STRING and Cytoscape for proteomics data analysis

Nadezhda T. Doncheva

nadezhda.doncheva@cpr.ku.dk

Integrative Biological Interpretation using Proteomics 2022

Instituto Gulbenkian de Ciência, Oeiras, Portugal

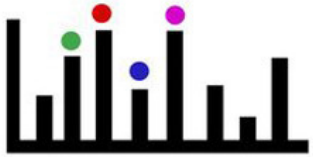
June 20 – June 24, 2022

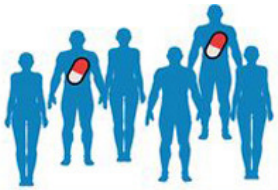
Intended learning outcomes

- Describe types of biological networks and give at least one example for such a network and where to find it
- Characterize a network in terms of what it represents
- Perform functional enrichment on a list of genes
- Analyze (your) high-throughput data using Cytoscape
  - Import your data into Cytoscape using the stringApp
  - Master network layouts and data visualization
  - Perform clustering and enrichment analyses
- Know where to find relevant documentation and tutorials

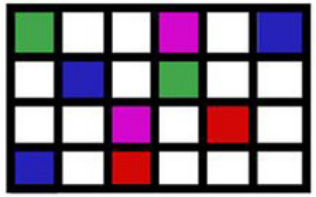
High-throughput technologies

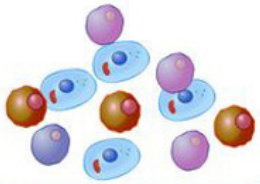
Proteomics



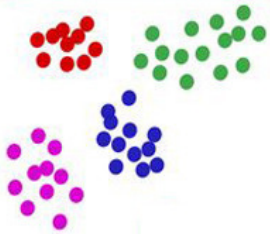


Microarray

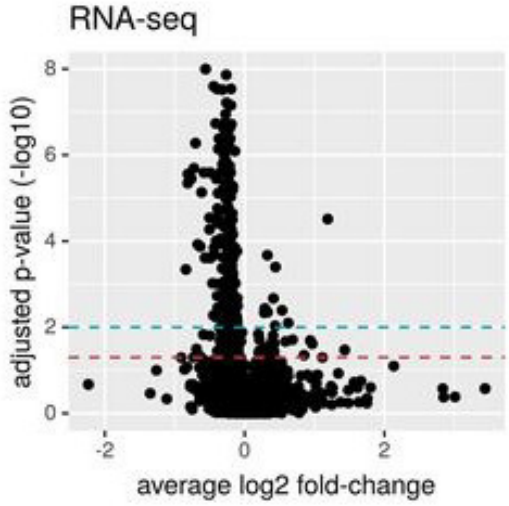




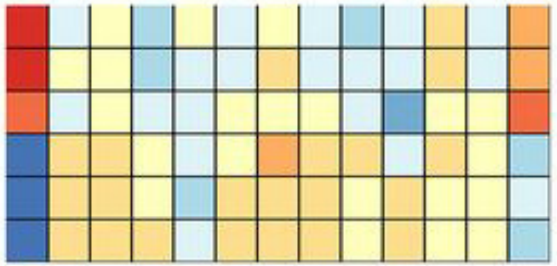
scRNA-seq



RNA-seq



Volcano plot



Heatmap

Adapted from Griss *et al.*, *Mol & Cell Prot,* 2020.

A typical table with data

- Temporal analysis by mass spectrometry of the proteome of neuroblastoma cells in response to nerve growth factor (NGF)
- Identification of 78 proteins that interact with tropomyosin-related kinase A (TrkA) upon NGF stimulation

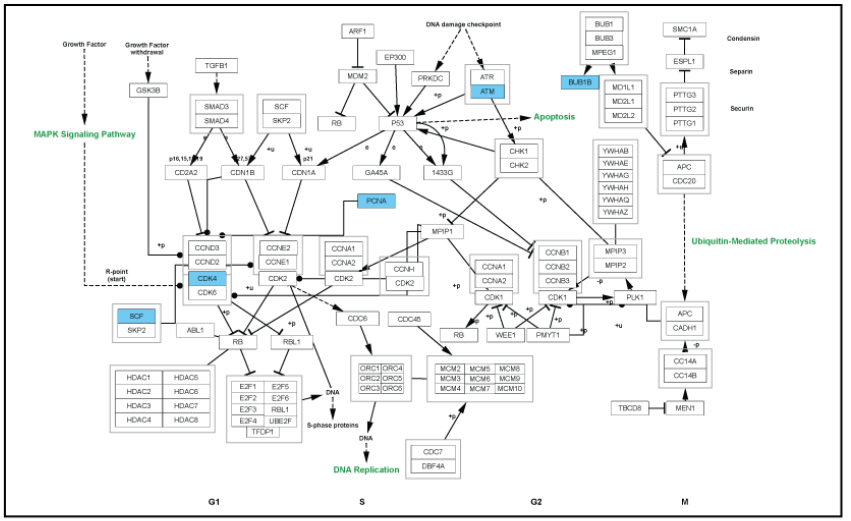
	A	B	C	D	G	J
1	**UniProt**	**Gene name**	**Peptides**	**Sequence coverage [%]**	**5 min log ratio**	**10 min log ratio**
2	Q99880	HIST1H2BL	5	35.7	-2.66	-2.66
3	Q8TER5	ARHGEF40	34	28.3	1.95	1.56
4	Q8IZ07	ANKRD13A	12	19.2	1.07	1.08
5	P62805	HIST1H4A	11	57.3	-2.31	-1.39
6	Q08380	LGALS3BP	14	28.2	-3.16	-2.98
7	O00750	PIK3C2B	35	24.2	2.21	2.31
8	O00443	PIK3C2A	29	17.8	1.13	1.26
9	Q9UJ41	RABGEF1	6	6.5	0.67	1.08
10	Q8TC07	TBC1D15	12	19.1	0.43	1.06

Emdal *et al.*, *Science Signaling,* 2015.

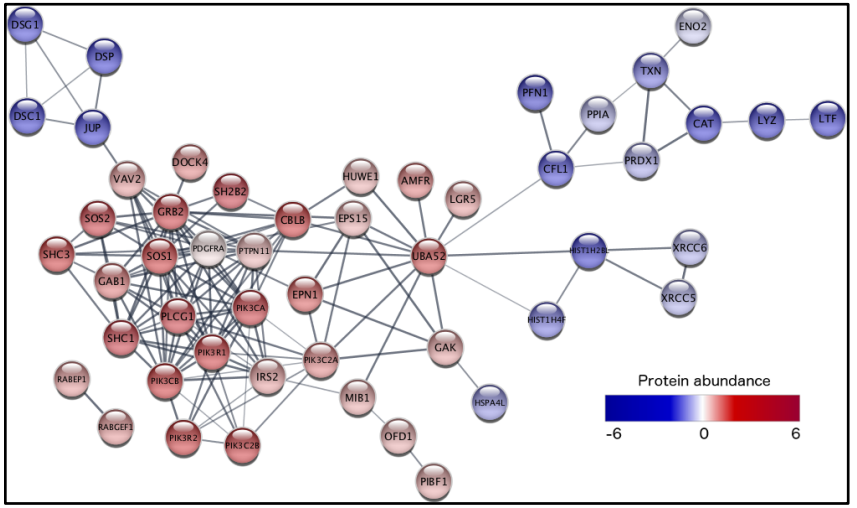
From gene lists to networks

UniProt
Q99880
Q8TER5
Q8IZ07
P62805
Q08380
O00750
O00443
Q9UJ41
Q8TC07

Identify relevant pathways & processes, e.g. KEGG & GO



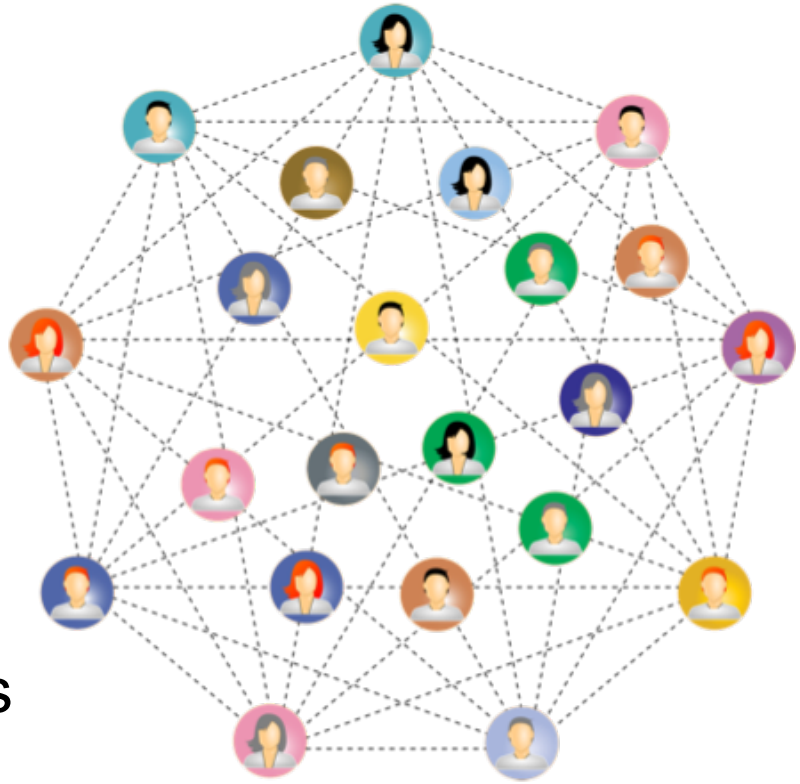
Construct networks and visualize data, e.g. STRING



What are networks?

- Consist of nodes (vertices, circles) and edges (links, lines)
- Represent relationships between entities

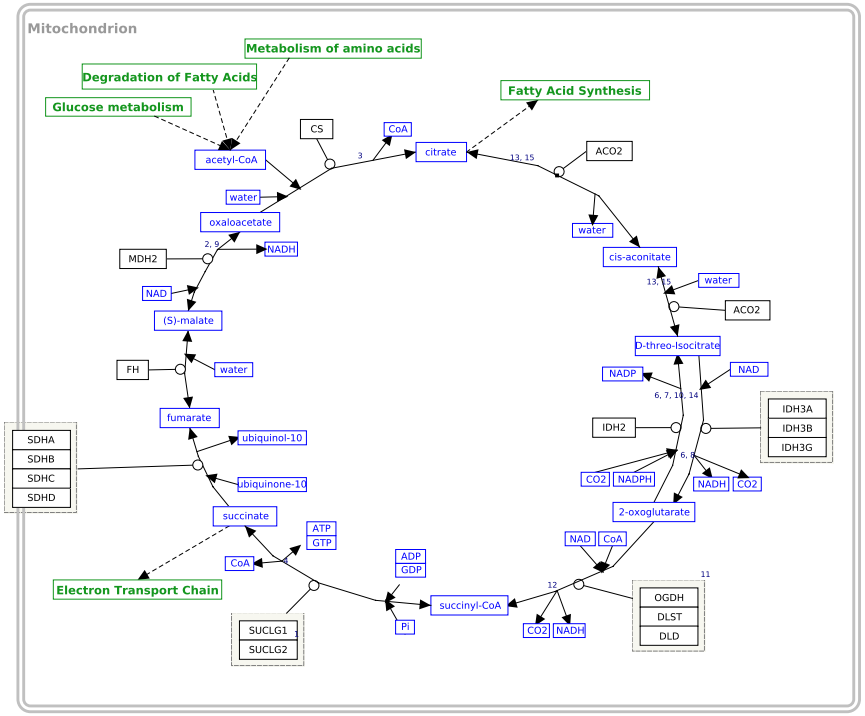
- Networks are everywhere…
  - Social networks
  - The power grid
  - The internet



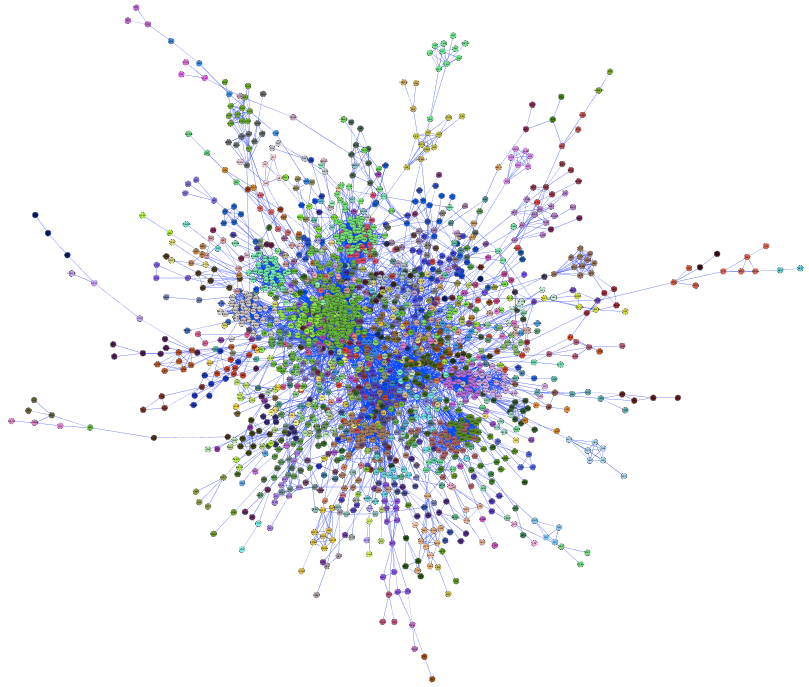
- Can you think of 1 or 2 examples for networks in your field?

Types of biological networks

- **Pathways**: metabolic, signaling, regulatory, etc.
- For example KEGG



- **Interaction networks**: protein-protein, protein-drug, etc.
- For example STRING



Sources of biological networks

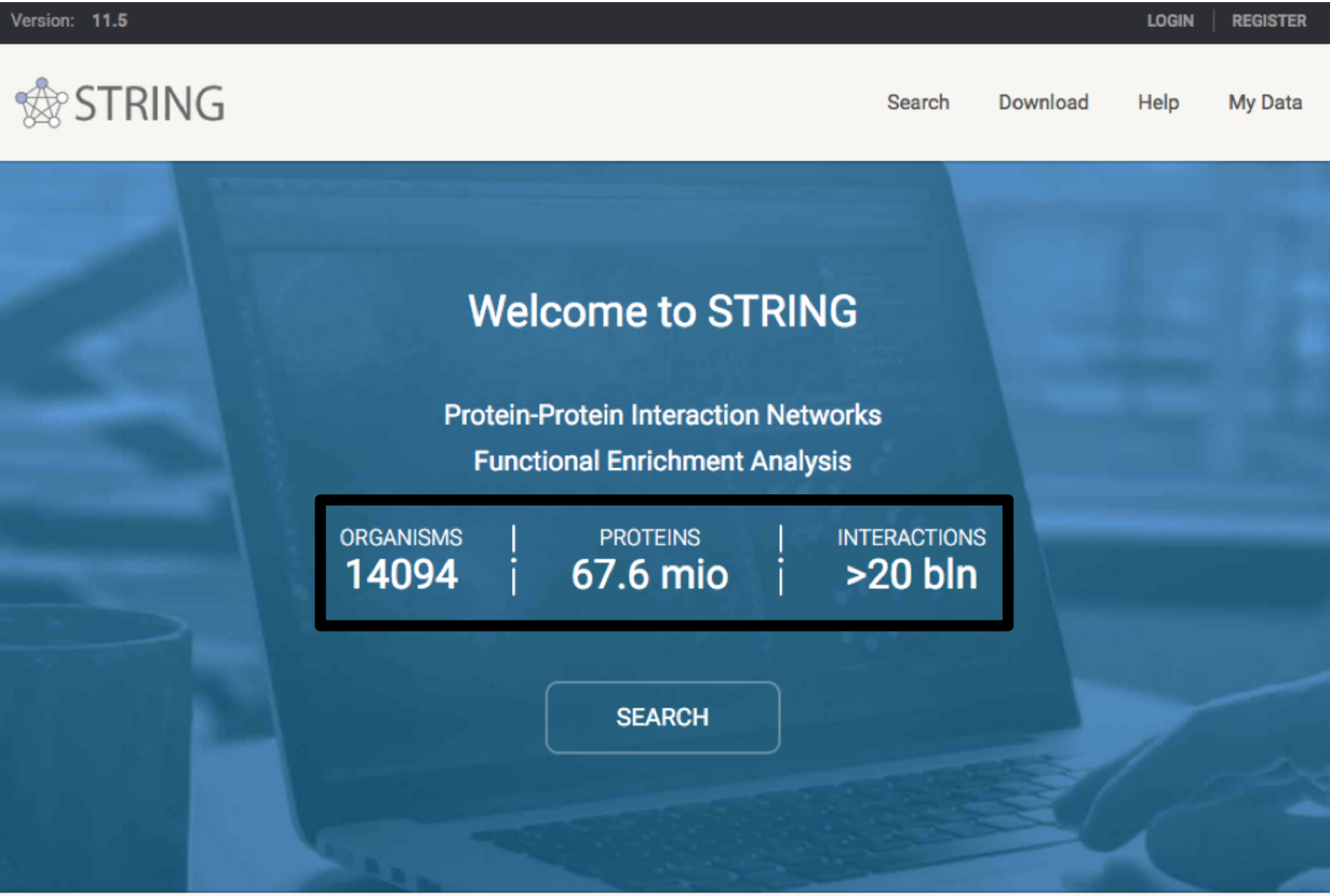
- There are hundreds of different interaction databases

→ It depends on your biological question and analysis plan

But know what you are getting...

- **Interaction networks**: broad coverage / lower resolution
  - STRING (https://string-db.org/)
  - IntAct (https://www.ebi.ac.uk/intact/)
  - BioGrid (https://thebiogrid.org/)

- **Pathways**: higher resolution / limited coverage
  - KEGG (https://www.kegg.jp/kegg/pathway.html)
  - Reactome (https://reactome.org/)
  - WikiPathways (https://www.wikipathways.org/)


Version: 11.5
LOGIN
REGISTER
STRING
Search
Download
Help
My Data
Welcome to STRING
Protein-Protein Interaction Networks
Functional Enrichment Analysis
ORGANISMS
14094
PROTEINS
67.6 mio
INTERACTIONS
>20 bln
SEARCH

STRING exercise 1 (15 min)

https://jensenlab.org/training/string/eubic/

In this exercise, we use the STRING database to create a network for one protein of interest and explore the different visual representation and the supporting evidence of the interactions.

Exercise 1.1: Single protein query

Question 1: *Why are there multiple lines connecting the same two proteins?*

Exercise 1.2: Visual representations

Question 2: *Which information is shown for the edges in each representation?*

Exercise 1.3: Evidence viewers

Question 3: *Which types of evidence support the interaction between insulin receptor (INSR) and insulin receptor substrate 1 (IRS1)?*

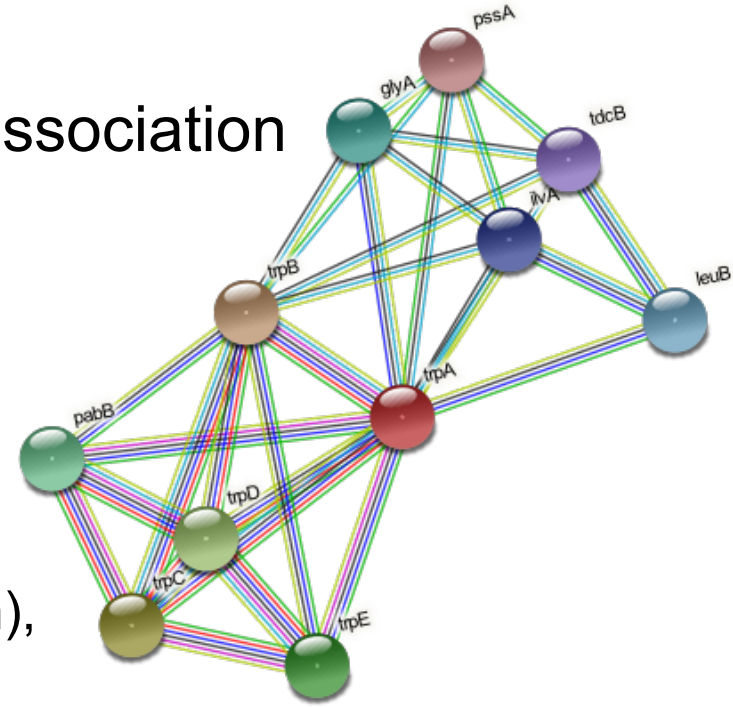
Question 4: *Which type of evidence gives the largest contribution to the confidence score 0.999?*

Question 5: *Which types of experiments support this interaction?*

STRING

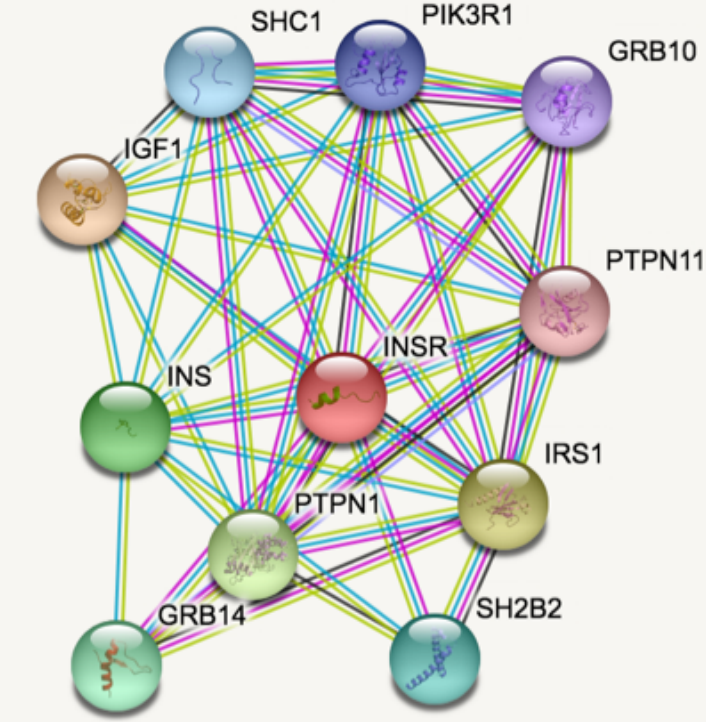
→ Integration of known and predicted functional protein associations

- Collect multiple types of evidence for known associations
- Predict new associations
- Transfer across species
- Assign confidence score to each association



Joint collaboration between the groups of Christian von Mering (University of Zurich), Lars Juhl Jensen (University of Copenhagen), and Peer Bork (EMBL Heidelberg)

STRING evidence channels



Known Interactions

- *from curated databases*
- *experimentally determined*

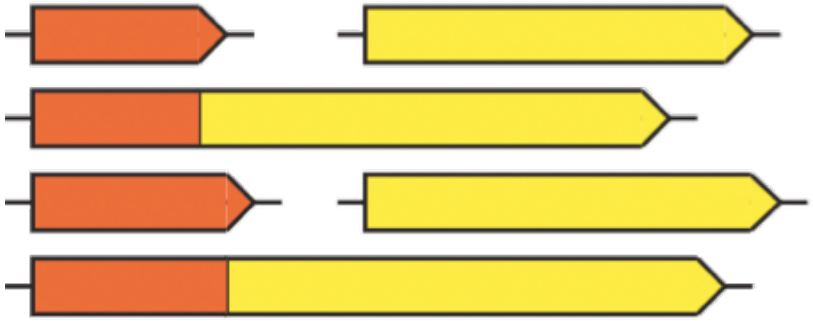
Predicted Interactions

- *gene neighborhood*
- *gene fusions*
- *gene co-occurrence*

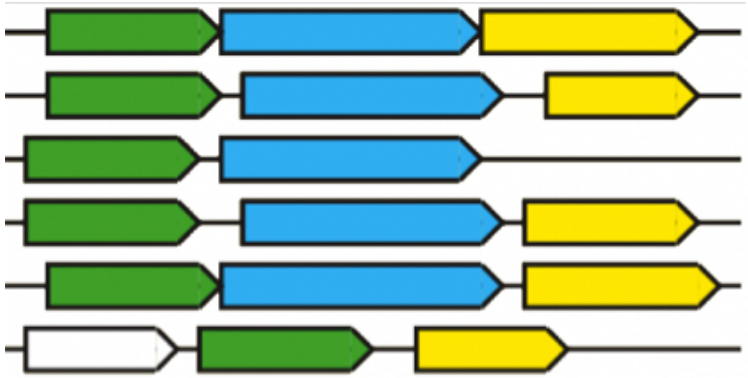
Others

- *textmining*
- *co-expression*
- *protein homology*

Predictions from genomic context

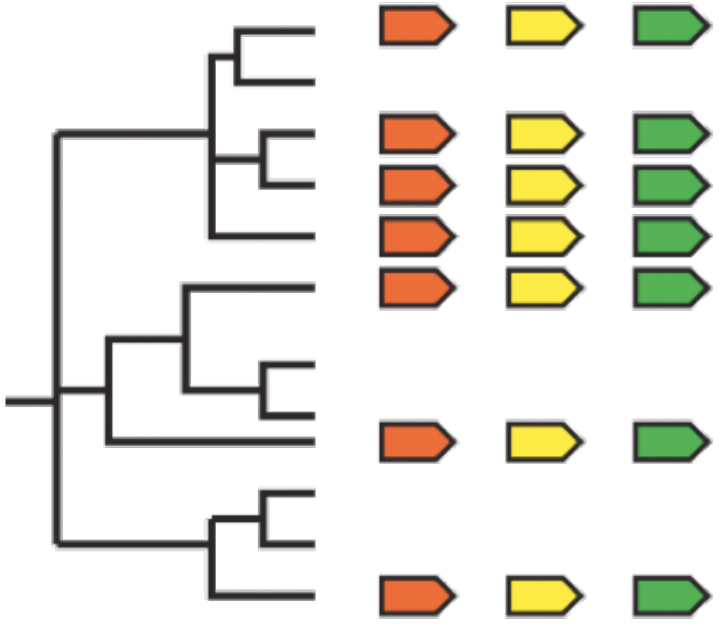


Gene fusion



Gene neighborhood

14,094 organisms

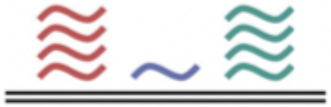


Gene co-occurrence
(phylogenetic profiles)

Korbel et al., *Nature Biotechnology*, 2004.

Experimental evidence

Experiments



Co-expression

Pair-wise interactions from experiments in curated databases like IntAct & BioGrid

Can be any type of *biochemical, biophysical or genetic interaction, like pull-down experiments*

Look for consistent similarities between expression profiles in many different conditions

Mainly RNA-based expression data + a bit of protein expression

“Higher-level” knowledge

Also known as *Databases*

Curated pathway databases like KEGG & Reactome

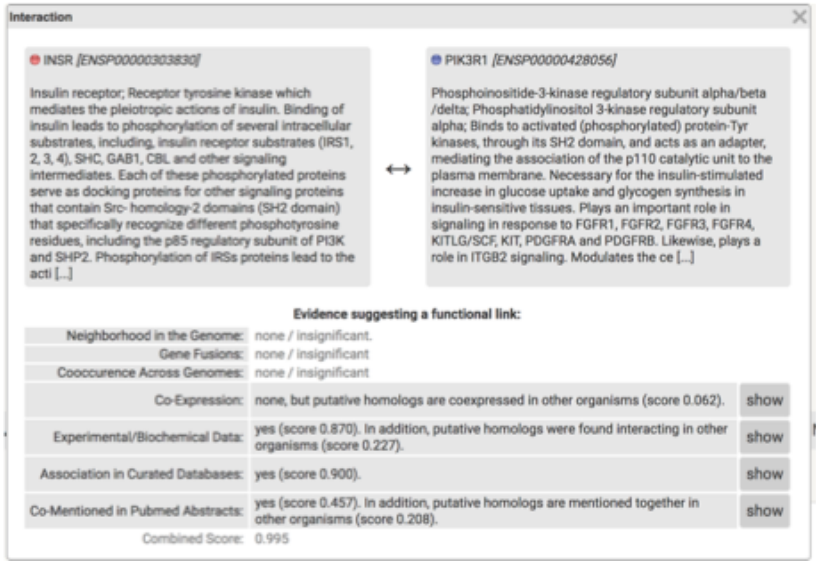
Co-occurrence text mining for functional associations

Natural language processing for physical interactions

STRING scores

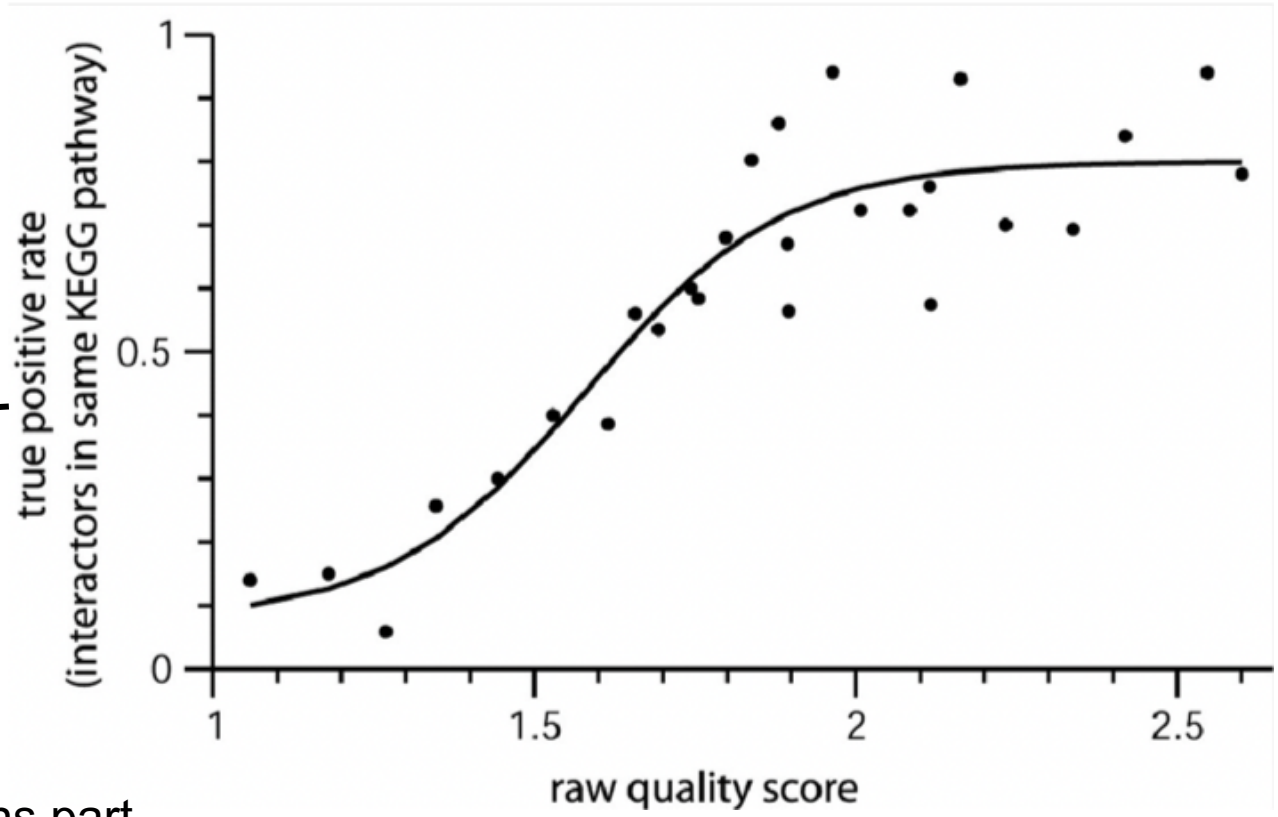
- However, it is not that simple:
  - Many databases
  - Different formats
  - Different identifiers
  - Variable quality
  - Not comparable
  - Not the same species

→ Quality scores [0,1] based on a gold standard
  - Common scale for comparison
  - Implicit weighting by quality



https://string-db.org/help//faq/#how-are-the-scores-computed

STRING score calibration

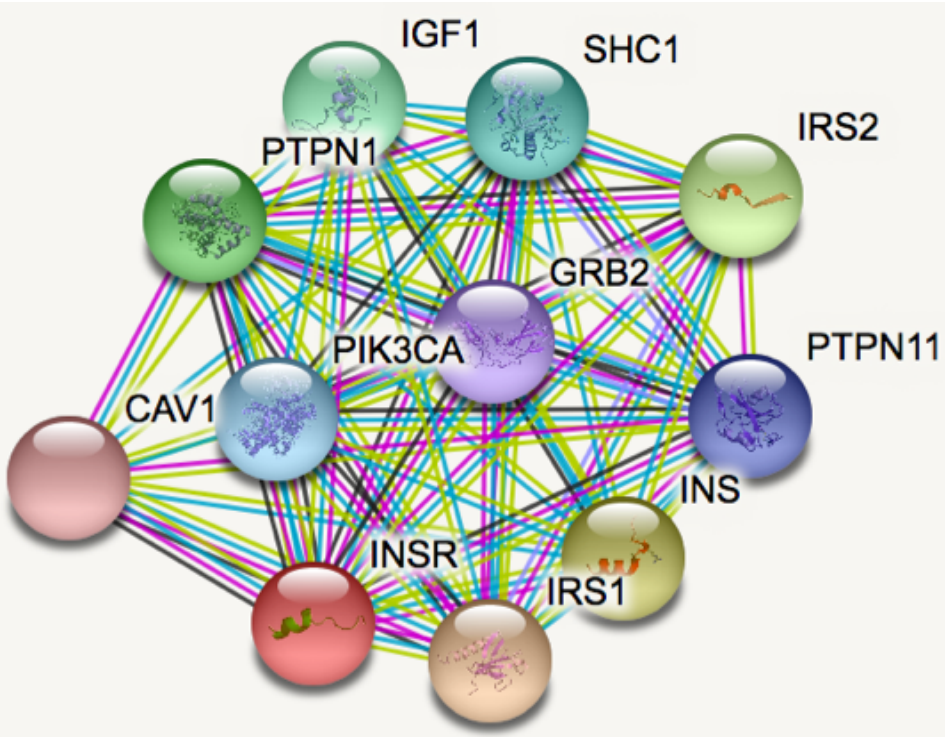


STRING confidence score
(how likely are two proteins part of the same KEGG pathway)

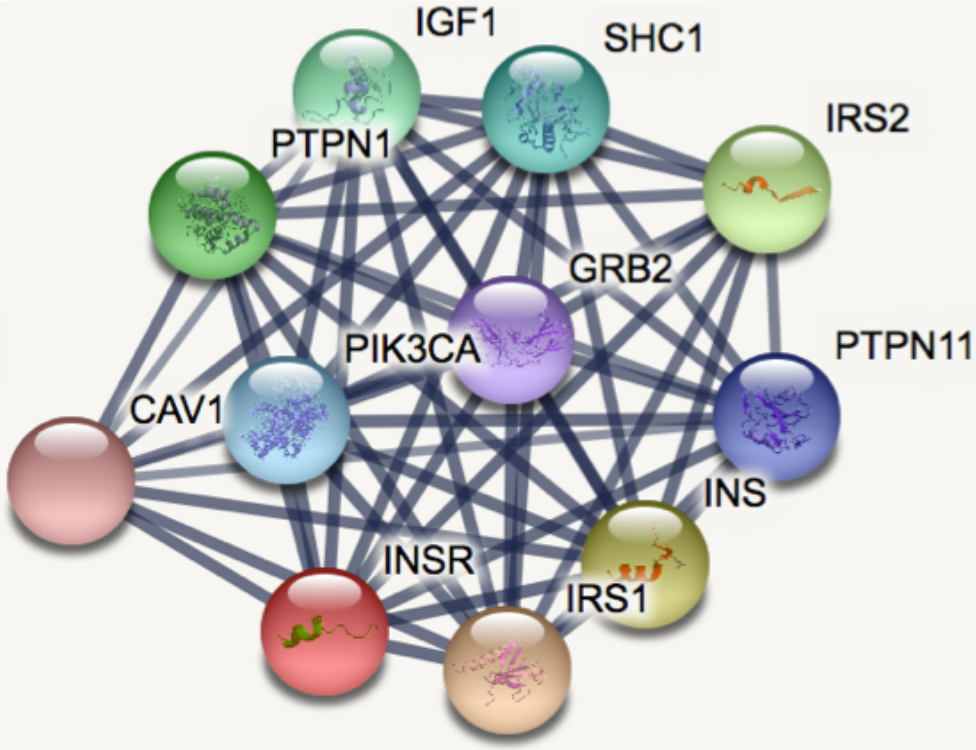
von Mering et al. *Nucleic Acids Research*, 2005.

STRING network views

What is the difference between these two views?



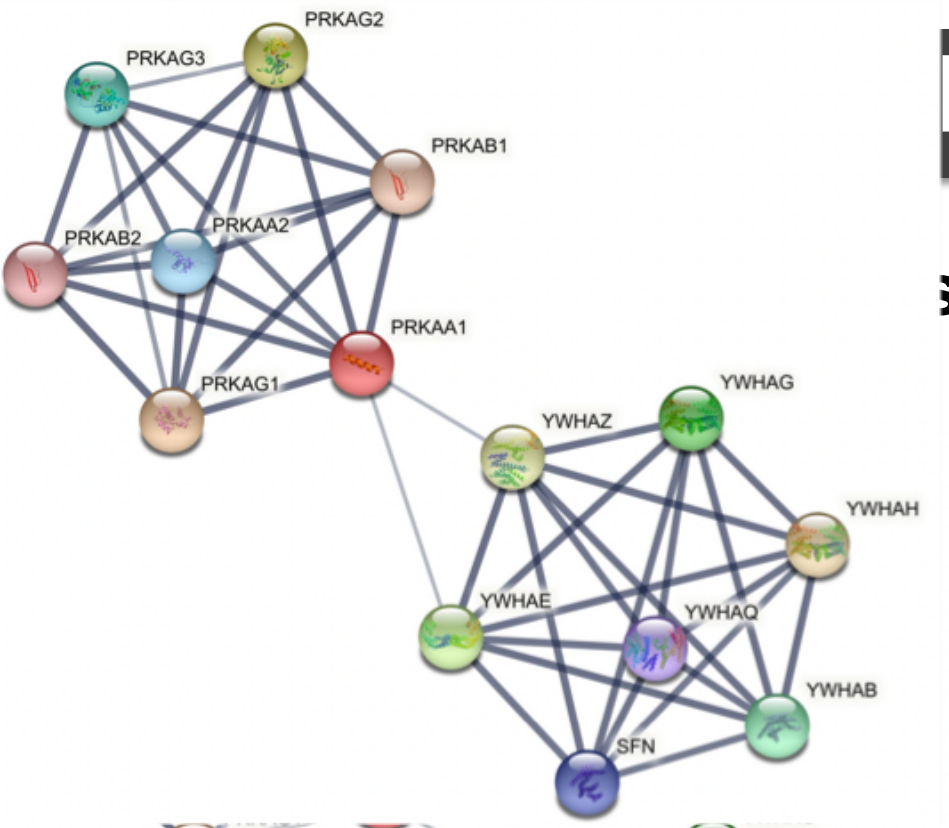
Evidence view



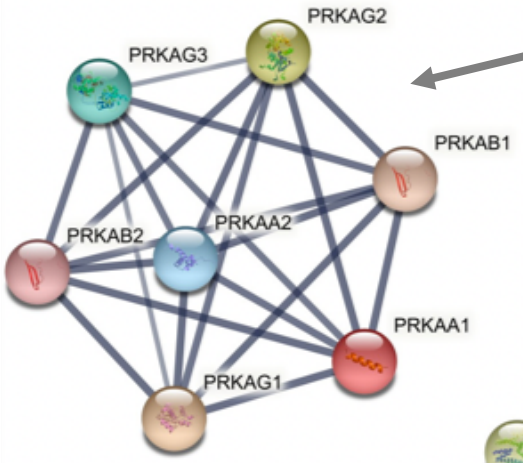
Confidence view

NG network types

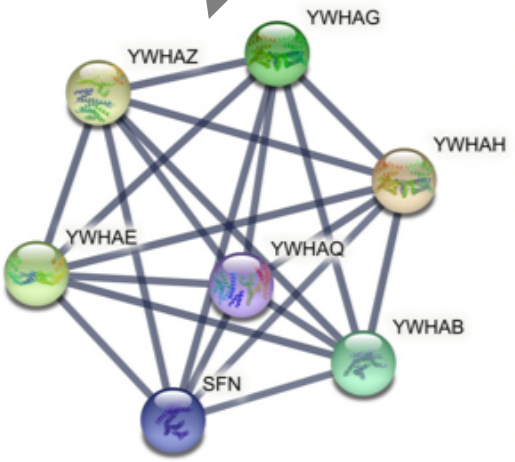
s vs. physical interactions



Indirect functional association



Physical complex formation



New since STRING v11.5

STRING exercise 2 (20 min)

https://jensenlab.org/training/string/eubic/

In this exercise, we will work with a list of proteins associated with epithelial ovarian cancer (EOC) in the study by Francavilla *et al.* and learn how to query STRING for multiple proteins, change the query parameters and do functional enrichment.

Exercise 2.1: Multiple proteins query

Question 1: *How many nodes and edges are in the resulting network?*

Exercise 2.2: Query parameters

Question 2: *How does changing the confidence or network type influence the set of interactions shown?*

Question 3: *What evidence types are available for the physical interactions? Are there more or fewer evidence types than in the full STRING network?*

Exercise 2.3: Functional enrichment

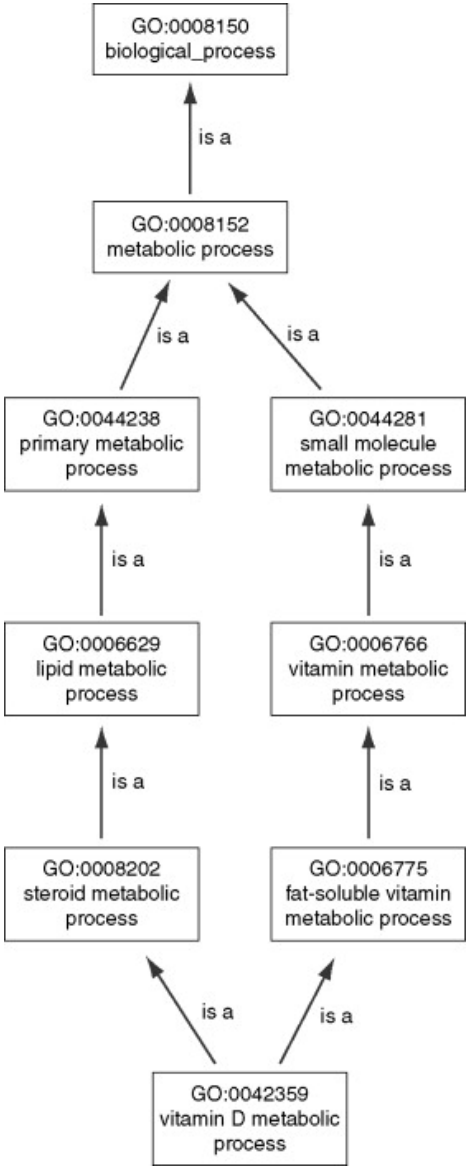
Question 4: *How many categories contain enriched annotation terms?*

Question 5: *What information is shown in each line of the table? How can you find out more about this annotation term?*

Question 6: *Do the pathways annotate the same set of genes or not?*

Functional enrichment analysis

- *aka* over-representation analysis
- A *term* is usually a *pre-defined group* of genes, with the same function or the same process, e.g. a pathway like *TCA cycle* or a Gene Ontology term like *mRNA processing*
- Identify terms that are statistically over-represented for a set of *regulated* genes compared to a background set of genes
- Fisher's exact test followed by multiple testing correction
- Choosing the *right* background is very important: genome-wide vs. user-defined



Gene ontology example

Enrichment in STRING

Viewers | Legend | Settings | Analysis | Exports | Clusters | More | Less

Network Stats

number of nodes:	11	expected number of edges:	14
number of edges:	48	PPI enrichment p-value:	1.05e-12
average node degree:	8.73		
avg. local clustering coefficient:	0.917		

your network has significantly more interactions than expected (what does that mean?)

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

explain columns

Biological Process (Gene Ontology)

GO-term	*description*	*count in network*	*strength*	*false discovery rate*
GO:1902202	regulation of hepatocyte growth factor receptor signaling pa...	2 of 4	2.95	3.45e-05
GO:0060267	positive regulation of respiratory burst	2 of 6	2.77	5.83e-05
GO:0045725	positive regulation of glycogen biosynthetic process	5 of 17	2.72	4.59e-11
GO:1990535	neuron projection maintenance	2 of 8	2.65	8.78e-05
GO:0032000	positive regulation of fatty acid beta-oxidation	2 of 9	2.6	0.00010

(more ...)

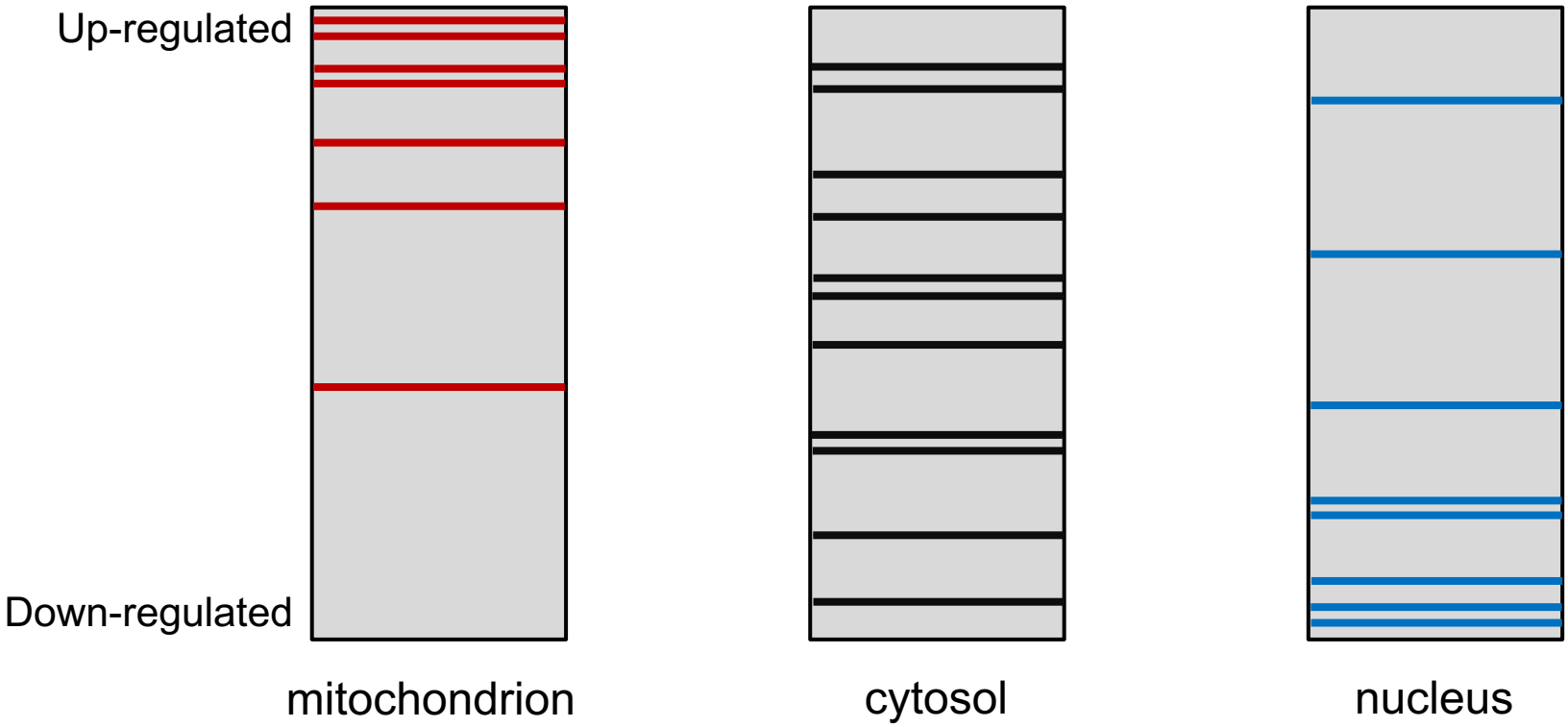
Molecular Function (Gene Ontology)

GO-term	*description*	*count in network*	*strength*	*false discovery rate*
GO:0043559	insulin binding	2 of 5	2.85	6.17e-05
GO:0005159	insulin-like growth factor receptor binding	5 of 16	2.74	2.40e-11
GO:0005158	insulin receptor binding	7 of 23	2.73	3.77e-16
GO:0043560	insulin receptor substrate binding	3 of 11	2.69	7.36e-07
GO:0031994	insulin-like growth factor I binding	2 of 9	2.6	0.00013

(more ...)

Gene set enrichment analysis

- *aka* GSEA is performed on a ranked list of all genes
- Kolmogorov-Smirnov test to identify which terms show a non-random distribution across the sorted gene list, followed by multiple testing correction



GSEA in STRING

Your input data

1: PKP1	-8.326649949152102
2: CDSN	-8.130157304186698
3: SERPINB5	-8.065760365992743
4: DSC1	-7.917077464751732
5: DSG1	-7.838328194641223
6: CALML5	-7.706114452582677
7: ZNF750	-7.5277671530949535
8: SERPINB7	-7.4978374812276632
9: LCE2B	-7.467221629957017 5
10: CHP2	-7.423878301199893
11: GJB6	-7.301189455146853
12: COL17A1	-7.2636604397081825
13: C19orf33	-7.1952071849953185
14: SBSN	-7.1404580970491 76
15: LY6D	-7.056120251292827
16: TRIM29	-7.034785864374081
17: FLG	-7.031575998772657
18: CRCT1	-7.0226906177601025
19: KRT15	-6.8670255485207 02
20: SPRR1A	-6.859561525754514
21: LOR	-6.848695263892816
22: CLCA2	-6.767725587791244
23: SLURP1	-6.7673021587277775
24: C1orf68	-6.695574595281212 5
25: LGALS7	-6.613240474340857 5
26: CST6	-6.585766047436771
27: LYPD3	-6.573128209505429 5
28: DMKN	-6.486748209061480 5
29: LCE1B	-6.460586585775241
30: WFDC5	-6.441728770048803
31: SPRR2G	-6.419209327245714 5
32: CNFN	-6.384859699255222

Your detected functional enrichments

Biological Process (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0070268	cornification	107	1.22e-13
GO:0031424	keratinization	180	1.73e-08
GO:0061436	establishment of skin barrier	19	7.48e-07
GO:0033561	regulation of water loss via skin	21	4.07e-06
GO:0050891	multicellular organismal water homeostasis	62	0.00070

(more ...)

Reference publications

publication	*(year) title*	*count in gene set*	*false discovery rate*
PMID:23921950	(2014) Highly rapid and efficient conversion of human fibroblasts to keratinocyte-like cells.	50	9.48e-11
PMID:26644517	(2015) A keratin scaffold regulates epidermal barrier formation, mitochondrial lipid	67	1.75e-07
PMID:27408699	(2016) Recent advances in understanding ichthyosis pathogenesis.	23	1.16e-06
PMID:25695600	(2015) Structural and biochemical changes underlying a keratoderma-like phenotype in mice	53	1.16e-06
PMID:9892899	(1998) All-trans retinoic acid compromises desmosome expression in human epidermis.	6	0.00013

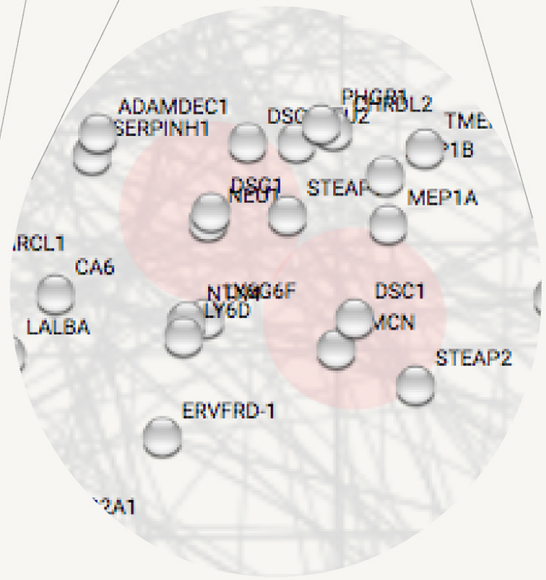
(more ...)

Cellular Component (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0001533	cornified envelope	51	9.48e-13
GO:0097209	epidermal lamellar body	4	2.65e-06
GO:0030056	hemidesmosome	7	2.65e-06
GO:0030057	desmosome	25	5.34e-05
GO:0097539	ciliary transition fiber	10	0.0407

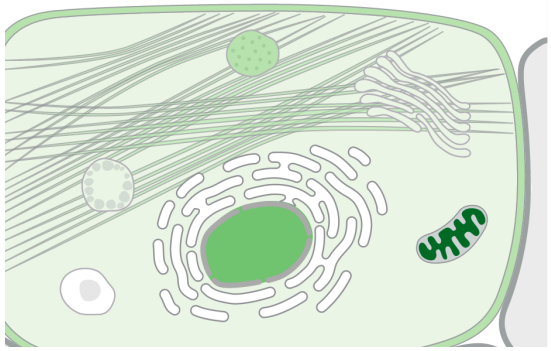
(more ...)

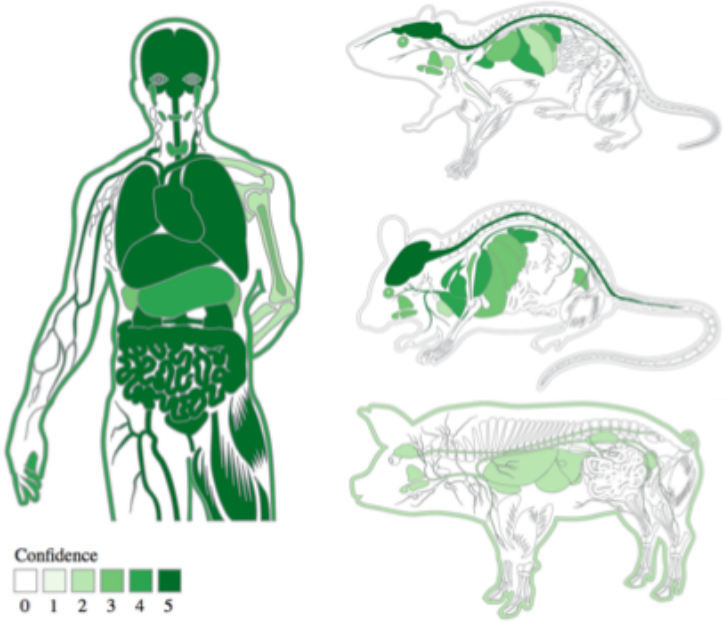
Full proteome network (Homo sapiens)



Related databases

- **COMPARTMENTS**: Subcellular localization database
- **TISSUES**: tissue expression database for human, mouse, rat and pig
- **DISEASES**: disease-gene associations mined from the literature
- All three provide confidence scores between *0 and 5 stars*

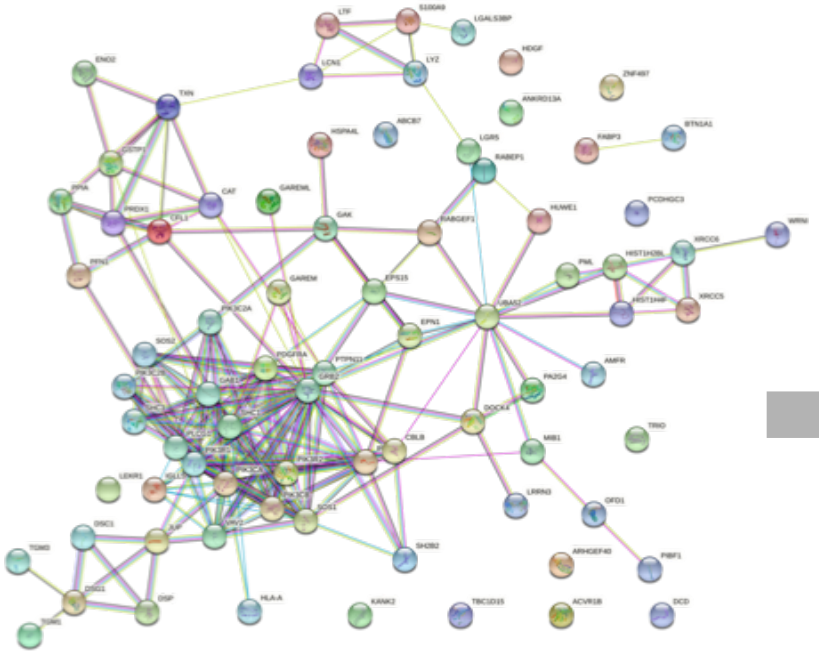


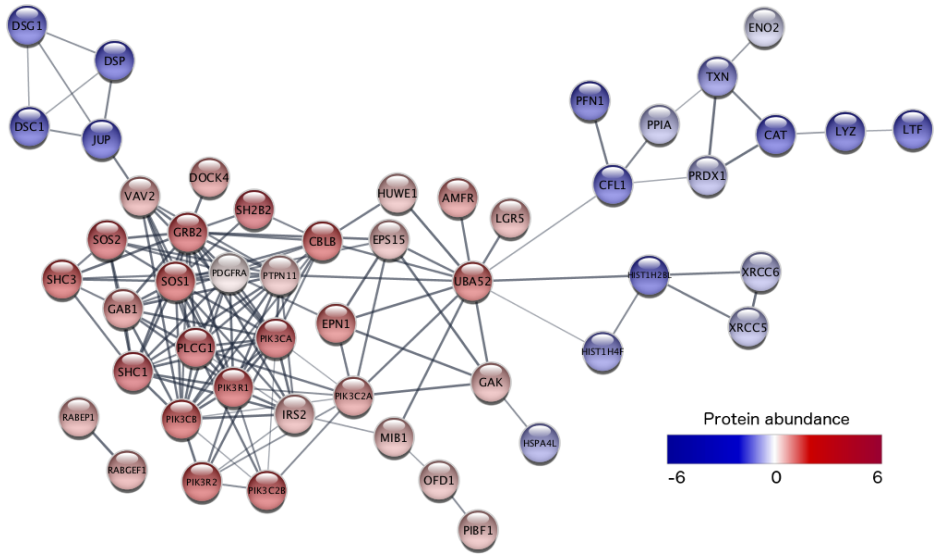


http://jensenlab.org/resources/

From STRING to Cytoscape

- Web-based network interfaces have limitations
  - Creating networks for large lists of genes
  - Integrating and showing additional experimental data
  - Having more powerful analysis and visualization options

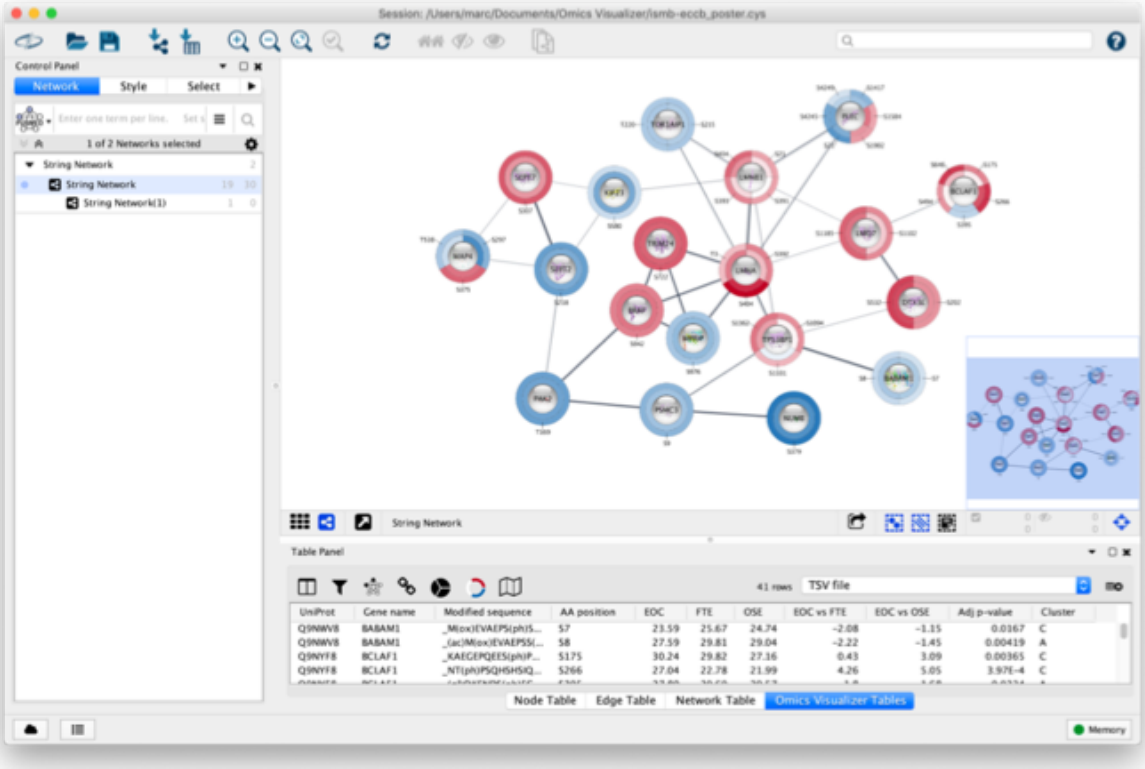




Cytoscape

- Open source tool for network analysis and visualization
- Large, active community of developers & users
- However, Cytoscape itself doesn't know any biology

→ **Cytoscape apps:** apps.cytoscape.org



All Apps

Categories

- collections
- data visualization
- network generation
- network analysis
- graph analysis
- online data import
- automation
- integrated analysis
- clustering
- systems biology
- utility
- enrichment analysis
- visualization
- data integration
- layout
- core app
- annotation
- ontology analysis
- pathway database
- network comparison

more »

Newest Releases

Get Started with the App Store »

BNMatch2 3.0+

BNMatch provides a reliable and optimized mapping between

DKernel 3.0+

DKernel uses Diffusion Kernel algorithm to propagate sub-

BioGateway Cytoscape Plugin 3.0+

An explorative network building plugin that works with the

IntAct App 3.0+

BETA: Build molecular interaction networks from IntAct database.

CyCommunityDetection 3.0+

Integrates multiscale community detection and functional

dot-app 3.0+

Import/export of Graphviz files in Cytoscape

more newest releases »

Top Downloaded Apps

ClueGO 3.0+

Creates and visualizes a functionally grouped network of

BiNGO 3.0+

Calculates overrepresented GO terms in the network and display

CluePedia 3.0+

CluePedia: A ClueGO plugin for pathway insights using integrated

GeneMANIA 3.0+

Imports interaction networks from public databases from a list of

stringApp 3.0+

Import and augment Cytoscape networks from STRING

MCODE 3.0+

Clusters a given network based on topology to find densely

more top downloads »

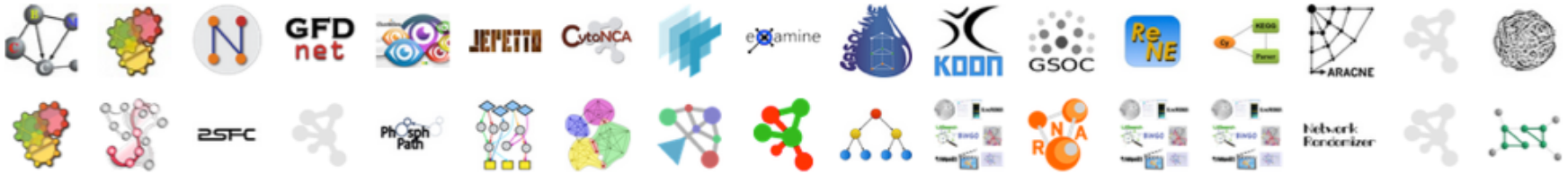
Wall of Apps 372 total

collections

data visualization

network generation

network analysis



stringApp

Import and augment Cytoscape networks from STRING

(22) 122414 downloads | citations | discussions

Details | Release History

Categories: annotation, automation, data visualization, disease, enrichment analysis, gene-disease association, gene function prediction, import, interaction database, network generation, online data import, PPI-network, visualization

stringApp imports functional associations or physical interactions between protein-protein and protein-chemical pairs from STRING, Viruses.STRING, STITCH, DISEASES and from PubMed text mining into Cytoscape. Users provide a list of one or more gene, protein, compound, disease, or PubMed queries, the species, the network type, and a confidence score and *stringApp* queries the database to return the matching network. Currently, four different queries are supported:

- STRING: protein query -- enter a list of protein names (e.g. gene symbols or UniProt identifiers/accession numbers) to obtain a STRING network for the proteins
- STRING: PubMed query -- enter a PubMed query and utilize text mining to get a STRING network for the top N proteins associated with the query
- STRING: disease query -- enter a disease name to retrieve a STRING network of the top N proteins associated with the specified disease
- STITCH: protein/compound query -- enter a list of protein or compound names to obtain a network for them from STITCH

CYTOSCAPE 3

Installed

Version 1.7.0

Released 13 Aug 2021

Works with Cytoscape 3.8

Download Stats Click here

RESOURCES

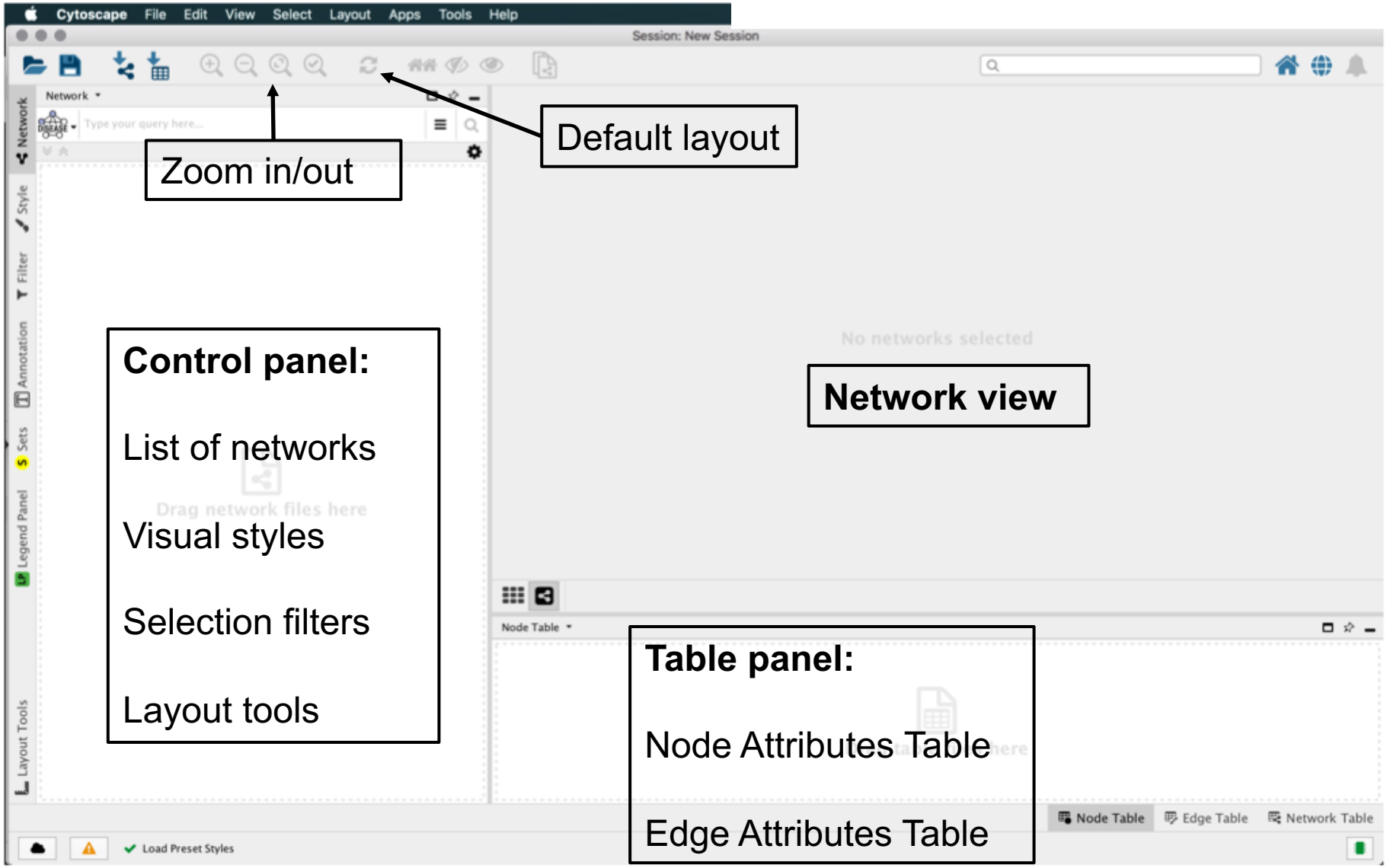
- Ask a question
- Search BioStars
- Website
- Tutorial
- Cite this App
- Code Repository
- Automation Support
- E-mail

Let's try it out!

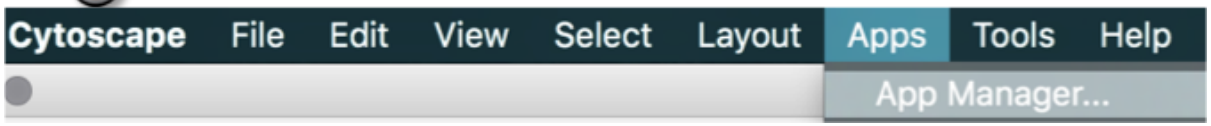
How many have installed Cytoscape **3.9.1**?

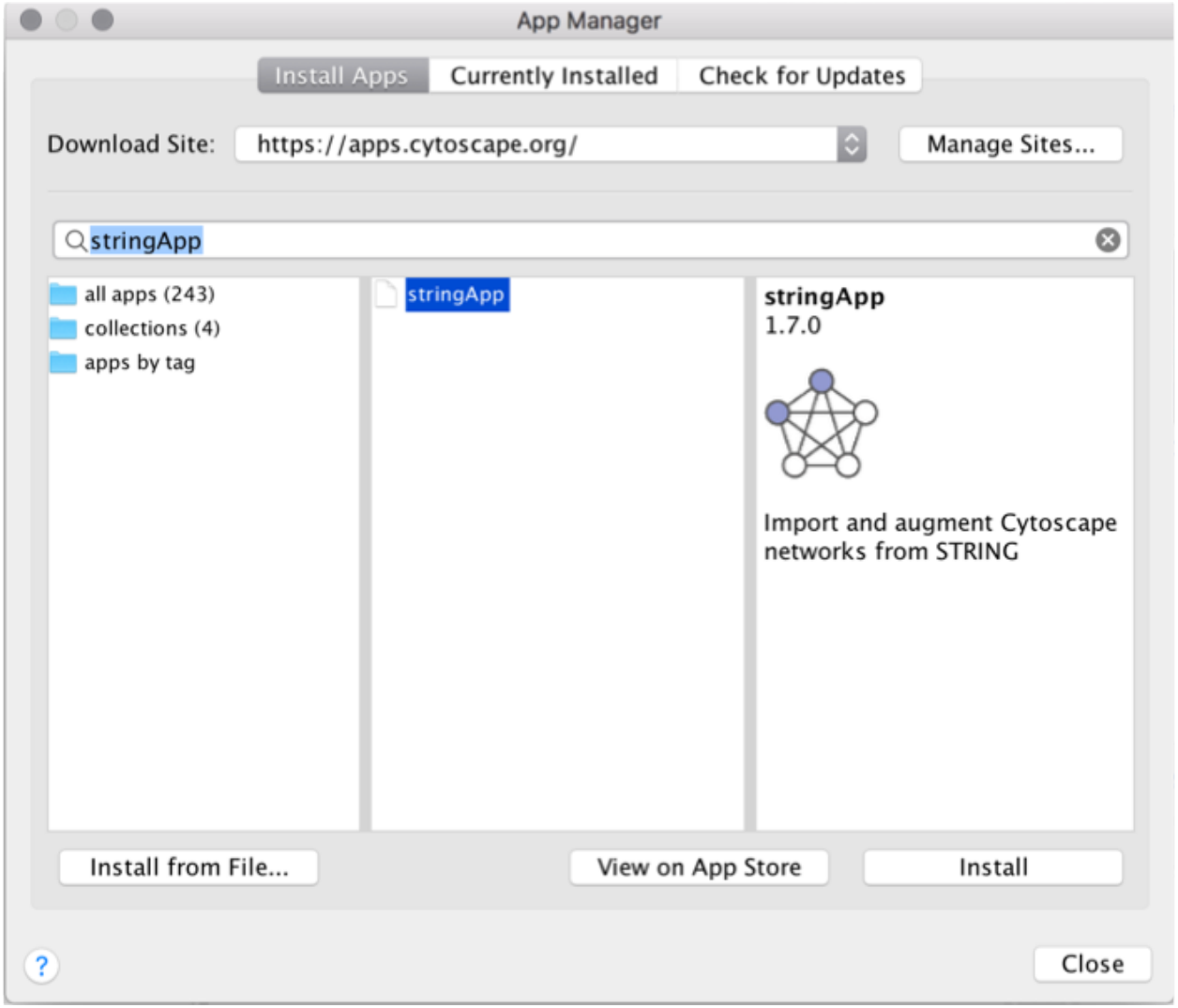
If not installed yet, get it from here:
http://cytoscape.org/download.php

Launch Cytoscape



Install stringApp v1.7.1





stringApp exercise 1 (10 min)

https://jensenlab.org/training/stringapp/eubic/

In this exercise, we will perform some simple queries to retrieve molecular networks in Cytoscape using the stringApp.

Exercise 1.1: Protein query

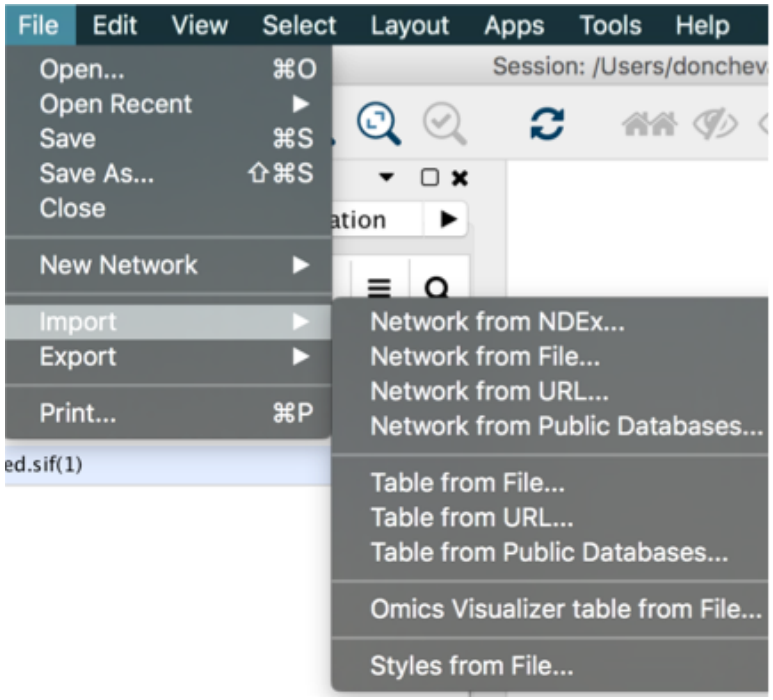
Question 1: *How many nodes are in the resulting network? What types of information do the* ***Node Table*** *and the* ***Edge Table*** *provide?*

Exercise 1.2: Disease query

Question 2: *Which additional attribute column do you get in the* ***Node Table*** *for a disease query compared to a protein query? Hint: check the last column.*

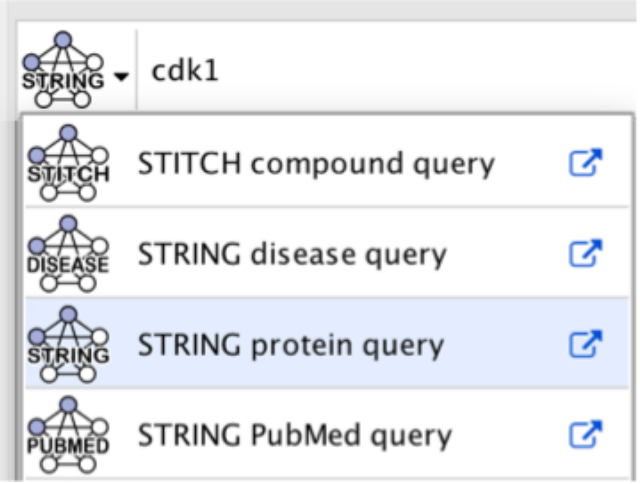
Import network (edge) data

- **Starting with a list of genes, no network data**
  - stringApp
  - GeneMANIA app
  - IntAct app
- **Pathway databases**
  - KEGGscape app
  - ReactomeFI app
  - WikiPathways app
- **Your own network data**
  - from files, e.g. Excel tables or text files
  - from R or Python via automation

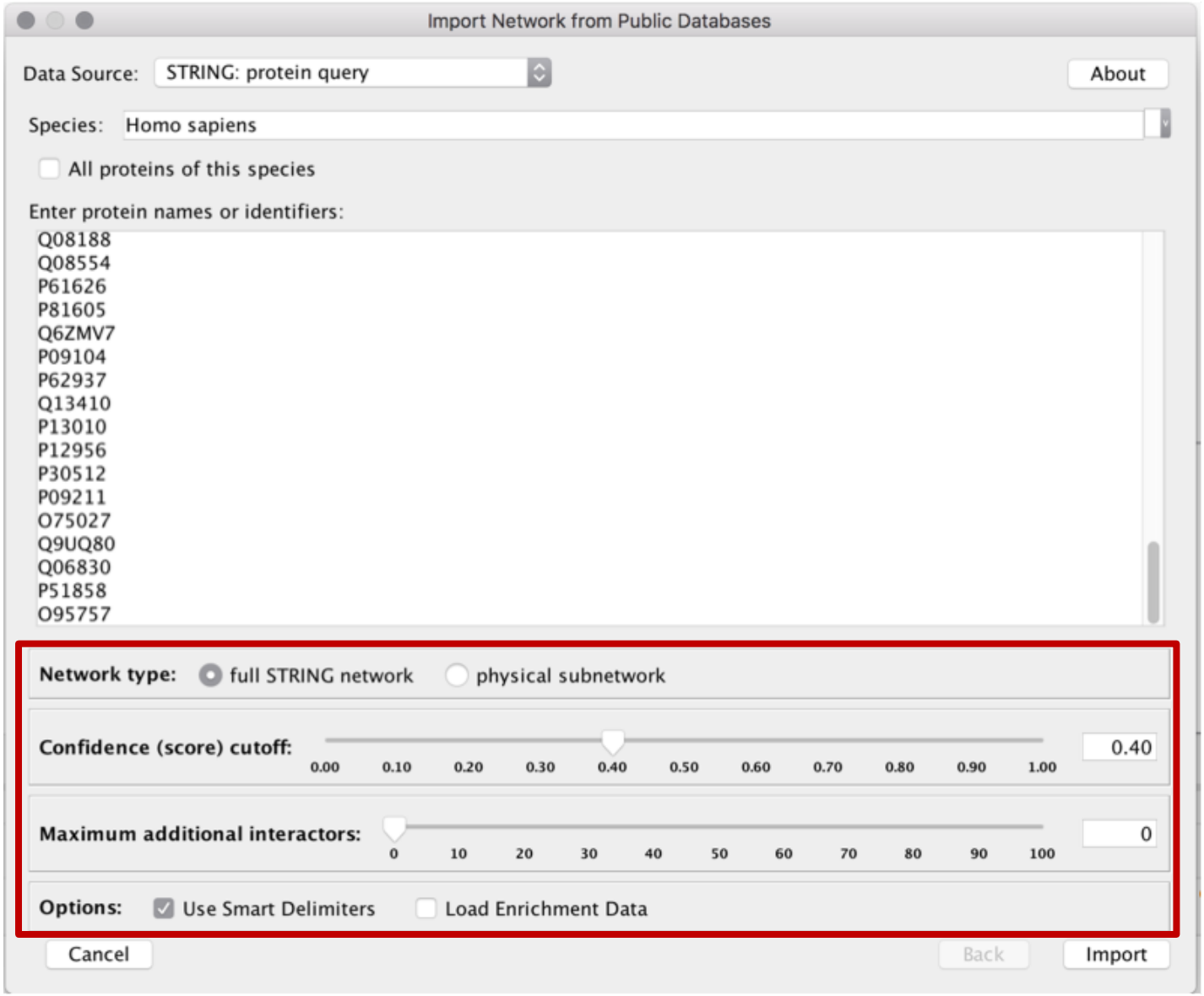


stringApp

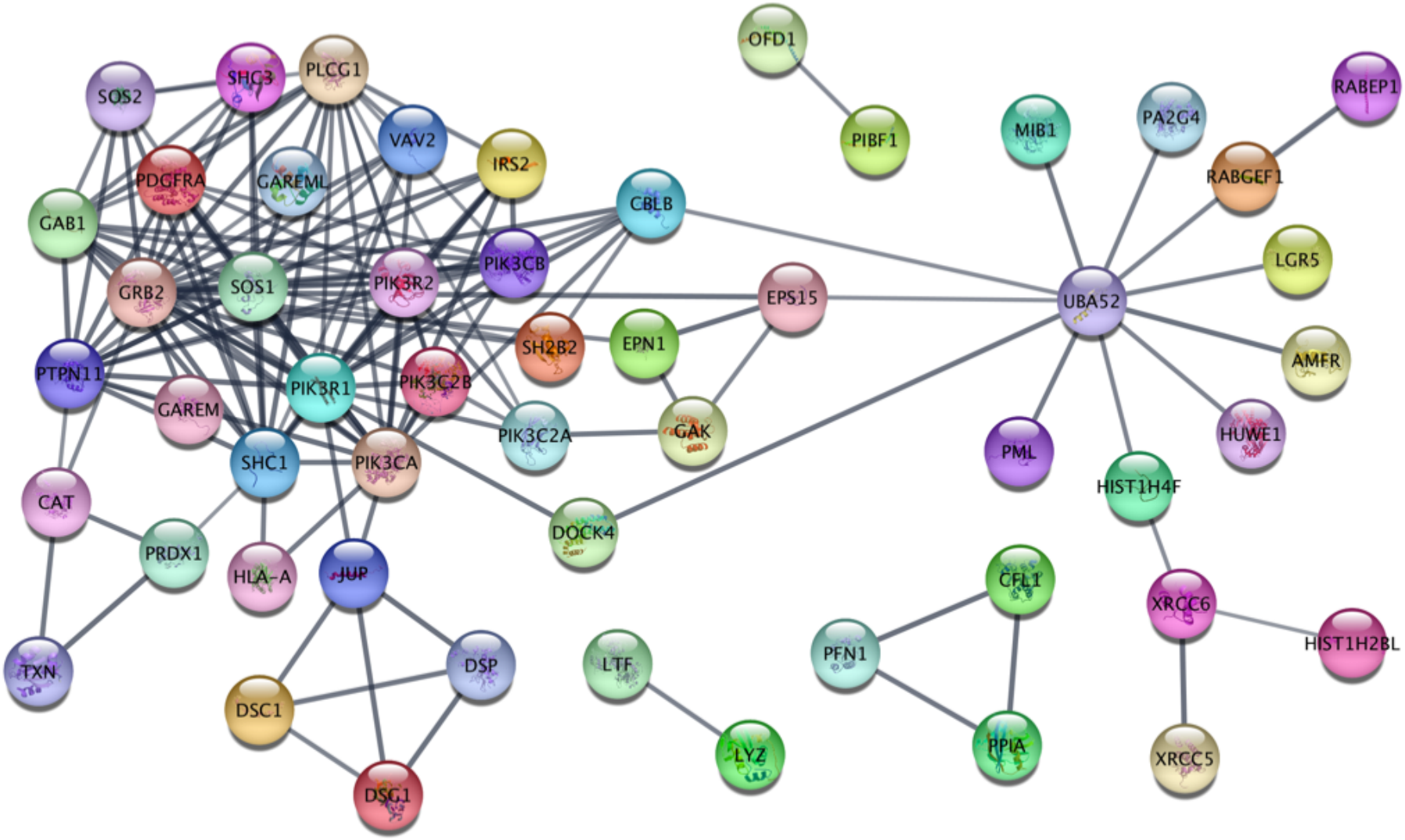
- STRING **protein** query
  - Queries for STRING interactions for **one** protein or for a **list** of identifiers
- STRING **compound** query
  - Queries for protein-compound interactions
- STRING **disease** query
  - Queries for disease-associated proteins from DISEASES and for STRING interactions between them
- STRING **PubMed** query
  - Retrieves STRING interactions for proteins co-occurring with the query term in PubMed



STRING protein query

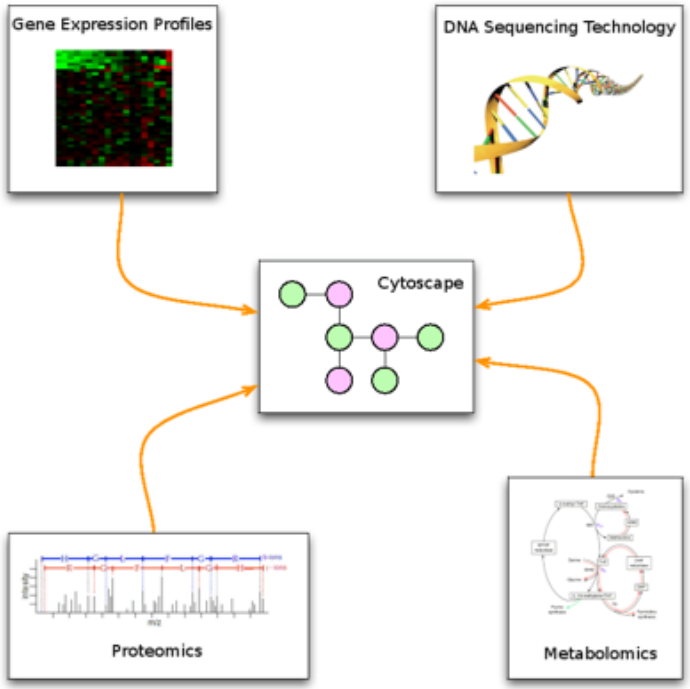


STRING network in Cytoscape

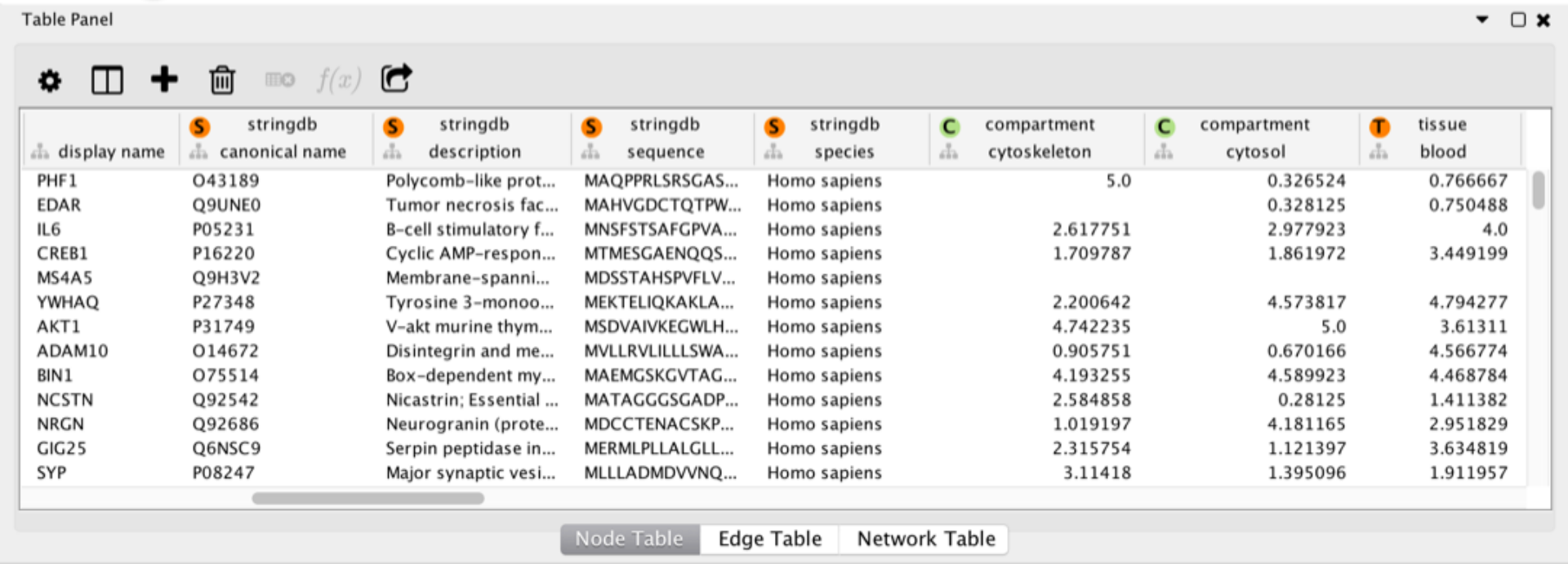


Attribute (table) data

- **Nodes and edges** can have data associated with them
  - Gene expression data
  - Mass spectrometry data
- **Import own data from files**, e.g. Excel sheets, or via automation from R or Python
- Import complex omics data (via Omics Visualizer app)
- **Identifiers have to match!**



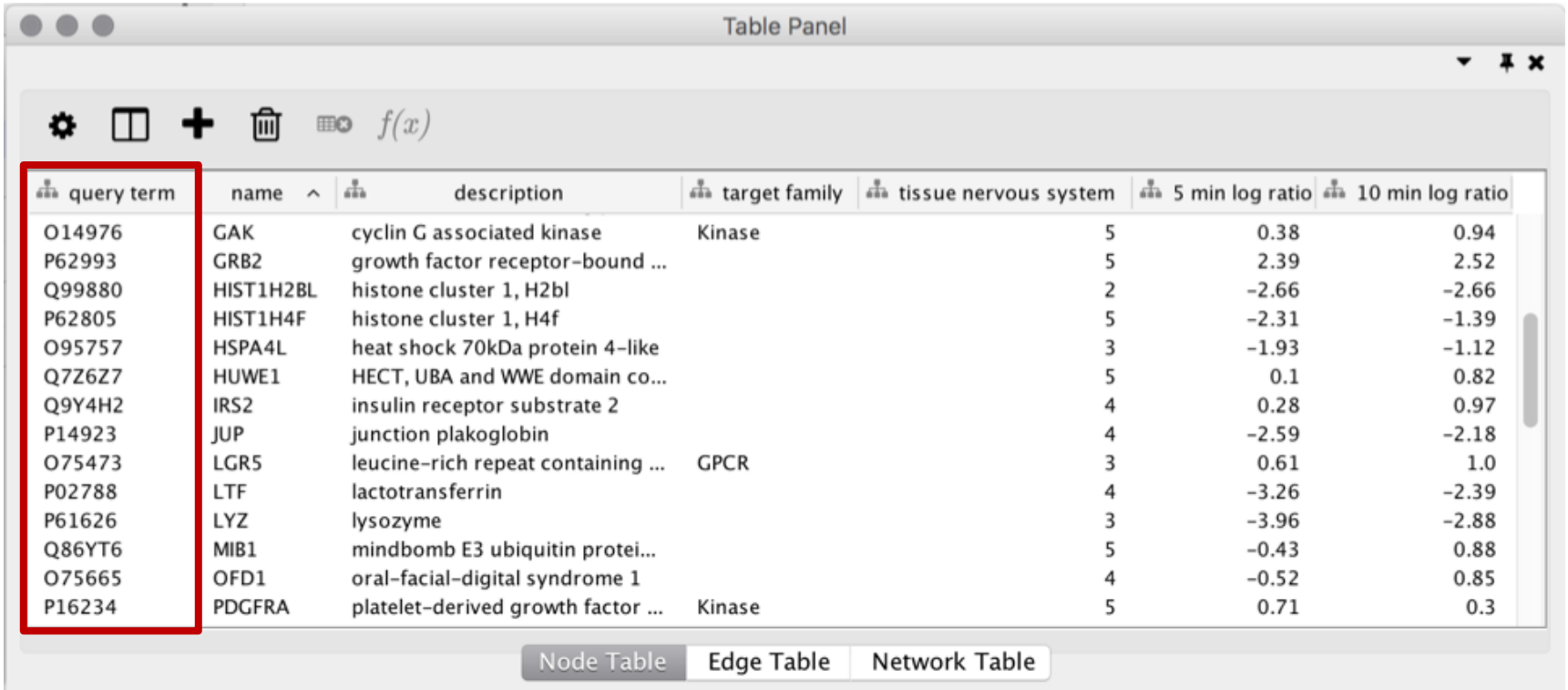
Node table



- Protein information from STRING
- Subcellular localization scores (https://compartments.jensenlab.org/)
- TISSUES expression scores (https://tissues.jensenlab.org/)
- IDG drug target information (https://pharos.nih.gov/)
- Experimental data from the original table

Know your identifiers

	A	B	C	D	G	J
1	**UniProt**	**Gene name**	**Peptides**	**Sequence coverage [%]**	**5 min log ratio**	**10 min log ratio**
2	Q99880	HIST1H2BL	5	35.7	-2.66	-2.66
3	Q8TER5	ARHGEF40	34	28.3	1.95	1.56
4	Q8IZ07	ANKRD13A	12	19.2	1.07	1.08
5	P62805	HIST1H4A	11	57.3	-2.31	-1.39
6	Q08380	LGALS3BP	14	28.2	-3.16	-2.98
7	O00750	PIK3C2B	35	24.2	2.21	2.31
8	O00443	PIK3C2A	29	17.8	1.13	1.26
9	Q9UJ41	RABGEF1	6	6.5	0.67	1.08
10	Q8TC07	TBC1D15	12	19.1	0.43	1.06



Cytoscape core concepts

Node Table

name	Degree	COMMON	gal1RGexp	gal1RGsig
YDL194W	1	SNF3	0.139	0.018043
YDR277C	2	MTH1	0.243	2.186E-5
YBR043C	1	YBR043C	0.454	5.373E-8
YPR145W	1	ASN1	-0.195	3.174E-5
YER054C	2	GIP2	0.057	0.16958
YBR045C	3	GIP1	0.786	5.5911E-6
YBL079W	1	NUP170	-0.186	2.5668E-4
YLR345W	1	YLR345W	0.108	0.012373
YIL052C	1	RPL34B	-0.258	3.7855E-5

Networks

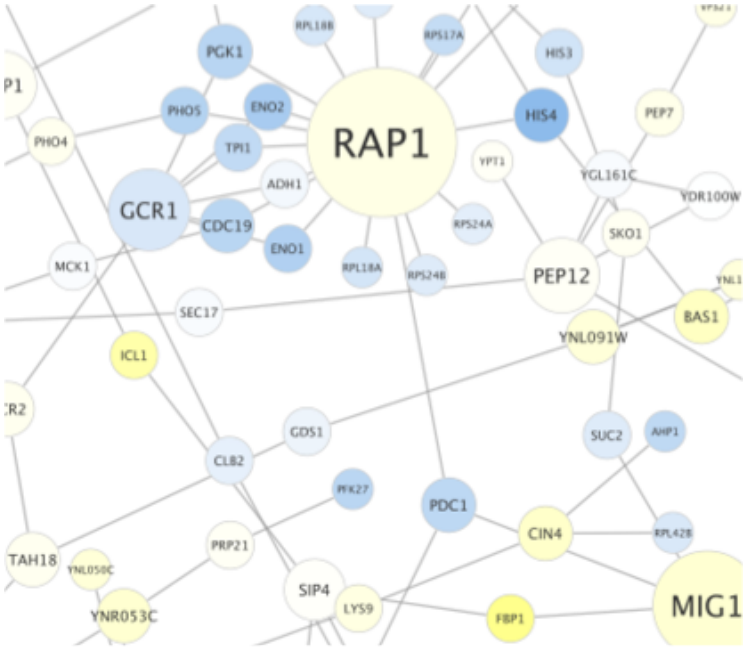
e.g., protein-protein interaction networks

Tables

e.g., actual network data or annotations

Visual Styles

Cytoscape core concepts



Node Table

name	Degree	COMMON	gal1RGexp	gal1RGsig
YDL194W	1	SNF3	0.139	0.018043
YDR277C	2	MTH1	0.243	2.186E-5
YBR043C	1	YBR043C	0.454	5.373E-8
YPR145W	1	ASN1	-0.195	3.174E-5
YER054C	2	GIP2	0.057	0.16958
YBR045C	3	GIP1	0.786	5.5911E-6
YBL079W	1	NUP170	-0.186	2.5668E-4
YLR345W	1	YLR345W	0.108	0.012373
YIL052C	1	RPL34B	-0.258	3.7855E-5

Networks

e.g., protein-protein interaction networks

Tables

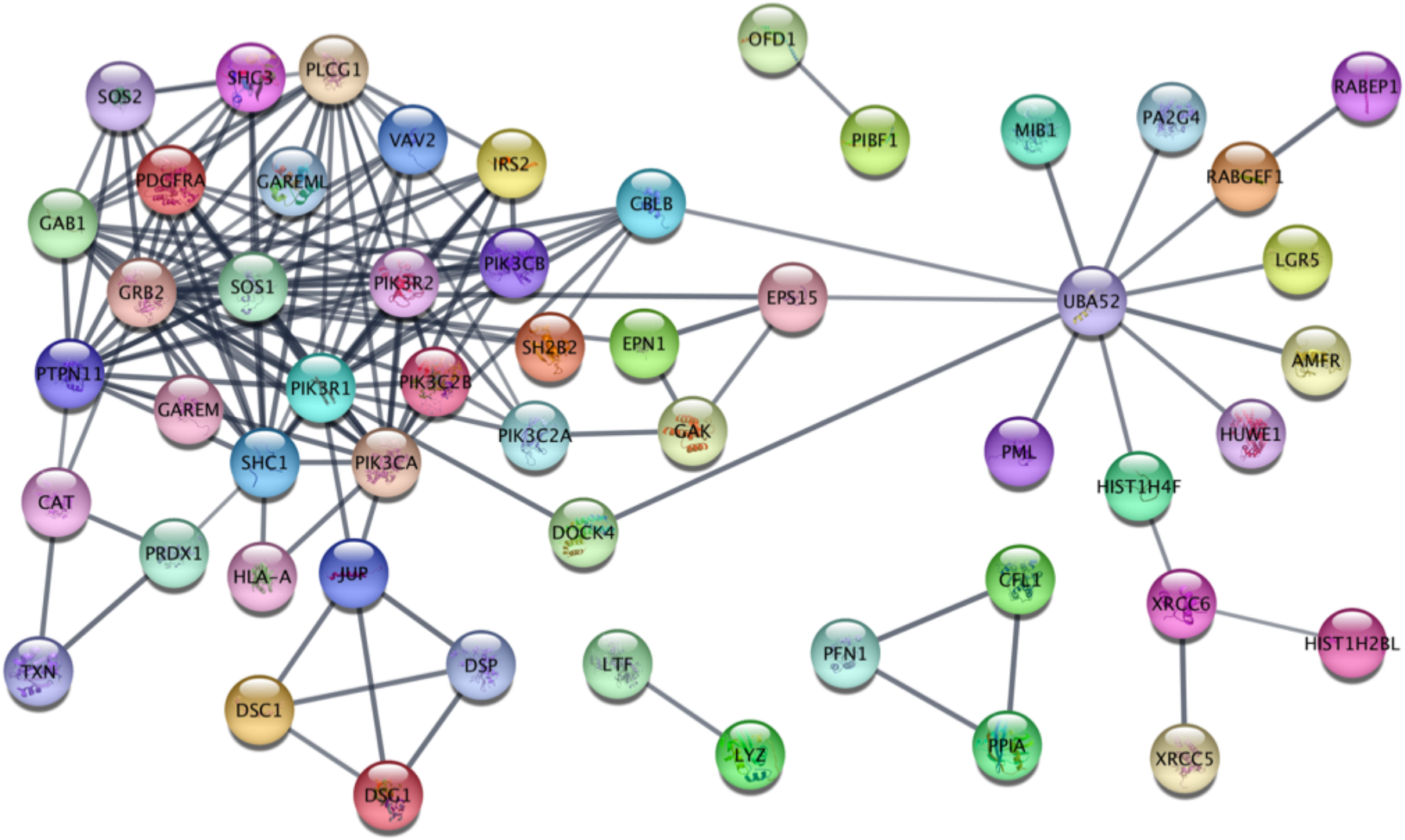
e.g., actual network data or annotations

Visual Styles

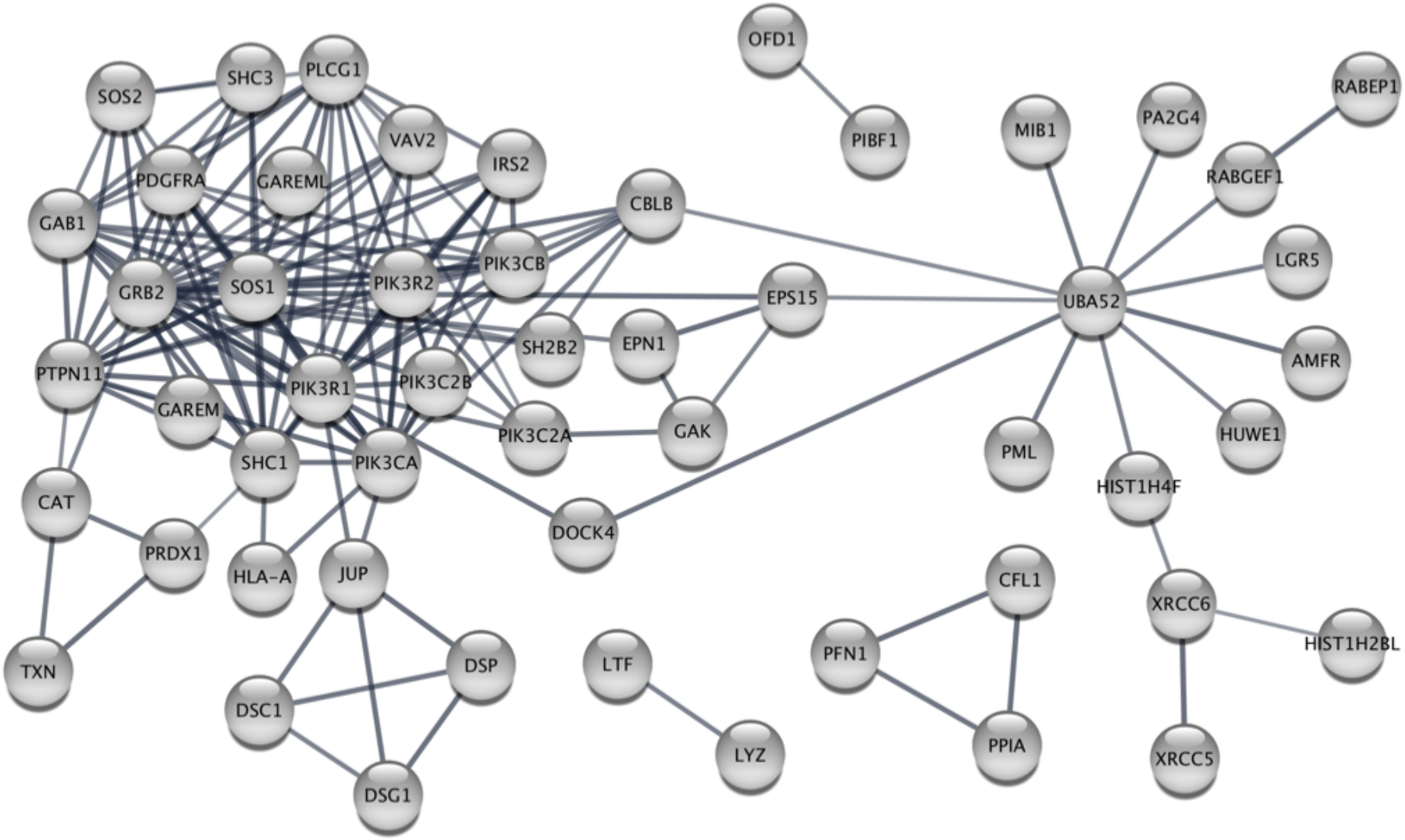
Visualize data using styles

- Visual attributes
  - **Nodes**: fill color, border color, border width, size, shape, opacity, label, etc.
  - **Edges**: line style, line color, line width, line opacity, ending type, ending color, etc.
- Mapping types
  - **Continuous (numeric values)**
    - Expression values, edge interaction scores
  - **Discrete (categories)**
    - Type of interaction, protein family
  - **Pass-through (labels)**
- Pre-defined visual styles

STRING network in Cytoscape

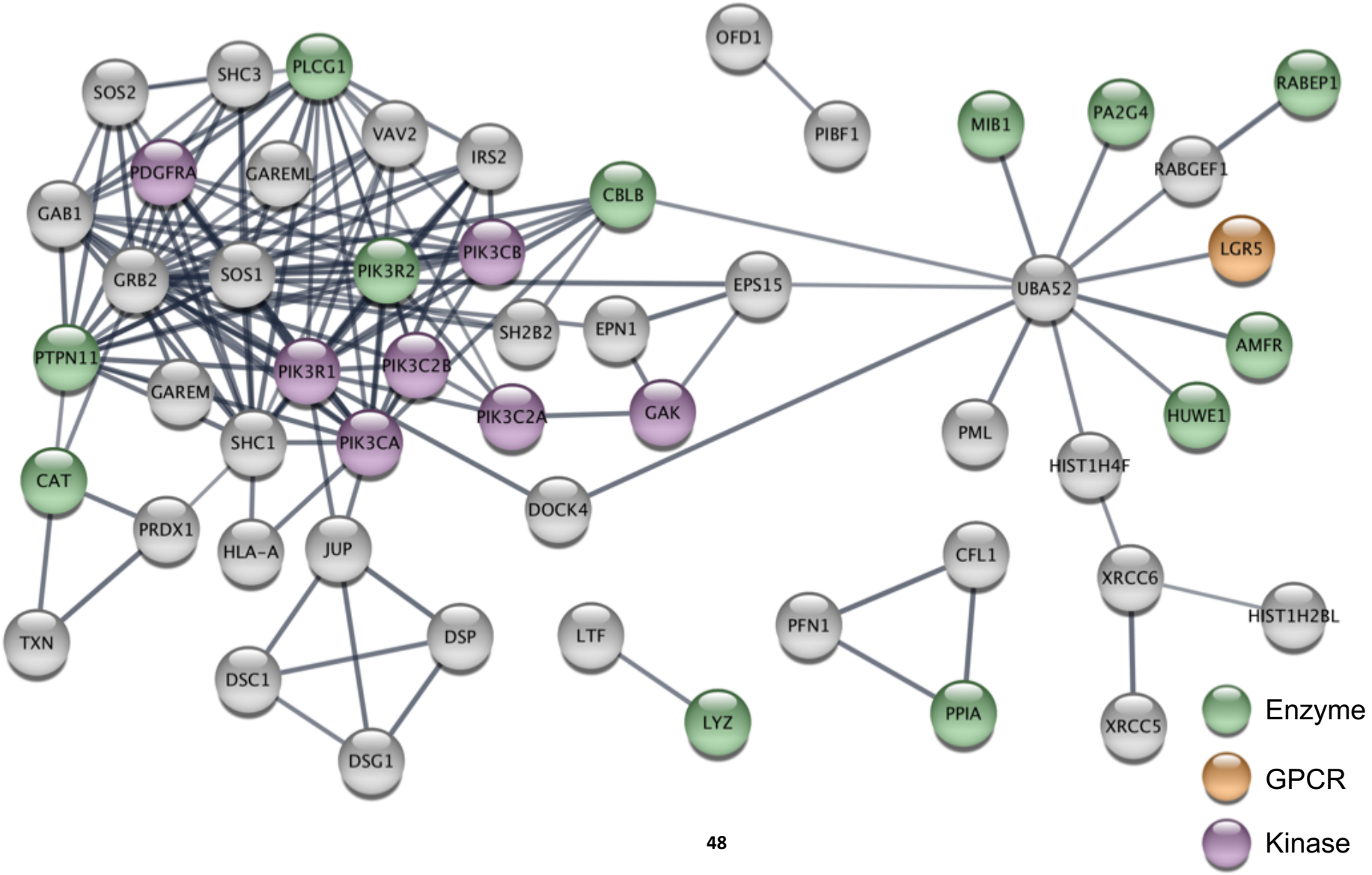


STRING network in Cytoscape



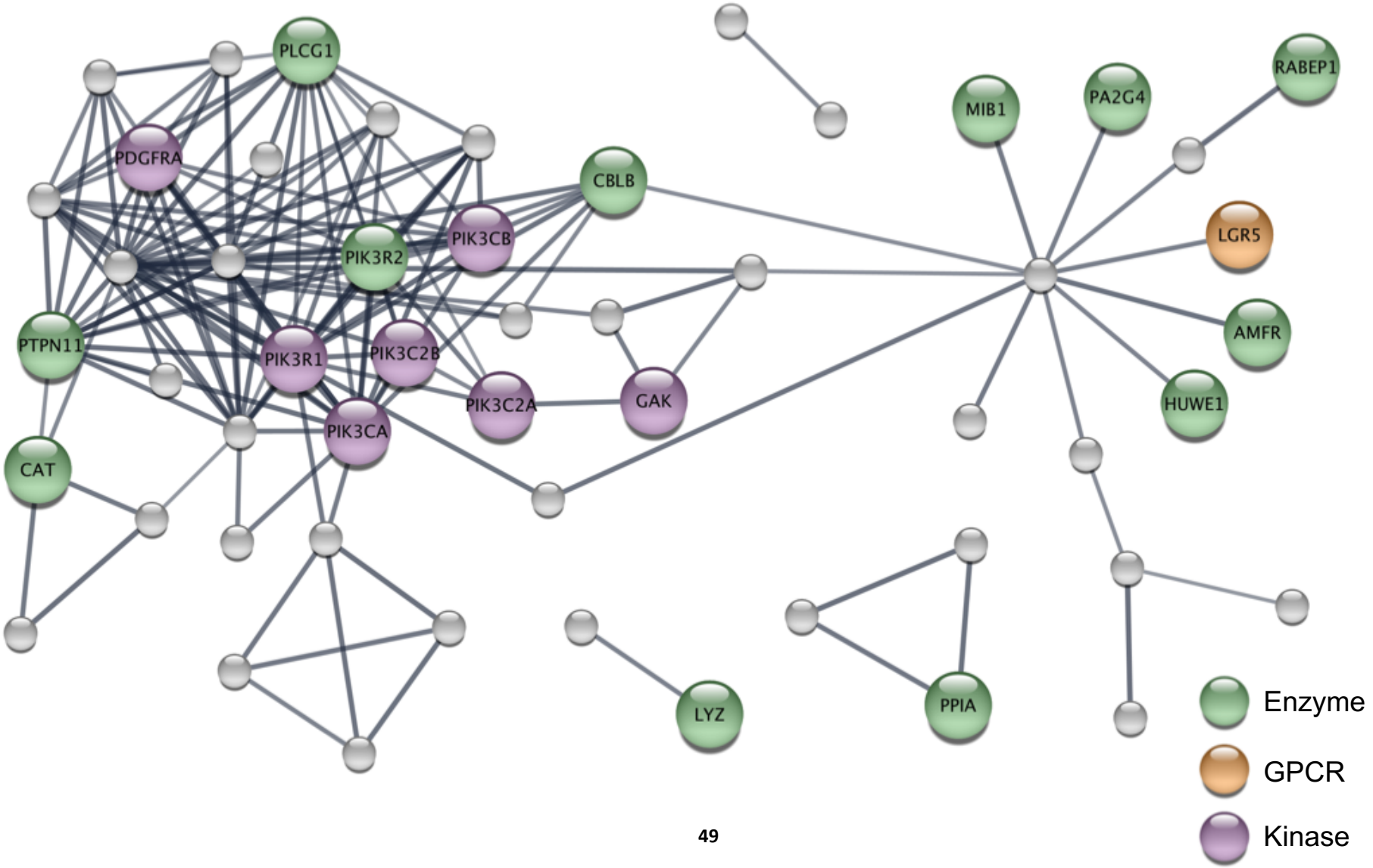
Pharos drug target information

Discrete mapping: node color

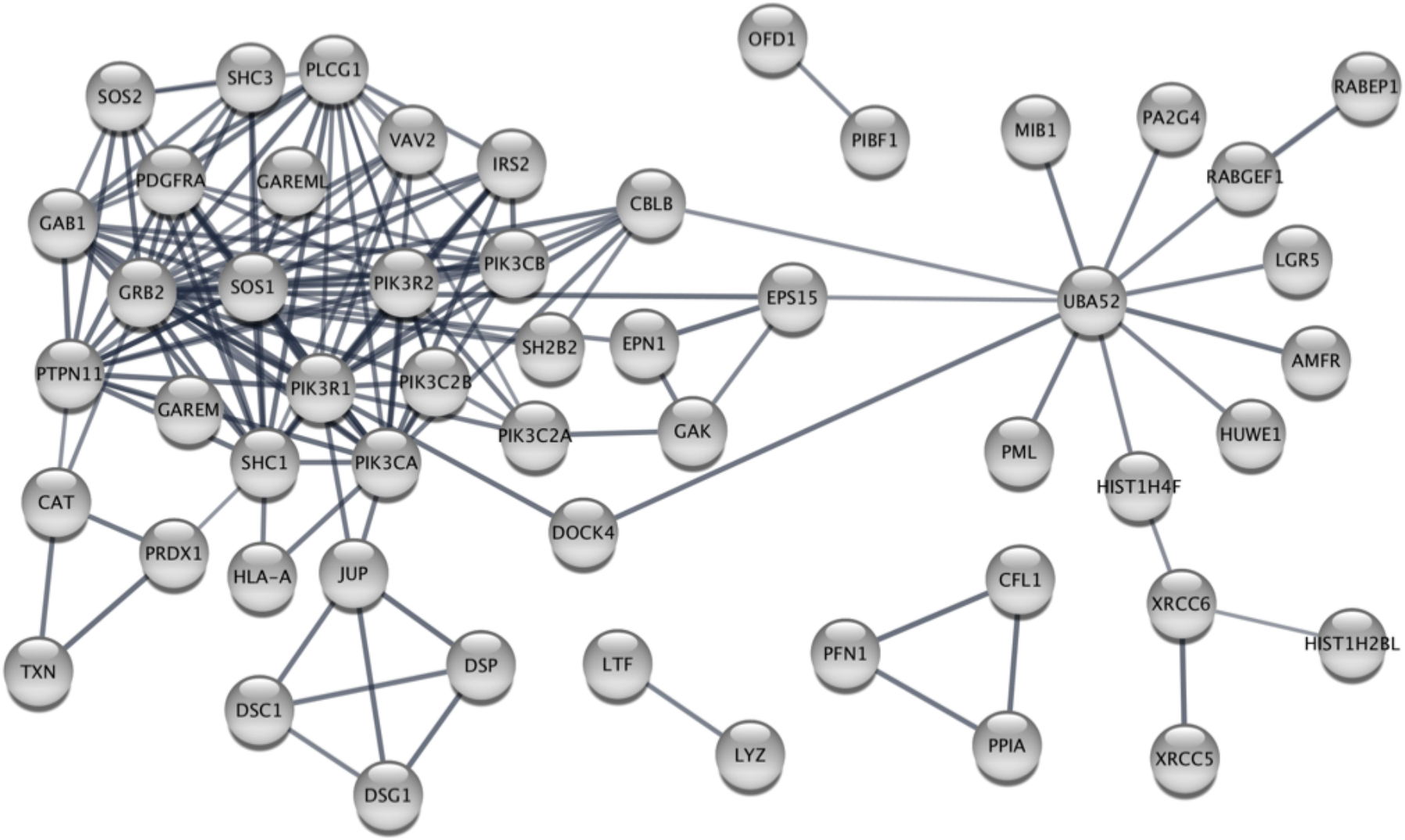


Pharos drug target information

Discrete mapping: node color & node size + Bypass: node labels

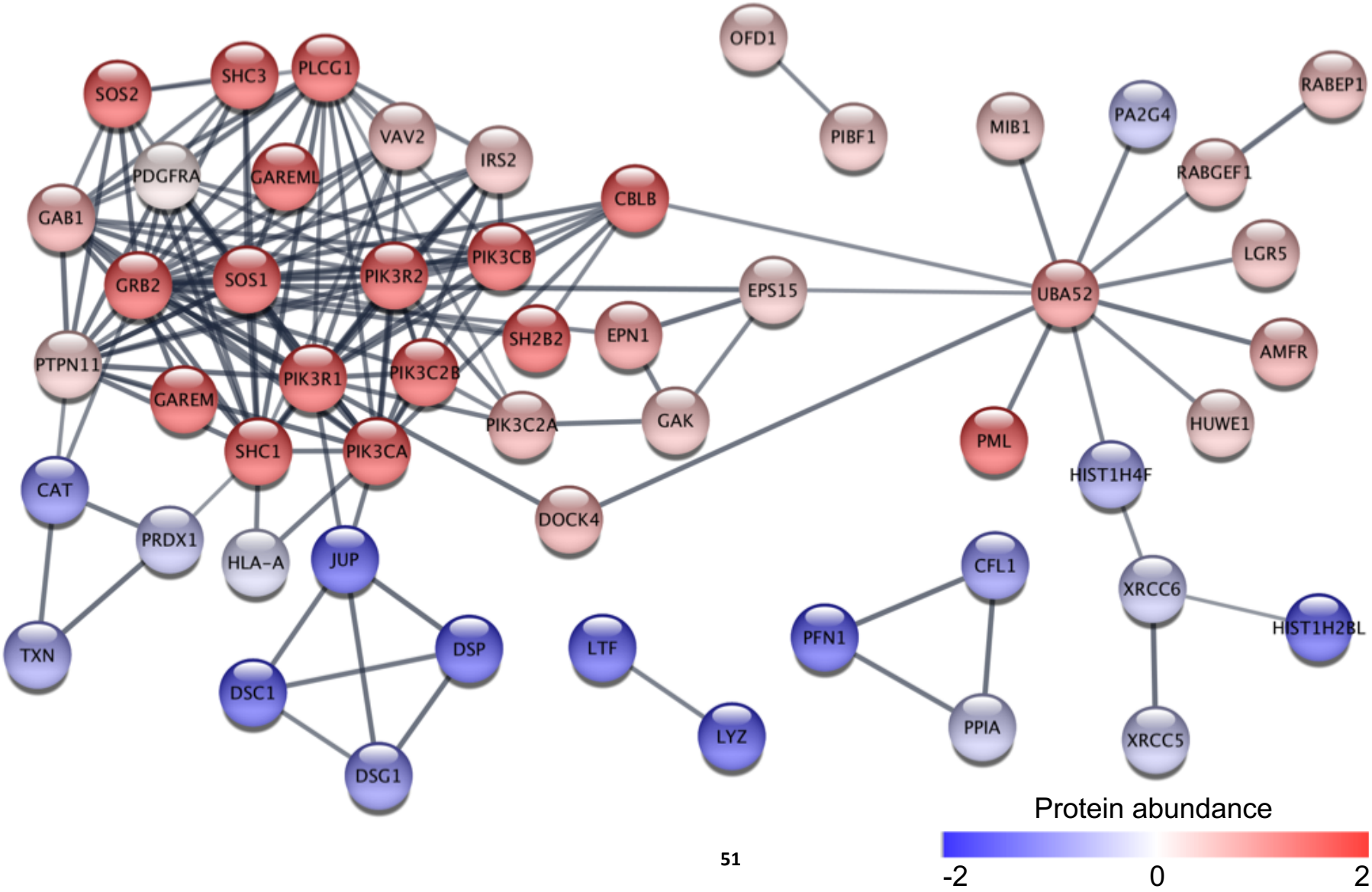


STRING network in Cytoscape



Expression data as node colors

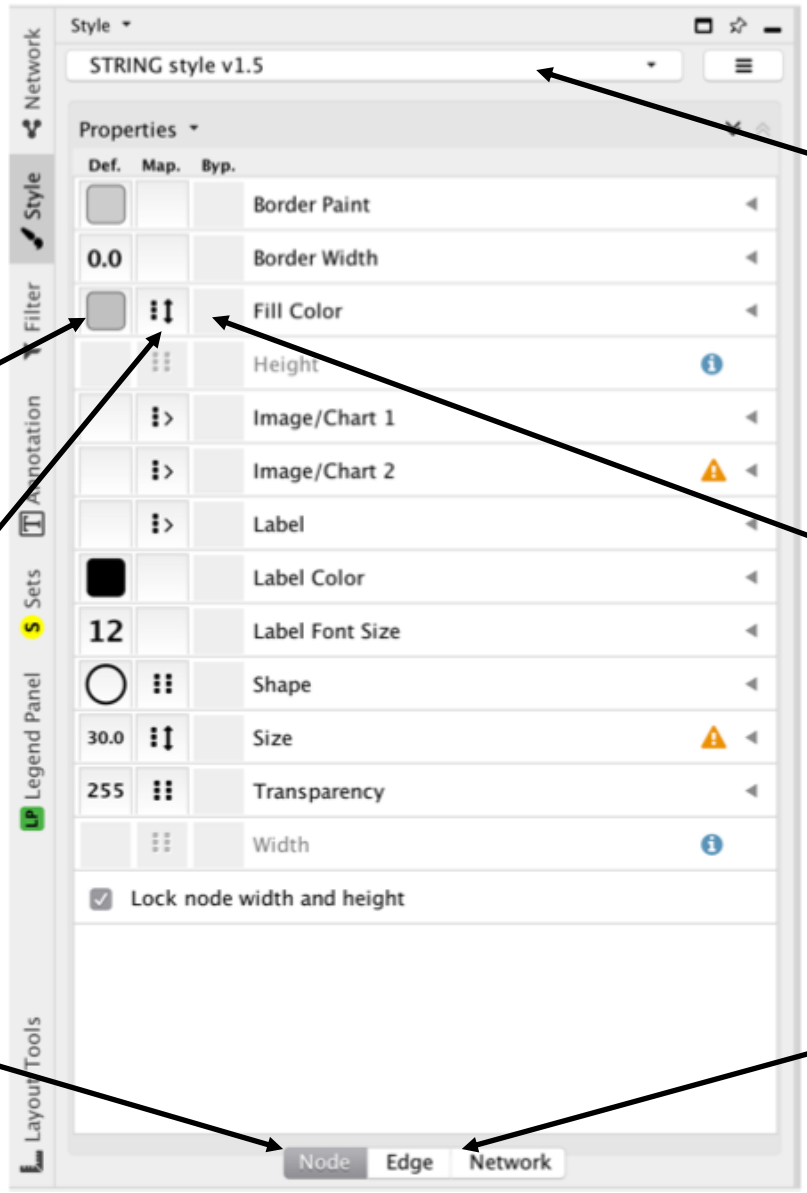
Continuous mapping: node color



Data Mapping: Tips

- Avoid cluttering your visualization with too much data
  - Highlight meaningful differences
  - Avoid confusing the viewer
  - Consider creating multiple network images

Styles: User interface



Styles tab

Available **styles**

Default fill color: grey

Bypass mapping

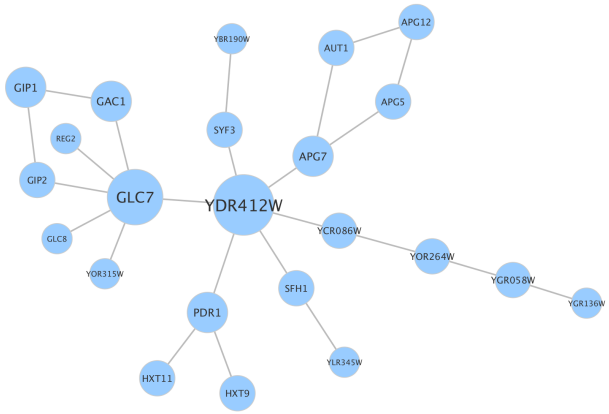
Fill color **mapping**: rainbow colors

Node tab

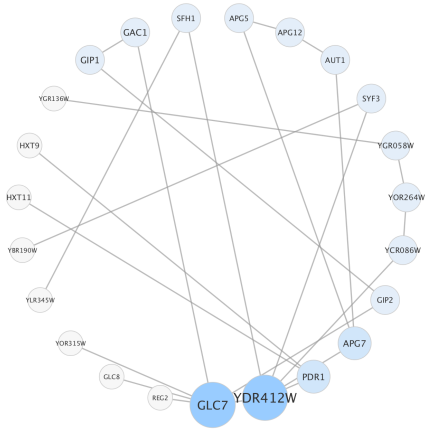
Edge tab

Layouts

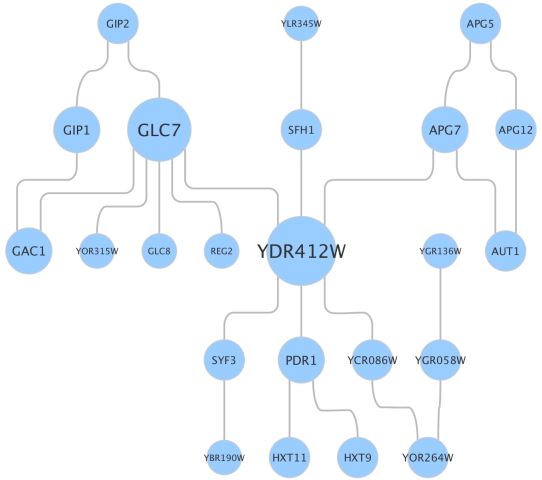
- Layouts determine the location of nodes and (sometimes) the paths of edges
- Recommended apps: yFiles Layout Algorithms, layoutSaver



Force-directed



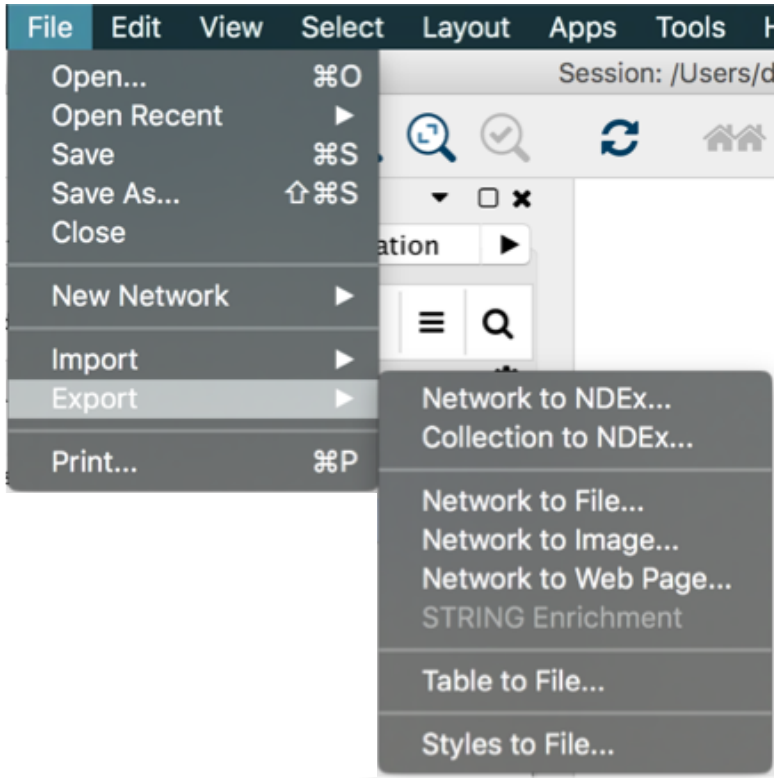
Circular



Hierarchical

Save data

- **Cytoscape sessions** save everything (.**cys** files)
- Export networks in different formats
- Export node & edge tables as text files
- Publication quality graphics in several formats



stringApp exercise 2 (20 min)

https://jensenlab.org/training/stringapp/eubic/

In this exercise, we will work with the list of proteins associated with epithelial ovarian cancer (EOC) in the study by Francavilla *et al.* to perform typical network imoprt and visualization tasks.

2.1 Protein network retrieval & layout

Question 1: *How many nodes and edges are there in the resulting network? Do the proteins all form a connected network? Why?*

Question 2: *Does any of the suggested layouts make patterns in the network easy to recognize? (Recommended: install the app* ***yFiles Layout Algorithms****)*

2.2 Discrete color mapping

Questions: How many of the proteins in the network are ion channels (IC) or GPCRs? How many kinases are in the network?

2.3 Data import

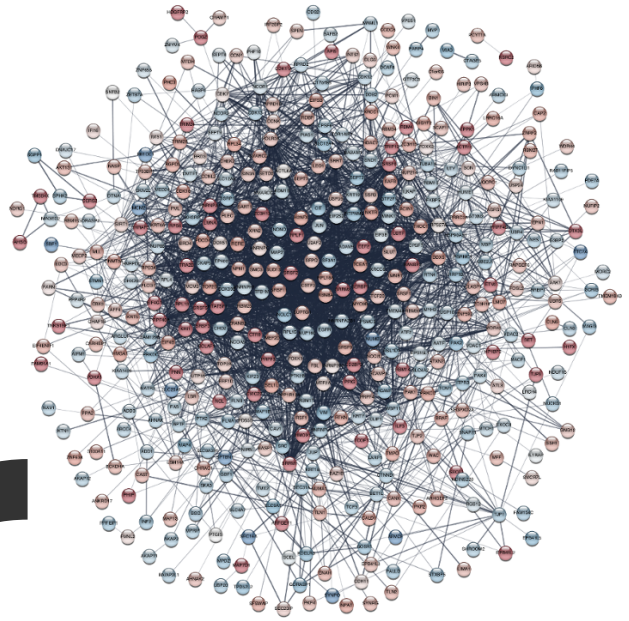
Question 4: *Do you see the columns from the Excel table in the Node Table?*

2.4 Continuous color mapping

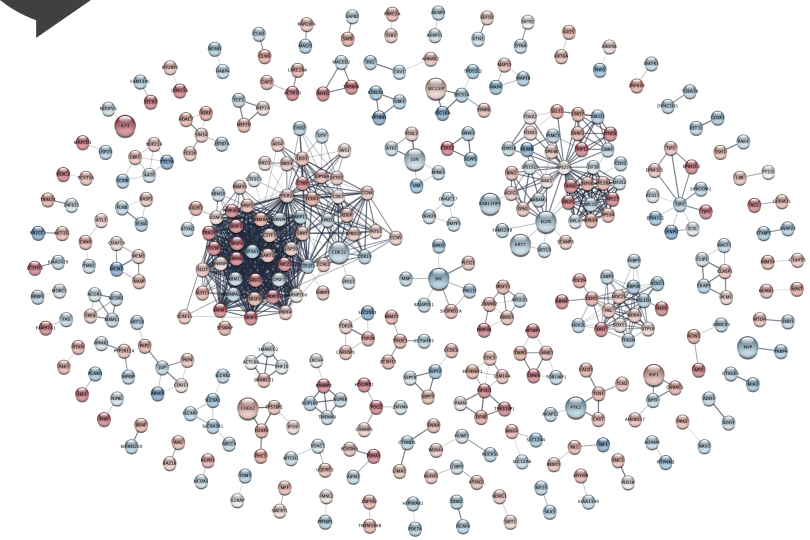
Question 5: *Are the up-regulated nodes grouped together? Do you see any issues with the color gradient and can you improve it?*

Why use (biological) networks?

- Networks are **powerful tools**
  - ✓ Reduce complexity
  - ✓ More efficient than tables
  - ✓ Great for data integration
  - ✓ Intuitive visualization



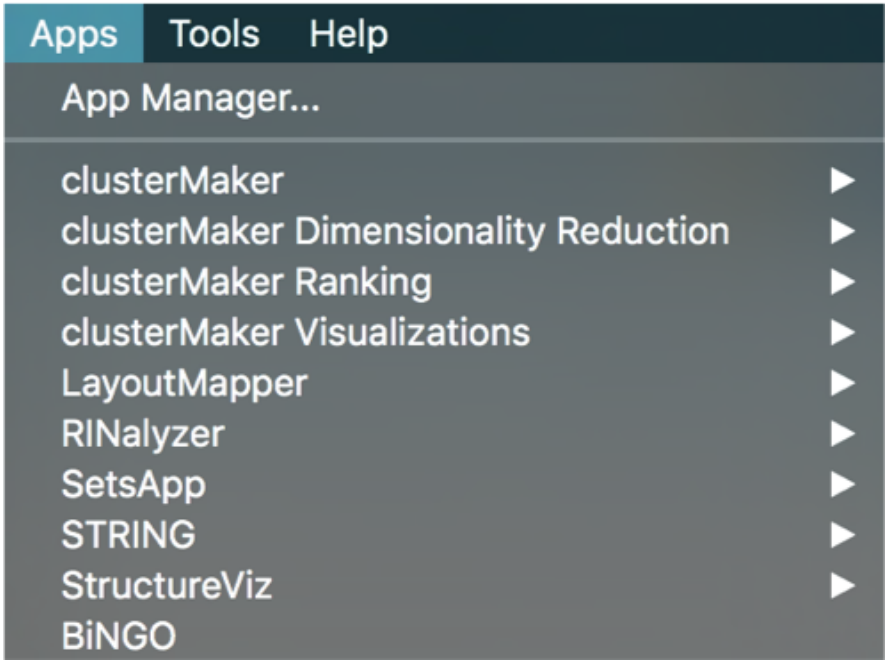
- But also… Challenging!



- Many different network analysis and visualization techniques available

Annotate & analyze the network

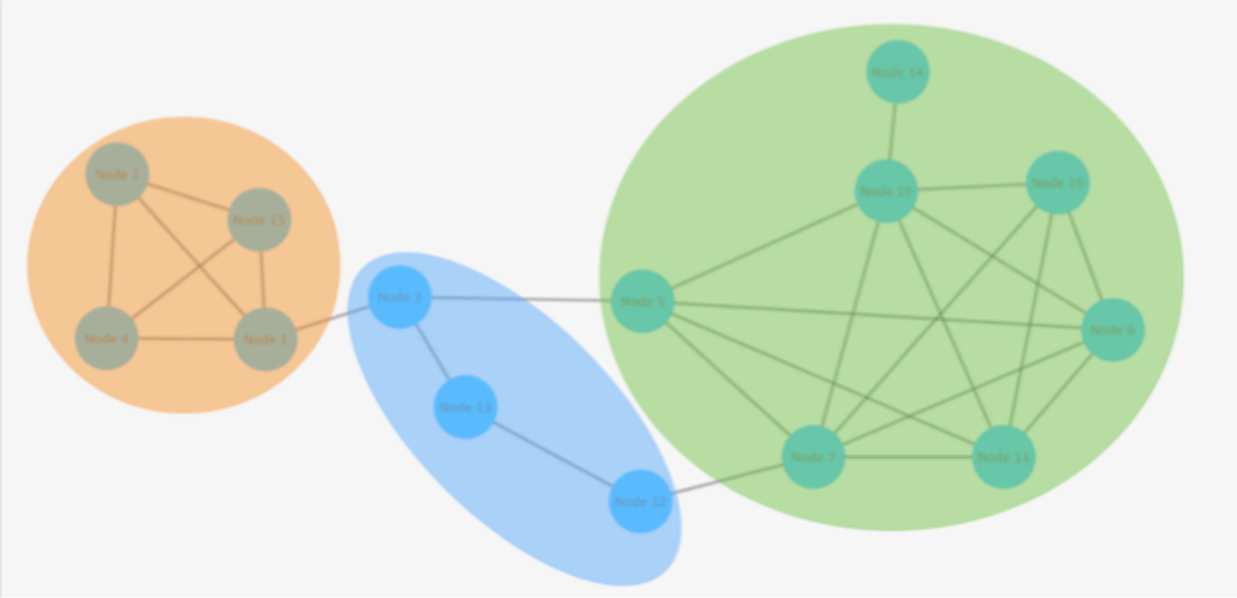
- Functional enrichment
- Topological analysis
- Clustering
- And many more…



→ Visit the Cytoscape app store at https://apps.cytoscape.org/

Network clustering

- Group nodes together based on some measure of distance or similarity between the nodes
- Makes the network easier to understand
- MCL (Markov CLustering)
  - Fast algorithm
  - No need to specify number of clusters



Clustering in Cytoscape

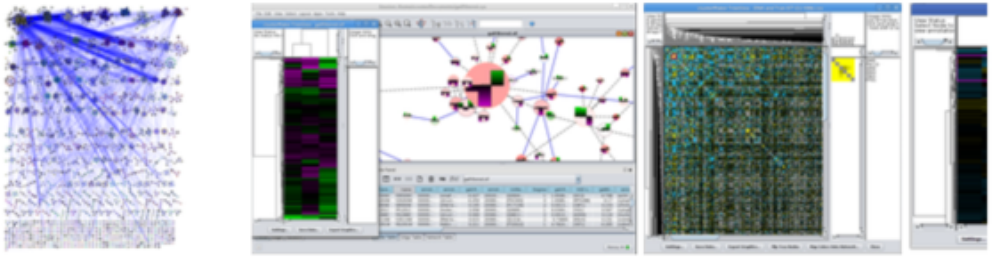
clusterMaker2

Multi-algorithm clustering app for Cytoscape

★★★★☆ (23) 89244 downloads | citations | discussions

Details | Release History

Categories: automation, clustering, data visualization, gene expression, grouping, heat map visualization, visualization



clusterMaker2 is the Cytoscape 3 version of the clusterMaker plugin. clusterMaker2 provides several clustering algorithms for clustering data within columns as well as clustering nodes within a network. This version also provides support for two new algorithms: Fuzzy C-Means and a new "Fuzzifier". In addition to providing clustering algorithms, clusterMaker2 provides heatmap visualization of both node data and edge data as well as the ability to create new networks based on the results of a clustering algorithm.

Current node attribute algorithms:

- Hierarchical
- K-Means
- K-Medoid

CYTOSCAPE 3

Download

Version 1.3.1

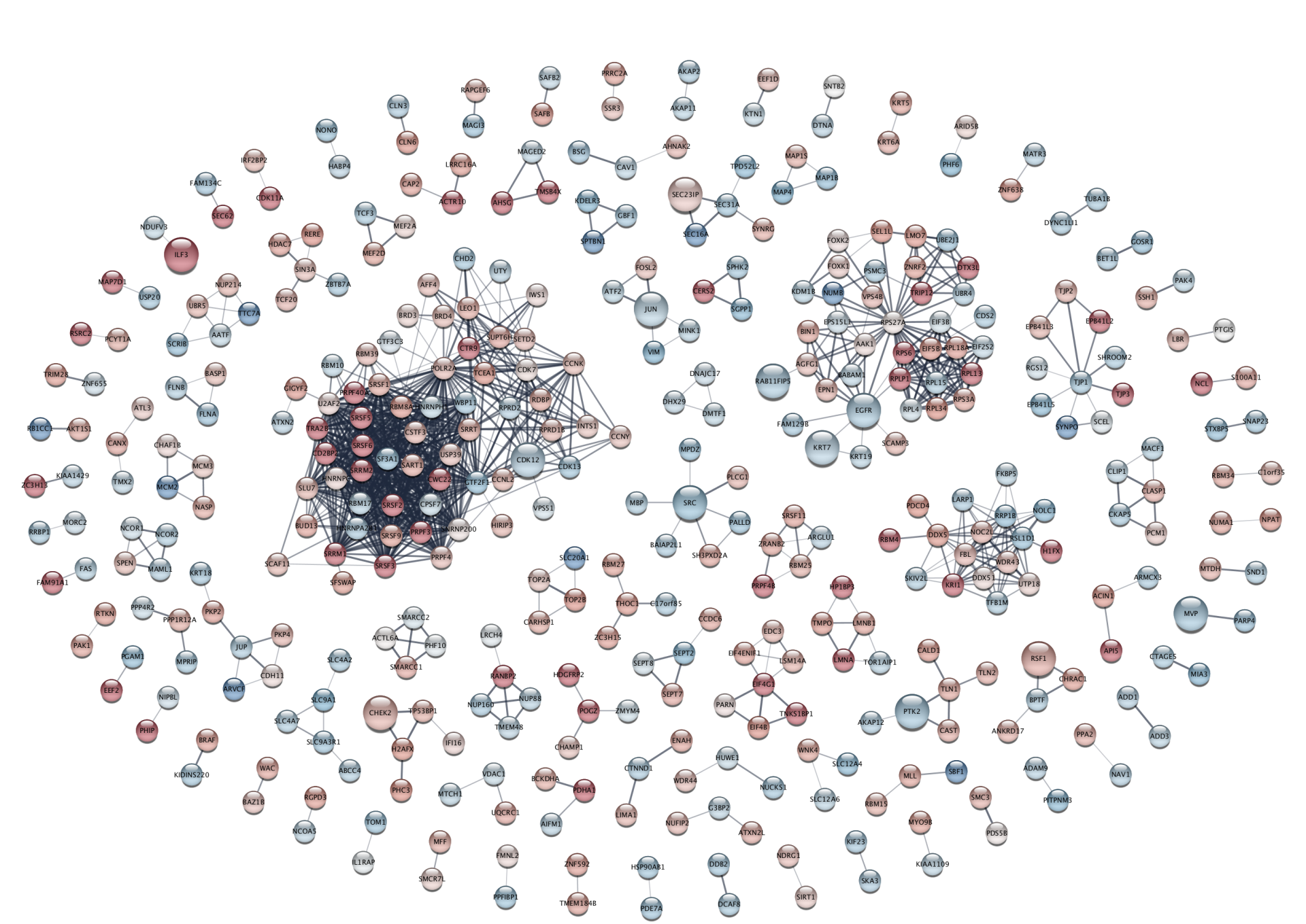
Released 30 Oct 2018

Works with Cytoscape 3.6

Download Stats Click here

RESOURCES

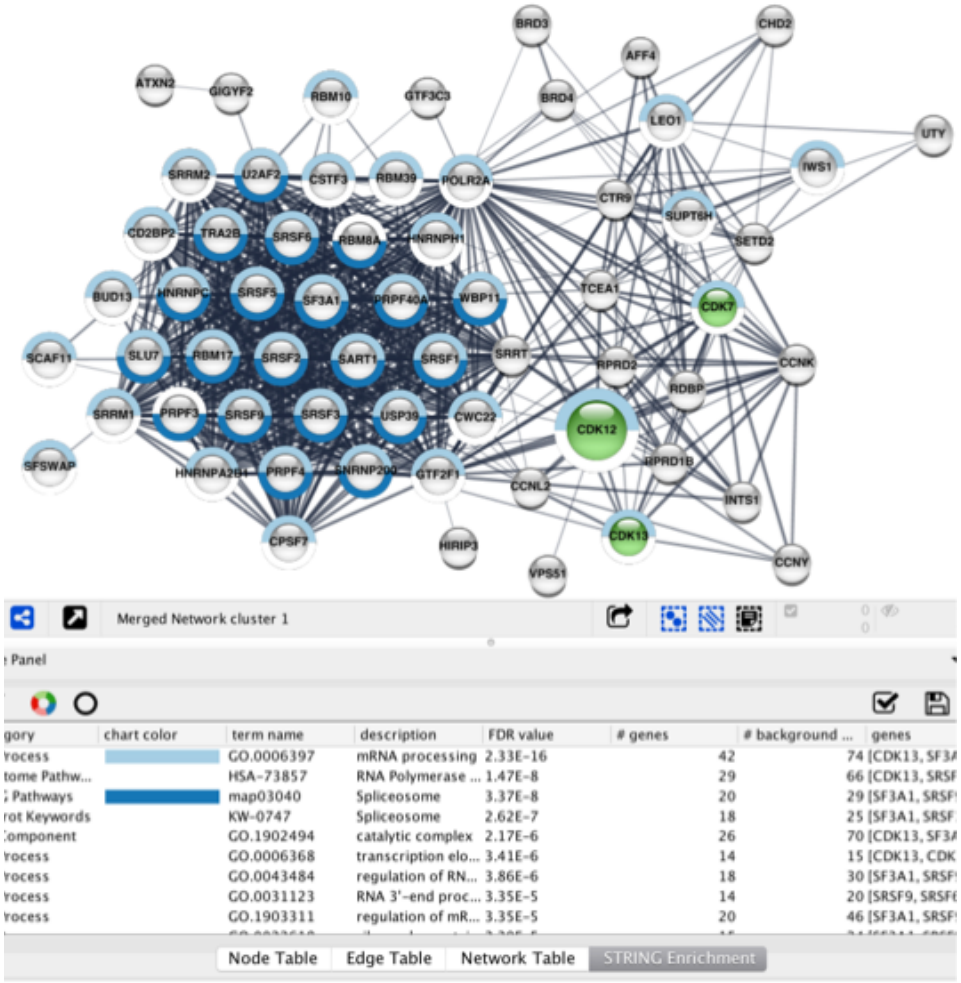
- Ask a question
- Search BioStars
- Website
- Tutorial
- Cite this App
- Code Repository
- Automation Support
- E-mail



Doncheva *et al.*, *J Proteome Research*, 2019

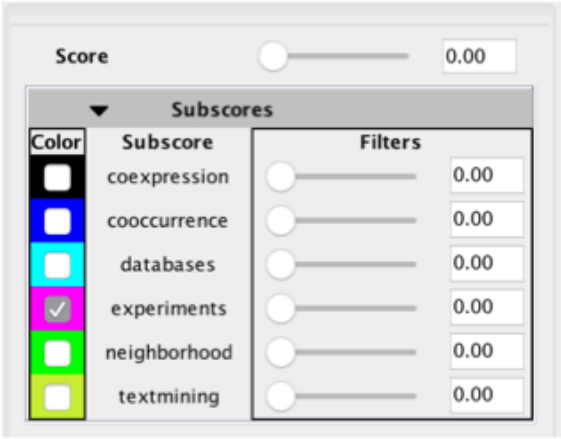
stringApp functional enrichment

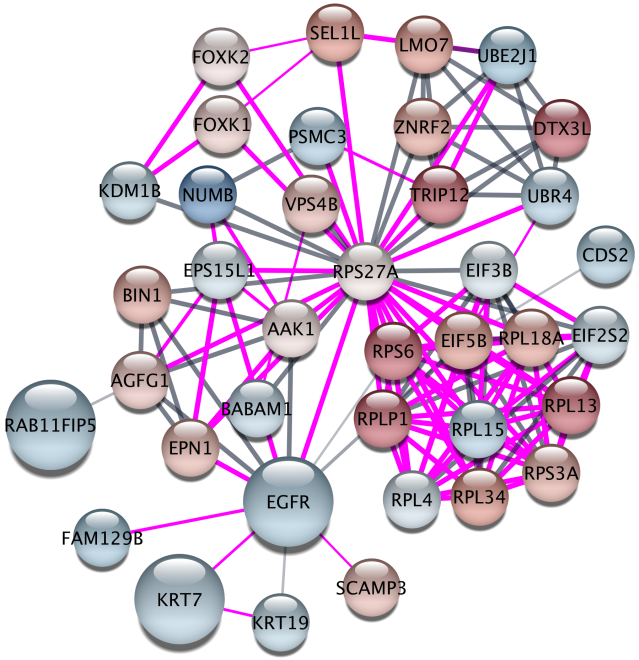
- Not really a network analysis technique, but very **useful for visualization**
- Filter by category, remove redundant terms
- Visualize significant terms
- Many categories: Gene Ontology terms, Reactome, KEGG & Wiki pathways, Tissues, Diseases, Subcellular localization, STRING clusters, Protein domains, Publications



More stringApp features

- Change confidence level of interactions → important when merging networks
- Expand network by a user defined number of additional interactors → for example compounds
- Filter nodes by tissue or compartment annotation
- Filter and color edges by evidence
- Stringify networks not created with the stringApp
- Query for virus-host interactions





Useful apps & features

- Functional enrichment: stringApp, clueGO & CluePedia, BiNGO, EnrichmentMap
- Clustering: clusterMaker2
- Merge networks: Cytoscape built-in (Tools → Merge)
- Omics data visualization: OmicsVisualizer
- Pathways: KEGGscape, WikiPathways, ReactomeFI, OmniPath
- Network topology: Cytoscape built-in (Tools → Analyze network)

stringApp exercise 3 (10 min)

https://jensenlab.org/training/stringapp/eubic/

In this exercise, we will work with the list of differentially abundant proteins from the study by Francavilla *et al.* and perform typical network analysis tasks.

Prerequisites: *install the app* ***ClusterMaker2***

3.1 Network clustering

Question 1: *How many clusters have at least 10 nodes?*

Question 2: *How many nodes and edges are there in this cluster?*

3.2 Functional enrichment

Question 3: *How many statistically significant terms are in the table? Which is the most significant term for each of the categories GO Biological Process, Reactome and KEGG Pathways?*

3.3 Functional enrichment extras

Question 4: *What is the title of the most recent publication?*

stringApp exercise 4 (10 min)

https://jensenlab.org/training/stringapp/eubic/

In this exercise, we will compare the network of differentially abundant proteins from the study by Francavilla *et al.* and the network of genes associated with the same disease based on literature and knowledge from the DISEASES database.

4.1 Overlap with DISEASES network

Question 1: *How many nodes are in the intersection?*

4.2 Integrate networks

Question 2: *Which protein from the experiment has the highest disease score?*

Question 3: *Can you find the protein with the highest disease score in the network view?*

Complex input data tables

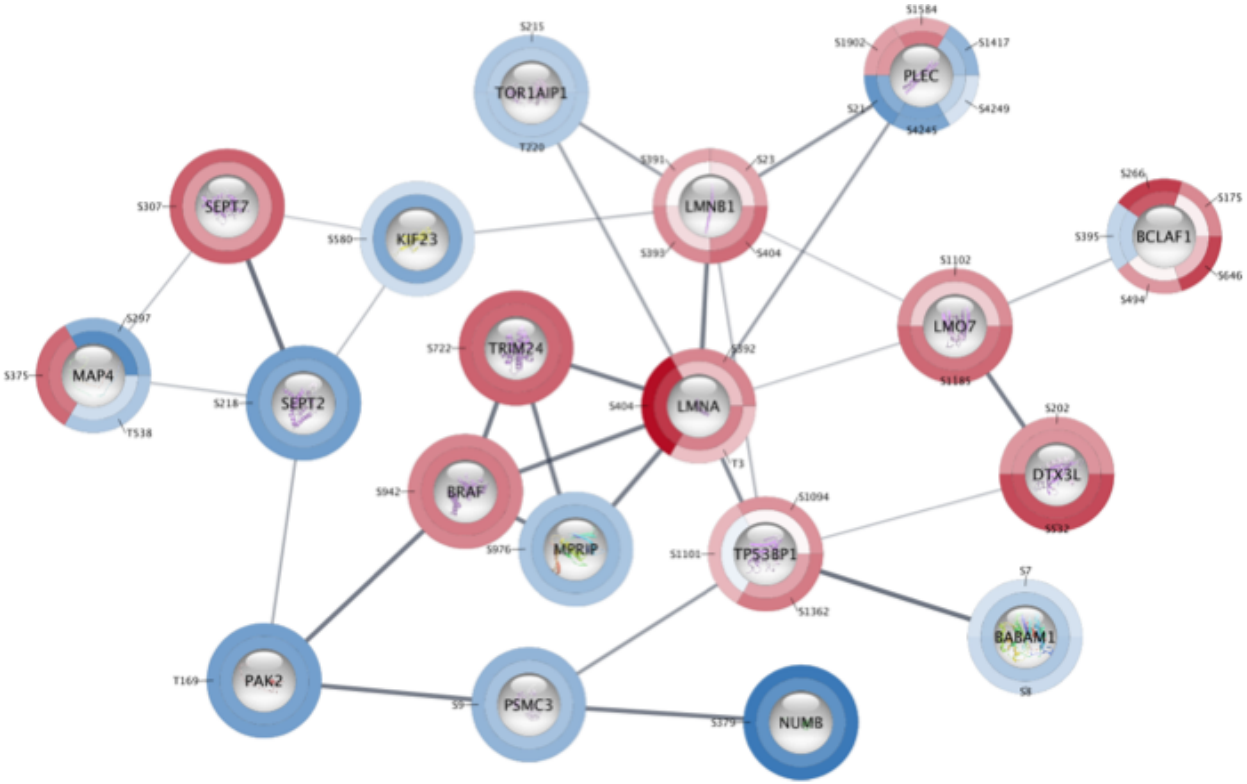
A1 | fx | UniProt

	A	B	C	D	E	F	G	H	I	J
1	UniProt	Gene name	Modified sequence	AA position	EOC	FTE	OSE	EOC vs FTE	EOC vs OSE	Adj p-value
2	O15439	ABCC4	_KDNEESEQPPVPGT(ph)PTLR_	T646	25.72	27.83	28.18	-2.11	-2.46	3.16E-02
3	Q9UKV3	ACIN1	_SSSISEEKGDS(ph)DDEKPR_	S118	24.82	25.54	22.18	-0.72	2.64	2.15E-02
4	Q9UKV3	ACIN1	_AAKLS(ph)EGS(ph)QPAEEEEDQETPSR_	S145	26.22	25.02	23.24	1.21	2.98	1.08E-02
5	Q9UKV3	ACIN1	_SKS(ph)PS(ph)PPRLTEDR_	S290	26	23.46	22.82	2.53	3.17	2.06E-02
6	Q9UKV3	ACIN1	_RLS(ph)QPESAEK_	S614	30.2	26.98	25.94	3.23	4.27	9.13E-04
7	Q9UKV3	ACIN1	_LQPERGS(ph)PK_	S633	27.31	24.19	22.94	3.12	4.37	4.03E-03
8	Q9UKV3	ACIN1	_GVPAGNS(ph)DTEGGQPGRK_	S742	25.39	21.05	21.62	4.33	3.77	1.89E-03
9	O96019	ACTL6A	_EAVREGS(ph)PANWK_	S233	25.07	25.71	23.7	-0.64	1.36	2.64E-02

- Usual workflow: given a list pf proteins, import a STRING network and add experimental data from the table
- But what to do with multiple lines of data for one node? Aggregate the data or select the "best" site?
- ⇒ How to visualize multiple lines of data for one node?

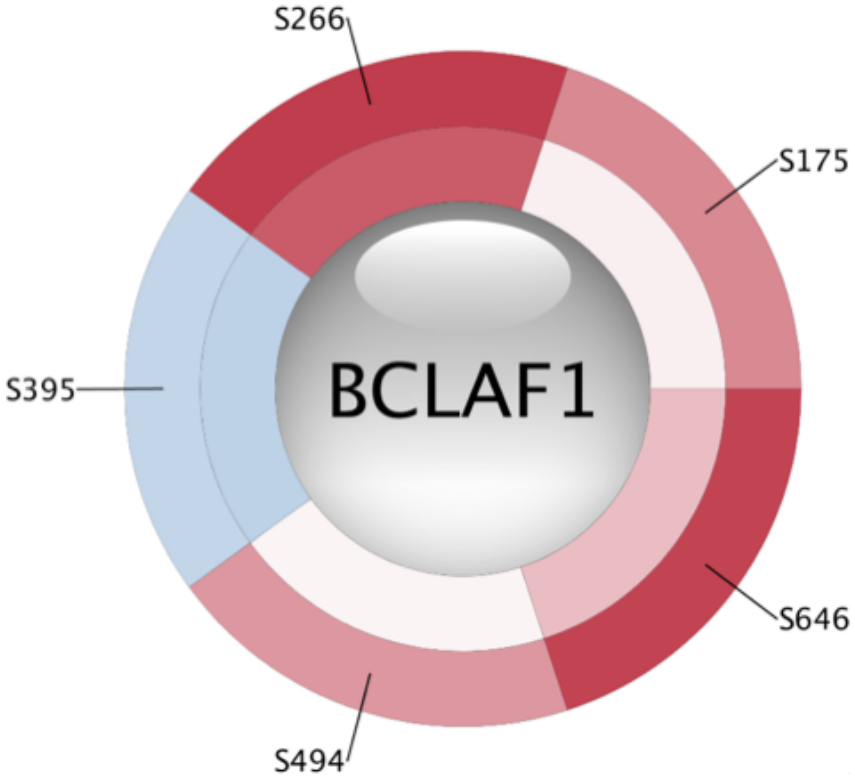
Omics Visualizer features

- Import multi-omics data as an **Omics Visualizer table**
- Retrieve a STRING network for the proteins in the table
- Visualize data (e.g. phospho-site specific) as pies inside the nodes or donuts around the nodes



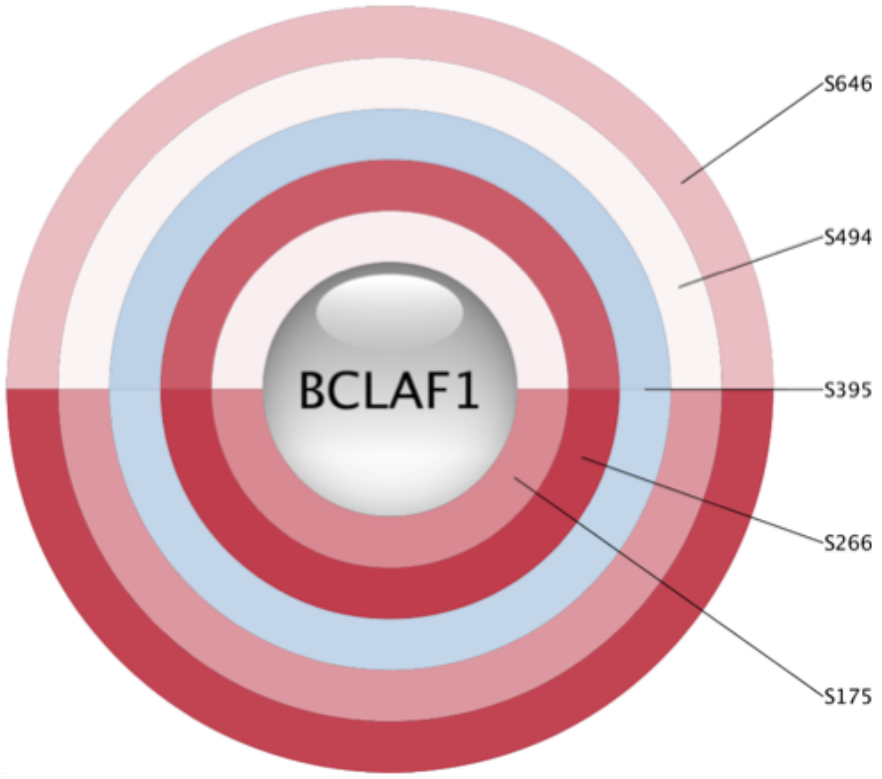
Donut visualization

Data: two different conditions as two columns, several phospho sites for each protein as different rows



Ring as column = condition

(Slice as row)

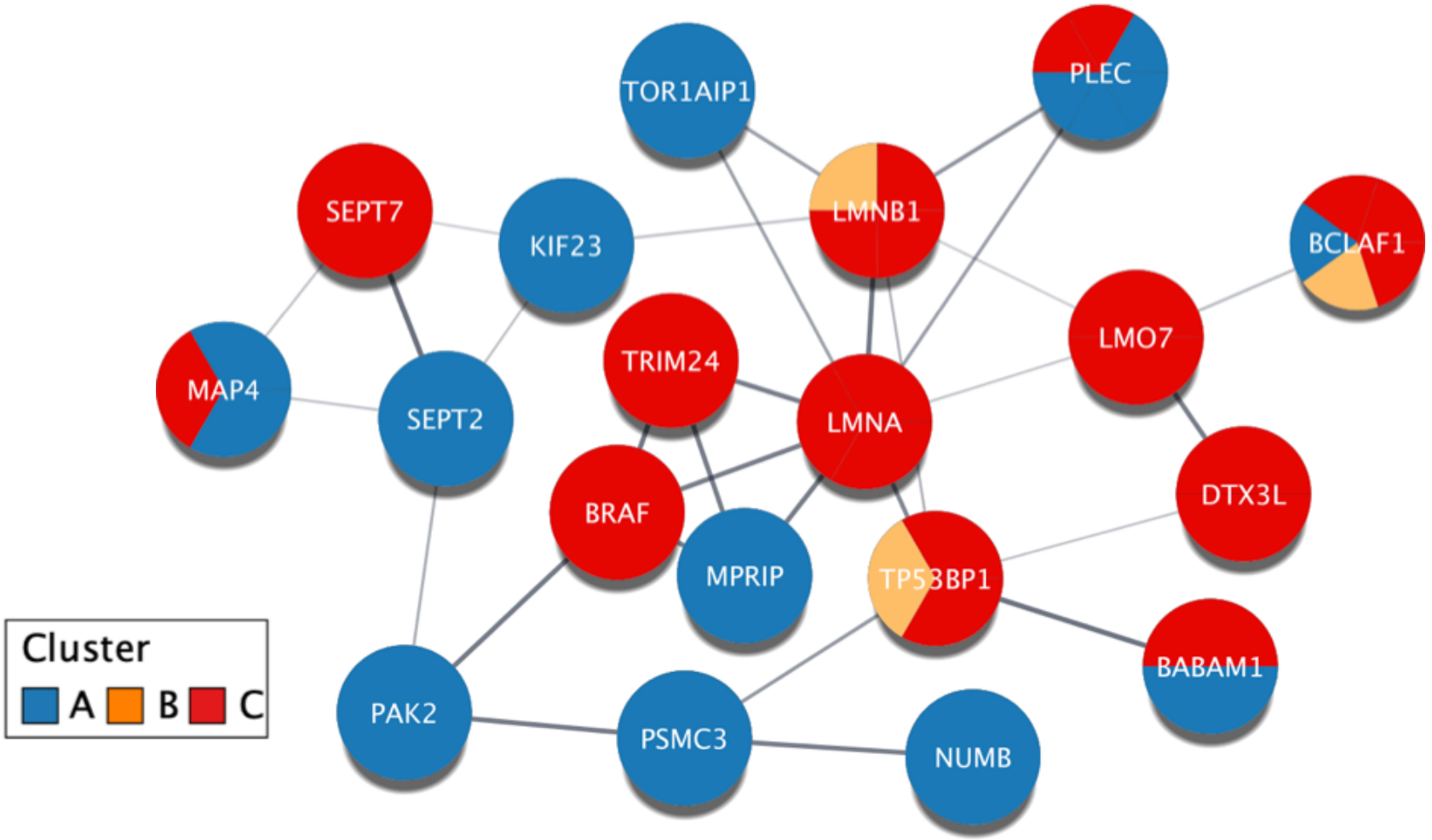


Ring as row = phospho side

(Slice as column)

Pie visualization

Note: only possible for one data column!



There is also an automatically generated legend!

Omics Visualizer ex. 1 (20 min)

https://jensenlab.org/training/omicsvisualizer/

In this exercise, we will work with the list of 541 differentially abundant proteins from the study by Francavilla *et al.* with focus on the phosphorylation data.

Prerequisites: *install the app* ***OmicsVisualizer***

1.1 Table import

1.2 Table row filtering

Question 1: *How many rows do you have in the table? How many rows are shown after you applied the filter?*

1.3 STRING network retrieval

Question 2: *How many nodes do you have in your network? Does it corresponds to the number of rows you queried?*

1.4 Donut visualization

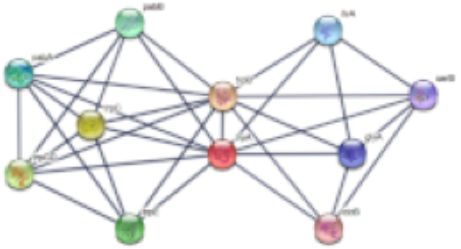
Question 3: *How are multiple sites on the same protein shown? Do different donut slices within a protein always show similar changes?*

1.5 Pie visualization

Question 5: *What are the advantages and disadvantages of donut vs. pie visualizations?*

Supporting lectures

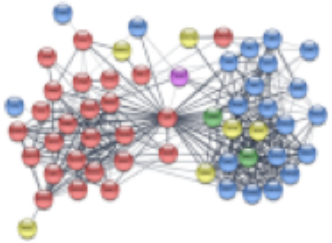
The STRING database



The Cytoscape platform

Cytoscape stringApp

The DISEASES database



Enrichment analysis

https://www.youtube.com/c/LarsJuhlJensen

Tutorials & getting help

- STRING, stringApp & Omics Visualizer:
  - YouTube videos: https://www.youtube.com/c/LarsJuhlJensen
  - Tutorials & exercises: https://jensenlab.org/training/
- Cytoscape
  - Online manual: http://manual.cytoscape.org/
  - Tutorials: https://github.com/cytoscape/cytoscape-tutorials/wiki
  - YouTube videos: https://www.youtube.com/channel/UCv6auk9FK4NgXiXiqrDLccw
  - Helpdesk mailing list: cytoscape-helpdesk@googlegroups.com
- Automation using R and Python
  - https://github.com/cytoscape/cytoscape-automation/wiki
  - https://github.com/scaramonche/EuBIC2020_Cytoscape