Community analysis and visualisation

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

rui.benfeitas@scilifelab.se

Overview

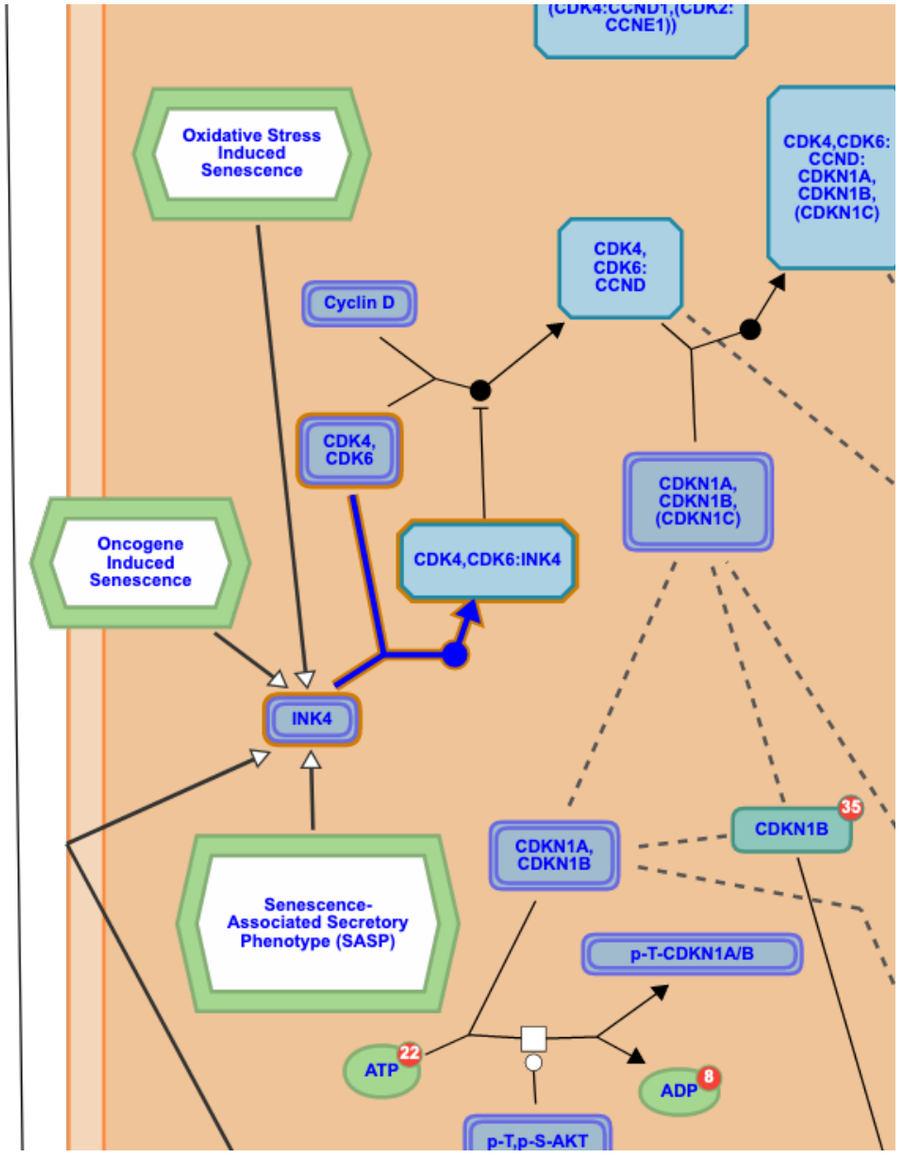
1. Introduction to network analysis
2. Terminology
3. Network inference
4. Key network properties
5. Community analysis

Community and functional analysis

What are modules?

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

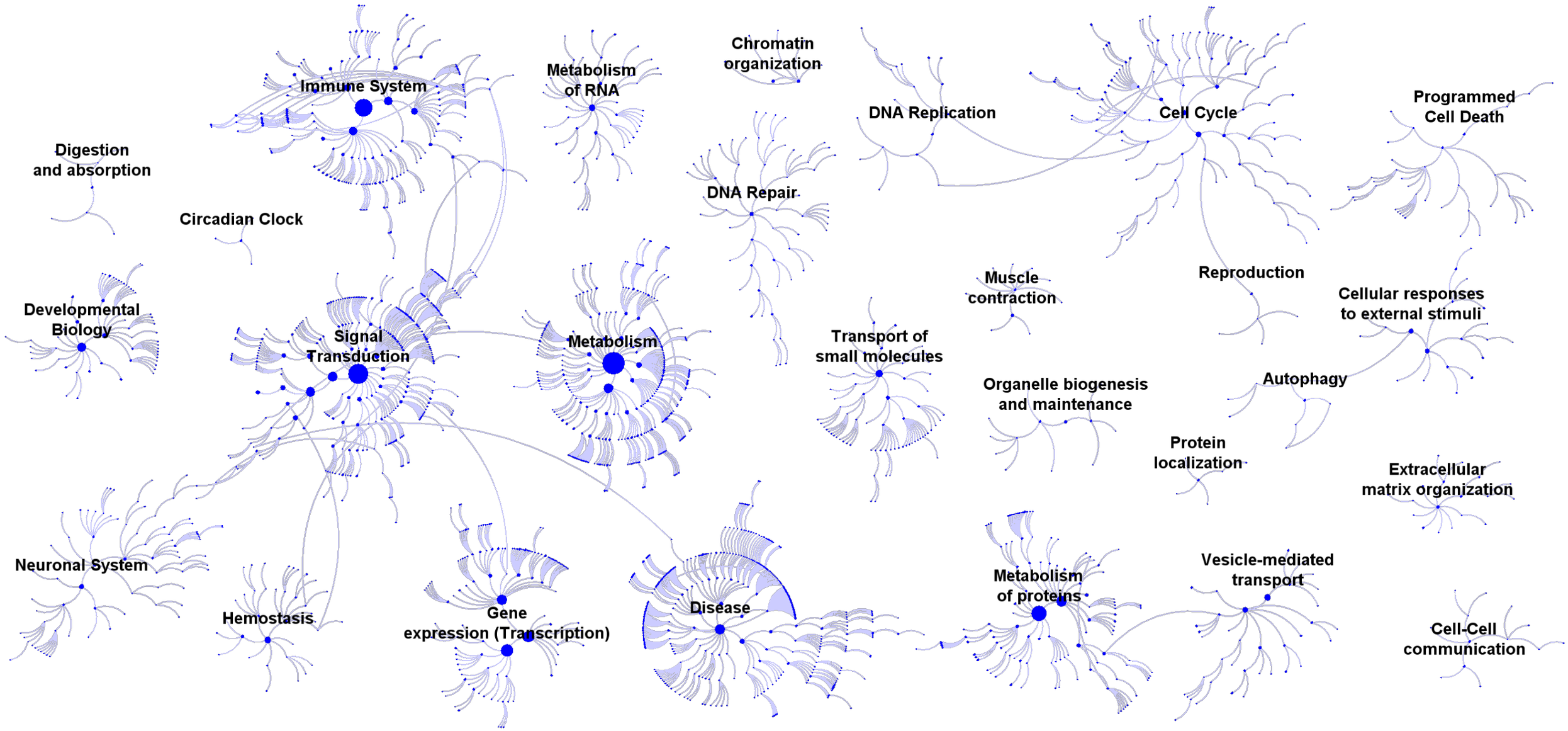
Protein complexes are physical modules



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

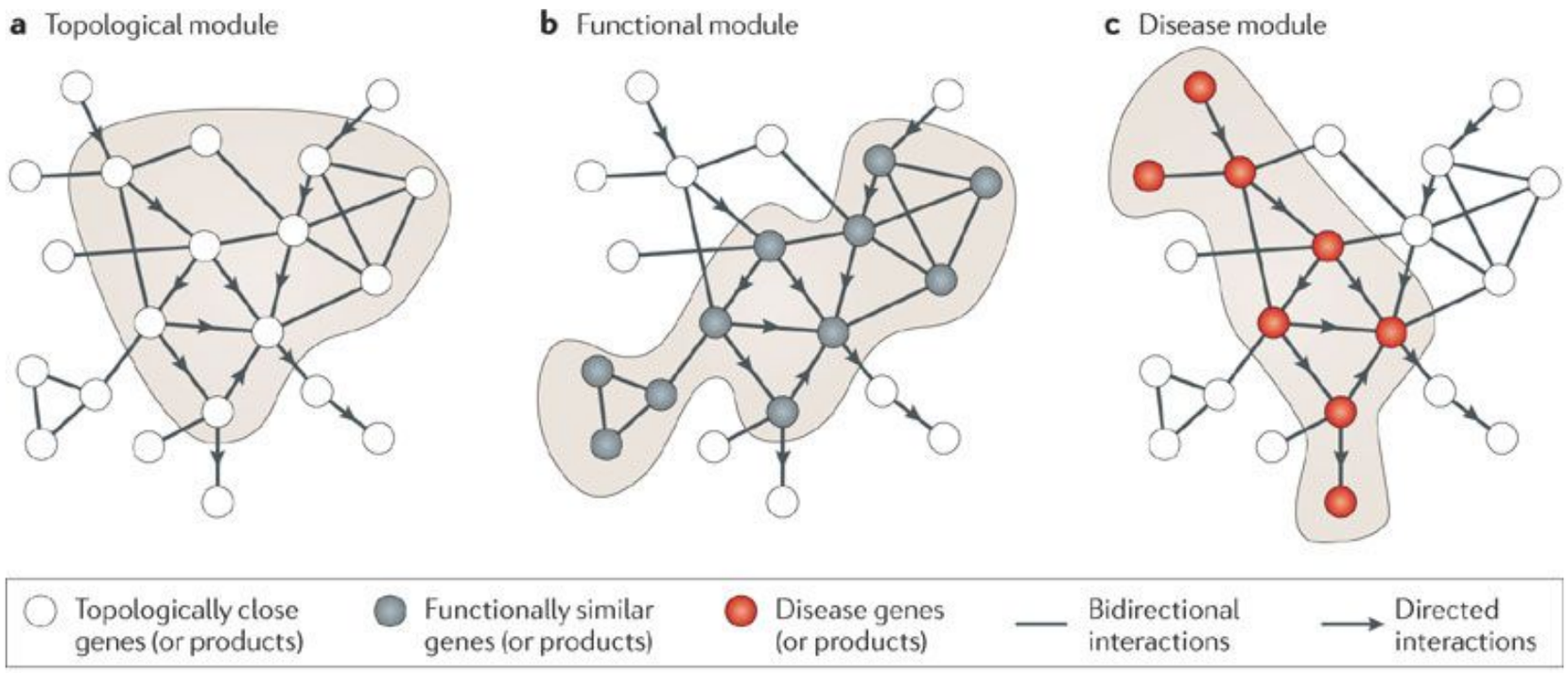
Disease: highly interconnected nodes associated with a disease response

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

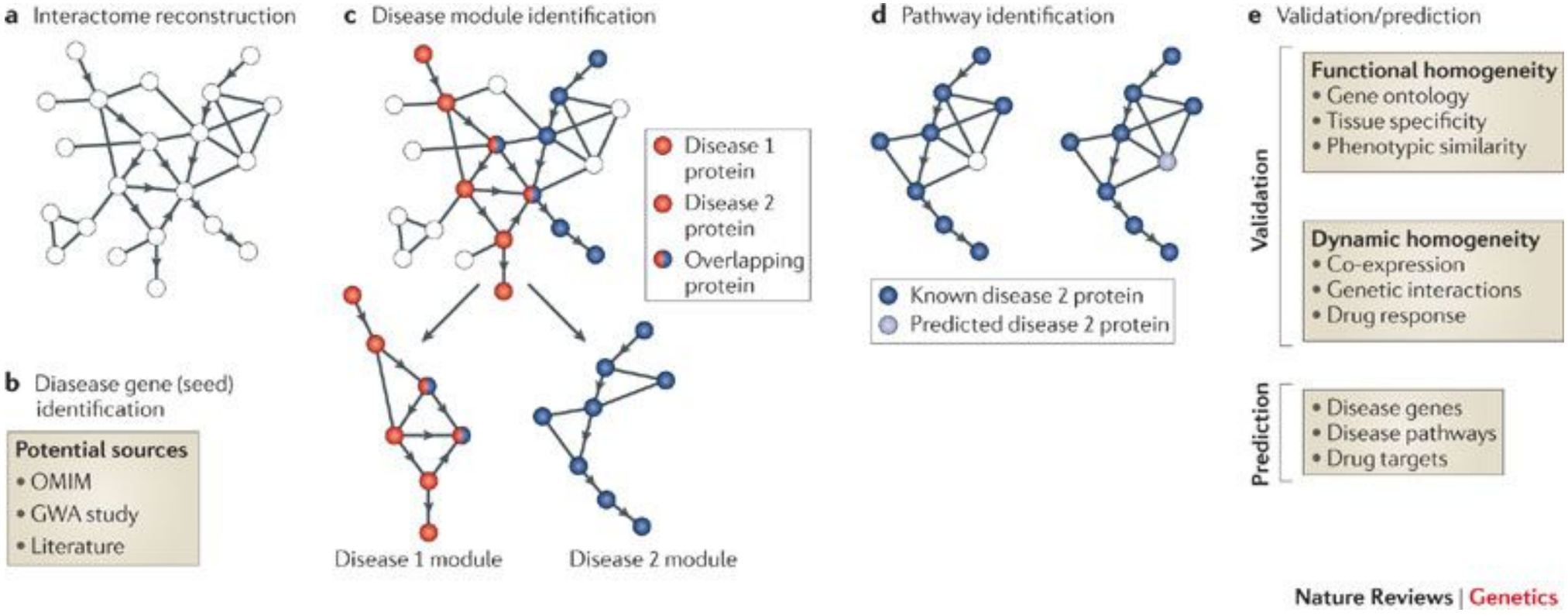
The challenge: identify and characterise modules

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



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

$$-1 < Q < 1$$

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

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

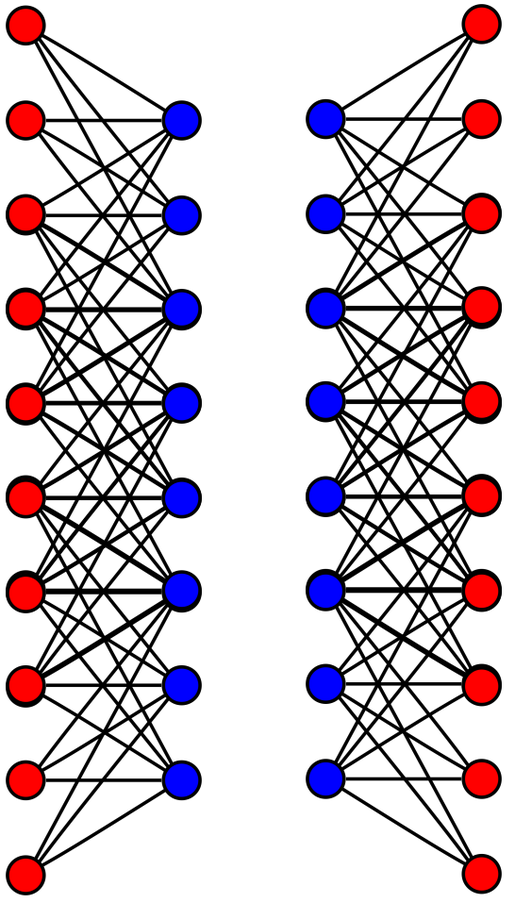
Q > 0.3 - 0.7 means significant community structure

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

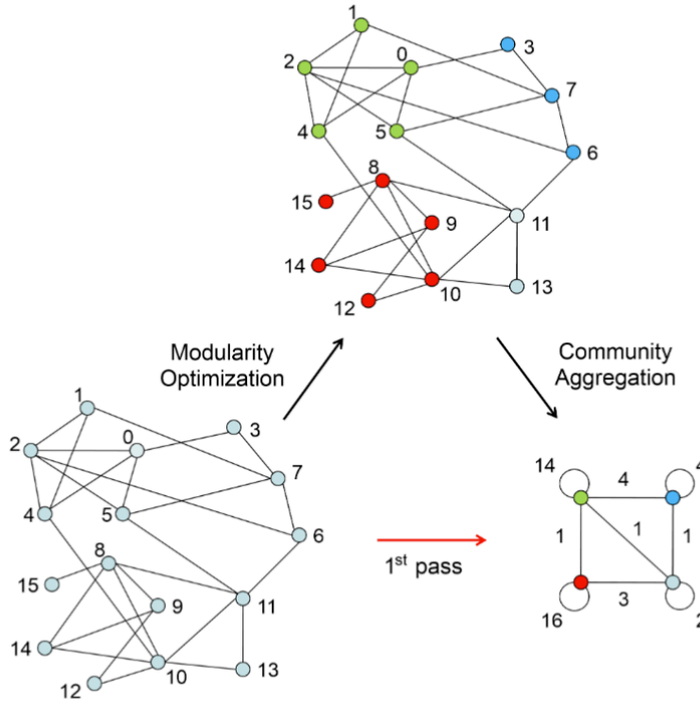
Module detection: Louvain algorithm

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 some known issues:

- Communities may be internally disconnected
- Misses smaller communities

Leiden algorithm

Community characterisation

Clustering coefficient and degree distribution

Enrichment analysis

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

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

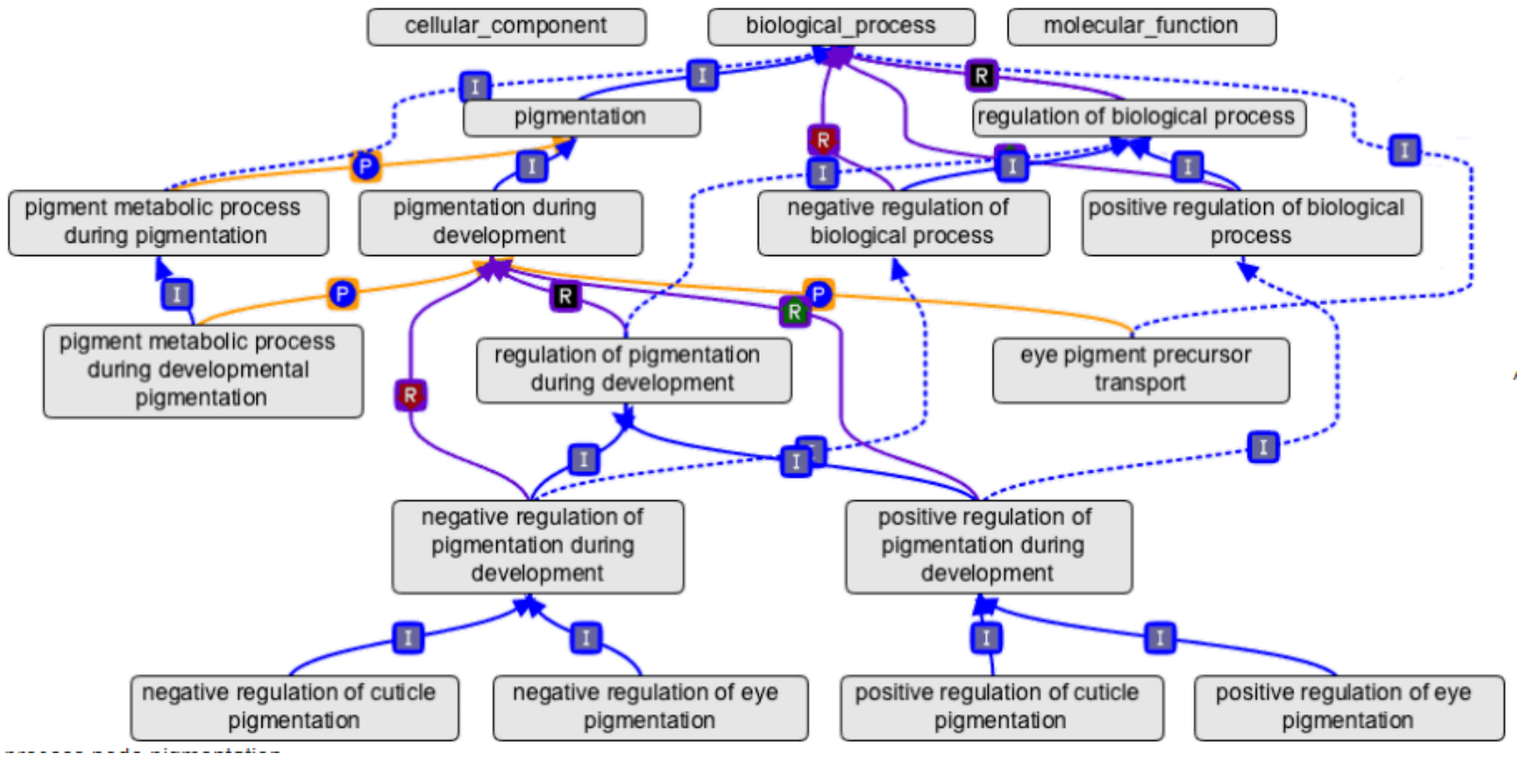
Mooney 2015

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

GSEA
Gene Set Enrichment Analysis

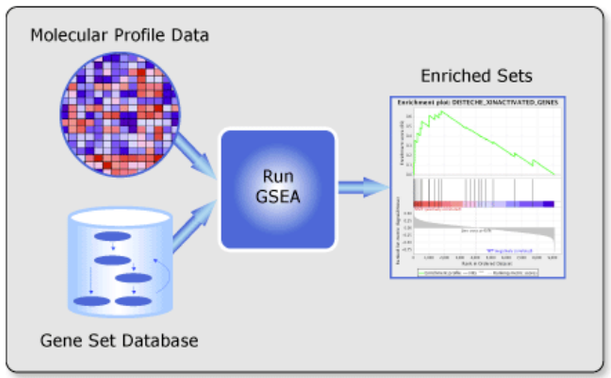
GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

- **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- **View documentation** describing GSEA and MSigDB.



What's New

20-Aug-2019: MSigDB 7.0 released. This is a major release that includes a complete overhaul of gene symbol annotations, Reactome and GO gene sets, and corrections to miscellaneous errors. See the release notes for more information.

20-Aug-2019: GSEA 4.0.0 released. This release includes support for MSigDB 7.0, plus major internal updates for Java 11 support and performance improvements. See the release notes for more information.

16-Jul-2018: MSigDB 6.2 released. This is a minor release that includes updates to gene set annotations, corrections to miscellaneous errors, and a handful of new gene sets. See the release notes for more information.

Follow @GSEA_MSigDB

License Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software, access our web tools, and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Contributors

GSEA and MSigDB are maintained by the GSEA team. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Citing GSEA

To cite your use of the GSEA software, please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

Enrichr

Enrichr

Login | Register

21,153,478 lists analyzed
307,486 terms
154 libraries

Analyze | What's New? | Libraries | Find a Gene | About | Help

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

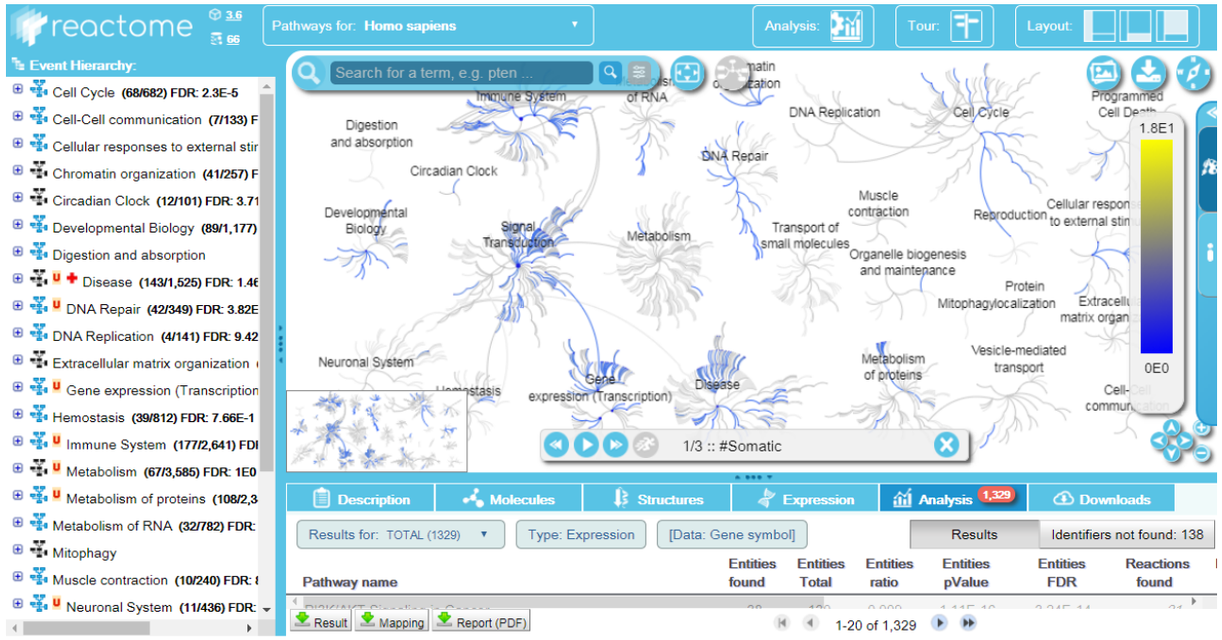
Mapping your data to common IDs

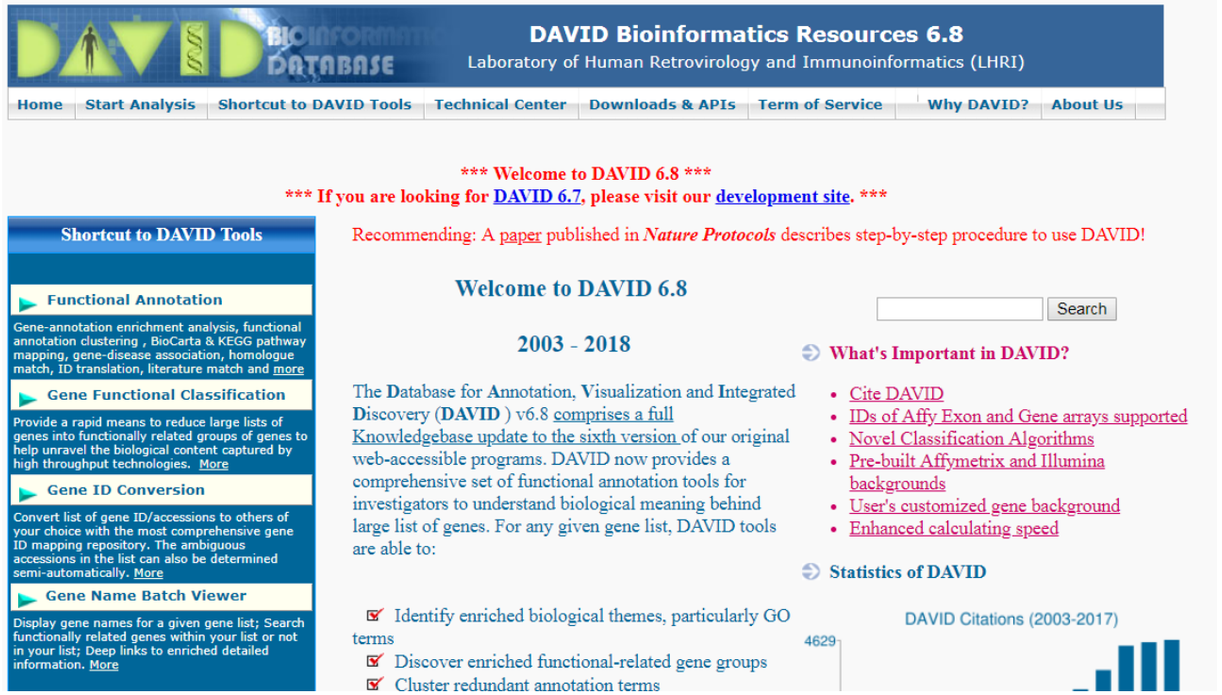
- Easy for genes and proteins: use DAVID, Biomart, or MyGene (in Python or R)
- Hard for other types

Tools for Enrichment analysis

Popular tools for GSEA:

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





References as hyperlinks

Tools for Enrichment analysis

STRING

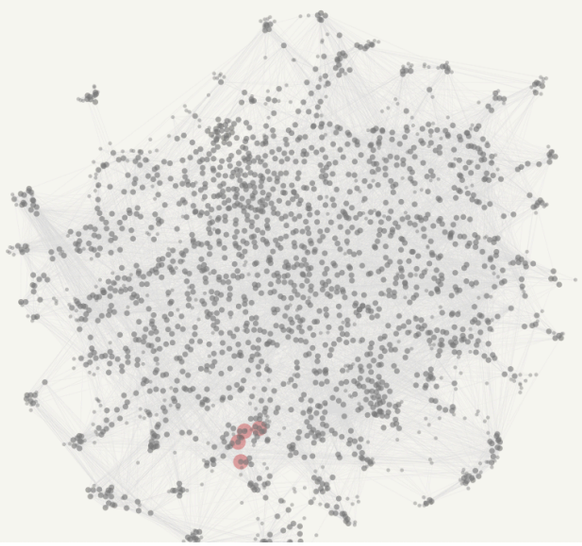
Search Download Help My Data

Your input data

GO:0000083 — 5 genes in your input

1: PKP1	Plakophilin-1	-8.326649
2: CDSN	Corneodesmosin	-8.130157
3: SERPINB5	Serpin B5	-8.065760
4: DSC1	Desmocollin-1	-7.917077
5: DSG1	Desmoglein-1	-7.838328
6: CALML5	Calmodulin-like protein 5	-7.706114
7: ZNF750	Zinc finger protein 750	-7.527767
8: SERPINB7	Serpin B7	-7.497837
9: LCE2B	Late cornified envelope protein 2E	-7.467221
10: CHP2	Calcineurin B homologous protein	-7.423878
11: GJB6	Gap junction beta-6 protein	-7.301189
12: COL17A1	Collagen alpha-1(XVII) chain	-7.263660
13: C19orf33	Immortalization up-regulated prot	-7.195207
14: SBSN	Suprabasin	-7.140458
15: LY6D	Lymphocyte antigen 6D	-7.056120
16: TRIM29	Tripartite motif-containing protein	-7.034785
17: FLG	Filaggrin	-7.031575
18: CRCT1	Cysteine rich C-terminal 1	-7.022690
19: KRT15	Keratin, type I cytoskeletal 15	-6.867025

Full proteome network



Your detected functional enrichments

Biological Process (GO)

GO-term	description	enrichment score	direction	pathway size	false discovery rate
GO:0000083	regulation of transcription involved in G1/S transition of mitotic cell cycle	5.93242	bottom of input	29	0.0044 (afc)
GO:0007094	mitotic spindle assembly checkpoint	5.84439	bottom of input	21	0.0055 (afc)
GO:0030071	regulation of mitotic metaphase/anaphase transition	5.83409	bottom of input	49	0.00045 (afc)
GO:0051983	regulation of chromosome segregation	5.76051	bottom of input	97	0.00018 (afc)
GO:0051784	negative regulation of nuclear division	5.68192	bottom of input	47	0.0023 (afc)

(more ...)

Molecular Function (GO)

GO-term	description	enrichment score	direction	pathway size	false discovery rate
GO:0030280	structural constituent of epidermis	6.53875	top of input	14	0.00043 (afc)
GO:0005198	structural molecule activity	2.84694	top of input	679	0.00043 (ks)
GO:0032559	adenyl ribonucleotide binding	2.1506	bottom of input	1514	0.00046 (ks)
GO:0030554	adenyl nucleotide binding	2.04171	bottom of input	1524	0.00071 (ks)
GO:0005524	ATP binding	2.03433	bottom of input	1462	0.0018 (ks)

(more ...)

Cellular Component (GO)

GO-term	description	enrichment score	direction	pathway size	false discovery rate
GO:0001533	cornified envelope	6.04702	top of input	64	4.13e-12 (ks)
GO:0009925	basal plasma membrane	5.97917	top of input	29	0.0080 (afc)
GO:0097209	epidermal lamellar body	5.83055	top of input	4	0.0093 (afc)
GO:0000794	condensed nuclear chromosome	5.66137	bottom of input	96	0.00092 (afc)
GO:0030057	desmosome	5.59763	top of input	25	1.63e-06 (afc)

(more ...)

Reference publications

publication	(year) title	enrichment score	direction	pathway size	false discovery rate
PMID:27426474	(2017) Clinical, microscopic and microbial characterization of exfoliative superficial pyoderma-associated epid...	9.43445	top of input	4	0.00018 (afc)
PMID:25496350	(2015) Expression patterns of superficial epidermal adhesion molecules in an experimental dog model of acut...	9.43445	top of input	4	0.00018 (afc)
PMID:24324345	(2013) Epidemiology of 'fragile skin': results from a survey of different skin types.	9.43445	top of input	3	0.00018 (afc)
PMID:23810772	(2013) Aberrant distribution patterns of corneodesmosomal components of tape-stripped corneocytes in atopi...	9.43445	top of input	3	0.00018 (afc)
PMID:23378711	(2012) Surgical Therapy by Sandwich Transplantation using a Dermal Collagen-Elastin Matrix and Full Thickne...	9.43445	top of input	4	0.00018 (afc)

(more ...)

local STRING network cluster

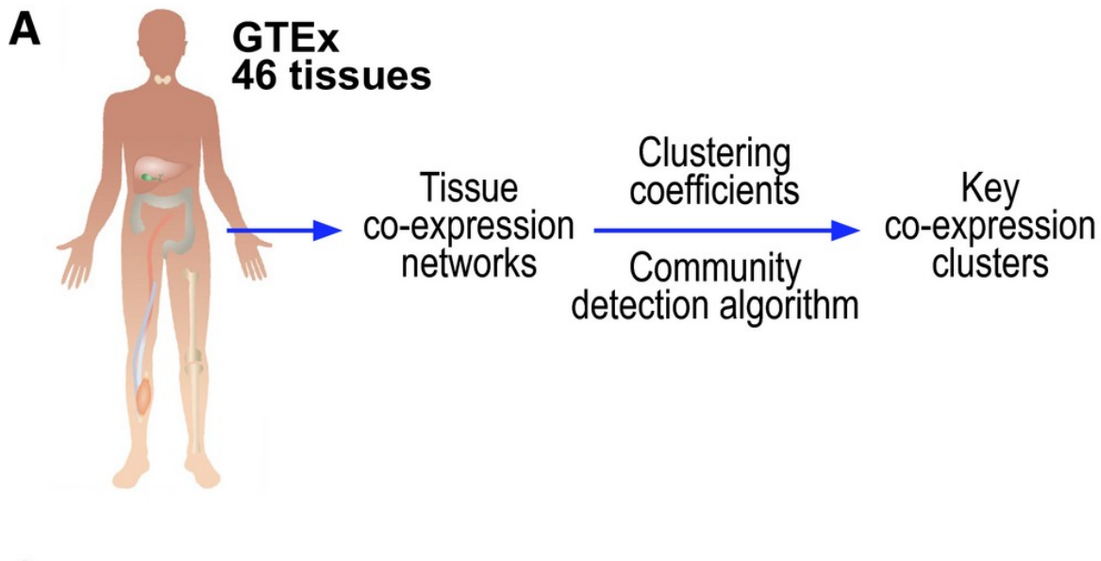
Considerations for enrichment analysis

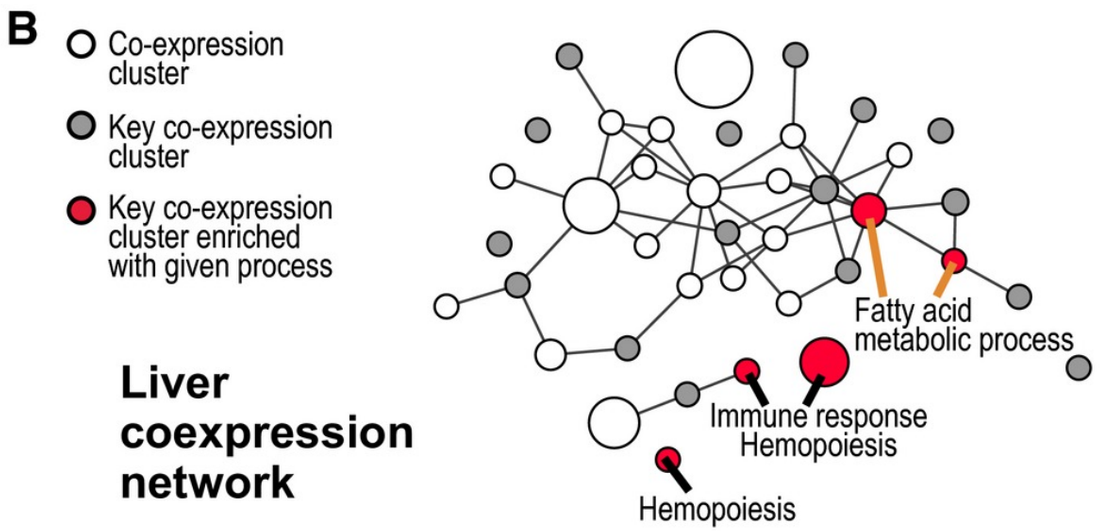
What background to consider?

Direction of change

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

Outcome





GO BPs from MSigDB

Network analyses identifies key stratifying genes

40-NN of FASN