Facultad de Ingeniería
UNER

Bioinformática

Análisis de Secuencias de Nuevas Tecnologías de Secuenciación en Paralelo

Trabajo Práctico Final: Análisis transcriptómico de pupas de *Drosophila melanogaster* crecidas en hipergravedad

Profesora:

Dra. Ileana Tossolini

Alumno:

Alejandro Escobar

Fecha de entrega:

5/06/2025

Índice

Introducción

El presente Trabajo Práctico corresponde a la instancia evaluativa final del Seminario. Para la realización del mismo, se brindarán datos reales y se requerirá la selección de herramientas para su análisis apropiado. A partir de los resultados obtenidos se elaborará un artículo científico (para ello se le brindará un documento modelo) y se realizará la presentación oral de los mismos en el congreso "Bioinformágica" a realizarse el día 6 de junio de 2025.

Problema biológico

El objetivo de este trabajo es estudiar el transcriptoma e investigar los cambios en los niveles de expresión génica de la mosca de la fruta, particularmente de pupas de *Drosophila melanogaster* crecidas en condiciones de hipergravedad. La gravedad alterada puede perturbar el desarrollo normal e inducir cambios correspondientes en la expresión génica. Comprender esta relación entre el entorno físico y la respuesta biológica es importante para los objetivos de viajes espaciales de la NASA.

Durante la etapa pupal, la mayoría de los insectos holometábolos experimentan una metamorfosis extensa de larva a adulto alado. Esta metamorfosis está rigurosamente regulada a nivel transcripcional y se conserva evolutivamente. También depende de la gravedad terrestre. Por ejemplo, la orientación de la pupa se basa en señales gravitatorias para una alineación adecuada, y la exposición a entornos de microgravedad o hipergravedad puede influir de manera profunda en la metamorfosis y alterar la expresión génica. El experimento consta de dos condiciones: **g3**, en la que las pupas se desarrollaron a tres veces la gravedad terrestre (3 g), y **g1**, el control, en la que se desarrollaron a la gravedad estándar de la superficie terrestre (1 g).

El set de datos está compuesto por 12 muestras, dos tratamientos con tres réplicas biológicas cada uno. Los datos son de tipo *paired-end*. Debido a que analizar los archivos originales llevaría mucho tiempo de procesamiento, se le brinda un set de datos reducido que incluye sólo las lecturas que alinean contra el cromosoma 4 de la mosca de la fruta (*Drosophila melanogaster*). Los datos de RNA-Seq completos están disponibles en la base de datos GEO de NCBI, bajo el número de acceso GSE80323.

A continuación se presenta la siguiente tabla con información (metadatos) de las muestras.

Tabla 1: Información de las muestras.

Muestra	Condición	Experimento	Réplica	SRA Accession	Protocolo preparación biblioteca	Total Read length	Equipo de secuenciación	Stranded?
g1_01	Control	RNA-seq	1	SRX1707053	Illumina mRNA TruSeq kit	50000	Illumina HiSeq 2000 100bp PE reads	No
g1_02	Control	RNA-seq	2	SRX1707048	Illumina mRNA TruSeq kit	50000	Illumina HiSeq-2000 100bp PE reads	No
g1_03	Control	RNA-seq	3	SRX1707049	Illumina mRNA TruSeq kit	50000	Illumina HiSeq-2000 100bp PE reads	No
g3_01	Tratamiento	RNA-seq	1	SRX1707051	Illumina mRNA TruSeq kit	50000	Illumina HiSeq-2000 100bp PE reads	No
g3_02	Tratamiento	RNA-seq	2	SRX1707052	Illumina mRNA TruSeq kit	50000	Illumina HiSeq-2000 100bp PE reads	No
g3_03	Tratamiento	RNA-seq	3	SRX1707047	Illumina mRNA TruSeq kit	50000	Illumina HiSeq-2000 100bp PE reads	No

Contamos las reads de cada uno con el comando:

```
for fastq in Datos_RNA_seq_mosca/*.fastq
do
  echo $fastq
  cat $fastq | awk 'BEGIN{i=0}{i++;}END{print i/4}'
done
```

Contamos las longitud de las secuencias con el comando:

```
for fastq in Datos_RNA_seq_mosca/*.fastq
do
  echo "$fastq"
  cat "$fastq" | paste - - - - | awk '{print length($2)}'
```

```
done
```

Se observa que todas tienen una longitud de 100 pb.

```
Datos_RNA_seq_mosca/g1_01_R1.fastq
50000
Datos_RNA_seq_mosca/g1_01_R2.fastq
50000
Datos_RNA_seq_mosca/g1_02_R1.fastq
50000
Datos_RNA_seq_mosca/g1_02_R2.fastq
50000
Datos_RNA_seq_mosca/g1_03_R1.fastq
50000
Datos_RNA_seq_mosca/g1_03_R2.fastq
50000
Datos_RNA_seq_mosca/g3_01_R1.fastq
50000
Datos_RNA_seq_mosca/g3_01_R2.fastq
50000
Datos_RNA_seq_mosca/g3_02_R1.fastq
50000
Datos_RNA_seq_mosca/g3_02_R2.fastq
50000
Datos_RNA_seq_mosca/g3_03_R1.fastq
50000
Datos_RNA_seq_mosca/g3_03_R2.fastq
50000
```

Control de Calidad

Se analizan los datos mediante FastQC y MultiQC. Los resultados se encuentran en el directorio: */reportes_datos_crudos*. Se emplearon los comandos:

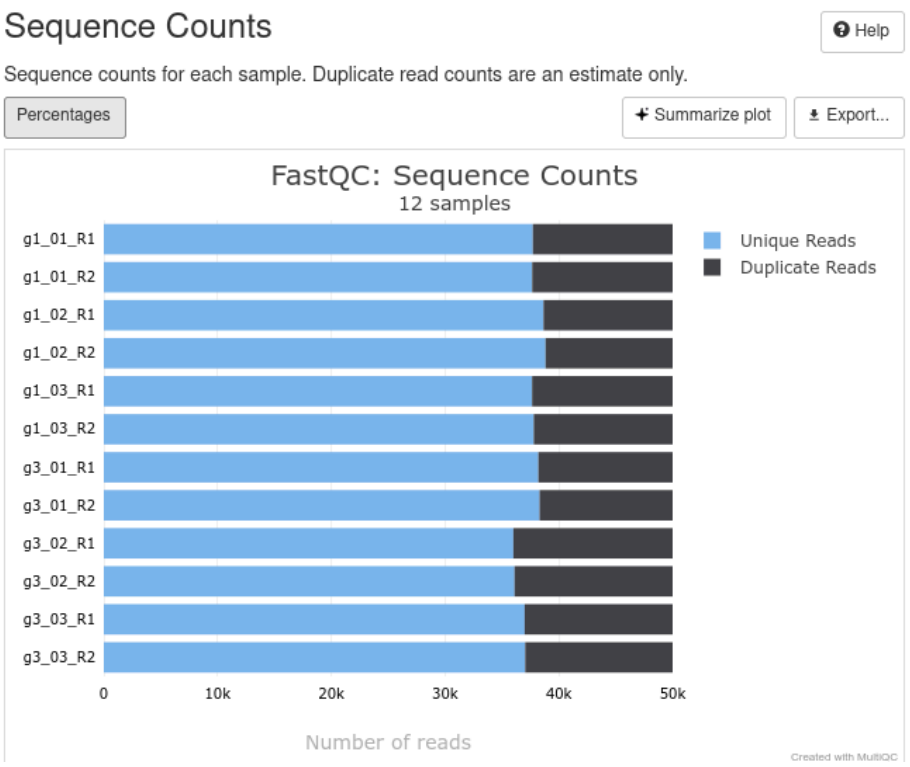
```
fastqc Datos_RNA_seq_mosca/*.fastq -o reportes_datos_crudos
multiqc --data-dir reportes_datos_crudos -o reportes_datos_crudos
```

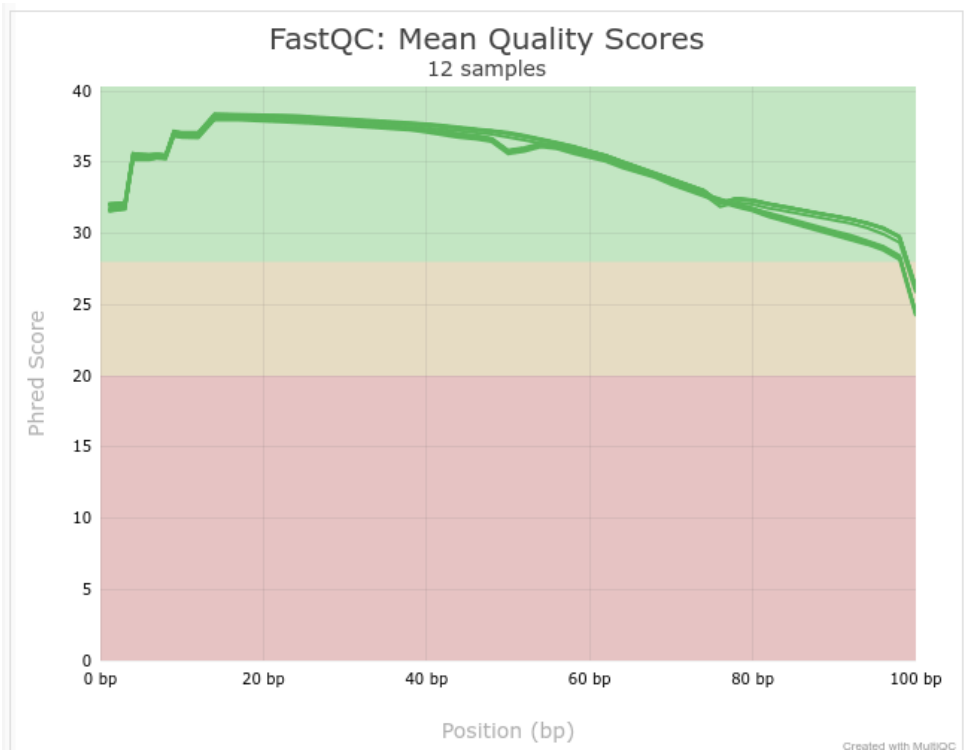
General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 12/12 rows and 3/6 columns. | Summarize table

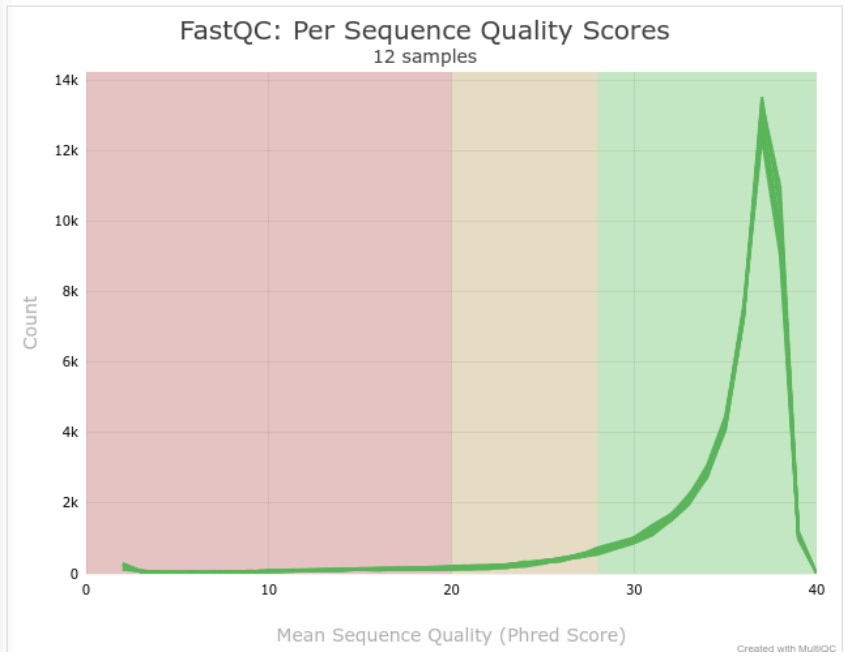
Sample Name	Dups	GC	Seqs
g1_01_R1	24.6%	41.0%	0.0 M
g1_01_R2	24.7%	41.0%	0.0 M
g1_02_R1	22.7%	41.0%	0.0 M
g1_02_R2	22.3%	41.0%	0.0 M
g1_03_R1	24.7%	41.0%	0.0 M
g1_03_R2	24.4%	41.0%	0.0 M
g3_01_R1	23.6%	41.0%	0.0 M
g3_01_R2	23.4%	41.0%	0.0 M
g3_02_R1	28.0%	40.0%	0.0 M
g3_02_R2	27.8%	40.0%	0.0 M
g3_03_R1	26.0%	40.0%	0.0 M
g3_03_R2	25.9%	40.0%	0.0 M

- El contenido de GC es del 41 o 42%, lo cuál se condice con lo reportado por el genoma del NCBI de 42%.





- Se observa que la calidad en la parte final de las lecturas (últimos 10-15 bp) tiende a disminuir un poco.



Per Base Sequence Content

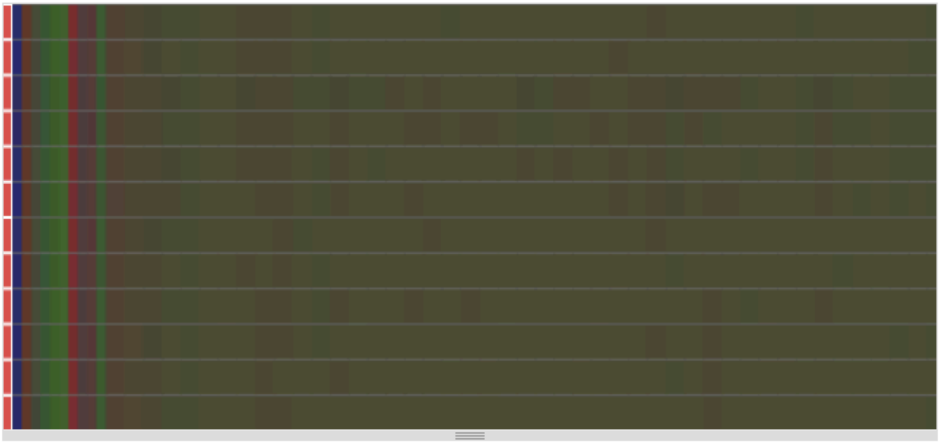
Help

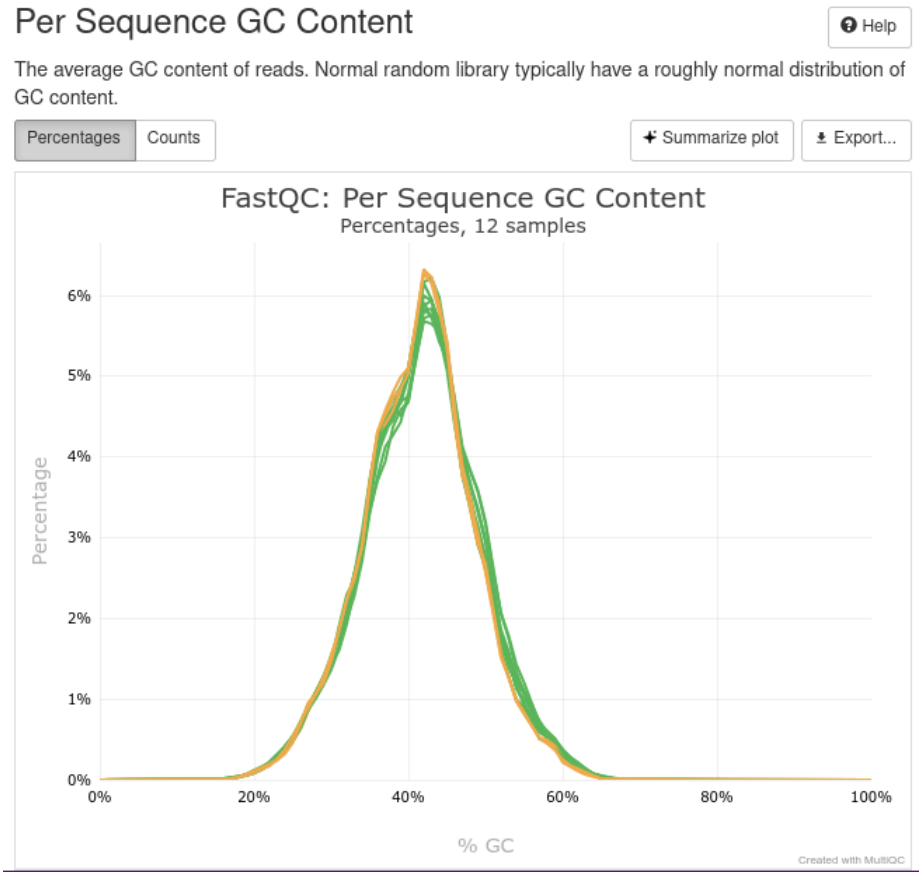
The proportion of each base position for which each of the four normal DNA bases has been called.

Click a sample row to see a line plot for that dataset.

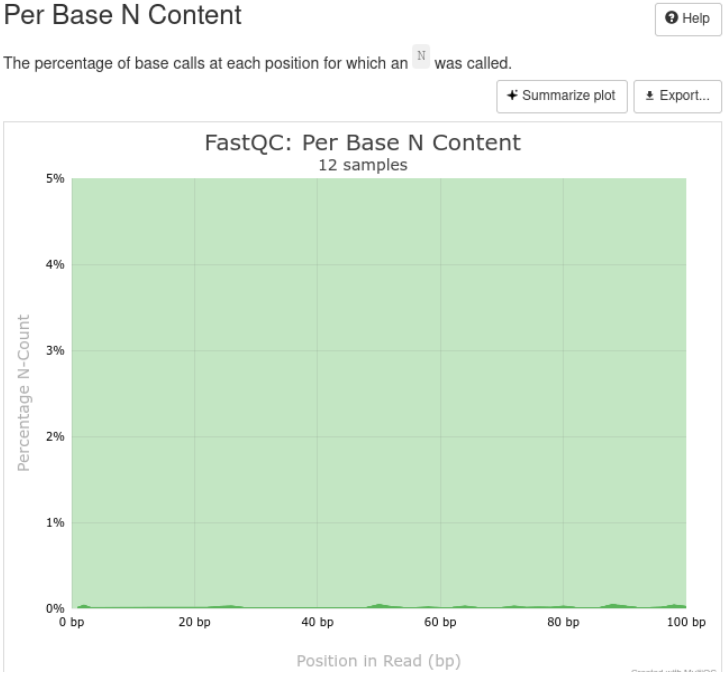
Rollover for sample name

Position: - %T: - %C: - %A: - %G: -





- Me llama la atención la distribución muy “picuda” o sesgada, puede indicar alguna contaminación o puede deberse a la presencia de los adaptadores.



Sequence Length Distribution

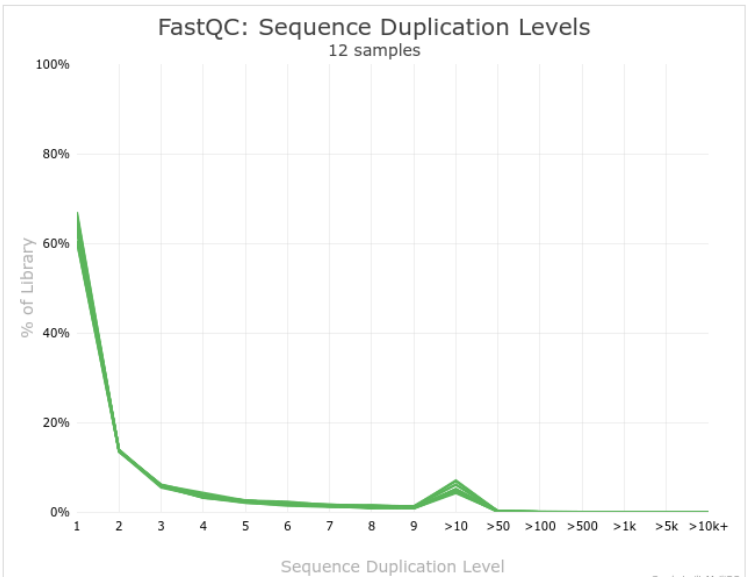
All samples have sequences of a single length (100bp)

Sequence Duplication Levels

Help

The relative level of duplication found for every sequence.

Summarize plot | Export...



- Los niveles altos de duplicación al principio puede ser normal en un análisis de RNA-seq por la alta expresión de algunos genes en el tratamiento.
- Estoy secuenciando muchas veces un mismo CDNA por la duplicación de los transcriptos, por eso es esperable que quede casi igual después de los trimmings.

Overrepresented sequences by sample

The total amount of overrepresented sequences found in each library.

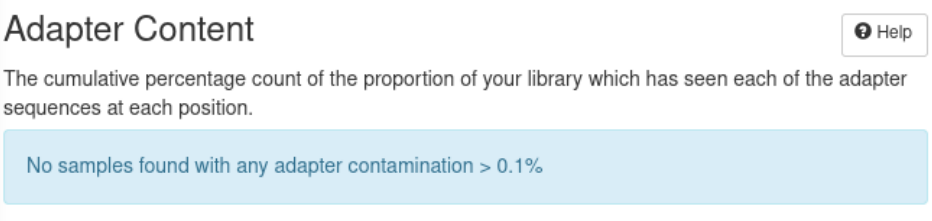
12 samples had less than 1% of reads made up of overrepresented sequences

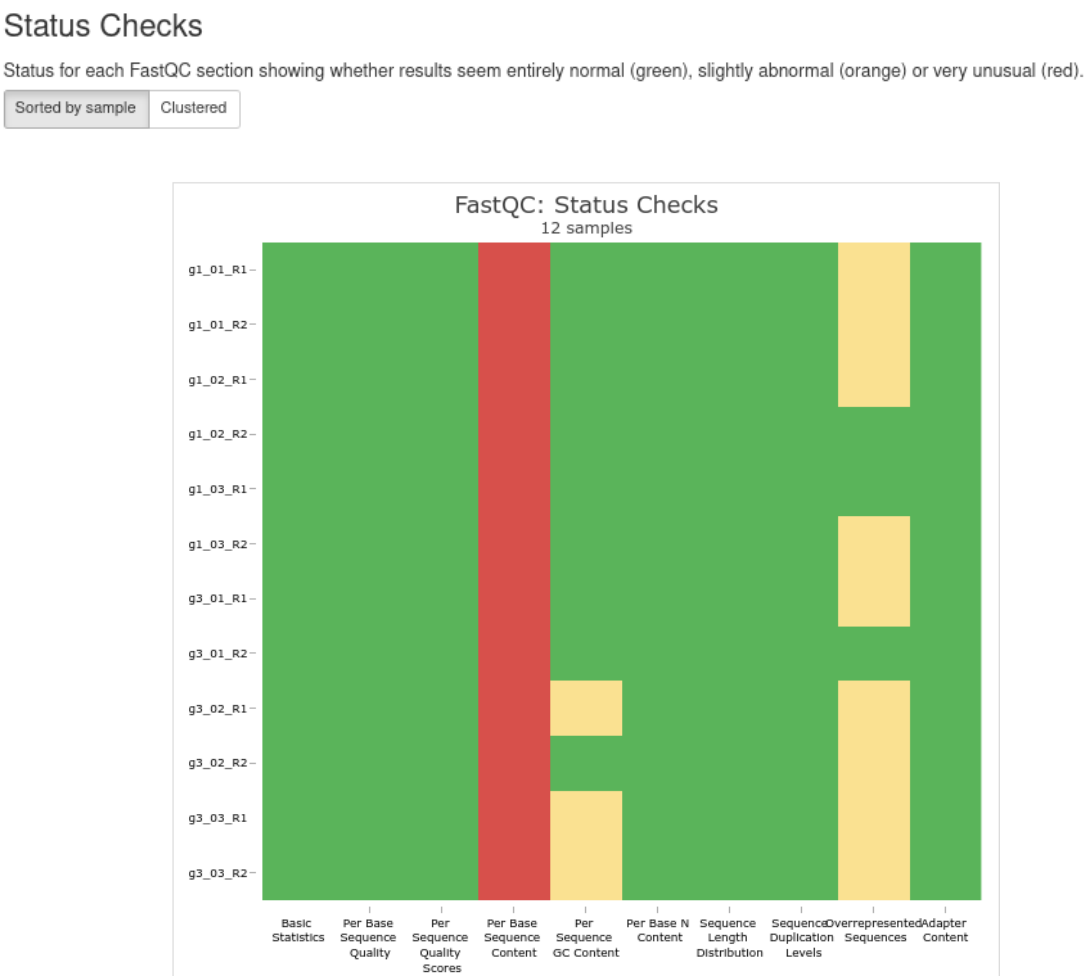
Top overrepresented sequences

Top overrepresented sequences across all samples. The table shows 20 most overrepresented sequences across all samples, ranked by the number of samples they occur in.

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 3/3 rows and 3/3 columns. | Summarize table

Overrepresented sequence	Reports	Occurrences	% of all reads
GTTGAACATGGCAGTCGGCAAAAATAAAGGTCTTTCCAAGGGTGGTAAGA	9	511	0.0852 %
CTCGAATCCAAGGTAATGAAATTGAAAGCCAGAGTCGATTCAATACCAAA	1	57	0.0095 %
TGAACATGGCAGTCGGCAAAAATAAAGGTCTTTCCAAGGGTGGTAAGAAG	1	53	0.0088 %





Se realizó un trimming de las lecturas mediante trimmomatic (v0.39) mediante el script de bash "trimming-trimmomatic.sh".

```
#!/bin/bash

# Crear carpeta de salida en el directorio anterior
mkdir -p ../trimming-trimmomatic

# Lista de muestras
```

```
samples=("g1_01" "g1_02" "g1_03" "g3_01" "g3_02" "g3_03")

# Bucle para procesar cada muestra
for sample in "${samples[@]}"
do
  echo "Procesando $sample..."

  trimmomatic PE -threads 16 \
    ${sample}_R1.fastq ${sample}_R2.fastq \
    ../trimming-trimmomatic/${sample}_R1_paired.fastq
../trimming-trimmomatic/${sample}_R1_unpaired.fastq \
    ../trimming-trimmomatic/${sample}_R2_paired.fastq
../trimming-trimmomatic/${sample}_R2_unpaired.fastq \
    ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 \
    LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:36

  echo "$sample procesado."
done

echo "Trimming finalizado. Archivos guardados en ../trimming-trimmomatic"
```

En donde los parámetros significan:

- **PE**: modo paired-end.
- **-threads 16**: usa 16 hilos para acelerar el proceso.
- **ILLUMINACLIP**: recorte de adaptadores (OJO: se debe copiar el archivo de los adaptadores *TruSeq3-PE.fa* en la carpeta).
    - **2**: número máximo de desajustes al buscar el adaptador.
    - **30**: puntuación mínima para mantener un emparejamiento.
    - **10**: número de bases para verificar en el palíndromo.
- **LEADING:3**: recorta bases del inicio si tienen calidad < 3.
- **TRAILING:3**: recorta del final si tienen calidad < 3.
- **SLIDINGWINDOW:4:20:** recorta cuando la media de una ventana de 4 bases cae por debajo de 20.
- **MINLEN:36**: descarta cualquier fragmento < 36 pb.

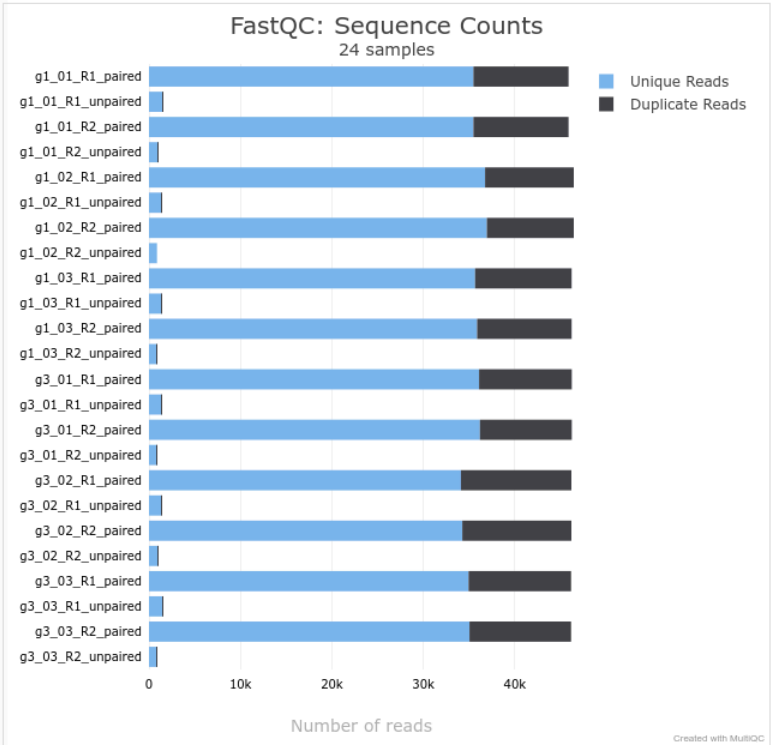
Se volvió a analizar la calidad de los datos mediante FastQC y MultiQC. Los resultados se encuentran en el directorio: */reportes_datos_trimming/trimmomatic*. Se emplearon los comandos:

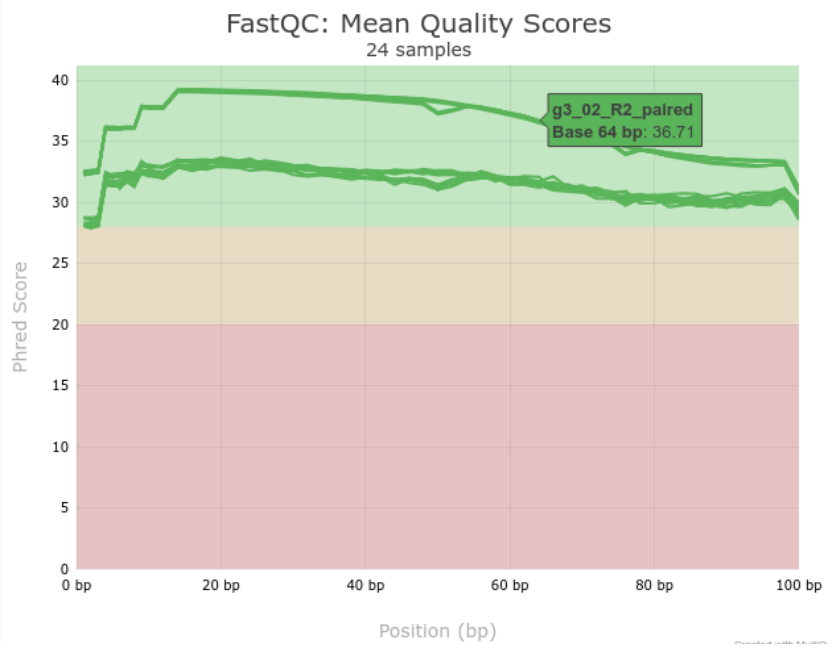
```
fastqc trimming-trimmomatic/*.fastq -o reportes_datos_trimming/trimommatic
multiqc --data-dir reportes_datos_trimming/trimommatic -o
```

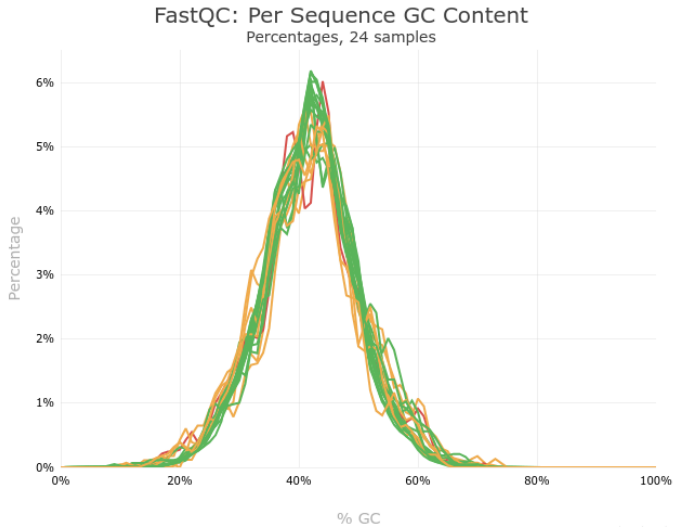
reportes_datos_trimming/trimommatic

Se muestra el resultado de este trimming:

Sample Name	Dups	GC	Median len	Seqs
g1_01_R1_paired	22.7 %	40.0 %	100 bp	0.0 M
g1_01_R1_unpaired	0.9 %	42.0 %	72 bp	0.0 M
g1_01_R2_paired	22.7 %	40.0 %	100 bp	0.0 M
g1_01_R2_unpaired	0.4 %	40.0 %	68 bp	0.0 M
g1_02_R1_paired	20.9 %	41.0 %	100 bp	0.0 M
g1_02_R1_unpaired	0.3 %	42.0 %	70 bp	0.0 M
g1_02_R2_paired	20.5 %	41.0 %	100 bp	0.0 M
g1_02_R2_unpaired	0.1 %	41.0 %	68 bp	0.0 M
g1_03_R1_paired	22.8 %	41.0 %	100 bp	0.0 M
g1_03_R1_unpaired	0.6 %	41.0 %	72 bp	0.0 M
g1_03_R2_paired	22.4 %	41.0 %	100 bp	0.0 M
g1_03_R2_unpaired	0.3 %	40.0 %	66 bp	0.0 M







Como la calidad de los datos no mejora significativamente, y las lecturas quedan bastante recortadas, se realizó otro trimming de las lecturas mediante skewer (v0.2.2) mediante el script de bash "trimming-skewer.sh".

```
#!/bin/bash

# Crear carpeta de salida en el directorio anterior
mkdir -p ../trimming-skewer

# Lista de muestras
samples=("g1_01" "g1_02" "g1_03" "g3_01" "g3_02" "g3_03")
```

```
# Bucle para procesar cada muestra
for sample in "${samples[@]}"
do
  echo "Procesando $sample..."

    skewer -m pe -q 20 -l 36 -t 16 \
    -o ../trimming-skewer/${sample} \
    ${sample}_R1.fastq ${sample}_R2.fastq

  echo "$sample procesado."
done

echo "Trimming finalizado. Archivos guardados en ../trimming-skewer"
```

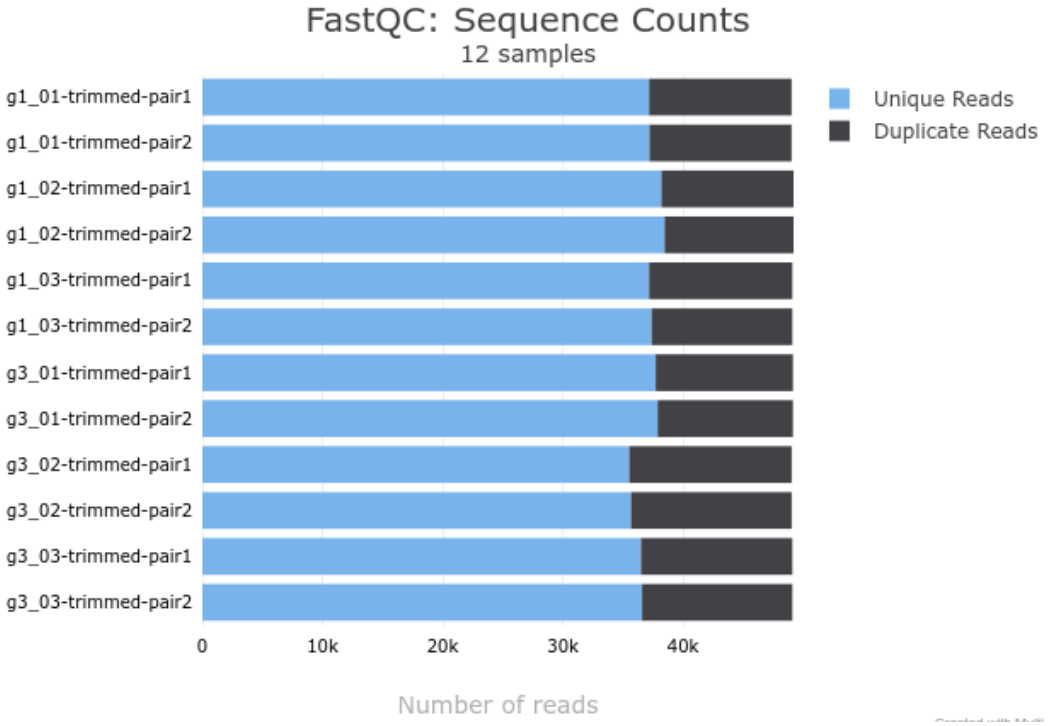
En donde los parámetros significan:

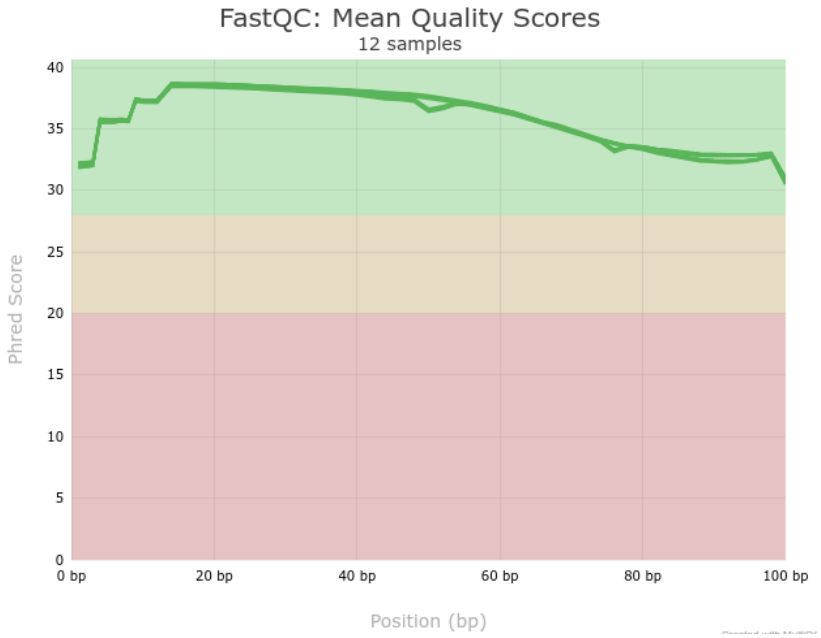
- **-m pe:** modo paired-end, indica que estás procesando datos emparejados (dos archivos: R1 y R2).
- **-q 20**: calidad mínima por base, elimina bases con calidad Phred < 20 desde los extremos de las lecturas. (Phred 20 = 1% de error esperado).
- **-l 36:** longitud mínima, descarta completamente las lecturas si, después del recorte, quedan con menos de 36 pb.
- **-t 16:** número de hilos (threads), usa 16 núcleos del procesador para paralelizar el análisis (mi compu tiene muchos hilos por ser un AMD Ryzen7 5700U).
- **-o trimming-skewer/${sample}** : prefijo de salida, el nombre base de los archivos de salida.
- **${sample}_R1.fastq ${sample}_R2.fastq:** archivos de entrada, son los archivos paired-end originales, lectura 1 y lectura 2.

Se volvió a analizar la calidad de los datos mediante FastQC y MultiQC. Los resultados se encuentran en el directorio: */reportes_datos_trimming/skewer*. Se emplearon los comandos:

```
fastqc trimming-skewer/*.fastq -o reportes_datos_trimming/skewer
multiqc --data-dir reportes_datos_trimming/skewer -o reportes_datos_trimming/skewer
```

Sample Name	Dups	GC	Seqs
g1_01-trimmed-pair1	24.2 %	40.0 %	0.0 M
g1_01-trimmed-pair2	24.1 %	40.0 %	0.0 M
g1_02-trimmed-pair1	22.3 %	41.0 %	0.0 M
g1_02-trimmed-pair2	21.8 %	41.0 %	0.0 M
g1_03-trimmed-pair1	24.3 %	41.0 %	0.0 M
g1_03-trimmed-pair2	23.8 %	41.0 %	0.0 M
g3_01-trimmed-pair1	23.2 %	41.0 %	0.0 M
g3_01-trimmed-pair2	22.9 %	40.0 %	0.0 M
g3_02-trimmed-pair1	27.6 %	40.0 %	0.0 M
g3_02-trimmed-pair2	27.3 %	40.0 %	0.0 M
g3_03-trimmed-pair1	25.6 %	40.0 %	0.0 M
g3_03-trimmed-pair2	25.5 %	40.0 %	0.0 M





Debido a que son lecturas muy cortas, y el trimmeado deja muchos desapareados en relación a la totalidad de la secuencia cuando se usa trimmomatic, y debido a que con skewer no se mejora en gran medida la calidad, se continuarán trabajando con los datos sin filtrado adicional.

Para proceso de mapeo se utilizará STAR ya que es un software "splice-aware" que permite mapear uniones de empalme descritas en la anotación o detectar nuevas.

Primero calculamos el tamaño del genoma del cromosoma 4 para pasar como parámetro a STAR. Para ello se empleó el comando de **Bioawk** (v20110810):

```
bioawk -c fastx '{print $name, length($seq)}' chr4.fa
```

```
…NGS\TP_Final\TP_Final_RNAseq_mosca_2025\Datos_RNA_seq_mosca\genome
 ⚡alejandro ›› bioawk -c fastx '{print $name, length($seq)}' chr4.fa
chr4    1351857
```

Como es un genoma chico, se calcula el parámetro de acuerdo al manual de STAR.

- --genomeSAindexNbases
  - default: 14
  - int: length (bases) of the SA pre-indexing string. Typically between 10 and 15. Longer strings will use much more memory, but allow faster searches. For small genomes, the parameter –genomeSAindexNbases must be scaled down to **min(14, log2(GenomeLength)/2 - 1)**.

Nos da un valor de 9.18.

Se utiliza el siguiente comando para crear el índice dentro de la carpeta */Mapping*

```
STAR --runMode genomeGenerate \
  --genomeDir Index_STAR_chr4 \
  --genomeFastaFiles ../Datos_RNA_seq_mosca/genome/chr4.fa \
  --sjdbGTFfile ../Datos_RNA_seq_mosca/genome/ensembl_dm3.chr4.gtf \
  --sjdbOverhang 99 \
  --runThreadN 16 \
  --genomeSAindexNbases 9.18 \
  --outFileNamePrefix dm3_chr4
```

- Se coloca un valor de overhang de 99 ya que es el tamaño de lectura (100) menos 1.

Ahora si generamos el mapeo teniendo en cuenta que se tratan de lecturas pareadas, entonces debemos pasarle primero la lectura *forward* y después la *reverse*, con el comando:

```
for sample in g1_01 g1_02 g1_03 g3_01 g3_02 g3_03 \
do
  echo "Mapeando muestra: $sample"

  STAR --runThreadN 16 \
    --genomeDir Index_STAR_chr4 \
    --readFilesIn ../Datos_RNA_seq_mosca/${sample}_R1.fastq
../Datos_RNA_seq_mosca/${sample}_R2.fastq \
    --outSAMtype BAM SortedByCoordinate \
    --quantMode GeneCounts \
    --outSAMstrandField intronMotif \
    --outFileNamePrefix alignments_STAR/${sample}_ \
done
```

Control de calidad de los mapeos

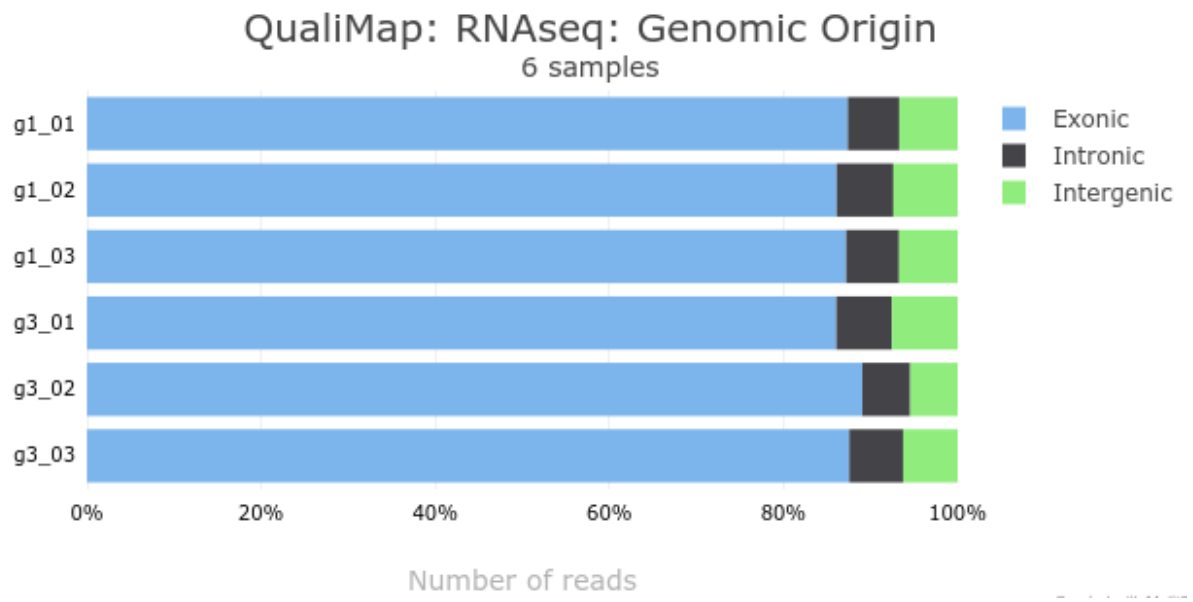
Se empleará quialimap con el comando:

```
for sample in g1_01 g1_02 g1_03 g3_01 g3_02 g3_03
do
  echo "Corriendo Qualimap para $sample..."
  qualimap rnaseq \
    -bam alignments_STAR/${sample}_Aligned.sortedByCoord.out.bam \
    -gtf ../Datos_RNA_seq_mosca/genome/ensembl_dm3.chr4.gtf \
    -outdir qc_qualimap/${sample} \
    -p non-strand-specific \
    -pe
done
```

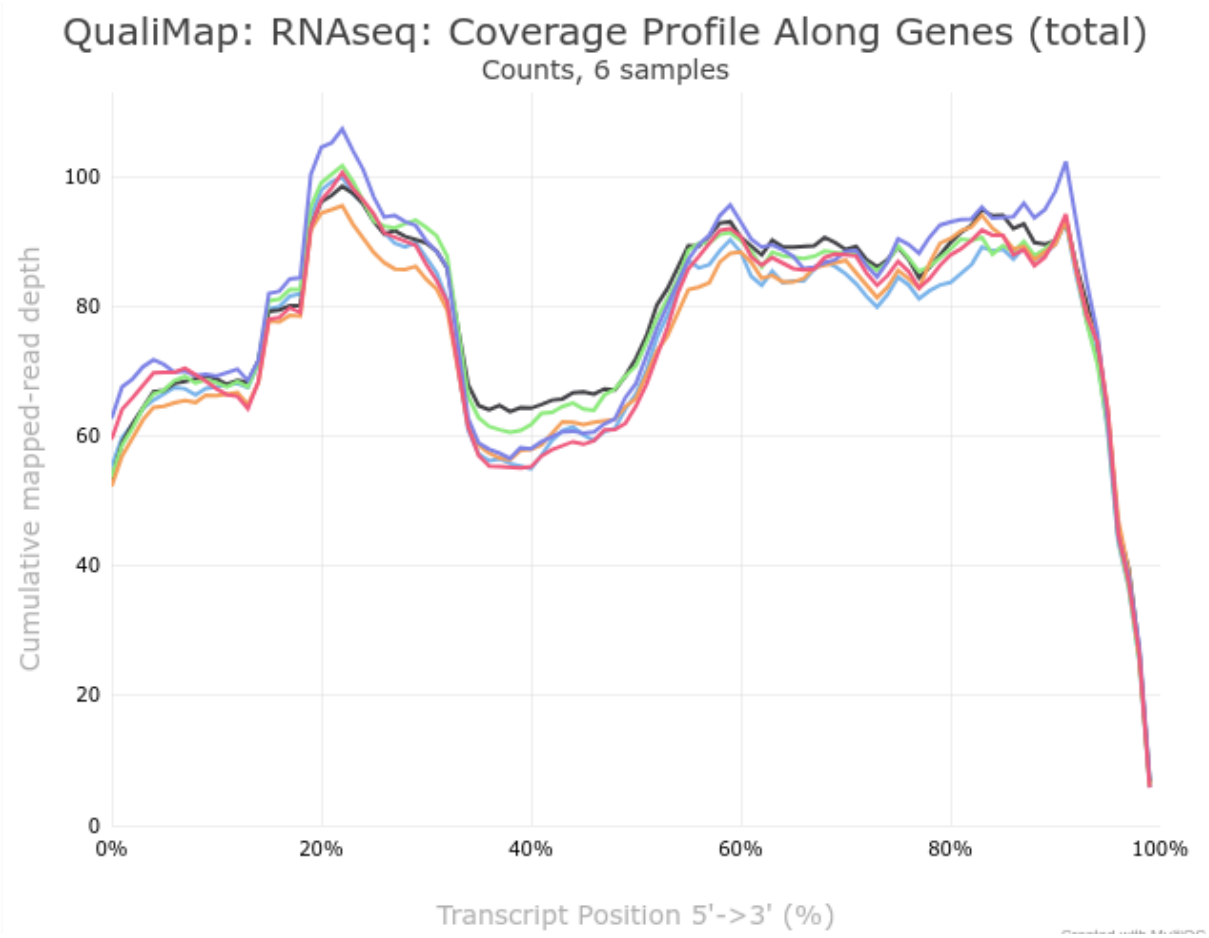
Los resultados se agruparon mediante MultiQC. Los resultados se encuentran en el directorio: */Mapping/qc_qualimap/*multiqc_report_qualimap. Se empleó el comando:

```
multiqc . -o multiqc_report_qualimap
```

Sample Name	5'-3' bias	M Aligned	Aligned	Uniq aligned
g1_01	1.95	0.0 M	99.7 %	94.8 %
g1_02	2.03	0.0 M	99.8 %	94.7 %
g1_03	1.95	0.0 M	99.8 %	94.9 %
g3_01	2.29	0.0 M	99.7 %	93.9 %
g3_02	1.77	0.0 M	99.7 %	95.8 %
g3_03	1.65	0.0 M	99.8 %	96.0 %



En todos los casos más del 86% fue de origen exónico.



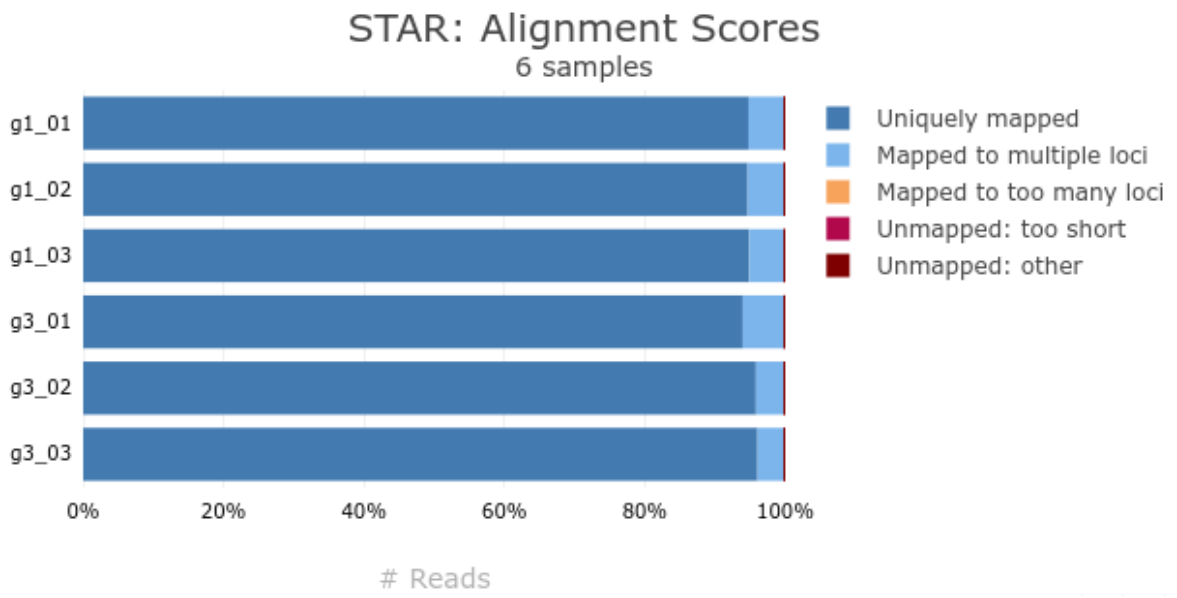
Mapeo con STAR

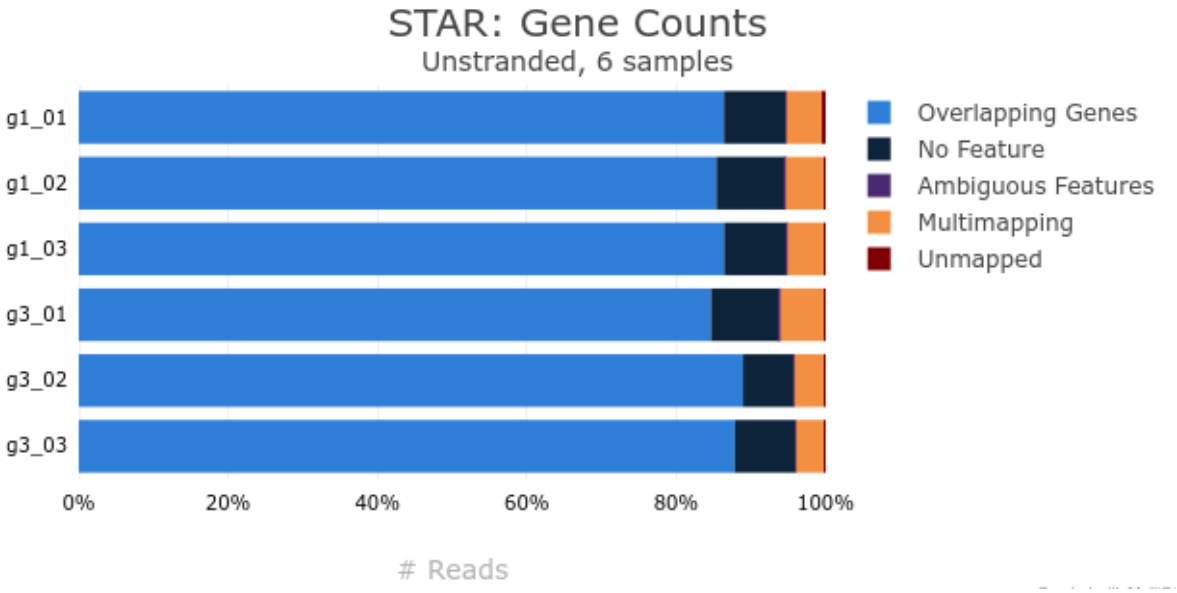
Summary Statistics

Summary statistics from the STAR alignment

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 6/6 rows and 10/19 columns. | Summarize table

Sample Name	Total reads	Aligned	Uniq aligned	Avg. mapped len	Annotated splices	Mismatch rate	Del rate	Del len	Ins rate	Ins len
g1_01	0.0 M	99.7 %	94.8 %	199.7 bp	0.0 M	0.2 %	0.0 %	1.2 bp	0.0 %	1.1 bp
g1_02	0.0 M	99.8 %	94.7 %	199.7 bp	0.0 M	0.2 %	0.0 %	1.2 bp	0.0 %	1.1 bp
g1_03	0.0 M	99.8 %	94.9 %	199.7 bp	0.0 M	0.2 %	0.0 %	1.2 bp	0.0 %	1.1 bp
g3_01	0.0 M	99.7 %	93.9 %	199.7 bp	0.0 M	0.2 %	0.0 %	1.2 bp	0.0 %	1.1 bp
g3_02	0.0 M	99.7 %	95.8 %	199.7 bp	0.0 M	0.2 %	0.0 %	1.2 bp	0.0 %	1.1 bp
g3_03	0.0 M	99.8 %	96.0 %	199.7 bp	0.0 M	0.2 %	0.0 %	1.3 bp	0.0 %	1.1 bp





Se generaron los archivos de cobertura mediante la herramienta bamCoverage de deepTools. Hasta este paso los archivos BAM no están indexados, y **bamCoverage** requiere un archivo de índice *.bai* para poder procesarlos. Para ello primero se generaron los índices mediante **samtools**:

```
for sample in g1_01 g1_02 g1_03 g3_01 g3_02 g3_03
do
  samtools index alignments_STAR/${sample}_Aligned.sortedByCoord.out.bam
done
```

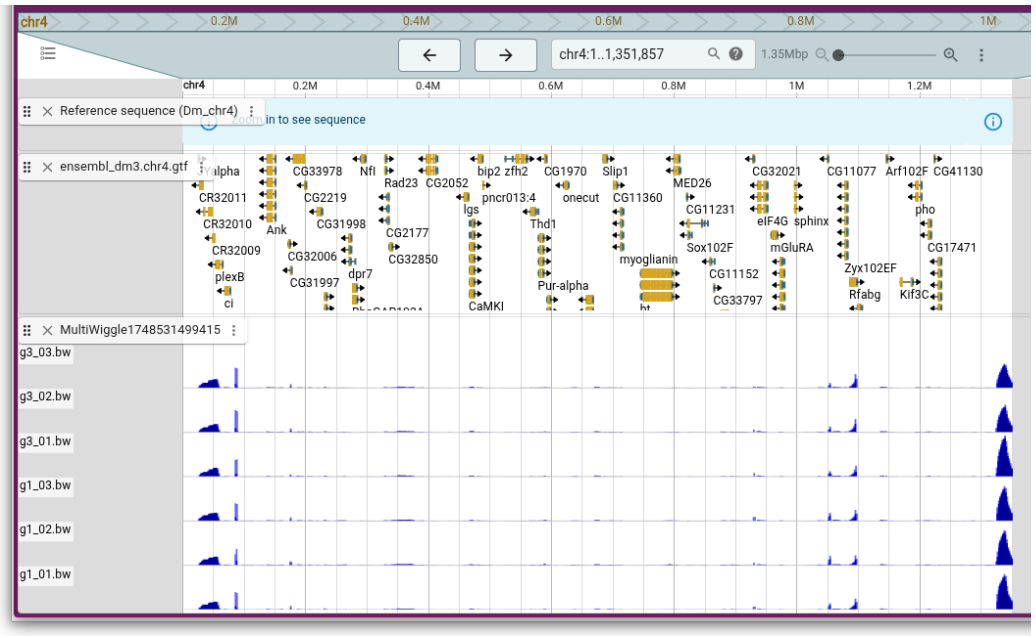
Ahora sí se puede ejecutar **bamCoverage**. Se empleó un valor de bin size igual a 5 dada a características de los datos reducidos. Los resultados se encuentran en el directorio: */Mapping/bw_tracks.* Se empleó el comando:

```
for sample in g1_01 g1_02 g1_03 g3_01 g3_02 g3_03
do
  echo "Generando .bw para $sample..."
  bamCoverage \
    -b alignments_STAR/${sample}_Aligned.sortedByCoord.out.bam \
    -o bw_tracks/${sample}.bw \
    --binSize 5 \
    --normalizeUsing CPM \
    --extendReads \
    --ignoreDuplicates \
    --numberOfProcessors 16
done
```

Donde:

- **-b:** Archivo BAM de entrada (ordenado por coordenadas).
- **-o:** Archivo .bw de salida.
- **--binSize 5**: Agrupa la señal cada 5 pb.
- **--normalizeUsing CPM**: Normaliza la cobertura por millones de lecturas mapeadas (útil para comparación entre muestras).
- **--extendReads**: Extiende las lecturas para simular el fragmento completo.
- **--ignoreDuplicates**: Ignora lecturas duplicadas.
- **--numberOfProcessors 16:** Usa 16 núcleos.

Se visualizan mediante JBrowse:



Como para este caso se empleó STAR con el parámetro ***--quantMode GeneCounts***, ya tengo los archivos de conteo por gen generados automáticamente por muestra. Estos archivos son los que tienen el formato *_ReadsPerGene.out.tab.* Se muestra un archivo a modo de ejemplo:

g1_01_ReadsPerGene.out.tab

1	N_unmapped	137	137	137
2	N_multimapping	2485	2485	2485
3	N_noFeature	4111	25647	25778
4	N_ambiguous	69	1	4
5	FBgn0040037	11	6	5
6	FBgn0052011	0	0	0
7	FBgn0052010	4	3	1
8	FBgn0052009	0	0	0
9	FBgn0025740	252	114	138
10	FBgn0004859	679	330	349
11	FBgn0017545	3776	1892	1884

Estos archivos tienen 4 columnas: la columna 1 muestra el ID del gen (desde el GTF), la columna 2 es *Unstranded* y muestra el conteo total sin distinguir hebra, la columna 3 corresponde al *Strand 1*, y la columna 4 corresponde al *Strand 2*; Como en este caso estoy trabajando con una librería no direccionada (*unstranded*), empleo los datos de la columna 2.

Se generó la matriz de conteos mediante un script de R. La misma se guardó en la carpeta: */Mapping/count_matrix_STAR.txt*.

count_matrix_STAR.txt

	GeneID	g1_01	g1_02	g1_03	g3_01	g3_02	g3_03
2	FBgn0002521	227	223	231	235	207	268
3	FBgn0004607	489	383	415	374	322	539
4	FBgn0004624	618	590	616	636	566	548
5	FBgn0004859	679	670	570	631	542	871
6	FBgn0005558	35	35	28	27	33	36
7	FBgn0005561	49	57	52	61	30	35
8	FBgn0005666	984	1757	1479	674	713	447
9	FBgn0010217	2815	4152	3529	2244	2431	1981
10	FBgn0011642	373	337	343	374	353	373
11	FBgn0011747	813	816	834	925	832	911
12	FBgn0013749	150	177	168	191	210	179
13	FBgn0016126	507	429	448	554	637	568
14	FBgn0017545	3776	3610	3819	3736	4647	4273
15	FBgn0019650	59	54	52	64	61	60
16	FBgn0019985	49	48	37	57	29	32
17	FBgn0022361	154	146	146	150	149	171
18	FBgn0023213	1163	1124	1123	1202	1293	1114

Para el análisis de expresión diferencial se usó **DESeq2**, el cuál utiliza modelos lineales generalizados binomiales negativos, basándose en la hipótesis de que la mayoría de los

genes no se expresan de manera diferencial. Los pasos seguidos se encuentran en el script de R Markdown en la carpeta: */expresion_diferencial/expresion_diferencial.Rmd*.

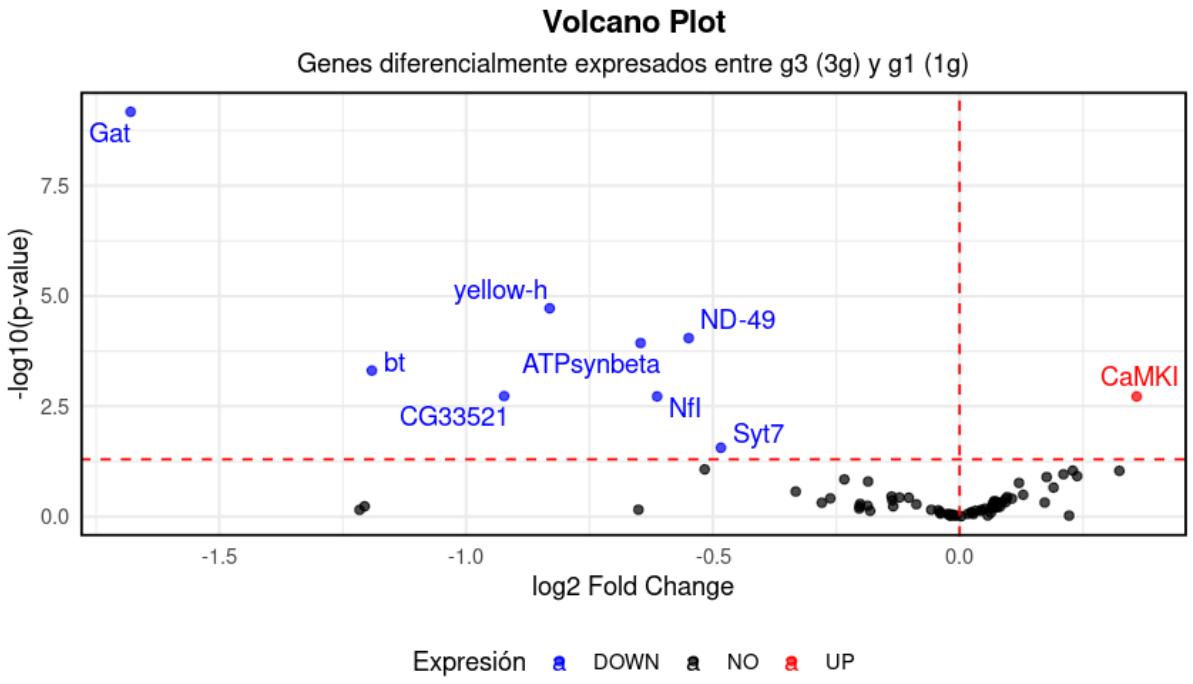
En el análisis inicial se utilizó un umbral de l**og2FoldChange > 0.5** para identificar genes diferencialmente expresados. Sin embargo, dada la naturaleza del experimento, se observó que pocos genes superaban ese umbral con significancia estadística (**p-value < 0.05**).

Teniendo en cuenta que el número total de genes analizados fue limitado (solo del cromosoma 4), el uso de un umbral estricto reducía excesivamente la cantidad de genes detectables, comprometiendo el poder del análisis funcional posterior.

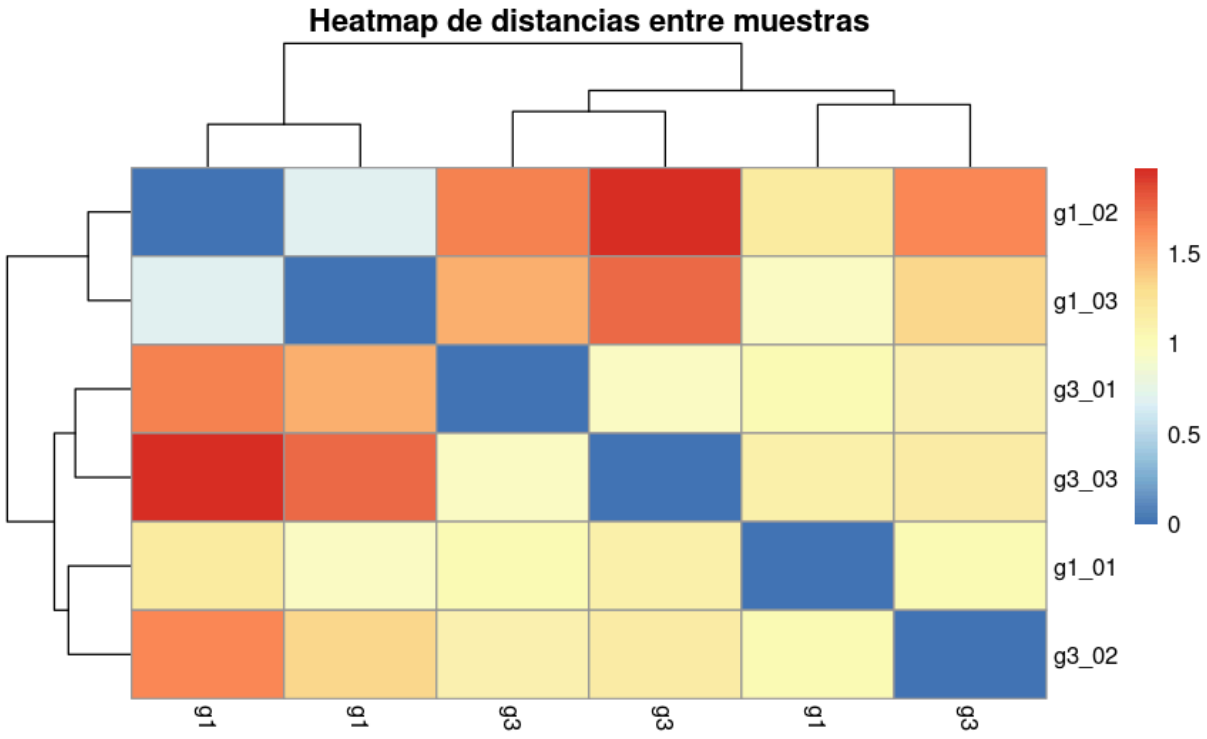
Para poder explorar con mayor sensibilidad los efectos relevantes de la condición experimental sobre la expresión génica, se optó por relajar el umbral de **log2FoldChange a 0**. Lo que permite incluir genes con cambios pequeños en la expresión, que podrían estar involucrados en procesos biológicos de baja magnitud pero alta especificidad.

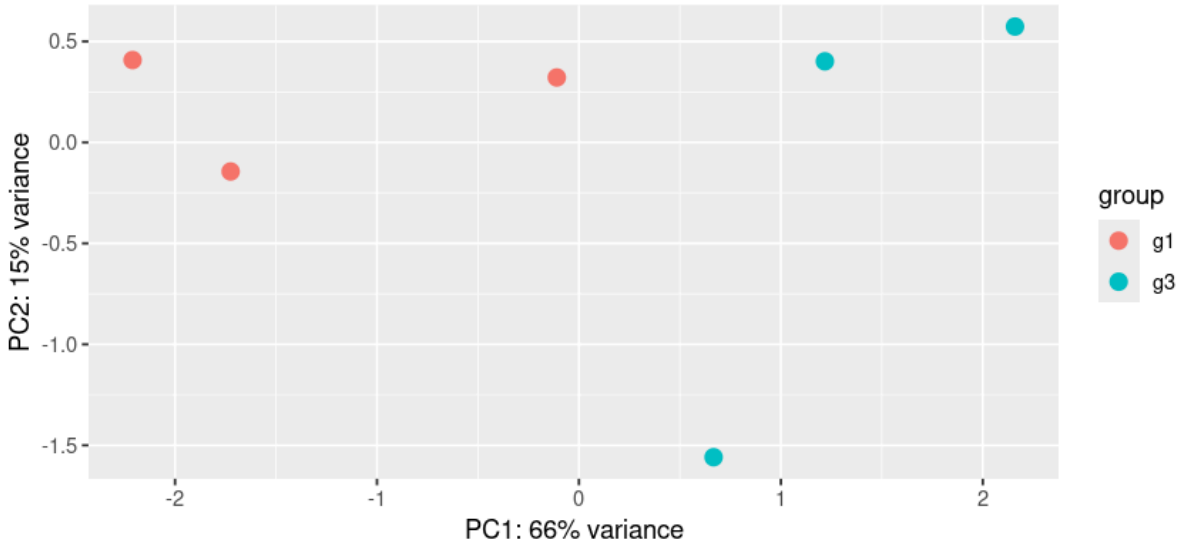
Tras realizar el análisis de expresión diferencial entre las condiciones de gravedad normal (1g) y gravedad aumentada (3g) en *Drosophila melanogaster*, se identificaron un conjunto de genes que mostraron cambios significativos en sus niveles de expresión entre condiciones (g3 vs g1):

external_gene_name	flybase_gene_id
Gat	FBgn0039915
Nfl	FBgn0042696
Syt7	FBgn0039900
CG33521	FBgn0250819
yellow-h	FBgn0039896
bt	FBgn0005666
ND-49	FBgn0039909
CaMKI	FBgn0016126
ATPsynbeta	FBgn0010217



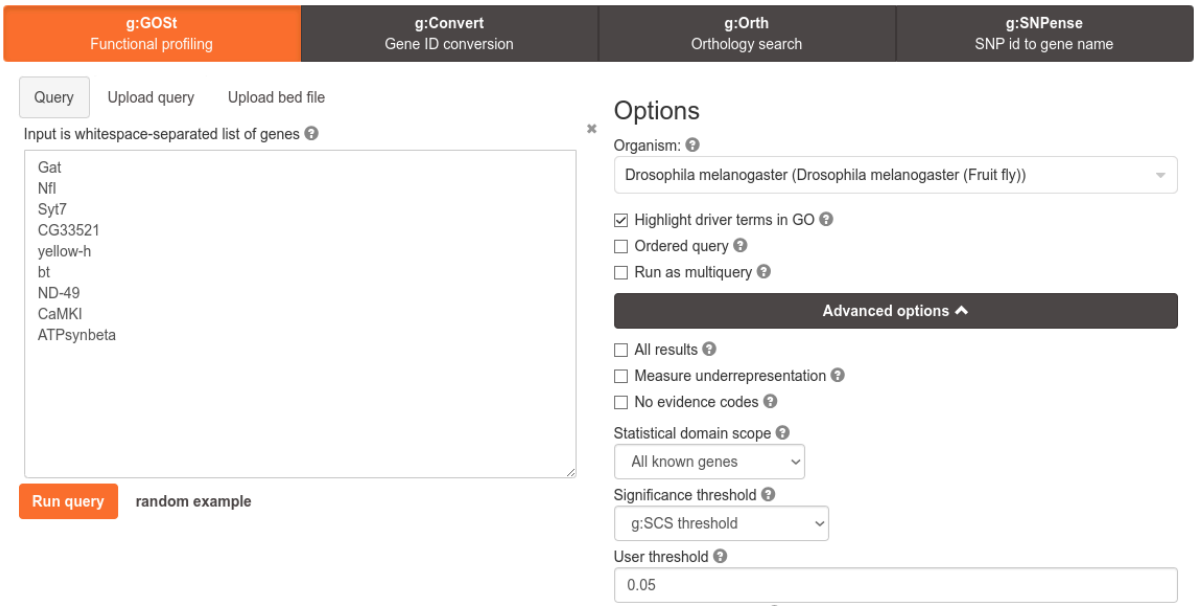
Análisis RNA-seq con DESeq2



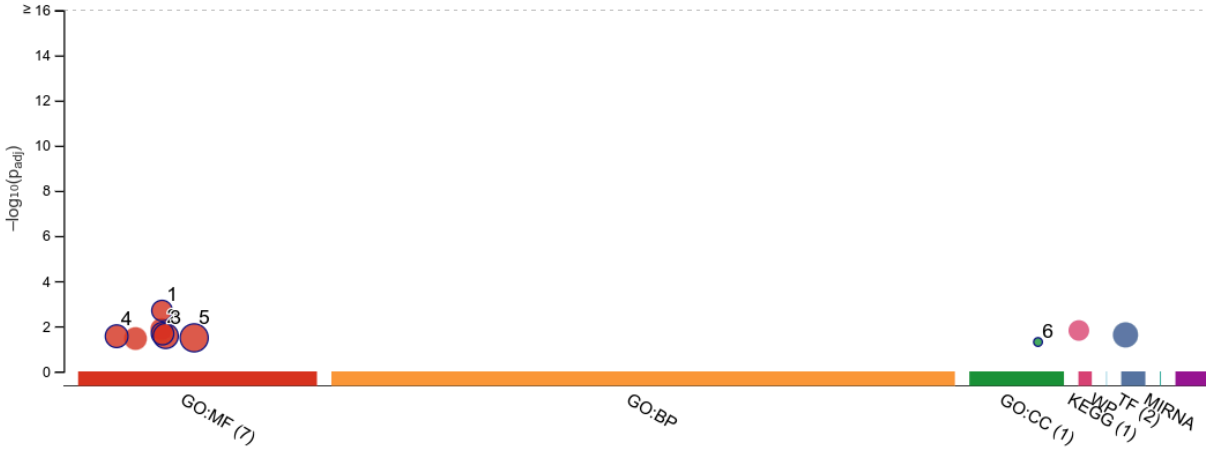


Como no me funcionó hacer el enriquecimiento funcional mediante Enrichr con R a pesar de haber convertido los IDs, se continuó con un enriquecimiento funcional mediante la plataforma de g:Profiler y se realizó un estudio para confirmar la función de los genes seleccionados.

Se empleó la función **g:GOSt**, esta realiza análisis de enriquecimiento funcional, también conocido como análisis de sobrerrepresentación (ORA) o análisis de enriquecimiento del conjunto de genes. Mapea genes a fuentes de información funcional conocidas y detecta términos con enriquecimiento estadísticamente significativo. Además de Gene Ontology, incluye vías de KEGG Reactome y WikiPathways; dianas de miRNA de miRTarBase y coincidencias de motivos reguladores de TRANSFAC; especificidad tisular de Human Protein Atlas; complejos proteicos de CORUM y fenotipos de enfermedades humanas de Human Phenotype Ontology. g:GOSt admite cerca de 500 organismos y acepta cientos de tipos de identificadores.



Se seleccionó como organismos a *Drosophila melanogaster.* Para encontrar más términos significativos se configuró el parámetros de *Statistical domain scope* como ***All know genes***, en lugar del por defecto de *Only annotated genes*.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0022853	active monoatomic ion transmembrane transpor…	2.027×10^{-3}
2	GO:MF	GO:0022890	inorganic cation transmembrane transporter act…	2.122×10^{-2}
3	GO:MF	GO:0030554	adenyl nucleotide binding	2.697×10^{-2}
4	GO:MF	GO:0008324	monoatomic cation transmembrane transporter …	2.771×10^{-2}
5	GO:MF	GO:0036094	small molecule binding	3.266×10^{-2}
6	GO:CC	GO:0097386	glial cell projection	4.996×10^{-2}

Las versiones para las búsquedas en las bases de datos son:

dmelanogaster (Drosophila melanogaster (Fruit fly)) - version: **BDGP6.46**

GO:MF – annotations: BioMart

classes: releases/2024-10-27
GO:CC – annotations: BioMart

classes: releases/2024-10-27
GO:BP – annotations: BioMart

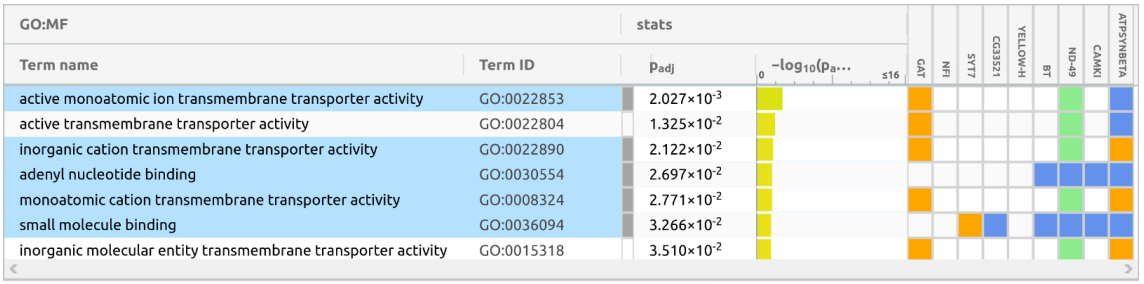
classes: releases/2024-10-27
KEGG – KEGG FTP Release 2024-01-22
REAC – annotations: BioMart

classes: 2025-2-3
WP – 20250110
TF – annotations: TRANSFAC Release 2023.2

classes: v2
MIRNA – Release 9.0
HPA – annotations: HPA website: 23-07-17

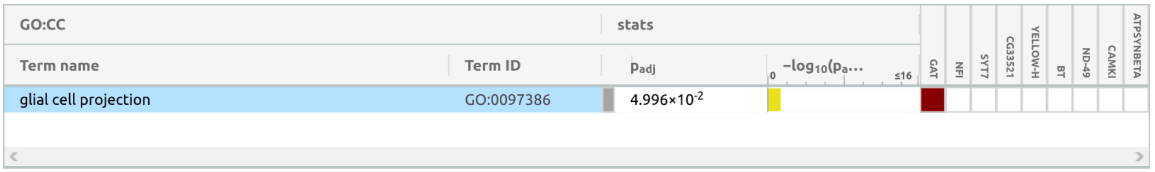
classes: script: 24-01-02
CORUM – 28.11.2022 Corum 4.1
HP – annotations: 02.2025

classes: None



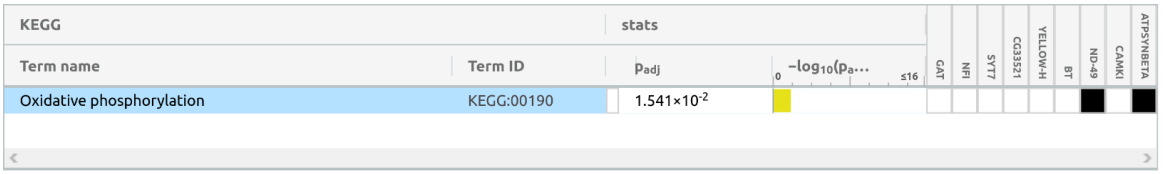
GO:MF		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_a\ldots$ (0 … ≤16)
active monoatomic ion transmembrane transporter activity	GO:0022853	2.027×10^{-3}	
active transmembrane transporter activity	GO:0022804	1.325×10^{-2}	
inorganic cation transmembrane transporter activity	GO:0022890	2.122×10^{-2}	
adenyl nucleotide binding	GO:0030554	2.697×10^{-2}	
monoatomic cation transmembrane transporter activity	GO:0008324	2.771×10^{-2}	
small molecule binding	GO:0036094	3.266×10^{-2}	
inorganic molecular entity transmembrane transporter activity	GO:0015318	3.510×10^{-2}	

GAT · NFI · SYT7 · CG33521 · YELLOW-H · BT · ND-49 · CAMKI · ATPSYNBETA



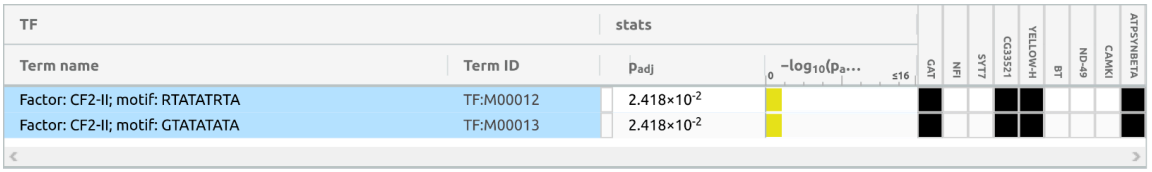
GO:CC		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_a\ldots$ (0 … ≤16)
glial cell projection	GO:0097386	4.996×10^{-2}	

GAT · NFI · SYT7 · CG33521 · YELLOW-H · BT · ND-49 · CAMKI · ATPSYNBETA



KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_a\ldots$ (0 … ≤16)
Oxidative phosphorylation	KEGG:00190	1.541×10^{-2}	

GAT · NFI · SYT7 · CG33521 · YELLOW-H · BT · ND-49 · CAMKI · ATPSYNBETA



TF		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_a\ldots$ (0 … ≤16)
Factor: CF2-II; motif: RTATATRTA	TF:M00012	2.418×10^{-2}	
Factor: CF2-II; motif: GTATATATA	TF:M00013	2.418×10^{-2}	

GAT · NFI · SYT7 · CG33521 · YELLOW-H · BT · ND-49 · CAMKI · ATPSYNBETA

El análisis de enriquecimiento funcional de los genes diferencialmente expresados entre pupas desarrolladas en hipergravedad (3g) y en gravedad estándar (1g) reveló funciones moleculares y rutas relevantes afectadas por la condición experimental. Dentro de la categoría GO:MF (*Molecular Function*), se identificaron términos significativamente enriquecidos asociados al transporte iónico, incluyendo "*active monoatomic ion transmembrane transporter activity*", "*inorganic cation transmembrane transporter activity*" y "*monoatomic cation transmembrane transporter activity*". Estos resultados indican una posible alteración en los mecanismos de transporte de iones a través de membranas, lo que sugiere un ajuste funcional en la homeostasis celular ante la exposición a hipergravedad.

Asimismo, se detectaron términos relacionados con la unión de nucleótidos y moléculas pequeñas ("*adenyl nucleotide binding*", "*small molecule binding*"), que podrían reflejar una modulación de procesos enzimáticos o de señalización celular vinculados al metabolismo energético y al control de funciones celulares esenciales.

En la categoría GO:CC (*Cellular Component*), el término "*glial cell projection*" apareció enriquecido, lo cual podría estar relacionado con cambios en la morfología celular o en la extensión de proyecciones gliales, especialmente bajo condiciones de estrés físico como la hipergravedad.

El análisis también reveló un enriquecimiento significativo en la vía de fosforilación oxidativa según KEGG, lo que es coherente con la regulación negativa de genes mitocondriales detectada en el análisis de expresión (como ND-49 y ATPsynbeta), implicando una potencial disminución de la actividad bioenergética.

Finalmente, se identificó la posible participación del factor de transcripción CF2-II, a través de dos motivos reguladores (RTATAT(R)TA y GTATATATA), sugiriendo un mecanismo de control transcripcional específico que podría estar regulando los genes afectados por la hipergravedad.

Conclusiones

Mediante este análisis se demuestra que la exposición a hipergravedad durante el desarrollo pupal de Drosophila melanogaster induce cambios transcriptómicos significativos. Se identificaron genes diferencialmente expresados relacionados con el metabolismo mitocondrial, el transporte iónico y la señalización neuronal, lo que sugiere una reprogramación funcional en respuesta al estrés gravitacional. Se refuerza el valor de Drosophila como modelo para estudiar la adaptación molecular a entornos físicos extremos. Estudios futuros podrían profundizar en los mecanismos reguladores que median esta respuesta, con aplicaciones en biología espacial y fisiología del desarrollo.