Community analysis and visualisation

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

rui.benfeitas@scilifelab.se

Overview

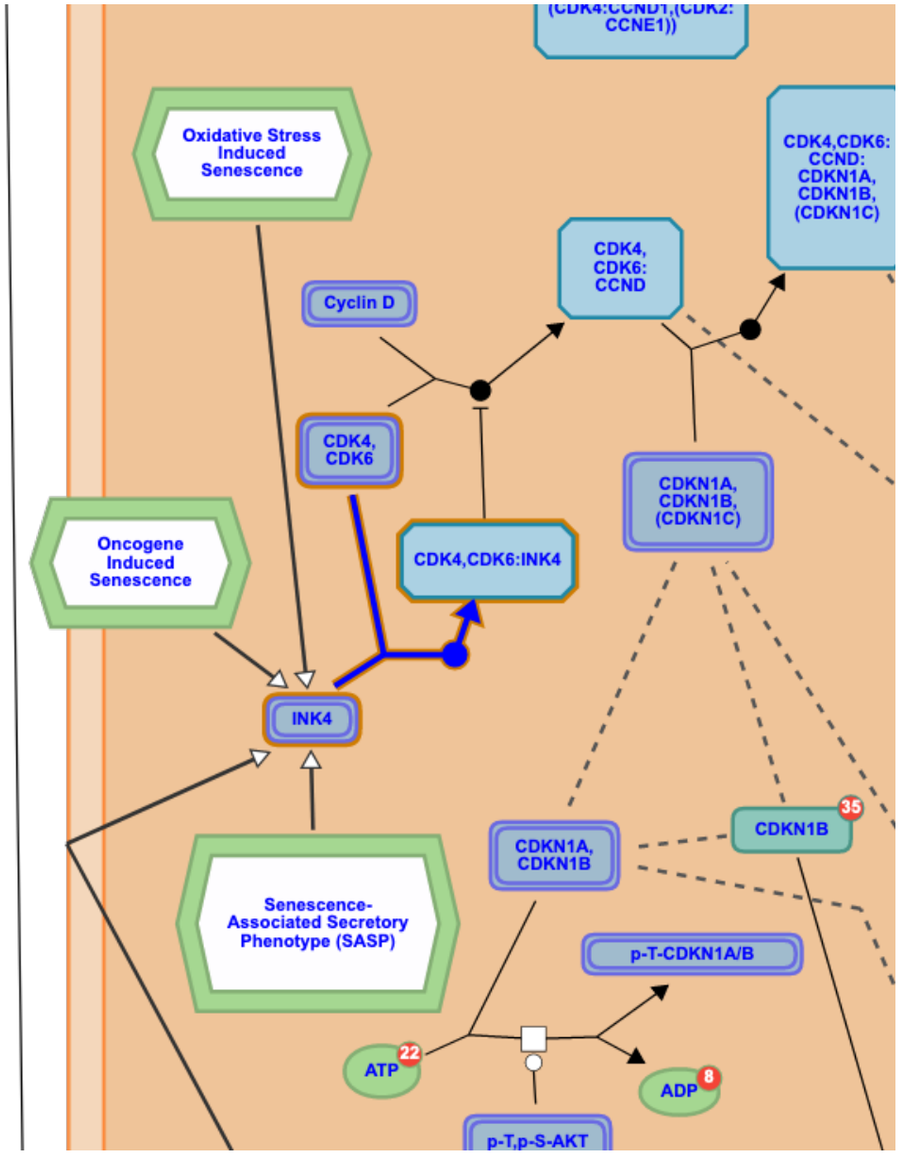
1. Introduction to network analysis
2. Terminology
3. Network inference
4. Key network properties
5. Community analysis
6. Visualization
7. Workshop

Community and functional analysis

What are modules?

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

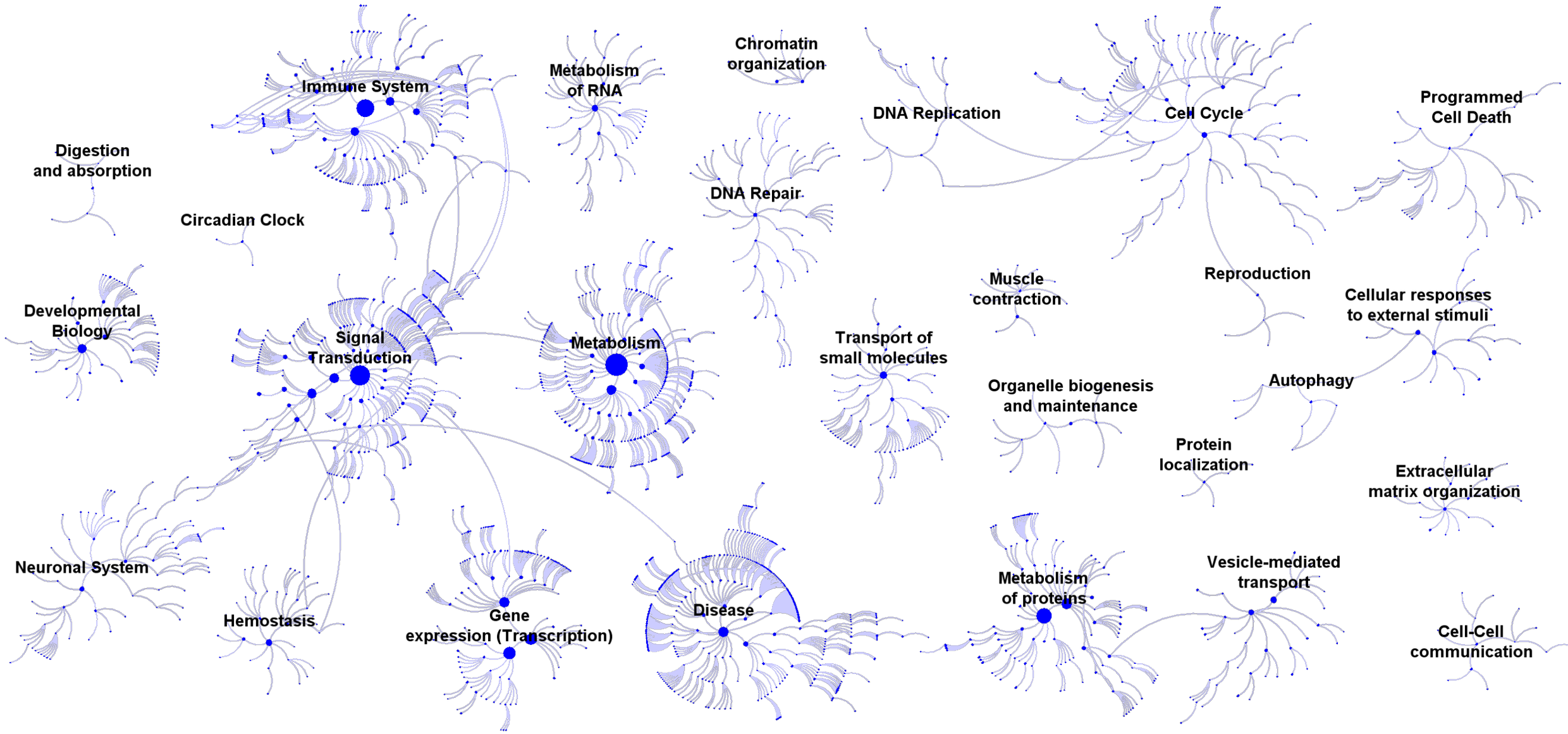
Protein complexes are physical modules



What are modules?

Pathway-associated proteins ***may*** represent functional modules

Gene Ontology



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

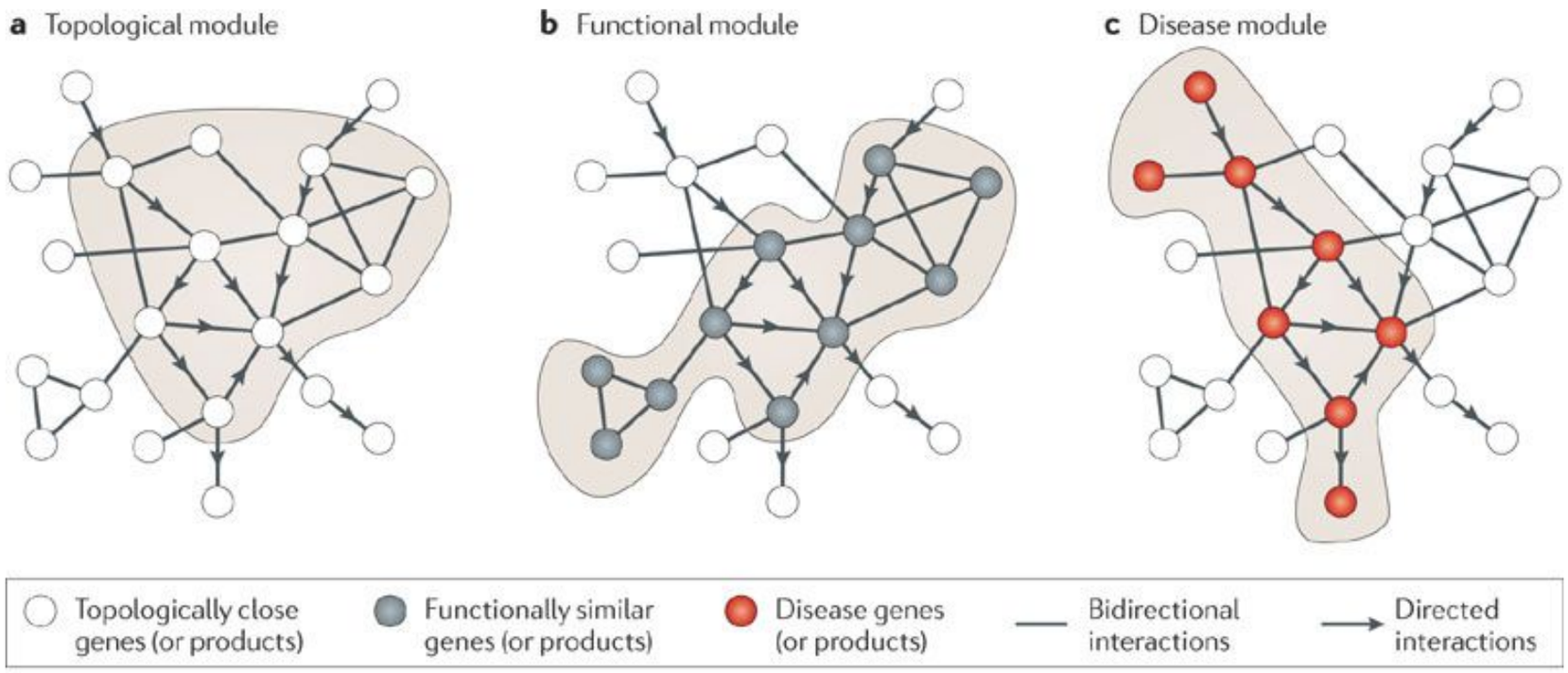
Disease: highly interconnected nodes associated with a disease response

Drug: highly interconnected nodes associated with a drug response

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

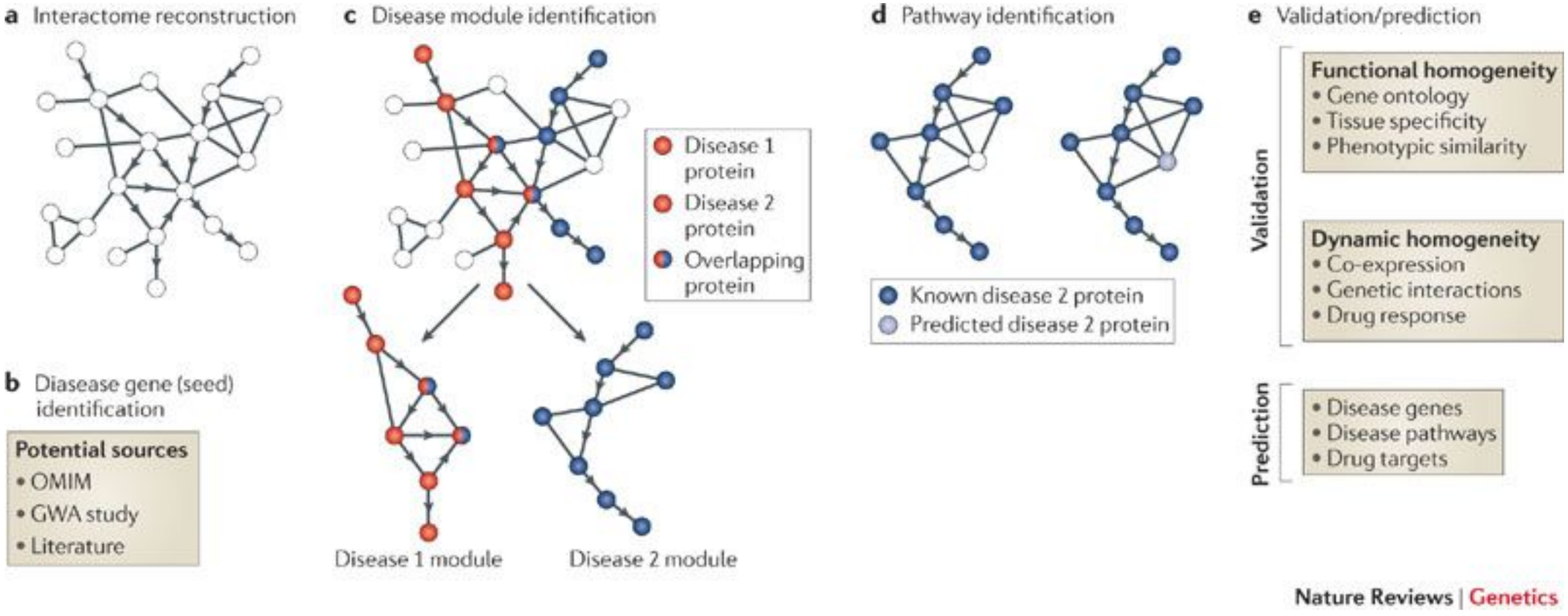
The challenge: identify and characterise modules

Moving from full network to modular characterisation

Different features (diseases, biological processes, etc.) may be associated with the same module

Prediction: *in silico*, relies on available knowledge

Validation: experimental responses



Barabasi 2011

Modularity

Modularity is a property of the 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

For a given partitioning of the network into individual groups s, compute

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

edges in group s

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

Modularity

Number of expected edges e if network is random, given the degree for its nodes

$$\frac{K_s}{2m}$$

← Sum(degrees of nodes in community)

← Total number of edges in community

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

edges in group s

Random network with
same number of nodes, edges

$$-1 < Q < 1$$

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

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

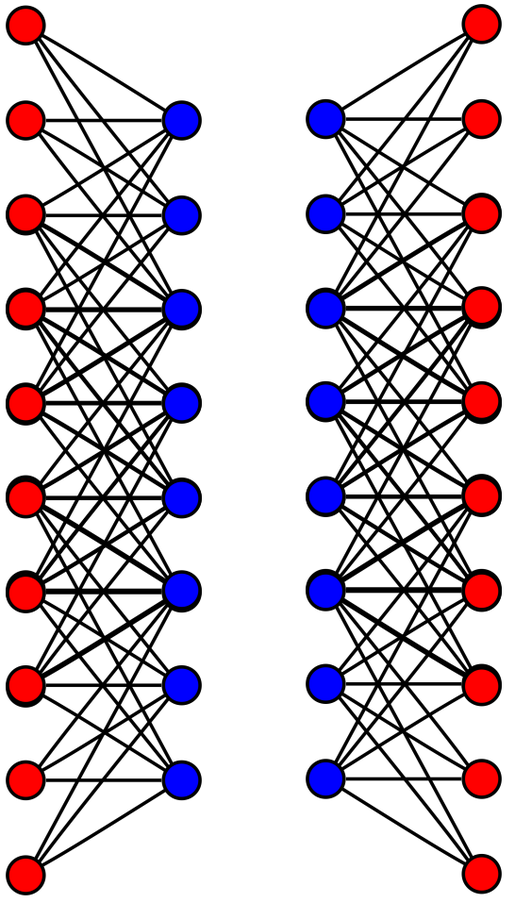
Q > 0.3 - 0.7 means significant community structure

Modularity

Modularity is different than **clustering coefficient**:

Graph composed of two bipartite complete subgraphs:

high Q but low connectivity (C)



Modules

A **module** (or **community**) is a set of nodes with a lot of **internal connections**, but **fewer external connections.**

How to identify modules? Maximise Q

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

Brute-force approach:

1. Start with 1 node/module
2. Compute distances between nodes
3. Join closest node
4. Re-compute distances between a 2n module and each 1n module
5. Join them if Q increases

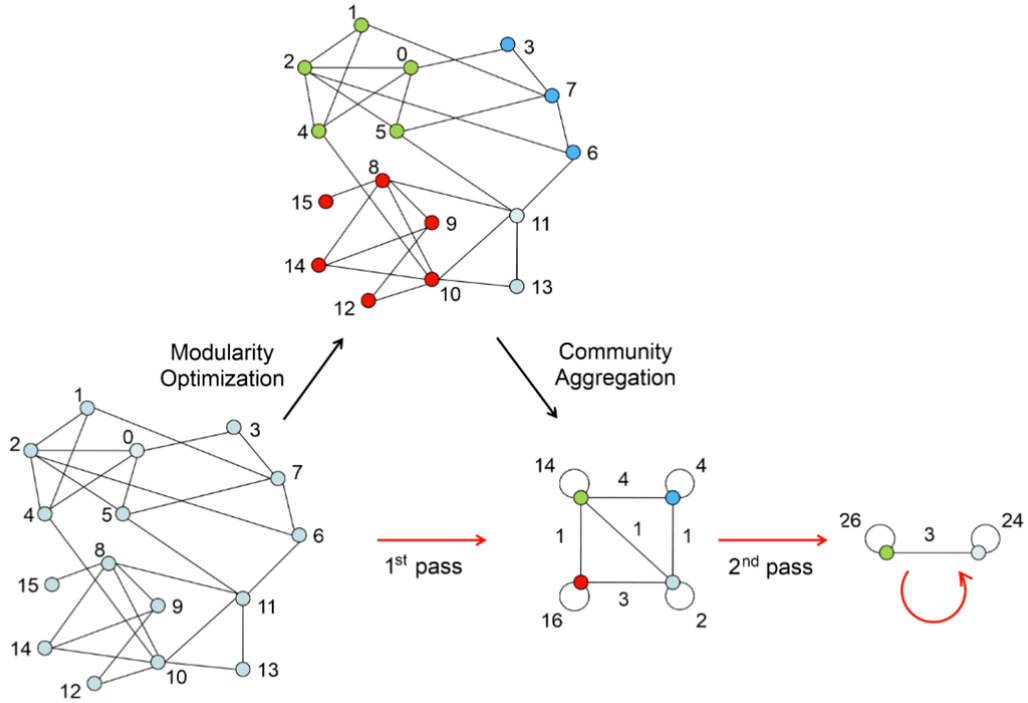
Module detection: Louvain algorithm

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
8. Repeat phase 1 on new network



Has <u>some known issues</u>:

- Communities may be internally disconnected
- Misses smaller communities

<u>Leiden algorithm</u>

Community characterisation

Clustering coefficient and degree distribution

Enrichment analysis

Assumption: community-associated features show coordinated (directly proportional) changes

Can significantly enriched biological processes serve as “validation”?

- Mutual feature associations may reinforce data characterisations not evident by individual features
- …or need of further network curation based on top biological terms

GSEA calculates overrepresentation by comparison of gene-level statistics against those of the gene-set, considering sample and feature permutation

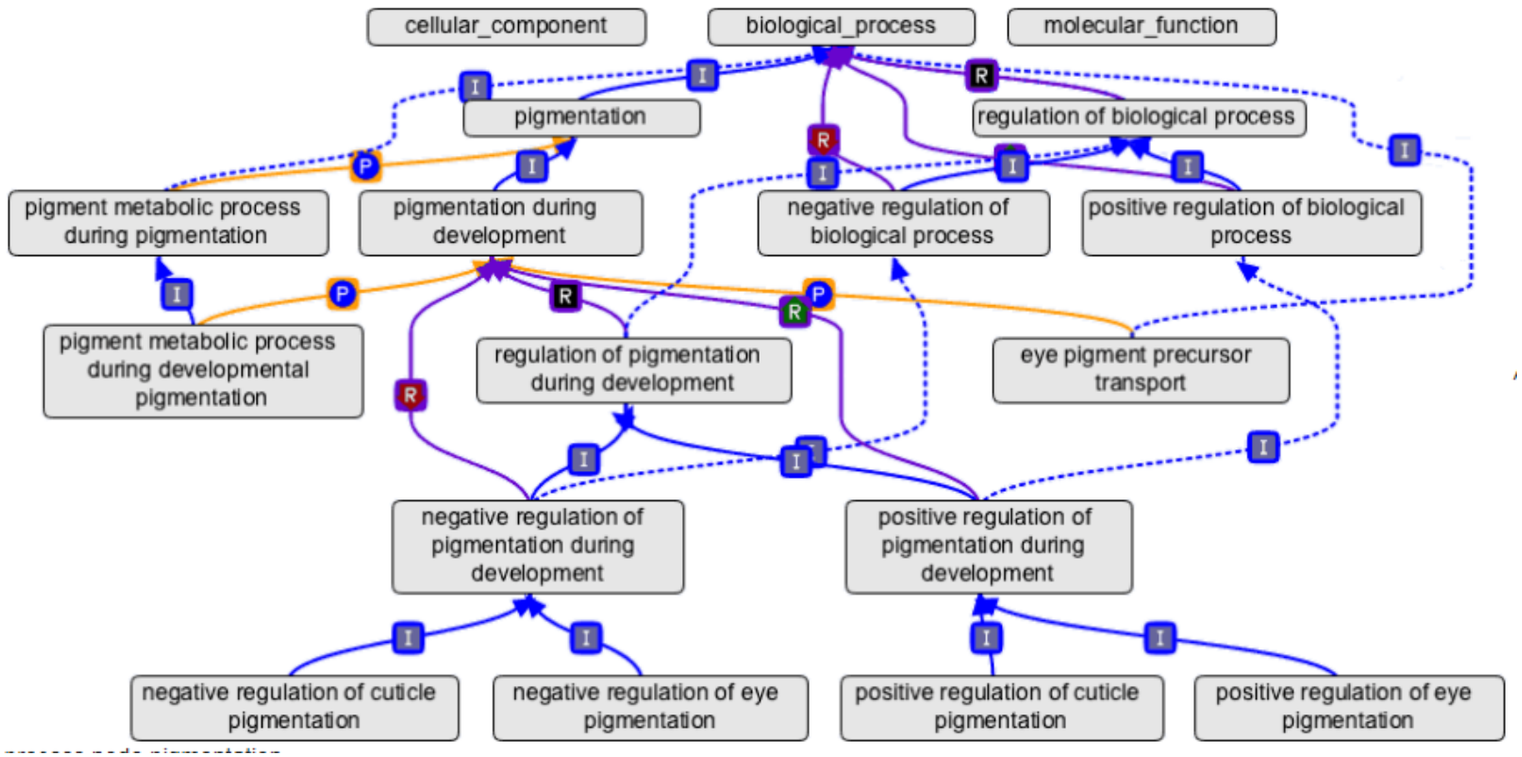
New to Gene Set Enrichment Analysis? See: Mooney 2015

Enrichment analysis

GO-terms, pathways, subcellular location, TF-targets, disease, drug, other?

Tests for significant overlap between groups

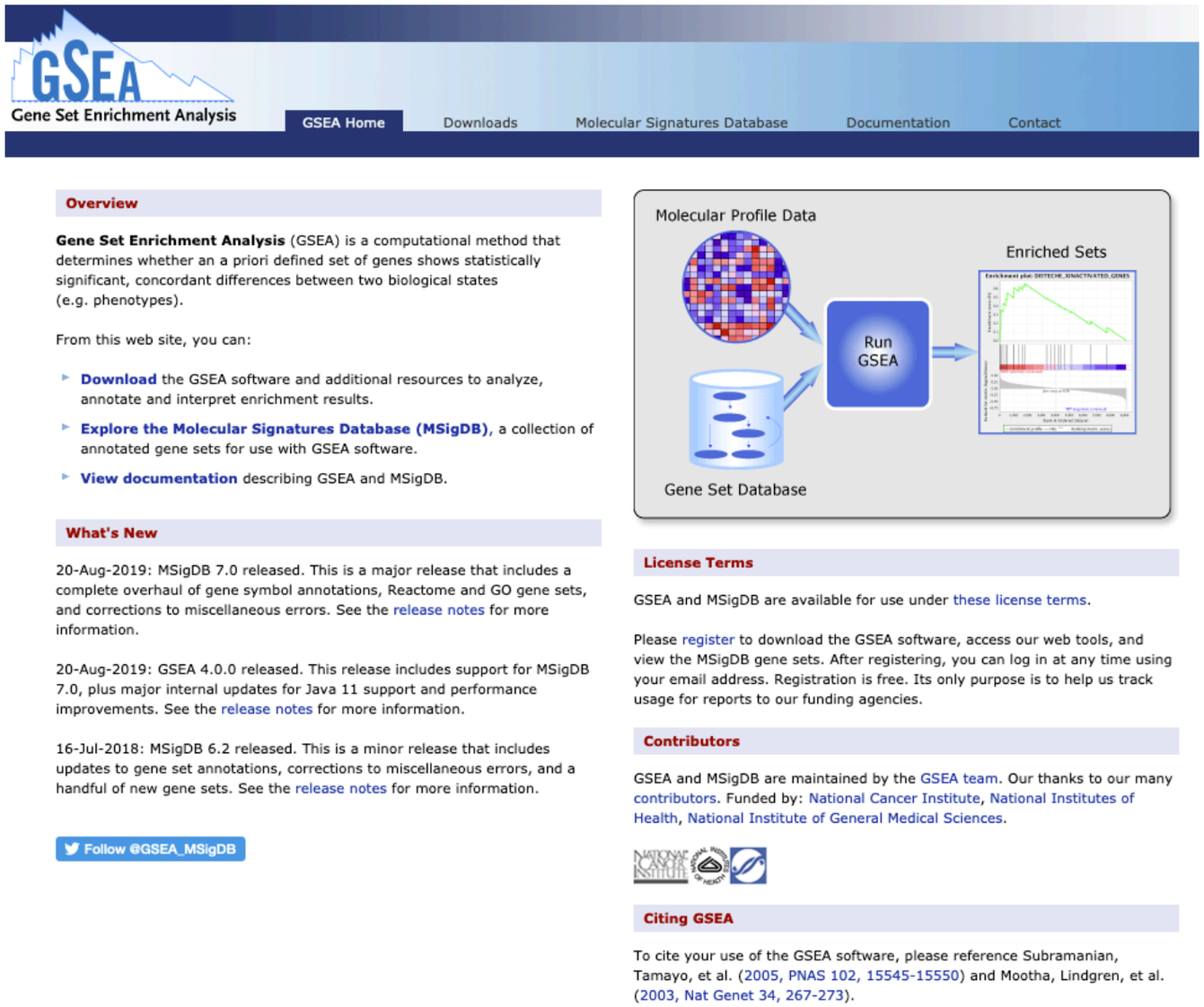
Some biological processes may have no biological meaning in your analysis



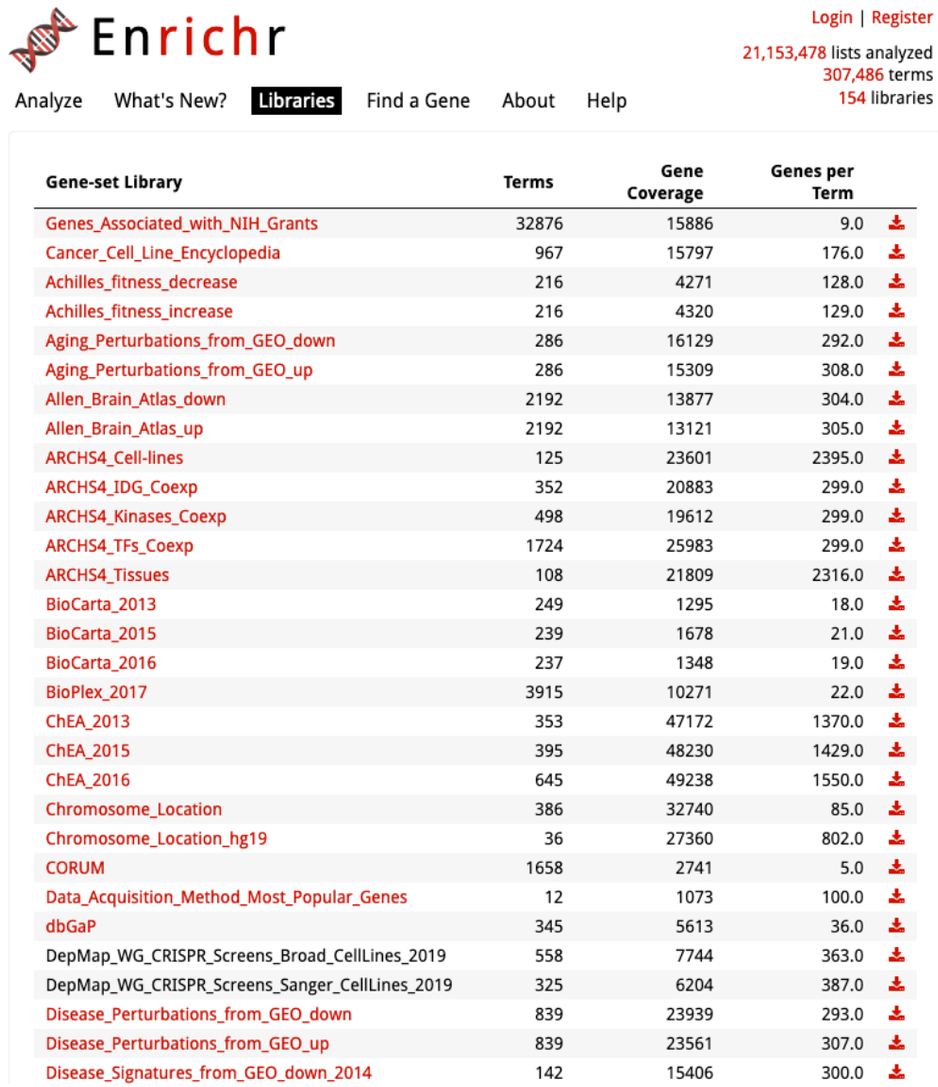
http://geneontology.org/

Enrichment analysis

MSigDB



Enrichr



Gene-set Library	Terms	Gene Coverage	Genes per Term
Genes_Associated_with_NIH_Grants	32876	15886	9.0
Cancer_Cell_Line_Encyclopedia	967	15797	176.0
Achilles_fitness_decrease	216	4271	128.0
Achilles_fitness_increase	216	4320	129.0
Aging_Perturbations_from_GEO_down	286	16129	292.0
Aging_Perturbations_from_GEO_up	286	15309	308.0
Allen_Brain_Atlas_down	2192	13877	304.0
Allen_Brain_Atlas_up	2192	13121	305.0
ARCHS4_Cell-lines	125	23601	2395.0
ARCHS4_IDG_Coexp	352	20883	299.0
ARCHS4_Kinases_Coexp	498	19612	299.0
ARCHS4_TFs_Coexp	1724	25983	299.0
ARCHS4_Tissues	108	21809	2316.0
BioCarta_2013	249	1295	18.0
BioCarta_2015	239	1678	21.0
BioCarta_2016	237	1348	19.0
BioPlex_2017	3915	10271	22.0
ChEA_2013	353	47172	1370.0
ChEA_2015	395	48230	1429.0
ChEA_2016	645	49238	1550.0
Chromosome_Location	386	32740	85.0
Chromosome_Location_hg19	36	27360	802.0
CORUM	1658	2741	5.0
Data_Acquisition_Method_Most_Popular_Genes	12	1073	100.0
dbGaP	345	5613	36.0
DepMap_WG_CRISPR_Screens_Broad_CellLines_2019	558	7744	363.0
DepMap_WG_CRISPR_Screens_Sanger_CellLines_2019	325	6204	387.0
Disease_Perturbations_from_GEO_down	839	23939	293.0
Disease_Perturbations_from_GEO_up	839	23561	307.0
Disease_Signatures_from_GEO_down_2014	142	15406	300.0

Enrichment analysis

Important databases with gene-sets:

- MSigDB (gene)
- Enrichr (gene)
- KEGG (metabolite, gene)
- DIANA (miRNA)
- MetaboAnalyst (metabolite)
- DAVID (web)
- Reactome (web)

Creating custom sets and joint sets

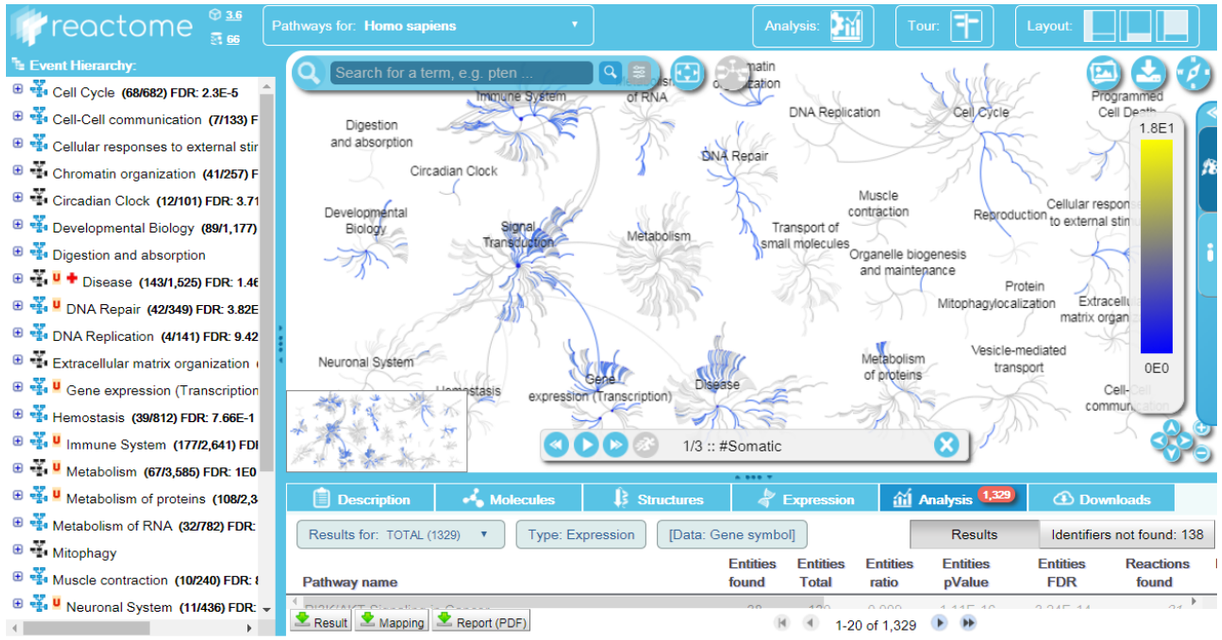
Mapping your data to common IDs

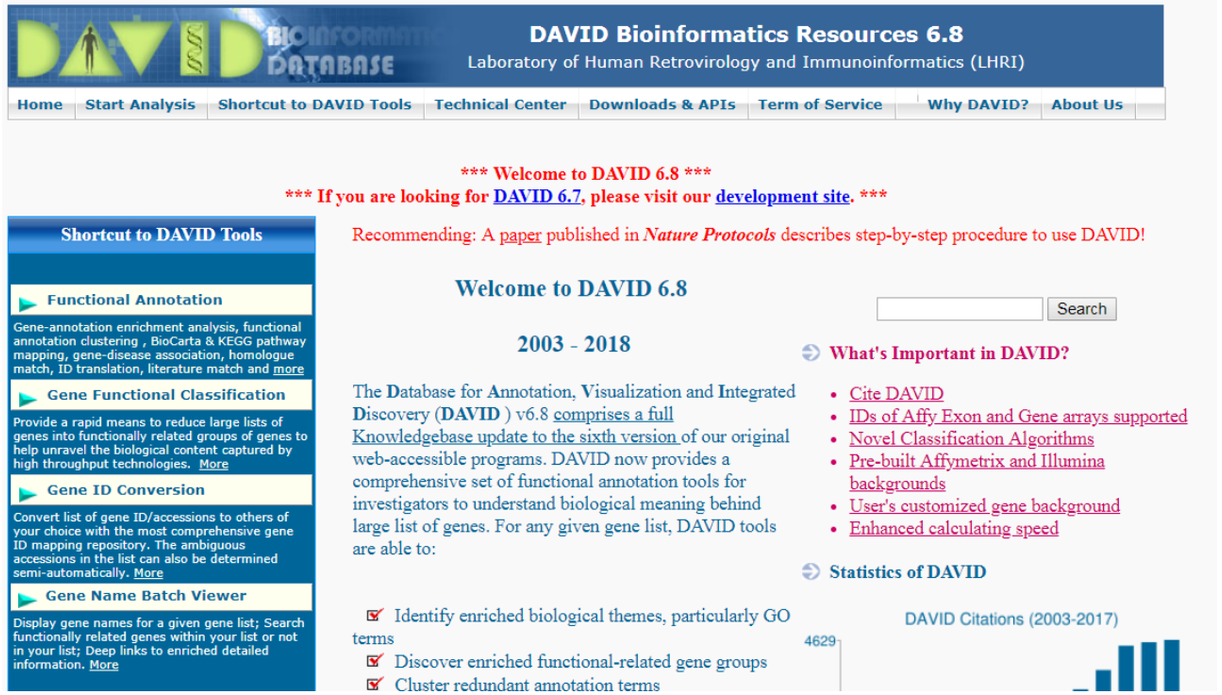
- Easy for genes and proteins: use DAVID, Biomart, or MyGene (in Python or R)
- Hard for other types

Tools for Enrichment analysis

Popular tools for GSEA:

- (most tools above)
- PIANO (highly recommended in R)
- Cytoscape (BINGO plugin)





References as hyperlinks

Considerations for enrichment analysis

What background to consider?

Direction of change

- DAVID, Enrichr, and others do not consider direction
- PIANO takes into account gene-level statistics including directions

Bonus: GSEA in PIANO

distinct-directional: takes direction of change. gene sets with both significantly up- and down-regulated will cancel out.

non-directional: disregards direction and uses absolute values of gene-level statistics

mixed-directional: considers up- and down-regulated subsets separately. Important when

```
[1] -4.0 -3.0  2.0  3.5
```

Gene-level statistics

```
> mean(c(-4, -3, 2.5, 4.5))
[1] 0
```

Distinct-directional

```
> mean(abs(c(-4, -3, 2.5, 4.5)))
[1] 3.5
```

non-directional

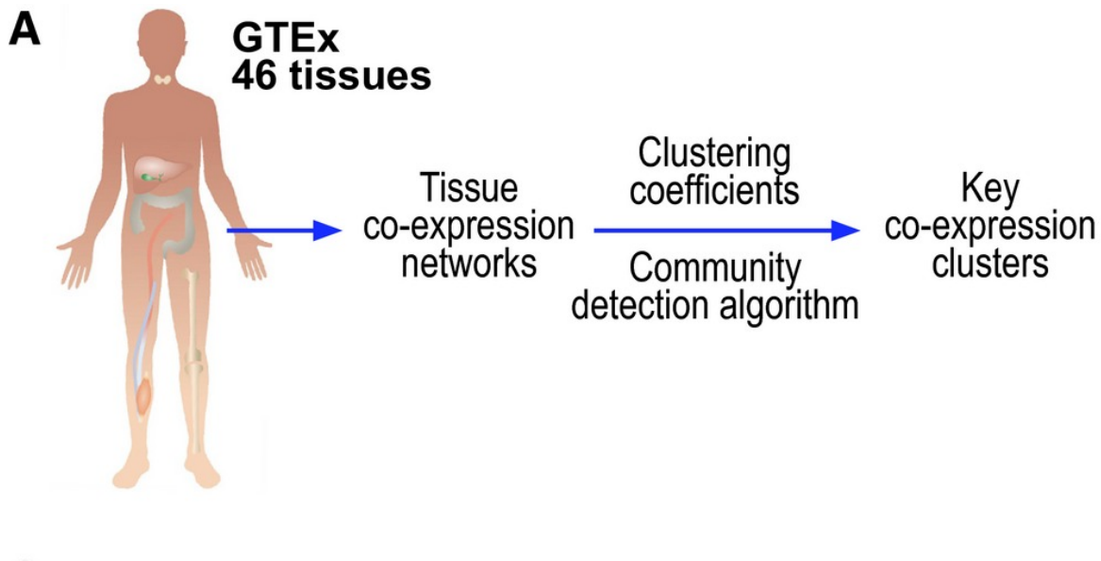
```
> mean(abs(c(-4, -3)))
[1] 3.5
> mean(abs(c(2.5, 4.5)))
[1] 3.5
```

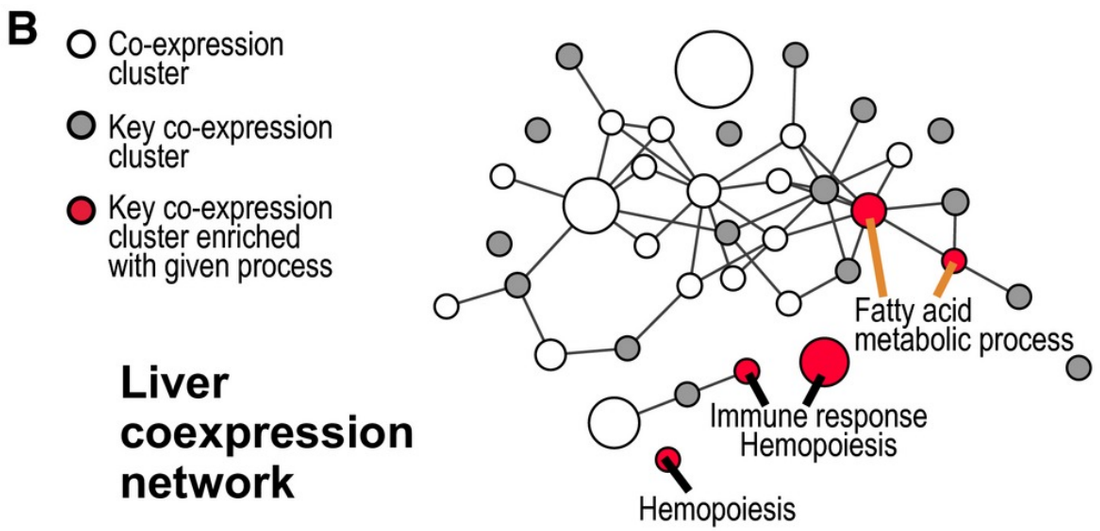
mixed-directional

```
> gs$stats
```

	Name	Value
1	Genes (tot)	243.00000
2	Stat (dist.dir.up)	0.47973
3	p (dist.dir.up)	0.03510
4	p adj (dist.dir.up)	0.40038
5	Stat (dist.dir.dn)	0.52027
6	p (dist.dir.dn)	0.96490
7	p adj (dist.dir.dn)	1.00000
8	Stat (non-dir)	0.15741
9	p (non-dir)	0.00000
10	p adj (non-dir)	0.00000
11	Genes (up)	127.00000
12	Stat (mix.dir.up)	0.15511
13	p (mix.dir.up)	0.00130

Outcome

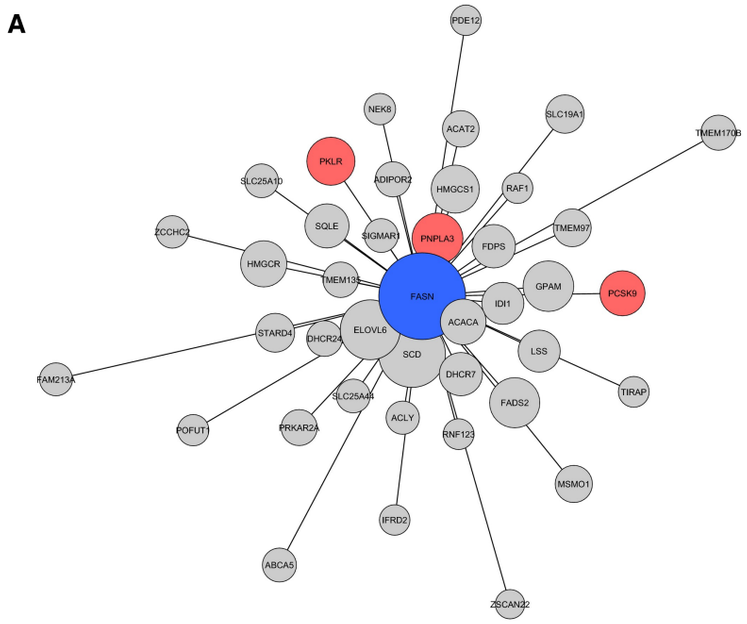




GO BPs from MSigDB

Network analyses identifies key stratifying genes

40-NN of FASN



Visualization

METABOLIC ATLAS

THE ATLAS FOR EXPLORATION OF METABOLISM

Welcome

GEM Browser

Interaction Partners

Map Viewer

Search

Export

Analyze

Metabolic Atlas integrates open source genome-scale metabolic models (GEMs) of human and yeast for easy browsing and analysis. It also contains many more genome scale metabolic models constructed by our organization.

Detailed biochemical information is provided for individual model components, such as reactions, metabolites, and genes. These components are also associated with standard identifiers, facilitating integration with external databases, such as the Human Protein Atlas.

Article under consideration

reactome

About Content Docs Tools Community Download

Find Reactions, Proteins and Pathways

e.g. O95631, NTN1, signaling by EGFR, glucose Go!

Pathway Browser

Visualize and interact with Reactome biological pathways

Analyze Data

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

ReactomeFIViz

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

Documentation

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

USE REACTOME GRAPH DATABASE IN YOUR PROJECT

LEARN MORE

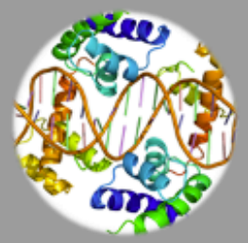
Omics Net - a network analytics platform for multi-omics integration

Home | Overview | FAQs | Tutorials | Gallery | About | Updates

Click an icon below to start

Genes/Proteins

- ID types: Entrez, Ensembl Gene/Transcript, Uniprot and official gene symbol

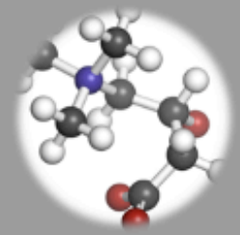


Transcription factors

- ID types: Entrez, Ensembl Gene/Transcript, Uniprot and official gene symbol

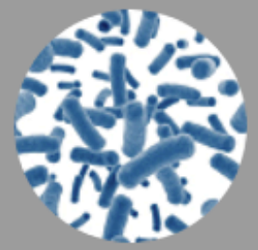
miRNAs

- ID types: miRBase Accession and miRBase ID



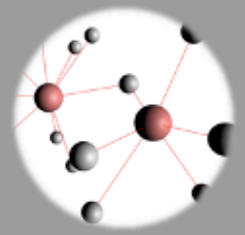
Metabolites

- ID types: KEGG ID, HMDB ID, PubChem and CHEBI ID



Microbiome

- ID types: KEGG Orthology (KO)



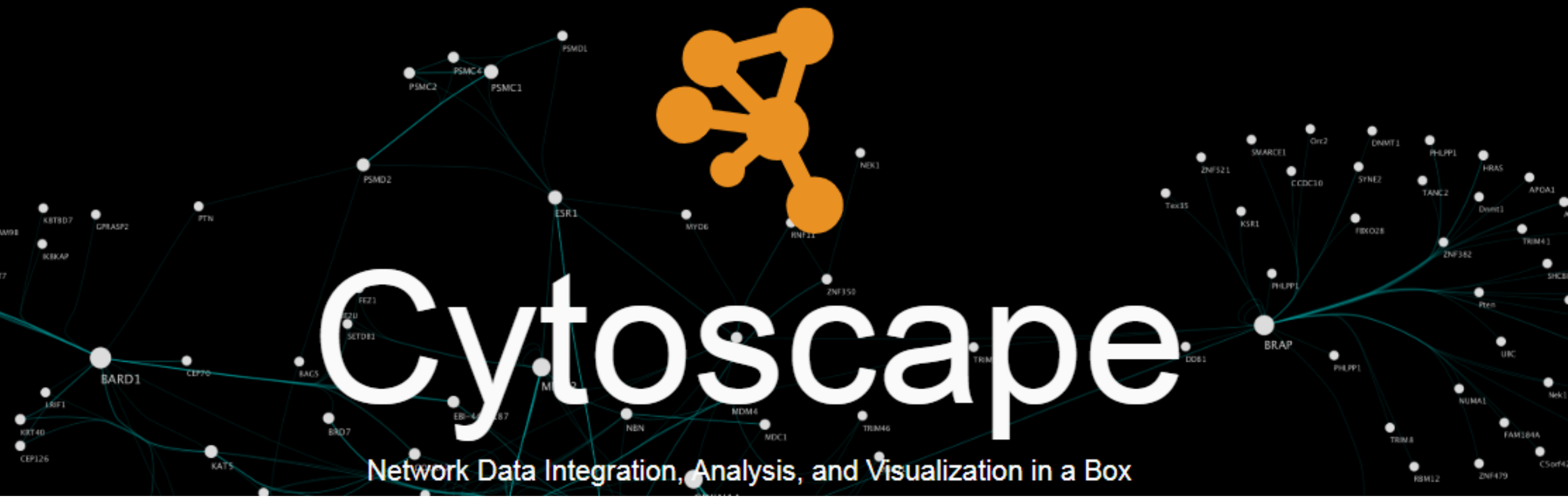
Graph file

- Formats: .sif, .graphml, .json or .txt (edge list)

Proceed

Project Objectives

OmicsNet has been developed to address three needs: 1) systems analysis of a single list of molecules; 2) integrative analysis of multiple lists of different types of molecules; and 3) intuitive web-based 2D/3D visualization. OmicsNet supports four types of molecular interactions - PPI, TF-gene, miRNA-gene and metabolite-protein. The 3D network visualization was implemented based on the innovative WebGL technology. Users can perform various style customization, enrichment analysis, targeted interactor search, and network toplology analysis for hypothesis generation and systems-level insights.



Input: coexpression_matrix.tsv

Tools > Network analyzer: Centrality

Node configuration: Centrality; Community

Edge weight configuration: Correlation

Module detection: ClusterMaker

GSEA: BINGO

Workshop objective:

Build and analyze a gene expression network