Desregulación molecular inducida por el alcohol durante la diferenciación neural de las células madre embrionarias humanas

Marina Ballesteros

4/26/2020

Contents

#Abstract

El estudio ha realizado un análisis de microarray de expresión génica estudiando la diferenciación de células precursoras neurales desde células madre embrionarias (ESC) en presencia o ausencia del tratamiento con etanol (EtOH). Los perfiles transcriptómicos de todo el genoma identificaron las alteraciones moleculares inducidas por la exposición al etanol durante la diferenciación neuronal de las ESC en rosetas neuronales y poblaciones de células precursoras neuronales.

Los datos y el código del análisis se encuentran en el siguiente repositorio github [https://github.com/marinabf93/Effect-of-alcool-in-in-ESC-differentiation-]

#Objetivos

En este estudio se pueden diferenciar dos claros objetivos:

a) Demostrar los posibles efectos teratogénicos del alcohol en el desarrollo fetal y, por tanto, los consecuentes defectos de desarrollo debidos al abuso del alcohol durante la gestación.

b) Determinar los mecanismos específicos por los que el alcohol media estas lesiones, demostrando que el alcohol tiene un efecto significativo en los mecanismos reguladores moleculares y celulares de la diferenciación de las células madre embrionarias (ESC), incluidos los genes que intervienen en el desarrollo neuronal.

#Materiales

El método de este trabajo ha sido analizar bioinformáticamente los datos de un experimento con microarrays. El experimento en el que he basado mi análisis se recoge en en dos artículos:

a) "Molecular effect of ethanol during neural differentiation of human embryonic stem cells in vitro" [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4114725/]

b) "Alcohol-Induced Molecular Dysregulation in Human Embryonic Stem Cell-Derived Neural Precursor Cells" [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5040434/]

##Diseño experimental

El **tipo de experimento** corresponde al análisis de microarrays, donde a través del diseño de un experimento se intenta responder a las cuestiones biológicas planteadas en los objetivos. Con el uso de la estadística y las diferentes herramientas bioinformáticas, se pretende procesar, analizar, visualizar y analizar los datos con el fin de responder a las cuestiones biológicas de partida.

En el estudio se investiga el efecto del alcohol (EtOH) en el desarrollo de células madre neurales derivadas desde células madre de embriones humanos. La metodología del experimento es la siguiente:

Primero se cultivan células madre embrionarias (ESC) durante cinco días en un medio de inducción neural (NIM). A continuación, los agregados neuronales que se formaron fueron sembrados en placas recubiertas con poli-L-ornitina/laminina y cultivados con NIM durante siete días para desarrollar la estructura de roseta neuronal. Al día siguiente de colocar los agregados neurales en dichas placas, se produjo el tratamiento con 20mM de EtOH. Las células fueron alimentadas con medio fresco todos los días alternando el tratamiento con 20 mM de etanol durante un día y dejando otro día de reposo sin tratamiento. Después de siete días, las rosetas neurales fueron desalojadas y luego replateadas en medio NIM para la expansión de las células precursoras neurales (NPC) durante 5 días, siguiendo el mismo procedimiento para el tratamiento con 20mM de EtOH.

##Datos

El material en el que he basado mi análisis ha sido descargado desde la página del Gene Expression Omnibus (GEO). El GEO es un depósito de datos públicos de genómica funcional donde se aceptan datos basados en matrices y secuencias. Además, se proporcionan herramientas para ayudar a los usuarios a consultar y descargar experimentos y perfiles de expresión génica corregidos.

El conjunto de datos utilizados en este análisis se identifica con el número de adhesión: **GSE56906**:[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56906]. En este caso se han analizado 10 muestras distintas.

El **tipo de microarray** utilizado ha sido del tipo Affymetrix Human Genome U133 Plus 2.0 Array, cuyo fabricante es *Affymetrix* uno de los principales vendedores de tecnología de microarray.

##Software

Para comenzar el análisis se necesita instalar **R statistical software** el cual permite hacer análisis estadísticos, representaciones gráficas y lectura y creación de documentos en diferentes formatos. El software se puede descargar en la página web [https://cran.r-project.org/index.html] y solo deben seguirse las instrucciones indicadas en función del tipo de software del ordenador que se utilice para el análisis. El análisis de microarray que se presenta en este informe ha sido desarrollado con la versión 3.6.2 y todos los análisis se han llevado a cabo con la interfaz *RStudio*. Esta interfaz puede descargarse desde la página principal[https://www.rstudio.com/]

#Métodos: Procedimiento general del análisis ("Workflow")

El flujo de trabajo se resumen en la imagen *Workflow* dentro del directorio **figures**. Esta imagen ha sido obtenida de los materiales del curso dentro del siguiente repositorio github [https://github.com/ASPteaching/Omics_Data_Analysis-Case_Study_1-Microarrays].

A continuación resumiré de forma muy general los métodos utilizados en cada paso del flujo de trabajo, así como los datos de entrada y los de salida. El desarrollo detallado de cada paso del análisis lo encontraréis en el archivo **Pipeline del análisis.Rmd** dentro del directorio principal del repositorio indicado al inicio de este informe.

Antes de empezar con el análisis y a manejar la enorme cantidad de datos y ficheros que ello conlleva, crearé tres carpetas para la organización del mismo:

- La carpeta principal del análisis será "Effect of alcool in ESC differentiation", la cual también será mi directorio de trabajo.
- Una carpeta llamada **data** para almecenar todo tipo de datos del experimento y en los cuales basaré mi análisis. En esta carpeta guardaré los archivos *.CEL* y el archivo *targets*, en el cual se decribirán los factores de estudio y sus niveles.
- En la carpeta **results** guardaré todos los resultados obtenidos en el análisis.
- La carpeta **figures** servirá para almacenar todo tipo de imágenes y figuras generadas durante el análisis.

##Identificación de los grupos y clasificación de las muestras

En este caso se han introducido los 10 archivos **.CEL** correspondientes a las 10 muestras de partida del experimento. Estos archivos contienen los datos en crudo originados tras el escaneo y preprocesado de los microarrays. Además, se ha leído el archivo **targets.csv**, el cual ha sido creado manualmente y en el que se incluye la información de los diferentes grupos y variables.

Tras la lectura conjunta de ambos archivos, se crea un objeto **rawData** con el fin de resumir toda la información anterior. El resultado es la siguiente tabla:

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

FileName	Group	CellType	Treatment	ShortName
GSM1371025	H1 EtOH 0 Rosett	Roseta diferenciada	Sin alcohol	Rosett.0.1
GSM1371026	H1 EtOH 0 Rosett	Roseta diferenciada	Sin alcohol	Rosett.0.2
GSM1371027	H1 EtOH 20 Rosett	Roseta diferenciada	Con alcohol	Rosett.20.1
GSM1371028	H1 EtOH 20 Rosett	Roseta diferenciada	Con alcohol	Rosett.20.2
GSM1371029	H1 p40 Und	ESC	Sin alcohol	P40.1
GSM1371030	H1 p40 Und	ESC	Sin alcohol	P40.2
GSM1371031	NPC H1 EtOH 0	NPC diferenciada	Sin alcohol	NPC.0.1
GSM1371032	NPC H1 EtOH 0	NPC diferenciada	Sin alcohol	NPC.0.2
GSM1371033	NPC H1 EtOH 20	NPC diferenciada	Con alcohol	NPC.20.1
GSM1371034	NPC H1 EtOH 20	NPC diferenciada	Con alcohol	NPC.20.2

##Control de calidad de los datos crudos

El objetivo de esta etapa es saber si los datos en crudo tienen la calidad suficiente como para ser normalizados. El control de calidad se lleva a cabo con el paquete **ArrayQualityMetrics**, este paquete efectúa distintos análisis con el fin de identificar los valores outliers a través de valores umbrales predefinidos.

Los datos de entrada son los datos en crudo *rawData* y se devuelve un informe del control de calidad llamado *index.html* guardado en el directorio **results**. En este informe se encontrarán boxplot de intensidad, análisis de componentes principales y MA plots entre otros. El criterio para eliminar un array del experimento es que este debe ser marcado tres veces como outlier. En la *Tabla 2* se muestra el resumen del control de calidad de los datos crudos.

##Normalización

El objetivo de la normalización es hacer comparables los arrays entre sí además de eliminar cualquier variabilidad en las muestras no debida a razones biológicas. Es decir, la normalización de los datos asegura que las diferencias de intensidades en las muestras se deban a diferencias en la expresión de los genes y no a sesgos debidos a cuestiones técnicas del experimento.

El proceso consta de tres etapas: eliminación del ruido de fondo, normalización y sumarización de los datos. Los tres procesos se llevan a cabo gracias al método **Robust Multichip Analysis** a través de la función *rma*.

Los datos de entrada son nuevamente los datos crudos *nuevame_rawData* y como output se obtiene un ExpressionSet llamado **eset_rma**, que contiene los datos normalizados.

##Control de calidad de los datos normalizados

Se realiza el mismo procedimiento que para el control de calidad de los datos en crudo; sin embargo, esta vez los datos de entrada es el vector *eset_rma* y como output se obtiene nuevamente una carpeta donde se encuentra informe del control de calidad *index.html*.

Se puede comprobar en la *Figura 3* como las intensidades de todas las muestras estan alineadas debido a que en el proceso de normaliación se incluye la normalización de los cuantiles, en el que la distribución empírica de todas las muestras se establece con los mismos valores.

	array	sampleNames	*1	*2	*3	SampleTitle	Classification
☐	1	GSM1371025				H1 EtOH 0 Rosett-1	Roseta diferencida sin alcohol
☐	2	GSM1371026				H1 EtOH 0 Rosett-2	Roseta diferencida sin alcohol
☐	3	GSM1371027				H1 EtOH 20 Rosett-1	Roseta diferencida con alcohol
☐	4	GSM1371028				H1 EtOH 20 Rosett-2	Roseta diferencida con alcohol
☐	5	GSM1371029				H1 p40 Und-1	ESC indiferenciada
☐	6	GSM1371030				H1 p40 Und-2	ESC indiferenciada
☑	7	GSM1371031			x	NPC H1 EtOH 0-1	NPC diferenciada sin alcohol
☐	8	GSM1371032				NPC H1 EtOH 0-2	NPC diferenciada sin alcohol
☑	9	GSM1371033			x	NPC H1 EtOH 20-1	NPC diferenciada con alcohol
☐	10	GSM1371034				NPC H1 EtOH 20-2	NPC diferenciada con alcohol

Figure 1: Tabla resumen del archivo index.html, generdo por el paquete arrayQualityMetrics en los datos crudos

	array	sampleNames	*1	*2	*3	Group	CellType	Treatment
☐	1	Rosett.0.1				H1 EtOH 0 Rosett-1	Roseta diferenciada	Sin alcohol
☐	2	Rosett.0.2				H1 EtOH 0 Rosett-2	Roseta diferenciada	Sin alcohol
☐	3	Rosett.20.1				H1 EtOH 20 Rosett-1	Roseta diferenciada	Con alcohol
☐	4	Rosett.20.2				H1 EtOH 20 Rosett-2	Roseta diferenciada	Con alcohol
☑	5	P40.1	x			H1 p40 Und-1	ESC	Sin alcohol
☑	6	P40.2	x			H1 p40 Und-2	ESC	Sin alcohol
☐	7	NPC.0.1				NPC H1 EtOH 0-1	NPC diferenciada	Sin alcohol
☐	8	NPC.0.2				NPC H1 EtOH 0-2	NPC diferenciada	Sin alcohol
☐	9	NPC.20.1				NPC H1 EtOH 20-1	NPC diferenciada	Con alcohol
☐	10	NPC.20.2				NPC H1 EtOH 20-2	NPC diferenciada	Con alcohol

Figure 2: Tabla resumen del archivo index.html, generdo por el paquete arrayQualityMetrics en los datos normalizados

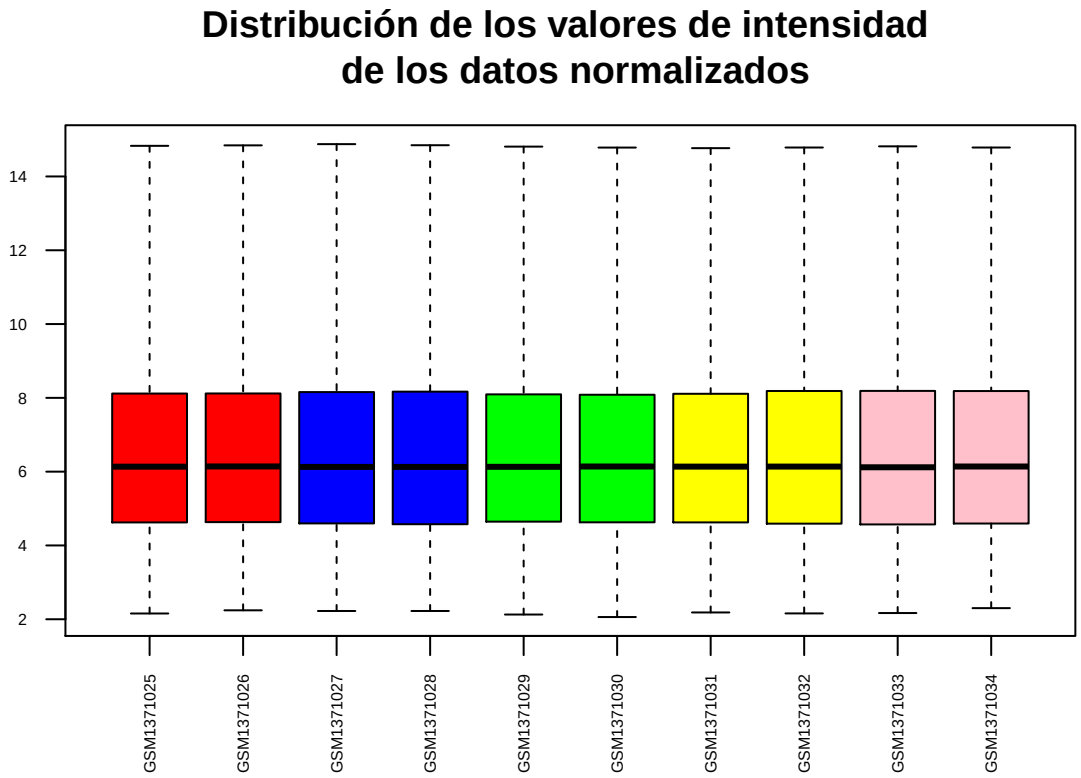

Figure 3: Boxplot para las intensidades de los arrays (Datos Normalizados)

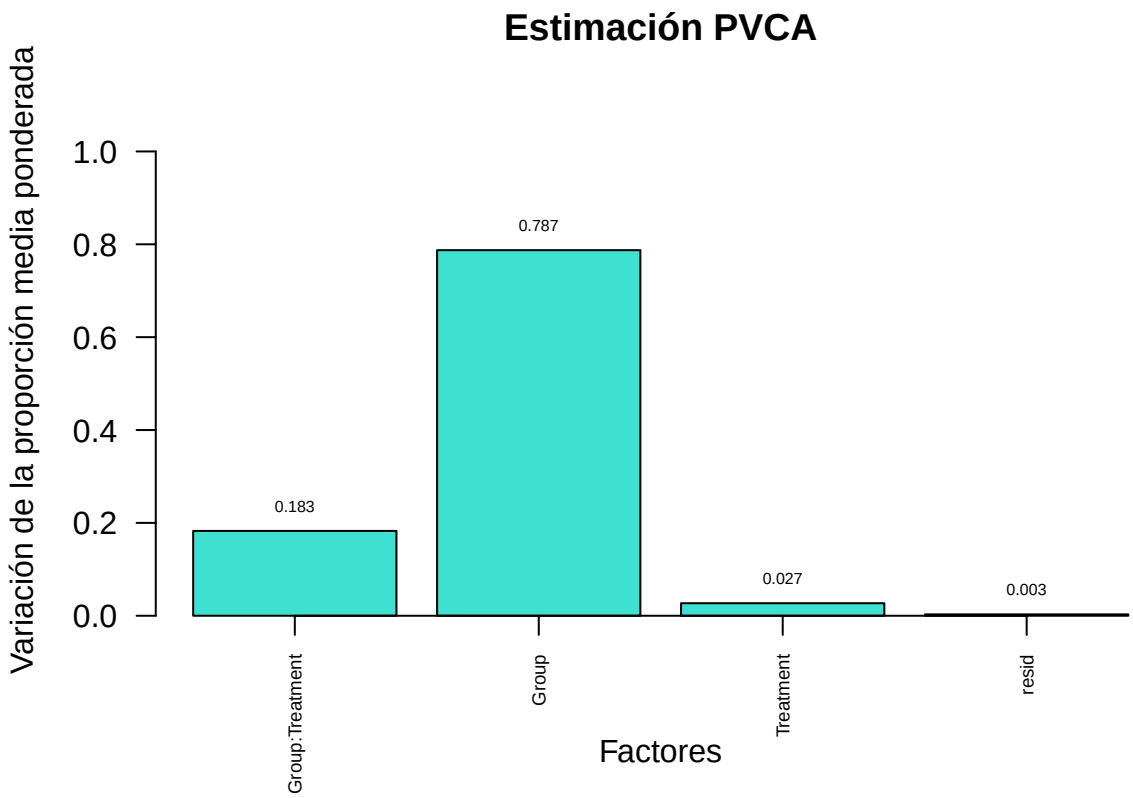

Figure 4: Relative importance of the different factors -genotype, temperature and interaction- affecting gene expression