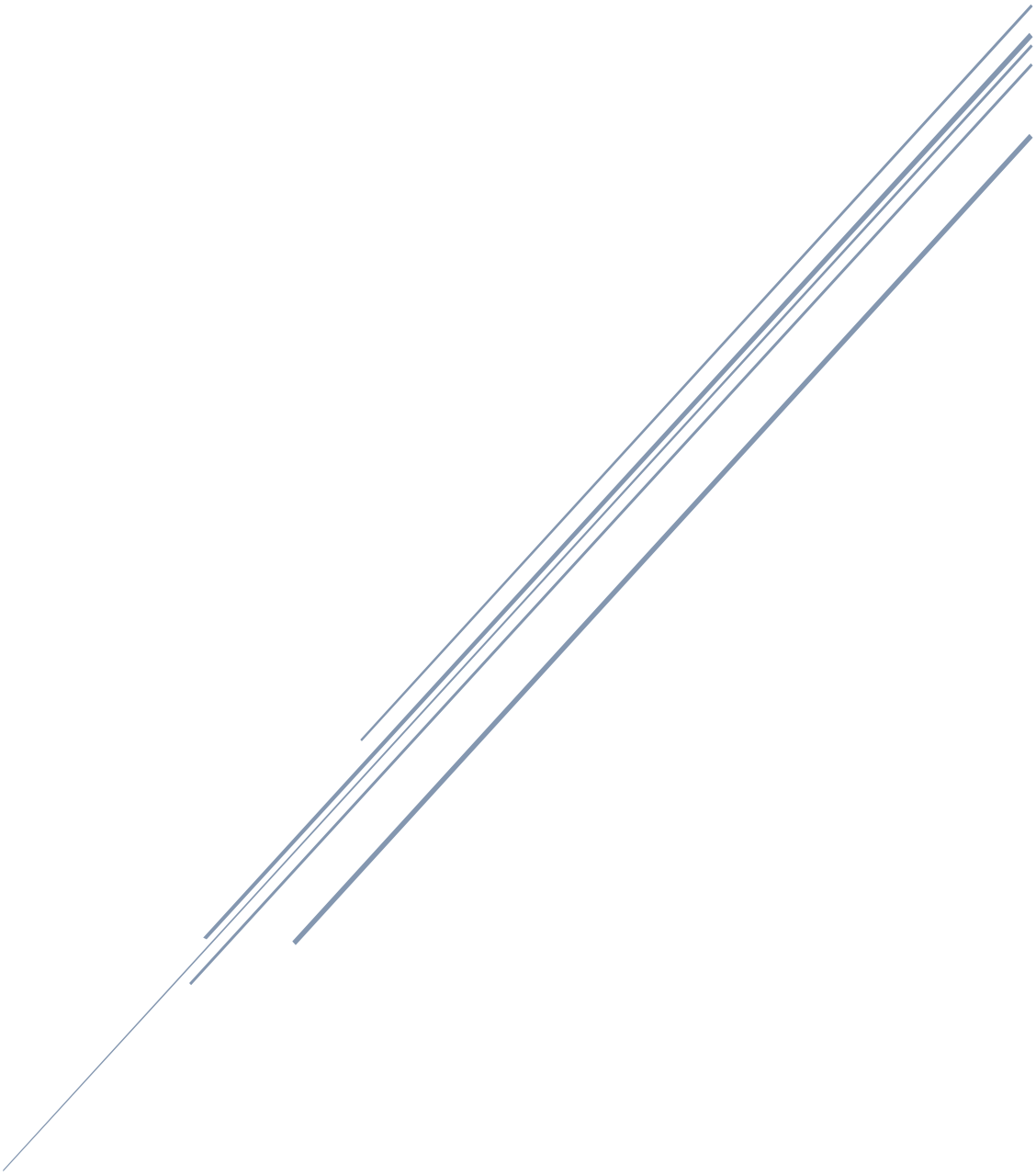
TRANSCRIPTÓMICA

Carlos Vivas Rodríguez



Instituto de Salud Carlos iii

Máster en Bioinformática aplicada a la Medicina Personalizada y la Salud 2023/24

1. APARTADO 1

En primer lugar, quiero destacar que este trabajo fue realizando en el ambiente "trabajo_transcriptomica" proporcionado en la carpeta "Ambiente" de la carpeta zip adjuntada. Igualmente se han usado muchos programas que comentaremos a lo largo del informe y de los que especificaremos la versión de uso de cada uno.

En este apartado hemos llevado a cabo un procedimiento basado en 3 pasos:

1- Un primer paso en el que nos hemos centrado en revisar la calidad de las muestras utilizando FastQC
2- Un segundo punto en el que se ha llevado a cabo el alineamiento utilizando HISAT2
3- Un último paso en el que se procedió al conteo de las lecturas con HTSEQ

i) CONTROL DE CALIDAD:

El control de calidad es un paso fundamental en el análisis de datos de secuenciación de ARN (RNA-seq). Nos permite evaluar la calidad de las lecturas de ARN de la secuenciación, crucial para asegurar la fiabilidad y precisión de los resultados de nuestro estudio. Mediante el control de calidad, podemos identificar posibles problemas técnicos que podrían afectar la interpretación de nuestros datos. Esto incluye la detección de errores de base, secuencias de baja calidad, adaptadores residuales, entre otros. Tales problemas podrían introducir sesgos o errores en nuestros datos, comprometiendo la validez de nuestras conclusiones biológicas.

Para llevar a cabo este procedimiento hemos hecho uso de FastQC, una herramienta de software ampliamente utilizada en el análisis de transcriptoma debido a su versatilidad y facilidad de uso. Nos evalúa la calidad de nuestras lecturas de ARN con varias métricas y gráficos. Una de las métricas más importantes que FastQC nos ofrece es la distribución de calidad de las bases. Esto nos permite evaluar la calidad de cada base a lo largo de nuestras lecturas, identificando posibles regiones de baja calidad que podrían requerir filtrado. Además, FastQC es una herramienta ampliamente utilizada y bien mantenida en la comunidad científica, lo que nos brinda confianza en la fiabilidad de sus resultados. En nuestro estudio, utilizamos la versión 0.12.1 de FastQC para llevar a cabo el control de calidad de nuestras muestras de RNA-Seq antes de realizar análisis posteriores. Obtenidas de la secuenciación, crucial para asegurar la fiabilidad y precisión.

Se ejecutó el comando fastqc desde la terminal de bash, esto nos abrió el programa FastQC. Una vez dentro, abrimos los cuatro archivos a analizar:

```
├── SRR479052.chr21_1.fastq
├── SRR479052.chr21_2.fastq
├── SRR479054.chr21_1.fastq
└── SRR479054.chr21_2.fastq
```

Una vez analizados, se generaron informes individuales para cada uno de ellos siendo guardados de dos maneras:

1. Un archivo .html que proporciona acceso a una dirección web donde se visualizan todos los diagramas relaccionados con la secuencia que ha sido analizada
2. Un archivo .zip que contendrá la misma información que el archivo .html, es decir los diagramas de la secuencia analizada.

Todos estos resultados podrán ser observados en: "`Apartado1/resultados/fastqc`"

Para llevar a cabo el análisis de los informes realizados por FastQC nos hemos fijado en:

1. Estadística Básica
2. Calidad de la secuencia por base
3. Puntuaciones de calidad por secuencia
4. Puntuaciones de calidad por base
5. Secuencias sobrerrepresentadas
6. Contenido de Adaptadores

1. Estadística Básica

Muestra	Total de Secuencias	Total de Bases	Secuencias de Baja Calidad	Longitud de Secuencia	%GC
SRR479052_1	15,340	1.5 Mbp	0	101	52
SRR479052_2	15,340	1.5 Mbp	0	101	52
SRR479054_1	9,746	984.3 kbp	0	101	51
SRR479054_2	9,746	984.3 kbp	0	101	51

Tabla1: Resumen de la estadística básica de las muestras, obtenida gracias a FastQC. Las muestras _1 y_2 en ambos casos presentan un número de secuencias muy parecidas, lo que representa una consistencia en la calidad y en la cantidad de los datos generados entre las réplicas técnicas de cada muestra. La longitud de la secuencia es de 101 bases, lo que indica que las lecturas de ARN tienen una longitud constante en todas las muestras analizadas. El porcentaje de FC es similar, consistencia en la composición de las secuencias de ARN entre las muestras. Ausencia de secuencias de baja calidad.

Tras analizar las estadísticas de las cuatro muestras de RNA-seq, podemos concluir que se han obtenido datos de alta calidad y consistencia. Las muestras tienen un número similar de secuencias y bases, y una longitud uniforme de 101 bases. Además, no se han detectado secuencias de baja calidad en ninguna de las muestras, lo que sugiere una calidad de secuenciación adecuada. Aunque existe una ligera variabilidad en el porcentaje de contenido GC entre las muestras, esta variación es mínima y no afecta significativamente a la calidad global de los datos. En conjunto, estos resultados respaldan la fiabilidad de los datos obtenidos y sientan las bases para análisis posteriores en busca de diferencias biológicas entre las condiciones experimentales analizadas.

2. Calidad de la secuencia por base

La calidad de la secuencia por base se refiere a la puntuación de calidad asignada a cada base individual en una secuencia de ADN o ARN durante el proceso de secuenciación.

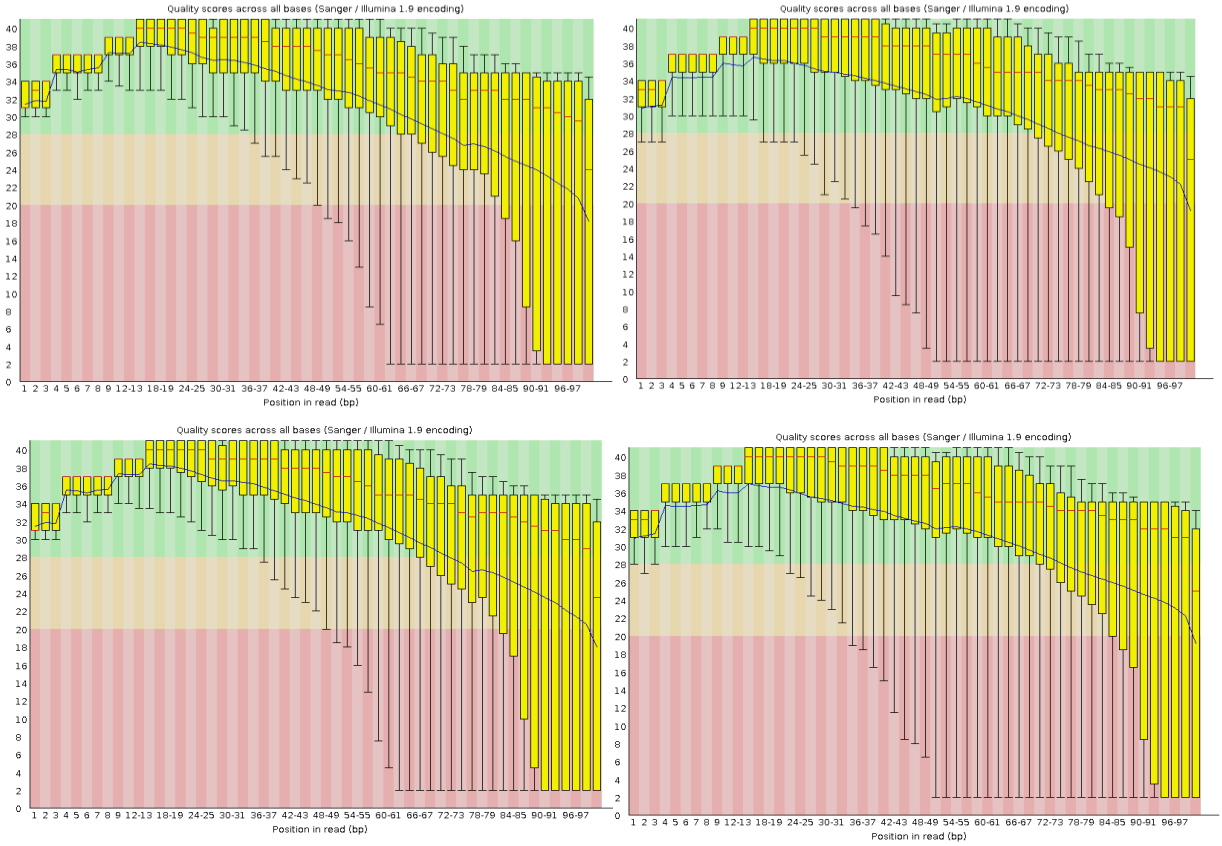


Figura 1: Per base sequence quality de FastQ

Como conclusión de estos gráficos observamos como a partir de la posición 66-67 y 72-73 la calidad de las secuencias deja de ser optima y pasa a ser de peor calidad. Y es a partir de la posición 84-85 que ya es mala calidad. Esto ocurre de igual manera en las 4 muestras. Esto lo que nos indica es que si queremos realizar un alineamiento correcto deberíamos recortar esas regiones.

3. Puntuaciones de calidad por secuencia

En el informe de FastQC, la puntuación de calidad por secuencia se refiere a la calidad de cada base individual en todas las secuencias presentes en el archivo de secuencia (por ejemplo, un archivo FASTQ). Esta puntuación de calidad se calcula utilizando el sistema de codificación de calidad utilizado en los datos de secuenciación, que suele ser Phred.

Cuanto mayor sea la puntuación de calidad, mayor será la confianza en la precisión de esa base. Por el contrario, una puntuación de calidad baja indica una mayor probabilidad de que la base sea incorrecta debido a ruido de fondo o artefactos de secuenciación.

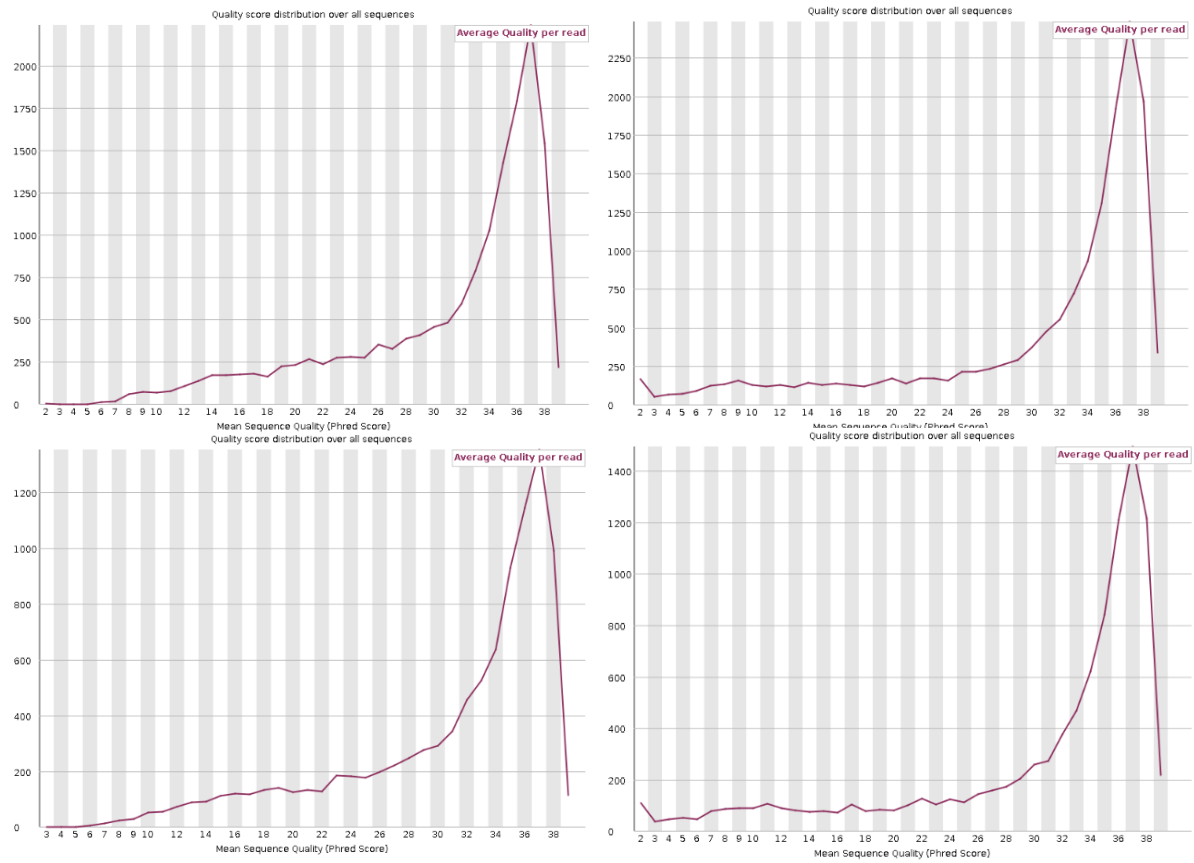


Figura 2: Se observa que hay una secuencia de alta calidad inciial, donde las bases tienen puntuaciones de calidad consistentemente altas. EL aumente repentino una vez llegado el valor 37 en el Phred Score indica una calidad muy buena de las secuencias obtenidas.

En conclusión, este patrón en el gráfico "Per sequence quality scores" respalda la confiabilidad de los datos de secuenciación y sugiere que son aptos para análisis detallados y precisos.

4. Puntuaciones de calidad por contenido

Este gráfico proporciona información sobre la distribución de las bases a lo largo de las secuencias analizadas. En él, el eje Y muestra el porcentaje acumulado de bases ordenadas por calidad, mientras que el eje X representa la posición de la base en la secuencia. Es esencial para identificar sesgos en la composición de las secuencias, como regiones ricas en GC o AT, que podrían indicar artefactos de secuenciación o problemas biológicos. Además, revela detalles sobre la estructura y composición del genoma estudiado. La distribución uniforme y equilibrada de los nucleótidos en las lecturas garantiza la calidad de los datos y una interpretación precisa de los resultados del análisis genómico, evitando sesgos que puedan afectar la representación precisa de la información genética.

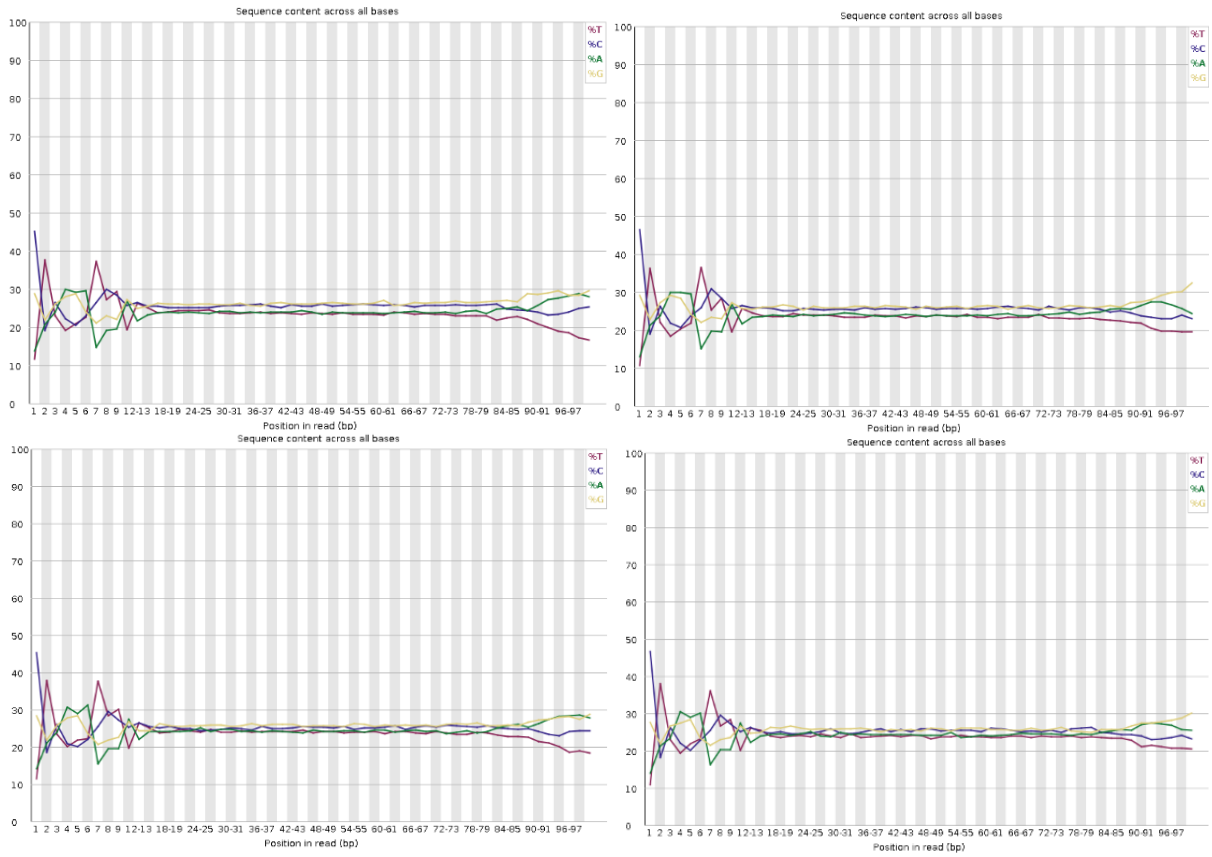


Figura 3: Las cuatro gráficas presentadas muestran características similares en cuanto a la distribución de las bases en las secuencias analizadas. Sin embargo, se observa una anomalía consistente en las primeras 12 o 13 pares de bases en todas las muestras. Estas anomalías podrían indicar la presencia de sesgos o distorsiones en esta región específica de las secuencias. Una posible explicación para estas distorsiones radica en la utilización de random hexamer primers durante el proceso de preparación de la librería de secuenciación. Los random hexamer primers son secuencias cortas de ADN sintético que se utilizan en la síntesis de ADN complementario (cDNA) durante la preparación de la librería para la secuenciación. Estas secuencias aleatorias se unen de manera no específica a las moléculas de ARN o ADN y se utilizan como cebadores para la síntesis de cDNA. Sin embargo, debido a su naturaleza aleatoria, los random hexamer primers pueden introducir sesgos en la secuencia durante la etapa de síntesis de cDNA, lo que puede resultar en distorsiones observadas en las primeras bases de las secuencias. Estas distorsiones podrían afectar la interpretación de los datos de secuenciación y requerir ajustes tanto en los protocolos de laboratorio como en el análisis bioinformático para mitigar sus efectos.

En conclusión, las gráficas presentan características similares en la distribución de las bases en las secuencias analizadas, excepto anomalías consistentes en las primeras 12 o 13 pares de bases. Estas anomalías podrían ser atribuibles a la presencia de sesgos introducidos por los random hexamer primers utilizados durante la preparación de la librería de secuenciación. Los random hexamer primers, al ser secuencias aleatorias, pueden generar distorsiones en la secuencia durante la síntesis de cDNA, lo que se refleja en las primeras bases de las secuencias obtenidas.

5. Secuencias sobrerrepresentadas

Las secuencias sobrerrepresentadas en los informes de FastQC son aquellas secuencias que aparecen con una frecuencia mucho mayor de lo esperado en el conjunto de datos de secuenciación. Esto puede ocurrir debido a varias razones, como la presencia de adaptadores de secuenciación, secuencias de control de calidad, contaminantes o secuencias repetitivas en el genoma o en la muestra analizada.

La identificación de secuencias sobrerrepresentadas es importante porque puede indicar la presencia de artefactos de secuenciación o contaminantes que podrían afectar la precisión y fiabilidad de los datos.

Sample	Sequence	Count	Percentage	Possible Source
SRR479052.chr21_1.fastq	CTTTTACTTCCTCTAGATAGTCAAGTTCGACCGTCTTCTCAGCGCTCCGC	21	13.6897	No Hit
SRR479052.chr21_2.fastq	CTAACACGTGCGCGAGTCGGGGGCTCGCACGAAAGCCGCCGTGGCGCAAT	20	13.0378	No Hit

Tabla 2: Secuencias sobrerrepresentadas

La tabla revela la presencia de dos secuencias distintas en los archivos SRR479052.chr21_1.fastq y SRR479052.chr21_2.fastq, con recuentos de 21 y 20, respectivamente. Esto sugiere diversidad genética en las muestras. Sin embargo, no se ha identificado una fuente específica para estas secuencias ("No Hit"), lo que indica que su origen es desconocido o no corresponde a secuencias conocidas. Estos hallazgos son relevantes para comprender la composición genética de las muestras, aunque requieren investigación adicional para determinar su significado biológico y potencial importancia.

6. Contenido de adaptadores

Se refiere a la proporción de secuencias en los datos de secuenciación que contienen adaptadores de secuenciación. Los adaptadores son secuencias cortas de ADN diseñadas para unirse específicamente a las secuencias de interés durante el proceso de preparación de la librería para la secuenciación.

Cuando se realiza la secuenciación, es importante que los adaptadores se eliminen o se secuencien por completo para evitar interferencias en la interpretación de los datos. En algunos casos, los adaptadores pueden permanecer unidos a las secuencias y detectarse durante el análisis de calidad de los datos de secuenciación.

Por lo tanto, el contenido de adaptadores en los informes de FastQC indica la proporción de secuencias que contienen adaptadores en los datos de secuenciación. Un alto contenido de adaptadores puede sugerir problemas durante el proceso de preparación de la librería, como una insuficiente eliminación de los adaptadores o una contaminación de la muestra con adaptadores. Esto puede afectar la calidad de los datos y requerir acciones correctivas durante el análisis de los datos de secuenciación.

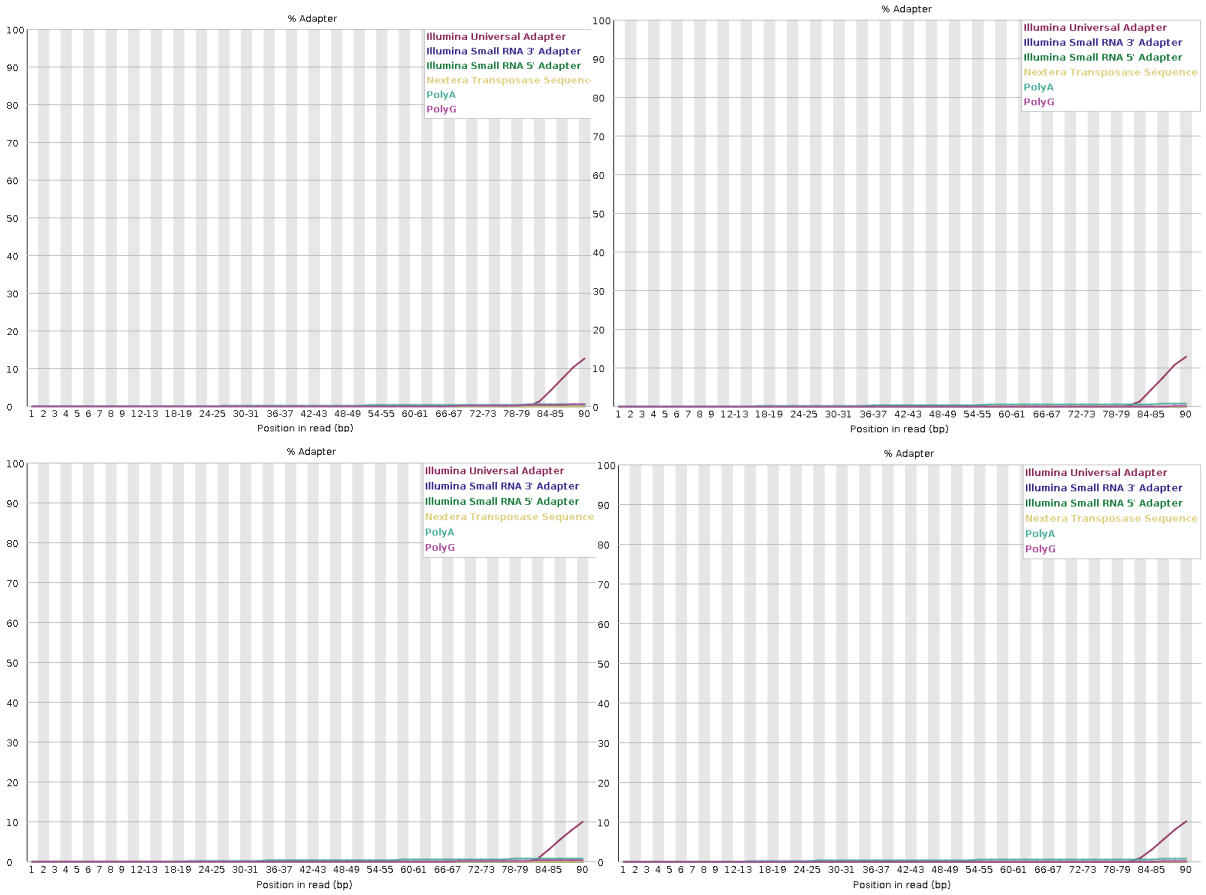


Figura 4: Gráficas de 'Adapter Content' mostrando la presencia del Illumina universal adapter a partir de la posición de la base 84-85 en todas las muestras analizadas. El aumento gradual del eje Y indica una proporción creciente de secuencias con adaptadores, lo que sugiere optimizar los pasos de eliminación durante la preparación de la librería para garantizar una secuenciación de alta calidad y resultados precisos del análisis genómico.

En conclusión, el análisis del "Adapter Content" revela una presencia significativa del Ilumina universal adapter a partir de la posición de la base 84-85 en todas las muestras, con un aumento gradual en la proporción de secuencias que contienen adaptadores. Estos hallazgos indican la necesidad de mejorar los procedimientos de eliminación de adaptadores durante la preparación de la librería para garantizar una secuenciación de alta calidad y resultados precisos en el análisis genómico posterior.

ii) ALINEAR EL GENOMA

EL proceso de alineamiento e indexado del genoma es una etapa fundamental en el análisis de datos de secuenciación a una referencia genómica conocida, lo que permite identificar la ubicación y la abundancia relativa de las secuencias en el genoma de interés.

Para este procedimiento se ha hecho uso del programa HISAT2 (versión 2.2.1):

Para llevar a cabo el alineamiento de las secuencias de ARN, se utilizó HISAT2, una herramienta de alineación de secuencias de ARN de alto rendimiento desarrollada específicamente para el mapeo de lecturas de ARN a un genoma de referencia.

HISAT2 utiliza un enfoque de mapeo basado en alineaciones de extremo dividido (split-read) y alineaciones de emparejamiento de extremos (paired-end), lo que le permite manejar de manera eficiente las lecturas cortas y largas producidas por las tecnologías modernas de secuenciación de ARN.

Su principal función es mapear las secuencias de ARN a la referencia genómica con alta precisión y sensibilidad, teniendo en cuenta las características específicas de las lecturas de ARN, como las lecturas divididas y las lecturas emparejadas. Al realizar el alineamiento con HISAT2, se puede determinar la ubicación precisa de cada secuencia en el genoma de referencia, lo que facilita el análisis posterior, como la cuantificación de la expresión génica y la identificación de variantes genéticas.

En primer lugar, antes de realizar el alineamiento, es necesario indexar el genoma, para ello se llevaron a cabo diferentes pasos.

En primer lugar, siempre después de haber generado la carpeta “index” hemos ejecutado el siguiente comando:

```
hisat2-build --seed 123 –p 2 input/Homo_sapiens.GRCh38.dna.chromosome.21.fa
resultados/index/chr21_index
```

Presenta la siguiente expliación:

- `hisat2-build:` Es el comando principal que invoca la herramienta HISAT2 para construir el índice.
- `--seed 123:` Especifica la semilla utilizada para la generación de números pseudoaleatorios. Esto puede ser útil para reproducir resultados o para controlar el comportamiento del algoritmo.
- `-p 2:` Especifica el número de subprocesos o hilos a utilizar durante la construcción del índice. En este caso, se utilizan 2 subprocesos para acelerar el proceso en sistemas con múltiples núcleos de procesamiento.
- `input/Homo_sapiens.GRCh38.dna.chromosome.21.fa:` Es la ruta al archivo FASTA que contiene la secuencia del cromosoma 21 del genoma humano de referencia GRCh38. Este archivo se utiliza como entrada para construir el índice.
- `resultados/index/chr21_index:` Especifica la ubicación y el nombre del archivo de salida que contendrá el índice generado. En este caso, el índice se guarda en el directorio "resultados/index" con el nombre "chr21_index".

Posteriormente realizamos la alineación del genoma:

Debemos tener en cuenta que tenemos 4 lecturas de 2 muestras, es decir, SRR479052 y SRR479054. Por lo tanto, tendremos dos códigos para cada una de las muestras.

```
hisat2 --new-summary --summary-file resultados/hisat2/SRR479052.hisat2.summary --rna-
strandness R --seed 123 --phred33 -p 2 -k 1 -x resultados/index/chr21_index -1
input/SRR479052.chr21_1.fastq -2 input/SRR479052.chr21_2.fastq -S
resultados/hisat2/SRR479052.chr21.sam
```

```
hisat2 --new-summary --summary-file resultados/hisat2/SRR479054.hisat2.summary --rna-
strandness R --seed 123 --phred33 -p 2 -k 1 -x resultados/index/chr21_index -1
input/SRR479054.chr21_1.fastq -2 input/SRR479054.chr21_2.fastq -S
resultados/hisat2/SRR479054.chr21.sam
```

Explicación del comando:

- `--new-summary:` Solicita que se genere un archivo de resumen de salida.
- `--summary-file resultados/hisat2/SRR479052.hisat2.summary:` Especifica la ruta y el nombre del archivo de resumen que se generará.
- `--rna-strandness R:` Indica la orientación esperada de las secuencias de ARN, en este caso, en la hebra inversa (reverse strand).
- `--seed 123`: Especifica la semilla utilizada para la generación de números pseudoaleatorios, asegurando reproducibilidad.
- `--phred33:` Indica que las secuencias de entrada están codificadas utilizando el esquema de puntuación Phred +33.
- `-p 2:` Establece el número de subprocesos o hilos para el procesamiento paralelo, en este caso, 2.
- `-k 1:` Indica que se debe informar solo un alineamiento por lectura.
- `-x resultados/index/chr21_index:` Especifica el índice genómico utilizado para el alineamiento.
- `-1 input/SRR479052.chr21_1.fastq -2 input/SRR479052.chr21_2.fastq:` Especifica los archivos FASTQ de las lecturas 1 y 2.
- `-S resultados/hisat2/SRR479052.chr21.sam:` Especifica la ruta y el nombre del archivo SAM de salida que contendrá los resultados del alineamiento.

Una vez se lleva a cabo la formación de los nuevos archivos, ". sam" se lleva a cabo de la misma manera dos archivos ".summary " los cuáles aportan información sobre la calidad del alineamiento.

Los archivos ".summary" generados por HISAT2 contienen información resumida sobre el proceso de alineamiento de las secuencias de ARN con respecto a la referencia genómica. Estos archivos son fundamentales para evaluar la calidad y el éxito del alineamiento, ya que proporcionan estadísticas clave y métricas relacionadas con el proceso de mapeo de las secuencias de ARN a la referencia genómica. Al analizar estos archivos, los investigadores pueden obtener una visión general de la eficiencia del alineamiento, la proporción de secuencias alineadas correctamente, la cobertura del genoma, la tasa de alineamiento y otros parámetros importantes que son cruciales para interpretar los resultados del análisis transcriptómico. Además, los archivos ".summary" facilitan la comparación entre diferentes muestras y condiciones experimentales, lo que permite identificar posibles variaciones en la calidad del alineamiento y en la expresión génica entre ellas.

En primer lugar:

```
SRR479052.hisat2.summary:

        Total pairs: 15340
                Aligned concordantly or discordantly 0 time: 6663 (43.44%)
Aligned concordantly 1 time: 7061 (46.03%)
Aligned concordantly >1 times: 0 (0.00%)
Aligned discordantly 1 time: 1616 (10.53%)
Total unpaired reads: 13326
Aligned 0 time: 6636 (49.80%)
Aligned 1 time: 6690 (50.20%)
Aligned >1 times: 0 (0.00%)
Overall alignment rate: 78.37%
```

El resumen generado por HISAT2 ofrece una visión general del proceso de alineamiento de las secuencias de ARN respecto a la referencia genómica. Se observan dos conjuntos de estadísticas: uno para los pares de secuencias y otro para las lecturas no pareadas. En cuanto a los pares de secuencias, se procesaron un total de 15,340 pares, de los cuales alrededor del 43.44% no se alinearon en ninguna ocasión, mientras que el 46.03% se alinearon correctamente una sola vez. No se encontraron alineamientos concordantes múltiples, pero alrededor del 10.53% se alinearon discordantemente una sola vez. En cuanto a las lecturas no pareadas, de un total de 13,326 lecturas, casi el 49.80% no se alinearon, mientras que aproximadamente el 50.20% se alinearon correctamente una sola vez. No hubo lecturas que se alinearon más de una vez. La tasa general de alineamiento fue del 78.37%, indicando un éxito significativo en el proceso de alineamiento. Estos datos son cruciales para evaluar la calidad y el rendimiento del proceso de mapeo de las secuencias de ARN, lo que es esencial para la interpretación precisa de los resultados del análisis transcriptómico.

En segundo lugar:

```
SRR479054.hisat2.summary:

        Total pairs: 9746

                Aligned concordantly or discordantly 0 time: 4043 (41.48%)
Aligned concordantly 1 time: 4852 (49.78%)
Aligned concordantly >1 times: 0 (0.00%)
Aligned discordantly 1 time: 851 (8.73%)
Total unpaired reads: 8086
Aligned 0 time: 4044 (50.01%)
Aligned 1 time: 4042 (49.99%)
Aligned >1 times: 0 (0.00%)
Overall alignment rate: 79.25%
```

El resumen de HISAT2 para este conjunto de datos presenta los siguientes hallazgos: Se procesaron un total de 9,746 pares de secuencias, de los cuales aproximadamente el 41.48% no se alinearon en ninguna ocasión, mientras que el 49.78% se alinearon correctamente una sola vez. No se encontraron alineamientos concordantes múltiples y alrededor del 8.73% se alinearon discordantemente una sola vez. En cuanto a las lecturas no pareadas, de un total de 8,086 lecturas, aproximadamente el 50.01% no se alinearon, mientras que casi el 49.99% se alinearon correctamente una sola vez. No hubo lecturas que se alinearon más de una vez. La tasa general de alineamiento fue del 79.25%, lo que indica un éxito notable en el proceso de alineamiento. Estos resultados son esenciales para evaluar la calidad y la eficacia del proceso de mapeo de las secuencias de ARN y son fundamentales para una interpretación precisa de los resultados del análisis transcriptómico.

A continuación, vamos a llevar a cabo la última sección del apartado 1. En este apartado, nos centraremos en la cuantificación de la expresión génica a partir de los archivos alineados obtenidos en el proceso anterior. Una vez que hemos alineado las secuencias de ARN con el genoma de referencia, es fundamental cuantificar la expresión de los genes para comprender mejor los patrones de expresión génica en nuestras muestras. Para lograr esto, utilizaremos herramientas de cuantificación como HTSeq, que nos permiten asignar las secuencias alineadas a las características genómicas, como exones o genes, y contar el número de lecturas que se superponen con estas características. Esta cuantificación nos proporcionará una medida de la expresión relativa de cada gen en nuestras muestras, lo que nos ayudará a identificar genes diferencialmente expresados y a comprender los procesos biológicos subyacentes en nuestro estudio. A lo largo de este apartado, detallaremos los comandos y parámetros utilizados en el proceso de cuantificación de expresión, justificando nuestras elecciones y asegurando una interpretación precisa de los resultados obtenidos.

Los comandos utilizados en este procedimiento son:

Para nuestra primera muestra SRR479052:

```
htseq-count --format=sam --stranded=reverse --mode=intersection-nonempty --minaqual=10 --type=exon --idattr=gene_id --additional-attr=gene_name resultados/hisat2/SRR479052_chr21.sam input/Homo_sapiens.GRCh38.109.chr21.gtf > resultados/htseq/SRR479052_chr21.htseq
```

Para nuestra segunda muestra SRR479054:

```
htseq-count --format=sam --stranded=reverse --mode=intersection-nonempty --minaqual=10 --type=exon --idattr=gene_id --additional-attr=gene_name resultados/hisat2/SRR479054_chr21.sam input/Homo_sapiens.GRCh38.109.chr21.gtf > resultados/htseq/SRR479054_chr21.htseq
```

- `htseq-count:` Es el comando principal de HTSeq para realizar el conteo de lecturas.
- `--format=sam:` Especifica que el archivo de entrada está en formato SAM.
- `--stranded=reverse:` Indica que las lecturas se alinearon en modo direccional inversa, lo que implica que la biblioteca de secuenciación es de tipo direccional.
- `--mode=intersection-nonempty:` Este modo de conteo asigna una lectura a una característica genómica si se superpone con al menos una base de esa característica, lo que evita contar características genómicas vacías.
- `--minaqual=10:` Establece el umbral mínimo de calidad de la lectura. Solo se contarán las lecturas con una calidad de al menos 10.
- `--type=exon:` Especifica que se realizará el conteo basado en exones.
- `--idattr=gene_id:` Indica que se utilizará el atributo 'gene_id' del archivo GTF como identificador del gen.
- `--additional-attr=gene_name:` Permite incluir atributos adicionales del archivo GTF en el archivo de salida. En este caso, se incluye el nombre del gen.
- `resultados/hisat2/SRR479054_chr21.sam:` Es el archivo SAM generado por HISAT2 que contiene las lecturas alineadas.
- `input/Homo_sapiens.GRCh38.109.chr21.gtf:` Es el archivo GTF que contiene la anotación genómica, necesario para asignar las lecturas a características genómicas.
- `resultados/htseq/SRR479054_chr21.htseq:` Es el archivo de salida que contendrá el conteo de lecturas para cada gen.

Para llevar a cabo el análisis del archivo ".htseq" generado previamente, hemos utilizado MultiQC, una herramienta que nos permite generar informes de calidad a partir de múltiples resultados de análisis bioinformáticos. En este caso, hemos utilizado MultiQC para recopilar y visualizar de manera integrada los datos de cuantificación de expresión génica obtenidos con HTSeq. Este proceso nos permite obtener una visión general de la calidad y los resultados del análisis transcriptómico en un informe único y fácilmente interpretable.

Después de ejecutar MultiQC sobre los resultados de HTSeq count, se obtiene el "HTSeq count assignments". Este resultado proporciona información detallada sobre el recuento de lecturas asignadas a cada característica genómica, como exones o genes, según lo realizado por HTSeq.

El HTSeq count assignments muestra el número de lecturas alineadas y asignadas a cada gen o característica genómica específica. Este archivo es esencial para comprender la expresión génica en el

conjunto de datos analizado. Cada línea del archivo representa una característica genómica, seguida por el número de lecturas asignadas a esa característica.

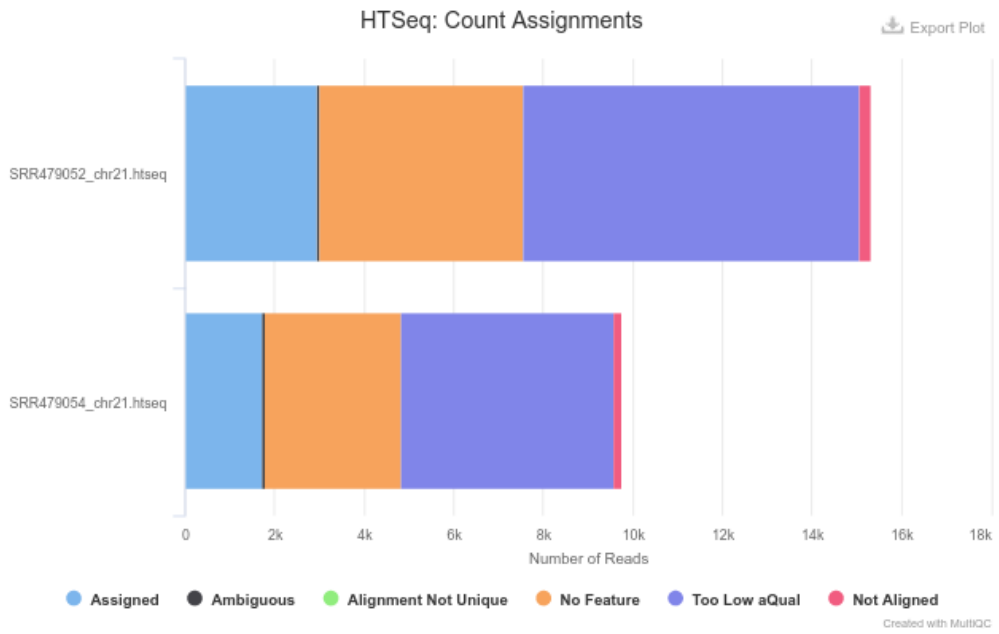


Figura 5 :Se percibe una predominancia de lecturas no asignadas debido a su baja calidad (resaltadas en morado como "Too Low aQual") o a la falta de asignación a una función específica (marcadas en naranja como "No Feature"). Por otro lado, se distinguen en tono azul claro las lecturas que han sido contabilizadas y asignadas a una anotación. Además, se observan en menor proporción lecturas no alineadas (en rosa como "Not Aligned") y lecturas ambiguas (en negro como "Ambiguous"). Esta distribución de colores y categorías revela la calidad y la asignación de las lecturas analizadas, ofreciendo una visión detallada pero concisa de los resultados.

iii) CONCLUSIÓN APARTADO 1

En esta primera parte del análisis transcriptómico, hemos llevado a cabo una serie de procesos que incluyen el control de calidad de las muestras, el alineamiento y la cuantificación de la expresión génica. Mediante herramientas como FastQC y MultiQC, hemos evaluado la calidad de los datos de secuenciación y hemos generado informes detallados para una comprensión exhaustiva de la integridad de nuestras muestras. Además, mediante el uso de HTSeq, hemos cuantificado la expresión génica y hemos obtenido información valiosa sobre la actividad transcriptómica en nuestro conjunto de datos. En conjunto, estos análisis nos han proporcionado una base sólida para avanzar en nuestra investigación transcriptómica, permitiéndonos identificar patrones de expresión génica y prepararnos para abordar preguntas más específicas en la siguiente etapa del estudio.

2. APARTADO 2

A continuación, nos centraremos en identificar los genes que muestran diferencias significativas en su expresión entre las muestras tratadas con OHT y las de control. Este análisis de los genes diferencialmente expresados (DEG) es fundamental para comprender como los tratamientos estan afectando a la actividad genética y como pueden llegar a proporcionar información alto valor sobre los mecanismos biológicos involucrados en la respuesta al tratamiento, así como conocer cuáles son las posibles dianas terapéuticas biomarcadores de respuesta al tratamiento.

i) ANÁLISIS DE LA EXPRESIÓN DIFERENCIAL DE GENES

Para llevar a cabo este análisis se ha hecho uso de RStudio en su versión 4.3.3, donde se usaron varios paquetes que nos permiten realizar estudios específicos y aportarnos una visualización mejor de los resultados. Entre ellos hemos utilizado:

I. DESeq2 en su versión 1.42.1, un paquete fundamental para el análisis de expresión génica diferencial que nos permite identificar los genes que muestran cambios significativos en su expresión entre condiciones.
II. Tidyverse en su versión 2.0.0 fue utilizado para realizar manipulaciones y visualizaciones de datos de manera más eficiente y coherente
III. Pheatmaap en su versión 1.0.12, que nos permite visualizar de manera efectiva la expresión diferencial de genes en forma de mapas de calor.
IV. RColorBrewer en su versión 1.1.3, un paquete que proporciona paletas de colores atractivas y fácilmente interpretables para nuestras gráficas.
V. Vsn en su versión 3.70.0 fue utilizado para normalizar los datos de expresión génica y corregir posibles sesgos técnicos en nuestros datos.
VI. Dplyr en su versión 1.1.4 proporciona una variedad de funciones que permiten realizar operaciones como filtrado, selección, reordenamiento y agregación de datos de una manera sencilla y eficiente.

Una vez establecidos cuáles fueron el programa y el paquete usado hablaremos de los diferentes puntos que hemos ido atravesando.

1. PRIMER TRATAMIENTO DE LOS DATOS

En primer lugar, se han cargado los datos que encontramos en los archivos “rawcounts.tsv” y “metadata.tsv”, matriz de cuentas crudas y metadatos del experimento, respectivamente. Una vez cargados hemos comprobado que coinicde las filas del metadata.tsv con las columnas del rawcounts.tsv.

Finalmente se estableció una nueva columna denominada “group” que combina las variables “agent” y “time”.

```
raw_data <- read.csv(file = "input/rawcounts.tsv", sep = "\t", row.names = 1)

colnames(raw_data)
experiment_data <- read.csv(file = "input/metadata.tsv", sep = "\t")
rownames(experiment_data) <- colnames(raw_data)
```

```
experiment_data <- mutate(.data = experiment_data,
                          X = NULL,
                          patient = as.factor(patient),
                          agent = as.factor(agent),
                          time = as.factor(time))
all(colnames(raw_data) %in% rownames(experiment_data))
all(colnames(raw_data) == rownames(experiment_data))
## Creación nueva variable
experiment_data$group <- as.factor(paste0(experiment_data$agent, experiment_data$time))
levels(experiment_data$group)
```

2. CREACIÓN DEL DESeqDataSet

En este punto, se describe en el párrafo es una metodología para diseñar un objeto DESeqDataSet, que es una estructura de datos utilizada en el paquete DESeq2 en R para el análisis de expresión génica diferencial en datos de secuenciación de ARN. En nuestro caso se modelarán los datos utilizando la variable "patient" (paciente) y la variable "group" que ha sido creada previamente. La variable "group" se ha formado combinando las variables de tratamiento y tiempo, lo que permite analizar simultáneamente el efecto del tratamiento y el tiempo sin la necesidad de especificar una interacción entre ambas variables en el diseño del análisis.

```
dds <- DESeqDataSetFromMatrix(countData = raw_data,
                              colData = experiment_data,
                              design = ~ patient + group)
dds
```

Una vez obtenido este dataset llevamos a cabo un primer filtrado, la eliminación de genes con menos de 10 lecturas se realiza para mejorar la calidad de los datos y la eficiencia del análisis. Los genes con lecturas muy bajas pueden ser poco fiables y añadir ruido al análisis. Al eliminarlos, se centra el análisis en genes más relevantes y se reduce la carga computacional. Esto simplifica el análisis y asegura que se estén considerando solo los genes más significativos para el estudio.

```
keep <- rowSums(counts(dds)) >= 10
dds2 <- dds[keep, ]
```

Una vez hemos obtenido el dataset y ya lo hemos filtrado de tal manera que nos hemos acabado quedando con 24416 genes, reduciendo práticamente a la mitad el número inicial de genes, reaalizaremos un análisis exploratorio.

3. TRANSFORACIÓN ESTABILIZADORA DE LA VARIANZA (VST)

En este estudio, hemos utilizado la transformación estabilizadora de la varianza (VST) como parte del análisis exploratorio de nuestros datos de expresión génica. La transformación VST es una herramienta utilizada en el análisis de datos de secuenciación de ARN para estabilizar la varianza según el nivel de expresión. Esto ayuda a corregir la heterocedasticidad y a mejorar la comparabilidad entre las muestras. Es una herramienta que nos ayudará normalizar y estabilizar la varianza y obtendremos una matriz homocedastica.

```
vsd <- vst(dds2, blind = TRUE)
```

- vsd: Se crea una nueva variable llamada vsd, que almacenará los resultados de la transformación estabilizadora de la varianza (VST) aplicada a los datos contenidos en el objeto dds2.
- vst(): Esta es una función del paquete DESeq2 en R que aplica la transformación estabilizadora de la varianza a un objeto DESeqDataSet. La VST es una técnica utilizada en el análisis de datos de expresión génica para estabilizar la varianza según su nivel.
- dds2: Es el objeto DESeqDataSet que contiene los datos de expresión génica que se utilizarán como entrada para la transformación VST.
- blind = TRUE: Este parámetro se refiere a si se debe realizar la transformación de manera "ciega" o no. Cuando blind = TRUE, la transformación se realiza sin tener en cuenta las etiquetas de las muestras, lo que es útil para evitar sesgos.

A continuación, normalizamos los datos y se formaron las gráficas de dispersión de media frente a desviación estándar, obteniendo esto:

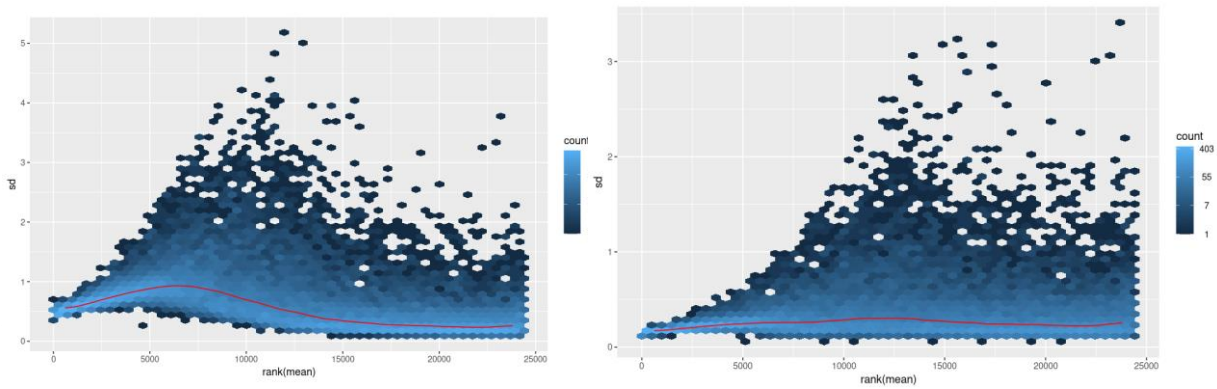


Figura 6: Transformación estabilizadora de la varianza (vst)

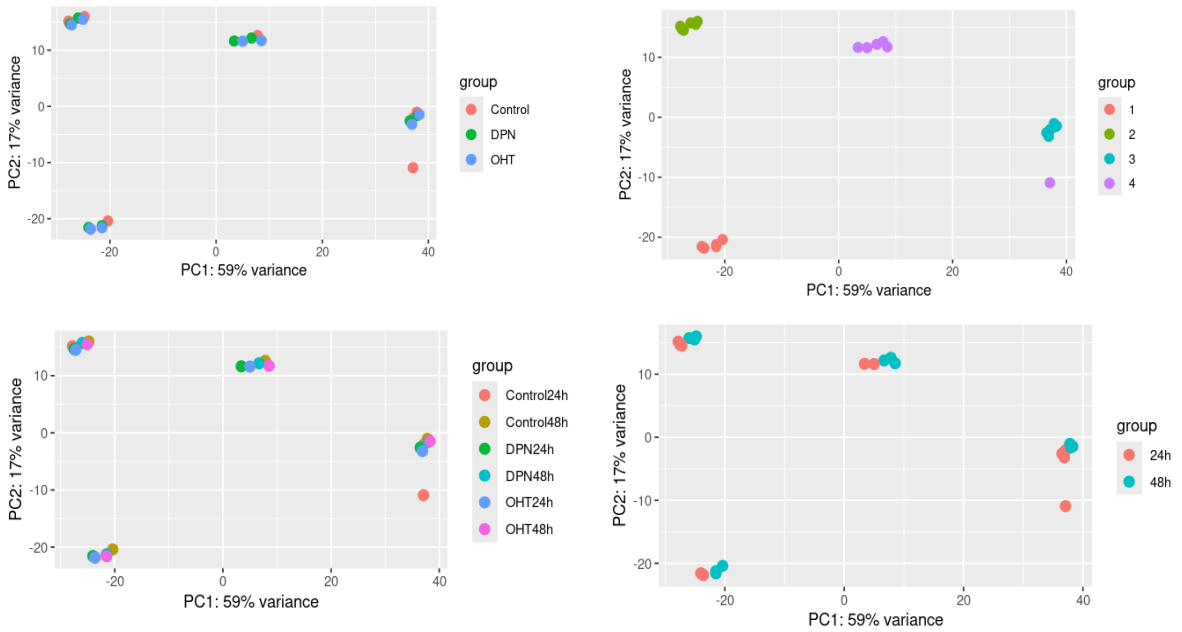
4. Exploración de la estructura y agrupamientos de muestras mediante Análisis de Componentes Principales (PCA)

Se realizó un análisis de componentes principales (PCA) para explorar la estructura de los datos de expresión génica después de aplicar la transformación estabilizadora de la varianza (VST).

El PCA es una técnica de reducción de dimensionalidad que ayuda a visualizar la variabilidad en los datos y a detectar patrones de agrupamiento entre las muestras. En este análisis, se representaron gráficamente los datos en un espacio de menor dimensión definido por los componentes principales.

Se realizaron cuatro gráficos PCA, cada uno coloreado según una variable diferente: "patient", "group", "agent" y "time". Estos gráficos permiten explorar cómo las muestras se agrupan o dispersan en función de estas variables, lo que puede proporcionar información sobre patrones biológicos o experimentales presentes en los datos.

```
plotPCA(vsd, intgroup = "patient")
plotPCA(vsd, intgroup = "group")
plotPCA(vsd, intgroup = "agent")
plotPCA(vsd, intgroup = "time")
```



Figuras 7: Gráficos de Análisis de Componentes Principales (PCA) coloreados según la variable de interés. Cada gráfico representa la distribución de las muestras en un espacio de menor dimensión definido por los componentes principales. La coloración de las muestras corresponde a las variables "patient", "group", "agent" y "time", respectivamente. Estos gráficos permiten visualizar la estructura y los patrones de agrupamiento presentes en los datos de expresión génica después de aplicar la transformación estabilizadora de la varianza (VST).

Los resultados del análisis de componentes principales (PCA) revelan una estructura en los datos de expresión génica en la que las muestras tienden a agruparse principalmente según la similitud entre los pacientes. Contrariamente, no se observa una separación clara basada en el tratamiento o el tiempo de exposición. La presencia de un outlier, correspondiente al paciente 4 con tratamiento control y tiempo 24 horas, sugiere la existencia de una muestra atípica que puede tener un impacto significativo en el análisis y debe ser considerada con precaución en interpretaciones posteriores

5. Matriz de Distancias

A continuación, llevamos a cabo la matriz de distancias que es una representación numérica de la similitud o diferencia entre pares de objetos en un conjunto de datos. En el contexto del análisis de datos de expresión génica, la matriz de distancias se utiliza para medir cuán similares o diferentes son las muestras entre sí en función de sus perfiles de expresión génica. Esto es fundamental para identificar patrones de agrupamiento o relaciones entre las muestras, lo que puede proporcionar información sobre la biología subyacente de las muestras o las condiciones experimentales.

```
sampleDistMatrix_outlier <- as.matrix( sampleDists_outlier )
rownames(sampleDistMatrix_outlier) <- paste( vsd_outlier$patient, vsd_outlier$group, sep =
" - " )
colnames(sampleDistMatrix_outlier) <- NULL
colors <- colorRampPalette( c("#E91E63", "#EC407A", "#F06292", "#F48FB1", "#FCE4EC"))(255)
pheatmap(sampleDistMatrix_outlier,
         clustering_distance_rows = sampleDists_outlier,
         clustering_distance_cols = sampleDists_outlier,
         col = colors)
```

El código se divide en varios puntos:

1. Creación de la matriz de distancias:
    a. sampleDists es una matriz que contiene las distancias entre las muestras, calculadas previamente.
    b. as.matrix(sampleDists) convierte esta matriz de distancias en una matriz numérica estándar para poder ser visualizada como un heatmap.
2. Asignación de nombres a las filas de la matriz de distancias:
    a. Se utilizan las etiquetas de los pacientes y los grupos (obtenidos del objeto vsd) para asignar nombres a las filas de la matriz de distancias.
    b. paste(vsd$patient, vsd$group, sep = " - ") combina los nombres de los pacientes y los grupos separados por un guion.
3. Asignación de colores para el heatmap:
    a. brewer.pal(9, "Greens") devuelve una paleta de colores de verde del paquete RColorBrewer.
    b. colorRampPalette(rev(...)) invierte la paleta de colores para que los colores más oscuros representen distancias más pequeñas (mayor similitud).
    c. (255) especifica el número de colores en la paleta resultante.
4. Generación del heatmap:
    a. pheatmap es una función que crea un heatmap a partir de una matriz de datos.
    b. sampleDistMatrix se pasa como argumento para especificar la matriz de datos que se va a visualizar.
    c. clustering_distance_rows y clustering_distance_cols se establecen en sampleDists para realizar clustering jerárquico en filas y columnas basado en las distancias entre las muestras.

El gráfico obtenido fue el siguiente:

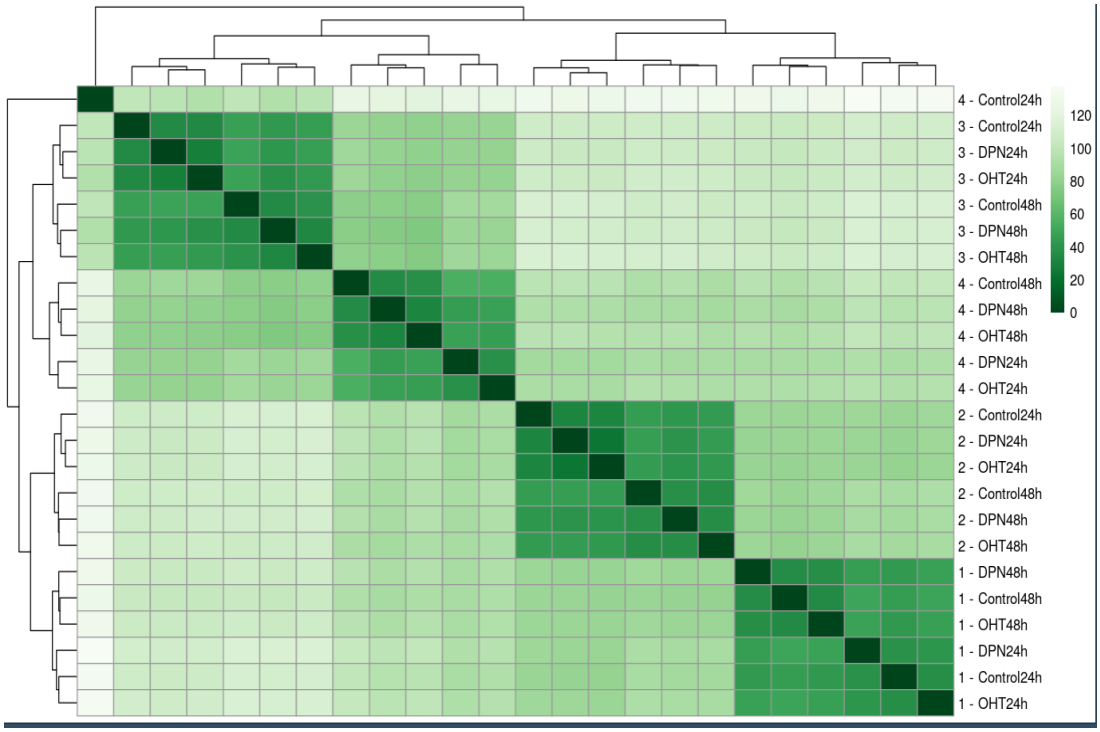


Figura 8: Esta matriz nos permite observar una coherente agrupación de los pacientes entre sí. Además, dentro de cada paciente, los distintos tiempos de muestreo parecen tender a formar grupos similares, como ocurre con las muestras tomadas a las 24 horas en comparación con las de 48 horas.En cuanto a las relaciones entre pacientes, se evidencia que los pacientes 1 y 2 muestran una mayor afinidad entre sí, al igual que sucede con los pacientes 3 y 4.Al analizar posibles valores atípicos, confirmamos su presencia utilizando este enfoque. Notamos que el valor atípico identificado se encuentra más cercano a los datos del paciente 3 que a los de su propio grupo, lo cual coincide con las observaciones realizadas en el análisis de componentes principales (PCA).

ii) Análisis de Expresión Diferencial

1. MANEJO DEL OUTLIER:

En esta sección del informe, nos adentramos en el análisis de expresión diferencial, una fase crucial en el estudio de datos de transcriptómica. Después de realizar el preprocesamiento de los datos y explorar la estructura subyacente utilizando métodos como el Análisis de Componentes Principales (PCA) y la generación de mapas de calor de las distancias entre las muestras, estamos listos para examinar cómo varía la expresión génica entre las condiciones experimentales.

El análisis de expresión diferencial nos permite identificar aquellos genes cuya expresión está significativamente alterada en diferentes tratamientos o condiciones. En nuestro estudio, este análisis nos proporcionará información sobre qué genes están asociados con las respuestas específicas a los tratamientos administrados y los tiempos de muestreo considerados.

Utilizando la metodología proporcionada por paquetes como DESeq2, exploraremos las diferencias en la expresión génica entre los grupos de interés, evaluando la significancia estadística de estas diferencias y generando listas de genes diferencialmente expresados. Estos resultados serán fundamentales para comprender los mecanismos subyacentes a los efectos observados en nuestro experimento y pueden proporcionar pistas importantes para futuras investigaciones y aplicaciones clínicas.

En esta sección, describiremos el procedimiento utilizado para analizar expresión diferencial, los resultados obtenidos y su interpretación en el contexto de nuestro estudio.

Primero de todo se llevó a cabo la eliminación del outliner para evitar afectar la calidad de los resultados obtenidos y hacer que el análisis resultante fuese más representativo.

```
patient_outlier <- which(experiment_data$patient == 4)
agent_outlier <- which(experiment_data$agent == "Control")
time_outlier <- which(experiment_data$time == "24h")
patient_agent <- intersect(patient_outlier, agent_outlier)
patient_time <- intersect(patient_outlier, time_outlier)
outlier <- intersect(patient_agent, patient_time)
experiment_data_clean <- experiment_data[-outlier,]
raw_data_clean <- raw_data[ ,-outlier]
```

Cuando ya hemos eliminado el outlier, deberemos crear el nuevo DESeqDataSet con las matrices de cuentas y metadatos sin él. Es decir, deberemos repetir todo el procedimiento anterior icnluyendo el código para eliminar los genes con menos de 10 lecturas, la transformación estabilizadora de la varianza y la normalización de los datos, hasta llegar a obtener las gráficas de la transformación de la vairnza, el PCA sin el outiler y el mapa de calor.

Los códigos realizados fueron:

```
## Creación nuevo DESeqDataSet
dds_outlier <- DESeqDataSetFromMatrix(countData = raw_data_clean,
                                      colData = experiment_data_clean,
                                      design = ~ patient + group)
dds_outlier
keep <- rowSums(counts(dds_outlier)) >= 10
dds2_outlier <- dds_outlier[keep, ]
dim(dds_outlier)
dim(dds2_outlier)
## VST
vsd_outlier <- vst(dds2_outlier, blind = TRUE)
## Visualización gráfica de VST
### Normalización de los datos de expresión
```

```
normal_data_outlier <- normTransform(dds2_outlier)
### Comparación efectos de transformación de la varianza
### frente a las muestras normalizadas
normal_graph_outlier <- meanSdPlot(assay(normal_data_outlier))
vsd_graph_outlier <- meanSdPlot(assay(vsd_outlier))
### Análisis de componentes principales (PCA)
plotPCA(vsd_outlier, intgroup = "patient")
plotPCA(vsd_outlier, intgroup = "group")
plotPCA(vsd_outlier, intgroup = "agent")
plotPCA(vsd_outlier, intgroup = "time")
### Matriz de distancias
#### Cálculo de las distancias entre muestras
sampleDists_outlier <- dist(t(assay(vsd_outlier)))
#### Creación de matriz de distancias
sampleDistMatrix_outlier <- as.matrix( sampleDists_outlier )
#### Asignación de nombres a las filas de la matriz de distancias
rownames(sampleDistMatrix_outlier) <- paste( vsd_outlier$patient, vsd_outlier$group, sep =
" - " )
colnames(sampleDistMatrix_outlier) <- NULL
#### Asignación de colores para el heatmap
colors <- colorRampPalette( rev(brewer.pal(9, "Greens")) )(255)
#### Generación del heatmap
pheatmap(sampleDistMatrix_outlier,
         clustering_distance_rows = sampleDists_outlier,
         clustering_distance_cols = sampleDists_outlier,
         col = colors)
```

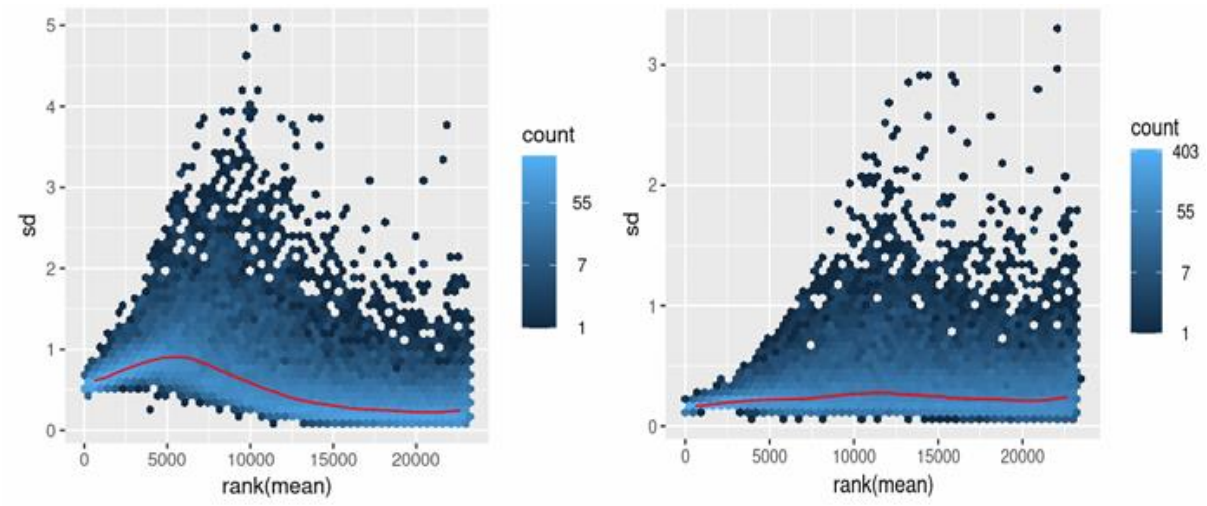


Figura 9. Comparación de los efectos causados por la transformación de la varianza.

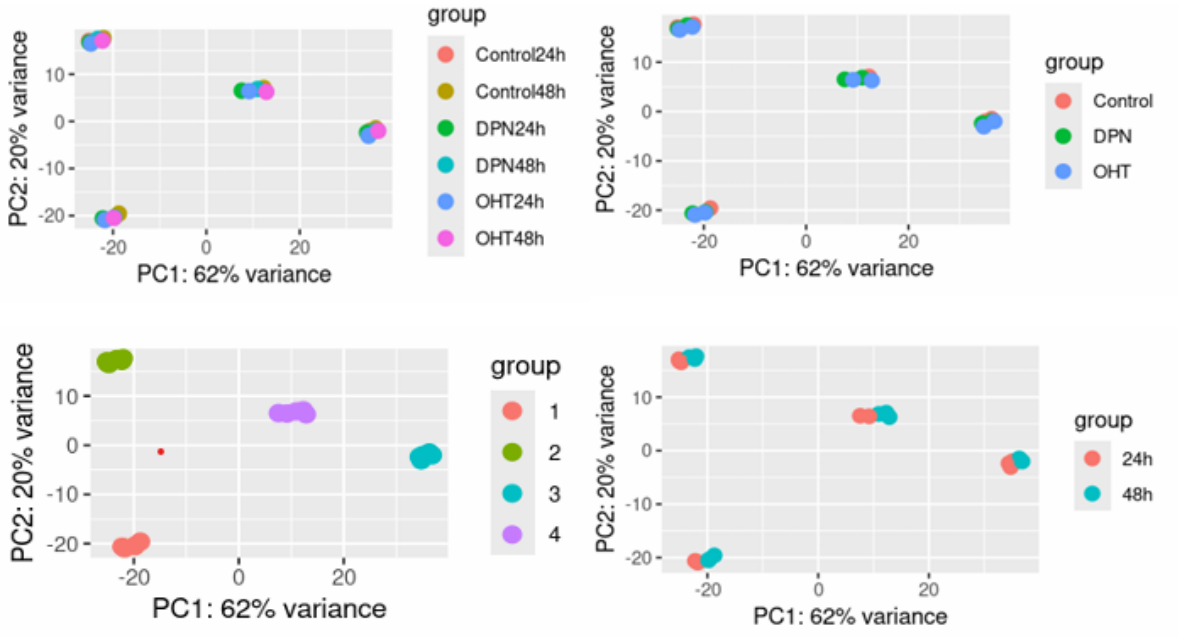


Figura 10. PCA de las muestras sin el outliner

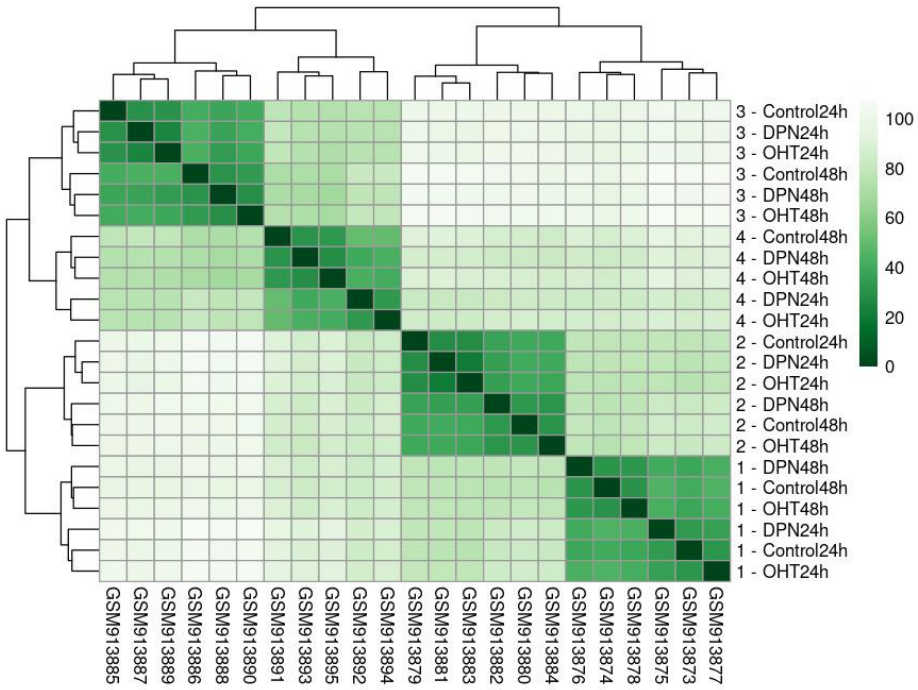


Figura 11: Mapa de calor de la matriz de distancias.

2. ANALISIS DE LA EXPRESIÓN DIFERENCIAL

Para llevar a cabo el análisis de expresión diferencial, se utiliza la función DESeq en R. Esta función realiza varias tareas esenciales, incluyendo la estimación de los size factors, que son utilizados para normalizar los recuentos de expresión génica, la estimación de la dispersión de los datos para evaluar la variabilidad entre las réplicas biológicas, y el ajuste de un modelo estadístico de distribución

binomial negativa. Además, se emplea el estadístico de Wald para calcular los valores de p y las puntuaciones de expresión diferencial.

Una vez completado el análisis, se generan dos gráficos para visualizar los resultados. El primero muestra la dispersión de los datos utilizando el método de la máxima probabilidad, lo que permite evaluar la variabilidad entre las muestras. El segundo gráfico, conocido como MA-plot, representa el cambio de plegamiento de cada gen en escala logarítmica en relación con la media de las cuentas normalizadas. Los genes con expresión diferencial se resaltan en colores distintos para facilitar su identificación.

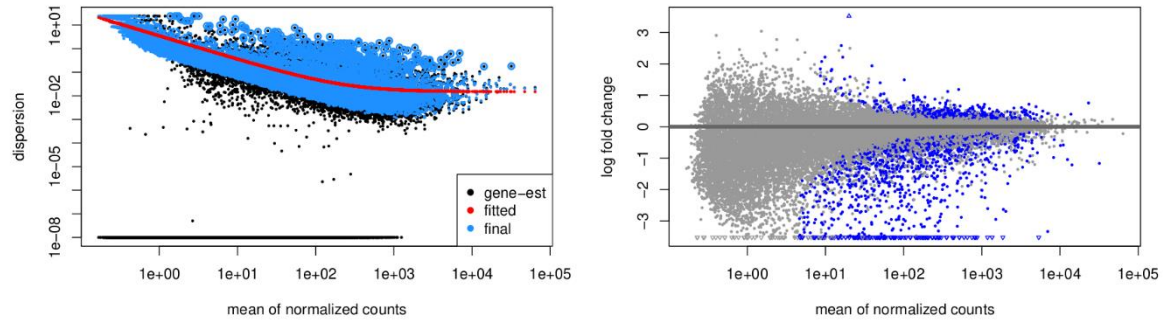


Figura 12. Estimación de la dispersión y resultados de expresión diferencial

Posteriormente se estudió el efecto que tiene el tratamiento con DPN frente al control en 24h sin aplicar un umbral.
Obteniéndose los siguientes resultados:

```
out of 23233 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 3, 0.013%
LFC < 0 (down)     : 2, 0.0086%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Posteriormente se repitió el análisis aplicandose un umbral (LFC=1) obtiendose en este caso los siguientes resultados:

```
out of 23233 with nonzero total read count
adjusted p-value < 0.05
LFC > 1.00 (up)    : 0, 0%
LFC < -1.00 (down) : 0, 0%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Se realizó el mismo procedimiento para el tratamiento con OHT para las muestras sin aplicar el umbral.

```
out of 23233 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 0, 0%
LFC < 0 (down)     : 0, 0%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

```
Y ahora aplicando el umbral:

out of 23233 with nonzero total read count
adjusted p-value < 0.05


out of 23233 with nonzero total read count
adjusted p-value < 0.05
LFC > 1.00 (up)    : 0, 0%
LFC < -1.00 (down) : 0, 0%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

En todos estos casos no se obtuvo ningún gen diferencialmente expresaado. Ocurriendo tanto sin el filtro del outilner como con él.

Para finalizar con el análisis se llevo a cabo una visualización de resultados, los denominados heatmap, se representaron los 30 genes con menor p-valor de los dos contrastes realizados, del DPN y el OHT.

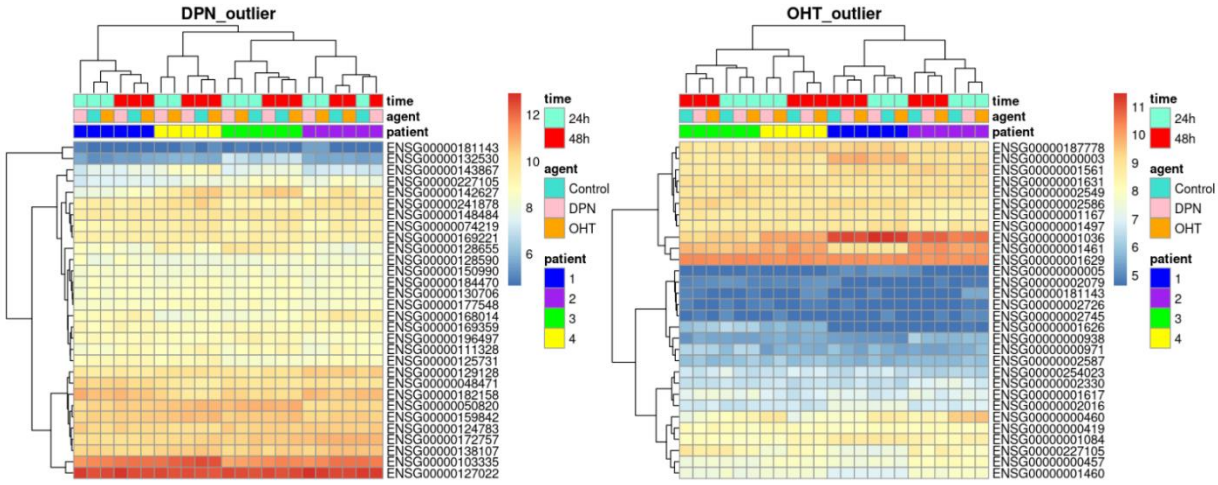


Figura 13. Heatmap de todos los genes diferencialmente expresados.

Para responder a la pregunta sobre qué genes están diferencialmente expresados entre las muestras tratadas con OHT y el control después de 24 horas, así como para las muestras tratadas con DPN después de 24 horas, se analizan los resultados obtenidos del análisis de expresión diferencial.

Para el tratamiento con OHT después de 24 horas, no se encontraron genes con una diferencia de expresión significativa en comparación con el control, ni siquiera al aplicar un umbral de log fold change (LFC) de ±1.

Para el tratamiento con DPN después de 24 horas, se identificaron 3 genes con una diferencia de expresión significativa con un LFC mayor que 0 (up-regulados) y 2 genes con una diferencia de expresión significativa con un LFC menor que 0 (down-regulados) sin aplicar un umbral. Sin embargo, al aplicar un umbral de LFC de ±1, no se encontraron genes diferencialmente expresados.

Es importante destacar que estos resultados están basados en un ajuste del valor p con un umbral de 0.05 y no se encontraron genes con ajuste significativo cuando se aplicó un umbral de LFC de ±1

iii) ANALISIS CON GSEA

En esta pregunta, nos proponemos explorar si el tratamiento con diarilpropionitrilo (DPN) tiene algún efecto en las muestras tratadas durante las primeras 24 horas de exposición. Para ello, nuestro colaborador ha comparado las muestras tratadas con DPN durante 48 horas con las muestras control. Sin embargo, surge la interrogante sobre si existen cambios significativos en la expresión génica en las primeras 24 horas de tratamiento con DPN.

Para abordar esta cuestión, llevaremos a cabo un análisis de Enriquecimiento de Genes (GSEA) utilizando un conjunto de genes previamente seleccionados para representar los cambios en la expresión génica inducidos por el tratamiento con DPN. A partir de los resultados de este análisis, podremos determinar si el tratamiento con DPN ejerce algún efecto sobre la expresión génica en las primeras 24 horas, y si estos cambios están asociados con algún pathway biológico específico. Mediante la comparación con las muestras control, evaluaremos la significancia de los cambios observados y extraeremos conclusiones sobre el efecto del DPN en las primeras etapas del tratamiento.

Este procedimiento se ha llevado a cabo con la misma versión de Rstudio que teniamos anteriormente y hemos hecho uso de tres paquetes:

1. DESeq2 (Versión 1.42.1):DESeq2 es un paquete de Bioconductor en R que se utiliza para el análisis de expresión génica diferencial en datos de secuenciación de ARN (RNA-seq). Permite identificar genes diferencialmente expresados entre diferentes condiciones experimentales, teniendo en cuenta la variabilidad biológica y técnica en los datos de RNA-seq. .

2. tidyverse (Versión 2.0.0): tidyverse es un conjunto de paquetes de R diseñados para realizar análisis de datos de una manera intuitiva y eficiente. Incluye una serie de paquetes que facilitan la manipulación, visualización y modelado de datos.

3. VennDiagram (Versión 1.7.3): VennDiagram es un paquete de R utilizado para crear diagramas de Venn y diagramas de Euler. Estos diagramas son útiles para visualizar la intersección y la unión de conjuntos de datos.

A continuación, vamos a llevar a cabo la creación del archivo. rnk que es necesario para realizar el análisis de GSEA (Enriquecimiento de Conjuntos de Genes) pre-rankeado. Este archivo contiene una lista de genes ordenados según su relevancia en términos de expresión diferencial entre condiciones experimentales y una puntuación de clasificación asociada, que generalmente se basa en alguna medida de cambio en la expresión génica, como el logFold Change (LFC).

El propósito principal de este archivo es proporcionar una entrada adecuada para el análisis de GSEA, donde los genes se clasifican según su asociación con la condición experimental de interés. Al ordenar los genes en función de su importancia relativa en la expresión diferencial, GSEA puede identificar conjuntos de genes que se enriquecen significativamente en una condición experimental en comparación con otra.

Una vez realizado se obtuvo el siguiente resultado:

```
out of 24416 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 13, 0.053%
LFC < 0 (down)     : 74, 0.3%
outliers [1]       : 0, 0%
low counts [2]     : 10888, 45%
(mean count < 28)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Cabe destacar que hay que reducir los datos para mejorar la precisión de las estimaciones de los parametros en el análisis. Para ello hicimos uso de “lfcShrink”. Se observa lo siguiente una vez que lo representamos:

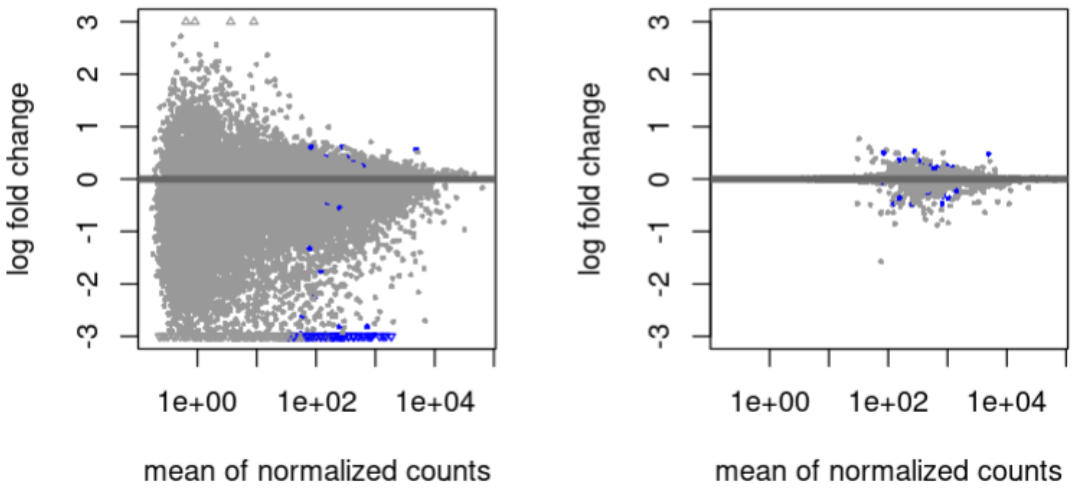


Figura 14: Se observan los datos reducidos a la derecha y sin reducir a la izquierda.

Para llevar a cabo el análisis de GSEA, vamos a hacer uso de la aplicación de escritorio GSEA, aqui cargamos los archivos necesarios haciendo uso del comando.

```
bash gsea-cli.sh GSEAPreranked -gmx
ftp.broadinstitute.org://pub/gsea/msigdb/human/gene_sets/h.all.v2023.2.Hs.symbols.gmt,"ruta
_del_archivo_gmt" Collapse -mode Abs_max_of_probes -rnd_seed 1000 -norm meandiv -collapse
149 -nperm 1000 -rnk "ruta_del_archivo_rnk" -rpt_label my_analysis -scoring_scheme weighted
-chip
ftp.broadinstitute.org://pub/gsea/msigdb/human/annotations/Human_Ensembl_Gene_ID_MSigDB.v20
23.2.Hs.chip -create_svgs -include_only_symbols true -set_max 500 -make_sets true -set_min
15 -plot_top_x 20 -out "ruta_de_salida" -zip_report "ruta_del_archivo_zip"
```

La explicación del comando es:

bash gsea-cli.sh: Inicia la ejecución del script gsea-cli.sh utilizando el intérprete de comandos bash.

GSEAPreranked: Especifica el tipo de análisis a realizar, en este caso, se está utilizando GSEA Preranked, que es una herramienta para el análisis de enriquecimiento de genes.

-gmx: Especifica la lista de conjuntos de genes predefinidos a utilizar en el análisis.

ftp.broadinstitute.org://pub/gsea/msigdb/human/gene_sets/h.all.v2023.2.Hs.symbols.gmt: Es la URL del archivo de conjuntos de genes predefinidos en formato GMT.

"ruta_del_archivo_gmt": Aquí se debe reemplazar con la ruta del archivo GMT que tienes en tu sistema.

Collapse: Especifica el método de colapso a utilizar en el análisis.

-mode Abs_max_of_probes: Define el modo de colapso como el máximo absoluto de sondas.

-rnd_seed 1000: Establece una semilla aleatoria para la reproducibilidad de los resultados.

-norm meandiv: Define el método de normalización como media/división.

-collapse 149: Define el tamaño máximo de genes para cada conjunto de genes enriquecidos.

-nperm 1000: Especifica el número de permutaciones a realizar para el cálculo de los valores p.

-rnk "ruta_del_archivo_rnk": Aquí se debe reemplazar con la ruta del archivo de clasificación (rnk) que tienes en tu sistema.

-rpt_label my_analysis: Etiqueta para la salida del informe de análisis.

-scoring_scheme weighted: Especifica el esquema de puntuación a utilizar en el análisis.

-chip
ftp.broadinstitute.org://pub/gsea/msigdb/human/annotations/Human_Ensembl_Gene_ID_MSigDB.v2023.2.Hs.chip: Define el archivo de anotación de chips.

-create_svgs: Indica que se deben crear archivos SVG para visualizar los resultados.

-include_only_symbols true: Especifica que solo se deben incluir símbolos en el análisis.

-set_max 500: Define el tamaño máximo de conjuntos de genes enriquecidos.

-make_sets true: Indica que se deben generar conjuntos de genes enriquecidos.

-set_min 15: Define el tamaño mínimo de conjuntos de genes enriquecidos.

-plot_top_x 20: Indica que se deben trazar los 20 mejores conjuntos de genes enriquecidos.

-out "ruta_de_salida": Aquí se debe reemplazar con la ruta donde deseas guardar los resultados de tu análisis.

-zip_report "ruta_del_archivo_zip": Aquí se debe reemplazar con la ruta donde deseas guardar el archivo comprimido que contiene los resultados.

El informe de GSEA proporciona una visión detallada del enriquecimiento de conjuntos de genes en diferentes fenotipos, lo que permite una mejor comprensión. La salida del informe de GSEA proporciona información detallada sobre el enriquecimiento de conjuntos de genes en dos fenotipos diferentes, denominados na_pos y na_neg. En el fenotipo na_pos, se observa que 19 de 50 conjuntos de genes están arriba regulados, de los cuales 6 son significativamente enriquecidos con un FDR (Tasa de Descubrimiento Falso) menor al 25%. Además, 1 conjunto de genes muestra una significancia más alta, con un valor p nominal inferior al 1%. Por otro lado, en el fenotipo na_neg, 31 de 50 conjuntos de genes están arriba regulados, con 19 de ellos significativamente enriquecidos a un nivel de FDR menor al 25%. También se observa una cantidad considerable de conjuntos de genes significativamente enriquecidos a niveles de significancia más estrictos (p-value < 1% y p-value < 5%).

El conjunto de datos utilizado contiene inicialmente 24,416 características nativas, que luego se reducen a 21,106 genes después de colapsar las características en símbolos de genes. Se aplicaron filtros de tamaño al conjunto de genes, lo que resultó en la eliminación de 2 de los 52 conjuntos de genes originales, dejando un total de 50 conjuntos de genes para el análisis.

Posteriormente hemos hecho un análisis de los genes más expresadas en las muestras tratadas y control. Para conocer los más expresados en las muestras tratadas nos debemos introducir en el archivo gsea_report_for_na_pos_1711477851437.tsv, en donde encontraremos la siguiente tabla:

NAME	**NES**	**FDR q-val**	**LEADING EDGE**
HALLMARK_OXIDATIVE_PHOSPHORYLATION	1.5675	0.0774	tags=60%, list=23%, signal=77%
HALLMARK_PROTEIN_SECRETION	1.4171	0.2087	tags=44%, list=14%, signal=50%
HALLMARK_INTERFERON_ALPHA_RESPONSE	1.3853	0.1921	tags=26%, list=14%, signal=30%
HALLMARK_HEME_METABOLISM	1.3853	0.1441	tags=36%, list=18%, signal=44%
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	1.3427	0.1771	tags=41%, list=17%, signal=50%
HALLMARK_MTORC1_SIGNALING	1.2869	0.2442	tags=25%, list=14%, signal=29%
HALLMARK_BILE_ACID_METABOLISM	1.2018	0.4030	tags=27%, list=15%, signal=32%
HALLMARK_GLYCOLYSIS	1.2008	0.3548	tags=23%, list=12%, signal=26%
HALLMARK_PEROXISOME	1.1390	0.4721	tags=41%, list=17%, signal=49%
HALLMARK_INTERFERON_GAMMA_RESPONSE	1.1177	0.4810	tags=16%, list=14%, signal=18%
HALLMARK_ADIPOGENESIS	1.0272	0.6996	tags=31%, list=15%, signal=36%
HALLMARK_MYC_TARGETS_V1	0.9847	0.7726	tags=52%, list=25%, signal=69%

HALLMARK_PI3K_AKT_MTOR_SIGNALING	0.9760	0.7394	tags=27%, list=15%, signal=32%
HALLMARK_ANDROGEN_RESPONSE	0.9643	0.7196	tags=22%, list=16%, signal=26%
HALLMARK_PANCREAS_BETA_CELLS	0.9571	0.6913	tags=12%, list=6%, signal=12%
HALLMARK_E2F_TARGETS	0.8621	0.8903	tags=17%, list=14%, signal=19%
HALLMARK_G2M_CHECKPOINT	0.8453	0.8749	tags=17%, list=13%, signal=20%
HALLMARK_WNT_BETA_CATENIN_SIGNALING	0.7122	1.0	tags=36%, list=23%, signal=46%
HALLMARK_IL6_JAK_STAT3_SIGNALING	0.7000	0.9718	tags=16%, list=15%, signal=18%

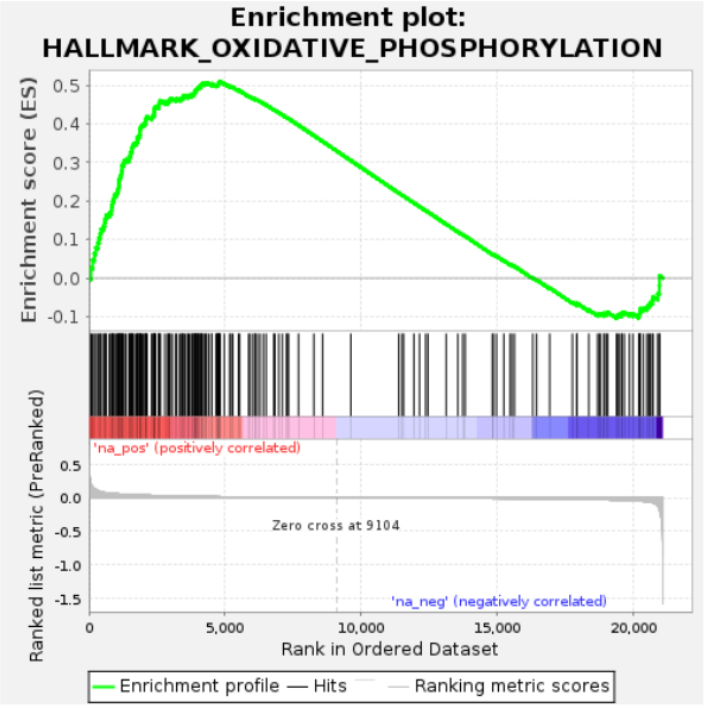
Esta tabla proporciona información sobre el enriquecimiento de los conjuntos de genes más destacados, destacando el valor NES (Normalized Enrichment Score), FDR q-val (False Discovery Rate), y detalles sobre el "Leading Edge", que indica qué porcentaje del conjunto de genes está contribuyendo al enriquecimiento.

Basándonos en la tabla proporcionada, podemos observar que el conjunto de genes "HALLMARK_OXIDATIVE_PHOSPHORYLATION" tiene el NES más alto (1.5675384), lo que indica que está altamente enriquecido en el fenotipo analizado. Esto sugiere que los genes incluidos en este conjunto se asocian con la respuesta biológica representada por el fenotipo.

Ahora, analicemos la última columna, que describe el "Leading Edge" o "Borde Principal". En este contexto, el "Borde Principal" se refiere a la parte del conjunto de genes que más contribuye al enriquecimiento observado. La información proporcionada en esta columna indica qué porcentaje del conjunto de genes está contribuyendo al enriquecimiento en términos de diferentes aspectos, como la cantidad de genes que se superponen con los genes de interés, la cantidad de genes presentes en la lista ordenada de genes, y la señal observada en la muestra analizada.

Por ejemplo, en el caso de "HALLMARK_OXIDATIVE_PHOSPHORYLATION", la entrada en la última columna dice "tags=60%, list=23%, signal=77%". Esto significa que el 60% de los genes en el conjunto están etiquetados como importantes, el 23% de los genes en la lista ordenada de genes contribuyen al enriquecimiento, y la señal observada en la muestra refleja el 77% del conjunto de genes. Esto sugiere que una proporción significativa de los genes en este conjunto está contribuyendo al enriquecimiento observado y, por lo tanto, pueden desempeñar un papel importante en la respuesta biológica asociada con el fenotipo analizado.

Una gráfica del enriquecimiento para este gen es esta:



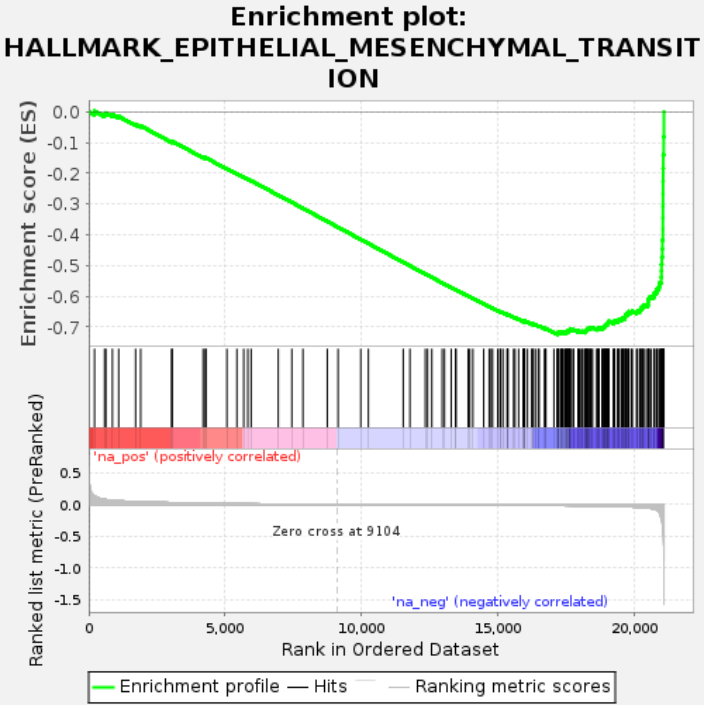
De la misma manera hicimos con las muestras sin tratar. Así que obtuvimos esta tabla.

NAME	NES	FDR q-val	LEADING EDGE
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	-2.3652	0.0	tags=65%, list=19%, signal=79%
HALLMARK_TNFA_SIGNALING_VIA_NFKB	-1.9777	0.0094	tags=45%, list=18%, signal=54%
HALLMARK_COAGULATION	-1.8421	0.0206	tags=33%, list=16%, signal=39%
HALLMARK_INFLAMMATORY_RESPONSE	-1.7609	0.0263	tags=33%, list=16%, signal=38%
HALLMARK_APICAL_SURFACE	-1.6928	0.0347	tags=19%, list=9%, signal=21%
HALLMARK_APICAL_JUNCTION	-1.6678	0.033	tags=33%, list=14%, signal=38%
HALLMARK_COMPLEMENT	-1.59	0.0511	tags=31%, list=16%, signal=36%
HALLMARK_TGF_BETA_SIGNALING	-1.5888	0.0447	tags=37%, list=9%, signal=41%
HALLMARK_MYOGENESIS	-1.5418	0.056	tags=20%, list=9%, signal=21%
HALLMARK_ALLOGRAFT_REJECTION	-1.5415	0.0504	tags=23%, list=14%, signal=26%
HALLMARK_KRAS_SIGNALING_UP	-1.5106	0.0581	tags=41%, list=20%, signal=52%
HALLMARK_UV_RESPONSE_DN	-1.4853	0.0632	tags=42%, list=15%, signal=49%

HALLMARK_HYPOXIA	-1.4424	0.0784	tags=40%, list=15%, signal=47%
HALLMARK_NOTCH_SIGNALING	-1.4152	0.0902	tags=41%, list=15%, signal=48%
HALLMARK_UV_RESPONSE_UP	-1.4061	0.0887	tags=30%, list=16%, signal=36%
HALLMARK_P53_PATHWAY	-1.3841	0.0971	tags=27%, list=13%, signal=31%
HALLMARK_MITOTIC_SPINDLE	-1.3779	0.096	tags=26%, list=9%, signal=28%
HALLMARK_ESTROGEN_RESPONSE_EARLY	-1.361	0.0998	tags=36%, list=15%, signal=43%
HALLMARK_XENOBIOTIC_METABOLISM	-1.3275	0.1175	tags=25%, list=16%, signal=30%
HALLMARK_APOPTOSIS	-1.1638	0.2902	tags=32%, list=14%, signal=37%
HALLMARK_ANGIOGENESIS	-1.1288	0.3336	tags=61%, list=18%, signal=74%
HALLMARK_CHOLESTEROL_HOMEOSTASIS	-1.0602	0.4495	tags=32%, list=12%, signal=36%
HALLMARK_IL2_STAT5_SIGNALING	-1.0453	0.4607	tags=40%, list=19%, signal=49%
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	-1.0102	0.5258	tags=27%, list=11%, signal=30%
HALLMARK_HEDGEHOG_SIGNALING	-0.9781	0.5869	tags=33%, list=12%, signal=38%
HALLMARK_ESTROGEN_RESPONSE_LATE	-0.9756	0.5707	tags=30%, list=15%, signal=35%
HALLMARK_FATTY_ACID_METABOLISM	-0.9191	0.6966	tags=19%, list=11%, signal=22%
HALLMARK_DNA_REPAIR	-0.9072	0.7043	tags=19%, list=10%, signal=21%
HALLMARK_MYC_TARGETS_V2	-0.8848	0.738	tags=22%, list=12%, signal=25%
HALLMARK_SPERMATOGENESIS	-0.8424	0.8164	tags=2%, list=1%, signal=2%
HALLMARK_KRAS_SIGNALING_DN	-0.6453	0.9872	tags=30%, list=

El gen más expresado en esta tabla es el HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION, ya que tiene el valor más bajo de NES (-2.3652), lo que indica que se expresa significativamente en la muestra analizada. El NES (Normalized Enrichment Score) es una medida utilizada en GSEA para cuantificar la importancia de un gen en una vía biológica dada. En este caso, un valor negativo de NES indica que el gen está suprimido o menos expresado en la muestra en comparación con otras muestras.

La gráfica obtenida en este caso es:



Este análisis revela que el conjunto de genes está enriquecido en los extremos inferiores de la tabla, lo que sugiere su supresión o baja regulación. Al examinar el archivo "HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION.html", se puede identificar una tabla que detalla los genes que conforman el "leading edge", marcados con un "sí" en la columna "core enrichment". Se nota que estos genes se encuentran predominantemente en la parte inferior de la tabla, lo que concuerda con su regulación a la baja.

iv) CONCLUSIONES

Tras analizar los datos preliminares, se puede concluir que las diferencias en la expresión génica no son atribuibles a las variaciones entre los distintos tratamientos, sino que parecen estar más relacionadas con las diferencias individuales entre los pacientes. Esto sugiere que la respuesta a los tratamientos puede ser altamente personalizada y dependiente de las características específicas de cada paciente. Además, el análisis realizado con DESeq2 indica que hay una cantidad mínima de genes que muestran una expresión diferencial significativa con ambos tratamientos a las 24 horas.

Los resultados obtenidos de los heatmap revelan que las agrupaciones de genes no se correlacionan con la similitud en los tratamientos o en los tiempos de tratamiento, sino más bien con similitudes entre los pacientes, como lo corroboran los análisis de PCA. La identificación de un número tan reducido de genes diferencialmente expresados podría deberse a la exclusión de un outlier. Para mejorar la precisión de los resultados, sería necesario ampliar la muestra para obtener múltiples datos

por condición y paciente, lo que permitiría una mejor comprensión de la variabilidad entre los individuos.

Adicionalmente, el análisis de GSEA revela que el tratamiento con DPN tiene un efecto significativo a lo largo del tiempo. Los genes asociados con el fenotipo "perturbed" están enriquecidos en la parte alta de la tabla, lo que indica una regulación al alza, mientras que los genes asociados con el fenotipo "unperturbed" están enriquecidos en la parte baja de la tabla, indicando una regulación a la baja. Esto sugiere que los cambios observados podrían ser más pronunciados después de 48 horas de tratamiento. Por lo tanto, se espera que el DPN produzca algún efecto en las primeras 24 horas, lo que justifica una mayor atención a los efectos a largo plazo del tratamiento.