

Bioinformatics I

Quantitative Proteomics:
Biological Interpretation:

Veit Schwämmle

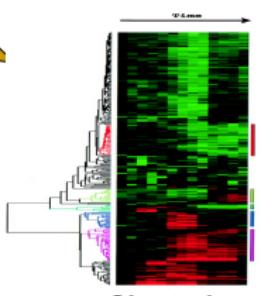
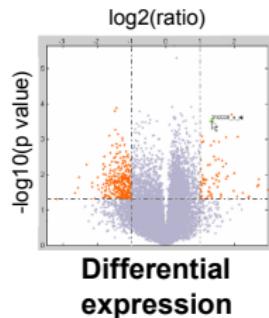
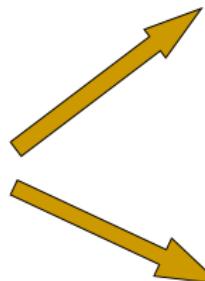
Protein Research Group
BMB
Syddansk Universitet

Omics studies generate lists of genes and proteins

Microarray

RNA-Seq

Proteomics



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

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

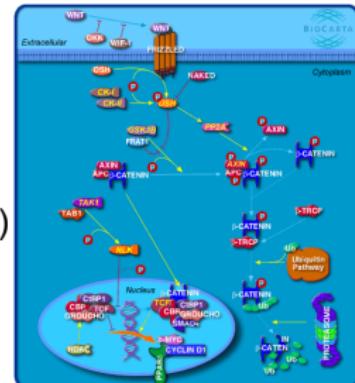
Remember

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

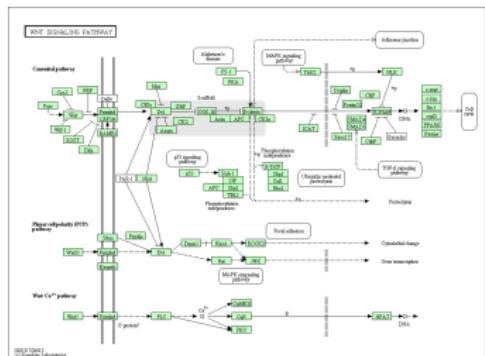
→ Expect partly wrong results

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



The image is a collage of various Reactome interface screenshots and logos, arranged in a grid-like structure. At the top left, there's a vertical column of 10 numbered boxes (1-10) containing pathway maps. A large blue arrow points from box 1 towards the center. Above the boxes are logos for A (Reactome logo), B (CSHL), C (CSHL), D (EBI), E (Ontario Institute for Cancer Research), and F (Reactome logo). To the right of the boxes is the main Reactome search bar with the placeholder "Find Reactions, Proteins and Pathways" and a search button. Below the search bar are four blue icons: "Pathway Browser" (hierarchy), "Analyze Data" (bar chart), "ReactomeFIViz" (two overlapping circles), and "Documentation" (document icon). Further down, there's a large central pathway map with the text "Reactome - a curated knowledgebase of human biological pathways and processes".

1

2

3

4

5

6

7

8

9

10

A B C D E F

reactome

Find Reactions, Proteins and Pathways

e.g. C00001, NTF1, signaling by EGFR, glucose

Pathway Browser

Analyze Data

ReactomeFIViz

Documentation

Visualize and interact with Reactome biological pathways

Merge pathway identifier mapping, over-representation, and expression analysis

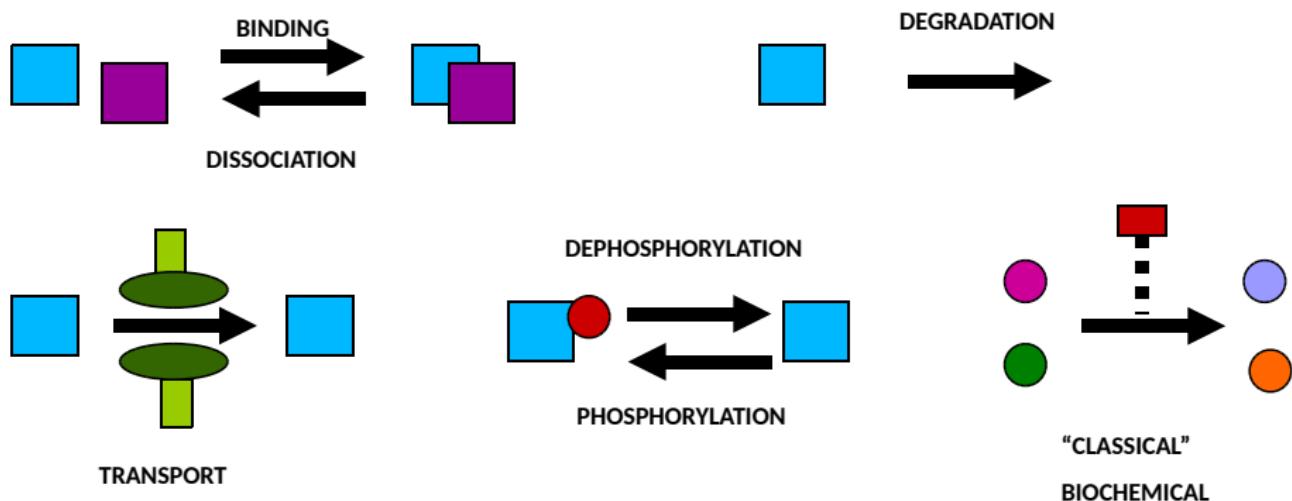
Designed to find pathways and network patterns related to cancer and other types of diseases

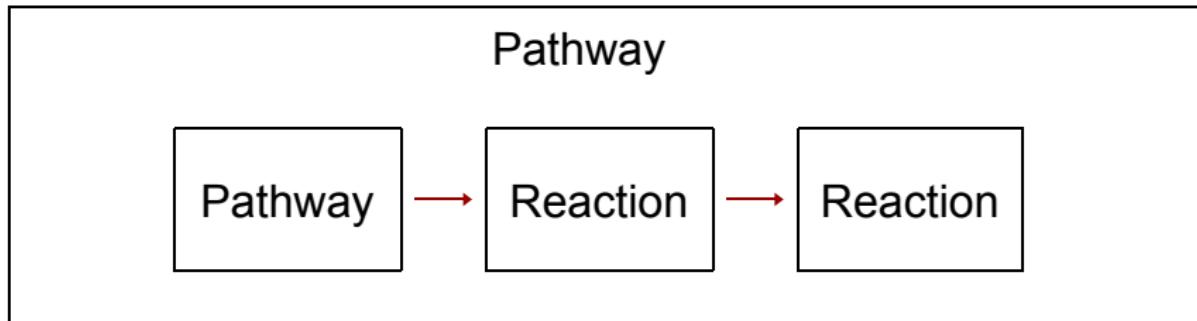
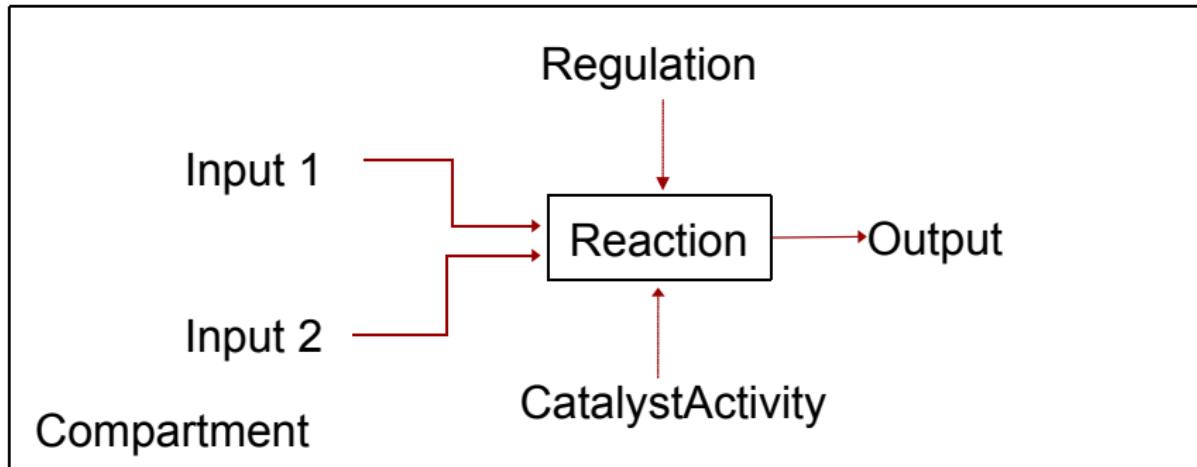
Information to browse the database and use its principal tools for data analysis

Reactome - a curated knowledgebase of human biological pathways and processes

Reactions

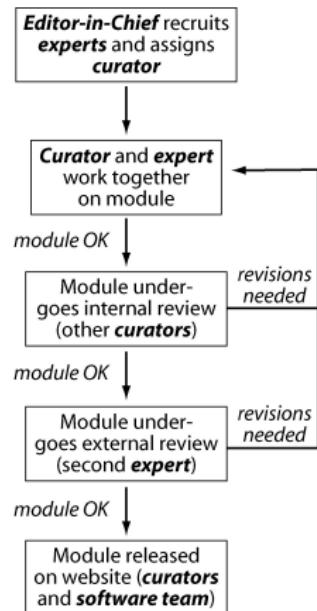
- Basic “unit” of Reactome
- Represents many events and states found in biology
- Human- and machine-readable





Where the Data Comes From

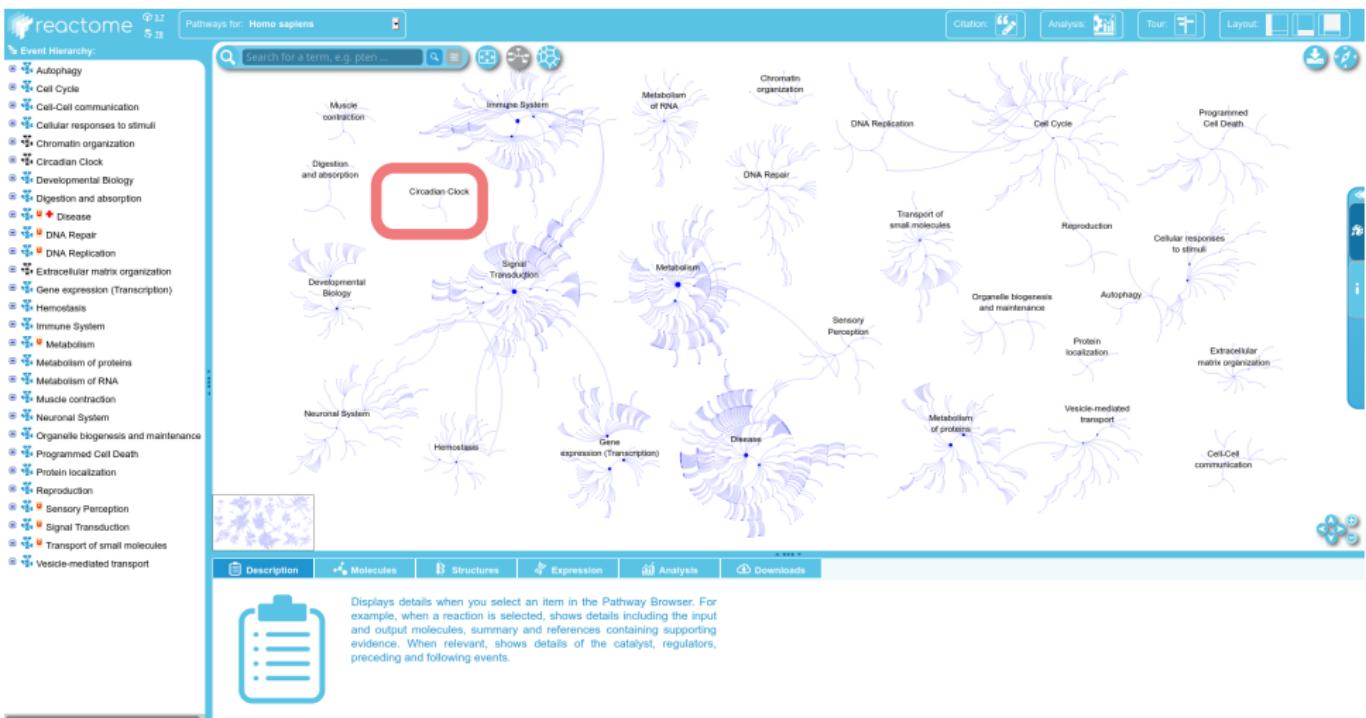
- 1) Field experts propose new modules.
- 2) Use a software tool to describe their pathway.
 - All molecules must be identified explicitly.
 - All assertions backed up by literature references.
- 3) Curators ensure consistency and completeness.
- 4) Peer review before publication.
- 5) Public Release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



Exploring the Circadian Clock

We want details about the 24h day-night cycle

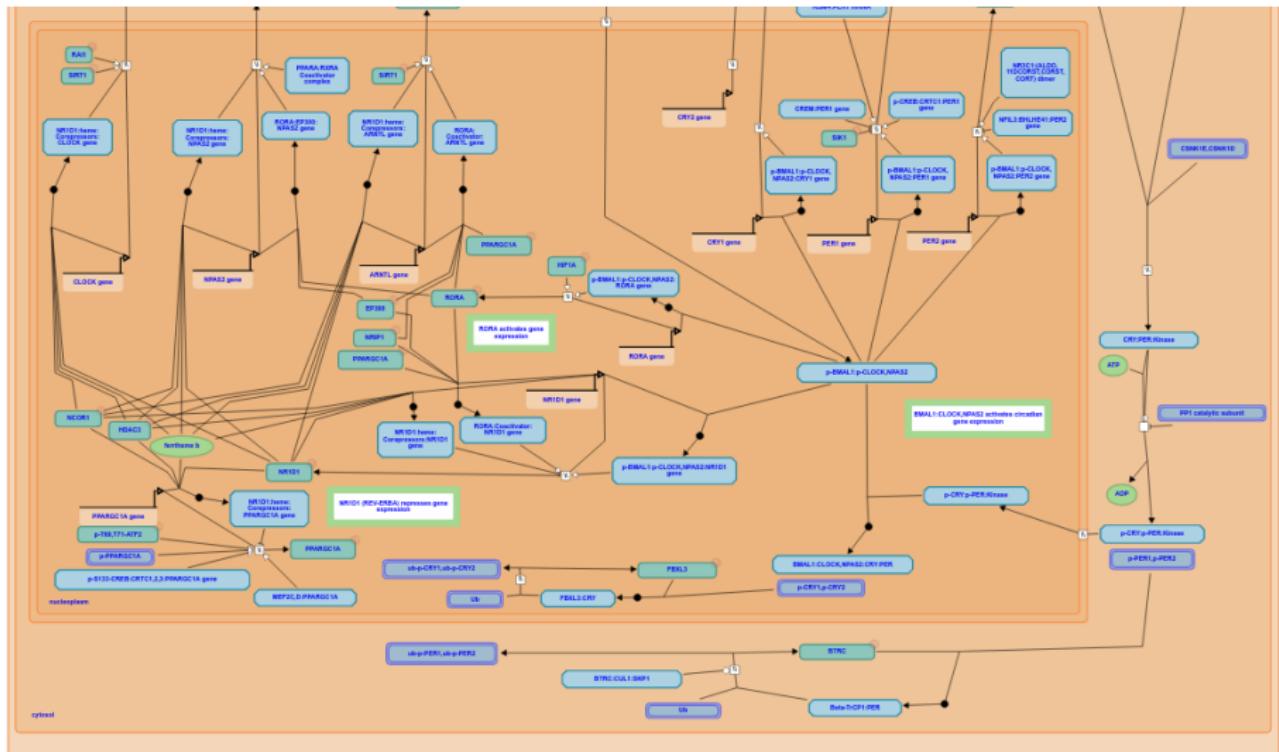
Let's zoom in



Circadian Clock: Details

SDU 

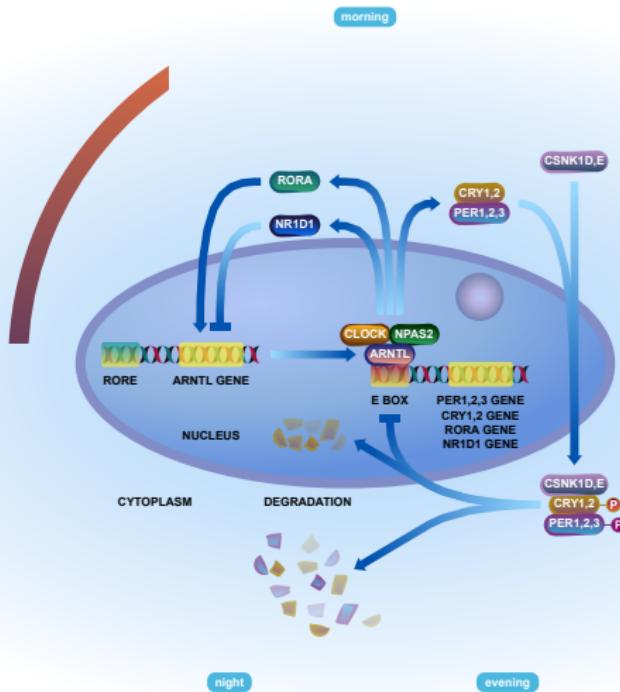
Visualization as graph



Circadian Clock: Details

As more comprehensive scheme

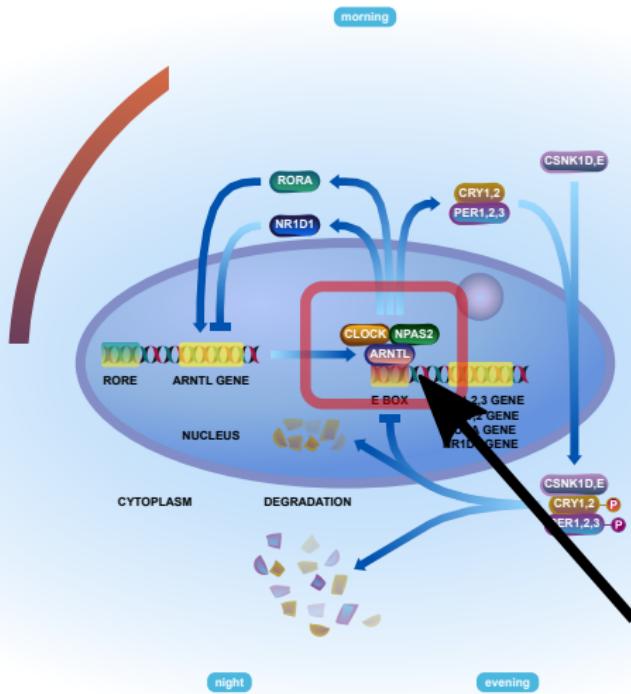
 reactome



Circadian Clock: Details

As comprehensive visualization

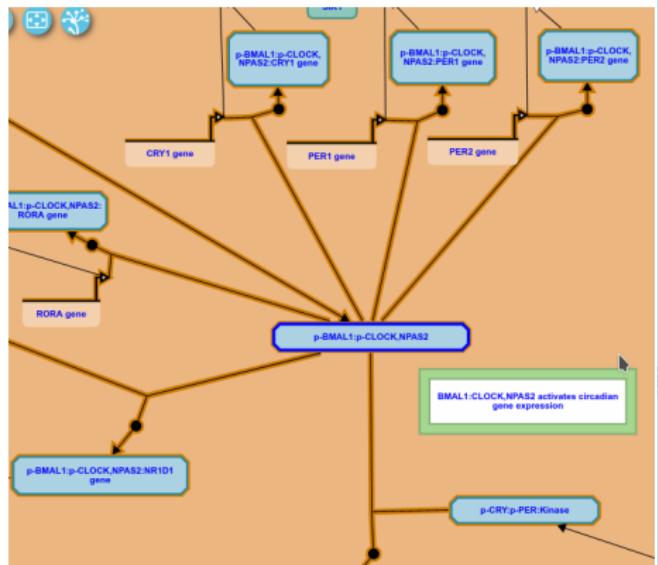
 reactome



What is the function of these 3 proteins?

The clock heterodimer

SDU



Extensive details with links to supporting literature

Description Molecules Structures 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]

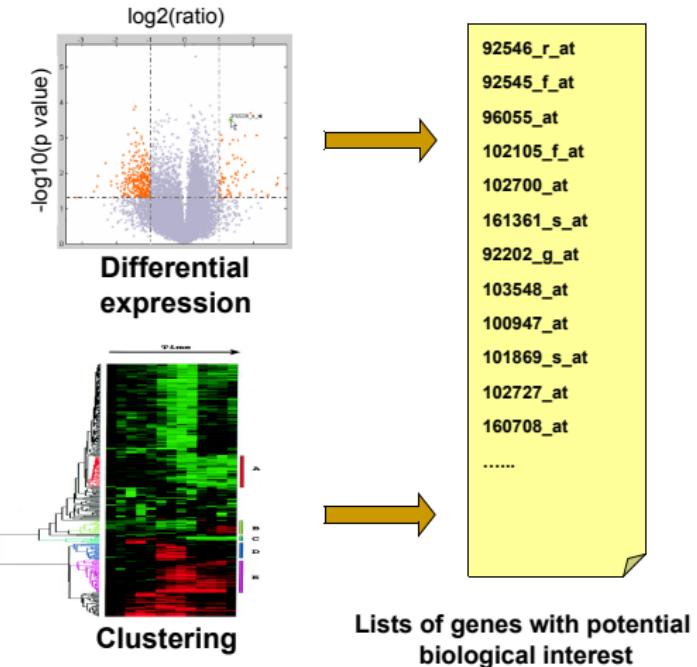
References
Cryptochromes impair phosphorylation of transcriptional activators in the clock: a general mechanism for circadian repression
CK2alpha phosphorylates BMAL1 to regulate the mammalian clock

Omics studies generate lists of genes and proteins

Microarray

RNA-Seq

Proteomics



? How to find the associated pathways? ?

How to assign the relevant pathways to the gene list?

- Gene 1
 - Pathway 1
 - Pathway 2
- Gene 2
 - Pathway 1
 - Pathway 2
 - 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

Contingency table

Immune system: 1191 proteins

Differentially regulated
treatment vs. control

	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 Ca ²⁺	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

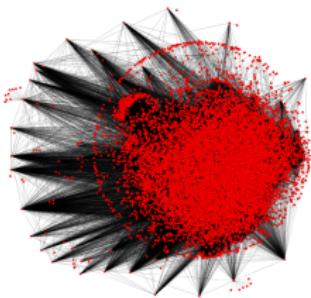
Graphs and biological interactions

	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

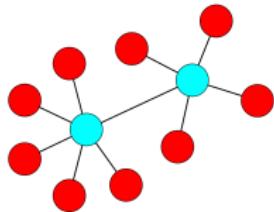
Reactome
KEGG



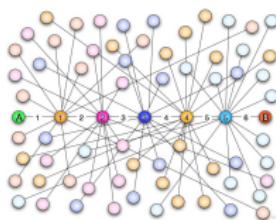
STRING



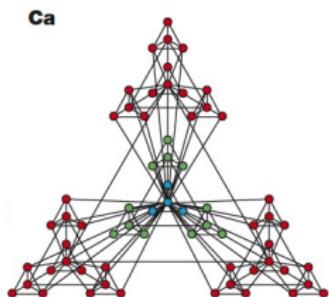
Human protein-protein interaction network
9,198 proteins and 36,707 interactions



Scale-free
(hubs)



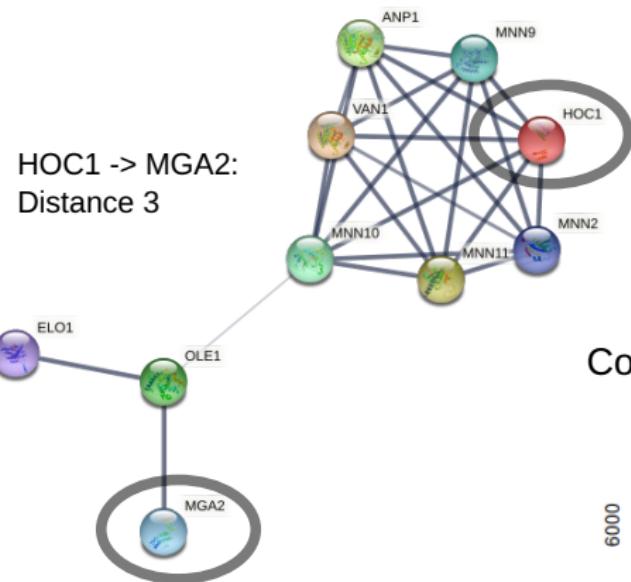
Small world
(six degree separation)



Hierarchical modular

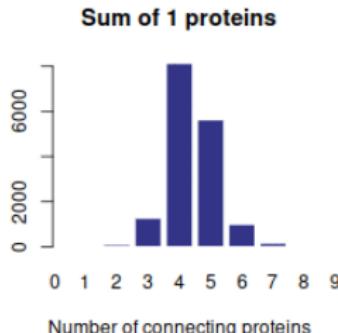
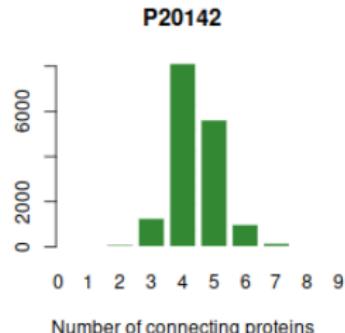
How much small world?

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



HOC1 -> MGA2:
Distance 3

Collecting some numbers (STRING network):



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

SDU

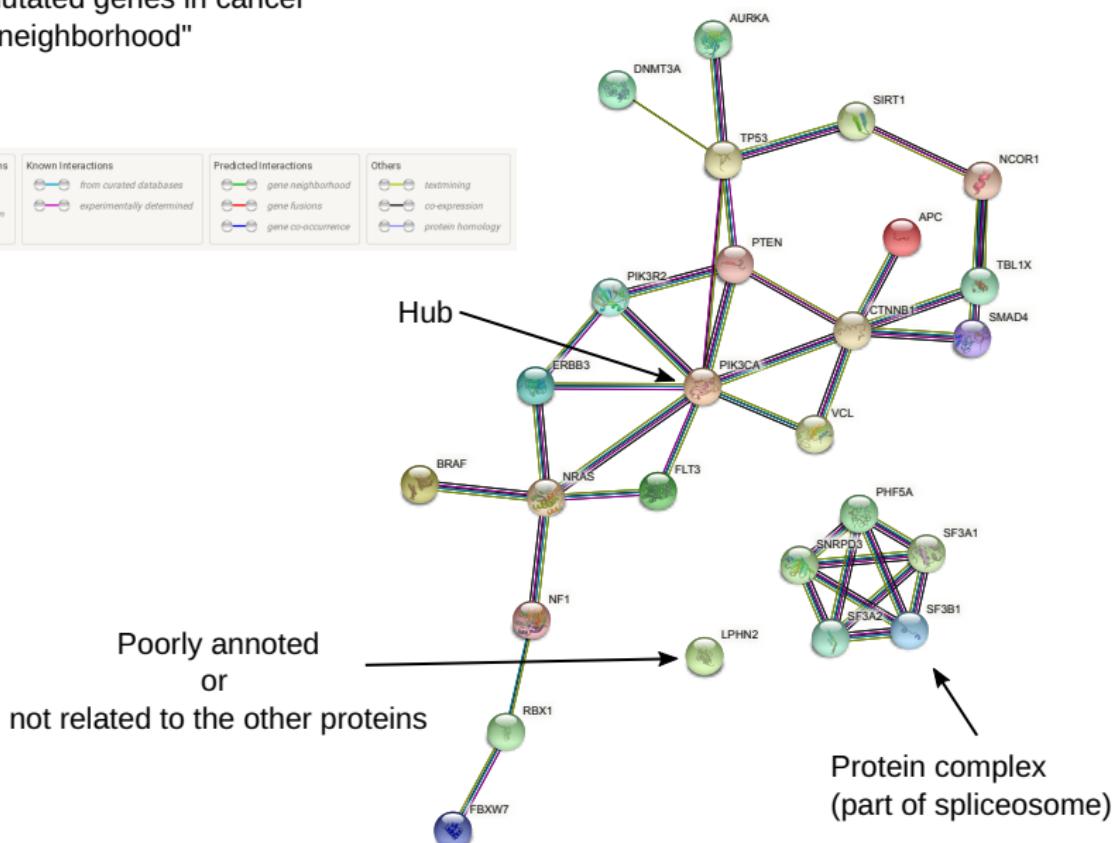
20 most mutated genes in cancer
and their "neighborhood"

Edges represent protein-protein associations
associations are meant to be specific and
meaningful, i.e. proteins jointly contribute to a
shared function; this does not necessarily mean
they are physically binding to each other.

Known Interactions
from curated databases
experimentally determined

Predicted Interactions
gene neighborhood
gene fusions
gene co-occurrence

Others
textmining
co-expression
protein homology



STRING: evidences details

SDU

20 most mutated genes in cancer and their "neighborhood"

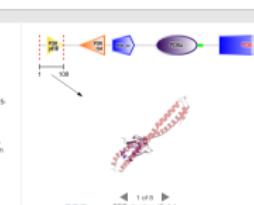
Edges represent protein-protein associations
associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding to each other.

Known Interactions
 from curated databases
 experimentally determined

Predicted Interactions
 gene neighborhood
 gene fusions
 gene co-expression
 protein homology

Others
 textmining
 co-expression
 protein homology

PIK3CA
Information:
 Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform; Phosphatidylinositol-4-phosphate (PI4K) that phosphorylates PI(4,5)P₂ to PI(3,4,5)P₃; Phosphatidylinositol 4,5-bisphosphate (PIP₂) and PI(4,5)P₂ (Phosphatidylinositol 3,4,5-trisphosphate (PIP₃)). PIP₃ plays a key role by recruiting PDZ-domain-containing proteins to the membrane, thus regulating various downstream signaling cascades involved in cell growth, survival, proliferation, motility and morphology. Participates in cellular responses to stress, PRKA, PI3K-Akt, Hoxo, Hoxo complex, PRKC, PI3K.



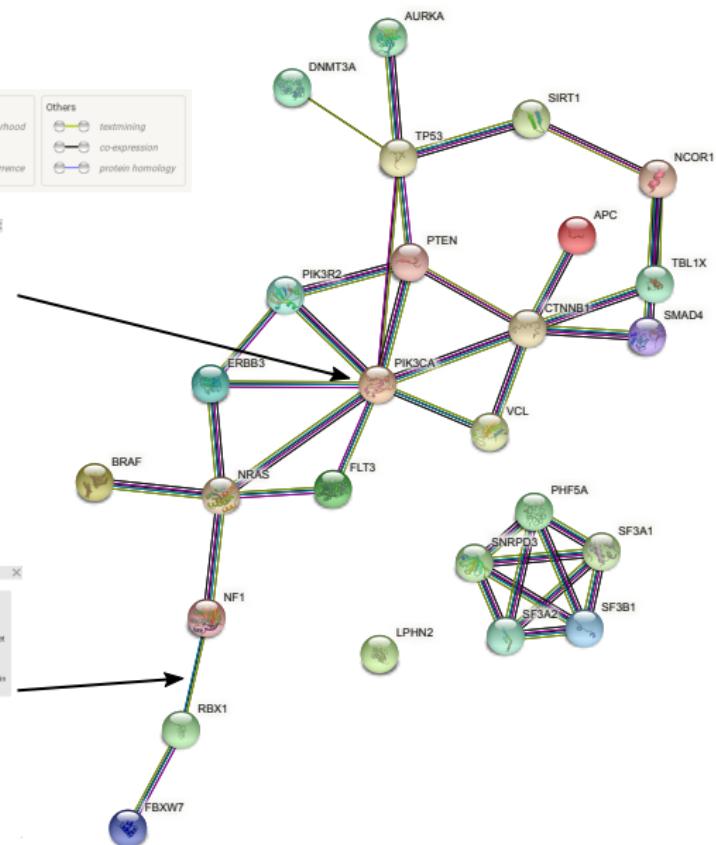
Action:
 - no center node set on this node
 - several nodes make up the input nodes
 - same protein neighbors
 - homologs among STRING organisms
 Pathways, Functions, Resources
 GeneCards

Highlighted enriched terms in the analysis results

Interaction:
 # NF1 [BMSNP00000351015]
 Description: Stimulates the GTPase activity of Ras. NF1 shows greater affinity for Ras GAP, but lower specific activity. Maybe a regulator of Ras activity. Annular-like helical domain containing

RBX1 [BMSNP00000276229]
 E3 ubiquitin-protein ligase RBX1. It disrupts target components of multiple multi-subunit E3 ubiquitin-protein ligase (S3X3) complexes which mediate the ubiquitination and subsequent proteasomal degradation of target proteins, including proteins involved in cell cycle progression, signal transduction, transcriptional regulation, DNA repair, chromatin remodeling, recombination repair, CRL3 complexes and ARH11 collaborates in tandem to mediate ubiquitination of target proteins. ARH11 mediating addition of the first ubiquitin on CRL3 targets. The functional specificity of the E3 ubiquitin-protein ligase

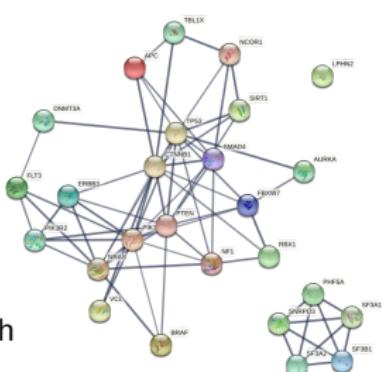
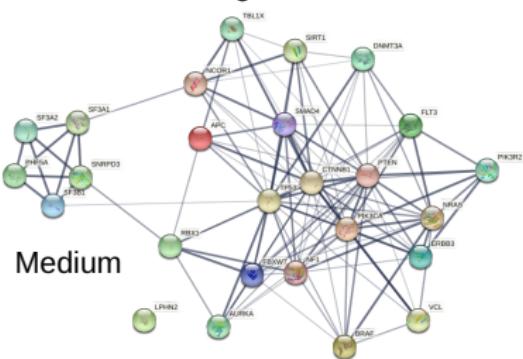
Evidence suggesting a functional link:
 Neighborhood in the Genome: co-epigenetic
 Gene Partners: co-regulated
 Coexpression Across Genomes: co-expressed
 Co-expression: co-expressed
 Experimental/Biochemical Data: co-regulated
 Association in Curated Databases: yes (xcore 0.900)
 Co-mentioned in PubMed Abstracts: yes (xcore 0.155)
 Combined Score: 0.911



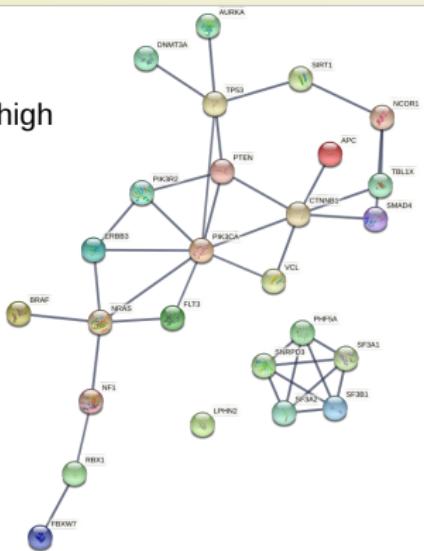
STRING: confidence levels

SDU

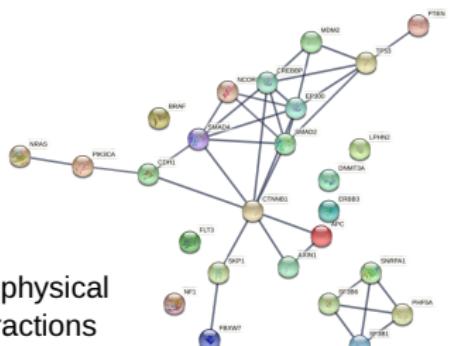
20 most mutated genes in cancer
and their "neighborhood"



Very high



Only physical
interactions



Network analysis: further possibilities

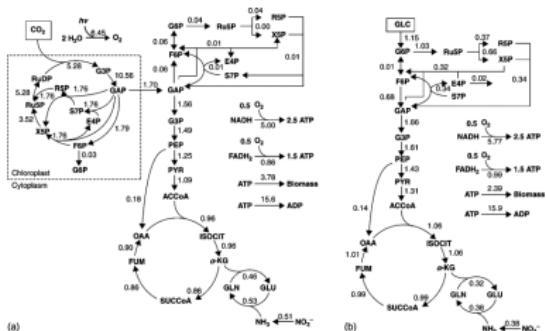
Software



Powerful software for network analysis
with many bioinformatics plugins

Analysis

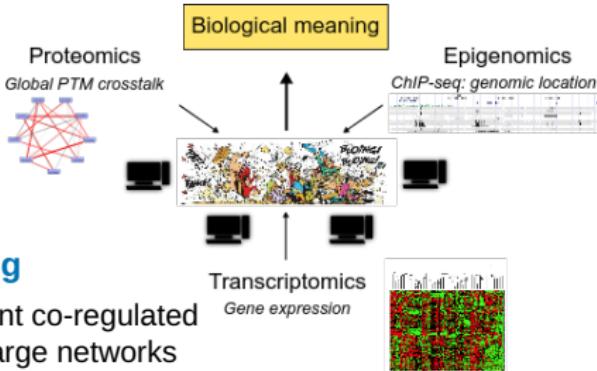
Popular in metabolomics and signalling
How does information propagate the cell?



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

