Functional Profiling

Raúl Pérez Moraga

Biomedical Imaging Unit FISABIO-CIPF

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

WODA

Web-based Omics Data Analysis

Unidad de
Bioinformática y
Bioestadística

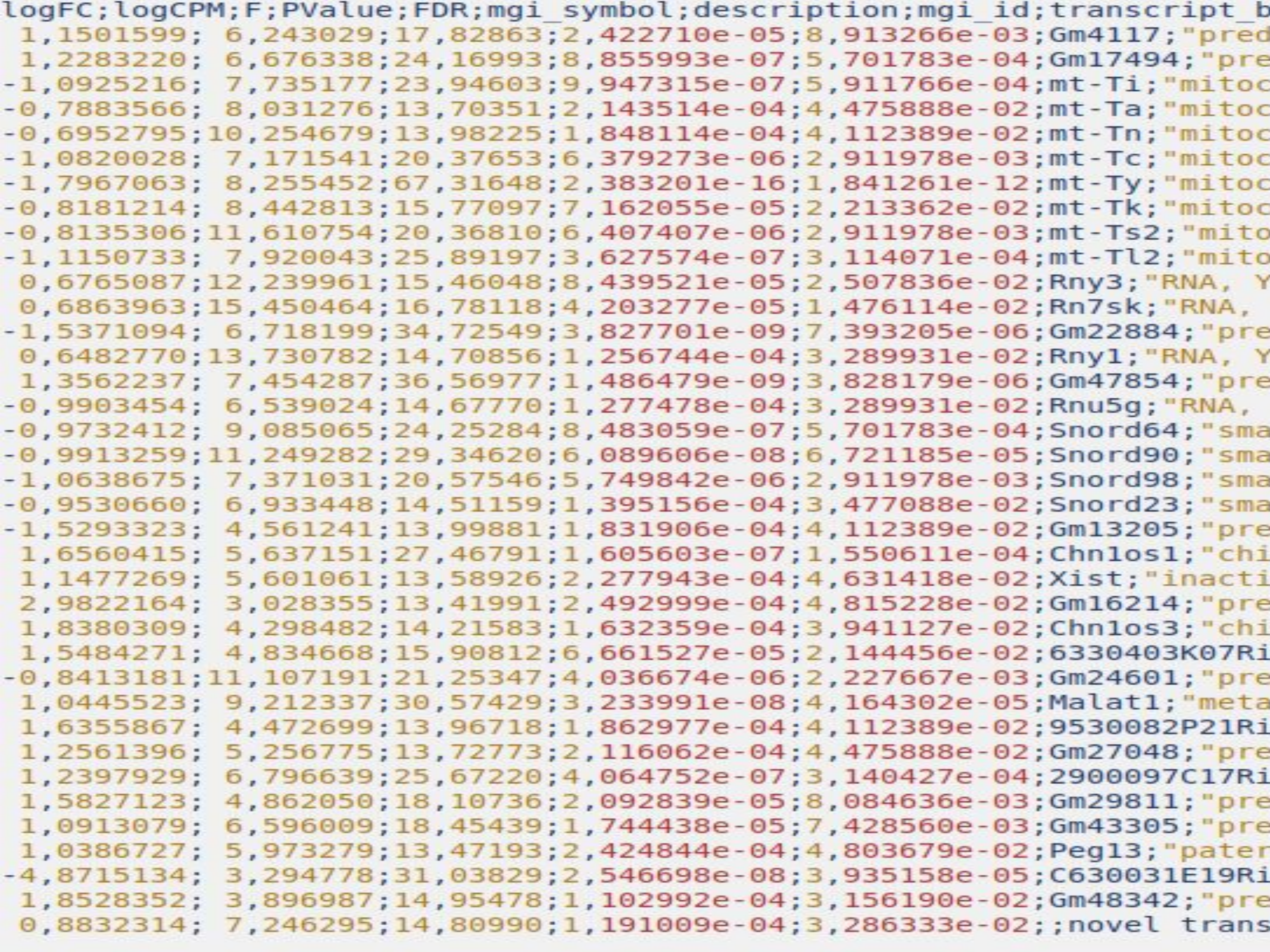
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Fundació per al Foment de la
Investigació Sanitària i Biomèdica
de la Comunitat Valenciana

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Outline

- Introduction
- Over-Representation Analysis (ORA)
- Gene Set Analysis (GSA)
- Network Analysis (NA)



```
logFC;logCPM;F;PValue;FDR;mgi_symbol;description;mgi_id;transcript_b
 1,1501599; 6,243029;17,82863;2,422710e-05;8,913266e-03;Gm4117;"pred
 1,2283220; 6,676338;24,16993;8,855993e-07;5,701783e-04;Gm17494;"pre
-1,0925216; 7,735177;23,94603;9,947315e-07;5,911766e-04;mt-Ti;"mitoc
-0,7883566; 8,031276;13,70351;2,143514e-04;4,475888e-02;mt-Ta;"mitoc
-0,6952795;10,254679;13,98225;1,848114e-04;4,112389e-02;mt-Tn;"mitoc
-1,0820028; 7,171541;20,37653;6,379273e-06;2,911978e-03;mt-Tc;"mitoc
-1,7967063; 8,255452;67,31648;2,383201e-16;1,841261e-12;mt-Ty;"mitoc
-0,8181214; 8,442813;15,77097;7,162055e-05;2,213362e-02;mt-Tk;"mitoc
-0,8135306;11,610754;20,36810;6,407407e-06;2,911978e-03;mt-Ts2;"mito
-1,1150733; 7,920043;25,89197;3,627574e-07;3,114071e-04;mt-Tl2;"mito
 0,6765087;12,239961;15,46048;8,439521e-05;2,507836e-02;Rny3;"RNA, Y
 0,6863963;15,450464;16,78118;4,203277e-05;1,476114e-02;Rn7sk;"RNA, 
-1,5371094; 6,718199;34,72549;3,827701e-09;7,393205e-06;Gm22884;"pre
 0,6482770;13,730782;14,70856;1,256744e-04;3,289931e-02;Rny1;"RNA, Y
 1,3562237; 7,454287;36,56977;1,486479e-09;3,828179e-06;Gm47854;"pre
-0,9903454; 6,539024;14,67770;1,277478e-04;3,289931e-02;Rnu5g;"RNA, 
-0,9732412; 9,085065;24,25284;8,483059e-07;5,701783e-04;Snord64;"sma
-0,9913259;11,249282;29,34620;6,089606e-08;6,721185e-05;Snord90;"sma
-1,0638675; 7,371031;20,57546;5,749842e-06;2,911978e-03;Snord98;"sma
-0,9530660; 6,933448;14,51159;1,395156e-04;3,477088e-02;Snord23;"sma
-1,5293323; 4,561241;13,99881;1,831906e-04;4,112389e-02;Gm13205;"pre
 1,6560415; 5,637151;27,46791;1,605603e-07;1,550611e-04;Chn1os1;"chi
 1,1477269; 5,601061;13,58926;2,277943e-04;4,631418e-02;Xist;"inacti
 2,9822164; 3,028355;13,41991;2,492999e-04;4,815228e-02;Gm16214;"pre
 1,8380309; 4,298482;14,21583;1,632359e-04;3,941127e-02;Chn1os3;"chi
 1,5484271; 4,834668;15,90812;6,661527e-05;2,144456e-02;6330403K07Ri
-0,8413181;11,107191;21,25347;4,036674e-06;2,227667e-03;Gm24601;"pre
 1,0445523; 9,212337;30,57429;3,233991e-08;4,164302e-05;Malat1;"meta
 1,6355867; 4,472699;13,96718;1,862977e-04;4,112389e-02;9530082P21Ri
 1,2561396; 5,256775;13,72773;2,116062e-04;4,475888e-02;Gm27048;"pre
 1,2397929; 6,796639;25,67220;4,064752e-07;3,140427e-04;2900097C17Ri
 1,5827123; 4,862050;18,10736;2,092839e-05;8,084636e-03;Gm29811;"pre
 1,0913079; 6,596009;18,45439;1,744438e-05;7,428560e-03;Gm43305;"pre
 1,0386727; 5,973279;13,47193;2,424844e-04;4,803679e-02;Peg13;"pater
-4,8715134; 3,294778;31,03829;2,546698e-08;3,935158e-05;C630031E19Ri
 1,8528352; 3,896987;14,95478;1,102992e-04;3,156190e-02;Gm48342;"pre
 0,8832314; 7,246295;14,80990;1,191009e-04;3,286333e-02;;novel trans
```

The problem

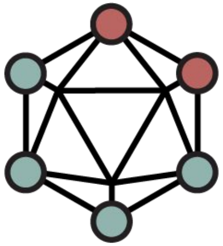
Bioinformatics raw results

Messy data

Difficult to interpret

Unbeatable by traditional methods (AKA: Excel)

Solution



STRING DB

Network analysis and functional enrichment

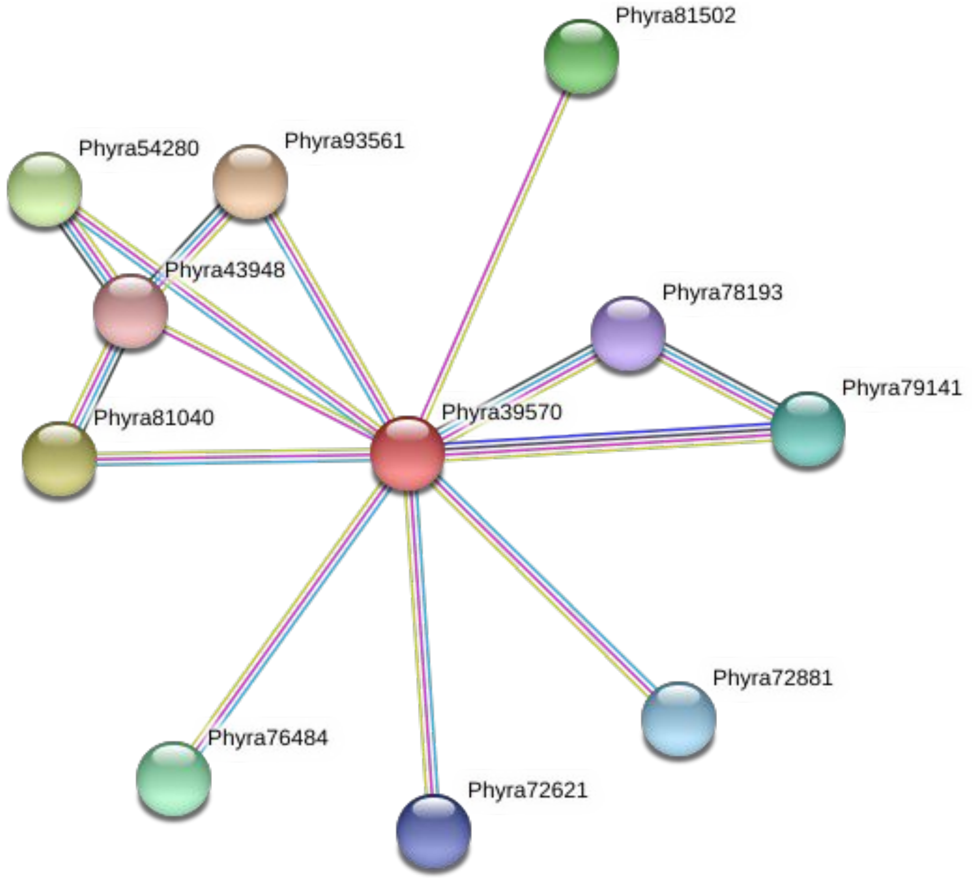
The problem

Protein-Protein Interactions (PPI)

Tidy data

It is easier to draw conclusions

Only need a web browser!



Protein-Protein Interactions (PPI)

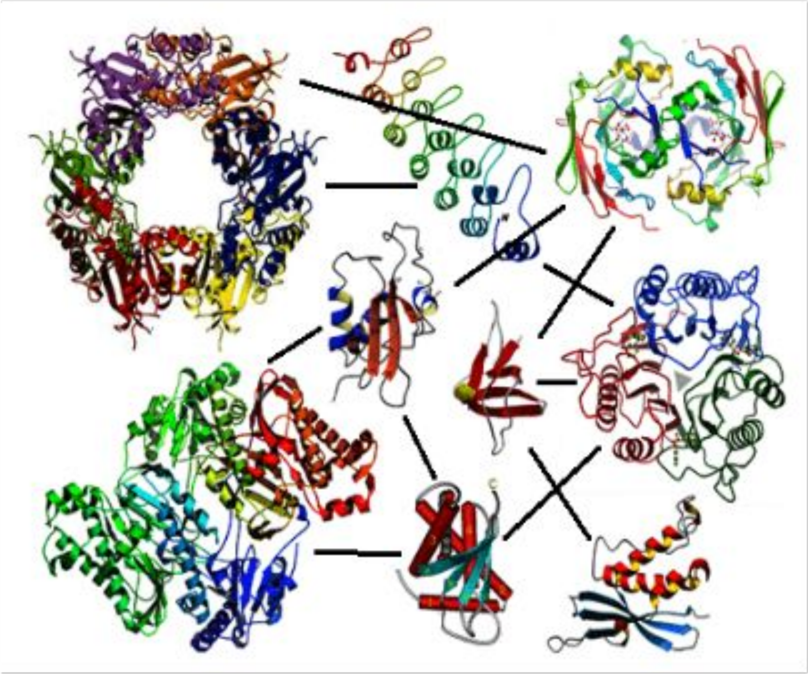
- How to extract information about sets of genes?

- How to perform functional enrichment analysis using protein-protein interactions as annotation source?

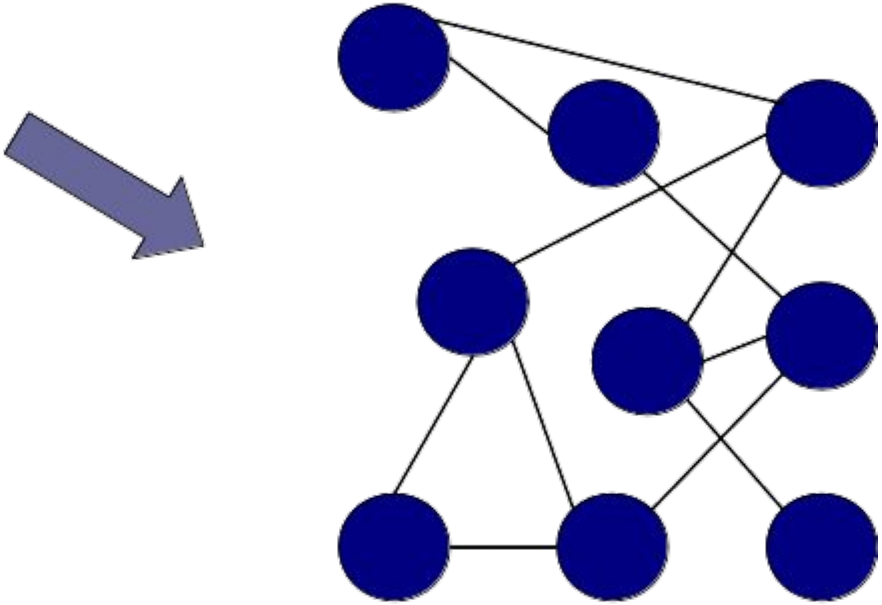
- How to prioritize candidate genes?

Graph Theory

Set of proteins interacting



Undirected graph



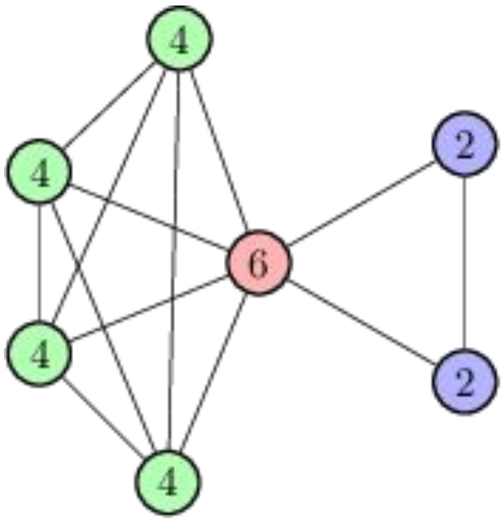
structured data

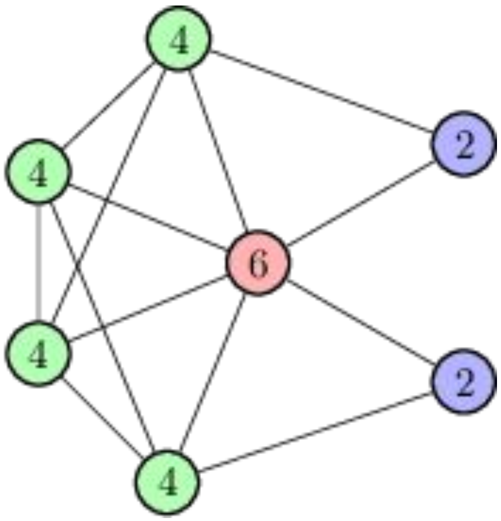
Nodes = proteins
Edges = interaction events

Graph Theory

Some Graph Theory concepts:

- **Degree (connectivity or connections):** Number of edges connected to a node. Nodes with degree are called hubs.





Graph Theory

Some Graph Theory concepts:

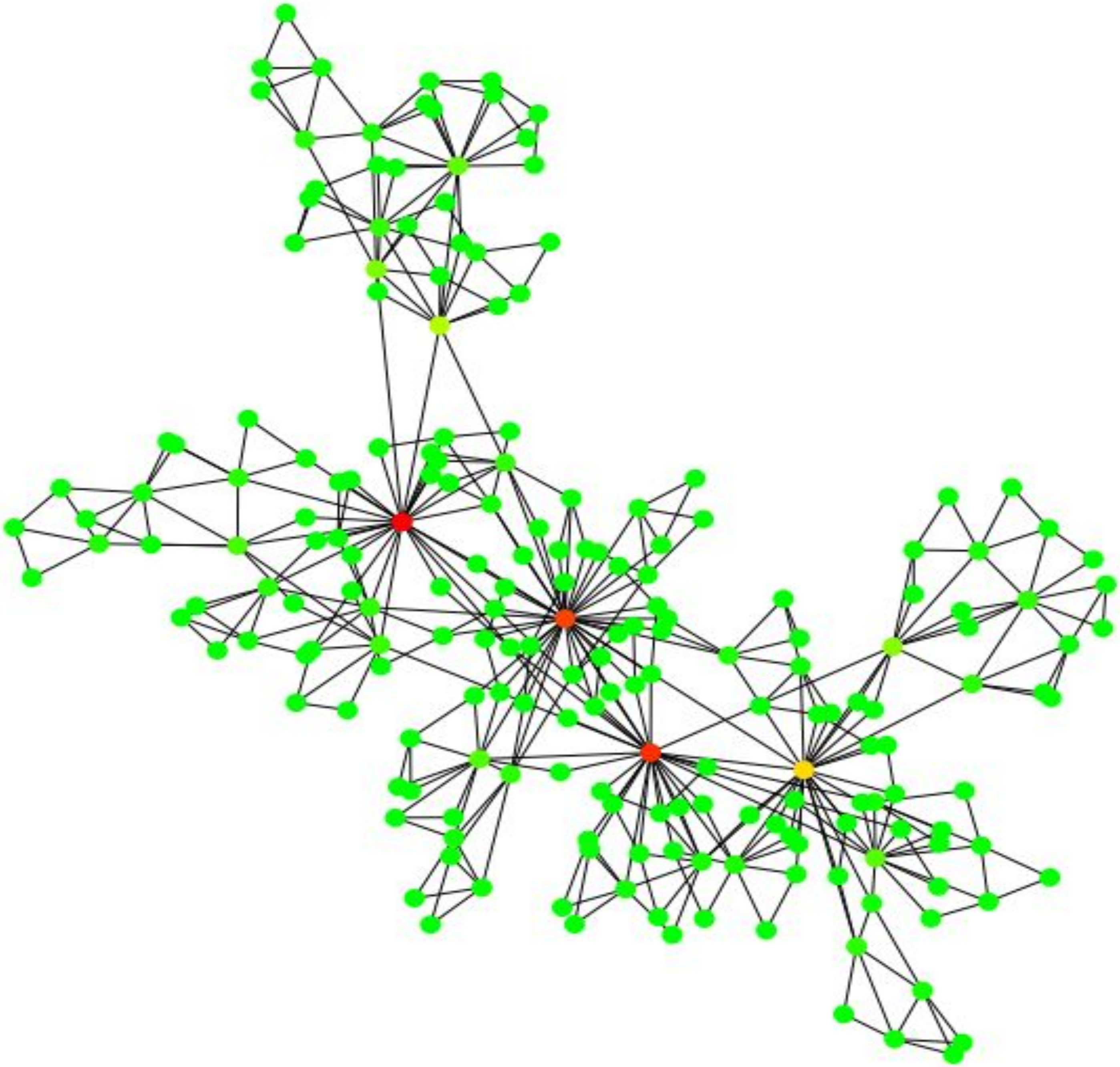
- **Degree (connectivity or connections):** Number of edges connected to a node. Nodes with degree are called hubs.
- **Shortest paths.**

Graph Theory

Some Graph Theory concepts:

- **Degree (connectivity or connections):** Number of edges connected to a node. Nodes with degree are called hubs.
- **Shortest paths.**
- **Betweenness:** A measure of centrality of a node by the number of shortest paths that pass through a node.
  - σ_{sr} is the **total** number of shortest paths in the graph.
  - σ_{sr} (V) is the number of shortest paths that pass through node V.

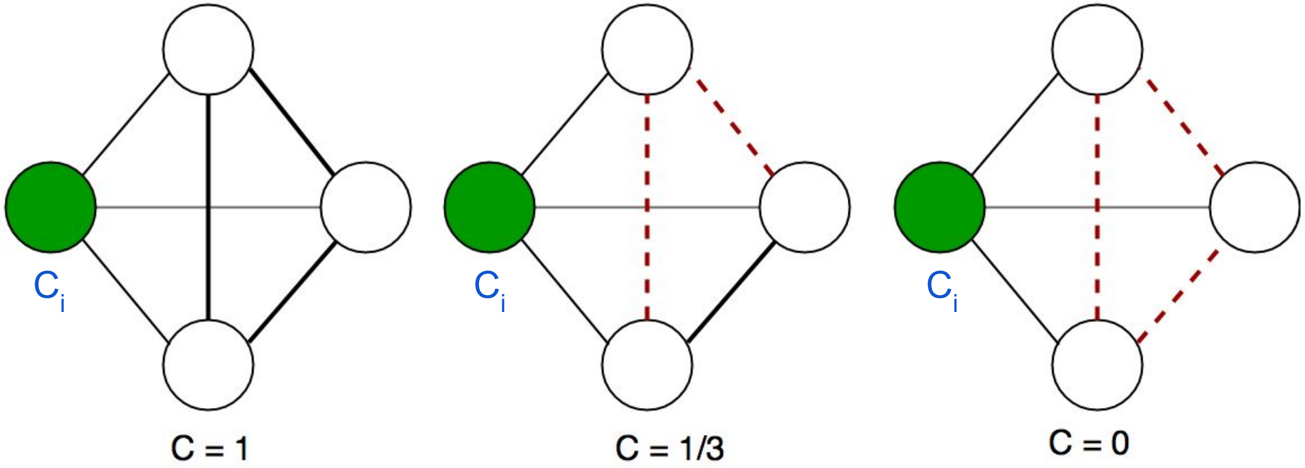
$$\sum_{s \neq v \neq r} \frac{\sigma_{sr}(v)}{\sigma_{sr}}$$



Graph Theory

- Clustering coefficient (a node): A measure of how interconnected the neighbours of that node are. Proportion of links between the nodes within its neighbourhood divided by the number of links that could possibly exist between them.
  - e_i is the number of edges among the nodes connected to node C_i.
  - n_i is the number of neighbours of node C_i.

$$C_i = \frac{2e_i}{n_i(n_i - 1)}$$



Graph Theory

- **Small world network: Typical organization of the biological networks.**
  - Few Hubs that connect all the network and a lot of nodes of low degree

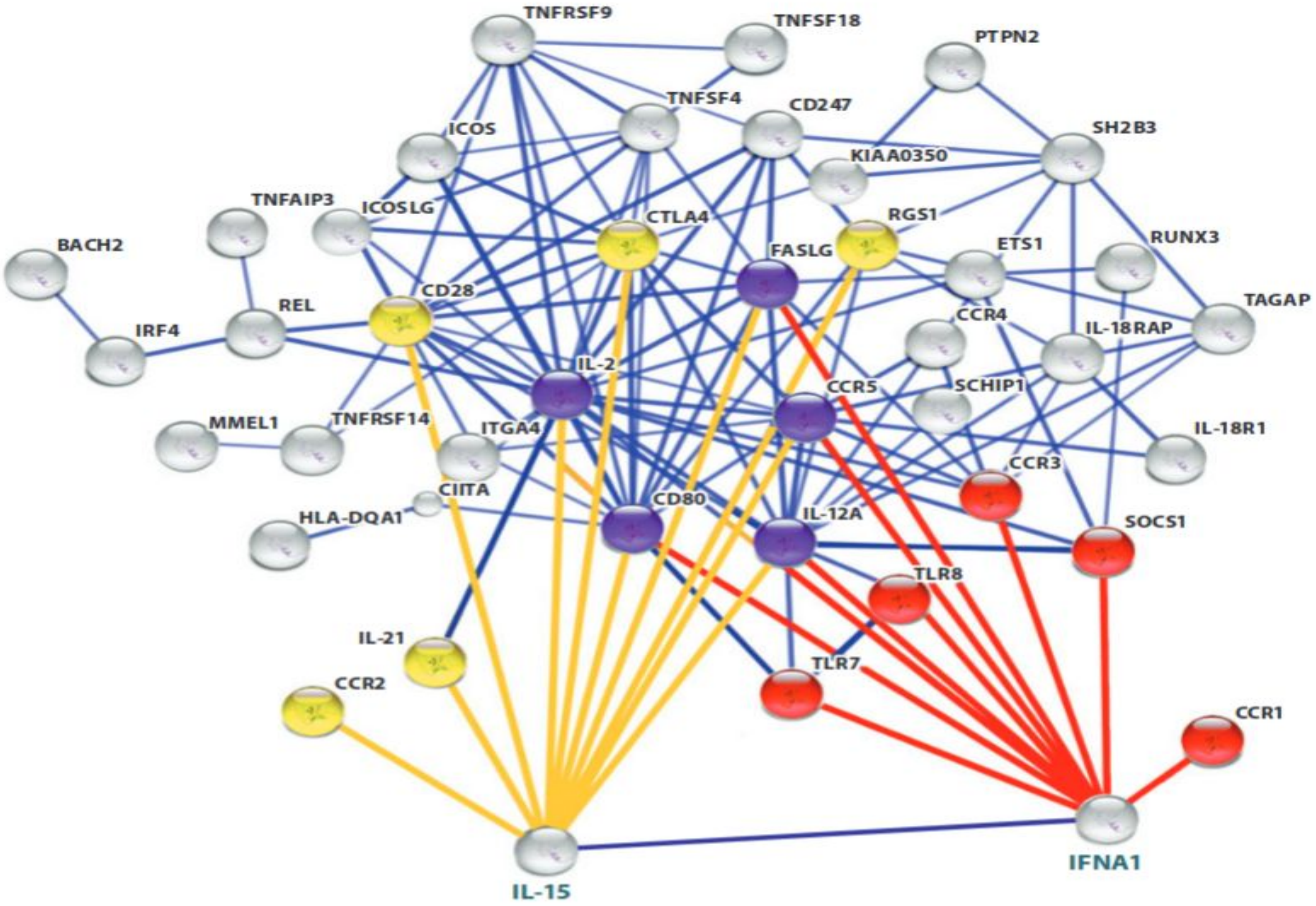


Figure 7

Network of known functional interactions between celiac disease (CD)–associated genes and key immunological markers of CD. We used the STRING database to look for known functional interactions among CD susceptibility genes, as well as functional interactions between CD susceptibility genes and interleukin (IL)-15 or interferon (IFN)-α. The STRING database assembles information about both known and predicted protein-protein interactions on the basis of numerous sources, including experimental repositories, computational prediction methods, and public text collections. Several CD susceptibility genes functionally interact with IL-15 (*yellow*), IFN-α (*red*), or both (*purple*).

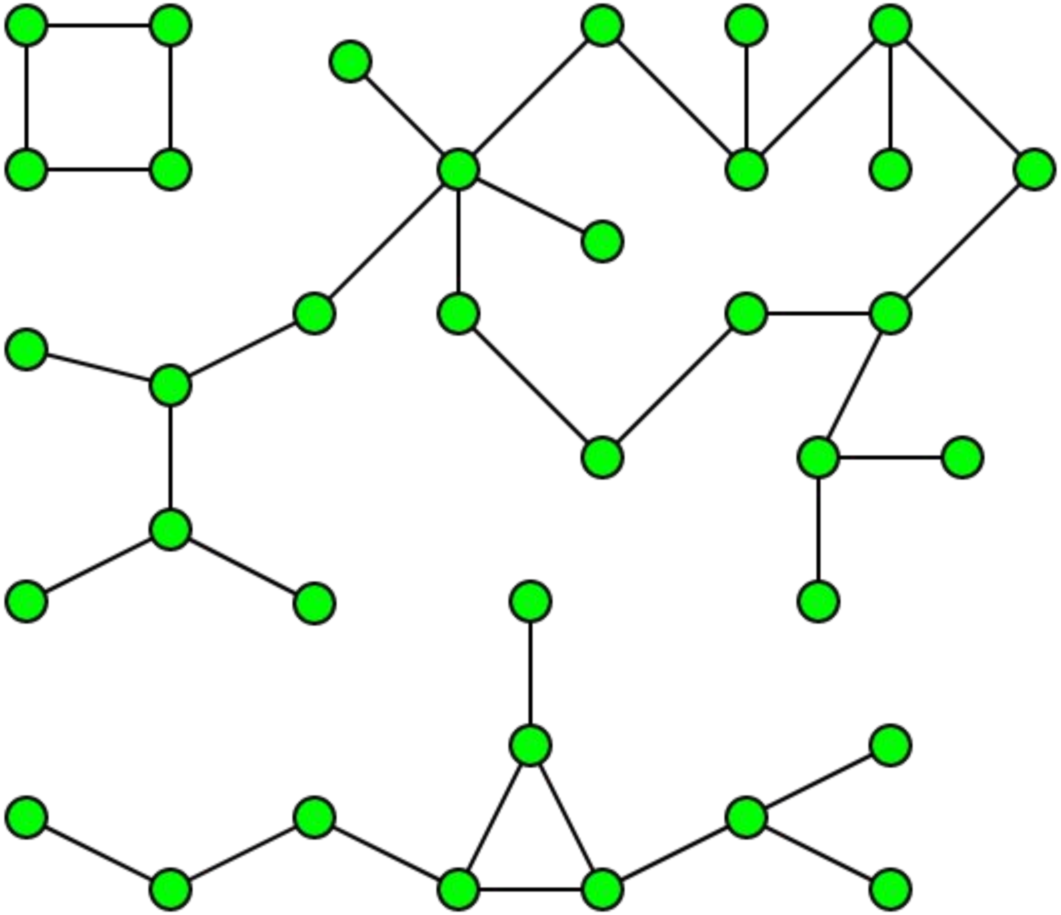
Graph Theory

Some Graph Theory concepts:

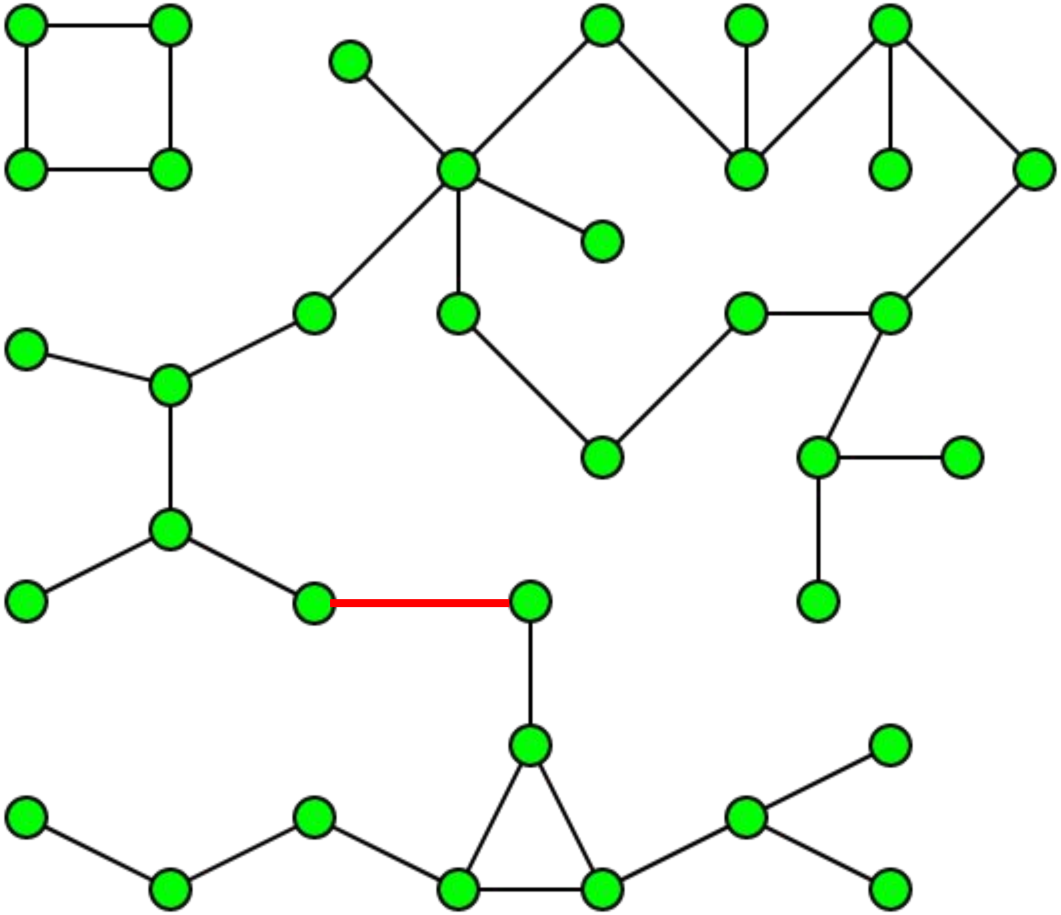
- **Small world network:** Biological like network organization.
  - Few Hubs that connect all the network and a lot of nodes of low degree

- **Component:** Node are connected to each other by paths and isolated from others clusters.

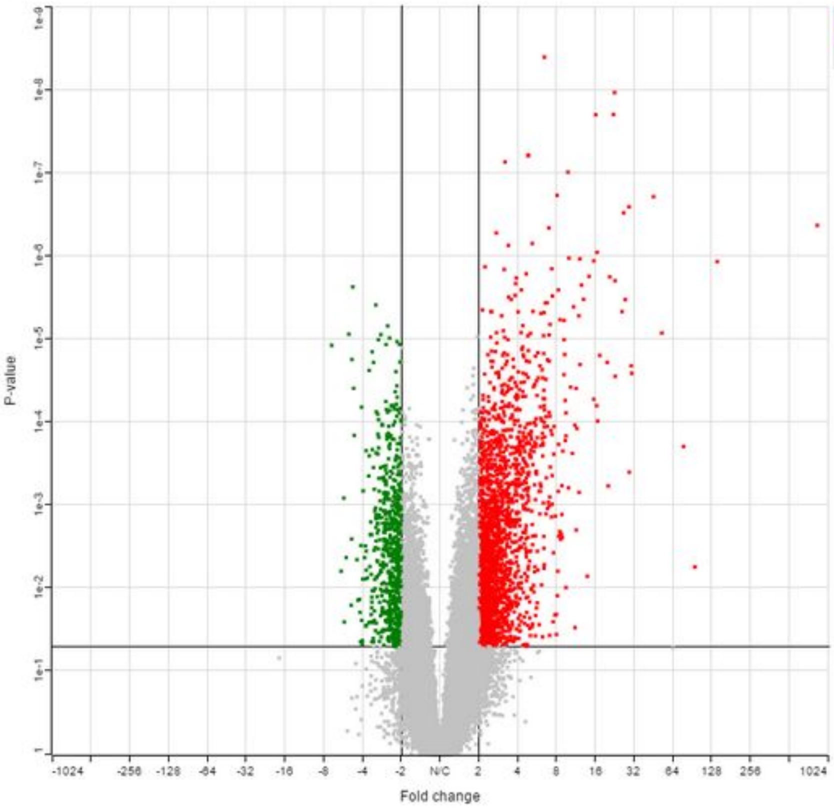
Graph Theory



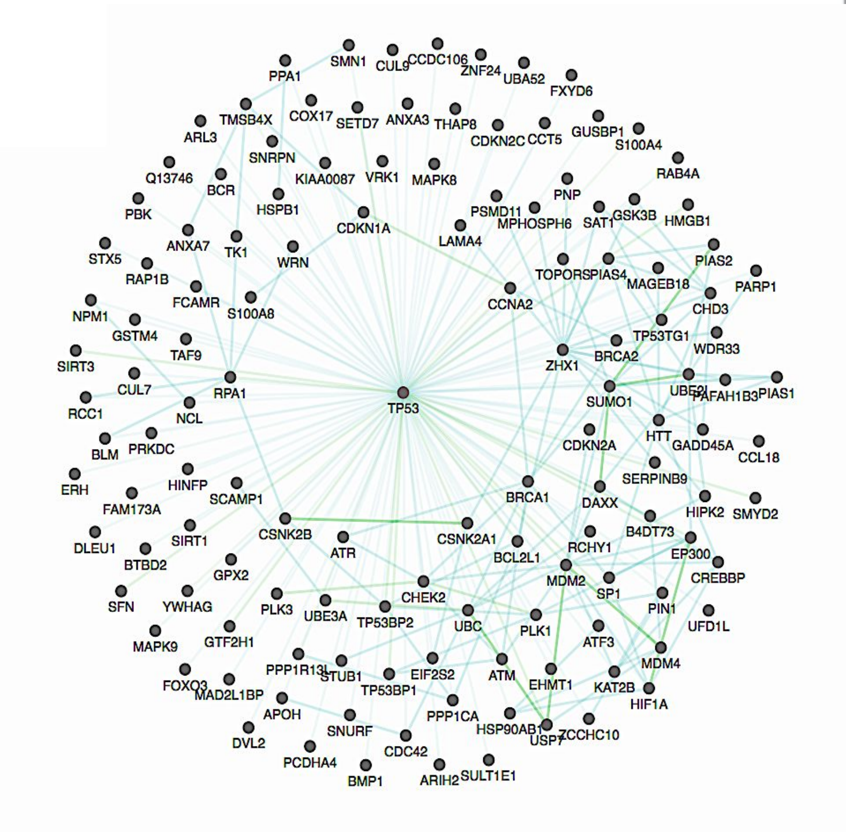
Graph Theory



Interactome & Transcriptome



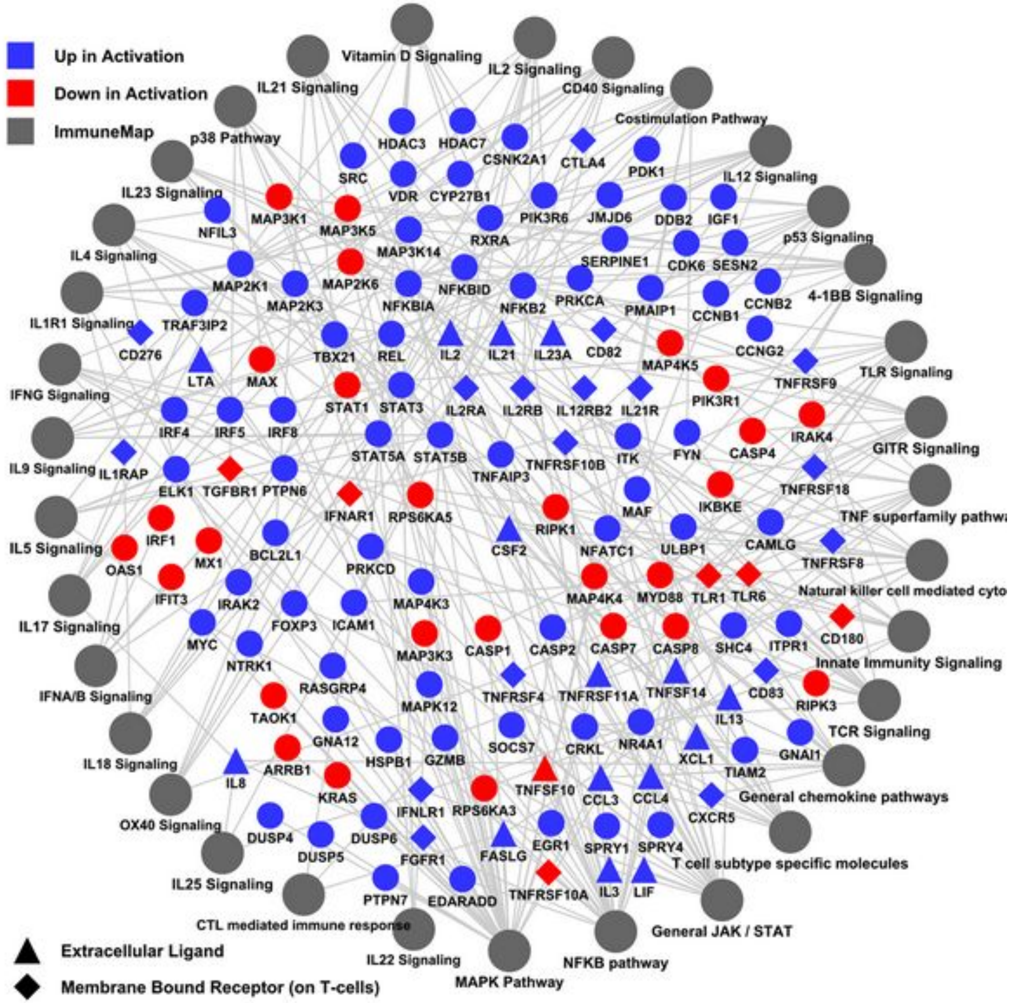
+



Transcriptome

Interactome

Interactome & Transcriptome



Any question?

Activities

1. Over-representation and GSEA exercises: http://bioinfo.cipf.es/WODA19/doku.php/bbdd

2. Protein-protein interaction exercises: http://bioinfo.cipf.es/WODA19/doku.php/ex_ppi