Tarea Unidad 4 - Sesion 3

Camilo Cabrera

30.12.2025

Contents


```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 8, fig.height = 4)
```

0.1 Introducción

La secuenciación de ARN (**RNA-seq**) representa el estándar actual para el estudio del transcriptoma a gran escala, permitiendo no solo la cuantificación de la abundancia de transcritos, sino también la comparación precisa de perfiles de expresión bajo condiciones que pueden ser comparadas. El trabajo en el que se basa el presente tutorial se centra en la arqueobacteria ***Sulfolobus acidocaldarius***, un modelo clave para entender la formación de biopelículas.

Para este análisis, se evaluó el impacto de una mutación tipo *knockdown* en el gen **Lrs14-like**, regulador fundamental en el desarrollo del fenotipo de biopelícula. El diseño experimental contempló un modelo factorial de cuatro librerías independientes:

1. **WildType_P** y **WildType_B**: Genotipo silvestre en medios planctónico y biopelícula, respectivamente.
2. **Mutant_P** y **Mutant_B**: Genotipo mutado bajo las mismas condiciones de cultivo.

Este reporte detalla el flujo de trabajo bioinformático empleado, con especial énfasis en el análisis de **expresión diferencial (Paso 8)**. No obstante, para garantizar la trazabilidad del proceso, se describen de manera sucinta las etapas preliminares de control de calidad, filtrado, alineamiento y cuantificación (Pasos 1–7) ejecutadas en el servidor institucional.

0.2 Métodos

0.2.1 Paso 1: Preparación del entorno de trabajo

Los análisis se realizaron en el servidor de uso común del curso (servidor bioinfo1), utilizando como entrada lecturas en formato **FASTQ** (4 librerías), el genoma de referencia de *S. acidocaldarius* en **FASTA** y su respectiva anotación en **GFF3**. El flujo de trabajo comenzó con la creación del entorno de trabajo, creando directorios independientes para garantizar la integridad y organización de los datos y resultados.

```
cd ccabrera/
mkdir -p RNA_seq/code
cd RNA_seq/code
```

0.2.2 Paso 2: Definición de carpetas de entrada

Se crearon unas variables que almacenan las rutas a las carpetas compartidas con los datos crudos, el genoma de referencia y la anotación génica. De esta manera reducir la escritura en los scripts y evitar posibles errores.

```
RAW=/home/bioinfo1/Tutorial_RNAseq/common/raw_data/
ANN=/home/bioinfo1/Tutorial_RNAseq/common/annot/
REF=/home/bioinfo1/Tutorial_RNAseq/common/ref_genome/
```

0.2.3 Paso 3: Definición de carpetas de salida

Luego, se definieron variables dirigidas a las carpetas donde se guardan los resultados generados en cada etapa del análisis (a excepción del análisis de expresión diferencial).

```
QC=../qc
FIL=../filtered
ALN=../alignment
CNT=../count
```

0.2.4 Paso 4: Control de calidad de las lecturas

Para el control de calidad de las lecturas crudas se usó el programa *IlluQC_PRLL.pl* del paquete **NGSQC Toolkit**. Como resultado, se obtuvieron reportes estadísticos y gráficos independientes para cada una de las cuatro librerías experimentales

```
mkdir $QC
mkdir "$QC/wild_planctonic" "$QC/wild_biofilm" "$QC/mut_planctonic" "$QC/mut_biofilm"

illuqc -se "$RAW/MW001_P.fastq" 5 A -onlystat -t 2 -o "$QC/wild_planctonic" -c 10 &
illuqc -se "$RAW/MW001_B3.fastq" 5 A -onlystat -t 2 -o "$QC/wild_biofilm" -c 10 &
illuqc -se "$RAW/0446_P.fastq" 5 A -onlystat -t 2 -o "$QC/mut_planctonic" -c 10 &
illuqc -se "$RAW/0446_B3.fastq" 5 A -onlystat -t 2 -o "$QC/mut_biofilm" -c 10 &
```

El análisis de los reportes obtenidos en el control de calidad permiten realizar el paso siguiente que es el filtrado de secuencias.

0.2.5 Paso 5: Filtrado de secuencias

Considerando el control de calidad, las lecturas se filtraron eliminando las que presentaron un puntaje de calidad PHRED menor de 20 en al menos un 80% del total de longitud. Este filtrado permitió mejorar la calidad de las lecturas utilizadas para el alineamiento.

```
mkdir $FIL
mkdir "$FIL/wild_planctonic" "$FIL/wild_biofilm" "$FIL/mut_planctonic" "$FIL/mut_biofilm"

illuqc -se "$RAW/MW001_P.fastq" 5 A -l 80 -s 20 -t 2 -o "$FIL/wild_planctonic" -c 1 &
illuqc -se "$RAW/MW001_B3.fastq" 5 A -l 80 -s 20 -t 2 -o "$FIL/wild_biofilm" -c 1 &
illuqc -se "$RAW/0446_P.fastq" 5 A -l 80 -s 20 -t 2 -o "$FIL/mut_planctonic" -c 1 &
illuqc -se "$RAW/0446_B3.fastq" 5 A -l 80 -s 20 -t 2 -o "$FIL/mut_biofilm" -c 1 &
```

0.2.6 Paso 6: Alineamiento al genoma de referencia

Las lecturas filtradas se contrastaron con la secuencia del genoma de referencia de ***S. acidocaldarius***, permitiendo determinar las coordenadas exactas de cada transcrito. Como resultado de esta fase, se generaron archivos en formato **SAM** (*Sequence Alignment Map*), los cuales contienen la información detallada de la posición y la calidad del mapeo para cada una de las librerías experimentales

```
mkdir $ALN

bwa078 mem "$REF/genome.fasta" -t 1 "$FIL/wild_planctonic/MW001_P.fastq_filtered" > "$ALN/MW001_P_aligne
bwa078 mem "$REF/genome.fasta" -t 1 "$FIL/wild_biofilm/MW001_B3.fastq_filtered" > "$ALN/MW001_B3_aligned
bwa078 mem "$REF/genome.fasta" -t 1 "$FIL/mut_planctonic/0446_P.fastq_filtered" > "$ALN/0446_P_aligned.s
bwa078 mem "$REF/genome.fasta" -t 1 "$FIL/mut_biofilm/0446_B3.fastq_filtered" > "$ALN/0446_B3_aligned.sa
```

0.2.7 Paso 7: Estimación de abundancia génica

Se cuantificó la expresión génica utilizando **HTSeq-count**, integrando los archivos de mapeo con la anotación genómica en formato **GFF3**. Este procedimiento derivó en tablas de conteo discretas por gen para cada una de las condiciones del estudio, facilitando la identificación de cambios en los niveles de expresión bajo diferentes contextos biológicos y genotípicos.

```
mkdir $CNT

python -m HTSeq.scripts.count -t Gene -i GenID "$ALN/MW001_P_aligned.sam" "$ANN/saci.gff3" > "$CNT/MW00
python -m HTSeq.scripts.count -t Gene -i GenID "$ALN/MW001_B3_aligned.sam" "$ANN/saci.gff3" > "$CNT/MW0
python -m HTSeq.scripts.count -t Gene -i GenID "$ALN/0446_P_aligned.sam" "$ANN/saci.gff3" > "$CNT/0446_
python -m HTSeq.scripts.count -t Gene -i GenID "$ALN/0446_B3_aligned.sam" "$ANN/saci.gff3" > "$CNT/0446
```

Los archivos de conteo creados durante este paso, se utilizarán para el análisis de expresión diferencial.

0.3 8. Prueba de expresión diferencial

0.3.1 8.1 Creación de directorios y carpetas de salida

```
# Definir la ruta donde estan los archivos .count
input_dir <- "/home/bioinfo1/RNA_seq/count"

# Definir el directorio personal de trabajo
user_dir <- "/home/bioinfo1/ccabrera/Tareas_BioinfRepro2025_CDCG/Tarea_4.3/RNA_seq"

# Definir y crear carpetas dentro del directorio
output_pseudo <- file.path(user_dir, "results", "pseudocounts")
output_histogram <- file.path(user_dir, "results", "histograms")
output_pvalue_fdr <- file.path(user_dir, "results", "pvalue_fdr")
output_table <- file.path(user_dir, "results", "tables")

# Crear las carpetas automáticamente si no existen
if(!dir.exists(output_pseudo))    dir.create(output_pseudo, recursive = TRUE)
if(!dir.exists(output_histogram)) dir.create(output_histogram, recursive = TRUE)
if(!dir.exists(output_pvalue_fdr)) dir.create(output_pvalue_fdr, recursive = TRUE)
if(!dir.exists(output_table))      dir.create(output_table, recursive = TRUE)

# Cargar la librería
library(edgeR)
```

0.3.2 8.2 Cargar y procesar archivos de entrada

```
# Leer archivos de entrada y asignarles nombres a sus columnas.
wild_p <- read.delim(file=file.path(input_dir, "MW001_P.count"), sep="\t", header = F, check=F); colnam
wild_b <- read.delim(file=file.path(input_dir, "MW001_B3.count"), sep="\t", header = F, check=F); colna
mut_p <- read.delim(file=file.path(input_dir, "0446_P.count"), sep="\t", header = F, check=F); colnames
mut_b <- read.delim(file=file.path(input_dir, "0446_B3.count"), sep="\t", header = F, check=F); colname

# Juntar los cuatro set de datos.
rawcounts <- data.frame(wild_p$Gen_ID, WildType_P = wild_p$Count, WildType_B = wild_b$Count, Mutant_P =

# Calcular RPKM
rpkm <- cpm(rawcounts)

# Remover filas que no serán utilizadas y aquellos genes con un valor de RPKM menor a 1, en tres de las
to_remove <- rownames(rawcounts) %in% c("__no_feature","__ambiguous","__too_low_aQual","__not_aligned",
keep <- rowSums(rpkm > 1) >= 3 & !to_remove
rawcounts <- rawcounts[keep,]
```

0.3.3 8.3 Expresión Diferencial para Medios de Cultivo

```
# Crear un vector que agrupará las muestras según el medio de cultivo a las que fueron sometidas.
group_culture <- c("planctonic","biofilm","planctonic","biofilm")

# Crear un objeto del tipo DGE, a través del cual se realizarán los cálculos para identificar genes dif
dge_culture <- DGEList(counts = rawcounts, group = group_culture)

# Calcular factor de normalización. Esto va a permitir, posteriormente, normalizar los valores de conte
dge_culture <- calcNormFactors(dge_culture)

# Estimar dos valores de dispersión, uno para cada gen y otro para cada librería. Esto es necesario par
dge_culture <- estimateCommonDisp(dge_culture)
dge_culture <- estimateTagwiseDisp(dge_culture)

# Realizar prueba de expresión diferencial.
de_culture <- exactTest(dge_culture, pair = c("planctonic","biofilm"))

# Obtener una tabla resumen de resultados.
results_culture <- topTags(de_culture, n = nrow(dge_culture))
results_culture <- results_culture$table

# Obtener ID de genes diferencialmente expresados
ids_culture <- rownames(results_culture[results_culture$FDR < 0.1,])
```

0.3.4 8.4 Expresión Diferencial para Genotipos

```
# Crear un set de conteos que no considere los genes diferencialmente expresados por Medio de Cultivo.
rawcounts_genotype <- rawcounts[!rownames(rawcounts) %in% ids_culture,]

# Replicar los pasos descritos en 8.3
group_genotype <- c("wildtype","wildtype","mutant","mutant")
dge_genotype <- DGEList(counts = rawcounts_genotype, group = group_genotype)
dge_genotype <- calcNormFactors(dge_genotype)
dge_genotype <- estimateCommonDisp(dge_genotype)
dge_genotype <- estimateTagwiseDisp(dge_genotype)
de_genotype <- exactTest(dge_genotype, pair = c("wildtype","mutant"))
results_genotype <- topTags(de_genotype, n = nrow(de_genotype))
results_genotype <- results_genotype$table
ids_genotype <- rownames(results_genotype[results_genotype$FDR < 0.1,])
```

0.3.5 8.5 Resultados

```r
# Definir vectores del tipo Booleano que, a partir del set completo de genes, etiquetará aquellos que p
de_genes_culture  <- rownames(rawcounts) %in% ids_culture
de_genes_genotype <- rownames(rawcounts) %in% ids_genotype


# Obtener pseudoconteos y transformarlos a escala logarítmica.
pseudocounts <- data.frame(rownames(rawcounts), WildType_P = log10(dge_culture$pseudo.counts[,1]), Wild

# Gráficar los pseudo conteos para cada gen.
# 1. Medio de Cultivo
png(filename = file.path(output_pseudo, "pair_expression_culture.png"),
    width = 8, height = 4, units = "in", res = 300) # Alta resolución (300 dpi)
par(mfrow = c(1,2))
plot(pseudocounts$WildType_P, pseudocounts$WildType_B, col = ifelse(pseudocounts$DE_C, "red", "blue"),
     main = "Wild Type", xlab = "Planctonic", ylab = "Biofilm")
abline(lsfit(pseudocounts$WildType_P, pseudocounts$WildType_B), col = "black")
plot(pseudocounts$Mutant_P, pseudocounts$Mutant_B, col = ifelse(pseudocounts$DE_C, "red", "blue"),
     main = "Mutant", xlab = "Planctonic", ylab = "Biofilm")
abline(lsfit(pseudocounts$Mutant_P, pseudocounts$Mutant_B), col = "black")
dev.off()
# 2. Genotipo
png(filename = file.path(output_pseudo, "pair_expression_genotype.png"),
    width = 8, height = 4, units = "in", res = 300)
par(mfrow = c(1,2))
plot(pseudocounts$WildType_P, pseudocounts$Mutant_P, col = ifelse(pseudocounts$DE_G, "red", "blue"),
     main = "Planctonic", xlab = "Wild Type", ylab = "Mutant")
abline(lsfit(pseudocounts$WildType_P, pseudocounts$Mutant_P), col = "black")
plot(pseudocounts$WildType_B, pseudocounts$Mutant_B, col = ifelse(pseudocounts$DE_G, "red", "blue"),
     main = "Biofilm", xlab = "Wild Type", ylab = "Mutant")
abline(lsfit(pseudocounts$WildType_B, pseudocounts$Mutant_B), col = "black")
dev.off()
# 3. Histogramas valor p
png(filename = file.path(output_histogram, "histograms_pvalue.png"),
    width = 8, height = 4, units = "in", res = 300)
par(mfrow = c(1,2))
hist(x = results_culture$PValue, col = "skyblue", border = "blue", main = "Culture", xlab = "P-value")
hist(x = results_genotype$PValue, col = "skyblue", border = "blue", main = "Genotype", xlab = "P-value"
dev.off()
# 4. P-value vs FDR
png(filename = file.path(output_pvalue, "pvalue_fdr.png"),
    width = 8, height = 4, units = "in", res = 300)
par(mfrow = c(1,2))
plot(results_culture$PValue, results_culture$FDR, col = "blue", main = "Culture", xlab = "P-value", ylab
plot(results_genotype$PValue, results_genotype$FDR, col = "blue", main = "Genotype", xlab = "P-value", y
dev.off()
```

0.3.6 8.5.1 Visualización de gráficos de pseudoconteos, histogramas de valores P y P vs FDR

A continuación se incluyen los gráficos generados durante el análisis (Paso 8):

```r
library(knitr)

figs_png <- c(
  file.path(user_dir, "results", "pseudocounts", "pair_expression_culture.png"),
  file.path(user_dir, "results", "pseudocounts", "pair_expression_genotype.png"),
  file.path(user_dir, "results", "histograms",  "histograms_pvalue.png"),
  file.path(user_dir, "results", "pvalue_fdr",  "pvalue_fdr.png")
)
# Incluir las figuras
knitr::include_graphics(figs_png)
```

0.3.7 8.5.2 Resultados: expresión diferencial por medio de cultivo

```r
culture_path <- file.path(output_table,"table_de_genes_culture.csv")

results_culture <- read.delim(
  culture_path,
  sep = "\t",
  header = TRUE,
  row.names = 1,
  check.names = FALSE
)

head(results_culture)
```

0.3.8 8.5.3 Resultados: expresión diferencial por genotipo

```r
genotype_path <- file.path(output_table,"table_de_genes_genotype.csv")

results_genotype <- read.delim(
  genotype_path,
  sep = "\t",
  header = TRUE,
  row.names = 1,
  check.names = FALSE
)

head(results_genotype)
```

0.4 Resultados y discusión

0.4.1 Expresión diferencial asociada al medio de cultivo (planctónico vs biopelícula)

El análisis transcriptómico mediante `edgeR` permitió desglosar la varianza de la expresión génica en dos componentes principales: el efecto del entorno (**medio de cultivo**) y el efecto del estado genético (**genotipo**).

0.4.2 3.1. Impacto Dominante del Medio de Cultivo (Planctónico vs. Biopelícula)

Los resultados indican que el medio de cultivo es el factor determinante en la reprogramación del transcriptoma de *S. acidocaldarius*.

- **Magnitud del Cambio:** Se observaron genes con valores de $|log_2FC|$ significativamente elevados. La presencia de genes con cambios de expresión tan marcados sugiere una respuesta biológica robusta ante la transición al estado sésil (biopelícula).
- **Significancia Estadística:** Los genes prioritarios presentan valores de $p-value$ extremadamente bajos (10^{-24} a 10^{-10}), los cuales mantienen su significancia tras la corrección por múltiples comparaciones ($FDR \ll 0.1$).
- **Coherencia Biológica:** Este perfil es consistente con la literatura; la formación de biopelículas requiere una reorganización celular profunda que abarca desde la síntesis de exopolisacáridos y proteínas de adhesión hasta ajustes metabólicos críticos para la supervivencia en la matriz extracelular.

Análisis Visual (Pseudoconteos y Distribución)

Los gráficos de dispersión de pseudoconteos confirman esta tendencia:

1. **Consistencia:** El patrón de expresión diferencial es análogo tanto en el genotipo *Wild Type* como en el *Mutante*, lo que posiciona al medio de cultivo como un factor de variabilidad "dominante" e independiente de la mutación en Lrs14-like.
2. **Distribución:** Mientras la mayoría de los genes mantienen una expresión basal (cerca de la diagonal), un subconjunto claro se desplaza significativamente, indicando una inducción (puntos sobre la diagonal) o represión (puntos bajo la diagonal) específica en biopelícula.
3. **Validación Global:** El histograma de $p-values$ para esta comparación muestra un sesgo pronunciado hacia el cero, comportamiento típico de un experimento con un efecto biológico real y masivo.

0.4.3 Expresión diferencial asociada al genotipo (wild type vs mutante)

A diferencia del impacto ambiental, la comparación entre el genotipo *Wild Type* y el mutante *knockdown* de Lrs14-like reveló un efecto mucho más sutil y localizado.

- **Evidencia Estadística Débil:** Tras excluir los genes afectados por el medio de cultivo, los genes restantes mostraron valores de $logFC$ de menor magnitud y valores de FDR menos competitivos.
- **Distribución de Probabilidades:** El histograma de $p-values$ para la comparación por genotipo presenta una distribución cuasi-uniforme. La ausencia de un pico claro cerca del cero sugiere que no existe una diferencia transcriptómica global masiva entre ambos genotipos bajo las condiciones probadas.

- **Interpretación Biológica:** Estos hallazgos sugieren que el rol del gen Lrs14-like podría estar restringido a rutas metabólicas muy específicas o a etapas muy concretas de la formación de la biopelícula, en lugar de actuar como un regulador maestro de amplio espectro.

0.5 Conclusiones

El flujo de trabajo implementado permitió concluir que:

1. **El entorno es el principal motor de cambio:** La transición entre medio planctónico y biopelícula induce una respuesta transcripcional masiva y consistente en *S. acidocaldarius*.
2. **Especificidad del Mutante:** El efecto del *knockdown* de Lrs14-like es secundario y discreto en comparación con el efecto del medio, sugiriendo un rol regulador más especializado o sutil.

Todos los datos crudos, tablas de expresión y visualizaciones que sustentan este análisis se encuentran disponibles en la carpeta `results/` del directorio de trabajo.