Biological networks: an approach to gene co-expression networks

II Escola Latino-Americana de Bioinformática

Dr. Edgardo Galán Vásquez

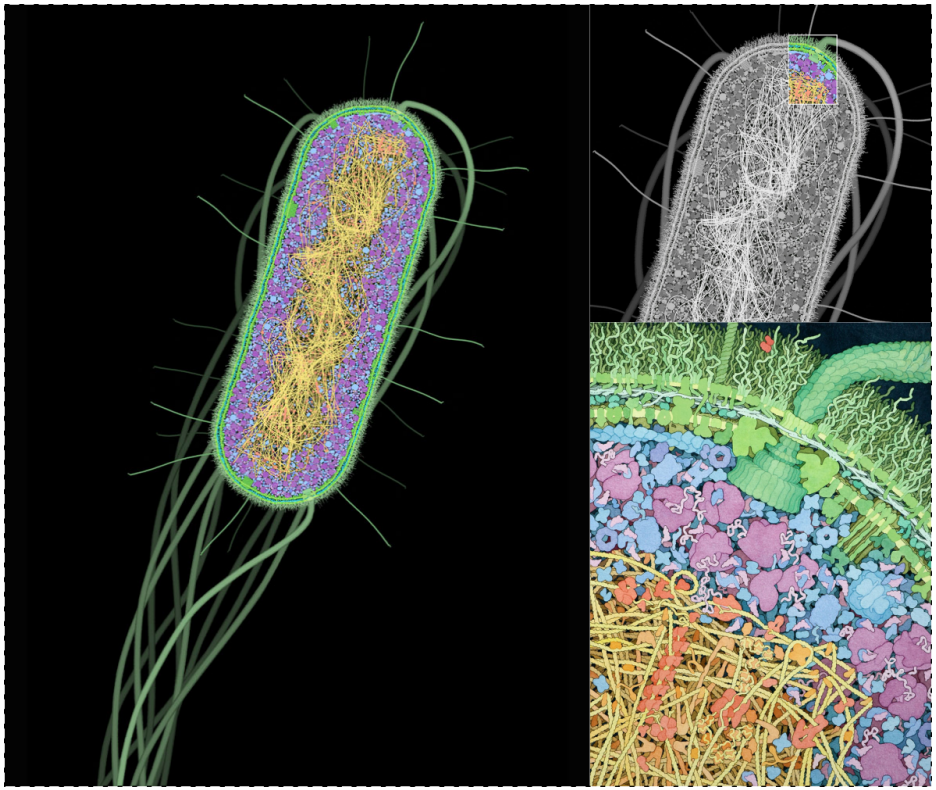
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

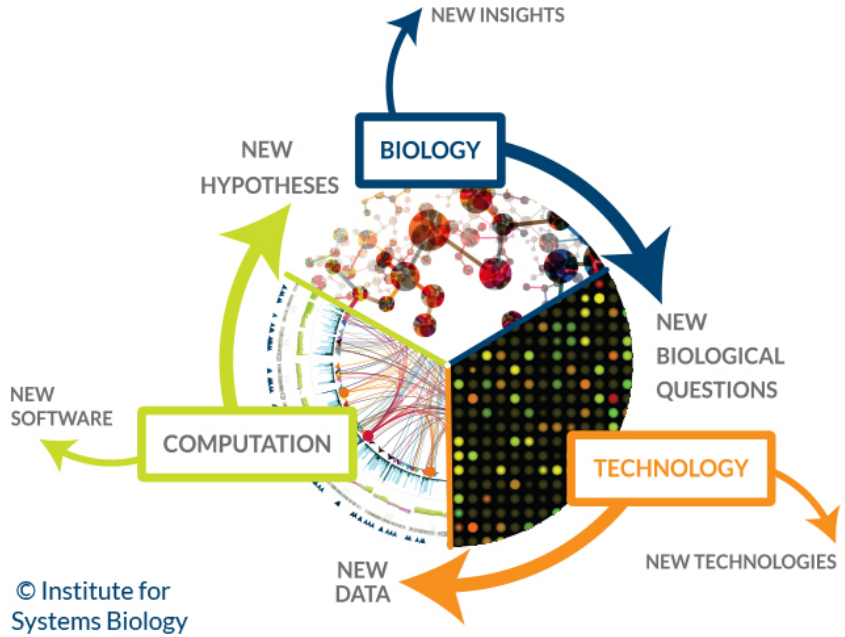
19 de Septiembre de 2024

Contents

Introduction

Organisms as complex systems

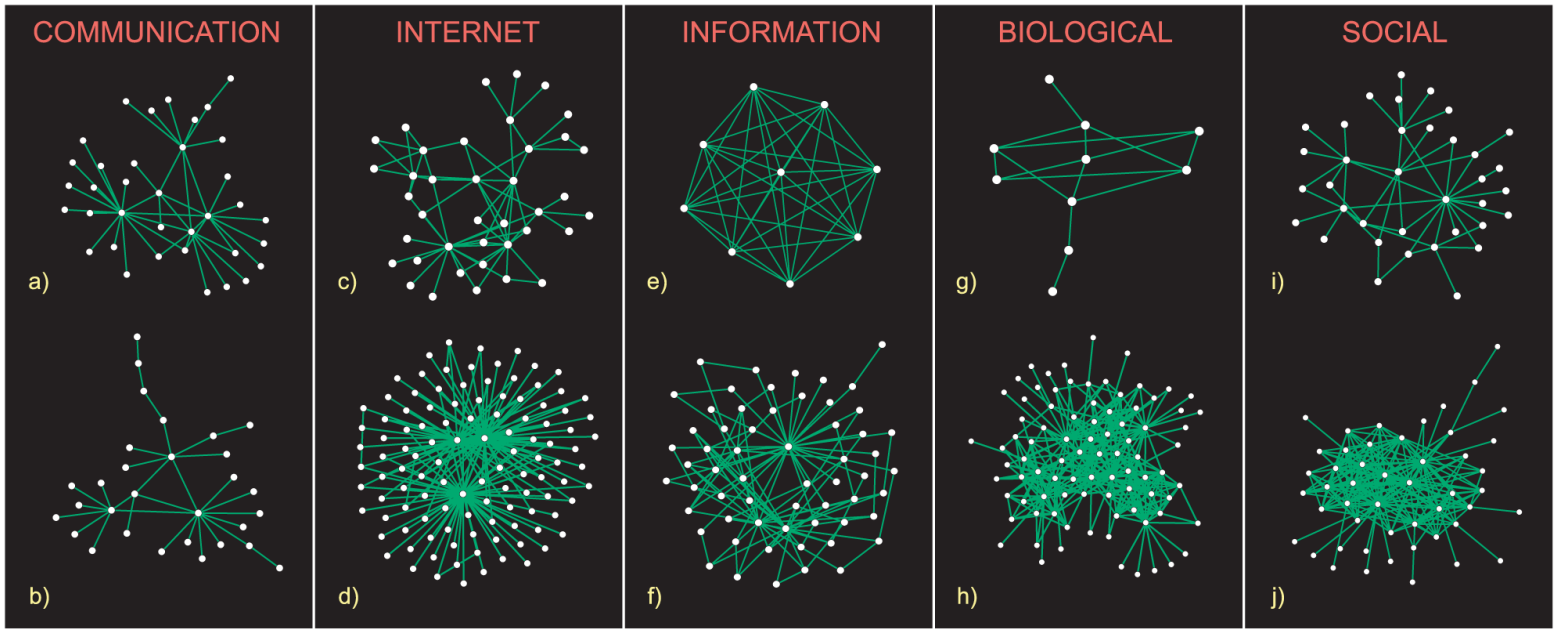




Goodsell David, 2010

Complex networks

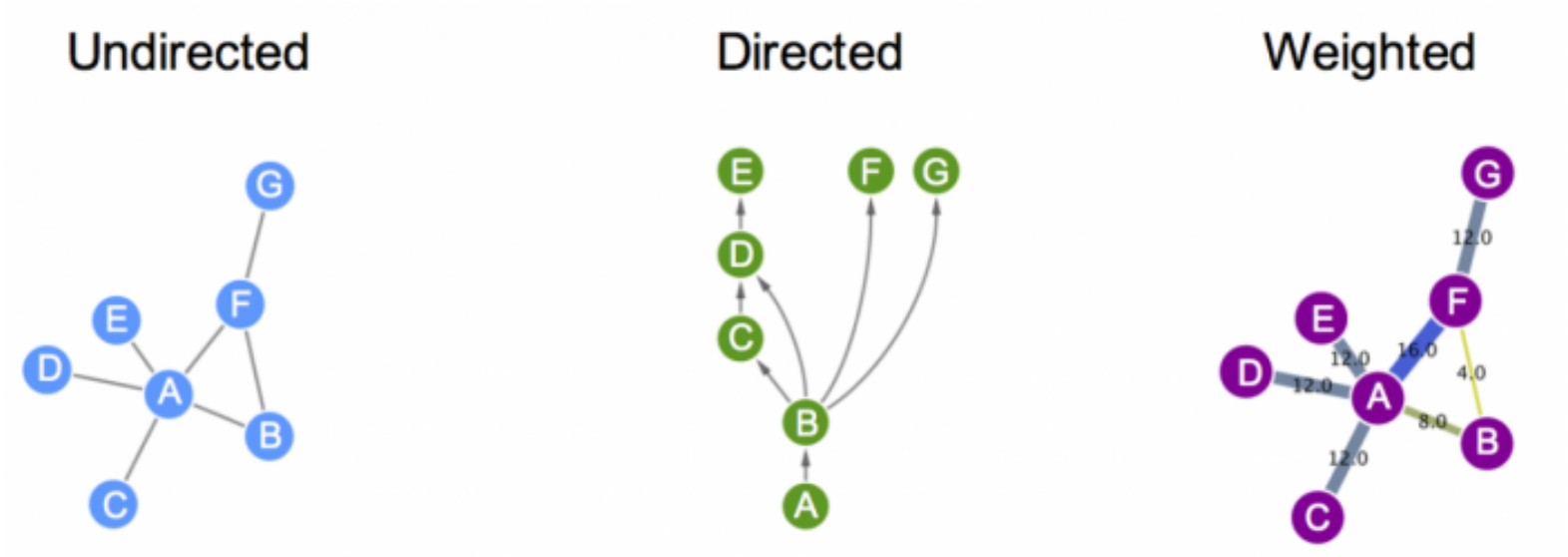
Complex networks can be employed to study a wide variety of natural phenomena, such as the World Wide Web, social networks, predator-prey relationships, biological processes, communication networks, academic citation networks, and many others.



Adamic, 1999; Lancichinetti et al. 2010

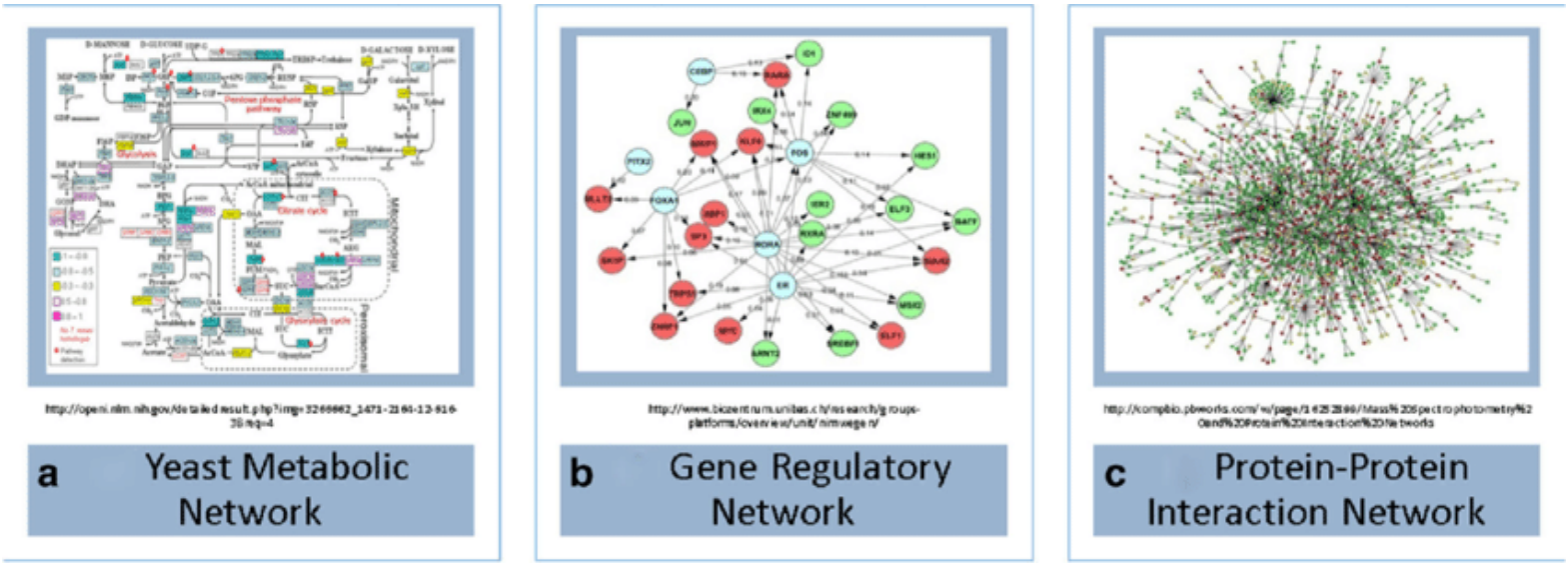
Types of networks

A graph is defined as $G = (V, E)$, where V is a set of nodes that represent the elements of the complexing system and E is a set of assigned edges that represent the relationship between each pair of elements.



Definition of Biological networks

Biological networks are representations of biological systems where nodes represent biological entities (e.g., genes, proteins, metabolites), and edges represent interactions or relationships (e.g., co-expression, regulation, metabolic flux).



a Yeast Metabolic Network

b Gene Regulatory Network

c Protein-Protein Interaction Network

Kim and Haukap 2017

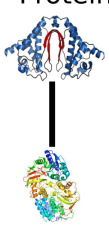
Types of Biological networks

Interacción de Proteína

Proteínas

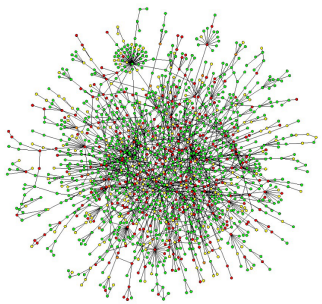
Interacción Física

Proteína-Proteína



A

B

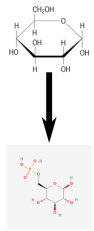


Metabolismo

Metabolitos

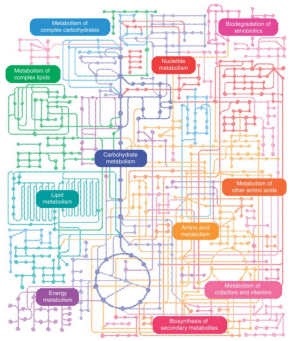
Conversión enzimática

Proteína-Metabolitos



A

B

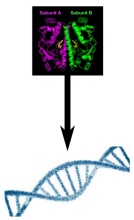


Transcripción

Factores de transcripción
Genes regulados

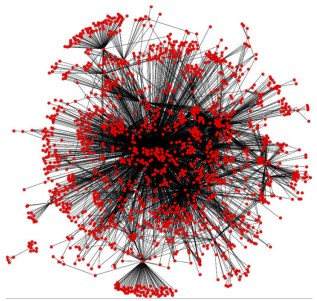
Interacción física

Proteína-DNA



A

B

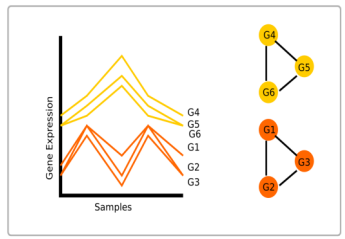


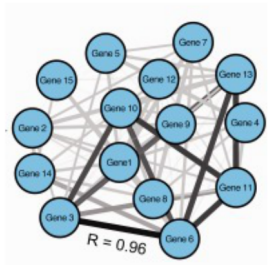
Co-expresión

Genes
Genes

Expresión similar

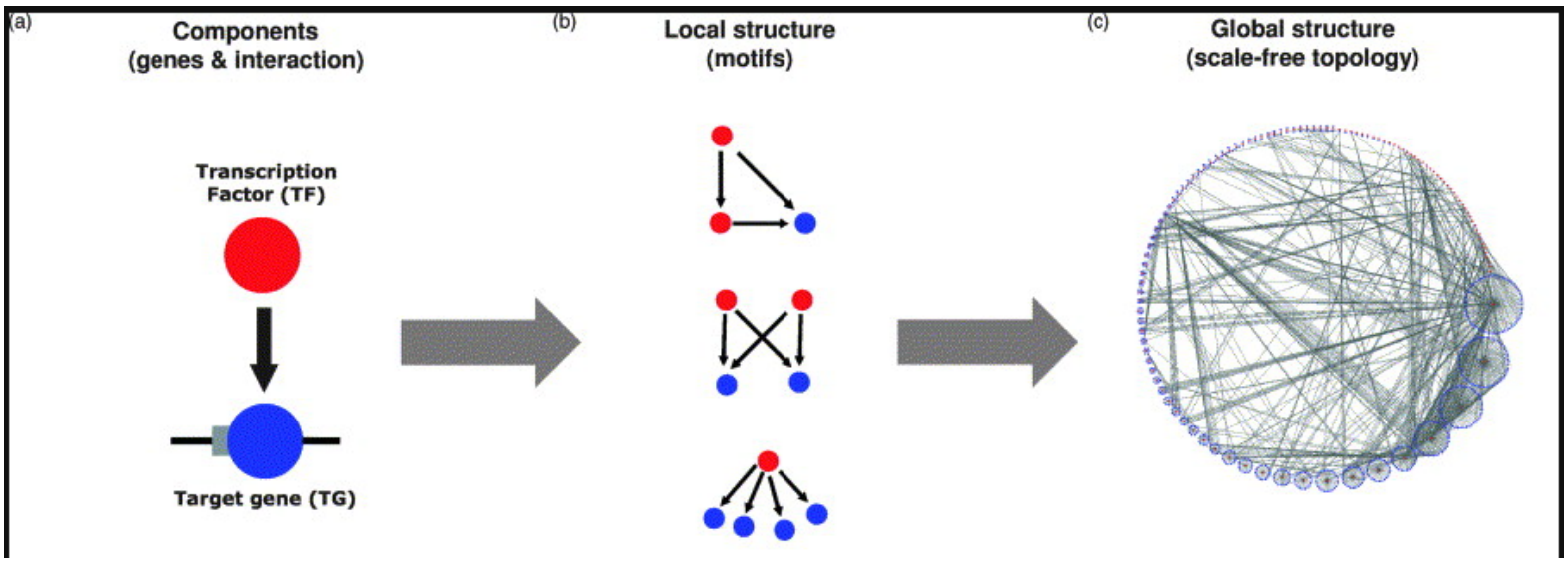
Gen-Gen





Structure of Biological nerworks

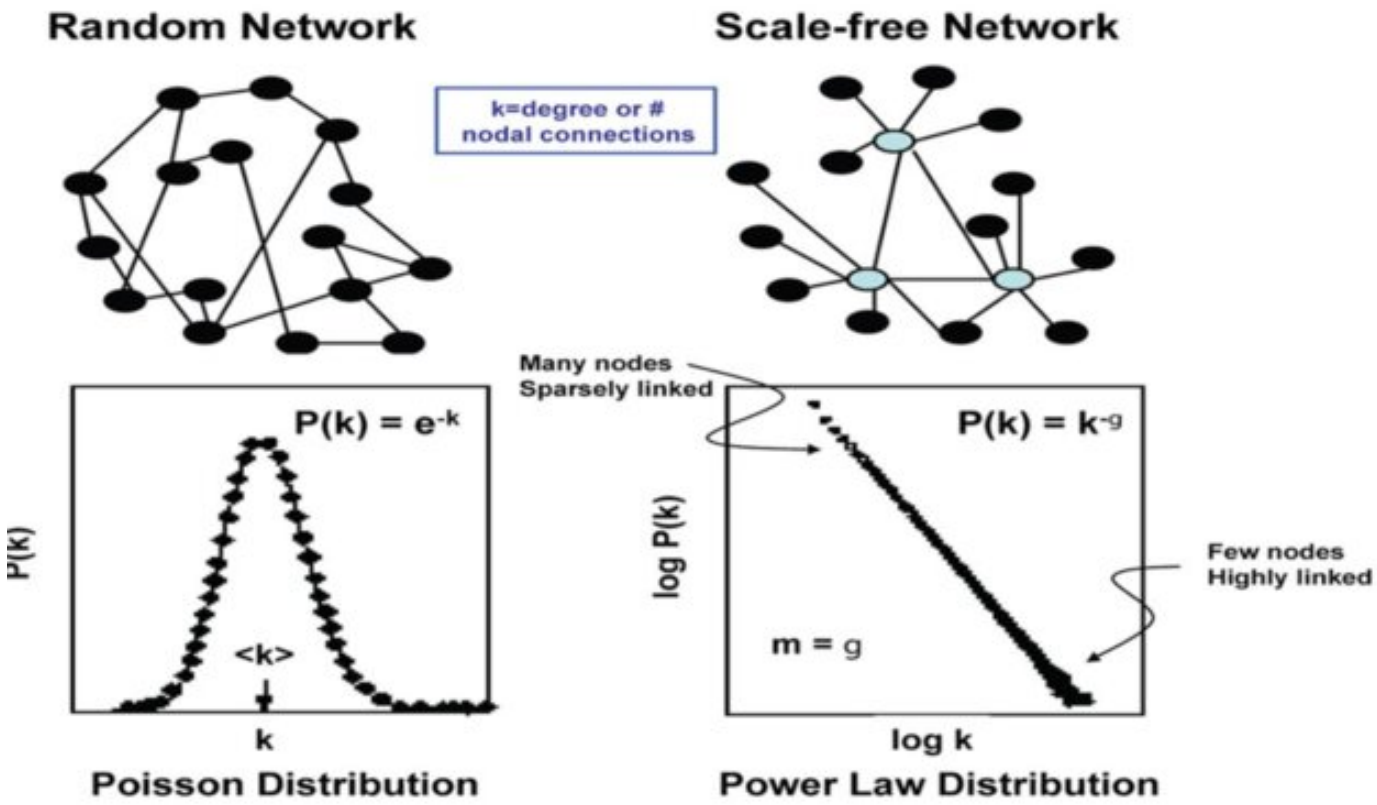
Las redes complejas pueden ser empleadas para estudiar una gran variedad de fenómenos naturales, tales como el World Wide Web, interacciones sociales, relaciones presa-predador, procesos biológicos, redes de comunicación, citas académicas, entre muchas otras.



Babu et al. 2006

Structure of Biological nerworks

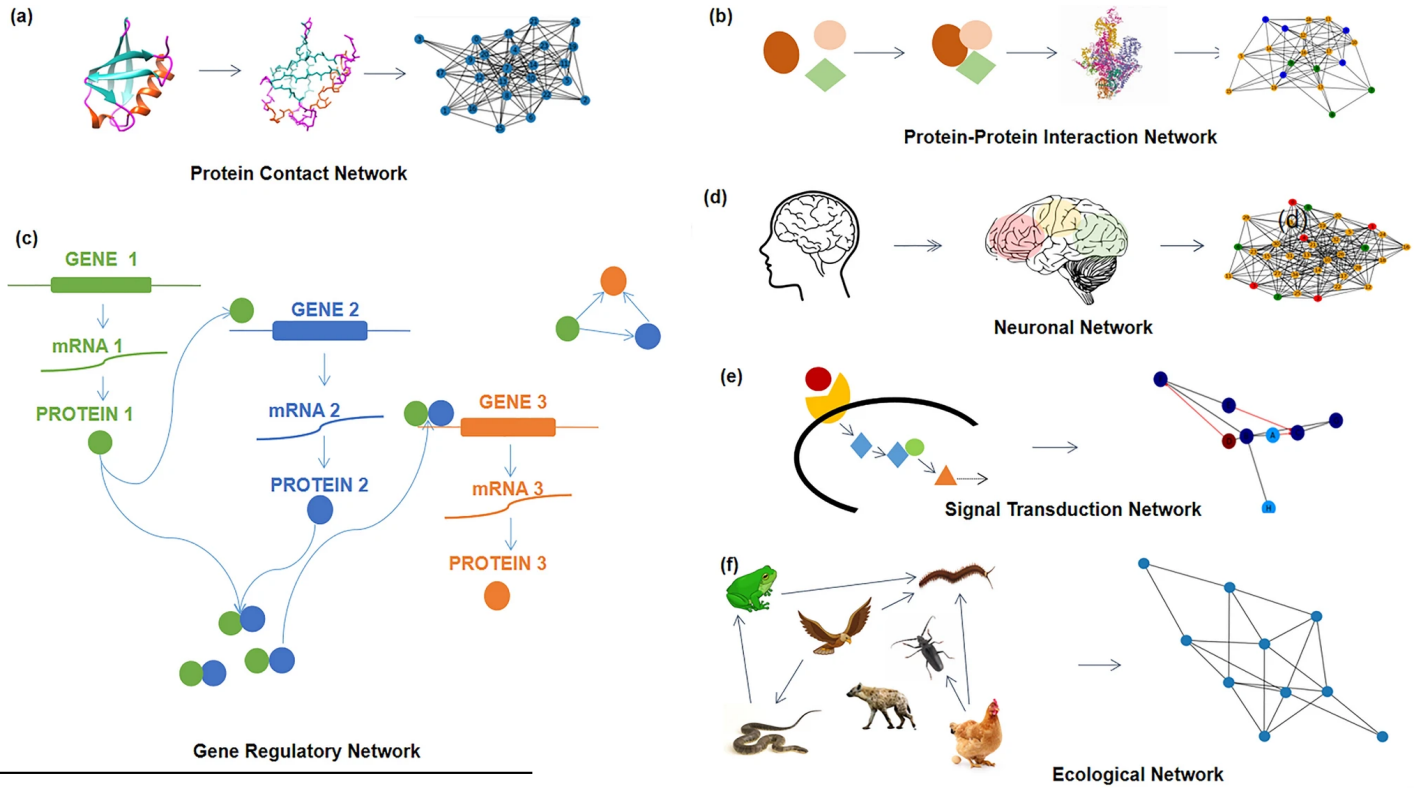
Scale-free networks have few nodes with a very large number of links (hubs) and meny nodes with only few links
Scale-free networks are characterized by a power-law distribution



Darrason 2014

Importance

Regulatory networks are crucial for understanding the complexity of biological systems because they represent the intricate relationships and interactions that control gene expression, protein activity, and cellular functions.

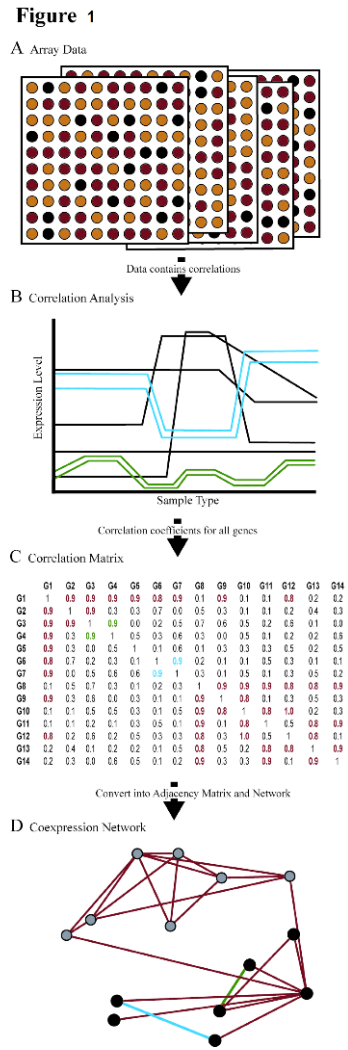


Chaudhuri and Srivastava 2022

Gene Co-expression Networks

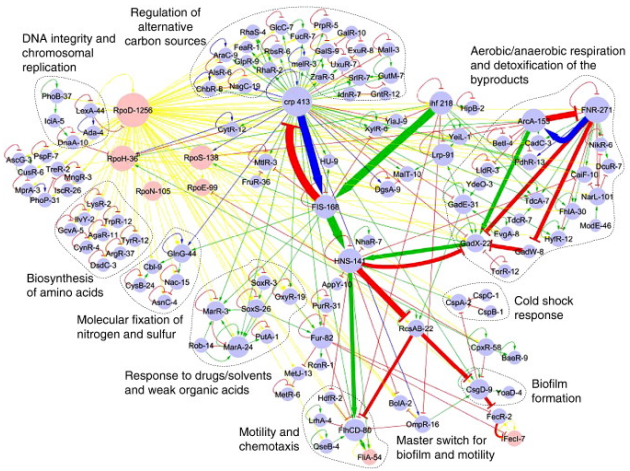
What is Gene Co-expression networks?

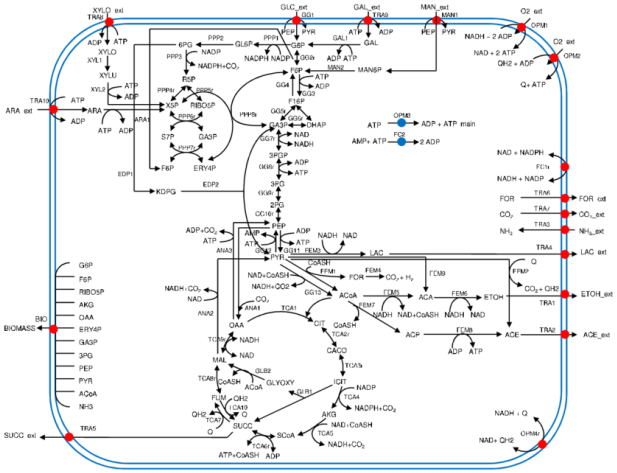
A Gene Co-expression Network is an undirected graph, where each node correpsonds to a gene, and the edges are a significant co-expression relationship between each pair of genes.

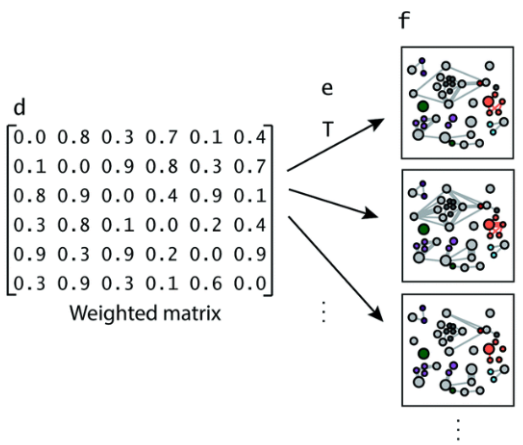


Stuart et al. 2003

GRN vs Metabolic Networks vs GCN





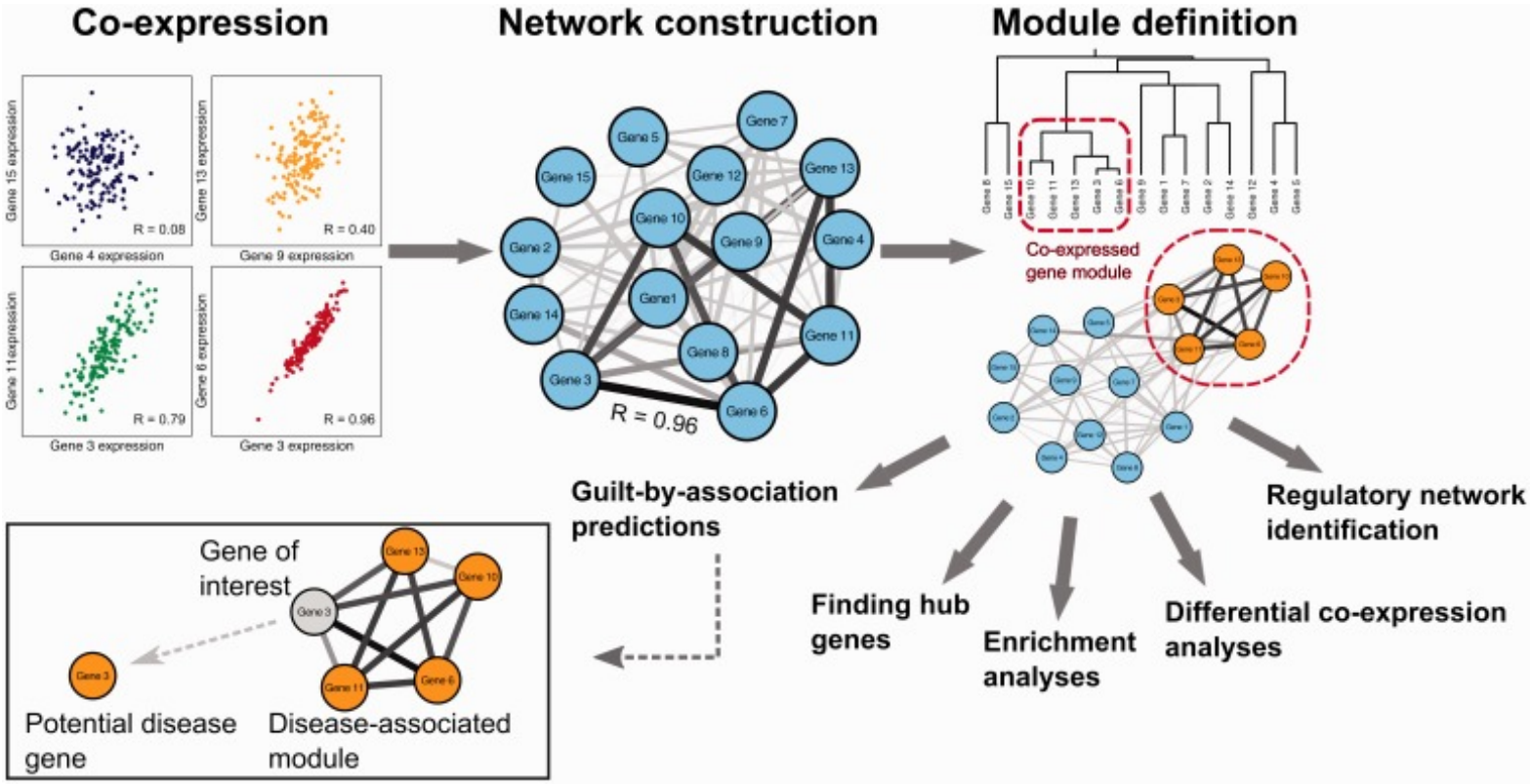


Villar et al. 2017

Applications

- Gene Function Prediction
- Identification of Gene Modules
- Discovery biomarkers
- Understanding Disease Mechanisms
- Drug Target Discovery
- Disease Classification
- Identifying Key Regulators
- Module conservation
- Time-series Analysis
- Studying Developmental Process
- among others.

Study of evolution



Ovens et al. 2021

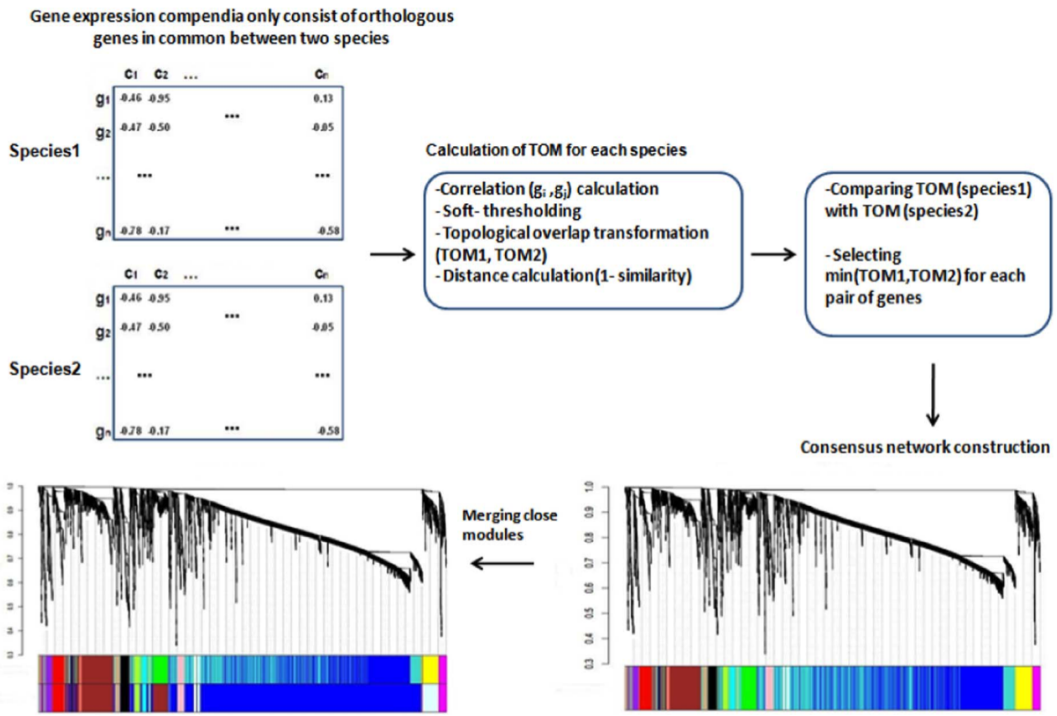
Module conservation

Research paper

Comparison of gene co-expression networks in *Pseudomonas aeruginosa* and *Staphylococcus aureus* reveals conservation in some aspects of virulence

Nazanin Hosseinkhan[a,c,*], Zaynab Mousavian[b,c], Ali Masoudi-Nejad[c,**]

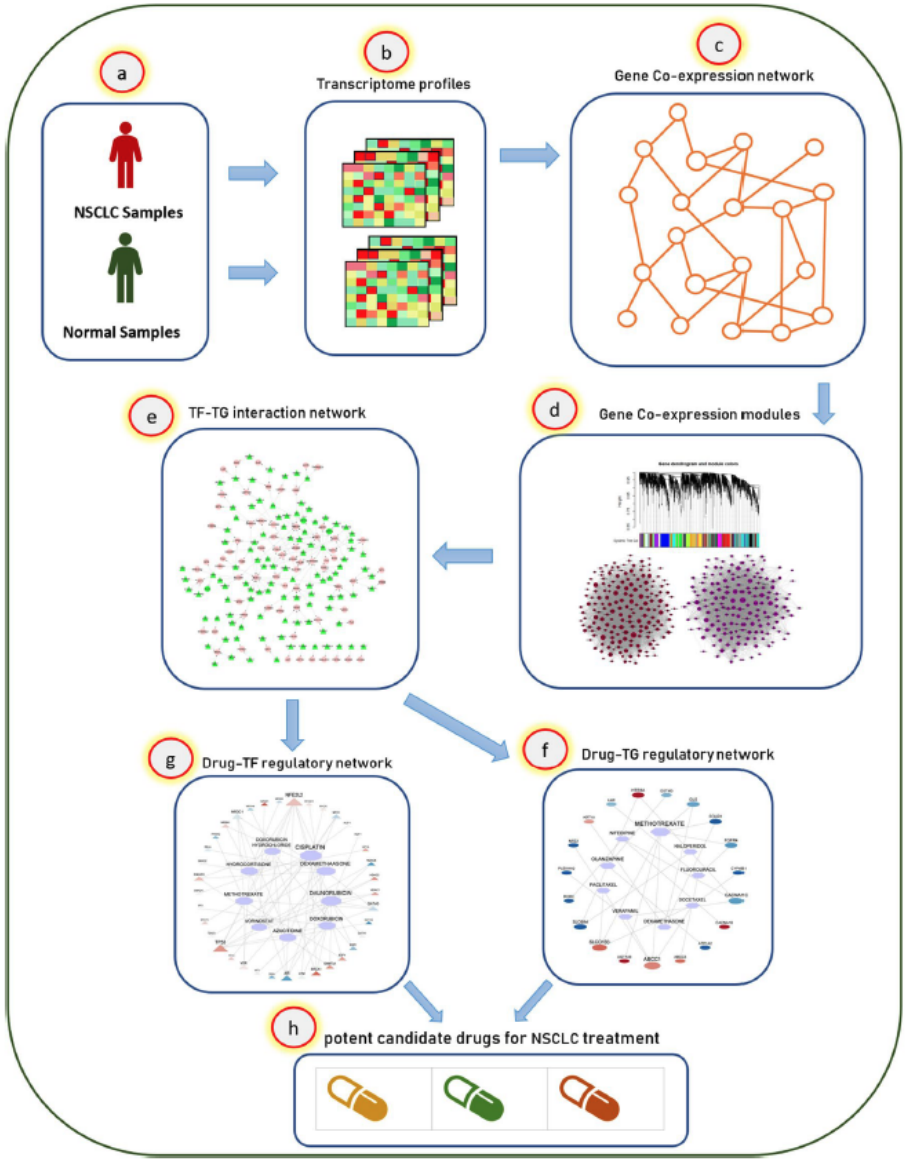
[a] Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[b] Department of Computer Science, School of Mathematics, Statistics, and Computer Science, University of Tehran, Tehran, Iran
[c] Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran



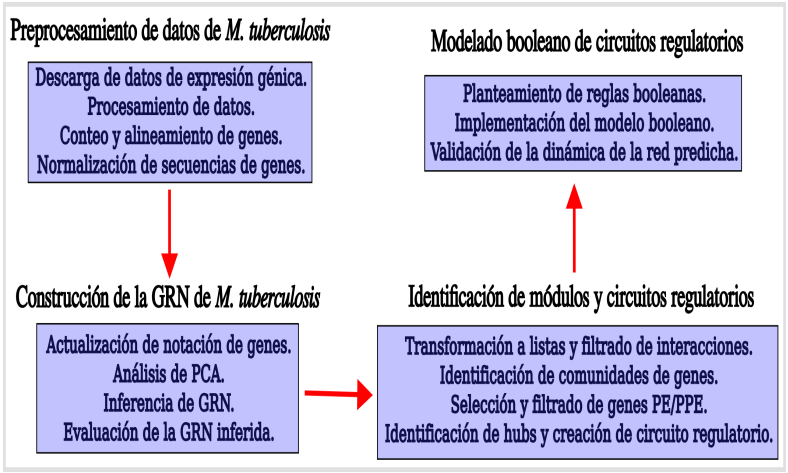
Drug Target Discovery

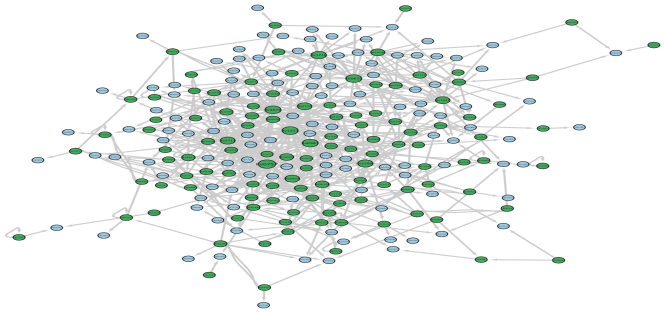
Drug repositioning in non-small cell lung cancer (NSCLC) using gene co-expression and drug–gene interaction networks analysis

Habib MotieGhader[1,2✉], Parinaz Tabrizi-Nezhadi[1,2], Mahshid Deldar Abad Paskeh[3], Behzad Baradaran[4], Ahad Mokhtarzadeh[4✉], Mehrdad Hashemi[3,5], Hossein Lanjanian[6], Seyed Mehdi Jazayeri[7], Masoud Maleki[1], Ehsan Khodadadi[8], Sajjad Nematzadeh[9], Farzad Kiani[10], Mazaher Maghsoudloo[5,11] & Ali Masoudi-Nejad[11]



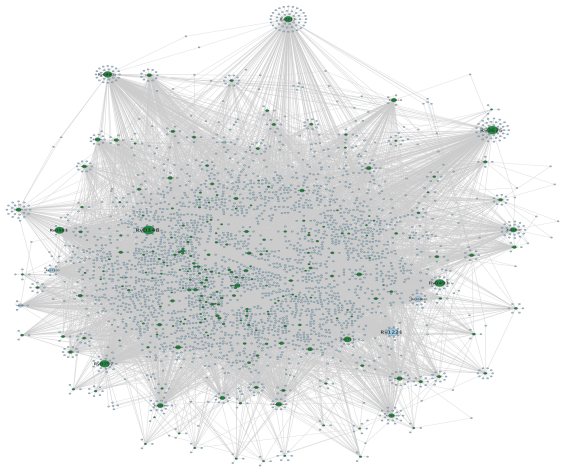
Gene Regulatory Networks: Dynamics of genetic regulatory circuits of Mycobacterium tuberculosis





32 series and 725 samples

124 of 154 genes of PE/PPE families

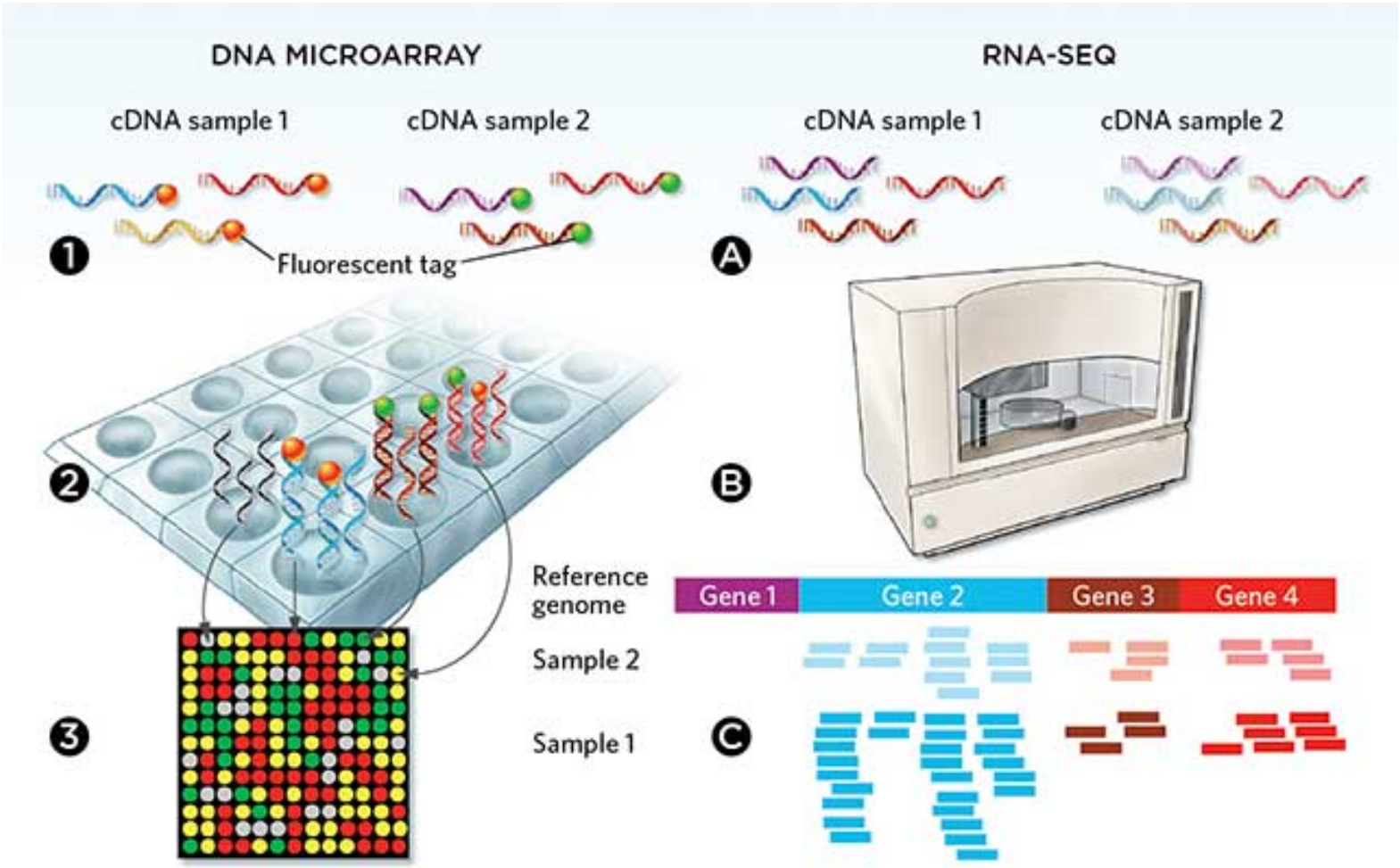


Grenedine: 3514 nodes and 10143 edges

BoolNet

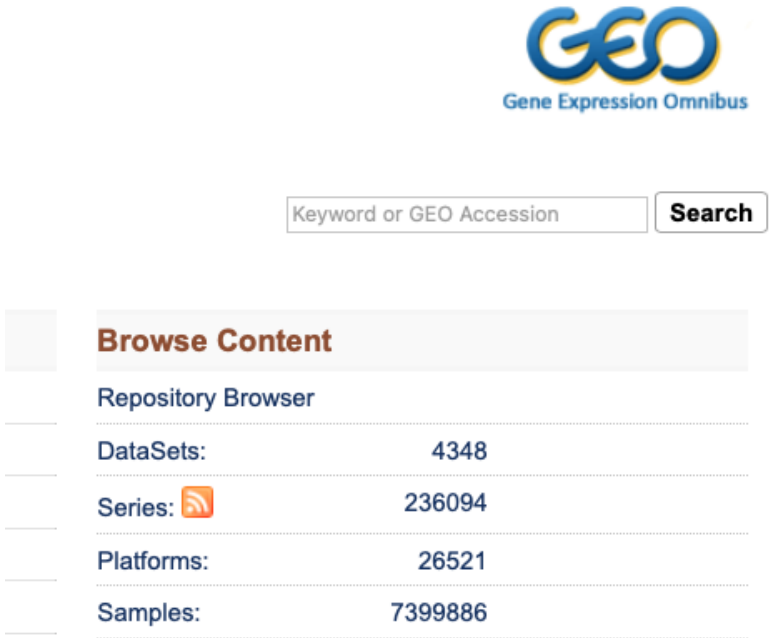
How do we build Gene Co-expression Networks?

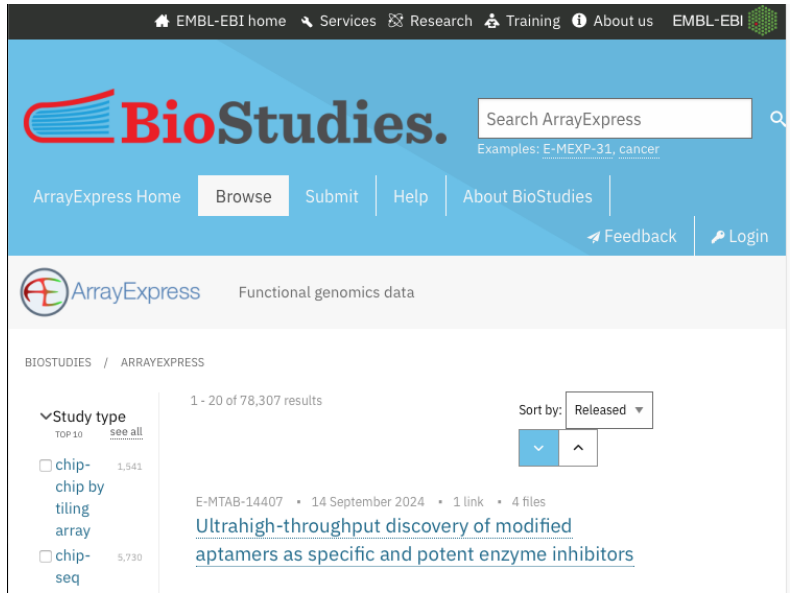
First, we need the expression data.



Devesa-Peiró et al. 2020

There are databases with genetic expression data

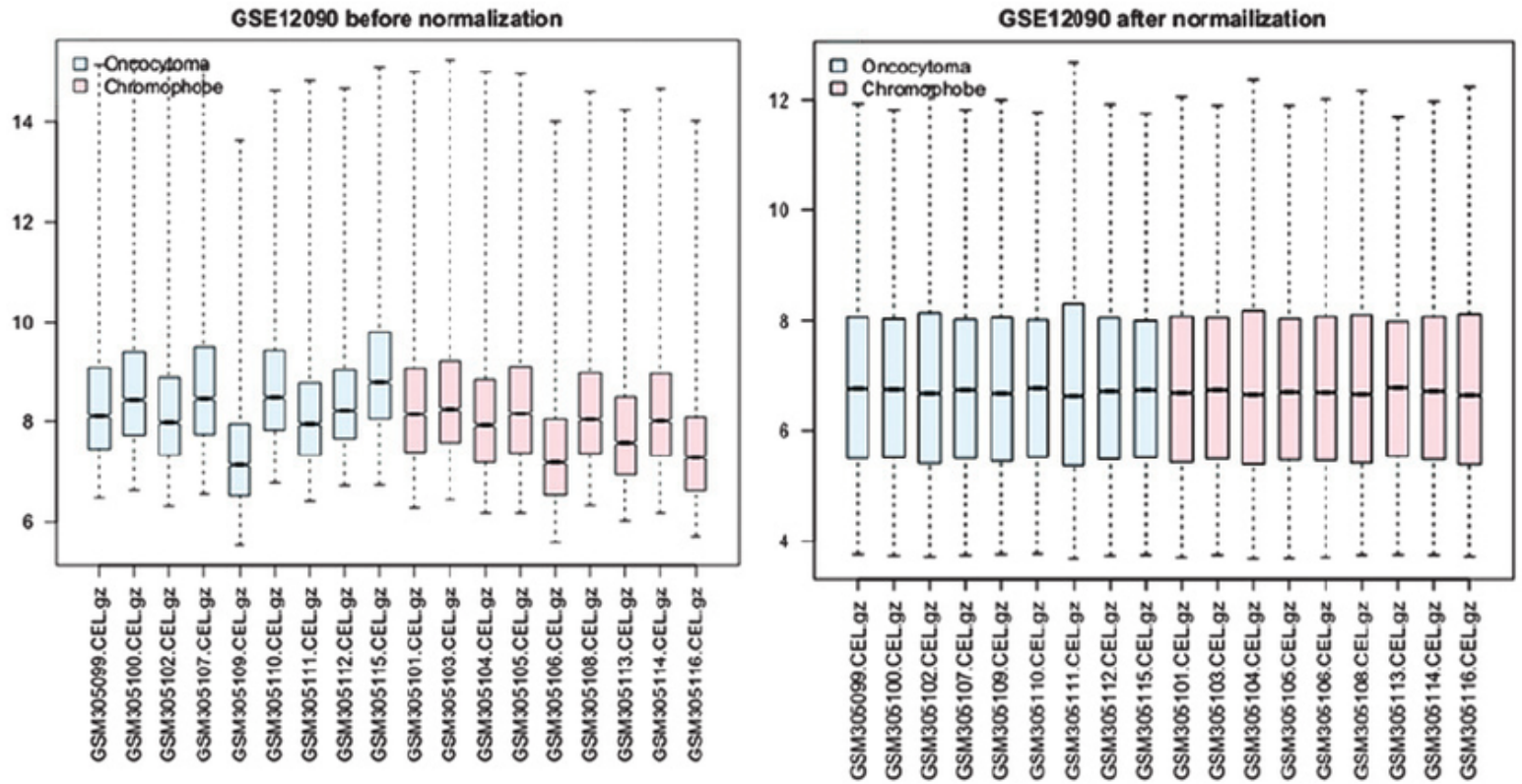




We can take advantage of published experiments

https://www.ebi.ac.uk/biostudies/arrayexpress/studies
https://www.ncbi.nlm.nih.gov/geo/

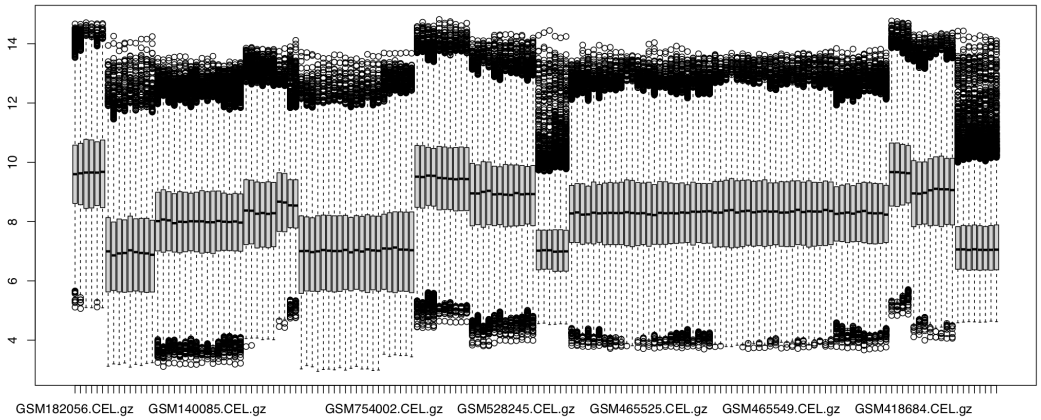
Data preprocessing: Normalization

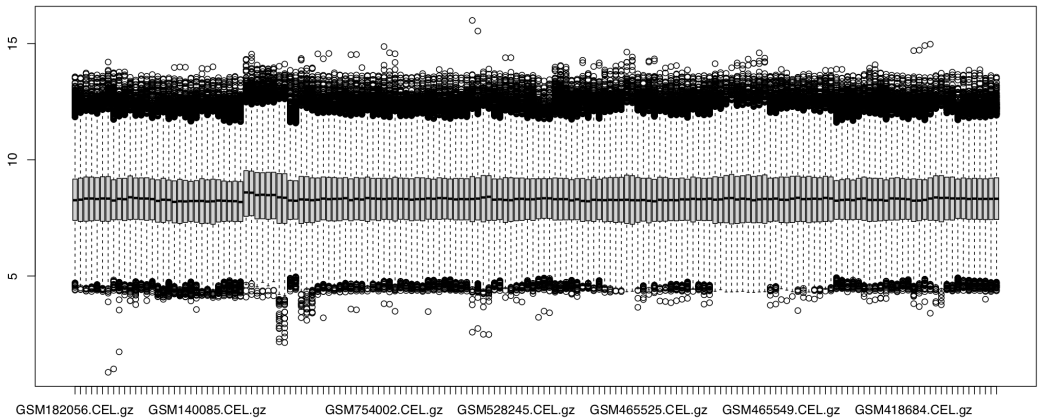


Liu et al. 2014

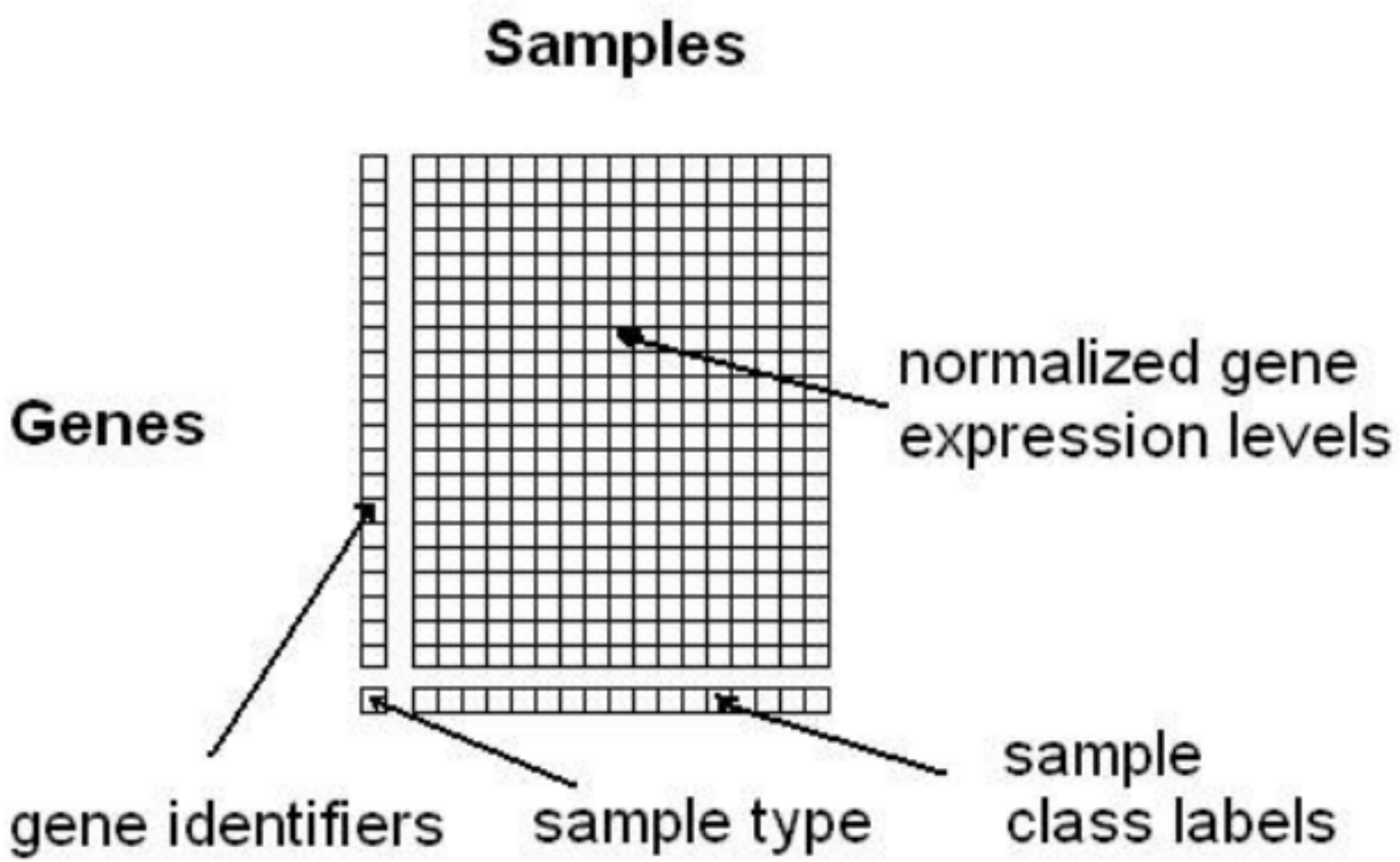
Data preprocessing: Batch ajustment

-ComBat: removes batch effects impacting both the means and variances of each gene across the batches.
-Limma: fits a linear model for each variable given a series of conditions as explanatory variables, including the batch effect and treatment effect.



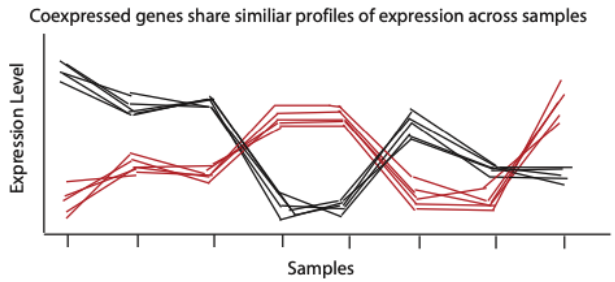


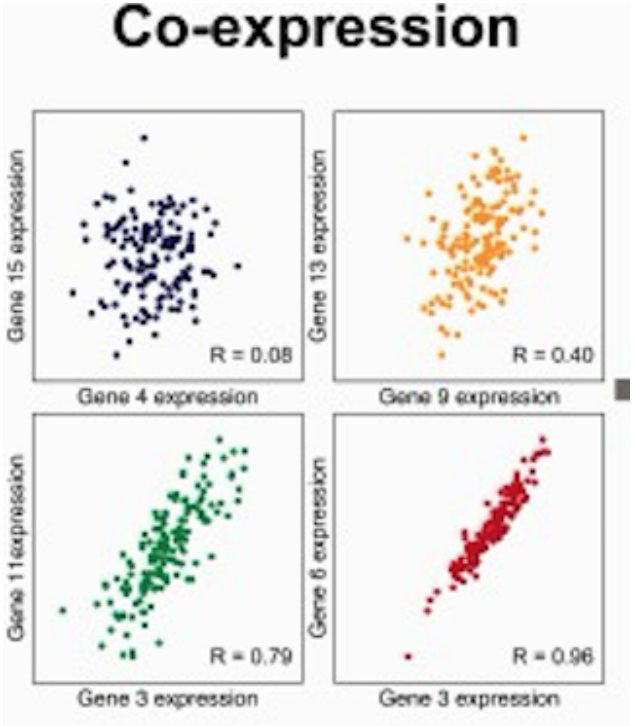
Gene expression matrix



Anil, 2017

Gene similarity matrix





	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	1.00	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	1.00	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	1.00	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	1.00	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	1.00	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	1.00	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	1.00	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	1.00	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	1.00	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	1.00

Similarity (Co-expression) score

Farber and Mesner, 2026; Anil, 2017

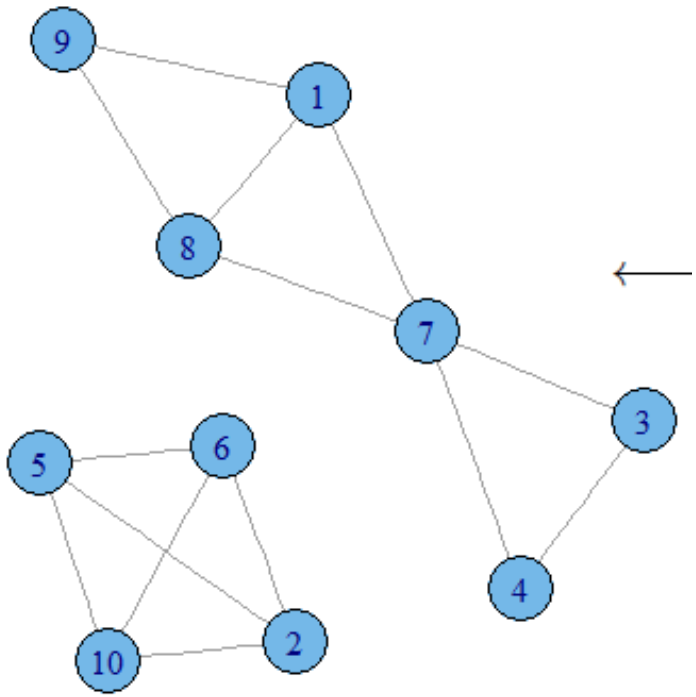
Measures

- Pearson correlation: "Standart" measure of linear correlation. Fastest, but sensitive to outliers:
- Biweight mid-correlation: robusts and less sensitive to outliers but much slower.
- Spearmen correlation: rank-basedworks even if relationship is not linear, less sensitive to gene expression differences

Thresholding

Ahora necesitamos definir un umbral para extraer la red de co-expresión genética.

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

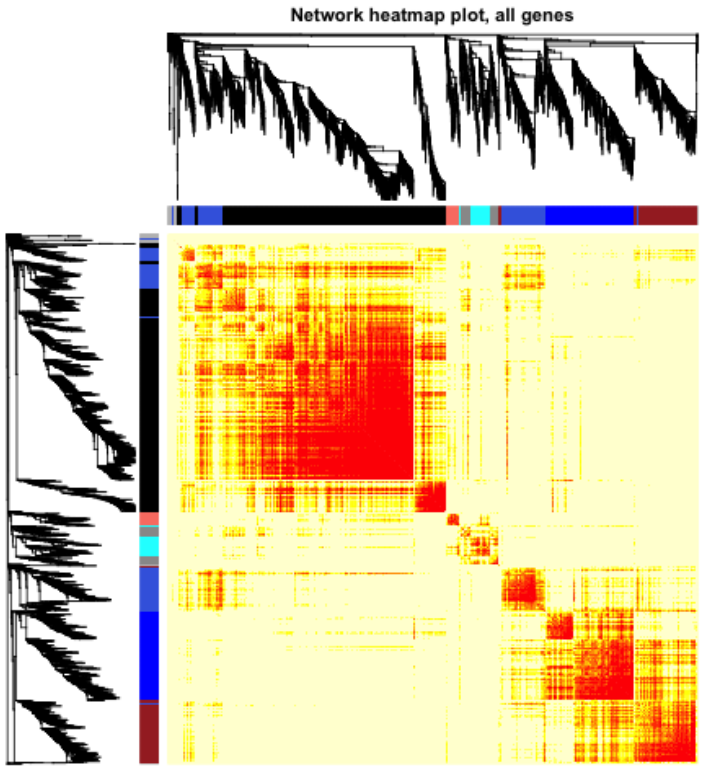


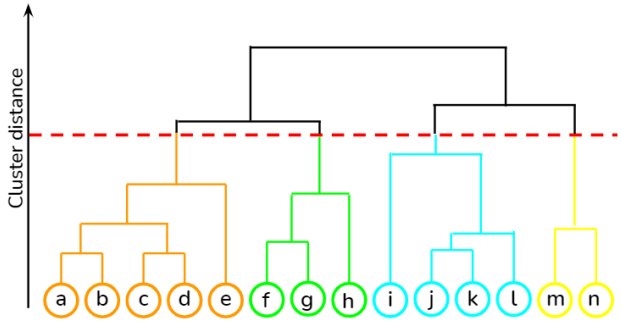
Adamic, 1999; Lancichinetti et al. 2010

Identifying Gene Modules

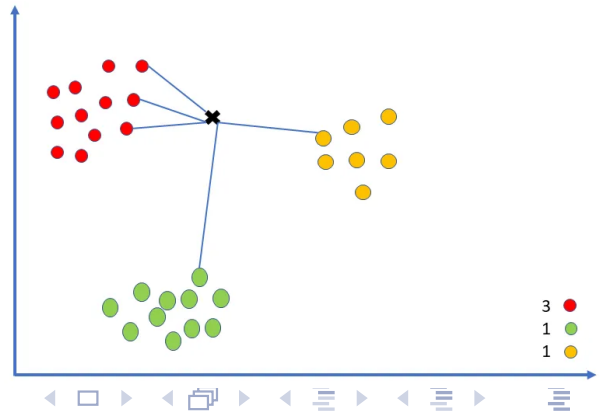
Además podemos clusterizan la matriz de similitud, para identificar módulos.

- Hierarchical Clustering



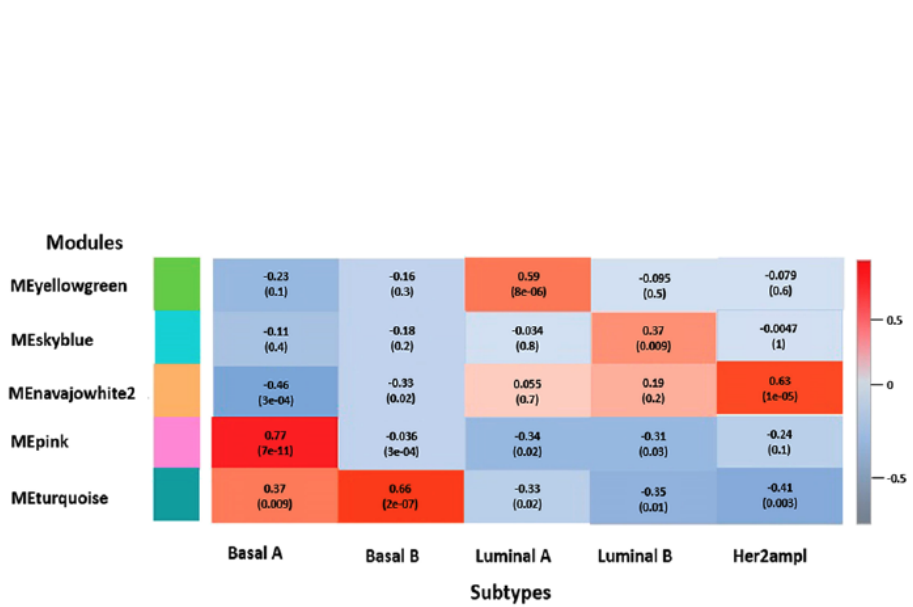


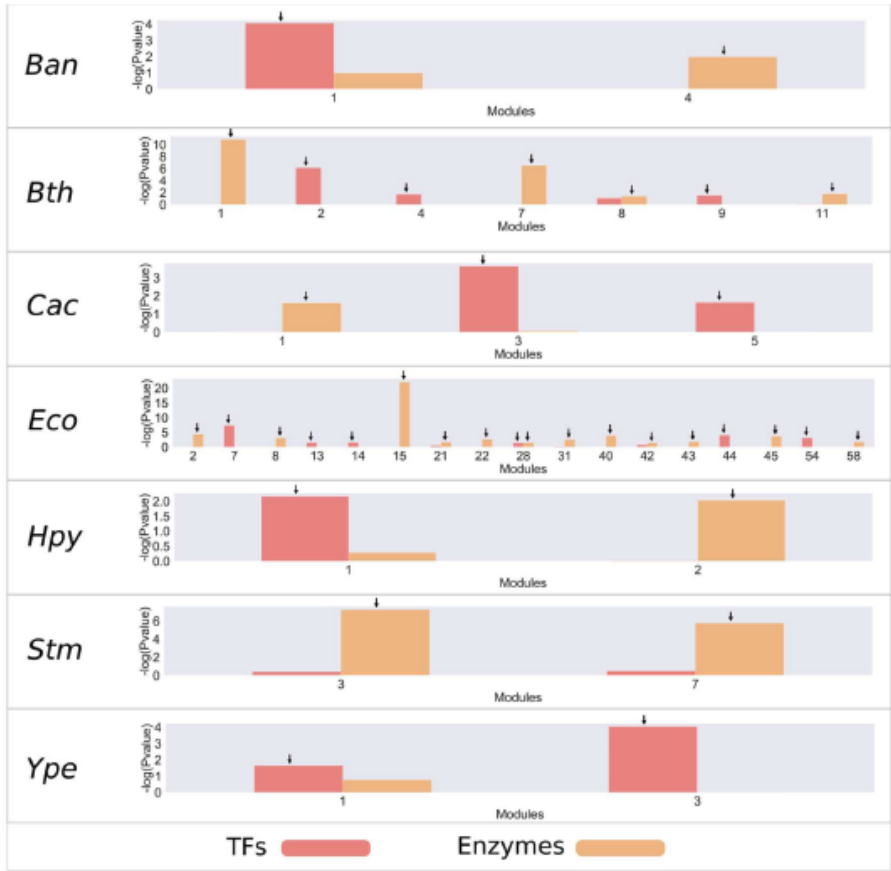
-K-Nearest Neighbors



Analyzing Gene Co-expression Networks

Identification of modules of interest

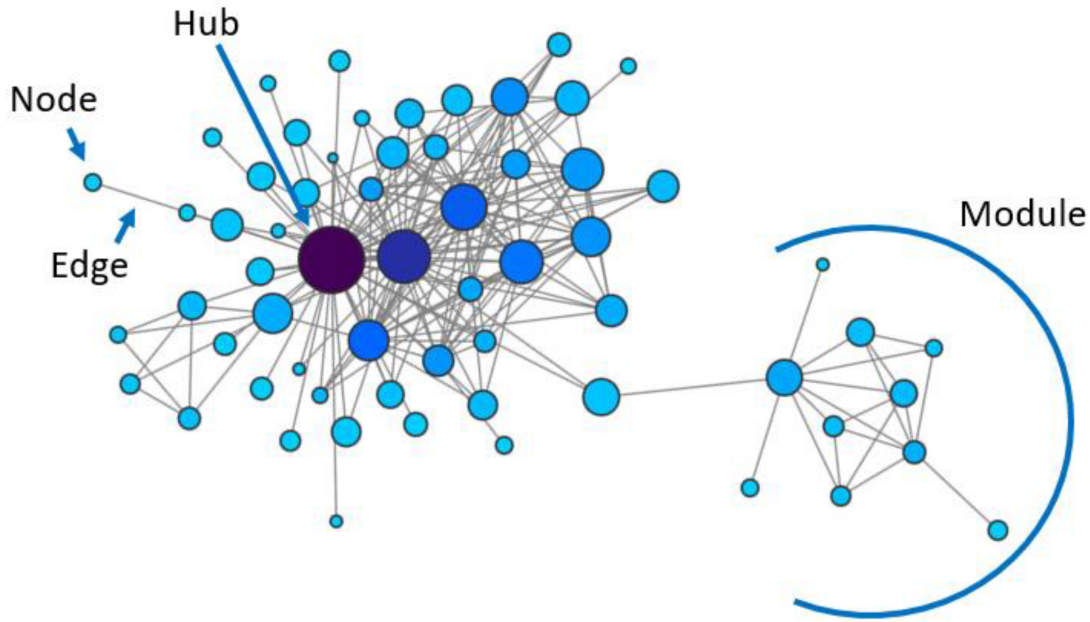


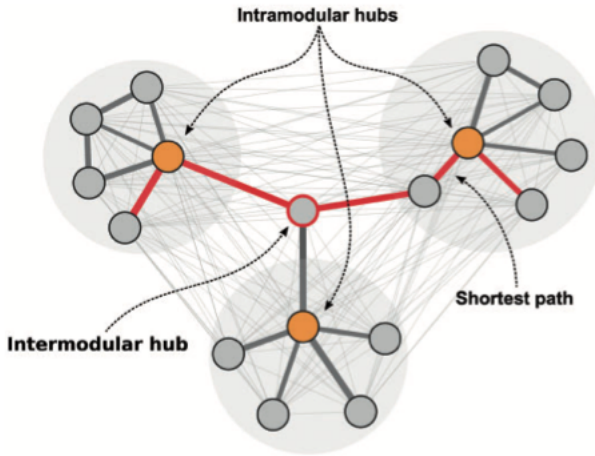


Galan-Vasquez and Perez-Rueda 2019; Mares-Quiñones et al. 2024

Hub genes

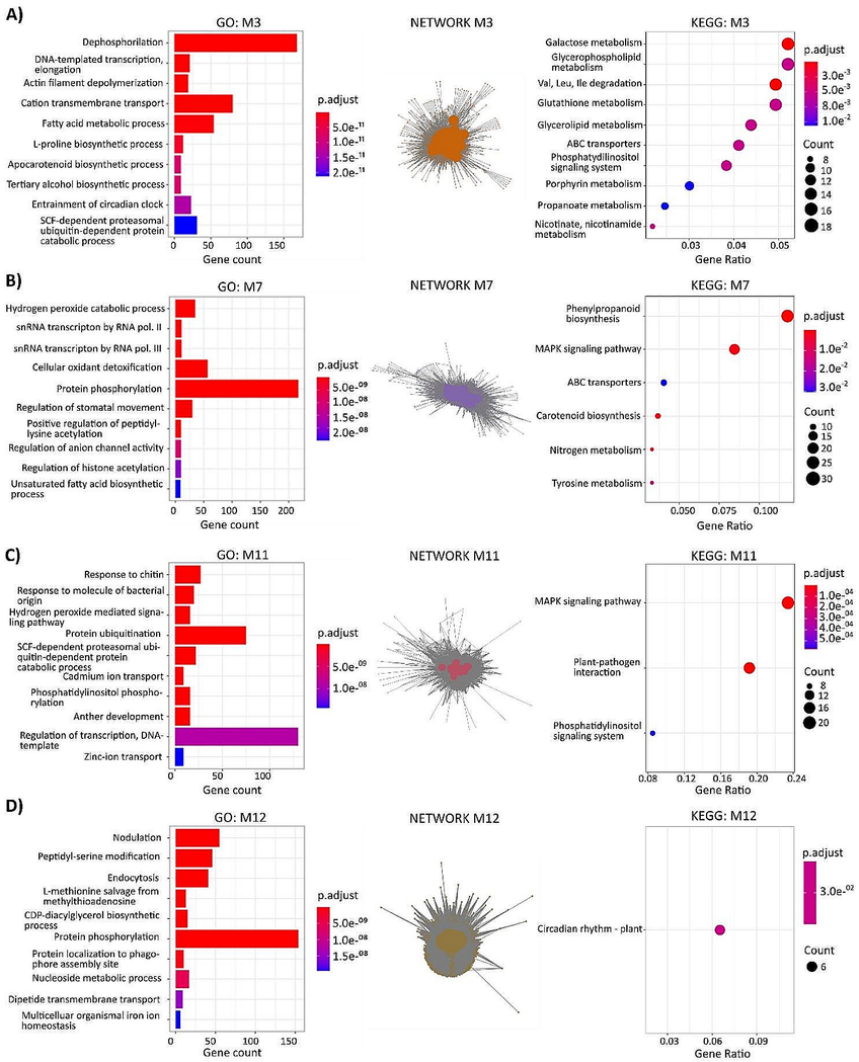
The hub gene is defined as the gene with the highest degree of connectivity in the hub module





van Dam et al, 2018

Gene Set Enrichment Analysis



Martha Sainz et al. 2024

Tools

I

Co-expression module detection	
WGCNA [54] https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/	• A tool that constructs a co-expression network using Pearson correlation (default) or a custom distance measure. • Uses hierarchical clustering and has various 'tree cutting' options to identify modules. + Most widely used tool, well supported and documented.
DiffCoEx [100]	• A method that uses a similar approach to WGCNA to identify and group differentially co-expressed genes instead of identifying co-expressed modules.
DICER [4]	• Identifies modules of genes that have the same different partners between different samples. • A method that identifies modules that correlate differently between sample groups, e.g. modules that form one large interconnected module in one group compared with several smaller modules in another group.
CoXpress [101] http://coxpress.sourceforge.net/	• A tool that identifies co-expression modules in each sample group and tests whether the genes within these modules are also co-expressed in other groups.
DINGO [102]	• DINGO is a more recent tool that groups genes based on how differently they behave in a particular subset of samples (representing e.g. a particular condition) from the baseline co-expression determined from all samples
GSCNA [103]	• A tool that tests whether a predefined defined gene set is differentially expressed between two sample groups.
GSVD [104]	• A method that identifies 'genelets', which can be interpreted as modules representing partial co-expression signals from multiple genes. These signals are then compared between two groups to identify genelets unique to samples and genelets that are shared between the two groups.
HO-GSVD [105] https://github.com/aanchan/hogsvd-python/blob/master/README.md	• A tool similar to GSVD, but that can be used across multiple sample groups rather than only two.
Biclustering [106]	• A group of methods that identify modules that are unique to a subpopulation of samples without the need for prior grouping of samples.

Van Dam et al. 2018

Tools

Regulatory network inference	
ARACNE [112]	• A tool that removes indirect connections between genes (i.e. partners of a gene that have a stronger correlation with each other than with the gene itself), leaving only those connections that are expected to be regulatory. + Creates directional networks.
Genie3 [113]	• A tool that incorporates TF information to construct a regulatory network by determining the TF expression pattern that best explains the expression of each of their target genes. + Creates directional networks. − Requires TF information.
CoRegNet [114]	• A tool that identifies co-operative regulators of genes from different data types.
cMonkey [115]	• Calculates joint bicluster membership probability from different data types by identifying groups of genes that group together in multiple data types.

Van Dam et al. 2018

Tools

Co-expression databases[a]	
COXPRESdb [60] http://coxpresdb.jp/	• A web resource incorporating 12 co-expression networks for different species created from ~157 000 microarrays and 10 000 RNA-seq samples. Has a focus on protein-coding RNAs.
GeneFriends [2] http://www.genefriends.org/	• Human and mouse gene and transcript co-expression networks. • Networks constructed from ~4000 RNA-seq samples each. + Includes a number of non-coding RNAs (~10 000 for mouse and ~25 000 for human).
GeneMANIA [118] http://www.genemania.org/	• Also includes physical and genetic interaction, co-localization, pathway and shared protein domain information data sets. + Networks for nine species.
GENEVESTIGATOR [119] https://genevestigator.com/gv/	• A database constructed using ~145 000 samples. + Curated database. + Networks for 18 species. + Multiple data types.
GIANT [120] http://giant.princeton.edu/	• Tissue-specific interaction network database. • Includes 987 Datasets encompassing 38 000 conditions describing 144 tissues types. + Integrates physical interaction, co-expression, miRNA binding motif and TF binding site data.

Van Dam et al. 2018

Case study

Modules in fungus

Construction and analysis of gene co-expression network in the pathogen *ilago maydis*

https://www.frontiersin.org/articles/10.3389/fmicb.2022.1048694/full

Cinthia V. Soberanes-Gutiérrez[1†], Alfredo Castillo-Jiménez[2†], Ernesto Pérez-Rueda[3] and Edgardo Galán-Vásquez[4*]

Soberanes-Gutierrez et al. 2022

Modules in fungus

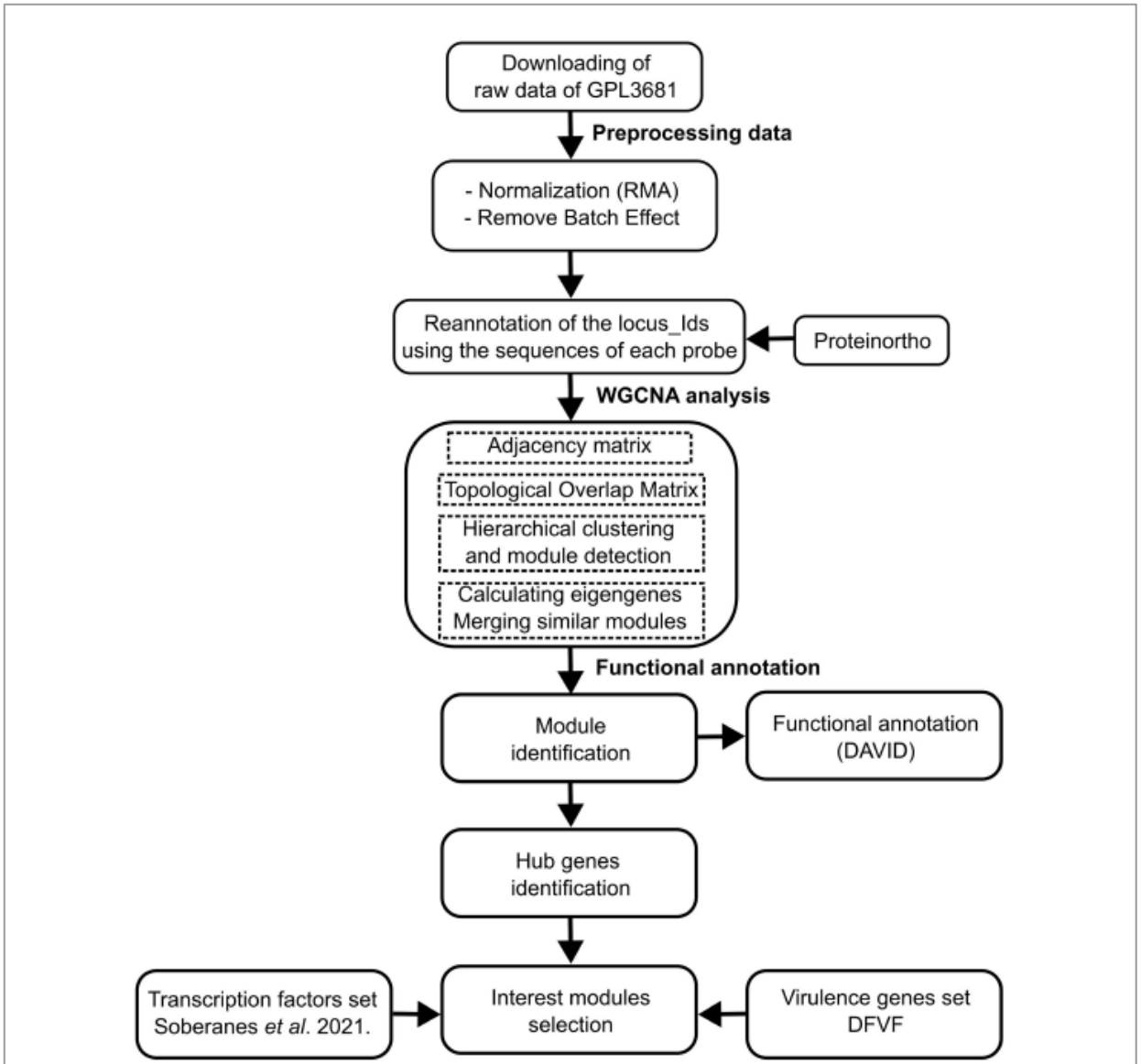


FIGURE 1
Schematic workflow of the data of preparation, processing, and analysis.

Modules in fungus

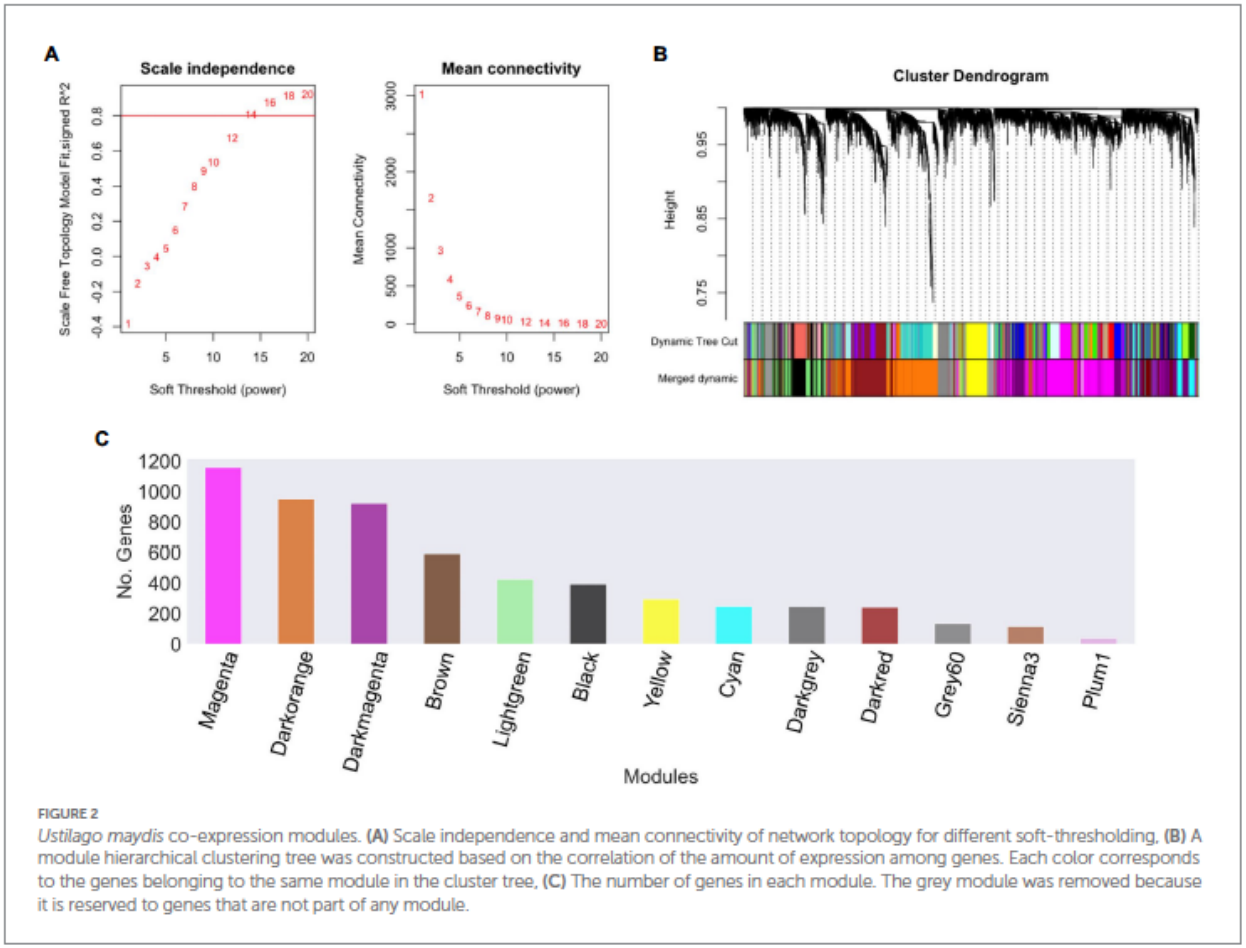
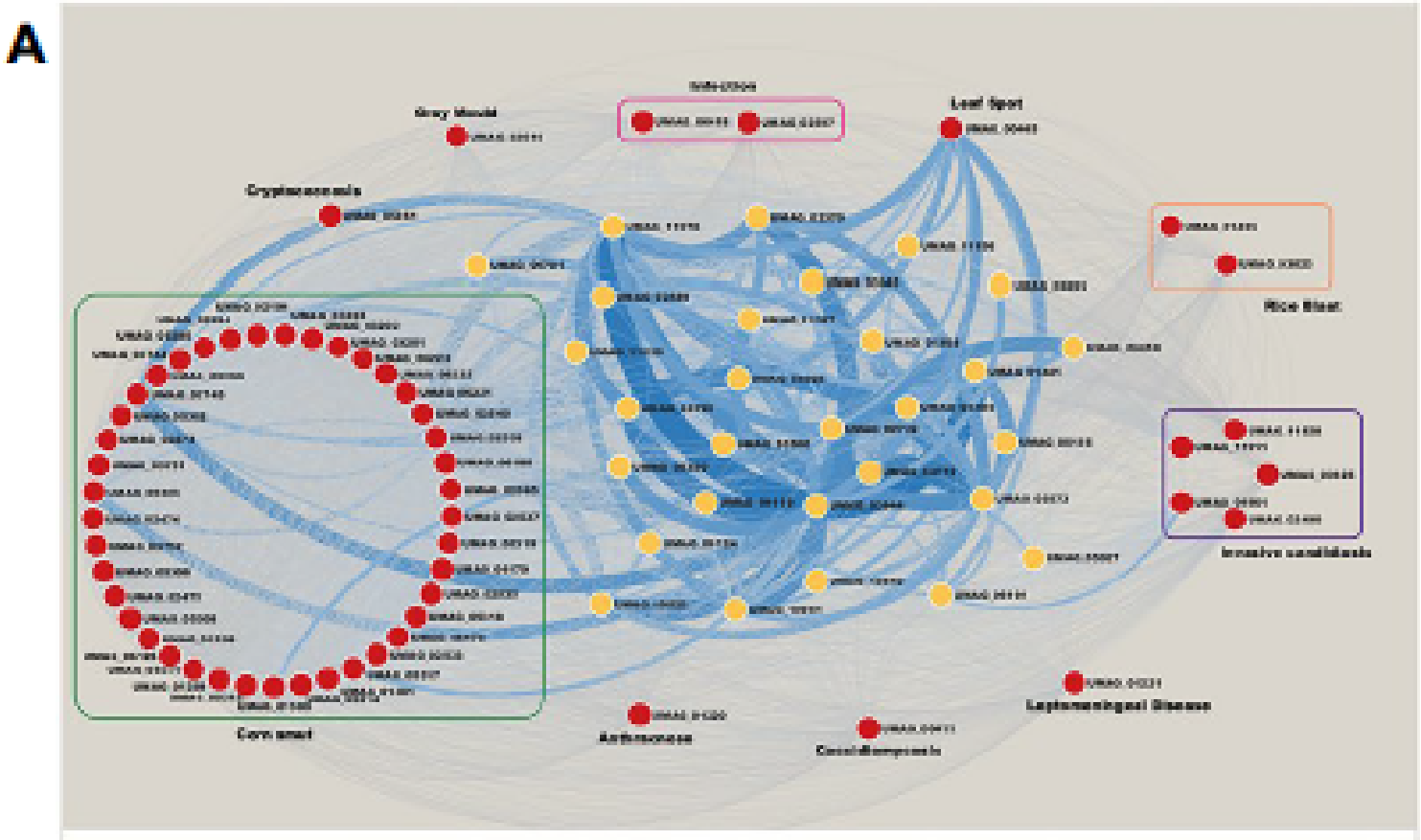


FIGURE 2
Ustilago maydis co-expression modules. **(A)** Scale independence and mean connectivity of network topology for different soft-thresholding, **(B)** A module hierarchical clustering tree was constructed based on the correlation of the amount of expression among genes. Each color corresponds to the genes belonging to the same module in the cluster tree, **(C)** The number of genes in each module. The grey module was removed because it is reserved to genes that are not part of any module.

Soberanes-Gutierrez et al. 2022

Modules in fungus



Soberanes-Gutierrez et al. 2022

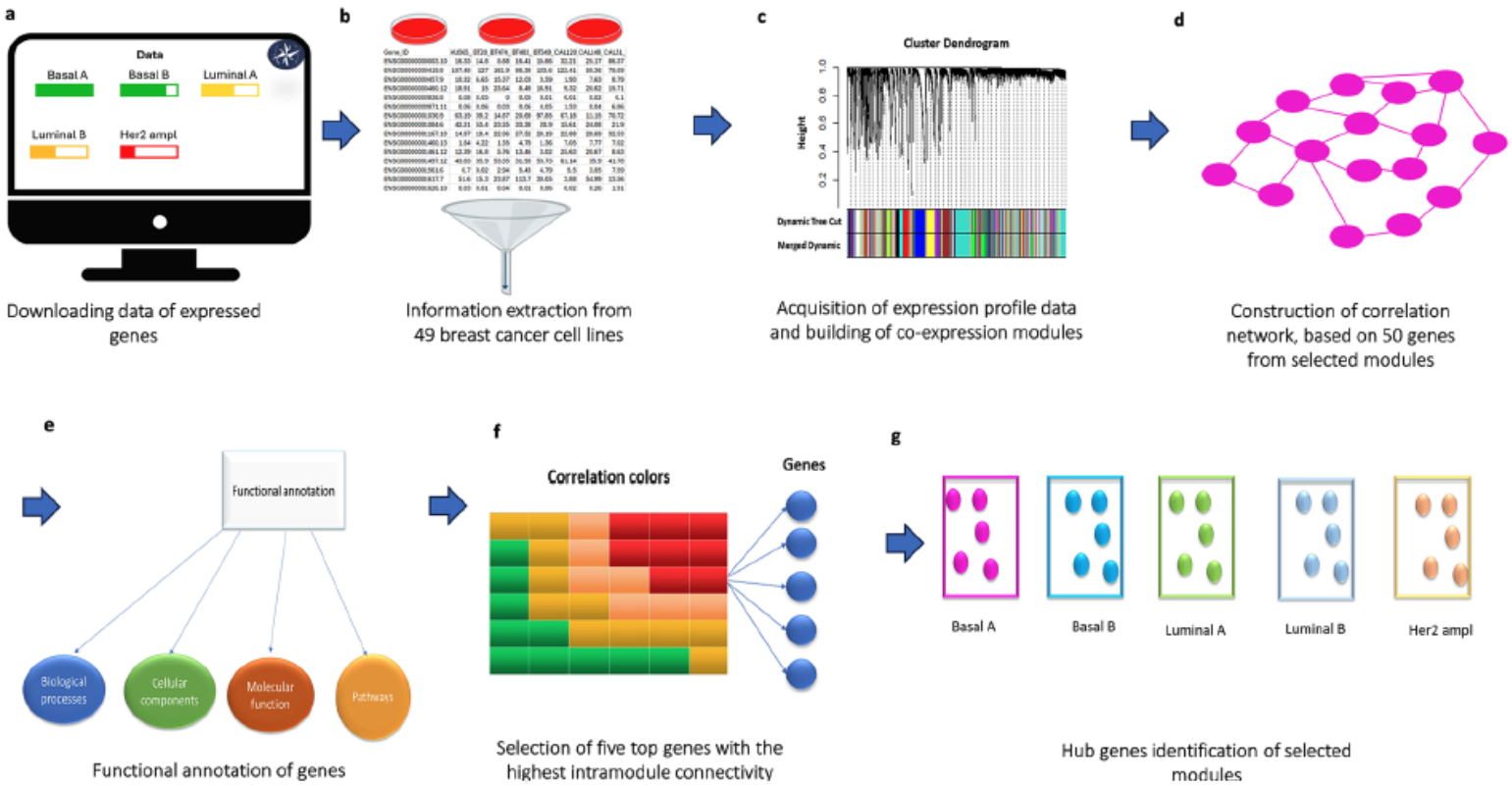
Application in Disease

Identification of modules and key genes associated with breast cancer subtypes through network analysis

María Daniela Mares-Quiñones[1], Edgardo Galán-Vásquez[2], Ernesto Pérez-Rueda[3], D. Guillermo Pérez-Ishiwara[1], María Olivia Medel-Flores[1] & María del Consuelo Gómez-García[1]✉

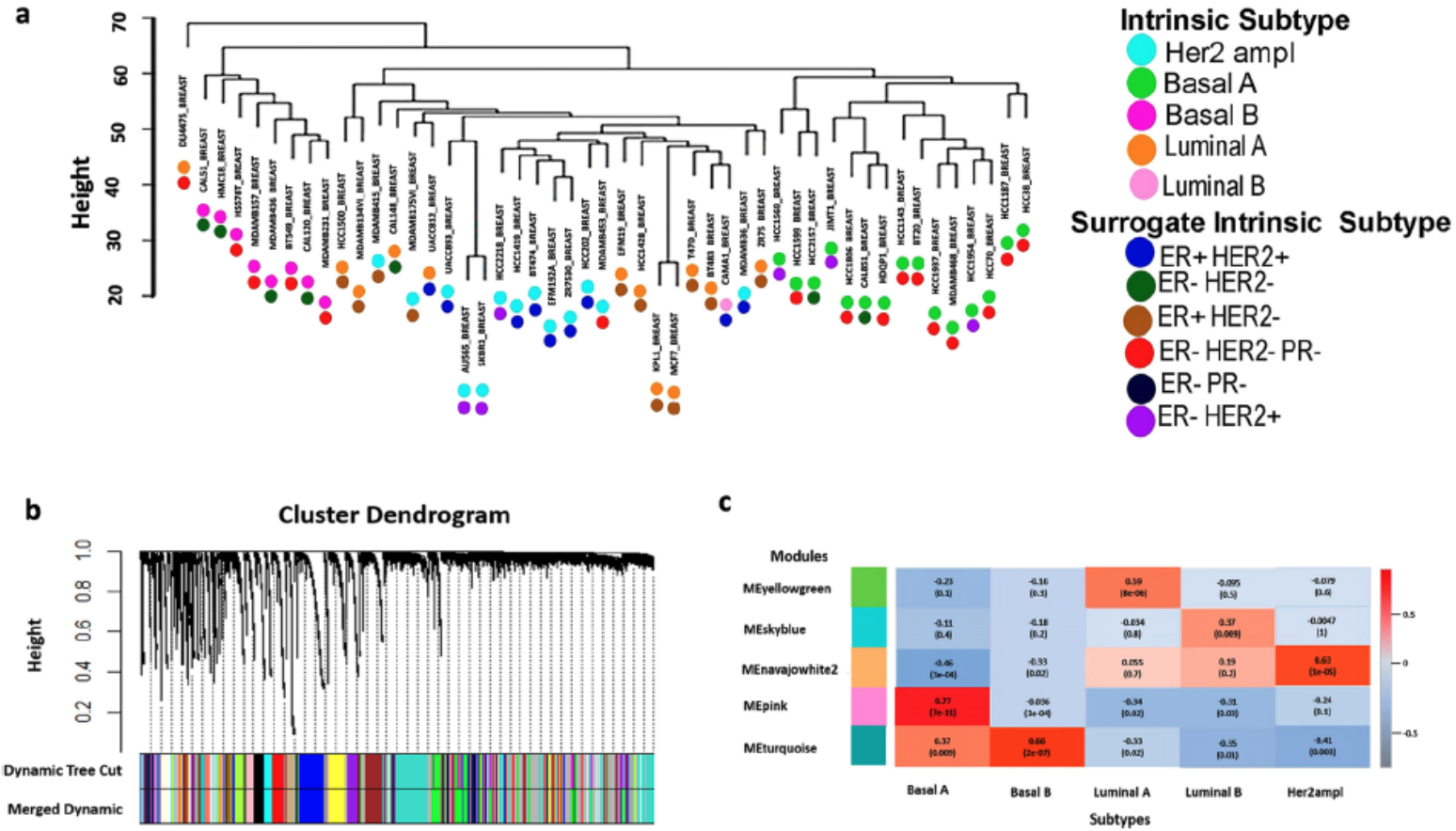
Mares-Quiñones et al. 2024

Application in Disease



Mares-Quiñones et al. 2024

Application in Disease



Mares-Quiñones et al. 2024

Application in Disease

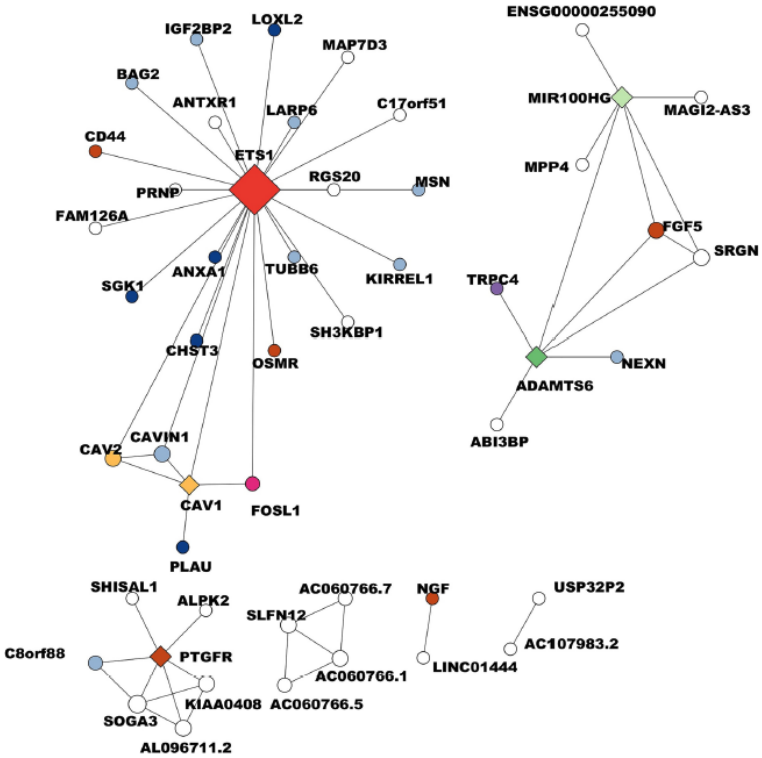
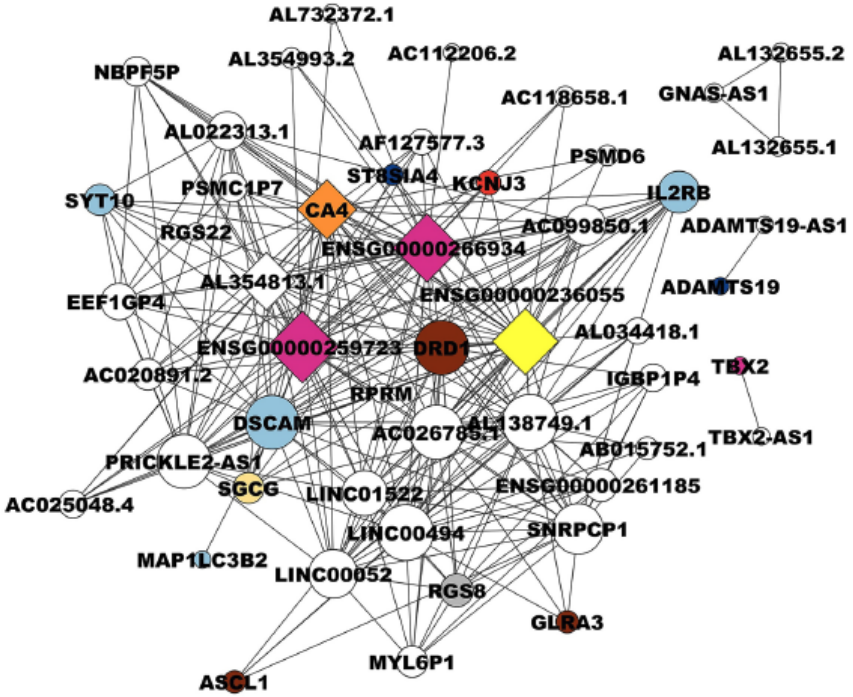


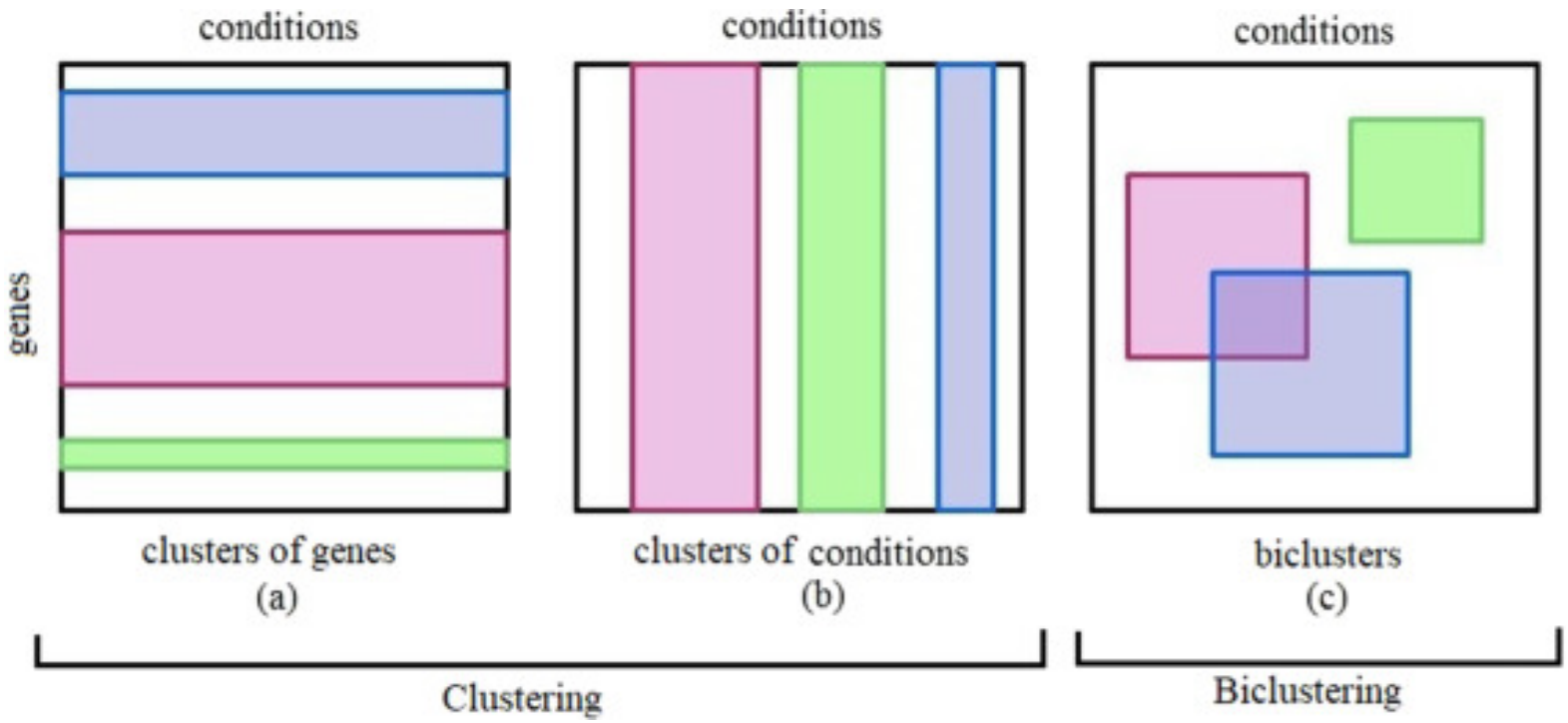
Figure 4. Basal B breast cancer subtype correlation network. The circles in the network symbolize different



Mares-Quiñones et al. 2024

Biclustering

Definitions of Biclustering



maatouk et al 2021

Challenger and Future

Challenger

- Noise in data: Gene expression data, especially from high-throughput technologies like RNA-seq, can contain noise, leading to false positive correlations.
- High Dimensionality: The sheer number of genes and samples makes network construction computationally expensive.
- Threshold Selection: Setting the right threshold for network construction is a critical and often subjective step.

Future Directions

The integration of multi-omics data (e.g., combining transcriptomics with proteomics or metabolomics) to create more comprehensive biological networks. Machine learning and AI techniques are improving the accuracy of gene regulatory network inference. Single-cell co-expression networks are gaining traction as single-cell technologies advance, offering insights into cellular heterogeneity.

WGCNA

WGCNA workflow

WGCNA is a compendium of methods to analyze "high-dimensional" data, as gene expression, methylation, proteomics, metabolomics, etc. measured across multiples samples (at least 20 but more is better).
WGCNA: Compute a correlation raised to a power between every pair of genes (i,j)
$a_{i,j} = |cor(i,j)|^{\beta}$

Langfelder and Horvath, 2008

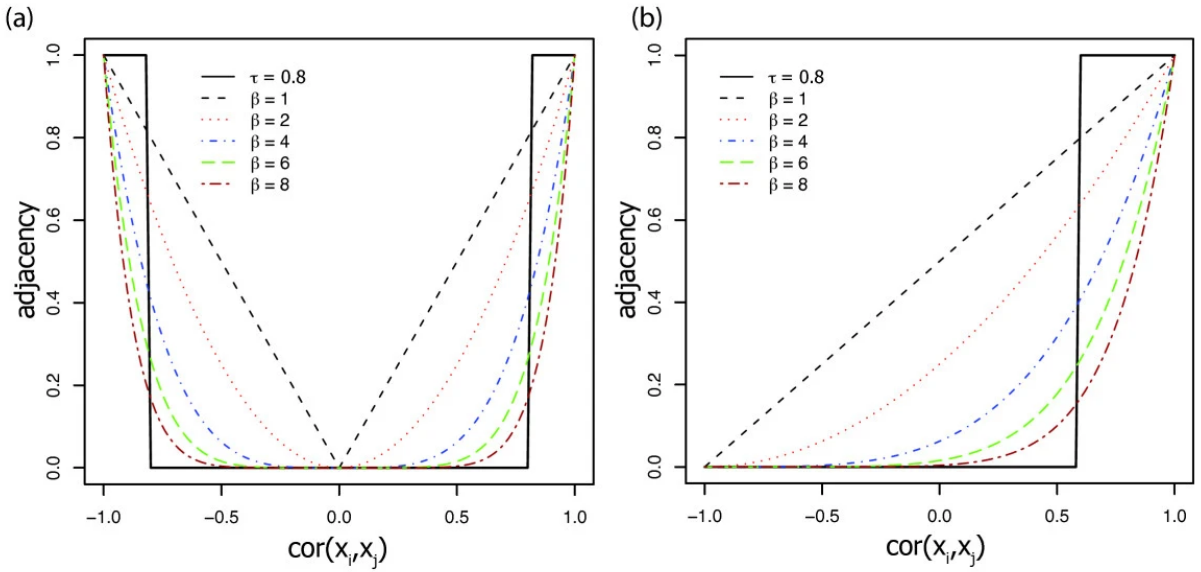
Effect of raising coorrelation to a power

Amplfies disparity between strong and weak correlations
Example: Power term $\beta = 4$
$cor(i, j) = 0.8 \longrightarrow |0.8|^{\beta} = 0.4096$
$cor(k, l) = 0.2 \longrightarrow |0.2|^{\beta} = 0.0016$
$0.8/0.2$: 4-fold difference $\longrightarrow$ $0.4096/0.0016$: 256-fold difference

Selecting a network type

Two types od weighted correlation networks

- Unsigned: absolute value. $a_{ij} = |cor(i,j)|^{\beta}$
- Signed: Preserves sing info. $a_{ij} = |0.5 + 0.5xcor(i,j)|^{\beta}$



Mason et al. 2009

Choosing a correlation method

- Fastest, but sensitive to outliers:
  - Pearson correlation cor(x)
    "standard" measure of linear correlation
- Less sensitive to outliers but much slower:
  - Biweight mid-correlation bicor(x)
    robust, recommended by the authors for most situations
  - Spearman correlation cor(x, method="spearman")
    rank-based, works even if relationship is not linear

Data cleaning and Proprocessing

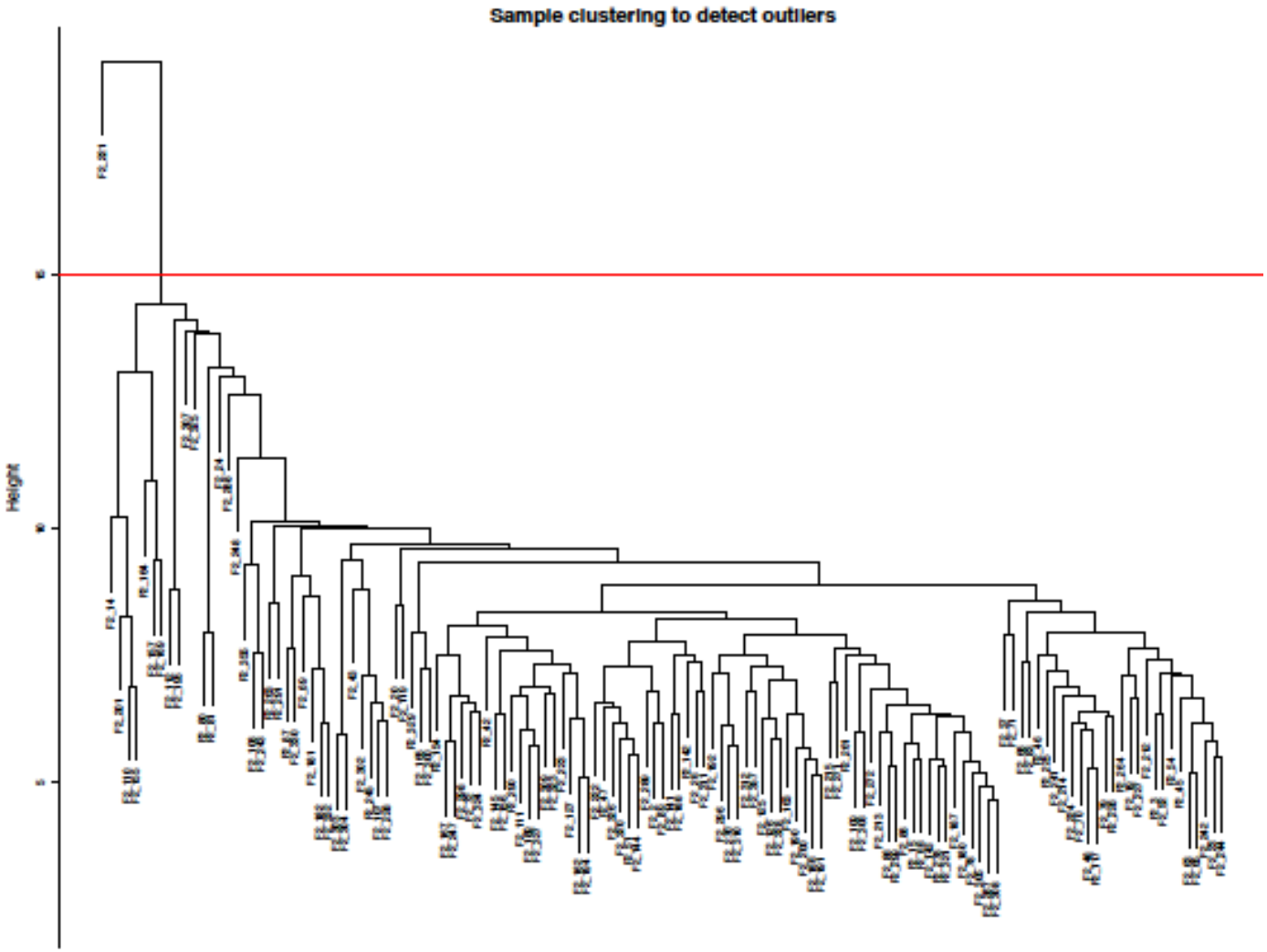


Figure 1: Clustering dendrogram of samples based on their Euclidean distance.

Data cleaning and Proprocessing

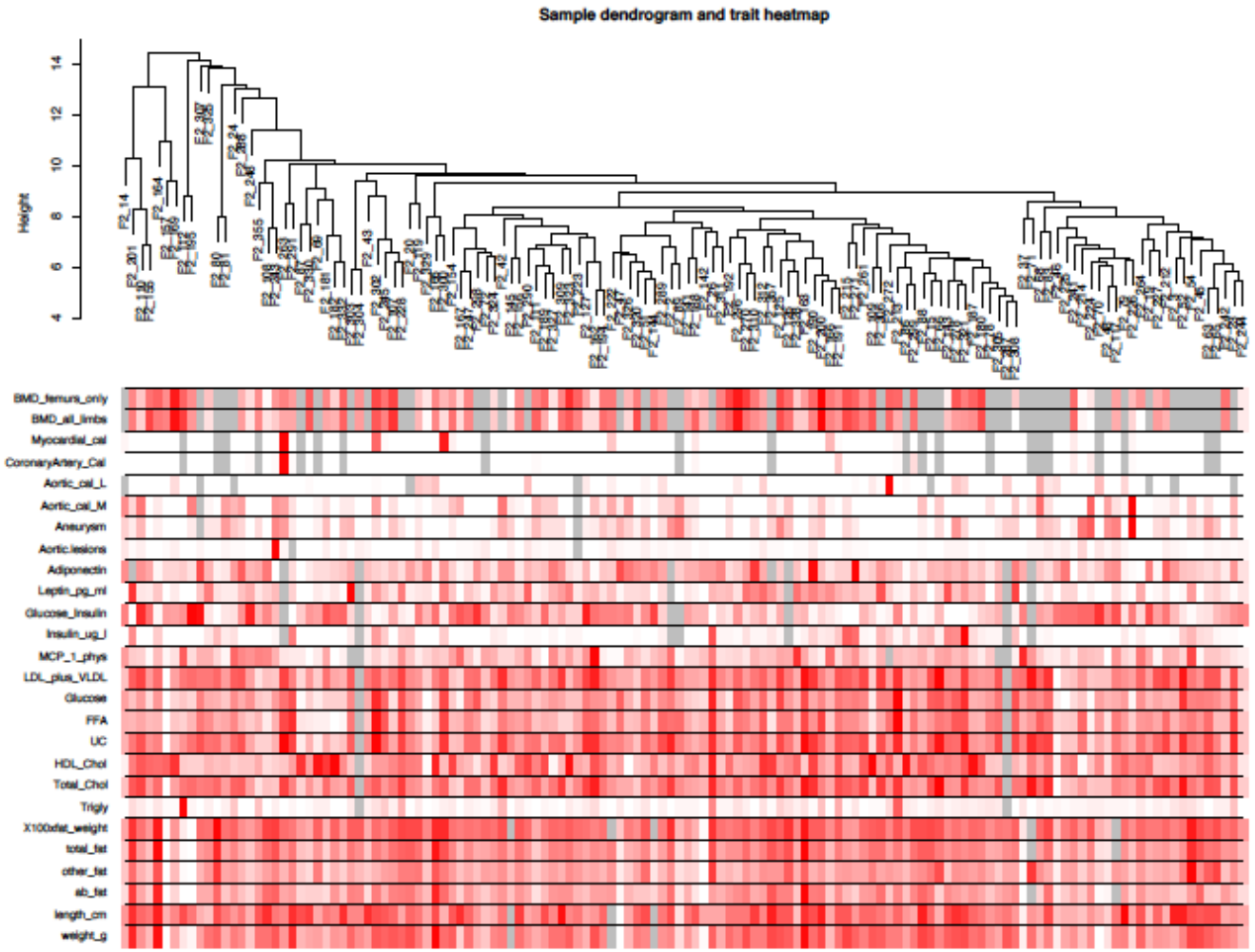
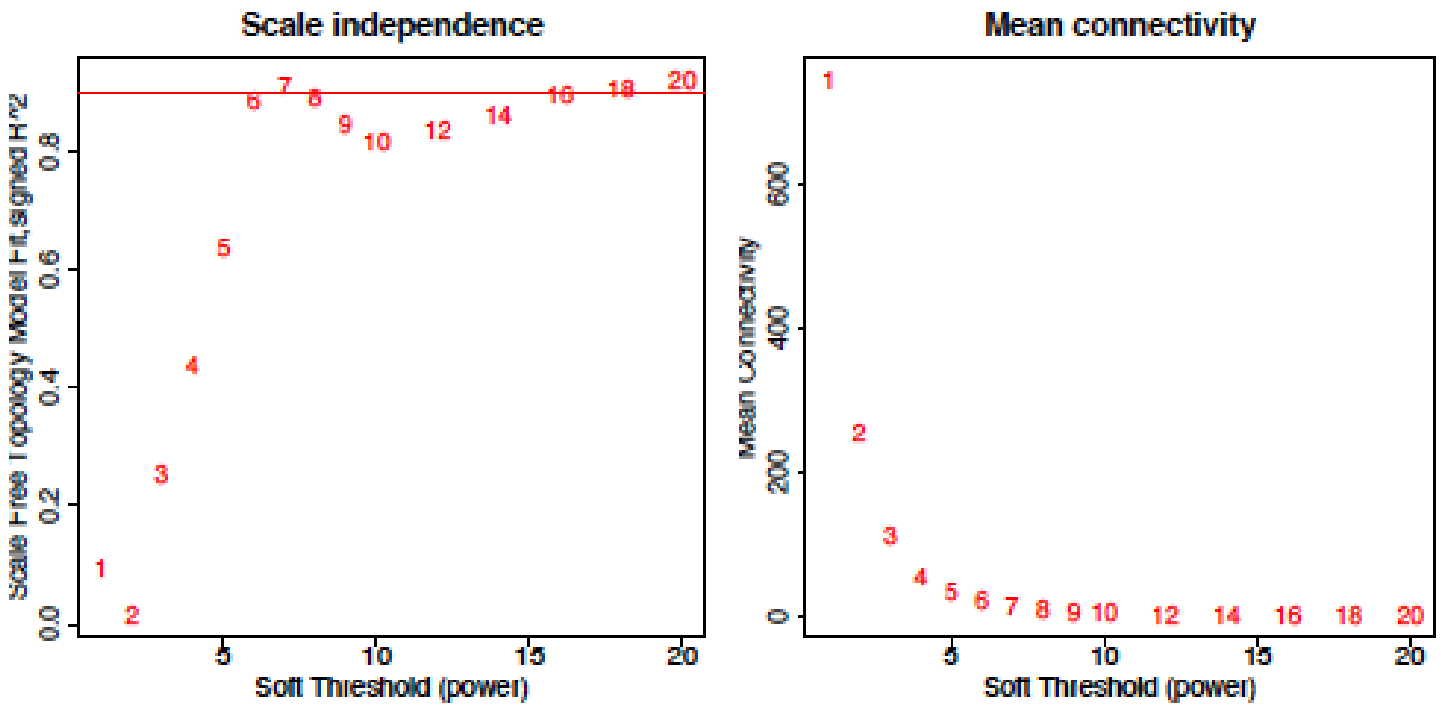


Figure 2: Clustering dendrogram of samples based on their Euclidean distance.

Softpower

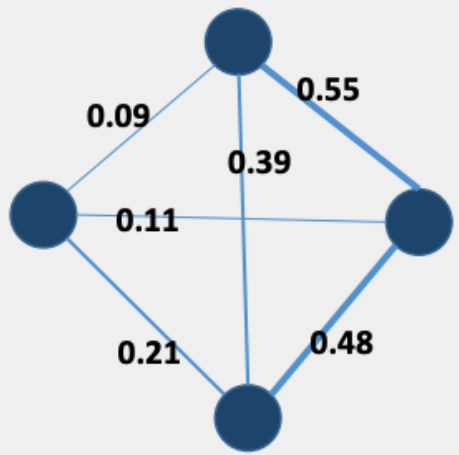

Scale independence
Scale Free Topology Model Fit,signed R^2
Soft Threshold (power)
Mean connectivity
Mean Connectivity
Soft Threshold (power)

Adjacency matrix

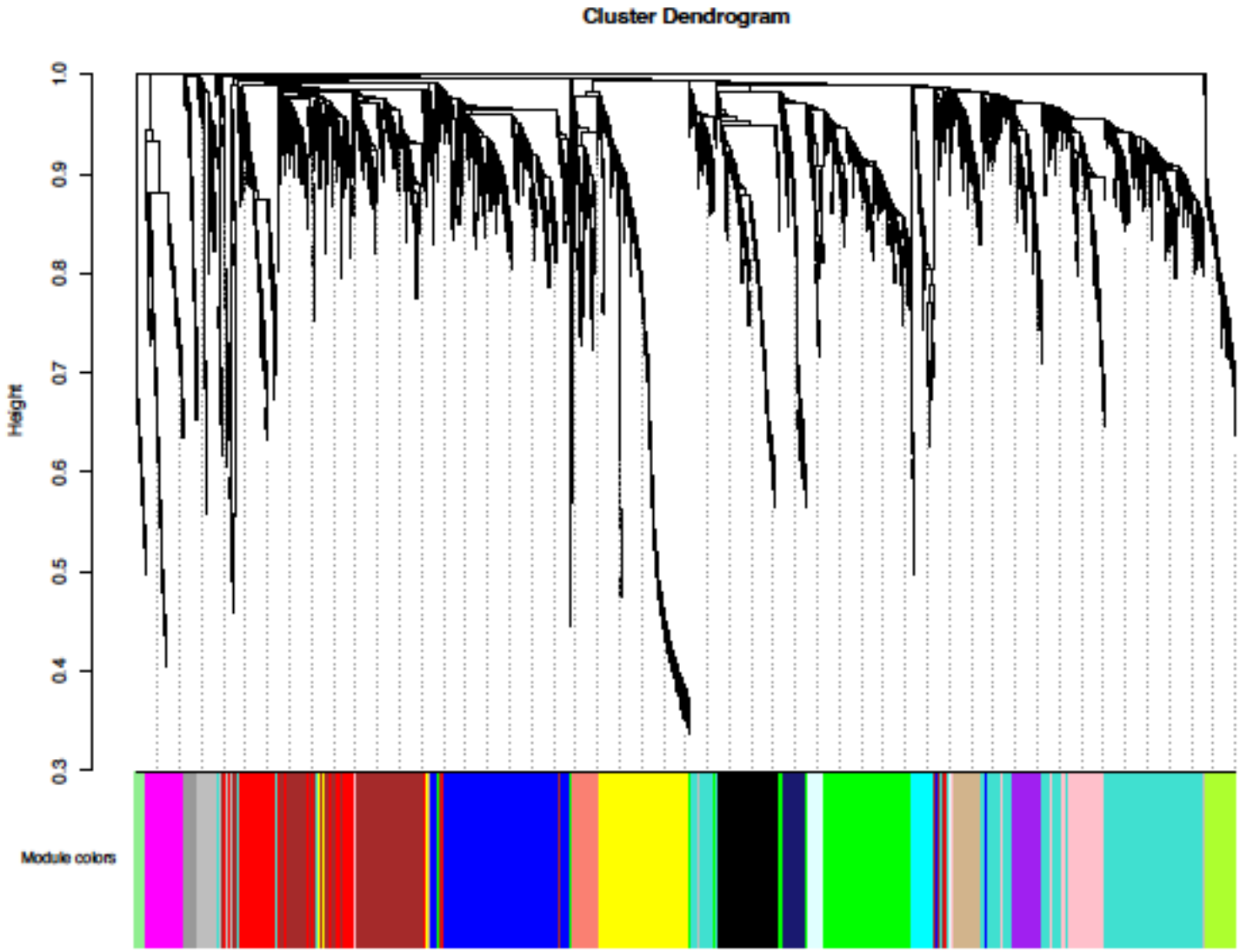
- Compute a correlation raised to a power between every pair of genes (i,j)
- Construct a full connected networks; Genes as nodes, a_{ij} as edges weights
- high correlation - strong connection
- low correlation - weak connection

Adjacency matrix of 4 genes

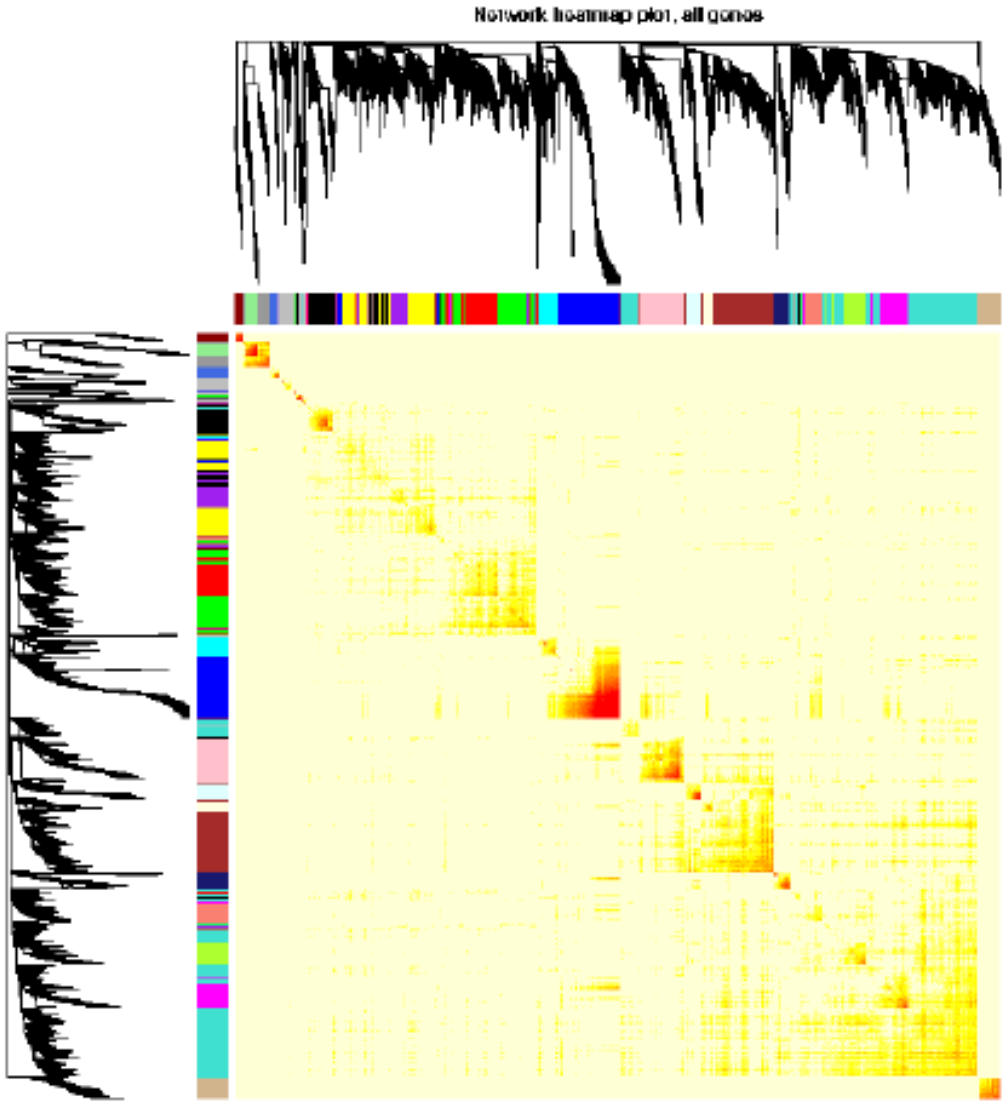
$a_{i,j}$	gene1	gene2	gene3	gene4
gene1	1	0.55	0.39	0.09
gene2	0.55	1	0.48	0.11
gene3	0.39	0.48	1	0.21
gene4	0.09	0.11	0.21	1



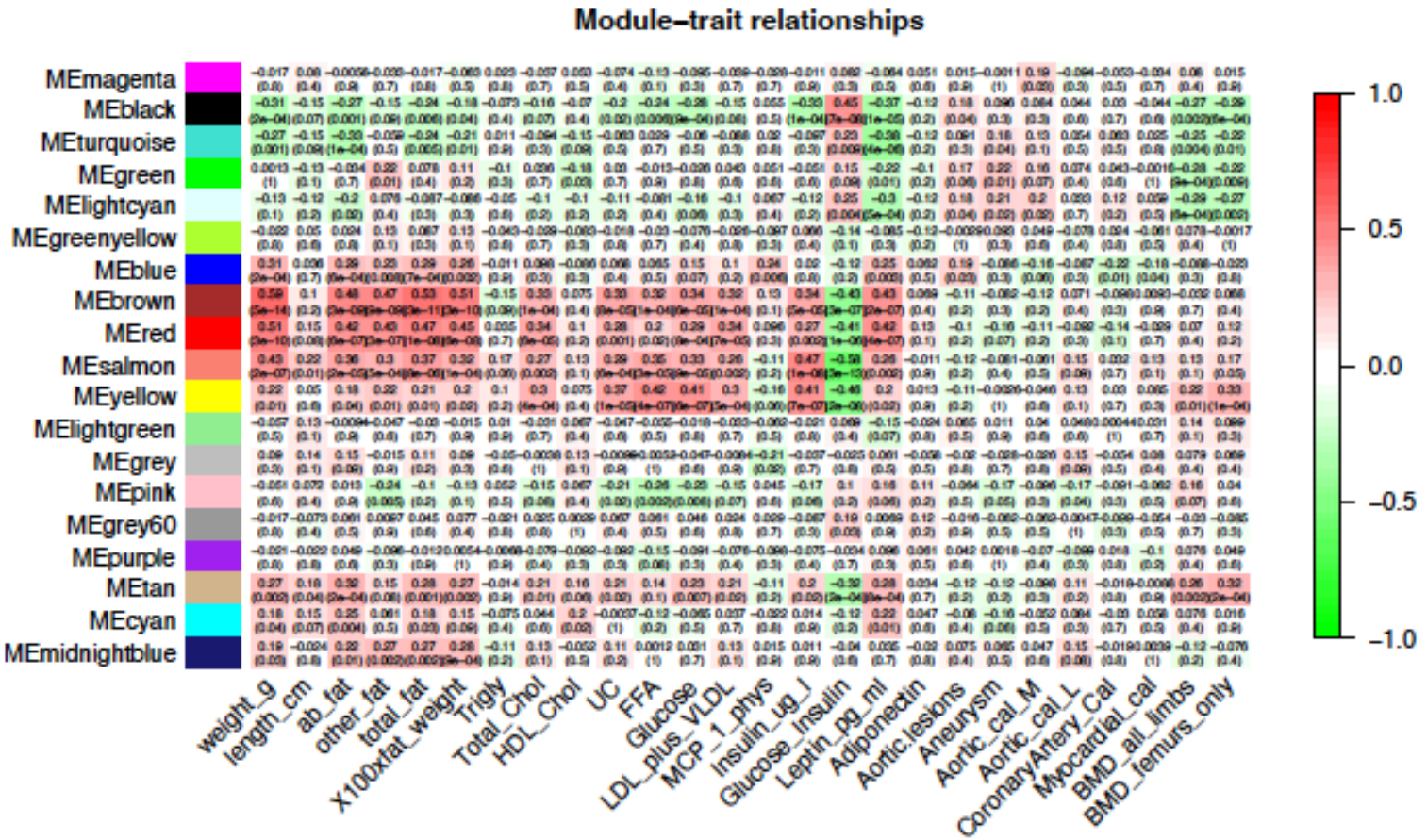
Identify modules



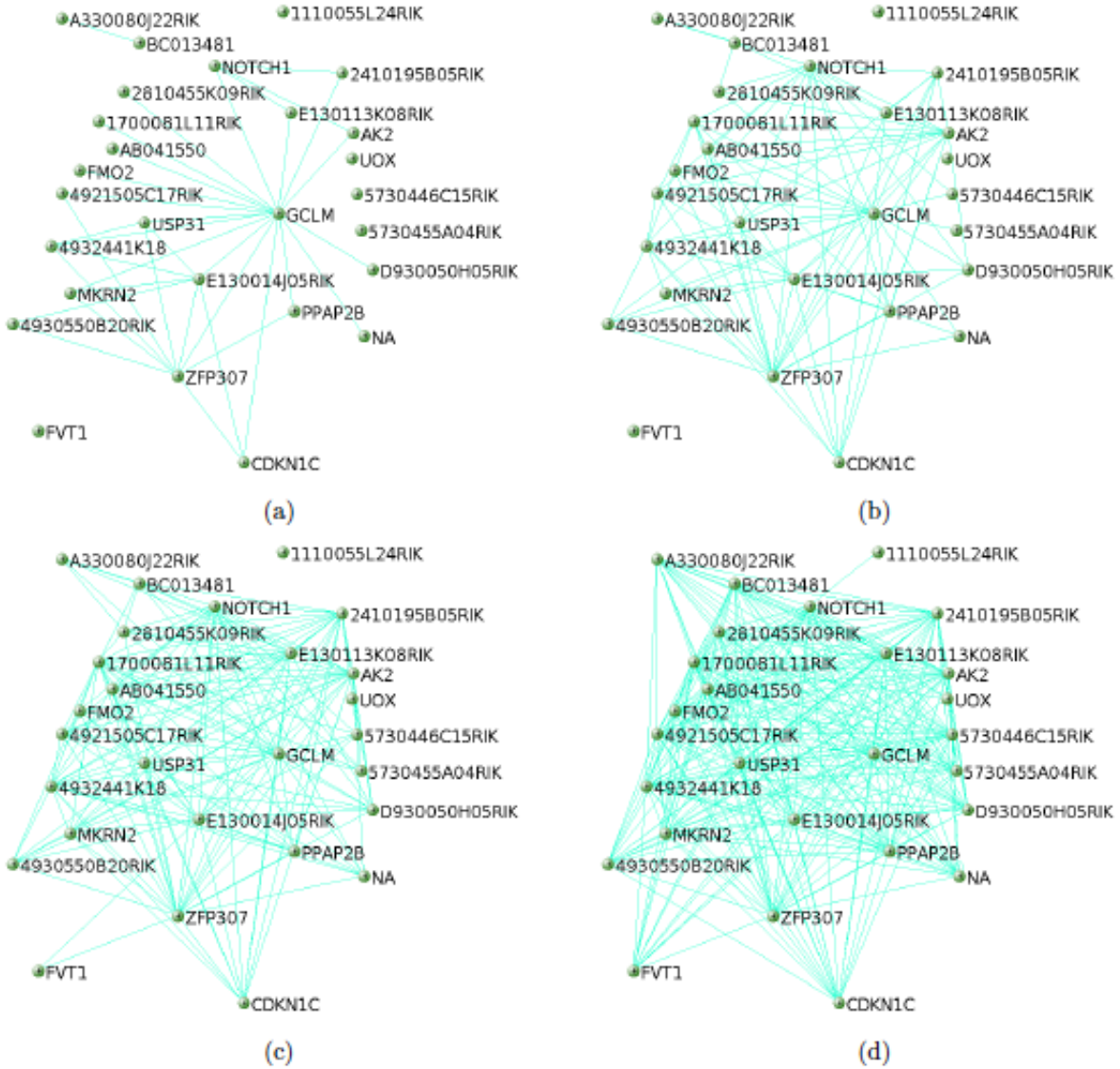
Module visualization



Relate modules to external information



Exporting a gene network



Q&A and Discussion

Contact

Dr. Edgardo Galán Vásquez
Departamento de Ingeniería de Sistemas Computacionales y Automatización
IIMAS
edgardo.galan@iimas.unam.mx
https://galanve.github.io/biominet/es/