Bioinformatics I

Quantitative Proteomics: Biological Interpretation:

Veit Schwämmle

Protein Research Group
BMB
Syddansk Universitet

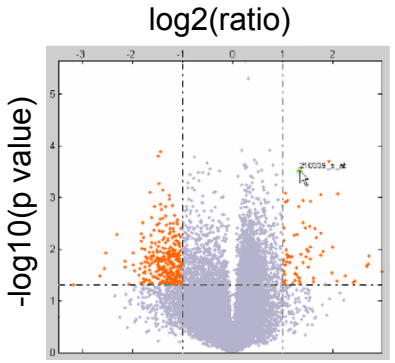
Gene lists and their interpretation

Omics studies generate lists of genes and proteins

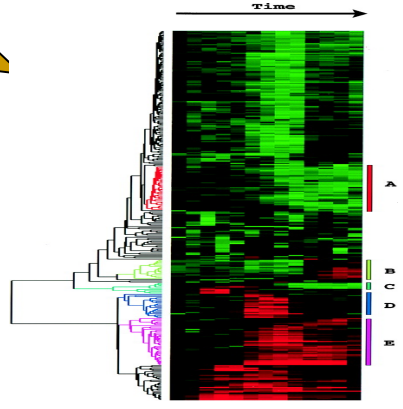
Microarray

RNA-Seq

Proteomics



Differential expression



Clustering

92546_r_at
92545_f_at
96055_at
102105_f_at
102700_at
161361_s_at
92202_g_at
103548_at
100947_at
101869_s_at
102727_at
160708_at
…...

Lists of genes with potential biological interest

Advantages of pathway analysis

Turning hundreds of protein names and IDs into meaningful information

- Compare own findings to "general knowledge"
- Extract relevant biological and functional themes from protein list
- Robust against noise to avoid misinterpretations
- Opportunity to reveal fine-grained and new details in a pathway

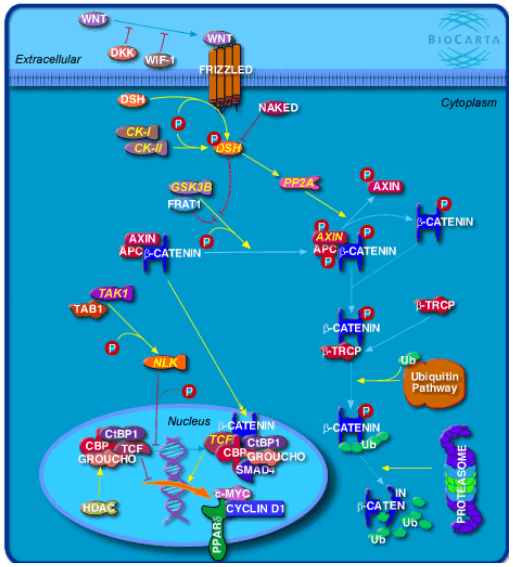
Remember

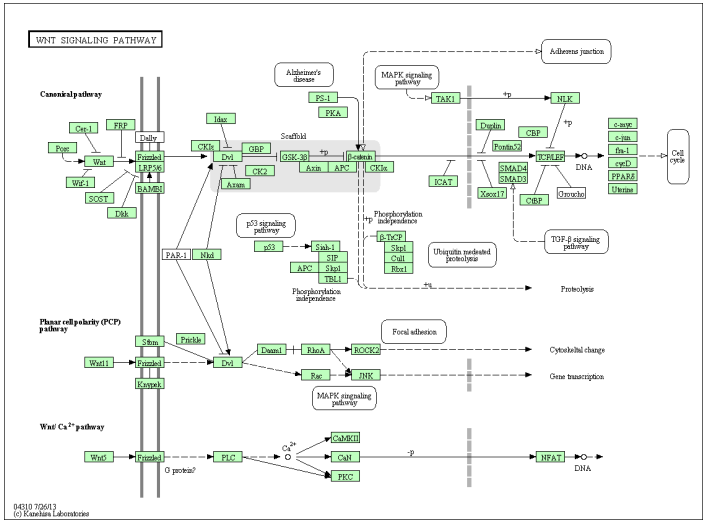
We apply false discovery rates (FDRs) of 1% or 5%

→ Expect partly wrong results

Pathway databases

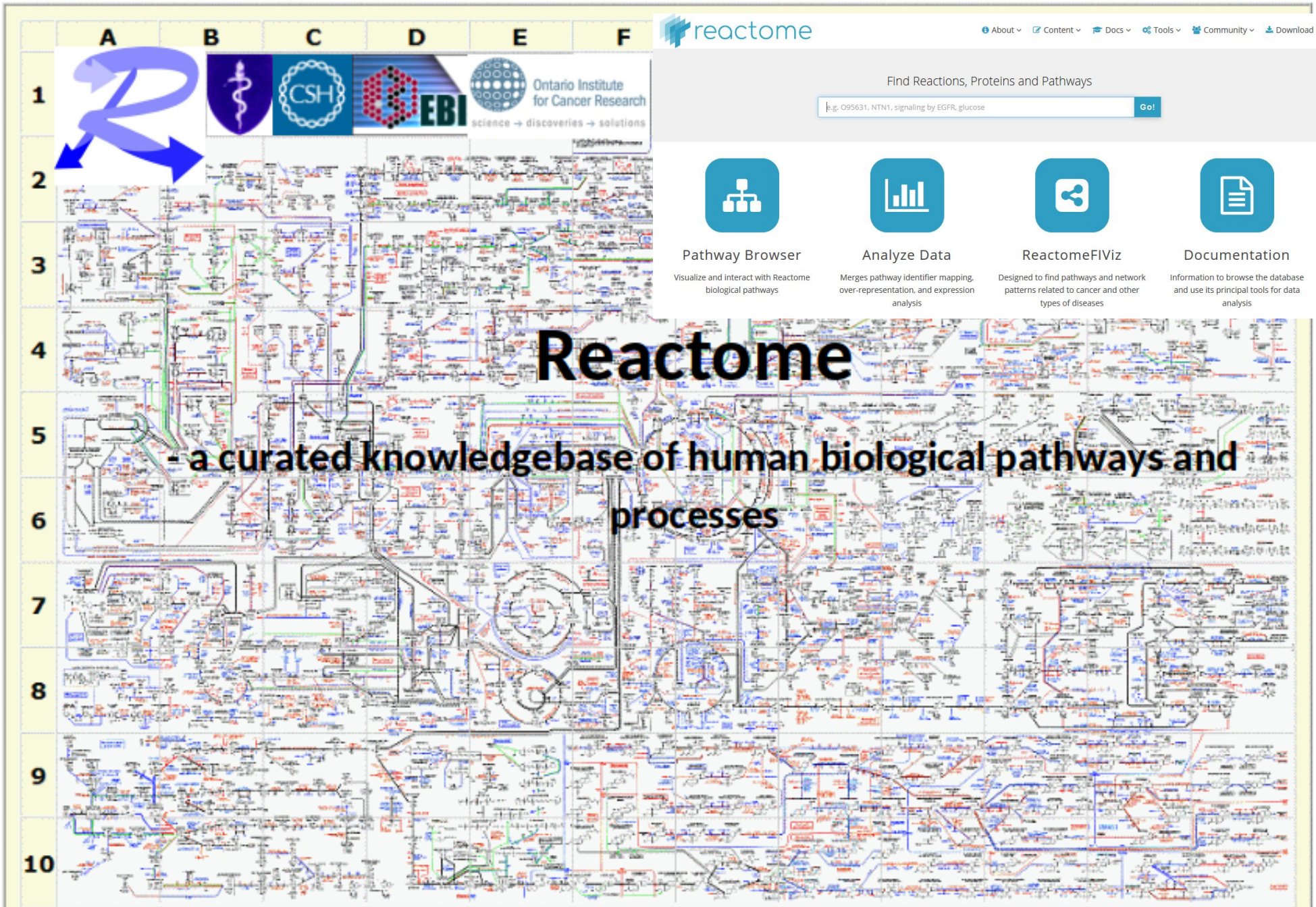
- Some major databases
  - BioCarta (http://www.biocarta.com/genes/index.asp)
  - KEGG (http://www.genome.jp/kegg/pathway.html)
  - MetaCyc (http://metacyc.org)
  - Pathway commons (http://www.pathwaycommons.org)
  - Reactome (http://www.reactome.org)
  - STKE (http://stke.sciencemag.org/cm)
  - Signaling Gateway (http://www.signaling-gateway.org)
  - Wikipathways (http://www.wikipathways.org)
- Limitation
  - Limited coverage
  - Inconsistency among different databases
  - Relationship between pathways is not defined





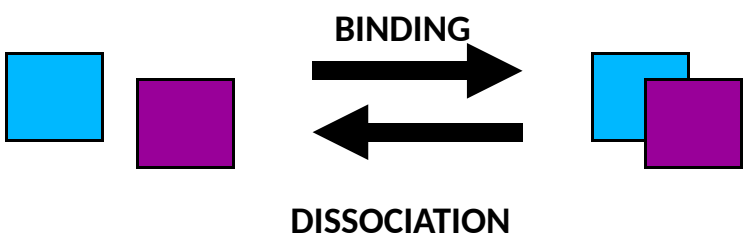
Reactome

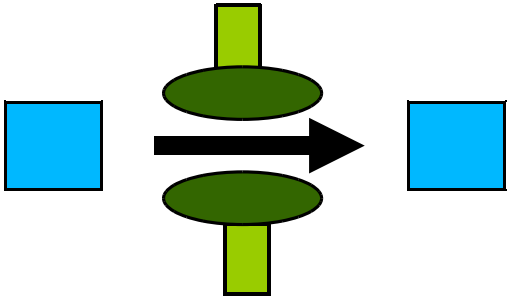
SDU

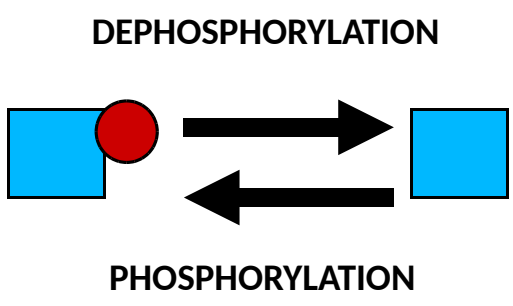


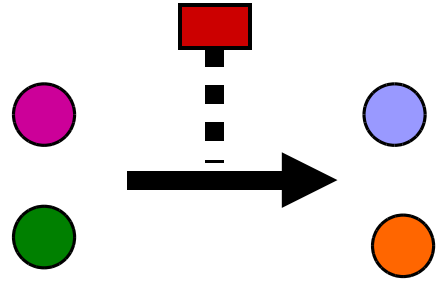
Reactions

- Basic “unit” of Reactome
- Represents many events and states found in biology
- Human- and machine-readible

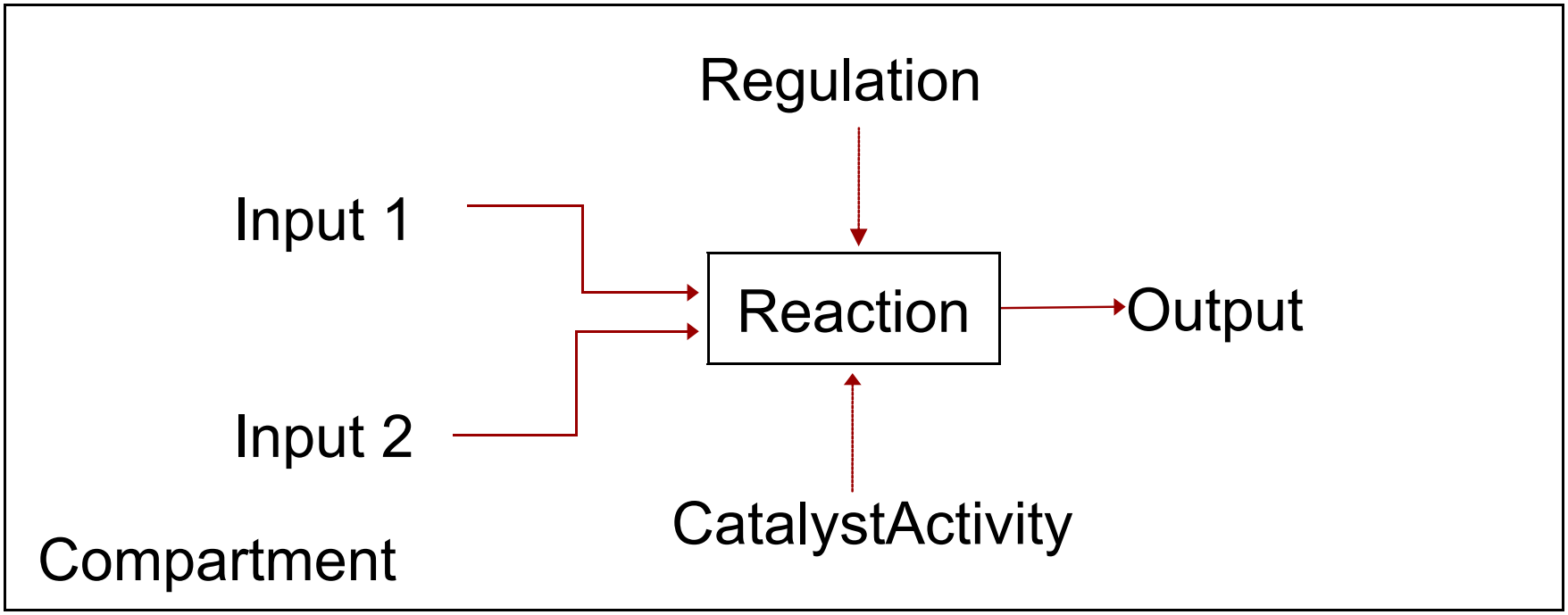


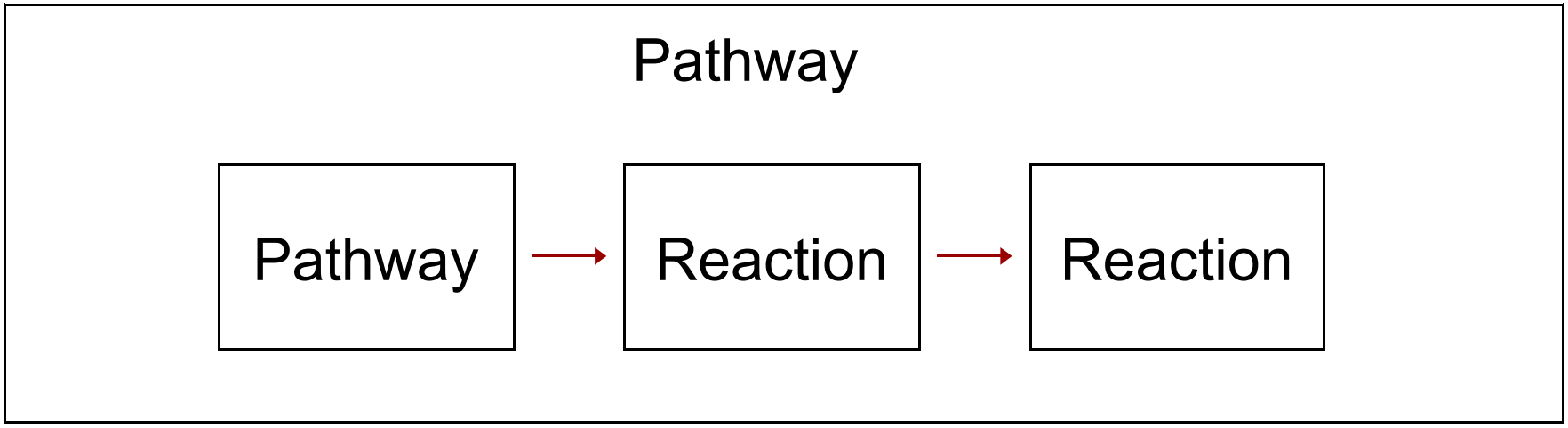






Reactions and pathways

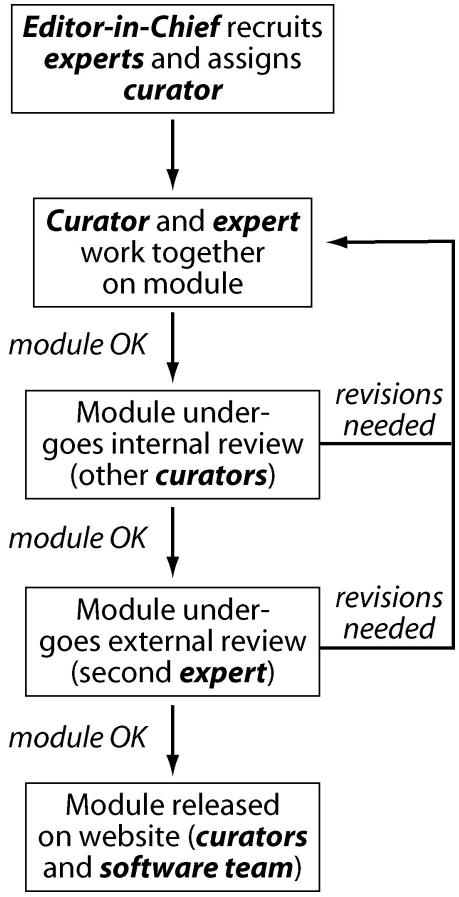




Curation sources

Where the Data Comes From

1) Field experts propose new modules.
2) Use a software tool to describe their pathway.
   - All molecules must be identified explicitly.
   - All assertions backed up by literature references.
3) Curators ensure consistency and completeness.
4) Peer review before publication.
5) Public Release of updated Reactome content every 3 months.
6) Rolling review every 2 years by expert.

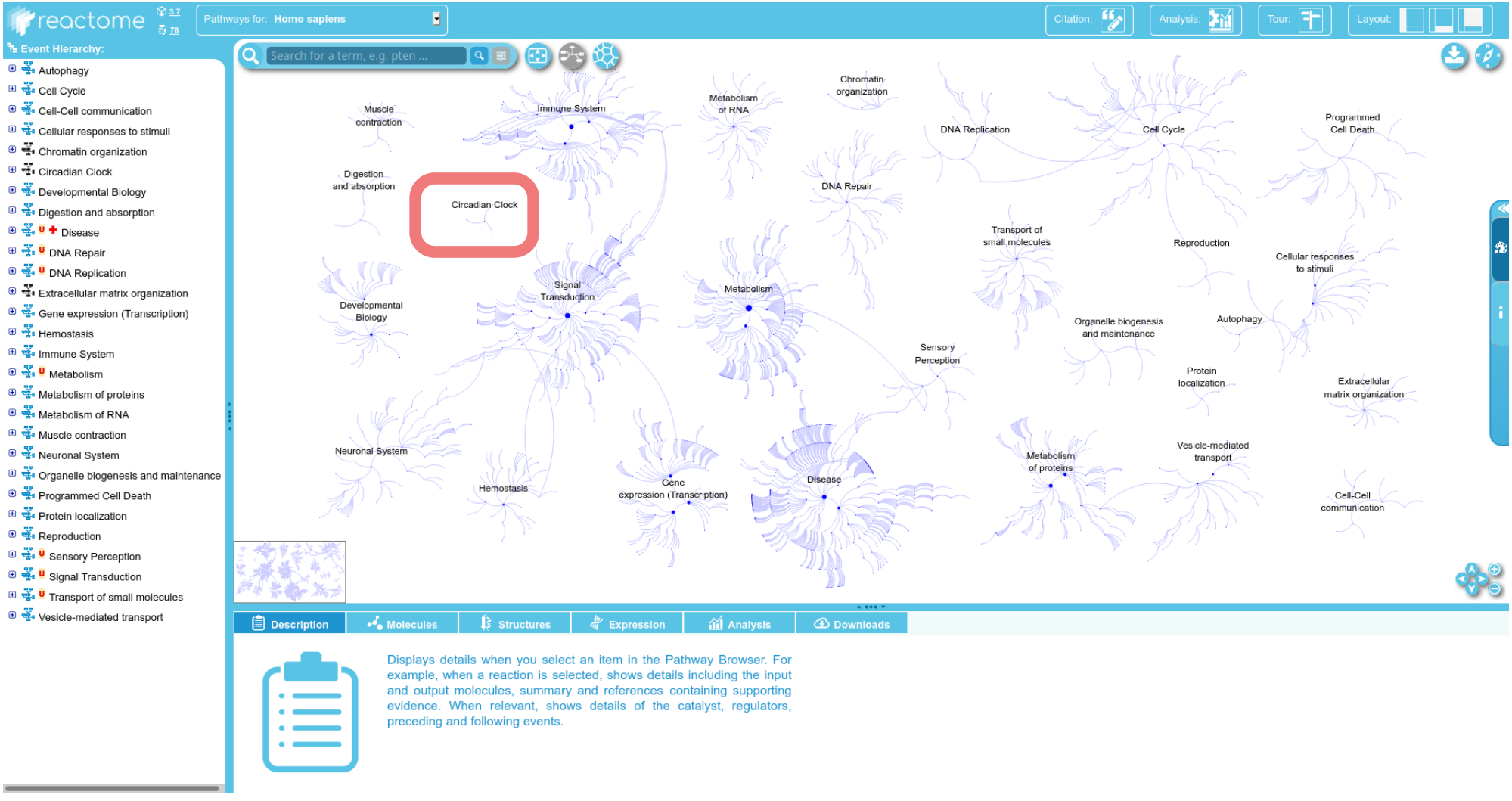


from original Reactome presentation

Exploring the Circadian Clock

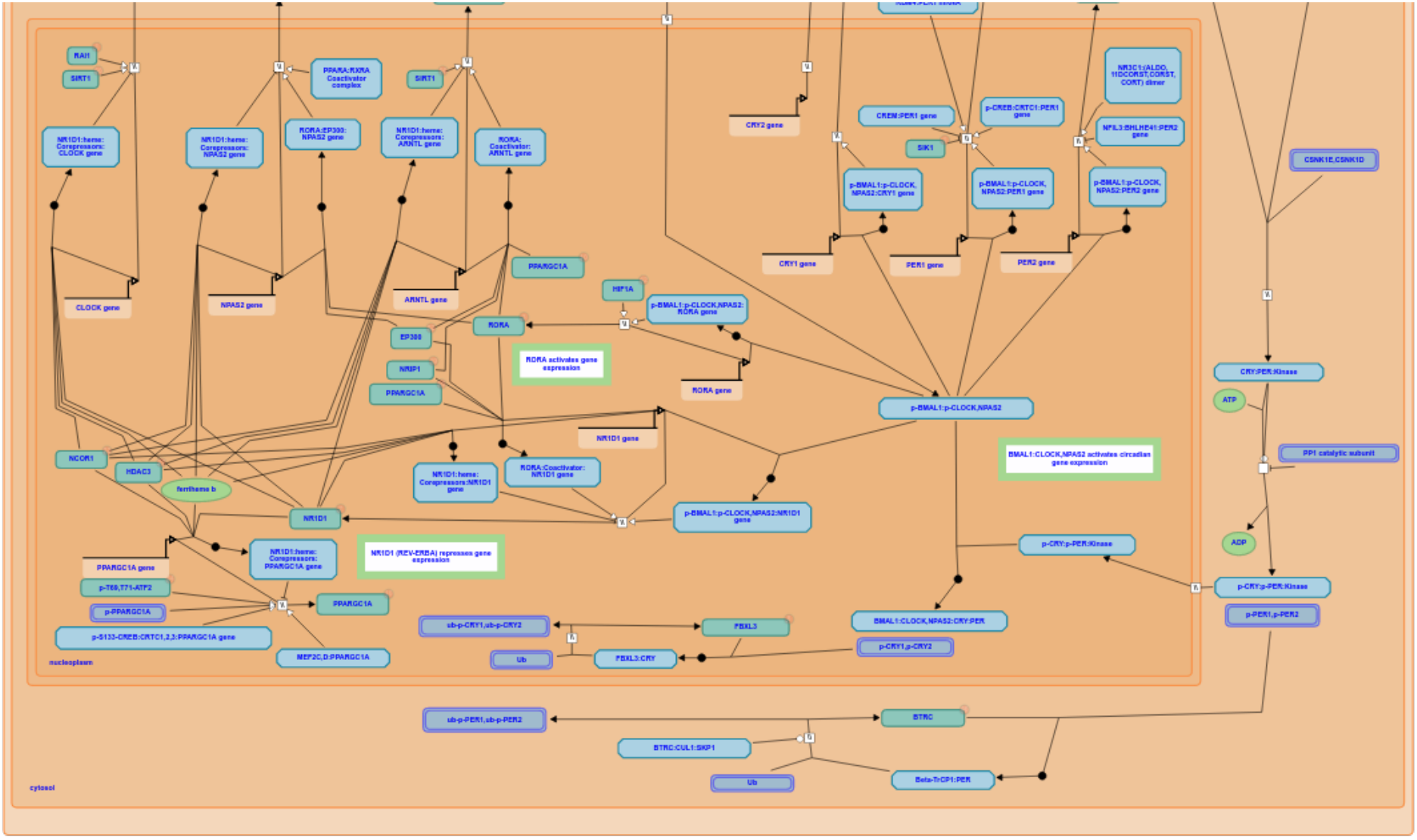
We want details about the 24h day-night cycle

Let's zoom in



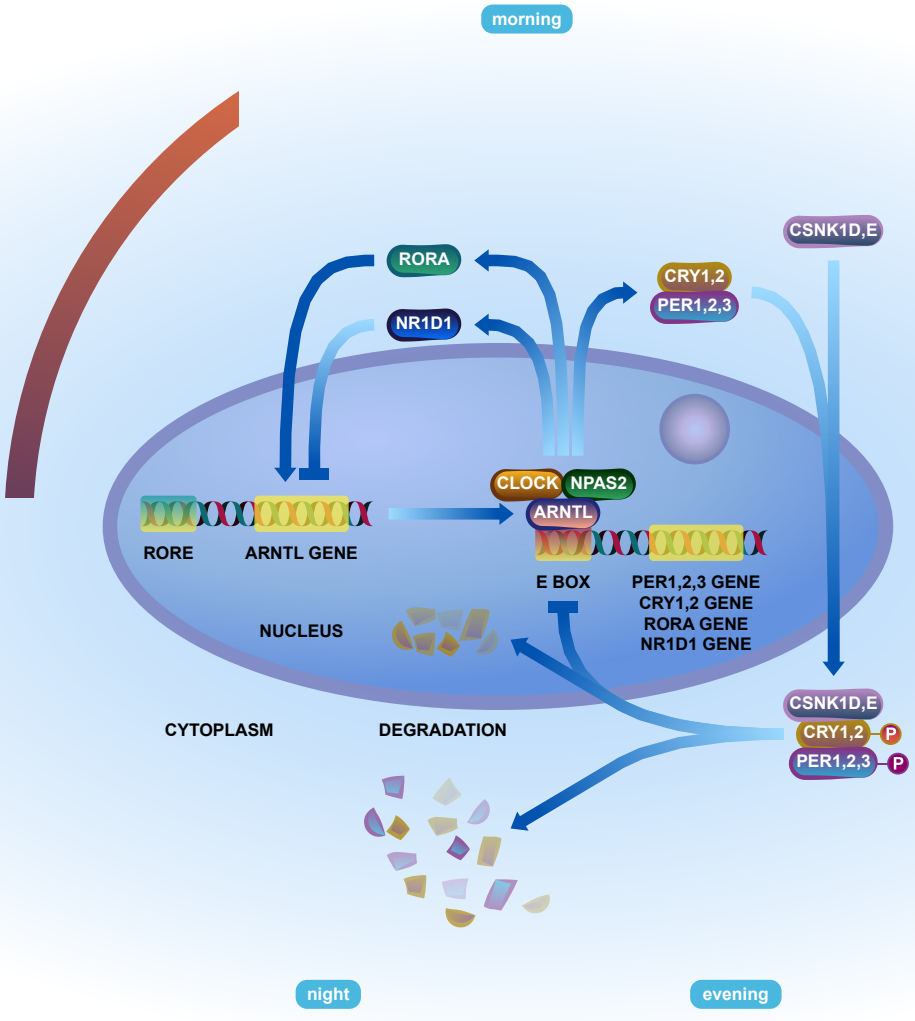
Circadian Clock: Details

Visualization as graph



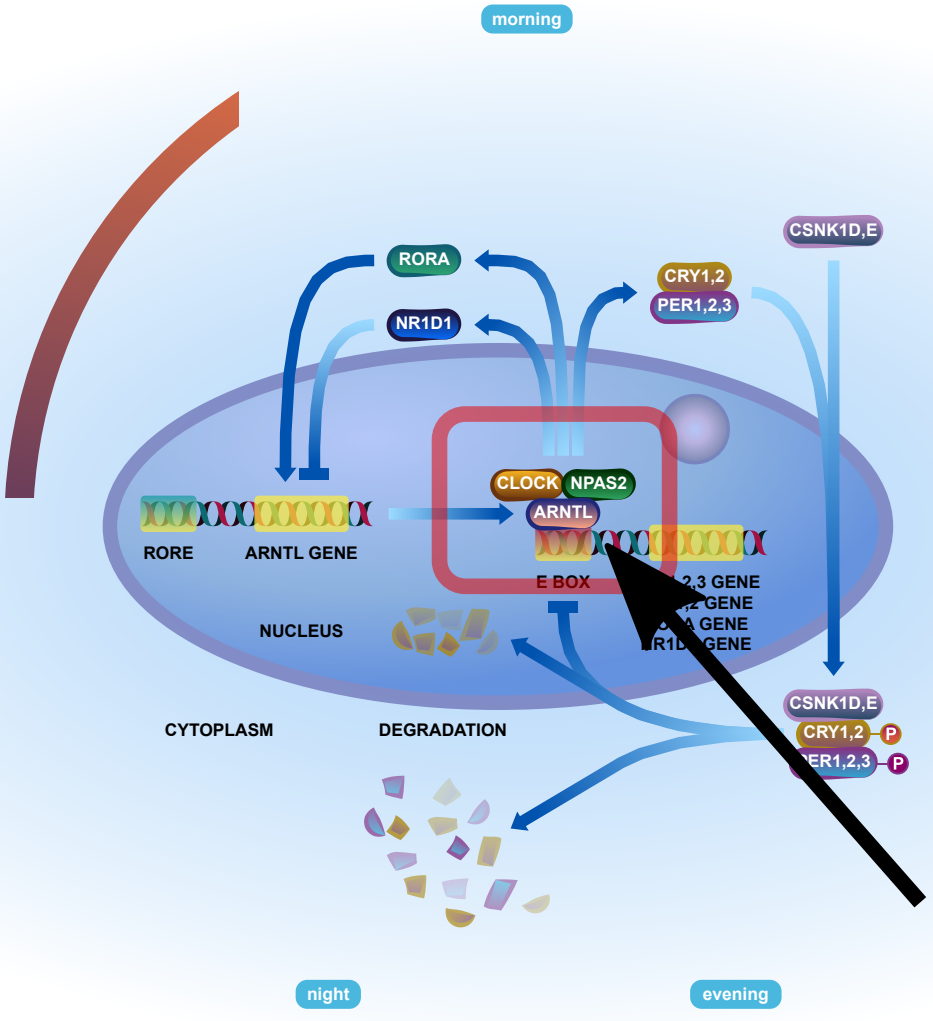
Circadian Clock: Details

As more comprehensive scheme



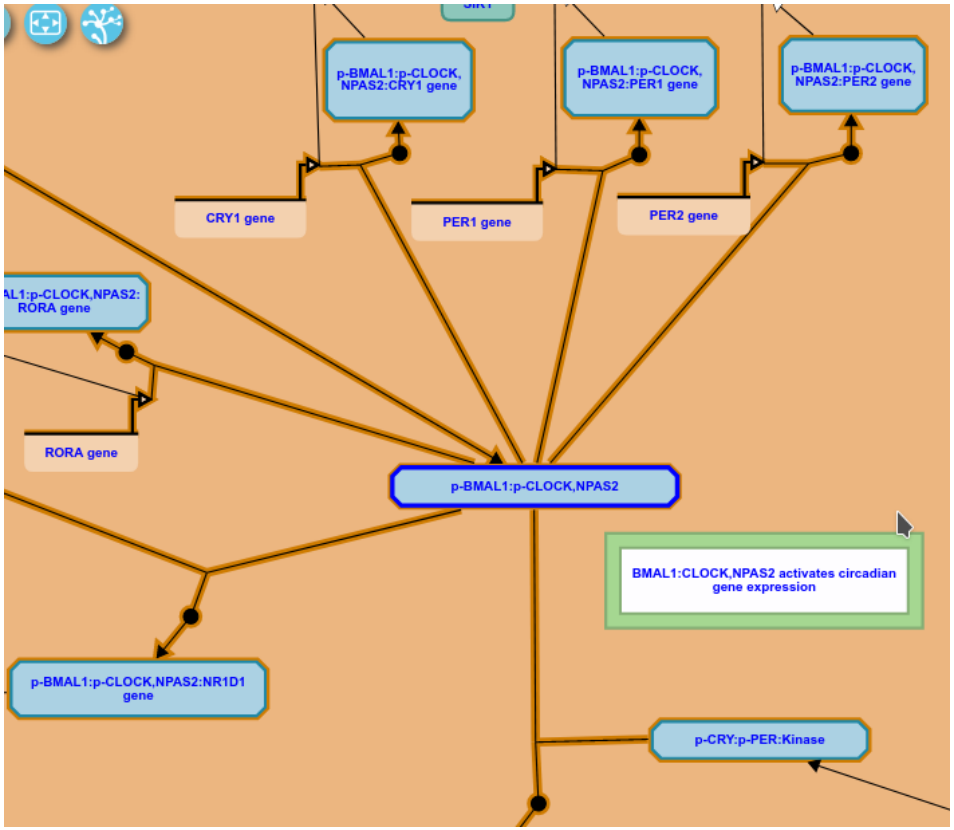
Circadian Clock: Details

As comprehensive visualization



What is the function of these 3 proteins?

The clock heterodimer



Extensive details with links to supporting literature

Description | Molecules | Structures 2/3 | Expression | Analysis | Downloads

p-BMAL1:p-CLOCK,NPAS2 [nucleoplasm] | Id: R-HSA-421315.1 | Species: Homo sapiens

Synonyms
p-ARNTL:p-CLOCK,NPAS2

Components
- p-S-ARNTL [nucleoplasm]
- p-CLOCK,p-NPAS2 [nucleoplasm]

Produced by
- BMAL1:CLOCK,NPAS2 heterodimer is phosphorylated and translocates to the nucleus [Homo sapiens]

Summation:
As inferred from mouse, BMAL1 (ARNTL), CLOCK, and NPAS2 are phosphorylated by unknown kinases. The phosphorylation is dependent on the heterodimerization of BMAL1 with CLOCK or NPAS2. Phosphorylated BMAL1:CLOCK/NPAS2 is a much stronger transactivator of gene expression than is unphosphorylated BMAL1:CLOCK/NPAS2.

Consumed by
- p-BMAL1:p-CLOCK,NPAS2 binds F7 gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds AVP gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds SERPINE1 gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds PPARA gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds BHLHE41 (DEC2) gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds CCRN4L (NOCTURNIN) gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds PER1 gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds RORA gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds NR1D1 gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds DBP gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds NAMPT gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds PER2 gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds CRY1 gene [Homo sapiens]
- CRY:PER heterodimer binds the BMAL1:CLOCK/NPAS2 heterodimer [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds BHLHE40 gene [Homo sapiens]
- p-BMAL1:p-CLOCK,NPAS2 binds KLF15 gene [Homo sapiens]

Deduced from the existence of
- p-Bmal1:p-Clock,Npas2 [nucleoplasm]

References
- Cryptochromes impair phosphorylation of transcriptional activators in the clock: a general mechanism for circadian repression
- CK2alpha phosphorylates BMAL1 to regulate the mammalian clock

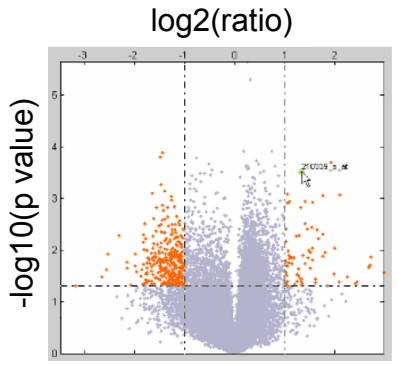
Gene lists and their interpretation

Omics studies generate lists of genes and proteins

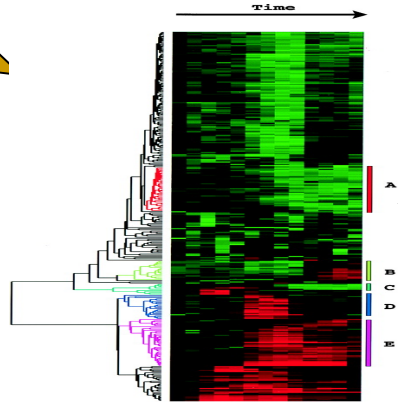
Microarray

RNA-Seq

Proteomics



Differential expression



Clustering

92546_r_at
92545_f_at
96055_at
102105_f_at
102700_at
161361_s_at
92202_g_at
103548_at
100947_at
101869_s_at
102727_at
160708_at
……

Lists of genes with potential biological interest

? How to find the associated pathways? ?

Complex relationships

How to assign the relevant pathways to the gene list?

- Gene 1
  - Pathway 1
  - Pathway 2
- Gene 2
  - Pathway 1
  - Pathway 3
- Gene 3
  - Pathway 2
  - Pathway 3
- Gene 4
  - Pathway 3

- Pathway 1
  - Gene 1
  - Gene 2
- Pathway 2
  - Gene 1
  - Gene 3
- Pathway 3
  - Gene 2
  - Gene 3
  - Gene 4

Note:
Protein pathways and properties are assigned to genes names!

Overrepresentation analysis

Example for overpresentation in this room:
Do we have a significantly larger proportion of people from Fyn?

Denmark:

Total population: ~5.2M

Population in Fyn: ~0.5M

Contingency table

	From Fyn	Not from Fyn
Denmark	0.5M	4.7M
You	?	?

Method:

Fisher's exact test: Statistical test and p-values

Output:

How likely is it to get these numbers by chance when there is no larger proportion?

Link to more detailed description

Overrepresentation analysis

Overrepresentation of a biological function (here pathway):
Is there an overly high proportion of genes associated with this pathway?

Human cells

Total population: 10720 annotated genes/proteins

Immune system: 1191 proteins

Contingency table

	Immune syst.	Not immune syst.
Human	10.72k	1,191
Differentially regulated treatment vs. control	?	?

Method:

Fisher's exact test: Statistical test and p-values

Output:

How likely is it to get these numbers by chance when the treatment does not affect this pathway?

Link to more detailed description

Example: Psoriaris

Comparing affected vs. healthy skin

Proteomics study reveals

128 differentially regulated proteins

Psoriaris disease:
Chronic skin condition

➔ Run Overrepresentation analysis in Reactome

The following table shows the 25 most relevant pathways sorted by p-value.

Pathway name	Entities				Reactions	
	found	ratio	p-value	FDR*	found	ratio
Neutrophil degranulation	20 / 480	0.043	8.27e-07	3.09e-04	7 / 10	7.37e-04
Metal sequestration by antimicrobial proteins	4 / 6	5.34e-04	9.02e-07	3.09e-04	4 / 5	3.68e-04
Innate Immune System	30 / 1,191	0.106	3.79e-05	0.009	62 / 710	0.052
Attenuation phase	3 / 14	0.001	6.29e-04	0.098	4 / 5	3.68e-04
Platelet degranulation	7 / 128	0.011	8.13e-04	0.098	3 / 11	8.10e-04
Antimicrobial peptides	6 / 95	0.008	9.28e-04	0.098	8 / 58	0.004
Response to elevated platelet cytosolic Ca2+	7 / 133	0.012	0.001	0.098	3 / 14	0.001

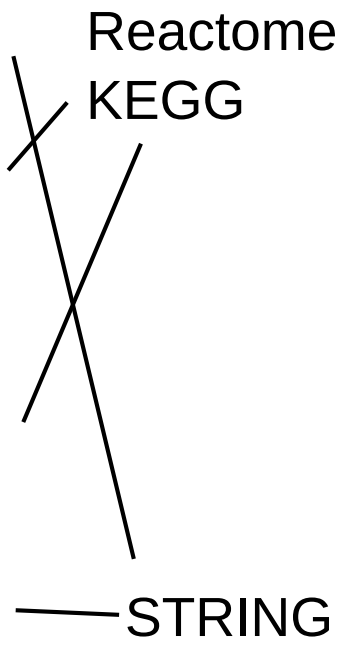
False discovery rate: How much do we expect this pathway when having a randomly generated list of proteins?

← Neutrophils play a crucial role in this disease

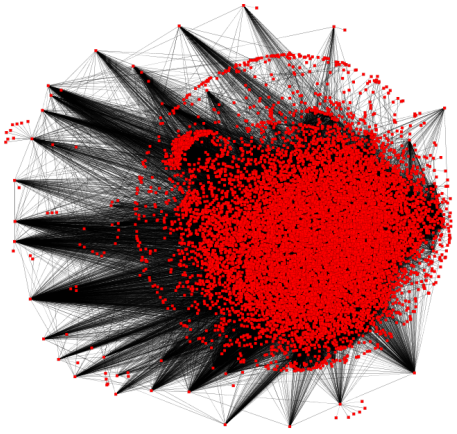
Data from paper: A proteomics approach to the identification of biomarkers for psoriasis utilising keratome biopsy

Graphs and biological interactions

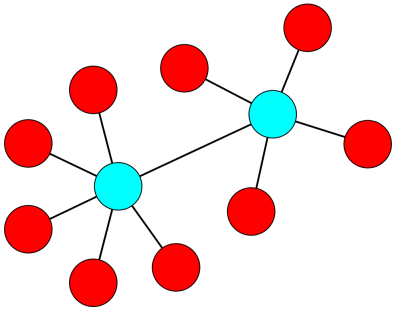
Networks		Nodes	Edges
Physical interaction networks	Protein-protein interaction network	Proteins	Physical interaction, undirected
	Signaling network	Proteins	Modification, directed
	Gene regulatory network	TFs/miRNAs Target genes	Physical interaction, directed
	Metabolic network	Metabolites	Metabolic reaction, directed
Functional association networks	Co-expression network	Genes/ proteins	Co-expression, undirected
	Genetic network	Genes	Genetic interaction, undirected



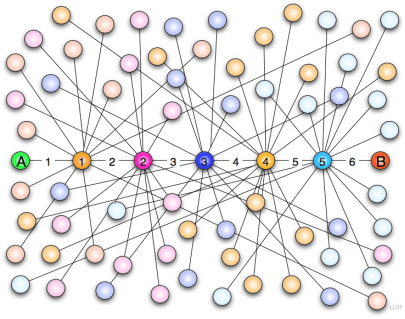
Networks/graphs, categories



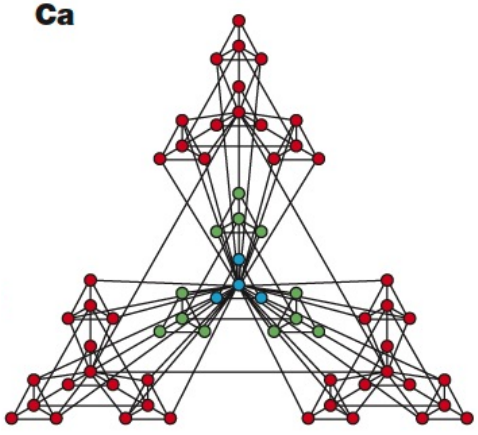
Human protein-protein interaction network
9,198 proteins and 36,707 interactions



Scale-free
(hubs)



Small world
(six degree separation)

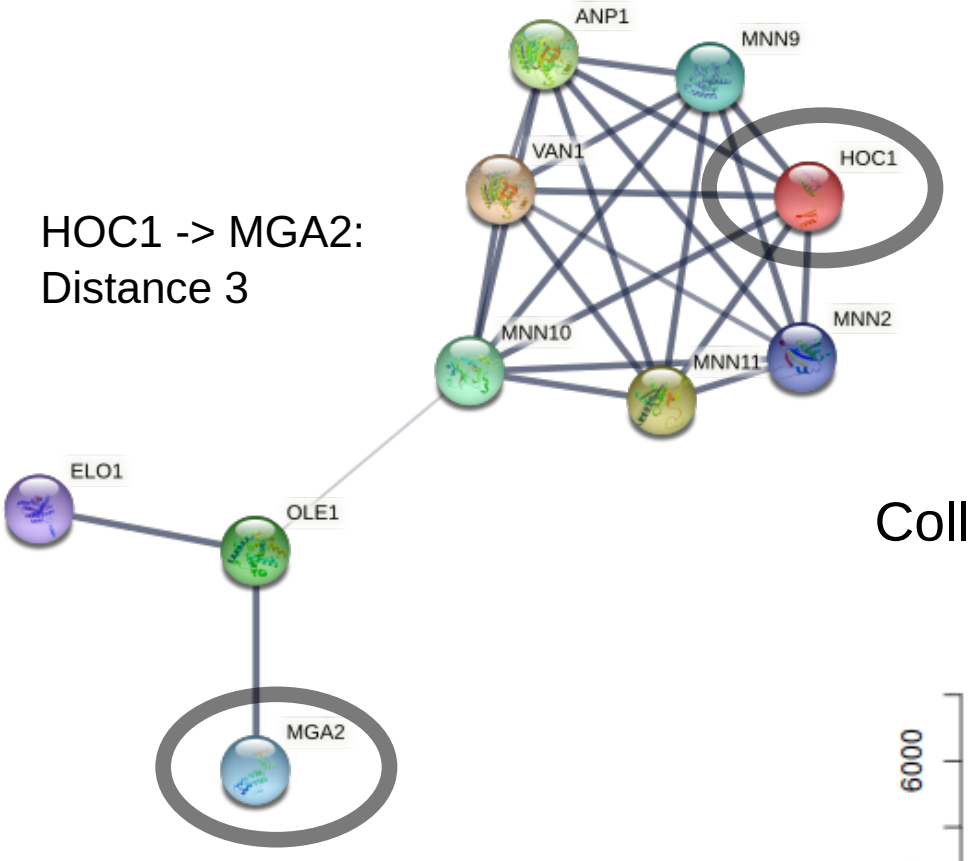


Hierarchical modular

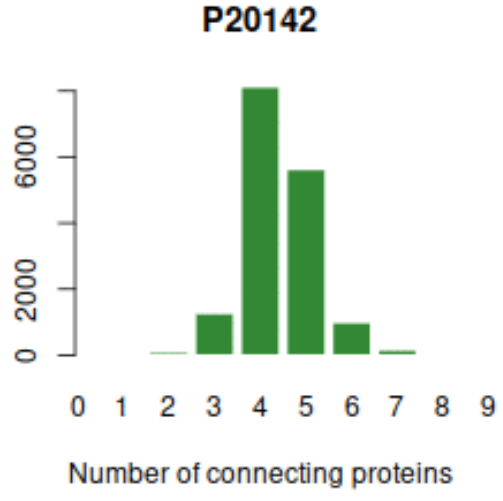
How much small world?

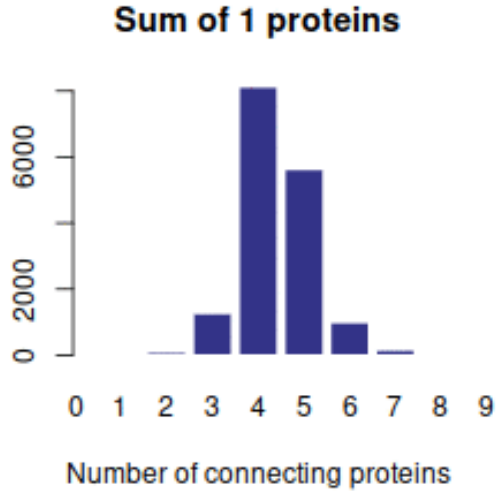
Take minimal number of genes/proteins
that connect two genes/proteins

HOC1 -> MGA2:
Distance 3



Collecting some numbers (STRING network):





The STRING database

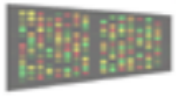
STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases.

Data Sources

Interactions in STRING are derived from five main sources:

Genomic Context Predictions

High-throughput Lab Experiments



(Conserved) Co-Expression

Automated Textmining

Previous Knowledge in Databases

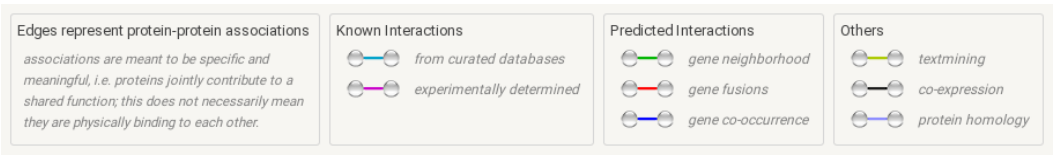
Coverage

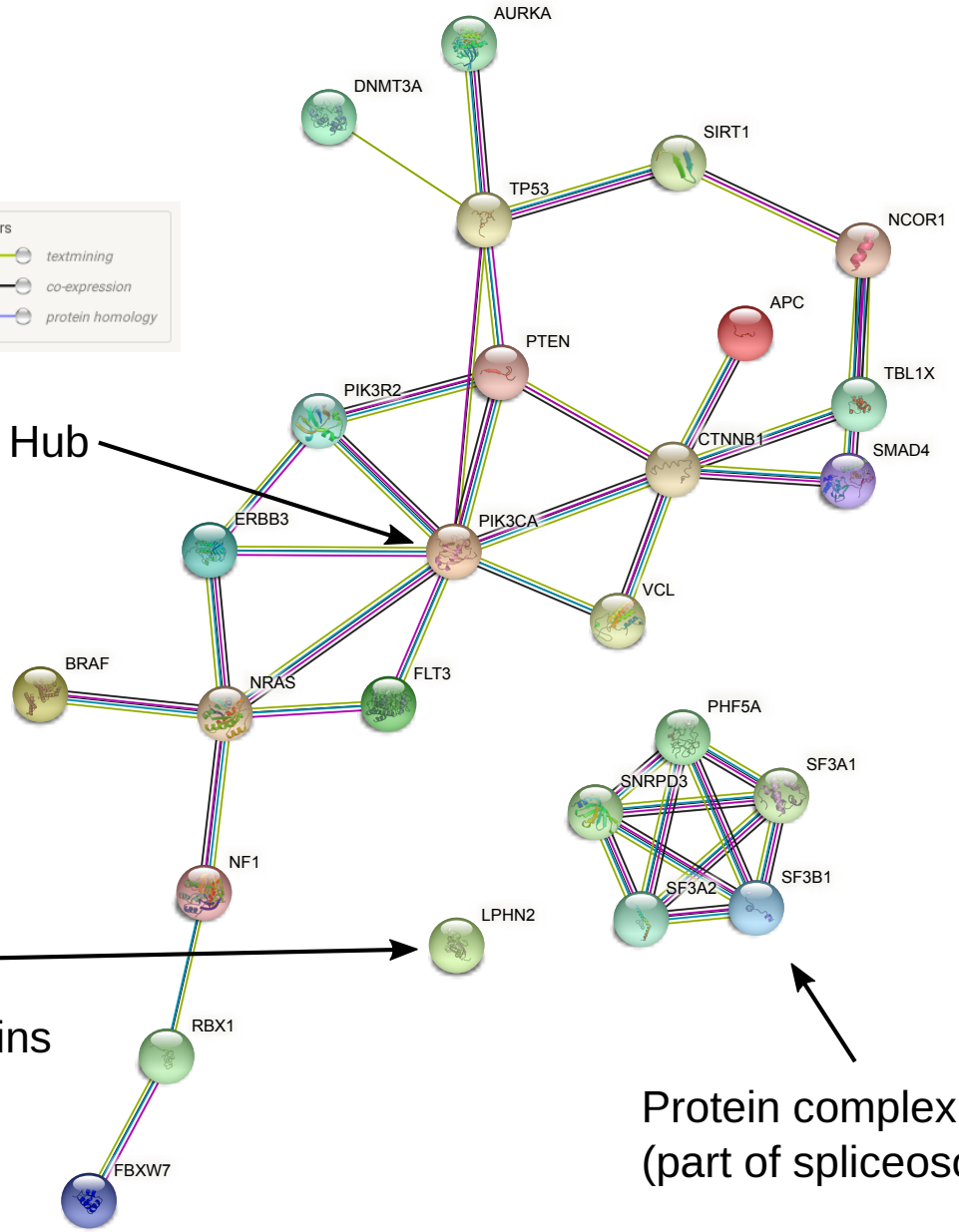
The STRING database currently covers 24'584'628 proteins from 5'090 organisms.

Relies on much more (automatically retrieved) information, no manual curation

STRING: evidences

20 most mutated genes in cancer
and their "neighborhood"





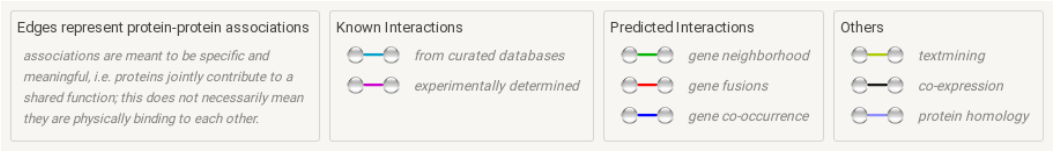
Hub

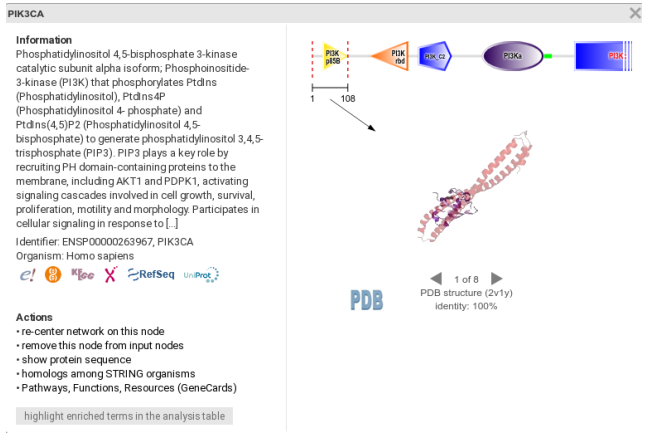
Poorly annoted
or
not related to the other proteins

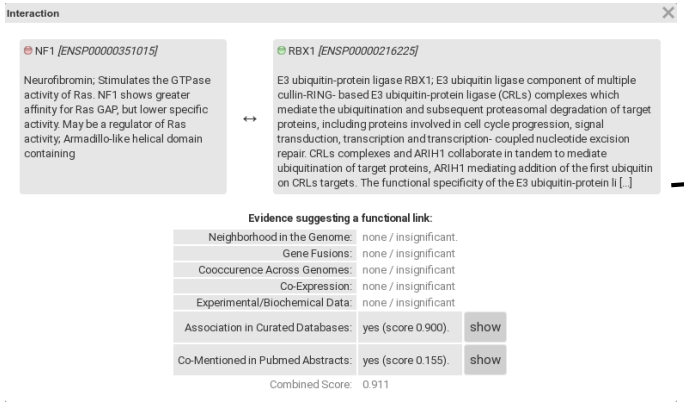
Protein complex
(part of spliceosome)

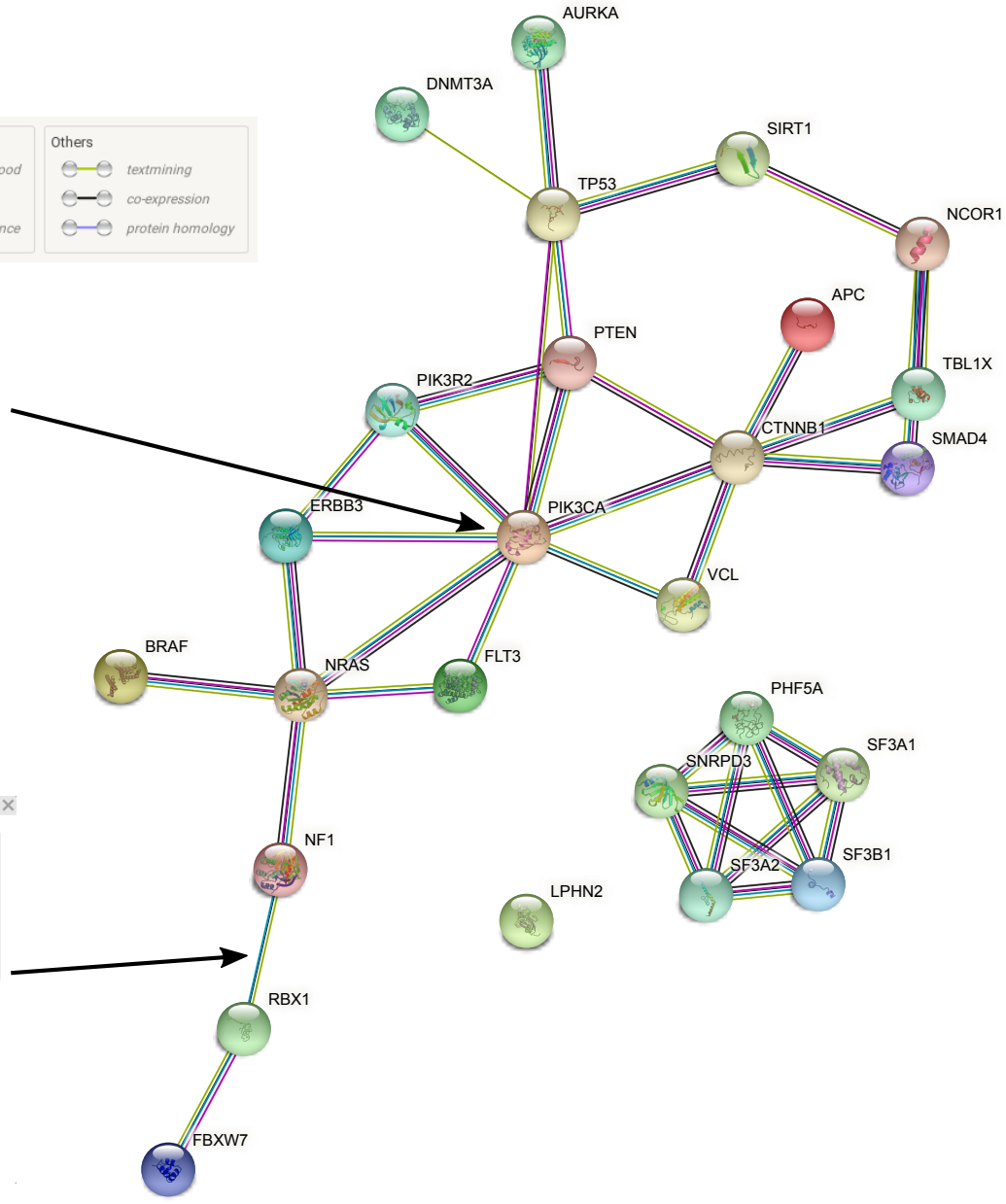
STRING: evidences details

20 most mutated genes in cancer and their "neighborhood"





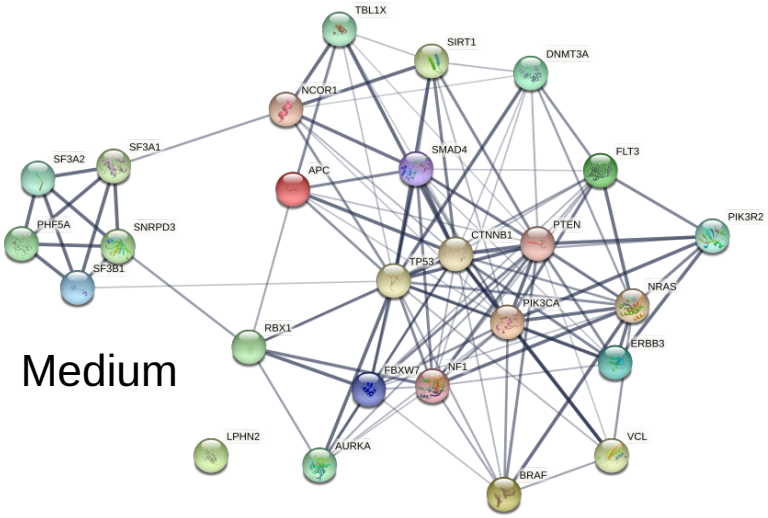




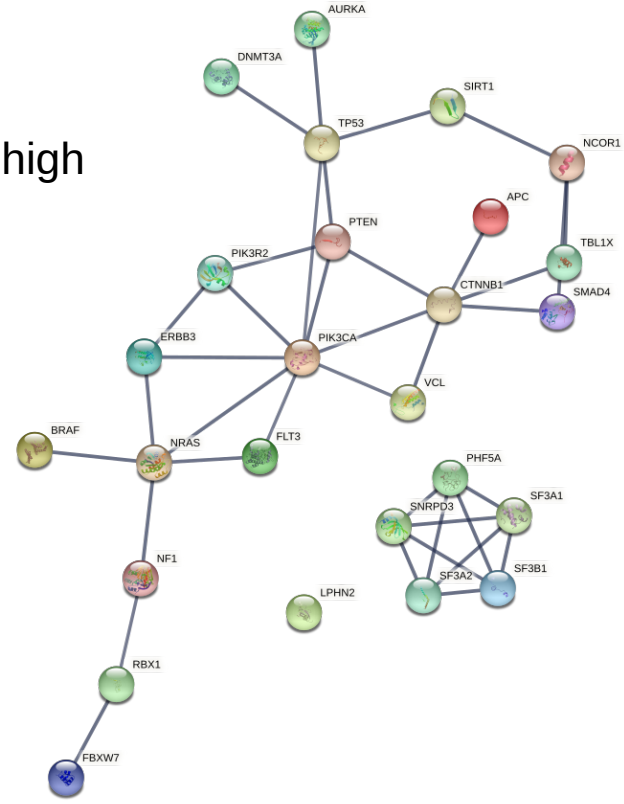
STRING: confidence levels

20 most mutated genes in cancer and their "neighborhood"

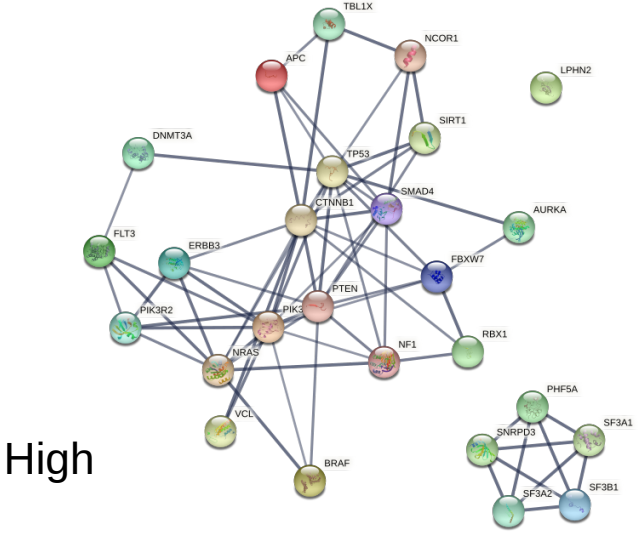
Medium



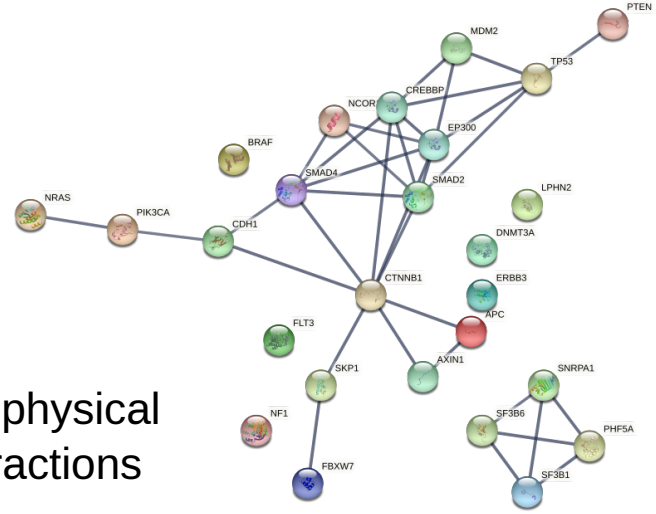
Very high



High



Only physical interactions



Network analysis: further possibilities

Software

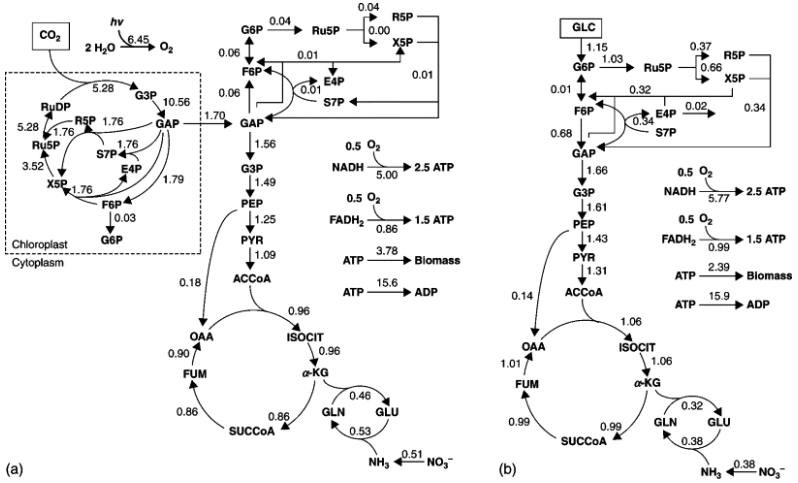
Cytoscape

Powerful software for network analysis with many bioinformatics plugins

Flux analysis

Popular in metabolomics and signalling

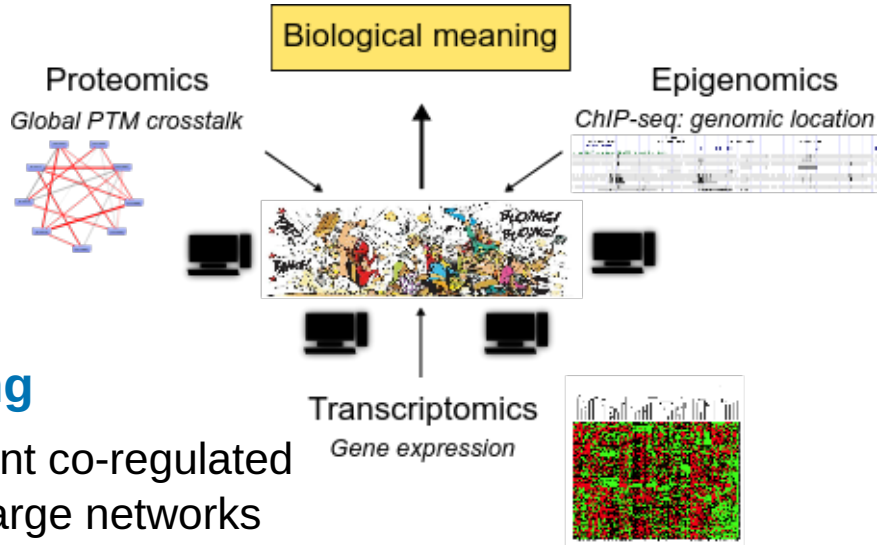
How does information propagate the cell?



Data integration

Combine data from different platforms, e.g. transcriptomics with DNA methylations

How much do your findings show in other data types?



Clustering

Find relevant co-regulated groups in large networks