Bioinformatical analysis of omics expression data
Part 5

Dr. Michael Turewicz[1,2]

[1]Institut für Klinische Biochemie und Pathobiochemie,
Deutsches Diabetes-Zentrum, Leibniz-Zentrum für Diabetesforschung an der Heinrich-Heine-Universität, Düsseldorf, Deutschland
[2]Deutsches Zentrum für Diabetesforschung (DZD), München-Neuherberg, Deutschland

Course schedule

- Part 1 (25.10.23)
  - Introduction (omics, example data, programming)
  - Data preprocessing (data inspection, normalization, missing values)
  - Exercises: R programming tutorial (part 1)
- Part 2 (08.11.23)
  - Differential expression analysis (statistics, volcano plot)
  - Exercises: R programming tutorial (part 2)
- Part 3 (15.11.23)
  - Machine learning I: Clustering (clustering, PCA)
  - Exercises: Customized hierarchical clustering & PCA in R
- Part 4 (22.11.23)
  - Overrepresentation analysis (GO, Reactome)
  - Exercises: Own GO- & Reactome analysis in R & other tools
- **Part 5 (29.11.23)**
  - **Network analysis (STRING, Cytoscape)**
  - **Exercises: Own network analysis in R & STRING**
- Part 6 (06.12.23)
  - Machine learning II: Classification algorithms

Recap of previous part

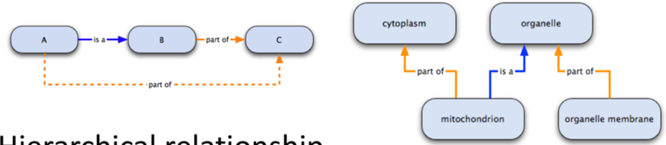
DDZ Deutsches Diabetes-Zentrum

- ## Gene Ontology (GO)
  - Organism-specific hierarchy of curated biol. terms
  - → directed acyclic graph (DAG) of terms (= nodes)
  - → edges: “is a”- & ”part of”-relationships
  - close to DAG-”root” general terms & terminal nodes most specific
  - Organized in 3 GO domains (separate DAGs): biological process, molecular function, cellular component

- ## Overrepresentation analysis (ORA)
  - **Basic idea:**
    1. Annotate input genes/proteins with the terms with which they are annotated in a biological database
    2. Return terms that are **statistically significantly (→ p-value) overrepresented** in input
  - **Statistical tests:** based on the urn model, e.g. Fisher’s exact test, Kolmogorov-Smirnov test, (…)
  - Popular biological knowledge DBs: GO, Reactome, KEGG, WikiPathways, PhosphoSitePlus, (…)

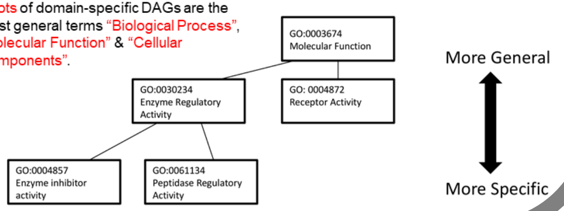
- ## Programming (example: R)
  - Own GO- & Reactome-based ORA + visualization in R & online tools

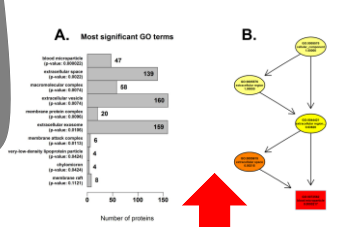
- Based on “is a” or “part of” relationship



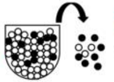
- Hierarchical relationship

Roots of domain-specific DAGs are the most general terms “Biological Process”, “Molecular Function” & “Cellular Components”.





- Fisher’s Exact Test (Hypergeometric Test)

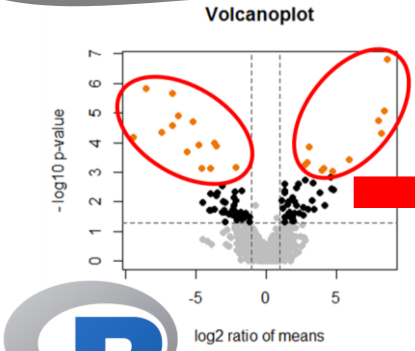


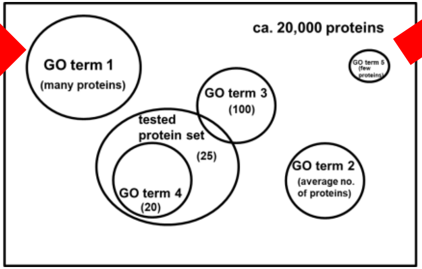
The test implements the urn model.

What is the probability of getting 7 or more black balls?

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

- **N:** total number of proteins
- **M:** total number of proteins annotated with this term
- **n:** number of proteins in the set (all balls drawn)
- **k:** number of proteins in tested set annotated with this term (black balls drawn)





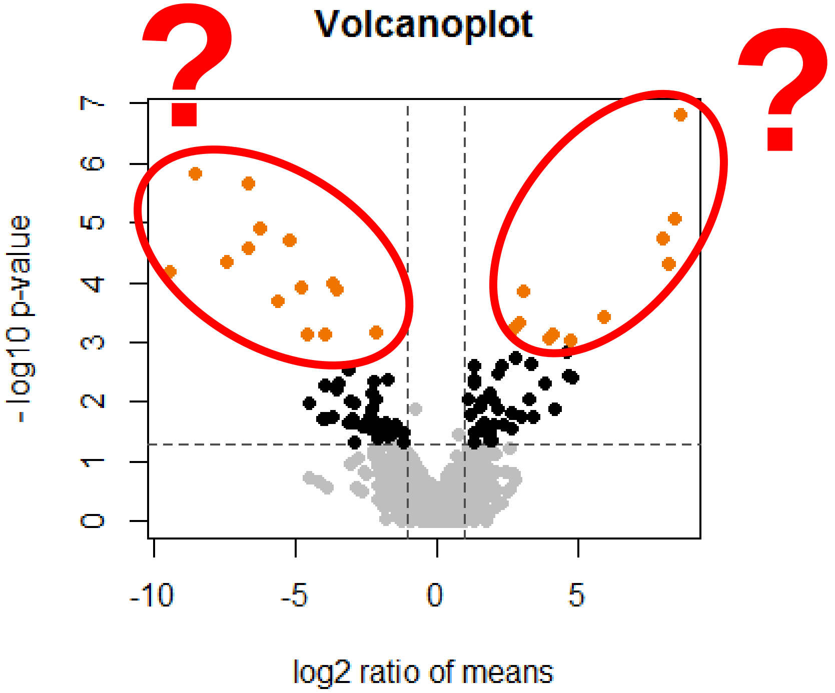
List of candidates: what's next?

- We have learned to find a list of statistically significant differential candidates with p-values and fold changes

- How can we interpret these biologically?

- Are there biological connections that could explain a common occurrence?

- Can a common function be assumed?

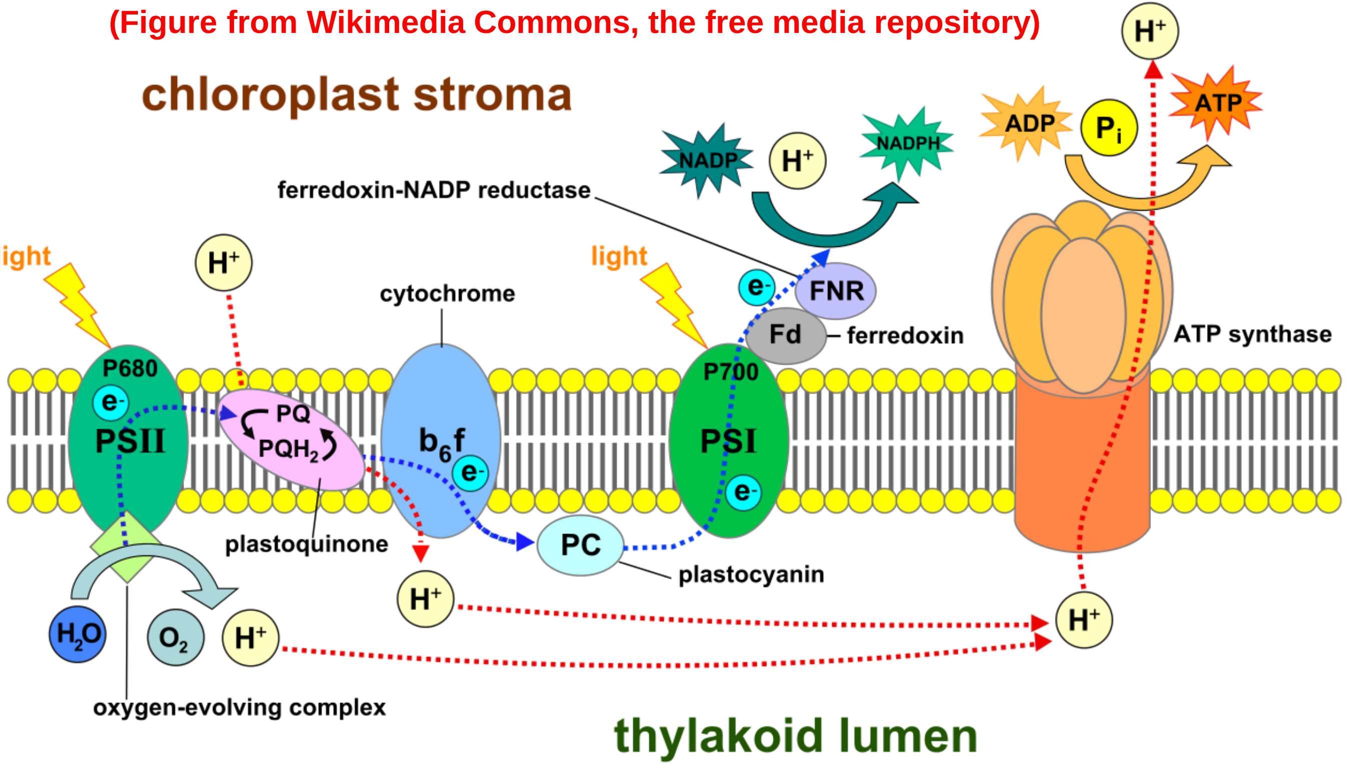


Motivation: Protein-protein interactions (PPIs)

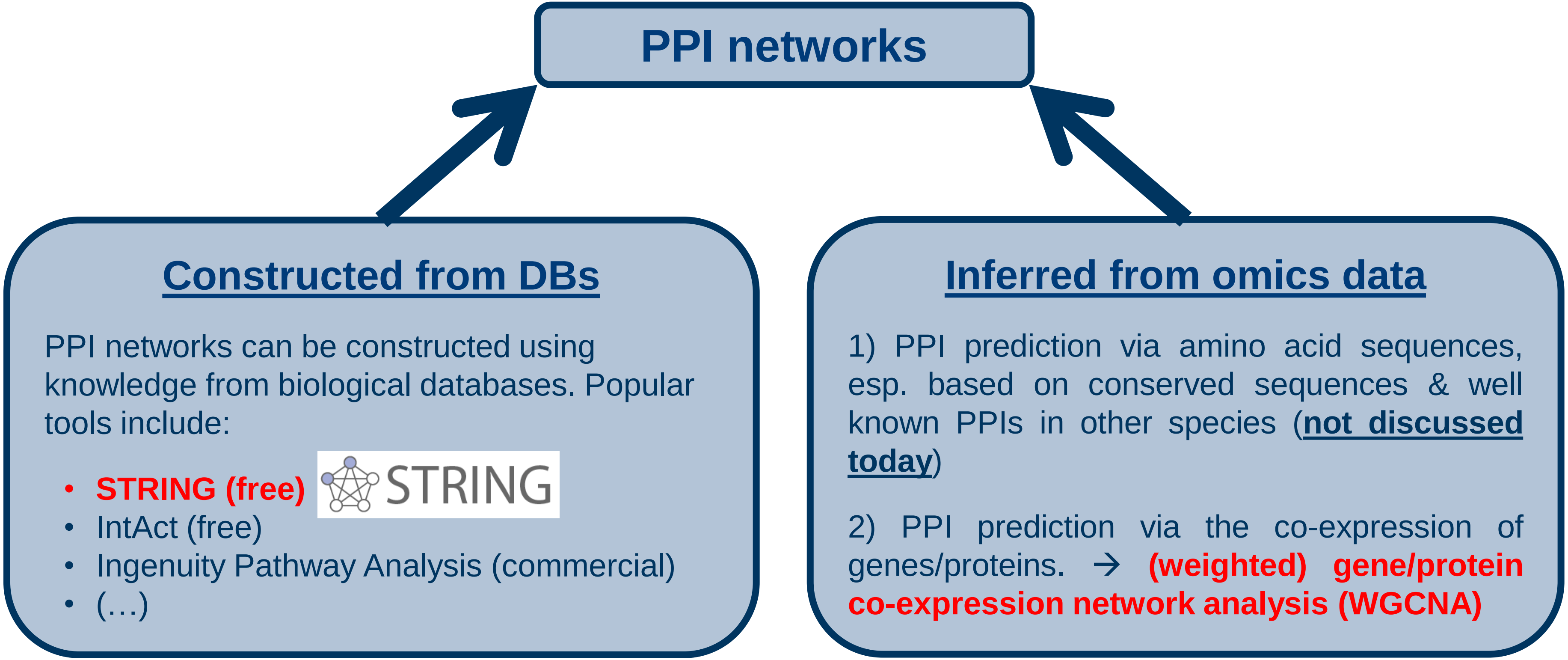
Deutsches Diabetes-Zentrum

Example: Photosynthetic electron transport chain of the thylakoid membrane.

(Figure from Wikimedia Commons, the free media repository)



PPI networks: basic concepts



PPI networks

Constructed from DBs

PPI networks can be constructed using knowledge from biological databases. Popular tools include:

- **STRING (free)**
- IntAct (free)
- Ingenuity Pathway Analysis (commercial)
- (…)

Inferred from omics data

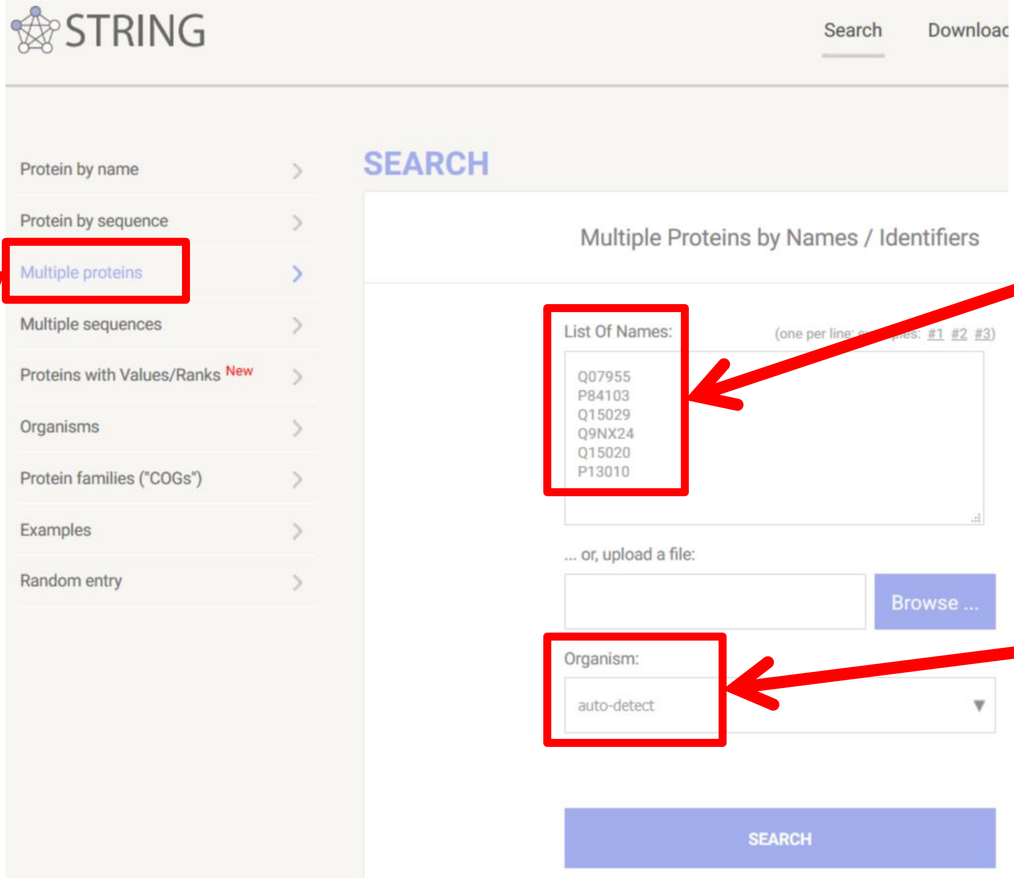
1) PPI prediction via amino acid sequences, esp. based on conserved sequences & well known PPIs in other species (**not discussed today**)

2) PPI prediction via the co-expression of genes/proteins. → **(weighted) gene/protein co-expression network analysis (WGCNA)**

Networks from DBs: STRING

- **STRING (https://string-db.org/):**
  - online tool for PPI network analysis (& more).
  - widely used & free.
- STRING is a DB of **known & predicted PPIs**.
- Interactions are derived from:
  - Genomic context predictions
  - High-throughput lab experiments
  - (Conserved) co-expression
  - Text mining (PubMed, OMIM, …)
  - Knowledge in DBs (Reactome, DIP, BioGRID, MINT, Gene Ontology, KEGG, PDB, UniProt, …)
- Version 11.0:
  - 5090 organisms
  - > 24.6 millions proteins
  - > 3,000 millions interactions
- **→ Search for PPIs for single or multiple proteins & visualization as a network.**

STRING: search form



List Of Names: (one per line, examples: #1 #2 #3)

Q07955
P84103
Q15029
Q9NX24
Q15020
P13010

Organism: auto-detect

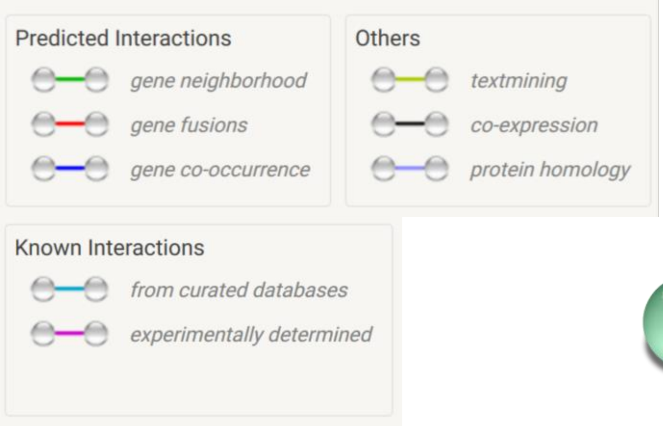
Various search modes available, e.g. “multiple proteins” to analyze list of candidates

Candidate list: gene/protein names or IDs (here: UniProt IDs)

Organism

STRING: network (evidence view)

Kinds of interaction:



Predicted Interactions

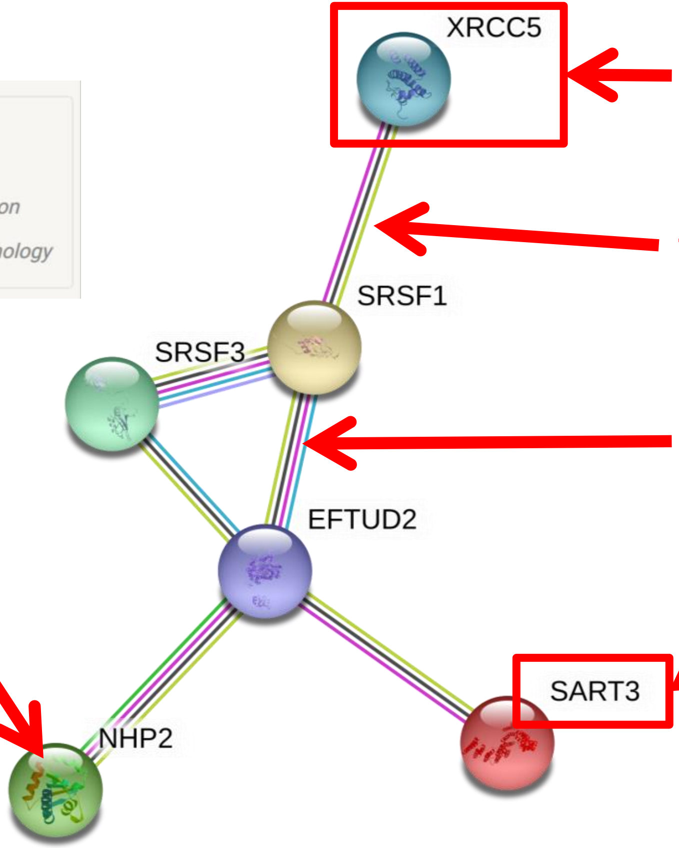
- 🟩 *gene neighborhood*
- 🟥 *gene fusions*
- 🟦 *gene co-occurrence*

Others

- 🟨 *textmining*
- ⬛ *co-expression*
- 🟪 *protein homology*

Known Interactions

- 🟦 *from curated databases*
- 🟪 *experimentally determined*



Nodes: Proteins from the given protein set.

Edges: Interactions between the proteins within the STRING database.

Multiple edges: Each color represents a different kind of interaction.

Gene names: Proteins are represented by their genes.

Preview of 3D-structure of the respective protein visualized in the node.

STRING: basic settings

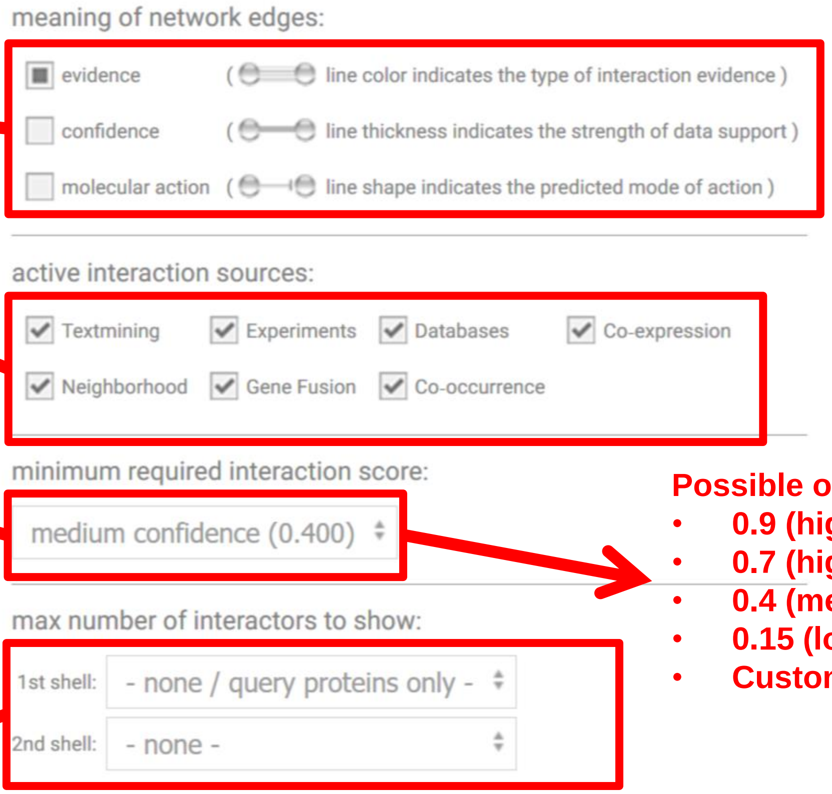
Different network views:

- <u>evidence view:</u> edges show types of interaction evidence
- <u>confidence view:</u> edges represent confidence score

Interaction sources can be included/excluded into/from analysis

Edges below this confidence score are not shown.

Optional network-extension by proteins not included in input list: in 2 layers a user-defined number of most confident interactors from STRING DB can be added.



meaning of network edges:

■ evidence (line color indicates the type of interaction evidence)

□ confidence (line thickness indicates the strength of data support)

□ molecular action (line shape indicates the predicted mode of action)

active interaction sources:

☑ Textmining ☑ Experiments ☑ Databases ☑ Co-expression

☑ Neighborhood ☑ Gene Fusion ☑ Co-occurrence

minimum required interaction score:

medium confidence (0.400)

max number of interactors to show:

1st shell: - none / query proteins only -

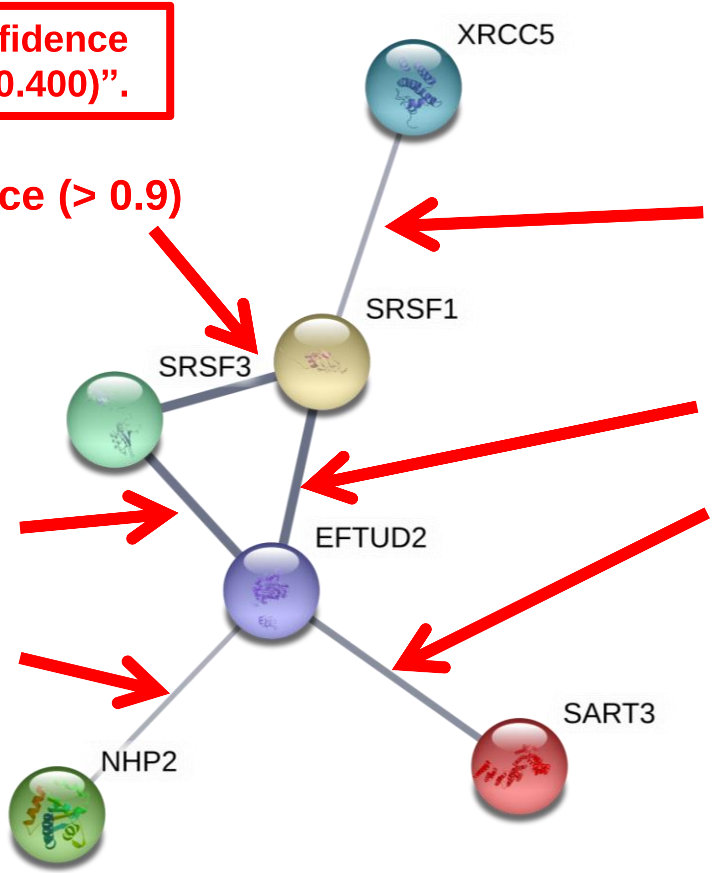
2nd shell: - none -

Possible options:

- 0.9 (highest)
- 0.7 (high)
- 0.4 (medium)
- 0.15 (low)
- Custom value

STRING: network (confidence view)

Confidence view with default confidence threshold: “medium confidence (0.400)”.



Highest confidence (> 0.9)

Medium confidence (> 0.4)

Highest confidence (> 0.9)

High confidence (> 0.7)

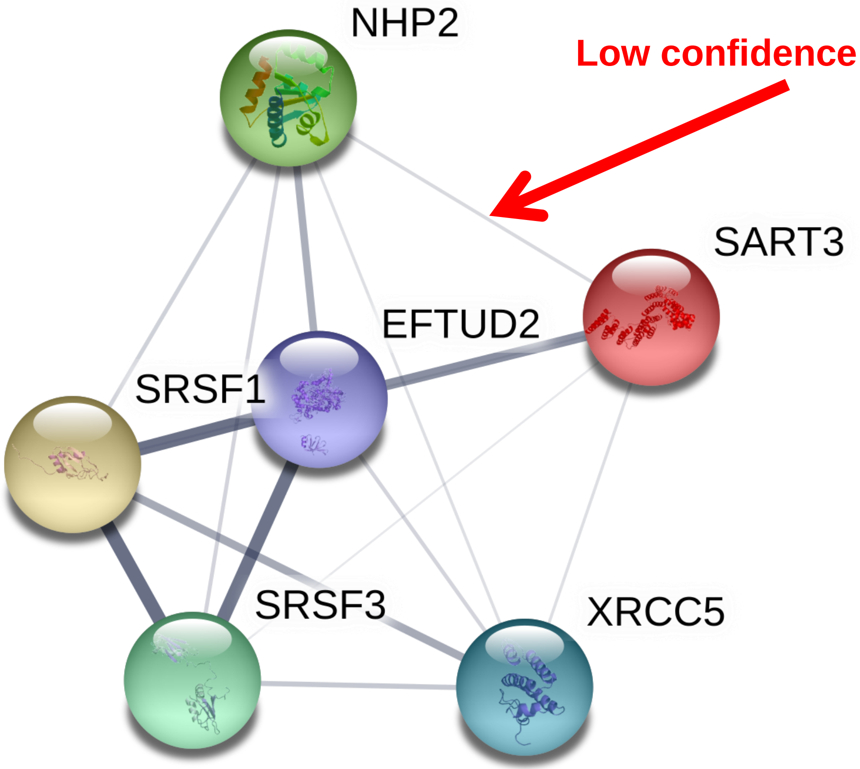
Highest confidence (> 0.9)

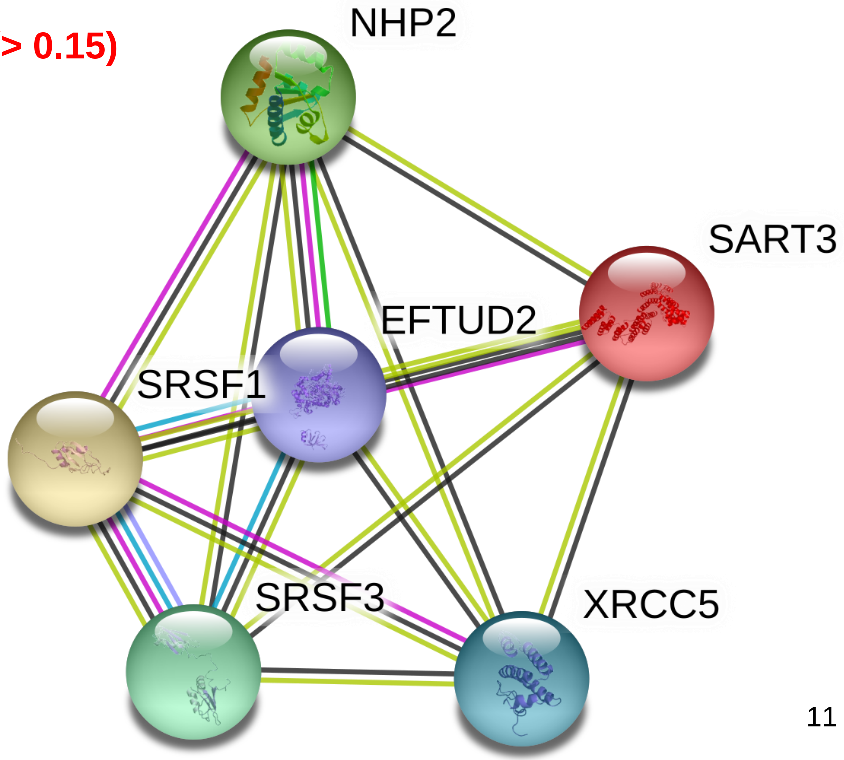
Medium confidence (> 0.4)

STRING: lax settings

Settings: “low confidence (0.150)” & all kinds of interactions

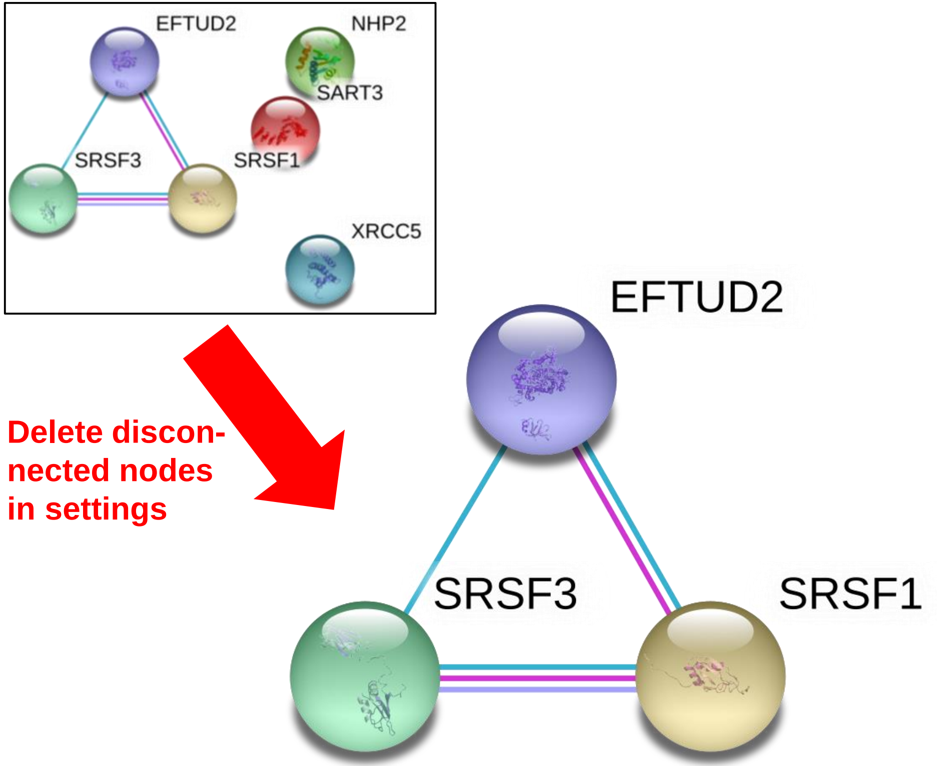
Low confidence (> 0.15)



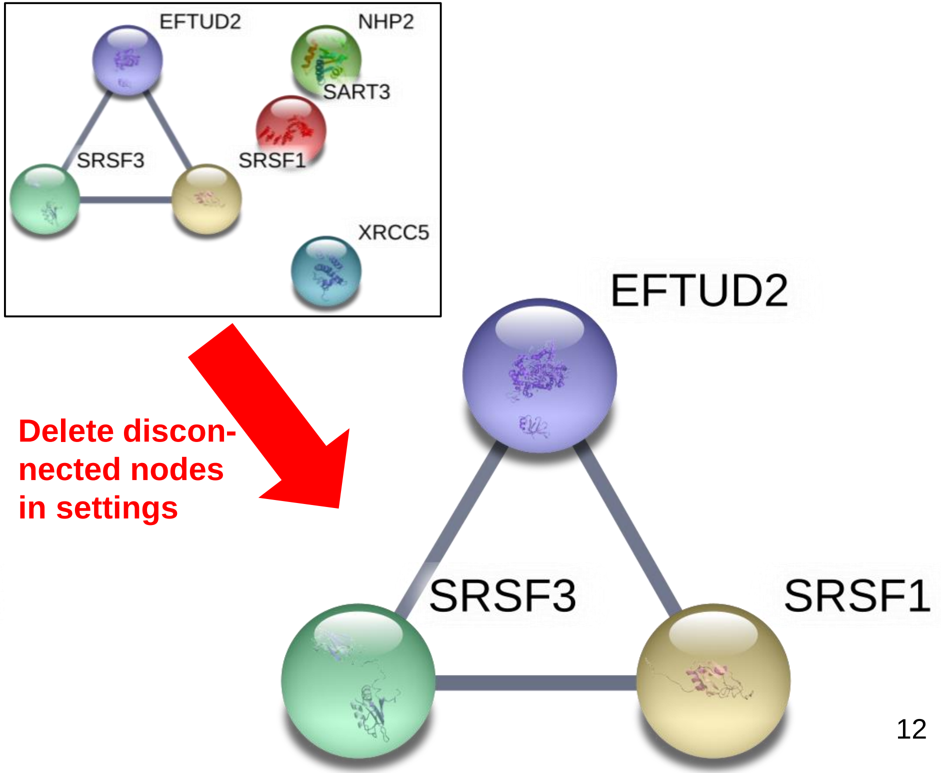


STRING: strict settings

Settings: "highest confidence (0.900)" & only known interactions (DBs & experiments)

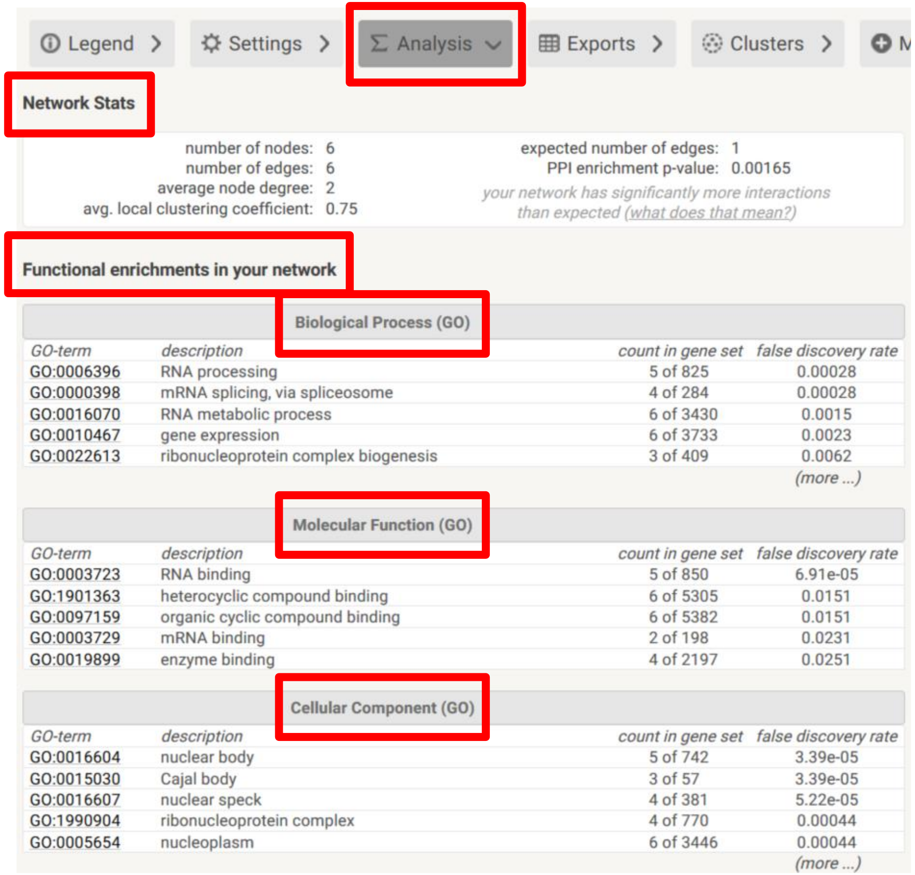


Delete disconnected nodes in settings



Delete disconnected nodes in settings

STRING: network statistics & ORA



Network Stats

number of nodes:	6	expected number of edges:	1
number of edges:	6	PPI enrichment p-value:	0.00165
average node degree:	2		
avg. local clustering coefficient:	0.75		

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

Functional enrichments in your network

Biological Process (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0006396	RNA processing	5 of 825	0.00028
GO:0000398	mRNA splicing, via spliceosome	4 of 284	0.00028
GO:0016070	RNA metabolic process	6 of 3430	0.0015
GO:0010467	gene expression	6 of 3733	0.0023
GO:0022613	ribonucleoprotein complex biogenesis	3 of 409	0.0062

(more ...)

Molecular Function (GO)

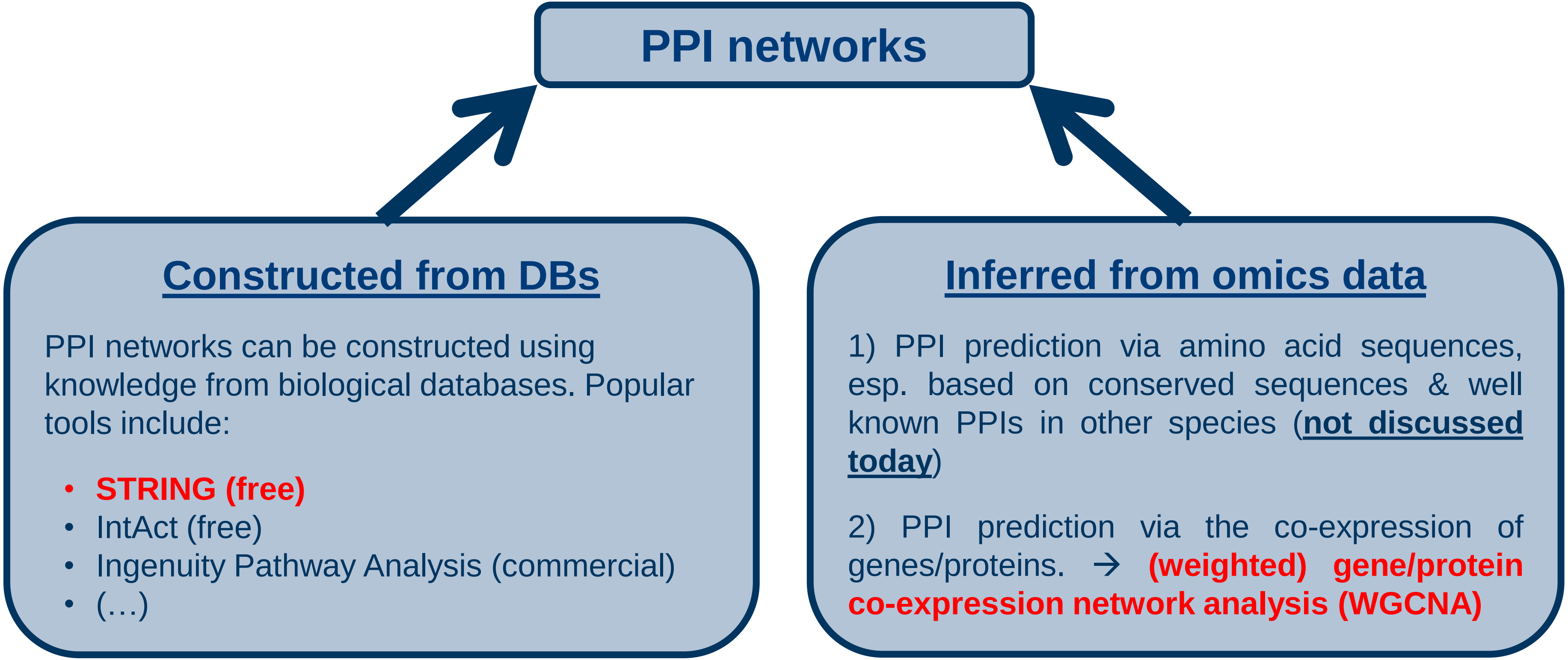
GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0003723	RNA binding	5 of 850	6.91e-05
GO:1901363	heterocyclic compound binding	6 of 5305	0.0151
GO:0097159	organic cyclic compound binding	6 of 5382	0.0151
GO:0003729	mRNA binding	2 of 198	0.0231
GO:0019899	enzyme binding	4 of 2197	0.0251

Cellular Component (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0016604	nuclear body	5 of 742	3.39e-05
GO:0015030	Cajal body	3 of 57	3.39e-05
GO:0016607	nuclear speck	4 of 381	5.22e-05
GO:1990904	ribonucleoprotein complex	4 of 770	0.00044
GO:0005654	nucleoplasm	6 of 3446	0.00044

(more ...)

PPI networks: basic concepts



PPI networks

Constructed from DBs

PPI networks can be constructed using knowledge from biological databases. Popular tools include:

- **STRING (free)**
- IntAct (free)
- Ingenuity Pathway Analysis (commercial)
- (…)

Inferred from omics data

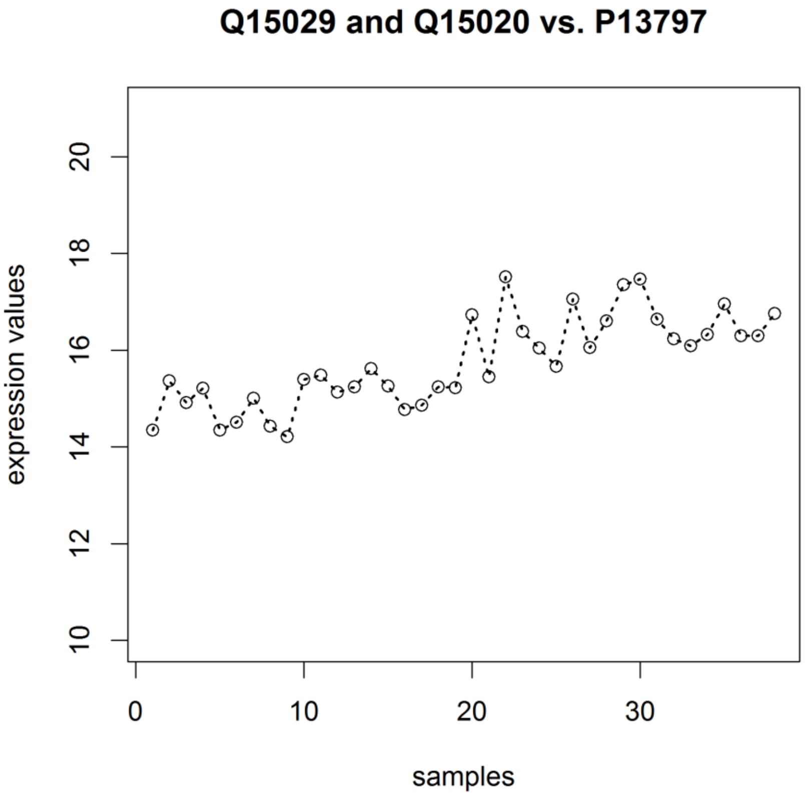
1) PPI prediction via amino acid sequences, esp. based on conserved sequences & well known PPIs in other species (**not discussed today**)

2) PPI prediction via the co-expression of genes/proteins. → **(weighted) gene/protein co-expression network analysis (WGCNA)**

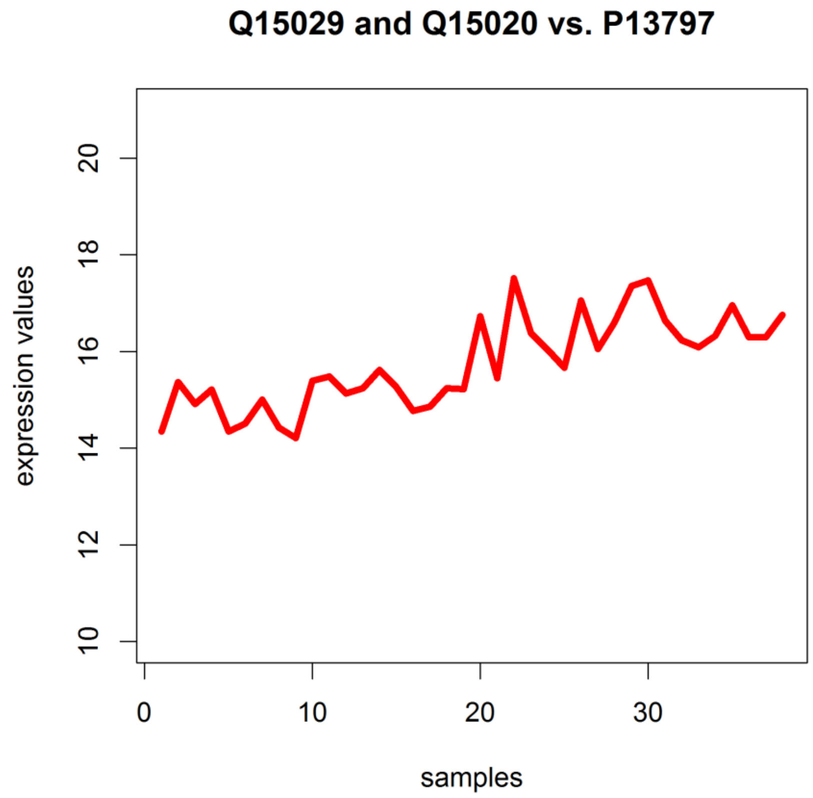
Example dataset

- Hepatocellular carcinoma (HCC) → liver cancer
- 19 HCC vs. 19 controls (C) → healthy samples
- Obtained from Naboulsi et al., J. Proteome Res. 2016
- PRIDE: PXD002171
- Label-free quantification of 2,736 proteins via Progenesis QI software

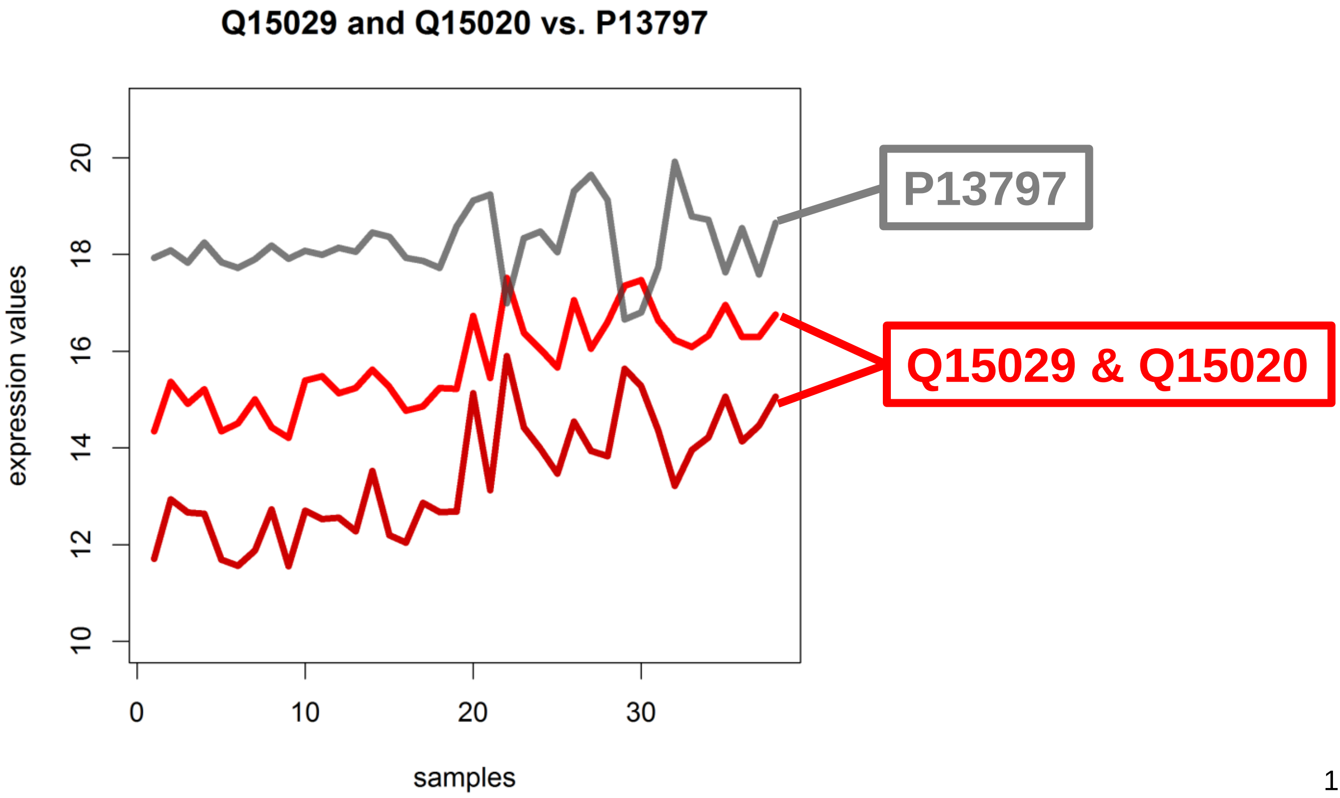
Protein expression profiles



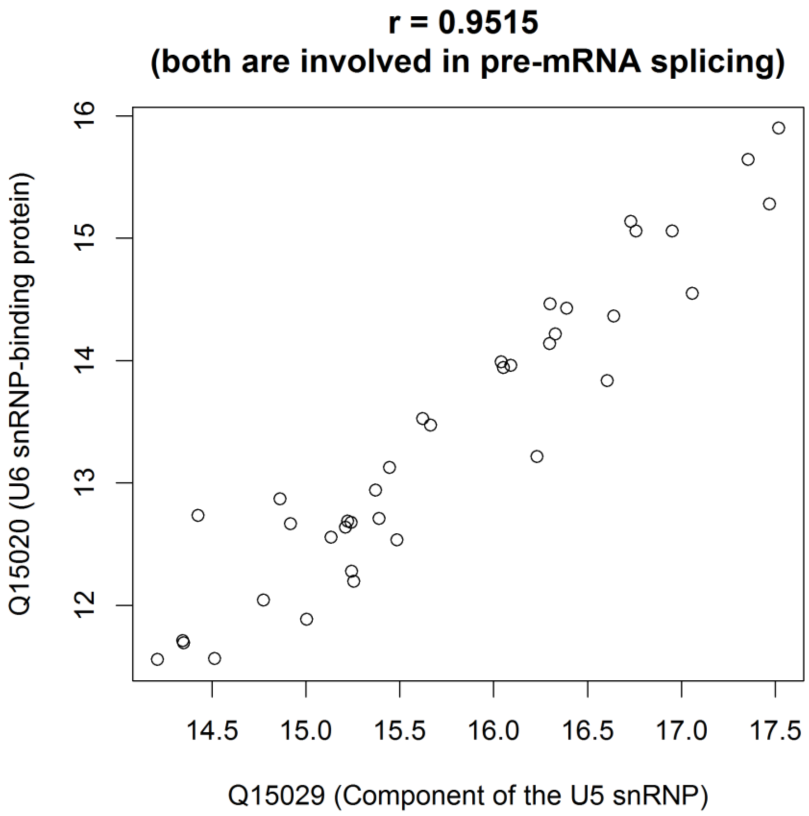
Protein expression profiles



Protein expression profiles



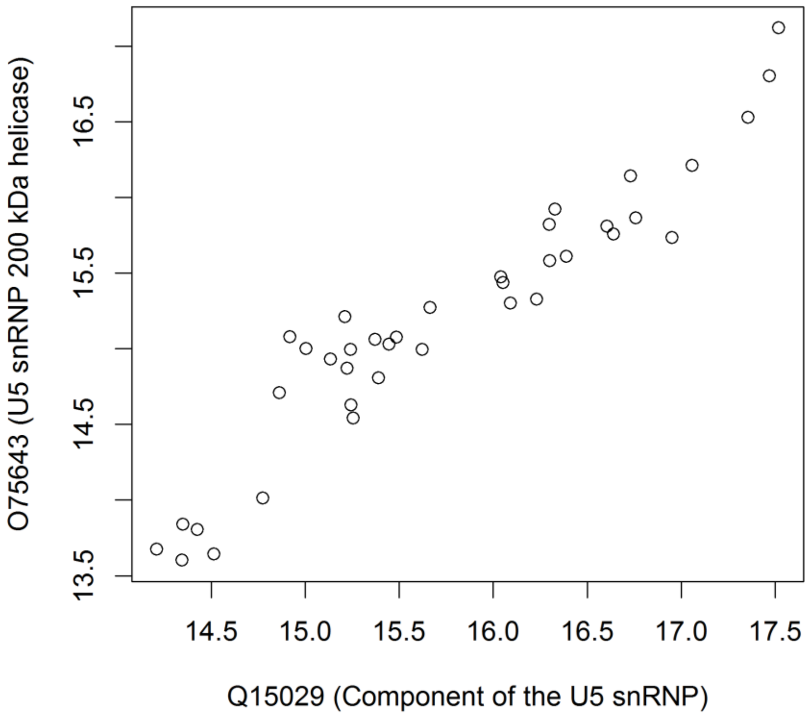
Measure for co-expression: Spearman’s rho



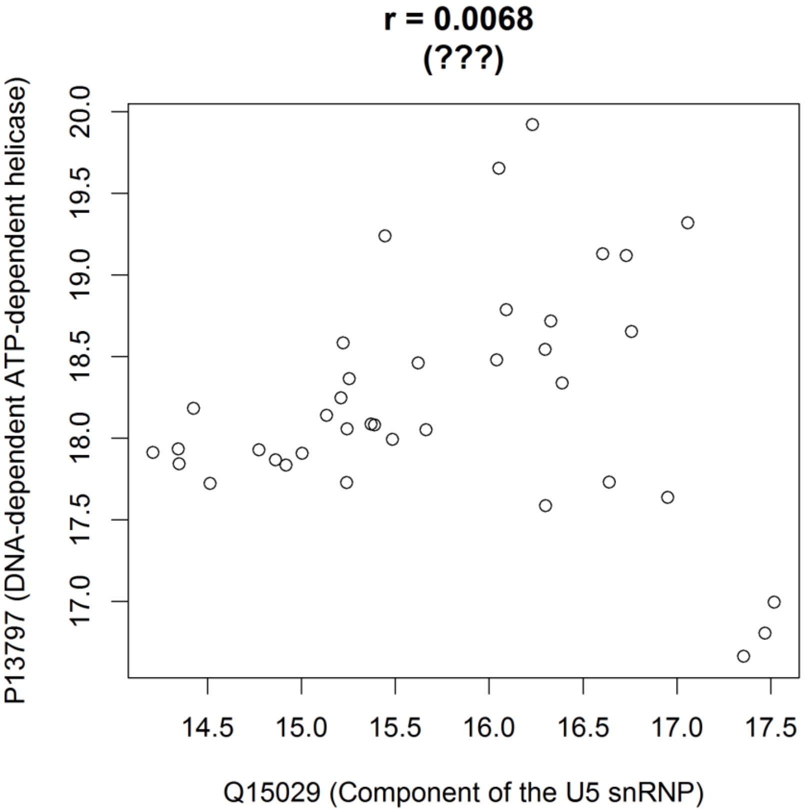
Q15020 vs. Q15029

Measure for co-expression: Spearman’s rho

r = 0.951
(both are involved in pre-mRNA splicing)



Measure for co-expression: Spearman’s rho



P13797 vs. Q15020

Protein co-expression network inference

- Let $P = \{p_1, \dots, p_5\}$ be a set of protein expression profiles (from quant. proteomics)
- Consider pairs of protein expression profiles $p_i, p_j \epsilon P$.
- Correlation-based measure for co-expression: $a_{ij} = \left|cor\left(p_i, p_j\right)\right|^{\beta}$, $\beta \epsilon \{1,2,3, \dots\}$
- Adjacency matrix: $\mathrm{A} = \left[a_{ij}\right] = \begin{pmatrix} a_{11} & \cdots & a_{15} \\ \vdots & \ddots & \vdots \\ a_{51} & \cdots & a_{55} \end{pmatrix}$
- Unweighted network: $G = (P, A)$ after setting $a_{ij} = \begin{cases} 1 \ \forall a_{ij} \geq \tau \\ 0 \ \forall a_{ij} < \tau \end{cases}$

1 = “edge”
0 = “no edge”

computing A

	p_1	p_2	p_3	p_4	p_5
p_1	1	0.3	0.6	0.4	0.81
p_2	0.3	1	0.5	0.3	0.94
p_3	0.6	0.5	1	0.6	0.86
p_4	0.4	0.3	0.6	1	0.87
p_5	0.81	0.94	0.86	0.87	1

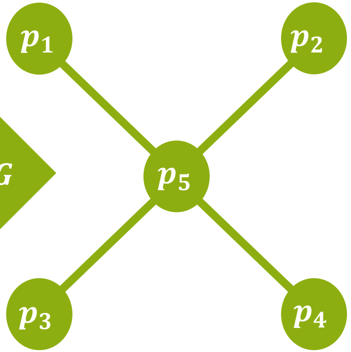
A

$\tau = 0.8$

	p_1	p_2	p_3	p_4	p_5
p_1	1	0	0	0	1
p_2	0	1	0	0	1
p_3	0	0	1	0	1
p_4	0	0	0	1	1
p_5	1	1	1	1	1

A

plotting G

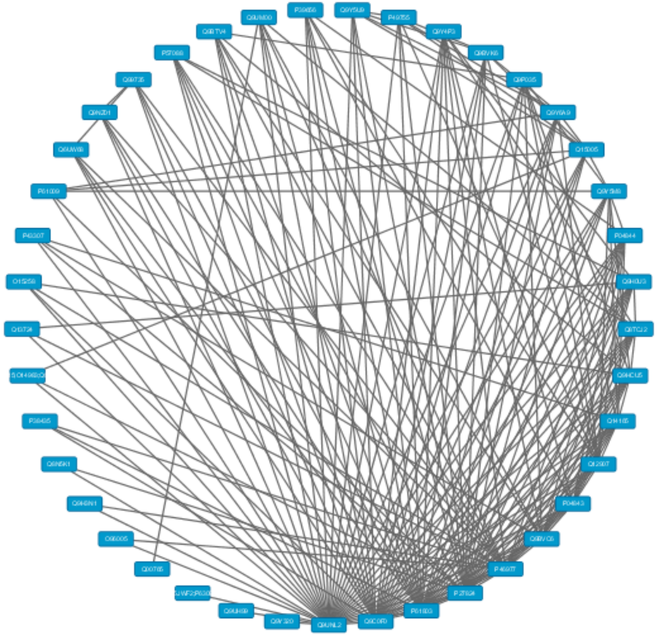


G

Protein co-expression networks

Deutsches Diabetes-Zentrum

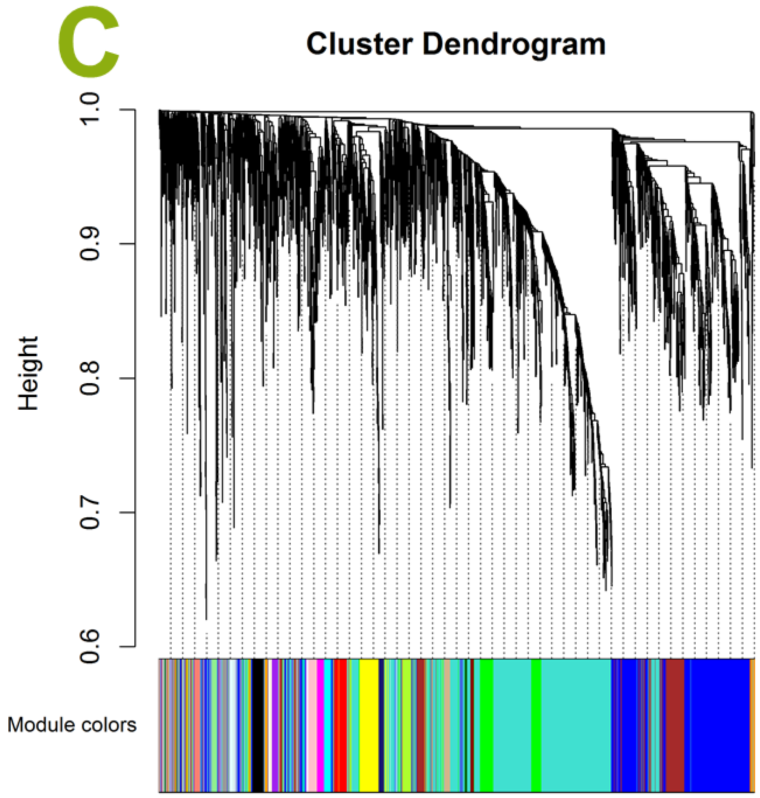
- **Module:** Cluster of interconnected (i.e., co-expressed) proteins.
- Modules can represent **pathways** & **hub proteins** (most interconnected proteins) can represent their modulators.
- **Topological Overlap:** Similarity measure for the “interconnectedness” between two proteins based on the number of shared neighbors.
- **TOM (Topological Overlap Matrix):** Matrix $\Omega = [\omega_{ij}]$ containing the topological overlap between all proteins. Used for the detection of modules.

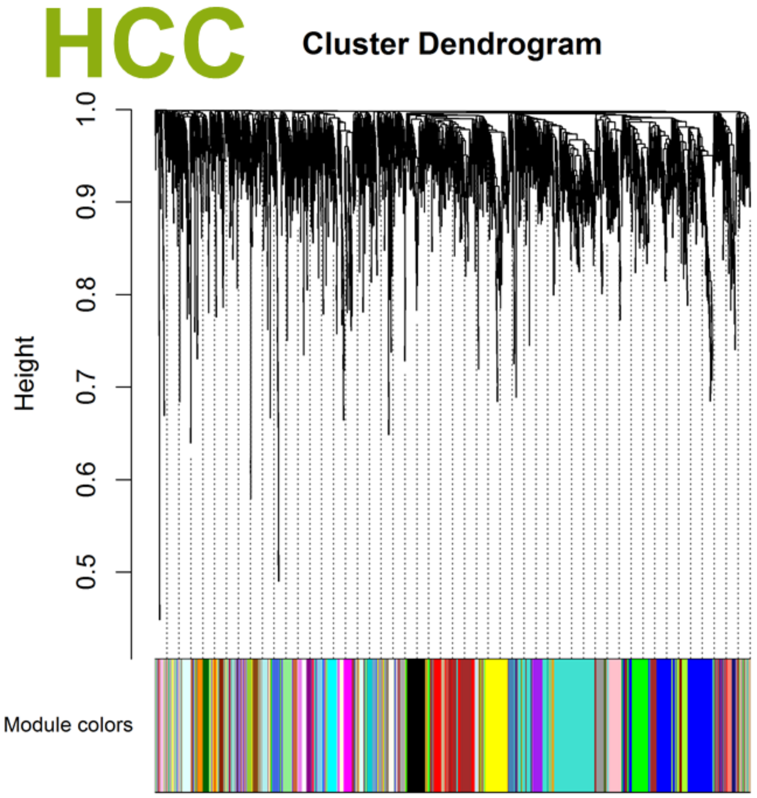


$$\boxed{\omega_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}}}$$

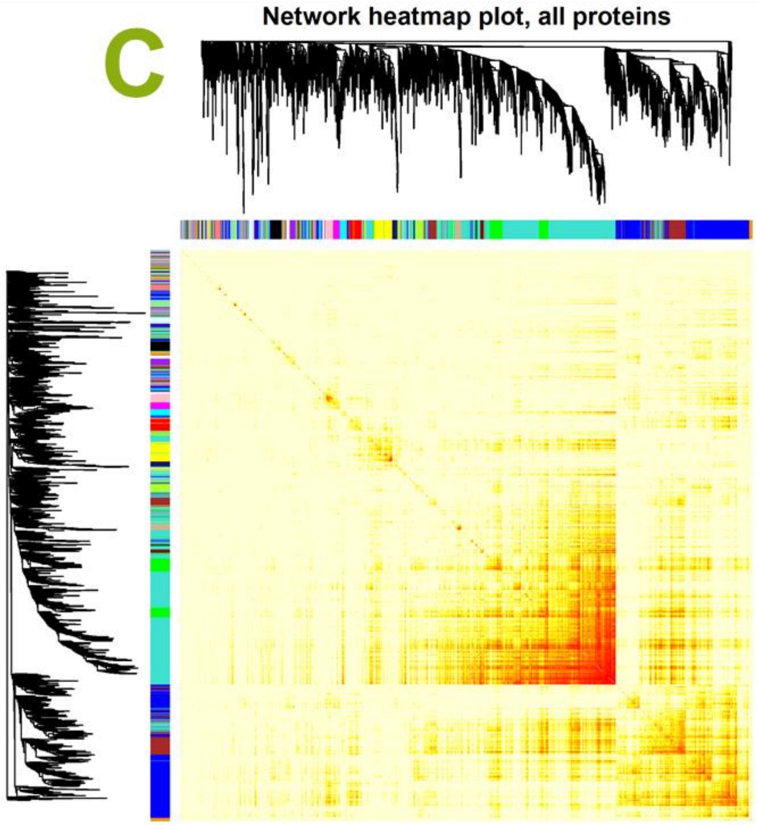
Distance metric for clustering.

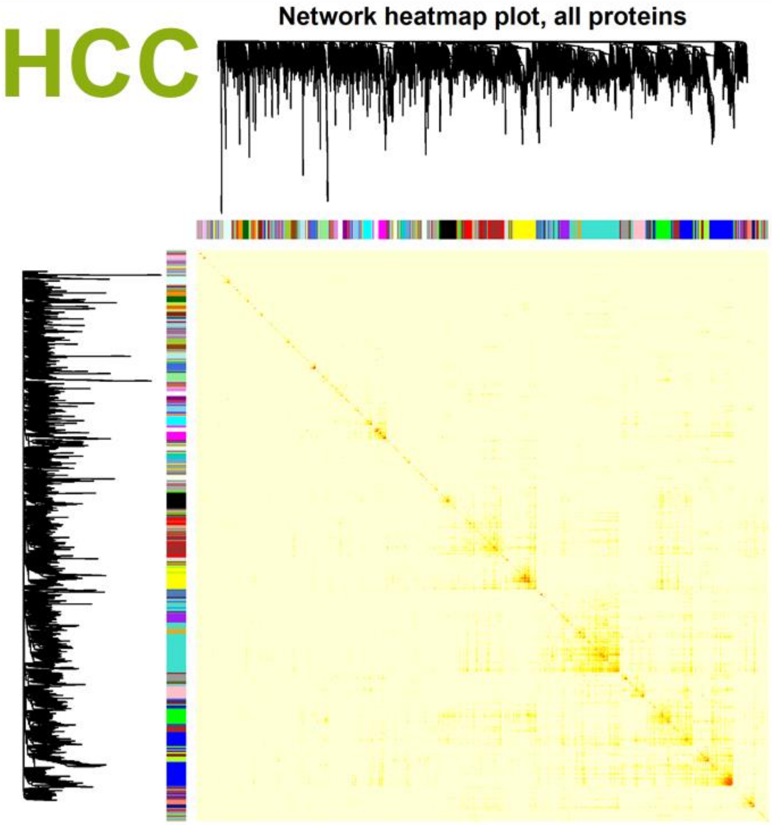
Comparing group-specific networks



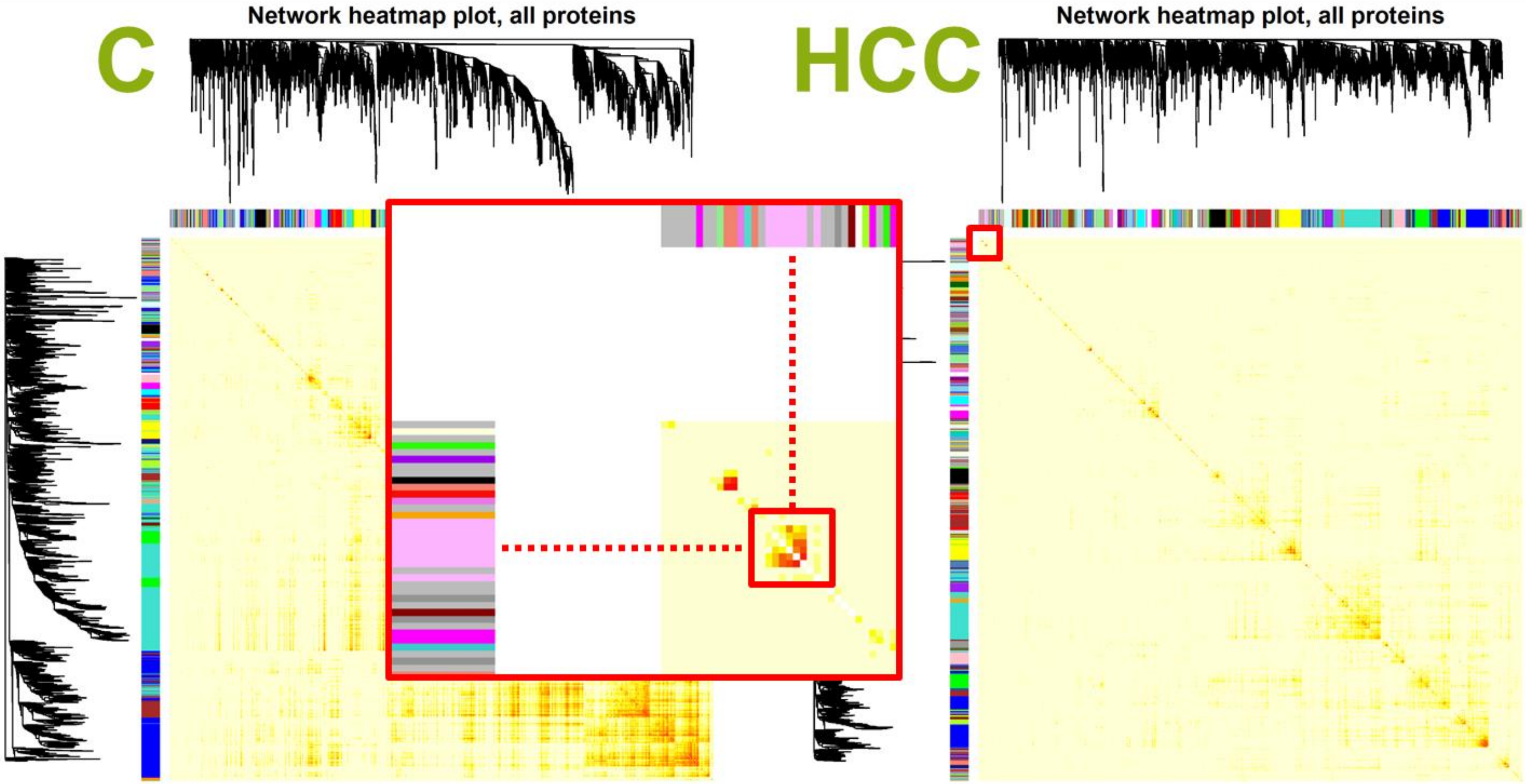


Comparing group-specific networks

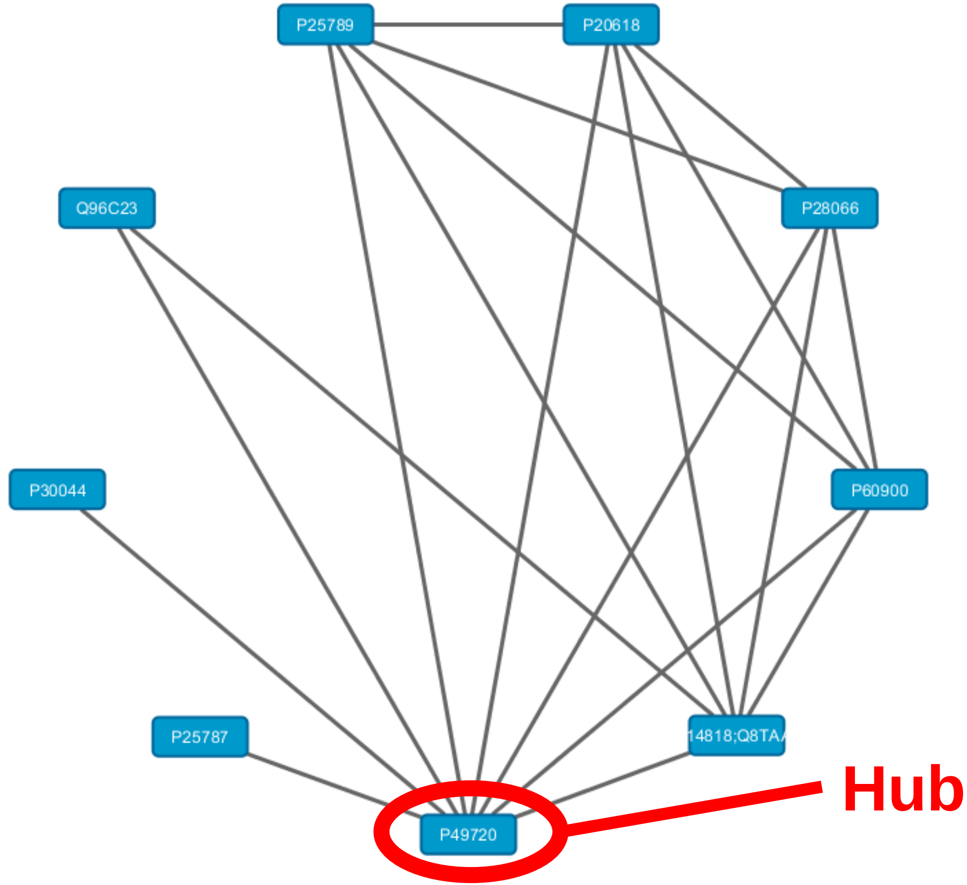




Comparing group-specific networks

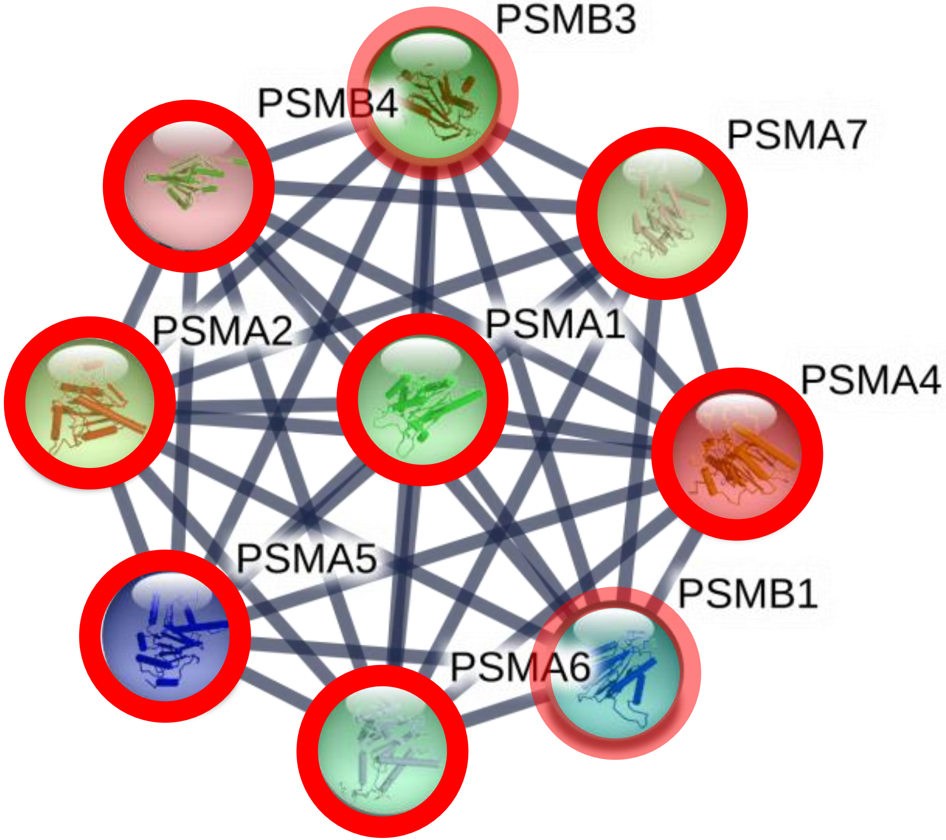


(Sub-)network visualization



Hub protein

Network interpretation (STRING)



Enrichment results (adjusted p-values):

- GO (cellular component): "**proteasome core complex, alpha subunit comples**" (p = 3.66e-14)
- KEGG pathway: "**proteasome**" (p = 1.54e-12)
- PFAM domains: "**proteasome subunit**" (p = 1.3e-17)
- INTERPRO domains: "**Proteasome, subunit alpha/beta**" (p = 2.77e-17)

Components of the ubiquitin-proteasome pathway are known targets for cancer therapy (proteasome inhibitors) – also in discussion for HCC…

○ **= proteasome subunit**

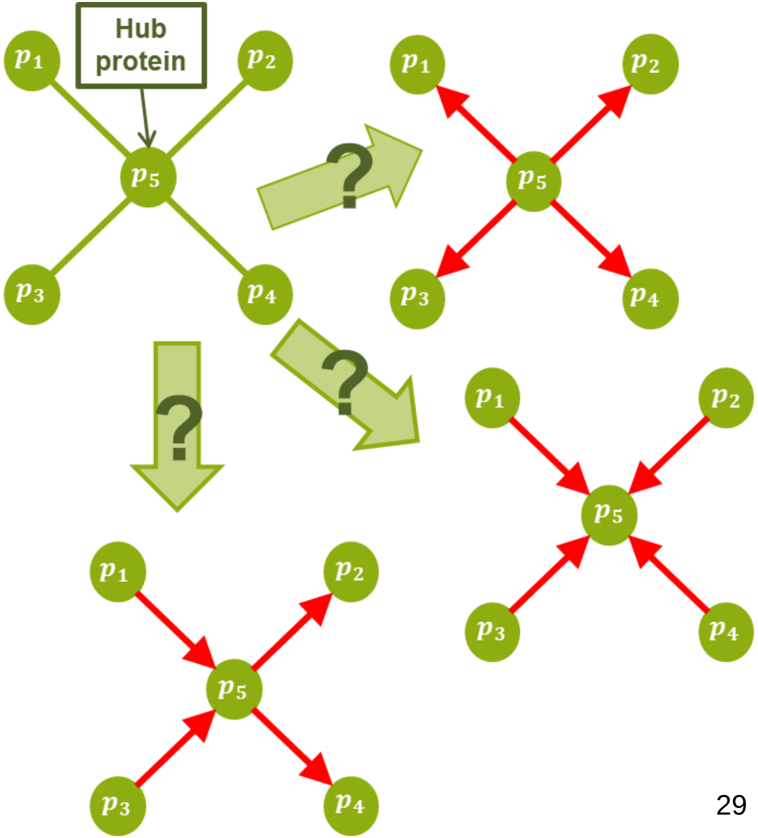
Network interpretation

Potential reasons for co-expression:

- Direct protein-protein interaction
- Target protein ←→ protease
- Target protein ←→ kinase/phosphorylase
- Receptor protein ←→ effector protein
- Signaling complexes
- Scaffold protein complexes
- Target protein ←→ transcription factor
- (…)

Elucidation by:

- Best way: experiments!
- Protein annotation (e.g., GO-/Reactome-ORA)
- PPI annotation
- Protein module identification & characterization
- Identification & characterization of hub proteins
- Differential network analysis
- (…)



Hands on part!

Exercises

- **Exercise 5**
  - https://drive.google.com/drive/folders/1vmewprs0gkpakU8idbgtexDIwmGVUJz3?usp=sharing
  - Use our example dataset from GitHub for the following exercises
  - **Exercise 5.1:** Perform an own STRING network analysis in R using differential candidates and confidence view (confidence threshold = 0.7). Visualize the network without not connected nodes.
  - **Exercise 5.2:** Perform an own WCGNA in R and find an biologically interesting module. You can use STRING analysis (via searching the UniProt IDs in the STRING web application) to quickly check whether a module is interesting (i.e. highly confident STRING interactions & interesting ORA results in STRING).
  - Please send me your solutions as an “.R”-file

Thank you!

UKD Universitätsklinikum Düsseldorf

Leibniz Gemeinschaft

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

Bundesministerium für Gesundheit

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

Bundesministerium für Bildung und Forschung

DZD Deutsches Zentrum für Diabetesforschung