PAC 2 Anàlisi de dades òmiques - informe: Resposta Transcriptòmica a Infeccions: Anàlisi d'RNA-seq per a la Detecció de Gens Diferencialment Expressats

Autor: Jordi Lladó Valero

Maig 2025

Contents

1. Abstract/Resum

Aquest estudi té com a objectiu caracteritzar la resposta transcriptòmica de pacients amb infecció per COVID-19 o pneumònia bacteriana, comparada amb individus sans, mitjançant dades d'RNA-seq de sang perifèrica. A partir de mostres seleccionades de l'estudi GSE161731, s'ha identificat un conjunt de gens diferencialment expressats que reflecteixen l'activació de mecanismes immunològics específics segons el tipus d'infecció. L'anàlisi exploratòria ha revelat patrons d'expressió diferenciats entre cohorts, especialment en el cas de les infeccions bacterianes, amb una resposta transcriptòmica més marcada. També s'ha detectat una influència de l'edat sobre el perfil d'expressió, que pot actuar com a factor confusor. L'anàlisi d'expressió

diferencial, basada en el mètode edgeR i ajustada per edat i sexe, ha permès identificar gens amb canvis d'expressió significatius, els quals s'han representat mitjançant diagrames de Venn i gràfics UpSet.

2. Objectius

L'objectiu principal d'aquesta activitat és realitzar una anàlisi completa d'expressió gènica diferencial a partir de dades RNA-seq, amb l'objectiu de resoldre preguntes biològiques relacionades amb la resposta transcriptòmica a diferents infeccions.

Els objectius específics són:

- Analitzar dades d'RNA-seq de sang perifèrica per caracteritzar el perfil transcriptòmic de pacients amb COVID-19, infeccions bacterianes i individus sans.
- Detectar gens diferencialment expressats entre cohorts utilitzant metodologies bioinformàtiques i estadístiques robustes.
- Aplicar tècniques d'anàlisi exploratòria multivariant (PCA, clustering, etc.) per examinar patrons globals d'expressió i identificar mostres atípiques.
- Realitzar una anàlisi funcional per determinar processos biològics sobreexpressats associats a la resposta a la infecció per SARS-CoV-2.

3. Metodologia

Per dur a terme aquesta anàlisi d'expressió gènica diferencial basada en dades RNA-seq, es van seguir les següents etapes metodològiques utilitzant R i paquets de Bioconductor:

3.1. Origen i naturalesa de les dades

Les dades provenen de l'estudi GEO amb identificador GSE161731 (M. McClain et al. (2021)), que inclou mostres de sang perifèrica de pacients amb infeccions per COVID-19, pneumònia bacteriana, individus sans entre altres. Es van descarregar de forma manual al els diferents arxius "GSE161731_counts.csv.gz" (comptatge) i "GSE161731_key.csv.gz" (metadades) que contenen els comptatges i metadades de les mostres. I posteriorment es va fer la lectura dels arxius amb RStudio.

3.2 Preprocessat i construcció de l'objecte

Després de la lectura incial, les dades es van netejar per eliminar mostres duplicades i assegurar la consistència entre noms i formats dels dos arxius (comptatge i metadades). Les accions que es van realitzar van ser: Els noms de mostra de la matriu de comptatges es van netejar eliminant prefixos innecessaris com "X" i substituint punts (".") per guions ("-") per tal de coincidir amb el format de la variable rna_id de les metadades. Es van eliminar les columnes amb sufix "_batch2", que representaven duplicats de processament tècnic, i es va filtrar la matriu per incloure només les mostres que també apareixien a les metadades. La coherència entre els dos conjunts es va validar, confirmant que no quedaven mostres desenllaçades. A continuació, es va construir un objecte SummarizedExperiment (se) contenint els comptatges bruts, les metadades i les coordenades genòmiques dels gens, obtingudes mitjançant el paquet EnsDb.Hsapiens.v86.

3.3 Filtrat i normalització

A continuació, es van seleccionar les mostres que tinguessin de cohort "COVID-19", "Bacterial" i "healthy", que són el motiu principal de l'estudi. I es va aplicar una llavor per tal de fer una selecció de totes les mostres identificades. El patró a seguir d'aquesta llavor va ser:

myseed <- sum(utf8ToInt("nomcognom1cognom2"))

set.seed(myseed)

El resultant d'aquest conjunt es va crear un nou objecte SummarizedExperiment (se_75). A partir d'aquí, es van eliminar els gens amb baixa expressió (menys d'1 CPM en almenys 3 mostres), i posteriorment es van calcular factors de normalització mitjançant la metodologia TMM (calcNormFactors). La matriu normalitzada es va transformar en log2(CPM) i es va afegir com a nou assay dins de l'objecte (se_75_edger).

3.4 Anàlisi exploratòria i detecció d'outliers

Un cop obtingut l'objecte final de treball, es van aplicar tècniques de reducció de dimensions (PCA i MDS) i de visualitzacions (heatmaps i clustering jeràrquic) per tal d'examinar l'estructura de les dades. També es van detectar i eliminar mostres atípiques segons la distància euclidiana en el primer component principal, així com es va avaluar la possible influència de covariables com l'edat i el sexe.

3.5 Anàlisi d'expressió diferencial

Abans de seguir amb l'anàlisi de l'expressió diferencial, es va aplicar la llavor:

set.seed(myseed)

sample(c("edgeR", "voom+limma", "DESeq2"), mida = 1)

per tal de seleccionar un mètode d'expressió. A continuació es va construir una matriu de disseny que incloïa la variable d'interès (cohort) i covariables confusores. El disseny del model es va realitzar sobre els comptatges filtrats i normalitzats de 75 mostres. Es va construir un model estadístic que inclou com a variable principal la cohort clínica (Healthy, COVID-19, Bacterial), i es van incorporar les covariables edat i sexe per controlar possibles efectes confusors identificats prèviament. La cohort Healthy es va definir com a nivell de referència. Es van definir dos contrastos: Bacterial vs Healthy i COVID-19 vs Healthy. Per últim el model es va ajustar mitjançant regressió lineal generalitzada (GLM), i es van aplicar tests de raó de versemblança per obtenir els gens diferencialment expressats. Els criteris de significació es van considerar significatius els gens amb log2FC superiors o iguals a 1.5 i FDR < 0.05. Aquest llindar selecciona gens amb canvis d'expressió robustos i control del fals positiu.

3.6 Anàlisi funcional

Finalment, es va fer una anàlisi d'enriquiment funcional utilitzant el domini "Biological Process" del Gene Ontology. Els resultats es van visualitzar amb eines com REVIGO per reduir la redundància dels termes.

3.7 Generació de documents i dades

Abans d'iniciar amb els resultats cal indicar que s'ha creat un repositori Github els diferents arxius demanats per a la realització de la PAC2. Aquest repositori té el següent enllaç:

https://github.com/JFox83/Llado-Valero-Jordi-PAC2

4. Resultats

Per tal de mostrar els diferents resultats, es seguirà el mateix ordre que en els apartats exposats en la metodologia, iniciant amb un breu text referent a l'origen i naturalesa de les dades.

4.1. Origen i naturalesa de les dades

Les dades analitzades en aquest treball provenen de l'estudi públic GSE161731, disponible a la base de dades Gene Expression Omnibus (GEO). Aquest estudi investiga la resposta transcriptòmica de la sang perifèrica de pacients amb diferents tipus d'infeccions respiratòries i controls sans.

En l'exploració inicial dels dos fitxers principals s'ha observat que; GSE161731_counts.csv.gz: és una matriu de comptatges de 60.675 gens (files) per a 201 mostres (columnes), i GSE161731_key.csv.gz: és un fitxer

amb metadades corresponents a 195 mostres, i 9 variables per mostra: rna_id, subject_id, age, gender, race, cohort, time_since_onset, hospitalized i batch.

Pel que fa a la distribució inicial de les mostres per cohort, s'ha obtingut la següent taula:

```
##
## Bacterial CoV other  COVID-19   healthy Influenza
##        23        59        77        19        17
```

Aquestes dades mostren una distribució desigual entre les cohorts, amb un predomini de mostres de COVID-19 i coronavirus estacional.

Per últim en la inspecció inicial dels arxius s'observa diferències entre els noms de les mostres de la matriu de comptatges i els identificadors de mostra (rna_id) de les metadades. Per tant, és necessari fer un primer pretractament de les dades abans de crear l'objecte summarizedExperiment.

4.2 Preprocessat i construcció de l'objecte

Després de fer el corresponent pretractament de les dades i la creació de l'objecte SummarizedExperiment s'observa que la dimensió de l'objecte conté 57.602 gens i 195 mostres. Per altra banda l'objecte inclou un assay anomenat "counts", i l'accés a la informació genòmica mostra que cada gen està vinculat a una regió concreta del genoma de referència.

4.3 Filtrat i normalització

Per últim, abans de procedir amb l'anàlisi exploratori de les dades, s'ha fet el pretractament de les metadades, escollint aquelles mostres les quals tingui el cohort"COVID-19", "Bacterial", "healthy", i no tinguin identificadors duplicats. EL nombre de mostres úniques obtingudes ha estat de 88. D'aquestes es van seleccionar 75 de forma aleatòria (llavor) i es va fer un nou objecte per tal de procedir amb la normalització de les dades.

El nou objecte, denominat se_75, contenia 57.602 gens i 75 mostres. A continuació, es va procedir a l'etapa de filtrat per expressió baixa. Es va calcular l'expressió relativa dels gens mitjançant l'estadística CPM (counts per million), i es van conservar només els gens amb almenys 1 CPM en 3 o més mostres. Aquest criteri va reduir considerablement el nombre de gens, passant de 57.602 a 19.137.

Posteriorment, es va aplicar la normalització TMM (Trimmed Mean of M-values) per corregir diferències en la profunditat de seqüenciació entre mostres. Els valors de comptatge es van transformar a log2(CPM) amb un pseudocomptatge de 1, per estabilitzar la variància en l'anàlisi posterior. Aquestes transformacions es van integrar en un nou objecte SummarizedExperiment anomenat se_75_edger, el qual inclou:

Assays:

- counts: matriu de comptatges filtrats.
- logCPM: matriu d'expressió normalitzada.
- ColData: 75 mostres amb 9 variables tècniques.
- RowRanges: 19.137 gens amb informació genòmica.

4.4 Anàlisi exploratòria i detecció d'outliers

Es va calcular una anàlisi de components principals (PCA) utilitzant la matriu d'expressió log 2 (CPM) de les 75 mostres seleccionades. Aquesta tècnica permet identificar patrons globals de variabilitat i relacions entre mostres.

A la Figura 1 es mostren tres representacions del pla PC1-PC2, on les mostres s'han acolorit segons la seva cohort, sexe i edat respectivament. Els dos primers components (PC1 i PC2) expliquen un 22.7% i un 11.5% de la variància total.

La primera representació mostra una separació significativa de les mostres per cohort en l'eix PC1. Les mostres Bacterial es concentren principalment al costat dret del gràfic, mentre que les mostres COVID-19 i Healthy apareixen agrupades a l'esquerra, amb certa superposició entre elles. Aquest patró suggereix que els perfils transcriptòmics de pacients amb COVID-19 i individus sans presenten una major similitud respecte al perfil dels pacients amb infecció bacteriana. Tot i aquesta proximitat al PCA, això no implica necessàriament una relació directa entre estar sa i una major susceptibilitat a la COVID-19, sinó que podria reflectir diferències en la intensitat o naturalesa de la resposta immunitària activada en cadascuna de les condicions. La segona representació, amb mostres acolorides per sexe, no mostra una separació evident, indicant que aquesta variable no té un efecte dominant en els primers components.

La tercera representació del PCA mostra les mostres acolorides segons l'edat en una escala contínua. Es pot observar una gradació clara de colors al llarg de l'eix PC1: les mostres amb valors més baixos de PC1 (extrem esquerre) corresponen majoritàriament a individus joves, mentre que les de valors alts (extrem dret) pertanyen a individus de més edat. Aquest patró coincideix amb la distribució de cohorts observada anteriorment, on les mostres Bacterial —situades a la dreta— corresponen també a pacients més grans, i les mostres COVID-19 i Healthy, a l'esquerra, corresponen a individus més joves. Aquesta coincidència suggereix que l'edat podria estar correlacionada amb el tipus de cohort, i per tant, pot actuar com a covariable confusora (pot tenir relació amb la cohort, amb la relació gènica i no és causada per la malaltia).

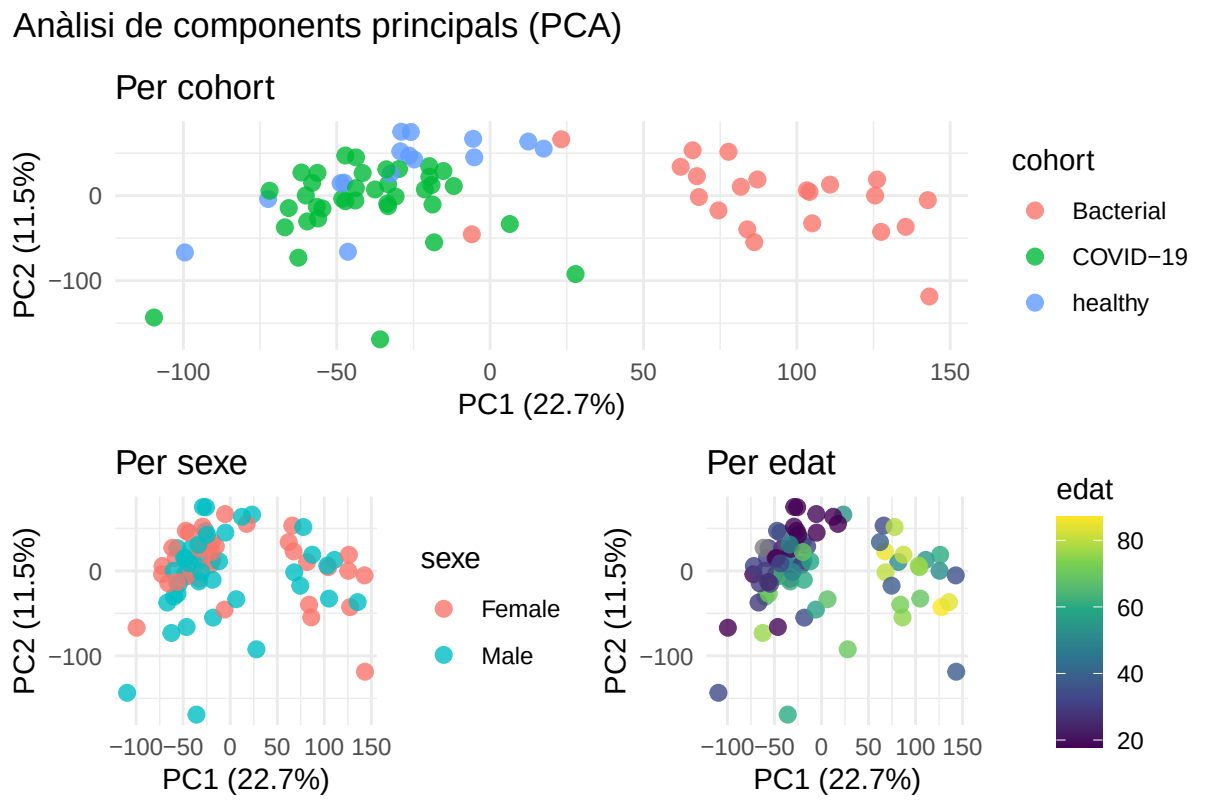


Figura 1: PCA aplicat a les dades de logCPM, acolorides per cohort, sexe i edat respectivament.

A part dels PCA, es va generar un gràfic MDS (annex) utilitzant el paquet limma, que mostra una representació alternativa de les distàncies entre mostres basada en la variació dels gens més diferencials. Els resultats confirmen les tendències observades al PCA, amb una agrupació parcial de les mostres per cohort.

També es va generar un heatmap amb els 500 gens més variables entre mostres (Figura 2), calculat segons la variància de l'expressió log2(CPM). La matriu d'expressió es va escalar per files (gens) per centrar els valors i fer comparables les diferències entre mostres. Les columnes del heatmap representen les 75 mostres, agrupades jeràrquicament segons similitud d'expressió, i es van annotar segons la cohort (COVID-19, Bacterial

o Healthy). Les files representen gens, també agrupats per similitud d'expressió.

Els colors del heatmap indiquen el nivell d'expressió relativa (Blau: expressió per sota de la mitjana (represió relativa), Vermell: expressió per sobre de la mitjana (sobreexpressió relativa)).

L'estructura del heatmap mostra clarament patrons diferenciats per cohort;les mostres del grup Bacterial presenten un patró distintiu, amb un grup de gens fortament sobreexpressats (en vermell intens) que no es detecten de la mateixa manera als altres grups. Les mostres COVID-19 i Healthy comparteixen més similitud, amb perfils d'expressió més homogenis, tot i que també es poden detectar grups de gens lleugerament diferencials. Aquests patrons visuals recolzen els resultats del PCA, i suggereixen que certes vies transcriptòmiques s'activen específicament en resposta a infeccions bacterianes.

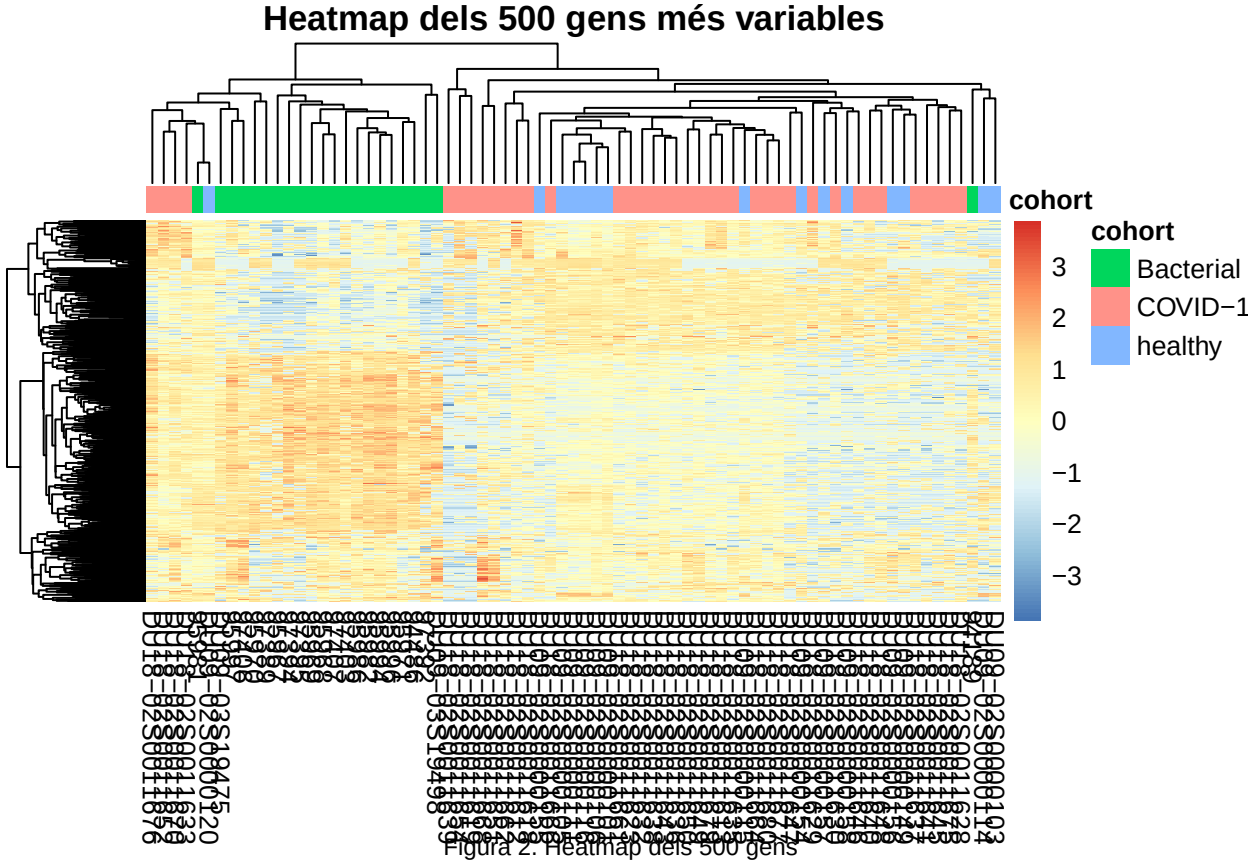


Figura 2. Heatmap dels 500 gens

Per últim, a partir dels resultats del PCA i del MDS, es va realitzar un anàlisi per identificar mostres potencialment atípiques. Es va calcular la distància euclidiana de cada mostra al valor mitjà del primer component principal (PC1). Les mostres amb una distància superior a 3 desviacions estàndard respecte a la mitjana es van considerar outliers transcriptòmics.

Aquest criteri va permetre identificar vuit mostres outliers:

```
## [1] "95969"           "95982"           "DU18-02S0011639" "95967"
## [5] "97394"           "95994"           "95996"           "97395"
```

Aquestes mostres es troben separades visualment de la resta al PCA i MDS, i podrien reflectir soroll tècnic, errors de processament o condicions biològiques atípiques. D'altra banda, les anàlisis exploratòries mostren que l'edat presenta una associació clara tant amb la cohort com amb la variabilitat transcriptòmica global, tal com es va evidenciar en la tercera visualització del PCA. Aquesta coincidència indica que l'edat pot actuar com a variable confusora, interferint en la detecció de diferències atribuïbles realment a la condició clínica.

4.5 Anàlisi d'expressió diferencial

Per identificar gens diferencialment expressats entre les diferents cohorts, es va seleccionar aleatòriament el mètode edgeR mitjançant una llavor generada a partir del nom complet de l'estudiant, per garantir la reproductibilitat del procés.

Els resultats obtinguts de les difernets comparacions indiquen que en la comparació Bacterial vs Healthy es van identificar diversos gens diferencialment expressats, destacant alguns amb valors molt elevats de log 2 FC (>10), com ENSG00000277452 i ENSG00000133063, indicant una marcada sobreexpressió en el grup Bacterial. Per altra banda la comparació COVID-19 vs Healthy s'hi van detectar també gens significatius, tot i que amb un nombre lleugerament inferior. Gens com ENSG00000183426 i ENSG00000171631 van mostrar valors de log2FC superiors a 8, indicant canvis importants en l'expressió gènica en resposta a la infecció per SARS-CoV-2.

Per analitzar la coincidència i especificitat dels gens diferencialment expressats entre les comparacions Bacterial vs Healthy i COVID-19 vs Healthy, s'han utilitzat dues visualitzacions complementàries: un diagrama de Venn (Figura 3) i un gràfic UpSet (Figura annex 2).

El diagrama de Venn mostra un total de 135 gens diferencials, distribuïts de la següent manera: 118 gens exclusius del contrast Bacterial vs Healthy, i 17 gens compartits entre Bacterial i COVID-19. Aquesta distribució indica que la resposta transcriptòmica associada a infecció bacteriana és molt més marcada que la de la infecció per SARS-CoV-2, almenys en les condicions i criteris de filtratge aplicats. Els 118 gens exclusius de la condició bacteriana podrien reflectir una activació immune més forta o sistèmica, consistent amb el tipus de resposta esperada en infeccions bacterianes greus.

Per altra banda, els 17 gens compartits podrien representar vies d'activació immunitària comunes, com la resposta inflamatòria inespecífica o la senyalització de citoquines, que s'activen tant en infeccions virals com bacterianes. Aquesta coincidència suggereix l'existència d'un nucli compartit de resposta immune, però també posa de manifest la distinció transcriptòmica entre els dos tipus d'infecció.

El gràfic UpSet confirma aquesta tendència, mostrant que la gran majoria dels gens DE són exclusius del contrast bacterià, i només una petita part són compartits (informació obtinguda similar).

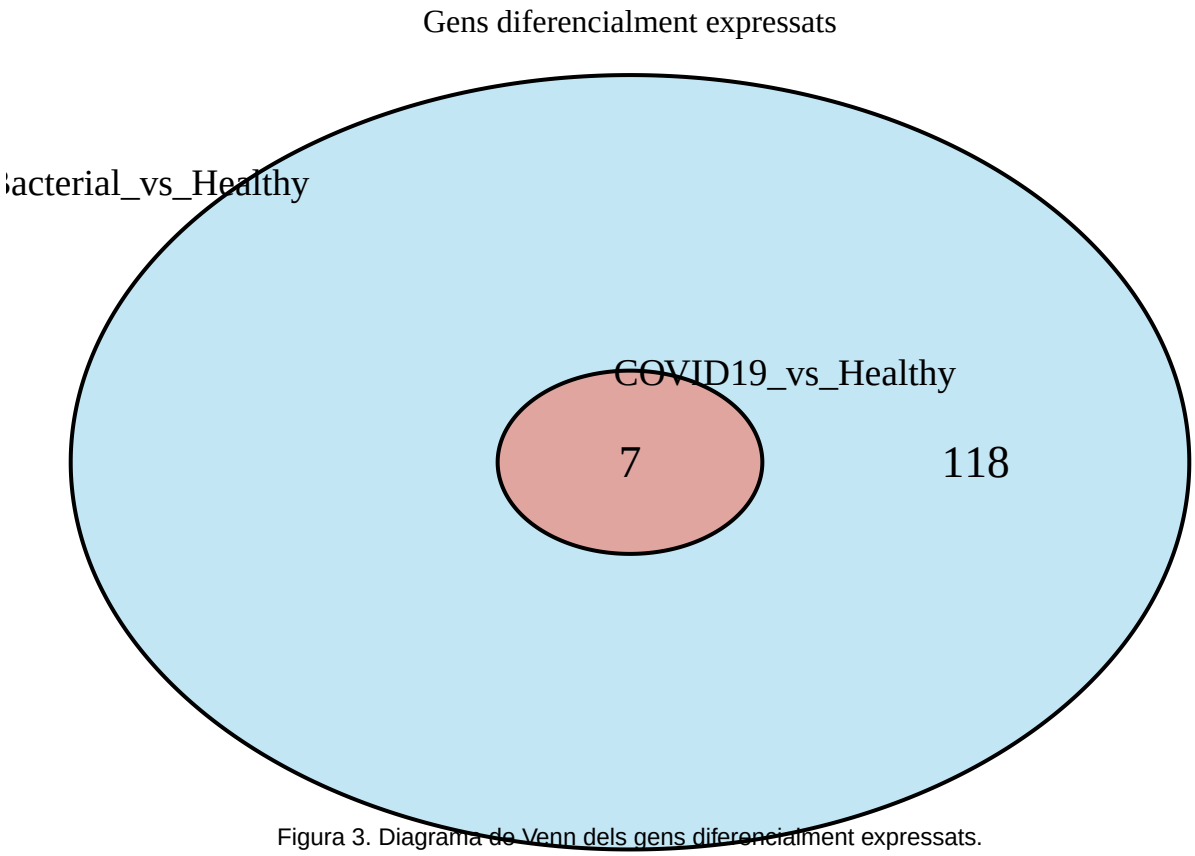


Figura 3. Diagrama de Venn dels gens diferencialment expressats.

5. Discussió

L'estudi dut a terme ha permès caracteritzar patrons d'expressió gènica diferencial entre individus sans i pacients amb infeccions per COVID-19 o pneumònia bacteriana, utilitzant dades d'RNA-seq de sang perifèrica. Tot i els resultats obtinguts, cal considerar diverses limitacions que afecten la interpretació dels resultats.

En primer lloc, el nombre limitat de mostres analitzades (75, després de fer filtres) pot reduir la potència estadística per detectar canvis subtils d'expressió, especialment quan es controlen covariables com l'edat i el sexe. A més, la selecció aleatòria de mostres introdueix variabilitat en els resultats obtinguts, i podria excloure mostres biològicament rellevants si no es té en compte l'equilibri entre cohorts.

Una altra limitació important és la heterogeneïtat clínica i tècnica de les dades, ja que les mostres provenen de diferents lots (batchs) i poden contenir informació incompleta o variables no mesurades (com comorbiditats o tractaments). Això pot introduir soroll i biaixos que són difícils de controlar amb les covariables disponibles.

Des del punt de vista biològic, la coincidència entre edat i cohort pot plantejar que moltes de les diferències observades podrien estar influïdes per l'edat dels pacients, especialment en la cohort bacteriana, que inclou individus més grans. L'ús d'un model estadístic ajustat per edat i sexe pot mitigar parcialment aquest problema, però no es pot eliminar completament l'efecte de variables confusores.

També cal destacar que l'anàlisi s'ha limitat a la identificació de gens diferencialment expressats, ja que no s'ha tingut temps a realitzar l'anàlisi funcional i d'enriquiment.

Finalment, l'activitat ha permès posar en pràctica un flux de treball bioinformàtic complet, des de la preparació de dades fins a la interpretació dels resultats. Malgrat les limitacions, el procés ha estat útil per comprendre la complexitat de l'anàlisi transcriptòmica i la importància d'un disseny experimental i estadístic rigorós.

6. Conclusions

LEs conclusions de l'activitat són:

- S'ha realitzat una anàlisi d'expressió gènica diferencial a partir de dades d'RNA-seq de sang perifèrica, centrada en la comparació entre individus sans i pacients amb COVID-19 o pneumònia bacteriana.
- L'anàlisi exploratòria (PCA, MDS i heatmap) ha revelat patrons d'expressió diferenciats entre cohorts, amb una segregació més clara del grup bacterià. També s'ha detectat un efecte important de l'edat, que ha estat considerat com a covariable en el model.
- El model estadístic utilitzat (edgeR) ha permès identificar un conjunt de gens diferencialment expressats amb canvis robustos ($\log2FC > 1.5$, $FDR < 0.05$), alguns dels quals són exclusius de cada infecció i d'altres compartits, suggerint respostes transcriptòmiques específiques i comunes.

7. Referències i enllaços

McClain M, Constantine F, Liu Y, Burke T M. McClain, F. COnstantine, Y. Liu, T. Burke.(2021) Dysregulated transcriptional responses to SARS-CoV-2 in the periphery support novel diagnostic approaches. Nat Commun 2021 Feb 17;12(1):1079. PMID: 33597532

M. McClain, F. Constantine, Y. Liu, T. Burke.(2021) Dysregulated transcriptional responses to SARS-CoV-2 in the periphery support novel diagnostic approaches. 'Dades' [Darrer accés: maig 2025] URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161731

Wang L, Balmat TJ, Antonia AL, Constantine FJ et al. An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility. Genome Med 2021 May 17;13(1):83. PMID: 34001247

rdrr.io (2021) cpm: Counts per Million or Reads per Kilobase per Million In edgeR: Empirical Analysis of Digital Gene Expression Data in R [darrer accés maig 2025] URL: https://rdrr.io/bioc/edgeR/man/cpm.html

ISCB Student Council (2025) RNA-seq: filtering, quality control and visualisation [darrer accés maig 2025] URL: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://combine-australia.github.io/RNAseq-R/slides/RNASeq_filtering_qc.pdf

QIAGEN (2013-25) TMM Normalization (Manual) [Darrer accés maig 2025] URL: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/1100/index.php?manual=TMM_Normalization.html

J.Llado (2025) PAC2 repositori Github [creació maig 2025] URL: https://github.com/JFox83/Llado-Valero-Jordi-PAC2

8. Annexes

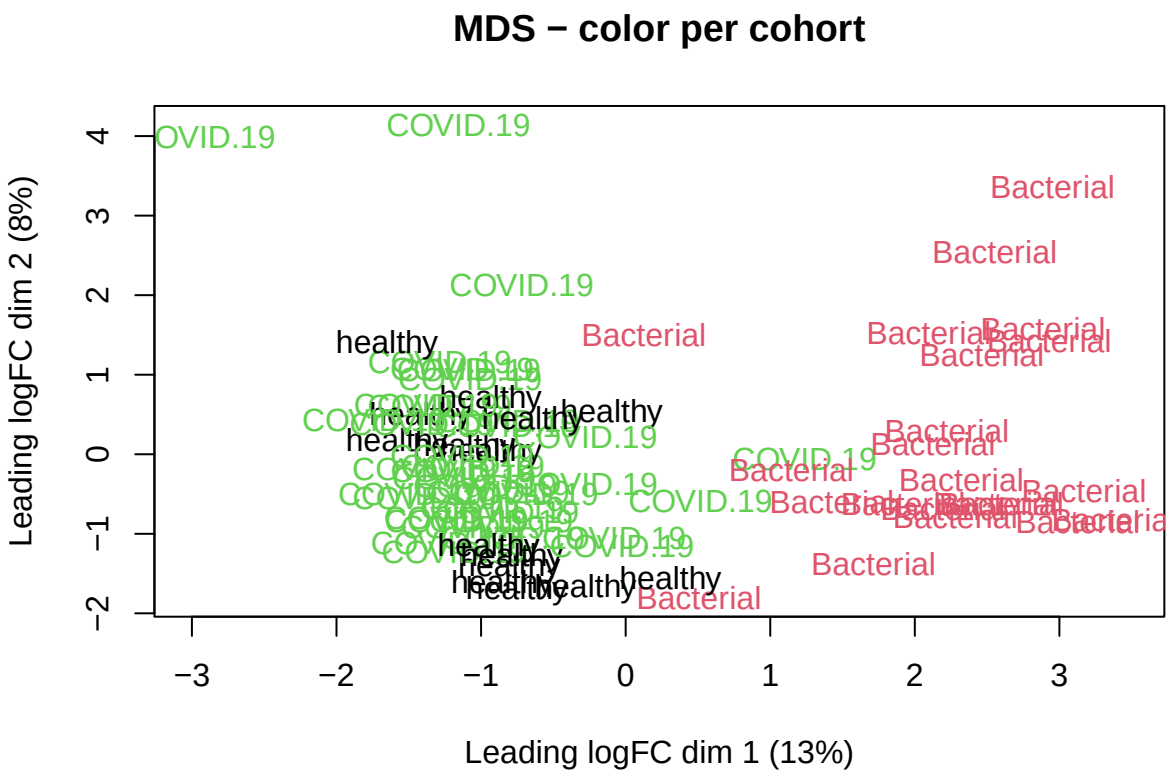
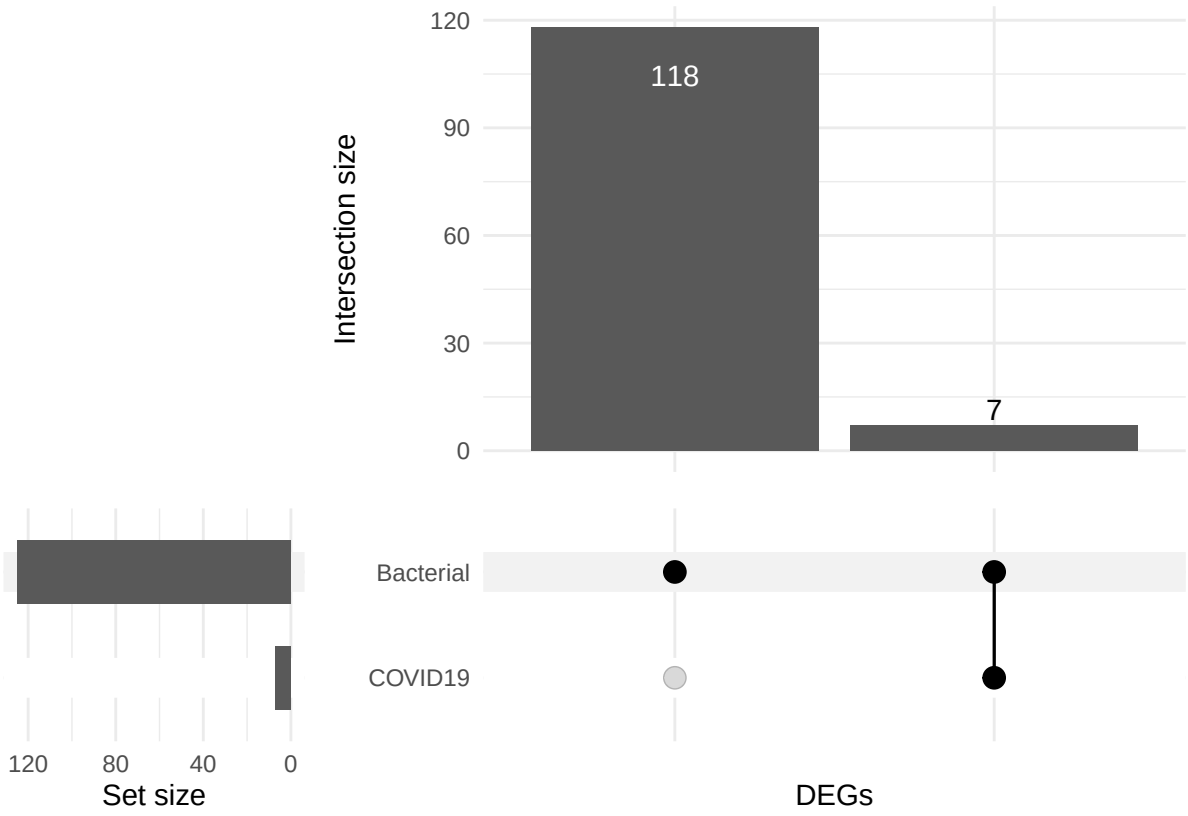


Figura Annex 1. Gràfic MDS: cada punt és una mostra.

```
##
## Attaching package: 'UpSetR'

## The following object is masked from 'package:ComplexUpset':
##
##     upset
```


120
90
60
30
0
Intersection size
118
7
Bacterial
COVID19
120
80
40
0
Set size
DEGs