PAC1 - Anàlisi de dades òmiques

Ricard Gonzalez

2025-03-29

Índex

1 Resum

2 Objectius

Aquest informe 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**, i la versió de **RStudio** *2024.12.1+563*. 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*, del què es van utilitzar els paquets **Biobase** en la versió *2.66.0* i **SummarizedExperiment** en la versió *1.36.0*. 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*, **kableExtra**, en versió *1.4.0* i **showtext**, en versió *0.9-7*.

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 exploratori 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.

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

Component	Valor
Metabolits	63
Mostres	77

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

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*.

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

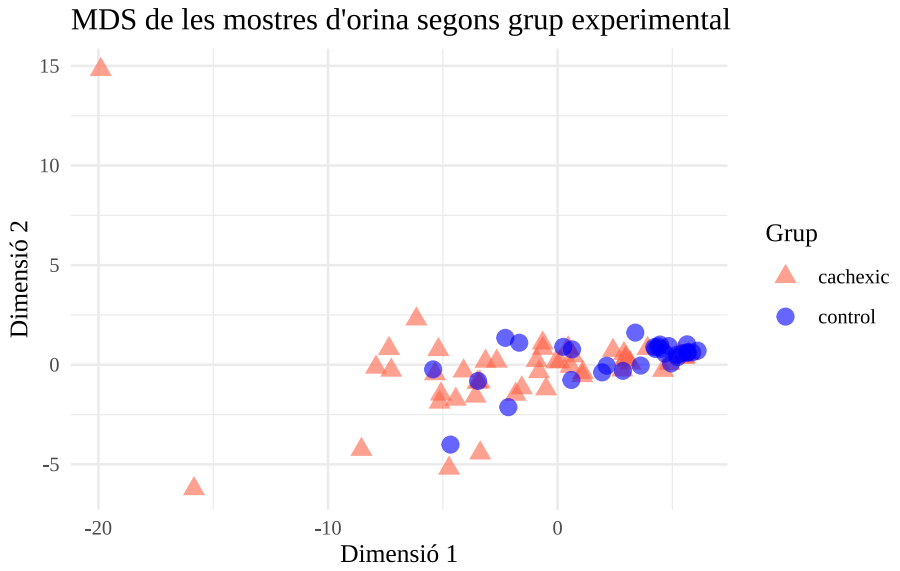

Figure 1: Diagrama de dispersió que mostra la projecció en 2 dimensions del dataset cachexia transformat via escalat multidimensional (MDS). Els pacients caquèxics apareixen com a triangles vermells, i els controls com a cercles blaus.

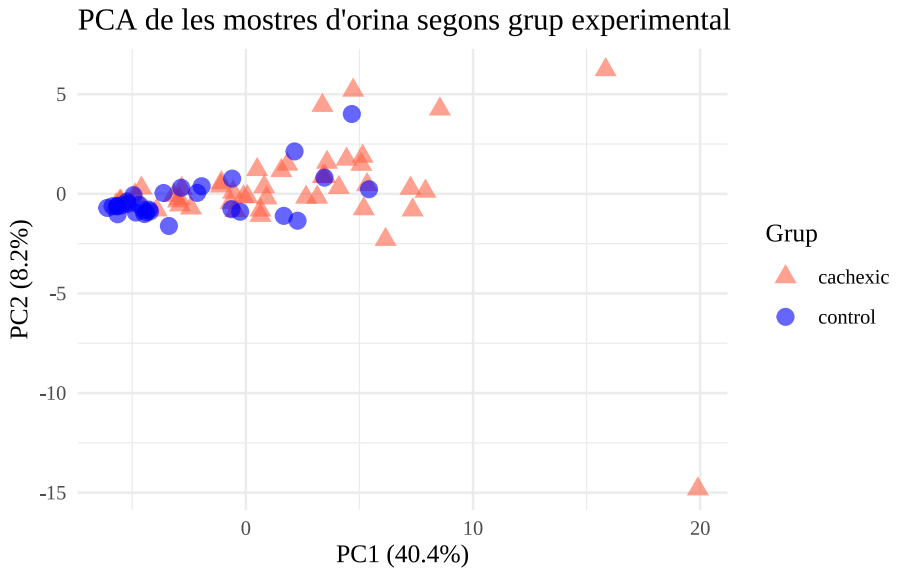

Figure 2: Diagrama de dispersió que mostra la projecció en 2 dimensions del dataset cachexia transformat via anàlisi de components principals (PCA). Cada eix inclou entre parèntesi la variància explicada per la corresponent Component Principal. Els pacients caquèxics apareixen com a triangles vermells,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

5 Discussió

6 Conclusions

7 Referències