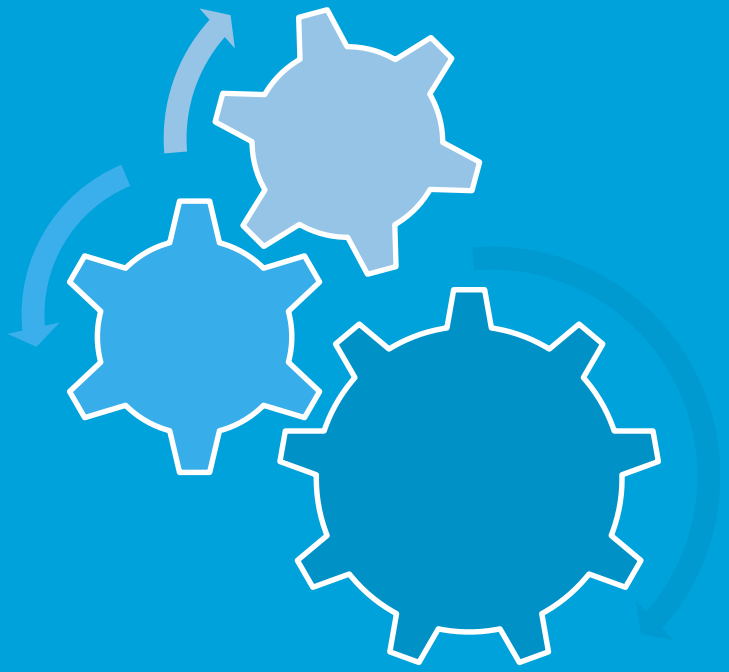
Network biology

Martina Summer-Kutmon, PhD
Assistant professor
MACSBIO

INT3007 - Systems Biology

31 October 2022



Maastricht University

Outline

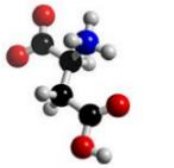
- General introduction
  - Context
  - Dataset
- Pathway analysis
  - Biological pathway models
  - Databases
  - Enrichment analysis
- Network analysis
  - Biological networks
  - Network sources
  - Software: Cytoscape

General introduction
Context and dataset

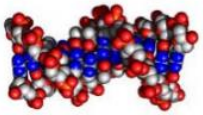
Images:
https://images.startups.co.uk/wp-content/uploads/2013/11/How_to_start_a_business_crop.jpg

Introduction

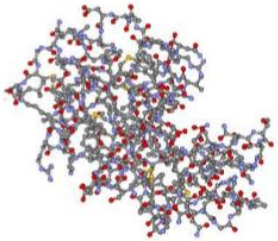
Molecules of life do not function in isolation ....



metabolites



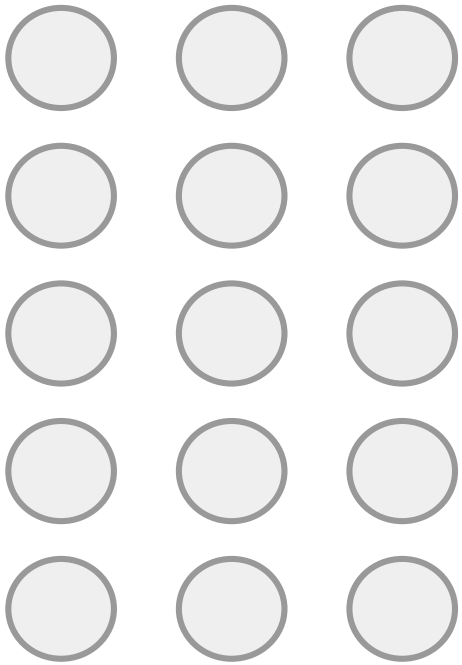
genes



proteins

but form complex networks that define a cell ....

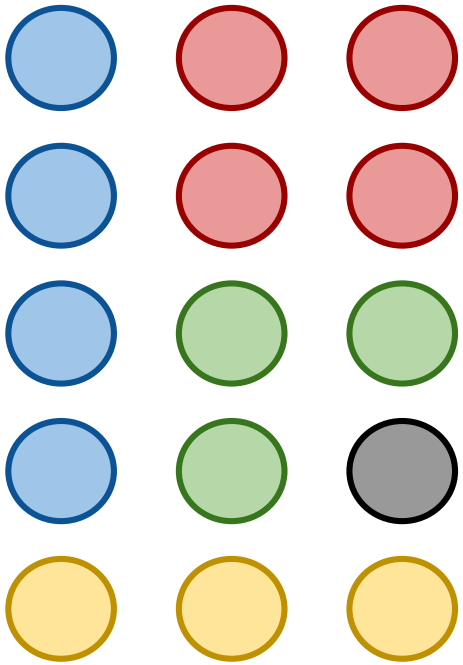
Introduction



Quantitative measurements

Isolated data points

Introduction

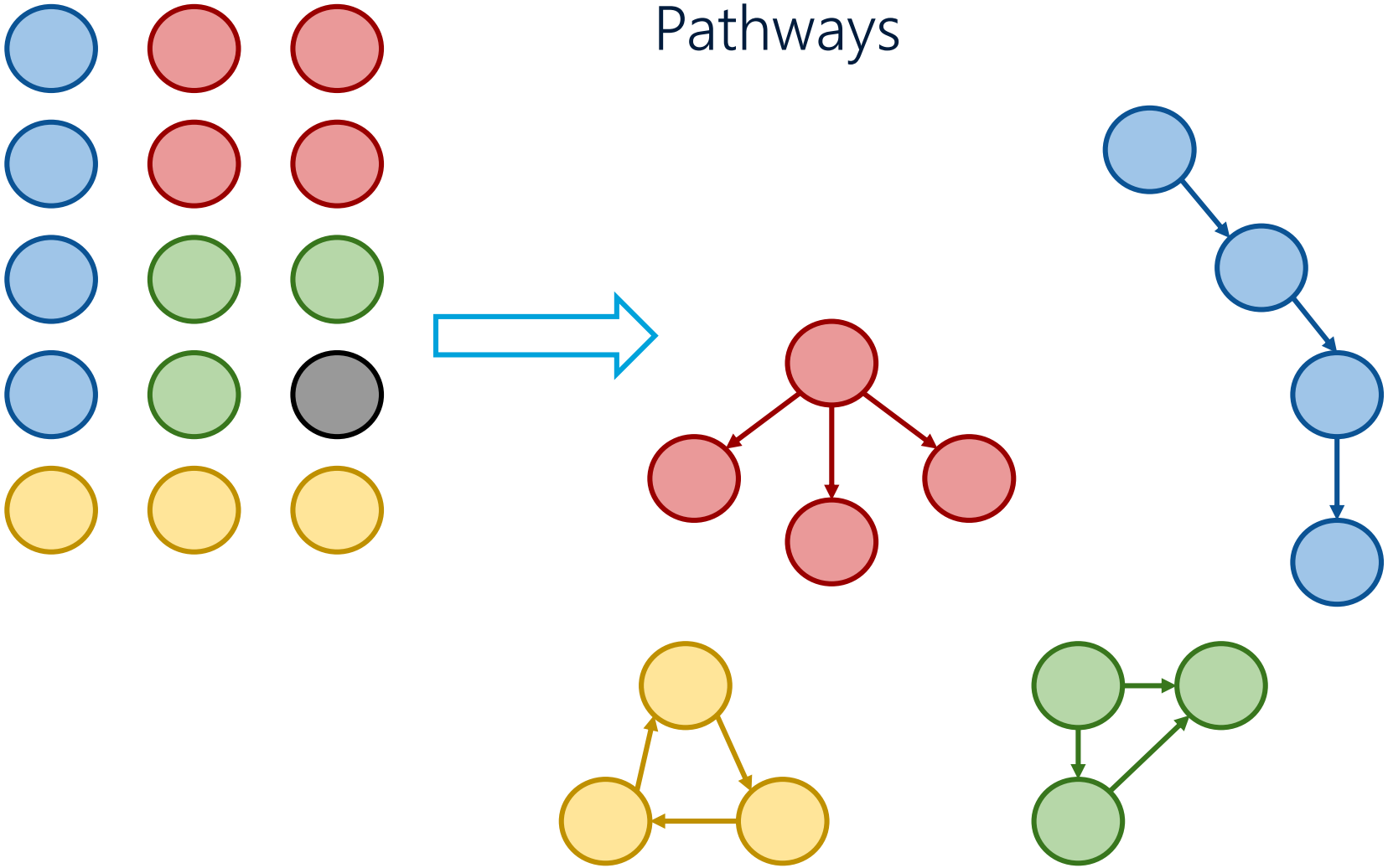


Comparative statistics
Isolated lists

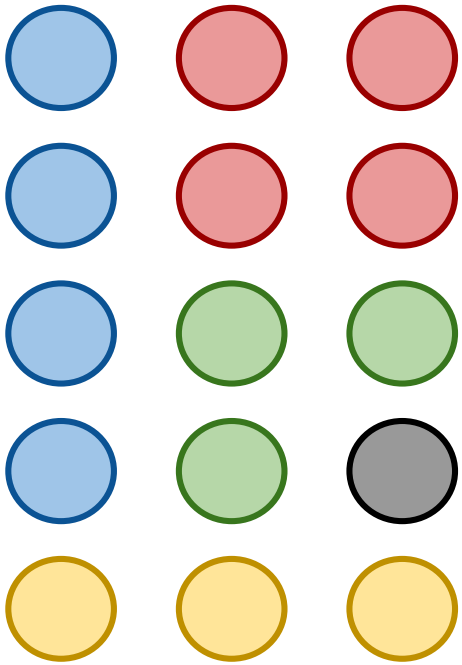
Clustering
Isolated groups

Gene sets
Functional groups

Introduction

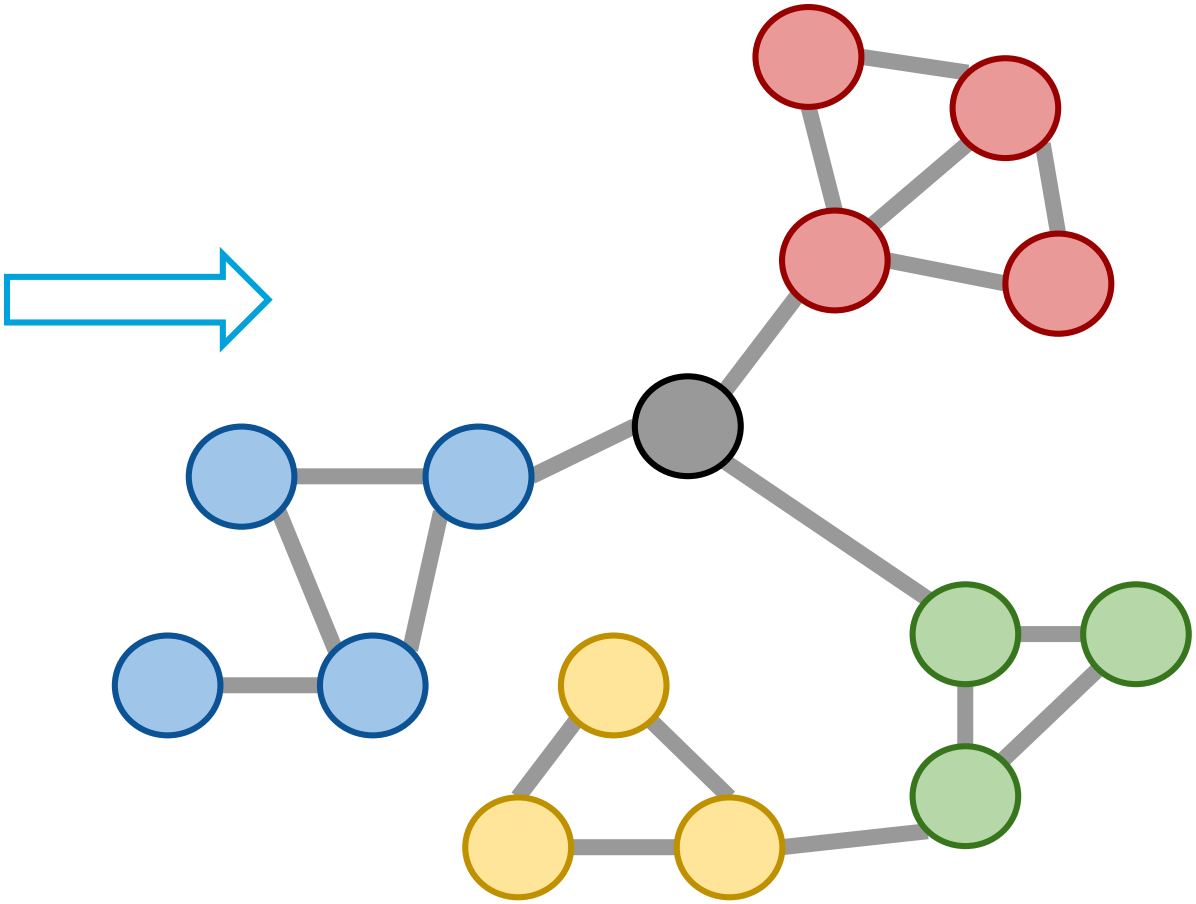


Introduction

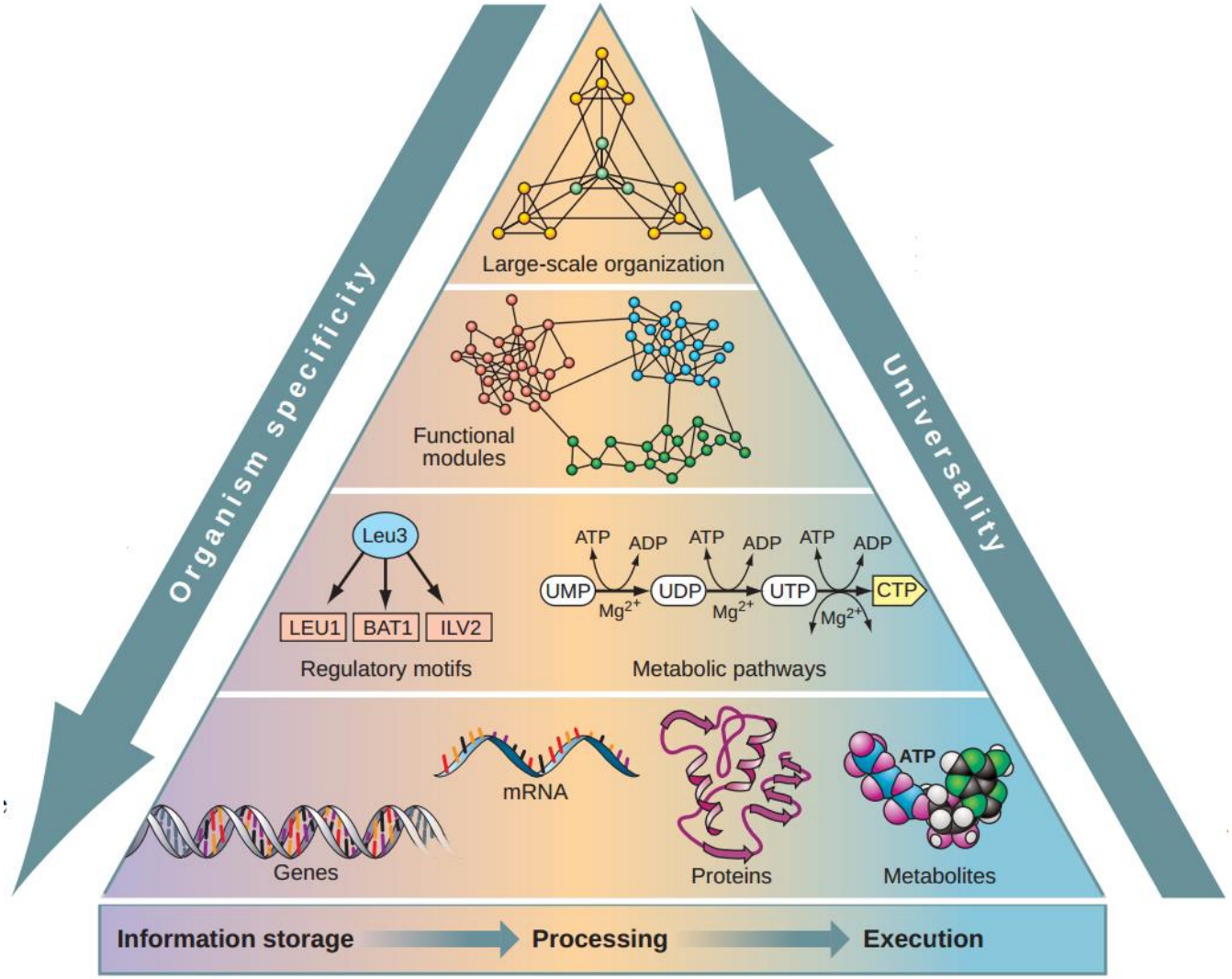


Systems organization

Networks

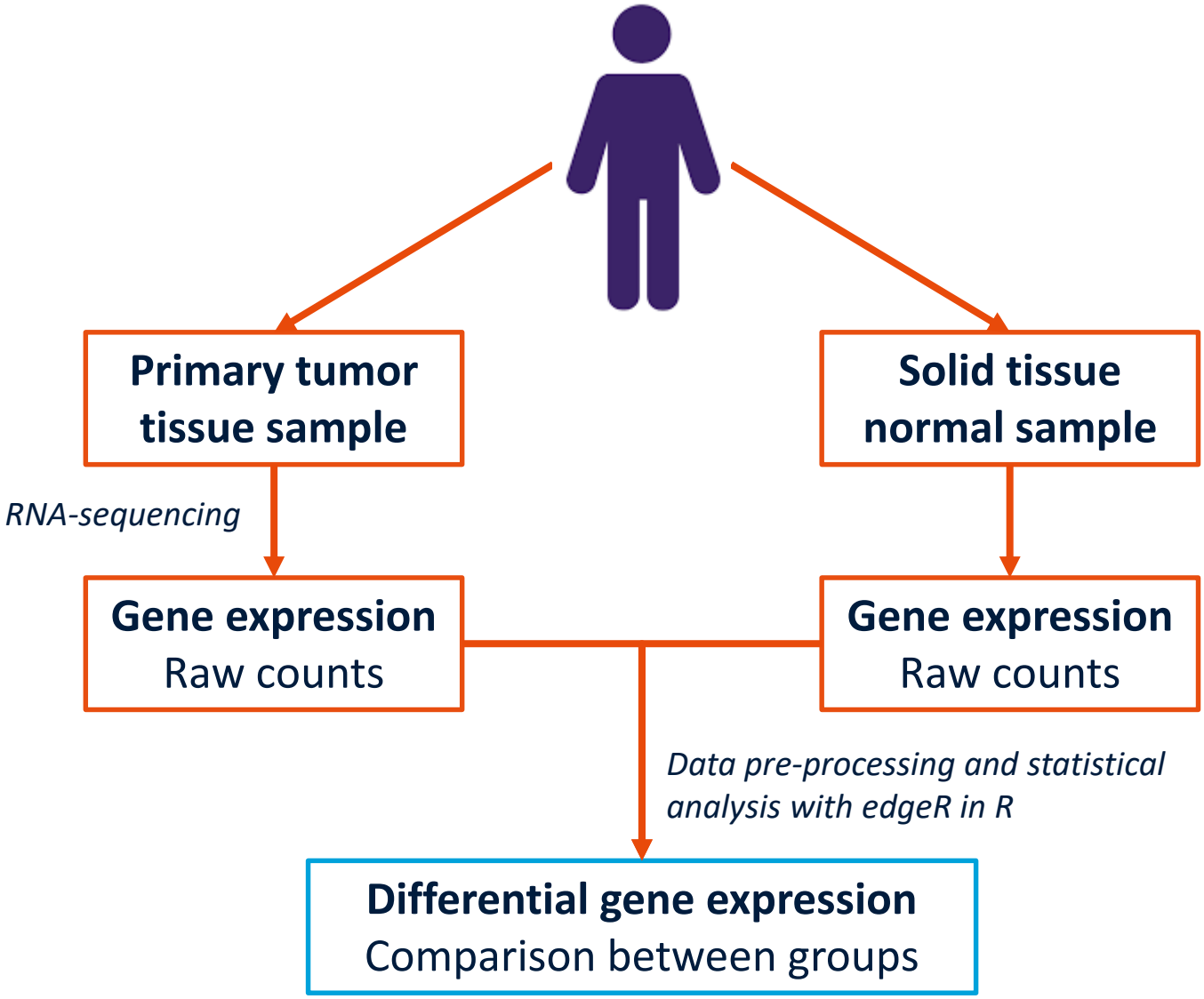


Introduction



doi: www.doi.org/10.1126/science.1078563

Experimental dataset



Experimental dataset

GeneID	GeneName	log2FC	P.Value	adj.P.Value
ENSG00000230657	PRB4	13.2739	0.0039	0.0978

- **GeneID** → identifier in online database
- **GeneName** → official gene symbol
- **log2FC** → log2 of fold change (ratio of the differences between cancer and healthy samples)
- **P.Value** → significance level of comparison
- **adj.P.Value** → corrected P.Value for multiple testing

Dataset

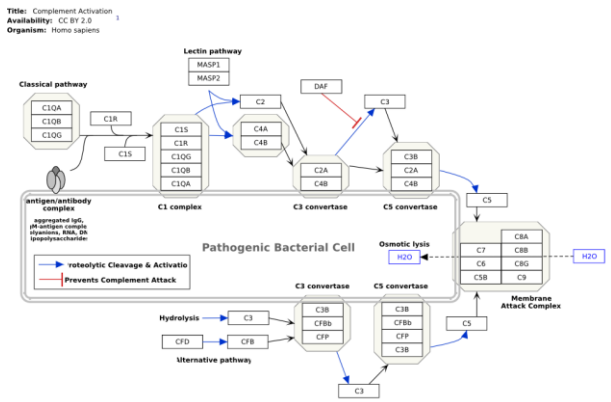
- log2FC
  - Is the gene more or less expressed in the cancer tissue sample compared to the normal tissue sample?

Negative log2FC
downregulation in cancer sample

Positive log2FC
upregulation in cancer sample

Example
log2FC = 1 → fold change = 2^1 = 2
gene is twice as expression cancer compared to healthy tissue sample

Why do we use the log2?
easier for interpretation – 0 = no chance, 1 is upregulation, -1 is downregulation



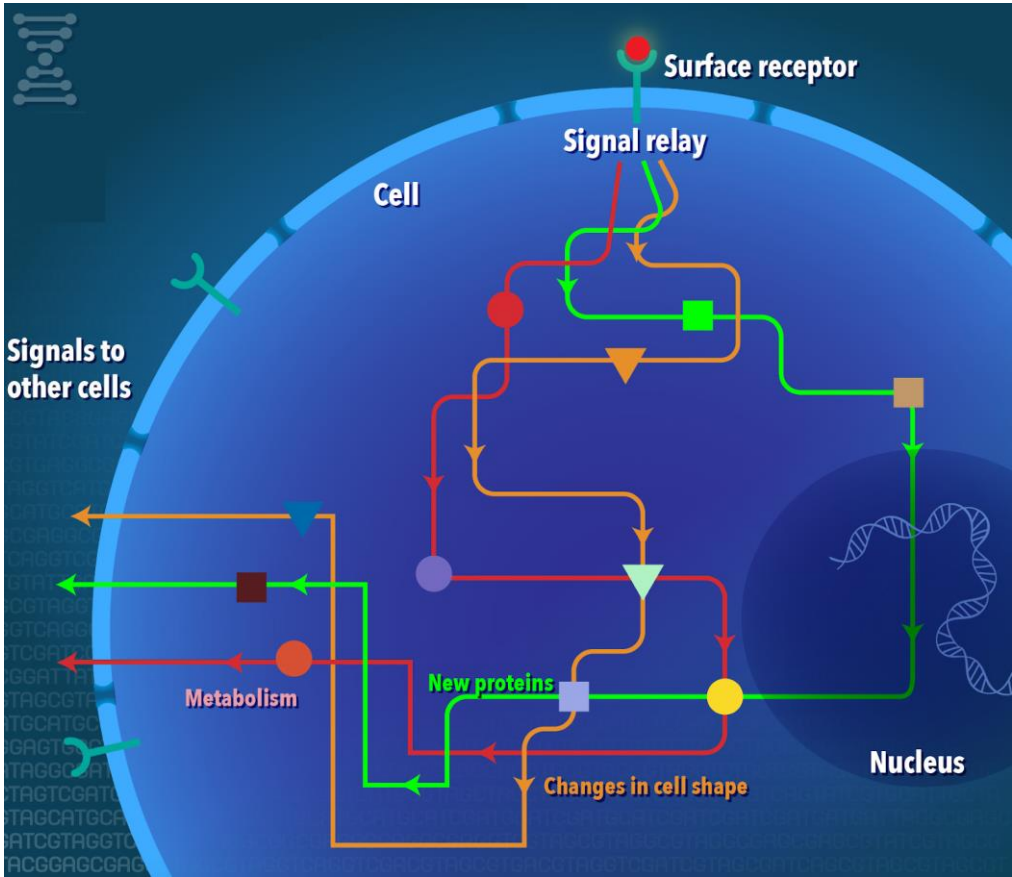
Pathway analysis

Biological pathways, online databases and enrichment analysis

Images:
https://www.wikipathways.org/index.php/Pathway:WP545

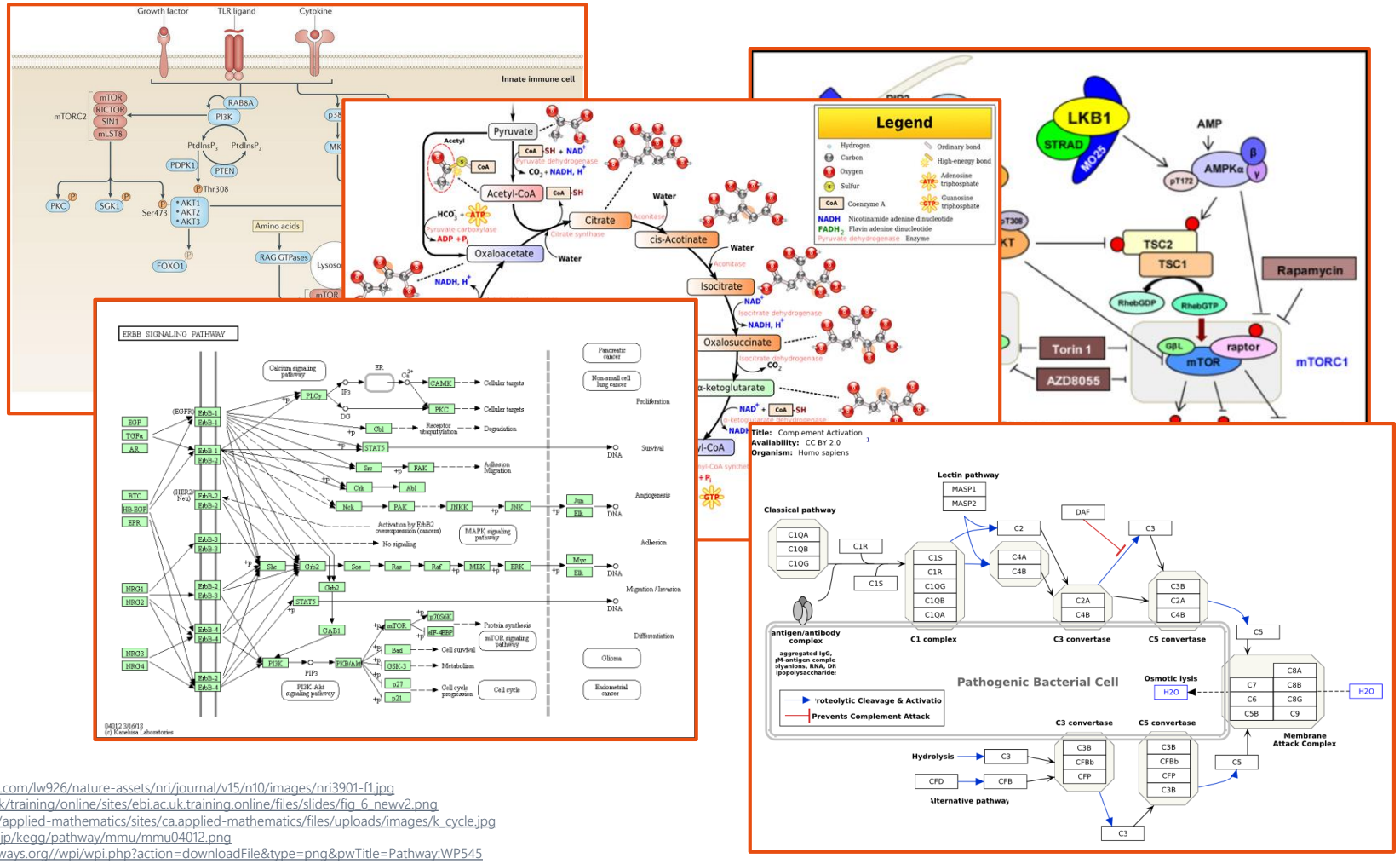
Biological pathways

- Signaling pathways
- Metabolic pathways
- Gene regulation pathways



Biological pathways

Pathway diagrams are found everywhere!



Images:
https://media.nature.com/lw926/nature-assets/nri/journal/v15/n10/images/nri3901-f1.jpg
https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/slides/fig_6_newv2.png
https://uwaterloo.ca/applied-mathematics/sites/ca.applied-mathematics/files/uploads/images/k_cycle.jpg
http://www.genome.jp/kegg/pathway/mmu/mmu04012.png
http://www.wikipathways.org//wpi/wpi.php?action=downloadFile&type=png&pwTitle=Pathway:WP545

Biological pathways

Pathway diagrams are found everywhere!

Utility to biologists as conceptual models is obvious

Biological pathways

Pathway diagrams are found everywhere!

⇩

Utility to biologists as conceptual models is obvious

⇩

If modeled properly - immensely useful for computational analysis and interpretation of large-scale experimental data

Online databases

- Where can we find these pathway models?
- Different online databases
- Most commonly used are:

Database	Link
KEGG	http://www.genome.jp/kegg/
Reactome	https://www.reactome.org
WikiPathways	https://www.wikipathways.org

WikiPathways

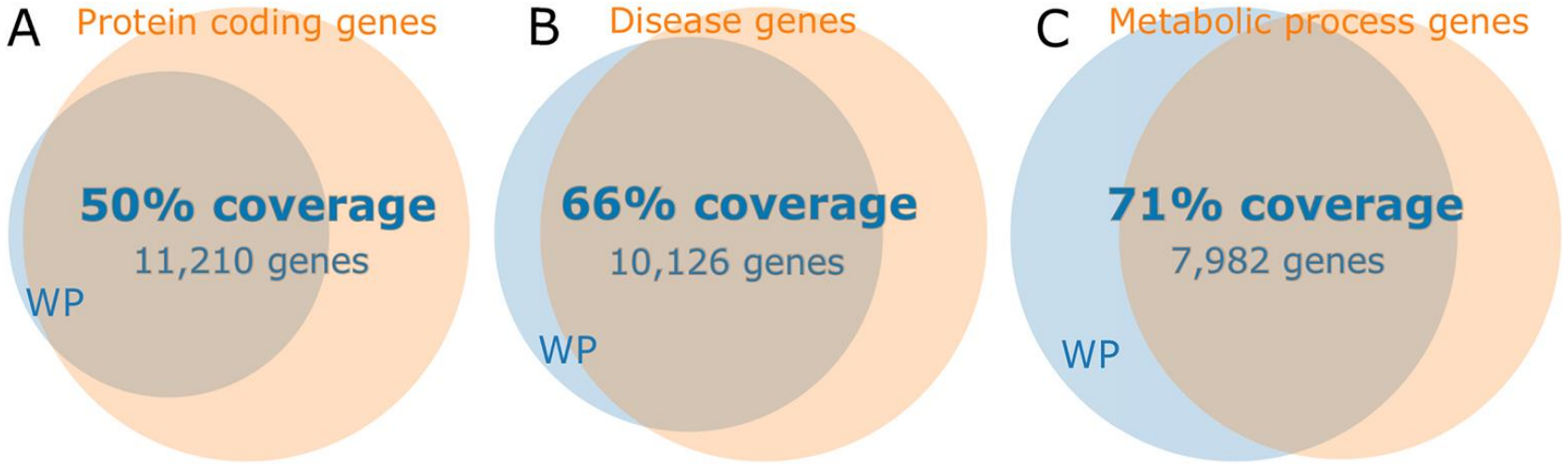
- Launched in 2008 as an experiment in community-based curation of biological pathways at Maastricht University together with the Gladstone Institutes in San Francisco

Too much data!
Difficult to keep knowledge up-to-date, accessible and integrated

Taking advantage of direct participation by a greater portion of the community (**crowdsourcing**)

WikiPathways

- Latest release on 10 October 2022
- 843 curators and >3000 pathways for ~30 different species



Images:
https://doi.org/10.1093/nar/gkx1064

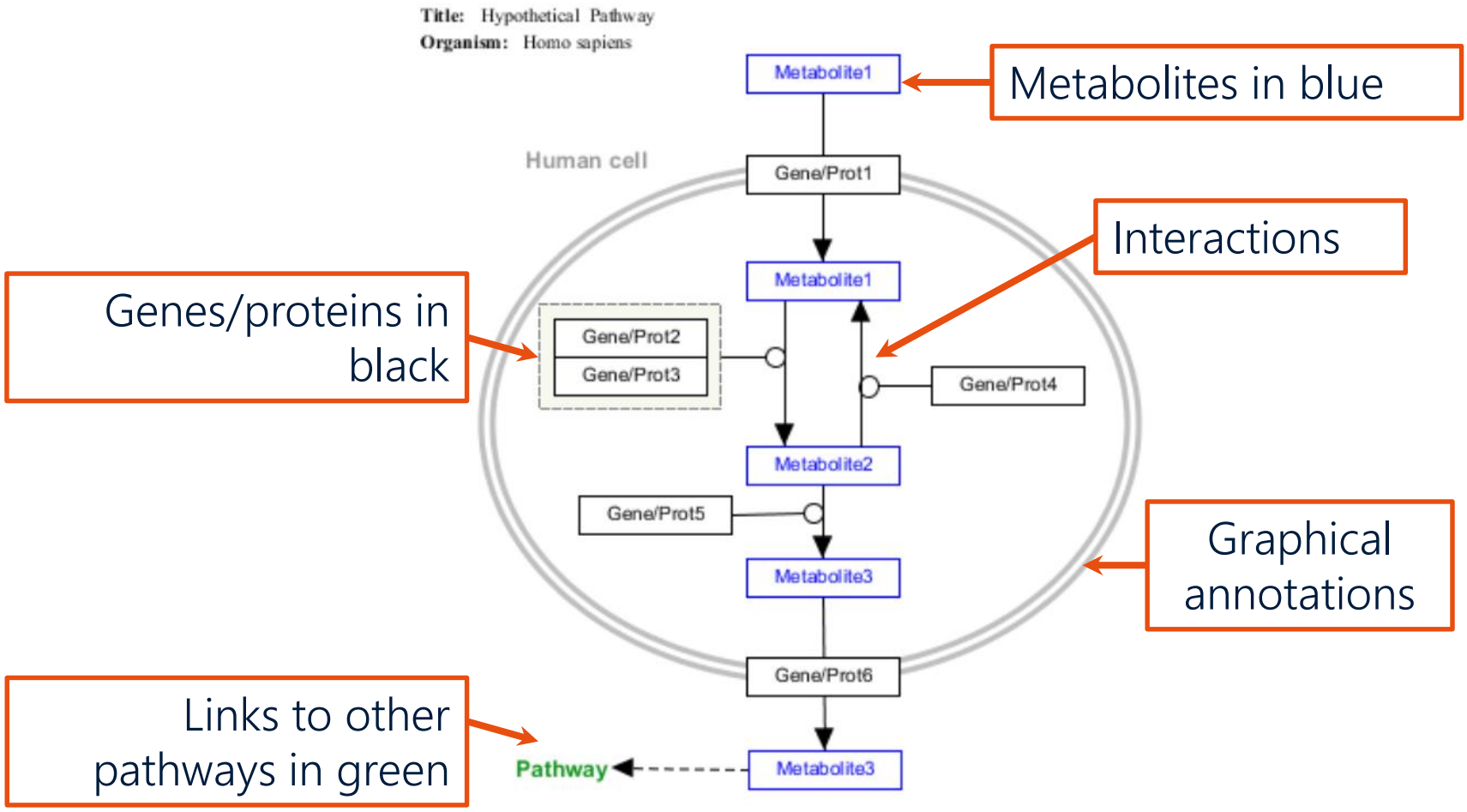
Why pathway analysis?

- "A picture is worth a thousand words."
  - Intuitive and simple
  - Puts data into a biological context → analysis on functional level
  - More efficient than looking up single gene information

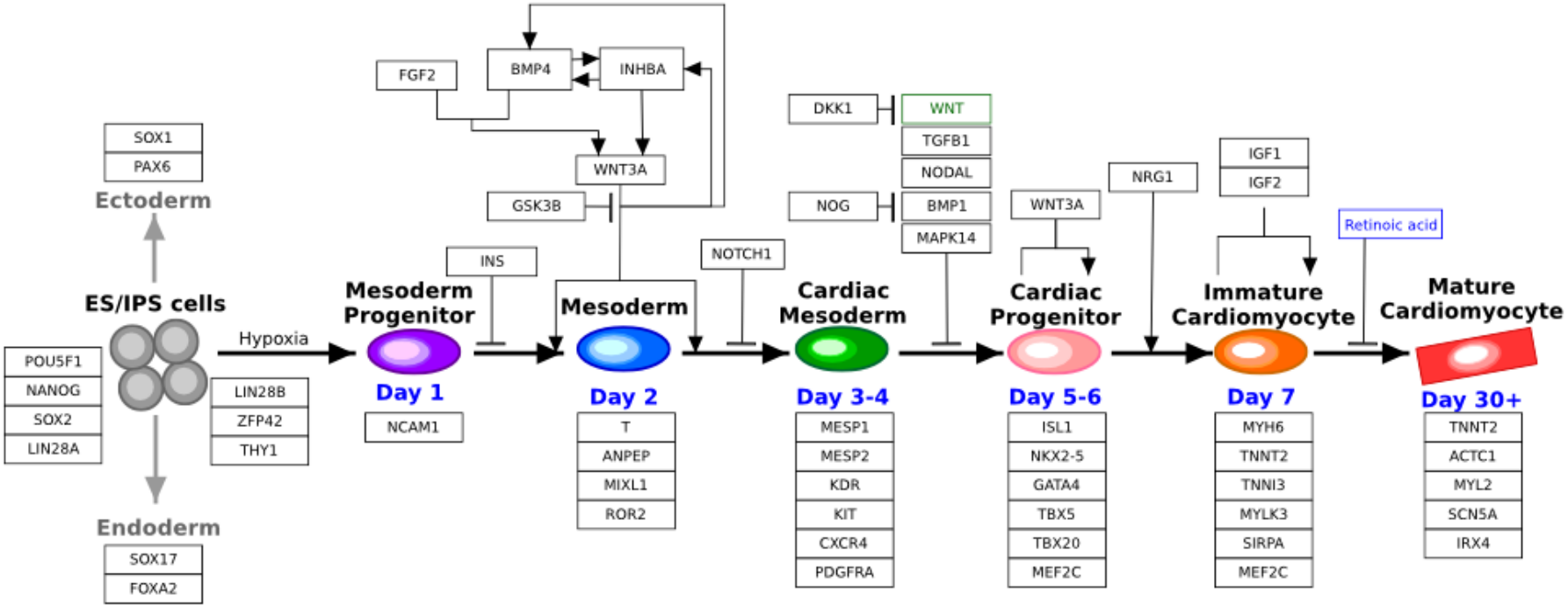
Why pathway analysis?

- "A picture is worth a thousand words."
  - Reduces complexity by grouping genes, proteins and other molecules → several hundred pathways instead of thousands of genes
  - Higher explanatory power than a simple gene list
  - Visual representation

Pathway models in WikiPathways



Pathway creation

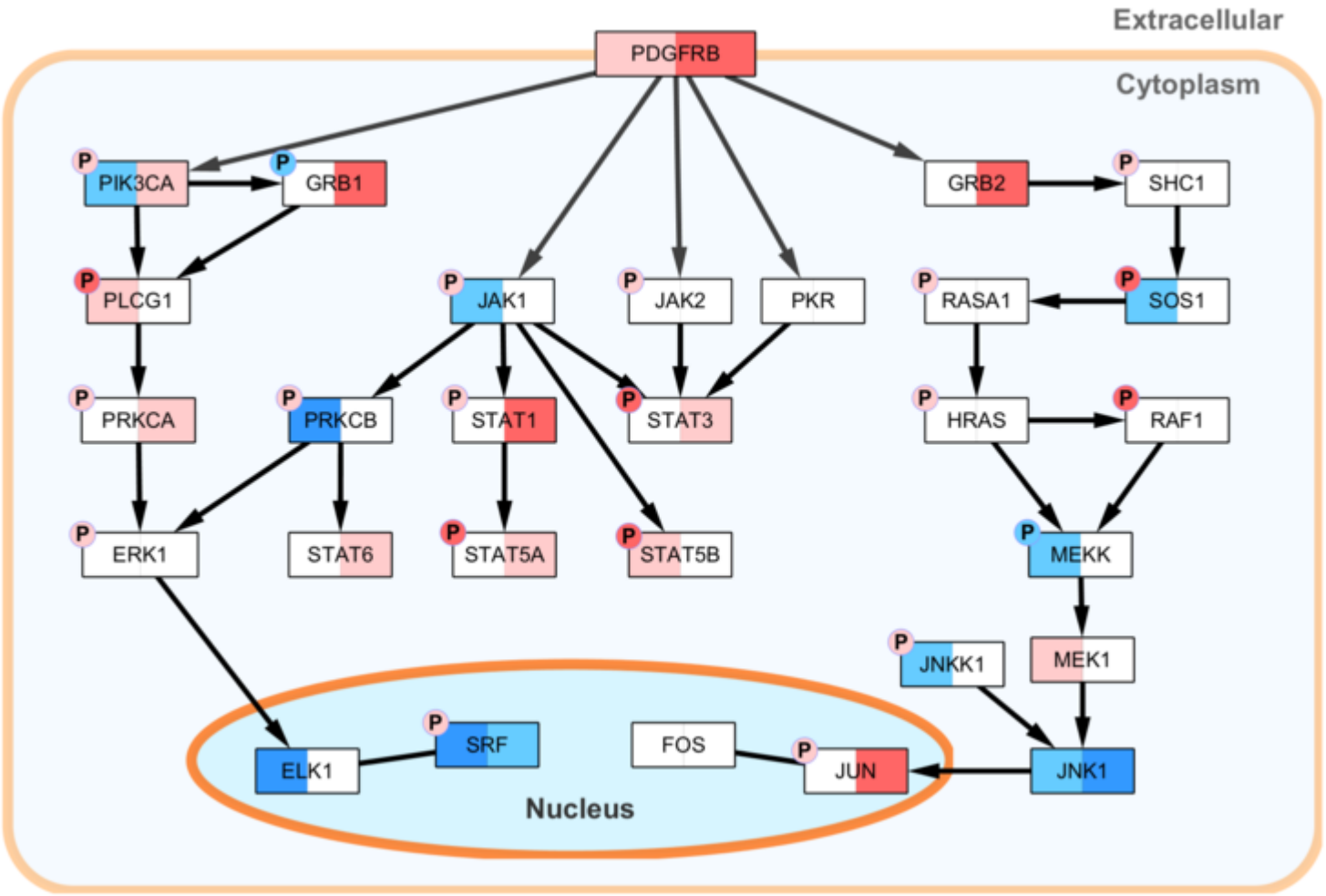


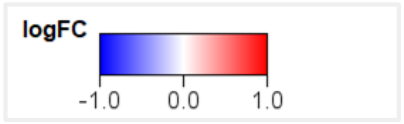
Human Cardiac Progenitor Differentiation Pathway:
https://www.wikipathways.org/instance/WP2406

Data visualization

- Data visualization on data nodes and interactions
- Color gradients and color rules
- Multi-omics visualization
- Time-series visualization

Data visualization

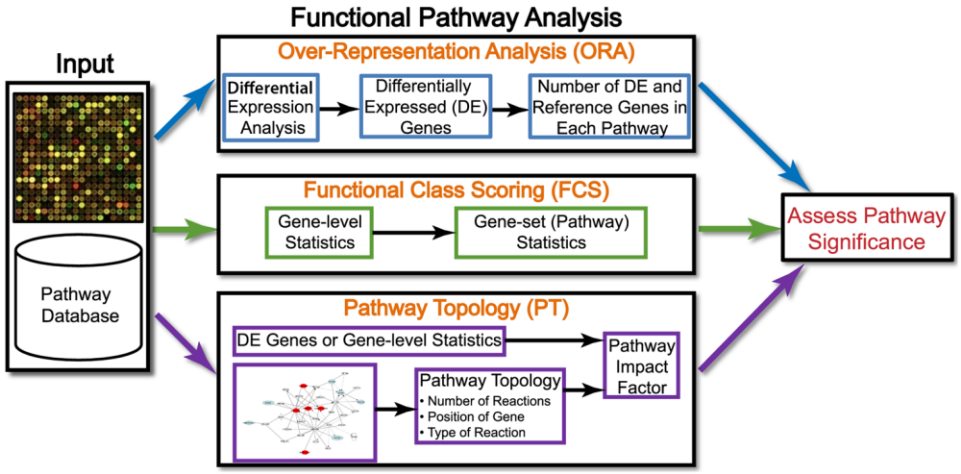




Data visualization on nodes in pathway

Pathway analysis methods

1. Overrepresentation analysis (ORA)
2. Functional Class Scoring (FCS)
3. Pathway Topology Based (PT)



Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Computational Biology, 8(2), e1002375.

Pathway analysis methods

- Overrepresentation analysis:
  - Input list → e.g. significantly up- or down-regulated genes
  - Background list → e.g. all measured genes
  - Statistical test → e.g. Fisher's exact test (hypergeometric test)

$$\text{Z-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

  - Z-Score is calculated for each pathway
    - Results in ranked list of pathways
  - Four variables in the formula: N, R, n, r

ORA

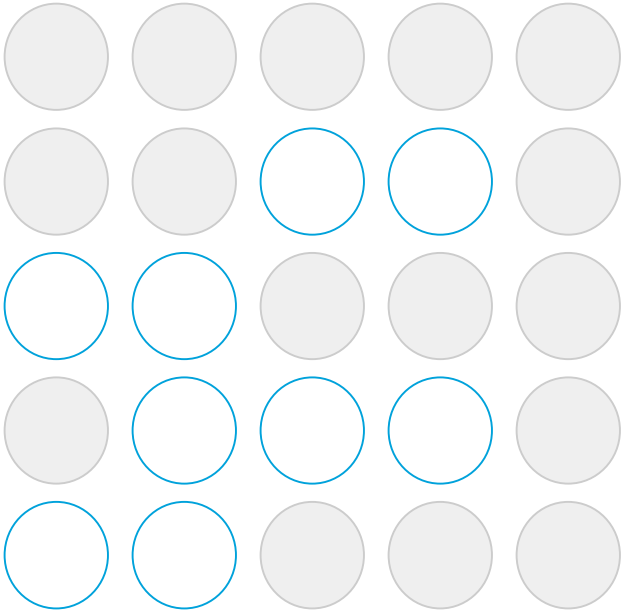
$$\text{Z-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

N = 25

background list (total number of measured genes in experiment)

ORA

$$\text{Z-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$



N = 25

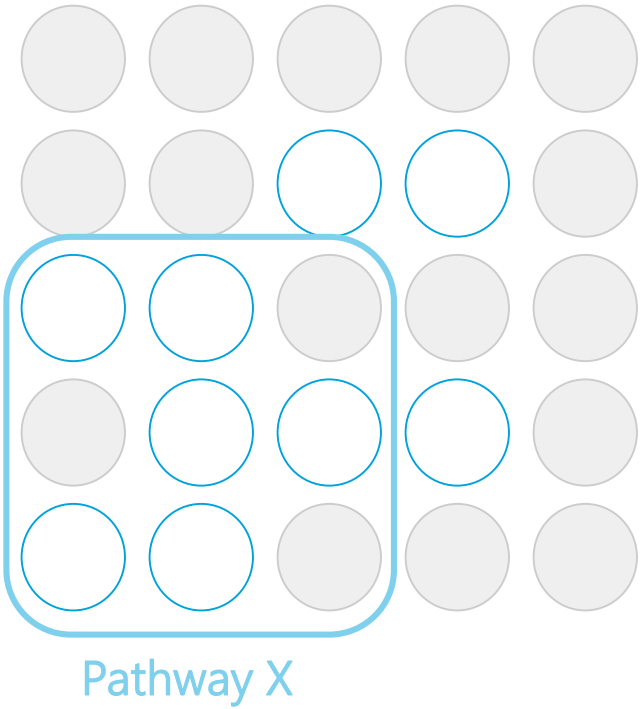
background list (total number of measured genes in experiment)

R = 9

input list (number of changed genes in experiment)

ORA

$$\text{Z-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$



N = 25

background list (total number of measured genes in experiment)

R = 9

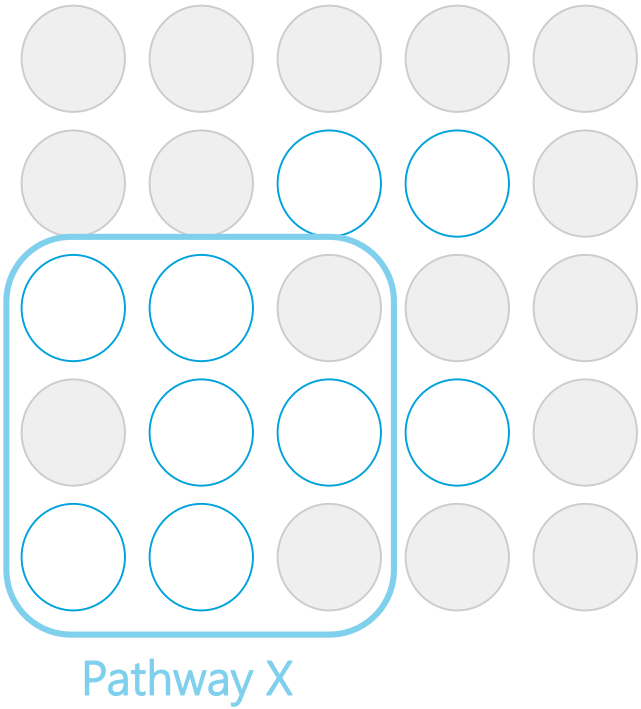
input list (number of changed genes in experiment)

n = 9

total number of genes in pathway

ORA

$$\text{Z-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$



N = 25

background list (total number of measured genes in experiment)

R = 9

input list (number of changed genes in experiment)

n = 9

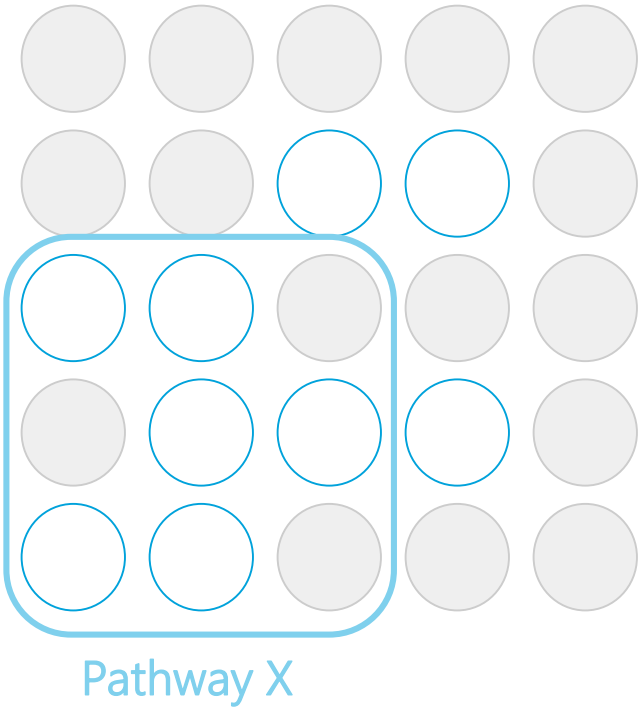
total number of genes in pathway

r = 6

number of changed genes in pathway

ORA

$$\text{Z-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$



N = 25

background list (total number of measured genes in experiment)

R = 9

input list (number of changed genes in experiment)

n = 9

total number of genes in pathway

r = 6

number of changed genes in pathway

Enrichment score for pathway X = 2.347

Permutation test to assess significance (p-value)

Z-Score ORA

Be aware!!

- What does the Z-Score tell you?
  - Z-Score > 1.96
    - Significantly more genes than expected are changed in the pathway → altered pathways in the experiment (different between the groups)
  - Z-Score = 0
    - Distribution of changed genes in the pathway is the same as in the complete dataset
  - Z-Score = < -1.96
    - Significantly less genes than expected are changed in the pathway → very stable pathway (not affected in experiment)

Z-Score ORA

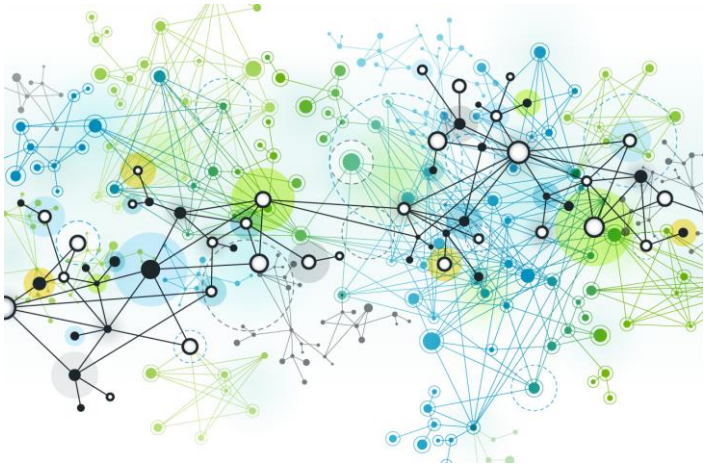
- Be aware!
  - ORA and FCS **do not** take pathway topology into account!
  - You don’t know yet where the changes occur in the pathway.
  - Always look at the pathway diagrams and study the changes to make the right conclusions!

Pathway analysis methods

- Overrepresentation analysis:
  - Ranked list of pathways

Up-regulated pathways (log2FC > 2, p-value < 0.05)	**Z-score**	**Perm. p-value**
Cell Cycle	6.12	0.001
G1 to S cell cycle control	4.26	0.002
Synaptic Vesicle Pathway	3.89	0.001
DNA Damage Response	3.88	0.002
ATM Signaling Pathway	3.80	0.001

Down-regulated pathways (log2FC < -2, p-value < 0.05)	**Z-score**	**Perm. p-value**
Complement and Coagulation Cascades	5.87	0.001
Complement Activation	5.84	0.001
Adipogenesis	5.49	0.001
Differentiation of white and brown adipocyte	5.44	0.001
Triacylglyceride Synthesis	4.53	0.001



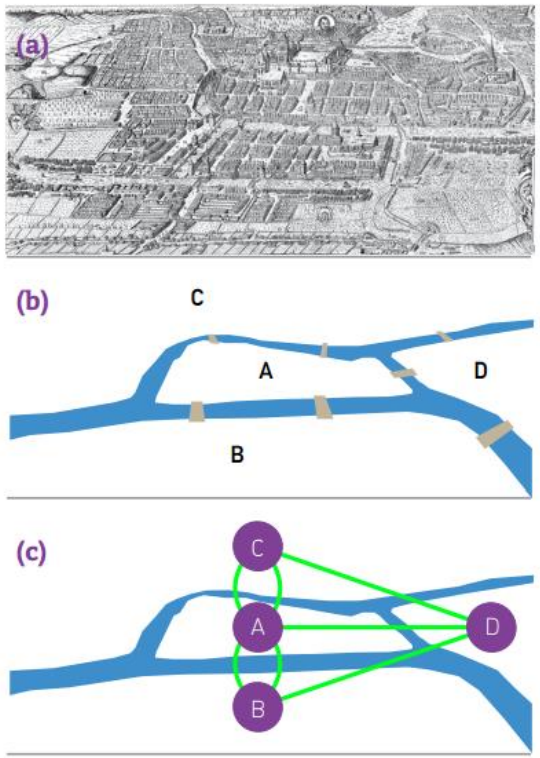
Network analysis

Networks, online databases, Cytoscape

Images:
http://www.ehealthresearchnetwork.com/wp-content/uploads/2017/03/Network-Design-A.jpg

Network science

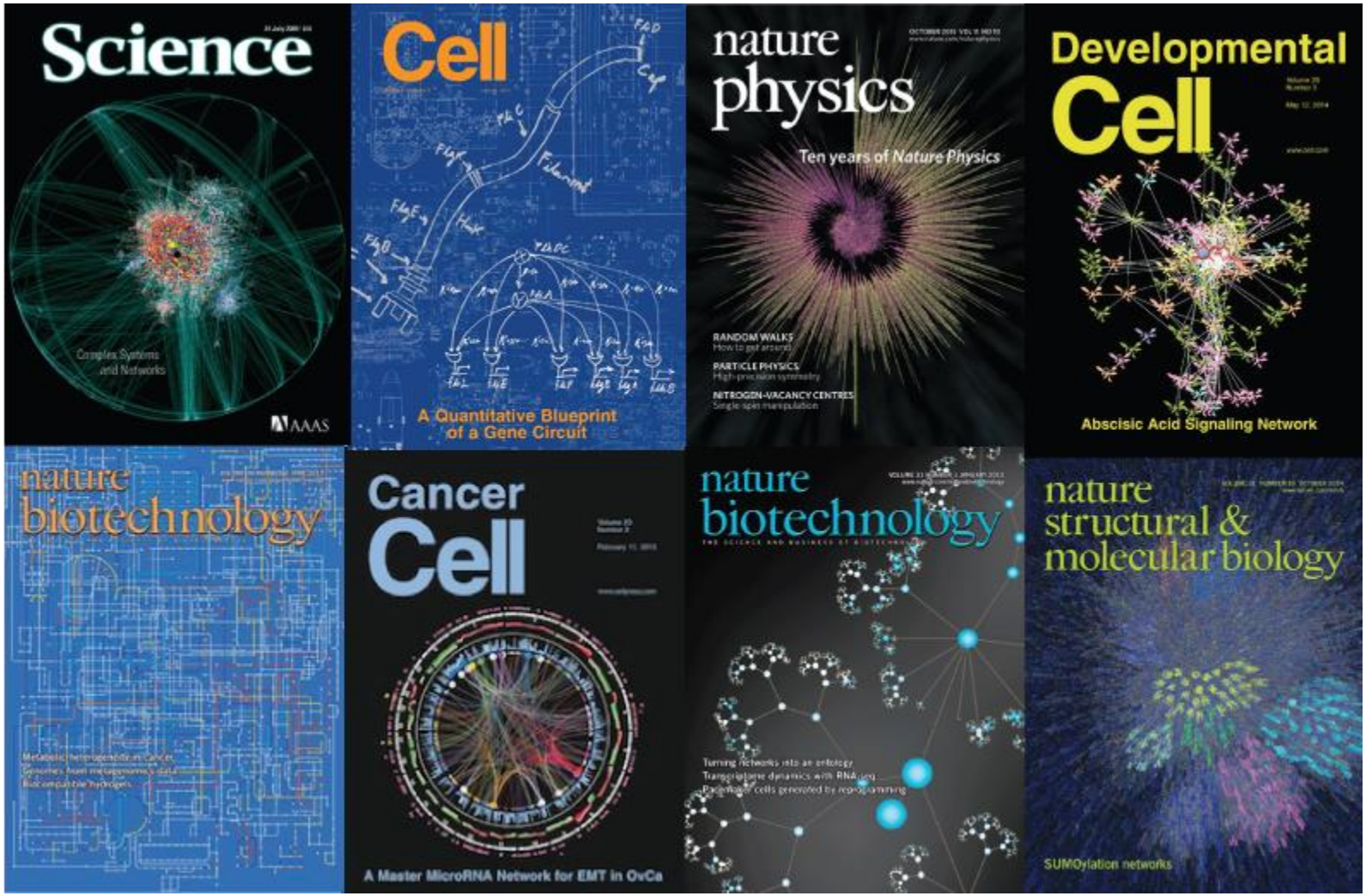
- Building on the field of graph theory
  - 1735 Koenigsberg (now Kaliningrad, Russia)
  - Leonard Euler (Swiss mathematician)
  - Walk across all seven bridges and never cross the same twice
  - Euler offered mathematical proof that such a path does not exist - using a graph representation



Network science

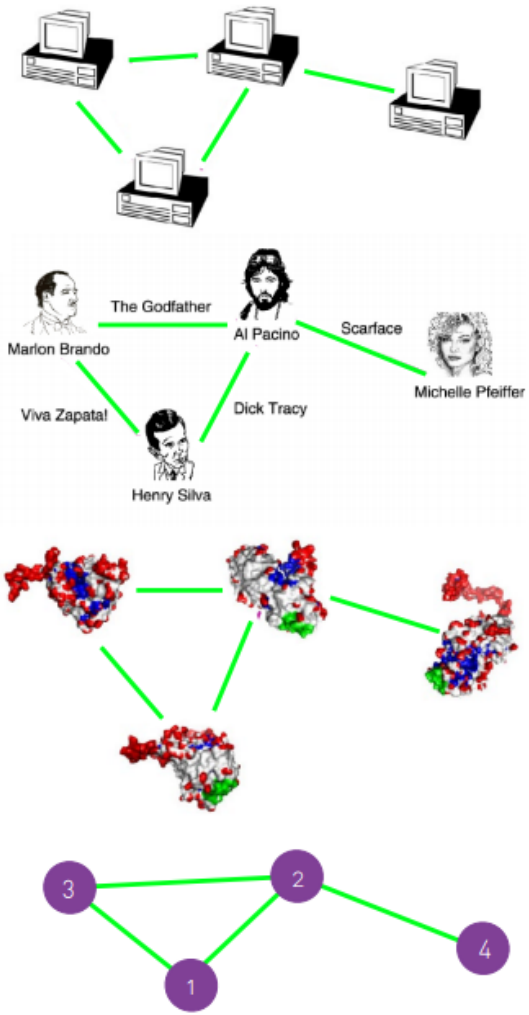
- Building on the field of graph theory

Networks in science



Network science

- Building on the field of graph theory



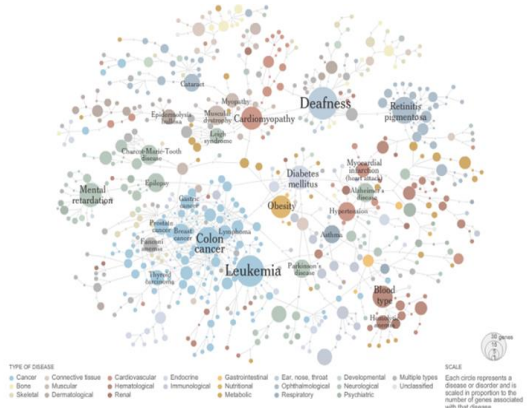
Network as universal concept

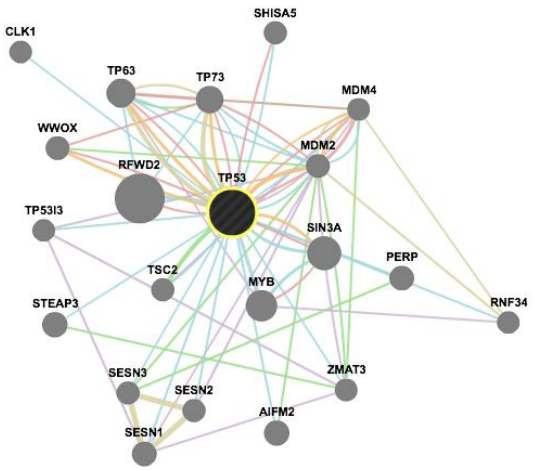
Why networks in biology?

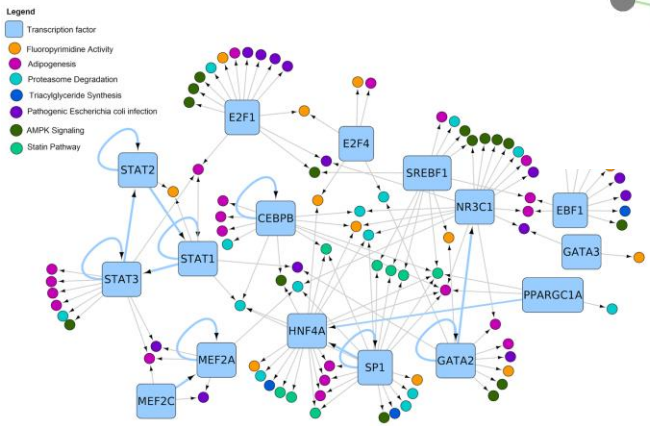
- Study biological complexity
- More efficient than tables
- Great for data integration
- Intuitive visualization

Why networks in biology?

- Types of networks
  - Molecular networks
    - Protein-protein interaction networks
    - Metabolic networks
    - Regulatory networks
  - Cell-cell communication
  - Nervous systems
  - Human disease network
  - Social networks

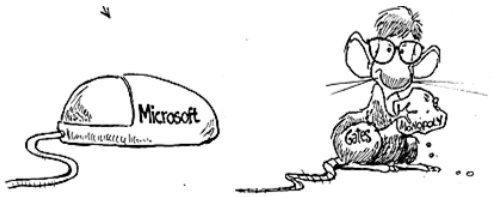






Terminology

- Let's make sure, we talk about the same things using the same terms!



Network

- A network is a graphical representation of a set of **objects** where some pairs of objects are connected by **links**.

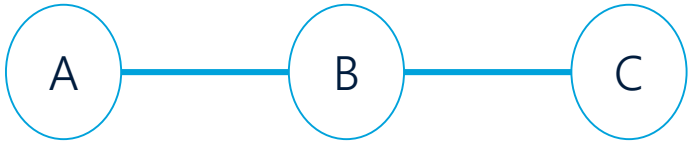
Network

- A network is a graphical representation of a set of **objects** where some pairs of objects are connected by **links**.

Objects in the network are called **nodes** (A and B).
Links in the network are called **edges** or interactions.

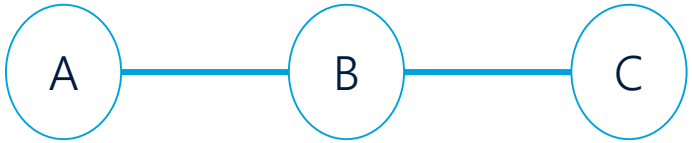
Neighbour

A neighbour is a node that is **linked** with another **node** through a **direct edge**.



Neighbour

A neighbour is a node that is **linked** with another **node** through a **direct edge**.



A is a neighbour of B but not of C.
B has two neighbours - A and C.
C is a neighbour of B but not A.

Path

- A path is a **sequence of edges** which connect a sequence of nodes.
- A path can intersect itself and pass through the same node/edge repeatedly.

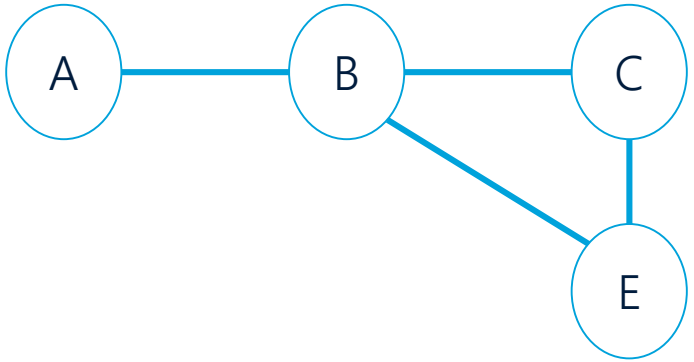
Path

- A path is a **sequence of edges** which connect a sequence of nodes.
- A path can intersect itself and pass through the same node/edge repeatedly.

Path(s) from A to E?

A - B - C - E

A - B - E

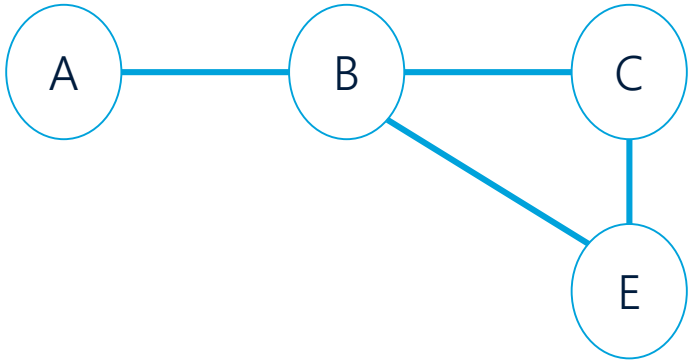


Distance

- Distance between two nodes is the number of edges along the path connecting them.
- If two nodes are disconnected, the distance is infinity.
- The **shortest path** is the path with the minimal number of edges necessary to get from one node to another.

Distance

- Distance between two nodes is the number of edges along the path connecting them.
- If two nodes are disconnected, the distance is infinity.
- The **shortest path** is the path with the minimal number of edges necessary to get from one node to another.

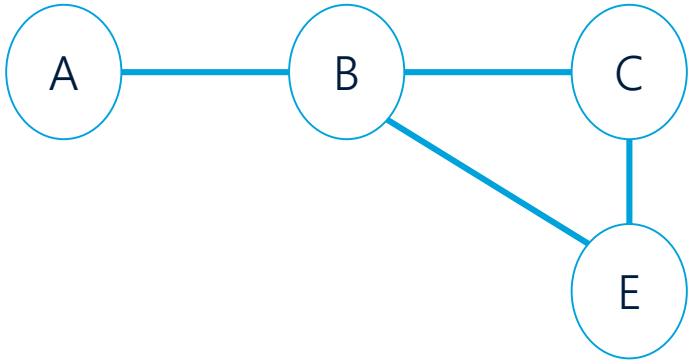


Distance from A to E?

A - B - C - E	3
A - B - E	2

Distance

- Distance between two nodes is the number of edges along the path connecting them.
- If two nodes are disconnected, the distance is infinity.
- The **shortest path** is the path with the minimal number of edges necessary to get from one node to another.



Distance from A to E?

A - B - C - E 3

A - B - E 2 (2 < 3 → shortest path)

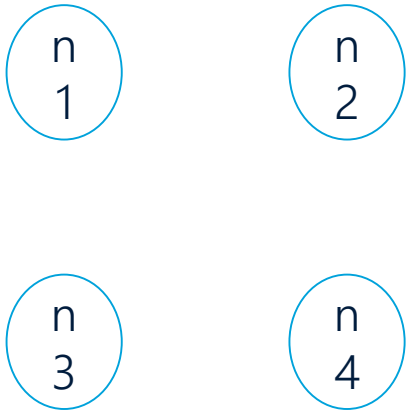
Adjacency matrix

Mathematical representation

$A_{ij} = 1$ there is an edge between node i and j

$A_{ij} = 0$ there is no edge between node i and j

$$A_{ij} = \begin{array}{c} \\ n1 \\ n2 \\ n3 \\ n4 \end{array} \begin{array}{c} \begin{array}{cccc} n1 & n2 & n3 & n4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$



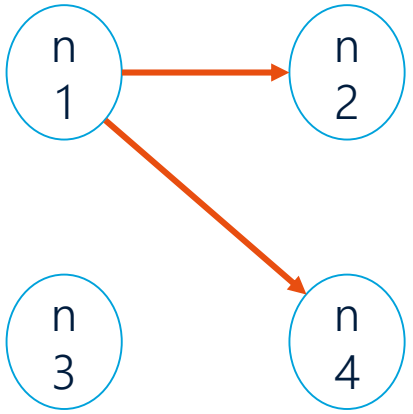
Adjacency matrix

Mathematical representation

$A_{ij} = 1$ there is an edge between node i and j

$A_{ij} = 0$ there is no edge between node i and j

$$A_{ij} = \begin{array}{c} \\ n1 \\ n2 \\ n3 \\ n4 \end{array} \begin{array}{c} \begin{array}{cccc} n1 & n2 & n3 & n4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$



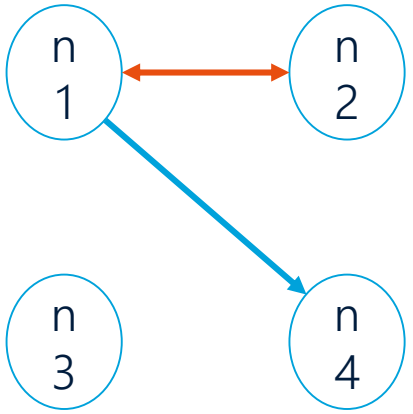
Adjacency matrix

Mathematical representation

$A_{ij} = 1$ there is an edge between node i and j

$A_{ij} = 0$ there is no edge between node i and j

$$A_{ij} = \begin{array}{c} \\ n1 \\ n2 \\ n3 \\ n4 \end{array} \begin{array}{c} \begin{array}{cccc} n1 & n2 & n3 & n4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$



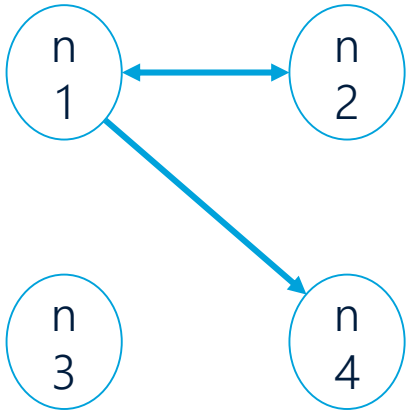
Adjacency matrix

Mathematical representation

$A_{ij} = 1$ there is an edge between node i and j

$A_{ij} = 0$ there is no edge between node i and j

$$A_{ij} = \begin{array}{c} \\ n1 \\ n2 \\ n3 \\ n4 \end{array} \begin{array}{c} \begin{array}{cccc} n1 & n2 & n3 & n4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$



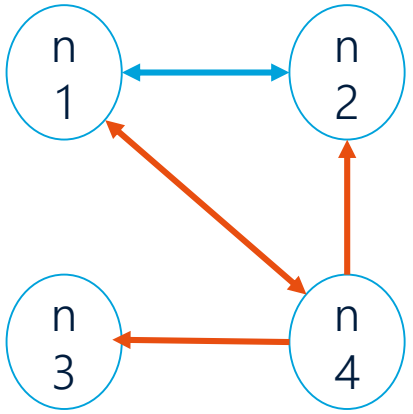
Adjacency matrix

Mathematical representation

$A_{ij} = 1$ there is an edge between node i and j

$A_{ij} = 0$ there is no edge between node i and j

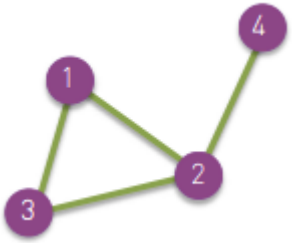
$$A_{ij} = \begin{array}{c} \\ n1 \\ n2 \\ n3 \\ n4 \end{array} \begin{array}{c} \begin{array}{cccc} n1 & n2 & n3 & n4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$



Undirected vs. directed networks

Undirected

Links: undirected (symmetrical)



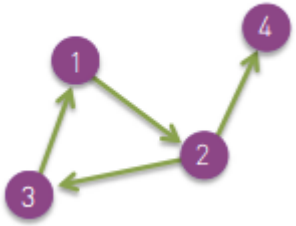
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Examples:

Coauthorship, actor network, protein interactions

Directed

Links: directed (arcs)



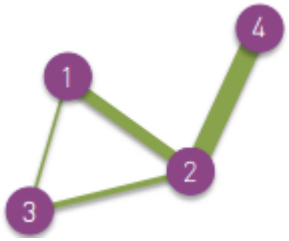
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Examples:

URLs (internet), phone calls, metabolic reactions

Weighted networks

Edges have a defined weight, strength or flow parameter



$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

Examples:

Correlation networks, route planning, mobile phone calls

Centrality measures

- Indicators which identify the **most important nodes** and/or **edges** in the network
  - Degree centrality
  - Betweenness centrality
  - Clustering coefficient
  - …

Help to answer the following questions:

How influential is a person?
How important is a room in a building?
How much influence has a mutation in a protein?

Degree centrality

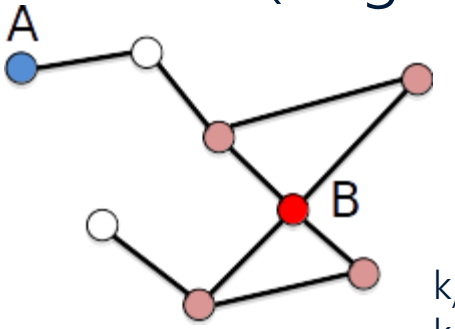
- Undirected:
  - node degree = number of edges connected to the node

- Directed:
  - in-degree = number of edges pointing towards a node (regulators)
  - out-degree = number of edges going out of a node (targets)

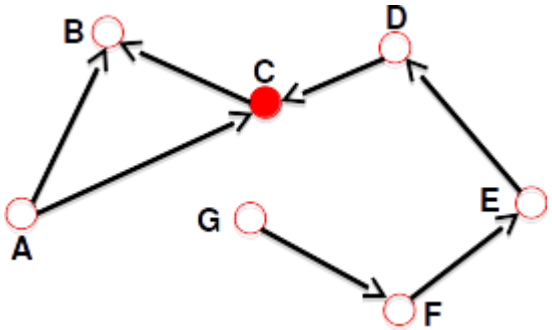
Degree centrality

- Undirected:
  - node degree = number of edges connected to the node

- Directed:
  - in-degree = number of edges pointing towards a node (regulators)
  - out-degree = number of edges going out of a node (targets)



$k_A = 1$
$k_B = 4$



$k_{C\_in} = 2$
$k_{C\_out} = 1$

Degree centrality

- Biological interpretation
  - Nodes with a high degree tend to be essential
  - Nodes with a high degree are also called **hub nodes**

  - Case 1

Betweenness centrality

- Betweenness = number of shortest paths going through a node

$$C_b(n) = \sum \frac{\delta_{st}(n)}{\delta_{st}}$$

δ_{st} = number of shortest path from s to t

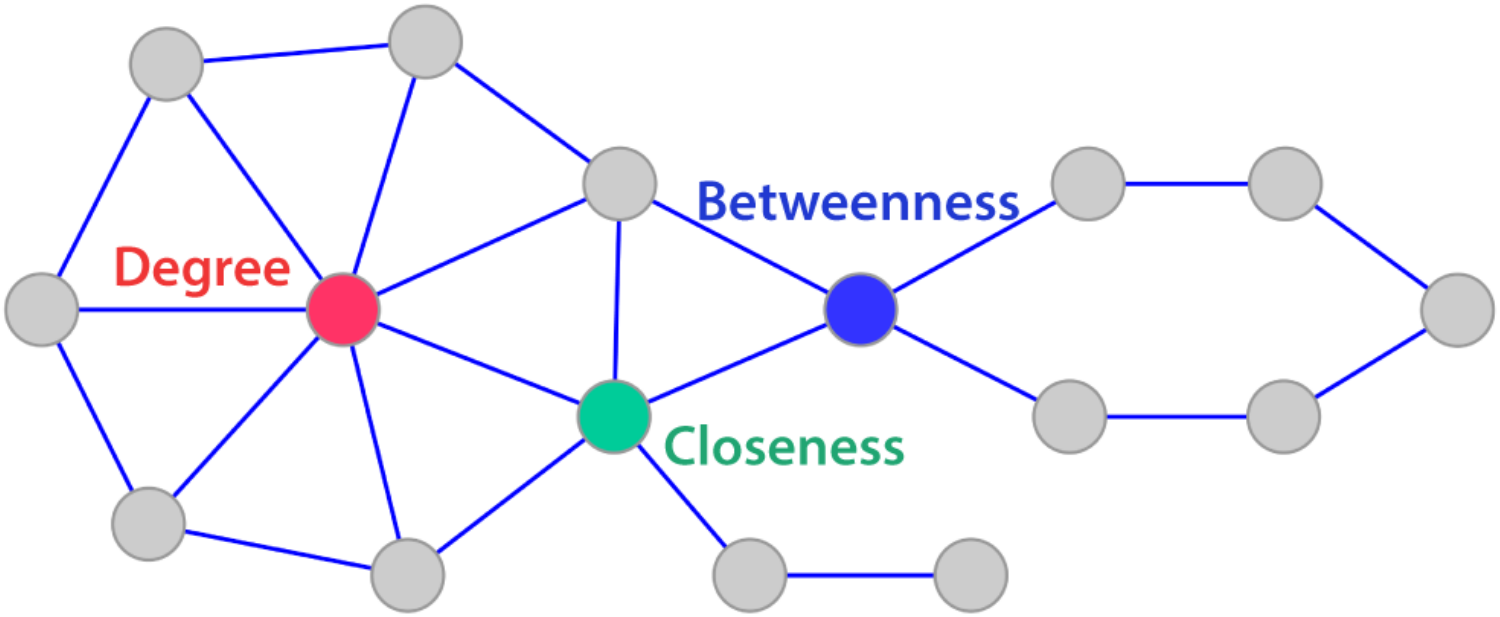
$\delta_{st}(n)$ = number of shortest path from s to t that go through n

- Betweenness = 0 no shortest paths go through this node
- Betweenness = 1 all shortest paths go through this node

Betweenness centrality

- Biological interpretation
  - Information load on a node
  - Control of the node over the connectivity of the network
  - Connection of two subnetworks
  - Weak links
  - Can be calculated for edges too

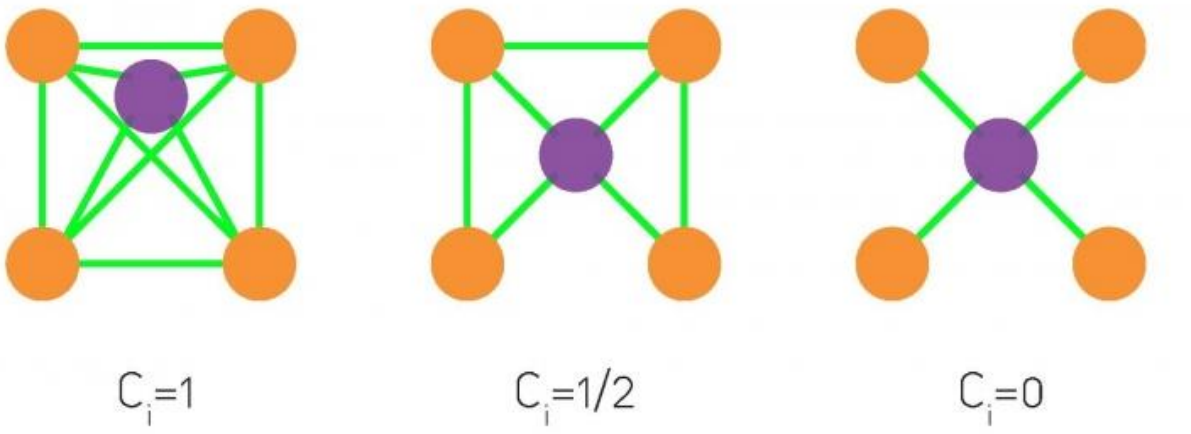
Betweenness centrality



Degree	ClosenessCentrality	BetweennessCentrality
7	0.45454545	0.29047619
5	0.51724138	0.42380952
4	0.48387097	0.4952381

Clustering coefficient

- Connectivity of the neighborhood - measure for the network's local edge density
  - How many of a nodes neighbors are connected to each other?

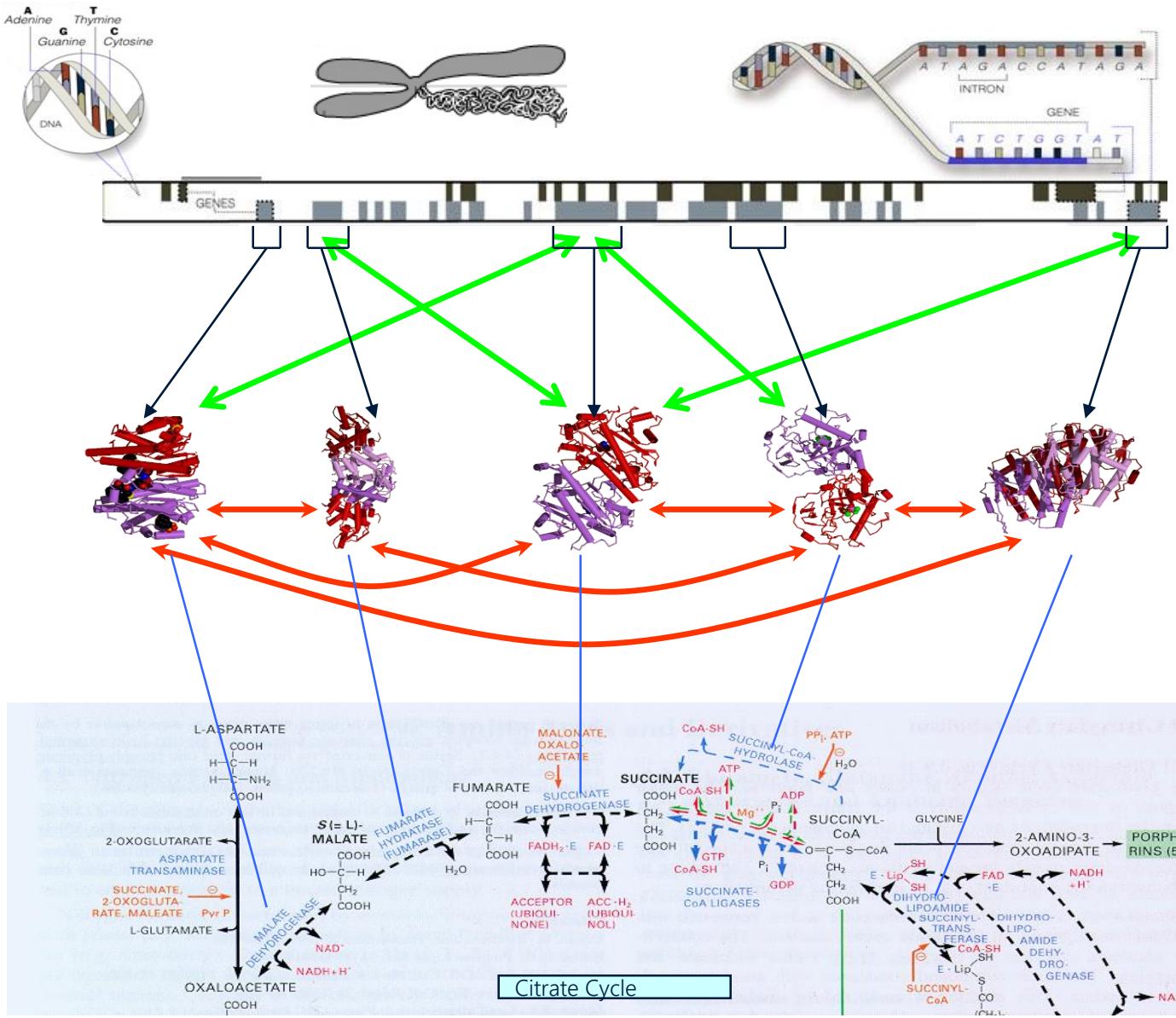


Where do I find *the* network?

- There is no such thing!
- >700 different interaction databases

www.pathguide.org

Molecular networks



GENOME

protein-gene interactions

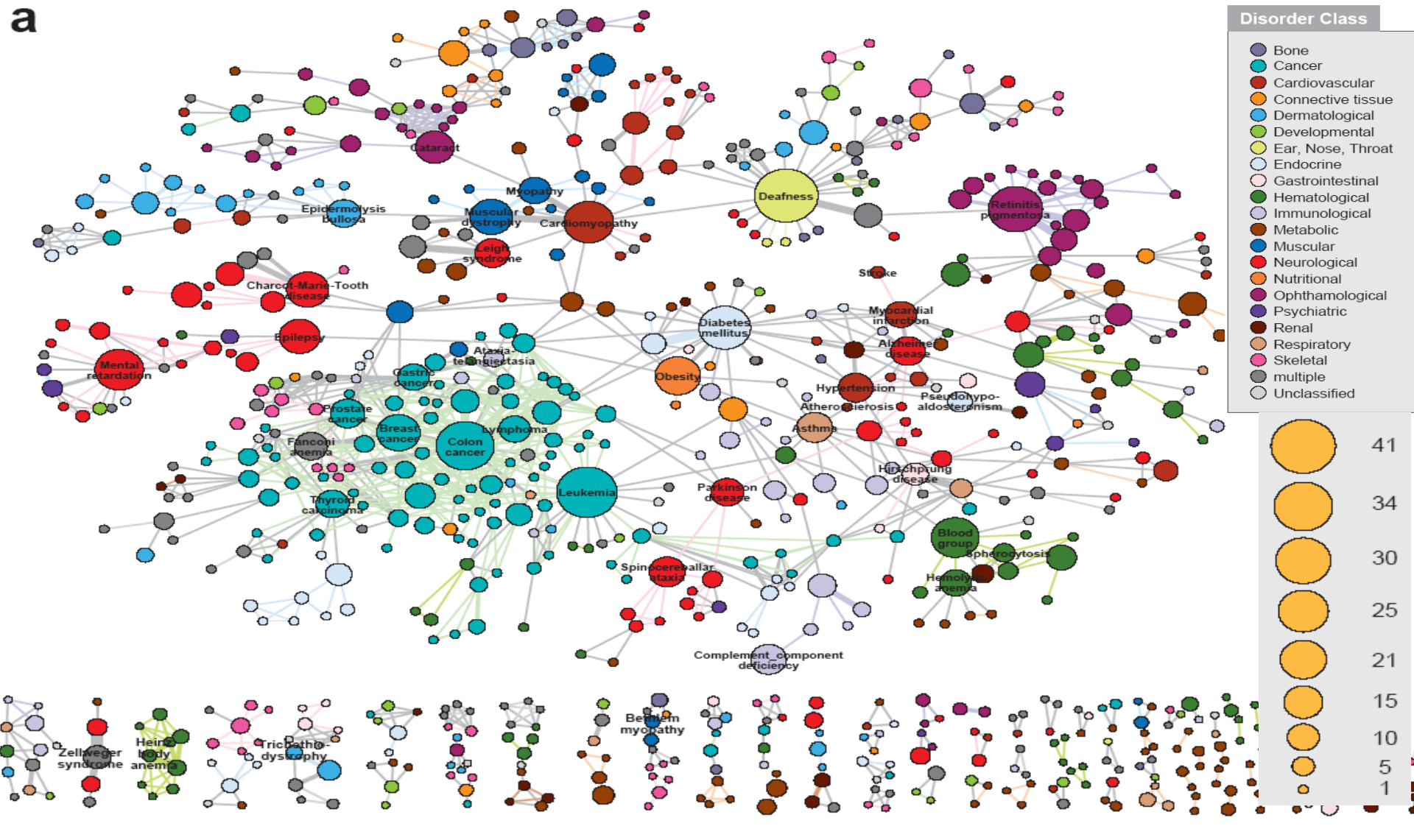
PROTEOME

protein-protein interactions

METABOLISM

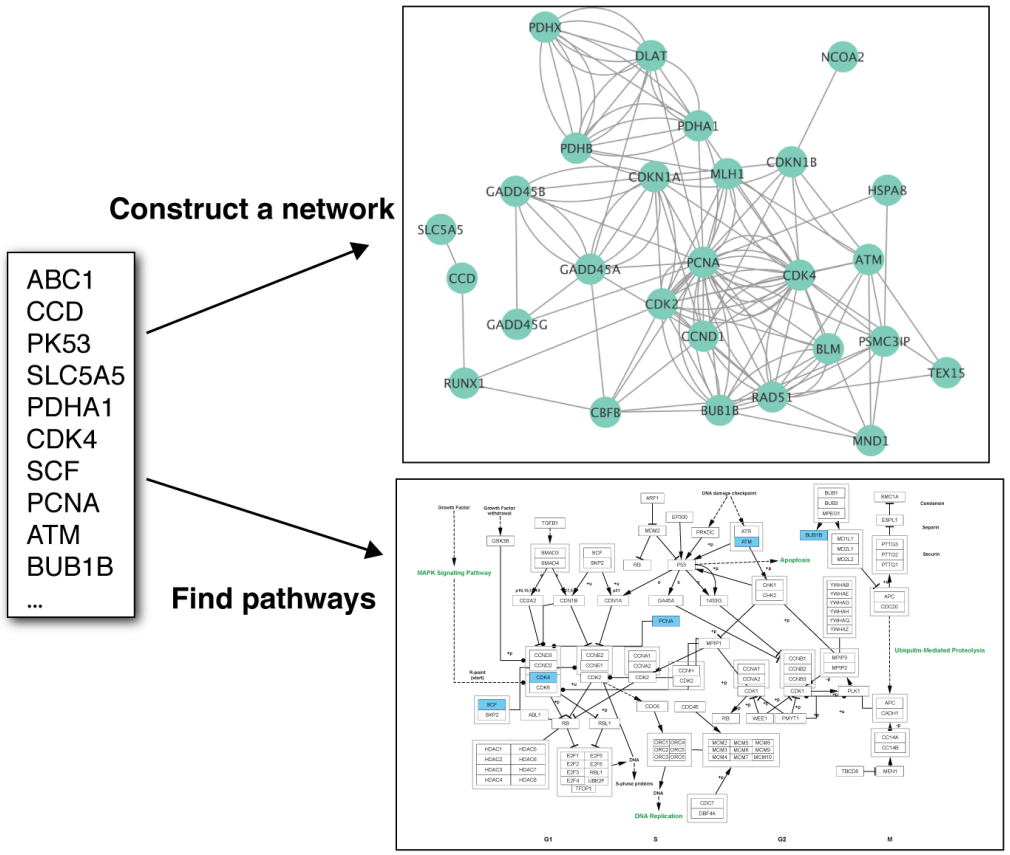
Bio-chemical reactions

Human disease network



Finding network sources

- Depends on biological question and analysis plan
- Typical start: **gene list**



Finding network sources

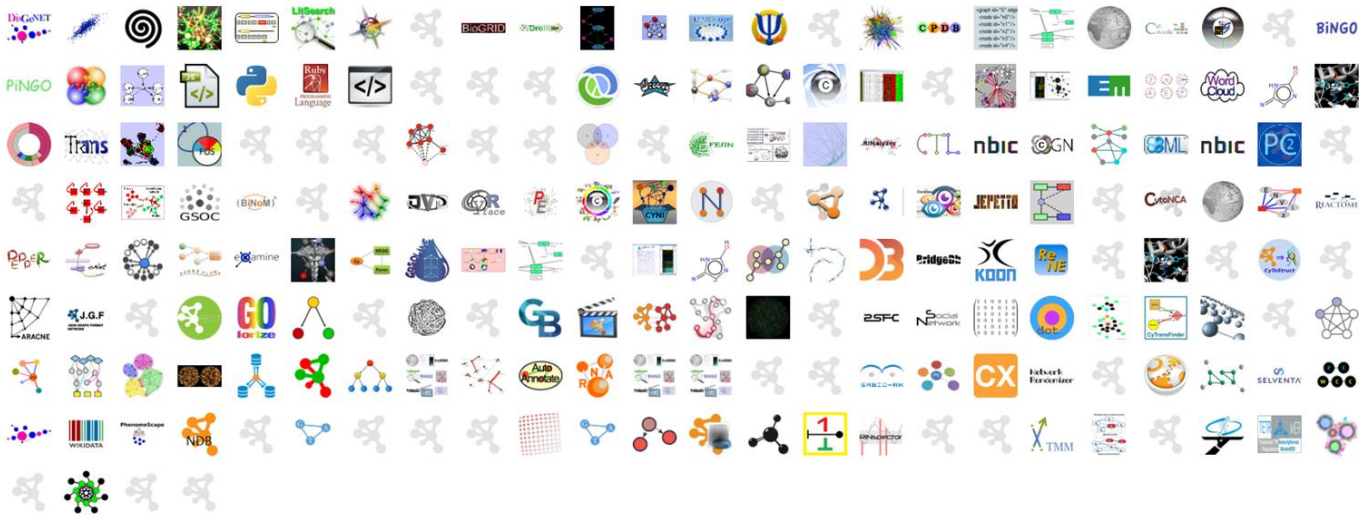
- Networks
  - Broad coverage / low resolution
  - Databases:
    - PSICQUIC, STRING, IntAct, GeneMANIA, NDEx, etc
  - Interaction types
    - Protein-protein interactions
    - Gene-regulatory interactions
    - Genetic interactions
    - Protein-compound interactions

Finding network sources

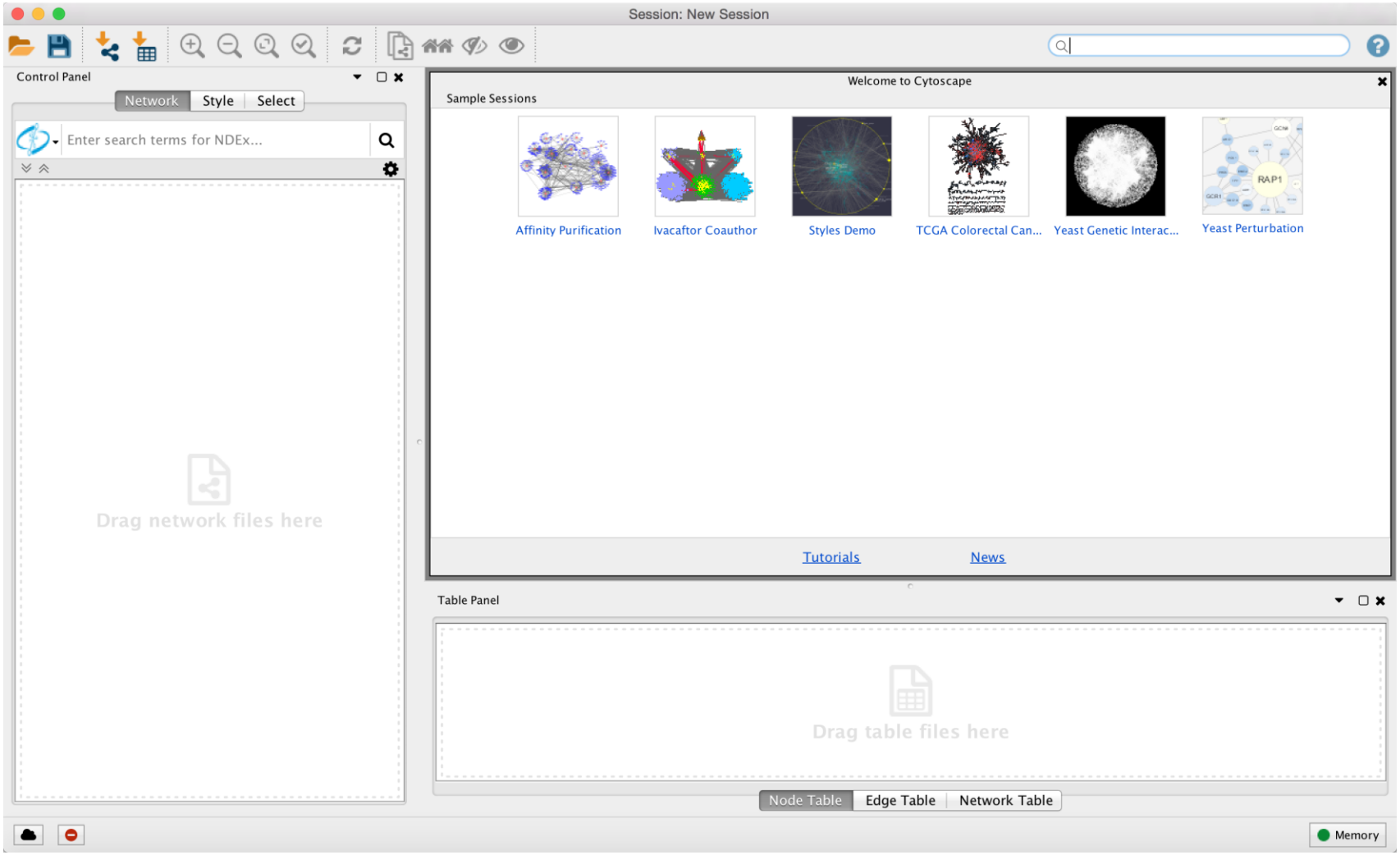
- **Pathways**
  - High resolution / limited coverage (~50% of genes)
  - **Databases:**
    - WikiPathways, Reactome, Pathway Commons, KEGG, etc
  - **Interaction types**
    - Signaling pathways
    - Metabolic pathways
    - Gene regulation pathways

Cytoscape

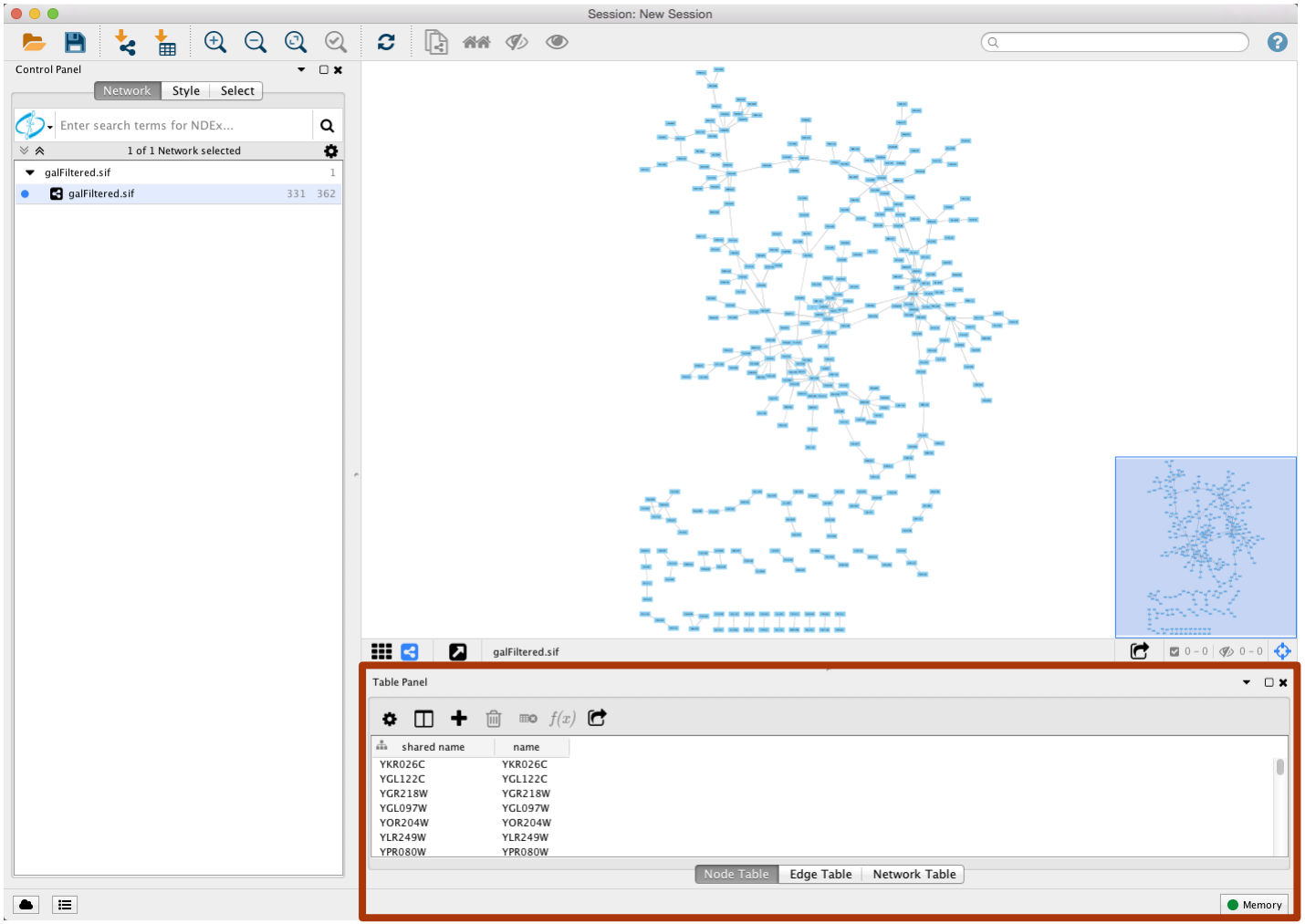
- Cytoscape (www.cytoscape.org) is a widely adopted network analysis and visualization toolbox
- Extendable with apps
  - 373 apps available (apps.cytoscape.org)



Cytoscape

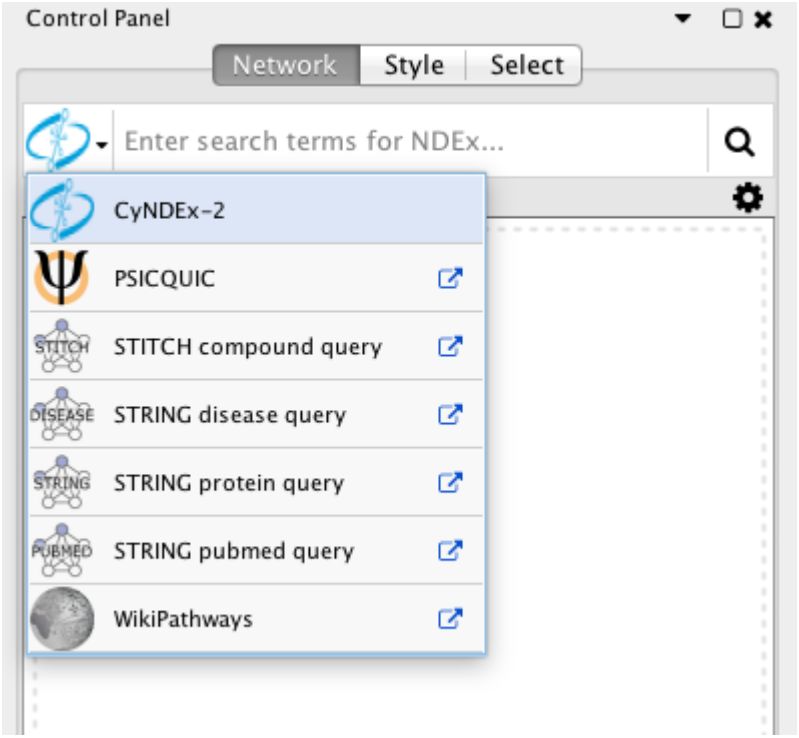

Session: New Session
Control Panel
Network
Style
Select
Enter search terms for NDEx...
Drag network files here
Welcome to Cytoscape
Sample Sessions
Affinity Purification
Ivacaftor Coauthor
Styles Demo
TCGA Colorectal Can...
Yeast Genetic Interac...
Yeast Perturbation
RAP1
Tutorials
News
Table Panel
Drag table files here
Node Table
Edge Table
Network Table
Memory

Cytoscape


Session: New Session
Control Panel
Network
Style
Select
Enter search terms for NDEx...
1 of 1 Network selected
galFiltered.sif
1
galFiltered.sif
331
362
galFiltered.sif
Table Panel
shared name
name
YKR026C
YKR026C
YGL122C
YGL122C
YGR218W
YGR218W
YGL097W
YGL097W
YOR204W
YOR204W
YLR249W
YLR249W
YPR080W
YPR080W
Node Table
Edge Table
Network Table
Memory

Cytoscape

- Import networks from public databases



Cytoscape

- There is a lot more functionality than what we will show you in the practical:
  - http://manual.cytoscape.org/en/stable/
  - Detailed documentation and examples

Practical computer session

Step 1: Investigate disease-associated proteins and their role as possible drug targets

- Create breast cancer specific protein-protein interaction network
- Analyze network and extend with drug-target information

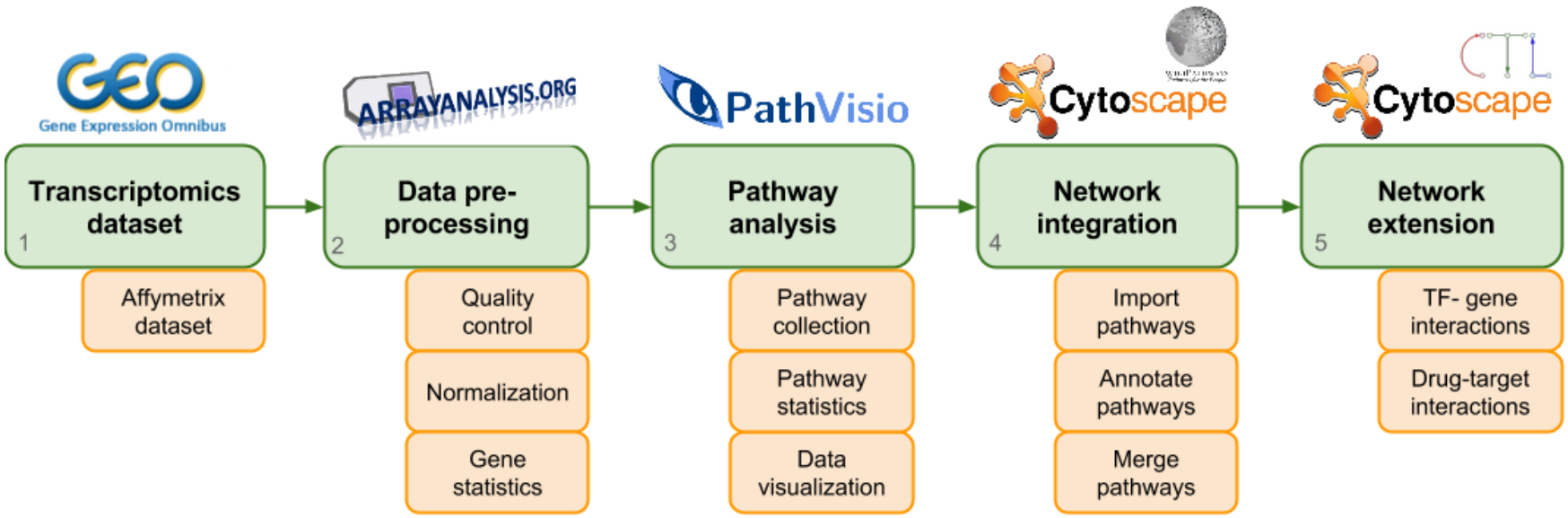
Step 2: Identify biological processes that are affected

- Perform pathway enrichment analysis
- Visualize data on biological pathways

Research examples

Images:
http://edulab.uoc.edu/wp-content/uploads/2017/11/sis-quantitative-market-research.jpg

Studying the diabetic liver



Kutmon et al. (2014) https://doi.org/10.1186/1471-2164-15-971

Studying the diabetic liver

Lean

N=5

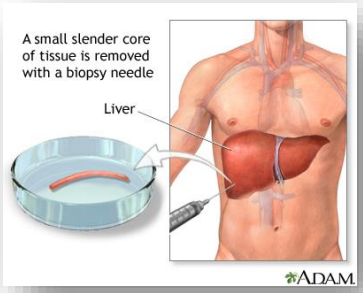
Elective cholecystectomy

Obese & T2DM

N=9

Gastric bypass

Open liver biopsie



* P < 0.05

	Lean	Obese & T2DM
Liver fat content (%)	2.4 ± 3.1	14.3 ± 9.3*

Isolation RNA

Affymetrix HG-U133A arrays

Gene expression

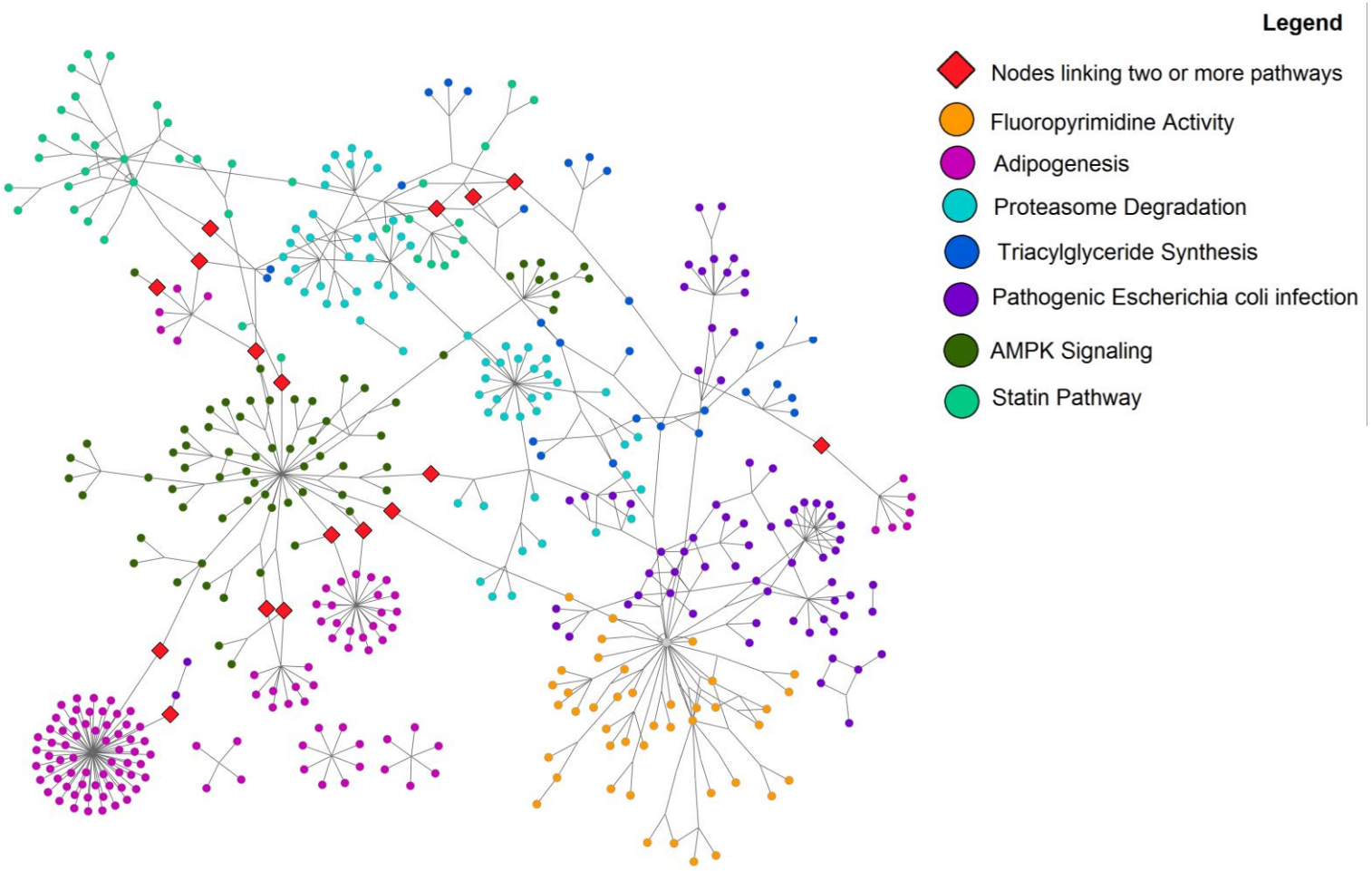
Publicly available at Gene Expression Omnibus (accession number GSE15653).

Pihlajamäki J et al. (2009) Thyroid hormone-related regulation of gene expression in human fatty liver. *J Clin Endocrinol Metab.* 94(9):3521-9. doi: 10.1210/jc.2009-0212.

Studying the diabetic liver

Pathway	Z-score	P-value	# genes	Differentially Expressed Genes
Triacylglyceride synthesis	3.78	0.001	3 (19)	↑ AGPAT2, GPD1, DGAT1
Proteasome degradation	3.32	0.006	5 (53)	↑ RPN1, PSMB3, HLA-B, HLA-E, HLA-J
Statin pathway	3.10	0.006	3 (25)	↑ DGAT1, APOA4, CYP7A1
Fluoropyrimidine activity	2.84	0.013	3 (28)	↑ SLC22A7 ↓ ABCG2, DPYD
Pathogenic E.coli infection	2.76	0.011	4 (46)	↑ ARPC1A, ARPC1B, ACTB ↓ ROCK1
Adipogenesis	2.41	0.016	7 (121)	↑ SREBF1, CDKN1A, NRIH3, PNPLA3, AGPAT2 ↓ CISD1, ZMPSTE24
AMPK signaling	2.38	0.029	4 (54)	↑ SREBF1, P21 ↓ LEPR, PFKFB3

Studying the diabetic liver

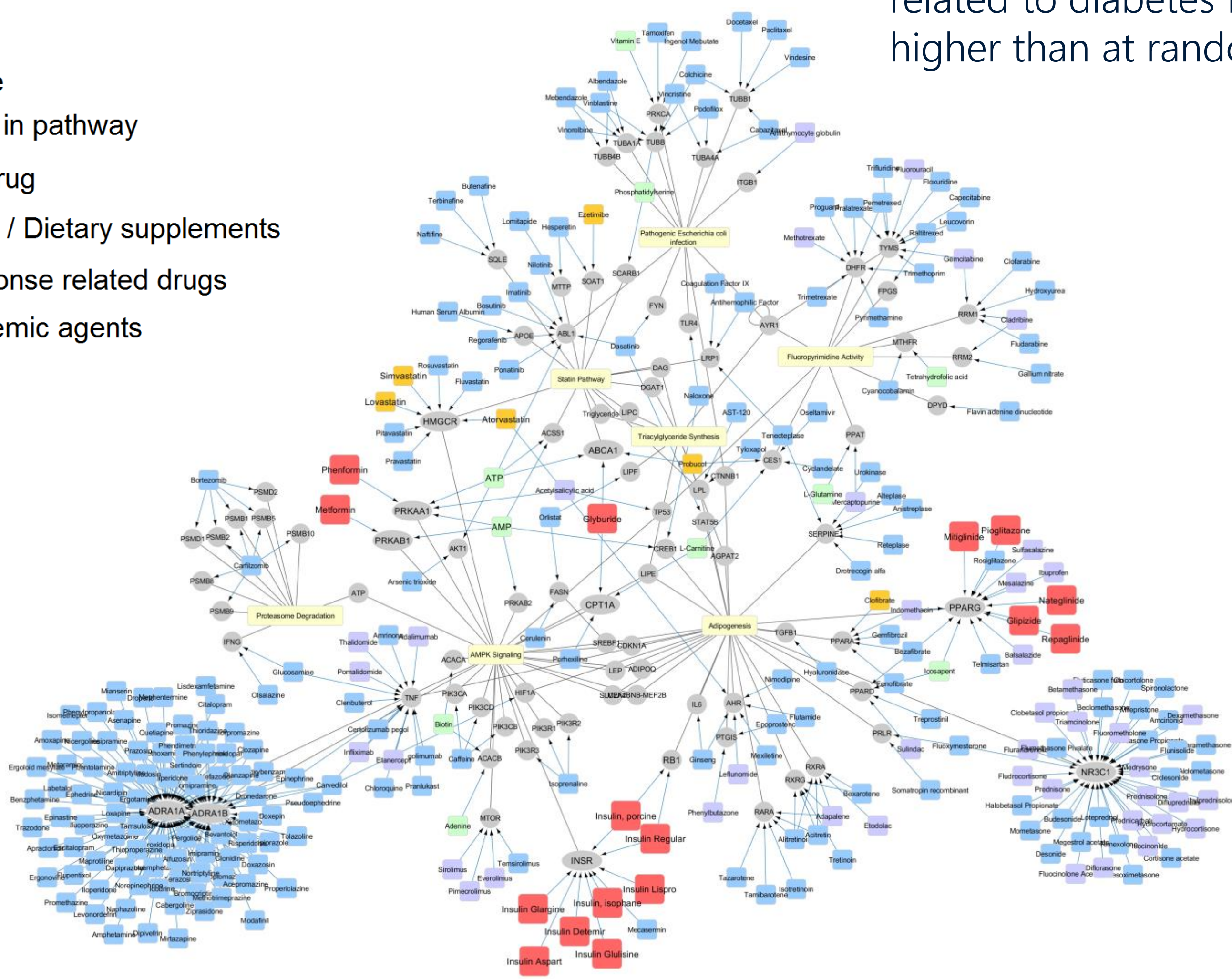


Studying the diabetic liver

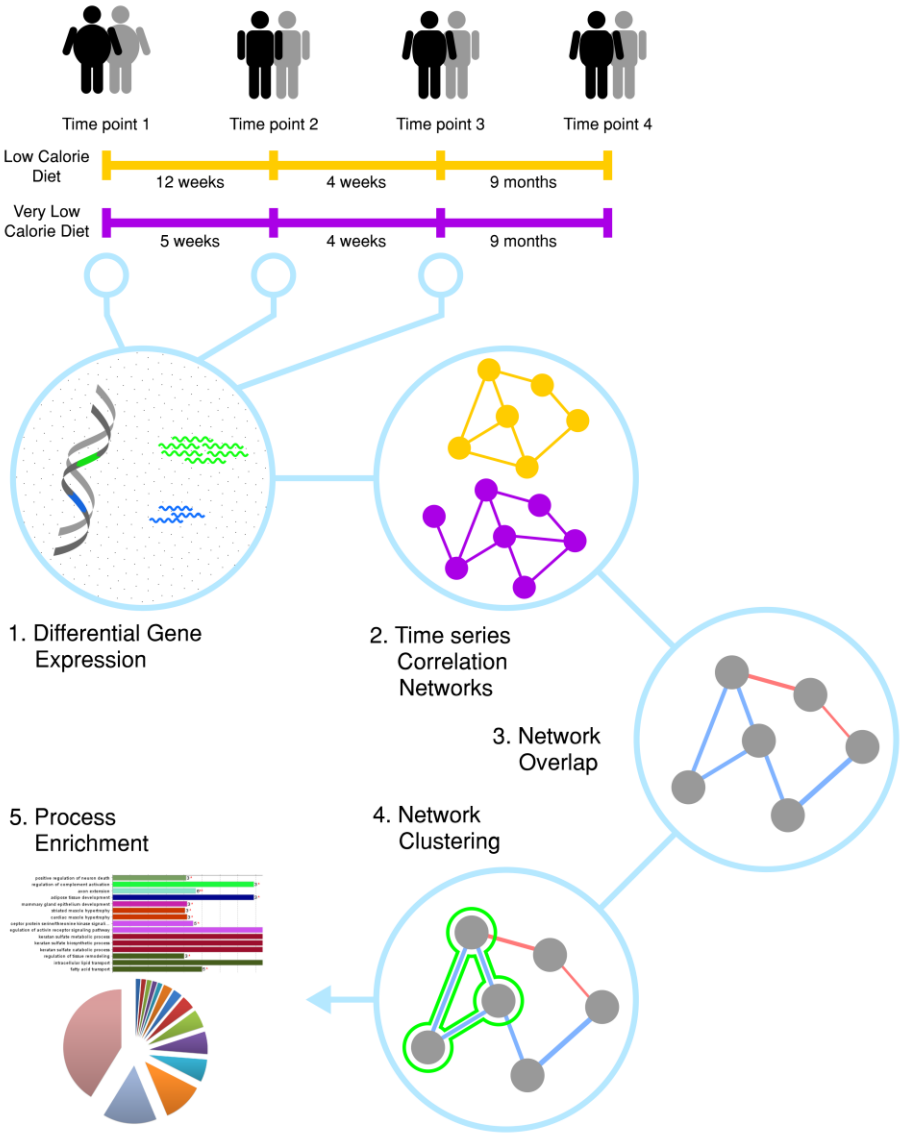
The amount of drugs related to diabetes is sign. higher than at random.

Legend

- Pathway node
- Gene product in pathway
- Antidiabetic drug
- Micronutrients / Dietary supplements
- Immune response related drugs
- Anticholesteremic agents
- Other drugs

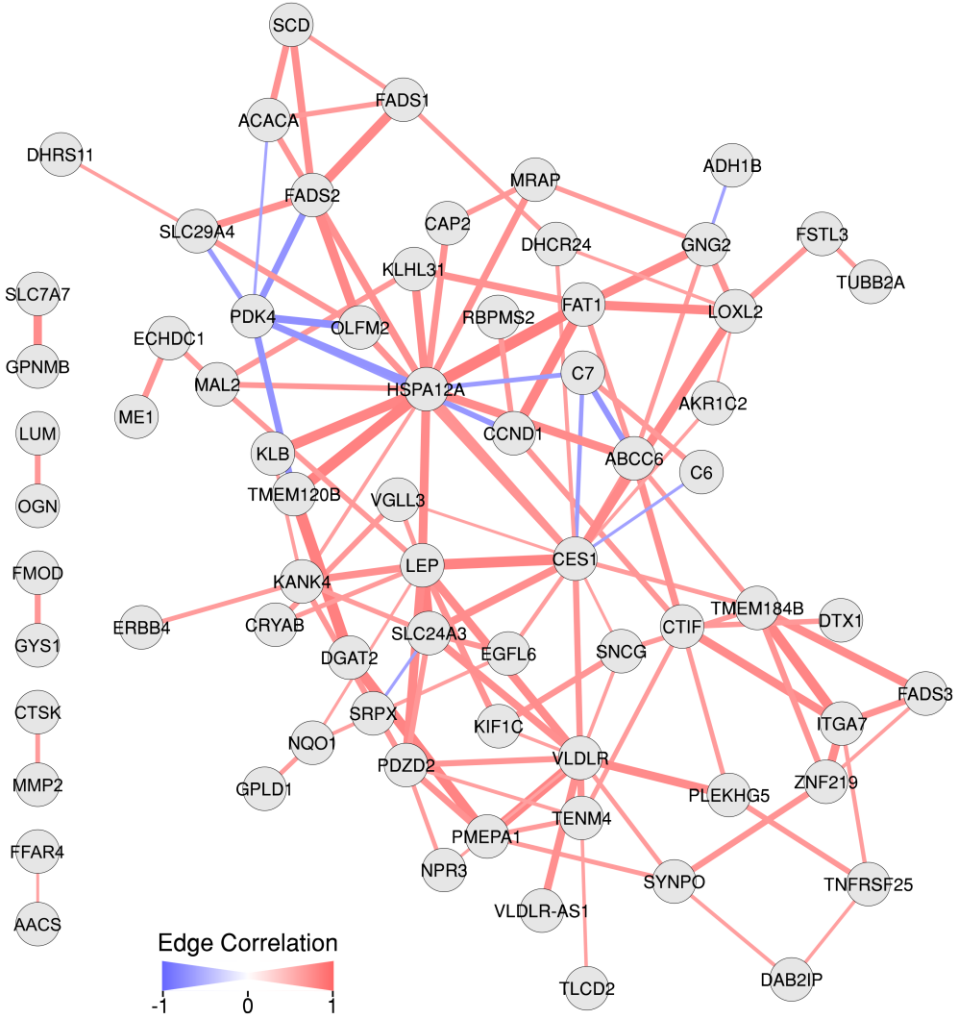


Diet study – network analysis

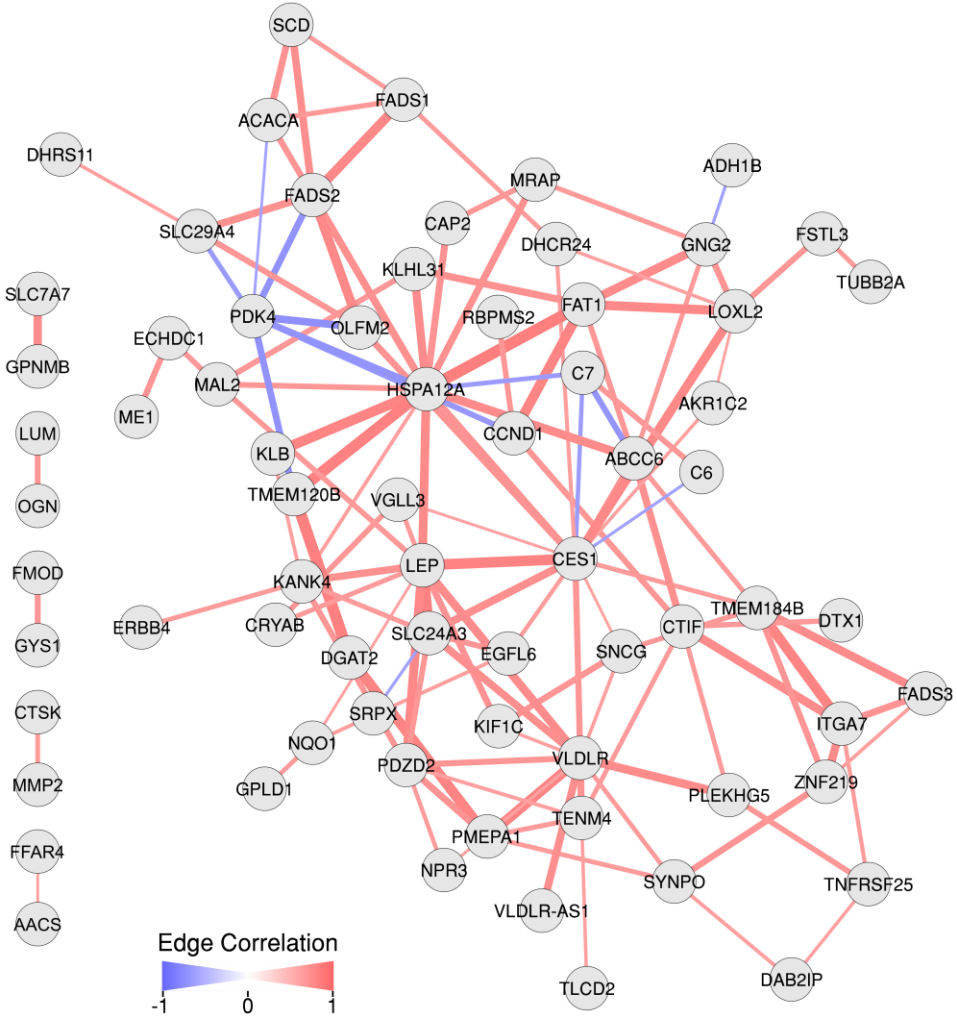


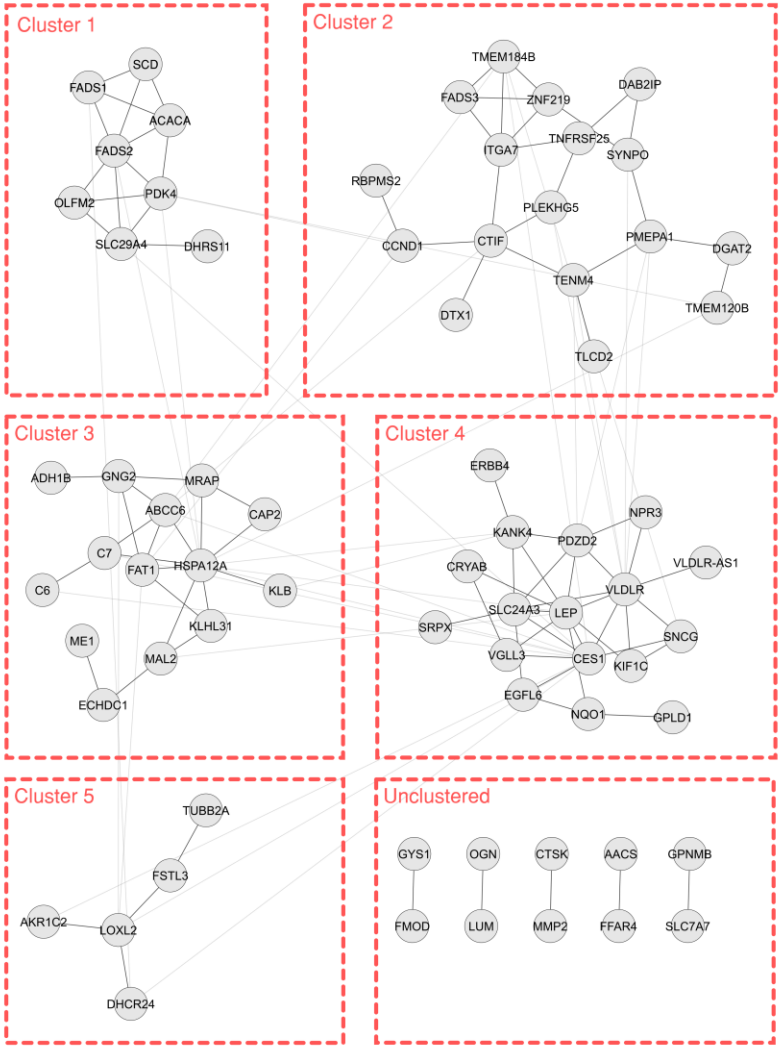
Tareen et al. (2020) https://doi.org/1038/s41598-020-58358-z

Diet study – network analysis



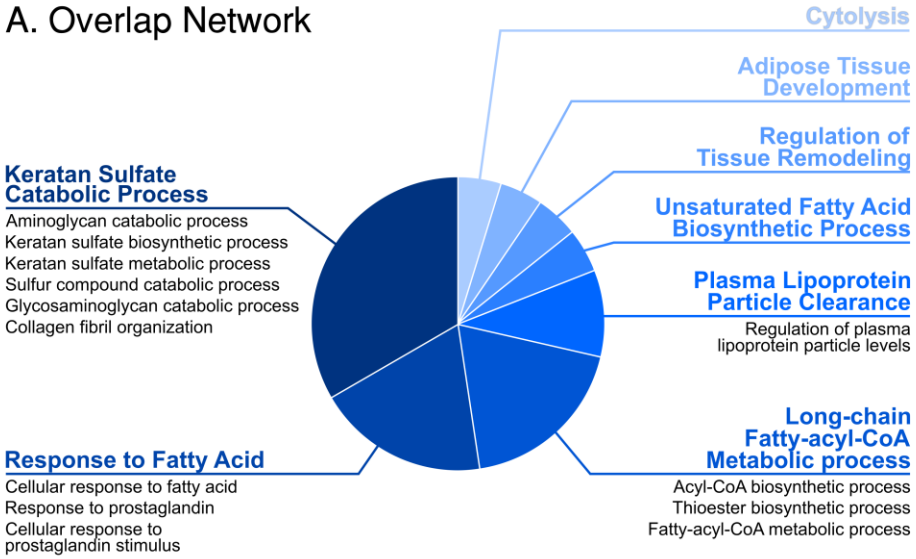
Diet study – network analysis



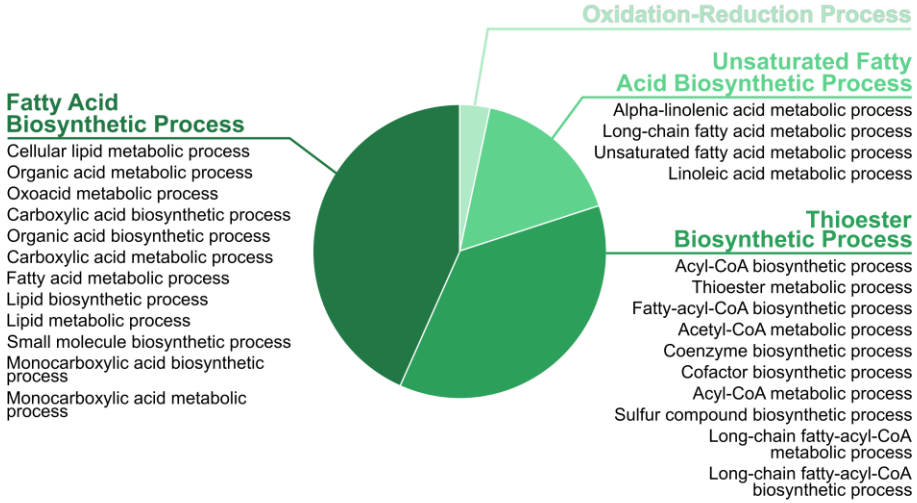


Diet study – network analysis

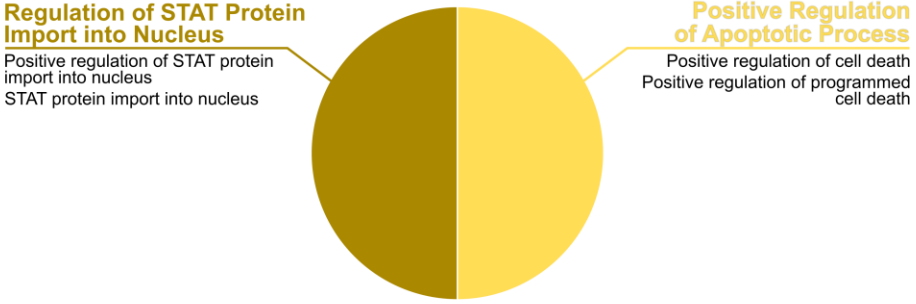
A. Overlap Network



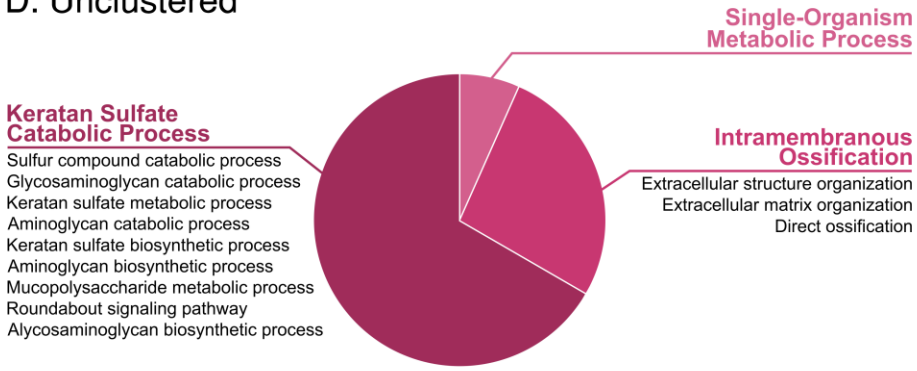
B. Cluster 1



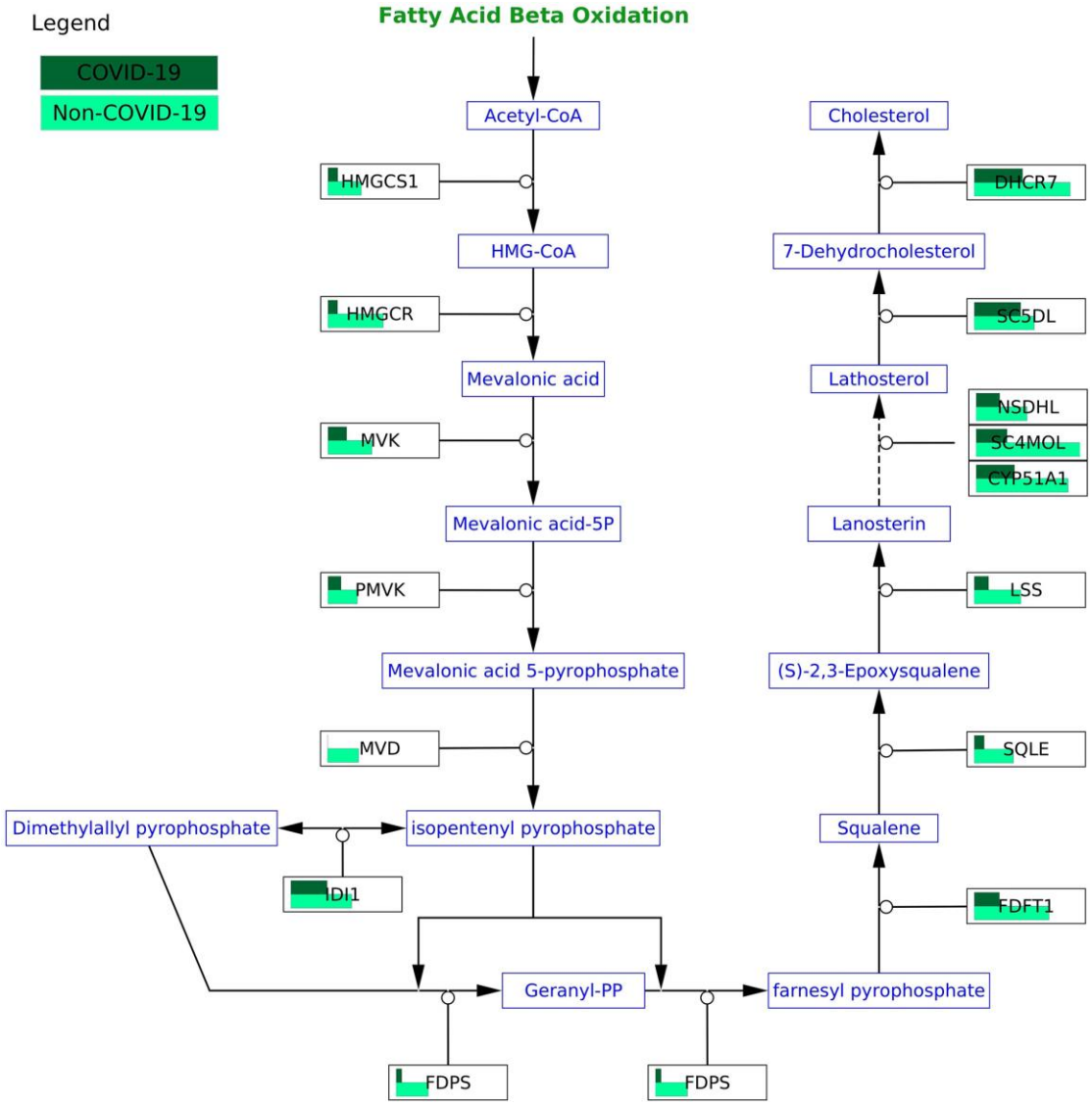
C. Cluster 4



D. Unclustered



Tissue-specific proteomics in COVID-19



Pham et al. (2022) https://doi.org/10.3389/fimmu.2022.963357

BTR internships

- Immunometabolism
- Pancreatic cancer / COVID19
- Chronic diseases

- Pathway curation
- Pathway analysis
- Network analysis
- Graph theory
- Network visualization

Feel free to contact me if you are interested!

Questions

Maastricht University

Martina Summer-Kutmon, PhD
email: martina.kutmon@maastrichtuniversity.nl

Images:
http://pngimg.com/uploads/question_mark/question_mark_PNG56.png