Análisis de datos de expresión de la serie GSE113834

Análisis Datos Ómicos - PEC 1

Jorge Vallejo Ortega

31 de October, 2020

Contents

Sumario **1**

Objetivos **1**

Materiales y métodos **2**
Muestras y datos de origen 2
Diseño experimental 2
Procedimiento seguido en el análisis 2

Resultados **18**

Discusión **19**

Apéndice A: Código **19**

Apéndice B: Reproducibilidad **19**

Notas **21**

References **21**

Sumario

Se cree que un gen, o una familia de genes, cuya función afecta a la poliadenilación del ARNm podría estar implicado en el riesgo de sufrir el trastorno de espectro autista (ASD). El examen mediante microarray de genes cuyo ARN tiene patrones de poliadenilación diferentes entre muestras de cerebro humano control, y muestras de personas con ASD, apunta a que genes implicados en el riesgo de ASD presentan estos patrones de poliadenilación diferencial con más frecuencia de lo que sería esperado por azar.

Objetivos

El objetivo, dentro del trabajo original (Parras et al. 2018), es estudiar el acortamiento (deadenilación) de la cola poli-A, provocada por las proteínas CPEB1-4 (*cytoplasmic polyadenylation element binding proteins*) en ARNm de genes asociados al riesgo de desarrollar trastorno del espectro autista ("autism spectrum disorder", ASD).

Mi objetivo en este análisis es, a partir de los datos de expresión detectados por las microarrays, extraer un listado de genes cuyos ARNm contengan una cola poli-A más larga, o más corta, al comparar muestras de sujetos control con muestras de sujetos con ASD. Una vez obtenido el listado, el objetivo final es averiguar, primero, qué términos de la ontología génica están significativamente enriquecidos en esa lista de genes. Y segundo, si nuestra lista de genes contiene más genes asociados al riesgo de ASD de lo que sería esperable por azar.

Materiales y métodos

Muestras y datos de origen

El material biológico de partida fueron muestras post-mórtem de córtex prefrontal de pacientes con trastorno del espectro autista (n = 5) y de controles (n = 4), todos ellos varones de 5-23 años de edad. A partir de las muestras de tejido se extrajo el ARN total. De cada muestra de ARN total se guardó una alíquota ("Input"), y el resto se separó por cromatografía en un fracción enriquecida en ARNm con colas poli-A largas ("Wash") y otra enriquecida en ARNm con colas poli-A cortas ("Eluted"). Las muestras de ARN fueron traducidas a ADNc, y este amplificado.

Las muestras de ADNc fueron hibridadas en arrays GeneChip Human PrimeView (Affymetrix, 901838); estos fueron leídos en un GeneChip Scanner GCS3000 (Affymetrix), y mediante Command Console (Affymetrix) los datos se almacenaron en archivos CEL. Estos archivos CEL base del presente análisis, y los datos necesarios para realizar la anotación de los genes destacados, fueron descargados desde la web de Gene Expression Omnibus (GEO) (Clough and Barrett 2016).

Diseño experimental

El conjunto de 27 ficheros CEL (cada uno procedente de una microarray) se puede clasificar según la condición del sujeto de origen ("Control" o "ASD"), y según el enriquecimiento en diferentes longitudes de colas poli-A en los ARNm ("Input", "Wash", o "Eluted").

Por combinación de ambas variables podemos dividir las muestras en seis grupos experimentales:

Control-Input	ASD-Input
Control-Wash	ASD-Wash
Control-Eluted	ASD-Eluted

Las muestras "Input" contienen el ARN total. Si comparamos las muestras Control-Input con las muestras ASD-Input, comprobaremos si podemos detectar expresión diferencial entre las condiciones Control y ASD.

Las muestras "Wash" están enriquecidas en ARN de cola poli-A corta, y las muestras "Eluted" en ARN de cola poli-A larga. Podemos hacer las siguientes comparaciones:

- Control-Wash vs Control-Eluted: Qué genes se transcriben a ARN de diferentes longitudes de cola poli-A en sujetos Control.
- ASD-Wash vs ASD-Eluted: Qué genes se transcriben a ARN de diferentes longitudes de cola poli-A en sujetos Control.
- Comparando los listados de genes de las dos comparaciones anteriores obtendremos un listado de los genes que se comportan de forma diferente en sujetos Control y sujetos ASD con respecto a las colas poli-A de sus ARN.

Procedimiento seguido en el análisis

Los pasos seguidos para realizar el presente análisis han sido los siguientes:

1. Obtenión de los datos de expresión en bruto.
2. Control de calidad de los datos brutos.
3. Normalización.

4. Control de calidad de los datos normalizados.

5. Filtraje no específico.

6. Identificación de genes diferencialmente expresados.

7. Anotación de los resultados.

8. Comparación entre comparaciones.

9. Análisis del enriquecimiento de rutas.

Obtención de los datos en bruto

Los datos en bruto usados en el análisis ha sido descargados de Gene Expression Omnibus, un repositorio público de datos de genómica funcional, donde los datos están catalogados bajo el código de acceso GSE113834.

Los datos de expresión han sido descargados de dicho repositorio en forma de ficheros CEL.

La relación entre cada fichero y el grupo experimental al que pertenece se ha extraido de la tabla "Samples", accesible en la misma página web que los ficheros CEL.

Los datos de anotación correspondientes al modelo de array usado, están catalogados en el repositorio GEO bajo el código de acceso GPL15207. Dichos datos han sido descargados en forma de fichero de texto, y usados para anotar los datos obtenidos de las sondas de las arrays.

Los diferentes archivos con los datos en bruto han sido manipulados utilizando R para realizar los controles de calidad y los análisis propiamente dichos. El código completo usado puede consultarse en el apéndice A.

Control de calidad de los datos en bruto

Con el control de calidad pretendemos averiguar si los datos de alguna de las muestras presentan defectos o sesgos que desaconsejen usarlos, antes de continuar con el análisis.

En este caso hemos usado examinado los datos de expresión mediante diferenes representaciones gráficas, en busca de anomalías.

Gráficos de densidad

Los gráficos de densidad nos informan acerca de la forma y posición de las señales sin normalizar.

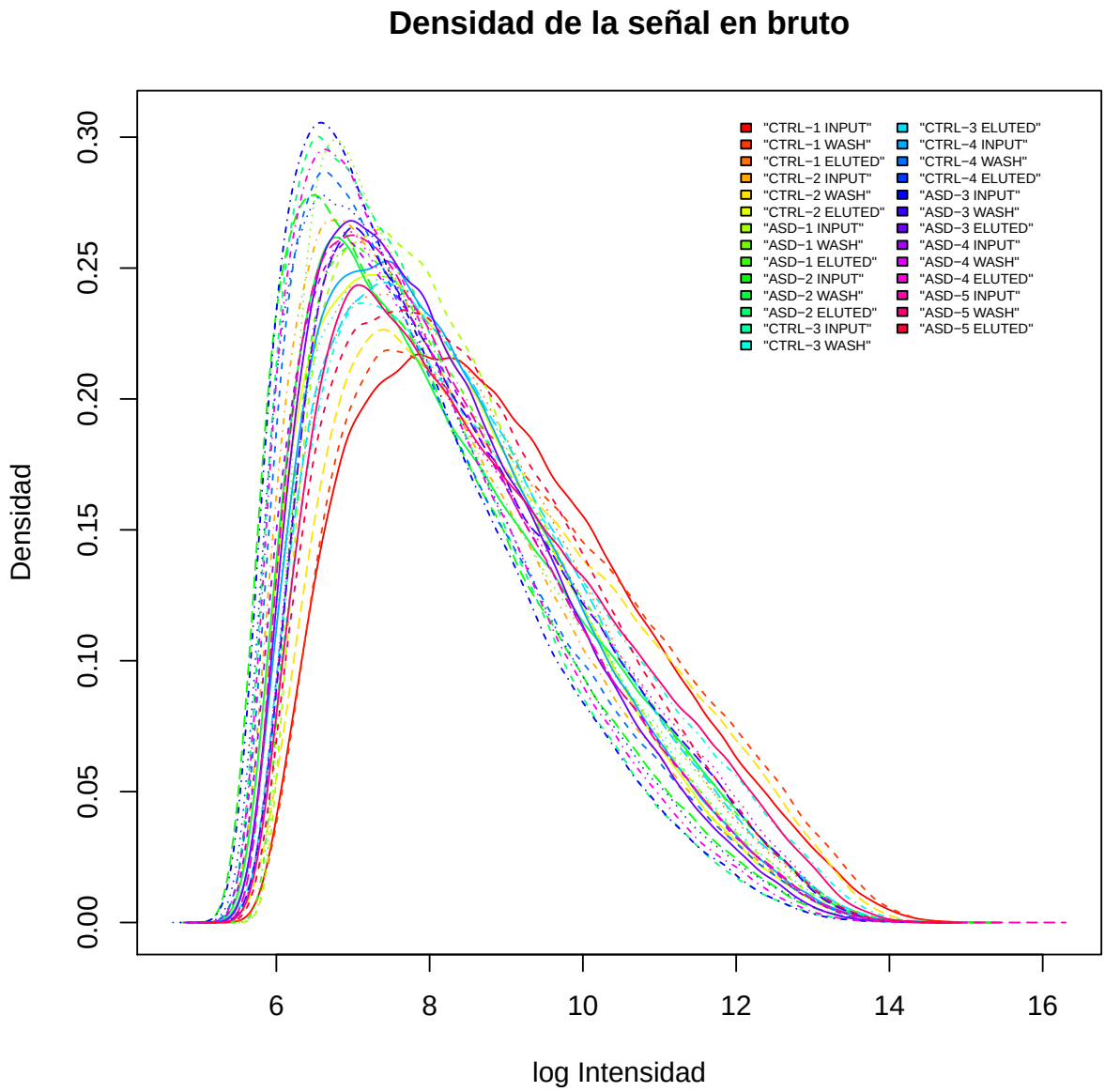


Figure 1: Gráfico de densidad de las señales sin normalizar. Los colores no son significativos.

Este caso vemos que la curva de densidad es similar en todas las muestras, sin mostrar grandes diferencias.

Diagrama de cajas

El gráfico de diagrama de cajas nos permite comparar la distribución de la intensidad entre las diferentes muestras.

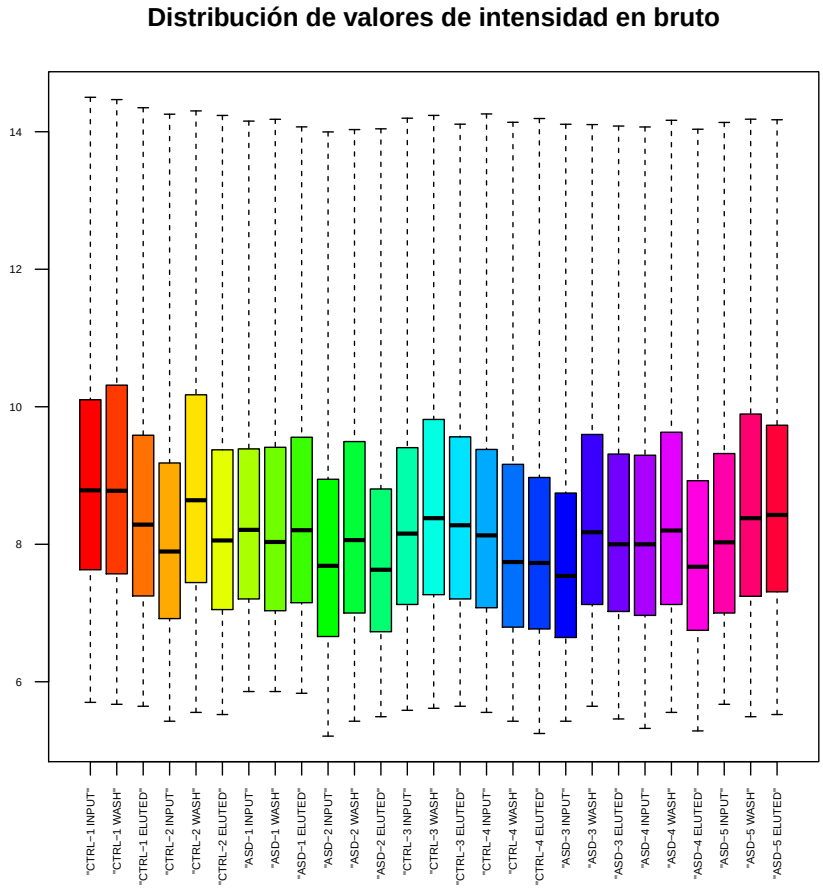


Figure 2: Diagrama de cajas de la intensidad de las muestras sin normalizar. Los colores no son significativos

Podemos ver que ninguna de las muestras destaca entre el resto. Hay pequeñas variaciones, pero es una característica esperable cuando comparamos los datos de intensidad en bruto.

Dendrograma del clúster jerárquico

El dendrograma nos ayuda a representar cómo se agrupan las muestras, y da pistas acerca de cuál es el factor experimental que determina las diferencias entre muestras. Aquellas muestras con datos más similares aparecerán agrupadas.

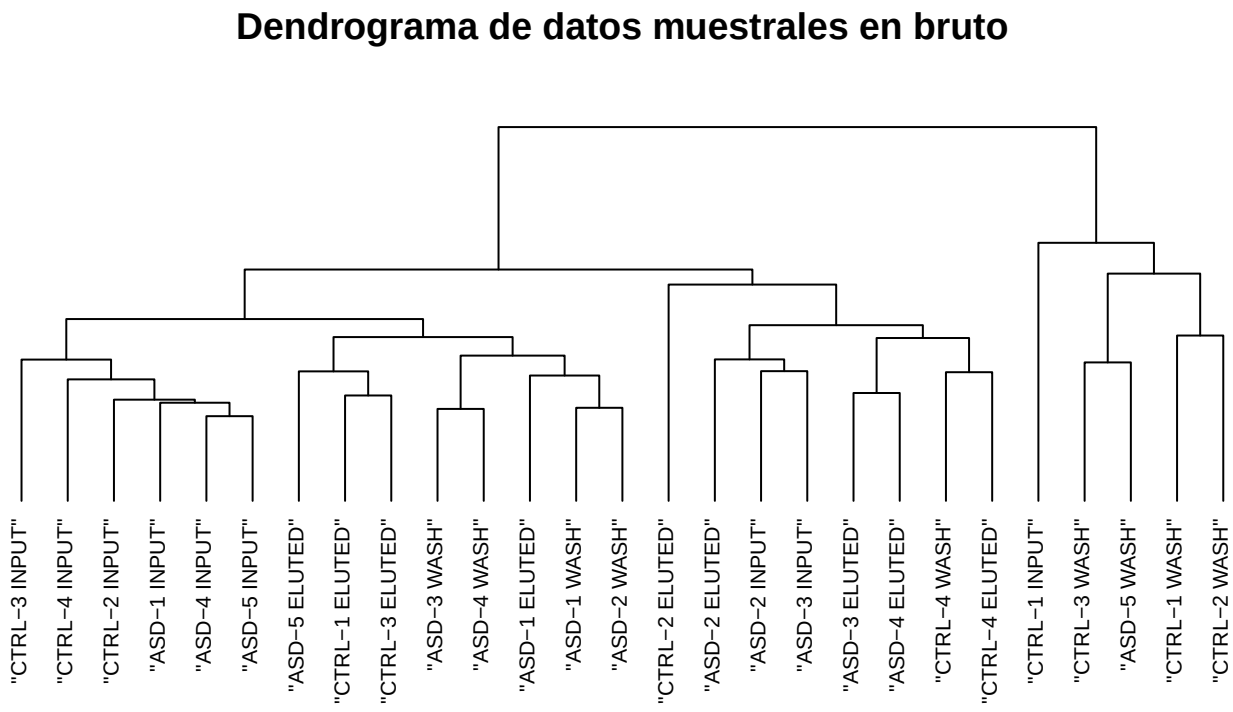


Figure 3: Dendrograma agrupando las muestras a partir de los datos de intensidad sin normalizar.

En este caso, y a primera vista, no parece haber un factor claro que haga que unas muestras están más cercanas entre sí que otras. Quizá el material (input/wash/eluted) en primer lugar, y el grupo (ctrl/asd) en segundo lugar; pero no es definitivo.

Componentes principales

El análisis de componentes principales nos puede servir para detectar si las muestras se agrupan con otras muestras procedentes del mismo grupo, o si no hay correspondencia entre muestras del mismo grupo.

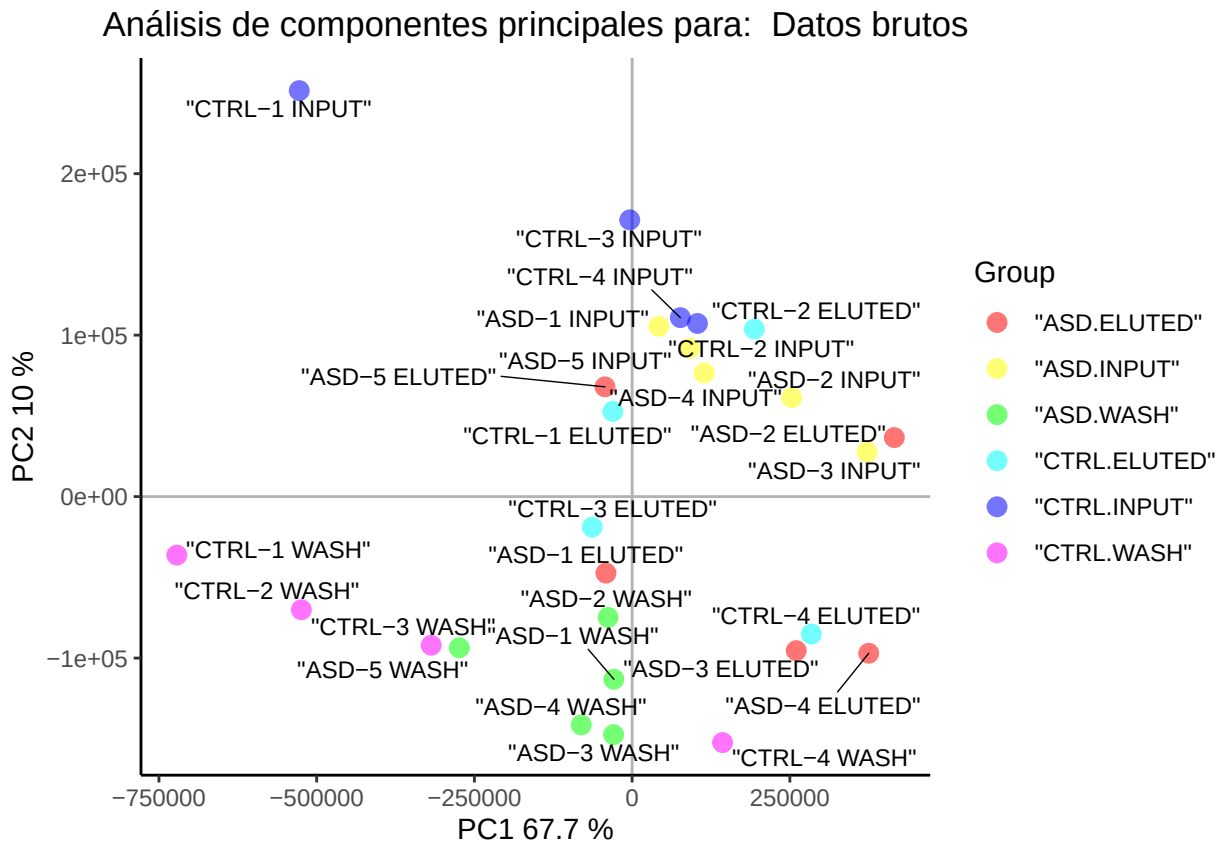


Figure 4: Gráfica de componentes principales a partir de los datos de intensidad sin normalizar. Atención a la muestra CTRL-1 INPUT en la esquina superior izquierda.

De forma similar a lo que veíamos en el dendrograma, las diferencias más importantes entre las muestras parecen deberse al material antes que al grupo.

La componente más importante explica el 67.7% de la variabilidad total de las muestras, y parece deberse principalmente al material; las muestras "Wash" se agrupan más a la izquierda, compartiendo la zona central con las muestras "Input", y las muestras "Eluted" agrupadas hacia la derecha del gráfico.

La muestra "CTRL-1 INPUT" aparece alejada, no sólo de las otras muestras del grupo CTRL-INPUT, sino de todo el resto de muestras. Esto por sí sólo no significa que la muestra sea defectuosa, pero sí que deberíamos fijarnos en ella y comprobar qué resultados obtiene en el resto de gráficos.

Imagen del array

Examinar la imagen del array nos permite hacer una evaluación de calidad a nivel "macro". Nos permite hacer una estimación a ojo de características como el balance del color, la uniformidad en la hibridación y en los spots, si el background es mayor de lo normal y la existencia de artefactos como el polvo o pequeñas marcas (rasguños).

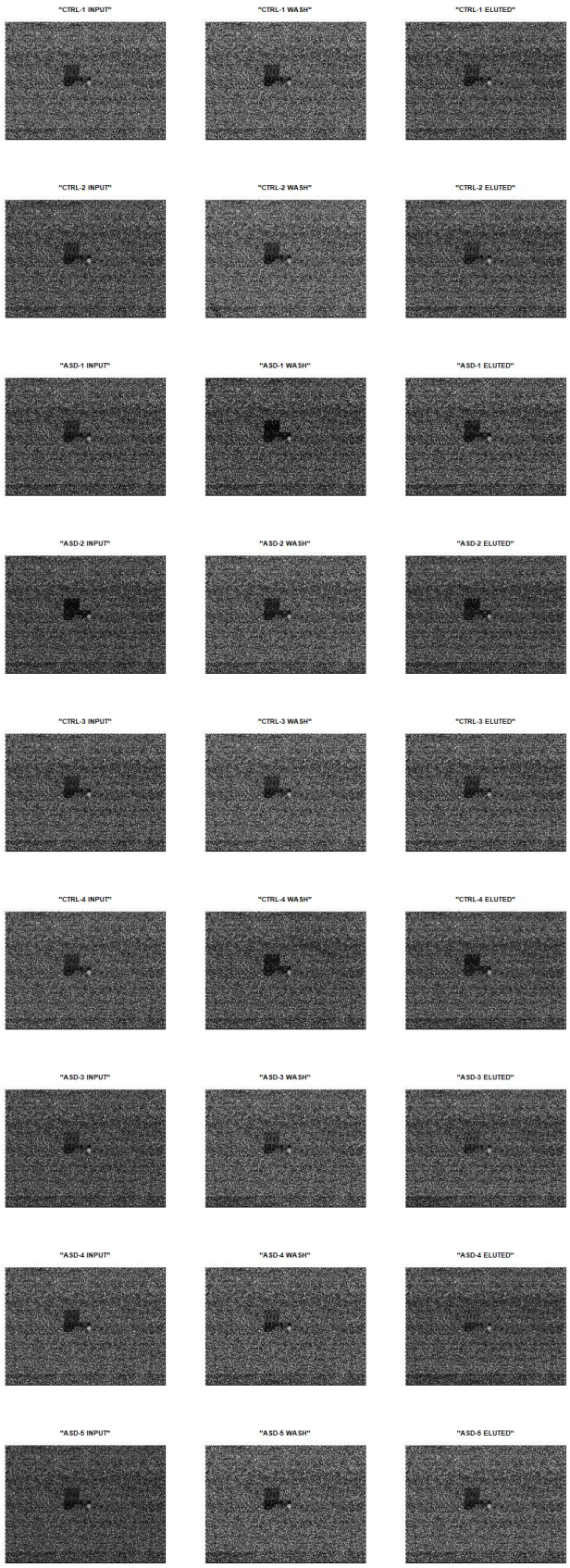


Figure 5: Imágenes de cada array a partir de los datos sin normalizar. No se detectan anormalidades.

A simple vista no se observa ningún gran defecto como roturas, burbujas o manchas.

Como conclusión del control de calidad de datos brutos, podemos decir que en ninguna de las gráficas hemos encontrado señales que nos hagan desconfiar de la calidad de ninguna de las muestras. La única excepción podría ser la muestra "CTRL-1 INPUT" en la gráfica de componentes principales; pero en el resto de gráficas no destaca de ninguna forma, así que la incluiremos con las demás en el resto del análisis.

Normalización de los datos

Antes de empezar el análisis de expresión es necesario procesar los datos brutos de forma que los datos de las diferentes muestras (micorarrays) sean comparables. El proceso de normalización intenta asegurarse de que las diferencias de intensidad reflejen la expresión diferencial de los genes, eliminando sesgos producidos por razones técnicas.

El método de normalización que hemos usado en este análisis es el método RMA (*robust multi-array average*), que es uno de los más usados en el ecosistema de Bioconductor.

Control de calidad sobre datos normalizados

Después de la normalización volvemos a realizar un control de calidad, para comprobar que el resultado de la normalización ha producido el efecto esperado en la distribución de los datos.

Diagrama de cajas

Con gráfico de diagrama de cajas volvemos a comparar la distribución de la intensidad entre las diferentes muestras.

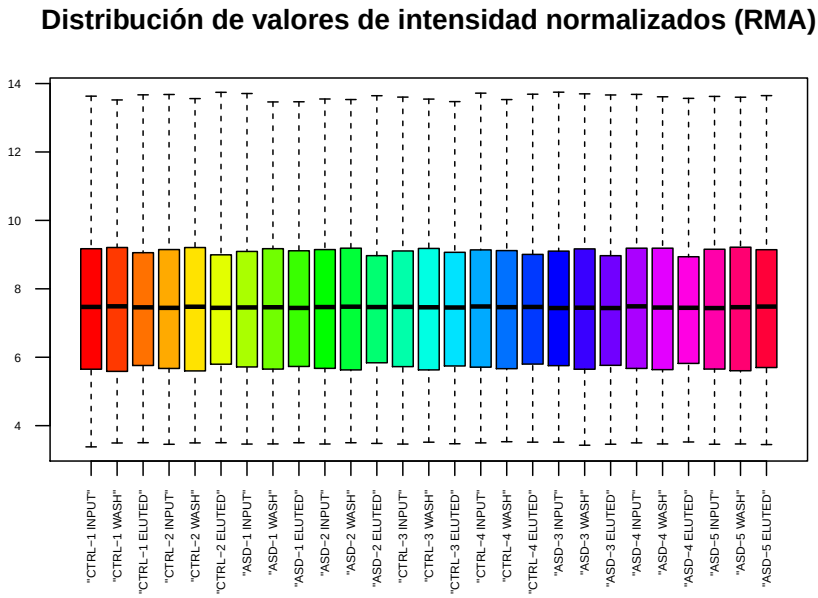


Figure 6: Diagrama de cajas de la distribución de intensidad, ya normalizada, de las muestras. Los colores no son significativos.

Si lo comparamos con el gráfico de antes de normalizar, en éste la distribución de intensidades es mucho más uniforme entre muestras.

Componentes principales

El análisis de componentes principales nos puede servir para detectar si las muestras se agrupan con otras muestras procedentes del mismo grupo o si no hay correspondencia entre muestras del mismo grupo.

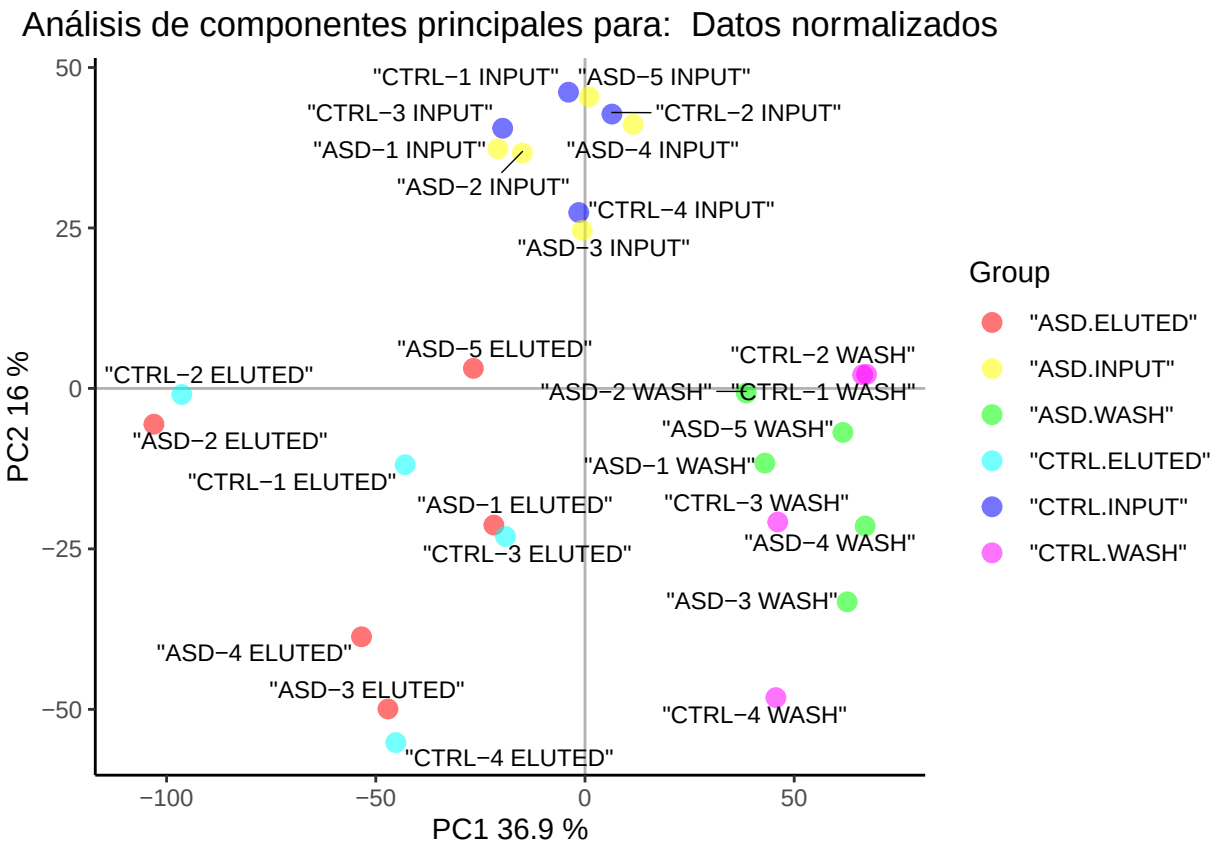


Figure 7: Gráfica de componentes principales creada a partir de los datos de intensidad normalizados.

Ahora el primer componente es responsable del 37% de la variabilidad total. Casi la mitad de lo que ocurría al utilizar los datos brutos. Las muestras se siguen separando según material, con "Input" en el centro, igual que antes; las muestras "Eluted" a un lado y las muestras "Wash" al otro.

La variabilidad explicada por la componente secundaria es el 16%, y también parece depender del material; separando claramente "INPUT" de los otros dos materiales ("ELUTED" y "WASH"). El grupo sin embargo - "CTRL" frente a "ASD" -, no parece tener una gran influencia en la variabilidad.

Un cambio notable en el diagrama de componentes es prinpales es la posición de la muestra "CTRL-1 INPUT", que ya no aparece separada del resto como sí ocurría al utilizar los datos brutos.

En conclusión, podríamos decir que la normalización de los datos ha producido el efecto esperado y no encontramos impedimento para proceder al análisis de los datos.

Filtraje no específico

Antes del análisis estadístico, para eliminar ruido y mejorar la sensibilidad, hemos eliminado datos procedentes de sondas que no aportan información o que pueden aportar información duplicada o confuso.

En concreto, se han eliminado datos de las siguientes sondas: Sondas cuya función específica es el control de calidad de las arrays y no se corresponden a genes, sondas para las que no se dispone de anotación, sondas duplicadas (diferentes sondas que corresponden al mismo gen)[1], y sondas que corresponden a más de un gen[2].

Después de esa primera selección hemos excluido también aquellas sondas con menor variabilidad de datos entre muestras[3]. Esto se ha hecho así porque no se espera que genes con poca variabilidad entre muestras presenten expresión diferencial, y reducir la cantidad de genes a examinar aumentará la sensibilidad del análisis.

Número inicial de sondas: 49 495

Sondas de control de calidad: 102
Sondas sin anotación: 430
Sondas duplicadas: 28777
Sondas que corresponden a más de un gen:1 249
Sondas excluidas por baja variabilidad: 9 419

Número de sondas restantes para análisis de expresión diferencial: 9 418

Identificación de genes diferencialmente expresados

Para el análisis de este ensayo, consideramos una matriz de diseño siguiendo un modelo de un factor con seis niveles, siendo esos niveles los grupos a los que está asignada cada muestra:
CTRL.INPUT, CTRL.WASH, CTRL.ELUTED, ASD.INPUT, ASD.WASH, ASD.ELUTED

Para cumplir con el objetivo del estudio, el interés se encuentra en averiguar qué genes presentan cambios significativos entre las muestras "Wash" y las muestras "Eluted". Y además, si esos cambios significativos se mantienen entre los grupos "Control" y "ASD".

Para replicar ese razonamiento, hemos organizado los siguientes contrastes:

WASH vs ELUTED en sujetos CONTROL
WASH vs ELUTED en sujetos ASD

También podría valer la pena comprobar si hay interacción entre condición y material.

[1] En estos casos, se ha comparado la variabilidad de los datos de cada sonda y se ha conservado sólo aquella con una mayor variabilidad entre muestras.

[2] En concreto, se trata de sondas que en la tabla de anotaciones están asociadas a más de una identificación Entrez. Puede ser interesante averiguar por qué ocurre esto; pueden explicarse por secuencias pertenecientes a familias génicas muy cercanas, pseudogenes, o por pertenecer a transcritos que 'corran' a lo largo de varios genes. En este caso, sin embargo, debido limitaciones de tiempo, nos hemos limitado a eliminar de la lista aquellas sondas que pueden identificar más de un gen.

[3] El criterio para eliminar sondas el análisis ha sido excluir aquellas sondas cuyo rango intercuartílico (IQR) estaba por debajo de la mediana del total de sondas.

Anotación de las listas de genes

A continuación presentamos una muestra de las listas de genes diferencialmente expresados para cada comparación, incluyendo el símbolo y nombre de cada gen, ordenados por p-valor ajustado.

Table 1: Comparación 1 (WvsE.CTRL): Genes diferencialmente expresados con colas poli-A cortas (WASH) frente a colas poli-A largas (ELUTED) en sujetos control (CTRL).

Gene.Title	Gene.Symbol	logFC	adj.P.Val
ERC2 intronic transcript 1	ERC2-IT1	-2.172550	8.202e-09
methionine sulfoxide reductase A	MSRA	-2.000875	8.202e-09
FK506 binding protein 11	FKBP11	-1.824279	4.846e-08
coiled-coil domain containing 65	CCDC65	-1.884547	1.746e-07
PMS1 homolog 2, mismatch repair system component pseudogene 9	PMS2P9	-1.693610	1.746e-07
cancer susceptibility candidate 4	CASC4	-1.986679	1.746e-07

Table 2: Comparación 2 (WvsE.ASD): Genes diferencialmente expresados con colas poli-A cortas (WASH) frente a colas poli-A largas (ELUTED) en sujetos autistas (ASD).

Gene.Title	Gene.Symbol	logFC	adj.P.Val
methionine sulfoxide reductase A	MSRA	-2.025492	7.976e-10
ERC2 intronic transcript 1	ERC2-IT1	-2.071131	1.867e-09
FK506 binding protein 11	FKBP11	-1.717882	1.522e-08
PMS1 homolog 2, mismatch repair system component pseudogene 9	PMS2P9	-1.675990	2.332e-08
cancer susceptibility candidate 4	CASC4	-1.957889	2.451e-08
coiled-coil domain containing 65	CCDC65	-1.771768	5.053e-08

Table 3: Comparación 3 (INT): Genes que cambian su expresión de forma significativa por interacción entre la condición del sujeto (CONTROL, ASD) y el material examinado (WASH, ELUTED).

Gene.Title	Gene.Symbol	logFC	adj.P.Val
tumor necrosis factor, alpha-induced protein 8-like 1	TNFAIP8L1	-0.0704	0.9999
histone cluster 1, H4e	HIST1H4E	0.6876	0.9999
histone cluster 1, H3a	HIST1H3A	0.3032	0.9999
histone cluster 1, H2bn	HIST1H2BN	-0.1529	0.9999
keratin associated protein 6-2	KRTAP6-2	-0.0198	0.9999
protocadherin gamma subfamily B, 3	PCDHGB3	0.2024	0.9999

Enlaces a las listas de genes en formato CSV

WvsE.CTRL

WvsE.ASD

Visualización (volcanoplots)

Para la visualización de estos datos elegimos las gráficas conocidas como volcanoplots. Cada gen está representado por un punto, con los cambios de expresión en el eje de abcisas y los p-valores ajustados en el eje de ordenadas. Se muestra en cada gráfica el nombre de los 5 genes más significativos. La línea roja horizontal corresponde al p-valor = 0.05.

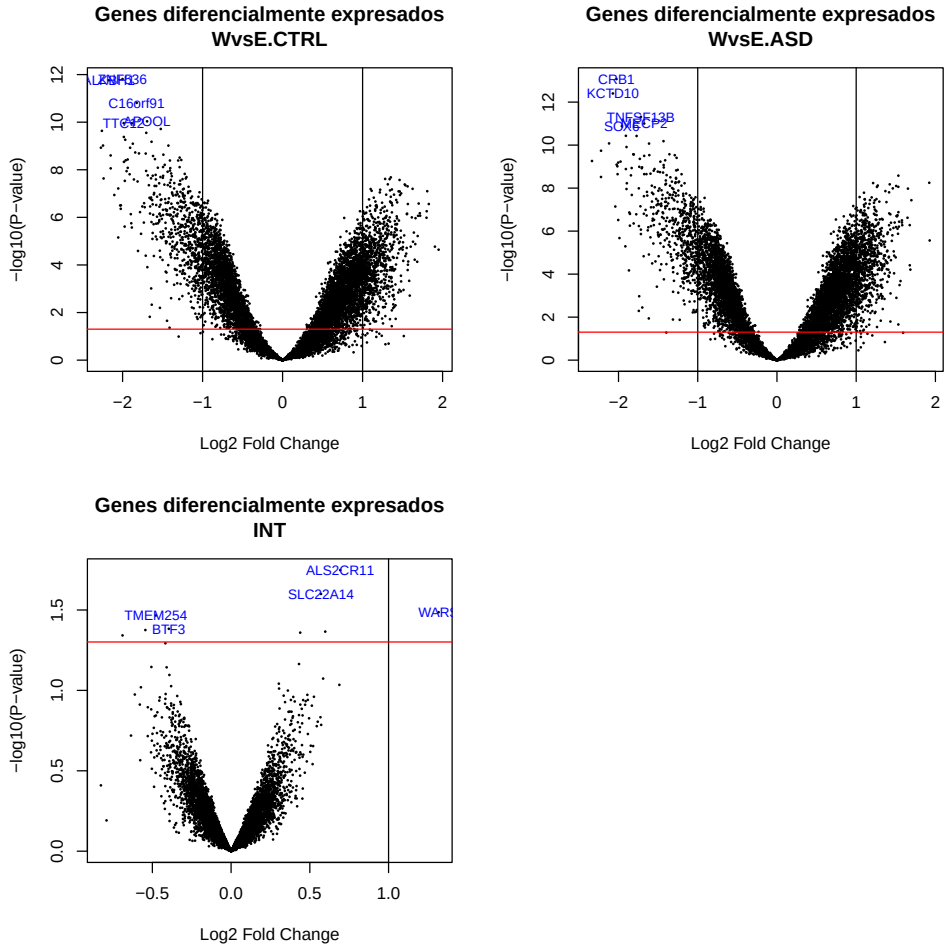


Figure 8: Volcanoplots de cada una de las listas anotadas de genes en cada una de las comparaciones. Se muestra en cada gráfica en nombre de los 5 genes más significativos. La línea roja horizontal corresponde al p-valor = 0.05.

Sólo examinando las gráficas podemos ver que no tiene sentido examinar los resultados de la comparación por interacción (INT), ya que apenas hay genes con expresión diferencial significativa; y estos, con sólo pequeñas diferencias de expresión.

Sí que son aparentes en cambio las diferencias de expresión significativas en las comparaciones entre "Wash" y "Eluted", aunque sólo con éstas no obtenemos respuesta a la pregunta de si hay diferencias entre "Control" y "ASD". Esto lo exploraremos en los siguientes apartados.

Comparaciones múltiples

Debido al diseño experimental y el objetivo del estudio, el interés final no está sencillamente en las listas de genes diferencialmente poliadenilados. Lo que queremos conocer es, de aquellos ARN que presentan

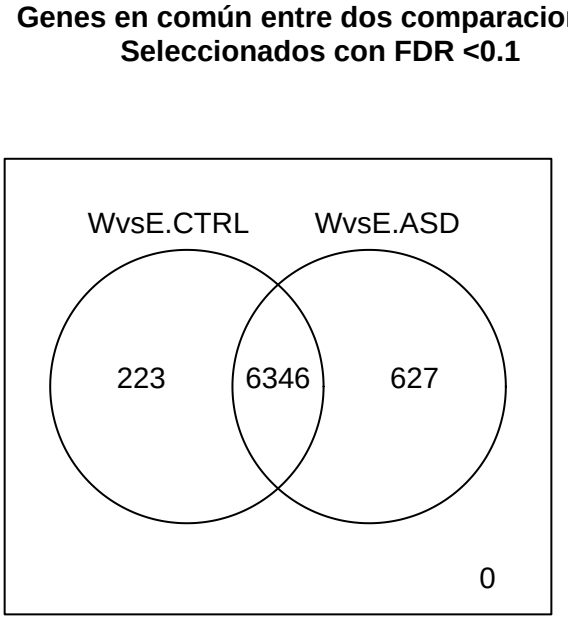
diferencias de poliadenilación, cuáles son exclusivos del grupo "Control" y cuáles del grupo "ASD".

Podemos hacer primero una comparación cuantitativa entre las comparaciónes; cuántos genes diferenciales son exclusivos de una de las comparaciones y cuántos son comunes.

Resumen de los resultados:

```
##        WvsE.CTRL WvsE.ASD  INT
## Down        2895     3017    0
## NotSig      2849     2445 9418
## Up          3674     3956    0
```

Diagrama de Venn



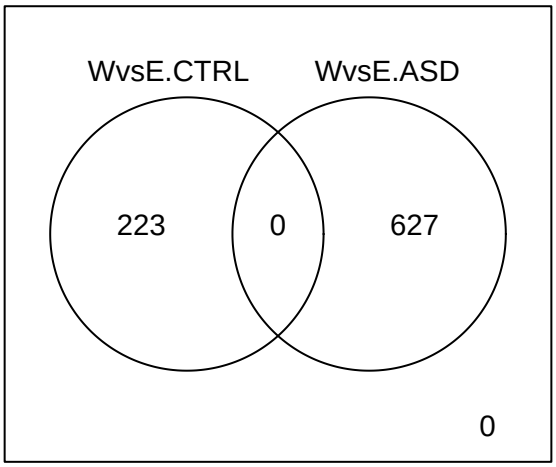
A partir del diagrama de Venn, vemos que **223** genes están diferencialmente poliadenilados sólo en las muestras "Control", y **627** sólo en las muestras "ASD". Son estos dos grupos de genes los que exploraremos para investigar las diferencias entre los grupos "Control" y "ASD".

Otro grupo de genes interesantes serían aquellos cuyo perfil de up/down regulación cambia en muestras de sujetos control y sujetos ASD.

Estos tres grupos son los que marcarán nuestra lista definitiva de genes de interés.

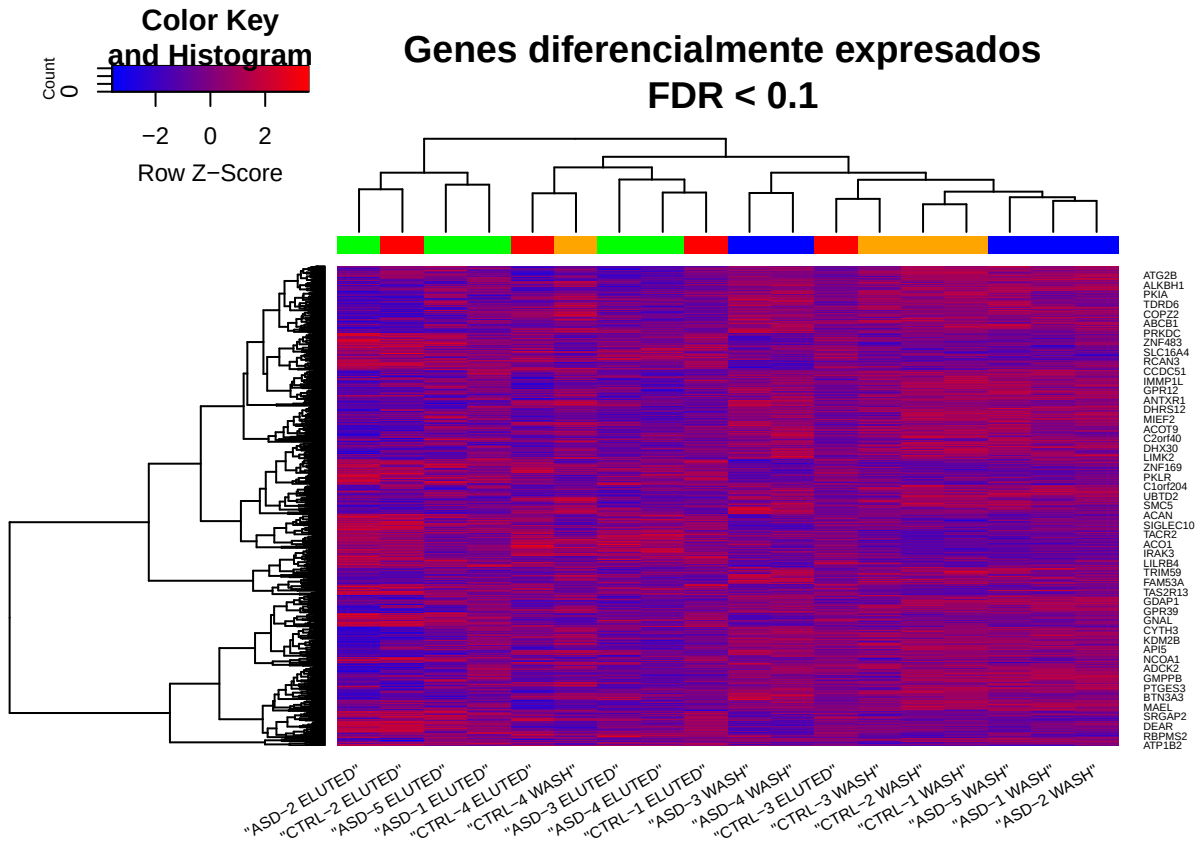
```
##        WvsE.CTRL WvsE.ASD
## Down          93      215
## NotSig       627      223
## Up           130      412
```

Genes en común entre dos comparaciones
Seleccionados con FDR <0.1



En resumen, encontramos **223** genes exclusivamente en sujetos “Control2 cuyos ARNm tienen colas diferencialmente largas o cortas.
Encontramos **627** genes exclusivamente en sujetos”ASD” cuyos ARNm tienen colas diferencialmente largas o cortas.
Ningún gen con diferenciación invertida entre comparaciones.

Visualización de perfiles de expresión mediante heatmap



Los mapas de calor (heatmaps) pueden servir para agrupar muestras y genes por similaridad en los patrones de expresión. En este caso, para simplificar, hemos eliminado las muestras del grupo "Input", que no nos están ofreciendo información útil.

Al igual que ocurría al examinar los componentes principales y el dendrograma, las muestras se agrupan sobre todo por material ("Wash" y "Eluted"). Dentro del grupo "Wash", sí que parece haber una separación entre "Control" y "ASD" (derecha del gráfico).

Clave de color para las muestras:
Control-Wash: naranja ASD-Wash: azul
Control-Eluted: rojo ASD-Eluted: verde

Significatividad biológica

Una vez tenemos nuestras listas de genes anotadas, un herramienta más para interpretar los resultados del estudio es el examen de la significatividad biológica. En este informe, lo que hemos hecho es, a partir de las listas de genes con comportamiento diferencial para las condiciones "Control" y "ASD", comprobar si existen funciones, procesos biológicos o rutas moleculares que aparezcan con más frecuencias en estas listas que en el resto de genes analizados.

Como listas de genes hemos utilizado las siguientes:
Común - lista completa de genes con comportamiento diferencias en grupo "Control" frente a "ASD".
Control - lista de genes con comportamiento diferencial exclusivos del grupo "Control" (es un subset de la lista "Común").

ASD - lista de genes con comportamiento diferencial exclusivos del grupo "ASD" (es un subset de la lista "Común").

Universo - lista de todos los genes detectables con el modelo de array usado en este estudio.

Test de sobrerrepresentación de términos GO

El análisis estadístico lo hemos realizado con la función *enrichGO()* del paquete *clusterProfiler* para el lenguaje R. Esta función devuelve un listado de términos GO estadísticamente más representados en nuestra lista de genes, con respecto a la lista Universo.

Para un valor de corte del p-valor de 0.05, y valor de corte del q-valor de 0.2; no se ha encontrado **ningún** término GO estadísticamente más representado en ninguna de las tres listas con respecto a la lista Universo.

Gene Set Enrichment Analysis

Este tipo de análisis está especializado en detectar situaciones en las que las diferencias de expresión son pequeñas, pero coordinadas para un grupo de genes relacionados. En este informe hemos utilizado la función *gseGO* del paquete *cluseterProfiler* para realizar este análisis.

El resultado ha sido el mismo que en el análisis anterior. No se ha detectado ningún grupo de genes especialmente representado en el listado de nuestros genes de interés con ARN diferencialmente poliadenilados.

Representación en la base de datos SFARI

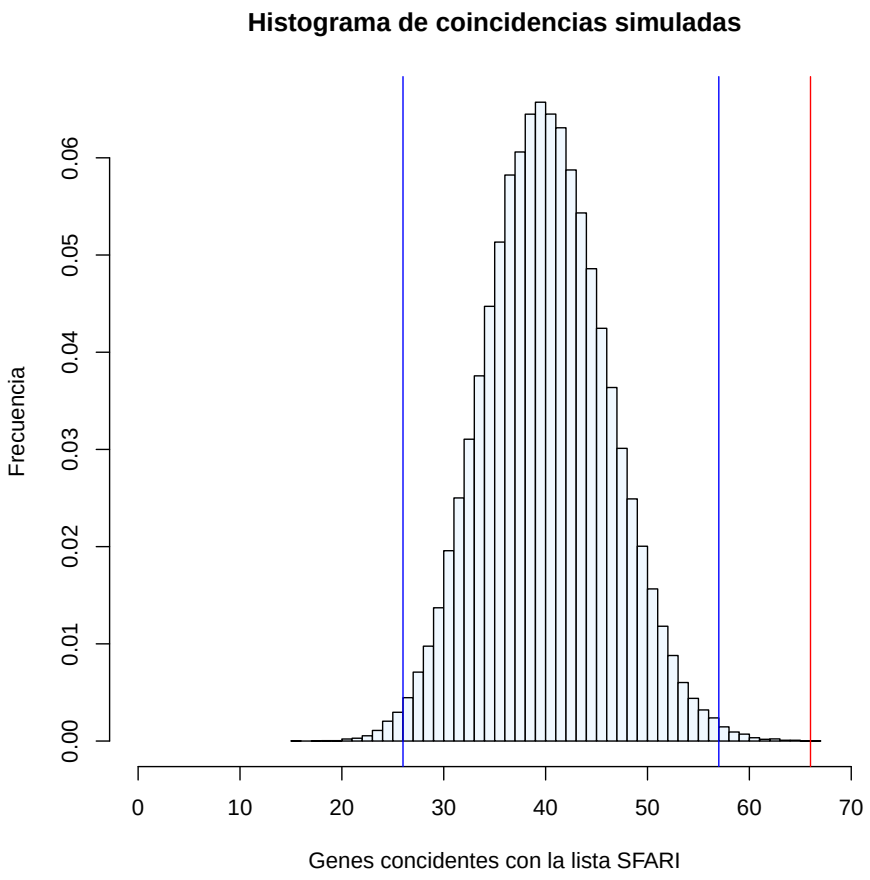
SFARI Gene es una base de datos al servicio de la investigación del autismo, centrada en genes implicados en susceptibilidad a dicho síndrome(SFARI 2019). Parte de la información que integra está compuesta por el módulo Human Gene, un set de datos que recopila centenares de genes humanos potencialmente relacionados con el ASD.

Como un último intento de investigar la significatividad biológica de nuestra lista de genes diferencialmente poliadenilados entre los grupos "Control" y "ASD", hemos comparado dicha lista con la lista de genes en el módula Human Gene de SFARI y extraído los genes que aparecen en ambas listas. Posteriormente, mediante simulaciones, hemos comprobado la probabilidad de obtener esa misma cantidad de genes muestreando al azar la lista total de genes que pueden detectarse con el modelo de array usado en el estudio.

El proceso de simulación ha consistido en, primeramente, extraer el mismo número de genes (850) que aparecen en nuestra lista de genes diferencialmente poliadenilados de entre la lista de genes que pueden ser detectados mediante el array (18 837. Después se ha comparado la lista aleatoria resultante con la lista de genes SFARI para averiguar la cantidad de genes en común. El proceso se ha repetido 100 000 veces.

Los resultados se resumen en el siguiente gráfico:

\begin{figure}[H]



{

}

\caption{Histograma con la distribución de 100 000 simulaciones. Las líneas verticales azules corresponden a los percentiles 0.5% y 99.5%. La línea roja corresponde a la cantidad de genes comunes entre nuestra lista diferencial y la lista de genes SFARI.} \end{figure}

Media: 41
Desviación típica: 6
Quantil 0.5%: 26
Quantil 99.5%: 57

Resultados

Al comparar los ARN diferencialmente poliadenilados en el estudio, obtenemos un total de 850 genes cuyo patrón de poliadenilación (cola poli-A larga o corta) es diferente en muestras “Control” y muestras “ASD”.

Buscando grupos de genes diferencialmente sobrerepresentados en este set de genes, no encontramos ninguno. Ni implicados en procesos biológicos comunes, componentes celulares, ni funciones moleculares.

Sin embargo, al examinar al examinar las coincidencias entre nuestra lista de genes, y la base de datos SFARI de genes relacionados con el ASD, sí encontramos algo interesante. El resultado de la simulación nos sugiere que el puro azar nos proporcionaría de media 41 genes coincidentes entre nuestra muestra y la lista SFARI de genes asociados a ASD. Sin embargo, en el estudio la cantidad de genes coincidentes es de 66, muy por encima del quantil 99.5% (57) de la distribución producida por la simulación. Esto nos hace

pensar que es muy poco probable que esa coincidencia de genes entre nuestra lista de genes diferencialmente poliadenilados, y la lista SFARI, sea debida al azar.

Discusión

Respecto al estudio original del que he podido analizar los datos de expresión, no tengo ningún comentario, tanto el diseño experimental como los métodos y los análisis parecen sólidos.

En cambio, en éste análisis he tenido grandes dudas al realizar la simulación. Especialmente al considerar si la población de genes de la que extraer las muestras debía ser la lista total de genes que el array es capaz de detectar, o sólo el subconjunto de genes que muestran un patrón de poliadenilación diferencial (incluyendo tanto los exclusivos de los grupos "Control" y "ASD" como los comunes). Finalmente me he decantado por la lista total de genes, pero llevado más por la intuición que por un razonamiento estadístico sólido.

Apéndice A: Código

El documento original en formato .Rmd, que incluye el código completo en lenguaje R usado para generar este informe, se puede consultar y descargar en el siguiente repositorio de Github:
jorgevallejo/analisis_GSE113834

Apéndice B: Reproducibilidad

```
sessionInfo() # For better reproducibility
```

```
## R version 3.6.3 (2020-02-29)
## Platform: i686-pc-linux-gnu (32-bit)
## Running under: Ubuntu 16.04.7 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] org.Hs.eg.db_3.10.0    AnnotationDbi_1.48.0   clusterProfiler_3.14.3
##  [4] gplots_3.0.4           limma_3.42.2           ggrepel_0.8.2
##  [7] affy_1.64.0            oligo_1.50.0           Biostrings_2.54.0
## [10] XVector_0.26.0         IRanges_2.20.2         S4Vectors_0.24.4
```

```
## [13] Biobase_2.46.0         oligoClasses_1.48.0    BiocGenerics_0.32.0
## [16] ggplot2_3.2.1          knitr_1.25
##
## loaded via a namespace (and not attached):
##   [1] httr_1.4.1                  jsonlite_1.6
##   [3] munsell_0.5.0               dplyr_1.0.0
##   [5] pillar_1.4.2                pkgconfig_2.0.3
##   [7] compiler_3.6.3              DBI_1.1.0
##   [9] lazyeval_0.2.2              cowplot_1.0.0
##  [11] hms_0.5.3                   tibble_2.1.3
##  [13] R6_2.4.0                    yaml_2.2.0
##  [15] affxparser_1.58.0           ggforce_0.3.2
##  [17] data.table_1.12.4           gdata_2.18.0
##  [19] tools_3.6.3                 stringr_1.4.0
##  [21] DOSE_3.12.0                 prettyunits_1.1.1
##  [23] foreach_1.5.1               digest_0.6.21
##  [25] KernSmooth_2.23-17          GO.db_3.10.0
##  [27] GenomeInfoDb_1.22.1         codetools_0.2-16
##  [29] vctrs_0.3.1                 fastmatch_1.1-0
##  [31] withr_2.1.2                 ff_2.2-14.2
##  [33] xfun_0.10                   triebeard_0.3.0
##  [35] tidyselect_1.1.0            europepmc_0.4
##  [37] scales_1.0.0                memoise_1.1.0
##  [39] stringi_1.4.3               xml2_1.3.2
##  [41] DO.db_2.9                   labeling_0.3
##  [43] highr_0.8                   purrr_0.3.4
##  [45] GenomeInfoDbData_1.2.2      lattice_0.20-41
##  [47] bit64_0.9-7                 Rcpp_1.0.2
##  [49] polyclip_1.10-0             caTools_1.17.1.2
##  [51] tweenr_1.0.1                enrichplot_1.6.1
##  [53] BiocManager_1.30.10         grid_3.6.3
##  [55] blob_1.2.1                  GenomicRanges_1.38.0
##  [57] preprocessCore_1.48.0       viridis_0.5.1
##  [59] plyr_1.8.4                  viridisLite_0.3.0
##  [61] DelayedArray_0.12.3         fgsea_1.12.0
##  [63] tidyr_1.1.0                 matrixStats_0.56.0
##  [65] ggridges_0.5.2              glue_1.4.1
##  [67] magrittr_1.5                ggplotify_0.0.5
##  [69] igraph_1.2.5                SummarizedExperiment_1.16.1
##  [71] farver_2.0.3                gridExtra_2.3
##  [73] affyio_1.56.0               GOSemSim_2.12.1
##  [75] lifecycle_0.2.0             htmltools_0.4.0
##  [77] RSQLite_2.2.0               crayon_1.3.4
##  [79] urltools_1.7.3              graphlayouts_0.7.0
##  [81] rvcheck_0.1.8               gtable_0.3.0
##  [83] ggraph_2.0.3                zlibbioc_1.32.0
##  [85] colorspace_1.4-1            MASS_7.3-53
##  [87] gtools_3.8.1                RCurl_1.98-1.1
##  [89] bitops_1.0-6                RColorBrewer_1.1-2
##  [91] progress_1.2.2              Matrix_1.2-18
##  [93] bit_4.0.4                   tidygraph_1.2.0
##  [95] reshape2_1.4.3              rmarkdown_1.17
##  [97] gridGraphics_0.5-0          iterators_1.0.13
##  [99] generics_0.0.2              splines_3.6.3
```

```
## [101] evaluate_0.14               rlang_0.4.6
## [103] BiocParallel_1.20.1         qvalue_2.18.0
```

Notas

References

Clough, Emily, and Tanya Barrett. 2016. "The Gene Expression Omnibus Database." In *Statistical Genomics*, 93–110. Springer.

Parras, Alberto, Héctor Anta, María Santos-Galindo, Vivek Swarup, Ainara Elorza, José L Nieto-González, Sara Picó, et al. 2018. "Autism-Like Phenotype and Risk Gene MRNA Deadenylation by Cpeb4 Mis-Splicing." *Nature* 560 (7719). Nature Publishing Group: 441–46.

SFARI, Gene. 2019. "Human Gene Module."