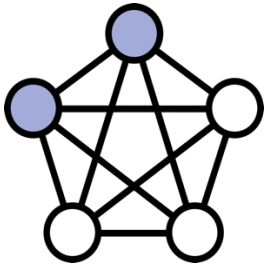
stringApp



Network biology approaches for high-throughput data analysis

Nadezhda T. Doncheva

nadezhda.doncheva@cpr.ku.dk

Statistical methods in bioinformatics

University of Copenhagen

May 1st, 2025

Who am I?

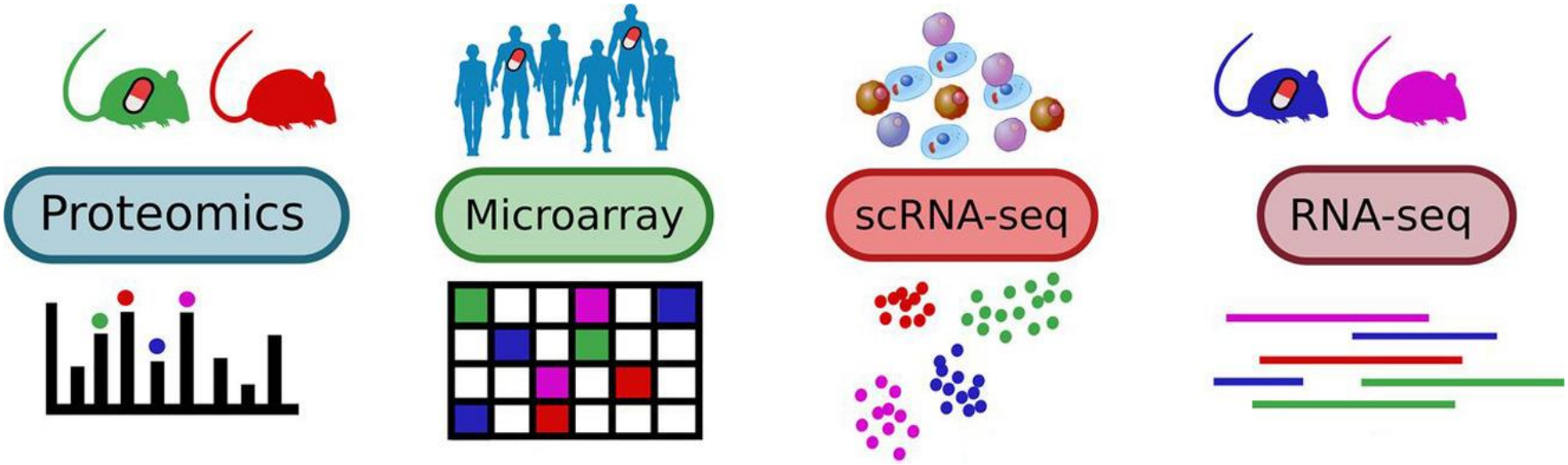
- Originally from Sofia, Bulgaria

- MSc in Bioinformatics at Saarland University (first Cytoscape app in 2007)
- PhD at Max Planck Institute for Informatics with external stay at University of California, San Francisco (meeting the Cytoscape team in 2013)

- Postdoc / Assistant professor at NNF Center for Protein Research, University of Copenhagen (working with STRING, Cytoscape & omics data since 2016)

Why networks?

High-throughput technologies

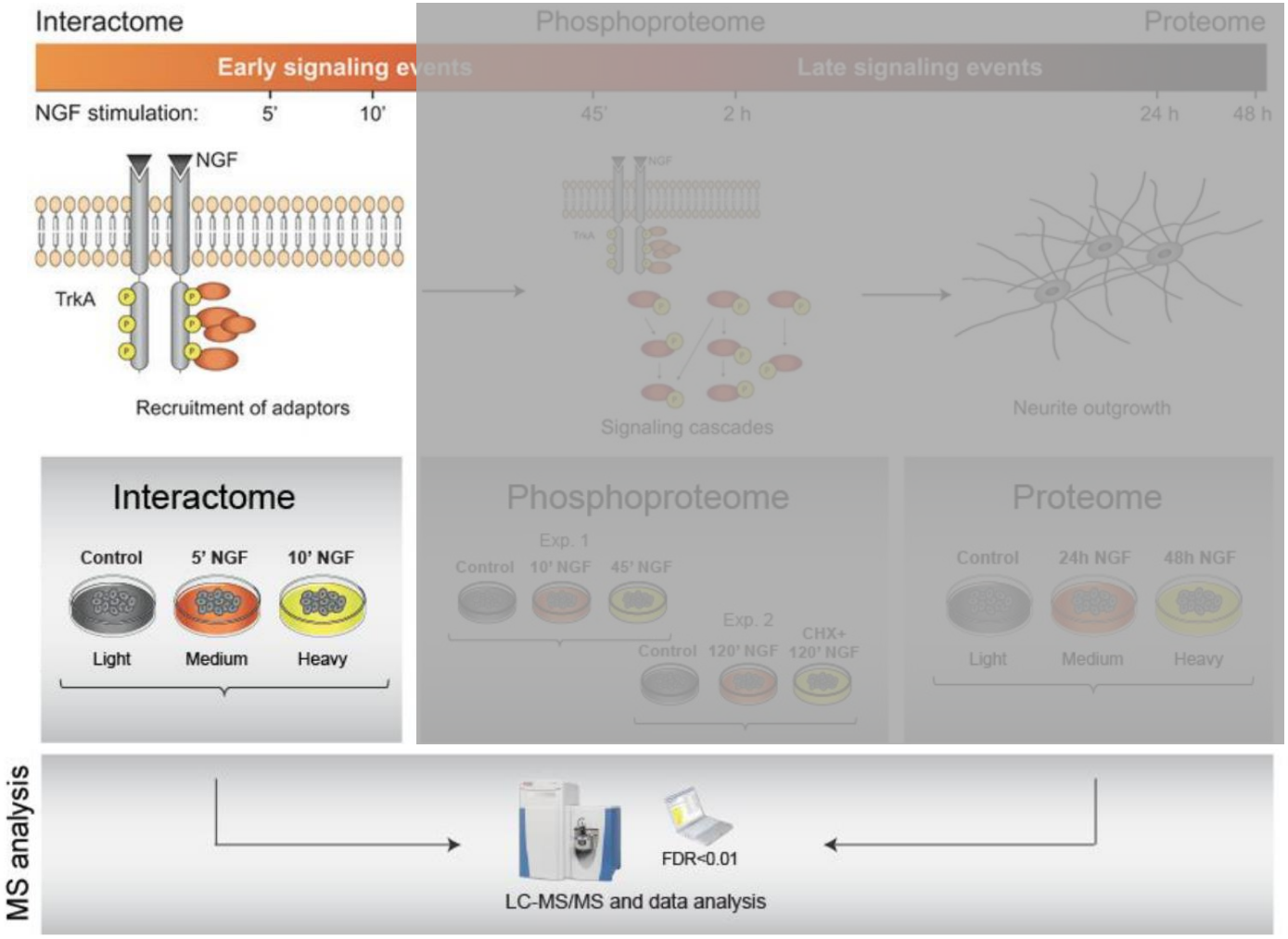


- Enable us to characterize genome- and proteome-wide expression changes
- Usually result in hundreds of regulated molecular players (genes, proteins, etc.)
- It is challenging to derive relevant biological insights from ‘omics data

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

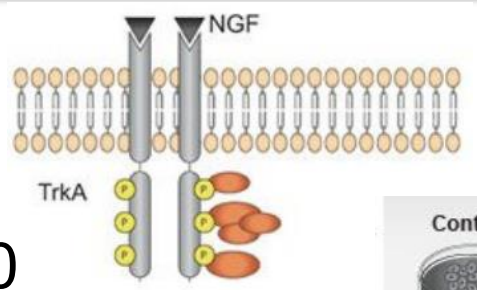
A typical proteomics dataset

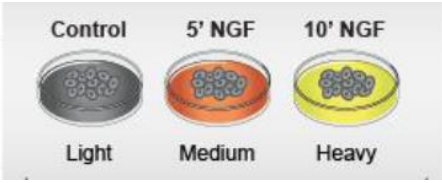
Temporal analysis of neuroblastoma cells in response to nerve growth factor (NGF) by mass spectrometry (Emdal *et al.*, *Science Signaling,* 2015).



A typical proteomics dataset

They identified 78 proteins that interact with TrkA (tropomyosin-related kinase A) after 5 min or 10 min of NGF stimulation.





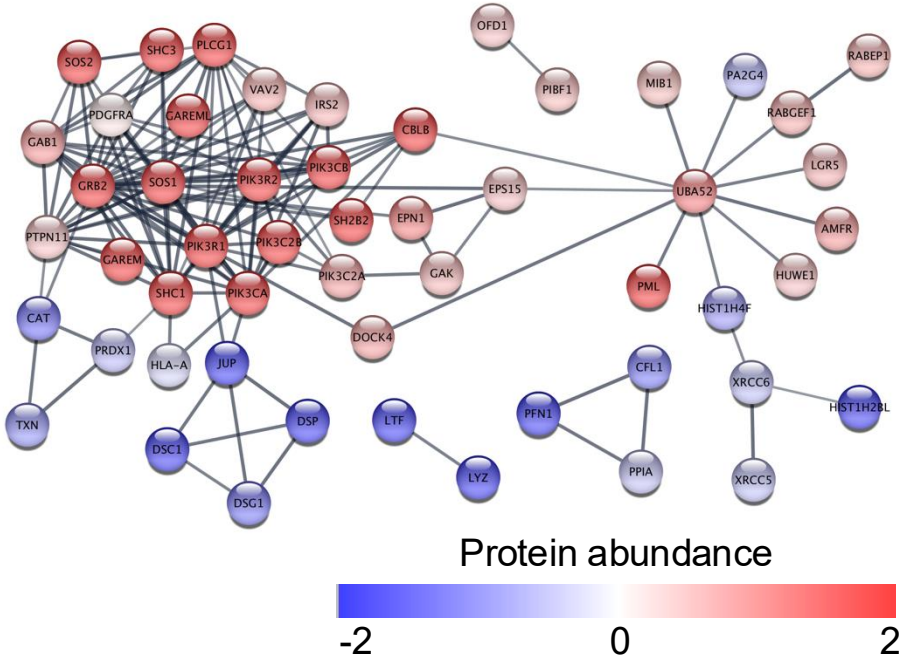
	A	B	C	D	G	J
1	**UniProt**	**Gene name**	**Peptides**	**Sequence coverage [%]**	**5 min log ratio**	**10 min log ratio**
2	Q99880	HIST1H2BL	5	35.7	-2.66	-2.66
3	Q8TER5	ARHGEF40	34	28.3	1.95	1.56
4	Q8IZ07	ANKRD13A	12	19.2	1.07	1.08
5	P62805	HIST1H4A	11	57.3	-2.31	-1.39
6	Q08380	LGALS3BP	14	28.2	-3.16	-2.98
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

$$\text{5 min log ratio} = \log_2 (\text{abundance}_{\text{5 min}} / \text{abundance}_{\text{control}})$$

From gene lists to networks

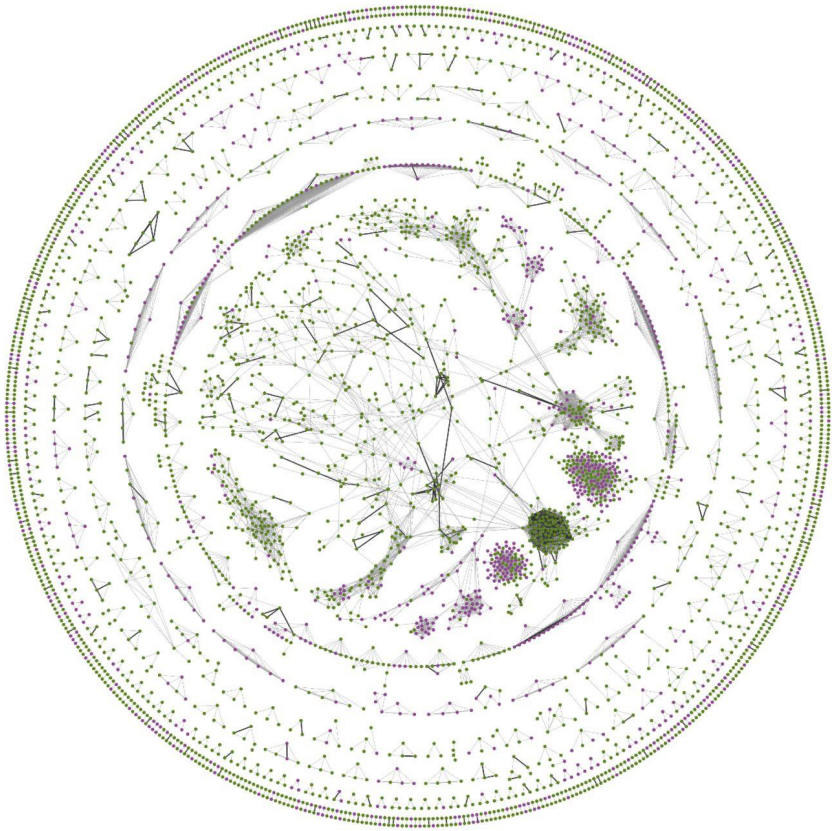
UniProt
Q99880
Q8TER5
Q8IZ07
P62805
Q08380
O00750
O00443
Q9UJ41
Q8TC07

Network analysis & visualization:
Identify target genes and their cellular context, e.g. STRING

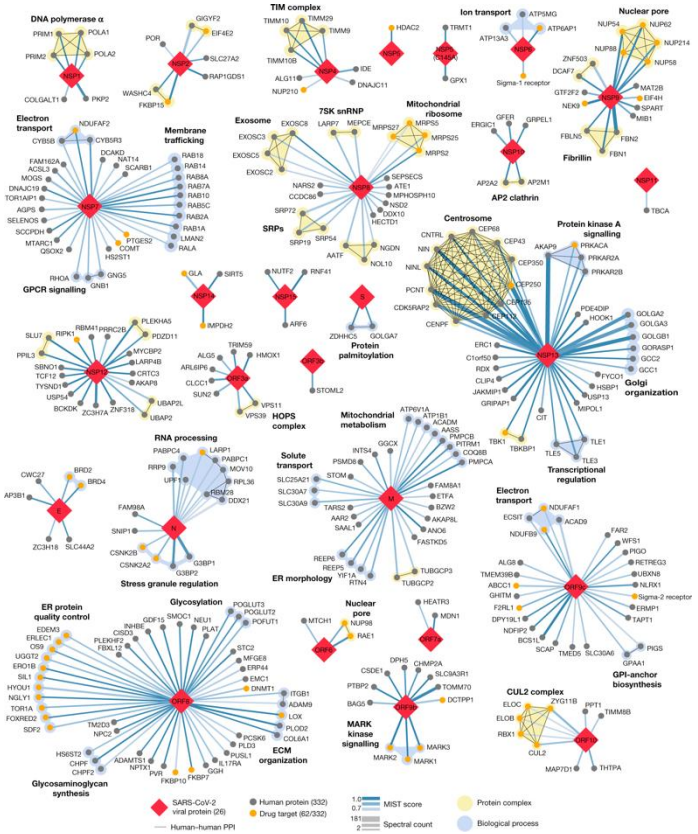


Protein interaction network with proteomics data visualized on the nodes

Networks are a **useful and intuitive abstraction** of complex biological systems that **lends itself to visualization**!



Koutrouli *et al.* (2024): Learning about understudied proteins through co-expression. *Bioinformatics*, 40.



Gordon *et al.* (2020): A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583.

Agenda

08:15 Introduction to biological networks + *exercise*

09:15 STRING & functional enrichment + *exercise*

10:15 Introduction to Cytoscape & stringApp + *exercise*

11:15 *stringApp demo & exercises*

12:00 Lunch break

13:00 Network visualization and analysis + *exercise*

14:00 *Hands-on exercises / Work with your own data*

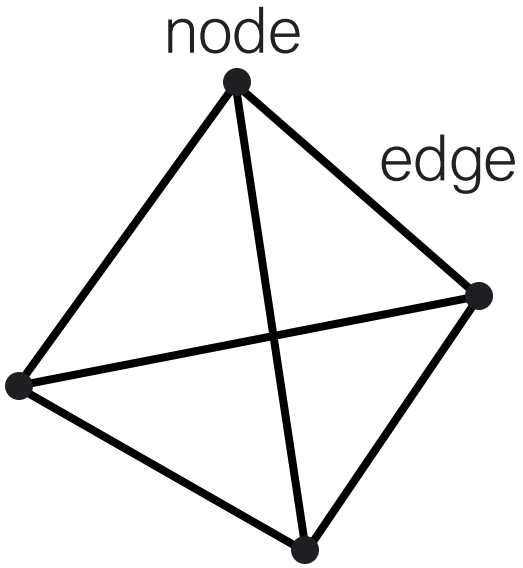
Materials: https://tinyurl.com/netbio2025

Intended learning outcomes

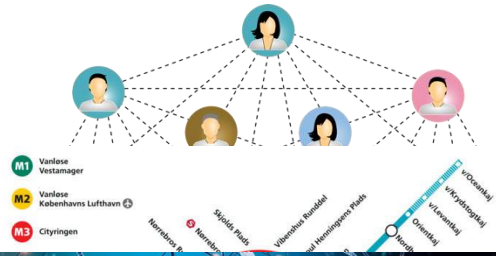
- Describe biological networks and give examples for the most common representatives
- Name and describe the sources of information integrated in the STRING database
- Perform functional enrichment analysis
- Analyze omics data using Cytoscape & stringApp
  - Import your data into Cytoscape using stringApp
  - Master network layouts and data visualization
  - Perform clustering and enrichment analyses
- Know where to find relevant documentation and tutorials

What are networks?

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



- Networks are everywhere…
  - Social networks (Facebook, LinkedIn)
  - Public transportation system
  - Nervous system
  - *… and many more*

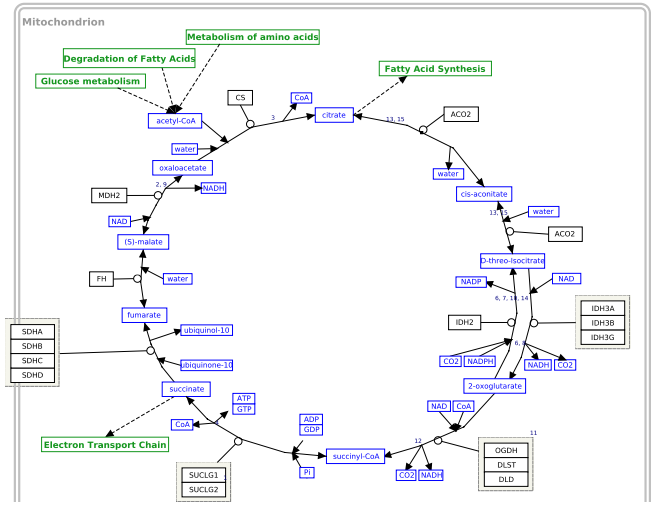


Biological networks

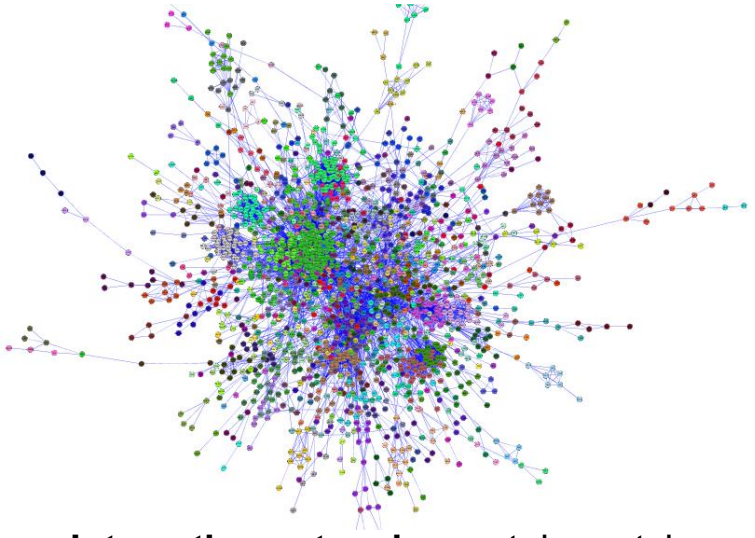
➔ Important to understand what the nodes and edges mean!

- **Nodes** can represent proteins, genes, metabolites, diseases, etc.

- **Edges** represent some kind of relationship between the nodes
  - Protein-protein interactions
  - Protein-ligand interactions
  - Metabolic reactions
  - Diseases comorbidities
  - Gene-disease associations

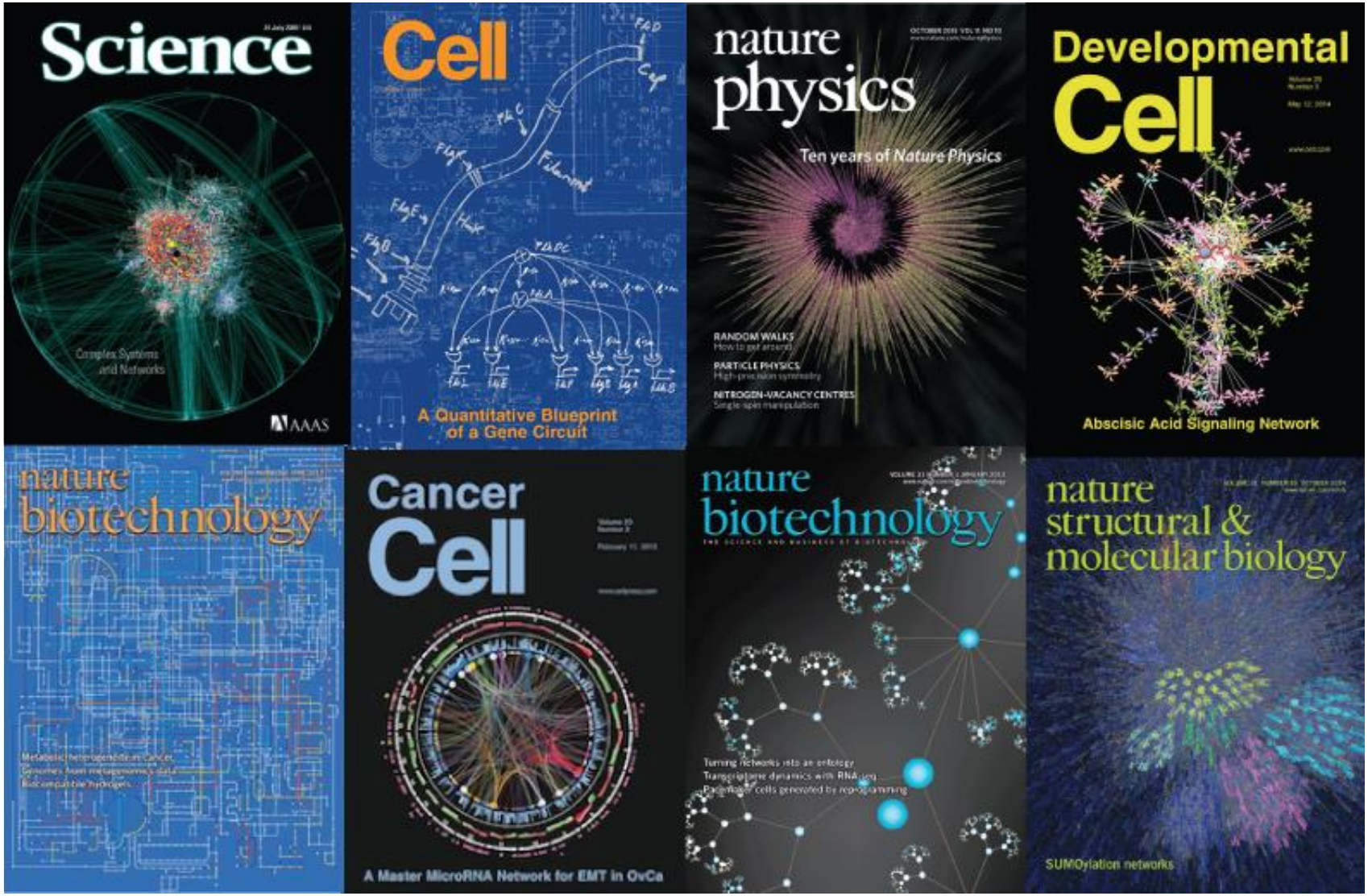


Pathways: metabolic, signaling, regulatory, such as KEGG



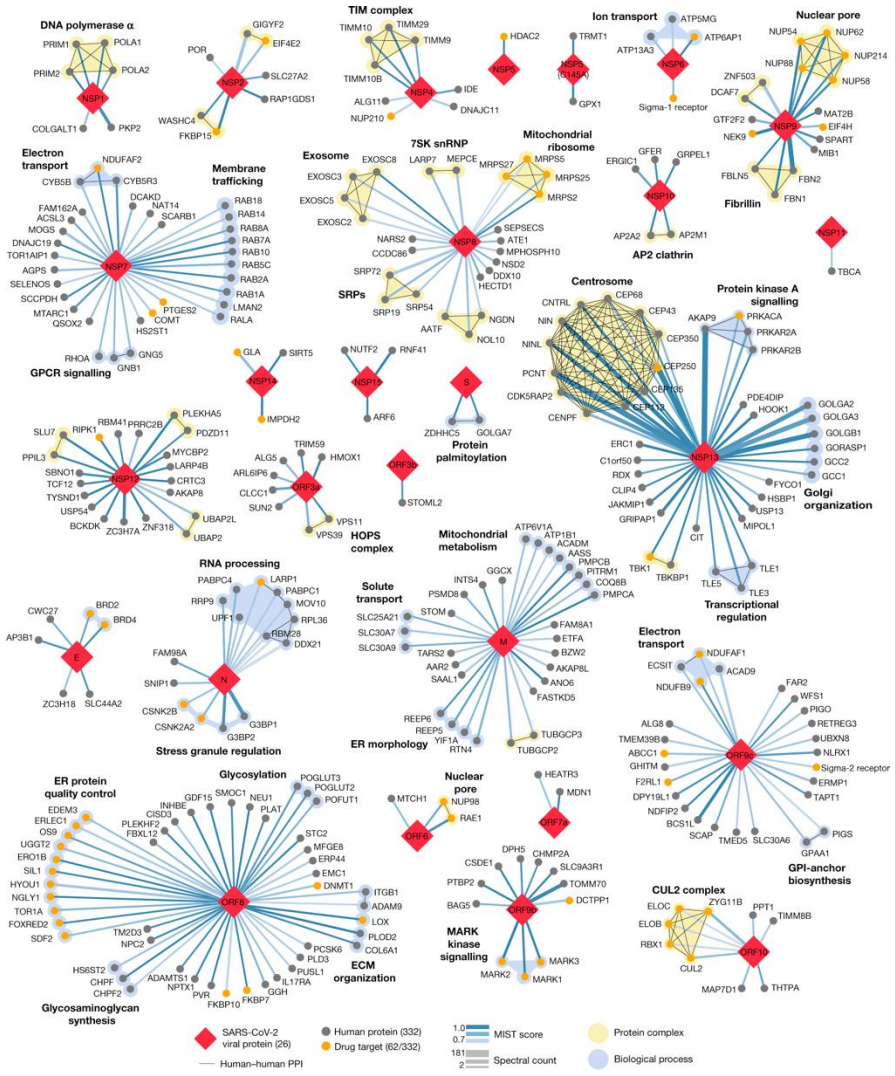
Interaction networks: protein-protein, protein-drug, such as STRING

Applications in Research



SARS-CoV-2-human network

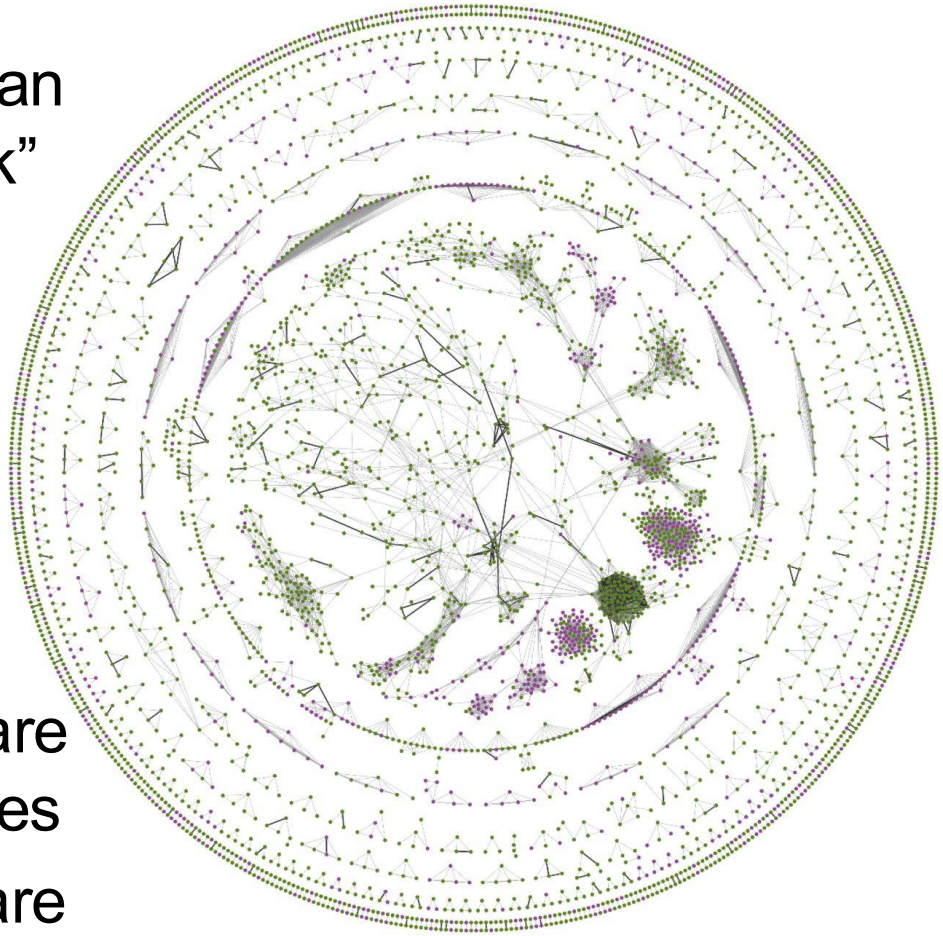
- AP-MS with 26 SARS-CoV-2 proteins reveals 332 interactions with human proteins
- Merged human-human physical protein interactions to identify complexes
- Used fill color to highlight known drug targets



Gordon *et al.* (2020): A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583, Fig. 3

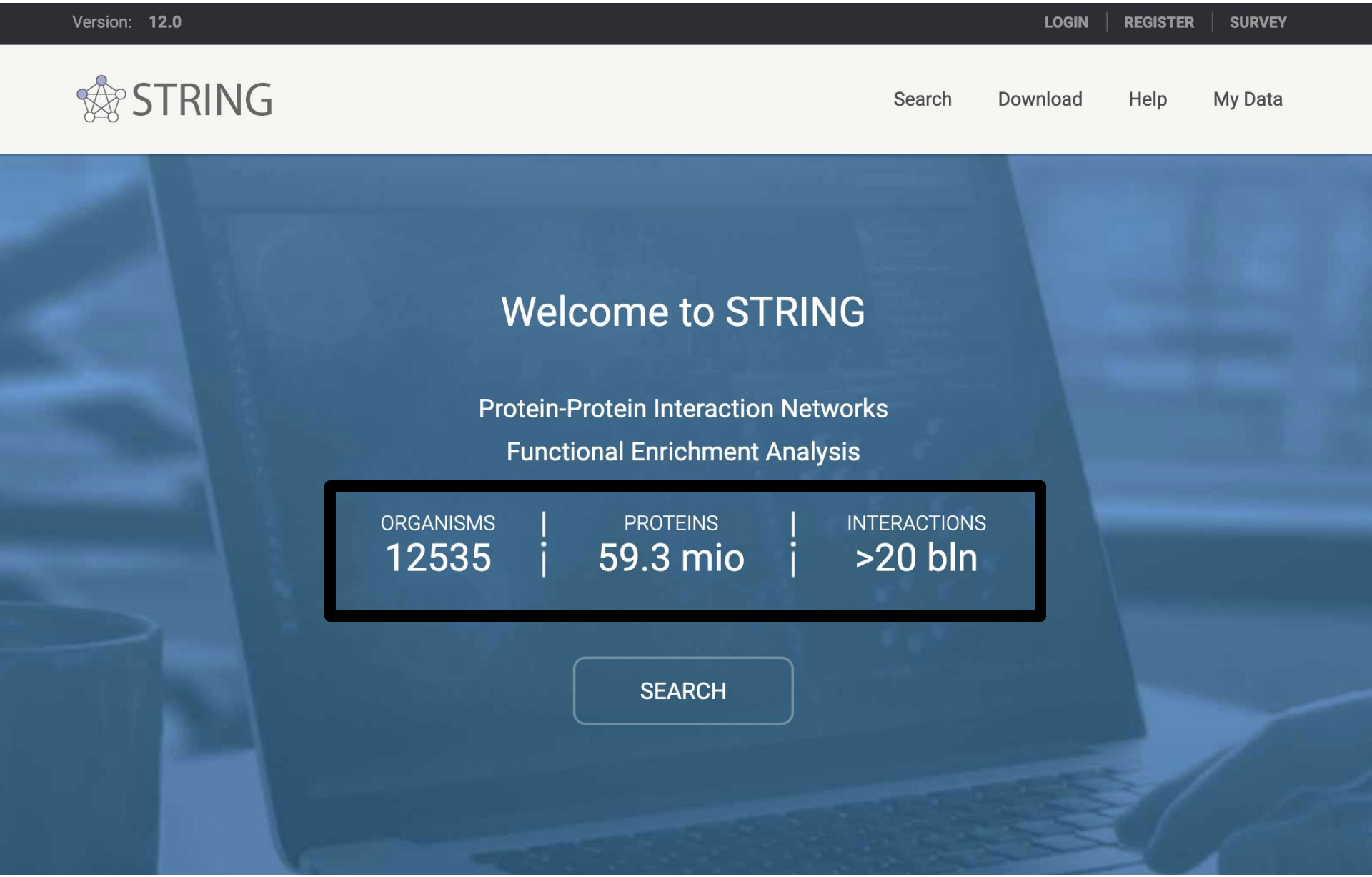
The bright side in the dark

- 30% of the genes in the human genome are considered "dark"
- FAVA: High-quality functional association network inferred from scRNA-seq and proteomics data
- *Better studied* proteins are green, *understudied* - purple, known physical interactions are shown as darker, thicker edges
- 1039 *understudied* proteins are connected to 611 *better studied* ones by predicted interactions.



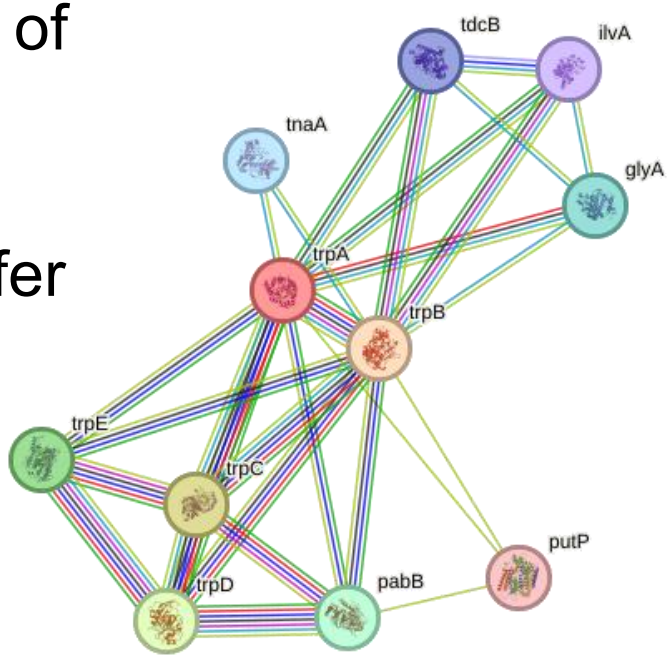
Koutrouli *et al.* (2024) Understudied proteins in the FAVA network. *Bioinformatics*, Fig. 5

Do you already have some ideas, if and how you can use networks in your project(s)?


Version: 12.0
LOGIN | REGISTER | SURVEY
STRING
Search
Download
Help
My Data
Welcome to STRING
Protein-Protein Interaction Networks
Functional Enrichment Analysis
ORGANISMS
12535
PROTEINS
59.3 mio
INTERACTIONS
>20 bln
SEARCH

STRING

- Collect and integrate multiple types of evidence for known protein-protein associations
- Predict new associations and transfer across species
- Assign confidence score to each association



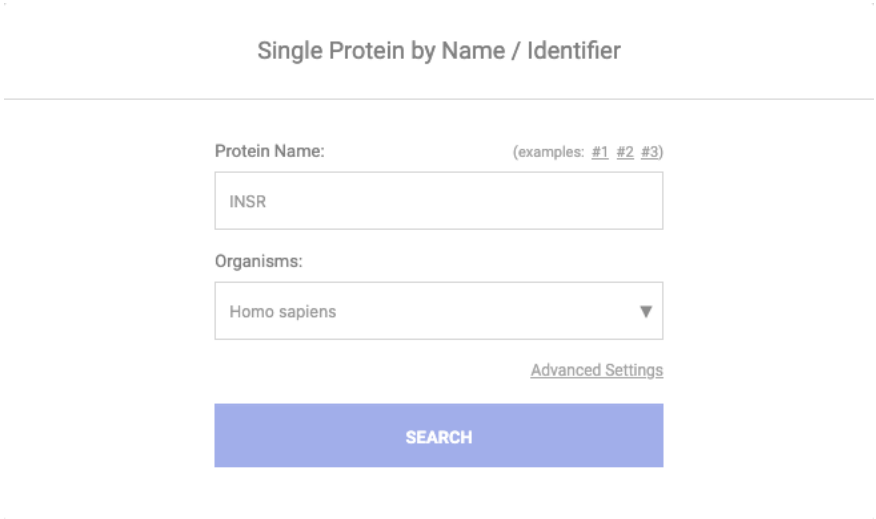
Query for protein trpA

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

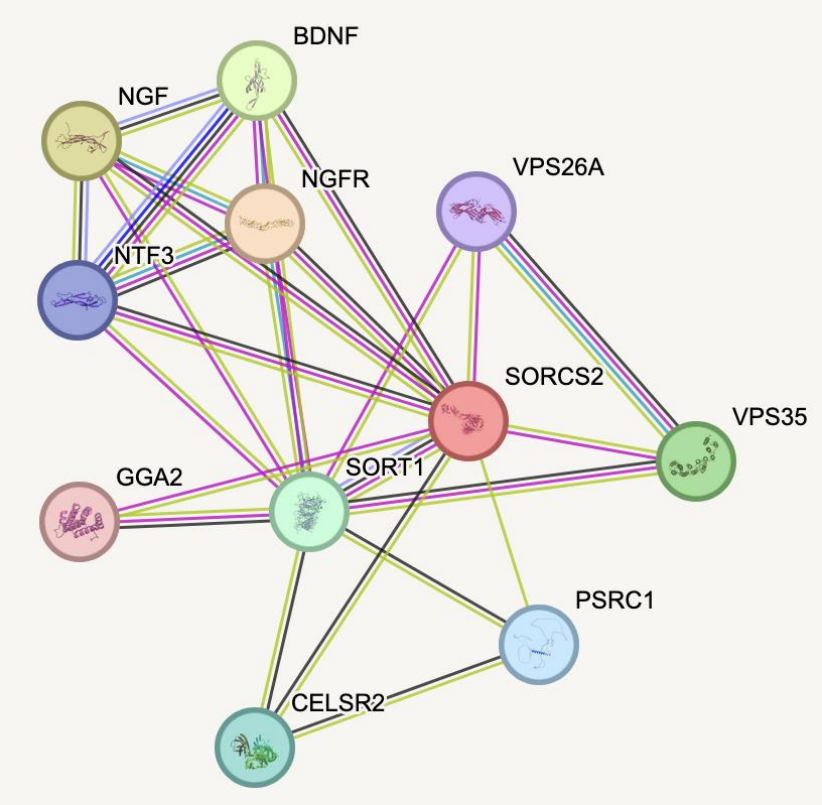
STRING exercise 1

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

- Query the database
- Inspect the evidence
- Change query parameters



STRING evidence channels



Known Interactions

- *from curated databases*
- *experimentally determined*

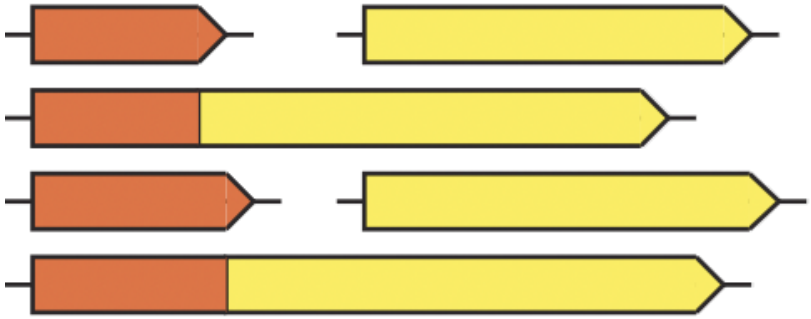
Predicted Interactions

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

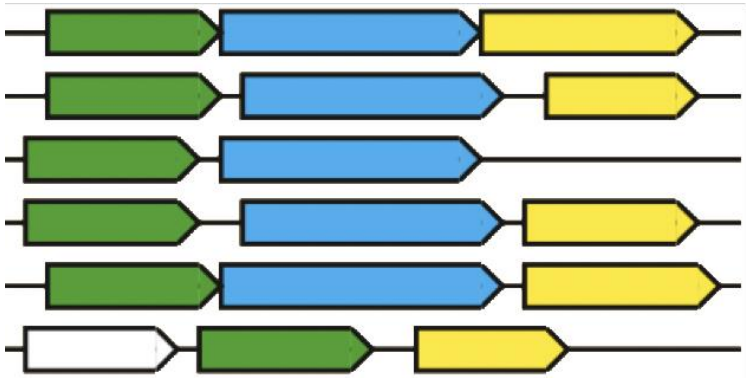
Others

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

Predictions from genomic context

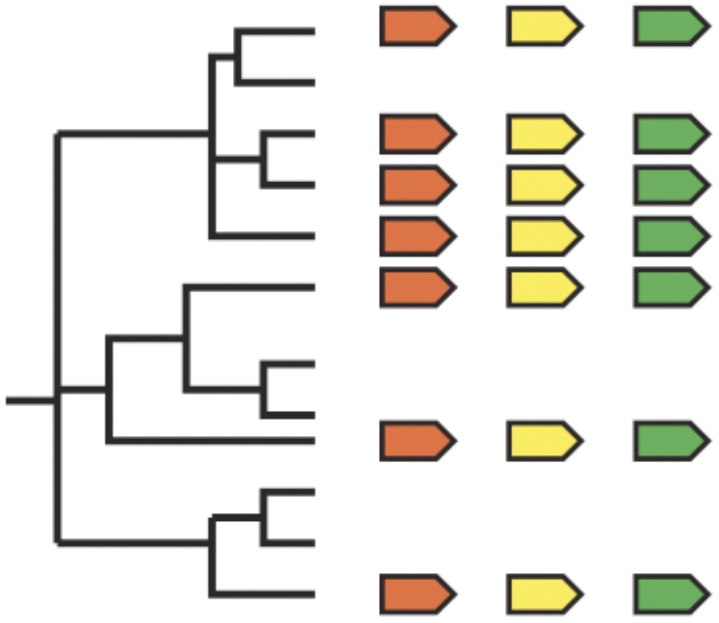


Gene fusion



Gene neighborhood

12,535 organisms

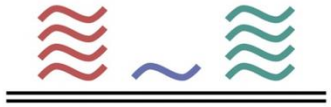


Gene co-occurrence
(phylogenetic profiles)

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

Experimental evidence

Experiments



Co-expression

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

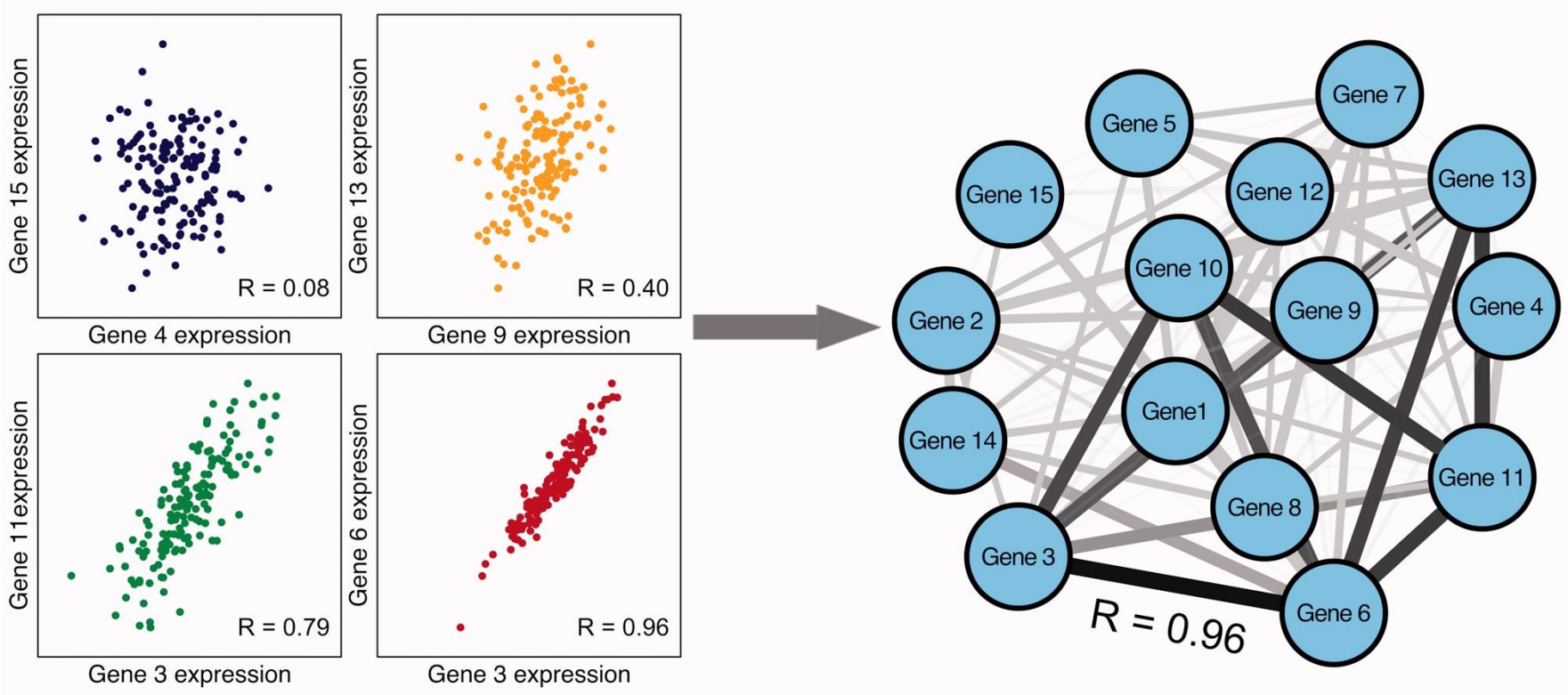
Look for consistent similarities between expression profiles in many different conditions

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

Includes both RNA-based expression data as well as protein expression

Molecule A	Molecule B	Identifier A	Identifier B	Type A	Type B	Species A	Species B	Host Organism	Positive Interaction	Detection Method	Publication	Interaction Type
MDM2	TP53	UniProt Q00987	UniProt P04637	protein	protein	Homo sapiens	Homo sapiens	In vitro	✓	pull down	10.1038/emboj.2008.25 18309296	physical association
MDM2	TP53	UniProt Q00987	UniProt P04637	protein	protein	Homo sapiens	Homo sapiens	In vitro	✓	molecular sieving	10.1038/emboj.2008.25 18309296	association
MDM2	TP53	UniProt Q00987	UniProt P04637	protein	protein	Homo sapiens	Homo sapiens	In vitro	✓	pull down	18485870	direct interaction

Gene co-expression networks



Commonly used measures: Pearson or Spearman correlation

Issues:

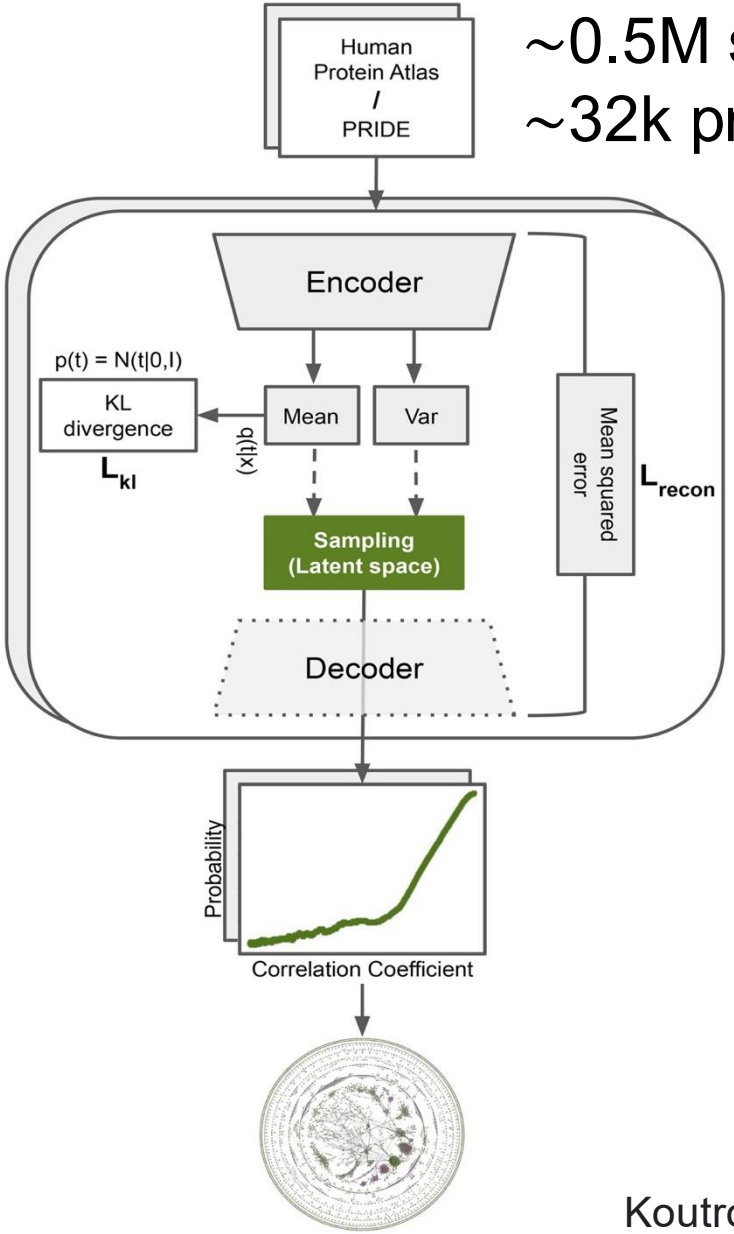
- Data sparsity and redundancy
- Information is not always linear

van Dam *et. al., Briefings in Bioinf*, 2017

FAVA

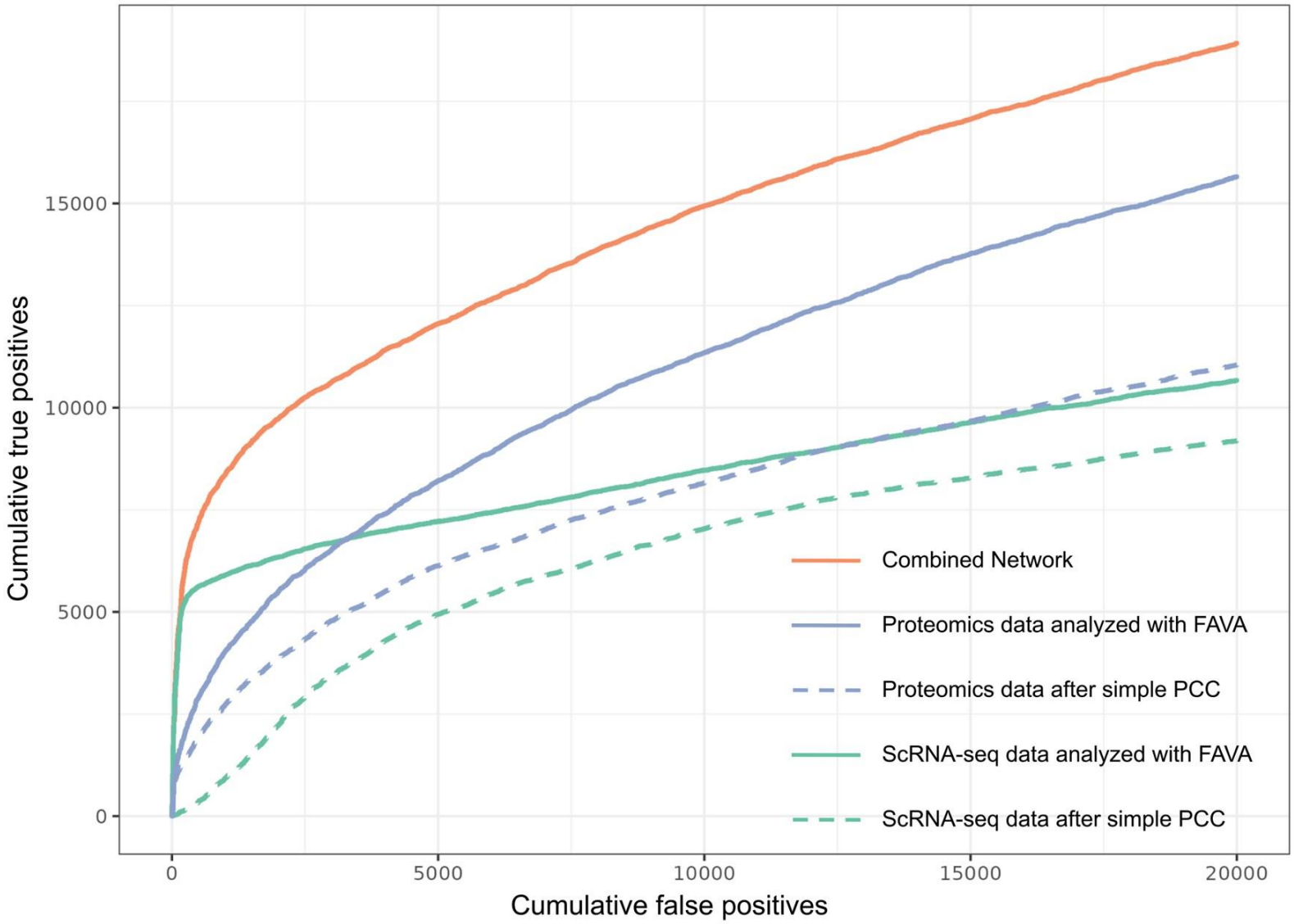
~0.5M single cells and
~32k proteomics studies

Dimensionality reduction using a variational auto-encoder



Koutrouli *et al.*, *Bioinformatics*, 2024

FAVA performance



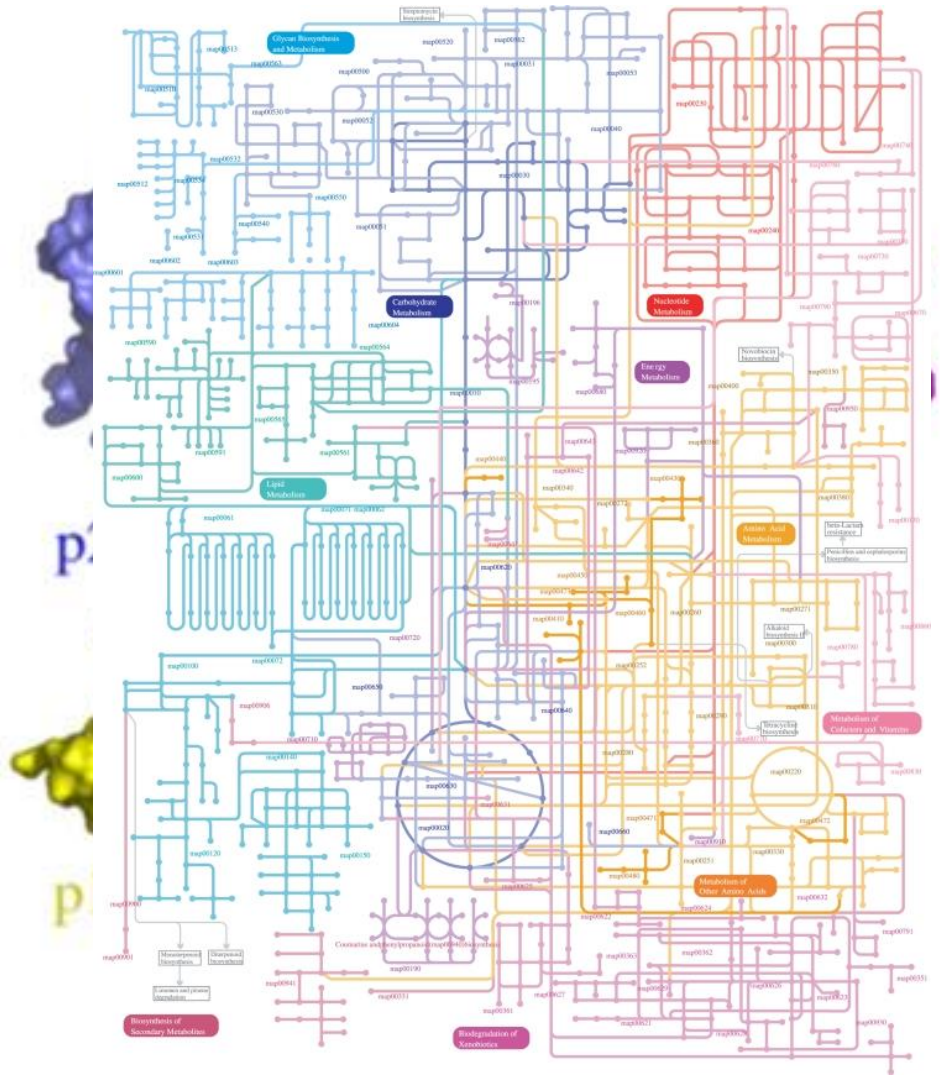
Koutrouli *et al.*, *Bioinformatics*, 2024

“Higher-level” knowledge

Also known as *Databases*

Curated pathway databases like KEGG & Reactome

Known protein complexes



“Higher-level” knowledge

Knowledge

Text mining

Also known as *Databases*

Curated pathway databases like KEGG & Reactome

Known protein complexes

Co-occurrence text mining for functional associations

Natural language processing using deep learning methods for physical interactions

STRING confidence scores

Your Input:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
SORCS2	*VPS10 domain-containing receptor SorCS2; The heterodimer formed by NGFR and SORCS2 functions as receptor for the precursor forms of NGF (proNGF) and BDNF (proBDNF). ProNGF and proBDNF binding both promote axon growth cone collapse (in vitro). Plays a role in the regulation of dendritic spine density in hippocampus neurons (By similarity). Required for normal neurite branching and extension in response to BDNF. Plays a role in BDNF-dependent hippocampal synaptic plasticity. Together with NGFR and NTRK2, is required both for BDNF-mediated synaptic long-term depression and long-term poten [...] (1159 aa)*									

Predicted Functional Partners:

NGFR	*Tumor necrosis factor receptor superfamily member 16; Low affinity receptor which can bind to NGF, BDNF, NTF3, and NTF4. F...*					●		●		0.939
NGF	*Beta-nerve growth factor; Nerve growth factor is important for the development and maintenance of the sympathetic and sens...*					●		●		0.927
BDNF	*Brain-derived neurotrophic factor; Important signaling molecule that activates signaling cascades downstream of NTRK2. Duri...*					●		●		0.890
VPS35	*Vacuolar protein sorting-associated protein 35; Acts as component of the retromer cargo-selective complex (CSC). The CSC is...*					●		●		0.761
SORT1	*Sortilin; Functions as a sorting receptor in the Golgi compartment and as a clearance receptor on the cell surface. Required for...*					●		●	●	0.748
CELSR2	*Cadherin EGF LAG seven-pass G-type receptor 2; Receptor that may have an important role in cell/cell signaling during nervou...*							●		0.677
PSRC1	*Proline/serine-rich coiled-coil protein 1; Required for normal progression through mitosis. Required for normal congress of chr...*							●		0.670
NTF3	*Neurotrophin-3; Seems to promote the survival of visceral and proprioceptive sensory neurons; Belongs to the NGF-beta family.*					●				0.567
VPS26A	*Vacuolar protein sorting-associated protein 26A; Acts as component of the retromer cargo-selective complex (CSC). The CSC i...*					●		●		0.552
GGA2	*ADP-ribosylation factor-binding protein GGA2; Plays a role in protein sorting and trafficking between the trans-Golgi network (T...*					●		●		0.493

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

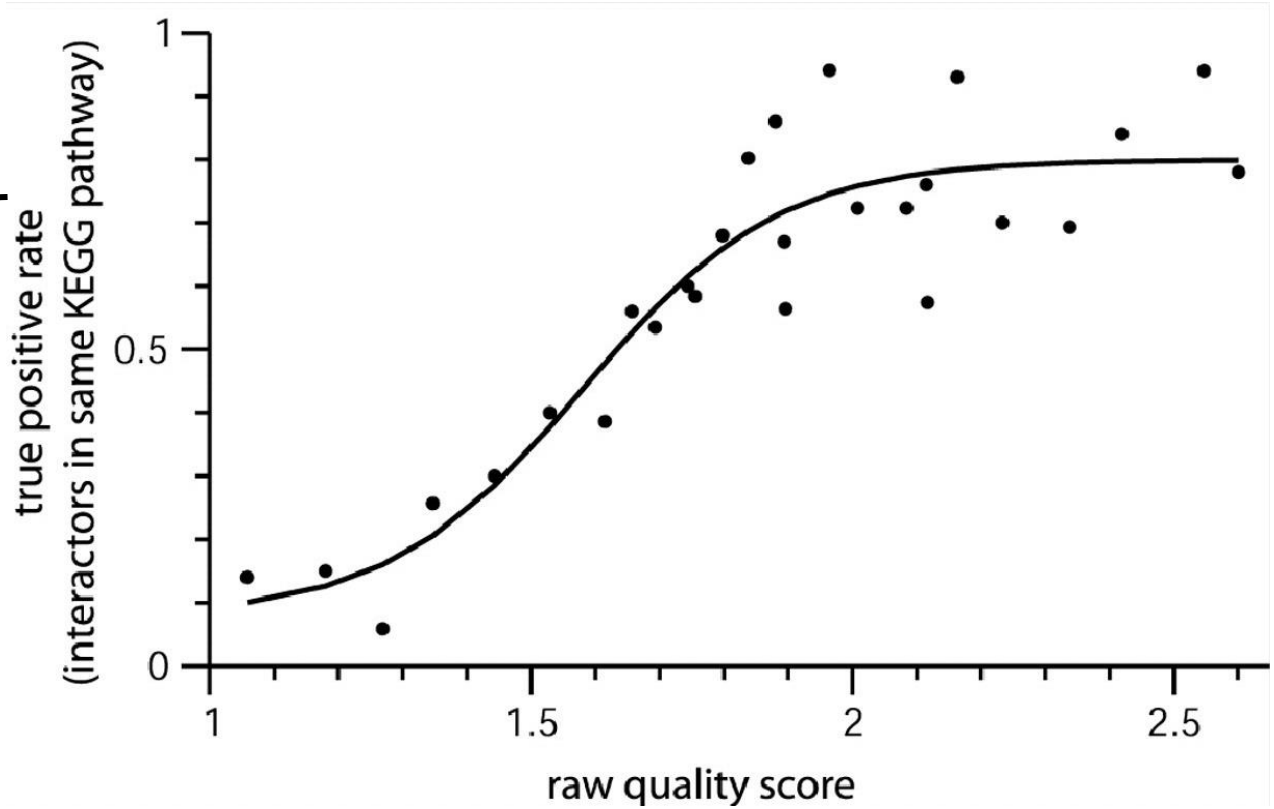
→ Parsers and mapping files

STRING score calibration

→ Quality scores [0,1] based on a gold standard

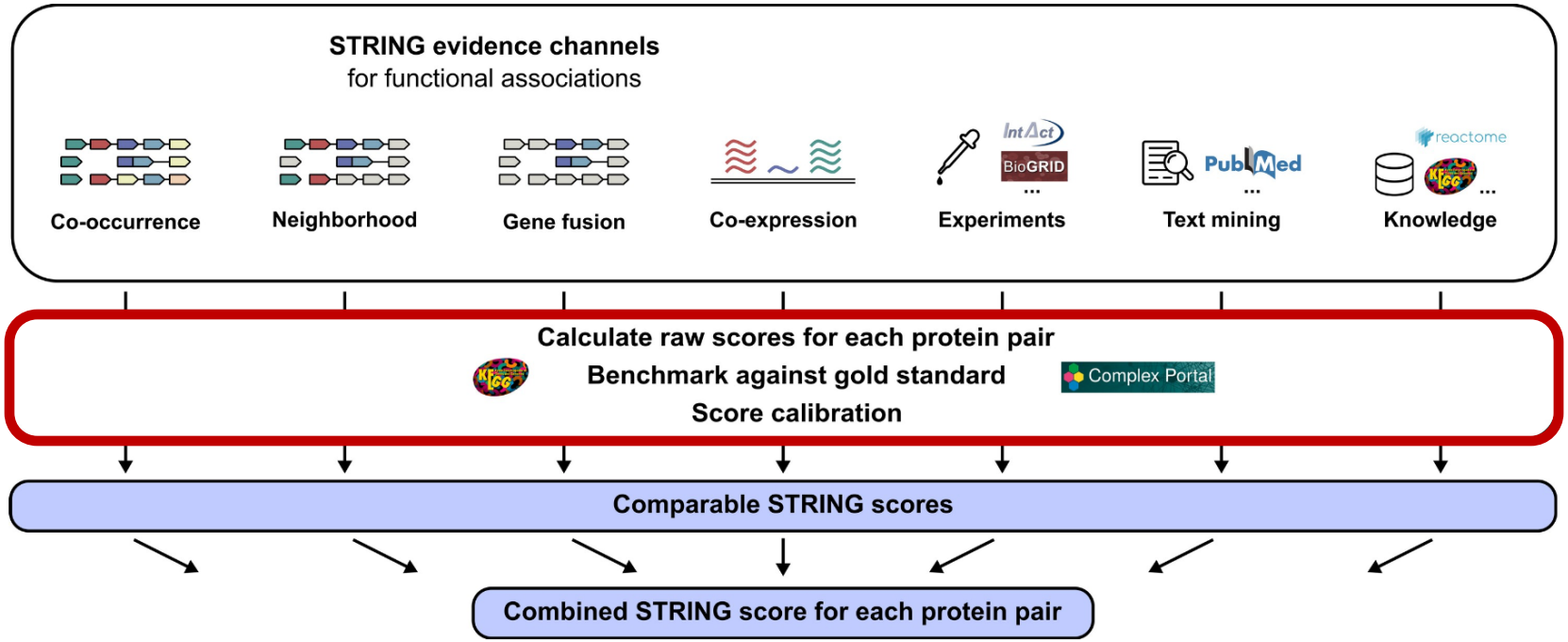
→ Common scale for comparison and implicit weight by quality

STRING score
(how likely are two proteins part of the same KEGG pathway or protein complex)



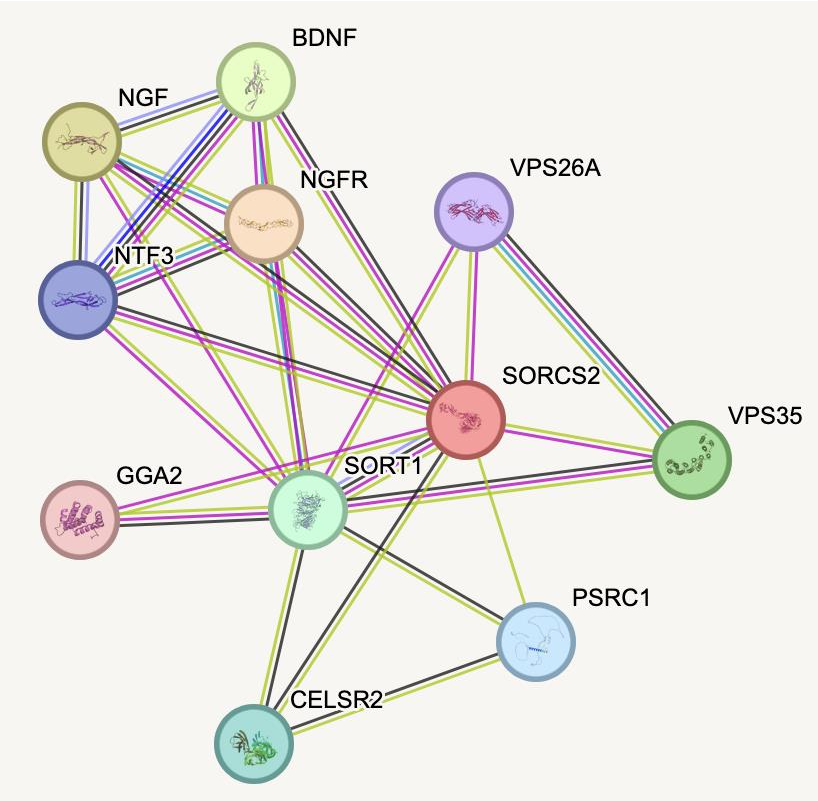
von Mering et al. *Nucleic Acids Research*, 2005

STRING confidence scores

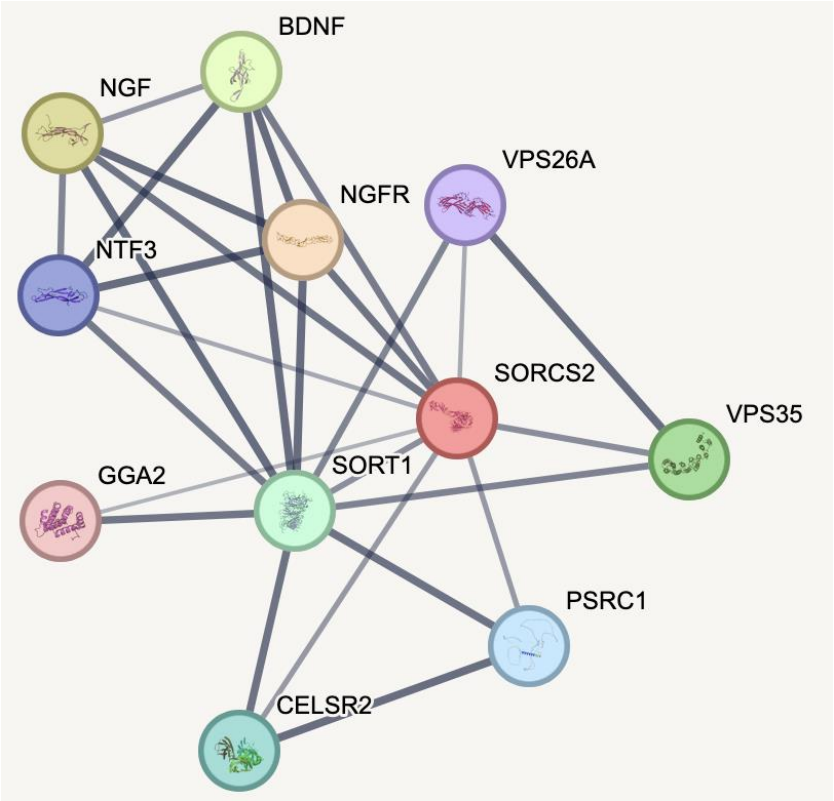


https://string-db.org/help//faq/#how-are-the-scores-computed

STRING network visualization



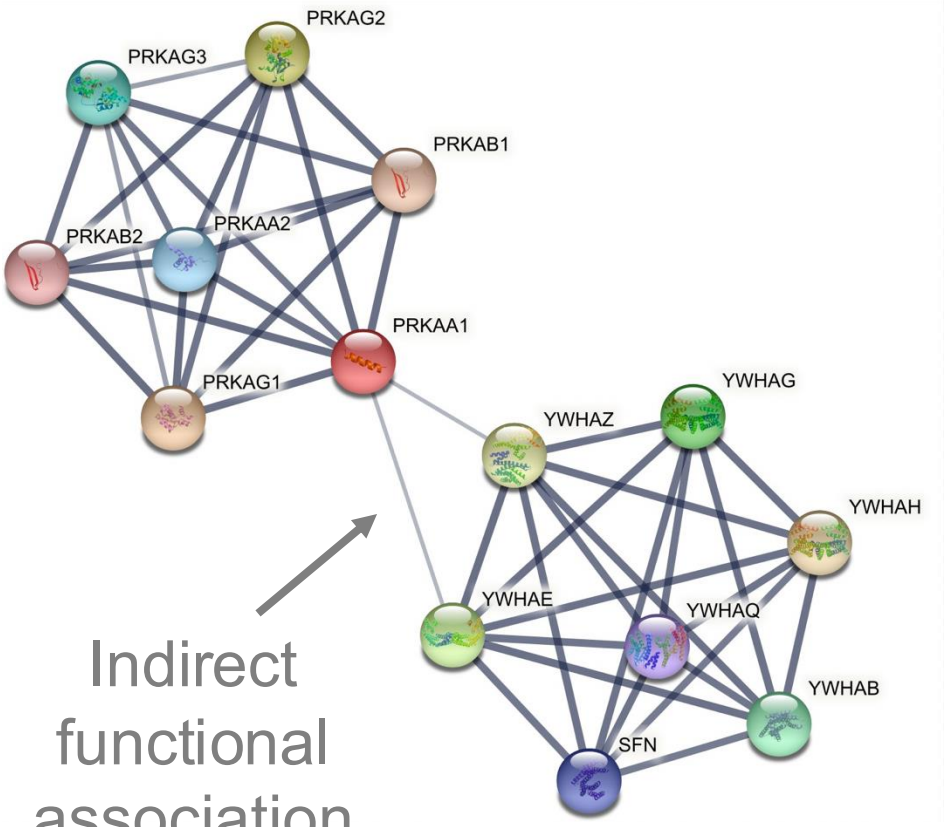
Evidence view



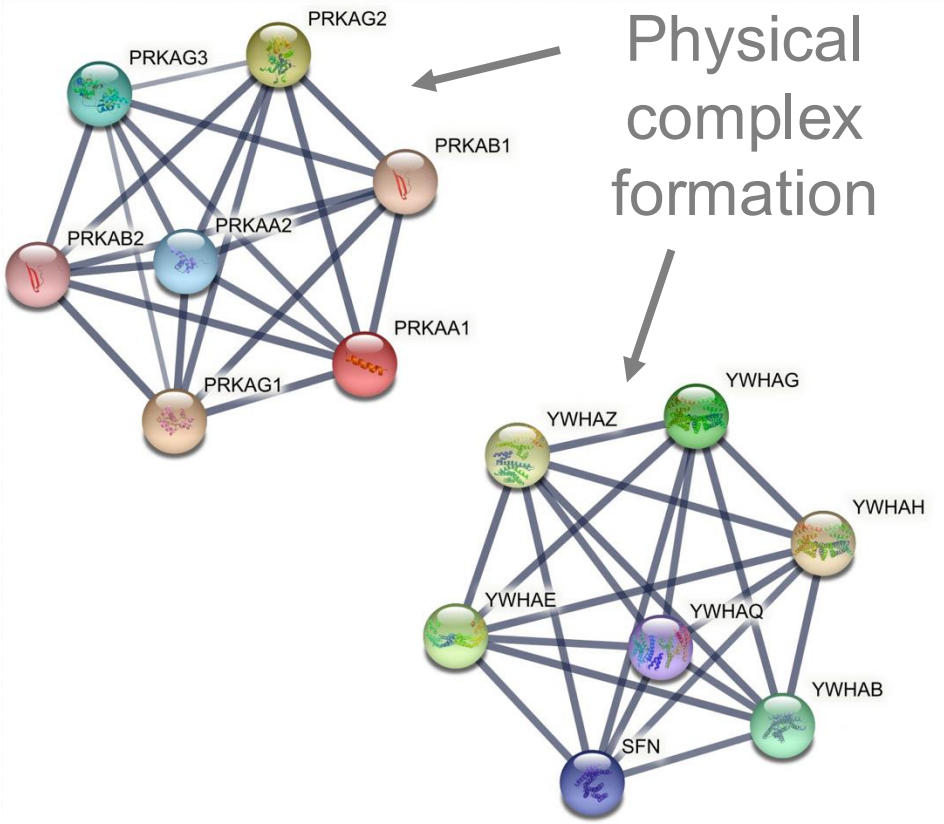
Confidence view

Type of interactions

Functional associations vs. physical interactions



Indirect functional association



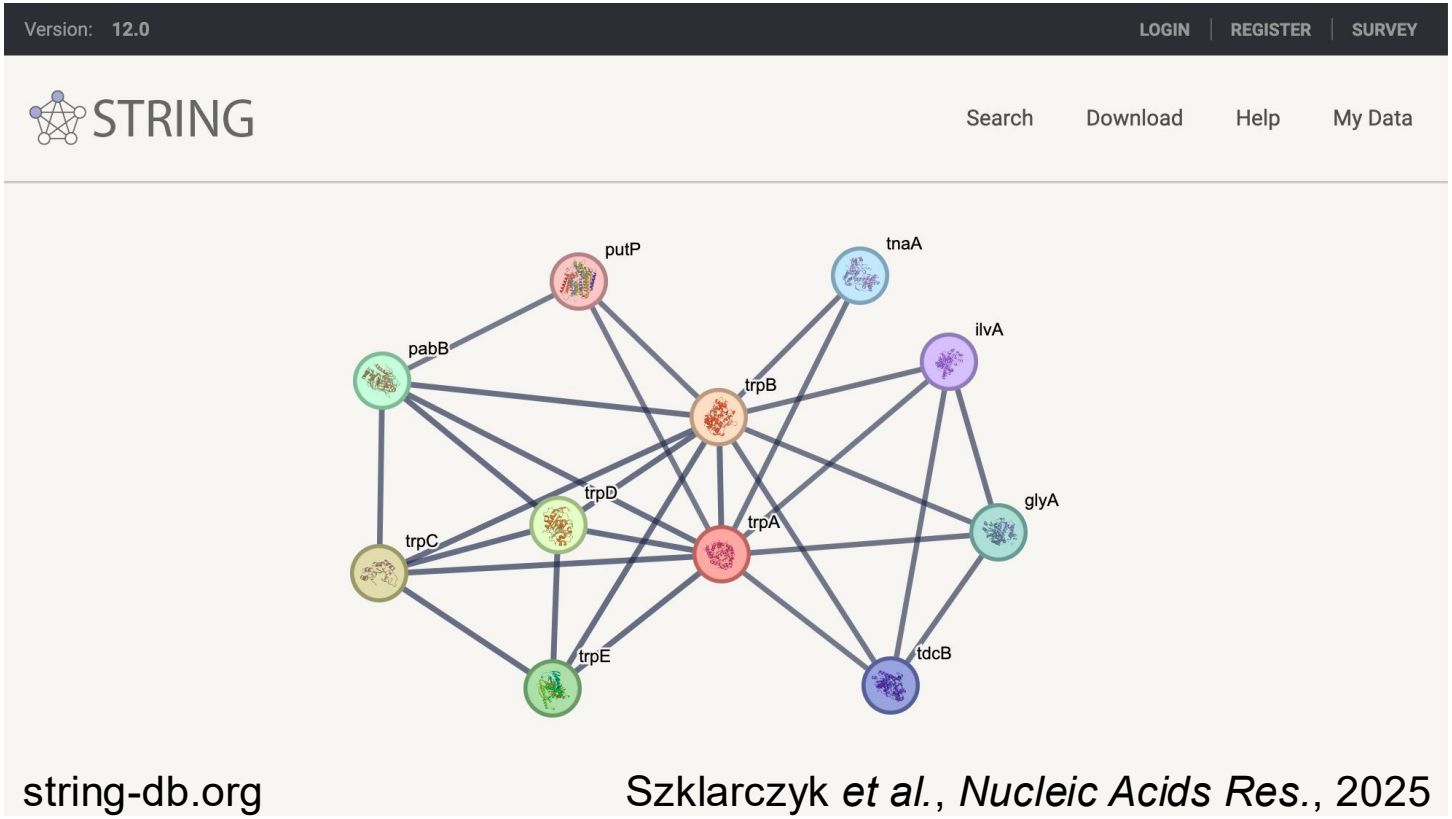
Physical complex formation

New since STRING v11.5

Access to STRING

- Web interface
- Evidence viewers
- Bulk download files

- Programmatic access
- Web services
- Cytoscape stringApp



string-db.org

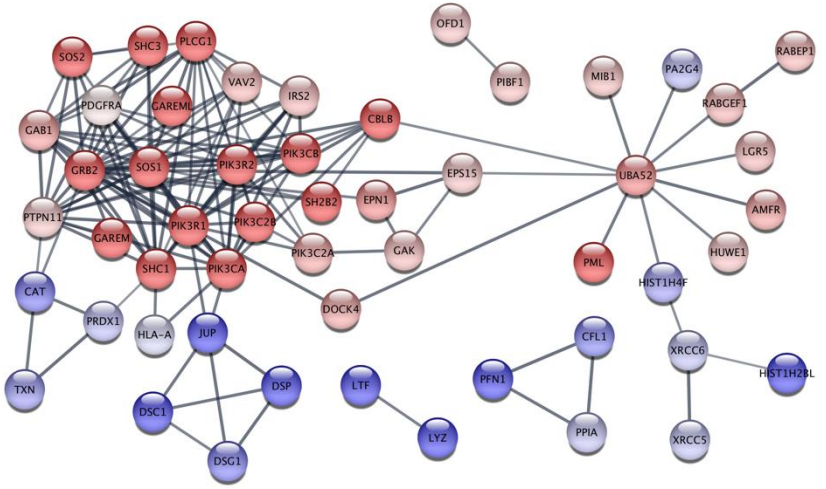
Szklarczyk *et al.*, *Nucleic Acids Res.*, 2025

Questions?

From gene lists to interpretation

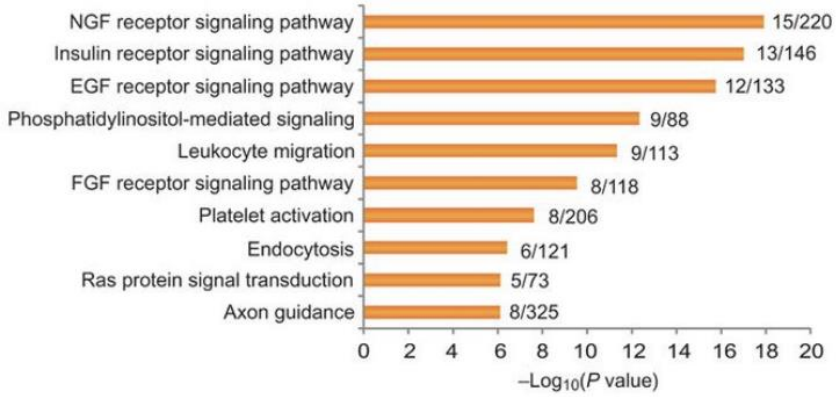
UniProt
Q99880
Q8TER5
Q8IZ07
P62805
Q08380
O00750
O00443
Q9UJ41
Q8TC07

Network analysis & visualization:
Identify target genes and their cellular context, e.g. STRING



Protein interaction network with proteomics data visualized on the nodes

Functional enrichment:
Identify relevant pathways & processes, e.g. KEGG & GO

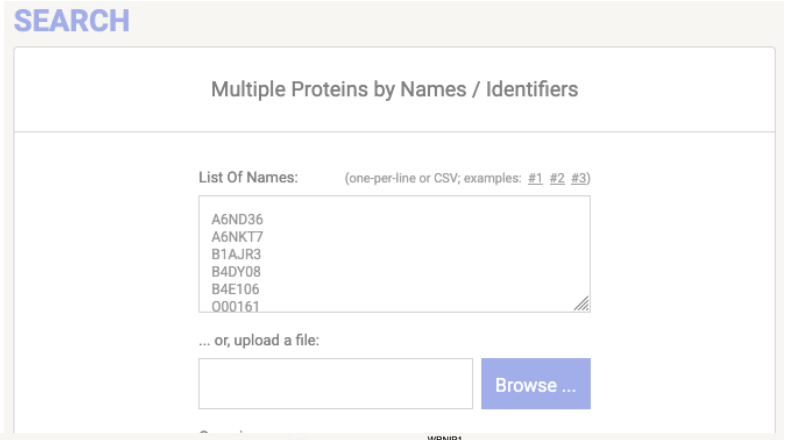


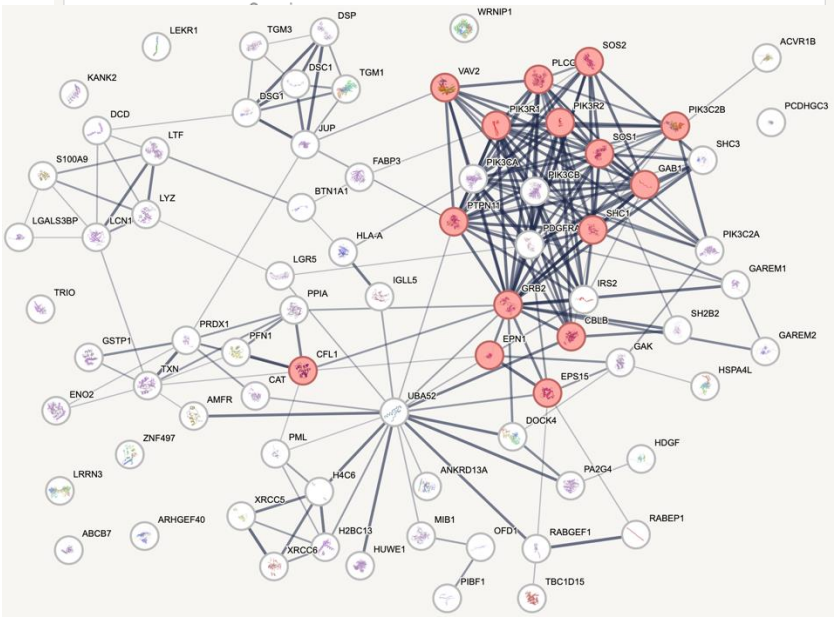
Significantly overrepresented GO terms for biological process

STRING exercise 2

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

- Data from a proteomics study
- Get network of multiple proteins
- Change query parameters
- Explore functional enrichment
- Show enrichment on network

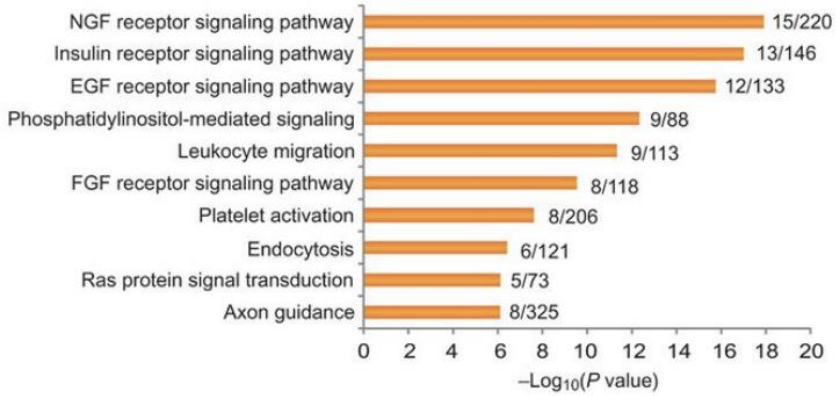




From gene lists to enrichment

UniProt
Q99880
Q8TER5
Q8IZ07
P62805
Q08380
O00750
O00443
Q9UJ41
Q8TC07

Functional enrichment:
Identify relevant pathways & processes, e.g. KEGG & GO



Significantly overrepresented GO terms for biological process

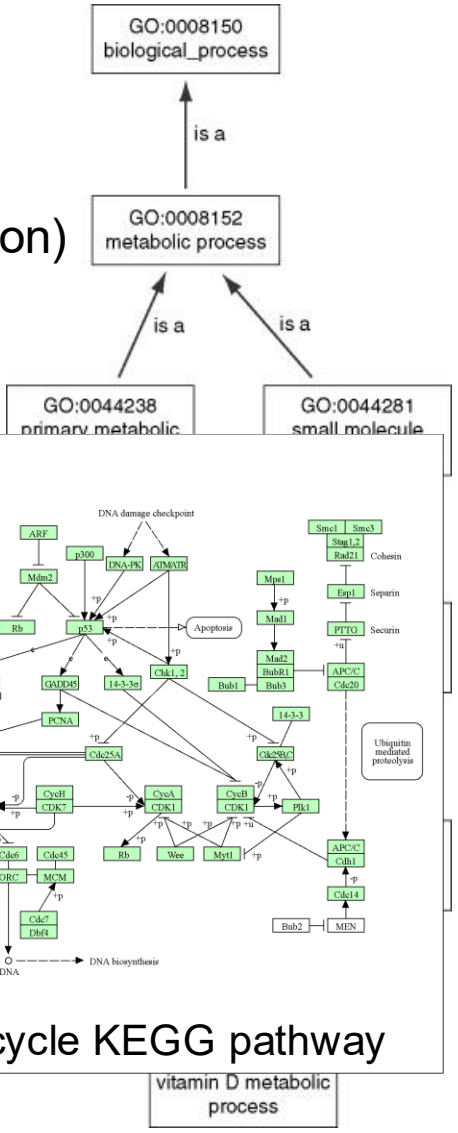
Enrichment analysis

Gene sets (pre-knowledge databases):

- Gene Ontology (GO)
  - Biological process (such as DNA repair, signal transduction)
  - Molecular function (such as catalysis, transport, binding)
  - Cellular component (such as mitochondria, ribosome)
- Pathways, e.g. KEGG or Reactome
- Genomic location
  - Chromosomal location
  - Enhancer regions/Transcription factors
  - Overlapping SNPs, etc.

Methodologies:

- Over-representation (enrichment) analysis
- Gene set enrichment analysis (GSEA)



Cell cycle KEGG pathway

Gene ontology example

Over-representation analysis

Given: a list of *regulated genes* or proteins and a list of *annotations (aka gene sets or enrichment terms)* such as GO biological processes

Goal: identify the *gene sets* that are statistically over-represented in the list of *regulated genes* compared to a *background* list of genes

How: use *Fisher's exact test* to calculate a *p-value* and correct for *multiple testing* to get a false discovery rate value

UniProt

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

O00750

Enrichment in STRING

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

Network Stats

number of n
number of e
average node d
avg. local clustering coeff

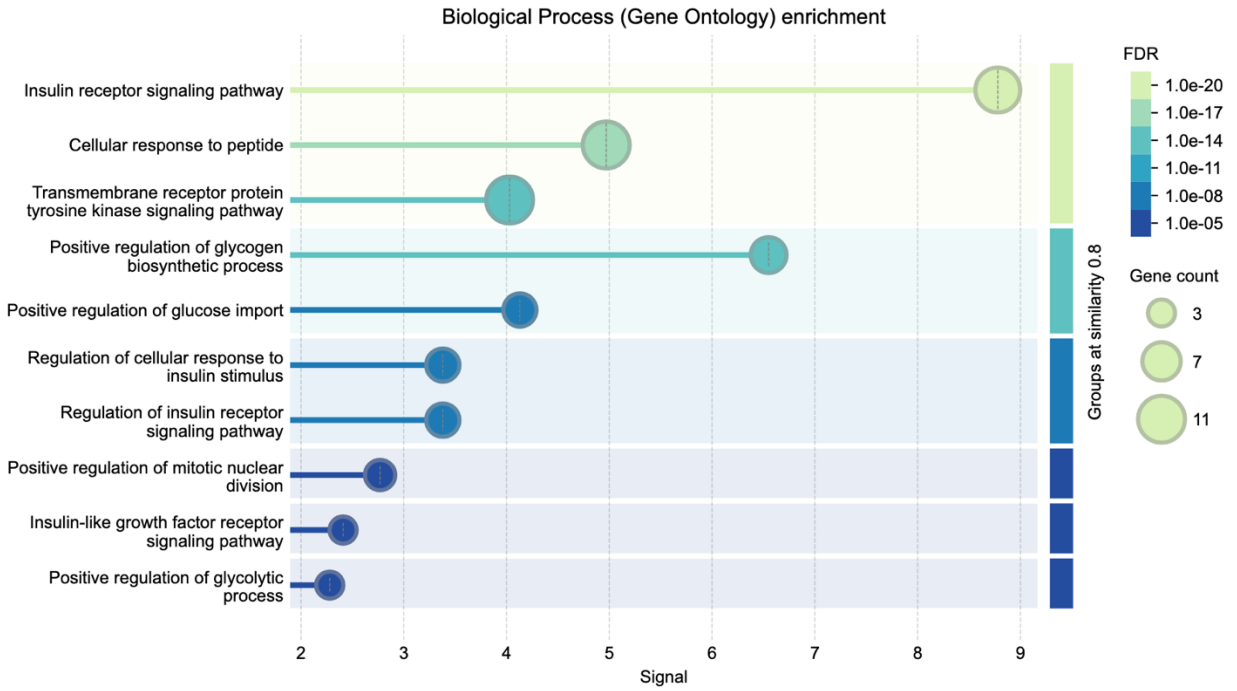
Functional enrichments in your net

Biological Process	
GO-term	*description*
GO:1902202	regulation of hepat
GO:0060267	positive regulation
GO:0045725	positive regulation
GO:1990535	neuron projection r
GO:0032000	positive regulation

Molecular Functio	
GO-term	*description*
GO:0043559	insulin binding
GO:0005159	insulin-like growth
GO:0005158	insulin receptor bir
GO:0043560	insulin receptor sul
GO:0031994	insulin-like growth

Functional enrichment visualization

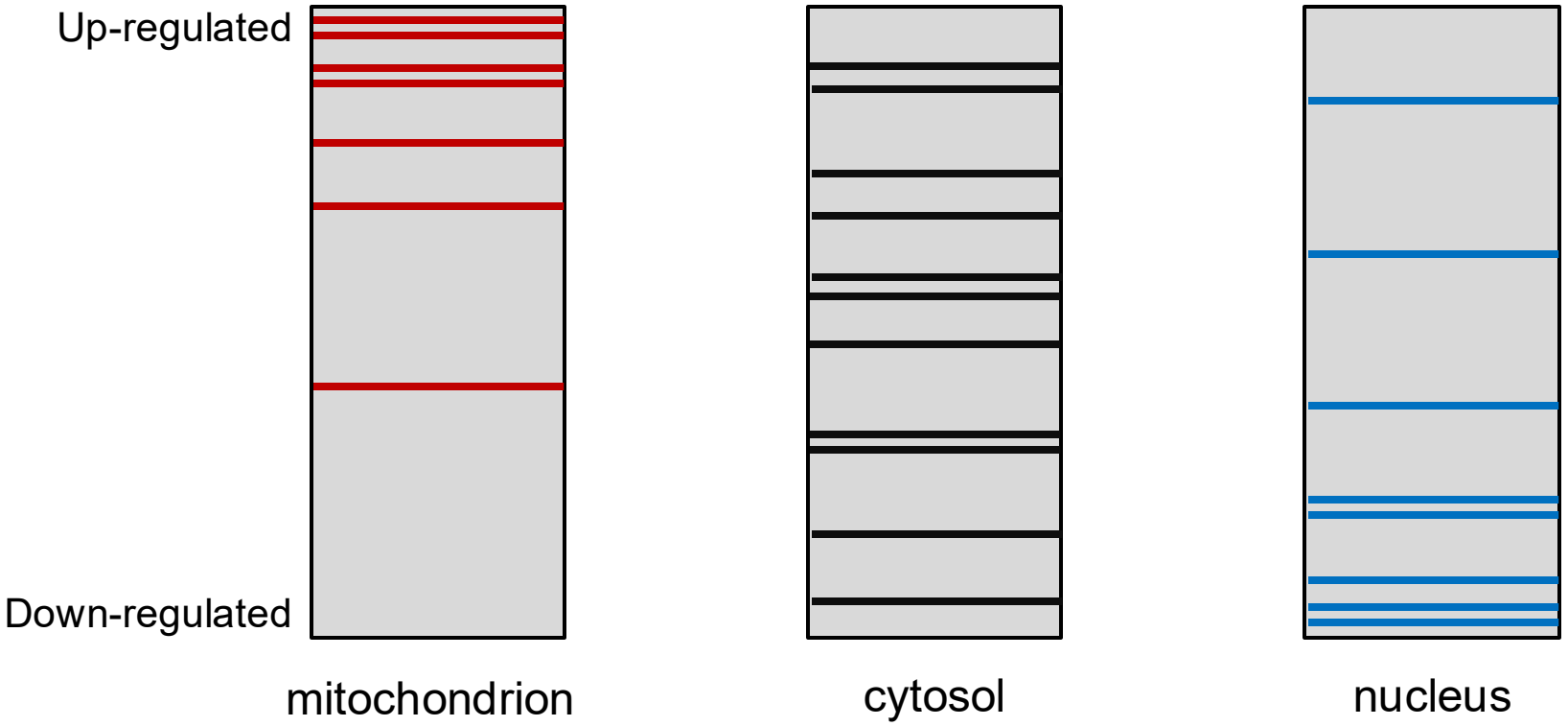
Category: Biological Process (Gene Ontology)
Group terms by: similarity >= 0.8
Sort terms by: Signal
Nr. of terms shown: 10
Color Palette: mint - blue
Highlight selected terms:



Download vector image (SVG) Download bitmap image (PNG)

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	-7.4672216299570175
10: CHP2	-7.423878301199893
11: GJB6	-7.301189455146853
12: COL17A1	-7.2636604397081825
13: C19orf33	-7.1952071849953185
14: SBSN	-7.140458097049176
15: LY6D	-7.056120251292827
16: TRIM29	-7.034785864374081
17: FLG	-7.031575998772657
18: CRCT1	-7.0226906177501025
19: KRT15	-6.8670255485207O2
20: SPRR1A	-6.859561525754514
21: LOR	-6.848595263892816
22: CLCA2	-6.767725587791244
23: SLURP1	-6.7673021587277775
24: C1orf68	-6.6955745952812125
25: LGALS7	-6.6132404743408575
26: CST6	-6.585766047436771
27: LYPD3	-6.5731282095054295
28: DMKN	-6.486748209061480S
29: LCE1B	-6.460586585775241
30: WFDC5	-6.441728770048803
31: SPRR2G	-6.4192093272457145
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

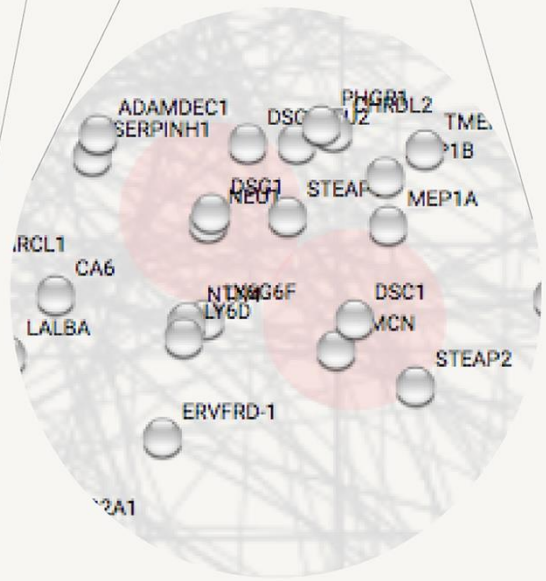
(more ...)

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)



Your input data

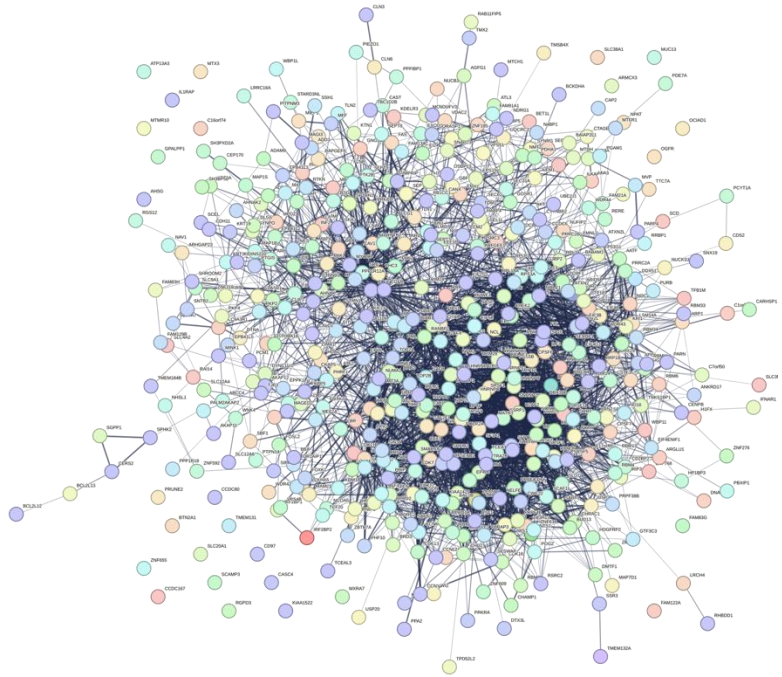
1: PKP1 -8.326649949
2: CDSN -8.1301573041
3: SERPINB5 -8.0657603659
4: DSC1 -7.9170774647
5: DSG1 -7.8383281946
6: CALML5 -7.706114452
7: ZNF750 -7.52776715?
8: SERPINB7 -7.4978374
9: LCE2B -7.46722?
10: CHP2 -7.42?
11: GJB6 -?

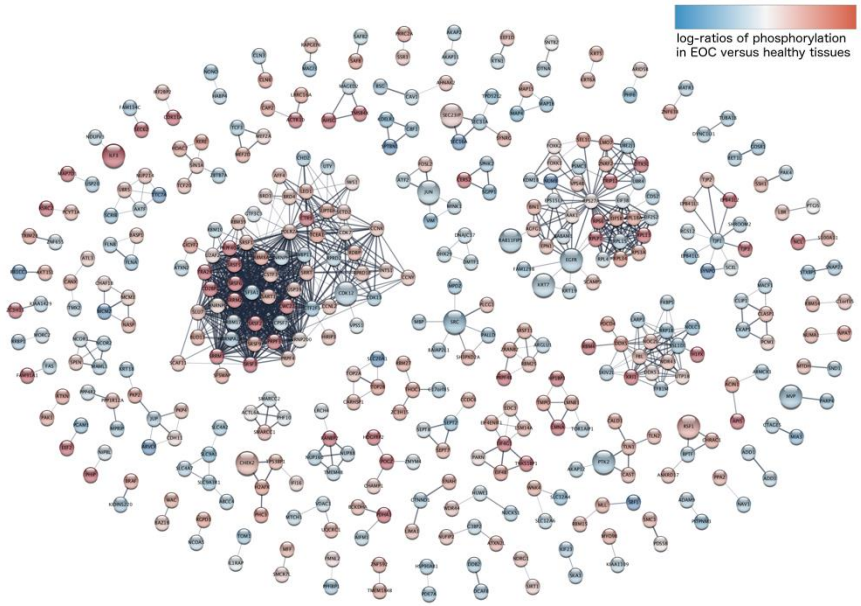
...rrier formation, ...
...nthyosis pathogenesis.
...s underlying a keratoderma-lik
... desmosome expression in h

...t in gene set | false discovery ...
50 | 9.48e-11
67 | 1.75e-07
23 | 1.16e-06
53 | 1.16e-06
6 | 0.00013
(more ...)

Questions?

From STRING to Cytoscape

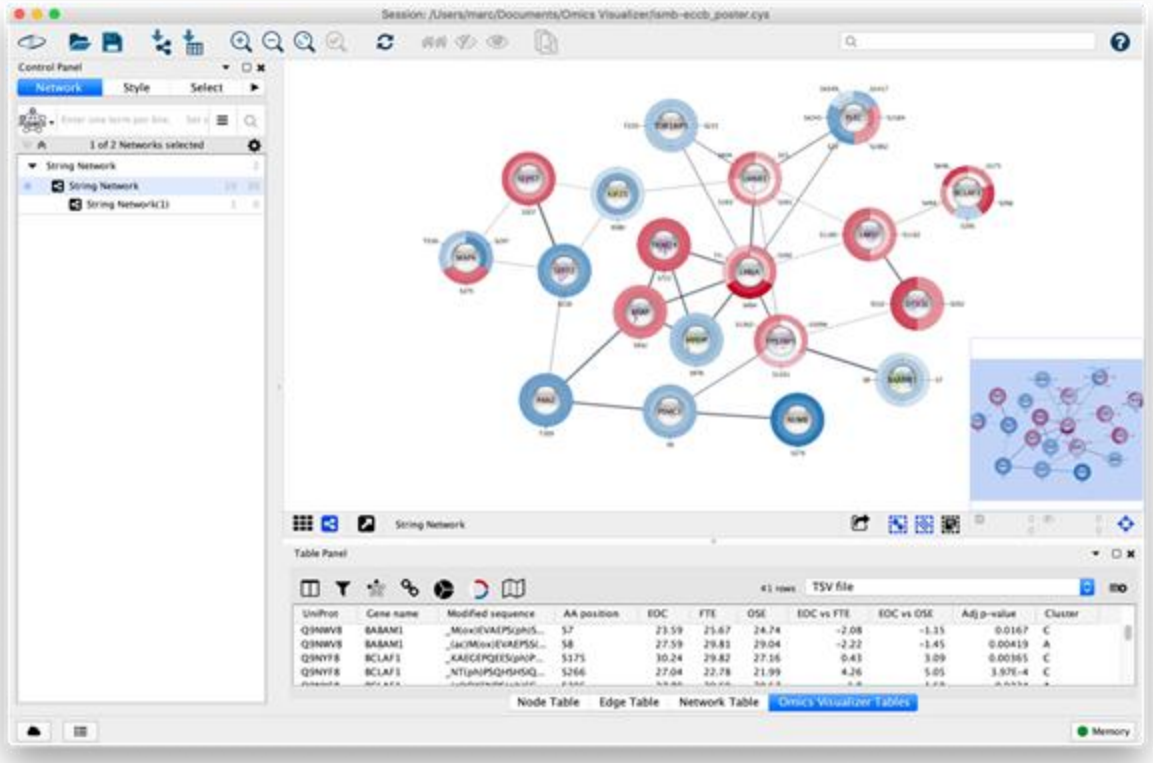




- What if we want to …
  - Create networks for large lists of genes
  - Integrate and easily show additional experimental data
  - Have more powerful analysis and visualization options

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
core app
layout
annotation
pathway database
ontology analysis
import

more »

Newest Releases

Get Started with the App Store »

AutoAnnotate
3.0+
Finds clusters and visually annotates them with labels and

MCODE
3.0+
Clusters a given network based on topology to find densely

ReactomeFIPlugIn
3.0+
Explore Reactome pathways and search for diseases related

ClueGO
3.0+
Creates and visualizes a functionally grouped network of

CluePedia
3.0+
CluePedia: A ClueGO plugin for pathway insights using integrated

mapSourceAndTarget
3.0+

more newest releases »

Top Downloaded Apps

ClueGO
3.0+
Creates and visualizes a functionally grouped network of

stringApp
3.0+
Import and augment Cytoscape networks from STRING

BiNGO
3.0+
Calculates overrepresented GO terms in the network and display

CluePedia
3.0+
CluePedia: A ClueGO plugin for pathway insights using integrated

MCODE
3.0+
Clusters a given network based on topology to find densely

yFiles Layout Algorithms
3.0+
Highly sophisticated algorithms for arranging networks.

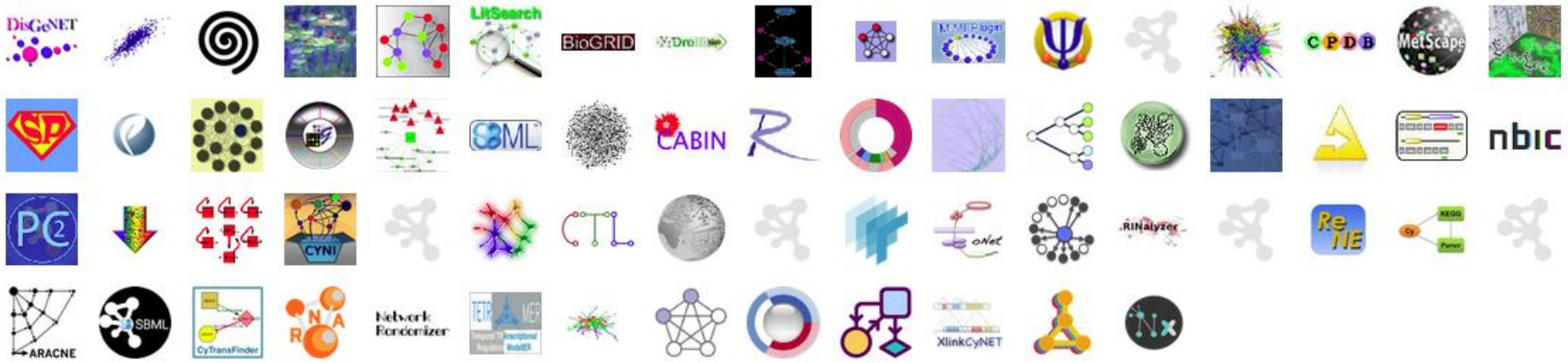
more top downloads »

Wall of Apps 379 total

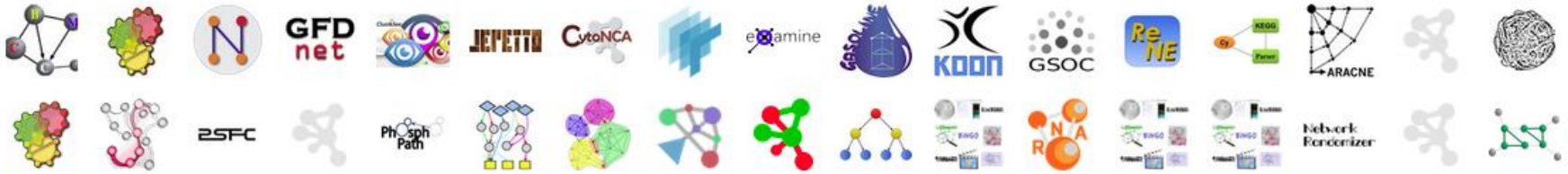
collections

data visualization

network generation



network analysis



stringApp

Import and augment Cytoscape networks from STRING

★★★★★ (34) 300746 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, five 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
- STRING: cross-species query -- choose two species to obtain a STRING network between and within the proteins of the interacting species

CYTOSCAPE 3

Download

Version 2.2.0

Released 30 Dec 2024

Works with Cytoscape 3.8

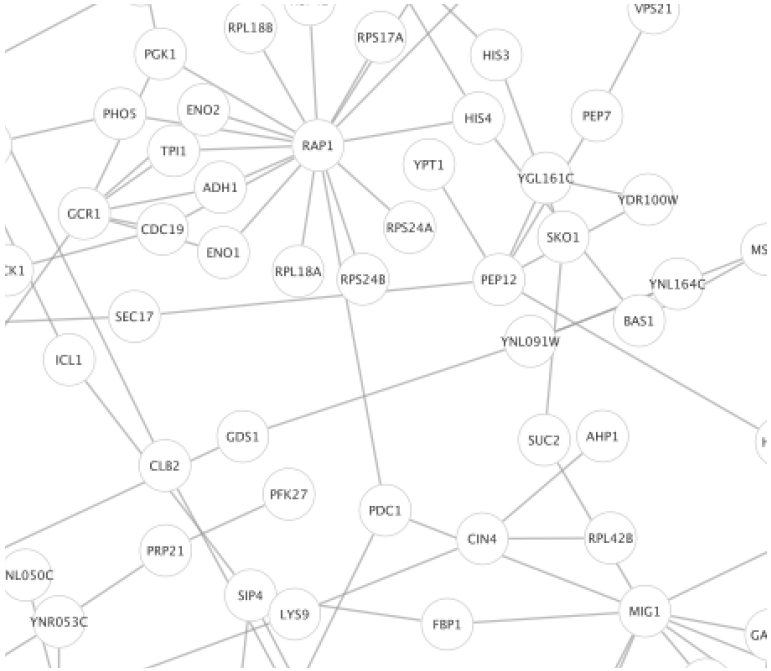
Download Stats Click here

cytoscape not running

RESOURCES

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

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

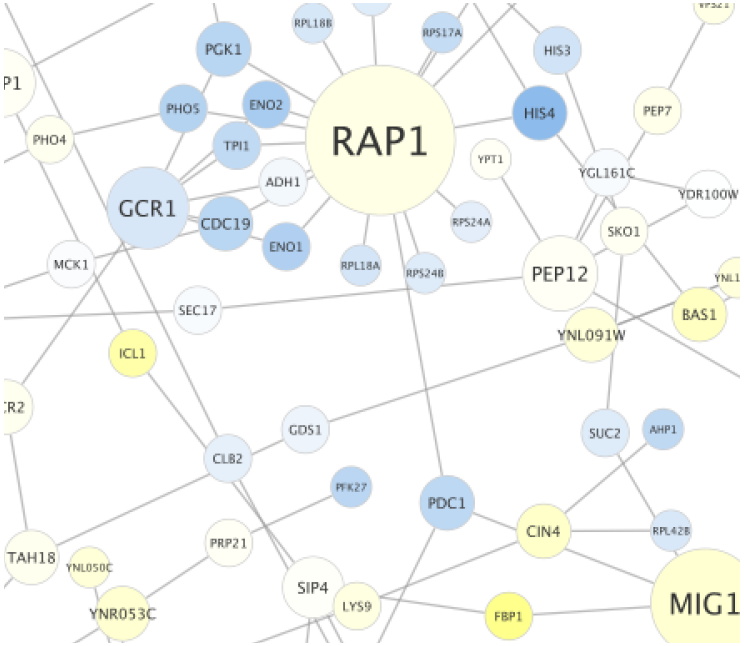
e.g., protein-protein interaction networks

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

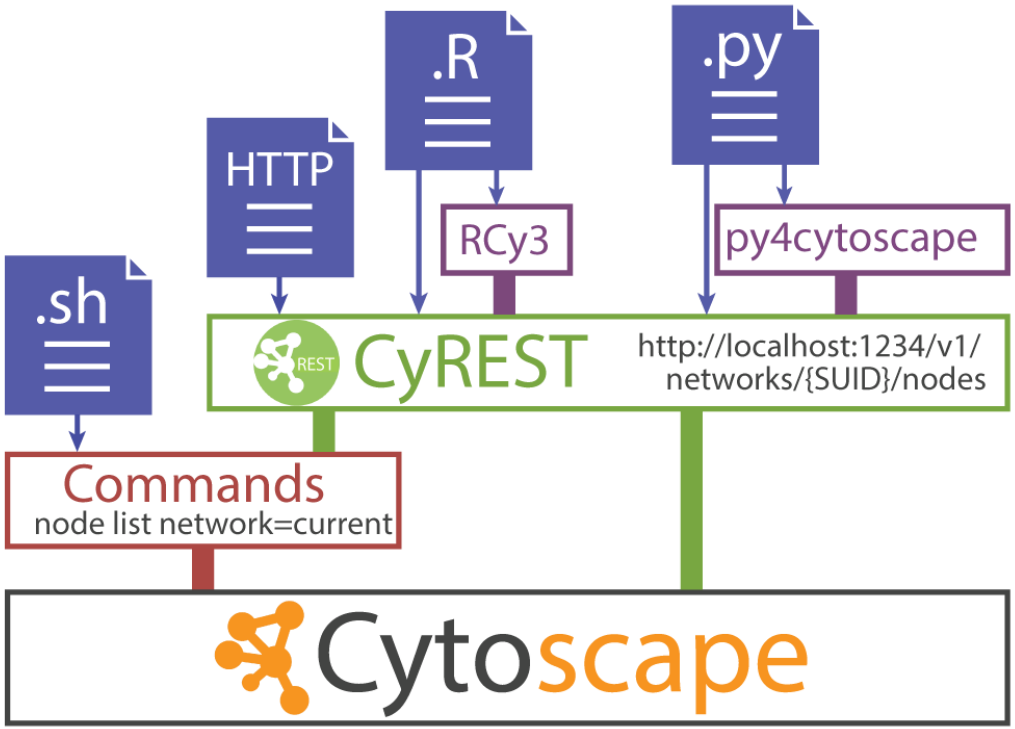
Tables

e.g., actual network data or annotations

Visual Styles

Cytoscape automation

- Use commands from R, Python, or JavaScript to execute Cytoscape, stringApp, and other apps' functionality



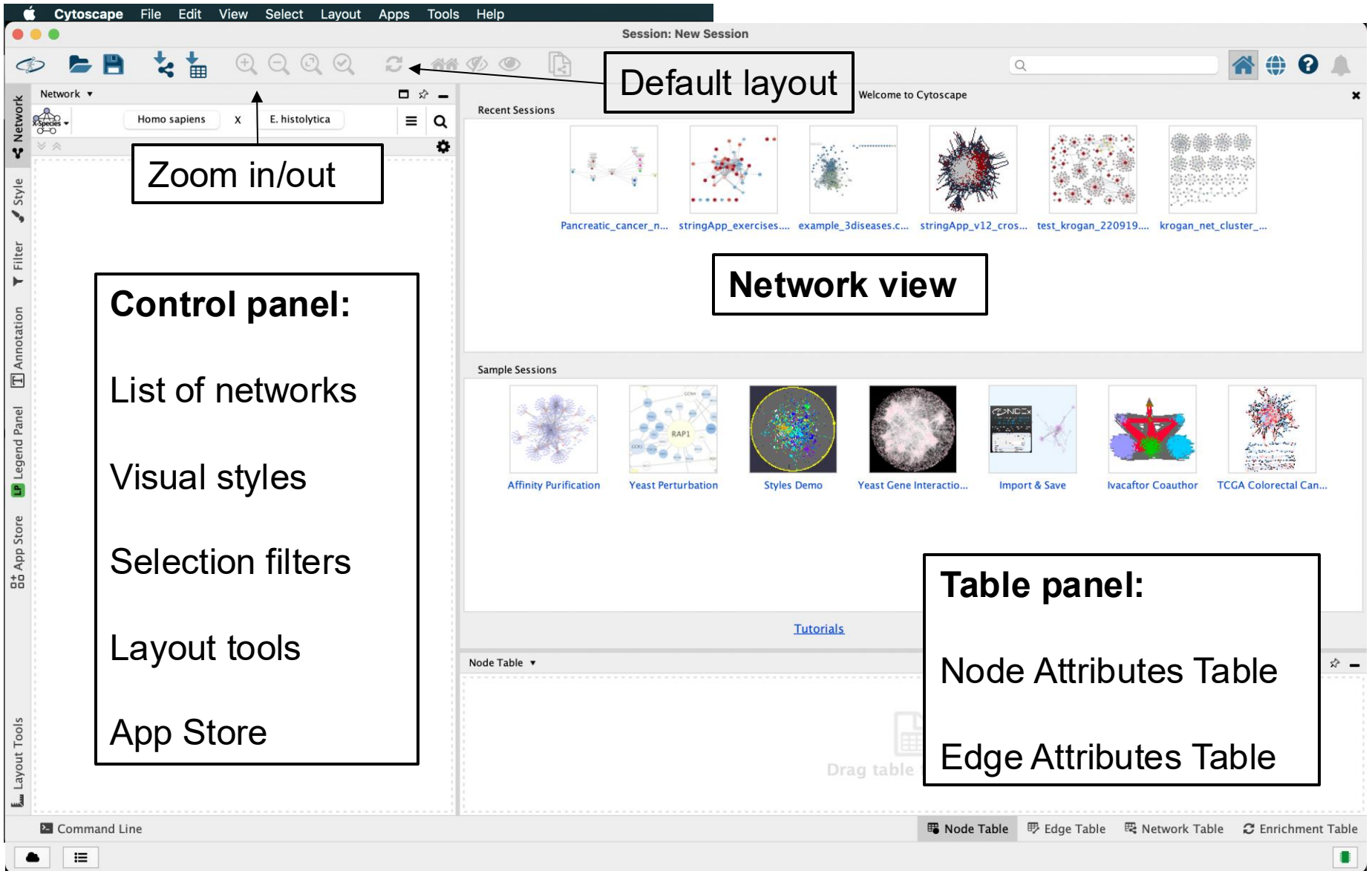
https://github.com/cytoscape/cytoscape-automation/wiki

Let's try it out!

How many have installed Cytoscape **3.10.3**?

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

Launch Cytoscape



Install stringApp v2.2

https://apps.cytoscape.org/

Troubleshooting: If your browser doesn't allow you to install the app directly from the App Store, you can still **download** it. Then, switch to **Cytoscape** and go to **Apps → App Store → Install apps from file.** Find the downloaded app in your files and press the **Open** button.

stringApp exercise 1

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

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

Pick two query types and try them out!

Exercise 1.1: Protein queries

Question 1: *How many nodes are in the resulting network? How does this compare to the maximum number of interactors you specified? What types of information do the* ***Node Table*** *and the* ***Edge Table*** *provide?*

Exercise 1.2: Compound queries

Question 2: *How is this network different from the protein-only network with respect to node types and the information provided in the Node Table?*

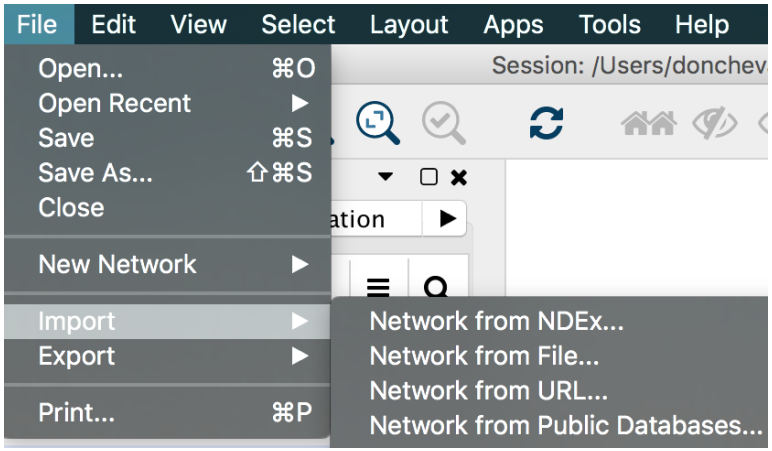
Exercise 1.3: Disease queries

Question 3: *Which additional attribute column do you get in the* ***Node Table*** *for a disease query compared to a protein query?*

Exercise 1.4: PubMed queries

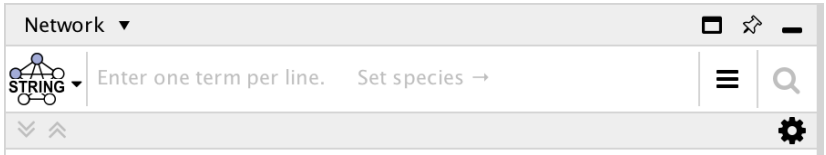
Question 4: *Which attribute column do you get in the* ***Node Table*** *for a PubMed query compared to a disease query?*

Import networks in Cytoscape

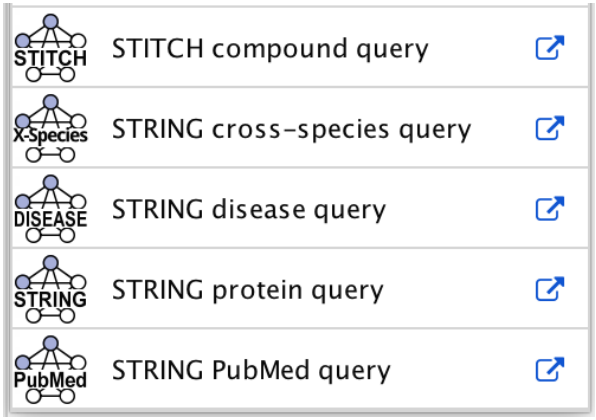


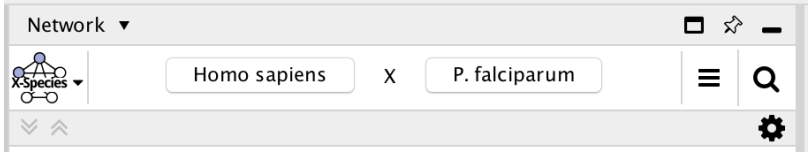
- **Starting with a list of genes and no network data**
  - stringApp
  - IntAct app
- **Starting with a pathway of interest**
  - KEGGscape app
  - ReactomeFI app
  - WikiPathways app
- **Starting with your own network data**
  - from files, e.g. Excel tables or text files
  - from R or Python via automation

stringApp

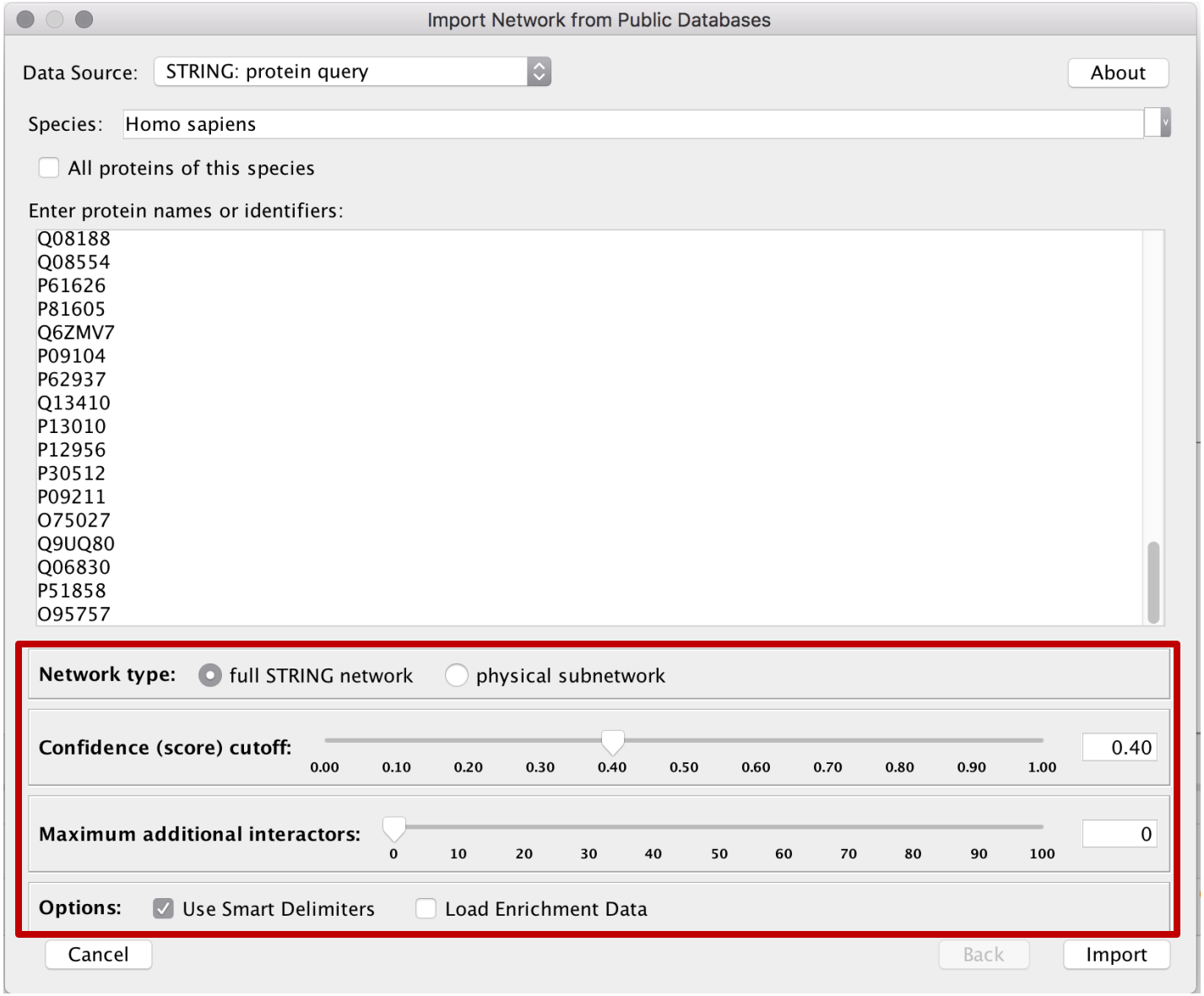


- STRING **protein** query
  - Queries for STRING interactions for **one** protein or for a **list** of identifiers
- STITCH **compound** query
  - Queries for protein-compound interactions
- STRING **disease** query
  - Queries for disease-associated proteins from DISEASES and for STRING interactions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

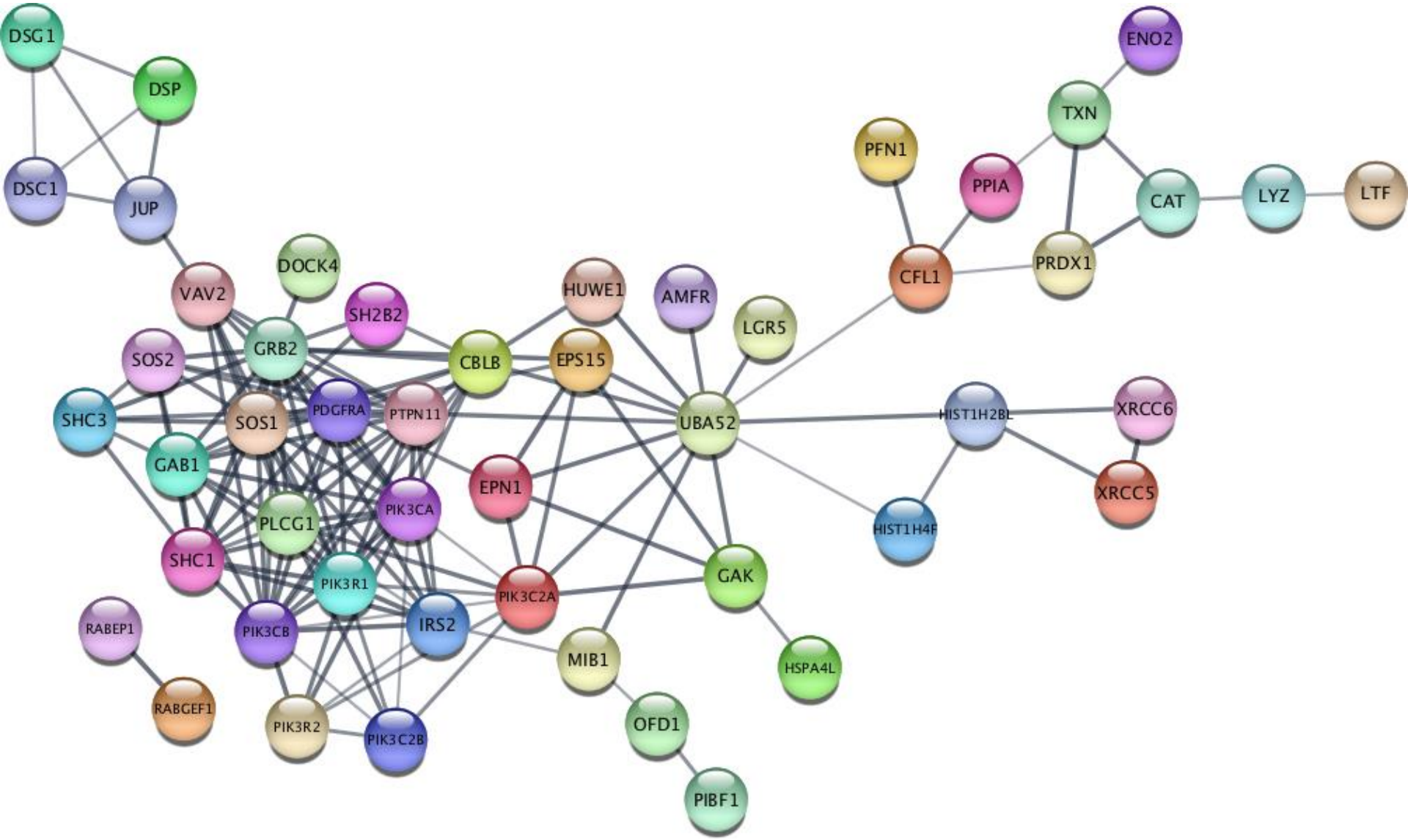




stringApp protein query

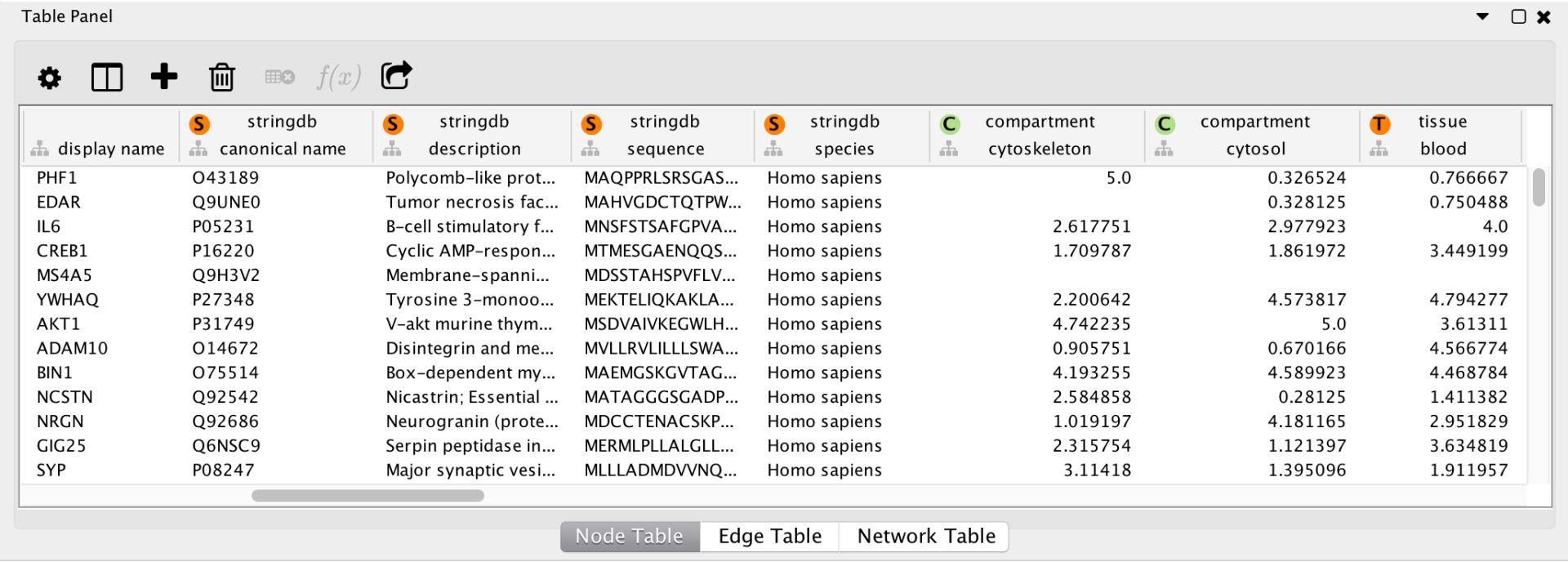


STRING network in Cytoscape



Node table (attributes)

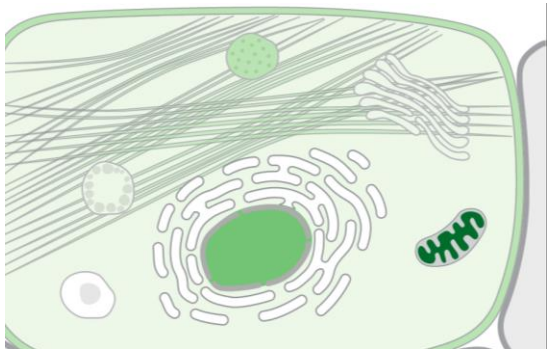
Nodes (and edges) can have data associated with them

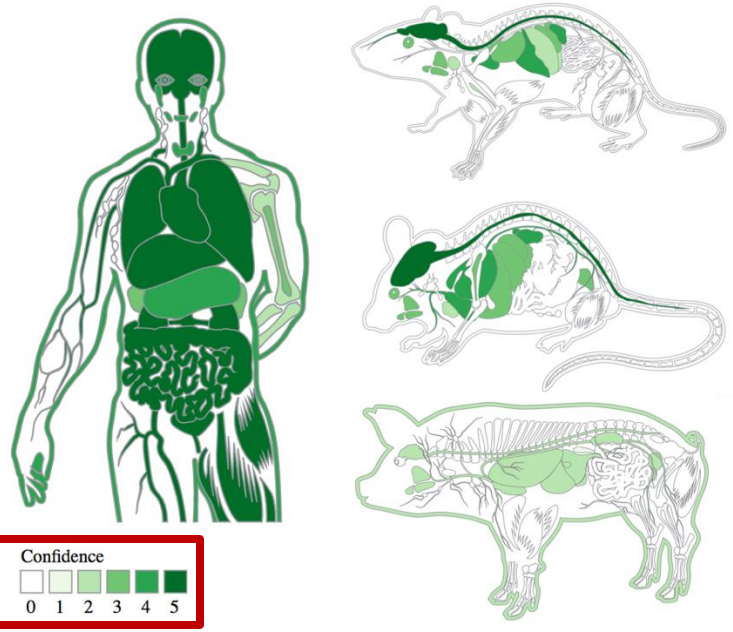


- Subcellular localization scores (https://compartments.jensenlab.org/)
- TISSUES expression scores (https://tissues.jensenlab.org/)
- Pharos drug target information (https://pharos.nih.gov/)

Related databases: Jensenlab

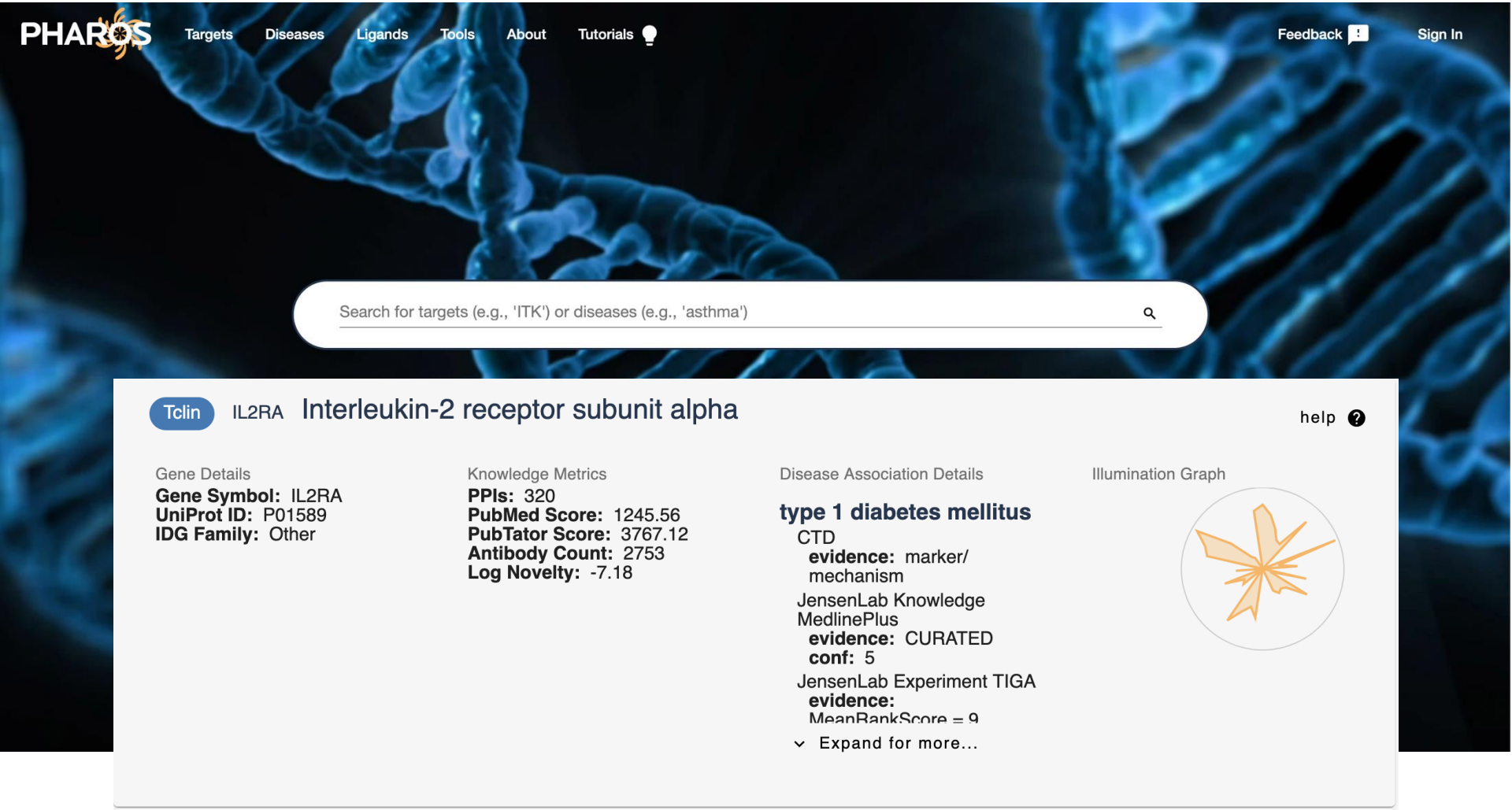
- **COMPARTMENTS**: Subcellular localization database
- **TISSUES**: tissue expression database for human, mouse, rat and pig
- **DISEASES**: disease-gene associations mined from the literature
- All three provide confidence scores between *0 and 5 stars*





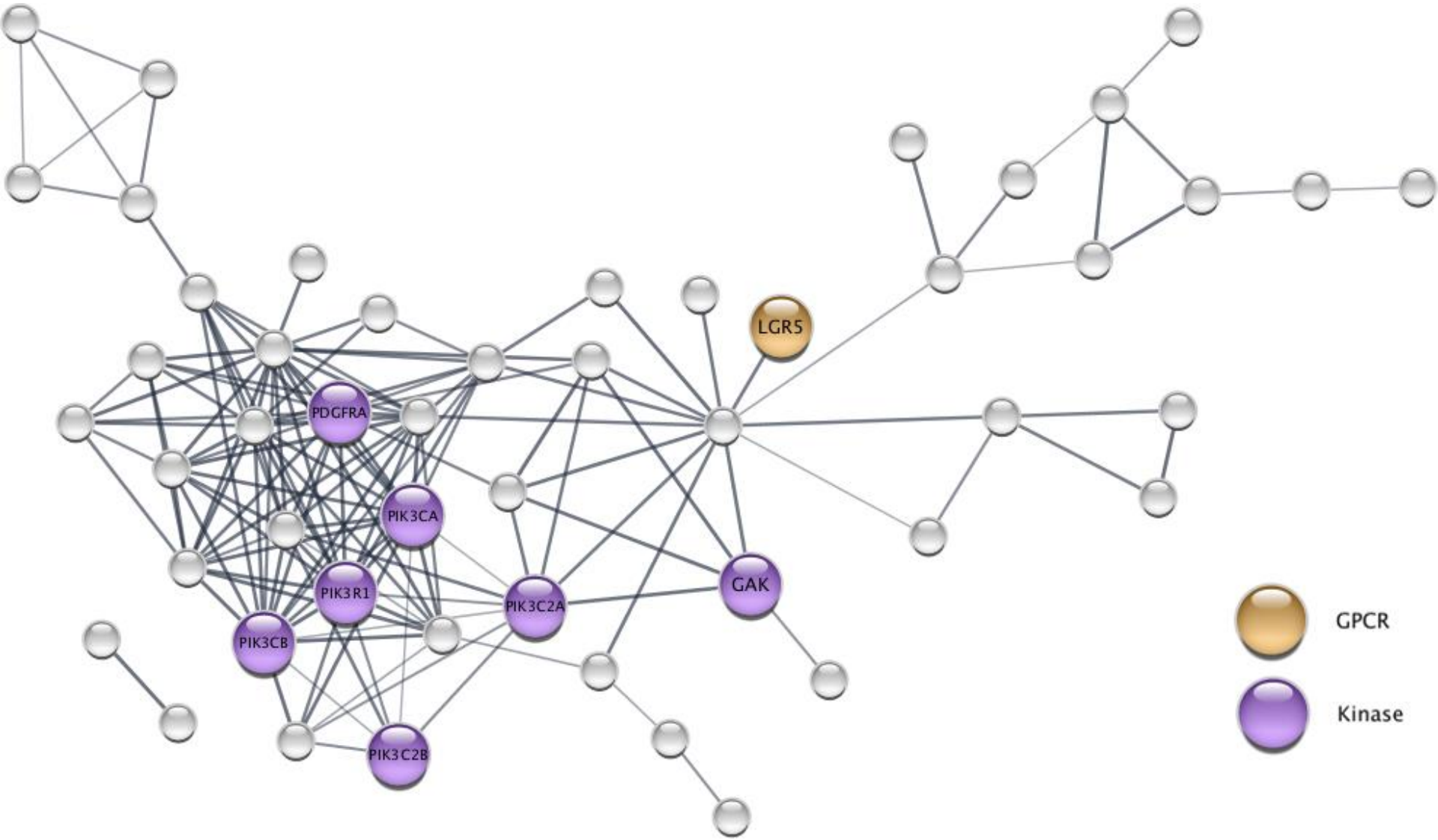
http://jensenlab.org/resources/

Related databases: Pharos



https://pharos.nih.gov/

Pharos drug target information

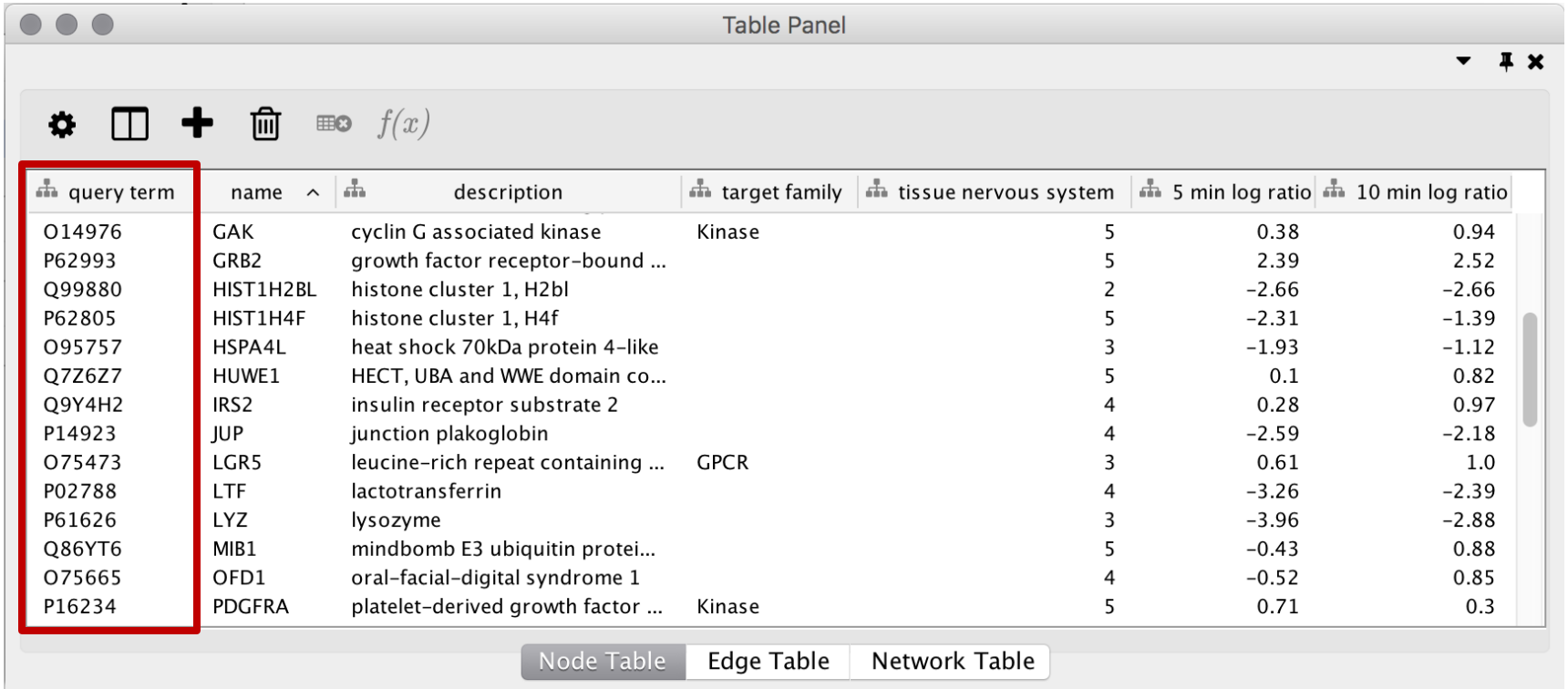


Visualize data using styles

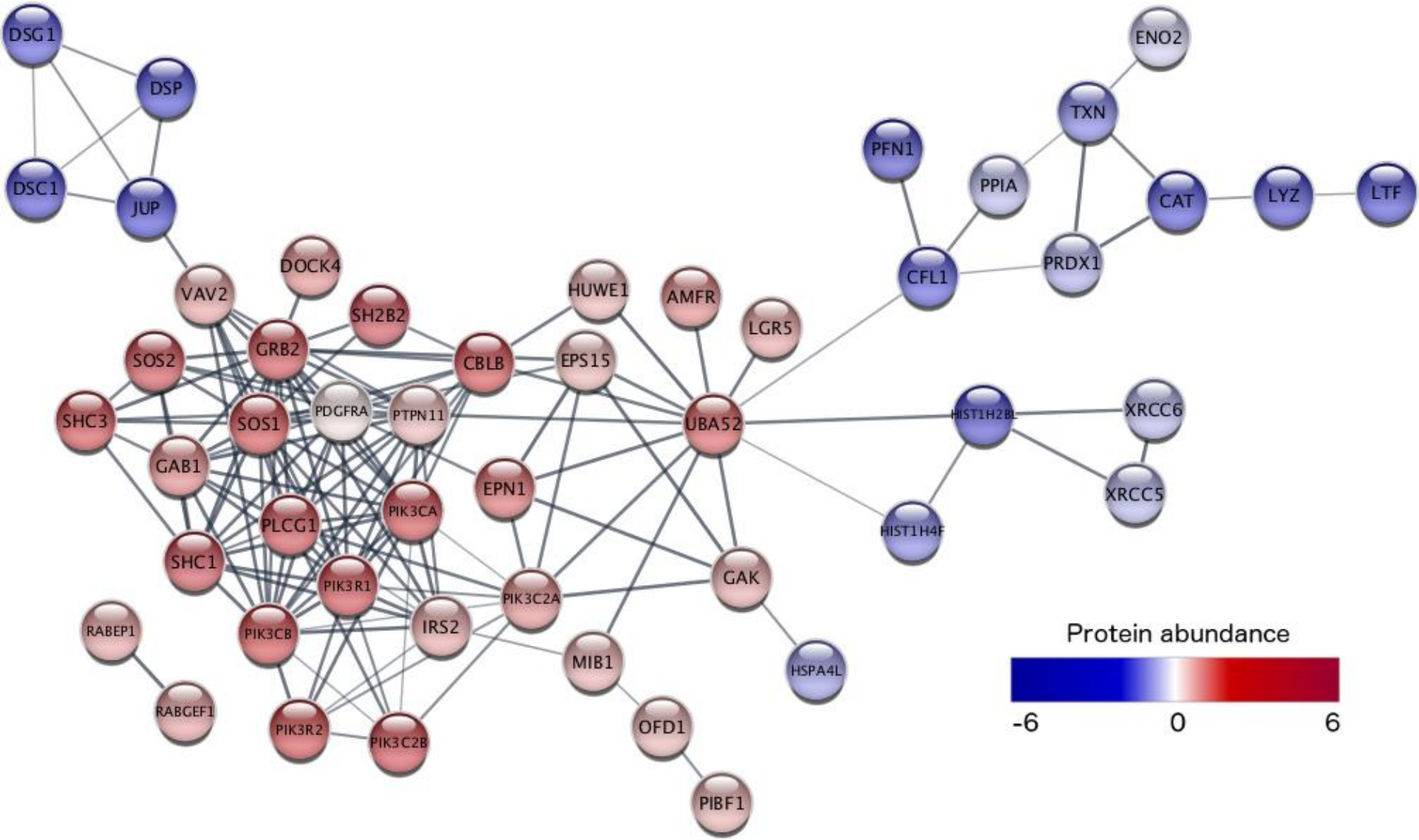
- Visual attributes
  - **Nodes**: fill color, border color, border width, size, shape, opacity, label, etc.
  - **Edges**: line style, line color, line width, line opacity, ending type, ending color, etc.
- Mapping types
  - **Continuous (numeric values)**
    - Expression values, edge interaction scores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

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

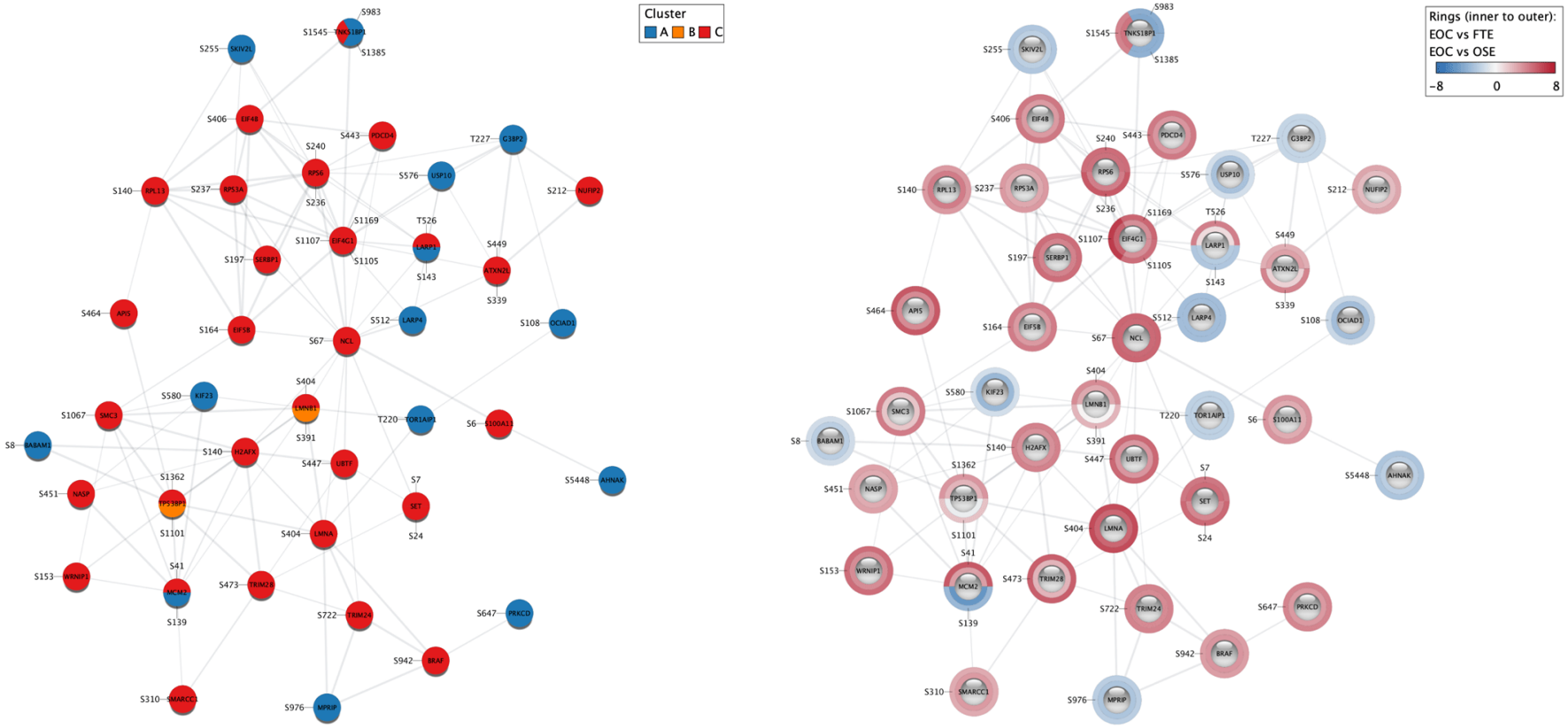


Expression data as node colors



Omics Visualizer app

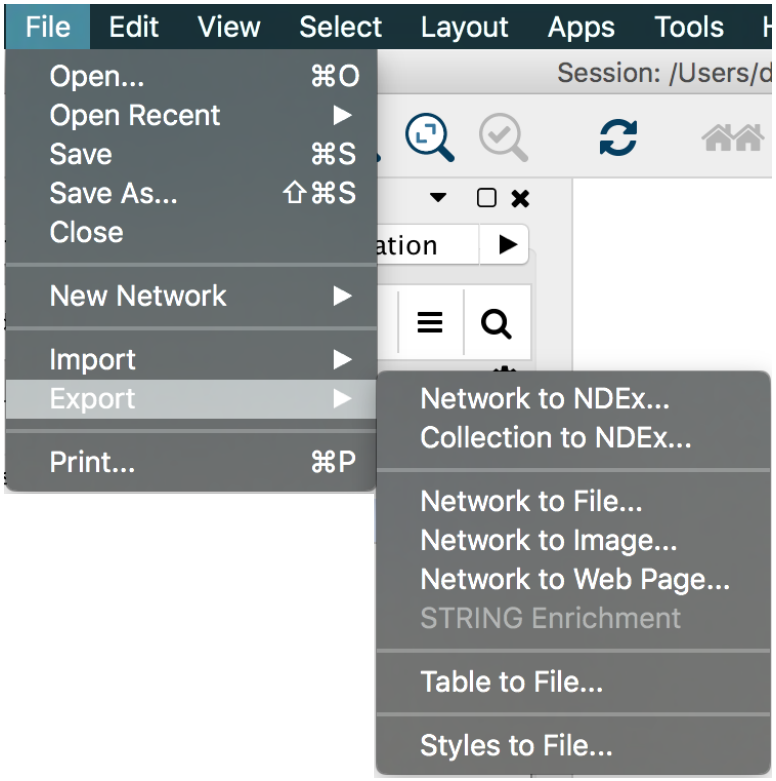
- Import your multi-omics data as an **Omics Visualizer table**
- Retrieve a STRING network for the proteins in the table
- Visualize as pies inside or donuts around the nodes



Legeay *et al.*, *F1000 Research*, 2020

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



Questions?

stringApp demo

stringApp exercise 2

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

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 import 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: *Do any of the suggested layouts help you to recognize patterns in the network? In which way?*

2.2 Discrete color mapping

Question 3: *How many of the proteins in the network are ion channels or GPCRs?*

Question 4: *How many kinases are in the network?*

2.3 Data import

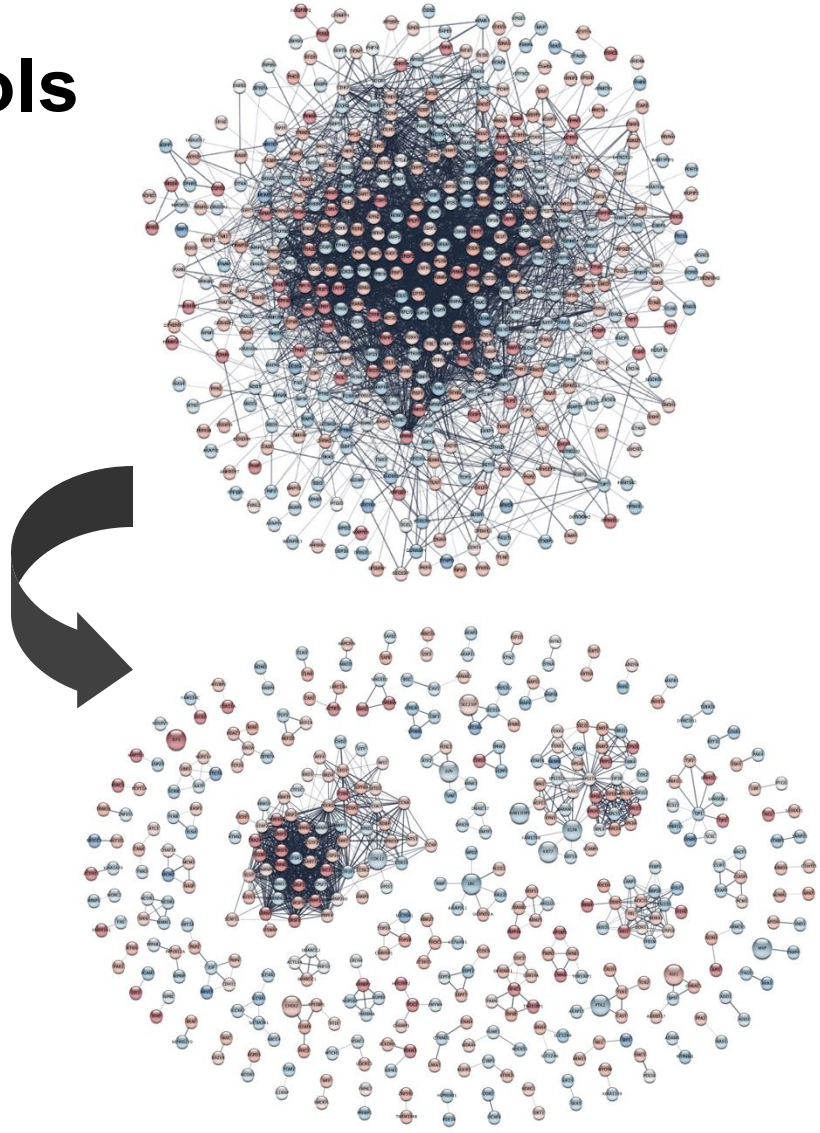
Question 5: *Do you see the columns from the Excel table in the Node Table?*

2.4 Continuous color mapping

Question 6: *Are the up-regulated nodes grouped together?*

Why use (biological) networks?

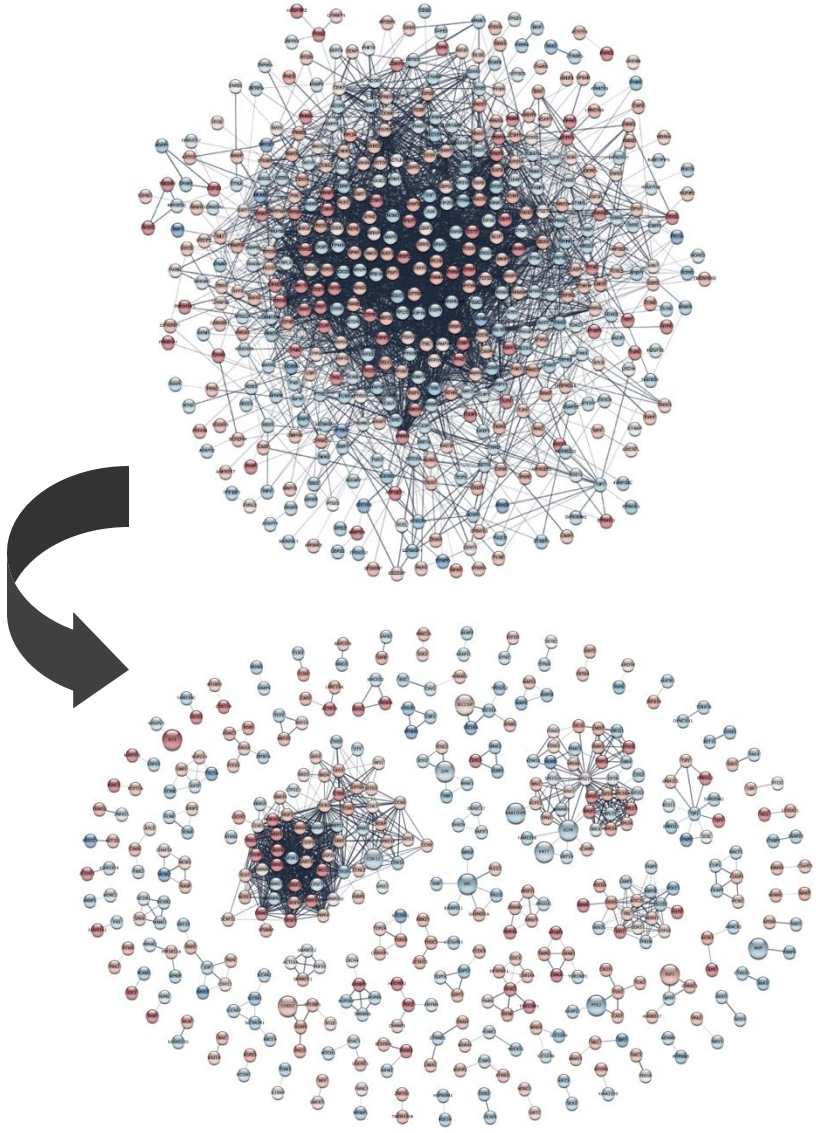
- Networks are **powerful tools**
  - ✓ Intuitive visualization
  - ✓ More efficient than tables
  - ✓ Reduce complexity
  - ✓ Great for data integration
- But also… Challenging!
- Many network analysis & visualization techniques available



Doncheva *et al.* (2019), *J Proteome Res*, 18(2): 623-632, Fig. 2 & 3.

Networks as tools

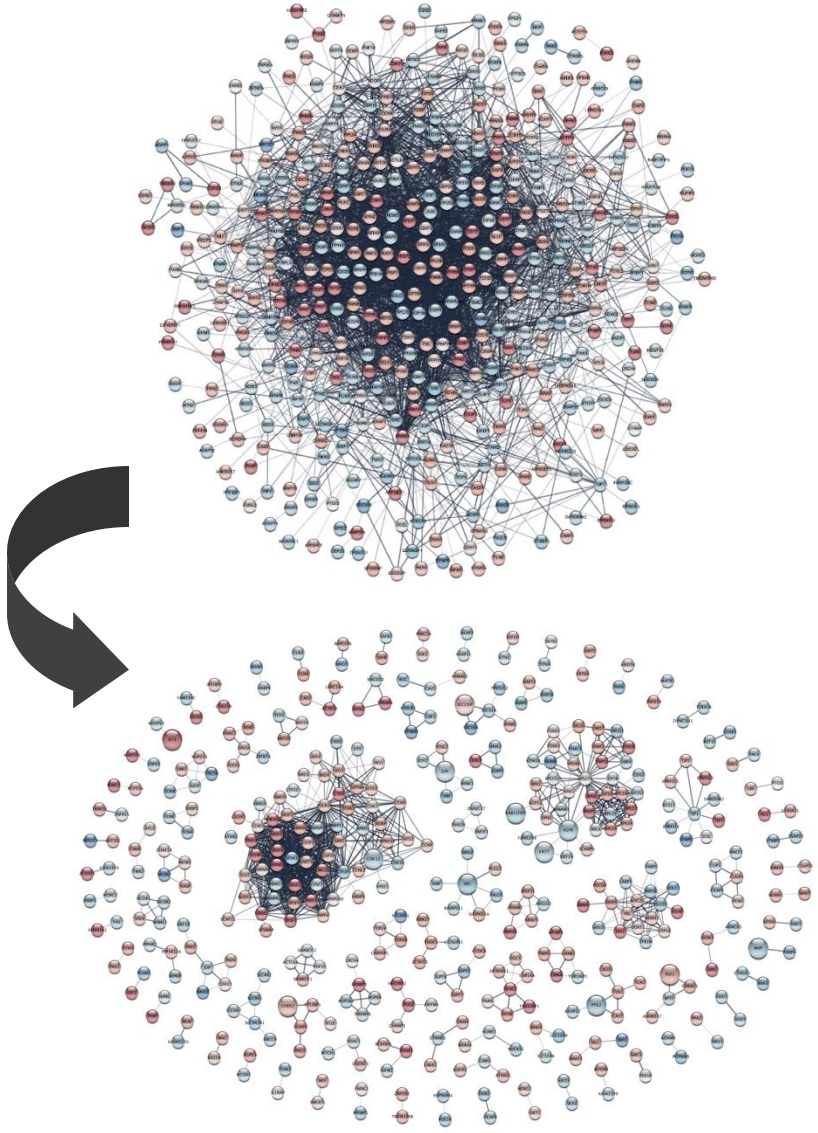
- **Visualization**
- **Analysis**



Doncheva *et al.* (2019), *J Proteome Res*, 18(2): 623-632, Fig. 2 & 3.

Networks as tools

- **Visualization** involves:
  - Data overlays
  - Layouts and animation
  - Exploratory analysis
  - Context and interpretation
- **Analysis** involves:
  - Topological properties
  - Hubs and robustness
  - Modularity/clusters
  - Data integration

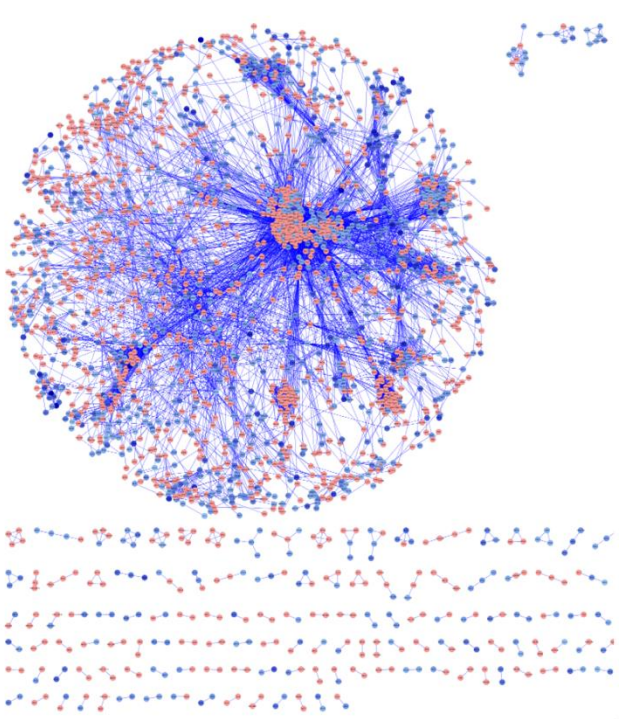


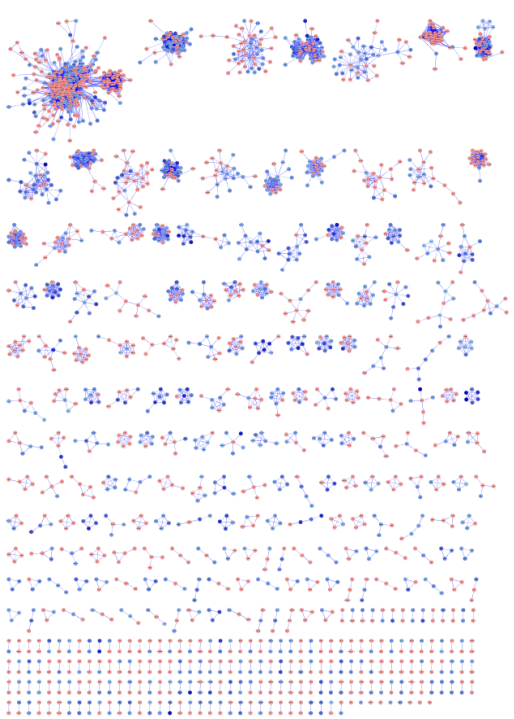
Doncheva *et al.* (2019), *J Proteome Res*, 18(2): 623-632, Fig. 2 & 3.

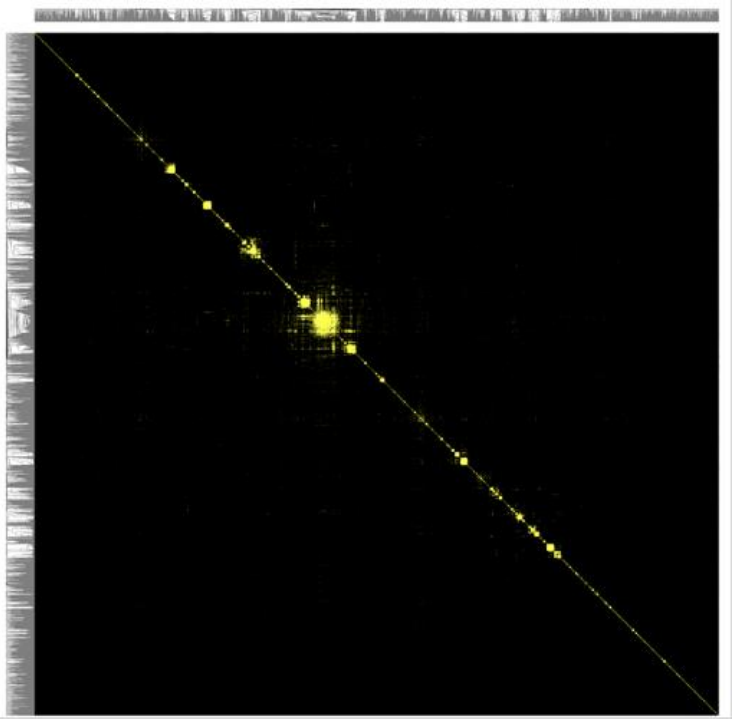
Depiction

Various ways to depict biological networks

- Node-Link (graph) representation
- Partitioned Node-Link representation
- Matrix representation

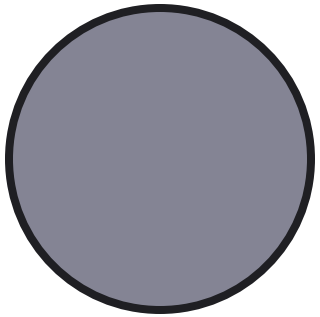


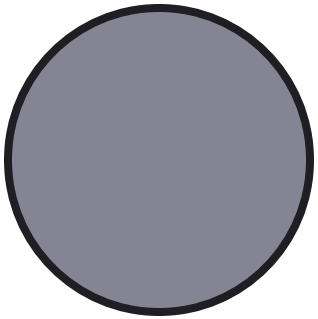


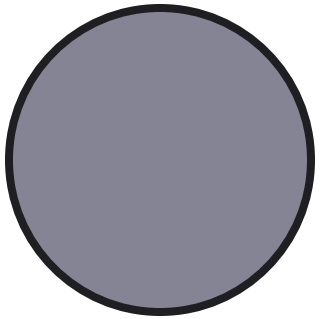


Visualize data on networks

- Visual variables: what can we vary to encode data?
  - Position

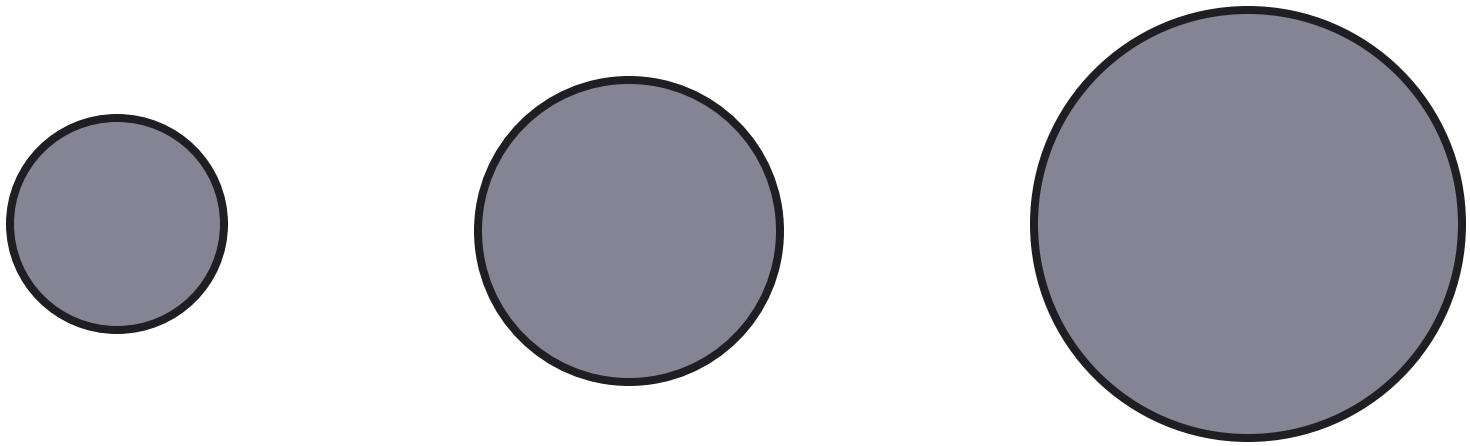






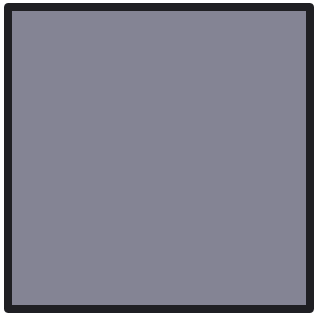
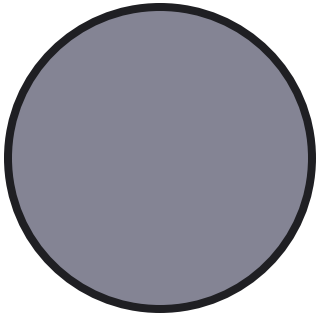
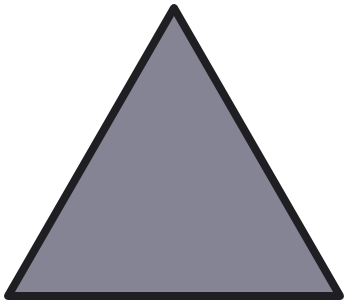
Visualize data on networks

- Visual variables: what can we vary to encode data?
  - Position
  - Size



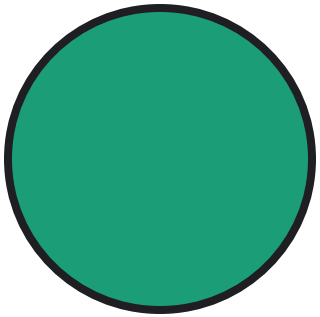
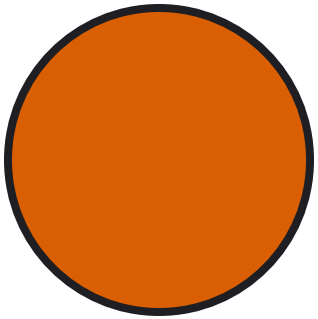
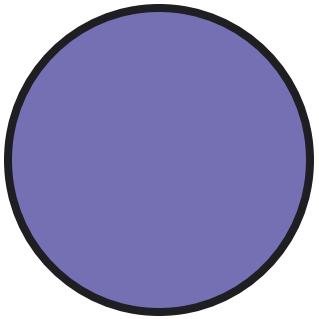
Visualize data on networks

- Visual variables: what can we vary to encode data?
  - Position
  - Size
  - Shape

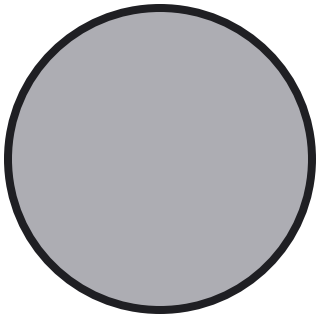
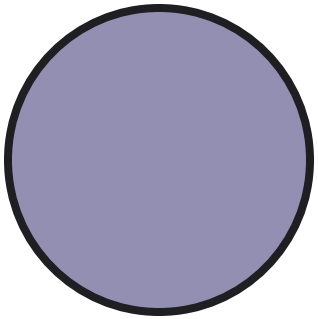
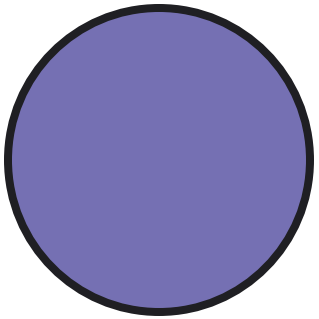
Visualize data on networks

- Visual variables: what can we vary to encode data?
  - Position
  - Size
  - Shape
  - Color
    - Hue

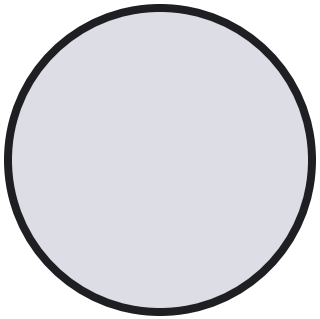
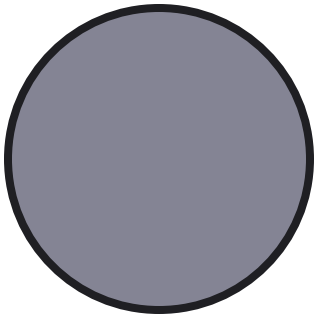
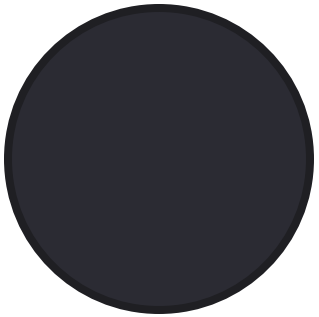
Visualize data on networks

- Visual variables: what can we vary to encode data?
  - Position
  - Size
  - Shape
  - Color
    - Hue
    - Saturation

Visualize data on networks

- Visual variables: what can we vary to encode data?
  - Position
  - Size
  - Shape
  - Color
    - Hue
    - Saturation
    - Brightness / lightness / value

Visualize data on networks

- Visual qualities: what are the visual variables good at?
  - Selective (easily spot groups with the same value)

  → all, except for shape

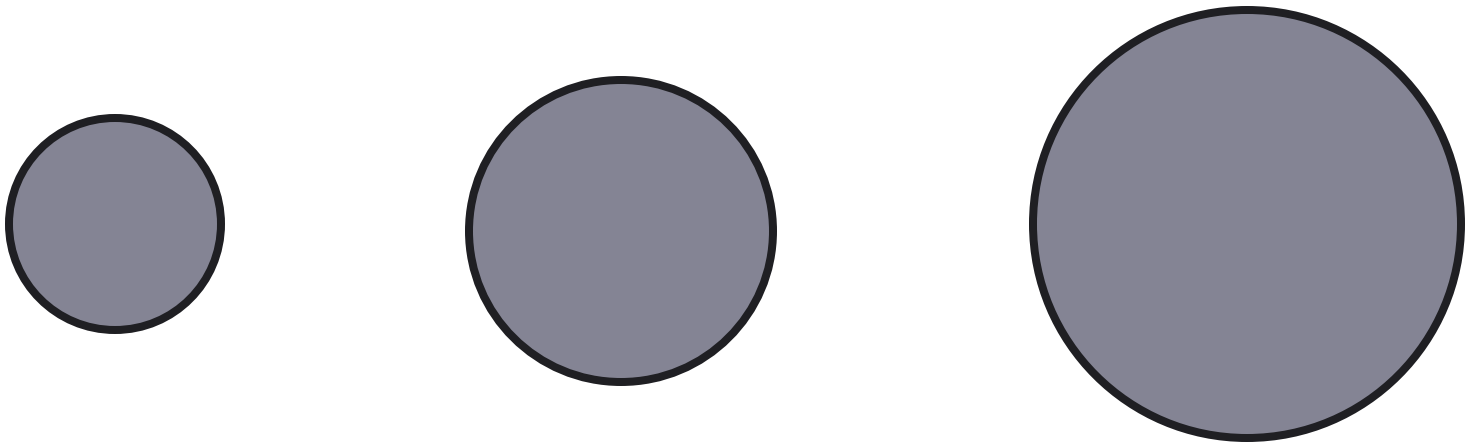
Visualize data on networks

- Visual qualities: what are the visual variables good at?
  - Selective (easily spot groups with the same value)

  → all, except for shape

  - Quantitative

  → position and size



Visualize data on networks

- Visual qualities: what are the visual variables good at?
  - Selective (easily spot groups with the same value)

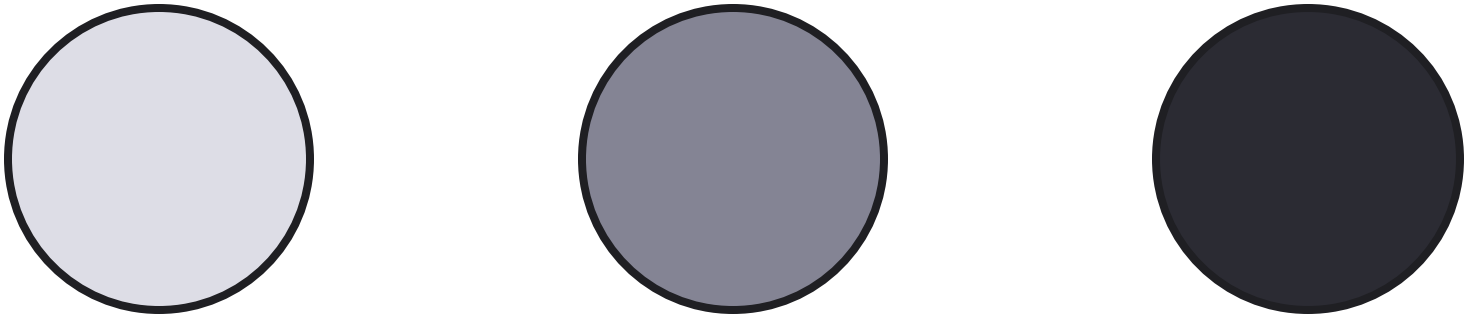
  → all, except for shape

  - Quantitative

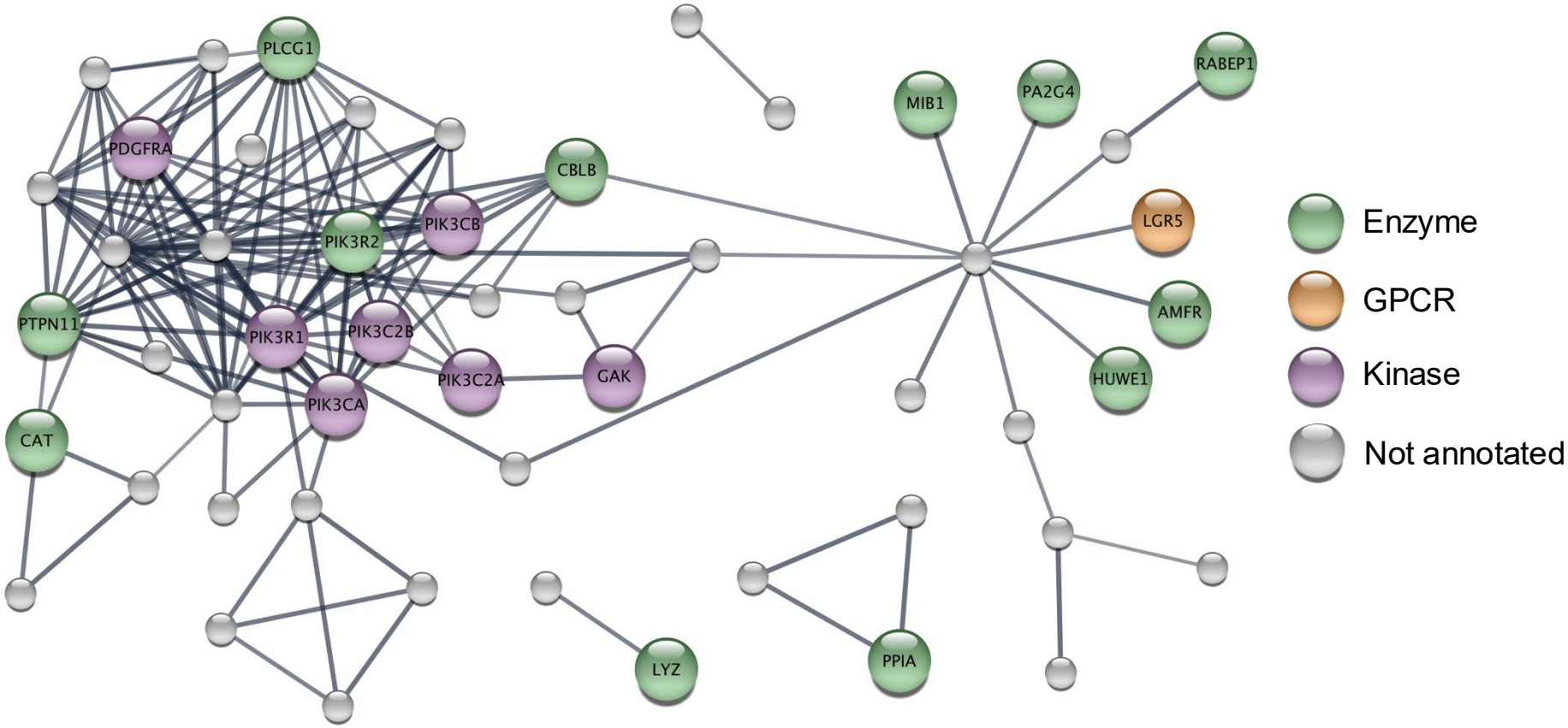
  → position and size

  - Ordered

  → saturation and brightness, but not hue or shape



Example: protein families



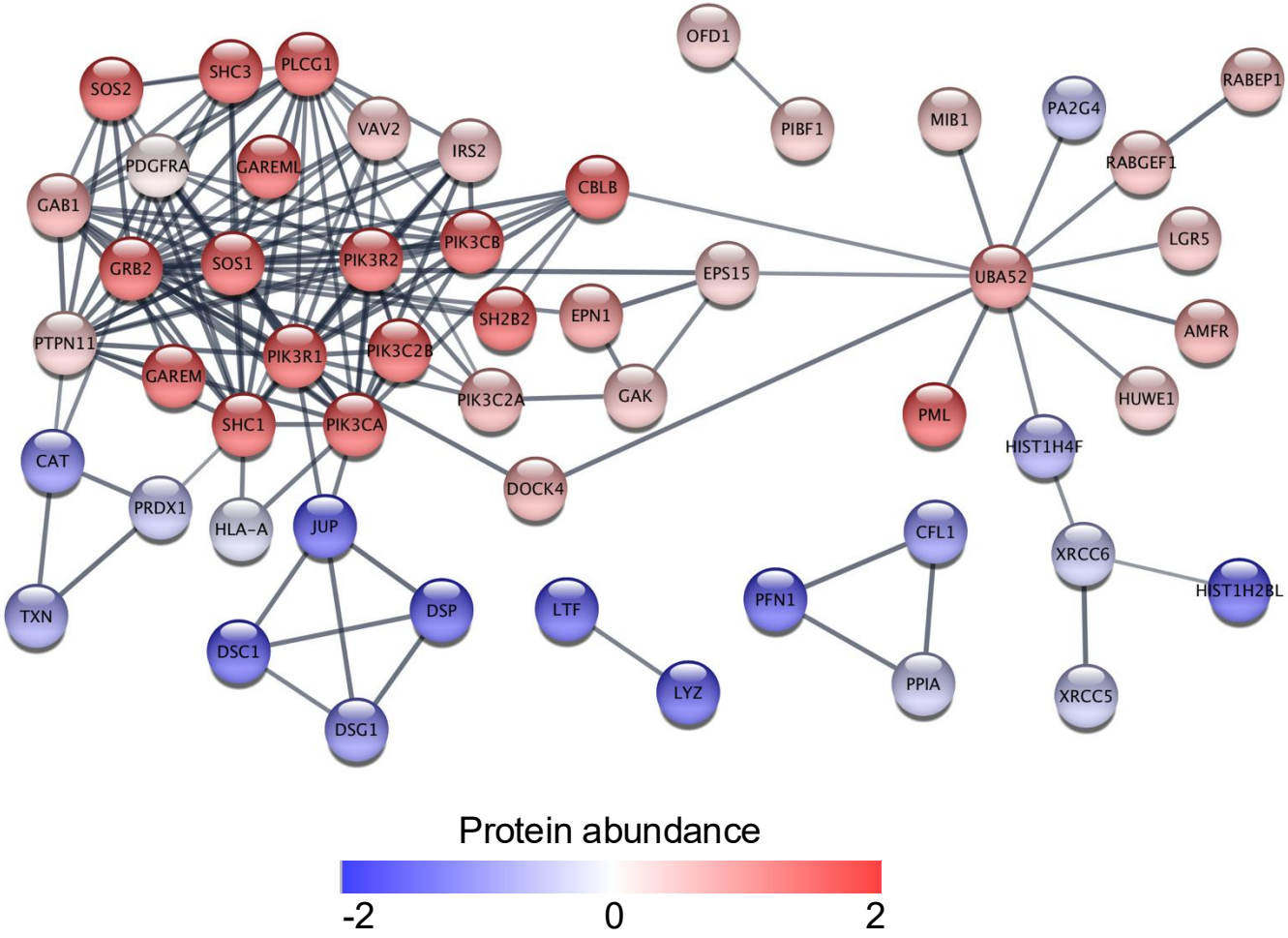
Network layout: node position

Annotated: node size, brightness, saturation

Protein family: node hue

Confidence of interactions: edge width & saturation

Example: proteomics data



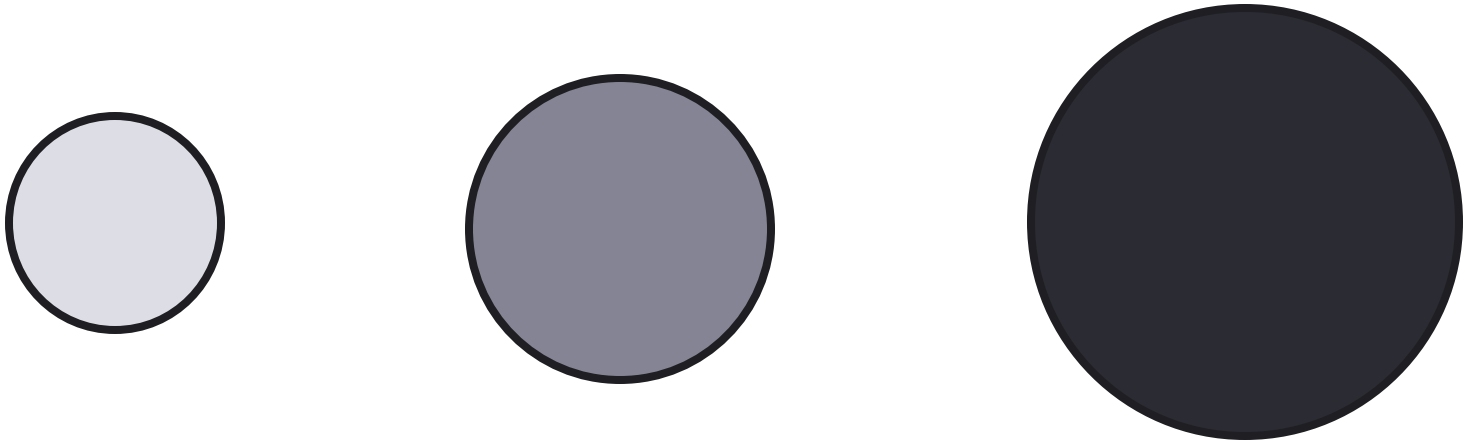
Direction:
node hue

Magnitude:
node brightness

Confidence:
edge width &
brightness

Tips & tricks

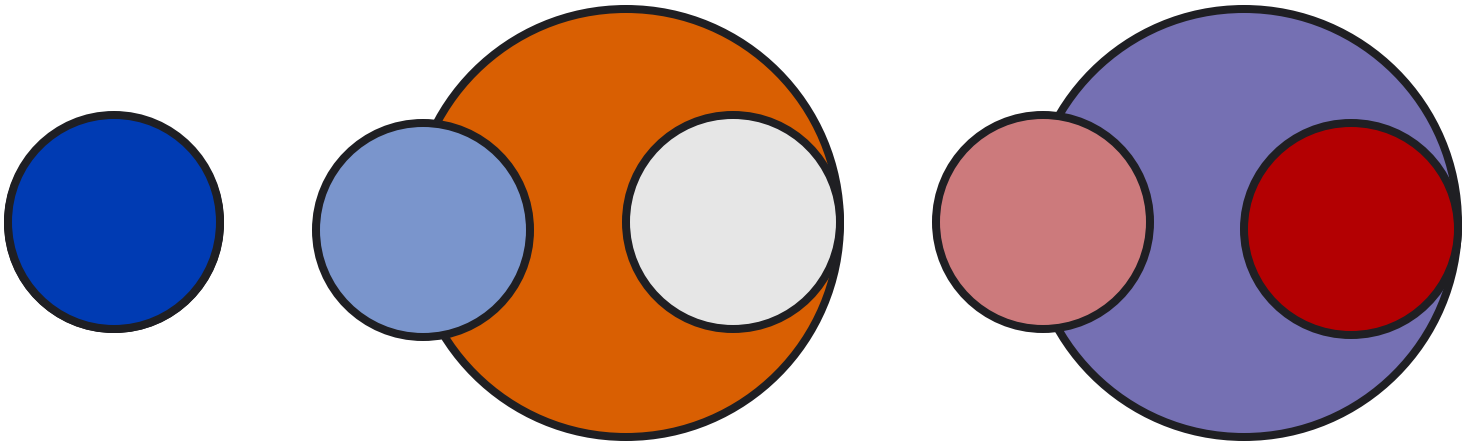
- Redundant encoding
  - E.g. size and brightness



Tips & tricks

- Redundant encoding
  - E.g. size and brightness

- Complementary encodings
  - E.g. size, brightness and hue (protein families example)
  - E.g. hue and brightness (proteomics data example)



Tips & tricks

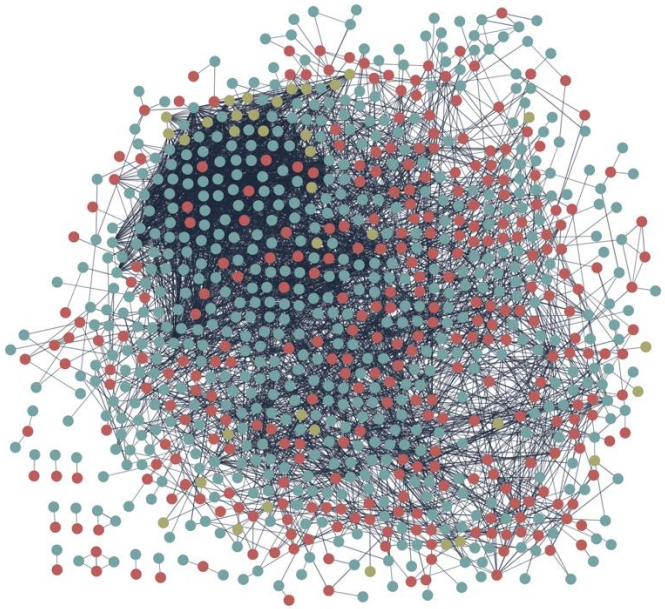
- Redundant encoding
  - E.g. size and brightness

- Complementary encodings
  - E.g. size, brightness and hue (protein families example)
  - E.g. hue and brightness (proteomics data example)

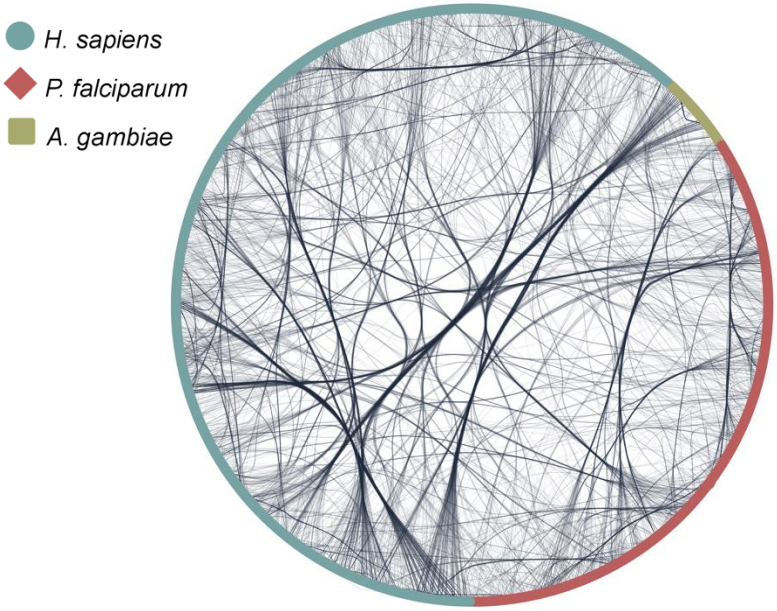
- Competing encodings
  - E.g. node hue and edge hue

Layouts

- Determine the location of nodes and (sometimes) the paths of edges on the 2D space
- Use them to emphasize the relationships between nodes
- There is not one *correct* layout → Try different things!



Force-directed



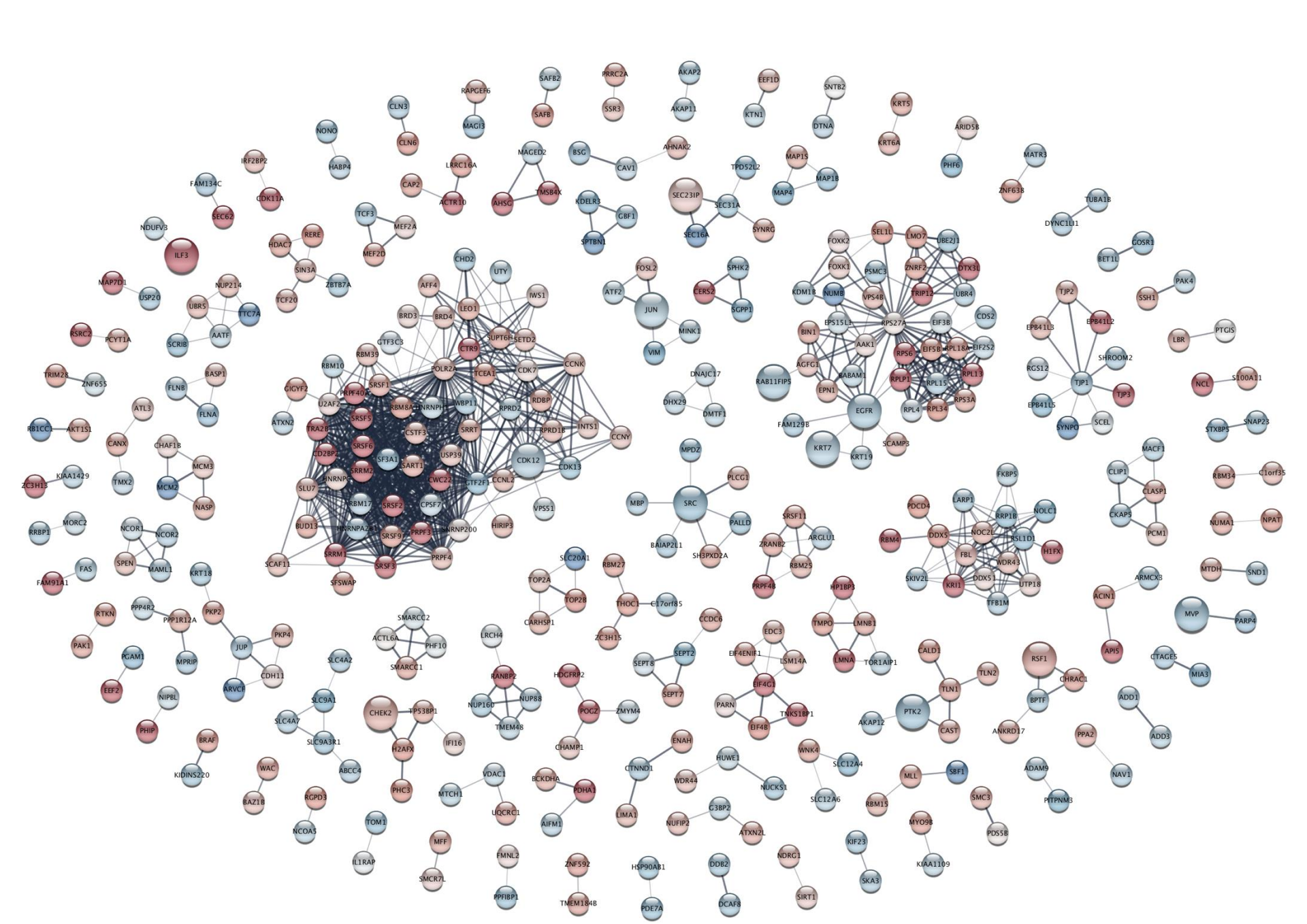
Circular

Cross-species network created with stringApp 2.0, Doncheva *et al., J Proteome Res,* 2023.

Network clustering

- Captures the structure of the network by identifying dense subgraphs, e.g.
  - Protein complexes in protein-protein interaction networks
  - Functional modules in functional association networks

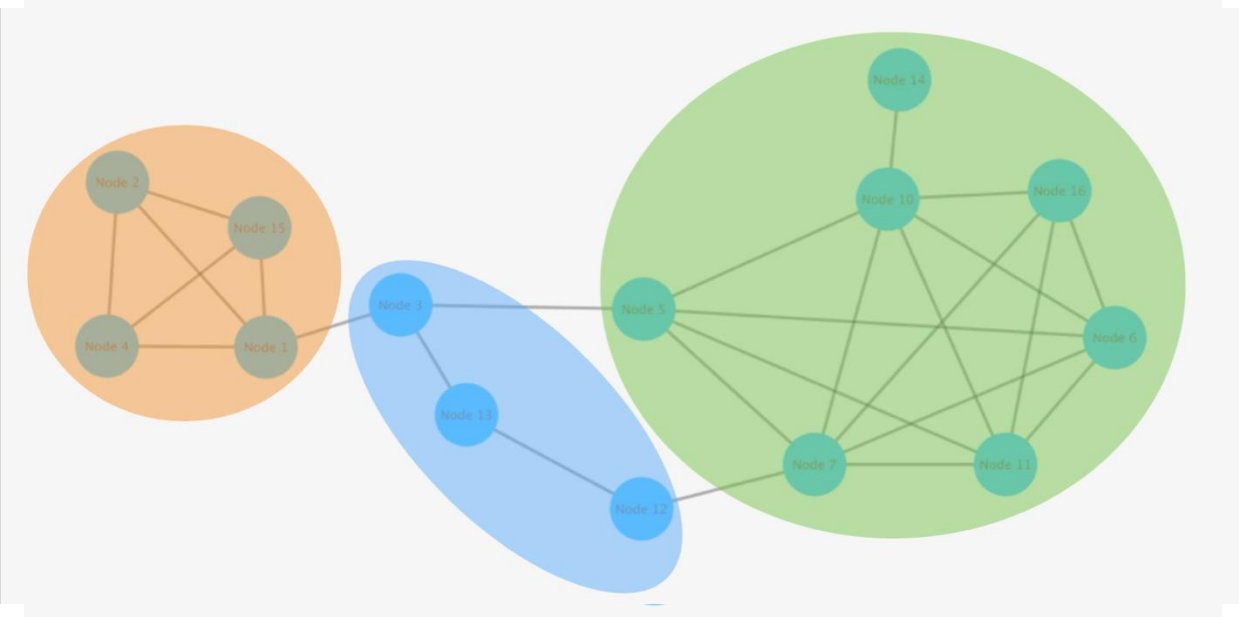
- Advantages for visualization → it can guide the layout of the network and be used for network simplification, e.g.
  - Represent each cluster as one node
  - Show only edges within clusters



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

Network clustering example

- Group nodes together based on a measure of similarity between the nodes, e.g. edges or edge weights
- MCL (Markov CLustering)
  - Fast algorithm
  - No need to specify number of clusters



Clustering in Cytoscape

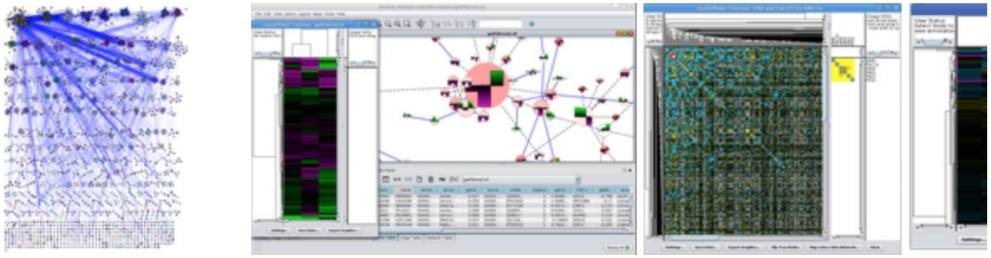
clusterMaker2

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

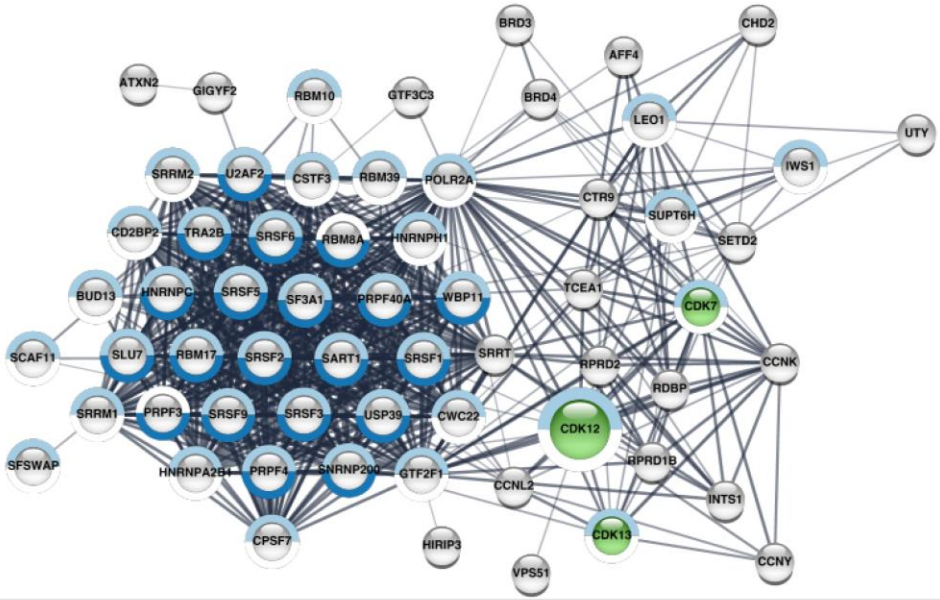
Download Stats Click here

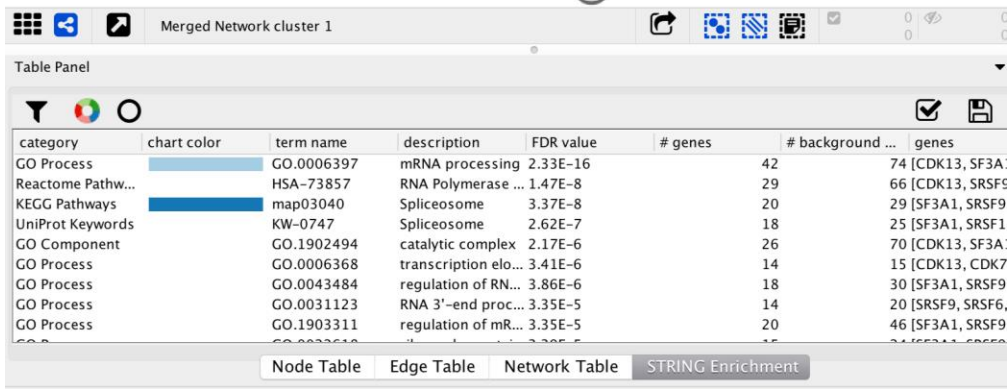
RESOURCES

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

stringApp functional enrichment

- Can be very **useful for visualization**
- Many categories: Gene Ontology, Pathways, Diseases & phenotypes, Tissues & subcellular localization, Protein domains, Publications
- **Visualize** significant terms
- **Filter** terms by category or remove redundant terms
- **Group-wise enrichment** on all clusters in the network

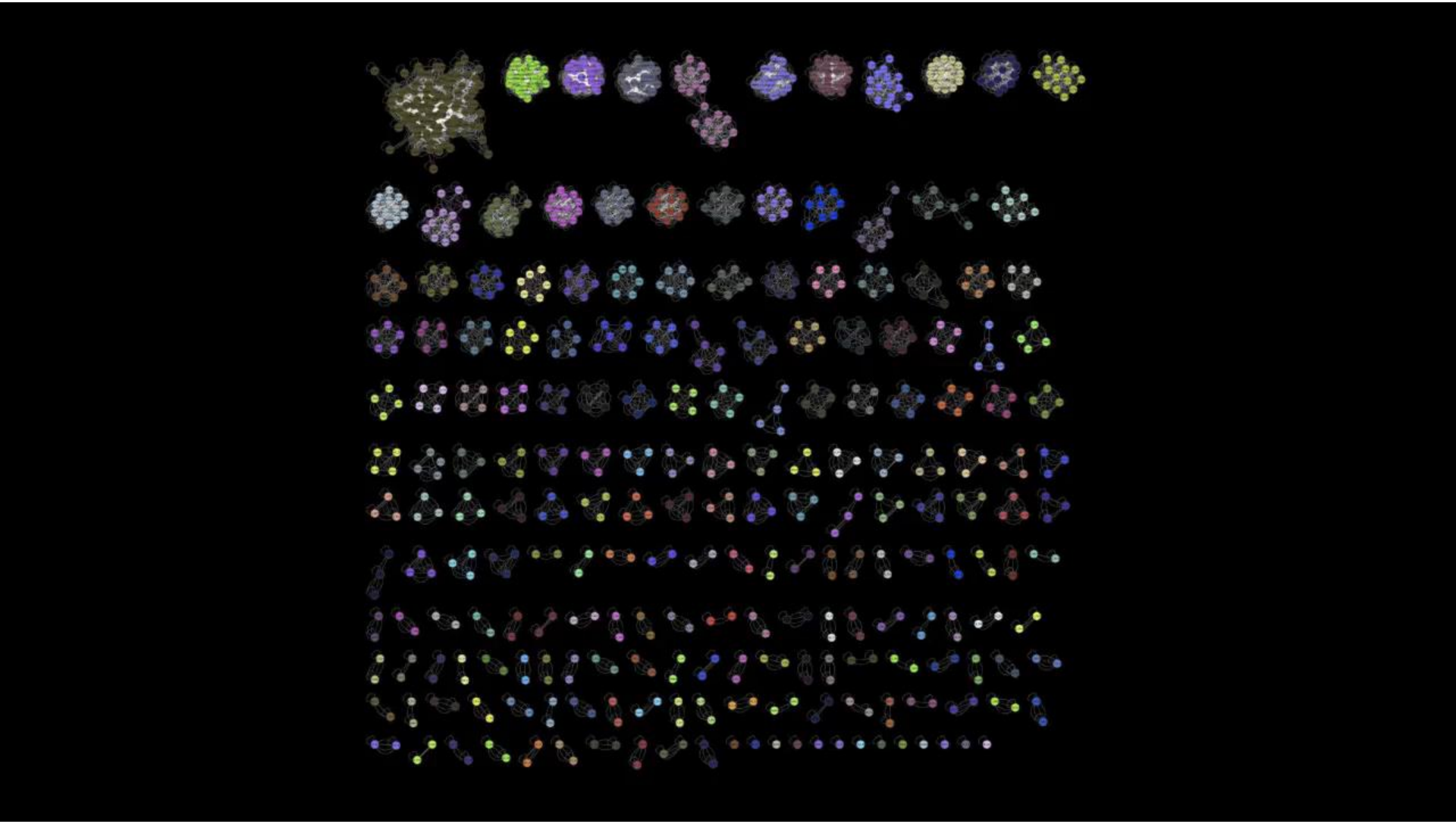




Animation

- Useful to show changes in a network:
  - Over a time series
  - Over different conditions
  - Between species

Animation



Questions?

stringApp exercise 3

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

In this exercise, we will continue to analyze the network of differentially abundant proteins from Francavilla *et al.* study.

Prerequisites: *install the app **ClusterMaker2** from the App store!*

3.1 Network clustering

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

3.2 Subnetworks and physical interactions

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

Question 3: *How many edges does the resulting network contain and why are there now fewer edges?*

3.3 Functional enrichment

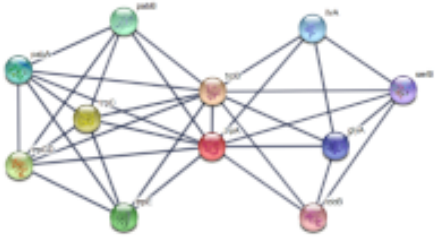
Question 4: *How many statistically significant terms are in the table? Which is the most significant term for each of the categories GO Biological Process, GO Molecular Function, and KEGG Pathways?*

3.4 Enriched publications

Question 5: *What is the title of the most recent publication?*

Supporting lectures

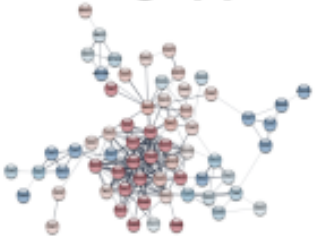
The STRING database



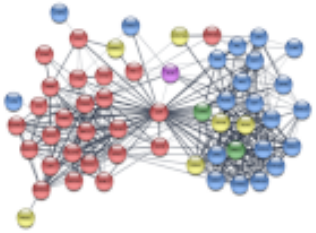
The Cytoscape platform

Cytoscape stringApp

Basic stringApp tutorial



The DISEASES database



Enrichment analysis

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

Tutorials & getting help

- STRING & stringApp:
  - YouTube videos: https://www.youtube.com/c/LarsJuhlJensen
  - Tutorials & exercises: https://jensenlab.org/training/
  - Automation repository: https://github.com/scaramonche/EuBIC2020_Cytoscape
- Cytoscape
  - Tutorials: https://github.com/cytoscape/cytoscape-tutorials/wiki
  - YouTube videos: https://www.youtube.com/channel/UCv6auk9FK4NgXiXiqrDLccw
  - Helpdesk mailing list: cytoscape-helpdesk@googlegroups.com
  - Publications using Cytoscape: https://cytoscape-publications.tumblr.com/
  - Automation: https://github.com/cytoscape/cytoscape-automation/wiki

Hands-on exercises

- Continue with the online exercises at https://jensenlab.org/training/
- Work with your own data by adapting the exercises
- Try out one of the Cytoscape tutorials at https://github.com/cytoscape/cytoscape-tutorials/wiki
  - E.g. Basic data visualization, RNA-seq data analysis, Using WikiPathways App
- Use R or Python to access Cytoscape
  - Official R2Cy tutorial using stringApp: https://cytoscape.org/cytoscape-automation/for-scripters/R/notebooks/stringApp.nb.html
  - More automation tutorials at https://github.com/cytoscape/cytoscape-automation/wiki