Universitat Oberta de Catalunya

uoc.edu

Análisis transcriptómico en pacientes con infección del virus SARS-CoV-2. Identificación de asociación de perfiles de riesgo/protectores frente a la enfermedad

Israel David Duarte Herrera
Máster en Bioinformática y Bioestadística
Área 4 - Subárea 4: Análisis de datos ómicos

Israel David Duarte Herrera

Consultor:
Guillem Ylla Bou
Profesor responsable de la asignatura:
Antoni Pérez Navarro
Tutor externo:
Juan Gómez De Oña

Diciembre 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis transcriptómico en pacientes con infección del virus SARS-CoV-2. Identificación de asociación de perfiles de riesgo/protectores frente a la enfermedad*
Nombre del autor:	*Israel David Duarte Herrera*
Nombre del consultor/a:	*Guillem Ylla Bou*
Nombre del PRA:	*Antoni Pérez Navarro*
Fecha de entrega (mm/aaaa):	*12/2021*
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*TFM-Bioinformática y Bioestadística Area 4 - Subárea 4: Análisis de datos ómicos*
Idioma del trabajo:	*Castellano*
Palabras clave	*RNA-seq, NGS, COVID-19*

Resumen del Trabajo

El SARS-CoV2 es un virus de la familia *Coronaviridae*, causante de la enfermedad COVID-19, capaz de producir un síndrome respiratorio agudo severo.

Este agente, es el responsable de la situación de pandemia mundial declarada en el primer trimestre del 2020 y que aún sigue vigente, en la que se han infectado más de 277 millones de personas en todo el mundo, falleciendo más de 5 millones como causa directa de la enfermedad.

Se ha establecido hasta el momento una serie de variables clínicas relacionadas con un peor pronostico de esta patología (HTA, dislipemia, sedentarismo, etc) y variables genéticas, tales como genes codificantes de HLA III, HLA-A, HLA-B, HLA-C, IFN, TNFα, IL-6 e IL-8 entre otros.

En este trabajo se estudian los cambios en la expresión generada en el transcriptoma de pacientes infectados por SARS-CoV2 y que han desarrollado diferentes grados de severidad del síndrome secundario a la infección. Se ha realizado un análisis mediante RNA-seq empleando diversas herramientas bioinformáticas.

Como resultado se obtienen 2042 genes diferencialmente expresados entre pacientes clasificados como grave (CV) frente a aquellos clasificados como críticos (UCI). De esta manera, encontramos algunos genes que pueden explicar distintos niveles de gravedad del COVID-19, tales como PGLYRP1, HDAC9 y FUT4 . Asimismo existen otros con potencial real para su futuro análisis: ABCF1, ABHD16A y IER3 entre otros.

Abstract (in English, 250 words or less):

SARS-CoV2 is a virus of the *Coronaviridae* family, which causes the COVID-19 disease, capable of producing a severe acute respiratory syndrome.
This agent is responsible for the global pandemic situation declared in the first quarter of 2020 and which is still in force, in which more than 277 million people have been infected worldwide, with more than 5 million dying as a direct cause. of the illness.

So far, a series of clinical variables related to a worse prognosis of this pathology (hypertension, dyslipidemia, sedentary lifestyle, etc.) and genetic variables, such as genes coding for HLA III, HLA-A, HLA-B, HLA- C, IFN, TNFα, IL-6 and IL-8 among others.

In this work we study the changes in the expression generated in the transcriptome of patients infected by SARS-CoV2 and who have developed different degrees of severity of the syndrome secondary to the infection. An RNA-seq analysis has been performed using various bioinformatics tools.

As a result, 2042 differentially expressed genes were obtained between patients classified as severe (CV) versus those classified as critical (UCI). In this way, we found some genes that can explain different levels of severity of COVID-19, such as PGLYRP1, HDAC9 and FUT4. There are also others with real potential for future analysis: ABCF1, ABHD16A and IER3 among others.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1. Contexto y justificación del Trabajo

El presente Trabajo de Fin de Máster (TFM) consiste en el análisis de datos ómicos procedentes de secuenciación masiva (NGS) obtenidos de pacientes con infección de SARS-COV-2, que han sido hospitalizados en el Hospital Universitario Central de Asturias (HUCA). Con el fin de buscar una posible relación entre las diversas variantes genéticas y la variabilidad fenotípica que muestran dichos pacientes.

Para ello se realizará análisis de transcriptoma mediante RNA-seq, con el objetivo de observar las posibles diferencias de expresión.

A pesar de los grandes avances en la adquisición de conocimiento científico que se ha desarrollado en los últimos meses, es destacable la actual carencia de explicación ante el desarrollo observado en la sintomatología y evolución de la infección en los distintos pacientes. Aún habiéndose encontrado diversas variables explicativas (HTA, dislipemia, etc) en este corto espacio de tiempo, se continúa apreciando un actual desconocimiento de gran parte de la variabilidad genotípica que podría explicar la diversidad observada en el desarrollo de la enfermedad.

1.2 Objetivos del Trabajo

Objetivos generales

1. Procesar los archivos fastq procedentes de la secuenciación masiva

2. Llevar a cabo un análisis de expresión diferencial y posterior enriquecimiento funcional

3. Consolidar conocimientos en el uso del lenguaje de programación R para el análisis de datos mediante RNA-seq

Objetivos específicos

1. Procesar los archivos fastq procedentes de la secuenciación masiva

    1.1. Realizar el control de calidad de los datos obtenidos tras la secuenciación mediante el uso de “FastQC”

    1.2. Realizar el trimado para ajustar los datos a los niveles de calidad deseados, se realizará mediante “BBduk”

    1.3. Llevar a cabo el alineamiento de la secuencia problema con una secuencia de referencia, agregando así contexto biológico a los datos. Este proceso se llevará a cabo mediante “Salmon”, debido a diversos factores, como la capacidad computacional

    1.4. Realizar la cuantificación, la cual consiste en el conteo del número de lecturas alineadas con una característica particular de interés

2. Llevar a cabo un análisis de expresión diferencial y posterior enriquecimiento funcional

2.1. Comprobar la expresión diferencial de los genes en ambo cohortes del estudio (divididos según severidad de la enfermedad)

2.2. Utilizar Rstudio para conocer dicha expresión diferencial. Para ello se utilizará el paquete de R/Bioconductor “DESeq2”

3. Consolidad conocimientos en el uso del lenguaje de programación R para el análisis de datos mediante RNA-seq

3.1. Realizar representaciones gráfica de los resultados obtenidos y de los datos durante el pipeline, tales como “PCA”, “Heatmap”, “Maplot”, “Volcanoplot”, etc

3.2. Comparar procedimientos realizados en la consola de LINUX con sus análogos en Rstudio

3.3. Llevar a cabo todos los análisis estadísticos necesarios para alcanzar una conclusión con sentido biológico

1.3. Enfoque y método seguido

Para alcanzar los objetivos previamente señalados, es de vital importancia comenzar con una detallada revisión bibliográfica, para ello se utilizará los diversos motores de búsqueda conocidos, tales como “NCBI” o “Scielo” entre otros. Ella se basará en recopilar la información actual sobre aquellas variables predictivas de severidad de la infección por SARS-COV-2, pero también en la búsqueda de los métodos más apropiados para llevar a cabo el mencionado flujo de trabajo necesario para la realización del RNA-seq que se pretende realizar.

Los datos que usaremos han sido extraídos de los archivos “fastq” procedentes de la secuenciación masiva realizada en el laboratorio de genética molecular del HUCA. Con ellos se estudiará el transcriptoma, utilizando en una primera fase el entrono Linux para efectuar un control de calidad y procesamiento de secuencia (tal y como se ha comentado en la sección de objetivos).
En un segundo tiempo se llevará a cabo un análisis estadístico y de expresión diferencial con R/Biocondutor empleando diferentes paquetes como “DESeq2”, “GenomicRangers”, “GenomicFeatures”, “ShortRead”, etc.

Los métodos de análisis propuesto se basan principalmente en el interés por fijar y poner en prácticas aquellos conocimientos y recursos impartidos durante el desarrollo del presente Máster, así como por las limitaciones computacionales actuales, haciendo poco probable poder aplicar algunos de los métodos de interés, como ocurriría en el uso de “STAR” para realizar el alineamiento de la secuencia problema.

1.4. Planificación del Trabajo

Tareas

1. Realizar la pertinente búsqueda bibliográfica para recopilar las variables que actualmente han sido identificadas como predictoras, y para realizar aquellos códigos y scripts necesarios para el posterior desarrollo analítico

1.1. Búsqueda bibliográfica sobre conceptos básicos en transcriptómica, realización y usos prácticos.

1.2. Búsqueda bibliográfica sobre los diferentes flujos de trabajo aplicados en Linux para realizar el preprocesado y procesado de archivos fastq, y posterior elección del más adecuado según diversos parámetros como comprensión, capacidad computacional, etc

1.3. Búsqueda bibliográfica sobre los diferentes métodos de análisis de expresión diferencial de interés, tras ello se escogerá el que suscite mayor inclinación y se procederá a realizar el pertinente script, necesario para el subsiguiente análisis

2. Recopilación de archivos y diseño experimental, Planta vs UCI

2.1. Recopilar todos los archivos necesarios para el trabajo. Esto se hará mediante el envío online, utilizando plataformas de envío gratuito que mejor se adapten al tamaño de los archivos (Smash)

2.2. Establecer el diseño experimental en el que se basará el trabajo, indicando los cohortes y variables a utilizar

3. Preprocesar y procesar las secuencias una vez se ha establecido los códigos a utilizar en Linux, a su vez, realizar comparativa de alguno de estos análisis con su análogo en R/Bioconductor

3.1. Realizar el control de calidad de los archivos sin procesar, mediante el uso de “FastQC” y “MultiQC”, donde se analizarán diferentes características tales como: contenido GC, contenido de N bases, secuencias duplicadas, etc

3.2. Realizar el trimado cuando sea necesario, con el objetivo de eliminar aquellas secuencias que no cumplen un mínimo de calidad establecido

3.3. Una vez obtenido el archivo con la calidad deseada, debido a las limitaciones computacionales, se opta por realizar un “quasi-mapping” a través de Salmon. Obviando así el proceso de cuantificación de archivos “Bam” con el uso de un referente GFF/GTF. Pues tras el proceso aplicado mediante Salmon se obtendrán los correspondientes archivos “quant.sf”.

3.4. Tras obtener los archivos “quant.sf” deseados, se procederá a realizar el conteo. Siendo el método de elección aquel que implica el uso del paquete de R/Bioconductor “tximport”

4. Estudio de la expresión diferencial mediante el uso de R/Bioconductor e interpretación biológica de los resultados

4.1. Control de calidad y normalización de los datos

4.2. Análisis exploratorio de los datos y transformación estabilizadora de varianza

4.3. Visualización de los datos (Heatmap, PCA, etc)

4.4. Análisis de la expresión diferencial

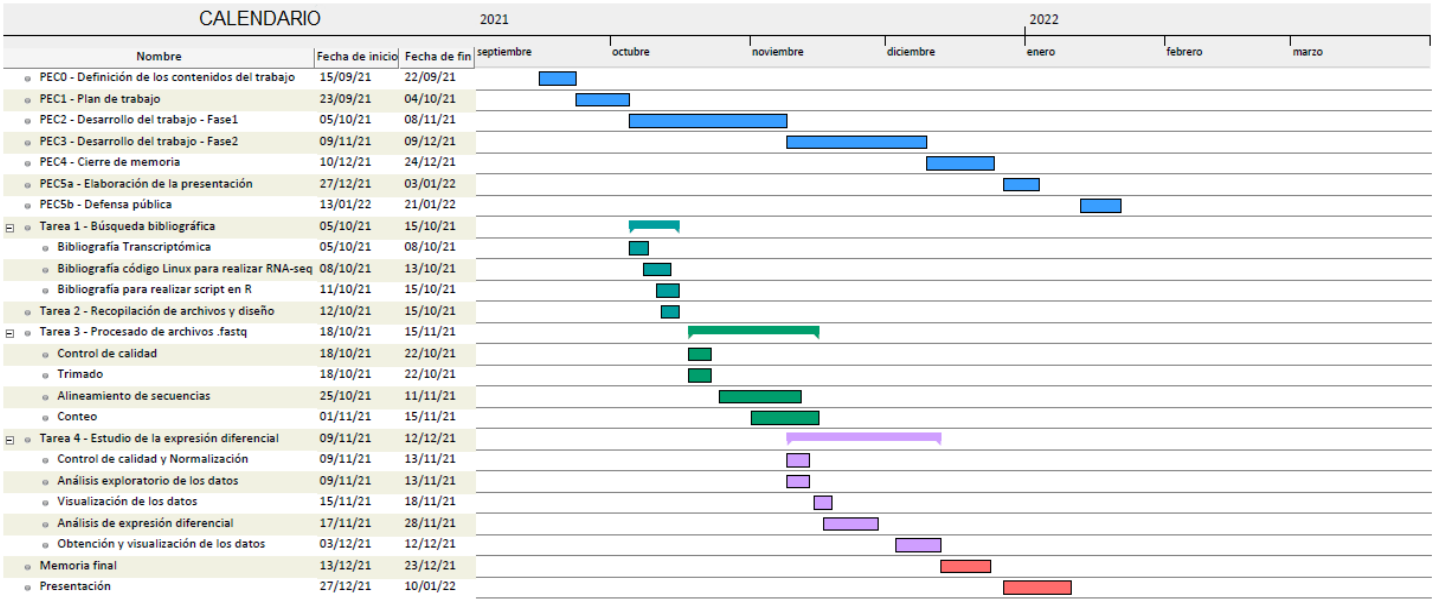
4.5. Obtención y visualización de los resultados

5. Elaboración de memoria final

6. Realización de la presentación para la posterior defensa pública

Calendario

Se muestra a continuación un diagrama de Gantt (realizado mediante el software "GanttProject"), en el que se representa visualmente la temporalidad de cada una de las tareas previstas en el desarrollo del presente TFM, indicándose en él las fechas de inicio y fin de cada una de dichas tareas, además de las pruebas de evaluación continua (PECs)

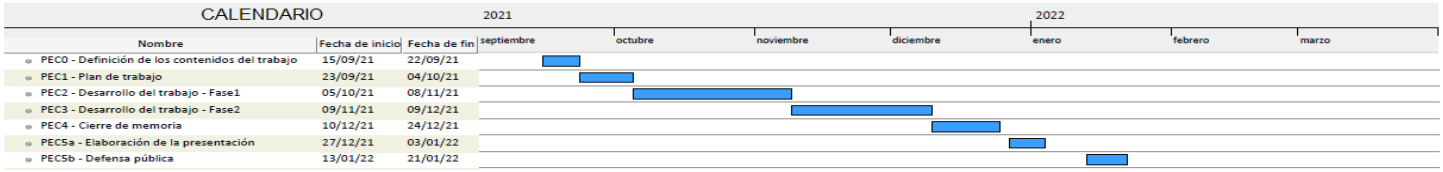


Nombre	Fecha de inicio	Fecha de fin
PEC0 - Definición de los contenidos del trabajo	15/09/21	22/09/21
PEC1 - Plan de trabajo	23/09/21	04/10/21
PEC2 - Desarrollo del trabajo - Fase1	05/10/21	08/11/21
PEC3 - Desarrollo del trabajo - Fase2	09/11/21	09/12/21
PEC4 - Cierre de memoria	10/12/21	24/12/21
PEC5a - Elaboración de la presentación	27/12/21	03/01/22
PEC5b - Defensa pública	13/01/22	21/01/22
Tarea 1 - Búsqueda bibliográfica	05/10/21	15/10/21
Bibliografía Transcriptómica	05/10/21	08/10/21
Bibliografía código Linux para realizar RNA-seq	08/10/21	13/10/21
Bibliografía para realizar script en R	11/10/21	15/10/21
Tarea 2 - Recopilación de archivos y diseño	12/10/21	15/10/21
Tarea 3 - Procesado de archivos .fastq	18/10/21	15/11/21
Control de calidad	18/10/21	22/10/21
Trimado	18/10/21	22/10/21
Alineamiento de secuencias	25/10/21	11/11/21
Conteo	01/11/21	15/11/21
Tarea 4 - Estudio de la expresión diferencial	09/11/21	12/12/21
Control de calidad y Normalización	09/11/21	13/11/21
Análisis exploratorio de los datos	09/11/21	13/11/21
Visualización de los datos	15/11/21	18/11/21
Análisis de expresión diferencial	17/11/21	28/11/21
Obtención y visualización de los datos	03/12/21	12/12/21
Memoria final	13/12/21	23/12/21
Presentación	27/12/21	10/01/22

Figura 1. Diagrama de Gantt

Hitos

A lo largo de este trabajo se ha diseñado y organizado una serie de tareas con una duración determinada. Aunque la intención de dicha organización es la de establecer las fechas en las que se debe finalizar cada uno de los objetivos, se define como hitos aquellas actividades de mayor relevancia que servirán como guía del seguimiento del plan previsto.
Estos hitos son los relativos a la presentación de las PECs, pues se consideran un óptimo indicativo del correcto seguimiento del trabajo. Teniendo en cuenta la inamovilidad de las fechas de presentación, al igual que el feedback recibido por el consultor para cada una de ellas, convirtiéndose así, en el mejor indicador de progreso del TFM.
Los mencionados hitos quedan señalados en el diagrama de Gantt adjunto (PEC0 – PEC5b).



Nombre	Fecha de inicio	Fecha de fin
PEC0 - Definición de los contenidos del trabajo	15/09/21	22/09/21
PEC1 - Plan de trabajo	23/09/21	04/10/21
PEC2 - Desarrollo del trabajo - Fase1	05/10/21	08/11/21
PEC3 - Desarrollo del trabajo - Fase2	09/11/21	09/12/21
PEC4 - Cierre de memoria	10/12/21	24/12/21
PEC5a - Elaboración de la presentación	27/12/21	03/01/22
PEC5b - Defensa pública	13/01/22	21/01/22

Figura 2. Hitos

Análisis de riesgos

Riesgo 1: Insuficiencia en la capacidad computacional. Se trabaja con un equipo casero de escasa memoria RAM, lo cual podría poner en riesgo la realización de alguno de los objetivos, principalmente aquellos relacionados con el procesado de archivos “fastq”. Suponiendo así la adquisición de un nuevo equipo (1200 – 1500 €) o el alquiler de un servidor VPS (140 – 160 €/mes)

Riesgo 2: Escaso tiempo de realización. Se pretende realizar una serie de análisis en los que se cumpla el proceso completo desde que se genera el archivo procedente de la secuenciación masiva hasta su interpretación biológica. Existe el riesgo de imposibilidad de realizar tal volumen de trabajo en el tiempo establecido, tendiendo así que reducir el número de pacientes de los cohortes.

Riesgo 3: Proceso de aprendizaje y decisión de métodos. Para cada una de las partes del flujo de trabajo propuesto hay que decidir el método más adecuado con los que realizar los diferentes análisis. Esto supone un previo proceso de consolidación de conocimientos y pruebas que se podrían dilatar en el tiempo.

Riesgo 4: Imprevistos no programados. Como en cualquier proceso analítico, existe la posibilidad de que surjan problemas no esperados (en la metodología, análisis, etc) o resultados con escaso interés biológico.

1.5. Breve sumario de contribuciones y productos obtenidos

Al finalizar este TFM se espera obtener una serie de entregables y resultados, los cuales se indican a continuación

Plan de trabajo

El plan de trabajo viene representado en el presente informe. Su ejecución supondrá la entrega de siete pruebas de evaluación continua, junto a una serie de documentos que se serán realizados y adjuntados a esta memoria final. Se resume a continuación en la siguiente tabla.

Entregable	Nombre	Descripción
PEC 0	Definición de los contenidos del trabajo	Definir la temática del trabajo, justificar su interés y/o relevancia e indicar la finalidad del TFM
PEC 1	Plan de trabajo	Definir líneas generales del TFM, incluyendo objetivos, temporalidad y riesgos
PEC 2	Desarrollo del trabajo (Fase 1)	Realizar un seguimiento del desarrollo del trabajo realizado hasta este momento
PEC 3	Desarrollo del trabajo (Fase 2)	Realizar un seguimiento del desarrollo del trabajo realizado hasta este momento
PEC 4	Cierre de memoria	Entrega del trabajo realizado
PEC 5a	Elaboración de la presentación	Realización de la presentación que posteriormente será expuesta
PEC 5b	Defensa pública	Presentación y defensa del TFM ante un tribunal
Anexo códigos Linux	Códigos Linux	Recopilación de todos los códigos utilizados en el procesado de archivos (calidad, alineamiento, etc)
Anexo scripts R	Scripts R	Recopilación de los scripts en R que han sido elaborados durante el desarrollo del TFM (expresión diferencial, enriquecimiento, etc)

Tabla 1: Plan de trabajo

Memoria

La memoria final consiste en un ensayo completo sobre el procedimiento llevado a cabo durante el desarrollo del presente TFM. En esta se pretende explicar cada uno de los pasos ejecutados para los diferentes análisis al igual que plasmar los resultados obtenidos. Igualmente, consiste en una exposición bibliográfica y en la discusión de las conclusiones alcanzadas.

Está compuesta por la siguiente estructura:

Estructura de la Memoria
1. Introducción
2. Estado del arte
3. Metodología
4. Resultados
5. Discusión
6. Conclusiones
7. Glosario
8. Bibliografía
9. Anexos

Tabla 2: Estructura de la memoria

Presentación virtual

La presentación virtual consistirá en un video de un máximo de veinte minutos donde se mostrarán aproximadamente veinte diapositivas.

En ella se expondrá de forma resumida el presente trabajo, junto a sus resultados y conclusiones.

Para la creación de la presentación se utilizará un software de edición de video, Camtasia.

1.6. Breve descripción de los otros capítulos de la memoria

El presente TFM tiene la estructura representada en la tabla 2, siendo una breve descripción del contenido la que se muestra a continuación

Contenido	Descripción
1. Introducción	Visión general del trabajo a realizar, indicando objetivos y tareas a desarrollar durante el proceso
2. Estado del arte	Revisión bibliográfica del contexto biológico del TFM, se expondrá todos los conceptos que posteriormente se trabajarán (NGS, transcriptómica, etc) junto al conocimiento actual de las variables predictoras de gravedad en la infección por SARS-COV-2
3. Metodología	Se desarrollará todos los análisis llevados a cabo, desde la obtención de archivos, pasando por su procesado y finalizando con el análisis de expresión diferencial
4. Resultados	Exposición de los resultados obtenidos en el desarrollo del trabajo
5. Discusión	Significación biológica de los resultados obtenidos (ej: genes diferencialmente expresados según cohorte)
6. Conclusiones	Deducciones obtenidas tras la discusión de resultados
7. Glosario	Definiciones de conceptos empleados
8. Bibliografía	Recopilación de recursos utilizados (*papers*, manuales, etc)
9. Anexos	Colección de códigos, figuras e imágenes que no han sido incluidos en el cuerpo de la memoria y que han sido claves para su desarrollo y/o interpretación

Tabla 3: Estructura del proyecto

2. Estado del arte

2.1. Introducción histórica

El COVID-19 es una enfermedad infecciosa causada por el virus SARS-COV-2, virus perteneciente a la familia de los coronavirus (*Coronaviridae*).

Desde diciembre de 2019 se han contabilizado más de 277 millones de infectados (John Hopkins University [JHU], Diciembre 2021) a nivel mundial, con más de 5 millones de fallecimientos, convirtiéndose en una pandemia con importantes consecuencias sanitarias y social-económicas.

El origen de la pandemia se remonta a la ciudad de Wuhan (China), donde se identifican los primeros contagios a mediados del mes, siendo la mayoría de estos pacientes trabajadores en un mercado de dicha ciudad.

A principios del mes siguiente, la OMS informa sobre la presencia de neumonía de causas desconocidas en dichos individuos, tras haberse descartado infección por otros agentes víricos ya conocidos (SARS, gripe, etc).

Es en enero cuando el gobierno chino anuncia el aislamiento y posterior secuenciación del genoma vírico, tras lo que laboratorios de todo el mundo produjeran pruebas diagnósticas con el uso de PCR.

En esos momentos, la OMS baraja diversas hipótesis sobre el origen del virus, aunque descartando cualquier otro origen diferente a la transmisión zoonótica. Entre dichas teorías destacan las siguientes (Organización Mundial de la Salud [OMS], Diciembre 2021):

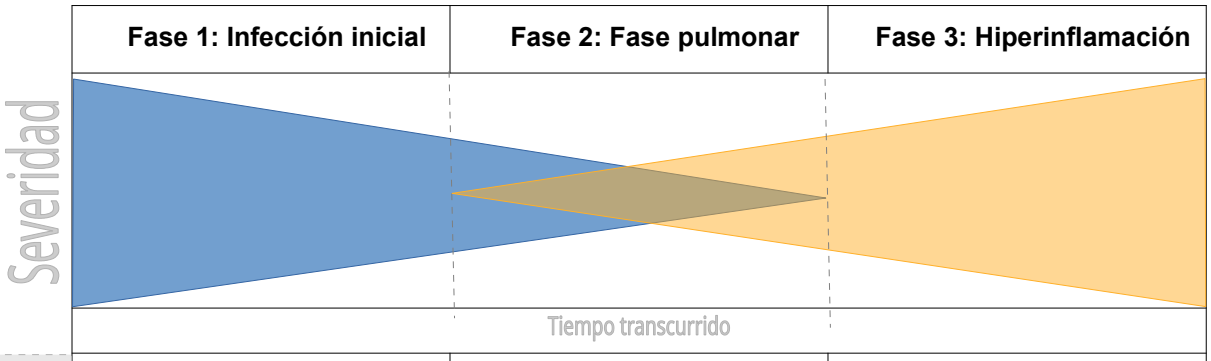
- animal – humano
- murciélago – animal intermedio – humano
- alimentos – humano

En ese mismo mes se localiza el primer caso de infección fuera de fronteras chinas, y poco después se informa de la presencia de nuevos casos de infección en otros países, con alta afluencia en Europa.

Finalmente, en Marzo de 2020 la OMS declara la situación de Pandemia.

2.2. Marco clínico

Ambos virus SARS-CoV (1 y 2), ingresan en la célula a través del receptor de la enzima convertidora de angiotensina 2 (ACE2). El SARS-CoV2 primero infecta las vías respiratorias inferiores, uniéndose a ACE2 en las células epiteliales alveolares (Jiang et al. 2019).
Una vez unido, una de las características más destacables del virus es su capacidad para inducir la producción de citocinas inflamatorias, pudiendo provocar la conocida “tormenta de citoquinas”, proceso causante de daño orgánico. Estos pasos conforman lo que se conoce como fases clínicas de la infección (Liu et al. 2020).



Síntomas	Síntomas constitucionales leves, anosmia, fiebre >38º, tos seca, diarrea	Disnea	Síndrome de insuficiencia respiratoria aguda (SIRA), Sepsis, lesión renal aguda, insuficiencia cardíaca aguda
Sígnos	Linfopenia, aumento leve de DHL, aumento dimero-D	SatO2 baja, radiografía anormal de tórax, aumento de: dimero-D, DHL y transaminasas	Aumento de: ferritina, dimero-D, proteína C reactiva, DHL, troponinas, creatitina y tiempos de coagulación

Tabla 4: Fases cínicas de infección por SARS-CoV2.

2.3. Marco inmunológico

De forma general y esquematizada, el sistema inmune humano está formado por la inmunidad innata y la inmunidad adquirida o adaptativa.

La inmunidad adquirida es la que se obtiene a lo largo de la vida (vacunas, infecciones, etc). Siendo su opuesta la inmunidad innata, que es aquella con la que se nace, formada por un conjunto de células que actúan como respuesta a la entrada de un patógeno (células NK, mastocitos, eosinófilos, células fagocíticas, etc).

Inmunidad innata

Es la primera línea de defensa del organismo, confiere inmunidad general (no específica) ante una agresión externa como es la entrada de un microorganismo patógeno.

En contraposición a la inmunidad adquirida, ésta no aumenta o mejora en sucesivos contactos con el patógeno. Su funciones principales son:

- Reclutamiento de células inmunes (producción de citoquinas)
- Identificación y eliminación de sustancias extrañas en el organismo
- Activación del sistema del complemento
- Inhibición directa de la replicación viral
- Activación de la respuesta adquirida (presentación de antígenos)

Inmunidad adquirida

Inmunidad mediada por células

La inmunidad mediada por células está regulada por linfocitos T, los cuales actúan reconociendo antígenos adheridos a la superficie de las células, como ocurre en el caso de células fagocitarias con microorganismos ingeridos, las cuales liberan antígenos hacia su superficie celular (al complejo mayor de histocompatibilidad, MHC), lugar donde los linfocitos T realizan el reconocimiento.

Existen diversos tipos de linfocitos T y de MHC, que de forma sintetizada se exponen en la siguiente tabla.

Tipo de linfocito T	Función	Interacción con MHC
Citotóxicos	Identifican y atacan antígenos virales que encuentran sobre las células infectadas	MHC-1 ---- CD8+
		MHC2 ---- CD4+, T_H1
Colaboradores	Identifican antígenos encontrados sobre la superficie de células presentadoras de antígenos (macrófagos, células B, etc). Tras esto, liberan *Interleucinas*	
De memoria	Identifican al antígeno en exposiciones sucesivas, disminuyendo los tiempos de respuesta	
Supresores	Disminuye la concentración de anticuerpos una vez eliminado el patógeno	

Tabla 5: Esquema sintetizado de inmunidad mediada por células

Inmunidad humoral

Al contrario de lo visto en la inmunidad celular, en la inmunidad humoral no son las células las encargadas de interaccionar con los patógenos, sino *anticuerpos* y el *sistema del complemento*.

Una vez los linfocitos T han “reconocido” al patógeno, serán los colabores los encargados de realizar la presentación a los linfocitos B. Mediante dicha unión, los linfocitos B se diferencian y realizan las siguientes funciones:

- Secreción de anticuerpos (inicia con IgM)
- Cambio de isotipo (IgG, IgA, etc)
- Maduración a anticuerpos de alta afinidad
- Conversión en linfocitos B de memoria

Dicha respuesta humoral está formada por dos fases, la primaria y la secundaria, durante la fase primaria se produce la liberación de anticuerpos y la posterior clonación de las células plasmáticas. Estos anticuerpos son mayoritariamente IgM, estando los IgG ausentes o en baja concentración.

En una respuesta humoral secundaria, la presencia del mismo antígeno produce la activación de los linfocitos de memoria (creados como consecuencia de la respuesta humoral primaria), reduciendo drásticamente el tiempo de respuesta frente al tiempo de respuesta ocurrido ante el primer contacto.

En esta fase, los IgM se encuentran ausentes. Siendo los IgG los anticuerpos predominantes (en ocasiones se encuentran IgA e IgE).

Por otro lado, dentro de la respuesta humoral se encuentra el Sistema del Complemento, que consiste en una cascada enzimática cuyas funciones principales son la lisis de células marcadas con anticuerpos y la inflamación.

Muchas de estas proteínas del complemento se encuentran en el suero como precursores enzimáticos inactivos (cimógenos). Otras sin embargo se establecen en las superficies celulares.

Existen tres vías de activación del complemento, la *clásica*, *de la lectina* y la *alternativa*. Sus componentes son proteínas nominadas de C1 a C9. Todas ellas implicadas en alguna de las mencionadas vías, las cuales tienen objetivos específicos diferentes (reconocimiento, inflamación, etc).

Inmunidad ante SARS-CoV2

Tal y como se ha expuesto anteriormente, existen diversas líneas de respuesta inmunitaria ante la presencia de un microorganismo patógeno. A renglón seguido se realiza una exposición general de los conocimientos actuales sobre la inmunidad ante la infección de SARS-CoV2.

Como primera línea de respuesta, una vez introducido el virus en el organismo, son los macrófagos y las células NK las encargadas de la primera fase de actuación. Los primeros fagocitan al virus, tras lo que sus lisosomas realizan la degradación del material vírico. Los restos resultante de esta degradación serán utilizados por el sistema inmune como herramienta para una posible infección posterior.

Si estas células son capaces de hacer frente a la infección, el cuadro clínico más probable es el de un paciente asintomático.

Si por el contrario, esta barrera es insuficiente, se produce la liberación de citoquinas (Tabla 6), acción que conlleva la producción de más células secretoras de más citoquinas.

A su vez, estas citoquinas estimulan la producción de ferritina (liberada al torrente sanguíneo) y de Proteína C Reactiva a nivel hepático (lo que conlleva a una apoptosis de células infectadas). Tanto la proteína C reactiva como la ferritina son reactantes de fase aguda, lo que permite monitorizar la evolución de la enfermedad.(Tabla 4).

Por otro lado, dentro de la primera fase de actuación se encuentra las células NK, encargadas de localizar y eliminar células infectadas.

Como segunda línea de respuesta se encuentran las células dendríticas, encargadas de realizar la presentación entre el virus y los linfocitos T colaboradores (CD4+). Éstos liberan citoquinas que activan a los linfocitos T citotóxicos (CD8+), encargados de la eliminación de células infectadas.

A su vez, actúan los linfocitos B, activados por las citoquinas liberadas o por contacto con el virus/antígeno.

Cuando un linfocito B se liga al virus, se inicia un clonado masivo de éstos, que se transforman posteriormente en células plasmáticas, cuyo destino serán los diferentes tejidos del organismo. Allí liberan anticuerpos, principalmente IgM (los 9-11 días post infección) e IgG (14-21 días post infección). El papel de estos anticuerpos es el de unirse al virus (a través de proteínas de la espícula) para convocar diversas células encargadas de su eliminación (como los macrófagos).

Estos anticuerpos también colaboran con las células NK mencionadas, facilitando su unión al virus y su posterior eliminación.

Los casos críticos de la enfermedad se imputan a la generación de tormentas de citoquinas, produciéndose un aumento de proteínas inflamatorias en el organismo.

Se muestra a continuación algunas moléculas identificadas en la diversa bibliografía como las más relevante en la infección por SARS-CoV2.

Elementos implicados en la infección por SARS-CoV2	
IFN-I Interferón (citoquinas)	Considerado el más importante en la infección por SARS-CoV2 (Vabret et al. 2020), observándose su implicación en la severidad de la enfermedad.
TNF-α, IL-1β, IL-1RA, IL-2RA, IL-6, IL-7, IL-8, IL-9, IL-10, IP-10, MCP-1, MIP-1a, MIP-1b, citoquinas	Se liberan juntos, potenciando la respuesta inmune adaptativa (Shuibing et al. 2020).
CCL2, CCL7, CXCL9 quimiocinas (citoquinas)	Reclutamiento de monocitos
IgM, IgG anticuerpos	Inmunidad humoral
IFITMs proteína de transmembrana/ receptores	Implicadas en la evolución de la enfermedad
TLRs , MDA5 Receptor de reconocimiento de patrones	Implicación en la severidad de la enfermedad

Tabla 6: Elementos implicados en la infección por SARS-CoV2

Según lo indicado, se ha optado por la realización de los objetivos previamente descritos, pues con esta base, se ha creado la hipótesis de la presencia de diversas variantes genéticas que impliquen un proceso infeccioso diverso en cuanto severidad.

A pesar de haberse establecido un importante número de genes implicados, se presupone la presencia de más genes explicativos. Por ello, se lleva a cabo una análisis de expresión diferencial entre los grupos descritos anteriormente.

2.4. Marco bioinformático

Bioinformática funcional

Los últimos años han supuesto un importante incremento en la información y el volumen de datos aportados por el ámbito biomédico. Para explorar dicha información es necesario la aplicación de técnicas computacionales que permitan la extracción, manipulación e interpretación de datos biológicos.

Dentro del contexto de este trabajo, es relevante la alusión del estudio de las ciencias ómicas, las cuales incluyen genómica, transcriptómica, proteómica y metabolómica entre otras.

Ciencias ómicas

Este término hace referencia al estudio de los sistemas celulares a un nivel específico. Éstas ciencias tienen una amplia variedad de aplicaciones como la industria farmacéutica, el diagnóstico clínico, la investigación biomédica, etc.

Transcriptómica

La transcriptómica es el estudio del transcriptoma de un organismo, es decir, la suma total de las transcripciones de ARN que en él existe.

Las tecnologías de transcriptómica proporcionan una amplia descripción de los procesos celulares activos e inactivos, por lo que se trata de técnicas cuya importancia radica en poder comprender cómo un mismo genoma puede dar lugar a diferentes tipos de células, y cómo se regula la expresión génica.

El entendimiento del transcriptoma no solo es esencial para la interpretación de elementos funcionales del genoma, sino también para comprender el origen y desarrollo de diversas enfermedades.

El estudio del transcriptoma se lleva a cabo con el uso de diversas tecnologías (ej: microarrays), pero es la secuenciación masiva (NGS) la que suscita especial interés para el desarrollo de esta memoria.

NGS: Next Generation Sequencing

NGS o secuenciación masiva, es una tecnología utilizada para determinar el orden de los nucleótidos en genomas completos o regiones específicas de ADN o ARN, mejorando notablemente el proceso de secuenciación original de *Sanger*.

Existen diversas tecnologías de secuenciación masiva, siendo las tres más utilizadas Illumina, *Roche 454* e *Ion torrent* (*Anexo 2*). Siendo ésta última la utilizada en la obtención de los datos presentes en este análisis.

Gracias al desarrollo de estas nuevas tecnologías de secuenciación se pudo originar un nuevo método para la cuantificación de transcriptomas, el **RNA-seq**, el cual posee fuertes ventajas frente a aproximaciones preexistentes como los microarrays.

3. Metodología

Se expone a continuación cada uno de los pasos desarrollados para la elaboración de el presente trabajo

3.1. Datos a integrar

El primer paso del proceso es la extracción de ARN de la sangre de los pacientes ingresados, tanto en planta como en UCI. Dicho protocolo está expuesto en el *Anexo 1*.
Los datos utilizados están compuestos por veinte pacientes ingresados en el Hospital Universitario Central de Asturias (HUCA). Estos están divididos a partes iguales según el nivel de gravedad de la enfermedad secundaria a la infección.

Para llevar a cabo dicha clasificación se ha utilizado la bibliografía presente y las escalas ya establecidas, en las que se consideran los siguientes criterios:

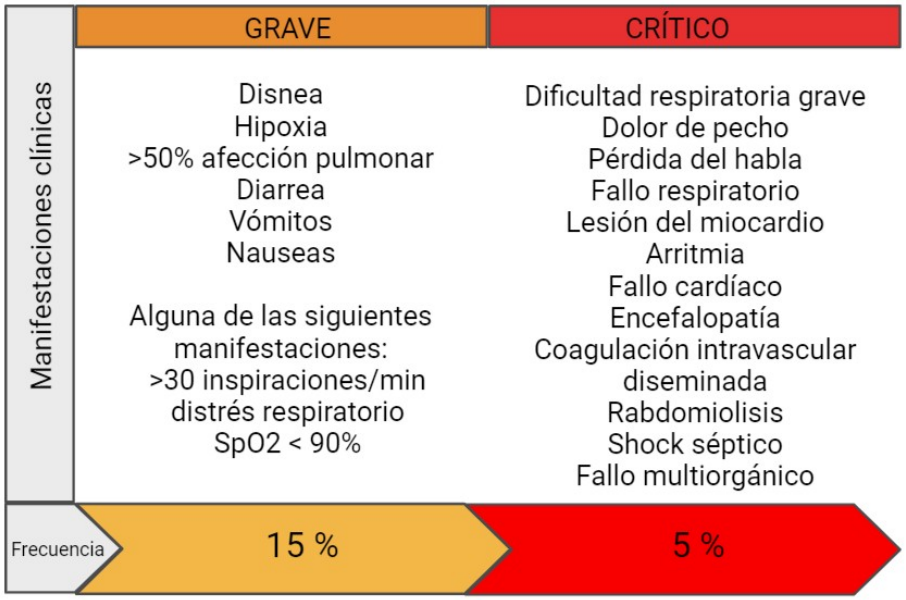


Figura 3. Evolución clínica de COVID-19. Data (1) and Clínical management of COVID-19, interim guidance (2). Creado con BioRender.com

De esta forma, tenemos los dos grupos presentes CV (grave) y UCI (crítico).

Muestras	Condición
CV-1	CV
CV-2	CV
CV-3	CV
CV-4	CV
CV-5	CV
CV-6	CV
CV-7	CV
CV-8	CV
CV-9	CV
CV-10	CV
UCI-1	UCI
UCI-2	UCI
UCI-3	UCI
UCI-4	UCI
UCI-5	UCI
UCI-6	UCI
UCI-7	UCI
UCI-8	UCI
UCI-9	UCI
UCI-10	UCI

Tabla 7: Estructura de los datos

3.2. Control de calidad

Una vez escogidos los datos que serán utilizados para el desarrollo del trabajo, el primer paso consistirá en determinar su calidad. Para ello, emplearemos el programa **FastQC**, estudiando uno por uno y de forma independiente los archivos disponibles, así como **MultiQC** con el fin de obtener una visión global de la calidad de nuestro dataset (*Anexo 3*).

La mayoría de secuenciadores generarán un informe de control de calidad como parte de su *pipeline*, sin embargo, generalmente se enfoca en la identificación de problemas generados por el propio secuenciador. FastQC nos proporcionará un informe de control de calidad, detectando tanto errores originados en el secuenciador, como los originados en la biblioteca de partida.

El análisis de calidad refierido anteriormente y elaborado con FastQC (V. 0.11.9) utiliza como input los archivos fastq de origen. Tras la ejecución del programa sobre dichos archivos, se genera un informe en html para cada uno de éstos en la que se resumen los datos, el cual estará compuesto por los siguientes apartados:

- Basic Statistics
- Per Base Sequence Quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per Base N Content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Una vez ultimado este proceso, se utiliza MultiQC para la obtención de un informe final, a partir del cual podamos inferir la calidad de la secuencia de la totalidad de archivos de forma global.

3.3. Trimado

El trimado consiste en el recorte de lectura, con el objetivo de eliminar las porciones de baja calidad mientras se conserva la parte más larga de alta calidad de una lectura NGS, aumentando la calidad y confiabilidad del análisis. De esta manera observaremos ganancias en términos de tiempo de ejecución y recursos computacionales necesarios.

La existencia de base de baja calidad puede suponer un problema potencial para cualquier análisis NGS, puesto que se corre el riesgo de agregación de secuencias poco confiables o aleatorias al conjunto de datos. Esto puede ser motivo de problemas en el canal de análisis ulterior, dando lugar a interpretaciones falsas de datos (Del Fabro et al. 2013)..

Para realizar este recorte de lectura utilizaremos aquella herramienta necesaria para tal fin, la cual encontraremos en **BBtools**, más concretamente, **BBduk** y **Dedupe**.

Bbduck se desarrolló para combinar las operaciones de recorte, filtrado y enmascaramiento más comunes relacionadas con la calidad de los datos. Es capaz de recortar y filtrar por calidad, recortar adaptadores, eliminar duplicados, filtrar contaminantes mediante coincidencia de kmer, enmascaramiento de secuencia, filtrado de GC, filtrado de longitud, filtrado de entropía, conversión de formato, generación de histogramas, submuestreo, recalibración de puntuación de calidad y varias otras operaciones en una sola pasada.

En el desarrollo de este trabajo, se realiza el trimado en dos pasadas, una primera para realizar un filtrado de calidad (BBduk), donde se elimina aquellas secuencias que no alcanzan un umbral de calidad mínimo. Y una segunda pasada, donde se elimina la alta presencia de duplicados (Dedupe) vistos previamente gracias al control de calidad. Tras su ejecución, obtenemos unos nuevos archivos fastq con mejores niveles de calidad.

Una vez realizado este paso, procedemos a ejecutar un nuevo control de calidad, con la finalidad de comprobar que se han solucionado los problemas de escasa calidad y exceso de duplicados.

3.4. Quasi-mapping con Salmon

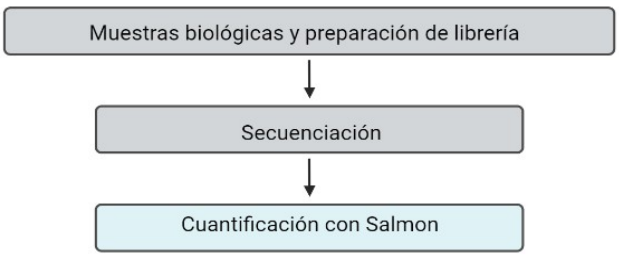


Figura 4. Flujo de trabajo simplificado 1. Creado con BioRender.com

Salmon es una herramienta cuya función es la cuantificación de la expresión de transcripciones mediante el uso de datos de RNA-seq. Es un cuasi-mapeador, ya que no produce las alineaciones de lectura (y no genera archivos BAM / SAM). Los “cuasi-mapas” de salmón se leen al transcriptoma en lugar de al genoma como lo hacen otros alineadores propiamente dicho, como por ejemplo STAR (Sato et al. 2018).

A diferencia de la pseudoalineación, el procedimiento de mapeo que utiliza Salmon rastrea, de forma predeterminada, la posición y orientación de todos los fragmentos mapeados. Esta información se utiliza junto con las abundancias de la inferencia en línea para calcular probabilidades condicionales por fragmento. Estas probabilidades se utilizan para estimar modelos auxiliares y términos de sesgo, y para actualizar las estimaciones de abundancia (Patro et al. 2017).

Salmon abarca tanto la "alineación" como la "cuantificación" en una sola herramienta. Tomando como entrada, en primer lugar un índice (para lo que necesitaremos la correspondiente secuencia en formato FASTA, obtenida de la base de datos de ensembl), y un conjunto de lecturas de secuenciación sin procesar (fastq). Una vez ejecutado este proceso, se lleva a cabo la cuantificación directamente sin generar ningún archivo de alineación intermedio. Esto ahorra mucho tiempo y espacio, ya que el cuasi-mapeo es más rápido que la alineación tradicional

El archivo resultante consiste en un archivo de cuantificación (*quant.sf*), es decir, un archivo de texto sin formato cuya información viene recogida en una serie de columnas, las cuales se interpretan tal y como indica la siguiente tabla.

Name	Length	EffectiveLength	TPM	NumReads
Nombre de la transcripción objetivo, proporcionada en la base de datos de transcritos de entrada (archivo FASTA)	Longitud de la transcripción diana en nucleótidos	Longitud efectiva calculada de la transcripción diana. Tiene en cuenta todos los factores que se modelan y que afectarán a la probabilidad de muestrear fragmentos de esta transcripción, incluida la distribución de la longitud del fragmento y el sesgo de fragmento GC	Estimación de la abundancia relativa	Estimación del número de lecturas asignadas a cada transcripción que se cuantificó

Tabla 8: Descripción del archivo de cuantificación generado por *Salmon*

Para la elaboración de este trabajo se utiliza Salmon v.1.4.0.

3.5. Análisis de la expresión diferencial

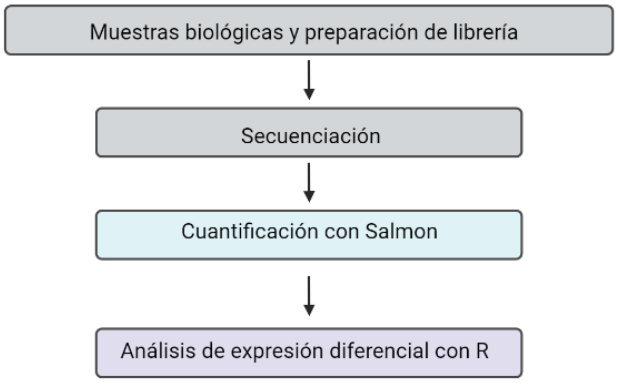


Figura 5. Flujo de trabajo simplificado 2. Creado con BioRender.com

El objetivo del análisis de expresión diferencial es determinar qué genes se expresan en diferentes niveles entre las diferentes condiciones. Estos genes pueden ofrecer información biológica sobre los procesos afectados por las condiciones de interés.

Para ello, es necesario seguir una serie de pasos que se ejecutarán mediante el software libre R/Bioconductor (v.4.1.2) y para lo que se deberá utilizar una serie de paquetes, todos imprescindibles para la realización del análisis de expresión diferencial.

R es un ambiente de programación (Open Source) cuyo origen se remonta al lenguaje S, formado por un conjunto de herramientas que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo funciones propias, cuya finalidad principal es el análisis estadístico y la representación gráfica.

Bioconductor es un proyecto que utiliza el lenguaje de programación R, cuya misión es la de desarrollar y difundir el código necesario para generar el análisis de determinados datos biológicos, siendo su objetivo inicial el análisis de Microarrays. Actualmente, es el análisis general de datos ómicos su papel principal.

Se mencionan a continuación aquellos paquetes de especial interés para el desarrollo del flujo de trabajo o workflow.

DESeq2

DESeq2 es un paquete (Delhomme et al. 2012) creado para realizar la normalización, visualización y el análisis diferencial de los datos de recuentos. Utiliza métodos Bayesianos para realizar las estimaciones pertinentes, mediante la estimación de la media de la varianza en función de los datos de recuento de alto rendimiento y determina la expresión diferencial basada en una distribución binomial negativa, partiendo de los datos de recuento sin procesar (sin normalizar).

Además de los datos de recuento sin procesar, DESeq2 necesita un archivo con la información de referencia, encargado de clasificar las muestras. En el presente análisis, las muestras se clasifican según las condiciones de severidad de la infección, aunque podría realizarse según otros criterios.

Tximport

El paquete de R/Bioconductor “tximport” está diseñado para realizar la importación de los datos de recuento sin procesar a nivel de TPM y resume (de forma predeterminada) abundancias, recuentos y longitudes de transcripción al nivel de gen.

La realización del pipeline mediante “cuasi-mapeo” y su posterior importación con tximport muestra una serie de beneficios o ventajas, tales como (Love et al. 2021):

- Corrige los cambios potenciales en la longitud de los genes en las muestras.
- Mayor sensibilidad
- Para su uso partimos de métodos de cuantificación ascendentes, como Salmon, lo que supone un considerable aumento en la velocidad de ejecución y un menor gasto de memoria, frente a otros métodos basados en la alineación (creación de archivos BAM).

3.6. Enriquecimiento

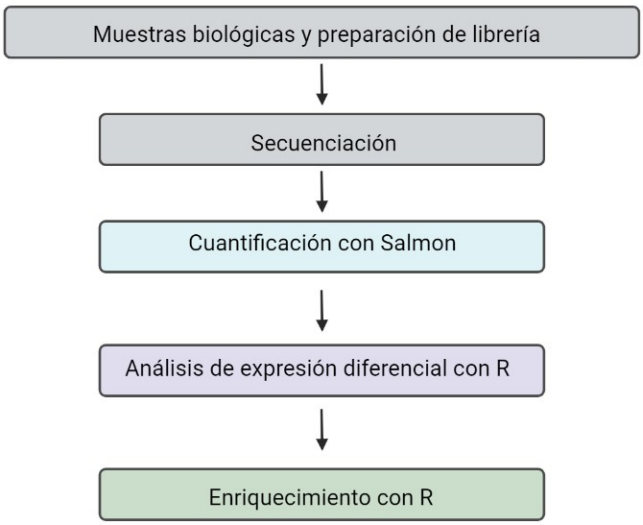


Figura 6. Flujo de trabajo simplificado 3. Creado con BioRender.com

El **análisis de enriquecimiento funcional** es un método cuya función se basa en identificar clases de genes que están sobre-representados en un gran conjunto de genes. El método utiliza enfoques estadísticos para identificar grupos de genes significativamente enriquecidos o empobrecidos.

Tras la realización del análisis de expresión diferencial, comparamos el conjunto de genes de entrada con cada uno de los términos (*bins*) en la ontología genética, consiguiendo así, recuperar un perfil funcional al conjunto de genes obtenidos en el análisis de expresión diferencial, y así comprender mejor los procesos biológicos subyacentes.

Dentro de los sistemas de anotación, el trabajo se centra en la ontología genética (GO) y la anotación de vías (KEGG). El **GO** define conceptos/clases que se utilizan para describir la función de los genes y las relaciones entre estos conceptos, clasificando las funciones en tres grupos:

- MF, función molecular: Funciones moleculares de los productos génicos.
- CC, Componente celular: Indica donde se encuentran activos los productos genéticos.
- BP, Proceso biológico: Vías y procesos más amplios formados por las actividades de múltiples productos génicos.

Por otro lado, **KEGG** es una colección de mapas de rutas dibujados “manualmente” que representan redes de interacción y reacción molecular. Estas vías cubren una amplia gama de procesos bioquímicos que se pueden dividir en 7 amplias categorías:

- Metabolismo
- Procesamiento de información genética
- Procesamiento de información ambiental
- Procesos celulares
- Sistemas
- Enfermedades humanas
- Desarrollo de fármacos

ClusterProfiler

Usaremos el paquete “ClusterProfiler” (v. 4.2.0) para realizar análisis de sobre-representación en términos de GO asociados con nuestra lista de genes significativos. Éste ofrece un método de clasificación de genes, llamado groupGO, para registrar genes en función de su proyección en un nivel específico del cuerpo GO, y proporciona funciones, enrichGO y enrichKEGG, para calcular la prueba de enriquecimiento para términos GO y rutas KEGG basadas en la distribución hipergeométrica (Guangchuang et al. 2012).

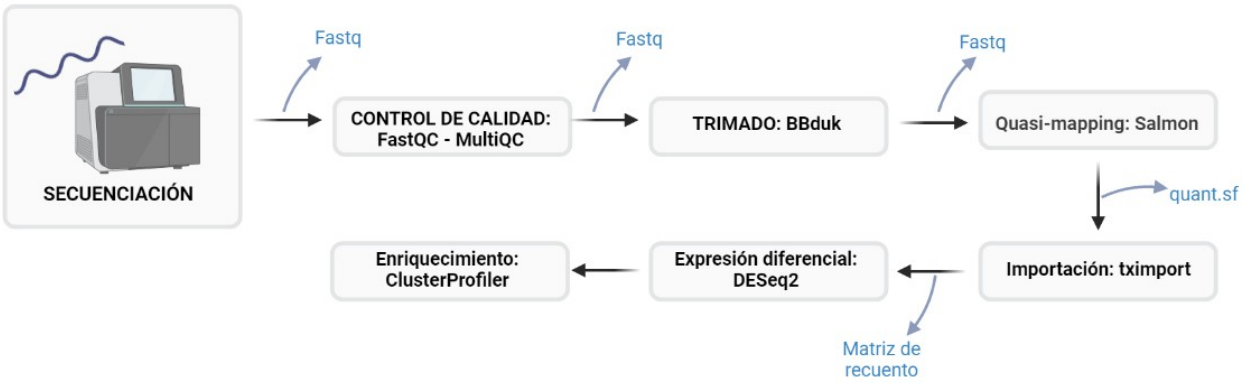


Figura 7. Flujo de trabajo. Creado con BioRender.com

4. Resultados

4.1. Control de calidad

La realización del análisis de calidad con FastQC y MultiQC muestra los diferentes parámetros de calidad de nuestros archivos fastq. Comenzamos analizando cada uno de los archivos mediante **FastQC**

- Basic Statistics: Genera un resumen general del análisis de calidad. Tales como "Nombre del archivo", "Tipo de archivo", "Codificación", "Número total de secuencias", "Secuencias marcadas como de mala calidad", "Rango de longitud de secuencias" y "%GC".

Measure	Value
Filename	RNA-seq_CoV26_UCI-14_540 06-10-20.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8156823
Sequences flagged as poor quality	0
Sequence length	25-397
%GC	51

Figura 8. FastQC: Basic Statistics

- Per Base Sequence Quality: muestra una descripción general del rango de valores de calidad en todas las bases en cada posición en el archivo FastQ. Para cada posición se dibuja un gráfico de tipo BoxWhisker.

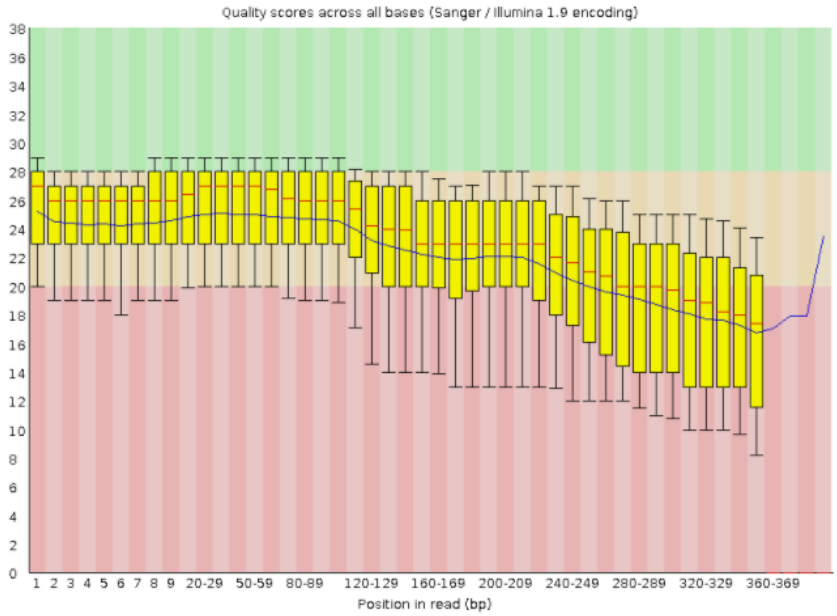


Figura 9. FastQC: Per Base Sequence Quality

El fondo del gráfico divide el eje y en muy buena calidad (verde), calidad razonable (naranja) y mala calidad (rojo). La calidad en la mayoría de las plataformas se degradará a medida que avanza la ejecución, por lo que es común ver que ésta disminuye hacia el final de una lectura. En el ejemplo aportado observamos como gran parte de las secuencias se encuentran en el rango de mala calidad.

- Per sequence quality scores: El informe de puntuación de calidad por secuencia permite ver si un subconjunto de sus secuencias tiene valores de calidad medios bajos. A menudo ocurre que un subconjunto de secuencias tendrá una calidad baja, porque tienen una imagen deficiente (en el borde del campo de visión, etc.), sin embargo, estas deben representar solo un pequeño porcentaje del total de secuencias.

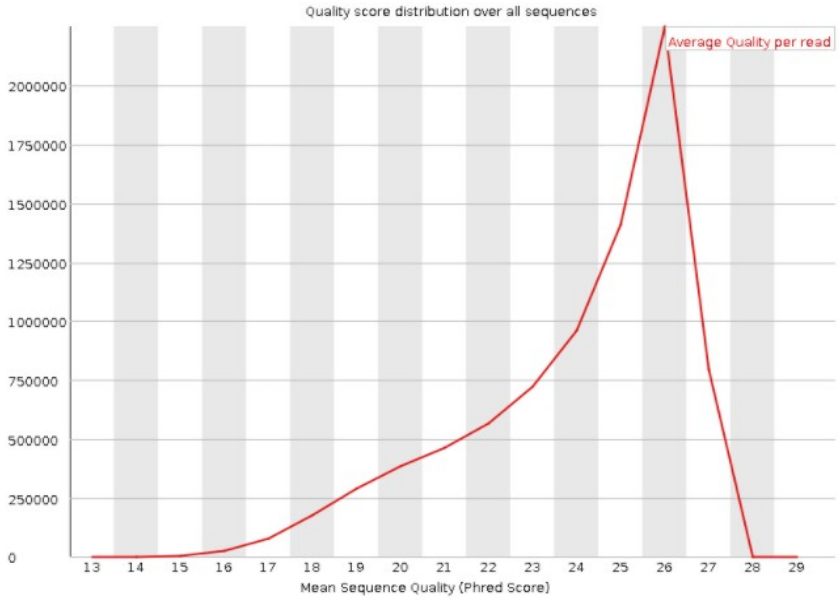


Figura 10. FastQC: Per sequence quality scores

- Per base sequence content: Indica el contenido de bases para cada posición que se observa en el archivo. En una biblioteca aleatoria, se esperaría que hubiera poca o ninguna diferencia entre las distintas bases de una secuencia, por lo que las líneas en esta gráfica deben correr paralelas entre sí. La cantidad relativa de cada base debe reflejar la cantidad total de estas bases en su genoma, pero en cualquier caso, no deben estar muy desequilibradas entre sí. La presencia de sesgos fuertes que cambian en diferentes bases, indica una secuencia sobre-representada que está contaminando la biblioteca.

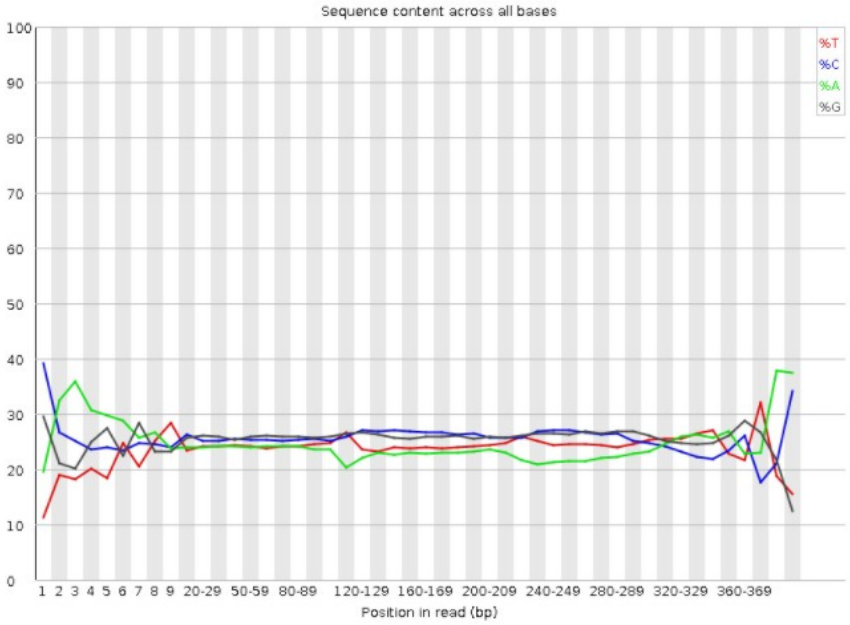


Figura 11. FastQC: Per base sequence content

- Per sequence GC content: Mide el contenido de GC en toda la longitud de cada secuencia en un archivo y lo compara con una distribución normal modelada. En una biblioteca aleatoria normal, se esperaría ver una distribución aproximadamente normal del contenido de GC, donde el pico central corresponde al contenido general de GC del genoma subyacente. Dado que no conocemos el contenido de GC del genoma, el contenido modal se calcula a partir de los datos observados y se utiliza para construir una distribución de referencia. Una distribución de forma inusual, como la encontrada, podría indicar una biblioteca contaminada o algún otro tipo de subconjunto sesgado.

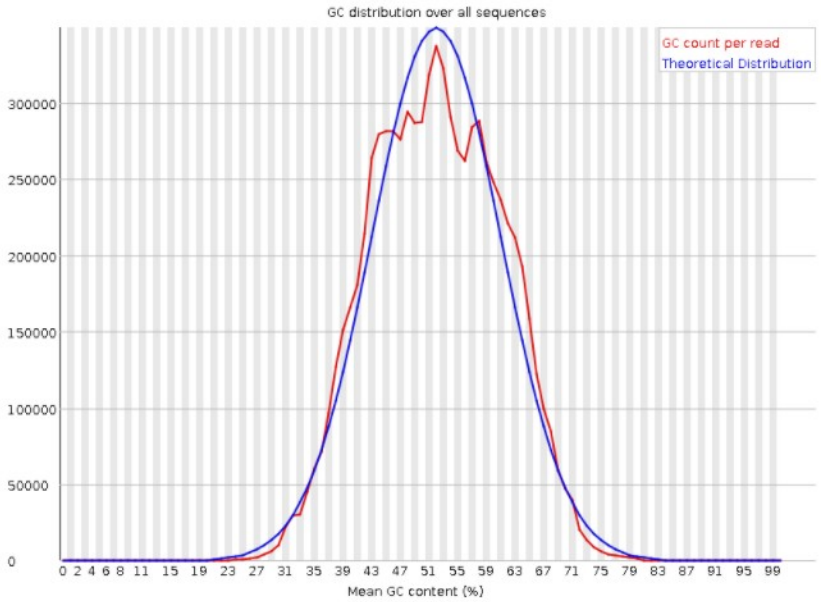


Figura 12. FastQC: Per sequence GC content

- Per Base N Content: Indica aquellas bases que no han podido ser nominadas, introduciendo una “N” en dichas posiciones. En el presente gráfico se observa que no ha habido bases sin nominar durante la secuenciación.

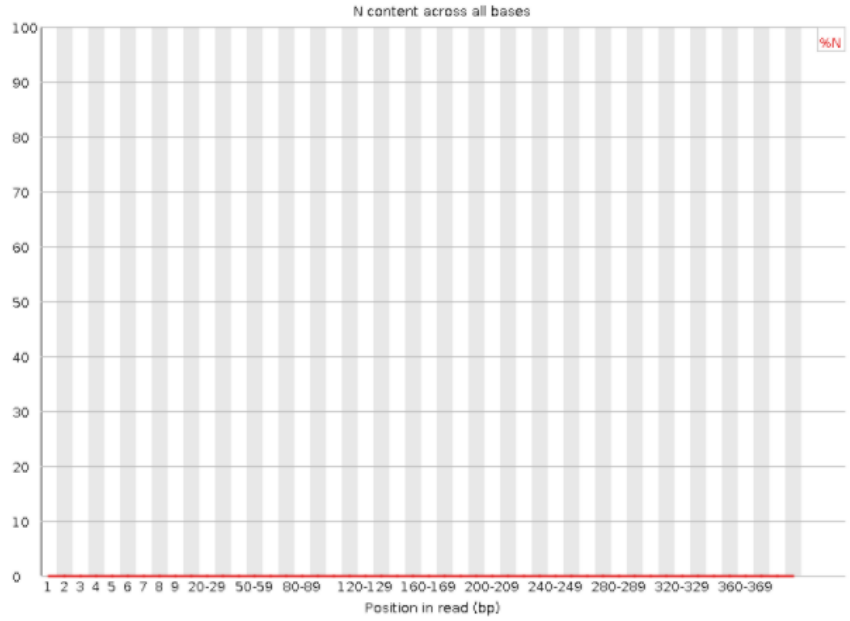


Figura 13. FastQC: Per Base N Content

- Sequence Length Distribution: Indica la distribución de longitudes de secuencias, generando un gráfico que muestra la distribución de tamaños de fragmentos en el archivo analizado.

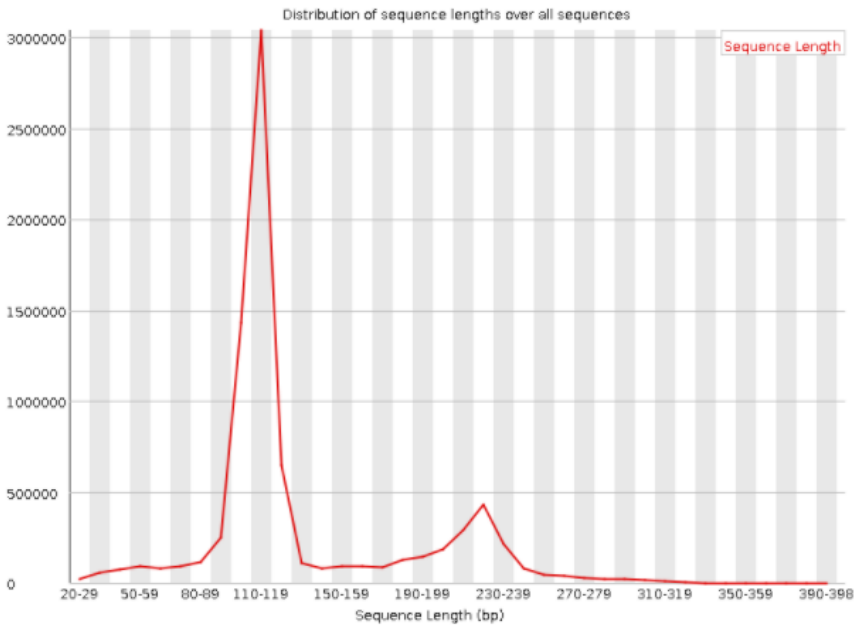


Figura 14. FastQC: Sequence Length Distribution

- Sequence Duplication Levels: Indica el grado de duplicación de cada secuencia en el conjunto y crea un gráfico que muestra el número relativo de secuencias con diferentes grados de duplicación. La mayoría de las secuencias se producirán solo una vez en el conjunto final. Un nivel bajo de duplicación puede indicar un nivel muy alto de cobertura de la secuencia diana, pero es más probable que un nivel alto de duplicación indique algún tipo de sesgo de enriquecimiento.

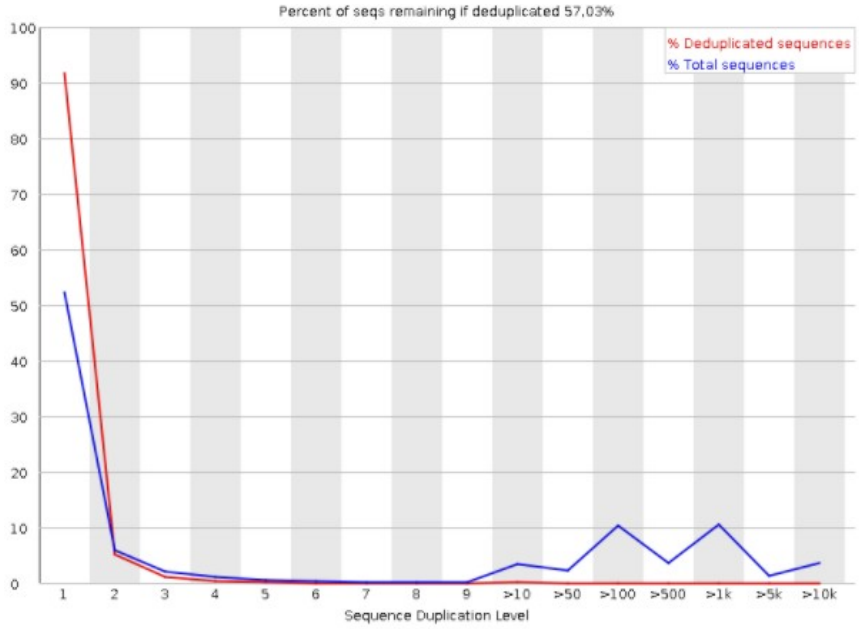


Figura 15. FastQC: Sequence Duplication Levels

- Overrepresented sequences: Muestra aquellas secuencias sobre-representadas. Descubrir que una sola secuencia está muy sobre-representada en el conjunto significa que es de gran importancia desde el punto de vista biológico o indica que la biblioteca está contaminada o no es tan diversa como se esperaba.

Sequence	Count	Percentage	Possible Source
GGCAATGAGCGGTTCCGCTGCCCTGAGGCACTCTTCCAGCCTTCCTTCCT	23876	0.29271200319045787	No Hit
GTCCACGTCACACTTCATGATGGAGTTGAAGGTAGTTTCGTGGATGCCAC	21314	0.261302715530299	No Hit
CTCTCTTTCTGGCCTGGAGGCTATCCAGCGTACTCCAAAGATTCAGGTTT	18893	0.23162204206220974	No Hit
GAAACCCAGACACATAGCAATTCAGGAAATTTGACTTTCCATTCTCTGCT	9493	0.11638109592423422	No Hit
CAATGGGGTACTTCAGGGTCAGGATGCCACGCTTGCTCTGGGCCTCGTCG	8405	0.10304256939251961	No Hit

Figura 16. FastQC: Overrepresented sequences

- Adapter Content: Gráfico acumulativo de la fracción de lecturas en las que se identifica la secuencia del adaptador de la biblioteca de secuencias en la posición base indicada. Idealmente no debería representar adaptadores.

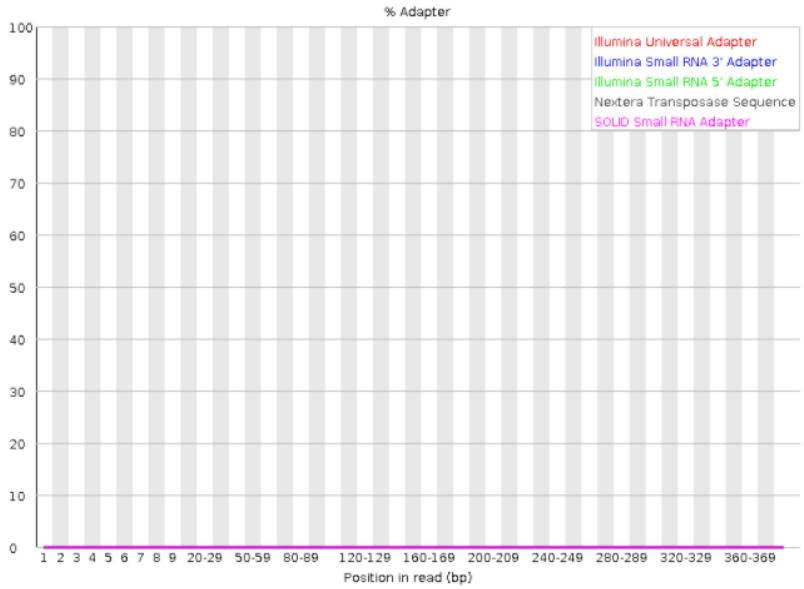


Figura 17. FastQC: Adapter Content

Este procedimiento se ha repetido para cada uno de los archivos fastq de partida, siendo el informe encargado de conglomerar esta información, aquel obtenido tras la ejecución de **MultiQC**.

A continuación, se genera una tabla resumen de parámetros indicadores de calidad de secuencia:

Sample Name	% Duplicate Reads	% GC	Average Length	Total Sequences (millions)
UCI-1	51.9%	50%	118 bp	9.3
UCI-2	54.1%	51%	116 bp	16.6
UCI-3	56.0%	50%	118 bp	11.0
UCI-4	53.6%	50%	119 bp	10.1
UCI-5	54.8%	51%	121 bp	10.7
UCI-6	60.3%	50%	119 bp	10.7
UCI-7	48.5%	51%	122 bp	12.8
UCI-8	48.4%	51%	124 bp	9.9
UCI-9	40.9%	51%	115 bp	12.1
UCI-10	43.0%	51%	134 bp	8.2
CV-10	64.3%	51%	115 bp	4.9
CV-9	63.6%	51%	116 bp	4.3
CV-8	52.1%	51%	120 bp	11.6
CV-7	49.2%	51%	117 bp	11.0
CV-6	42.3%	50%	118 bp	31.8
CV-5	56.1%	53%	117 bp	8.4
CV-4	48.3%	51%	120 bp	10.3
CV-3	49.9%	50%	120 bp	11.4
CV-2	52.3%	50%	120 bp	9.1
CV-1	49.4%	51%	114 bp	7.2

Tabla 9: Resumen resultados MultiQC

La tabla anterior muestra una descripción general de los valores clave, extraídos de todos los módulos. Su objetivo es reunir las estadísticas de cada muestra de todo el análisis. Además, nos aporta información de gran valor, especialmente en la localización de posibles archivos inválidos.

En concreto observamos tres archivos cuyo tamaño de secuencia se desvía considerablemente de los valores esperados, ya que se ha utilizado un panel de transcriptómica en el que se introducen 8 muestras por chip, el cual es de 60 millones. Con ello sabemos que cada muestra tendrá un mínimo de 7 millones de secuencias (60/8). Además, se observa una muestra con un exceso tamaño de secuencia (CV-6), muy por encima de la media, posiblemente por un error de lecturas.

Esto nos hará sustituir tres muestras para el futuro desarrollo del análisis, acción que se llevará a cabo junto al trimado y cuyos archivos serán nominados de igual modo, las cuales son:

- CV-10
- CV-9
- CV-6

El informe generado aporta, además, la siguiente información:

- Recuento de secuencias

La información sobre el tamaño de secuencia se encuentra representada en el siguiente gráfico

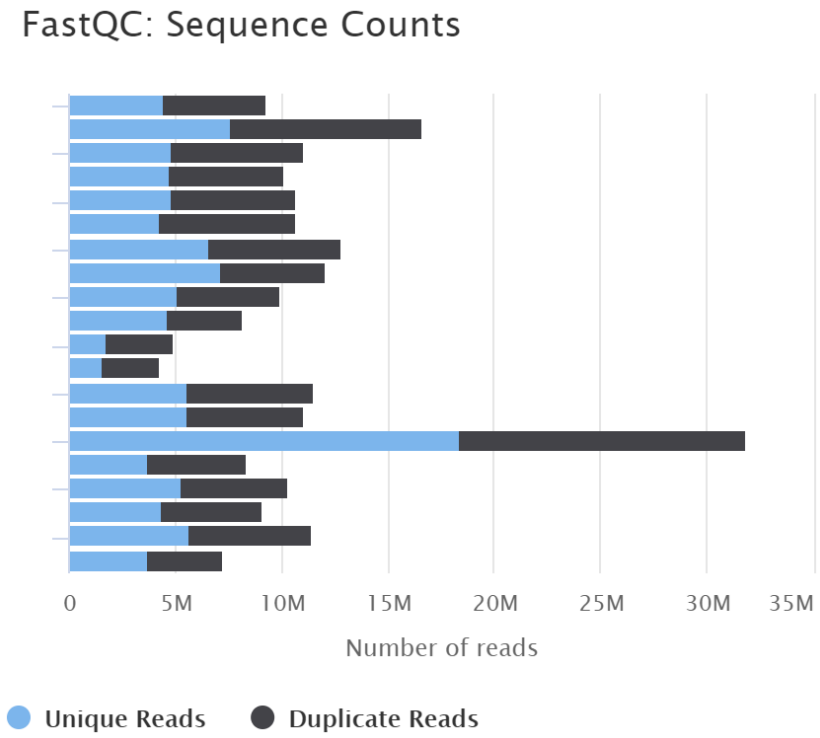


Figura 18. Recuento de secuencias

Donde vemos gráficamente la información presente en la tabla anterior.

- Calidad media de las secuencias

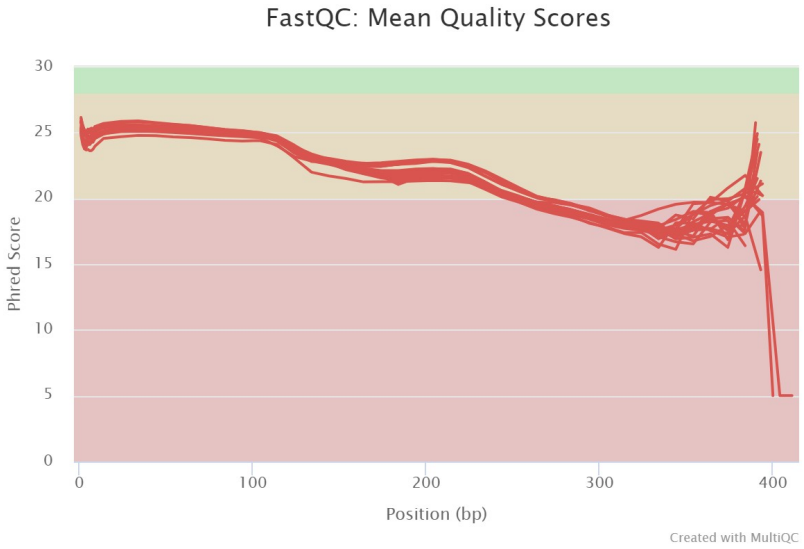


Figura 19. Calidad media de las secuencias

El programa devuelve como fallidas todas las secuencias. Esta pérdida de calidad observada suele corresponder a una degradación en la química de la secuenciación. Vemos que la pérdida de calidad de las secuencias ocurre a medida que avanza la ejecución, donde la mayor parte de la secuencia se encuentra en un rango de calidad aceptable (>20).

- Puntuaciones de calidad por secuencia

Indica el número de lecturas con puntajes de calidad promedio, mostrando si un subconjunto de lecturas tiene mala calidad.

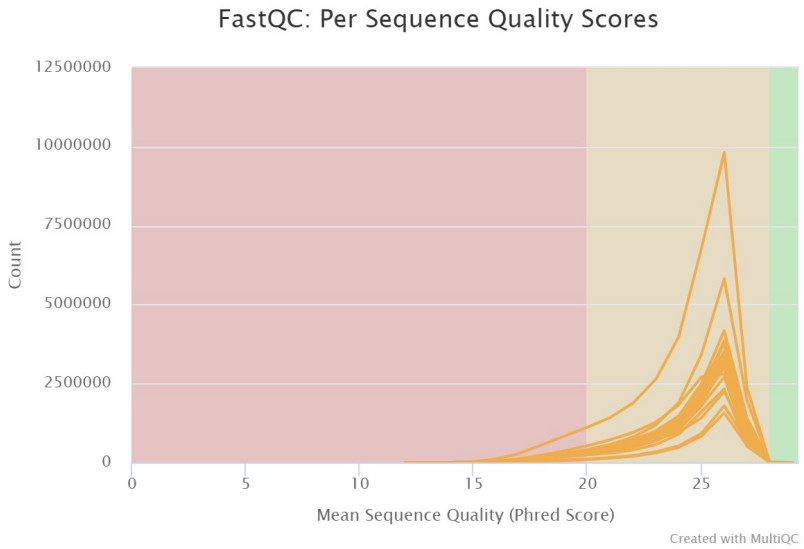


Figura 20. Puntuaciones de calidad por secuencia

- Contenido en bases

Apreciamos un primer gráfico donde cada fila corresponde a uno de los archivos fastq

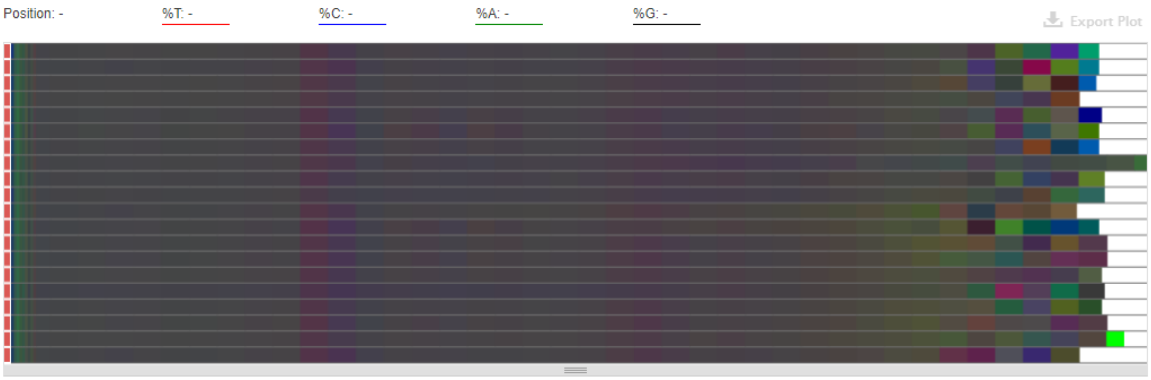


Figura 21. Contenido en bases (global)

Tal y como observamos en el ejemplo aportado por FastQC, se aprecia un sesgo generalizado en cuanto al contenido en bases (%) presentes. Existen diversos factores que pueden generar esta desviación, tales como:

- Secuencias sobre-representadas: si hay secuencias sobre-representadas pueden sesgar la composición general.

- Fragmentación sesgada: Se debe a una selección sesgada de cebadores aleatorios, pero no representa ninguna secuencia sesgada individualmente. Casi todas las bibliotecas de RNA-Seq fallarán en este módulo debido a este sesgo, pero este no parece afectar negativamente la capacidad de medir la expresión.

- Contenido GC por secuencias

Donde comprobamos la presencia de sesgo con respecto a una distribución normal en varias secuencias.

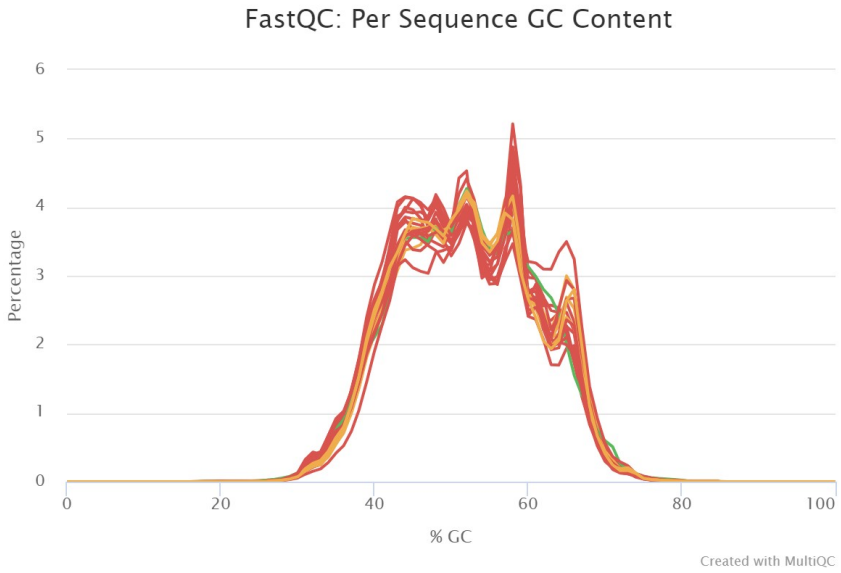


Figura 22. Contenido GC por secuencias

- Contenido en “N”

Comprobamos a continuación la ausencia de bases que no hayan sido identificadas

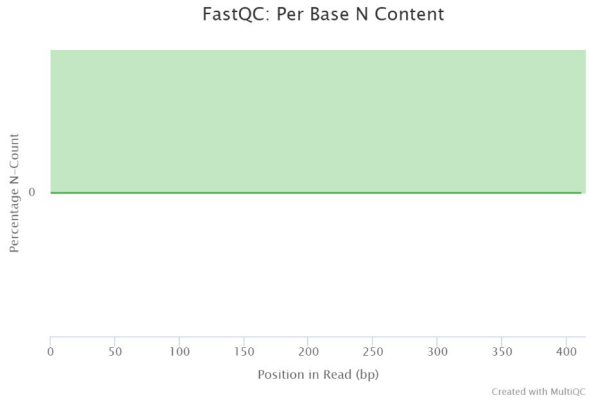


Figura 23. Contenido en “N”

- Distribución de la longitud de secuencias

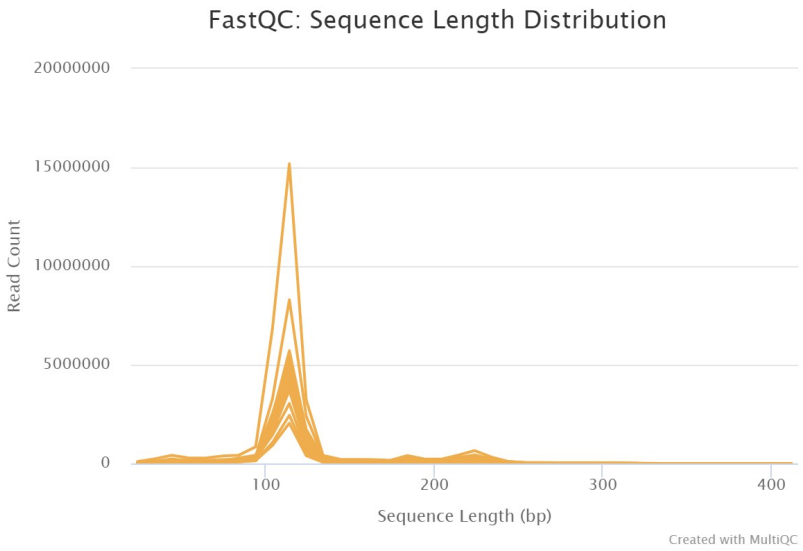


Figura 24. Distribución de la longitud de secuencias

Se aprecia que la distribución en la longitud de las secuencias se ajusta a una distribución válida, con una pérdida en los índices de calidad causada por los “picos” observados entorno a las 200-250 bp.

- Nivel de duplicados

Se evidencia en el siguiente gráfico la presencia de duplicados en todas las secuencias en estudio, siendo importante en varias de ellas.

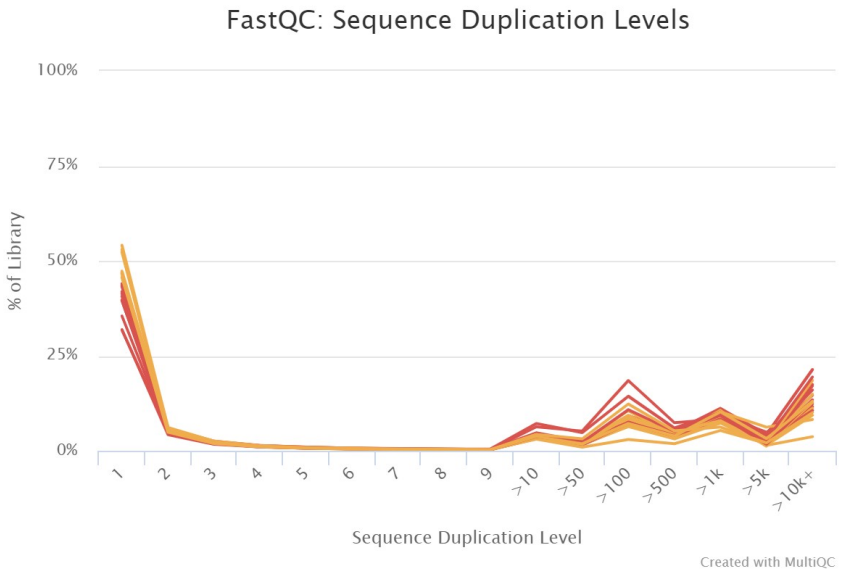


Figura 25. Nivel de duplicados

Sin embargo, cabe destacar que en las bibliotecas de RNA-Seq, las secuencias de diferentes transcripciones estarán presentes a niveles tremendamente diferentes en la población inicial. Por lo tanto, para poder observar transcripciones de baja expresión, es común sobre-secuenciar en gran medida las transcripciones de alta expresión, y esto potencialmente creará un gran conjunto de duplicados.

- Secuencias sobre-representadas

Los siguientes gráficos muestran aquellas secuencias que se encuentras sobre-representadas

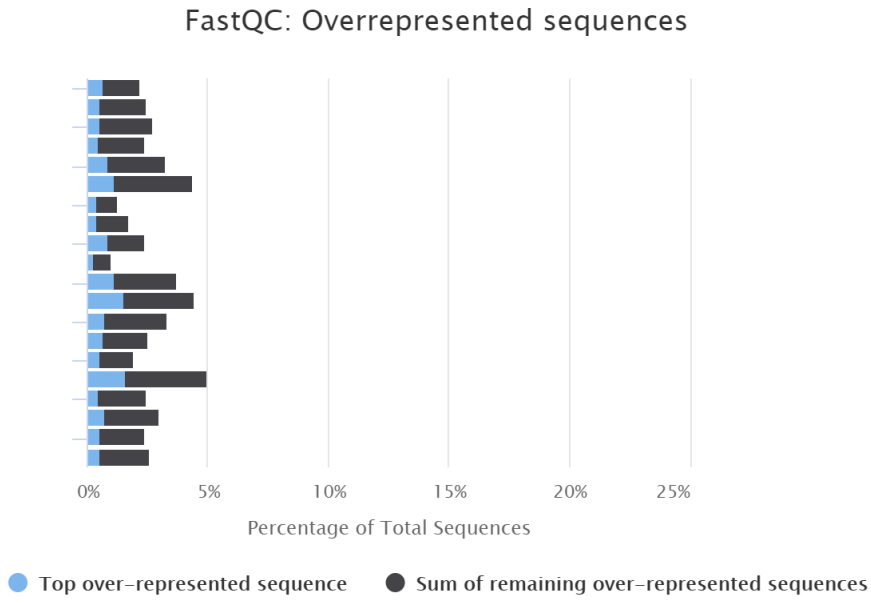


Figura 26. Secuencias sobre-representadas (global)

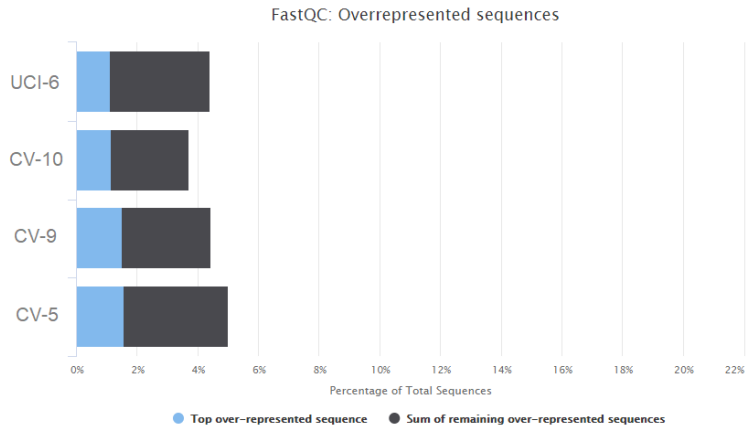


Figura 27. Secuencias sobre-representadas (específico)

Comprobamos que existen 4 secuencias sobre-expresadas, de las cuales, dos corresponden a aquellas secuencias que anteriormente definimos como inválidas.
Las otras dos secuencias sobre-expresadas pueden indicar una contaminación en la biblioteca, una baja diversidad o un alto valor biológico.

Finalmente, se genera un cuadro resumen con el estado de cada sección de FastQC que muestra si los resultados parecen completamente normales (verde), ligeramente anormales (naranja) o muy inusuales (rojo). Comprobamos los resultados que han sido comentados en cada sección:

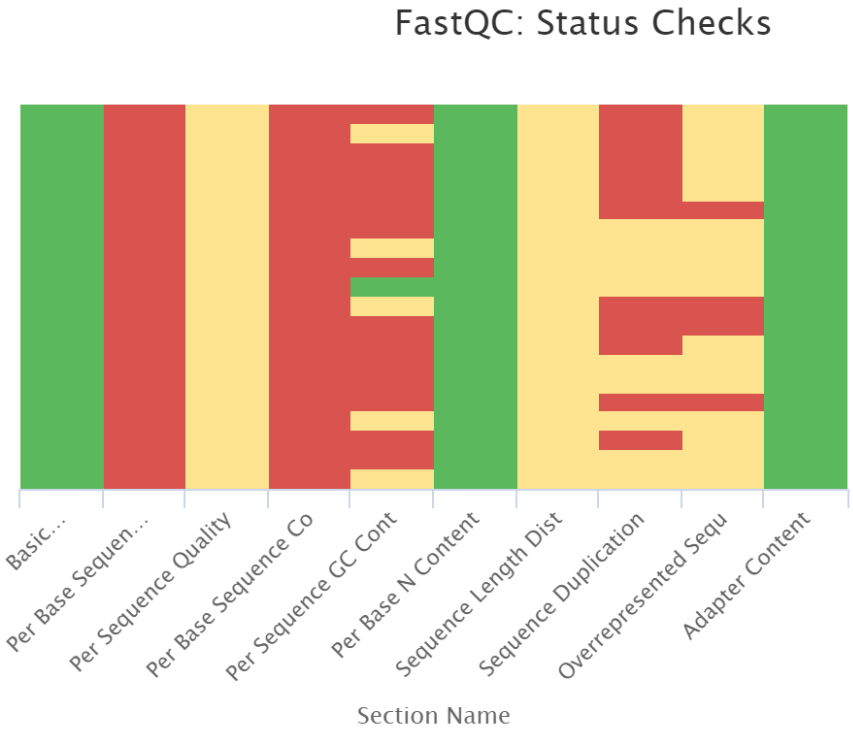


Figura 28. Estado general de la calidad de los datos

De esta forma se aporta una importante visión general sobre el estado de las secuencias problemas.

En éste, se aprecia una escasa calidad general, siendo especialmente llamativo la calidad de las bases por secuencia, el contenido en bases, la concentración de GC y la cantidad de secuencias duplicadas. Todo ello, hace catalogar el presente conjunto de archivos como de baja calidad aparente. Por ello se decide la realización del siguiente apartado, el trimado.

4.2. Trimado

Aunque, habitualmente, la función principal del trimado es la eliminación o remoción de adaptadores. En nuestro caso (datos secuenciados con la técnica *Ion torrent*) no hay adaptadores presentes, tal y como hemos visto en el apartado anterior.
Sin embargo, a pesar de dicha ausencia, cabe destacar la escasa calidad general de las secuencias, por lo que se realiza seguidamente un trimado, cuyo objetivo principal será filtrar secuencias según su calidad (eliminando bases con calidad menor a un "valor Q" umbral) y eliminar los duplicados. Pues es en estos dos ámbitos donde se presupone, principalmente, una importante influencia en cuanto a la valoración final aportada por el análisis de calidad.

Para realizar este procedimiento se ha optado por emplear las herramientas **BBtools**, tal y como mencionamos anteriormente. Con él, realizaremos una serie de modificación de los archivos fastq problemas.
Para tal fin, se ejecuta BBduk en dos ocasiones o vueltas.
En la primera vuelta se ejecuta con la intención de realizar los siguientes procesos:

- Eliminar las bases de baja calidad:

A cada base de la lectura se le asigna un valor Q, que se define como el logaritmo negativo de la probabilidad de que la base se haya llamado incorrectamente. El valor Q tiende a disminuir (la calidad empeora) hacia el final 3' de la lectura. Estas regiones de menor calidad pueden afectar negativamente los análisis posteriores, como el mapeo.

El recorte conduce a una disminución en el número de lecturas, pero aumenta la proporción de lecturas mapeables, y aunque no siempre es recomendado en los flujos de trabajo de RNA-seq por la posible desviación de procesos posteriores, es necesario cuando la calidad de los datos es muy baja (Williams et al. 2016).

Se realiza un recorte de calidad en *Q=10*, mediante el uso del algoritmo Phred. Indicando que los cortes se realizarán a ambos lados, tanto a la izquierda como a la derecha.

- Recorte de la secuencia para mejorar el contenido en bases:

Tal y como se visualiza en las figuras 11 y 20, el contenido en GC sufre una mayor desviación en ambos extremos. Por ello, se opta por realizar recortes a las secuencias.
Se realiza un corte en la posición 20 por la izquierda, eliminando las secuencias desde la posición 0 a 20. También se lleva a cabo un corte por la derecha, en concreto en la posición 300, pues es donde se estima que se inicia un importante sesgo en cuanto a contenido GC.

En una segunda vuelta, utilizamos **Dedupe** para:

- Eliminación de duplicados:

El análisis de calidad muestra una elevada presencia de duplicados, que junto con el resto de resultados, hace que concluyamos que partimos de unos archivos con baja calidad.
Por ello, utilizamos Dedupe para eliminar contigs duplicados. Éste, además puede encontrar todas las secuencias contenidas y superpuestas en un conjunto de datos, lo que permite un número específico de sustituciones.

Una vez realizado el pertinente trimado, se realiza un nuevo análisis de calidad con la intención de comprobar cambios en la calidad de las secuencias.
Para ello se utiliza nuevamente los programas FastQC y MultiQC, tras lo que podemos observar los siguientes cambios.

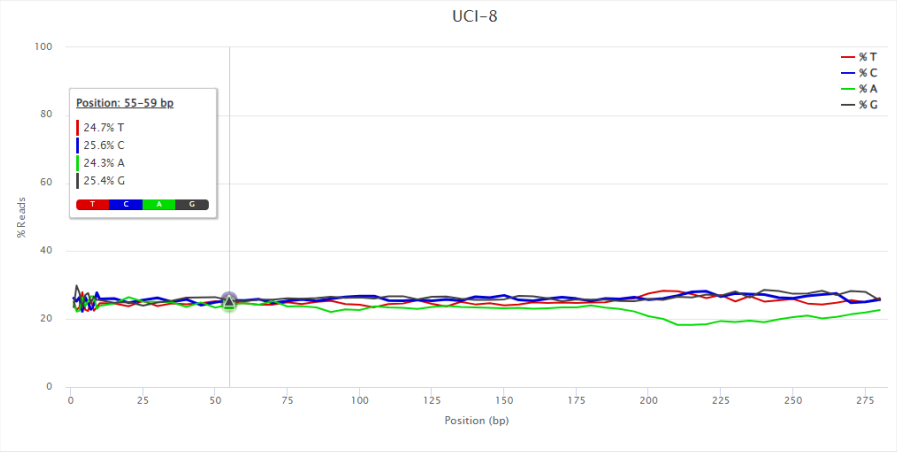


Figura 29. Contenido en bases post-trimado

A contrario de lo ocurrido anteriormente, una vez realizados los cortes mencionados se puede atisbar cómo las secuencias mantienen, en gran parte de las posiciones, un correcto contenido en bases.

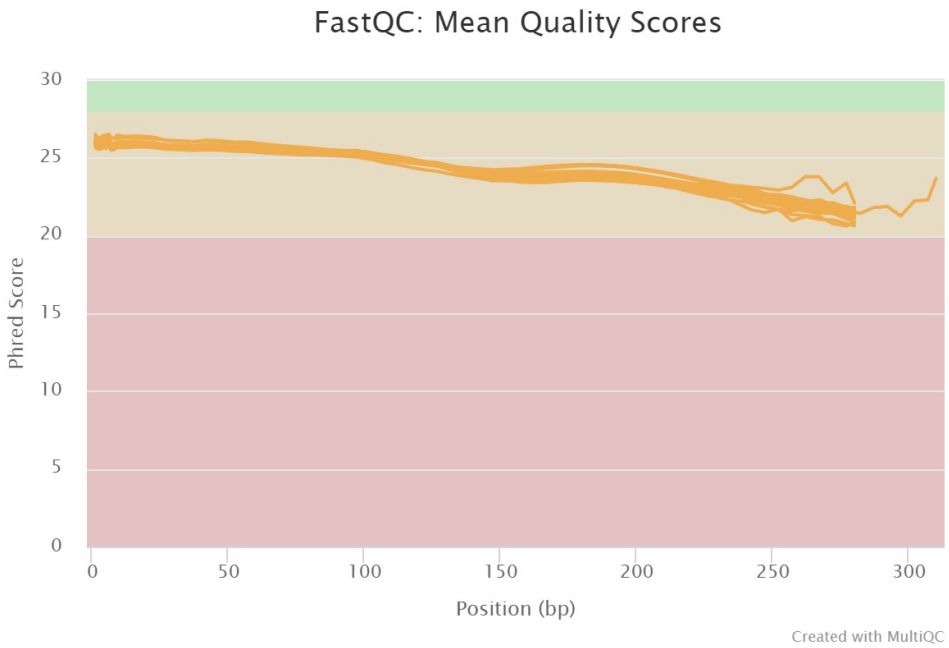


Figura 30. Valores medios de calidad post-trimado

Se aprecia cómo también se ha producido un importante aumento en la calidad media de las secuencias, pues aunque sigue siendo una calidad discreta, contrario a lo visto anteriormente, se observa que ninguna secuencia entra en el rango de “mala calidad”, es decir, por debajo del umbral (20)

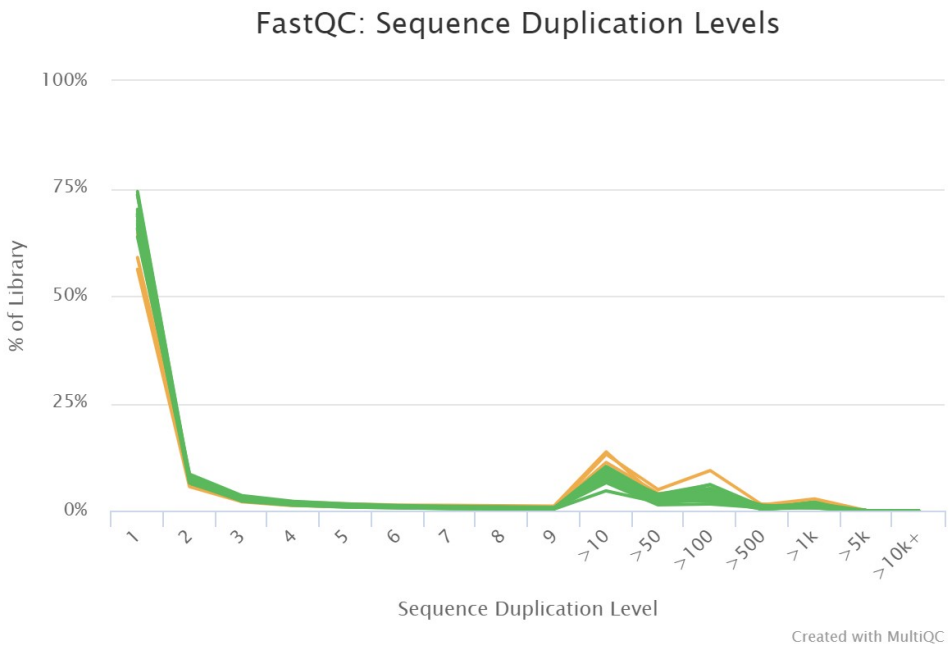


Figura 31. Nivel de secuencias duplicadas post-trimado

Una vez más se constata una mejora posterior al trimado, viendo como el exceso de secuencias duplicadas se ha reducido drásticamente hasta niveles óptimos para la mayoría de las secuencias, con niveles discretos para tres de ellas.

Finalmente, observamos estos y otros parámetros en el resumen final del control de calidad.

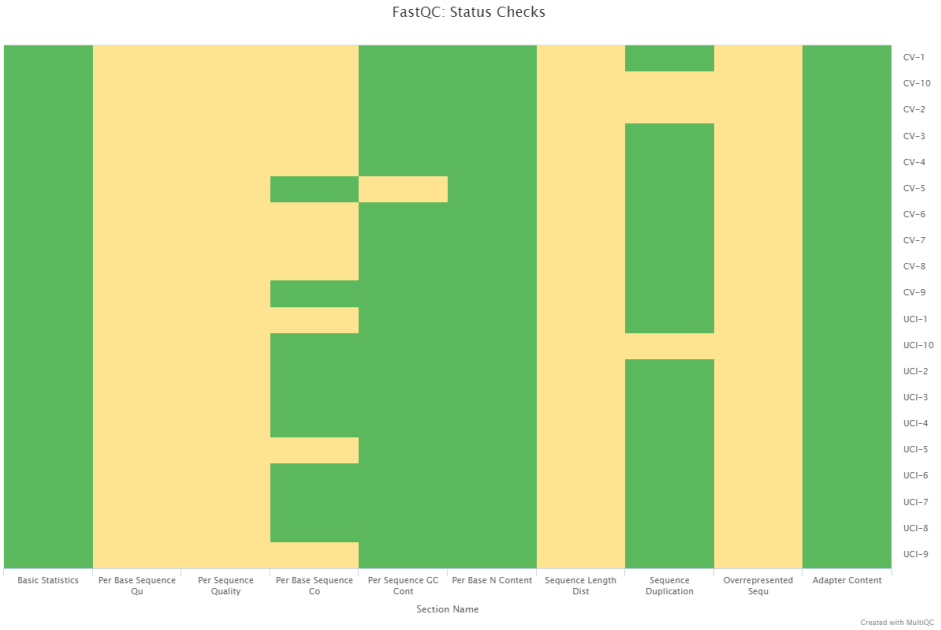


Figura 32. Estado general de la calidad de los datos post-trimado

De dicho análisis concluimos que la calidad de los datos ha mejorado sustancialmente tras la realización del trimado. Es por ello, que proseguimos con el “quasi-mapping”

4.3. Quasi-mapping

El objetivo fundamental de este paso es conocer la ubicación de las correspondientes lecturas con respecto a un transcriptoma de referencia. Para ello, en la elaboración del presente trabajo se ha utilizado *Salmon*, con la finalidad de realizar el mencionado "cuasi-mapeo".

Creamos el índice utilizando la requerida secuencia de referencia en formato fasta, obtenida en ensembl (https://www.ensembl.org/Homo_sapiens/Info/Index).

Como se pretende asignar las lecturas a los genes, además de las transcripciones, hay que proporcionar un archivo GTF correspondiente a la versión de la transcripción que se utilizó para construir el índice Salmon, el cual se descarga de la misma fuente que el archivo de referencia anteriormente indicado.

Con estos dos pasos listos, se procede a realizar el "mapeo", con el que se consigue pasar de un archivo fastq a un archivo quant.sf. Éste consiste en un archivo tabulado que contiene los recuentos de lectura de los genes.

Se toma la información sobre recuentos de lectura de genes de archivos *quant.sf* para el análisis de expresión diferencial (DE)

4.4. Procesado y visualización de los datos

Una tarea esencial en el análisis de los datos de recuento de RNA-seq es la detección de genes expresados diferencialmente. Los datos de recuento se presentan como una tabla que informa, para cada muestra, el número de fragmentos de secuencia que se han asignado a cada gen.

Una cuestión de análisis importante es la cuantificación y la inferencia estadística de cambios sistemáticos entre condiciones, en comparación con la variabilidad dentro de las condiciones. Tal y como se mencionó en apartados anteriores, el paquete DESeq2 proporciona métodos para probar la expresión diferencial mediante el uso de modelos lineales generalizados binomiales negativos.

Antes de comenzar con el análisis de expresión diferencial, es necesario realizar una serie de procesamientos a los archivos provenientes del "cuasi-mapeo". El primer paso es la instalación y carga de todas las librerías que se utilizarán posteriormente y la creación de los objetos iniciales, siendo uno de ellos el *coldata*, es decir, la información relativa al diseño experimental llevado a cabo.

	names	condition	files
CV-1	CV-1	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-2	CV-2	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-3	CV-3	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-4	CV-4	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-5	CV-5	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-6	CV-6	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-7	CV-7	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-8	CV-8	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-9	CV-9	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
CV-10	CV-10	CV	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-1	UCI-1	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-2	UCI-2	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-3	UCI-3	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-4	UCI-4	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-5	UCI-5	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-6	UCI-6	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-7	UCI-7	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-8	UCI-8	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-9	UCI-9	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...
UCI-10	UCI-10	UCI	C:/Users/israe/Desktop/Master_Bioinformatica_y_Bio...

Figura 33. *Coldata*. Diseño experimental

El otro paso clave para iniciar el análisis de expresión diferencia es la exportación de la matriz de conteo al entorno R/Bioconductor. Para ello se utilizan los paquetes *tximport* y *DESeq2*, a partir de los cuales se obtiene una primera matriz, con datos sin procesar.

$abundance

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
ENSG00000000003	0.076340	0.074121	0.141384	0.233424	0.085935	0.028438	0.161940	0.153767
ENSG00000000005	0.000000	0.000000	0.000000	0.029330	0.000000	0.000000	0.008976	0.000000
ENSG00000000419	69.234909	66.432783	74.504850	56.271188	65.287733	44.459620	63.880607	57.316438
ENSG00000000457	3.453722	4.384591	2.779065	4.246130	2.986123	2.518673	3.906874	3.314732
ENSG00000000460	0.424110	0.703089	0.709139	0.385324	0.650709	0.259363	0.336984	0.335311
ENSG00000000938	41.448890	54.701794	57.926808	96.797906	63.145872	156.689683	106.942190	78.913353
ENSG00000000971	5.264666	9.135209	0.987275	1.693703	3.354243	1.144406	0.816988	5.798026
ENSG00000001036	11.333654	10.553243	6.101159	4.407463	7.695826	3.075766	8.695824	8.253877
ENSG00000001084	30.124221	37.548077	82.560979	40.209977	28.951046	8.575852	36.382596	36.073959
ENSG00000001167	8.513343	12.213564	11.696491	17.379840	10.572539	15.475742	12.717589	14.332180
ENSG00000001460	2.515192	5.496754	3.408570	1.759330	3.583593	0.982591	2.782617	4.078203
ENSG00000001461	9.035420	24.001168	11.545228	6.834258	12.851302	4.009136	6.284540	12.648676
ENSG00000001497	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ENSG00000001561	2.410773	2.824959	2.273368	2.204266	3.188012	0.354913	1.826627	3.013436
ENSG00000001617	0.000000	0.000000	0.000000	0.000000	0.186356	1.299674	0.184235	0.000000
ENSG00000001626	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ENSG00000001629	31.112601	41.394509	44.711437	31.469844	28.738090	14.775036	45.206711	39.972066
ENSG00000001630	40.163403	42.944848	70.156330	21.875448	45.716066	13.069930	64.171688	33.780039
ENSG00000001631	0.000000	0.037232	0.000000	0.666722	0.381622	1.050227	0.382608	0.670012
ENSG00000002016	5.790913	4.480927	6.266863	4.107036	5.681416	3.497702	4.235488	5.704750

Figura 34. Matriz de cuentas sin procesar

A continuación, se realiza siguiente tratamiento a los datos de partida.

Normalización

Los recuentos sin procesar representan el número de lecturas que se alinean con cada gen y que debe ser proporcional a la expresión del ARN en la muestra. Sin embargo, hay factores, además de la expresión del ARN, que pueden influir en el número de lecturas.

Si bien la normalización es esencial para el análisis de expresión diferencial, también lo es para realizar una visualización de los datos y el análisis exploratorio de estos, acción previa al propio análisis de expresión diferencial.

Podemos ajustar los datos de recuento para eliminar la influencia de estos factores en los recuentos generales mediante el uso de métodos de normalización.

Desde una visión puramente teórica, la normalización que realiza DESeq2 se ejecuta en los siguientes pasos:

- Creación de una muestra de “pseudo-referencia”, que es igual a la media geométrica de todas las muestras.

- Cálculo de la relación de cada muestra frente a la pseudo-referencia. Para cada gen de una muestra, se calcula las proporciones (muestra/pseudo-referencia). Destacarán aquellos genes con expresión diferencial, que serán los que tengan proporciones dispares.

- Cálculo del factor de normalización para cada muestra. La mediana de todas las proporciones para una muestra dada se toma como *factor de normalización* para dicha muestra.

- Cálculo de los valores de recuento normalizado utilizando el *factor de normalización* previamente calculado. Para ello se divide cada valor de recuento sin procesar entre dicho factor, generando así recuentos normalizados.

Una vez realizados dichos pasos, obtenemos la mencionada matriz de recuentos normalizados.

	CV-1	CV-2	CV-3	CV-4	CV-5	CV-6	CV-7	CV-8
ENSG00000000003	2.566512e+00	2.148801e+00	4.810881e+00	8.812346e+00	2.704619e+00	1.180509e+00	5.421567e+00	5.046612e+00
ENSG00000000005	0.000000e+00	0.000000e+00	0.000000e+00	1.017234e+00	0.000000e+00	0.000000e+00	2.751609e-01	0.000000e+00
ENSG00000000419	8.101488e+02	6.707350e+02	8.831905e+02	7.381409e+02	7.158011e+02	6.693230e+02	7.402142e+02	6.547921e+02
ENSG00000000457	1.105202e+02	1.209900e+02	9.000915e+01	1.521668e+02	8.945538e+01	1.037315e+02	1.237571e+02	1.035498e+02
ENSG00000000460	3.860696e+01	5.519047e+01	6.533608e+01	3.928138e+01	5.545221e+01	3.038095e+01	3.036437e+01	2.979763e+01
ENSG00000000938	1.403857e+03	1.597635e+03	1.985743e+03	3.671540e+03	2.002163e+03	6.828979e+03	3.585300e+03	2.609197e+03
ENSG00000000971	7.594581e+01	1.136364e+02	1.441468e+01	2.736168e+01	4.529726e+01	2.124317e+01	1.166582e+01	8.165061e+01
ENSG00000001036	4.941214e+02	3.967488e+02	2.692215e+02	2.151913e+02	3.140968e+02	1.725530e+02	3.752674e+02	3.512919e+02
ENSG00000001084	2.143408e+02	2.303792e+02	5.945686e+02	3.203818e+02	1.928394e+02	7.851153e+01	2.562412e+02	2.505707e+02
ENSG00000001167	2.973984e+02	3.679145e+02	4.135496e+02	6.799182e+02	3.457501e+02	6.956577e+02	4.397540e+02	4.887616e+02
ENSG00000001460	3.311837e+01	6.241228e+01	4.542590e+01	2.588317e+01	4.417338e+01	1.656668e+01	3.614533e+01	5.242192e+01
ENSG00000001461	2.774634e+02	6.355586e+02	3.588339e+02	2.350287e+02	3.694442e+02	1.584214e+02	1.910281e+02	3.791829e+02
ENSG00000001497	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
ENSG00000001561	2.327617e+02	2.351977e+02	2.221563e+02	2.383370e+02	2.881506e+02	4.409430e+01	1.745707e+02	2.840295e+02
ENSG00000001617	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	8.755358e-01	8.393183e+00	9.152170e-01	0.000000e+00
ENSG00000001626	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
ENSG00000001629	5.186287e+02	5.950160e+02	7.543489e+02	5.874719e+02	4.484589e+02	3.169235e+02	7.459149e+02	6.504639e+02
ENSG00000001630	4.651732e+02	4.286123e+02	8.217570e+02	2.838329e+02	4.952418e+02	1.940513e+02	7.351728e+02	3.814554e+02
ENSG00000001631	0.000000e+00	4.451207e-01	0.000000e+00	1.033344e+01	4.786132e+00	1.861572e+01	5.212314e+00	9.039349e+00
ENSG00000002016	4.608086e+01	3.074730e+01	5.047282e+01	3.659943e+01	4.232276e+01	3.581474e+01	3.336132e+01	4.431562e+01

Figura 35. Matriz de cuentas normalizada

Análisis exploratorio de los datos normalizados

Se realiza una breve exploración de los datos, que se inicia realizando agrupaciones jerárquicas, tal y como se muestra en el siguiente dendrograma.

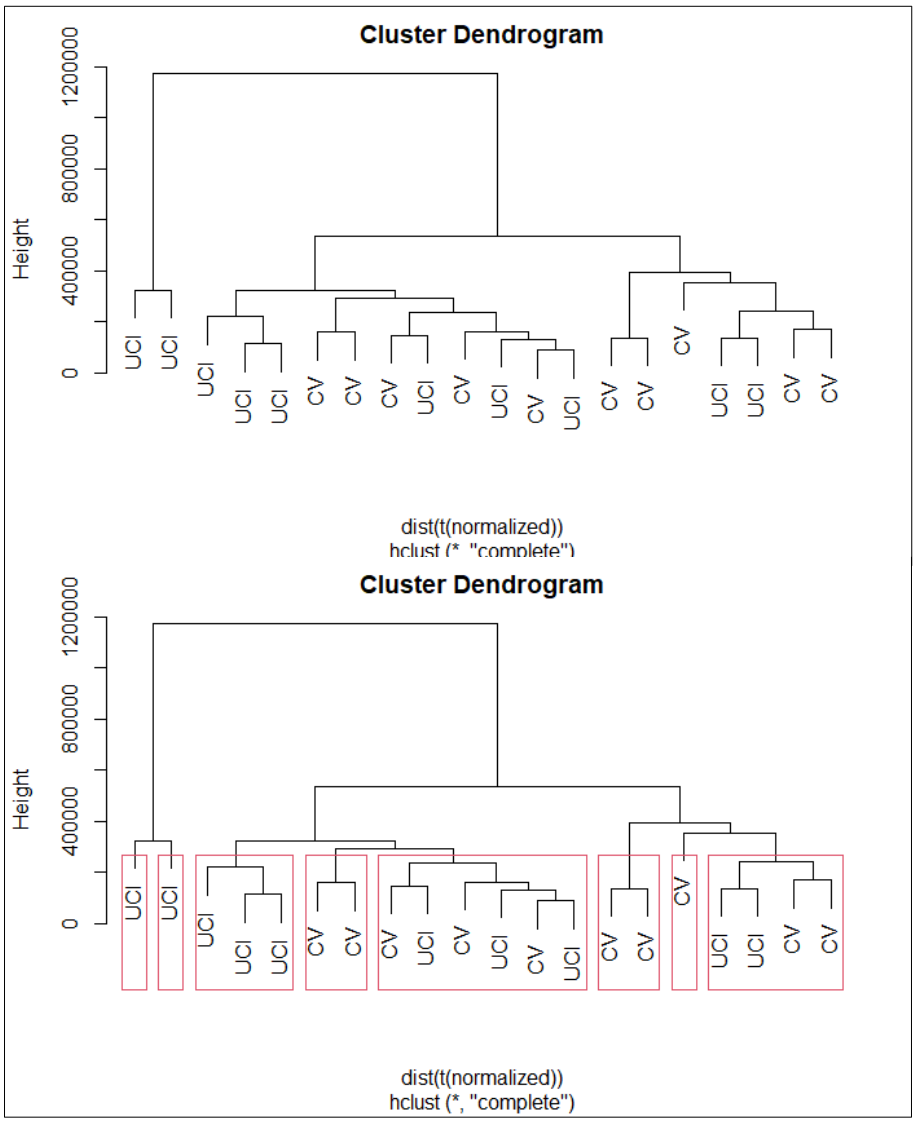


Figura 36. Dendrogramas (CV-UCI)

Un dendrograma es un diagrama formado mediante el uso de un algoritmo de clustering jerárquico, en el que se calculan las distancias para cada par de clases en un archivo de firma de entrada. Tras esto, se fusionan iterativamente aquellos pares de clases más cercanos, proceso que realiza sucesivamente (siempre con el par de clase más cercano) hasta que todos los pares se encuentran fusionados.

El gráfico se genera calculando las distancias existente entre todos estos pares de clases.

Salvo algunas excepciones, se aprecia que en los niveles bajos del dendrograma dado, cada una de las condiciones tiende a agruparse entre ellas más fuertemente que entre la otra condición problema.

Continuamos realizando la **transformación estabilizadora de la varianza**, con la finalidad de reducir los valores de la muestra de aquellos genes poco expresados y con alta varianza. Esta transformación se realiza para evitar que la variabilidad de los valores esté relacionada con su valor medio (Durbin et al. 2002).

	CV-1	CV-2	CV-3	CV-4	CV-5	CV-6	CV-7	CV-8	CV-9	CV-10
CV-1	1.0000000	0.9692306	0.9719701	0.9437776	0.9792478	0.9107169	0.9495537	0.9702278	0.9586652	0.9458900
CV-2	0.9692306	1.0000000	0.9598242	0.9365297	0.9651217	0.8968063	0.9401534	0.9650722	0.9617587	0.9536379
CV-3	0.9719701	0.9598242	1.0000000	0.9572848	0.9737660	0.9078488	0.9547917	0.9745770	0.9675473	0.9429722
CV-4	0.9437776	0.9365297	0.9572848	1.0000000	0.9597447	0.9546031	0.9691257	0.9635736	0.9627411	0.9467402
CV-5	0.9792478	0.9651217	0.9737660	0.9597447	1.0000000	0.9318725	0.9610044	0.9764208	0.9640771	0.9533113
CV-6	0.9107169	0.8968063	0.9078488	0.9546031	0.9318725	1.0000000	0.9423514	0.9231543	0.9212455	0.9250617
CV-7	0.9495537	0.9401534	0.9547917	0.9691257	0.9610044	0.9423514	1.0000000	0.9616826	0.9598063	0.9499131
CV-8	0.9702278	0.9650722	0.9745770	0.9635736	0.9764208	0.9231543	0.9616826	1.0000000	0.9793080	0.9450902
CV-9	0.9586652	0.9617587	0.9675473	0.9627411	0.9640771	0.9212455	0.9598063	0.9793080	1.0000000	0.9449949
CV-10	0.9458900	0.9536379	0.9429722	0.9467402	0.9533113	0.9250617	0.9499131	0.9450902	0.9449949	1.0000000

Figura 37. Correlación tras *Transformación estabilizadora de la varianza*, se muestra el grupo CV

Una vez obtenidos los recuentos normalizados se procede a la comparación de éstos entre las diferentes muestras, donde se podrá apreciar que tan similares son éstas entre si con respecto a la expresión génica, evaluando así la calidad del experimento.

Su realización se lleva a cabo mediante métodos de visualización para el análisis de agrupamiento no supervisado, concretamente con Heatmap y PCA. Con ellos se adquiere una inducción de cuán similares son las réplicas biológicas entre sí y se podrá identificar muestras atípicas, al igual que las principales fuentes de variación en el dataset.

Se visualiza a continuación, a través de “heatmaps”, los valores de correlación entre muestras :

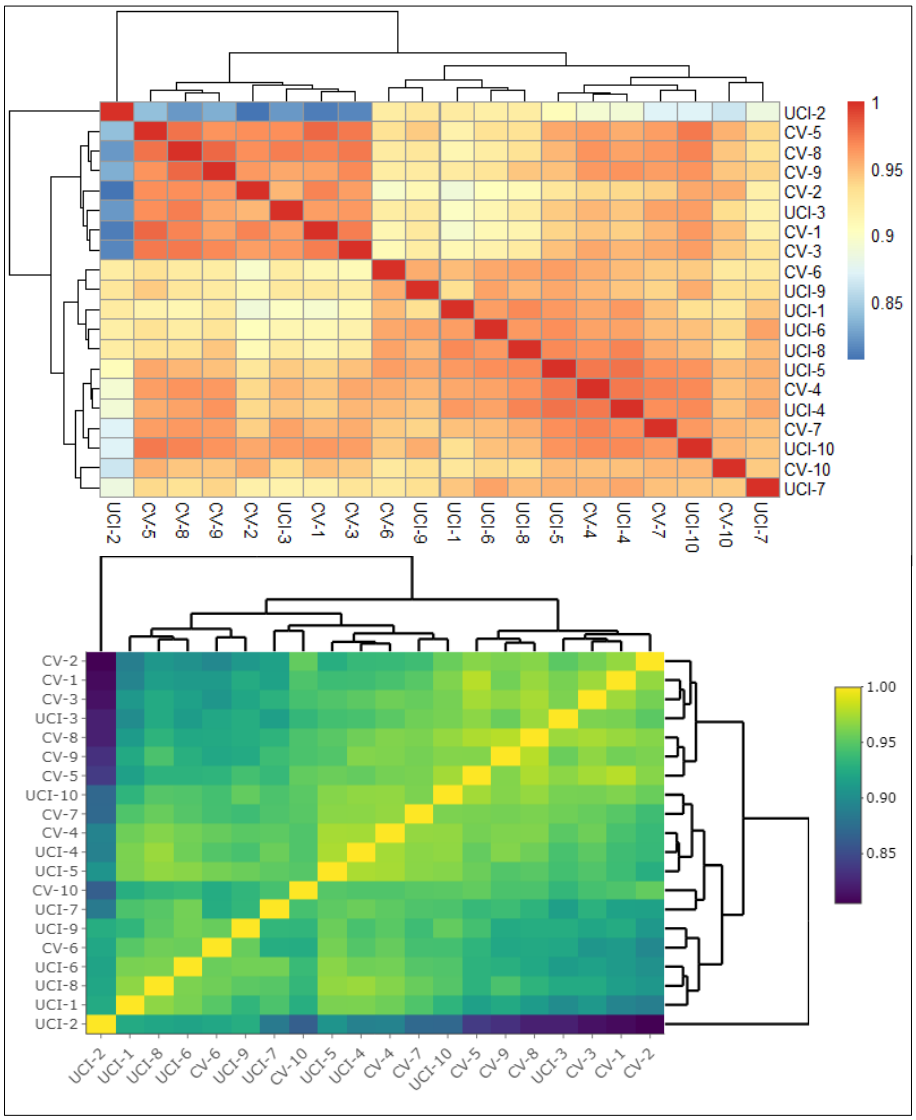


Figura 38. Heatmaps (CV-UCI)

Con ellos podemos evaluar las similitudes y diferencias en la expresión génica entre las diferentes muestras del conjunto de datos.

Se genera posteriormente el mencionado gráfico PCA. El Análisis de Componentes Principales (PCA) sirve para verificar el agrupamiento de muestras, además de confirmar si cumplen con el diseño experimental.
Es una técnica estadística utilizada para describir un conjunto de datos en términos de una nueva variable no correlacionada, pues reduce la bidimensionalidad de los datos, aumentando así la interpretabilidad y a la vez minimizando la pérdida de información. Para ello, crea nuevas variables no correlacionadas que maximizan sucesivamente la varianza (Jollife et al. 2016).

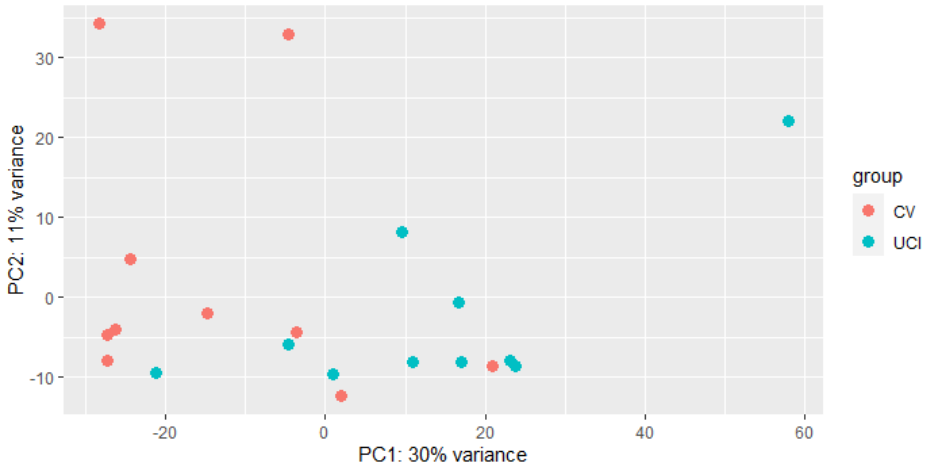


Figura 39. PCA (CV-UCI)

Existe un agrupamiento discreto, sin embargo, si se observa como la mayoría de los datos pertenecientes a las condiciones “UCI” y “CV” tienden a permanecer en regiones separadas del gráfico.
Esto podría indicar que la expresión genética puede verse influenciada por los grupos (condiciones), aunque el valor de PC1 (30%) parece indicar que existen otras fuentes de variación importantes además de las condiciones mencionadas.

4.5. Análisis de expresión diferencial

Generación de resultados

El análisis de expresión diferencial propiamente dicho se inicia ajustando los conteos sin procesar al modelo DESeq2, mediante la función *DESeq* aplicada al objeto generado anteriormente con la información de recuentos.

En los estudios de RNA-seq se espera que la varianza aumente con la expresión media de cada gen, por lo que para observar dicha relación, calcularemos las medias y varianzas para cada gen, tras lo que crearemos un marco de datos para trazar un gráfico de dispersión a partir de dichos valores.

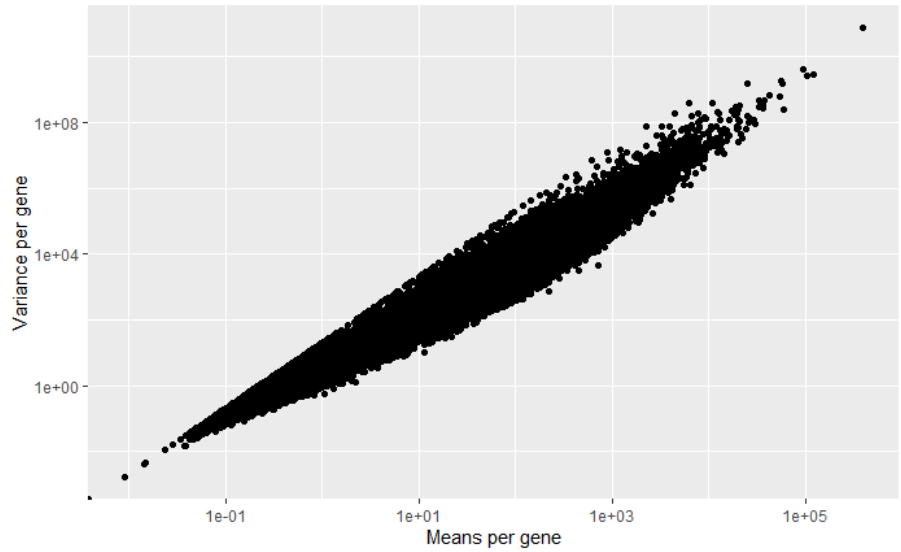


Figura 40. Gráfico de dispersión

En el gráfico de dispersión generado, cada punto es la representación de un gen. Se observa que la varianza en la expresión génica aumenta a medida que aumenta la media. Una medida de la varianza para una media dada se escribe mediante la métrica llamada “dispersión en el modelo DESeq2”, el cual usa la dispersión para evaluar la variabilidad en la expresión. Por lo tanto, un aumento en la varianza produce un aumento en la dispersión.

A continuación se realiza el gráfico de dispersión estimada.

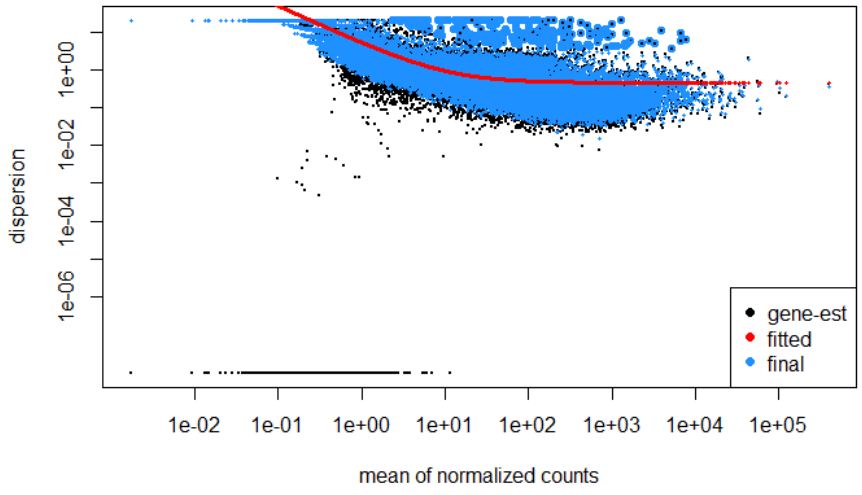


Figura 41. Gráfico de dispersión estimada

El presente gráfico sirve para comprobar el ajuste de los datos al modelo. La curva que se muestra como una línea roja, tiene la estimación del valor de dispersión esperado para los genes de un valor de expresión dado.

Cada uno de los puntos negros dibujados representan genes con valores medios y de dispersión asociados, donde se aprecia que la dispersión disminuye a medida que aumenta la media (comportamiento esperado).

Los puntos azules representan la estimación más precisa de la dispersión. Se comprueba que además del mencionado comportamiento de la dispersión, existe un buen ajuste de la curva.

La función DESeq realizada anteriormente aplica la fórmula de Wald, donde lleva a cabo comparaciones por pares, con el fin de probar diferencias en la expresión entre dos grupos de muestra para las diferentes condiciones. Posteriormente, se extraen los resultados de la prueba con el uso de la función *result*.

Con estos resultados disponible, se pasa realizar la visualización del cambio de pliegue sobre el nivel de expresión promedio de todas las muestras, para ello se grafica un MA-plot.

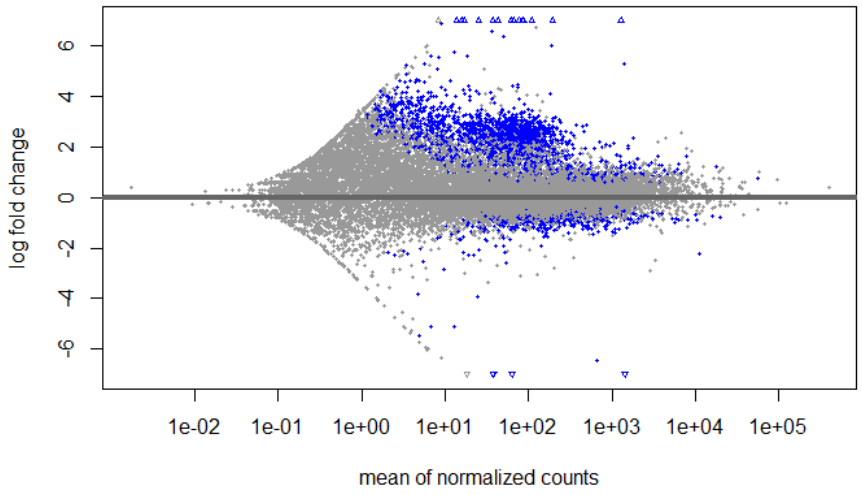


Figura 42. MA-plot

Éste muestra la media de los recuentos normalizados frente a los cambios log2 veces para todos los genes probados.

En azul se evidencian aquellos genes significativamente diferenciados, donde se aprecia la presencia de mayor número de genes con expresión diferencial para valores positivos (UP).

Análisis de resultados

Seguidamente se exploran los resultados obtenidos tras aplicar los diferentes tratamientos.

```
```{r warning=FALSE, message=FALSE}
res<-results(dds,alpha = 0.05)
head(res)
```
log2 fold change (MLE): condition UCI vs CV
Wald test p-value: condition UCI vs CV
DataFrame with 6 rows and 6 columns
                   baseMean log2FoldChange   lfcSE   stat   pvalue   padj
                  <numeric>   <numeric> <numeric> <numeric> <numeric> <numeric>
ENSG00000000003   4.445260   0.531889 0.636491 0.835659 0.40334685 0.6032788
ENSG00000000005   0.351123   0.744828 1.922481 0.387431 0.69843749   NA
ENSG00000000419 742.936462   0.168565 0.130457 1.292110 0.19631908 0.3931323
ENSG00000000457 119.327732   0.172668 0.136422 1.265685 0.20562581 0.4041080
ENSG00000000460   50.426158   0.283779 0.233940 1.213044 0.22511298 0.4267799
ENSG00000000938 3928.653699   0.804573 0.304526 2.642054 0.00824049 0.0540181
```

Figura 43. Resultados

Este código aporta como ejemplo un fragmento de los resultados obtenidos tras ejecutar el análisis de expresión diferencial mediante DESeq2, donde se visualiza una serie de columnas con la siguiente información:

- `baseMean`: Muestra la media de lecturas normalizadas de cada tránscrito

- `log2FoldChange`: Muestra el cambio en la proporción de lecturas, en función del log_2, para ambas condiciones

  Valores `log2FoldChange` positivos indica genes sobre-expresados

  Valores `log2FoldChange` negativos indica genes infra-expresados

- `lfcSE`: Indica el error estándar

- `stat`: Estadístico correspondiente al test utilizado. En este caso se ha utilizado el test paramétrico de Wald para comprobar la expresión diferencial de cada tránscrito

- `pvalue`: Valor de *p* según distribución binomial

- `padj`: Valor de *p* ajustado

Mediante el uso de la función *summary* se obtiene un resumen de los resultados, gracias al cual se puede comprobar el número de tránscritos expresados de forma dispar según las dos condiciones presentes. En dicho resumen se observa el número de tránscritos que presentan un log2FoldChange (LFC) positivo o negativo con un p-valor ajustado inferior al nivel de significación α=0,05.

```
out of 22198 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 1996, 9%
LFC < 0 (down)     : 406, 1.8%
outliers [1]       : 0, 0%
low counts [2]     : 5473, 25%
(mean count < 1)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Figura 44. Resumen de resultados

Se observa el porcentaje de genes (tanto regulados hacia arriba, UP; como hacia abajo, DOWN) que se expresan diferencialmente, donde se puede comprobar la presencia **2402** genes diferencialmente expresados, **1996** "hacia arriba" y **406** "hacia abajo" (visualizado en Figura 38).

Una vez ordenados estos resultados (según el p-valor ajustado) es posible realizar una representación gráfica de la expresión diferencial vista.

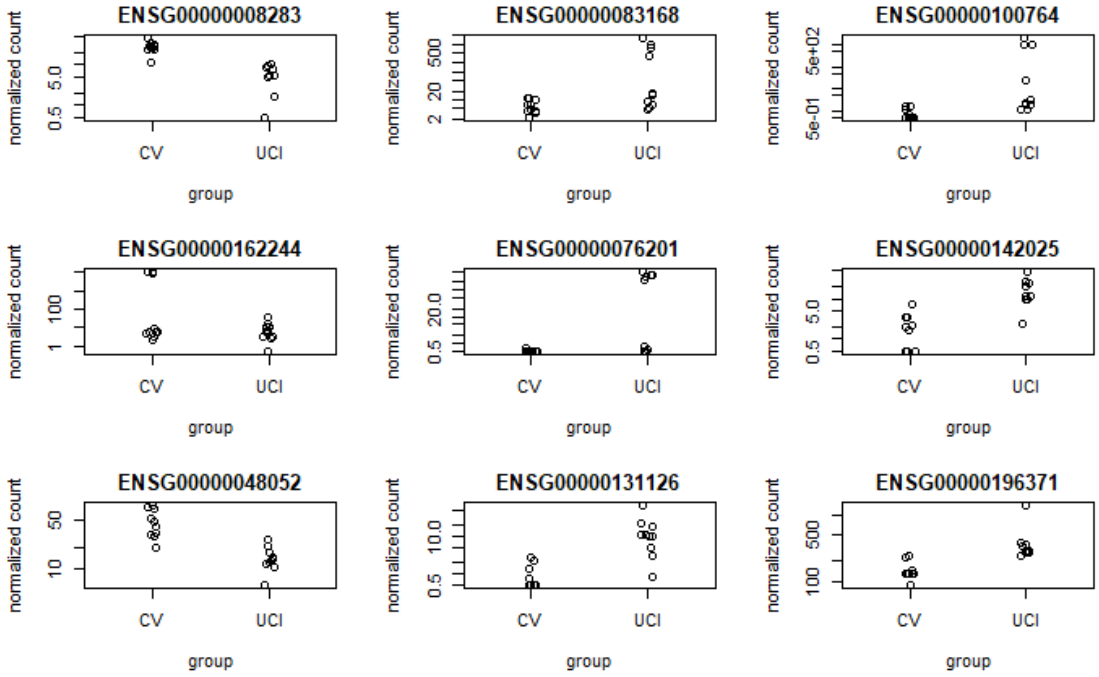


Figura 45. Expresión diferencial de diversos tránscritos

Se representan varios tránscritos con diferencias significativas en la expresión, gracias a lo que se puede inferir en qué condición se produce una sobre-expresión (tomando de base la condición “UCI”) del gen presente.

Anotación

Para comprender a que genes corresponden los resultados, se utiliza el paquete *org.Hs.eg.db*, que nos permite la obtención de los nombres de los diferentes datos a través de ensembl.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	entrez	genname	ensembl
ENSG00000118680	1276.683891	28.7523208	2.9566268	9.724704	2.365924e-22	3.974989e-18	MYL12B	103910	myosin light chain 12B	ENSG00000118680
ENSG00000225989	63.346428	-25.1406118	2.9569233	-8.502288	1.858907e-17	1.053490e-13	ABCF1	23	ATP binding cassette subfamily F member 1	ENSG00000225989
ENSG00000206403	86.593534	25.0003681	2.9568318	8.455120	2.788003e-17	1.053490e-13	ABHD16A	7920	abhydrolase domain containing 16A, phospholipase	ENSG00000206403
ENSG00000275610	86.239021	24.9616826	2.9568324	8.442035	3.118603e-17	1.053490e-13	DUX4L21	102723518	double homeobox 4 like 21 (pseudogene)	ENSG00000275610
ENSG00000258834	77.912286	24.8435121	2.9568560	8.402003	4.389279e-17	1.053490e-13	DUX4L4	441056	double homeobox 4 like 4 (pseudogene)	ENSG00000258834
ENSG00000274599	77.912286	24.8435121	2.9568560	8.402003	4.389279e-17	1.053490e-13	DUX4L24	102723449	double homeobox 4 like 24 (pseudogene)	ENSG00000274599
ENSG00000280337	77.912286	24.8435121	2.9568560	8.402003	4.389279e-17	1.053490e-13	DUX4L25	102723423	double homeobox 4 like 25 (pseudogene)	ENSG00000280337
ENSG00000281652	62.589541	24.5386268	2.9569157	8.298724	1.052358e-16	2.210083e-13	DUX4L7	653543	double homeobox 4 like 7 (pseudogene)	ENSG00000281652
ENSG00000235030	38.118521	-24.4493756	2.9571260	-8.267952	1.362799e-16	2.429068e-13	IER3	8870	immediate early response 3	ENSG00000235030
ENSG00000214425	37.579173	-24.4286304	2.9571393	-8.260900	1.445788e-16	2.429068e-13	*NA*	*NA*	*NA*	ENSG00000214425
ENSG00000164597	66.966245	23.6425207	2.9568949	7.995726	1.288127e-15	1.967439e-12	COG5	10466	component of oligomeric golgi complex 5	ENSG00000164597
ENSG00000276681	25.332162	23.2737844	2.9573698	7.869758	3.553282e-15	4.974892e-12	LENG8	114823	leukocyte receptor cluster member 8	ENSG00000276681

Figura 46. Transcritos ordenados por padj y nominados por ensembl

Visualización de resultados

A continuación se visualizan los resultados a través de un Volcano-plot, que permite la identificación visual rápida de genes con grandes cambios de pliegues que también son estadísticamente significativos. Los puntos serán coloreados según si el gen está o no expresado de forma significativa, donde los más regulados al alza están hacia la derecha, los más regulados hacia abajo están hacia la izquierda y los genes más estadísticamente significativos están hacia la parte superior del gráfico.

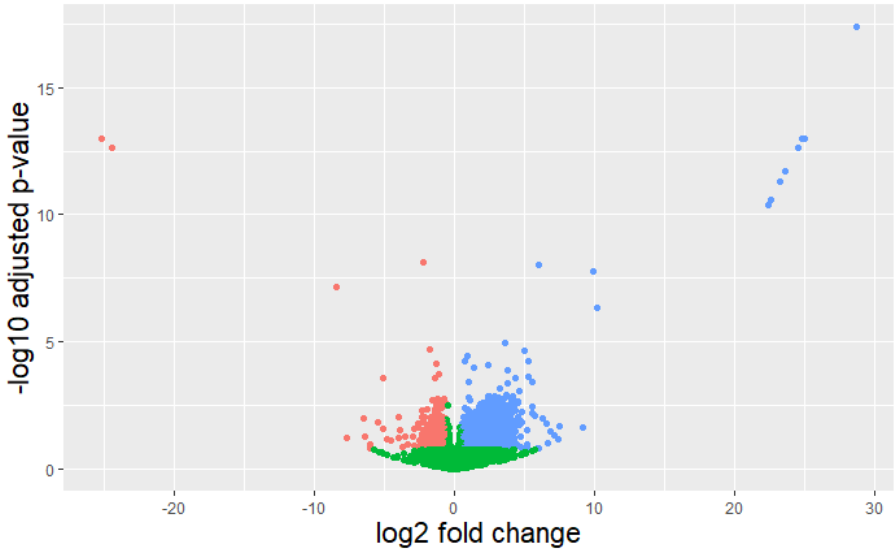


Figura 47. Volcano plot de resultados (en azul: expresados UP, en rojo: expresados DOWN, en verde: sin expresión diferencial)

Se visualiza a continuación aquellos genes más representativos en cuanto a las diferencias en expresión.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	entrez	genname	ensembl
ENSG00000118680	1276.68389	28.75232	2.956627	9.724704	2.365924e-22	3.974989e-18	MYL12B	103910	myosin light chain 12B	ENSG00000118680
ENSG00000225989	63.34643	-25.14061	2.956923	-8.502288	1.858907e-17	1.053490e-13	ABCF1	23	ATP binding cassette subfamily F member 1	ENSG00000225989
ENSG00000206403	86.59353	25.00037	2.956832	8.455120	2.788003e-17	1.053490e-13	ABHD16A	7920	abhydrolase domain containing 16A, phospholipase	ENSG00000206403
ENSG00000275610	86.23902	24.96168	2.956832	8.442035	3.118603e-17	1.053490e-13	DUX4L21	102723518	double homeobox 4 like 21 (pseudogene)	ENSG00000275610
ENSG00000258834	77.91229	24.84351	2.956856	8.402003	4.389279e-17	1.053490e-13	DUX4L4	441056	double homeobox 4 like 4 (pseudogene)	ENSG00000258834
ENSG00000274599	77.91229	24.84351	2.956856	8.402003	4.389279e-17	1.053490e-13	DUX4L24	102723449	double homeobox 4 like 24 (pseudogene)	ENSG00000274599
ENSG00000280337	77.91229	24.84351	2.956856	8.402003	4.389279e-17	1.053490e-13	DUX4L25	102723423	double homeobox 4 like 25 (pseudogene)	ENSG00000280337
ENSG00000281652	62.58954	24.53863	2.956916	8.298724	1.052358e-16	2.210083e-13	DUX4L7	653543	double homeobox 4 like 7 (pseudogene)	ENSG00000281652
ENSG00000235030	38.11852	-24.44938	2.957126	-8.267952	1.362799e-16	2.429068e-13	IER3	8870	immediate early response 3	ENSG00000235030
ENSG00000214425	37.57917	-24.42863	2.957139	-8.260900	1.445788e-16	2.429068e-13	*NA*	*NA*	*NA*	ENSG00000214425

	CV.1	CV.2	CV.3	CV.4	CV.5	CV.6	CV.7	CV.8	CV.9	CV.10	UCI.1	UCI.2	UCI.3	UCI.4	UCI.5	UCI.6	UCI.7	UCI.8	UCI.9	UCI.10
ENSG00000118680	0.0000	0.0000	0.0000	0.0000	0	0.0000	0.0000	0	0	0	0	8924.5050	0.0000	9097.1489	0	0	0.000	7512.0239	0	0.0000000
ENSG00000225989	518.7119	0.0000	416.8539	0.0000	0	0.0000	331.3627	0	0	0	0	0.0000	0.0000	0.0000	0	0	0.000	0.0000	0	0.0000000
ENSG00000206403	0.0000	0.0000	0.0000	0.0000	0	0.0000	0.0000	0	0	0	0	0.0000	385.4281	660.3109	0	0	0.000	685.4500	0	0.6815895
ENSG00000275610	0.0000	0.0000	0.0000	0.0000	0	0.0000	0.0000	0	0	0	0	557.0053	0.0000	0.0000	0	0	744.645	423.1302	0	0.0000000
ENSG00000258834	0.0000	0.0000	0.0000	0.0000	0	0.0000	0.0000	0	0	0	0	265.4321	0.0000	0.0000	0	0	744.645	548.1687	0	0.0000000
ENSG00000274599	0.0000	0.0000	0.0000	0.0000	0	0.0000	0.0000	0	0	0	0	265.4321	0.0000	0.0000	0	0	744.645	548.1687	0	0.0000000
ENSG00000280337	0.0000	0.0000	0.0000	0.0000	0	0.0000	0.0000	0	0	0	0	265.4321	0.0000	0.0000	0	0	744.645	548.1687	0	0.0000000
ENSG00000281652	0.0000	0.0000	0.0000	0.0000	0	0.0000	0.0000	0	0	0	0	201.0849	0.0000	0.0000	0	0	589.564	461.1419	0	0.0000000
ENSG00000235030	107.8797	140.7902	0.0000	261.8741	0	0.0000	251.8264	0	0	0	0	0.0000	0.0000	0.0000	0	0	0.000	0.0000	0	0.0000000
ENSG00000214425	337.1647	282.9493	0.0000	0.0000	0	131.4695	0.0000	0	0	0	0	0.0000	0.0000	0.0000	0	0	0.000	0.0000	0	0.0000000

Figura 48. Visualización de resultados (10 primeros genes listados por p-value y 10 primeros genes con los valores del conteo para cada muestra)

Estos genes, y su expresión diferencial pueden ser visualizados a partir del siguiente gráfico de expresión.

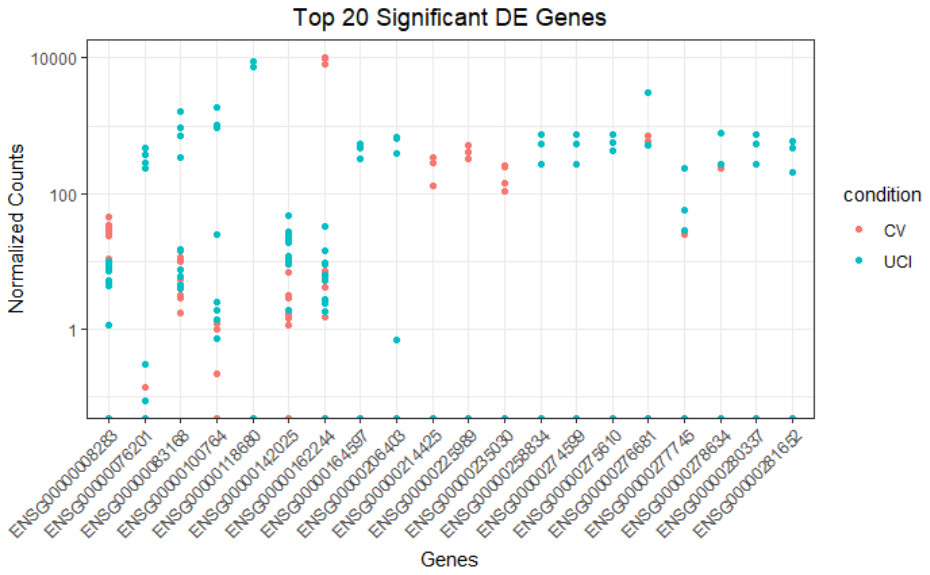


Figura 49. Expresión diferencial de los 20 genes estadísticamente más significativos

Se aprecia cómo, de forma general (con algunas excepciones), los conteos se agrupan según condición para cada uno de los genes, lo que indica que en el gen representado, la expresión difiere entre pacientes provenientes de planta hospitalaria (CV) y aquellos provenientes de la Unidad de Cuidados Intensivos (UCI).

4.6. Enriquecimiento

El enriquecimiento comienza con la carga de todas aquellas bibliotecas que se utilizarán junto con la creación del objeto “OrgDb”, el cual representa al organismo problema (*Homo sapiens*). Además del objeto “gene”, que extrae de los resultados, aquella información de interés para el posterior enriquecimiento (principalmente el código entrez de los genes ordenados según el p-valor).

En *clusterProfiler*, la función *groupGO* está diseñada para la clasificación de genes basada en la distribución de Gene Ontology (GO) a un nivel específico. Se muestra a continuación la clasificación GO, es decir, una serie de ejemplos indicativos de diferentes niveles encontrados en GO.

- BP. Proceso biológico

	ID	Description	Count	GeneRatio	geneID
GO:0019953	GO:0019953	sexual reproduction	885	885/25168	DMRTC2/TEX101/CATSPER4/ODF4/HSPA2/SPACA6/P2...
GO:0019954	GO:0019954	asexual reproduction	0	0/25168	
GO:0022414	GO:0022414	reproductive process	1487	1487/25168	RPL29/DMRTC2/TEX101/EXD1/CATSPER4/ODF4/HSP...
GO:0032504	GO:0032504	multicellular organism reproduction	865	865/25168	DMRTC2/CATSPER4/ODF4/HSPA2/NCOR2/SPACA6/P2...
GO:0032505	GO:0032505	reproduction of a single-celled organism	2	2/25168	RTF2/ZNF445
GO:0075325	GO:0075325	spore dispersal	0	0/25168	
GO:0001776	GO:0001776	leukocyte homeostasis	77	77/25168	ANXA1/ADA/SPNS2/TNFAIP3/CXCL6/SLC7A11/ZC3H8...
GO:0002200	GO:0002200	somatic diversification of immune receptors	73	73/25168	PTPRC/SUPT6H/TBX21/KMT5C/CD40LG/SAMHD1/CT...
GO:0002252	GO:0002252	immune effector process	1465	1465/25168	LILRA2/AMPD3/PECAM1/CD96/SLC22A13/SLAMF6/IF...
GO:0002253	GO:0002253	activation of immune response	639	639/25168	LILRA2/PSMC1/CARD11/MUC16/IGLC1/PTPRC/ADA/G...

Figura 50. Gene Ontology . BM

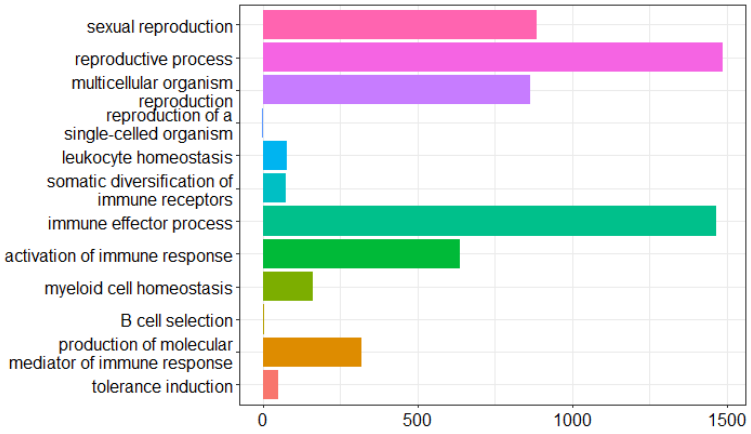


Figura 51. Gene Ontology . BM, plot

- CC. Componente celulares

	ID	Description	Count	GeneRatio	geneID
GO:0044423	GO:0044423	virion part	0	0/25168	
GO:0000133	GO:0000133	polarisome	0	0/25168	
GO:0000408	GO:0000408	EKC/KEOPS complex	7	7/25168	OSGEPL1/OSGEP/LAGE3/TP53RK/GON7/GON7/TPRKB
GO:0000417	GO:0000417	HIR complex	1	1/25168	HIRA
GO:0000444	GO:0000444	MIS12/MIND type complex	4	4/25168	DSN1/PMF1/MIS12/NSL1
GO:0000808	GO:0000808	origin recognition complex	9	9/25168	ORC5/ORC2/REPIN1/ORC1/ORC4/LRWD1/ORC6/MCM...
GO:0000930	GO:0000930	gamma-tubulin complex	24	24/25168	MZT2A/BLOC1S2/TUBG1/TUBGCP3/TOPORS/TUBG2/...
GO:0000939	GO:0000939	condensed chromosome inner kinetochore	5	5/25168	ORC2/DSN1/CENPO/CENPK/CENPA
GO:0000940	GO:0000940	condensed chromosome outer kinetochore	15	15/25168	PLK1/NUP133/CENPF/SPDL1/SKA3/CCNB1/NDC80/BO...
GO:0000974	GO:0000974	Prp19 complex	13	13/25168	CTNNBL1/PRPF19/CRNKL1/CDC5L/HSPA8/PLRG1/POL...
GO:0001114	GO:0001114	protein-DNA-RNA complex	0	0/25168	
GO:0001534	GO:0001534	radial spoke	7	7/25168	CFAP251/CFAP206/RSPH9/CFAP91/RSPH6A/RSPH4A/...
GO:0001535	GO:0001535	radial spoke head	0	0/25168	

Figura 52. Gene Ontology . CC

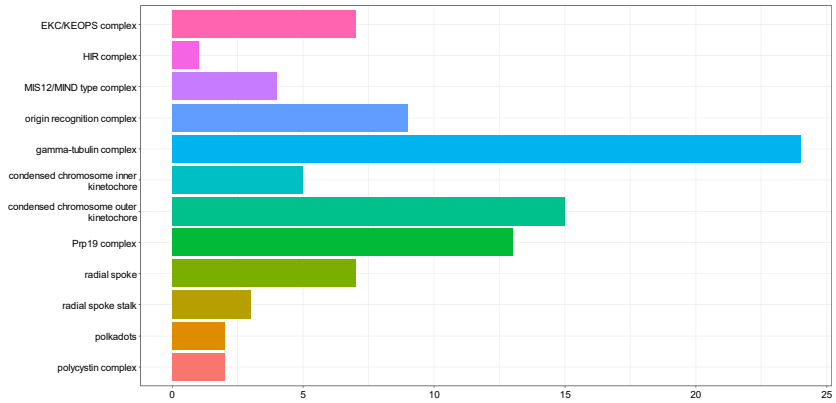


Figura 53. Gene Ontology .CC, plot

- MF. Función molecular

	ID	Description	Count	GeneRatio	geneID
GO:0004133	GO:0004133	glycogen debranching enzyme activity	1	1/25168	AGL
GO:0009975	GO:0009975	cyclase activity	22	22/25168	ADCY6/NPR2/NPR1/ADCY9/TKFC/GUCY2D/ADCY7/AD...
GO:0016491	GO:0016491	oxidoreductase activity	772	772/25168	CYB561/OGFOD3/CRYM/NXN/HSD3B1/NDUFA3/COX1...
GO:0016740	GO:0016740	transferase activity	2359	2359/25168	KAT6A/FUT4/CILK1/CHRAC1/GLT8D1/COX10/IFNB1/C...
GO:0016787	GO:0016787	hydrolase activity	2681	2681/25168	ABCF1/ABHD16A/PSMC1/PTPN23/HDAC9/TUBB6/ATP...
GO:0016829	GO:0016829	lyase activity	198	198/25168	SDS/CA1/HTD2/CD38/ADCY6/RPS3/ECHS1/CENPVL3/...
GO:0016853	GO:0016853	isomerase activity	163	163/25168	HSD3B1/DSE/PPIAL4D/ERO1A/PPIAL4H/IDI1/PPIAL4G...
GO:0016874	GO:0016874	ligase activity	180	180/25168	LARS1/ATP5F1D/TTLL1/SUCLG2/ATP5F1A/SLC27A5/...
GO:0032451	GO:0032451	demethylase activity	39	39/25168	KDM4E/RIOX2/MMACHC/KDM3A/ALKBH3/KDM2B/KD...
GO:0061783	GO:0061783	peptidoglycan muralytic activity	14	14/25168	LYG1/PGLYRP1/LYZ/PGLYRP4/LYZL4/LYZL1/PGLYRP2/L...
GO:0106265	GO:0106265	THPH synthase activity	0	0/25168	
GO:0140096	GO:0140096	catalytic activity, acting on a protein	2371	2371/25168	ABHD16A/KAT6A/PTPN23/HDAC9/CILK1/USP28/C1GA...

Figura 50. Gene Ontology . MF

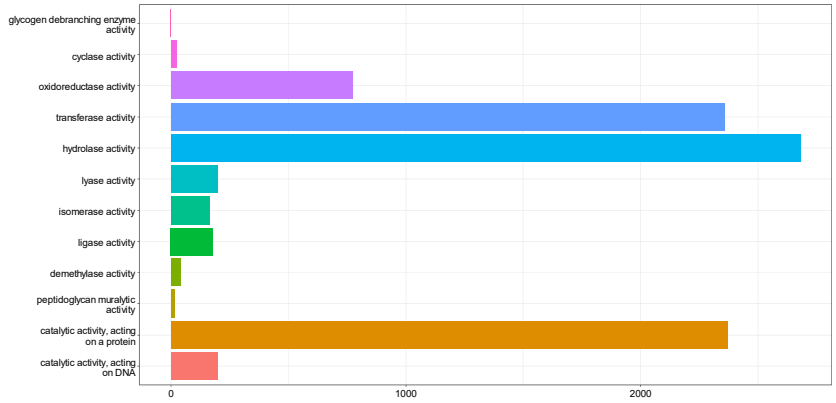


Figura 54. Gene Ontology . MF, plot

Seguidamente se realiza el análisis de enriquecimiento para asignar a nuestros datos la clasificación correspondiente. Para ello se utiliza la prueba hipergeométrica para sobre-representación como método estadístico, empleado para verificar si los genes de interés se asocian con más frecuencia a determinadas funciones biológicas de lo que cabría esperar en un conjunto aleatorio de genes.

Una vez establecidas las clasificaciones GO se procede a clasificar los genes obtenidos tras el análisis de expresión diferencial. Esto se llevará a cabo en las tres categorías recientemente mencionadas, mediante la función *enrichGO*

- BP. Proceso biológico

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0016569	GO:0016569	covalent chromatin modification	460/17108	461/18862	7.933362e-19	5.033718e-15	3.496525e-15	KAT6A/HDAC9/WD...	460
GO:0016570	GO:0016570	histone modification	447/17108	448/18862	2.833356e-18	8.988823e-15	6.243823e-15	KAT6A/HDAC9/WD...	447
GO:0023061	GO:0023061	signal release	471/17108	475/18862	1.101906e-15	2.044764e-12	1.420336e-12	PTPN23/SLC44A4/...	471
GO:0002446	GO:0002446	neutrophil mediated immunity	494/17108	499/18862	1.289056e-15	2.044764e-12	1.420336e-12	AMPD3/PECAM1/P...	494
GO:0000280	GO:0000280	nuclear division	433/17108	436/18862	3.330239e-15	3.607429e-12	2.505795e-12	HSPA2/HNRNPU/M...	433
GO:0002283	GO:0002283	neutrophil activation involved in immune response	483/17108	488/18862	3.472970e-15	3.607429e-12	2.505795e-12	LILRA2/AMPD3/PE...	483
GO:0048285	GO:0048285	organelle fission	481/17108	486/18862	4.157342e-15	3.607429e-12	2.505795e-12	HSPA2/HNRNPU/C...	481
GO:0043312	GO:0043312	neutrophil degranulation	480/17108	485/18862	4.548374e-15	3.607429e-12	2.505795e-12	AMPD3/PECAM1/P...	480
GO:0042119	GO:0042119	neutrophil activation	494/17108	500/18862	1.044819e-14	7.365973e-12	5.116557e-12	LILRA2/AMPD3/PE...	494

Figura 55. Genes clasificados según BP. GO

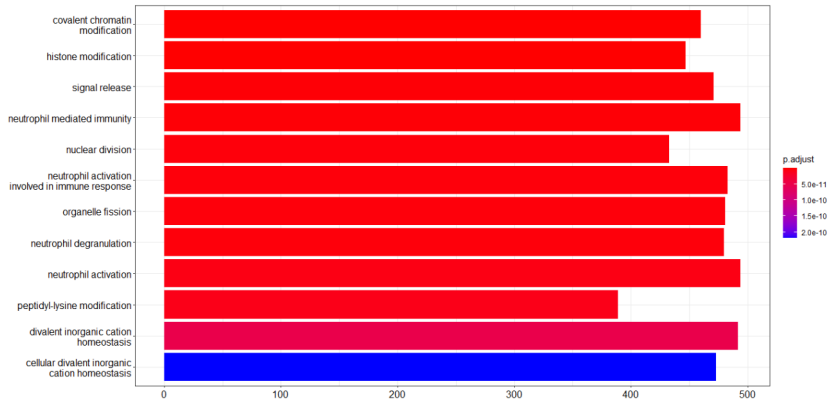


Figura 56. Genes clasificados según BP. GO. Bar-plot

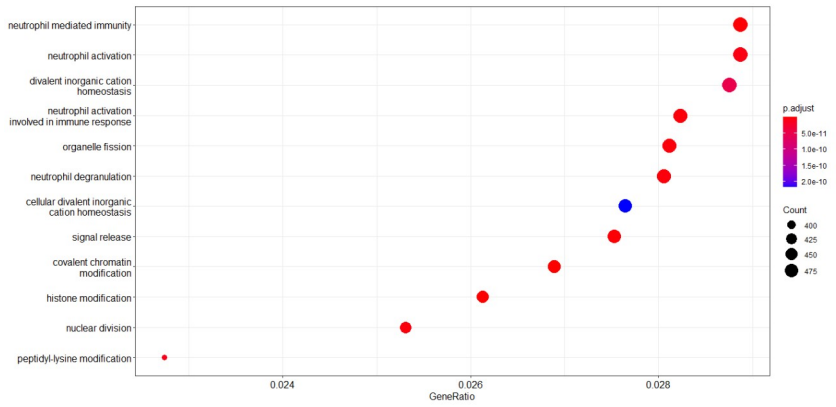


Figura 57. Genes clasificados según BP. GO. Dot-plot

Los gráficos arriba expuestos muestran el número de genes asociados con los primeros términos (representado según el tamaño) y los p-valores ajustado para estos términos (representado según el color). El Dotplot muestra los 12 genes principales por “geneRatio” (genes relacionados con el término GO / número total de genes sig), no por el p-valor ajustado.

Con ello presente, se observa que dos de los términos más llamativos, según el número de genes asociados y el p-valor de dicha asociación serían “neutrophil mediated inmunity” y “neutrophil activation”.

Se hallan reflejados un importante número de genes asociados a términos referidos a la implicación de neutrófilos, lo cual concuerda con la respuesta fisiológica ante una infección vírica.

- CC. Componente celulares

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0031253	GO:0031253	cell projection membrane	337/17962	337/19520	5.190955e-13	1.732234e-10	1.108859e-10	KCNJ11/RPS3/SCIMP/...	337
GO:0045177	GO:0045177	apical part of cell	412/17962	414/19520	5.354079e-13	1.732234e-10	1.108859e-10	SLC44A4/SLC15A1/S...	412
GO:0031252	GO:0031252	cell leading edge	409/17962	411/19520	6.810880e-13	1.732234e-10	1.108859e-10	KCNJ11/CTNND1/RPS...	409
GO:0098793	GO:0098793	presynapse	482/17962	487/19520	2.109040e-12	4.022993e-10	2.575248e-10	GRM7/SYT5/P2RY1/R...	482
GO:0016324	GO:0016324	apical plasma membrane	350/17962	351/19520	5.080116e-12	7.752257e-10	4.962471e-10	SLC44A4/SLC15A1/S...	350
GO:0005911	GO:0005911	cell-cell junction	479/17962	485/19520	1.788020e-11	2.273766e-09	1.455511e-09	PGM5/PECAM1/ANXA...	479
GO:0005874	GO:0005874	microtubule	419/17962	423/19520	3.186936e-11	3.283969e-09	2.102175e-09	TUBB6/HNRNPU/MAP...	419
GO:0005759	GO:0005759	mitochondrial matrix	470/17962	476/19520	3.443218e-11	3.283969e-09	2.102175e-09	SDHAF1/DHRS2/HTD...	470
GO:0043025	GO:0043025	neuronal cell body	468/17962	474/19520	3.981673e-11	3.375574e-09	2.160814e-09	SYT5/KCNJ11/AGRP/R...	468

Figura 58. Genes clasificados según CC. GO

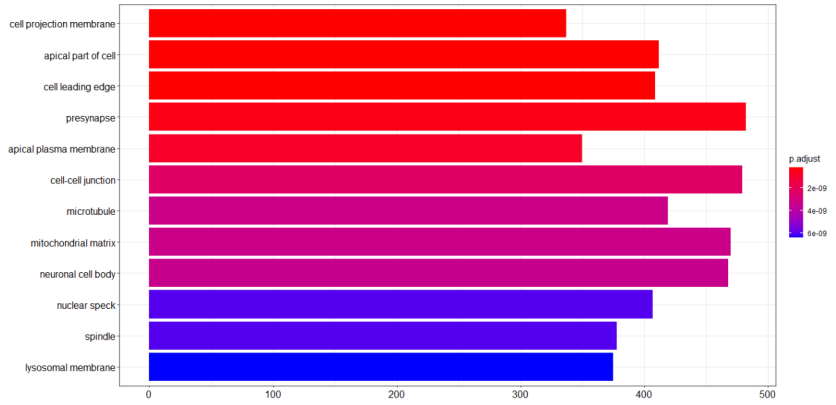


Figura 59. Genes clasificados según CC. GO. Bar-plot

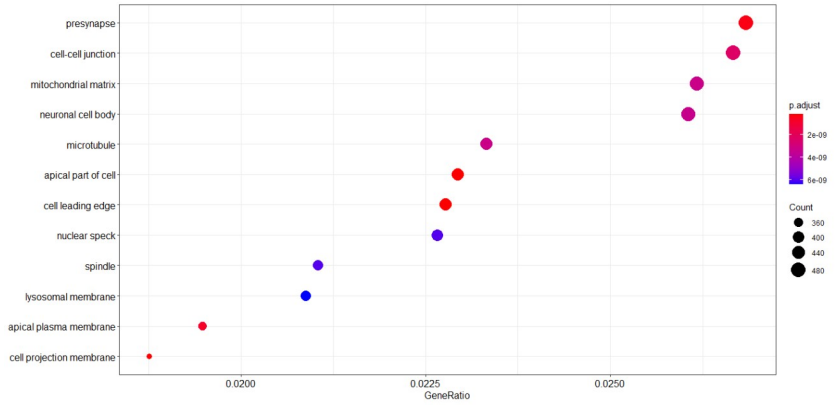


Figura 60. Genes clasificados según CC. GO. Dot-plot

Se observa la clasificación de los genes resultantes de la expresión diferencial, ordenados por p-valor ajustado para una clasificación ontológica según su localización en componentes celulares.

- MF. Función molecular

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0046873	GO:0046873	metal ion transmembrane trans...	422/17421	425/18337	6.066704e-07	0.0007340712	0.0006571198	CATSPER4/ATP2A1/KC...	422
GO:0003712	GO:0003712	transcription coregulator activity	475/17421	480/18337	1.733839e-06	0.0010489725	0.0009390106	KAT6A/HDAC9/CRYM/...	475
GO:0022804	GO:0022804	active transmembrane transpor...	300/17421	301/18337	3.002037e-06	0.0012108215	0.0010838932	SLC15A1/ABCG4/SLC...	300
GO:0022836	GO:0022836	gated channel activity	335/17421	337/18337	4.872642e-06	0.0014196157	0.0012707999	CATSPER4/KCNJ11/KC...	335
GO:0005543	GO:0005543	phospholipid binding	446/17421	451/18337	5.866180e-06	0.0014196157	0.0012707999	AMER2/SYT5/ANXA1/...	446
GO:0015631	GO:0015631	tubulin binding	365/17421	368/18337	7.893875e-06	0.0015919314	0.0014250521	MAP10/FEZ1/KIF3A/M...	365

Figura 61. Genes clasificados según MF. GO

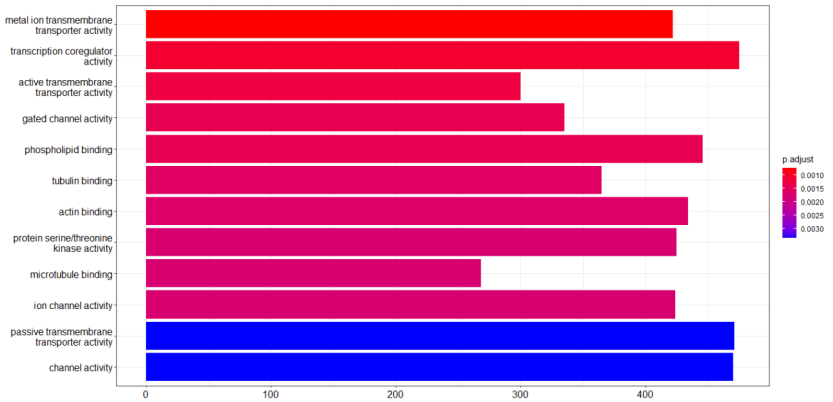


Figura 62. Genes clasificados según MF. GO. Bar-plot

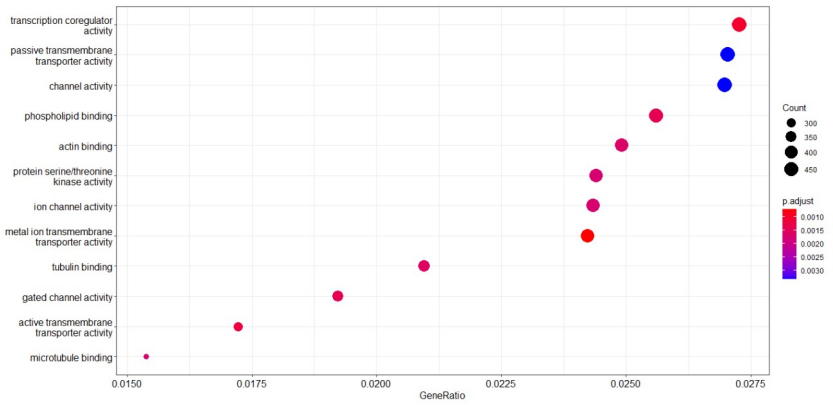


Figura 63. Genes clasificados según MF. GO. Dot-plot

Se puede objetivar la ontología de nuestros genes problemas según su función molecular. Destacan por número de recuentos y por el p-valor, aquellos genes implicados en la corregulación de la transcripción.

Es destacable como según su importancia en la interpretación que atañe al presente trabajo, la clasificación de mayor relevancia es aquella que aporta datos sobre el rol de los genes en los procesos biológicos (BP), pues es de ella de donde se puede inferir información relacionada al objetivo principal de este estudio.

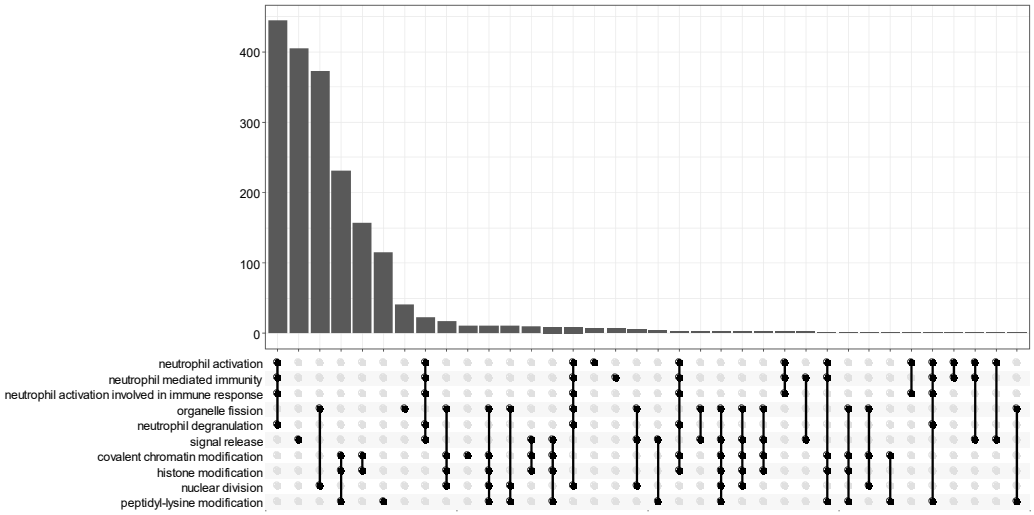


Figura 64. Superposición de genes en diferentes funciones biológicas

Tras esto, se realiza el “Mapa de enriquecimiento”. El mapa de enriquecimiento organiza los términos enriquecidos en una red con bordes que conectan conjuntos de genes superpuestos. De esta manera, los conjuntos de genes que se superponen mutuamente tienden a agruparse, lo que facilita la identificación de módulos funcionales

En este caso se observa la relación entre los 15 términos GO enriquecidos más significativamente (según padj.) por agrupamiento de términos similares. Se realiza para la ontología referente a las funciones biológicas, por lo explicado anteriormente.

El color representa los p-valores en relación con los otros términos mostrados (el rojo más brillante es más significativo) y el tamaño de los términos representa el número de genes que son significativos de nuestra lista.

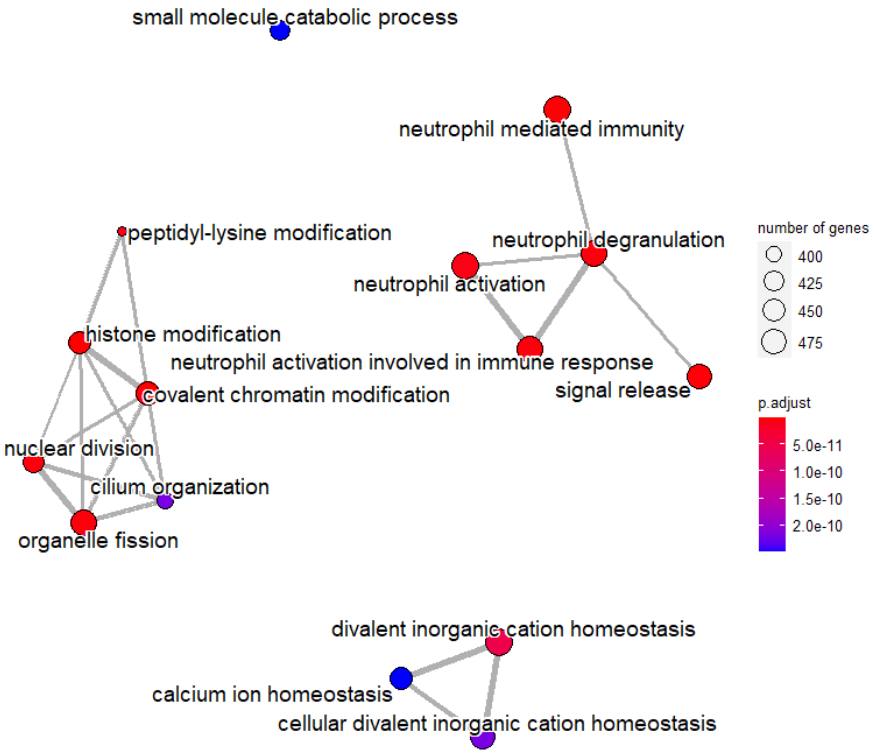


Figura 65. Mapa de enriquecimiento

A continuación un Goplot aportará una visión general de los cuatro términos más relevantes.

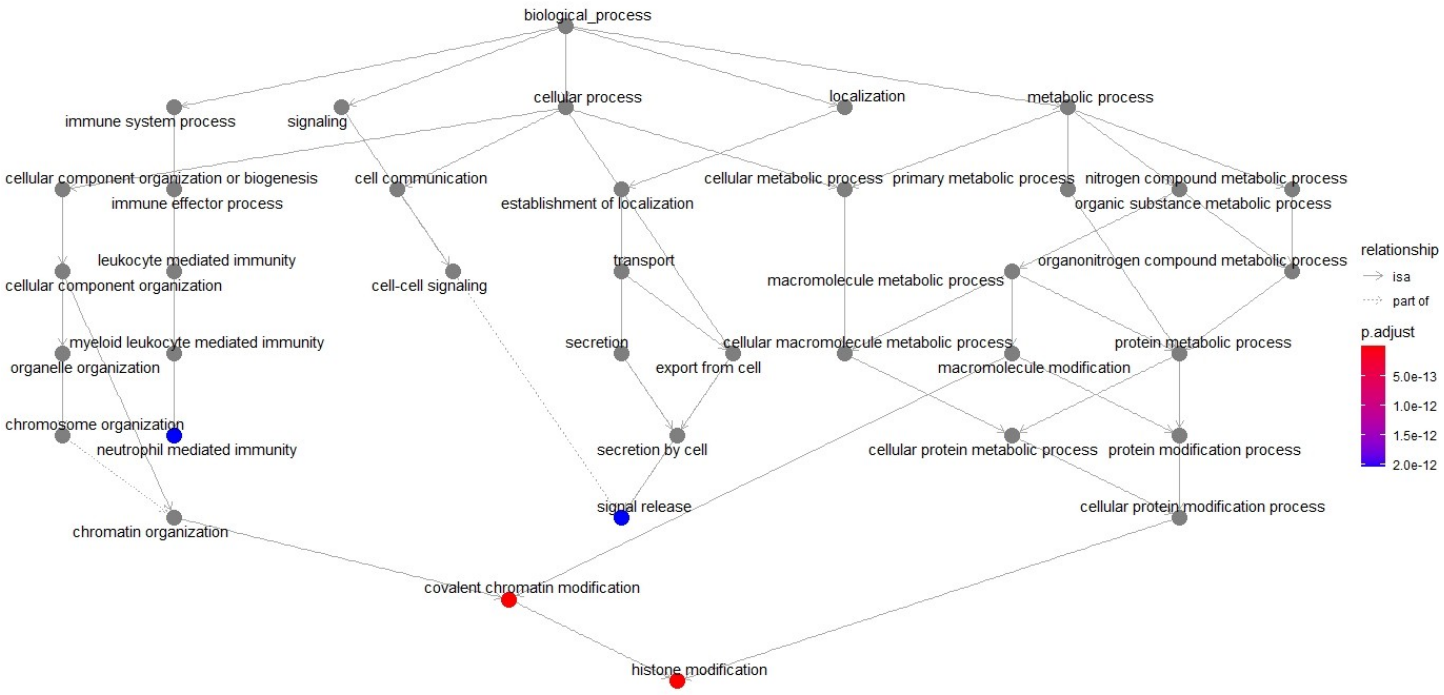


Figura 66. Visualización multicapa del análisis funcional

Por último, el gráfico de red de categorías muestra las relaciones entre los genes asociados con los cinco términos GO más importantes y los cambios de pliegues de los genes importantes asociados con estos términos (color). El tamaño de los términos GO refleja sus p-valores, siendo los términos más significativos aquellos de mayor embergadura.

Esta gráfica es particularmente útil para la generación de hipótesis en la identificación de genes que pueden ser importantes para varios de los procesos más afectados.

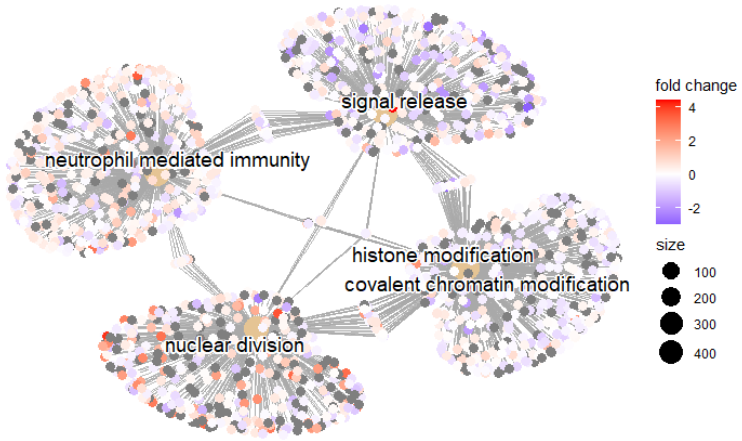


Figura 67. Red de categorías (GO)

El estudio continúa mediante el análisis de enriquecimiento de conjuntos de genes, a través de los paquetes *clusterProfiler* y *Pathview*.

El análisis de sobre-representación es solo un tipo único de método de análisis funcional que está disponible para separar los procesos biológicos importantes para su condición de interés. Otros tipos de pruebas pueden ser igualmente importantes o informativas, incluidos los métodos de puntuación de clases funcionales y de topología de vías.

Las herramientas de puntuación de clase funcional (FCS) utilizan con mayor frecuencia las estadísticas a nivel de gen o cambios log2 para todos los genes a partir de los resultados de expresión diferencial, luego, observan si los conjuntos de genes para vías biológicas particulares están enriquecidos entre los grandes positivos. o cambios de pliegue negativos.
La hipótesis de los métodos FCS postura que, aunque grandes cambios en genes individuales pueden tener efectos significativos en las vías, cambios más débiles pero coordinados en conjuntos de genes relacionados funcionalmente, también pueden tener efectos significativos.

Por lo tanto, en lugar de establecer un umbral arbitrario para identificar "genes significativos", todos los genes se consideran en el análisis. Las estadísticas a nivel de gen del conjunto de datos se agregan para generar una única estadística a nivel de vía y se informa la significación estadística de cada vía.

Por ello, se repite el enriquecimiento, pero esta vez utilizando la base de datos "Kyoto Encyclopedia of Genes and Genomes", KEGG. De esta forma, se seleccionan conjuntos de genes que interactúan en la misma ruta biológica.

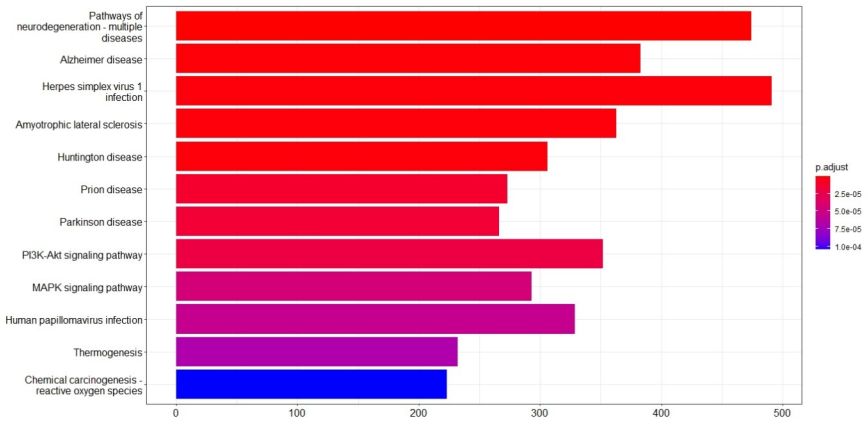


Figura 68. Gráfico de barras (KEGG)

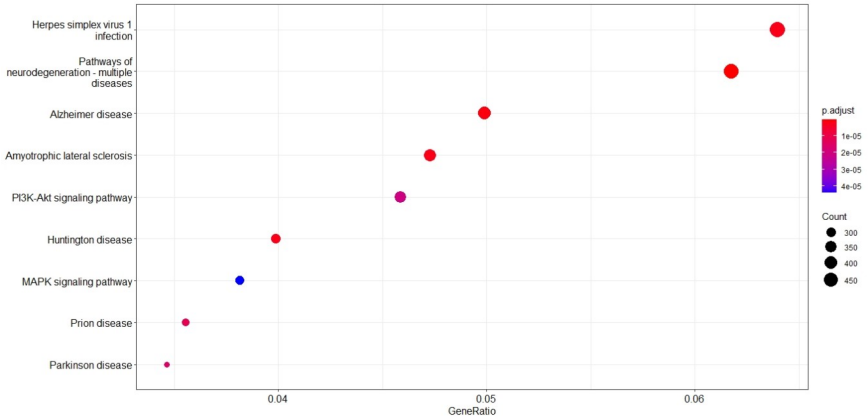


Figura 69. Dotplot (KEGG)

El mayor conteo de genes y con el p-valor más significativo se encuentra en la ruta correspondiente a la infección por “virus herpes simple 1”, que se muestra en el siguiente esquema

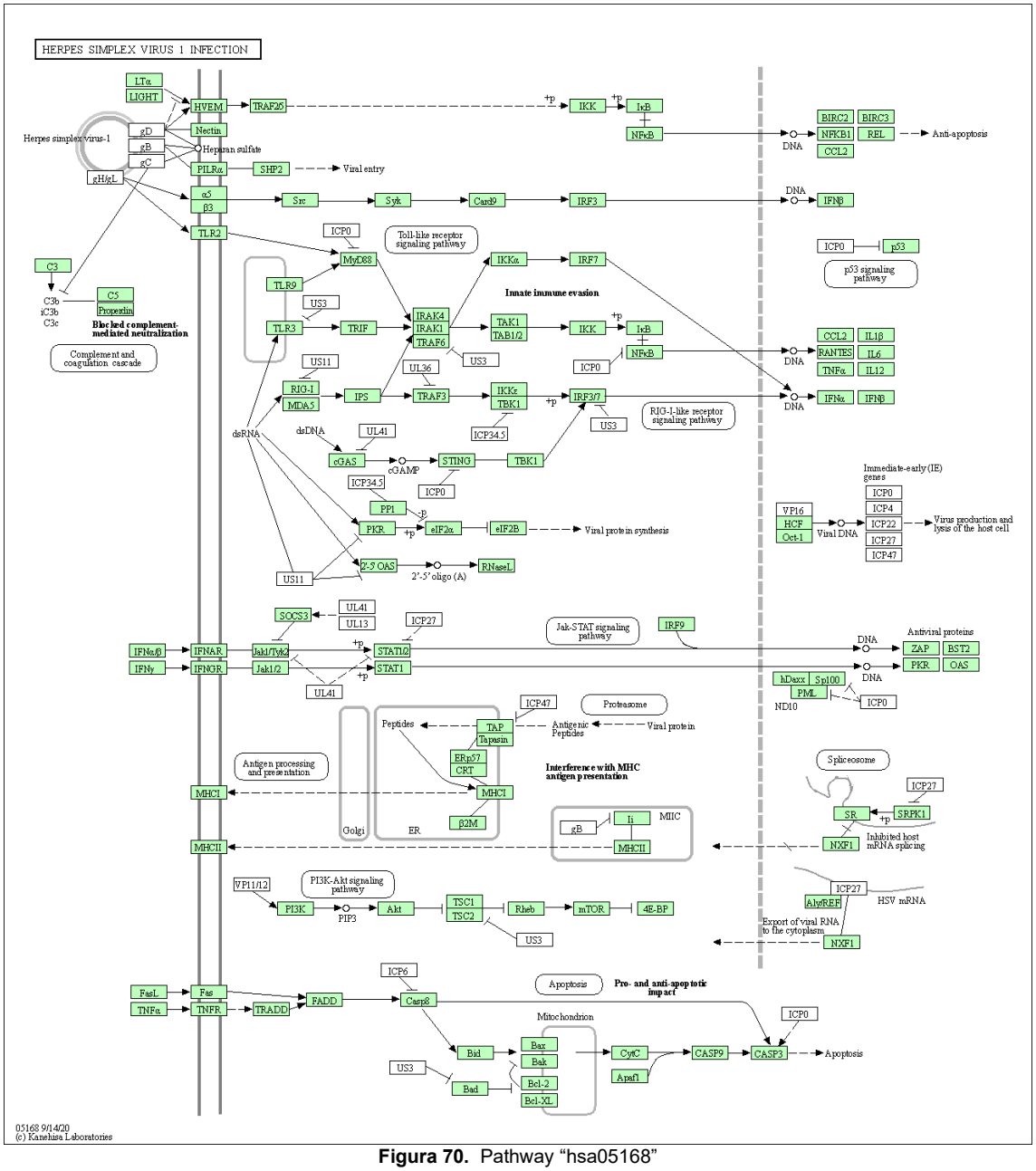


Figura 70. Pathway “hsa05168”

Donde se aprecia la implicación de diversas moléculas y procesos que también se encuentran implicados en la severidad de la infección por Sars-cov-2 (IL6, TNF, IL1, entre otros) tal y como se indica en el apartado “Estado de arte”.

5. Discusión

Existe una urgente necesidad de establecer aquellas causas que puedan explicar los diferentes cursos clínicos vistos en el desarrollo de la infección por SARS-CoV2. Aproximadamente el 20% de los pacientes desarrolla un cuadro clínico grave. De ellos, el 5% se convierten en críticos, precisando ingreso en UCI.

Por ello, a lo largo de este trabajo se ha buscado, con el uso de la transcriptómica, variantes explicativas de la alta diversidad vista en el desarrollo de la enfermedad.

Es de especial relevancia para esta discusión, los resultados obtenidos mediante el análisis de expresión diferencial plasmados en el apartado 4.5. En él se objetivan una serie de genes con expresión diferencial significativa. Se muestran a continuación aquellos que se consideran de mayor interés para los objetivos marcados:

- **PGLYRP1** (Proteína de reconocimiento de peptidoglicano 1, TAG7): Se observa una expresión diferencial de este gen entre los dos grupos en estudio.

  La proteína TAG7 tiene diversas funciones, siendo las más destacable su participación en la respuesta inmune innata y como regulador negativo de la respuesta inflamatoria.

  Teniendo en cuenta dichas funciones y los resultados observados, se presupone un papel relevante en el desarrollo de la enfermedad COVID-19, conjetura ya realizada y analizada por *Tatiana et al. 2021*, donde se evidencia que las proteínas TAG7 interactúan con TNFR1 y TREM-1 (ambos amplificadores de la respuesta inflamatoria), reduciendo significativamente la concentración plasmática de IL-6, IL-1B, TNFα e IFNγ, todas ellas implicadas en la cascada de citocinas previamente mencionada.

- **HDAC9** (Histona deacetilasa 9): Implicada en la activación y diferenciación de linfocitos B y respuesta inflamatoria (regulación negativa de la producción de citocinas) entre otros procesos biológicos.

  Los resultados aportados en este análisis muestran de nuevo una expresión diferencial significativa entre los pacientes en planta y los pacientes en UCI. Una revisión bibliográfica pone de manifiesto esta diferenciación, considerándola un potencial candidato al desarrollo farmacológico en la lucha contra el SARS-CoV2 (Selvaraj et al. 2021).

- **FUT4** (Alfa 1-3-fucosiltransferasa 4): Su proceso biológico más relevante para nuestro fin, es la regulación en la adhesion célula-célula de los linfocitos. Esta función posee una relación lógica en el desarrollo de COVID-19.

  FUT4 se considera un gen proviral, en el que su regulación a la baja podría estar implicada en la inhibición del eje PD1-PDL1. Esto concuerda con la presencia de una mayor expresión de este gen en pacientes críticos ingresados en UCI (Galimberti et al. 2020). Se evidencia en estudios previos cómo su expresión disminuye significativamente ante el tratamiento del paciente con *imatinib*.

Estos tres genes han sido nombrados como discusión final en este trabajo debido a los resultados aportados (todos con expresión diferencial significativa), y por poseer evidencias bibliográficas que apoyan la conclusión.

Sin embargo, con la finalidad de buscar aportación extra por parte de este TFM, se mencionan a continuación algunos genes con expresión diferencial significativa y relacionados con la

actividad inmunológica y/o inflamatoria, que no están respaldados de momento por la bibliografía existente, y que se presupone implicación en el proceso.

- **ABCF1**: Gen implicado en la respuesta inflamatoria (GO:0006954).
  Estudios previos ponen de manifiesto su relación en la respuesta a infecciones víricas (Cao et al. 2020), con lo que su implicación en la infección de SARS-CoV2 (teniendo en cuenta la DE) se supone algo probable.
- **ABHD16A**: Posee actividades de lipasa de acilglicerol y fosfatidilseria, teniendo su ubicación genética en el grupo de genes del MHC III. Ésto sugiere la participación de la proteína en la inmunomodulación del organismo (Xu et al. 2018)
- **IER3**: Factor regulador clave en la respuesta inmune, en el que se ha observado un aumento notable de su expresión en infecciones víricas (Villalba et al. 2017). Se ha evidenciado como IER3 modula la expresión de interleucinas ante la presencia de material vírico, por lo que se cree probable, posee un papel clave en la infección que nos atañe.

Estos son algunos de los genes resultantes del análisis realizado en los que se estima puedan tener una implicación en el desarrollo clínico de la COVID-19. Dichos resultados aportaron una considerable cantidad de genes significativos (2042), por lo que su estudio y evaluación es una actividad que sobrepasa los objetivos de este TFM.

6. Conclusiones

6.1. Conclusiones

Tras la realización del análisis completo de la secuenciación masiva realizada, se ha alcanzado las siguientes conclusiones:

- 2402 genes se han expresado diferencialmente de forma significativa, teniendo 1996 sobre-expresión y los 406 restantes infra-expresión.

- El enriquecimiento de los genes con expresión diferencial, muestra como gran parte de ellos están implicados en la regulación de la respuesta inmune, especialmente aquella relacionada con la actividad neutrofílica.

- Las secuencias analizadas muestran la relación entre diversos genes y el desarrollo de la enfermedad, siendo aquellos con evidencia bibliográfica previa: PGLYRP1, HDAC9 y FUT4.

- Se ha observado la presencia de otros genes envueltos en la actividad inmunológica, con expresión diferencial significativa y no descritos previamente en la bibliografía encontrada: ABCF1, ABHD16A, IER3, etc.

Dichas conclusiones concuerdan con los objetivos del TFM establecidos previamente, con lo que se consideran objetivos alcanzados.

6.2. Líneas de futuro

Como se ha anotado en apartados anteriores, existen 2042 genes con expresión diferencial significativa en los datos de origen. Se han evaluado de forma teórica aquellos con referencia bibliográfica y/o con implicación en procesos inmunes y/o inflamatorios.

Dicha evaluación se ha realizado teniendo en cuenta la significación aportada por el p-valor ajustado y su log2FoldChange, con lo que se estima que son los más relevantes.

Sin embargo, como línea de investigación futura se prevé práctico e informativo, realizar una revisión y análisis completo de aquellos genes con mayores índices de significación.

Además, es importante reflejar como limitación principal de este TFM la carencia en potencia computacional. Por ello, podría ser interesante realizar estos análisis con el conjunto total de muestras disponibles y utilizando otras herramientas de alineado, como STAR.

6.3. Seguimiento de la planificación

Se evalúa a continuación aquellos objetivos previamente establecidos, analizando su cumplimiento en cada caso.

> *1. Procesar los archivos fastq procedentes de la secuenciación masiva*: El presente objetivo ha sido alcanzado con éxito, se han procesado todos los archivos fastq de interés a través de una serie de pasos (control de calidad, trimado, “cuasi-mapeo”, etc)
>
> *2. Llevar a cabo un análisis de expresión diferencial y posterior enriquecimiento funcional*: Objetivo llevado a cabo con éxito mediante el uso de R/Bioconductor para realizar el análisis DE y su posterior enriquecimiento

3. Consolidar conocimientos en el uso del lenguaje de programación R para el análisis de datos mediante RNA-seq: Objetivo alcanzado con la lectura bibliográfica (artículos, manuales, scripts, etc) y la puesta en práctica, analizando cada una de las posibles opciones para optar por la que mejor se ajustase a la intención del trabajo y a las capacidades computacionales presentes.

Además, se evidencia cómo se ha realizado el total de las tareas previstas dentro de los tiempos establecidos. Ha de tenerse en cuenta algunas desviaciones ocurridas en diversas tareas:

- La obtención de la matriz de recuento supuso una importante desviación en la planificación, pues en principio se realizó conforme a ésta, siendo evidenciada posteriormente la necesidad de repetir el procesado a través de *Salmon*, debido a la baja calidad del total de los archivos pero, sobre todo, a la presencia de tres archivos con valores fuera de lo normal. Estos *fastq* fueron eliminados y sustituidos por otros, lo que supuso la repetición del “cuasi-mapeo”

- Otra importante desviación en la planificación de las tareas supuso el interés original de utilizar el paquete de R *tximeta* en lugar de *tximport*, debido a las mejoras del primero. Un desconocido error en la aplicación del paquete de interés, supuso una importante pérdida de tiempo en la búsqueda de solución, que finalmente fue descartado y sustituido por el paquete previamente conocido y utilizado *tximport*

7. Glosario

ADN: Ácido desoxi-ribonucleico

ARN: Ácido ribonucleico

CCDC: Centro Chino para el Control y Prevención de Enfermedades

GO: Gene Ontology.

HTA: Hipertensión Arterial

KEGG: Kyoto Encyclopedia of Genes and Genomes

OMS: Organización Mundial de la Salud

PCA: Análisis de componentes principales

PCR: Reacción en Cadena de la Polimerasa (Polymerase Chain Reaction)

8. Bibliografía

1. Baj J, Karakuła-Juchnowicz H, Teresiński G, Buszewicz G, Ciesielka M, Sitarz E, et al. COVID-19: Specific and Non-Specific Clinical Manifestations and Symptoms: The Current State of Knowledge. J Clin Med (2020) 5 (9):1753. doi: 10.3390/jcm9061753

2. World Health Organization Clinical management of COVID-19. Interim guidance. Disponible en https://www.who.int/publications/i/item/clinical#management-of-covid-19 (Accessed on Aug 6, 2020).

3. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013) An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE 8(12): e85024. https://doi.org/10.1371/journal.pone.0085024

4. Wilk, A. J., Rustagi, A., Zhao, N. Q., Roque, J., Martínez-Colón, G. J., McKechnie, J. L., Ivison, G. T., Ranganath, T., Vergara, R., Hollis, T., Simpson, L. J., Grant, P., Subramanian, A., Rogers, A. J., & Blish, C. A. (2020). A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nature Medicine, 26(7), 1070–1076. https://doi.org/10.1038/s41591-020-0944-y

5. Fricke-Galindo, I., & Falfán-Valencia, R. (2021). Genetics Insight for COVID-19 Susceptibility and Severity: A Review. Frontiers in Immunology, 12(April), 1–11. https://doi.org/10.3389/fimmu.2021.622176

6. Bost, P., Giladi, A., Liu, Y., Bendjelal, Y., Xu, G., David, E., Blecher-Gonen, R., Cohen, M., Medaglia, C., Li, H., Deczkowska, A., Zhang, S., Schwikowski, B., Zhang, Z., & Amit, I. (2020). Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients. Cell, 181(7), 1475-1488.e12. https://doi.org/10.1016/j.cell.2020.05.006

7. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

8. Delhomme N, Padioleau I, Furlong EE, Steinmetz LM: easyRNASeq: a Bioconductor package for processing RNA-seq data . Bioinformatics. 2012, 28: 2532-2533. 10.1093/bioinformatics/bts477.

9. Charlotte Soneson, Michael I. Love, Mark D. Robinson (2015): Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research http://dx.doi.org/10.12688/f1000research.7563.1

10. Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2016). Trimming of sequence reads alters RNA- Seq gene expression estimates. BMC Bioinformatics, 1–13. http://doi.org/10.1186/s12859-016-0956-2

11. MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. Frontiers in Genetics, 5, 13. http://doi.org/10.3389/fgene.2014.00013

12. Durbin, B. P., Hardin, J. S., Hawkins, D. M., & Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics, 18(SUPPL. 1), 105–110. https://doi.org/10.1093/bioinformatics/18.suppl_1.S105

13. Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065). https://doi.org/10.1098/rsta.2015.0202

14. John Hopkins University (Diciembre 2021). CORONAVIRUS RESOURCE CENTER. https://coronavirus.jhu.edu/data/disparity-explorer

15. Organización Mundial de la Salud (Diciembre 2021): Nuevo coronavirus 2019. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019

16. Jiang, F., Deng, L., Zhang, L., Cai, Y., Cheung, C. W., & Xia, Z. (2020). Review of the Clinical Characteristics of Coronavirus Disease 2019 (COVID-19). Journal of General Internal Medicine, 35(5), 1545–1549. https://doi.org/10.1007/s11606-020-05762-w

17. Moraleda, L., Escosa, T., Sainz, D., Aguilera, L., Espinosa, M., Barrio, I., & María, J. (2020). Manejo clínico del COVID-19: atención hospitalaria. Ministerio de Sanidad, Giobierno de España, 1–28. https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos/20200224.Preguntas_respuestas_COVID-19.pdf?utm_source=rss&utm_medium=rss

18. Liu, K., Fang, Y. Y., Deng, Y., Liu, W., Wang, M. F., Ma, J. P., Xiao, W., Wang, Y. N., Zhong, M. H., Li, C. H., Li, G. C., & Liu, H. G. (2020). Clinical characteristics of novel coronavirus cases in tertiary hospitals in Hubei Province. Chinese Medical Journal, 133(9), 1025–1031. https://doi.org/10.1097/CM9.0000000000000744

19. Vabret, N., Britton, G. J., Gruber, C., Hegde, S., Kim, J., Kuksin, M., Levantovsky, R., Malle, L., Moreira, A., Park, M. D., Pia, L., Risson, E., Saffern, M., Salomé, B., Esai Selvan, M., Spindler, M. P., Tan, J., van der Heide, V., Gregory, J. K., … Laserson, U. (2020). Immunology of COVID-19: Current State of the Science. Immunity, 52(6), 910–941. https://doi.org/10.1016/j.immuni.2020.05.002

20. Shuibing Chen, Liuliu Yang, Benjamin Nilsson-Payant, Yuling Han, Fabrice Jaffré, Jiajun Zhu, Pengfei Wang, Tuo Zhang, David Redmond, Sean Houghton, Rasmus Møller, Daisy Hoagland, Shu Horiuchi, Joshua Acklin, Jean Lim, Yaron Bram, Chanel Richardson, Vasuretha Chandar, Alain Borczuk, Yaoxing Huang, Jenny Xiang, David Ho, Robert Schwartz, Benjamin tenOever, Todd Evans. (2020). SARS-CoV-2 Infected Cardiomyocytes Recruit Monocytes by Secreting CCL2. Doi:10.21203/rs.3.rs-94634/v1

21. Wang, Z., Gerstein, M., & Snyder, M. (2010). Nihms229948. 10(1), 57–63. https://doi.org/10.1038/nrg2484.RNA-Seq

22. Gupta, A. K., & Gupta, U. D. (2013). Next Generation Sequencing and Its Applications. In Animal Biotechnology: Models in Discovery and Translation. Elsevier. https://doi.org/10.1016/B978-0-12-416002-6.00019-5

23. LaRossa, R. A. (2013). Transcriptome. Brenner's Encyclopedia of Genetics: Second Edition, 101–103. https://doi.org/10.1016/B978-0-12-374984-0.01553-9

24. Srivastava, A., George, J., & Karuturi, R. K. M. (2018). Transcriptome analysis. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 1–3, 792–805. https://doi.org/10.1016/B978-0-12-809633-8.20161-1

25. Zhang, J. Y., Wang, X. M., Xing, X., Xu, Z., Zhang, C., Song, J. W., Fan, X., Xia, P., Fu, J. L., Wang, S. Y., Xu, R. N., Dai, X. P., Shi, L., Huang, L., Jiang, T. J., Shi, M., Zhang, Y., Zumla, A., Maeurer, M., … Wang, F. S. (2020). Single-cell landscape of immunological responses in patients with COVID-19. Nature Immunology, 21(9), 1107–1118. https://doi.org/10.1038/s41590-020-0762-x

26. Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., De Domenico, E., Wendisch, D., Grasshoff, M., Kapellos, T. S., Beckstette, M., Pecht, T., Saglam, A., Dietrich, O., Mei, H. E., … Ziebuhr, J. (2020). Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. Cell, 182(6), 1419-1440.e23. https://doi.org/10.1016/j.cell.2020.08.001

27. Andrews, S. (2010). FastQC 1 . 1 What is FastQC 2 . Basic Operations 2 . 1 Opening a Sequence file. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

28. Objectives, L. (n.d.). MultiQC - fastQC summary tool -- GVA2020 Installing multiqc Checking installation Generating FastQC commands. 3–6.

29. Sato, S. A., & Xiao-Min Tong. (2018). SALMON document.

30. Mikeliunas, A. S. (2010). COMANDOS. 1–61.

31. Cuaresma, S. B. (2005). Manual básico Ubuntu GNU/Linux. Manual Básico Ubuntu GNU/Linux, 7. http://www.uls.edu.sv/pdf/ubuntu.pdf

32. Love, M., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. In bioRxiv. https://doi.org/10.1101/002832

33. Michael, A., Ahlmann-eltze, C., Forbes, K., & Anders, S. (2021). Package ‘ DESeq2 .’

34. Klaus, B. (2014). Differential expression analysis of RNA–Seq data using DESeq2. 1–24.

35. Singh, R. R. (2017). Next generation sequencing technologies. Comprehensive Medicinal Chemistry III, 2–8, 354–361. https://doi.org/10.1016/B978-0-12-409547-2.12327-3

36. Almeida, N. (2016). A short introduction to counselling. British Journal of Guidance & Counselling, 44(3), 365–367. https://doi.org/10.1080/03069885.2016.1176123

37.Folk, M. (2005). Data Formats. Hydroinformatics, 117–134. https://doi.org/10.1201/9781420038002.ch8

38. Seaby, E. G., Pengelly, R. J., & Ennis, S. (2016). Exome sequencing explained: A practical guide to its clinical application. Briefings in Functional Genomics, 15(5), 374–384. https://doi.org/10.1093/bfgp/elv054

39. Sánchez, A. (n.d.). Omics : The -not so new- technologies Pre-genomics vision.

40. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology, 17(1), 1–19. https://doi.org/10.1186/s13059-016-0881-8

41. Chen, M., & Dall, G. (2021). Package ‘ clusterProfiler .’

42. Rodríguez-Alonso, G., & Shishkova, S. (2018). Estudio del transcriptoma mediante rna-seq con énfasis en las especies vegetales no modelo. Revista de Educación Bioquímica, 3(37), 75-88. http://www.medigraphic.com/pdfs/revedubio/reb-2018/reb183c.pdf

43. Sánchez Santana, S. del C. (2015). Análisis de datos de RNA-Seq: comparación de métodos para el estudio de expresión génica diferencial. 63. https://idus.us.es/xmlui/handle/11441/40809

44. Project, B., Library, N., & Information, B. (2010). Bioconductor : Annotation Package Overview. October, 2002, 1–6.

45. Package, T., Disease, T., Semantic, O., & Annotationdbi, I. (2021). Package ‘ DOSE .’

46. Visualization, T., Enrichment, F., Version, R., The, D., Vignettebuilder, E., Artistic-, L., Annotation, V., Utf-, V. E., Yu, A. G., Hu, E., & Yu, M. G. (2021). Package ‘ enrichplot .’

47. April, G. (2006). Basic GO Usage. Nature Genetics, 1–7.

48. Package, T., & Luo, A. W. (2021). Package ‘ pathview .’

49. Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

50. Tatiana N. Sharapova, Elena A. Romanova, Aleksandr S. Chernov, Alexey N. Minakov, Vitaly A. Kazakov, Anna A. Kudriaeva, Alexey A. Belogurov, Jr., Olga K. Ivanova, Alexander G. Gabibov, Georgii B. Telegin, Denis V. Yashin, and Lidia P. Sashchenko. (2021). *Protein PGLYRP1/Tag7 Peptides Decrease the Proinflammatory Response in Human Blood Cells and Mouse Model of Diffuse Alveolar Damage of Lung through Blockage of the TREM-1 and TNFR1 Receptors*. doi: 10.3390/ijms222011213

51. Selvaraj G, Kaliamurthi S, Peslherbe GH and Wei DQ. Identifying potential drug targets and candidate drugs for COVID-19: biological networks and structural modeling approaches [version 3; peer review: 3 approved]. F1000Research 2021, 10:127 (https://doi.org/10.12688/f1000research.50850.3)NOTE: it is important to ensure the information in square b

52. Galimberti, S., Petrini, M., Baratè, C., Ricci, F., Balducci, S., Grassi, S., Guerrini, F., Ciabatti, E., Mechelli, S., Di Paolo, A., Baldini, C., Baglietto, L., Macera, L., Spezia, P. G., & Maggi, F. (2020). Tyrosine Kinase Inhibitors Play an Antiviral Action in Patients Affected by Chronic Myeloid Leukemia: A Possible Model Supporting Their Use in the Fight Against SARS-CoV-2. Frontiers in Oncology, 10(September), 1–9. https://doi.org/10.3389/fonc.2020.01428

53. Cao, Q. T., Aguiar, J. A., Tremblay, B. J. M., Abbas, N., Tiessen, N., Revill, S., Makhdami, N., Ayoub, A., Cox, G., Ask, K., Doxey, A. C., & Hirota, J. A. (2020). ABCF1 Regulates dsDNA-induced Immune Responses in Human Airway Epithelial Cells. Frontiers in Cellular and Infection Microbiology, 10(September), 1–17. https://doi.org/10.3389/fcimb.2020.00487

54. Xu J, Gu W, Ji K, Xu Z, Zhu H, Zheng W. 2018 Sequence analysis and structure prediction of ABHD16A and the roles of the ABHD family members in human disease. Open Biol. 8: 180017. http://dx.doi.org/10.1098/rsob.180017

55. Villalba M, Fredericksen F, Otth C, Olavarría VH. Molecular characterization of the bovine IER3 gene: Down-regulation of IL-8 by blocking NF-κB activity mediated by IER3 overexpression in MDBK cells infected with bovine viral diarrhea virus-1. Mol Immunol. 2017 Dec;92:169-179. doi: 10.1016/j.molimm.2017.10.012. Epub 2017 Nov 2. PMID: 29101849.

9. Anexos

Anexo 1. Extracción ARN total de muestras de sangre

TUBOS -TEMPUS™ BLOOD RNA TUBE *(cat. # 4342792)*

MATERIALES Y REACTIVOS

- Tubos falcon 15 ml estériles
- Tubos eppendorf 2 ml (RNAasa free)
- Tubos eppendorf 1,5 ml (RNAasa free)
- Centrifuga con control de temperatura y rotor con cestillos para tubos falcom 15 ml
- Centrifuga con control de temperatura y rotor para tubos eppendorf 1,5/2 ml
- Hielo a demanda
- PBS 1X
- TRIzol™ Reagent (*cat. #15596026*)
- Cloroformo
- Isopropanol (2-Propanol) ≥99.8% para biología molecular
- Glicogeno
- Etanol para biología molecular preparado al 70%

PROTOCOLO

1. Descongelamos en hielo la muestra de sangre almacenada a -80^{o}C .

Desde este momento trabajaremos en hielo y con centrífugas a 4^{o}C

2. Vertemos su contenido total (9 ml) a un tubo falcon de 15 ml.

3. Añadimos 3 ml de PBS a cada tubo falcon.

4. Centrifugamos a 4^{o}C y 3000 rpm durante 30 min.

5. Retornamos las muestras al hielo asegurándonos de que se mantienen de forma vertical durante 10 min.

6. Decantamos el contenido del falcon "sin forzar" y nos quedamos con el volumen del fondo.

--> El color del contenido es marrón oscuro y no veremos pellet

7. Añadimos un 1 ml de TRIzol y lo re-suspendemos suavemente con la pipeta hasta que el color rosado del TRIzol desaparezca.

8. Pasamos la mezcla a un eppendorf de 2 ml y añadimos 200 µl de Cloroformo.

9. Agitamos fuertemente la mezcla hasta torne a un color marrón mas claro y centrifugamos a 4oC y 12.000 rpm durante 15 min.

--> Mientras se centrifuga preparamos 500 µl de Isopropanol en un eppendorf de 1,5 ml.

10. De los tubos centrifugados con tres fases bien diferenciadas, recuperamos la fase superior acuosa (500 *µl* aprox.) que contendrá en RNA total y la llevamos al tubo eppendorf de 1,5 con el Isopropanol añadido previamente y lo agitamos bien.

--> Por si nos interesara, la capa blanca intermedia sería el DNA.

--> En este paso podríamos detener la extracción y almacenar las muestras a -20ºC.

11. A la solución de RNA con Isopropanol, le añadimos 1 *µl* de Glicogeno e invertiremos un par de veces el tubo para que este se disuelva adecuadamente.

12. Centrifugamos las muestras a 4^{0}C y 12.000 rpm durante 10 min.

13. Retiramos todo el volumen del tubo y nos quedamos solamente con el pellet.

--> Hemos de tener en cuenta que los pellets pueden ser móviles.

14. Echamos 800 *µl* de Etanol al 70% y añadimos 1 *µl* mas de Glicógeno. De nuevo invertiremos un par de veces el tubo para que este se disuelva adecuadamente.

15. Centrifugamos a 4^{0}C y 12.000 rpm durante 5 min.

16. Retiramos muy bien todo el sobrenadante y nos quedamos con el pellet lo mas seco posible.

--> Hemos de tener en cuenta que los pellets pueden ser móviles.

17. Con la tapa abierta de los tubos dejamos que los restos de etanol se evaporen a temperatura ambiente y una vez que el pellet esta seco (transparente), le añadimos 10* *µl* de agua destilada estéril y le damos un *spin* a 4^{0}C.

--> Este volumen podrá ser superior si el pellet lo requiriera.

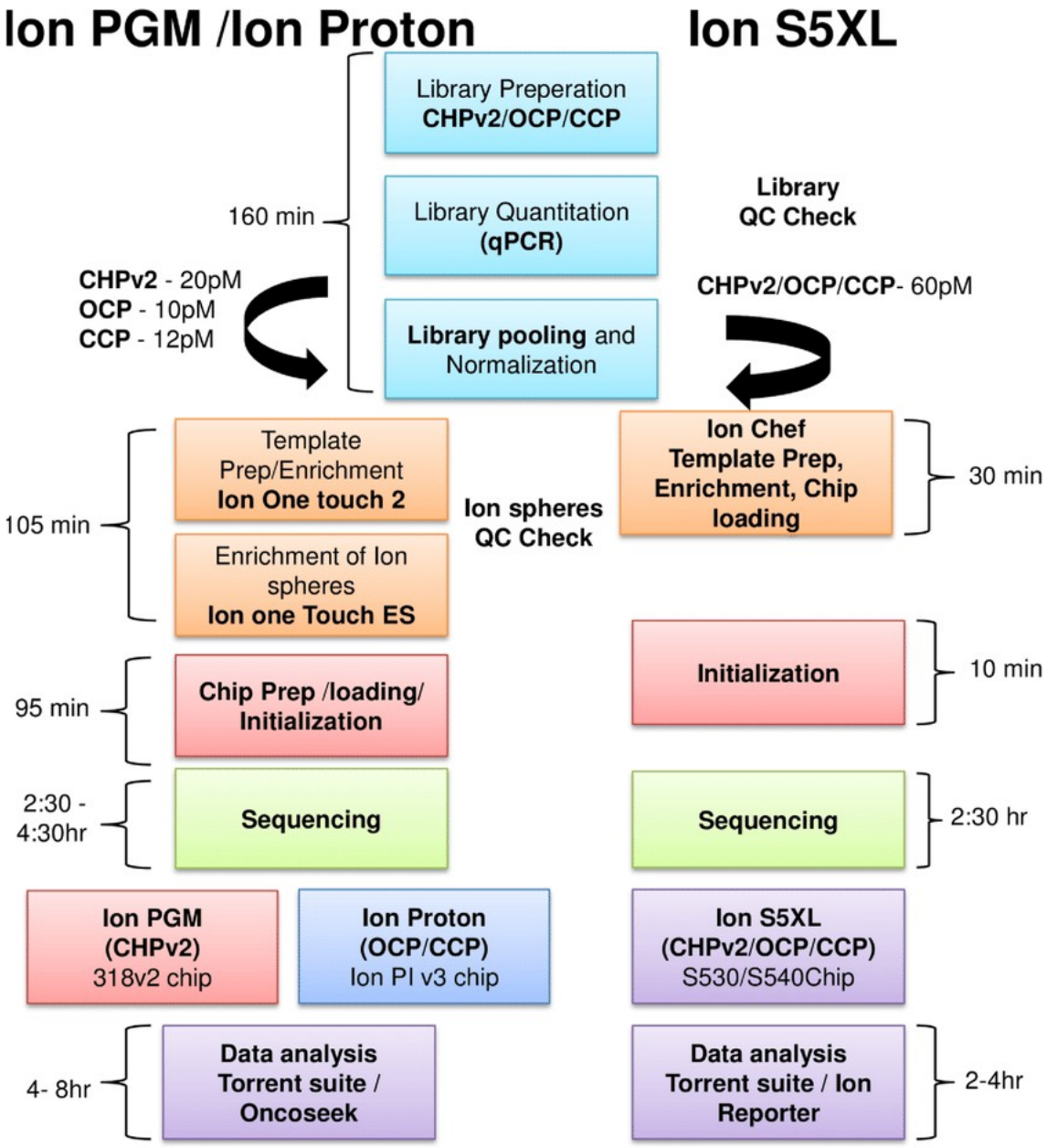
--> El RNA se mantendrá en hielo si va a ser posteriormente utilizado:

- *Cuantificación (NanoDrop, TapeStation, QUBIT...)*
- *Retro-Transcripción*
- *Creación de Librerías para RNAseq*

--> Si su uso no va a ser inmediato, conviene almacenarlo a -80 ºC.

--> Se recomienda hacer alícuotas para evitar descongelaciones y recongelaciones del RNA.

Anexo 2. Flujo de trabajo Ion Torrent



Nota. Versatile ion S5XL sequencer for targeted next generation sequencing of solid tumors in a clinical laboratory - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/NGS-workflow-used-for-different-Ion-Torrent-platforms-in-a-clinical-laboratory-Ion-PGM_fig1_318869700 [accessed 10 Dec, 2021]

Anexo 3. Procesado de datos: Códigos

Preparación del entorno Linux (Oracle VM VirtualBox)

```
mkdir -p ~/workspace/rnaseq/
export RNA_HOME=~/workspace/rnaseq
echo $RNA_HOME

export RNA_HOME=~/workspace/rnaseq
export RNA_EXT_DATA_DIR=/home/israel/RNA_data
export RNA_DATA_DIR=$RNA_HOME/data
export RNA_DATA_TRIM_DIR=$RNA_DATA_DIR/trimmed
export RNA_REFS_DIR=$RNA_HOME/refs
export RNA_REF_INDEX=$RNA_REFS_DIR/Homo_sapiens.GRCh38.dna.alt
export RNA_REF_FASTA=$RNA_REF_INDEX.fa
export RNA_REF_GTF=$RNA_REF_INDEX.gtf

cd $RNA_HOME
mkdir student_tools
cd student_tools
```

Archivos de referencia

```
cd $RNA_HOME
echo $RNA_REFS_DIR
mkdir -p $RNA_REFS_DIR
wget
http://ftp.ensembl.org/pub/release-104/gtf/homo_sapiens/Homo_sapiens.GRCh38.104.gtf.gz
gzip -d Homo_sapiens.GRCh38.104.gtf.gz
```

Control de calidad

FastQC

```
Instalación:
cd $RNA_HOME/student_tools/
wget  https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.8.zip  --
no-check-certificate
unzip fastqc_v0.11.8.zip
cd FastQC/
chmod 755 fastqc
./fastqc -help

Ejecución:
cd $RNA_HOME/data
fastqc *.fastq.gz
```

MultiQC

```
Instalación:
sudo apt install python3-pip
pip3 install multiqc
multiqc -help

Ejecución:
cd $RNA_HOME/data
multiqc .
```

Trimado

```
Instalación:
https://sourceforge.net/projects/bbmap/

Ejecución 1° vuelta:
export PATH=/home/israel/workspace/rnaseq/student_tools/bbmap:$PATH
bbduk.sh -Xmx1g in=reads.fq out=clean.fq qtrim=rl trimq=10 ftl=20 ftr=300 maq=20

Ejecución 2° vuelta:
dedupe.sh -Xmx2g in=<file or stdin> out=<file or stdout>
```

Cuasi-Mapeo (Salmon)

Creación del índice:

```
salmon index \
     -t $RNA_REFS_DIR/Homo_sapiens.GRCh38.cdna.all.fa \
     -i salmon_index \
     -k 31
```

Ejecución:

```
salmon quant -i /home/israel/workspace/rnaseq/refs/salmon_index \
 -l A \
 -r /home/israel/workspace/rnaseq/data/entrada.fastq \
 -o salida.subset.salmon \
 --seqBias \
 --validateMappings \
 --useVBOpt
```

Anexo 4. Análisis de expresión diferencial

```
BiocManager::install("ellipsis")
BiocManager::install("readODS")
BiocManager::install("ensembldb")
BiocManager::install("GenomicFeatures")
BiocManager::install("AnnotationHub")
BiocManager::install("tximport")
BiocManager::install("tximportData")
BiocManager::install("tximeta")
BiocManager::install("DESeq2")
BiocManager::install("pheatmap")
BiocManager::install("tidyverse")
BiocManager::install("EnsDb.Hsapiens.v86")
BiocManager::install("RColorBrewer")
BiocManager::install("ggnewscale")

library(tximeta)
library(DESeq2)
library(pheatmap)
library(tidyverse)
library(readODS)
library(GenomicFeatures)
library(ensembldb)
library(xlsx)
library(rjson)
library(readr)
library(RColorBrewer)
library(tibble)



path<-"C:/Users/israe/Desktop/Master_Bioinformatica_y_Bioestadistica/10-TFM/00-
SCRIPTS_Y_CODIGOS/quants"
files <- list.files(path=path, pattern="quant.sf", full.names = TRUE, recursive = TRUE)
covid <- file.path("C:/Users/israe/Desktop/Master_Bioinformatica_y_Bioestadistica/10-
TFM/00-SCRIPTS_Y_CODIGOS/quants/metadata.csv")
coldata <- read.csv(covid, row.names=c("CV-1","CV-2","CV-3","CV-4","CV-5","CV-6","CV-
7","CV-8","CV-9","CV-10","UCI-1","UCI-2","UCI-3","UCI-4","UCI-5","UCI-6","UCI-7","UCI-
8","UCI-9","UCI-10"), stringsAsFactors=FALSE)
coldata$files <- file.path(files, coldata$names, "quant.sf")
all(rownames(coldata) == colnames(files)) #TRUE: Confirmamos la igualdad en el orden de
                                          las variables


library(EnsDb.Hsapiens.v86)
edb <- EnsDb.Hsapiens.v86
txs<-transcripts(edb,return.type="data.frame")
tx2gene<-txs[,c("tx_id", "gene_id")]
library(tximport)
txi<-tximport(files,type = "salmon", tx2gene = tx2gene, ignoreTxVersion = T)
dds <- DESeqDataSetFromTximport(txi, colData=coldata, design= ~condition)


dds <- estimateSizeFactors(dds) # Calculamos los factores de normalización
normalized <- counts(dds, normalized=T)
head(normalized)

vsd <- vst(dds, blind = T)
vsd_mat <- assay(vsd)
vsd_cor <- cor(vsd_mat)

plot(hclust(dist(t(normalized))), labels=colData(dds)$condition)
pheatmap(vsd_cor, annotation = dplyr::select(coldata, condition))
plotPCA(vsd, intgroup = "condition")


# ANÁLISIS DE EXPRESIÓN DIFERENCIAL
dds <- DESeq(dds)

mean_counts<-apply(normalized[,1:20],1,mean)
variance_counts<-apply(normalized[,1:20],1,var)
df←data.frame(mean_couns,variance_counts)


#Gráficado
```

```
ggplot(df)+
geom_point(aes(x=mean_couns, y=variance_counts))+
scale_y_log10()+
scale_x_log10()+
xlab("Means per gene")+
ylab("Variance per gene"

plotDispEsts(dds)

#Resultados

res<-results(dds,alpha = 0.05)

plotMA(res, ylim=c(-7,7) )
mcols(res)
head(res,n=5)
summary(res) # Genes diferencialmente expresados significativamente - Resumen
sum(res$padj < 0.05, na.rm=TRUE)# 2402


#Procesado y visualización de resultados

resSort <- res[order(res$pvalue),]
head(resSort)

library("AnnotationDbi")
library("org.Hs.eg.db")
columns(org.Hs.eg.db)

ens.str <- substr(rownames(resSort), 1, 15)
resSort$symbol <- mapIds(org.Hs.eg.db,
                         keys=ens.str,
                         column="SYMBOL",
                         keytype="ENSEMBL",
                         multiVals="first")
resSort$entrez <- mapIds(org.Hs.eg.db,
                         keys=ens.str,
                         column="ENTREZID",
                         keytype="ENSEMBL",
                         multiVals="first")
resSort$genname <- mapIds(org.Hs.eg.db,
                         keys=ens.str,
                         column="GENENAME",
                         keytype="ENSEMBL",
                         multiVals="first")
resSort$ensembl <- mapIds(org.Hs.eg.db,
                         keys=ens.str,
                         column="ENSEMBL",
                         keytype="ENSEMBL",
                         multiVals="first")

head(resSort)


res_all<-data.frame(resSort) %>%
rownames_to_column(var = "ensgene") %>%
mutate(threshold = padj<0.05)

ggplot(res_all,aes(x=log2FoldChange,y=-log10(padj),color=diffexpressed))+
geom_point()+
xlab("log2 fold change")+
ylab("-log10 adjusted p-value")+
theme(legend.position = "none",
plot.title = element_text(size = rel(1.5),hjust = 0.5),
axis.title = element_text(size = rel(1.5)))

resultados_significativos<-data.frame(resSort)[1:10, ]
conteos_significativos<-data.frame(gen_info2)[1:10, ]


top20<-data.frame(gen_info)[1:20, ] %>%
  rownames_to_column(var = "ensgene")
names(top20)
top20<-rename(top20,         "CV-1"="CV.1","CV-2"="CV.2","CV-3"="CV.3","CV-4"="CV.4","CV-
5"="CV.5",  "CV-6"="CV.6","CV-7"="CV.7","CV-8"="CV.8","CV-9"="CV.9","CV-10"="CV.10","UCI-
```

```
1"="UCI.1","UCI-2"         ="UCI.2","UCI-3"="UCI.3","UCI-4"="UCI.4","UCI-5"="UCI.5","UCI-
6"="UCI.6","UCI-7"="UCI.7","UCI-8"="UCI.8","UCI-9"="UCI.9","UCI-10"="UCI.10")

top20<-gather(top20, key = "samplename", value = "normalized_counts",2:21)

coldat2<-coldata
coldat2<-rename(coldat2,samplename=names)

top20<-inner_join(top20,
                  rownames_to_column(coldata, var = "samplename"),
                  by="samplename")

ggplot(top20)+
  geom_point(aes(x=ensgene, y =normalized_counts, color= condition))+
  scale_y_log10()+
  xlab("Genes")+
  ylab("Normalized Counts")+
  ggtitle("Top 20 Significant DE Genes")+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  theme(plot.title = element_text(hjust = 0.5))


# ENRIQUECIMIENTO

library(org.Hs.eg.db)
library(DOSE)
library(pathview)
library(clusterProfiler)
library(AnnotationHub)
library(ensembldb)
library(tidyverse)
library(ggnewscale)
library(enrichplot)

OrgDb <- org.Hs.eg.db


geneList <- as.vector(resSort$log2FoldChange)
names(geneList) <- resSort$entrez
gene <- na.omit(resSort$entrez)


# Grupos GO

#BP
ggo.bp <- clusterProfiler::groupGO(gene     = gene,
                                   OrgDb    = OrgDb,
                                   ont      = "BP",
                                   level    = 3,
                                   readable = TRUE)


view(ggo.bp)
barplot(ggo.bp, drop=TRUE, showCategory=12)

#CC

ggo.cc<- clusterProfiler::groupGO(gene     = gene,
                                    OrgDb    = OrgDb,
                                    ont      = "CC",
                                    level    = 3,
                                    readable = TRUE)

view(ggo.cc)
barplot(ggo.cc, drop=TRUE, showCategory=12)


#MF

ggo.mf<- clusterProfiler::groupGO(gene     = gene,
                                    OrgDb    = OrgDb,
                                    ont      = "MF",
                                    level    = 3,
                                    readable = TRUE)
```

```
view(ggo.mf)
barplot(ggo.mf, drop=TRUE, showCategory=12)


# GO Test de Sobrerrepresentación

#BP
ego.bp <- clusterProfiler::enrichGO(gene          = gene,
                                    OrgDb         = OrgDb,
                                    ont           = "BP",
                                    pAdjustMethod = "BH",
                                    pvalueCutoff  = 0.05,
                                    qvalueCutoff  = 0.05,
                                    readable      = TRUE)

summary(ego.bp)
head(summary(ego.bp)[,-8])
view(ego.bp)


barplot(ego.bp, showCategory=12)
clusterProfiler::dotplot(ego.bp, showCategory=12)



#CC
ego.cc <- clusterProfiler::enrichGO(gene          = gene,
                                    OrgDb         = OrgDb,
                                    ont           = "CC",
                                    pAdjustMethod = "BH",
                                    pvalueCutoff  = 0.05,
                                    qvalueCutoff  = 0.05,
                                    readable      = TRUE)

summary(ego.cc)
head(summary(ego.cc)[,-8])
view(ego.cc)

barplot(ego.cc, showCategory=12)
clusterProfiler::dotplot(ego.cc, showCategory=12)


#MF
ego.mf <- clusterProfiler::enrichGO(gene          = gene,
                                    OrgDb         = OrgDb,
                                    ont           = "MF",
                                    pAdjustMethod = "BH",
                                    pvalueCutoff  = 0.05,
                                    qvalueCutoff  = 0.05,
                                    readable      = TRUE)

summary(ego.mf)
head(summary(ego.mf)[,-8])
view(ego.mf)

barplot(ego.mf, showCategory=12)
clusterProfiler::dotplot(ego.mf, showCategory=12)

#Mapa de enriquecimiento

d <- GOSemSim::godata("org.Hs.eg.db", ont = "BP")
compare_cluster_GO_emap    <-    enrichplot::pairwise_termsim(ego.bp,    semData    =    d,
method="Wang")
emapplot(compare_cluster_GO_emap, showCategory = 15)

#GO-plot

goplot(ego.bp, showCategory = 4)


cnetplot(ego.bp, categorySize="pvalue", foldChange=geneList, showCategory = 5)


#KEGG

kk <- clusterProfiler::enrichKEGG(gene= gene,
                                  organism     = 'hsa',
```

```
                                    pAdjustMethod = "BH",
                                    pvalueCutoff = 0.05,
                                    qvalueCutoff  = 0.05)

barplot(kk, showCategory=12)
clusterProfiler::dotplot(kk, showCategory=9)


#Pathway

library(pathview)
pathview(gene.data = gene,
         pathway.id = "hsa05168",
         species = "hsa",
         limit = list(gene = 2,
                      cpd = 1))
```