PAC 2

Mar Torné Farré 2025-05-05

Índex

Resum

Estudi d'anàlisi d'expressió gènica diferencial en R/Bioconductor a partir del dataset GSE161731

Iniciem l'estudi fent una revisió de l'article **Dysregulated transcriptional responses to SARS-CoV-2 in the periphery support novel diagnostic approaches**, identificat amb el codi GSE161731 a la pàgina de **Gene Expression Omnibus** (https://www.ncbi.nlm.nih.gov/geo/).

Amb l'objectiu de descobrir nous aspectes de la resposta de l'hoste al SARS-CoV-2, l'estudi va dur a terme la seqüenciació d'ARN en 77 mostres de sang perifèrica de 46 subjectes amb COVID-19 i les va comparar amb subjectes amb coronavirus estacional, grip, pneumònia bacteriana i controls sans. El transcriptoma de la sang perifèrica revela aspectes únics de la resposta immunitària a la COVID-19 i proporciona nous enfocaments de diagnòstic basats en biomarcadors.

Per definir la resposta transcripcional de la sang perifèrica en subjectes amb infecció per SARS-CoV-2, es va dur a terme la seqüenciació d'ARN en mostres de 46 individus simptomàtics amb PCR positiva, 14 dels quals van ser mostrejar en múltiples moments. Els subjectes es van inscriure quan es van presentar per a atenció clínica i es va registrar el temps transcorregut des de l'aparició dels símptomes per a cada mostra individual recollida (rang 1-35 dies). Els subjectes amb COVID-19 es van classificar segons la gravetat de la malaltia i el temps des de l'inici dels símptomes. Com a grups comparadors, es van analitzar mostres de sang emmagatzemades de pacients que es van presentar al Servei d'Urgències amb infecció respiratòria aguda (IRA) a causa de coronavirus estacional (n=59), grip (n=17) o pneumònia bacteriana (n=20), així com de controls sans aparellats (n=19).

Objectiu

L'objectiu d'aquest treball és investigar els canvis en l'expressió gènica associats a diferents infeccions respiratòries mitjançant una anàlisi completa d'expressió gènica diferencial a partir de dades d'ARN-seq, per tal de respondre preguntes biològiques relacionades amb el perfil transcriptòmic de la resposta immunitària.

En primer lloc, realitzarem un anàlisi exploratori que ens permeti obtenir una visió general del conjunt de dades. Per fer-ho, crearem un objecte que contingui les dades i les metadades de l'estudi, i aplicarem processos de depuració, filtratge i revisió. Tot seguit, abordarem l'anàlisi diferencial pròpiament dita.

Mètodes

L'estudi està disponible a **Gene Expression Omnibus** (https://www.ncbi.nlm.nih.gov/geo/) amb l'identificador GSE161731.

L'objecte **SummarizedExperiment** que utilitzarem té les següents característiques:

- Files: Cada fila correspon a un gen, representant el nivell d'expressió d'aquest gen en les diferents mostres.
- Columnes: Cada columna representa una mostra biològica, corresponent a un pacient o control.
- Metadades: Les metadades associades a les mostres inclouen informació sobre el cohort (COVID-19, Bacterial o Healthy), així com dades clíniques i demogràfiques com el gènere, raça, temps des de l'inici dels símptomes, hospitalització i el batch experimental.

Per al processament i l'anàlisi de les dades, utilitzarem **Bioconductor**, una plataforma R especialitzada en l'anàlisi de dades genòmiques i transcriptòmiques d'alt rendiment. Per a la detecció de gens diferencialment expressats aplicarem els mètodes **DESeq2** i **limma-voom**, mentre que per a l'anàlisi funcional utilitzarem **clusterProfiler** emprant com a base d'anotació **org.Hs.eg.db** i l'ontologia de **Gene Ontology - Biological Process**. La visualització de resultats es durà a terme mitjançant els paquets **ggplot2**, **VennDiagram**, **UpSetR** i **heatmap**. A més, es duran a terme anàlisis de sobrerepresentació funcional sobre els gens diferencialment expressats, emprant l'ontologia de **Gene Ontology - Biological Process**.

Resultats

Descàrrega del dataset GSE161731

A la pàgina de **GEO** es pot trobar diversa informació sobre l'estudi. A l'apartat **Downloas RNA-seq counts** es poden descarregar els següents fitxers:

Fitxers proporcionats pels autors:

• Series SOFT file (GSE161731_family.soft.gz): Fitxer en format **SOFT** (llegible) que conté les metadades de la sèrie, com ara informació sobre l'estudi, les mostres, els protocols utilitzats, els processos i anàlisis realitzats.

• Series MINiML file (GSE161731_family.xml.tgz): Fitxer en format **XML** (per processar) amb les mateixes metadades que el fitxer SOFT.

• Series Matrix file (GSE161731_series_matrix.txt.gz): Fitxer tabular de **GEO** amb l'expressió gènica normalitzada o els valors processats per mostra.

Fitxers suplementaris:

• GSE161731_counts.csv.gz: Fitxer de format **CSV** amb la matriu de comptatges bruts per gen i per mostra. Conté el nombre de lectures (reads) assignades a cada gen.

• GSE161731_counts_key.csv.gz: Fitxer **CSV** associat al fitxer de comptatges, que conté el diccionari de mostres amb 198 registres. Aquestes claus permeten identificar les mostres a les matrius de comptatges.

• GSE161731_key.csv.gz: Fitxer **CSV** amb un altre diccionari de mostres, amb 196 registres (claus úniques), també útils per identificar les mostres dins les matrius de dades.

• GSE161731_xpr_nlcpm.csv.gz: Expressions normalitzades en log2 counts per milió (logCPM).

• GSE161731_xpr_tpm_geo.txt.gz: Expressions en TPM (Transcripts Per Million).

Fitxers NCBI - Generats automàticament per GEO

• GSE161731_raw_counts_GRCh38.p13_NCBI.tsv.gz: Matriu de comptatges bruts per gen i per mostra, alineats contra el genoma de referència GRCh38.p13.

• GSE161731_norm_counts_FPKM_GRCh38.p13_NCBI.tsv.gz: Matriu de comptatges normalitzats en FPKM (Fragments Per Kilobase Million).

• GSE161731_norm_counts_TPM_GRCh38.p13_NCBI.tsv.gz: Matriu de comptatges normalitzats en TPM (Transcripts Per Million).

• Human.GRCh38.p13.annot.tsv.gz: Taula d'anotació de gens utilitzada per **GEO** per etiquetar les matrius de dades. Inclou informació com l'identificador i el nom del gen, les coordenades genòmiques, el tipus de gen …

Per fer l'anàlisi, s'han seleccionat els fitxers següents: **GSE161731_counts.csv.gz**, **GSE161731_counts_key.csv.gz** i **GSE161731_key.csv.gz**, ja que contenen la matriu de comptatges bruts i les claus d'identificació de les mostres, elements essencials per a l'anàlisi d'expressió gènica diferencial. Aquest fitxers s'han copiat al directori **dades** del **GitHUb**.

Instal·lació i càrrega de paquets

Per tal de garantir que tots els paquets necessaris estiguin disponibles, implementem una funció que comprova si cada paquet està instal·lat i, en cas contrari, el descarrega automàticament. A continuació, els carreguem per disposar de les funcions requerides per a l'anàlisi. També creem els directoris necessaris al repositori i descarreguem els fitxers, que emmagatzemen al repositori **GitHub**.

Construcció de l'objecte SummarizedExperiment

Un cop descarreguem la matriu d'expressió i les metadades de **GEO** i les carreguem a R, construïm un objecte **SummarizedExperiment** que contingui les dues. A més, hi afegim les coordenades gèniques com **rowRanges**, utilitzant les coordenades dels gens humans amb **gens(EnsDb.Hsapiens.v86)**.

Per poder construir l'objecte, eliminem tres mostres marcades amb el sufix **_batch2** a la variable **rna_id**, ja que no disposem de la seva informació al fitxer **counts_key**.

```
## [1] "DU09-02S0000150_batch2" "DU09-02S0000154_batch2" "DU09-02S0000158_batch2"
```

Neteja de les dades i selecció de mostres

Tot seguit, explorem i netegem les metadades. Seleccionem les cohorts **COVID19**, **Bacterial** i **healthy**. Eliminem les mostres d'un mateix individu, conservant només la primera entrada. Assignem el format corresponent a cadascuna de les variables: convertim **age** a variable numèrica, i les variables **gender**, **race**, **cohort**, **time_since_onset**, **hospitalized** i **batch** a factors.

A més, substituïm els caràcters espai en blanc (" "), guions ("-") i barres ("/") per guions baixos ("_") per uniformitzar els valors.

Finalment, seleccionem aleatòriament 75 mostres utilitzant una llavor amb un format específic per garantir la reproductibilitat.

Pre-processat inicial de les dades

Filtrat de gens amb baixa expressió

Abans de realitzar l'anàlisi d'expressió diferencial, fem un primer filtratge per eliminar aquells gens amb una expressió molt baixa en totes les mostres, ja que aporten poca informació i poden interferir en les aproximacions estadístiques que s'utilitzaran més endavant. A més, aquests gens penalitzen els ajustos per comparacions múltiples, reduint la potència estadística per detectar gens realment diferencialment expressats.

Per això, abans de filtrar, convertim els comptatges bruts a **Counts Per Million (CPM)**. Aquesta transformació normalitza els comptatges en funció de la profunditat de seqüenciació de cada mostra, evitant que els valors baixos siguin deguts a diferències en la grandària de les biblioteques. Un cop normalitzats els comptatges, retenim aquells gens que mostren una expressió superior a 0,5 CPM en almenys dues mostres, per garantir que només es considerin gens amb una expressió mínima suficient.

```
# Expresem el comptatges en CPM i els seleccionem segons criteri
cpm_matriu<-cpm(assay(SE_GSE161731))
gens_guardats<-rowSums(cpm_matriu>0.5)>=2 # Seleccionem els gens amb cpm>0.5 en almenys 2 mostres
SE_GSE161731<-SE_GSE161731[gens_guardats,] # Apliquem el filtratge a l'objecte SummarizedExperiment
SE_GSE161731
```

```
## class: RangedSummarizedExperiment
## dim: 24128 75
## metadata(0):
## assays(1): counts
## rownames(24128): ENSG00000223972 ENSG00000227232 ... ENSG00000275063
##   ENSG00000271254
## rowData names(6): gene_id gene_name ... symbol entrezid
## colnames(75): 94189 DU18-02S0011619 ... DU09-02S0000156 DU09-02S0000153
## colData names(9): rna_id subject_id ... hospitalized batch
```

Normalització dels comptatges per profunditat de seqüenciació

Després de filtrar els gens amb baixa expressió, normalitzem els comptatges bruts per compensar les diferències en la profunditat de seqüenciació entre mostres. Aquest procés assegura que les diferències observades en l'expressió gènica no siguin atribuïbles a diferències en el nombre total de lectures entre mostres, sinó a canvis biològics reals. Per fer-ho, primer creem un objecte **DESeqDataSet** a partir de l'objecte **SummarizedExperiment**, especificant una fórmula de disseny neutra ~1, ja que en aquest punt només volem estimar els factors de mida de les biblioteques sense ajustar cap variable explicativa. A continuació, utilitzem la funció **estimateSizeFactors()** per calcular els factors de normalització per a cada mostra, que es basa en la mediana de les ràtios dels comptatges respecte a un pseudo-genoma de referència. Això permet corregir les diferències en profunditat de seqüenciació.

Finalment, apliquem una transformació de variància estabilitzada (VST) mitjançant la funció **vst()**. Aquesta transformació redueix la dependència de la variància respecte als comptatges mitjans i produeix dades més aproximades a una distribució normal, facilitant la visualització i anàlisi exploratòria.

La matriu de comptatges normalitzada la guardem com un nou assay dins l'objecte **SummarizedExperiment** i actualitzem la llista d'assays disponibles per verificar que s'ha afegit correctament.

```
# Normalitzem els comptatges bruts per les diferents profunditats de secuenciació per cada mostra
dds<-DESeqDataSet(SE_GSE161731,design=~1)
dds<-estimateSizeFactors(dds) # Estimem els factors de normalització
vst<-vst(dds) # Apliquem una transformació de variància estabilitzada
assay(SE_GSE161731,"vst_counts")<-assay(vst) # Afegim la matriu normalitzada com un nou assay
assayNames(SE_GSE161731)
```

```
## [1] "counts"     "vst_counts"
```

Control de qualitat de les dades normalitzades

Un cop pre-processades i normalitzades les dades d'expressió, realitzem un control de qualitat. Per fer-ho, utilitzem la funció **arrayQualityMetrics()**, que genera un informe interactiu amb diverses mètriques de qualitat. Creem un objecte **ExpressionSet**, que conté la matriu de comptatges normalitzats (VST) i les metadades associades a cada mostra. Aquest format és compatible amb la funció **arrayQualityMetrics()** del paquet homònim.

L'informe de qualitat generat amb **arrayQualityMetrics** proporciona diverses visualitzacions per avaluar l'estat de les mostres després de la normalització:

Figura 1. Distances between arrays. Mostra les distàncies absolutes mitjanes (distància L1) entre mostres. Es detecten dues mostres atípiques amb una suma de distàncies elevades respecte a la resta, que podrien considerar-se potencialment problemàtiques.

Figura 2. Outlier detection for Distances between arrays. Mostra un gràfic de barres amb els valors Sa de cada mostra. S'estableix un llindar de 69,1, i dues mostres el superen, confirmant el resultat de la Figura 1.

Figura 3. Principal Component Analysis. La majoria de mostres s'agrupen correctament, però es detecta un valor aïllat que podria correspondre a una de les mostres atípiques identificades anteriorment.

Figura 4. Boxplots. Les caixes són, en general similars, però s'identifica una mostra amb una distribució anòmala que podria indicar problemes tècnics.

Figura 5. Outlier detection for Boxplots. Complementa el boxplot, mostrant les mostres amb distribucions significativament diferents respecte a la població global. Es marca un llindar (0,222) i es confirmen mostres candidates a excloure.

Figura 6. Density plots. Permet observar la forma de la distribució per mostra. La majoria de mostres tenen formes semblants, però alguna presenta desviacions, com un pic més aixecat o cua anòmala.

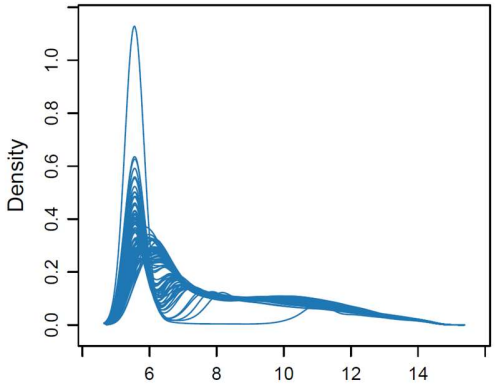
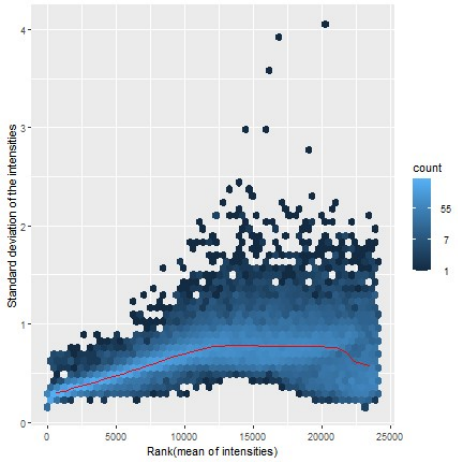


Figura 7. Standard deviation versus rank of the mean. Mostra la relació entre la mitjana i la desviació estàndard d'intensitat. La línia vermella (mediana mòbil) es manté estable en general, tot i que s'observa lleuger augment en alguna mostra cap a valors alts, suggerint una possible saturació o dispersió excessiva.

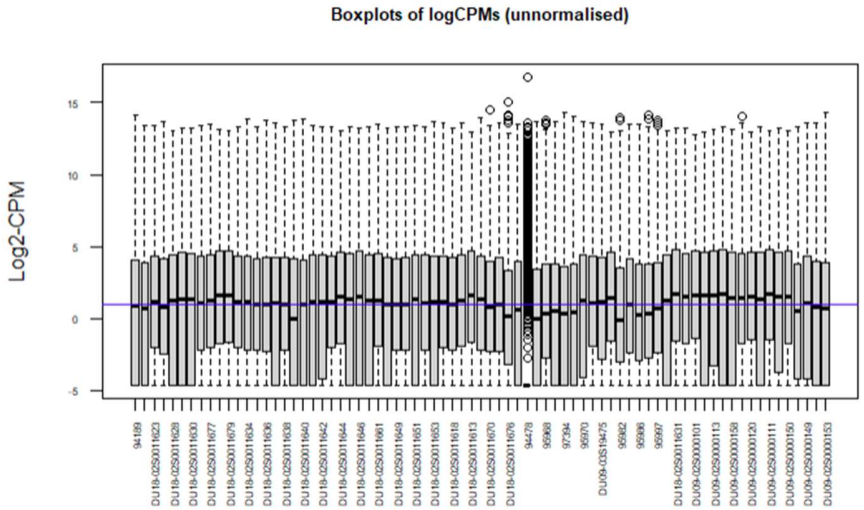


Anàlisi exploratòria de les dades transformades/normalitzades

Identificació i eliminació de mostres atípiques

Per tal d'identificar possibles mostres atípiques abans de l'anàlisi diferencial, representem boxplots dels valors de log2-CPM dels comptatges no normalitzats. Aquesta visualització ens permet detectar mostres amb distribucions anòmales. Per fer-ho, creem un objecte **DGEList** amb els comptatges originals, calculem els Counts Per Million (CPM) en escala log2 i generem el boxplot per visualitzar la distribució global per mostra.

En aquest gràfic s'identifica una mostra clarament discrepant (mostra 94478), amb una mediana i una distribució de valors notablement diferents respecte a la resta. La mostra 94478 es confirma com a outlier i l'eliminem de l'anàlisi i de totes les matrius i objectes rellevants per garantir la coherència de les dades en els passos posteriors.

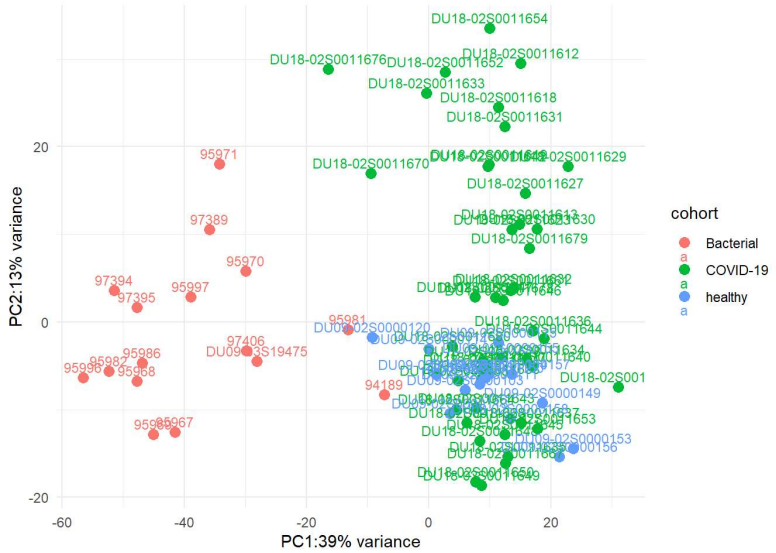


Anàlisi de components principals

Realitzem una **Anàlisi de Components Principals (PCA)** i un **clustering jeràrquic** basat en la distància euclidiana entre mostres. En primer lloc, calculem i representem la PCA a partir dels valors normalitzats amb VST, pintant les mostres segons la cohort i etiquetant-les pel seu nom. Aquest anàlisi ens permet identificar patrons d'agrupació associats a les cohorts i detectar possibles mostres aïllades.

Pel que fa al clustering jeràrquic, calculem la distància entre mostres a partir dels valors VST i representem un mapa de calor (heatmap) de la matriu de distàncies. Aquesta visualització permet confirmar les agrupacions observades a la PCA i detectar patrons coherents o possibles anomalies.

Per detectar mostres atípiques, calculem la distància de cada mostra al centroide de la PCA (mitjana de les coordenades de PC1 i PC2) i considerem com a possibles outliers aquelles mostres amb una distància superior a la mitjana + 2 desviacions estàndard. En aquest cas, s'han identificat algunes mostres fora del llindar, però no les eliminem perquè considerem que poden reflectir variabilitat biològica i perquè no s'han confirmat com a problemàtiques amb l'anàlisi prèvia de control de qualitat.

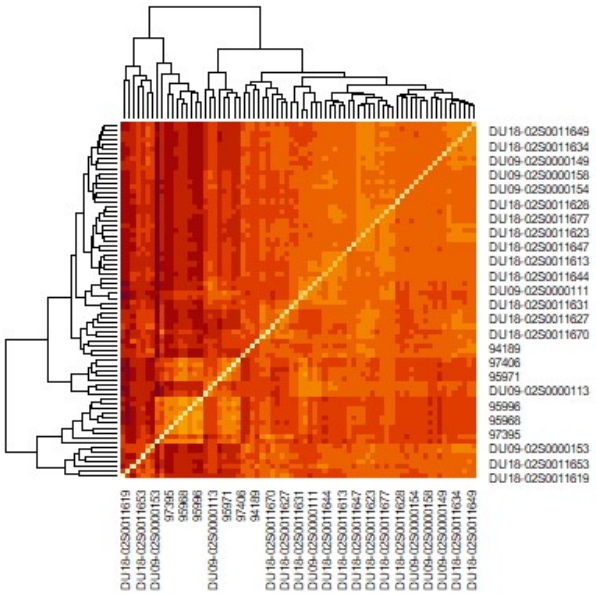


El gràfic de l'anàlisi de components principals (PCA) revela una clara separació de les mostres segons la seva cohort, mostrant com la variabilitat en els dos primers components principals (PC1 i PC2) reflecteix diferències biològiques rellevants.

L'eix PC1, que explica el 39% de la variància total, sembla capturar principalment la separació entre les mostres de la cohort **Bacterial** i la resta. Les mostres bacterianes es localitzen majoritàriament a la part esquerra de l'eix PC1, mentre que les mostres **COVID-19** i **healthy** es troben a la part dreta, amb les mostres sanes formant un grup més compacte. Aquesta separació suggereix que PC1 reflecteix canvis globals en l'expressió gènica associats específicament a la presència d'infecció bacteriana, diferenciant-les clarament de les altres mostres.

En canvi, l'eix PC2, que explica un 13% addicional de la variància, sembla captar la variabilitat dins de les condicions patològiques, especialment en la cohort **COVID-19**. Les mostres d'aquesta cohort s'estenen al llarg de l'eix PC2, indicant diferències internes probablement relacionades amb factors clínics com la severitat de la malaltia, el temps des de l'inici dels símptomes o la resposta immune individual.

Pel que fa a la cohort **healthy**, les mostres es mantenen agrupades a la part dreta de PC1 i presenten poca variabilitat en PC2, indicant perfils transcriptòmics més homoginis, característics d'estats fisiològics estables.



El heatmap amb clustering jeràrquic complementa aquests resultats mostrant agrupacions coherents de mostres amb perfils d'expressió similars. Els blocs de colors clars indiquen clústers de mostres altament correlacionades, com les mostres DU09 (**healthy**) i DU18 (**COVID-19**), que formen grups ben definits, mentre que les mostres bacterianes apareixen més disperses, cosa que podria reflectir una heterogeneïtat més gran deguda a diferents agents infecciosos o respostes inflamatòries diverses.

Aquest patró global de separació en PCA i clustering reforça la idea que les signatures transcriptòmiques identificades són específiques per a cada condició, amb potencial per a la classificació i la interpretació biològica dels diferents estats de salut i malaltia.

Identificació de possibles variables confusores

Per detectar variables que podrien introduir biaixos en l'anàlisi d'expressió gènica diferencial, realitzem un anàlisi de components principals (PCA) sobre les dades transformades mitjançant la variància estabilitzada (vst). En cada cas, representem les mostres en funció dels dos primers components principals i assignem colors a les mostres segons les diferents categories de les variables: **gender**, **race**, **time_since_onset**, **hospitalized** i **batch**.

Addicionalment, explorem la distribució de les categories de cada variable respecte a les cohorts mitjançant taules de contingència, per identificar possibles desequilibris en la distribució de les mostres.

Un cop representem els gràfics, analitzem si alguna variable mostra agrupacions clares. Aquesta evidència indicaria que la variable té un efecte sistemàtic sobre els patrons d'expressió gènica, independentment de la

condició experimental principal (cohort). En aquest cas, és recomanable incloure-la com a covariable en la matriu de disseny per controlar aquest efecte i evitar biaixos. Si, en canvi, no s'observen agrupacions evidents, podem considerar que la variable no introdueix un efecte rellevant sobre l'expressió i, per tant, no la inclourem.

A més, si les taules de contingència mostren un desequilibri important entre cohorts i categories d'una variable, convé tenir-ho en compte, ja que aquesta variable podria actuar com a factor confusor i afectar la interpretació dels resultats. En aquests casos, també seria prudent incloure-la al model.

En concret, pel que fa a les variables analitzades:

- **gender**: les proporcions de sexe estan relativament equilibrades dins de cada cohort i, al PCA, no es detecten agrupacions clares. Per tant, no es considera necessari incloure-la com a covariable.
- **race**: presenta una associació clara amb la cohort. La majoria de mostres **bacterial** són **Black/African American**; les de **COVID-19** són principalment **White**, amb més diversitat; i les **healthy** són majoritàriament **White** i **Asian**. Aquesta associació, sumada als indicis del PCA, justifica la seva inclusió com a covariable.
- **time_since_onset**: només és aplicable als pacients **COVID-19**. per la qual cosa no es pot incloure com a covariable global.
- **hospitalized**: també només s'informa per **COVID-19**, amb majoria de casos "No", i per tant, no és aplicable globalment.
- **batch**: Batch 1 conté gairebé totes les mostres **Bacterial** i **Healthy**, i la major part de les **COVID-19**; Batch 2 inclou algunes mostres **COVID-19** i poques de les altres cohorts. Tot i que no s'observen agrupacions clares al PCA, hi ha una associació notable entre cohort i batch, fet que podria introduir biaixos per l'efecte de lot. Per precaució, la inclourem com a covariable al model.

Construcció de la matriu del disseny i ajust del model

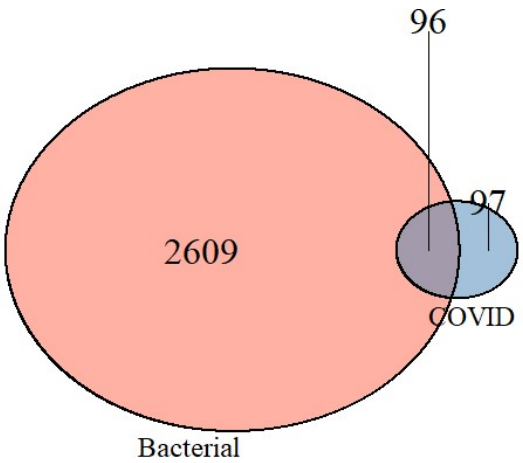
Per analitzar l'expressió gènica diferencial en les comparacions **Bacterial vs Healthy** i **COVID-19 vs Healthy**, definim un model de disseny que inclou la variable principal d'interès **cohort** i les variables **race** i **batch**, identificades prèviament com a potencials factors confusors. La inclusió d'aquestes variables permet controlar possibles efectes de composició poblacional que podrien interferir en la detecció de diferències genuïnes associades a la condició clínica.

Per dur a terme l'anàlisi, utilitzarem el paquet **DESeq2**, seleccionat aleatòriament entre: **edgeR**, **voom+limma** i **DESeq2**.

Finalment, establim uns llindars per definir la significació estadística i rellevància biològica: considerem com a diferencialment expressats els gens amb un valor ajustat de p (padj) inferior a 0,05 i un canvi en l'expressió absoluta (|log2 fold change|) superior a 1,5. Aquests criteris asseguren que les diferències identificades siguin, a més d'estadísticament significatives, biològicament rellevants.

Comparació dels resultats dels contrastos

Per avaluar el solapament de gens diferencialment expressats entre les dues comparacions (**Bacterial vs Healthy** i **COVID-19 vs Healthy**), representem un **diagrama de Venn** i un **UpSet plot** a partir dels gens significatius.



Obtenim els següents resultats:

- **Bacterial vs Healthy**: 2.705 gens diferencialment expressats.
- **COVID-19 vs Healthy**: 193 gens diferencialment expressats.
- Gens compartits entre ambdues condicions: 96 gens.

Observem que **Bacterial** provoca molts més canvis en l'expressió gènica respecte als controls sans (2.705 gens) que la infecció per **COVID-19** (193 gens). Dels 193 gens diferencialment expressats en **COVID-19**, 96 també ho són en la infecció bacteriana. Això implica que gairebé la meitat dels gens diferencials associats a **COVID-19** també responen a infeccions bacterianes, suggerint una possible resposta comuna a la infecció. Alhora, s'observa que 2.609 gens són específics de **Bacterial** i 97 són exclusius de **COVID-19**, reflectint les diferències biològiques entre ambdues condicions.

El gràfic UpSet reforça aquesta interpretació, mostrant visualment el nombre de gens exclusius de cada condició i els que són compartits.

Enriquiment funcional GO dels gens sobreexpressats en la infecció per COVID-19

Finalment, realitzem un anàlisi de sobrerepresentació per identificar les funcions biològiques enriquides entre els gens sobreexpressats en pacients amb **COVID-19** en comparació amb els controls sans. Els resultats que obtenim reflecteixen molt bé els processos activats en pacients infectats i en destaquem aquestes àrees funcionals:

- Resposta immunitària mediada per cèl·lules B i immunoglobulines GO:0016064 immunoglobulin mediated immune response GO:0019724 B cell mediated immunity

Aquests termes indiquen una activació notable de la resposta adaptativa, especialment a través de les cèl·lules B i la producció d'anticossos. És un resultat esperable en el context d'una infecció viral, on l'organisme intenta generar anticossos específics per neutralitzar el virus.

- Processos relacionats amb la divisió cel·lular i segregació cromosòmica GO:0140014 mitotic nuclear division GO:0000070 mitotic sister chromatid segregation GO:0007088 regulation of mitotic nuclear division GO:0007059 chromosome segregation GO:0000819 sister chromatid segregation GO:0098813 nuclear chromosome segregation

Aquest conjunt de termes apunta a una activació o desregulació dels processos de divisió cel·lular.

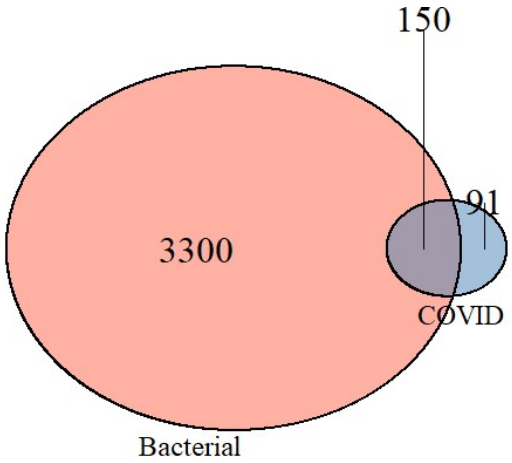
- Traducció citoplasmàtica GO:0002181 cytoplasmic translation

Que suggereix una activació de la síntesi proteica, possiblement associada a la producció de proteïnes antivirals, citocines i altres mediadors implicats en la resposta a la infecció.

En resum, aquesta anàlisi reflecteix tres grans àrees biològiques activades en la **COVID-19**, com són la resposta immunitària adaptativa, l'alteració de processos de divisió cel·lular i regulació de la síntesi proteica.

Replicació de l'anàlisi amb un mètode alternatiu

Repliquem l'anàlisi d'expressió gènica diferencial utilitzant un enfocament estadístic alternatiu: **voom+limma**.



Quan comparem els resultats d'expressió diferencial obtinguts amb **DESeq2** i **voom+limma**, ajustant per **race** i **batch**, observem que **voom+limma** detecta un nombre més elevat de gens diferencialment expressats en ambdós contrastos respecte a **DESeq2**.

- **Bacterial vs Healthy**: 3.450 gens (**voom+limma**) vs 2.705 gens (**DESeq2**)
- **COVID-19 vs Healthy**: 241 gens (**voom+limma**) vs 193 gens (**DESeq2**)

El nombre de gens comuns entre ambdues condicions també és lleugerament superior amb **voom+limma** (150 gens) que amb **DESeq2** (96 gens). Aquesta diferència és esperable, ja que **voom+limma** tendeix a ser més sensible en la detecció de canvis d'expressió, mentre que **DESeq2** és més conservador.

Pel que fa a l'anàlisi de sobre-representació destaquen dues grans categories funcionals:

- Divisió cel·lular i segregació cromosòmica.
- Traducció citoplasmàtica.

Les anàlisis de sobre-representació funcional amb **DESeq2** i **voom+limma** han coincidit en identificar processos relacionats amb la divisió cel·lular i la traducció citoplasmàtica com a diferencials entre les condicions estudiades. Addicionalment, l'anàlisi amb **DESeq2** ha posat de manifest la implicació de la resposta immunitària mediada per cèl·lules B i immunoglobulines, suggerint que podria tractar-se d'un procés rellevant però que no assoleix significació amb l'estratègia **voom+limma**.

Referències

ADO-2.3.4-Análisis de datos de RNA

ADO-2.3.5-Guión del ejemplo de análisis de datos de RNA-Seq con R

ADO-04-Análisis de Significación Biologica-2: Métodos de Análisis

https://aspteaching.github.io/An-Introduction-to-Pathway-Analysis-with-R-and-Bioconductor/

https://github.com/ASPteaching/Analisis_de_datos_omicos-Ejemplo_2-RNASeq

Analysis of the termogenic program in mice

L'enllaç al repositori de **GitHub** que conté tota la documentació del projecte és:
https://github.com/MarTorne/Torne-Farre-Mar-PEC2.git