CZI Omics Integration
18 March 2022

Application of integrated network analyses in disease characterization

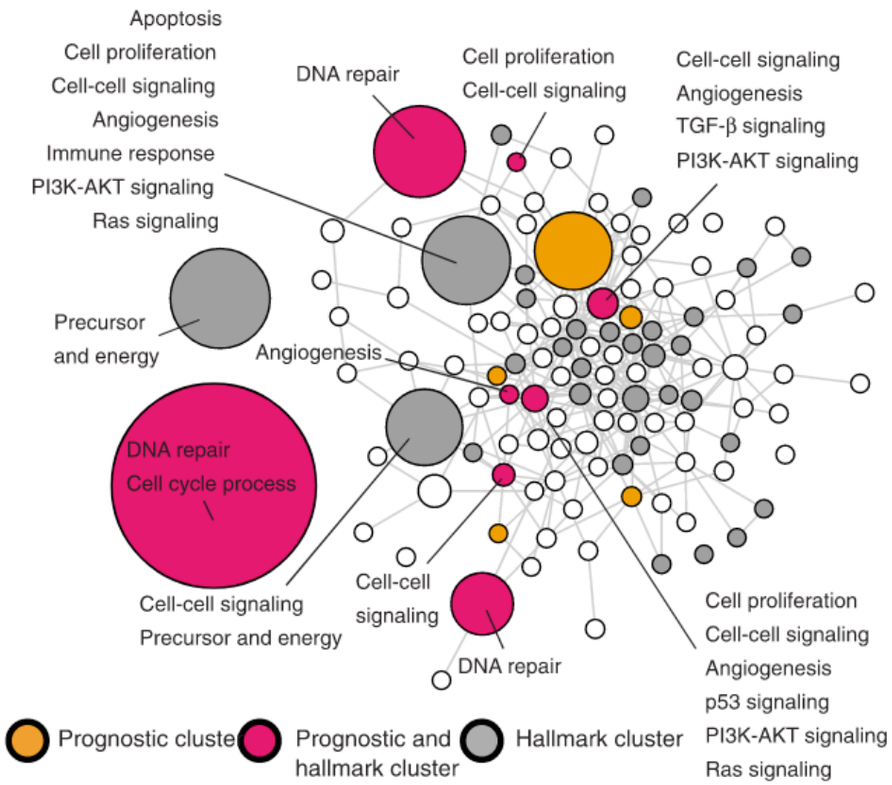
Rui Benfeitas

NBIS - National Bioinformatics Infrastructure Sweden
Science for Life Laboratory, Stockholm

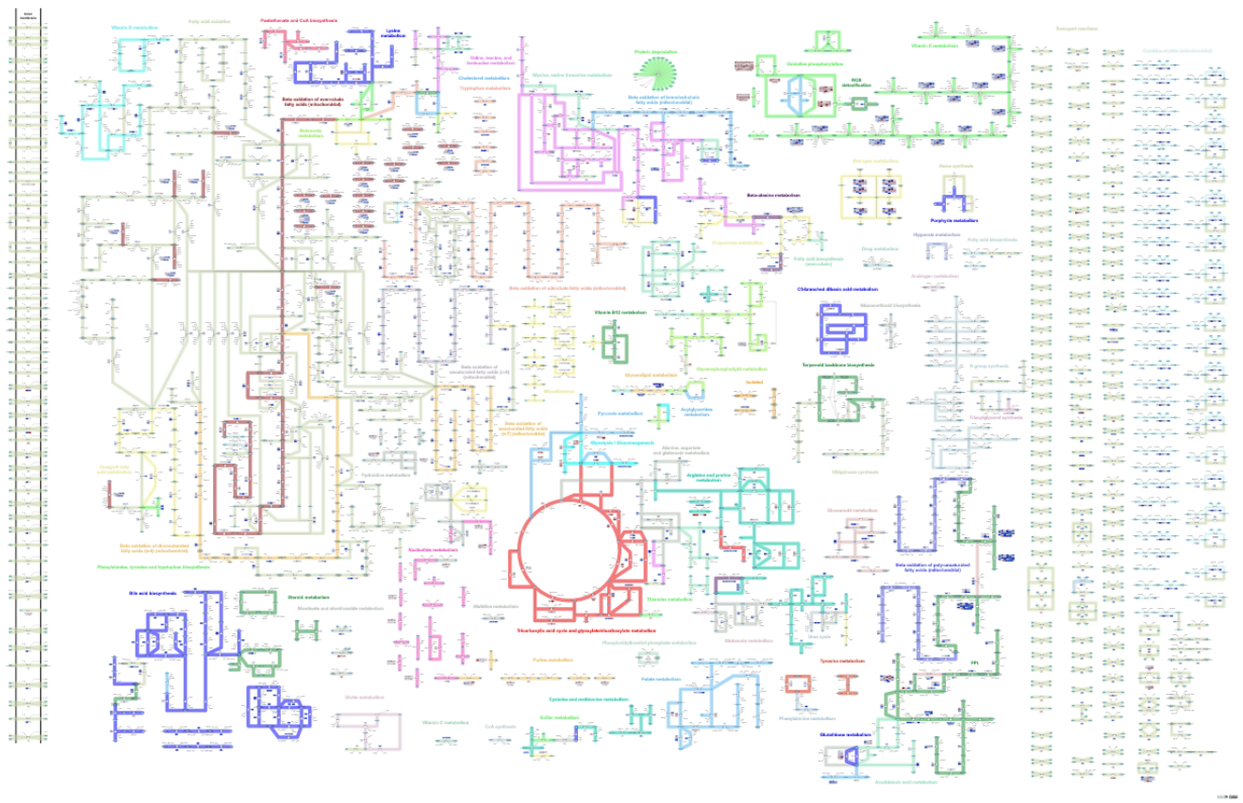
rui.benfeitas@scilifelab.se

Frameworks for biological network analysis in health and disease

Introduction to application of graph analysis in disease



Genome-scale metabolic modeling for data integration and simulation



Uhlen 2017
https://metabolicatlas.org/

Companion resources

Pre-course information and installation instructions

Data pre-processing notebook: link

Integrated network analysis notebook: link

Rui Benfeitas, Scilifelab, NBIS National Bioinformatics Infrastructure Sweden

rui.benfeitas@scilifelab.se

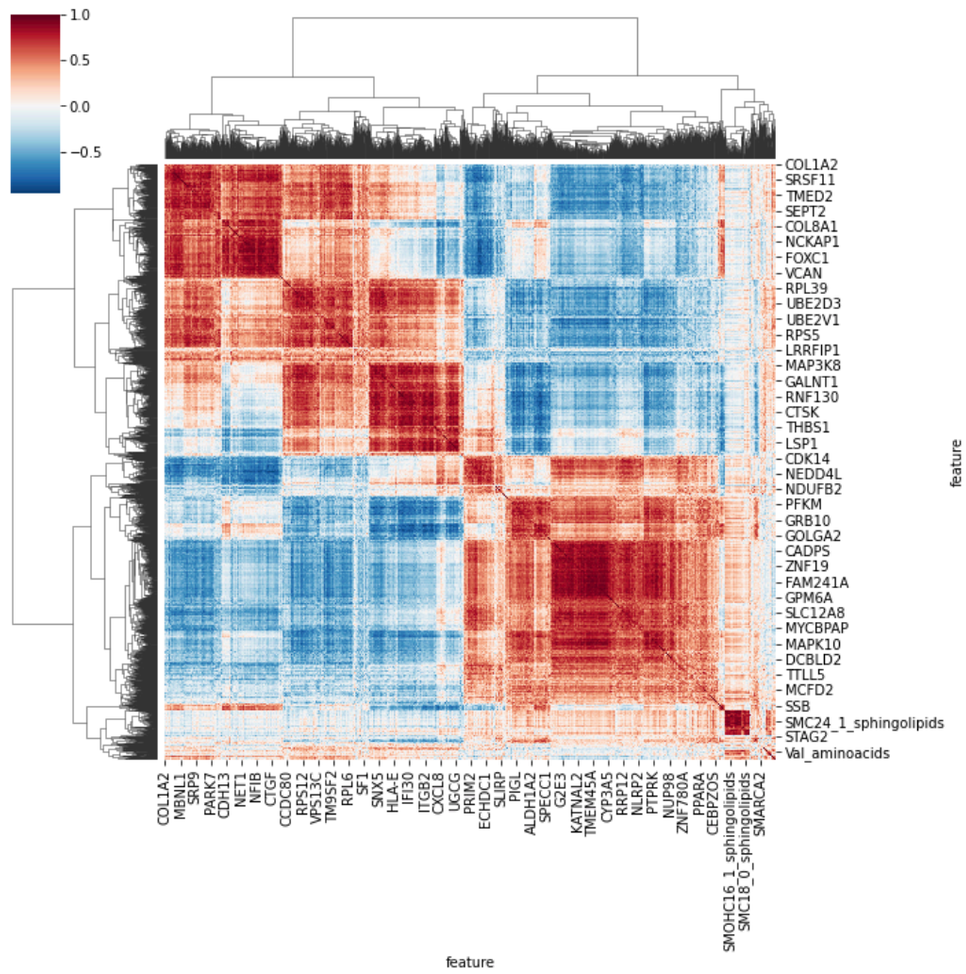
Abstract

In this notebook we will examine several properties in preparing our data for integration. However, it should be noted that a correct preparation of your dataset will depend on your data's technology among other factors.
We start from an overview of our dataset, followed by removing redundant or uninformative features. We then tackle different methods of data imputation, outlier detection, and a quick data profiling of feature behavior and quality. Finaly, we look into data transformation, including rescaling, normalization and batch correction.

This notebook should be seen as a reminder to several factors that we need to consider before downstream analyses, and there are any alternative approaches to tackle a specific problem such as missingness or outlier detection. As such, this guide should **not** be seen as an exhaustive benchmarking notebook, **nor** should it be taken as replacement for dedicated QC and pre-processing methods or pipelines. For more information, refer to NBIS workshops or Scilifelab courses.

Preamble

Before starting, it is crucial that we have as much information about our dataset as possible, and that we have a good idea of the analyses that we will perform afterwards. Some of the questions to bear in mind before starting:



Outline

1. Networks as frameworks for phenotypic characterization

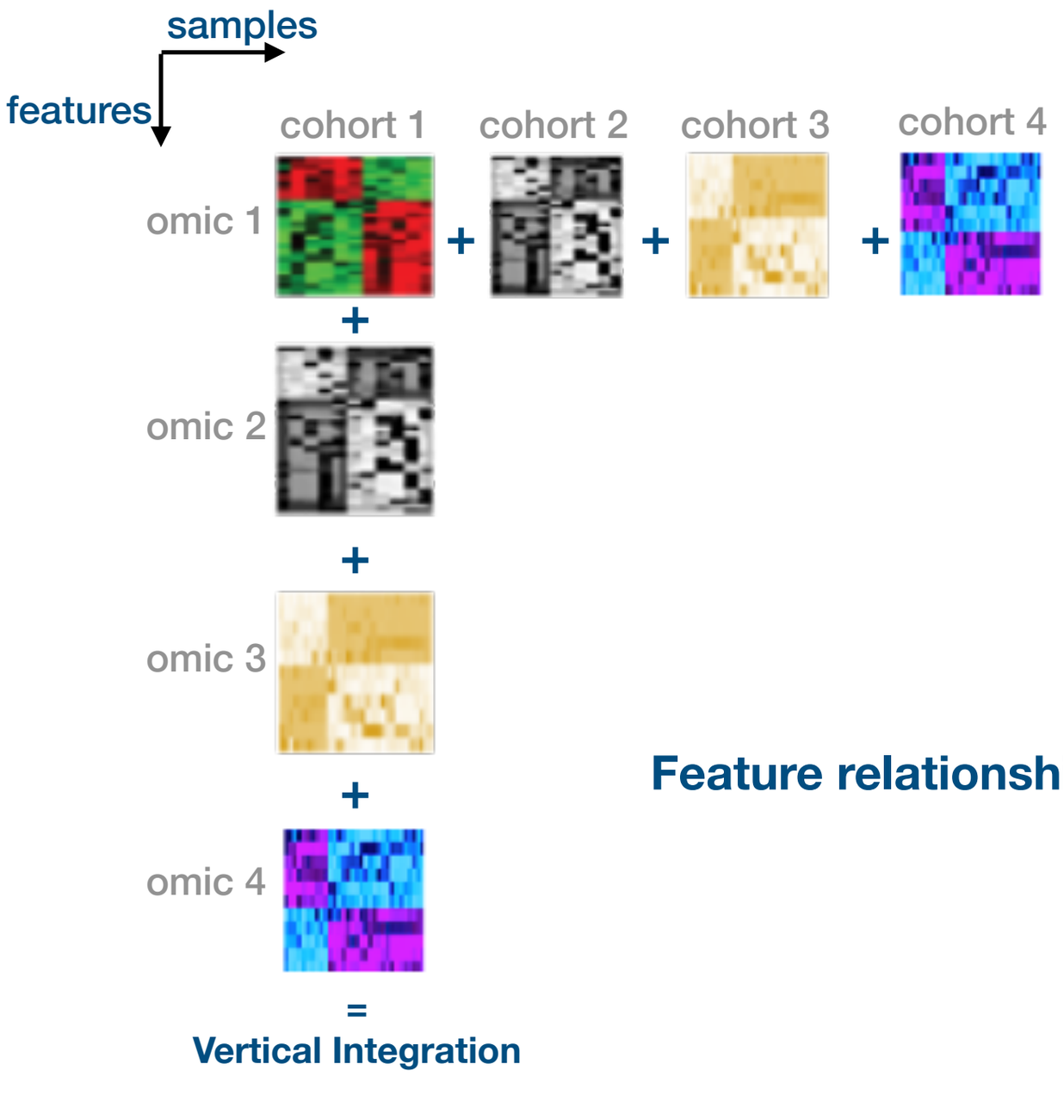
2. Network inference

3. Key properties in biological network characterisation

4. Communities and functional analysis

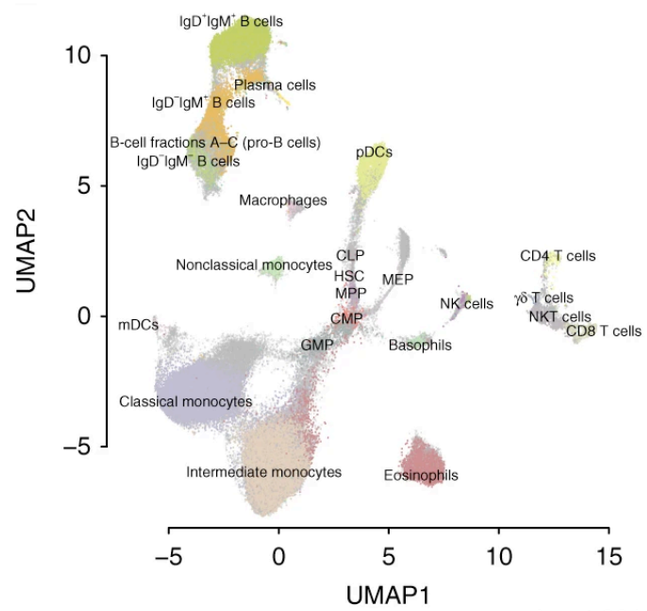
Original sources of images provided as reference and hyperlinks where applicable.

Techniques may be employed at feature and sample



Feature relationships

Application at sample and cell level



Becht 2018

Network formalisms balance parameter number and size

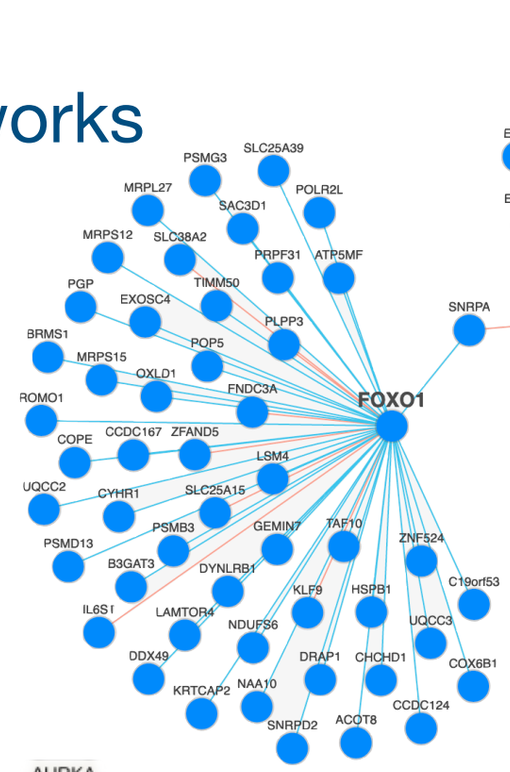
	Pros	Cons
Kinetic models	Detailed Quantitative Dynamic / Steady state	Small Requires detailed parameterization
Stoichiometric **GEMs**	Large Semi-quantitative Steady state	Static
Topological **Graphs**	Comprehensive No extensive parameterisation required	Static

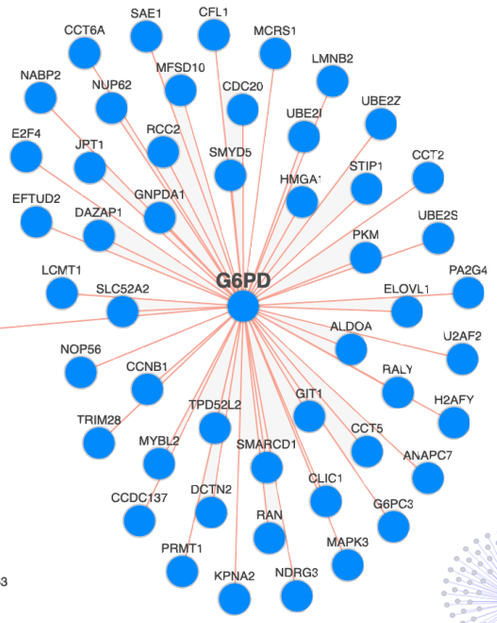
Adapted from Hartmann et al.

What are biological networks?

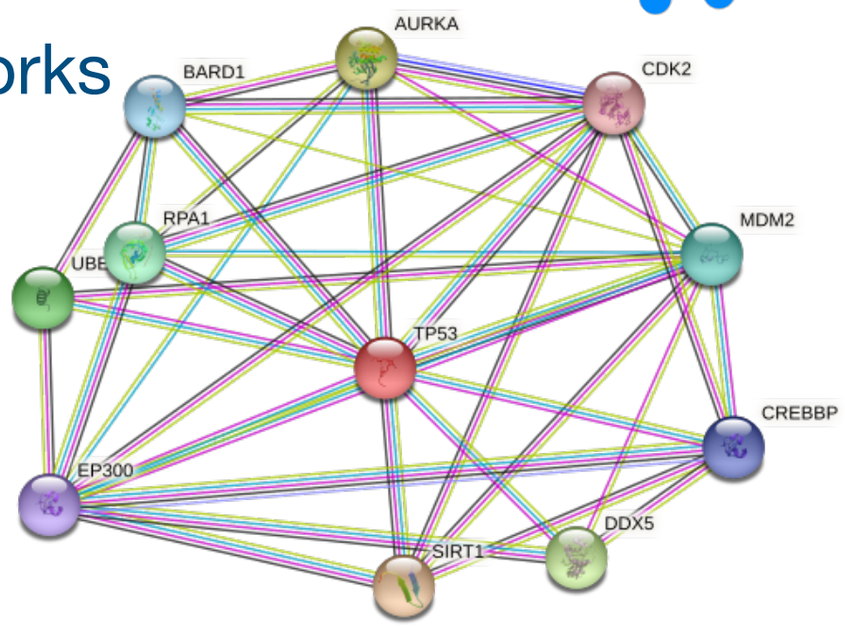
Comprehensive adimensional representation of feature associations

Co-expression networks

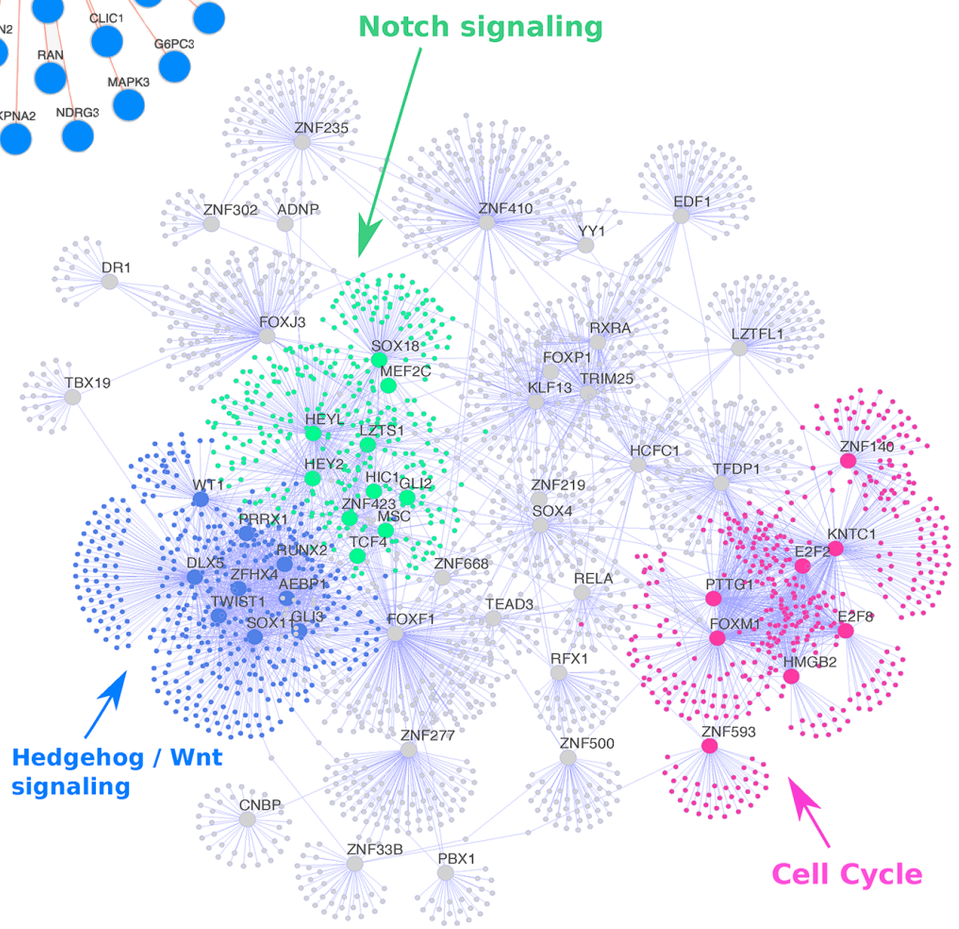




PPI networks



TF regulatory networks



Sivakumar 2017
inetmodels.com
string-db.org

What are biological networks?

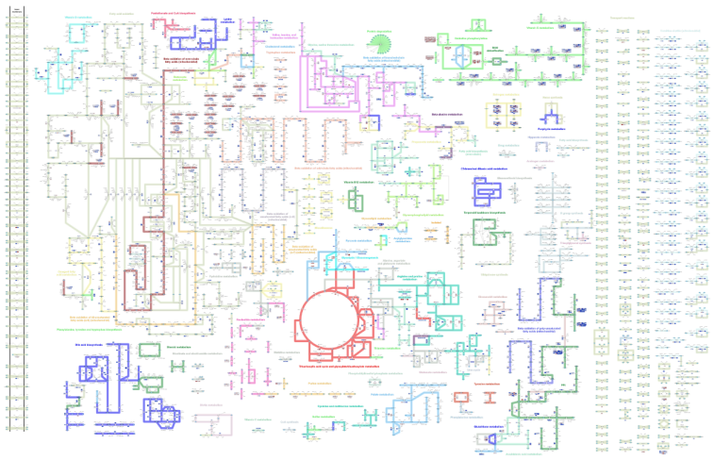
Metabolite - Reaction - Genes (GEMs)

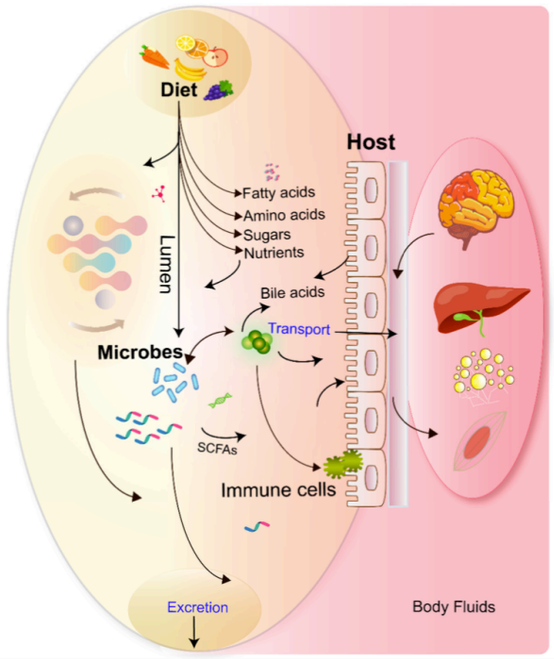
Multi-tissue networks

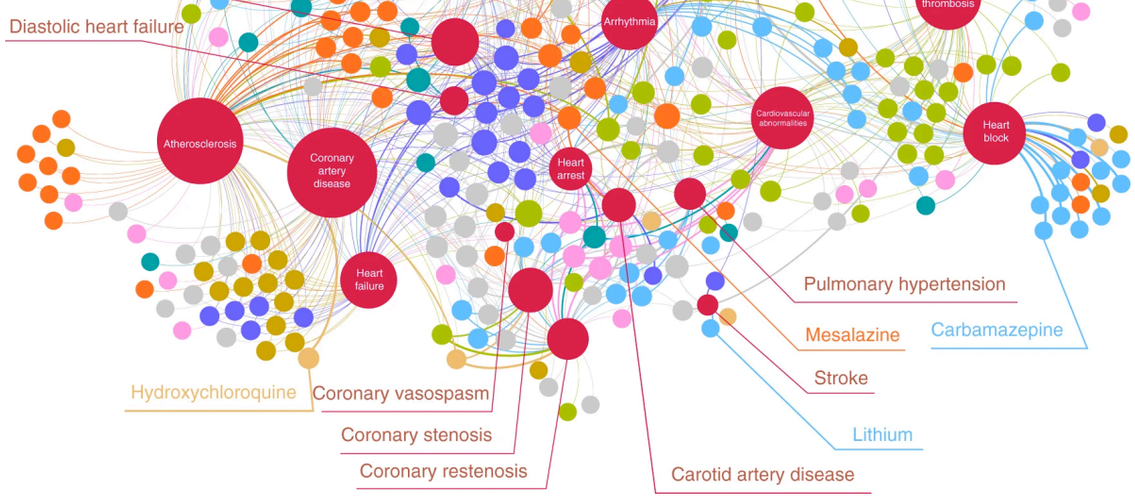
Multi-species networks

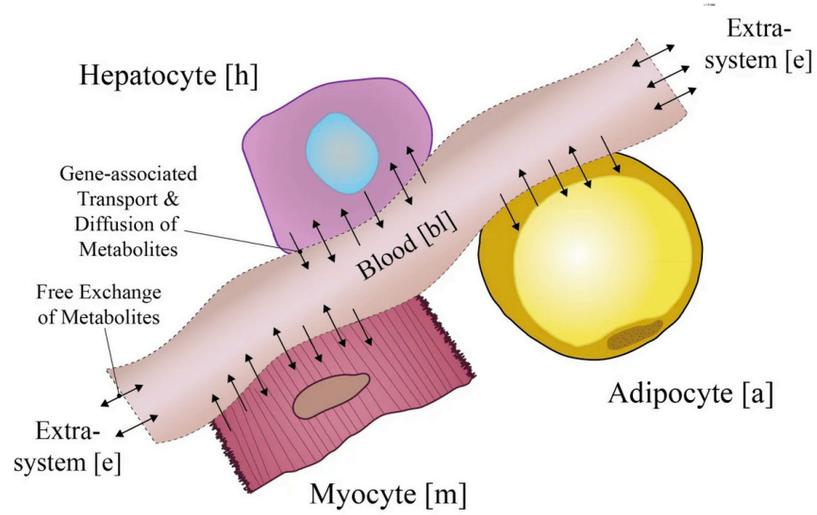
Drug-Disease networks

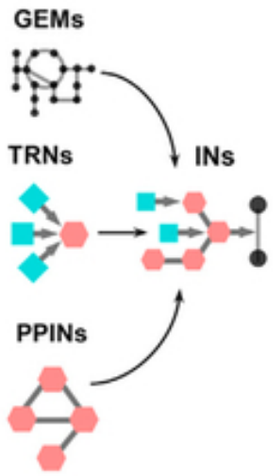
Integrated networks





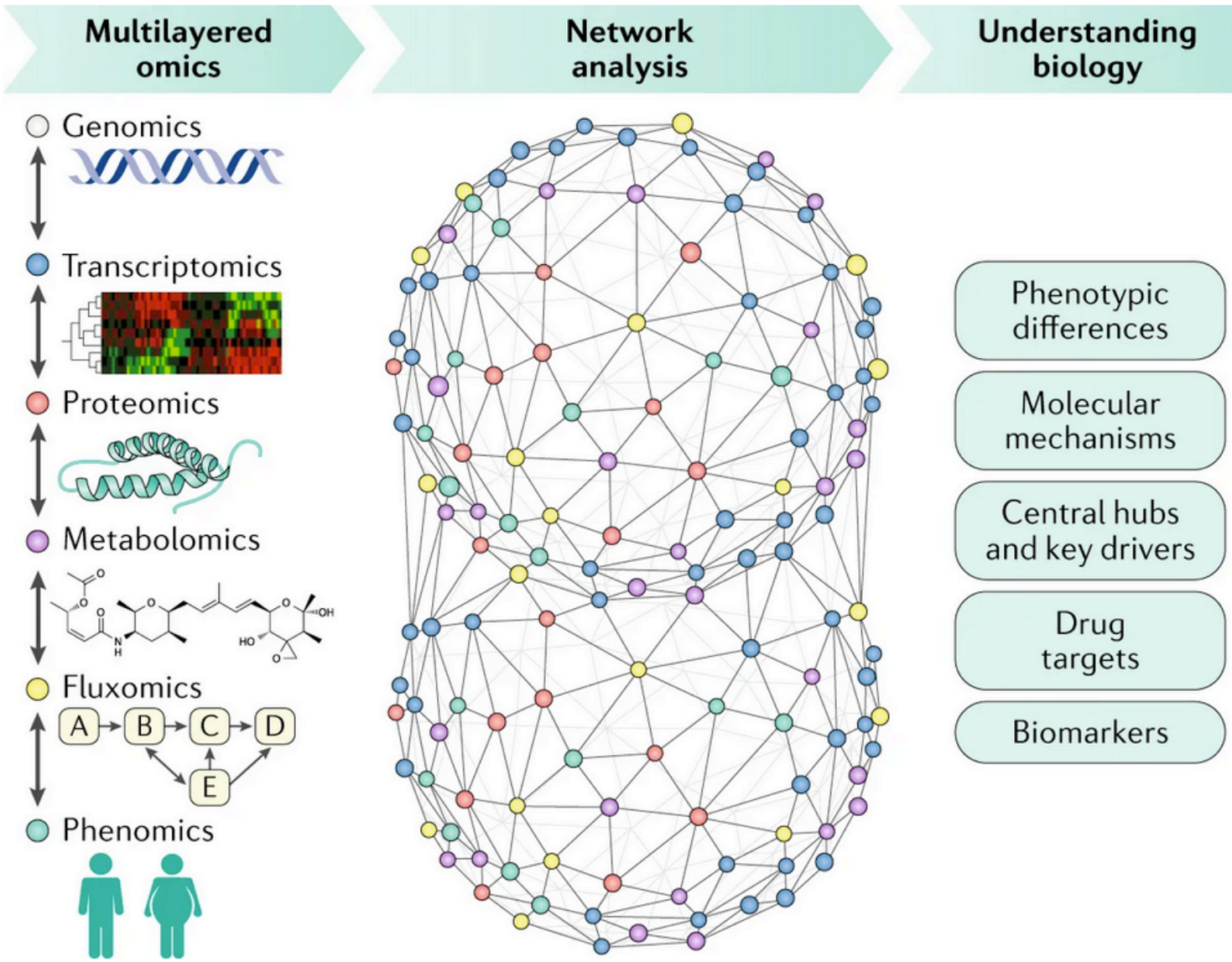






https://metabolicatlas.org/
Bordbar et al 2011
Sen & Oresic 2019
Cheng 2018
Lee et al 2016

Network as integrative frameworks



Systematically examine coordinated patterns

Phenotypic subgroup specificity

Feature prioritisation and target prediction

Pathway dysregulation

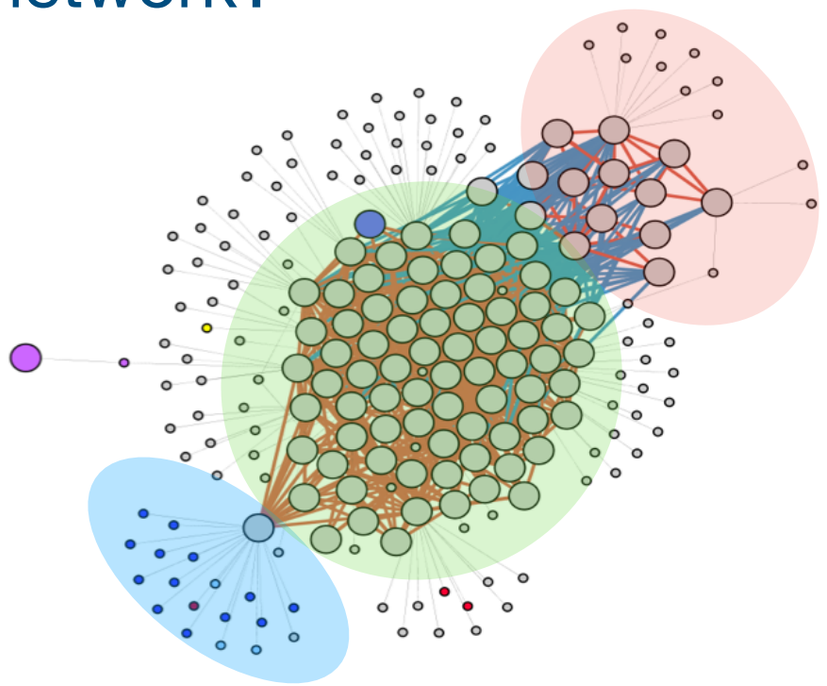
In silico simulation

Mardinoglu 2018

Analysis of network topology for disease characterization

Examples:

- What features are associated with my gene of interest?
- What are the feature communities in my network?
- What possible functional relationships do they share?
- What are the key elements in a community?
- How connected is a feature to the rest of the network?



Outline

1. Networks as frameworks for phenotypic characterization

2. Network inference

3. Key properties in biological network characterisation

4. Communities and functional analysis

Original sources of images provided as reference and hyperlinks, where applicable.

Networks may be inferred based on interaction evidence from various resources

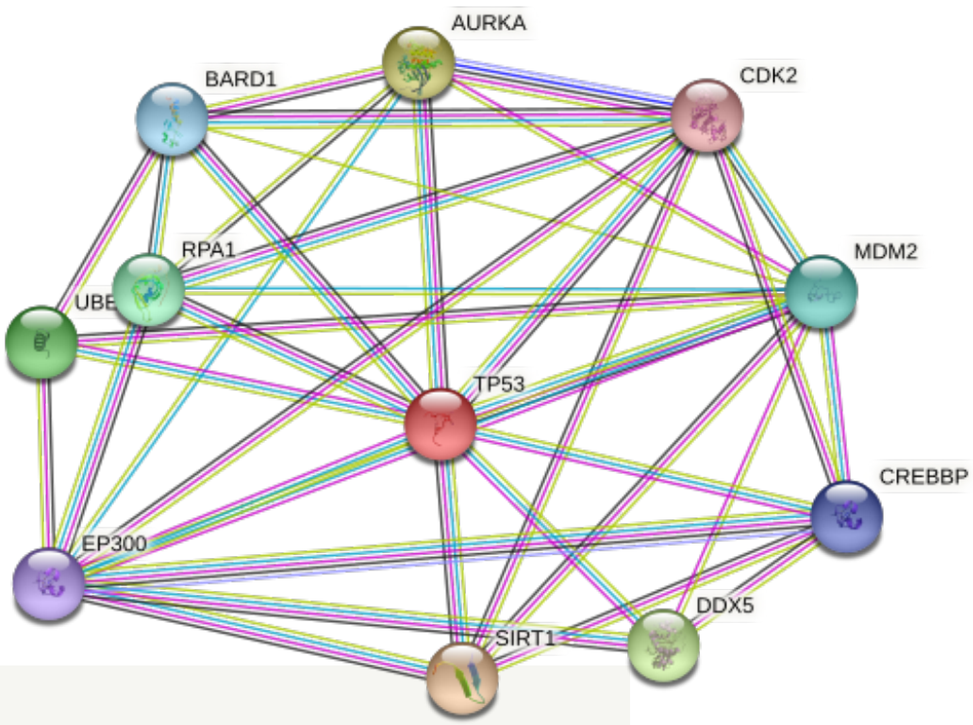
Any distance / association matrix may be translated to a network format

Limitations:

- Some of the functional analyses depend on annotation

- Sample size / Effect size / False discovery

TP53 PPI network in human



Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

Edges:

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

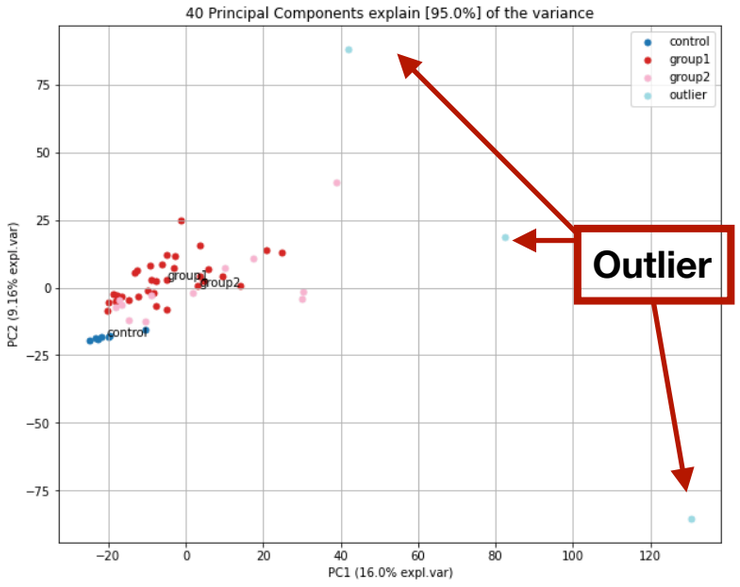
textmining

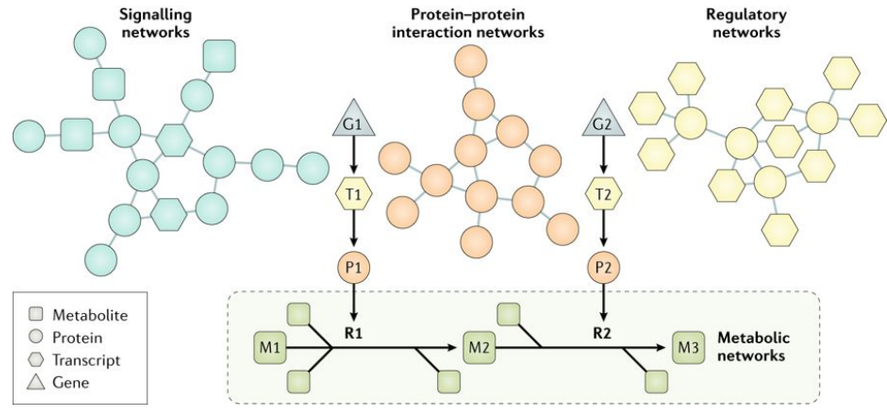
co-expression

protein homology

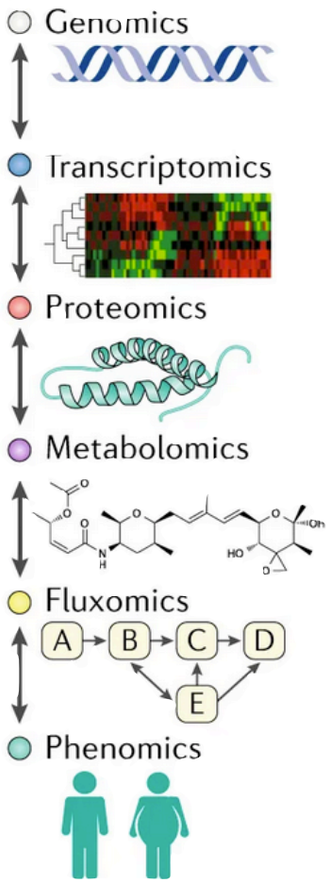
string-db.org

Network inference and analysis workflow





Raw → Pre-processing → **Distance calculation** → Graph analysis

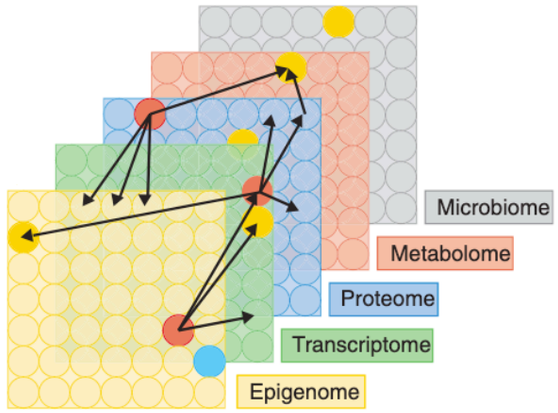


Batch effect correction

Feature selection

Anomaly detection

...



Intra-/Interomic

Hasin 2017

Piening 2018

Mardinoglu 2018

Different approaches for network inference

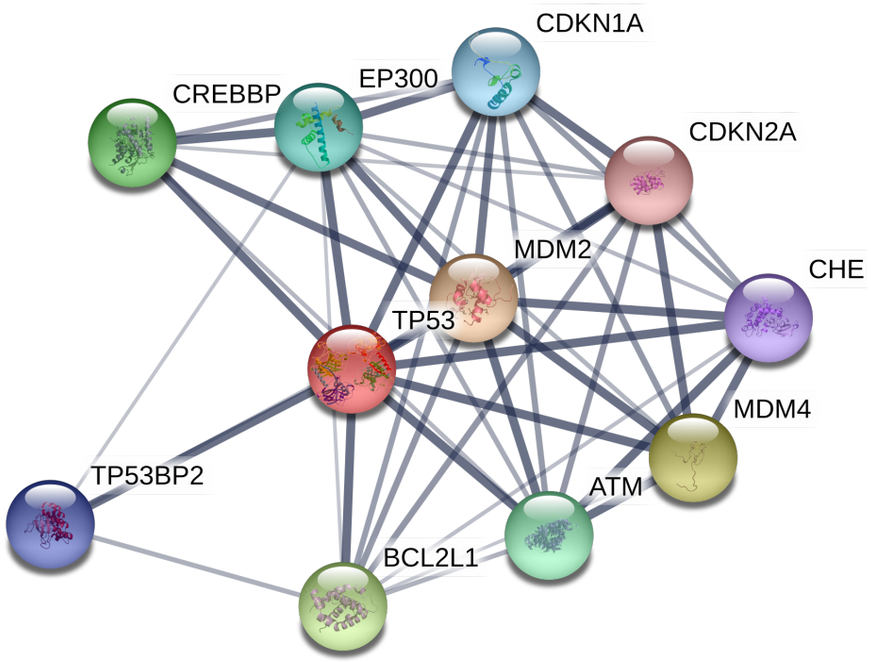
1. Feature association (e.g. correlation, kNN-g)
2. Knowledge-based
3. Genome-scale metabolic models

No prior graph structure

Based on prior knowledge

Nodes and edges in weighted networks

STRING-db.org: TP53 as queried node



Weighted non-directional network

1. Association analysis

Balanced dataset for group sizes

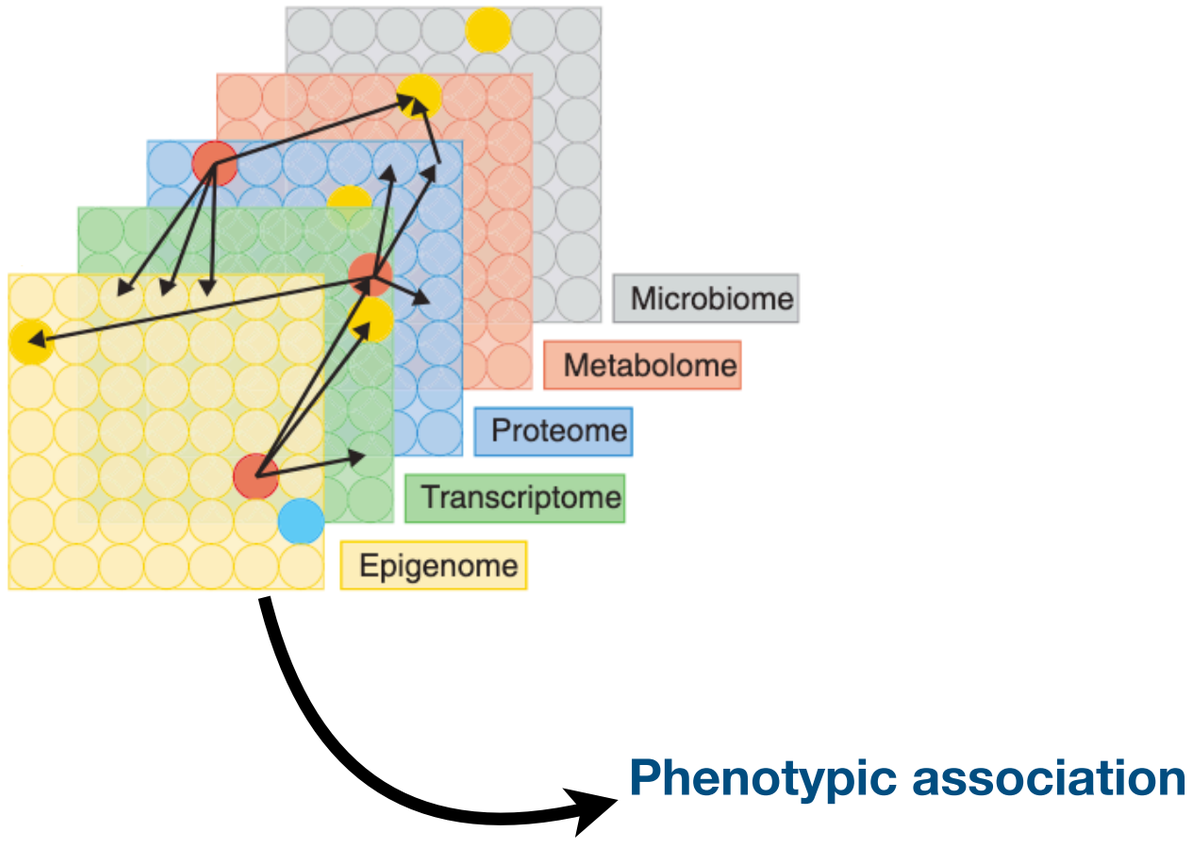
Healthy group (80 samples) vs Disease group (20 samples)

Healthy group (50 samples) vs Disease group (50 samples)

Common approach: compute correlations between features

- Spearman
- Pearson
- WGCNA

Extend known associations



Adapted from Piening 2018

1. Association analysis

Correlations are easy to interpret

Unweighted vs weighted ($-1 \leq \rho \leq 1$)

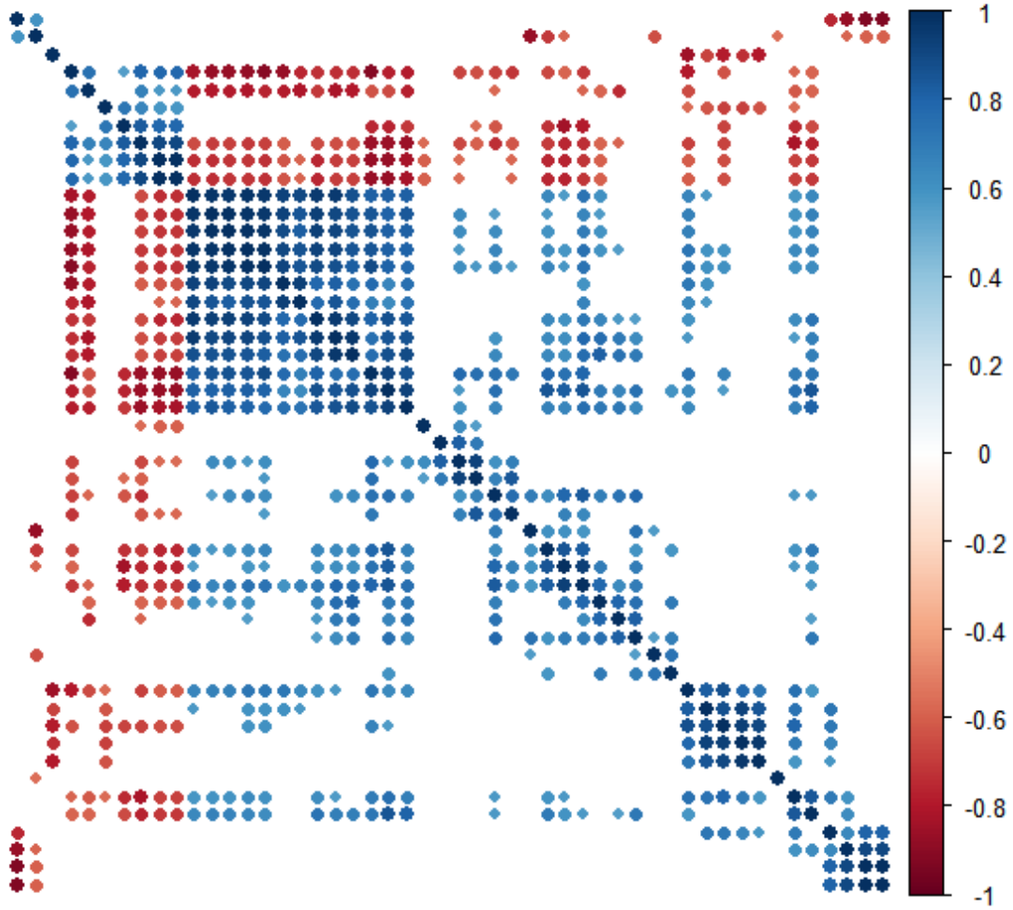
Unbalanced networks vs KNN-g

Prone to type I errors

Filtering

- FDR / Bonferroni
- Correlation coefficient cutoff

Need adjustment to possible confounding factors



Wu 2019
Uhlen 2017

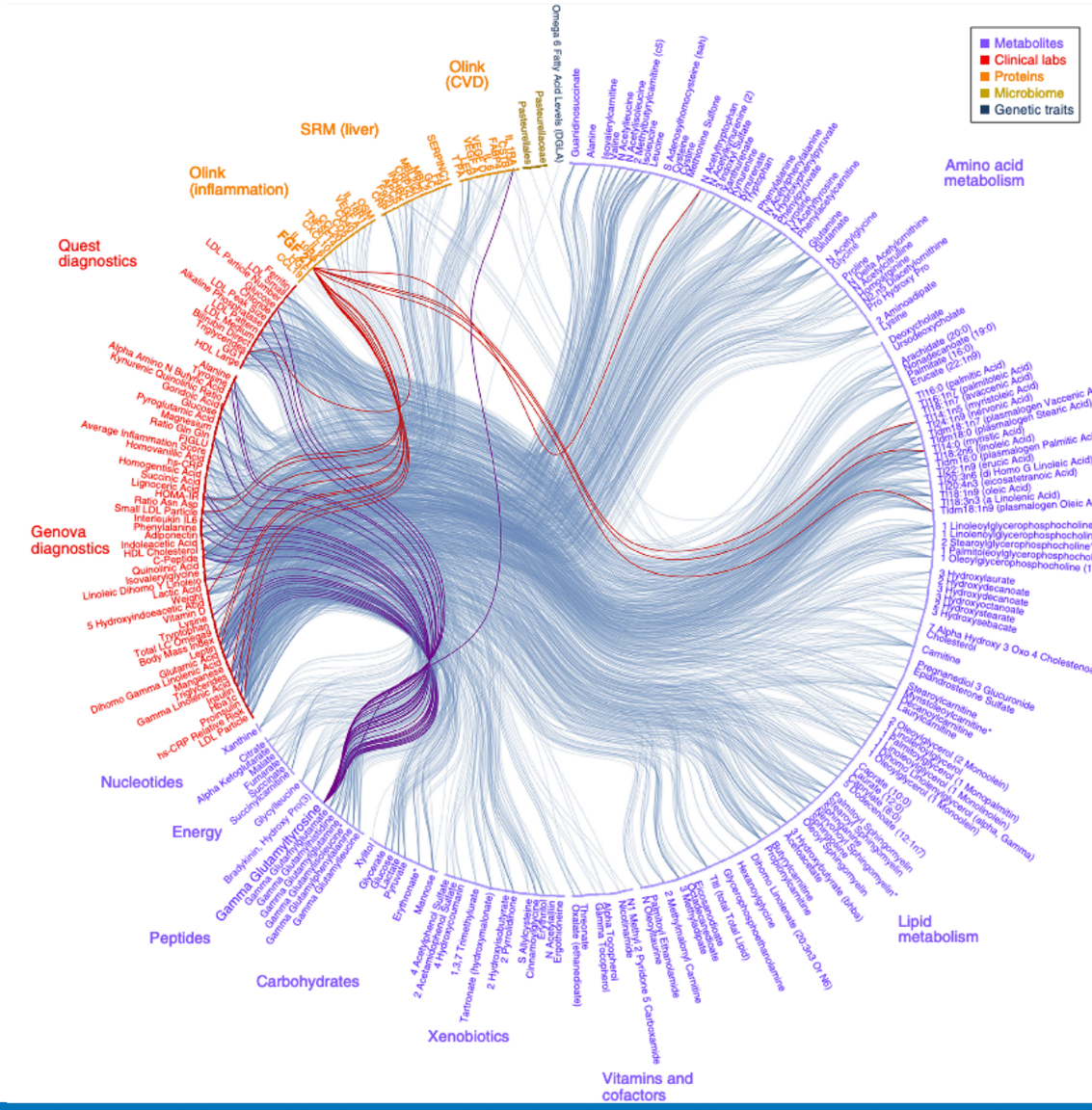
1. Association analysis

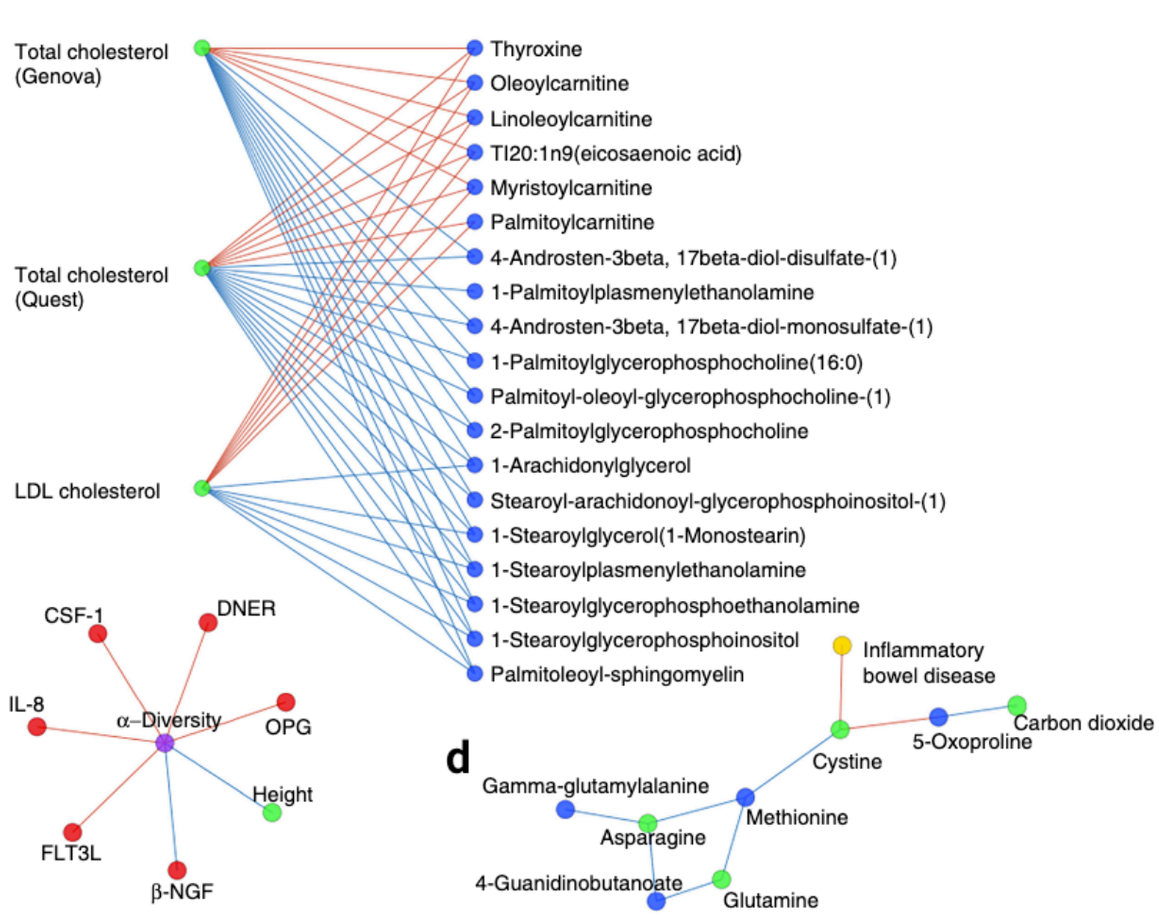
Adjusting for confounding factors

Below:

- gender and age are known confounding factors

- feature regression on confounding factors, followed by correlation on the residuals of each model





Price 2016

2. Knowledge-based graph extraction

Many reference databases

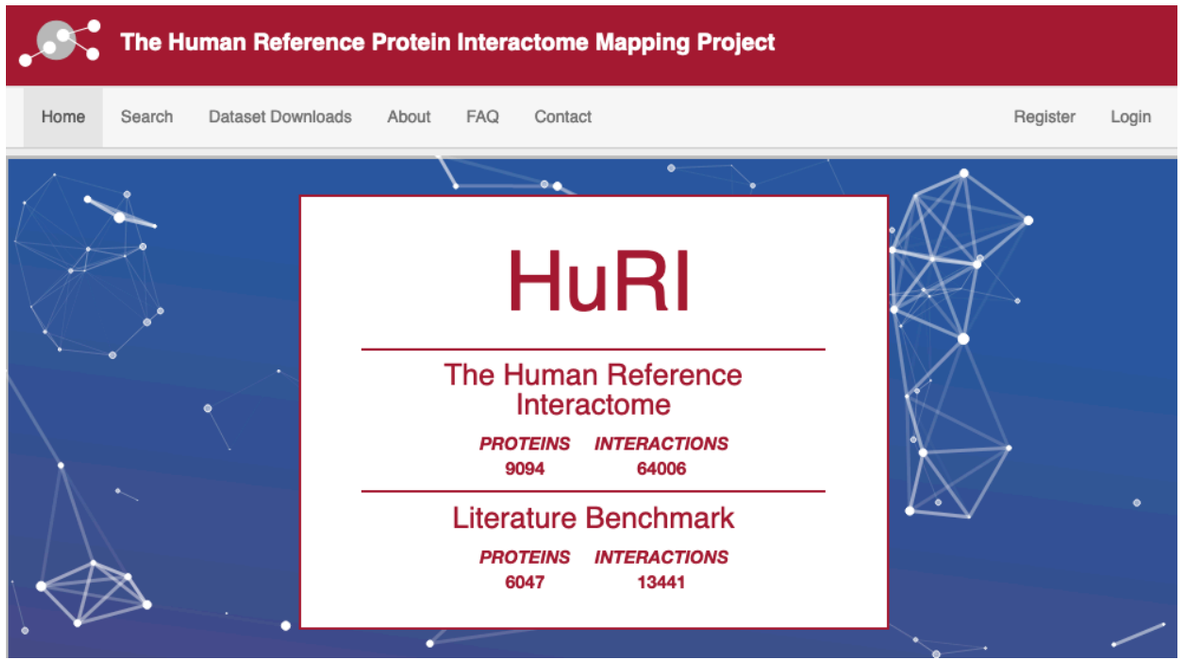
KEGG

Reactome

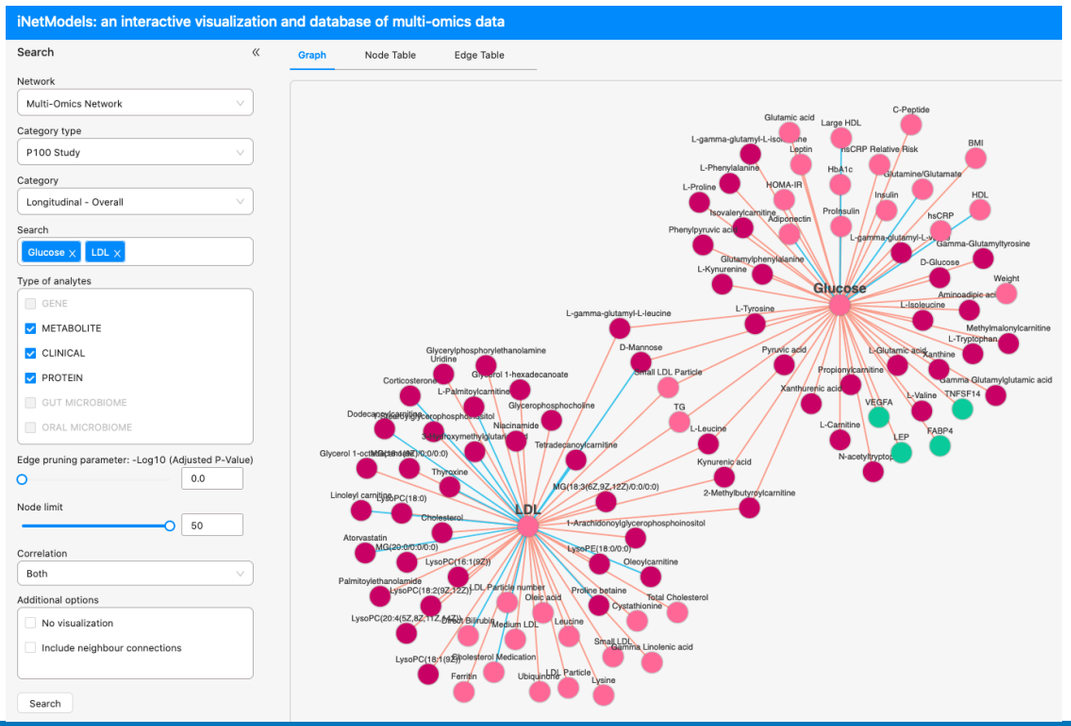
WikiPathways

String-DB

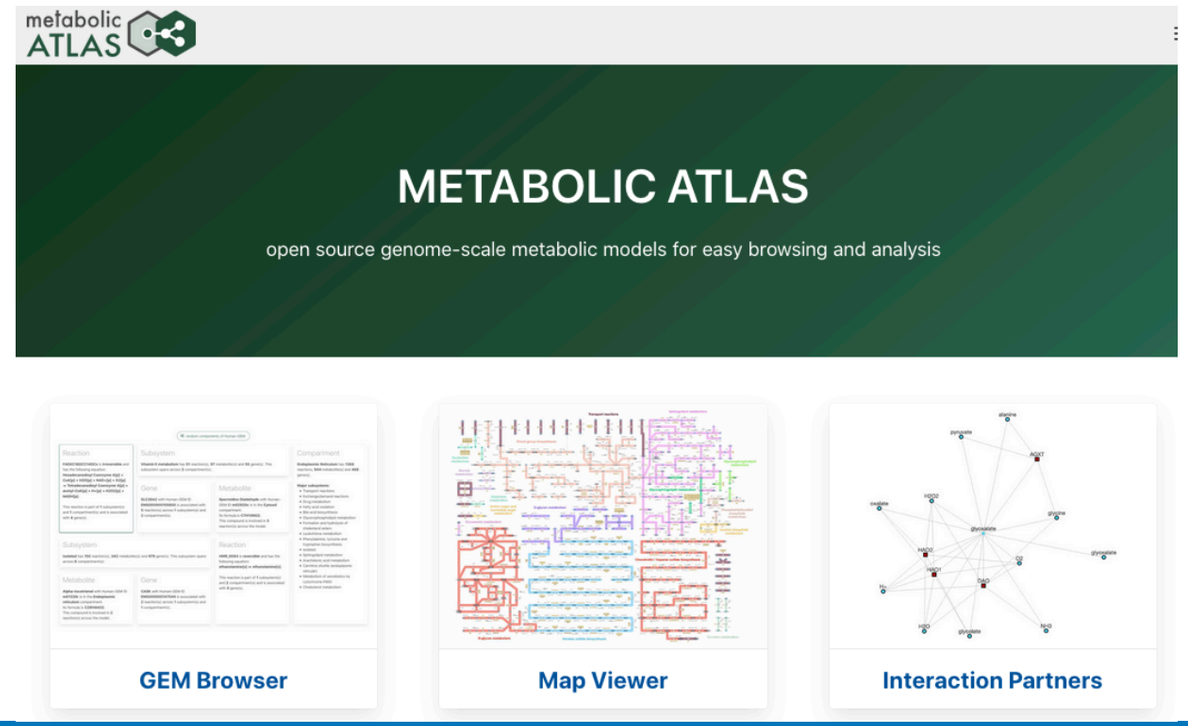
Interactome



Inetmodels



Metabolic Atlas



Outline

1. Networks as frameworks for phenotypic characterization

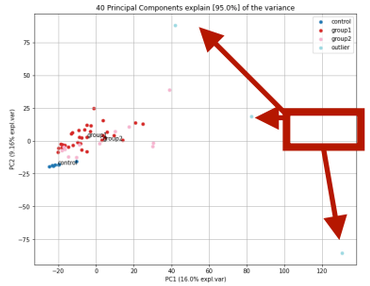
2. Network inference

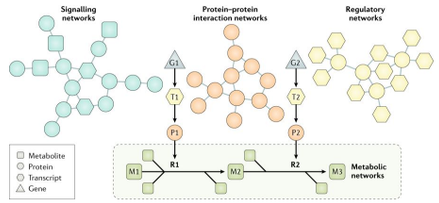
3. Key properties in biological network characterisation

4. Communities and functional analysis

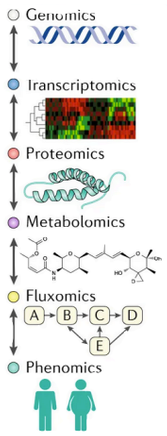
Original sources of images provided as reference and hyperlinks, where applicable.

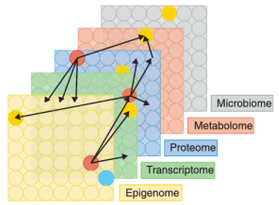
Network inference and analysis workflow





Raw → Pre-processing → Distance → **Graph analysis**





- Computing global properties (e.g. clustering coef., average path, …)
- Comparison against null models
- Identification of key nodes
- Module assessment
- Phenotypic interrogation
- Experimental validation of key drivers

Hasin 2017

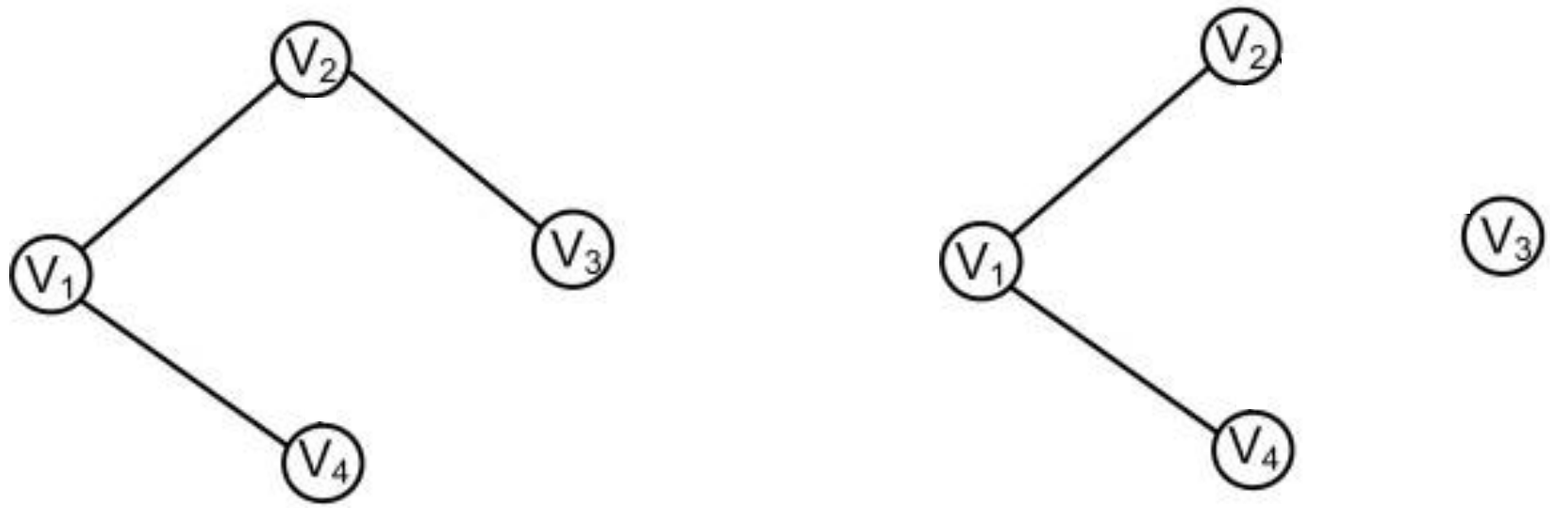
Piening 2018

Mardinoglu 2018

Connected vs disconnected networks

Connected network: there is at least 1 path connecting all nodes in a network

Disconnected network: some of the nodes are unreachable



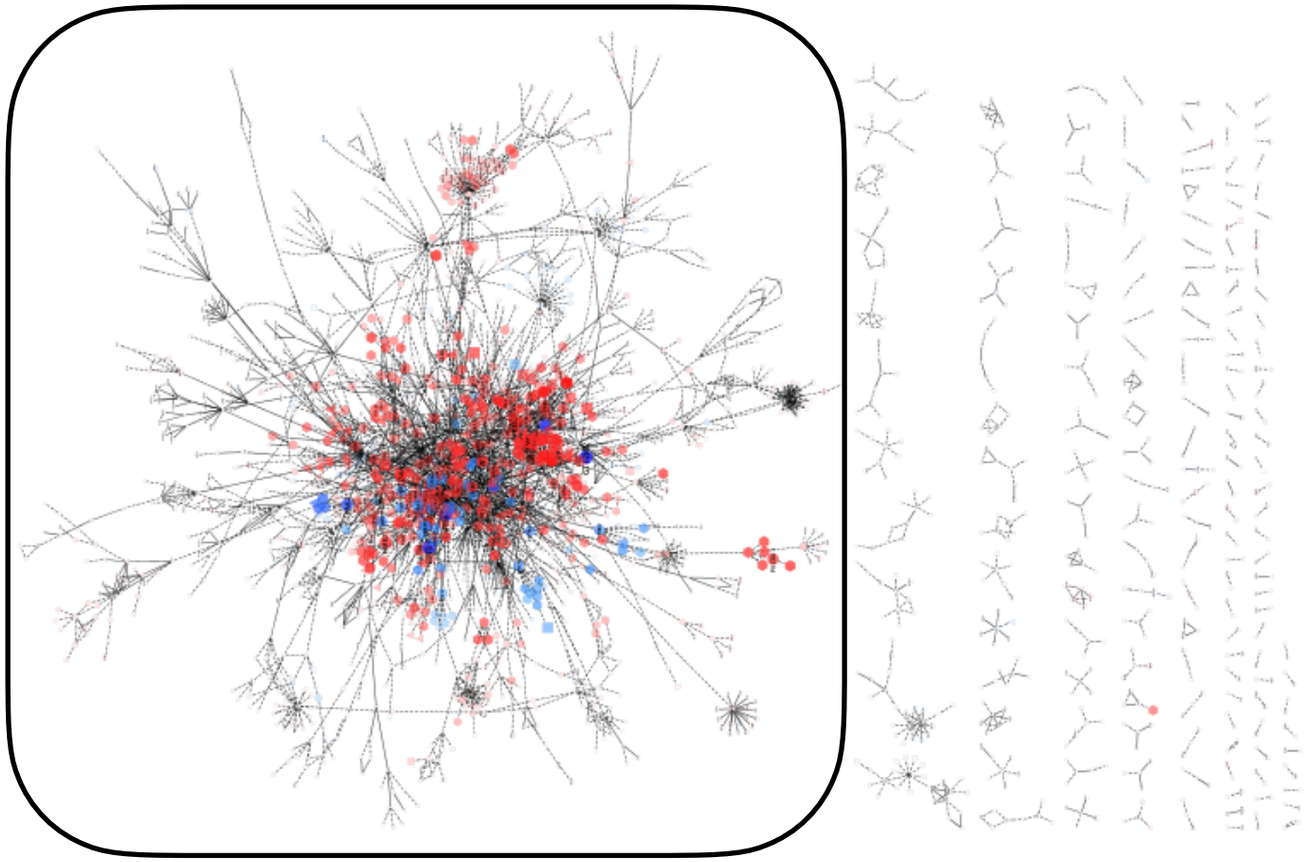
Biological networks are often disconnected

Null models (Erdos-Renyi graphs) are often connected

Connected components

Connected components are those where all nodes of each subgraph are connected.

In biological networks it is often useful to examine the **largest connected component(s)**



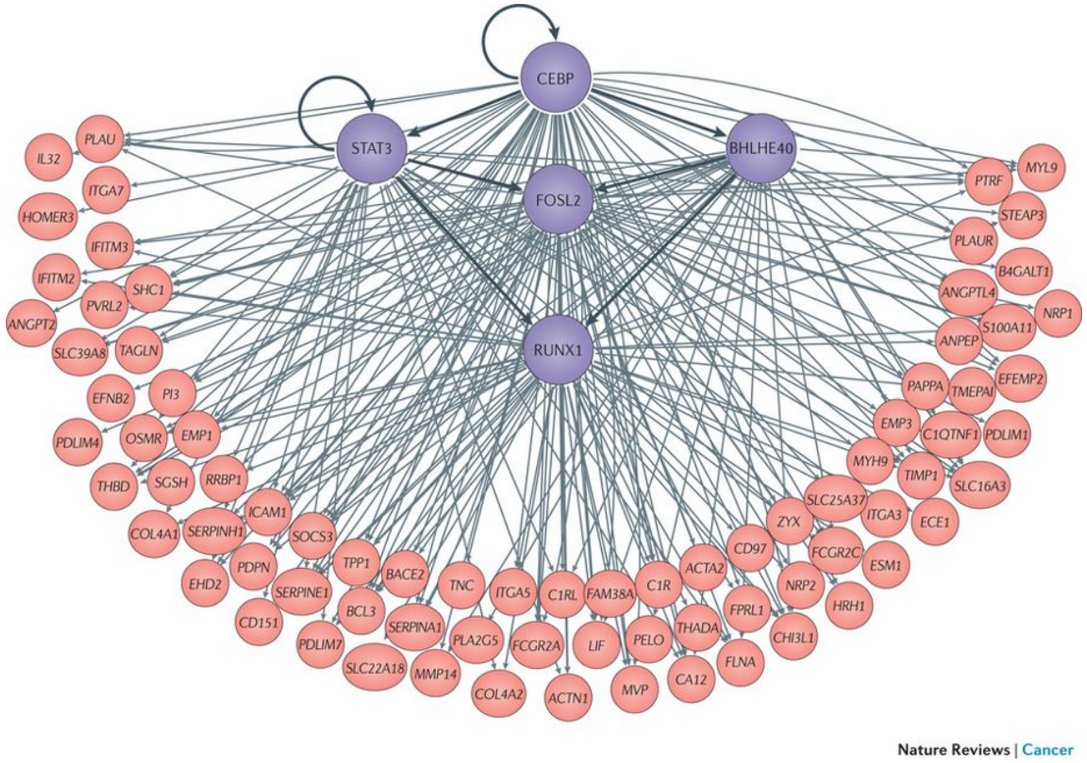
Why centrality is important

Indicate the most central nodes in a network

Why look at the central nodes?

Hubs

Example: Transcription Factor Master Regulators



Tyagarajan 2014

Centrality metrics

Indicate the most central nodes in a network

Hypothesis: central nodes are important in the network

There are many different measures of centrality:

- Eccentricity

- Degree

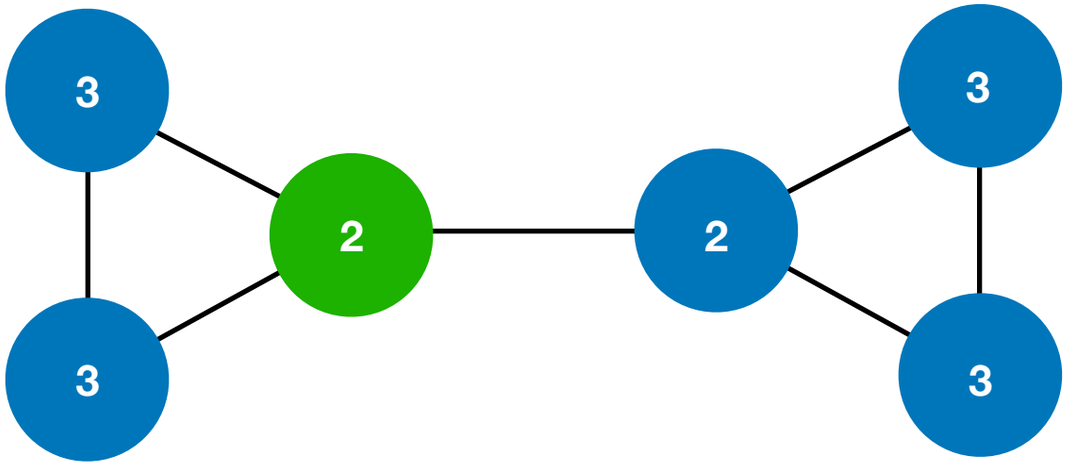
- *Betweenness*
- *Closeness*
- *Eigenvector*
- PageRank
- Katz
- Percolation
- Cross-clique

…

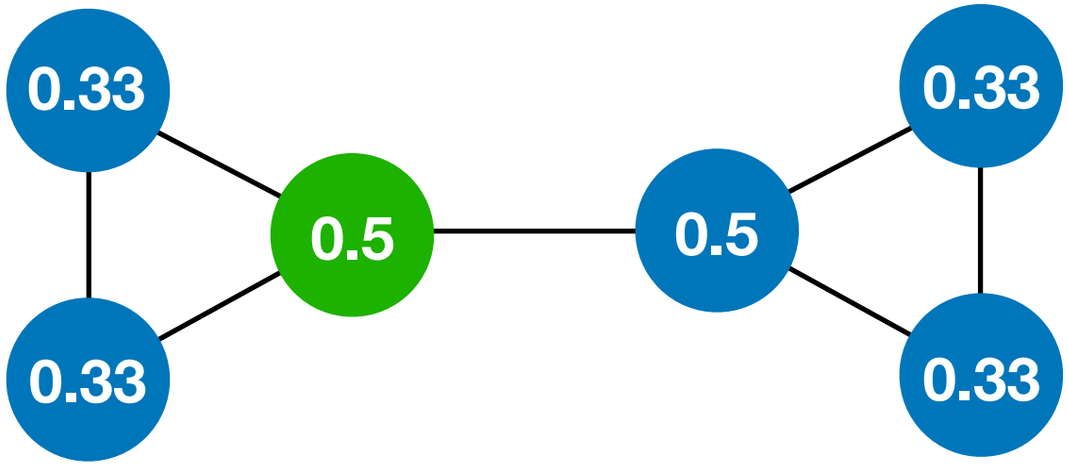
Eccentricity centrality

Eccentricity considers the maximum shortest path passing through a node

[max shortest path]



$$\frac{1}{\text{[max shortest path]}}$$



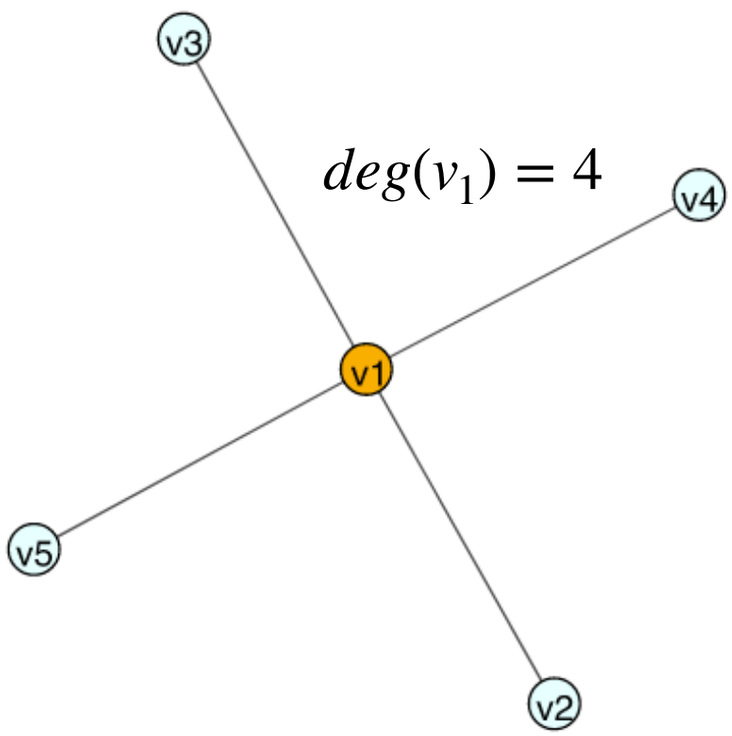
Hage 1995

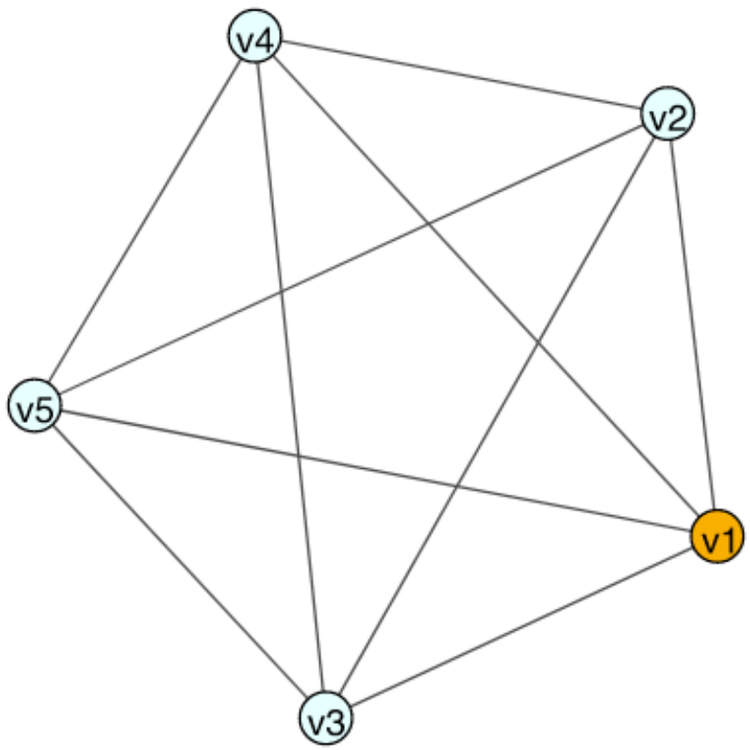
Degree centrality

Degree indicates the number of connections with a node

$$d(v) = |N(i)|$$

where *N(i)* is the number of 1st neighbours of a node.





Degree centrality

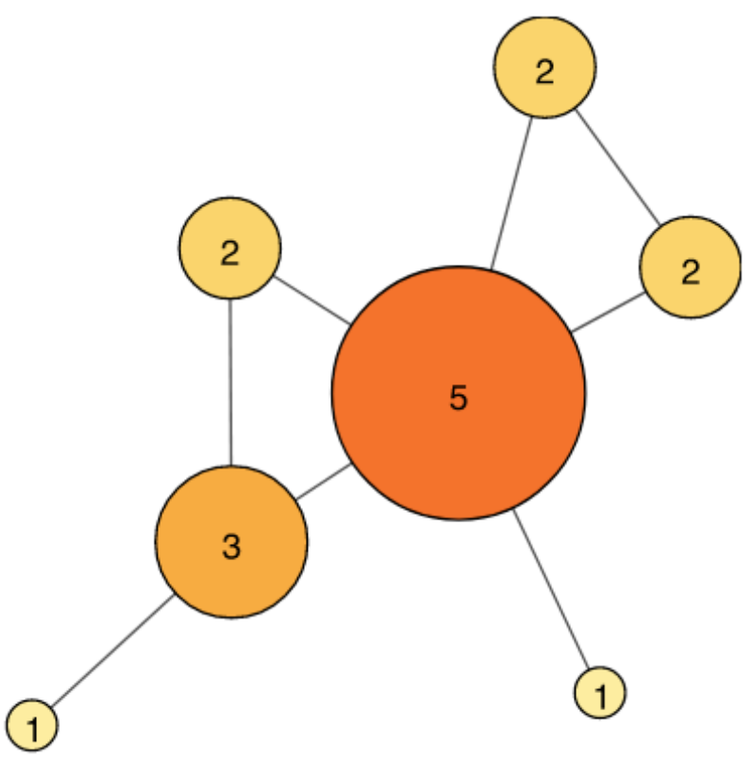
Degree centrality

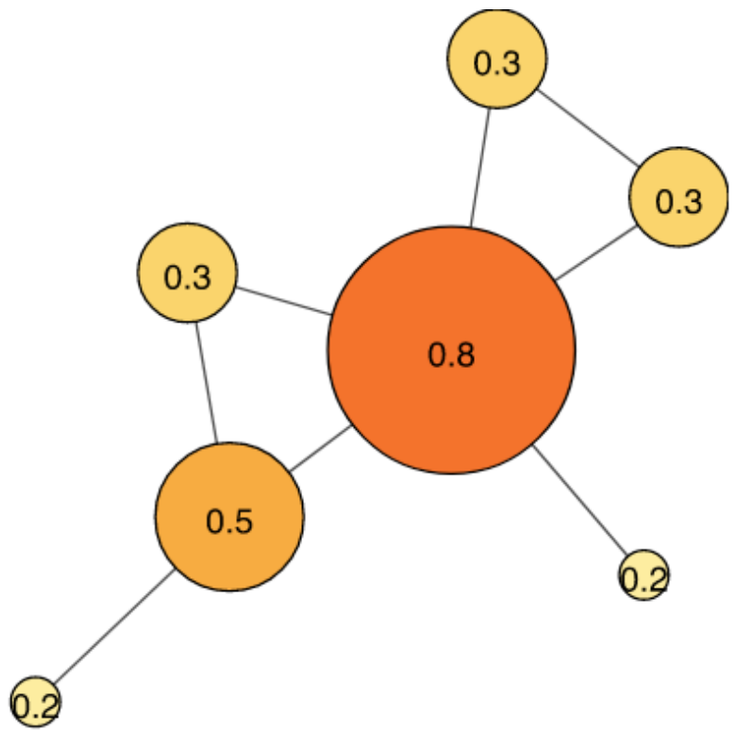
$$C_D(v_i) = \sum_{j=1}^{N} e_{ij}$$

Normalized degree centrality

$$C_D(v_i) = \frac{\sum_{j=1}^{N} e_{ij}}{N-1}$$

Centrality normalization allows for comparison between networks of different sizes

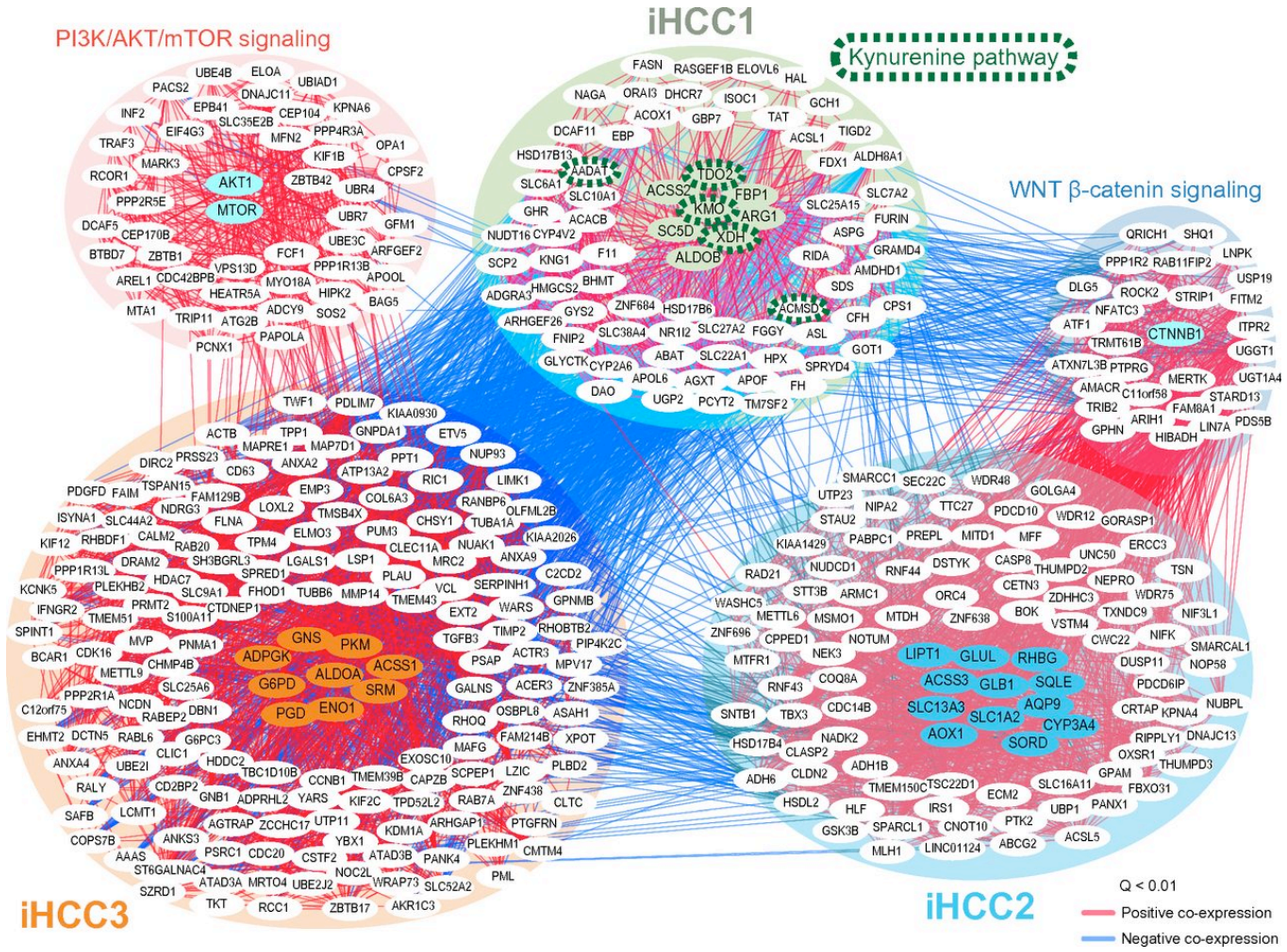


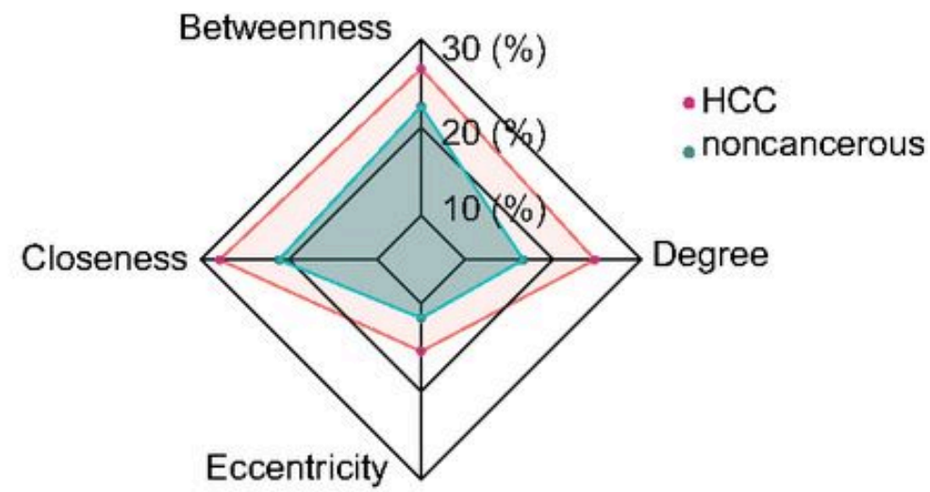


Centrality: good practices

Should be:

1. Compute multiple metrics and understand their differences
2. Find nodes with largest ranks





3. Compute node **influence,** modifications of centrality

Measure **information transmission** rather than *connectiveness*

Bidkhori 2018

Clustering coefficient

How likely is it that two connected nodes are part of a highly connected group of nodes?

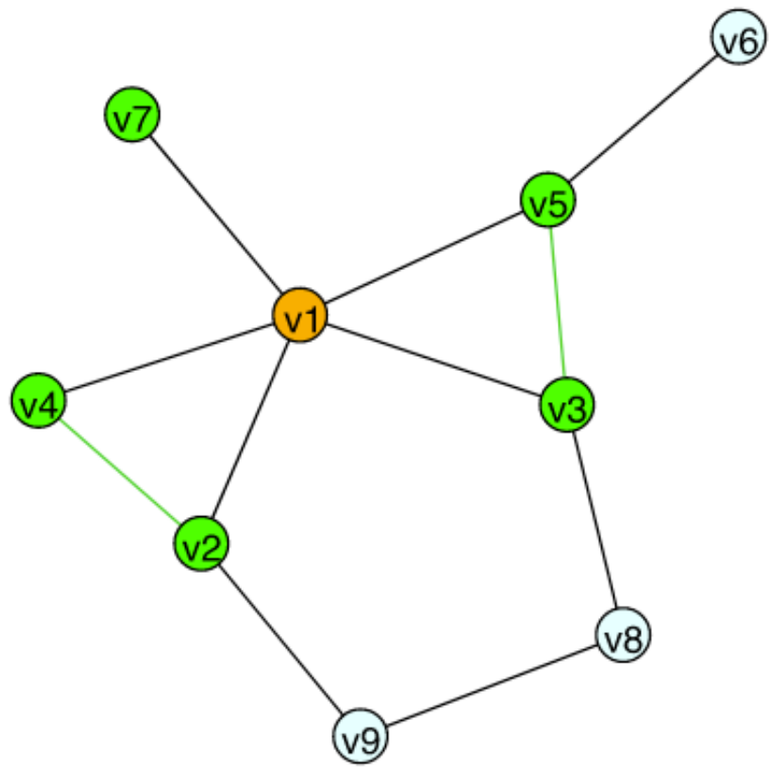
Takes into account degree of a node and the degree of its 1st neighbours

For node v_1

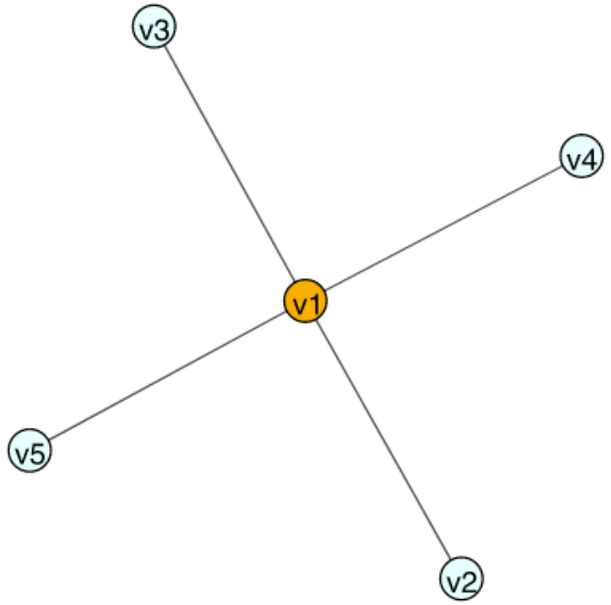
- $deg(v_1) = k = 5$
- n connections between 1st neighbours of $v1 = 2$

$$C_i = \frac{n}{\frac{k \cdot (k-1)}{2}}$$

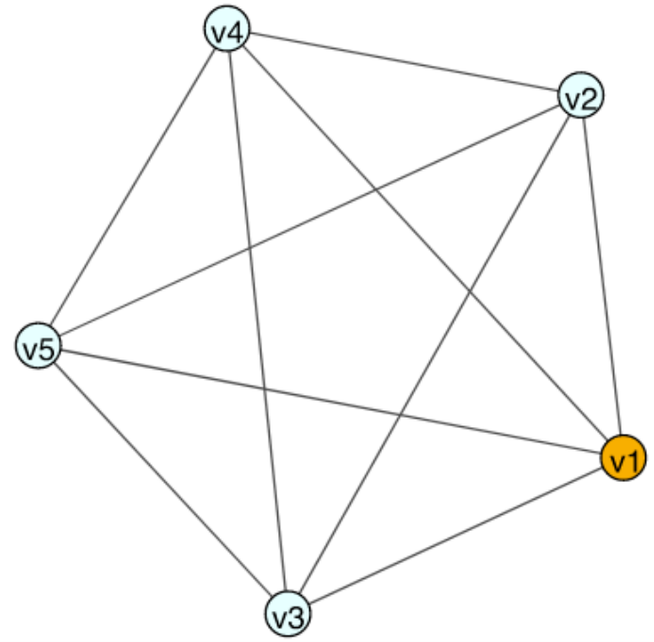
$$C(v_1) = \frac{2}{\frac{5 \cdot 4}{2}} = 0.2 \qquad C(v_7) = \frac{2}{0} = ND \text{ or } 0$$



Clustering coefficient

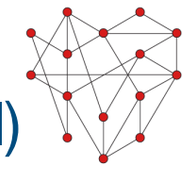
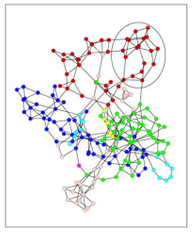


$$0 \leq C_i \leq 1$$



The global clustering coefficient $C(G)$ is the average of its clustering coefficients

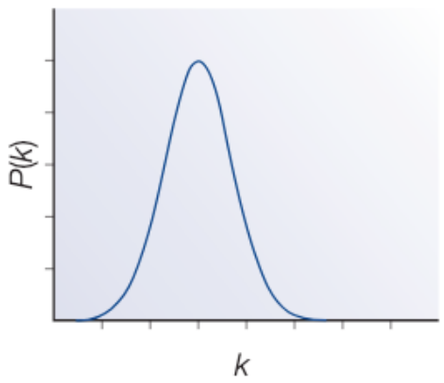
What distinguishes biological networks from random?

	Random network (e.g. Erdös-Rényi model) 		**Metabolic network** (hierarchical organization) 
Node number	N	=	N
Edge number	E	=	E
Density	D	=	D
Average shortest path	L	<	$\boldsymbol{L}$

Node failure easily propagates

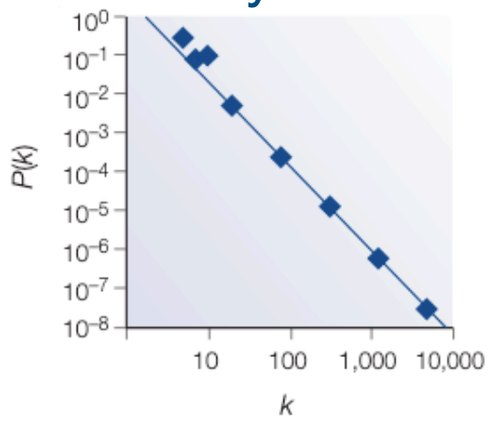
Degree distribution

no highly connected nodes



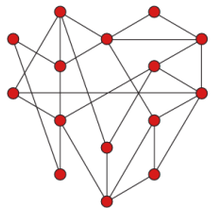
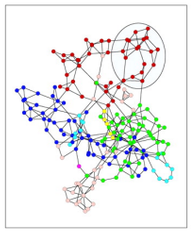
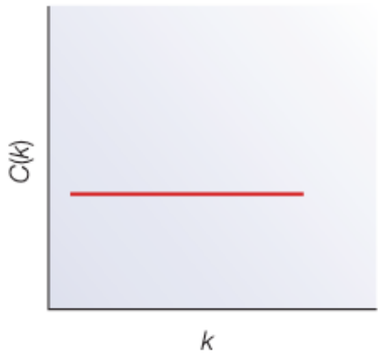
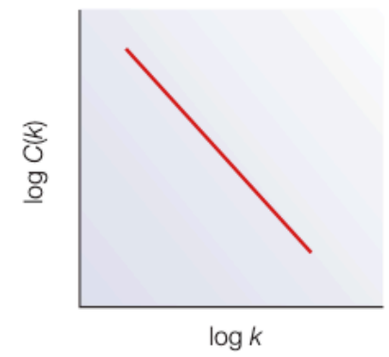
Most nodes have ~ <*k*>

many with low degrees
few very central



Barabasi 2004
Jeong 2000
Ravasz 2002

What distinguishes biological networks from random?

	Random network (e.g. Erdös-Rényi model)	**Metabolic network** (hierarchical organization)
		
Clustering coefficient	 Degree	 Degree
Organization	No modular organisation	Sparsely connected nodes in highly modular areas Communication between modules maintained by few hubs

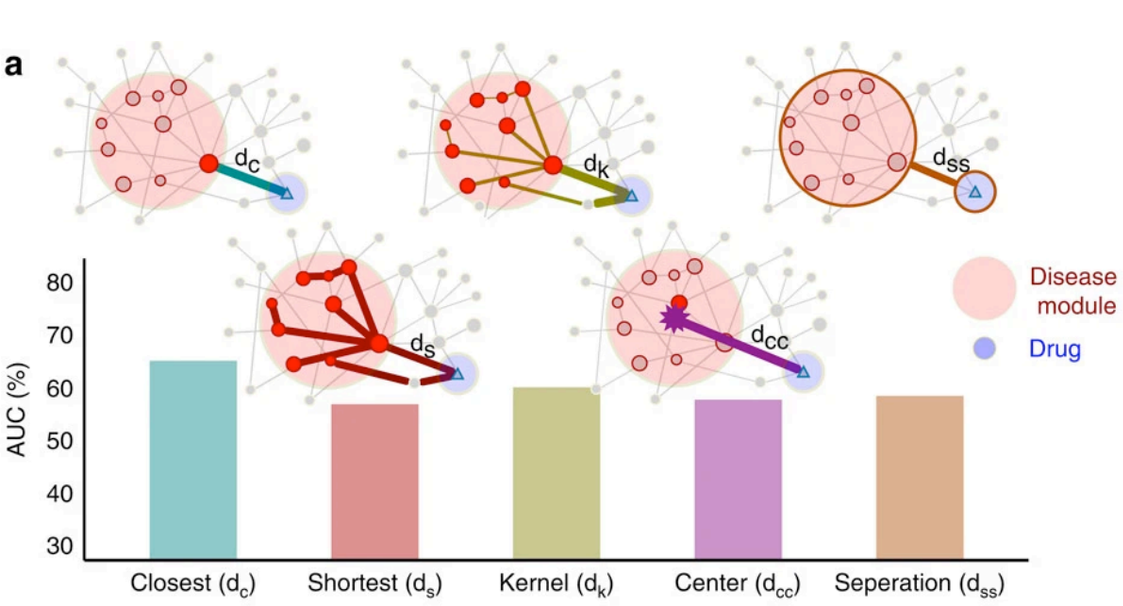
high robustness to node failure:
removal of <80% nodes still retains paths between remaining nodes

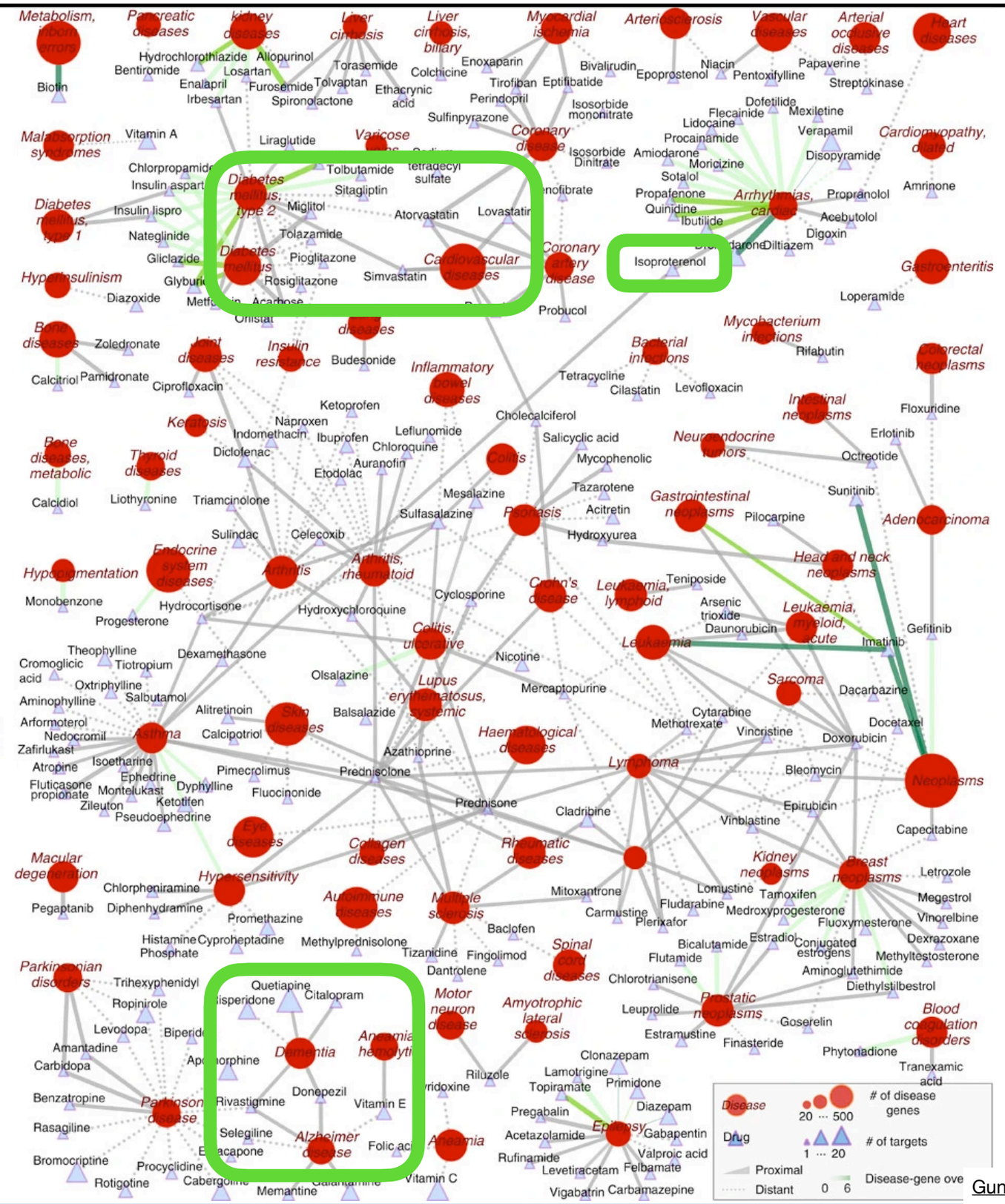
Barabasi 2004
Jeong 2000
Ravasz 2002

Centrality applied to drug repositioning

Centrality used to prioritise drug associations

Identified suitable candidates for repositioning





Guney 2016

Outline

1. Networks as frameworks for phenotypic characterization

2. Network inference

3. Key properties in biological network characterisation

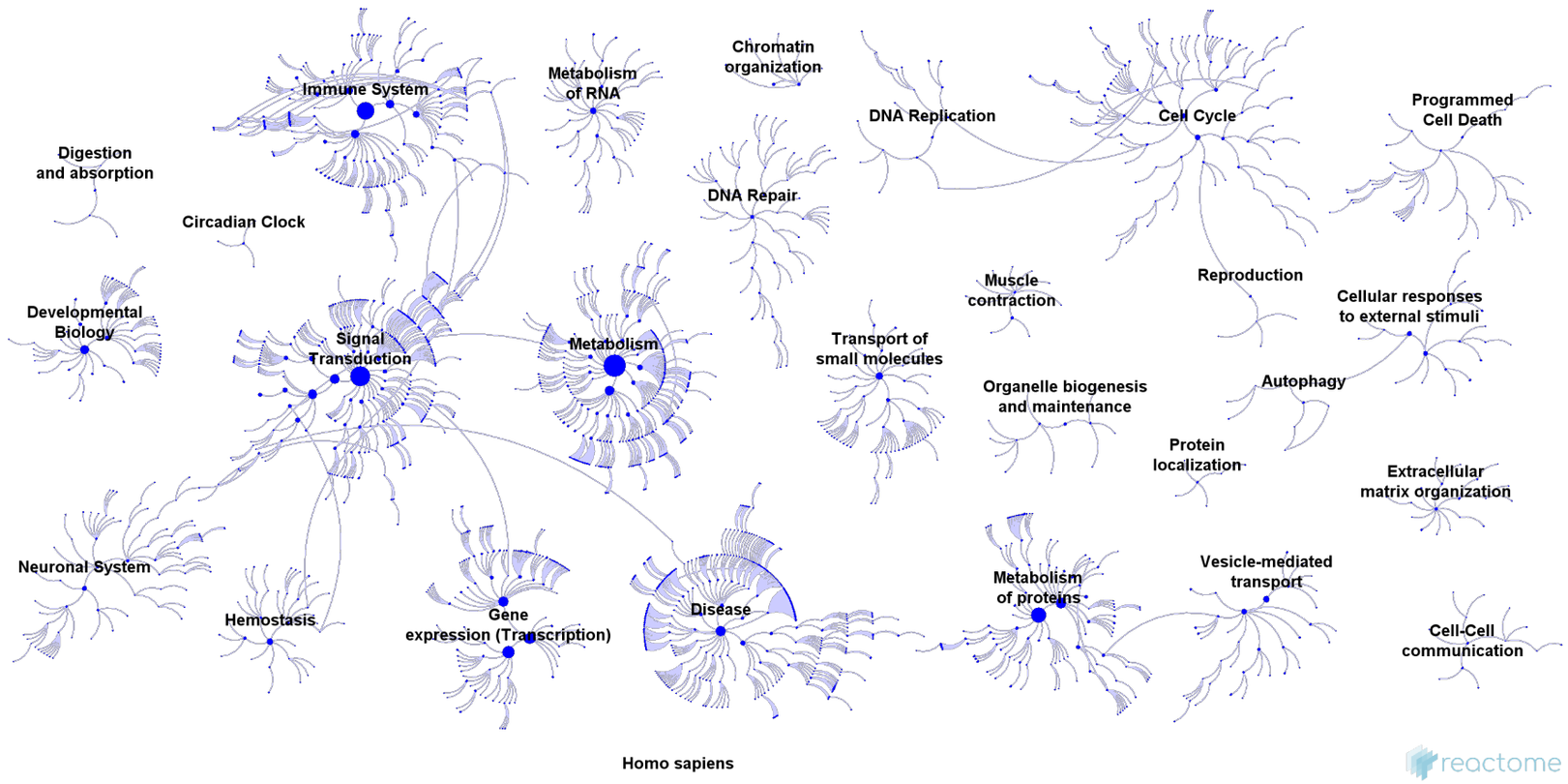
4. Communities and functional analysis

Original sources of images provided as reference and hyperlinks, where applicable.

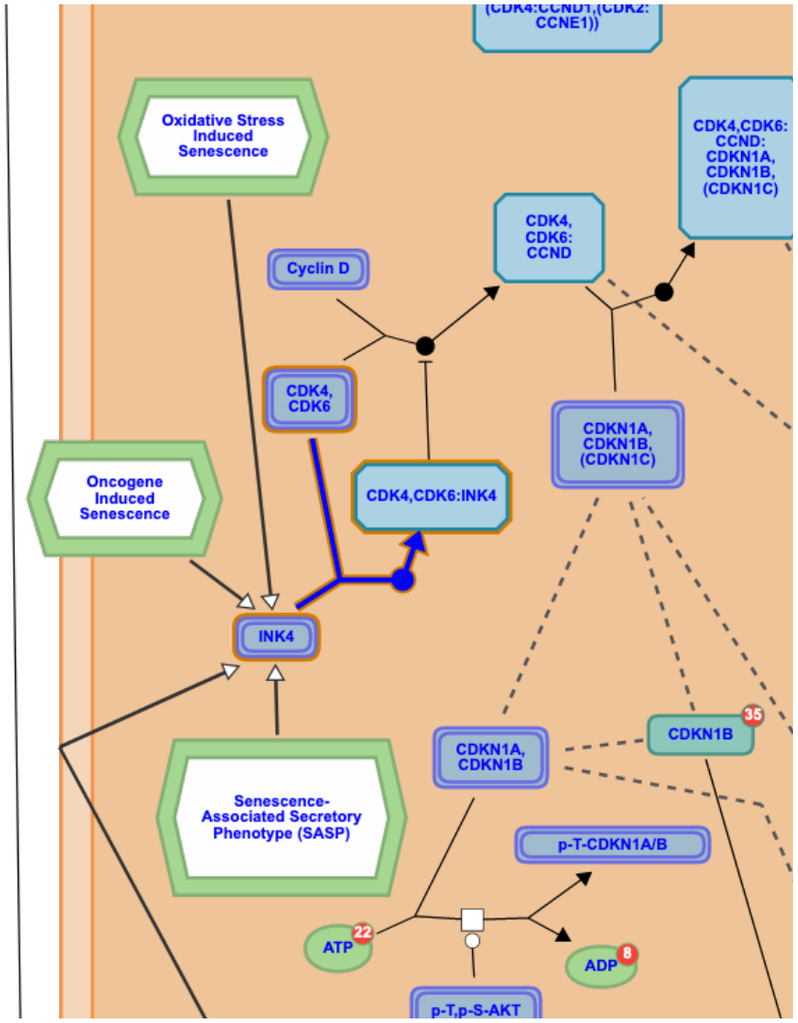
What are modules?

Modules are physically or functionally associated nodes
that work together to achieve a given function

Pathways = functional modules



Protein complexes = physical modules



reactome.org

What are modules?

In addition to physical or functional modules, one may identify other types of modules

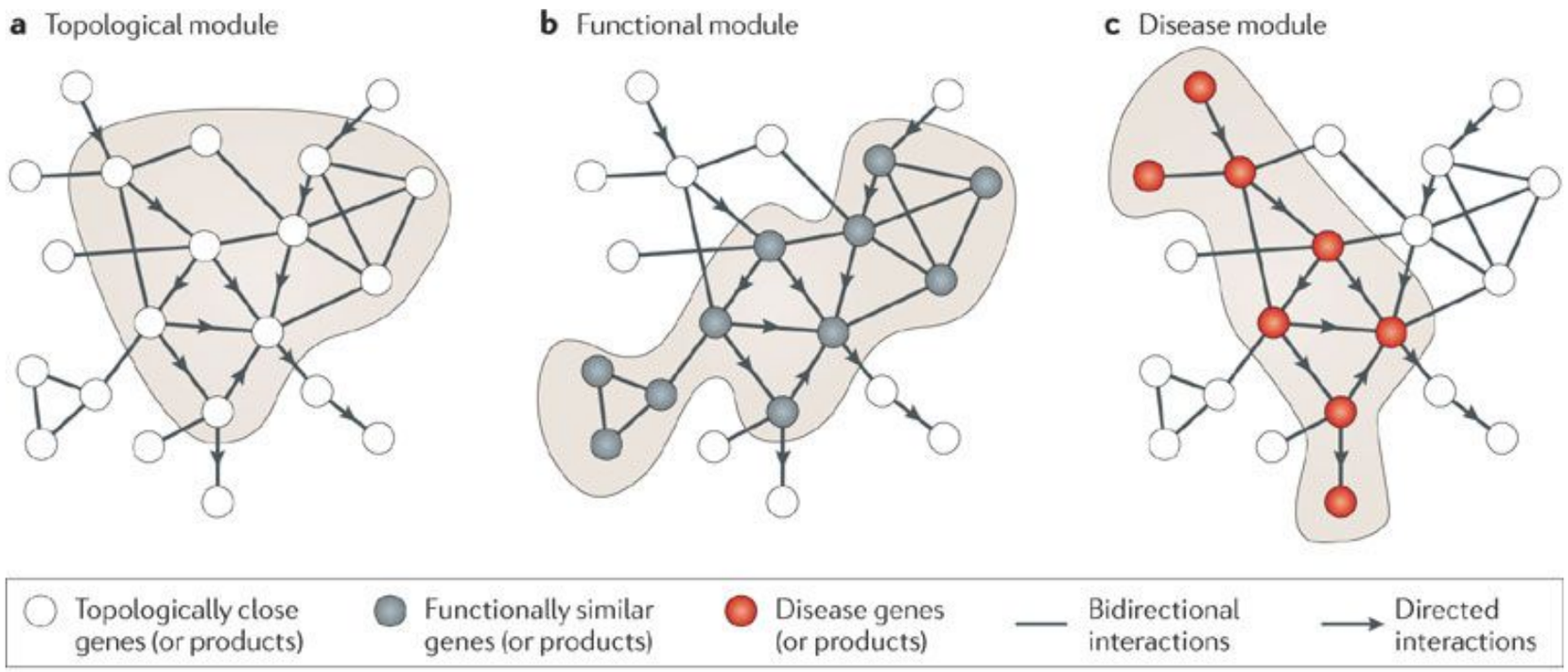
Topological: derived from their high within-module degree

Disease/drug: highly interconnected nodes associated with a disease response or drug

Subgroup: highly interconnected nodes associated with a sample subgroup (e.g. cancer subtype)

Tissue-, cell-type-specific: highly interconnected nodes associated with a specific tissue or cell type

Highly interlinked local regions of a network



Barabasi 2011

Modularity

Modularity is a property of the global network

Modularity (Q) measures the tendency of a graph to be organised into modules

Modules computed by comparing probability that an edge is in a module vs what would be expected in a random network

$$Q \propto \sum_{s \in S} [(e_s) - (\text{expected } e_s)]$$

edges in group s

Random network with same number of nodes, edges and degree per node

$-1 < Q < 1$

Q = 1: much higher number of edges than expected by chance

Q = -1: lower number of edges than expected by chance

Q > 0.3 - 0.7 means significant community structure

Module detection: Louvain algorithm

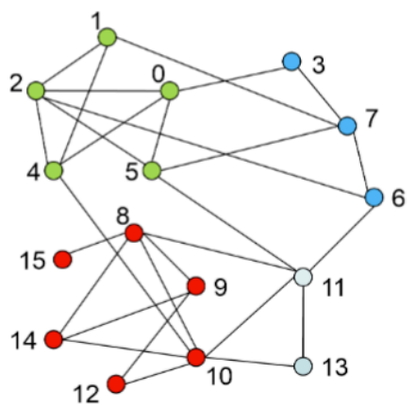
Phase 1: greedy modularity optimisation

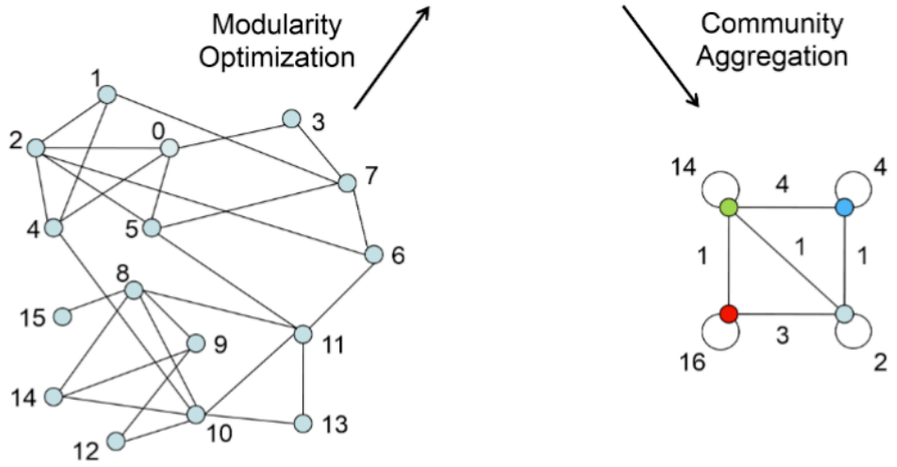
1. Start with 1n/community
2. Compute Q by moving *i* to the community of *j*
3. If ΔQ>1, node is placed in community
4. Repeat 1-3 until no improvement is found. Ties solved arbitrarily

Phase 2: coarse grained community aggregation

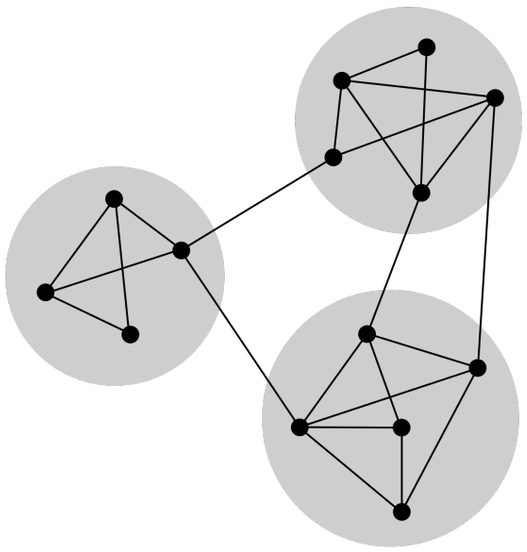
5. Link nodes in a community into single node.
6. Self loops show intra-community associations
7. Inter-community weights kept

Second pass: repeat phase 1 on the new network





Other methods:
Walktrap
Label propagation
…
(benchmarking)



Campigotto 2014
Traag 2019

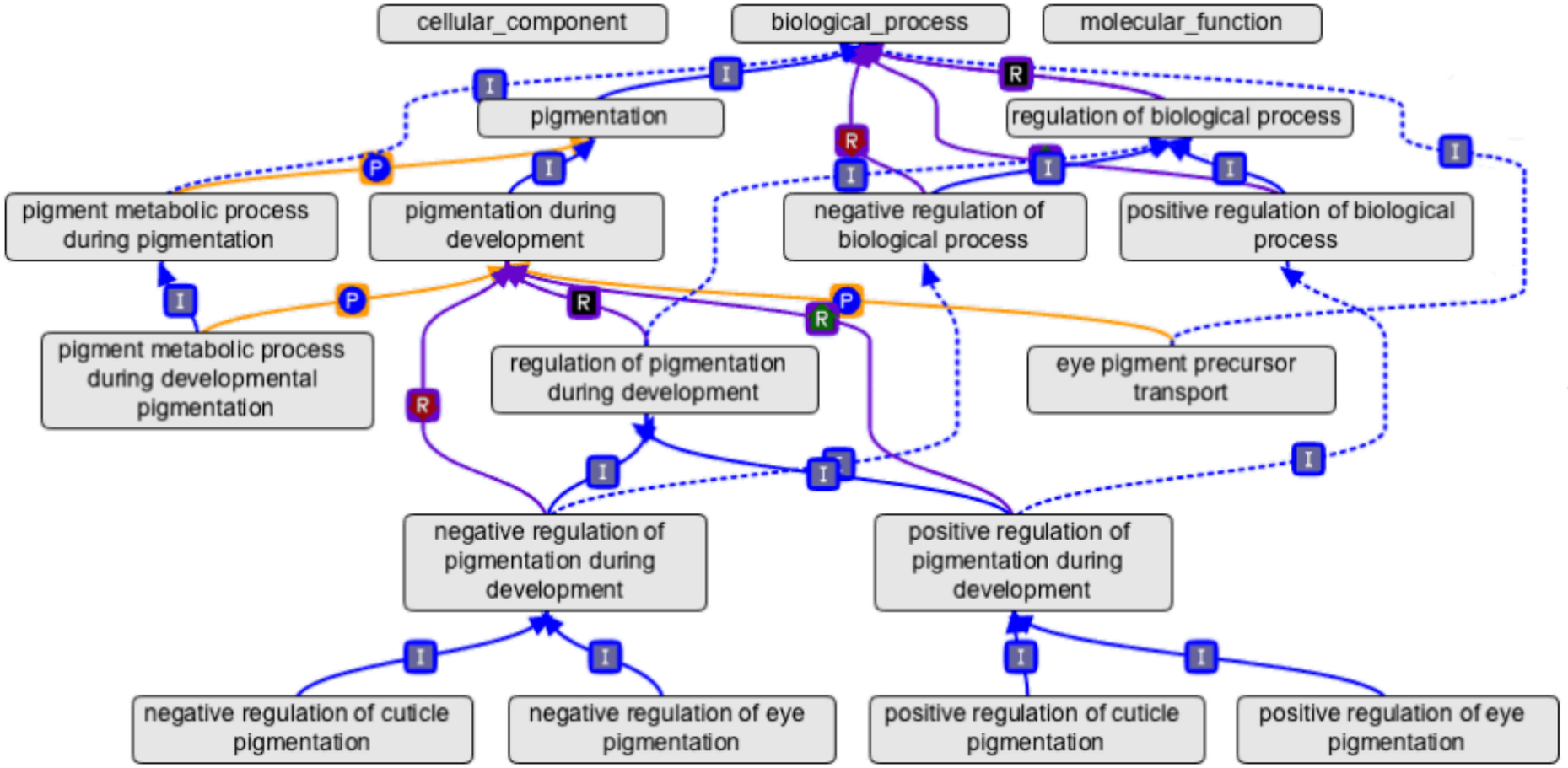
Community characterisation

Clustering coefficient and degree distribution

Enrichment analysis

Phenotypic association

Hypothesis: community-associated features show coordinated changes related with common biological processes or phenotypes

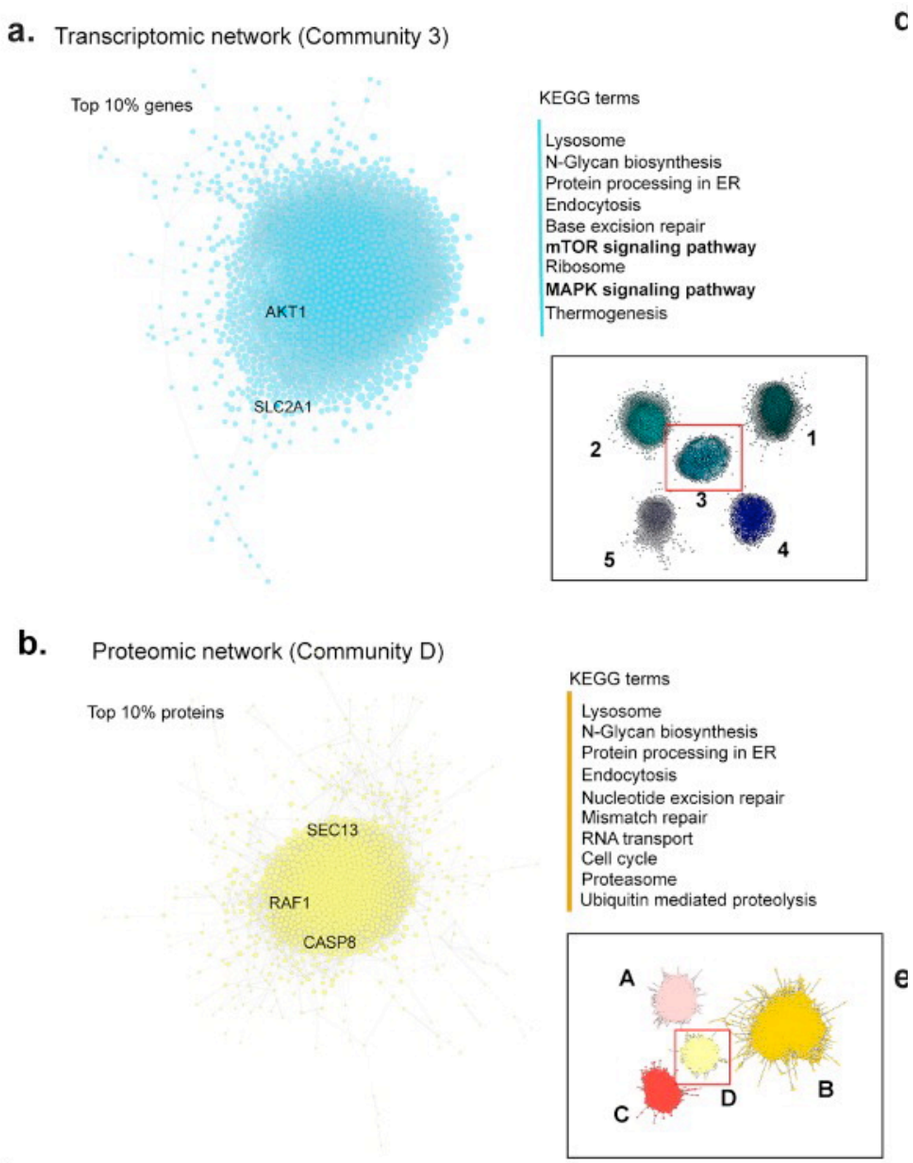


http://geneontology.org/

Multi-tissue network analysis of proteo-transcriptomics data in response to Covid-19 infection

Graph analysis of proteo-transcriptomic data

Centrality and Community characterization



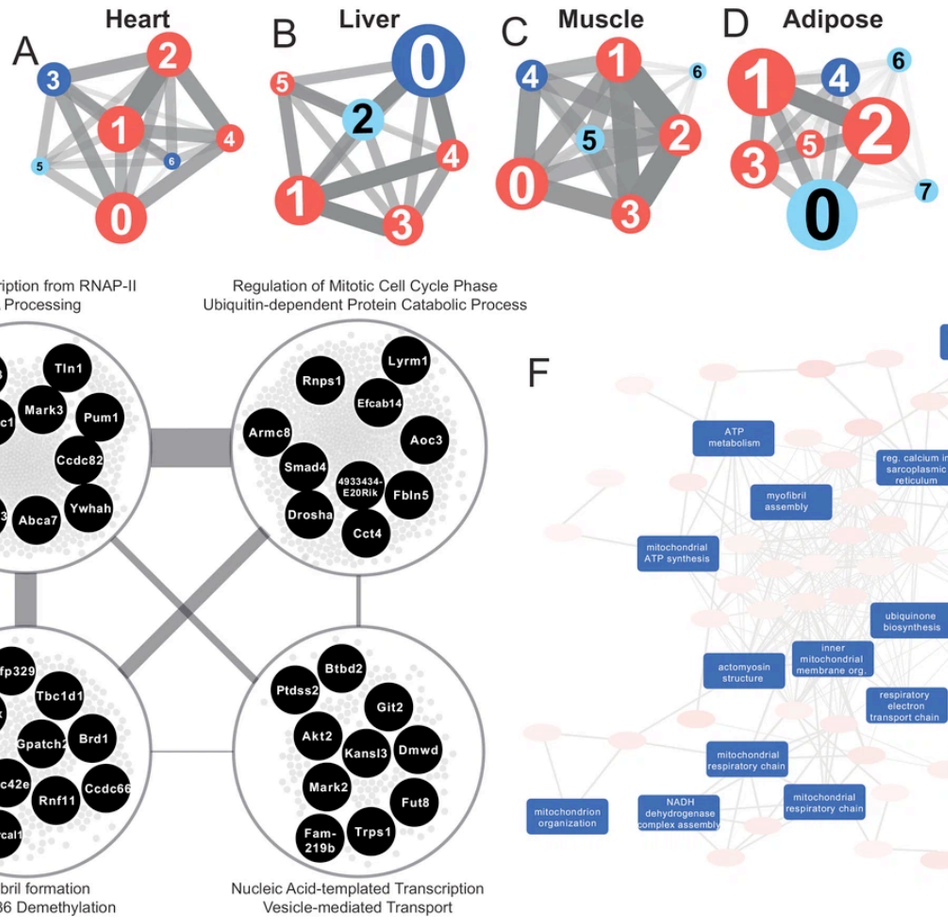
Appelberg 2020

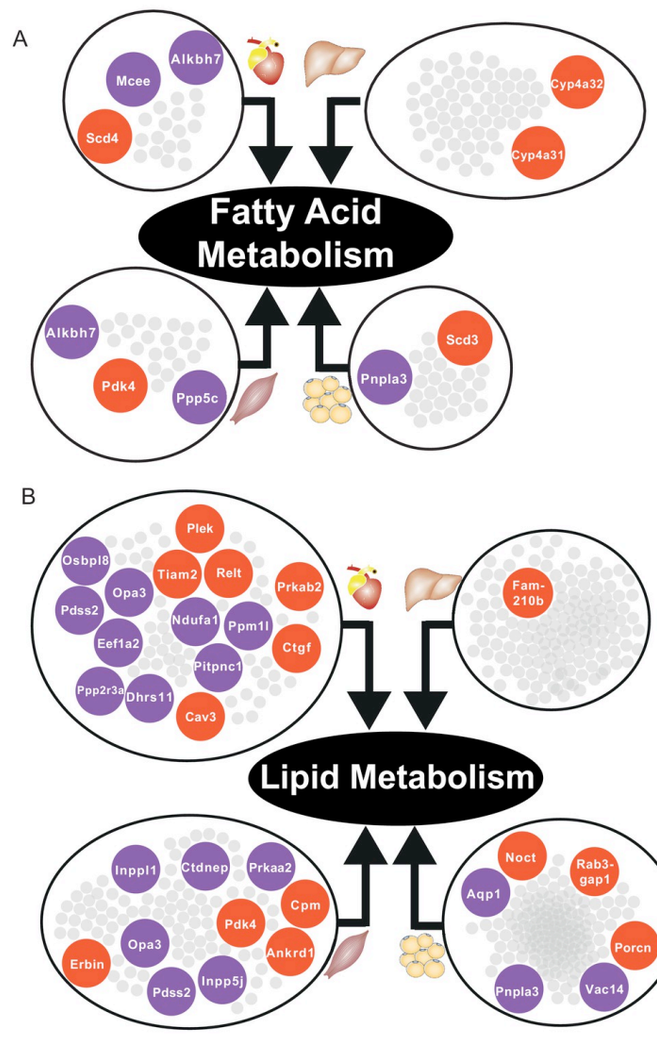
Multi-tissue network analysis of RNAseq data in CVD

Graph analysis (centrality, modularity, cluster coefficients, …)

Community identification and characterization

Tissue-specific and shared responses to myocardial infraction





Arif 2021

Conclusions

Graph analysis can be used to examine coordinated patterns of variation at feature or sample levels

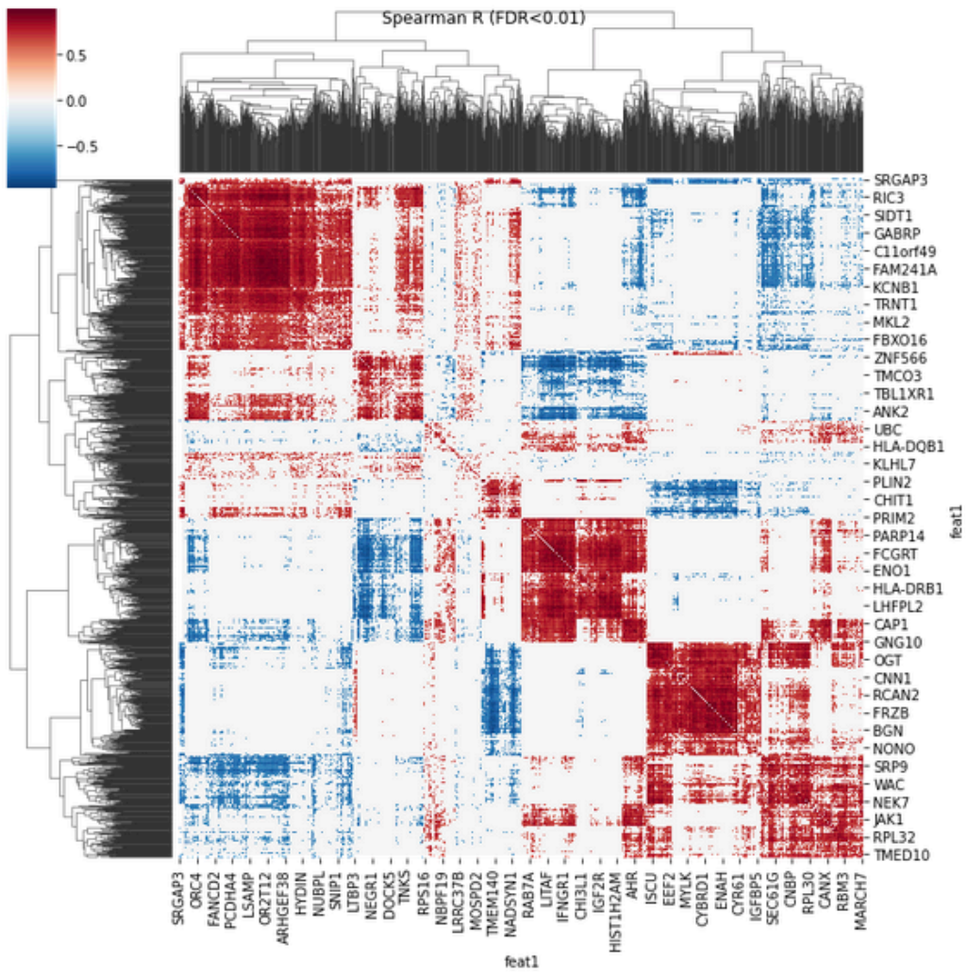
Technical biases need to be tackled, and graphs should be compared against matching null models

Centrality and module assessment may be used to characterise and prioritise features or their groups

Lab: Network construction and analysis of a transcriptomic and metabolomic dataset

The following notebook can be found in `/session_topology/lab.ipynb`. Please use the jupyter container to run it.

In [16]:
```
g=sns.clustermap(Rmatrix_fdr_top, cmap="RdBu_r", center=0);
g.fig.suptitle('Spearman R (FDR<0.01)');
plt.show()
```



The plots above show that the Bonferroni correction is only selecting very high (absolute) correlations. This should remove false positives, but it may also remove weaker correlations that are biologically relevant and true positives. The Bonferroni correction also removes most of the negatively-associated features. Notice this from the distribution of correlation coefficients:

In [17]:
```
shortPR=PRmatrix.copy().loc[:,['feat1','feat2','R (padj)','R (fdr)']]
shortPR=shortPR.loc[shortPR.feat1!=shortPR.feat2]

fig=plt.figure(figsize=(8,4))
p=sns.histplot(shortPR['R (padj)'][shortPR['R (padj)']!=0], color='black', label='Bonferroni (<0.01)', kde=True, bins=100);
```