PEC2_Análisis_datos_ómicos

Samuel Pis Vigil

2025-05-08

Resumen

En este Trabajo se realiza un análisis transcriptómico de la respuesta molecular humana al proceso infectivo por parte del virus del Covid-19 y se compara con la respuesta que se da en los procesos de neumonía bacteriana. Para ello se hace uso de un repositorio público (GSE161731, NCBI) que contiene los datos de secuenciación masiva de ARN de 77 muestras de sangre periférica de 46 pacientes con coronavirus y la comparación con datos de pacientes de una serie de otras enferemedades respiratorias y de controles sanos.

Objetivos

El principal objetivo de este Trabajo es la identificación de las rutas moleculares que se vean alteradas por la infección por Covid-19 en pacientes humanos.

Los objetivos secundarios de este Trabajo son la identificación del conjunto de genes cuya expresión se ve alterada en pacientes humanos de Covid-19 y de infecciones bacterianas, comprobar si existe un solapamiento entre ambos grupos de genes y comparar las rutas moleculares que están exacerbadas en infecciones de neumonías bacterianas y de infecciones por Covid-19

NOTA Hemos decidido organizar los contenidos de los apartados *Métodos* y *Resultados* de la siguiente manera. En *Métodos* se explican los datos de origen, los paquetes usados y el método estadístico de reducción dimensional usado. En *Resultados* se han ido dando, a parte de los resultados en sí mismos, explicaciones metodológicas en mayor detalle de algunos pasos. El motivo de incluir explicaciones metodológicas en el apartado *Resultados* radica en que consideramos que es más entendible explicar los análisis conforme se van ejecutando que ponerlo todo junto al principio.

Métodos

Los datos usados para este Trabajo están extraídos del repositorio público de datos *GSE161731* del *Gene Expression Omnibus* del NCBI, el cual contiene los datos de 77 muestras de RNAseq de sangre periférica de 46 pacientes con Covid-19, 59 muestras de pacientes de coronavirus estacional, 17 muestras de pacientes de influenza, 20 de pacientes de neumonía bacteriana y 19 controles sanos.

Para realizar el trabajo se han usado, a parte del conjunto de paquetes de R cargados por defecto en Rstudio, una serie de librerías más específicas. Se ha usado el paquete *SummarizedExperiment* para generar el objeto "SummarizedExperiment" que agrupa los datos y metadatos provenientes de los experimentos de RNAseq que engloba *GSE161731*. Para procesar y descomprimir los archivos descargados desde el NCBI se ha usado el paquete R.utils; para obtener los rangos genómicos de los transcritos se ha empleado el paquete *GenomicRanges*; para almacenar, interpretar y traducir los IDs de Ensembl asociados a las anotaciones de las lecturas y transcritos de ARN se ha empleado la librería *EnsDb.Hsapiens.v86*; para realizar el análisis de

expresión diferencial se han usado los paquetes *edgeR* y *limma*; finalmente, para la representación visual de los gráficos se ha empleado *ggplot2* (paquete estándar de graficación), *ComplexHeatmap* (para la realización de gráficos de mapas de calor), *circlize* (en este caso se ha usado por ser auxiliar a *ComplexHeatmap*) y nVennR (para la realización de diagramas de Venn con conjuntos cuasi-proporcionales). Cabe destacar que se ha empleado *nVennR* frente a otros paquetes más conocidos y extendidos de generación de diagramas de Venn (y más fáciles de descargar ya que se encuentran en CRAN y no es necesaria su instalación manual desde GitHub, como es el caso de *nVennR*) como *VennDiagram* debido a que es el único que realiza representaciones cuasi-proporcionales de los conjuntos según el tamaño de las listas que forman dichos conjuntos, lo que da una imagen mucho más intuitiva e informativa de la realidad del diagrama.

En cuanto a la metodología estadística empleada, cabe destacar que se ha empleado el método de Voom-Limma (se ha escogido siguiendo la directriz de pseudoaleaterización marcado en el enunciado de la PEC) para la realización del análisis de expresión diferencial. Este método realiza una transformación de Voom en primer lugar para poder, posteriormente, realizar un ajuste a un modelo lineal que permite extraer datos de FoldChange (indica en términos relativos cuánto varía la expresión de un gen) y de p-valor crudo y ajustado asociado a cada comparación de RNAseqs.

Resultados

Generación del *SummarizedExperiment*

Tras descargar los datos desde el NCBI y generar el objeto *SummarizedExperiment* (al que hemos llamad0 *se*) se comprueba si se ha creado correctamente.

```
# Comprobación de que se ha hecho bien el se
dim(se)
```

```
## [1] 57602   195
```

```
identical(rownames(se), names(rowRanges(se)))
```

```
## [1] TRUE
```

```
identical(colnames(se), rownames(colData(se)))
```

```
## [1] TRUE
```

Se verifica que nombres de las muestras y los genes son iguales y están alineados entre los datos y los metadatos (si no da errores *a posteriori*). Por otra parte, se tiene que en los datos hay 57.602 genes y 195 muestras; el número de muestras coincide con lo expresado en la descripción del dataset en GEO, y el número de genes está en el rango más o menos normal para un conjunto de datos de RNAseq.

Limpieza de los datos y selección pseudoaleatoria de muestras

Para la selección pseudoaleatoria del conjunto de muestras seguimos el plante de semilla indicado en el enunciado de la PEC:

```
myseed <- sum(utf8ToInt("samuelpisvigil"))
set.seed(myseed)
```

Para comprbar si el nuevo objeto se ha generado correctamente, se comprueba en un vistazo que los datos se correspondan solo a cohortes de "COVID_19", "healthy" y "Bacterial". También se comprueba que ninguna columna se haya corrompido en el proceso intermedio y que el número de genes siga siendo 57.602 y el número de muestras 75:

```
# Comprobación de que "se75" es correcto
colData(se75)
```

```
## DataFrame with 75 rows and 8 columns
##                   subject_id       age      gender                  race
##                  <character> <numeric> <character>           <character>
## DU18-02S0011612      936128        64        Male                 White
## DU09-02S0000149      6AA2B5        18      Female                 Asian
## DU18-02S0011641      6BDA69        29      Female                 White
## 95967                9C7138        38      Female Black/African American
## DU18-02S0011629      B85D75        31      Female                 White
## ...                     ...       ...         ...                   ...
## DU18-02S0011676      12B6DB        70        Male Black/African American
## DU09-03S19498        AB87B2        72      Female Black/African American
## DU18-02S0011679      03ADC4        32      Female Black/African American
## DU18-02S0011623      0B943B        33        Male                 White
## DU18-02S0011677      41EE2E        30      Female                 Asian
##                      cohort time_since_onset hospitalized     batch
##                 <character>      <character>  <character> <integer>
## DU18-02S0011612    COVID-19           middle          Yes         1
## DU09-02S0000149     healthy               NA           NA         1
## DU18-02S0011641    COVID-19            early           No         1
## 95967              Bacterial               NA           NA         1
## DU18-02S0011629    COVID-19            early           No         1
## ...                     ...              ...          ...       ...
## DU18-02S0011676    COVID-19            early          Yes         1
## DU09-03S19498     Bacterial               NA           NA         1
## DU18-02S0011679    COVID-19           middle           No         1
## DU18-02S0011623    COVID-19            early           No         2
## DU18-02S0011677    COVID-19           middle           No         1
```

```
dim(se75)
```

```
## [1] 57602    75
```

Preprocesado inicial de los datos

En el preprocesado inicial de los datos . También se normalizan los datos restantes y se realiza una trasformación logarítmica a los conteos por millón (CPM) para manejarlos mejor en los posteriores análisis sin perder la representatividad de estos. De manera arbitraria, hemos escogido como límite defiltrado que un gen tenga más de 1 CPM en al menos la mitad de las muestras.

Comprobamos que se ha añadido correctamente el assay filtrado al nuevo objeto *se* y echamos un ojo al número de genes que quedan tras el filtrado

```
assays(se_filt)
```

```
## List of length 2
## names(2): counts logCPM
```

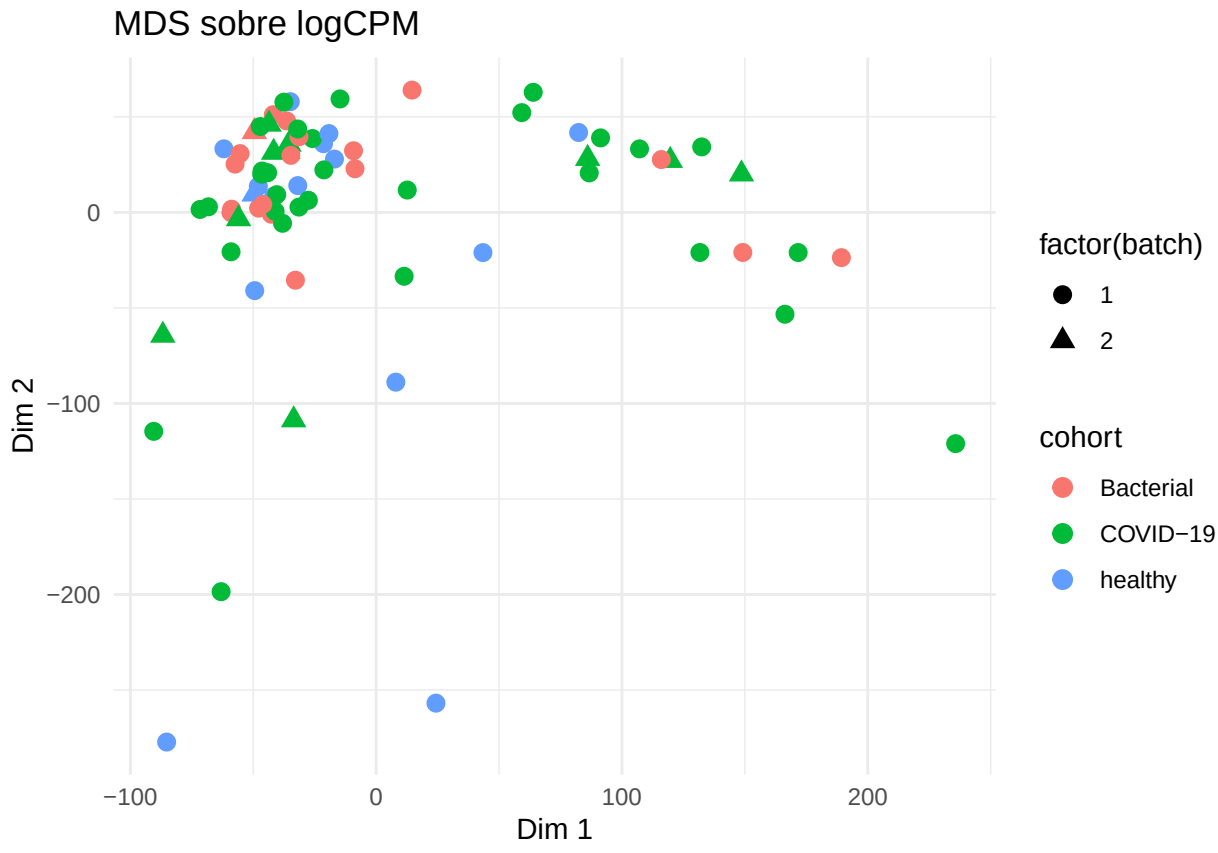
```
dim(se_filt)
```

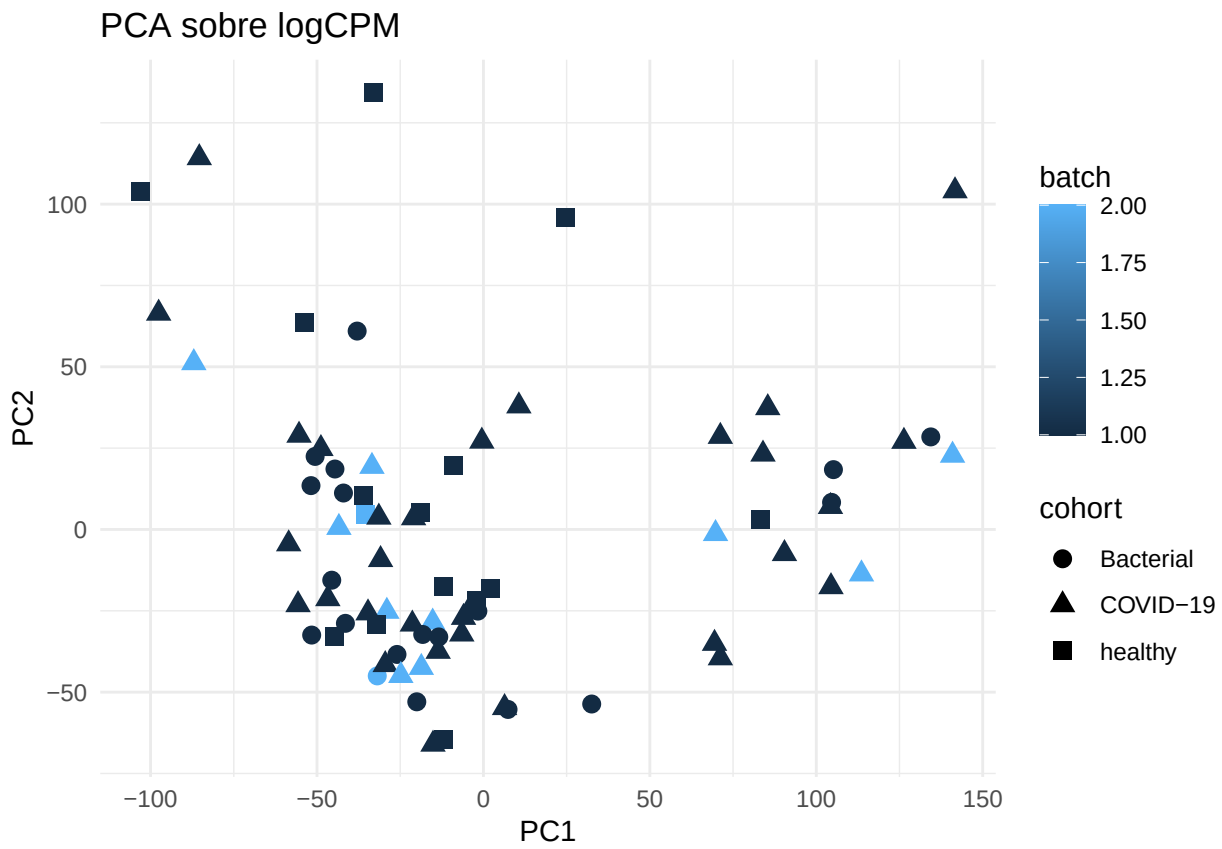
```
## [1] 13592    75
```

El número de genes se ha reducido muy notablemente desde 57.602 hasta 13.592.

Análisis exploratorio

Para realizar un análisis exploratorio de los datos en primer lugar vamos a hacer dos análisis de reducción dimensional, uno mediante el método MDS y otro mediante el método PCA. Como ambos análisis nos permiten discernir en las graficaciones dos tipos de información en la representación de cada punto (una mediante su color y otro mediante su forma) hemos decidido que se representen la cohorte (para ver si se agrupan las muestras dependiendo de si provienen de individuos sanos o de pacientes con neumonía bacteriana o covid) y el batch (para ver si hay un *batch effect* claro en los gráficos). Hemos decidido representar el batch y no cualquier otra variable que pudiera ser confusora porque la observación de un *batch effect* nos indicaría un sesgo técnico que habría que atajar antes de continuar con los análisis

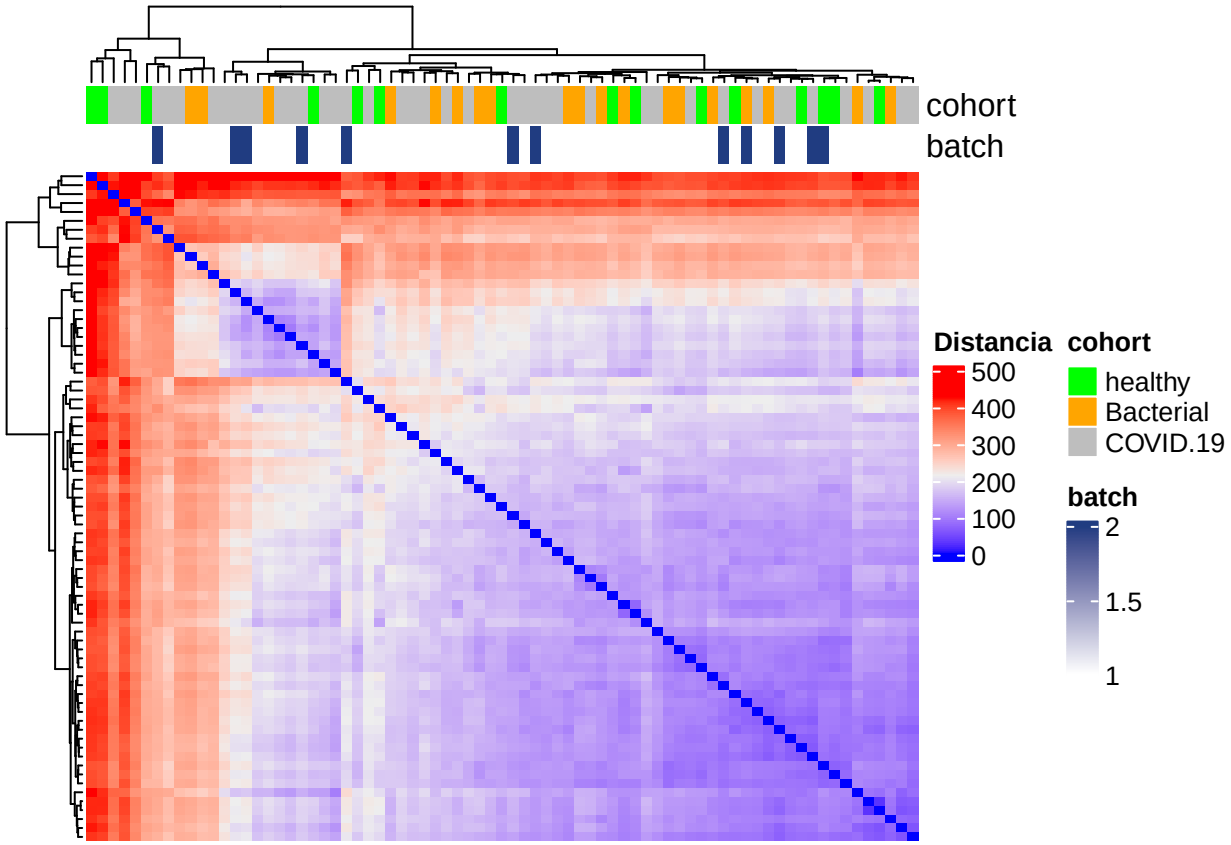




Como sepuede observar, ni en la graficación del MDS ni en la del PCA se observan claras agrupaciones en torno a la variable de la cohorte ni del batch. Esto nos indicaría que no hay un *batch effect* muy aparente en las muestras y que no existen diferencias lo suficientemente generalizadas en los perfiles transcriptómicos de los pacientes de las dos enfermedades y en los controles sanos como para observarlas con una reducción dimensional.

Otra aproximación sería realizar un heatmap con dendrograma asociado para comprobar si las muestras se agrupan entre sí. La idea de fondo es similar a la de los dos gráficos anteriores, comprobar si la expresión transcriptómica general de las muestras es suficientemente diferencial como para distinguirlas entre sí previamente a un análisis de expresión diferencial *per se*. La diferencia es que antes hemos usado dos aproximaciones de reducción dimensional y ahora usaremos un enfoque de agrupación jerárquica.

```
## [1] "healthy"   "Bacterial" "COVID.19"
```



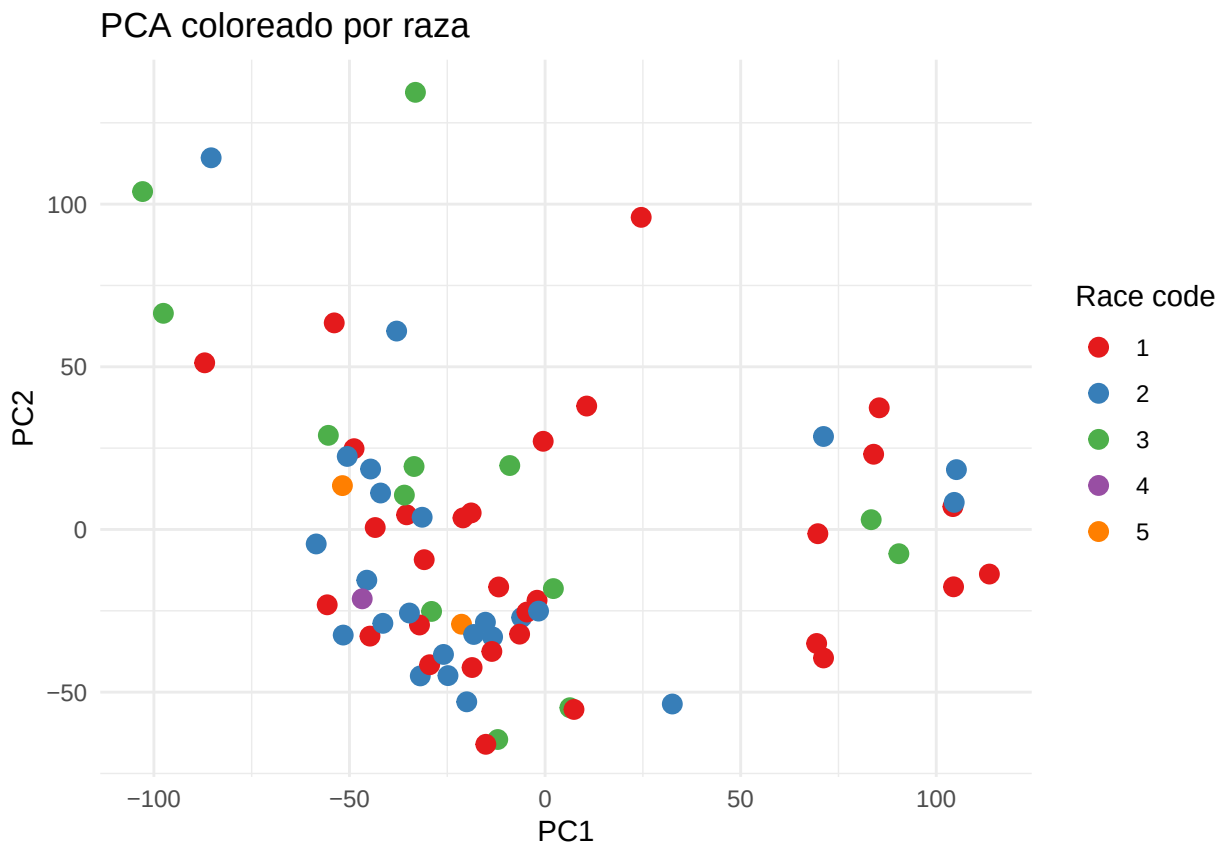
Como se puede observar en el dendrograma asociado al heatmap, no existe a simple vista una agrupación demasiado clara entre muestras de la misma cohorte ni con el mismo batch.

A simple vista parece que el batch no está actuando como una variable confusora, es decir, parece ser que no hay un *batch effect*. Para analizar matemáticamente si hay variables confusoras vamos a realizar el siguiente procedimiento: en primer lugar identificaremos y eliminaremos los outliers, poniendo como definición de outlier aquel punto que esté más allá del doble de la desviación estándar en el PC1 y/o el PC2 del PCA. Una vez eliminados realizaremos un análisis de correlación entre el PC1 y las variables que son candidatas a ser confusoras. Elegimos hacer la correlación solo con el PC1 porque es el componente que explica una mayor proporción de la variabilidad de las muestras, por lo que deducimos que de haber efecto confusor este se mostraría con mayor claridad en el primer componente principal. De entre todo el conjunto de datos hemos elegido como variables candidatas a ser confusoras el batch, la edad, el sexo y la raza. Se han escogido la raza, el sexo y la edad por ser las variables no técnicas que, en términos generales, generan una mayor variabilidad biológica entre individuos; por otra parte, se incluye el batch para cercionarnos de que efectivamente no hay un *batch effect*.

```
## batch vs PC1 correlation: -0.04
## race_num vs PC1 correlation: -0.22
## age vs PC1 correlation: 0.13
## gender_num vs PC1 correlation: 0.04
```

Se observa que no existe apenas correlación entre el sexo y el PC1 y el batch y el PC1. La correlación entre edad y PC1 es algo mayor pero sigue siendo baja. En cuanto a la correlación de raza y PC1, el coeficiente de correlación es ya algo mayor, de -0,22, lo que indica la presencia de una correlación negativa y débil. Aunque sea débil, es un efecto notablemente mayor al que ejercen el resto de variables candidatas a ser confusoras.

Para observar visualmente este efecto podemos volver a representar el gráfico PCA pero esta vez plasmando la pertenencia a cada raza



A simple vista no se observa el efecto confusor. No obstante, debido a su coeficiente de correlación incluiremos a la variable *raza* dentro de la matriz de diseño.

Análisis de expresión diferencial

Para construir la matriz de diseño y evaluar expresión génica diferencial de *Bacterial* frente a *Healthy* y de *Covid* frente a *Healthy* vamos utilizar el método que se elija de manera pseudoaleatoria por el código del enunciado de la PEC:

```
set.seed(myseed)
sample(c("edgeR", "voom+limma", "DESeq2"), size = 1)
```

```
## [1] "voom+limma"
```

En este caso vamos a usar el método voom+limma.

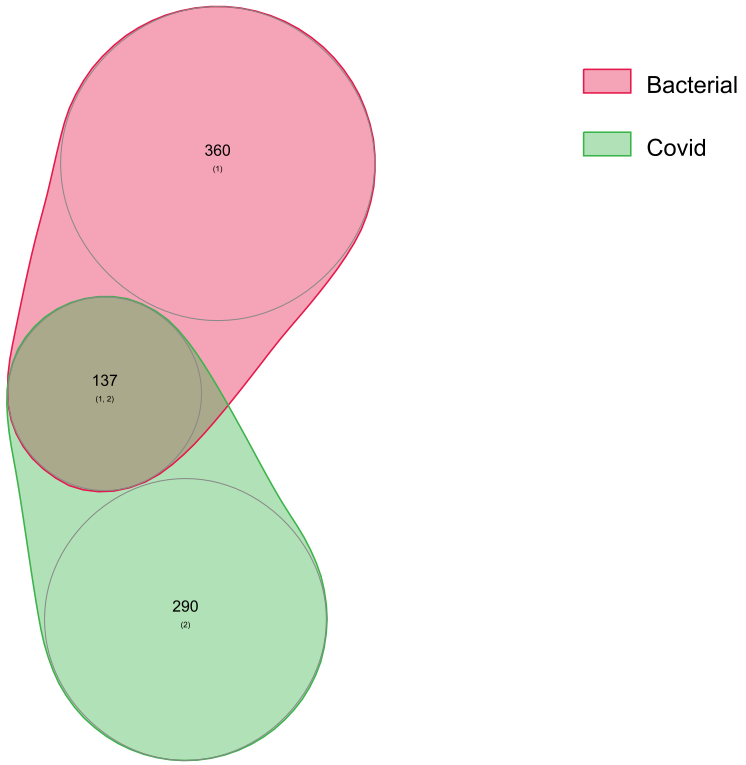
NOTA En el modelo que hemos calculado no salía ningún gen diferencialmente expresado si poníamos como umbrales o 0,05 de p-valor ajustado y 1,5 de log2FC. Para poder continuar con los análisis de la PEC hemos usado como umbral el p-valor crudo, no el ajustado, para que así haya genes a analizar. Esto no es lo correcto (somos conscientes de ello), se debería usar el valor ajustado, pero de esta manera podemos continuar con un análisis de expresión diferencial.

```
## Genes DE en Bacterial vs Healthy: 497
```

```
## Genes DE en COVID-19 vs Healthy: 427
```

Así pues, tendríamos un total de 497 genes diferencialmente expresados entre bacterial y healthy un un total de 427 genes diferencialmente expresados entre Covid y healthy.

Para comprobar hasta qué punto estos genes diferencialmente expresados podrían ser comunes en ambas comparaciones podemos realizar un diagrama de Venn y ver hasta qué punto solapan ambos conjuntos. Un gran solapamiento nos podría indicar que los genes cuya transcripción se ve más alterada en ambos casos podrían pertenecer a vías moleculares generales relacionadas con la respuesta inmunitaria o la respuesta a daño intracelular, ambas relacionadas con procesos infectivos.



Análisis funcional

Para realizar un análisis funcional de los genes diferencialmente expresados vamos a obtener los términos GO compartidos por los genes de cada comparativa y los vamos a ordenar en base a su p-valor asociado. Vamos a ranquear los diez términos GO más significativos por comparativa.

Términos GO Covid vs Healthy

```
##       term_id                                        term_name    p_value
## 1  GO:0006366             transcription by RNA polymerase II 0.02300536
## 2  GO:0006793                   phosphorus metabolic process 0.02300536
## 3  GO:0006796 phosphate-containing compound metabolic process 0.02300536
## 4  GO:0007264       small GTPase-mediated signal transduction 0.02300536
## 5  GO:0016192                     vesicle-mediated transport 0.02300536
## 6  GO:0016197                           endosomal transport 0.02300536
## 7  GO:0035556              intracellular signal transduction 0.02300536
```

```
## 8  GO:0050794                  regulation of cellular process 0.02300536
## 9  GO:0043412                      macromolecule modification 0.02627956
## 10 GO:0046907                         intracellular transport 0.02627956
##    intersection_size
## 1                 49
## 2                 47
## 3                 47
## 4                 17
## 5                 34
## 6                 12
## 7                 53
## 8                152
## 9                 54
## 10                30
```

En el caso del Covid se observa que las rutas más alteradas están relacionadas con la trascirpción, el transporte intracelular (en particular el endosomal), la transducción de señales y los procesos metabólicos de compuestos fosfatados.

Términos GO Covid vs Healthy

```
##       term_id                                                term_name
## 1  GO:0035556                      intracellular signal transduction
## 2  GO:0007264              small GTPase-mediated signal transduction
## 3  GO:0080090                regulation of primary metabolic process
## 4  GO:0141124                       intracellular signaling cassette
## 5  GO:0019222                  regulation of metabolic process
## 6  GO:0150104                transport across blood-brain barrier
## 7  GO:0009889                  regulation of biosynthetic process
## 8  GO:0010232                                vascular transport
## 9  GO:0065007                             biological regulation
## 10 GO:0010604 positive regulation of macromolecule metabolic process
##        p_value intersection_size
## 1  0.002088295                59
## 2  0.002088295                20
## 3  0.002088295                92
## 4  0.005749500                42
## 5  0.015543304               107
## 6  0.016344563                 7
## 7  0.016344563                91
## 8  0.016344563                 7
## 9  0.016344563               166
## 10 0.016344563                60
```

En cuanto a la infección bacteriana, los procesos más alterados tienen que ver con la transducción de señales, los procesos metabólicos primarios y el transporte vascular.

Discusión

En base a los resultados podemos deducir varias cuestiones de interés. En primer lugar que en lo que respecta a la respuesta biológica a nivel transcritómico a infecciones bacterianas y por Covid existe poca influencia

de factores como el sexo y la edad, tal y como se puso de manifiesto en el análisis de correlación. A este respecto cabe señalar que, aunque la influencia sea estadísticamente pequeña en el nivel trnascriptómico, a nivel tisular y de organismo estas pequeñas diferencias pueden tener consecuencias importantes.

De igual manera, la no agrupación de manera nítida de las muestras por cohorte en los análisis de reducción dimensional y agrupamiento jerárquico nos indica que las infecciones bacterianas y por covid no estarían produciendo cambios transcriptómicos a gran escala, al menos no lo suficientemente fuertes como para separar los grupos en dichos análisis.

Finalmente, del análisis funcional de los genes diferencialmente transcritos cabe destacar que en el caso de la infección por Covid se ven alteradas preferencialmente las vías de transducción de señales y de transporte intracelular, mientras que en la infección bacteriana se alteran los procesos metabólicos primarios, el transporte vascular y la transducción de señales. Conviene resaltar que en ambas comparaciones resulta alterada la vía de transducción de señales mediada por GTP-asas pequeñas

Conclusiones

1 La edad y el sexo son variables que afectan poco a la respuesta transcriptómica frente a la infección por Covid y la neumonía bacteriana

2 Ambos tipos de infección no generan respuestas transcriptómicas generalizadas y a gran escala al menos en muestras de sangre periférica

3 Las vías preferencialmente alteradas en ambas infecciones están relacionadas con la transducción de señales. En el caso del Covid se añadiría el transporte intracelular y en el caso de la infección bacteriana los procesos metabólicos primarios y el trasporte vascular.

Referencias

Hemos decidido subir el código de esta PEC en formato Rmarkdown para facilitar su lectura. Para acceder a él se ha de visitar el siguiente repositorio de GitHub: https://github.com/spisv/Pis_Vigil_Samuel_PEC2

Artículos científicos

Dyer, S. C., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Barrera-Enriquez, V. P., Becker, A., Bennett, R., Beracochea, M., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Cortes, L. A., ... Yates, A. D. (2025). Ensembl 2025. Nucleic acids research, 53(D1), D948–D957. https://doi.org/10.1093/nar/gkae1071

Pérez-Silva, J. G., Araujo-Voces, M., & Quesada, V. (2018). nVenn: generalized, quasi-proportional Venn and Euler diagrams. Bioinformatics (Oxford, England), 34(13), 2322–2324. https://doi.org/10.1093/bioinformatics/bty109

Tutoriales web

https://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/GenomicRanges/doc/GenomicRangesIntroduction.pdf

https://aspteaching.github.io/Analisis_de_datos_omicos-Ejemplo_2-RNASeq/Workflow_basico_de_RNASeq.html#7_An%C3%A1lisis_de_expresi%C3%B3n_diferencial_con_limma-voom

https://aspteaching.github.io/AMVCasos/#an%C3%A1lisis-de-componentes-principales

https://jokergoo.github.io/ComplexHeatmap-reference/book/