Transcriptómica

Analisis de RNA-Seq

Tamara Said El Artah

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

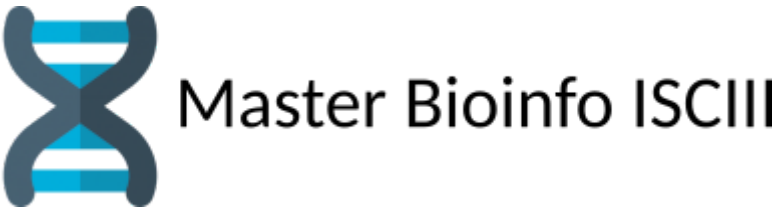


Tabla de contenidos

Introducción

El objetivo de este informe es demostrar el conocimiento adquirido sobre el análisis de RNA-seq. Este trabajo consta de dos secciones en las cuales se utilizarán datos de un experimento con 24 cultivos primarios de tumores paratiroideos negativos para el receptor alfa de estrógeno (ERα). Las muestras, procedentes de cuatro pacientes diferentes, han sido tratadas con dos fármacos diferentes: el diarilpropionitrilo (DPN) o el 4-hidroxitamoxifeno (OHT) durante 24 o 48 horas. El DPN es un agonista del ERα, mientras que el OHT es un inhibidor competitivo de los receptores de estrógeno.

La primera parte describe los pasos de control de calidad y fuentes de contaminación, trimming, alineación y cuantificación para obtener recuentos crudos y normalizados a partir de un subconjunto de archivos fastq. La segunda parte se basa en la matriz completa de recuentos crudos y se centra en realizar un control de calidad biológico, detectar genes diferencialmente expresados entre condiciones y las vías enriquecidas en cada una de ellas.

Metodología

Primera Parte

El database consta de 27 muestras emparejadas depositadas en SRA, sin embargo, para los propósitos de este trabajo, se han seleccionado únicamente dos muestras, SRR479052 y SRR479054, con secuencias de forward y reverse. La referencia genómica utilizada corresponde al cromosoma 21 humano (ensamblaje GRCh38), acompañada de un archivo GTF que contiene la anotación genética de los genes presentes en el cromosoma 21 (GRCh38.ensembl.109). Este enfoque garantiza la precisión y relevancia de los datos analizados, permitiendo una interpretación adecuada de los resultados obtenidos en las etapas posteriores del proyecto.

Para completar la primera parte del proyecto, se desarrolló un script de shell bash, se muestra en el Anexo 1. Este script contiene los pasos necesarios para el control de calidad, recorte, alineación y cuantificación de un subconjunto de archivos fastq.

Las herramientas empleadas en este proceso incluyen FASTQC para la evaluación de la calidad, HISAT2 para indexar y alinear las muestras con el cromosoma 21 de referencia, Samtools para generar estadísticas de alineación y HTSeq para cuantificar los recuentos de expresión.
FASTQC tiene como objetivo realizar controles de calidad integrales en datos de secuencia sin procesar, proporcionando información valiosa sobre problemas potenciales antes de realizar más análisis. Mientras, HISAT2 se destaca como una herramienta rápida y de alineación capaz de mapear lecturas de secuenciación de próxima generación a una población o a un genoma de referencia único. Samtools, por otro lado, comprende un conjunto de programas diseñados para la manipulación eficaz de datos de secuenciación de alto rendimiento. Finalmente, se emplea HTSeq, un paquete de Python, para el análisis detallado de datos de secuenciación de alto rendimiento, lo que contribuye significativamente a las etapas posteriores del proyecto. [1-4]

Quality Control: FASTQC

FASTQC tiene como objetivo realizar controles de calidad integrales en datos de secuencia sin procesar, proporcionando información valiosa sobre problemas potenciales antes de realizar más análisis.
La herramienta procesa las muestras en función de varias categorías:

1. Basic Statistics, que crea estadísticas de composición básicas para la lectura que se analiza.
2. Per Base Sequence Quality, que muestra la calidad de los pares de bases en función de la puntuación matemática y estadística y se representa en un gráfico. El gráfico está codificado por colores y dividido en 3 partes: verde, amarillo y rojo. El verde refleja puntuaciones de calidad altas y muy buena, el amarillo representa una calidad razonable y el rojo representa puntuaciones de calidad bajas, por lo tanto, muy mala calidad. Un cuadro amarillo rectangular muestra rangos estadísticos cuartiles e intercuartílicos. Y finalmente, la fina azul fina representa la cualidad media.
3. Per Base Sequence Quality Scores representan gráficamente los valores de baja calidad.
4. Per Base Sequence Content, traza las proporciones de las 4 bases de nucleótidos (A, C, G, T) y sus posiciones.
5. Per Sequence GC Content, representa gráficamente el contenido de GC y se compara con una distribución normal del contenido de GC.
6. Per Base N Content, representa cuando se sustituye N si no se proporciona una base nucleotídica.
7. Sequence Length Distribution, que representa la distribución de los tamaños de secuencia.

8. Duplicate Sequences, cuenta y representa el número de duplicaciones de las secuencias.
9. Overrepresented Sequences y 10. Adapter Content, enumera todas las secuencias que coinciden con la base de datos, ya que existe la posibilidad de que la secuencia que se utiliza para el análisis esté contaminada.

Cada módulo está representado por 3 símbolos principales: una marca verde, un signo de exclamación naranja y una x roja. Cada uno tiene su propio significado, la marca verde significa que el módulo pasó y no hubo problemas, un signo de exclamación naranja significa que hay algo desconocido en la lectura y se debe verificar, y una x roja significa un desconocimiento muy poco común, por lo tanto, incorrecto leer en el módulo.

Basic Statistics

En Table 1, el análisis realizado con FASTQC revela las Basic Statistics de los datos de secuenciación obtenidos de las muestras SRR479052 y SRR479054.

Table 1 Basic Statistics Comparison

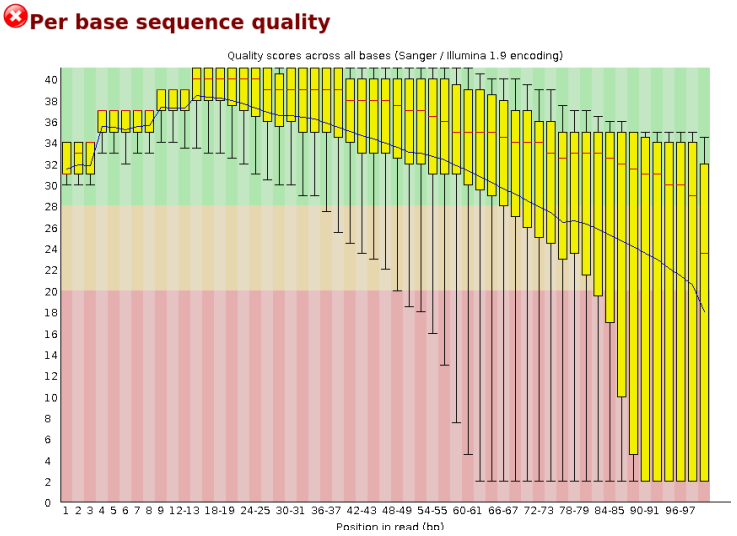
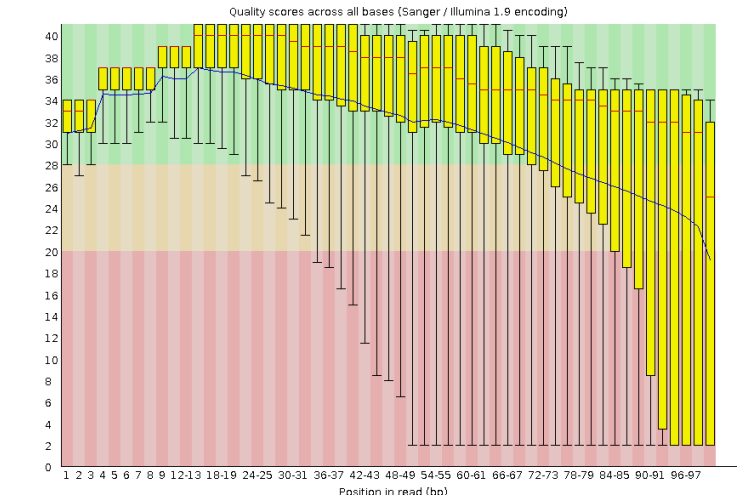
	SRR479052.chr21_1 y SRR479052.chr21_2	**SRR479054.chr21_1 y SRR479054.chr21_2**
Basic Statistics	**Basic Statistics** Measure / Value Filename: SRR479052.chr21_2.fastq File type: Conventional base calls Encoding: Sanger / Illumina 1.9 Total Sequences: 15340 Sequences flagged as poor quality: 0 Sequence length: 101 %GC: 52	**Basic Statistics** Measure / Value Filename: SRR479054.chr21_1.fastq File type: Conventional base calls Encoding: Sanger / Illumina 1.9 Total Sequences: 9746 Sequences flagged as poor quality: 0 Sequence length: 101 %GC: 51

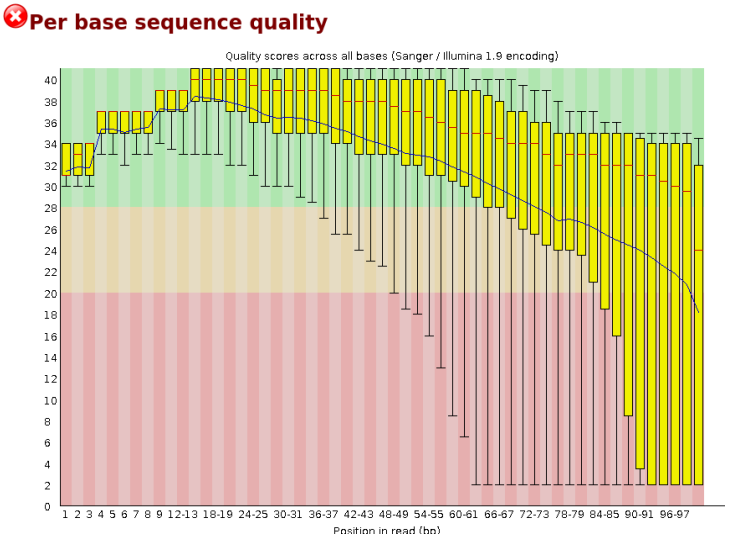
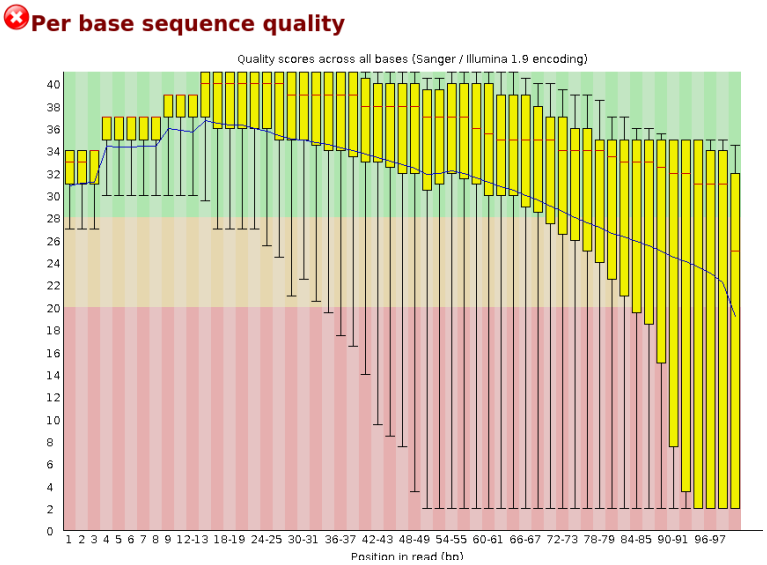
Para los archivos SRR479052.chr21_1.fastq y SRR479052.chr21_2.fastq, se observa que presentan llamadas de bases convencionales con un formato de codificación Sanger/Illumina 1.9. En total, se detectan 15,340 secuencias, sin ninguna marcada como de calidad deficiente. Cada secuencia tiene una longitud de 101 nucleótidos, con un contenido de GC del 52%. De manera similar, para los archivos SRR479054.chr21_1.fastq y SRR479054.chr21_2.fastq, se observan las mismas características, con un total de 9,746 secuencias, todas superando el umbral de calidad. La longitud de la secuencia y el contenido de GC se mantienen consistentes en 101 nucleótidos y 51%, respectivamente. Estas estadísticas proporcionan información valiosa sobre la calidad general y la composición de los datos de secuenciación, asegurando la fiabilidad de los análisis e interpretaciones posteriores.

Per Base Sequence Quality

En Table 2, el análisis realizado con FASTQC revela las Per Base Sequence Quality de las muestras SRR479052 y SRR479054.

Table 2 Per Base Sequence Quality Comparison

	Per Base Sequence Quality
SRR479052.chr21_1	
SRR479052.chr21_2	

SRR479054.chr21_1	
SRR479054.chr21_2	

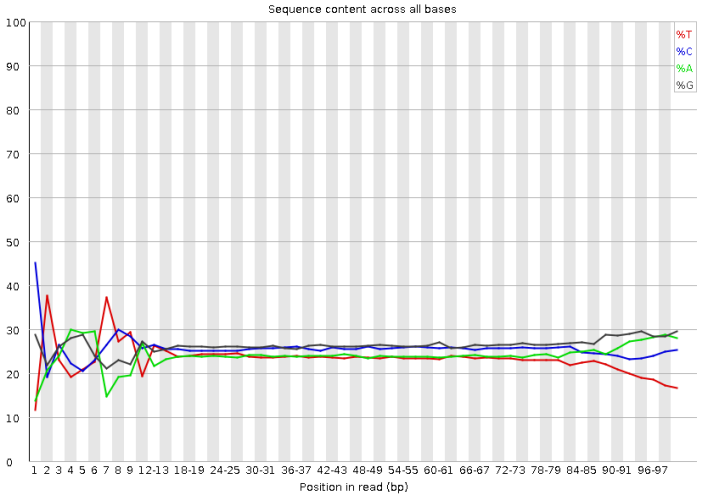
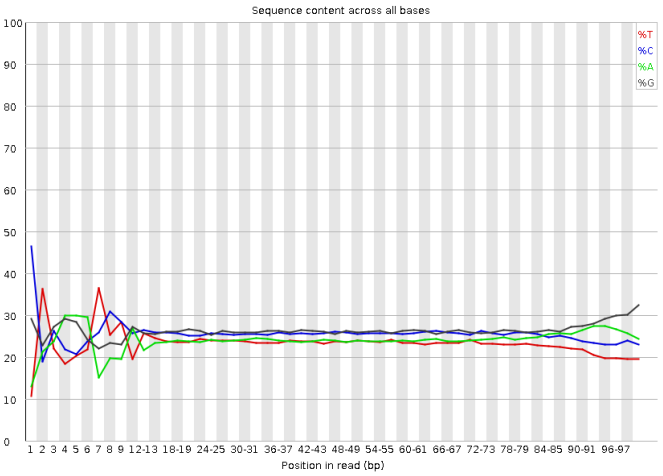
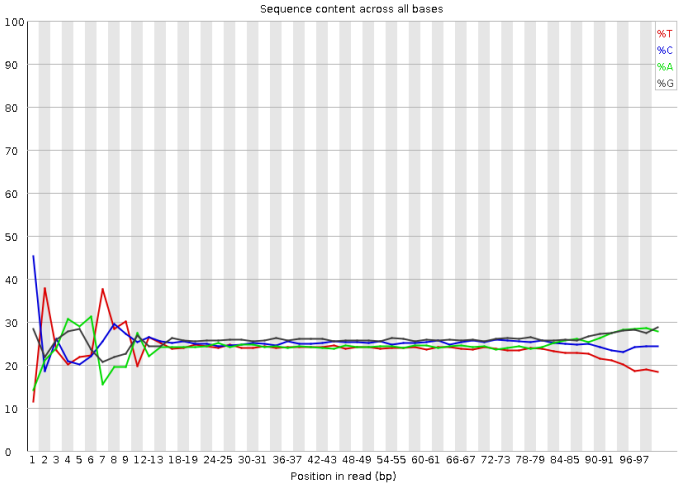
En el caso que nos ocupa, es evidente que la calidad de las secuencias ya no se considera buena a partir de aproximadamente 65-70 pb. Sin embargo, en la región amarilla de aproximadamente 72-84 pb, se observa una calidad deficiente que requiere ser corregida o mejorada para evitar que afecte los resultados debido a la baja calidad.

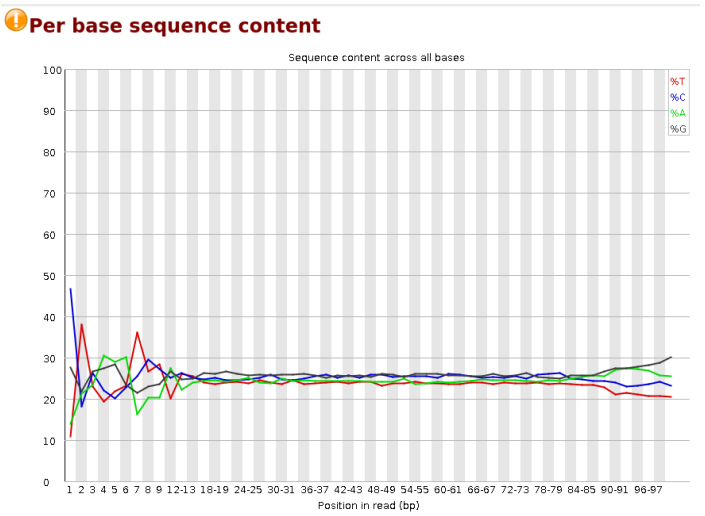
Per Base Sequence Content

En Table 3, el análisis realizado con FASTQC revela las Per Base Sequence Content de las muestras SRR479052 y SRR479054.

Table 3 Per Base Sequence Content

	Per Base Sequence Content

SRR479052.chr21_1	**Per base sequence content** 
SRR479052.chr21_2	**Per base sequence content** 
SRR479054.chr21_1	**Per base sequence content** 

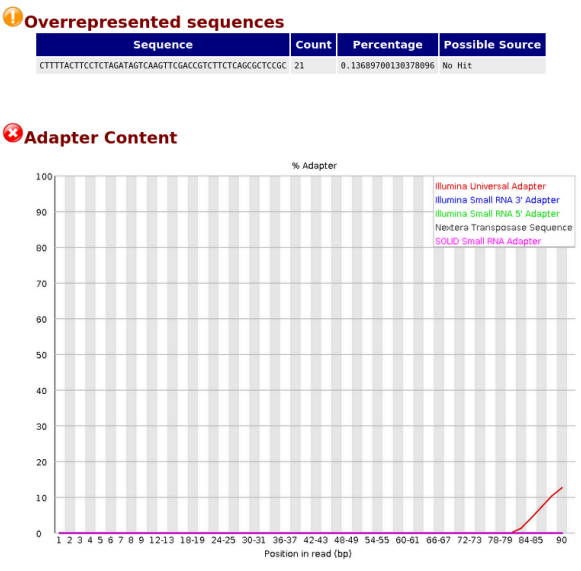
SRR479054.chr21_2	

Este análisis puede revelar problemas durante la preparación de la muestra o incluso durante el proceso de secuenciación, como la presencia de adaptadores o contaminantes. La evaluación integral de esta métrica proporciona una visión detallada de la calidad de las secuencias, lo que permite identificar y abordar cualquier anomalía que pueda surgir en el proceso de secuenciación.

Overrepresented Sequences y Adapter Content

En Table 4, el análisis realizado con FASTQC revela las Overrepresented Sequences y Adapter Content de las muestras SRR479052 y SRR479054.

Table 4 Overrepresented Sequences y Adapter Content

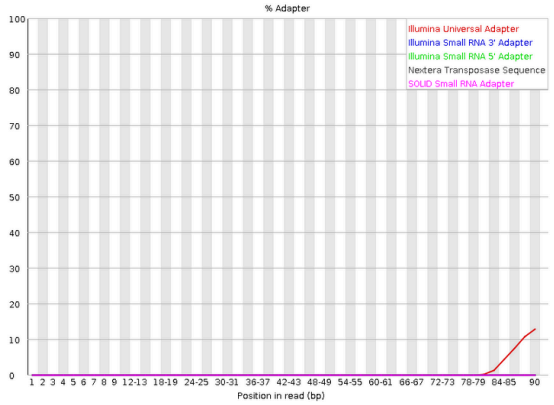
	Overrepresented Sequences y Adapter Content
SRR479052.chr21_1	

SRR479052.chr21_2

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTAACACGTGCGCGAGTCGGGGGCTCGCACGAAAGCCGCCGTGGCGCAAT	20	0.1303780964797914	No Hit

Adapter Content

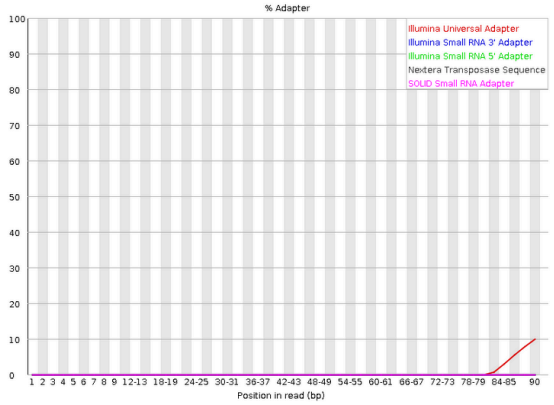


SRR479054.chr21_1

Overrepresented sequences

No overrepresented sequences

Adapter Content

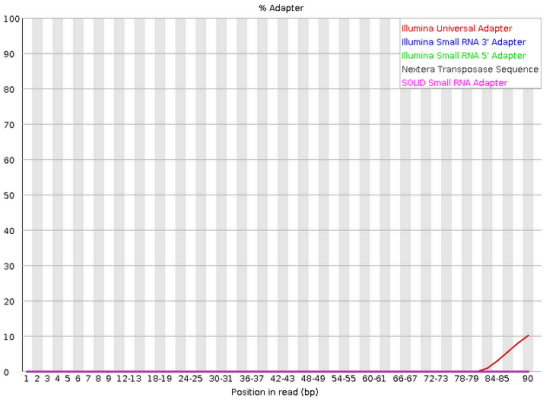


SRR479054.chr21_2

Overrepresented sequences

No overrepresented sequences

Adapter Content



Las secuencias sobrerrepresentadas, en ocasiones, pueden originarse a partir de adaptadores residuales o secuencias contaminantes. Como se puede apreciar en el análisis de la muestra SRR479052, se han detectado secuencias sobrerrepresentadas en ambas secuencias, lo que sugiere la presencia de fragmentos adicionales que pueden no estar relacionados con el material genético de interés. La identificación temprana de estas secuencias sobrerrepresentadas es fundamental para aplicar estrategias de eliminación o corrección que minimicen su impacto en la interpretación de los resultados obtenidos del análisis genómico.

Finalmente, se identificó la presencia de secuencias adaptadoras en todas las muestras analizadas, lo que sugiere la presencia de adaptadores universales de Illumina. Los adaptadores, juegan un papel fundamental durante la etapa de preparación de la muestra para la secuenciación, facilitando la unión de los fragmentos de interés al soporte sólido utilizado en el proceso. Sin embargo, es importante resaltar que estos adaptadores no forman parte del material genético que se pretende analizar. La detección de estos adaptadores en las secuencias secuenciadas es crucial para el posterior análisis de datos, ya que su presencia puede influir en la calidad y precisión de los resultados obtenidos. Por lo tanto, la identificación temprana y la eliminación adecuada de estos adaptadores son pasos críticos en el proceso de análisis de datos de secuenciación de próxima generación.

HISAT2, Samtools y HTSeq

Como se indicó anteriormente, HISAT2 es una herramienta de alineación capaz de mapear lecturas de secuenciación de próxima generación en una población o un único genoma de referencia; Samtools comprende un conjunto de programas diseñados para la manipulación eficiente de datos de secuenciación de alto rendimiento, y HTSeq, un paquete de Python, se emplea para el análisis detallado de datos de secuenciación de alto rendimiento, lo que contribuye significativamente a las etapas posteriores del proyecto. Los comandos y parámetros utilizados en el script son los siguientes:

1. Indexación y Alineación con HISAT2:
    - `hisat2-build input/Homo_sapiens.GRCh38.dna.chromosome.21.fa input/Homo_sap_index`: este comando indexa el archivo del genoma de referencia `Homo_sapiens.GRCh38.dna.chromosome.21.fa` para crear un índice llamado `Homo_sap_index ` para una alineación eficiente.
    - `hisat2 -x input/Homo_sap_index -1 input/SRR479052.chr21_1.fastq -2 input/SRR479052.chr21_2.fastq -S input/SRR479052.chr21_alignments.sam`: este comando realiza la alineación de lecturas de extremos emparejados de los archivos fastq `SRR479052.chr21_1.fastq` y `SRR479052.chr21_2.fastq` al genoma

de referencia indexado. Las lecturas alineadas se guardan en formato SAM como `SRR479052.chr21_alignments.sam`.

i. Resultados procesaron lecturas de secuenciación y proporcionaron estadísticas de alineación para dos muestras:
    1. Para `SRR479052`, de 15.340 lecturas emparejadas, el 53,87 % se alineó 0 veces, el 37,66 % se alineó exactamente una vez, con una tasa de alineación general del 78,61 %.
    2. Para `SRR479054`, de 9746 lecturas emparejadas, el 50,10 % se alineó 0 veces, el 39,79 % se alineó exactamente una vez, con una tasa de alineación general del 79,44 %.

ii. Estos resultados indican variaciones en la eficiencia de alineación entre las dos muestras, donde SRR479052 muestra una tasa de alineación general ligeramente menor en comparación con SRR479054. Además, la proporción de lecturas que se alinean concordantemente exactamente una vez es mayor en SRR479052, lo que sugiere una calidad de alineación potencialmente mejor en esta muestra. Una investigación más profunda sobre las discrepancias de alineación y los pares discordantes puede proporcionar información sobre los factores subyacentes que afectan el rendimiento de la alineación.

2. Estadísticas de alineación con SAMTools y cuantificación de expresión con HTSeq:
    - `vista de samtools -@ 4 -Sb input/SRR479052.chr21_alignments.sam | samtools sort -@ 4 -o input/SRR479052.chr21_alignments.bam`: este comando convierte el archivo SAM `SRR479052.chr21_alignments.sam` al formato BAM, lo ordena y lo guarda como `SRR479052.chr21_alignments.bam`. También utiliza subprocesos múltiples con `-@ 4` para mejorar el rendimiento.
    - `htseq-count -f bam -r pos -s no -t exon -i gene_id input/SRR479052.chr21_alignments.bam input/Homo_sapiens.GRCh38.109.chr21.gtf > input/SRR479052.chr21_counts.txt`: este comando cuantifica la expresión de lecturas alineadas con características en el genoma de referencia utilizando la herramienta HTSeq. Especifica parámetros como el formato del archivo de entrada (`-f bam`), el tipo de característica (`-t exon`), el atributo a contar (`-i gene_id`) y el nombre del archivo de salida (`SRR479052.chr21_counts.txt`).
        i. Resultados de los comandos `htseq-count` generaron recuentos de genes a partir de los archivos de alineación para dos

muestras, `SRR479052` y `SRR479054`, frente al archivo de anotación de referencia `Homo_sapiens.GRCh38.109.chr21.gtf`

1. Para "SRR479052", los recuentos revelaron que 1758 lecturas no tenían características, 358 eran ambiguas, 7290 tenían una calidad de alineación demasiado baja y 1759 lecturas no estaban alineadas con ninguna característica. No hubo lecturas con alineaciones no únicas.
2. De manera similar, para `SRR479054`, 1267 lecturas no tenían características, 213 eran ambiguas, 4436 tenían una calidad de alineación demasiado baja y 1145 lecturas no estaban alineadas con ninguna característica. Nuevamente, no hubo lecturas con alineaciones no únicas.

ii. Estos resultados proporcionan información sobre la distribución de lecturas entre características genómicas y ayudan a evaluar la calidad de la alineación y la cuantificación de cada muestra.

Estos comandos y parámetros se eligieron para realizar de manera eficiente la indexación, alineación y cuantificación de la expresión, asegurando un análisis preciso de los datos de secuenciación. Además, se incorporan comprobaciones de la existencia de ficheros para evitar tratamientos redundantes.

Segunda Parte

Primero, se incluye la matriz de cuentas crudas correspondiente a los 24 cultivos analizados. Esta matriz constituye la base de datos primaria sobre la cual se realizarán las comparaciones y análisis pertinentes para determinar los genes diferencialmente expresados entre diferentes grupos de muestras. Además, se dispone de un data frame que contiene los metadatos asociados al experimento. Estos metadatos proporcionan información crucial sobre las características de cada muestra, como el paciente, el tratamiento aplicado y el tiempo de exposición. Esta información contextual es fundamental para contextualizar los resultados y realizar análisis diferenciados según las variables de interés.

Por último, se suministra un archivo GMT, que será utilizado para llevar a cabo un Análisis de Enriquecimiento de Conjuntos de Genes (GSEA). Este análisis permitirá identificar patrones de expresión génica asociados a diferentes condiciones experimentales, lo que nos ayudará a comprender mejor los efectos de los tratamientos aplicados.

DESeq2

Se buscará determinar qué genes muestran una expresión diferencial entre las muestras tratadas con OHT y el control después de 24 horas, así como entre las muestras tratadas con DPN y el control durante el mismo período. Se entregarán tablas detalladas que incluyan los genes diferencialmente expresados, junto con los criterios utilizados para filtrar los resultados y cualquier gráfico relevante generado durante el análisis. El código se puede ver en el Anexo 2:

1. Cargar bibliotecas
   Se cargan las bibliotecas necesarias para realizar análisis de expresión diferencial y visualización de datos. Estas bibliotecas incluyen DESeq2 para análisis estadístico, ggplot2 y pheatmap para gráficos y mapas de calor, y otras herramientas útiles para manipular y representar datos.
2. Cargar datos
   Se cargan los datos necesarios para el análisis. Esto incluye metadatos que describen las características de la muestra y los recuentos sin procesar obtenidos de la secuenciación de ARN.
3. Modificación de datos
   La preparación de los datos se realiza antes de realizar el análisis. Se ajustan los nombres de filas y columnas, las variables relevantes se convierten en factores y se crea un nuevo 'grupo' de variables para la interacción entre las condiciones de tratamiento y el tiempo.
4. Análisis de expresión diferencial
   El análisis de expresión diferencial se realiza utilizando el paquete DESeq2. Esto implica crear un conjunto de datos DESeq2, filtrar genes con baja expresión y ajustar un modelo de regresión lineal generalizado para cada gen.
5. Calcular Variance-stabilizing transformation por DESeq y dibujar mapa de calor
   VST normaliza los datos de recuento y estabiliza la varianza utilizando factores de tamaño. ‘pheatmap’ genera mapas de calor agrupados con control mejorado sobre los parámetros gráficos. Como se puede mostrar en la figura 1.

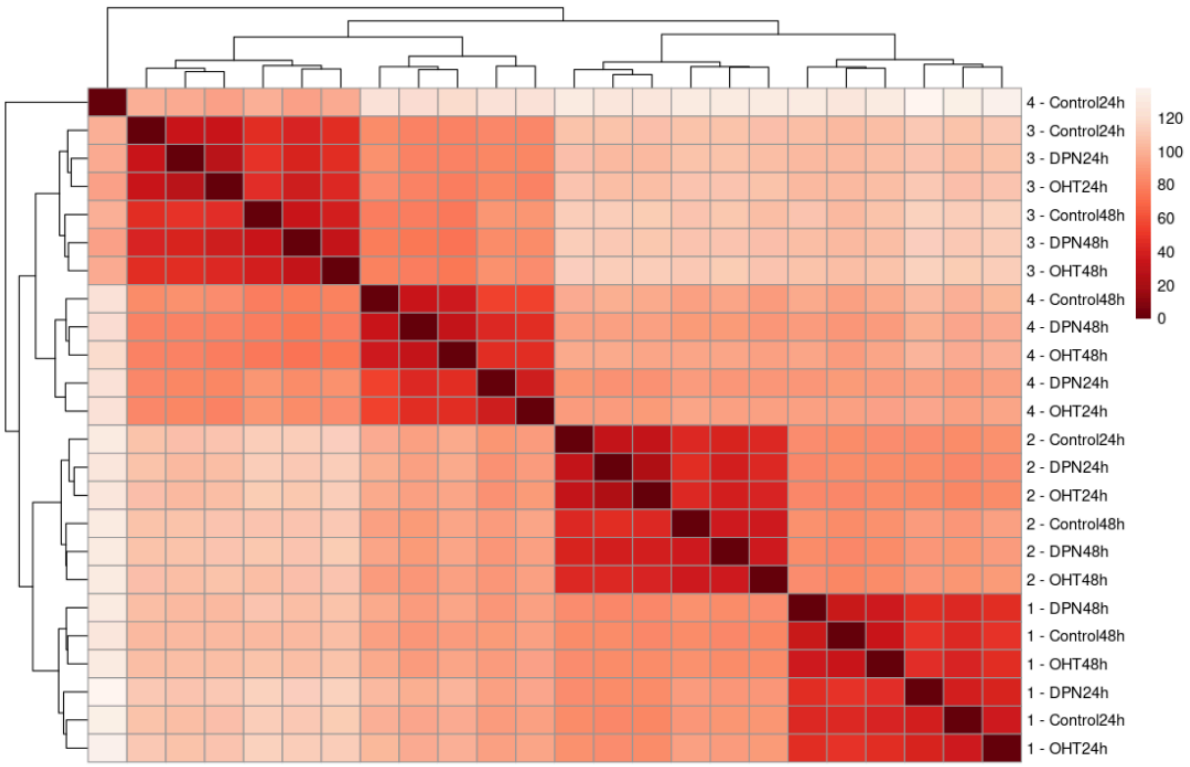


Figura 1 mapa de calor

6. <u>Análisis de expresión diferencial para tratamientos OHT y DPN después de 24h</u>

   Los análisis de expresión diferencial se realizan por separado para los tratamientos OHT y DPN después de 24 horas. Esto nos permite identificar genes que se expresan diferencialmente entre los grupos de tratamiento y control en cada tratamiento, en el que también podemos visualizar utilizando un mapa de calor, como se puede mostrar en la figura 2 y 3.

   - ☐ out of 24416 with nonzero total read count
     adjusted p-value < 0.05
     LFC > 0 (up) : 74, 0.3%
     LFC < 0 (down) : 13, 0.053%
     outliers [1] : 0, 0%
     low counts [2] : 10888, 45%
     (mean count < 28)
     [1] see 'cooksCutoff' argument of ?results
     [2] see 'independentFiltering' argument of ?results
   - ☐ out of 24416 with nonzero total read count
     adjusted p-value < 0.05
     LFC > 0 (up) : 44, 0.18%
     LFC < 0 (down) : 4, 0.016%
     outliers [1] : 0, 0%
     low counts [2] : 5681, 23%
     (mean count < 2)
     [1] see 'cooksCutoff' argument of ?results
     [2] see 'independentFiltering' argument of ?results

Estos resultados representan el análisis diferencial de expresión génica entre las muestras de control y las tratadas con DPN (DPN24h) y OHT (OHT24h) respectivamente. Para el tratamiento con DPN, se identificaron 74 genes con un aumento significativo en la expresión (upregulated) y 13 genes con una disminución significativa en la expresión (downregulated), todo con un nivel de significancia ajustado de $p < 0.05$ utilizando el método de corrección de Benjamini-Hochberg. No se encontraron valores atípicos, pero el número de recuentos bajos es considerable (45% con un promedio de recuento < 28). Para el tratamiento con OHT, se identificaron 44 genes con un aumento significativo en la expresión y 4 genes con una disminución significativa en la expresión, también con un nivel de significancia ajustado de $p < 0.05$. En este caso, el porcentaje de recuentos bajos es menor (23% con un promedio de recuento < 2).

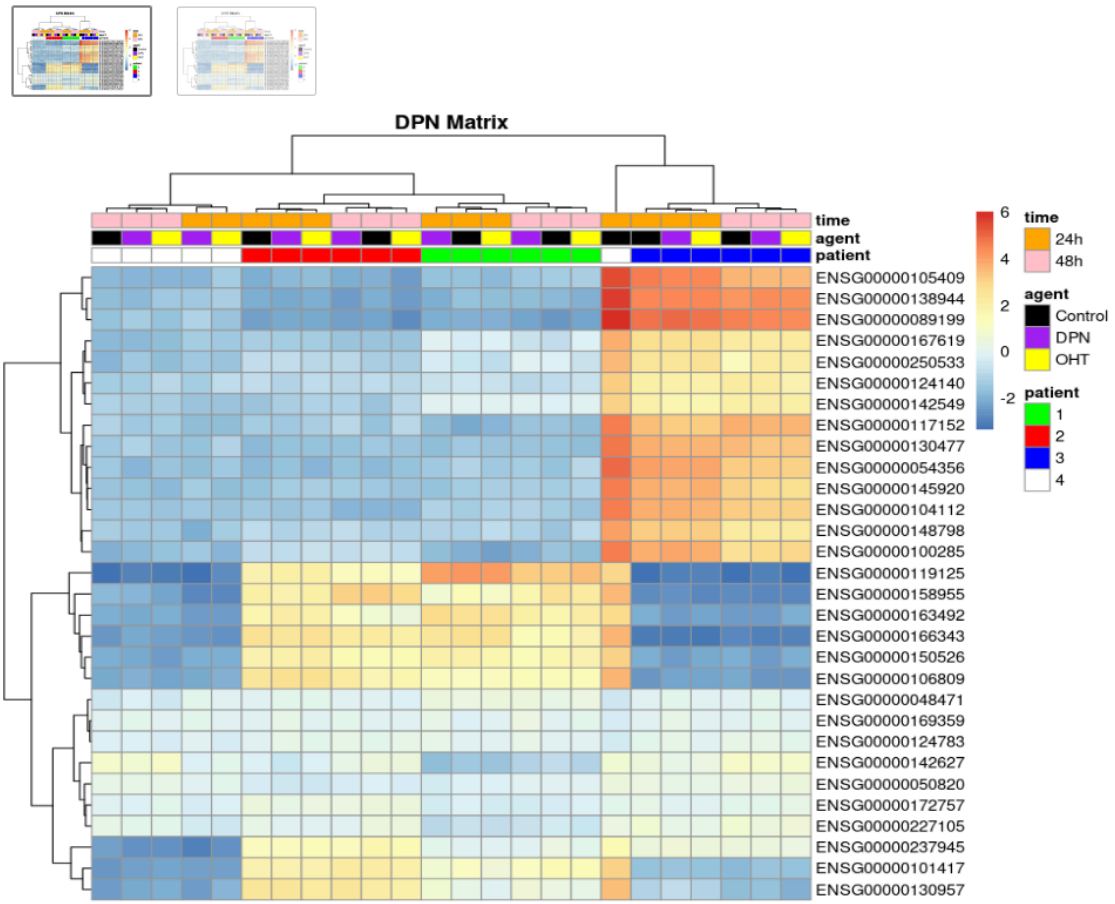


Figura 2 Matriz de DPN

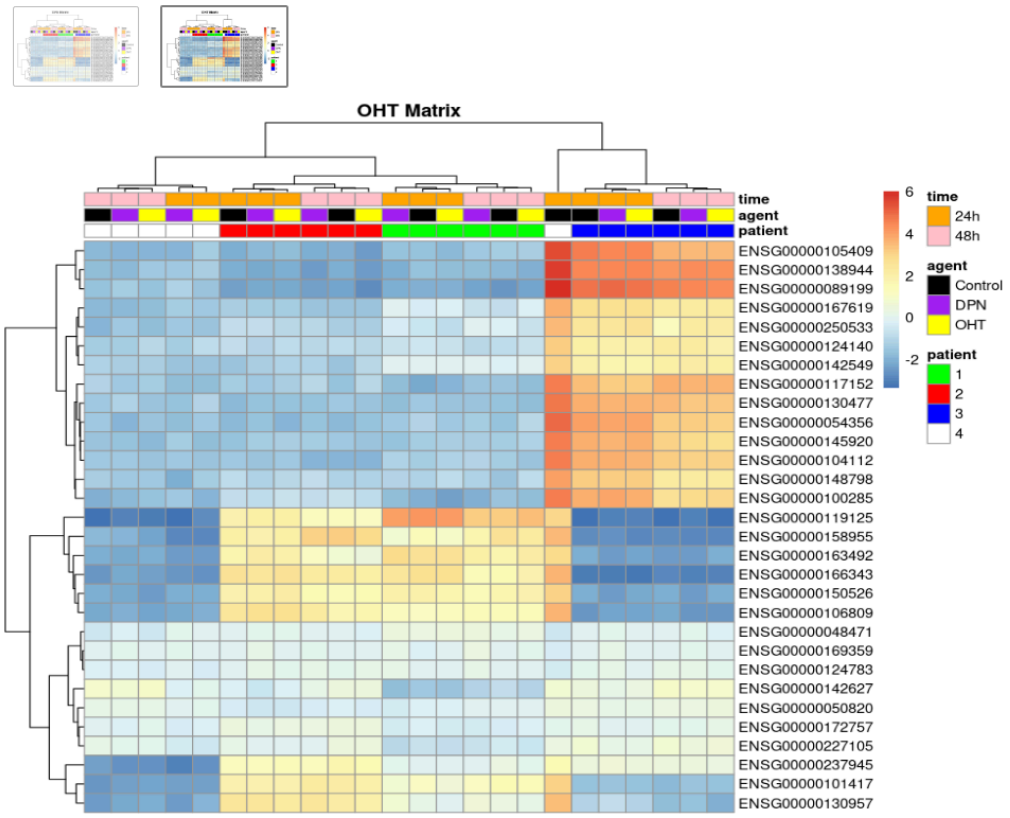


Figura 3 Matriz de OHT

7. <u>Guardar análisis de expresión diferencial</u>
   Los resultados del análisis de expresión diferencial se guardan en un archivo para su posterior referencia y análisis.
8. <u>Crear rango para análisis GSEA</u>
   Se crea un archivo de clasificación que contiene los genes clasificados según su cambio de expresión entre el grupo tratado con DPN y el grupo de control después de 24 horas. Este archivo se utiliza para realizar análisis de enriquecimiento genético (GSEA) para identificar vías biológicas asociadas con cambios en la expresión genética.

   ☐ 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

   using 'apeglm' for LFC shrinkage. If used in published research, please cite:
   Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895

☐ 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

Los resultados muestran el análisis diferencial de expresión génica entre las muestras tratadas con DPN durante 24 horas y las muestras de control tomadas en el mismo intervalo de tiempo. Se observa que de un total de 24,416 genes con recuentos totales no nulos, se identifican 13 genes con un incremento significativo en la expresión (upregulated) con un fold change ajustado positivo y 74 genes con una disminución significativa en la expresión (downregulated) con un fold change ajustado negativo, todo con un nivel de significancia de $p < 0.05$. Además, se señala que no hay valores atípicos ni recuentos bajos después del ajuste. La técnica utilizada para el ajuste de fold change es "apeglm", con el fin de mejorar la precisión del análisis.

GSEA

Se explorará si el tratamiento con DPN tiene algún efecto en las primeras 24 horas de exposición, antes de que los cambios de expresión sean más evidentes a las 48 horas. Se realizará un análisis de GSEA utilizando el GMT proporcionado, y se examinarán los resultados para determinar conclusiones significativas sobre el efecto del tratamiento a las 24 horas. Se presentarán tablas con los resultados del análisis, resaltando las columnas clave utilizadas para extraer conclusiones, así como los gráficos característicos de este tipo de análisis.

Al generar el GSEA Preranked, se genera un archivo lleno de gráficos del análisis. Un informe proporciona detalles sobre el enriquecimiento de conjuntos de genes en dos fenotipos distintos, uno identificado como "perturbado" (na_pos), en el que se identificaron 19 de los 50 conjuntos de genes y otro como "imperturbado" (na_neg), en el que 31 de los Se identificaron 50 conjuntos de genes.
Por los "perturbado" (na_pos), se identifican 6 conjuntos de genes significativos con un FDR < 25%, mientras que 1 conjunto de genes muestra un enriquecimiento significativo con un valor de p nominal < 1%. Adicionalmente, 4 conjuntos de genes presentan un enriquecimiento significativo con un valor de p nominal < 5%.
Mientras, se identifican 19 conjuntos de genes con un enriquecimiento significativo con un FDR < 25%, mientras que 9 conjuntos de genes presentan un enriquecimiento significativo con un valor de p nominal < 1%. Asimismo, se observa que 17 conjuntos de genes muestran un enriquecimiento significativo con un valor de p nominal < 5%.

"Perturbado" (na_pos):

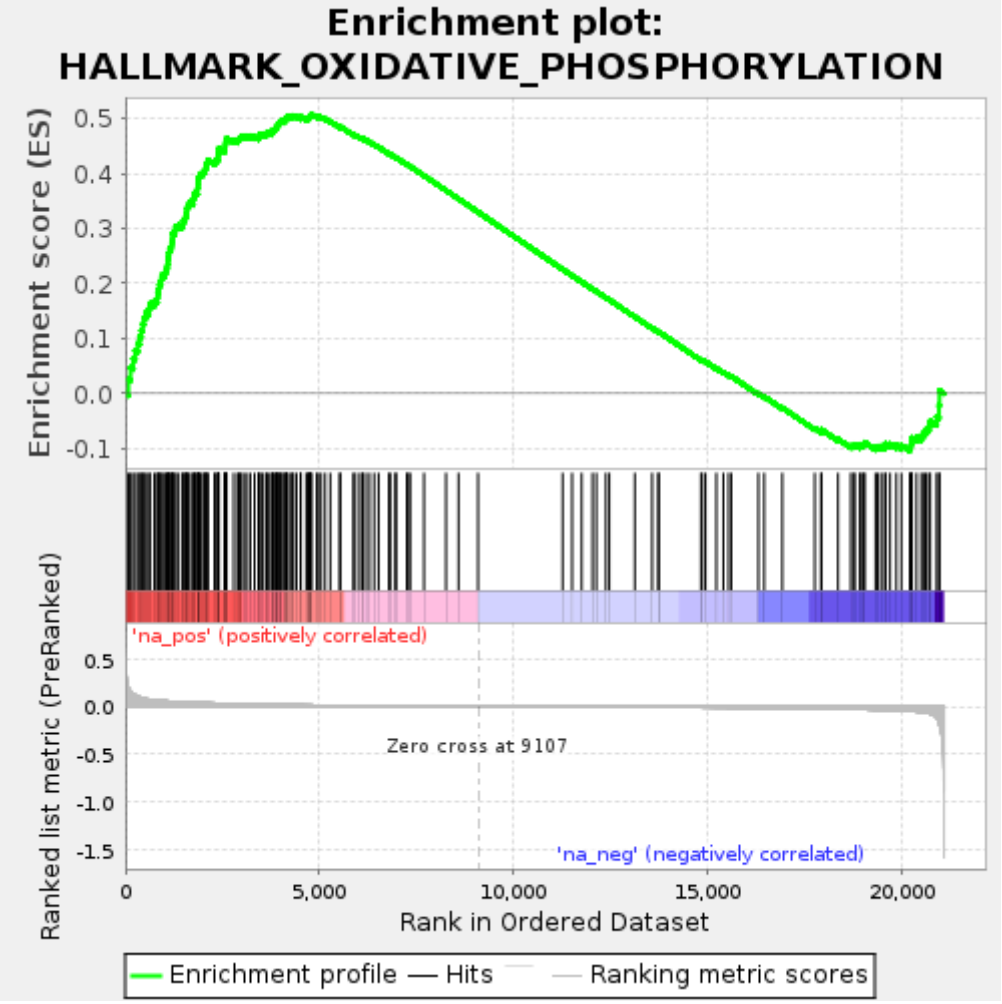


Figura 4

El conjunto de datos de enriquecimiento más alto se estableció como:"HALLMARK_OXIDATIVE_PHOSPHORYLATION". El archivo proporciona detalles específicos sobre el conjunto de genes "HALLMARK_OXIDATIVE_PHOSPHORYLATION", incluyendo su Puntuación de Enriquecimiento (ES) de 0.5080438 y la Puntuación de Enriquecimiento Normalizada (NES) de 1.5623937. Además, se presenta un valor de p nominal de 0.0014430014, indicando una significancia estadística en la asociación del conjunto de genes con el fenotipo de interés. El Valor q de FDR (False Discovery Rate) es de 0.082092226, mientras que el Valor p de FWER (Family-Wise Error Rate) es de 0.134, lo que proporciona información adicional sobre la fiabilidad de los resultados obtenidos. Marcar los genes más altos son: mitochondrial ribosomal protein S30 [Source:HGNC Symbol;Acc:HGNC:8769] *MRPS30* y HtrA serine peptidase 2 [Source:HGNC Symbol;Acc:HGNC:14348] *HTRA2*.

"Imperturbado" (na_neg):

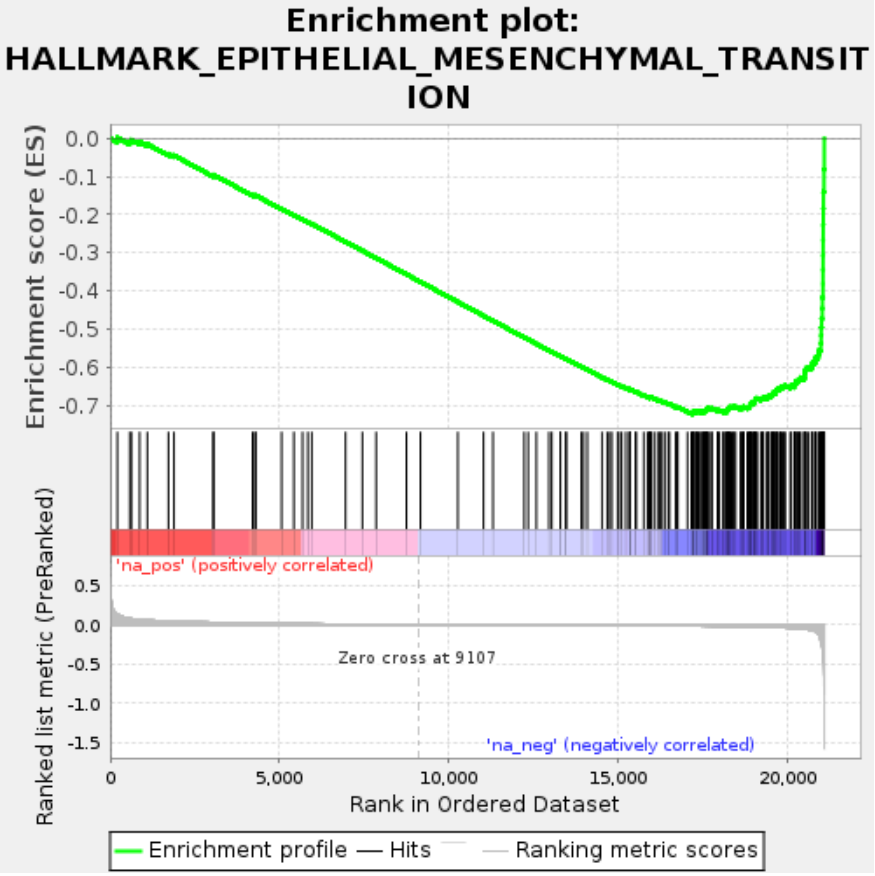


Figura 5

EL conjunto de datos de enriquecimiento más alto se estableció como:"HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION". El archivo detalla el conjunto de genes "HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION", con un Puntaje de Enriquecimiento (ES) de -0.72311187 y un Puntaje de Enriquecimiento Normalizado (NES) de -2.3637903. Además, presenta un valor de p nominal de 0.0, lo que indica una fuerte asociación del conjunto de genes con el fenotipo de interés. Tanto el Valor q de FDR (False Discovery Rate) como el Valor p de FWER (Family-Wise Error Rate) también son 0.0, lo que sugiere una alta confiabilidad en los resultados obtenidos. Marcar los genes más altos son: vascular endothelial growth factor A [Source:HGNC Symbol;Acc:HGNC:12680] *VEGFA* y solute carrier family 6 member 8 [Source:HGNC Symbol;Acc:HGNC:11055] *SLC6A8.*

Conclusión

En conclusión, los datos preparatorios sugieren que el desequilibrio en la expresión genética no es atribuible a variaciones en los tratamientos. El análisis con DESeq2 revela una pequeña cantidad de genes con expresión diferencial a las 24 horas para ambos tratamientos, y los patrones observados en los mapas de calor indican que los grupos de genes están relacionados con las características individuales de la muestra. La GSEA indica que el tratamiento con DPN tiene un impacto significativo a

lo largo del tiempo, reflejado en la regulación positiva de genes asociados con el fenotipo "perturbado" y la regulación negativa de genes relacionados con el fenotipo "no perturbado". Estos hallazgos sugieren que los efectos de la DPN pueden ser más notorios después de 48 horas, lo que indica un posible efecto de la DPN dentro de las primeras 24 horas.

Referencia

1. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data
2. GitHub - samtools/samtools: Tools (written in C using htslib) for manipulating next-generation sequencing data
3. HISAT2 graph-based alignment of next generation sequencing reads to a population of genomes
4. HTSeq
5. GSEA

Anexo 1

```
#!/bin/bash

# Quality Control with FastQC and contamination analysis with FastQScreen
# Checking existence of FASTQC files
if [ ! -f "fastqc/SRR479052.chr21_1_fastqc.html" ] || [ ! -f "fastqc/SRR479052.chr21_2_fastqc.html" ]
|| [ ! -f "fastqc/SRR479054.chr21_1_fastqc.html" ] || [ ! -f "fastqc/SRR479054.chr21_2_fastqc.html"
]; then
        # Executing FASTQC if files don't exist
        fastqc input/SRR479052.chr21_1.fastq input/SRR479052.chr21_2.fastq
input/SRR479054.chr21_1.fastq input/SRR479054.chr21_2.fastq -o fastqc
else
        echo "FASTQC files already exit. Continue..."
fi

# Checking existence of FASTQScreen reports
if [ ! -f "fastq_screen_v0.13.0/fastq_screen.summary.txt" ]; then
        # Executing FastQScreen if files don't exist
        perl fastq_screen_v0.13.0/fastq_screen --aligner bowtie2 input/SRR479052.chr21_1.fastq
input/SRR479052.chr21_2.fastq input/SRR479054.chr21_1.fastq input/SRR479054.chr21_2.fastq
else
        echo "FastQScreen report already exists. Continue..."
fi

# Indexing and Alignment with HISAT2
# Checking existance of alignment files
if [ ! -f "input/SRR479052.chr21_alignments.sam" ] || [ ! -f "input/SRR479054.chr21_alignments.sam"
]; then
        # Indexing chromosome 21
        hisat2-build input/Homo_sapiens.GRCh38.dna.chromosome.21.fa input/Homo_sap_index
```

```
        # Alignment of samples with chromosome 21
        hisat2 -x input/Homo_sap_index -1 input/SRR479052.chr21_1.fastq -2
input/SRR479052.chr21_2.fastq -S input/SRR479052.chr21_alignments.sam
        hisat2 -x input/Homo_sap_index -1 input/SRR479054.chr21_1.fastq -2
input/SRR479054.chr21_2.fastq -S input/SRR479054.chr21_alignments.sam
else
        echo "Alignment files already exists. Continue..."
fi

# Alignment statistics with SAMTools and quantification of expression with HTSeq
# Checking existence of files
if [ ! -f "input/SRR479052.chr21_counts.txt" ] || [ ! -f "input/SRR479054.chr21_counts.txt" ]; then
        # Alignment statistics with SAMTools
        samtools view -@ 4 -Sb input/SRR479052.chr21_alignments.sam | samtools sort -@ 4 -o
input/SRR479052.chr21_alignments.bam
        samtools view -@ 4 -Sb input/SRR479054.chr21_alignments.sam | samtools sort -@ 4 -o
input/SRR479054.chr21_alignments.bam
        # Quantification of counts expression with HTSeq
        htseq-count -f bam -r pos -s no -t exon -i gene_id input/SRR479052.chr21_alignments.bam
input/Homo_sapiens.GRCh38.109.chr21.gtf > input/SRR479052.chr21_counts.txt
        htseq-count -f bam -r pos -s no -t exon -i gene_id input/SRR479054.chr21_alignments.bam
input/Homo_sapiens.GRCh38.109.chr21.gtf > input/SRR479054.chr21_counts.txt
else
        echo "Counts file already exists. Continuando..."
fi

echo 'Process Complete'
```

Anexo 2

Cargar paquetes necesarias
```{r echo=FALSE}
library("DESeq2")         # Para análisis de expresión diferencial
library("ggplot2")        # Para visualización de datos
library("ggplotify")      # Para integrar gráficos ggplot2 en otros formatos
library("tidyverse")      # Para manipulación de datos
library("pheatmap")       # Para generar mapas de calor
library("RColorBrewer")   # Para generar paletas de colores
```

Cargar Data
```{r}
# Cargar metadata
metadata <- read.csv(file = "input/metadata.tsv", sep = "\t")

# Cargar raw counts
raw_counts <- read.csv(file = "input/rawcounts.tsv", sep = "\t", row.names = 1)

```

Modificaccion de datos
```{r}
# Establecer los nombres de fila como los nombres de columna de los raw counts
rownames(metadata) <- colnames(raw_counts)

# Convertir las variables pertinentes en factores
metadata <- mutate(.data = metadata,
                   X = NULL,  # Eliminar columna no deseada
                   patient = as.factor(patient),
                   agent = as.factor(agent),
                   time = as.factor(time))

# Comprobar si los nombres de columna coinciden con los de fila
identical(colnames(raw_counts), rownames(metadata))

# Crear una nueva columna 'group' para la interacción entre 'agent' y 'time'
metadata$group <- as.factor(interaction(metadata$agent, metadata$time, sep = ""))

```

Análisis de expresión diferencial
```{r}
# Crear un conjunto de datos de DESeq2
dds <- DESeqDataSetFromMatrix(countData = raw_counts,
                              colData = metadata,
                              design = ~ patient + group)

# Filtrar genes con menos de 10 recuentos
keep <- rowSums(counts(dds)) >= 10
dds2 <- dds[keep, ]

# Ajuste del modelo GLM: Estadística de Wald
dds3 <- DESeq(dds2, test = "Wald")
summary(dds3)
```

```{r}
# Variance-stabilizing transformation por DESeq
vst_m <- vst(dds3)
```

Variance-stabilizing transformation por DESeq y mapa de calor
```{r}
# Mapa de calor
# Calcular distancias de muestra
sampleDists <- dist(t(assay(vst_m)))

# Convertir objeto de distancia en matriz
sampleDistMatrix <- as.matrix(sampleDists)
```

```
# Asignar nombres de filas según el paciente y el grupo
rownames(sampleDistMatrix) <- paste(vst_m$patient, vst_m$group, sep = " - ")

# Borrar nombres de columnas
colnames(sampleDistMatrix) <- NULL

# Definir colores para el mapa de calor
colors <- colorRampPalette(rev(brewer.pal(9, "Reds")))(255)

# Dibujar mapa de calor agrupado
pheatmap(sampleDistMatrix,
         clustering_distance_rows = sampleDists,
         clustering_distance_cols = sampleDists,
         col = colors)
```

Análisis de expresión diferencial para tratamientos con OHT y DPN después de 24h
```{r}
# Resultados para DPN
res_DPN <- results(object = dds3,
                   contrast = c("group", "Control24h", "DPN24h"),
                   alpha = 0.05,
                   pAdjustMethod = "BH")

summary(res_DPN)

# Resultados para OHT
res_OHT <- results(object = dds3,
                   contrast = c("group", "Control24h", "OHT24h"),
                   alpha = 0.05,
                   pAdjustMethod = "BH")

summary(res_OHT)
```

Mapa de color por visualicacion
```{r}
# Extraiga las 30 filas superiores de la matriz de datos de expresión transformada para la condición DPN
DPN_matrix <- assay(vst_m)[head(order(res_DPN$padj), 30), ]

# Restar los medios de fila de la matriz de expresión DPN para centrar los datos
DPN_matrix <- DPN_matrix - rowMeans(DPN_matrix)

# Extraiga las 30 filas superiores de la matriz de datos de expresión transformada para la condición OHT
OHT_matrix <- assay(vst_m)[head(order(res_DPN$padj), 30), ]

# Restar los medios de fila de la matriz de expresión OHT para centrar los datos
```

```
OHT_matrix <- OHT_matrix - rowMeans(OHT_matrix)

# Extraiga columnas de metadatos de paciente, agente y tiempo y guárdelas
annotat <- as.data.frame(colData(vst_m)[ , c("patient", "agent", "time")])

# Definir colores para las columnas de anotaciones
colors <- list(
        patient = c("1" = "green", "2" = "red", "3" = "blue", "4" = "white"),
        agent = c(Control = "black", DPN = "purple", OHT = "yellow"),
        time = c("24h" = "orange", "48h" = "pink")
)

# Crear mapa de calor para la condición DPN con columnas anotadas y colores especificados
pheatmap(mat = DPN_matrix, annotation_col = annotat, show_colnames = FALSE,
        annotation_colors = colors, main = "DPN Matrix")

# Crear mapa de calor para la condición OHT con columnas anotadas y colores especificados
pheatmap(mat = OHT_matrix, annotation_col = annotat, show_colnames = FALSE,
        annotation_colors = colors, main = "OHT Matrix")
```
Guardar análisis de expresión diferencial
```{r}
saveRDS(object = dds3, file = "input/dds3.rds")
```

Crear rango para análisis GSEA
```{r}
# Resultados para DPN para GSEA
res <- results(dds3, alpha = 0.05, contrast = c("group", "DPN24h", "Control24h"))
summary(res)

# Ajustar el LFC para GSEA
res.ape <- lfcShrink(dds = dds3, coef = "group_DPN24h_vs_Control24h", type = "apeglm",
                res = res)
summary(res.ape)

# Crear archivo de rango para GSEA
rnk <- data.frame(Feature = rownames(res.ape), LFC = res.ape$log2FoldChange)
write.table(rnk, file = "input/ranked_DNP_24h.rnk", sep = "\t", quote = FALSE,  col.names = FALSE,
row.names = FALSE)

```
```
```