Sanchez-Martin-Maria-PEC1

María Sánchez Martín

2024-11-06

Contents

Abstract

Esta primera PEC pretende ser un método de introducción a las ómicas mediante un ejercicio de repaso y ampliación. En ella se ha empleado el paquete `SummarizedExperiment` con la finalidad de crear un contenedor con datos de origen ómico, en este caso se ha seleccionado el dataset de `2024-Cachexia`. Se ha realizado un visionado de la distribución datos de expresión y se ha determinado que contiene información de 63 metabolitos o características diferentes para un total de 77 muestras, de las cuales 30 son controles y 47 enfermos. Además, se ha determinado que un total de 16 de estos metabolitos podrían ser candidatos como indicadores de la enfermedad

Objetivos del estudio

El objetivo de la realización de esta PEC es la introducción a las ómicas mediante un ejercicio de repaso y ampliación que permita trabajar con algunas de las herramientas y contenidos que se han trabajado a lo largo de este primer reto.

Durante este primer reto, se han llevado a cabo un total de 3 actividades. La primera introdujo el uso de la herramienta `Bioconductor` y las diversas utilidades de la misma, así como con el uso del paquete `GEOquery`, `Biobase` y empleando `ExpressionSets`.

La segunda actividad introdujo a la ultrasecuenciación y se empleó el paquete `ShortRead` de `Bioconductor` para abrir archivos `FastQ` y la herramienta `Galaxy` para hacer control de calidad de los datos.

Y en la última actividad, se comentaron diversos artículos que empleasen tecnologías ómicas, principalmente, para poder discutir el proceso de análisis de datos ómicos.

El objetivo principal de esta PEC es planificar y ejecutar una versión simplificada del proceso de análisis de datos ómicos empleando y ampliando las herramientas y métodos aprendidos.

Se ha seleccionado el dataset de metabolómica `2024-Cachexia`. De acuerdo con la información dada, Cachexia es un síndrome metabólico asociado a otras enfermedades y se caracteriza por la pérdida de masa muscular con o sin pérdida de masa grasa. Se quiere realizar una exploración de los datos de este dataset, así como determinar si algunos de los metabolitos podrían ser indicativos esenciales de esta enfermedad.

Materiales y métodos

Materiales

- R y RStudio.
- Dataset de metabolómica: `human_cachexia.csv`.
- Paquete de `Bioconductor`: `SummarizedExperiment`.
- GitHub.

Métodos

Selección del dataset: Se ha seleccionado el *dataset de metabolómica* del repositorio de GitHub. Ha sido clonado en la carpeta destinada a la PEC1 de esta asignatura de acuerdo con el siguiente tutorial:

- https://github.com/nutrimetabolomics/metaboData.git

Procedimiento:

1. Encima de la lista de archivos, haz clic en <> **Código**.
2. Copia la dirección URL del repositorio.
3. Abrir el terminar.
4. Cambia el directorio de trabajo actual a la ubicación en donde quieres clonar el directorio.
5. Escriba `git clone` y pegue la dirección URL que ha copiado antes.
6. Presione **Enter** para crear el clon local.

Una vez el repositorio ha sido clonado, se procede a la visualización de los archivos que contiene y selección del que presente mayor interés.

Tras leer los archivos `README.html` y `Data_Catalog.xlsx`, se ha decidido seleccionar el dataset `2024-Cachexia`. De acuerdo con la información dada, Cachexia es un síndrome metabólico asociado a otras enfermedades y se caracteriza por la pérdida de masa muscular con o sin pérdida de masa grasa.

```
data <- read.csv("metaboData/Datasets/2024-Cachexia/human_cachexia.csv", row.names = 1) # La primera co
```

Creación del contenedor `SummarizedExperiment` Información sobre el paquete de `Bioconductor SummarizedExperiment`:

https://bioconductor.org/packages/release/bioc/html/SummarizedExperiment.html

https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

https://lcolladotor.github.io/rnaseq_LCG-UNAM_2021/objetos-de-bioconductor-para-datos-de-expresi%C3%B3n.html

De acuedo con el uso del paquete, se han de tener los datos de expresión y los metadatos separados.

```
metadata <- as.data.frame(data$Muscle.loss) # Metadatos, la primera columna contiene inrformación sobre
expressiondata <- t(as.matrix(data[,-1])) # Convertimos los datos de expresión en matriz
rownames(metadata) <- rownames(data)
colnames(metadata) <- 'Muscle.loss'
colnames(expressiondata) <- rownames(data)
```

La creacción del contenedor se ha basado en el ejemplo del tutorial de uso del paquete `SummarizedExperiment`.

Creación del perfil de GitHub Ya que no tenía cuenta en GitHub me he creado una. Se ha creado un repositorio con el nombre `Sanchez-Martin-Maria-PEC1` con la idea de subir los archivos que se especifican en el enunciado de la PEC1.

Guardar el objeto contenedor con los datos y metadatos:

```
# save(se, file = "Sanchez-Martin-Maria-PEC1-SummarizedExperiment.Rda")
```

Se han ido generando los archivos necesarios de acuerdo a lo que se exige en el enunciado de la PEC1:

- Informe: `Sanchez-Martin-Maria-PEC1.Rmd`, `Sanchez-Martin-Maria-PEC1.pdf`, `Sanchez-Martin-Maria-PEC1.html`.
- El código R para la exploración de los datos: `Sanchez-Martin-Maria-PEC1-RScript.R`
- Los metadatos acerca del dataset: `Sanchez-Martin-Maria-PEC1-README.Rmd`
- El objeto contenedor con los datos y los metadatos en formato binario: `Sanchez-Martin-Maria-PEC1-SummarizedExper`
- Los datos en formato texto: `human_cachexia.csv`

Enlace GitHub: https://github.com/mariasanchez13/Sanchez-Martin-Maria-PEC1.git

Resultados

En una primera exploración del dataset se ha determinado que contiene un total de 77 observaciones (muestras) y 64 variables (características). De estas 77 muestras, se sabe que 47 pertenecen a pacientes con cachexia y 30 a controles, por lo que tendremos dos grupos bien diferenciados. En cuanto a las 64 caracteerísticas, sabemos que la primera columna son metadatos que indica si se trata de un paciente (etiquetado con `cachexic`) o un individuo del grupo de control (etiquetado con `cachexic`), las otras 63 columnas contienen información de niveles de diversos metabolitos como pueden ser `Acetate`, `cis.Aconitate`, `N.N.Dimethylglycine` o `Pyruvate`, entre otros.

De acuerdo con el tutorial de uso de `SummarizedExperiment`, los datos de expresión y los metadatos deben encontrarse en archivos separados, por lo que a partir de este archivo original se han generado dos,

`expressiondata` y `metadata`. Además, se ha de tener en cuenta que el número de columnas de datos de expresión debe coincidir con el número de filas de metadatos, además, deberían tener el mismo nombre, para facilitar la interacción con las mismas.

Para la utilización del paquete `SummarizedExperiment`, se ha determinado que de acuerdo al léxico, `assays` será `expressiondata` y `colData` será `metadata`.

Una vez generado el contenedor, la información general del mismo es:

```
library(SummarizedExperiment)
se <- SummarizedExperiment(
  assays = SimpleList(counts = expressiondata),
  colData = metadata
)
```

```
se
```

```
## class: SummarizedExperiment
## dim: 63 77
## metadata(0):
## assays(1): counts
## rownames(63): X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide ...
##   pi.Methylhistidine tau.Methylhistidine
## rowData names(0):
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(1): Muscle.loss
```

- class: `SummarizedExperiment`. Indica la clase.
- dim: 63 77. Indica las dimensoones del experimento.
- metadata: 0
- assays: 1 (`expressiondata`)

Se han empleado otras funciones y expresiones basadas en el tutorial de uso de `SummarizedExperiment` de `Bioconductor` para la exploración específica del dataset, se pueden ver en el archivo `Sanchez-Martin-Maria-PEC1-RScript.R`

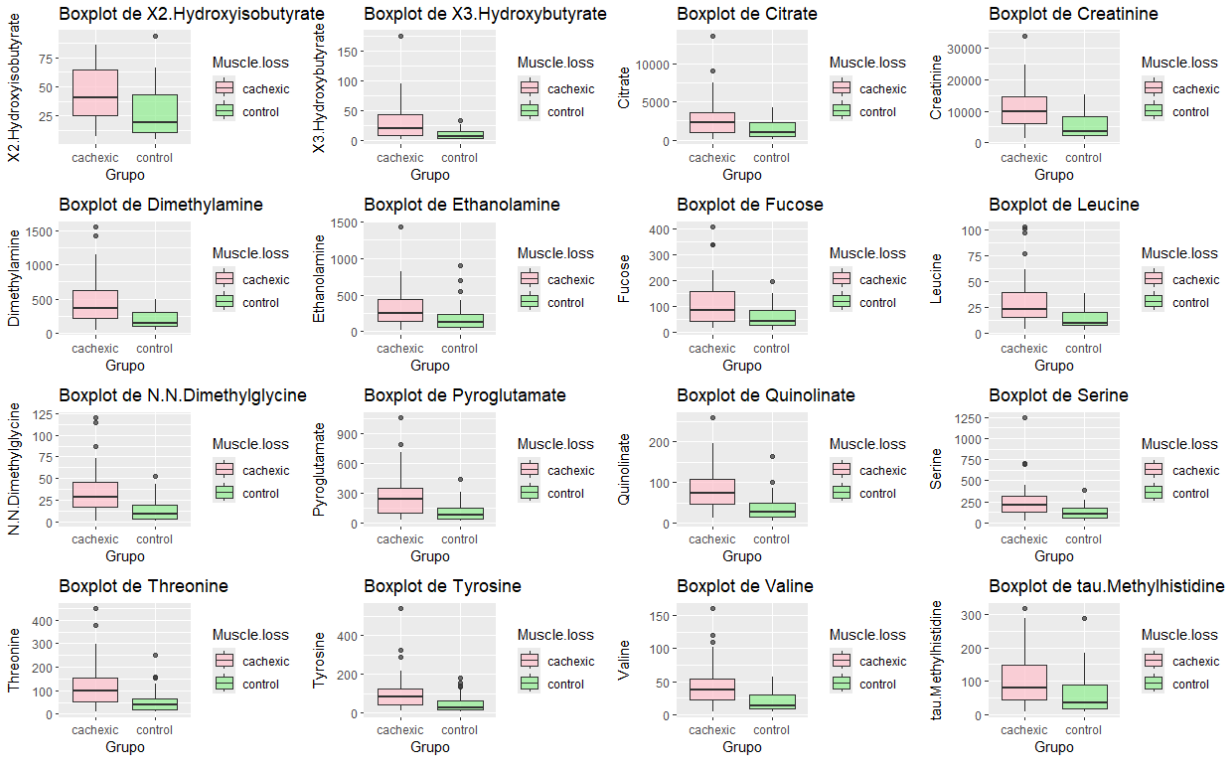
Una vez se ha resalizado una exploración general del archivo, se ha determinado que el volumen de datos es demasiado elevado como para realizar una representación visual, por lo que se ha determinado que el próximo paso lógico a seguir será un filtrado de los datos, de manera que se puedan eliminar algunos.

Se quiere realizar una comparación en la expresión de diferentes metabolitos tanto en pacientes como individuos control para determinar la influencia de estos en el padecimiento de la enfermedad.

Como ya se ha mencionado con anterioridad, se aprecia la existencia de dos grupos, los individuos `control` y los individuos `cachexic`. Por lo que se ha pensado en la realización de un ANOVA (Análisis de varianza) para determinar si hay o no diferencias significativas entre las medias de ambos grupos.

Previamente, se debe asegurar que los datos siguen una distribución normal, por lo que se realizará una prueba Kolgomorov-Smirnov. De las 63 características, únicamente 19 siguen una distribución normal, por lo que se realizará un test ANOVA sobre estas. De estas 19, 16 de ellas presentan un p-value inferior a 0.05, por lo que se catalogarán como significativas y por ende, podremos determinar que existen diferencias significativas entre las medias los grupos para estos metabolitos. Son: `X2.Hydroxyisobutyrate`, `X3.Hydroxybutyrate`, `Citrate`, `Creatinine`, `Dimethylamine`, `Ethanolamine`, `Fucose`, `Leucine`, `N.N.Dimethylglycine`, `Pyroglutamate`, `Quinolinate`, `Serine`, `Threonine`, `Tyrosine`, `Valine` y `tau.Methylhistidine`.

Se ha realizado un diagrama de cajas y bigotes para la visualización de las diferencias entre las medias de ambos grupos.



Discusión y limitaciones y conclusiones del estudio

El objetivo principal del estudio era la construcción de un contenedor `SummarizedExperiment` con la finalidad de practicar y trabajar con dicha función, así como realizar una exploración de los datos de expresión de metabolómica y una introducción al trabajo con los mismos. Se ha trabajado con el archivo y enfrentado situaciones, que en el día a día como bioinformático podrían o no ser reales. Se han eliminado datos que no sirven, realizado filtrados de información y enfocado el trabajo.

Además, se han determinado un conjunto de características o metabolitos de interés, ya que podrían emplearse como predictores o al menos indicadores de la enfermedad. De los 63 inicales, se han determinado que un total de 16 siguen una distribución normal y presentan diferencias significativas. Tal y como se puede observar en la figura, los individuos catalogados como `cachexic` presentan un nivel medio superior de los diferentes metabolitos.

Una de las limitaciones a comentar podría ser, se han determinado que 16 de ellos podrían ser válidos, ¿qué sucederá con los 47 restantes? Un total de 44 no siguen distribución normal, por eso no se han incluido en el estudio, pero podría ser interesante la realización del homólogo no psaramétrico del ANOVA (test de U de Mann-Whitney o Kruskal-Wallis) sobre los mismos para determinar si alguno más podría ser de interés.

Por otro lado, se podrían realizar pruebas sobre un mayor número de indivuiduos y tener en cuenta más metabolitos, quizás focalizados en los relacionados con los 16 que se han seleccionado.