PAC1

Judit Tella Vila

2025-04-02

S'ha triat un data set del repositori de GitHub: "CIMBC" de l'estudi Chan et al. (2016) en el British Journal of Cancer.

Tot el procediment de programació en Rstudio es troba en Annexes, ja que també hi ha l'output dels mateixos cosa que en algunes ocasions no serà estrictament necessari per a l'informe (amb un màxim de 10 pàgines).

En l'inici veiem que el data set es troba en dues fulles de l'Excel i obtenim informació de mostres i dels metabòlits implicats.

Fent un primer anàlisi exploratori es veu que hi ha un total de 140 mostres, de les quals com a pacients es consideren aquelles etiquetades com a "Sample", quedant un total de 123 mostres. Referent als metabòlits analitzats n'hi ha 149, etiquetats com a M + un número. Per tenir la informació de cada metabòlit s'ha d'accedir a la part "peak" de l'objecte creat.

En el propi data set hi ha valors NA (967), els quals no s'han eliminat ja que no és que sigui la presència de NA en algunes mostres, sinó que sembla ser per raons de metadata, és a dir, les mostres són un agrupament de dades existents en les quals no s'han categoritzat els mateixos metabòlits en totes les mostres, a part de que son 967 NAs vs. 18.327 registres en la taula.

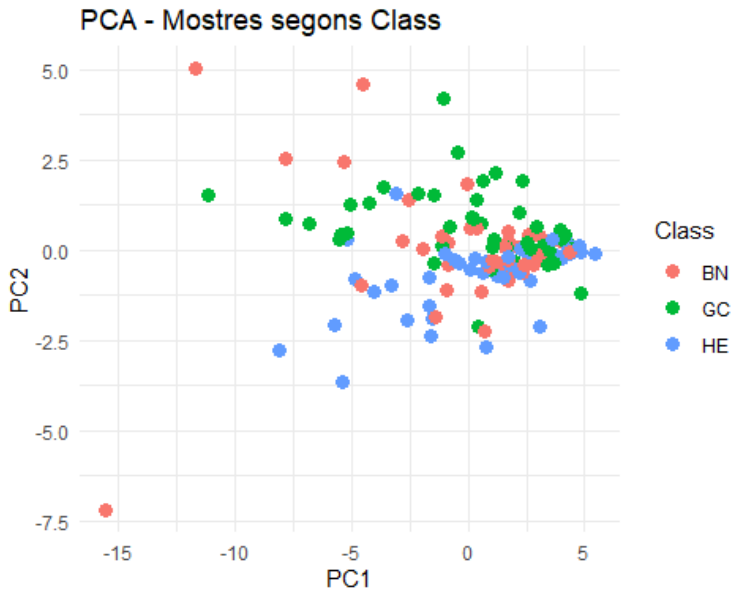
Les mostres tenen 3 categories (Class) que classifiquen les mostres en 40 tumors benignes (benign gastric disease; BN), 43 càncer gàstric (gastric càncer; GC) i 40 controls sans (healthy; HE).

Una de les preguntes que ens podem fer és: hi ha diferència entre els grups clínics i els metabòlits presents en la mostra?

Per respondre aquesta pregunta podem realitzar una PCA i una ANOVA.

Per veure la distribució de les mostres segons la classe podem realitzar un Anàlisi de Components Principals (Principal Component Analisis, PCA) per reduir

la dimensionalitat de les dades i poder veure en dos components principals el major percentatge de variancia:



Imatge 1 PCA (Principal Component Analysis) de les dades escollides de metabolòmica escollides

En aquesta PCA no veiem una clara separació de les mostres, per tant, no sembla haver una clara diferenciació de perfils metabòlics entre grups. A part, veiem també que el percentatge explicat per cada un dels components és relativament baix.

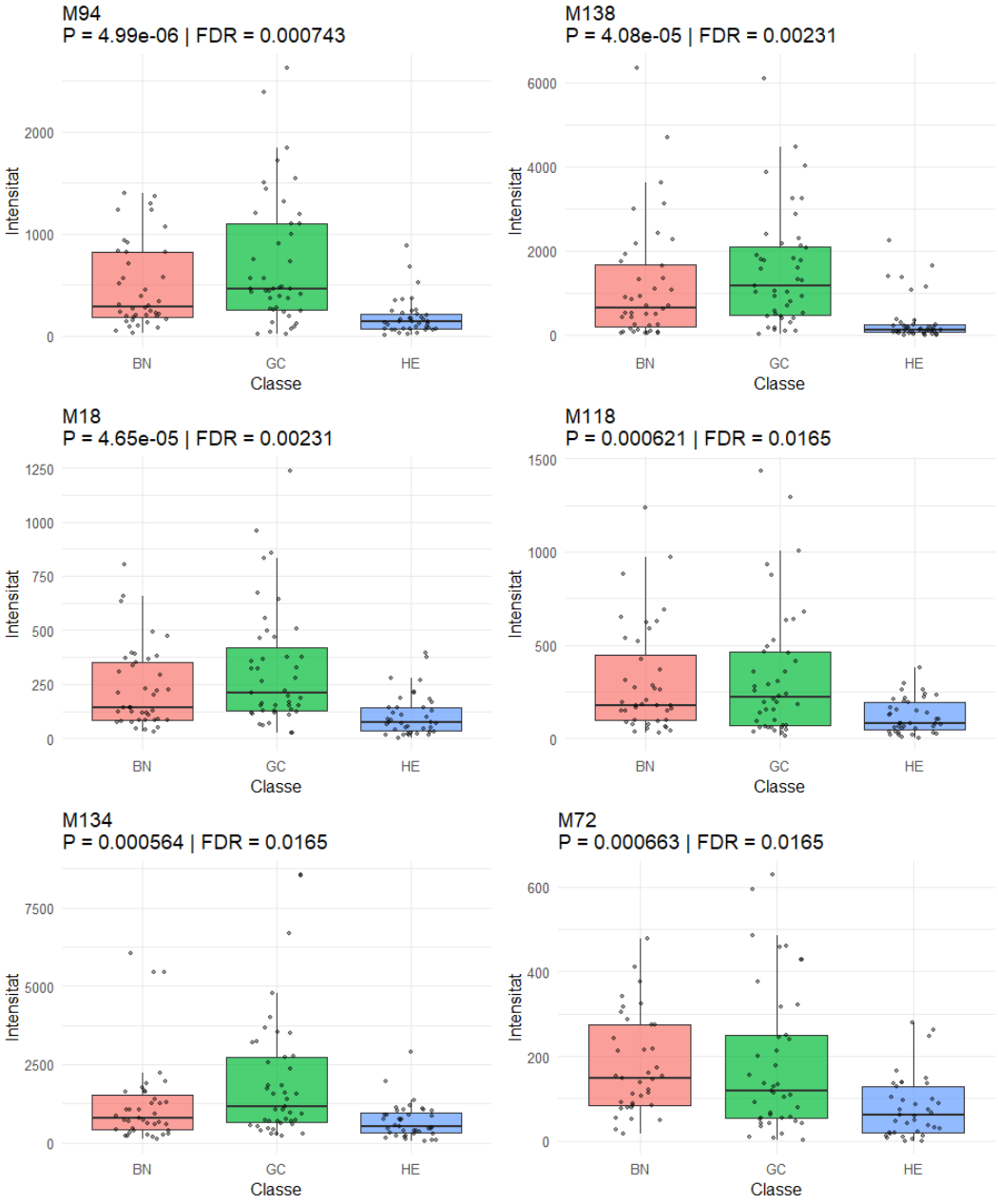
En canvi, en els resultats de l'ANOVA, podem mirar si algun dels metabòlits és diferent en algun dels grups, és a dir, si es troben diferències significatives. En l'ANOVA realitzada surten 10 metabòlits significativament diferents:

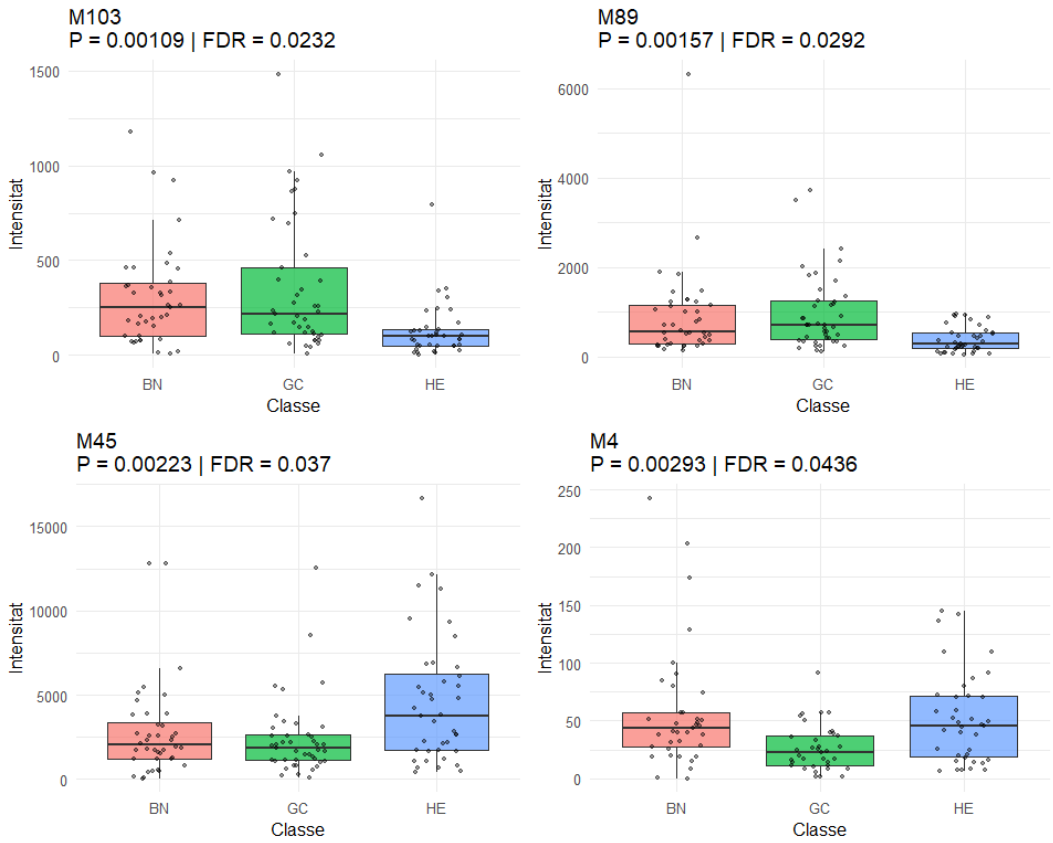
Taula 1 Metabòlits obtinguts en l'anàlisi ANOVA resultant com a significatius després de la correcció FDR.

Name (MX)	p-value	fdr	Label
M94	4.986988e-06	0.0007430612	N-Phenylacetylglycine
M138	4.078888e-05	0.0023093032	u233
M18	4.649604e-05	0.0023093032	3-Indoxylsulfate
M118	6.206694e-04	0.0164697126	Tropate
M134	5.641083e-04	0.0164697126	u144

M72	6.632099e-04	0.0164697126	Indole-3-acetate
M103	1.089648e-03	0.0231939299	Phenylalanine
M89	1.570093e-03	0.0292429748	N-AcetylglutamineDerivative
M45	2.232977e-03	0.0369681684	Citrate
M4	2.926107e-03	0.0435989976	1-Methylnicotinamide

Aquests metabòlits són els que han resultat significatius en quant a diferències entre els grups mencionats anteriorment, però per veure millor com són aquestes diferències podem fer els següents plots:



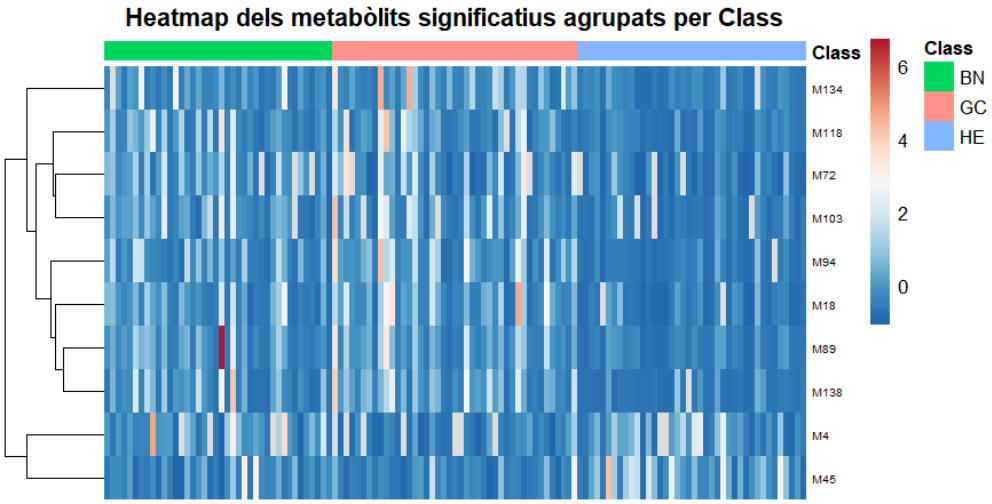


Imatge 2 Boxplots generats a partir dels metabòlits significatius, mostrant com són aquestes diferències entre grups. Veiem els controls en blau (HE), els tumors benignes en una tonalitat rosada (BN) i els casos de càncer gàstric en verd (GC).

Veiem com en M4 i M45 el metabòlit és més freqüent en el grup control (1-Methylnicotinamide i Citrate, respectivament), mentre que en la resta de gràfics veiem que els metabòlits dels grups control són menors que en els grups GC i BN. En alguns casos sembla veure's una diferència entre GC i BN, com ara en M94, M138 i M134 (N-Phenylacetylglycine, u233 i u144, respectivament), però no tant clar com en el cas dels control.

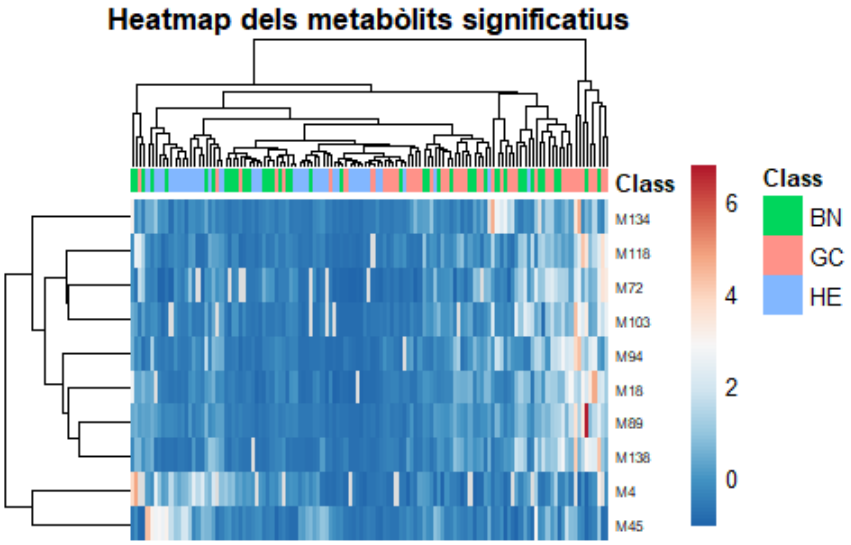
Una altra pregunta que podem fer-nos si es trobessin diferències entre grups clínics és: existeixen perfils metabòlics característics dels grups clínics?

Un tipus comú d'anàlisi també són els Heatmaps, cosa que ens permetrà veure si hi ha presència de perfils similars o agrupacions interessants. Segons hem vist en la PCA, esperem no trobar un perfil clar per cada grup clínic. Això és confirma veient el següent heatmap agrupat per aquests grups:



Imatge 3 Heatmap obtingut suposant que la informació clínica de la mostra podria ser una bona clusterització dels perfils metabòlics d'aquells que havien resultat significatius.

Ara, si no especifiquem l'agrupació mirem si hi ha o no perfils metabòlics clars en algun altre tipus d'agrupació:



Imatge 4 Heatmap utilitzant clusterització per veure si hi ha perfils metabòlics clars en els que hi ha diferència entre grups.

En aquest cas tampoc es veu una agrupació clara de les mostres i tampoc perfils marcats en els grups, sinó que sembla que tenim bastanta heterogeneïtat de perfils per a cada un dels metabòlits que són significativament diferents entre classes.

Conclusions

Mitjançant l'anàlisi ANOVA realitzat i segons els resultats obtinguts (Taula 1), s'observen diferències metabòliques entre el grup control (HE) i els grups amb presència de tumor (tant benigne com cancerós) (Imatge 2). En concret, s'han identificat 10 metabòlits amb diferències significatives que podrien actuar com a possibles marcadors de presència tumoral.

Tot i així, l'anàlisi dels heatmaps (Imatges 3 i 4) no permet concloure l'existència d'un perfil metabòlic clarament diferenciador entre els grups clínics (GC i BN). No s'han observat diferències estadísticament significatives entre tumors benignes i cancerosos, fet que impedeix considerar aquests metabòlits com a possibles biomarcadors per a distingir la naturalesa del tumor.

Cal considerar també les limitacions associades tant a la mostra com a la pròpia anàlisi. Dels 149 metabòlits analitzats, només 10 han mostrat diferències significatives entre grups, i cap d'ells permet discriminar entre tumors benignes i cancerosos. A més, la mida de la mostra, tot i ser de 123 individus, presenta una distribució ajustada entre grups (40 HE, 43 GC i 40 BN), cosa que limita el poder estadístic de l'estudi. També cal destacar la presència de 967 valors absents (NAs), que poden haver afectat negativament els resultats de l'ANOVA.

En comparar els resultats d'aquest estudi amb els de Chan et al. (2016), s'observen algunes discrepàncies. Tot i que també s'han detectat diferències entre els grups HE, GC i BN, no s'han trobat diferències significatives entre GC i BN, a diferència de l'estudi de referència. Dels metabòlits destacats per Chan et al., només el 3-indoxylsulfate ha resultat significatiu en aquest cas; altres, com l'alanina o el 2-hydroxyisobutyrate, no han estat identificats en la Taula 1.

Aquesta discrepància de resultats pot atribuir-se principalment a les diferències metodològiques. L'estudi de Chan et al. utilitza tests no paramètrics (com el Mann–Whitney U-test amb correcció per múltiples comparacions), així com anàlisis estadístiques multivariants com el Partial Least Squares Discriminant Analysis (PLS-DA) i l'Orthogonal PLS-DA (O-PLS-DA), enfocats a trobar correlacions complexes entre les dades.

Com a limitació principal d'aquest estudi preliminar, cal destacar la manca de temps per aprofundir en metodologies multivariants, la diferència d'experiència respecte als professionals implicats en l'estudi original i la manca d'un equip interdisciplinari que aporti robustesa i qualitat a l'anàlisi.

Tot i això, aquest conjunt de dades ha resultat molt útil per adquirir experiència amb la plataforma Bioconductor i iniciar-se en l'anàlisi estadística de dades metabolòmiques.

Repositori GitHub: https://github.com/Judit-tv/Tella_Vila_Judit_PAC1