Análisis de clustering jerárquico de expresión génica

Autor: Roberto Naranjo

1. introducción y objetivos

Este análisis corresponde a la continuación del estudio de expresión diferencial realizado sobre datos de microarreglos Illumina MouseRef-8. A partir de los resultados obtenidos previamente con **limma**, se seleccionaron únicamente los genes que mostraron expresión diferencial significativa, con el fin de explorar patrones globales de similitud tanto entre **muestras** como entre **sondas (genes)**.

El objetivo específico de esta etapa es:

- Evaluar si las **muestras** se agrupan de forma coherente utilizando **clustering jerárquico con distancia euclidiana**.
- Evaluar si las **sondas** presentan patrones coordinados de expresión utilizando **clustering jerárquico con distancia basada en el complemento de la correlación de Pearson (1 − r)**.
- Justificar la elección del número de clústeres mediante **gráficos de suma de cuadrados intra-clúster (SSQ)**.
- Presentar dendrogramas finales con los clústeres seleccionados.

El análisis permite evaluar si los genes identificados como diferencialmente expresados presentan patrones coordinados de expresión, y si dichas estructuras se reflejan también a nivel de las muestras, reforzando la interpretación biológica del análisis de expresión diferencial previo.

---

2. Selección de genes para el clustering

Para lograr este trabajo se utilizaron exclusivamente los genes seleccionados como diferencialmente expresados en el análisis previo. En particular, se optó por usar el efecto de **interacción Genotipo × Tratamiento**, ya que este efecto captura diferencias dependientes simultáneamente de ambas variables experimentales y es el más informativo dentro del diseño factorial 2×2.

Se consideraron genes significativos con un umbral de **FDR ≤ 0.19**, de acuerdo con lo solicitado en la tarea.

```r
# Selección de genes diferencialmente expresados por interacción
keep <- results$FDR.Int <= 0.19
X <- normdata[keep, , drop = FALSE]
```

Bajo este criterio se seleccionaron *N* genes (o sondas), los cuales fueron utilizados como base para todos los análisis de clustering posteriores.

---

3. Preprocesamiento de los datos para clustering

Antes de realizar los análisis de clustering, los datos fueron transformados para asegurar que la distribución y la escala fueran apropiadas para las métricas de distancia utilizadas:

- Los datos ya se encontraban **normalizados por cuantiles** desde el análisis previo.
- Para el clustering, cada gen fue **centrado y escalado** (media 0, desviación estándar 1) de manera independiente.

Este paso se aplicó para evitar que genes con mayor varianza dominen las medidas de distancia, especialmente en el clustering de muestras basado en distancia euclidiana.

```r
# Estandarización por gen
Xz <- t(scale(t(X)))
Xz[is.na(Xz)] <- 0
```

---

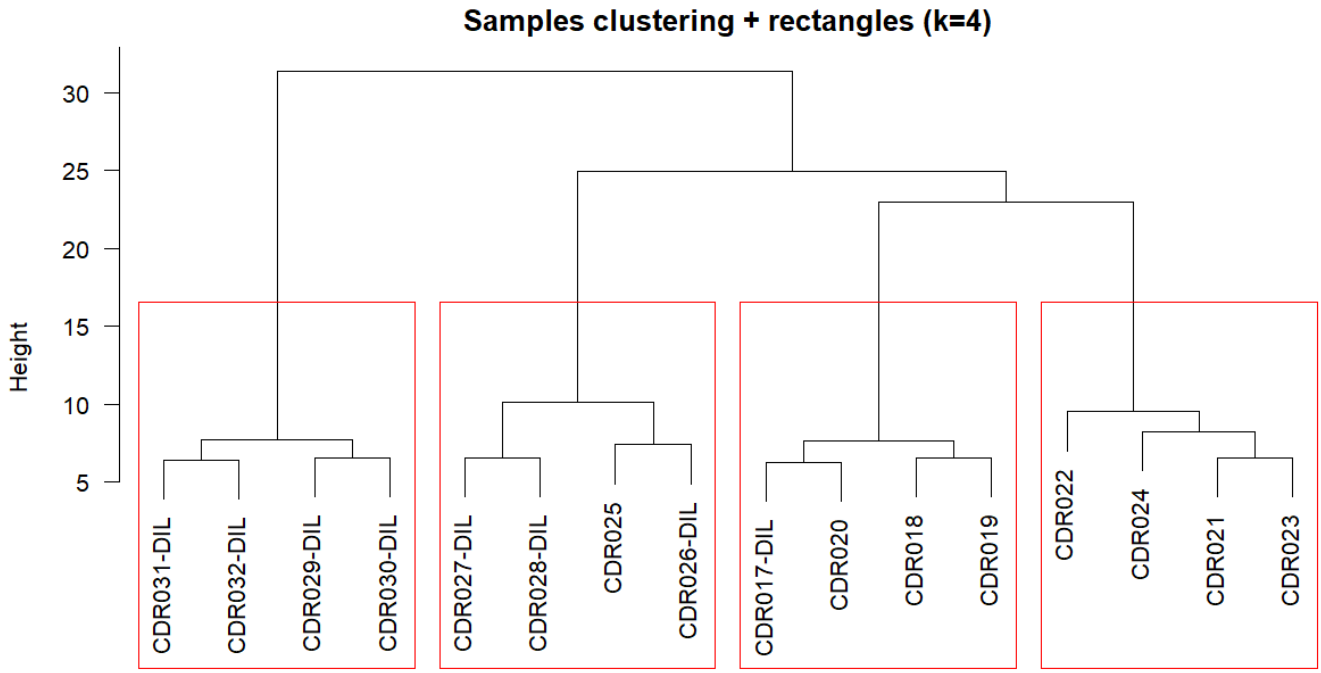
4. Clustering jerárquico de muestras

El clustering de muestras se realizó utilizando:

- **Distancia:** Euclidiana
- **Método de enlace:** Ward.D2

La distancia se calculó sobre la matriz estandarizada, considerando a cada muestra como un vector de expresión génica.

```r
# Clustering de muestras
d_samples <- dist(t(Xz), method = "euclidean")
hc.samples <- hclust(d_samples, method = "ward.D2")
```



`hclust_samples_rect.png`

(Dendrograma de clustering jerárquico de muestras con rectángulos, k = 4)

La estructura del dendrograma muestra una separación clara entre grupos de muestras, consistente con el diseño experimental, sugiriendo que las diferencias asociadas a genotipo y tratamiento contribuyen de manera importante a la variabilidad global de expresión.

---

5. Clustering jerárquico de sondas (genes)

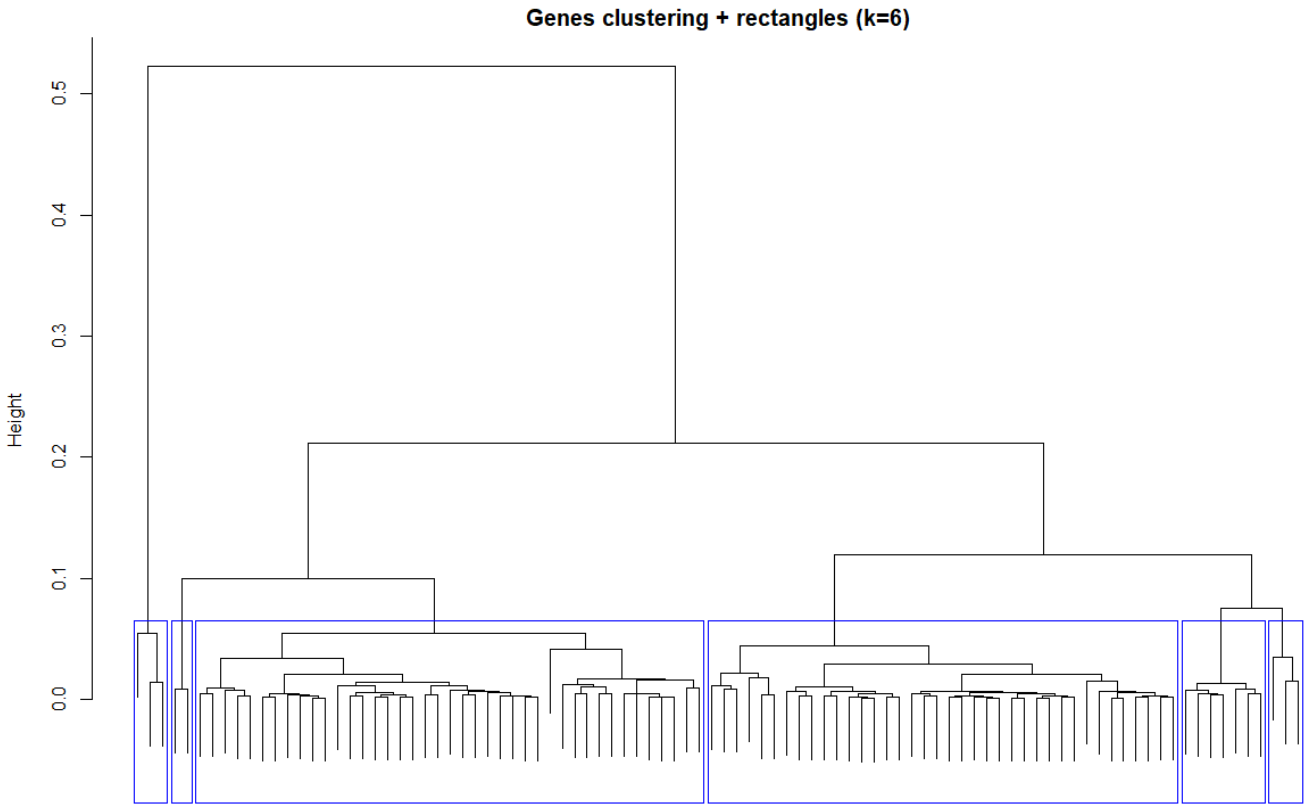
El clustering de sondas se realizó con el objetivo de identificar conjuntos de genes con perfiles de expresión similares a través de las muestras. Para ello se utilizó:

- **Distancia:** 1 − correlación de Pearson
- **Método de enlace:** Ward.D2

Esta métrica permite agrupar genes según la similitud en la forma de sus perfiles de expresión, independientemente de su magnitud absoluta.

```r
library(amap)

d_genes <- Dist(Xz, method = "pearson")
hc.genes <- hclust(d_genes, method = "ward.D2")
```



`hclust_genes_rect.png`

(Dendrograma de clustering jerárquico de genes con rectángulos, k = 6)

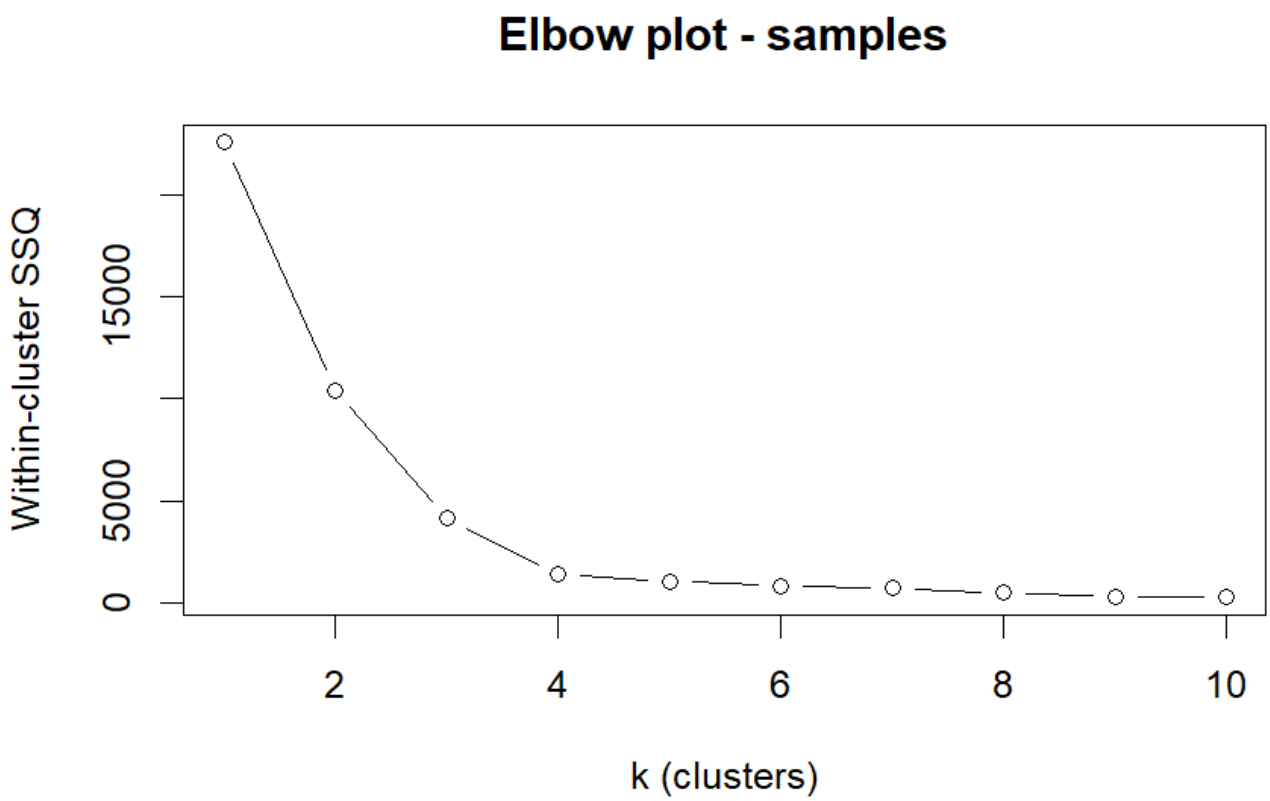
La presencia de múltiples clústeres sugiere que los genes diferencialmente expresados por la interacción no responden de forma homogénea, sino que se organizan en módulos con respuestas transcripcionales diferenciadas

---

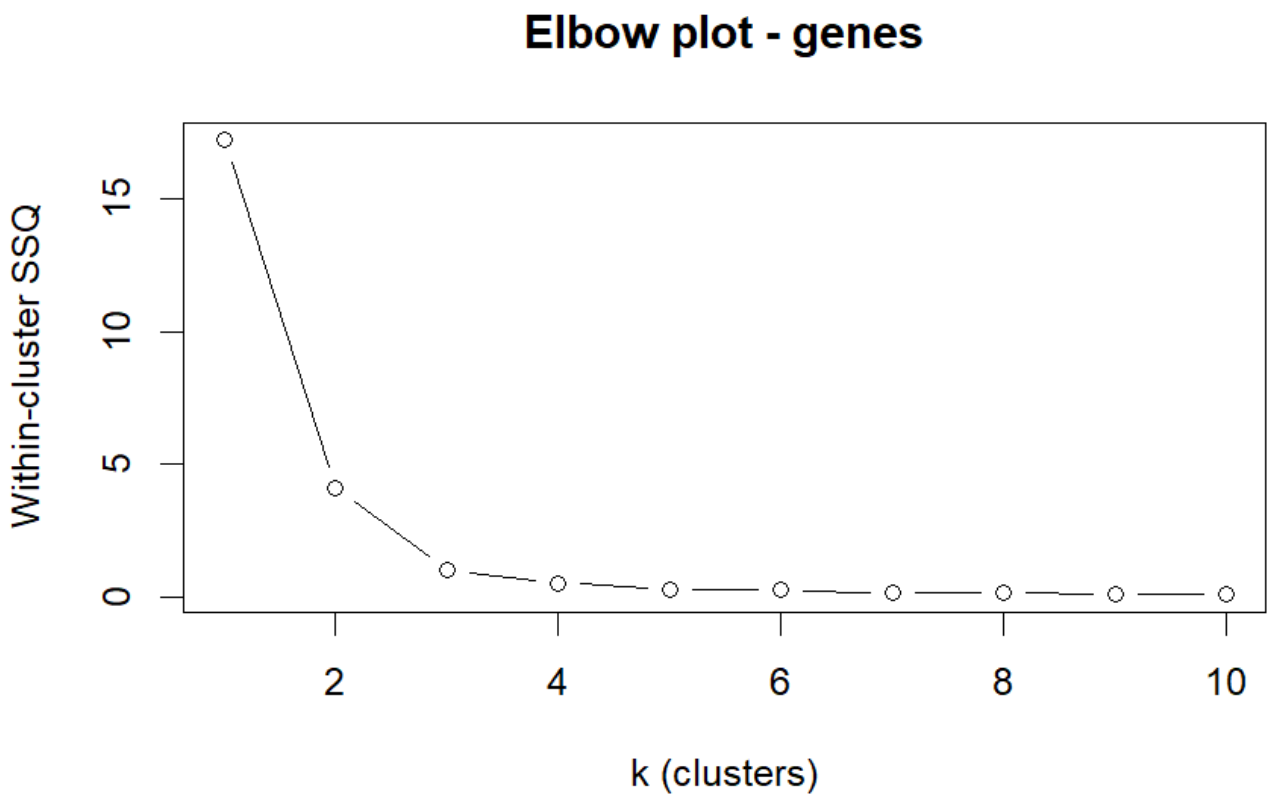
6. Determinación del número de clústeres (SSQ)

Para seleccionar el número apropiado de clústeres, se calcularon gráficos de **suma de cuadrados intra-clúster (SSQ)** para distintos valores de *k* tanto en muestras como en genes.

```r
# Función para calcular SSQ
wss_from_hclust <- function(hc, D, kmax = 10) {
  out <- numeric(kmax)
  for (k in 1:kmax) {
    cl <- cutree(hc, k = k)
    ss <- 0
    for (g in unique(cl)) {
      idx <- which(cl == g)
      if (length(idx) > 1) {
        ss <- ss + sum(as.matrix(D)[idx, idx]^2) / 2
      }
    }
    out[k] <- ss
  }
  out
}
```



`SSQ_samples.png` *(Gráfico SSQ para clustering de muestras)*



`SSQ_genes.png` *(Gráfico SSQ para clustering de genes)*

En ambos casos se observa un cambio marcado en la pendiente de la curva… No obstante, el punto de inflexión ocurre antes en muestras que en genes, reflejando una mayor heterogeneidad en los perfiles de expresión génica.

- **k = 4** clústeres para muestras
- **k = 6** clústeres para genes

---

7. Dendrogramas finales con rectángulos

A partir de los valores seleccionados de *k*, se generaron los dendrogramas finales incorporando rectángulos para resaltar visualmente los clústeres definidos.

```
# Rectángulos en dendrogramas
rect.hclust(hc.samples, k = 4, border = "red")
rect.hclust(hc.genes, k = 6, border = "blue")
```

(En el informe se presentan únicamente los dendrogramas finales con rectángulos, de acuerdo a lo que se pidio en la tarea)

---

8. Conclusiones

El análisis de clustering jerárquico realizado sobre genes diferencialmente expresados permitió identificar patrones claros de agrupamiento tanto a nivel de muestras como de genes. La coherencia observada en el clustering de muestras respalda la calidad del experimento y la relevancia del diseño experimental, mientras que los clústeres de genes sugieren la existencia de módulos de coexpresión potencialmente asociados a mecanismos biológicos dependientes del genotipo y el tratamiento.

En conjunto, estos resultados complementan el análisis de expresión diferencial y el análisis funcional, aportando evidencia adicional de que la interacción genotipo–tratamiento estructura la respuesta transcriptómica tanto a nivel de genes individuales como de patrones globales de expresión.