Fecha de publicación del enunciado: 24/10/2024

Fecha límite para presentar la PEC: 6/11/20241

Descripción de la PEC El objetivo de esta PEC es que planifiquéis y ejecutéis una versión simplificada del proceso de análisis de datos ómicos, a la vez que practicáis con algunas de las herramientas y métodos que hemos trabajado.

En concreto lo que tendréis que hacer es:

1. **Seleccionar un dataset de metabolómica que podéis obtener de**
    a. **Este repositorio de github: https://github.com/nutrimetabolomics/metaboData/**
    b. **Si lo preferís podéis usar algún dataset del repositorio metabolomicsWorkbench**

Selecciono el dataset "2024-Cachexia" que se encuentra en el repositorio github, arriba mencionado. Este análisis busca identificar diferencias en el perfil metabolómico entre pacientes con caquexia y pacientes de control, utilizando técnicas exploratorias como el análisis de componentes principales (PCA) y la visualización de distribuciones de metabolitos.

2. **Una vez descargados los datos cread un contenedor del tipo *SummarizedExperiment* que contenga los datos y los metadatos (información acerca del dataset, las filas y las columnas). La clase *SummarizedExperiment* es una extensión de *ExpressionSet* y muchas aplicaciones o bases de datos (como metabolomicsWorkbench) lo utilizan en vez de usar expressionSet.**

Ejecuto el siguiente código para instalar *SummarizedExperiment* y otras dependencias:

```
> if (!requireNamespace("SummarizedExperiment", quietly = TRUE)) {
+     install.packages("BiocManager")
+     BiocManager::install("SummarizedExperiment")
+ }
```

Cargo el dataset en R a partir del archivo que he cargado en mi repositorio de github:

```
> dataset <- read.csv("https://raw.githubusercontent.com/aramonarr/RAMON-ARRUFAT-ALBERT-PEC1/main/human_cachexia.csv", row.names = 1)
```

Seleccionamos solo las columnas numéricas

```
numeric_data <- dataset[, sapply(dataset, is.numeric)]
```

Una vez cargado el dataset, creo un objeto *SummarizedExperiment* utilizando los datos del estudio y sus metadatos:

```
> library(SummarizedExperiment)
> colData <- DataFrame(SampleID = colnames(numeric_data), Condition = "CondiciónEjemplo")
> rowData <- DataFrame(FeatureID = rownames(numeric_data), Description = "DescripciónEjemplo")
> se <- SummarizedExperiment(assays = list(counts = as.matrix(numeric_data)), rowData = rowData, colData = colData)
```

```
> colData
DataFrame with 64 rows and 2 columns
             SampleID       Condition
          <character>     <character>
1          Muscle.loss CondiciónEjemplo
2  X1.6.Anhydro.beta.D... CondiciónEjemplo
3   X1.Methylnicotinamide CondiciónEjemplo
4       X2.Aminobutyrate CondiciónEjemplo
5   X2.Hydroxyisobutyrate CondiciónEjemplo
...               ...           ...
60        cis.Aconitate CondiciónEjemplo
61         myo.Inositol CondiciónEjemplo
62       trans.Aconitate CondiciónEjemplo
63     pi.Methylhistidine CondiciónEjemplo
64    tau.Methylhistidine CondiciónEjemplo


> rowData
DataFrame with 77 rows and 2 columns
     FeatureID       Description
    <character>      <character>
1      PIF_178 DescripciónEjemplo
2      PIF_087 DescripciónEjemplo
3      PIF_090 DescripciónEjemplo
4    NETL_005_V1 DescripciónEjemplo
5      PIF_115 DescripciónEjemplo
...        ...           ...
73  NETCR_019_V2 DescripciónEjemplo
74   NETL_012_V1 DescripciónEjemplo
75   NETL_012_V2 DescripciónEjemplo
76   NETL_003_V1 DescripciónEjemplo
77   NETL_003_V2 DescripciónEjemplo

> se
class: SummarizedExperiment
dim: 77 64
metadata(0):
assays(1): counts
rownames(77): PIF_178 PIF_087 ...
  NETL_003_V1 NETL_003_V2
rowData names(2): FeatureID Description
colnames(64): Muscle.loss
  X1.6.Anhydro.beta.D.glucose ...
  pi.Methylhistidine tau.Methylhistidine
colData names(2): SampleID Condition
```

3. **Llevad a cabo una exploración del dataset que os proporcione una visión general del mismo en la línea de lo que hemos visto en las actividades**

Este dataset contiene datos metabolómicos de pacientes con y sin caquexia (cachexia) muscular. Las muestras se analizaron para evaluar las concentraciones de varios metabolitos en diferentes condiciones. Su estructura muestra lo siguiente:

Columnas Principales:

- **Patient ID**: Identificador único de cada paciente.
- **Muscle loss**: Condición del paciente (`cachexic` o `control`), indicando si el paciente tiene pérdida muscular (caquexia) o pertenece al grupo de control.
- **Metabolitos**: Las columnas restantes corresponden a las concentraciones de diferentes metabolitos, como:
  - 'X1.6.Anhydro.beta.D.glucose'
  - `X2.Aminobutyrate`
  - `Citrate`
  - `Lactate`
  - ... y muchos otros.

Información Adicional

- **Tipo de datos**: Todas las concentraciones están en formato numérico y representan la cantidad relativa del metabolito en cada muestra.
- **Número de Muestras**: 77 muestras en total.
- **Número de Metabolitos**: 63 metabolitos cuantificados para cada muestra.

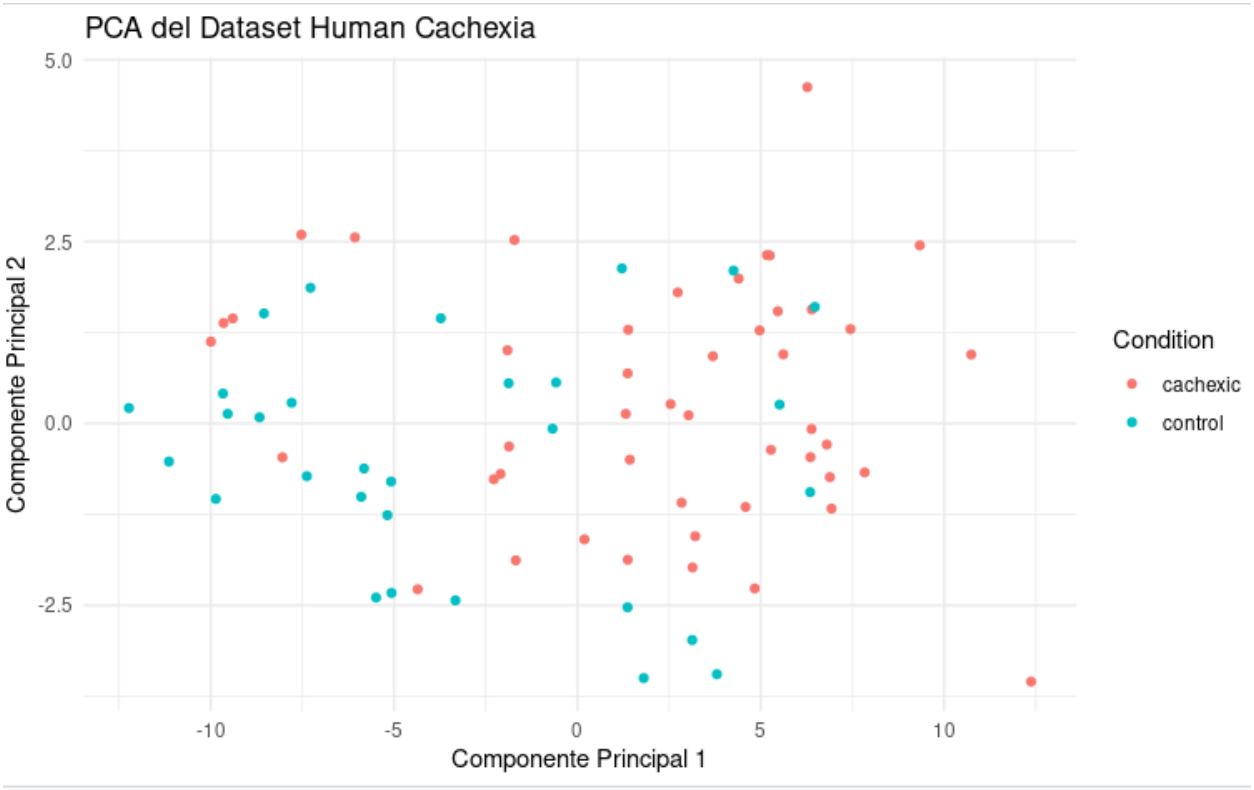
Realizo también un resumen estadístico del dataset estudiado:

```
X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide
 Min.   :  4.71              Min.   :   6.42
 1st Qu.: 28.79              1st Qu.:  15.80
 Median : 45.60              Median :  36.60
 Mean   :105.63              Mean   :  71.57
 3rd Qu.:141.17              3rd Qu.:  73.70
 Max.   :685.40              Max.   :1032.77
 X2.Aminobutyrate X2.Hydroxyisobutyrate
 Min.   :  1.28   Min.   : 4.85
 1st Qu.:  5.26   1st Qu.:15.80
 Median : 10.49   Median :32.46
 Mean   : 18.16   Mean   :37.25
 3rd Qu.: 19.49   3rd Qu.:54.60
 Max.   :172.43   Max.   :93.69
 X2.Oxoglutarate   X3.Aminoisobutyrate
 Min.   :   5.53   Min.   :   2.61
 1st Qu.:  22.42   1st Qu.:  11.70
 Median :  55.15   Median :  22.65
 Mean   : 145.09   Mean   :  76.76
 3rd Qu.:  92.76   3rd Qu.:  56.26
 Max.   :2465.13   Max.   :1480.30
.....

.....
```

A continuación, realizamos el análisis de componentes más importantes y creamos un gráfico de dispersión para visualizar las muestras en los primeros dos componentes principales.

Realizamos la transformación logarítmica para reducir el sesgo y también evitar valores cero o negativos sumando 1 antes de aplicar el logaritmo. Finalmente creamos un dataframe con los resultados de los componentes principales y graficamos:

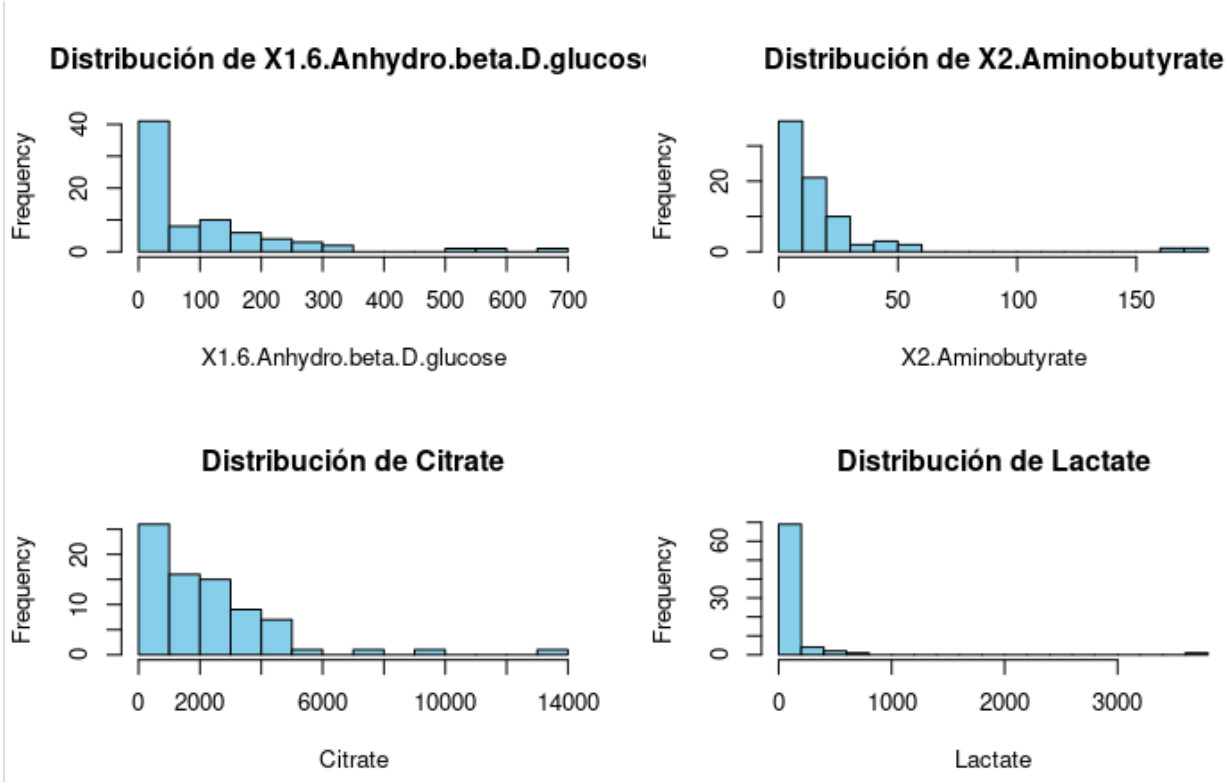
```
> transformed_data <- log(numeric_data + 1)

>pca <- prcomp(transformed_data, scale. = TRUE)
>pca_df <- data.frame(PC1 = pca$x[, 1], PC2 = pca$x[, 2], Condition = dataset$Muscle.loss)

> install.packages("ggplot2")
> library(ggplot2)
 > ggplot(pca_df, aes(x = PC1, y = PC2, color = Condition)) +
   geom_point() +
   labs(title = "PCA del Dataset Human Cachexia", x = "Componente Principal 1", y =
 "Componente Principal 2") +
   theme_minimal()
```



El gráfico muestra una separación entre las muestras de las condiciones cachexic y control en los componentes principales, lo cual sugiere variabilidad en los datos entre estas dos condiciones.

Realizamos también de las variables seleccionadas los histogramas correspondientes:

```
> selected_vars <- c("X1.6.Anhydro.beta.D.glucose", "X2.Aminobutyrate", "Citrate", "Lactate")
> par(mfrow = c(2, 2))
> for (var in selected_vars) {
+     hist(numeric_data[[var]], main = paste("Distribución de", var), xlab = var, col = "skybl
ue", breaks = 15)
+ }
> par(mfrow = c(1, 1))
>
```



Los gráficos muestran claramente la dispersión de valores para los metabolitos X1.6-Anhydro-beta-D-glucose, X2-Aminobutyrate, Citrate, y Lactate, con una distribución sesgada hacia valores más bajos en la mayoría de los casos, lo que es común en datos de metabolómica.

Realizamos un análisis de correlación

```
> correlation_matrix <- cor(transformed_data)
> print(correlation_matrix)
```

	X1.6.Anhydro.beta.D.glucose	X1.Methylnicotinamide	X2.Aminobutyrate
X1.6.Anhydro.beta.D.glucose	1.00000000	0.2465217	0.4008973
X1.Methylnicotinamide	0.24652171	1.0000000	0.3314584
X2.Aminobutyrate	0.40089732	0.3314584	1.0000000
X2.Hydroxyisobutyrate	0.45394245	0.6155661	0.5260492
X2.Oxoglutarate	0.16870610	0.4529416	0.4371957
X3.Aminoisobutyrate	0.32571459	0.2757309	0.5868155
X3.Hydroxybutyrate	0.42466694	0.5556548	0.6181006
X3.Hydroxyisovalerate	0.48014288	0.5800126	0.4270324
X3.Indoxylsulfate	0.41348265	0.4742505	0.4957325
X4.Hydroxyphenylacetate	0.54486461	0.5012370	0.4562886
Acetate	0.42269175	0.3874655	0.2718100
Acetone	0.04765826	0.1264387	0.4361913
Adipate	0.49527425	0.6032129	0.5245596
Alanine	0.49183050	0.6288308	0.5977992
Asparagine	0.52774253	0.5819800	0.6274036

	X2.Hydroxyisobutyrate	X2.Oxoglutarate	X3.Aminoisobutyrate	X3.Hydroxybutyrate
X1.6.Anhydro.beta.D.glucose	0.4539425	0.1687061	0.3257146	0.4246669
X1.Methylnicotinamide	0.6155661	0.4529416	0.2757309	0.5556548
X2.Aminobutyrate	0.5260492	0.4371957	0.5868155	0.6181006
X2.Hydroxyisobutyrate	1.0000000	0.6200256	0.4318539	0.6887889
X2.Oxoglutarate	0.6200256	1.0000000	0.4299625	0.5751608
X3.Aminoisobutyrate	0.4318539	0.4299625	1.0000000	0.6065301
X3.Hydroxybutyrate	0.6887889	0.5751608	0.6065301	1.0000000
X3.Hydroxyisovalerate	0.5920512	0.2912305	0.3153972	0.6109552
X3.Indoxylsulfate	0.5779861	0.3832053	0.4608024	0.5469607
X4.Hydroxyphenylacetate	0.6639955	0.4780489	0.4646047	0.6022109
Acetate	0.5097321	0.1520278	0.2810022	0.5241748
Acetone	0.1910115	0.1281810	0.2808751	0.3436830
Adipate	0.5721619	0.4445545	0.4781481	0.6736276
Alanine	0.7999070	0.6265798	0.5594458	0.7679271
Asparagine	0.7597226	0.5935035	0.6544770	0.7676097

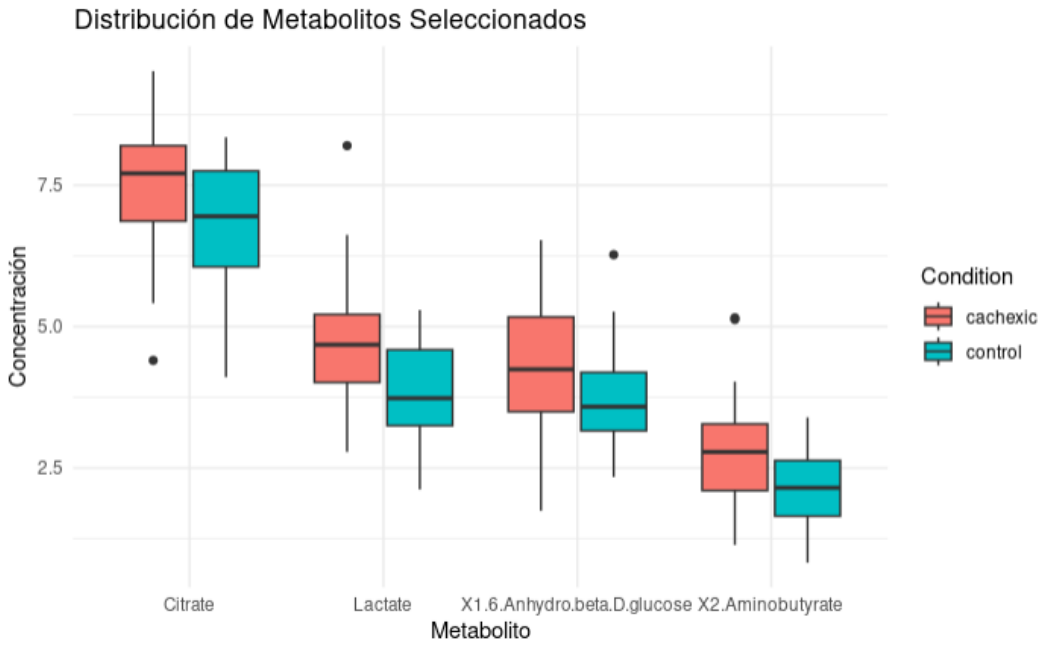
	X3.Hydroxyisovalerate	X3.Indoxylsulfate	X4.Hydroxyphenylacetate	Acetate
X1.6.Anhydro.beta.D.glucose	0.4801429	0.41348265	0.5448646	0.42269175
X1.Methylnicotinamide	0.5800126	0.47425050	0.5012370	0.38746546

.....

.....

Realizaremos gráficos de caja para los metabolitos seleccionados

```
> install.packages("tidyr")
>library(tidyr)
>numeric_data_long <- transformed_data %>%
  mutate(Condition = dataset$Muscle.loss) %>%
  pivot_longer(cols = -Condition, names_to = "Metabolite", values_to = "Value")

>ggplot(numeric_data_long %>% filter(Metabolite %in% selected_vars), aes(x = Metabolite, y = V
alue, fill = Condition)) +
  geom_boxplot() +
  labs(title = "Distribución de Metabolitos Seleccionados", x = "Metabolito", y = "Concentraci
ón") +
  theme_minimal()
```



4. **Elaborad un informe que describa el proceso que habéis realizado, incluyendo la descarga de los datos, la creación del contenedor, la exploración de los datos y la reposición de los datos en github. El nombre del repositorio tiene que ser el siguiente: APELLIDO1-Apellido2-Nombre-PEC1. Por ejemplo, en mi caso el repositorio se llamaría: "Sanchez-Pla-Alex-PEC1"**

   Realizado

1. **Cread un repositorio de github que contenga: aramonarr/RAMON-ARRUFAT-ALBERT-PEC1: PEC1:**

   a. **el informe**
      Realizado y adjuntado en la plataforma de la UOC
   b. **el objeto contenedor con los datos y los metadatos en formato binario (.Rda)**
      Realizado con el nombre "exploración_dataset_cachexia.R" y cargado en mi repositorio github
   c. **el código R para la exploración de los datos o los datos en formato texto**
   d. **los metadatos acerca del dataset en un archivo markdown.**
      Realizado con el nombre "metadatos_dataset_cachexia.R" y cargado en mi repositorio github

La dirección (url) del repositorio deberá estar incluida en la última sección del informe de forma clara.

1. **Repositorio: aramonarr/RAMON-ARRUFAT-ALBERT-PEC1: PEC1**