PEC 1:
Análisis de Datos Ómicos

Sara Álvarez González
Máster universitario en Bioinformática y Bioestadística
(Dated: April 26, 2020)

La aparición de de los Microarrays supuso la revolución en el estudio del transcriptoma. Especialmente ha sido relevante su aplicación en el campo de la expresión génica, como en una de sus aplicaciones principales, en el estudio escogido se han aplicado para la búsqueda de un patrón de reconocimiento en el proceso de retención y codificación de nueva información en las subregiones hipocampales (CA1, CA3 y DG). Se buscaron por tanto una lista de genes diferencialmente expresados (DE) en las secciones estudiadas analizando también sujetos control que no se sometieron al protocolo de aprendizaje diseñado. Así pues, el estudio se llevó a cabo a través de librerías de Bioconductor para el programa R, principalmente Affy y Limma. Se encontraron un total de 258 genes DE, así como 377 términos GO con una significación menor $p < 0.01$.

Título original: Rapid encoding of new information alters the profile of plasticity-related mRNA transcripts in the hippocampal CA3 region.

Información GEO: GSE11476
Plataforma GPL341 Affymetrix Rat Expression 230A Array

I. RESUMEN DEL ESTUDIO

El artículo[1] elegido versa sobre la teorización de que existe una función en el lóbulo temporal para la memoria. Para ello, recoge información de tres subregiones (CA3, CA1 y DG) para evaluar su relevancia para una prueba de aprendizaje espacial diseñada. A partir de la recogida de 40 muestras, una por cada una de las 40 ratas noruegas empleadas, con una parte control y otra experimental, se realiza la comparación de genes diferencialmente expresados. Se puede concluir que, efectivamente, la región CA3 va a desempeñar mecanismos de remodelación sináptica bases para la codificación de información en la memoria a largo plazo, debido a los resultados de sobreexpresión de los genes y el análisis de caminos biológicos que éstos recorren.

II. OBJETIVOS

Implementar un script en lenguaje R para el análisis de los datos obtenidos en los Microarrays con la meta de:

- Obtener un listado de genes DE en cada una de las subregiones estudiadas de los sujetos control vs los experimentales.
- Analizar los genes DE que sean confluyentes en estas zonas haciendo un estudio de las rutas metabólicas implicadas y con relación a esa lista de genes obtenidos.

III. MATERIALES Y MÉTODOS

En esta sección se van a ir describiendo por secciones diferentes materiales y métodos empleados para la realización de este trabajo. Es de comentar, sin embargo, que detalles concretos sobre la ejecución del código o la razón de la utilización de cada comando, junto con una explicación de los datos obtenidos en cada parte, se encontrará en los dos scripts adjuntados a este trabajo (`codigo.R`, `PEC1_pipeline.Rmd` y `PEC1_pipeline.html`). Solo una parte breve de los mismos se traerá a este documento.

A. Descripción de los datos empleados

El conjunto de datos `.CEL` empleado fue descargado de la plataforma: **GPL341 Affymetrix Rat Expression 230A Array**. Encontrado en la serie de base de datos **GSE11476**. En ella encontramos las 40 muestras con las que trabajamos en el estudio: 14 de ellas corresponden al área CA3 (7 control y 7 experimentales), 12 del área CA1 (6 control y 6 experimentales) y las últimas 14 corresponderían a la tercera área, DG (7 control y 7 experimentales). Además, el organismo empleado como unidad experimental fue la especie *Rattus norvegicus*. La página puede ser consultada a través de la URL: GSE11476.

Este estudio, como se acaba de comentar, se llevó a cabo con Microarrays de un color con el objetivo de hacer una comparación de grupos con dos factores: área (CA3, CA1 y DG) y estado (control o experimental), con 3 y 2 niveles respectivamente. Sin embargo, para el análisis nos interesa detectar los genes DE en cada área comparando los genes de ambos estados.

B. Métodos empleados en el análisis

El programa empleado para este trabajo fue `RStudio`. Se trata de un ambiente de desarrollo bajo el lenguaje R. Se empleó bajo su versión para Windows y versión 3.6. Para su realización se emplearon tanto un script `.R` para registrar el código al completo, como un `RMarkdown` para

añadir anotaciones.

Se llevó a cabo además como principal gestor de librerías, **Bioconductor**, debido a que posee la mayor colección de librerías R para analizar datos biológicos, especialmente destacables fueron:

- **Affy**[2]: Diseñado para trabajar con arrays de oligonucleótidos de Affymetrix.
- **Limma**[3]: Para llevar a cabo análisis de datos, modelos lineales y expresión diferencial para los datos de microarrays.
- **topGO**[4]: Para conseguir la significación biológica a través de mapas de anotación de la "Gene Ontology" para los genes DE.

C. Descripción de los datos empleados

Se puede decir que el pipeline seguido para llevar a cabo este análisis se puede comprender en tres secciones (correspondientes a las principales utilizaciones de los paquetes previamente comentados):

1. *Importación y preprocesado de los datos* (paquete Affy)

En primer lugar, tras haber descargado las 40 muestras en formato `.CEL` a nuestro directorio de trabajo, se procedió a importarlos, obteniendo un objeto de **clase AffyBatch**. Para identificar los grupos existentes y a qué grupo pertenecía cada muestra, fue preciso extraer los datos de expresión y los datos fenotípicos. Este proceso fue llevado a cabo de dos maneras diferentes, renombrando tanto la información fenotípica de los archivos, como a través del archivo `.txt` también encontrado en la plataforma del estudio. Los datos de identificación de cada muestra fueron extraídos de la página fuente, donde al lado de cada archivo `.CEL` ponía el área y estado al que correspondía.

En segundo lugar, se procedió a llevar a cabo un control de calidad de los datos crudos, necesario para establecer un primer contacto visual con los datos que se nos presentan, con el fin de decidir si parecen correctos o van a presentar anomalías que precisan ser corregidas antes de comenzar con el tratamiento de estos. Para este estudio se llevaron a cabo estudios gráficos (boxplot, histograma y clúster jerárquico) que dejaban clara la necesidad de una normalización previa para poder ser comparados estando en una escala equivalente y eliminar posibles sesgos técnicos. Además, también se pasó un control de calidad en el que se pudo ver que efectivamente, ninguna de las muestras va a mostrar una clara degradación del ARN y no sería necesario eliminar ninguna. Incluso se lleva a cabo una muestra cómo hacer un análisis de calidad de la fluorescencia representado la imagen de cada muestra con todos sus genes. En tercer lugar, habiendo determinado la importancia de esta metodología, se lleva a cabo la normalización. Se ha elegido el método de **RMA**, debido que se trata de una muestra de microarray de un color, concretamente Affymetrix. Este método de resumen y normalización va a aportar mejores antes las deficiencias de los anteriores métodos (**MAS4** y **MAS5**). Los pasos que va a llevar a cabo este **Robust multi-array average** serán: un ajuste del ruido de fondo, la toma de logaritmos de base 2 ajustados por el anterior, la normalización por cuantiles de los valores anteriores y estimación de las intensidades para, en este caso, cada gen para cada área en su correspondiente estado (control o experimental).

En cuarto lugar, una vez llevado a cabo, se representaron de nuevo gráficamente los resultados para un control de calidad de estos, comparándolos con los anteriores de los datos en crudo, viendo una clara diferencia.

2. *Análisis estadísticos para la obtención de los genes DE* (paquete Limma)

Para la identificación de los **genes DE**, se decidió en primer lugar analizar cada una de las áreas por separado con sus correspondientes estados, debido a que tras realizar un gráfico MDS y ver que cada una de las áreas suponía una fuente de variedad de los datos muy grande, siendo por tanto preciso incidir principalmente en las diferencias entre haber realizado la prueba cognitiva las ratas o no.

Así pues, utilizando los datos extraídos de la función RMA anterior, se creó un modelo en el que se iban a comparar el ser control o experimental, y se fue aplicando a cada una de las áreas gracias a la **interacción área-estado** establecida previamente. Mediante la aplicación de la función de hacer contrastes de la *librería Limma*, junto con la comparación de los dos grupos a través de un *empirical Bayes* (para reducir la desviación estándar), se pudo obtener un listado con todos los genes ordenados por su significación o pvalue, siendo para cada área, guardadas esas significaciones en un documento `.txt`.

Los resultados obtenidos fueron posteriormente comparados entre las dos áreas que mostraron genes significativos con un umbral de 0.05 (es decir, para las áreas CA3 y DG), comprobando que los genes DE obtenidos en el área CA3, también fueron DE en el área de DG.

3. *Búsqueda de la significancia biológica de los resultados obtenidos* (paquete topGO)

Finalmente, se llevó a cabo una **prueba de enriquecimiento de términos GO**, junto con los datos de los genes DE, empleando el estadístico de la *prueba de Fisher*. Para ellos, primero se creó un objeto tipo **topGOdata** con todos los ID de los genes DE y sus puntuaciones, con las anotaciones GO, la estructura jerárquica GO y el resto de información necesaria para llevar a cabo este análisis.

Con la *librería biomaRt* se obtuvieron los símbolos de los genes y los GO asociados, convirtiendo primero los genes DE a un ID comparable y después al conjunto de genes con los que hemos trabajado la ID comparable con sus GO asociados. Finalmente graficando el resultado de la prueba de enriquecimiento y los **principales procesos biológicos más significativos** asociados a estos genes DE.

IV. RESULTADOS Y DISCUSIÓN

Tras realizar una primera visual y ver que los datos se encontraban correctos para trabajar con todos ellos, lo podemos ver en la Fig 1.

FIG. 1. Calidad de los datos

La normalización consiguió que fueran comparables y se pudieran establecer relaciones y análisis evitando datos disruptivos propios de la recogida o de la disparidad en la medición por ser diferentes áreas o estados, siendo posible observar, una vez han sido normalizadas, homogeneidad en todas las muestras, a través de los gráficos de boxplot (Fig 2), histograma (Fig 3) y clúster jerárquico (Fig 4). En este último, a parte, es interesante ver cómo es capaz de distinguir fácilmente las tres áreas cerebrales estudiadas.

Tras decidir hacer los análisis de manera separada debido a la clara diferencia entre las áreas (Fig 5) dejando el primer análisis a las diferencias con los estados.

Se obtuvieron en efecto, **317 genes DE coincidentes tanto en el área CA3 como en el área DG**, habiendo sido descartada el área de CA1 debido a no superar el

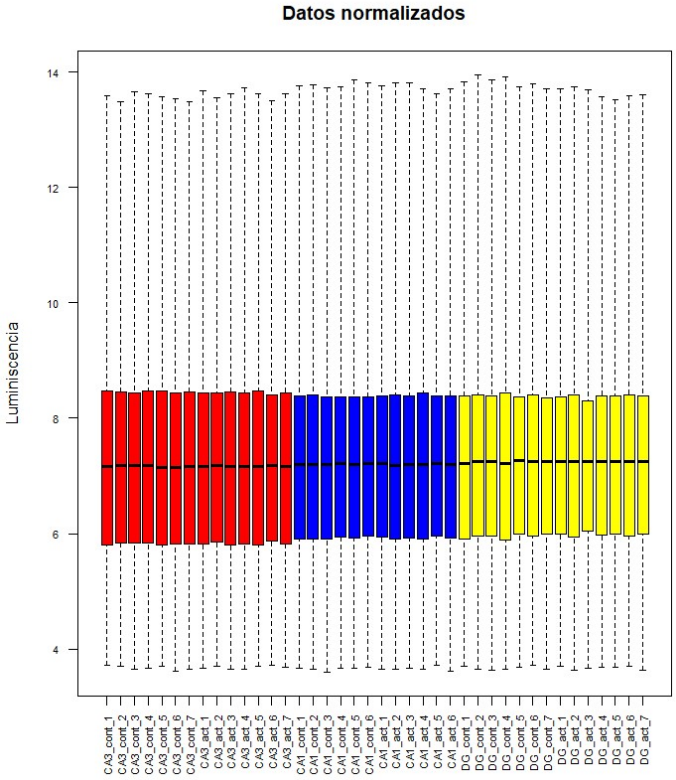


FIG. 2. Boxplot datos normalizados

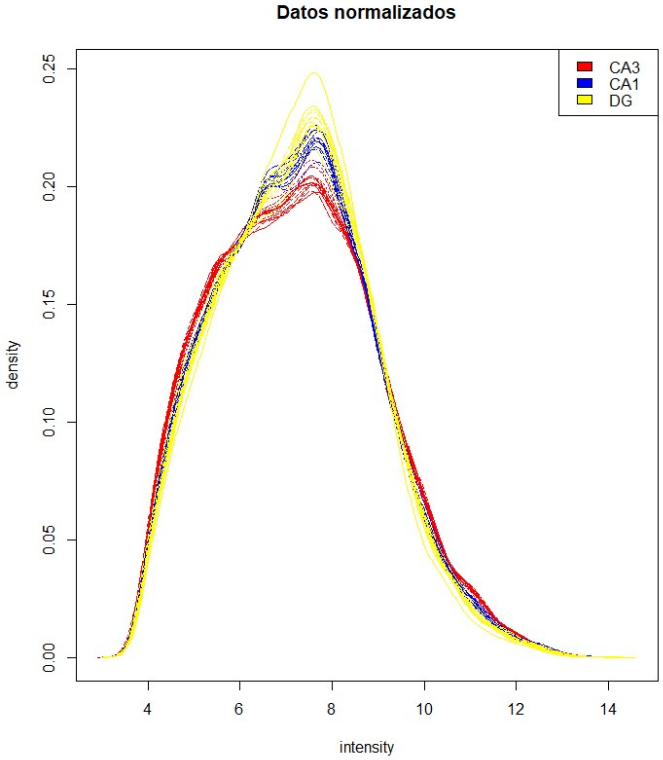


FIG. 3. Histograma datos normalizados

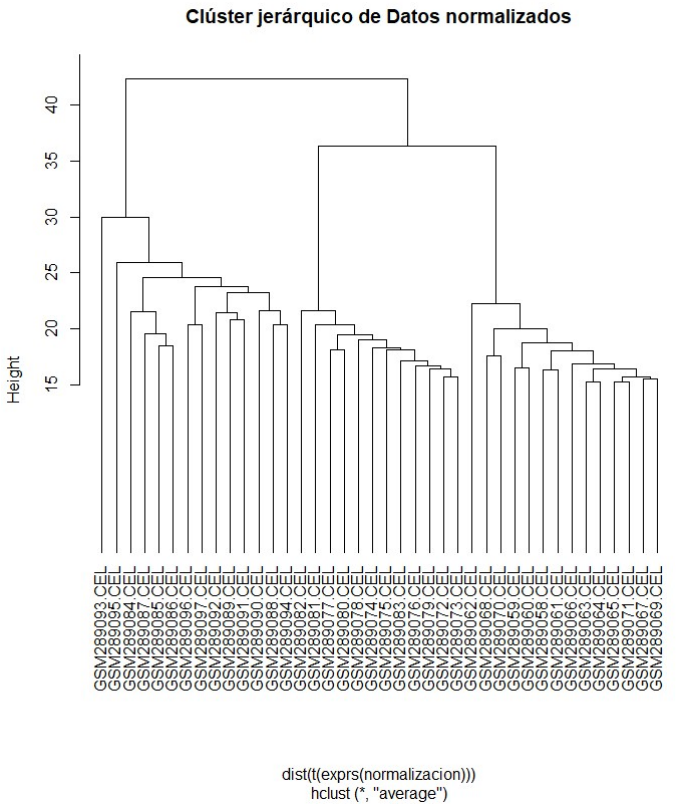


FIG. 4. Clúster datos normalizados

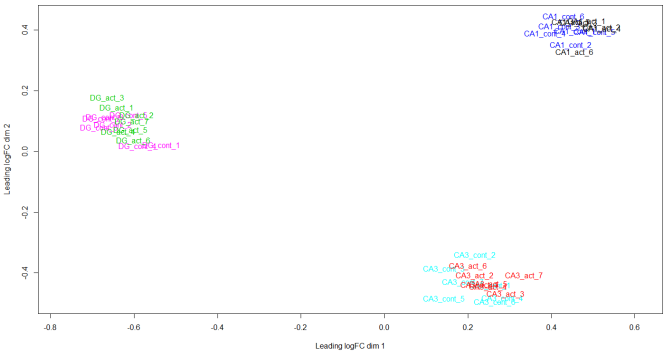
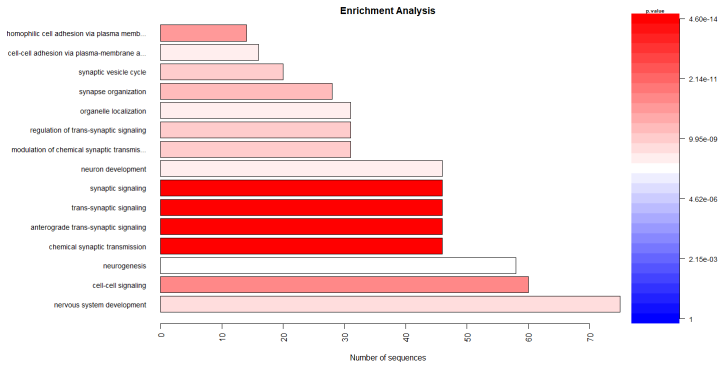


FIG. 5. Comparativa de los grupos de muestras

umbral de 0.05 la expresión diferencial de sus genes en los diferentes estados.

Una vez obtenido esto, se llevó a cabo la prueba de enriquecimiento para comprobar si éstos iban a corresponder a procesos biológicos relacionados con las hipótesis establecidas. Obteniendo así **258 genes significativos**, con **377 términos GO** y proceso biológico asociado. Destacando los 15 primeros para que fuera más sencilla la visualización, como muestra la Fig IV



Prueba de enriquecimiento

V. CONCLUSIÓN

Aquí podemos ver cómo los procesos biológicos más sobreexpresados serán los relacionados con la **señal sináptica**, la **señalización trans-sináptica**, la **señalización anterógrada sináptica** y la **transmisión química sináptica**, entre otros procesos también relacionados con la sinapsis y el procesamiento de transmisión de información. En definitiva, que tal y como se comenta en el paper asociado a estos datos, los genes DE en la región CA3 van a indicar que esta región va a mostrar mecanismos de remodelación sináptica, lo que puede servir como base para la codificación rápida de nueva información en la memoria a largo plazo.

[1] R. P. Haberman, H. J. Lee, C. Colantuoni, M. T. Koh, and M. Gallagher, Rapid encoding of new information alters the profile of plasticity-related mrna transcripts in the hippocampal ca3 region, Proceedings of the National Academy of Sciences **105**, 10601 (2008).

[2] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, affy—analysis of affymetrix genechip data at the probe level, Bioinformatics **20**, 307 (2004).

[3] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, limma powers differential expression analyses for rna-sequencing and microarray studies, Nucleic acids research **43**, e47 (2015).

[4] A. Alexa and J. Rahnenführer, Gene set enrichment analysis with topgo, Bioconductor Improv **27** (2009).