PAC1 - Anàlisi de dades òmiques

Ricard Gonzalez

2025-03-29

Índex

1 Resum

Aquest treball presenta la construcció i validació d'un objecte de tipus SummarizedExperiment a partir d'un conjunt de dades metabolòmiques obtingudes d'anàlisis d'orina de pacients amb caquèxia i controls sans. Les dades, prèviament netes i sense valors perduts, han estat integrades en una estructura òmica moderna que permet encapsular simultàniament les mesures i les metadades de cada mostra.

A partir d'aquest objecte, s'han aplicat dues tècniques d'anàlisi exploratòria —escalat multidimensional (MDS) i anàlisi de components principals (PCA)— amb l'objectiu de detectar possibles patrons diferenciadors entre grups. Els resultats obtinguts mostren una manca de separació clara, si bé es poden observar tendències parcials d'agrupament en alguns pacients caquèxics.

2 Objectius

Aquest treball està dedicat al compliment d'una sèrie d'objectius, esmentats a continuació:

1. Seleccionar un conjunt de dades ("*dataset*") de metabolòmica per al seu ús durant l'informe.
2. Construïr, a partir d'aquest *dataset*, un objecte de tipus *SummarizedExperiment*.
3. Comparar, a nivell conceptual, l'objecte de tipus *SummarizedExperiment* amb el seu predecessor *ExpressionSet*.
4. Dur a terme una anàlisi exploratòria de les dades.
5. Organitzar la informació, mètodes emprats, resultats de l'anàlisi i discussió i conclusions sobre aquests en un informe estructurat.
6. Creació d'un repositori de GitHub, esmentat al propi informe, on s'hi recullin totes les dades i materials necessaris per a replicar l'anàlisi descrit a l'informe.

Alguns dels objectius tindràn cabuda explícita a l'informe (per exemple, l'anàlisi exploratòria de les dades). Altres, en canvi (com la creació del propi informe, o la generació d'un repositori de GitHub) son purament operatius i per tant se'n presentaràn els entregables directament (el propi informe, o un enllaç d'accés al repositori, respectivament).

3 Mètodes

3.1 Tria del conjunt de dades

Per aquest informe, es van considerar diversos conjunts de dades metabolòmiques presents al repositori proveït per l'assignatura.

Entre d'altres, hi havia conjunts de dades sobre estudis de cancer gàstric, assajos de fosfoproteòmica o pèrdua de massa muscular (caquèxia).

El conjunt de dades escollit ha estat el "2024-Cachexia", obtingut del repositori en format valors separats per comes (.CSV, per les seves sigles en anglès). Al llarg de l'informe, s'anomenarà *cachexia* per facilitar la lectura.

Els motius que justifiquen la tria d'aquest conjunt de dades han estat el seu bon balanç cardinalitat/dimensionalitat (més mostres que dimensions), el fet que totes les variables no-metadades eren numèriques, i la claredat del conjunt de dades (sense valors perduts).

Altres conjunts de dades, tot i interessants, tenien especificitats que els feien poc pràctics per un exercici centrat en entendre aspectes bàsics de Bioconductor, la programació orientada a objectes i la familiarització bàsica amb dades òmiques.

Aquest conjunt de dades recull anàlisis d'orina de 77 pacients, 47 dels quals presenten caquèxia i els 30 restants son controls. De cada mostra, s'hi han determinat experimentalment els valors de 63 metabòlits.

3.2 Paquets de software utilitzats

Per a la interacció amb les dades, es va utilitzar la versió *4.4.2* de **R** [1], i la versió de **RStudio** *2024.12.1+563* [2]. L'informe es va generar mitjançant el paquet **quarto** en la seva versió *1.4.4*.

Es van importar les dades mitjançant el paquet **data.table**, en la seva versió *1.17.0*.

Respecte la manipulació de les dades, es va fer servir **Bioconductor** en la seva versió *3.20*[3], del què es van utilitzar els paquets **Biobase** en la versió *2.66.0* [4] i **SummarizedExperiment** en la versió *1.36.0*[5]. També es van utilitzar funcions del paquet **S4Vectors**, en la versió *0.44.0*.

Per a l'anàlisi de les dades només es van utilitzar funcions del paquet **stats**, inclòs amb R. Per tant, no és requisit configurar versions addicionals de paquets per a l'anàlisi més enllà de la versió de R.

Finalment, la presentació i visualització de les dades en figures i taules es va fer amb els paquets **ggplot2**, en versió *3.5.1* [6], **kableExtra**, en versió *1.4.0* [7] i **showtext**, en versió *0.9-7*[8].

3.3 Creació de l'objecte SummarizedExperiment

Per a la generació d'un objecte de tipus *SummarizedExperiment* a partir del conjunt de dades *cachexia*, es va seguir una estratègia metòdica, centrada en la identificació i organització progressiva del contingut del conjunt de dades per a estructurar-lo en parts de format desitjat.

En primer lloc, es van separar les metadades (presents a les dues primeres columnes del conjunt, corresponent a l'identificador únic del pacient, "Patient ID", i al Grup experimental "Muscle loss").

Les dades restants es van transformar en matriu transposada: aquest pas és clau, atès que els objectes *SummarizedExperiment* requereixen les variables com a files, i les observacions com a columnes. Per assegurar la integritat de les dades i la consistència relacional, es va asignar identificadors de pacient tant a la matriu de resultats metabolòmics com a la taula de metadades.

Es va construïr un objecte *SummarizedExperiment* assignant la matriu transposada i etiquetada a l'*slot* "*assays*", amb el nom de "*counts*". A l'*slot* "*colData*" s'hi van assignar les metadades.

L'objecte es va desar en format binari (.Rda) per facilitar-ne la reutilització.

3.4 Anàlisi exploratòria de les dades

3.4.1 Objectius i preparació de les dades

Per a una exploració preliminar de les dades, es va decidir examinar si era possible capturar els dos "Grups de tractament" (els pacients caquèxics versus els pacients control) en funció de la seva distància inter-pacient. De manera similar, també es va intentar capturar la variabilitat entre mostres (i, si fos possible, Grups de tractament) en funció de la covariància.

Per a resoldre tots dos punts, es va procedir obtenint la matriu de metabòlits en la seva estructura original (amb els pacients com a files i els valors de metabòlit com a columnes). Es va etiquetar degudament la matriu amb codis de pacient i es va normalitzar amb la funció *scale()*, que per defecte centra i escala les dades a $\mu = 0$ i $\sigma = 1$.

Addicionalment, es va calcular la matriu de distàncies entre pacients a partir de la matriu normalitzada, mitjançant la mètrica de distància Euclídia.

3.4.2 Escalat Multidimensional

Per a intentar distingir els pacients a través de la seva distància, es va utilitzar la matriu de distàncies euclídies per a Escalat Multidimensional (MDS, per les seves sigles en anglès), mitjançant la funció *cmdscale()*. Aquesta funció es va utilitzar per obtenir una representació bidimensional de les dades projectades a través de les seves distàncies euclídies.

Aquest resultat es va etiquetar posteriorment amb el Grup de tractament de cada pacient mitjançant el seu identificador, i es van visualitzar les dades.

3.4.3 Anàlisi de Components Principals

Finalment, per capturar la variabilitat entre pacients a través de la seva covariància, es va fer servir un Anàlisi de Components Principals (PCA).

Amb aquesta finalitat, la matriu normalitzada obtinguda anteriorment es va sotmetre a descomposició en components principals mitjançant la funció *prcomp()*. Es va emprar per extreure la projecció de les dades en cadascuna de les components principals, i les dues primeres components (PC1 i PC2) es van utilitzar,

amb les dades etiquetades, per visualitzar els pacients en funció del seu Grup de tractament.

4 Resultats

4.1 Obtenció de l'objecte SummarizedExperiment

Seguint la estratègia esmentada als Mètodes, s'ha obtingut un objecte SummarizedExperiment a partir del conjunt de dades *cachexia*.

Les dimensions originals del conjunt de dades *cachexia* són 77 observacions i 65 columnes, de les quals 63 corresponen a variables metabolòmiques.

L'objecte creat conté per separat les metadades de la matriu de mesures metabolòmiques. Per tant, les dimensions originals de *cachexia* passen a estar repartides. Les dimensions de l'*slot* "*assay*", que conté les dades metabolòmiques, son les següents:

Table 1: Dimensions de l'objecte SummarizedExperiment 'dades'.

Component	Valor
Metabolits	63
Mostres	77

Cal recordar que la matriu està transposada respecte el conjunt de dades original, ja que *SummarizedExperiment* espera les observacions com a columnes i les variables com a files.

Addicionalment, l'objecte conté a *assay* les dades:

Table 2: Primeres 5 files i 5 columnes de la matriu de metabolits de l'objecte 'dades'.

	PIF_178	PIF_087	PIF_090	NETL_005_V1	PIF_115
1,6-Anhydro-beta-D-glucose	40.85	62.18	270.43	154.47	22.20
1-Methylnicotinamide	65.37	340.36	64.72	52.98	73.70
2-Aminobutyrate	18.73	24.29	12.18	172.43	15.64
2-Hydroxyisobutyrate	26.05	41.68	65.37	74.44	83.93
2-Oxoglutarate	71.52	67.36	23.81	1199.91	33.12

I a l'*slot* "*colData*" les metadades:

Table 3: Metadades de les primeres 5 mostres de l'objecte 'dades'.

	Patient.ID	Muscle.loss
PIF_178	PIF_178	cachexic
PIF_087	PIF_087	cachexic
PIF_090	PIF_090	cachexic
NETL_005_V1	NETL_005_V1	cachexic
PIF_115	PIF_115	cachexic

4.2 Diferències entre SummarizedExperiment i ExpressionSet

Per completar l'objectiu de la PAC, s'ha dut a terme una comparació conceptual vía la secció d'ajuda de RStudio entre l'objecte SummarizedExperiment i el seu antecessor ExpressionSet, tots dos definits com a classes de Bioconductor.

Tots dos formats permeten encapsular dades experimentals juntament amb informació de mostra i metadades, però es diferencien en alguns aspectes:

• **ExpressionSet** agrupa totes les dades en una estructura més rígida, i permet una única matriu de dades a *assay*.

• **SummarizedExperiment**, en canvi, està dissenyat per a ser més flexible: pot gestionar múltiples assays (mesures), associar metadades enriquides i adaptar-se millor a tipus de dades diversos com ara proteòmica, metabolòmica o epigenètica (no només transcriptòmica).

En aquest projecte, la diferència pràctica entre tots dos formats ha estat limitada ja que s'ha utilitzat únicament *SummarizedExperiment*. No obstant, el seu ús sembla extès a diversitat de paquets de Bioconductor i millor alineat amb les bones pràctiques d'anàlisi en el camp de la bioinformàtica. Per exemple, si haguéssim disposat també de dades de proteòmica, el nostre SummarizedExperiment hagués pogut gestionar-les simultàniament. En canvi, un ExpressionSet hagués estat limitat a una única matriu.

4.3 Resultats de l'anàlisi per MDS

La següent figura projecta les mostres en dues dimensions mantenint les distàncies euclídies originals.

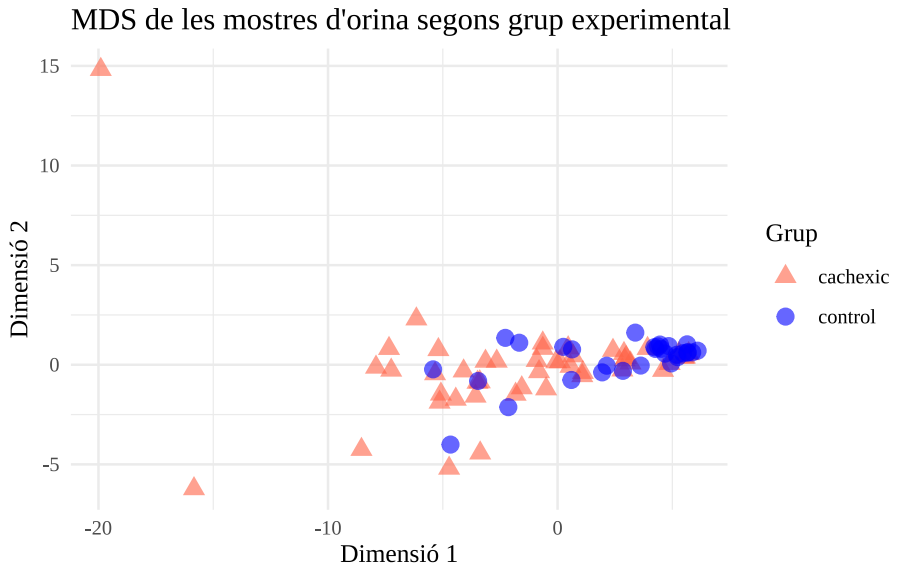

Figure 1: Diagrama de dispersió que mostra la projecció en 2 dimensions del dataset cachexia transformat via escalat multidimensional (MDS). Els pacients caquèxics apareixen com a triangles vermells, i els controls com a cercles blaus.

No s'observa una separació clara entre mostres caquèxiques i controls, però s'hi observen algunes agrupacions parcials. Els pacients caquèxics, en particular, presenten tendència a agrupar-se cap a valors més negatius de la primera dimensió, tot i que tots dos grups coexisteixen a regions comunes de l'espai reduït.

La Dimensió 1 sembla capturar una part notable de la variabilitat entre pacients, però no permet distingir de manera precisa entre grups.

4.4 Resultats de l'anàlisi per PCA

Les dues components principals expliquen, respectivament, el 40% i el 8% de la variància total de les dades.

Es pot apreciar, com a la projecció de l'MDS anterior, una manca de separació nítida entre grups. Tot i així, també s'hi detecta una certa tendència de pacients caquèxics allunyats en direcció als extrems, separats de les regions de la projecció on coexisteixen tots dos grups d'individus.

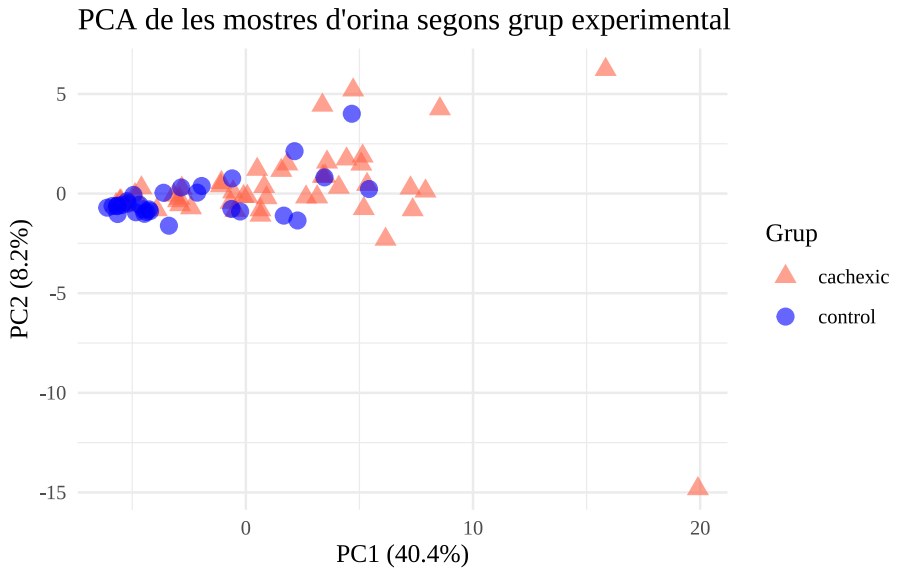

Figure 2: Diagrama de dispersió que mostra la projecció en 2 dimensions del dataset cachexia transformat via anàlisi de components principals (PCA). Cada eix inclou entre parèntesi la variància explicada per la corresponent Component Principal. Els pacients caquèxics apareixen com a triangles vermells, i els controls com a cercles blaus.

De la mateixa manera que a l'escalat multidimensional, s'observa una elevada dispersió.

En conjunt, les anàlisis exploratòries vía MDS i PCA no mostren una separació clara entre els grups de les primeres dimensions projectades, tot i la presència de tendències parcials d'agrupament en pacients caquèxics.

5 Discussió

L'objectiu principal d'aquest treball ha estat construïr un objecte de tipus SummarizedExperiment a partir d'un conjunt de dades òmiques i aplicar tècniques d'anàlisi exploratòria per avaluar la presència de patrons associats a dos grups experimentals: pacients amb caquèxia i pacients control. Amb aquest propòsit, s'han aplicat tècniques de reducció de la dimensionalitat (MDS i PCA) sobre dades metabolòmiques derivades d'anàlisis d'orina, prèviament normalitzades.

El conjunt de dades escollit (*cachexia*) presentava diverses característiques favorables: una mida moderada (77 mostres, 63 metabòlits), absència de valors perduts i una classificació binària clara de les mostres. Aquestes propietats el feien especialment adequat per a una anàlisi exploratòria supervisada i per la construcció d'objectes de tipus SummarizedExperiment, permetent integrar tant les dades com les metadades en un format estructurat i reutilitzable. Malgrat això, la baixa separació observada entre grups indica que, en aquest cas, la

informació disponible pot no ser suficient per capturar els patrons distintius de la caquèxia amb claredat mitjançant reducció de dimensionalitat.

Els resultats obtinguts a través de PCA [9] i MDS [10] apunten en una direcció similar: la separació entre grups no és nítida. Si bé s'observen tendències parcials d'agrupament —particularment en alguns pacients caquèxics que es distribueixen cap a valors més extrems de la primera dimensió—, no es poden identificar clústers ben definits ni en l'espai de distàncies (MDS) ni en l'espai de variància (PCA).

Aquest resultat pot ser interpretat com una mostra de l'heterogeneïtat pròpia de la caquèxia, en ser una síndrome multifactorial amb manifestacions clíniques variables, o com una limitació del tipus de dades disponibles. És possible que altres fonts de variabilitat —com l'estat nutricional, la teràpia farmacològica, la composició corporal o el perfil inflamatori— exerceixin un paper important però no estiguin reflectides en les mesures incloses.

La comparació entre MDS i PCA reforça la validesa dels resultats, ja que totes dues tècniques projecten les mostres de manera coherent, indicant una manca de separació estructural robusta entre grups. El fet que tant una tècnica basada en distàncies com una basada en variància arribin a conclusions convergents suggereix que la distribució observada és una propietat intrínseca del conjunt de dades, i no un artefacte d'algun mètode específic. Això aporta un cert grau de robustesa a la conclusió que, amb les dades actuals, no es poden distingir clarament els dos grups.

Malgrat l'ús de tècniques clàssiques de reducció de dimensionalitat, és probable que l'aplicació de mètodes alternatius com ara *t-distributed Stochastic Neighbor Embedding* (t-SNE) [11] o *Uniform Manifold Approximation and Projection* (UMAP) [12] pogués millorar la separació entre grups en cas que hi hagi estructures locals difícils de capturar linealment. De la mateixa manera, si l'anàlisi hagués estat *supervisat* (per exemple, emprant tècniques de regressió compatibles amb alta dimensionalitat) en comptes de *no-supervisat* (basant-se en anàlisi de conglomerats), podria ser que s'haguessin descobert relacions i tendències als pacients que escapin a una representació basada només en els dos primers components. Aquestes opcions podrien ser considerades en futurs treballs.

Finalment, cal destacar que el format SummarizedExperiment ha demostrat ser una estructura eficaç per integrar les dades analitzades. A diferència del seu antecessor, l'objecte ExpressionSet, el SummarizedExperiment separa clarament els assays i les metadades, i permet compatibilitat amb nombrosos paquets moderns de Bioconductor. En aquest cas, tot i que la diferència funcional entre ambdós formats no ha estat determinant per al tipus d'anàlisi realitzat, l'ús de SummarizedExperiment representa una pràctica alineada amb els estàndards actuals d'anàlisi òmica reproducible.

Tot i així, l'estudi té algunes limitacions, que s'han de tenir en compte.

En primer lloc, la mida del conjunt de dades (77 observacions per 63 variables numèriques) és limitant: la seva cardinalitat pot ser insuficient per capturar patrons complexos, i pot haver contribuït a una no-discriminació artefactual entre grups de tractament i haver ocultat diferències biològicament rellevants.

De la mateixa manera, les tècniques emprades son purament no-supervisades,

pel què poden no ser les més adïents. Tot i que l'objectiu principal no era discriminar entre grups directament sino *observar la presència orgànica de separació*, una tècnica supervisada amb un objectiu més dirigit (com ara un arbre de decisió) podria haver facilitat la descoberta de patrons rellevants.

Aquestes limitacions han de ser tingudes en compte a l'hora d'interpretar els resultats i extreure conclusions generals.

6 Conclusions

En aquest treball s'ha aconseguit construir correctament un objecte de tipus SummarizedExperiment a partir d'un conjunt de dades metabolòmiques real, integrant tant les mesures quantitatives com les metadades experimentals en un format estructurat i reutilitzable. Aquesta estructura ha facilitat l'aplicació d'anàlisis exploratòries clàssiques, com l'escalat multidimensional (MDS) i l'anàlisi de components principals (PCA), per avaluar la presència de patrons associats a la caquèxia.

Els resultats han mostrat una manca de separació clara entre grups, tant en termes de variància com de distància interindividual. Tot i que s'han identificat tendències parcials d'agrupament en alguns pacients caquèxics, aquestes no han estat suficients per establir una discriminació robusta. Aquest fet pot reflectir la naturalesa difusa de la caquèxia o bé limitar-se a la informació proporcionada per les dades analitzades.

Metodològicament, s'ha comparat SummarizedExperiment amb el seu antecessor ExpressionSet, valorant-ne les diferències estructurals i la seva rellevància en contextos d'anàlisi òmica moderna.

En conjunt, el treball ha permès posar en pràctica les competències associades a la manipulació d'objectes òmics en R, la normalització de dades, l'aplicació d'anàlisis exploratòries i la valoració crítica dels resultats obtinguts.

7 Referències

Els materials emprats, dades i arxius amb codi necessaris per a la reproducció del treball es poden trobar en aquest repositori de GitHub[13]. En aquest, s'hi pot trobar addicionalment un arxiu en format Markdown que hi conté metadades del conjunt de dades emprat.

[1] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2024.

[2] RStudio Team. RStudio: Integrated development environment for r. Boston, MA: Posit Software, PBC; 2024.

[3] Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 2004;5:R80. https://doi.org/10.1186/gb-2004-5-10-r80.

[4] Gentleman R, Carey V, Morgan M, et al. Biobase: Base functions for bioconductor. 2024. https://doi.org/10.18129/B9.bioc.Biobase.

[5] Morgan M, Obenchain V, Hester J, et al. SummarizedExperiment: A container for matrix-like assays. 2024. https://doi.org/10.18129/B9.bioc.SummarizedExperiment.

[6] Wickham H. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York; 2016.

[7] Zhu H. kableExtra: Construct complex table with 'kable' and pipe syntax. 2024. https://doi.org/10.32614/CRAN.package.kableExtra.

[8] Qiu Y. Showtext: Using fonts more easily in r graphs. 2023. https://doi.org/10.32614/CRAN.package.showtext.

[9] Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 1901;2:559–72.

[10] Torgerson WS. Multidimensional scaling: I. Theory and method. Psychometrika 1952;17:401–19.

[11] Maaten L van der, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research 2008;9:2579–605.

[12] McInnes L, Healy J, Saul N, et al. UMAP: Uniform manifold approximation and projection. Journal of Open Source Software 2018;3:861. https://doi.org/10.21105/joss.00861.

[13] Gonzalez R. Gonzalez-rodriguez-ricard-PEC1 (PAC1) 2025.