CARACTERIZACIÓN TRANSCRIPTÓMICA DEL CÁNCER DE ENDOMETRIO Y SU IMPORTANCIA EN EL DISEÑO DE NUEVAS ESTRATEGIAS TERAPÉUTICAS

Mónica Parra Grande
Máster en Bioinformática y Bioestadística
Área: Estadística y Bioinformática 32

Consultor: Jeroni Luna Cornadó
Profesor responsable de la asignatura: David Merino Arranz

Fecha Entrega: 5 de Junio de 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Caracterización transcriptómica del cáncer de endometrio y su importancia en el diseño de nuevas estrategias terapéuticas*
Nombre del autor:	*Mónica Parra Grande*
Nombre del consultor/a:	*Jeroni Luna Cornadó*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	06/2018
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Estadística y Bioinformática 32*
Idioma del trabajo:	*Castellano*
Palabras clave	*RNA-Seq, endometrial cancer, targeted therapy*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Contexto: El cáncer de endometrio es el tumor ginecológico más frecuente en países desarrollados. Actualmente existe una falta de opciones terapéuticas en este tipo de tumores, lo que origina un gran interés en la búsqueda de dianas terapéuticas que permitan personalizar el tratamiento en beneficio del paciente. El objetivo del proyecto fue buscar si había diferencias significativas en la expresión de los genes de pacientes con cáncer de endometrio MSI-H frente a MSI-L y MSS, con la finalidad de encontrar posibles alternativas terapéuticas.

Metodología: Para llevar a cabo el estudio se obtuvieron datos del análisis de RNA-seq de muestras de pacientes con cáncer de endometrio del TCGA. Mediante el paquete limma se realizó el análisis de expresión diferencial. Posteriormente se llevó a cabo un análisis de enriquecimiento de pathways y de co-expresión de genes mediante la aplicación ReactomeFIPlugin a través de Cytoscape versión 3.6.1. Por último, el estudio se completó con un análisis de supervivencia de ambos grupos.

Resultados: Se encontraron diferencias significativas en más de 2000 genes o transcritos. En el grupo MSi-H se encontraron genes o transcritos relacionados con pathways del ciclo celular, mientras que en el grupo MSI-L y MSS se encontraron genes o transcritos relacionados con pathways de la vía de señalización p53. En cuanto al análisis de supervivencia no hubo diferencias significativas entre ambos grupos.

Conclusiones: Según los resultados encontrados en las pacientes con cáncer de endometrio MSI-H una alternativa al tratamiento complementario con quimioterapia o radioterapia podrían ser los inhibidores de CDKs.

Abstract (in English, 250 words or less):

Background: Endometrial cancer is the most frequent gynecological tumor in development countries. Currently there is a lack of therapeutic options in this type of tumors, which gives rise to a great interest in the search for therapeutic targets that allows the treatment to be customized for the benefit of the patient. The objective of the project was to find out significant differences in the expression of the genes in patients with MSI-H endometrial cancer against MSI-L and MSS, with the objective of finding possible therapeutics options.

Methods: To carry out the study, data were obtained from the RNA-seq analysis of samples from patients with TCGA endometrial cancer. The differential expression analysis was carried out using the limma package. Subsequently an analysis of pathway enrichment and gene co-expression was carried out using the ReactomeFIPlugin application through cytoscape version 3.6.1. Finally, the study was completed with a survival analysis of both groups.

Results: Significant differences were found in more than 2000 genes or transcripts. Genes or transcripts related to pathways of the cell cycle were found in the MSI-H group, while genes or transcripts related to pathways of the p53 signaling pathway were found in the group MSI-L and MSS. In regard to the survival analysis, there were no significant differences between the two groups.

Conclusions: According to the results found in patients with MSI-H endometrial cancer an alternative to the complementary treatment with chemotherapy or radiotherapy could be the CDKs inhibitors.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

El concepto de medicina de precisión[1] definido como "el uso de la información de genes, proteínas y otras características del cáncer de una persona a fin de determinar el diagnóstico o el tratamiento de la enfermedad" no es algo nuevo, siempre ha estado, pero gracias a los adelantos recientes en ciencia y tecnología han ayudado a acelerar los avances en este campo de investigación. Actualmente el objetivo que se persigue es conseguir el máximo nivel de eficacia en los tratamientos contra el cáncer.

La terapia dirigida es la base de la medicina de precisión y con ella se pretende actuar sobre los cambios que promueven el crecimiento, la división y diseminación de las células cancerosas consiguiendo crear tratamientos prometedores que actúen sobre estos cambios o bloqueen sus efectos.

La identificación de nuevas mutaciones y perfiles moleculares debería mejorar nuestra capacidad de personalizar el tratamiento con terapias dirigidas guiadas por el genoma.

Debido a la creciente necesidad por encontrar tratamientos complementarios en el cáncer de endometrio sumado a la creciente disponibilidad de multitud de datos ómicos me han parecido una excelente oportunidad de incorporarme en este campo de investigación con el objetivo de no solo adquirir un adecuado manejo de este tipo de estudios sino además de poder contribuir a la necesidad de encontrar nuevas estrategias terapéuticas.

1.2 Objetivos del Trabajo

Objetivo general

- Identificar alteraciones en la expresión diferencial de los genes de pacientes con cáncer de endometrio.

Objetivos específicos

- Estratificar aquellos genes alterados en subgrupos según su estatus mutacional.
- Análisis de supervivencia de los subgrupos identificados.
- Identificar alteraciones en los pathways según los subgrupos identificados.
- Análisis de coexpresión de datos
- Identificar potenciales dianas terapéuticas para cada subgrupo con el fin de llevar a cabo un tratamiento personalizado y dirigido.

1.3 Enfoque y método seguido

Los datos que utilizaremos para realizar el estudio proceden del análisis de RNA-seq de muestras de cáncer de endometrio del TCGA. Para llevar a cabo el análisis utilizaré el lenguaje de programación estadístico R versión 3.5.0.

Una vez descargados los datos tendremos millones de secuencias obtenidas del cDNA generado del RNA extraído de las muestras. Nuestro objetivo principal es buscar si hay genes desregulados. Para ello primero realizaremos un paso de preprocesado de los datos para evaluar la calidad de las secuencias y normalizar los datos (identificar y eliminar las diferencias técnicas entre las muestras). Posteriormente realizaremos el análisis estadístico para identificar los genes diferencialmente expresados según el estatus mutacional. Finalmente completaremos el estudio realizando un análisis de supervivencia, un análisis de significación biológica de pathways y un análisis de coexpresión génica.

1.4 Planificación del Trabajo

	Meses											
	MARZO		ABRIL				MAYO				JUNIO	
ACTIVIDAD / TAREA	19-25	26-1	2-8	9-15	16-22	23-29	30-6	7-13	14-20	21-27	28-3	4-10
Familiarización y puesta a punto del paquete de R TCGAbiolinks	✓											
Obtención de los datos de RNASeq de cáncer de endometrio	✓											
Normalización de los datos		✓	✓									
Análisis de expresión diferencial		✓	✓									
Visualización de los resultados		✓	✓									
Análisis estadístico y visualización de clústeres de genes según "driver mutation" asociada				✓								
Análisis de supervivencia: Regresion cox					✓							
Análisis significación biológica de Pathways						✓						
Análisis coexpresión de genes							✓					
Búsqueda de posibles alternativas terapéuticas								✓	✓			
Elaboración de la memoria										✓	✓	✓

1.5 Breve sumario de productos obtenidos

Los resultados obtenidos son:

- Memoria
- Pipeline: Con la implementación del código usado durante el análisis. Queda presentado como un anexo a la memoria.
- Producto: Una vez considerada la importancia y novedad de los resultados obtenidos en este estudio se valorará hacer públicos los resultados a la comunidad científica a través de una comunicación en un congreso nacional o internacional así como a través de un artículo original publicado en una revista científica.
- Presentación virtual

1.6 Breve descripción de los otros capítulos de la memoria

La memoria se encuentra dividida en cuatro bloques:

En el primer bloque, cáncer de endometrio y técnicas de secuenciación masiva, se hará una breve descripción de las principales características que presenta este tipo de cáncer y de la aplicación de las nuevas técnicas de secuenciación masiva (transcriptómica) en el mismo.

En el segundo bloque, material y métodos, se describirá detalladamente toda la metodología que se ha llevado a cabo para realizar el proyecto.

El tercer bloque, resultados, se presentaran todos los resultados obtenidos.

El cuarto bloque, discusión, se interpretaran los resultados obtenidos.

Y por último, en el quinto bloque, conclusiones, se hará una descripción de las conclusiones obtenidas y una valoración crítica de los logros conseguidos y de las perspectivas futuras.

2. Cáncer de endometrio y técnicas de secuenciación masiva

2.1 Cáncer de endometrio

2.2.1 Incidencia, factores de riesgo y clasificación

El cáncer de endometrio es el tumor ginecológico más frecuente en países desarrollados. En España, la incidencia del carcinoma de endometrio es de 11.13 por 100.000 habitantes/año, con una mortalidad de 2.34 por 100.000 habitantes/año[2].

La mayoría de los casos se diagnostican en mujeres posmenopáusicas, entre 55-65 años. Sólo el 20% de los casos se presentan en mujeres premenopáusicas y el 5% en mujeres menores de 40 años. La mediana de edad al diagnóstico se estima en 63 años[3].

En el cáncer de endometrio existe un desarrollo y proliferación anormal de las células del revestimiento interno del útero (endometrio). La causa real es desconocida, aunque se ha relacionado frecuentemente con la exposición excesiva y/o continua a estrógenos, tanto endógenos como exógenos[4]. La nuliparidad, menarquia precoz y menopausia tardía estarían englobadas en el grupo de causas relacionadas con la exposición elevada a estrógenos. La diabetes mellitus, hipertensión arterial, obesidad y el cáncer de colon hereditario no polipósico (síndrome de Lynch) son otras causas que se citan como factores de riesgo.

Clasificación clínica:

- Tipo I: asociado a la estimulación de estrógenos, obesidad, e hiperlipidemia (niveles altos de colesterol). Se suelen encontrar en estadios iniciales, son de bajo grado histológico, se asocian a lesiones pre-malignas (hiperplasia endometrial) y suelen presentar buen pronóstico. Representan el 70- 80% del total de los casos diagnosticados [4,5].

- Tipo II se caracteriza por ser principalmente de histología serosa, asociado a edad avanzada, pacientes no obesas, presentan mayor riesgo de metástasis y un peor pronóstico. Tiene un comportamiento más agresivo, no suele presentar lesiones premalignas y normalmente no son dependientes de hormonas (estrógenos)[6].

Clasificación histológica:

- Adenocarcinoma endometrioide (75%)
- Mixto; definido por la presencia de dos tipos de células carcinomatosas (10%)
- Seroso papilar uterino (<10%)
- Células claras (4%)
- Carcinosarcoma (3%)

- Mucinoso (10%)
- Celulas escamosas (<1%)
- Indiferenciado (<1%)

2.2.2 Métodos de diagnóstico y tratamiento

El síntoma más importante de presentación es el sangrado vaginal anormal, fundamentalmente peri o posmenopáusico, ya que tres cuartas partes de estos cánceres se presentan en mujeres en este intervalo de edad, y que aproximadamente el 90% debutan con este síntoma. El sangrado posmenopáusico tiene más posibilidades de ser debido a un adenocarcinoma de endometrio cuando su presentación se produce en edad de menopausia avanzada.

Entre los métodos de diagnósticos destacar[7]:

- Examen pélvico
- Ecografía transvaginal
- Biopsia de endometrio
- Dilatación y legrado
- Histerectomía

La piedra angular del tratamiento de esta enfermedad es la cirugía, siendo la histerectomía total más salpingooforectomía bilateral y el muestreo linfático ilíaco bilateral (en casos seleccionados) el tratamiento estandar[4].

La necesidad de tratamientos complementarios a la cirugía, como la quimioterapia o radioterapia adyuvante, se basa fundamentalmente teniendo en cuenta factores clínico-patológicos. Actualmente existe mayor evidencia acerca de la importancia en la identificación de biomarcadores moleculares pronósticos y predictivos de respuesta que nos permita personalizar el tratamiento complementario.

En cuanto al grupo de pacientes con enfermedad metastásica o recurrente las opciones terapéuticas son limitadas. El tratamiento se valora según la localización de la enfermedad y los síntomas. No existe ningún tratamiento estándar determinado en este escenario de la enfermedad[8].

Dada la falta de opciones terapéuticas en esta enfermedad existe un gran interés en la identificación de marcadores moleculares predictivos de respuesta a tratamientos. La identificación de mutaciones y perfiles moleculares mediante el estudio del genoma permitirá personalizar el tratamiento en beneficio de los pacientes, mejorando los resultados y reduciendo la toxicidad[9].

2.2.3 The Cancer Genome Atlas

Bajo la dirección conjunta del Instituto Nacional del Cáncer (NCI) y del Instituto Nacional de Investigación del Genoma Humano (NHGRI) se estableció el Atlas del Genoma del Cáncer (TCGA) (https://cancergenome.nih.gov) con el objetivo de crear

mapas multidimensionales completos que contengan los cambios genómicos claves en los tipos y subtipos principales de cáncer.

En 2013 la Red de Investigación del Atlas del Genoma del Cáncer publicó un estudio que establecía las bases para la clasificación genómica del cáncer de endometrio[10].

Identificaron cuatro subtipos genómicos de cáncer de endometrio, lo que permitiría nuevos enfoques de diagnóstico y tratamiento. Cada uno de los cuatro subtipos genómicos se agrupó y fue denominado por su principal característica:

- **POLE ultramutado**: presentaba mutaciones inusualmente altas en el gen POLE.

- I**nestabilidad de microsatélites hipermutados** (MSI): presentaba una alta inestabilidad de microsatélites.

- **Bajo número de copias** (CNL): caracterizado por presentar la mayor estabilidad de microsatélites, bajo número de copias y una alta frecuencia de mutaciones en CTNNB1, un gen crítico para mantener los revestimientos del endometrio.

- **Alto número de copias** (CNH): presentaba un alto número de copias y mutaciones en TP53, muy característico de los tumores serosos.

2.2 Técnicas de secuenciación masiva

2.2.1 Transcriptómica

Dentro de las diversas áreas de investigación que contempla la Bioinformática se conoce como transcriptómica el estudio de todas las moléculas de RNA en una célula, tejido u órgano.

La expresión génica es el proceso mediante el cual el DNA se transcribe en RNA para posteriormente llevar a cabo la síntesis de proteínas. Por tanto una secuencia de RNA es un reflejo de la secuencia de DNA de la que fue transcrito. Si analizamos la colección completa de secuencias de RNA en una célula, los investigadores pueden determinar cuándo y dónde está activado cada gen en las células y tejidos de un organismo[11].

Al comparar los transcriptomas de distintos tipos de células, se puede adquirir un conocimiento más profundo sobre las funciones de una célula y cómo los cambios en su actividad génica pueden contribuir a la aparición de enfermedades.

Entre los métodos para medir los niveles de expresión génica, actualmente encontramos:

- Microarrays
- Análisis de secuenciación masiva RNA-Seq

El análisis de RNA-Seq a diferencia del análisis de datos mediante microarrays presenta las siguientes ventajas[12]:

- No requiere sondas específicas de especie o transcripción
- Puede detectar nuevas transcripciones
- Permite una detección más fácil de transcripciones raras y de baja abundancia
- Bajo ruido de fondo
- Capacidad para cuantificar un amplio rango dinámico de niveles de expresión con valor absoluto en lugar de valores relativos
- Mayor sensibilidad y especificidad para una mejor detección de genes, transcripciones y expresión diferencial

La única desventaja es su mayor costo.

La elección entre una técnica u otra se debe realizar en función de nuestros objetivos. Por tanto si nuestro proyecto requiere de alguna de las consideraciones previas, como es nuestro caso, nos decantaremos por un análisis de RNA-Seq.

Fundamento del análisis de RNA-seq

Como ya hemos comentado el RNA-Seq es una herramienta transcriptómica que está basada en la secuenciación de DNAc.

Primero se extrae el RNAm maduro (contiene solo la secuencia codificante), se fragmenta y mediante una transcripción reversa se convierte en DNAc. Este DNAc se secuencia. Posteriormente se procede a alinear las secuencias obtenidas con las secuencias de genomas de referencia, reconstruyendo todas las regiones del genoma que se transcriben.

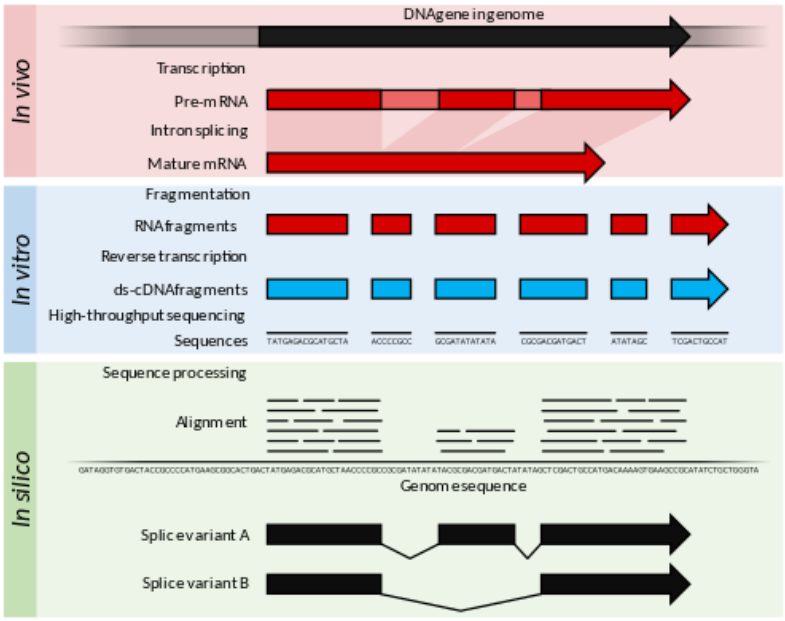

Figura 1. Resumen análisis RNA-Seq

2.2.2 Importancia de la transcriptómica en el cáncer

Gracias a las nuevas técnicas de secuenciación masiva y la disponibilidad de datos de secuencias de RNA de múltiples tumores gracias a proyectos como el Atlas del Genoma del Cáncer se ha puesto en evidencia la importancia del transcriptoma en el desarrollo de tumores.

Aunque el cáncer se origina por mutaciones en el DNA, estas presentan un impacto en el transcriptoma que pueden inducir mecanismos vinculados con el desarrollo del cáncer[13]. Por tanto además de saber si un gen ha mutado o es normal, también es importante saber si ese gen se expresa y cómo lo hace.

Con estas nuevas técnicas se abren nuevos vías para entender la biología del cáncer y muy importante buscar nuevas estrategias terapéuticas.

Los estudios de expresión génica se han ido incrementando cada vez más en el ámbito de la oncología, con el objetivo de identificar marcadores específicos de cada tipo de tumor y como comentábamos buscar nuevas dianas terapéuticas.

3. Material y métodos

3.1 Software de base para el análisis

A la hora de seleccionar el software con el que realizaría los análisis, me he decantado por **el lenguaje de programación estadístico R versión 3.5.0**.

R es un entorno de software libre para computación y gráficos estadísticos (http://R-project.org).

Los motivos por los que me he decantado por el uso de este lenguaje de programación son los siguientes:

- Amplia gama de herramientas estadísticas
- Capacidad de generar gráficos de alta calidad
- Eficacia en el manejo de grandes volúmenes de datos
- Compatibilidad con diferentes software (Windows)
- Posibilidad de cargar paquetes como Bioconductor que nos permite realizar el análisis y la comprensión de datos genómicos de alto rendimiento
- Actualmente cuenta con el paquete *TCGAbiolinks* para poder descargar directamente los datos del TCGA.

3.2 Obtención de muestras

Para llevar a cabo el proyecto propuesto se han utilizado los datos del TCGA. Los datos correspondientes con el cáncer de endometrio se engloban dentro del proyecto: **TCGA-UCEC**.

En concreto seleccionamos:

- data.category = “Transcriptome Profiling”
- data.type = “Gene Expression Quantification”
- experimental.strategy = “RNA-Seq”
- workflow.type = “HTSeq –Counts”
- sample.type = “Primary solid Tumor”

En cuanto al estatus mutacional seleccionado para posteriormente clasificar nuestras muestras en subgrupos, en el TCGA únicamente encontramos información de acceso abierto (Nivel 3) de la inestabilidad de satélites (MSi-MSS).

Microsatellite Instability (MSI)

Data Subtype	Cancer Types Applicable	Data Type Name	Level 1	Level 2	Level 3	Important Metadata	How to Retrieve Data Files
	COAD, READ, UCEC	Fragment Analysis Results	Markers indicating presence or absence of a MSI shift, allele homozygosity/heterozygosity, and loss of heterozygosity (LOH) observed in the tumor sample for each participant File types: fragment analysis trace file (.fsa) and tab-delimited (.txt) file summarizing the trace file (Controlled-access)	n/a	Classifications of microsatellite instability detected for each participant's tumor sample File type: auxiliary.xml	Level 1 data are submitted as part of a standard **MAGE-TAB** archive Level 3 data are contained in the BCR clinical data archives (see above)	Level 1: Data Matrix & Bulk Download: Select 'Fragment Analysis Results' for Data Type File Search: Select 'Other' for Data Category

Figura 2. Información disponible en el TCGA del estatus mutacional MSi-MSS

Dada la baja frecuencia de tumores MSI-L y su similitud con los tumores MSS, he decidido clasificar las muestras en dos grupos:

- G1: MSI-H
- G2: MSI-L+MSS

Mediante la librería *TCGAbiolinks* descargamos la matriz de conteo sin normalizar que contiene el número de lecturas para cada gen o transcrito. Encontrando en las filas los genes o transcritos y en las columnas nuestras muestras.

Partimos inicialmente:

Muestras	Genes o transcritos
375	56830

Una vez que tenemos nuestra matriz de conteo, creamos una lista con la función DGEList del paquete EdgeR para almacenar nuestra matriz de conteo y un data.frame con información del grupo al que pertenece cada muestra.

3.3 Pre procesado de los datos

3.3.1 Filtrado

Los genes o transcritos con recuentos muy bajos en todas las librerías proporcionan poca evidencia en la expresión diferencial e interfieren con algunas aproximaciones estadísticas que se utilizan más adelante. Además de reducir la potencia para detectar genes o transcritos diferencialmente expresados.

Para filtrar estos genes o transcritos, el método más utilizado el CPM, elimina aquellos genes o transcritos que presentan un número de lecturas inferior a un determinado número de lecturas por millón, dependiendo del total de lecturas en cada muestra, dicho límite corresponde a un número diferente. Como regla general se considera que los genes se expresan si su valor de CPM está por encima de 1 (equivale a un valor de log-CPM de 0).

En nuestro caso utilizamos como criterio para hacer el filtrado un valor mínimo de CPM de 1 en al menos 15 muestras.

Tras realizar el filtrado nos quedamos con:

Muestras	Genes o transcritos
375	21452

3.3.2 Control de calidad

Antes de continuar con nuestro análisis es importante realizar un control de la distribución de nuestras muestras.

Primero observamos la distribución del tamaño de nuestras librerías.

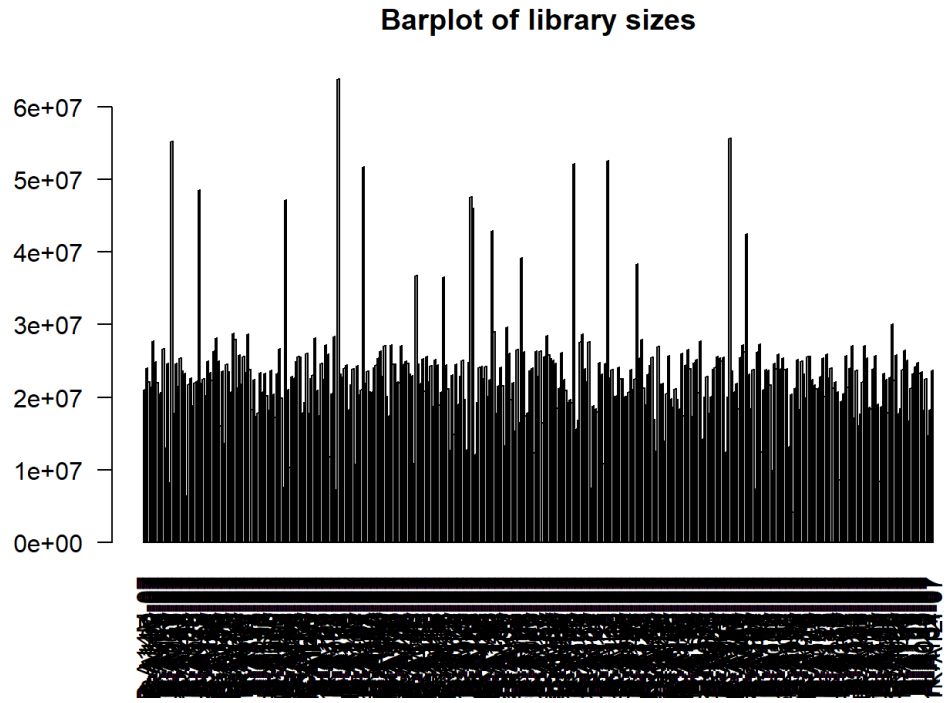

Figura 3. Distribución del tamaño de las librerías de nuestras muestras

A continuación vamos a examinar la distribución de los recuentos sin procesar mediante un diagrama de cajas en escala log2.

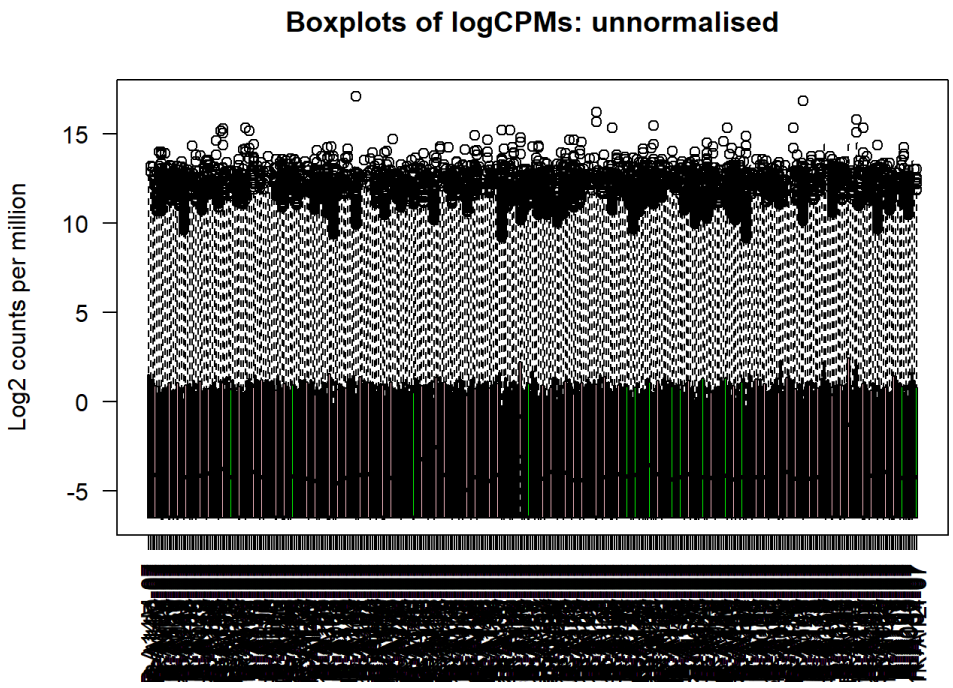

Figura 4. Distribución de las muestras sin normalizar en escala log2

En ambos gráficos se ve reflejada la necesidad de normalizar nuestros datos.

3.3.3 Normalización

Con este proceso no buscamos transformar nuestros datos para que sigan una distribución normal. Buscamos principalmente dos objetivos:

- Poner en una misma escala todas las muestras o individuos correspondientes a cada grupo para evitar falsos positivos, ya que una muestra o librería con mayor profundidad de secuenciación tiene más probabilidad de tener más genes diferencialmente expresados respecto a otra, sin deberse estas diferencias a la condición bajo estudio.

- Eliminar variaciones biológicas entre las muestras

Utilizaremos la función calcNormFactors del paquete EdgeR que proporciona un conjunto de factores de normalización que minimizan el log-fold-change (cambio en la proporción de lecturas) entre las muestras, según el total de lecturas de cada una, para la mayoría de genes. El cálculo de dichos factores de normalización se realiza por TMM entre cada par de muestras.

A continuación mostramos el resultado de la normalización de las 10 primeras muestras:

```
##                              group lib.size norm.factors
## TCGA-D1-A1NW-01A-11R-A14M-07    G2 21010497    1.1387169
## TCGA-D1-A1NZ-01A-21R-A14D-07    G1 24008791    1.1525806
## TCGA-BS-A0UL-01A-11R-A109-07    G1 22052752    1.1153037
## TCGA-B5-A0KB-01B-11R-A14D-07    G2 21307732    0.9969682
## TCGA-D1-A15V-01A-11R-A118-07    G2 27703818    0.9261388
## TCGA-BG-A18A-01A-21R-A12I-07    G2 24786266    0.8459126
## TCGA-AX-A06J-01A-11R-A00V-07    G2 21955971    0.7419865
## TCGA-BG-A0M7-01A-11R-A040-07    G2 20376253    0.9025180
## TCGA-B5-A11U-01A-11R-A118-07    G1 20728006    1.0326801
## TCGA-D1-A1NY-01A-11R-A16F-07    G1 26631577    1.0188571
```

Figura 5. Tabla de los resultados obtenidos en las 10 primeras muestras tras el proceso de normalización

Cuando obtenemos un factor de normalización inferior a 1 nos indica que el tamaño de la librería se reducirá debido a que hay más supresión en esa librería en comparación con el resto. Esto equivale a escalar los recuentos hacia arriba en esa muestra. Mientras que un factor de normalización superior a uno indica que el tamaño de la librería se incrementará, lo que equivale a reducir la escala de los recuentos.

Si volvemos a representar la gráfica de distribución usando la lista de datos normalizados, deberíamos ver que el problema de sesgo de composición se ha resuelto.

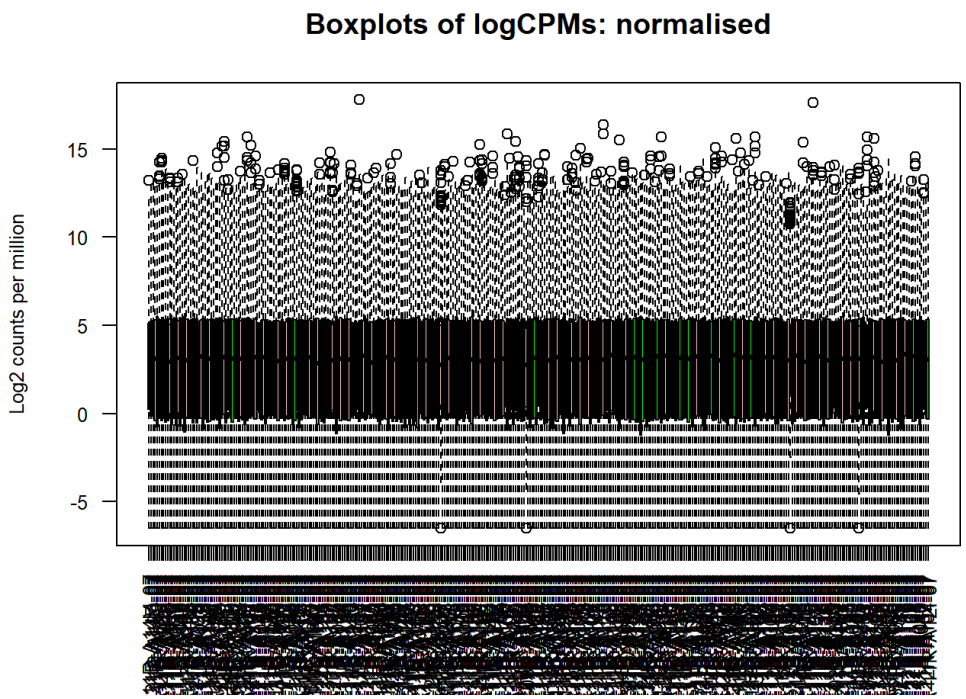

Figura 6. Distribución de las muestras tras normalizar en escala log2

3.5 Análisis de expresión diferencial

Para realizar el análisis de expresión diferencial utilizaremos el paquete Limma de Bioconductor. Este paquete se utiliza en el análisis de expresión diferencial de genes para datos de Microarray. Pero actualmente se ha adaptado para el análisis de datos de RNA-seq. Su metodología está basada en el uso de modelos lineales.

Lo primero para poder aplicar esta función es crear la matriz de diseño con las condiciones a comparar.

```
        Contrasts
Levels G1vsG2
    G1      1
    G2     -1
```

Figura 7. Matriz de diseño

Para trabajar con las funciones del paquete Limma, los datos de conteo deben de ser transformados mediante la función voom, que convierte estos datos en logaritmos del número de lecturas por millón (log-cpm). Cada dato dispondrá de un peso de precisión, que constituye la inversa de la dispersión estimada de dicha observación.

A continuación mostramos un gráfico donde observamos una visión general de la dispersión de nuestros datos.

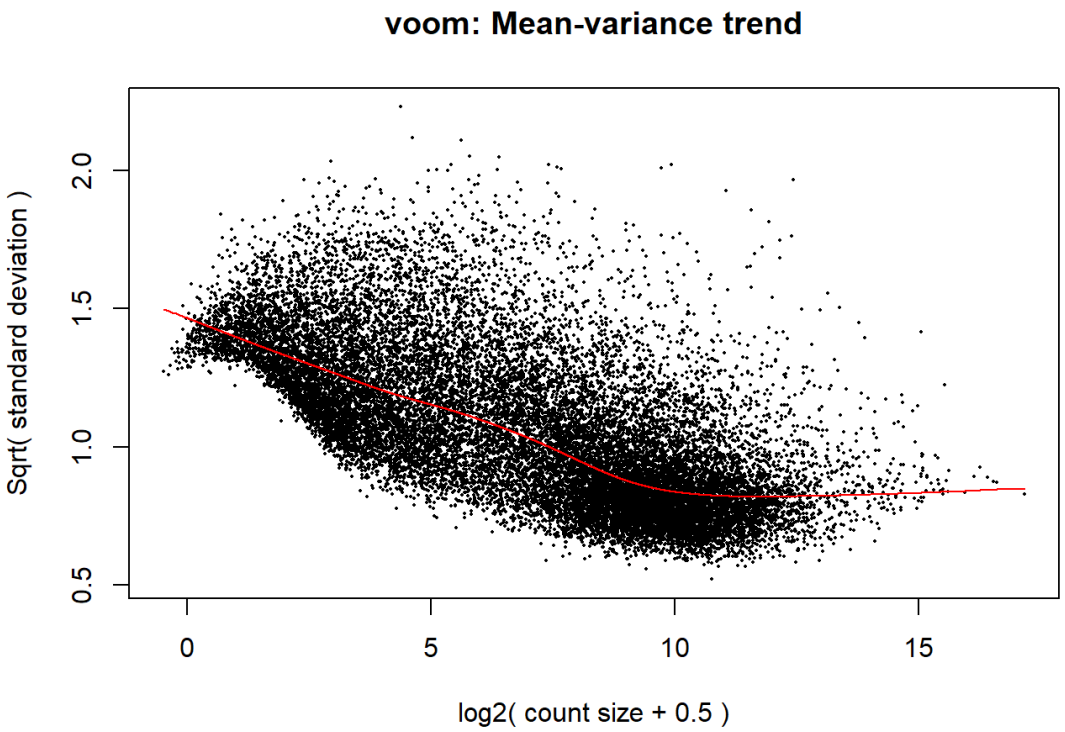

Figura 8. Función voom: Tendencia de la media-varianza

Una línea recta nos indicaría que la media y la varianza tienden a ser iguales, sin embargo, en nuestro caso obtenemos una curva, lo que nos indica que nuestros datos presentan una sobredispersión (varianza superior a la media).

Con la función voom obtenemos un objeto Elist con el que realizamos el estudio de expresión diferencial de los genes. Para llevar a cabo dicho estudio utilizamos la función lmFit donde ajustamos los datos de expresión de cada gen o transcrito del objeto que hemos llamado v a un modelo lineal teniendo en cuenta la matriz de diseño. A continuación con la función eBayes realizamos el estudio de expresión diferencial basado en métodos bayesianos.

Volvemos a representar el gráfico de tendencia de la media-varianza:

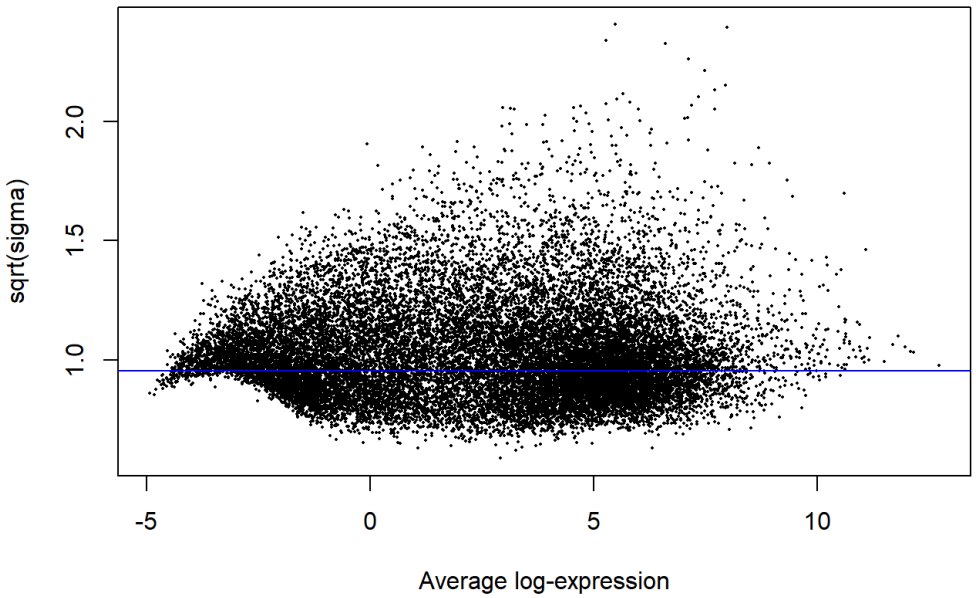

Figura 9. Tendencia de la media-varianza tras el realizar el estudio de expresión diferencial

Ahora observamos que la varianza ya no depende del nivel medio de expresión.

Por último con la función topTable podemos examinar de forma individual los genes diferencialmente expresados. Este tipo de análisis involucra gran cantidad de test, es necesario realizar un ajuste para evitar que el número de falsos positivos se incremente demasiado. Utilizaremos el criterio FDR que no es demasiado estricto.

A continuación mostramos los 10 genes o transcritos más diferencialmente expresados:

	entrezid	symbol	logFC	AveExpr	t	P.Value
ENSG00000076242	4292	MLH1	-2.49	3.94	-23.05	0.00
ENSG00000178567	9852	EPM2AIP1	-2.71	4.26	-20.14	0.00
ENSG00000184719	55328	RNLS	-2.21	1.61	-12.41	0.00
ENSG00000198134			1.07	0.88	10.48	0.00
ENSG00000178458			1.63	-1.90	10.41	0.00
ENSG00000231503			1.29	1.69	9.96	0.00
ENSG00000145908	91975	ZNF300	-2.35	1.75	-9.56	0.00
ENSG00000115970	63892	THADA	-0.71	5.61	-9.35	0.00
ENSG00000243943	84450	ZNF512	-1.08	4.51	-9.34	0.00
ENSG00000250486	152756	FAM218A	-2.02	-1.45	-9.36	0.00

Figura 10. Tabla con los 10 primeros genes o transcritos diferencialmente expresados

Otra manera de examinar el número de genes diferencialmente expresados es con la función decideTest. Con esta función obtenemos una tabla donde para cada comparación obtenemos un 1 si el gen está sobre-expresado (up), un 0 si no hay cambios significativos o un -1 si está sub-expresado (down).

Como criterios seleccionamos aquellos genes o transcritos con un p-valor inferior a 0.001 y ajustamos con el criterio FDR.

	G1vsG2
Down	1194
NotSig	19311
Up	947

Figura 11. Genes Diferencialmente expresados (función *decideTest*)

4. Resultados

Con la función topTable obtenemos 2141 genes o transcritos donde se observan diferencias significativas entre G1 y G2.

Con la función decideTest obtenemos 2141 genes o transcritos donde se observan diferencias significativas entre G1 y G2. A continuación utilizamos la función plotMD para ver gráficamente estas diferencias, se genera un gráfico que muestra log-FC del ajuste del modelo lineal frente a los valores log-cpm promedio:

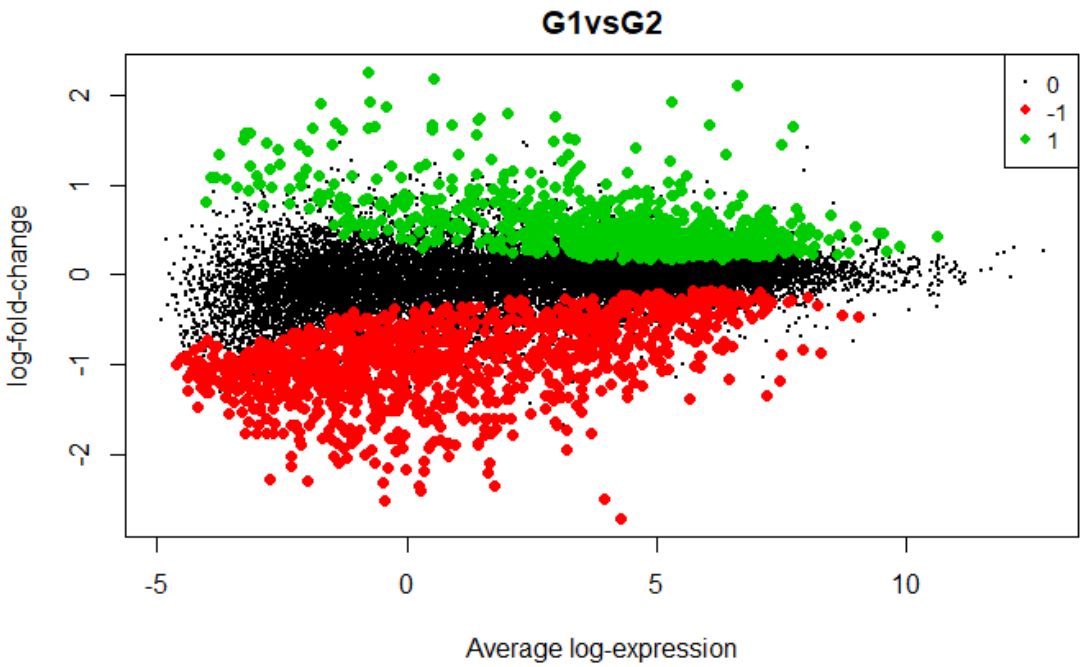

Figura 12. plotMD G1 vs G2

En verde observamos los genes o transcritos sobre-expresados, en rojo los genes o transcritos sub-expresados y en negro los genes o transcritos que no hay diferencias significativas.

A continuación, podemos observar la expresión de un subconjunto de genes en cada muestra mediante un heatmap (mapa de calor). Este tipo de gráficos, cuando se examinan patrones promediados sobre miles de genes al mismo tiempo, pierden resolución. En nuestro caso representamos los primeros 30 genes o transcritos:

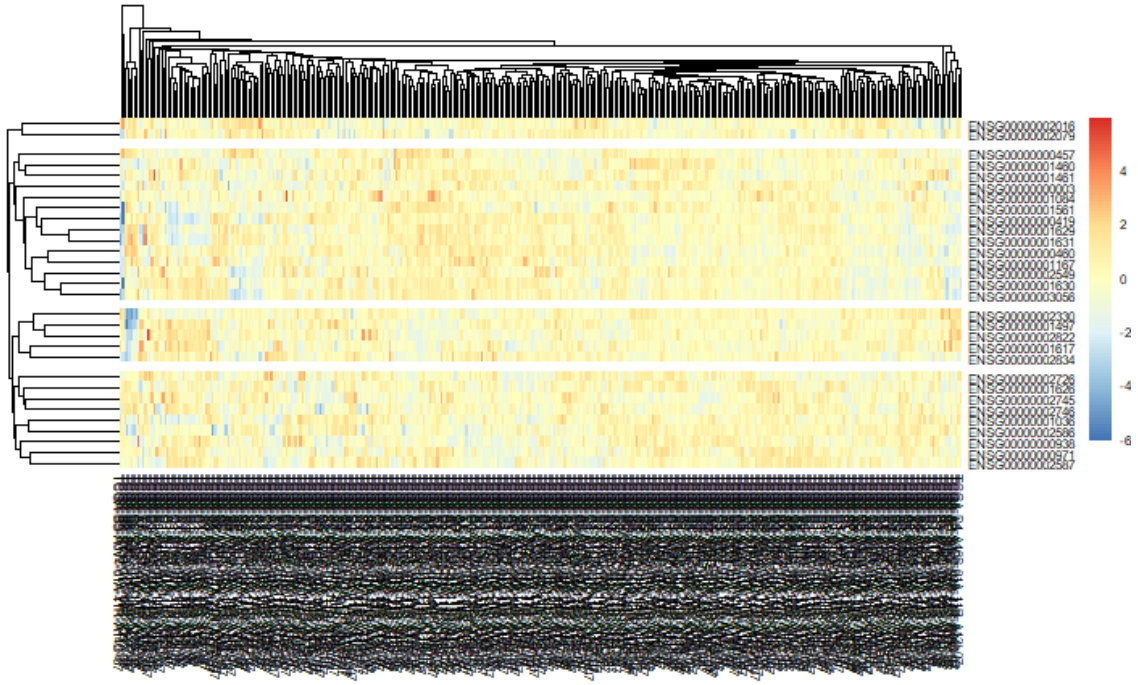

Figura 13. Heatmap 30 primeros genes o transcritos.

No se observan patrones claramente distinguibles.

Una vez que hemos obtenido los genes o transcritos diferencialmente expresados en ambos grupos vamos a ver a que pathways se asocian y realizar un análisis de coexpresión para además observar en que vías están integrados estos genes o transcritos. Para ello utilizamos la aplicación Reactome FIPlugin mediante Cytoscape versión 3.6.1.

Cytoscape[14] es una plataforma de software de código abierto para visualizar redes de interacción molecular y vías biológicas e integrar estas redes con anotaciones, perfiles de expresión génica y otros datos.

Reactome FIPlugin[15] es una aplicación diseñada para encontrar vías y patrones de red relacionados con el cáncer y otras enfermedades. Esta aplicación accede a rutas almacenadas en la base de datos Reactome, que lo ayudan a realizar un análisis de enriquecimiento para un conjunto de datos e investigar las relaciones funcionales entre los genes en las vías de acceso. La aplicación también puede acceder a la red Reactome Functional Interaction, una red de interacción funcional proteica basada en vías altamente confiables que cubre más del 60 % de las proteínas humanas y le

permite construir una subred de FI basada en conjunto de genes, analizar módulos de redes de grupos de genes altamente interactivos, realizar análisis de enriquecimiento funcional, expandir la red mediante la búsqueda de genes relacionados y mostrar diagramas de ruta entre otras muchas más utilidades. Recientemente también han agregado características para ayudar a visualizar fármacos contra el cáncer aprobados por la FDA en los contextos de la red FI y las vías Reactome.

Por tanto mediante la aplicación ReactomeFI realizamos un análisis de datos basado en redes FI para nuestra lista de genes o transcritos diferencialmente expresados (Gene Set/Mutation Analysis). Al cargar nuestra lista utilizamos la Reactome FI Network versión 2016. Obtenemos nuestra red de interacción funcional:

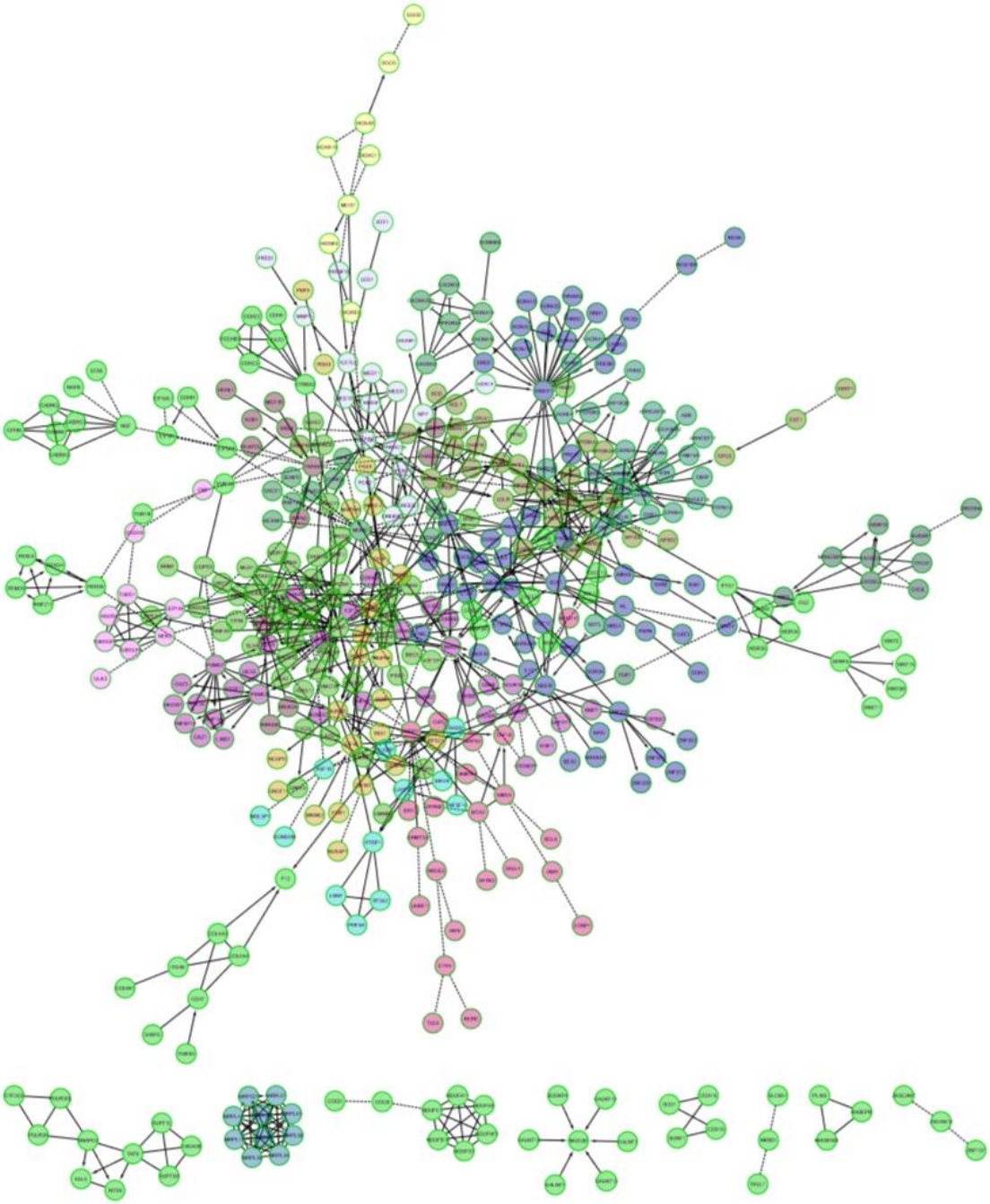

Figura 14. Red de interacción funcional de genes o transcritos diferencialmente expresados.

A la hora de interpretar estos resultados, tendríamos que ir viendo cada gen o transcrito donde está más diferencialmente expresado para asociarlo a posibles dianas terapéuticas en cada grupo. Para facilitar este trabajo lo que vamos hacer es realizar este mismo análisis por separado, es decir primero en la lista de genes o transcritos sobre-expresados y posteriormente en la lista de genes o transcritos sub-expresados.

GENES O TRANSCRITOS SOBRE-EXPRESADOS

La red de interacción funcional que obtenemos utilizando los mismos criterios que hemos comentado es la siguiente:

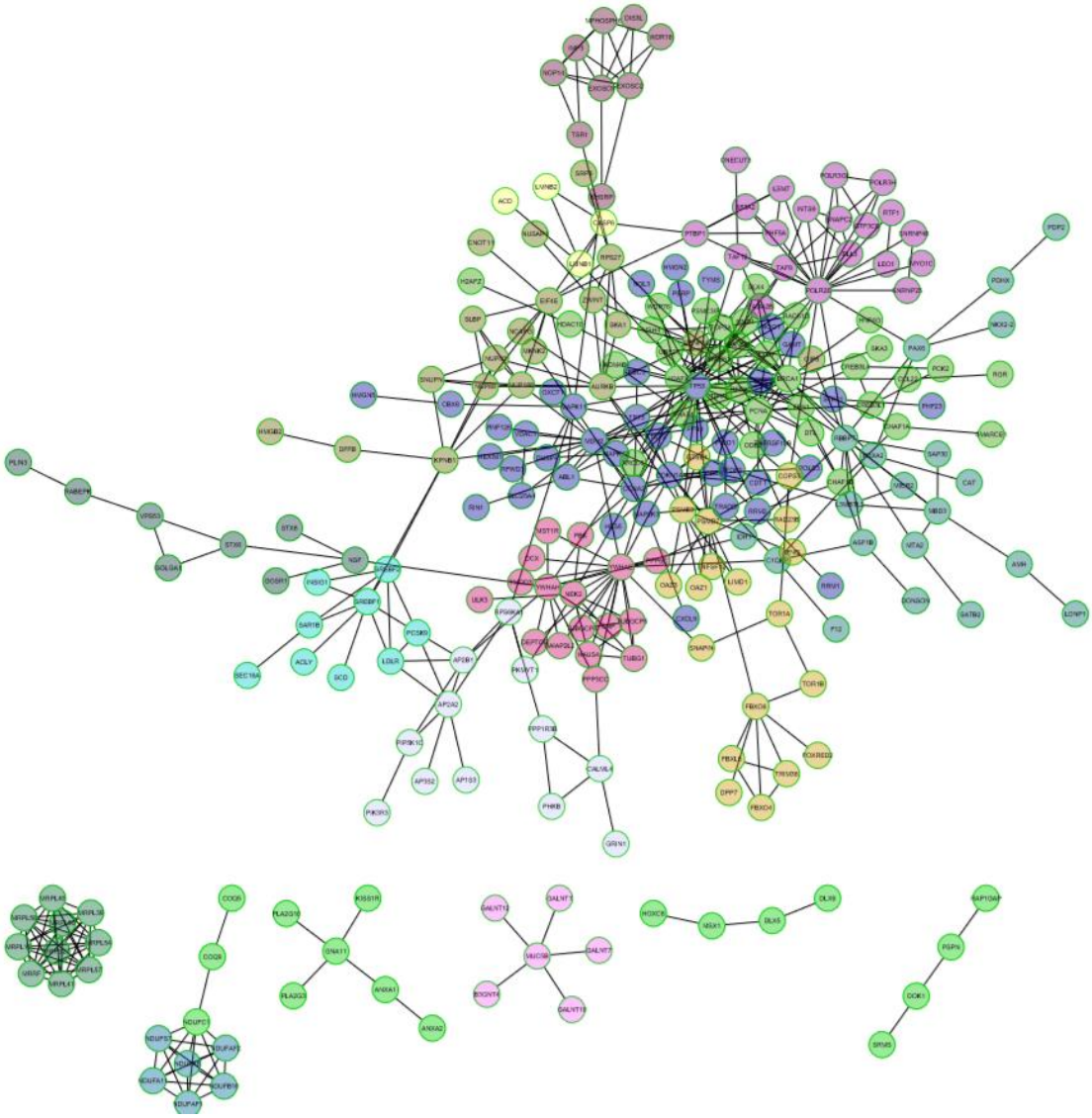

Figura 15. Red de interacción funcional de genes o transcritos sobre-expresados.

Los pathways asociados a nuestros genes o transcritos sobre-expresados los mostramos a continuación. Utilizamos como criterio FDR = 0.01 y únicamente seleccionamos los 5 más significativos.

GeneSet	Ratio of protein in Gene Set	Nº protein in Gene Set	Protein from Network	P-value	FDR	Nodes
Generic Transcription Pathway (R)	0.045	494	36	2,57E-08	1,61E-05	SCO1,TP53,SURF1,TP73,CDKN1A,HNF4G,BRCA1,BBC3,TOP3A,DDB2,PCNA,PERP,MBD3,STK11,BNIP3L,COX16,COX18,MED31,FANCI,TNFRSF10B,CNOT11,MDM2,YWHAE,PIDD1,CASP6,EXO1,YWHAH,RFC2,TFDP1,TAF9,AURKB,POLR2E,RBBP7,MTA2,TAF12,MAPK11
The citric acid (TCA) cycle and respiratory electron transport (R)	0.014	161	19	6,09E-08	1,90E-05	PDHX,SCO1,ATP5D,SURF1,NDUFAF2,NDUFAF1,UQCR11,PDP2,NDUFA11,COX16,NDUFB10,COX18,UQCRQ,ETFA,ATP5G1,NDUFC1,NDUFS7,ETFDH,NDUFA6
p53 signaling pathway (K)	0.006	69	12	3,41E-07	7,11E-05	GTSE1,TP53,TP73,CDKN1A,BBC3,DDB2,PERP,TNFRSF10B,MDM2,PIDD1,EI24,RRM2
Direct p53 effectors (N)	0.012	132	16	4,84E-07	7,55E-05	TP53,TADA2B,TP73,CDKN1A,BBC3,DDB2,PCNA,PERP,BNIP3L,TNFRSF10B,MDM2,PIDD1,CASP6,TFDP1,TAF9,E2F2
Cell Cycle Checkpoints (R)	0.015	164	16	7,51E-06	7,82E-04	PSMD7,GTSE1,PSME3,TP53,CDC6,CDKN1A,BRCA1,TOP3A,PIAS4,PKMYT1,MDM2,YWHAE,EXO1,YWHAH,RFC2,H2AFX

Tabla1. Pathways asociados a genes o transcritos sobre-expresados.

Ahora representamos la red de interacción funcional únicamente con los genes asociados a estas pathways.

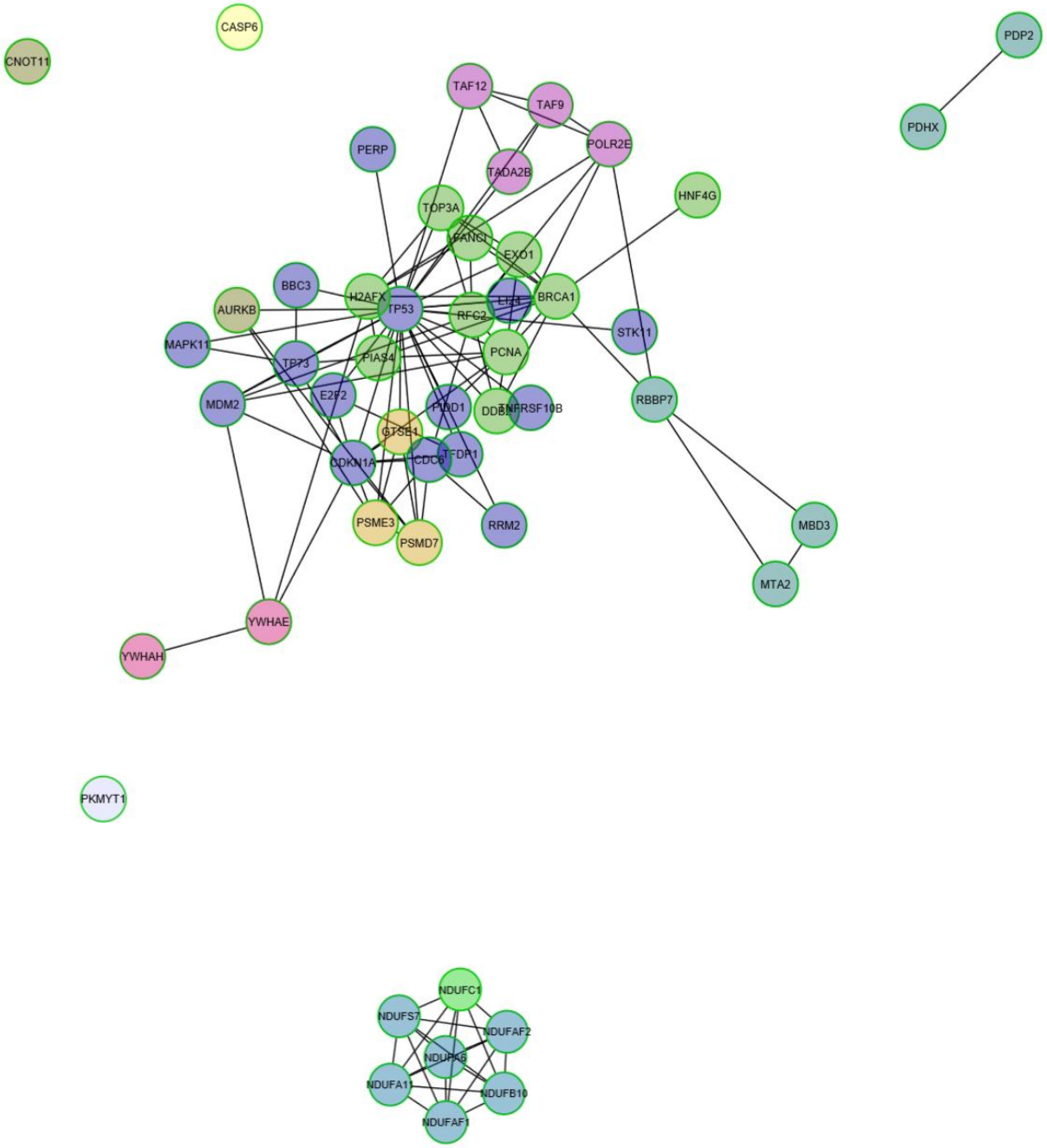

Figura 16. Red de interacción funcional de genes o transcritos sobre-expresados asociados a las pathways.

GENES O TRANSCRITOS SUB-EXPRESADOS

La red de interacción funcional que obtenemos ahora es la siguiente:

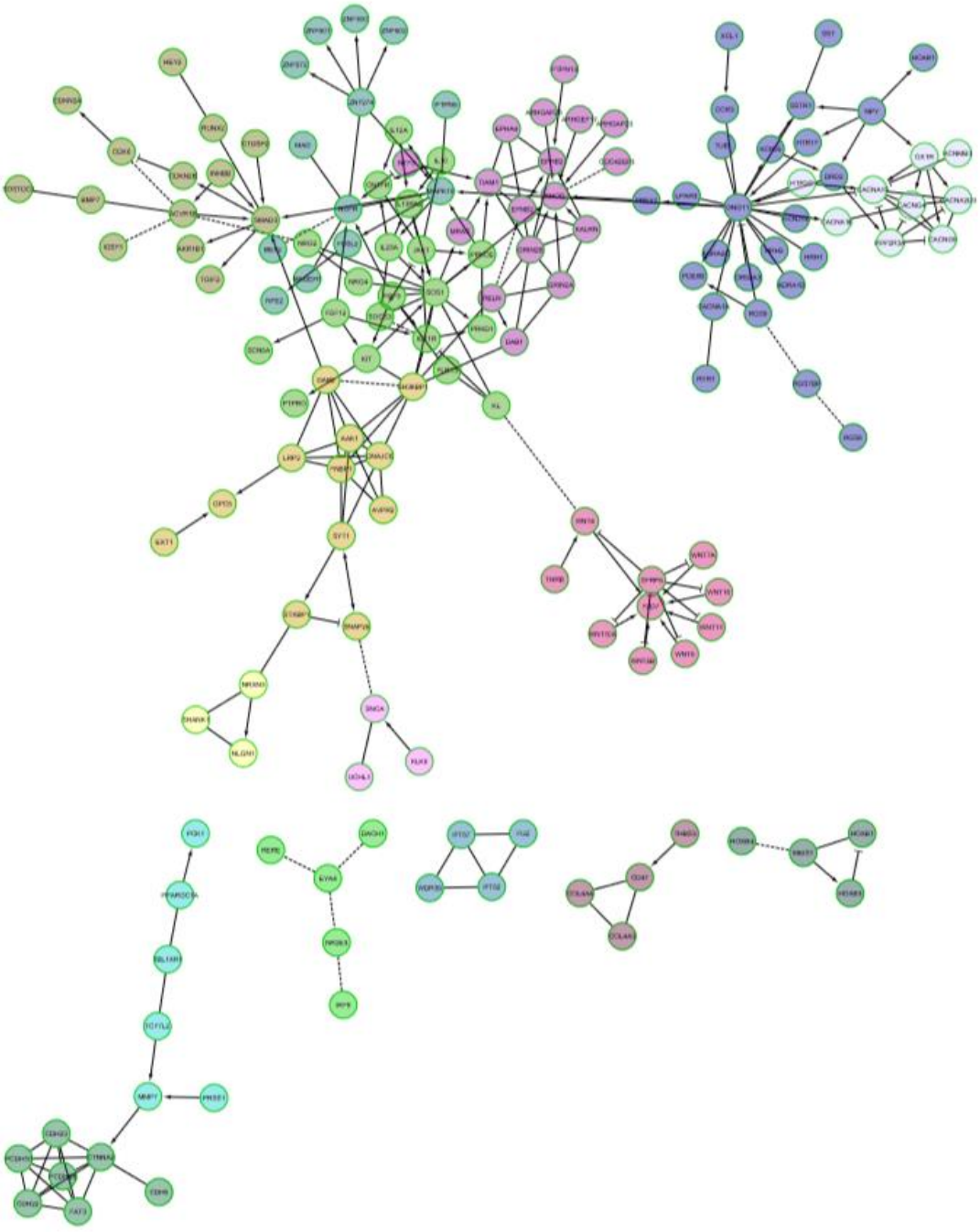

Figura 17. Red de interacción funcional de genes o transcritos infra-expresados.

Los pathways asociados a nuestros genes o transcritos sub-expresados los mostramos a continuación. Utilizamos al igual que antes como criterio FDR = 0.01 y únicamente seleccionamos los 5 más significativos.

GeneSet	Ratio of protein in Gene Set	Nº protein in Gene Set	Protein from Network	P-value	FDR	Nodes
Pathways in cancer (K)	0.036	397	27	3,18E-10	1,68E-07	FGF3,IGF1R,WNT5B,COL4A4,COL4A3,LPAR3,WNT6,WNT4,WNT10A,SMAD3,FZD7,CDK6,FGF12,LAMC3,GNGT1,JAK1,WNT16,KIT,SOS1,LAMA2,WNT11,CTNNA2,TCF7L2,CDKN2B,CDKN2A,WNT7A,MAPK10
Cadherin signaling pathway (P)	0.009	100	14	1,32E-09	3,52E-07	WNT5B,PCDH10,WNT6,CDH22,CDH23,WNT4,WNT10A,FZD7,PCDHB5,FAT3,WNT16,WNT11,CTNNA2,WNT7A
Breast cancer (K)	0.013	146	16	2,73E-09	4,16E-07	FGF3,IGF1R,WNT5B,WNT6,WNT4,WNT10A,FZD7,CDK6,FGF12,HEY2,WNT16,KIT,SOS1,WNT11,TCF7L2,WNT7A
Wnt signaling pathway (P)	0.024	268	21	3,15E-09	4,16E-07	CDH6,WNT5B,PRKCE,TBL1XR1,SFRP5,PCDH10,WNT6,CDH22,CDH23,WNT4,WNT10A,FZD7,PCDHB5,FAT3,MMP7,WNT16,ACVR1B,WNT11,CTNNA2,TCF7L2,WNT7A
Signaling pathways regulating pluripotency of stem cells (K)	0.013	142	15	1,41E-08	1,49E-06	IGF1R,WNT5B,HOXB1,WNT6,WNT4,WNT10A,SMAD3,FZD7,INHBB,JAK1,WNT16,ACVR1B,WNT11,WNT7A,MEIS1

Tabla2. Pathways asociados a genes o transcritos infra-expresados.

Representamos la red de interacción funcional únicamente con los genes asociados a estas pathways.

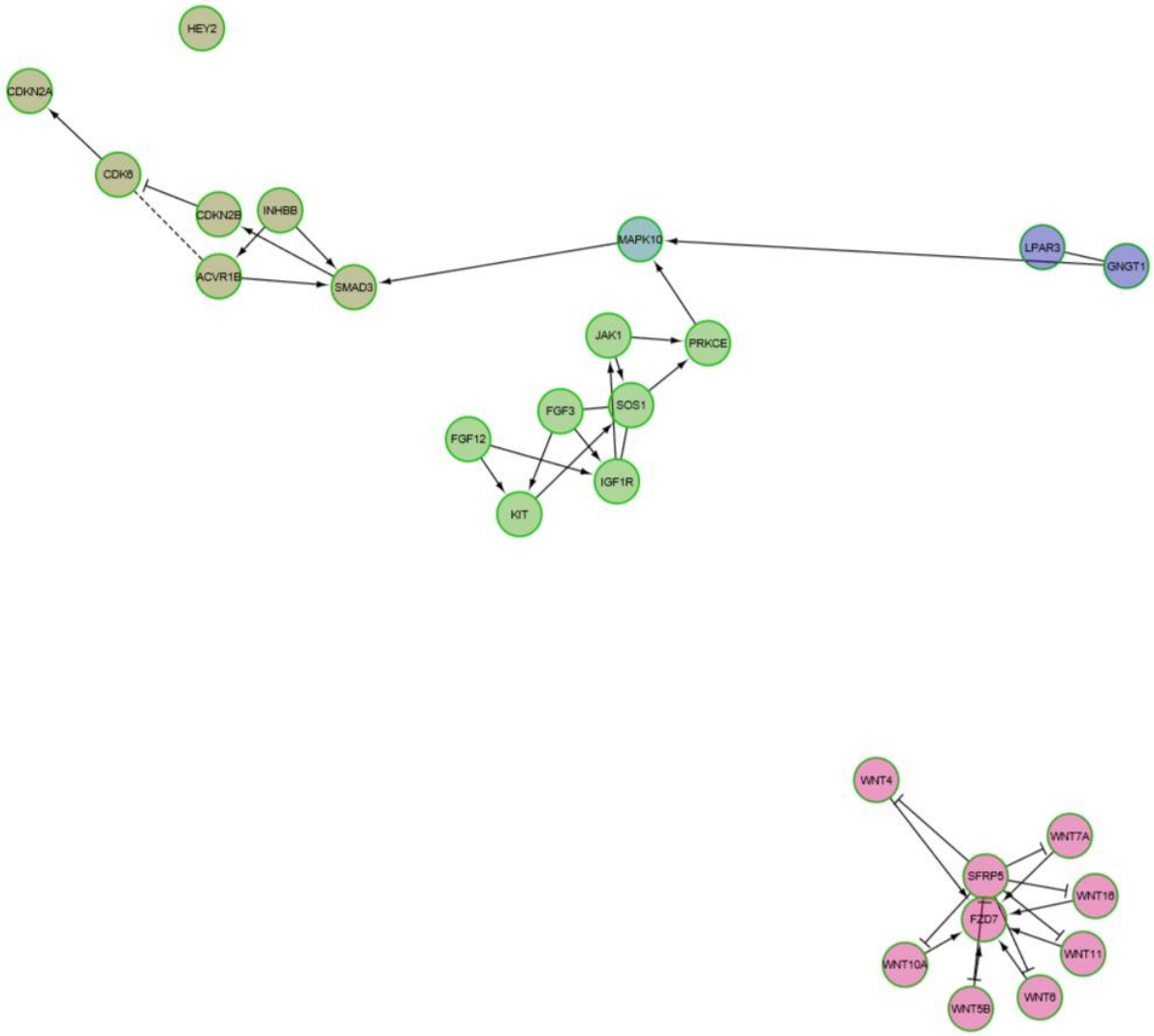

Figura 18. Red de interacción funcional de genes o transcritos infra-expresados asociados a las pathways.

Por último, realizamos un análisis de supervivencia comparando nuestros dos grupos. Utilizamos el método de Kaplan-Meier para estimar las probabilidades de supervivencia en los instantes en los que ha ocurrido el evento. Este análisis lo realizamos con la función survfit de la librería survival. Y para comparar las curvas de supervivencia de ambos grupos utilizamos el modelo de riesgos proporcionales de Cox. Para llevarlo a cabo utilizamos la función coxph de la librería survminer.

Los resultados obtenidos son los siguientes:

Diagrama de Kaplan Meier

My_data.subtype_msi_status=G1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	127	0	1.000	0.0000	1.000	1.000
365	108	5	0.956	0.0190	0.920	0.995
730	76	5	0.905	0.0290	0.850	0.963
1095	53	5	0.838	0.0394	0.765	0.919
1460	46	0	0.838	0.0394	0.765	0.919

My_data.subtype_msi_status=G2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	248	0	1.000	0.0000	1.000	1.000
365	220	11	0.953	0.0138	0.926	0.981
730	171	9	0.910	0.0193	0.873	0.948
1095	115	10	0.847	0.0263	0.797	0.900
1460	85	4	0.813	0.0304	0.755	0.875

Figura 19. Diagrama de Kaplan Meier

Análisis de coxph

	coef	exp(coef)	se(coef)	z	p
My_data.subtype_msi_statusG2	0.236	1.266	0.291	0.81	0.42

Likelihood ratio test=0.68 on 1 df, p=0.41
n= 375, number of events= 58

Figura 20. Modelo de riesgos proporcionales de Cox

Curvas de supervivencia

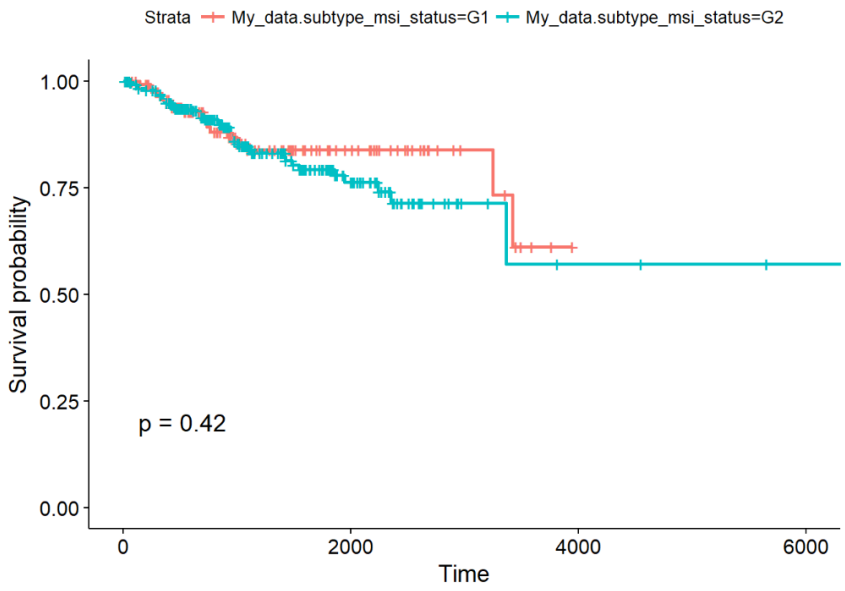

Figura 21. Curvas de supervivencia

5. Discusión

Si observamos los resultados obtenidos del análisis de pathways y coexpresión de genes, en el primer caso, que corresponde con los genes o transcritos sobre-expresados en el grupo 2 (MSS+MSS-L), podemos ver que existe una asociación significativa con pathways relacionados con la vía de señalización p53.

Sería interesante explotar la vía de señalización p53 como una posible diana terapéutica en pacientes con cáncer de endometrio MSS y MSI-l. Actualmente están pendiente de aprobación fármacos que actúen sobre esta via.

En cuanto a los genes o transcritos sobre-expresados en el grupo 1 (MSI-H), podemos destacar la mayor expresión de genes o transcritos asociados a pathways relacionados con el ciclo celular, en concreto ha llamado nuestra atención, los asociados con las ciclinas.

En la siguiente imagen podemos ver algunos de los pathways asociados al cáncer:

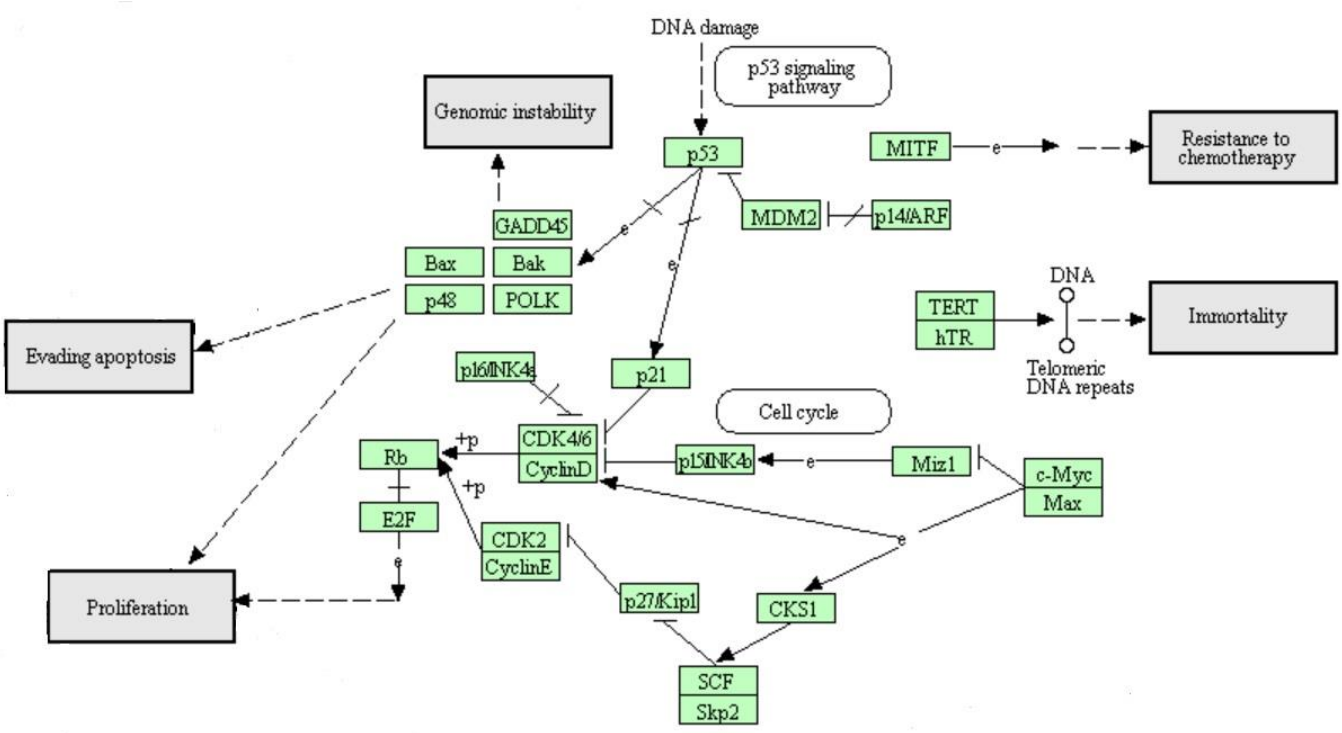

Figura 22. Pathways en cáncer.

Las ciclinas se encuentran entre los reguladores centrales del ciclo celular más importantes. Los niveles de ciclinas se encargan de marcar el ritmo de la división y mantener a la célula en la fase necesaria del ciclo. Por lo tanto existen diferentes tipos de ciclinas que marcan el inicio de la síntesis del DNA y el inicio de la mitosis.

Las ciclinas actúan mediante la regulación de la actividad de las enzimas quinasas dependientes de ciclinas (CDKs), uniéndose a ellas y abriendo su centro activo. Todo este proceso está controlado por numerosos activadores e inhibidores que regulan de forma controlada la progresión del ciclo celular.

Las células cancerosas pueden ser portadoras de mutaciones que modifican estos controles, bloqueando los puntos de control y promoviendo la división celular lo que conduce a una proliferación anormal de las células[16]. Por esto, las ciclinas y las CDKs se consideran importantes dianas terapéuticas.

Debido a la importancia de la actividad de CDK4 / 6 en las células cancerosas, los inhibidores de CDK4 / 6 se han convertido en fármacos prometedores en el tratamiento del cáncer[17].

La ciclina D1 y las CDK4/6 son factores en los que confluyen múltiples vías de señalización que conducen a la proliferación celular. A través de la inhibición de CDK4/6, palbociclib reduce la proliferación celular mediante el bloqueo de la progresión de la célula de la fase G1 a la fase S del ciclo celular.

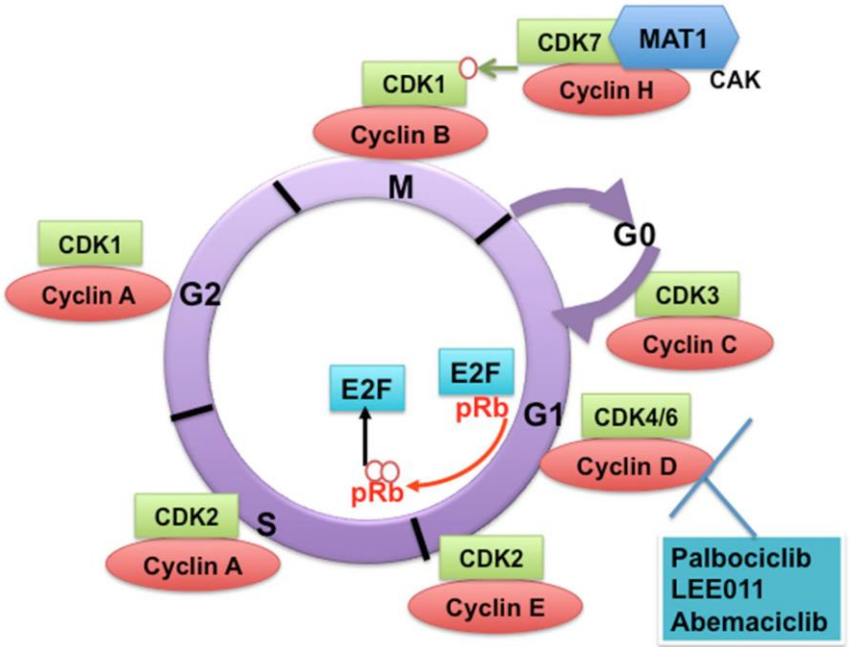

Figura 23. Cyclin-dependent kinases and their cyclin regulatory subunits.

Actualmente existe 3 fármacos desarrollados: Palbociclib, Ribociclib y Abemaciclib. Palbociclib fue el primer fármaco, dentro de la familia de los inhibidores de CDKs, aprobados tanto por la FDA como por la EMA. Palbociclib es un inhibidor altamente selectivo y reversible de las CDKs 4 y 6.

Palbociclib está indicado para el tratamiento del cáncer de mama metastásico o localmente avanzado, positivo para el receptor hormonal y negativo para el receptor

2 del factor de crecimiento epidérmico humano (HER2) en combinación con hormonoterapia debido a que consigue aumentos significativos en la supervivencia libre de progresión tanto en primera línea (24,8 meses en el grupo de palbociclib + letrozol frente a 14,5 meses en el grupo de letrozol; HR: 0,58)[18], como en pacientes previamente tratados (11,2 meses en el grupo de palbociclib + fulvestrant frente a 4,6 meses en el grupo de fulvestrant; HR: 0,49)[19] sin un efecto sobre la calidad de vida.

Dada la evidencia de la actividad de estos fármacos en el tratamiento de pacientes cáncer de mama, se ha analizado la actividad de Palbociclib en otros tipos de tumores, dentro de los que se encuentra el cáncer de endometrio. Palbociclib ha demostrado actividad en líneas celulares de cáncer de endometrio con deficiencia de *PTEN*[20] y en aquellas con expresión de la proteína Rb[21].

Teniendo en cuenta estos resultados, la vía de las ciclinas puede ser un nuevo objetivo donde centrar la estrategia en investigación en el cáncer de endometrio. Actualmente está en marcha un ensayo clínico fase II aleatorizado donde analiza la actividad de Palbociclib combinado con Letrozol o placebo en pacientes con cáncer de endometrio avanzado con expresión de receptores hormonales (NCT02730429).

Con respecto al análisis de supervivencia realizado en nuestro proyecto, no se objetivan diferencias estadísticamente significativas entre ambos grupos como hemos podido ver. Estos datos son semejantes a los obtenidos en el análisis de supervivencia del TCGA teniendo en cuenta el estatus de estabilidad o inestabilidad de microsatelites.

6. Conclusiones

En el campo de la oncología el estudio de nuevas dianas terapéuticas cada vez resulta algo más imprescindible, ya que se busca conseguir tratamientos dirigidos contra las alteraciones genéticas o moleculares específicas del tumor.

En nuestro caso hemos visto como los tumores de endometrio con un estatus mutacional de MSI-H presentan una sobreexpresión de genes o transcritos implicados en las vías de las ciclinas dependientes de quinasas. Por tanto, basándonos en los resultados que hemos obtenido, podríamos plantearnos que este tipo de tumores se podrían beneficiar de un tratamiento complementario con los inhibidores de ciclinas dependientes de quinasa (actualmente aprobados únicamente en cáncer de mama) como alternativa al tratamiento actual con quimioterapia o radioterapia.

En cuanto al seguimiento de la planificación y metodología todo se realizó según lo previsto, únicamente se decidió completar el análisis de enriquecimiento de pathways con un análisis de coexpresión de redes, con el objetivo de integrar las vías en las que participan nuestros genes o transcritos de interés, facilitando la búsqueda de dianas terapéuticas. Dado el buen seguimiento a lo largo del proyecto se han conseguido alcanzar todos los objetivos propuestos.

En cuanto a las líneas de trabajo futuro a corto plazo se está valorando la posibilidad de hacer públicos dichos resultados dada su transcendencia en la búsqueda de nuevas estrategias terapéuticas, más a largo plazo sería interesante confirmar estos resultados obtenidos mediante la realización de un ensayo clínico, donde se pudiese confirmar nuestra hipótesis. Además comentar que dada la limitación de la información disponible acerca del estatus mutacional en los datos del TCGA, solo hemos podido comparar MSI-H frente a MSS y MSI-L, sería interesante ampliar el estudio teniendo en cuenta la clasificación en los cuatro subgrupos moleculares del TCGA.

7. Glosario

CDKs: Cyclin-dependent kinases

cDNA: DNA complementario

CNH: Copy number high

CNL: Copy number low

CPM: Counts per millon

FDR: False Discovery Rate

FI: Functional Interaction

HER2: Receptor 2 del factor de crecimiento epidérmico humano

HR: Hazard Ratio

Log-FC: Log-Fold Change

MSI: microsatellite instability

MSI-H: High microsatellite instability

MSI-L: Low microsatellite instability

MSS: microsatellite stability

NCI: Instituto Nacional del Cáncer

NHGRI: Instituto Nacional de Investigación del Genoma Humano

Rb: Retinoblastoma

RNA-seq: RNA-sequencing

RNAm: RNA mensajero

TCGA: The Cancer Genome Atlas

TMM: Trimmed mean of M-values

8. Bibliografía

1. Targeted Therapy to Treat Cancer was originally published by the National Cancer Institute.

2. Galcerán J. Carcinoma de endometrio: Incidencia y mortalidad. Ginecologia y obstetricia clinica. 2003; 4: 8-11

3. Hayat MJ, Howlader N, Reichman ME, et al. Cancer statistics, trends, and multiple primary cancer analysis from the surveillance, epidemiology, and end results (SEER) program. The Oncologist. 2007; 12 :20-37.

4. Morice, Philippe et al. Cancer endometrial. The Lancet, 2015; 387: 1094-1108.

5. Lax SF. Molecular genetic pathways in various types of endometrial carcinoma: From a phenotypical to a molecular-based classification. Virchoes Archiv. 2004; 444: 213-223.

6. Okuda T, Sekizawa A, Purwosunu Y, et al. Genetics of endometrial cancers. Obstet Gynecol Int. 2010; 984013.

7. Endometrial Cancer Treatment (PDQ®)–Health Professional Version was originally published by the National Cancer Institute.

8. Stubert J, Gerber B. Current issues in the diagnosis and treatment of endometrial carcinoma. Geburtshilfe Frauenheilkd. 2016; 76: 170-175.

9. Binder PS, Mutch DG: Update on prognostic markers for endometrial cancer. Womens Health. 2014; 10: 277-88.

10. The Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Endometrial Carcinoma. Nature. 2013; 497: 67–73.

11. https://www.genome.gov/27562853/transcriptoma/ (Abril-Mayo)

12. RNA-Seq Data Comparison with Gene Expression Microarrays. A cross-platform comparison of differential gene expression analysis. White Paper: Sequencing. Illumina

13. Sebestyén E, Singh B, Miñana B, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. 2016

14. http://www.cytoscape.org/ (Abril-Mayo)

15. Wu G, Feng X , Stein L . A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 2010;11(5)

16. https://basessobreelcancer.weebly.com/ciclinas-y-cdk.html (Abril-Mayo)

17. Hamilton E. Infante JR. Targeting CDK4/6 in patients with cancer. Cancer Treat Rev. 2016; 45:129-38

18. Finn RS, Martin M, Rugo HS, et al. Palbociclib and Letrozole in Advanced Breast Cancer. N Engl JMed. 2016; 375: 1925-1936

19. Cristofanilli M, Turner NC, Bondarenko I, et al. Fulvestrant plus palbociclib versus fulvestrant plus placebo for treatment of hormone-receptor-positive, HER2-negative metastatic breast cancer that progressed on previous endocrine therapy (PALOMA-3): final analysis of the multicentre, double-blind, phase 3 randomised controlled trial. Lancet Oncol. 2016; 17: 425–39.

20. Dosil MA,Mirantes C, Eritja N, et al. Palbociclib has antitumour effects on Pten-deficient endometrial neoplasias. J Pathol. 2017; 242: 152-164.

21. Tanaka T, Terai Y, Ashihara K, et al. The efficacy of the cyclin-dependent kinase 4/6 inhibitor in endometrial cancer. Castresana JS, ed. *PLoS ONE*. 2017; 12: e0177019.

9. Anexos

ANEXO 1. Pipeline con el código R implementado

```
# Directorio de trabajo

setwd("C:/Users/MONICA/Desktop/MASTER/TFM")
workingDir <- getwd()


# Directorio carpeta resultados

resultsDir<-file.path(workingDir, "Results")

setwd(workingDir)
```

Obtención datos TCGA

```
library(TCGAbiolinks)
library (SummarizedExperiment)
library(DT)

# Selección de nuestros datos
query <- GDCquery(project = "TCGA-UCEC",
                  data.category = "Transcriptome Profiling",
                  data.type = "Gene Expression Quantification",
                  experimental.strategy = "RNA-Seq",
                  workflow.type = "HTSeq - Counts",
                  sample.type = c ("Primary solid Tumor"))
```

```
GDCdownload(query)
```

```
data <- GDCprepare(query)
```

```
# Selecciono muestras con datos de MSI-MSS
mss<-which(substr(data$subtype_msi_status_7_marker_call,1,7)=="MSS")
msiH<-which(substr(data$subtype_msi_status_7_marker_call,1,7)=="MSI-H")
msiL<-which(substr(data$subtype_msi_status_7_marker_call,1,7)=="MSI-L")

My_samples<-sort(c(mss,msiH, msiL))

My_data<-data[,My_samples]

# Creo la variable subtype_msi_status con mis dos grupos: MSI-H: G1 y MSS+MSI.L: G2
My_data$subtype_msi_status[My_data$subtype_msi_status_7_marker_call=="MSI-H"]<-"G1"
My_data$subtype_msi_status[My_data$subtype_msi_status_7_marker_call=="MSI-L"]<-"G2"
My_data$subtype_msi_status[My_data$subtype_msi_status_7_marker_call=="MSS"]<-"G2"


# Matriz
Matrix <- assay(My_data,"HTSeq - Counts")

# Dimensión
dim(Matrix)
```

```
# Lista de Datos
library(edgeR)
group<-as.factor(My_data$subtype_msi_status)
ListaDatos<-DGEList(Matrix, group=group)

# Anotaciones
library(org.Hs.eg.db)
symbol <- mapIds(org.Hs.eg.db,keys=rownames(ListaDatos),keytype="ENSEMBL",column="SYMBOL")

entrezid <- mapIds(org.Hs.eg.db,keys=rownames(ListaDatos),keytype="ENSEMBL",column="ENTREZID")
ListaDatos$genes <- data.frame(entrezid=entrezid, symbol=symbol)
```

Filtrado para eliminar los genes o transcritos poco expresados

```
# Calculo cpm y lcpm
cpm<-cpm(ListaDatos)
lcpm<-cpm(ListaDatos, log=TRUE)
```

```
# Filtrado
datos<-rowSums(cpm>1)>=15
ListaDatosF<-ListaDatos[datos, keep.lib.sizes = FALSE]
dim(ListaDatosF)
```

Tamaño de librerias y gráficos de distribución

```
# Gráfico del tamaño de la libreria
barplot(ListaDatosF$samples$lib.size,names=colnames(ListaDatosF),las=2)+
title("Barplot of library sizes")
```

```
lib.col <- c("green","pink")[group]
boxplot(lcpm, ylab="Log2 counts per million",col=lib.col,las=2, main="Boxplots of logCPMs: unnormalised")
```

Normalización

```
# Normalización
ListaDatosN<-calcNormFactors(ListaDatosF, method = "TMM")
head(ListaDatosN$samples,10)
```

```
lcpmN<-cpm(ListaDatosN, log=TRUE)

lib.col <- c("green","pink")[group]
boxplot(lcpmN, xlab="", ylab="Log2 counts per million", col= lib.col, las=2, main="Boxplots of logCPMs: normalised",pars=lis
t(cex.lab=0.8,cex.axis=0.8))
```

Análisis de expresión diferencial

```
# Matriz de diseño
design<-model.matrix(~0+group)
colnames(design) <- gsub("group", "", colnames(design))

# Matriz de contraste
contr.matrix<-makeContrasts(
  G1vsG2 = G1-G2,
  levels = c("G1","G2"))
contr.matrix
```

```
# Función Voom
v<-voom(ListaDatosN, design, plot=TRUE)
```

```
vfit<-lmFit(v, design)
vfit.main<-contrasts.fit(vfit, contrasts=contr.matrix)
efit<-eBayes(vfit.main)
plotSA(efit)
```

```
# topTable: para cada contrate se genera una lista de más a menos diferencialmente expresados)
topTabG1vsG2<-topTable(efit, n=nrow(ListaDatosN$counts),coef="G1vsG2", adjust = "fdr")

Genes<-topTabG1vsG2[topTabG1vsG2$adj.P.Val<=0.001,]

cat("Numero de genes con un p-valor inferior a 0.001 en la comparacion 'G1 vs G2': ", sum(topTabG1vsG2$adj.P.Val<=0.001),
"\n")
```

```
# 10 genes más diferencialmente expresados
library(Biobase)
library(xtable)
G1vsG2.10<-xtable(topTabG1vsG2[1:10,1:6],
    label='topTabG1vsG2',
    caption='10 genes mas expresados diferencialmente en la comparacion "G1 vs G2"')
print(G1vsG2.10, tabular.environment='longtable',floating=FALSE)
```

```
res<-decideTests(efit, method="separate", adjust.method="fdr", p.value = 0.001, lfc=1)
print(summary(res))
```

```
# Que filas tienen como mínimo una celda distinta de cero
sum.res.rows<-apply(abs(res),1,sum)
res.selected<-res[sum.res.rows!=0,]
```

Visualización de los perfiles de expresión

```
plotMD(efit, status=res[,1], main=colnames(efit)[1])
```

```
library(pheatmap)

o <- order(topTabG1vsG2$P.Value)
logCPM <- lcpmN[o[1:30],]
logCPM <- t(scale(t(logCPM)))

pheatmap(mat = logCPM, scale = "none", clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean", clustering_method = "average",
         cutree_rows = 4, fontsize = 6)
```

Análisis de supervivencia

```
library(survival)
library(survminer)
```

```
My_data$time<-ifelse(is.na(My_data$days_to_death), My_data$days_to_last_follow_up, My_data$days_to_death)

d<-data.frame(My_data$bcr_patient_barcode, My_data$subtype_msi_status, My_data$vital_status, My_data$time)

# Modelo coxph
xtabs(~My_data.vital_status+My_data.subtype_msi_status, data=d) %>% addmargins()
```

```
# Diagrama de Kaplan meier
sfit <- survfit(Surv(My_data.time,as.numeric(My_data.vital_status))~My_data.subtype_msi_status, data=d)
summary(sfit, times=seq(0,365*4,365))
```

```
# Curvas de supervivencia
ggsurvplot(sfit, conf.int=F, pval=TRUE)
```