Análisis del patrón de expresión de la tolerancia a endotoxinas en un modelo de monocitos humanos mediante RNA-seq

Nombre Estudiante:
Víctor Manuel Toledano Real
Máster en Bioinformática y Bioestadística

Área del trabajo final

Nombre Consultores
Dr. José Luis Villanueva Cañas
D. Enrique Vázquez de Luis
Genómica comparativa y evolución

Nombre Profesor responsable de la asignatura
Dr. Carles Ventura Royo

Fecha Entrega 02/01/2019

A) Creative Commons:

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis del patrón de expresión de la tolerancia a endotoxinas en un modelo de monocitos humanos mediante RNA-seq*
Nombre del autor:	*Víctor Manuel Toledano Real*
Nombre del consultor/a:	*José Luis Villanueva Cañas* *Enrique Vázquez de Luis*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	*01/2019*
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*área2: Genómica comparativa y evolución*
Idioma del trabajo:	*Castellano*
Palabras clave	*Sistema Inmune, Tolerancia a Endotoxinas, RNA-seq*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Este trabajo se centra en la respuesta del Sistema Inmune enfocándose en la tolerancia a endotoxinas. Esta se describe como la reprogramación del sistema inmune innato contra las infecciones causadas por patógenos y es uno de los mecanismos de protección más importantes descritos en la biomedicina.
En este estudio se desarrollaron tres modelos con monocitos/macrófagos humanos, los cuales simulan un donante sano, un paciente inflamado y el estado de tolerancia a endotoxinas. Para realizar un estudio exhaustivo diferencial de los tres estados utilizamos la técnica RNA-seq, ya que ésta cuantifica la presencia de ARN mensajero en una muestra biológica, y la utilizamos para detectar los cambios en el transcriptoma entre condiciones, obteniendo así características específicas y cambios genéticos en la tolerancia a endotoxinas.
Una vez desarrollado el análisis, utilizando diversas herramientas informáticas, podemos concluir que se comprueban los diferentes patrones de expresión de cada modelo, aclarando así la independencia de cada estado. Además obtenemos listados de genes de expresión diferencial y sus asociaciones a procesos y funciones biológicas, que pueden ser de gran interés en futuras investigaciones. A parte generamos un listado de inmunocheckpoints asociados al dominio v-set por indicación propia del grupo de investigación, dotándonos de posibles moléculas “diana” para futuros tratamientos.

Abstract (in English, 250 words or less):

The aims of this work is to evaluate the response of the Immune System focusing on the Endotoxin Tolerance. This is described as the reprogramming of the innate immune system against infections caused by pathogens and is one of the most important mechanisms of protection described in biomedicine.

In this study, three models were developed with human monocytes/macrophages. These models simulate a healthy donor, an inflamed patient and the state of endotoxin tolerance. To carry out a comprehensive differential study of the three groups, we used the RNA-seq technique, since it quantifies the presence of messenger RNA in a biological simple. Moreover we use it to detect changes in the transcriptome between these conditions, thus obtaining specific characteristics and genetic changes in endotoxin tolerance.

Once the analysis is developed using various computer tools, we can conclude that the different expression patterns of each model are checked. This fact allows clarifying the independence of each state and obtaining listings of differential expression genes and their associations to biological processes and functions that may be of great interest in future research. We also generated a list of immunocheckpoints associated with the v-set domain by the research group's own indication, donating possible "target" molecules for future treatments.

Índice

Lista de figuras.

Lista de tablas.

1. Introducción

1.1 Contexto y justificación del Trabajo

Introducción:

Este trabajo se centra en la respuesta del Sistema Inmune enfocándose en la Tolerancia a Endotoxinas (TE). esta se describe ampliamente como un fenómeno clínico[1]. Se trata de reprogramar el sistema inmune innato (SII) contra las infecciones causadas por patógenos y es uno de los mecanismos de protección más importantes descritos en biomedicina [1,2]. Durante la TE las células están en un estado latente que impide responder de manera convencional a la presencia de agentes patógenos o a sus derivados (Pathogens-associated molecularpatterns– PAMPs). Varios estudios confirman que ciertas células del SII, en particular los monocitos / macrófagos, entran en un estado de TE o reprogramación desarrollando una actividad alternativa caracterizada por: i) la regulación negativa de la respuesta inflamatoria, ii) una mayor capacidad de fagocitosis y iii) baja presentación antigénica[1,2,3,4]. Las características I y II pueden facilitar la resolución de enfermedades como la sepsis y pueden hacer que el paciente evolucione positivamente, evitando daños en los tejidos y eliminando de manera eficaz los patógenos [5]. Sin embargo, este estado puede comprometer la defensa contra infecciones secundarias, aunque cabe destacar que la TE no significa una inmunoparálisis del SII [1,2,4].

Este fenómeno está bien caracterizado en modelos experimentales; primero con un tratamiento previo con la endotoxina lipopolisacárido (LPS, este es un componente mayoritario de la membrana externa de las bacterias Gram negativas) reprogramando la respuesta del SII y posteriormente con un segundo contacto con LPS [3], este caso específico se conoce como homotolerancia.

Todos los estudios de homotolerancia en humanos están realizados en modelos *ex vivo* e i*n vitro* y con aislamiento de células mononucleares de sangre periférica (PBMCs) [2], donde se observa la disminución de expresión de HLA-DR en estos pacientes, también exhiben una capacidad disminuida de liberación de las citoquinas proinflamatorias TNF, IL-1, IL-6 e IL-12 después de un estímulo con LPS, lo que indica que la señalización intracelular modifica la producción de los mediadores antiinflamatorios asociados, también existe un aumento de citoquinas antiinflamatorias como IL10. La importancia de esta inmunomodulación es la reversibilidad del estado de TE, debido a esta afirmación, diferentes grupos de investigación sugieren que si determinadas moléculas/compuestos pueden facilitar la transición a un estado de TE controlado [1], se puede favorecer la evolución con un mejor pronóstico para los pacientes sépticos [6] y en el futuro se podrían desarrollar terapias induciendo la TE a estos pacientes pudiendo mejorar su tratamiento [7].

Para poder realizar un estudio exhaustivo diferencial de este estado utilizamos una técnica denominada RNA sequencing (RNA-seq) [8,9], una técnica de secuenciación de nueva generación (NGS) que cuantifica la presencia de ARN mensajero en una muestra biológica. Con ella podemos detectar detectar los cambios en el transcriptoma entre condiciones, realizar un estudio detallado de sus características específicas y cambios genéticos en la TE [10].

Para explicar la técnica de RNA-seq nos remitimos a la Figura 1, esta parte inicialmente de una fragmentación del ARN en secuencias cortas, llamadas lecturas, cada una de éstas es secuenciada y posteriormente se enfrenta a un genoma de referencia en nuestro caso de humano, obteniendo un alineamiento.

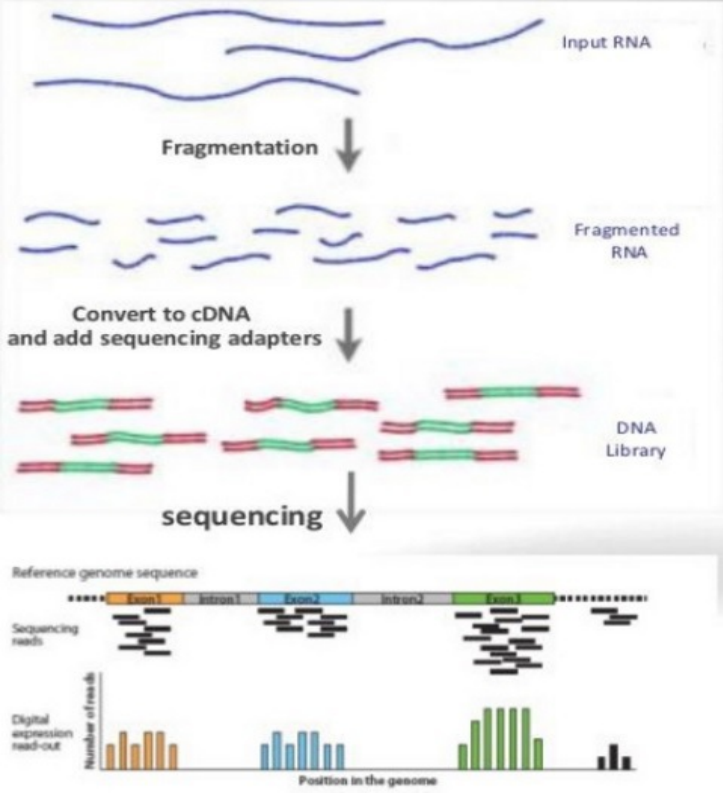


Figura1. Esquema de un experimento de RNA-seq [10]

Partiendo de estos alineamientos obtenemos la expresión génica, ya que cada gen tendrá una mayor o menor expresión dependiendo del número de lecturas que se hayan alineado frente a ese gen, y se podrá comparar entre diferentes condiciones experimentales.

El análisis de todo el procedimiento se realiza con herramientas informáticas utilizándolas acorde de nuestras muestras, equipos disponibles y conocimientos, finalizando este análisis de datos con una expresión diferencial de nuestros genes. En este caso el programa más utilizado es R con la librería de Bioconductor [8,10] y sus diferentes herramientas.

Se eligió esta área y tema por la disponibilidad de datos experimentales de investigación habiéndose generado en el grupo de investigación donde trabaja. Aplicando la tecnología de RNA-seq en la investigación de TE y respuesta inmune, tratando de avanzar con patrones genéticos en el diagnóstico y prognosis en las enfermedades que existe este estado de reprogramación del SII. También se pretenden buscar posibles inmunocheckpoints,es decir puntos de control inmunológico, que son reguladores del sistema inmune, esto nos podría determinar dianas terapéuticas frente a diferentes enfermedades donde existe esta TE.

1.2 Objetivos

1.Diseñar y analizar un experimento utilizando técnicas de NGS como el RNA-seq. En este primer objetivo general existen los siguientes específicos:

1.1.Desarrollar un pipeline adecuado
1.2.Realizar el control de calidad
1.3.Realizar el alineamiento y mapeo de los “reads”
1.4.Generar un listado de genes diferencialmente expresados

2. Determinar el patrón de expresión de la tolerancia a endotoxinas en un modelo con monocitos humanos. En este segundo objetivo general existen los siguientes específicos:

2.1. Comparar nuestro modelo de TE frente al control
2.2. Comparar nuestro modelo de Inflamación frente al TE
2.3. Obtener un patrón de expresión de TE definido

3. Buscar nuevos inmunocheckpoints mediante datos de RNA-seq con diferentes condiciones experimentales. En este tercer objetivo general existen los siguientes específicos:

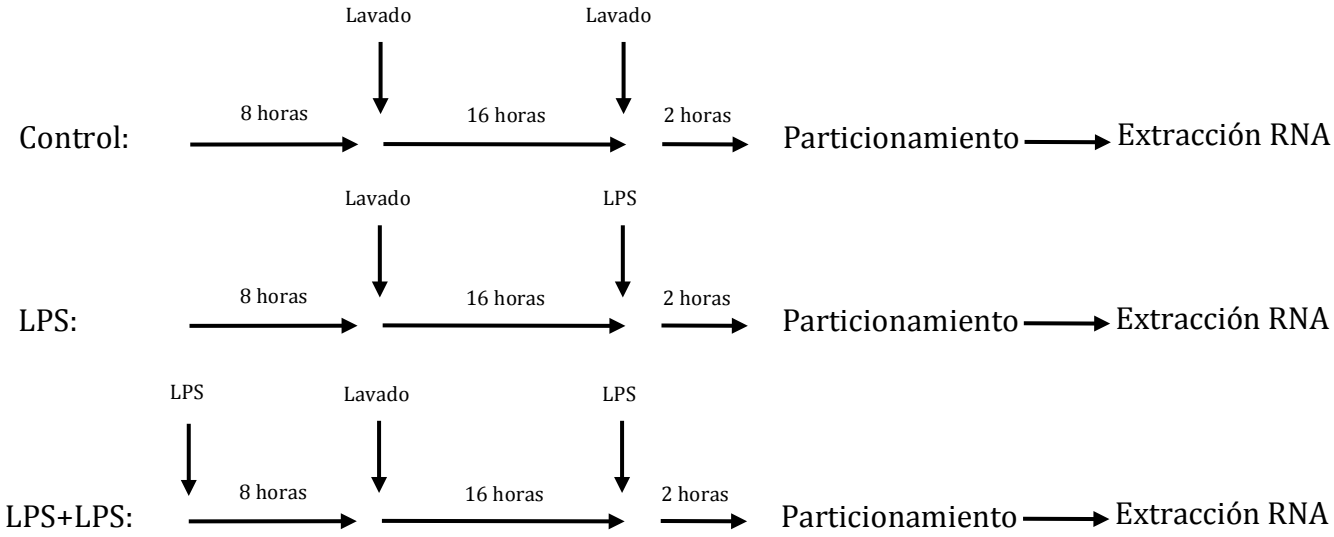
3.1. Asociar nuestros genes expresados diferencialmente con procesos biológicos
3.2. Buscar genes específicos en regiones de interés

1.3 Enfoque y método seguido

El desarrollo de este trabajo esta dividido en varias partes. Primero ha existido una parte técnica de laboratorio para obtener las muestras que posteriormente se secuenciaron. En esta parte inicial partimos de sangre de donantes sanos donde se realizaron diferentes modelos *in vitro* de TE, con ello podemos simular una infección con LPS. Estos modelos nos dan la ventaja de recrear situaciones que existen en determinadas enfermedades como la sepsis, debido a que no se dispone fácilmente de muestras de pacientes, pero con éstas aproximaciones se puede avanzar en las investigaciones de diferentes enfermedades.

En la parte técnica, lo primero que realizamos es una extracción de PBMCs con Ficoll (lo que conseguimos con ello es la separación de estas células por

gradiente de densidad), posteriormente por adherencia en las placas enriquecemos nuestro cultivo para que haya el mayor porcentaje de monocitos posibles, ya que la tolerancia a endotoxinas esta descrita en estas células [1]. Posteriormente se hacen varios modelos (Control, LPS, LPS+LPS=TE) se describe en el siguiente esquema:



De cada grupo se realizaron 3 réplicas biológicas, realizamos una extracción de ARN, en este caso hicimos un particionamiento por lisis secuencial por detergentes [11], con esto queremos conseguir el ARN asociado a polisomas de la membrana del retículo endoplásmico (ER). Este es el centro de la síntesis de proteínas de membrana y lípidos. Con esta extracción queremos centrarnos en los genes de membrana para buscar ahí nuestras diferencias y los inmunocheckpoint.

Finalmente, para resolver nuestras preguntas biológicas comparando entre distintos perfiles de expresión genética. Se opto por el método de secuenciación masiva RNA-seq debido a que permite analizar el transcriptoma, es decir, el conjunto de ARN mensajero transcrito.

1.4 Planificación del Trabajo

Para alcanzar los objetivos propuestos realizamos un plan de trabajo dividido en dos fases y tareas específicas descritas a continuación:

Tareas:

F1.-Fase 1:

T1. Instalación y prueba del software seleccionado para el análisis con el RNA-seq.

T2. Estudio y descripción de los tipos de muestras y el método utilizado de extracción de RNA para el posterior análisis del RNA-seq

T3. Desarrollo de la pipeline.

T4. Estudio y selección de métodos para el análisis del control de calidad de nuestras muestras de reads. Utilizando FASTQC y utilizando la depuración Cutadapt.
T5. Estudio y selección de métodos para el alineamiento y mapeo de los "reads". Utilizando Bowtie, TopHat2 o Kallisto.
T6. Estudio y selección de métodos para estimas de expresión génica. HTSeq, Cufflinks, Cutmerge, Cuffdiff.
T7. Entrega Fase 1.

F2.-Fase2

T1. Estudio de métodos para el inicio de análisis de expresión génica diferencial EdgeR, Cufflinks.
T2. Análisis de expresión diferencial, EdgeR.
T3. Asociación de nuestros genes expresados diferencialmente con procesos biológicos GOrilla y Uniprot.
T4. Análisis de significación entre nuestros grupos de muestras utilizando R.
T5. Entrega Fase 2.

Redacción de la Memoria.
Elaboración de la presentación.
Defensa pública.

Calendario:

A continuación mostramos una gráfica donde se define el calendario de tareas mediante un diagrama de Gantt, y podemos observar el tiempo transcurrido de cada tarea.

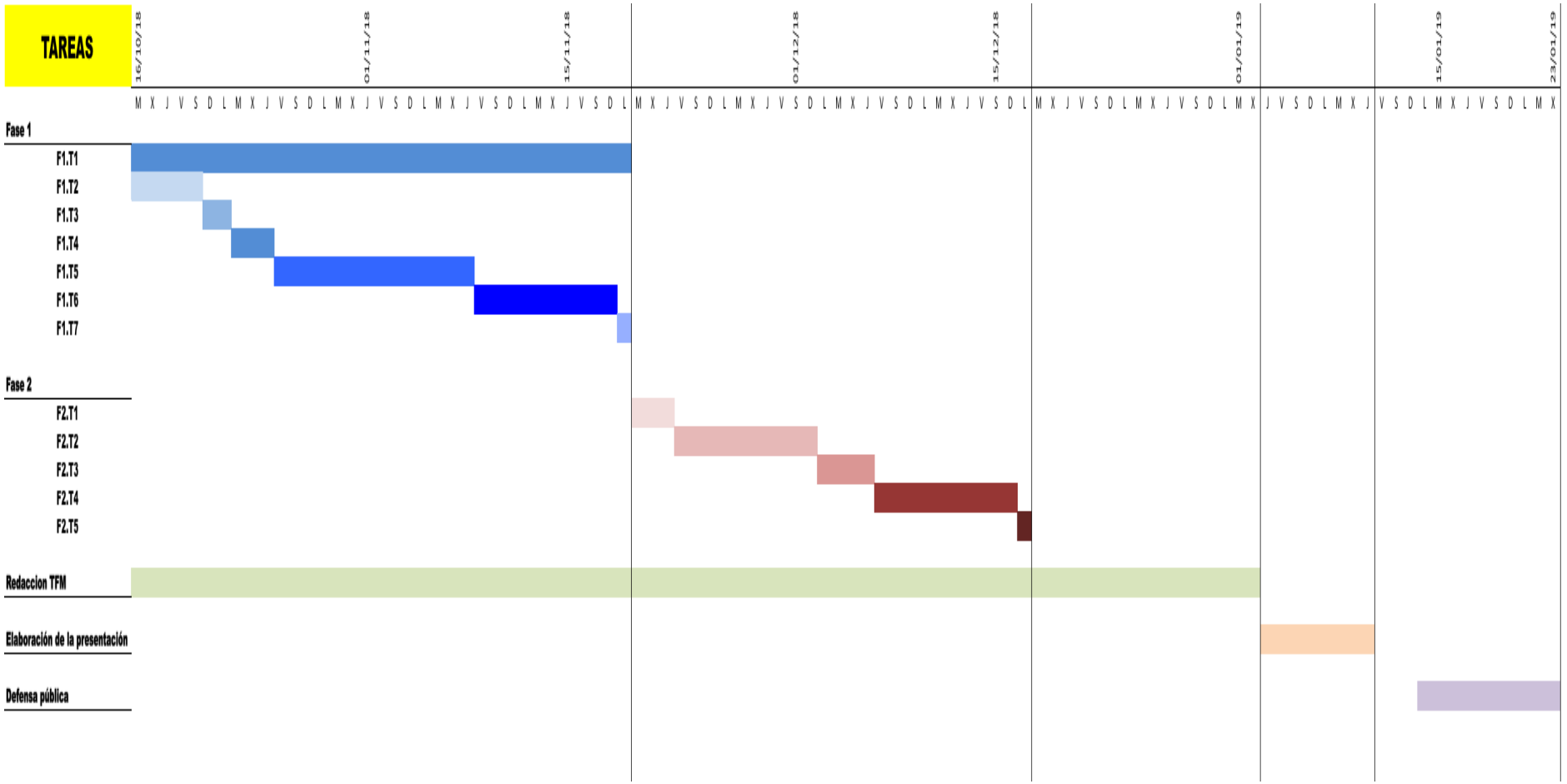


Figura2. Diagrama de Gantt.

Hitos: Durante todo el desarrollo del trabajo hemos planteado varios hitos

1.Entrega del Plan de Trabajo.
2.Entrega de la Fase 1, consistente en la pipeline, un informe del control de calidad de nuestras muestras de RNA-seq, el alineamiento y mapeo de los "reads", estimas de expresión génica y obtención de una la lista de genes diferencialmente expresados.
3.Entrega de la Fase 2, compuesta por el análisis de expresión diferencial entre grupos experimentales, listado de procesos biológicos con los que se asocian nuestros genes expresados diferencialmente, análisis de significación entre nuestros grupos y lista de genes candidatos a inmunocheckpoint.
4.Entrega de memoria y presentación del TFM.

Análisis de riesgos: En la planificación del trabajo también constan los análisis de riesgos que hemos podido tener.

1. Gestión del tiempo. Una mala estimación de la duración de las tareas puede impedir alcanzar los objetivos propuestos. El cálculo de cada tarea es estimado debido a la inexperiencia en este tipo de trabajo, pero se pretende compensar los tiempos entre tareas planificadas si fuera necesario como factor de corrección.

2.Problemas de recursos informáticos. En el caso del análisis del RNA-seq y sabiendo los recursos de los que disponemos cabe la posibilidad que en determinados procesos como puede ser el alineamiento y mapeo de los "reads" al ser archivos muy grandes, pueden aparecen problemas de ejecución o una gran ralentización de las tareas.

3. Elección errónea de métodos. Al existir una gran variedad de software en los diferentes procesos para el análisis específico del RNA-seq. Puede existir una mala comprensión y elección del método informático adecuado pudiendo comprometer el tiempo definido de nuestras tareas retrasándolas.

4. Estrategia de análisis errónea. Al ser un análisis que engloba muchas posibilidades en el tratamiento de datos, una elección equivocada tanto a nivel bioinformático como estadístico puede ocultar resultados relevantes para la investigación y una correcta finalización del TFM.

5. Errores en nuestros datos. Este punto se centra en el análisis final de los datos como que nos falle una replica biológica o que no obtengamos genes diferencialmente expresados. En este caso habrá fallado la planificación e idea inicial, y en este caso se afectaría a la ejecución del proyecto.

6. Falta de planificación y actualización de la Memoria. Es importante llevar al día la realización del documento a entregar permitiendo realizar la entrega en plazo establecido. La menospreciación de esta tarea puede relegar el trabajo a una mala valoración por no tener la calidad adecuada e incluso a comprometer su entrega.

7. Retrasos inesperados. Existen imprevistos de diferentes índoles que se escapan a nuestro control. Si en algún momento ocurriera esto, realizaría un ajuste en la planificación del TFM ajustando los tiempos y objetivos si fuera necesario, aunque este riesgo podría comprometes la entrega del documento.

1.5 Breve sumario de productos obtenidos

Durante el desarrollo de este trabajo se han obtenido los siguientes documentos oficiales:

PEC0- Definición de los contenidos del trabajo.
PEC1- Plan de trabajo
PEC2- Desarrollo del trabajo FASE1
PEC3- Desarrollo del trabajo FASE2
PEC4- Redacción de la Memoria
PEC5- Elaboración de la presentación y Defensa pública

Como productos obtenidos hemos generado:

1. Memoria del TFM y presentación del mismo.
2. Listados de genes diferencialmente expresados. Para la definición de los diferentes estados planteados en el trabajo.
4. Listado de genes candidatos a inmunochekpoint.
5. Información y gráficos tanto de los genes diferencialmente expresados como de las asociaciones de estos a procesos, funciones y componentes.
6. Script desarrollado de todo el proceso.
7. Autoevaluación del proyecto. Documento con cuestiones objetivas para una autoevaluación lo más correcta posible.

1.6 Breve descripción de los otros capítulos de la memoria

En el apartado 2 de capítulos de la memoria, se desarrollará todo el proceso de este proyecto con los siguientes apartados.

Desarrollo: en este apartado se explicará que herramientas y como se utilizaron para desarrollar el proyecto. En este caso, se encuentra intrínseco el apartado habitual de Material y métodos.

Resultados y discusión: en este punto se muestran detallados los resultados obtenidos y ensamblado la discusión pertinente de ellos para resolver los objetivos biológicos planteados.

2. Desarrollo

En este apartado se describirá todo el desarrollo llevado a cabo en este trabajo y se explicarán las tomas de decisiones por las cuales se utilizaron unas u otras herramientas.

2.1 Proceso de obtención de muestras

Inicialmente nos referenciamos al punto de: “Enfoque y método”. Seguido para describir únicamente el sistema utilizado para la extracción de RNA. Y el tipo de muestras de las cuales disponemos.

Primero obtenemos muestras de sangre de donantes sanos en tubos con EDTA, de estos aislamos los PBMC por centrifugación con Ficoll-Histopaque (Biosciences). Las células serán cultivadas con medio RPMI, con estas realizaremos los modelos con diferentes estímulos de LPS especificados en el punto “Enfoque y método”. Obtendremos tres réplicas biológicas, y de cada una de estas tres modelos (Control, LPS y L+L).
La obtención de nuestras muestras será mediante raspado de las placas. Y posteriormente, realizaremos el un protocolo de separación del ARN asociado a polisomas de la membrana del retículo endoplásmico, descrito en: https://pdfs.semanticscholar.org/4705/ae906b05d024a24f0a850e66e32c08521b0e.pdf.
Una vez obtenida las muestras de cada una de las condiciones, el ARN será extraído con el kit “High Pure Isolation kit” (Roche Diagnostics). De estas muestras se realizo el RNA-seq en el servicio de secuenciación del Centro Nacional de Investigaciones Cardiovasculares (CNIC).

2.2 Pipeline

Para el procesamiento del RNA-seq, lo dividimos en cuatro grandes grupos (preprocesado de las lecturas, mapeo, expresión diferencial y anotación funcional), en los siguientes apartados iremos desgranando todo el proceso y las herramientas asociadas a este.

En el siguiente Pipeline podemos ver todo el flujo de trabajo desarrollado, y las diferentes alternativas que nos planteamos en utilizar, hasta obtener la adecuada para nuestro conocimiento y equipamiento.

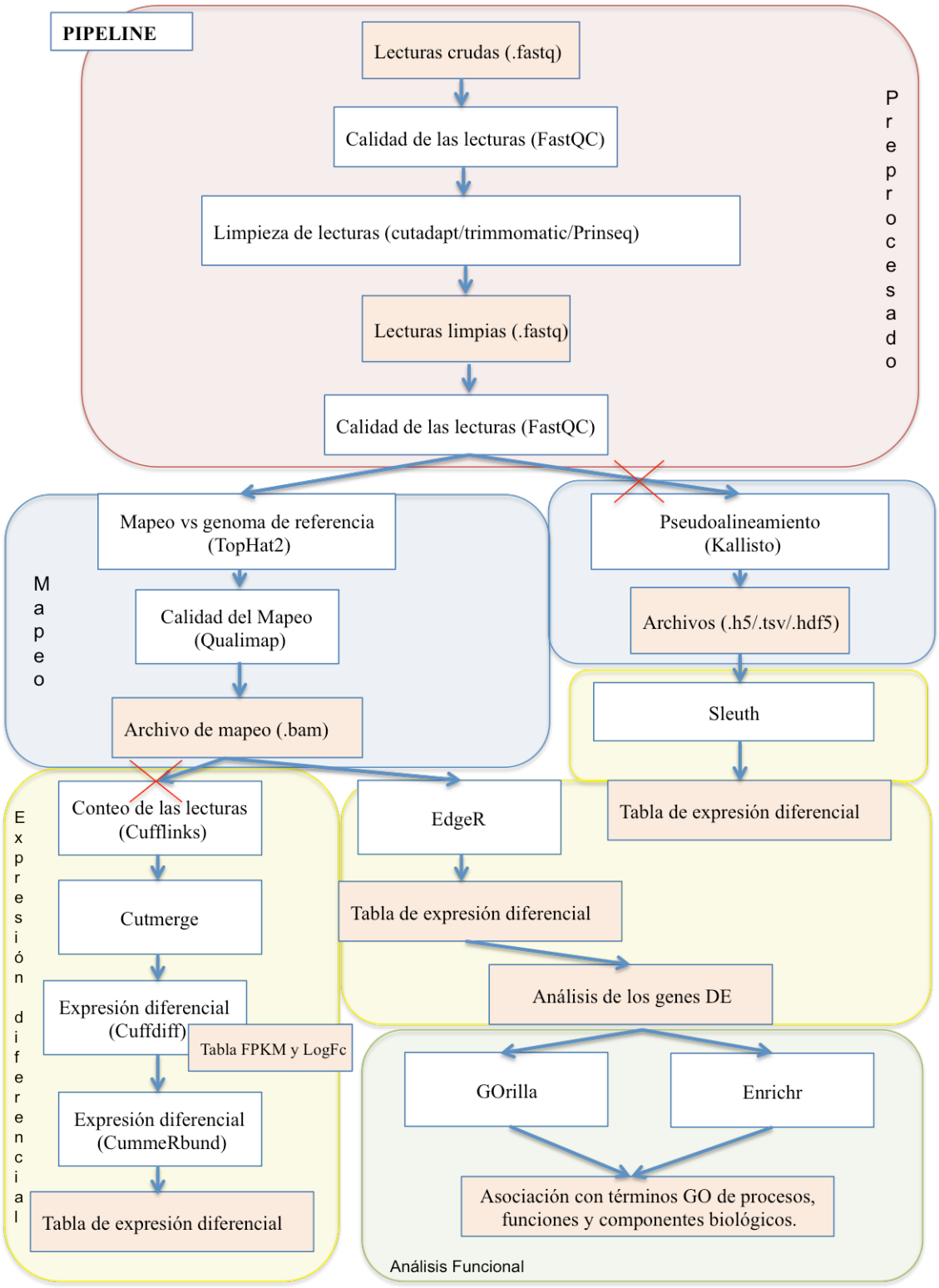


Figura3. Pipeline

2.2 Preprocesado del RNA-seq.

2.2.1 Preprocesamiento de lecturas.

Los secuenciadores actuales generan millones de datos. Estos son imágenes de microscopía tomadas durante la fase de ultrasecuenciación, donde el propio secuenciador es capaz de convertirlos en ficheros de texto en un formato llamado fastq, con los que generalmente se trabaja para el análisis bioinformático. Este fichero contiene la secuencia biológica y la información de la calidad de cada uno de los núcleótidos.

La información de los ficheros fastq contienen las lecturas de secuenciación con la siguiente estructura:

• Una primera línea que determina la identificación de la secuencia, está empieza con el símbolo "@"

• Una segunda línea que contiene la secuencia de la lectura, esta está compuesta por los nucleótidos: Adenina (A), Guanina (G), Citosina (C) y Timina (T), y en caso que aparezca una "N" significa que el nucleótido no se ha podido identificar y se le asigna este carácter.

• Una tercera línea con el símbolo "+" este puede ser único o venir complementado con otra información como tipo de secuenciador, el número de la carrera, el ID del ""flow cell", etc.

• Y una cuarta línea, que nos informa de la calidad de la secuencia para cada una de las bases, codificadas en caracteres ASCII.

Podemos ver nuestros ficheros fastq con el siguiente código:
less mRNA_EQ_LPS_S203_L005_R1_001.fastq

```
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
#==BBEGGGGFGGGGGGGGGGGGGGGGBDGGGGGGGGGEGGGGGGGGGGCG
@D00607:171:CCBMMANXX:5:2301:1174:2110 1:N:0:ATTCCT
NAAGAAATGGGCTACATTTTCTACCCCAGAAAACTACGATAGCCCTTATGA
+
#<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@D00607:171:CCBMMANXX:5:2301:1218:2242 1:N:0:ATTCCT
AGAAGATTTATTGCCTTTAGAACGAATGGGCACAAAAAAAGTAGAAAACTT
+
CCCCCEGGGGGGGGGGGCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@D00607:171:CCBMMANXX:5:2301:1271:2109 1:N:0:ATTCCT
NAATGACCTCCGTTTTTATTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
#=ABB0>?FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGDGGGG
@D00607:171:CCBMMANXX:5:2301:1472:2169 1:N:0:ATTCCT
AAAACATTCTCCTCCGCATAAGCCTGCGTCAGATTAAAACACTGAACTGAC
```

Figura4. Imagen del contenido de un fichero .fastq.

2.2.2 Análisis de calidad

En este apartado se estudia un primer control de calidad de los datos crudos para poder así tratar las lecturas y evitar futuros errores en las restantes fases del análisis RNA-seq. El programa más utilizado para esta labor es FastQC, una herramienta de control de calidad de datos de NGS.

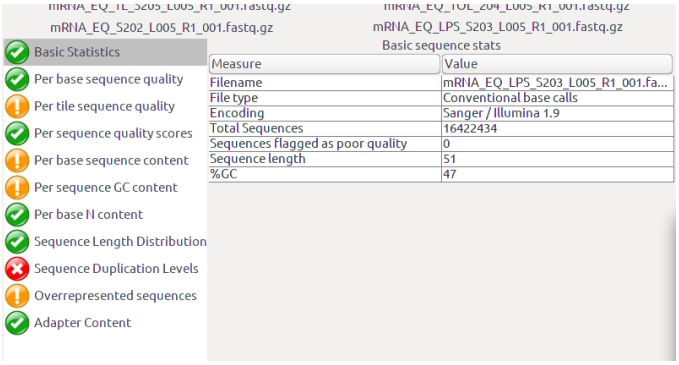


Figura5. Imagen de FastQC

FASTQC es una herramienta cuyo objetivo realizar comprobaciones intuitivas y sencillas de control de calidad de datos procedentes de secuenciación masiva, en nuestro caso se utilizamos la tecnología Illumina. FASTQC utiliza una serie de módulos que mediante gráficos y tablas obtenemos la calidad de nuestras "reads", permitiendo con FASTX Toolkit o Trimmomatic la depuración de estos si fuera necesario.

A continuación describiremos las 12 secciones donde se comprueban la calidad de los datos.

1. Basic Statistics.

La primera sección genera una tabla con el nombre y tipo de archivo, el programa con el que se ha secuenciado (codificación ASCII), el número total de secuencias procesadas, el número de secuencias con baja calidad, el tamaño de cada secuencia (es aconsejable que la longitud sea la misma para evitar complicaciones) y el porcentaje de bases GC (este % es correcto cuando es mayor del 45%).

Measure	Value
Filename	mRNA_EQ_LPS_S203_L005_R1_001.fa...
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	16422434
Sequences flagged as poor quality	0
Sequence length	51
%GC	47

Figura6. Imagen de FastQC (Basic Statistics)

2. Per Base Sequence Quality.

En esta sección mostramos un gráfico que permite ver la calidad a lo largo de todas las bases en cada posición (nucleótido) de nuestras secuencias. El valor de calidad se denomina Q y puede ir de 0 a 40, considerándose aceptable los valores de Q iguales o mayores de 25.

En cada posición se dibuja un box plot donde la línea roja representa la mediana, la caja amarilla el rango intercuartílico 25-75%, las líneas superiores e inferiores el 10% y el 90% y la línea azul la calidad media. En el eje Y,

observamos el valor Q de calidad . Los colores del gráfico representan la calidad: Color verde muy buena calidad , color naranja calidad razonable y color rojo baja calidad.

La calidad se suele degradar a medida que la secuenciación avanza a lo largo de la lectura, debido al aumento de incorporación de más nucleótidos mayor posibilidad de error. En nuestras lecturas podemos observar una muy alta calidad.

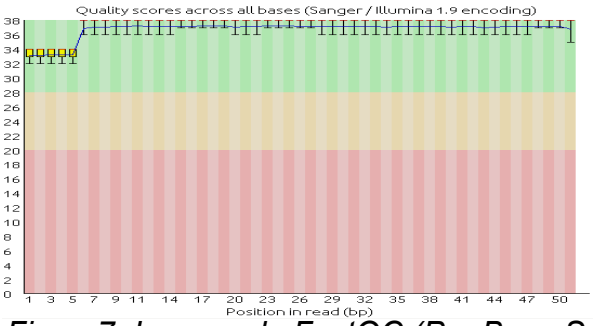


Figura7. Imagen de FastQC (Per Base Sequence Quality)

3. Per Tile Sequence.

Esta sección nos informa de una correcta reacción dentro de la célula de flujo. Cuando la reacción se produce correctamente el gráfico solo muestra un fondo azul, si aparecen manchas indican las interferencias en la reacción, estas suelen ser debidas a burbujas o suciedad.

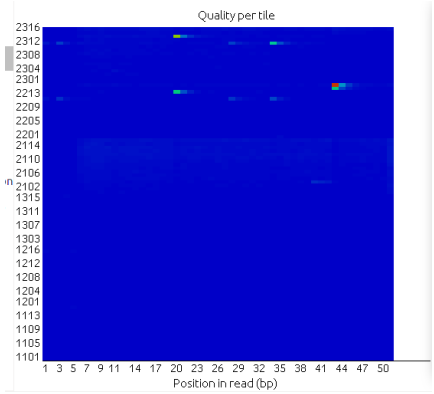


Figura8. Imagen de FastQC (Per Tile Sequence.)

4. Per Sequence QualityScores.

En esta sección observamos un gráfico de valores de calidad en nuestros conjunto de datos. Donde el eje de las x es el valor de Q, anteriormente explicado y en el eje y es el número de secuencias. En este caso la calidad es incontestable, si existiera algún conjunto de baja calidad aparecería un pico por debajo de un valor de Q de 25.

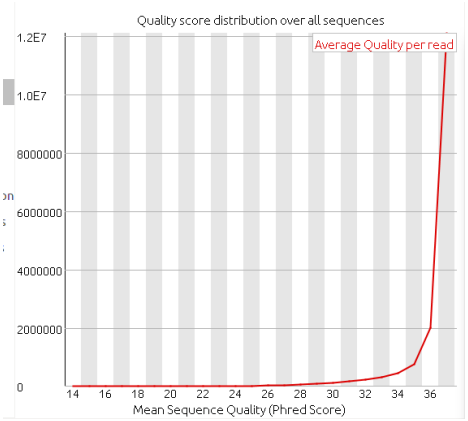


Figura9. Imagen de FastQC (Per Sequence QualityScores)

5. Per Base SequenceContent.

Este apartado nos muestra la información del porcentaje de cada nucleótido en la secuencia, posición por posición. Quitando la desproporción inicial, probablemente debido al inicio del proceso, vemos que las líneas son paralelas esto refleja un equilibrio entre las bases en el genoma, y demuestra la calidad de nuestras muestras. El programa nos daría un señal de advertencia cuando la diferencia entre bases sea mayor de un 10%, y si supera el 20% sería inaceptable. También cabe destacar que generalmente las CG deberían estar por encima de las AT, esta curiosidad es debida al tipo de muestra secuenciada por eso no tendremos ningún problema en validar la calidad del proceso.

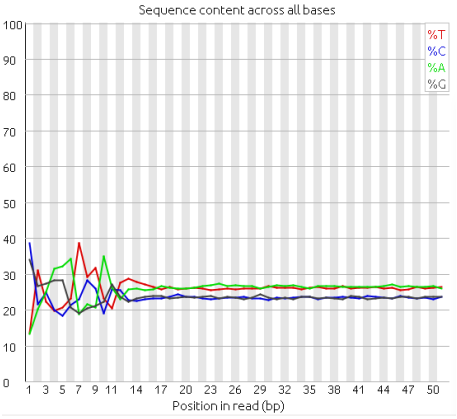


Figura10. Imagen de FastQC (Per Base SequenceContent)

6. Per SequenceGC Content.

En este punto observamos el porcentaje de CG de nuestras secuencias. La línea roja muestra el contenido real de CG frente a la línea azul que muestra la distribución teórica a la que debe asemejarse dicho porcentaje. Nuestras muestras cumplen con el criterio de calidad, si existiera algún pico fuera de esta campana de Gauss podría deberse a una contaminación con ADN exógeno, dimerización de cebadores o a que una mala calidad de la muestra se esta leyendo como CG, etc.

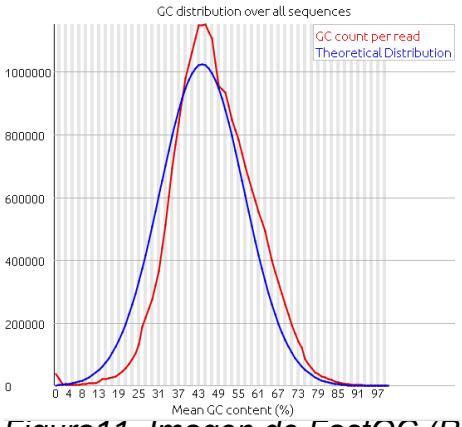


Figura11. Imagen de FastQC (Per SequenceGC Content)

7. Per Base N Content.

Este apartado nos indica el %N colocado para cada posición de las secuencias, esto quiere decir que cuando la calidad de la secuenciación es mala y no es capaz de asignar una base este introduce una N. En nuestro caso no tenemos ninguna N.

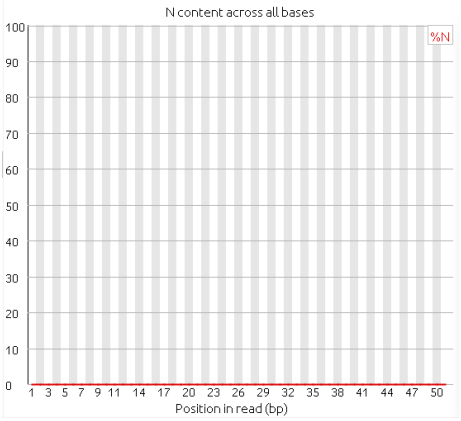


Figura12. Imagen de FastQC (Per Base N Content)

8. Sequence Length Distribution.

Este apartado nos muestra la longitud de las secuencias analizadas. En nuestro caso las secuencias son de 51pb de longitud.

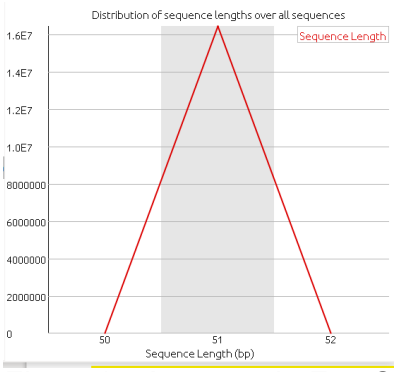


Figura13. Imagen de FastQC (Sequence Length Distribution)

9. Sequence Duplication levels.

En este punto observamos el porcentaje de secuencias que se encuentran duplicadas y el nivel de duplicación de las mismas. En el eje y se representa el porcentaje de reads repetidas y en el eje x, el número de repeticiones.

En nuestras muestras observamos que aparecen picos azules alrededor del 10% y nos informa que nuestras secuencias se repiten más de 10 veces y hasta más de 500 veces. Estas duplicaciones podrían deberse a algún artefacto de la PCR, pero en nuestro caso debemos recordar la procedencia de nuestra muestra, ya que al haber sido obtenida con esa extracción diferencial , asumimos la normalidad de este dato.

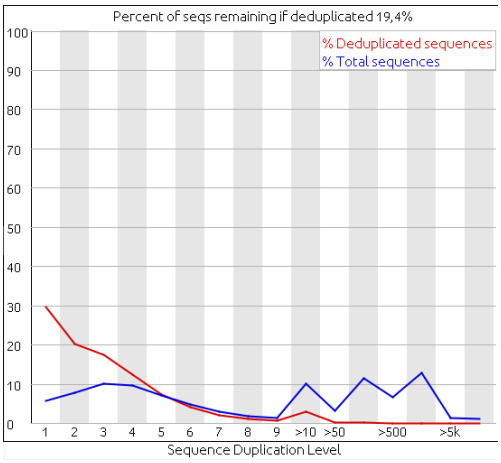


Figura14. Imagen de FastQC (Sequence Length Distribution)

10. Overrepresented Sequences.

En esta sección obtenemos una tabla de aquellas secuencias que se repiten un número elevado de veces. Este dato nos informa de posibles dimerizaciones de adaptadores usados en la secuenciación, cuando existe esta secuenciación de adaptadores puede surgir un ensamblado falso de las secuencias debido a que este fragmento no pertenece al genoma.

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TTGTACTCTAGTT...	52601	0,32	No Hit
TTTTTTTTTTTTT...	21013	0,128	No Hit
AAAAAAAAAAA...	16555	0,101	No Hit

Figura15. Imagen de FastQC (Overrepresented Sequences)

En nuestro caso existen tres secuencias sobrerrepresentadas ya que superan el 0,1% de representación del total de secuencias. En el caso de un porcentaje mayor del 1%, FastQC lo señalaría como erróneo.

11. Adapter Content.

El último apartado muestra el contenido de adaptadores, en nuestras muestras las secuencias no nos muestran adaptadores existentes.

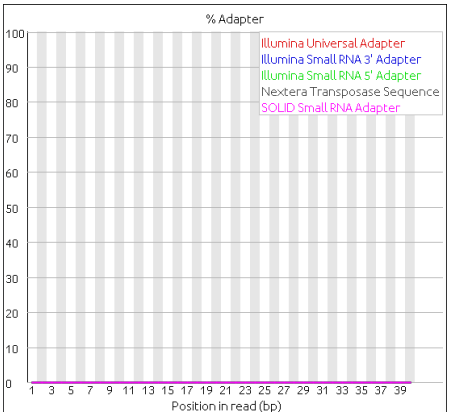


Figura16. Imagen de FastQC (Adapter Content)

Los resultados son guardados en un archivo HTML, donde podemos ver un resumen de cada una de las secciones , lo realizamos para cada muestra.

A continuación muestro una tabla resumen de los datos obtenidos de FastQC, informándonos de la calidad de nuestras muestras.

Nombre Muestra	Nº Secuencias	Tamaño Secuencias	%GC
EQ_Control (C1)	20345902	51	47
JJ_Control (C2)	14175948	51	47
R_Control (C3)	26911891	51	47
EQ_LPS (LPS1)	16422434	51	47
JJ_LPS (LPS2)	15312231	51	47
R_LPS (LPS3)	17263668	51	47
EQ_TL (L+L1)	11230653	51	47
JJ_TL (L+L2)	17656650	51	47
R_TL (L+L3)	18745745	51	47

Tabla 1. Tabla informativa de nuestras muestras.

2.2.3 Depuración de secuencias.

En este proceso de depuración utilizo dos herramientas las cuales pueden utilizarse con la misma finalidad.

1. Eliminación de adaptadores de secuenciación (Cutadapt).

En el proceso de secuenciación utilizamos una librería de adaptadores (secuencias) donde el material de entrada de doble cadena es ligado a estos, los cuales son complementarios a oligos fijados en la superficie del flowcell. Puede ocurrir que una molécula de ARN/ADN o el fragmento de ADN sea más

corto que el número de ciclos, y que se secuencie parte del 3' adaptador, en este caso debemos localizar esta secuencia y eliminarla de la lectura por eso en muchos de los casos la calidad de la secuenciación va deteriorándose hacia el extremo 3'. Esta eliminación de adaptadores es crucial para la calidad del alineamiento ya que si no se realiza puede existir perdidas o alineamientos erróneos.
Observando la calidad de nuestras muestras, es posible que no sea necesario la eliminación de adaptadores, aunque con la iniciativa de aprendizaje de este paso en el filtrado y calidad de nuestros datos vamos a utilizar la herramienta Cutadapt para la eliminación de estos.

Quitaremos los adaptadores con el programa cutadapt de nuestras secuencias en ficheros fastq. Utilizando el siguiente código.

Cutadapt -a <Adaptador> -e 0.1 -m 30 mRNA_EQ_LPS_S203_L005_R1_001.fastq -o New.fastq

Los caracteres del código anterior tienen el siguiente significado:
-a: Indica la secuencia nucleotídica del adaptador, la secuencia de los adaptadores viene en las especificaciones de la librería, en nuestro caso son los adaptadores estándar de Illumina. Hay veces que en los fastqc podemos verlos como "Overrepresented sequences" (cuando aparecen en esta sección los marca como adaptador de illumina e incluso muestra el índice que tenía).
-e: Ratio de error, 0,1.
-m: Minimun lenght, longitud mínima de lectura. En este caso ponemos un límite inferior de 30 bp para no quedarnos con lecturas muy pequeñas así que los "reads" filtrados tienen una longitud [30-51]
-o: Directorio donde indicaremos el archivo que se guardara.

En el siguiente link mostramos la guía del programa Cutadapt:
http://cutadapt.readthedocs.io/en/stable/guide.html

Para una visión de los código del programa siempre podemos teclear en el terminal:
cutadapt --help

Posteriormente, volveremos a comprobar con FastQC el control de calidad para cerciorarnos de la eliminación de estos adaptadores.

2. Refinamiento de lecturas (Prinseq).

En el desarrollo del pipeline inicial se propuso un paso de refinamiento de lecturas, observadas los informes de calidad de nuestra experimentación del programa FastQC, este paso no es necesario hacerlo en este caso, una vez habiendo procesado las muestras con el programa Cutadapt. Aunque si consideramos necesario describirlo. Este punto se centra en eliminar las bp de baja calidad y lecturas que en su totalidad no son aptas para su alineamiento. Con ello evitamos inferencias en el correcto alineamiento de las lecturas frente al genoma de referencia.

En este paso utilizaríamos la herramienta Prinseq, con esta herramienta podemos realizar varias funciones como eliminar secuencias de menos de unos nucleótidos determinados (-min_len 30), seleccionar la calidad media de las lecturas es decir que la calidad sea mayor o igual a 30 por ejemplo (-min_qual_mean 30) o eliminar nucleótidos de un extremo con calidad menor de 25 (-trim_qual_right 25) . También en este proceso puede darnos las lecturas sin adaptadores, las lecturas desechadas e incluso un informe gráfico.

Posteriormente, comprobamos de nuevo la calidad con FastQC, donde deberíamos observar grandes mejoras en las secuencias si estas lo requieren y así facilitar la fase de alineamiento posterior.

Nombre Muestra	Nº Secuencias	Tamaño Secuencias	%GC
EQ_Control (C1)	20332804	30-51	47
JJ_Control (C2)	14170869	30-51	47
R_Control (C3)	26877005	30-51	47
EQ_LPS (LPS1)	16398318	30-51	47
JJ_LPS (LPS2)	15301231	30-51	47
R_LPS (LPS3)	17251539	30-51	47
EQ_TL (L+L1)	11197511	30-51	47
JJ_TL (L+L2)	17650326	30-51	47
R_TL (L+L3)	18724556	30-51	47

Tabla 2. Tabla informativa realizada con los datos de FastQC de nuestras muestras depuradas.

2.3 Mapeo de las lecturas.

2.3.1 Mapeo de las lecturas (TopHat2 y Kallisto)

Posterior al análisis de calidad de las lecturas, junto con la depuración de las mismas, procedemos al alineamiento y mapeo frente a nuestro genoma de referencia que es el Hg38 (humano). En este paso el objetivo fundamental es tratar de conocer nuestras lecturas ubicándolas con la ayuda de nuestro genoma de referencia. Para la realización de este procedimiento utilizamos las herramientas : TopHat2 y Kallisto. A posteriori describimos cada una de ellas. La decisión de utilizar diferentes herramientas en todo el procesamiento es debido al cumplimiento del objetivo1.

1. TopHat2

Primero, describimos el procesamiento con la herramienta TopHat2, este es un mapeador de unión de empalme rápido para lecturas RNA-seq. En este caso alinea las lecturas de RNA-seq con el genoma Hg38 de humano utilizando bowtie2 que es un alineador de lectura corta de alto rendimiento, posteriormente analiza los resultados del mapeo para identificar uniones de empalme entre exones. TopHat2 es la herramienta de mapeado más ampliamente utilizada.

El programa realiza optimizados donde se mejora la velocidad y trata de evitar el alineamiento de lecturas en pseudogenes, estos son genes que no se expresan ya que aunque su secuencia sea similar a la de un gen “normal” sus productos no son funcionales. El funcionamiento se puede resumir en tres pasos principales,
1) Primer alineamiento de lecturas frente a regiones cuya información de anotación está disponible (archivo GTF/GFF).
2) Aquellas lecturas que no alinean completamente frente al transcriptoma se alinean al genoma, utilizando bowtie2.
3) Las lecturas que han quedado finalmente sin alinear son fragmentadas en segmentos menores (25bp por defecto) y mapeadas de nuevo frente al genoma. TopHat2 buscará ahora splice junctions, indels y puntos de fusión.

El proceso bioinformático que seguimos para la aplicación de este software es:

Preparación del genoma de referencia:

Para encontrar uniones con TopHat2, primero deberá instalar un índice Bowtie para el organismo en su experimento RNA-seq, en este caso para humanos. Las lecturas que corresponden al organismo de humano las descargue en: http://hgdownload.soe.ucsc.edu/downloads.html#human, utilizando el Genoma Dec. 2013 (GRCh38/hg38)
Lo descomprimimos con Ubuntu, en este caso directamente en las ventanas, aunque lo podemos hacer en un entorno Linux con el comando:

gzip -d hg38.fa.gz

Renombramos a Hg38 con el código.

mv Hg38.fa hg38

Utilizaremos Bowtie2 para poder indexar el genoma. El código sería el siguiente:

bowtie2-build hg38.fa hg38

La herramienta Bowtie2 ordena una serie de datos o informaciones de acuerdo a un criterio común a todos ellos, para facilitar su consulta y análisis. Este código anterior nos genera ficheros binarios los cuales usaremos a posteriori para realizar el alineamiento. Estos ficheros generados tienen una extensión .bt2.

Figura17. Imagen de los ficheros generados con bowtie2.

A continuación indexamos el archivo de anotación con extensión .GTF (también obtenido de la página UCSC) con nuestro genoma de referencia en nucleótidos, acelerando el proceso de mapeo. Utilizando el siguiente código.

tophat2 --G hg38.gtf --transcriptome-index=hg38.tr hg38.fa

```
victor@victor-MacBookPro:~/Escritorio/data/working$ tophat2 --G hg38.gtf -transc
riptome-index=hg38.tr hg38.fa
tophat: option -t not recognized
    for detailed help see http://ccb.jhu.edu/software/tophat/manual.shtml
victor@victor-MacBookPro:~/Escritorio/data/working$ tophat2 --G hg38.gtf --trans
criptome-index=hg38.tr hg38.fa
```

Figura18. Imagen del código de uso de TopHat2.

Finalmente realizamos el mapeo de nuestras muestras. Con el siguiente código:

tophat2 –o carpeta_salida –p 4 –-transcriptome-index= hg38.gtf hg38 New.fastq

Los parámetros del comando anterior indican:
-o/-- output-dir: Nombre del directorio de salida, es decir indica donde TopHat2 guardará todos tus archivos de salida.
-p/--num-threads: Determina el número de “threads” para alinear lecturas, por defecto es 1, en este caso utilizamos 30.
--transcriptome-index <dir/prefix>: Cuando se suministra a TopHat2 un documento conocido transcrito (parámetro –G/--GTF), se produce un transcriptoma y se crea un índice de Bowtie para alinear las lecturas con los tránscritos conocidos.

Durante el proceso de mapeo TopHat2 desecha las lecturas menores de 12bp. En este proceso se generan diferentes archivos de salida para cada muestra, en la carpeta que hemos indicado, en este caso (carpeta_salida):

```
victor@victor-MacBookPro:~/Escritorio/data/working/bow/tophat_out$ ls
accepted_hits.bam      align_summary.txt  insertions.bed  logs            unmapped.bam
accepted_hits.bam.fai  deletions.bed      junctions.bed   prep_reads.info
```

Figura19. Imagen de los archivos de salida obtenidos de TopHat2.

deletions.bed : este archivo muestra las delecciones sobre el genoma.
junctions.bed : este archivo muestra las uniones de exón.
insertions.bed : este archivo muestra las inserciones.
prep_reads.info : archivo de control de preparación de lecturas previo al mapeo tophat.
align_summary.txt : archivo resumen del alineamiento (número de lecturas de inicio, número de lecturas mapeadas, % mapeo).
Directorio logs : archivos donde se muestran los pasos del proceso.
Unmapped.bam : archivo binario donde se muestran las lecturas no mapeadas.

El archivo BAM es la versión binaria comprimida del formato de Alineación / Mapa de Secuencias (SAM), una representación compacta e indexable de las alineaciones de secuencias de nucleótidos. Muchas herramientas de análisis y

secuenciación de próxima generación funcionan con SAM / BAM. Y el archivo BED sirve para definir las líneas de datos que son mostradas al usuario. Cada línea BED tiene tres campos obligatorios (cromosoma, comienzo y fin) y nueve que son adicionales/opcionales (nombre, puntuación, hebra, etc.).

Con el alineamiento, obtenemos diferentes tipos de lecturas las cuales las podemos ver en el archivo align_summary.txt :
1. No alineadas.
2. De alineamiento único.
3. Lecturas multimapeo, que se alinean en más de una localización del genoma.

Es importante conocer estas lecturas ya que un alto porcentaje de lecturas con alineamiento múltiple puede llevarnos a alineamientos erróneos y aumentar la cobertura de las lecturas sobre el genoma dando un resultado irreal.

Nombre Muestra	Mapeo total %	Multimapeo %	Mapeo único%
EQ_Control (C1)	88.1	7.6	80.5
JJ_Control (C2)	93.3	7.3	86.0
R_Control (C3)	80.0	10.2	69.8
EQ_LPS (LPS1)	89.9	9.9	80.0
JJ_LPS (LPS2)	93.2	7.5	85.7
R_LPS (LPS3)	90.9	10.3	80.6
EQ_TL (L+L1)	90.2	9.1	81.1
JJ_TL (L+L2)	93.6	6.2	87.4
R_TL (L+L3)	90.2	7.8	82.4

Tabla3. Porcentaje de lecturas mapeadas con TopHat2

2. Kallisto

En este procesamiento de mapeo también utilizamos la herramienta Kallisto , esta utiliza un método de cuantificación de RNA-seq "libre de alineación", programa que en un principio nos debe agilizar este proceso. Kallisto es un programa para cuantificar abundancias de transcripciones a partir de datos de RNA-seq, o más generalmente de secuencias objetivo que utilizan lecturas de secuenciación de alto rendimiento. Se basa en la novedosa idea de pseudoalineación para determinar rápidamente la compatibilidad de las lecturas con los objetivos, sin la necesidad de alineación. En los puntos de referencia con datos estándar de RNA-seq, Kallisto puede cuantificar millones de lecturas humanas utilizando solo las secuencias de lectura y un índice de transcriptoma. La pseudoalineación de las lecturas conserva la información clave necesaria para la cuantificación, por lo que no solo es rápida, sino que también es tan precisa como las herramientas de cuantificación existentes. De hecho, debido a que el procedimiento de pseudoalineación es robusto a los errores en las

lecturas, en muchos puntos de referencia kallisto supera significativamente las herramientas existentes.

En el proceso hemos instalado Kallisto y nos iniciamos en su uso.

Realizo el índice con index con el siguiente código:
kallisto index -i transcripts.idx transcripts.fasta.gz

```
victor@victor-MacBookPro:~/kallisto/test$ kallisto index -i transcripts.idx transcripts.fasta.gz

[build] loading fasta file transcripts.fasta.gz
[build] k-mer length: 31
[build] counting k-mers ... done.
[build] building target de Bruijn graph ...  done
[build] creating equivalence classes ...  done
[build] target de Bruijn graph has 27 contigs and contains 22118 k-mers
```

Figura20. Imagen del código de uso de Kallisto.

con kallisto quant y los archivos test corro la cuantificación de estos archivos.
kallisto quant -i transcripts.idx -o output reads1.fastq.gz reads□.fastq.gz

```
victor@victor-MacBookPro:~/kallisto/test$ kallisto quant -i transcripts.idx -o output reads_1.fastq.gz reads_2.fastq.gz

[quant] fragment length distribution will be estimated from the data
[index] k-mer length: 31
[index] number of targets: 14
[index] number of k-mers: 22,118
[index] number of equivalence classes: 20
[quant] running in paired-end mode
```

Figura21. Imagen del código de uso de Kallisto.

En este procesamiento incurrimos en uno de los riesgos propuestos debido a que el equipo utilizado no es lo suficiente potente para el procesamiento de nuestras muestras. Podemos corroborar esto con los test del propio programa y con la prueba de código de nuestro consultor.

A continuación muestro el código utilizado para su uso y el error pertinente.
Realizo el índice con index con el siguiente código
kallisto index -i indice.idx Homo_sapiens.GRCh38.cdna.all.fa
Debería continuar con este código y todas mis muestras.
Kallisto quant -i indice.idx -o carpeta_salida New1.fastq New2.fastq

```
victor@victor-MacBookPro:~/kallisto/test$ kallisto index -i indice.idx Homo_sapiens.GRCh38.cdna.all.fa

[build] loading fasta file Homo_sapiens.GRCh38.cdna.all.fa
[build] k-mer length: 31
[build] warning: clipped off poly-A tail (longer than 10)
        from 1421 target sequences
[build] warning: replaced 3 non-ACGUT characters in the input sequence
        with pseudorandom nucleotides
[build] counting k-mers ... terminate called after throwing an instance of 'std::bad_alloc'
  what():  std::bad_alloc
Abortado (`core' generado)
```

Figura22. Imagen del código de uso y error existente en Kallisto.

Realizamos varias pruebas intentando crear el índice con varios archivos del trancriptoma humano y siempre obtenemos el mismo error, parece ser la limitación de mi equipo. Kallisto se nutriría para la cuantificación de un conjunto de secuencias FASTA de nuestras muestras, es decir, no realiza el ensamblado de la transcripción y no puede cuantificar la expresión de las

nuevas transcripciones que no están en el índice de la transcripción que le proporcionamos. Entonces debido a este error no podemos continuar con este desarrollo.

Si quisiéramos seguir con este procesado con kallisto deberíamos utilizar posteriormente Sleuth, este es un programa para el análisis de experimentos RNA-seq para los cuales se han cuantificado las abundancias de transcritos con kallisto. Este proporciona herramientas para el análisis de datos exploratorios utilizando Shiny by RStudio, e implementa algoritmos estadísticos para el análisis diferencial que aprovechan las estimaciones de Kallisto. Las características clave de la herramienta son: La capacidad de realizar tanto análisis de nivel de transcripción como de nivel de gen. Compatibilidad con kallisto que permite un flujo de trabajo rápido y preciso desde lecturas a resultados. El uso de boostraps para determinar y corregir variaciones técnicas en experimentos.

Finalmente, este proceso se realizo con Tophat2 pudiendo salvar los riegos de equipamiento con la cesión de una “work station” del tutor externo. Todos los códigos aquí descritos son perfectamente funcionales, pero debido a la limitación del equipo se comprobaron con una reducción de nuestras muestras.

2.3.2Análisis de la calidad del mapeo (Qualimap).

Una vez obtenidas las lecturas frente a nuestro genoma de referencia también disponemos de su localización y la anotación frente al genoma. Antes de realizar el análisis de expresión diferencial debemos chequear la calidad de estas lecturas para comprobar si ha habido problemas de saturación, en la profundidad de secuenciación, si estas lecturas se distribuyen uniformemente en el genoma, etc.

Para ello utilizamos Qualimap que es una aplicación escrita en Java y R que proporciona tanto una interfaz gráfica de usuario (GUI) como una interfaz de línea de comandos para facilitar el control de calidad de los datos de secuencia de alineación y sus derivados, como el conteo de características. Esta herramienta soporta diferentes tipos de experimentos: Secuenciación del genoma completo, secuenciación de todo el exoma, RNA-seq o ChIP-seq.

Esta examina los datos de alineación de la secuenciación en los archivos SAM / BAM de acuerdo con las características de las lecturas asignadas y proporciona una vista general de los datos que ayuda a detectar sesgos en la secuenciación y/o el mapeo de los datos y facilita la toma de decisiones para análisis posteriores. Incluso las nuevas versiones nos pueden proporcionar una comparación de múltiples muestras de alineación y datos de conteos. Se puede ejecutar por comandos o ahora mismo también tenemos un interfaz de ventanas. Mostramos a modo ejemplo algunas de las gráficas realizada con una de nuestras muestras.

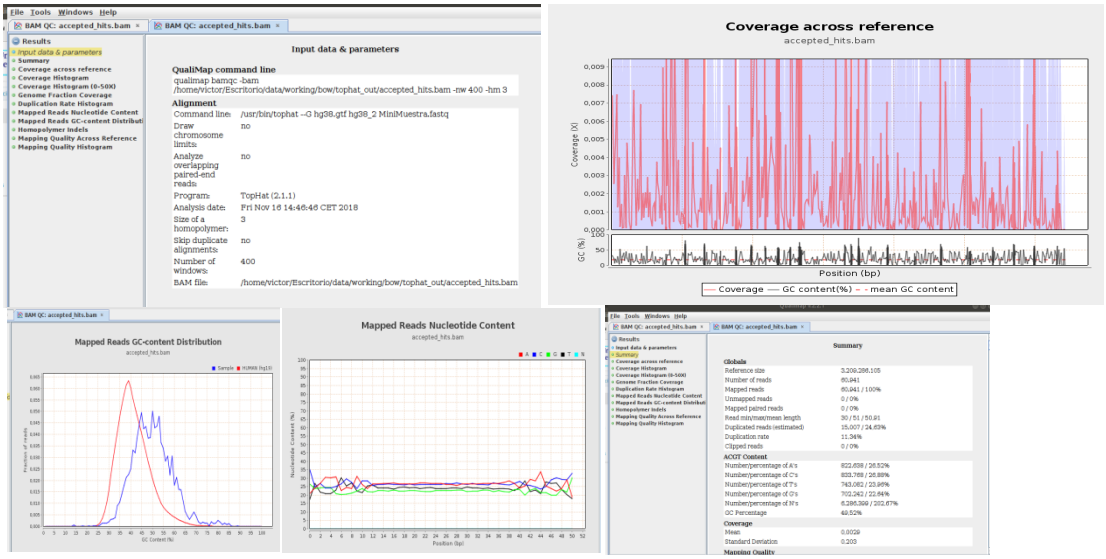


Figura23. Imágenes de Qualimap, con una de nuestras muestras.

Con Qualimap obtenemos: Análisis rápido a través de la referencia de cobertura de genoma y distribución de nucleótidos; resumen fácil de interpretar de las principales propiedades de los datos de alineación; análisis de las lecturas asignadas dentro / fuera de las regiones provistas en formato GFF; cálculo y análisis de los recuentos de lectura obtenidos de la intersección de alineaciones de lectura con características genómicas; análisis de la adecuación de la profundidad de secuenciación en los experimentos RNA-seq; comparación multimuestra de alineación y datos de conteos; agrupación de perfiles epigenómicos.

En relación a nuestras lecturas podemos decir que tienen una buena calidad para continuar con el procesamiento.

2.4 Análisis de expresión diferencial.

El siguiente punto a realizar es el análisis de la expresión diferencial. En este caso describimos el procesamiento con EdgeR una herramienta de Bioconductor. De nuevo en este apartado realizamos el proceso con varias herramientas (EdgeR y Cufflinks;Cuffmerge;Cuffdiff), debido a la importancia que dada al objetivo1. Aunque el procesamiento con Cufflinks; Cuffmerge; Cuffdiff se encuentra descrito en el Anexo III.

2.4.1 Análisis de Expresión Diferencial

Para el proceso de análisis de expresión diferencial voy a utilizar nuestras lecturas y la herramienta EdgeR, debido a un mayor conocimiento de ésta y mejor desarrollo y rapidez del proyecto, a continuación definimos todos los procesos llevados a cabo en R con estas librerías y obtendremos finalmente varias gráficas y tablas de resultados del proyecto.

En los pasos anteriores obtuvimos un conteo de lecturas, es decir una matriz de conteo la cual utilizaremos para el análisis de expresión genética diferencial. Como hemos referido anteriormente existen numerosas herramientas para realizar este análisis, las diferencias entre ellas son los métodos de normalización, el filtrado de características (genes, tránscritos) con bajo nivel de expresión, los modelos probabilísticos, los correspondientes test para el análisis de la expresión diferencial, etc.

En este proceso utilizamos R que es un entorno y lenguaje de programación, de software libre, de los más utilizados en investigación biomédica y bioinformática por su enfoque al análisis estadístico. Este dispone de la posibilidad de cargar diferentes librerías/paquetes con funcionalidades de cálculo y graficación, enfocadas a diferentes campos.

En este proceso utilizamos Bioconductor que es un proyecto basado en el lenguaje de R para el análisis de datos en genómica. Y para llevar a cabo dicho análisis utilizaremos de este sus diferentes librerías con el objetivo de comparar el conteo de lecturas para cada tránscrito/gen con diferentes condiciones biológicas, mediante test estadísticos. Dichos paquetes son: EdgeR, NOISeq, DESEq2 y Limma. Nosotros utilizaremos EdgeR

EdgeR es un paquete de software Bioconductor para examinar la expresión diferencial de datos de conteo replicados. Utiliza un modelo de Poisson sobredispersado para tener en cuenta la variabilidad tanto biológica como técnica. Los métodos empíricos de Bayes se utilizan para modelar el grado de sobredispersión en las transcripciones, lo que mejora la confiabilidad de la inferencia. Esta librería modela los datos de conteo mediante una distribución Binomial Negativa. La metodología se puede utilizar incluso con los niveles más mínimos de replicación, siempre que se replique al menos un fenotipo o condición experimental.

Un vez definidas las herramientas a usar, describimos paso a paso el proceso a seguir para obtener la tabla de datos expresados diferencialmente, y en siguientes apartados desarrollaremos las diferencias significativas que pueden existir en estos datos y en cada condición biológica.

Lectura de la muestra.

Primero instalo Bioconductor y todas las librerías necesarias para el desarrollo del proceso.

```
source("http://bioconductor.org/biocLite.R")
require(biocLite)
biocLite("edgeR")
biocLite("DESeq2")
require(edgeR)
require(DESeq2)
install.packages("DescTools")
require("DescTools")
install.packages("ggplot2")
```

require("ggplot2")

Leemos nuestro dataset, recordamos que en este caso son las lecturas.

```
y <- read.delim("~/Desktop/results.txt", header=FALSE, row.names=1)
> dim(y)
[1] 46189    10
```

Posteriormente definimos los grupos, en este caso tenemos tres grupos (Control, LPS, L+L) y tres replicas biológicas por cada grupo. Construimos el data.frame "targets" con los etiquetas de cada muestra:

```
Lane <- c(1,1,1,2,2,2,3,3,3)
Treatment <- c("C","C","C","LPS","LPS","LPS","L+L","L+L","L+L")
Label <- c("C1","C2","C3","LPS1","LPS2","LPS3","L+L1","L+L2","L+L3")
targets <- data.frame(Lane,Treatment,Label)
rownames(targets) <- Label
targets
```

A continuación utilizamos DGEList para crear un objeto donde asociamos nuestros datos con las etiquetas que hemos creado

```
> y <- DGEList(counts=y[,2:10], group=targets$Treatment)
> colnames(y) <- targets$Label
> colnames(y)
[1] "C1"   "C2"   "C3"   "LPS1"
[5] "LPS2" "LPS3" "L+L1" "L+L2"
[9] "L+L3"
```

Una vez obtenido el objeto con la clase DGEList se puede comenzar con el análisis. Se podrán hacer estudios estadísticos de este nuevo objeto al igual que se podían hacer antes a la matriz conteo, ya que no se ha perdido ninguna información.

Filtrado de genes.

Este paso es importante para continuar con el análisis, ya que su cometido es eliminar los genes con mínimos conteos o sin ellos, debido a su baja expresión o incluso nula en algunas de nuestras condiciones de estudio. Con ello evitamos posibles problemas en el posterior uso de las funciones logarítmicas del paquete EdgeR.

Este método realizado es el llamado conteos por millón (CPM), consiste simplemente en eliminar aquellos tránscritos/genes que presentan un número de lecturas inferior a un determinado número de CPM. Debido a la agrupación de nuestras muestras en tres grupos, seleccionamos los genes que logran al menos un cpm (conteos por millón) de tres.

Primero seleccionamos estos genes. Donde la función cpm(-) calcula los conteos por millón de cada muestra, es decir, la suma de cada columna tras aplicar esta función será de un millón.

```
> keep <- rowSums(cpm(y)>1) >= 3
```

El vector keep contiene los genes seleccionados, pudiendo eliminar el resto del objeto que estamos estudiando.

```
> y <- y[keep,]
> dim(y)
[1] 14848     9
```

Aquí observamos la reducción del número de filas de nuestra base de datos después del filtrado. Hemos realizado una depuración de nuestros datos como se haría en cualquier estudio estadístico, donde se analiza que hacer con aquellos datos que se identifican como outlier o que representan algún tipo de problema al estudio.

Normalización.

Posteriormente debemos normalizar nuestros datos. En RNA-seq la normalización se realiza para minimizar el ruido técnico introducido en los datos durante el proceso de secuenciación con el fin de volverlos comparables entre sí. Buscando dos objetivos, el primero es obtener todos nuestros datos en una misma escala para evitar falsos positivos, ya que una muestra o librería con mayor profundidad de secuenciación tiene más probabilidad de tener genes expresados diferencialmente respecto a otra, sin deberse estas diferencias en nuestros grupos. El segundo objetivo es eliminar las variaciones biológicas entre las muestras.

En este caso EdgeR utiliza la función calcNormFactors, la cual proporciona un conjunto de factores de normalización que minimizan el log-Fold-change(cambio en la proporción de lecturas) entre las muestras, según el total de lecturas de cada uno de ellos, para la mayoría de tránscritos. El cálculo de dichos factores de normalización se realiza por la media truncada de los valores M (trimmed mean of M values, TMM) entre cada par de muestras. Obtenemos factores de normalización cercanos a 1, lo cual nos indica que no existen grandes diferencias entre ellas.

Estimamos los factores de normalización que harán que las muestras sean comparables:

```
> y<- calcNormFactors(y)
> y$samples
     group lib.size norm.factors
C1       C 15374649    1.0343128
C2       C 11008262    1.0829204
C3       C 18083094    1.1577212
LPS1   LPS 12086248    0.8496729
```

LPS2 LPS 11677884 0.8297256
LPS3 LPS 13064018 0.9555541
L+L1 L+L 8494744 1.0429317
L+L2 L+L 13929602 0.9783697
L+L3 L+L 14544456 1.1218819

Posteriormente, utilizamos el gráfico MDS (Multidimensional scaling plot of distances) donde se muestra la relación entre todas las muestras. El análisis que realiza este gráfico es similar a un PCA, es decir, al análisis de componentes principales.

```
> plotMDS(y)
```

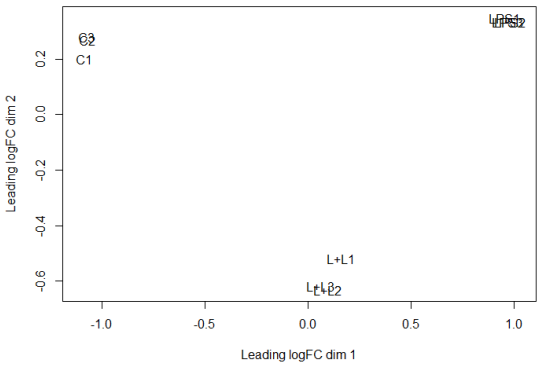


Figura24. Plot MDS.

Podemos resaltar del gráfico la separación entre los tres grupos. Las muestras estimuladas con LPS, se encuentran en valores positivos del eje X, y las muestras del grupo control en valores negativos, esto supone la mayor diferencia en el gráfico. Observando este gráfico podemos intuir la existencia de genes diferencialmente expresados debido a la clara separación entre los grupos.

Una obtenida la distribución de los genes con la función plotMDS realizamos un dendograma, para conocer mejor nuestras muestras.

```
> normalized.counts=cpm(y)
> transposed=t(normalized.counts)
> distance=dist(transposed)
> clusters=hclust(distance)
> plot(clusters)
```

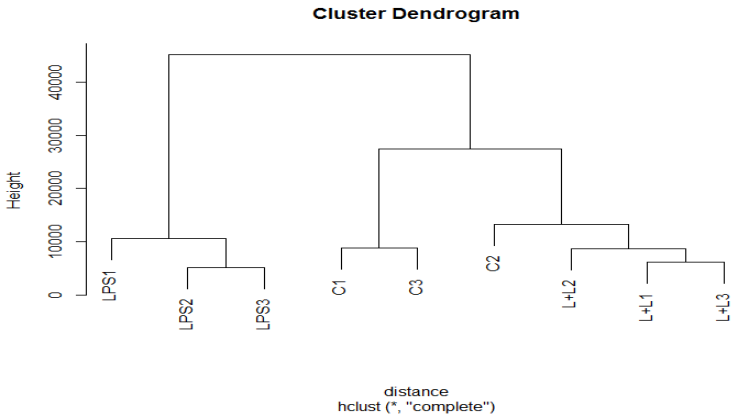


Figura25. Dendograma.

De este gráfico dedujimos que efectivamente disponemos de diferentes grupos. Aunque nos genera una pequeña intranquilidad porque el plot anterior los separaba de una manera tajante.

Estimación de la dispersión.

EdgeR se basa en que el supuesto de los datos de conteo son modelados mediante una Binomial Negativa, tenemos que estimar el parámetro de dispersión antes de realizar el test para determinar los tránscritos diferencialmente expresados. Utilizando la función estimateCommonDisp estimamos la dispersión común para todos los tránscritos , es decir, este parámetro nos proporciona una idea general de la variabilidad de nuestros datos. Esta función calcula el coeficiente de variación biológica (BCV), que constituye la raíz cuadrada de dicho parámetro de dispersión y representa el coeficiente de variación entre réplicas de la misma condición. Finalmente con la función estimateTagwiseDisp estimamos la dispersión para cada uno de los tránscritos.

En el código R al calcular la dispersión común añadiremos como atributo en la función verbose = TRUE para que muestre los valores del coeficiente de variación biológico (BCV) y la dispersión común.

```
> y <- estimateCommonDisp(y,verbose=TRUE)
Disp = 0.18016 , BCV = 0.3844
> y <- estimateTagwiseDisp(y)
```

Ya están estimadas las dispersiones, y como se puede ver en la salida obtenida al estimar la dispersión común, esta es de 0,18 y el coeficiente de variación biológica (BCV) de un 0,38.

Representamos con un gráfico BCV las dispersiones estimadas, y así podemos comprobar si la dispersión común representa realmente la dispersión existente entre los genes.

```
> plotBCV(y)
```

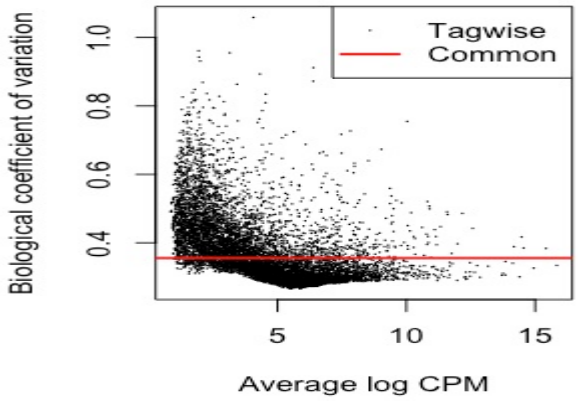


Figura26. Plot BCV.

2.4.2 Comparaciones entre grupos.

Obtenidos unos datos que están bajo la clase DGEList y normalizados, podemos realizar comparaciones entre sí y se han ajustado una distribución teórica, la distribución binomial negativa. Procediendo a identificar los genes diferencialmente expresados (DE).

Con la función exactTest , somos capaces de realizar pruebas estadísticas por parejas para la expresión diferencial entre dos grupos.
exactTest(object, pair=1:2, dispersion = “auto”, big.count=900, prior.count=0.125)

Volveremos introducir nuestro objeto, de clase DGEList , ya normalizado y con las estimaciones calculadas. Con esta función obtenemos tres elementos:

1. Bajo el dominio ...$table habrá un data.frame donde se pueden observar por columnas el log 2 fold – change (logFC) , los logaritmos en base 2 de la media de los cpm (conteos por millon) de cada gen (logCPM) y el p-valor bilateral (PValue)
2. Bajo ...$comparison se puede ver un vector de caracteres con las condiciones de los dos grupos que se comparan, nosotros diponemos de tres grupos y vamos a hacer comparaciones dos a dos.
3. Bajo...$genes un data.frame donde aparecerá la información que se le haya dado, en caso de haberlo hecho, de los genes al crear el objeto de la clase DGEList.

Posteriormente, aplicamos la función topTags(), y de ella obtenemos varios elementos, una tabla, similar a la obtenida con la función anterior, con una columna en la que encontraremos el p-valor ajustado según el método elegido, que en nuestro caso será el método de B-H, y en caso de que al crear el objeto DGEList se haya incluido alguna información de los genes, también nos aparecerá una columna con dicha información. Es para lo único que serviría incluir información complementaria de los genes, para que cuando se obtenga esta tabla aparezca. topTags(object, n=10, adjust.method="BH", sort.by="PValue")

Esta función nos permite ordenar los resultados por columna, por defecto lo ordena según la columna PValue. Esta función calcula los valores para todos los genes aunque solo nos muestra los 10 primeros con un p-value menor, aunque modificando el atributo n de la función podemos cambiarlo. También es capaz de proporcionarnos vectores donde se muestra el método elegido para el cálculo del p-valor ajustado, el tipo de test que se ha realizado y los nombres de los dos grupos comparados.

Una vez descrito, escribo el código utilizando la función exactTest(·):

```
> et <- exactTest(y)
> et
An object of class "DGEExact"
$table
      logFC   logCPM     PValue
3  0.3000989 4.378591 0.38305848
4 -0.4856810 3.718754 0.19814221
5 -0.2897353 2.254014 0.60116674
6  0.1974462 9.865107 0.49188673
8  0.5715941 5.584735 0.08729471
14843 more rows ...

$comparison
[1] "C"   "L+L"

$genes
NULL
```

Podemos observar a continuación la comparación de otros grupos.

```
> et_2 <- exactTest(y, pair=2:3)
> et_2
An object of class "DGEExact"
$table
       logFC   logCPM     PValue
3 -0.20166677 4.378591 0.56227064
4 -0.68085346 3.718754 0.07957936
5  0.08866152 2.254014 0.87594818
6  0.17684168 9.865107 0.53832802
8 -0.49568164 5.584735 0.13871433
```

14843 more rows ...

$comparison
[1] "L+L" "LPS"

$genes
NULL

Utilizamos topTags(-), para obtener también las razones de falsos descubrimientos (FDRs) y ordenar la tabla según la columna“PValue”.

```
> top <- topTags(et)
> top
Comparison of groups:  L+L-C
       logFC    logCPM       PValue          FDR
1239  5.843041  8.111915 1.079971e-64 1.603541e-60
4342  6.553471  7.195795 1.200512e-60 8.912605e-57
3567  9.395037  6.592454 2.475265e-57 1.225091e-53
42506 5.895442  9.665337 1.035169e-44 3.842549e-41
5652  5.239523 10.767988 5.373084e-43 1.595591e-39
43448 4.160103 11.037550 1.352594e-36 3.347220e-33
42812 4.163389  9.440249 7.370047e-36 1.563292e-32
7224  8.355671  5.926937 1.228031e-34 2.279226e-31
5997  5.267735  5.557799 1.141245e-33 1.882801e-30
683   5.818394  4.971068 1.682960e-32 2.498859e-29

> dim(y)
[1] 14848     9
> top2 <- topTags(et,n=14848)
> table(top2$table$FDR <0.05)
FALSE  TRUE
13712  1136
> table(top2$table$FDR <0.05)/nrow(top2$table)
    FALSE       TRUE
0.92349138 0.07650862
```

Al estar ordenados según la columna “PValue”, los primeros elementos serán los que menor p-valor tengan, que coincide, lógicamente, con los menores FDRs. Aquí solo mostramos diez genes existentes del total. Para observar la proporción de DGE utilizamos nuevos gráficos. El histograma nos representa todos los genes con top2.

```
> hist(top2$table$FDR, breaks=100, main="Histograma de FDR")
> abline(v=0.05, col="red", lwd=3)
```

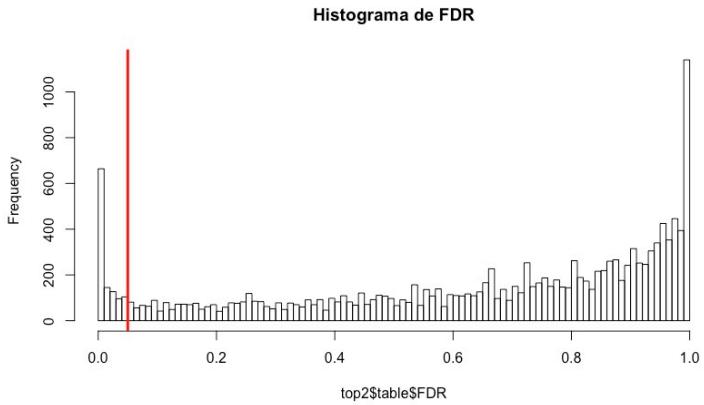


Figura27. Histograma.

En esta descripción establecimos el corte para p-valor ≤ 0.05 para identificar los genes que están diferencialmente expresados. Y podemos ver las diferencias notorias entre grupos.

Posteriormente realizamos el plotSmear; este gráfico representa el log2fold − change frente a la media del logCPM.

```
> de <- decideTestsDGE(et)
> summary(de)
   [,1]
-1   436
0  13712
1    700
> detags <- rownames(y)[as.logical(de)]
> plotSmear(et, de.tags=detags, main="plotSmear")
> abline(h=c(-1,1), col="blue")
```

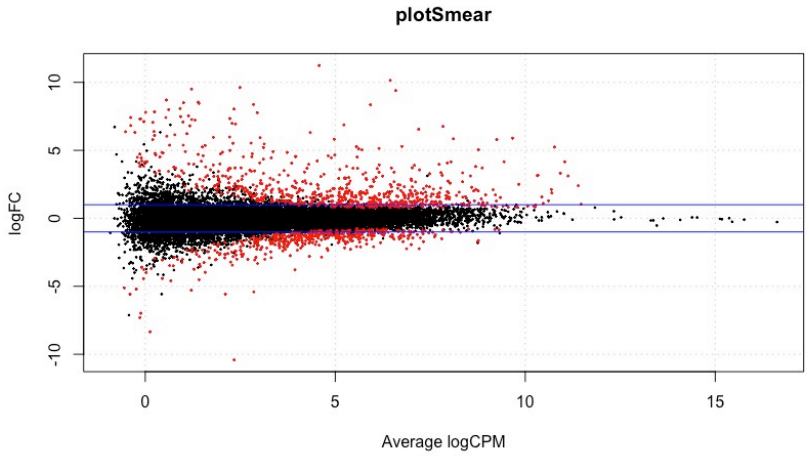


Figura28. plotSmear.

Primero defino el objeto “de”, donde están los genes clasificados atendiendo a los siguientes criterios, primero que el p-valor ajustado, es decir el FDR sea menor de 0.05, y luego que el logFC sea menor que -1 o mayor que 1. Si

observamos la tabla creada anteriormente, ya intuíamos el número de genes que tenían un p-valor ajustado menor que 0.05, lo que se realiza ahora es la separación del resto según el logFC. Tras este proceso se guarda en cada gen si pertenece a la fila central del summary o no, para poder representarlos con distintos colores en un gráfico. Este gráfico representará el log2fold − change frente a la media del logCPM, como vemos en el gráfico. En el gráfico aparecen en color rojo los genes DE, y los demás en negro. Podemos ver que entre los grupos C y L+L existen muchos de los genes diferencialmente expresados. También hemos realizado el mismo proceso sacando el resto de comparaciones y posteriormente genero un tabla de todos ellos. En este caso guardamos nuestras tablas resultantes y utilizamos Excel para varios procesamientos. En el apartado de resultados definiremos mejor éstos, debido a que generamos una gran cantidad de información y variaciones de análisis.

2.5 Anotación Funcional

Para realizar la anotación funcional utilizamos un análisis de enriquecimiento de conjuntos de genes (GSEA), este método identifica clases de genes o proteínas que están representados en exceso en un gran conjunto de estos, y pueden ser asociados con diferentes fenotipos. El método utiliza enfoques estadísticos para identificar grupos de genes significativamente enriquecidos. Las tecnologías de transcriptómica y los resultados de proteómica a menudo identifican miles de genes que se utilizan para el análisis [12].Con ello pretendemos conseguir un perfil funcional de nuestros conjunto de genes, para comprender mejor los procesos biológicos subyacentes.

Para este proceso utilizamos GOrilla por la sencillez y la rapidez de obtención de datos. Es una herramienta web para identificar y visualizar términos GO enriquecidos en listas clasificadas de genes. Se puede ejecutar en uno de los dos modos: Buscar términos GO enriquecidos que aparezcan densamente en la parte superior de una lista clasificada de genes o buscar términos GO enriquecidos en una lista objetivo de genes en comparación con una lista de genes de fondo.

En nuestro caso realizamos un análisis con GOrilla para cada uno de los listados independientemente de las comparaciones para ver que posibles procesos están expresados, de ahí unificamos los datos y obtenemos un listado de genes, para dos objetivos diferenciar y posiblemente obtener un patrón de genes que definan la TE, y también obtener ese listado de imnocheckpoint.

También utilizamos Enrichr, que es una herramienta web de análisis de enriquecimiento de conjuntos de genes específica de mamíferos. Contiene bibliotecas de fondo para la regulación de la transcripción, vías e interacciones de proteínas, ontologías que incluyen GO y las ontologías de fenotipo humano, etc. [13]. Este programa será utilizado para asociar las funciones biológicas de nuestro listado de genes de posibles inmunocheckpoint.

3. Resultados y Discusión.

Una vez descrito todo desarrollo del procesamiento del RNA-seq hasta llegar a la expresión diferencial nos centramos en el análisis de los resultados. En este proyecto se plantearon varios objetivos biológicos: Determinar el patrón de expresión de la TE en un modelo con monocitos humanos, y buscar nuevos inmunocheckpointsmediante datos de RNA-seq de nuestras diferentes condiciones experimentales.

Para alcanzar los objetivos, analizamos nuestros tres grupos que disponen de tres réplicas biológicas, donde su diseño experimental fue crear un modelo de TE que contenía tres grupos:

1. Control; Cultivo (anteriormente descrito) sin estímulo.
2. LPS; Cultivo con un estimulo de 5ng/ml de LPS durante 2 horas.
3. TE: Cultivo con un estimulo inicial de 5ng/ml de LPS durante 8 horas (el cual genera la TE), 16 horas de descanso, finalmente un estímulo de LPS durante 2 horas.

Para realizar estas comparaciones hemos utilizado varias herramientas descritas anteriormente como Bioconductor de R, Microsoft Excel o el programa web GOrilla.

La primera observación que queremos destacar es el comportamiento de las muestras en su conjunto. Para ello utilizamos la herramienta EdgeR, después de su normalización muestro los gráficos de plotMDS y dendograma.

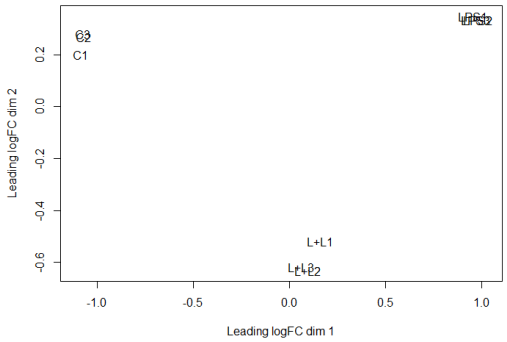


Figura29. Plot MDS.

En la figura 27, mostramos un gráfico de tipo Multi-Dimensional Scaling (MDS), este se basa en la distribución de las muestras teniendo en cuenta la varianza en su expresión. Como podemos observar las réplicas de cada grupo se mantienen agrupadas, y nuestros datos positivos en el eje X son de los dos grupos que se estimularon con LPS. La información obtenida de este gráfico nos indicaría el efecto que genera cualquier estímulo de LPS a nuestros monocitos/macrófagos e indicaría la activación del SII y también la diferencia entre un modelo de inflamación (LPS) y el modelo de TE (L+L) [6,14]

También queremos observar el comportamiento de nuestras muestras con un dendograma. Este es un gráfico el cual utilizando algoritmos de clustering

podemos ver las relaciones entre muestras y grupos por medio de distancias y similitudes entre ellas.

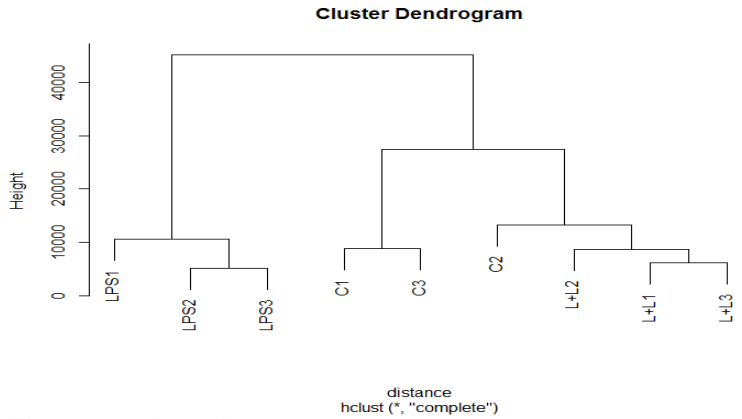


Figura30. Dendograma.

En la figura 28, podemos observar que existen dos ramas diferenciales una que agrupa el grupo de LPS y otra de estas que aúna los controles y L+L. Aunque podemos ver el nivel de similitud de cada grupo prácticamente diferenciado, observamos que una de nuestras réplicas Control se agrupa con el grupo L+L (nuestros tolerantes), pese a ello no se obtiene ninguna asociación atípica. Ya que según la bibliografía los patrones de expresión del grupo Control debe ser más parecido al grupo L+L que al LPS.[14, 15]

En la figura 29, mostramos los MAplots donde se observan los cambios de expresión entre cualquier par de muestras (en este caso lo determinamos entre grupos). Estos comparan el log de la Abundancia (cuentas de lecturas mapeadas) en el eje X contra el log *Fold-change*(diferencia de expresión) en el eje Y.

CvsLL CvsLPS LLvsLPS

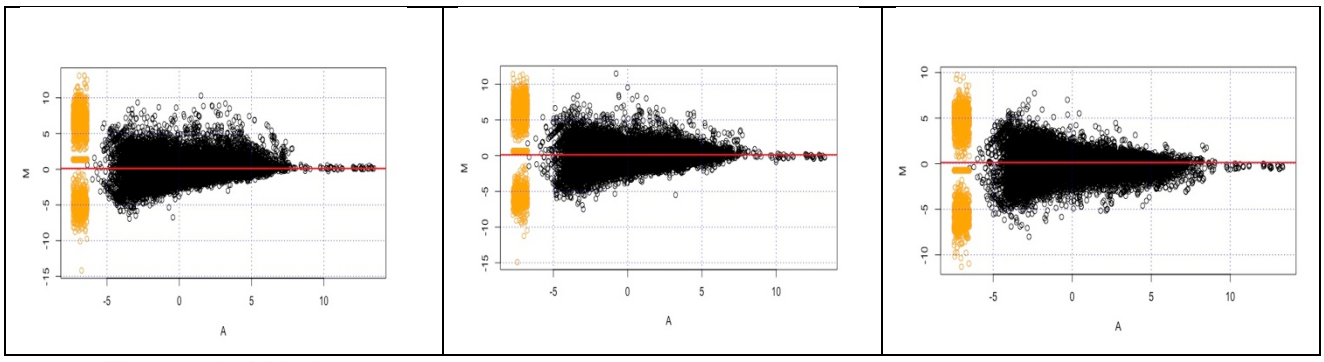


Figura31. MAplots.

Podemos observar en los MAplots que el grueso de los genes parece tener log Fold-changes en positivo aunque mínimo, lo cual nos informa que los genes se expresan más en los grupos estimulados con LPS y L+L comparado con el grupo control, esta es una conclusión muy lógica biológicamente. También observamos como en la representación de los grupos L+L vs LPS, existe poca

variación entre ellos pero si se puede decir que hay más genes expresados en el grupo LPS.

Posteriormente, utilizamos los valores normalizados de todos los genes para poder visualizar bien las diferencias entre grupos, lo que se realizo fue seleccionarlos, ordenarlos (por un mayor número de lecturas) y trazar con ellos un mapa de calor de su expresión usando heatmaps aquí observamos de una manera más visual las distancias entre muestras y los patrones de expresión en cada una. El tipo de selección para realizar este heatmap ha sido por un mayor número de lecturas normalizadas en el grupo LPS.

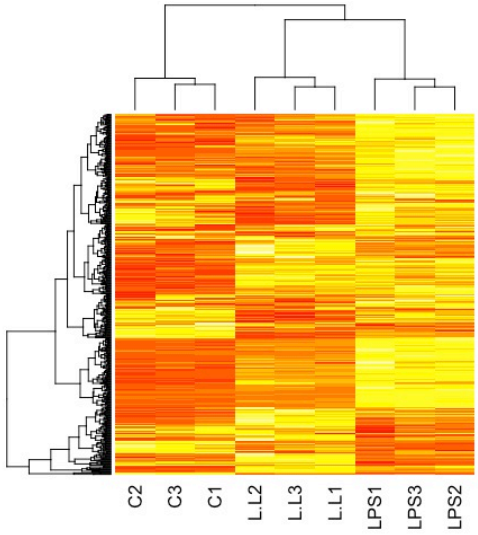


Figura 32. Heatmap de los 500 genes definidos por mayor número de lecturas del grupo LPS.

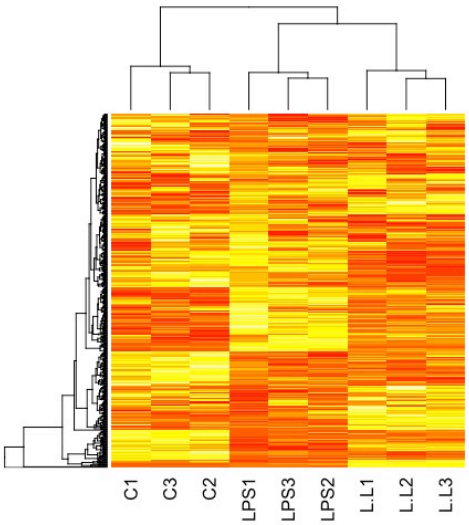


Figura 33. Heatmap definido por un número medio de lecturas del grupo LPS.

Podemos ver en estos heatmap como se asocian los diferentes genes creando un mapa de expresión definido de cada grupo, y si nos enfocamos en los dendogramas que generan estos, se definen mucho mejor los grupos con respecto al dendograma anterior esto es debido a que este se realizo con la

totalidad de genes expresados y los correspondientes de los heatmaps anteriores con 500 de genes representativos. Con estas gráficas estamos corroborando la activación de los monocitos/macrófagos en respuesta al LPS y las diferencias existentes frente al TE, y su posible asociación a los diferentes estados de estos en enfermedades estudiadas como la sepsis o fibrosis quística [4,5,14,16]

De hecho, para comprobar la variabilidad genética de cada muestra, escogimos los genes con menor número de lecturas y podemos ver en este nuevo heatmap como desaparecen esos patrones definidos. Con ello demostramos que el efecto de una mayor o menor expresión es determinado por el LPS y en los genes con un número bajo de lecturas no existen diferencias de patrones de expresión entre nuestros grupos.

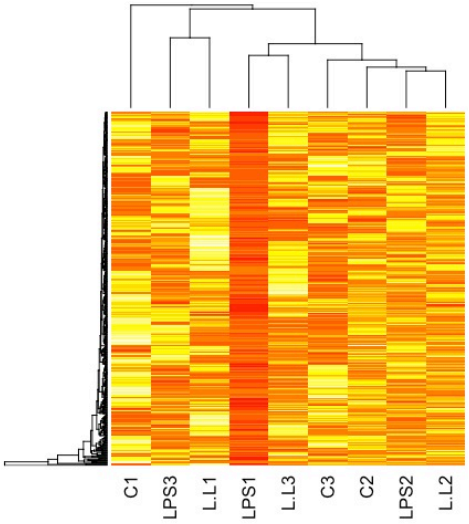


Figura 34. Heatmap definido por un número bajo de lecturas del grupo LPS

Con estos resultados de comportamiento de nuestras muestras y de sus agrupaciones, podemos concluir que tanto el diseño experimental como el experimento biológico en sí se ha realizado correctamente. Además podemos seleccionar a continuación un determinado número de genes expresados diferencialmente para ver su importancia y asociarlos a funciones biológicas.

Posteriormente de acuerdo con el punto 2.4 del desarrollo obtenemos un listado de las diferencias entre genes de nuestros grupos. Obteniendo con ello:

1. Genes que se expresan en monocitos de manera diferencial con diferentes estímulos de LPS.
2. Genes específicos de cada grupo para definir un listado de posibles inmunocheckpoint.

Para tratar de responder al punto numero 1, muestro las tablas de expresión diferencial obtenida de la función exacTest() de EdgeR. En estas muestro las diferencias entre los grupos.

1.Comparación CvsLL

```
> et
An object of class "DGEExact"
$table
                  logFC   logCPM    PValue
ENSG00000121410  0.11386244 4.511140 1.0000000
ENSG00000268895  0.05883717 4.462670 1.0000000
ENSG00000175899  0.11228725 8.491402 0.6124715
ENSG00000245105 -0.38562354 3.991767 1.0000000
ENSG00000128274  0.95448779 4.547620 0.4533589
14694 more rows ...

$comparison
[1] "C"   "L+L"

$genes
NULL
```

2.Comparación LLvsLPS

```
> et2
An object of class "DGEExact"
$table
                  logFC   logCPM     PValue
ENSG00000121410  0.04035168 4.511140 1.00000000
ENSG00000268895 -0.04785568 4.462670 1.00000000
ENSG00000175899  0.53032393 8.491402 0.00232808
ENSG00000245105 -0.10640284 3.991767 1.00000000
ENSG00000128274 -0.63336958 4.547620 0.72671840
14694 more rows ...

$comparison
[1] "L+L" "LPS"

$genes
NULL
```

3.Comparación CvsLPS

```
> et3
An object of class "DGEExact"
$table
                  logFC   logCPM       PValue
ENSG00000121410  0.15539697 4.472136 1.0000000000
ENSG00000268895  0.01209881 4.421647 1.0000000000
ENSG00000175899  0.64396334 8.531021 0.0002568673
ENSG00000245105 -0.49101990 3.963687 1.0000000000
ENSG00000128274  0.32144986 4.394481 1.0000000000
14901 more rows ...

$comparison
```

[1] "C"   "LPS"

$genes
NULL

Posteriormente de cada tabla utilizamos la función topTags(), para obtener también las razones de falsos descubrimientos (FDRs) y ordenar la tabla según la columna “PValue”. Al estar ordenados según la columna “PValue”, los primeros elementos serán los que menor p-valor tengan, que coincide, lógicamente, con los menores FDRs. Aquí solo mostramos el listado de los diez genes con mayor diferencia significativa.

```
> topTags(et)
Comparison of groups:  L+L-C
                 logFC   logCPM       PValue          FDR
ENSG00000102962          3.262336 11.057438 2.742610e-274 4.031362e-270
ENSG00000184371          3.074189 10.669353 5.411796e-144 3.977399e-140
ENSG00000169429          2.434700 11.338617 7.081173e-130 3.469539e-126
ENSG00000122641          5.731221  9.473413 5.504883e-120 2.022907e-116
ENSG00000277632          4.149276 10.855827 1.603651e-111 4.714414e-108
ENSG00000136689          2.122257 10.617770  3.521372e-93  8.626773e-90
ENSG00000118503          3.164042 10.481175  4.842877e-93  1.016935e-89
ENSG00000112096          3.140750 10.633151  1.603258e-83  2.945786e-80
ENSG00000125538          5.106087 10.292441  5.221829e-79  8.528407e-76
ENSG00000113657          3.907458  8.464539  4.313233e-57  6.340022e-54

> topTags(et2)
Comparison of groups:  LPS-L+L
                 logFC   logCPM       PValue          FDR
ENSG00000151726          2.543508  10.260183  1.784681e-145  2.623302e-
141
ENSG00000185215          2.465553  10.420636  2.257117e-141  1.658868e-
137
ENSG00000102962         -1.753559  11.057438  1.055194e-123  5.170099e-
120
ENSG00000277632          1.717475  10.855827  3.832600e-113  1.408385e-
109
ENSG00000125538          1.910330 10.292441  1.890303e-99  5.557113e-96
ENSG00000275302          2.353447  9.365019  2.930827e-78  7.180036e-75
ENSG00000112096          1.490979 10.633151  9.652971e-78  2.026986e-74
ENSG00000276085          2.562546  9.106391  9.316425e-75  1.711777e-71
ENSG00000162645          4.055409  8.234293  5.176457e-69  8.454305e-66
ENSG00000067082          1.646518 10.143485  1.088809e-63  1.600440e-60

> topTags(et3)
Comparison of groups:  LPS-C
                 logFC   logCPM       PValue          FDR
ENSG00000277632          5.868005 11.025932  0.000000e+00  0.000000e+00
ENSG00000125538          7.017518 10.500681 3.438367e-302 2.562615e-298
ENSG00000112096          4.633048 10.756007 4.037315e-284 2.006007e-280
```

ENSG00000169429	3.159974	11.260107	4.246100e-258	1.582309e-254
ENSG00000151726	4.111641	10.589484	1.934925e-232	5.768400e-229
ENSG00000118503	4.203717	10.480851	2.561228e-222	6.362945e-219
ENSG00000026508	3.008746	10.755966	7.773321e-183	1.655273e-179
ENSG00000275302	7.945985	9.650754	1.130502e-182	2.106409e-179
ENSG00000276085	6.606850	9.423798	3.316703e-150	5.493198e-147
ENSG00000100906	4.289845	9.794724	5.707933e-148	8.508245e-145

Con este análisis anterior de los 10 genes diferencialmente expresados entre grupos realizamos tres heatmaps. Con ello demostramos de nuevo las diferencias con un análisis de diferencia de expresión entre grupos.

CvsLL **CvsLPS** **LLvsLPS**

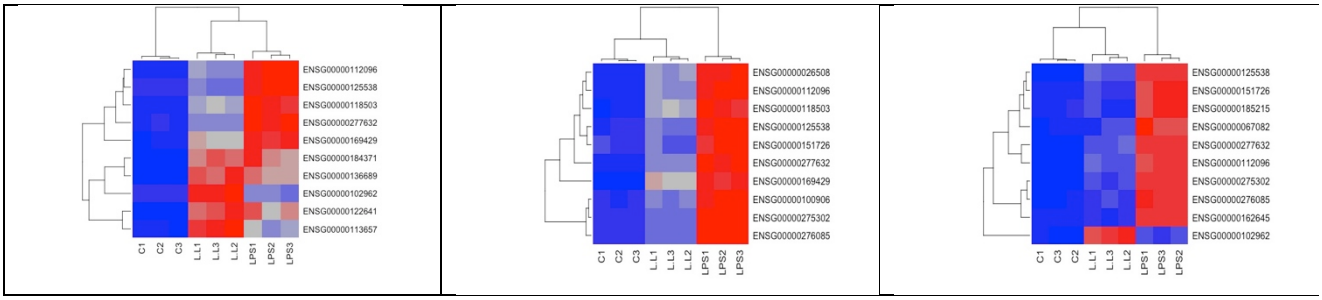


Figura 35. Heatmps del los listados de genes obtenidos de topTags()

Estos genes obtenidos nos dan pistas de un posible listado que caracterice la tolerancia a endotoxinas, de hecho el heatmap de la comparativa de genes Control vs LPS, nos reporta tres patrones muy diferenciados. Las siguientes tablas de genes obtenida de la comparativa obtenida de topTest de los grupos CvsLPSvsLL puede ser una buen inicio para determinar diferencias entre los tres grupos, y en un futuro buscarlos en pacientes con TE.

Entry name	Protein names	Gene names
CCL3_HUMAN	C-C motif chemokine 3 (G0/G1 switch regulatory protein 19-1) (Macrophage inflammatory protein 1-alpha) (MIP-1-alpha) (PAT 464.1) (SIS-beta) (Small-inducible cytokine A3) (Tonsillar lymphocyte LD78 alpha protein) [Cleaved into: MIP-1-alpha(4-69) (LD78-alpha(4-69))]	CCL3 G0S19-1 MIP1A SCYA3
IL1B_HUMAN	Interleukin-1 beta (IL-1 beta) (Catabolin)	IL1B IL1F2
SODM_HUMAN	Superoxide dismutase [Mn], mitochondrial (EC 1.15.1.1)	SOD2
IL8_HUMAN	Interleukin-8 (IL-8) (C-X-C motif chemokine 8) (Chemokine (C-X-C motif) ligand 8) (Emoctakin) (Granulocyte chemotactic protein 1) (GCP-1) (Monocyte-derived neutrophil chemotactic factor) (MDNCF) (Monocyte-derived neutrophil-activating peptide) (MONAP) (Neutrophil-activating protein 1) (NAP-1) (Protein 3-10C) (T-cell chemotactic factor) [Cleaved into: MDNCF-a (GCP/IL-8 protein IV) (IL8/NAP1 form I); Interleukin-8 ((Ala-IL-8)77) (GCP/IL-8 protein II) (IL-8(1-77)) (IL8/NAP1 form II) (MDNCF-b); IL-8(5-77); IL-8(6-77) ((Ser-IL-8)72) (GCP/IL-8 protein I) (IL8/NAP1 form III) (Lymphocyte-derived neutrophil-activating factor) (LYNAP) (MDNCF-c) (Neutrophil-activating factor) (NAF); IL-8(7-77) (GCP/IL-8 protein V) (IL8/NAP1 form IV); IL-8(8-77) (GCP/IL-8 protein VI) (IL8/NAP1 form V); IL-8(9-77) (GCP/IL-8 protein III) (IL8/NAP1 form VI)]	CXCL8 IL8
ACSL1_HUMAN	Long-chain-fatty-acid--CoA ligase 1 (EC 6.2.1.3) (Acyl-CoA synthetase 1) (ACS1) (Long-chain acyl-CoA synthetase 1) (LACS 1) (Long-chain acyl-CoA synthetase 2) (LACS 2) (Long-chain fatty acid-CoA ligase 2) (Palmitoyl-CoA ligase 1) (Palmitoyl-CoA ligase 2)	ACSL1 FACL1 FACL2 LACS LACS1 LACS2
TNAP3_HUMAN	Tumor necrosis factor alpha-induced protein 3 (TNF alpha-induced protein 3) (EC 2.3.2.-) (EC 3.4.19.12) (OTU domain-containing protein 7C) (Putative DNA-binding protein A20) (Zinc finger protein A20) [Cleaved into: A20p50; A20p37]	TNFAIP3 OTUD7C
CD44_HUMAN	CD44 antigen (CDw44) (Epican) (Extracellular matrix receptor III) (ECMR-III) (GP90 lymphocyte homing/adhesion receptor) (HUTCH-I) (Heparan sulfate proteoglycan) (Hermes antigen) (Hyaluronate receptor) (Phagocytic glycoprotein 1) (PGP-1) (Phagocytic glycoprotein I) (PGP-I) (CD antigen CD44)	CD44 LHR MDU2 MDU3 MIC4
CCL4_HUMAN	C-C motif chemokine 4 (G-26 T-lymphocyte-secreted protein) (HC21) (Lymphocyte activation gene 1 protein) (LAG-1) (MIP-1-beta(1-69)) (Macrophage inflammatory protein 1-beta) (MIP-1-beta) (PAT 744) (Protein H400)	CCL4 LAG1 MIP1B SCYA4

	(SIS-gamma) (Small-inducible cytokine A4) (T-cell activation protein 2) (ACT-2) [Cleaved into: MIP-1-beta(3-69)]	
CL3L1_HUMAN	C-C motif chemokine 3-like 1 (G0/G1 switch regulatory protein 19-2) (LD78-beta(1-70)) (PAT 464.2) (Small-inducible cytokine A3-like 1) (Tonsillar lymphocyte LD78 beta protein) [Cleaved into: LD78-beta(3-70); LD78-beta(5-70)]	CCL3L1 D17S1718 G0S19-2 SCYA3L1; CCL3L3
IKBA_HUMAN	NF-kappa-B inhibitor alpha (I-kappa-B-alpha) (IkB-alpha) (IkappaBalpha) (Major histocompatibility complex enhancer-binding protein MAD3)	NFKBIA IKBA MAD3 NFKBI

Tabla4. Tabla de genes obtenida de la comparativa obtenida de topTest de los grupos CvsLPS

Entry name	Protein names	Gene names
CCL22_HUMAN	C-C motif chemokine 22 (CC chemokine STCP-1) (MDC(1-69)) (Macrophage-derived chemokine) (Small-inducible cytokine A22) (Stimulated T-cell chemotactic protein 1) [Cleaved into: MDC(3-69); MDC(5-69); MDC(7-69)]	CCL22 MDC SCYA22 A-152E5.1
CSF1_HUMAN	Macrophage colony-stimulating factor 1 (CSF-1) (M-CSF) (MCSF) (Lanimostim) [Cleaved into: Processed macrophage colony-stimulating factor 1]	CSF1
IL8_HUMAN	Interleukin-8 (IL-8) (C-X-C motif chemokine 8) (Chemokine (C-X-C motif) ligand 8) (Emoctakin) (Granulocyte chemotactic protein 1) (GCP-1) (Monocyte-derived neutrophil chemotactic factor) (MDNCF) (Monocyte-derived neutrophil-activating peptide) (MONAP) (Neutrophil-activating protein 1) (NAP-1) (Protein 3-10C) (T-cell chemotactic factor) [Cleaved into: MDNCF-a (GCP/IL-8 protein IV) (IL8/NAP1 form I); Interleukin-8 ((Ala-IL-8)77) (GCP/IL-8 protein II) (IL-8(1-77)) (IL8/NAP1 form II) (MDNCF-b); IL-8(5-77); IL-8(6-77) ((Ser-IL-8)72) (GCP/IL-8 protein I) (IL8/NAP1 form III) (Lymphocyte-derived neutrophil-activating factor) (LYNAP) (MDNCF-c) (Neutrophil-activating factor) (NAF); IL-8(7-77) (GCP/IL-8 protein V) (IL8/NAP1 form IV); IL-8(8-77) (GCP/IL-8 protein VI) (IL8/NAP1 form V); IL-8(9-77) (GCP/IL-8 protein III) (IL8/NAP1 form VI)]	CXCL8 IL8
INHBA_HUMAN	Inhibin beta A chain (Activin beta-A chain) (Erythroid differentiation protein) (EDF)	INHBA
CCL3_HUMAN	C-C motif chemokine 3 (G0/G1 switch regulatory protein 19-1) (Macrophage inflammatory protein 1-alpha) (MIP-1-alpha) (PAT 464.1) (SIS-beta) (Small-inducible cytokine A3) (Tonsillar lymphocyte LD78 alpha protein) [Cleaved into: MIP-1-alpha(4-69) (LD78-alpha(4-69))]	CCL3 G0S19-1 MIP1A SCYA3
IL1RA_HUMAN	Interleukin-1 receptor antagonist protein (IL-1RN) (IL-1ra) (IRAP) (ICIL-1RA) (IL1 inhibitor) (Anakinra)	IL1RN IL1F3 IL1RA
TNAP3_HUMAN	Tumor necrosis factor alpha-induced protein 3 (TNF alpha-induced protein 3) (EC 2.3.2.-) (EC 3.4.19.12) (OTU domain-containing protein 7C) (Putative DNA-binding protein A20) (Zinc finger protein A20) [Cleaved into: A20p50; A20p37]	TNFAIP3 OTUD7C
SODM_HUMAN	Superoxide dismutase [Mn], mitochondrial (EC 1.15.1.1)	SOD2
IL1B_HUMAN	Interleukin-1 beta (IL-1 beta) (Catabolin)	IL1B IL1F2
DPYL3_HUMAN	Dihydropyrimidinase-related protein 3 (DRP-3) (Collapsin response mediator protein 4) (CRMP-4) (Unc-33-like phosphoprotein 1) (ULIP-1)	DPYSL3 CRMP4 DRP3 ULIP ULIP1

Tabla5. Tabla de genes obtenida de la comparativa obtenida de topTest de los grupos LPSvsLL

Entry name	Protein names	Gene names
ACSL1_HUMAN	Long-chain-fatty-acid--CoA ligase 1 (EC 6.2.1.3) (Acyl-CoA synthetase 1) (ACS1) (Long-chain acyl-CoA synthetase 1) (LACS 1) (Long-chain acyl-CoA synthetase 2) (LACS 2) (Long-chain fatty acid-CoA ligase 2) (Palmitoyl-CoA ligase 1) (Palmitoyl-CoA ligase 2)	ACSL1 FACL1 FACL2 LACS LACS1 LACS2
TNAP2_HUMAN	Tumor necrosis factor alpha-induced protein 2 (TNF alpha-induced protein 2) (Primary response gene B94 protein)	TNFAIP2
CCL22_HUMAN	C-C motif chemokine 22 (CC chemokine STCP-1) (MDC(1-69)) (Macrophage-derived chemokine) (Small-inducible cytokine A22) (Stimulated T-cell chemotactic protein 1) [Cleaved into: MDC(3-69); MDC(5-69); MDC(7-69)]	CCL22 MDC SCYA22 A-152E5.1
CCL3_HUMAN	C-C motif chemokine 3 (G0/G1 switch regulatory protein 19-1) (Macrophage inflammatory protein 1-alpha) (MIP-1-alpha) (PAT 464.1) (SIS-beta) (Small-inducible cytokine A3) (Tonsillar lymphocyte LD78 alpha protein) [Cleaved into: MIP-1-alpha(4-69) (LD78-alpha(4-69))]	CCL3 G0S19-1 MIP1A SCYA3
IL1B_HUMAN	Interleukin-1 beta (IL-1 beta) (Catabolin)	IL1B IL1F2
CCL4_HUMAN	C-C motif chemokine 4 (G-26 T-lymphocyte-secreted protein) (HC21) (Lymphocyte activation gene 1 protein) (LAG-1) (MIP-1-beta(1-69)) (Macrophage inflammatory protein 1-beta) (MIP-1-beta) (PAT 744) (Protein H400) (SIS-gamma) (Small-inducible cytokine A4) (T-cell activation protein 2) (ACT-2) [Cleaved into: MIP-1-beta(3-69)]	CCL4 LAG1 MIP1B SCYA4
SODM_HUMAN	Superoxide dismutase [Mn], mitochondrial (EC 1.15.1.1)	SOD2
CL3L1_HUMAN	C-C motif chemokine 3-like 1 (G0/G1 switch regulatory protein 19-2) (LD78-beta(1-70)) (PAT 464.2) (Small-inducible cytokine A3-like 1) (Tonsillar lymphocyte LD78 beta protein) [Cleaved into: LD78-beta(3-70); LD78-	CCL3L1 D17S1718

	beta(5-70)]	G0S19-2 SCYA3L1; CCL3L3
GBP2_HUMAN	Guanylate-binding protein 2 (EC 3.6.5.-) (GTP-binding protein 2) (GBP-2) (HuGBP-2) (Guanine nucleotide-binding protein 2) (Interferon-induced guanylate-binding protein 2)	GBP2
KLF6_HUMAN	Krueppel-like factor 6 (B-cell-derived protein 1) (Core promoter element-binding protein) (GC-rich sites-binding factor GBF) (Proto-oncogene BCD1) (Suppressor of tumorigenicity 12 protein) (Transcription factor Zf9)	KLF6 BCD1 COPEB CPBP ST12

Tabla6. Tabla de genes obtenida de la comparativa obtenida de topTest de los grupos CvsLPS

Como podemos observar entre las tablas es que nueve de estos genes coinciden en dos o tres de nuestras comparativas entre grupos, estos serían: CCL3, IL1B, SODM, IL8, ACSL1, TNAP, CCL4, CL3L1 y CCL22

Algunos de ellos se encuentran descritos en la TE como TNAP que se encuentra disminuida con dos estímulos de LPS frente a una sola exposición de la endotoxina[2]. O esos genes regulados negativamente en TE que codifican para citoquinas inflamatorias como IL8 o CCL4[14].

El resto de genes obtenidos CD44, IKBA, CSF1, IL1RA, DPYL3, GBP2 y KLF6, podrían ser más característicos de algunos de los estados que planteamos, como CD44 que se ve implicado en la señalización de TLR4 en sepsis [17]

Aunque del análisis anterior hemos podido reportar un listado de genes interesante para definir la TE. A continuación representamos los plotSmear estableciendo un corte estricto con los genes que tienen un p-valor ajustado de menos de 0.01 y una magnitud de cambio (log Fold-change) mayor a 1 (fuera de escala log2, mayor a 2). Y así podemos observar cuantos de estos genes serían los más representativos en nuestra comparativa por grupos.

CvsLL **CvsLPS** **LLvsLPS**

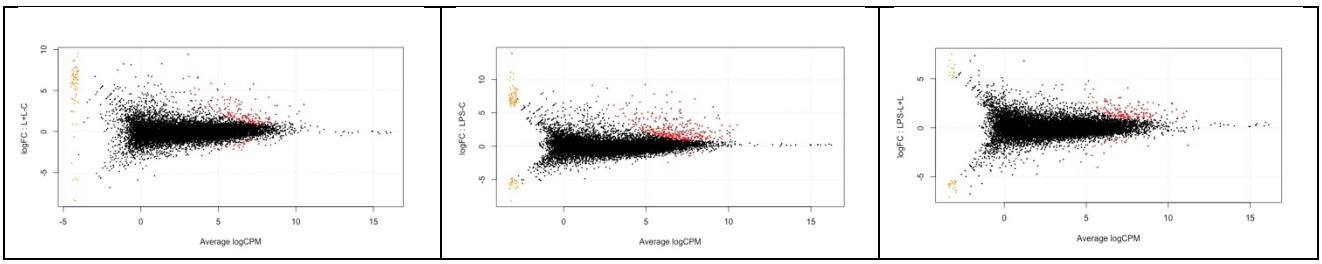


Figura 36. PlotSmear.

```
> length(deGenes1)
[1] 196
> length(deGenes2)
[1] 363
> length(deGenes3)
[1] 192
```

Podemos observar con el dato de número de genes que el grupo Control en comparación con el grupo LPS es donde se encuentran más genes DE.

Con todos estos resultados a parte de demostrar las diferencias entre nuestros diferentes estados (grupos) de patrones de expresión. Logramos proponer un listado de genes “referencia” para caracterizar la TE (adjunto listado de todos los genes expresados diferencialmente en los archivos del TFM). A posteriori de este trabajo, se pretende buscar estos patrones de expresión en diferentes tipos de pacientes que pudieran cumplir las características de estos estados como podrían ser pacientes sépticos [18,19] y observar si existe alguna correlación de estos con evolución del paciente. Con los listados obtenidos pretendemos comprobar el comportamiento de estos genes con q-PCR en nuestros modelos, y ver su existencia de nuevo en pacientes.

Con los listados de genes vamos a comprobar en que procesos biológicos están implicados estos, de ahí pretendemos obtener otros genes de interés que pudieran ser inmunocheckpoint.

En este caso voy a utilizar la herramienta web GOrilla. En nuestro caso realizamos un GOrilla, primero con el listado total de genes que hemos obtenido del RNA-seq. Este análisis es global, es decir queremos mostrar en que procesos, funciones y componentes están involucrados el listado total de genes obtenidos. Para posteriormente, desarrollar un análisis de los genes DE frente a nuestro listado global.

Mostramos los mapas del listado total de genes normalizados y que cumplían con el criterio explicado en el desarrollo, es decir los genes existentes con más de 1 cpm en los tres grupos que tenemos.

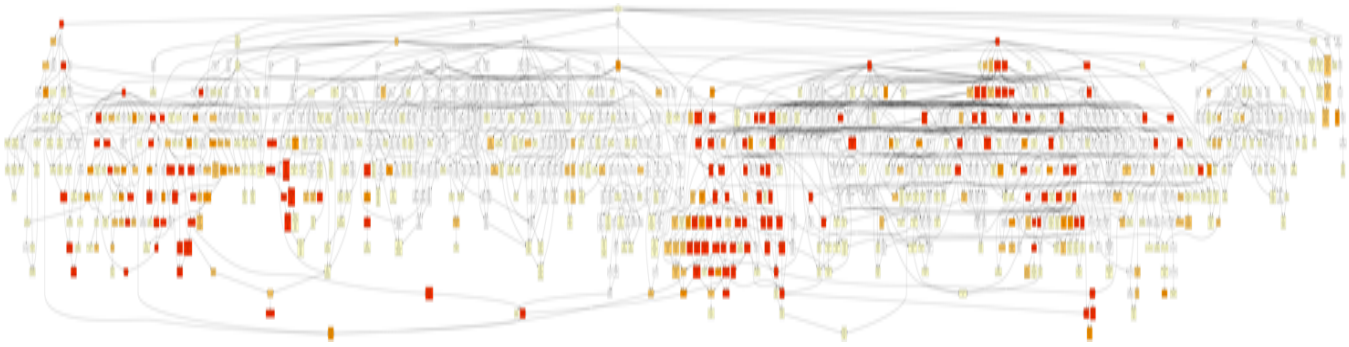


Figura37. Mapa de procesos obtenido de GOrilla con el listado de genes totales del RNA-seq.

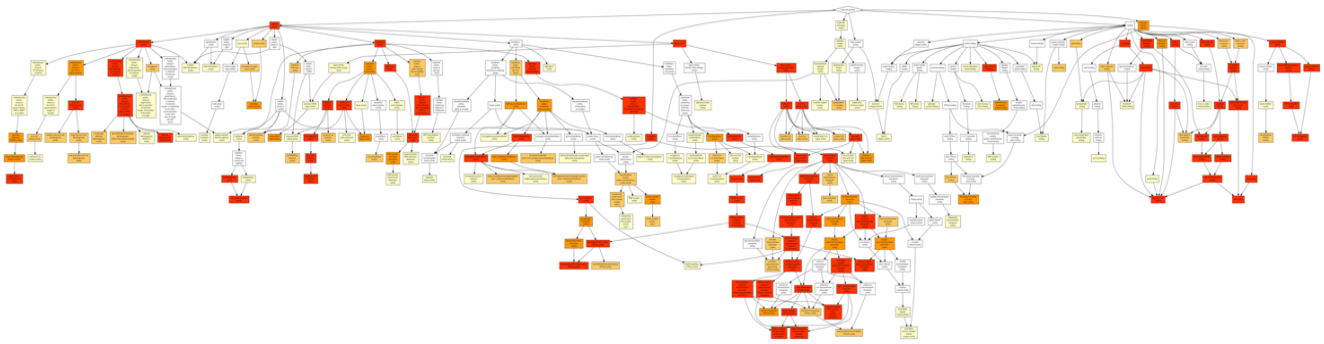


Figura38. Mapa de funciones obtenido de GOrilla con el listado de genes totales del RNA-seq

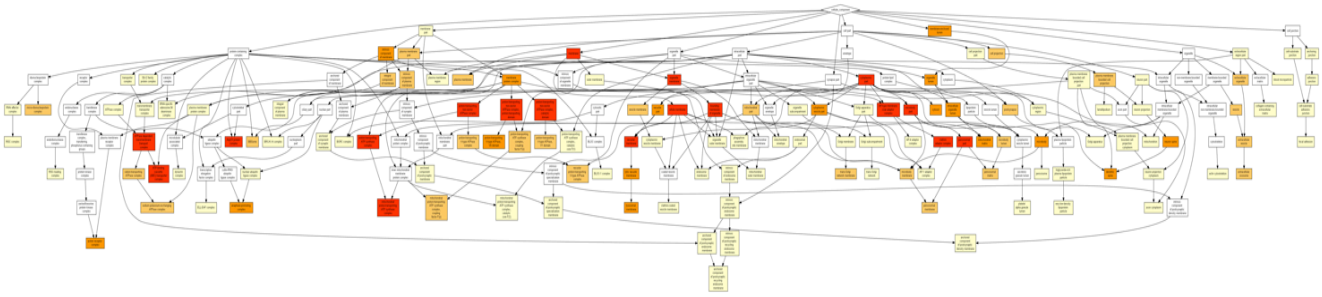


Figura39. Mapa de componentes obtenido de GOrilla con el listado de genes totales del RNA-seq.

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0099132	ATP hydrolysis coupled cation transmembrane transport	6.67E-41	9.09E-37	13.17 (10920,39,808,38)	[+] Show genes
GO:0044281	small molecule metabolic process	1.75E-36	1.19E-32	3.30 (10920,1147,392,136)	[+] Show genes
GO:0006631	fatty acid metabolic process	2.25E-36	1.02E-32	15.89 (10920,193,146,41)	[+] Show genes
GO:0019752	carboxylic acid metabolic process	2.2E-33	7.51E-30	4.52 (10920,576,382,91)	[+] Show genes
GO:0032787	monocarboxylic acid metabolic process	1.43E-32	3.9E-29	11.00 (10920,306,146,45)	[+] Show genes
GO:0006082	organic acid metabolic process	2.58E-31	5.86E-28	4.17 (10920,637,382,93)	[+] Show genes
GO:0043436	oxoacid metabolic process	6.98E-31	1.36E-27	4.17 (10920,631,382,92)	[+] Show genes
GO:0006637	acyl-CoA metabolic process	1.22E-25	2.08E-22	26.47 (10920,65,146,23)	[+] Show genes
GO:0035383	thioester metabolic process	1.22E-25	1.85E-22	26.47 (10920,65,146,23)	[+] Show genes
GO:0006629	lipid metabolic process	1.46E-23	1.99E-20	5.41 (10920,746,146,54)	[+] Show genes
GO:0034032	purine nucleoside bisphosphate metabolic process	2.1E-23	2.6E-20	20.17 (10920,89,146,24)	[+] Show genes
GO:0033865	nucleoside bisphosphate metabolic process	2.1E-23	2.38E-20	20.17 (10920,89,146,24)	[+] Show genes

Tabla7. Tabla de procesos obtenido de GOrilla con el listado de genes totales del RNA-seq.

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0043492	ATPase activity, coupled to movement of substances	2.94E-83	1.19E-79	12.49 (10920,87,814,81)	[+] Show genes
GO:0042626	ATPase activity, coupled to transmembrane movement of substances	4.95E-71	1E-67	12.60 (10920,74,808,69)	[+] Show genes
GO:0015405	P-P-bond-hydrolysis-driven transmembrane transporter activity	4.47E-67	6.05E-64	11.96 (10920,78,808,69)	[+] Show genes
GO:0015399	primary active transmembrane transporter activity	3.34E-66	3.39E-63	11.80 (10920,79,808,69)	[+] Show genes
GO:0022853	active ion transmembrane transporter activity	1.1E-47	8.89E-45	13.21 (10920,45,808,44)	[+] Show genes
GO:0042625	ATPase coupled ion transmembrane transporter activity	1.1E-47	7.41E-45	13.21 (10920,45,808,44)	[+] Show genes
GO:0019829	cation-transporting ATPase activity	1.49E-46	8.63E-44	13.21 (10920,44,808,43)	[+] Show genes
GO:0016887	ATPase activity	1.87E-36	9.49E-34	4.37 (10920,304,814,99)	[+] Show genes
GO:0042623	ATPase activity, coupled	1.48E-32	6.67E-30	4.60 (10920,245,814,84)	[+] Show genes
GO:0032559	adenyl ribonucleotide binding	1.56E-31	6.31E-29	4.67 (10920,1080,169,78)	[+] Show genes
GO:0030554	adenyl nucleotide binding	3.68E-31	1.36E-28	4.61 (10920,1093,169,78)	[+] Show genes
GO:0022804	active transmembrane transporter activity	5.9E-31	2E-28	29.56 (10920,196,49,26)	[+] Show genes
GO:0044769	ATPase activity, coupled to transmembrane movement of ions, rotational mechanism	1.38E-28	4.31E-26	13.05 (10920,28,807,27)	[+] Show genes

Tabla8. Tabla de funciones obtenido de GOrilla con el listado de genes totales del RNA-seq.

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0030119	AP-type membrane coat adaptor complex	4.87E-23	8.81E-20	14.11 (10920,39,496,25)	[+] Show genes
GO:0016469	proton-transporting two-sector ATPase complex	5.76E-21	5.2E-18	10.84 (10920,30,806,24)	[+] Show genes
GO:0033177	proton-transporting two-sector ATPase complex, proton-transporting domain	1.38E-15	8.3E-13	12.16 (10920,18,798,16)	[+] Show genes
GO:0030131	clathrin adaptor complex	2.88E-15	1.3E-12	14.80 (10920,24,492,16)	[+] Show genes
GO:0044444	cytoplasmic part	4.73E-14	1.71E-11	1.21 (10920,6517,988,713)	[+] Show genes
GO:0005885	Arp2/3 protein complex	2.06E-13	6.19E-11	16.67 (10920,11,655,11)	[+] Show genes
GO:0043190	ATP-binding cassette (ABC) transporter complex	1.25E-12	3.22E-10	253.95 (10920,5,43,5)	[+] Show genes
GO:0098588	bounding membrane of organelle	1.27E-12	2.87E-10	1.68 (10920,1318,1046,212)	[+] Show genes
GO:0033178	proton-transporting two-sector ATPase complex, catalytic domain	2.2E-12	4.42E-10	13.53 (10920,11,807,11)	[+] Show genes
GO:0016020	membrane	3.4E-12	6.14E-10	1.29 (10920,4494,887,470)	[+] Show genes
GO:0045259	proton-transporting ATP synthase complex	7.38E-12	1.21E-9	10.21 (10920,19,788,14)	[+] Show genes
GO:0005753	mitochondrial proton-transporting ATP synthase complex	7.38E-12	1.11E-9	10.21 (10920,19,788,14)	[+] Show genes
GO:0098533	ATPase dependent transmembrane transport complex	1.7E-11	2.37E-9	13.03 (10920,12,768,11)	[+] Show genes
GO:0098805	whole membrane	2.49E-11	3.22E-9	1.75 (10920,902,1046,151)	[+] Show genes
GO:0005774	vacuolar membrane	5.29E-11	6.37E-9	2.58 (10920,294,863,60)	[+] Show genes
GO:0044438	microbody part	2.21E-10	2.49E-8	13.15 (10920,76,142,13)	[+] Show genes
GO:0044439	peroxisomal part	2.21E-10	2.35E-8	13.15 (10920,76,142,13)	[+] Show genes
GO:0031090	organelle membrane	3.64E-10	3.66E-8	1.47 (10920,1942,932,244)	[+] Show genes
GO:0005765	lysosomal membrane	2.41E-9	2.3E-7	2.34 (10920,268,1046,60)	[+] Show genes
GO:0098796	membrane protein complex	2.43E-9	2.2E-7	1.81 (10920,674,1000,112)	[+] Show genes
GO:0098852	lytic vacuole membrane	2.78E-9	2.39E-7	2.33 (10920,269,1046,60)	[+] Show genes
GO:0044437	vacuolar part	3.15E-9	2.59E-7	2.22 (10920,409,819,68)	[+] Show genes

Tabla 9. Tabla de componentes obtenido de GOrilla con el listado de genes totales del RNA-seq.

Debido a la gran cantidad de datos en estos listados y a la inmensidad de nuestros mapas el análisis realizado se enfoca únicamente en procesos y funciones más expresados. Con ellos debemos hacer a posteriori un análisis más exhaustivo.

En el análisis de componentes aparecen más expresados los términos GO de membrana, orgánulos y mitocondrias. A nivel funciones nos aparece un gran listado donde vemos donde están más implicados nuestros genes, como: Actividad del transportador transmembrana, actividad de la ligasa, formación de enlaces carbono-azufre, etc. Y en los procesos estos están implicados en: metabolismo de ácidos grasos, metabolismo del ácido carboxílico, metabolismo del ácido monocarboxílico, proceso metabólico ácido orgánico, etc.

Debido a esa gran cantidad de datos obtendría únicamente una conclusión de estas gráficas y tablas es en los componentes, pudiendo observar que la mayoría de los genes se relacionan con componentes extracelulares, vacuolas o lisosomas este dato nos corrobora que nuestra extracción esta bien hecha. Ya que se realizó una separación inicial de estos componentes y con ello se extrajo el RNA. A nivel de procesos y funciones observando bien los datos no podríamos concretar bien sus asociaciones, aunque a nivel de procesos indican que nuestros genes intervienen de manera recurrente en metabolismo.

A continuación mostramos un ejemplo de cómo obtendríamos un listado de genes, para posteriormente estudiarlos uno a uno.

GO:0022857	transmembrane transporter activity	1.41E-19	1.5E-17	11.38 (10920,509,49,26)	[-] Hide genes ABCC2 - atp-binding cassette, sub-family c (cftr/mrp), member 2 ABCB8 - atp-binding cassette, sub-family b (mdr/tap), member 8 ABCA1 - atp-binding cassette, sub-family a (abc1), member 1 ABCG2 - atp-binding cassette, sub-family g (white), member 2 ABCA2 - atp-binding cassette, sub-family a (abc1), member 2 ABCA3 - atp-binding cassette, sub-family a (abc1), member 3 ABCB7 - atp-binding cassette, sub-family b (mdr/tap), member 7 ABCB4 - atp-binding cassette, sub-family b (mdr/tap), member 4 ABCB1 - atp-binding cassette, sub-family b (mdr/tap), member 1 ABCD4 - atp-binding cassette, sub-family d (ald), member 4 ABCD3 - atp-binding cassette, sub-family d (ald), member 3 ABCG1 - atp-binding cassette, sub-family g (white), member 1 ABCA6 - atp-binding cassette, sub-family a (abc1), member 6 ABCA5 - atp-binding cassette, sub-family a (abc1), member 5 ABCC10 - atp-binding cassette, sub-family c (cftr/mrp), member 10 ABCB9 - atp-binding cassette, sub-family b (mdr/tap), member 9 ABCC6 - atp-binding cassette, sub-family c (cftr/mrp), member 6 ABCC4 - atp-binding cassette, sub-family c (cftr/mrp), member 4 ABCC1 - atp-binding cassette, sub-family c (cftr/mrp), member 1 ABCC5 - atp-binding cassette, sub-family c (cftr/mrp), member 5 ABCB6 - atp-binding cassette, sub-family b (mdr/tap), member 6 ABCA9 - atp-binding cassette, sub-family a (abc1), member 9 ABCC3 - atp-binding cassette, sub-family c (cftr/mrp), member 3 ABCB10 - atp-binding cassette, sub-family b (mdr/tap), member 10 ABCA10 - atp-binding cassette, sub-family a (abc1), member 10 ABCA7 - atp-binding cassette, sub-family a (abc1), member 7

Tabla 10. Tabla de GOrilla con el listado de genes de un termino GO desplegado.

Como obtuvimos, un listado de genes DE y tenemos el listado global utilizado en el análisis de GOrilla anterior, vamos a realizar de nuevo estos análisis viendo la implicación de esos genes DE en la totalidad de los obtenidos en el RNA-seq. En este nos focalizamos en la importancia de los genes DE. De nuevo aunque no mostramos estos análisis la inmensidad de datos es abrumadora para sacar conclusiones exhaustivas en el tiempo disponible. Por ello decidimos realizar un nuevo análisis exclusivamente con los genes DE. Una vez obtenidas las tablas de los DE las exportamos con write.table(), de esta extraemos el listado y lo copiamos en la herramienta.

Las gráficas y tablas mostradas a continuación son para ver la implicación de esos genes obtenidos del análisis de expresión diferencial, en los procesos, funciones y componentes obtenidos de GOrilla.

GOrilla de Listado DE CvsL+L

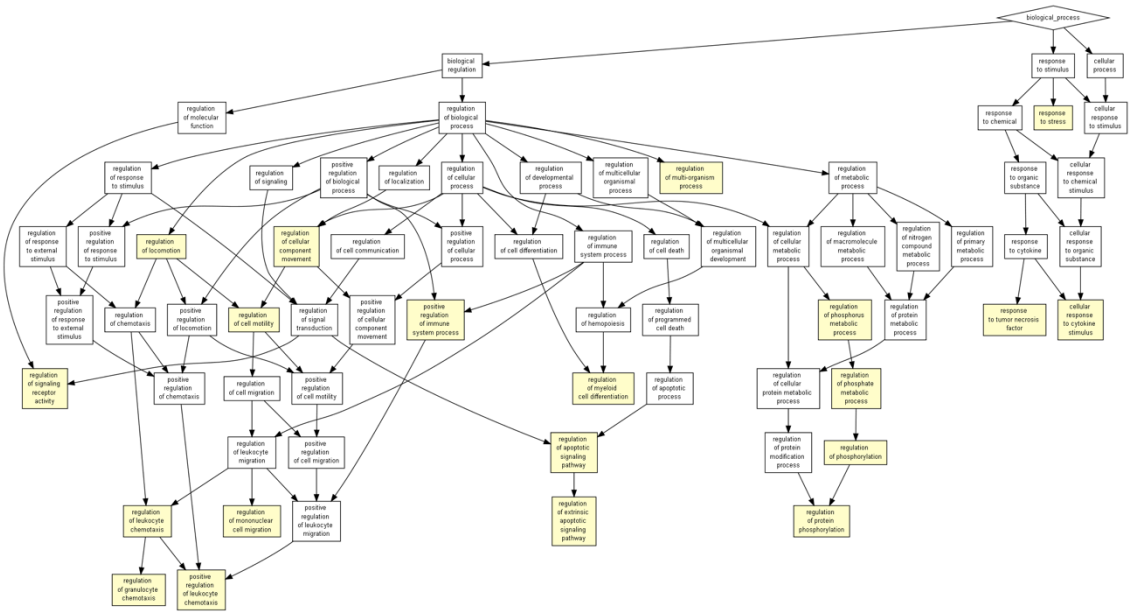


Figura 40. Mapa de procesos obtenido de GOrilla con el listado de genes diferencialmente expresados CvsL+L.

Process | Function | Component | Back to GOrilla

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0006950	response to stress	4.72E-5	1.79E-1	2.39 (168,57,21,17)	[+] Show genes
GO:2001233	regulation of apoptotic signaling pathway	1.09E-4	2.07E-1	4.70 (168,13,22,8)	[+] Show genes
GO:0002688	regulation of leukocyte chemotaxis	2.23E-4	2.81E-1	7.00 (168,6,20,5)	[+] Show genes
GO:0043900	regulation of multi-organism process	2.24E-4	2.13E-1	12.44 (168,9,6,4)	[+] Show genes
GO:0071675	regulation of mononuclear cell migration	2.44E-4	1.85E-1	8.40 (168,4,20,4)	[+] Show genes
GO:0002684	positive regulation of immune system process	3.18E-4	2.01E-1	3.53 (168,28,17,10)	[+] Show genes
GO:0001932	regulation of protein phosphorylation	4.11E-4	2.22E-1	2.74 (168,38,21,13)	[+] Show genes
GO:0019220	regulation of phosphate metabolic process	4.11E-4	1.95E-1	2.74 (168,38,21,13)	[+] Show genes
GO:0042325	regulation of phosphorylation	4.11E-4	1.73E-1	2.74 (168,38,21,13)	[+] Show genes
GO:0051174	regulation of phosphorus metabolic process	4.11E-4	1.56E-1	2.74 (168,38,21,13)	[+] Show genes
GO:0071345	cellular response to cytokine stimulus	4.59E-4	1.58E-1	3.28 (168,16,32,10)	[+] Show genes
GO:2000145	regulation of cell motility	4.7E-4	1.48E-1	3.14 (168,28,21,11)	[+] Show genes
GO:0010469	regulation of signaling receptor activity	5E-4	1.46E-1	7.37 (168,19,6,5)	[+] Show genes
GO:0034612	response to tumor necrosis factor	5.38E-4	1.45E-1	5.04 (168,8,25,6)	[+] Show genes
GO:0002690	positive regulation of leukocyte chemotaxis	7.13E-4	1.8E-1	8.40 (168,5,16,4)	[+] Show genes
GO:0040012	regulation of locomotion	7.34E-4	1.74E-1	3.03 (168,29,21,11)	[+] Show genes
GO:0051270	regulation of cellular component movement	7.34E-4	1.64E-1	3.03 (168,29,21,11)	[+] Show genes
GO:0071622	regulation of granulocyte chemotaxis	7.83E-4	1.65E-1	37.33 (168,3,3,2)	[+] Show genes
GO:0045637	regulation of myeloid cell differentiation	8.64E-4	1.72E-1	5.60 (168,6,25,5)	[+] Show genes
GO:2001236	regulation of extrinsic apoptotic signaling pathway	8.68E-4	1.64E-1	3.17 (168,7,53,7)	[+] Show genes

Tabla 11. Tabla de procesos obtenido de GOrilla con el listado de genes diferencialmente expresados CvsL+L.

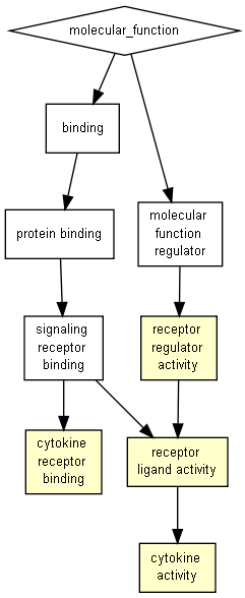


Figura 41. Mapa de funciones obtenido de GOrilla con el listado de genes diferencialmente expresados CvsL+L

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0005125	cytokine activity	1.64E-4	9.83E-2	8.75 (168,16,6,5)	[-] Hide genes INHBA - inhibin, beta a CSF1 - colony stimulating factor 1 (macrophage) IL8 - interleukin 8 IL1B - interleukin 1, beta CCL22 - chemokine (c-c motif) ligand 22
GO:0005126	cytokine receptor binding	2.36E-4	7.05E-2	8.24 (168,17,6,5)	[+] Show genes
GO:0048018	receptor ligand activity	5E-4	9.97E-2	7.37 (168,19,6,5)	[+] Show genes
GO:0030545	receptor regulator activity	6.14E-4	9.17E-2	7.00 (168,20,6,5)	[+] Show genes

Tabla 12. Tabla de funciones obtenido de GOrilla con el listado de genes diferencialmente expresados CvsL+L.

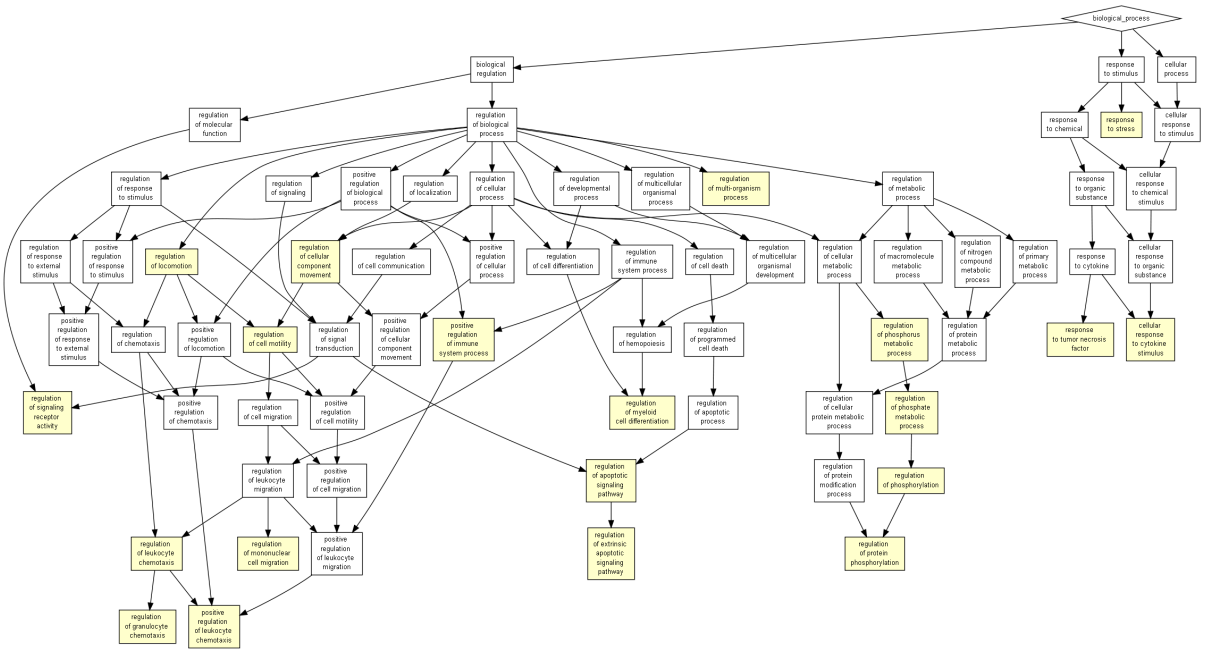


Figura 42. Mapa de componentes obtenido de GOrilla con el listado de genes diferencialmente expresados CvsL+L

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0005615	extracellular space	4.88E-4	1.82E-1	1.83 (168,33,64,23)	[+] Show genes
GO:0000323	lytic vacuole	6.24E-4	1.17E-1	9.33 (168,6,12,4)	[+] Show genes
GO:0005773	vacuole	6.24E-4	7.78E-2	9.33 (168,6,12,4)	[+] Show genes
GO:0005764	lysosome	6.24E-4	5.84E-2	9.33 (168,6,12,4)	[+] Show genes
GO:0005576	extracellular region	8.81E-4	6.59E-2	3.18 (168,33,16,10)	[+] Show genes

Tabla 13. Tabla de componentes obtenido de GOrilla con el listado de genes diferencialmente expresados CvsL+L.

GOrilla de Listado DE LPSvsC

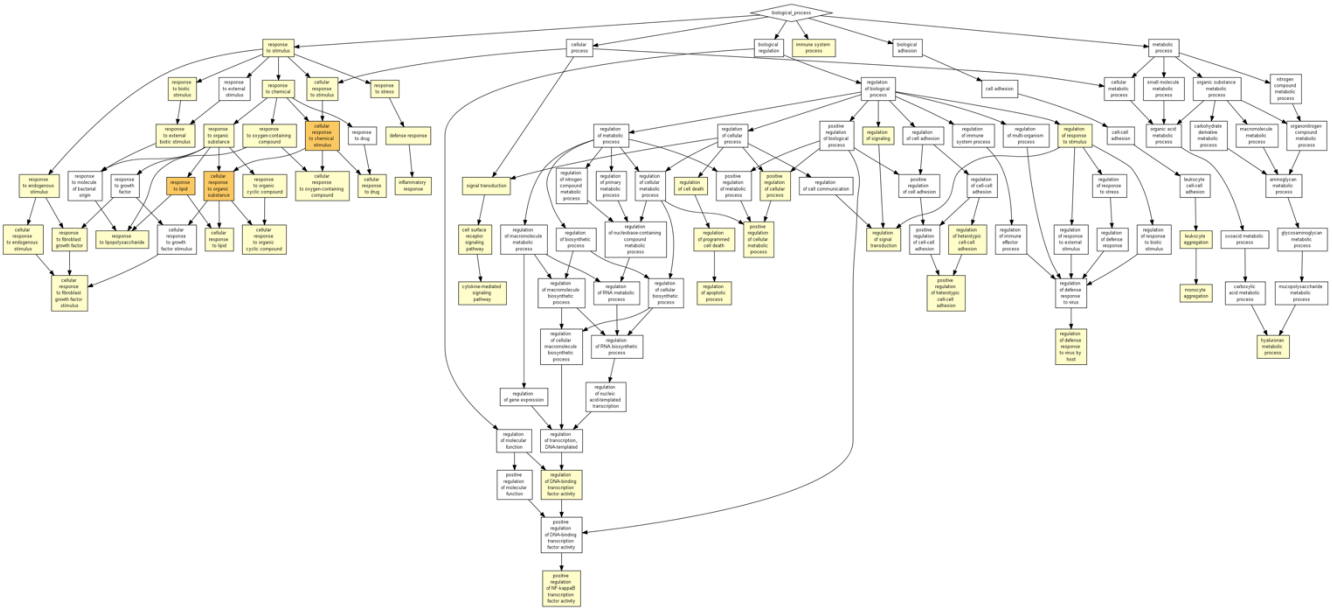
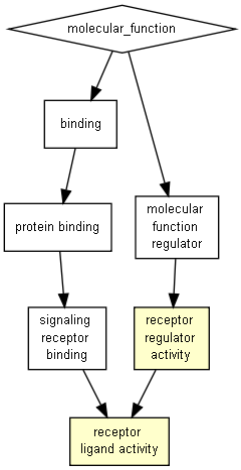


Figura 43. Mapa de procesos obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsC

Process | Function | Component | Back to GOrilla

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0071310	cellular response to organic substance	1.67E-6	8.31E-3	1.33 (312,73,218,68)	[+] Show genes
GO:0070887	cellular response to chemical stimulus	5.15E-6	1.28E-2	1.31 (312,80,218,73)	[+] Show genes
GO:0033993	response to lipid	6.66E-6	1.1E-2	2.50 (312,41,70,23)	[+] Show genes
GO:0051716	cellular response to stimulus	1.3E-5	1.62E-2	1.25 (312,108,218,94)	[+] Show genes
GO:0010033	response to organic substance	1.39E-5	1.38E-2	1.50 (312,104,122,61)	[+] Show genes
GO:0019221	cytokine-mediated signaling pathway	1.69E-5	1.4E-2	1.43 (312,54,198,49)	[+] Show genes
GO:1901700	response to oxygen-containing compound	2.28E-5	1.61E-2	2.06 (312,65,70,30)	[+] Show genes
GO:1901701	cellular response to oxygen-containing compound	2.31E-5	1.43E-2	2.33 (312,46,70,24)	[+] Show genes
GO:0050896	response to stimulus	4.59E-5	2.53E-2	1.27 (312,164,147,98)	[+] Show genes
GO:0007166	cell surface receptor signaling pathway	5.98E-5	2.96E-2	1.27 (312,98,205,82)	[+] Show genes
GO:0048583	regulation of response to stimulus	9.47E-5	4.27E-2	1.19 (312,146,210,117)	[+] Show genes
GO:0007165	signal transduction	1.01E-4	4.18E-2	1.19 (312,153,206,120)	[+] Show genes
GO:0010941	regulation of cell death	1.05E-4	4.01E-2	1.26 (312,74,227,68)	[+] Show genes
GO:0071396	cellular response to lipid	1.06E-4	3.76E-2	1.44 (312,30,216,30)	[+] Show genes
GO:0002376	immune system process	1.3E-4	4.28E-2	1.23 (312,103,216,88)	[+] Show genes
GO:0006952	defense response	1.43E-4	4.43E-2	1.57 (312,62,138,43)	[+] Show genes
GO:0006950	response to stress	1.6E-4	4.66E-2	1.37 (312,112,140,69)	[+] Show genes
GO:0071407	cellular response to organic cyclic compound	1.6E-4	4.42E-2	1.55 (312,24,201,24)	[+] Show genes
GO:0042221	response to chemical	1.79E-4	4.67E-2	1.52 (312,119,86,50)	[+] Show genes
GO:0071495	cellular response to endogenous stimulus	1.82E-4	4.51E-2	2.75 (312,33,55,16)	[+] Show genes
GO:0048522	positive regulation of cellular process	1.87E-4	4.41E-2	1.21 (312,169,173,113)	[+] Show genes
GO:0009719	response to endogenous stimulus	2.16E-4	4.87E-2	1.34 (312,41,227,40)	[+] Show genes
GO:0043067	regulation of programmed cell death	2.54E-4	5.47E-2	1.26 (312,71,227,65)	[+] Show genes

Tabla 14. Tabla de procesos obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsC



GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0048018	receptor ligand activity	1E-3	9.14E-1	1.62 (312,18,192,18)	[+] Show genes
GO:0030545	receptor regulator activity	1E-3	4.57E-1	1.62 (312,18,192,18)	[+] Show genes

Figura 44. Mapa de funciones obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsC

Tabla 15. Tabla de funciones obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsC.

Process | Function | Component

No GO Enrichment Found

There are no GO terms with an enrichment p-value above the value you specified

Tabla 16. Tabla de componentes obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsC. En este caso no existen términos GO con valores significativos.

GOrilla de Listado DE LPSvsL+L

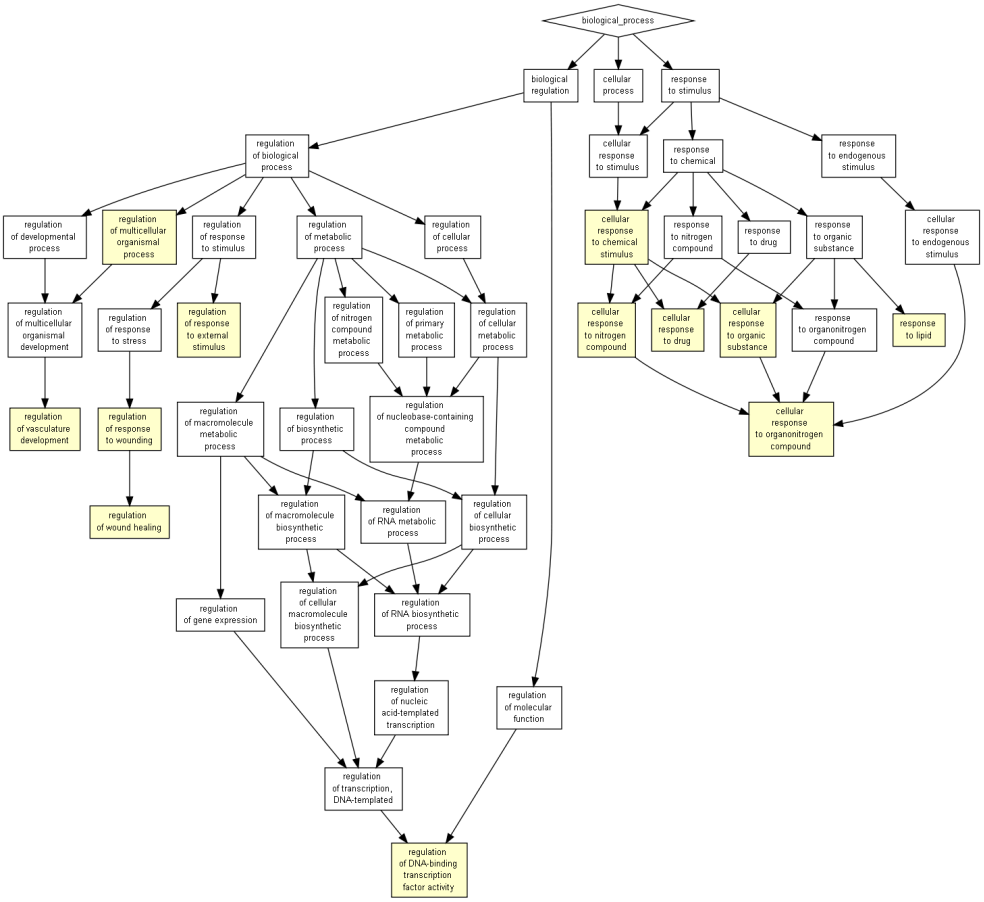


Figura 45. Mapa de procesos obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsL+L

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0051090	regulation of DNA-binding transcription factor activity	1.95E-5	7.08E-2	3.08 (161,17,40,13)	[+] Show genes
GO:0061041	regulation of wound healing	1.99E-4	3.62E-1	5.75 (161,8,21,6)	[+] Show genes
GO:1903034	regulation of response to wounding	1.99E-4	2.41E-1	5.75 (161,8,21,6)	[+] Show genes
GO:0070887	cellular response to chemical stimulus	2.04E-4	1.85E-1	1.27 (161,52,119,49)	[+] Show genes
GO:0051239	regulation of multicellular organismal process	2.16E-4	1.57E-1	1.60 (161,58,59,34)	[+] Show genes
GO:1901342	regulation of vasculature development	6.7E-4	4.06E-1	2.88 (161,13,43,10)	[+] Show genes
GO:0035690	cellular response to drug	7.18E-4	3.73E-1	3.18 (161,12,38,9)	[+] Show genes
GO:0032101	regulation of response to external stimulus	7.57E-4	3.44E-1	2.62 (161,25,32,13)	[+] Show genes
GO:0071310	cellular response to organic substance	7.98E-4	3.22E-1	1.30 (161,45,116,42)	[+] Show genes
GO:0033993	response to lipid	8.97E-4	3.26E-1	2.41 (161,24,39,14)	[+] Show genes
GO:0071417	cellular response to organonitrogen compound	9.58E-4	3.17E-1	2.27 (161,10,71,10)	[+] Show genes
GO:1901699	cellular response to nitrogen compound	9.58E-4	2.9E-1	2.27 (161,10,71,10)	[+] Show genes

Tabla 17. Tabla de procesos obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsL+L

Process | Function | Component | Back to GOrilla

No GO Enrichment Found

There are no GO terms with an enrichment p-value above the value you specified

Tabla 18. Tabla de funciones obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsL+L. En este caso no existen términos GO con valores significativos.

Process	Function	Component	Back to GOrilla

No GO Enrichment Found

There are no GO terms with an enrichment p-value above the value you specified

Tabla 19. Tabla de componentes obtenido de GOrilla con el listado de genes diferencialmente expresados LPSvsL+L. En este caso no existen términos GO con valores significativos.

De estas gráficas y tablas volvemos a obtener una gran cantidad de información, y de ella:

Observamos que en el apartado de procesos es donde el listado es mucho mayor. En este caso nuestros genes ,aun existiendo varios grupos hay términos GO que se repiten como respuesta al estrés, respuesta a lípidos, respuesta celular a estímulos, etc. De estos procesos repetidos, en nuestras diferentes comparaciones, podemos obtener varios listados de genes de interés como nuevos inmunocheckpoint.

Deteniéndonos en las funciones y creando un compendio de estas, obtenidas de todas la comparaciones de nuestros grupos, aparecen:
- Actividad del ligando receptor. Descripción: Interactúa con los receptores de manera que aumenta la proporción de receptores en la forma activa.
- Actividad del regulador receptor. Descripción: La función de interactuar (directa o indirectamente) con los receptores de modo que se modifique la proporción de receptores en la forma activa.
- Actividad de citoquinas. Descripción: Funciones para controlar la supervivencia, crecimiento, diferenciación y función efectora de tejidos y células.
- Unión al receptor de citoquinas. Descripción: Interactuar selectivamente y no covalentemente con un receptor de citoquinas.

De ellas podemos obtener un listado de genes que podrían ser importantes en la TE. Las funciones asociadas son coherentes debido a que en la TE todo este listado anterior se ve afectado en la reprogramación del SII, de hecho una de las alteraciones más importantes es la disminución de la tormenta de citoquinas [1,2]

Finalmente, volvemos a concluir que los listados de los genes DE se relacionan con componentes extracelulares, vacuolas o lisosomas este dato sigue corroborando el tipo de extracción.

Posteriormente, debido a la gran cantidad de resultados que hemos obtenido y que su análisis completo nos ocupará una gran cantidad de tiempo. Nos disponemos a centrar el análisis con asesoramiento del grupo de investigación en la búsqueda en nuestros listados de genes diferencialmente expresados centrándonos en un dominio o región que por experiencia puede ser importante y ver el patrón de expresión ahí, en este caso obtenemos los genes del dominio V-Set debido a que en este se encuentran moléculas como PDL1, manualmente buscando en google (“v-set domain protein”).
Posteriormente utilizando la web Uniprot, en la opción Retrieve/ID mapping introducimos el listado, seleccionando en los apartados disponibles nombre del

gen y el organismo (homo sapiens). Seleccionamos la salida de los genes siempre comprobándolos antes, y finalmente los enfrento a mis datos, para ver si ese dominio puede tener repercusión en mis grupos experimentales. Este procesamiento se realizó manualmente con la herramienta de filtrado de Excel.

Posteriormente, generamos un heatmap en el cual volvemos a ver un patrón de expresión definido por grupos, lo que nos indica que la selección de genes podría ser correcta para indagar en ellos y estudiar sus posibles implicaciones en la TE.

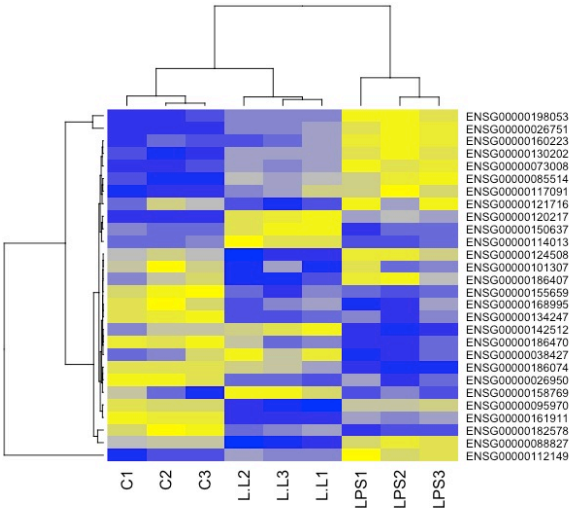


Figura 46. Heatmap de los genes del dominio V-set con expresión diferencial.

A continuación presento el listado de estos genes.

Entry name	Protein names	Gene names
SLAF7_HUMAN	SLAM family member 7 (CD2 subset 1) (CD2-like receptor-activating cytotoxic cells) (CRACC) (Membrane protein FOAP-12) (Novel Ly9) (Protein 19A) (CD antigen CD319)	SLAMF7 CS1 UNQ576/PRO1138
CD83_HUMAN	CD83 antigen (hCD83) (B-cell activation protein) (Cell surface protein HB15) (CD antigen CD83)	CD83
PVR_HUMAN	Poliovirus receptor (Nectin-like protein 5) (NECL-5) (CD antigen CD155)	PVR PVS
CD48_HUMAN	CD48 antigen (B-lymphocyte activation marker BLAST-1) (BCM1 surface antigen) (Leukocyte antigen MEM-102) (SLAM family member 2) (SLAMF2) (Signaling lymphocytic activation molecule 2) (TCT.1) (CD antigen CD48)	CD48 BCM1 BLAST1
PILRA_HUMAN	Paired immunoglobulin-like type 2 receptor alpha (Cell surface receptor FDF03) (Inhibitory receptor PILR-alpha)	PILRA
PILRB_HUMAN	Paired immunoglobulin-like type 2 receptor beta (Activating receptor PILR-beta) (Cell surface receptor FDFACT)	PILRB FDFACT PP1551
SHPS1_HUMAN	Tyrosine-protein phosphatase non-receptor type substrate 1 (SHP substrate 1) (SHPS-1) (Brain Ig-like molecule with tyrosine-based activation motifs) (Bit) (CD172 antigen-like family member A) (Inhibitory receptor SHPS-1) (Macrophage fusion receptor) (MyD-1 antigen) (Signal-regulatory protein alpha-1) (Sirp-alpha-1) (Signal-regulatory protein alpha-2) (Sirp-alpha-2) (Signal-regulatory protein alpha-3) (Sirp-alpha-3) (p84) (CD antigen CD172a)	SIRPA BIT MFR MYD1 PTPNS1 SHPS1 SIRP
ICOSL_HUMAN	ICOS ligand (B7 homolog 2) (B7-H2) (B7-like protein Gl50) (B7-related protein 1) (B7RP-1) (CD antigen CD275)	ICOSLG B7H2 B7RP1 ICOSL KIAA0653
NECT2_HUMAN	Nectin-2 (Herpes virus entry mediator B) (Herpesvirus entry mediator B) (HveB) (Nectin cell adhesion molecule 2) (Poliovirus receptor-related protein 2) (CD antigen CD112)	NECTIN2 HVEB PRR2 PVRL2
SN_HUMAN	Sialoadhesin (Sialic acid-binding Ig-like lectin 1) (Siglec-1) (CD antigen CD169)	SIGLEC1 SN
BT2A2_HUMAN	Butyrophilin subfamily 2 member A2	BTN2A2 BT2.2 BTF2
CLM2_HUMAN	CMRF35-like molecule 2 (CLM-2) (CD300 antigen-like family member E) (CMRF35-A5) (Immune receptor expressed on myeloid cells 2) (IREM-2) (Polymeric immunoglobulin receptor 2) (PIgR-2) (PIgR2) (Poly-Ig receptor 2) (CD antigen CD300e)	CD300E CD300LE CLM2 CMRF35A5 IREM2
CD226_HUMAN	CD226 antigen (DNAX accessory molecule 1) (DNAM-1) (CD antigen CD226)	CD226 DNAM1

PD1L1_HUMAN	Programmed cell death 1 ligand 1 (PD-L1) (PDCD1 ligand 1) (Programmed death ligand 1) (B7 homolog 1) (B7-H1) (CD antigen CD274)	CD274 B7H1 PDCD1L1 PDCD1LG1 PDL1
JAM1_HUMAN	Junctional adhesion molecule A (JAM-A) (Junctional adhesion molecule 1) (JAM-1) (Platelet F11 receptor) (Platelet adhesion molecule 1) (PAM-1) (CD antigen CD321)	F11R JAM1 JCAM UNQ264/PRO301
CSPG2_HUMAN	Versican core protein (Chondroitin sulfate proteoglycan core protein 2) (Chondroitin sulfate proteoglycan 2) (Glial hyaluronate-binding protein) (GHAP) (Large fibroblast proteoglycan) (PG-M)	VCAN CSPG2
CD86_HUMAN	T-lymphocyte activation antigen CD86 (Activation B7-2 antigen) (B70) (BU63) (CTLA-4 counter-receptor B7.2) (FUN-1) (CD antigen CD86)	CD86 CD28LG2
SIG10_HUMAN	Sialic acid-binding Ig-like lectin 10 (Siglec-10) (Siglec-like protein 2)	SIGLEC10 SLG2 UNQ477/PRO940
SIRB1_HUMAN	Signal-regulatory protein beta-1 (SIRP-beta-1) (CD172 antigen-like family member B) (CD antigen CD172b)	SIRPB1
SIRBL_HUMAN	Signal-regulatory protein beta-1 isoform 3 (SIRP-beta-1 isoform 3)	SIRPB1
TREM2_HUMAN	Triggering receptor expressed on myeloid cells 2 (TREM-2) (Triggering receptor expressed on monocytes 2)	TREM2
TRML1_HUMAN	Trem-like transcript 1 protein (TLT-1) (Triggering receptor expressed on myeloid cells-like protein 1)	TREML1 TLT1 UNQ1825/PRO3438
BT3A1_HUMAN	Butyrophilin subfamily 3 member A1 (CD antigen CD277)	BTN3A1 BTF5
BT3A2_HUMAN	Butyrophilin subfamily 3 member A2	BTN3A2 BT3.2 BTF3 BTF4
CLM1_HUMAN	CMRF35-like molecule 1 (CLM-1) (CD300 antigen-like family member F) (Immune receptor expressed on myeloid cells 1) (IREM-1) (Immunoglobulin superfamily member 13) (IgSF13) (NK inhibitory receptor) (CD antigen CD300f)	CD300LF CD300F CLM1 IGSF13 IREM1 NKIR UNQ3105/PRO10111
CSF1R_HUMAN	Macrophage colony-stimulating factor 1 receptor (CSF-1 receptor) (CSF-1-R) (CSF-1R) (M-CSF-R) (EC 2.7.10.1) (Proto-oncogene c-Fms) (CD antigen CD115)	CSF1R FMS
VSIG4_HUMAN	V-set and immunoglobulin domain-containing protein 4 (Protein Z39Ig)	VSIG4 CRIg Z39IG UNQ317/PRO362
FPRP_HUMAN	Prostaglandin F2 receptor negative regulator (CD9 partner 1) (CD9P-1) (Glu-Trp-Ile EWI motif-containing protein F) (EWI-F) (Prostaglandin F2-alpha receptor regulatory protein) (Prostaglandin F2-alpha receptor-associated protein) (CD antigen CD315)	PTGFRN CD9P1 EWIF FPRP KIAA1436
SIGL7_HUMAN	Sialic acid-binding Ig-like lectin 7 (Siglec-7) (Adhesion inhibitory receptor molecule 1) (AIRM-1) (CDw328) (D-siglec) (QA79 membrane protein) (p75) (CD antigen CD328)	SIGLEC7 AIRM1

Tabla 20. Tabla de posibles inmunocheckpoints del dominio V-set.

Para completar el análisis, utilizamos una nueva herramienta, Enrichr, para asociar nuestra selección de genes a los procesos y funciones pertinentes.

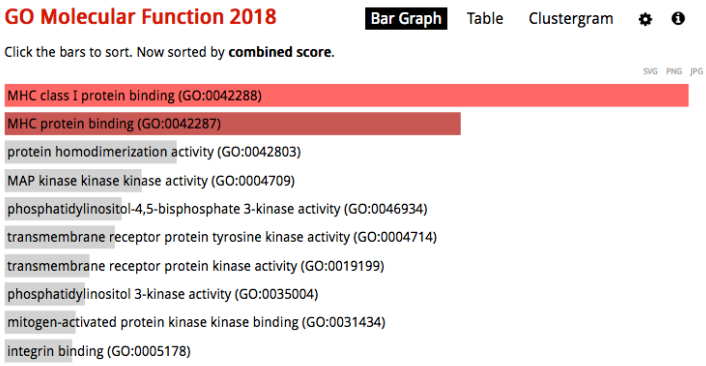


Figura 47. GO Molecular Function de nuestro listado de Inmunocheckpointscon Enrichr.

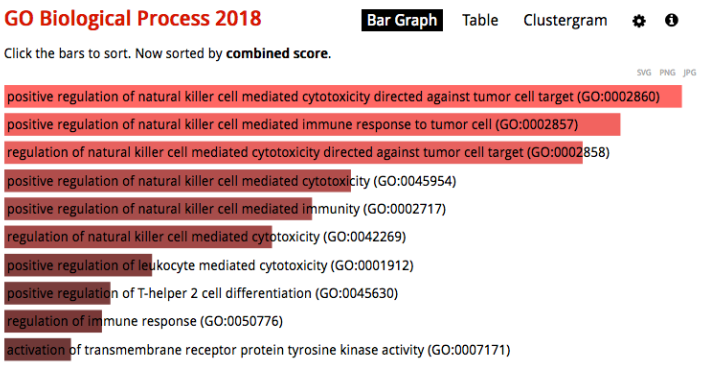


Figura 48. GO Biological Process de nuestro listado de Inmunocheckpointscon Enrichr.

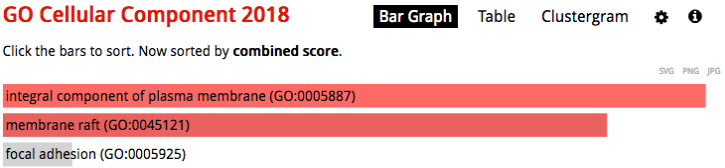


Figura 49. GO Celular Component de nuestro listado de Inmunocheckpointscon Enrichr.

Para finalizar nuestro proyecto podemos concluir que hemos obtenido un listado de inmunocheckpointsdel dominio v-set que están implicados en diferentes funciones asociadas a la unión a proteínas del complejo mayor de Histocompatibilidad (MHC). Esta reportado que el complejo formado por el antígeno y el MHC es reconocido por los linfocitos T iniciando la respuesta adaptativa[3,14,20]. Además es responsable de activar diversas vías de transducción de señales que inducen la liberación de citoquinas IL6, IL8, IL12 y el factor de necrosis tumoral α [23].

Si nos centramos en los procesos biológicos podríamos decir que estos genes están involucrados en la regulación de las “natural killer” (NK), es lógico ya que las células presentadoras de antígeno secretan IL12 e IL18, incrementando la función citotóxica de las células NK y los linfocitos T CD8+. Todos estos procesos están alterados en la TE [21,22,24].

En la TE se describe una reducida expresión de las moléculas del MHC de clase II, HLA-DQ y HLA-DR se ha relacionado con incapacidad para la presentación antigénica que se observa en los monocitos presentes en diversas patologías como la sepsis, la Fibrosis Quística, etc. [1,2,14,19]. Con ello queremos concluir que este listado podría ser un buen punto de partida para aclarar la TE y obtener moléculas “diana” para futuros tratamientos.

4. Conclusiones

4.1 Conclusiones del estudio

1. En este trabajo se ha desarrollado un análisis completo de RNA-seq obteniendo un conocimiento muy completo de la técnica y de las herramientas bioinformáticas disponibles para desarrollar este proceso.

2. Hemos comprobado que nuestros grupos tienen patrones de expresión diferentes (Control: simulando el donante sano, LPS: simulando una inflamación, y L+L: simulando la TE). Indicando la independencia de estos para posteriormente poderlos asociar a diferentes estados en una enfermedad.

3. Hemos obtenido un listado de genes para caracterizar distintos estados asociados a enfermedades como la sepsis, con tres modelos “in vitro” desarrollados con monocitos/macrófagos de donantes sanos. Con esta información obtenemos valiosos genes “diana” para nuevos estudios.

4. Los genes obtenidos en nuestros diferentes grupos están asociados a procesos de respuesta al estrés, respuesta a lípidos, respuesta celular a estímulos, etc. Debemos indagar más en estos listados para realizar un análisis exhaustivo y ver de una manera específica la implicación de nuestros grupos en estos procesos.

5. Los genes obtenidos se asocian a las funciones de actividad del ligando receptor, actividad del regulador receptor, actividad de citoquinas y unión al receptor de citoquinas. Las funciones asociadas son coherentes debido a que en la TE, este listado anterior se ve afectado en la reprogramación del SII, de hecho una de las alteraciones más importantes es la disminución de la tormenta de citoquinas.

6. Hemos obtenido un listado de inmunocheckpointsasociado al dominio v-set. Por indicación propia del grupo de investigación debido a que en este se encuentran moléculas como PDL1 catalogado como inmunocheckpoint.

7. El listado de genes de inmunocheckpointspodría ser un buen punto de partida para aclarar la TE y obtener moléculas “diana” para futuros tratamientos. Debido a que están implicados en diferentes funciones asociadas a la unión a proteínas del complejo mayor de Histocompatibilidad (MHC). Y estos están involucrados en procesos biológicos de regulación de las NK.

4.2 Cumplimiento de planificación y Autoevaluación

El proyecto ha finalizado cumpliendo todos los objetivos previstos. En todo el proceso ha habido una gran parte de estudio, uso de herramientas bioinformáticas y también realización de diferentes procedimientos para obtener resultados y análisis de estos, acorde a nuestro conocimiento y equipamiento. Hemos tenido que modificar los tiempos de algunas tareas debido al desconocimiento del campo y de las herramientas a utilizar, pero finalmente se alcanzó en mayor o menor medida todos los objetivos planteados. Debido a estas variaciones es posible que resulte escaso el análisis completo del RNA-seq, ya que disponiendo de mayor tiempo se podría obtener una información más completa e incluso enfocarnos en genes específicos e interesantes para la TE. También se podría esperar un análisis estadístico más exhaustivo por ejemplo de los tres grupos en su totalidad y no solo las diferencias dos a dos mostradas. Con ello nos evaluamos correctamente al alcanzar los objetivos, pero comprendemos las flaquezas del trabajo pudiendo mejorarse con un análisis más exhaustivo.

4.3 Futuro

El paso siguiente a realizar es un análisis más exhaustivo de cada gen con una posible implicación en TE. Y obteniendo muestras de pacientes, poder comprobar estos estados en ellos, con una posible asociación a buena o mala evolución.

Con el listado de inmunocheckpointsse deben plantear nuevos estudios para ver si realmente esos genes regulan o no el sistema inmune. Y que implicación tienen en diferentes enfermedades que se pueda dar la TE.

5. Glosario

TE: Tolerancia a Endotoxinas
SII: Sistema inmune innato
LPS: Lipopolisacárido
PBMC: Células mononucleares de sangre periférica
RNA-seq: RNA sequencing
NGS: Secuenciación de nueva generación
retículo endoplásmico (ER)
A: Adenina
G: Guanina
C: Citosina
T: Timina
CPM: Conteos por millón
TMM: Trimmed mean of M values
MDS: Multidimensional scaling plot of distances
PCA: Análisis de componentes principales
BCV: coeficiente de variación biológica
DE: genes diferencialmente expresados

FDR: Razón de falsos descubrimientos
C: grupo Control
L+L: Grupo L+L
GSEA: Análisis de enriquecimiento de conjuntos de genes
NK: Natural killer
MHC: Complejo mayor de Histocompatibilidad
ADN: Ácido Desoxirribonucleico
ARN: Ácido ribonucleico

6. Bibliografía

Artículos científicos.

1. López-Collazo E, del Fresno C (2013) Pathophysiology of endotoxin tolerance: mechanisms and clinical consequences. Crit Care 17: 242.
2. Biswas SK, López-Collazo E (2009) Endotoxin tolerance: new mechanisms, molecules and clinical significance. Trends Immunol 30: 475-487.
3. Escoll P, del Fresno C, Garcia L, Valles G, Lendinez MJ, et al. (2003) Rapid up-regulation of IRAK-M expression following a second endotoxin challenge in human monocytes and in monocytes isolated from septic patients. Biochem Biophys Res Commun 311: 465-472.
4. López-Collazo E, Gomez-Pina V, Arnalich F (2010) Understanding immune dysfunctions in sepsis patients. Crit Care 14: 435.
5. Hayden WR (1993) Sepsis and organ failure definitions and guidelines. Crit Care Med 21: 1612-1613.
6: Cubillos-Zapata C, Hernandez-Jimenez E, Toledano V, Esteban-Burgos L, Fernandez-Ruiz I, et al. (2014) NFkappaB2/p100 is a key factor for endotoxin tolerance in human monocytes: a demonstration using primary human monocytes from patients with sepsis. J Immunol 193: 4195-4202.
7: Medzhitov R (2013) Septic shock: on the importance of being tolerant. Immunity 39: 799-800.
8: Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A,
Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. (2016) A survey of best
practices for RNA-seq data analysis. Genome Biol. Jan 26;17:13
9: Costa-Silva J, Domingues D, Lopes FM. (2017) RNA-Seq differential expression
analysis: An extended review and a software tool. PLoS One. Dec
21;12(12):e0190152
10:Carrasco Carrasco,Sara. Técnicas de análisis de expresión diferencial basadas en conteos para el estudio de datos de RNA-seq usando R y Bioconductor. TFG. Unv. De Sevilla.(2015)

11: Stoltenburg R., Wartmann T., Kunze I., and Kunze G. (1995) Reliable method to prepare RNA from free and membrane-bound polysomes from different yeast species. Biotechniques 18: 564-568
12: Subramanian, A; Tamayo, P; Mootha, Vamsi K.; Mukherjee, Sayan; Ebert, Benjamin L.; Gillette, Michael A.; Paulovich, Amanda; Pomeroy, Scott L.; Golub, Todd R. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 102 43: 15545–15550.
13: Ma'ayan Laboratory - Computational Systems Biology - Icahn School of medicine at Mount Sinaí". labs.icahn.mssm.edu.
14: del Fresno C, Garcia-Rio F, Gomez-Pina V, Soares-Schanoski A, Fernandez-Ruiz I, et al. (2009) Potent phagocytic activity with impaired antigen presentation identifying lipopolysaccharide-tolerant human monocytes: demonstration in isolated monocytes from cystic fibrosis patients. J Immunol 182: 6494-6507.
15: Jurado-Camino T, Cordoba R, Esteban-Burgos L, Hernandez-Jimenez E, Toledano V, et al. (2015) Chronic lymphocytic leukemia: a paradigm of innate immune cross-tolerance. J Immunol 194: 719-727.
16: Shalova IN, Lim JY, Chittezhath M, Zinkernagel AS, Beasley F, et al. (2015) Human monocytes undergo functional re-programming during sepsis mediated by hypoxia-inducible factor-1alpha. Immunity 42: 484-498.
17: Muto J, Yamasaki K, Taylor KR, Gallo RL. 2009. Engagement of CD44 by hyaluronan suppresses TLR4 signaling and the septic response to LPS. Molecular immunology 47:449.
18: Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, et al. (1992) Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. Chest 101: 1644-1655.
19: Gomez HG, Gonzalez SM, Londono JM, Hoyos NA, Nino CD, et al. (2014) Immunological characterization of compensatory anti-inflammatory response syndrome in patients with severe sepsis: a longitudinal study*. Crit Care Med 42: 771-780.
20:Fernandez-Ruiz I, Arnalich F, Cubillos-Zapata C, Hernandez-Jimenez E, Moreno-Gonzalez R, et al. (2014). Mitochondrial DAMPs Induce Endotoxin Tolerance in Human Monocytes: An Observation in Patients with Myocardial Infarction. PloS one 9:e95073
21:Fujiwara N, Kobayashi K. (2005). Macrophages in inflammation. Current drug targets. Inflammation and allergy 4:281-6
22. Puccetti P, Bella donna ML, Grohmann U. (2002). Effects of IL-12 and IL-23 on antigen-presenting cells at the interface between innate and adaptive immunity. Critical reviews in immunology 22:373-90
23: Álvarez E, Toledano V, Morilla F, Hernández-Jiménez E, Cubillos-Zapata C, Varela-Serrano A, Casas-Martín J, Avendaño-Ortiz J, Aguirre LA, Arnalich F, Maroun-Eid C, Martín-Quirós A, Quintana Díaz M, López-Collazo E. (2017)A System Dynamics Model to Predict the Human Monocyte Response to Endotoxins. Front Immunol. Aug 3;8:915
24: Hernández-Jiménez E, Cubillos-Zapata C, Toledano V, Pérez de Diego R, Fernández-Navarro I, Casitas R, Carpio C, Casas-Martín J, Valentín J, Varela-Serrano A, Avendaño-Ortiz J, Alvarez E, Aguirre LA, Pérez-Martínez A, De

Miguel MP, Belda-Iniesta C, García-Río F, López-Collazo E. (2017) Monocytes inhibit NK activity via TGF-β in patients with obstructive sleep apnoea. Eur Respir J. Jun 15;49(6).

Paginas web utilizadas durante todo el desarrollo de este trabajo:

- RNA-seq Blog: https://www.rna-seqblog.com/ 09/2018
- Bioinformatics babraham. FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ 09/2018
- Trimmomatic: A flexible read trimming tool for Illumina NGS data: http://www.usadellab.org/cms/?page=trimmomatic 09/2018
- Cutadapt. https://cutadapt.readthedocs.io/en/stable/guide.html 09/2018
- PrinSeq: http://prinseq.sourceforge.net/manual.html 09/2018
- TopHat: https://ccb.jhu.edu/software/tophat/index.shtml 09/2018
- HISAT2: https://ccb.jhu.edu/software/hisat2/index.shtml 09/2018
- Kallisto: https://pachterlab.github.io/kallisto/ 10/2018
- IGV: http://software.broadinstitute.org/software/igv/ 10/2018
- Qualimap: http://qualimap.bioinfo.cipf.es/ 10/2018
- Cufflinks, Cuffmerge, Cuffdiff: http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/ 10/2018
- R: https://www.r-project.org/ 10/2018
- Bioconductor: https://www.bioconductor.org/ 10/2018
- EdgeR: https://bioconductor.org/packages/release/bioc/html/edgeR.html 10/2018
- GOrilla: http://cbl-gorilla.cs.technion.ac.il/ 11/2018
- Enrichr: http://amp.pharm.mssm.edu/Enrichr/ 12/2018

7. Anexos

Anexo I: Listado de documentos entregados

Archivos entregados	Tipo de Documento
Memoria final del TFM	Documento pdf
Listados de genes DE	Archivos .txt obtenidos con R
Listado de genes candidatos a inmunocheckpoint	Archivo .xls
Listados de genes que definen los patrones de expresión de cada grupo	Incluido en documento de memoria final y documento Excel: Tabla_top10_DE
Información y gráficos de los resultados del TFM	Incluido en documento de memoria final
Script utilizados	Adjunto en el Anexo II

Anexo II: Script del desarrollo del proyecto

Para todo el procedimiento se utilizó el sistema operativo de Ubuntu.

#En este caso con el terminal de UBUNTU
#Podemos ver nuestros ficheros fastq generados del equipo con el siguiente código:

less mRNA_EQ_LPS_S203_L005_R1_001.fastq

#Posteriormente instalamos FastQC y cargamos el programa con el código:
.\fastqc
#Con esta herramienta realizamos el análisis de control de calidad, en este caso con el software utilizando ventanas.

#Posteriormente quitaremos los adaptadores con el programa cutadapt, de nuestras secuencias en ficheros fastq, en este caso cada muestra una a una. Utilizando el siguiente código.

Cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -e 0.1 -m 30 mRNA_EQ_LPS_S203_L005_R1_001.fastq -o EQ_LPS.fastq

#Posteriormente instalamos varios programas como Bowtie2 y TopHat2.
#Necesitamos también las lecturas que corresponden al organismo de humano descargándola de: http://hgdownload.soe.ucsc.edu/downloads.html#human, utilizando el Genoma Dec. 2013 (GRCh38/hg38)

#Utilizando el terminal lo descomprimimos:

gzip -d hg38.fa.gz

#Renombramos a Hg38 con el código.

mv Hg38.fa hg38

#Utilizaremos Bowtie2 para poder indexar el genoma. El código sería el siguiente:

bowtie2-build hg38.fa hg38

#A continuación indexamos el archivo de anotación con extensión .GTF (también obtenido de la página UCSC) con nuestro genoma de referencia en nucleótidos, Utilizando el siguiente código.

tophat2 --G hg38.gtf --transcriptome-index=hg38.tr hg38.fa

#Realizamos el mapeo de nuestras muestras, una a una:

tophat2 –o carpeta_salida –p 4 –-transcriptome-index= hg38.gtf hg38 EQ_LPS.fastq

#Instalamos Qualimap y nuevamente con un formato de ventanas realizamos el análisis de calidad.
#Generamos un archivo con el conjunto de los resultados de mis muestras con las reads de cada grupo. Este documento fue generado con la ayuda de mi consultor **Enrique Vázquez de Luis** utilizando RSEM, por eso no describo el código ya que este paso del proceso no es mérito mio.

#EN ESTE SIGUIENTE CODIGO SE GENERARON MUCHOS SCRIPT Y VARIACIONES DE ESTE, PARA OBTENER TODOS LOS RESULTADOS. EL DESCRITO A CONTINUACIÖN ES UN COMPENDIO DE ELLOS, POR DECIRLO ASI EL DEFINITIVO PARA GENERAR LOS RESULTADOS PRESENTADOS.

#Primero instalo Bioconductor y todas las librerías necesarias para el desarrollo del proceso
```
source("http://bioconductor.org/biocLite.R")
require(biocLite)
biocLite("edgeR")
biocLite("DESeq2")
require(edgeR)
require(DESeq2)
install.packages("DescTools")
require("DescTools")
install.packages("ggplot2")
require("ggplot2")
```

#Leemos el archivo generado anteriormente.
```
y <- read_excel("~/Desktop/count_victor.xlsx",
sheet = "ExpresionNormalizada", col_types = c("text",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric"))
View(count_norm)
```

#Lo hago data frame
```
y <- data.frame(y)
```

#Miro las dimensiones
```
dim(y)
[1] 46189    10
```

#Le digo que la columna uno son los nombres
```
rownames(y)<- y[,1]
```

#Lo revisamos y aparecen los nombre y la columna1 que tenemos que extraer
```
head(y)
y<- y[,2:10]
head(y)
```

#Finalmente, ya obtenemos los datos para realizar todo el procesamiento.
counts<-y[rowSums(cpm(y) >= 1) >= 3,]
> dim(counts)
[1] 14848   9
>colnames(counts)

#Definimos los grupos
grupo <- c("C","C","C","LPS","LPS","LPS","L+L","L+L","L+L")
dge <- DGEList(counts= counts, group=grupo)

#Sacamos los plotMDS, y después hacemos la normalización
dge = calcNormFactors(dge)
dge$samples
maPlot(dge$counts[,1], dge$counts[,2])

#Podemos realizar un diagnóstico visual con figuras de tipo MA donde comparamos el log de la Abundancia (cuentas de lecturas mapeadas) en el eje X contra el log Fold-change(diferencia de expresión) en el eje Y.

#Para facilitar la interpretación podemos agregar una gradilla y una línea de tendencia, mostrando la diferencia entre los dos factores de normalización de las mismas bibliotecas
grid(col = "blue")

maPlot(dge$counts[,1], dge$counts[,4])
maPlot(dge$counts[,1], dge$counts[,6])

#Con esto he comparado columnas de un mismo grupo, y con otros dos grupos restantes
#Vemos que si son de un mismo grupo tenemos menos dispersión que con columna de distintos grupos
dge = estimateCommonDisp(dge)

#Con la función anterior vamos a ver la dispersión de nuestros datos
#Para hacer un análisis de expresión diferencial necesitamos estimar la dispersión de nuestros datos, y luego hacer una prueba para buscar diferencias entre nuestros grupos
et = exactTest(dge, dispersión=dge$common.dispersion)
et
et2 = exactTest(dge, pair= 2:3, dispersión=dge$common.dispersion)
et2

#Después podemos interrogar los genes más diferencialmente expresados de nuestras dos comparaciones.
topTags(et)
topTags(et2)

#Estos pocos genes tienen una confianza estadística excelente. Probablemente queramos ver una lista más grande y completa de genes. Generalmente es más fácil obtener la evaluación estadística de todos los genes y después elegir los

umbrales de corte a utilizar.
deTab <- topTags(et, n=Inf)$table
head(deTab)
deTab2 <- topTags(et2, n=Inf)$table
head(deTab2)

#Podemos elegir ahora genes que tienen un P-valor ajustado de menos de 0.01 (esto representa la fracción de falsos positivos) y una magnitud de cambio (log Fold-change) mayor a 1 (fuera de escala log2, mayor a 2).
deGenes = rownames(deTab)[deTab$FDR < 0.01 & abs(deTab$logFC) > 1]
head(deGenes)
length(deGenes)
plotSmear(dge, de.tags=deGenes)

deGenes2 = rownames(deTab2)[deTab2$FDR < 0.01 & abs(deTab2$logFC) > 1]
head(deGenes2)
length(deGenes2)
plotSmear(dge, de.tags=deGenes2)

#Finalmente, podemos guardar la tabla completa de nuestros resultados de expresión diferencial. Podemos hacerlo así:
write.table(deTab, file="DiffExp_edgeR_C_LL.txt", sep="\t")
write.table(deTab2, file="DiffExp_edgeR_LL_L.txt", sep="\t")

#Realizar heatmap de los counts iniciales ordenados
counts_o<- counts[order(counts$LPS1, counts$LPS2, counts$LPS3),]
heatmap(as.matrix(counts_o[1:500,]))
heatmap(as.matrix(counts_o[13500:14000,]))

#Para la comparación CvsLPS, Modifique el archivo inicial excel eliminando el grupo L+L, porque no era capaz de realizar esta comparativa, y volví a hacer el mismo código anterior, y así obtuve la comparativa CvsLPS

#Para hacer los heatmaps de los TOP de cada DE (en este caso solo muestro uno de ellos pero fue el mismo código para cada uno de ellos)

TopDE_CvsLL <- read_excel("~/Desktop/TopDE_CvsLL.xlsx",
sheet = "Hoja1", col_types = c("text",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric"))
View(TopDE_CvsLL)
y3<- data.frame(TopDE_CvsLL)
rownames(y3)<- y3[,1]
count_3<- y3[,2:10]
counts_o_3<- count_3[order(count_3$L.L1),]

colfunc2 <- colorRampPalette(c("blue","grey", "red"))
heatmap(as.matrix(counts_o_3),col=colfunc2(15),scale="row", trace="none")

Anexo III. Descripción del proceso de Análisis de expresión diferencial con la herramientas Cufflinks; Cuffmerge; Cuffdiff

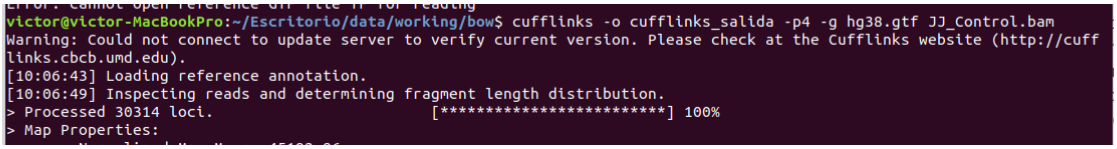
Para obtener la expresión génica diferencial explicamos los procesos a seguir con la herramientas Cufflinks; Cuffmerge; Cuffdiff, este proceso es una medición cuantitativa de las diferencias entre los tránscritos de nuestro diferentes de grupos

Recuento de las lecturas.

El primer paso después del alineamiento, consiste en la comparación de los recuentos de las lecturas de cada transcrito, es decir realizar un conteo por gen o transcrito de cada una de nuestras muestras. Este proceso genera una cuantificación de nuestras lecturas en relación a los genes o transcritos de nuestro genoma de referencia. Para este procedimiento utilizaremos Cufflinks.

Cufflinks es una herramienta que ensambla alineamientos de lecturas de RNA en transcritos, calcula estimaciones de su abundancia y prueba la expresión diferencial y la regulación del transcriptoma.
Esta se basa en un algoritmo basado en grafos dirigidos acíclicos que buscan el conjunto mínimo de transcritos independientes que pueden explicar las lecturas de nuestro RNA-seq, agrupa las lecturas en clusters que se mapean a la misma región del genoma y va etiquetando como incompatibles a los que no pueden venir del mismo transcrito. Una vez identificado el número mínimo de transcritos posibles, se asigna cada lectura a uno o más transcritos.

En este caso utilizamos el siguiente código:

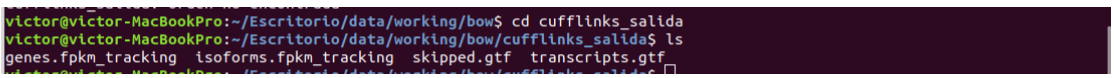


cufflinks -o cufflinks_salida -p4 -g hg38.gft JJ_Control.bam

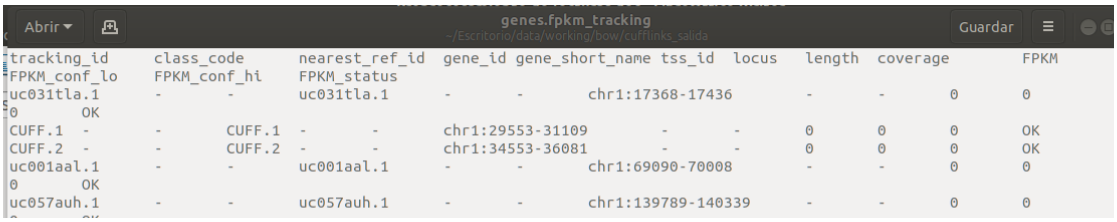
Explicamos a continuación los parámetros utilizados en nuestro código:

-o: Carpeta donde se guardarán los archivos, en este caso cufflinks_salida
-p: Número de “threads” para alinear las lecturas.
-g: Archivo GFF/GFT del genoma de referencia a utilizar.

Este nos genera los siguiente archivos:



Veamos el contenido de uno de los ficheros: el fichero genes.fpkm_tracking que nos muestra la expresión diferencial de los genes



Los ficheros de salida que produce Cufflinks son:

•file: transcripts.gtf: Este archivo GTF contiene las isomorfas ensambladas de Cufflinks.
•file: isoforms.fpkm_tracking: Contiene los valores estimados de los niveles de expresión de las isomorfas en el formato de seguimiento FPKM.
•file: genes.fpkm_tracking: Contiene los valores estimados de los niveles de expresión de los genes en el formato de seguimiento FPKM.

Creación de una anotación especifica

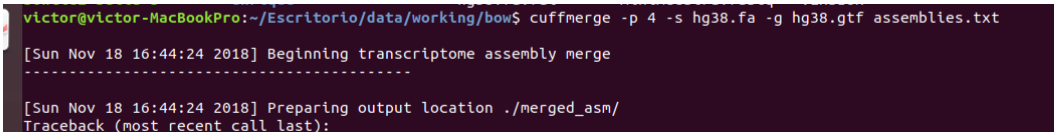
El siguiente paso se realizaría con la herramienta Cuffmerge, y se utilizaría para fusionar los archivos ya ensamblados con la anotación del transcriptoma de referencia y crear una anotación unificada de nuestras muestras para posteriormente, ser analizada.

El primer paso es utilizar los archivos transcripts.gtf generados anteriormente de todas nuestras muestras, en uno solo en formato .txt

Contents of 'assemblies.txt":
cufflinks_salida/transcripts.gtf
cufflinks_salida2/transcripts.gtf

Una vez realizado este paso y con el código:

cuffmerge -p 4 -s hg38.fa -g hg38.gtf assemblies.txt



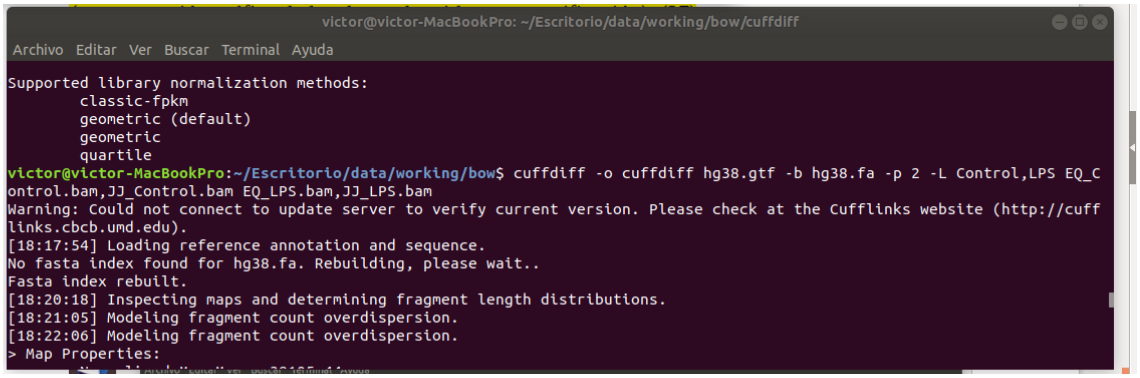
El fichero de salida que produce es un .gtf que contiene un ensamblado que fusiona los ensamblados de nuestras muestras. merged.gtf

Con este paso obtenemos un fichero de ensamblaje .gft que nos sirve para realizar el análisis de expresión diferencial de genes y transcritos de nuestras muestras. Si hubiéramos obviado los pasos de cufflinks y cuffmerge nuestro análisis solo sería de genes con el genoma de referencia.

Expresión génica diferencial entre muestras

Realizaremos nuestro análisis con la herramienta Cuffdiff es un programa es capaz de calcular niveles de expresión de transcritos y genes utilizando el motor de cuantificación Cufflinks explicado anteriormente. Este lo utilizaremos para encontrar genes y transcritos expresados diferencialmente.

Con Cuffdiff utilizamos un archivo .gft de transcritos como entrada en este caso del genoma Humano, junto con nuestros archivos .bam de nuestras muestras alineadas.
El código empleado es:

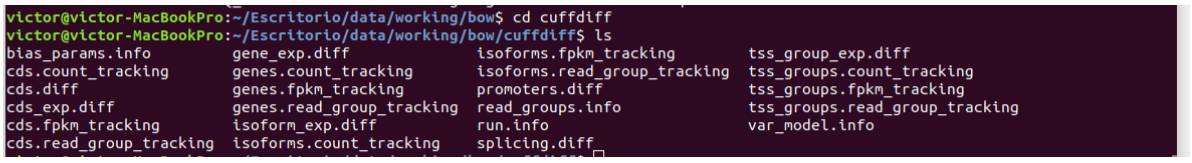


cuffdiff -o cuffdiff/ hg38.gtf -b hg38.fa -p 2 -L Control,LPS EQ_Control.bam,JJ_Control.bam EQ_LPS.bam,JJ_LPS.bam

Explicamos a continuación los parámetros utilizados en nuestro código:

-o: Carpeta donde se guardarán los archivos, en este caso cuffdiff
-p: Número de “threads” para alinear las lecturas. El mínimo por defecto es 1. Utilizamos 2.
-L: Este comando asigna una etiqueta para cada muestra.

Después de correr el comando de Cuffdiff obtendremos los siguientes archivos de salida:



Estos ficheros de salida producidos contienen resultados de las pruebas de los cambios en los niveles de expresión de transcritos, transcritos primarios y genes. También obtenemos un registro de cambios en la abundancia de los transcritos con una misma localización de inicio, y en la abundancia relativa de los transcritos primarios de cada gen. El seguimiento de los primeros permite ver los cambios en la unión, y los segundos ver los cambios en el uso del promotor dentro de un gen.

Archivos esenciales de salida:

•file: read_groups.info : Este archivo nos muestras observamos la etiqueta de nuestras muestras, y se dispone formato tabla. En este archivo cada muestra de entrada .bam tiene su etiqueta, numero de replica, número de lecturas, el número total de lecturas normalizadas y las escalas de normalización.

• file: gene_exp.diff : Éste contiene la información sobre la expresión diferencial, el test estadístico aplicado y si cada uno de los genes del genoma es significativo o no. Aunque tiene un formato texto podemos abrirlo en Excel. La tabla generada contiene:

- Tested_id: Un único identificador que describe el tránscrito, el gen, el tránscrito primario o los CDS que están siendo probados.
- gene_id: El nombre de los genes.
- Locus: Posición del gen dentro del genoma.
- sample_1; Etiquetas de la primera muestra.
- sample_2: Etiquetas de la segunda muestra.
- Status: Puede tomar diversos valores. Ok indica que la prueba ha sido correcta, NOTEST indica que no ha habido suficientes alineamientos para llevarlo a cabo, LOWDATA indica que es demasiado complejo o pobremente secuenciado, HIDATA indica que hay demasiados fragmentos en el locus y FAIL cuando no se produce la prueba.
- log2(fold_change): Logaritom en base 2 del “fold change” x/y.
- test_stat: El valor del test estadístico empleado para calcular la significación de los cambios observados en FPKM.
- p_value: El p-valor estadístico sin corregir.
- q_value : El p-valor ajustado al FDR
- significant: Puede ser “yes” o “no”, en función de si p es mayor que FDR después de la corrección Benjamini-Hochberg.

Este paso fue finalizado en la Fase1, pero en la Fase2 nos decantamos por el uso de EdgeR de Bioconductor, el cual fue el seleccionado para la obtención de los resultados finales. Esta selección fue preferencia del alumno en el uso del programa R.