Reprogramación transcriptómica inducida por hiperglucemia en células de cáncer de mama: análisis de expresión diferencial y enriquecimiento funcional en rutas proliferativas y de daño genómico

Yixi Zhang, Pablo Pérez Rodríguez
22 de junio de 2025

Introducción

El cáncer de mama representa una de las principales causas de mortalidad a nivel global, siendo la neoplasia maligna más común en mujeres, con más de 2,3 millones de casos diagnosticados en 2020 y más de 680.000 muertes anuales, según estimaciones de la Organización Mundial de la Salud (OMS, 2024). La etiopatogenia del cáncer de mama involucra una compleja interacción de factores genéticos, hormonales, ambientales y epigenéticos, que dan lugar a un espectro heterogéneo de subtipos moleculares (como los luminales, HER2-positivos y triple negativos), con distinto pronóstico, evolución clínica y respuesta terapéutica (Perou et al., 2000; Yersal & Barutca, 2014).

En las últimas décadas, se ha acumulado la evidencia acerca del papel de enfermedades metabólicas como la diabetes mellitus tipo 2 (DM2) como comorbilidades que no solo aumentan el riesgo de desarrollar ciertos tipos de cáncer, sino que también modifican su biología tumoral y complican su tratamiento (Giovannucci et al., 2010; Wang et al., 2012). Estudios epidemiológicos han identificado una asociación significativa entre la presencia de DM2 y una mayor incidencia, progresión y letalidad del cáncer de mama (Lipscombe et al., 2006; Autier et al., 2010). Esta relación se ha atribuido a múltiples mecanismos interrelacionados, incluyendo la hiperglucemia crónica, la hiperinsulinemia, la inflamación sistémica, el estrés oxidativo y la desregulación de rutas metabólicas implicadas en el crecimiento y la supervivencia celular (Zhao et al., 2022).

A pesar de estos hallazgos, aún sigue sin estar completamente esclarecido cómo la diabetes modifica de forma tan específica el transcriptoma tumoral, y qué implicaciones funcionales tienen estos cambios a nivel celular y terapéutico. En particular, la posibilidad de que el

microambiente hiperglucémico pueda inducir vulnerabilidades específicas en células cancerosas (por ejemplo, déficits en la reparación del DNA) plantea oportunidades en el desarrollo de estrategias terapéuticas dirigidas. Estudios previos han demostrado que la glucosa podría actuar como una señal metabólica regulando genes implicados en la proliferación celular, ciclo celular, apoptosis y la respuesta al daño genético, pero los datos siguen siendo fragmentarios y dependientes del contexto celular (Lei et al., 2023; Masoud & Pagès, 2021). En particular, algunos estudios apuntan a que un entorno hiperglucémico podría inducir un efecto premetafásico caracterizado por una mayor capacidad de invasión y migración celular, a la vez que comprometer mecanismos clave en el mantenimiento genómico, como la reparación del DNA. Esta doble alteración (una mayor agresividad junto a una menor fidelidad replicativa) plantea una paradoja funcional con importantes implicaciones terapéuticas

El presente trabajo se basa en el estudio desarrollado por Panigrahi et al. (2023), quienes utilizaron un enfoque transcriptómico de RNA-seq para caracterizar el impacto de la hiperglucemia en líneas celulares de cáncer de mama tipo Hs578T, bajo condiciones de baja (5mM) y alta glucosa (25mM) durante 48 horas; analizando las consecuencias funcionales de los cambios de la expresión génica mediante análisis bioinformáticos y validación de modelos xenográficos murinos.

Los resultados revelaron que la exposición a altas concentraciones de glucosa induce un profundo cambio en la expresión génica, caracterizado por la activación de vías asociadas a la transición epitelio-mesenquimal (EMT), la adquisición de fenotipos de células madre tumorales y, notablemente, una marcada desregulación de genes implicados en la reparación del DNA. Este último hallazgo sugiere que la hiperglucemia puede inducir un estado de deficiencia en la reparación genómica lo cual, paradójicamente, podría aumentar la sensibilidad de las células cancerosas a ciertos tratamientos como los inhibidores de PARP (poly-ADP-ribose-polymerase) (Panigrahi et al., 2023; Lord & Ashworth, 2017).

La posibilidad de que condiciones del microambiente (en este caso, la glucosa) modulen directamente la eficacia de fármacos dirigidos a rutas de reparación del DNA abre nuevas líneas de investigación en la oncología personalizada y la farmacogenómica. Asimismo, la identificación de firmas transcriptómicas asociadas a la DM2 podría permitir una mejor estratificación de pacientes, y e diseño de intervenciones terapéuticas específicas para subgrupos de pacientes con cáncer de mama y comorbilidad metabólica.

En este contexto, el presente trabajo se propone reproducir el análisis de expresión génica diferencial (DEA) realizado en el citado estudio de Panigrahi, empleando los datos de secuenciación RNA-Seq depositados en el repositorio público SRA (Sequence Read Archive) bajo el BioProject PRJNA990784. En concreto se han seleccionado tres réplicas biológicas cultivadas en condiciones normoglucémicas (SRR25118847, SRR25118848, SRR25118849) y tres en condiciones hiperglucémicas (SRR25118851, SRR25118852, SRR25118853). A

partir de estas muestras se llevará a cabo el pipeline completo de un análisis bioinformático, que incluirá la evaluación de calidad, el alineamiento al genoma de referencia, la obtención de matrices de conteo y el análisis de la expresión diferencialmente mediante la herramienta DESeq2.

Posteriormente, los genes diferencialmente expresados (DEGs) se someterán a un análisis funcional de sobrerrepresentación (ORA) mediante el paquete ClusterProfiler (Yu et al., 2014) de Bioconductor, con el fin de identificar términos enriquecidos del GeneOntology (GO) y rutas biológicas alteradas por la hiperglucemia. Este enfoque permitirá confirmar si, efectivamente, la exposición a altas concentraciones de glucosa modula genes clave implicados en la reparación del DNA, el metabolismo celular, la plasticidad fenotípica y otros procesos relevantes en la progresión tumoral.

Materiales y métodos

a. Protocolo de Secuenciación Masiva

El presente trabajo se basa en el análisis de datos transcriptómicos derivados del estudio publicado por Panigrahi et al. (2023), quienes emplearon un diseño experimental basado en el tratamiento de líneas celulares de cáncer de mama humano con distintos niveles de glucosa. Concretamente, las muestras analizadas corresponden a la línea celular Hs578T, un modelo ampliamente empleado en la investigación la biología del cáncer de mama triple negativo (TNBC, por sus siglas en inglés), caracterizado por la ausencia de expresión en receptores hormonales y HER2, y asociado a un fenotipo particularmente agresivo y resistente hacia terapias convencionales.

Las células Hs578T fueron cultivadas en un medio DMEM (Dulbecco's Modified Eagle Medium), suplementado con 10% de FBS (Suero Bovino fetal), 2mM de L-Glutamina, y penicilina/estreptomicina (100 U/mL y 100 µg/mL respectivamente). Se emplearon dos condiciones experimentales diferenciadas: una con glucosa baja (5 mM) y otro con glucosa alta (25 mM), las cuales simularon un entorno normoglucémico e hiperglucémico respectivamente; durante 48 horas previo a la extracción de RNA.

La extracción de RNA total se realizó empleando RNeasy Plus Mini Kit (Qiagen, Hilden, Alemania), un método basado en columnas de membrana de sílice y acompañado de un cartucho gDNA Eliminator que permite la eliminación directa y eficiente de ADN genómico sin necesidad de tratamiento con DNasa. Según información del fabricante, este método garantiza la obtención de ARN de alta calidad, con un RIN cercano a 10 (Figura 1); apto para técnicas sensibles como RNA-seq y RT-qPCR, incluso a partir de pocos tipos celulares (Quiagen, 2025).

Aunque el artículo no especifica el protocolo exacto de la extracción de RNA ni la herramienta de control de calidad utilizada, los datos de secuenciación están depositados en el repositorio Sequence Read Archive (SRA) bajo el BioProject PRJNA990784, que agrupa seis muestras: tres correspondientes a la condición normoglucémica (SRR25118847, SRR25118848, SRR25118849) y tres a la condición hiperglucémica (SRR25118851, SRR25118852, SRR25118853); asegurando la robustez de los análisis bioinformáticos subsecuentes.

Según los metadatos disponibles en el SRA, la construcción de las librerías de RNA-Seq se llevó a cabo utilizando el kit NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs), que permite la preparación de librerías direccionales a partir de RNA mensajero enriquecido por selección de poly-A. Las muestras fueron secuenciadas por Novogene Corporation (China) en la plataforma Illumina NovaSeq 6000, en un esquema de lectura paired-end (2x150pb). Esta configuración garantiza una alta calidad de lectura, adecuada para la cuantificación de genes y el análisis diferencial de expresión,

El rendimiento promedio de secuenciación fue de aproximadamente 30 a 35 millones de lecturas por muestra, con una tasa de bases de calidad al 90% (Q30), según los reportes de Novogene disponibles en el SRA. Este volumen de datos proporciona una profundidad suficiente para realizar análisis transcriptómicos de alta resolución.

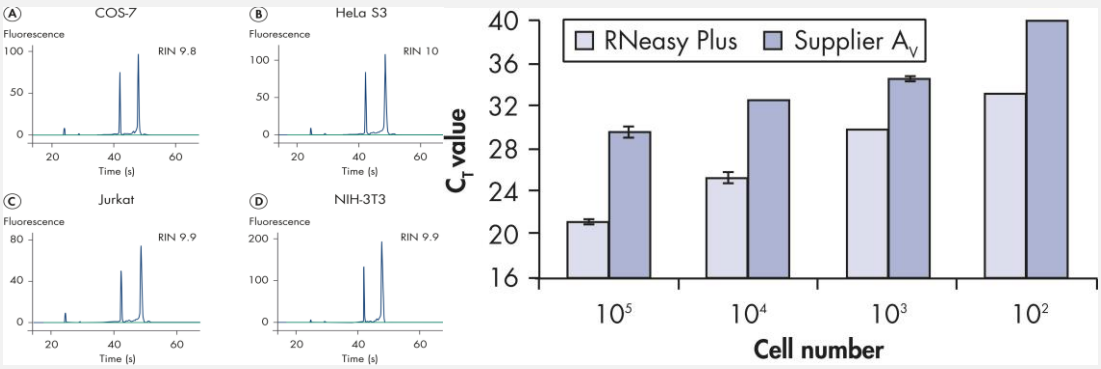


Figura 1: Representación de perfiles de calidad de RNA total obtenidos con el kit RNeasy Plus Mini Kit (Qiagen), basada en la información técnica del fabricante.

b. Workflow Bioinformático

El análisis bioinformático se estructuró en varias etapas secuenciales siguiendo un pipeline estandarizado para datos de secuenciación RNA-Seq, desde la descarga de datos crudos (RAW DATA) hasta la obtención e interpretación funcional de genes diferencialmente expresados (DEGs). El proceso fue ejecutado en el entorno GNU/Linux utilizando herramientas ampliamente validadas en transcriptómica, con versiones específicas detalladas en la Tabla 1.

Tabla 1: Herramientas bioinformáticas utilizadas en el análisis de RNA-Seq, con versiones y propósito funcional dentro del pipeline.

Herramienta	Versión	Propósito
`fastq-dump`	v3.2.1	Descarga de archivos `.sra` desde el repositorio NCBI SRA
`prefetch`	v.3.2.1	Conversión de archivos `.sra` a formato `.fastq` legible
`fastqc`	v0.12.1	Evaluación de calidad de lecturas crudas (contenido GC, calidad por base, duplicados)
`multiqc`	v1.27.1	Agregación e integración de múltiples reportes de calidad en un único archivo
`trimmomatic`	v0.36	Recorte de adaptadores y filtrado de baja calidad en lecturas FASTQ.
`hisat2`	v.2.2.1	Alineación de lecturas contra el genoma de referencia, soportando empalmes
`samtools`	v1.2.1	Conversión de archivos `.sam` a `.bam`
`featureCounts`	v.2.2.1	Cuantificación de lecturas alineadas por gen a partir de archivos BAM y anotaciones

La descarga de datos brutos a partir de un archivo `.txt` que contenía los identificadores de acceso SRA (SRR25118847, SRR25118848, SRR25118849, SRR25118851, SRR25118852, SRR25118853). Para este propósito se desarrolló un script Bash automatizado `00_get_raw_data.sh` en el entorno conda activo `environmentYP`.

El script implemente una serie de entrada asegurando que los ID de acceso existan y tengan extensión `.txt`, y crea un directorio logs automáticamente en la carpeta de salida especificada. La ejecución del flujo comienza con la activación del entorno Conda y la descarga de los archivos `.sra` mediante `prefetch` (parte del paquete NCBI SRA Toolkit [https://github.com/ncbi/sra-tools/blob/master/CHANGES.md]). Posteriormente, con la herramienta `fastq-dump` se convirtieron los archivos a formato FASTQ, generando archivos `.fastq` para cada muestra.

A continuación, los archivos FASTQ fueron sometidos a una evaluación de calidad preliminar en un script `01_qc_raw_reads.sh` usando `fastqc` (Andrews, 2010) que permitió inspeccionar indicadores como la calidad promedio por base, contenido de GC, presencia de adaptadores y niveles de duplicación. Los informes individuales generados fueron integrados mediante `multiqc` (Ewels et al., 2016), facilitando una visión global de la calidad de todas las muestras.

Las lecturas crudas fueron entonces sometidas a un filtrado y recorte mediante la herramienta `trimmomatic` (Bolger et al., 2014) en el script `02_trimming.sh`. Este procedimiento tuvo en cuenta parámetros como la calidad mínima por base (`SLIDINGWINDOW:4:25`), la eliminación de bases iniciales o terminales de baja calidad (`LEADING:3` y `TRALING:3`) y una longitud mínima aceptable de las lecturas (`MINLEN:36`). Además se eliminó la contaminación por adaptadores mediante el módulo ILLUMINACLIP, empleando el archivo `TruSeq3-PE.fa`. Se adaptaron entradas tanto de lecturas single-end como paired-end, procesando las lecturas emparejadas de forma coordinada.

Una vez aplicado el trimming, se realizó un segundo control de calidad (script `03-post_fastqc.sh`), para verificar la efectividad del trimming asegurando la idoneidad de las lecturas para su posterior alineamiento. Esta etapa garantizó la integridad de los datos procesados.

Para el alineamiento, se utilizó el genoma humano *Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa* y su anotación *Homo_sapiens.GRCh38.114.gtf* descargados de Ensembl (https://ftp.ensembl.org/pub/release-114/), y la herramienta alineadora HISAT2 (Kim et al., 2019) gracias a un script `04_reads_alignment.sh`**.** Este cuenta con parámetros por defecto, incluyendo indexación previa al genoma con `hisat2-build` y el uso de múltiples hilos (`--threads 8`) para paralelización (`-p 4`). Las lecturas SE o PE se gestionaron mediante `-U` (single-end) o `-1/-2` (paired-end). Tras el alineamiento con HISAT2, los archivos en formato SAM fueron convertidos y ordenados a formato BAM mediante `samtools sort` (Li et al., 2009). Este procedimiento, implementado internamente mediante el módulo `bam_sort_core` de la biblioteca HTSlib, permitió generar archivos BAM ordenados por coordenadas genómicas, optimizados para análisis posteriores.

La cuantificación de los niveles de expresión se realizó a partir de los archivos `.bam` generados por HISAT2, utilizando la herramienta `featureCounts` (Liao et al., 2014), especificando los archivos de anotación del genoma (`.gtf`) correspondientes al genoma de referencia. Esta etapa generó una matriz de conteo de genes por muestra, que sirvió como entrada para el análisis de expresión diferencial posterior

El análisis de expresión diferencial (DEA) se realizó utilizando el paquete DESeq2 (Love et al., 2014) en R, partiendo de la matriz de conteos generada por `featureCounts` y un archivo `.csv` de metadatos que incluía el identificador de la muestra y la condición experimental (sampleID, condition). Como paso previo se filtraron las matrices para que solo queden aquellas filas con más de 10 lecturas en total. Se estableció la condición "LowGlucose" como referencia. Posteriormente, se utilizaron funciones del paquete ClusterProfiler para realizar tanto el análisis de enriquecimiento de términos de GeneOntolgy (GO) como el de conjuntos de genes (Gene Set Enrichment Analysis, GSEA). Para el análiss GO se empleó la función `enrichGO` con los siguientes parámetros:

```
gene = interesting_set,
OrgDb = org.Hs.eg.db
keyType = "ENSEMBL",
ont = "BP",
pvalueCutoff = 0.01,
pAdjustMethod = "BH",
universe = background_genes,
qvalueCutoff = 0.05
```

Se realizaron dos análisis por separado, uno para los genes up-regulados y otro para los down-regulados.
El análisis GSEA se llevó a cabo con anotaciones HallMark con todos los genes. Se descargó la anotación mediante `msigdbr` (Dolgalev, 2025) (`collection = "Homo Sapiens", category = "H"`), y se ejecutó sobre un listado ordenado de genes con el conjunto de anotaciones correspondientes (gene_list, TERM2GENE = anotación_HallMark_descargada, pvalueCutoff = 0.05)
La visualización de los resultados se generó con funciones de los paquetes `ggplot2` (Wickham, 2016) y `enrichplot` (Yu, 2025), incluyendo gráficos de barras, dotplots y diagramas de enriquecimiento.

Resultados

Todo el análisis fue realizado en un entorno de cómputo de alto rendimiento, dentro del clúster institucional, utilizando scripts reproducibles almacenados en el directorio: `/data/2025/grado_biotech/pablo.perezr/proyecto_final/SRV_01`

Antes de proceder con el filtrado y recorte de las secuencias, se realizó un análisis de calidad sobre las lecturas crudas empleando la herramienta FastQC, y se integraron los resultados mediante MultiQC. El conjunto de datos evaluados corresponde a seis muestras biológicas con lecturas paired-end, obtenidas mediante tecnología Illumina, con una longitud uniforme de 151 pares de bases por lectura, evidenciando un correcto desempeño del sistema de secuenciación.

Tal y como se observa en la Figura 2, el número total de secuencias por archivo osciló entre 11.6 y 12.8 millones de lecturas, lo cual proporciona una profundidad adecuada para el análisis de expresión diferencial. Sin embargo, se observó un porcentaje de duplicación elevado en todos los archivos, con valores comprendidos entre el 50.5 y el 60.3 %, situándose por encima del umbral de advertencia establecido por FastQC (Figura 3).

Sample Name	Dups	GC	Seqs
SRR25118847_1	58.8%	47.0%	12.8M
SRR25118847_2	51.6%	47.0%	12.8M
SRR25118848_1	58.8%	47.0%	12.2M
SRR25118848_2	51.7%	47.0%	12.2M
SRR25118849_1	55.8%	48.0%	11.7M
SRR25118849_2	50.5%	48.0%	11.7M
SRR25118851_1	58.9%	49.0%	11.6M
SRR25118851_2	56.1%	49.0%	11.6M
SRR25118852_1	60.3%	49.0%	12.5M
SRR25118852_2	57.2%	49.0%	12.5M
SRR25118853_1	59.1%	49.0%	12.2M
SRR25118853_2	56.1%	50.0%	12.2M

Figura 2: Tabla resumen de MultiQC con porcentaje de duplicación, contenido en GC y número de lecturas por muestra.

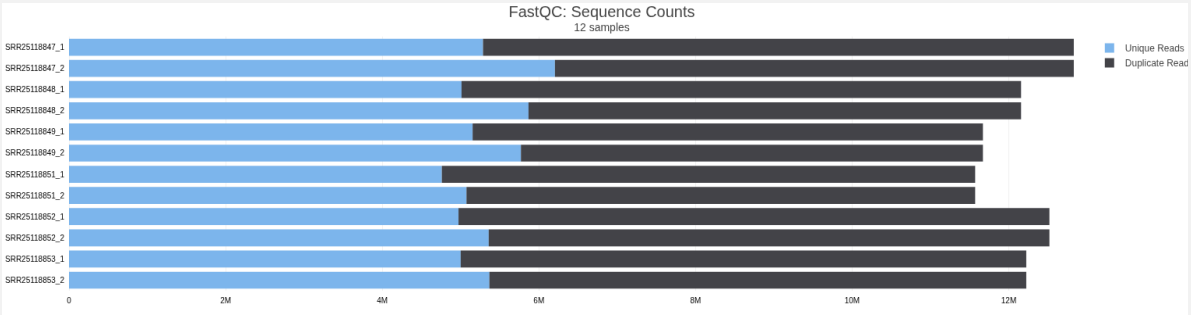


Figura 3: Gráfico de barras de FastQC con conteo de lecturas únicas y duplicadas por muestra (12 samples).

Este hallazgo fue consistente en todas las muestras, conduciendo a la marcación del módulo “Sequence Duplication Levels” como [FAIL] (Figura 4). No obstante, niveles elevados de duplicación pueden ser esperables en transcriptómica, especialmente si hay genes altamente

expresados o si la biblioteca presenta baja complejidad [Figura picos rojos]. Además, el contenido GC de las muestras osciló entre el 47 y el 50 % con una distribución unimodal simétrica y consistente entre archivos (Figura 1), lo cual es esperable en bibliotecas transcriptómicas humanas.

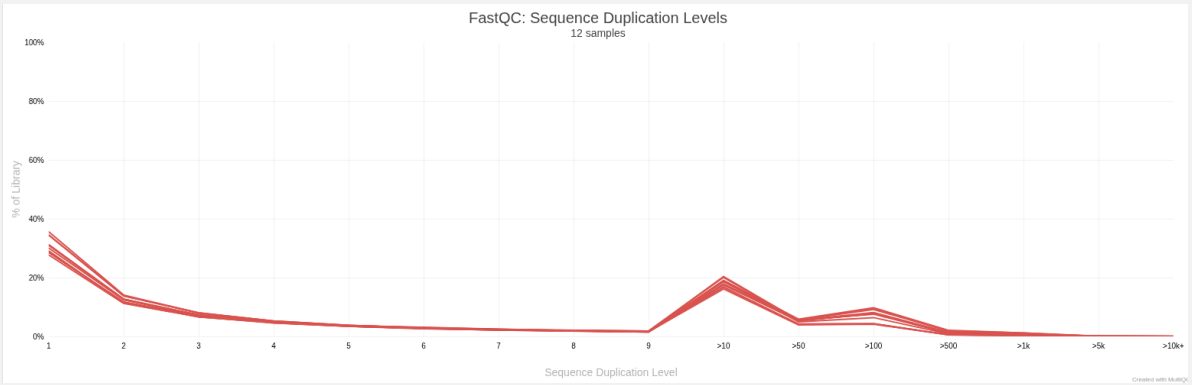


Figura 4: Distribución del porcentaje de secuencias duplicadas por muestra, indicando niveles elevados de duplicación (FastQC).

El análisis de calidad promedio por posición muestra valores de Phred Score iguales o mayores a 35 a lo largo de prácticamente toda la longitud de la lectura (Figura 5). Esto se traduce en una probabilidad de error menor del 0.001%, lo cual denota una excelente calidad de secuenciación. Este perfil es característico de una secuenciación eficiente, sin caída de calidad hacia los extremos 3', lo cual sugiere que el proceso de secuenciación fue técnicamente robusto.

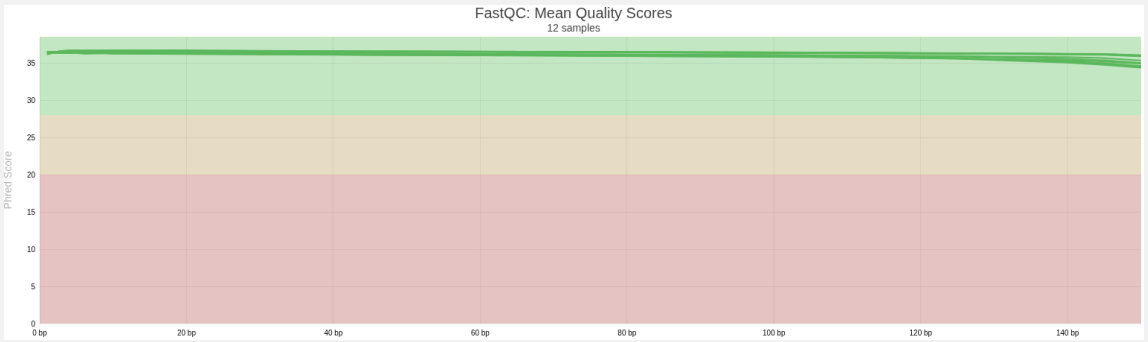


Figura 5: Gráfico Mean Quality Scores por posición (FastQC)**.**

El porcentaje de bases no identificadas (N) fue residual en todas las muestras, manteniéndose por debajo del 0.05% en todas las posiciones, sin ninguna advertencia crítica registrada por FastQC (Figura 6).

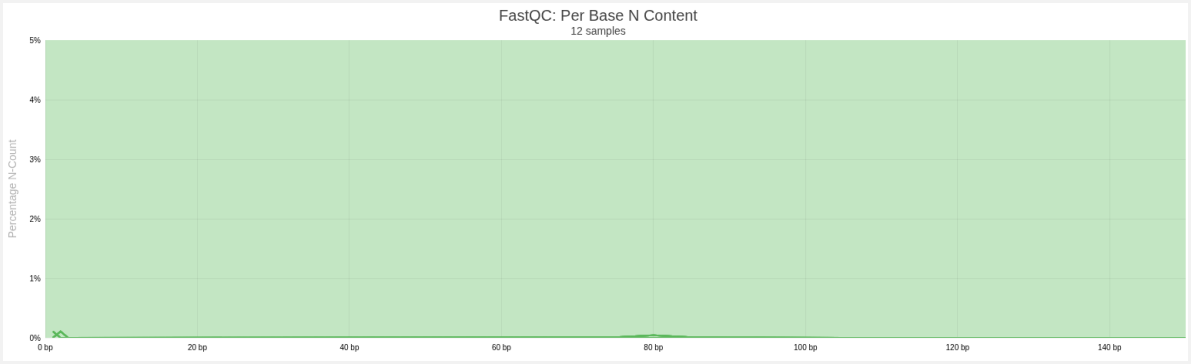


Figura 6: Porcentaje de bases ambiguas (N) por posición en las lecturas, sin alertas significativas (FastQC).

El contenido de adaptadores previo al recorte también fue evaluado. El gráfico de "Adapter Content" (Figura 7) muestra un incremento sostenido en la proporción acumulada de adaptadores a partir de los ~80 bp, alcanzando valores superiores al 20% en los extremos 3' de múltiples muestras, lo cual indica una alta probabilidad de solapamiento de adaptadores no eliminados durante la preparación de librerías. La presencia de adaptadores en una proporción tan elevada compromete el análisis posterior, ya que introduce sesgos en el alineamiento y cuantificación. Esto justifica la implementación de un proceso de recorte sistemático con herramientas como Trimmomatic en etapas posteriores del pipeline, para asegurar la integridad y calidad del dataset.

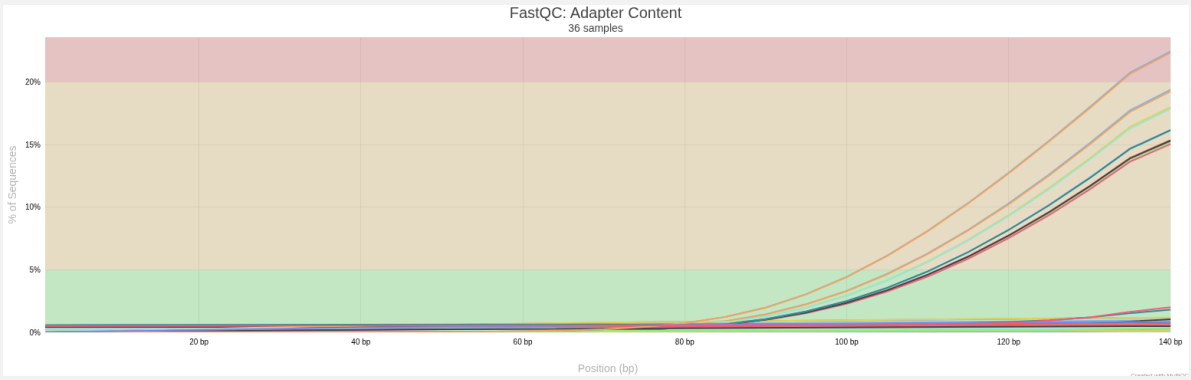


Figura 7: Contenido de adaptadores a lo largo de la longitud de las lecturas antes del recorte (FastQC).

Paralelamente, el resumen del control de calidad, "Status Checks" (Figura 8), reveló que todas las muestras (n=12, pares forward y reverse) fallaron en el parámetro específico de contenido de adaptadores, y mostraron advertencias o fallos adicionales en "Per base sequence content" (atribuibles al sesgo de iniciación asociados a los primers de la PCR de amplificación) y "Sequence Duplication Levels".
La presencia de adaptadores en una proporción tan elevada compromete el análisis posterior, ya que introduce sesgos en el alineamiento y cuantificación. Estos hallazgos justifican la

implementación de un proceso de recorte sistemático con herramientas como Trimmomatic en etapas posteriores del pipeline, para asegurar la integridad y calidad del dataset

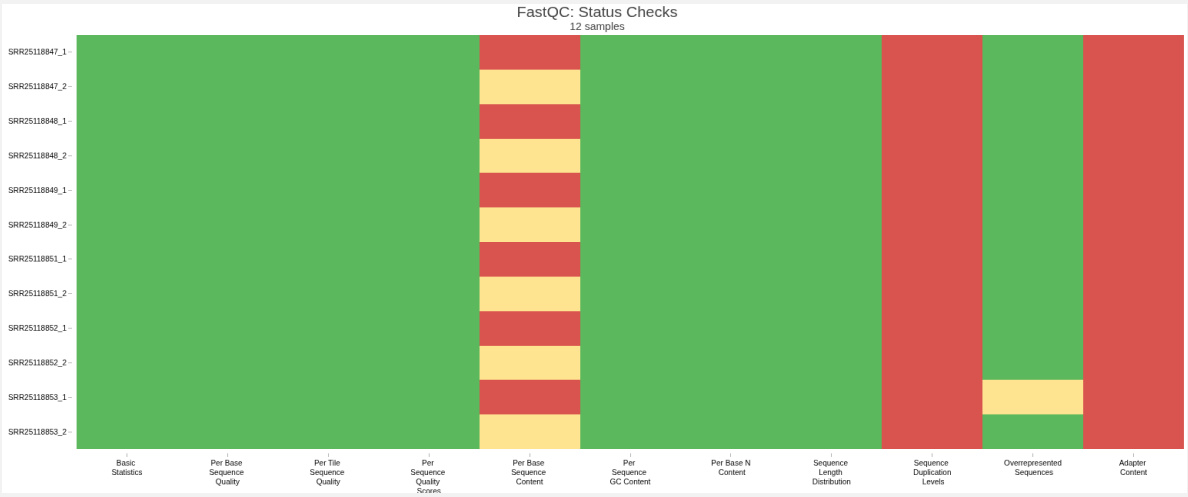


Figura 8: Resumen de los controles de calidad (Status Checks) según FastQC, mostrando fallos recurrentes en adaptadores y niveles de duplicación.

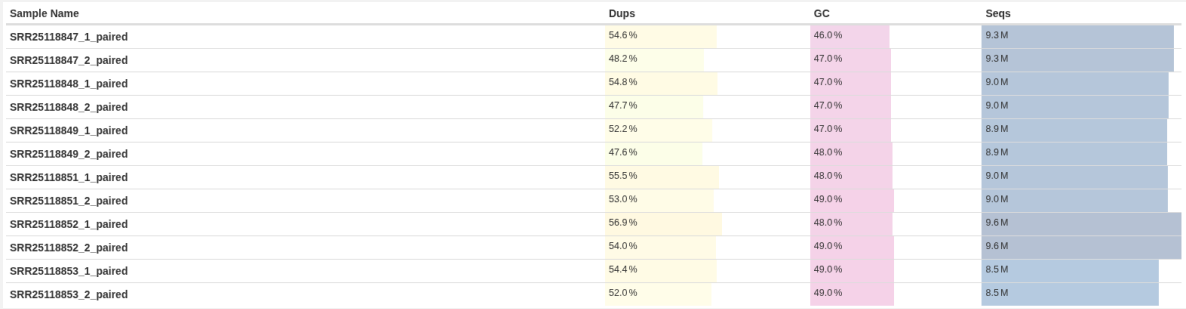
Tras aplicar el proceso de recorte con Trimmomatic, se conservaron entre el 69,82% y el 77,65% de las lecturas originales, lo que refleja una retención eficiente de secuencias de alta calidad. Una proporción relevante de las lecturas se mantuvo *single-end*, especialmente en las muestras SRR25118847 y SRR25118853, las cuales presentaron un mayor porcentaje de lecturas "Forward only". El porcentaje de lecturas descartadas fue bajo en todas las muestras (<2.5 %), lo que indica una buena calidad inicial de las secuencias. La Tabla 2 resume el número de pares de lecturas crudas, las lecturas conservadas emparejadas, uniparadas y descartadas para cada muestra.

Tabla 2: Resultados del recorte de secuencias con `trimmomatic`. Número de pares de lecturas crudas, lecturas emparejadas conservadas (*both surviving)*, uniparadas (*forward/reverse only*) y descartadas (*dropped)*, expresadas en valores absolutos y porcentuales por muestra.

Muestra	Lecturas crudas (pairs)	Both surviving (%)	Forward only (%)	Reverse only (%)	Dropped (%)
SRR25118847	12,830,179	9,257,272 (72,15 %)	3,115,800 (24,28 %)	160,735 (1,25 %)	296,372 (2,31 %)
SRR25118848	12,156,645	9,025,413 (74,24 %)	2,817,010 (23,17 %)	140,879 (1,16 %)	173,343 (1,43 %)

SRR25118849	11,669,727	8,927,397 (76,50 %)	2,383,795 (20,43 %)	149,438 (1,28 %)	209,097 (1,79 %)
SRR25118851	11,570,587	8,984,159 (77,65 %)	2,279,382 (19,70 %)	147,324 (1,27 %)	159,722 (1,38 %)
SRR25118852	12,518,735	9,627,853 (76,91 %)	2,581,807 (20,62 %)	151,411 (1,21 %)	157,664 (1,26 %)
SRR25118853	12,223,077	8,533,830 (69,82 %)	3,359,048 (27,48 %)	138,516 (1,13 %)	191,683 (1,57 %)

Posteriormente, se aplicó un nuevo control de calidad, evaluando la efectividad del trimming sobre parámetros clave de calidad. En general se observó una mejoría sustancial en la mayoría de las métricas evaluadas.
El contenido GC se mantuvo relativamente constante y dentro de los rangos esperados (~48–52 %), sin anomalías post-trimming (Figura 9). Esto indica que el recorte no introdujo sesgos sistemáticos relacionados con la composición de nucleótidos.



Sample Name	Dups	GC	Seqs
SRR25118847_1_paired	54.6 %	46.0 %	9.3 M
SRR25118847_2_paired	48.2 %	47.0 %	9.3 M
SRR25118848_1_paired	54.8 %	47.0 %	9.0 M
SRR25118848_2_paired	47.7 %	47.0 %	9.0 M
SRR25118849_1_paired	52.2 %	47.0 %	8.9 M
SRR25118849_2_paired	47.6 %	48.0 %	8.9 M
SRR25118851_1_paired	55.5 %	48.0 %	9.0 M
SRR25118851_2_paired	53.0 %	49.0 %	9.0 M
SRR25118852_1_paired	56.9 %	48.0 %	9.6 M
SRR25118852_2_paired	54.0 %	49.0 %	9.6 M
SRR25118853_1_paired	54.4 %	49.0 %	8.5 M
SRR25118853_2_paired	52.0 %	49.0 %	8.5 M

Figura 9: Tabla resumen de MultiQC posterior al trimming con porcentaje de duplicación, contenido en GC y número de lecturas por muestra.

En cuanto a la calidad por posición, las puntuaciones *Phred* se mantuvieron consistentemente por encima de Q30 en las lecturas, tanto en las cadenas forward como reverse. A diferencia del perfil pre-trimming, no se observó caída de calidad en los extremos 3', evidenciando la eliminación efectiva de regiones degradadas o con errores de secuenciación (Figura 10).

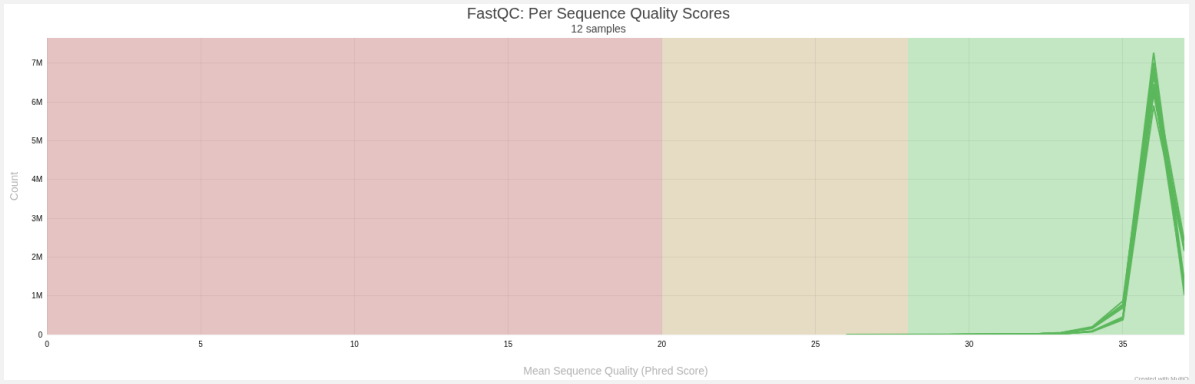


Figura 10: Calidad promedio por posición post-trimming, con valores Phred consistentemente superiores a Q30 (FastQC).

Asimismo, se detectó una estabilización del contenido en A, T, G y C en las secuencias (Figura 11). Mientras que en el análisis inicial se observaban fuertes desviaciones al inicio de la lectura (indicativas de sesgo de cebadores o fragmentos artefactuales), tras el trimming las curvas de las cuatro bases son prácticamente paralelas, especialmente en las primeras 50 posiciones.

Figura 11: Contenido de bases por posición tras el trimming, con estabilización en la composición de nucleótidos.

El trimming también produjo una disminución importante en la proporción de secuencias duplicadas. Mientras que en la etapa cruda varios datasets superaban el 60–70 % de duplicación (probablemente por sesgo de amplificación o baja complejidad), tras el recorte los valores cayeron por debajo del 40 %, acercándose al umbral de advertencia estándar (Figura 12).

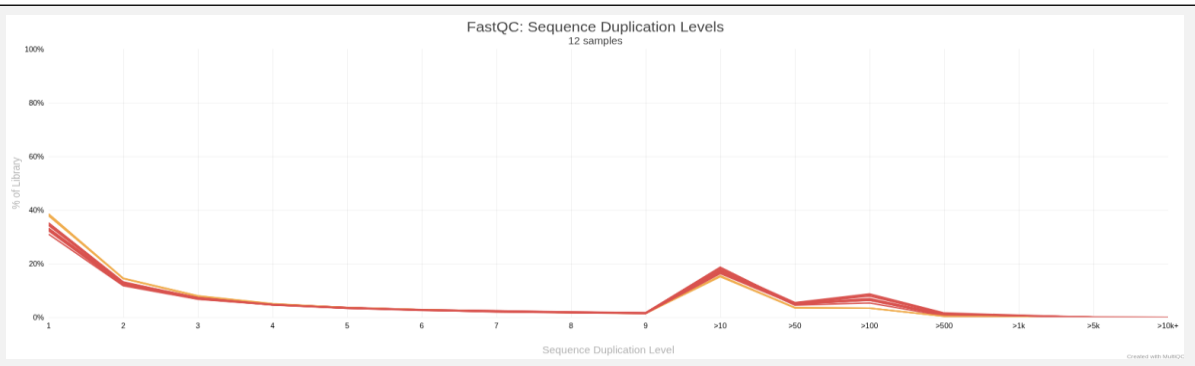


Figura 12: Distribución del porcentaje de secuencias duplicadas por muestra post-trimming

En lo referente al contenido de adaptadores, las gráficas muestran niveles de contaminación residuales, por debajo del 1% en todas las posiciones, respecto a los niveles superiores al 20% registrados en la etapa cruda. Finalmente, los resúmenes de "Status Checks" mostraron un patrón global de éxito (Figura 13), reflejando el impacto positivo del trimming y la idoneidad del dataset para el análisis transcriptómico.

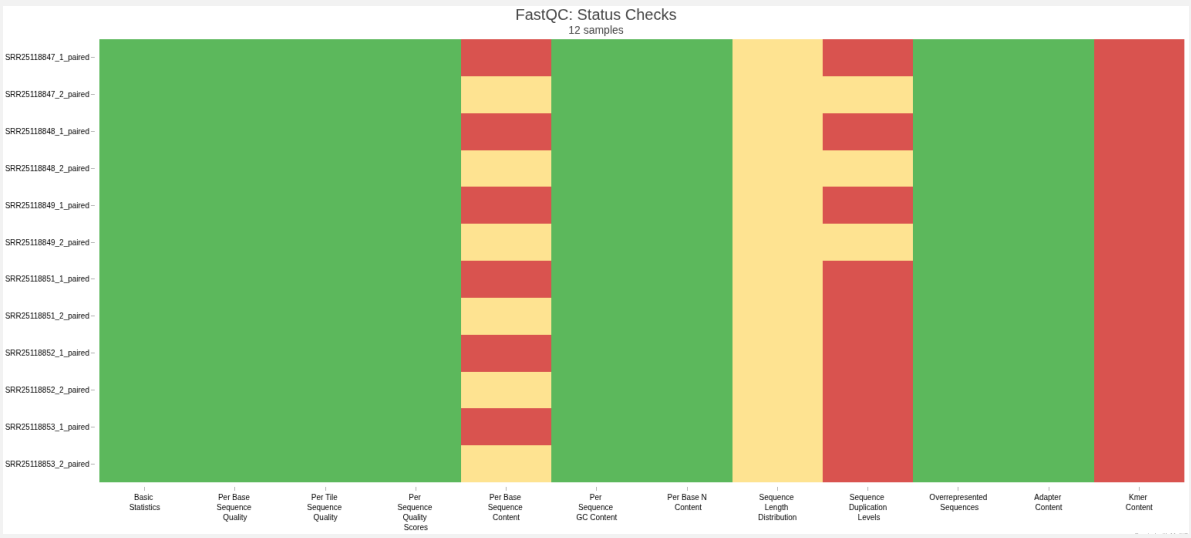


Figura 13: Resumen de controles de Calidad (*Status Checks)* tras el trimming, mostrando éxito global en todas las muestras.

Las secuencias emparejadas (paired-end) obtenidas tras el control de calidad y el preprocesamiento fueron alineadas frente al genoma de referencia utilizando el alineador HISAT2. Dado que el índice del genoma se encontraba precompilado, no fue necesario realizar un nuevo proceso de indexación, optimizando el tiempo computacional de análisis. Los resultados del proceso de alineamiento se resumen en la Tabla 3. El número total de muestras osciló entre las 8,5 y 9,6 millones. En todos los casos, más del 99,4% de las lecturas se alinearon correctamente con el genoma de referencia, evidenciando una excelente calidad del alineamiento. Concretamente, más del 91% de las lecturas emparejadas se alinearon concordantemente una sola vez, lo que refleja una tasa baja de ambigüedad en el mapeo. Por otra parte, entre los pares de lecturas que no lograron un alineamiento concordante, aproximadamente un 60% se alinearon de manera discordante, lo cual podría deberse a la presencia de translocaciones, reordenamientos o artefactos de amplificación. El resto de lecturas no alineadas representó un porcentaje residual (<0.6%) (Kim et al., 2019).

Tras el alineamiento, los archivos SAM generados fueron automáticamente ordenados y convertidos a formato BAM, sirviendo como base para el análisis de cuantificación.

Tabla 3: Estadísticas de alineamiento con HISAT2 por muestra. Se incluye el número total de lecturas emparejadas, porcentaje de alineamientos concordantes (únicos y múltiples), discordantes y tasa global de alineamiento.

Muestra	Lecturas totales	Alineamiento concordante 1 vez (%)	Alineamiento Concordante >1 veces (%)	Discordante (%)	Tasa global de alineamiento
sample_1.bam	9,257,272	93.65%	2.85%	62.92%	**99.49%**
sample_2.bam	9,025,413	93.41%	3.33%	63.03%	**99.53%**
sample_3.bam	8,927,397	93.59%	3.39%	61.33%	**99.54%**
sample_4.bam	8,984,159	93.00%	4.15%	59.41%	**99.54%**
sample_5.bam	9,627,853	92.67%	4.36%	59.73%	**99.54%**
sample_6.bam	8,533,830	91.70%	4.85%	60.03%	**99.47%**

La proporción de lecturas asignadas unívocamente osciló entre el 67.8% y el 75.3%; un rendimiento globalmente robusto en términos de asignación (Tabla 4), con un porcentaje de lecturas asignadas superior al 71% en la mayoría de las muestras. Las principales pérdidas se atribuyen a mapeo múltiple y ambigüedad; típicos en estudios transcriptómicos con abundancia de isoformas. La proporción de lecturas no alineadas fue mínima en todas las muestras (< 0.5%), confirmando la calidad del alineamiento obtenido con HISAT2

Tabla 4: Resultados del análisis de cuantificación. Porcentaje de lecturas asignadas a *features*, lecturas con mapeo múltiples, no asignadas por falta de anotación (*No Features*), ambigüedad y lecturas no alineadas.

Muestra	Asignadas (%)	Multi-mapping (%)	No Features (%)	Ambiguas (%)	No mapeadas (%)
Sample 1	75.34 %	7.30 %	3.32 %	9.04 %	0.46 %
Sample 2	73.84 %	9.28 %	3.18 %	8.98 %	0.42 %
Sample 3	72.95 %	9.47 %	2.96 %	9.10 %	0.40 %
Sample 4	71.80 %	12.01 %	2.34 %	9.81 %	0.40 %
Sample 5	72.41 %	13.11 %	2.45 %	9.87 %	0.41 %

Sample 6	67.84 %	14.61 %	2.16 %	9.51 %	0.45 %

Con el objetivo de evaluar la coherencia entre réplicas biológicas y explorar patrones de expresión globales, se realizaron análisis exploratorios sobre los datos transformados por estabilización de la varianza.
En el análisis de componentes principales (PCA, Figura 14), se observó una separación clara y robusta entre las condiciones de alta glucosa (HighGlucose) y normoglucosa (LowGlucose) a lo largo del primer componente PC1, que explica el 99% de la varianza. Estos resultados sugieren una profunda remodelación transcriptómica, inducida por la exposición a glucosa elevada.

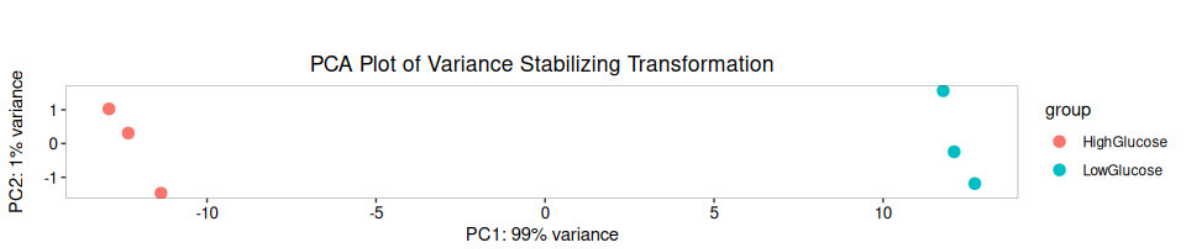


Figura 14: Análisis de componentes principales (PCA) sobre datos transformados mediante estabilización de varianza (VST). Se observa una separación clara entre las condiciones de alta glucusa y normoglucosa.

Este patrón fue corroborado por el agrupamiento jerárquico (Figura 15, a la izquierda), donde no se identificaron outliers y las réplicas biológicas por grupo se agruparon de forma coherente por condición. Además, las matrices de correlación por muestras (Figura 15, a la derecha), mostraron alta consistencia intragrupo.

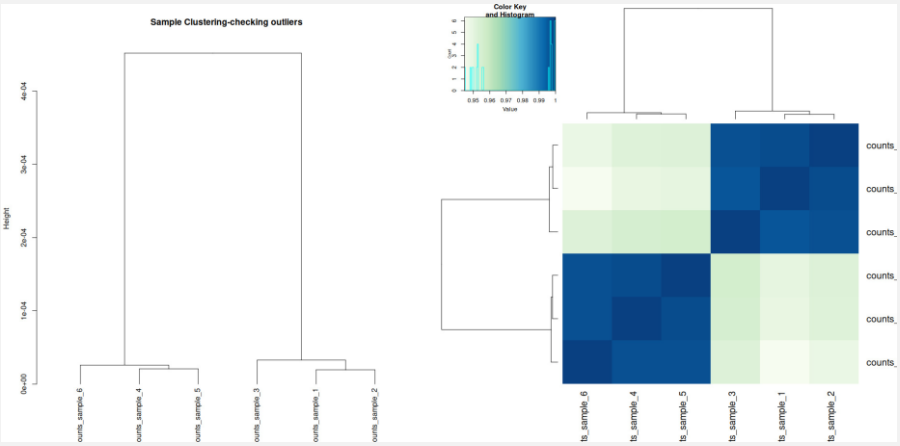


Figura 15: a la izquierda, Clustering jerárquico de las muestras biológicas con distancia euclídea y método de enlace completo. A la derecha, mapa de calor (*Heatmap)* de correlaciones por muestra. Las muestras se agrupan coherentemente por condición, mostrando alta consistencia intragrupo (correlación de Pearson entre datos VST)

El análisis con DESeq2 identificó un conjunto de genes con expresión diferencial significativa entre ambas condiciones. El Vulcano Plot (Figura 16, a la izquierda) resume estos resultados, mostrando numerosos genes sobreexpresados (en rojo) y subexpresados (en azul) en la condición hiperglucémica.
Para complementar esta visualización discreta de los genes diferencialmente expresados, se incluyó un MA plot (Figura 16, a la derecha), en el cual se representa el log2 del cambio de expresión (log2FC) frente a la media de cuentas normalizadas para cada gen. Este gráfico permite apreciar el patrón global de expresión diferencial, evidenciando cóm tanto genes altamente expresados como de baja abundancia presentan regulación significativa. El perfil simétrico en torno a la línea basal refuerza la hipótesis de una reprogramación molecular sustancial inducida por la glucosa elevada, sin sesgo hacia la sobreexpresión o represión.

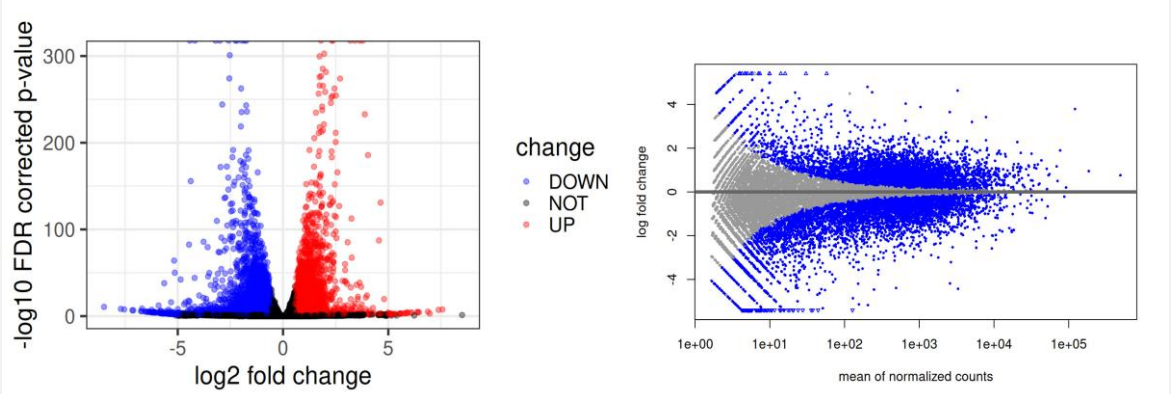


Figura 16: a la izquierda, Volcano plot del análisis de expresión diferencial (DESeq2). En rojo, genes sobreexpresados en HighGlucose; en azul, genes subexpresados; gris: genes no significativos. Criterios: $|\log_2 FC| > 1$ y padj < 0.05.
A la derecha, MA plot del análisis de expresión diferencial. Representación del log2 Fold Change frente a la media de cuentas normalizadas por gen. Genes significativos, en azul (padj<0.05); no significativos en gris.

A nivel funcional, la sobrerrepresentación funcional de los DEGs destacó procesos críticos vinculados a la agresividad tumoral:
Los genes infraexpresados en HighGlucose (Figura 17, izquierda) se asociaron con términos de control del ciclo celular (*cell cycle phase transition, checkpoint signaling, DNA-templated replication…*). Esta represión funcional sugiere una alteración en los mecanismos de vigilancia genómica, particularmente aquellos que limitan la proliferación y aseguran la integridad del genoma. Estos hallazgos son consistentes con el fenotipo DDR-deficient (deficiencia en la reparación del DNA) descrito en tumores de pacientes con diabetes mellitus tipo 2, en el cual se documente una supresión coordinada de genes implicados en recombinación homóloga (HR), unión de extremos NHEJ y reparación por desajustes (MMR), aumentando la susceptibilidad hacia inhibidores de PARP, ATR o CHK1 (Panigrahi et al., 2023).
Los genes sobreexpresados (Figura 17, derecha) mostraron enriquecimiento en biogénesis ribosomal (*rRNA processing, translation, ribosome assembly…*), y activación de señalización

de apoptótica intrínseca. Esto implica un aumento en la capacidad biosintética celular y carga proteica, junto a un estrés celular que puede desencadenar apoptosis compensatoria, relacionada consistentemente como signo de daño genotóxico.

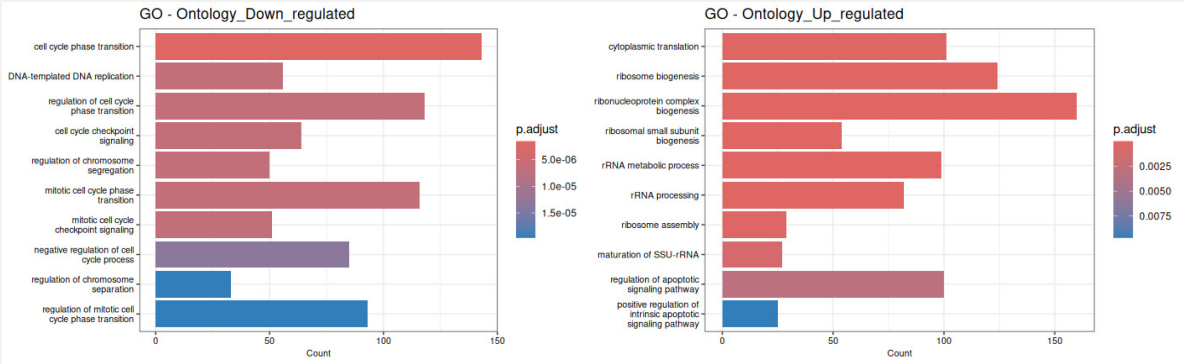


Figura 17: Análisis de enriquecimiento GO de genes diferencialmente expresados en HighGlucose. A la izquierda, genes subexpresados; a la derecha, genes sobreexpresados. El eje Y muestra términos biológicos en GO; el eje X indica el número de genes enriquecidos (Count). El gradiente de color representa el valor ajustado de p (padj).

El análisis de enriquecimiento funcional con conjuntos de genes HallMark proporcionó una visión sistémica de las rutas moleculares afectadas (Figura 18). Entre las firmas trasncriptómicas más significativamente activadas en la condición HighGlucose se encuentran los *MYC targets v1* y *v2*, *E2F targets* y *G2M Checkpoint*, todas ellas representativas de un estado de activación proliferativa sostenida. Las puntuaciones de *MYC targets v1 y v2* se correlacionan significativamente con la agresividad tumoral, un peor pronóstico y un aumento en la mutación genómica y carga proliferativa en cáncer de mama ER-positivo (Schulze et al., 2020). Por tanto, la coactivación en nuestro modelo experimental de glucosa alta sugiere que la hiperglucemia no solo induce cambios en el metabolismo energético y biosintético, sino que activa rutas implicadas en la progesteración tumoral y la invasividad.
Los factores E2F impulsan la entrada en fase S, reforzando un fenotipo celular altamente replicativo, con potencial transformante y riesgo de inestabilidad cromosómica (Hernando et al. 2004).

Finalmente, la activación de la ruta *p53 pathway* sugiere un intento compensatorio de respuesta al daño genómico, posiblemente inducido por el daño oxidativo y replicativo . En este contexto, si esta compensación falla (como ocurre en tumores con deficiencia en reparación de DNA) puede favorecer un fenotipo de progresión agresiva (Moon et al., 2019). Además, la interrelación funcional de p53 con rutas como la *Oxidative phosphorylation*, la *Unfolded Protein Response* y la regulación de la traducción, sugiere que la vía p53 no solo estaría actuando como indicador de daño, sino también como modulador central del estado metabólico y la integridad proteica.

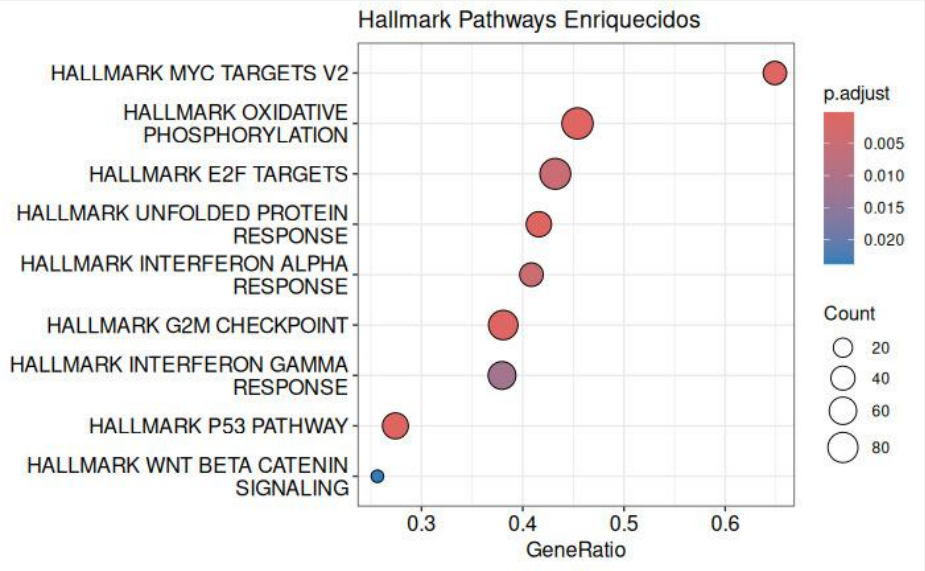


Figura 18: Análisis de enriquecimiento funcional con conjuntos HallMark (MsigDB) a partir de genes diferencialmente expresados (DEGs). El eje Y muestra las rutas enriquecidas; el eje X representa el GeneRatio (proporción de genes por término). El tamaño de los puntos indica el número de genes por categoría, y el gradiente de color representa el valor ajustado de p (padj).

Referencias

1. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

2. Autier, P., Boniol, M., LaVecchia, C., Vatten, L., Gavin, A., Héry, C., & Heanue, M. (2010). Disparities in breast cancer mortality trends between 30 European countries: Retrospective trend analysis of WHO mortality database. BMJ (Online), 341(7768), 335. https://doi.org/10.1136/bmj.c3620

3. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

4. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

5. Giovannucci, E., Harlan, D. M., Archer, M. C., Bergenstal, R. M., Gapstur, S. M., Habel, L. A., … Yee, D. (2010). Diabetes and Cancer: A Consensus Report. CA: A Cancer Journal for Clinicians, 60(4), 207–221. https://doi.org/10.3322/caac.20078

6. Hernando, E., Nahlé, Z., Juan, G., Diaz-Rodriguez, E., Alaminos, M., Hemann, M., … Cordon-Cardo, C. (2004). Rb inactivation promotes genomic instability by uncoupling cell cycle progression from mitotic control. Nature, 430(7001), 797–802. https://doi.org/10.1038/nature02820

7. Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

8. Lei, P., Wang, W., Sheldon, M., Sun, Y., Yao, F., & Ma, L. (2023, July 1). Role of Glucose Metabolic Reprogramming in Breast Cancer Progression and Drug Resistance. Cancers. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/cancers15133390

9. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

10. Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

11. Lord, C. J., & Ashworth, A. (2017). PARP inhibitors: Synthetic lethality in the clinic. *Science*, *355*(6330), 1152–1158. https://doi.org/10.1126/science.aam7344

12. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12). https://doi.org/10.1186/s13059-014-0550-8

13. Masoud, R., & Pagès, G. (2021). Targeting the tumor microenvironment of triple-negative breast cancer. *Cancers*, *13*(3), 411. https://doi.org/10.3390/cancers13030411

14. Moon, S. H., Huang, C. H., Houlihan, S. L., Regunath, K., Freed-Pastor, W. A., Morris, J. P., … Prives, C. (2019). p53 Represses the Mevalonate Pathway to Mediate Tumor Suppression. Cell, 176(3), 564-580.e19. https://doi.org/10.1016/j.cell.2018.11.011

15. Panigrahi, G., Candia, J., Dorsey, T. H., et al. (2023). Diabetes-associated breast cancer is molecularly distinct and shows a DNA damage repair deficiency. *JCI Insight*, *8*(23), e170105. https://doi.org/10.1172/jci.insight.170105

16. Perou, C. M., Sørlie, T., Eisen, M. B., et al. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–752. https://doi.org/10.1038/35021093

17. Schulze, A., Oshi, M., Endo, I., & Takabe, K. (2020). Myc targets scores are associated with cancer aggressiveness and poor survival in er-positive primary and metastatic breast cancer. International Journal of Molecular Sciences, 21(21), 1–13. https://doi.org/10.3390/ijms21218127

18. Wang, Y. Y., Lehuédé, C., Laurent, V., Dirat, B., Dauvillier, S., Bochet, L., Le Gonidec, S., Escourrou, G., Valet, P., & Muller, C. (2012). Adipose tissue and breast epithelial cells: a dangerous dynamic duo in breast cancer. Cancer letters, 324(2), 142–151. https://doi.org/10.1016/j.canlet.2012.05.019

19. Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag. https://ggplot2.tidyverse.org

20. World Health Organization. (2024). *Breast cancer: Key facts*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

21. Yersal, O., & Barutca, S. (2014, August 10). Biological subtypes of breast cancer: Prognostic and therapeutic implications. World Journal of Clinical Oncology. Baishideng Publishing Group Co., Limited. https://doi.org/10.5306/wjco.v5.i3.412

22. Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. OMICS A Journal of Integrative Biology, 16(5), 284–287. https://doi.org/10.1089/omi.2011.0118

23. Zhao, Y., Luo, Y., Wang, X., et al. (2022). Metabolic regulation of cancer immunotherapy. *Nat Rev Immunol*, *22*(9), 540–558. https://doi.org/10.1038/s41577-022-00731-5

24. Qiagen. (n.d.). RNeasy Plus Kits: Total RNA purification. https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/rna-purification/total-rna/rneasy-plus-kits