

# Datos Omicos PEC1

Prueba de evaluación continua 1

Komal Malik Khanam

2024-11-07

Configuro el entorno R, instalando y cargando el paquete “SummarizedExperiment”

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("SummarizedExperiment")
library(SummarizedExperiment)
```

## Summarized Experiment

Se descarga el dataset del repositorio indicado en la PEC. Se descarga el archivo zip y se exploran los diferentes datasets. Escogeré el dataset “MetabotypingPaper” para el ejercicio, aunque no se realizará un análisis estadístico tan exhaustivo como se puede realizar.

```
# Cargo los datos
data_values <- read.csv("/Users/Komal/Documents/UOC/Datos
  ↳ ómicos/PEC1/metaboData-main/Datasets/2018-MetabotypingPaper/DataValues_S013.csv",
  ↳ row.names = 1)
data_info <- read.csv("/Users/Komal/Documents/UOC/Datos
  ↳ ómicos/PEC1/metaboData-main/Datasets/2018-MetabotypingPaper/DataInfo_S013.csv",
  ↳ row.names = 1)
```

Comprobamos la estructura de datos

```
dim(data_values)
```

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

```
dim(data_info)
```

```
## [1] 695 3
```

Se puede observar que se trata de un conjunto de datos con 695 variables de 39 sujetos. También disponemos de información como tipo de variable y una descripción sobre cada variable. Opcionalmente podemos observar los nombres de todas las variables. Para el ejercicio sólo muestro el nombre de las primeras columnas.

```
head(colnames(data_values))
```

```
## [1] "SUBJECTS" "SURGERY" "AGE" "GENDER" "Group" "MEDDM_T0"
```

```
head(colnames(data_info))
```

```
## [1] "VarName" "varTpe" "Description"
```

```
# Tras observar las variables, convertiré la columna 'SUJETOS' en nombres de fila
rownames(data_values) <- data_values$SUBJECTS
```

```
# Elimino la columna 'SUBJECTS' de data_values
data_values <- data_values[, -which(colnames(data_values) == "SUBJECTS")]
```

```
# Compruebo correcta asignación y la eliminación
head(rownames(data_values))
```

```
## [1] "1" "2" "3" "4" "5" "6"
```

```
dim(data_values)
```

```
## [1] 39 694
```

Ahora usaré la columna ‘SURGERY’ que define el tipo de cirugía a la que se sometió a cada sujeto y ‘GENDER’, para definir información del grupo al que pertenece cada sujeto.

```
# Crearé col_data usando la columna 'SURGERY' como información de grupo
```

```
col_data <- DataFrame(
  Group = data_values$SURGERY,
  Gender = data_values$GENDER
)
```

```
# Eliminamos las columnas 'SURGERY' y 'GENDER' de data_values para que solo queden las
↪ variables de medición
```

```
data_values <- data_values[, !(colnames(data_values) %in% c("SURGERY", "GENDER"))]
```

```
# Confirmo que colData y data_values están alineados
```

```
dim(col_data)
```

```
## [1] 39 2
```

```
dim(data_values)
```

```
## [1] 39 692
```

‘col\_data’ muestra correctamente una fila por sujeto y data\_values ahora muestra 692 variables tras eliminar ‘SURGERY’.

Por otro lado, el conjunto que contenía información sobre las variables ‘data\_info’ aún incluye ‘SUBJECTS’, ‘SURGERY’ y ‘GENDER’ como variables. Las elimino antes de proceder.

```
# Filtrar data_info para excluir 'SUBJECTS' y 'SURGERY'
```

```
data_info_filtered <- data_info[!(rownames(data_info) %in% c("SUBJECTS", "SURGERY",
  ↪ "GENDER")), ]
```

```
# Creamos row_data usando los datos filtrados anteriores
```

```
row_data <- DataFrame(data_info_filtered)
```

```
# Confirmación
```

```
dim(row_data)
```

```
## [1] 692 3
```

A continuación, crearé el objeto Summarized Experiment (se). Añado información de sujetos con col\_data y información de las variables con row\_data. Se transpone la matriz de data\_values ya que en formato de SE las columnas deben mostrar información sobre sujetos.

```
# El objeto SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(counts = as.matrix(t(data_values))),
  colData = col_data,
  rowData = row_data)
```

```
# Mostramos el objeto SummarizedExperiment
print(se)
```

```
## class: SummarizedExperiment
## dim: 692 39
## metadata(0):
## assays(1): counts
## rownames(692): AGE Group ... SM.C24.0_T5 SM.C24.1_T5
## rowData names(3): VarName varTpe Description
## colnames(39): 1 2 ... 38 39
## colData names(2): Group Gender
```

Voy a añadir descripción de los datos que obtengo de su paper publicado (Metabotypes of response to bariatric surgery independent of the magnitude of weight loss)

```
# Información del estudio
```

```
Info <- list(
  Name = "Magali Palau-Rodriguez, Sara Tulipani, Anna Marco-Ramell, Antonio Miñarro,
    ↪ Olga Jauregui, Alex Sanchez-Pla, Bruno Ramos-Molina, Francisco J. Tinahones,
    ↪ Cristina Andres-Lacueva",
  Contact = "candres@ub.edu",
  Title = "Metabotypes of response to bariatric surgery independent of the magnitude of
    ↪ weight loss",
  DOI = "https://doi.org/10.1371/journal.pone.0198214",
  PMID = "29856816",
  URL =
    ↪ "https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0198214&type=printable"
)
```

```
# Crearé el objeto MIAME con la información
```

```
myDesc <- new("MIAME",
  name = Info[["Name"]],
  contact = Info[["Contact"]],
  title = Info[["Title"]],
  pubMedIds = Info[["PMID"]],
  url = Info[["URL"]])
print(myDesc)
```

```
## Experiment data
```

```
## Experimenter name: Magali Palau-Rodriguez, Sara Tulipani, Anna Marco-Ramell, Antonio Miñarro, Olga
## Laboratory:
## Contact information: candres@ub.edu
## Title: Metabotypes of response to bariatric surgery independent of the magnitude of weight loss
## URL: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0198214&type=printable
## PMIDs: 29856816
## No abstract available.
```

```
# Lo añadimos en metadata
```

```
metadata(se) <- list(MIAME = myDesc)
```

```
str(se)
```

```
## Formal class 'SummarizedExperiment' [package "SummarizedExperiment"] with 5 slots
##   ..@ colData      :Formal class 'DFrame' [package "S4Vectors"] with 6 slots
##   .. .. ..@ rownames      : chr [1:39] "1" "2" "3" "4" ...
##   .. .. ..@ nrows         : int 39
##   .. .. ..@ elementType   : chr "ANY"
##   .. .. ..@ elementMetadata: NULL
##   .. .. ..@ metadata      : list()
##   .. .. ..@ listData      :List of 2
##   .. .. .. ..$ Group : chr [1:39] "by pass" "by pass" "by pass" "by pass" ...
##   .. .. .. ..$ Gender: chr [1:39] "F" "F" "F" "F" ...
##   ..@ assays        :Formal class 'SimpleAssays' [package "SummarizedExperiment"] with 1 slot
##   .. .. ..@ data:Formal class 'SimpleList' [package "S4Vectors"] with 4 slots
##   .. .. .. ..@ listData      :List of 1
##   .. .. .. .. ..$ counts: num [1:692, 1:39] 27 1 0 0 0 1 85 11.4 2.4 NA ...
##   .. .. .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. .. ..$ : chr [1:692] "AGE" "Group" "MEDDM_TO" "MEDCOL_TO" ...
##   .. .. .. .. ..$ : chr [1:39] "1" "2" "3" "4" ...
##   .. .. .. ..@ elementType   : chr "ANY"
##   .. .. .. ..@ elementMetadata: NULL
##   .. .. .. ..@ metadata      : list()
##   ..@ NAMES          : chr [1:692] "AGE" "Group" "MEDDM_TO" "MEDCOL_TO" ...
##   ..@ elementMetadata:Formal class 'DFrame' [package "S4Vectors"] with 6 slots
##   .. .. ..@ rownames      : NULL
##   .. .. ..@ nrows         : int 692
##   .. .. ..@ elementType   : chr "ANY"
##   .. .. ..@ elementMetadata: NULL
##   .. .. ..@ metadata      : list()
##   .. .. ..@ listData      :List of 3
##   .. .. .. ..$ VarName     : chr [1:692] "AGE" "Group" "MEDDM_TO" "MEDCOL_TO" ...
##   .. .. .. ..$ varType     : chr [1:692] "integer" "integer" "integer" "integer" ...
##   .. .. .. ..$ Description: chr [1:692] "dataDesc" "dataDesc" "dataDesc" "dataDesc" ...
##   ..@ metadata        :List of 1
##   .. ..$ MIAME:Formal class 'MIAME' [package "Biobase"] with 13 slots
##   .. .. ..@ name          : chr "Magali Palau-Rodriguez, Sara Tulipani, Anna Marco-Ramell, Ant
##   .. .. ..@ lab           : chr ""
##   .. .. ..@ contact       : chr "candres@ub.edu"
##   .. .. ..@ title         : chr "Metabotypes of response to bariatric surgery independent of t
##   .. .. ..@ abstract      : chr ""
##   .. .. ..@ url           : chr "https://journals.plos.org/plosone/article/file?id=10.1371/jou
##   .. .. ..@ pubMedIds      : chr "29856816"
##   .. .. ..@ samples       : list()
##   .. .. ..@ hybridizations : list()
##   .. .. ..@ normControls   : list()
##   .. .. ..@ preprocessing : list()
##   .. .. ..@ other          : list()
##   .. .. ..@ __classVersion__:Formal class 'Versions' [package "Biobase"] with 1 slot
##   .. .. .. ..@ .Data:List of 2
##   .. .. .. .. ..$ : int [1:3] 1 0 0
##   .. .. .. .. ..$ : int [1:3] 1 1 0
##   .. .. .. ..$ names: chr [1:2] "MIAXE" "MIAME"
```

## Analisis Estadístico básico

Ahora, procederé a realizar un pequeño análisis descriptivo general de los datos en el momento de la inclusión. Primero, observo las variables recogidas con `colnames()`, pero no se muestra el resultado completo de 692 variables en el documento.

Por la disposición de datos, puedo observar que T0 se refiere a los datos en el momento de inicio de estudio. Por lo tanto, datos basales. Crearé un subset con datos descriptivos y antropomórficos. Aquí asumo que las variables “AGE”, “PESO\_T0”, “BMI\_T0”, “CAD\_T0”, “CINT\_T0”, “CC\_T0” se refieren respectivamente a (edad, genero, peso, IMC, circunferencia de cadera, circunferencia de cintura, ratio cadera-cintura)

```
# Selecciono las variables de interés
datos_iniciales <- c("AGE", "PESO_T0", "bmi_T0", "CAD_T0", "CINT_T0", "CC_T0")

# Creo el subset
subset_se <- se[datos_iniciales, ]

# Verifico las dimensiones del subset y los nombres de las variables
dim(subset_se)
```

```
## [1] 6 39
```

```
rownames(subset_se)
```

```
## [1] "AGE" "PESO_T0" "bmi_T0" "CAD_T0" "CINT_T0" "CC_T0"
```

Para analizar calcularé las medias de cada variable en grupo bypass y tubular, para luego compararlas con Test T de student. Esto me dará un análisis descriptivo muy básico de los pacientes en el momento de iniciar el estudio.

```
# Genero un nuevo conjunto extrayendo los datos mencionados anteriormente
subset_data <- assay(subset_se)
dim(subset_data)
```

```
## [1] 6 39
```

```
rownames(subset_data)
```

```
## [1] "AGE" "PESO_T0" "bmi_T0" "CAD_T0" "CINT_T0" "CC_T0"
```

```
colnames(subset_data)
```

```
## [1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" "15"
## [16] "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30"
## [31] "31" "32" "33" "34" "35" "36" "37" "38" "39"
```

```
# Transpongo los datos para analizar
subset_data_transposed <- t(subset_data)
colnames(subset_data_transposed)
```

```
## [1] "AGE" "PESO_T0" "bmi_T0" "CAD_T0" "CINT_T0" "CC_T0"
```

```
str(subset_data_transposed)
```

```
## num [1:39, 1:6] 27 19 42 37 42 24 33 55 40 47 ...
## - attr(*, "dimnames")=List of 2
## ..$ : chr [1:39] "1" "2" "3" "4" ...
## ..$ : chr [1:6] "AGE" "PESO_T0" "bmi_T0" "CAD_T0" ...
```

```
# Los transformo en datos numéricos
subset_data_transposed <- data.frame(apply(subset_data_transposed, 2, as.numeric))
```

```
# Compruebo de nuevo
str(subset_data_transposed)
```

```
## 'data.frame': 39 obs. of 6 variables:
## $ AGE : num 27 19 42 37 42 24 33 55 40 47 ...
## $ PESO_TO: num 151 139 84 136 121 148 109 109 114 120 ...
## $ bmi_TO : num 62.9 47 29.8 53.1 46.6 48.8 43.7 41.8 44 40.6 ...
## $ CAD_TO : num 167 NA 126 162 132 148 141 136 148 NA ...
## $ CINT_TO: num 116 NA 90 157 123 110 122 124 136 NA ...
## $ CC_TO : num 0.7 NA 0.7 1 0.9 0.7 0.9 0.9 0.9 NA ...
```

```
# Extraigo la información de los grupos
group <- colData(subset_se)$Group
```

```
# Creo un data frame con los datos transpuestos y el grupo
data_with_group <- data.frame(subset_data_transposed, Group = group)
```

```
# Calcular la media de cada variable para el grupo 'bypass'
mean_bypass <- round(colMeans(data_with_group[data_with_group$Group == "by pass",
↪ -ncol(data_with_group)], na.rm = TRUE), 2)
```

```
# Calcular la media de cada variable para el grupo 'tubular'
mean_tubular <- round(colMeans(data_with_group[data_with_group$Group == "tubular",
↪ -ncol(data_with_group)], na.rm = TRUE), 2)
```

```
# Muestro las medias para ambos grupos
mean_bypass
```

```
## AGE PESO_TO bmi_TO CAD_TO CINT_TO CC_TO
## 39.73 145.77 52.82 150.35 137.30 0.93
```

```
mean_tubular
```

```
## AGE PESO_TO bmi_TO CAD_TO CINT_TO CC_TO
## 42.92 128.54 45.93 140.83 132.67 0.94
```

```
# Aplico ahora la prueba t- test de student
t_test_results <- apply(data_with_group[, -ncol(data_with_group)], 2, function(x) {
  t.test(x ~ data_with_group$Group)$p.value
})
```

```
# Mostrar los valores p de cada variable
t_test_results <- round(t_test_results, 2)
```

Creo una tabla para mostrar los resultados

```
if (!requireNamespace("knitr", quietly = TRUE))
  install.packages("knitr")
```

```
library(knitr)
```

```
# Primero una tabla con los resultados
results_table <- data.frame(
```

```

    Mean_Bypass = mean_bypass,
    Mean_Tubular = mean_tubular,
    p_value = t_test_results
  )

kable(results_table, format = 'markdown', caption = "Medias por Grupo y Resultados del
  ↪ Test t de Student")

```

Table 1: Medias por Grupo y Resultados del Test t de Student

	Mean_Bypass	Mean_Tubular	p_value
AGE	39.73	42.92	0.32
PESO_T0	145.77	128.54	0.06
bmi_T0	52.82	45.93	0.01
CAD_T0	150.35	140.83	0.13
CINT_T0	137.30	132.67	0.51
CC_T0	0.93	0.94	0.82

Se puede observar que en ambos grupos en el momento de iniciar el estudio, las características de los pacientes son similares sin diferencias estadísticas (considerando  $p < 0.05$  para significación estadística).

Si el interés fuera estudiar sólo las pacientes de genero femenino, se pueden aislar esos datos.

```

subset_female <- se[, colData(se)$Gender == "F"]

# Confirmo que el subset contiene solo pacientes femeninas
colData(subset_female)$Gender

## [1] "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F"
## [20] "F" "F" "F" "F" "F" "F" "F" "F"

```

Se puede hacer un pequeño analisis descriptivo en este subconjunto. Para ello escojo las pacientes mujeres y las variables anteriormente seleccionados.

```

dim(subset_female)

## [1] 692 27

# Selecciono las variables de interés previamente seleccionados en 'datos_iniciales'
subset_fem <- subset_female[datos_iniciales, ]
dim(subset_fem)

## [1] 6 27

```

```

# Aplico los pasos de antes
subset_fem <- assay(subset_fem)
subset_fem_t <- t(subset_fem)
subset_fem_t <- data.frame(apply(subset_fem_t,2,as.numeric))
group_fm <- colData(subset_female)$Group
data_with_group_fm <- data.frame(subset_fem_t, Group = group_fm)

# Calculo las medias
mean_bypass_fm <- round(colMeans(data_with_group_fm[data_with_group_fm$Group == "by
  ↪ pass", -ncol(data_with_group_fm)], na.rm = TRUE),2)
mean_tubular_fm <- round(colMeans(data_with_group_fm[data_with_group_fm$Group ==
  ↪ "tubular", -ncol(data_with_group_fm)], na.rm = TRUE), 2)

```

```

# Aplico la prueba t- test de student
t_test_results_fm <- apply(data_with_group_fm[, -ncol(data_with_group_fm)], 2,
  ↪ function(x) {
    t.test(x ~ data_with_group_fm$Group)$p.value
  })

# Mostro los resultados en tabla
results_table_fm <- data.frame(
  Mean_Bypass = mean_bypass_fm,
  Mean_Tubular = mean_tubular_fm,
  p_value = round(t_test_results_fm, 2)
)

kable(results_table_fm, format = 'markdown', caption = "Medias por Grupo y Resultados del
  ↪ Test t de Student en pacientes de genero femenino")

```

Table 2: Medias por Grupo y Resultados del Test t de Student en pacientes de genero femenino

	Mean_Bypass	Mean_Tubular	p_value
AGE	40.25	46.57	0.18
PESO_T0	136.20	111.86	0.00
bmi_T0	52.20	43.01	0.00
CAD_T0	148.72	137.86	0.08
CINT_T0	129.89	123.00	0.19
CC_T0	0.89	0.90	0.92

En este subgrupo, hay diferencia significativa en BMI en el momento de iniciar el estudio. Por lo tanto, puede generar un sesgo en el momento de interpretación de reultados.

Tras estos dos ejemplos sencillos de análisis descriptivo inicial, se puede ampliar mucho más el análisis estadístico o crear más subsets, pero no es el objetivo de la actividad. A continuación, se crea el informe que se guarda en pdf con nombre de “datos-omicos\_PEC1.pdf”, se guarda el archivo .rmd para subir en el repositorio y creo un pequeño archivo de texto README.txt con descripción general de la actividad.