PAC1 - Anàlisi de dades òmiques

Estel · la Micó Frau

2025-04-02

Contents

Resum

Aquest treball té com a objectiu la manipulació i anàlisi de dades de metabolòmica mitjançant la classe "SummarizedExperiment" i altres eines en R. Per a això, s'ha utilitzat un conjunt de dades obtingut del repositori GitHub metaboData, que conté resultats de l'anàlisi metabolòmica de mostres intestinals abans i després d'un trasplantament en sis pacients (12 mostres totals). Inicialment, s'ha generat un objecte "SummarizedExperiment" estructurat amb les dades experimentals, informació de les mostres i metadades. Posteriorment, s'ha dut a terme una anàlisi exploratòria de les dades, incloent un Anàlisi de Components Principals (PCA) i un càlcul de fold-change per identificar diferències en els metabòlits entre les condicions pre- i postoperatori. Els resultats obtinguts han mostrat diferències en els perfils metabòlics dels dos grups, la qual cosa és coherent amb la hipòtesi inicial de l'estudi. Finalment, aquesta tasca s'ha dut a terme mitjançant l'ús de GitHub i el controlador de versions Git.

Objectius

1. Familiaritzar-nos amb la classe "SummarizedExperiment" generant un objecte d'aquesta classe a partir de dades reals d'un experiment de metabolòmica.

2. Aplicar diferents eines de R per tal d'analitzar aquestes dades i extreure'n informació rellevant. Aplicat, a les dades seleccionades, l'objectiu és estudiar si hi ha diferències en el perfil metabolòmic abans i després d'una operació.

3. Millorar aptituds pel que fa a l'ús de Git i GitHub per a la gestió i la compartició de dades òmiques.

Mètodes

Al llarg de la realització de tota la prova s'ha utilitzat el programari RStudio/2024.12.0+467 amb R 4.4.2 i Git com a eina de control de versions. El codi complet amb menció als paquets utilitzats està disponible a l'script (Script-PAC1-MicoFrau). Aquest i els diferents fitxers associats a la resolució de la prova estaran disponible al repositori a partir del dia d'entrega de l'activitat: https://github.com/esmifrau/Mico-Frau-Estella-PAC1/

Les dades seleccionades han sigut 2023-UGrX-4MetaboAnalystTutorial, accessibles des del repositori de GitHub esmentat en l'enunciat. Aquestes dades provenen d'un estudi publicat el 2009 on s'estudia si el microbioma intestinal humà canvia després d'un transplantament menut. S'hipotetitzava que el metabolisme d'aquestes dues comunitats (abans i després de l'operació) són diferents.

Les dades provenen de l'anàlisi metabolòmica de dues mostres intestinals (abans i després de la intervenció quirúrgica) de sis pacients (12 mostres totals). En aquest treball, es va utilitzar la tècnica d'espectrometria de masses acoblada a cromatografia de gasos (GC-TOF) amb els instruments Agilent 6890N, per a la cromatografia de gasos, i Leco Pegasus III GC TOF per a l'anàlisi espectromètrica. Les unitats utilitzades per a la mesura són l'altura dels pics de la intensitat espectral, les quals són comparables a la concentració dels metabòlits.

Al repositori disposem del fitxer original amb informació de metadades abans de la taula, aquest, ja ha estat modificat: s'han eliminat aquestes dades i dades del final i s'han recodificat les mostres i ells factors. Per tant, el fitxer del qual es partirà en aquesta prova (format .csv) conté 143 files i 13 columnes. La primera fila correspon als grups (After o Before) i la resta són els resultats per als 142 metabòlits estudiats. La primera columna correspon a la nomenclatura dels metabòlits i la resta de columnes són les medicions per a les 12 mostres.

Resultats

Dscàrrega d'un dataset de metabolòmica

Tal i com s'ha esmentat en l'apartat anterior, per a aquest treball s'ha utilitzat el conjunt de dades del repositori GitHub metaboData, 2023-UGrX-4MetaboAnalystTutorial.

Creació d'un objecte SummarizedExperiment

En primer lloc, s'han carregat les dades des del nostre ordinador amb la funció read.csv. En aquest pas és important especificar que el separador de les diferents columnes és el tabulador, per tal d'evitar problemes de lectura. Al final obtenim un dataframe de les característiques comentades en l'apartat anterior.

```
data <- read.csv("ST000002_AN000002_clean.csv", sep="\t", stringsAsFactors=FALSE)
data <- as.data.frame(data)
```

Abans de començar amb la creació de l'objecte s'han realitzat algunes modificacions: s'ha afegit el nom dels metabòlits com a nom de les files (rownames) i s'ha eliminat la columna d'aquests.

SummarizedExperiment, igual que ExpressionSet, és una classe d'objectes d'R que permet desar de forma estructurada informació d'experiments, en general, d'expressió gènica i inclouen una matriu d'expressió, dades sobre les mostres i sobre els atributs(features).

Tot i les semblances en la funcionalitat i l'estructura, també presenten diferències.

ExpressionSet	**SummarizedExperiment**
Generalment utilitzat per a dades d'arrays	Generalment utilitzat per a dades de seqüenciació (RNA-Seq)
La matriu de dades es desa en exprs	La matriu de dades es desa en assays (com a counts)
La informació de les mostres es desa a pData (phenotypical data)	La informació de les mostres es desa a colData
La informació dels gens es desa a fData	La informació dels gens es desa a GRangesList o rowData
Contenen una matriu	Poden contindre més d'una matriu (de les mateixes dimensions)

Tenint açò en compte s'ha procedit a generar l'objecte amb la funció SummarizedExperiment() d'un paquet amb el mateix nom.

Dades experimentals: en primer lloc, s'ha eliminat la fila que conté informacions sobre el grup i s'ha convertit el data frame a matriu (s'han convertit també els valors a numèrics).

colData: en segon lloc, s'ha creat un DataFrame per a la informació de la mostra i s'ha inclós com a informació l'ID i el factor que indica el grup al qual pertanyen (transplantation). També s'han afegit metadades addicionals (ID de l'estudi, origen de la mostra i tipus de mostra).

rowData: en tercer lloc, s'ha creat un DataFrame amb informació dels metabòlits estudiats. No utilitzem GRangesList perquè no es tracta de dades genòmiques.

Metadata: en quart lloc, s'han afegit informacions sobre el projecte i l'estudi (metodologia, instruments d'anàlisi). Açò s'ha afegit després de crear l'objecte.

Objecte SE: S'ha creat l'objecte amb la funció SummarizedExperiment() i amb el codi següent s'ha desat l'objecte com a fitxer .Rda (disponible al repositori).

```
save(se, file = "se-PAC1-MicoFrauEstella.Rda")
```

Aquest fitxer permet carregar de forma automàtica a R les dades de l'experiment.

Anàlisi exploratòria de les dades

Abans de començar amb l'anàlisi exploratòria, s'han fet alguns canvis amb les dades. Per tal de facilitar l'anàlisi s'ha transposat el dataframe de dades i s'ha configurat la variable Groups com a factor, així doncs, les mostres han passat a ser les files i els metabòlits i el factor grup, columnes.

Resum de les dades

Amb la funció summary() hem obtingut un resum de les dades:

Anàlisi de components principals (PCA)

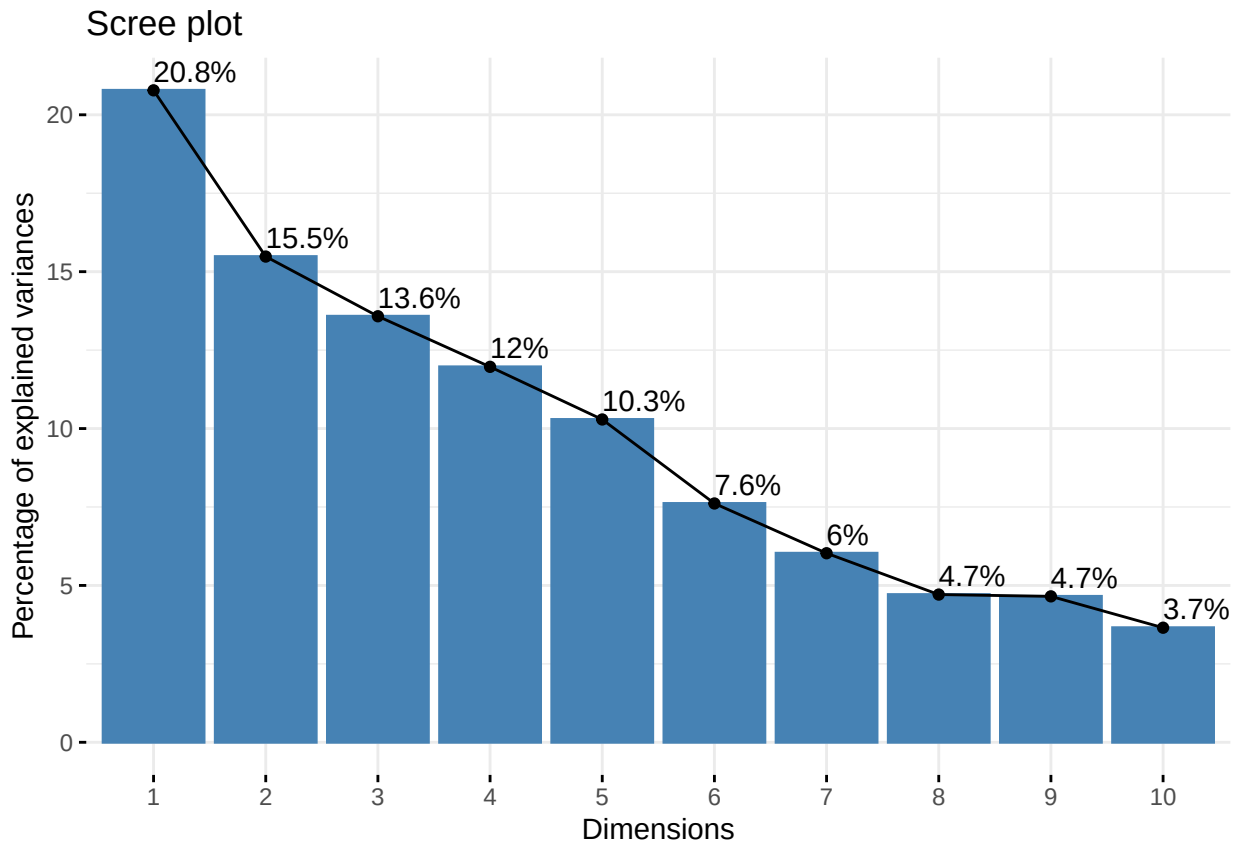
Per al PCA, hem utilitzat la funció prcomp(). En aquesta s'especifica que se centren i escalen les dades.

```
grp <- as.factor(data_t$Groups)
```

```
data_t.pca <- prcomp(data_t[,-1], center = TRUE, scale = TRUE)
```

Una vegada realitzat, hem representat diferents gràfics per tal d'entendre les dades. En primer lloc, amb la funció fviz_eig() s'ha representat el percentatge de variància explicat per a cada component a partir dels valors Eigen.

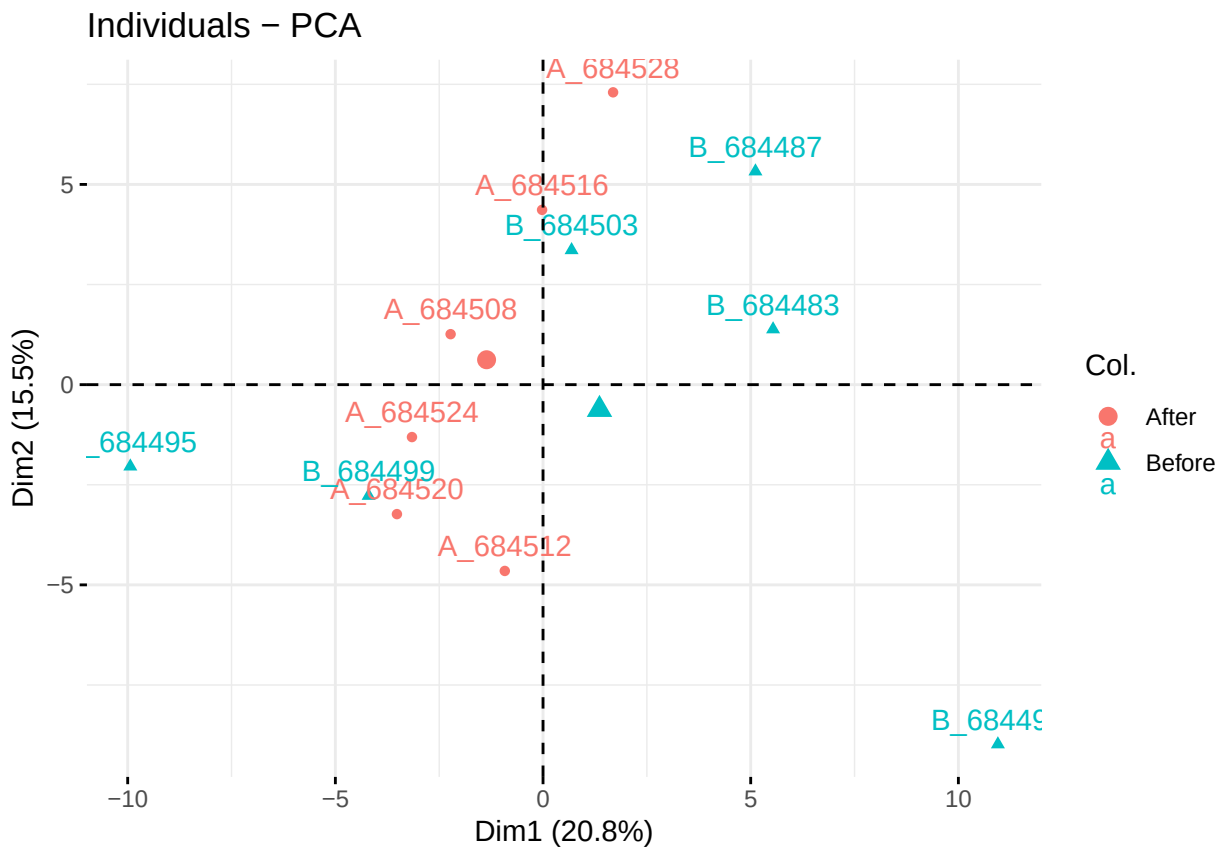
```
fviz_eig(data_t.pca, addlabels = TRUE)
```



En aquest observem que a partir del sisé component se supera el 80% de la variància.

Per al següent gràfic hem utilitzat la funció fviz_pca_ind(), la qual representa els resultats per a cada una de les mostres (coordenades, correlació, etc.). S'han utilitzat dos colors per diferenciar les mostres de A i de B.

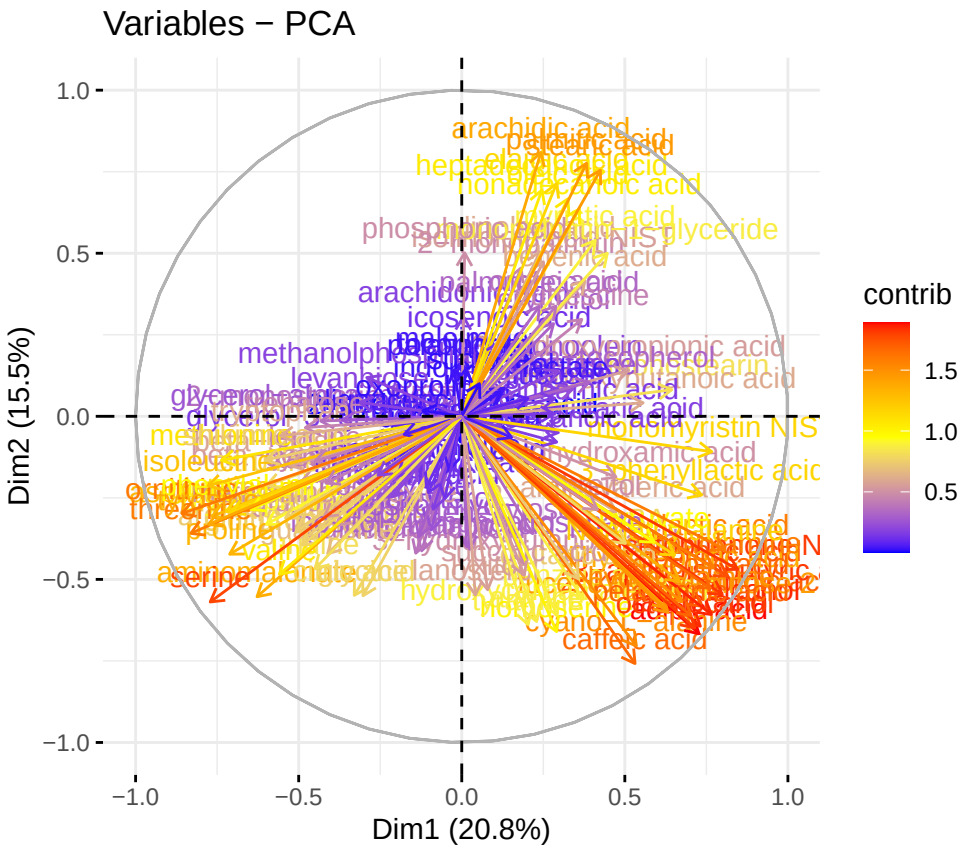
```
fviz_pca_ind(data_t.pca, col.ind = grp, addEllipses = FALSE)
```



Observem que les dades corresponents al grup A (després de l'operació) estan agrupats i per tant es veu una diferència entre els dos grups.

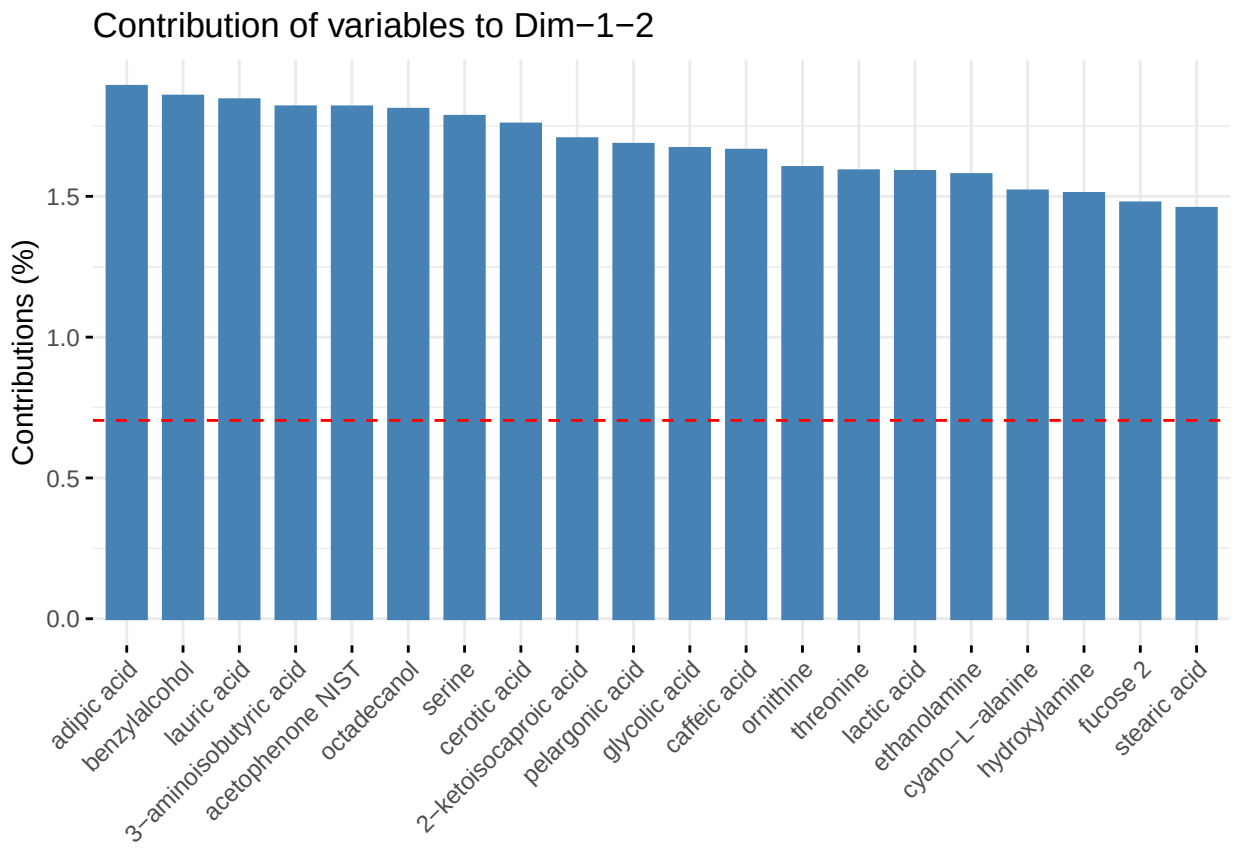
També es poden representar les variables amb fviz_pca_var. Les variables positivament correlacionades apunten cap al mateix quadrant del gràfic i estan agrupades, les que no, apunten a quadrants oposats. col.var = "contrib" ens permet representar amb colors la contribució d'aquestes variables. Com major és aquest valor, més contribueix la variable al component. Això s'ha fet en el següent pas.

```
fviz_pca_var(data_t.pca, col.var = "contrib", gradient.cols = c("blue", "yellow", "red"))
```



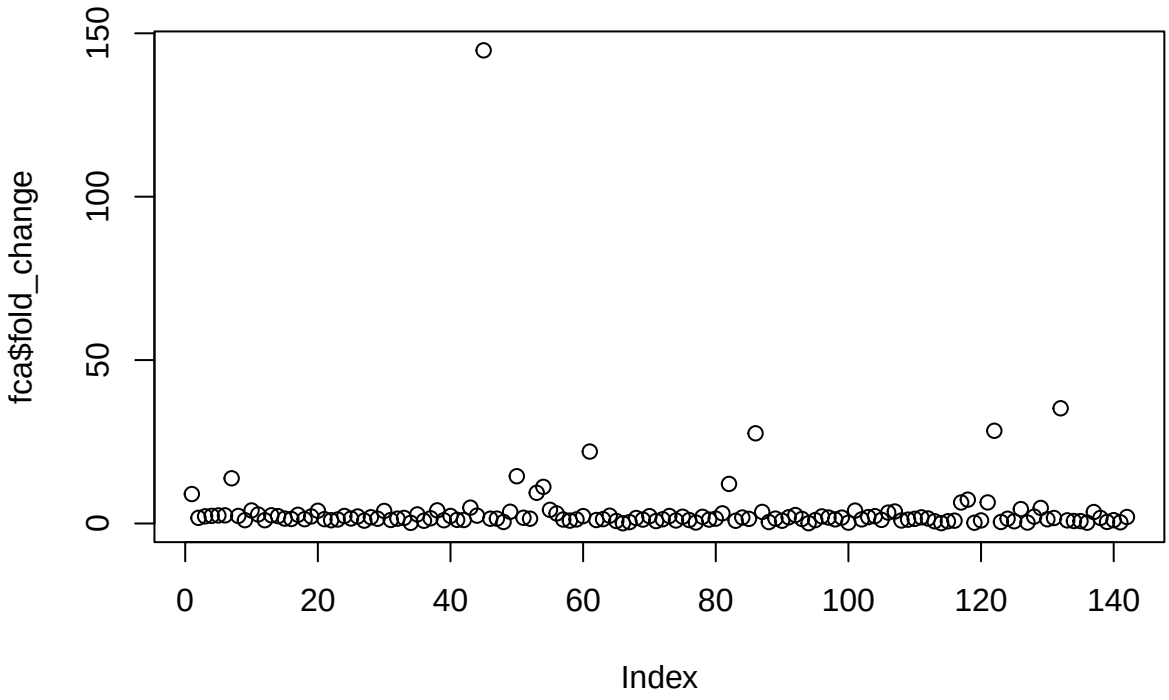
A continuació, es mostren els 20 metabòlits que més contribueixen a ambdós components.

```
fviz_contrib(data_t.pca, choice = "var", axes = 1:2, top = 20)
```



Anàlisi de fold-change (FCA)

Un altre enfocament a l'anàlisi de les dades és utilitzar el fold-change per mesurar el canvi de les mesures abans i després del transplantament. Per fer-ho s'ha calculat la mitjana dels grups After i Before i se n'ha calcular el quocient. La gràfica obtinguda mostra el valor de fold-change per a cada metabòlit.



Addicionalment, s'han extret els metabòlits amb un fold-change en intensitat major, tant positiu com negatiu.

```
fca_top <- arrange(fca, desc(fold_change))
head(fca_top, 5)
```

```
##                               fold_change
## fructose                        144.78488
## trehalose                        35.24820
## sorbitol                         28.36865
## levanbiose                       27.59182
## glycerol-3-galactoside NIST      22.00275
```

```
tail(fca_top, 5)
```

```
##                    fold_change
## ribose              0.22964798
## cholesterol         0.19744760
## phosphoric acid     0.10951327
## methanolphosphate   0.08377482
## guanine             0.07232225
```

Creació del repositori GitHub

Per tal de disponibilitzar els resultats d'aquesta prova s'ha creat un repositori ("Mico-Frau-Estella-PAC1") i per gestionar-lo i afegir els documents pertinents s'ha creat un projecte de R amb control de versions de

Git.

Discussió

D'una banda, pel que fa al primer apartat de la prova, s'ha mostrat que la generació d'objectes poden facilitar la lectura de dades d'experiments que generen grans quantitats d'informació, com són els assajos en ciència de dades òmiques. SummarizedExperiment és una bona eina per agrupar tota la informació d'un experiment, tot i que, sembla que està més desenvolupat per a l'anàlisi de dades d'expressió gènica (ExpressionSet) i dades de transcriptòmica (SummarizedExperiment). En aquest cas, per exemple, no s'ha pogut utilitzar la funció de GRanges, ja que fa referència a les regions cromosòmiques.

D'altra banda, la segona part de la prova ha permés aplicar diferents funcionalitats en R per entendre les dades que s'estan tractant. L'anàlisi de components principals ens ha permés discernir un patró d'agrupació per als diferents grups, la qual cosa pot ser indicatiu de la diferència en el metabolisme de la microbiota en les dues condicions, tal i com s'hipotetizava per a aquest treball. A més, amb l'anàlisi de fold-change, s'han identificat alguns dels metabòlits que tenen una diferència d'intensitat major entre ambdos grups. Abans de l'operació, hi ha una major presència de metabòlits lligats al metabolisme dels sucres, la levanbiosa i el glicerol-3-galactòsid, tot i que no tenen un rol clar i la fructosa, relacionada també amb el cicle de Krebs (una ruta clau). En canvi, després de l'operació, s'han detectat compostos aromàtics (com el naftalé), esterols (colesterol), uracil (cicle de la urea). Tot i que no es coneix el rol exacte d'aquests metabòlits en les lesions, sí que s'observa un canvi en el perfil metabolòmic general.

A l'article del qual s'han obtingut les dades, s'han realitzat altres estudis de qPCR i seqüenciació d'ADNr 16S, i la valoració de les dades conjuntament porta el grup de recerca a hipotetitzar un canvi en la comunitat microbiana de l'intestí, d'una comunitat annaeròbia facultativa abans de l'operació (per això s'han detectat alguns metabòlits importants per al cicle de Krebs) a una comunitat amb més microorganismes estrictament anaeròbics.

També és interessant comentar que, aquestes dades han estat utilitzades per fer un tutorial de la utilització de la plataforma MetaboAnalyst, la qual permet des d'una mateixa pàgina realitzar múltiples anàlisis exploratòries i estadístiques. Açò, per a dades que no són molt pesades, resulta d'especial interés, ja que no requereix carregar múliples paquets de R i fer diferents passos per representar gràfiques o obtindre resultats de proves analítiques. Tot i així, R continua sent una millor opció per treballar amb dades pesades, ja que té més poder computacional.

Per últim, s'ha demostrat que Git i GitHub són eines molt interessants per compartir informació amb la comunitat científica en l'àmbit de la bioinformàtica. A més a més, la possibilitat de penjar els fitxers amb els quals s'està treballant directament des de R o RStudio, facilita aquesta tasca.

Conclusió

En conclusió, amb aquesta activitat s'ha evidenciat la utilitat de la classe "SummarizedExperiment" en la gestió de dades de metabolòmica de manera estructurada. A més, s'han aplicat tècniques d'anàlisi exploratòria de dades, com la PCA i el fold-change, les quals ajuden a l'obtenció d'informació a partir de dades de mesura com les d'espectrometria de masses. L'exemple triat ha permés aplicar a un context real aquestes ferramentes i obtindre conclusions sobre el tema d'estudi, ja que s'han detectat diferències en el metabolisme de la microbiota abans i després de la intervenció, tal i com, també s'havia detectat (amb mètodes més exhaustius i rigorosos) en l'article original. També s'ha millorat la comprensió del funcionament dels repositoris GitHub.

Referències

Micó-Frau, E. (2025). Mico-Frau-Estella-PAC1 [RStudio]. GitHub. https://github.com/esmifrau/Mico-Frau-Estella-PAC1/

Hartman, A. L., Lough, D. M., Barupal, D. K., Fiehn, O., Fishbein, T., Zasloff, M., & Eisen, J. A. (2009). Human gut microbiome adopts an alternative state following small bowel transplantation. Proceedings of the National Academy of Sciences of the United States of America, 106(40), 17187–17192. https://doi.org/10.1073/pnas.0904847106