PEC 1. Análisis de datos ómicos

Yeison Santamaría Alza

Tabla de contenido:

Resumen

El linfoma es una neoplasia de linfocitos que puede comprometer el tejido linfático y extra linfático. Típicamente se clasifican en Linfoma tipo Hodgkin y no Hodgkin(1). Dentro de los no Hodgkin se encuentran los linfomas de zona marginal y dentro de este tipo existe el linfoma de zona marginal esplénico, que representa cerca del 2% del total de linfomas, el cual tiene importante compromiso del bazo(2). En los linfomas no Hodgkin se han encontrado diferentes mutaciones, sin embargo, en el linfoma de zona marginal esplénico se ha encontrado una mutación genética característica: la deleción 7q, la cual podría tener implicaciones en la oncogénesis de este tipo de tumor(3).

1. Resumen del estudio elegido

Usando el buscador de Gene Expression Omnibus (GEO) del Instituto Nacional de Salud de Estados unidos, seleccioné el estudio llamado "An integrated genomic and expression analysis of 7q deletion in splenic marginal zone lymphoma (Affymetrix HG-U133plus2 gene expression microarray)", con el código de acceso **GSE35426.** El estudio fue realizado por un grupo colaborativo que incluyó centros en Estados Unidos, China, Suiza y Reino Unido. Los datos que se encontraron en GEO, correspondieron a 24 muestras, de las cuales 14 fueron de pacientes con linfoma de zona marginal esplénico (LZME), 5 de linfoma folicular (LF) y 5 de linfoma de células del manto (LCM). El total de muestras del estudio no están incluidas en esa base de datos.

Teniendo en cuenta hallazgos previos del mismo grupo, en donde se encontró que la deleción 7q32 es una alteración frecuente en el LZME, el objetivo del estudio fue evaluar la expresión diferencial de genes en tres tipos de linfoma (LZM, LF y LCM), para determinar si alguno de estos genes correspondiera a un gen supresor de tumor. En los resultados que publicaron los autores, los cuales incluyen más muestras que las que se contó para hacer la práctica de análisis, se encontró un número de genes con expresión a la baja en pacientes con la deleción mencionada previamente, de los cuales varios presentaron hipermetilación. También se encontraron 8 miRNA con expresión a la baja en pacientes con LZME y deleción 7q32 y 3 miRNA con expresión a la baja en pacientes con LZME en comparación con otros tipos de linfoma (4).

2. Objetivo de la práctica:

- Realizar el proceso de extracción de datos de GEO y exportarlos a R
- Ejecutar el análisis estadístico de Microarray
- Encontrar los genes diferencialmente expresados en tres tipos de linfoma (LZME, LF y LCM)
- Determinar la significancia biológica de los resultados

3. Materiales y métodos

Para una mejor comprensión de este apartado, se explicarán inicialmente los métodos del estudio tomado y posteriormente se mostrará en detalle los métodos del análisis realizado en la práctica.

3.1 Métodos del estudio

En el estudio se obtuvieron los datos de 95 muestras de pacientes con LZME. Todos los empecines fueron de *homo sapiens.* Para la extracción de RNA se empleó el kit RNEse en muestras de tejido congelado que tuviera más del 70% de células tumorales. La integridad del RNA fue evaluada mediante el bionalizador Agilent 2100. Luego se realizó la síntesis de cDNA con RNA de 2 µg usando GeneCHip, luego se realizó transcripción in vitro con nucleótidos biotinilados. Finalmente se purificó cRNA biotinilado y se hibrido a Affimetrix HG-U11 Plus 2.0.

Dado que los microarrays utilizados fueron de un solo color (Affimetrix), estos fueron analizados en la plataforma Affimetrix HG-133 Plus 2.0 para evaluar la expresión diferencial de genes en pacientes con LZME con y sin deleción 7q. En un apartado del análisis también incluyen pacientes con LF y LCM, pero en los métodos no mencionan el momento en que fueron incluidos.

3.2 Métodos del análisis realizado en la práctica

Como se mencionó previamente, los datos incluidos en el informe final de los autores difieren en cantidad a los que se encuentran en GEO. Por lo tanto, para esta práctica se utilizaron 24 muestras distribuidas de la siguiente manera:

- 14 muestras provenientes de pacientes con LZME
- 5 muestras provenientes de pacientes con LF
- 5 muestras provenientes de pacientes con LCM

El análisis se realizó usando el software RStudio, con lenguaje de R, en la versión 4.0 para Windows. Para toda la ejecución del proceso se empleó RMarkdown. El gestor de librerías usado fue Bioconductor. Las librerías de Bioconductor usadas para el análisis fueron:

- affy
- arrayQualityMetrics
- lima
- GEOquery
- ClusterProfiler

A continuación, se describe el paso a paso de la actividad realizada:

3.2.1 Importación de datos

Este proceso es posible realizarlo de dos formas: mediante importación de los archivos .CEL descardados de GEO o directamente de la página principal GEO. La diferencia entre estas dos formas es que, con la segunda, los datos obtenidos se encuentran normalizados. El proceso realizado fue el siguiente:

3.2.1.1 Importación desde archivos descargados con extensión .CEL

Para este proceso es necesaria la librería affy y con el comando ReadAffy ( ), se realiza la importación de los datos. Luego es necesario renombrar el fenotipo de los datos, que en este caso fue identificar a que tipo histológico de linfoma correspondía cada caso. Estos datos importados son objeto de clase AffyBatch

3.2.1.2 Importación de datos directamente de la página GEO

Para esta importación se requiere la librería GEOquery. En la página del estudio en GEO, se da click en el botón "Analyze with GEO2R" y posteriormente en el botón R script. En ese apartado aparece el comando requerido que debe correrse en RStudio. En este caso el comando usado fue getGEO ( ). Los datos obtenidos son objeto clase ExpressionSet

3.2.2 Control de datos no normalizados

Para este apartado, lo primero que realicé fue un boxplot para evaluar la distribución de las muestras, usando la función boxplot ( ).

Posteriormente usando la librería y la función con nombre arrayQualityMetrics ( ), se realizó el análisis de datos no normalizados.

3.2.3 Normalización de datos

Para la normalización se utilizó la función rma ( ) de la librería affy. La normalización se efectuó en los datos con característica AffyBatch, ya que los que tienen característica ExpressionSet ya se encontraban normalizados.

Luego se realizó análisis de control de datos normalizados, tal como se realizó para los no normalizados y se análisis de componentes principales.

3.2.4 Identificación de genes diferencialmente expresados

Para este apartado se requiere el uso de la librería limma

3.2.4.1 Diseño de la matriz

Primero se diseña una matriz con la función model.matrix que incluya la variable donde se encuentran los tipos histológicos de linfoma de la base de datos normalizados, así mismo se le da nombre a las columnas para identificarlas posteriormente

3.2.4.2 Matriz de contraste

Para el diseño de la matriz de contraste se usó la función makeContrast ( ), en donde se incluyeron las comparaciones entre los tres tipos de linfoma y la intersección entre ellos

3.2.4.3 Modelo de estimación y selección de genes

Una vez se tienen las dos matrices, se puede realizar el modelo de estimación y la selección de los genes diferencialmente expresados. Para esto se requiere la librería limma y las funciones lmFit ( ), contrasts.fit ( ) y eBayes ( ). El resultado obtenido es un objeto de clase MArrayLM.

3.2.4.4 Obtención de lista de genes diferencialmente expresados

Para este objetivo se obtuvo la lista con la función topTable para cada una de las comparaciones realizadas en el estudio, así como de la intersección de las tres comparaciones: LF vs MCL, FL vs SMZL y MCL vs SMZL. El resultado obtenido es una tabla en donde se muestran las estadísticas más importantes de los genes que se encuentran diferencialmente expresados en cada comparación.

3.2.4.5 Anotación de resultados

Una vez que se obtienen los genes diferencialmente expresados en cada una de las comparaciones, se contrasta esta información, con lo disponible para el genoma de la especie en que se realizó el estudio, en este caso *homo sapiens.* Para esto es necesario la librería hgu133plus2. Dicha evaluación se realizó para cada una de las comparaciones del estudio.

3.2.4.6 Visualización de la expresión diferencial

Se realizó un volcanoplot para evaluar visualmente como se expresan diferencialmente los genes en los grupos del estudio. La función empleada fue volcanoplot ( )

3.2.5 Múltiples comparaciones

Para evaluar de manera simultánea las tres comparaciones realizadas previamente y la intersección, se realizó este análisis. Se requiere la librería limma y se usaron las funciones decideTests ( ) y para su visualización la función vennDiagram ( )

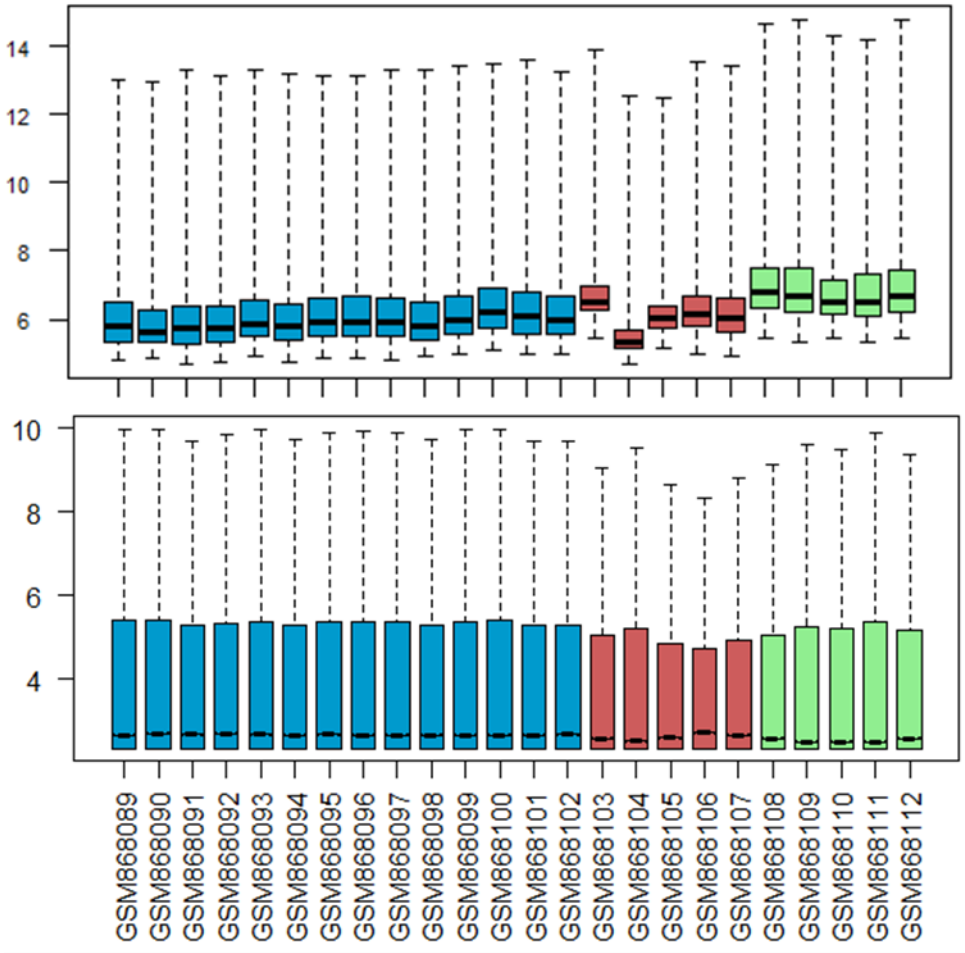
3.2.6 Análisis de significancia biológica

Lo primero que se crea es una lista de genes seleccionados usando las tablas de las tres comparaciones y de la intersección. Posteriormente re requiere de la librería clusterProfiler, en donde se obtienen los procesos biológicos que se han visto involucrados con la expresión de los genes que diferencialmente fueron encontrados en el estudio. Luego se realiza un gráfico donde se muestran los procesos más frecuentemente afectados por esos genes, en contraste con cada una de las comparaciones del estudio usando la función dotplot ( )

4. Resultados

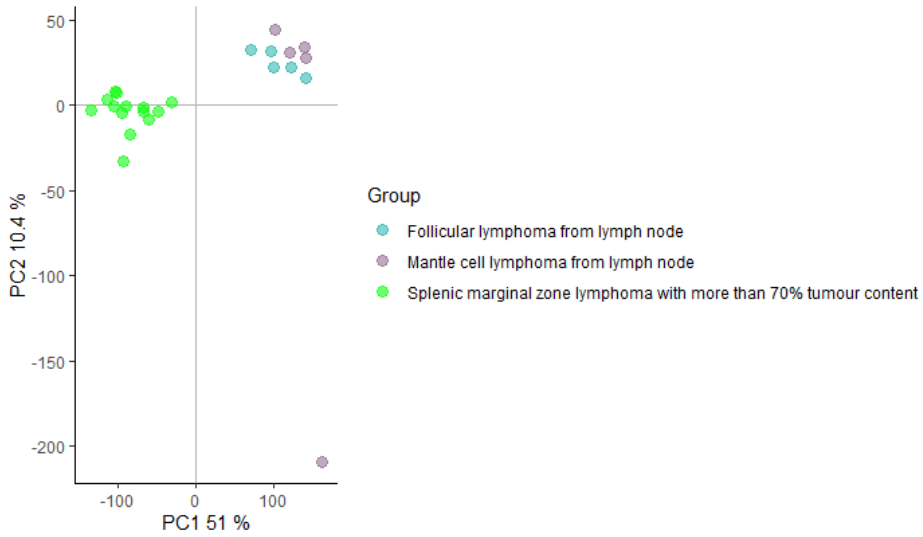
Lo primero que se realizó fue la importación de los datos. Luego de importarlos hubo necesidad de normalizarlos. En la gráfica 1 se muestra la comparación gráfica entre los datos crudos y normalizados.

Gráfica 1: Representación visual de datos no normalizados y normalizados



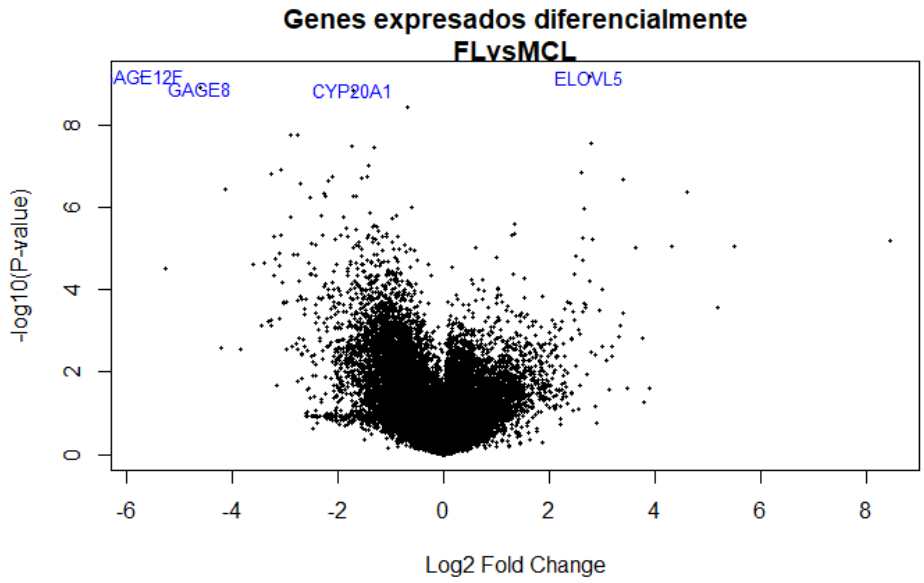
Posteriormente se realizó análisis de control de calidad de los datos, lo que se ejecutó con la función arratQualityMetrics. En el informe de incluye la imagen de análisis de componentes principales (gráfica 2), pero como adjunto se pondrán todos los análisis de calidad de datos. Con el análisis de componentes principales, se encuentra que el primer componente aporta el 51% de la variabilidad de las muestras y con la evaluación del gráfico se muestra que esta variabilidad está dada por la variedad histológica del linfoma, ya que todos los pacientes con linfoma de zona marginal esplénico se encuentran a la izquierda de la gráfica.

Gráfica 2: Análisis de componentes principales



Luego se evaluaron los genes diferencialmente expresados para cada uno de los grupos con las funciones mencionadas previamente. Para la visualización de estos resultados se realizó un volcano plot que se muestra en la gráfica 3.

Gráfica 3: Volcano plot de genes diferencialmente expresados

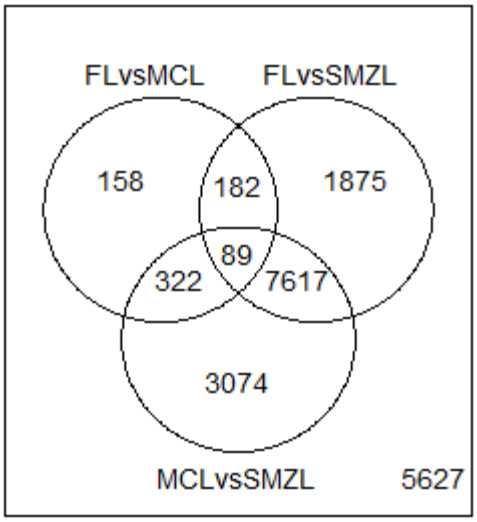


Posteriormente se quiso definir si los genes diferencialmente expresados eran comunes entre las diferentes comparaciones y la intersección de lo mismo. Para lo anterior se realizó un diagrama de Venn, en donde se encontró que entre FL y MCL hay 158 genes diferencialmente expresados, entre FL y SMZL hay 1875 genes diferencialmente expresados, entre MCL y SMZL hay 3074 genes diferencialmente expresados.

Luego al comparar las comparaciones se encontró que entre la compasión FL-MCL y FL-SMZL se encontraron 182 genes diferencialmente expresados, entre FL-SMZL y MCL-SMZL hay 7617 genes

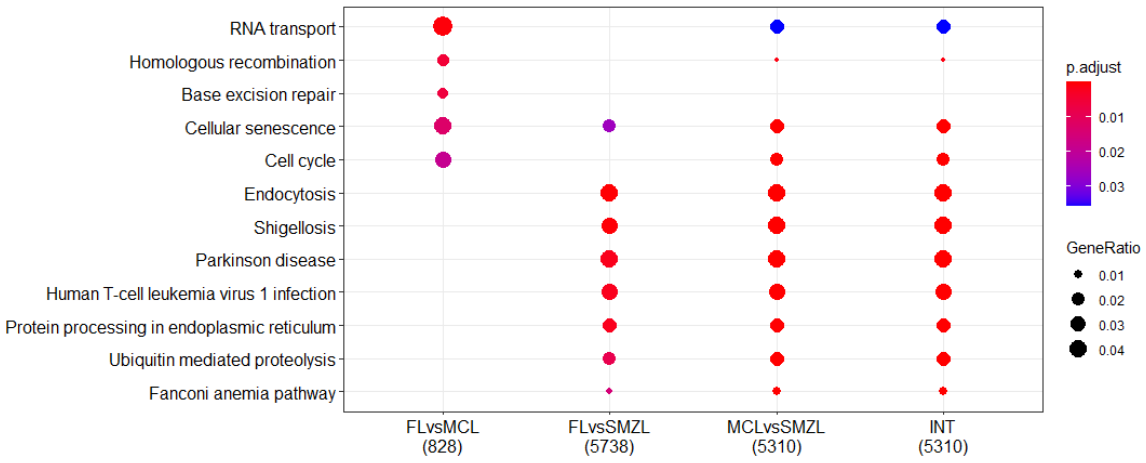
diferencialmente expresados y entre MCL-SMZL y FL-MCL se encontraron 322 genes diferencialmente expresados. Finalmente, en la intersección de las tres comparaciones se encontraron 89 genes diferencialmente expresados. En la gráfica 4, se muestra el diagrama de Venn.

Gráfica 4: Diagrama de Venn



Para finalizar en el análisis de significancia biológica, se determinó la importancia de estos genes diferencialmente expresados. Para esto se realizó un análisis con compareCluster y se graficó en un dotplot (gráfica 5)

Gráfica 5. Dot plot para análisis de significancia biológica



5. Discusión

Si bien los linfomas son proliferación anormal de linfocitos que pueden llevar a la muerte, hay diferentes tipos de linfoma, que típicamente se han evaluado según los hallazgos en histoquímica(5). La genética ha tenido aun auge muy importante en las patologías oncológicas, ya que la totalidad de las células tumorales tienen alteraciones en el código genético que causan en ellas aumento en la tasa de replicación y generación de mecanismos que las hacen resistentes a la eliminación por el sistema inmune del organismo(6).

Por lo anterior ha surgido un interés en evaluar la expresión diferencial de genes en los diferentes tumores, con el objetivo de identificar posibles alteraciones que permitan un mejor entendimiento de la fisiopatología tumoral y finalmente concluir en una terapia dirigida para ese tumor, de tal manera que pudiera generarse una mejor respuesta terapéutica.

En este estudio se evaluó la expresión génica diferencial en tres tipos de linfomas: linfoma folicular, linfoma de células del manto y linfoma de zona marginal esplénico. En los resultados es llamativo que el número de genes con expresión diferencial al comparar linfoma folicular y linfoma de células del manto es mucho menor que los encontrados cuando se compara cualquiera de los linfomas con el linfoma de zona marginal esplénico. Esto nos ayuda a pensar que genéticamente hay una mayor diferencia en el LZME comparado con los otros tipos de linfoma. Así mismo, esa diferencia genera la hipótesis de que, al ser genéticamente diferente, el comportamiento clínico y la respuesta farmacológica va a ser diferente en el LZME respecto a otros tipos de linfoma

Inicialmente se tenía la hipótesis de que los genes que diferencialmente se expresaban en LZME eran genes supresores de tumor, sin embargo, en el análisis de significancia biológica se encontraron genes relacionados con procesos como endocitosis, infección por virus humano de leucemia de células T, procesamiento de proteínas en retículo endoplásmico y proteólisis mediada por ubiquitina, se expresaron diferencialmente en las comparaciones que incluyeron LZME. Así mismo, se identificaron genes diferencialmente expresados al comparar FL y MCL en procesos relacionados con recombinación y reparación de escisión de bases. Por su parte genes relacionados con ciclo celular y transporte de RNA se encontraron diferencialmente expresados en las comparaciones que incluyeron MCL.

6. Conclusión

Al comparar la expresión génica de los tres tipos histológicos de tumores, se encontró una gran cantidad de genes que están de manera diferencial en cada tipo de tumor. Si bien en lo encontrado no hay genes relacionados con supresión de tumor expresados diferencialmente en el LZME, si se encontraron muchos otros que tienen esta condición y que tienen función clave en la función celular.

Así mismo fue posible identificar que esta expresión diferencial tiene importantes discrepancias cuando la comparación incluía pacientes con LZME en relación cuando no se incluía, lo que genera la hipótesis de que las alteraciones genéticas en este tipo de tumor son marcadamente diferentes a las que se encuentran en otro tipo de linfomas.

Enlace github: https://github.com/yeison1807/PEC1_Analisis_datos_omicos/

7. Bibliografía

1. Shanbhag S, Ambinder RF. Hodgkin lymphoma: A review and update on recent progress. CA Cancer J Clin [Internet]. 2018 Mar 1 [cited 2020 May 3];68(2):116–32. Available from: http://doi.wiley.com/10.3322/caac.21438

2. Bello A, De La Vega F, Redondo K, Riuz K, Mendoza L, Lora M. Linfoma esplénico de zona marginal. Vol. 40, Acta Médica colombiana. 2015.

3. James Watkins A, Huang Y, Ye H, Chanudet E, Johnson N, Hamoudi R, et al. Splenic marginal zone lymphoma: Characterization of 7q deletion and its value in diagnosis. J Pathol. 2010 Mar;220(4):461–74.

4. Watkins AJ, Hamoudi RA, Zeng N, Yan Q, Huang Y, Liu H, et al. An Integrated Genomic and Expression Analysis of 7q Deletion in Splenic Marginal Zone Lymphoma. El-Maarri O, editor. PLoS One [Internet]. 2012 Sep 13 [cited 2020 May 3];7(9):e44997. Available from: https://dx.plos.org/10.1371/journal.pone.0044997

5. Rao IS. Role of immunohistochemistry in lymphoma. Indian J Med Paediatr Oncol. 2010 Oct;31(4):145–7.

6. Skibola CF, Curry JD, Nieters A. Genetic susceptibility to lymphoma. Haematologica. 2007 Jul;92(7):960–9.