Manzano-Rubio-Robert-PEC1

Robert Manzano Rubio

2025-03-31

Resum

El present estudi té com a objectiu analitzar dades metabolòmiques d'esquirols vermells americans mitjançant el paquet `SummarizedExperiment` de R. Es busca explorar la relació entre els metabolits i l'edat dels esquirols, amb l'objectiu de descobrir possibles biomarcadors per datar altres individus. S'han obtingut dades de metabolits en plasma de 33 individus, 15 joves, 17 adults i un desconegut. Després de realitzar una imputació de valors faltants mitjançant el mètode *knn* i una normalització amb log Pareto, es fa una Anàlisi de Components Principals (PCA) i un dendrograma. Els resultats no mostren una clara separació dels individus per edat o sexe. Això suggereix que altres factors podrien estar influint més que l'edat en la concentració dels metabolits observada.

Objectius

L'objectiu general d'aquest estudi és realitzar una anàlisi exploratòria d'unes dades de metabolòmica d'esquirols vermells americans (*Tamiasciurus hudsonicus*), centrant-se en l'ús del paquet de R `SummarizedExperiment`. Concretament, els objectius són:

1. Obtenir les dades metabolòmiques de l'experiment seleccionat i crear una classe `SummarizedExperiment` amb elles.
2. Realitzar una exploració bàsica de les dades per detectar possibles problemes, valors faltants, etc.
3. Realitzar una Anàlisi de Components Principals per identificar possibles agrupacions de les dades i variables especialment importants en descriure la variabilitat observada a les dades.
4. Realitzar un anàlisi d'agrupacions geràrquic (dendrograma) per aprofundir en la identificació de possibles grups a les dades.

La classe `SummarizedExperiment` és similar a la classe `ExpressionSet` en el sentit que són classes dissenyades per treballar amb dades òmiques de manera eficient. No obstant, hi ha dues diferències principals (entre d'altres):

1. `SummarizedExperiment` és més flexible, permetent treballar amb múltiples assays del mateix experiment, múltiples condicions experimentals o diferents tipus de dades, amb un sol objecte de R.
2. `SummarizedExperiment` treballa amb dos objectes principals de nom diferent als de `ExpressionSet`: `colData` (mostres) i `rowData` (característiques de les observacions).

Mètodes

Les dades analitzades s'han obtingut de l'estudi ST000724, titulat *Red squirrels age related changes*, de la base de dades del Metabolomics Workbench, disponible aquí. L'estudi en qüestió busca trobar canvis en la metabolòmica associats a l'edat de l'individu. L'interés d'aquest estudi és que, en cas de detectar certs metabolits associats a l'edat, aquests poden servir per després utilitzar-los com a biomarcadors i datar individus d'edat desconeguda. Concretament, les dades corresponen a la concentració d'un llistat concret de metabolits en plasma de sang de diferents individus joves i adults.

Les dades s'han obtingut a partir del paquet de R `metabolomicsWorkbenchR`, que permet extreure directament les dades del Metabolomics Workbench en un objecte de classe `SummarizedExperiment` a través de la funció `do_query`:

```
SE<-do_query(
    context = 'study',
    input_item = 'study_id',
    input_value = 'ST000724',
    output_item = 'SummarizedExperiment' # or 'DatasetExperiment'
    )
```

L'estudi referenciat conté dades de dues anàlisis, segons el tipus d'ionització utilitzat per quantificar els metabolits en l'espectrometria de masses, AN001134 (negativa) i AN001135 (positiva), de 240 features (metabolits) i 33 mostres (individus) i 416 features i 33 mostres, respectivament. Per tal de reduir la mida de l'informe, aquest es centrarà en l'anàlisi AN001134.

```
summary(SE)
```

```
##          Length Class                  Mode
## AN001134 240    SummarizedExperiment S4
## AN001135 416    SummarizedExperiment S4
```

```
SE1<-SE$AN001134
metadata(SE1)$analysis_summary
```

```
## [1] "Reversed phase NEGATIVE ION MODE"
```

```
dim(SE1)
```

```
## [1] 240  33
```

```
SE2<-SE$AN001135
metadata(SE2)$analysis_summary
```

```
## [1] "Reversed phase POSITIVE ION MODE"
```

```
dim(SE2)
```

```
## [1] 416  33
```

Les anàlisis s'han fet, primer, mitjançant cromatografia de fase revertida amb l'instrument *Agilent* per la separació de metabolits i espectrometria de masses per la seva quantificació.

Els resultats contenen, primer, una primera part de descripció estadística i gràfica de les dades, calculant la mitjana de les mostres de l'estudi i la seva desviació estàndard. Després, es realitza una exploració de les dades per la detecció de valors faltants, els quals s'imputen mitjançant el mètode de veins més propers (*knn*) ja que aquest mètode té els següents avantages per treballar amb dades de metabòlits:

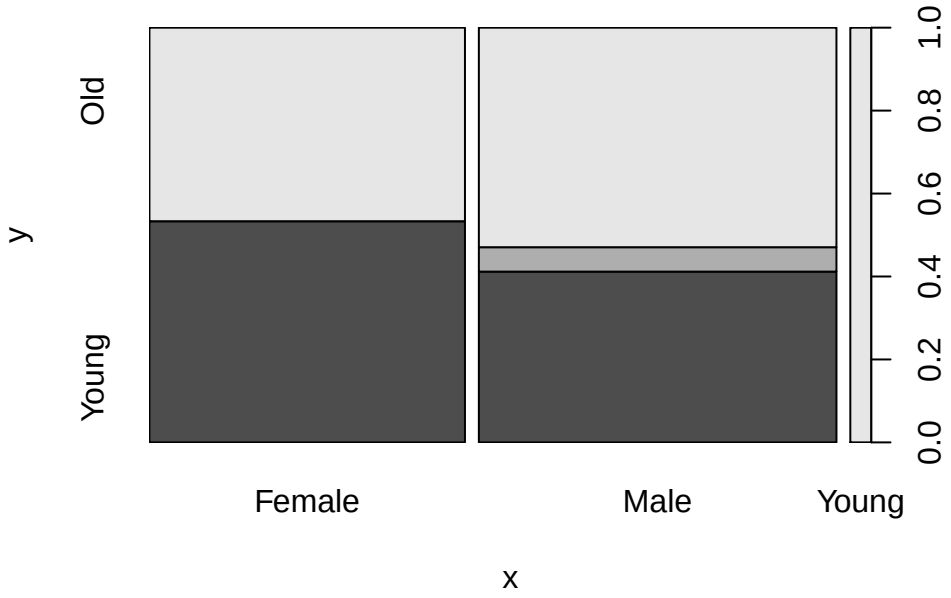
1. És efectiu treballant amb dades de moltes dimensions (variables), preservant les possibles relacions que poden haver entre els diferents metabolits.
2. És més efectiu en mantenir la variabilitat de les dades que altes mètodes, com la mitjana o la mediana.
3. No assumeix linearitat.

Després, s'ha realitzat una normalització de les dades amb el mètode log Pareto, el qual és especialment recomanable per una posterior anàlisi de dades multivariant mitjançant PCA. Finalment, es realitza una PCA amb les dades normalitzades i per mètode de descomposició de valors singulars i, també, es realitza un dendograma per distàncies euclídies, per investigar si, tal i com vol esbrinar l'estudi d'origen, l'edat està associada a una concentració de metabolits en particular.

Resultats

Les mostres de sang s'han obtingut de 33 individus amb una representació equitativa d'edat i sexe: 15 individus són joves, 17 adults i 1 indeterminat, i 15 individus són femelles, 17 són mascles i 1 indeterminat.

```
plot(SE1$Sex_of_individual,SE1$Age_Category)
```



Les dades provenen de 240 metabolits mesurats als 33 individus.

```r
dim(SE1)
```

```
## [1] 240  33
```

Es fa una exploració de les dades per detectar valors faltants (NAs) i es detecten 3755 valors faltants. S'utilitza la funció `PomaImpute` i el mètode de veins més propes (*knn*) per intentar imputar-los, sempre i quan corresponguin a features amb un 20% màxim de valors faltants. En cas que sigui superior, el feature directament s'elimina, resultant en 89 features restants.

```r
sum(is.na(assay(SE1)))
```

```
## [1] 3755
```

```r
feature_names<-rownames(assay(SE1))

imputed<-SE1 %>%
  PomaImpute(method = "knn", zeros_as_na = TRUE, remove_na = TRUE, cutoff = 20)
```

```
## 151 features removed.
```

```r
sum(is.na(assay(imputed)))
```

```
## [1] 0
```

Es realitza una exploració de la variabilitat de les dades, la qual és molt gran i pot generar problemes en anàlisis posteriors. Per exemple, les mitjanes de concentració entre metabolits varien en l'ordre de 1000 vegades respecte la mitjana més petita i la desviació estàndar en l'ordre de 2000 vegades respecte la desviació més petita.

```r
means<-c()
for (i in 1:nrow(imputed)) {
  means<- c(means,mean(assay(imputed[i])))
}

sd<-c()
for (i in 1:nrow(imputed)) {
  sd<- c(sd,sd(assay(imputed[i])))
}

range(means)
```

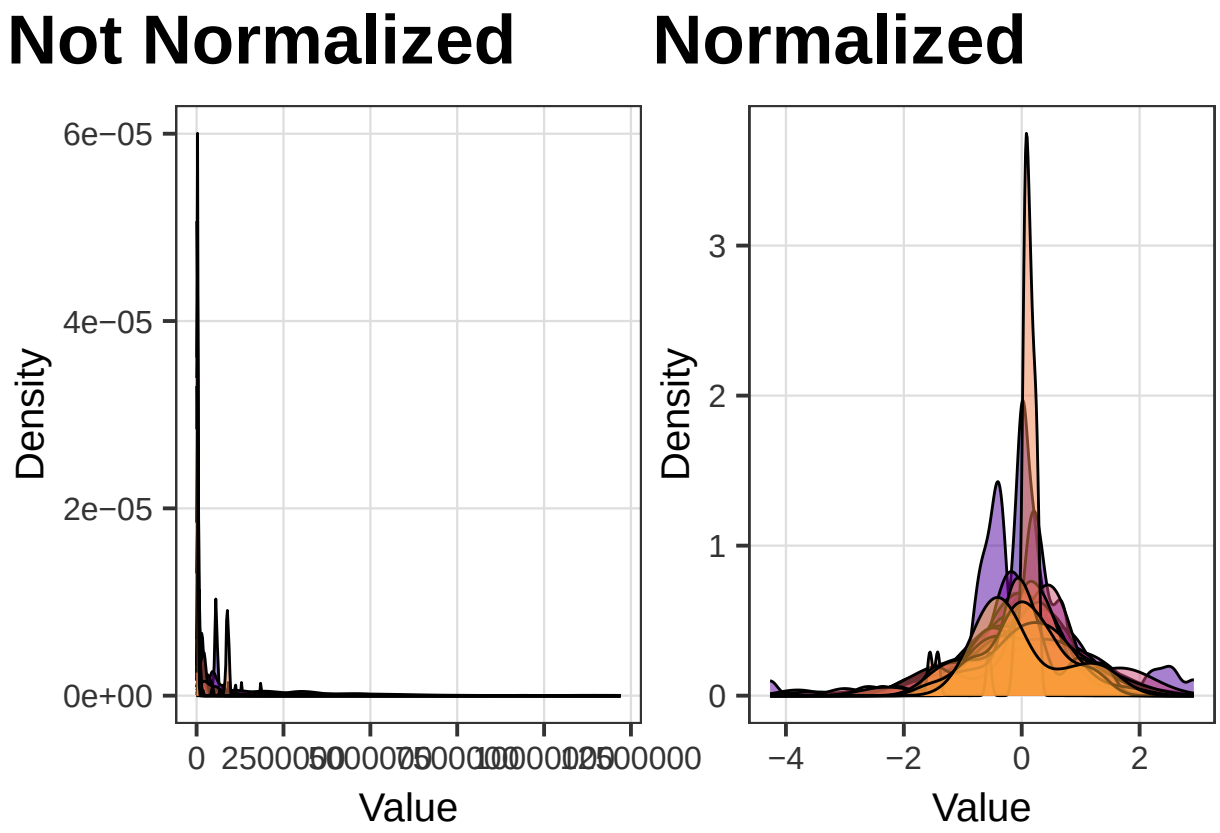
```
## [1]    6617.636 7098645.939
```

```r
range(sd)
```

```
## [1]    2421.259 5022340.797
```

Per solucionar el problema, es realitza una normalització de les dades mitjançant el mètode log pareto. L'efecte de la normalització s'il · lustra mitjançant els següents gràfics de densitat comparatius. Per facilitar la visualització, només es mostren gràfics pels primers 30 metabolits.

```
normalized <- imputed %>%
  PomaNorm(method = "log_pareto")

i1<-PomaDensity(imputed[1:30], x = "features",theme_params = list(legend_title = FALSE,
legend_position = "none")) + ggplot2::ggtitle("Not Normalized")

n1<-PomaDensity(normalized[1:30], x = "features", theme_params = list(legend_title = FALSE,
legend_position = "none")) + ggplot2::ggtitle("Normalized")
i1|n1
```

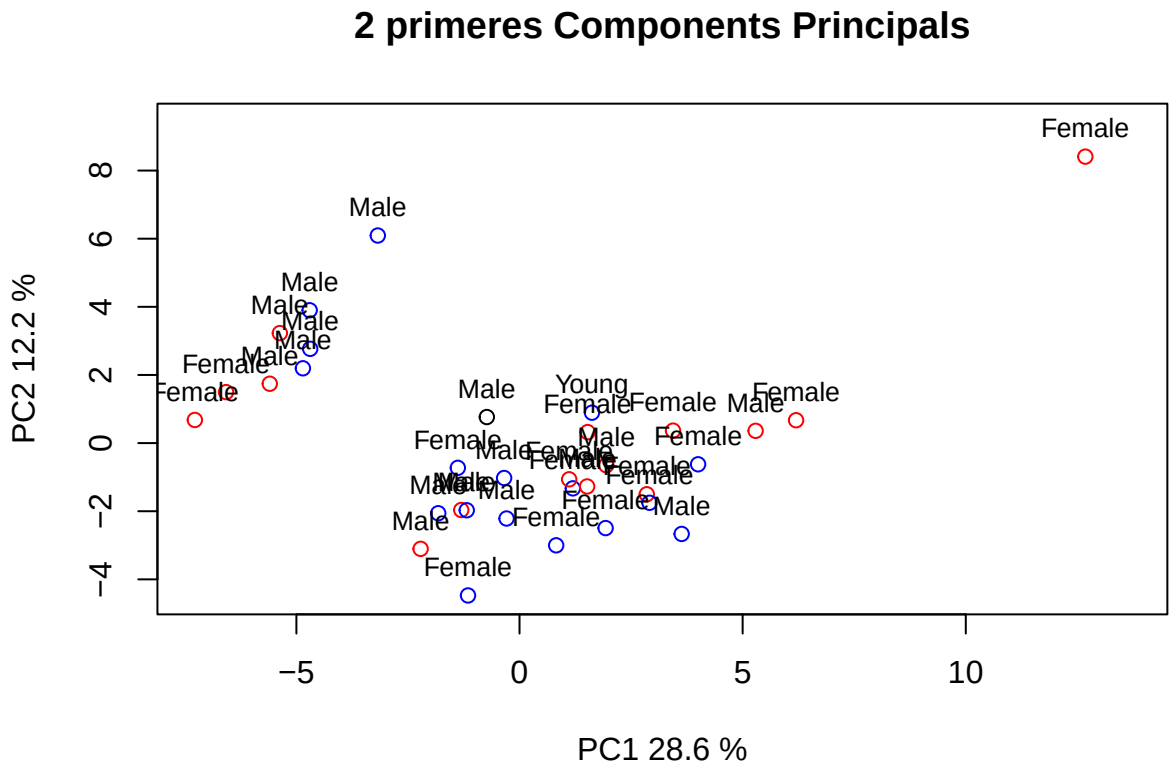


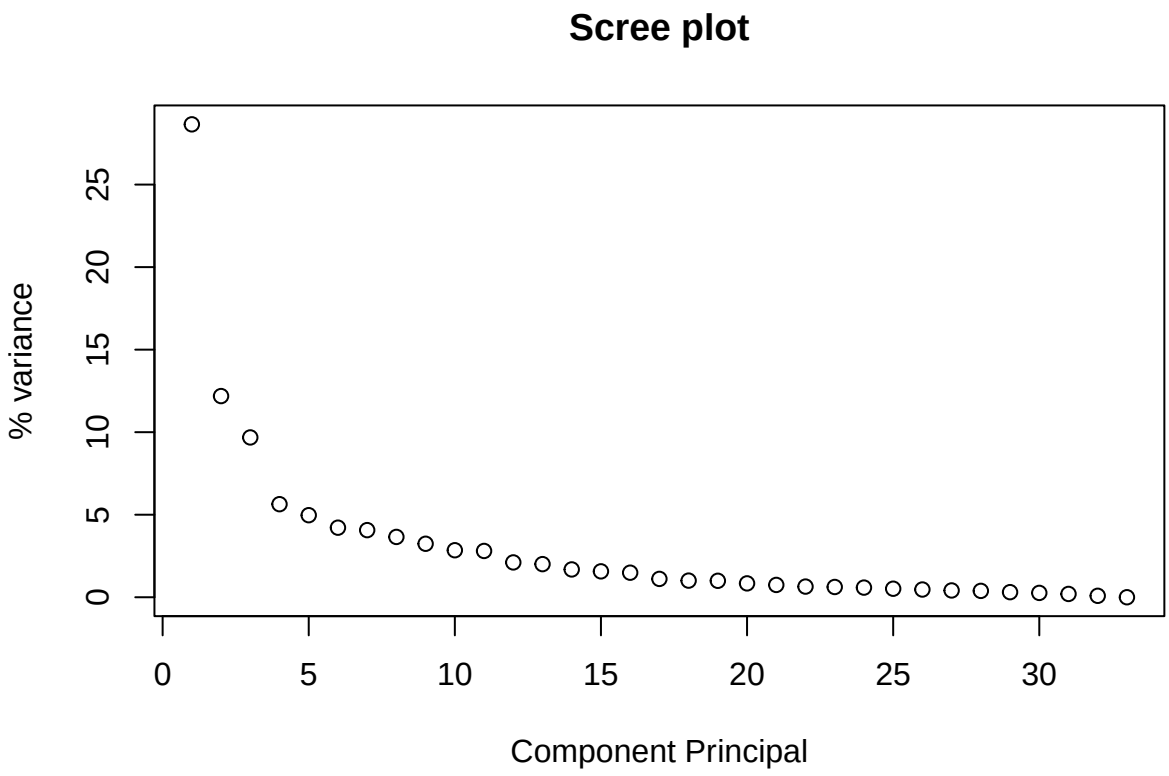
Amb les dades normalitzades, primer es genera una categorització de colors en funció de l'edat, útil en la interpretació posterior de la PCA. També, s'extreuen els sexes en un nou vector per la mateixa raó.

```
age<-as.vector(colData(SE1)$Age_Category)
sex<-as.vector(colData(SE1)$Sex_of_individual)

age[age=="Young "]<-"red"
age[age=="Old "]<-"blue"
age[age=="Unknown "]<-"black"
```

Es realitza la PCA, on s'observa que el primer component principal PC1 representa gairebé el 30% de la variabilitat de les dades, mentres que la resta de components principals representa, cadascú, valors per sota del 15% d'aquesta. No obstant, no s'observa que els dos components principals PC1 i PC2 ajudin a separar les dades de manera lògica d'acord amb el sexe o l'edat. En canvi, s'observen dos grups de mostres principals, més una mostra aïllada.

```
data<-data.frame(assay(normalized))

pcX<-prcomp(t(data), scale=FALSE)
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$x[,1:2],xlab=xlab,ylab=ylab, col=age,
     xlim=c(min(pcX$x[,1]), max(pcX$x[,1])+1),
     ylim=c(min(pcX$x[,2]), max(pcX$x[,2])+1))
text(pcX$x[,1],pcX$x[,2], sex, pos=3, cex=0.8)
title(paste("2 primeres Components Principals", sep=" "), cex=0.8)
plot(1:length(pcX$sdev),100*pcX$sdev^2/sum(pcX$sdev^2),ylab="% variance",
     xlab="Component Principal",main="Scree plot")
```





En quant a la contribució de cada metabolit al Component Principal 1, el que més variabilitat explica, aquest ve representat positivament pels metabolits V64, V126 i V114 i negativament pels metabolits V239, V98 i V240. Aquest corresponen, respectivament, a àcid cítric (M-H)-, àcid isocírtric (M-H)- i àcid hippúric (M-H)- i a zeatina [ISTD] (M+CL)-, glucosa (M+CL)- i a zeatina [ISTD] (M-H)-.

```
rotation1<-data.frame(pcX$rotation[,1])
rownames(rotation1)[order(rotation1$pcX.rotation...1., decreasing = TRUE)[1:3]]
```

```
## [1] "V239" "V98"  "V240"
```

```
rownames(rotation1)[order(rotation1$pcX.rotation...1., decreasing = FALSE)[1:3]]
```

```
## [1] "V64"  "V126" "V114"
```

```
mtb_PC1_pos<-c()
for (i in c(64,126,114)) {
  mtb_PC1_pos<-c(mtb_PC1_pos,
                 rowData(SE1)$metabolite_name[rowData(SE1)$metabolite_id==rownames(assay(SE1))[i]])}

mtb_PC1_neg<-c()
for (i in c(239,98,240)) {
  mtb_PC1_neg<-c(mtb_PC1_neg,
                 rowData(SE1)$metabolite_name[rowData(SE1)$metabolite_id==rownames(assay(SE1))[i]])}

mtb_PC1_pos
```

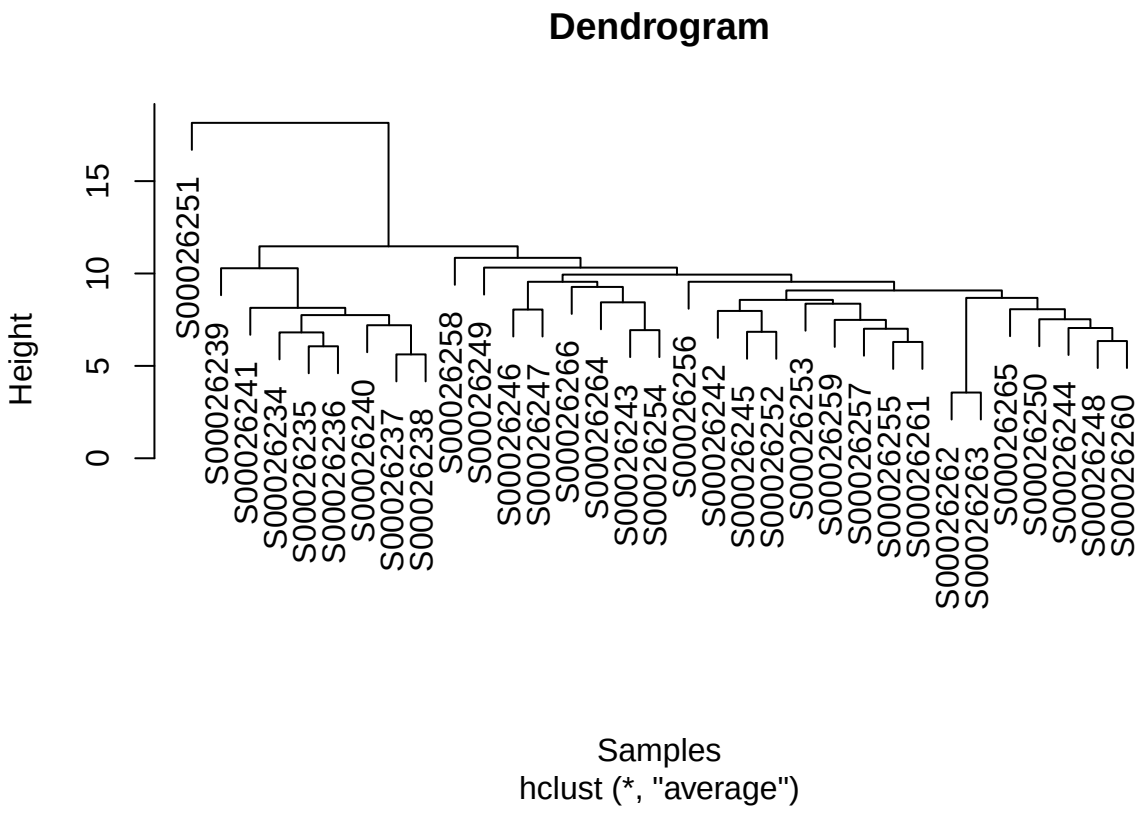
```
## [1] "CITRIC ACID (M-H)-"    "ISOCITRIC ACID (M-H)-" "HIPPURIC ACID (M-H)-"
```

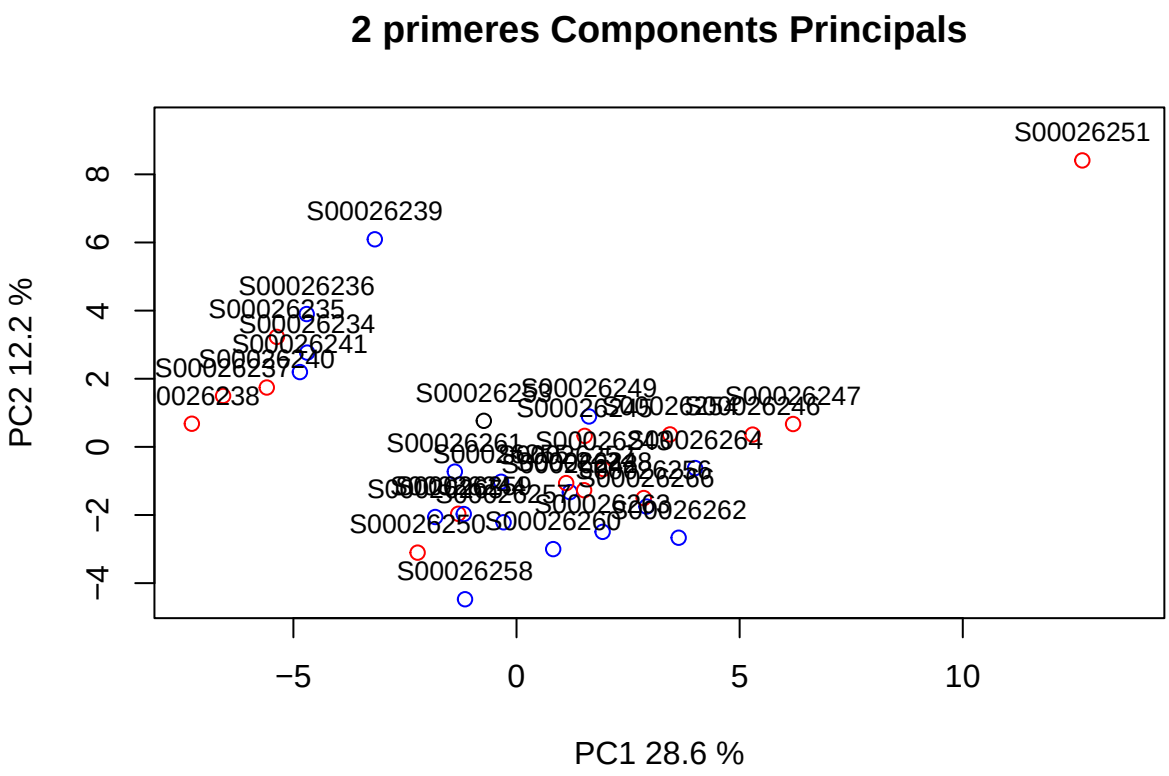
```
mtb_PC1_neg
```

```
## [1] "ZEATIN [ISTD] (M+CL)-" "GLUCOSE (M+CL)-"       "ZEATIN [ISTD] (M-H)-"
```

Per últim, es genera un dendograma per identificar possibles grups a les mostres. Aquest dendrograma coincideix amb el PCA i indica que existeixen tres grups principals de mostres: un corresponent a la mostra S0026251 (la més aïllada de la resta en el PCA) i altres dos grups amb la resta de mostres.

```
distmeth <- c("euclidian")
Distan <- dist(t(data), method=distmeth)
treemeth <- c("average")
hc <- hclust(Distan, method=treemeth)
plot(hc, main="Dendrogram", xlab="Samples", ylab="Height")

plot(pcX$x[,1:2],xlab=xlab,ylab=ylab, col=age,
     xlim=c(min(pcX$x[,1]), max(pcX$x[,1])+1),
     ylim=c(min(pcX$x[,2]), max(pcX$x[,2])+1))
text(pcX$x[,1],pcX$x[,2], colnames(data), pos=3, cex=0.8)
title(paste("2 primeres Components Principals", sep=" "), cex=0.8)
```





Discussió

L'anàlisi ha revelat una limitació important de les dades d'origen, amb molts valors faltants, que segurament condicionen la possibilitat d'obtenir conclusions respecte a la possible associació entre concentració de metabolits i l'edat o el sexe de l'esquirol vermell americà. L'anàlisi de PCA revela que un 28,6% ve explicada per la PC1, la qual està representada positivament amb àcid cítric, àcid isocítric i àcid hippúric i negativament per la zeatina i la glucosa. Si bé la concentració d'aquests metabolits, relacionats amb la producció d'energia, el sistema renal i hepàtic, o la senescència, poden variar amb l'edat, per exemple amb una possible disminució d'àcid cítric o isocítric a mesura que l'organisme envelleix, els resultats de l'anàlisi de PCA no revelen aquesta associació. El dendograma mostra una agrupació que tampoc correspon amb l'edat o el sexe dels 33 individus analitzats. Un possible motiu sigui l'existència d'altres factors que determinen, amb més força que els factors d'interès, la concentració dels metabolits observada. Per exemple, podria haver un efecte del batch o de la fitness de cada individu. No obstant, les metadades dels individus mostrejats són escassos i per tant no permeten aprofundir més en aquest aspecte.

Conclusions

L'anàlisi realitzat correspon a una exploració bàsica de les dades de l'estudi sobre relació entre metabolits i l'edat de l'esquirol vermell americà, corresponent a dues anàlisis de 33 individus, amb representació equitativa d'edat i sexe. Les dades representen molts valors faltants, els quals dificulten l'anàlisi posterior. També mostren força variabilitat i és per això que es normalitzen. El PCA i el dendograma realitzats no revelen agrupacions associades al sexe o l'edat, si bé mostren dos grups més una mostra aïllada. Caldria repetir el

mateix estudi per les dades obtingudes de l'espectrometria de ionització positiva per veure si els metabolits quantificats amb aquest mètode sí revelen patrons associats a l'edat.

Referències

link a repositori GitHub