Anásisis de microarrays

https://github.com/IagoLast/ADO_PEC_1

Iago Lastra Rodríguez

Resumen

En este trabajo se realiza un análisis de microarrays utilizando los archivos .CEL originales de un estudio publicado en el año 2019 que investiga los posibles efectos secundarios de los tratamientos mediante inhibidores de factores de necrosis tumoral (TNF). Para realizar el análisis se han descargado y verificado datos correspondientes a dos grupos de ratones tratados y sin tratar. Se han filtrado aquellos que proporcionan poca información y se han aplicado diferentes técnicas para comprobar que genes presentan una variabilidad estadísticamente significativa. Finalmente se ha aplicado un análisis de significación biológica para descubrir las vías afectadas por el tratamiento.

Índice

1. Objetivos

Aunque los inhibidores de factores de necrosis tumoral TNF son utilizados en el tratamiento de enfermedades inflamatorias crónicas no existe demasiada información acerca de cómo pueden afectar estos tratamientos al funcionamiento normal del sistema nervioso central.

En este trabajo se analizarán 8 Microarrays de ARN para buscar diferencias estadísticamente significativas entre muestras sin tratar (WT) y muestras sometidas a tratamientos de inhibición de TNF.

Una vez obtenidas estas diferencias se realiza una interpretación biológica de los resultados comparando los niveles de expresión obtenidos contra las bases de datos disponibles para averiguar qué vías son as más afectadas por estos tratamientos.

2. Materiales y Métodos

Este trabajo es un estudio de comparación de dos grupos (class comparison) donde se han tomado muestras correspondientes al dia 13.5 de la fase embrionaria (E13.5) al séptimo día de vida (P7) y en adultos de 2 y 4 meses de vida (A2 y A4 respectivamente) de un grupo de control (WT) de ratones C57BL/6 y un segundo grupo de ratones tratados (TNF-/-).

Los microarrays utilizados son del modelo GeneChip Mouse Gene 1.0 ST Array de Affymetrix que según su especificación contienen aproximadamente 25 sondas (probes) diseñadas para cubrir 28,853 genes bien conocidos y anotados.

Para el análisis se ha utilizado el software R siguiendo los pasos que se detallan a continuación y todo el código está disponible en el repositorio original.

Figura 1: GeneChip Mouse Gene 1.0 ST

3. Procedimiento de trabajo

En un primer paso se analizaron gráficamente los archivos .CEL originales buscando posibles errores en los datos. Aunque tanto el histograma como el boxplot mostraron datos de intensidades bastante uniformes se realizó una comprobación adicional utilizando el paquete `arrayQualityMetrics` para verificar que los datos no contenían errores (Ver resultados en detalle).

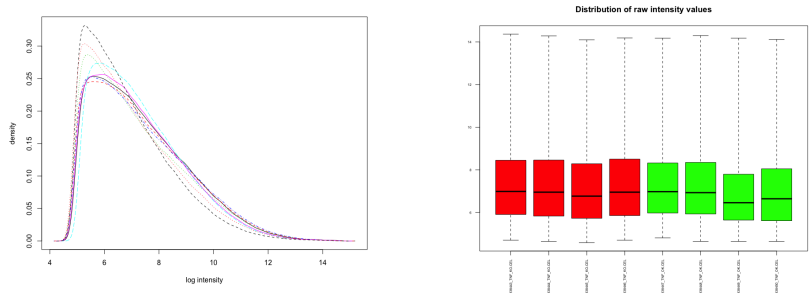

Figura 2: Distribución de intensidades en los datos originales.

3.1. Normalización

Para poder realizar un análisis de la expresión diferencial de los datos es necesario transformar los datos para que sean comparables entre sí.

Esta transformación se ha realizado utilizado el algorimo Robust Multichip Analysis (RMA) que a grandes rasgos corrige el ruido de forndo, normaliza los datos y realiza una estimación final de la intensidad.

Una vez obtenidos los datos normalizados se repite el control de calidad sobre los mismos. (Ver resultados en detalle)

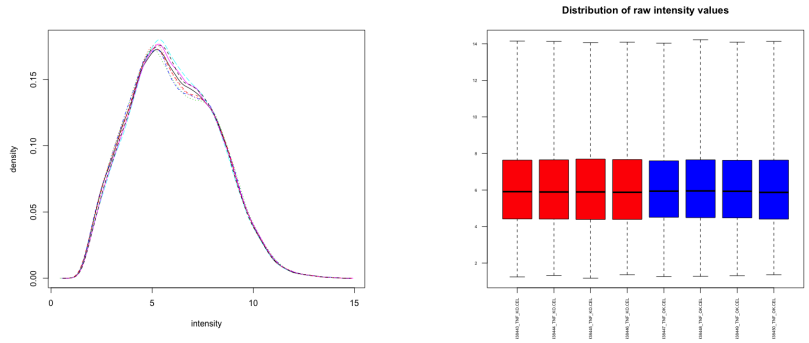

Figura 3: Distribución de intensidades en los datos normalizados.

3.2. Filtrado

Antes de empezar el análisis es interesante eliminar los genes cuya variabilidad puede ser consecuencia de un ruído aleatorio y aquellos de los que no se dispone de anotaciones.

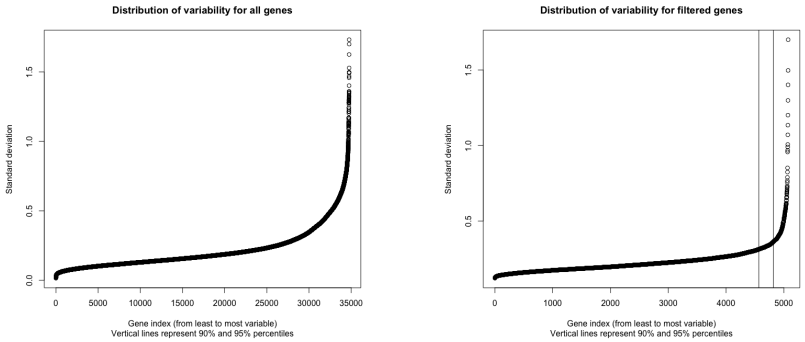

Figura 4: Variabilidad de los genes antes y después del filtrado.

Para ello se ha utilizado la función `nsFilter` paquete `geneFilter` para eliminar 1594 genes duplicados, 15225 con una variabilidad irrelevante y 12866 de los que actualmente no se disponen anotaciones

numDupsRemoved	1594
numLowVar	15225
numRemoved.ENTREZID	12866

3.3. Selección de genes

Para escoger los genes se han utilizado dos aproximaciones.

Por un lado se ha realizado un t-test (`rowttest`) y por otro se ha utilizado el método de Smyth (`limma`) visto en prácticas previas.

Debido al grán número de genes procesados se ha optado por aplicar una corrección sobre el p-valor. Dado que estamos dispuestos a asumir falsos positivos a cambio de maximizar los genes candidatos el método seleccionado es el de Benjamini & Hochber. Los genes seleccionados mediante rowtest y limma respectivamente han sido:

SYMBOL	BH
9430060I03Rik	0.0431307
Gm10787	0.0431307
Gm10024	0.0431307
Gm11696	0.0431307
Dnmt3aos	0.0491421
Gm10782	0.0431307
BC025933	0.0431307
Gm10536	0.0431307
Gm10532	0.0431307
Gm10857	0.0431307
Gm10804	0.0491421
Gm10714	0.0431307
Oog3	0.0431307
Gm10369	0.0431307
Gm10445	0.0491421
Gm10610	0.0431307
Fam129c	0.0431307
Gm10655	0.0431307

PROBE_ID	SYMBOL	adj.P.Val
10423836	Cthrc1	0.0311910
10418205	Plac9b	0.0311910
10566326	Trim12a	0.0311910
10471675	Glo1	0.0311910
10398432	Mir377	0.0311910
10572130	Lpl	0.0361161
10412394	Nnt	0.0410250

3.4. Interpretación biológica de los resultados

A partir de las listas obtenidas en el paso anterior se puede realizar un Pathway Enrichment Analysis para identificar las funciones biológicas afectadas por el tratamiento entre las que se pueden destacar las siguientes:

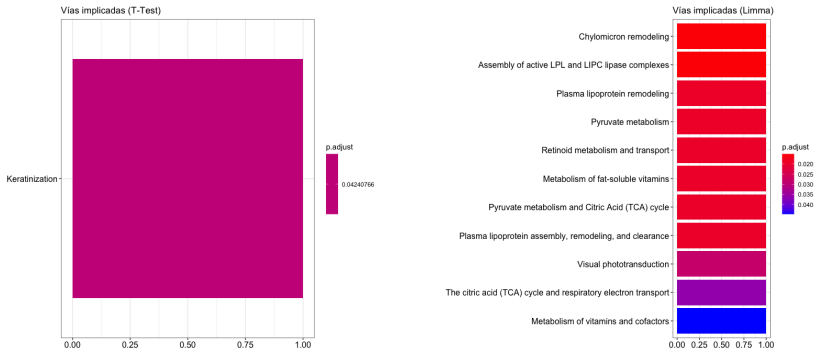

Figura 5: Vías con mayor significancia entre los grupos estudiados.

4. Resultados

Aunque los resultados han de ser interpretados por un profesional con los conocimientos adecuados, me ha paredico interesante comprobar que la literatura confirma las relaciones obtenidas en este trabajo:m

- Keratinization:
  - Se relaciona los anti-TNF con los keratinocitos
- Chylomicron remodeling
  - Se relaciona con TNF y LPS
- Assembly of active LPL and LIPC lipase complexes
  - La inactivación del TNF afecta el metabolismo lipídico
  - La inactivación del TNF afecta a LPL
- Plasma lipoprotein remodeling
  - La sobreactivación de TNF disminuye la creación de lipoproteinas
- Pyruvate metabolism
  - Se relacionan los TNF con el metabolismo del “Pyruvate”
- Retinoid metabolism and transport
  - Los retinoides suprimen la respuesta TNF

5. Discusión

Aunque supongo que es bastante habitual, la principal limitación que veo en el estudio es la poca cantidad de muestras utilizadas.

Por otra parte es una pena que no todos los probes del microarray tengan asociado un símbolo, algunos de ellos han tenido que ser descartados por este motivo pese a tener un nivel de significancia elevado.

La parte positiva es que teniendo los datos en crudo se podría repetir el análisis en el futuro confiando en tener unas bases de datos más completas.

Por último destacar la importancia de las técnicas elegiadas a la hora de obtener resultados. La Keratinization sólamente entra en escena cuando se aplica (t-test) y pasaría inadvertida si sólo se utiliza limmma. Aunque en el apartado 4 se puede ver que están todas relacionadas me parece interesante como bioinformático ofrecer al investigador el máximo número de opciones.

6. Apéndice

- **ID:** GSE134178
- Link al estudio Original
- Link al BioProject
- Link al repositorio en github
- Link al Código en R del estudio [Github]