Análisis de expresión diferencial por RNA-seq en Sulfolobus acidocaldarius

Roberto Naranjo Partarrieu

Table of Contents

Introducción

El análisis de expresión diferencial mediante RNA-seq permite identificar genes cuya abundancia transcripcional varía entre condiciones experimentales. En este estudio se analizan datos de RNA-seq de *Sulfolobus acidocaldarius*, comparando el crecimiento planctónico y en biopelícula, así como genotipos wildtype y mutante para el gen *Lrs14-like*. El objetivo es identificar genes diferencialmente expresados asociados al medio de cultivo y al genotipo.

Materiales y Métodos

Control de calidad de lecturas

Las lecturas crudas en formato FASTQ fueron evaluadas mediante el programa IlluQC del paquete NGSQC Toolkit para caracterizar la calidad de las secuencias, su distribución de puntajes PHRED y el contenido de GC.

Filtrado de secuencias

Las lecturas fueron filtradas eliminando aquellas con baja calidad (PHRED < 20 en más del 20% de su longitud), conservando únicamente secuencias de alta calidad para los análisis posteriores.

Alineamiento al genoma de referencia

Las lecturas filtradas fueron alineadas contra el genoma de referencia de *Sulfolobus acidocaldarius* DSM 639 utilizando el algoritmo BWA-MEM.

Estimación de abundancia génica

La cuantificación de lecturas por gen se realizó mediante HTSeq-count, empleando un archivo de anotación en formato GFF3.

Análisis de expresión diferencial

```
library(edgeR)
input_dir <- "count"

wild_p <- read.delim(file.path(input_dir, "MW001_P.count"), header =
FALSE)
wild_b <- read.delim(file.path(input_dir, "MW001_B3.count"), header =
FALSE)
mut_p  <- read.delim(file.path(input_dir, "0446_P.count"), header =
FALSE)
mut_b  <- read.delim(file.path(input_dir, "0446_B3.count"), header =
FALSE)

colnames(wild_p) <- c("Gen_ID","Count")
colnames(wild_b) <- c("Gen_ID","Count")
colnames(mut_p)  <- c("Gen_ID","Count")
colnames(mut_b)  <- c("Gen_ID","Count")

rawcounts <- data.frame(
  wild_p$Gen_ID,
  WildType_P = wild_p$Count,
  WildType_B = wild_b$Count,
  Mutant_P   = mut_p$Count,
  Mutant_B   = mut_b$Count,
  row.names = 1
)

rpkm <- cpm(rawcounts)

to_remove <- rownames(rawcounts) %in% c(
  "__no_feature","__ambiguous",
  "__too_low_aQual","__not_aligned",
  "__alignment_not_unique"
)

keep <- rowSums(rpkm > 1) >= 3 & !to_remove
rawcounts_f <- rawcounts[keep,]
```

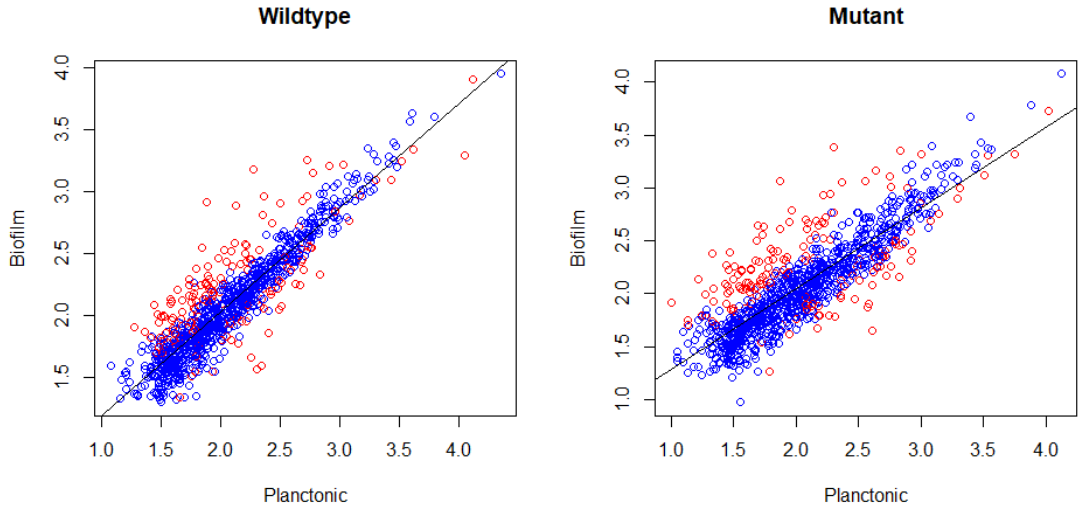
```
group_culture <- c("planctonic","biofilm","planctonic","biofilm")

dge_culture <- DGEList(counts = rawcounts_f, group = group_culture)
dge_culture <- calcNormFactors(dge_culture)
dge_culture <- estimateCommonDisp(dge_culture)
dge_culture <- estimateTagwiseDisp(dge_culture)

de_culture <- exactTest(dge_culture, pair = c("planctonic","biofilm"))
results_culture <- topTags(de_culture, n = nrow(dge_culture))$table
ids_culture <- rownames(results_culture[results_culture$FDR < 0.1,])

rawcounts_genotype <- rawcounts_f[!rownames(rawcounts_f) %in%
ids_culture,]

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(dge_genotype))$table
ids_genotype <- rownames(results_genotype[results_genotype$FDR < 0.1,])
```

Resultados y Discusión

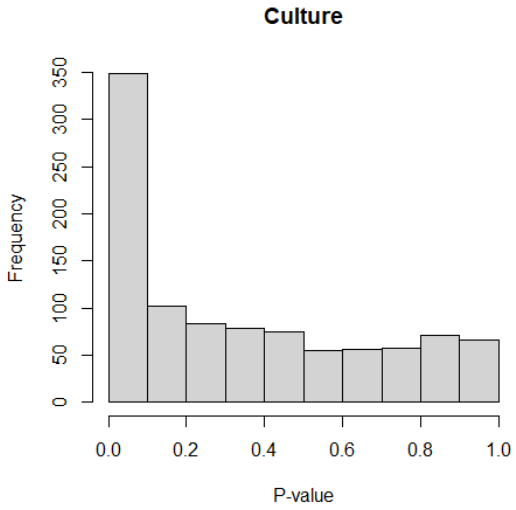
Se identificaron 163 genes diferencialmente expresados asociados al medio de cultivo (planctónico vs biopelícula), mientras que solo un gen mostró expresión diferencial significativa entre genotipos. Estos resultados indican que el modo de crecimiento ejerce un efecto transcriptómico más amplio que la mutación evaluada, sugiriendo que la transición a biopelícula constituye el principal factor regulador de la expresión génica bajo las condiciones estudiadas.

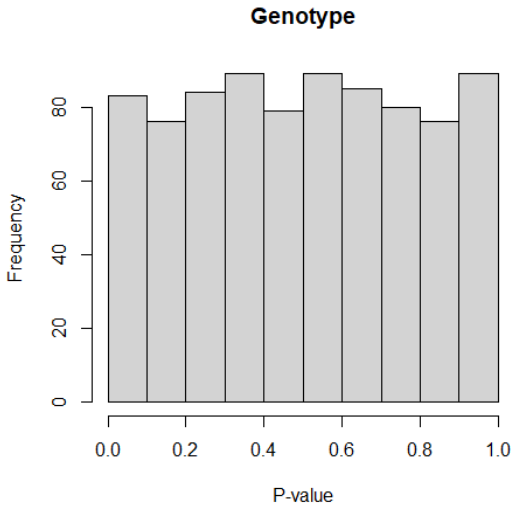
```
de_genes_culture <- rownames(rawcounts_f) %in% ids_culture

pseudocounts <- data.frame(
WildType_P = log10(dge_culture$pseudo.counts[,1]),
WildType_B = log10(dge_culture$pseudo.counts[,2]),
Mutant_P = log10(dge_culture$pseudo.counts[,3]),
Mutant_B = log10(dge_culture$pseudo.counts[,4]),
DE_C = de_genes_culture
)

par(mfrow=c(1,2))

plot(pseudocounts$WildType_P, pseudocounts$WildType_B,
```

```
col = ifelse(pseudocounts$DE_C, "red", "blue"),
main = "Wildtype",
xlab = "Planctonic",
ylab = "Biofilm")
abline(lsfit(pseudocounts$WildType_P, pseudocounts$WildType_B))

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))
```




```
par(mfrow=c(1,2))
hist(results_culture$PValue, main="Culture", xlab="P-value")
hist(results_genotype$PValue, main="Genotype", xlab="P-value")
```





##

Conclusiones

El análisis de expresión diferencial demuestra que el medio de cultivo tiene un efecto predominante sobre la expresión génica en Sulfolobus acidocaldarius, mientras que la mutación en el gen Lrs14-like presenta un impacto transcriptómico limitado y específico.

rmarkdown::render("RNAseq_DE_analysis.Rmd")