Informe

Noé

2025-04-02

Contents

1 Abstract

En aquest estudi es pretèn realitzar un estudi exploratòri de dades metabòliques de l'estudi ***"Differential Metabolites and Disturbed Metabolic Pathways Associated with chlorpromazine Poisoning"*** amb referència **ST001739** en MetabolomicsWorkbench.

Aquest estudi es basa en un analisi descriptiu i dels principals components estadístics que proporcioni una visió general del dataset. On s'ha hagut de realitzar anteriorment, una construcció de l'objecte que contingui la informació de l'estui: SummarizedExperiment.

I una posterior implementació de tècniques exploratòries, com PCA i un Agrupament jeràrquic. Ambdues tècniques, han mostrat una tendència de que les diferents mostres de ratolí s'agrupin en diferents grups. Aquest grups estan conformats per els efectes de la intoxicació per cloropromazina, fet que podría indicar una alteració en els perfils metabòlics de manera diferencial, i per tan, una alteració en les vies metabòliques de les diferents mostres analitzades, que defineixin cada cas.

2 Objectius

En aquest projecte es pretèn realitzar un anàlisi exploratòri amb l'objectiu de poder comprendre en profunditat l'estructura que tenen les dades emprades, ja sigui per entendre com aquestes dades estan estructurades, les

seves principals característiques estadístiques (estadístis), i si cal fer un procés de filtratge de dades donada la presència de valors perduts. A més, d'una visualització d'aquestes per poder veue si cal aplicar alguna transformació a les dades.

Per una altre banda, amb l'ús de diverses tècniques com una PCA o un agrupament jeràrquic, poder determinar si les mostres emprades es troben diferenciades unes de les altres, donada l'alteració dels metabòlits en el torrent sanguini per culpa de l'enverinament per cloropromizina. D'aquesta manera, gràcies als agrupaments, poder determinar si les mostres es troben ben diferenciades unes de les altres.

3 Mètodes

Per aquest estudi, s'han utilitzat dades provinents de **Metabolomics Workbench**, específicament de l'estudi **ST001739** titulat ***"Differential Metabolites and Disturbed Metabolic Pathways Associated with chlorpromazine Poisoning"***.

Els investigadors que han desenvolupat aquest estudi, han emprat i analitzat mostres biològiques de ratolins, els quals sel's hi ha subministrat clorpromazina, un atipsicòtic, en mateixes dosis. D'aquests animals, van extreure mostres sanguínees dels morts per intoxicació per clorpromazine, mostres de ratolins intoxicats però sense morir, i mostres de ratolins no intoxicats, com a control.

Per poder detectar els metabòlits, s'ha combinat la cromatografia líquida d'alt rendiment i l'espectrometria de masses d'alta resolució (UPLC-HRMS), amb la finalitat de poder identrificar els metabòlits diferencials associats amb la intoxicacó letal per cloropromazina (CPZ), i per tant, quines són les víes metabòliques que s'han trobat alterades per aquesta intoxicació.

Per realitzar l'estudii que es pugui entendre bé el que s'ha fet, s'ha separat en 3 etapes, que corresponen als diferents nivells de l'estudi. També s'ha emprat com a guia l'exemple el document: https://aspteaching.github.io/AMVCasos/#ejemplo-pca-1-bÃžsqueda-de-factores-latentes-en-datos-ecolÃ§gicos1

3.1 Construcció del SummarizedExperiment:

Per poder treballar s'ha hagut de construïr un SummarizedExperiment, una classe que permet guardar dades respecte l'estudi, provinents de tècniques com microarray o RNA-seq, en aquest cas d'espectrometria de masses.

Per poder realitzar el SummarizedExperiment, s'ha emprat els paquets ***"SummarizedExperiment"*** i ***"metabolomicsWorkbenchR"***, respectivament per generar la classe i descarregar l'experiment de la plataforma de MetabolomicsWorkbench, a partir de la funció [***do__query()***].

3.2 Descripció i resum numèric univariant:

Per poder treballar amb la matriu de dades i relitzar un anàlisi descriptiu, primer de tot s'ha hagut d'extreure la matriu del SummarizedExperiment, continguda com a assay. Aquesta matriu resultant és practicament amb la que es treballarà al llarg de l'estudi.

Per comprendre l'estructura de les dades s'ha implementat la funció [***str()***], per determinar el tipus de dades amb les que es treballen, com també el tipus de variables que analitzem. Totes elles numèriques.

També, he mirat si hi ha presència de números perduts (Na's) o números nuls. Aquests, a l'hora de realitzar posteriors processos poden ser molests per l'estudi, i entorpir el procés. Per tant, han sigut eliminades les files amb valsors Na's o nuls.

Finalment, s'ha mirat la dimensió de les dades amb les que es treballa després de la modificació, 325 (metabòlits) * 30 (mostres). I s'ha realitzat un resum estadístic, per explorar ràpidament els estadístics principals de les diferents mostres en funció dels diferents metabòlits en sang.

Un cop realitzada l'exploració, per contrastar els resultats, he construit uns histogràmes, un per cada mostra per veure quina és la distribució de les concentracions metabòliques, i un diagrama de caixes (BoxPlot), que

permet veure si escal alguna transformació de les dades.

3.3 PCA i Agrupament Jeràrquic:

Per realiztar aquests dos processos primer de tot s'han normalitzat les dades, per si fós el cas que les dades no estan en la mateixa escala, doncs així que quedin escalades, i treure les diferències sistemàtiques, com és el cas de valors molt alts o baixos de outliers.

A continuació, he formulat l'Analisi de Components Principals (PCA) amb [prcom()], on s'han estipulat els colors segons el fenotip de les mostres (Blau = Control, Vermell = Morts per enverinament, Verd = només intoxicats), així serà més fàcil catalogar les mostres en la PCA, i també s'han etiquetat cada mostra corresponent.

Pel que fa l'agrupament jeràrquic, ha servit com a contrast amb la PCA, per veure si l'agrupament que aquesta tècnica em generava, era igual a les tendències intrinseques que mostrava la PCA amb les mostres.

4 Resultats

4.1 Anàlisi descriptiu i resum numèric univariant:

Amb les dades que treballem, són les que es mostren a continuació:

```
#SummarizedExperiment
library(metabolomicsWorkbenchR)
library(SummarizedExperiment)
SE_experiment <- do_query(context = 'study',
                          input_item = 'study_id',
                          input_value = 'ST001739',
                          output_item = 'SummarizedExperiment')

#Generació de taula de dades:
matriz_metabolitos<-assay(SE_experiment)
matriz <- data.frame(matriz_metabolitos)
rownames(matriz) <- rownames(matriz_metabolitos)
```

De la taula de dade creada anomenada "***matriz***", no hi ha cap valor nul, però si que n'hi han valors perduts (Na's), en concret 30 Na's.

```
matriz_neta <- na.omit(matriz) #eliminació Na's
```

Després d'eliminar-los, estem treballant amb un data.frame (***matriz_neta***) que conté 325 files, que corresponen als diferents metabòlits, i 30 columnes, corresponents a les diferents mostres de ratolí.

Comparant la dimenssió inicial de les dades, on era una taula de 326 files x 30 columnes, s'ha reduït una fila amb l'eliminació de Na's.

Es pot observar que en les dades inicials, aquests 30 Na's corresponien a un metabòlit en concret, sense medicions. A l'eliminar aquesta fila, 30 Na's, un per columna, la taula resultant de l'estudi ha quedat amb un metabòlit menys (fila).

A més, el data.frame està composat únicament de variables de tipus numèriques, sent 30 variables en total, amb una longitud de 325 valors per variable.

```
str(matriz_neta)
```

```
## 'data.frame':    325 obs. of  30 variables:
##  $ FCS1 : num  7332284 6732751 1333260 11921369 4646279 ...
##  $ FCS2 : num  13104559 8115000 809485 15646333 1713192 ...
##  $ FCS3 : num  12599217 10072704 2752612 18209421 2827638 ...
```

```
##  $ FCS4 : num  12722891 9351066 5524092 12591201 2253902 ...
##  $ FCS5 : num  13241862 9591180 4309110 12715117 2212733 ...
##  $ FCWS1: num  9123494 10222020 731131 12898007 4160956 ...
##  $ FCWS2: num  8912003 9905821 535413 17126930 3037486 ...
##  $ FCWS3: num  12340101 9655406 1064938 17521896 2567837 ...
##  $ FCWS4: num  8964435 10164078 2347810 18003334 3502713 ...
##  $ FCWS5: num  11585131 12515860 481818 15759903 2798733 ...
##  $ FT1  : num  37325253 18526512 13350945 35663582 265087 ...
##  $ FT2  : num  43144954 11105144 1875753 18047322 128042 ...
##  $ FT3  : num  45015529 12029915 1723597 18957784 457161 ...
##  $ FT4  : num  30738872 12360240 1695841 16589103 345151 ...
##  $ FT5  : num  48193641 12710292 1930660 14881781 184972 ...
##  $ MCS1 : num  4960220 6577836 11931404 9655376 41728 ...
##  $ MCS2 : num  8713577 6560774 2012238 11204891 317153 ...
##  $ MCS3 : num  12558654 10565535 22914995 15059623 123420 ...
##  $ MCS4 : num  11642838 7877545 4268995 16734553 56341 ...
##  $ MCS5 : num  12918220 10235496 24934258 14174203 225728 ...
##  $ MCWS1: num  10069604 9349532 1397265 17065936 35526 ...
##  $ MCWS2: num  11599981 9060284 2116767 12371992 95283 ...
##  $ MCWS3: num  4736159 4619325 640011 10041332 1523313 ...
##  $ MCWS4: num  8570822 8155701 3504633 12365753 37494 ...
##  $ MCWS5: num  12356870 8561216 2093485 13051378 137944 ...
##  $ MT1  : num  19922939 12448424 961028 21241487 72759 ...
##  $ MT2  : num  25415963 9234215 740795 15113523 39532 ...
##  $ MT3  : num  21022265 7636000 1554244 15758415 56064 ...
##  $ MT4  : num  20290369 17882917 803779 21134059 31997 ...
##  $ MT5  : num  19636156 8073882 1730313 14535747 90787 ...
##  - attr(*, "na.action")= 'omit' Named int 239
##   ..- attr(*, "names")= chr "ME403893"
```

Per últim, pel que fa els estadístics de cada variable, tenen una mediana al voltant de 1.5e+07 (varien segons la mostra), on el valor mínim més petit de totes les mostres correspon al de la mostra MCWS3, i el valor màxim més elevat de totes les mostres correspon a la mostra de MCWS1, amb un valor de 1.54e+10. Aquesta gran diferència entre els valor màxims i mínims podría indicar una gran dispersió dels valors, o presència de outliers molt elevats o petits.

Gràcies als histogrames graficats, s'ha vist que la gran majoria dels valors dels metabòlits per mostra, es troben concentrats en valors relativament baixos comparat amb els màxims, establint una distribució desplaçada en el cantó esquerre. A continuació, es mostra un histograma de la mostra FCS1, ja que totes les altres mostres, tenen una representació similar.

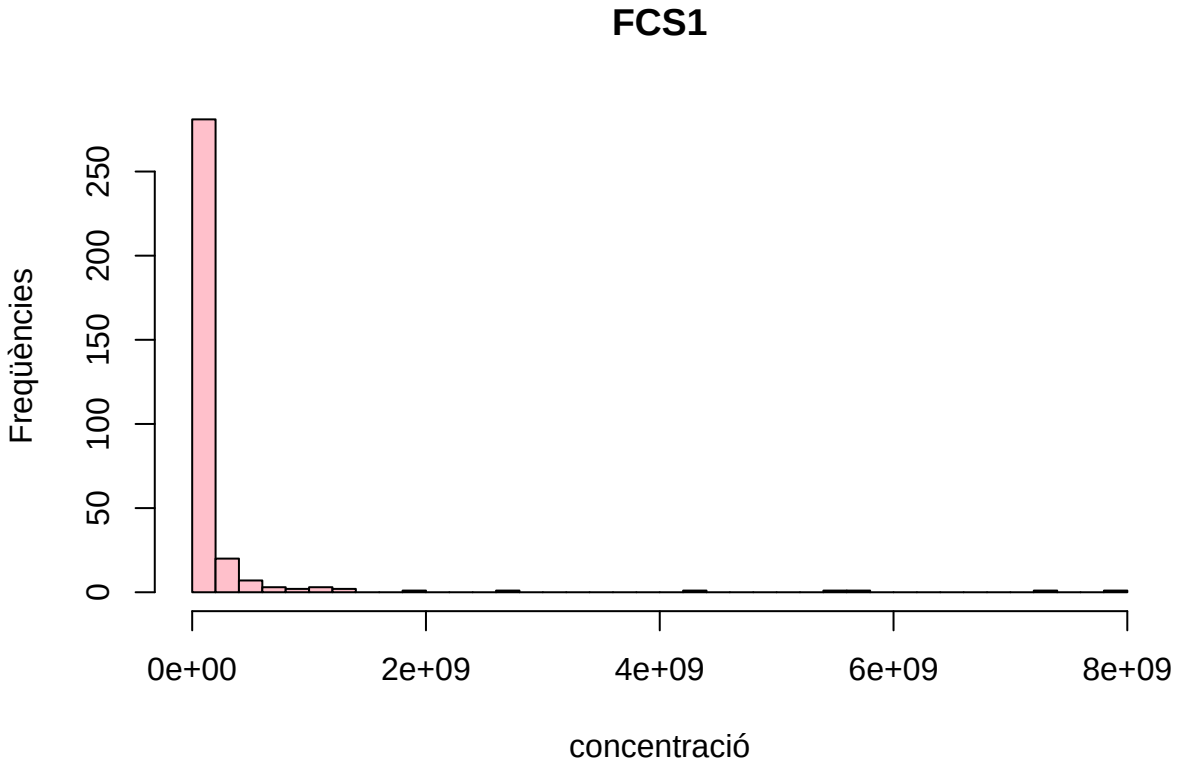

Figure 1: Histograma de la primera mostra

Per finalitzar en la graficació, s'han visualitzat amb uns boxplots totes les mostres, on es pot veure un número elevat de outliers que impedeixen visualitzar correctament les mostres. Per això, també s'ha procedit a normalitzar les dades, per evitar que aquests outliers tan elevats puguin arribar a afectar a l'estudi.

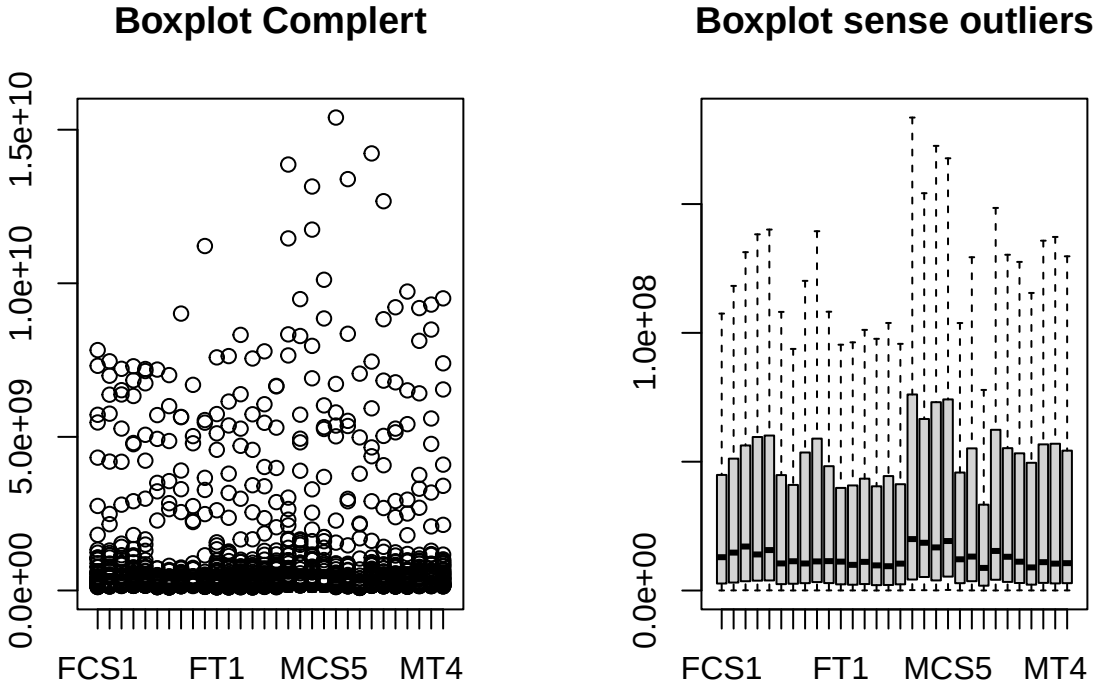

Figure 2: BoxPlot després de Normalitzar. Boxplot de l'esquerre complert, i el boxplot de la dreta sense outliers per millorar l'interpretació

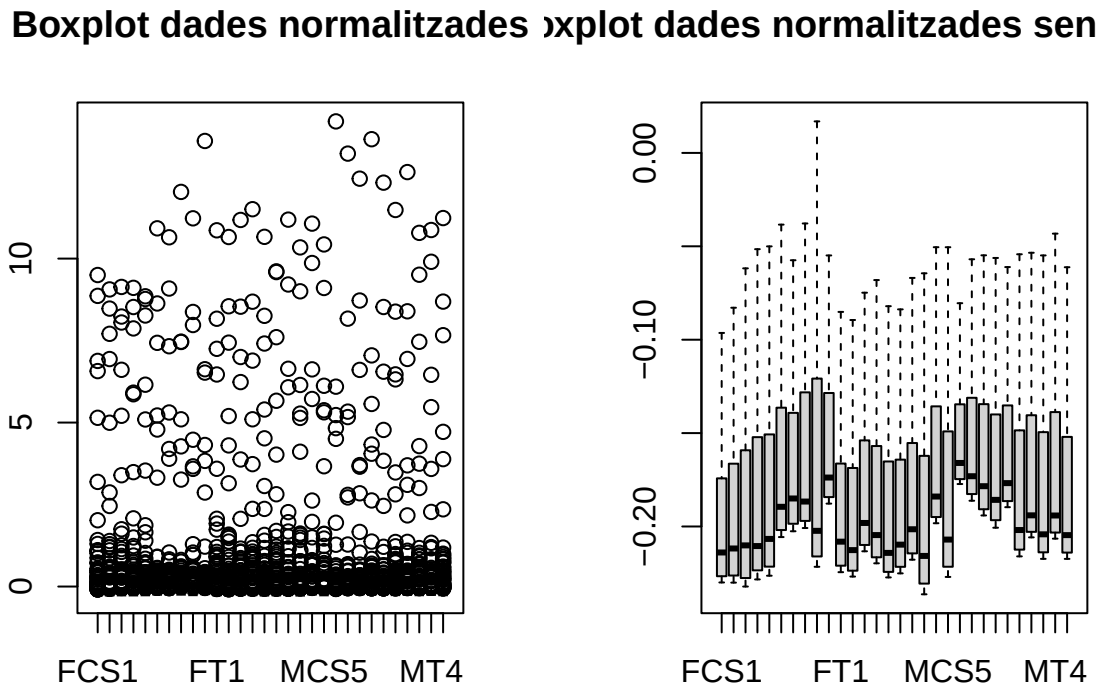

Figure 3: BoxPlot després de Normalitzar. Boxplot de la esquerre completo, i el boxplot de la dreta sense outliers per millorar l'interpretació

4.2 PCA i Agrupament Jeràrquic:

Com es pot veure en la PCA, la desviació estàndard és major en el primer component (8.57), mentre que a mida que es van passant als següents components aquesta va decaient (7.70, 6.03, 5.10...).

A continuació, es mostren les característiques dels 5 primers components principals, encara que en el codi original penjat al github, es pot veure tots els components principals: (https://github.com/Noe81018/NOE_PAC1_DADESOMIQUES/blob/main/Cosi_Exploratori.R).

```
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     8.576865 7.704339 6.037306 5.106528 4.155187
## Proportion of Variance 0.226350 0.182640 0.112150 0.080240 0.053120
## Cumulative Proportion  0.226350 0.408980 0.521130 0.601370 0.654490
```

En la graficació de la PCA s'ha vist una tendència subyacent de les mostres, on aquestes s'agrupen en funció del seu fenotip. És a dir, si aquestes corresponen a mostres control, si després d'intoxicació per antipsicòtic han mort, o si només han quedat intoxicades.

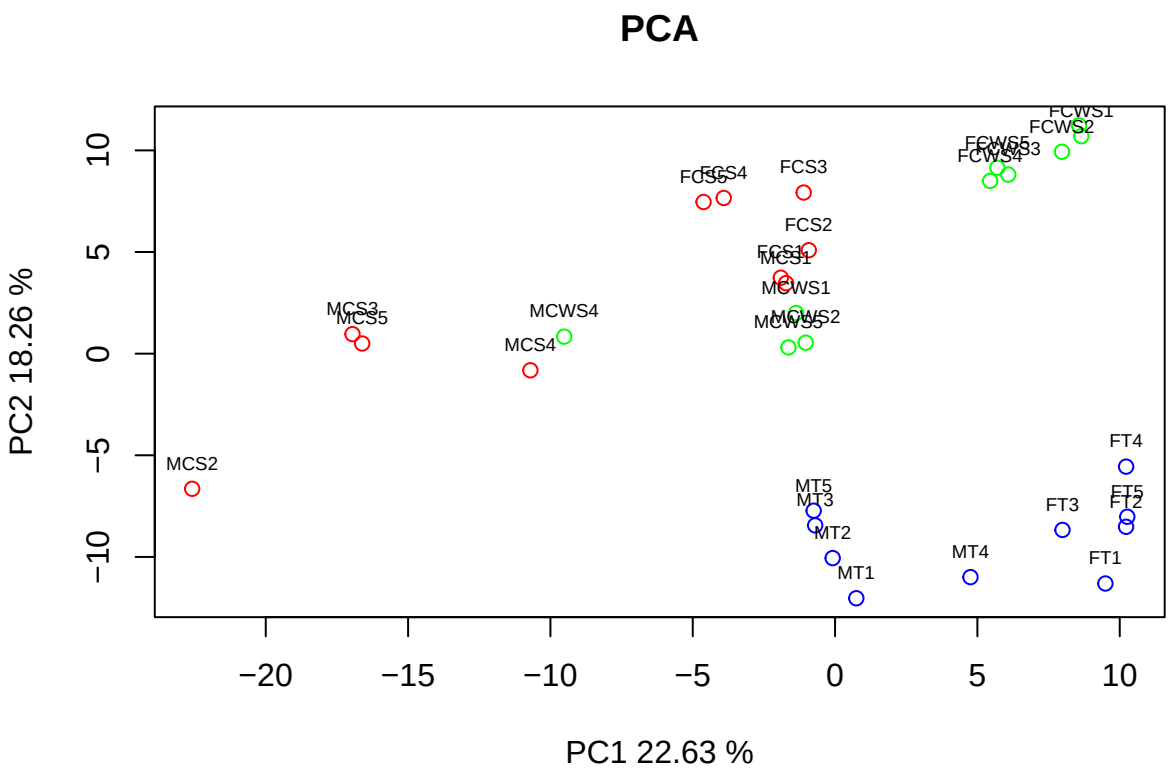

Figure 4: PCA a partir de les dades normalitzades

També es pot observar en la gràfica, que el PC1 explica un 22.63% de la varianza explicada total, mentres que el PC2 un 18.26%. Si sumem els 2 primers components dona un 40.89%, donant indícis que només amb dos components no són suficients per definir gran part de la variança explicada. Podria ser, que donada la complexitat de les variables, no es pugui representar la complexitat només amb dos compponents.

Per donar contrast als resultats obtinguts a la PCA, mitjançant un agrupament jeràrquic, s'ha vist també que les mostres tendeixen a agrupar-se en funció del fenotip d'aquestes. El que podrien indicar que, les concentracions dels metabòlits en les diferents mostres, està relacinat amb el seu fenotip.

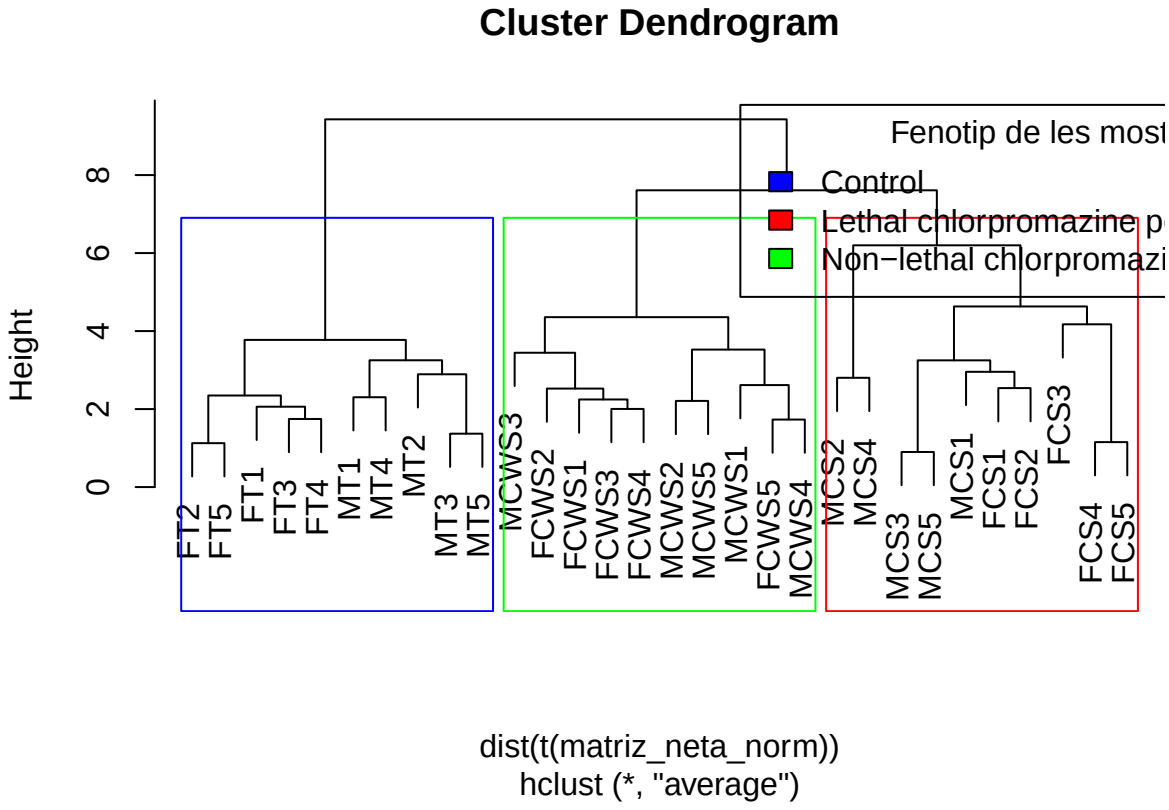

Figure 5: Agrupament jeràrquic de les mostres

5 Discussió

En aquest estudi realitzat, s'ha abordat desde una perspectiva una mica "superficial", pel que fa un estudi metabòlic, per tenir una visió general de com les dades treballades es troben estructurades, tant a nivell descriptiu d'aquestes, com pot ser la seva organització, estructura, i els principals trets estadístics, com la visualització d'aquestes per poder veure si escalen algun tipus de transformació necessaria.

A més, d'implementar tècniques multivariants, com la PCA i un agrupament jeràrquic per poder esbrinar si n'hi han patrons subyacents, gràcies als agrupament de les mostres.

Uns factors a tenir en compte en aquest estudi, és la falta d'experiència i coneixement, per tant, podría ser que calgués algun tipus de transformació necessaria de les dades, a banda de la normalització, com pot ben ser un logaritme. En principi, no s'ha considerat l'efecte Batch, ja que sembla que les mostres s'han agrupat correctament per el tipus fenotípic d'aquestes, la qual cosa no s'ha fet cap correcció.

Una limitació que també s'ha trobat, és que la PCA perd variança explicada, probablement per la complexitat de les dades, per tant, caldria més components principals per poder explicar més quantitat de variança.

En principi, els resultats obtinguts, han servit per comendre d'una forma clara i entenedora, de com les dades treballades es troben estructurades i que en un principi semblen diferenciar-se unes mostres de les altres.

6 Conclusions

Encara que s'ha vist en un principi que les dades semblaven no estar preprocessades, ja que contenien valors perduts, i no estar normalitzades. Gràcies al preprocessament, i a les tècniques de PCA i Agrupament Jeràrquic, s'ha pogut revelar possibles alteracions en els nivells dels metabòlits de les mostres dels diferents

ratolins, probablement a causa d'una alteració en les vies metabòliques per cloropromazina, abocant a l'altreació d'unes determinades vies metabòliques o unes altres, depenent si aquests han mort per intoxicació o no, diferenciant els perfils metabòlics en sang de les diferents mostres en tres agrupacions: Control, intoxicat per cloropromazina letal o intoxicació amb cloropromazina no letal.

Per poder realitzar estudis futurs, es podria fer un estudi en profunditat, de quines podrien ser aquestes rutes metabòliques que es troben altrades, i com s'han alterat, "*upregulated*" o "*downregulated*", per acabar d'entendre l'efecte de la droga en el sistema metabòlic.

7 Referències

- Per accedir al codi realitzat en l'estudi, es troba ubicat en el GitHub, en el documen **Cosi_Exploratori.R**:

https://github.com/Noe81018/NOE_PAC1_DADESOMIQUES.git

- El material emprat en aquesta PAC ha sigut el seguent:

https://aspteaching.github.io/Analisis_de_datos_omicos-Ejemplo_0-Microarrays/ExploreArrays.html

https://aspteaching.github.io/AMVCasos/#ejemplo-pca-1-bÃžsqueda-de-factores-latentes-en-datos-ecolÃşgicos1

https://mixomicsteam.github.io/mixOmics-Vignette/id_03.html

https://www.bioconductor.org/packages/release/bioc/vignettes/metabolomicsWorkbenchR/inst/doc/Introduction_to_metabolomicsWorkbenchR.html

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

https://www.bioconductor.org/packages/release/bioc/vignettes/metabolomicsWorkbenchR/inst/doc/example_using_structToolbox.html

https://www.bioconductor.org/packages/release/bioc/vignettes/Biobase/inst/doc/ExpressionSetIntroduction.pdf

https://aspteaching.github.io/AMVCasos/

8 Anex

Quina són les diferències principals entre SummarizedExperiment i ExpressionSet?

Tan el SummarizedExperiment i el ExpressionSet, són dues classes que permeten guardar informació respecte estudis. Mentres que el ExpressionSet permet guardar diferents fonts d'informació d'un microarray, el SummarizedExperiment permet guardar matrius de resultats d'experiments de microarray o de seqüenciació, no obstant, el que diferència aquest últim objecte, es que és més flexible en la informació de les seves files, permetent l'ús de GRanges com descripcions basades en DataFrames arbitraris. Aquesta característica, permet utilitzar aquesta classe en una varietat molt gran d'experiments, com RNA-seq i ChIP-Seq.

No obstant, una altre característica que es troba, és la manera i els noms que reben els slots en els diferents objectes.

En l'ExpressionSet els slots que el composen, corresponen a una matriu de dades de microarray de l'experiment (assayData), una metadata que descriu les mostres en l'experiment (phenoData), les anotacions (annotation) i la metada corresponent al chip o tecnologia emprada en l'experiment (featureData), informació relacionada sobre el protocol (protocolData), i finalment, una estructura que descriu l'experiment(experimentData).

Mentres que el SummarizedExperiment esta compossat per diferents slots, un que conté les dades experimentals (assay, pot haver-hi més d'un), on les columnes corresponen a les mostres i les files, en aquest cas als metabòlits,

una metadata que dona informació extra sobre les files (rowData, en aquest projecte sobre els metabòlits), també hi ha una metadata que guarda informació descriptiva sobre les mostres (colData), i per últim, una metadata que descriu els mètodes experimentals i publicacions de referència (metadata).