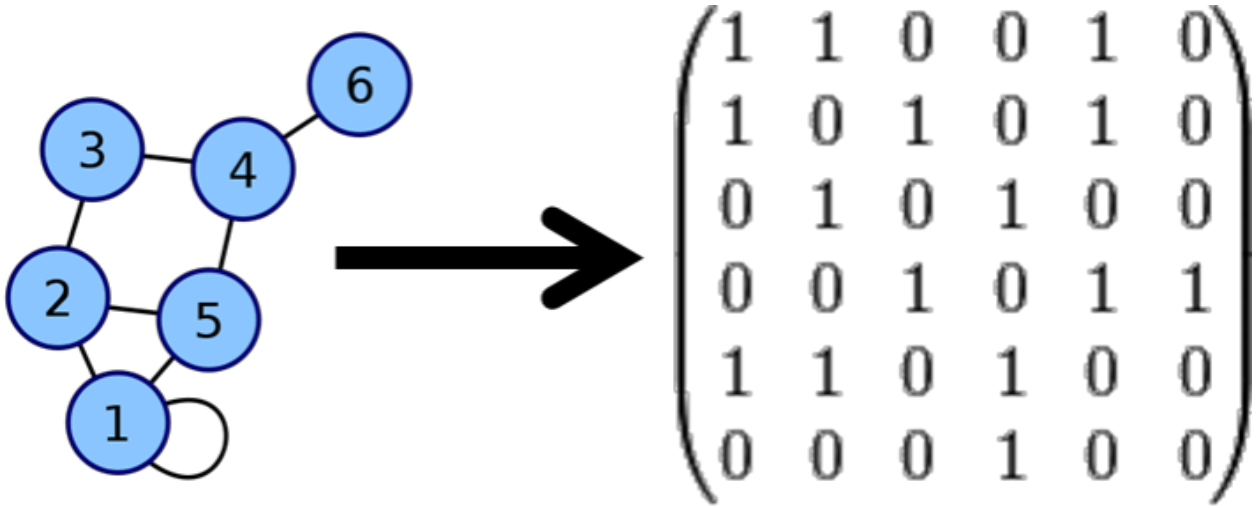
Biological Networks

John Platig
Assistant Professor
Department of Genome Sciences
Department of Biomedical Engineering

What is a network?

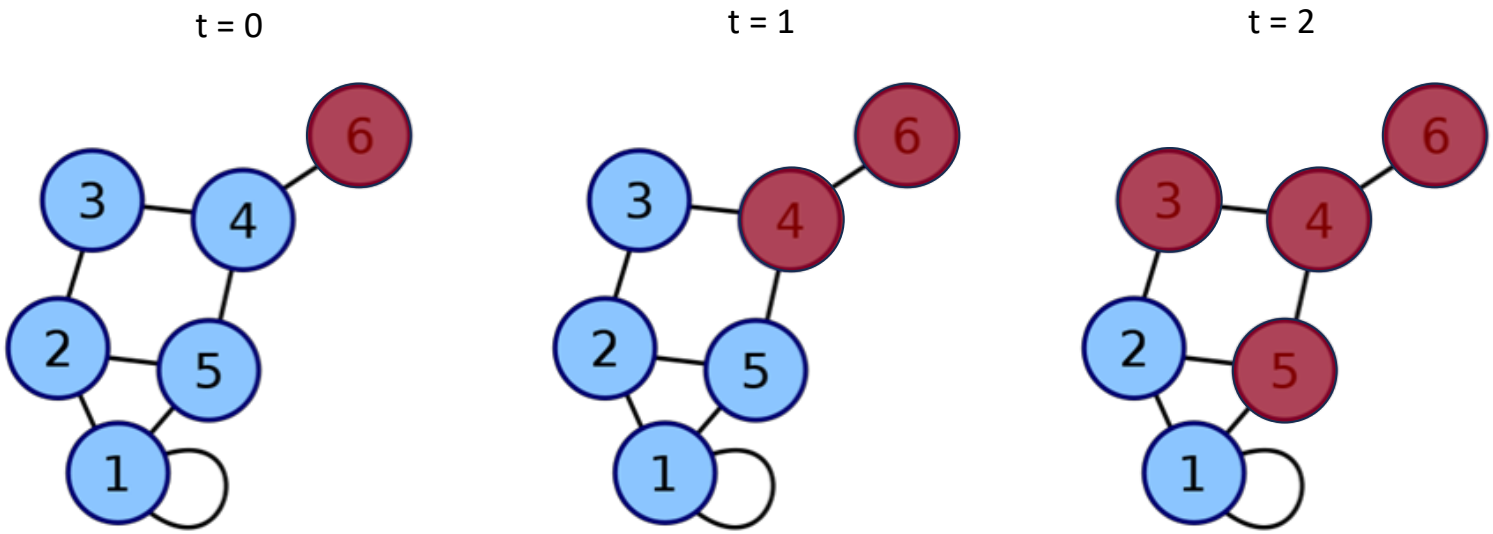
- A network is a way of representing patterns of connectivity (links) between entities (nodes).
- Networks can be abstract or represent physical/regulatory interactions.



$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

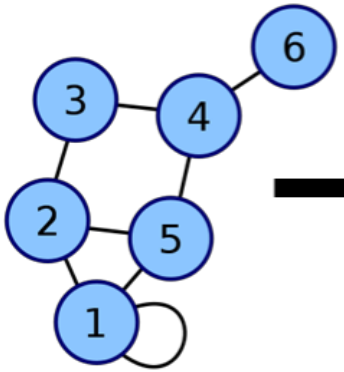
Why build a network?

- Can capture the *collective and nonlinear* effects of a system
- Network science literature tends to focus on two areas:
  - Network structure
  - Network processes (i.e., dynamics)

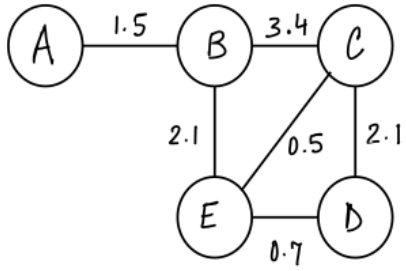


Part I: Network Theory

Matrix representation of Networks



$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

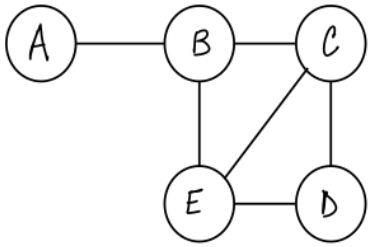


	A	B	C	D	E
A	0	1.5	0	0	0
B	1.5	0	3.4	0	2.1
C	0	3.4	0	2.1	0.5
D	0	0	2.1	0	0.7
E	0	2.1	0.5	0.7	0

This is a very flexible representation:

- Weights
- Directions
- Multi-edges
- Self-loops

Network Degree

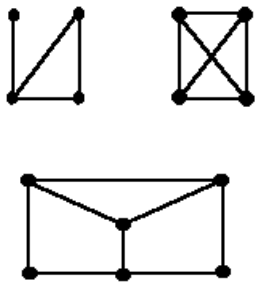


	A	B	C	D	E	Degree
A	0	1	0	0	0	1
B	1	0	1	0	1	3
C	0	1	0	1	1	3
D	0	0	1	0	1	2
E	0	1	1	1	0	3

What happens if the network is directed?

Networks (graphs)

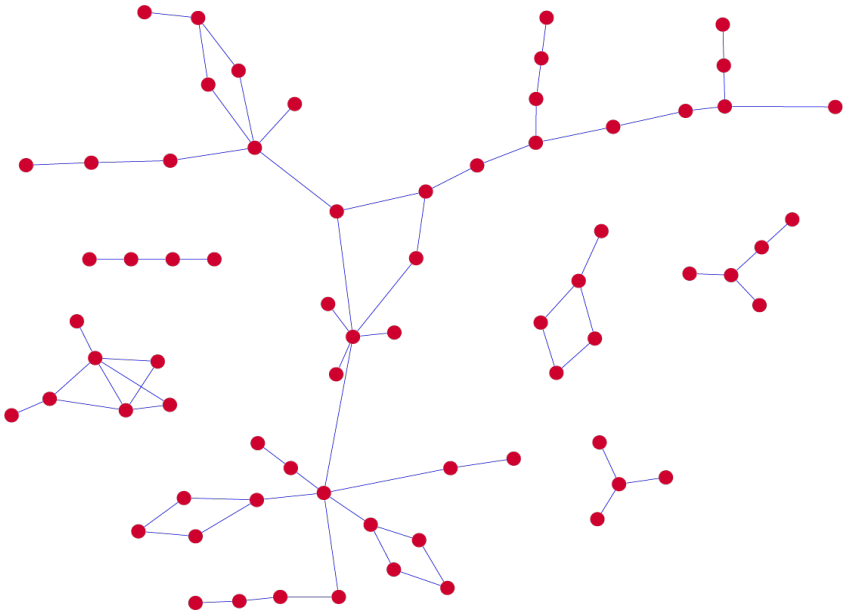
- Graph $G=(V,E)$ is a set of vertices V and edges E
- A subgraph G' of G is induced by some $V' \subset V$ and $E' \subset E$
- Graph properties:
  - Connectivity (node degree, paths)
  - Cyclic vs. acyclic
  - Directed vs. undirected



Sparse *vs* Dense

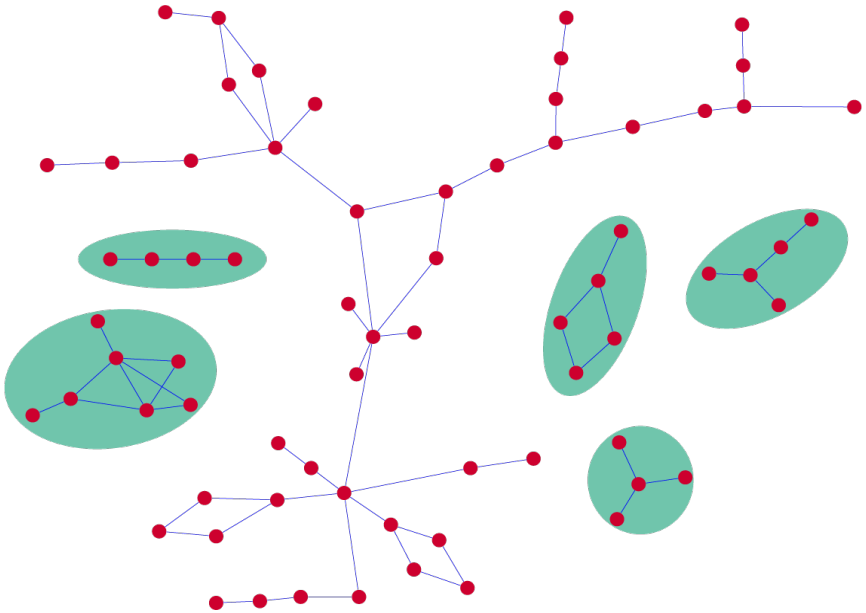
- G(V, E) where |V|=n, |E|=m the number of vertices and edges
- Graph is **sparse** if m~n
- Graph is **dense** if m~n^2
- **Complete graph** when m=n(n-1)/2

Connected Components



- G(V,E)
- |V| = 69
- |E| = 71

Connected Components

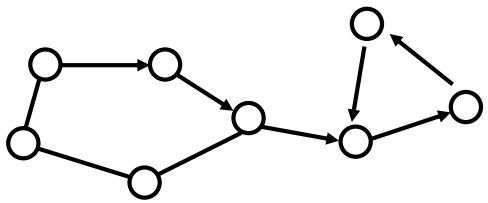


- G(V,E)
- |V| = 69
- |E| = 71
- 6 connected components

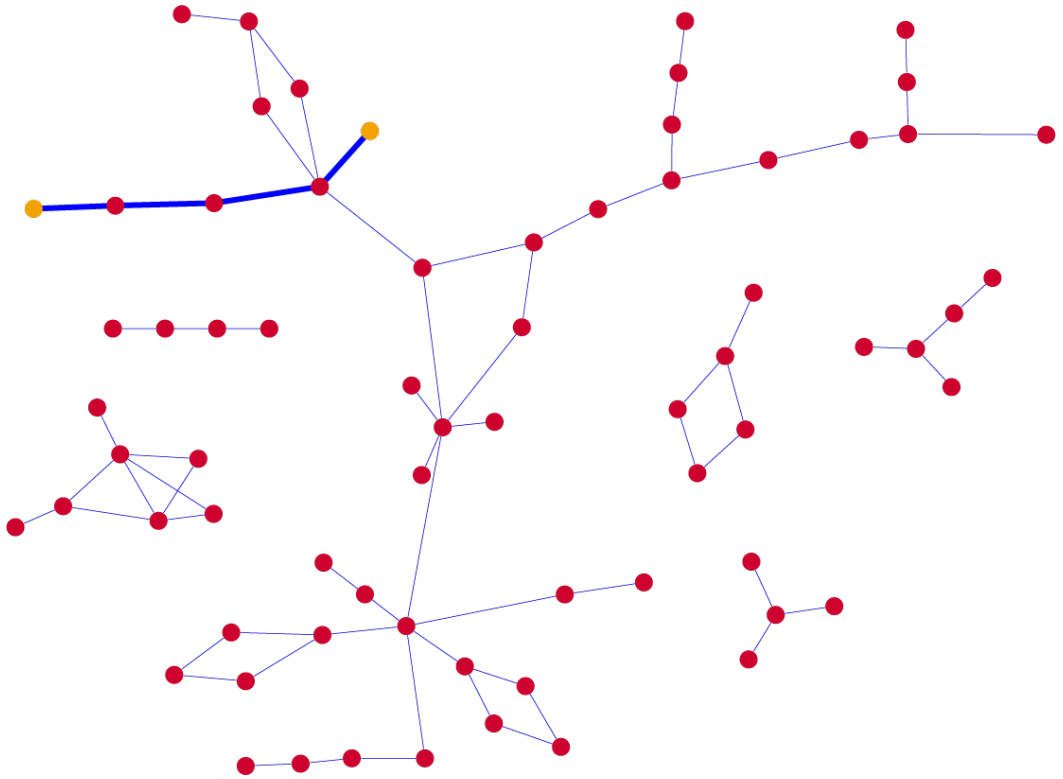
Paths

- A **path** is a sequence $\{x_1, x_2, \ldots, x_n\}$ such that (x_1,x_2), (x_2,x_3), …, (x_{n-1},x_n) are **edges** of the graph.

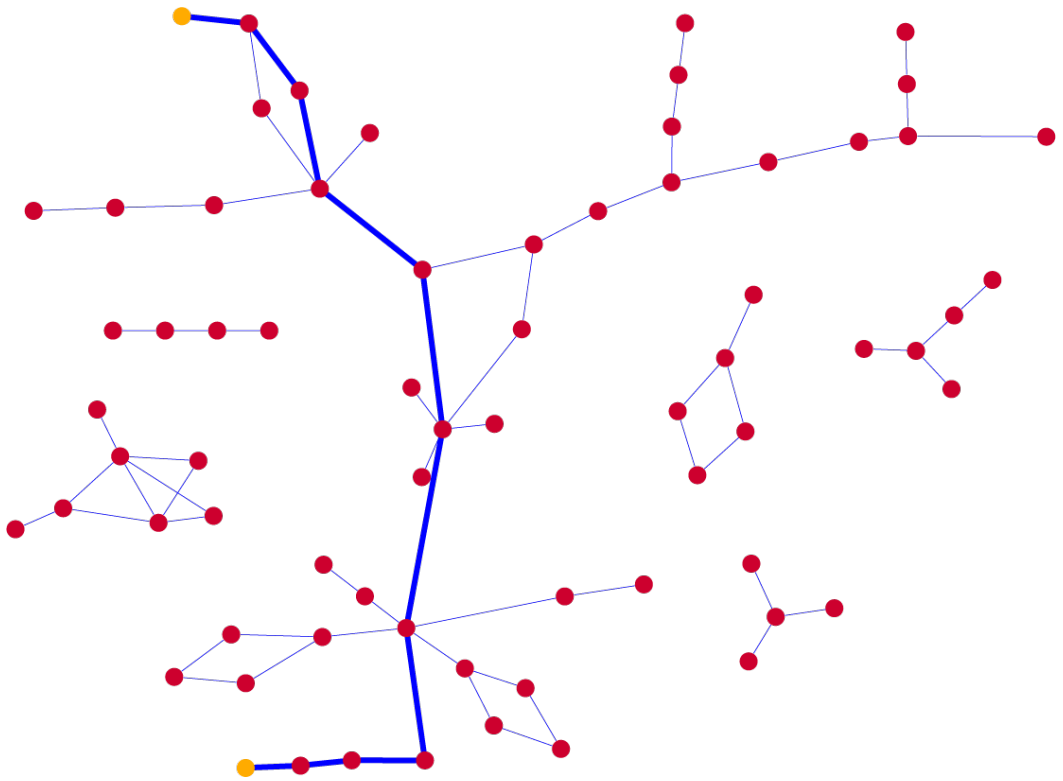
- A closed path $x_n=x_1$ on a graph is called a **graph cycle** or **circuit**.



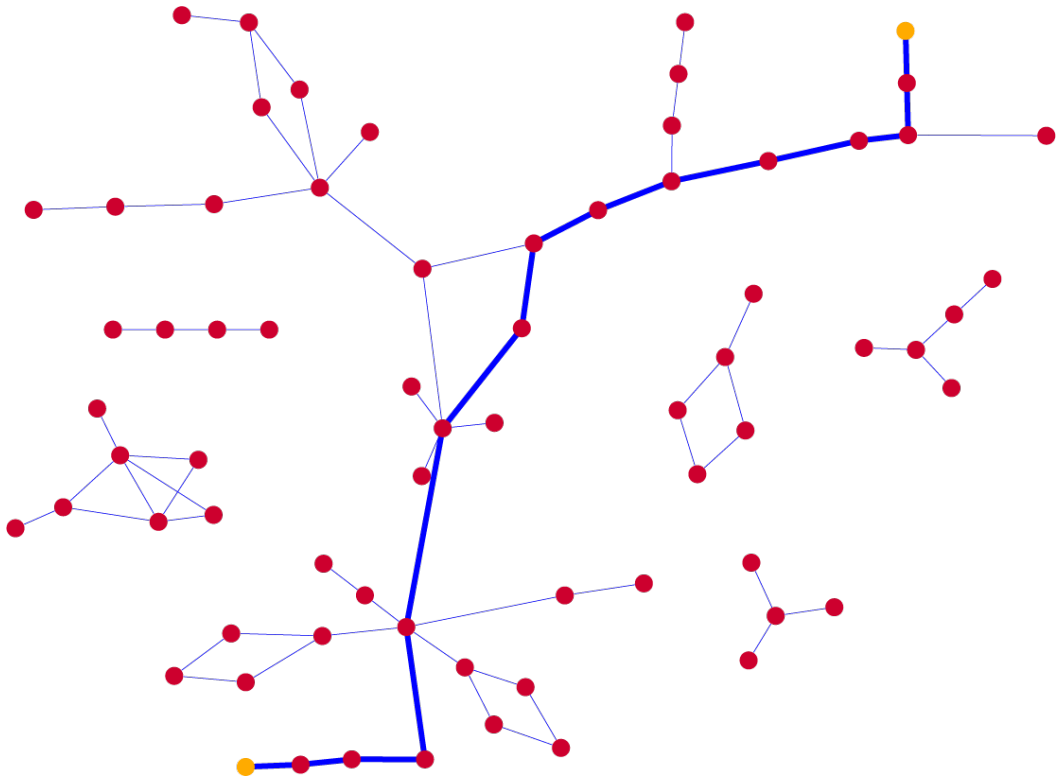
Shortest-Path between nodes



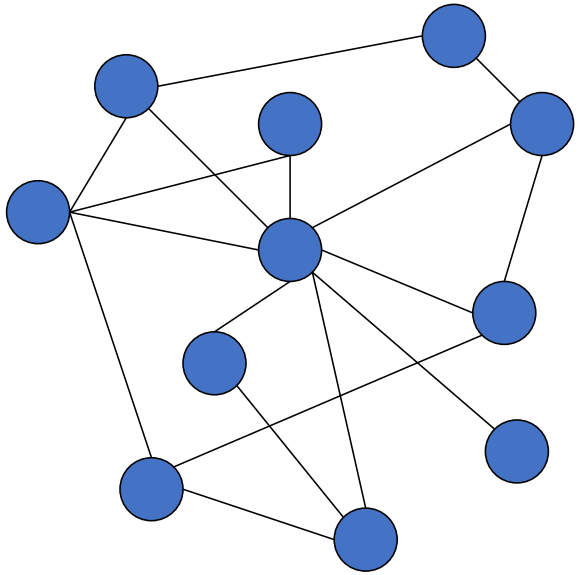
Shortest-Path between nodes



Longest Shortest-Path between nodes



Network paths and diameter



Shortest path:

Connect two nodes by as few edges as possible

Network diameter:

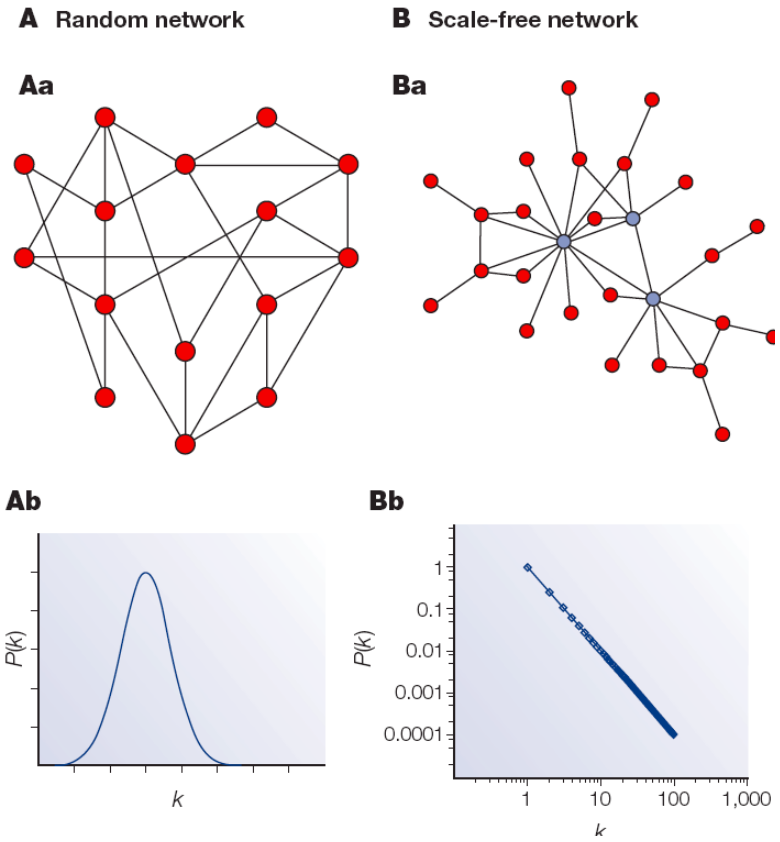
The longest shortest path in the network

The network diameter is often very short: ‘Small world network’

Small-World network

- Every node can be reached from every other by a small number of hops or steps
- High clustering coefficient and low mean-shortest path length
  - Random graphs don't necessarily have high clustering coefficients
- Social networks, the Internet, and biological networks all exhibit small-world network characteristics

Random *vs* scale-free networks



P(k) is probability of each degree k, i.e fraction of nodes having that degree.

For random networks, P(k) is Poisson distributed.

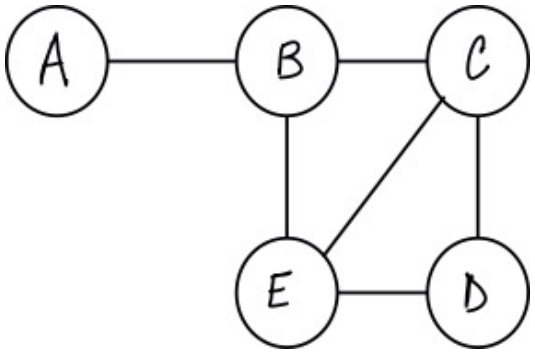
For real networks the distribution is often power-law-like:

$$P(k) \sim k^{-\gamma}$$

Such networks are said to be scale-free

Betweenness Centrality of a Node

The number of shortest paths in the graph that pass through the node divided by the total number of shortest paths.



$$BC(k) = \sum_i \sum_j \frac{\rho(i,k,j)}{\rho(i,j)}, i \neq j \neq k$$

Shortest paths: AB,ABC,ABCD,ABE,BC,BCD,BE,CD,CE,DE

$BC(B)$= 3/10

Nodes with a high betweenness centrality control information flow in a network.

Closeness Centrality

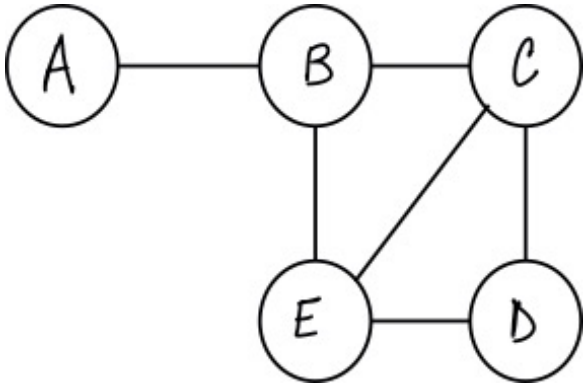
Measures the mean shortest distance from a node to all other nodes
Let d_{ij} = shortest path from *i* to *j*. Then the mean shortest path is

$$l_i = \frac{1}{n}\sum_j d_{ij}$$

And the closeness centrality is…

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}}$$

Note: $d_{ii} = 0$ by def. here.



$C_E = 1$, $C_A = 5/8$

Eigenvector Centrality

Make x_i proportional to the average of the centralities of i's neighbors:

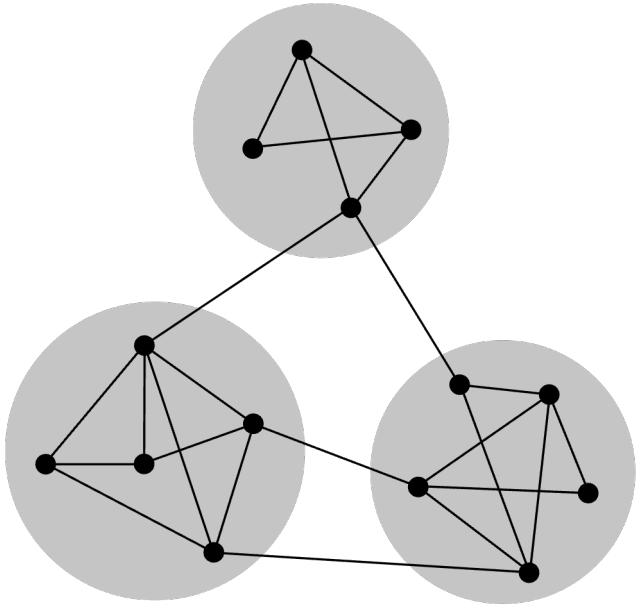
$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j$$

$$X = \frac{1}{\lambda} A X$$

Connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

Network clusters (communities)

- Identify groups of nodes that are densely connected within the group
  - Clique based methods
  - Quality function maximization (modularity)
  - Hierarchical clustering
  - Statistical inference



https://en.wikipedia.org/wiki/Community_structure

Network Clustering

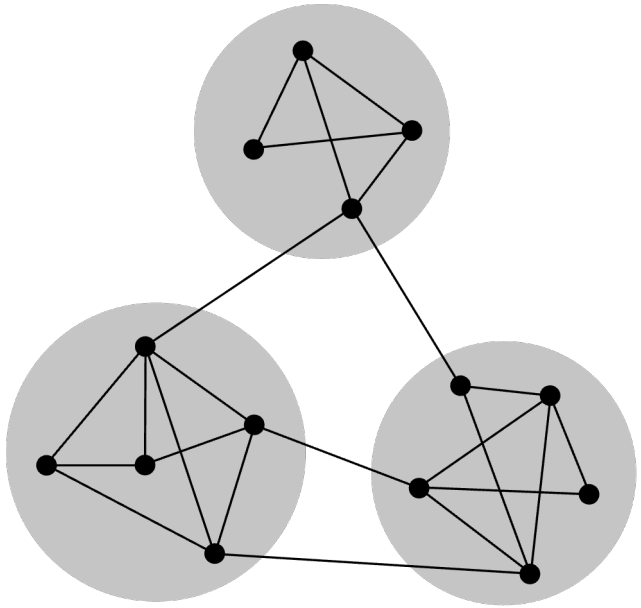
Most popular approach is modularity maximization:

$$Q(\boldsymbol{A}, \boldsymbol{b}) = \frac{1}{2E} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2E} \right) \delta_{b_i, b_j}$$

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmax}} \; Q(\boldsymbol{A}, \boldsymbol{b})$$

There are drawbacks to this method:

- Resolution limit
- Does not provide statistical significance of a partition/community



Community detection in networks: A user guide
https://doi.org/10.1016/j.physrep.2016.09.002

Network Clustering

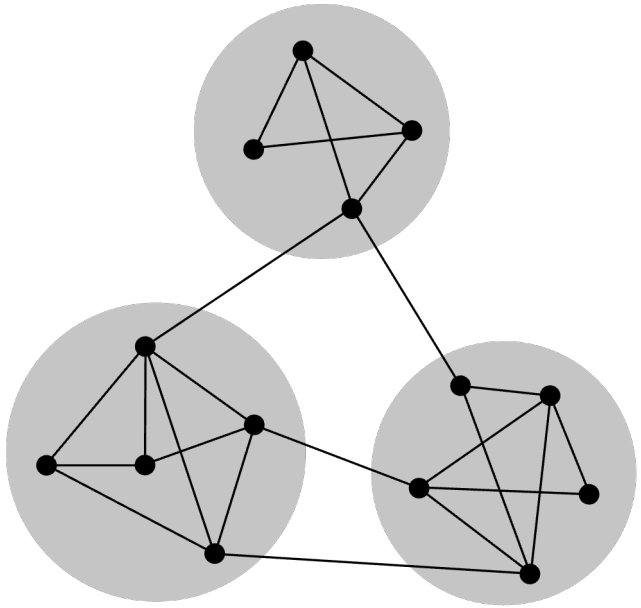
What the network theory people like: Stochastic block models

$$\mathcal{L}(G|g) = \sum_{rs} \frac{m_{rs}}{2m} \log \frac{m_{rs}/2m}{(\kappa_r/2m)(\kappa_s/2m)}$$

G – graph
g – membership vector of each node
m – num. edges in the network
r,s – groups (communities)
κ – stubs in the group

There are drawbacks to this method:

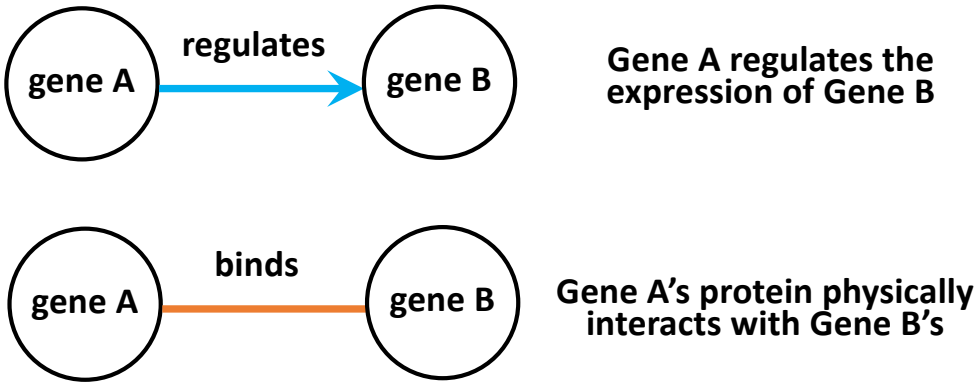
- You need to tell it how many groups there are



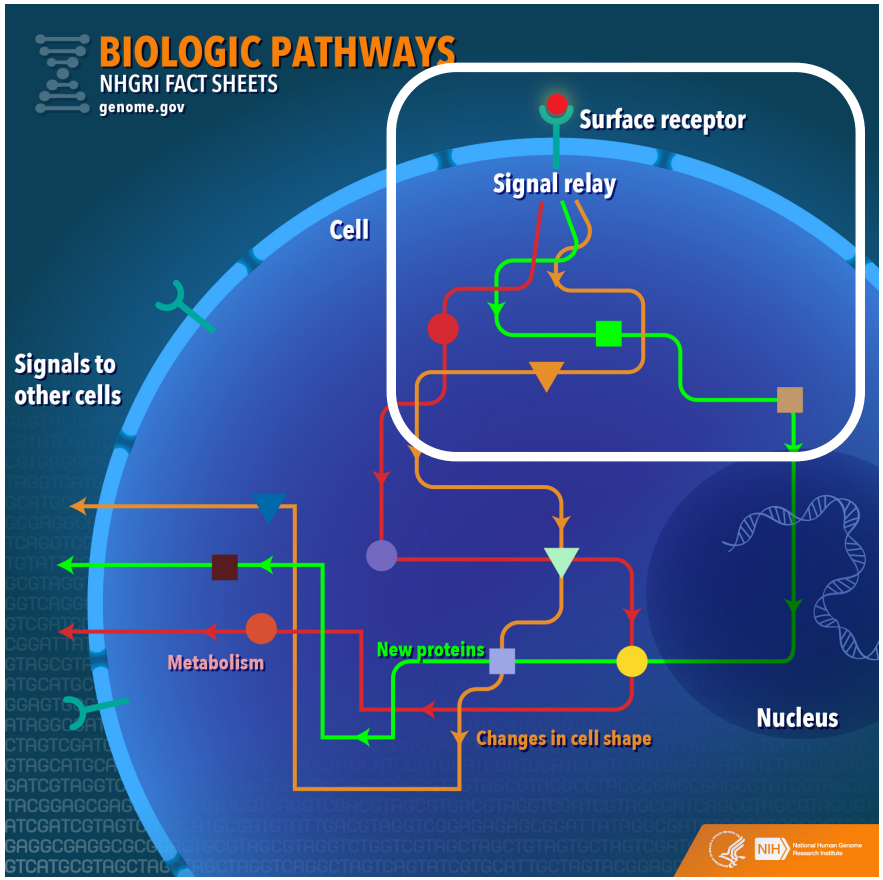
Community detection in networks: A user guide
https://doi.org/10.1016/j.physrep.2016.09.002

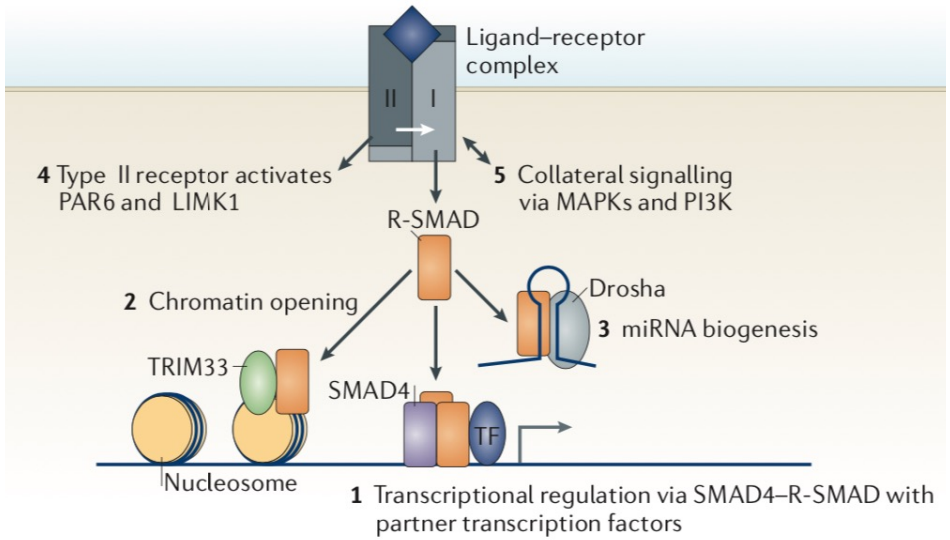
Part II: Networks in Biology

Biological network representations



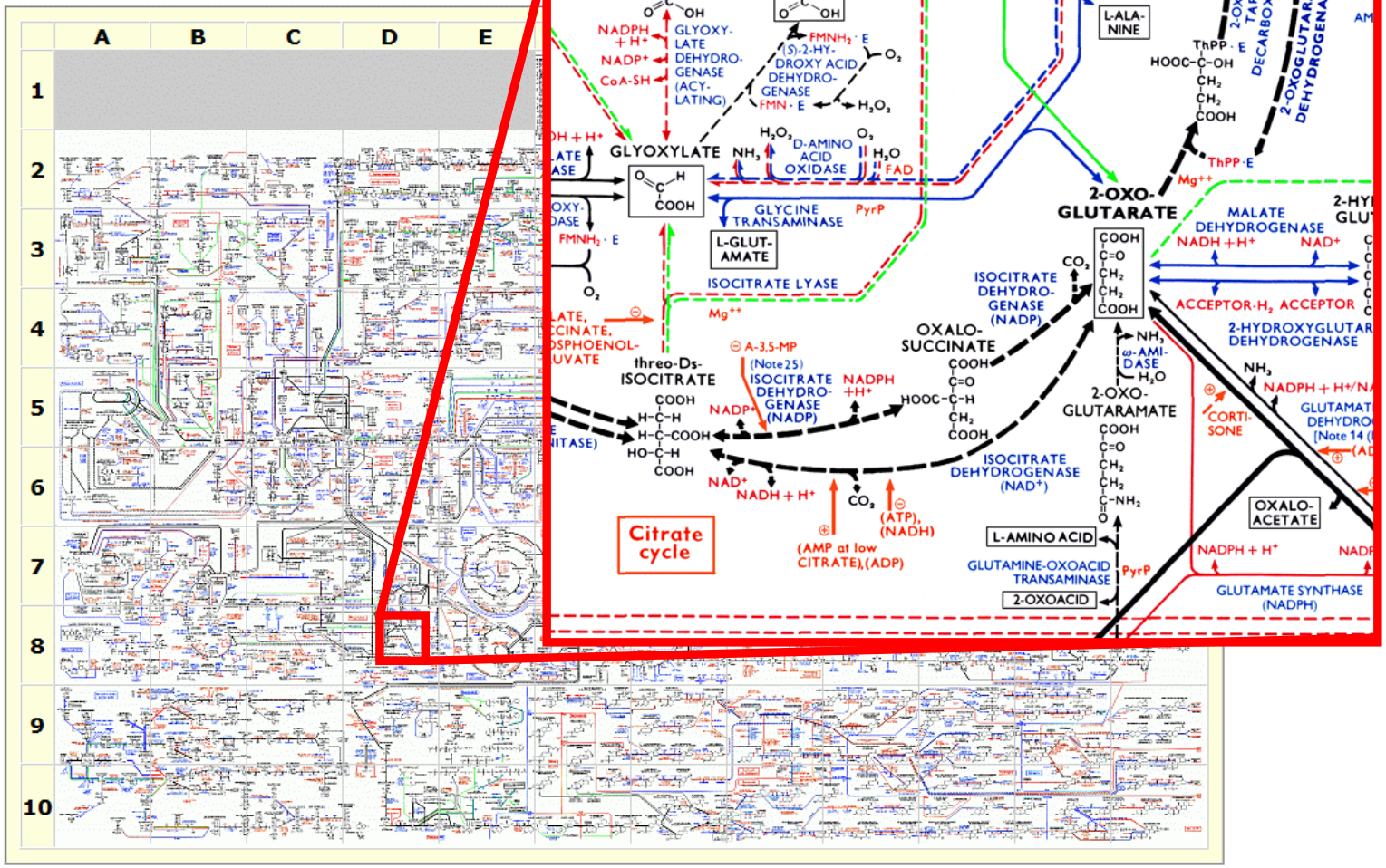
Networks: Signal Transduction



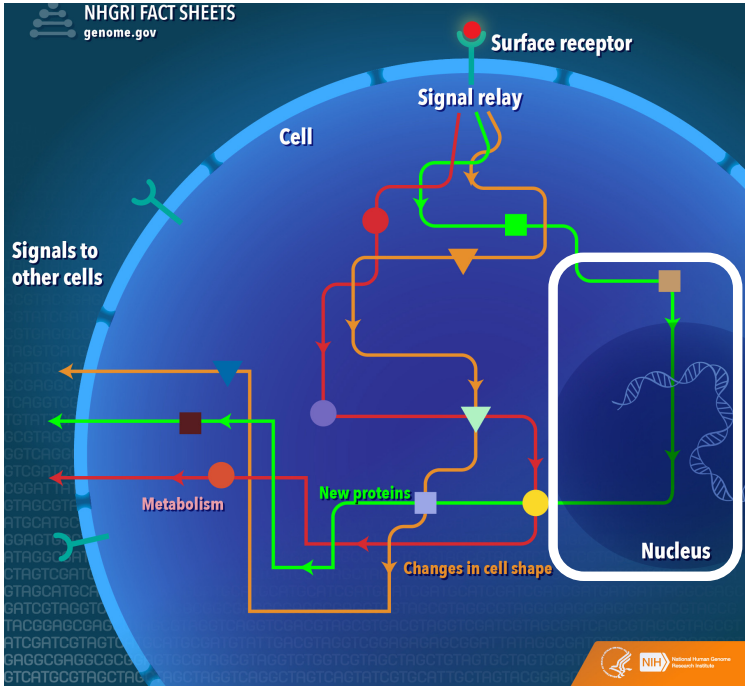


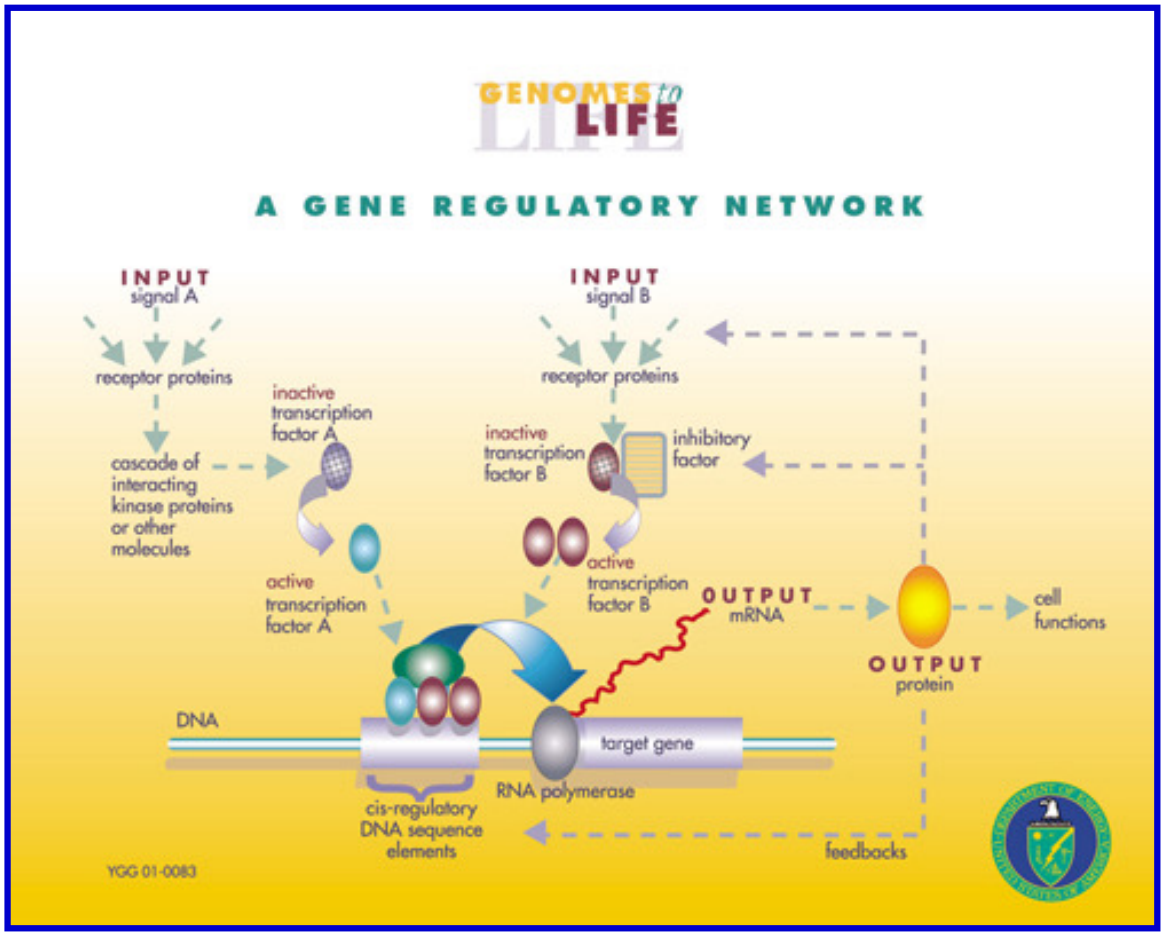
Massagué, *TGF-Beta signaling in context.*
PMID: 22992590

Networks: Metabolic

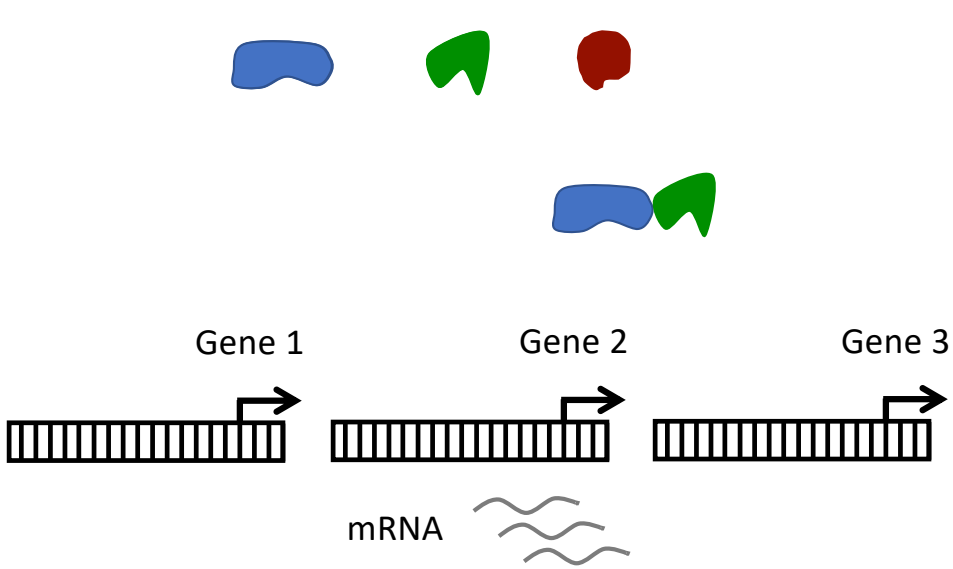


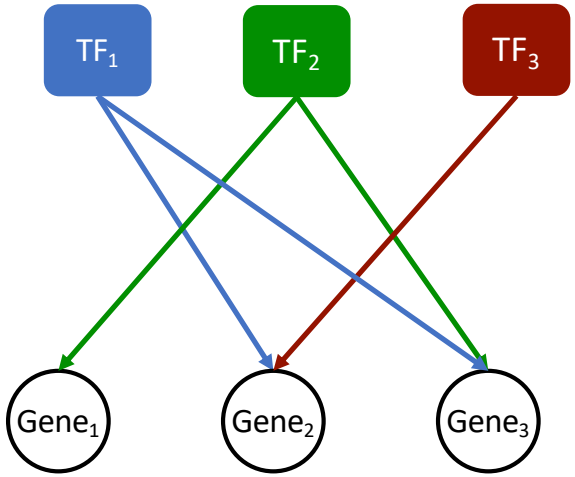
Networks: Gene Regulatory Processes





Transcription factor (TF) regulation of gene transcription

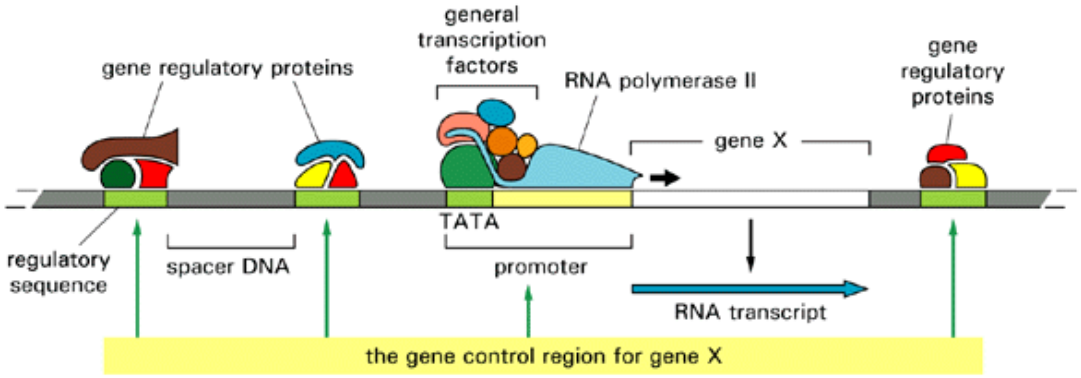




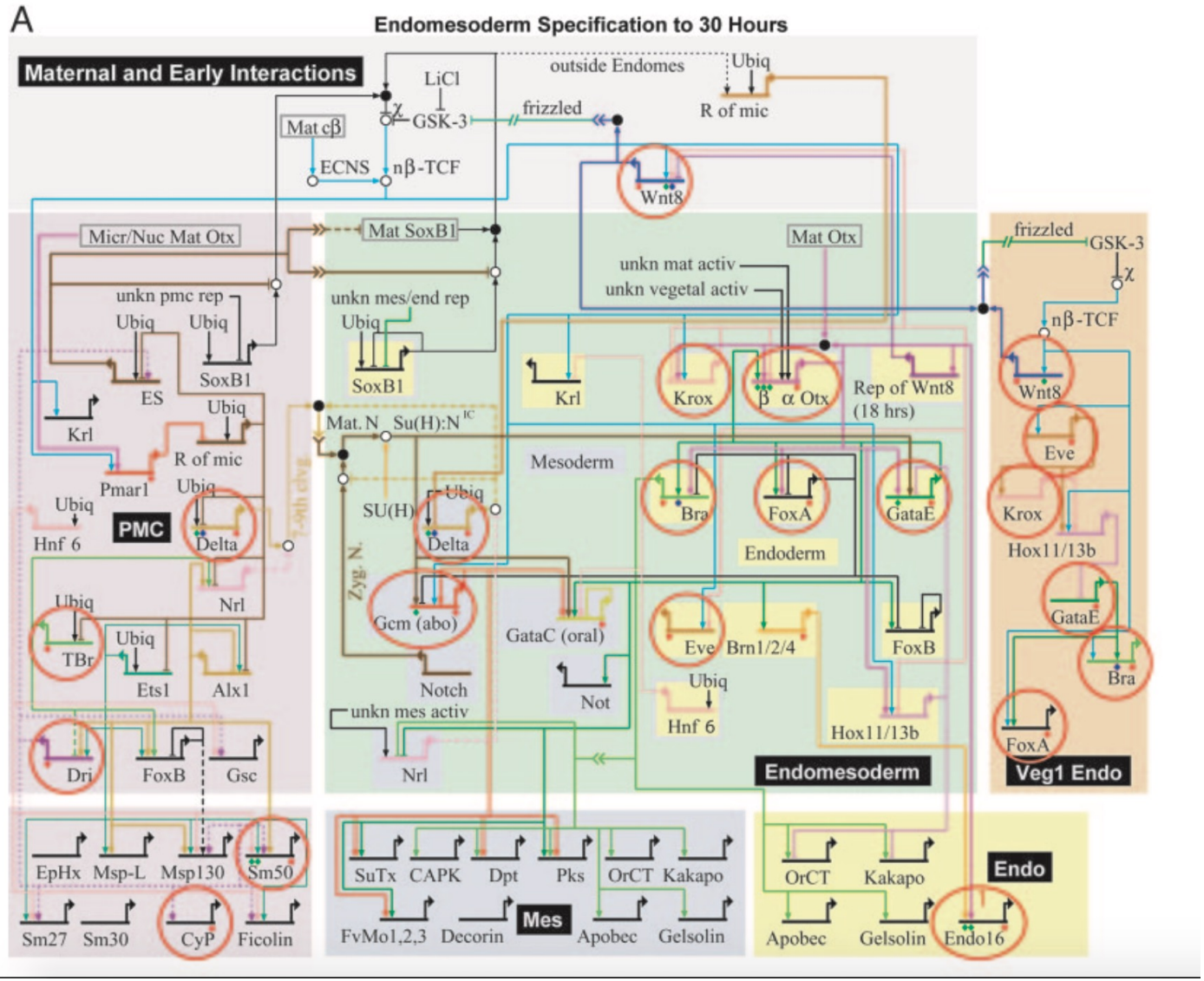
These TF → gene edges depend on the cell's environment and state

What counts as an “edge”?

- In developmental biology, an edge is supported by a focused experiment demonstrating the relationship between linked nodes
  - “Much of the architecture of the sea urchin network…is based directly on cis-regulatory experimentation, and the same is to a large extent true of all of the GRNs* included in this Special Feature.”
    - Michael Levine and Eric Davidson, PNAS 2005



*GRN – Gene Regulatory Network (TF -> gene edges)


A
Endomesoderm Specification to 30 Hours
Maternal and Early Interactions
LiCl
GSK-3
Mat cβ
ECNS
nβ-TCF
outside Endomes
Ubiq
R of mic
frizzled
Wnt8
Micr/Nuc Mat Otx
Mat SoxB1
Mat Otx
unkn mat activ
unkn vegetal activ
unkn pmc rep
Ubiq
SoxB1
ES
Krl
R of mic
Pmar1
PMC
Hnf 6
Delta
Nrl
TBr
Ets1
Alx1
Dri
FoxB
Gsc
unkn mes/end rep
Mat. N
Su(H):N IC
SU(H)
Zyg. N.
Gcm (abo)
Notch
unkn mes activ
Mesoderm
Krox
β α Otx
Rep of Wnt8
(18 hrs)
Bra
FoxA
GataE
Endoderm
GataC (oral)
Not
Eve
Brn1/2/4
FoxB
Hnf 6
Hox11/13b
Endomesoderm
Veg1 Endo
Krox
Hox11/13b
EpHx
Msp-L
Msp130
Sm50
Sm27
Sm30
CyP
Ficolin
SuTx
CAPK
Dpt
Pks
OrCT
Kakapo
FvMo1,2,3
Decorin
Mes
Apobec
Gelsolin
Endo
Endo16

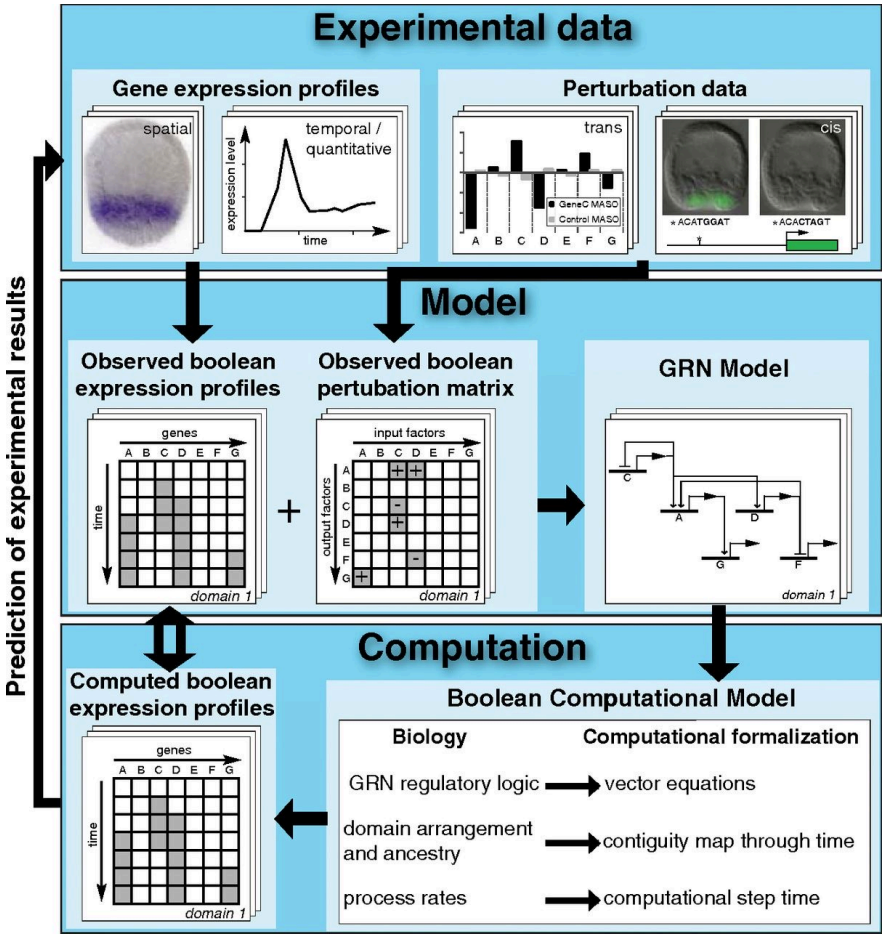
Predicting outputs from inputs

Boolean logic tables

TF1	TF2	G1
1	1	1
1	0	1
0	1	0
0	0	0

TF1	TF2	G1
1	1	1
1	0	0
0	1	0
0	0	0

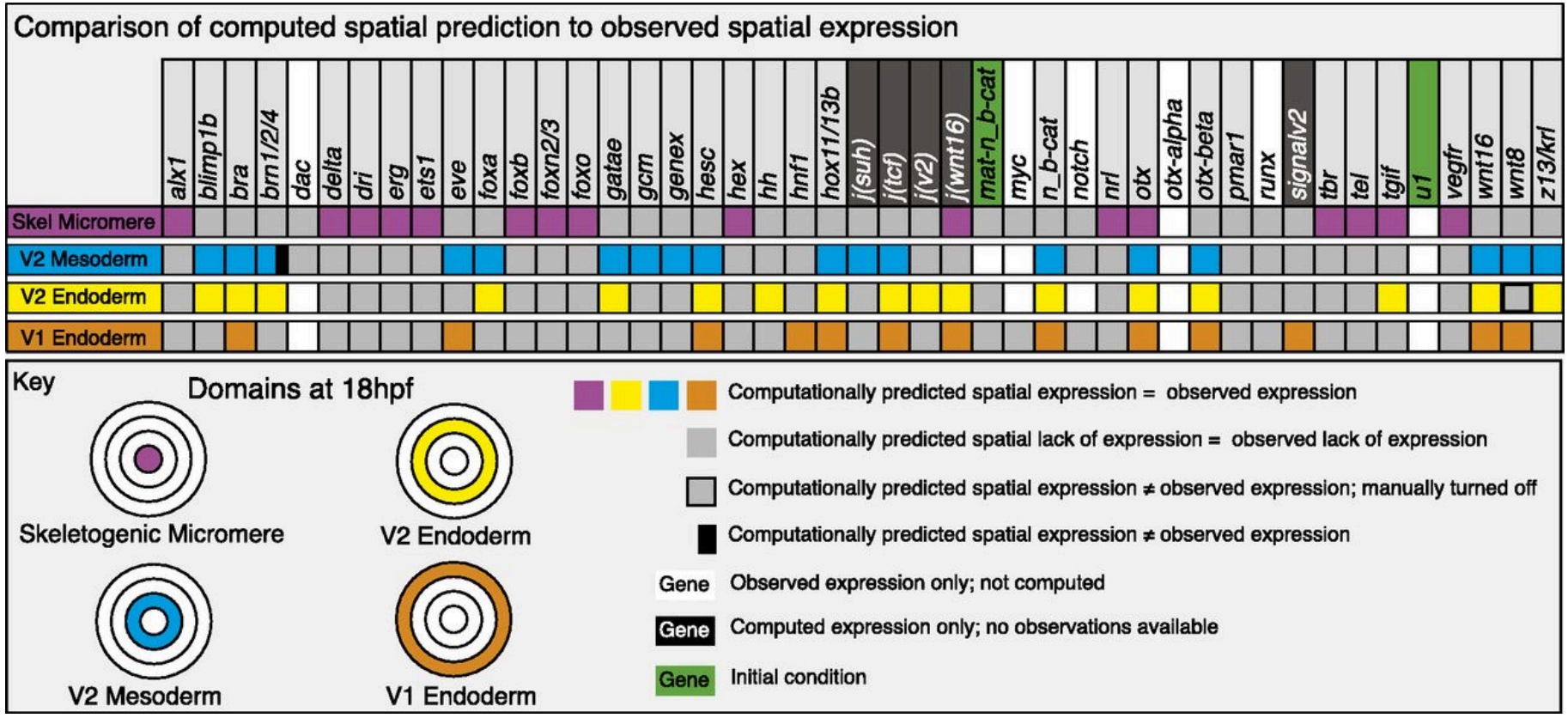
Predicting outputs from inputs



"This Boolean model computes spatial and temporal gene expression according to the regulatory logic and gene interactions specified in a GRN model for embryonic development in the sea urchin."

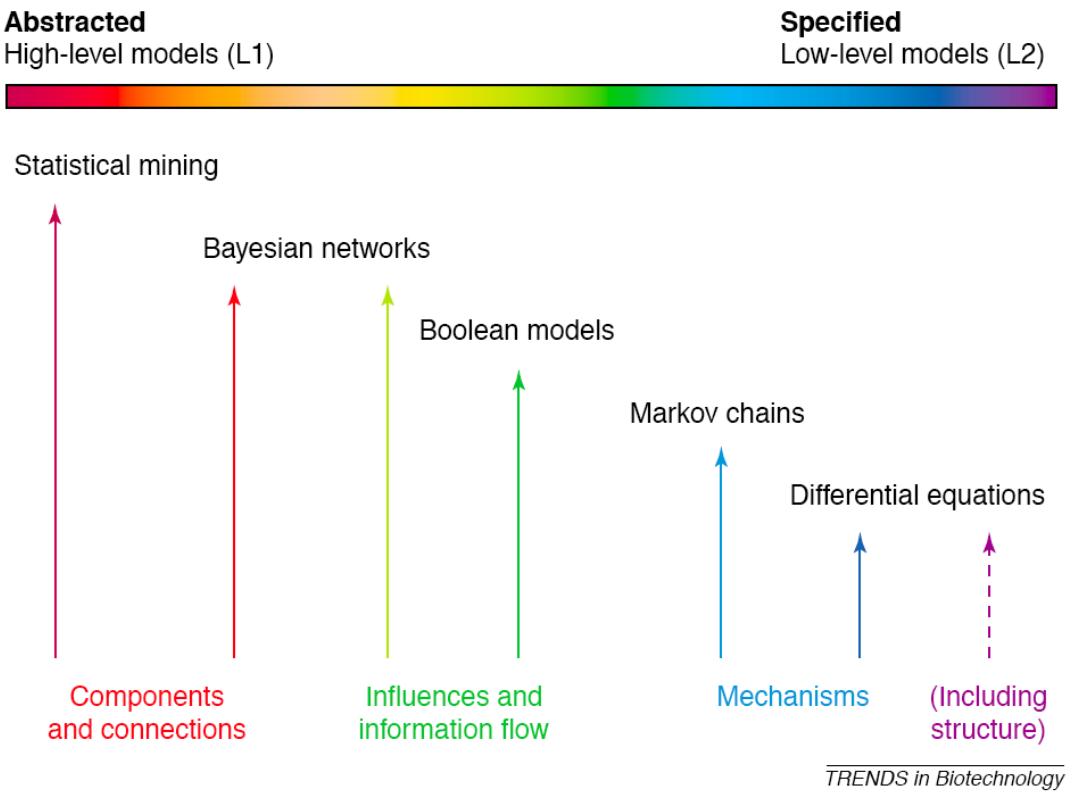
Peter et. al. PNAS 2012

Predicting outputs from inputs



Peter et. al. PNAS 2012

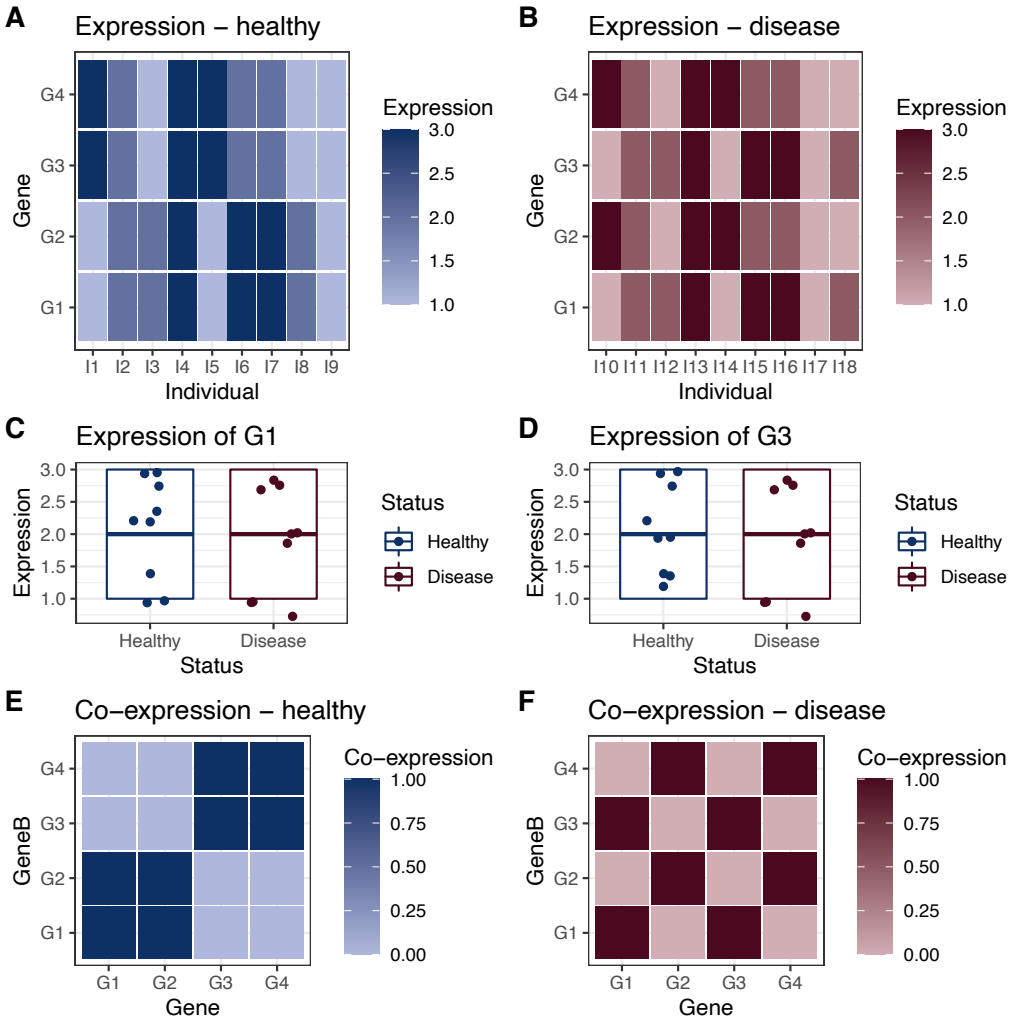
Hierarchy of models



Ideker & Lauffenburger
TRENDS in Biotechnology
2003

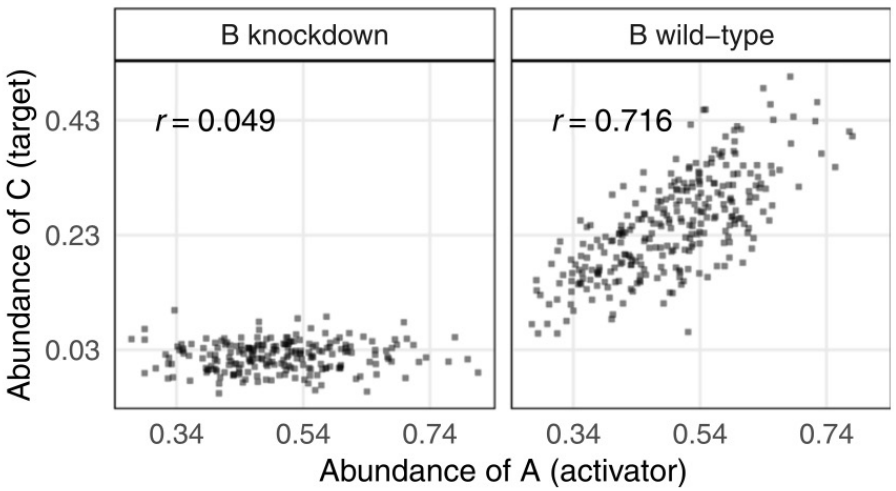
GRNs: The population genomics perspective

- How do we identify processes that matter in human health and disease based on many observations of all genes' mRNA levels
  - Is the expression of a gene different between two conditions?
- Differential expression is only part of the story, some genes don't change their mean expression, but their correlation with other genes changes



Glass et al. 2014

A coexpression-based regulatory network model of TFs



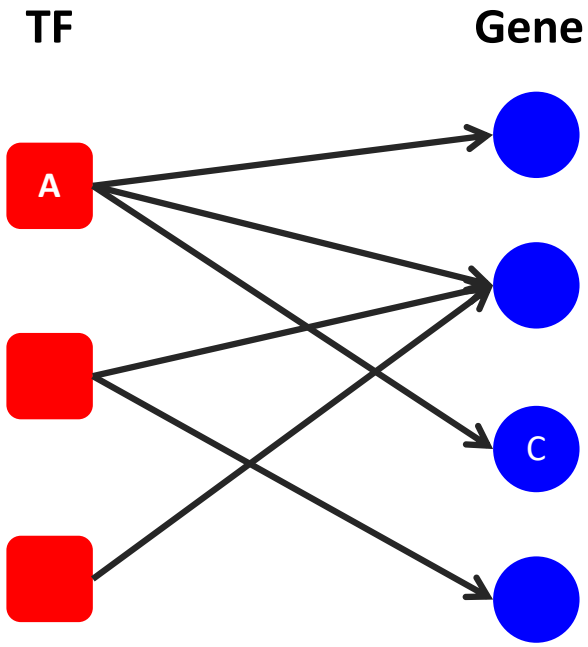
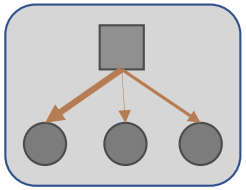
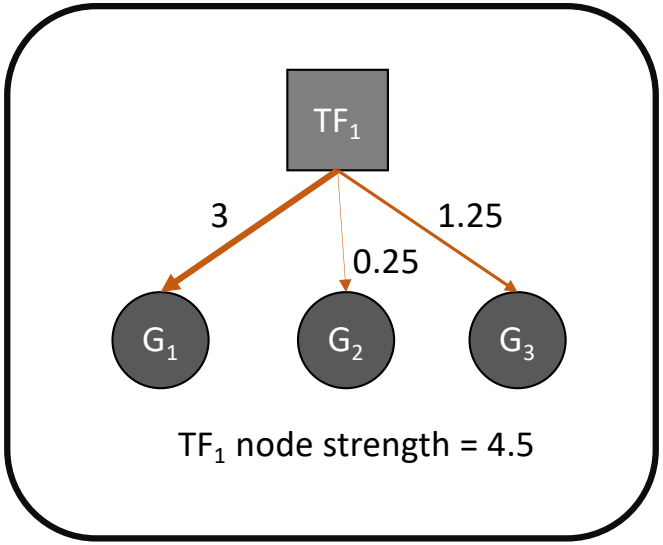


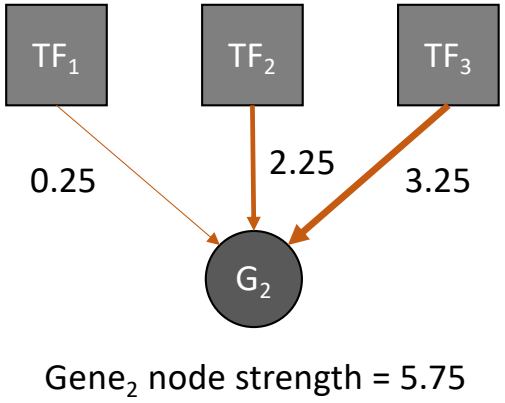
Fig 1. Bhuva, D.D., Cursons, J., Smyth, G.K. *et al.* Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. *Genome Biol* **20**, 236 (2019). https://doi.org/10.1186/s13059-019-1851-8

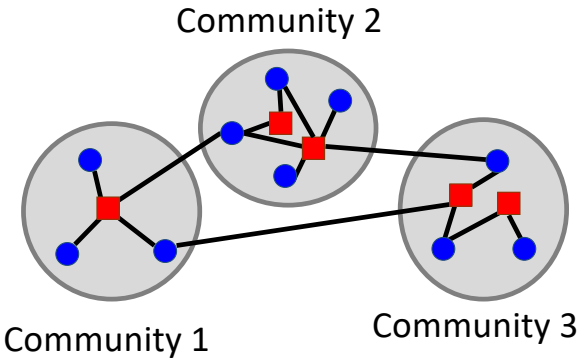


Three levels of network statistics

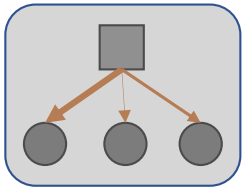
- TF node strength/degree (sum of outgoing edge weights)
- Gene node strength/in-degree (sum of incoming edges)
- Modules (network communities)



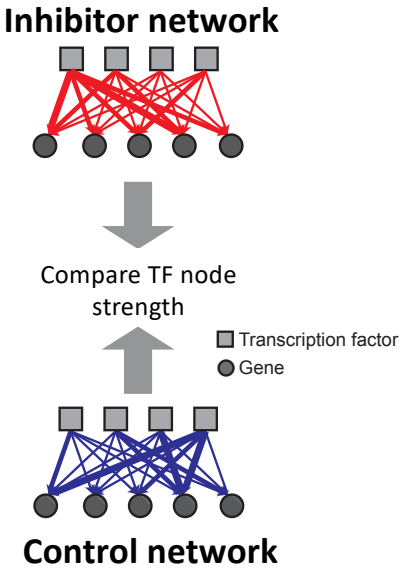


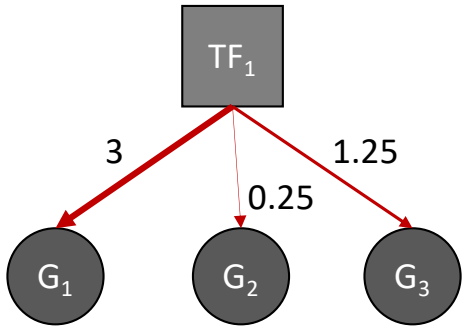


Differences in TF node strength

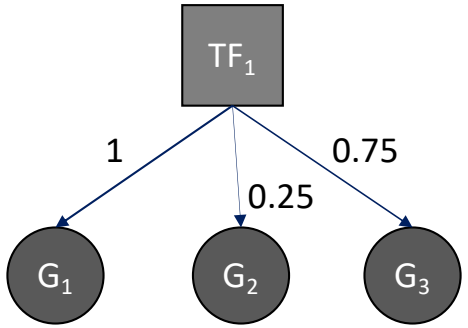


Compare the sum of outgoing edge weights for each TF in each condition

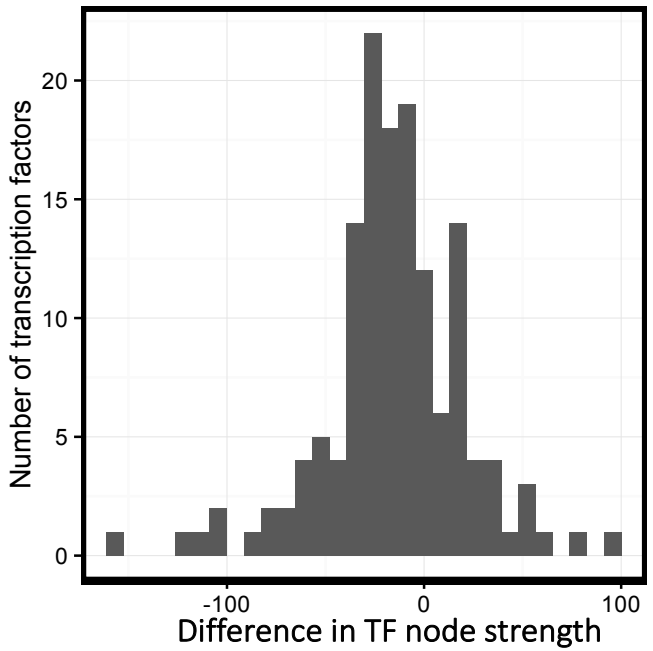


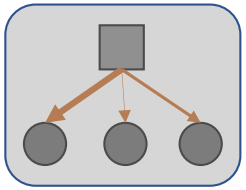


TF_1 node strength = 4.5



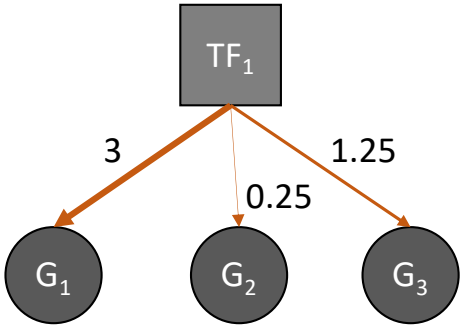
TF_1 node strength = 2



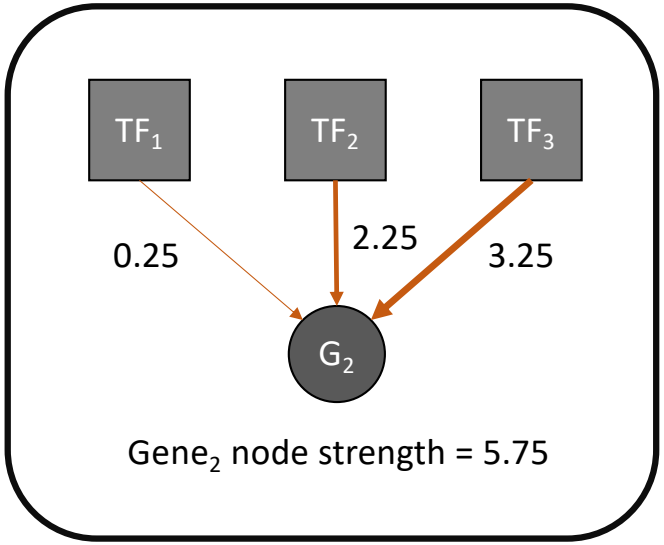


Three levels of network statistics

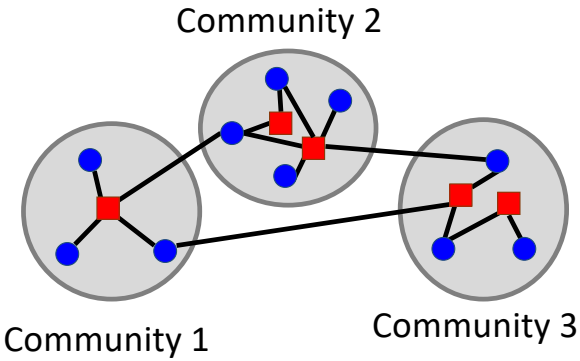
- TF node strength/degree (sum of outgoing edge weights)
- Gene node strength/in-degree (sum of incoming edges)
- Modules (network communities)



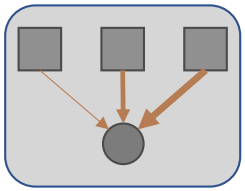
TF_1 node strength = 4.5



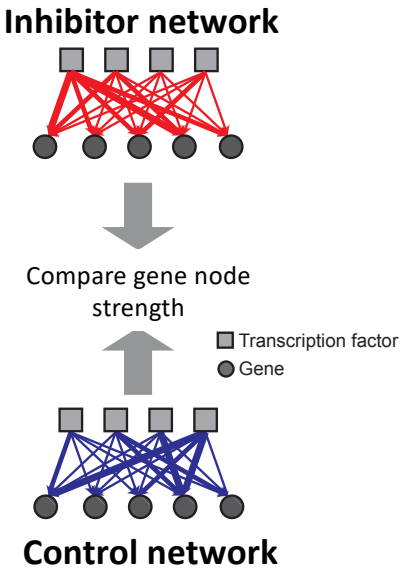
$Gene_2$ node strength = 5.75

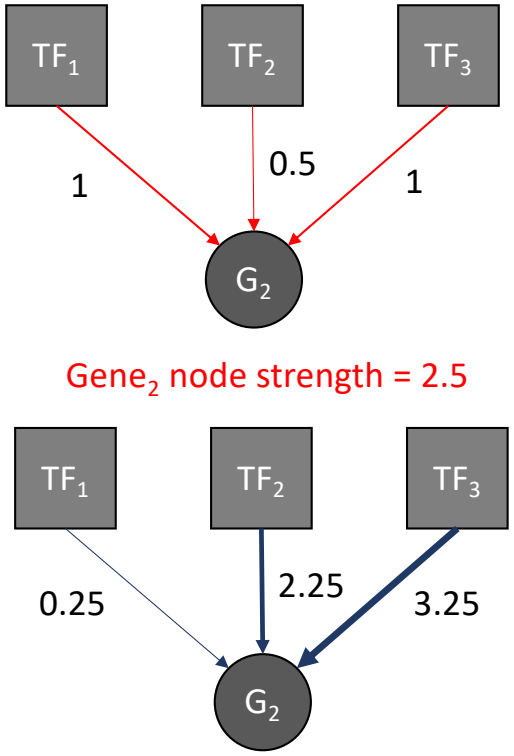


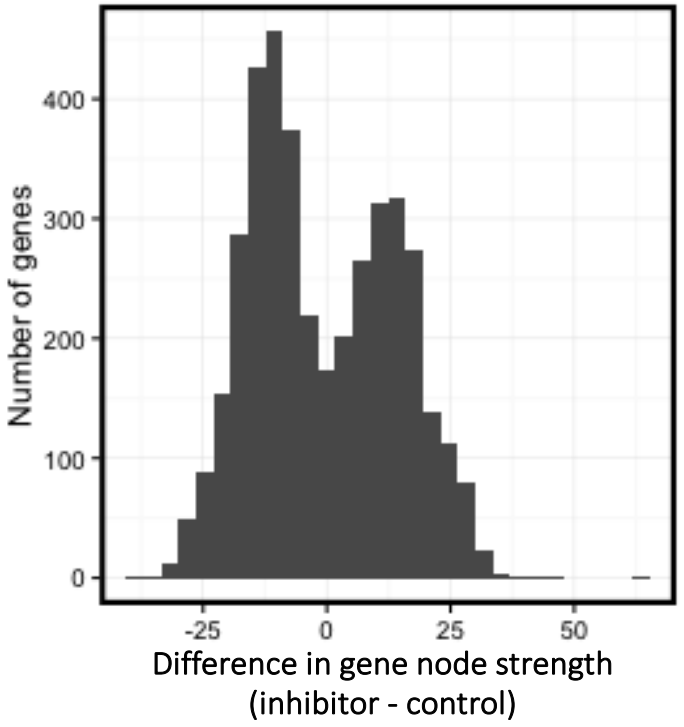
Differences in gene node strength

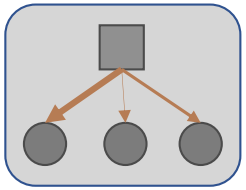


Compare the sum of incoming edge weights for each gene in each condition



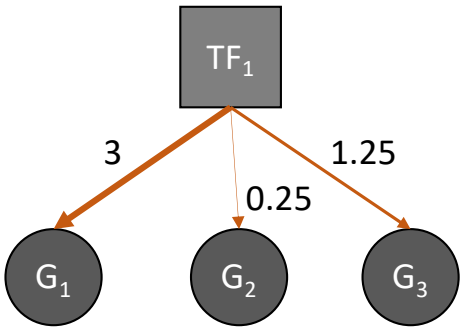




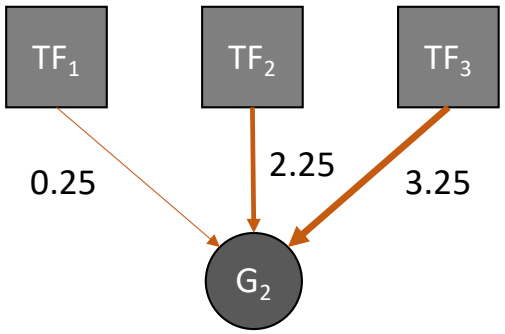


Three levels of network statistics

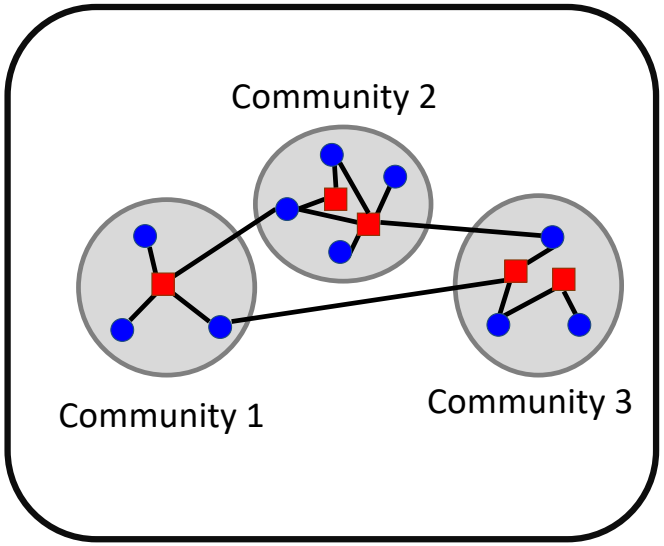
- TF node strength/degree (sum of outgoing edge weights)
- Gene node strength/in-degree (sum of incoming edges)
- Modules (network communities)



TF_1 node strength = 4.5

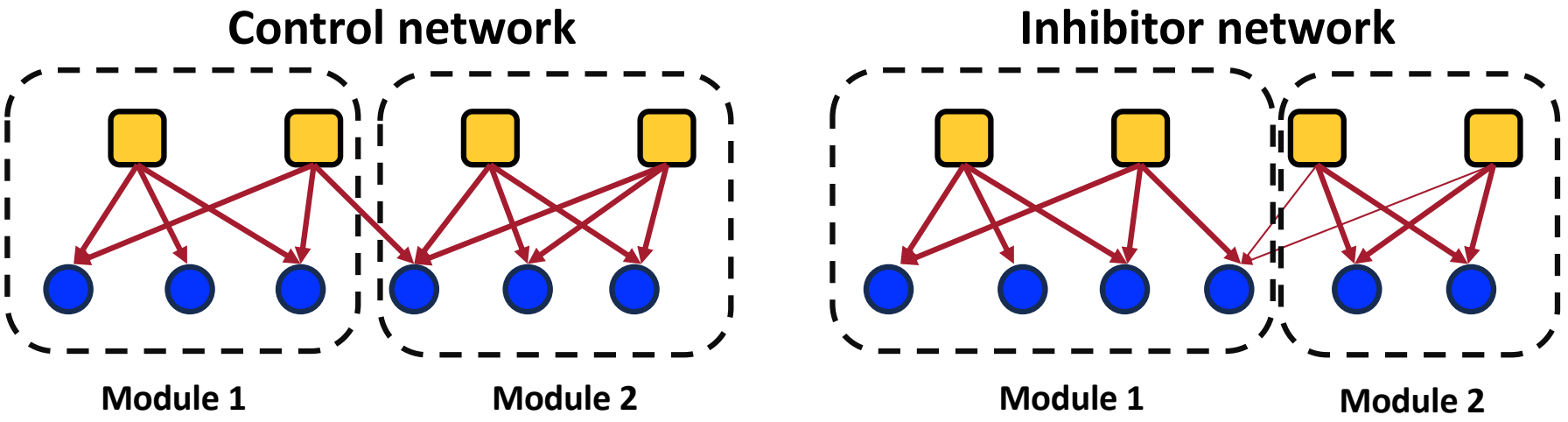


$Gene_2$ node strength = 5.75



Differential network clustering

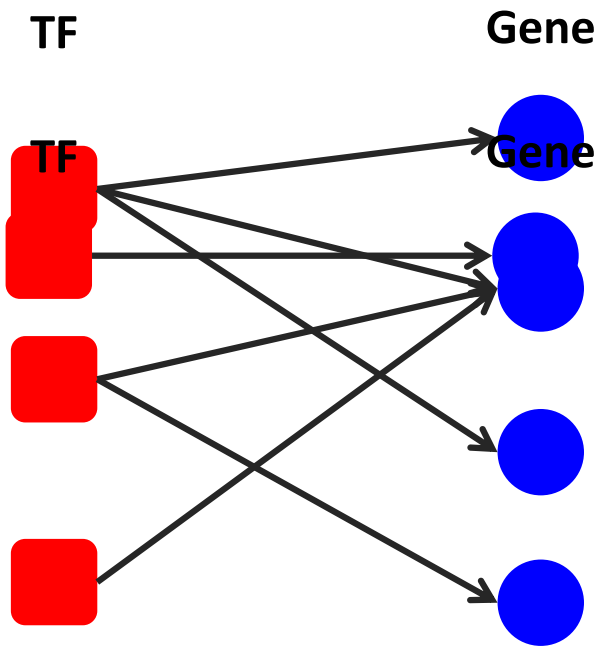
The pattern of connectivity in regulatory networks can reveal the groups of regulators and regulatory elements that are active in a given condition



Padi, M., Quackenbush, J. Detecting phenotype-driven transitions in regulatory network structure. *npj Syst Biol Appl* **4**, 16 (2018). https://doi.org/10.1038/s41540-018-0052-5

Network inference methods

Basic premise: The network structure itself is informative



- **Gene expression correlation**
  - Requires only a couple dozen expression profiles
  - No directionality, TFs are lowly expressed
- **DNase footprinting**
  - Requires TFs with motifs
- **Multi-omic and machine learning methods**
  - Often uses ChIP-seq from large databases like ENCODE or TF motifs

GRNs: The population genomics perspective

- Now edges between nodes are the result of statistical inference
- They often come from observational studies
  - No causal element to study design (case vs. control)
  - Mixed cell types
  - Other confounders

Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome

Matthew N. McCall,[1,*] Peter B. Illei,[2] and Marc K. Halushka[2,*]

To summarize, in a collection of normal lung samples, we found that tissue heterogeneity caused by harvesting location (medial or lateral lung) and late therapeutic intervention (mechanical ventilation) were major contributors to expression variation. These unexpected sources of variation were the result of altered cell ratios in the tissue samples, an underappreciated source of expression variation.

concerned genes. One large cluster included surfactant genes (*SFTPA1*, *SFTPA2*, and *SFTPC*), which are expressed exclusively in type II pneumocytes, cells that proliferate in ventilator associated lung injury. High surfactant expression was strongly associated with death on a ventilator and type II pneumocyte hyperplasia. A second large cluster included dynein (*DNAH9* and *DNAH12*) and mucin (*MUC5B* and *MUC16*) genes, which are exclusive to the respiratory epithelium and goblet cells of bronchial structures. This indicates heterogeneous bronchiole sampling due to the harvesting location in the lung. A small cluster included acute-phase reactant genes (*SAA1*, *SAA2*, and *SAA2–SAA4*). The final two small clusters were technical and gender related. To summarize, in a collection of normal lung samples, we found that tissue heterogeneity caused by harvesting location (medial or lateral lung) and late therapeutic intervention (mechanical ventilation) were major contributors to expression variation. These unexpected sources of variation were the result of altered cell ratios in the tissue samples, an underappreciated source of expression variation.

Cell type proportion accounts for a substantial amount of co-expression in bulk samples

Untangling the effects of cellular composition on coexpression analysis

Marjan Farahbod[1,2,3] and Paul Pavlidis[1,2]

[1]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; [2]Department of Psychiatry, University of British Columbia, Vancouver, British Columbia V6T 2A1, Canada; [3]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, British Columbia V5T 4S6, Canada

"Our conclusion is that the dominant coexpression signal in brain, blood, and likely, other complex tissues can be attributed to cellular compositional effects, rather than intra-cell-type regulatory relationships. *These results have implications for the relevance and interpretation of coexpression analysis.*"

Cell type proportion accounts for a substantial amount of co-expression in bulk samples

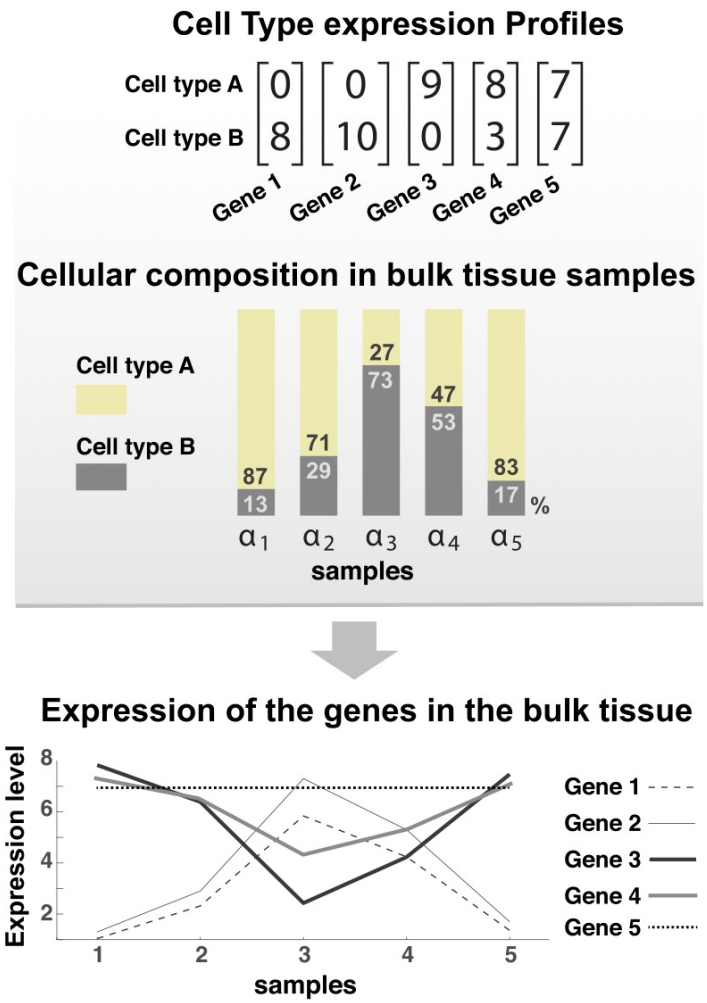


Fig. 1 from Farahbod and Pavlidis

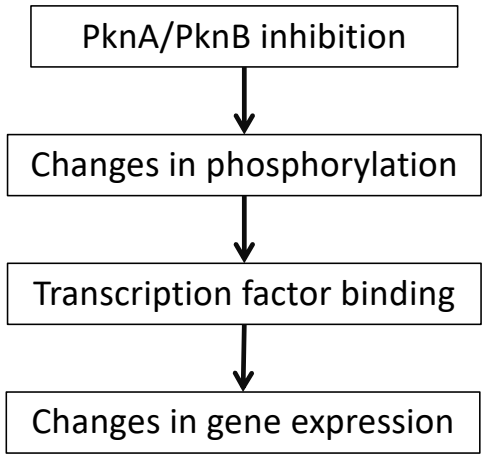
How do you know you're right?

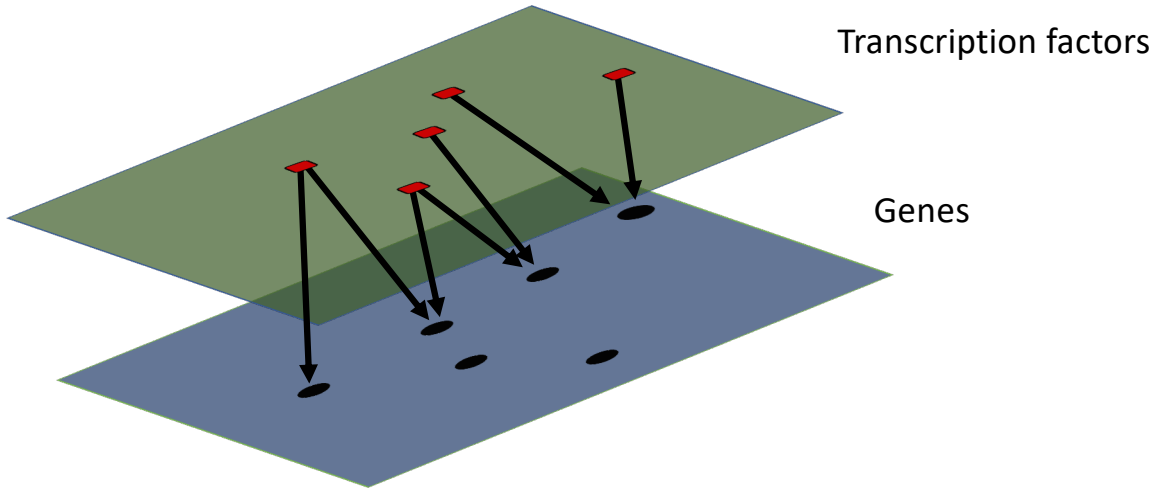
- You don't, especially if there's no strong scientific hypothesis (e.g., case/control)
- Replication
- TF knockout/knock-down in a "relevant" phenotype
- Look for confirmatory evidence in literature ("biopoetry")
  - The interesting TF I found for COPD is implicated in COPD based on other data
- If you have a strong scientific hypothesis, this is much easier
  - Does TF X differentially regulate its targets after to cigarette smoke exposure in alveolar type 2 cells?

Regulatory network response to PknA/PknB inhibition

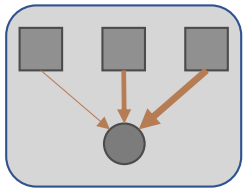
Experimental set-up:

Inhibit PknA/PknB signal transduction using a synthetic small molecule in a virulent TB lab strain (H37Rv)





Gene In-degree



Which pathways are most differentially regulated?

- Ranked genes by in-degree difference, ran Gene Set Enrichment Analysis

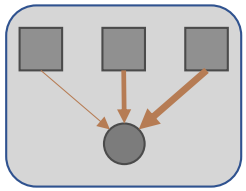
Positively enriched functional categories

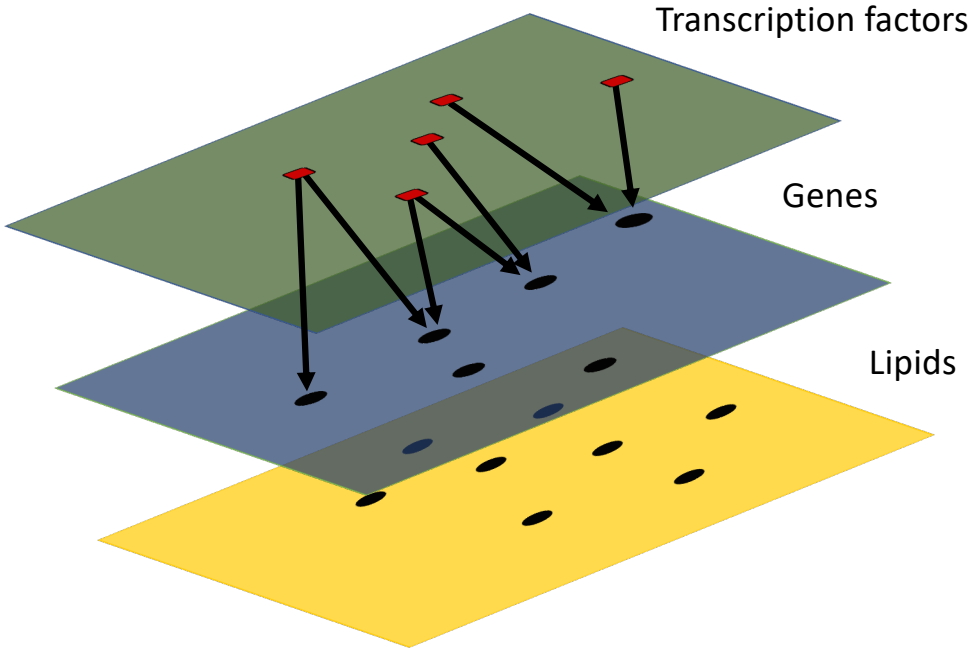
SIG.ID	SIGNATURE	SOURCE	SIZE	ES	FDR
MYCOBACTIN BIOSYNTHESIS	Mycobactin biosynthesis	MANUAL_CURATION	12	0.76	0.0019
ESX-1 LOCUS	ESX-1 LOCUS	MANUAL_CURATION	18	0.58	0.015
GO:0031177	phosphopantetheine binding	GO	15	0.6	0.036

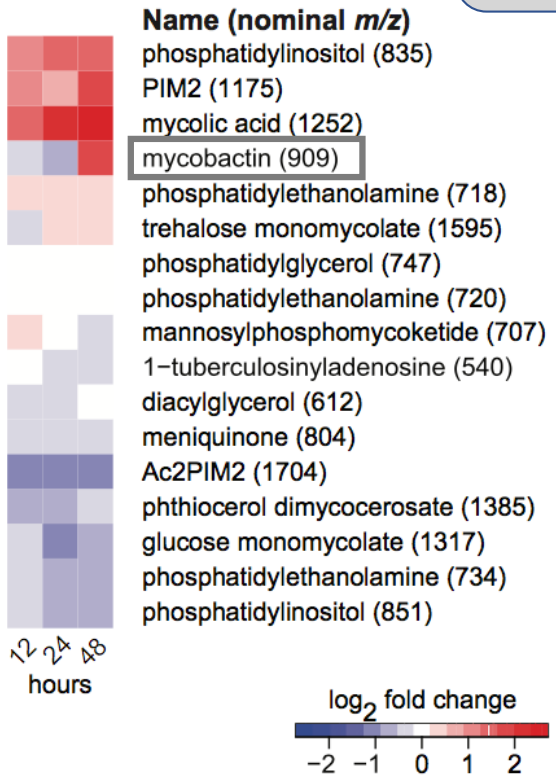
Negatively enriched functional categories

SIG.ID	SIGNATURE	SOURCE	SIZE	ES	FDR
190	Oxidative phosphorylation	KEGG	40	-0.49	0.011
NADH DEHYDROGENASE	NADH dehydrogenase	MANUAL_CURATION	14	-0.62	0.018
GO:0048038	quinone binding	GO	12	-0.61	0.096

Connecting mycobactin gene targeting with pathway output







Mycobactin levels are up 48 hours after PknA/PknB inhibition