

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

Resolución Actividad 2, Secuenciación y Ómicas de Próxima Generación, Máster Bioinformática UNIR (2025). Análisis de expresión diferencial de genes relacionados con la obesidad mediante RNA-seq

Oriana Batista Ceballos

2025-12-21

I. Introducción

- En esta primera parte de la resolución de la actividad dos del curso de Secuenciación y Ómicas de Próxima Generación se plasmará el contenido relacionado con la ejecución de los scripts, los cuales serán útiles para lograr cada uno de los objetivos establecidos para el análisis de la expresión diferencial de genes relacionados con la obesidad mediante el método de RNA-seq para los grupos obesos uno y dos, de cuya expresión génica se hará un análisis comparativo y, además, se tratará de establecer asociaciones entre los genes en estudio y la obesidad. La primera parte que contiene los scripts, escritos en código R, consta de tres documentos para evitar conflictos entre las bibliotecas. En el primer documento denominado `Batista_CeballosAct2PruebaFINAL1.Rmd` se incluye el script 1 con información acerca del control de calidad, alineamiento, cuantificación, normalización y la expresión diferencial de los genes relacionados con la obesidad incluyendo la visualización de la misma mediante los gráficos Volcano Plot y MA-Plot. En un segundo documento, con el nombre `Batista_CeballosAct2PruebaFINAL2B.Rmd`, se incluye el Script 2 con información para la visualización de la

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

expresión génica mediante el gráfico de Heatmap y en un tercer documento, Batista_CeballosAct2PruebaFINAL3A.Rmd, se incluye el Script 3, el cual presenta información acerca de los genes correlacionados y la asociación de éstos con las características fenotípicas, así como un análisis de enriquecimiento funcional y rutas metabólicas. La totalidad de los resultados generados, de acuerdo con cada uno de los objetivos planteados, serán presentados en la parte dos, de la resolución de esta actividad, en la forma de un póster científico.

A. Objetivos

- Esta actividad tiene como propósito familiarizar al estudiante con el análisis de expresión diferencial a partir de datos de RNA-seq, aplicando una comparativa sobre un conjunto de muestras simuladas.
 1. Realizar un análisis de calidad y alineamiento de lecturas RNA-seq de los grupos de obesos uno y dos.
 2. Cuantificar niveles de expresión génica y normalizar los datos de los grupos de obesos uno y dos.
 3. Identificar genes diferencialmente expresados entre los grupos de obesos uno y dos.
 4. Interpretar perfiles de expresión en función del fenotipo o condición en los grupos de obesos uno y dos.
 5. Analizar la funcionalidad de los genes estudiados en los grupos de obesos uno y dos y su participación en rutas metabólicas.
 6. Representar los resultados obtenidos en los grupos de obesos uno y dos en un póster científico.

B. Pautas de elaboración

Por ser una estudiante internacional me correspondió analizar y comparar la expresión génica de los grupos obesos uno y dos.

1. **Herramientas recomendadas:** se recomendó el uso de las siguientes herramientas para la realización de las etapas de análisis:

a. Control de calidad

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

FastQC, MultiQC, Fastp, Trimmomatic

b. Alineamiento

Salmon, STAR, Bowtie2, HISAT2

c. Cuantificación

Salmón, featureCounts, tximport

d. Expresión diferencial

DESeq2, EdgeR, limma-voom

e. Visualización

Ggplot2, EnhancedVolcano, pheatmap

f. Enriquecimiento

clusterProfiler, Enrichr, gprofiler

C. Instrucciones

Datos proporcionados. Se entregó un conjunto de archivos de secuencias con formato fastq, los cuales simulan datos de RNA-seq de 5 personajes del universo de los Simpson que forman parte de los grupos obesos uno y dos. Cada archivo representa una muestra con un perfil de expresión distinto, generado artificialmente para modelar diferencias genéticas relacionadas con la obesidad. Adicionalmente, se facilitaron otros archivos que incluyeron una secuencia de referencia, transcritos correspondientes a la secuencia de referencia, el listado de 37 genes utilizados en la simulación, una tabla con información biológica y extensión cvs y un archivo denominado Transcrito_a_Gen.tsv. Los personajes agrupados en los perfiles metabólicos mencionados son los siguientes:

- Grupo Obeso1 (Familia 1, Sobrepeso/Obeso): Abraham Simpson y Homer Simpson.
- Grupo Obeso2 (Familia 2, Sobrepeso/Obeso2): Marge Simpson, Patty Bouvier y Selma Bouvier.

- **Dataset de expresión de genes:** para la realización de esta parte de la actividad, se deben considerar las secuencias de RNAseq con formato fastq de los personajes de los grupos obesos uno y dos.

El análisis se centrará en genes relacionados con la obesidad (ver en el fichero de listado de genes), cuya expresión ha sido simulada con perfiles específicos y consta de las siguientes secciones:

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

a. Control de calidad

- Determinación del control de calidad de las secuencias de RNA-seq.

Solución: Se eligió utilizar la biblioteca FASTQC y el análisis se corrió en Linux utilizando Oracle VirtualBox.

Resultados: Se analizaron los 10 informes de FastQC correspondientes a los 5 individuos de la familia Simpson/Bouvier pertenecientes a los grupos obesos uno y dos. Los resultados en forma de tabla se presentan en el póster. Los informes originales generados por la herramienta FASTQC para cada personaje integrante de los grupos obesos uno y dos están en GitHub. Los resultados revelaron las siguientes observaciones más relevantes:

- 1) Uniformidad de las muestras o secuencias analizadas debido a que todos los individuos presentaron exactamente 50,000 lecturas, lo cual sugiere un submuestreo previo para este análisis. El contenido de GC se mantiene estable entre el 48% y 50%, valores esperados para genomas humanos o similares.
- 2) Calidad R1 vs R2: * En la secuenciación Illumina, es normal observar una degradación de la calidad hacia el final de la lectura (extremo 3') debido al agotamiento de reactivos y a la desincronización de la señal química. En este juego de datos, aunque las secuencias Reverse (R2) muestran una dispersión ligeramente mayor en los últimos ciclos en comparación con las Forward (R1), ambos juegos mantienen una mediana de calidad de Phred > 30 hasta el final (151 pb). Por ello, todos los informes han recibido el estatus PASS.
- 3) Longitud de Lectura: La distribución de longitudes (35-151 pb) indica que las secuencias ya han pasado por un proceso de limpieza de adaptadores o trimming, eliminando bases de baja calidad o secuencias contaminantes en los extremos. Aunque como conocemos que estas secuencias fueron simuladas no es de esperar la presencia de adaptadores. Los archivos de control de calidad para cada personaje, el resumen de calidad en forma de tabla, así como los escritos en código R, incluidos en documentos Markdown, fueron subidos al repositorio de GitHub cuyo link es el siguiente: <https://github.com/obatista0115/Metodo-RNA-seq> (<https://github.com/obatista0115/Metodo-RNA-seq>)

b. Mapeo y cuantificación de las lecturas.

Asignación de las lecturas a genes o transcritos

- Pseudoalineamiento de las secuencias de RNA-seq a la secuencia de referencia

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

- Cuantificación de la abundancia de los transcritos
- Obtención de la matriz

Solución:

1. Indexación del genoma de referencia con salmon en Linux utilizando Oracle VirtualBox

El código utilizado para la indexación, así como la cuantificación de las secuencias fue facilitado por el profesor de la materia, doctor Sergio Buenestado Serrano. El código utilizado para la indexación fue el siguiente:

```
salmon index -t Referencia.fasta -i index -p 4
```

2. Cuantificación de la expresión génica de cada individuo con Salmon para lograr un pseudomapeo, el cual es más eficiente en términos de computación.

```
salmon quant -i index -l A -1 AbrahamSimpson_R1 -2  
AbrahamSimpson_R2 -p 4 -o Salmon/Abraham
```

```
salmon quant -i index -l A -1 HomerSimpson_R1 -2  
HomerSimpson_R2 -p 4 -o Salmon/Homer
```

```
salmon quant -i index -l A -1 MargeSimpson_R1 -2  
MargeSimpson_R2 -p 4 -o Salmon/Marge
```

```
salmon quant -i index -l A -1 PattyBouvier_R1 -2 PattyBouvier_R2  
-p 4 -o Salmon/Patty
```

```
salmon quant -i index -l A -1 SelmaBouvier_R1 -2  
SelmaBouvier_R2 -p 4 -o Salmon/Selma
```

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

CARGAR LIBRERÍAS

```
library(tximport)
library(DESeq2)
library(ggplot2)
library(pheatmap)
library(dplyr)
```

```
# ===== 1. CONFIGURACIÓN =====
=====
```

```
samples <- c("AbrahamSimpson", "HomerSimpson", "MargeSimpson", "PattyBouvier", "SelmaBouvier")
files <- file.path("Salmon", samples, "quant.sf")
names(files) <- samples
```

```
# ===== 2. LEER DATOS =====
===
```

```
tx2gene <- read.table("Transcrito_a_Gen.tsv", header = TRUE, sep = "\t")
txi <- tximport(files, type = "salmon", tx2gene = tx2gene, countsFromAbundance = "no")
counts_matrix <- round(txi$counts)
```

```
cat("Datos cargados:", nrow(counts_matrix), "genes\n")
```

```
## Datos cargados: 37 genes
```

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

```
# ===== 3. METADATOS =====  
==  
colData <- data.frame(  
  row.names = samples,  
  Persona = samples,  
  Grupo = factor(c("obeso1", "obeso1", "obeso2", "obeso2", "obeso2"))  
)  
  
# ===== 4. ANÁLISIS DIFERENCIAL =====  
=====  
dds <- DESeqDataSetFromMatrix(  
  countData = counts_matrix,  
  colData = colData,  
  design = ~ Grupo  
)  
  
dds <- DESeq(dds)  
res <- results(dds, contrast = c("Grupo", "obeso1", "obeso2"))  
  
# Preparar resultados  
res_df <- as.data.frame(res)  
res_df$gene <- rownames(res_df)  
res_df$significant <- ifelse(  
  !is.na(res_df$padj) & res_df$padj < 0.05 & abs(res_df$log2FoldChange) > 1,  
  "Significativo",  
  "No significativo"  
)  
  
# ===== 5. GRÁFICA 1: VOLCANO PLOT =====  
=====  
cat("\nGRÁFICA 1: Volcano Plot\n")
```

```
##
```

```
## GRÁFICA 1: Volcano Plot
```

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

```
volcano_data <- res_df
volcano_data$log10pvalue <- -log10(volcano_data$pvalue)

volcano_plot <- ggplot(volcano_data, aes(x = log2FoldChange, y = log10pvalue,
                                          color = significant)) +
  geom_point(size = 3) +
  scale_color_manual(values = c("No significativo" = "gray", "Significativo" = "red")) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "blue") +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "green") +
  labs(title = "Volcano Plot: Obeso 1 vs Obeso 2",
       x = "Log2 Fold Change", y = "-Log10 p-value") +
  theme_minimal()

if(sum(volcano_data$significant == "Significativo", na.rm = TRUE) > 0) {
  sig_genes <- volcano_data[volcano_data$significant == "Significativo", ]
  volcano_plot <- volcano_plot +
    geom_text(data = sig_genes, aes(label = gene), vjust = -0.5, size = 3)
}

print(volcano_plot)
```



I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

```
ggsave("01_volcano_plot.png", width = 10, height = 8)
```

```
# ===== 6. GRÁFICA 2: MA-PLOT =====  
=====  
cat("\nGRÁFICA 2: MA-Plot\n")
```

```
##
```

```
## GRÁFICA 2: MA-Plot
```

```
png("02_ma_plot.png", width = 800, height = 600)  
plotMA(res, main = "MA-Plot: Obeso 1 vs Obeso 2", ylim  
= c(-3, 3))  
dev.off()
```

```
## png
```

```
## 2
```

I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

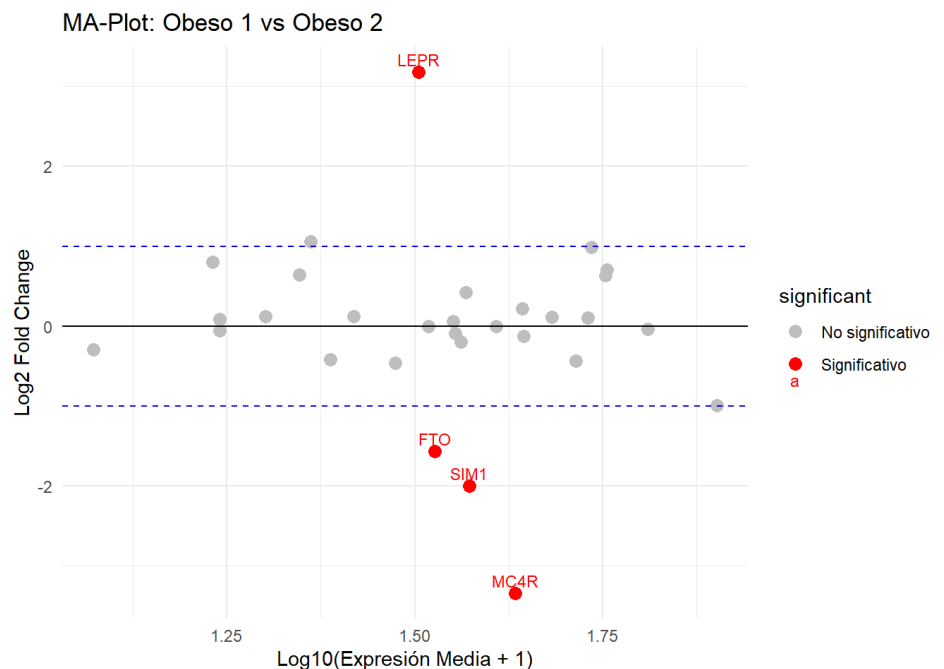
b. Mapeo y cuantificación de las lecturas.

```
# Versión ggplot2
normalized_counts <- counts(dds, normalized = TRUE)
ma_data <- res_df
ma_data$baseMean <- rowMeans(normalized_counts)

ma_plot <- ggplot(ma_data, aes(x = log10(baseMean +
1), y = log2FoldChange,
                                color = significant)) +
  geom_point(size = 3) +
  geom_hline(yintercept = 0, color = "black") +
  geom_hline(yintercept = c(-1, 1), linetype = "dashed", color = "blue") +
  scale_color_manual(values = c("No significativo" =
"gray", "Significativo" = "red")) +
  labs(title = "MA-Plot: Obeso 1 vs Obeso 2",
        x = "Log10(Expresión Media + 1)", y = "Log2 Fold
Change") +
  theme_minimal()

if(sum(ma_data$significant == "Significativo", na.rm =
TRUE) > 0) {
  sig_genes_ma <- ma_data[ma_data$significant == "Sign
ificativo", ]
  ma_plot <- ma_plot +
    geom_text(data = sig_genes_ma, aes(label = gene),
              vjust = -0.5, size = 3)
}

print(ma_plot)
```



I. Introducción

A. Objetivos

B. Pautas de elaboración

C. Instrucciones

a. Control de calidad

b. Mapeo y cuantificación de las lecturas.

```
ggsave("02_ma_plot_ggplot.png", width = 10, height = 8)
```