

Network Medicine Approaches for the Analysis of Metabolomic Data



JESSICA LASKY-SU, SCD
RACHEL KELLY, PHD

NETWORK APPROACHES WITH
METABOLOMICS
JUNE 25, 2017



Objectives



- To introduce the concepts and theories underlying network based approaches in systems biology
- To explore the utilities, advantages and potential pitfalls of network approaches in the study of metabolomic data

Learning Outcomes



This workshop will provide attendees with the basic skills necessary to apply network approaches in their analysis of metabolomics data moving forward, specifically to;

- *Understand* the concept of network medicine and its role in the study of biology and disease
- *Explore* how metabolic pathways can be related to metabolomic networks
- *Identify* dynamic networks in large metabolomic profiling datasets, and *distinguish* the hubs, nodes and edges comprising these networks
- *Recognize* the limitations and potential pitfalls of these approaches, and consider how the continued development of the field is aiming to address these challenges

Agenda



Time	Presenter	Title, Org, Country	Topic
15:30 – 16:30	Jessica Lasky-Su, Sc.D. (moderator)	Associate Professor Channing Division of Network Medicine, Brigham and Womens Hospital Harvard Medical School	Network theory and the concepts of network medicine
16:30 – 17:30	Rachel Kelly, Ph.D.	Postdoctoral Research Fellow Channing Division of Network Medicine, Brigham and Womens Hospital Harvard Medical School	The application of network medicine to metabolomics datasets

Network Concepts Outline

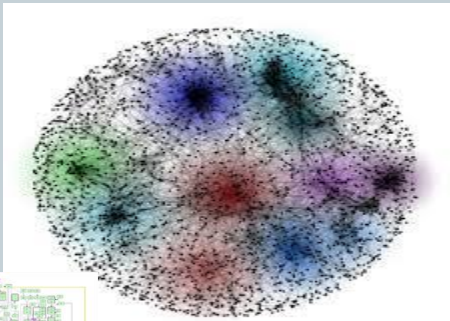


- What are Networks?
- Biological Networks
- Network Terminology & Characteristics
 - Edges, Nodes, Hubs
 - Connectivity, Betweenness
 - Scale Free Networks
- General Approaches for Network Creation
 - Weighted Correlation Gene Network Analysis
 - Bayesian Networks
- Precision Medicine and Networks
 - Redefine phenotypes
 - Novel Disease Treatments
 - DIAMoND

What Is a Network?

- A collection of points (Nodes) that are joined in pairs by lines (Edges). A graphical approach to visualizing and analyzing relationships between variables of interest.

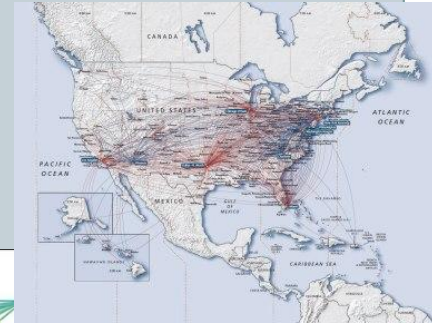
(Adapted From M. Newman, Networks: An Introduction, 2010)



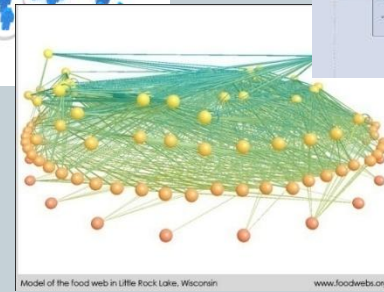
Biological
Networks



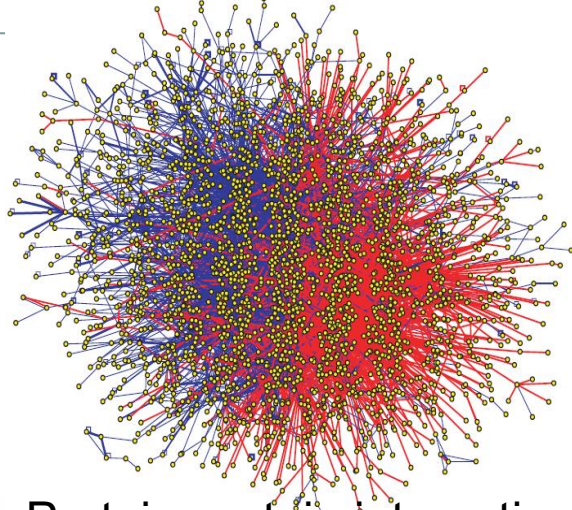
Social Networks



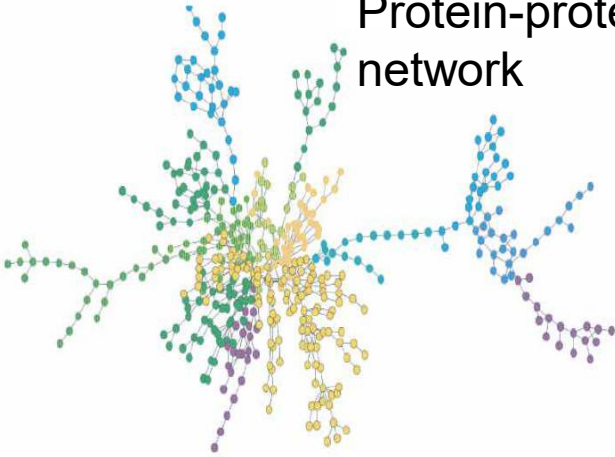
Transportation
Network



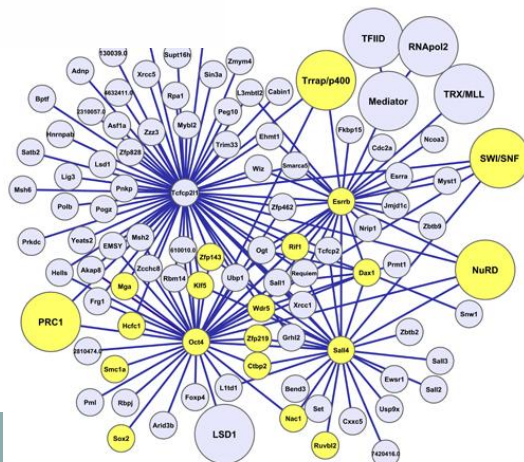
Ecological Network



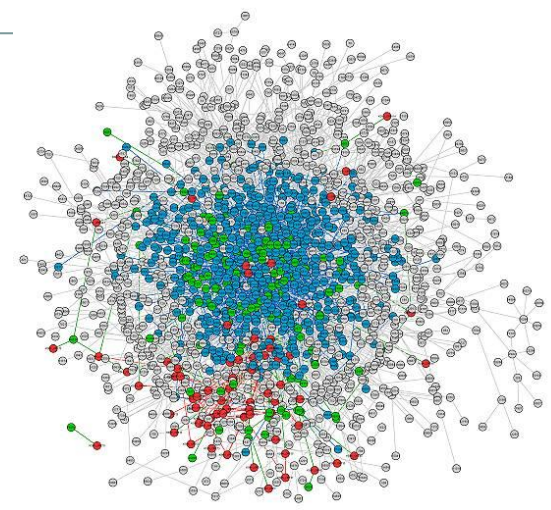
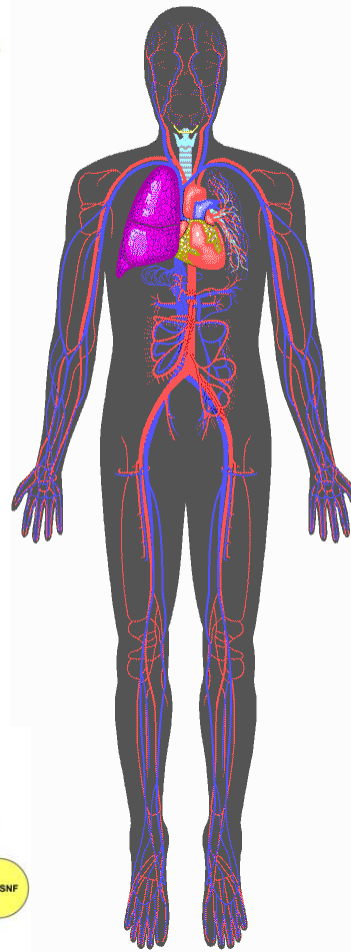
Protein-protein interaction network



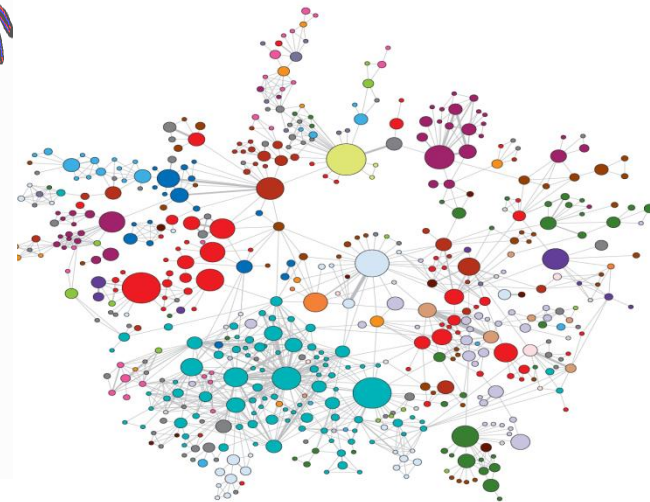
Metabolic network



Transcriptional regulatory network



Virus-host networks



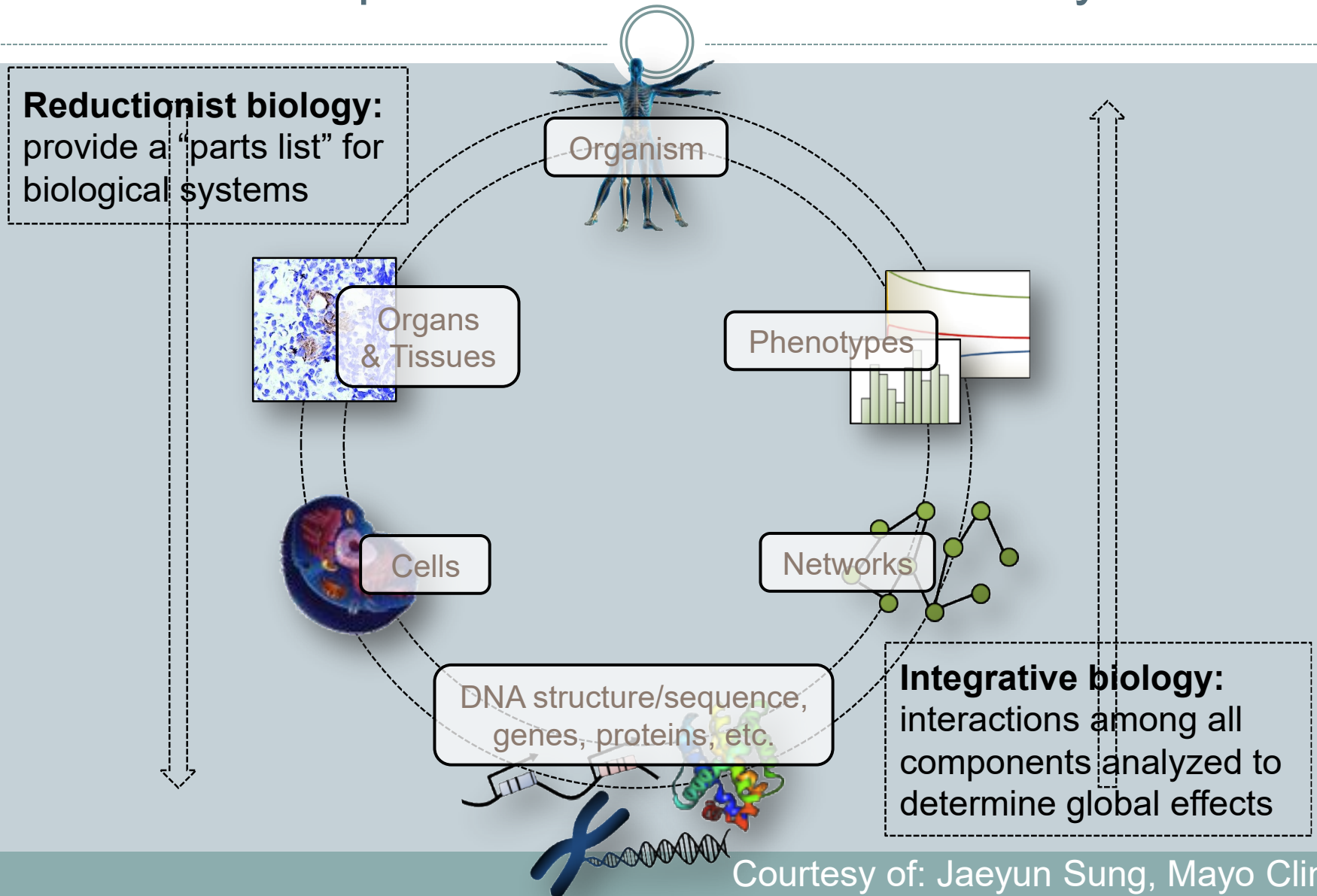
Gene-disease networks

What Is Network Medicine?



- The study of cellular, disease, and social networks which aims to quantify the complex interlinked factors contributing to individual diseases. (Adapted from Barabasi, NEJM 2007; 357:404)
- **Key components of Network Medicine:**
 - Holistic rather than reductionist approach
 - Construction of molecular disease networks
 - Non-linear responses of complex systems
 - Emergent properties from entire network
 - Investigates response of network to various types of perturbation
 - Employs systems biology methods

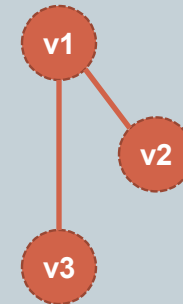
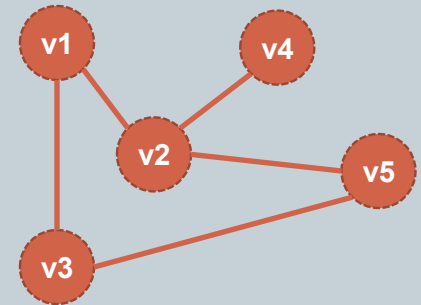
Paradigm of systems biology: From parts to networks to discovery



Network Terminology



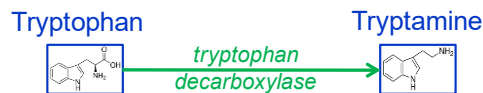
- A “network” is a set of “nodes” connected by a set of “edges”
- A “subnetwork”: a subset of the nodes in the network and all the edges between them.



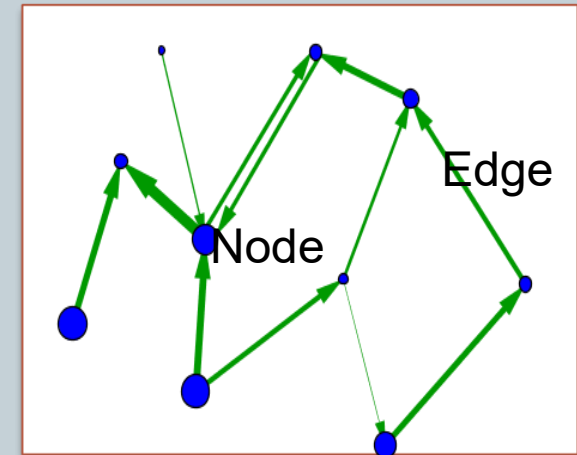
Pathways versus Networks



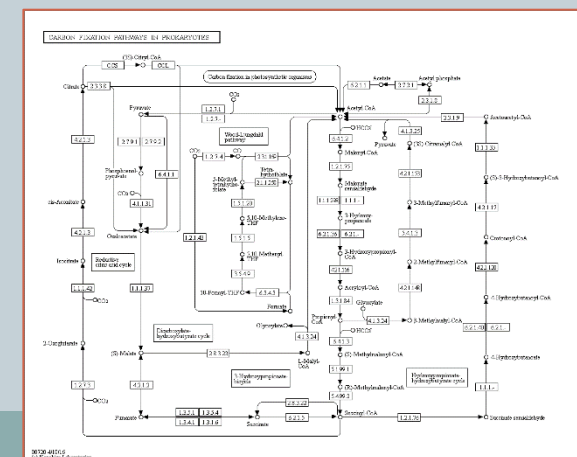
- Networks are data dependent
- Biological pathways are predefined units based on human knowledge
- The two approaches are complimentary



Network

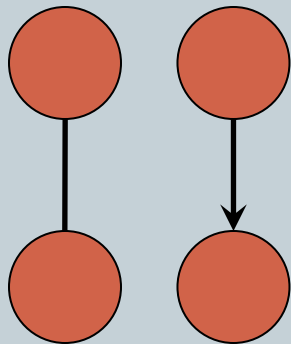


Biological Pathway

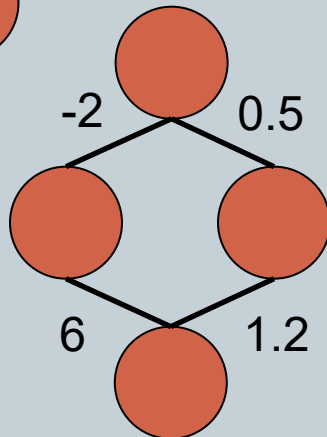


Common Network Structures/Features

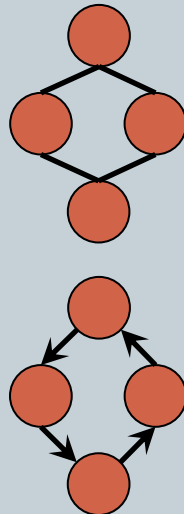
Directed



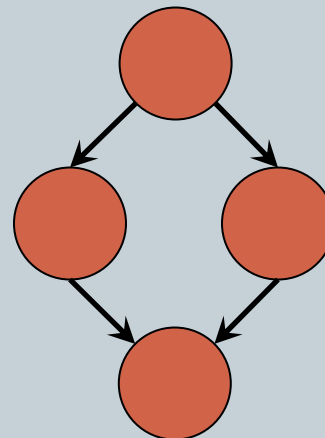
Undirected



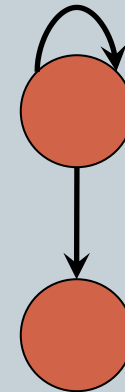
Weighted



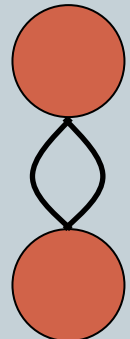
Cyclic



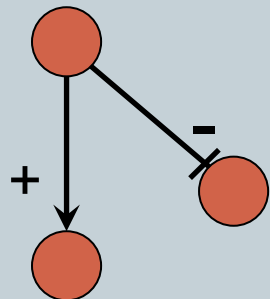
Acyclic
(DAG)



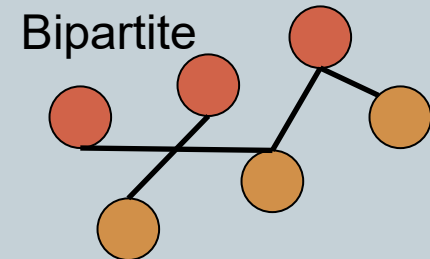
Loops
(Self-connections)



Multigraph



Signed



Bipartite

Directed Acyclic Graph

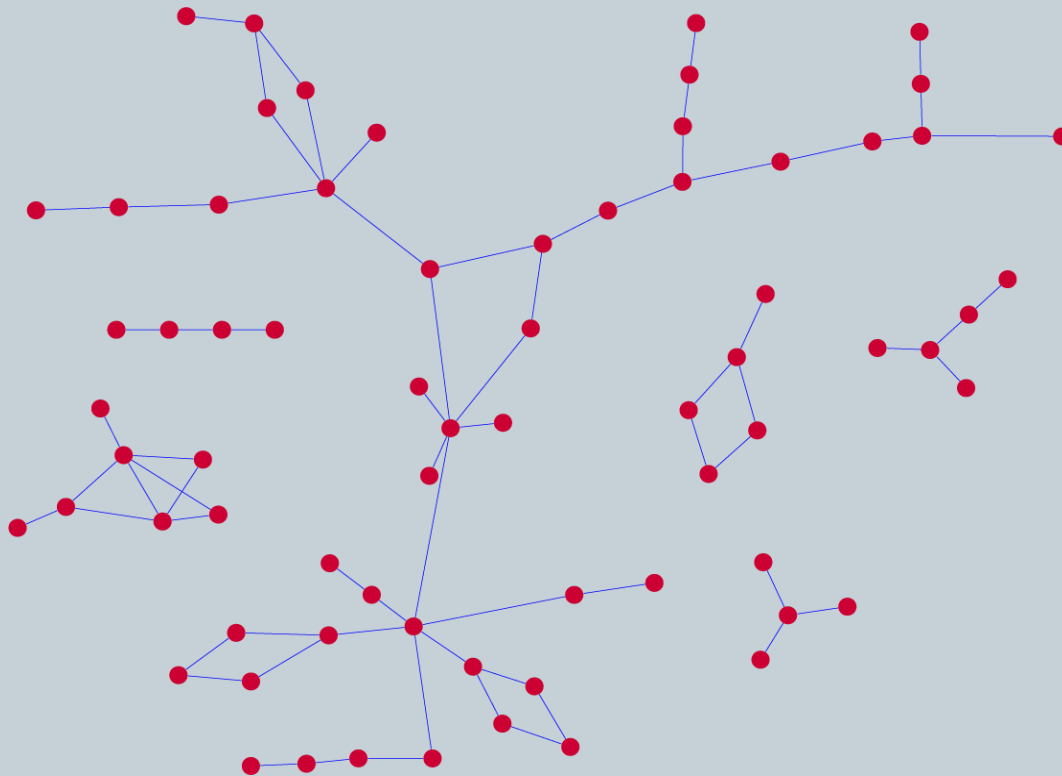
Courtesy of Quackenbush

Topological Properties



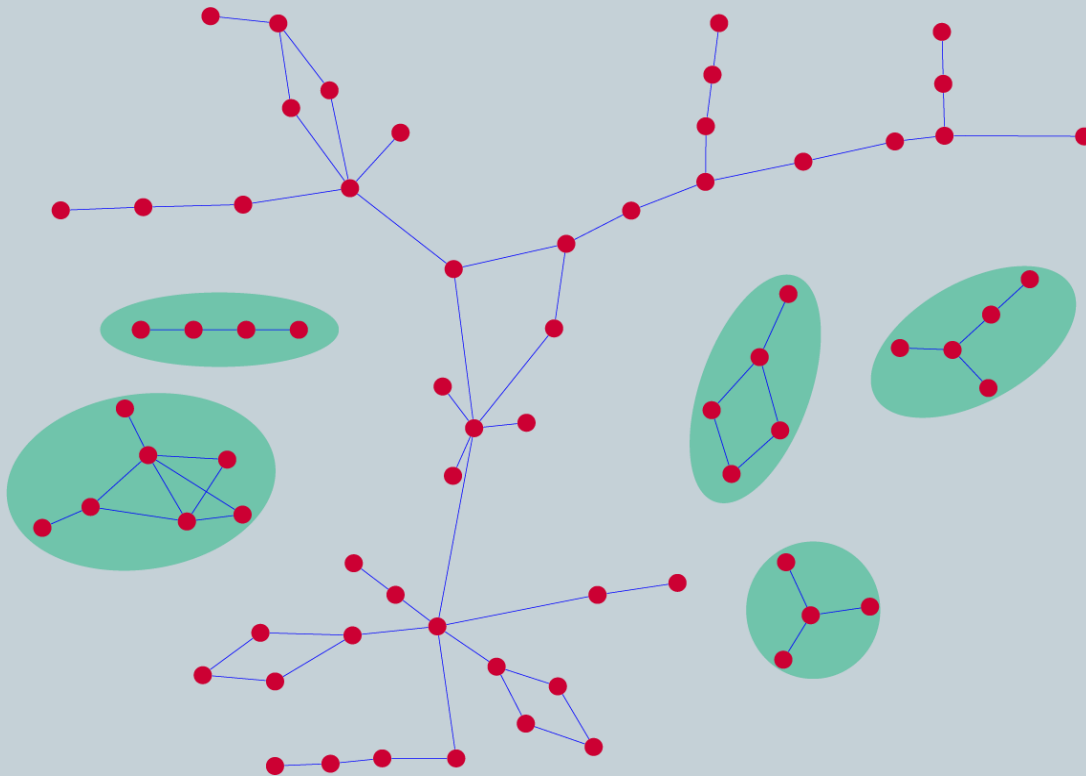
- Importance of a node is determined by the position in the network
- Hubs: highly connected nodes
 - **Degree centrality (connectivity)** measures the number of links that connect to a node within a pathway;
 - **Closeness centrality** measures the overall distance from a given node to all other nodes in a pathway;
 - **Betweenness centrality** measures the number of shortest paths from all nodes to all the others that pass through a given node within a pathway.

Describing the Network Connectivity: Connected Components



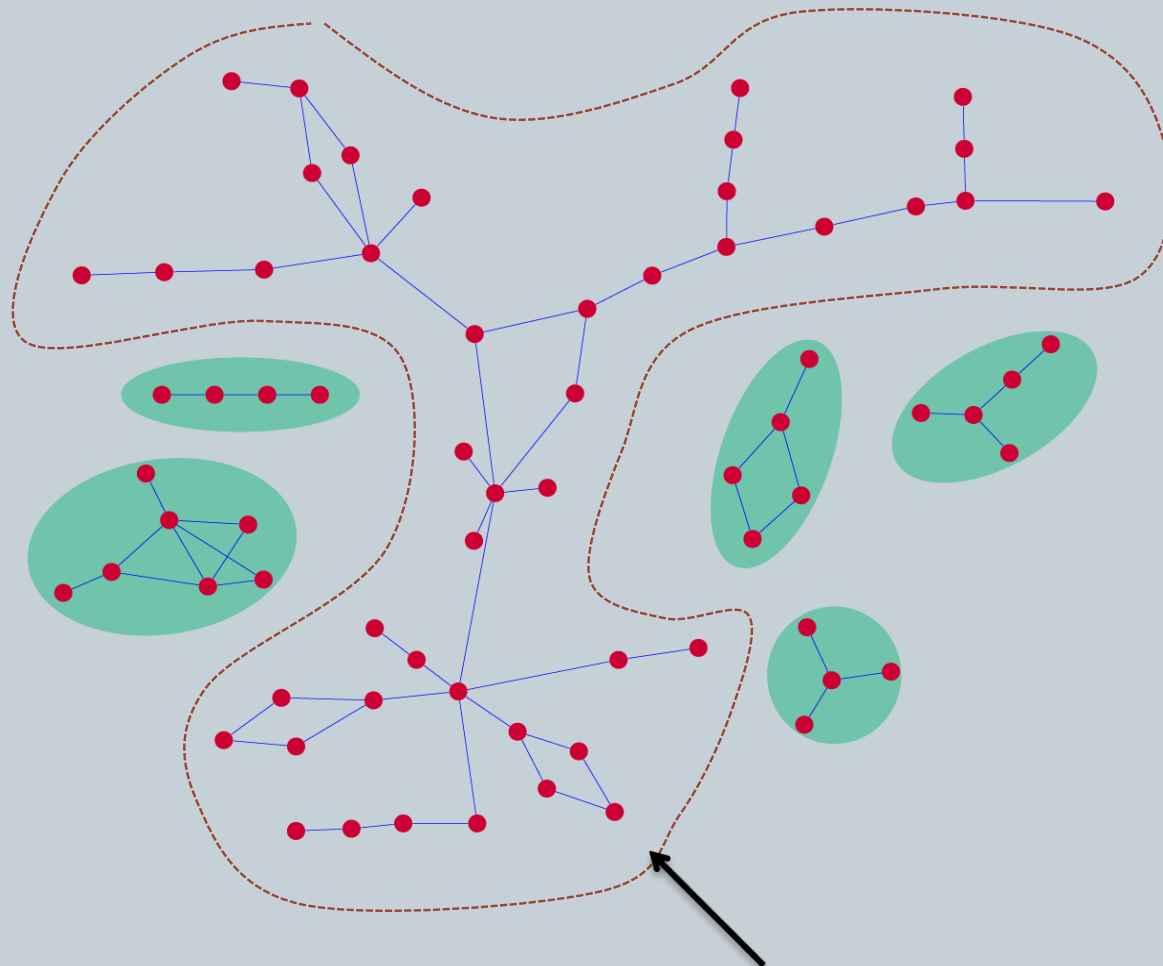
- #nodes = 69
- #edges = 71

Describing the Network Connectivity: Connected Components



- #nodes = 69
- #edges = 71
- 6 connected components

Describing the Network Connectivity: Connected Components



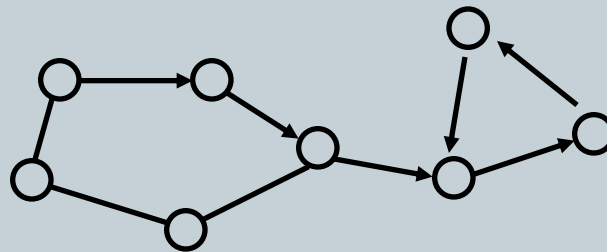
- #nodes = 69
- #edges = 71
- 6 connected components

Largest Connected Component

Describing the Network Connectivity: Network Paths

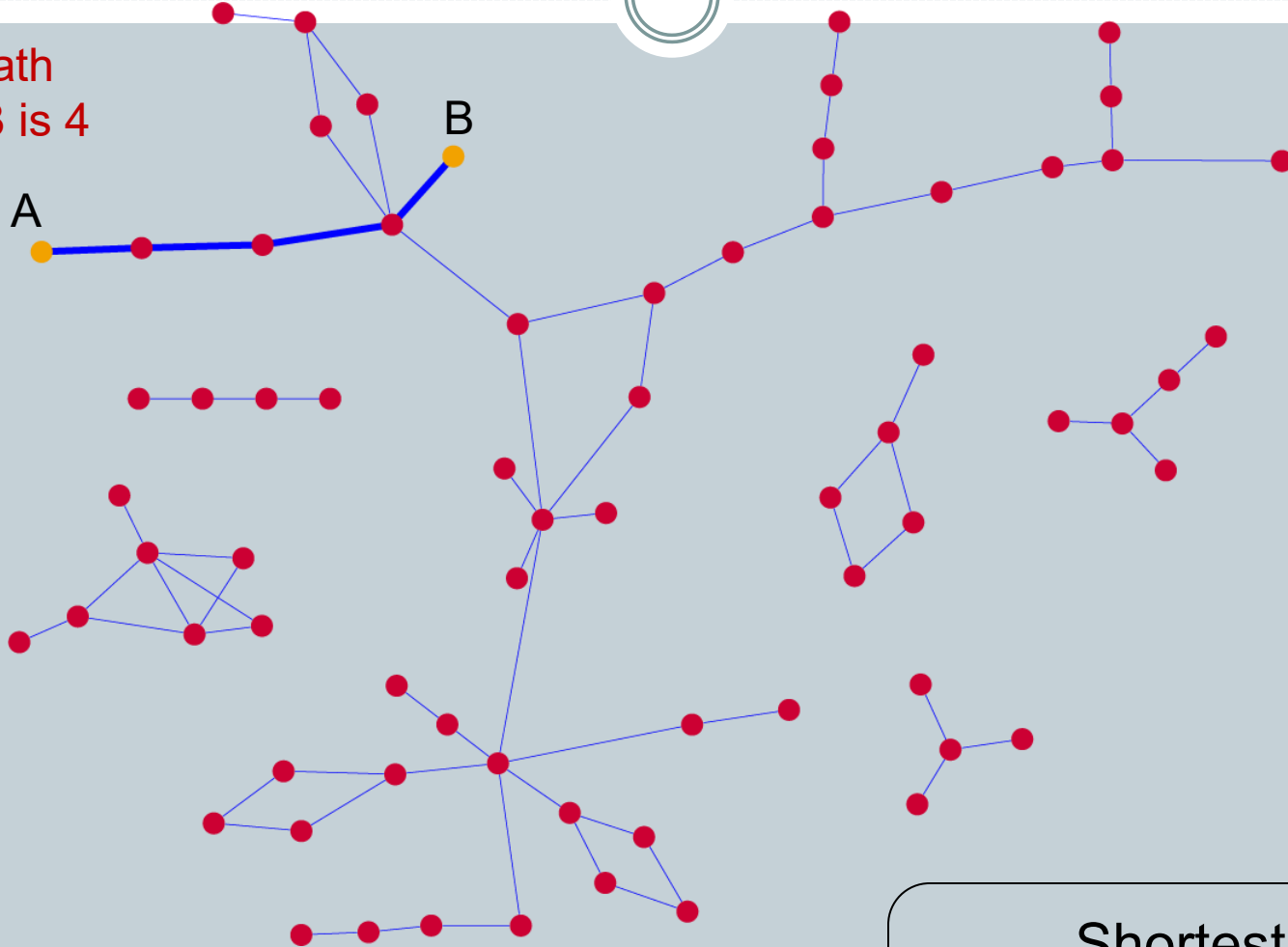
A **path** is a sequence of **edges** in the network

A closed path on a network is called a **graph cycle** or **circuit**.



Shortest-Path between nodes

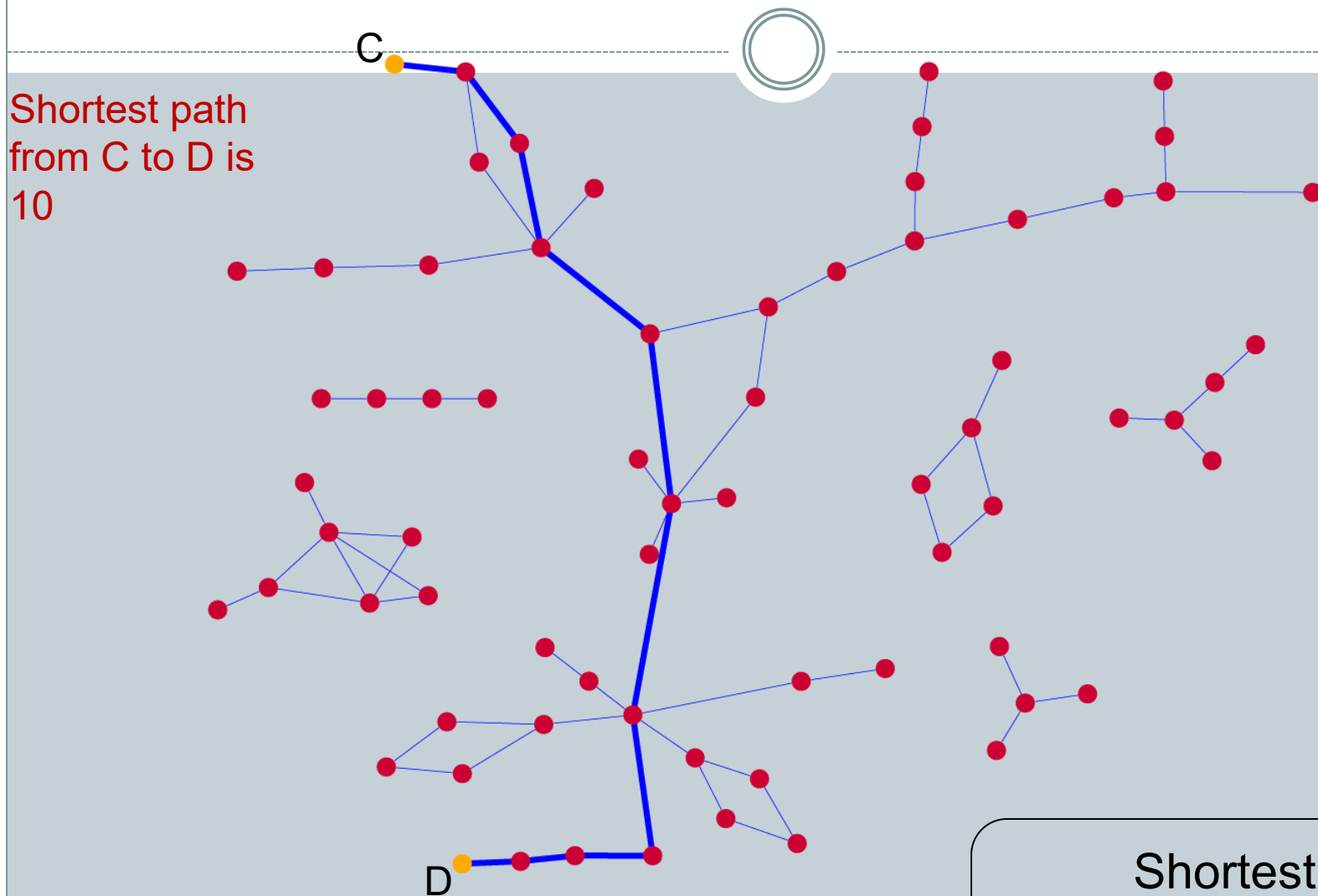
Shortest path
from A to B is 4



Shortest path:

Connect two nodes by as few
edges as possible

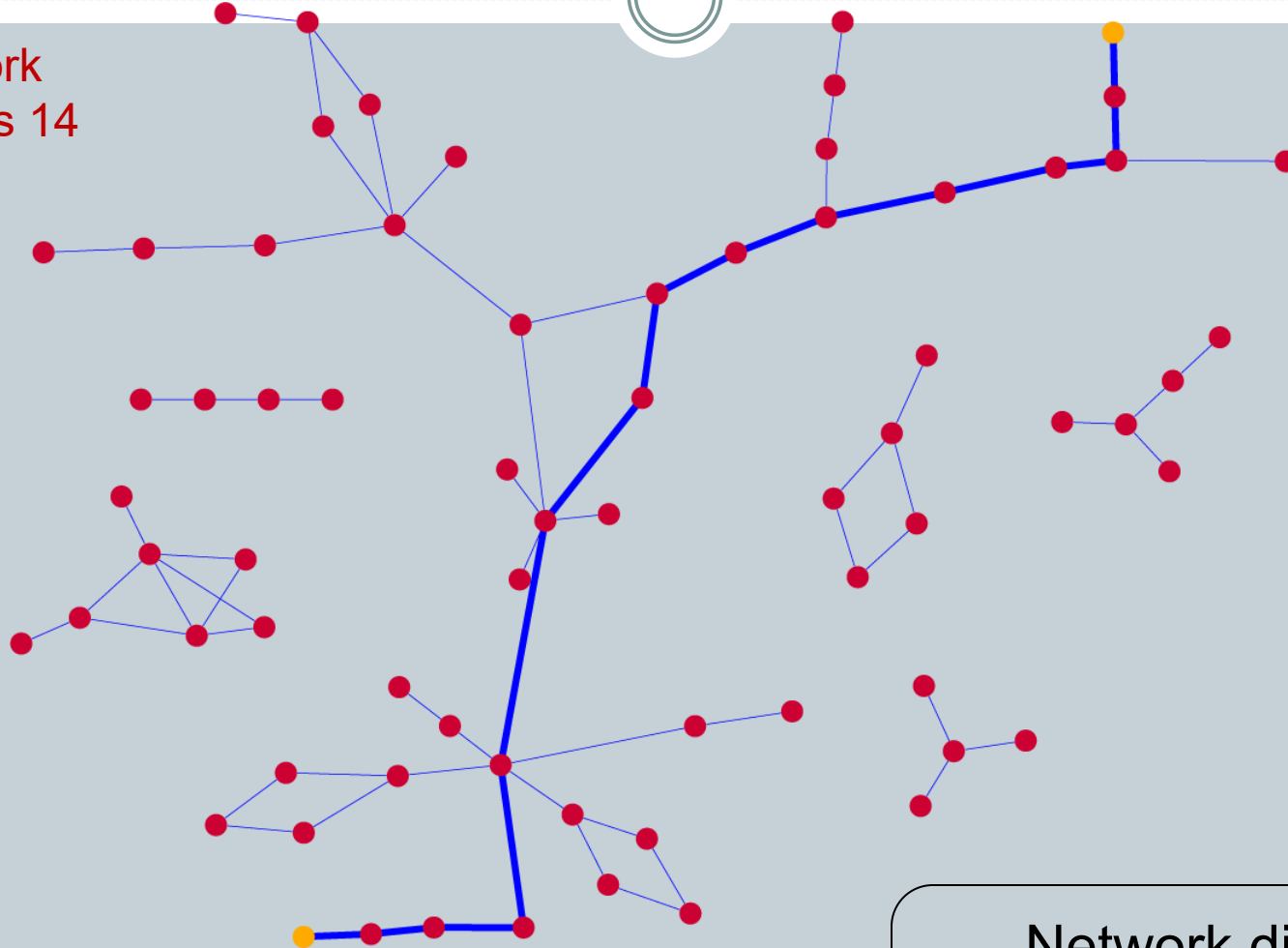
Shortest-Path between nodes



Shortest path:
Connect two nodes by as few
edges as possible

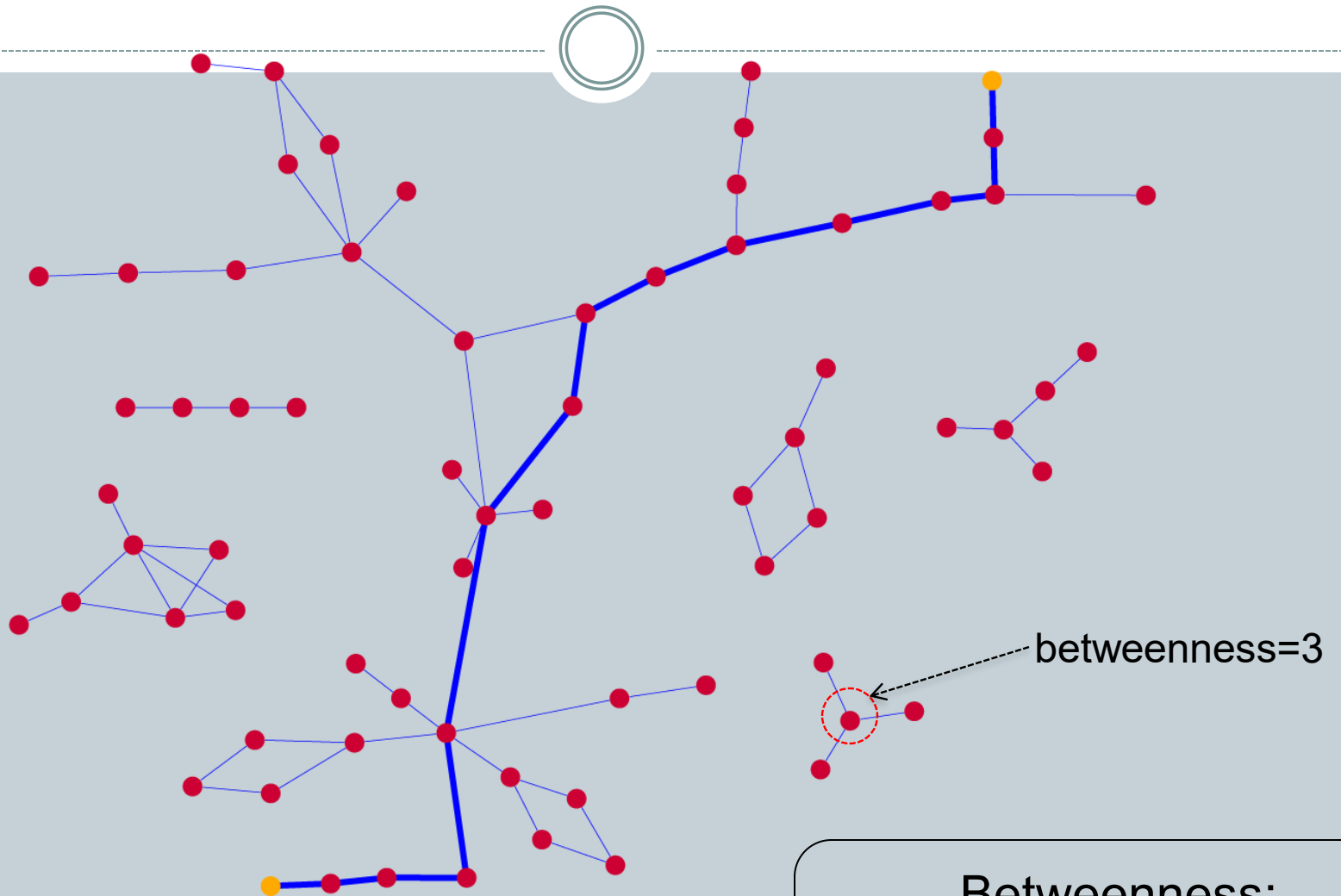
Longest Shortest-Path

The Network
Diameter is 14



Network diameter:
The longest shortest path in
the network

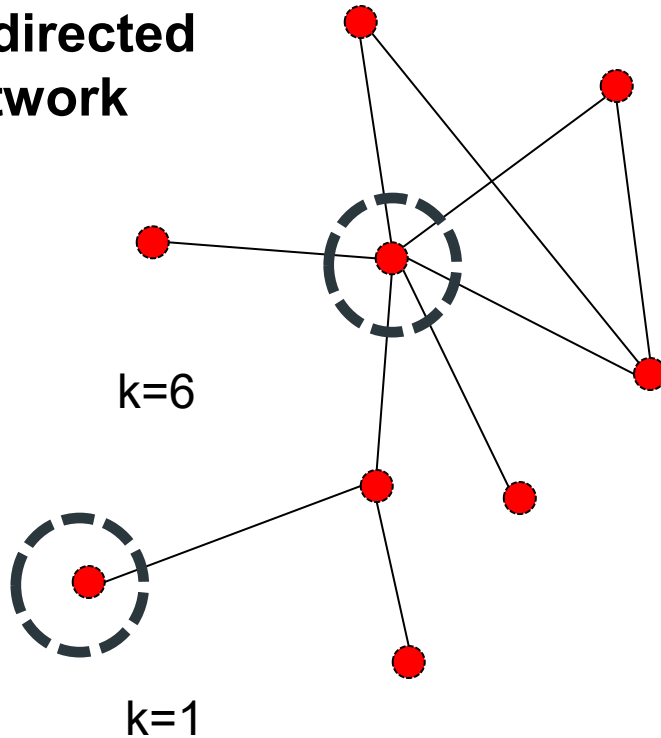
Betweenness



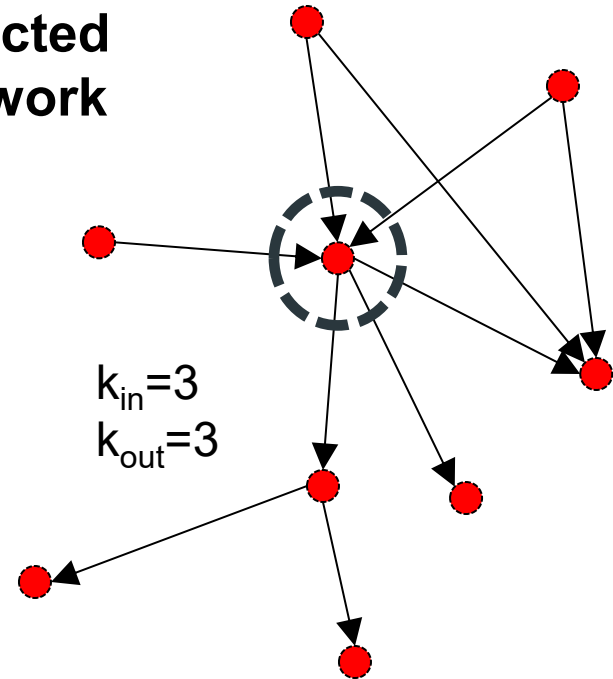
Betweenness:
Number of Shortest paths
through a node/edge

Node Connectivity: Degree

Undirected Network

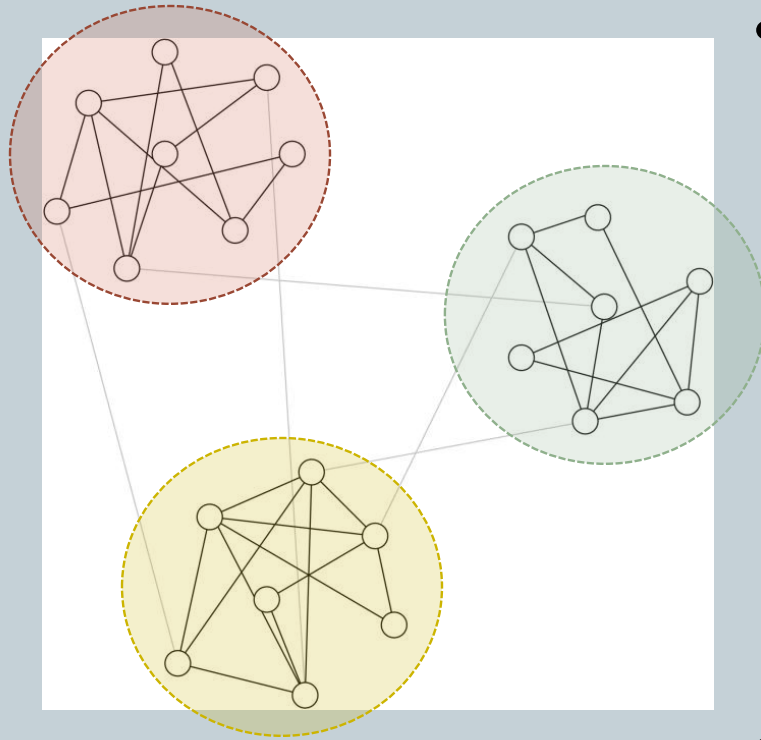


Directed Network



Degree: The number of edges associated with a node
Hub: a node with a large number of edges

Community Structure

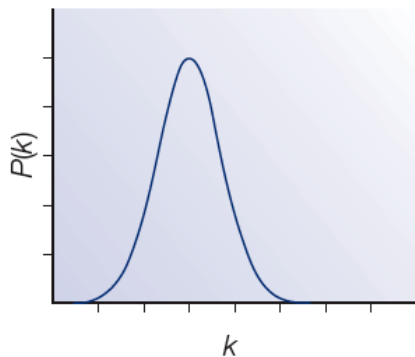
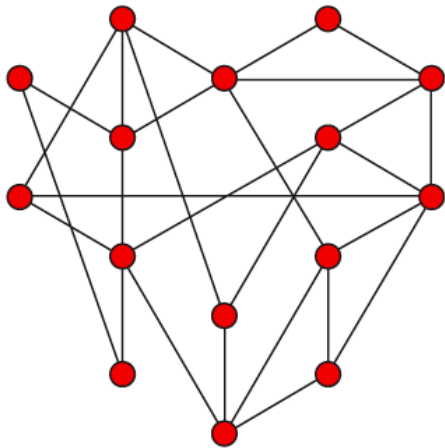


- **Networks Communities:** sets of nodes within which there are many more connections than expected by chance, but between which there are many fewer.
- These communities are so known as **Modules**

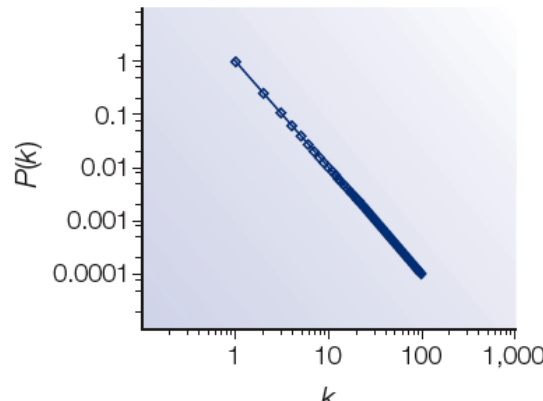
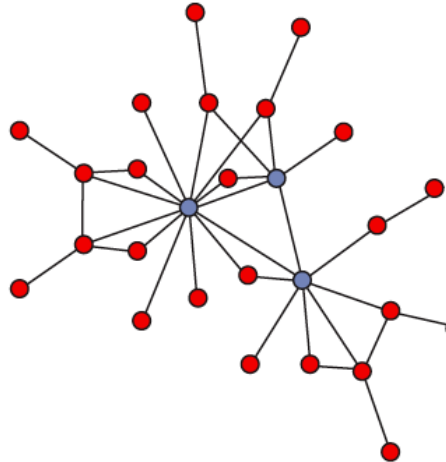
Community structure in social and biological networks

Network Degree Distribution

A Random network



B Scale-free network



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

For random networks, $P(k)$ is normally distributed.

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

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

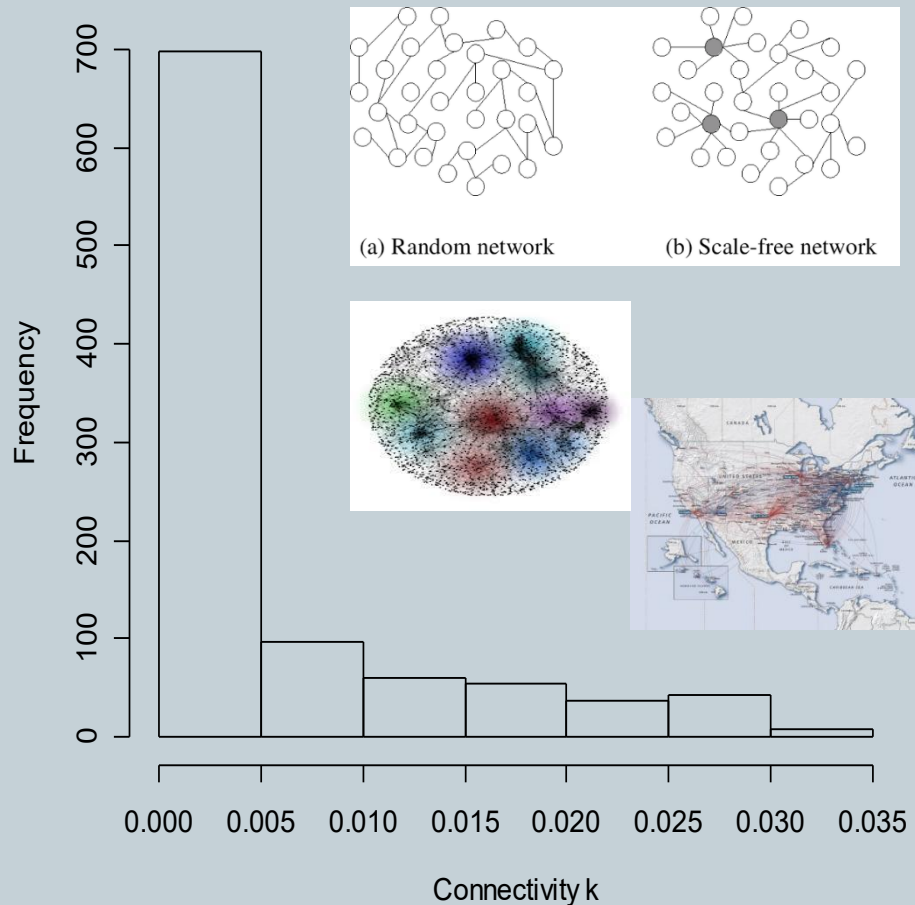
Such networks are said to be **scale-free**

Metabolomic networks are scale-free

$P(k)$ vs k in scale free networks

$P(k)$

Frequency Distribution of Connectivity



- Barabasi has identified approximate scale free topology as a fundamental property of networks
 - The existence of **hubs** that are connected through multiple **nodes**
 - Robust to the removal of **nodes**
 - Sensitive to the removal of **hubs**

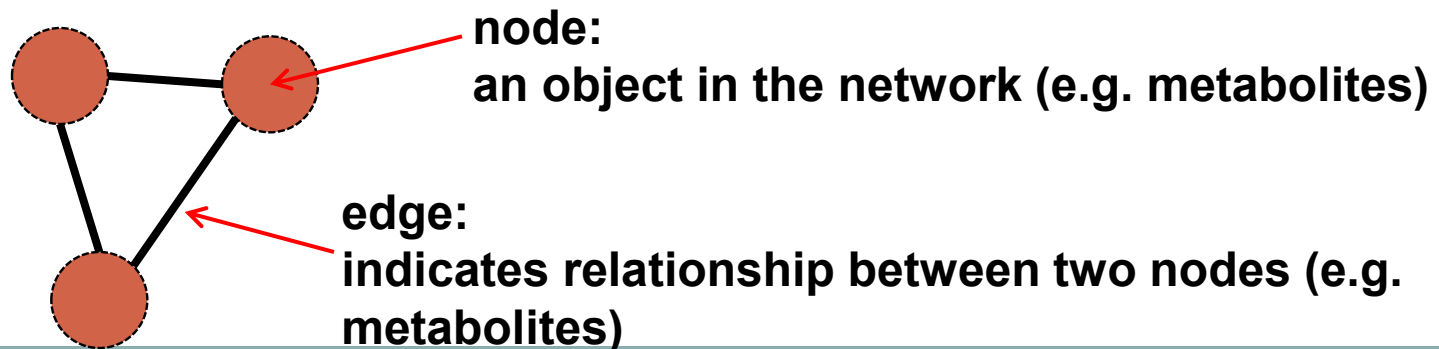




How to define a metabolite co-expression network?

Building & Interpreting Biological Networks

- How we construct a network depends on what data we have **AND** what we want the edges in the network to represent.
⇒ Influences which method we use to reconstruct a network.
- The meaning of the edges in a reconstructed network depend on the method used to generate those edges.
⇒ Influences how we interpret the interactions in a network.



Interpreting the “edges” in Biological Networks

Co-expression Networks



Correlation Network

- Undirected (non-causal relationships)
- Nodes all of same “type”
- Edges can have “signs”

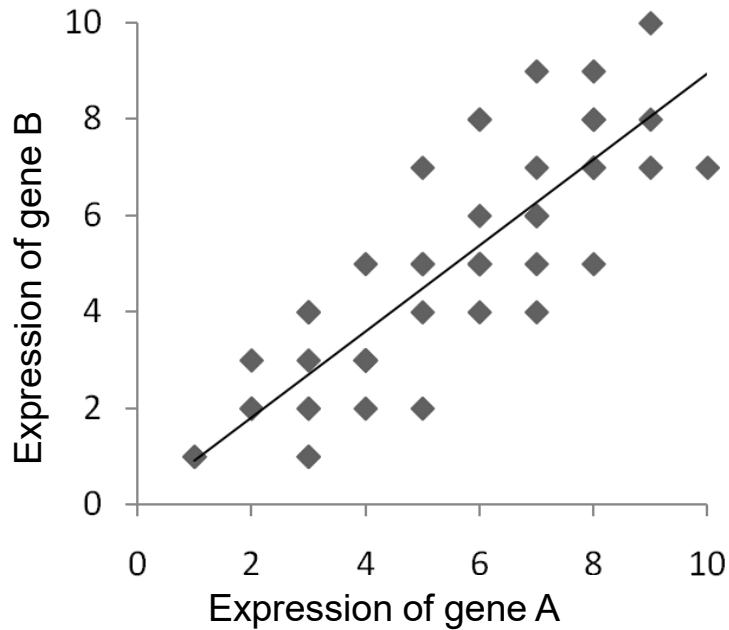
Example: When metabolite **A** changes, so does metabolite **B**.

Data:

high-throughput data

(metabolomics, expression, methylation, etc.)

Correlation is not Causation



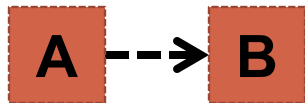
Co-expression occurs when:

(1) One metabolite regulates another.

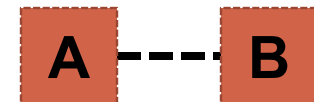
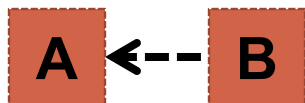
(2) Two metabolites are targeted by the same regulator.

Both are represented the same way in a co-expression network.

Co-expression Network

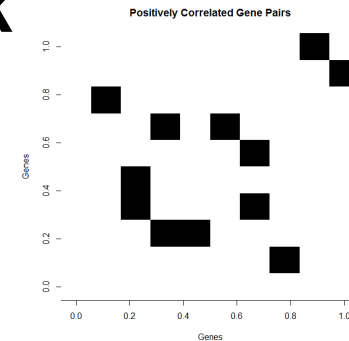
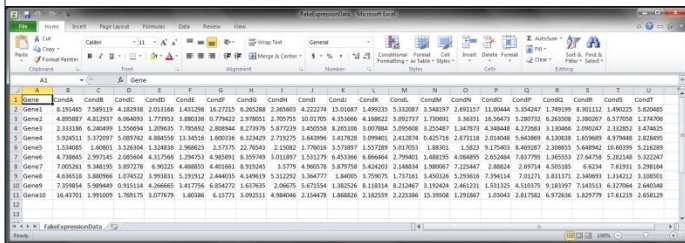


OR

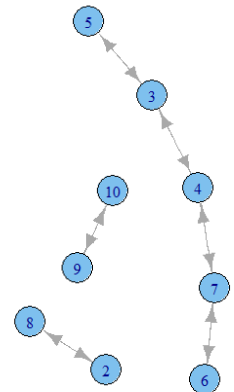
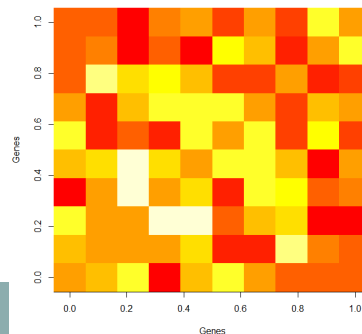


Four Main Steps for Building a Co-expression Network

- 1) Read in the metabolomic data
- 2) Calculate correlation between all metabolite-pairs
- 3) Select top co-expressed metabolite-pairs
- 4) Visualize/Analyze the Network

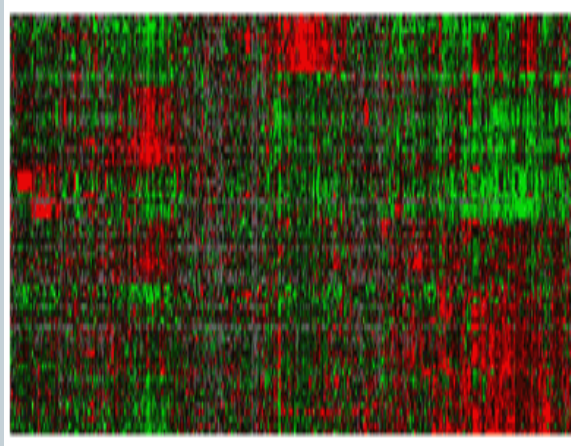


Pearson Correlation between Gene Pairs



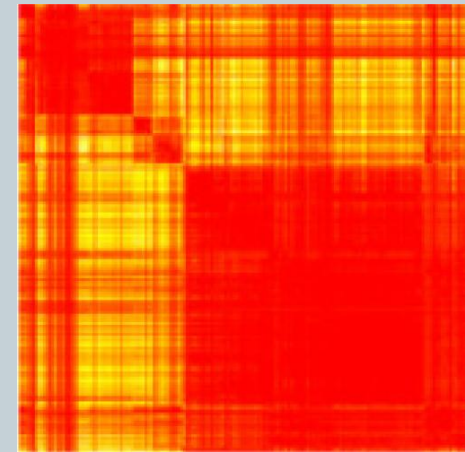
Pearson Correlation is one of the simplest metrics for co-expression

conditions



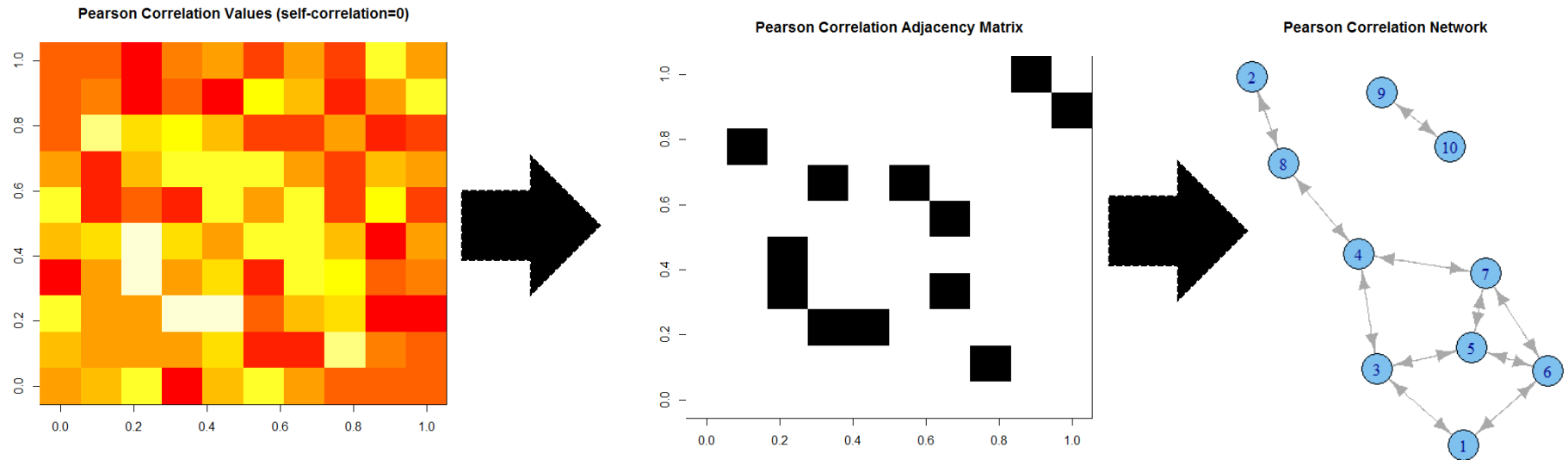
metabolites

metabolites



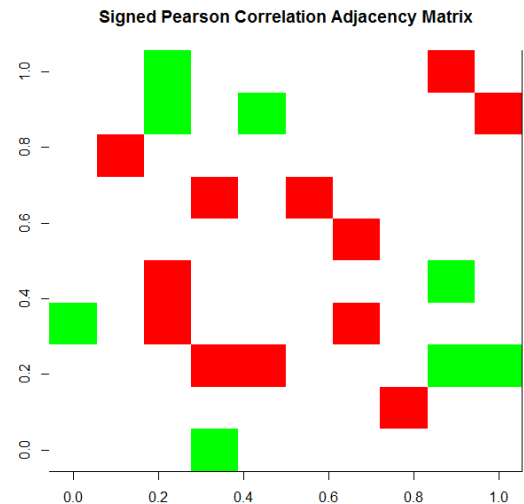
$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

We can turn correlation values into a co-expression network by thresholding



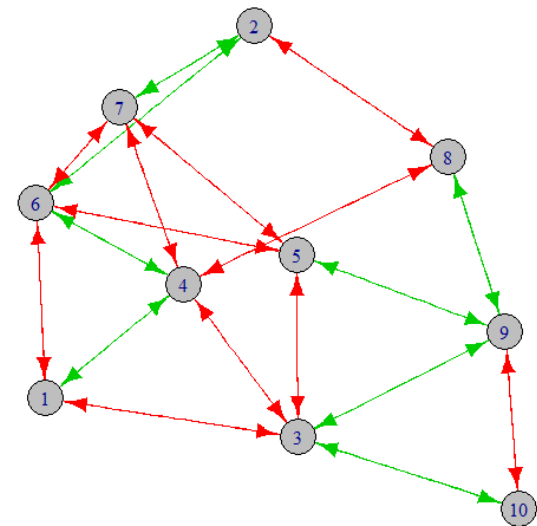
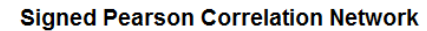
Building a Signed Co-expression Network

“Signed” adjacency matrix



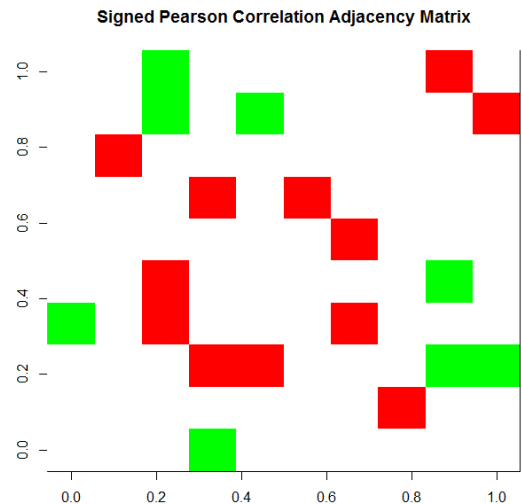
■ 1 (positively correlated)

■ -1 (negatively correlated)



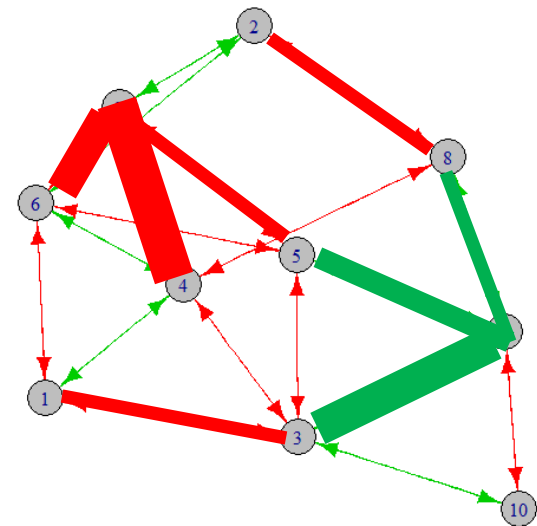
Building a Signed & Weighted Co-expression Network

“Signed” adjacency matrix



- 1 (positively correlated)
- 1 (negatively correlated)

Signed Pearson Correlation Network



WGCNA

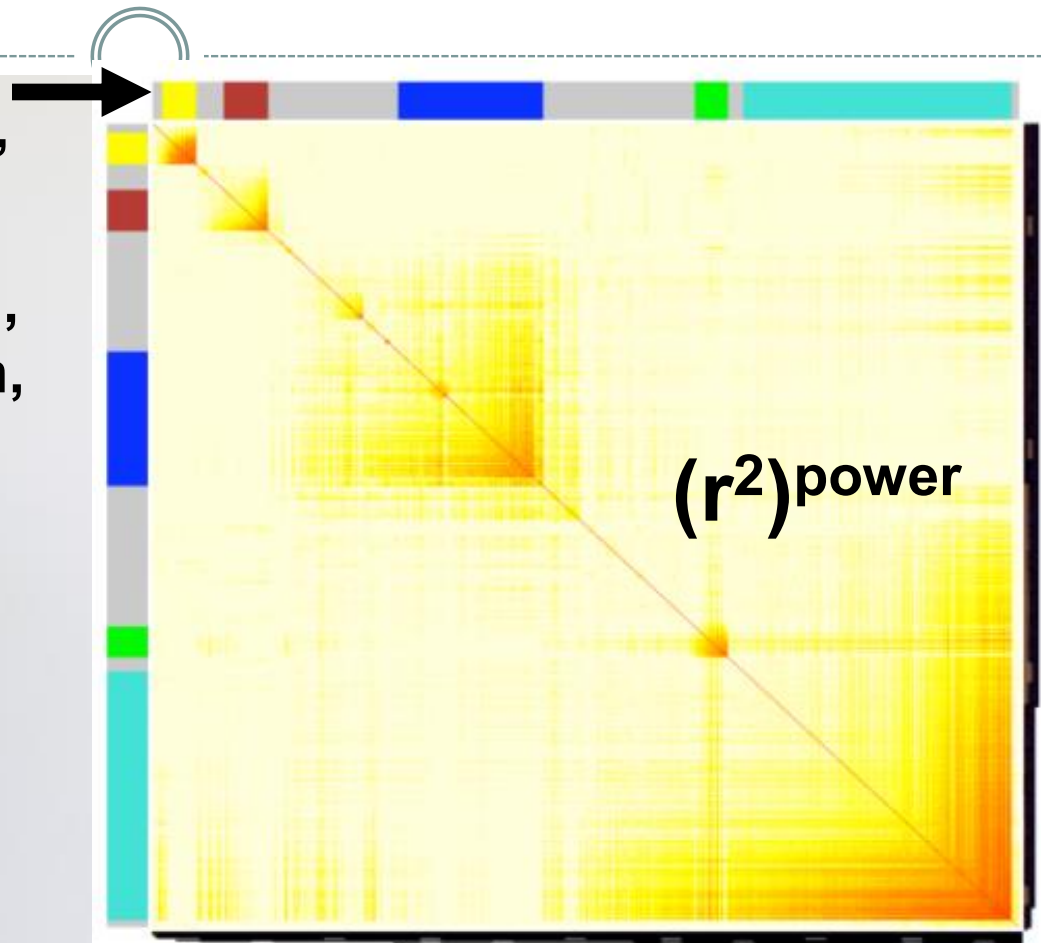
(Weighted Gene Co-expression Network Analysis)

Zhang, Horvath. *Stat Appl Genet Mol Biol*. 2005;4:Article17. Epub 2005 Aug 12..

Key Idea: Scale co-expression for biological topology.



five gene
“modules”
detected:
yellow, red,
blue, green,
cyan.



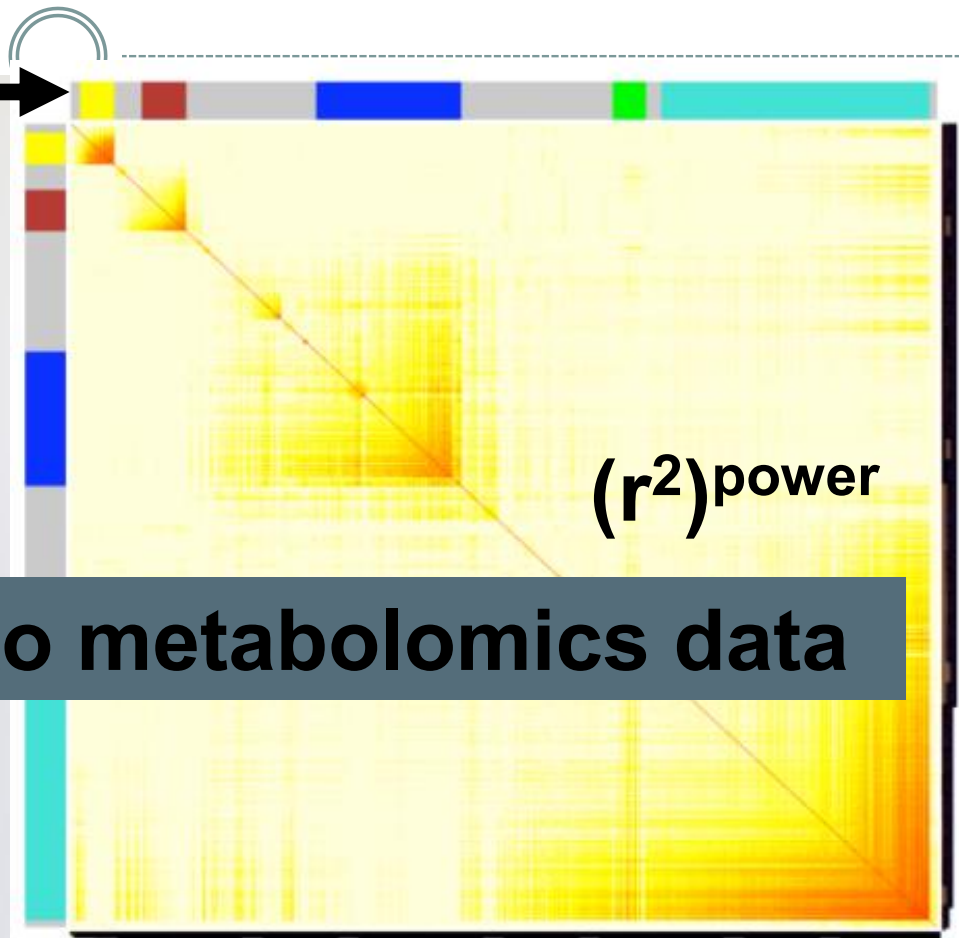
WGCNA

(Weighted Gene Co-expression Network Analysis)

Zhang, Horvath. *Stat Appl Genet Mol Biol*. 2005;4:Article17. Epub 2005 Aug 12..

Key Idea: Scale co-expression for biological topology.

five gene
“modules”
detected:
yellow, red,
blue, green,
cyan.



Can also be applied to metabolomics data

Construct a network

Rationale: make use of interaction patterns between genes

Identify modules

Rationale: module (pathway) based analysis

Find the key drivers of *interesting* modules

Rationale: experimental validation, therapeutics, biomarkers

Relate modules to external information

Array Information: Clinical data, SNPs, proteomics

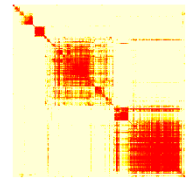
Gene Information: gene ontology, EASE, IPA

Rationale: find biologically interesting modules

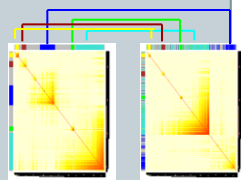
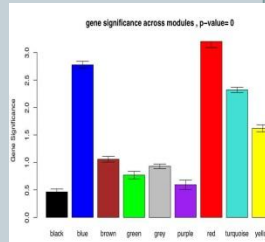
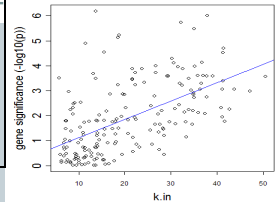
Study Module Consensus and Preservation across different data

Rationale:

- Same data: to check robustness of module definition
- Different data: to find interesting modules

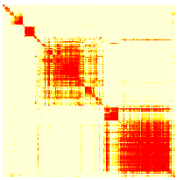


correlation=0.56, pvalue=2.2e-16



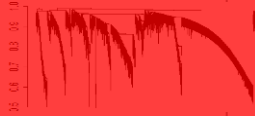
Construct a network

Rationale: make use of interaction patterns between genes



Identify modules

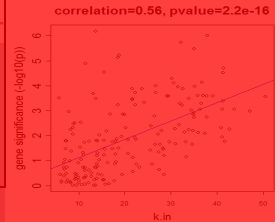
Rationale: module (pathway) based analysis



Find the key drivers of *interesting* modules

Rationale: experimental validation of the module, high links

Measures of Connectivity

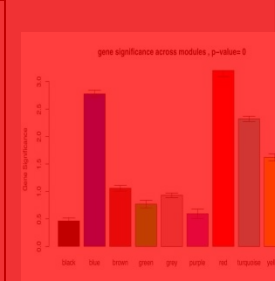


Relate modules to external information

Array Information: Clinical data, SNPs, proteomics

Gene Information: gene ontology, EASE, IPA

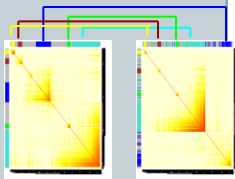
Rationale: find biologically interesting modules



Study Module Consensus and Preservation across different data

Rationale:

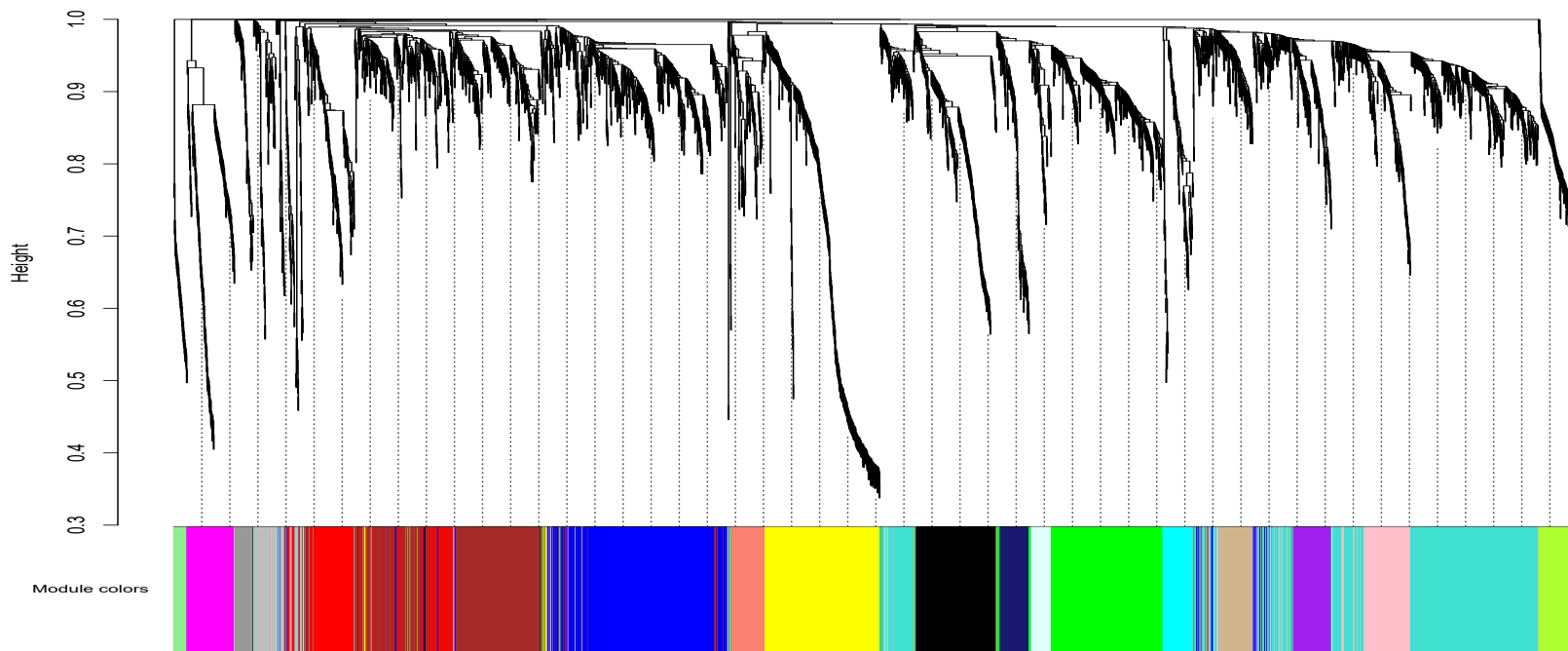
- Same data: to check robustness of module definition
- Different data: to find interesting modules



Generating the network modules



Cluster Dendrogram



- Modules are identified using measures of connectivity of metabolites

What is clustering?

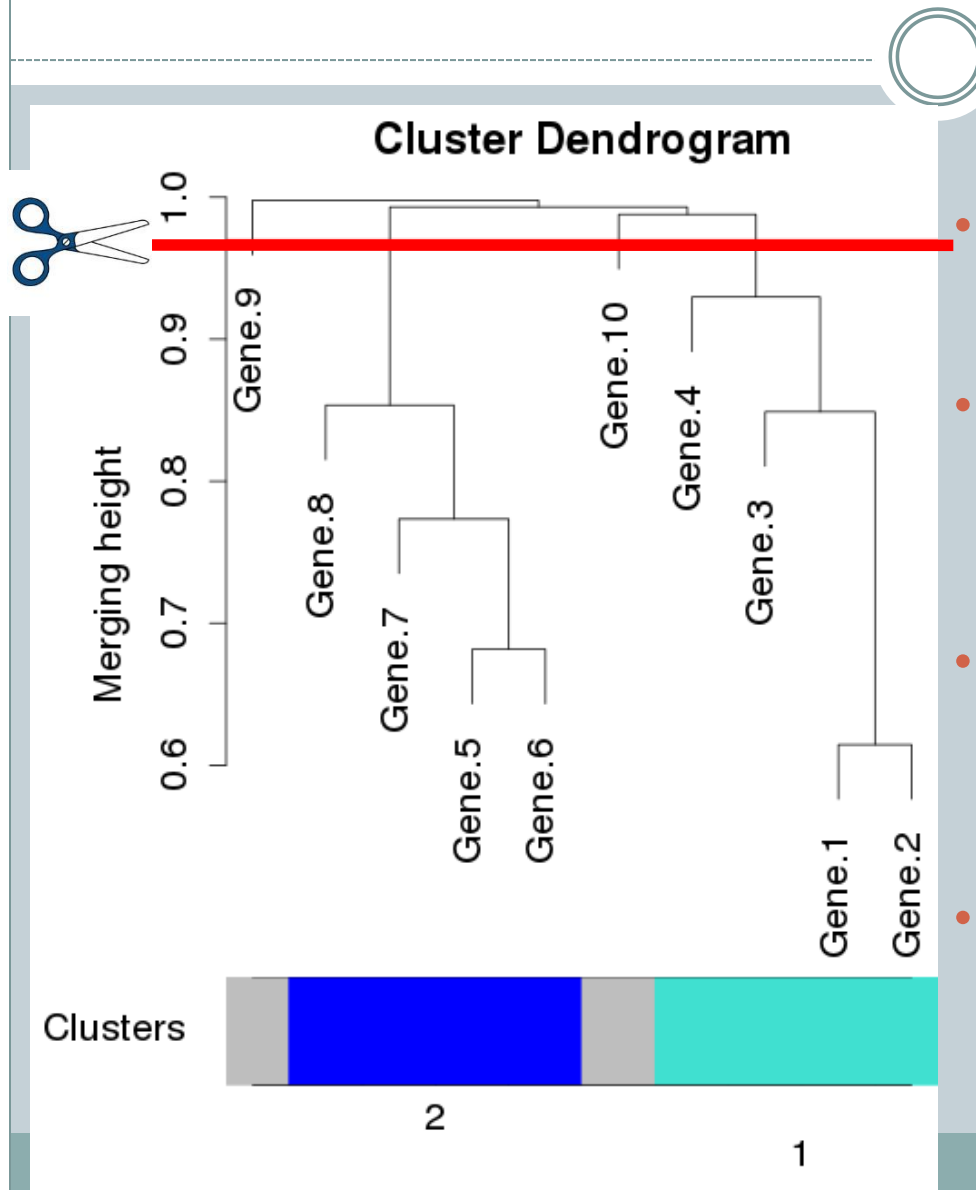


- Input: pair-wise dissimilarities between objects
- Aim: find clusters (groups) of objects that are closely related according to the (dis-)similarity measure
- Output: a cluster label for each object

(r^2)power

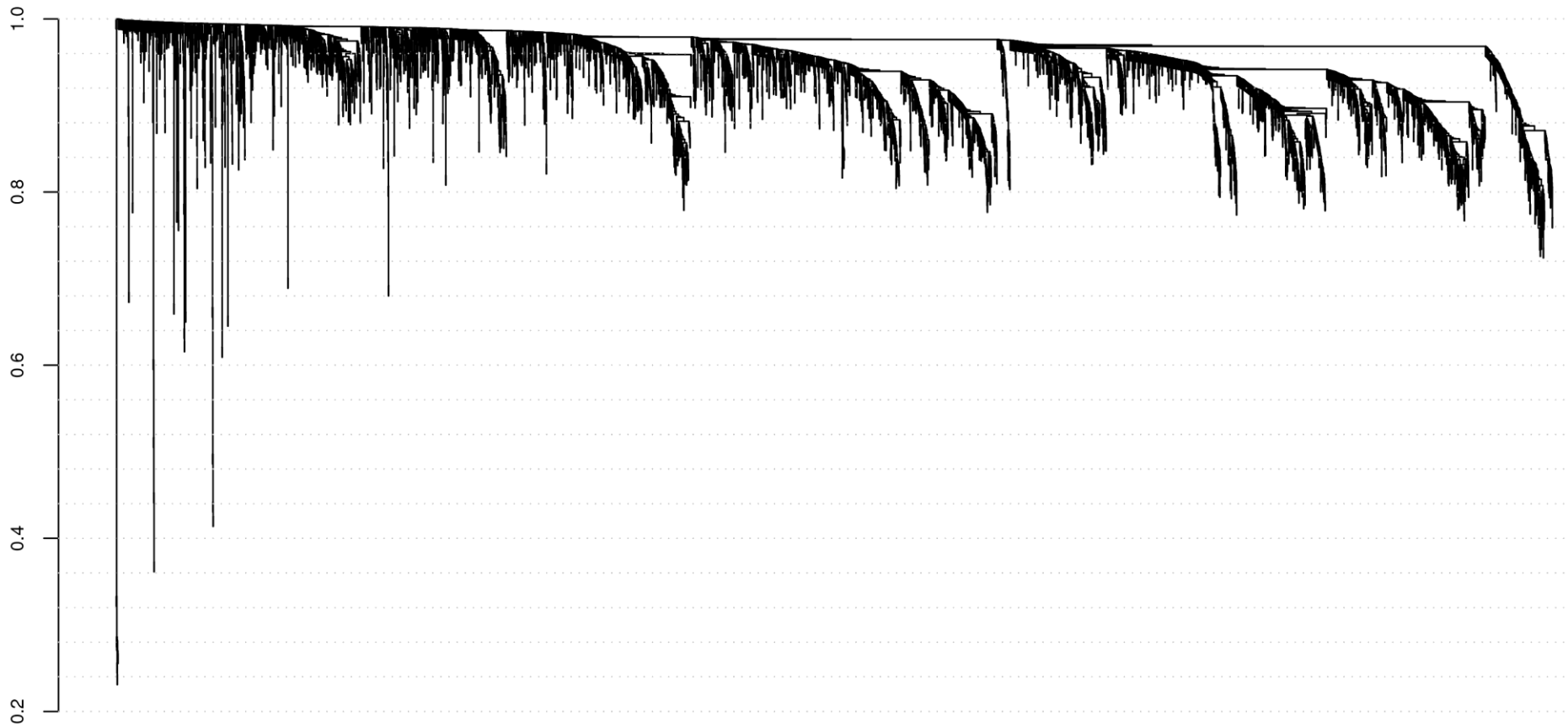
Gene.9	0	1	1	1	1	0.9	0.9	0.98	1	1
Gene.8	1	0	0.63	0.55	0.69	1	0.94	0.82	0.92	0.98
Gene.7	1	0.63	0	0.49	0.56	1	1	0.76	1	1
Gene.5	1	0.55	0.49	0	0.44	1	0.94	0.88	1	1
Gene.6	1	0.69	0.56	0.44	0	0.95	1	0.9	1	0.91
Gene.10	0.9	1	1	1	0.95	0	1	1	0.88	0.82
Gene.4	0.9	0.94	1	0.94	1	1	0	0.76	0.82	0.65
Gene.3	0.98	0.82	0.76	0.88	0.9	1	0.76	0	0.67	0.56
Gene.1	1	0.92	1	1	1	0.88	0.82	0.67	0	0.38
Gene.2	1	0.98	1	1	0.91	0.82	0.65	0.56	0.38	0

Cut height must be chosen carefully!



- Different cut heights lead to very different results
- Proper cut height setting requires an intelligent operator
- In general, each application will require a different cut height
- This is a major disadvantage of constant-height tree cut

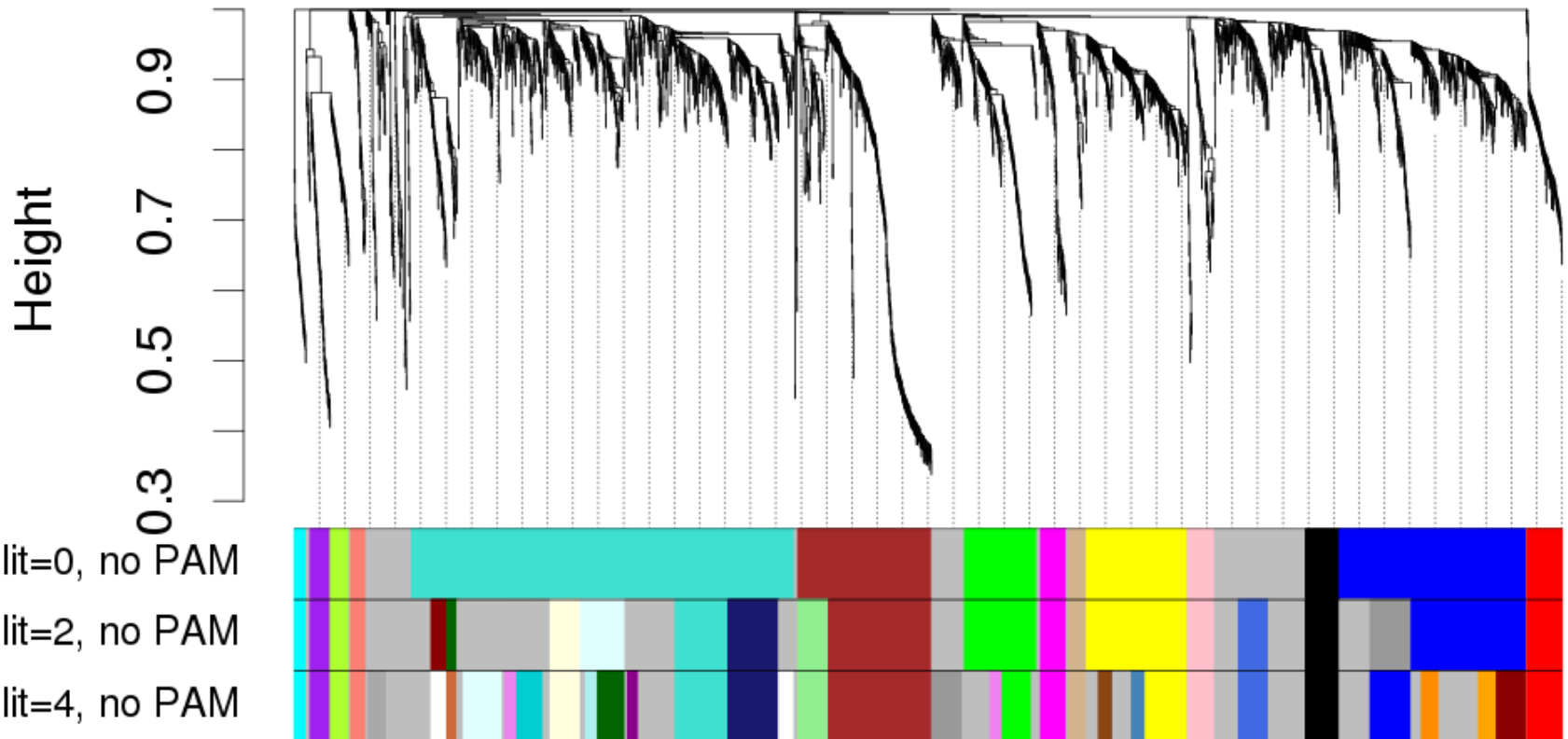
The simple solution does not always work for complicated trees



- Clustering of mouse adipose expression data: no single cut height captures the prominent branches
- Dynamic Tree Cut *Langfelder P, Zhang B, Horvath S (2008), [Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R](#), Bioinformatics 24:719*

Module selection is dependent on objectives

- `deepSplit` controls how finely the branches should be split
- Higher values give more smaller modules, low values (0) give fewer larger modules



Why (not) Build a Co-expression Network using Pearson Correlation?

Reasons to use Pearson correlation to estimate co-expression:

- Easy to understand
- Simple to calculate

Limitations of using Pearson Correlation:

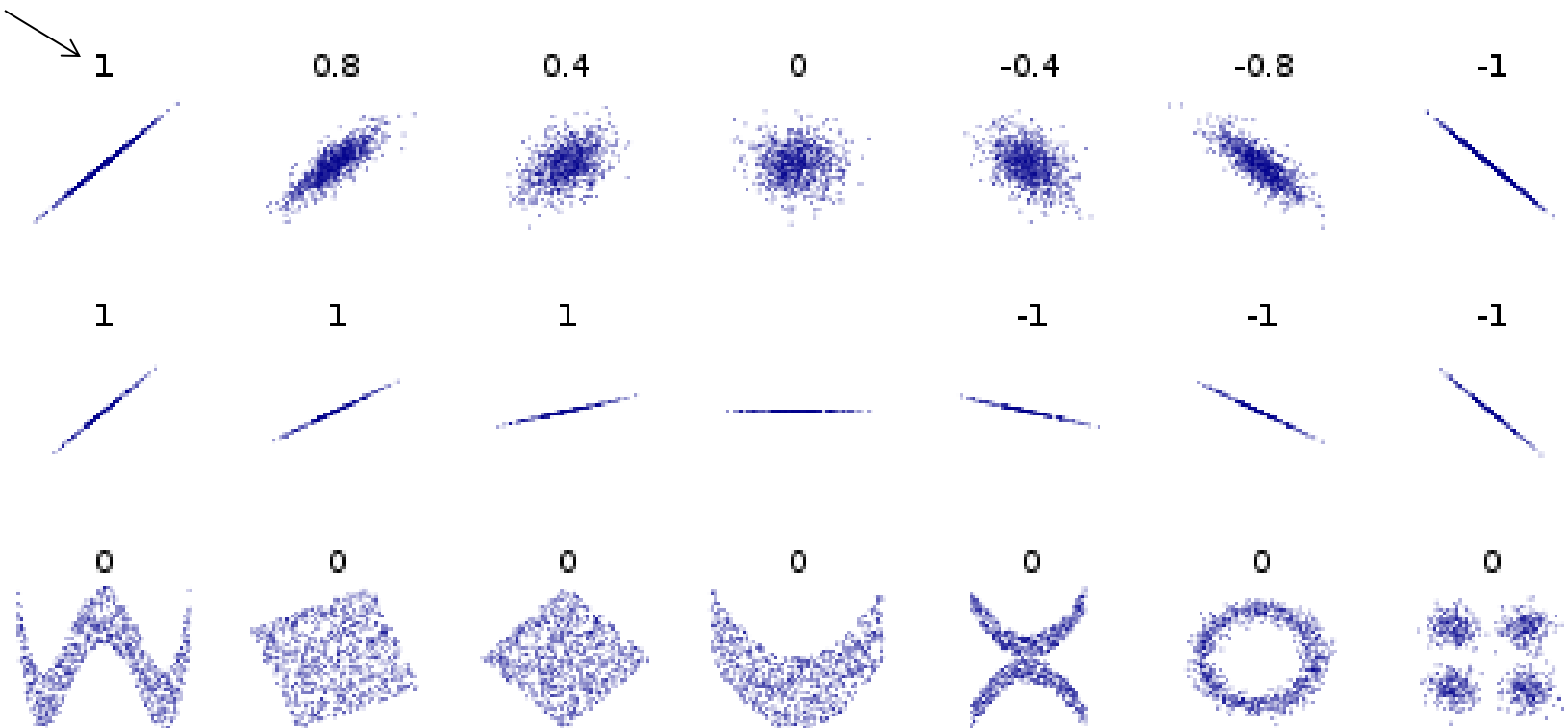
- Unable to capture complex relationships between genes
- Doesn't consider local network structure

Other Methods for Building Co-expression Networks:

- Mutual Information (MI): a measure of non-linear correlation
- ARACNe: prunes edges that are more likely to have a high MI due to indirect effects.

Co-expression may be non-linear

Pearson
correlation
value



Source: http://en.wikipedia.org/wiki/Correlation_and_dependence

Mutual Information (MI)

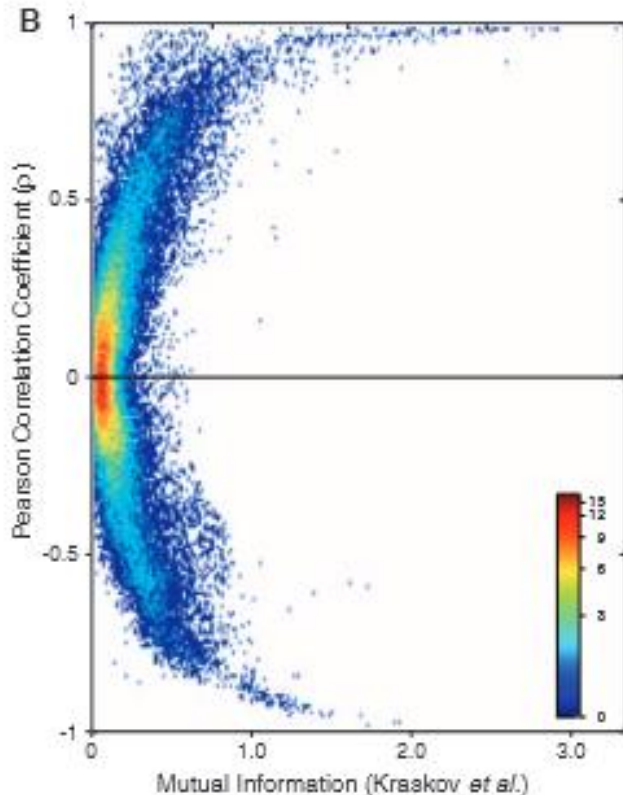


Definition $I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$

Properties $MI = I(X;Y) = H(X) - H(X|Y)$

- Measures how much knowing one of these variables reduces uncertainty about the other
- Positive and symmetric
- Invariant under nonlinear transformation

Understanding Nonlinear Relationships with Mutual Information



Some things to keep in mind:

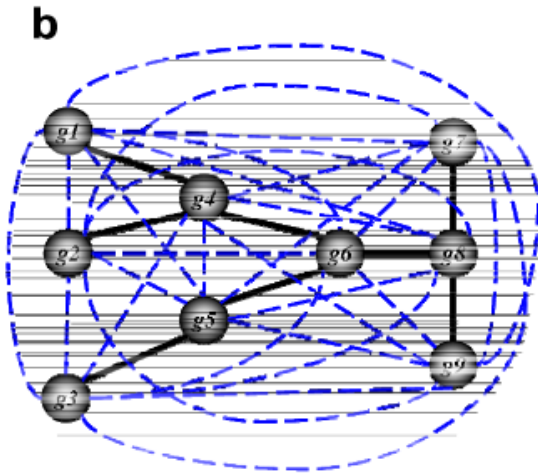
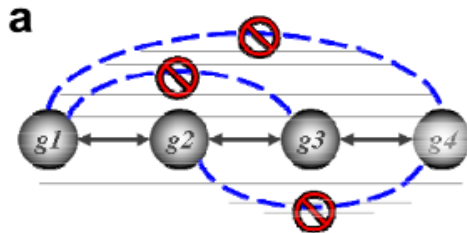
- Mutual Information only takes positive values.
- It has a very different overall distribution compared to Pearson correlation.
- It relies on estimating an underlying probability distribution.

ARACNe

(Algorithm for the Reconstruction of Accurate Cellular Networks)

Margolin, Califano et al. BMC Bioinformatics. 2005 Mar 20;7 Suppl 1:S7.

Key Idea: Remove indirect relationships.



- If g_1 and g_2 are correlated, and g_2 and g_3 are correlated, then g_1 and g_3 are also correlated, but this may not be meaningful.
- Uses the data processing inequality to remove these types of indirect relationships.
- Limitation: no “triangles” remain in the final network

How to relate modules to each other?



- **Question:** How does one summarize the metabolomic profiles in a module?

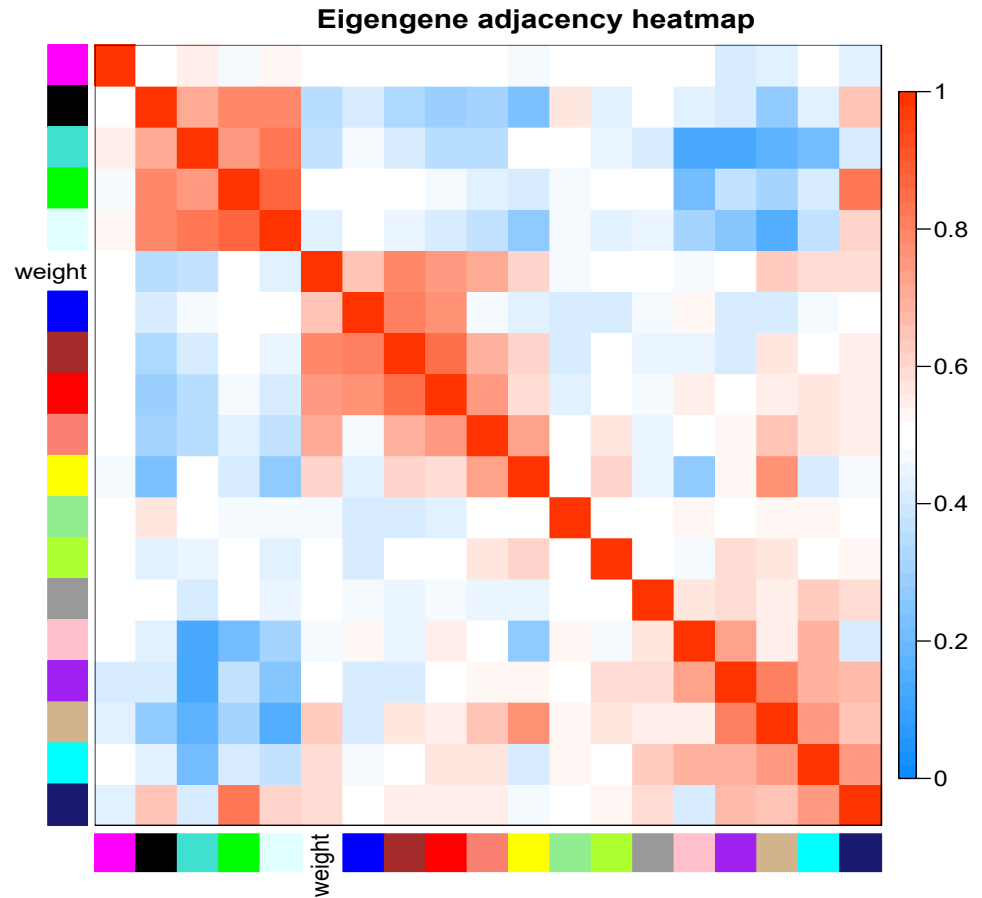
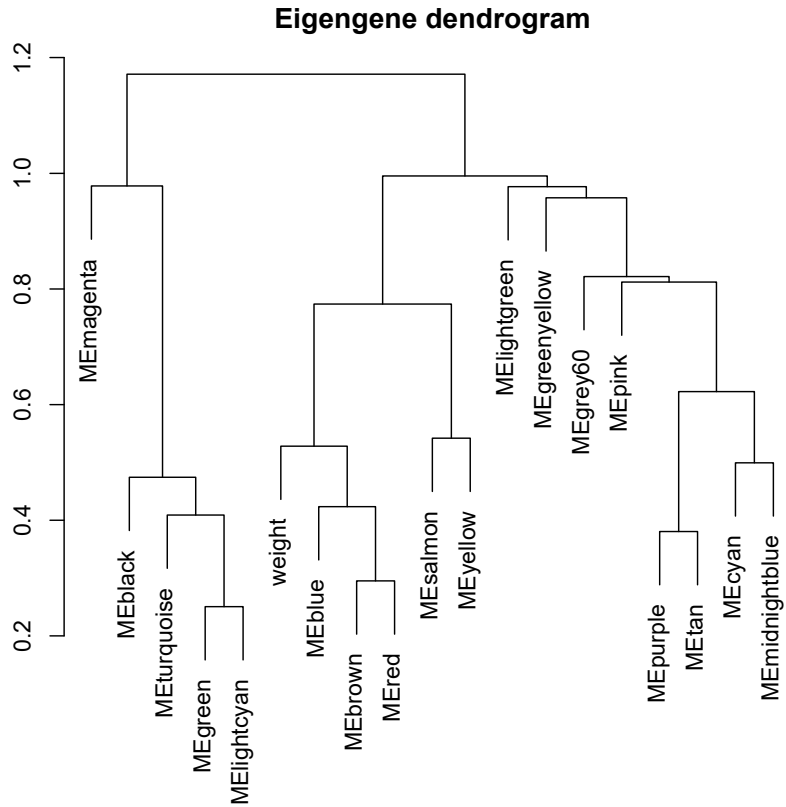
Answer: This has been solved.

Math answer: module eigengene = first principal component

Network answer: the most highly connected intramodular hub gene

Both turn out to be equivalent

Visualizing network of eigengenes

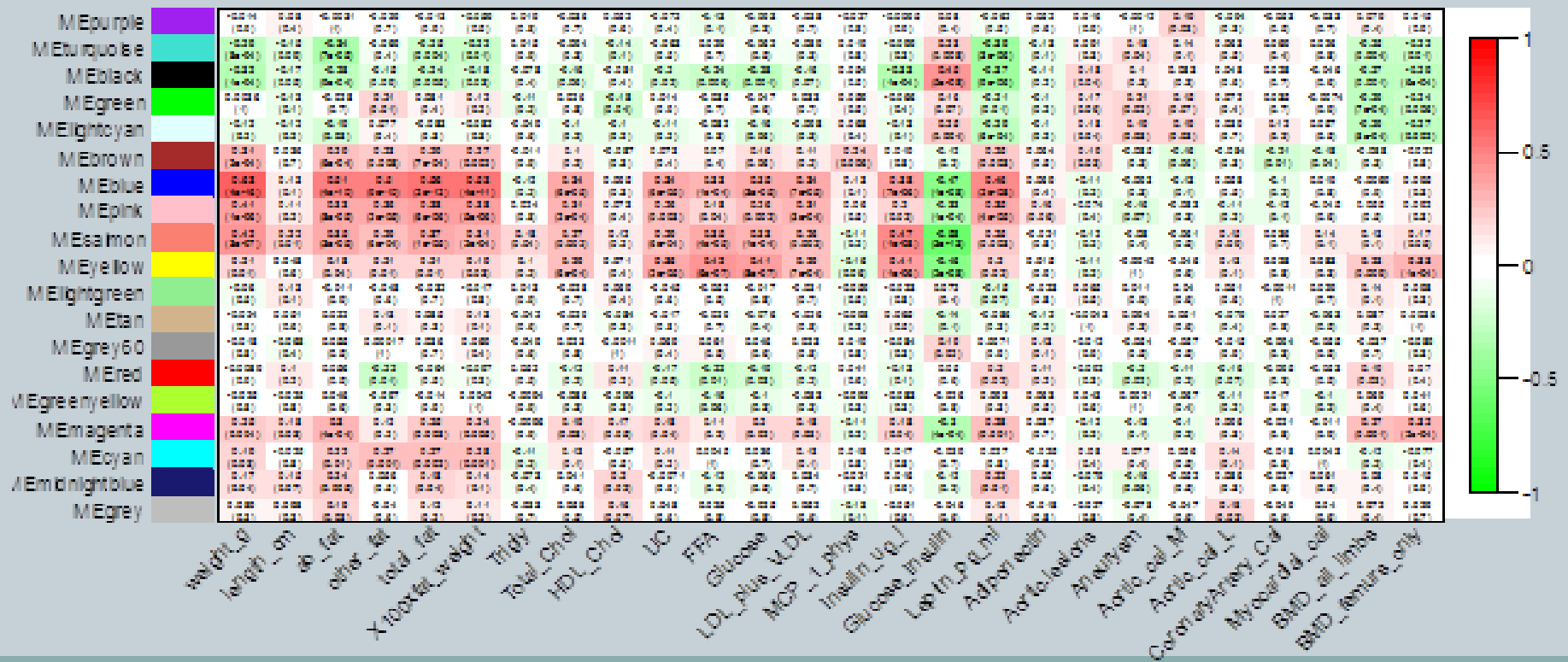


How to relate modules to external data?



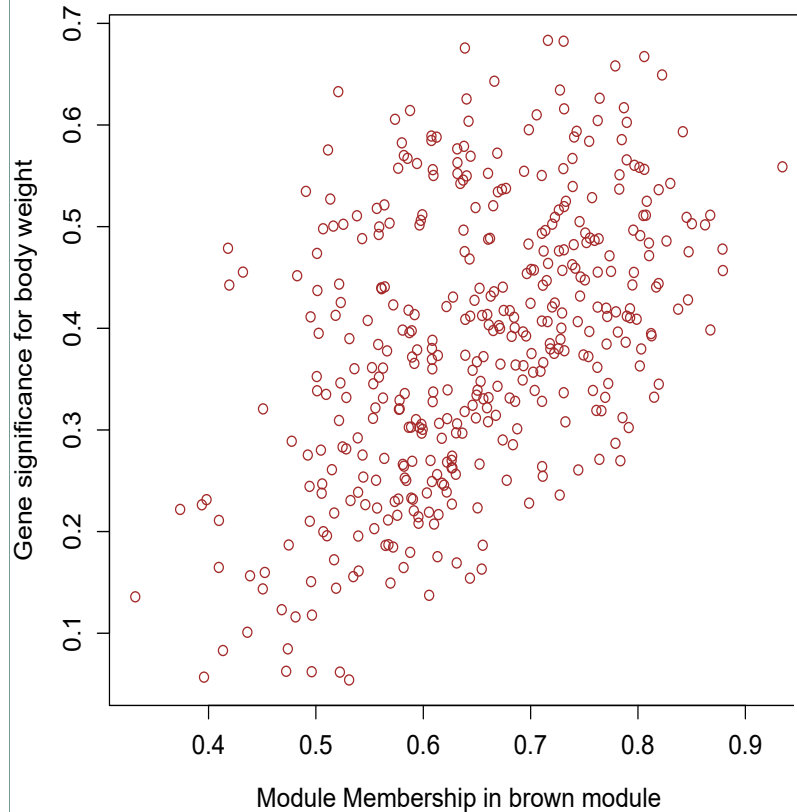
Table of module-trait correlations and p-values.

Each cell reports the correlation (and p-value) resulting from correlating module eigengenes (rows) to traits (columns). The table is color-coded by correlation according to the color legend.

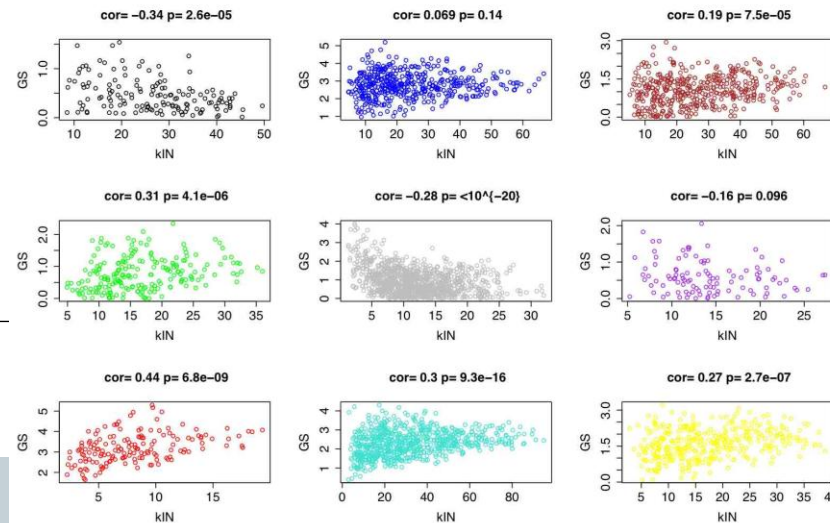


How to identify hub genes in a module

Module membership vs. gene significance
 $\text{cor}=0.48, p=5.8\text{e-}25$



- Note the relatively high correlation between gene significance and intramodular connectivity in some modules
- A combination of GS and intramodular connectivity is used to select important hub genes.



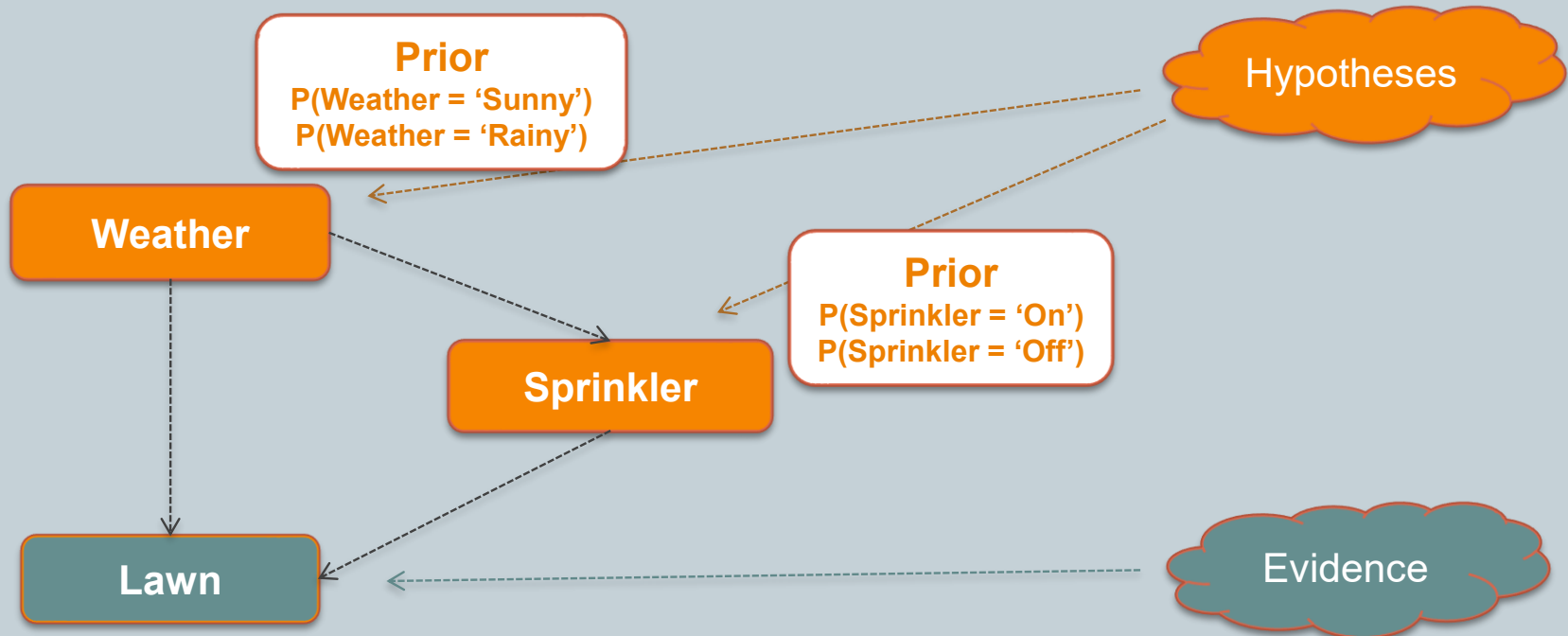
Bayesian Networks



- Both a Machine learning methodology and a Network methodology
- Predicts a disease outcome
- Process of building/searching for a network maximizes the probability of the dataset
 - Has incidental property of predicting an outcome (ie, asthma status, asthma exacerbations)
- Bayesian Priors
 - Bias toward expectation or seed network;
 - Or provide complexity penalty to enable parsimonious models

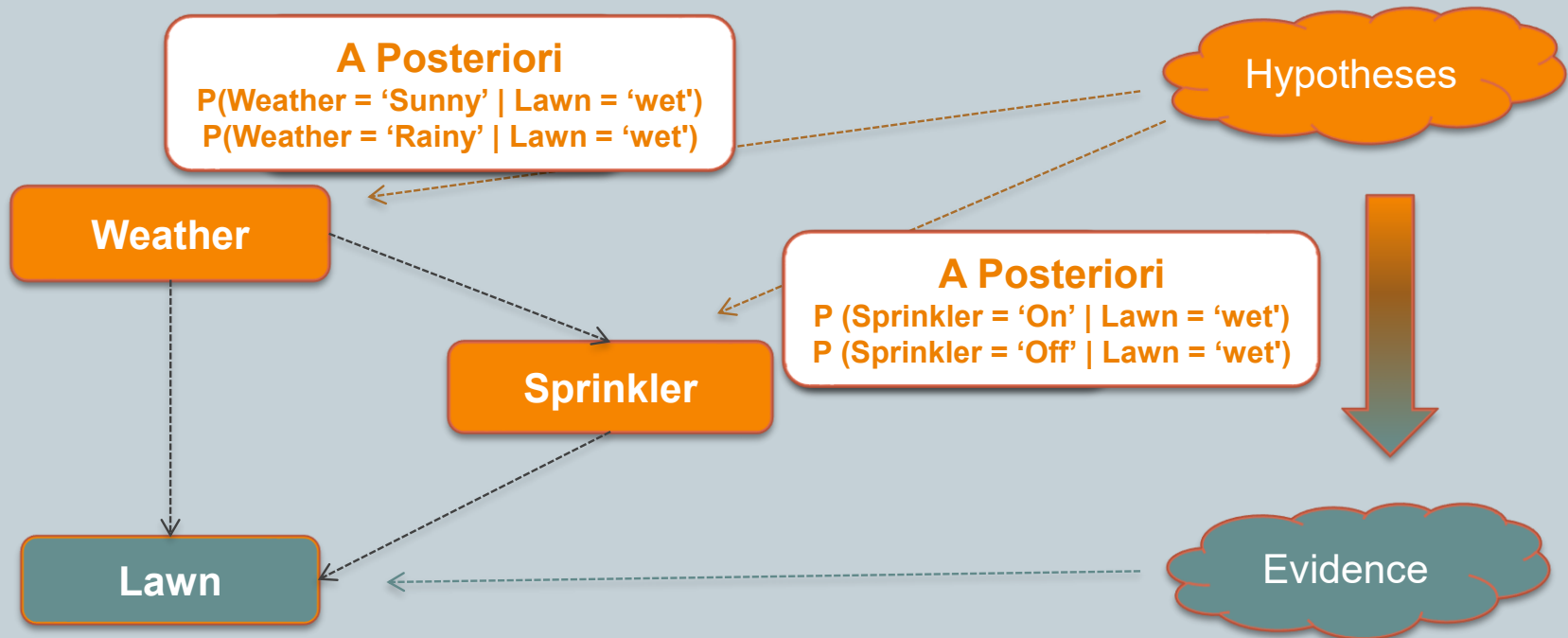
Bayesian Inference – Belief Propagation

- Given that the lawn is wet, what could be the reason for that?
 - Rainy weather? or
 - A turned-on sprinkler?



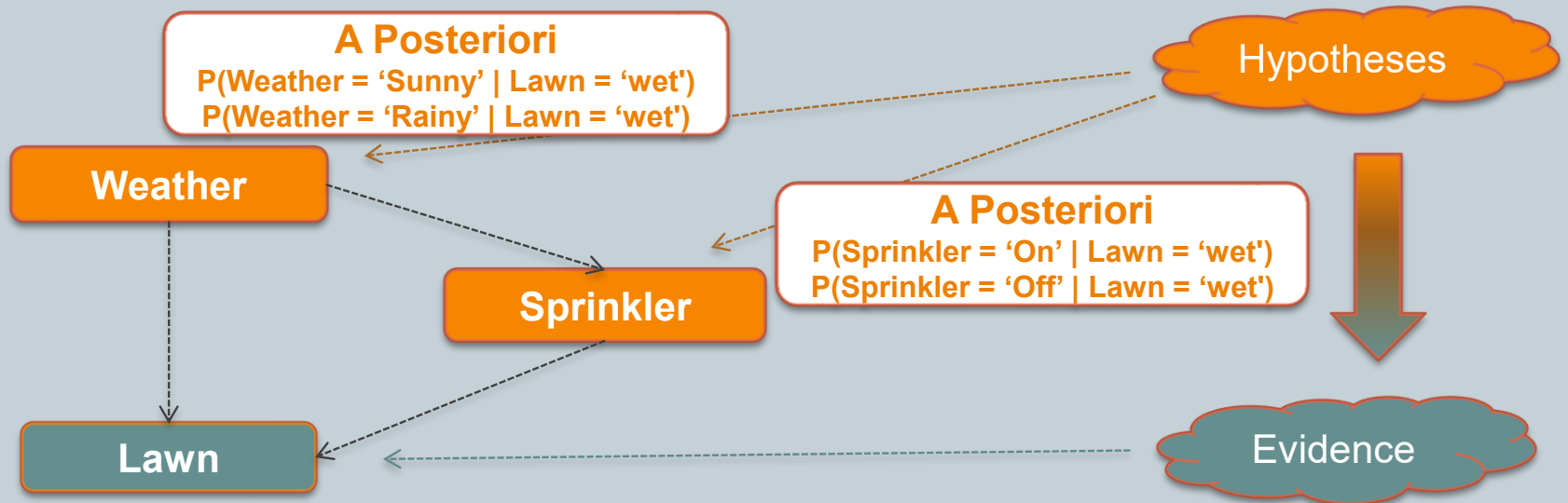
Bayesian Inference – Belief Propagation

- Given that the lawn is wet, what could be the reason for that?
 - Rainy weather? or
 - A turned-on sprinkler?



MAP = Bayes Decision Rule

- So what to predict? Rainy weather or turned-on sprinkler?
- MAP: choose Maximum A posteriori Probability
 - For $P(\text{Weather} = \text{'rainy'} \mid \text{Lawn} = \text{'wet'}) = 0.1$; $P(\text{Sprinkler} = \text{'On'} \mid \text{Lawn} = \text{'wet'}) = 0.08$
 - ✦ Choose **Weather = 'rainy'**, i.e. given the lawn is wet it's more probable that a rainy weather caused it rather than a turned-on sprinkler



Differences vs Other Network methods



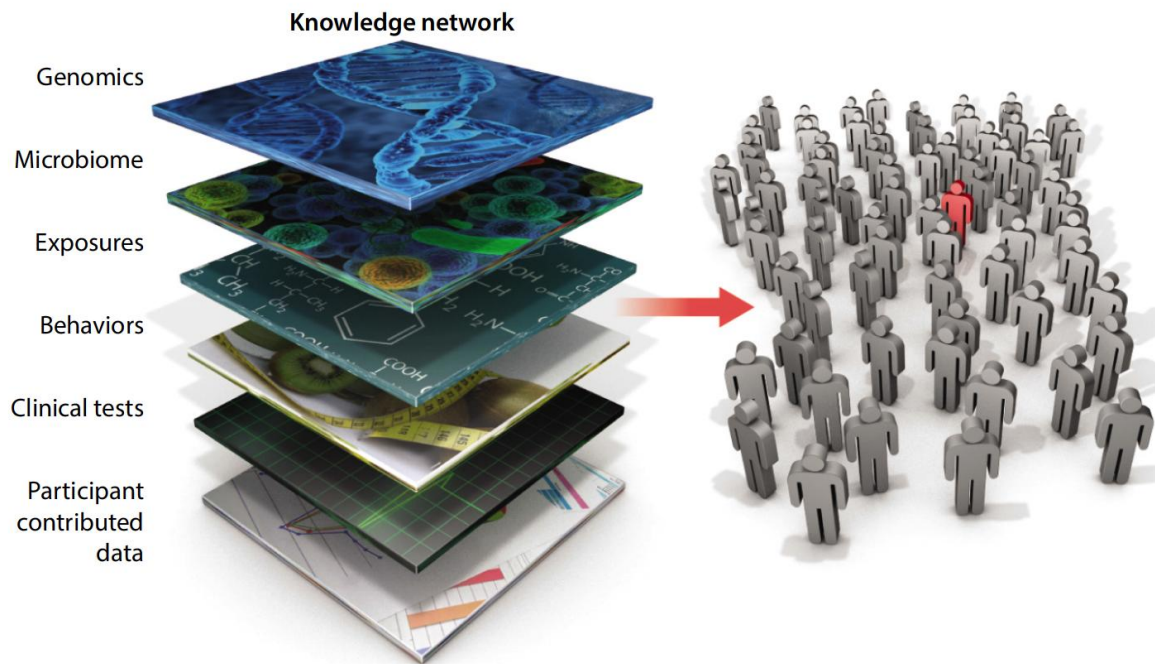
- Nonlinear interactions
- Predict ALL variables simultaneously
- Robust to missing data
- Robust to correlated predictors
- Network Search
 - Variety of search algorithms
 - find and identify the most important (nonlinear) predictors (not just main effects) in an unbiased way
- Prediction
 - The network actually *does* something
- How good is the network? Comparable to other networks?
 - Network quality score
 - posterior likelihood of data given network structure
- Parsimonious models
 - Doesn't require ad hoc p-value or thresholding of edges

When to and not to use BNs



- **Small Sample size**
 - Replication requires:
 - ✦ Large sample size or
 - ✦ Large effect sizes (eg, metabolomics)
- **Many thousands of variables**
 - Computationally slow
 - Prone to overfitting
- **Ideally suited for metabolomics**
 - Number of metabolites is ideal
 - Knowledge about Bayesian Prior

Networks and Precision Medicine



Population Medicine



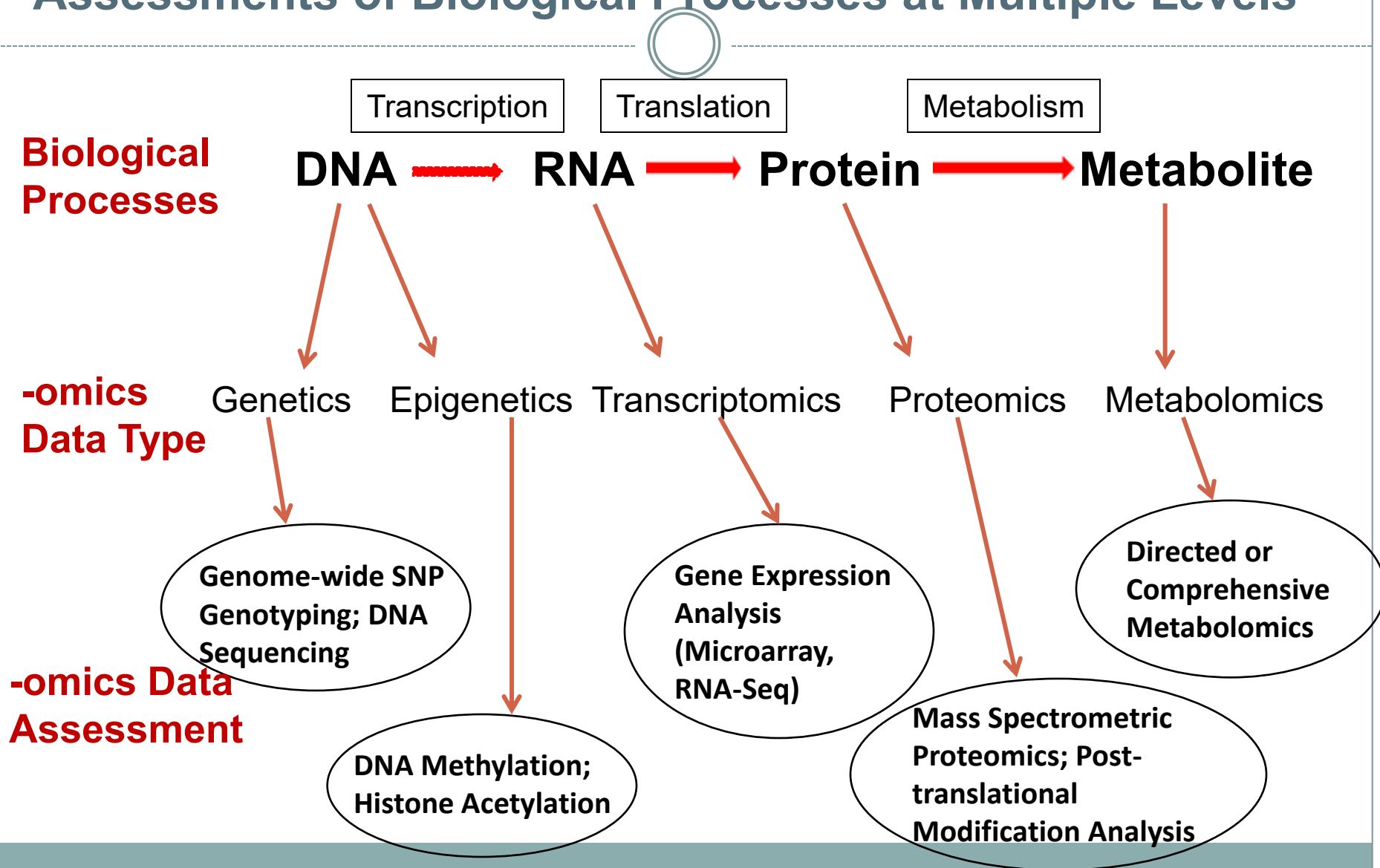
Precision Medicine



Personalized Medicine



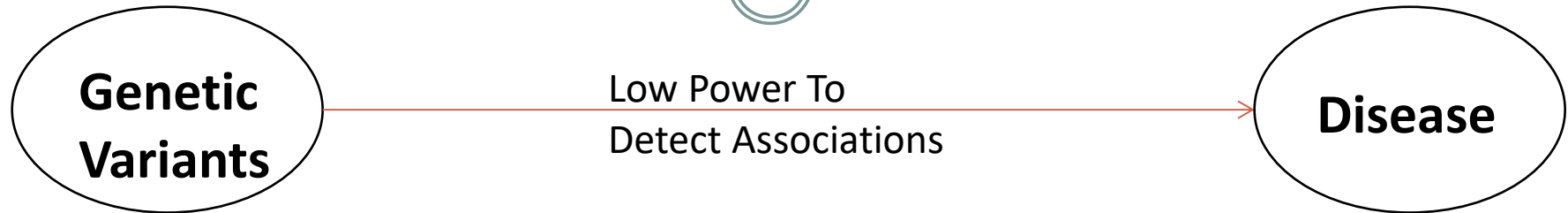
Technological Advances Have Enabled Large-Scale Assessments of Biological Processes at Multiple Levels



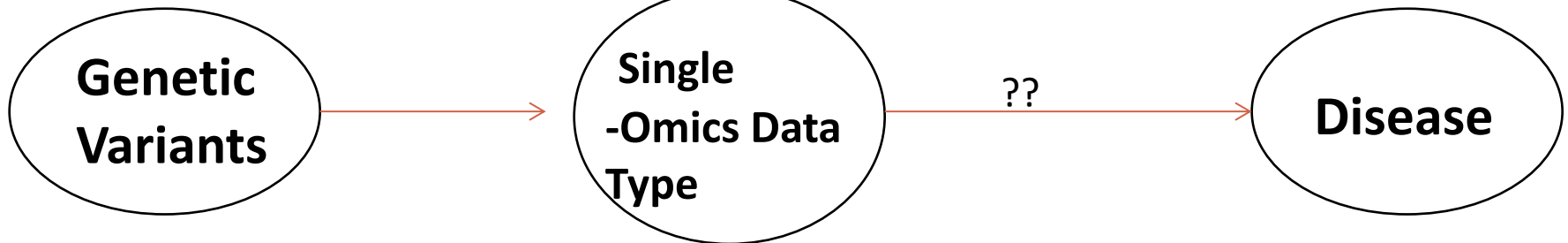
Evolution of Complex Disease Genetic Studies

(Silverman/Loscalzo, Discovery Med 2012; 14: 143)

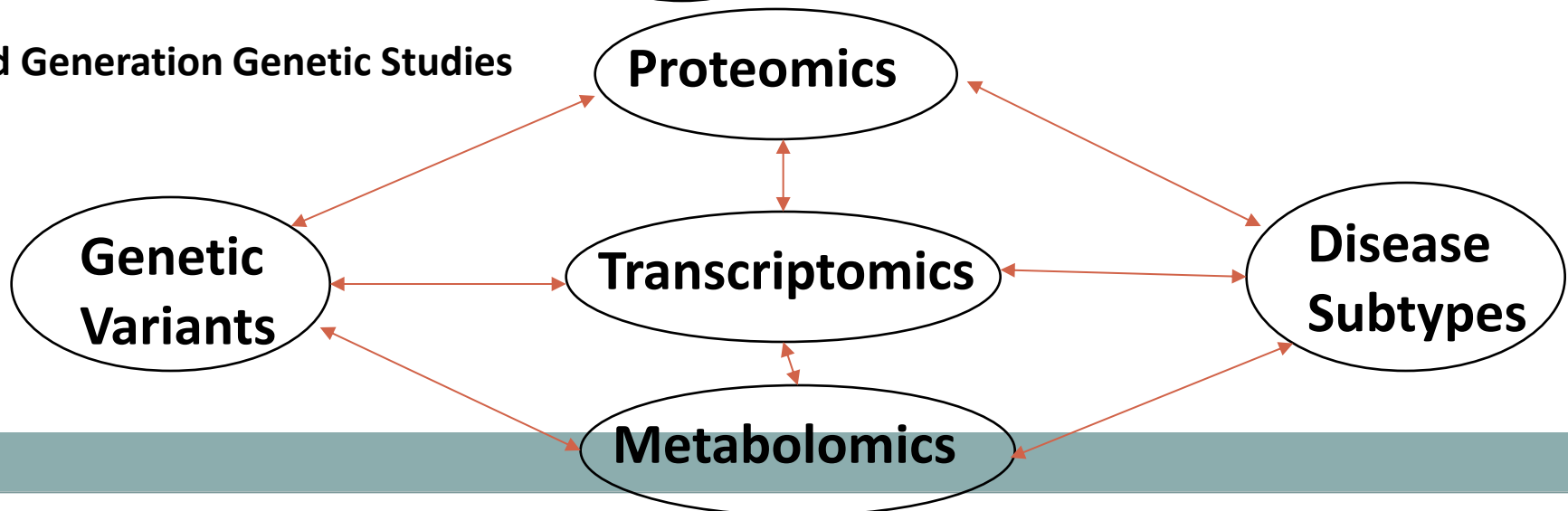
First Generation Genetic Studies



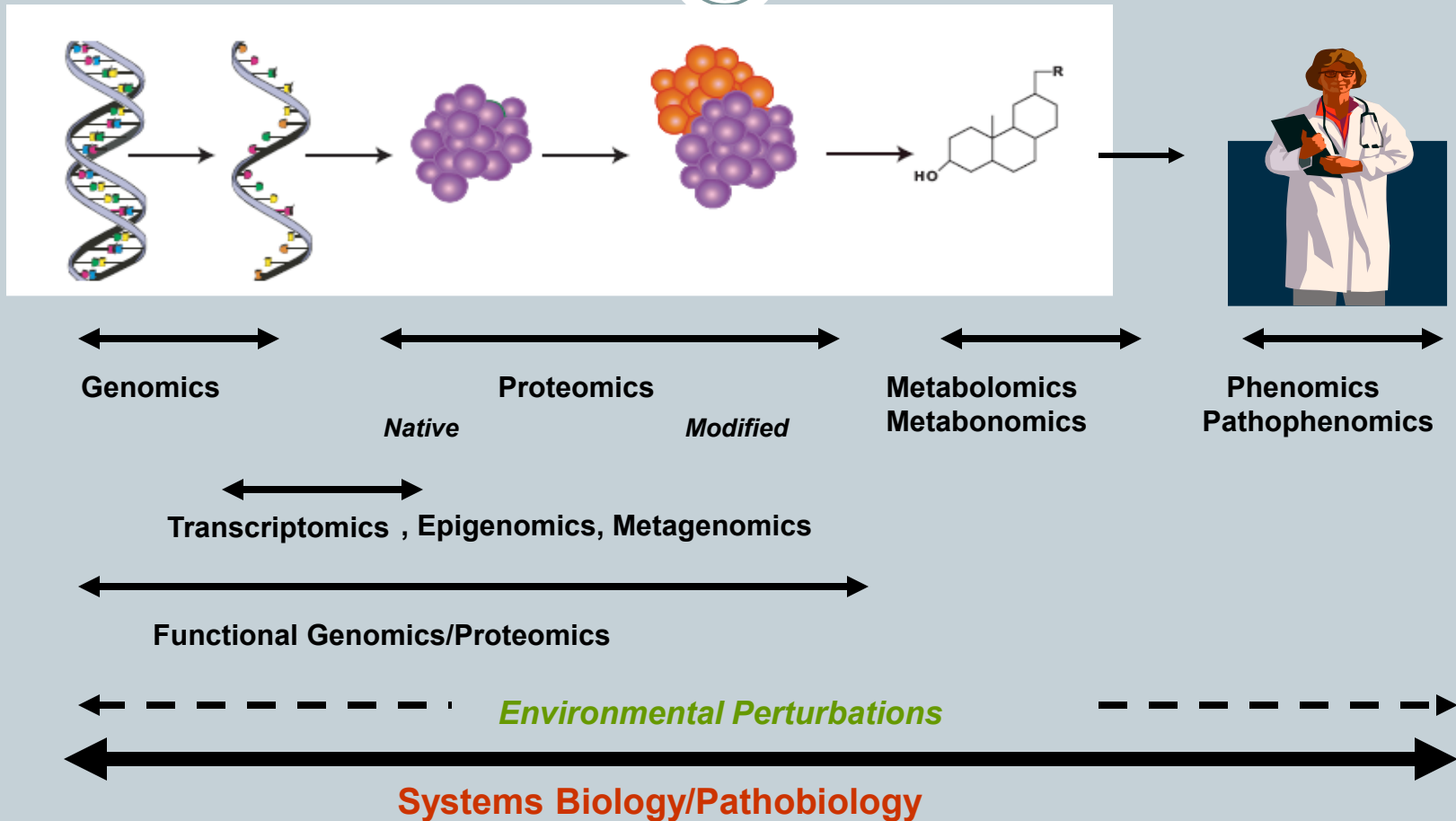
Second Generation Genetic Studies



Third Generation Genetic Studies



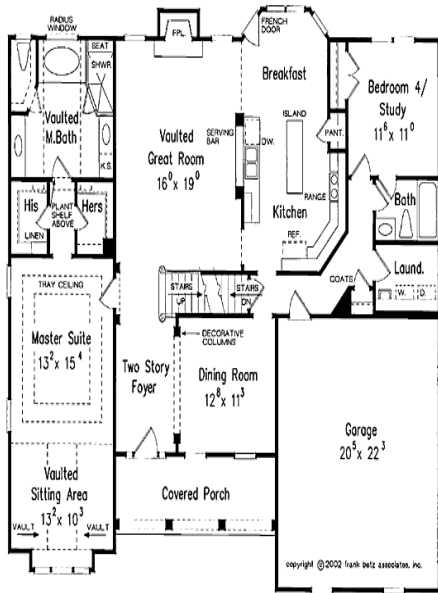
Universe of Precision Medicine



Genome-Proteome-Metabolome Relationship

Genome *implies* phenotype.

Metabolome *is* phenotype.



Genome



Proteome

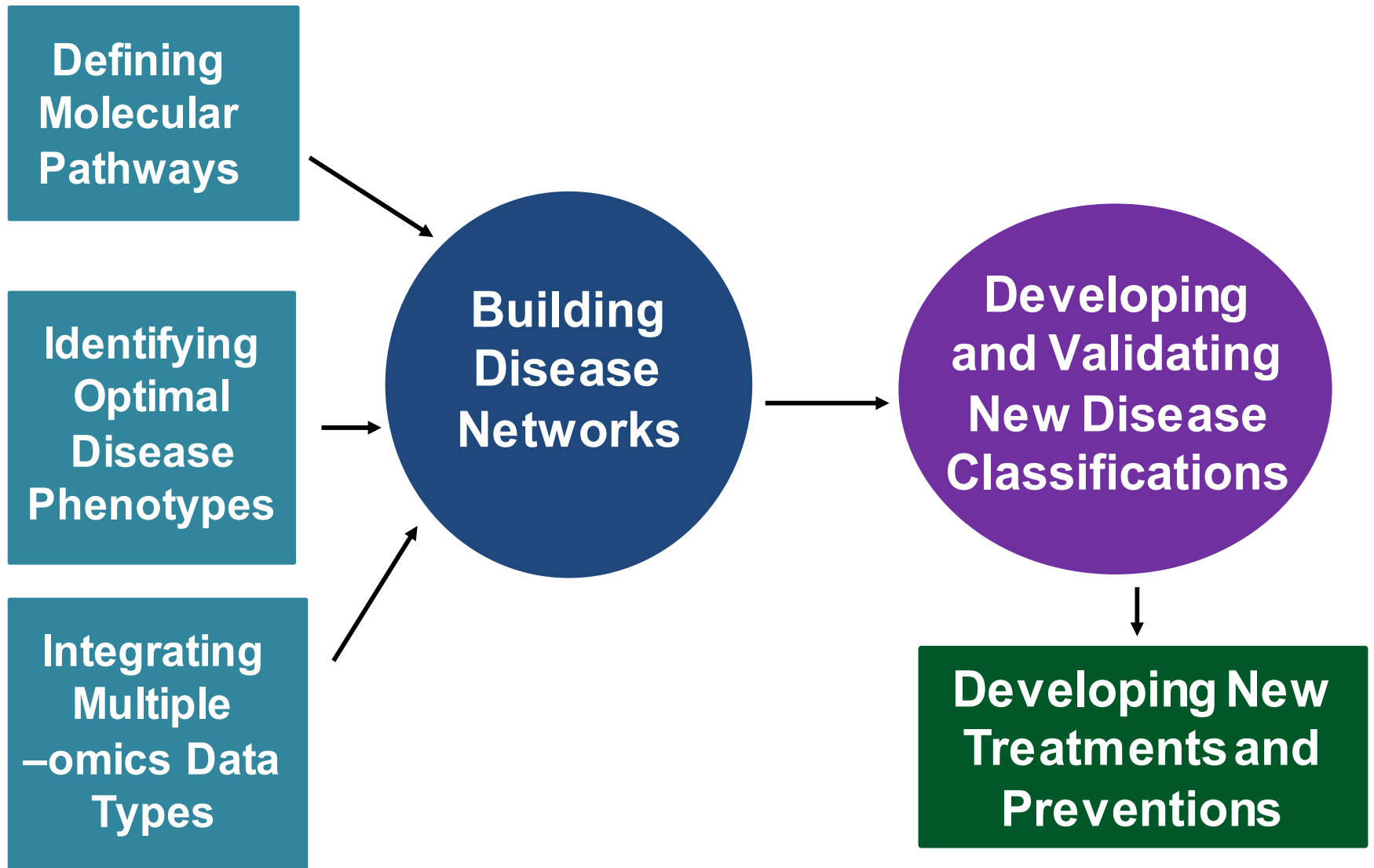


Metabolome



Environmental Determinants

Approaches to Complex Diseases in Network Medicine



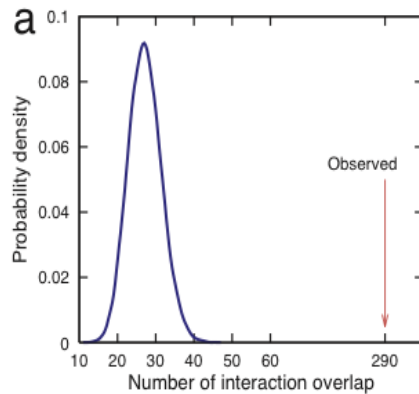
Challenges to Conventional Disease Classification

- Many distinct diseases have common phenotypes
- Common treatments, disease co-occurrence, and GWAS suggest common underlying mechanisms for many 'distinct' diseases.

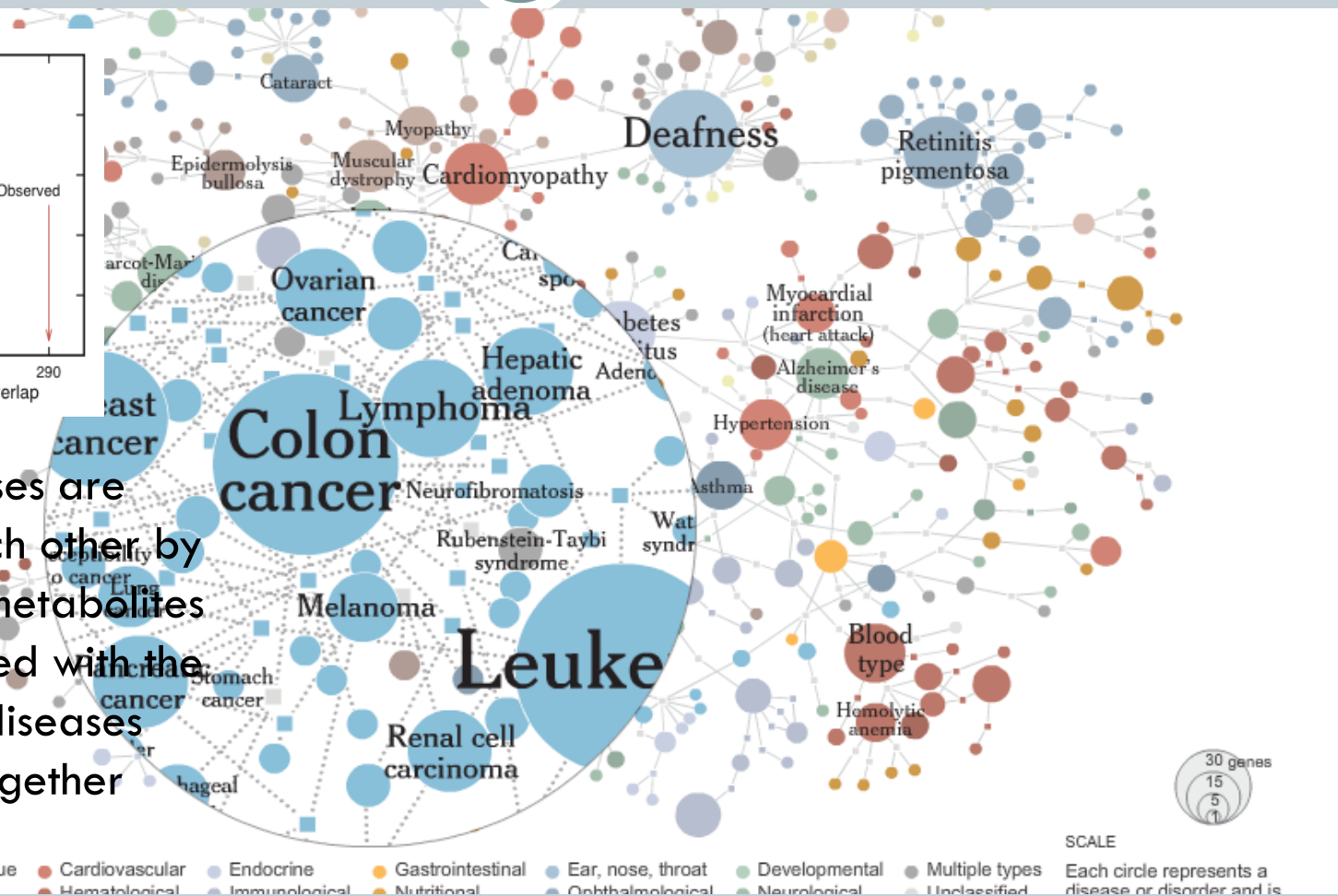
“A new disease taxonomy should be developed that would:

- describe and define diseases based on their intrinsic biology in addition to traditional physical ‘signs and symptoms’;
- go beyond description and be directly linked to a deeper understanding of disease mechanisms, pathogenesis, and treatments; and
- be highly dynamic...continuously incorporating newly emerging...information.”

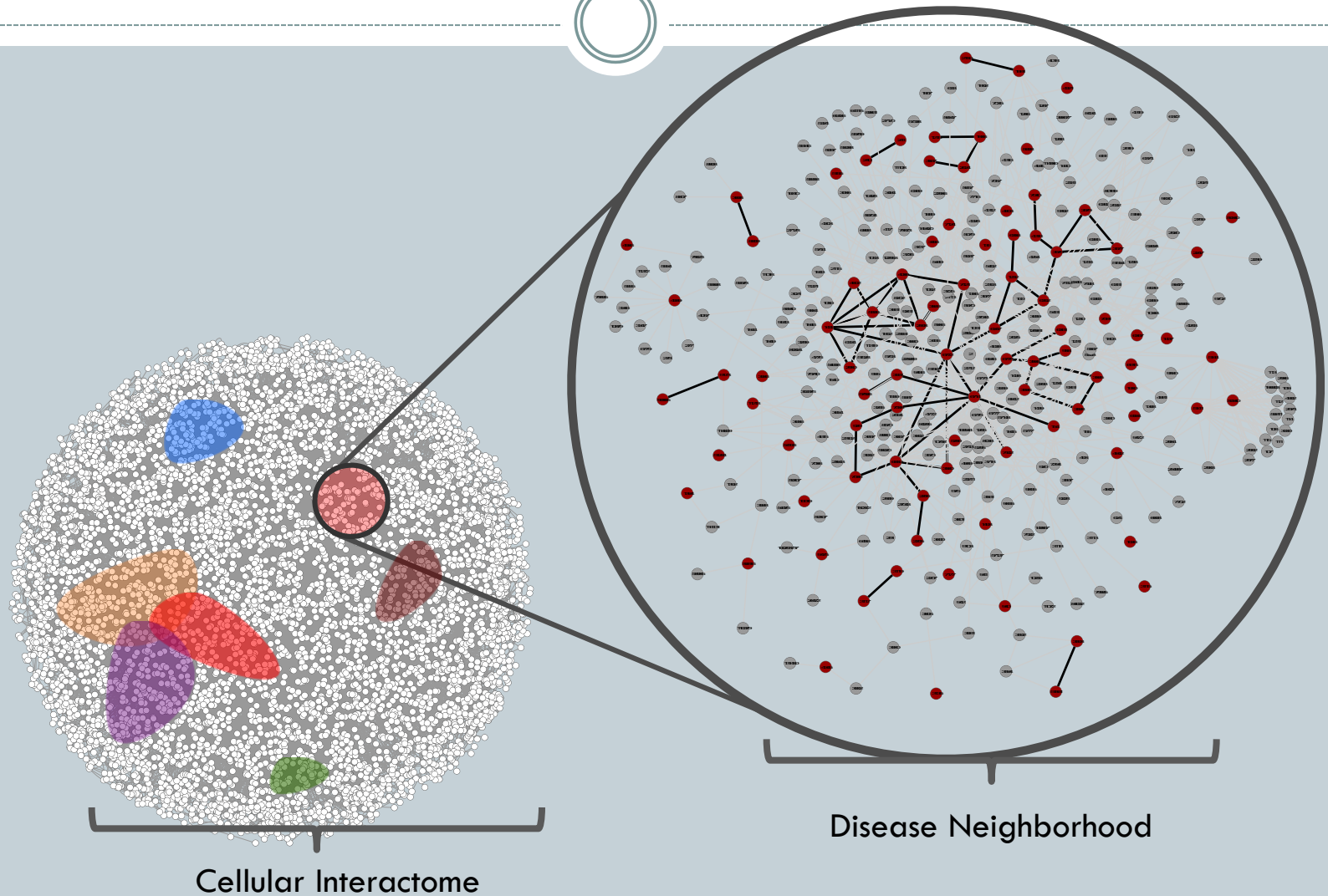
Human Disease Network



- Different diseases are connected to each other by common genes/metabolites
- Genes associated with the same or similar diseases tend to cluster together



Disease module

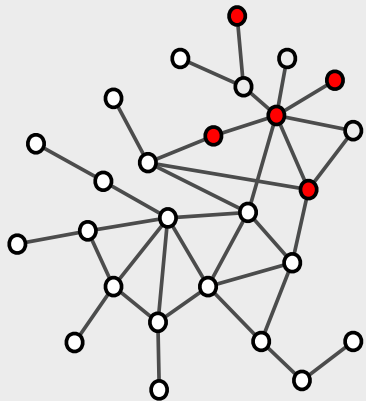


Disease module identification approach

DIAMOnD – Disease Module Detection algorithm

I Interactome construction

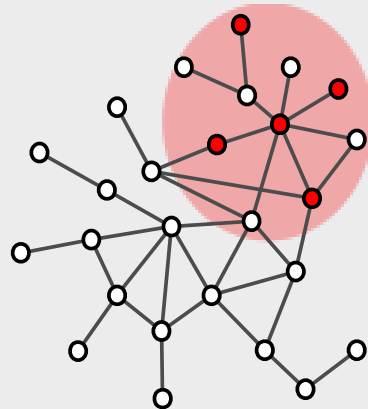
Binary interactions, metabolic couplings,
regulatory interactions ...



Seed gene selection

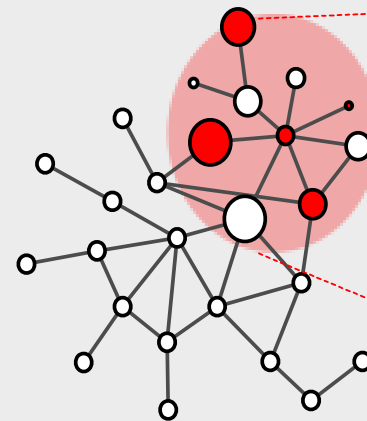
OMIM, GWAS, literature

II Disease Module Identification



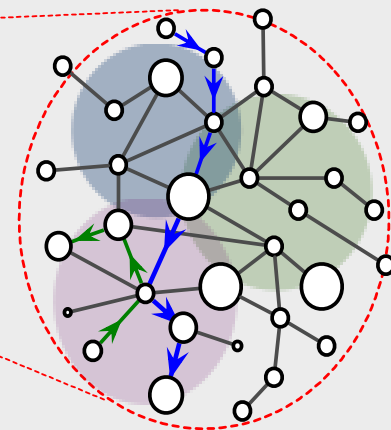
DIAMOnD: Disease Module
Detection Algorithm

III Validation



Gene expression data
Gene Ontologies
Pathways
Comorbidity

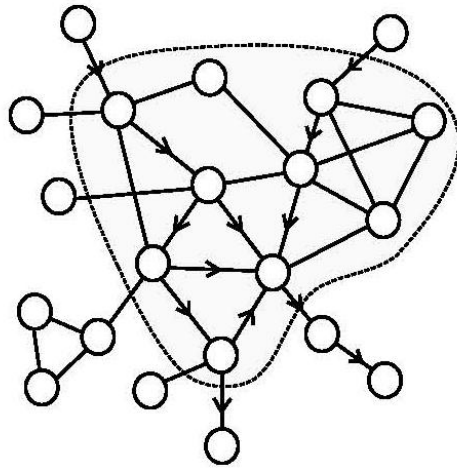
IV Biological interpretation



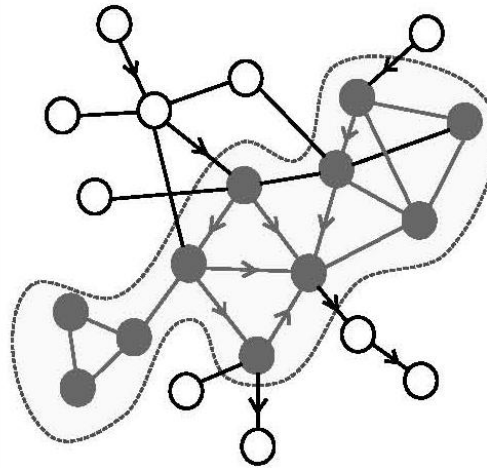
Pathway prioritization
Disease heterogeneity
Drug target identification

Disease Modules and Therapeutics

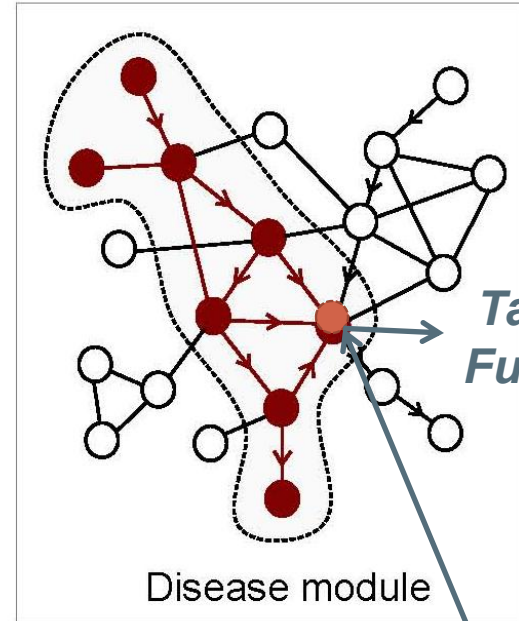
Drug targets are typically characterized in isolation from the disease module.



Topological module



Functional module

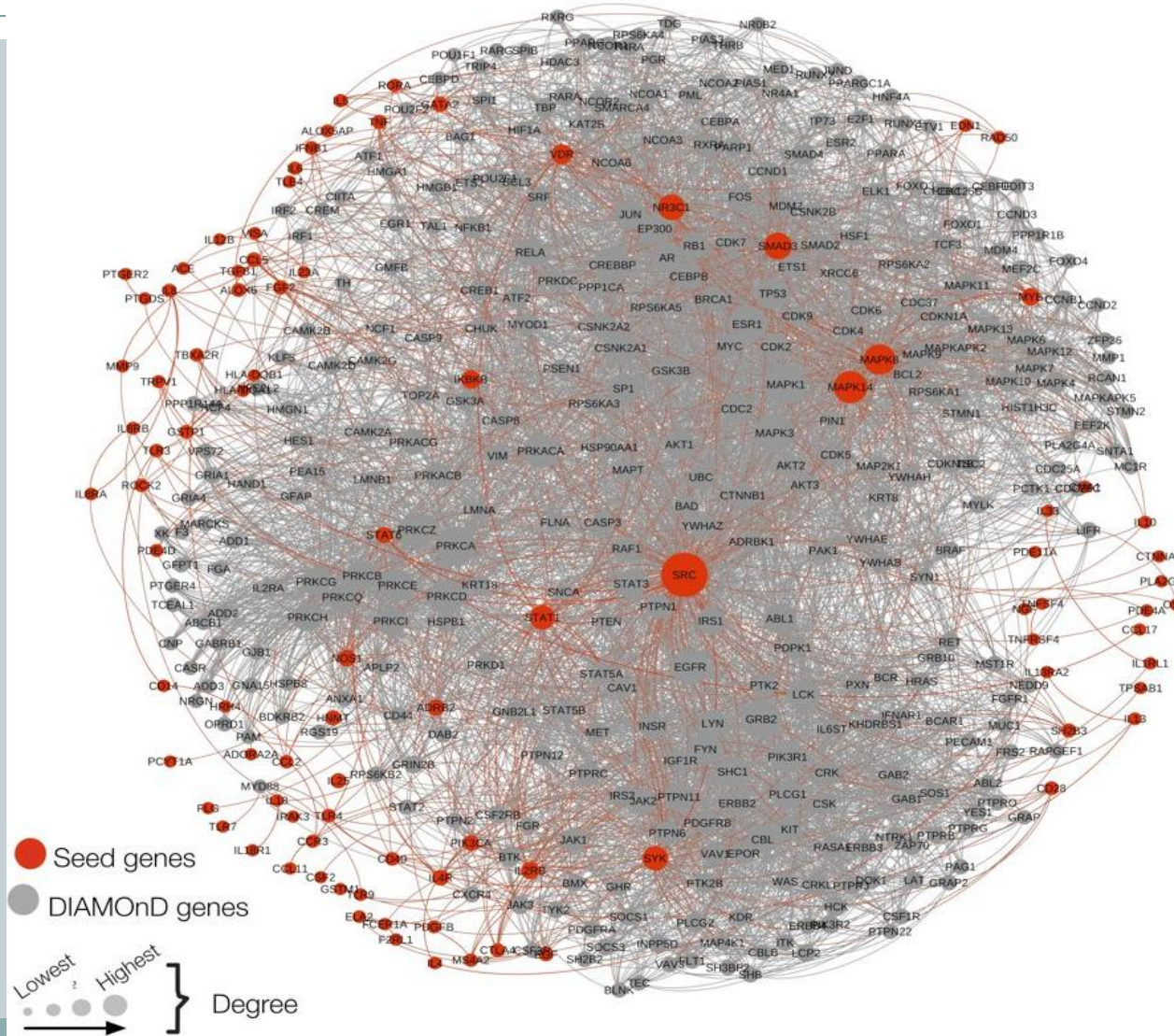


Disease module

- Topologically close genes (or products)
- Functionally similar genes (or products)
- Disease genes (or products)
- Undirectional interactions
- Directional interactions

Drug Target

Asthma Module

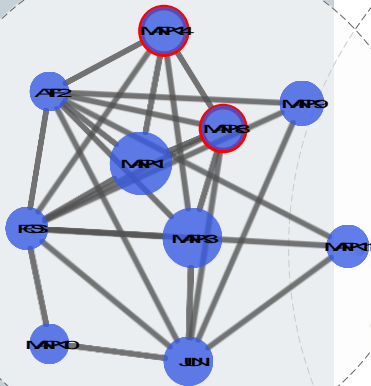
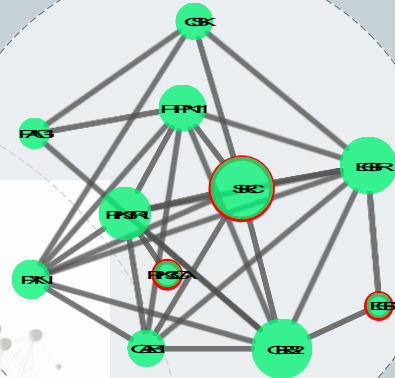


Prioritization of pathways

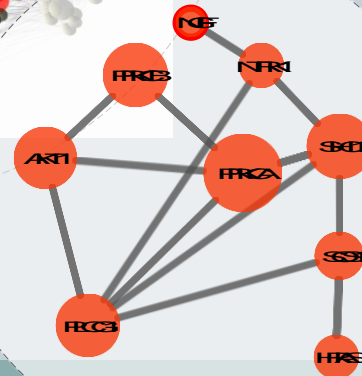
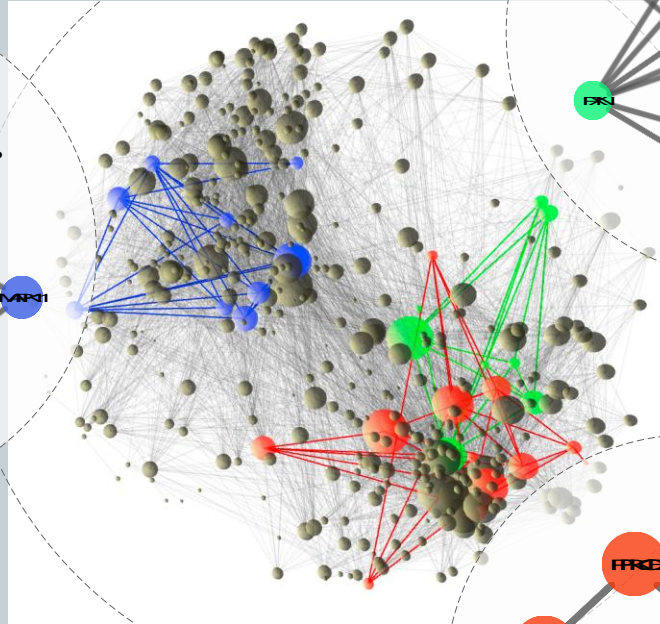
Fully embedded pathways



GAB1 Signosome pathway



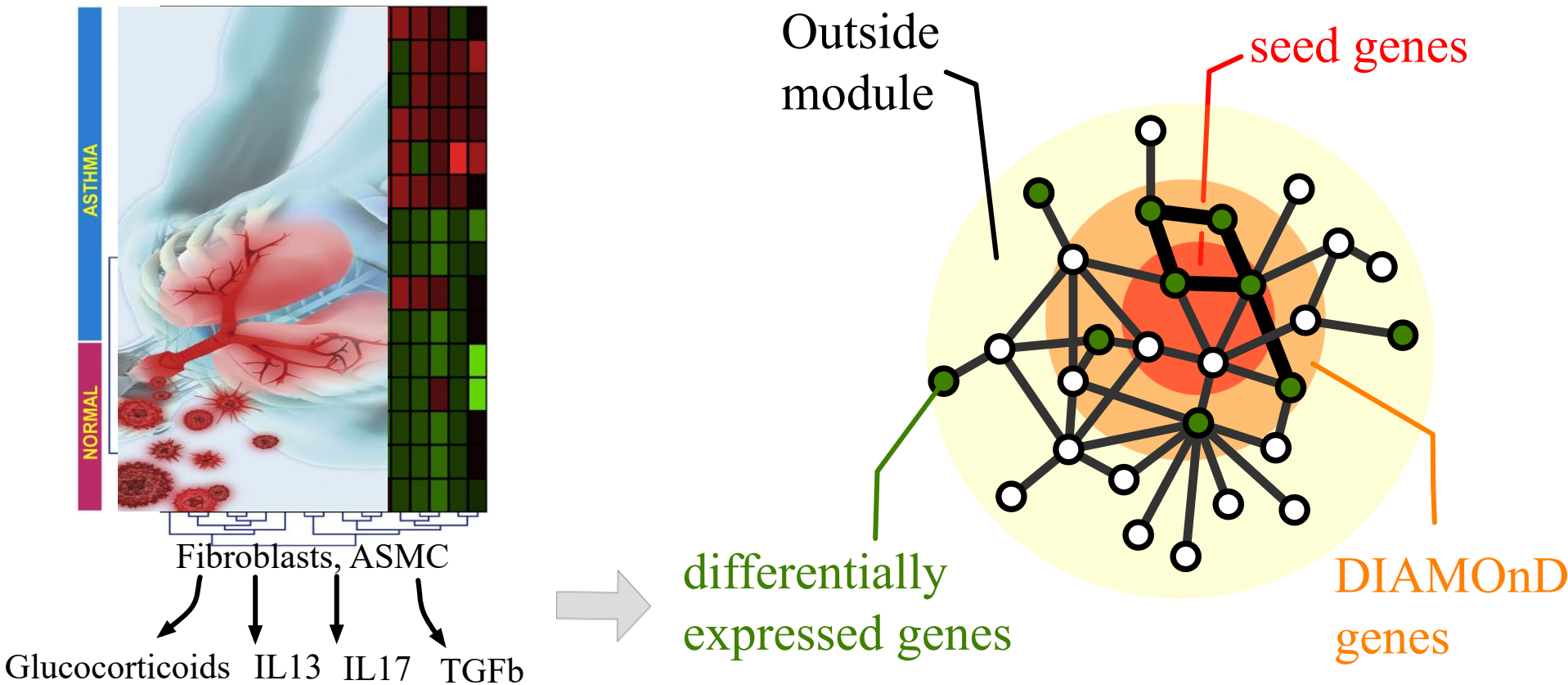
AP-1 Activation pathway



TRKA pathway

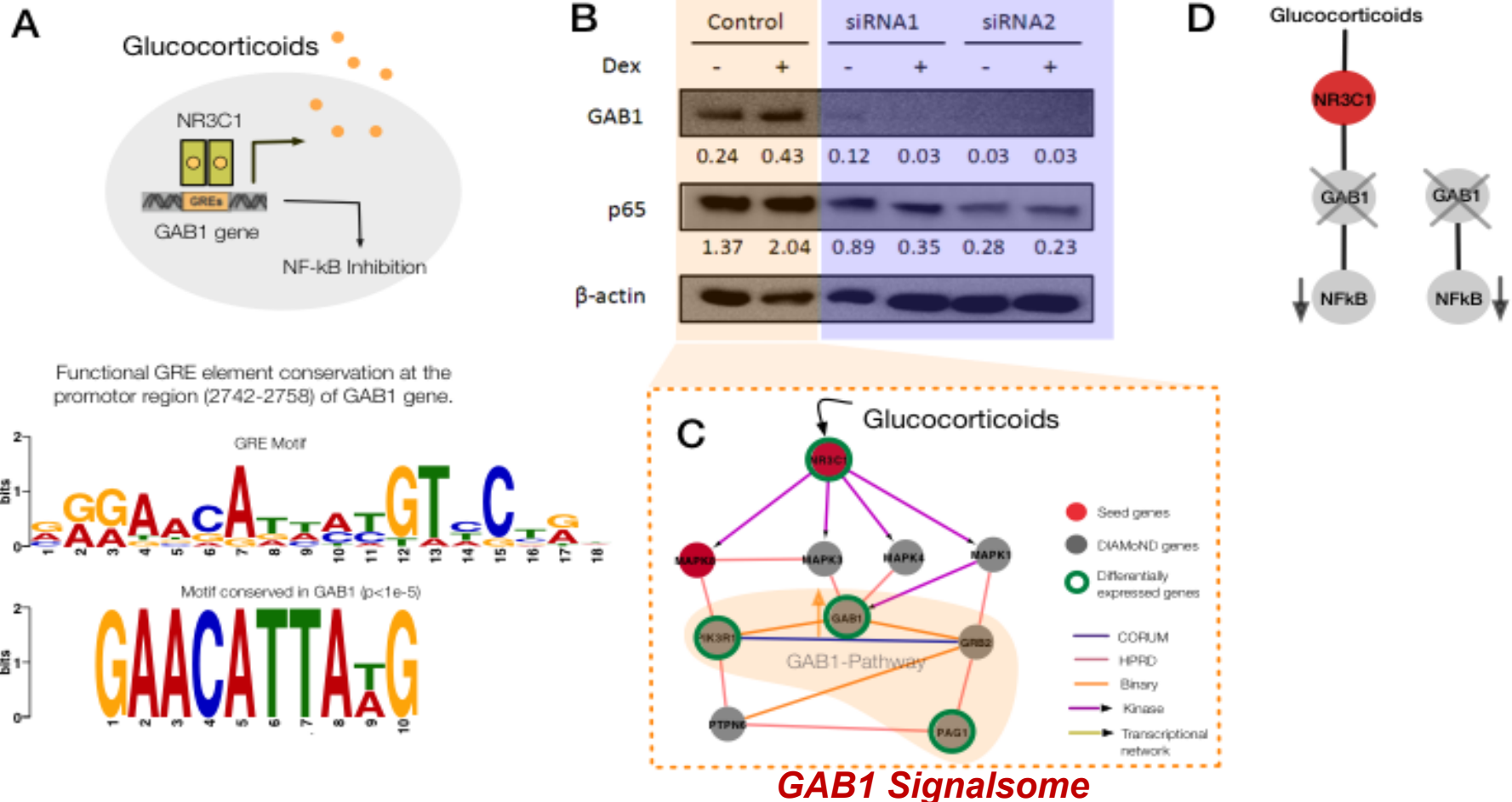
Differential gene expression from drug treatment and the disease module

Gene-expression profile



Novel Regulatory Pathway in Asthma: GAB1

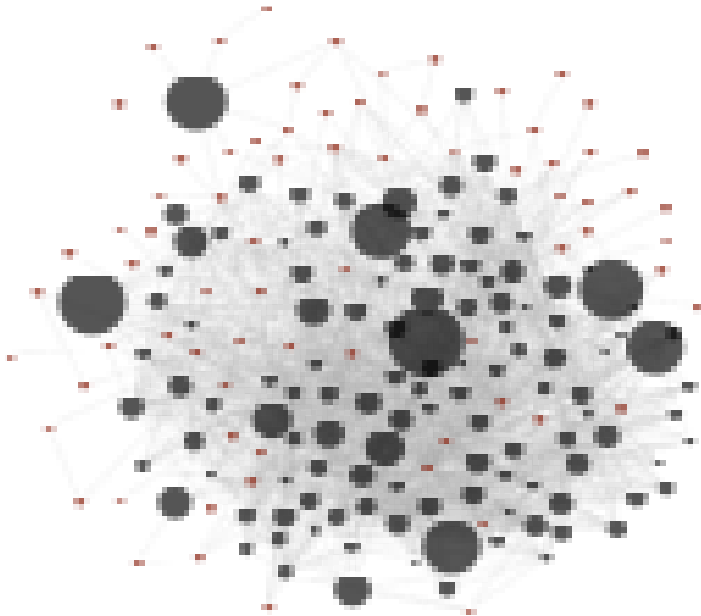
Novel inflammatory pathway in asthma derived from disease module



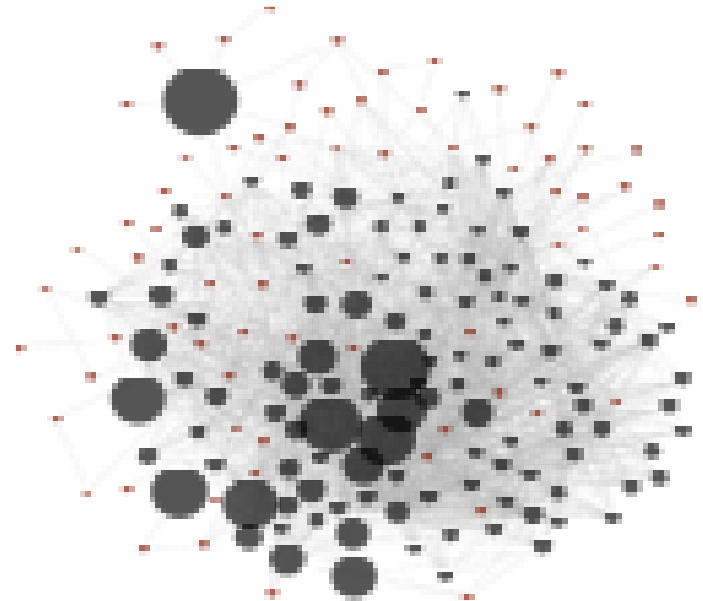
Prioritization of pathways

Evidence from multi-omics

gene expression



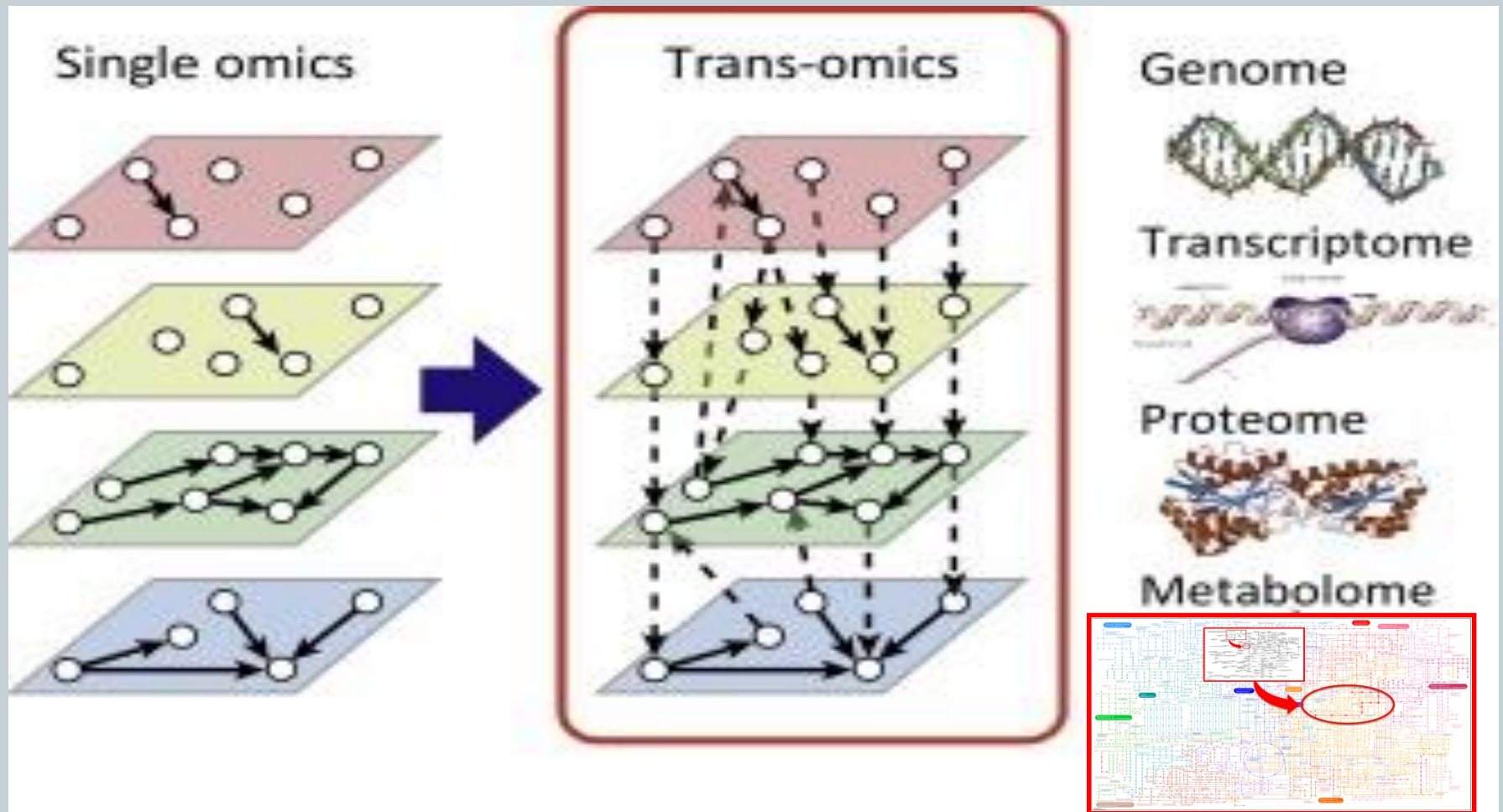
Metabolomic data



+

GSE3183 (human airway hyperresponsiveness)
GSE473 (Human CD4+ lymphocytes)
GSE470 (Human asthma exacerbatory factors)GSE3004
(Human airway epithelial)
GSE2125 (alveolar macrophages; $p < 0.02$) and human
primary cell lines of NHBE, NHLF and BSMC, exposed to
cytokines or PBS ($p < 0.05$)

The Future: Multidimensional Networks



Adapted from: Yugi et al. *Trends in Biotechnology* (2016)

Summary of Network Medicine



- Biological Networks
- Network Terminology & Characteristics
 - Edges, Nodes, Hubs
 - Connectivity, Betweenness
 - Scale Free Networks
- General approaches for Network creation
 - Weighted Correlation Gene Network Analysis
 - Bayesian Networks
- Precision Medicine and Networks
 - Redefine phenotypes
 - Novel disease treatments
 - DIAMoND