Actividad 1- *From reads to gene counts*

Objetivo

El propósito de esta actividad es que el estudiante demuestre que ha adquirido las habilidades y competencias necesarias para realizar el **preprocesamiento inicial de datos de RNA-seq.** Esto incluye desde la caracterización y evaluación de calidad de las lecturas crudas hasta su filtrado, alineación y generación de la matriz final de recuentos génicos.

Obtención de los datos

La actividad consiste en el análisis de una muestra real de RNA-seq, que forma parte de un estudio más amplio que examina los perfiles de expresión entre poblaciones de células basales y luminales en glándulas mamarias de ratones en diferentes estadios (vírgenes, gestantes y lactantes).

La publicación de referencia es la siguiente:

Fu, Nai Yang, et al. "EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival." *Nature Cell Biology* 17.4 (2015): 365-375.

Todo lo necesario para poder realizar esta actividad ha sido previamente abordado en clase y, por tanto, todo el material (incluyendo el entorno de trabajo conda 05MBIF y los archivos necesarios) puede encontrarse en la sección de "Recursos y Materiales/Materiales del profesor/Proyecto_JULIO_2024".

Formato de la entrega

- La entrega se realizará utilizando este documento como plantilla, que se convertirá en PDF y lo adjuntará a la actividad correspondiente dentro del Campus VIU.
- Las respuestas se presentarán de forma clara y concisa, justificando su contenido.
- Se deberá adicionar SIEMPRE los comandos empleados, las capturas de pantalla que muestren su ejecución y los gráficos generados que apoyen sus repuestas:
    - Para salidas estándar de larga extensión de algún comando, incluir solo las primeras líneas que muestren su correcta ejecución.
    - No es necesario explicar con detalle los comandos, opciones y/o argumentos empleados.

- Los gráficos deberán tener una calidad suficiente para su lectura y serán acompañados de un pie de figura explicativo con el número de imagen (Figura X: "…").

Preguntas

Para alcanzar la nota máxima, se deberán contestar y justificar cada una de las siguientes preguntas.

P1. Responde brevemente a las siguientes preguntas sobre la caracterización inicial del estudio (1 pts)

- ¿Cuántas réplicas biológicas por grupo se utilizaron en este estudio? ¿Es este un número óptimo para experimentos de RNA-seq? Justifica tu respuesta.
  2 réplicas biológicas por grupo.
  No es un número óptimo para experimentos de RNA-seq ya que tan pocas réplicas hace que tenga poca potencia estadística y poca confianza en los resultados obtenidos, por la variabilidad entre individuos, que puede ser más alta. Cuanto mayor número de réplicas mejor para reducir esa variabilidad entre individuos. Además, el número mínimo de réplicas recomendado en estudios con muestras de animales es de 3 réplicas.
- Describe el protocolo de secuenciación utilizado en este estudio.
  Primero se extrajo el RNA de las glándulas mamarias para generar la biblioteca siguiendo el protocolo de TruSeq RNA v2 de Illumina de preparación de muestras y las bibliotecas fueron secuenciadas con Illumina HiSeq 2000.
  Las lecturas se alinearon con el genoma del ratón mm10 utilizando Rsubread versión 13.25.
  Después para el conteo de la cantidad de lecturas que se superponían a cada gen utilizaron la anotación de genes RefSeq y featureCounts. Para el filtrado de los genes expresados y su normalización utilizaron el paquete edgeR. Posteriormente, evalúan la expresión diferencial utilizando el método TREAT.

P2. Emplea la herramienta FastQC para evaluar la calidad de las lecturas del archivo SRR1552444.fastq.gz. A continuación, utiliza la herramienta TrimGalore para eliminar adaptadores y extremos de baja calidad. Reporta los comandos utilizados y responde a las siguientes preguntas (3 pts).

- ¿Cuántas lecturas iniciales había en el archivo SRR1552444.fastq.gz y cuántas permanecen tras la utilización de TrimGalore?
  El número de lecturas iniciales en el archivo fastq es de 27919481 lecturas.
  El comando empleado para obtener el número de lecturas es: zgrep -c "^@SRR" SRR1552444.fastq.gz
  Comando empleado para análisis FastQC: fastqc SRR1552444.fastq.gz –o ../Processed/01.Quality_control/

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Raw]$ zgrep -c "^@SRR" SRR1552444.fastq.gz
27919481
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Raw]$ fastqc SRR1552444.fastq.gz --o ../Proc
essed/01.Quality_control/
Started analysis of SRR1552444.fastq.gz
Approx 5% complete for SRR1552444.fastq.gz
Approx 10% complete for SRR1552444.fastq.gz
Approx 15% complete for SRR1552444.fastq.gz
Approx 20% complete for SRR1552444.fastq.gz
Approx 25% complete for SRR1552444.fastq.gz
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 01.Quality_control]$ firefox SRR1552444_fast
qc.html
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 01.Quality_control]$
```

Comando empleado para la eliminación de adaptadores y extremos de baja calidad con TrimGalores: trim_galore SRR1552444.fastq.gz -o ../Processed/02.Trimming/

Número de lecturas tras la utilización de TrimGalore: 27906762 lecturas.

Comando empleado: zgrep -c "^@SRR" SRR1552444_trimmed.fq.gz

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Raw]$ trim_galore SRR1552444.fastq.gz -o ../
Processed/02.Trimming/
Multicore support not enabled. Proceeding with single-core trimming.
Path to Cutadapt set as: 'cutadapt' (default)
Cutadapt seems to be working fine (tested command 'cutadapt --version')
Cutadapt version: 3.4
single-core operation.
igzip command line interface 2.30.0
igzip detected. Using igzip for decompressing

No quality encoding type selected. Assuming that the data provided uses Sanger encoded Phred scores (def
ault)

Output will be written into the directory: /home/msevillanogonzalez/Documentos/Analisis_transcriptomicos
/Proyecto_JULIO_2024/Data/Processed/02.Trimming/


AUTO-DETECTING ADAPTER TYPE
```

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 02.Trimming]$ fastqc SRR1552444_trimmed.fq.gz
Started analysis of SRR1552444_trimmed.fq.gz
Approx 5% complete for SRR1552444_trimmed.fq.gz
Approx 10% complete for SRR1552444_trimmed.fq.gz
Approx 15% complete for SRR1552444_trimmed.fq.gz
Approx 20% complete for SRR1552444_trimmed.fq.gz
Approx 25% complete for SRR1552444_trimmed.fq.gz
```

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 02.Trimming]$ zgrep -c "^@SRR" SRR1552444_tr
immed.fq.gz
27906762
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 02.Trimming]$
```

- Comenta las diferencias que observas en los reportes de FastQC antes y después del filtrado (recuerda reportar los gráficos resultantes), específicamente en:
    - **Per base sequence quality**

      En las gráficas de las figuras 1 y 2 se puede ver que la calidad de las secuencias por base, tanto antes como después del filtrado es alta, como indica el valor Phred score (>30). Al recortar extremos se observa una pequeña diferencia en la figura 2, todos los bigotes de las cajas entran en el rango de calidad alta, y la línea azul, que indica la calidad promedio para el nucleótido en todas las lecturas, también mejora.

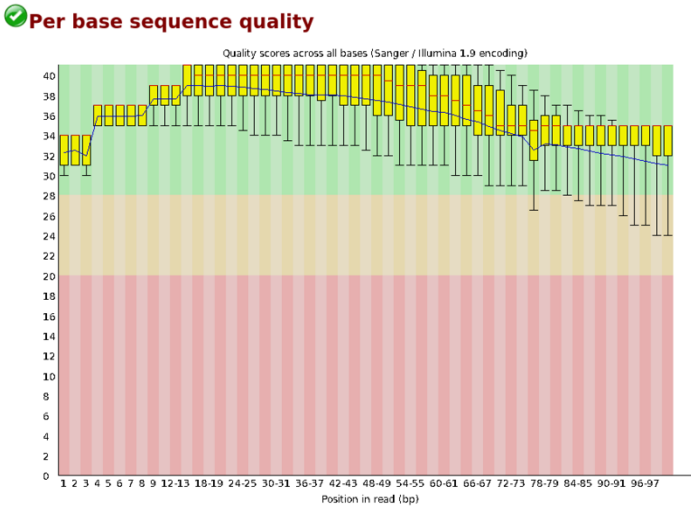


Figura 1. Gráfico de calidad de las secuencias por base del archivo SRR155244.fastq.gz antes del filtrado de calidad

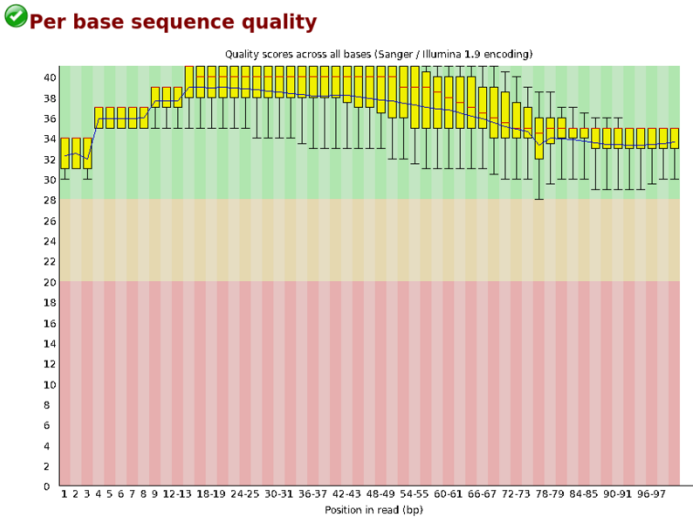


Figura 2. Gráfico de calidad de las secuencias por base del archivo SRR155244.trimmed.fq.gz después del filtrado de calidad

- **Sequence Length Distribution**

Estos gráficos representan la distribución de la longitud de las secuencias. En la figura 3, que corresponde con el archivo antes de filtrar, se observa un pico en el 100, lo que quiere decir que la longitud de las lecturas es de un tamaño de 100 nucleótidos. Sin embargo, en la figura 4 se observa otra distribución tras el filtrado de calidad y recorte, por lo que la longitud de las lecturas varía entre 20 y 100, lo que disminuye un poco su calidad al haber lecturas más cortas.

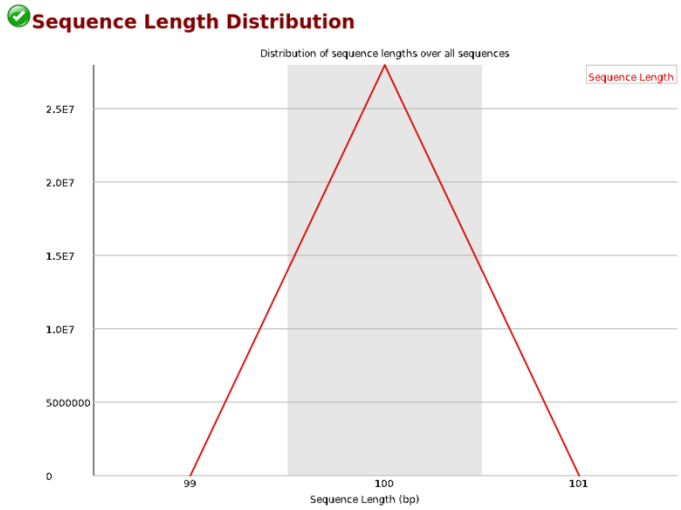


Figura 3. Gráfico de distribución de longitud de las secuencias del archivo SRR155244.fastq.gz antes del filtrado de calidad

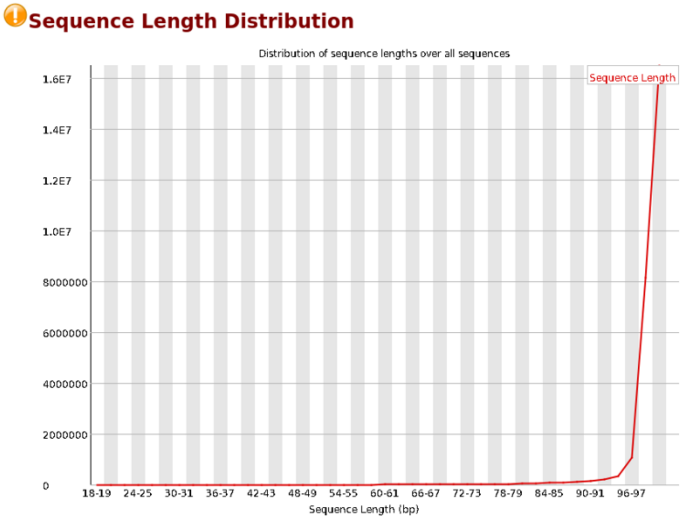


Figura 4. Gráfico de distribución de longitud de las secuencias del archivo SRR155244.trimmed.fq.gz después del filtrado de calidad

- **Adapter Content.** ¿Qué secuencia se determina como el adaptador principal? ¿En cuántas lecturas ha encontrado dicho adaptador TrimGalore?

```
=== Adapter 1 ===

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 9664562 times
```

Secuencia que se determina como el adaptador principal: AGATCGGAAGAGC
Lecturas en las que ha encontrado dicho adaptador: 9664562 lecturas.

En la figura 5, se muestra el gráfico del contenido de adaptadores del archivo sin filtrar, por lo que al final de las lecturas aparece una pequeña cantidad de adaptadores, que podría ser porque los fragmentos son más cortos que la longitud de la lectura. En el gráfico 6, tras

el filtrado de calidad y recorte de adaptadores se observa la línea horizontal en el cero, que indica la eliminación de esos adaptadores.

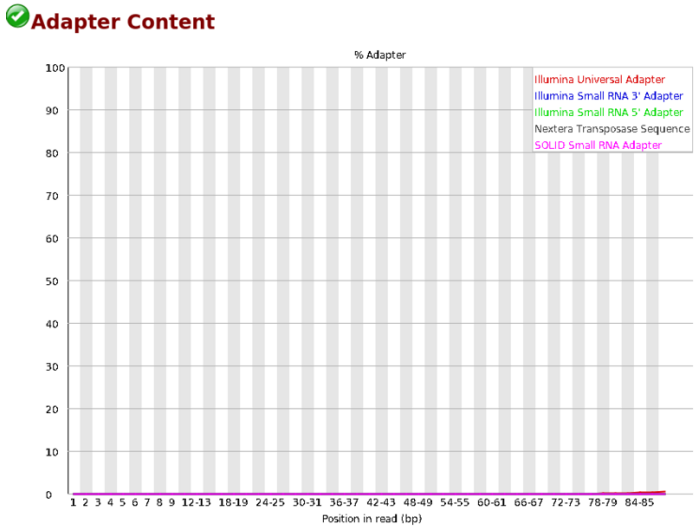


Figura 5. Gráfico del contenido de adaptadores del archivo SRR155244. fastq.gz antes del filtrado de calidad

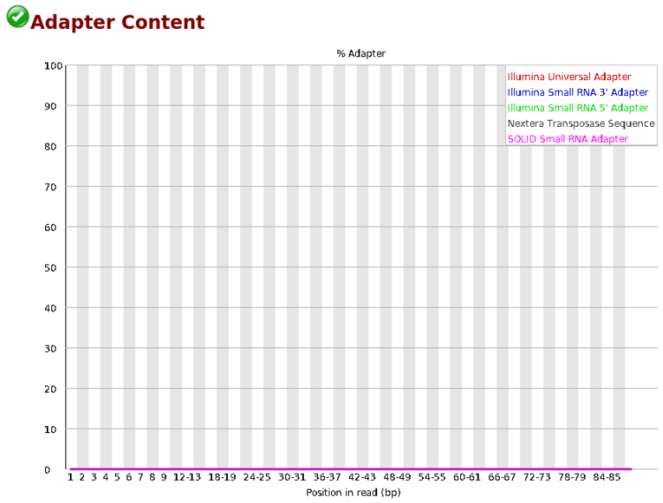


Figura 6. Gráfico del contenido de adaptadores del archivo SRR155244.trimmed.fq.gz después del filtrado de calidad

P3. Alinea las lecturas depuradas anteriormente sobre el genoma de referencia indexado empleando la herramienta Hisat2 con la opción (-k = 1). Reporta los comandos empleados y contesta a las siguientes preguntas. (1,5 pts)

Comando para el alineamiento de las lecturas con -k=1:

hisat2 -k1 -U ../02.Trimming/SRR1552444_trimmed.fq.gz -x ../../Reference_genome/mm10/genome -S SRR1552444_hisat2.sam

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 03.Alignment]$ hisat2 -k1 -U ../02.Trimming/SRR1552444_trimmed.fq.gz -x
../../Reference_genome/mm10/genome -S SRR1552444_hisat2.sam
27906762 reads; of these:
  27906762 (100.00%) were unpaired; of these:
    817651 (2.93%) aligned 0 times
    27089111 (97.07%) aligned exactly 1 time
    0 (0.00%) aligned >1 times
97.07% overall alignment rate
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 03.Alignment]$
```

Comando para el alineamiento de las lecturas con -k=5:

hisat2 -k5 -U ../02.Trimming/SRR1552444_trimmed.fq.gz -x ../../Reference_genome/mm10/genome -S SRR1552444_hisat2k5.sam

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 03.Alignment]$ hisat2 -k5 -U ../02.Trimming/
SRR1552444_trimmed.fq.gz -x ../../Reference_genome/mm10/genome -S SRR1552444_hisat2k5.sam
27906762 reads; of these:
  27906762 (100.00%) were unpaired; of these:
    797754 (2.86%) aligned 0 times
    24348188 (87.25%) aligned exactly 1 time
    2760820 (9.89%) aligned >1 times
97.14% overall alignment rate
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 03.Alignment]$
```

- Después del alineamiento con Hisat2, ¿cuál fue el número total de lecturas mapeadas y no mapeadas en la muestra?
  Número total de lecturas mapeadas: 27089111 lecturas.
  Número total de lecturas no mapeadas: 817651 lecturas.

- Recuerda que uno de los parámetros de Hisat2 es el valor de -k, que indica el número máximo de alineamientos por lectura a generar. Si empleas un valor de **k = 5**, ¿Cómo afecta esto a los valores finales de alineamiento obtenidos y al tiempo de computación?
  Como antes el número máximo de alineamientos por lectura a generar era de 1 (k=1) no había lecturas con más de un alineamiento. En este caso, al aumentar el número máximo de alineamientos por lectura a 5 (k=5), tanto el número de lecturas no mapeadas como el número de lecturas que alinean una vez ha disminuido, mientras que ahora si hay lecturas que alinean más de una vez, en concreto, 2760820 lecturas.
  El tiempo de computación también ha aumentado respecto al anterior.

P4. Utiliza SAMtools para convertir el archivo SAM (generado con el valor de -k igual a 1) a BAM, ordenarlo por coordenadas e indexarlo. Reporta los comandos utilizados y contesta a las siguientes preguntas. (1,5 pts)

Comandos utilizados:
samtools view -Sbh SRR1552444_hisat2.sam > SRR155244_hisat2.bam
samtools sort SRR15samtools index SRR1552444_hisat2.sorted.bam
samtools index SRR1552444_hisat2.sorted.bam

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Data]$ samtools view -Sbh SRR1552444_hisat2.sam > SRR1552444_hisat2.bam
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Data]$ samtools sort SRR1552444_hisat2.bam -o SRR1552444_hisat2.sorted.b
am
```

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Data]$ samtools index SRR1552444_hisat2.sorted.bam
```

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Data]$ samtools stats SRR1552444_hisat2.sorted.bam > SRR1552444_hisat2.s
orted.bam.stats
```

- ¿Cuántas FLAGs distintas se encuentran en el archivo **SRR1552444_hisat2.sam**? Indica cuáles son y sus cantidades.
  Comando para contar el número de FLAGS:
  grep -v "^@" SRR1552444_hisat2.sam | cut -f2 | sort -n | uniq -c

  Hay 3 FLAGs distintas en el archivo, que son: 0, 4 y 16.
  La FLAG 0 aparece 13569495 veces.
  La FLAG 4 aparece 817651 veces.
  Y la FLAG 16 aparece 13519616 veces.

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Processed]$ grep -v "^@" SRR1552444_hisat2.sam | cut -f2 | sort -n | uni
q -c
13569495 0
 817651 4
13519616 16
```

- ¿Cuántos valores distintos de MAPQ hay en el archivo **SRR1552444_hisat2.sam**? Indícalos y cuenta cuántos son.
  Comando para contar los valores de MAPQ:
  grep -v "^@" SRR1552444_hisat2.sam | cut -f5 | sort -n | uniq -c

  Hay 3 valores distintos de MAPQ: 0, 1 y 60.
  El valor 0 aparece 907534 veces.
  El valor 1 aparece 3171570 veces.
  El valor 60 aparece 23827658 veces.

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Processed]$ grep -v "^@" SRR1552444_hisat2.sam | cut -f5 | sort -n | uni
q -c
 907534 0
3171570 1
23827658 60
```

- ¿Cuántas letras distintas del alfabeto encontramos en la columna CIGAR del archivo **SRR1552444_hisat2.sam**? Indica cuáles son, sus cantidades y qué información proporcionan.
  Comando para contar los valores de CIGAR:
  grep -v "^@" SRR1552444_hisat2.sam |cut -f6 | grep -o [[:alpha:]] | sort -V | uniq -c

  Aparecen 5 letras distintas en la columna CIGAR: D,I,M,N y S.
  La letra D aparece 140911 veces, que significa que hay una deleción en la secuencia de lectura respecto al genoma de referencia.
  La letra I aparece 252141 veces, que significa que hay una inserción en la secuencia de lectura respecto al genoma de referencia.
  La letra M aparece 35510494 veces, que significa que hay alineamiento entre el genoma de referencia y la lectura en esa posición.
  La letra N aparece 8028904 veces, que significa que hay saltos o regiones omitidas de la referencia.
  La letra S aparece 1897487 veces, que significa que hay secuencias recortadas presentes.

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Processed]$ grep -v "^@" SRR1552444_hisat2.sam |cut -f6 | grep -o [[:alp
ha:]] | sort -V | uniq -c
 140911 D
 252141 I
35510494 M
8028904 N
1897487 S
```

P5. Como hemos visto en el archivo BAM, hay una ubicación cromosómica para cada lectura asignada. Ahora que ya hemos descubierto de dónde proviene cada lectura en el genoma, necesitamos anotar dicha información. Para ello obtén el archivo de anotaciones en formato GTF y responde a las siguientes cuestiones. (1,5 pts)

- ¿Cuántos y qué programas y bases de datos se emplearon para anotar este archivo? Para responder a esta pregunta, interroga la columna SOURCE.
  Comando para obtener los datos la columna SOURCE:
  grep -v "^##" gencode.vM10.annotation.gtf | cut -f2 | sort -u

  Dos tipos de programas y bases de datos: ENSEMBL y HAVANA
  239249 ENSEMBL
  1377387 HAVANA

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Annotation]$ grep -v "^##" gencode.vM10.annotation.gtf | cut -f2 | sort
-u
ENSEMBL
HAVANA
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Annotation]$
```

- ¿Cuáles y cuántas FEATURES podemos encontrar en el archivo de anotaciones?
  Comando para obtener las FEATURES:
  grep -v "^##" gencode.vM10.annotation.gtf | cut -f3 | sort -u
  Se encuentran 8 tipos distintos de FEATURES: CDS, exón, gene, selenocysteine, start_codon, stop_codon, transcript, UTR.

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Annotation]$ grep -v "^##" gencode.vM10.annotation.gtf | cut -f3 | sort
-u
CDS
exon
gene
Selenocysteine
start_codon
stop_codon
transcript
UTR
```

- ¿Qué porcentaje de GENES en el archivo GTF están ubicados en el cromosoma X?
  Comandos para obtener el porcentaje de genes que están ubicados en el cromosoma X:
  awk '$1 == "chrX" && $3 == "gene" {print $1, $3}' gencode.vM10.annotation.gtf | wc -l
  awk '$3 == "gene"' gencode.vM10.annotation.gtf | wc -l
  echo "scale=2; (2613/48440) *100" | bc

  Un 5% de los genes están ubicados en el cromosoma X.

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Annotation]$ awk '$1 == "chrX" && $3 == "gene" {print $1, $3}' gencode.v
M10.annotation.gtf | wc -l
2613
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Annotation]$ awk '$3 == "gene"' gencode.vM10.annotation.gtf | wc -l
48440
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Annotation]$ echo "scale=2; (2613/48440)*100" | bc
5.00
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Annotation]$
```

P6. Finalmente, ejecuta el recuento de los alineamientos sobre el archivo BAM con htseq-count. Reporta los comandos utilizados y computa sobre el archivo tsv final, el porcentaje total de lecturas asignadas, no_feature, ambiguous, too_low_aQual y not_aligned. Una vez computado estos porcentajes, muéstralos con un gráfico (usando cualquier lenguaje de programación) y comenta brevemente su resultado. (1,5 pts)

Comando utilizado para el recuento de los alineamientos:

htseq-count -t exon -i gene_id --stranded=no -f bam -r pos -s no SRR1552444_hisat2.sorted.bam ../../Annotation/gencode.vM10.annotation.gtf > ../../../Results/SR1552444_counts.tsv

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 03.Alignment]$ htseq-count -t exon -i gene_id --stranded=no -f bam -r po
s -s no SRR1552444_hisat2.sorted.bam ../../Annotation/gencode.vM10.annotation.gtf > ../../../Results/SR1552444_counts.tsv
100000 GFF lines processed.
200000 GFF lines processed.
300000 GFF lines processed.
400000 GFF lines processed.
```

Porcentaje de lecturas asignadas: 73,00%

Porcentaje no_feature: 8%

Porcentaje ambiguous: 3%

Porcentaje too_low_aQual: 11%

Porcentaje not_aligned: 2%

Comandos utilizados:

$ lecturas_asignadas=$(grep -v "^__" SRR1552444_counts.tsv | awk '{sum+=+2} END {print sum}')

$ lecturas_totales=$(awk '{sum+=+2} END {print sum}' SRR1552444_counts.tsv)

porcentaje= $(echo "scale=2; ($lecturas_asignadas/$lecturas_totales)*100" | bc)

echo $porcentaje

no_features=$(grep "__no_feature" SRR1552444_counts.tsv | awk '{print $2}')

porcentaje_no_feature=$(echo "scale=2; ($no_feature/$lecturas_totales)*100 | bc)

ambiguous=$(grep "ambiguous" SRR1552444_counts.tsv | awk '{print $2}')

porcentaje_ ambiguous =$(echo "scale=2; ($ambiguous /$lecturas_totales)*100 | bc)

too_low_aQual=$(grep "too_low_aQual" SRR1552444_counts.tsv | awk '{print $2}')

porcentaje_ too_low_aQual =$(echo "scale=2; ($too_low_aQual /$lecturas_totales)*100 | bc)

not_aligned=$(grep "not_aligned" SRR1552444_counts.tsv | awk '{print $2}')

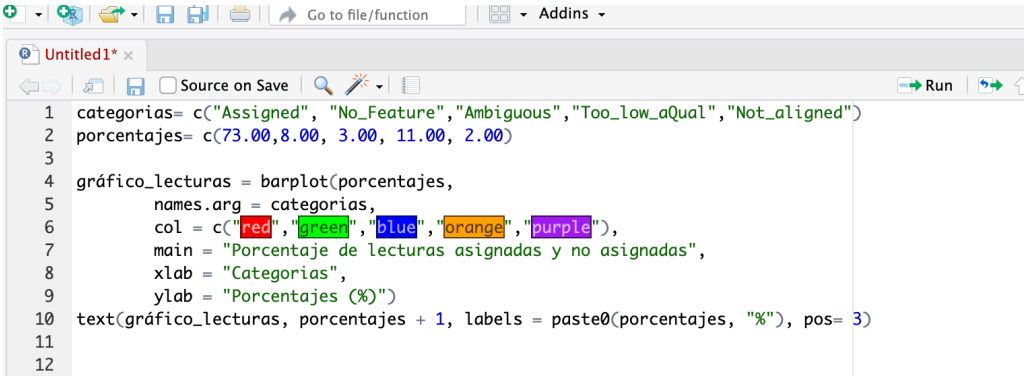
porcentaje_ not_aligned =$(echo "scale=2; ($not_aligned /$lecturas_totales)*100 | bc)

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ lecturas_asignadas=$(grep -v "^__" SR1552444_counts.tsv | awk
'{sum+=$2} END {print sum}')
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ echo $lecturas_asignadas
20482711
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ lecturas_totales=$(awk '{sum+=$2} END {print sum}' SR1552444_c
ounts.tsv)
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ echo $lecturas_totales
27906762
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ porcentaje=$(echo "scale=2; ($lecturas_asignadas/$lecturas_tot
ales)*100" | bc)
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ echo $porcentaje
73.00
```

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ no_feature=$(grep "__no_feature" SR1552444_counts.tsv | awk '{
print $2}')
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ ambiguous=$(grep "ambiguous" SR1552444_counts.tsv | awk '{prin
t $2}')
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ too_low_aQual=$(grep "too_low_aQual" SR1552444_counts.tsv | aw
k '{print $2}')
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ not_aligned=$(grep "not_aligned" SR1552444_counts.tsv | awk '{
print $2}')
```

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ porcentaje_no_feature=$(echo "scale=2; ($no_feature/$lecturas_
totales)*100" | bc)
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ echo $porcentaje_no_feature
8.00
```

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ porcentaje_ambiguous=$(echo "scale=2; ($ambiguous/$lecturas_to
tales)*100" | bc)
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ echo $porcentaje_ambiguous
3.00
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ porcentaje_too_low_aQual=$(echo "scale=2; ($too_low_aQual/$lec
turas_totales)*100" | bc)
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ echo $porcentaje_too_low_aQual
11.00
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ porcentaje_not_aligned=$(echo "scale=2; ($not_aligned/$lectura
s_totales)*100" | bc)
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ echo $porcentaje_no
$porcentaje_no_feature    $porcentaje_not_aligned
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$ echo $porcentaje_not_aligned
2.00
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Results]$
```



En el gráfico de la figura 7 se puede observar que el mayor porcentaje (un 73%) son lecturas que han alineado, es decir, que se han mapeado correctamente a una característica genómica, lo cual indica una buena calidad de las lecturas. El 8% de las lecturas no se han asignado a ninguna característica genómica. El 3% de las lecturas se clasificaron como ambiguas, es decir, que pueden alinear con varias características genómicas (varios genes, por ejemplo) y no se puede determinar a cuál pertenece, lo que genera ambigüedad, pero es un porcentaje muy bajo. El 11% de las lecturas no se han contado debido a que su calidad de alineamiento era baja, por lo que se podría revisar el filtrado de calidad para mejorarlo. Y solo un 2% de las lecturas no han alineado, es decir, un porcentaje muy bajo del total de lecturas no han alineado, el resto tienen una posible alineación.

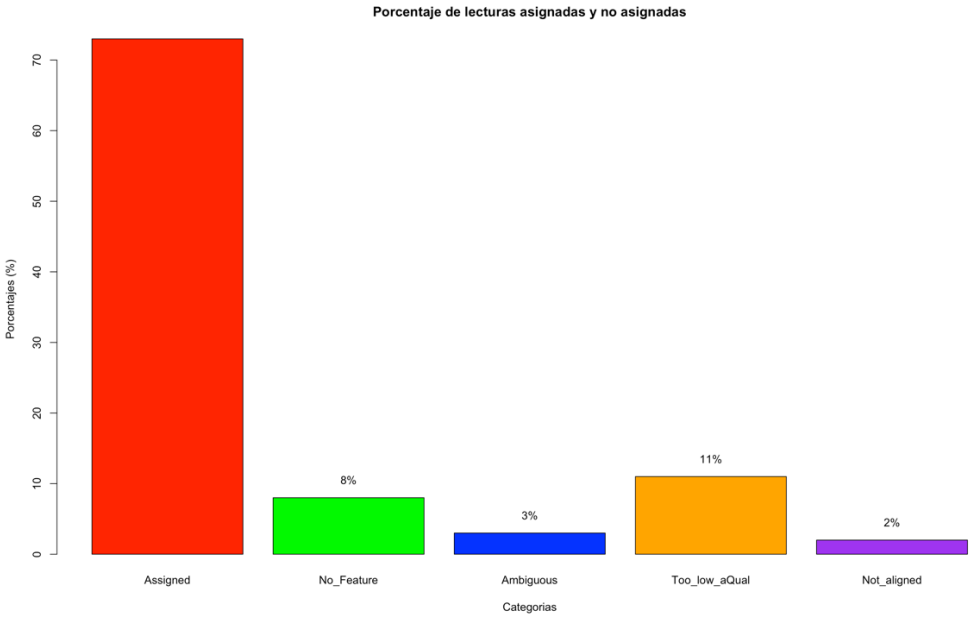


Figura 7. Gráfico del porcentaje de lecturas alineadas y el porcentaje de lecturas no asignadas.

P7 OPCIONAL. A lo largo del flujo de trabajo, existen etapas que no se han abordado en clase, como la identificación y marcaje de duplicados con MarkDuplicates(Picard), la evaluación de calidad del alineamiento con RseQC, o el uso de herramientas alternativas para el recorte, filtrado, mapeado y recuento de lecturas. En esta parte opcional de la actividad (que puede aportar hasta 0,5 puntos adicionales), se brinda al estudiante la libertad de seleccionar y aplicar alguna de estas etapas adicionales. Se solicita que, además de implementar la etapa seleccionada, se comparta tanto el código desarrollado como las observaciones principales derivadas de su aplicación.

Código empleado para utilizar la herramienta de marcaje de duplicados:

Picard MarkDuplicates I=SRR15524444_hisat2.sorted.bam 0=SRR1552444_hisat2.dedup.bam M=SRR1552444_hisat2.dedup.bam -M SRR1552444_hisat2_dedup.metrics.txt

cat SRR1552444_hisat2_dedup.metrics.txt

Al examinar el archivo de las métricas se obtienen algunos datos relevantes como el número de lecturas no emparejadas que se marcaron como duplicadas (unpaired_read_duplicates) que es de 15532298, el cual es alto. El porcentaje de lecturas duplicadas (percent_duplication) es de 57,3% (0.573378), es decir, más de la mitad de las lecturas se han marcado como duplicadas, lo cual es un porcentaje alto.

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 03.Alignment]$ picard MarkDuplicates I=SRR1552444_hisat2.sorted.bam O=SR
R1552444_hisat2.dedup.bam M=SRR1552444_hisat2_dedup.metrics.txt
INFO    2024-09-08 12:51:42     MarkDuplicates

********** NOTE: Picard's command line syntax is changing.
**********
********** For more information, please see:
********** https://github.com/broadinstitute/picard/wiki/Command-Line-Syntax-Transition-For-Users-(Pre-Transition)
**********
********** The command line looks like this in the new syntax:
**********
**********    MarkDuplicates -I SRR1552444_hisat2.sorted.bam -O SRR1552444_hisat2.dedup.bam -M SRR1552444_hisat2_dedup.metrics.txt
**********
```

```
(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 03.Alignment]$ cat SRR1552444_hisat2_dedup.metrics.txt
## htsjdk.samtools.metrics.StringHeader
# MarkDuplicates INPUT=[SRR1552444_hisat2.sorted.bam] OUTPUT=SRR1552444_hisat2.dedup.bam METRICS_FILE=SRR1552444_hisat2_dedup.metric
s.txt    MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000 MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=8000 SORTING_COLLECTION_SIZE_RATIO=0.25 TAG_D
UPLICATE_SET_MEMBERS=false REMOVE_SEQUENCING_DUPLICATES=false TAGGING_POLICY=DontTag CLEAR_DT=true DUPLEX_UMI=false ADD_PG_TAG_TO_RE
ADS=true REMOVE_DUPLICATES=false ASSUME_SORTED=false DUPLICATE_SCORING_STRATEGY=SUM_OF_BASE_QUALITIES PROGRAM_RECORD_ID=MarkDuplicat
es PROGRAM_GROUP_NAME=MarkDuplicates READ_NAME_REGEX=<optimized capture of last three ':' separated fields as numeric values> OPTICA
L_DUPLICATE_PIXEL_DISTANCE=100 MAX_OPTICAL_DUPLICATE_SET_SIZE=300000 VERBOSITY=INFO QUIET=false VALIDATION_STRINGENCY=STRICT COMPRES
SION_LEVEL=5 MAX_RECORDS_IN_RAM=500000 CREATE_INDEX=false CREATE_MD5_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.json USE_JDK_DEF
LATER=false USE_JDK_INFLATER=false
## htsjdk.samtools.metrics.StringHeader
# Started on: Sun Sep 08 12:51:42 CEST 2024

## METRICS CLASS        picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS_EXAMINED     SECONDARY_OR_SUPPLEMENTARY_RDS  UNMAPPED_READS  UNPAIRED_READ_DUPLICATES   R
EAD_PAIR_DUPLICATES     READ_PAIR_OPTICAL_DUPLICATES    PERCENT_DUPLICATION     ESTIMATED_LIBRARY_SIZE
Unknown Library 27089111        0       0       817651  15532298        0       0       0.573378


(05MBIF) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 03.Alignment]$
```