Tarea 4.3 – RNA-seq Differential Expression

Pamela González

2025-12-17

Contents

0.1 Introducción

La secuenciación de RNA (RNA-seq) es una de las técnicas más utilizadas actualmente para estudiar la expresión génica a gran escala, ya que permite cuantificar la abundancia de transcritos y comparar cómo varía la expresión de los genes bajo distintas condiciones experimentales. A partir de este tipo de análisis es posible identificar genes diferencialmente expresados y relacionarlos con procesos biológicos específicos.

En este trabajo se analizan datos de RNA-seq correspondientes a la arqueobacteria *Sulfolobus acidocaldarius*. En este organismo se introdujo una mutación tipo *knockdown* en el gen *Lrs14-like*, el cual ha sido previamente descrito como un regulador importante en la formación de biopelículas. El objetivo del estudio es evaluar el impacto de esta mutación a nivel transcripcional y determinar qué genes se asocian al crecimiento en biopelícula de manera dependiente o independiente del gen mutado.

Para ello, *S. acidocaldarius* fue cultivada bajo dos condiciones experimentales distintas: crecimiento en medio planctónico y crecimiento en biopelícula. Cada una de estas condiciones fue evaluada tanto en el genotipo *wildtype* como en el mutante, generando cuatro librerías independientes: WildType_P, WildType_B, Mutant_P y Mutant_B. A partir de estas librerías se aplicó un flujo de trabajo estándar de análisis de RNA-seq que permitió, finalmente, identificar genes diferencialmente expresados entre condiciones y genotipos.

Si bien el foco principal de este informe corresponde al análisis de expresión diferencial (Paso 8), a continuación se describen también las etapas previas del pipeline (Pasos 1–7), las cuales fueron ejecutadas previamente en el servidor, con el fin de contextualizar el análisis y entregar una visión completa del flujo de trabajo utilizado.

0.2 Métodos

0.2.1 Paso 1: Preparación del entorno de trabajo

El análisis se realizó en un servidor de cómputo que cuenta con los programas y dependencias necesarias previamente instaladas. Los datos de entrada incluyen cuatro librerías de lecturas en formato FASTQ, un

genoma de referencia de *S. acidocaldarius* en formato FASTA y un archivo de anotación génica en formato GFF3.

Como primer paso, cada usuario creó su propio directorio de trabajo, dentro del cual se organizaron los archivos y resultados del análisis.

```
cd <mi_usuario>
mkdir -p RNA_seq/code
cd RNA_seq/code
```

0.2.2 Paso 2: Definición de carpetas de entrada

Se definieron variables que almacenan las rutas a las carpetas compartidas que contienen los datos crudos, el genoma de referencia y la anotación génica.

```
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

Posteriormente, se definieron variables que apuntan a las carpetas donde se almacenarían los resultados generados en cada etapa del análisis.

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

0.2.4 Paso 4: Control de calidad de las lecturas

El control de calidad de las lecturas crudas se realizó utilizando el programa *IlluQC_PRLL.pl* del paquete **NGSQC Toolkit**. Esta herramienta permite evaluar parámetros como la calidad de las bases, el contenido de GC y la distribución de calidades a lo largo de las secuencias. Para cada librería se generó un reporte independiente.

```
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 &
```

Los reportes generados fueron utilizados para definir los criterios de filtrado aplicados en el paso siguiente.

0.2.5 Paso 5: Filtrado de secuencias

En base a los resultados del control de calidad, las lecturas fueron filtradas eliminando aquellas que presentaban un puntaje de calidad PHRED menor a 20 en al menos el 80% de su 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 fueron alineadas contra el genoma de referencia de *S. acidocaldarius* utilizando el algoritmo **BWA-MEM**, generando archivos de alineamiento en formato SAM para cada librería.

```
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_aligne
bwa078 mem "$REF/genome.fasta" -t 1 "$FIL/mut_planctonic/0446_P.fastq_filtered" > "$ALN/0446_P_aligned.
bwa078 mem "$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

La cuantificación de lecturas asignadas a cada gen se realizó utilizando el programa **HTSeq-count**, empleando el archivo de anotación génica en formato GFF3 como referencia. Este paso permitió generar archivos de conteo por gen para cada una de las condiciones experimentales, los cuales fueron utilizados posteriormente como entrada para el análisis de expresión diferencial.

```
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 generados en esta etapa constituyen la base para la prueba de expresión diferencial desarrollada en el Paso 8.

0.3 8. Prueba de expresión diferencial

0.3.1 8.1 Directorios y verificación de resultados

```
## Directorio de resultados (buscando aquí):

## C:/Users/pame5/Desktop/Tarea4.3/results/diff_expr

## ¿Existe output_base?:  TRUE

## Contenido de output_base:

## [1] "histograms"   "pseudocounts" "pvalue_fdr"   "tables"

##
## Contenido de tables:

## [1] "table_de_genes_culture.csv"  "table_de_genes_genotype.csv"

##
## Contenido de pseudocounts:

## [1] "pair_expression_culture.pdf"  "pair_expression_genotype.pdf"

##
## Contenido de histograms:

## [1] "histograms_pvalue.pdf"
```

0.3.2 8.2 Resultados: expresión diferencial por medio de cultivo

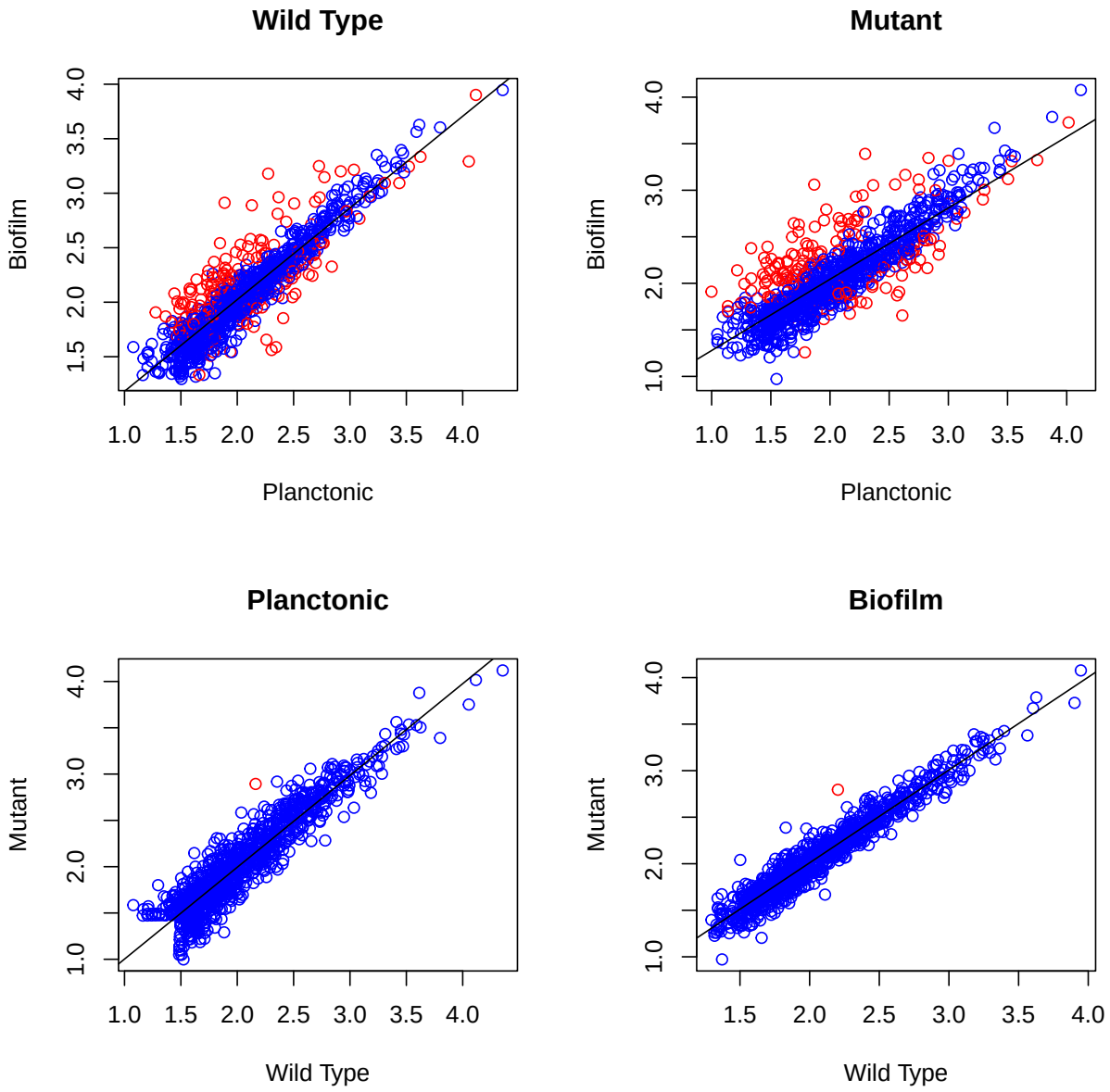
```
##               logFC    logCPM       PValue          FDR
## Saci_1717  3.703873 11.036075 7.905020e-24 7.849685e-21
## Saci_1078  3.361133 12.065425 1.573246e-19 7.811168e-17
## Saci_2035 -2.898926  9.374041 4.461542e-13 1.476770e-10
## Saci_1952  2.393023  9.191711 1.165563e-10 2.432183e-08
## Saci_1953  2.154735 11.067516 1.224664e-10 2.432183e-08
## Saci_1226  2.378578 11.133507 4.026676e-10 6.664149e-08
```

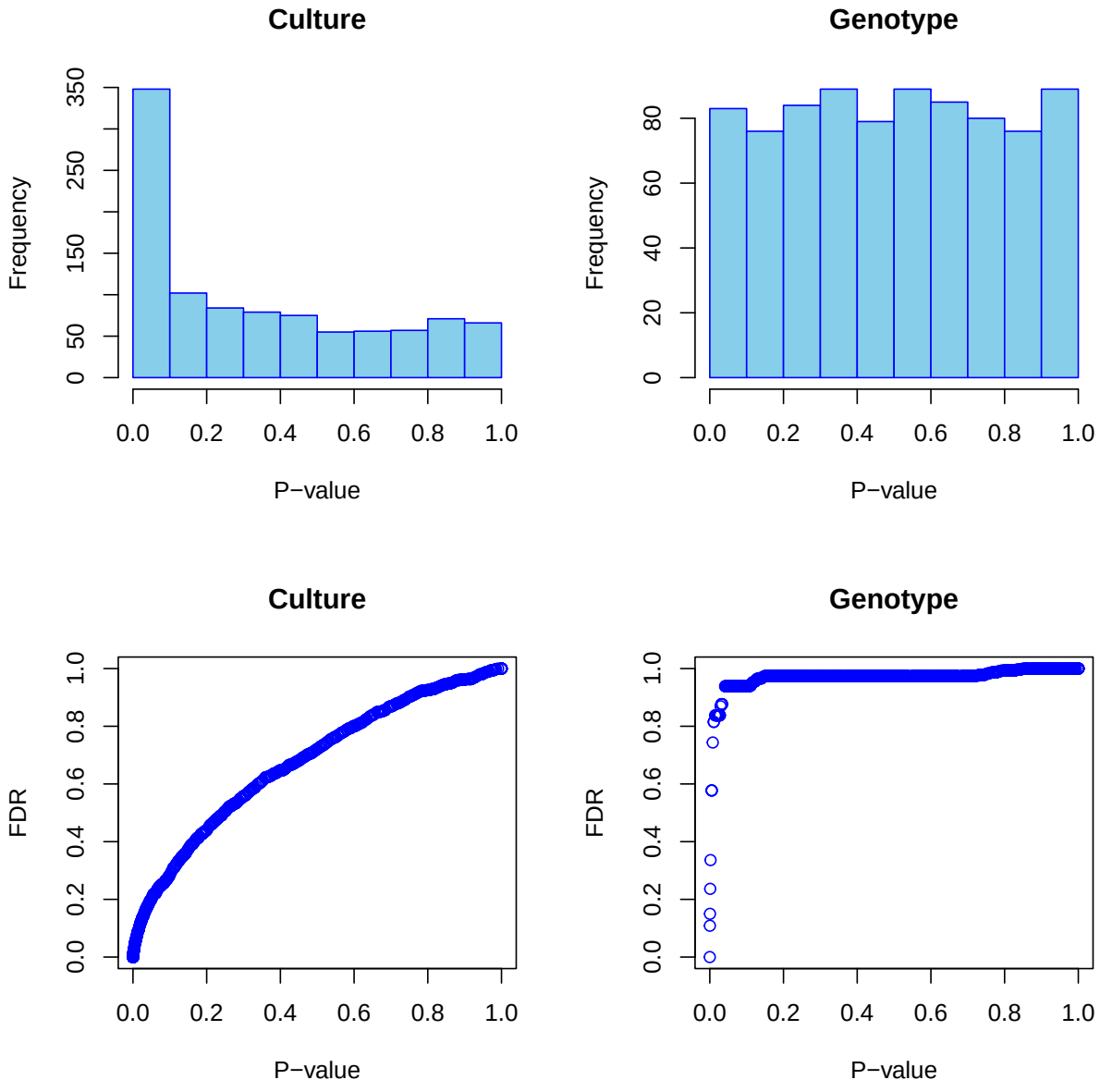
0.3.3 8.3 Resultados: expresión diferencial por genotipo

```
##               logFC    logCPM       PValue          FDR
## Saci_2195  2.206850 11.092771 1.029273e-10 8.542966e-08
## Saci_2207  1.369328  9.894604 2.617151e-04 1.086118e-01
## Saci_2057 -1.143487  8.872415 5.416287e-04 1.498506e-01
## Saci_1180  1.487972  8.178501 1.139288e-03 2.364022e-01
## Saci_1894 -1.099476 12.152496 2.024716e-03 3.361029e-01
## Saci_0317 -1.245367  7.653952 4.711503e-03 5.774394e-01
```

0.3.4 8.4 Visualización

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





0.4 Resultados y discusión

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

El análisis de expresión diferencial realizado con **edgeR** evidencia que el **medio de cultivo** constituye el principal factor asociado a cambios en la expresión génica. En la tabla de resultados correspondiente a la comparación planctónico vs biopelícula se observan genes con valores de **logFC elevados**, tanto positivos como negativos, lo que indica cambios de expresión de magnitud considerable entre ambas condiciones.

Los valores de **PValue** para estos genes son extremadamente bajos (del orden de 10^{-2} a 10^{-1} en las primeras filas), y se mantienen significativos luego de la corrección por múltiples comparaciones, con **FDR muy**

inferiores al umbral de 0.1. Esto indica que una fracción relevante de los genes presenta diferencias de expresión robustas y estadísticamente confiables asociadas al cambio de condición de cultivo.

Este resultado es coherente con el contexto biológico del experimento, ya que el crecimiento en biopelícula implica una reorganización profunda del estado celular, incluyendo procesos como adhesión, formación de matriz extracelular, cambios metabólicos y adaptación al estrés ambiental. Por lo tanto, es esperable que la transición desde un crecimiento planctónico a biopelícula se refleje en un perfil transcripcional marcadamente distinto.

Los gráficos de dispersión de pseudoconteos (Wild Type y Mutant) refuerzan esta observación. En ambos casos, la mayoría de los genes se distribuye alrededor de la diagonal, lo que representa niveles de expresión similares entre planctónico y biopelícula. Sin embargo, se observa un subconjunto claro de genes que se separa de la diagonal y que corresponde a genes diferencialmente expresados, marcados visualmente en un color distinto. Los genes ubicados por encima de la diagonal presentan mayor expresión relativa en biopelícula, mientras que aquellos por debajo muestran mayor expresión en planctónico.

Es importante destacar que este patrón se observa tanto en el genotipo **wild type** como en el **mutante**, lo que sugiere que el efecto del medio de cultivo es consistente e independiente del estado del gen *Lrs14-like*. En este sentido, el medio de cultivo emerge como un factor dominante en la variabilidad global de la expresión génica.

El histograma de valores p para la comparación por medio de cultivo muestra una **acumulación pronunciada de valores cercanos a cero**, lo cual es característico de escenarios donde existe un número considerable de genes verdaderamente diferencialmente expresados. En conjunto, tanto las tablas como las visualizaciones apoyan la conclusión de que el crecimiento en biopelícula induce cambios transcripcionales amplios y estadísticamente significativos en *S. acidocaldarius*.

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

En contraste con los resultados obtenidos para el medio de cultivo, la comparación por **genotipo** (wild type vs mutante), realizada tras excluir los genes asociados al efecto de culture, muestra un patrón mucho menos marcado. Si bien se identifican genes con valores de logFC distintos de cero, la magnitud de estos cambios es, en general, menor que la observada en la comparación por medio de cultivo.

Además, los valores de **FDR** para los genes listados en la tabla de resultados por genotipo no alcanzan niveles tan bajos como en la comparación por culture, lo que sugiere que la evidencia estadística para expresión diferencial asociada exclusivamente al genotipo es más débil. Esto se ve reflejado también en el histograma de valores p para "Genotype", el cual presenta una distribución cercana a uniforme, sin una acumulación evidente de valores bajos.

Desde un punto de vista biológico, estos resultados indican que el knockdown del gen *Lrs14-like* no genera, al menos en estas condiciones experimentales, una reprogramación transcripcional global comparable a la inducida por el cambio de medio de cultivo. Es posible que el efecto del gen mutado esté restringido a un subconjunto específico de genes o rutas funcionales, particularmente relacionadas con la formación de biopelículas, y que dichos efectos sean más sutiles a nivel transcriptómico global.

Otro aspecto relevante es que el diseño experimental no incluye réplicas biológicas por condición, lo que limita la capacidad estadística para detectar cambios de expresión de menor magnitud. En este contexto, es esperable que solo efectos fuertes sean identificados como estadísticamente significativos, especialmente en la comparación por genotipo.

En conjunto, los resultados sugieren que el **medio de cultivo** explica la mayor parte de la variabilidad observada en la expresión génica, mientras que el **genotipo** presenta un efecto secundario y más acotado bajo el criterio estadístico utilizado en este análisis.

0.4.3 Consideraciones y limitaciones

La ausencia de réplicas biológicas constituye una limitación importante del análisis, ya que restringe la estimación adecuada de la variabilidad y reduce la potencia estadística, especialmente para detectar cambios moderados de expresión. No obstante, el efecto del medio de cultivo es lo suficientemente fuerte como para ser detectado de manera clara y consistente, lo que refuerza la solidez de las conclusiones asociadas a esta comparación.

0.5 Conclusiones

El análisis de expresión diferencial permitió identificar genes con cambios significativos asociados al medio de cultivo y al genotipo. Los resultados y visualizaciones incluidos en este informe corresponden a las tablas y figuras generadas durante la ejecución del paso 8 y almacenadas en la carpeta `results/diff_expr`.