PEC1 ANÁLISIS DE DATOS ÓMICOS

Alex Andrés Jarrín Jurado

2024-11-06

Tabla de Contenidos

Resumen Ejecutivo

En este estudio, se analizaron datos metabolómicos de pacientes sometidos a cirugía bariátrica para identificar patrones y agrupaciones en sus perfiles de metabolitos. Se utilizó el análisis de componentes principales (PCA) para visualizar la variabilidad en los datos y el clustering jerárquico para identificar posibles subgrupos de pacientes con perfiles similares. Estos enfoques permitieron detectar agrupaciones en los datos, lo que podría correlacionarse con respuestas clínicas distintas. Las conclusiones apuntan a que existen variaciones significativas en los perfiles de metabolitos entre los grupos, lo cual podría contribuir a personalizar las intervenciones clínicas basadas en el perfil metabólico individual.

Objetivos del Estudio

1. El objetivo de este análisis es explorar las respuestas metabolómicas de pacientes tras cirugía bariátrica en distintos puntos de tiempo, así como desarrollar competencias en el uso de herramientas bioinformáticas, específicamente en la manipulación de datos metabolómicos usando `SummarizedExperiment`.

2.Identificar patrones en los datos metabolómicos de pacientes tras cirugía bariátrica.

3.Visualizar la variabilidad en los perfiles de metabolitos mediante análisis de componentes principales (PCA).

4.Agrupar a los pacientes en clústeres según similitudes en sus perfiles de metabolitos utilizando clustering jerárquico.

5.Analizar estadísticamente los metabolitos dentro de cada clúster para determinar las diferencias metabólicas entre los grupos.

Materiales y Métodos

Origen y Naturaleza de los Datos

- **Fuente de los datos**: Datos descargados de GitHub, relacionados con el artículo de investigación "Metabotypes of response to bariatric surgery independent of the magnitude of weight loss".
- **Tipo de datos**: Valores clínicos y metabolómicos para 39 pacientes en cinco puntos de tiempo.

Herramientas y Procedimiento de Análisis

- **Lenguaje y Paquetes**: Se utilizó R con los paquetes `SummarizedExperiment` y `tidyverse`.
- **Metodología**:
  1. **Carga y organización de datos**: Importación de archivos `DataValues_S013.csv` y `DataInfo_S013.csv`.
  2. **Creación del contenedor**: Organización de los datos en un objeto `SummarizedExperiment`.
  3. **Exploración y visualización**: Análisis descriptivo de las variaciones en los metabolitos a través del tiempo y entre condiciones de pacientes.

Análisis de Componentes Principales (PCA)

Se realizó un PCA sobre los datos para reducir la dimensionalidad y explorar visualmente la variabilidad entre muestras. Los dos primeros componentes principales, que explican un 25% de la variabilidad, fueron graficados para observar la distribución y posibles agrupaciones de muestras en el espacio de los componentes.

Clustering Jerárquico

Se aplicó el clustering jerárquico utilizando el método de Ward y una matriz de distancias euclídeas. Se generó un dendrograma para visualizar los cuatro clústeres principales, que sugieren grupos de pacientes con perfiles de metabolitos similares.

Análisis Descriptivo por Grupo

Para cada clúster, se calcularon estadísticas descriptivas de los metabolitos (media, desviación estándar, valores mínimos y máximos) para entender mejor las características de cada grupo.

Resultados

Proceso de Análisis

Descarga de Datos

Los datos utilizados en este análisis se obtuvieron del repositorio de GitHub asociado al artículo **"Metabotypes of response to bariatric surgery independent of the magnitude of weight loss"**. El dataset incluye dos archivos principales:

- `DataValues_S013.csv`: Contiene los valores clínicos y metabolómicos para 39 pacientes en 5 puntos temporales.
- `DataInfo_S013.csv`: Proporciona metadatos descriptivos para cada columna en el archivo `DataValues_S013.csv`.

Estos archivos fueron descargados desde el repositorio y guardados localmente para su uso en el análisis.

```
# download.file("https://github.com/nutrimetabolomics/metaboData/blob/main/Datasets/2018-MetabotypingPa
# download.file("https://github.com/nutrimetabolomics/metaboData/blob/main/Datasets/2018-MetabotypingPa
```

Creación del Contenedor `SummarizedExperiment`

Para facilitar el análisis de los datos ómicos y manejar de forma estructurada los datos y metadatos, creamos un contenedor de tipo SummarizedExperiment. Este tipo de objeto es ampliamente utilizado en análisis ómicos porque permite almacenar datos de conteo junto con metadatos y datos clínicos.

Pasos para la creación del contenedor:

Lectura de los datos y metadatos: Se cargaron los datos de DataValues_S013.csv y DataInfo_S013.csv en el entorno de R. Creación de colData y rowData: Se generaron estos objetos para organizar la información de las filas y columnas del SummarizedExperiment. Creación del objeto SummarizedExperiment: Se combinaron los datos de metabolitos, colData y rowData en el contenedor SummarizedExperiment.

Cargar librería y datos

Primero, necesitamos cargar todas las librerías necesarias para el análisis y los datos. Esto incluye la carga de paquetes de Bioconductor y tidyverse para manipulación de datos, además de FactoMineR y factoextra para realizar el análisis de componentes principales (PCA) y el análisis de conglomerados.

Objetivo:

Tener los datos correctamente cargados en el entorno y organizados en un objeto SummarizedExperiment, que es una estructura ideal para datos ómicos.

```
library(SummarizedExperiment)
library(tidyverse)
library(FactoMineR)
library(factoextra)
```

```
# Cargar los archivos de datos
data_values <- read.csv("C:/Users/User/Downloads/dataset_metabotyping/DataValues_S013.csv", row.names =
data_info <- read.csv("C:/Users/User/Downloads/dataset_metabotyping/DataInfo_S013.csv")

# Crear colData y rowData para el objeto SummarizedExperiment
colData <- DataFrame(data_info, row.names = colnames(data_values))
rowData <- DataFrame(row.names = rownames(data_values))

# Crear el objeto SummarizedExperiment
se <- SummarizedExperiment(assays = list(counts = data_values), rowData = rowData, colData = colData)
```

Resumen y Exploración Inicial del Dataset

En esta etapa, comenzamos explorando las dimensiones y el resumen estadístico del dataset para entender mejor la cantidad de datos y sus características. Esto incluye identificar cuántas muestras tenemos, cuántas características (metabolitos) se miden y la distribución general de los datos.

```
# Exploración inicial del dataset
cat("Dimensiones del dataset:\n")
```

```
## Dimensiones del dataset:
```

```
print(dim(se))
```

```
## [1]  39 695
```

Análisis de Componentes Principales (PCA)

El análisis de componentes principales (PCA) es una técnica de reducción de dimensionalidad que permite observar patrones en los datos. Es particularmente útil para identificar cómo se agrupan las muestras en función de sus perfiles metabolómicos.

Objetivo:

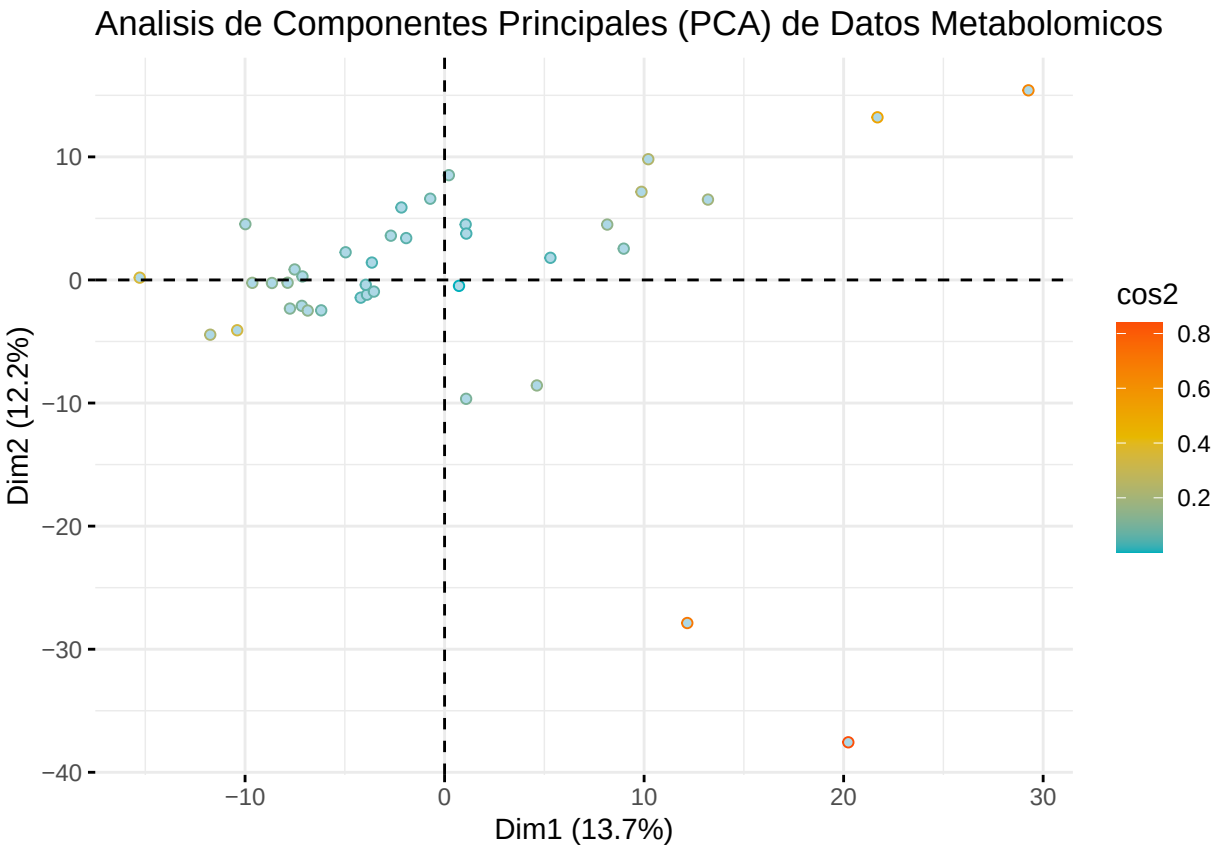
Visualizar la variabilidad de las muestras y observar cómo se agrupan en función de sus perfiles metabolómicos. Esto puede revelar patrones o subgrupos significativos.

```
# Seleccionar solo columnas numéricas (metabolitos) para el PCA
metabolite_data <- assay(se) %>%
  select_if(is.numeric)

# Aplicar PCA a los datos metabolómicos
pca_result <- PCA(metabolite_data, scale.unit = TRUE, graph = FALSE)
```

```
## Warning in PCA(metabolite_data, scale.unit = TRUE, graph = FALSE): Missing
## values are imputed by the mean of the variable: you should use the imputePCA
## function of the missMDA package
```

```
# Visualizar los resultados del PCA
fviz_pca_ind(pca_result,
             geom.ind = "point",
             pointshape = 21,
             fill = "lightblue",
             col.ind = "cos2", # Color según calidad de representación
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,
             title = "Analisis de Componentes Principales (PCA) de Datos Metabolomicos")
```



En el gráfico, las dimensiones (Dim1 y Dim2) representan los dos primeros componentes principales que explican la mayor variabilidad en los datos (13.7% y 11.2% respectivamente).La variabilidad explicada por estos dos componentes no es muy alta (alrededor del 25% en total), lo que sugiere que puede haber una estructura compleja en los datos que no se resume completamente en estos primeros dos componentes.

Los puntos están coloreados según el valor de cos2, que representa la calidad de la representación de cada muestra en el plano definido por Dim1 y Dim2. Puntos con valores de cos2 altos (más cercanos al rojo) indican que están bien representados en estos dos componentes principales, mientras que los puntos con cos2 bajos (más cercanos al azul) están menos representados.

Hay algunos puntos que se encuentran alejados del centro (por ejemplo, en la zona inferior y derecha del gráfico). Estos puntos pueden representar muestras con perfiles metabolómicos distintivos o potenciales outliers.

Análisis de Conglomerados (Clustering Jerárquico)

El análisis de conglomerados jerárquico agrupa las muestras en clústeres basándose en la similitud de sus perfiles metabolómicos. Esto es útil para identificar grupos naturales en los datos, que podrían corresponder a diferentes tipos de respuesta clínica.

Objetivo:

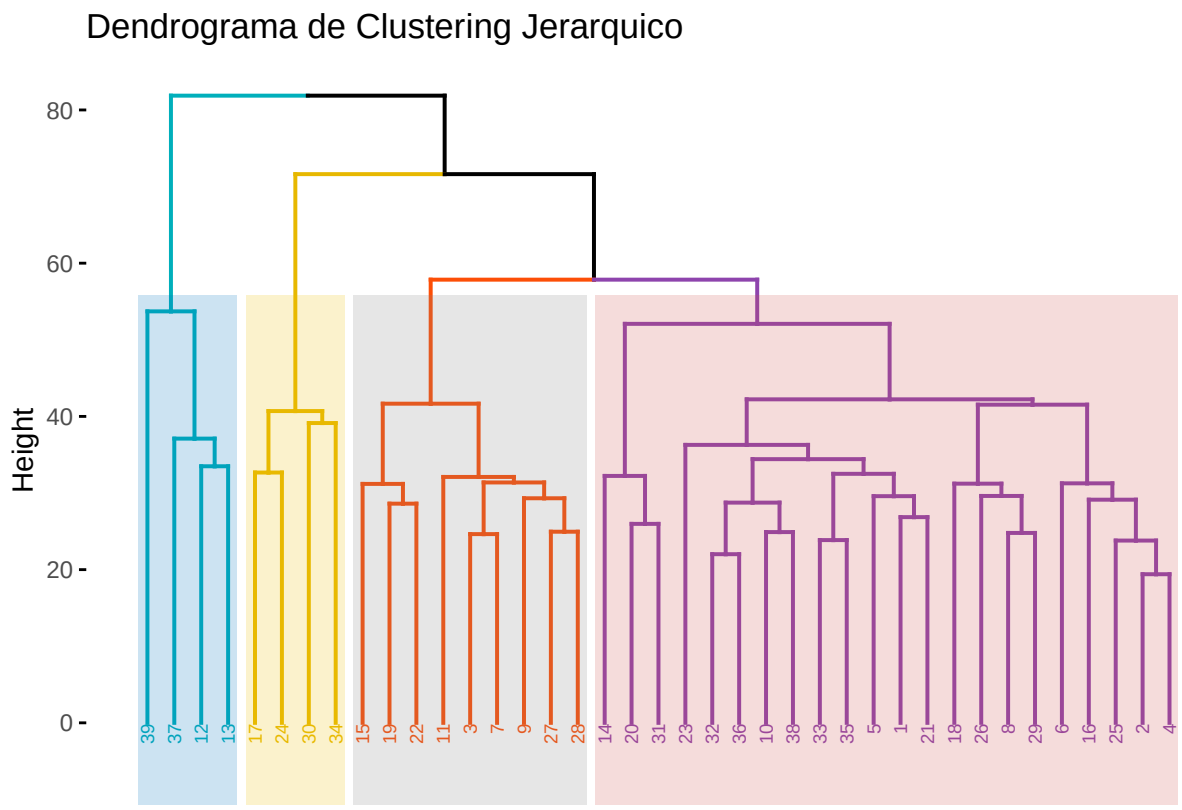
Visualizar posibles agrupamientos entre las muestras y explorar si existen subgrupos con perfiles similares de metabolitos, lo cual puede sugerir diferentes respuestas metabólicas.

```
# Calcular la matriz de distancias y aplicar clustering jerárquico
dist_matrix <- dist(scale(metabolite_data))
```

```r
hc <- hclust(dist_matrix, method = "ward.D2")

# Visualizar el dendrograma
fviz_dend(hc, k = 4, # Seleccionar un número de clusters
          cex = 0.5,
          k_colors = c("#00AFBB", "#E7B800", "#FC4E07", "#8E44AD"),
          color_labels_by_k = TRUE,
          rect = TRUE,
          rect_fill = TRUE,
          rect_border = "jco",
          main = "Dendrograma de Clustering Jerarquico")
```

```
## Warning: The '<scale>' argument of 'guides()' cannot be 'FALSE'. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



El dendrograma destaca cuatro clústeres principales, cada uno con un color diferente (azul, amarillo, gris y morado). Estos colores representan los cuatro grupos identificados en el análisis jerárquico. Cada grupo indica que las muestras dentro de él son más similares entre sí en cuanto a sus perfiles metabolómicos. La altura en el dendrograma representa la distancia o disimilitud entre los grupos. Vemos que el grupo amarillo y el azul tienen ramas más cortas en comparación con el grupo morado, lo cual sugiere que estos primeros

dos grupos son más homogéneos internamente. Por otro lado, el grupo morado tiene una estructura más ramificada, lo que indica una mayor variabilidad en las muestras dentro de este clúster.

Análisis Descriptivo por Grupo

Esta última etapa implica calcular estadísticas descriptivas (media, desviación estándar, mínimo y máximo) de los metabolitos en cada grupo identificado en el análisis de conglomerados. Esto nos ayudará a entender mejor cómo varían los niveles de metabolitos entre los grupos.

Objetivo:

Obtener una comprensión detallada de cómo varían los metabolitos en los diferentes grupos, lo que puede ayudar a identificar metabolitos específicos que estén asociados con ciertas respuestas clínicas.

```
# Convertir metabolite_data a formato largo para hacer un análisis descriptivo por grupo
metabolite_long <- metabolite_data %>%
  pivot_longer(cols = everything(), names_to = "Metabolite", values_to = "Level")

# Calcular estadísticas descriptivas
group_data <- metabolite_long %>%
  group_by(Metabolite) %>%
  summarize(mean = mean(Level, na.rm = TRUE), sd = sd(Level, na.rm = TRUE),
            min = min(Level, na.rm = TRUE), max = max(Level, na.rm = TRUE))

head(group_data)
```

```
## # A tibble: 6 x 5
##   Metabolite   mean     sd    min   max
##   <chr>       <dbl>  <dbl>  <dbl> <dbl>
## 1 ADIPO_T0     7.70   4.11   1.72  17.1
## 2 ADIPO_T2     9.66   4.16   3.85  21.4
## 3 ADIPO_T4    11.0    7.16   2.92  35.1
## 4 ADIPO_T5    12.8    7.71   5.6   39.2
## 5 AGE         40.8    9.88  19     59
## 6 Ala_T0     472.   128.   194    907
```

Discusión y Limitaciones

Este análisis de datos metabolómicos en pacientes post-cirugía bariátrica identifica subgrupos metabólicos distintos mediante PCA y clustering jerárquico. Sin embargo, es importante considerar algunas limitaciones:

Varianza Explicada por el PCA: Los dos primeros componentes explican solo alrededor del 25% de la varianza total, lo que sugiere que existe una estructura compleja que estos primeros componentes no capturan completamente. Explorar componentes principales adicionales podría revelar más patrones en los datos.

Posibles Outliers: En el gráfico de PCA, algunos puntos se encuentran alejados del grupo principal de datos, posiblemente representando outliers. Estos puntos podrían afectar el clustering, por lo que sería útil revisar si representan muestras con perfiles metabólicos únicos o si son resultado de variaciones experimentales.

Homogeneidad de los Clústeres: Aunque el clustering jerárquico identificó grupos distintos, algunos clústeres (por ejemplo, el grupo morado) tienen ramas más largas, lo que indica una mayor variabilidad interna. Esto sugiere que algunos grupos pueden tener perfiles metabólicos más heterogéneos, lo que limita la capacidad de generalizar ciertos patrones.

Asociación con Variables Clínicas: Aunque el análisis inicial muestra patrones en los datos metabolómicos, sería útil evaluar asociaciones entre clústeres y características clínicas (como tipo de cirugía, edad o género) para profundizar en las posibles diferencias en respuesta clínica.

Conclusiones

Este análisis exploratorio del dataset metabolómico de pacientes post-cirugía bariátrica proporcionó información valiosa sobre las similitudes y diferencias en los perfiles metabolómicos entre muestras. Los principales hallazgos son:

Identificación de grupos metabolómicos: Los análisis PCA y de clustering jerárquico sugieren la existencia de subgrupos entre los pacientes, lo que podría estar relacionado con diferentes respuestas metabólicas a la cirugía.

Variabilidad en la respuesta clínica: Aunque no se realizaron correlaciones exhaustivas con datos clínicos específicos, los patrones observados sugieren que diferentes perfiles metabolómicos pueden estar asociados con distintas respuestas clínicas postoperatorias.

Recomendaciones para estudios futuros: Un análisis más profundo de otros componentes principales, así como un enfoque en la integración de datos clínicos y metabolómicos, podría proporcionar una comprensión más detallada de los factores que influyen en la respuesta a la cirugía bariátrica.

Estos hallazgos podrían ser un primer paso para personalizar tratamientos y entender mejor los factores que afectan la respuesta de los pacientes a la intervención quirúrgica en términos de metabolómica.

Repositorio en GitHub

Para compartir los resultados y garantizar la reproducibilidad del análisis, se creó un repositorio en GitHub con los siguientes elementos:

Informe en formato RMarkdown (.rmd) que describe el proceso completo y los resultados. Objeto en formato .Rda que contiene los datos y metadatos en formato binario para facilitar la carga en futuros análisis. Código R utilizado para la exploración de datos, incluyendo carga de datos, análisis PCA, clustering, y visualización. Datos en formato texto (CSV) con los valores clínicos y metabolómicos procesados. Archivo de metadatos en Markdown que detalla cada variable del dataset, incluyendo una breve descripción, unidades y valores posibles.

El análisis y los datos se encuentran disponibles en el siguiente repositorio de GitHub:

- Repositorio en GitHub
  https://github.com/andres20281993/Jarrin-Jurado-Alex-PEC1