RESEARCH

Red de los targets de SARS-CoV2

Irene Romero Granados*, Paula Andújar Zambrano*, Rosario García Morales* and Soledad del Castillo Carrera*

*Correspondence:
ireero99@uma.es;
paulandujar@uma.es;
0619884107@uma.es;
delcastillosoledad@uma.es
ETSI Informática, Universidad de Málaga, Málaga, España
Full list of author information is available at the end of the article

Abstract

Este proyecto pretende estudiar las interacciones que se producen entre las 29 proteínas del virus SARS-COV-2 y el interactoma humano. Para ello, se recopilarán los datos de interacción comentados, con los que se procederá a realizar un análisis sobre ellos y construir la red de proteínas que conforman los objetivos principales para el virus. En cuanto a las herramientas que se utilizarán, serán la base de datos biológica UniProt para la obtención de los datos de entrada, y el lenguaje de R para los métodos de análisis.

Keywords: SARS-COV-2; interactoma; R

1 Introducción

La familia de los coronavirus son virus infecciosos a los que se llama así debido a que en su superficie tienen puntas en forma de corona. A esta familia se les unió en 2019 el conocido SARS-CoV-2, que ha dado lugar al coronavirus 2 o COVID-19. Esta enfermedad es una enfermedad infecciosa que afecta a las vías respiratorias, de manera leve a moderada. Sin embargo esta enfermedad en personas mayores o con patologías previas puede hacer que se desarrolle la enfermedad con consecuencias o síntomas más graves, pudiendo producir hasta la muerte.

El coronavirus actualmente es considerado un problema de salud global, ya que debido a esta pandemia se han contagiado hasta ahora unas 369.955.862 personas y han fallecido un total de 5.650.738 personas.

Es por esto que es esencial el estudio de este virus, tanto de sus genes, sus proteínas o como interacciona con el ser humano.

A día de hoy tras toda la inversión mundial que se ha hecho para poder poner fin a este virus, se sabe que el SARS-CoV-2 está formado por 29 proteínas que interactúan con las células del ser humano pudiendo producir síntomas respiratorios graves hasta poder causar la muerte. A estas interacciones moleculares binarias proteína-proteína se les llama interactoma.

El interactoma sirve como de punto de partida para estudiar los posibles fármacos que podrían bloquear dichas interacciones y así evitar que el virus entre a la célula y se replique. Gracias al estudio del interactoma ha sido posible la realización de vacunas contra el COVID-19.

En este proyecto vamos a crear y estudiar la red de interacciones de las proteínas del SARS-CoV-2 con las proteínas humanas, y así poder obtener cuales son las principales funciones biológicas humanas en las que este virus interviene y relacionarlo con la realidad. Todos los recursos usados para la obtención de dicha información la podremos encontrar en el GitHub proporcionado.

2 Materiales y métodos

2.1 Carga de librerías y datos

Para poder llevar a cabo este trabajo, el primer paso a realizar es la carga de librerías necesarias y la carga de datos. Antes de cargar los datos, estos han sido descargados de Uniprot (https://www.uniprot.org/) en formato .csv para poder llevar a cabo el análisis de la red. Una vez ha sido añadido este fichero al directorio correspondiente (data), se procederá a la carga de librerías. Las librerías que han sido utilizadas en este proyecto son las siguientes:

- igraph: Esta librería permite realizar análisis de redes, por lo cual, proporciona funciones para manipular gráficos con facilidad.
- dplyr: Esta librería proporciona métodos para poder manejar los ficheros de datos.
- ggplot2: Esta librería es un paquete de visualización de datos.
- zoo: Esta librería está especialmente dirigida a series temporales irregulares de vectores/matrices y factores numéricos.
- STRINGdb: Este paquete proporciona una interfaz para la base de datos STRING de interacciones proteína-proteína.

Después de tener las librerías necesarias y saber la funcionalidad de cada una de ellas, se procederá a cargar el archivo en una variable llamada "data" mediante el método "read.csv()"

Seguidamente, se filtrarán las entradas utilizando el paquete "dplyr" en las que la columna Entry.Name contenga en su nombre "HUMAN" ya que estos son los datos que interesan en esta práctica.

El código que ha sido implementado para la carga de librerías y datos es el siguiente:

```
library(igraph)
library(dplyr)
library(ggplot2)
library(zoo)
library(STRINGdb)
library(linkcomm)

# Cargamos el archivo de datos
data <- read.csv("code/data/uniprot1.csv", header=T, sep=";")

# Filtramos las entradas y cambiamos el Entry name
data2 <- dplyr::filter(data, grepl("HUMAN",Entry.Name))
data2$Entry.Name2 <- sapply(data2$Entry.Name, function(i) gsub("_
    HUMAN", "", i))
contenidos...
```

2.2 Mapeo y primera capa de la red

A continuación, utilizando la librería STRINGdb, se realiza un mapeo con los datos ya filtrados. Se guardará los hits de string en una imagen png en el directorio de los resultados (results)

Seguidamente, se creará la primera capa de la red y se guardará el resultado de esta primera capa en una imagen png.

2.3 Grado de Distribución

El grado de un vértice en una red es el número de conexiones asociadas a un vértice. Haciendo un recuento en una red del número de nodos por cada grado se tiene el grado de distribución. Este es entendido igualmente como la distribución de probabilidad de un grado en la red.

En esta práctica, se ha obtenido el grado de la red mediante el método "degree" y después se ha obtenido el grado de distribución mediante el método "degree_distribution".

Seguidamente se ha obtenido el coeficiente de agrupamiento mediante el método "transitivity" y por último, ha sido calculada la distancia euclídea.

2.4 Robustez

Por último, utilizando los métodos proporcionados en el campus virtual, se ha calculado la robustez de la red de genes. Esta funciona para conocer si la red que se está estudiando es un sistema fuerte y si esta sigue manteniendo sus funciones en la presencia de "ataques" (errores o fallos). Además, ha sido calculada frente a ataques aleatorios como a ataques dirigidos, pero también, han sido combinados ambos ataques.

2.5 Linked Communities

Una vez se tiene los datos ya mapeados y filtrados, se pasa a realizar Linked Communities, para ello, ha sido utilizado el paquete Linkcomm. Este paquete, proporciona las herramientas necesarias para generar, visualizar y analizar comunidades dentro de un grafo.

Al obtener las comunidades vinculadas, estas serán guardadas en la carpeta de results y, además, se han obtenido los tamaños de los clusters. Seguidamente, se ha obtenido la modularidad de las comunidades y se ha guardado el resultado de un cluster aleatorio. Además, se han obtenido las comunidades completamente anidadas dentro de la comunidad más grande de nodos.

2.6 Enriquecimiento Funcional

Al realizar Linked Communities, es decir, el agrupamiento de estas comunidades, se realizará el enriquecimiento funcional. Este enriquecimiento es utilizado para obtener los procesos biológicos de los grupos que han sido formados en el clustering. En primer lugar, ha sido diseñado el enriquecimiento mediante STRINGdb. Después, se ha realizado el enriquecimiento con GO, es decir, con una ontología génica y también, se ha realizado el enriquecimiento con la ontología KEGG.

El enriquecimiento ha sido utilizado con los clusters de mayor tamaño y de mayor modularidad.

3 Resultados

3.1 Red de interacciones y robustez

En la imagen Figure 1 se muestra la red de interacciones del ser humano con las proteínas del SARS-CoV. Como podemos ver el SARS-CoV interacciona con 89 proteínas humanas, produciendo un total de 475 interacciones.

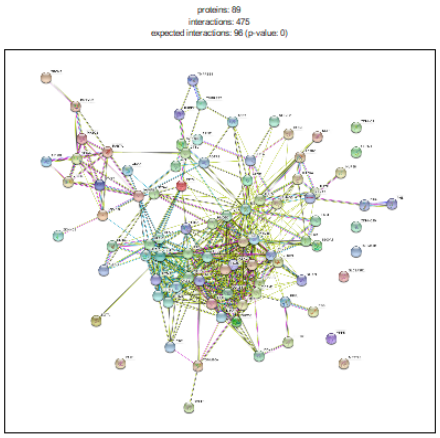


Figure 1 *Red de interacciones del SARS-CoV con las proteínas humanas*

Tras eliminar los nodos que no están conectados, hemos obtenido la red real de interacciones que podemos ver a continuación. Sin embargo hay demasiadas conexiones como para poder distinguir los nodos. Es por ello que realizaremos los pasos siguientes de clustering, para así poder extraer la información relevante de la red. Podemos observar la red de interacciones en la figura Figure 2.

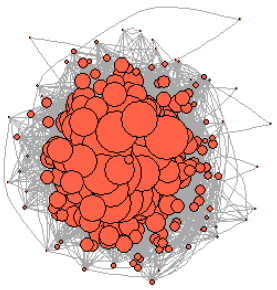


Figure 2 *Red de interacciones del SARS-CoV con las proteínas humanas tras un proceso de filtrado*

Antes de empezar con ese proceso vamos a estudiar diferentes aspectos de nuestra red. En primer lugar si observamos la imagen Figure 3, vemos que el la distribución de grado sigue la ley de potencias, por lo tanto nuestra red sigue un modelo de free-scale, lo cual era predecible al estar tratando con una red real.

Podemos ver una gran cantidad de hubs.

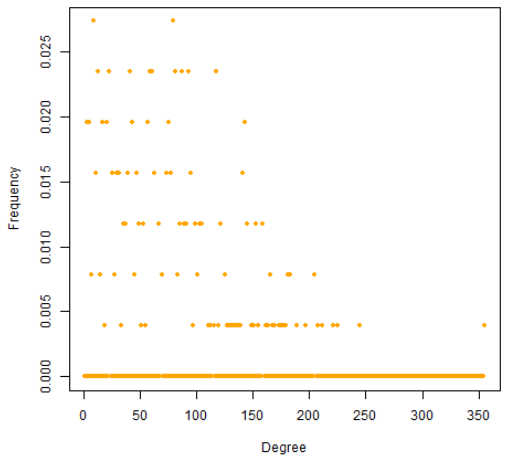


Figure 3 *Distribución de grado*

El coeficiente medio de agrupamiento es de 0.605, lo cual es bastante alto. Además se puede observar la característica propia de las redes reales la cual afirma que conforme el grado de los nodos aumenta, el coeficiente de agrupamiento disminuye. La gráfica del coeficiente de agrupamiento esta en la Figure 4.

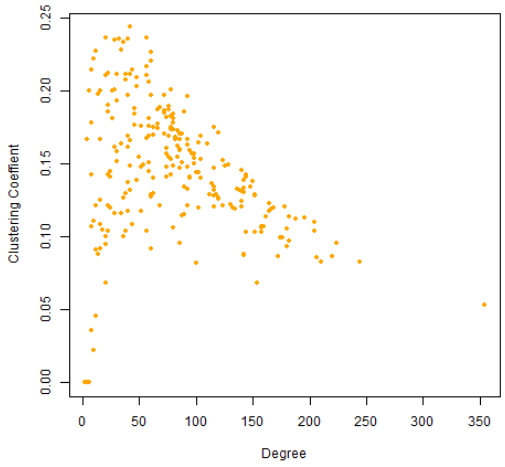


Figure 4 *Coeficiente de Agrupamiento*

La distancia media entre nodos es de 2.03, una medida muy pequeña que puede significar que los nodos tienen un alto índice de conexiones. Esta medida es la que le da la propiedad de mundo pequeño, es decir, la distancia entre nodos elegidos al azar en una red es muy pequeña. Esto se puede observar de forma visual en la Figure 5.

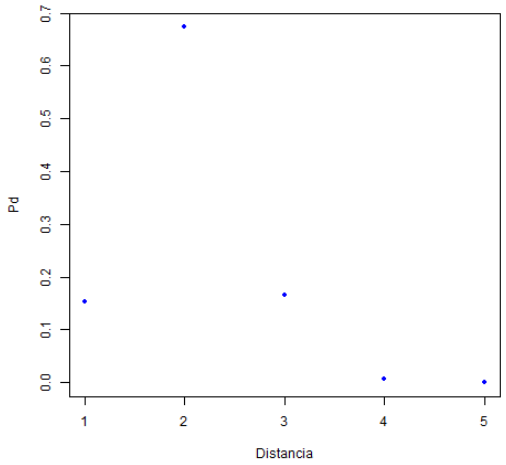


Figure 5 *Distancia entre nodos*

Por último vamos a estudiar la robustez de nuestra red. Para poder estudiar cual es la capacidad de nuestra red de mantener sus funciones frente a la presencia de "ataques" y ver cuán de adaptable es, usamos la robustez. Podemos observar en la Figure 6 que para ataques aleatorios es bastante robusta, mientras que para ataques dirigidos es más débil. Pues a que tenemos una red real, estos resultados resultan obvios, ya que es más fácil destruir una red si atacas a puntos estratégicos como son los hubs, dónde el tamaño de la componente conexa se reduce drásticamente cuando eliminamos una pequeña fracción de los nodos (hubs).

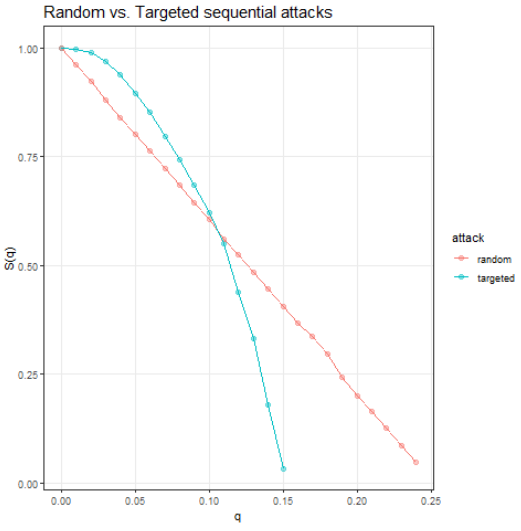


Figure 6 *Robustez frente a ataques dirigidos y aleatorios*

3.2 Linked Communities

En esta sección se explicarán los resultados obtenidos aplicar los métodos de comunidades enlazadas a nuestra red, para los cuales hemos usado el paquete linkcomm. Primero de todo, mencionar que las comunidades de una red son subredes que están altamente relacionados entre sí debido a que sus proteínas poseen algunas características similares. Estas pueden ser funciones biológicas, coeficiente de clustering, o tienen perturbaciones en su secuencia genética que pueden ser enlazadas a una enfermedad común.

Para la identificación de las comunidades más relevantes en nuestra red, hemos aplicado una función predefinida que devuelve el conjunto de las comunidades identificadas mediante la aplicación de un algoritmo de 'single clustering'. Se ha guardado una imagen Figure 7 del resumen de estas comunidades para un resultado más visual.

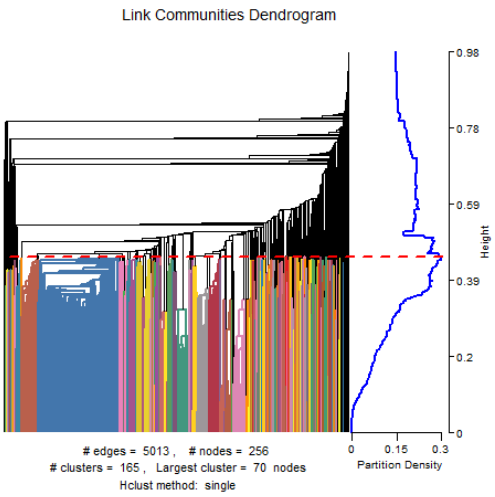


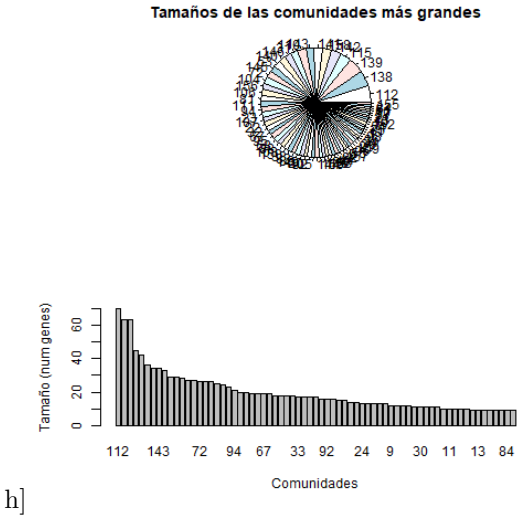
Figure 7 *Esquema de las comunidades obtenidas*

Podemos observar que este método ha definido 165 comunidades diferentes, la mayor de estas tiene 70 nodos lo cual podemos deducir que hay varias comunidades que comparten nodos en nuestra red.

Para poder analizar las comunidades y obtener las que consideremos más relevantes, han sido filtradas por tamaño y modularidad. Lo cual nos ha facilitado la búsqueda de la comunidad más grande y dos comunidades que tienen mayor modularidad (y respectivamente). Las gráficas generadas para este propósito son Figure 8 y Figure 9.

Hemos podido extraer la comunidad 112 siendo esta la más grande y las comunidades 78 y 22 con una mayor modularidad. A estas comunidades encontradas se les realizará un enriquecimiento funcional para observar las funciones y características comunes que poseen y así poder sacar conclusiones adecuadas.

Para una mejor visualización de los resultados, se ha cambiado el diseño de la gráfica al de Fruchterman Reingold (se puede observar en la Figure 10), mostrando solo nodos que pertenezcan a 10 o más comunidades.



h]

Figure 8 *Comunidades por tamaño*

Por otra parte, se han obtenido las comunidades anidadas y se han filtrado para que se muestren las que son independientes de las demás. Estas se observar en la Figure 11.

3.3 Enriquecimiento funcional

En esta sección se van a mostrar los resultados obtenidos al realizar el enriquecimiento funcional con GO y con KEGG, mediante el uso de STRINGdb, para los clústeres elegidos. Se ha guardado la información del enriquecimiento en archivos de tipo csv, y se va a mostrar una imagen de los mismos.

3.3.1 Clúster 112

Enriquecimiento con GO En la siguiente Figure 12 se puede observar que las funciones biológicas visibles están relacionadas con la defensa del organismo, por lo que se puede deducir que son proteínas que forman parte del sistema inmunológico de nuestro organismo. No obstante, en la imagen solo aparecen algunas funciones biológicas. Si se sigue observando el archivo generado nos encontramos con los resultados localizados en la Figure 13:

Aquí observamos que otras tantas funciones biológicas están asociadas con la regulación de diversos procesos biológicos.

Enriquecimiento con KEGG En la Figure 14 se puede observar que las funciones biológicas están relacionadas con enfermedades/infecciones (malaria, hepatitis B, tuberculosis) por lo que se podríamos deducir que dichas proteínas forman parte de la respuesta inmunitaria del organismo ante dichas enfermedades o que las provocan. Además de esto, se observan funciones biológicas relacionadas con vías biológicas del organismo, como son la vía de señalización de quimioquinas, la de detección de ADN ctosólico...

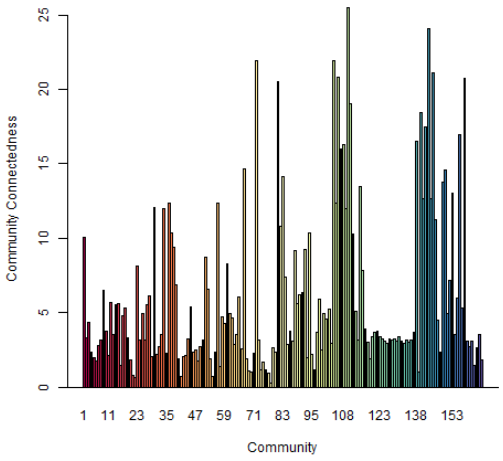


Figure 9 *Comunidades por modularidad*

3.3.2 Clúster 78

Enriquecimiento con GO En la Figure 15 , GO no ha encontrado funciones biológicas asociadas al clúster elegido.

Enriquecimiento con KEGG En este caso, en la Figure 16 , KEGG ha encontrado dos funciones biológicas asociadas al clúster: cáncer de tiroides y vías en el cáncer. Es por tanto que deducimos que las proteínas que forman parte de este clúster se ocupan de las vías principales del cáncer de tiroides, ya sea para detectarlo o provocarlo.

3.3.3 Clúster 22

Enriquecimiento con GO En este caso, en la Figure 17 , se puede observar que las principales funciones biológicas de las proteínas pertenecientes a este clúster son de regulación, transporte y recepción, por lo que desempeñan funciones muy importantes en las rutas metabólicas del organismo.

""	"term"	"number$_o f_g enes$"	"number$_o f_g enes_i n_b ackground$"	"ncbiTaxonId"	"inputGenes"	"descri

Enriquecimiento con KEGG En la Figure 18 , las funciones biológicas obtenidas por KEGG son más concretas: reabsorción de calcio regulada por factores endocrinos, ciclo de vesículas sinápticas, endocitosis, lisosoma, enfermedad de Huntington, invasión bacteriana de las células epiteliales y vía de señalización de la fosfolipasa D. Esta última está relacionada con la traducción de señales. Otras están relacionadas con enfermedades y sus causas.

4 Discusión

5 Conclusiones

Abreviaciones

Disponibilidad de datos y materiales

https://github.com/Paulandujar/project_template

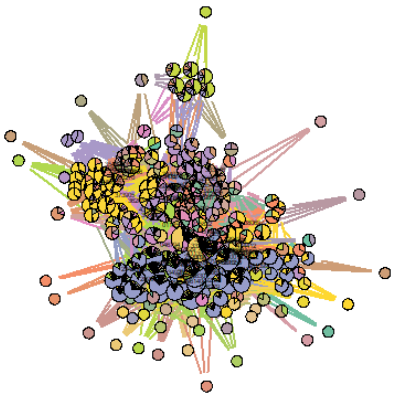


Figure 10 *Fruchterman Reingold graph*

Contribución de los autores

¡¡¡¡¡¡¡ Updated upstream I.R.G: Encargada del análisis de la red (distancia entre nodos, distribución de grado y coeficiente de agrupamiento), cálculo de la robustez, escritura del launch.sh y escritura de los resultados de estos cálculos en el report; P.A.Z: Encargada del cálculo linked communities, posterior análisis de esos resultados en el report y escritora del abstract; R.G.M: Encargada de las funciones de Robustez y escritura del apartado Materiales y Métodos; S.dC.C: Encargada del enriquecimiento funcional, análisis de esos resultados en el report y escritura del setup.sh =======
I.R.G: Encargada del análisis de la red (distancia entre nodos, distribución de grado y coeficiente de agrupamiento), cálculo de la robustez, escritura del launch.sh y escritura de los resultados de estos cálculos en el report; P.A.Z: Encargada del cálculo linked communities, posterior análisis de esos resultados en el report y escritora del abstract; R.G.M: Encargada de las funciones de Robustez y escritura del apartado Materiales y Métodos; S.dC.C: Encargada del enriquecimiento funcional, análisis de esos resultados en el report y escritura del setup.sh
¿¿¿¿¿¿¿ Stashed changes

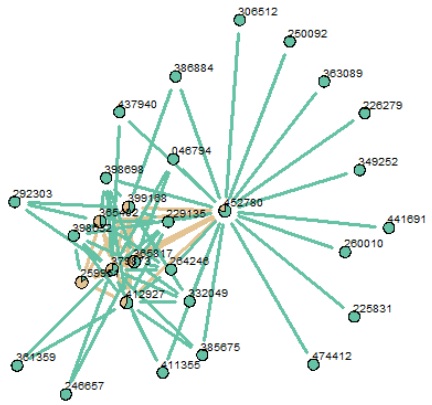


Figure 11 *Comunidades anidadas*

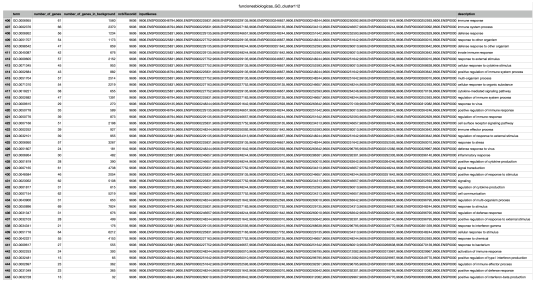


Figure 12 *Funciones biológicas clúster 112 con GO*

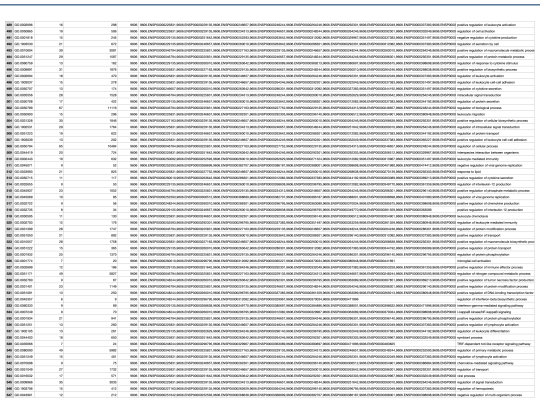


Figure 13 *Funciones biológicas clúster 112 con GO*

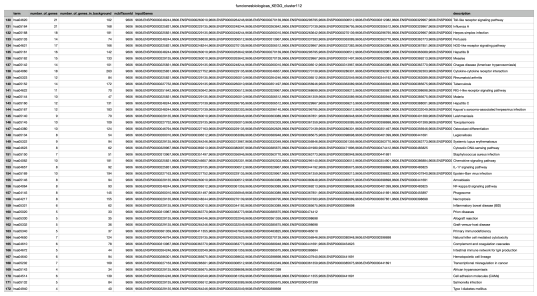


Figure 14 *Funciones biológicas clúster 112 con KEGG*

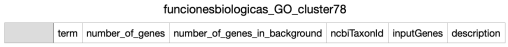


Figure 15 *Funciones biológicas clúster 78 con GO*

Figure 16 *Funciones biológicas clúster 78 con KEGG*

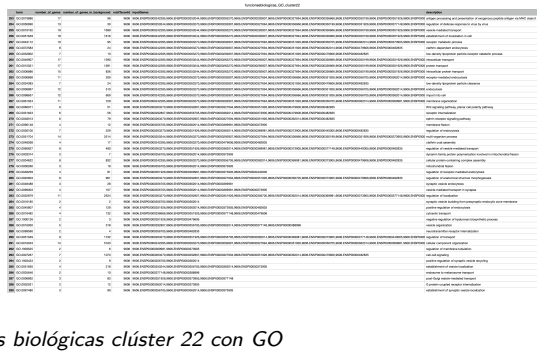


Figure 17 *Funciones biológicas clúster 22 con GO*

Figure 18 *Funciones biológicas clúster 22 con KEGG*