PEC2

Bruno Bel

2025-12-14

Contents

1 Abstract

Este estudio presenta un análisis transcriptómico comparativo de muestras de sangre periférica procedentes de pacientes con COVID-19, infecciones bacterianas e individuos sanos, utilizando datos del estudio de McClain et al. El objetivo principal del estudio fue identificar la firma molecular específica de la infección por SARS-CoV-2 y los procesos biológicos.

En nuestra exploración tras un preprocesamiento y normalización de datos de RNA-seq, se realizó un análisis de expresión diferencial mediante modelos lineales. Los resultados revelaron una respuesta inmunitaria distintiva en COVID-19, una menor expresión génica en comparación a las infecciones bacterianas. Sin embargo, los genes sobreexpresados mostraron un enriquecimiento significativo en la respuesta inmune adaptativa y la regulación del ciclo celular. Estos hallazgos sugieren que, a pesar de una supresión transcriptómica generalizada, existe una activación específica de mecanismos de defensa celular y división nuclear esenciales para la respuesta del huésped ante el virus.

2 Objetivos

El propósito central de este trabajo es caracterizar la respuesta molecular y célular del paciente frente a la infección por SARS-CoV-2 mediante técnicas de bioinformática y estadística. Los objetivos específicos son:

- Identificar patrones de agrupamiento y posibles fuentes de variación técnica (efecto batch) mediante técnicas de reducción de la dimensionalidad.
- Identificar genes diferencialmente expresados. Determinar la firma de expresión específica de pacientes con COVID-19 y compararla con la respuesta ante infecciones bacterianas para discernir patrones comunes y exclusivos.
- Caracterizar funcionalmente la respuesta biológica. Interpretar biológicamente los genes identificados mediante el análisis de enriquecimiento.

3 Métodos

Los datos analizados provienen de un subconjunto del estudio clínico de McClain et al., depositado en repositorios públicos de expresión génica. La naturaleza de los datos es transcriptómica, obtenida mediante RNA-seq de sangre periférica. El dataset original incluye recuentos de lecturas, junto con metadatos clínicos.que categorizan a los individuos en tres grupos: COVID-19, infección bacteriana y controles sanos.

Se utiliza el paquete SummarizedExperiment para la gestión de datos y metadatos. Se aplicó un filtrado de genes con baja expresión para reducir el ruido y tener en cuenta genes con expresión real. La normalización se realizó mediante transformación logarítmica para estabilizar la varianza.

Empleamos el Análisis de Componentes Principales (PCA) y Heatmaps para evaluar la homogeneidad de los grupos y visualizar diferencias según variables categóricas. Acorde a esto definimos la matriz de diseño y contrastes.

Se utiliza la librería voom-lima, fijando un umbral de P.adj < 0.05 y logFC de 1.5.Para la interpretación biológica, se utiliza el paquete topGO.

Durante el análisis se incluyen gráficos asociados a los distintos objetos, generados por las funciones de las própias librerias o utilizando ggplot2.

4 Introducción

Los archivos y el código generado para realizar los análisis posteriores se encuentran en el repositorio llamado "Analisis-Datos-Omicos—PEC-2" al que se puede acceder mediante el siguiente enlace https://github.com /bruniix/Analisis-Datos-Omicos---PEC-2.

Para dicha resolución se ha trabajado con R y utilizando RMarkdown para la escritura y ejecución.

5 Obtención de los datos

En primer lugar, tal como se indica en la propuesta del ejercicio, los datos de el estudio se encuentran en la plataforma Gene Expression Omnibus mediante el identificador GSE161731.

Como se describe en la plataforma, en el estudio dispusieron de 46 pacientes con COVID-19 alguno de ellos tomando varias muestras en momentos distintos llegando al número total de 77 muestras de sangre periférica.

Para la comparación transcriptómica y de la expressión génica compararon las muestras con las almacenadas de otros pacientes diagnosticados con infección respiratoria aguda debída a coronavirus estacional, al virus de la gripe, neumonía bacteriana o pacientes control.

Además los pacientes de estudio fueron divididos en grupos según la gravedad de la infección aparente debido a lo sintomas expresados y mediante tiempo en relación al desarrollo de la enfermedad.

Para empezar vamos a descargar los datos mediante el paquete GEOquery que nos permite acceder directamente a la base de datos.

5.1 Descarga datos GEO, preparación datos contaje y metadatos

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 0 features, 198 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM4913486 GSM4913487 ... GSM4913683 (198 total)
##   varLabels: title geo_accession ... tissue:ch1 (67 total)
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
##   pubMedIds: 33597532
## 34001247
## Annotation: GPL24676
```

```
## Habiendo descargado los datos mediante el paquete GEOquery, obtenemos eluna lista que contiene ub ob
```

```
##
##  Utilizando la funcion experimentData() visualizamos la información asociada al caso de GEO y el enl
```

```
##
##  Las dimensiones del objeto descargado que contiene los datos de expresión son: 60675 201
##  Podemos hacernos una idea de la distribución de los datos viendo la expresión genica para algunas d
```

	DU09-941890	DU09-03S0000604	DU09-03S0000611	DU09-10592	DU09-03S0000774	DU09-03S0000775	DU09-03S19478	DU14-03S0000878	DU14-03S0000889	DU18-02S0011619
ENSG00000223972	2	3	49	6	46	35	26	14	0	
ENSG00000227232	242	152	246	586	570	213	559	369	192	
ENSG00000278267	72	29	44	104	69	58	73	56	0	
ENSG00000243485	5	0	0	1	1	0	5	0	0	
ENSG00000274890	0	0	0	0	0	0	0	0	0	
ENSG00000237613	3	0	0	0	0	0	0	0	0	
ENSG00000268020	0	0	0	0	0	0	0	0	0	
ENSG00000240361	1	0	7	0	0	0	0	0	0	
ENSG00000186092	2	0	0	0	0	0	0	0	0	

	94189	DU09-03S0000604	DU09-03S0000611	105920	DU09-03S0000774	DU09-03S0000775	DU09-03S19478	DU14-03S0000878	DU14-03S0000889	DU18-02S0011619
ENSG00000223972	29	72	64	31	17	54	32	33	6	102

```
##
##  Las dimensiones del objeto importado de la base de datos de GEO
## , que contiene los metadatos (información en relación a las muestras y los pacientes) son 198 67 Pod
```

Vemos que los metadatos importados de la base de datos GEO tienen su identificador própio, lo que nos impide relacionarlos con las muestras correspondientes. Modificamos el nombre de las filas para que corresponda con los carácteres de interés de la variable "titulo".

Para preparar el objeto SummarizedExperiment debemos en primer lugar asegurar que disponemos de información para las muestas de las cuales se ha obtenido el trasnscriptoma y la anotación genica para los genes cuantificados.

Ya hemos visto que las dimensiones eran distintas para ambos objetos, por lo que cabe suponer que hay 3 muestras de las cuales no se dispone información (ncol(count_matrix)-nrow(metadata)).

```
## Al haber realizado la intersección tenemos el mismo identificador, y por tanto información completa,
```

5.2 Descarga datos genes y filtrado según genes en común

```
## Después de haber seleccionado los genes y carácteristicas comunes las dimensiones de la matriz de co
```

5.3 Creación obtejo SummarizedExperiment

Una vez tenemos la matriz de contajes con las observaciones correspondientes, los metadatos con el nombre de fila correcto y hemos comprobado la coincidencia y orden correcto, podemos crear el objeto SummarizedExperiment para conseguir tener todos los datos enlazados.

```
## class: RangedSummarizedExperiment
## dim: 57602 194
## metadata(0):
## assays(1): counts
## rownames(57602): ENSG00000223972 ENSG00000227232 ... ENSG00000277475
##   ENSG00000268674
## rowData names(6): gene_id gene_name ... symbol entrezid
## colnames(194): 94189 DU09-03S0000604 ... 105847 105848
## colData names(67): title geo_accession ... time_since_onset:ch1
##   tissue:ch1
```

Como se observa al mostrar el SummarizedExperiment, está formado por 57602 genes a los cuáles se dispone de registro de expressión para 194 muestras, con 67 columnas con información de los metadatos y 6 variables en relación a los genes.

6 Limpieza y selección de datos

6.1 Selección de cohorte

En primer lugar seleccionamos la variables de interés para los metadatos y cambiamos el nombre para mejor interpretación.

```
## Columnas de metadatos seleccionadas:
```

```
##  [1] "title"            "geo_accession"    "age"              "batch"
##  [5] "cohort"           "gender"           "hospitalized"     "race"
```

```
##  [9] "id"               "time_since_onset"
```

```
## Formal class 'DFrame' [package "S4Vectors"] with 6 slots
##   ..@ rownames       : chr [1:194] "94189" "DU09-03S0000604" "DU09-03S0000611" "105920" ...
##   ..@ nrows          : int 194
##   ..@ elementType    : chr "ANY"
##   ..@ elementMetadata: NULL
##   ..@ metadata       : list()
##   ..@ listData       :List of 10
##   .. ..$ title           : chr [1:194] "whole blood, 94189" "whole blood, DU09-03S0000604" "whole blo
##   .. ..$ geo_accession   : chr [1:194] "GSM4913486" "GSM4913487" "GSM4913488" "GSM4913489" ...
##   .. ..$ age             : chr [1:194] "57" "19" "14" "21" ...
##   .. ..$ batch           : chr [1:194] "1" "2" "2" "1" ...
##   .. ..$ cohort          : chr [1:194] "Bacterial" "Influenza" "Influenza" "Influenza" ...
##   .. ..$ gender          : chr [1:194] "Female" "Male" "Male" "Female" ...
##   .. ..$ hospitalized    : chr [1:194] "NA" "NA" "NA" "NA" ...
##   .. ..$ race            : chr [1:194] "Black/African American" "Black/African American" "White" "Whi
##   .. ..$ id              : chr [1:194] "A1BD46" "44DF6B" "658A11" "61DE97" ...
##   .. ..$ time_since_onset: chr [1:194] "NA" "NA" "NA" "NA" ...
```

Seleccionamos las cohortes en las que estamos interesados, cogeremos solo las muestras que corresponden clasificadas como: COVID19, Bacterial y healthy.

```
## Actualmente las clasificaciones para las variables de la cohorte son:
##  Bacterial Influenza CoV other COVID-19 healthy
```

```
##
## Bacterial  COVID-19   healthy
##        24        77        16
```

Vemos que del número inicial de muestras quedan 117 según las cohortes y pacientes con el estado inmunológico de interés-

6.2 Revisión datos

A continuación antes de passar al análisis de los datos, revisamos que no hayan individuos repetidos y que las variables sean del tipo que corresponde

```
## El número de identificadores de las muestras y de entradas coincide: FALSE
```

```
## Habiendo eliminado los duplicados, ahora el número total de observaciones / muestras es de: 86
```

Eliminamos espacios en blanco de la variable raza y cambiamos el tipo de variable: edad a númerica y batch, cohorte, genero y raza a factor.

Al transformar la variable edad a númerica aparece el aviso de que se han introducido valores NA's. Buscamos los valores únicos para observar cuales no pueden ser asociados a un número entero.

```
## Todos los valores de la variable 'age' : 57 60 33 28 27 31 30 32 29 26 50 46 56 54 51 20 61 63 52 59
```

```
##
## >89  18  19  20  26  27  28  29  30  31  32  33  35  37  38  39  43  46  50  51
##   1  11   5   1   1   1   4   5   3   3   2   4   2   1   2   1   1   1   2   2
##  52  53  54  55  56  57  59  60  61  62  63  64  67  68  69  70  71  72  74  75
##   2   1   1   1   1   1   1   3   1   1   1   1   1   1   2   1   1   1   1   1
##  76  79  80  81  84  85  87  88
##   2   1   1   1   1   1   1   1
```

```
## El valor que se introduce como NA es la edad registrada como >89, puesto que es la única observación
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.00   37.50   44.69   61.75   89.00
```

6.3 Semilla aleatoria y selección de muestras.

```
## La semilla generada para la selección aleatoria de los datos es: 1403
```

```
## Posterior a la selección aleatoria la distribución de muestras queda de la siguiente manera.
```

```
##
## Bacterial  COVID-19   healthy
##        19        41        15
```

7 Preprocesado, selección genes y transformación

7.1 Filtrado de genes

Antes de realizar el análisis y clasificación debemos hacer una selección de los genes a contabilizar. Genes con una muy baja expresión o sin expresión no son relevantes para realizar la comparación entre grupos de distinta cohorte.

Para ello utilizamos el paquete de Bioconductor edgeR, de este podemos utilziar la función cpm() para visualizar los contajes por millón y filtrar. Una función útil del paquete utilizado es la función filterByExpr() la cual realiza el cálculo mencionado anteriormente y simultáneamente seleccióna las filas o genes a seleccionar.

```
##
##  Habiendo filtrado los genes con una menor o nula expresión
##  nos queda el siguiente resumen el objeto:
```

```
## [1] "RangedSummarizedExperiment object of length 20953 with 6 metadata columns"
```

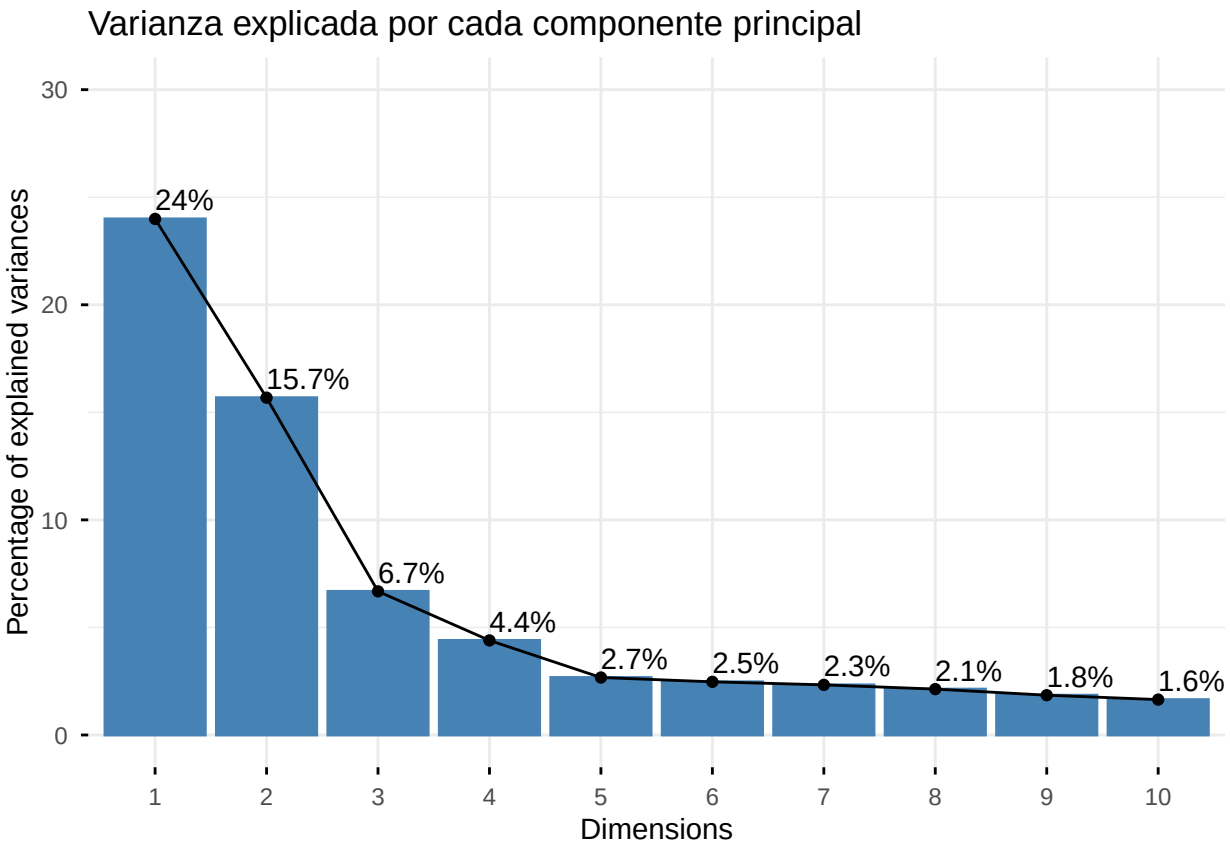
7.2 Normalización

A continuación, habiendo seleccionado los genes expresados, es necesario normalizar los recuentos para que las observaciones sean comparables. Para ello utilizamos las funciones de edgeR para tener presente las lecturas de la secuenciación y buscar un factor para igualar la matriz de conteos. Con esto pretendemos que la diferencia sea únicamente biológica y no de carácter técnico. Realizamos la normalización mediante el contaje por millon y aplicamos escala logaritmica para minimizar las diferencias y que el rango de los datos sea menor, por lo tanto más comparable.

8 Análisis exploratorio

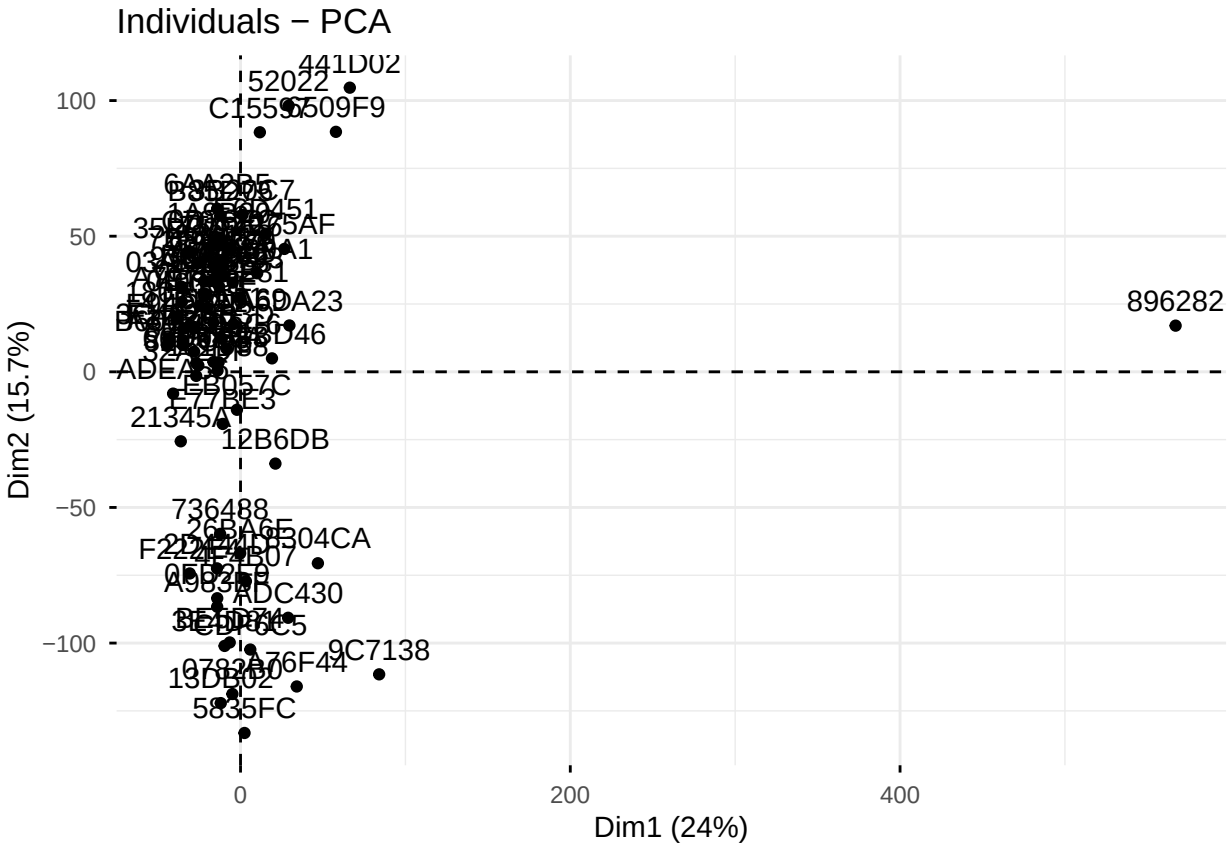
8.1 PCA

En primer lugar para visualizar la distribución de los datos generamos un análisis de componentes principales. Podemos utilizar el paquete factoextra para visualzar fácilmente las muestras y sus características.



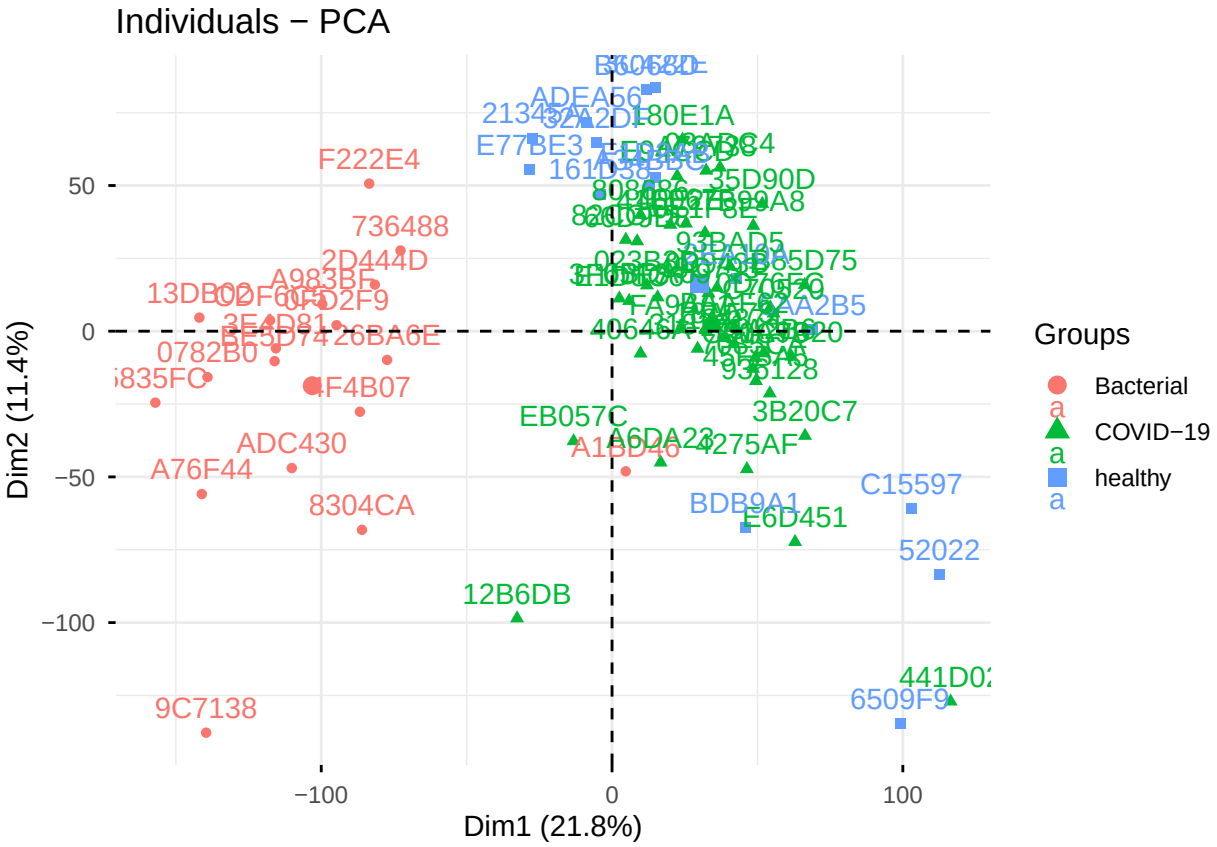
Grácias al gráfico anterior entendemos que dimensiones o ejes explican la variabilidad de la expressión, ya normalizada y en escala logarítmica.

Con el primer y segundo componente principal ya obtenemos un 40% de toda la variabilidad, lo cuál es considerable para poder gráficar y observar la distribución de las muestras. En caso de requerir información adicional podemos tener en cuenta hata la 3a y 4a dimensión.



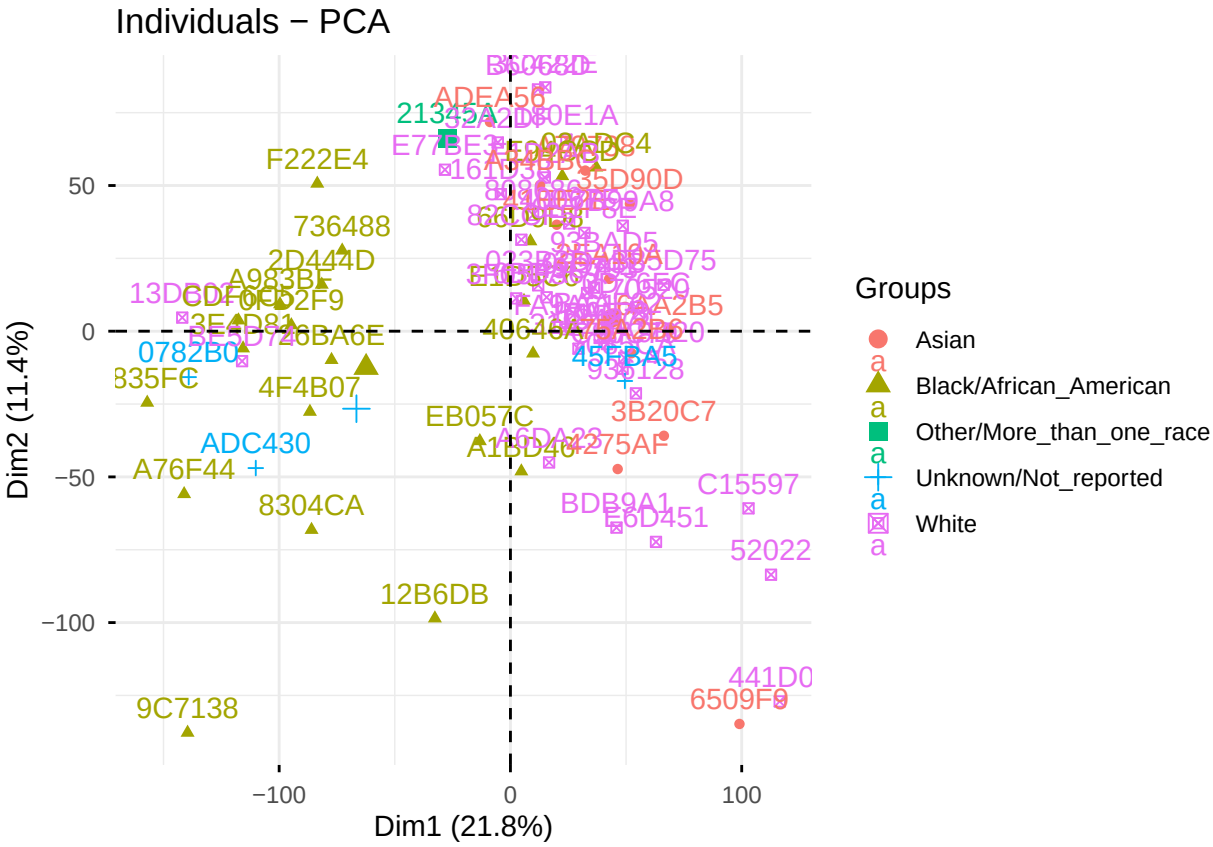
En primer lugar, al visualizar la distribución de las muestras en las dos dimensiones graficadas vemos una muestra atípica que impide una correcta visualización y muede sesgar el análisis. Procedemos a eliminarla y visualizar los componentes principales según algunas de las variables de interés como la cohorte, la raza o el sexo.

8.1.1 Cohorte



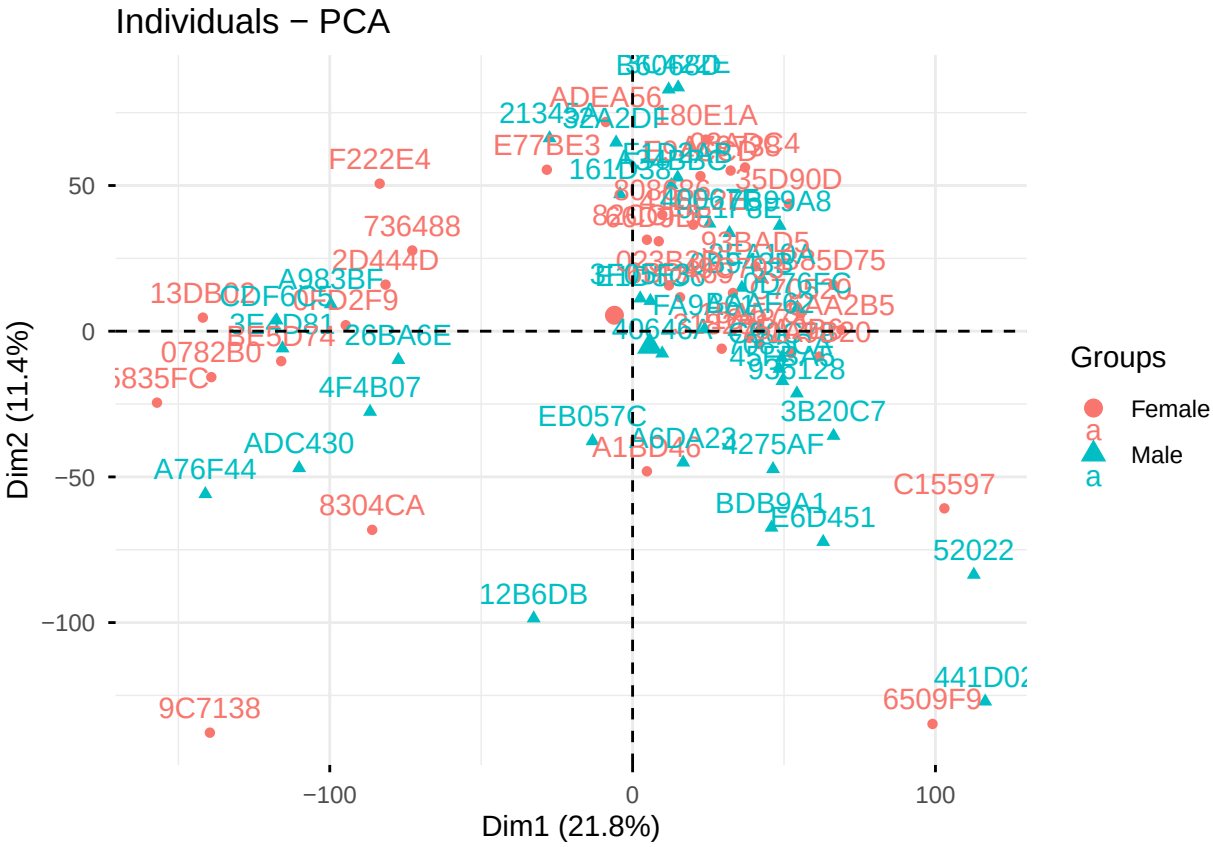
En primer lugar por lo que respecta a la cohorte o el grupo clínico al que los pacientes estaban asociados vemos una separación clara para el primer componente principal ente individuos con una infección bacteriana frente a una infección vírica como el Sars-CoV2 y los individuos sanos. De los pacientes del estudio con COVID-19 y los sanos no se puede ver una separación en etos componentes.

8.1.2 Raza



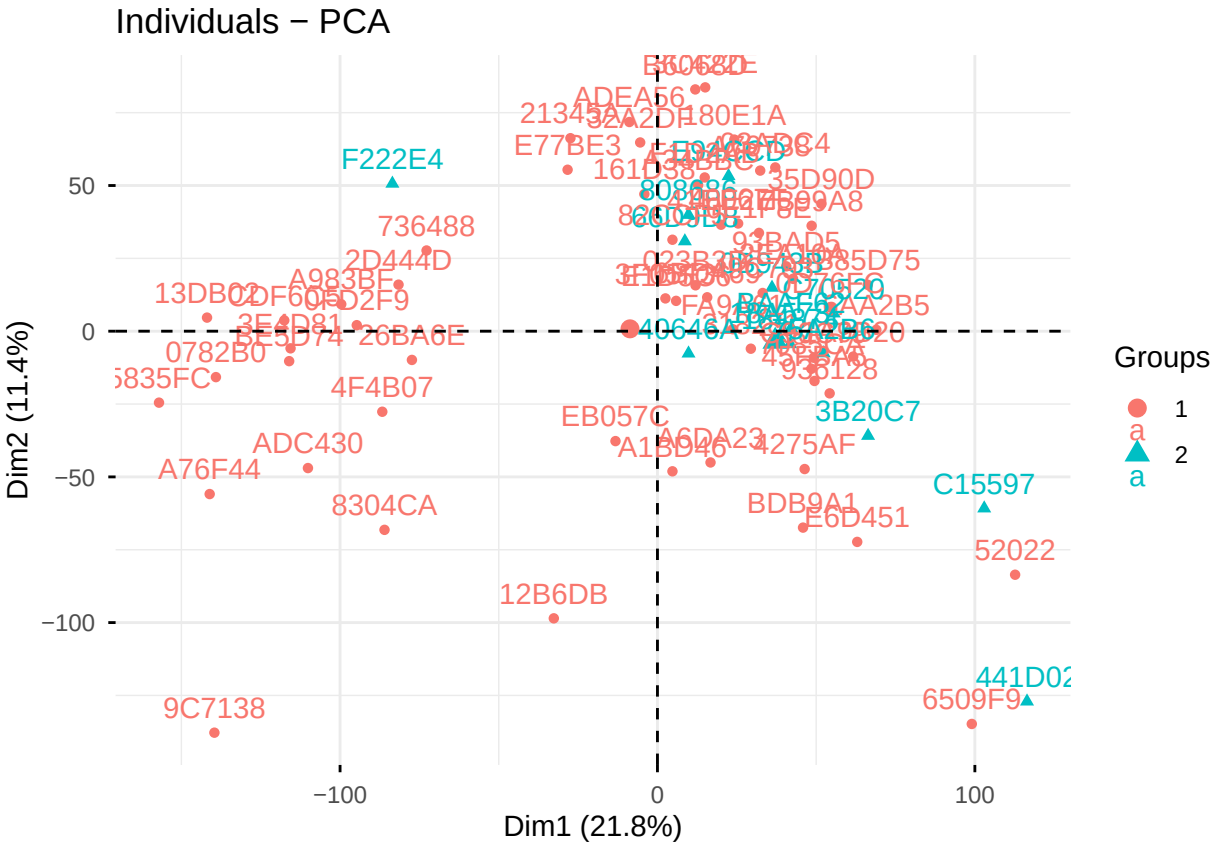
Por lo que refiere a la entina de los pacientes no se ve una división en los componenetes. Si bien es cierto que parece que pacientes africoamericanos son más presentes en una región respecto al eje principal podría ser debido a sesgo durante la obtención de las muestras para el banco.

8.1.3 Sexo



No se observa ningúna distribución según el sexo de los pacientes.

8.1.4 Batch

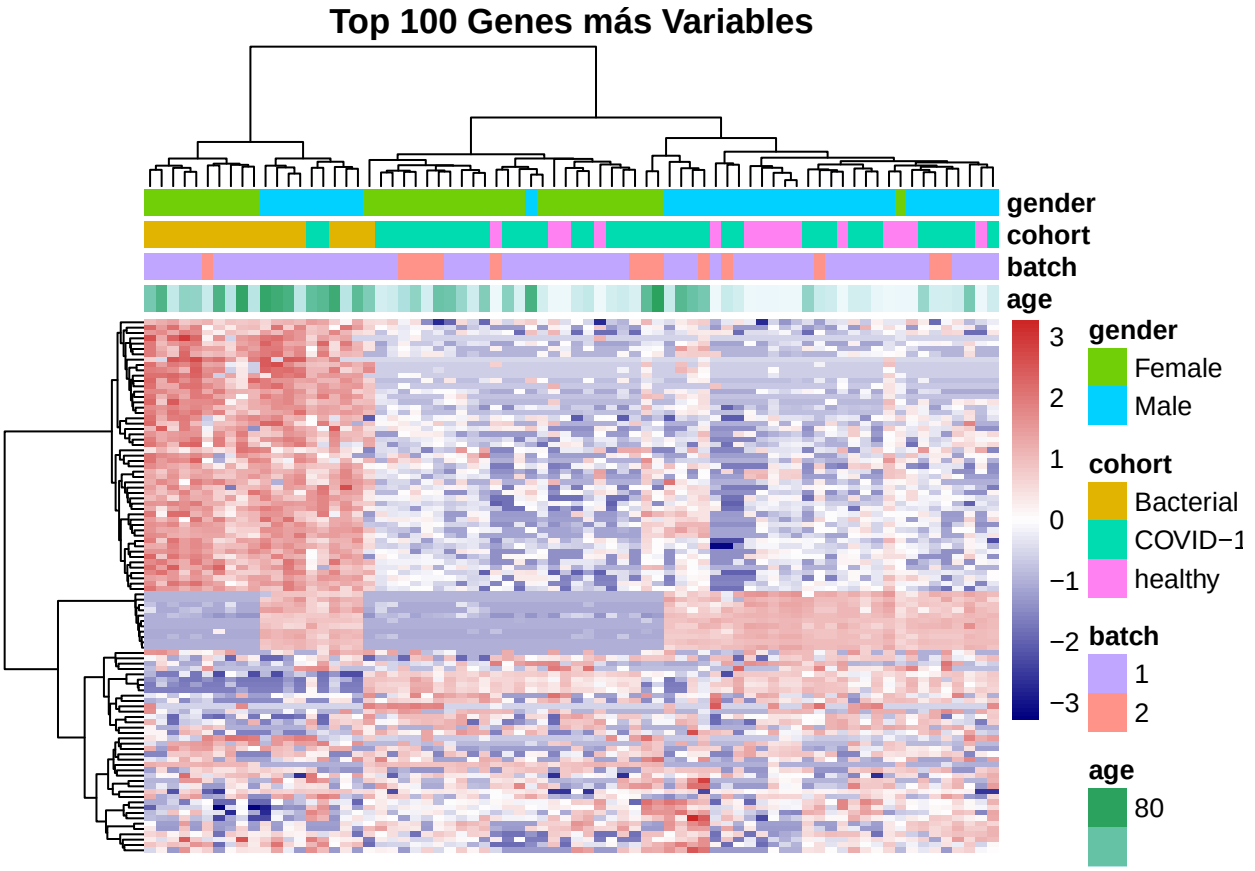



```
##
##  1  2
## 60 14
```

Si bien es cierto que parece haber una distribución en la variabilidad debida al primer componente principal para muestras del segundo batch, debido a la diferencia de número de muestras entre ambos grupos y la presencia heterogenea del primer batch no podemos afirmar diferencias.

8.2 Heatmap y clustering

Teniendo presente la agrupación de las muestras por lo que refiere a variabilidad realizamos un agrupamiento jerárquico según las similitudes de las muestras.



Al visualizar miles de genes se genera un mapa demasiado saturado donde no es posible obtener información relevante. En este caso hemos expresado los 100 genes más variables. De el clustering y mapa de calor podemos obtener ides similares a las obtenidas mediante el PCA.

Hay más similitud entre muestras o individuos que comparten cohorte y cierta coincidencia con la etnia de origen. En este caso si que se aprecian diferencias en la expresión para el genero, dentro de cada gran grupo del clustering hay un cambio marcado según si las muestras pertenecen a un paciente masculino o femenino, por lo que está característica resulta relevante.

Por lo que refiere a la edad, aunque no se configura un patrón claro, para la expresión dentro de los grupos generados parece observarse un cambio en la cuantificaicón según la edad a la que pertenecen. Además, siguiendo los conocimientos biológicos disponibles, podemos asumir que el estado fisiológico e inmunitario de un individuo puede ir relacionado con la edad, y por lo tanto la respueta frente a una infección.

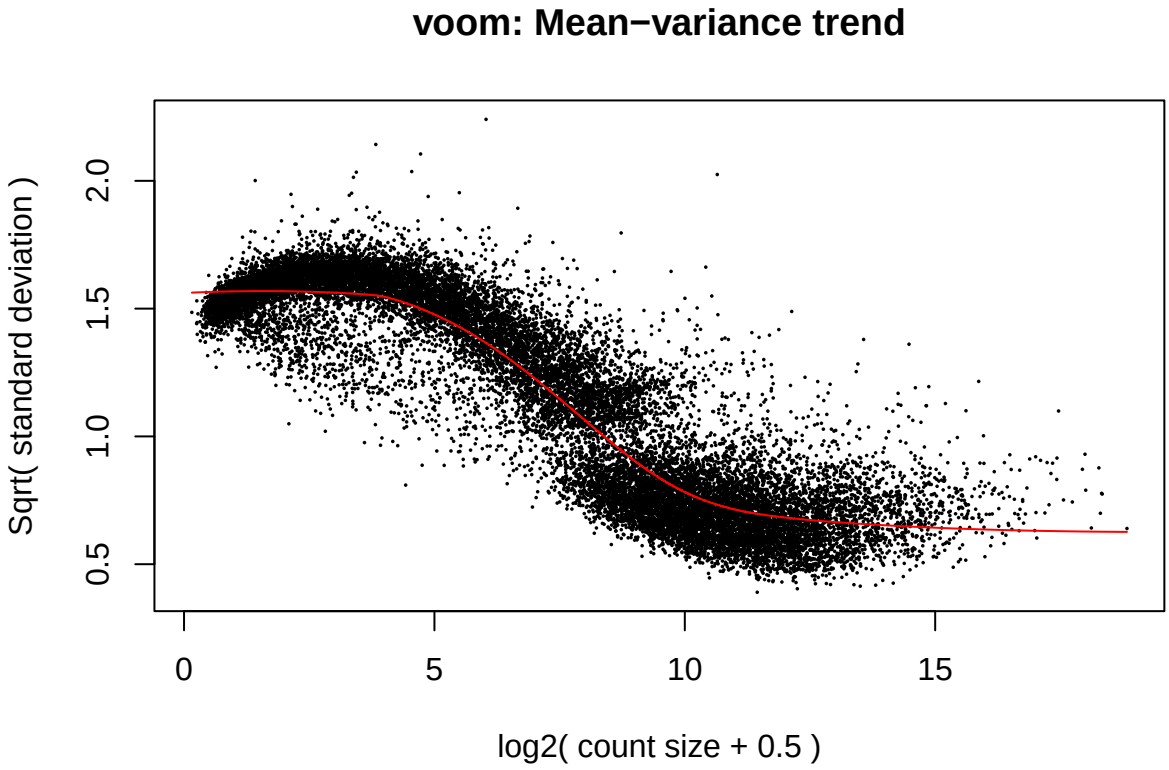
9 Expresión diferencial

Habiendo explorado los datos y las variables de interés procedemos con el análisis de expresión diferencial.

```
## El método seleccionado para el análisis según la semilla generada anteriormente es: voom+limma
```

En primer lugar creamos la matriz de diseño con las variables de interés: cohorte, sexo, etnia y edad. Y aplicamos el modelo voom lima para ver desviación de la media respecto a los contajes transformados.

```
## [1] "cohortBacterial"              "cohortCOVID-19"
## [3] "cohorthealthy"                "raceBlack/African_American"
## [5] "raceOther/More_than_one_race" "raceUnknown/Not_reported"
## [7] "raceWhite"                    "sexoMale"
## [9] "age"
```



Mencionar que hemos limpiado los niveles de la variable categórica etnia, ya que había configurado un nivel que no ha sido seleccionado en nuestro subgrupo.

Al haber utilizado la función voom vemos el gráfico esperable. Con un mayor contaje de la expresión (formato log2) obtenemos una menor desviación estándar y por lo tanto mayor estabilidad.

Ajustamos al modelo lineal y definimos los contrastes a realizar en una matriz de contrastes, en este caso Bacterial vs Healthy y COVID-19 vs Healthy. Luego ajustamos el modelo a dichos contrastes.

```
## [1] "cohortBacterial"              "cohortCOVID.19"
## [3] "cohorthealthy"                "raceBlack.African_American"
## [5] "raceOther.More_than_one_race" "raceUnknown.Not_reported"
## [7] "raceWhite"                    "sexoMale"
## [9] "age"
```

```
## La matriz de contrastes es:
```

```
##                                Contrasts
## Levels                          COVID_vs_Healthy Bacterial_vs_Healthy
##   cohortBacterial                              0                    1
##   cohortCOVID.19                               1                    0
##   cohorthealthy                               -1                   -1
##   raceBlack.African_American                   0                    0
##   raceOther.More_than_one_race                 0                    0
##   raceUnknown.Not_reported                     0                    0
##   raceWhite                                    0                    0
##   sexoMale                                     0                    0
##   age                                          0                    0
```

9.1 Comparativa COVID19 vs Healthy / Bacterial vs Healthy

```
## Primeras observaciones de los resultados de la comparación COVID vs Healthy:
```

```
##                     logFC  AveExpr         t      P.Value    adj.P.Val        B
## ENSG00000138326 -2.401211 6.078235 -9.277974 6.498135e-14 1.361554e-09 21.30597
## ENSG00000145425 -2.918927 6.155220 -8.177627 7.259920e-12 5.586412e-08 16.64757
## ENSG00000133112 -1.969766 9.465118 -8.155068 7.998490e-12 5.586412e-08 16.46472
## ENSG00000071082 -2.728704 5.382610 -8.051545 1.247596e-11 6.535220e-08 16.15941
## ENSG00000163682 -2.775337 6.047680 -7.970737 1.764980e-11 7.396326e-08 15.77702
## ENSG00000171863 -2.564476 5.665018 -7.889907 2.496878e-11 8.068250e-08 15.45606
```

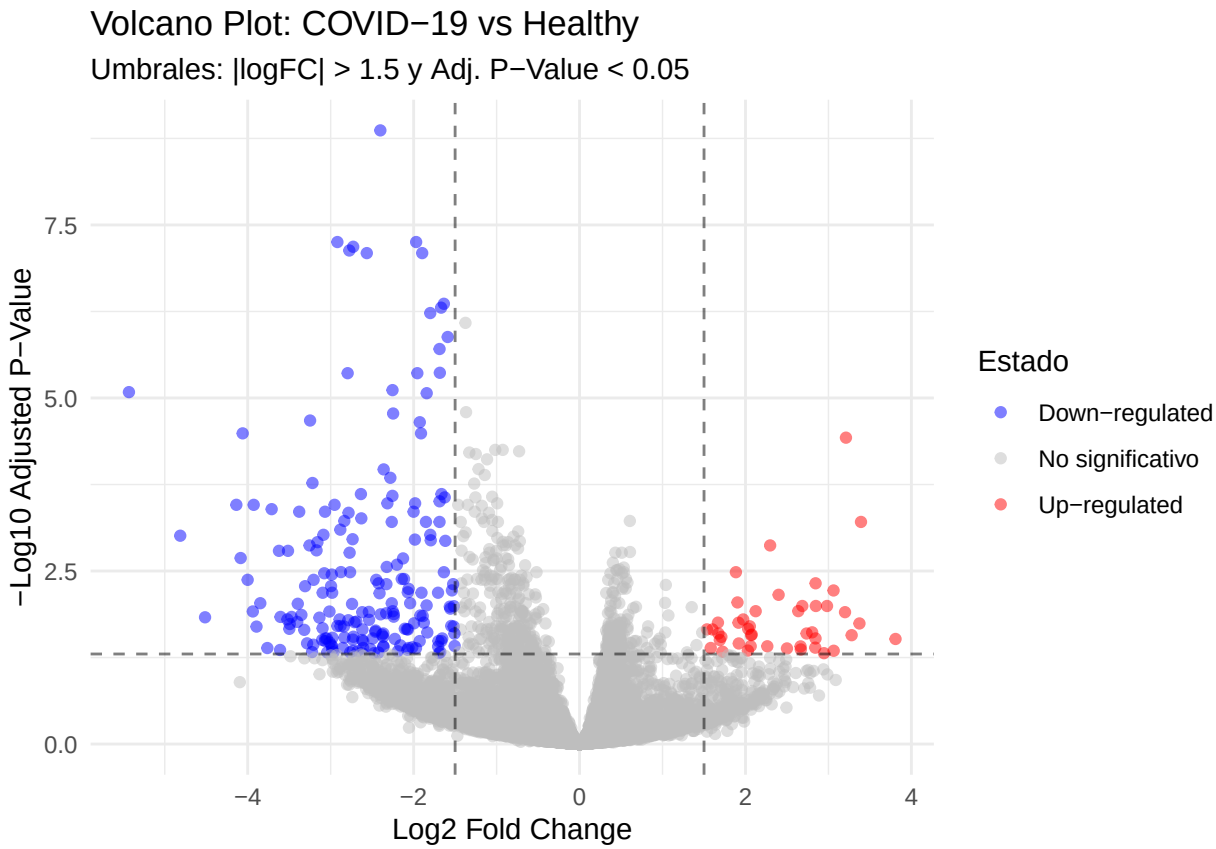
```
## Primeras observaciones de los resultados de la comparación Bacterial vs Healthy:
```

```
##                     logFC  AveExpr         t      P.Value    adj.P.Val        B
## ENSG00000188559  2.890308 6.318396  9.670735 1.221418e-14 1.534880e-10 22.94899
## ENSG00000114942 -3.171974 6.047745 -9.627888 1.465070e-14 1.534880e-10 22.75542
## ENSG00000196396  1.772180 6.713572  9.423315 3.497076e-14 1.958595e-10 21.94590
## ENSG00000157557  2.927563 6.075768  9.399919 3.863539e-14 1.958595e-10 21.82424
## ENSG00000231500 -3.124449 8.177072 -9.355244 4.673783e-14 1.958595e-10 21.67700
## ENSG00000101152  1.448018 7.713533  9.195380 9.245222e-14 2.872470e-10 21.01067
```

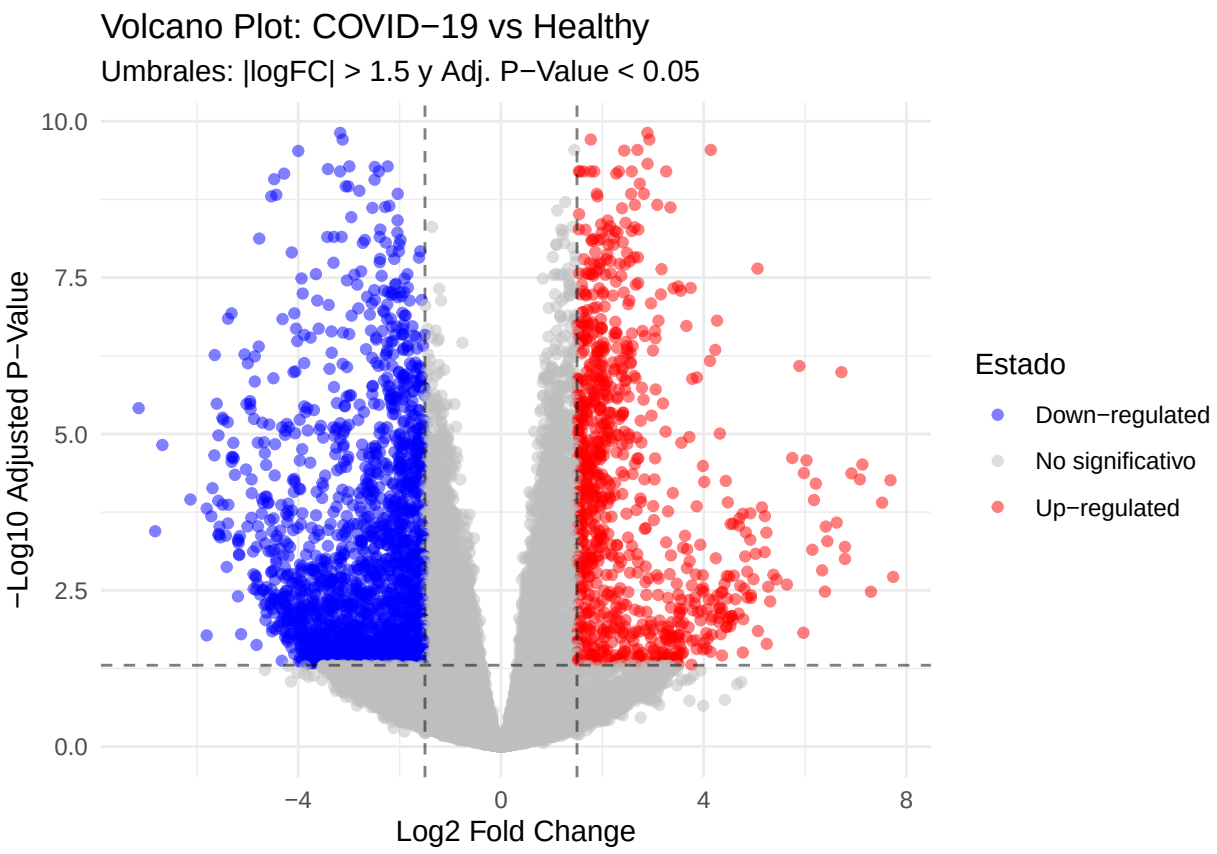
```
## Utilizamos la función decidetest() para ver que genes está sobreexpresados
##  o por debajo para ambas comparaciones. Establecemos los parametros lfc a 1.5 y p.value a 0.05
```

```
##        COVID_vs_Healthy Bacterial_vs_Healthy
## Down                190                1726
## NotSig            20719               18343
## Up                   44                 884
```

9.1.1 Visualización COVID vs Healthy



9.1.2 Visualización Bacterial vs Healthy



10 Comparación

En los diferentes análisis hemos observado lo siguiente:

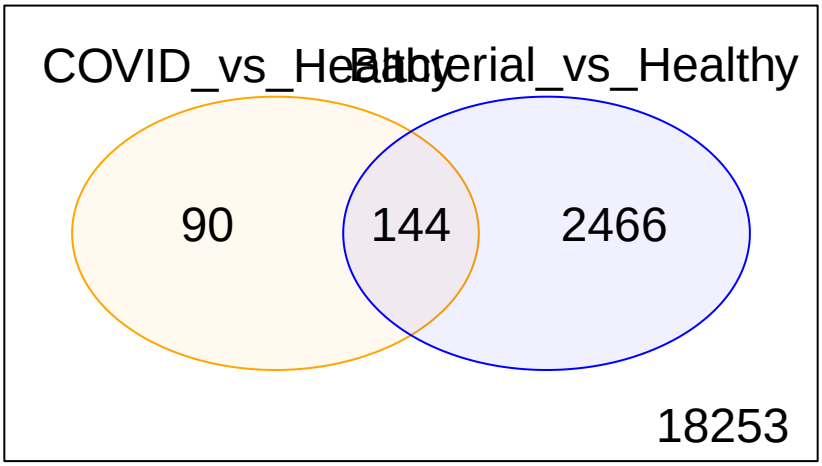
- Hay expresión diferencial entre individuos que presentan una infección y los individuos sanos.
- Las muestas de individuos con una infección de origen bacteriana presentan una expresión diferencial mucho mayor, tanto en aumento como disminución de dicha expresión.

Recordamos la tabla vista anteriormente.

```
##        COVID_vs_Healthy Bacterial_vs_Healthy
## Down                190                1726
## NotSig            20719               18343
## Up                   44                 884
```

Vemos que en ambos casos hay muchos más genes con menos expresión. Es decir que de manera global, pese a los genes más expresados, en sagre periférica se observa una menor expresión de genes a causa de infecciones tanto víricas como bacterianas. Podemos visualizar estos datos con un diagrama de Venn.

Genes Diferenciales Comunes y Específicos



Del total de genes diferencialmente expresados, 144 son comunes entre ambas infecciones mientras que 90 son exclusivamente de los pacientes con COVID-19 y 2466 de los pacientes con neumonia.

11 Sobrerepresentración

Por último vamos a analizar desde un punto de vista biológico los genes sobreexpresados para pacientes con COVID-19, que es la base y el objetivo final por parte de los datos de estudio.

Table 2: Análisis de Sobrerepresentación: Procesos Biológicos en COVID-19 (44 genes)

GO.ID	Term	Annotated	Significant	Expected	classicFisher
GO: 0002250	adaptive immune response	659	17	1.89	6.1e-13
GO: 0006955	immune response	1721	18	4.93	2.6e-07
GO: 1901970	positive regulation of mitotic sister ch…	20	4	0.06	2.7e-07
GO: 0016064	immunoglobulin mediated immune response	176	7	0.50	5.8e-07
GO: 0019724	B cell mediated immunity	179	7	0.51	6.5e-07
GO: 0010965	regulation of mitotic sister chromatid s…	61	5	0.17	7.8e-07
GO: 0000070	mitotic sister chromatid segregation	189	7	0.54	9.3e-07
GO: 0000280	nuclear division	381	9	1.09	9.5e-07
GO: 0051306	mitotic sister chromatid separation	64	5	0.18	9.9e-07
GO: 0007051	spindle organization	194	7	0.56	1.1e-06

GO.ID	Term	Annotated	Significant	Expected	classicFisher
GO: 0098813	nuclear chromosome segregation	290	8	0.83	1.3e-06
GO: 0002376	immune system process	2392	20	6.86	1.5e-06
GO: 1905820	positive regulation of chromosome separa…	30	4	0.09	1.5e-06
GO: 1905818	regulation of chromosome separation	74	5	0.21	2.1e-06
GO: 0048285	organelle fission	426	9	1.22	2.4e-06
GO: 1903047	mitotic cell cycle process	706	11	2.02	2.9e-06
GO: 0000819	sister chromatid segregation	226	7	0.65	3.1e-06
GO: 0051304	chromosome separation	81	5	0.23	3.2e-06
GO: 1901992	positive regulation of mitotic cell cycl…	88	5	0.25	4.9e-06
GO: 1902850	microtubule cytoskeleton organization in…	156	6	0.45	4.9e-06

Del resultado de la anotación génica y ver los genes sobreexpresados, obtenemos lo siguiente.

Con significación estadística vemos que, respecto a los genes sobreexpresados 44 total, hay un gran porcentaje que estan involucrados en la respuesta inmune. Era de esperar y tiene sentido obtener estos datos de pacientes con COVID-19.

A su vez hay un gran número de genes involucrados en procesos como la mitosiss y el ciclo celular (regulation of mitotic sister chromatida, mitotic cell cycle process, cytoskeleton organization).Este resultado es realmente interesante porque puede significar que el virus está alterando el ciclo célular para una mayor división y por lo tanto conseguir más cópias del virus y así aumentar la cantidad de tejido infectado y una mayor propagación. También podemos suponer que las células del sistema inmunitario, que ya hemos visto que aumenta la respuesta, están más activas y por lo tanto aumentando el número celular para hacer frente a la respuesta (desde limfócitos T hasta B o células plasmáticas).

11.1 REVIGO

La plataforma Revigo resulta una opción visual para observar la organización de la clasificación de la función biológica obtenida anteriormente.

De la misma manera que hemos mencionado antes, visualizamos la próximidad entre funcioes relacionadas y el peso que ocupan. Destacando la división celular y la respuesta inmune.

12 Conclusión

En este estudio hemos analizado un subgrupo de muestras del conjunto de datos de McClain et al., evaluando el perfil de expresión diferencial bajo distintas condiciones clínicas.

Tras el preprocesamiento y filtrado de los datos, el Análisis de Componentes Principales (PCA) y el clustering jerárquico revelaron que casi el 50% de la variabilidad de las muestras queda explicada por las dos primeras componentes. Estas técnicas de exploración permitieron validar la calidad de las muestras y confirmar que los grupos biológicos eran el principal factor de diferenciación, permitiendo así definir una matriz de contrastes robusta.

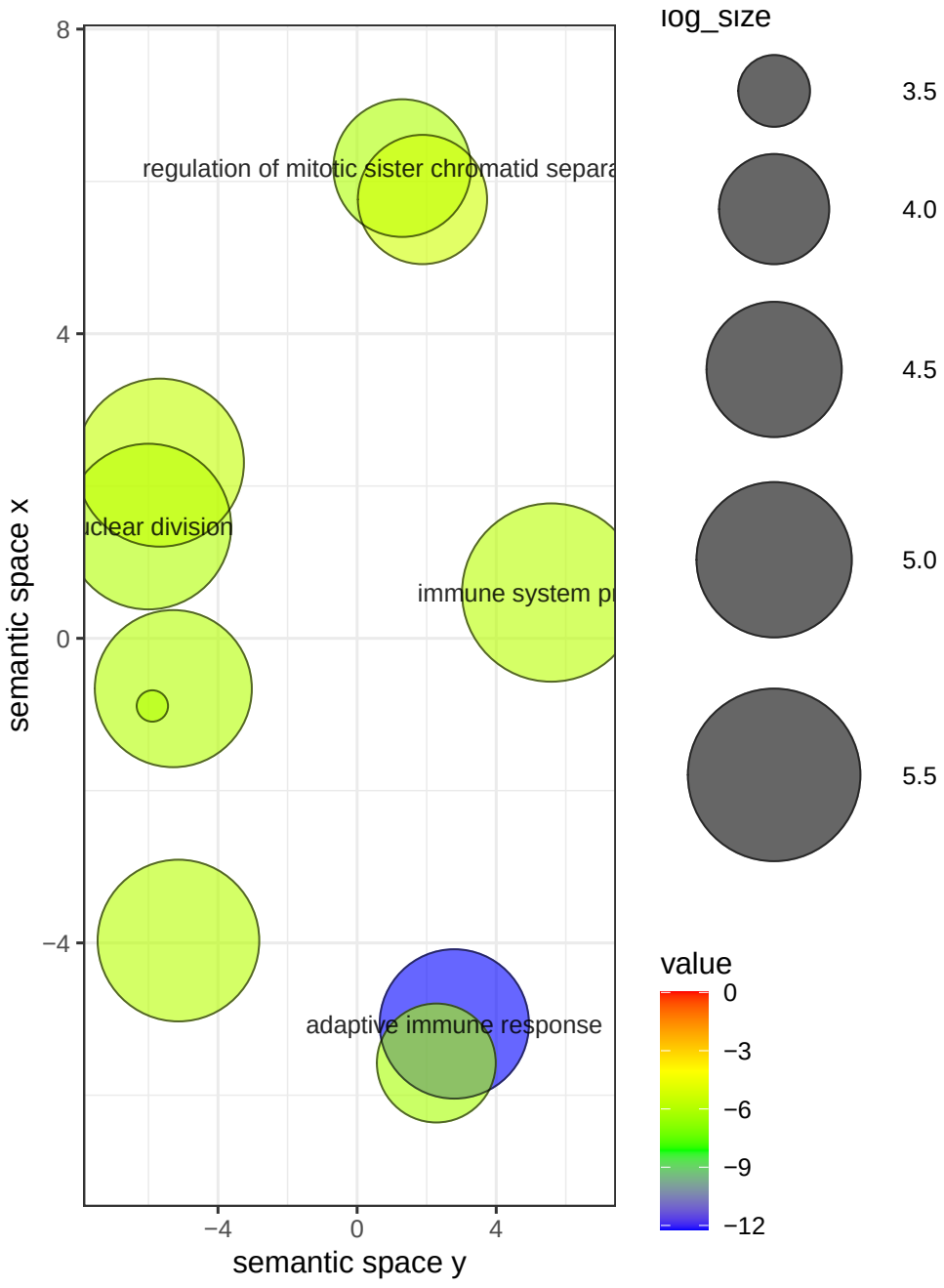


Figure 1: Visualización de la redundancia de términos GO mediante REVIGO

Se realizó un análisis de expresión diferencial comparando pacientes con COVID-19 y pacientes con infección bacteriana, utilizando en ambos casos individuos sanos como grupo control. Los resultados mostraron una respuesta transcriptómica significativa en ambos contrastes, aunque el número de genes diferencialmente expresados (DEGs) fue notablemente menor en el grupo de COVID-19. En ambas patologías se observó una tendencia hacia la infraexpresión génica (down-regulation) frente a la sobreexpresión.

Finalmente, el análisis de sobrerepresentación funcional de los genes sobreexpresados en pacientes con COVID-19 identificó una firma biológica clara. Los procesos biológicos enriquecidos se centran principalmente en la respuesta inmune adaptativa, la regulación del ciclo celular y procesos de división nuclear, sugiriendo una activa expansión clonal y una respuesta coordinada del sistema inmune frente a la infección viral.