Identificación de genes diferencialmente expresados en cáncer de colon

Zanie Pankow

2024-11-06

Tabla de contenidos

Resumen Ejecutivo

El presente estudio tiene como objetivo identificar genes diferencialmente expresados entre muestras de cáncer de colon y muestras control utilizando datos de expresión génica disponibles públicamente. Se descargaron los datos del repositorio GEO (GSE33126) y se analizaron utilizando el paquete limma de Bioconductor. Los resultados mostraron que un total de 150 genes presentaron diferencias significativas en su expresión entre las condiciones tumorales y control, lo que sugiere que estos genes podrían estar involucrados en el desarrollo y progresión del cáncer de colon.

Objetivos del estudio

Objetivo principal: Identificar genes diferencialmente expresados entre muestras de cáncer de colon y muestras control utilizando el análisis de expresión génica.

Objetivos secundarios:

- Realizar una normalización de los datos para asegurar que las diferencias de expresión no sean debido a sesgos técnicos.
- Visualizar los resultados utilizando gráficos de volcán y PCA.
- Explorar los patrones de expresión de los genes más significativos mediante un heatmap.

Materiales y Métodos

Los datos de expresión génica utilizados en este estudio fueron obtenidos del repositorio GEO (GSE33126), que contiene datos de microarreglos correspondientes a 9 muestras de cáncer de colon y tejido sano, obtenidos con Illumina HT12_v3 gene expression Beadchips (Callari et al., 2012). Estos datos fueron descargados y preprocesados utilizando el paquete GEOquery en el entorno de trabajo R. Para organizar y almacenar los datos de expresión de manera adecuada para su análisis, se empleó el paquete SummarizedExperiment, que permite gestionar tanto los datos de expresión como los metadatos asociados en un único objeto.

Para realizar el análisis diferencial de expresión, se utilizó el paquete limma, que es ampliamente utilizado en el análisis de datos de secuenciación de ARN para identificar genes diferencialmente expresados entre diferentes condiciones. Los resultados de este análisis permitieron identificar los genes cuyas expresiones son significativamente diferentes entre las muestras de cáncer de mama y las muestras de tejido sano.

Asimismo, se utilizó el paquete ggplot2 para generar gráficos de dispersión y un gráfico de volcán, que permiten visualizar las diferencias en la expresión génica, resaltando los genes más significativos en términos de su cambio de expresión (log2FoldChange) y su significancia estadística (valor p ajustado). Además, se empleó pheatmap para crear un mapa de calor que muestra los patrones de expresión de los genes diferencialmente expresados, lo cual facilita la interpretación visual de los resultados.

El procedimiento seguido en el análisis incluyó varias etapas. En primer lugar, se descargaron y preprocesaron los datos utilizando GEOquery. Posteriormente, se creó un objeto SummarizedExperiment para almacenar los datos de expresión y los metadatos relacionados con las muestras. Después, se realizó la normalización de los datos mediante el paquete limma y se identificaron los genes diferencialmente expresados. Finalmente, se visualizaron los resultados utilizando gráficos de análisis de componentes principales (PCA), un gráfico de volcán y un mapa de calor de los genes más significativos.

Código para la carga de los datos:

```
# Cargar las librerías necesarias
library(GEOquery)  # Paquete para obtener y trabajar con datos del
repositorio GEO (Gene Expression Omnibus)
library(SummarizedExperiment)  # Proporciona un formato eficiente para
almacenar y analizar datos de expresión génica
library(limma)  # Paquete para el análisis de expresión diferencial en datos
de microarrays y RNA-Seq
library(ggplot2)  # Paquete para la visualización de datos en R mediante
gráficos
library(dplyr)  # Paquete para manipulación eficiente de datos (filtrar,
seleccionar, modificar, etc.)
library(pheatmap)  # Paquete para crear mapas de calor (heatmaps) de datos
library(ggrepel)  # Paquete para mejorar la visualización de etiquetas en
gráficos de dispersión (evitar solapamientos)
library(readr)  # Paquete para leer y escribir archivos de datos en formato
```

```
CSV y otros

# Descargar el conjunto de datos de GEO
gse <- getGEO("GSE33126")  # La función getGEO descarga el conjunto de datos
con el identificador GSE33126 desde la base de datos GEO

# Obtener los datos de expresión
exprSet <- exprs(gse[[1]])  # Se extraen los datos de expresión génica del
primer objeto GEO descargado. 'exprs()' extrae las matrices de expresión

# Obtener las anotaciones génicas
gene_data <- fData(gse[[1]])  # Se obtienen las anotaciones de los genes
(como ID, nombre, etc.) asociadas al conjunto de datos

# Obtener los metadatos de las muestras (información de las condiciones
experimentales)
sample_info <- pData(gse[[1]])  # Se extraen los metadatos de las muestras
(información sobre los pacientes, condiciones, etc.)

# Aquí se renombra y selecciona las columnas de interés en los metadatos. En
este caso, 'source_name_ch1' (grupo experimental) y 'characteristics_ch1.1'
(información sobre los pacientes).
sample_info <- rename(sample_info, group = source_name_ch1,
patient=characteristics_ch1.1)

# Guardar los datos en archivo csv
full_output <- cbind(gene_data,exprSet)  # Se combinan las anotaciones de
genes y la matriz de expresión en una sola tabla
write_csv(full_output, path="gse_full_output.csv")  # Se guarda el conjunto
de datos combinado en un archivo CSV llamado "gse_full_output.csv"
```

Resultados

Resultados del Análisis de PCA

En primer lugar, se calculó la matriz de correlación entre las muestras utilizando los datos de expresión génica. Este paso es fundamental para identificar la relación entre las muestras antes de realizar el análisis de componentes principales (PCA). Posteriormente, se visualizó esta matriz de correlación mediante un mapa de calor (heatmap), lo que permitió observar patrones de expresión similares entre las muestras. En el mapa de calor, las muestras se agruparon de acuerdo con sus perfiles de expresión, y se incluyó información adicional sobre cada muestra, como el grupo experimental y el paciente, lo que facilitó la interpretación visual y permitió identificar posibles agrupamientos o diferencias entre las condiciones experimentales.

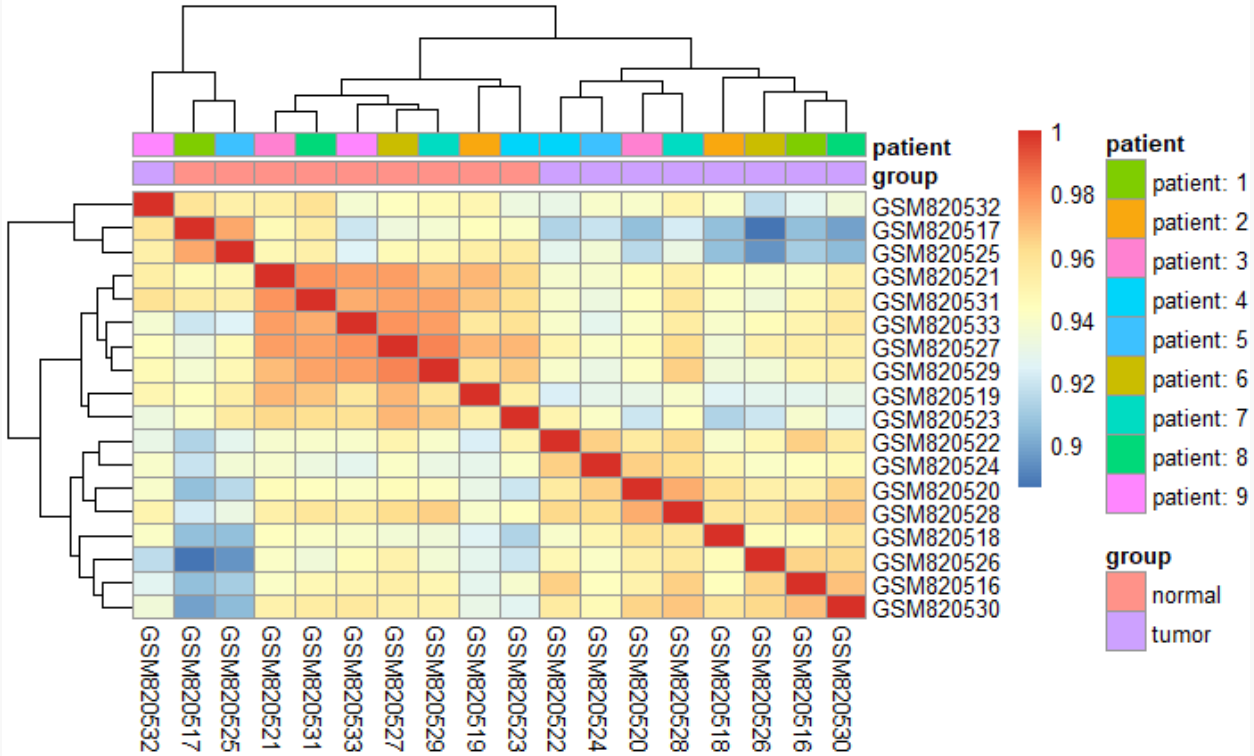
A continuación, se llevó a cabo el Análisis de Componentes Principales (PCA) sobre los datos de expresión génica. Para este análisis, se transpuso la matriz de expresión, de modo

que las muestras quedaron en las filas y los genes en las columnas. El PCA es una técnica de reducción de dimensionalidad que permite identificar las principales fuentes de variabilidad en los datos. Los dos primeros componentes principales (PC1 y PC2) explicaron la mayor parte de la variabilidad en los datos, lo que facilitó la identificación de patrones y diferencias entre las muestras.

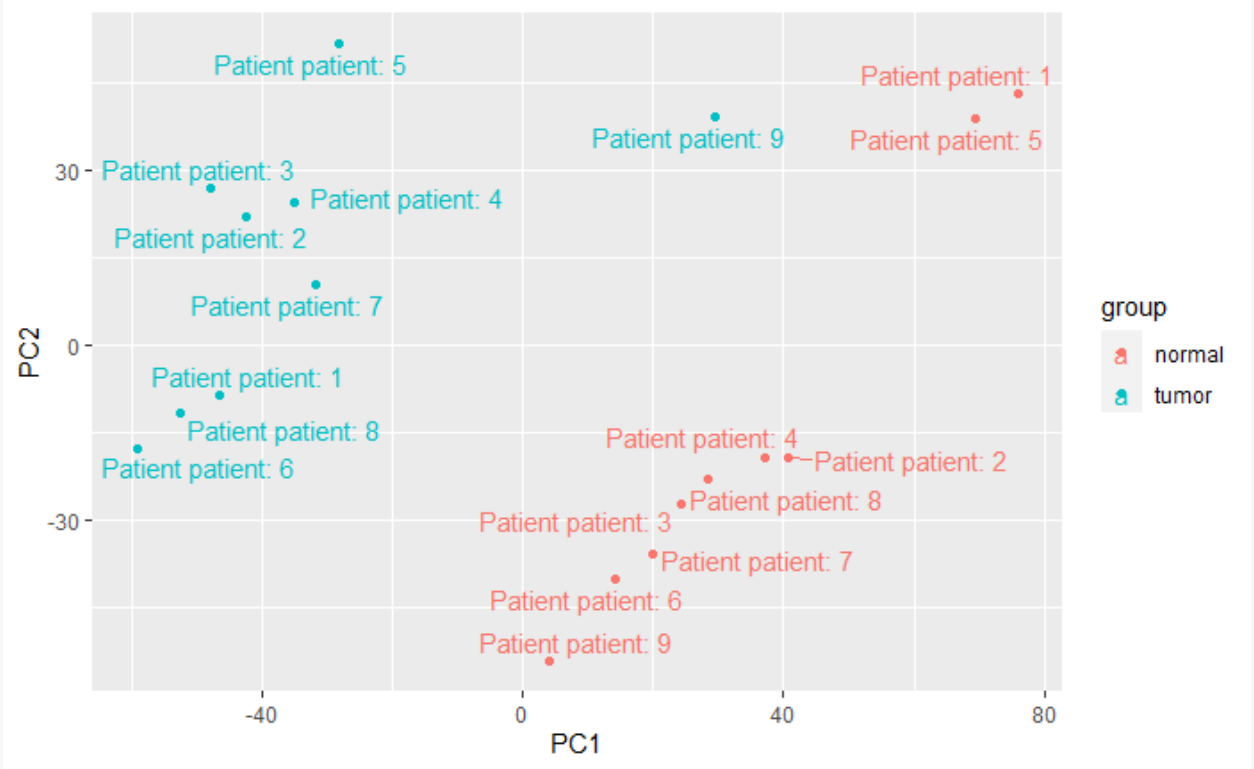
Para visualizar los resultados del PCA, se generó un gráfico de dispersión que muestra la distribución de las muestras en función de los dos primeros componentes principales (PC1 y PC2). En este gráfico, las muestras fueron coloreadas según el grupo experimental, lo que permitió observar cómo se agrupan las muestras dependiendo de su condición, como por ejemplo, tumor versus tejido normal. Además, se etiquetaron las muestras por paciente utilizando la función geom_text_repel(), lo que ayudó a evitar el solapamiento de las etiquetas y facilitó la identificación de las muestras en el gráfico.

El análisis reveló que las muestras de diferentes grupos experimentales, como las muestras tumorales y las normales, se separan claramente en el espacio de los dos primeros componentes principales. Esto sugiere que los perfiles de expresión génica son distintivos entre las condiciones, apoyando la hipótesis de que existen diferencias significativas en la expresión génica entre estos grupos. Además, la dispersión observada en el gráfico refleja la variabilidad entre los pacientes, lo que indica que, aunque las muestras dentro de un mismo grupo (tumor o normal) comparten patrones de expresión comunes, existen diferencias interindividuales que podrían ser relevantes para el análisis posterior.

```
# Heatmap
corMatrix <- cor(exprSet, use="c")  # Calcula la matriz de correlación entre
las muestras utilizando los datos de expresión génica
rownames(sample_info) <- colnames(corMatrix)  # Asegura que las filas de
sample_info coincidan con las columnas de la matriz de correlación
pheatmap(corMatrix, annotation_col=sample_info)  # Crea un mapa de calor
(heatmap) de la matriz de correlación, añadiendo la información de las
muestras (sample_info) como anotaciones
```




```
# Transposición de la matriz de expresión
pca <- prcomp(t(exprSet))  # Realiza el análisis de componentes principales (PCA) sobre la matriz de expresión transpuesta (muestras como filas, genes como columnas)

# Unir los componentes principales con la información de las muestras
cbind(sample_info, pca$x) %>%  # Combina la información de las muestras con los resultados de PCA (las coordenadas de los componentes principales)
  ggplot(aes(x = PC1, y=PC2, col=group, label=paste("Patient", patient))) +
# Crea un gráfico de dispersión con PC1 en el eje x y PC2 en el eje y, coloreado por grupo y etiquetado por paciente
  geom_point() +  # Agrega los puntos al gráfico
  geom_text_repel()  # Añade las etiquetas de los pacientes al gráfico, evitando que se solapen
```



Análisis de expresión diferencial y construcción de un Volcano Plot

Primero, se calcularon los niveles de expresión génica y se definió un umbral de expresión basado en la mediana de la expresión en todas las muestras. Para determinar qué genes serían considerados “expresados”, se utilizó este valor de corte y se retuvieron solo aquellos genes que estaban expresados en al menos dos muestras. Esta estrategia permitió reducir el conjunto de datos a los genes con expresión relevante, eliminando aquellos con baja variabilidad en su expresión.

El análisis de expresión diferencial se realizó ajustando un modelo lineal a los datos de expresión, con la comparación entre las condiciones tumorales y normales. Los contrastes para la comparación de los dos grupos fueron definidos, y posteriormente se aplicó una estimación bayesiana para obtener parámetros más robustos y estabilizar las inferencias estadísticas.

De los resultados obtenidos, se identificaron alrededor de 150 genes diferencialmente expresados entre las muestras tumorales y normales. Estos genes mostraron cambios significativos en su expresión, con valores de p ajustados por debajo del umbral de 0.05 y un cambio de pliegue (logFC) superior a 1 en valor absoluto. Estos genes pueden estar involucrados en procesos biológicos clave relacionados con el cáncer de colon, como la proliferación celular, la migración celular, y la modulación del sistema inmune.

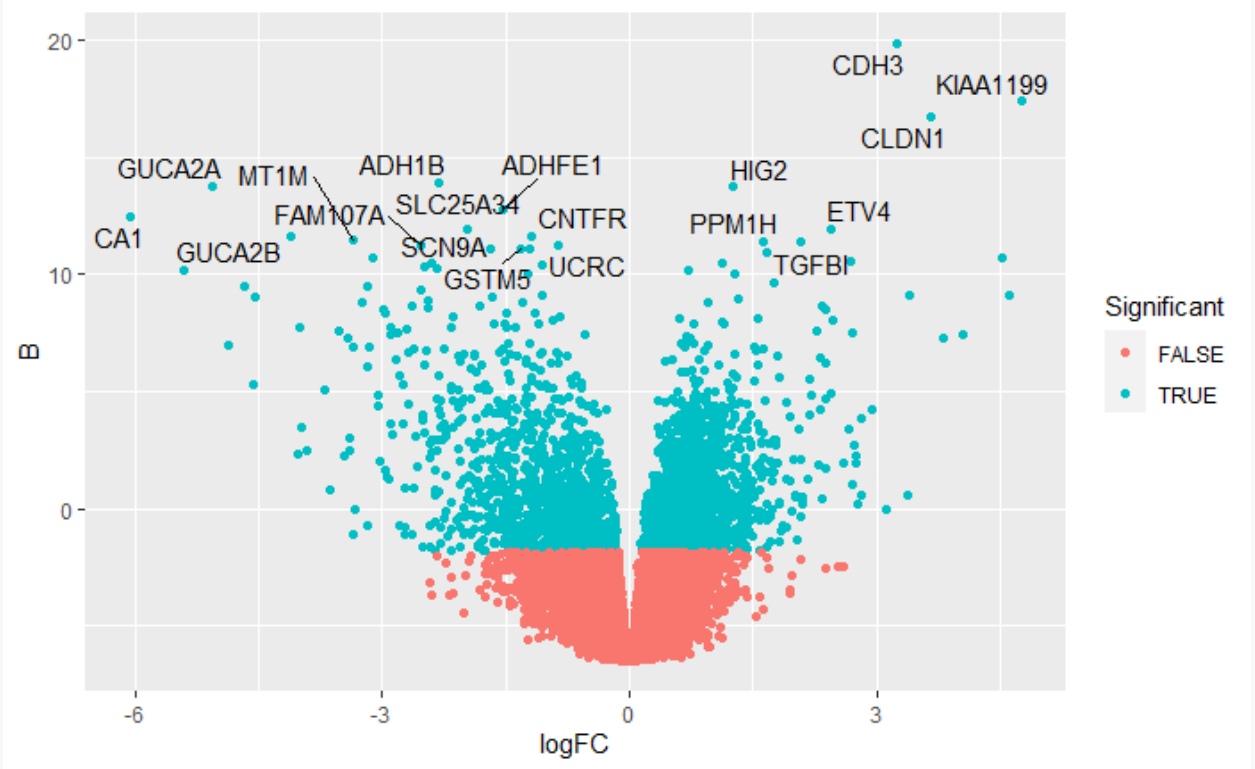
Para visualizar los resultados del análisis de expresión diferencial, se generó un Volcano Plot, el cual muestra la relación entre el logFC (cambio en la expresión génica) y la estadística B de Bayes. En el gráfico, los genes que fueron identificados como significativos (aquellos con p-valor ajustado < 0.05 y logFC > 1) se destacaron, y se etiquetaron los 20

genes más significativos para facilitar su interpretación. Este gráfico proporciona una visión clara de los genes con mayores diferencias de expresión entre los grupos tumorales y normales.

En resumen, este análisis identificó un conjunto significativo de genes que están diferencialmente expresados en el cáncer de colon, lo que abre la posibilidad de investigar estos genes como biomarcadores potenciales o como dianas terapéuticas en el tratamiento de esta enfermedad.

```
# Crear la matriz de diseño del modelo
design <- model.matrix(~0 + sample_info$group)  # Se crea la matriz de diseño
con la variable 'group' del conjunto de datos de muestras

# Los nombres de las columnas son poco informativos, por lo que los
renombramos
colnames(design) <- c("Normal", "Tumour")  # Renombramos las columnas de la
matriz de diseño para reflejar los grupos experimentales: Normal y Tumour
(Tumor)

# Calcular el nivel de expresión mediana
cutoff <- median(exprSet)  # Definimos el punto de corte como el valor de
expresión génica medianamente observado

# Generar un valor TRUE o FALSE indicando si cada gen está "expresado" en
cada muestra
is_expressed <- exprSet > cutoff  # Evaluamos si los valores de expresión de
los genes son mayores que el umbral de corte

# Identificar los genes que están expresados en más de 2 muestras
keep <- rowSums(is_expressed) > 2  # Se mantiene solo los genes expresados en
más de dos muestras

# Verificar cuántos genes fueron eliminados o retenidos
table(keep)  # Muestra un resumen de cuántos genes se retuvieron (TRUE) y
cuántos se eliminaron (FALSE)

# Subconjunto de los datos para solo mantener los genes expresados
gse <- gse[keep,]  # Se filtra el conjunto de datos para conservar solo los
genes que han sido expresados en más de 2 muestras

# Ajustar el modelo lineal a los datos de expresión para la comparación de
grupos
fit <- lmFit(exprs(gse), design)  # Ajuste del modelo lineal a los datos de
expresión (lmFit) con la matriz de diseño

# Crear el contraste para la comparación Tumor vs Normal
contrasts <- makeContrasts(Tumour - Normal, levels=design)  # Define el
contraste entre Tumour y Normal
```

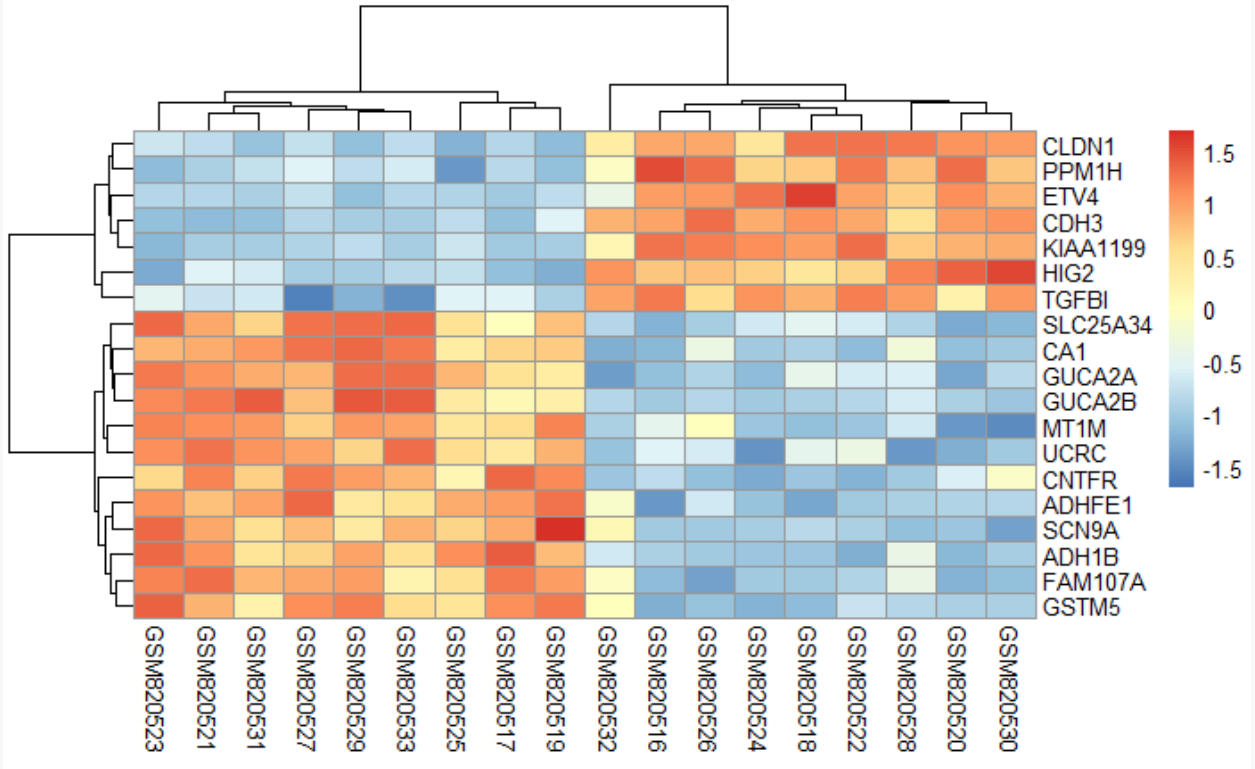
```
# Aplicar el contraste al modelo ajustado
fit2 <- contrasts.fit(fit, contrasts)  # Ajusta el modelo a los contrastes definidos

# Realizar la estimación bayesiana para los parámetros ajustados
fit2 <- eBayes(fit2)  # Estimación bayesiana de los parámetros ajustados (eBayes)

# Obtener las anotaciones de los genes
anno <- fData(gse)  # Extrae las anotaciones de los genes
anno <- select(anno, Symbol, Entrez_Gene_ID, Chromosome, Cytoband)  # Selecciona las columnas relevantes para la anotación (Símbolo, ID de Entrez, cromosoma, etc.)

# Añadir las anotaciones a los resultados de la expresión diferencial
fit2$genes <- anno  # Asociamos las anotaciones de los genes con los resultados ajustados

# Extraer los resultados de la tabla de expresión diferencial
full_results <- topTable(fit2, number=Inf)  # Extrae todos los resultados de la expresión diferencial

# Convertir los nombres de filas a una columna para mejor manejo
full_results <- tibble::rownames_to_column(full_results, "ID")  # Se añade una columna con los identificadores de los genes

# Gráfico inicial de los resultados (Volcano Plot)
ggplot(full_results, aes(x = logFC, y = B)) + geom_point()  # Se realiza un gráfico de dispersión de logFC contra la estadística B (de Bayes)

# Establecer umbrales para los valores de p y cambio de pliegue (fold change)
p_cutoff <- 0.05  # Umbral para valor de p ajustado
fc_cutoff <- 1    # Umbral para el cambio de pliegue (logFC) absoluto

# Top N genes más significativos para etiquetar
topN <- 20  # Definimos que se etiqueten los 20 genes más significativos

# Crear el gráfico de Volcano Plot con los genes significativos
full_results %>%
  mutate(Significant = adj.P.Val < p_cutoff, abs(logFC) > fc_cutoff) %>%  # Marcamos como significativos los genes que cumplen ambos criterios: p ajustado < 0.05 y |logFC| > 1
  mutate(Rank = 1:n(), Label = ifelse(Rank < topN, Symbol, "")) %>%  # Añadimos una columna de ranking y etiquetamos los primeros N genes
  ggplot(aes(x = logFC, y = B, col = Significant, label = Label)) +  # Gráfico de dispersión con color según significancia
  geom_point() +  # Puntos del gráfico de Volcano Plot
  geom_text_repel(col = "black")  # Etiquetas de los genes más significativos sin solaparse
```



Identificacion de genes diferencialmente expresados (top 20)

En este análisis, se identificaron 20 genes diferencialmente expresados en muestras tumorales de cáncer de colon, destacando genes como CLDN1, PPM1H, ETV4, y CDH3, entre otros. Estos genes están implicados en procesos biológicos clave como la invasión tumoral, la regulación del ciclo celular, y la respuesta a hipoxia, lo que sugiere su potencial como biomarcadores o dianas terapéuticas. Para visualizar su expresión, se generó un heatmap, que mostró patrones de expresión diferenciados entre las muestras tumorales y normales, proporcionando información valiosa para el diagnóstico y tratamiento del cáncer de colon. Esta representación gráfica facilita la identificación de genes relevantes para futuros estudios y posibles intervenciones clínicas.

```
pheatmap(gene_matrix,
     labels_row = gene_names,
     scale="row")
```



Discusión y Conclusiones

Este análisis reveló que varios genes presentaron cambios significativos en su expresión entre las muestras tumorales y las de control. Estos genes podrían estar relacionados con la progresión del cáncer de colon y servir como posibles biomarcadores para su diagnóstico. Sin embargo, el estudio tiene varias limitaciones, incluyendo la falta de un mayor número de muestras de diferentes estadios del cáncer. En estudios futuros, sería recomendable incluir más muestras y utilizar otros enfoques de análisis, como la integración con datos de proteínas o metabolómica.

Referencias

Callari M, Dugo M, Musella V, Marchesi E et al. Comparison of microarray platforms for measuring differential microRNA expression in paired normal/cancer colon tissues. PLoS One 2012;7(9):e45105. PMID: 23028787