Análisis de datos ómicos - Segunda prueba de evaluación continua

Aràntzazu Alonso Carrasco

2023-05-16

Índice

Introducción y objetivos

Para este informe, analizaremos el conjunto de datos `GDS2107` con la serie `GSE3311`, titulado *Long-term ethanol consumption effect on pancreas*. Este estudio se llevó a cabo utilizando muestras de rata común (*Rattus norvegicus*). El conjunto de datos proporciona información detallada sobre los perfiles de expresión génica en el páncreas de ratas sometidas a consumo prolongado de etanol. A través de este análisis, se busca comprender los cambios moleculares y los procesos biológicos involucrados en la respuesta del páncreas al consumo crónico de etanol.

Con todo, el objetivo de este estudio fue investigar los efectos del consumo crónico de etanol en el tejido pancreático.

Métodos

En este estudio se consideran dos tratamientos (control/etanol) en una única cepa de ratas, por lo que se trata de un diseño en bloques aleatorizados puesto que podemos escoger de forma aleatoria a qué animal se le asigna cada tratamiento.

Resultado

Discusión

Referencias

Apéndice

```
#-------------------------------------------------------
# Instalación de paquetes necesarios
#--------------------------------------------------------

if (!require(BiocManager)) install.packages("BiocManager")

installifnot <- function (pkg){
  if (!require(pkg, character.only=T)){
    BiocManager::install(pkg)
  }
}
installifnot("pd.mogene.1.0.st.v1")
installifnot("mogene10sttranscriptcluster.db")
installifnot("oligo")
installifnot("limma")
installifnot("Biobase")
installifnot("arrayQualityMetrics")
installifnot("genefilter")
installifnot("annotate")
installifnot("xtable")
installifnot("gplots")
installifnot("GOstats")
installifnot("gplots")
installifnot("GEOquery")
installifnot("rae230a.db")
```

```
workingDir <-getwd()
dataDir <- file.path(workingDir, "dades")
resultsDir <- file.path(workingDir, "results")
```

```
library(Biobase)
#TARGETS
targets <-read.csv(file=file.path(dataDir,"targets.csv"), header = TRUE, sep=";")
#DEFINE SOME VARIABLES FOR PLOTS
sampleNames <- as.character(targets$ShortName)
# Creamos un objeto AnnotatedDataFrame
targets <- AnnotatedDataFrame(targets)
```

```
CELfiles <- targets$fileName
rawData <- read.celfiles(file.path(dataDir,CELfiles), phenoData=targets)
```

```
## Loading required package: pd.rae230a
## Platform design info loaded.
```

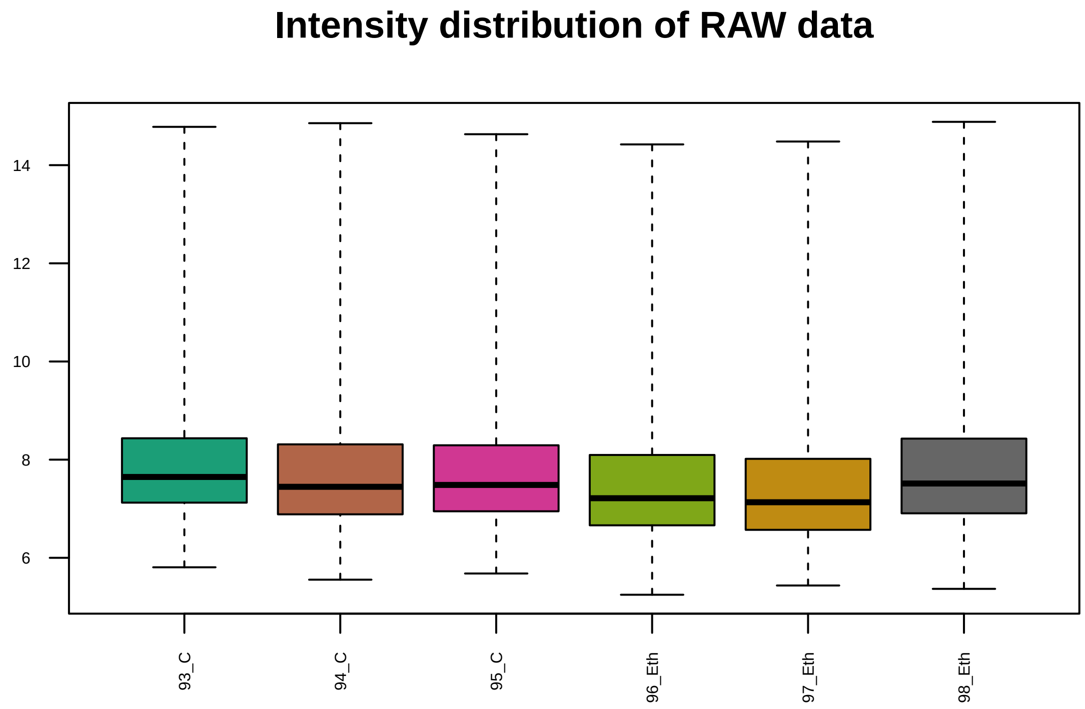
```
## Reading in : G:/Bioinformàtica/Curs 2022-2023/2n semestre/Anàlisi de dades òmiques/PAC2/Analisis_de_
## Reading in : G:/Bioinformàtica/Curs 2022-2023/2n semestre/Anàlisi de dades òmiques/PAC2/Analisis_de_
## Reading in : G:/Bioinformàtica/Curs 2022-2023/2n semestre/Anàlisi de dades òmiques/PAC2/Analisis_de_
## Reading in : G:/Bioinformàtica/Curs 2022-2023/2n semestre/Anàlisi de dades òmiques/PAC2/Analisis_de_
## Reading in : G:/Bioinformàtica/Curs 2022-2023/2n semestre/Anàlisi de dades òmiques/PAC2/Analisis_de_
## Reading in : G:/Bioinformàtica/Curs 2022-2023/2n semestre/Anàlisi de dades òmiques/PAC2/Analisis_de_
```

```
## Warning in read.celfiles(file.path(dataDir, CELfiles), phenoData = targets):
## 'channel' automatically added to varMetadata in phenoData.
```

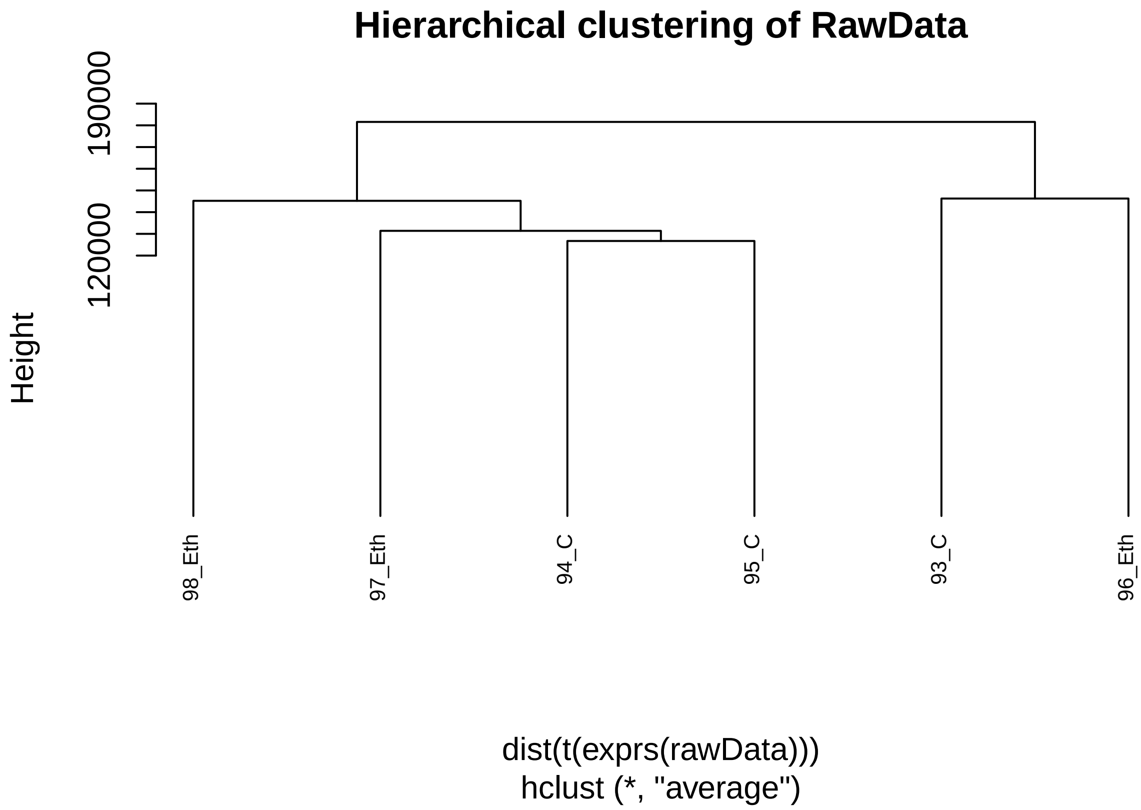
```
rawData
```

```
## ExpressionFeatureSet (storageMode: lockedEnvironment)
## assayData: 362404 features, 6 samples
##   element names: exprs
## protocolData
##   rowNames: 1 2 ... 6 (6 total)
##   varLabels: exprs dates
##   varMetadata: labelDescription channel
## phenoData
##   rowNames: 1 2 ... 6 (6 total)
##   varLabels: fileName grupos ShortName
##   varMetadata: labelDescription channel
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: pd.rae230a
```

```
#BOXPLOT
boxplot(rawData, which="all",las=2, main="Intensity distribution of RAW data",
        cex.axis=0.5, names=sampleNames)
```

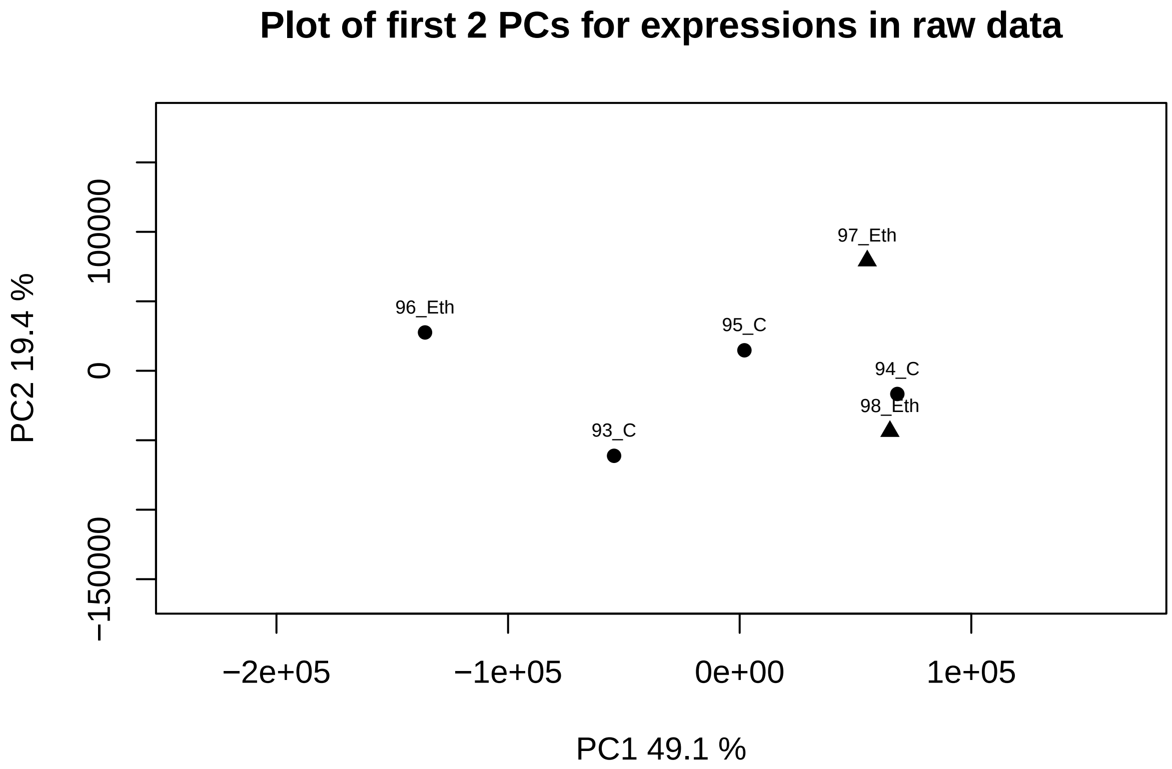


```
#HIERARQUICAL CLUSTERING
clust.euclid.average <- hclust(dist(t(exprs(rawData))),method="average")
plot(clust.euclid.average, labels=sampleNames, main="Hierarchical clustering of RawData", cex=0.7,  hang
```



```
#PRINCIPAL COMPONENT ANALYSIS
plotPCA <- function ( X, labels=NULL, colors=NULL, dataDesc="", scale=FALSE,
                      formapunts=NULL, myCex=0.8,...)
{
  pcX<-prcomp(t(X), scale=scale) # o prcomp(t(X))
  loads<- round(pcX$sdev^2/sum(pcX$sdev^2)*100,1)
  xlab<-c(paste("PC1",loads[1],"%"))
  ylab<-c(paste("PC2",loads[2],"%"))
  if (is.null(colors)) colors=1
  plot(pcX$x[,1:2],xlab=xlab,ylab=ylab, col=colors, pch=formapunts,
       xlim=c(min(pcX$x[,1])-100000, max(pcX$x[,1])+100000),
       ylim=c(min(pcX$x[,2])-100000, max(pcX$x[,2])+100000))
  text(pcX$x[,1],pcX$x[,2], labels, pos=3, cex=myCex)
  title(paste("Plot of first 2 PCs for expressions in", dataDesc, sep=" "), cex=0.8)
}

plotPCA(exprs(rawData), labels=sampleNames, dataDesc="raw data",
        formapunts=c(rep(16,4),rep(17,4)), myCex=0.6)
```



```
# Avoid re-running it each time  the script is executed.
rerun <- FALSE
if(rerun){
  arrayQualityMetrics(eset,  reporttitle="QC_RawData", force=TRUE)
}
```

```
# Normalización
eset<-rma(rawData)
```

```
## Background correcting
## Normalizing
## Calculating Expression
```

```
write.exprs(eset, file.path(resultsDir, "NormData.txt"))
eset
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 15923 features, 6 samples
##   element names: exprs
## protocolData
##   rowNames: 1 2 ... 6 (6 total)
##   varLabels: exprs dates
##   varMetadata: labelDescription channel
## phenoData
##   rowNames: 1 2 ... 6 (6 total)
##   varLabels: fileName grupos ShortName
##   varMetadata: labelDescription channel
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: pd.rae230a
```

```
# Filtrado
library(genefilter)
library(rae230a.db)
annotation(eset) <- "rae230a.db"
eset_filtered <- nsFilter(eset, var.func=IQR,
         var.cutoff=0.75, var.filter=TRUE, require.entrez = TRUE,
         filterByQuantile=TRUE)
#NUMBER OF GENES REMOVED
print(eset_filtered)
```

```
## $eset
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 2690 features, 6 samples
##   element names: exprs
## protocolData
##   rowNames: 1 2 ... 6 (6 total)
##   varLabels: exprs dates
##   varMetadata: labelDescription channel
## phenoData
##   rowNames: 1 2 ... 6 (6 total)
##   varLabels: fileName grupos ShortName
##   varMetadata: labelDescription channel
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: rae230a.db
##
## $filter.log
## $filter.log$numDupsRemoved
## [1] 2537
##
## $filter.log$numLowVar
## [1] 8070
##
## $filter.log$numRemoved.ENTREZID
## [1] 2620
##
## $filter.log$feature.exclude
## [1] 6
```

```
#NUMBER OF GENES IN
print(eset_filtered$eset)
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 2690 features, 6 samples
##   element names: exprs
## protocolData
##   rowNames: 1 2 ... 6 (6 total)
##   varLabels: exprs dates
##   varMetadata: labelDescription channel
## phenoData
##   rowNames: 1 2 ... 6 (6 total)
##   varLabels: fileName grupos ShortName
##   varMetadata: labelDescription channel
## featureData: none
## experimentData: use ’experimentData(object)’
## Annotation: rae230a.db
```

```
filteredEset <- eset_filtered$eset
filteredData <- exprs(filteredEset)
colnames(filteredData) <- pData(eset_filtered$eset)$ShortName
```

```
# Matriz de diseño
library(limma)
treat <- pData(filteredEset)$grupos
lev <- factor(treat, levels = unique(treat))
design <-model.matrix(~0+lev)
colnames(design) <- levels(lev)
rownames(design) <- sampleNames
print(design)
```

```
##        control_diet ethanol_diet
## 93_C              1            0
## 94_C              1            0
## 95_C              1            0
## 96_Eth            0            1
## 97_Eth            0            1
## 98_Eth            0            1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$lev
## [1] "contr.treatment"
```

```
#COMPARISON
cont.matrix1 <- makeContrasts(
        Ethanol.vs.control = ethanol_diet-control_diet,
        levels = design)
comparisonName <- "Efecto del etanol"
print(cont.matrix1)
```

```
##               Contrasts
## Levels         Ethanol.vs.control
##   control_diet                 -1
##   ethanol_diet                  1
```

```
#MODEL FIT
fit1 <- lmFit(filteredData, design)
fit.main1 <- contrasts.fit(fit1, cont.matrix1)
fit.main1 <- eBayes(fit.main1)
```

```
topTab <-  topTable (fit.main1, number=nrow(fit.main1), coef="Ethanol.vs.control", adjust="fdr",lfc=1,
dim(topTab)
```

```
## [1] 29  6
```

```
head(topTab)
```

```
##                 logFC   AveExpr          t      P.Value    adj.P.Val        B
## 1388271_at -2.301132 10.478841 -14.652405 1.437301e-19 3.866340e-16 33.84419
## 1387930_at -2.746280  7.733064 -11.929641 3.976415e-16 5.348278e-13 26.33449
## 1387874_at -2.150355  7.631634 -11.316954 2.694968e-15 2.416488e-12 24.50182
## 1367725_at  1.643399  8.988216   9.770816 4.165251e-13 2.801131e-10 19.64728
## 1387116_at -1.575883  6.899841  -9.500717 1.035018e-12 5.568396e-10 18.76720
## 1390249_at  1.771694  6.002867   9.226933 2.625876e-12 1.177268e-09 17.86615
```

```
# Anotar
library(AnnotationDbi)
keytypes(rae230a.db)
```

```
##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
##  [6] "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
## [11] "GENETYPE"     "GO"           "GOALL"        "IPI"          "ONTOLOGY"
## [16] "ONTOLOGYALL"  "PATH"         "PFAM"         "PMID"         "PROBEID"
## [21] "PROSITE"      "REFSEQ"       "SYMBOL"       "UNIPROT"
```

```
valid_keys <- keys(org.Rn.eg.db)
anotaciones <- AnnotationDbi::select(rae230a.db, keys = rownames(filteredData), columns = c("ENTREZID",
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:graph':
##
##     union
```

```
## The following object is masked from 'package:AnnotationDbi':
##
##     select
```

```
## The following object is masked from 'package:oligo':
##
##     summarize
## The following object is masked from 'package:Biobase':
##
##     combine

## The following objects are masked from 'package:Biostrings':
##
##     collapse, intersect, setdiff, setequal, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following object is masked from 'package:XVector':
##
##     slice

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
topTabAnotada <- topTab %>%
  mutate(PROBEID=rownames(topTab)) %>%
  left_join(anotaciones) %>%
  arrange(P.Value) %>%
  select(7,8,9, 1:6)
```

```
## Joining with `by = join_by(PROBEID)`
```

```
head(topTabAnotada)
```

```
##        PROBEID ENTREZID     SYMBOL     logFC   AveExpr          t      P.Value
## 1 1388271_at    689415       Mt2A -2.301132 10.478841 -14.652405 1.437301e-19
## 2 1387930_at    171162      Reg3a -2.746280  7.733064 -11.929641 3.976415e-16
## 3 1387874_at     24309        Dbp -2.150355  7.631634 -11.316954 2.694968e-15
## 4 1367725_at    64534       Pim3  1.643399  8.988216   9.770816 4.165251e-13
## 5 1387116_at    24908     Dnajb9 -1.575883  6.899841  -9.500717 1.035018e-12
## 6 1390249_at   315702 C8h15orf39  1.771694  6.002867   9.226933 2.625876e-12
##      adj.P.Val        B
## 1 3.866340e-16 33.84419
## 2 5.348278e-13 26.33449
## 3 2.416488e-12 24.50182
## 4 2.801131e-10 19.64728
## 5 5.568396e-10 18.76720
## 6 1.177268e-09 17.86615
```