PAC 1. Les Òmiques

Cheyenne Romero Freijo

Novembre 2024

Abstract

Aquest estudi analitza la diferenciació entre dos tipus de tumors, els grups MSS i PD, basant-se en dades de fosfoproteòmica. Mitjançant espectrometria de masses (MS), s'han quantificat aproximadament 1400 fosfopèptids. Els valors recollits inclouen mostres replicades dels dos grups tumorals.

L'anàlisi inicial mitjançant boxplots indica una coherència en les mitjanes entre repeticions, tot i la presència de valors extrems que podrien generar soroll. Els tests t de Student emprats per comparar repeticions i per investigar possibles diferències entre grups, han revelat que només el 19,6% de les mostres presenten diferències significatives entre MSS i PD. Aquesta proporció indica una baixa diferenciació entre els tumors basada en les abundàncies de fosfopèptids. També s'ha realitzat un PCA dels dos grups tumorals el que ha confirmat la tendència d'agregació dels grups, encara que aquesta es veu afectada per la presència d'outliers.

Les conclusions apunten que, tot i l'evidència de segregació parcial entre grups, l'eliminació de valors extrems podria millorar la precisió de la diferenciació entre els dos tumors.

Índex

1 Objectius de l'estudi

Dins de les dades de l'experiment es registren dos grups de tumors; el primer denominat com grup MSS i el segon com a grup PD. L'objectiu principal és veure si hi ha diferències entre els dos grups de mostres que siguin significatives entre els dos tumors. Per identificar aquestes disparitats es registren els fosfopèptids característics trobats a cada grup.

2 Materials i mètodes

El conjunt de dades d'estudi prové d'un expriment de fosfoproteòmiques. Les dades contene abundàncies normalitzades registrades a partir de senyals MS (espectrometria de masses) provinents d'uns 1400 fosfopèptids. Tots els registres s'han guardat en un excel (TIO2+PTYR-human-MSS+MSIvsPD.XLSX) on es poden consultar diferents característiques:

- SequenceModification: Conté els valors d'abundància per a cada fosfopèptid.
- Accession: Codis d'accés únic de UniProt.
- Description: Descripció de la proteïna associada al fosfopètid.
- Score: Un valor numèric.
- M1_1_MSS: Valor extret de la mostra M1 del grup MSS.
- M1_2_MSS: Segona extracció del valor de la mostra M1 del grup MSS.
- M5_1_MSS: Valor extret de la mostra M5 del grup MSS.
- M5_2_MSS: Segona extracció del valor de la mostra M5 del grup MSS.
- T49_1_MSS: Valor extret de la mostra T49 del grup MSS.
- T49_2_MSS: Segona extracció del valor de la mostra T49 del grup MSS.
- M42_1_PD: Valor extret de la mostra M42 del grup PD.
- M42_2_PD: Segona extracció del valor de la mostra M42 del grup PD.
- M43_1_PD: Valor extret de la mostra M43 del grup PD.
- M43_2_PD: Segona extracció del valor de la mostra M43 del grup PD.
- M64_1_PD: Valor extret de la mostra M64 del grup PD.
- M64_2_PD: Segon valor extret de la mostra M64 del grup PD.
- CLASS: Conté els valors H i C.
- PHOSPHO: S'especifica quin tipus de fosforilaió s'ha fet servir. Y en cas de fosforilació de la tirosina i S/T en cas de fosforilació de la terina (S) o de la treonina (T).

De totes les variables del conjunt de dades, només es faran servir SequenceModfication i les variables acabades en MSS (grup de variables d'un dels tumors) i les acabades PD (grup de variable de l'altre tumor).

2.1 Preparació de les dades

El primer pas per a l'anàlisi de les dades es la preparació de les mateixes. En aquest cas es treballarà amb un objecte de tipus SummarizedExperiment. Per la creació d'aquest objecte es necessita la llibreria *SummarizedExperiment* del paquet *BiocManager* (Bioconductor) de R. A més, com el programari fet servir per a l'anàlisi es el R, es necesitará la llibreria *readxl* per llegir el fitxer de tipus xlsx que conté les dades de l'experiment.
Per crear l'objecte d'emmagatzematge, SummarizedExperiment, a partir d'ara el referenciarem com a SE, s'han d'extreure diferents blocs de dades del fitxer original.

- assay: Conté les dades principals de l'experiment, poden ser conteigs, volums, nivells d'abundància (com en el nostre cas), etc. Pot ser diferents objectes com ara una matriu o un data frame. Nosaltres recopilem els valors de les columnes d'abundància d'ambdos grups de tumors, M1_1_MSS, M1_2_MSS, M5_1_MSS, M5_2_MSS, T49_1_MSS, T49_2_MSS, M42_1_PD, M42_2_MPD, M43_1_PD, M43_2_MPD, M64_1_PD, M64_2_MPD.

- rowData: Conté la metadata de les files; dades extra que aporten informació sobre les files. En el nostre cas, aquest data frame estarà compost per la variable *SequenceModifications*.

- colData: Conté la metadata de les columnes; dades que aporten informació extra sobre cada grup de mostres. En el nostre cas, aquest data frame està compost per una columna de *sample_id* on es troben els noms de les columnes de l'assay; una columna de *group* on s'indica a quin grup, MSS o PD, pertany cada columna; i la columna *repeticions* on s'indica a quina repetició, 1 o 2, pertany cada columna.

Els tres data frames creats es passen per la funció *SummarizedExperiment* que crearà un objecte SE.

Tant la lectura del fitxer de dades com la creació de l'objecte d'emmagatzematge es poden consultar a l'Annex.

3 Resultats

Un cop creat l'objecte SE ja es pot començar l'anàlisi de les dades de l'experiment. Primerament, es mira la quantitat de dades que conté el conjunt de dades, on es troba que hi ha 1438 files i 12 columnes. Les columnes són les que s'han seleccionat durant la creació del data frame del component *assay* i les files són els nivells d'abundància obtinguts en cada mesurament.

Seguidament, es comprova la distribució de les diferents columnes. Es vol comprovar si els màxims i els mínims dels valors de cada columna coincideixen entre repeticions, és a dir, que coincideixin, per exmeple, màxims i mínims entre M1_1_MSS i M1_2_MSS.

```
   M1_1_MSS          M1_2_MSS          M5_1_MSS          M5_2_MSS
Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
1st Qu.:    5653   1st Qu.:    5497   1st Qu.:    2573   1st Qu.:    3273
Median :   30682   Median :   26980   Median :   20801   Median :   26241
Mean   :  229841   Mean   :  253151   Mean   :  232967   Mean   :  261067
3rd Qu.:  117373   3rd Qu.:  113004   3rd Qu.:  113958   3rd Qu.:  130132
Max.   :16719906   Max.   :43928481   Max.   :15135169   Max.   :19631820
  T49_1_MSS         T49_2_MSS         M42_1_PD          M42_2_PD
Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
1st Qu.:    9306   1st Qu.:    8611   1st Qu.:    5341   1st Qu.:    4216
Median :   55641   Median :   46110   Median :   36854   Median :   30533
Mean   :  542449   Mean   :  462616   Mean   :  388424   Mean   :  333587
3rd Qu.:  223103   3rd Qu.:  189141   3rd Qu.:  180252   3rd Qu.:  152088
Max.   :49218872   Max.   :29240206   Max.   :48177680   Max.   :42558111
   M43_1_PD          M43_2_PD          M64_1_PD          M64_2_PD
Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
1st Qu.:   19641   1st Qu.:   17299   1st Qu.:   11038   1st Qu.:    8660
Median :   67945   Median :   59607   Median :   52249   Median :   47330
Mean   :  349020   Mean   :  358822   Mean   :  470655   Mean   :  484712
3rd Qu.:  205471   3rd Qu.:  201924   3rd Qu.:  209896   3rd Qu.:  206036
Max.   :35049402   Max.   :63082982   Max.   :71750330   Max.   :88912734
```

Com es pot observar, els mínims són tots 0, això és degut a que els valors de l'abundància no poden ser negatius, per tant, ens fixarem en els valors del primer quartil. Aquests són bastant propers entre les repeticions, d'igual manera passa amb els tercers quartils, que són també semblants entre repeticions. Per altra banda, els màxims són bastant dispars però això pot ser degut a valors extrems que estan afegint soroll a la mostra.
Per comprobar que les distirbucions són iguals es crean boxplots de cada mostra i es comprova visualment si les mitjes coincideixen i els valors màxims són realment valors extrems (outliers).

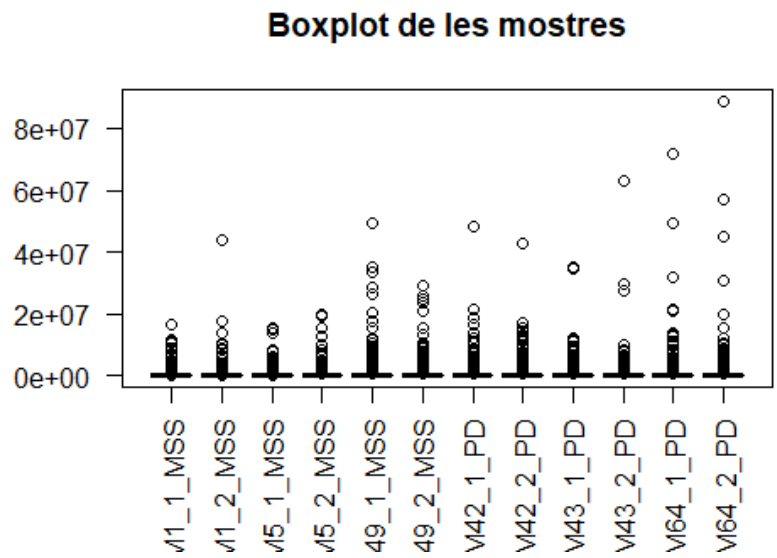

Figure 1: Boxplot de les dades crues

Es pot observar com les mitjanes semblen estar totes centrades al voltant del mateix valor però, no

es pot comprovar correctament visualment perquè els valors extrems fan que el gràfic sigui molt allargat. Tot i això, sí que es pot veure com sembla que els valors màxims de les mostres es deuen a outliers, i, per tant, no es descabellat pensar que les mitjanes siguin iguals i, les repeticions, aportin informació rellevant a l'estudi. Per poder fer una correcta comparació es passen els valors a escala logarítmica en base 10 i es torna a crear el boxplot:

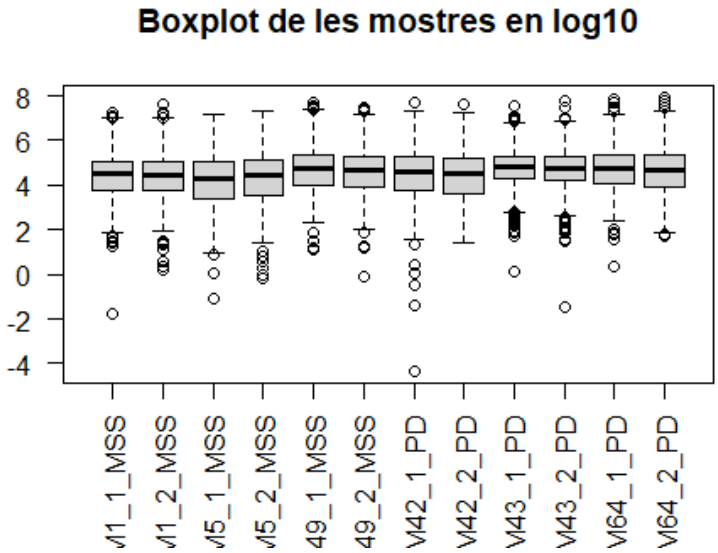

Figure 2: Boxplot de les dades en logaritma base 10

Ara sí, es veu més recolzada la hipòtesis que les mitjanes entre repeticions són semblants i de que són els outliers els que provoquen les diferències.

S'ha de tenir en compte que, en cas que les repeticions no fossin iguals en distribució, alguna cosa estaria inroduïnt biaix a les mostres com ara un error humà o un error en la maquinària que calcula l'abundància de fosfopèptids en les mostres.

Per corroborar que les disitrbucions de les mostres són iguales, s'aplica un test t de Student. Aquest test té com a hipòtesis nula que les mitjanes de les distribucions són iguals. Es considera valor llindar el p-valor 0.05, és a dir, si el p-valor d'un test és inferior a 0.05, es rebutjarà la hipòtesi nula d'igualtat entre les mitjanes de les mostres. Es mostra, a continuació, una taula dels p-valors entre les diferents repeticions.

	M1_MSS	M5_MSS	T49_MSS	M42_PD	M43_PD	M64_PD
p_values	0.614	0.474	0.369	0.404	0.888	0.902

Cap dels p-valors trobats en els test és inferior a 0.05, per tant, no tenim cap raó per pensar que les distribucions entre repeticions siguin diferents, tenint fins i tot en compte els outliers.

Un cop comprovat que totes les repeticions segueixen les mateixes ditribucions, es passa a comprovar si hi ha diferències significatives entre els dos grups de mostres, és a dir, si hi ha diferències entre les abundàncies dels fosfopèptids entre els dos grups de mostres de tumors. Per realitzar aquesta prova, es tornen a seguir les passes anteriors, es tornen a revisar els boxplots i s'aplica un test t Student aprofitant que les mostres són independents.

La nova hipòtesis es que els grups MSS i PD tinguin distribucions diferents d'abundàncies. Segons els boxplots observats amb antelació a la figura 2 no sembla que hi hagi una diferència significativa entre les mitjanes de les observacions, per confirmar-ho, com ja s'ha mencionat, es realitza un test t de Student entre els dos grups. Aquest test és més complexe que els anteriors on es comparava

per columnes, ara comparem els valors de cada fila, és a dir, els tres registres d'abundància corresponents a les tres mesures de MSS d'un fosfopèptid contra les tres mesures de PD. Per tant, es retornen 1438 p-valors dels quals 284 són inferiors a 0.05, en altres paraules, 284 files rebutjen la hipòtesis nula d'igualtat de mitjanes.

Aquest resultat ens indica que de 1438 registres d'abundància només 284 rebutjen la hipòtesis nula. Només són 284 registres els que es diferencien entre un tumor i l'altre de forma significativa.

Per acabar de comprobar la diferenciació entre grups es realitza un PCA per trobar els principals components del conjunt de dades. En la gràfica es poden veure els dos grups separats per aquests components, els corresponents a MSS es mostren en negre i els corresponents a PD es mostren en salmó.

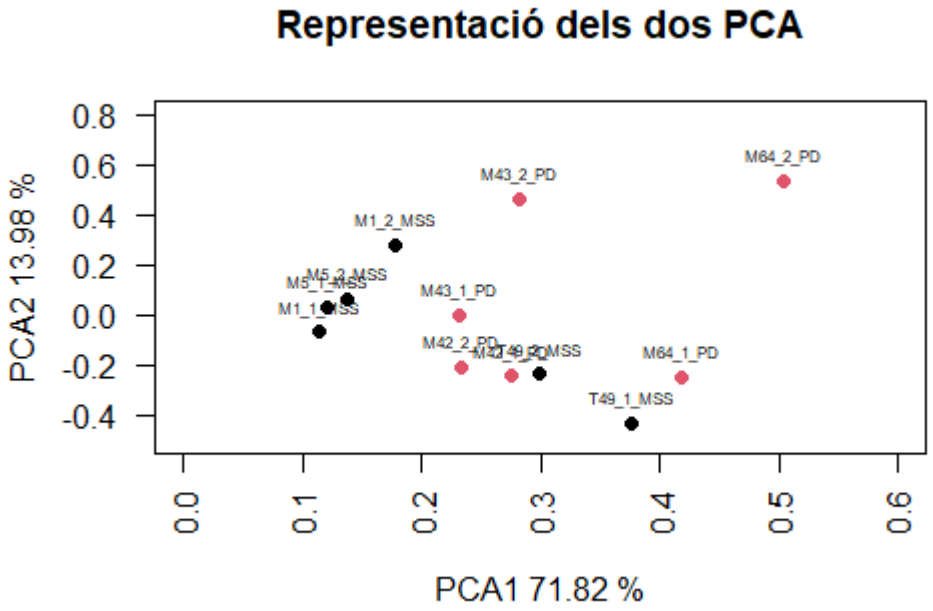

Figure 3: PCA de l'abundància entre els grups

D'una PCA s'espera veure grups separats dels diferents components. En el nostre cas, el grup de MSS sembla agregar-se al costat esquerra mentre que el grup PD s'agrega al dret. Tot i això, els grups no estan separats per complet, sino que hi ha mostres que es superposen entre elles. De fet, les dues repetecions de T49_MSS es troben agregadades amb les PD. Tot i les discrepances, les altres repeticiones sí que es troben segregades per color.

És interessant com el M64_2_PD es troba tant allunyat de la resta, sobretot tant allunyat de la seva primera repetició però, si ens fixem en el primer boxplot de la figura 1, es pot veure com és el conjunt d'observacions amb més valors extrems. Són justament aquest valors els que provoquen la seva separació del conjunt.

Es pot trobar el codi emprat per la creació dels boxplots, l'aplicació dels test t de Student i la execució i creació del gràfic de la PCA a l'Annex.

4 Discussió, limitacions i conclusions de l'estudi

Segons les probes realitzades, s'ha vist que hi ha grups d'observacions esbiaixats per culpa dels valors extrems extrets de l'experiment. Tot i això, els grups que presenten menys otuliers sí que s'ha vist a la PCA que s'agrupen correctament diferenciant els dos tipus de tumors.

A més, cal també mencionar els test t de Student realitzats, on s'ha comprovat que tot i no haver retirat els valors extrems de les mostres, les repeticions presenten les mateixes mitjanes i, per tant, distribucions semblants. Els boxplots, també han reforçat la hipòtesi de la igualtat de distribucions.

D'altra banda, els test t de Student ens han permés veure que només el 19.6% (284) de les mostres són significativament diferentes entre ambdos grups.

Un cop finalitzat l'anàlisi de les dades proporcionades, cal mencionar que es podria realitzar un estudi més extens retirant els valors extrems i tornant a realitzar l'anàlisi tant estadístic com de visualització, el que permetria treure soroll de les mostres i, potser, trobar diferències en els resultats finals.

Tota la documentació, dades i fitxer d'anàlisi es poden trobar al següent GitHub: https://github.com/CheyenneRomero
Freijo-Cheyenne-PEC1

Annex

Càrrega del fitxer de dades:

```
library(readxl)
ds <- read_xlsx("TIO2+PTYR_human_MSS+MSIvsPD.XLSX")
```

Creació de l'objecte *SummarizedExperiment*:

```
BiocManager::install("SummarizedExperiment")
library(SummarizedExperiment)
assay <- data.frame(ds[,5:16])
row_data <- ds[,1]
col_data <- data.frame(
  sample_id = colnames(assay),
  group = c(rep('MSS',6), rep('PD',6)),
  replica = c(rep(c(1,2),6))
)

se <- SummarizedExperiment(
  assays = list(counts = assay),
  rowData = row_data,
  colData = col_data
)
```

Creació de boxplots:

```
counts <- assay(se)
groups <- colData(se)$group
repeticions <- colData(se)$replica

# Boxplots
library(ggplot2)
par(las = 2)
boxplot(counts, main = 'Boxplot de les mostres')

# Es passen els valors a escala logar tmica en base 10
par(las = 2)
boxplot(log10(counts), main = 'Boxplot de les mostres en log10')
```

Aplicació de Student t-test:

```
# Es volen comparar mostres independents: MSS1 vs MSS2
c_mss_1 <- counts[groups == "MSS" & repeticions == 1]
c_mss_2 <- counts[groups == "MSS" & repeticions == 2]
c_pd_1 <- counts[groups == "PD" & repeticions == 1]
c_pd_2 <- counts[groups == "PD" & repeticions == 2]

p_mss <- c()
for (i in 1:dim(c_mss_1)[2]){
  p_mss <- c(p_mss,t.test(c_mss_1[,i], c_mss_2[,i])$p.value)
}

p_pd <- c()
for (i in 1:dim(c_pd_1)[2]){
  p_pd <- c(p_pd,t.test(c_pd_1[,i], c_pd_2[,i])$p.value)
}

# Diferencies entre els dos grups MSS i PD per repetici  si
#s n diferents, general si son iguals
c_mss <- counts[groups == "MSS"]
```

```
c_pd <- counts[groups == "PD"]

p <- c()
for (i in 1:dim(c_mss)[1]){
  p <- c(p,t.test(c_mss[i,], c_pd[i,])$p.value)
}

# Es comproba quants registres rebutjen la H_0
length(which(p < 0.05))
```

Creació i execució de PCA:

```
# S'aplica una PCA
pcs <- prcomp(counts)

# Es ploteja la PCA
plot(pcs$rotation[,1], pcs$rotation[,2],
     main="Representaci dels dos PCA",
     xlab = paste("PCA1", round(c(pcs$sdev^2 / sum(pcs$sdev^2))[1]*100,2), "%"),
     ylab = paste("PCA2", round(c(pcs$sdev^2 / sum(pcs$sdev^2))[2]*100,2), "%"),
     xlim = c(0, 0.6), ylim = c(-0.5, 0.8),
     col = as.factor(groups), pch = 19)
text(pcs$rotation[,1], pcs$rotation[,2], colnames(counts), cex=0.5, pos=3)
```