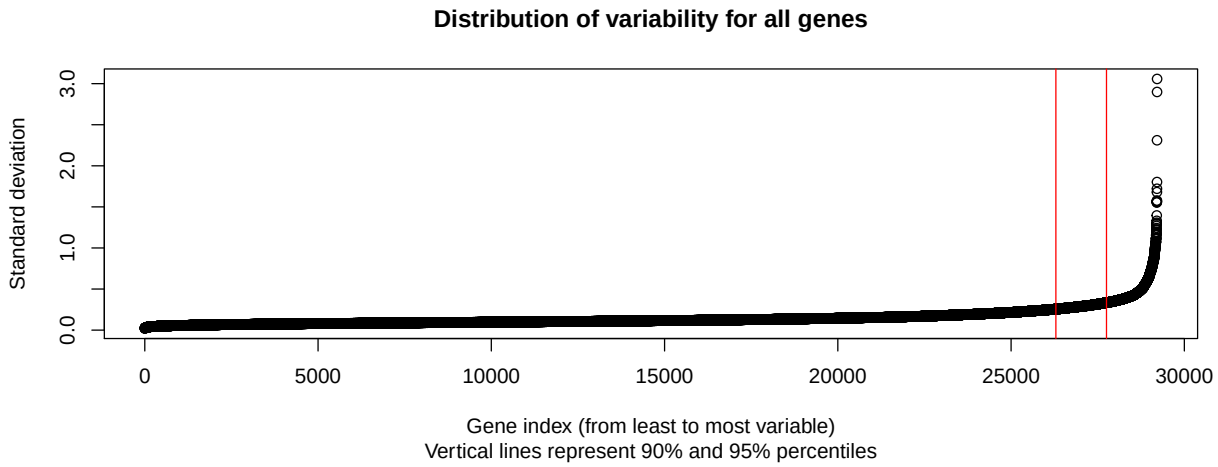

Figure 9: Desviación estándard para todos los genes de todas las muestras ordenados de menor a mayor

Se considera un gen diferencialmente expresado si hay una cierta diferencia en la varianza entre los grupos de estudio. En el gráfico (ver Figura 9) se traza la variabilidad general de todos los genes, con el fin de tener una idea del porcentaje de genes que presentan variabilidad no atribuible a una variación aleatoria. Por ello, en este gráfico se representa la desviación estándard de todos los genes ordenada de forma ascendente. Los genes más variables son los que se encuentran con una desviación estándar superior al 90-95%, es decir, todos aquellos genes a la derecha de las líneas verticales rojas, en este caso, corresponde a algo menos de 5000 genes.

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

```
## $numDupsRemoved
## [1] 1030
##
## $numLowVar
## [1] 13393
##
## $numRemoved.ENTREZID
## [1] 10327
```

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

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 4464 features, 20 samples
##   element names: exprs
## protocolData
##   rowNames: "LEW.STD.1" "LEW.STD.2" ... "WKY.CAF.5" (20 total)
##   varLabels: exprs dates
##   varMetadata: labelDescription channel
## phenoData
##   rowNames: 1 2 ... 20 (20 total)
##   varLabels: FileName Group ... ShortName (5 total)
```

```
##   varMetadata: labelDescription channel
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: ragene10sttranscriptcluster.db
```

En consecuencia, resulta importante filtrar aquellos genes cuya variabilidad en la expresión génica es atribuible a la aleatoriedad y no como resultado del diseño experimental. Para ello, descargamos la base de datos de anotaciones en función del tipo de microarray empleado. Por lo tanto, dado que se ha empleado el microarray *Affymetrix Rat Gene 1.0 ST Array*, se ha buscado en la web de Bioconductor la base de datos correspondiente, `ragene10sttranscriptcluster.db`, en este caso. A continuación, se han filtrado los genes con la función `nsFilter` y se han obtenido 4464 genes con potencial para ser diferencialmente expresados, eliminando aquellos que presentaban variaciones aleatorias.

- **Matriz de diseño**

La matriz de diseño permite mostrar la asignación de cada muestra a un grupo experimental. En este caso tenemos una matriz de 20 filas, una por muestra, y 4 columnas, una por grupo. Cada fila contiene un 1 en la columna del grupo al que pertenece la muestra y un 0 en los demás.

```
##    LEW.STD WKY.STD LEW.CAF WKY.CAF
## 1        0       1       0       0
## 2        0       1       0       0
## 3        0       1       0       0
## 4        0       1       0       0
## 5        0       1       0       0
## 6        0       0       0       1
## 7        0       0       0       1
## 8        0       0       0       1
## 9        0       0       0       1
## 10       0       0       0       1
## 11       1       0       0       0
## 12       1       0       0       0
## 13       1       0       0       0
## 14       1       0       0       0
## 15       1       0       0       0
## 16       0       0       1       0
## 17       0       0       1       0
## 18       0       0       1       0
## 19       0       0       1       0
## 20       0       0       1       0
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$Group
## [1] "contr.treatment"
```

- **Matriz de contraste**

En este estudio el objetivo principal es analizar el efecto de la dieta en la expresión génica por cada cepa de rata (*i.e.*, WKY.STD vs WKY.CAF – LEW.STD vs LEW.CAF), así como la interacción entre la dieta y la cepa. Por lo tanto, la matriz de contraste quedaría del siguiente modo:

```
cont.matrix <- makeContrasts (STDvsCAF.LEW = LEW.STD - LEW.CAF,
                              STDvsCAF.WKY = WKY.STD - WKY.CAF,
                              INT = (LEW.STD - LEW.CAF) - (WKY.STD - WKY.CAF),
                              levels=designMat)
print(cont.matrix)
```

```
##          Contrasts
## Levels    STDvsCAF.LEW STDvsCAF.WKY INT
##   LEW.STD            1            0   1
##   WKY.STD            0            1  -1
##   LEW.CAF           -1            0  -1
##   WKY.CAF            0           -1   1
```

- **Estimación del modelo y selección de genes diferenciados**

Realizamos los tests de significancia con el paquete `limma` para seleccionar aquellos genes diferencialmente expresados, definidos en las anteriores comparaciones. Se realiza también el ajuste del p-valor con el método Benjamini and Hochberg con el fin de evitar la aparición de un alto número de falsos positivos debido al alto número de cálculos de contrastes simultáneos en los mismos datos.

```
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"
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

El volcano plot (ver Figura 10) permite visualizar de forma gráfica los genes que presentan expresión diferenciada para cada una de las comparaciones. En estos gráficos se muestra el fold-change, es decir, la escala logarítmica del efecto biológico y si éste es significativo en función del p-valor. Además, se destaca el nombre de los 5 genes más significativos para cada una de las comparaciones. Observamos que los volcano plot son muy equivalentes para la comparación del efecto de la dieta administrada en cada una de las cepas. En cambio, el volcano plot para la interacción, presenta valores muy bajos para el logaritmo del p-valor y muy agrupados entorno a 0 para el fold-change.

- **Lista de genes diferencialmente expresados**

Con la función `topTable` del paquete `limma` que permite obtener una lista ordenada según el p-valor del contraste ordenada de forma ascendente.

a) LEW.STD vs LEW.CAF: analizar el efecto de la dieta en la expresión génica para la cepa LEW

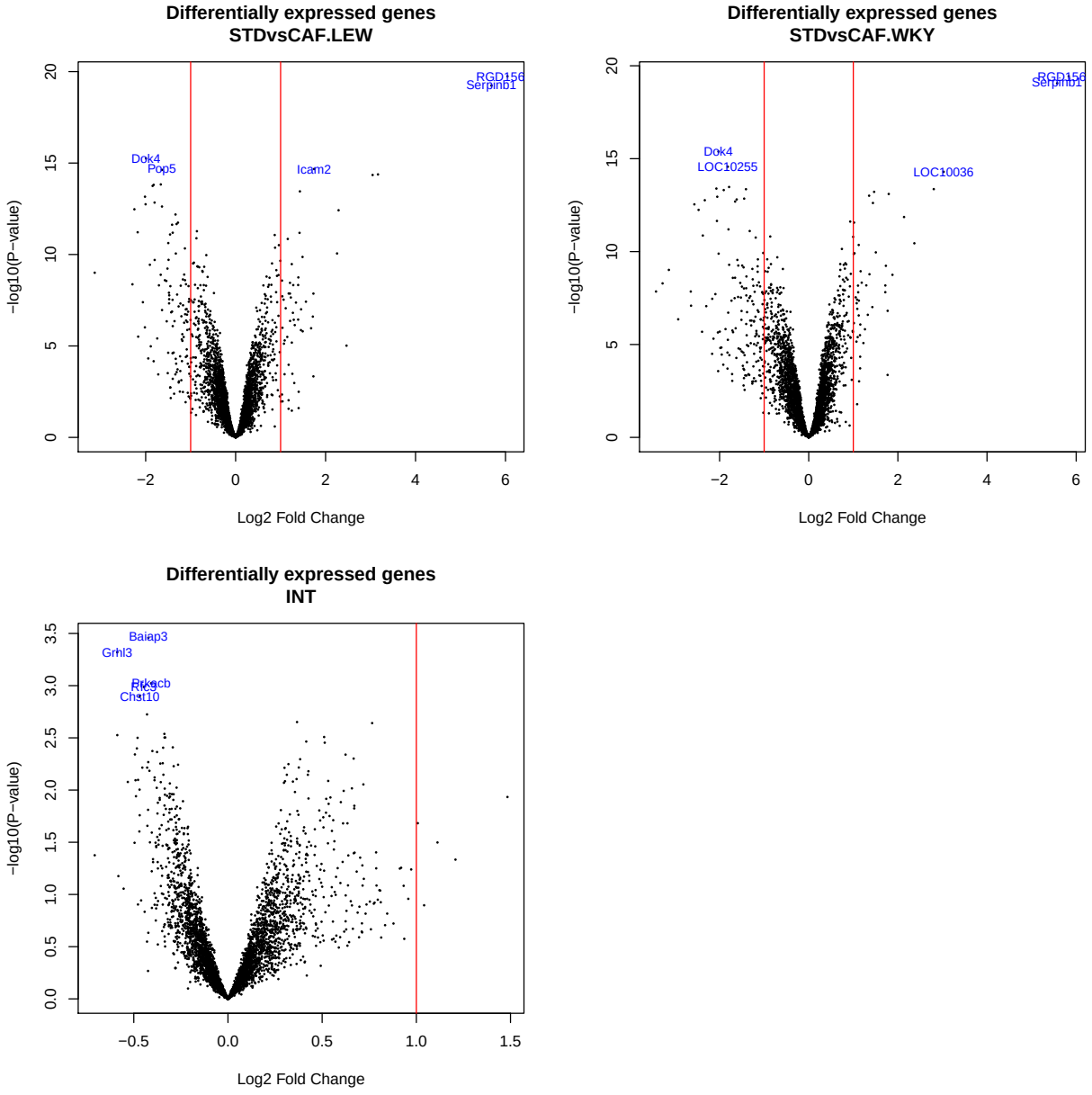

Figure 10: Gráfico volcán para las diferentes comparaciones considerando la cepa de rata y la dieta administrada

Table 2: Genes que modifican su expresión en función de la dieta administrada en la cepa de rata LEW (tabla ordenada según el p-valor ajustado)

	logFC	AveExpr	t	P.Value	adj.P.Val	B
10908861	6.043696	10.166702	38.32264	1.810911e-20	8.083905e-17	35.82878
10763351	5.687649	7.605733	36.16993	5.785832e-20	1.291398e-16	34.87237
10805976	-1.993683	8.470568	-22.80078	5.599568e-16	8.332158e-13	26.64241

b) WKY.STD vs WKY.CAF: analizar el efecto de la dieta en la expresión génica para la cepa WKY

Table 3: Genes que modifican su expresión en función de la dieta administrada en la cepa de rata WKY (tabla ordenada según el p-valor ajustado)

	logFC	AveExpr	t	P.Value	adj.P.Val	B
10908861	5.833674	10.166702	36.99091	3.686532e-20	1.645668e-16	35.32694
10763351	5.572807	7.605733	35.43961	8.712811e-20	1.944700e-16	34.60361
10805976	-2.024637	8.470568	-23.15478	4.136852e-16	6.155635e-13	26.95194

c) INT: interacción entre la dieta administrada y la cepa de rata

Table 4: Genes que modifican su expresión en función de la dieta administrada y de la cepas de rata (tabla ordenada según el p-valor ajustado)

	logFC	AveExpr	t	P.Value	adj.P.Val	B
10741373	-0.42248	8.89352	-4.28583	0.00035	0.73918	-2.81147
10880627	-0.58788	7.68335	-4.15012	0.00048	0.73918	-2.88581
10940456	-0.40789	10.26574	-3.85978	0.00095	0.73918	-3.04984

Las tablas anteriores (Tablas 2, 3 y 4) muestran los genes que en el contraste de significancia presentan un p-valor ajustado menor para cada una de las comparaciones establecidas. La primera columna de estas tablas indica el código del gen de acuerdo con el distribuidor del microarray (Affimetrix en este caso). Destacamos dos aspectos de las tablas, en primer lugar, las tablas 2 y 3 presentan los mismo genes para el efecto de la dieta en cada una de las cepas, lo cual nos indica que estos genes son relevantes dado que se han expresado de forma diferencial en dos cepas distintas. En segundo lugar, vemos que la interacción entre cepa y dieta muestra p-valores ajustados muy altos, lo que nos llevará a eliminar esta comparación del resto del estudio, ya que no son significativos.

Anotación de los resultados

A partir de la base de datos de anotaciones correspondiente al microarray que se ha usado (*i.e.*, `ragene10sttranscriptcluster.db`), podremos relacionar el código comercial de la primera columna con identificadores estándars de los nombres de los genes correspondientes (*i.e.*, Gene Symbol or Entrez gene identifier).

De modo que volvemos a mostrar las tablas anteriores, esta vez para los 10 primeros genes de cada comparación y añadiendo los códigos estándard:

a) LEW.STD vs LEW.CAF: analizar el efecto de la dieta en la expresión génica para la cepa LEW

Table 5: Genes que modifican su expresión en func
administrada en la cepa de rata LEW (tabla ord
p-valor ajustado)

PROBEID	SYMBOL	ENTREZID	GENENAME
10908861	RGD1562690	500965	similar to L-lactate dehydrogenase A chain (LDH-A) (LDH muscle subunit) (
10763351	Serpinb11	304689	serpin family B member 11
10805976	Dok4	361364	docking protein 4
10748306	Icam2	360647	intercellular adhesion molecule 2
10762717	Pop5	117241	POP5 homolog, ribonuclease P/MRP subunit
10866076	Klri2	503650	killer cell lectin-like receptor family I member 2
10936923	LOC100363171	100363171	histone variant H2al2-like
10729409	Tjp2	115769	tight junction protein 2
10817987	LOC102551064	102551064	glutamic acid-rich protein-like
10827582	Tmlhe	170898	trimethyllysine hydroxylase, epsilon

b) WKY.STD vs WKY.CAF: analizar el efecto de la dieta en la expresión génica para la cepa WKY

Table 6: Genes que modifican su expresión en func
administrada en la cepa de rata WKY (tabla ord
p-valor ajustado)

PROBEID	SYMBOL	ENTREZID	GENENAME
10908861	RGD1562690	500965	similar to L-lactate dehydrogenase A chain (LDH-A) (LDH muscle subunit) (
10763351	Serpinb11	304689	serpin family B member 11
10805976	Dok4	361364	docking protein 4
10817987	LOC102551064	102551064	glutamic acid-rich protein-like
10936923	LOC100363171	100363171	histone variant H2al2-like
10762717	Pop5	117241	POP5 homolog, ribonuclease P/MRP subunit
10827582	Tmlhe	170898	trimethyllysine hydroxylase, epsilon
10845143	Neb	311029	nebulin
10866076	Klri2	503650	killer cell lectin-like receptor family I member 2
10876769	Abca1	313210	ATP binding cassette subfamily A member 1

c) INT: interacción entre la dieta administrada y la cepa de rata

Table 7: Genes que modifican su expresión en función de la dieta administrada y de la cepas de rata (tabla ordenada según el p-valor ajustado)

PROBEID	SYMBOL	ENTREZID	GENENAME	logFC
10703930	Isoc2b	361501	isochorismatase domain containing 2b	0.36424
10708015	Isg20	293052	interferon stimulated exonuclease gene 20	0.66725

PROBEID	SYMBOL	ENTREZID	GENENAME	logFC
10708587	LOC103691190	103691190	uncharacterized LOC103691190	0.53224
10709200	Pde2a	81743	phosphodiesterase 2A	-0.34282
10711268	Itgam	25021	integrin subunit alpha M	0.62427
10718504	Lilrc2	690906	leukocyte immunoglobulin-like receptor, subfamily C, member 2	0.61431
10718954	Lilrb4	292594	leukocyte immunoglobulin like receptor B4	0.29724
10719148	RGD1564801	499084	similar to hepatic multiple inositol polyphosphate phosphatase	-0.49090
10722992	Anpep	81641	alanyl aminopeptidase, membrane	0.65801
10731643	Ppl	302934	periplakin	-0.40127

Comparación entre comparaciones

```
##        STDvsCAF.LEW STDvsCAF.WKY  INT
## Down            114          161    0
## NotSig         4295         4257 4464
## Up               55           46    0
```

En la tabla se representan en las columnas el nº de comparaciones, mientras que las filas hacen referencia a la expresión genética, siendo *up-regulated*, *down-regulated* o sin diferencia significativa estableciendo un cutoff de 0.1 del p-value y un log2-fold-change mínimo de 1. Aplicando este filtro obtenemos un conjunto de genes *down-regulated* alto en los grupos STDvCAF para ambas cepas de rata, mientras que genes *up-regulated* se encuentran la mitad o la tercera parte de los anteriores. Sin embargo, sorprenden los datos del grupo INT en los que no tenemos ningún gen significativo, es un dato que ya preveíamos de las tablas anteriores debido a los valores altos de p-valor ajustado. De modo que eliminiamos esta comparación del estudio.

- **Diagrama de Venn**

En el diagrama de Venn (ver Figura 11) representamos aquellos genes que se encuentran en común entre las 2 comparaciones anteriores de tipo de dieta administrada en función de la cepa de rata. En este caso, vemos que en la comparación de la dieta STDvsCAF para cada una de las cepas, comparten 140 genes *up* o *down-regulated* y que cada una de ellas presenta 29 y 67 genes que se expresan de forma diferenciada entre la cepa LEW y WKY, respectivamente.

- **Heatmap**

En el heatmap (ver Figura 12) observamos los genes expresados significativamente destacados en una gradación de color (azul a rojo para *down-regulated* y *up-regulated*, respectivamente). Los genes seleccionados según el criterio anterior (*i.e.*, FDR < 0.1 and logFC > 1) se encuentran ordenados de acuerdo a su similitud.

Análisis de significación biológica

Una vez obtenemos la lista de genes expresados diferencialmente entre dos condiciones, debemos interpretar su relevancia biológica, es decir, conocer en que rutas metabólicas están implicados para conocer su función. Este análisis lo realizamos con la ayuda del paquete `clusterProfiler`que nos permite conocer dichas rutas de acuerdo al `Entrez ID` de cada gen.

En este caso, se recomienda según el protocolo de la asignatura [2], relajar los criterios de selección del p-valor y del fold change para seleccionar la lista de genes representativos. Sin embargo, dado que los p-valores ajustados son muy bajos, se ha establecido un *cutoff* del p-valor ajustado < 0.1, obteniendo el siguiente número de genes significativos para cada comparación.

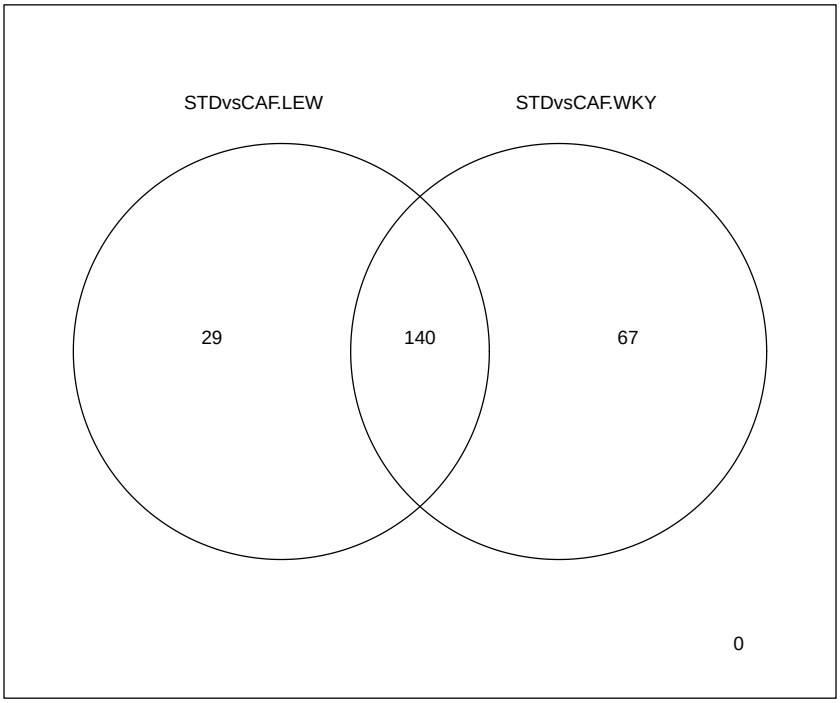

Figure 11: Genes en común entre las 3 comparaciones realizadas previamente (Selección FDR < 0.1 and logFC > 1)

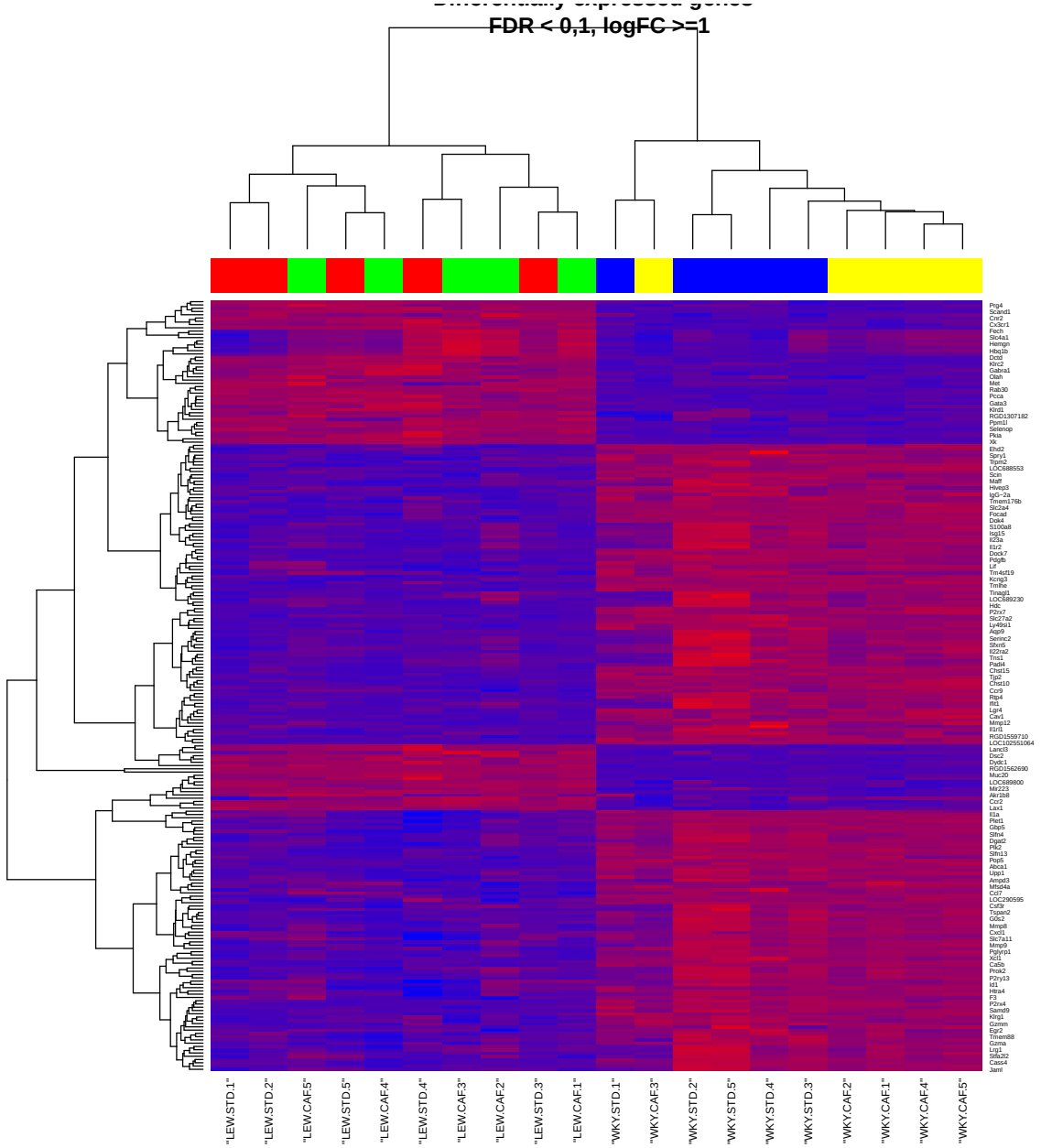

Figure 12: Heatmap de la expresión de genes agrupados por su similitud

```
## STDvsCAF.LEW STDvsCAF.WKY
##         2158         2545
```

- **Barplot para los términos EnrichGO**

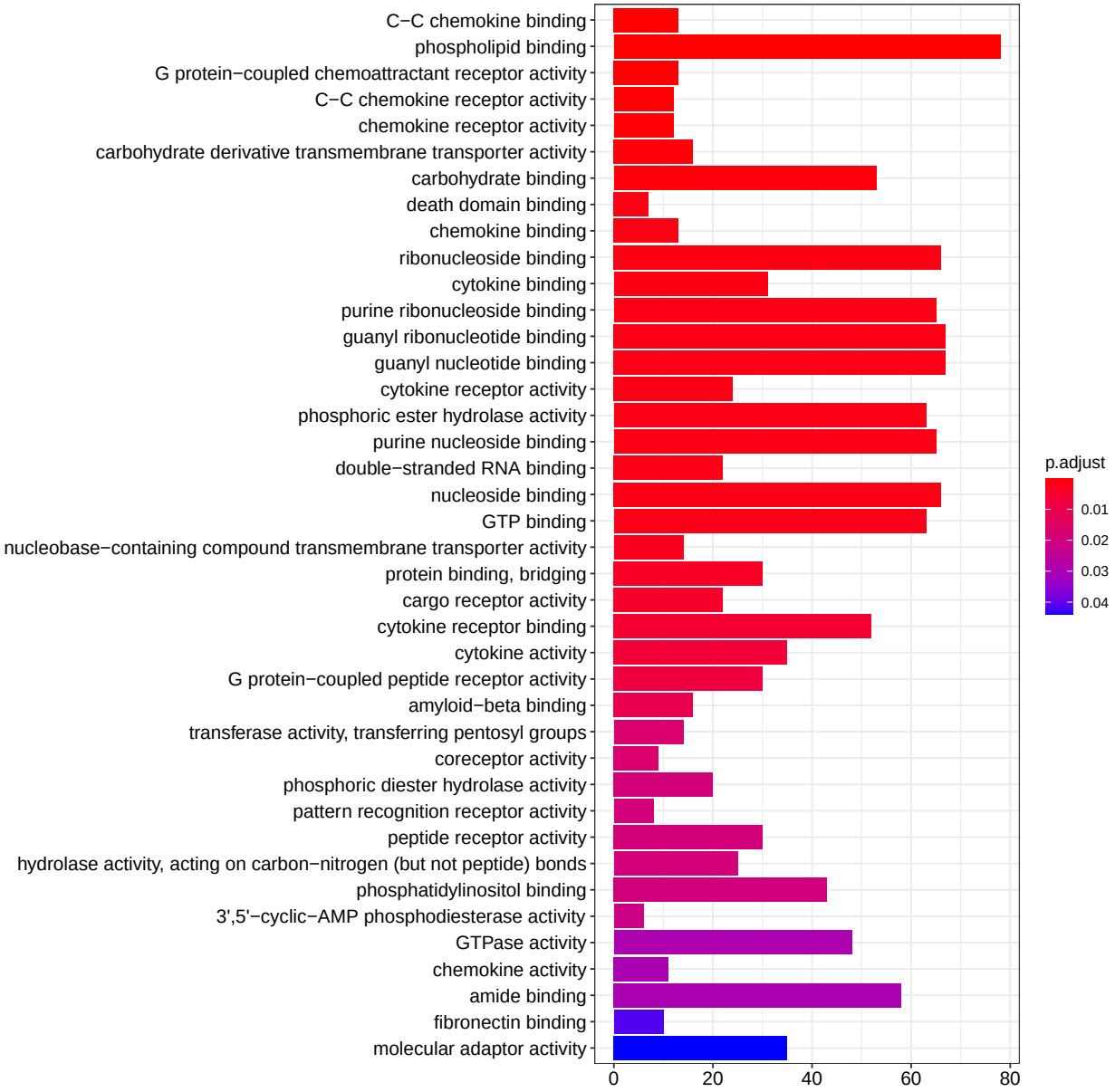

Figure 13: Barplot para los términos de EnrichmentGO para el análisis el efecto de la dieta en la expresión génica para la cepa de rata LEWIS

En el barplot (ver Figura 13) vemos que el 80% de los genes con p-valor significativo de la cepa LEW participan en la unión a fosfolípidos, así como un 65% aprox también se ve implicado en a la unión a carbohidratos y a diferentes nucleótidos y ribonucleótidos. La relación con los fosfolípidos y los carbohidratos podría estar justificada dada la relación de la dieta de cafetería que tiene como objetivo provocar síndrome metabólico y por lo tanto, provoca un aumento del número de lípidos. En el caso de las ratas WTY (ver

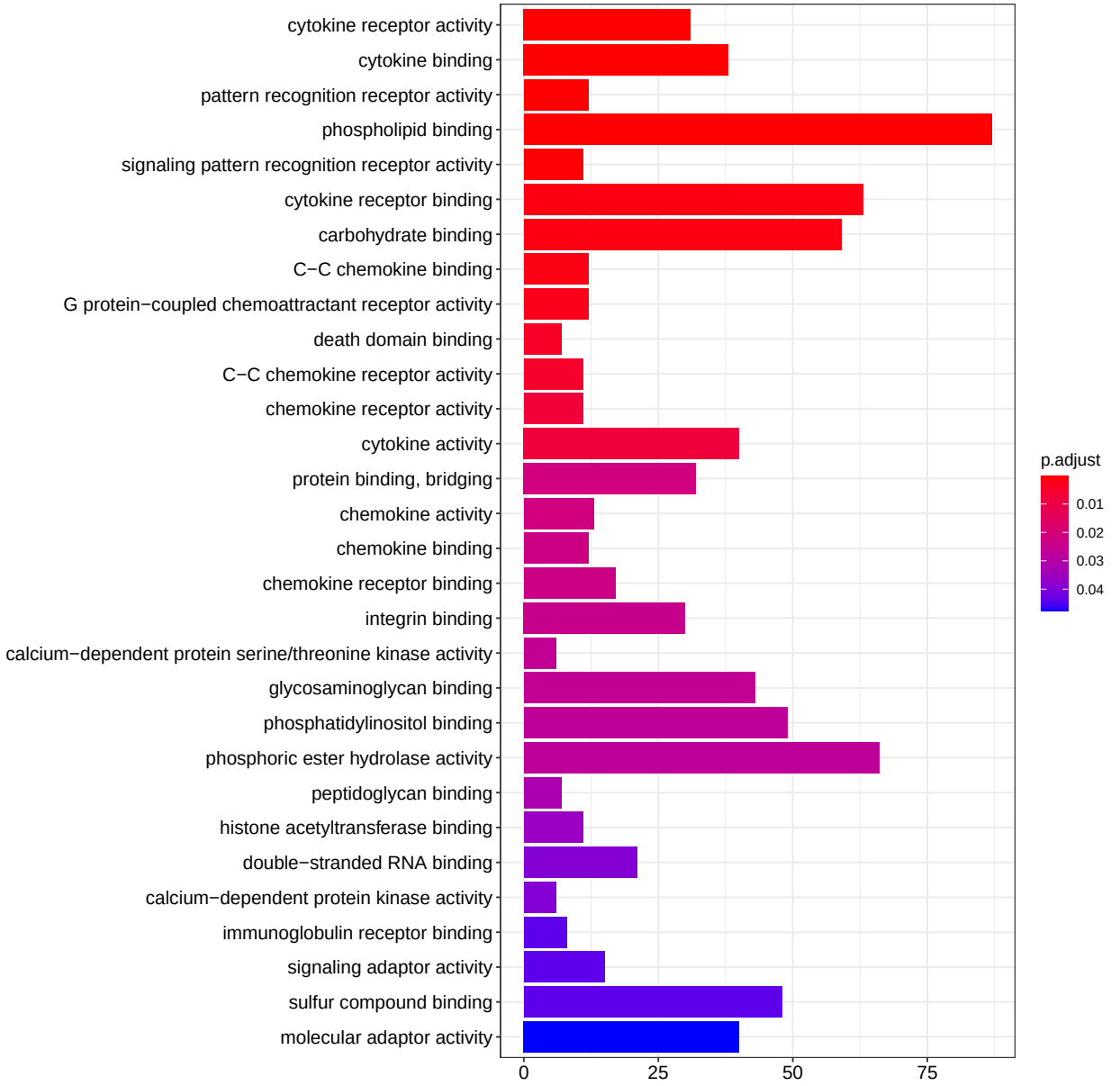

Figure 14: Barplot para los términos de EnrichmentGO para el análisis el efecto de la dieta en la expresión génica para la cepa de rata WISTAR KYOTO

Figura 14) también se visualiza la alta participación de los genes en la unión a fosfolípidos y carbohidratos, sin embargo no se ve una alta implicación en la unión a nucleótidos. Por el contrario, aparece una participación en la enzima hidrolasa ester-fosfórica y en la unión al fosfatidilinositol, otro tipo de lípido.

- **Gene-Concept Network**

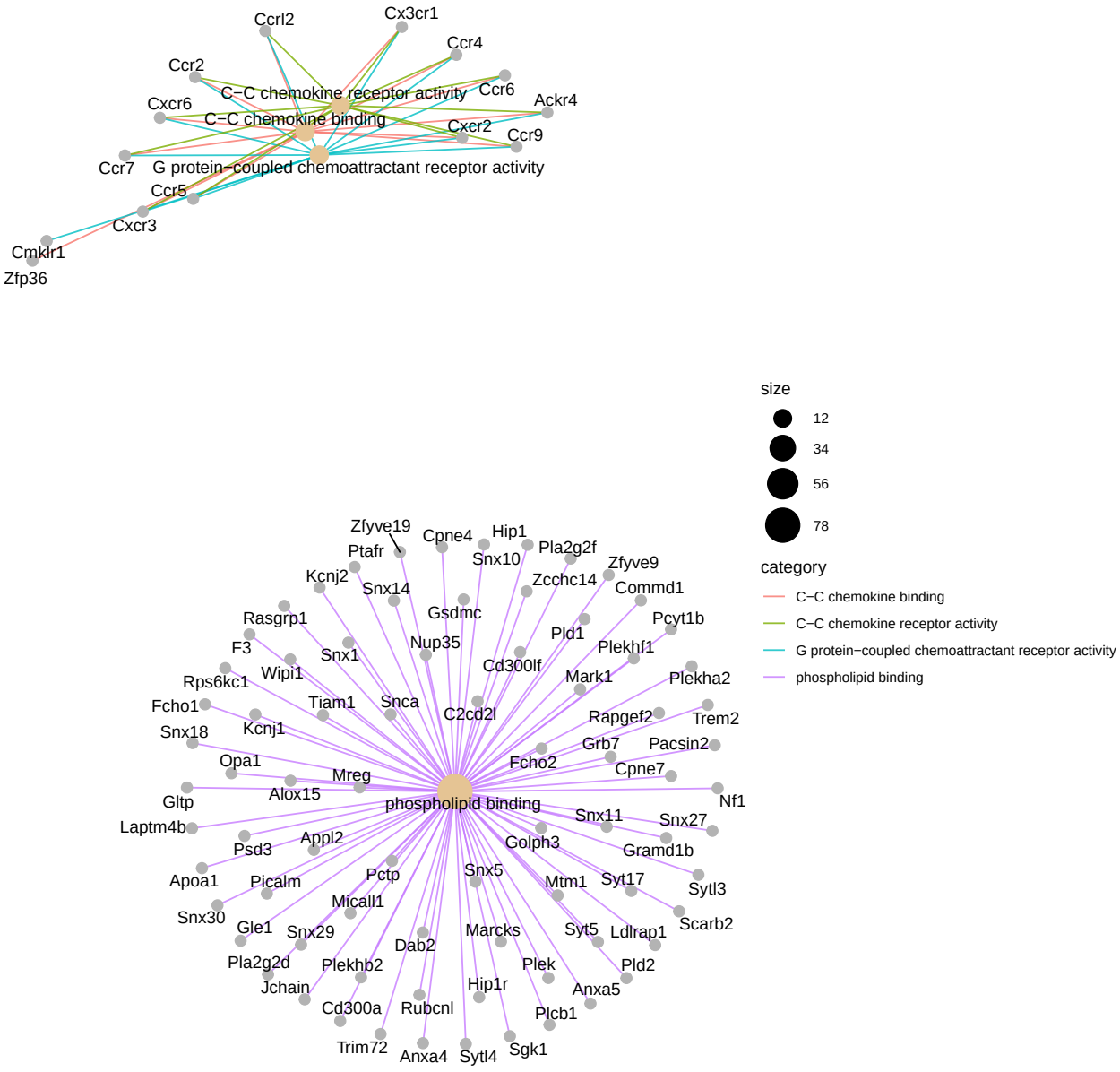

Figure 15: Red para los términos de EnrichmentGO para el análisis el efecto de la dieta en la expresión génica para la cepa de rata LEW

Como hemos visto anteriormente, la red para agrupar los términos de Enrichment de los genes significativos (ver Figuras 15 y 16) tiene una alta participación en la unión a fosfolípidos para ambas cepas. Por otro lado, obtenemos otra red en la que se ven implicadas 3 dianas que participan en el sistema immune (*i.e.*, CCR5, una proteína de membrana de la família GPCR y dos quimiocinas).

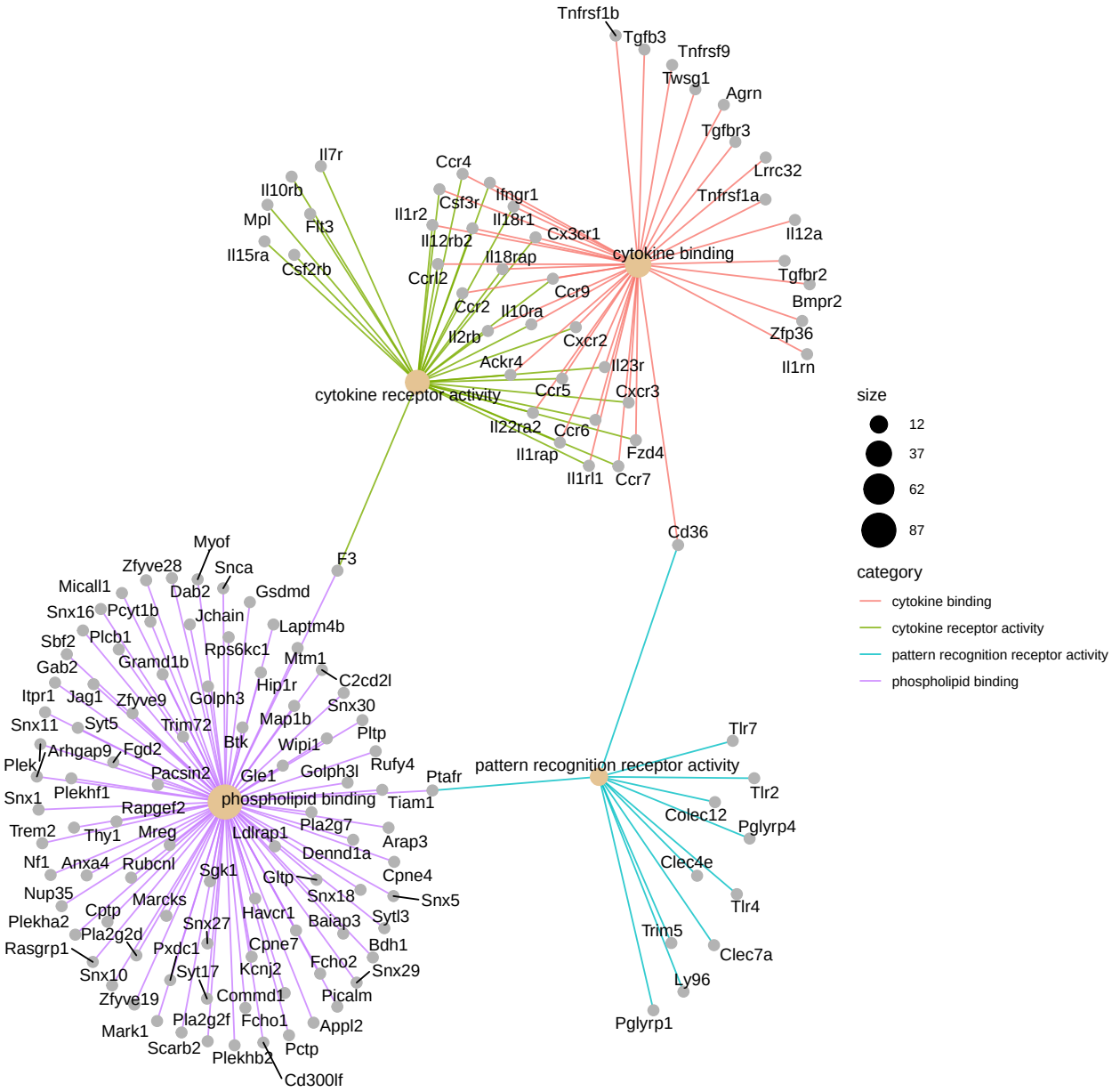

Figure 16: Red para los términos de EnrichmentGO para el análisis el efecto de la dieta en la expresión génica para la cepa de rata WKY

INFORME DEL ANÁLISIS

Identificación de la expresión transcripcional en monocitos mediante el uso de modelos de ratas endogámicas en una dieta de cafetería

Abstract

La obesidad se ha convertido en una enfermedad de alta prevalencia a nivel mundial. El modelo animal nos permite estudiar esta enfermedad a diferentes niveles, por ello, se ha demostrado que una dieta alta en grasas y calorías reproduce los síntomas y factores propios de la obesidad en humanos. En este estudio se emplean dos cepas de rata (Wistar Kyoto (WKY) y Lewis (LEW) dado que presentan diferente respuesta a la dieta de cafetería alterando el transcriptoma de monocitos. Tras el análisis de la expresión de genes, se ha obtenido que hay una transcripción significativa de genes implicados en ruta metabólicas de unión a fosfolípidos y carbohidratos. Estos resultados reflejan la importancia de analizar el background genético como respuesta a diferentes ingestas nutricionales.

Introducción y Objetivo

La obesidad representa un importante problema de salud con una alta tasa de crecimiento en nuestra sociedad. Por esta razón, es necesario disponer de modelos animales que reproduzcan de forma óptima las principales características de la obesidad humana (*i.e.*, aumento de peso, adipogénesis y dislipidemia). Con este fin, en el presente estudio se emplean dos cepas de ratas que presentan dos perfiles genéticos que determinan una respuesta fenotípica y metabólica diferente tras la inducción producida por una dieta obesogénica. De modo que en el caso de las ratas Lewis (LEW) metabolizan preferentemente carbohidratos, mientras que las ratas Wistar Kyoto (WKY) metabolizan los lípidos y además, se observan variaciones en la regulación de la leptina [3]. La dieta de cafetería (CAF) administrada a las ratas es considerada un buen modelo para provocar el síndrome metabólico y las patalogías relacionadas, entre las que se encuentran la obesidad. La dieta de cafetería consiste en la ingesta voluntaria de productos muy calóricos y energéticos.

Los monocitos son células immunitarias de particular interés dado que son circulantes se encuentran expuestas a factores metabólicos y la produccion de citoquinas pro-inflamatorias. En consecuencia, la regulación del perfil de expresión génica de monocitos podría reflejar el estado fisiológico de todo el organismo.

Por lo tanto, el análisis con microarrays en este estudio originalmente presenta dos objetivos:

1) comparar los perfiles de expresión de monocitos entre las ratas WKY y LEW para identificar genes expresados por la adaptación genotípica inducida por la obesidad (dieta de cafetería).
2) analizar el efecto de la dieta (STD vs cafetería) en la expresión genética de cada tipo de rata.

Sin embargo, en nuestro estudio nos hemos centrado en el análisis del segundo objetivo con el fin de simplificar y entender el análisis del microarray.

Materiales y Métodos

- Diseño experimental

Tras un periodo de adaptación cada cepa de rata fue distribuida de forma aleatoria a uno de los dos grupos de dieta administrada durante 7 semanas. La unidad observacional corresponde a los monocitos circulantes aislados de la extracción de sangre procedente de la aorta abdominal tras las 7 semanas de tratamiento.

- Diseño computacional

El microarray empleado en este experimento es del tipo Rat Gene 1.0 ST arrays perteneciente a la casa comercial Affymetrix. El análisis del microarray se ha analizado con la versión 3.6.1 de R (en el artículo se usa la versión 3.4.4) y empleando funciones y paquetes pertenecientes al proyecto Bioconductor. El dataset seleccionado se encuentra identificada con el ID de la base de datos *Gene Expression Omnibus*: GSE85167. El dataset consta de dos factores con dos niveles cada una correspondientes a la cepa de rata (LEW y WKY) y la dieta administrada (STD y CAF). El número de muestras es de 20 repartidas en 4 grupos con 5 réplicas cada uno.

Resultados y Discusión

El análisis del microarray se ha llevado a cabo en primer lugar analizando la calidad de los datos. Como resultado se obtiene un potencial *outlier* en la muestra WKY.CAF.5, pero tras la normalización se ha observado una mejora en la calidad de los datos. A continuación, se ha realizado un filtraje no específico para eliminar los genes que presentaban varianza aleatoria y no como consecuencia del tratamiento del estudio. De este filtro, obtenemos 4464 genes que potencialmente pueden ser significativos. Seguidamente, se ha realizado una matriz de contraste en la que se ha especificado que los grupos a comparar son en función del tipo de dieta para cada una de las cepas (STD vs CAF en LEW y WKY). Como resultado del test de significancia y aplicando cutoff para el fold change y el p-valor ajustado con el método Benjamini and Hochberg. Por último, se han analizado los términos de enrichment de GO database para agrupar los genes significativos en función de la ruta metabólica en la se ven implicada dichos genes.

De acuerdo con los resultados publicados en el artículo orginal, se observó que la respuesta fenotípica en función de la dieta administrada en las ratas LEW resultaron significativas en la expresión de genes de monocitos, obteniendo 228 y 195 transcritos significativos up y down-regulated, respectivamente. Sin embargo, los términos de GO y KEGG no revelaron ningún enrequecimiento relativo a las rutas metabólicas relacionadas con la obesidad. En cambio, la modulación de la expresion de monocitos en ratas WKY obtuvo un cambio significativo obteniendo 483 y 449 transcritos up- y down-regulated.

En nuestro caso, no se han obtenido los mismos valores de transcritos dado que hemos sido más restrictivos aplicando un cutoff de 0.1 del p-value y un log2-fold-change mínimo de 1, pero si que se obtiene que las ratas WKY presentan valores más elevados de transcripts significativos.

No obstante, los genes significativos en cada cepa de rata no coinciden entre el estudio original y el realizado en el presente informe. Esto puede ser debido a múltiples factores como son el paquete empleado o el límite empleado para el p-valor y el fold-change. En todo caso, obtenemos que muchos de los genes siginificativos participan en el metabolismo de fosfolípidos y carbohidratos implicados activamente en la obesidad. También encontramos quimiocinas que participan en el sistema immune, lo cual es implicito en el hecho que estudiamos la expresión de monocitos.

Conclusión

Tras este informe, se ha conseguido analizar un microarray con el fin de valorar la respuesta fenotípica en dos cepas de rata a la dieta de cafetería. Se ha comprendido la importancia de cada paso del análisis y se han obtenido múltiples figuras que nos dan una imagen de la calidad de los datos, de la significancia de los test de comparación y de la participación de genes sigficativos en rutas metabólicas. No obstante, este primer paso de aproximación a un array se mejorará con el análisis de nuevas funciones y argumentos disponibles en Bioconductor. Así como también se requiere una mejor comprensión de toda la información extraída de cada gráfico y un análisis más detallado de cada uno de los genes significativos.

References

[1] Neus Martínez-Micaelo, Noemi González-Abuín, Ximena Terra, Ana Ardévol, Montserrat Pinent, Enrico Petretto, et al. Identification of a nutrient-sensing transcriptional network in monocytes by using inbred rat

models on a cafeteria diet. Disease Models & Mechanisms 2016;9:1231–9. https://doi.org/10.1242/dmm.025528.

[2] Ricardo Gonzalo, Sanchez-Pla A. Statistical analysis of microarray data 2019. https://github.com/ASPteaching/Omics_Data_Analysis-Case_Study_1-Microarrays.

[3] Neus Martínez-Micaelo, Noemi González-Abuín, Ana Ardévol, Montserrat Pinent, Enrico Petretto, Jacques Behmoaras, et al. Leptin signal transduction underlies the differential metabolic response of lew and wky rats to cafeteria diet. Journal of Molecular Endocrinology 2016;56:1–10. https://doi.org/10.1530/JME-15-0089.