ANÁLISIS DE DATOS ÓMICOS

Primera prueba de evaluación continua

Autor: Julián David Sánchez Bautista, Médico y cirujano, candidato a especialista en Medicina Interna.

Título de trabajo original: Genome-wide CpG island methylation analysis in non-small cell lung cancer patients [Affymetrix expression data]. **GEO:** GSE32496

RESUMEN:
La metilación del ADN hace parte de la regulación epigenética, la cual es relevante para la patogénesis del cáncer. Se evaluó la expresión de genes después de administrarse 5-asa-2'-deoxycytidine (Aza-dC) y/o trichostatin A (TSA), comparado con células sin tratamiento, en 3 líneas celulares de tumor de células no pequeña pulmonar (A549, H1993, H2073)(1).

TABLA DE CONTENIDO:

1. OBJETIVOS:

- Analizar usando el lenguaje R los datos que se obtuvieron en los Microarrays
- Evaluar la lista de genes expresados según si no recibían tratamiento, si lo recibían o si recibían la combinación de las dos drogas

2. MATERIALES Y MÉTODOS:

Se realizará una descripción de las herramientas usadas y los procedimientos para la obtención de resultados. Los detalles exactos y el código usado se presentan como adjuntos.

2.1 Descripción de los datos empleados:

En la web Gene Expression Omnibus, se realizó la búsqueda. Lo datos usados para el análisis fueron descargados de la plataforma **GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array,** con el **GEO:** GSE32496.

El organismo como unidad experimental fue el *Homo sapiens* y se uso un microarray de color. Con esta búsqueda se obtuvieron 18 muestras: 6 muestras con el linaje celular A549, 6 muestras con el linaje H1993 y 6 muestras con el linaje H2073. Dos de cada set de muestras correspondían a las diferentes intervenciones: 2 tratadas con 0.5 mcg Aza-dC por 6 días, 2 tratadas con Aza-dC más TSA 100 ng/mL por 6 días y 2 no tratadas.

2.2 Métodos usados para el análisis:

El software usado para el análisis fue R en su versión 4.0 para macOS; para hacer anotaciones se uso RMarkdown de R studio.
El principal gestor de paquetes usado fue **BioConductor**, y las librerías destacadas usadas para el análisis fueron:

Limma: generación de modelos lineales, análisis y procesamiento de datos, y expresión diferencial para los microarrays.

Clusterprofiles: análisis de significancia biológica

2.3 Descripción de los métodos usados:

2.3.1 Importación de datos:

Se extrajeron las muestras en formato .CEL usando el código **GEO:** GSE32496, de la web Gene Expression Omnibus, a través de la librería **GEOquery** y se generó una tabla para observar los grupos y qué grupo pertenecía cada muestra.

Normalmente seguido de la extracción de datos crudos y observar las características de los grupos y a qué grupo pertenecían se debe realizar normalización de estos. En el caso de este trabajo en particular, al hacer extracción con la función GEOquery, se hizo la extracción de los datos ya normalizados; para comprobar esto se uso la función *class()* ***que arrojó output*** *como ExpressionSet.*

2.3.2 Control de calidad de datos:

Para obtener representación visual de los datos para asegurarnos que estén correctamente normalizados se realizó este paso. Para hace el control de calidad de datos se uso la librería

arrayQualityMetrics, se generaron plot y se realizó una gráfica de análisis de componentes principales. (las demás gráficas están adjuntas en el RMarkDown y el arrayQualityMetrics report):

2.3.3 Obtención de genes diferencialmente expresados:

Inicialmente se creó una matriz para asignar cada grupo según el tipo de intervención o sino tenía tratamiento. Posteriormente se creó la matriz con los contrastes entre los grupos y a partir de esta, con la librería ***lima*** se realizaron contrastes y con ***ebayes*** comparaciones entre los grupos, obteniéndose así el listado de genes ordenados por su significancia estadística o valor p. Con estos resultados se compararon ntre los grupos con ajuste *fdr*. Para la visualización de la expresión diferencial se realizó un *volcanoplot*.

2.3.4 Anotación de los resultados

Una vez obtuvimos la tabla con los genes diferencialmente expresados, se proveyó información de características que han sido seleccionadas, para asociarlas con los identificadores que aparecen en los datos. Se usaron símbolos (*symbol*), los id (*entrezid*) y los nombres (*genename*).

2.3.5. Comparaciones múltiples

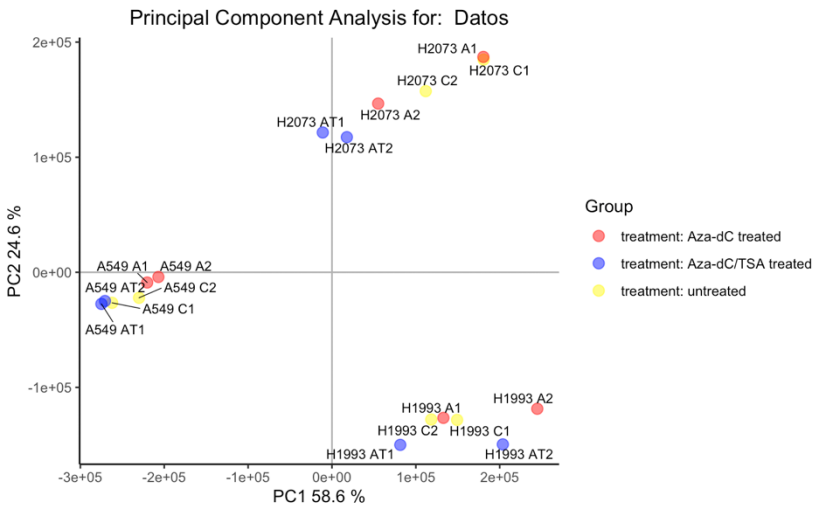
Para conocer los genes que han sido seleccionados en cada comparación se realizaron comparaciones múltiples usando la librería **limma** y se gráfico usando un *venndiagram.*

2.3.6. Significancia biológica de los resultados obtenidos.

Se llevó a cabo una prueba de enriquecimiento de datos usando la librería **clusterprofiles**, se compararon los grupos con la función *enrichKEGG.* Finalmente, se gráficaron los principales procesos biológicos más significativos asociados a los genes diferencialmente expresados.

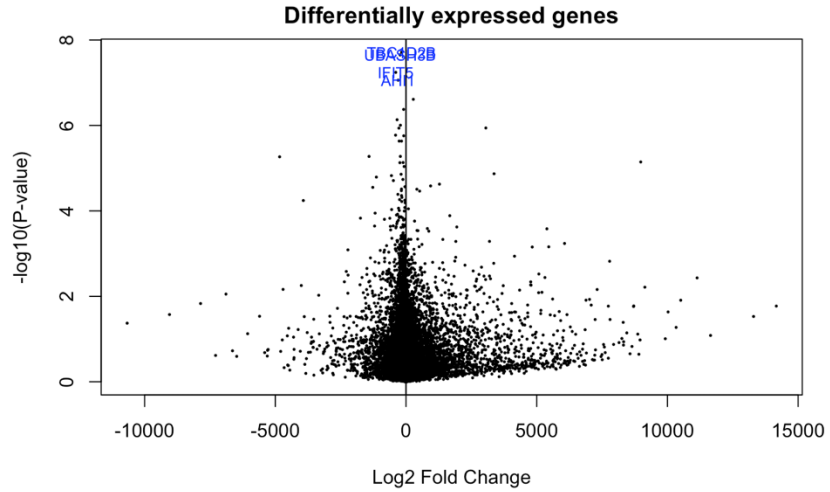
3. RESULTADOS Y DISCUSIÓN

Al ver una primera aproximación gráfica de los resultados, se observó que eran aptos para trabajar con ellos. Se realizó un análisis de componentes principales (gráfica 1), donde se observa que la contribución a estos va a ser más por el tipo celular que por el tipo de intervención. Las demás gráficas están en el archivo adjunto de *arrayQualityMetrics*.



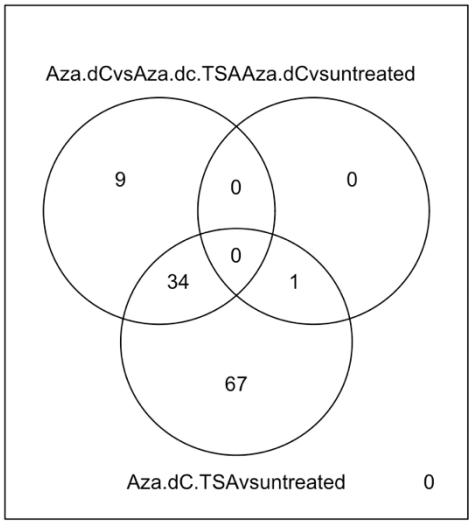
Gráfica 1. Análisis de componentes principales

Cuando se obtuvieron los genes diferencialmente expresados se obtuvo de forma gráfica un volcanoplot. Gráfica 2. Vemos que la mayoría de los genes no cumplieron la significancia estadística.



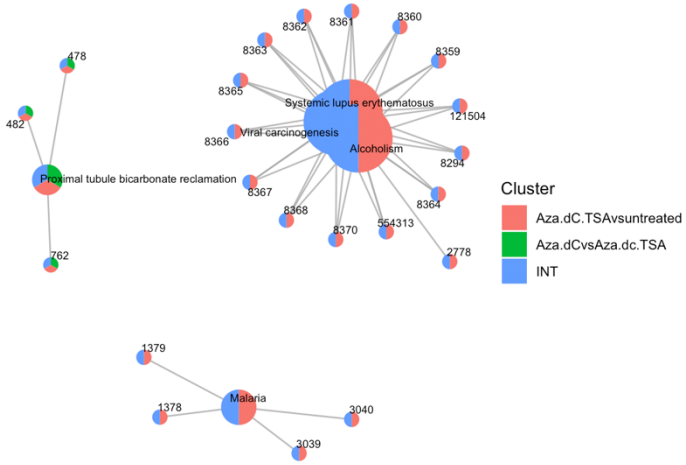
Gráfica 2. Genes diferencialmente expresados

Al realizar la comparación de múltiples grupos, se obtuvo de forma gráfica la información de qué genes diferencialmente expresados pertenecían a cada grupo. Gráfica 3. Vemos que la mayoría se presentaron en la comparación que contenía TSA.

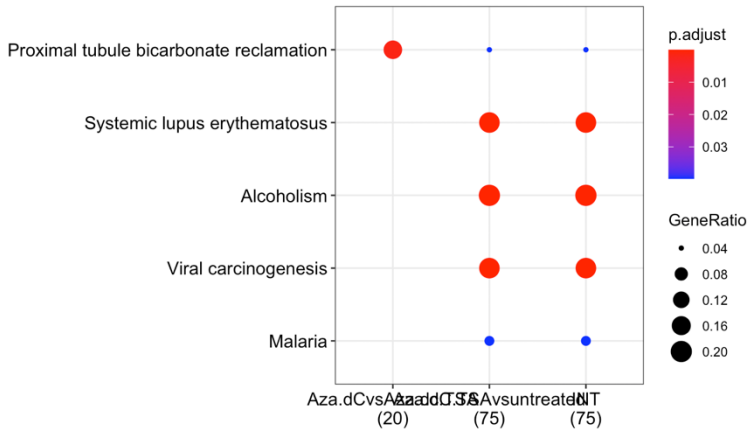


Gráfica 3. Comparación entre múltiples grupos.

Una vez obtenidos estos datos, se llevo a cabo la prueba de enriquecimiento en busca de la significancia biológica, y a que procesos biológicos correspondían estos. Gráfica 4 y Gráfica 5.



Gráfica 4. Significancia biológica



Gráfica 5. Significancia biológica con su significancia estadística.

4. CONCLUSIONES

Los procesos biológicos más sobreexpresados fueron los relacionados con la reabsorción proximal de bicarbonato, el lupus, alcoholismo y la carcinogénesis viral, procesos que se espera pasen pues están relacionados con la respuesta inmune y la exposición a medicamentos.

5. BIBLIOGRAFÍA

1. Heller G, Altenberger C, Steiner I, Topakian T, Ziegler B, Tomasich E, et al. DNA methylation of microRNA-coding genes in non-small-cell lung cancer patients. J Pathol. 2018 Aug 1;245(4):387–98.

6.ENLACE DE GITHUB

https://github.com/juliansanchez8/PEC1_ADO