Aplicación de la biología de sistemas al estudio de la malaria y búsqueda de biomarcadores y dianas terapéuticas

Autora: Mireia Ferrer Almirall
Máster en Bioinformática y Bioestadística
Area 1-Bioinformática farmacéutica

Tutores: Melchor Sánchez Martínez y Alex Sánchez Pla
Profesor responsable de la asignatura: Carles Ventura Royo

02/01/2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Aplicación de la biología de sistemas al estudio de la malaria y búsqueda de biomarcadores y dianas terapéuticas*
Nombre del autor:	*Mireia Ferrer Almirall*
Nombre del consultor/a:	*Melchor Sánchez Martínez y* *Alex Sánchez Pla*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	01/2019
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*1-Bioinformática farmacéutica*
Idioma del trabajo:	*castellano*
Palabras clave	*Malaria, Biología-de-sistemas,* *dianas-terapéuticas*

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

La finalidad de este trabajo es aplicar herramientas de biología de sistemas para investigar los mecanismos implicados en la infección por el parásito de la malaria e identificar posibles biomarcadores y dianas terapéuticas.

Se ha partido de una serie temporal de datos de microarrays del bazo de ratones infectados con dos cepas del parásito (NL y L) para determinar los genes que se encuentran diferencialmente expresados (DEG) respecto a ratones control. A partir de las listas de DEG obtenidas, se han utilizado herramientas de biología de sistemas en combinación con análisis de significación biológica para obtener una visión integrada de los procesos biológicos que se encuentran alterados en la enfermedad e identificar posibles biomarcadores/dianas terapéuticas. Con este fin, se han modelado las redes de interacción proteína-proteína de los DEG y se han identificado los módulos altamente interconectados como posibles unidades funcionales. También se han caracterizado los DEG en base a su patrón de expresión temporal para descubrir módulos corregulados y caracterizar los procesos involucrados en los distintos días post-infección. Por otro lado, se han realizado análisis topológicos para determinar los DEG más influyentes de la red y de los distintos módulos analizados, obteniéndose un listado de posibles candidatos a dianas terapéuticas. Finalmente, se ha realizado un análisis de druggability para validar la lista de candidatos obtenida.

En conjunto, estos análisis ofrecen una visión integrada de los procesos biológicos que determinan el curso de la enfermedad, más allá de los genes

individuales, así como la identificación de posibles dianas terapéuticas y biomarcadores de la enfermedad.

Abstract (in English, 250 words or less):

The purpose of this work is to apply systems biology tools to investigate the mechanisms involved in malaria parasite infection and identify potential biomarkers and therapeutic targets.

We have started from a time series of microarray data of the spleen of mice infected with two strains of the parasite (NL and L) to determine the genes that are differentially expressed (DEG) with respect to control mice. From the DEG lists obtained, systems biology tools have been used in combination with biological significance analyses to obtain an integrated view of the biological processes that are altered in the disease and to identify possible biomarkers/therapeutic targets. To this end, the protein-protein interaction networks of the DEG have been modeled and the highly interconnected modules have been identified as possible functional units. DEG have also been characterized based on their temporal expression pattern to discover co-regulated modules and characterize the processes involved in the different post-infection days. On the other hand, topological analyzes have been carried out to determine the most influential DEG of the network and of the different modules analyzed, obtaining a list of possible candidates for therapeutic targets. Finally, a druggability analysis has been carried out to validate the list of candidates obtained.

Together, these analyzes offer an integrated view of the biological processes that determine the course of the disease, beyond the individual genes, as well as the identification of possible therapeutic targets and biomarkers of the disease.

Índice

Índice de figuras

Índice de cuadros

1. Introducción

1.1. Contexto y justificación del trabajo

La malaria es un problema de salud global asociado a gran morbilidad, principalmente en los países en vías de desarrollo. Está causada por parásitos del género *Plasmodium* que se transmiten a los humanos a través de la picadura de mosquitos *Anopheles* infectados. En los humanos, el parásito se desarrolla silenciosamente en el hígado antes de llegar a la etapa sanguínea, donde la invasión cíclica, el daño y ruptura de los glóbulos rojos conlleva complicaciones clínicas de leves a graves, incluida la muerte. De las cinco cepas que infectan a los humanos, *P. falciparum* y *P. vivax* son responsables de la mayor mortalidad y morbilidad, siendo *P. falciparum* la más letal. La falta de una vacuna eficaz contra la enfermedad, así como la emergencia de cepas resistentes a múltiples fármacos, revelan la necesidad de encontrar nuevas estrategias terapéuticas [1].

Durante las etapas sanguíneas de la infección, el bazo es el principal órgano implicado en el desarrollo de la respuesta inmune y en la eliminación de los eritrocitos parasitados. Sin embargo, los parásitos han desarrollado estrategias para establecer infecciones crónicas a través de la evasión y la modulación de la respuesta inmune y mediante la remodelación del bazo, a veces provocando respuestas inmunes desequilibradas que pueden causar enfermedad grave [2]. Esta doble función del bazo en protección vs. patología pone de relieve la necesidad de comprender el papel del bazo en la infección, así como los factores del huésped y del parásito que determinan el resultado de la infección.

Durante mi doctorado (2008-2012), investigamos la respuesta del bazo en un modelo murino de malaria comparando la infección causada por una cepa letal (L) de malaria, similar a *P. falciparum*, con una no letal (NL), similar a *P. vivax*. Nuestros resultados demostraron la citoadherencia de la cepa NL en el bazo, así como la remodelación del bazo mediante la formación de unas barreras de origen fibroblástico que podrían proteger a los parásitos NL del ataque de los macrófagos [3]. Mediante análisis de expresión génica diferencial en el bazo de ratones infectados con la cepa NL vs L, pudimos identificar un factor de fibroblasto diferencialmente expresado en el bazo de ratones infectados por la cepa NL que podría estar implicado en la remodelación del bazo. En este trabajo me propongo reanalizar estos datos desde una perspectiva de biología de sistemas para tener una visión más amplia e integrada de los procesos/vías que se encuentran alterados en la infección NL con el fin de identificar mecanismos clave en el curso de la enfermedad y posibles dianas terapéuticas.

1.2. Objetivos

El objetivo principal de este trabajo es aplicar herramientas de biología de sistemas para investigar los procesos biológicos implicados en la infección por el parásito de la malaria `NL` vs `L`, con el fin de identificar posibles dianas terapéuticas y entender los mecanismos involucrados en la patogénesis. A continuación se listan los objetivos específicos:

Objetivo 1

Modelado de redes de interacción proteína-proteína en la malaria `NL` y `L`

Objetivo 2

Identificación de los mecanismos/vías que determinan el resultado de la enfermedad y las posibles dianas terapéuticas

1.3. Enfoque y método seguido

El estudio de las enfermedades ha experimentado recientemente un cambio de paradigma debido a la explosión de datos "-ómicos" generados en las últimas décadas y el desarrollo de enfoques más sistémicos para el análisis integrativo de los datos, más allá del análisis a nivel de un solo gen/proteína. Estos enfoques, entre los que destacan la biología de sistemas y los análisis de significación biológica, mejoran nuestra comprensión de los

sistemas biológicos en estados fisiológicos y de enfermedad, y permiten el descubrimiento de biomarcadores y nuevas dianas terapéuticas [4, 5].

En este trabajo, me he basado en datos de expresión temporal del bazo de ratones infectados con dos cepas del parásito de la malaria (NL y L) para determinar los genes que se encuentran diferencialmente expresados (DEG) en la enfermedad. A partir de las listas de DEG obtenidas, se han modelado las redes de interacción proteína-proteína de los DEG con *Cytoscape* (https://cytoscape.org/), una herramienta de libre acceso que permite crear y visualizar redes de interacción molecular y vías biológicas e integrar estas redes con anotaciones, perfiles de expresión génica y otros datos. También se han explorado diferentes aplicaciones de *Cytoscape* como `clusterMaker` [6], `TiCoNe` [7] y `ClueGO` [8] para identificar y caracterizar las vías/módulos de la red que determinan el curso de la enfermedad así como las posibles dianas terapéuticas. Para determinar el potencial de estos últimos a ser atacados por fármacos, se ha realizado una búsqueda exhaustiva de los posibles candidatos en bases de datos como *DrugBank* (https://www.drugbank.ca/) y se ha calculado su *druggability* con la herramienta *DrugEBIlity* del EMBL (https://www.ebi.ac.uk/chembl/drugebility).

El informe se ha elaborado en *Rmarkdown* (https://rmarkdown.rstudio.com/).

1.4. Plan de trabajo

A continuación se detalla el plan de trabajo seguido con el desglose de tareas planeadas para alcanzar los objetivos propuestos.

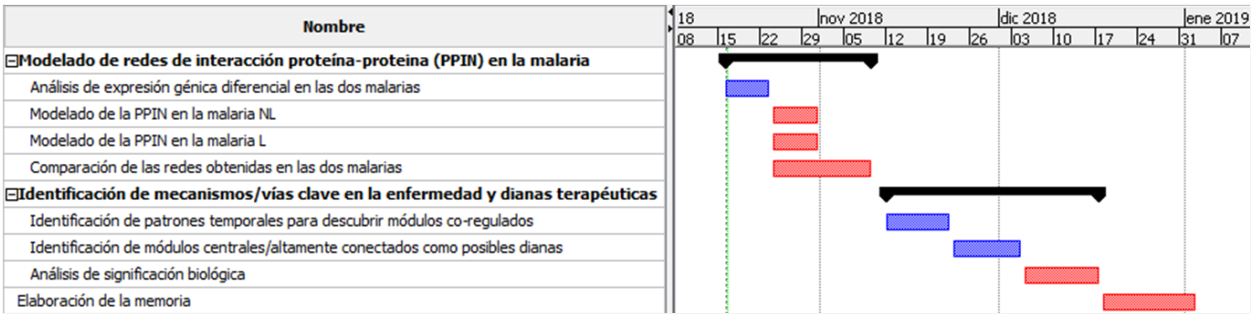


Figura 1: Plan de trabajo

1.5. Sumario de los resultados obtenidos

Se han identificado los principales procesos biológicos involucrados en la malaria NL y L así como posibles biomarcadores y dianas terapéuticas contra las que actuar para modular el curso de la enfermedad.

1.6. Descripción de los capítulos de la memoria

Los capítulos de la memoria se han estructurado en tres partes principales: antecedentes, resultados y conclusión. La metodología utilizada para generar los resultados se describe dentro de su capítulo respectivo. Además, al inicio de cada capítulo se incluye un pequeño párrafo introductorio con la descripción del capítulo y un apartado final con las conclusiones específicas del capítulo, donde se discuten también sus limitaciones y posibles variaciones.

A continuación se detalla el contenido de los capítulos contenidos en las diferentes secciones:

I. Antecedentes (sección 2): Breve revisión bibliográfica sobre los aspectos de interés para el trabajo, como son la biología de sistemas, los análisis de significación biológica y el descubrimiento de biomarcadores/dianas terapéuticas. Los aspectos clave de la enfermedad de estudio están descritos en el apartado 1.1, por lo que no se repiten aquí.

II. Resultados:

- Sección 3.1: Análisis de expresión diferencial en los bazos de ratones infectados con las cepas NL y L respecto a los ratones no infectados (control).
- Sección 3.2: Caracterización de la malaria NL mediante redes de interacción proteína-proteína, con diferentes subapartados correspondientes a los distintos análisis realizados (modelado de la red, identificación de nodos centrales, análisis de módulos topológicos/temporales y análisis de significación biológica).
- Sección 6.1: Caracterización de la malaria L. Esta sección se incluye como anexo para facilitar la lectura de la memoria, ya que los análisis realizados son esencialmente los mismos que los realizados para la malaria NL.
- Sección 3.3: Comparación entre las dos malarias.
- Sección 3.4: Selección de posibles candidatos a biomarcadores y/o dianas terapéuticas en la malaria y análisis de su *druggability*.

III. Conclusión (sección 4): Conclusiones generales del trabajo.

2. Antecedentes

2.1. Biología de sistemas

La biología de sistemas tiene como objetivo comprender las entidades biológicas a nivel sistémico, analizándolas no sólo como componentes individuales, sino también como sistemas interactivos y sus propiedades emergentes. Relacionado con esto está la biología de redes, que permite la representación y el análisis de sistemas biológicos utilizando herramientas derivadas de la teoría de grafos [9].

Los sistemas complejos se pueden ver como *redes* en las que sus componentes se pueden representar como *nodos* y se vinculan a través de sus interacciones, que se denominan *ejes*. Los nodos pueden representar diferentes entidades (por ejemplo, genes o proteínas) y los ejes transmiten la información sobre cómo se vinculan los nodos (por ejemplo, interacción física, coexpresión, etc.). Este tipo de representación matemática permite integrar diferentes capas de datos ómicos, lo que facilita la extracción de información nueva y fisiológicamente relevante de los datos.

2.1.1. Propiedades de las redes

Las redes, también denominadas *grafos*, tienen algunas propiedades que son muy útiles para revelar la información que contienen. El análisis de las características topológicas de una red, es decir, la forma en que se organizan los nodos y los ejes dentro de ésta, permite identificar componentes y subestructuras relevantes que pueden ser de importancia biológica [9–11]. En la siguiente tabla se detallan las principales propiedades de las redes y sus componentes:

Propiedades	Descripción
Nodos	
Grado	número de ejes que se conectan a un nodo.
Camino más corto	distancia más corta entre dos nodos cualesquiera, medida en el número de ejes que los conecta.
Centralidad de intermediación	número de caminos más cortos que pasan por un nodo de entre todos los caminos más cortos que conectan todos los pares de nodos posibles.
Tipo	información representada por los nodos (por ejemplo: genes, proteínas, metabolitos, anotación de la ontología de genes, etc.).
Peso	valor cuantitativo asociado a un nodo, por ejemplo, el nivel de expresión de un gen.
Ejes	
Centralidad	estima la importancia de un eje para la conectividad o el flujo de información de la red. Como en los nodos, existen diferentes medidas de centralidad.
Tipo	indica la relación entre el par de nodos unidos por el eje (por ejemplo, interacción física, coexpresión, similitud, etc).
Direccionalidad	indica el sentido del flujo de la información a través de este eje. Esta puede ser no dirigida (conexión simple, sin un flujo dado) o dirigida (implica un flujo de señal y la red se puede organizar jerárquicamente).
Peso	valor cuantitativo asociado al eje. Se utiliza para describir conceptos como la confiabilidad de una interacción, el cambio de expresión cuantitativo que un gen induce sobre otro o incluso qué tan estrechamente relacionados están dos genes en términos de similitud de secuencia.
Red	
Camino más corto característico	camino más corto promedio entre todos los pares de nodos que componen la red.
Diámetro	camino más corto de mayor longitud.
Distribución de la conectividad	distribución de frecuencias del grado de los nodos que componen la red.
Centralización	medida que muestra si la topología de la red es en forma de estrella (hay pocos nodos con grado alto y muchos con grado bajo) o si, por el contrario, los nodos de la red tienen en promedio el mismo grado.
Coeficiente de agrupamiento	medida de la tendencia de los nodos a agruparse.
Modularidad	representa los módulos topológicos o clústeres de la red, que son grupos de nodos estrechamente interconectados, o dicho de otra manera, áreas densas de conectividad separadas por regiones de baja conectividad.

Los nodos con alto grado se denominan *hubs* y su eliminación tiene un gran impacto en la topología de la red. Además, las medidas de centralidad proporcionan una estimación de la importancia de un nodo/eje para la conectividad o el flujo de información de la red. Los nodos con una alta centralidad de intermediación son interesantes porque se encuentran en las rutas de comunicación y pueden controlar el flujo de información. Estos nodos pueden representar proteínas importantes en las vías de señalización y pueden formar dianas para el descubrimiento de fármacos [11].

La mayoría de las redes biológicas moleculares son redes centralizadas, libres de escala, donde la distribución del grado de los nodos es exponencial. Este tipo de topología hace que las redes sean robustas a perturbaciones aleatorias, pero a su vez, la afectación de los *hubs* conlleva graves consecuencias para la red. Por otro lado, las redes biológicas tienen un coeficiente de agrupamiento promedio significativamente más alto en comparación con las redes aleatorias, lo que demuestra su naturaleza modular [9–11]. Encontrar estos clústeres/módulos es muy importante, ya que pueden reflejar unidades funcionales (p. ej. complejos proteicos) relevantes para la biología de la red y ayudan a reducir su complejidad.

2.1.2. Tipos de redes en biología

Diferentes tipos de datos pueden ser representados mediante redes donde los nodos y ejes integran múltiples capas de información. El tipo de datos determinará las características globales de la red y debe por lo tanto tenerse en cuenta durante el análisis. Algunos de los tipos más comunes de redes biológicas son: (i) las redes de interacción proteína-proteína, (ii) las redes de coexpresión, (iii) las redes de señalización celular y (iv) las

redes metabólicas. Mientras que las dos primeras son no dirigidas, en las redes de regulación y metabólicas los ejes son dirigidos y el signo de las flechas puede usarse para indicar estimulación o inhibición. A continuación se ofrece una descripción más detallada de los dos primeros tipos de redes, ya que son los más pertinentes para este trabajo.

Redes de interacción proteína-proteína (PPIN)

Las PPIN son una representación matemática de las interacciones específicas entre proteínas. Los nodos representan las proteínas y los ejes reflejan la presencia de una interacción entre éstas. Estas interacciones pueden ser transitorias (p. ej. quinasas o transportadores) o estables (ej. complejos proteicos) y sirven diferentes funciones biológicas. Los análisis de PPIN son cruciales para comprender la fisiología celular en estados normales y de enfermedad y se utilizan para el desarrollo de fármacos dirigidos a bloquear las interacciones entre proteínas diana. Además, se pueden usar para asignar roles putativos a proteínas no caracterizadas, caracterizar vías de señalización o caracterizar las relaciones entre proteínas que forman complejos multimoleculares [12].

Existen diferentes bases de datos como *String* (http://string-db.org) o *IntAct* (http://www.ebi.ac.uk/intact/) que proporcionan la información de las interacciones entre las proteínas para diferentes organismos. Esta información puede provenir de evidencias experimentales obtenidas con tecnologías de alto rendimiento tales como la purificación por afinidad unida a espectrometría de masas o los ensayos de dos híbridos en levadura; de sets de datos curados manualmente disponibles en la literatura, o de predicciones computacionales basadas en ortólogos o información estructural de proteínas [13]. Generalmente, las interacciones van acompañadas de un índice o *score* que mide la confiabilidiad de que dicha interacción represente una interacción biológica "real" [14].

En cuanto a sus propiedades topológicas, las PPIN son redes no dirigidas y se caracterizan principalmente por tener un diámetro pequeño (~ 6 ejes de longitud), un coeficiente de agrupamiento elevado (contienen comunidades de nodos que están más conectados internamente que el resto de la red) y ser libres de escala (contienen un pequeño número de nodos centrales o *hubs*) [9]. Dadas sus características, los métodos topológicos más utilizados para analizar las PPIN se basan en las medidas de centralidad para identificar los *hubs* y en la detección de clústeres que puedan representar complejos de proteínas y/o maquinarias celulares.

Redes de coexpresión

Una red de coexpresión identifica qué genes tienen una tendencia a mostrar un patrón de expresión coordinado a través de un grupo de muestras. En este tipo de red, cada nodo representa un gen y cada eje representa la presencia y la fuerza de la relación de coexpresión, que se define en base a medidas de correlación. También se pueden representar mediante matrices de similitud gen-gen. Los análisis de *clustering* basados en *k-means* o *clustering jerárquico* permiten identificar módulos de genes coexpresados en las distintas muestras, que pueden utilizarse para inferir sobre la funcionalidad de estos genes en un proceso biológico o para identificar unidades transcripcionales reguladas [15].

2.2. Análisis de significación biológica

El análisis de significación biológica, o análisis de enriquecimiento funcional, es uno de los métodos más populares para comprender el contexto biológico de las redes de interacción proteína-proteína. Aunque no es estrictamente una herramienta de análisis de redes, a menudo se utiliza en combinación con ésta para ayudar a caracterizar la red o subconjuntos de la misma [10].

Este tipo de análisis se basa en la información disponible en bases de datos como *Gene Ontology* (*GO*, http://www.geneontology.org/) o *Reactome* (https://reactome.org/) para inferir qué procesos biológicos/vías metabólicas están sobrerrepresentadas en una lista de genes/proteínas. Las anotaciones pueden ser asignadas por un curador humano, que realiza una anotación cuidadosa y manual, o mediante predicciones computacionales. La base principal del análisis de enriquecimiento es que si un proceso biológico es anormal en un estudio dado, los grupos de genes que funcionan conjuntamente en ese proceso deberían tener una mayor probabilidad de ser seleccionados como un grupo relevante por las tecnologías de detección de alto rendimiento.

Este enriquecimiento puede medirse cuantitativamente mediante tests estadísticos como los basados en la distribución hipergeométrica [16].

Aunque estos análisis proporcionan valiosa información para ayudar a la interpretación biológica/funcional de grandes listas de genes, también presentan ciertas limitaciones. Por un lado, ciertas áreas de la biología están más anotadas y mejor descritas que otras, proporcionando más detalle y términos más precisos para los procesos mejor conocidos. Esto introduce un cierto sesgo en el análisis estadístico. Por otro lado, la complejidad y el detalle de las anotaciones asociadas a grandes conjuntos de genes conlleva la generación de redes de términos interrelacionados y similares extremadamente complicadas y abrumadoras [16]. Existen diferentes métodos para tratar de reducir la complejidad y redundancia de los resultados de enriquecimiento. El más simple es utilizar ontologías simplificadas, como *GOslims*, donde los términos detallados son eliminados y asignados a términos parentales más amplios y generales. Otras herramientas, como las aplicaciones de Cytoscape *BiNGO* o *ClueGO* (https://apps.cytoscape.org/), permiten representar los resultados del enriquecimiento como una red de términos, donde los ejes dirigidos representan las relaciones jerárquicas de los términos. Además, las aplicaciones *ClueGO* y *EnrichmentMap* permiten agrupar los términos por similaridad mediante el cálculo del coeficiente *kappa* de Cohen [8], facilitando así la interpretación de los resultados.

2.3. Aplicación de la biología de sistemas al descubrimiento de biomarcadores y dianas terapéuticas

La biología de sistemas proporciona información valiosa sobre las moléculas y procesos biológicos relevantes en la enfermedad, lo que a su vez permite el descubrimiento de mejores biomarcadores y dianas terapéuticas.

Los biomarcadores son parámetros medibles, ya sean bioquímicos, fisiológicos o morfológicos, que se asocian a un determinado estado biológico; pudiéndose utilizar como indicadores de procesos biológicos normales, patogénicos o respuestas farmacológicas a una intervención terapéutica [17]. Así pues, el descubrimiento de biomarcadores es de gran interés, tanto en investigación básica para caracterizar el objeto de estudio, como en clínica para el pronóstico/diagnóstico de enfermedades o en campos como el de la farmacogenómica [5]. En los últimos años, el uso de tecnologías ómicas de alto rendimiento ha llevado al rápido descubrimiento de muchos biomarcadores candidatos; sin embargo, estos pueden ser difíciles de validar y requieren diferentes niveles de validación dependiendo de su uso previsto [18]. Por otro lado, el análisis integrativo de diferentes ómicas mediante redes permite el descubrimiento de biomarcadores mejores y más precisos que permitan monitorear la integridad funcional de la red perturbada por las enfermedades, llevando a una mejor clasificación de las enfermedades y allanando el camino a terapias personalizadas [4].

El término diana terapéutica se usa frecuentemente en investigación farmacéutica para describir la molécula del organismo (p.ej. proteína o ácido nucleico) donde un fármaco ejerce su acción. La identificación de dianas suele ser el punto de partida en el descubrimiento de fármacos. Para que una molécula se convierta en diana terapéutica, no sólo debe ser relevante en la enfermedad: también debe ser específica de la condición de estudio, y accesible y modulable con fármacos basados en moléculas pequeñas u otros reactivos para lograr un efecto terapéutico deseado (lo que se conoce como *druggable*) [19]. Sin embargo, se cree que sólo el 10-15 % de todos los genes humanos son *druggable* (p. ej. codifican proteínas similares en secuencia a las que ya han sido atacadas con moléculas pequeñas), y que el solapamiento entre los genes *druggable* y los relacionados con enfermedades conocidas es sólo del 25 % [19, 20]. Existen estrategias para tratar de ampliar este rango. Por un lado, la optimización de las librerías basadas en compuestos sintéticos, productos naturales y otros tipos de moléculas puede permitir alcanzar proteínas que se consideraban *undruggable* con compuestos tradicionales. Por otro lado, los análisis basados en la interconectividad funcional de las redes intracelulares permiten encontrar dianas alternativas (en sentido ascendente, descendente o en paralelo al gen causante de la enfermedad) que sean *druggable* y donde su modulación influya indirectamente en el proceso de la enfermedad [19].

3. Resultados

3.1. Análisis de expresión génica diferencial en las dos malarias

En este capítulo, se analizan los datos de los microarrays de bazos de ratones infectados con la malaria `NL` o `L` realizados durante mi doctorado. El objetivo de este análisis es obtener listas de genes diferencialmente expresados (DEG) en las malarias `NL` y `L` respecto a los ratones no infectados (`Ctrl`). Estas listas se utilizarán para modelar las redes de interacción proteína-proteína (PPIN) en las dos malarias y para los análisis de significación biológica.

3.1.1. Datos de partida

Partimos de la serie temporal de microarrays `GSE17603` [3], disponible en la base de datos de *Gene Expression Omnibus* (GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17603).

Se trata de microarrays de dos colores ("Agilent Whole Mouse Genome G4122A") del bazo de ratones infectados con las cepas `NL` y `L` a días 3, 4 y 5 post-infección, así como del bazo de ratones no infectados (`Ctrl`). Cada array ha sido cohibridado con una muestra `NL` marcada con *Cy3* y con una muestra `L` marcada con *Cy5*, para los distintos días post-infección. A continuación se muestra la información fenotípica de las muestras:

Cuadro 1: Información fenotípica de las muestras

	GEO_Access	Array	Strain	Time	Dye	SampleName
1	GSM439422	251269423040	Ctrl	0	Cy3	Ctrl1_Cy3
2	GSM439423	251269423042	NL	3	Cy3	NL_d3_Cy3
3	GSM439424	251269423043	NL	4	Cy3	NL_d4_Cy3
4	GSM439425	251269423044	NL	5	Cy3	NL_d5_Cy3
7	GSM439428	251269423040	Ctrl	0	Cy5	Ctrl2_Cy5
8	GSM439429	251269423042	L	3	Cy5	L_d3_Cy5
9	GSM439430	251269423043	L	4	Cy5	L_d4_Cy5
10	GSM439431	251269423044	L	5	Cy5	L_d5_Cy5

3.1.1.1. Normalización de los datos

Partimos de la matriz de valores normalizados disponible online, que contiene un total de 35088 genes anotados. Para la normalización, se utilizó el paquete de Bioconductor `limma` utilizando la corrección de fondo ("normexp") y la normalización por cuantiles entre las matrices para los valores finales (las intensidades de señal de *Cy3* y *Cy5* se normalizaron por separado como matrices de un solo color) [3].

3.1.2. Control de calidad

Como parte del control de calidad, analizamos la distribución de los datos normalizados mediante diagramas de cajas y gráficos de densidad para las distintas muestras:

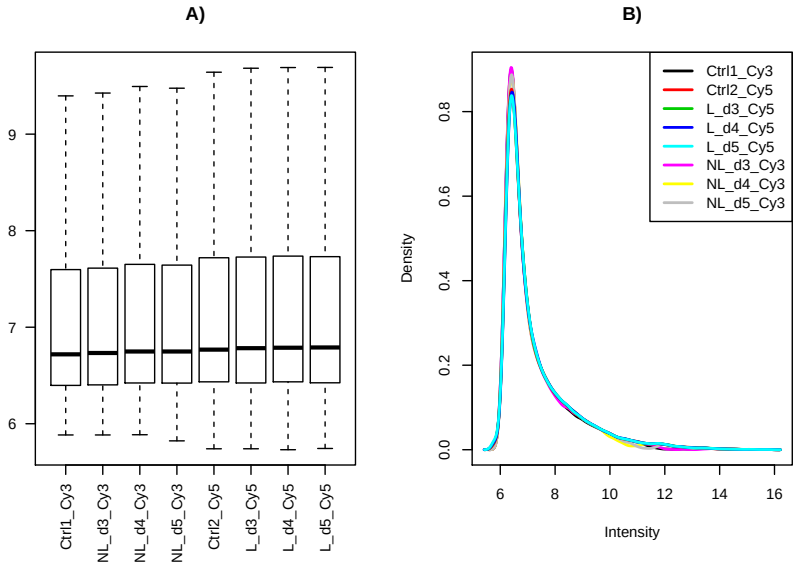


Figura 2: Diagrama de cajas (A) y gráficos de densidad (B) de los datos normalizados

En general, la distribución de la señal es homogénea, sugiriendo que la normalización ha funcionado bien. Se observan ligeras diferencias en la distribución de la señal de las muestras marcadas con *Cy3* y las marcadas con *Cy5*, probablemente debido al mayor ruido de fondo asociado al marcaje con *Cy5*. Sin embargo, como las comparaciones se realizarán separadamente para cada canal, no se espera que estas ligeras diferencias afecten al resultado.

A continuación analizamos la distribución de las muestras mediante clustering jerárquico y análisis de componentes principales (PCA).

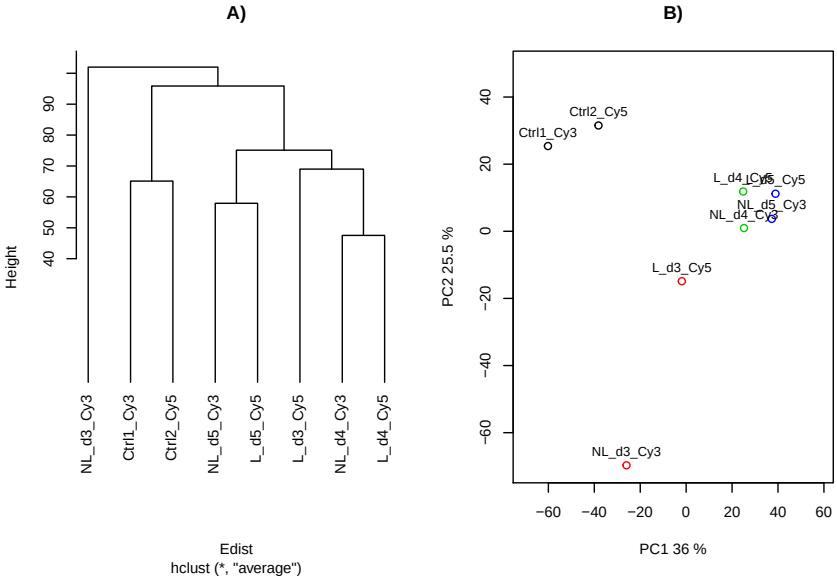


Figura 3: Dendograma (A) y gráfico de las 2 primeras componentes del PCA (B) para los datos normalizados

Se observa una distribución de las muestras agrupadas por días post-infección, correspondientes a los distintos arrays. Por otro lado, la muestra `NL_d3_Cy3` aparece separada del resto, sugiriendo grandes diferencias en esta muestra.

3.1.3. Selección de genes diferencialmente expresados

Con el fin de encontrar los genes diferencialmente expresados (DEG) en la malaria `NL` y `L` con respecto a los controles (`Ctrl`) en los diferentes días post-infección, realizamos las siguientes comparaciones:

1. Cambios en los distintos días post-infección en la malaria `NL`:
   - NLvsCtrl_d3 = NL_d3_Cy3 - Ctrl1_Cy3
   - NLvsCtrl_d4 = NL_d4_Cy3 - Ctrl1_Cy3
   - NLvsCtrl_d5 = NL_d5_Cy3 - Ctrl1_Cy3
2. Cambios en los distintos días post-infección en la malaria `L`:
   - LvsCtrl_d3 = L_d3_Cy5 - Ctrl2_Cy5
   - LvsCtrl_d4 = L_d4_Cy5 - Ctrl2_Cy5
   - LvsCtrl_d5 = L_d5_Cy5 - Ctrl2_Cy5

Aunque los arrays son de dos colores, se realiza un análisis por colores separados. Esto permitirá comparar las distintas malarias por separado con respecto a los controles y evitar sesgos debidos a los diferentes marcajes utilizados.

Utilizamos el paquete de Bioconductor `limma` [21], que se basa en la construcción de modelos lineales para calcular las diferencias de expresión en las distintas comparaciones. Con sólo una muestra por grupo, no es posible obtener una medida estadística para estimar la significancia de los DEG encontrados (no hay grados de libertad), de manera que sólo podrán calcularse los cambios de expresión (*log2-fold-changes* (logFC)) para los distintos contrastes, contenidos en los coeficientes del modelo lineal.

Las siguiente tablas resumen muestran el número de genes up/down-regulados (logFC absoluto mayor o igual a 1) para los distintos contrastes realizados:

Cuadro 2: Recuento de genes up/down en los distintos contrastes

	NLvsCtrl_d3	**NLvsCtrl_d4**	**NLvsCtrl_d5**	**LvsCtrl_d3**	**LvsCtrl_d4**	**LvsCtrl_d5**
Down	707	692	871	356	560	803
0	33453	33204	33060	34399	34142	33793
Up	928	1192	1157	333	386	492

Como resultado del análisis obtenemos las listas de DEGs y sus logFC para las dos comparaciones realizadas. Estas listas servirán de *input* para modelar las redes de interacción proteína-proteína en las dos malarias. A continuación se muestra el número de genes obtenido con estos criterios para cada comparación:

- NL vs Ctrl a días 3,4,5 p.i.: 3417 genes
- L vs Ctrl a días 3,4,5 p.i.: 1671 genes

3.1.4. Comparación entre los días post-infección

Para visualizar los genes comunes en los distintos días post-infección, realizamos diagramas de Venn con los DEGs seleccionados para cada contraste:

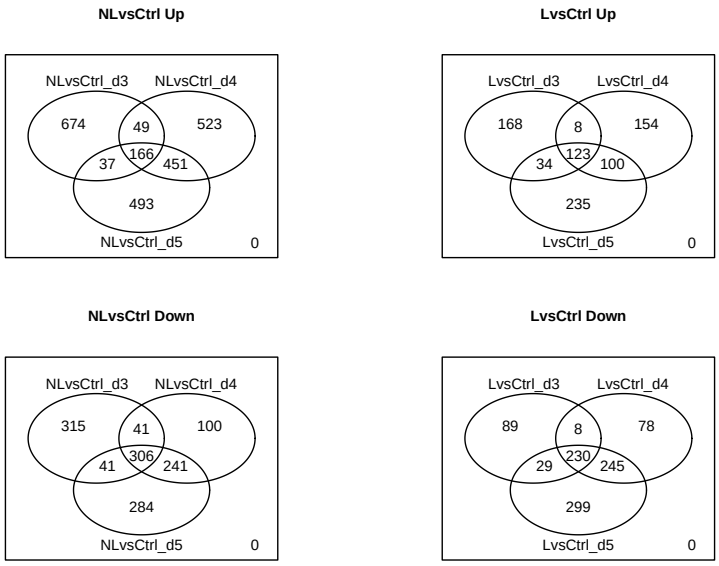


Figura 4: Diagramas de Venn para las comparaciones temporales en NLvsCtrl y LvsCtrl.

En general, se observan pocos genes up-regulados comunes entre los tres días post-infección.

3.1.5. Perfiles de expresión

Finalmente, realizamos un *heatmap* con los genes con logFC absoluto mayor o igual a 1 en alguno de los días en NLvsCtrl o en LvsCtrl.

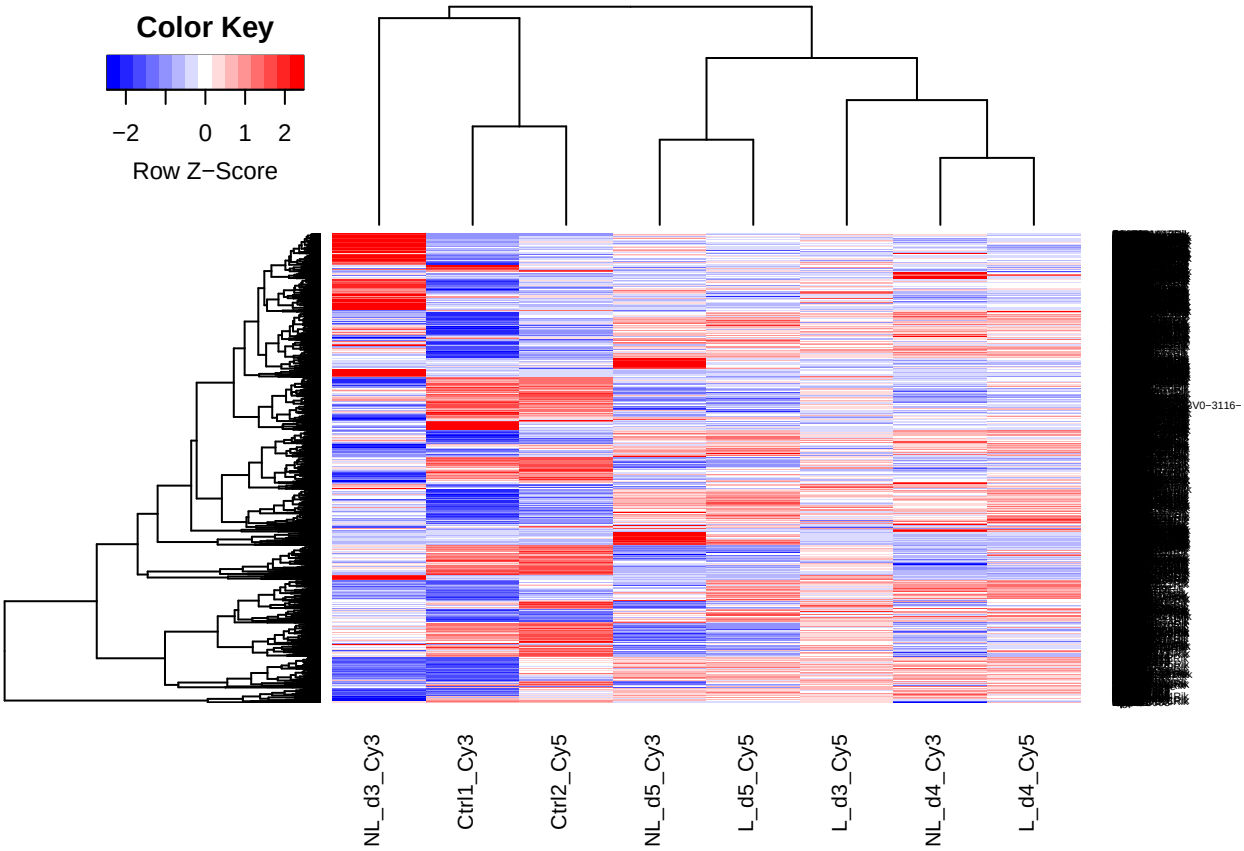


Figura 5: Perfiles de expresión de los genes seleccionados.

Los perfiles de expresión obtenidos muestran tres patrones diferenciados, uno con las muestras `NL_d3_Cy3`, `Ctrl1_Cy3` y `Ctrl2_Cy5`; otro con las muestras `L_d3_Cy5`, `NL_d4_Cy3` y `L_d4_Cy5`, y un tercer grupo con las muestras `NL_d5_Cy3` y `L_d5_Cy5`. Esto refleja las diferencias de expresión entre los diferentes días post-infección, que parecen conservarse entre las dos malarias. De interés, a día 3 post-infección sí se observan diferencias en los perfiles de expresión de la malaria NL vs L, donde `NL_d3_Cy3` muestra un perfil más similar a los controles y `L_d3_Cy5` tiene un perfil más similar a las muestras de día 4 post-infección.

3.1.6. Conclusiones

- Tanto el clustering jerárquico como el PCA muestran una agrupación de las muestras por día post-infección. La muestra `NL_d3_Cy3` aparece separada del resto, sugiriendo mayores diferencias en esta muestra. Dada la limitación en el número de muestras, se decide incluir todas las muestras en el análisis.
- Para cada comparación múltiple, se han seleccionado los genes con logFC absoluto mayor o igual a 1 para alguno de los días post-infección. Con estos criterios, se han obtenido 3417 genes diferencialmente expresados (DEGs) en la comparación `NLvsCtrl`, y 1671 DEGs en la comparación `LvsCtrl`. Estas listas servirán de *input* para el modelado de redes.
- Los diagramas de Venn muestran pocos genes up-regulados comunes entre los distintos días post-infección para cada malaria, sugiriendo cambios de expresión en el curso de la infección.
- Los perfiles de expresión o *heatmaps* muestran perfiles de expresión similares para las dos malarias en los días 4 y 5 post-infección, mientras que las muestras a día 3 muestran mayores diferencias.

Limitaciones y posibles variaciones

La falta de réplicas biológicas, así como de arrays con marcajes invertidos, son las principales limitaciones de los datos de partida. Aunque el análisis de microarrays carece de una validación estadística, puede ser interesante explorar estos datos con técnicas de análisis más integrativas, como son las aproximaciones de biología de sistemas o análisis de significación biológica, donde se tengan en cuenta unidades más grandes que el gen individual.

Por otro lado, existen sets de datos similares en la web de GEO que podrían ser analizados en paralelo y comparados con los datos del estudio para aumentar la robustez del análisis.

3.2. Caracterización de la malaria NL mediante redes de interacción proteína-proteína (PPIN)

En este capítulo se realiza un análisis de redes a partir de los datos de expresión diferencial obtenidos en el capítulo anterior. El objetivo de este análisis es pasar de una lista de genes individuales a visualizarlos en su conjunto dentro de un contexto de interacción. De esta manera se ponen de relevancia qué genes son más influyentes dentro del conjunto, así como la presencia de subconjuntos regulados entre sí que puedan representar ciertas funciones biológicas.

Existen diferentes herramientas para la representación y visualización integrativa de los datos en forma de red, como son Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/), GraphViz (http://www.graphviz.org/) o Cytoscape (https://cytoscape.org/). Para este trabajo me he basado en Cytoscape, una herramienta bioinformática de libre acceso que permite la visualización y análisis de redes moleculares, así como la integración de datos de expresión génica o de otros tipos. Se compone de una gran variedad de plugins o aplicaciones que permiten abordar diferentes tipos de análisis.

3.2.1. Modelado de la PPIN en la malaria NL

Para el modelado de la red de interacciones proteína-proteína de la malaria NL, partimos de la lista de DEGs obtenidos en el análisis de expresión diferencial de microarrays para la comparación NLvsCtrl. Utilizamos la aplicación de Cytoscape `stringApp` para importar los datos de interacción proteína-proteína contenidos en la base de datos *String* (https://string-db.org/) para los DEG seleccionados. Las interacciones en *String* provienen de diferentes fuentes de evidencia, tales como estudios experimentales de interacciones físicas, coexpresión en conjuntos de datos públicos, co-cita en la literatura y evidencia de interacción funcional o física extraída de bases de datos públicas. Cada interacción tiene asociado un *score* combinado que indica la probabilidad de que dicha interacción sea "verdadera" [22]. Seleccionamos las interacciones de ratón (*Mus musculus*) con un *score* combinado >=0.4, que refleja un nivel de confianza medio o superior y es el generalmente utilizado en la literatura.

A continuación se muestra la red de interacciones obtenida a partir de la lista de DEGs seleccionados en la malaria NL, según la información disponible en la base de datos de *String*.

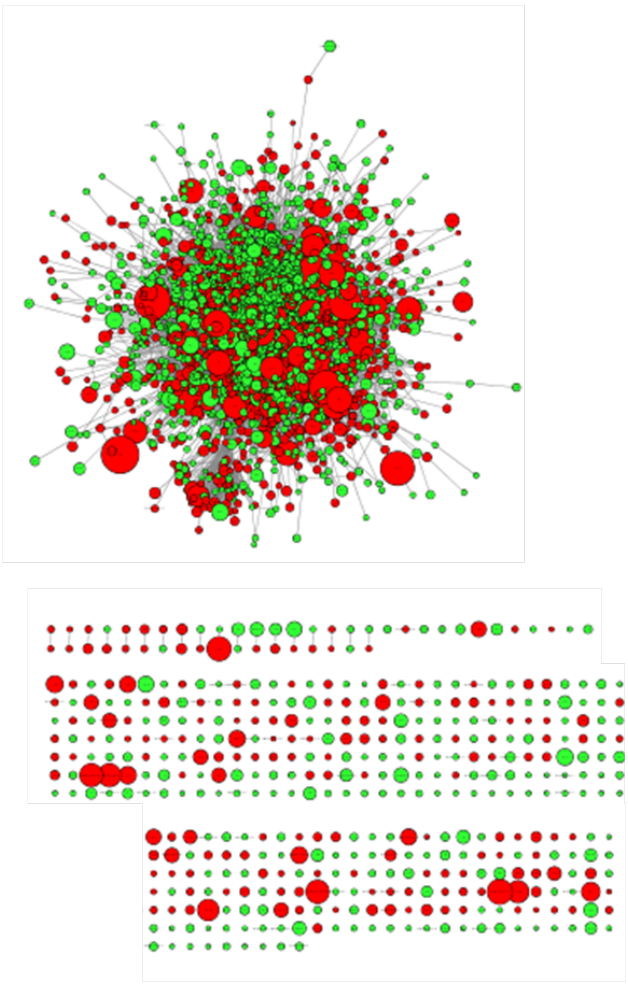


Figura 6: PPIN de los genes diferencialmente expresados en la malaria NLvsCtrl

Obtenemos una red con 3233 nodos (genes) y 36757 ejes (interacciones), de los cuales sólo 5492 interacciones han sido validadas experimentalmente con un *score*>=0.4. En verde representamos los 2058 genes sobreexpresados (logFC>=1 para alguno de los días post-infección) y en rojo los 1188 genes subexpresados (logFC<=-1 para alguno de los días post-infección). El tamaño de los nodos corresponde a la variación de la expresión acumulada en los 3 días. Se observa una componente conectada principal de 2832 nodos y 36758 ejes, que es la que utilizaremos para los subsiguientes análisis.

A continuación se muestran algunos descriptivos globales de la red:

Parámetro	Valor
Núm. nodos	2832
Núm. ejes	36758
Grado máximo	366
Diámetro	9
Centralización	0.122
Camino más corto característico	3.2

La distribución del grado de los nodos sigue una distribución exponencial (figura 7), correspondiente a una topología "libre de escala" como se esperaría en redes biológicas.

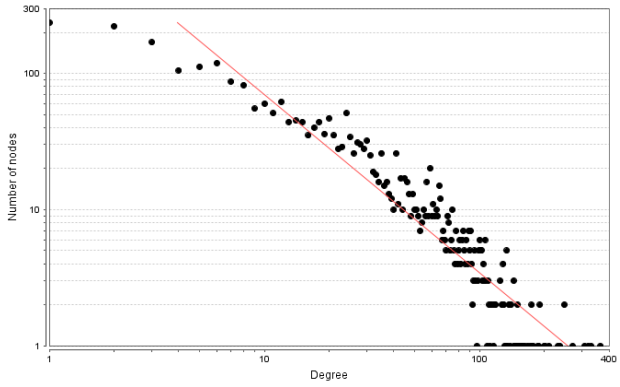


Figura 7: Distribución del grado de los nodos de la PPIN NLvsCtrl

3.2.2. Identificación de nodos centrales o *hubs* en la malaria NL

Para identificar genes centrales, utilizamos las medidas de centralidad, principalmente la centralidad de intermediación o *betweenness centrality*. Los genes con una alta centralidad de intermediación son importantes como conectores de camino más corto a través de una red.

La figura 8 muestra la distribución del grado en función de la medida de centralidad de intermediación. Se observa que, en general, los nodos con grado más alto se corresponden con los de mayor centralidad.

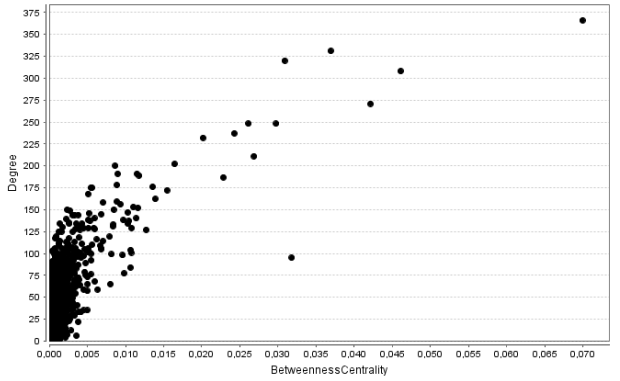


Figura 8: Gráfico de dispersión del grado de los nodos en función de su centralidad en NLvsCtrl

Por otro lado, como se observa en la figura 9, los nodos de mayor centralidad no coinciden con los de mayor variación acumulada a lo largo de los días post-infección.

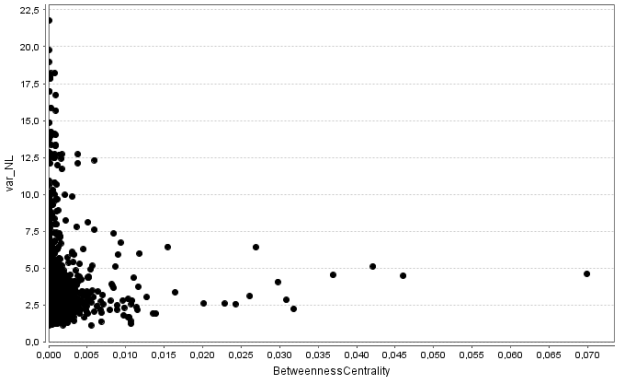


Figura 9: Gráfico de dispersión de la centralidad de los nodos en función de la variación acumulada en NLvsCtrl

A continuación creamos una subred con los 20 nodos de mayor grado/centralidad. El grosor de los ejes indica el score calculado para la evidencia experimental de la interacción.

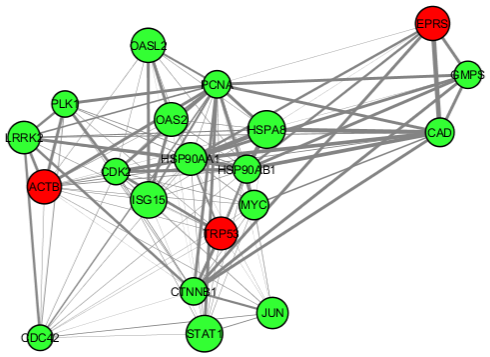


Figura 10: Subred con los 20 nodos con grado más alto (NLvsCtrl)

Nótese que los nodos centrales no tienen por qué estar altamente conectados entre ellos.

En la siguiente tabla se detalla la información de los nodos seleccionados, entre los que se encuentran varios oncogenes y *heat shock proteins*:

Cuadro 4: Lista de nodos con grado más alto en NLvsCtrl

Entrez	Symbol	description	Degree	Centrality	logFC.d3	logFC.d4	logFC.d5	var
22059	TRP53	transformation related protein 53	366	0.070	-0.990	-1.682	-1.946	4.617
15519	HSP90AA1	heat shock protein 90, alpha (cytosolic), class A member 1	331	0.037	0.271	2.146	2.117	4.534
15516	HSP90AB1	heat shock protein 90 alpha (cytosolic), class B member 1	320	0.031	-0.252	1.621	0.967	2.840
66725	LRRK2	leucine-rich repeat kinase 2	308	0.046	1.499	1.607	1.376	4.482
11461	ACTB	actin, beta	271	0.042	-0.702	-2.402	-2.013	5.118
16476	JUN	Jun oncogene	248	0.030	1.273	1.585	1.218	4.077
69719	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	248	0.026	0.949	1.117	1.037	3.103
229363	GMPS	guanine monophosphate synthetase	237	0.024	0.590	1.037	0.948	2.574
18538	PCNA	proliferating cell nuclear antigen	232	0.020	-0.013	1.424	1.174	2.611
15481	HSPA8	heat shock protein 8	211	0.027	1.384	2.706	2.326	6.416
17869	MYC	myelocytomatosis oncogene	202	0.016	0.898	1.344	1.126	3.368
107508	EPRS	glutamyl-prolyl-tRNA synthetase	200	0.009	-1.871	-1.598	-1.652	5.121
12566	CDK2	cyclin-dependent kinase 2	191	0.011	-0.048	1.134	1.000	2.182
53606	ISG15	ISG15 ubiquitin-like modifier	191	0.009	1.923	3.396	0.600	5.918
20846	STAT1	signal transducer and activator of transcription 1	189	0.012	2.057	2.802	1.138	5.996
12387	CTNNB1	catenin (cadherin associated protein), beta 1	187	0.023	0.546	0.960	1.085	2.591
18817	PLK1	polo-like kinase 1	178	0.009	0.143	1.524	0.532	2.199
12540	CDC42	cell division cycle 42	176	0.014	0.088	0.754	1.108	1.950
246728	OAS2	2-5 oligoadenylate synthetase 2	175	0.005	1.266	2.526	1.109	4.902
23962	OASL2	2-5 oligoadenylate synthetase-like 2	175	0.006	1.343	2.303	1.512	5.158

El grado mínimo de estos nodos seleccionados es 175, es decir, en la red existen 20 nodos con al menos 175 vecinos. Si ampliamos el criterio de selección de los nodos centrales para que tengan al menos 1500 vecinos a distancia <=2, obtenemos 54 nodos, de los cuales 18 coinciden con los nodos de grado más alto. Nótese que cada uno de estos nodos tiene al menos la mitad de los nodos de la red a distancia <=2. De hecho, seleccionando los vecinos más cercanos de estos nodos a distancia 1 obtenemos una subred de 1642 nodos, y a distancia 2 de 2644 nodos, casi toda la red. Esto demuestra que los nodos centrales son una buena representación de la red.

El nodo con el máximo número de vecinos a distancia <=2 es `TRP53` que coincide con el de grado máximo. Este nodo tiene 2100 vecinos a distancia <=2 (un 75 % de los nodos de la red), por lo que parece un nodo muy influyente en la red. El gen murino *trp53* codifica para la proteína supresora de tumores p53, que responde a diversos estreses celulares para regular los genes diana que inducen la detención del ciclo celular, la apoptosis, la senescencia, la reparación del ADN o los cambios en el metabolismo. Este gen se encuentra subexpresado en los bazos de ratones infectados con la cepa NL con respecto a los no infectados. De interés, se ha reportado que en ratones deficientes para este gen, la infección por malaria promueve un tipo de linfoma de células B similar al linfoma de Burkitt en humanos, lo que podría explicar por qué este tipo de cáncer es común en áreas donde la malaria es endémica [23]. Por otro lado, otro estudio ha revelado que la supresión de p53 es crítica para el estadío hepático de la infección por *Plasmodium* [24].

3.2.3. Identificación de módulos altamente conectados en la malaria NL

Los análisis de clustering en las PPIN sirven para encontrar grupos de genes altamente interconectados, también llamados módulos o clusters. Estos pueden representar procesos biológicos o unidades funcionales, por lo que su caracterización puede ayudar a reducir la complejidad de la red. Además, pueden ser específicos de fenotipo. Existen diferentes algoritmos de clustering para analizar las PPIN basados en la topología de la red, disponibles en la aplicación de Cytoscape `clusterMaker` [6]. Dos de los métodos más usados son:

- Algoritmo de Newman-Girvan [25]: identifica los clusters o comunidades mediante el uso de la medida de centralidad de intermediación de los ejes (*edge betweenness centrality*). Los ejes que conectan diferentes clusters tienen valores de centralidad más altos, ya que una gran proporción de los caminos más

cortos los atravesarán. Para definir las comunidades, primero clasifica los ejes en base a su medida de centralidad de intermediación para eliminar los más centrales y luego vuelve a calcular los puntajes de intermediación hasta que no queden ejes. Los ejes afectados por la eliminación se consideran parte de la misma comunidad.

- Algoritmo de detección de complejos moleculares (MCODE) [26]: es un método más estringente que el algoritmo de Newman-Girvan, ya que tiene como objetivo encontrar sólo aquellas subredes que están muy altamente interconectadas, que representan complejos de proteínas relativamente estables que funcionan como una entidad única en el tiempo y el espacio. El algoritmo utiliza un proceso de tres etapas: (1) Ponderación: se otorga una puntuación más alta a aquellos nodos cuyos vecinos están más interconectados; (2) Predicción del complejo molecular: partiendo del nodo de mayor ponderación (semilla), se va moviendo recursivamente hacia afuera, agregando nodos al complejo que están por encima de un umbral determinado; (3) Postprocesamiento: aplica filtros para mejorar la calidad del grupo.

En este trabajo me he basado en el primer método, implementado en la aplicación de Cytoscape `clusterMaker` (método `GLay`) [6], ya que es más sencillo de utilizar (y computacionalmente más ligero) y, al ser menos estringente, produce clústeres más grandes que MCODE y lo hace más adecuado para redes grandes y para facilitar la interpretación funcional de la red [27]. A continuación se muestran los clusters obtenidos mediante este método con la PPIN NLvsCtrl. En amarillo marcamos los nodos centrales identificados anteriormente (genes con >=1500 vecinos a distancia <=2):

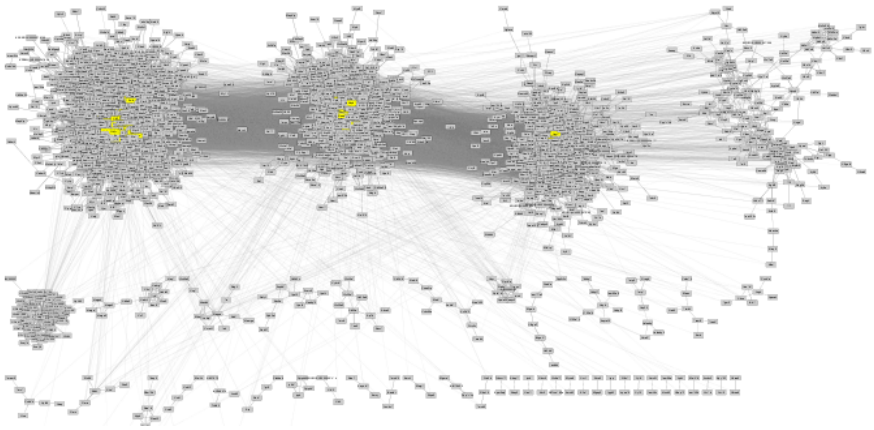


Figura 11: Clústers topológicos de la PPIN NLvsCtrl

En total se obtuvieron 46 clústeres, 3 con un gran número de nodos y otros clusters más pequeños. Se observan los nodos centrales identificados anteriormente al centro de los 3 primeros clústers. La siguiente tabla resume el número de nodos y ejes obtenidos para los 5 primeros clusters:

Cluster	Nodos	Ejes
1	1112	10345
2	748	12259
3	518	4958
4	153	515
5	67	380

Cabe destacar el gran número de ejes del cluster 2 y la forma en estrella del clúster 5.

3.2.3.1. Nodos centrales de los módulos topológicos

Mientras que los *hubs* inter-modulares son centrales para toda la red, los *hubs* intra-modulares son centrales para módulos específicos en la red y su identificación es importante ya que suelen tener alta relevancia biológica [15]. Las siguientes tablas muestran los 5 nodos con grado más alto para cada clúster.

Cuadro 6: Lista de nodos con grado más alto en los clústers de NLvsCtrl

Entrez	Symbol	description	Degree	Centrality	logFC.d3	logFC.d4	logFC.d5	var
Cluster1								
22059	TRP53	transformation related protein 53	366	0.070	-0.990	-1.682	-1.946	4.617
66725	LRRK2	leucine-rich repeat kinase 2	308	0.046	1.499	1.607	1.376	4.482
11461	ACTB	actin, beta	271	0.042	-0.702	-2.402	-2.013	5.118
16476	JUN	Jun oncogene	248	0.030	1.273	1.585	1.218	4.077
18538	PCNA	proliferating cell nuclear antigen	232	0.020	-0.013	1.424	1.174	2.611
Cluster2								
15519	HSP90AA1	heat shock protein 90, alpha (cytosolic), class A member 1	331	0.037	0.271	2.146	2.117	4.534
15516	HSP90AB1	heat shock protein 90 alpha (cytosolic), class B member 1	320	0.031	-0.252	1.621	0.967	2.840
69719	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	248	0.026	0.949	1.117	1.037	3.103
229363	GMPS	guanine monophosphate synthetase	237	0.024	0.590	1.037	0.948	2.574
15481	HSPA8	heat shock protein 8	211	0.027	1.384	2.706	2.326	6.416
Cluster3								
53606	ISG15	ISG15 ubiquitin-like modifier	191	0.009	1.923	3.396	0.600	5.918
246728	OAS2	2'-5' oligoadenylate synthetase 2	175	0.005	1.266	2.526	1.109	4.902
23962	OASL2	2'-5' oligoadenylate synthetase-like 2	175	0.006	1.343	2.303	1.512	5.158
231655	OASL1	2'-5' oligoadenylate synthetase-like 1	168	0.005	0.714	1.207	-0.011	1.932
20878	AURKA	aurora kinase A	158	0.007	0.257	1.344	0.959	2.560
Cluster4								
14104	FASN	fatty acid synthase	43	0.003	0.136	1.610	0.519	2.264
12613	CEL	carboxyl ester lipase	41	0.001	-2.744	-3.551	-3.380	9.675
109660	CTRL	chymotrypsin-like	41	0.002	-2.283	-3.015	-2.944	8.242
22072	PRSS2	protease, serine, 2	38	0.001	-4.063	-5.265	-4.690	14.018
67373	AV072249	RIKEN cDNA 2210010C04 gene	38	0.001	-4.863	-6.539	-5.342	16.743
Cluster5								
109689	ARRB1	arrestin, beta 1	95	0.032	-1.192	-0.146	-0.896	2.234
14708	GNG7	guanine nucleotide binding protein (G protein), gamma 7	77	0.010	1.883	-0.357	-0.088	2.327
258749	OLFR556	olfactory receptor 556	25	0.000	3.077	0.009	0.437	3.524
258572	OLFR1032	olfactory receptor 1032	25	0.001	-1.017	-0.488	-0.443	1.949
258582	OLFR1022	olfactory receptor 1022	25	0.000	-1.188	-1.133	-1.463	3.784

3.2.4. Análisis de módulos temporales en la malaria NL

Para integrar la información temporal de los datos de expresión de los DEG, podemos realizar un clustering de los genes basado en los patrones de expresión en los diferentes días post-infección. Para ello, utilizamos la aplicación de Cytoscape `TiCoNe`, que permite (i) agrupar los genes en base a su perfil de coexpresión temporal, (ii) analizar las interacciones entre los diferentes patrones temporales identificados y (iii) comparar los perfiles temporales entre diferentes condiciones experimentales [7].

En primer lugar, realizamos un agrupamiento de los genes en base a su perfil de expresión temporal, de manera que los perfiles temporales en un grupo sean más similares entre sí que a los perfiles temporales de otros grupos. Utilizamos la distancia euclídea como función de similaridad. Para la partición, se utiliza el método *PAMK* o de los *k-medoides* [28], un método similar al de las *k-medias* pero menos sensible a los *outliers*. Se realiza una partición inicial en 20 clústers seguida por una serie de iteraciones para minimizar el coste (suma de distancias de los objetos a su medoide) hasta convergencia. Elegimos un número elevado de clústers como punto de partida ya que se esperan encontrar al menos 14 perfiles temporales distintos (ver dibujos figura 12). Posteriormente se realiza una optimización manual de los clusters dividiendo los clusters con diferentes perfiles y fusionando aquellos con prototipos similares. Para cada clúster, `TiCoNe` muestra un patrón consenso, llamado prototipo. Para cada clúster, se calcula el coeficiente de correlación de Pearson medio de los objetos respecto a su prototipo, así como un p-valor empírico basado en permutaciones [7].

A continuación se muestra el resultado del análisis de clustering temporal realizado con `TiCoNe`:

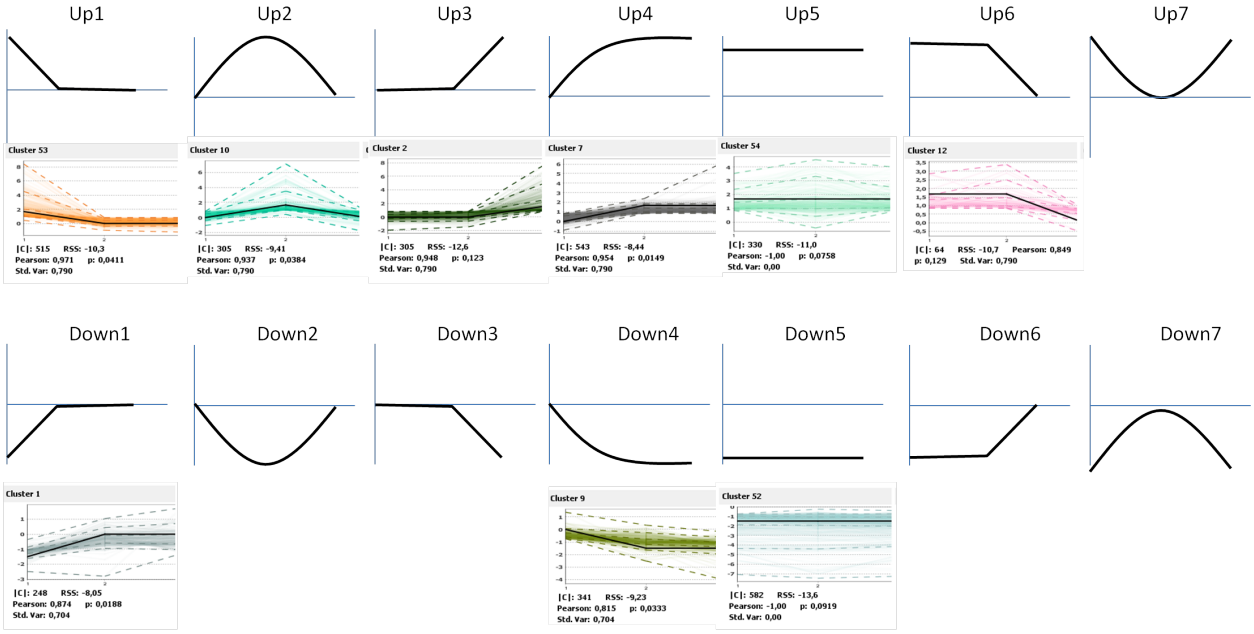


Figura 12: Clústers temporales de los DEG en NLvsCtrl

Se han ordenado los patrones encontrados en función de los perfiles temporales esperados. Por ejemplo, el patrón *Up1* corresponde a los genes que se sobreexpresan a día 3 pero no están alterados a días 4 y 5. En cambio, el patrón *Down1* indicaría los genes que están subexpresados a día 3 pero no están alterados a días 4 y 5. En la siguiente tabla se resumen las características de los patrones encontrados.

Patrón	ClusterID	Nº genes	Pearson	RSS	P-valor
Up1	53	515	0,985	-10,3	0,041*
Up2	10	305	0,968	-9,41	0,038*
Up3	2	305	0,974	-12,6	0,123
Up4	7	543	0,977	-8,4	0,015*
Up5	54	330	0	-11	0,0758
Up6	12	64	0,925	-10,7	0,129
Down1	1	248	0,937	-8,05	0,019*
Down4	9	341	0,907	-9,23	0,034*
Down5	52	582	0	-13,6	0,092

Se han identificado 9 patrones distintos de entre los 14 esperados, 5 de los cuales muestran un coeficiente de correlación (Pearson) elevado y un p-valor menor a 0.05 (en la tabla se indican con un asterisco).

Comparación con los diagramas de Venn

La siguiente tabla muestra una comparación entre los genes clasificados en los diferentes patrones temporales por `TiCoNe` y las comparaciones temporales realizadas anteriormente con diagramas de Venn (ver sección 3.1.4).

Patrón	Nº Objetos (TiCoNe)	Nº Objetos (Venn)	P-valor (TiCoNe)
Up1	515	674	0,041*
Up2	305	523	0,038*
Up3	305	493	0,123
Up4	543	451	0,015*
Up5	330	166	0,0758
Up6	64	49	0,129
Down1	248	315	0,019*
Down2	-	100	-
Down3	-	284	-
Down4	341	241	0,034*
Down5	582	306	0,092

A diferencia de los diagramas de Venn obtenidos anteriormente, la clasificación de los genes en función de su perfil temporal realizada con `TiCoNe` no se restringe a valores absolutos de logFC sino que permite un ajuste más amplio en base a patrones o tendencias de expresión entre los distintos días post-infección. Aunque esto puede resultar ventajoso con patrones temporales complejos, hay que tener en cuenta que el margen de error es amplio y con redes grandes obtenemos perfiles bastante ruidosos.

3.2.4.1. Análisis de conectividad entre los clústeres temporales

A continuación analizamos las interacciones entre genes a lo largo del tiempo con `TiCoNe` mediante el cálculo de la probabilidad de observar más (o menos) interacciones entre pares de patrones de tiempo por casualidad. La siguiente red muestra la relación entre los clusteres temporales, donde los ejes representan los enriquecimientos (verde) y depleciones (azul) significativos ($p < 0.01$).

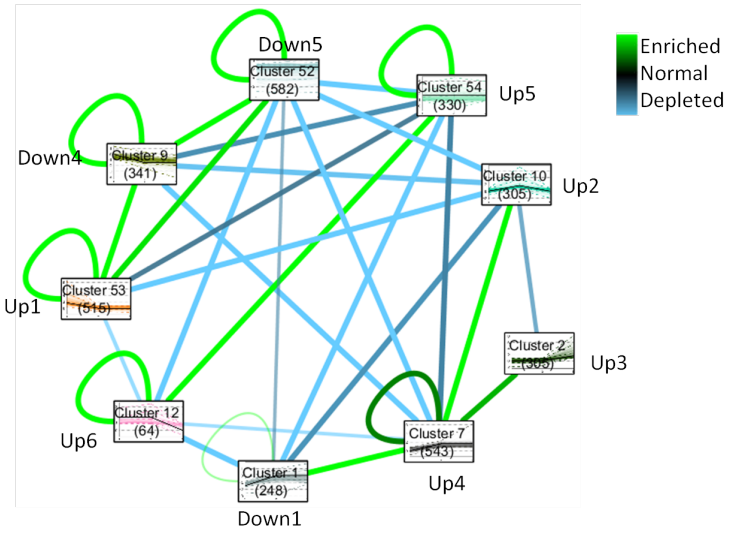


Figura 13: Conectividad entre los clústers temporales de NLvsCtrl

Se observan algunos clústeres temporales más conectados entre sí que lo esperado por azar, sugiriendo la presencia de subconjuntos co-regulados más grandes. Encontramos 3 subconjuntos formados por los siguientes clústers:

1) `Up4-Up3-Up2-Down1` (1401 nodos)
2) `Up5-Up6` (394 nodos)
3) `Down5-Down4-Up1` (1438 nodos)

Nótese que en el subconjunto 1, el clúster `Up4` aparece como nodo central, mientras que el subconjunto 3 es una gráfica completa (todos están conectados entre sí). Por otro lado, los *self-loops* verdes indicarían que los genes dentro del cluster están también más conectados de lo esperado por azar.

3.2.4.2. Comparación de los clusteres topológicos con los temporales

A continuación comparamos los clústeres temporales con los clústers topológicos obtenidos previamente. Para ello, coloreamos los nodos de la red de clústeres topológicos en función de los patrones temporales obtenidos.

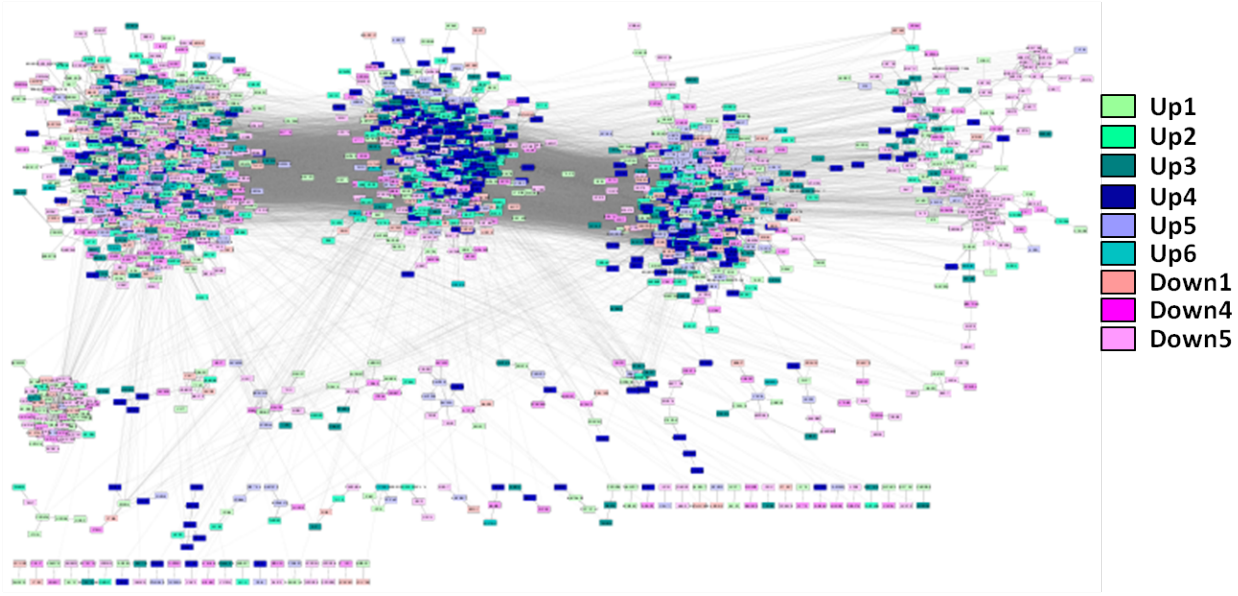


Figura 14: Visualización de los clústers temporales con los clústers topológicos en la PPIN NLvsCtrl

La siguiente tabla resume la repartición de los patrones temporales entre los distintos clústeres (se muestra el porcentaje de nodos de cada cluster que corresponden al patrón indicado).

Cuadro 9: Distribución de los patrones temporales de NLvsCtrl en los distintos clusters

	Clust1	**Clust2**	**Clust3**	**Clust4**	**Clust5**
Up1	17.3	10.0	12.9	12.4	34.3
Up2	7.4	14.2	10.8	7.2	4.5
Up3	11.2	7.6	9.3	7.8	3.0
Up4	9.8	32.4	23.0	8.5	0.0
Up5	10.3	10.3	15.6	3.9	3.0
Up6	2.2	1.1	3.3	0.7	0.0
Down1	8.5	8.2	7.1	2.0	11.9
Down4	12.1	5.9	7.3	9.8	11.9
Down5	21.2	10.4	10.6	47.7	31.3

Se observa que el `clúster 2` está principalmente representado por genes con el patrón `Up4`, el `clúster 4` por genes con el patrón `Down5` y el `clúster 5` por genes con el patrón `Up1` y `Down5`.

3.2.4.3. Coloración de los nodos centrales de la red en función de los perfiles temporales

Por otro lado, podemos colorear los 54 nodos centrales de la red obtenidos en la sección 3.2.2 en función de su perfil temporal para determinar si alguno de ellos se encuentra más representado (figura 15).

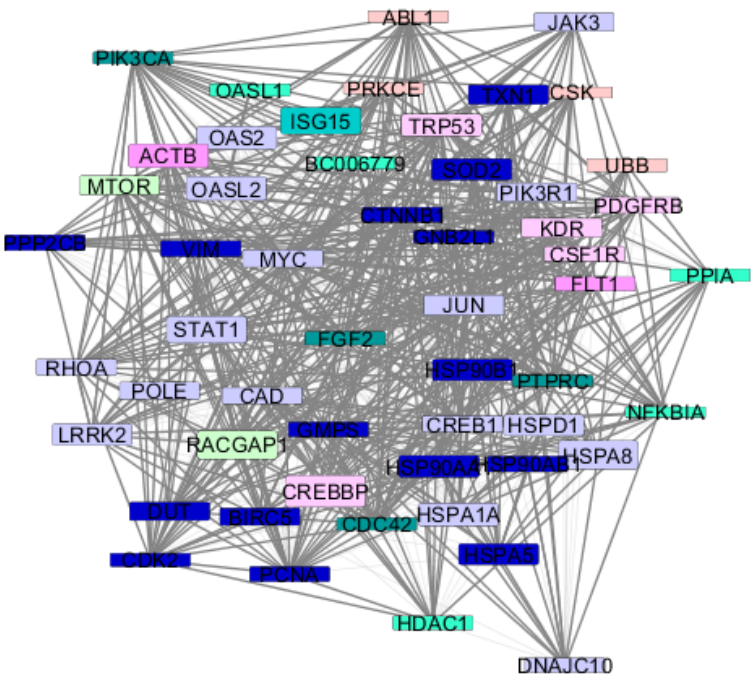


Figura 15: Visualización de los nodos centrales de la red NLvsCtrl en función del perfil temporal (color) y la variación acumulada (altura caja)

La siguiente tabla muestra el número de nodos centrales observado para cada perfil temporal así como el que se esperaría dado el tamaño del cluster y de la red. De interés, los clústeres temporales `Up4` y `Up5` se encuentran sobrerrepresentados entre los nodos centrales (ratio > 1.2). Estos perfiles temporales corresponden a genes que permanecen activados a días 4 y 5 post-infección (`Up4`) o durante los tres días post-infección (`Up5`), por lo que podrían tener una función importante.

Cluster	Num. Objetos en el cluster	Num. Objetos en la red	Num. hubs en la red	Num. hubs del cluster observados	Num. hubs del cluster esperados	Ratio Observ./Esperados
Up1	515	2796	54	2	9,9	0,20
Up2	305	2796	54	5	5,9	0,85
Up3	305	2796	54	4	5,9	0,68
Up4	543	2796	54	15	10,5	1,43
Up5	330	2796	54	16	6,4	2,51
Up6	64	2796	54	1	1,2	0,81
Down1	248	2796	54	4	4,8	0,84
Down4	341	2796	54	2	6,6	0,30
Down5	582	2796	54	5	11,2	0,44

3.2.4.4. Nodos centrales de los clusters temporales

Las siguientes tablas muestran los 5 nodos con grado más alto para cada clúster temporal.

Cuadro 11: Lista de nodos con grado más alto en los clústers temporales de NLvsCtrl

Entrez	Symbol	description	Degree	Centrality	logFC.d3	logFC.d4	logFC.d5	var
Up1								
26934	RACGAP1	Rac GTPase-activating protein 1	172	0.015	6.312	0.076	0.062	6.450
56717	MTOR	mechanistic target of rapamycin (serine/threonine kinase)	152	0.012	3.222	0.380	0.147	3.749
12615	CENPA	centromere protein A	130	0.004	1.418	0.060	0.442	1.919
231889	BUD31	BUD31 homolog (yeast)	111	0.004	2.794	0.734	0.683	4.210
11820	APP	amyloid beta (A4) precursor protein	104	0.011	1.042	0.084	0.098	1.224
Up2								
18817	PLK1	polo-like kinase 1	178	0.009	0.143	1.524	0.532	2.199
231655	OASL1	2'-5' oligoadenylate synthetase-like 1	168	0.005	0.714	1.207	-0.011	1.932
229003	BC006779	cDNA sequence BC006779	162	0.014	-0.160	1.672	-0.078	1.910
67891	RPL4	ribosomal protein L4	150	0.002	-0.591	1.349	0.171	2.110
246727	OAS3	2'-5' oligoadenylate synthetase 3	144	0.003	0.823	1.462	0.786	3.072
Up3								
12540	CDC42	cell division cycle 42	176	0.014	0.088	0.754	1.108	1.950
18706	PIK3CA	phosphatidylinositol 3-kinase, catalytic, alpha polypeptide	138	0.010	0.019	0.777	1.007	1.803
19264	PTPRC	protein tyrosine phosphatase, receptor type, C	137	0.010	-0.003	0.587	1.075	1.665
19989	RPL7	ribosomal protein L7	134	0.001	-0.314	0.761	1.046	2.121
20873	PLK4	polo-like kinase 4	117	0.003	0.289	-0.066	3.078	3.432
Up4								
15519	HSP90AA1	heat shock protein 90, alpha (cytosolic), class A member 1	331	0.037	0.271	2.146	2.117	4.534
15516	HSP90AB1	heat shock protein 90 alpha (cytosolic), class B member 1	320	0.031	-0.252	1.621	0.967	2.840
229363	GMPS	guanine monophosphate synthetase	237	0.024	0.590	1.037	0.948	2.574
18538	PCNA	proliferating cell nuclear antigen	232	0.020	-0.013	1.424	1.174	2.611
12566	CDK2	cyclin-dependent kinase 2	191	0.011	-0.048	1.134	1.000	2.182
Up5								
66725	LRRK2	leucine-rich repeat kinase 2	308	0.046	1.499	1.607	1.376	4.482
16476	JUN	Jun oncogene	248	0.030	1.273	1.585	1.218	4.077
69719	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	248	0.026	0.949	1.117	1.037	3.103
15481	HSPA8	heat shock protein 8	211	0.027	1.384	2.706	2.326	6.416
17869	MYC	myelocytomatosis oncogene	202	0.016	0.898	1.344	1.126	3.368
Up6								
53606	ISG15	ISG15 ubiquitin-like modifier	191	0.009	1.923	3.396	0.600	5.918
77579	MYH10	myosin, heavy polypeptide 10, non-muscle	98	0.010	1.197	0.888	0.735	2.821
12524	CD86	CD86 antigen	96	0.004	1.029	0.861	0.426	2.316
67138	HERC6	hect domain and RLD 6	92	0.003	0.858	1.598	0.805	3.260
24110	USP18	ubiquitin specific peptidase 18	70	0.001	1.375	3.062	0.816	5.253
Down1								
19983	RPL5	ribosomal protein L5	149	0.003	-0.785	-0.245	-1.183	2.213
22187	UBB	ubiquitin B	140	0.006	-1.295	-0.179	-0.595	2.069
11350	ABL1	c-abl oncogene 1, non-receptor tyrosine kinase	134	0.010	-1.244	-0.081	-0.333	1.658
20103	RPS5	ribosomal protein S5	125	0.001	-1.098	0.424	0.267	1.788
20104	RPS6	ribosomal protein S6	125	0.002	-1.247	1.038	0.900	3.186
Down4								
11461	ACTB	actin, beta	271	0.042	-0.702	-2.402	-2.013	5.118
14254	FLT1	FMS-like tyrosine kinase 1	104	0.005	-0.560	-1.123	-0.754	2.437
170758	RAC3	RAS-related C3 botulinum substrate 3	103	0.003	-0.290	-0.824	-1.193	2.308
12048	BCL2L1	BCL2-like 1	88	0.002	-0.501	-0.282	-1.135	1.917
20637	SNRNP70	small nuclear ribonucleoprotein 70 (U1)	88	0.002	-0.401	-1.090	-1.159	2.649
Down5								
22059	TRP53	transformation related protein 53	366	0.070	-0.990	-1.682	-1.946	4.617
107508	EPRS	glutamyl-prolyl-tRNA synthetase	200	0.009	-1.871	-1.598	-1.652	5.121
12914	CREBBP	CREB binding protein	150	0.008	-2.464	-2.483	-2.399	7.346
16542	KDR	kinase insert domain protein receptor	131	0.008	-1.680	-1.112	-1.011	3.803
53607	SNRPA	small nuclear ribonucleoprotein polypeptide A	107	0.003	-1.659	-1.376	-1.501	4.536

3.2.5. Análisis de significación biológica (ABS) en la malaria NL

Para inferir en las funciones biológicas afectadas por la malaria NL, realizamos un análisis de significación biológica sobre las diferentes listas de genes seleccionadas en los apartados anteriores. El análisis se basa en un análisis de enriquecimiento para diferentes anotaciones disponibles en bases de datos como la *Gene Ontology* (GO, http://www.geneontology.org/) o *Reactome Pathway* (https://reactome.org/). El objetivo de este análisis es realizar una de las pruebas estadísticas disponibles para determinar si un conjunto de genes dado (p.ej. una categoría particular del GO), está sobrerrepresentada en la lista de genes seleccionados con respecto a un conjunto de referencia. Como conjunto de referencia suelen tomarse todos los genes del array o los que están anotados en la base de datos del organismo utilizado.

Existen muchas herramientas para realizar análisis de enriquecimiento funcional, disponibles vía servidores web (`DAVID`, `g:Profiler`) o implementadas en R como paquetes (`GOstats`, `clusterProfiler`). En este trabajo, me he basado en la aplicación de Cytoscape `ClueGO` [8], la cual realiza un análisis de enriquecimiento basado en la distribución hipergeométrica y ofrece una visualización organizada y agrupada de los términos por similaridad funcional que facilitan su interpretación.

Realizamos la búsqueda conjunta para términos GO relacionados con procesos biológicos y vías metabólicas anotadas en Reactome. Para reducir la redundancia de los términos encontrados y facilitar así su interpretación, agrupamos los términos por similaridad mediante el cálculo del estadístico *kappa* (*kappa score* > 0.4) y fusionamos los términos con más de la mitad de genes comunes dentro de la jerarquía padre-hijo. Como resultado, obtenemos un listado con los términos/grupos enriquecidos (p-valor ajustado < 0.05), un diagrama de barras con el porcentaje de genes/término, así como un diagrama de sectores con los grupos enriquecidos. Finalmente, `ClueGO` ofrece también una representación en forma de red de los términos agrupados por similaridad. En cada caso, el nombre del grupo corresponde al del término más significativo del grupo.

3.2.5.1. ABS de los top 300 genes con más variación acumulada

En primer lugar, realizamos un análisis de enriquecimiento funcional para los top 300 genes con más variación acumulada (en valores absolutos) a lo largo de los tres días post-infección. En la figura 16 se muestra el diagrama de sectores con los grupos de procesos/vías significativos (p-valor ajustado <0.05) encontrados. En el anexo se incluye el diagrama de barras con la relación de términos encontrados para cada grupo (anexo 7.3).

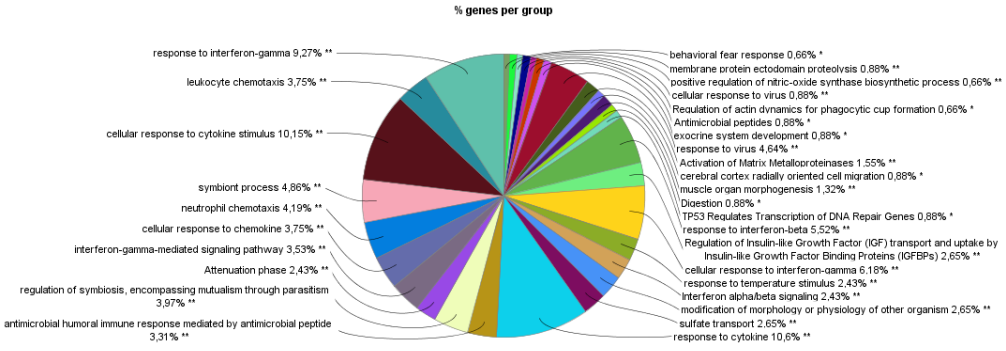


Figura 16: Diagrama de sectores de los procesos/vías enriquecidos en los top 300 DEGs de NLvsCtrl

Se observan un elevado porcentaje de genes/términos relacionados con la respuesta inmune (respuesta a citoquinas y al IFN-γ, así como con la interacción huésped-patógeno (adhesión, simbiosis) y la hematopoyesis.

3.2.5.2. ABS de los nodos centrales

También podemos realizar el análisis de enriquecimiento sobre los 54 nodos centrales encontrados anteriormente. Este subset permite tener una representación simplificada de toda la red. A pesar de ser un grupo pequeño de genes, se obtiene un buen número de procesos/vías enriquecidas, reflejando la centralidad de estos nodos. El siguiente diagrama muestra la distribución de los grupos enriquecidos encontrados. Se observan enriquecidos términos relacionados con la proliferación de células del tejido muscular liso y la angiogénesis, así como el metabolismo del ADN.

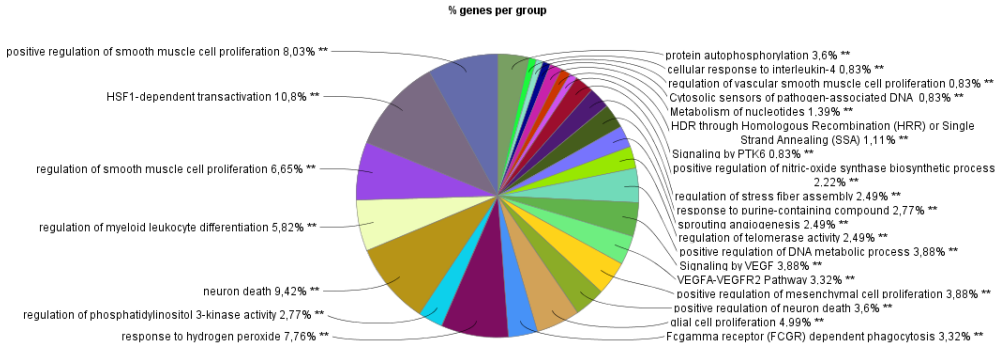


Figura 17: Diagrama de sectores de los procesos/vías enriquecidos en los nodos centrales de NLvsCtrl

3.2.5.3. ABS de los módulos topológicos

A continuación se realiza un análisis de enriquecimiento funcional para los módulos principales detectados. La tabla 12 muestra los top 5 grupos significativos (p-valor ajustado < 0.05) encontrados para cada clúster.

Cuadro 12: Principales procesos/vías enriquecidos en los clústers topológicos de NLvsCtrl

ID	Término	P.Valor	% Genes asociados	Num. Genes
Cluster1				
GO:0032879	regulation of localization	4,50E-49	11,19	332,00
GO:0048583	regulation of response to stimulus	5,57E-67	10,82	452,00
GO:0051239	regulation of multicellular organismal process	1,08E-58	11,32	379,00
GO:0010646	regulation of cell communication	8,36E-55	10,77	388,00
GO:0006952	defense response	2,62E-36	12,34	217,00
Cluster2				
GO:0022613	ribonucleoprotein complex biogenesis	8,02E-63	24,61	111,00
GO:0034641	cellular nitrogen compound metabolic process	6,70E-69	6,79	442,00
R-MMU:72766	Translation	5,70E-29	28,75	46,00
GO:0009199	ribonucleoside triphosphate metabolic process	1,20E-20	17,56	49,00
GO:0070925	organelle assembly	4,47E-02	4,56	37,00
Cluster3				
R-MMU:1640170	Cell Cycle	2,80E-40	14,56	83,00
GO:0051276	chromosome organization	1,45E-50	10,75	132,00
GO:0006281	DNA repair	3,38E-24	11,94	59,00
GO:0071824	protein-DNA complex subunit organization	1,07E-23	17,06	43,00
R-MMU:983169	Class I MHC mediated antigen processing & presentation	3,17E-24	14,17	51,00
Cluster4				
R-MMU:556833	Metabolism of lipids	5,16E-19	4,93	32,00
GO:0046486	glycerolipid metabolic process	2,13E-11	5,23	18,00
R-MMU:211859	Biological oxidations	2,25E-10	5,86	15,00
R-MMU:8957322	Metabolism of steroids	2,93E-08	7,52	10,00
GO:0006066	alcohol metabolic process	1,27E-10	5,07	17,00
Cluster5				
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	1,99E-69	5,10	59,00

Se observa una segregación de los tipos de procesos/vías por clúster. Aunque aquí sólo se muestran los top 5 más representativos para cada clúster, recorriendo la lista completa observamos una tendencia en la presencia de términos relacionados con la migración celular y respuesta a estímulos en el `Cluster1`; términos relacionados con biogénesis y metabolismo de RNA y proteínas en el `Cluster2`; términos relacionados con el ciclo celular en el `Cluster3`, y términos relacionados con el metabolismo de lípidos en el `Cluster4`. El `Cluster5` contiene un solo término significativamente enriquecido relacionado con el olfato (en la figura 11 se observa que este clúster es pequeño y casi todos los genes son de la familia *OLF*). Esta segregación de términos por clúster es consistente con la idea de que los clústeres topológicos (unidades fuertemente interconectadas) tienen un significado biológico y representan unidades funcionales. Así pues, atacando los *hubs* de estos clústeres estaremos afectando diferentes funciones biológicas.

3.2.5.4. ABS de los módulos temporales

A continuación se realiza un análisis de enriquecimiento funcional para los módulos temporales detectados. La tabla 13 muestra los top 5 grupos significativos (p-valor ajustado < 0.05) encontrados para cada clúster.

De interés, entre los genes con patrones `Up`, encontramos términos principalmente relacionados con la respuesta inmune, aunque los grupos son un poco heterogéneos. Un estudio detallado de los procesos inmunes involucrados en los diferentes patrones `Up` podría ayudar a entender la temporalización en la inducción de la respuesta inmune en la malaria NL. Por ejemplo, en los patrones `Up1`, `Up2` y `Up3`, correspondientes a genes sobreexpresados a días 3, 4 o 5 post-infección, encontramos enriquecidos términos relacionados con la respuesta inmune humoral, la activación de células T o la diferenciación de monocitos, respectivamente. El patrón `Up5`, que corresponde a los genes que se mantienen sobreexpresados en los tres días post-infección, hay un enriquecimiento claro en genes relacionados con la respuesta a patógenos. Por otro lado, entre los genes con patrones `Down`, encontramos enriquecidos términos relacionados con la hemostasis y la vascularización.

Cuadro 13: Principales procesos/vías enriquecidos en los clústers temporales de NLvsCtrl

ID	Término	P.Valor	% Genes asociados	Num. Genes
Up1				
GO:0007018	microtubule-based movement	4,96E-04	6,51	19,00
GO:0009755	hormone-mediated signaling pathway	2,92E-02	4,84	12,00
GO:0021987	cerebral cortex development	1,44E-02	6,71	10,00
GO:0043500	muscle adaptation	1,11E-02	7,09	10,00
GO:0002923	regulation of humoral immune response mediated by circulating immunoglobulin	2,09E-02	25,00	3,00
Up2				
GO:0048525	negative regulation of viral process	4,36E-06	11,11	11,00
R-MMU:453279	Mitotic G1-G1/S phases	3,32E-03	5,38	7,00
R-MMU:73856	RNA Polymerase II Transcription Termination	2,75E-06	15,00	9,00
R-MMU:3700989	Transcriptional Regulation by TP53	2,16E-05	5,51	15,00
GO:0002711	positive regulation of T cell mediated immunity	2,82E-04	11,32	6,00
Up3				
GO:0045655	regulation of monocyte differentiation	2,76E-03	21,05	4,00
GO:0019722	calcium-mediated signaling	2,82E-03	5,03	10,00
GO:0071426	ribonucleoprotein complex export from nucleus	3,86E-03	7,23	6,00
GO:0002221	pattern recognition receptor signaling pathway	1,06E-02	4,61	7,00
GO:0060707	trophoblast giant cell differentiation	1,00E-02	14,29	3,00
Up4				
R-MMU:8953854	Metabolism of RNA	8,86E-29	13,28	66,00
GO:0034641	cellular nitrogen compound metabolic process	3,17E-35	4,39	286,00
GO:0022613	ribonucleoprotein complex biogenesis	2,95E-22	12,20	55,00
GO:0006996	organelle organization	2,87E-21	4,70	178,00
GO:0006259	DNA metabolic process	1,43E-13	6,78	65,00
Up5				
GO:0034097	response to cytokine	5,25E-18	5,65	57,00
GO:0045087	innate immune response	1,21E-11	4,71	44,00
GO:0042832	defense response to protozoan	3,23E-07	22,22	8,00
GO:0043903	regulation of symbiosis, encompassing mutualism through parasitism	2,67E-05	6,14	14,00
GO:0002697	regulation of immune effector process	5,52E-07	5,42	22,00
Up6				
GO:0002294	CD4-positive, alpha-beta T cell differentiation involved in immune response	4,91E-04	4,55	3,00
Down1				
GO:0090287	regulation of cellular response to growth factor stimulus	1,76E-03	4,14	12,00
GO:2001026	regulation of endothelial cell chemotaxis	3,07E-03	13,79	4,00
GO:0008064	regulation of actin polymerization or depolymerization	2,50E-03	4,37	8,00
GO:0030218	erythrocyte differentiation	5,56E-03	4,29	6,00
R-MMU:72613	Eukaryotic Translation Initiation	2,72E-03	7,81	5,00
Down4				
GO:0048488	synaptic vesicle endocytosis	4,21E-03	7,81	5,00
GO:1901215	negative regulation of neuron death	6,53E-04	4,96	13,00
GO:0010039	response to iron ion	1,91E-03	11,11	5,00
GO:0007599	hemostasis	1,53E-03	5,26	10,00
R-MMU:140877	Formation of Fibrin Clot (Clotting Cascade)	2,65E-04	15,38	6,00
Down5				
GO:0042127	regulation of cell proliferation	4,97E-07	4,11	79,00
GO:0016477	cell migration	4,54E-06	4,07	62,00
GO:0090066	regulation of anatomical structure size	1,15E-03	4,49	26,00
GO:0003018	vascular process in circulatory system	1,06E-02	5,24	11,00
GO:0030216	keratinocyte differentiation	2,30E-04	8,39	13,00

3.2.6. Conclusiones

- La red de interacciones proteína-proteína modelada a partir de los 3417 genes diferencialmente expresados en la malaria NL está altamente conectada, con una componente conectada principal de 2832 nodos y 36758 ejes.
- Los nodos con mayor centralidad de intermediación coinciden, en general, con los de grado más alto, pero no se corresponden con los DEG con más variación acumulada a lo largo de los días post-infección.
- El nodo con el máximo número de vecinos a distancia <=2 es `TRP53`, que coincide con el de máximo grado/centralidad.
- Se han identificado 5 clústeres topológicos principales en la red. El análisis de significación biológica para los diferentes clústers muestra una segregación de términos por clúster, consistente con la idea de que los clústeres topológicos tienen un significado biológico y representan unidades funcionales.
- Se han identificado 9 clústeres temporales distintos, de acuerdo a su perfil de expresión en los diferentes días post-infección. El ABS detallado de los diferentes patrones temporales proporciona información sobre la regulación temporal de la respuesta del huésped en la infección. Por otro lado, el análisis topológico entre los diferentes clústers temporales revela la agrupación de estos en 3 grandes conjuntos que podrían indicar subunidades correguladas más grandes.
- Se han identificado los nodos centrales de la red, así como de los clústers topológicos y temporales. Estos constituyen posibles candidatos a dianas terapéuticas de cara a atacar toda la red o módulos específicos.

Limitaciones y posibles variaciones

- La representación de un elevado número de DEG mediante redes permite obtener una visión amplia de las moléculas/módulos involucrados en los distintos procesos, sin embargo, también ha dificultado su manejo, visualización e interpretación. Esto se podría resolver aplicando unos umbrales más restringentes, ya sea seleccionando sólo los DEG con abs(logFC)>2 para limitar el número de genes de partida, o considerando sólo las interacciones validadas experimentalmente (o con *score* más elevado) para tener una componente principal conectada más pequeña.
- En mi opinión, el análisis de clústeres temporales con `TiCoNe` es laborioso y genera resultados ruidosos. Se podrían explorar otros métodos que permitan integrar la información temporal con la topológica, por ejemplo mediante análisis de redes con pesos en los ejes de acuerdo a su correlación temporal (método *WGCNA* [29]).
- Aunque la aplicación `ClueGO` implementa algoritmos para agrupar los términos por similaridad y reducir así la redundancia de las anotaciones funcionales, se observan ciertas limitaciones asociadas a la nomenclatura de los grupos. Por un lado, al recibir el grupo el nombre del término más significativo, el hecho de que un mismo término pueda aparecer en distintos grupos si comparte similaridad con los otros términos del grupo puede hacer que varios grupos reciban el mismo nombre. Por otro lado, el nombre del grupo no siempre es representativo de los términos que aparecen dentro, resultando en una nomenclatura confusa y en pérdida de información. Creo que los esfuerzos de anotación y reducción de redundancia de las anotaciones funcionales aún es un campo en desarrollo necesario para facilitar la interpretación biológica de los resultados a los investigadores.

3.3. Comparación de las PPIN en las dos malarias

Con el fin de comparar la malaria NL con la L, modelamos la PPIN para los genes diferencialmente expresados en el contraste LvsCtrl y repetimos algunos de los análisis efectuados para la malaria NL (ver anexo 6.1).

3.3.1. Intersección entre las 2 PPIN: Nodos comunes y específicos

Para comparar los DEG entre la malaria NL y L, obtenemos la intersección entre la PPIN NLvsCtrl y la PPIN LvsCtrl (genes comunes) así como la substracción de ambas (genes únicos para cada malaria). La siguiente tabla muestra las características de las 3 subredes obtenidas:

Subred	Nodos	Ejes	Nodos componente ppal
Intersección NLvsL	1234	6445	939
- Up-Up	560		
- Down-Down	632		
- Up-Down	33		
- Down-Up	9		
Sólo NL	1898	13661	1517
Sólo L	312	400	170

Para cada subred, se obtiene una componente principal conectada con el número de nodos indicado en la tabla. Sorprendentemente, la mayoría de los DEG de LvsCtrl se encuentran también en NLvsCtrl. Entre los nodos comunes, se han desglosado entre los que muestran el mismo sentido de sobreexpresión (Up-Up) o subexpresión (Down-Down) entre las dos malarias, denominados con "correlación positiva", así como los que muestran una regulación en sentido inverso (Up-Down/Down-Up), denominados con "correlación negativa". Se observa que la mayoría van en el mismo sentido. El subset de nodos regulados en sentidos opuestos, así como los nodos específicos para cada malaria (Sólo NL/Sólo L), son interesantes ya que podrían estar asociados a diferencias en el fenotipo observado entre las dos malarias. En un análisis más fino, sería interesante también analizar las diferencias en los perfiles temporales de los genes comunes encontrados. Un análisis a nivel general se muestra en la sección 3.3.2.

Por otro lado, un análisis de clustering topológico de las diferentes subredes podría revelar módulos funcionales específicos/comunes para cada malaria. En la figura 18 se muestra el clustering topológico obtenido anteriormente para la red NLvsCtrl, y en verde se indican los genes comunes correlacionados positivamente con la malaria L, en rojo los correlacionados negativamente y en gris los genes únicos para la malaria NL. A rasgos generales, se observa que los nodos comunes se encuentran repartidos entre los cuatro primeros clusters, y tienen una representación importante en el tercer clúster, sugiriendo la contribución de este módulo en funciones comunes a las dos malarias.

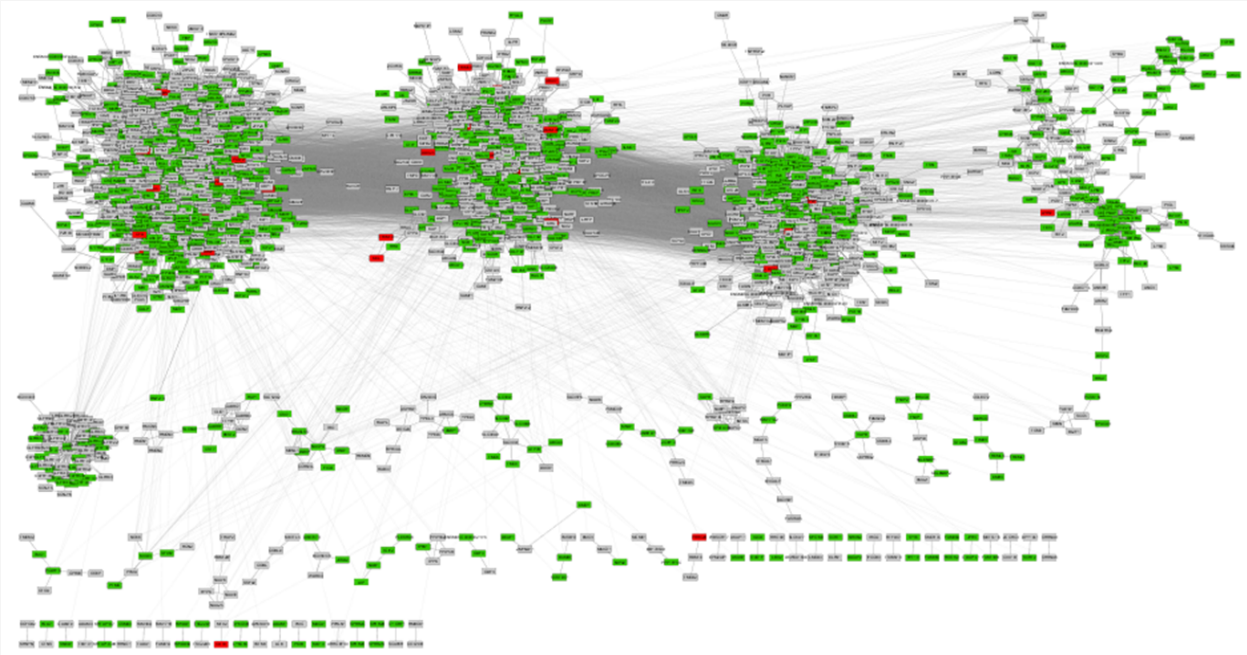


Figura 18: Visualización de los DEG comunes/específicos de NL/L en los clústers topológicos de NLvsCtrl

3.3.1.1. Nodos centrales comunes/específicos para cada malaria

Si comparamos los 20 nodos centrales (grado más alto) identificados anteriormente para cada malaria, obtenemos que 11 son comunes en las dos malarias. En las tablas 15-17 se listan los nodos centrales comunes y específicos para cada malaria junto con su patrón de expresión y variación acumulada.

Cuadro 15: Lista de nodos con grado más alto comunes en las dos malarias

Entrez	Symbol	description	var_NL	Pattern_NL	var_L	Pattern_L
11461	ACTB	actin, beta	5.118	Down4	6.044	Down5
15481	HSPA8	heat shock protein 8	6.416	Up5	3.001	Up4
17869	MYC	myelocytomatosis oncogene	3.368	Up5	3.024	Up5
20846	STAT1	signal transducer and activator of transcription 1	5.996	Up5	6.838	Up5
22059	TRP53	transformation related protein 53	4.617	Down5	6.323	Down5
23962	OASL2	2'-5' oligoadenylate synthetase-like 2	5.158	Up5	6.064	Up5
53606	ISG15	ISG15 ubiquitin-like modifier	5.918	Up6	6.853	Up5
66725	LRRK2	leucine-rich repeat kinase 2	4.482	Up5	2.921	Up5
107508	EPRS	glutamyl-prolyl-tRNA synthetase	5.121	Down5	4.933	Down5
229363	GMPS	guanine monophosphate synthetase	2.574	Up4	2.126	Up4
246728	OAS2	2'-5' oligoadenylate synthetase 2	4.902	Up5	5.918	Up5

Cuadro 16: Lista de nodos con grado más alto diferencialmente expresados sólo en la malaria NL

Entrez	Symbol	description	var	Pattern
15519	HSP90AA1	heat shock protein 90, alpha (cytosolic), class A member 1	4.534	Up4
15516	HSP90AB1	heat shock protein 90 alpha (cytosolic), class B member 1	2.840	Up4
16476	JUN	Jun oncogene	4.077	Up5
69719	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	3.103	Up5
18538	PCNA	proliferating cell nuclear antigen	2.611	Up4
12566	CDK2	cyclin-dependent kinase 2	2.182	Up4
12387	CTNNB1	catenin (cadherin associated protein), beta 1	2.591	Up4
18817	PLK1	polo-like kinase 1	2.199	Up2
12540	CDC42	cell division cycle 42	1.950	Up3

Cuadro 17: Lista de nodos con grado más alto diferencialmente expresados sólo en la malaria L

Entrez	Symbol	description	var	Pattern
21926	Tnf	tumor necrosis factor	2.213	Up1
15978	Ifng	interferon gamma	7.977	Up5
16183	Il2	interleukin 2	2.723	Down4
15945	Cxcl10	chemokine (C-X-C motif) ligand 10	9.467	Up5
20848	Stat3	signal transducer and activator of transcription 3	1.769	Up1
23918	Impdh2	inosine 5'-phosphate dehydrogenase 2	1.613	Up4
231655	Oasl1	2'-5' oligoadenylate synthetase-like 1	4.466	Up5
20296	Ccl2	chemokine (C-C motif) ligand 2	7.973	Up5
17329	Cxcl9	chemokine (C-X-C motif) ligand 9	10.322	Up5

Se observa que los nodos centrales comunes en las dos malarias se encuentran diferencialmente expresados en el mismo sentido (Up-Up/Down-Down).

Las listas proporcionadas en estas tablas pueden ser de interés de cara a encontrar dianas terapéuticas comunes/específicas para cada malaria (ver sección 3.4.2).

3.3.2. Comparación de los perfiles temporales en las dos malarias

Finalmente, utilizamos la funcionalidad de comparación de fenotipos de `TiCoNE` para buscar grupos de genes que se comportan temporalmente diferente en las dos infecciones. El resultado obtenido muestra los grupos conectados por ejes en función del número de objetos compartido entre cada par de grupos. Como se observa en la figura 19, mediante este método no detectamos grupos de genes con una marcada diferencia en su perfil temporal entre los dos grupos.

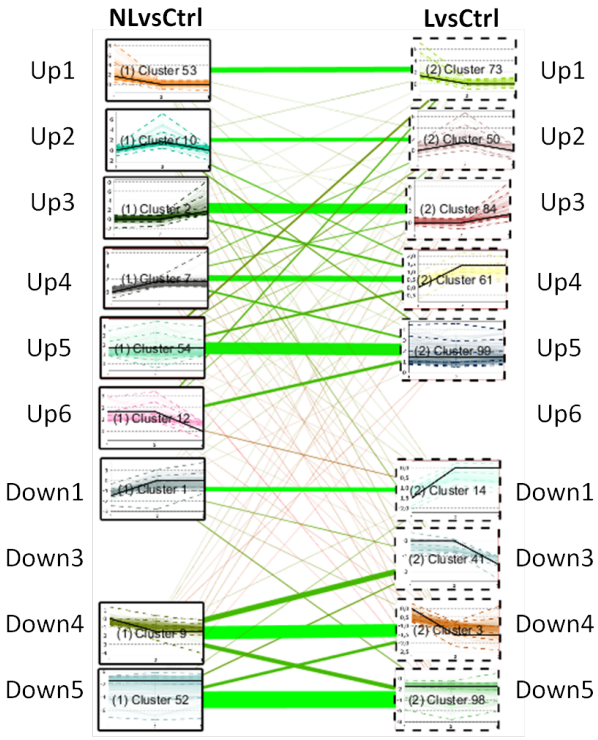


Figura 19: Comparación de los clústers temporales obtenidos en las dos malarias

Aunque a nivel general no se observan diferencias en los patrones de expresión de los genes entre las dos malarias, podría ser interesante comparar los perfiles temporales de los genes entre las dos malarias para subgrupos funcionales específicos, p.ej. genes relacionados con la respuesta inmune.

3.3.3. ABS de los genes comunes/específicos entre las dos malarias

Para el análisis de enriquecimiento, seleccionamos los top 300 genes con más variación acumulada para cada subred: (i) Intersección con correlación positiva (Up-Up/Down-Down), (ii) Intersección con correlación negativa (Up-Down/Down-Up), (iii) Sólo NL y (iv) Sólo L.

Entre los términos encontrados para los DEGs comunes con correlación positiva (figura 20), encontramos procesos relacionados con la respuesta a citoquinas y al IFN-γ. Este diagrama se asemeja bastante al obtenido con los top 300 genes con más variación acumulada en NLvsCtrl (figura 16), sugiriendo que los genes que más varían en la malaria NL son comunes con la L. Por otro lado, entre los DEGs comunes con correlación negativa, sólo se encontraron dos términos enriquecidos: uno relacionado con la respuesta inmune humoral y el otro con la eliminación de células de otros organismos, indicando una regulación diferencial/inversa de estos procesos entre las dos malarias.

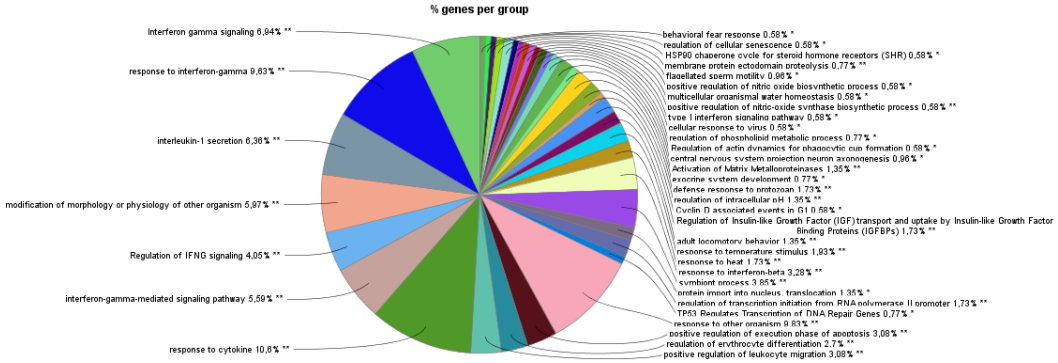


Figura 20: Diagrama de sectores de los procesos/vías enriquecidos en los top 300 DEGs comunes en las dos malarias (Up-Up/Down-Down)

Por otro lado, entre los top 300 DEGs específicos para la malaria NL, se observan enriquecidos términos relacionados con el ciclo celular y el metabolismo del ADN, así como la regulación negativa de la respuesta inmune (figura 21).

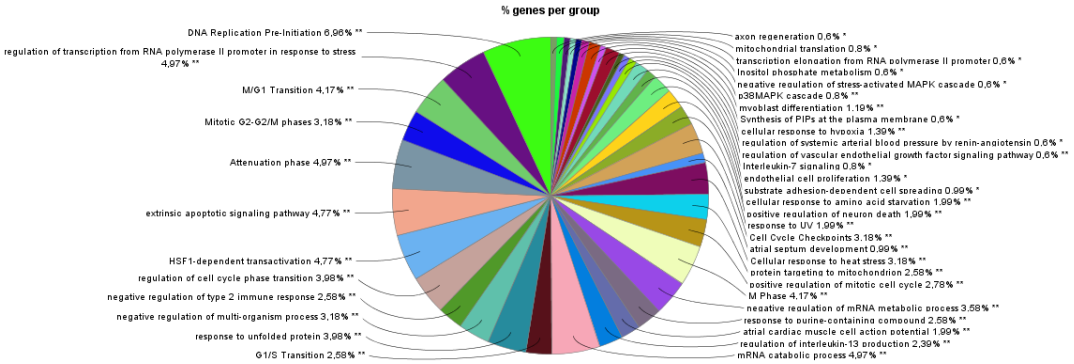


Figura 21: Diagrama de sectores de los procesos/vías enriquecidos en los top 300 DEGs específicos para la malaria NL

En cambio, entre los DEGs únicos para la malaria letal, destacan procesos relacionados con la activación de la respuesta inmune innata, el metabolismo de proteínas y la regulación epitelial (figura 22).

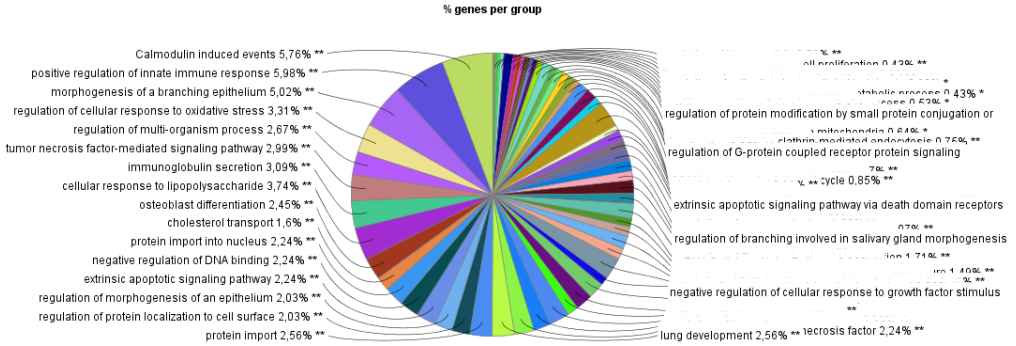


Figura 22: Diagrama de sectores de los procesos/vías enriquecidos en los top 300 DEGs específicos para la malaria L

3.3.4. Conclusiones

- La mayoría de los DEG de LvsCtrl se encuentran también en NLvsCtrl y muestran perfiles temporales similares.
- Se observa que los nodos comunes están sobrerrepresentados en el tercer clúster de la PPIN NLvsCtrl, sugiriendo la contribución de este módulo en funciones comunes a las dos malarias.
- El análisis de significación biológica para los DEG comunes/específicos de cada malaria permite diferenciar entre los procesos que se encuentran alterados en las dos malarias y los que se asocian a un fenotipo específico.
- De los top 20 nodos con grado más alto en cada malaria, 11 son comunes en las dos malarias y se encuentran diferencialmente expresados en el mismo sentido (Up-Up/Down-Down).
- Los nodos centrales comunes/específicos para cada malaria pueden ser de interés de cara a encontrar dianas terapéuticas generales/específicos para los diferentes fenotipos de la enfermedad.

Limitaciones y posibles variaciones

Aunque estos análisis permiten tener una idea general de los genes y procesos más influyentes en la malaria NL respecto a la L, seguramente podrían ser mejorados empleando análisis más finos/herramientas más sofisticadas. Por ejemplo, sería interesante investigar las diferencias topológicas entre las PPIN NLvsCtrl y LvsCtrl para identificar cambios en la interacción entre moléculas, en lugar de cambios en moléculas individuales (lo que se conoce como *Differential network analysis* [30]).

Por otro lado, aunque a nivel general no se observan diferencias en los patrones de expresión de los genes entre las dos malarias, podría ser interesante comparar los perfiles temporales de los genes entre las dos malarias para subgrupos funcionales específicos, p.ej. genes relacionados con la respuesta inmune.

3.4. Selección de candidatos a biomarcadores/dianas terapéuticas en la malaria

En este último capítulo se han seleccionado posibles candidatos a biomarcadores y/o dianas terapéuticas derivados de los análisis anteriores. Como candidatos a biomarcadores, se han considerado los genes que se encuentran más diferencialmente expresados en los ratones infectados respecto a los no infectados, ya que su presencia/ausencia está asociada a la enfermedad. Por otro lado, los nodos centrales representan posibles candidatos a dianas terapéuticas, ya que son nodos altamente influyentes en la topología de la red y por lo tanto su alteración puede modular el curso de la enfermedad. Para determinar el potencial de estos últimos a ser atacados por fármacos, se realiza un análisis de *druggability*.

3.4.1. Candidatos a biomarcadores en la malaria

Como candidatos a biomarcadores, seleccionamos los top 10 DEG con patrón de expresión `Up5` o `Down5`, ya que nos interesa que estén presentes/ausentes en los tres días post-infección, priorizados de mayor a menor variación acumulada. Estos pueden ser comunes para las dos malarias (tabla 18), reflejando características generales de la malaria respecto a los ratones no infectados; o específicos para cada malaria (tabla 19 para NL y tabla 20 para L), los cuales pueden estar asociados a las diferencias de fenotipo entre las dos malarias (supervivencia, parasitemia, respuesta inmune, etc).

Cuadro 18: Lista de candidatos a biomarcadores comunes en la malaria

Entrez	Symbol	Description	var_NL	var_L	Pattern_NL	Pattern_L
Up						
60440	IIGP1	interferon inducible GTPase 1	10.007139	10.637476	Up5	Up5
17329	CXCL9	chemokine (C-X-C motif) ligand 9	9.956739	10.321547	Up5	Up5
24108	Ubd	ubiquitin D	9.849104	10.458819	Up5	Up5
58203	ZBP1	Z-DNA binding protein 1	9.665327	10.647887	Up5	Up5
14468	GBP1	guanylate binding protein 1	9.559009	10.122701	Up5	Up5
110454	Ly6a	lymphocyte antigen 6 complex, locus A	9.419547	11.458898	Up5	Up5
12702	Socs3	suppressor of cytokine signaling 3	9.298753	9.848213	Up5	Up5
20715	SERPINA3G	serine (or cysteine) peptidase inhibitor, clade A, member 3G	8.789248	9.486104	Up5	Up5
12703	Socs1	suppressor of cytokine signaling 1	8.636214	9.084379	Up5	Up5
100702	Gbp6	guanylate binding protein 6	8.263010	8.389245	Up5	Up5
Down						
81840	Sorcs2	sortilin-related VPS10 domain containing receptor 2	21.782811	17.874903	Down5	Down5
70297	Gcc2	GRIP and coiled-coil domain containing 2	19.765592	14.230475	Down5	Down5
269585	ZSCAN20	zinc finger and SCAN domains 20	18.986254	14.122522	Down5	Down5
11722	AMY1	amylase 1, salivary	18.209965	15.978263	Down5	Down5
76701	CTRC	chymotrypsin C (caldecrin)	18.192061	15.138040	Down5	Down5
76703	CPB1	carboxypeptidase B1 (tissue)	17.948676	15.081467	Down5	Down5
67868	GM13011	predicted gene 13011	17.853368	14.718332	Down5	Down5
69060	PNLIP	pancreatic lipase	17.845859	14.852693	Down5	Down5
11723	AMY2A2	amylase 2a2	16.979260	15.412294	Down5	Down5
67373	AV072249	RIKEN cDNA 2210010C04 gene	16.743289	13.226090	Down5	Down5

Cuadro 19: Lista de candidatos a biomarcadores específicos de la malaria NL

Entrez	Symbol	Description	var_NL	Pattern_NL
Up				
102124	ENKD1	enkurin domain containing 1	5.081783	Up5
102093	PHKB	phosphorylase kinase beta	4.989282	Up5
193740	HSPA1A	heat shock protein 1A	4.829270	Up5
70292	AFAP1	actin filament associated protein 1	4.793259	Up5
26992	BRD7	bromodomain containing 7	4.735113	Up5
21917	TMPO	thymopoietin	4.533754	Up5
19183	PSMC3IP	proteasome (prosome, macropain) 26S subunit, ATPase 3, interacting protein	4.527728	Up5
76246	RTF1	Rtf1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)	4.367458	Up5
12457	CCRN4L	CCR4 carbon catabolite repression 4-like (S. cerevisiae)	4.266715	Up5
319362	ENSMUSG00000079242	RIKEN cDNA C730034F03 gene	4.192951	Up5
Down				
22138	TTN	titin	12.709647	Down5
71886	AV083437	RIKEN cDNA 2310002L09 gene	12.140901	Down5
20309	CXCL15	chemokine (C-X-C motif) ligand 15	12.115369	Down5
53603	TSLP	thymic stromal lymphopoietin	11.971906	Down5
74894	ENSMUSG00000053603	RIKEN cDNA 4930442H23 gene	11.898437	Down5
114585	D17H6S53E	DNA segment, Chr 17, human D6S53E	11.092548	Down5
258282	OLFR801	olfactory receptor 801	10.931069	Down5
77733	RNF170	ring finger protein 170	10.815221	Down5
328108	FAM179B	family with sequence similarity 179, member B	10.790421	Down5
20710	SERPINB9E	serine (or cysteine) peptidase inhibitor, clade B, member 9e	10.676609	Down5

Cuadro 20: Lista de candidatos a biomarcadores específicos de la malaria L

Entrez	Symbol	Description	var_L	Pattern_L
Up				
55932	Gbp4	guanylate binding protein 4	8.416613	Up5
64382	Ms4a6d	membrane-spanning 4-domains, subfamily A, member 6D	4.669320	Up5
56066	Cxcr3	chemokine (C-X-C motif) receptor 3	3.643167	Up5
16852	Lgals1	lectin, galactose binding, soluble 1	3.642837	Up5
434223	Gm1966	predicted gene 1966	3.610826	Up5
16185	Il2rb	interleukin 2 receptor, beta chain	3.610523	Up5
15039	H2-T22	histocompatibility 2, T region locus 22	3.503367	Up5
22380	Wbp4	WW domain binding protein 4	3.335651	Up5
20529	Slc31a1	solute carrier family 31, member 1	3.322398	Up5
21936	Tnfrsf18	tumor necrosis factor receptor superfamily, member 18	3.093300	Up5
Down				
20905	Sumf1	sulfatase modifying factor 1	7.650596	Down5
18573	Pde1a	phosphodiesterase 1A, calmodulin-dependent	4.710251	Down5
53604	Zpbp	zona pellucida binding protein	4.349986	Down5
16679	Krt86	keratin 86	3.904443	Down5
15424	Hoxc5	homeobox C5	3.858193	Down5
66583	Exosc1	exosome component 1	3.725349	Down5
226594	Rcsd1	RCSD domain containing 1	3.704142	Down5
11303	Abca1	ATP-binding cassette, sub-family A (ABC1), member 1	3.675953	Down5
13190	Dct	dopachrome tautomerase	3.663575	Down5
114301	Palmd	palmdelphin	3.652818	Down5

Los genes con un patrón *Up* serían marcadores de la enfermedad por su presencia/sobreexpresión, mientras

que los genes *Down* lo serían por su ausencia/subexpresión, respecto a un estado fisiológico *normal* (control).

Para validar la lista de candidatos a biomarcadores obtenida, por un lado, habría que validar experimentalmente sus niveles de expresion en el bazo e idealmente en sangre periférica, de cara a su uso translacional en humanos. Por otro lado, también habría que validar su poder pronóstico/diagnóstico con otros datasets mediante métodos de clasificación/predicción basados en *machine learning*, por ejemplo. En un análisis más fino, se podría seleccionar una combinación de varios genes para definir patrones más complejos de la enfermedad.

3.4.2. Candidatos a dianas terapéuticas en la malaria NL

Como candidatos a dianas terapéuticas en la malaria NL, seleccionamos los nodos centrales. Estos pueden ser de toda la red (tabla 4), de los módulos topológicos (tabla 6), de los módulos temporales (tabla 11) o de procesos biológicos/vías metabólicas específicos, según lo que se quiera atacar. A su vez, estos se pueden subdividir entre los que están afectados en las dos malarias (tabla 15) o específicos para cada malaria (tablas 16 y 17). Para simplificar, y como prueba de concepto, en este trabajo sólo se analiza la *druggability* de los top 20 nodos centrales (grado más alto) de la red NLvsCtrl.

Existen diferentes métodos para medir la *druggability* de una diana. Estos son, ordenados de menor a mayor confiabilidad, los basados en (i) la secuencia de la proteína, (ii) en su estructura, (iii) en la unión a ligandos conocidos (ya sean endógenos o farmacológicos) y (iii) los basados en evidencias de ensayos clínicos u otros que avalen que es *druggable*. De hecho, si una proteína es diana de un fármaco ya aprobado, esto da un grado muy alto de confianza en su *druggability*, aunque no es garantía de su éxito en la enfermedad/condición de estudio. Esta información se puede obtener de bases de datos como *DrugBank* (https://www.drugbank.ca/), *SuperTarget* (http://bioinformatics.charite.de/supertarget/) o *ChEMBL* (https://www.ebi.ac.uk/chembl/). Para predecir la *druggability* a partir de la secuencia/estructura de la proteína, existen diferentes herramientas como *DrugEBIlity* (https://www.ebi.ac.uk/chembl/drugebility) o *MitOpenScreen* (http://bioserv.rpbs.univ-paris-diderot.fr/services/MTiOpenScreen/).

En este trabajo, me he basado en la búsqueda de evidencias previas mediante *DrugBank* y el cálculo de la *druggability* con *DrugEBIlity* para validar la lista de candidatos a dianas terapéuticas. Este cálculo se basa en algoritmos de *machine learning* entrenados con estructuras del *Protein Data Bank (PDB)* que se sabe que se unen a ligandos con determinadas propiedades. Como resultado obtenemos el porcentaje promedio de *druggability* de todas las estructuras disponibles para el dominio de unión de la proteína de interés (https://www.ebi.ac.uk/chembl/drugebility/faq).

La siguiente tabla resumen muestra el resultado obtenido para los top 20 nodos centrales de la red NLvsCtrl, ordenados por su porcentaje promedio de *druggability* predicho. La tabla incluye la siguiente información:

- *Gene*: Nombre del gen seleccionado
- *Pattern*: Perfil de expresión del gen en NLvsCtrl, obtenido con `TiCoNe`.
- *Prot.Sym*: Nombre de la proteína más frecuente codificada por el gen seleccionado, obtenido mediante búsqueda en la base de datos *UniprotKB* (https://www.uniprot.org/). Por convención, éste suele conservarse entre especies.
- *Prot.Acc*: Identificador único de la proteína codificada en humanos, obtenida mediante búsqueda en *UniprotKB* a partir del gen murino.
- *Drugg %*: Porcentaje promedio de *druggability* calculado con *DrugEBIlity*.
- *DrugBank*: Existencia de evidencias previas en *DrugBank* (fármacos aprobados, en fase investigacional o experimental).
- *Compounds*: Compuestos aprobados encontrados en *DrugBank* para la diana de interés.
- *Actions*: Acción del fármaco descrita en *DrugBank*.

Cuadro 21: Lista de candidatos a dianas terapéuticas en la malaria NL y análisis de druggability

Gene	Pattern	Prot.Sym	Prot.Acc	Drugg %	DrugBank	Compounds	Action
HSP90AB1	Up4	HS90B	P08238	100	13 experimental, 2 investigational		
CDK2	Up6	CDK2	P24941	87	1 approved, 133 experimental, 3 investigational	Bosutinib	inhibitor
HSP90AA1	Up4	HS90A	P07900	51	3 approved, 43 experimental, 3 investigational	Rifabutin, Nedocromil, Copper	Unknown
CTNNB1	Up4	CTNB1	P35222	42	1 approved	Urea	
GMPS	Up4	GUAA	P49915	33	2 approved	Glutamic acid, L-Glutamine	substrate
CDC42	Up3	CDC42	P60953	30	2 experimental		
PLK1	Up2	PLK1	P53350	26	1 approved, 5 experimental	Fostamatinib	inhibitor
STAT1	Up5	STAT1	P42224	25	not found??		
TRP53	Down5	P53	P04637	6	4 approved, 2 experimental, 1 investigational	Acetylsalicylic acid, Zinc, Zinc acetate, Zinc chloride	Acetylation, unknown
LRRK2	Up5	LRRK2	Q5S007	0	1 approved	Fostamatinib	Inhibitor
ACTB	Down4	ACTB	P60709	0	1 experimental, 1 investigational		
JUN	Up5	JUN	P05412	0	3 approved, 1 investigational	Vinblastine, Irbesartan, Arsenic trioxide	unknown, inducer
CAD	Up5	DFFB[CAD]	O76075	0	2 approved, 1 investigational	L-Aspartic acid, L-Glutamine	substrate
PCNA	Up4	PCNA	P12004	0	1 approved	Liothyronine	
HSPA8	Up5	HSP7C	P11142	0	2 approved, 1 experimental, 2 investigational	Copper, Dasatinib	
MYC	Up5	MYC	P01106	0	1 approved	Nadroparin	
EPRS	Down5	SYEP	P07814	0	2 approved, 3 experimental	Proline, Glutamic Acid	
ISG15	Up4	ISG15	P05161	0	not found		
OAS2	Up5	OAS2	P29728	0	not found		
OASL2	Up5	OASL2	Q15646	0	not found		

En la tabla se muestra la proteína más frecuente codificada por el gen murino seleccionado y su homóloga en humano, que es la que utilizamos para evaluar la idoneidad de la proteína como diana terapéutica.

Los cálculos realizados con *DrugEBIlity* revelan 3 proteínas (`HSP90B`, `CDK` y `HSP90A`) con una probabilidad de *druggability* elevada (>50 %), 5 con una *druggability* intermedia (25-50 %) y 12 con baja (<6 %). Para la proteína OAS2 no se obtiene ningún resultado basado en su estructura, por lo que se utiliza una búsqueda por similaridad basada en su secuencia. Esta búsqueda nos devuelve similaridades de alrededor del 50 % con dominios de la proteína OAS1 de *Sus scrofa*, para la que se obtiene un 0 % de *druggability*.

Por otro lado, se observa que para la mayoría de proteínas seleccionadas existen evidencias previas de fármacos en *DrugBank*. En concreto, para 13 de las 20 dianas preseleccionadas existen evidencias de fármacos aprobados, lo que indicaría que son *druggable*, a pesar de haber obtenido una baja predicción con *DrugEBIlity*. Para la proteína `HSP90B` no se han obtenido evidencias de fármacos aprobados, pero muestra un elevado porcentaje promedio de *druggability*, por lo que también podría ser considerada como buena candidata a diana terapéutica. En cambio, para las proteínas `ISG15`, `OAS2` y `OASL2` no se han obtenido evidencias en *DrugBank* y su *druggability* es del 0 %, por lo que no serían buenos candidatos.

Finalmente, habría que ver si los fármacos encontrados se han descrito para el tratamiento de la malaria y si su acción terapéutica es la deseada. Cabe comentar, que el hecho de atacar estos nodos centrales no es garantía de que revirtamos el fenotipo de la enfermedad a un estado “sano”, pero es un punto de partida

para su validación experimental/clínica.

3.4.3. Conclusiones

- Se han seleccionado los top 10 DEG con más variación acumulada, de entre los DEG comunes/específicos para cada malaria, como posibles candidatos a biomarcadores. Los DEG específicos de cada malaria podrían ayudar a monitorizar procesos específicos asociados las diferentes patologías.
- Los nodos centrales de la red, así como de los diferentes módulos topológicos/temporales, representan posibles candidatos a dianas terapéuticas.
- Como prueba de concepto, se ha realizado un análisis de *druggability* con los candidatos a dianas terapéuticas identificados para la malaria NL. De las 20 proteínas candidato, sólo 3 tienen un promedio de *druggability* superior al 50 %, según el cálculo realizado con la herramienta *DrugEBIlity*. Sin embargo, se han encontrado evidencias en *DrugBank* de fármacos aprobados para 13 de las proteínas seleccionadas, sugiriendo que son *druggable*.

Limitaciones y posibles variaciones

Se observan ciertas discrepancias entre los cálculos de *druggability* realizados con *DrugEBIlity* y la búsqueda de evidencias previas en *DrugBank*, por lo que sería interesante comparar con otras herramientas/bases de datos para respaldar los resultados obtenidos.

Aunque los análisis de *druggability* son un buen filtro para validar dianas terapéuticas, comentar que habría otras maneras de atacar los genes candidatos, no sólo con *small molecule drugs*, sino atacando las interacciones proteína-proteina con inhibidores o anticuerpos, o utilizando estrategias de terapia génica/siRNA para bloquear la expresión del gen de interés; por lo que el hecho de que no sean *druggable* no los descartaría por completo como diana terapéutica. Por otro lado, para las dianas que no son *druggable*, los análisis de biología de sistemas pueden proporcionar alternativas válidas, ampliando así el "universo *druggable*" de la enfermedad.

4. Conclusión

En este trabajo se han se han utilizado herramientas de biología de sistemas en combinación con análisis de significación biológica para obtener una visión integrada de los procesos biológicos que se encuentran alterados en la malaria, más allá de los análisis basados en genes individuales, e identificar posibles biomarcadores/dianas terapéuticas.

Se han logrado alcanzar los objetivos inicialmente planteados, explorando el potencial de diferentes herramientas para, a partir de unos datos de microarray, extraer el máximo de información desde diferentes puntos de vista e integrarlos de cara a identificar procesos clave en la enfermedad.

En el futuro, sería interesante comparar los resultados obtenidos en el modelo murino con el conocimiento disponible para la enfermedad humana y estudiar en más profundidad las diferentes estrategias terapéuticas.

5. Glosario

Abreviaciones

- *DEG*: Gen diferencialmente expresado
- *NL*: No-letal
- *L*: Letal
- *Ctrl*: Control
- *vs*: versus
- *PCA*: Análisis de componentes principales
- *PPIN*: Red de interacción proteína-proteína
- *ABS*: Análisis de significación biológica
- *GO*: Gene Ontology

Conceptos

- *Up-regulado*: sobreexpresado.
- *Down-regulado*: subexpresado.
- *Red de interacción proteína-proteína*: representación matemática de las interacciones específicas entre proteínas. Los nodos representan las proteínas y los ejes reflejan la presencia de una interacción entre éstas.
- *Nodo central o hub*: nodo con alto número de conexiones que es importante para la conectividad de la red.
- *Módulo/clúster topológico*: grupo de nodos estrechamente interconectados, definido por áreas densas de conectividad separadas por regiones de baja conectividad.
- *Módulo/clúster temporal*: grupo de genes con un perfil de expresión más similar entre sí que al de otros grupos.
- *Análisis de significación biológica*: análisis de enriquecimiento basado en tests estadísticos para determinar si un conjunto de genes dado, p.ej. una categoría particular del *GO*, está sobrerrepresentada en la lista de genes seleccionados ("muestra") con respecto a un conjunto de referencia ("población") de donde fue seleccionada.
- *Biomarcador*: indicador medible de un estado biológico.
- *Diana terapéutica*: molécula del organismo donde un fármaco ejerce su acción.
- *Druggability*: capacidad de un objetivo biológico de unirse a un fármaco (normalmente pequeña molécula) y ser modulado por éste.

Herramientas

- *R* (https://www.r-project.org/): software libre para computación estadística y gráficos.
- *Limma*: paquete de R para el análisis de datos de expresión génica obtenidos mediante tecnologías de microarray o RNA-Seq.
- *Cytoscape* (https://cytoscape.org/): herramienta para la representación y visualización integrativa de los datos en forma de red.
- *stringApp* (http://apps.cytoscape.org/apps/stringApp): aplicación de Cytoscape para importar los datos de interacción proteína-proteína contenidos en la base de datos *String*.

- *clusterMaker* (http://apps.cytoscape.org/apps/clustermaker2): aplicación de Cytoscape para la agrupación/clustering de la red usando diferentes algoritmos.
- *TiCoNe* (http://apps.cytoscape.org/apps/ticone): herramienta implementada como aplicación de Cytoscape para agrupar los nodos de la red en base a su perfil temporal.
- *ClueGO* (http://apps.cytoscape.org/apps/cluego): aplicación de Cytoscape que permite realizar análisis de enriquecimiento para diferentes anotaciones disponibles en bases de datos como *GO* o *Reactome* y ofrece una visualización organizada y agrupada de los términos por similaridad funcional.
- *DrugEBIlity* (https://www.ebi.ac.uk/chembl/drugebility): herramienta del EBI-EMBL para predecir la *druggability* de una proteína a partir de su secuencia/estructura.

Bases de datos

- *String* (https://string-db.org/): proporciona información sobre las interacciones proteína-proteína conocidas y predichas para diferentes especies.
- *GO* (http://www.geneontology.org/): provee un vocabulario controlado (ontología) que describe el gen y los atributos del producto génico (anotaciones) en cualquier organismo. La información se divide en tres ontologías, cada una de las cuales representando un concepto clave en biología molecular: la función molecular de los productos génicos; su rol en los procesos biológicos, y su localización en componentes celulares.
- *Reactome* (https://reactome.org/): provee información sobre las vías y reacciones biológicas que abarcan una amplia gama de procesos biológicos como la señalización, el metabolismo, la regulación transcripcional, la apoptosis y la transmisión sináptica.
- *DrugBank* (https://www.drugbank.ca/): provee información sobre medicamentos y sus objetivos farmacológicos.

6. Anexo

6.1. Caracterización de la malaria L mediante PPIN

A continuación se muestran los principales resultados obtenidos para la malaria L.

6.1.1. Modelado de la PPIN en la malaria L

A continuación se muestra la red de interacciones obtenida a partir de la lista de DEGs seleccionados en la malaria L, según la información disponible en la base de datos de STRING.

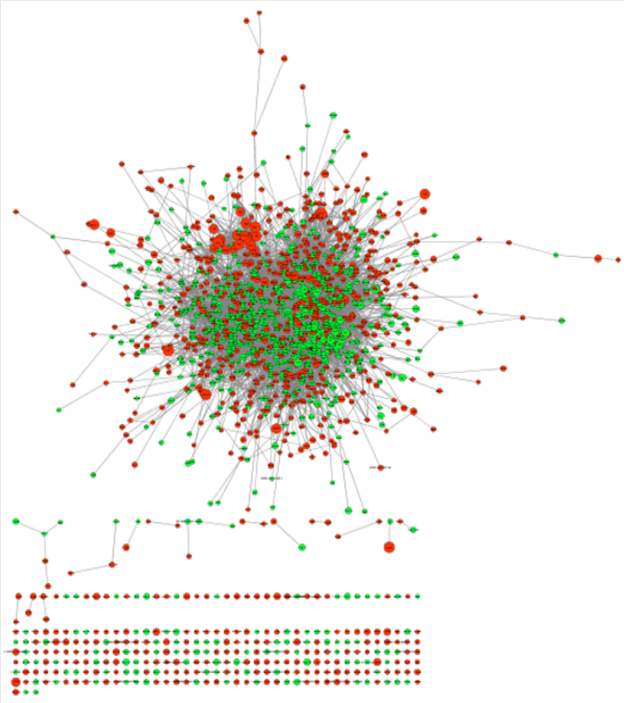


Figura 23: PPIN de los genes diferencialmente expresados en la malaria LvsCtrl

Obtenemos una red con 1546 nodos (genes) y 9824 ejes (interacciones), de los cuales sólo 883 interacciones han sido validadas experimentalmente con un *score*>=0.4. En verde representamos los 700 genes sobreexpresados (logFC>=1 para alguno de los días post-infección) y en rojo los 846 genes subexpresados (logFC<=-1 para alguno de los días post-infección). El tamaño de los nodos corresponde a la variación de la expresión acumulada en los 3 días. Se observa una componente conectada principal de 1226 nodos y 9805 ejes, que es la que utilizaremos para los subsiguientes análisis. La siguiente tabla muestra algunos descriptivos globales de la red:

Parámetro	Valor
Núm. nodos	1226
Núm. ejes	9805
Grado máximo	203
Diámetro	11
Centralización	0.153
Camino más corto característico	3.3

6.1.2. Identificación de nodos centrales en la malaria L

A continuación creamos una subred con los 20 nodos de mayor grado/centralidad. El grosor de los ejes indica el score calculado para la evidencia experimental de la interacción.

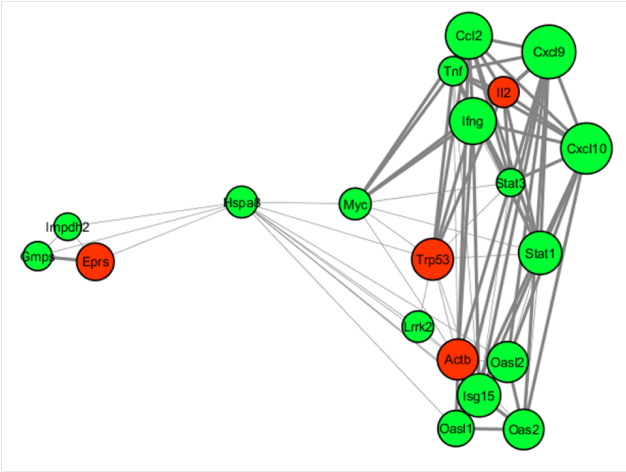


Figura 24: Subred con los 20 nodos con grado más alto (LvsCtrl)

En la siguiente tabla se detalla la información de los nodos seleccionados, entre los que se encuentran varias citoquinas y factores de transcripción:

Cuadro 23: Lista de nodos con grado más alto en LvsCtrl

Entrez	Symbol	description	Degree	Centrality	logFC.d3	logFC.d4	logFC.d5	var
22059	Trp53	transformation related protein 53	203	0.128	-1.402	-2.305	-2.615	6.323
21926	Tnf	tumor necrosis factor	178	0.068	1.090	0.424	0.699	2.213
66725	Lrrk2	leucine-rich repeat kinase 2	138	0.067	0.905	0.808	1.208	2.921
20846	Stat1	signal transducer and activator of transcription 1	135	0.024	2.800	2.076	1.962	6.838
11461	Actb	actin, beta	131	0.062	-0.872	-2.790	-2.382	6.044
15978	Ifng	interferon gamma	121	0.024	3.230	2.314	2.433	7.977
17869	Myc	myelocytomatosis oncogene	112	0.049	1.078	0.952	0.993	3.024
16183	Il2	interleukin 2	110	0.020	-0.748	-0.934	-1.041	2.723
15945	Cxcl10	chemokine (C-X-C motif) ligand 10	107	0.011	3.767	2.482	3.218	9.467
20848	Stat3	signal transducer and activator of transcription 3	106	0.027	1.195	0.239	0.335	1.769
53606	Isg15	ISG15 ubiquitin-like modifier	106	0.015	2.906	2.372	1.575	6.853
229363	Gmps	guanine monophosphate synthetase	105	0.032	0.384	1.022	0.720	2.126
15481	Hspa8	heat shock protein 8	103	0.041	-0.736	0.995	1.271	3.001
23962	Oasl2	2'-5' oligoadenylate synthetase-like 2	99	0.009	1.939	2.056	2.069	6.064
246728	Oas2	2'-5' oligoadenylate synthetase 2	96	0.008	2.133	1.744	2.041	5.918
23918	Impdh2	inosine 5'-phosphate dehydrogenase 2	92	0.022	0.194	1.048	0.371	1.613
231655	Oasl1	2'-5' oligoadenylate synthetase-like 1	92	0.007	2.221	1.226	1.018	4.466
20296	Ccl2	chemokine (C-C motif) ligand 2	90	0.011	3.189	2.191	2.593	7.973
107508	Eprs	glutamyl-prolyl-tRNA synthetase	89	0.015	-1.926	-1.364	-1.643	4.933
17329	Cxcl9	chemokine (C-X-C motif) ligand 9	87	0.007	4.463	2.808	3.051	10.322

El grado mínimo de estos nodos seleccionados es 87, es decir, en la red existen 20 nodos con al menos 87 vecinos. Si ampliamos el criterio de selección de los nodos centrales para que tengan al menos 600 vecinos a distancia <=2, obtenemos 35 nodos, de los cuales 13 coinciden con los nodos de grado más alto. Como habíamos observado en el caso de la malarial NL, el nodo con el máximo número de vecinos a distancia <=2 es `TRP53` que coincide con el de grado máximo, por lo que parece ser también un gen central en la malaria L.

6.1.3. Identificación de módulos altamente conectados en la malaria L

A continuación se muestra el resultado de la clusterización por el método de Glay con la PPIN LvsCtrl. En amarillo marcamos los nodos centrales identificados anteriormente:

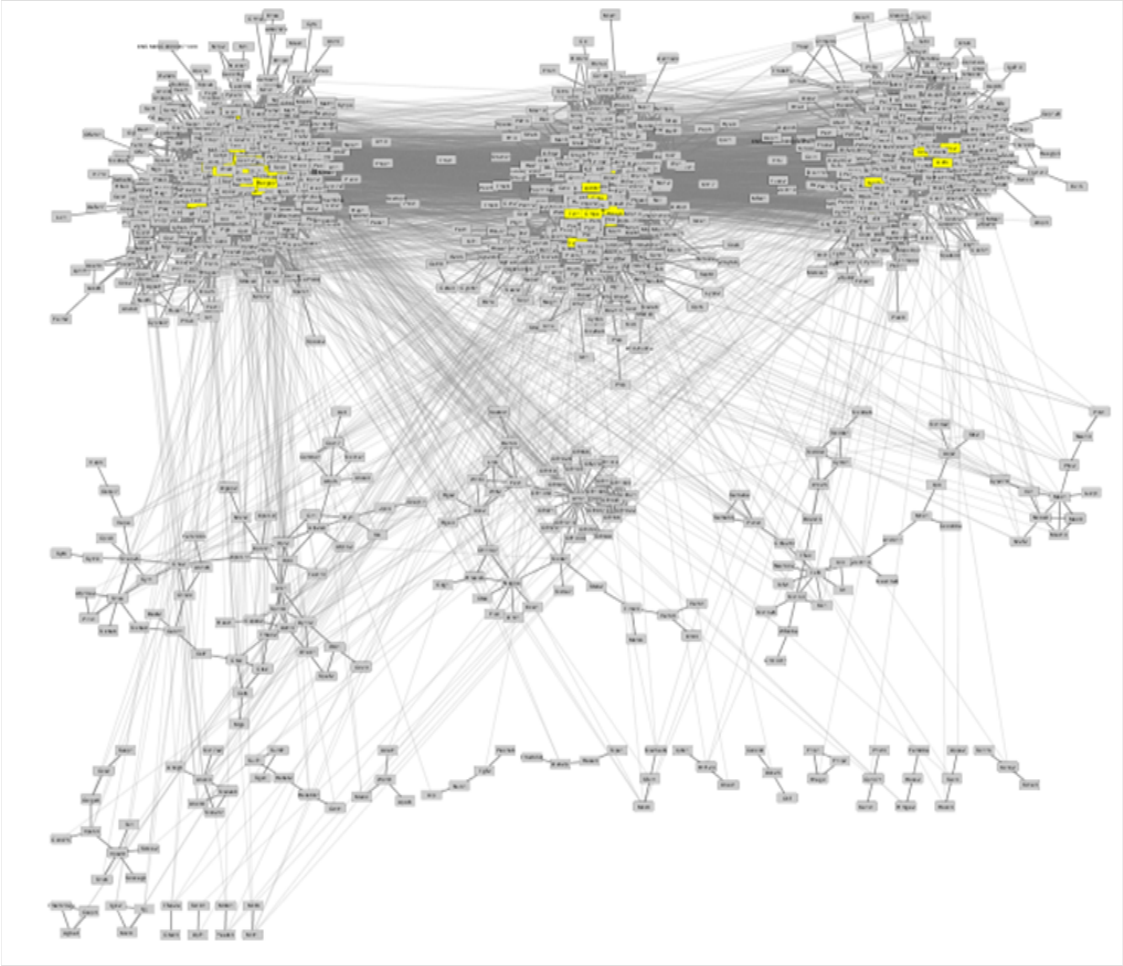


Figura 25: Clústers topológicos en la PPIN LvsCtrl

En total se obtuvieron 27 clústeres, 3 con un gran número de nodos y otros clusters más pequeños. Se observan los nodos centrales identificados anteriormente al centro de los 3 primeros clústers. La siguiente tabla resume el número de nodos y ejes obtenidos para los 5 primeros clusters:

Cluster	Nodos	Ejes
1	408	3106
2	304	2170
3	297	2211
4	56	78
5	48	83

6.1.3.1. Nodos centrales de los módulos

Las siguientes tablas muestran los 5 nodos con grado más alto para cada clúster.

Cuadro 25: Lista de nodos con grado más alto en los clústers de LvsCtrl

Entrez	Symbol	description	Degree	Centrality	logFC.d3	logFC.d4	logFC.d5	var
Cluster1								
22059	Trp53	transformation related protein 53	203	0.128	-1.402	-2.305	-2.615	6.323
21926	Tnf	tumor necrosis factor	178	0.068	1.090	0.424	0.699	2.213
20846	Stat1	signal transducer and activator of transcription 1	135	0.024	2.800	2.076	1.962	6.838
15978	Ifng	interferon gamma	121	0.024	3.230	2.314	2.433	7.977
17869	Myc	myelocytomatosis oncogene	112	0.049	1.078	0.952	0.993	3.024
Cluster2								
229363	Gmps	guanine monophosphate synthetase	105	0.032	0.384	1.022	0.720	2.126
15481	Hspa8	heat shock protein 8	103	0.041	-0.736	0.995	1.271	3.001
23918	Impdh2	inosine 5'-phosphate dehydrogenase 2	92	0.022	0.194	1.048	0.371	1.613
107508	Eprs	glutamyl-prolyl-tRNA synthetase	89	0.015	-1.926	-1.364	-1.643	4.933
20020	Polr2a	polymerase (RNA) II (DNA directed) polypeptide A	72	0.020	-0.132	-0.831	-1.056	2.018
Cluster3								
66725	Lrrk2	leucine-rich repeat kinase 2	138	0.067	0.905	0.808	1.208	2.921
11461	Actb	actin, beta	131	0.062	-0.872	-2.790	-2.382	6.044
53606	Isg15	ISG15 ubiquitin-like modifier	106	0.015	2.906	2.372	1.575	6.853
23962	Oasl2	2'-5' oligoadenylate synthetase-like 2	99	0.009	1.939	2.056	2.069	6.064
246728	Oas2	2'-5' oligoadenylate synthetase 2	96	0.008	2.133	1.744	2.041	5.918
Cluster4								
80297	Spnb4	spectrin beta 4	19	0.001	-1.506	-2.164	-2.530	6.200
20617	Snca	synuclein, alpha	14	0.006	-0.344	-0.849	-1.539	2.731
68097	Dynll2	dynein light chain LC8-type 2	13	0.003	0.773	1.167	0.926	2.866
67665	Dctn4	dynactin 4	12	0.002	1.384	0.204	0.126	1.714
64138	Ctsz	cathepsin Z	11	0.002	0.881	1.144	1.028	3.053
Cluster5								
109689	Arrb1	arrestin, beta 1	47	0.011	-0.413	-0.577	-1.392	2.382
27426	Nagpa	N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase	20	0.005	-0.199	-0.913	-1.242	2.354
20544	Slc9a1	solute carrier family 9 (sodium/hydrogen exchanger), member 1	13	0.005	-0.355	-0.590	-1.151	2.096
23805	Apc2	adenomatosis polyposis coli 2	12	0.001	-0.414	-1.072	-0.899	2.385
258582	Olfr1022	olfactory receptor 1022	11	0.000	-0.625	-1.242	-1.474	3.340

6.1.4. Análisis de módulos temporales en la malaria L

A continuación se muestra el resultado del análisis de clustering temporal realizado con `TiCoNe` para los genes diferencialmente expresados en LvsCtrl:

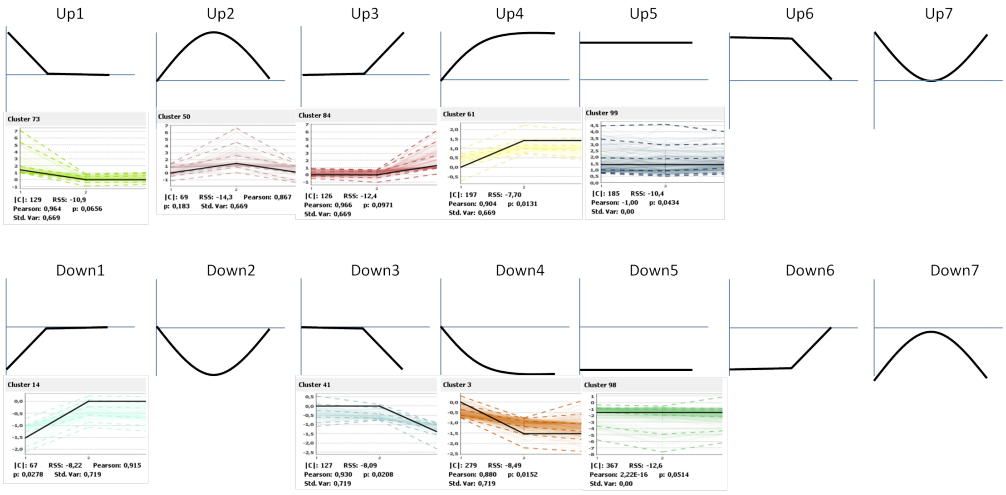


Figura 26: Clústers temporales de los DEG en LvsCtrl

En la siguiente tabla se resumen las características de los patrones encontrados.

Patrón	ClusterID	Nº genes	Pearson	RSS	P-valor
Up1	73	129	0,964	-10,9	0,066
Up2	50	69	0,867	-14,3	0,183
Up3	84	126	0,966	-12,4	0,097
Up4	61	197	0,904	-7,7	0,013*
Up5	99	185	-1	-10,4	0,043
Down1	14	67	0,915	-8,22	0,028*
Down3	41	127	0,93	-8,09	0,021*
Down4	3	279	0,88	-8,49	0,015*
Down5	98	367	0	-12,6	0,051

Se han identificado 9 patrones distintos de entre los 14 esperados, 4 de los cuales muestran un coeficiente de correlación (Pearson) elevado y un p-valor menor a 0.05 (en la tabla se indican con un asterisco).

6.1.4.1. Análisis de conectividad entre los clústeres temporales

La siguiente red muestra la relación entre los clusteres temporales, donde los ejes representan los enriquecimientos (verde) y depleciones (azul) significativos ($p < 0.01$).

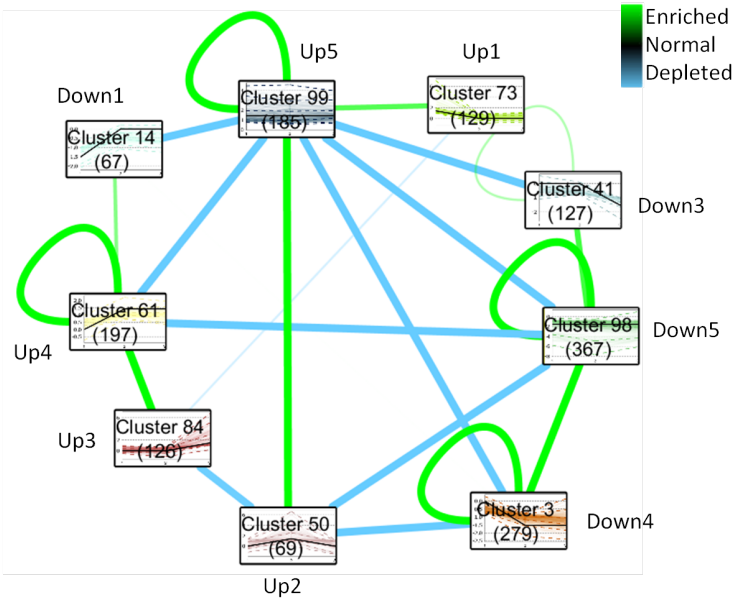


Figura 27: Conectividad entre los clústers temporales de LvsCtrl

Se observan algunos clústeres temporales más conectados entre sí que lo esperado por azar, sugiriendo la presencia de subconjuntos co-regulados más grandes. Encontramos 3 subconjuntos formados por los siguientes clústers:

1) `Up2-Up5-Up1` (383 nodos)
2) `Up3-Up4` (323 nodos)
3) `Down5-Down4-Down3` (773 nodos)

Comparación de los clusteres topológicos con los temporales

A continuación comparamos los clústeres temporales con los clústers topológicos obtenidos previamente. Para ello, coloreamos los nodos de la red de clústeres topológicos en función de los patrones temporales obtenidos.

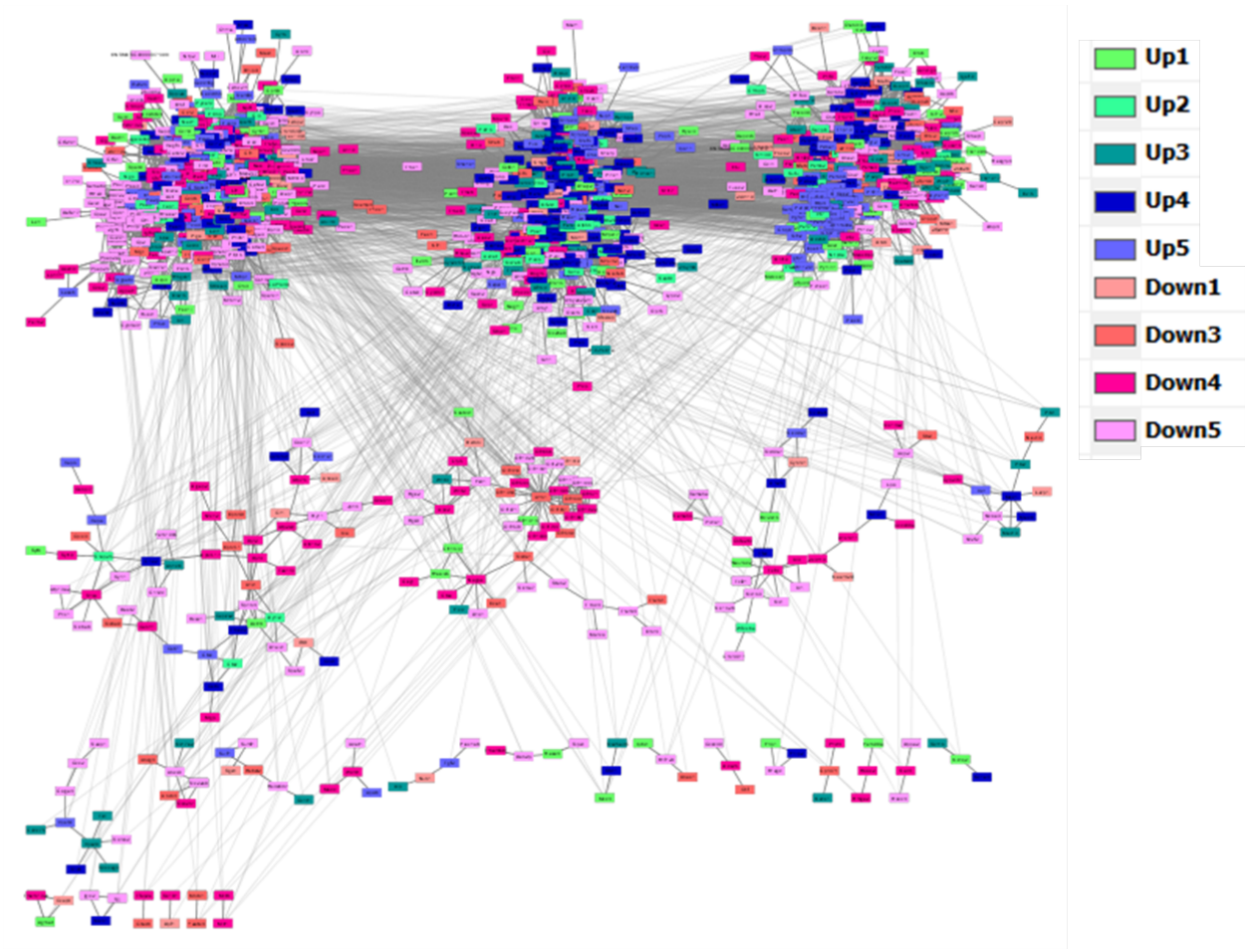


Figura 28: Visualización de los clústers temporales con los clústers topológicos en la PPIN LvsCtrl

La siguiente tabla resume la repartición de los patrones temporales entre los distintos clústeres (se muestra el porcentaje de nodos de cada cluster que corresponden al patrón indicado).

Cuadro 27: Distribución de los patrones temporales de LvsCtrl en los distintos clusters

	Clust1	**Clust2**	**Clust3**	**Clust4**	**Clust5**
Up1	8.6	6.6	9.4	3.6	8.3
Up2	4.4	6.2	5.4	5.4	0.0
Up3	5.4	9.2	9.1	3.6	4.2
Up4	11.0	27.0	11.8	10.7	0.0
Up5	13.0	11.8	21.5	8.9	0.0
Down1	3.4	3.3	7.1	5.4	4.2
Down3	7.4	5.6	7.1	10.7	18.8
Down4	16.2	13.5	12.8	26.8	29.2
Down5	30.6	16.8	15.8	25.0	35.4

Se observa un subcluster con los genes `Down5` en el `Cluster1` y de los genes `Up5` en el `Cluster3`. El `Cluster2` está enriquecido en genes con el patrón `Up4` y el `Cluster4` y `Cluster5` en los patrones `Down4` y `Down5`.

Coloración de los nodos centrales de la red en función de los perfiles temporales

Por otro lado, podemos colorear los 35 nodos centrales de la red (>=600 vecinos a distancia <=2) en función de su perfil temporal para determinar si alguno de ellos se encuentra más representado (figura 29).

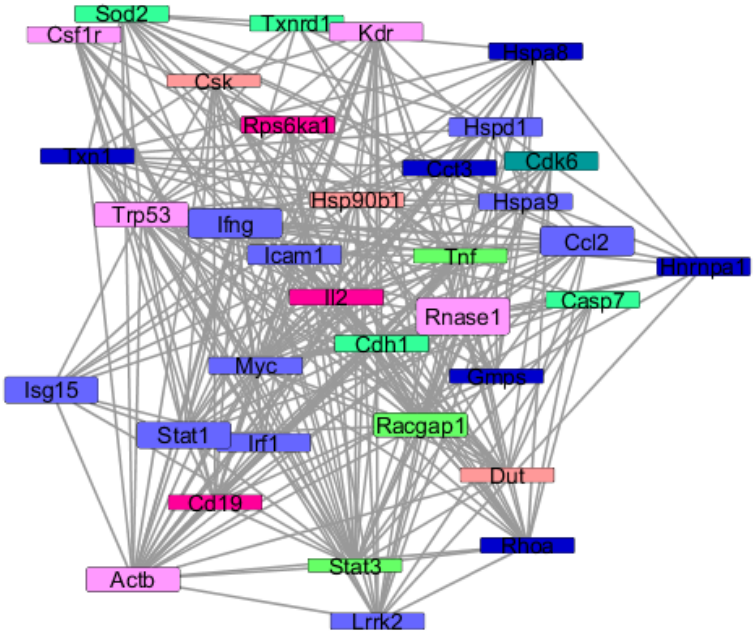


Figura 29: Visualización de los nodos centrales de la red LvsCtrl en función del perfil temporal (color) y la variación acumulada (altura caja)

6.1.4.2. Nodos centrales de los clusters temporales

Las siguientes tablas muestran los 5 nodos con grado más alto para cada clúster temporal.

Cuadro 28: Lista de nodos con grado más alto en los clústers temporales de LvsCtrl

Entrez	Symbol	description	Degree	Centrality	logFC.d3	logFC.d4	logFC.d5	var
Up1								
21926	Tnf	tumor necrosis factor	178	0.068	1.090	0.424	0.699	2.213
20848	Stat3	signal transducer and activator of transcription 3	106	0.027	1.195	0.239	0.335	1.769
26934	Racgap1	Rac GTPase-activating protein 1	80	0.032	5.495	0.150	-0.206	5.851
66222	Serpinb1a	serine (or cysteine) peptidase inhibitor, clade B, member 1a	53	0.005	1.112	0.443	0.915	2.469
70110	Ifi35	interferon-induced protein 35	49	0.000	1.049	0.482	0.532	2.063
Up2								
54123	Irf7	interferon regulatory factor 7	81	0.006	1.086	1.610	0.804	3.500
16912	Psmb9	proteasome (prosome, macropain) subunit, beta type 9 (large multifunctional peptidase 2)	59	0.002	1.311	1.168	1.161	3.639
20656	Sod2	superoxide dismutase 2, mitochondrial	52	0.008	0.807	1.137	1.029	2.974
12550	Cdh1	cadherin 1	47	0.008	0.135	2.310	0.497	2.942
50493	Txnrd1	thioredoxin reductase 1	44	0.012	0.759	1.095	0.869	2.723
Up3								
12571	Cdk6	cyclin-dependent kinase 6	54	0.012	0.045	0.045	3.800	3.891
230073	Ddx58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	52	0.001	0.146	0.116	3.515	3.777
228889	Ddx27	DEAD (Asp-Glu-Ala-Asp) box polypeptide 27	40	0.006	0.541	0.583	1.181	2.305
56095	Ftsj3	FtsJ homolog 3 (E. coli)	36	0.003	0.456	0.528	1.037	2.022
14745	Lpar1	lysophosphatidic acid receptor 1	35	0.005	0.185	0.646	1.026	1.857
Up4								
229363	Gmps	guanine monophosphate synthetase	105	0.032	0.384	1.022	0.720	2.126
15481	Hspa8	heat shock protein 8	103	0.041	-0.736	0.995	1.271	3.001
23918	Impdh2	inosine 5'-phosphate dehydrogenase 2	92	0.022	0.194	1.048	0.371	1.613
14688	Gnb1	guanine nucleotide binding protein (G protein), beta 1	86	0.033	0.245	1.113	0.787	2.144
11848	Rhoa	ras homolog gene family, member A	76	0.018	0.662	1.105	0.878	2.646
Up5								
66725	Lrrk2	leucine-rich repeat kinase 2	138	0.067	0.905	0.808	1.208	2.921
20846	Stat1	signal transducer and activator of transcription 1	135	0.024	2.800	2.076	1.962	6.838
15978	Ifng	interferon gamma	121	0.024	3.230	2.314	2.433	7.977
17869	Myc	myelocytomatosis oncogene	112	0.049	1.078	0.952	0.993	3.024
15945	Cxcl10	chemokine (C-X-C motif) ligand 10	107	0.011	3.767	2.482	3.218	9.467
Down1								
68556	Uckl1	uridine-cytidine kinase 1-like 1	61	0.020	-1.034	-0.482	-0.384	1.900
22027	Hsp90b1	heat shock protein 90, beta (Grp94), member 1	59	0.012	-1.441	0.054	-0.787	2.282
110355	Adrbk1	adrenergic receptor kinase, beta 1	52	0.013	-1.097	-0.223	-0.188	1.508
110074	Dut	deoxyuridine triphosphatase	48	0.006	-1.443	0.195	-0.543	2.182
67653	ENSMUSG00000036463	RIKEN cDNA 4930544G11 gene	48	0.005	-1.001	0.067	-0.253	1.321
Down3								
19983	Rpl5	ribosomal protein L5	49	0.007	-0.123	-0.462	-1.468	2.053
53607	Snrpa	small nuclear ribonucleoprotein polypeptide A	48	0.007	-0.853	-0.661	-1.028	2.542
109689	Arrb1	arrestin, beta 1	47	0.011	-0.413	-0.577	-1.392	2.382
19088	Prkar2b	protein kinase, cAMP dependent regulatory, type II beta	35	0.010	0.123	-0.621	-1.286	2.030
66156	Anapc11	anaphase promoting complex subunit 11	31	0.002	-0.504	-0.736	-1.066	2.307
Down4								
16183	Il2	interleukin 2	110	0.020	-0.748	-0.934	-1.041	2.723
20020	Polr2a	polymerase (RNA) II (DNA directed) polypeptide A	72	0.020	-0.132	-0.831	-1.056	2.018
12478	Cd19	CD19 antigen	63	0.008	-0.388	-0.981	-1.007	2.376
170758	Rac3	RAS-related C3 botulinum substrate 3	52	0.011	-0.386	-0.808	-1.248	2.443
18709	Pik3r2	phosphatidylinositol 3-kinase, regulatory subunit, polypeptide 2 (p85 beta)	45	0.005	-0.549	-0.820	-1.085	2.454
Down5								
22059	Trp53	transformation related protein 53	203	0.128	-1.402	-2.305	-2.615	6.323
11461	Actb	actin, beta	131	0.062	-0.872	-2.790	-2.382	6.044
107508	Eprs	glutamyl-prolyl-tRNA synthetase	89	0.015	-1.926	-1.364	-1.643	4.933
16542	Kdr	kinase insert domain protein receptor	74	0.018	-1.441	-1.130	-1.171	3.741
20315	Cxcl12	chemokine (C-X-C motif) ligand 12	64	0.007	-2.230	-3.072	-3.452	8.754

6.1.5. ABS en la malaria L

6.1.5.1. ABS de los top 300 genes con más variación acumulada

El siguiente diagrama muestra la distribución de los grupos enriquecidos encontrados en los top 300 DEGs con más variación acumulada a lo largo de los días.

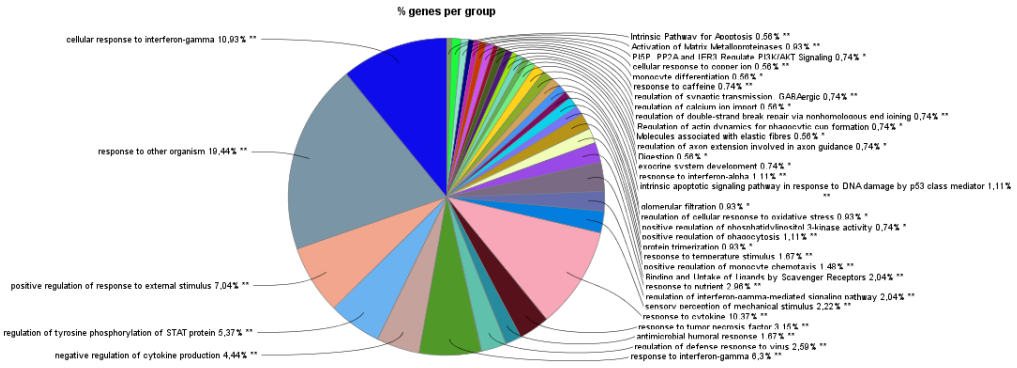


Figura 30: Diagrama de sectores de los procesos/vías enriquecidos en los top 300 DEGs de LvsCtrl

6.1.5.2. ABS de los nodos centrales

El siguiente diagrama muestra la distribución de los grupos enriquecidos encontrados en los 35 nodos centrales anteriormente identificados (>=600 vecinos a distancia 2).

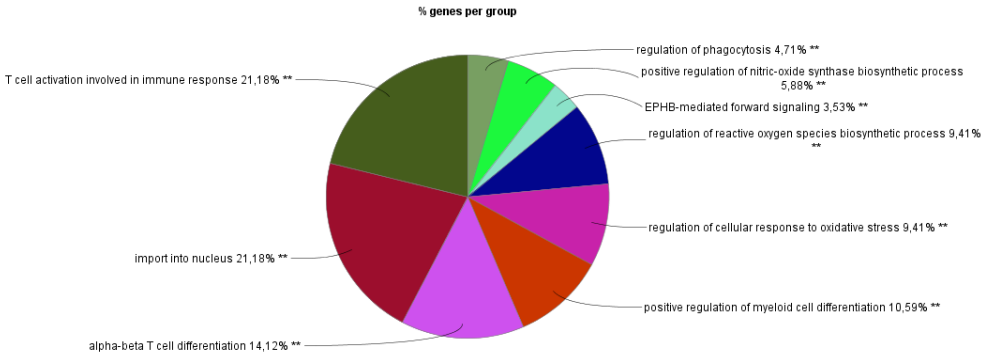
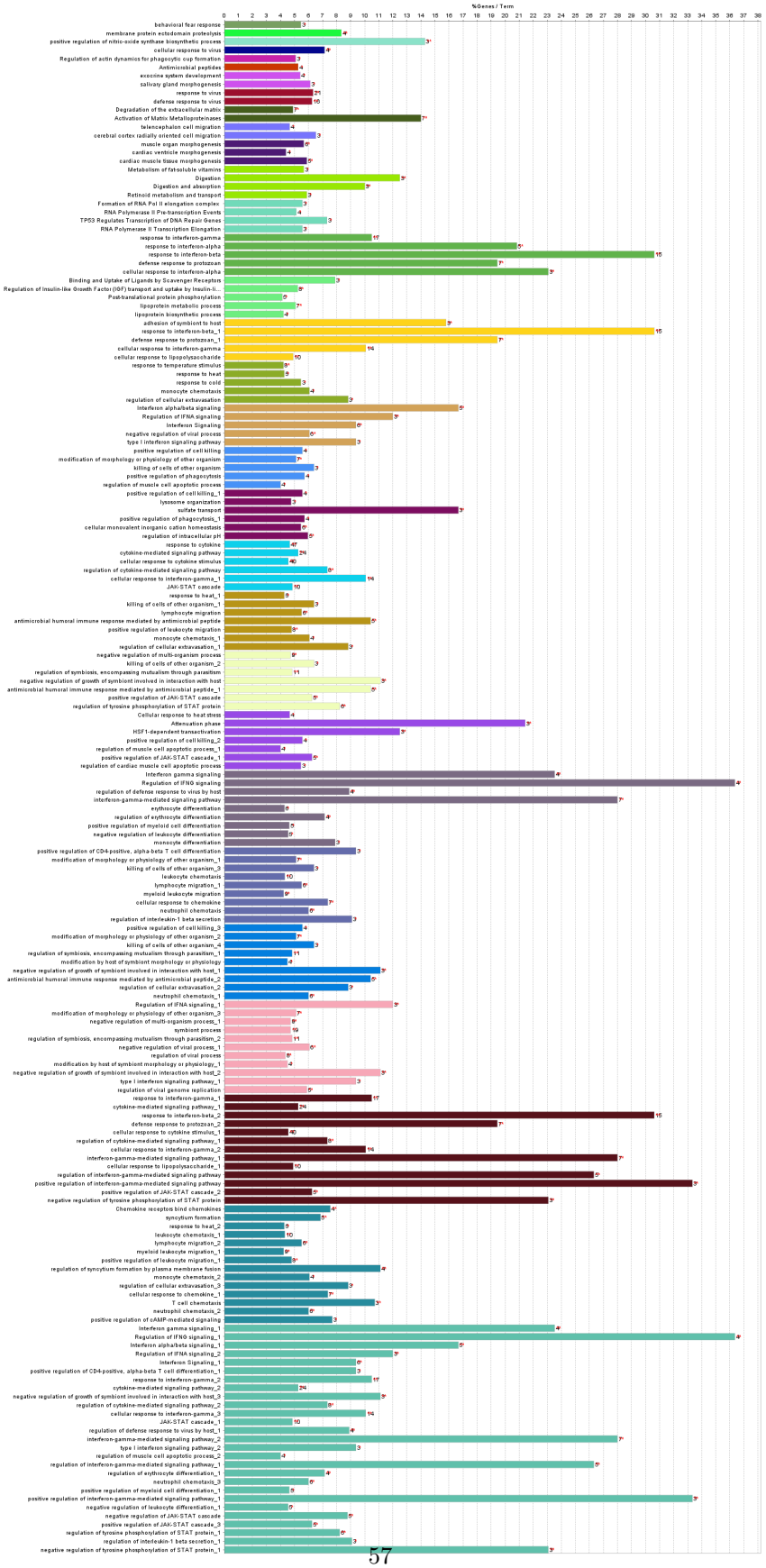


Figura 31: Diagrama de sectores de los procesos/vías enriquecidos en los nodos centrales de LvsCtrl

6.2. Código utilizado para el análisis de microarrays con R

El código utilizado para el análisis de microarrays con R está disponible en *github* en el siguiente *enlace*.

6.3. Procesos/vías enriquecidos en los top 300 DEGs de NLvsCtrl (barplot)



7. Bibliografía

‘

1. Mueller I, Galinski M, Baird J, Carlton J, Kochar D, Alonso P, Del Portillo H. Key gaps in the knowledge of plasmodium vivax, a neglected human malaria parasite. Lancet Infect Dis. 2009;9:555–66.

2. Del Portillo H, Ferrer M, Brugat T, Martin-Jaular L, Langhorne J, Lacerda M. The role of the spleen in malaria. Cellular Microbiology. 2012;14:343–55.

3. Martin-Jaular L, Ferrer M, Calvo M, Rosanas-Urgell A, Kalko S, Graewe S, Soria G, Cortadellas N, Ordi J, Planas A, Burns J, Heussler V, Del Portillo HA. Strain-specific spleen remodelling in plasmodium yoelii infections in balb/c mice facilitates adherence and spleen macrophage-clearance escape. Cellular microbiology. 2011;13:109–22.

4. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. Nat Rev Genet. 2011;12:58–68.

5. Rodin AS, Gogoshin G, Boerwinkle E. Systems biology data analysis methodology in pharmacogenomics. Pharmacogenomics. 2011;12:1349–60.

6. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE. ClusterMaker: A multi-algorithm clustering plugin for cytoscape. BMC Bioinformatics. 2011;12.

7. Wiwie C, Rauch A, Haakonsson A, Nandez IB-H, Blagoev B, Mandrup S, Röttger R, Baumbach J. Elucidation of time-dependent systems biology cell response patterns with time course network enrichment. arXiv: 171010262. 2017.

8. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J. ClueGO: A cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25:1091–3.

9. Barabási AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nature Reviews Genetics. 2004;5:101–13.

10. Ma'ayan A. Introduction to network analysis in systems biology. Sci Signal. 2011;4.

11. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG. Using graph theory to analyze biological networks. BioData Mining. 2011;4.

12. Sevimoglu T, Arga KY. The role of protein interaction networks in systems biomedicine. Computational and Structural Biotechnology Journal. 2014;11:22–7.

13. Koh GC, Porras P, Aranda B, Hermjakob H, Orchard SE. Analyzing protein-protein interaction networks. Journal of Proteome Research. 2012;11:2014–31.

14. Kamburov A, Stelzl U, Herwig R. IntScore: A web tool for confidence scoring of biological interactions. Nucleic Acids Research. 2012;40:140–6.

15. Dam S van, Võsa U, Graaf A van der, Franke L, Magalhães JP de. Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in Bioinformatics. 2017;19:575–92.

16. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009;37:1–13.

17. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010;5:463–6.

18. Quezada H, Guzmán-Ortiz A, Díaz-Sánchez H, Valle-Rios R, Aguirre-hernández J. Omics-based biomarkers : Current status and potential use in the clinic. Bol Med Hosp Infant Mex. 2017;74:219–26.

19. Dixon S, Stockwell B. Identifying druggable disease-modifying gene products. Curr Opin Chem Biol.

2009;13:549–55.

20. Hopkins A, Groom C. The druggable genome. Nature Reviews Drug Discovery. 2002;1:727–30.

21. ME R, B P, D W, Y H, CW L, W S, GK S. Limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Research. 2015;43:e47.

22. Mering C von, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Research. 2005;33:433–7.

23. Robbiani DF, Deroubaix S, Feldhahn N, Oliveira TY, Wang Q, Jankovic M, Silva IT, Rommel PC, Bosque D, Eisenreich T, Nussenzweig A, Nussenzweig MC. Plasmodium infection promotes genomic instability and aid dependent b cell lymphoma. Cell. 2015;162:727–37.

24. Kaushansky A, Ye AS, Austin LS, Mikolajczak SA, Vaughan AM, Camargo N, Metzger PG, Douglass AN, MacBeath G, Kappe SH. Interrogation of infected hepatocyte signaling reveals that suppression of host p53 is critical for plasmodium liver stage infection. Cell Rep. 2013;3:630–7.

25. M. G, J. NME. Community structure in social and biological networks. Proc Natl Acad Sci. 2002;99:7821–6.

26. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4.

27. Su G, Kuchinsky A, Morris JH, States DJ, Meng F. GLay: Community structure analysis of biological networks. Bioinformatics. 2010;26.

28. Kaufman L, Rousseew P. Finding groups in data: An introduction to cluster analysis. “John Wiley & Sons, Inc.”; 2005.

29. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology. 2005;4.

30. Gambardella G, Moretti M, Cegli R de, Cardone L, Peron A, Bernardo D di. Differential network analysis for the identification of condition-specific pathway activity and regulation. Bioinformatics. 2013;29:1776–85.