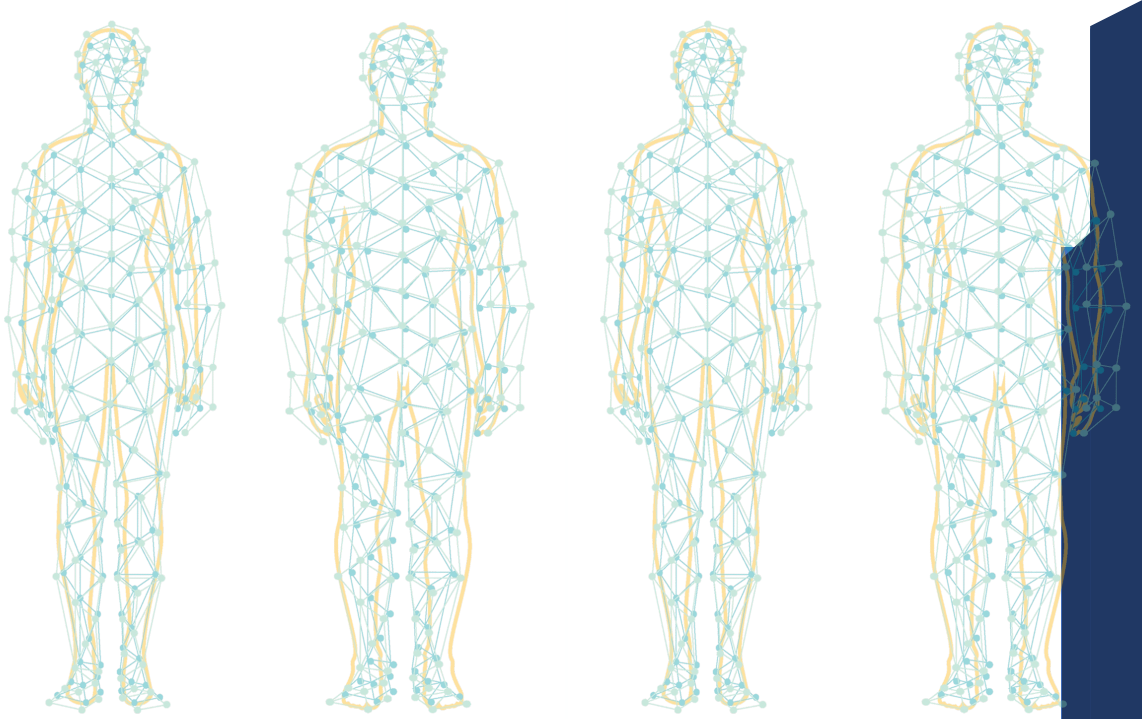


Computational methods for data integration

Xin Hu, PhD

Assistant Professor of Medicine
Division of Pulmonary, Allergy, Critical Care and Sleep Medicine
Emory University

Outline

Overview of systems biology and common methods for omics studies

Design of integrative omics study and major approaches:
1. Pathway-based approach
2. Network-based approach

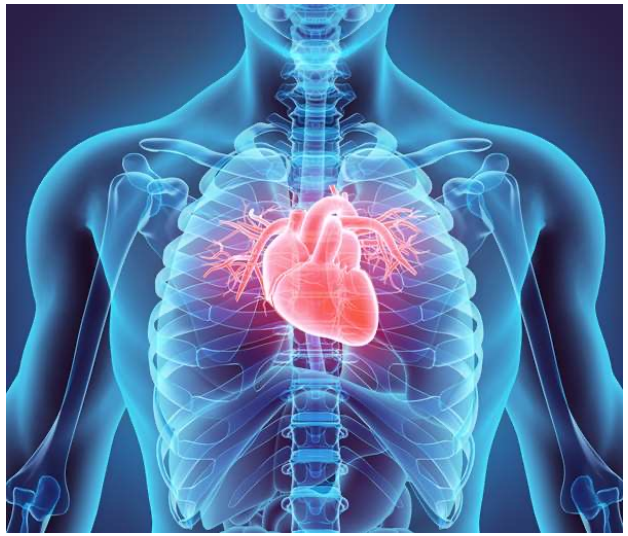
xMWAS

Other advanced machine learning tools

Introduction: A Systems Biology Framework

Systems biology studies biological systems, not isolated components but their interactions as well as emergent behavior.
Approach: Holistic as opposed to subsystems
Characteristics: Computational and mathematical modeling; can provide quantitative analysis

Beyond reductionism – systems biology gets dynamic

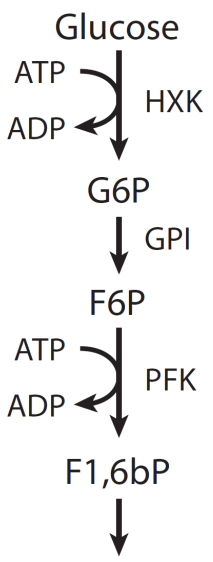


http://www.wiringthebrain.com/2019/09/beyond-reductionism-systems-biology.html

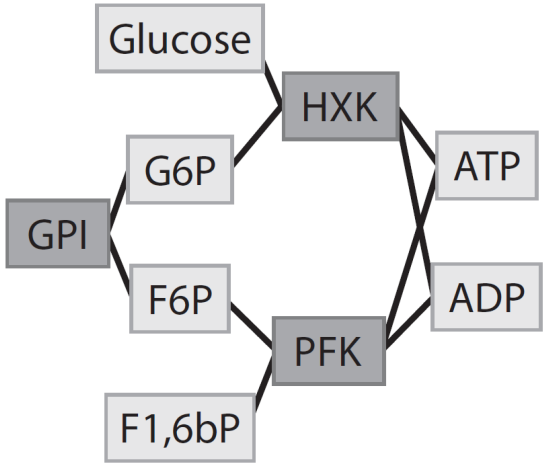
Complex and Dynamic Nature of Biological Systems

Study of the individual parts of metabolism is difficult because perturbing a single metabolic pathway may impact the function of a large part of the complete network.

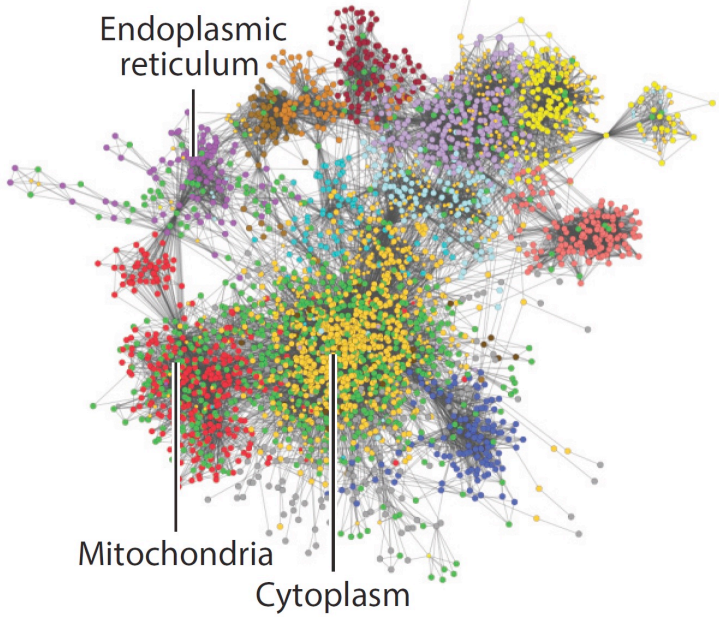
Canonical pathway



High degree of connectivity



Complex network at genome scale



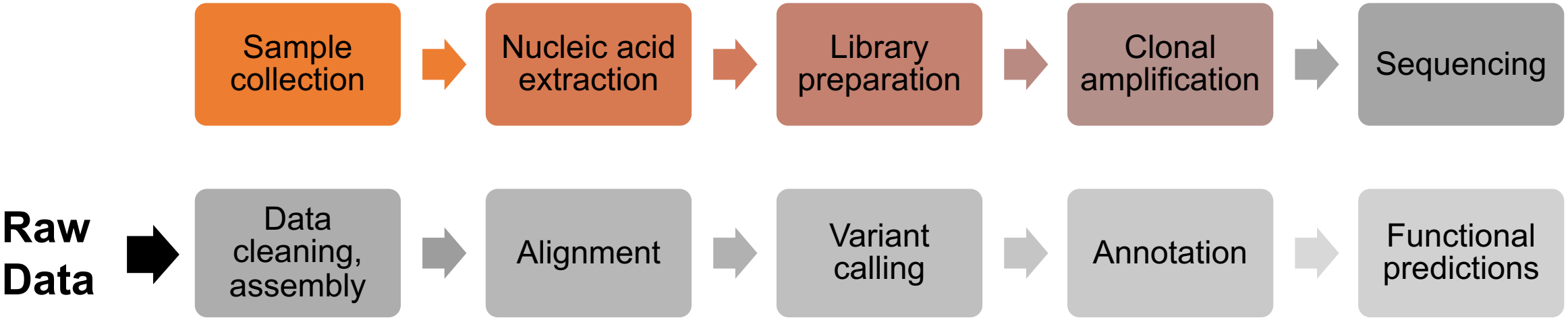
Nielsen 2017 Systems Biology of Metabolism. Annu Rev Biochem

Top-down Approach in Systems Biology with Omics

Genomics: Single nucleotide polymorphisms (SNPs), copy number variants (CNVs), loss of heterozygosity variants, genomic rearrangements and rare variants.

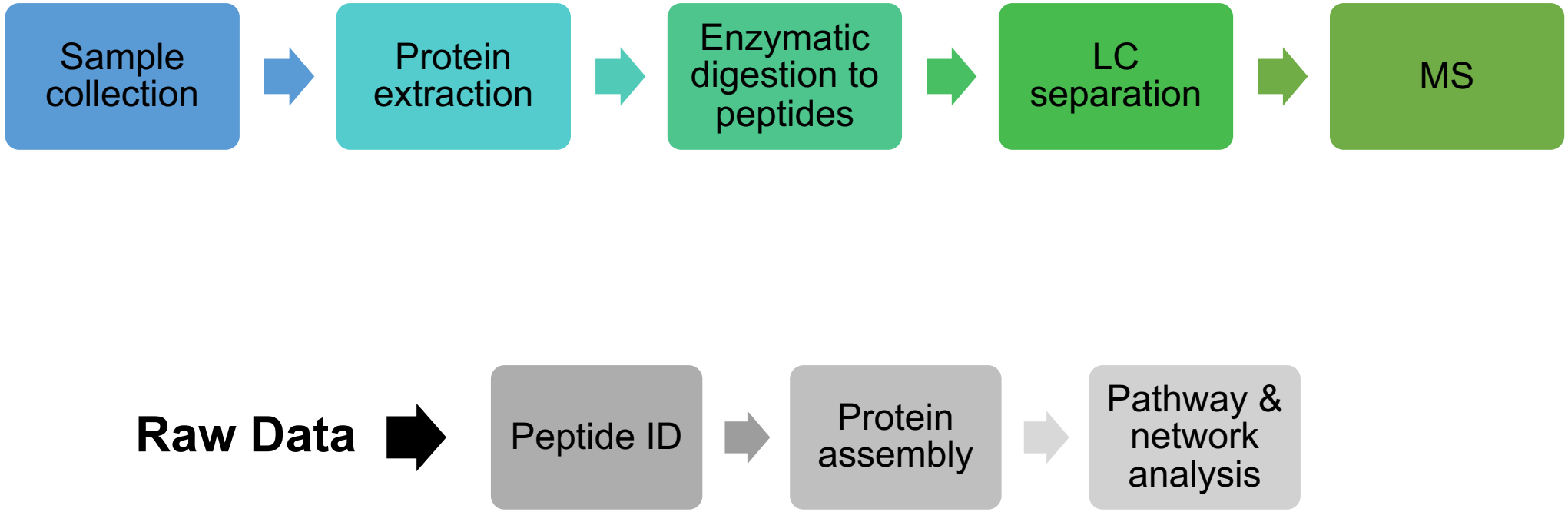
Epigenomics: DNA methylation, histone modification, chromatin accessibility, transcription factor binding.

Transcriptomics: Gene expression, alternative splicing, long-coding RNA, microRNA



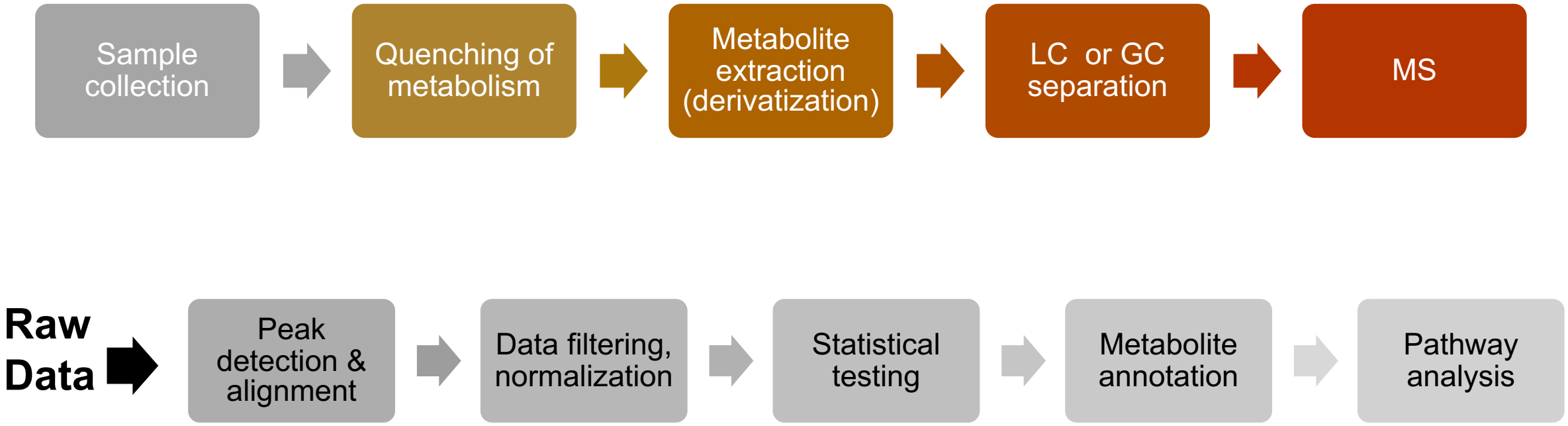
Top-down Approach in Systems Biology with Omics

Proteomics: Protein abundances, identification and quantification of post-translational modifications



Top-down Approach in Systems Biology with Omics

Metabolomics: Catabolic products, anabolic precursors, intermediates, nutrients, environmental chemicals, microbiome products.



Challenges of OMICS Data

The human body contains almost ~20,000 proteins, 20,000–22,000 protein-coding genes, ~30,000 mRNAs, 2300 miRNAs, and 114,100 metabolites, respectively.

"**Information Overload**": >10,000 variables per –omics experiment

The number of functionally relevant interactions between the components of this network (i.e. the links) is expected to be much larger.

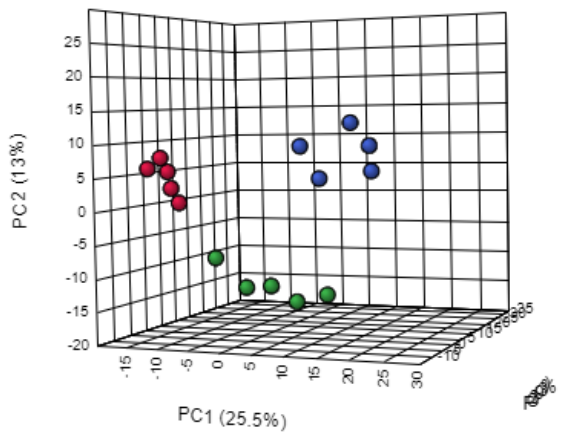
Biswas and Chakrabarti 2020 Front Oncol.

OMICS Data Reduction and Inference

"large p, small n" problem: Number of variables (*p*) measured >> Number of experimental units (*n*) (Johnstone and Titterington, 2009)

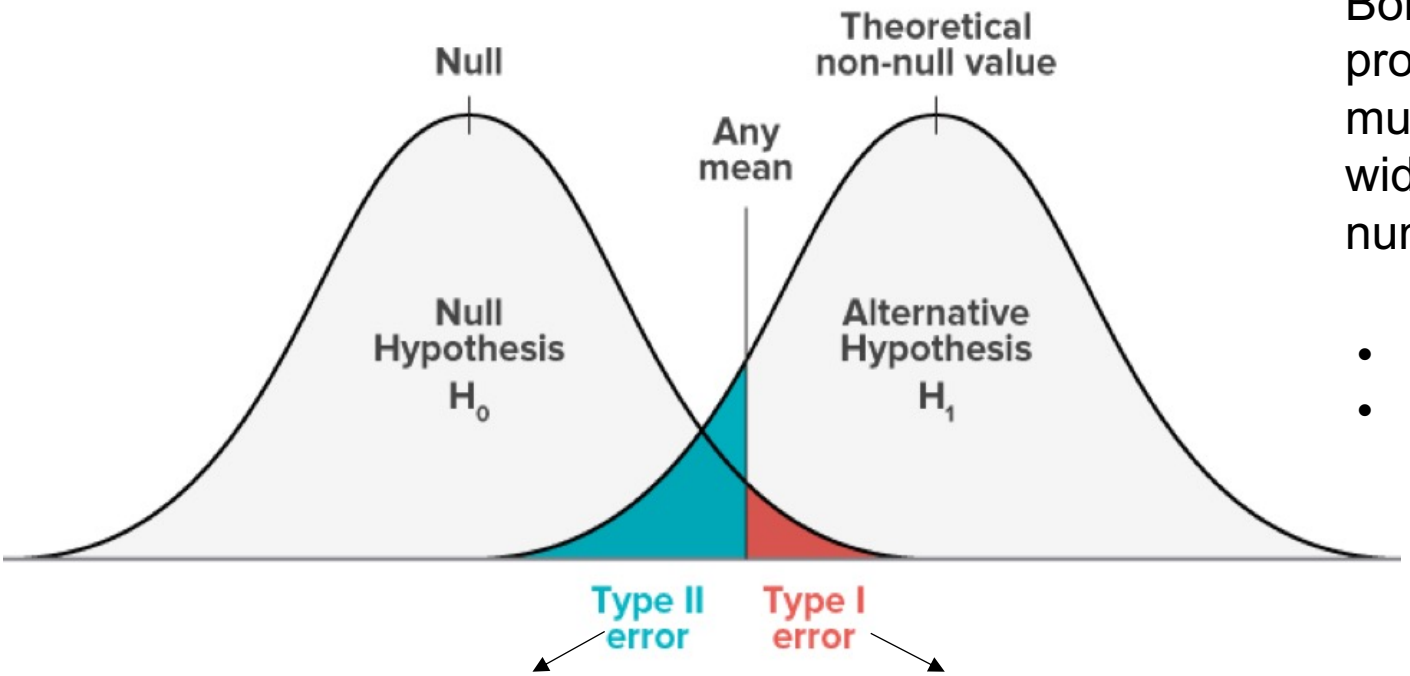
Data (dimension) Reduction

- **Principal component analysis (PCA)**: Orthogonally transforms the original coordinates of a data set into a new set of coordinates called principal components (PC). PC1 has the largest possible variance; each succeeding PC has the highest possible variance under the constraint that it is orthogonal to (i.e. uncorrelated with) with preceding components.



OMICS Data Reduction and Inference

Null Hypothesis Significance Testing



False negative

False positive

False discovery rate control for Type I error

Bonferroni, Benjamini-Hochberg or other procedures reduce the α for each test to a value much smaller than 0.05, so that the experiment-wide error is not as inflated due to the large number of comparisons being made.

- Increase the probability of **Type II errors**
- Implies that to conclude that a metabolite is affected by a treatment when it is not is much **worse** than to conclude that a metabolite is not affected by a treatment when, in reality, it is.

Biomarker discovery

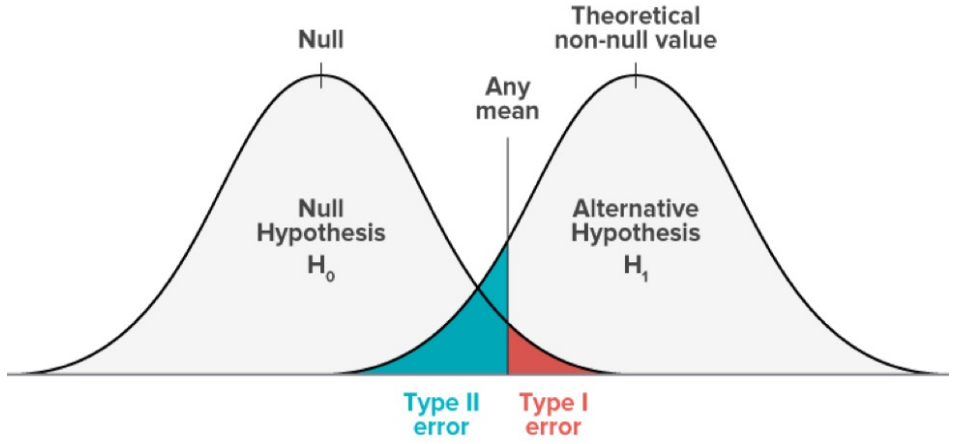
Biology

Statistical Significance vs Biological Significance

A two-step statistical strategy:

Step 1: Avoid type 2 error by selecting all features at raw P<0.05.

Step 2: Follow this with pathway enrichment, which uses permutation testing (pathway P<0.05) to determine whether the features initially selected at raw P<0.05 are enriched in pathways.



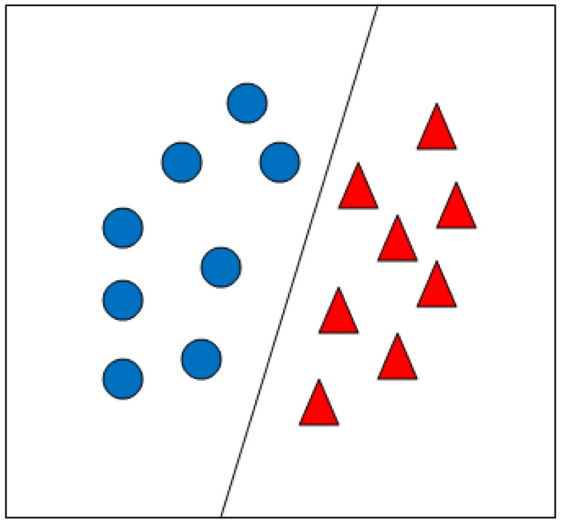
This second test protects against type 1 error because multiple metabolites in the pathway are changed, thereby providing confidence that the metabolites in this pathway are relevant.

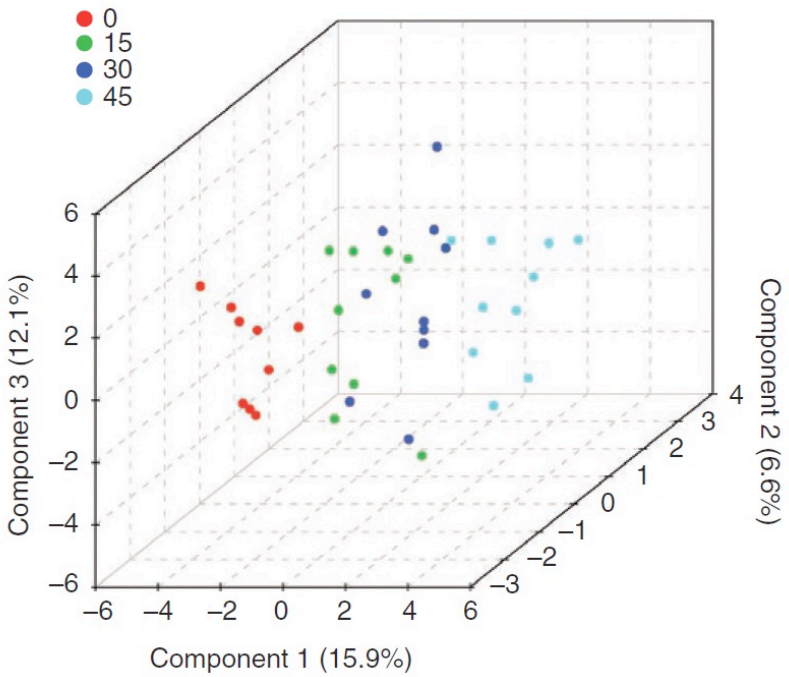
Alternatively, ranking and prioritizing of candidates based on cumulative evidence across data types and their variable can support more robust feature selection, such as GSEA (Subramanian et al. 2005), integRATE (Eidem et al. 2018), but are limited to gene/protein data analysis.

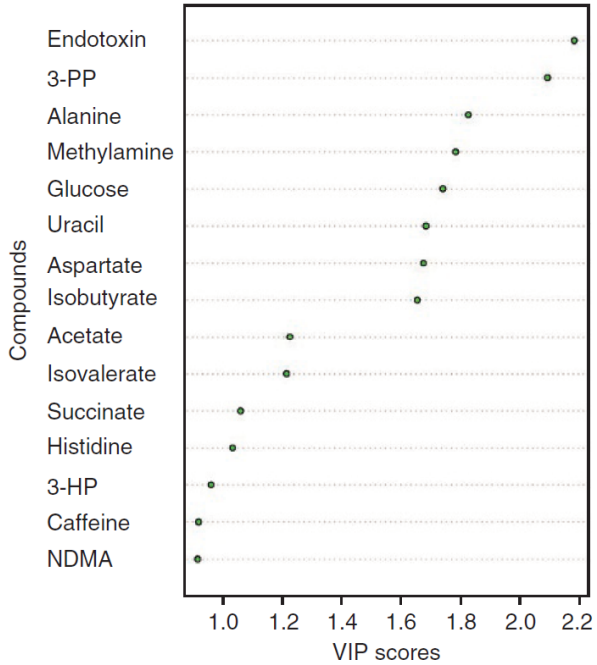
Data Reduction and Inference: Selecting Discriminatory Features

Contribution: Features contributing the most to separation of groups
Different from statistical significance!!

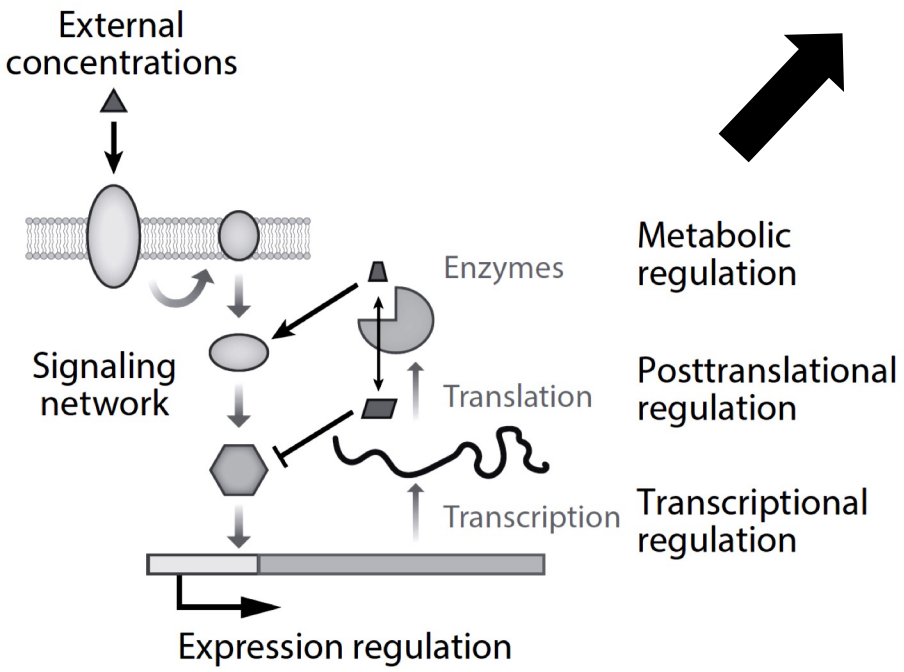
PLS-DA (Partial least squares discriminant analysis) is a supervised classification method. This means that class labels (*Y*) is used during the classification process. PLS-DA projects the data (*X*) into a low-dimensional space that maximizes the separation between different groups of data in the first few dimensions. PLS-DA also produces variable importance measures (**VIP**).

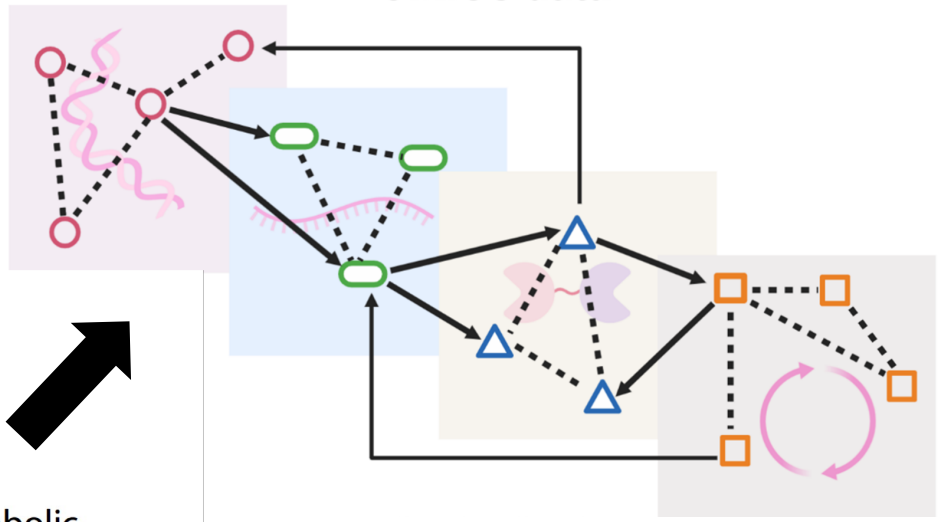






Why OMICS Data Integration?





Non-canonical data, e.g. immunomics, microbiome

Hu et al. 2020 Omics integration for mitochondrial systems biology. Antioxid Redox Signal
Nielsen 2017 Systems biology of metabolism. Annu Rev Biochem

Why OMICS Data Integration?

The regulatory mechanisms are distributed across different types of bio-entities.

Systems level analysis provides:

- More comprehensive overview of underlying mechanisms.
- Interactions between biomedical entities.

Combining multiple types of data collected on the same subjects compensate for noise or unreliable information in a single data type.

More confidence in results if multiple sources of evidence corroborate (orthogonal).

Design of Integrative Omics Study

Vertical or paired design

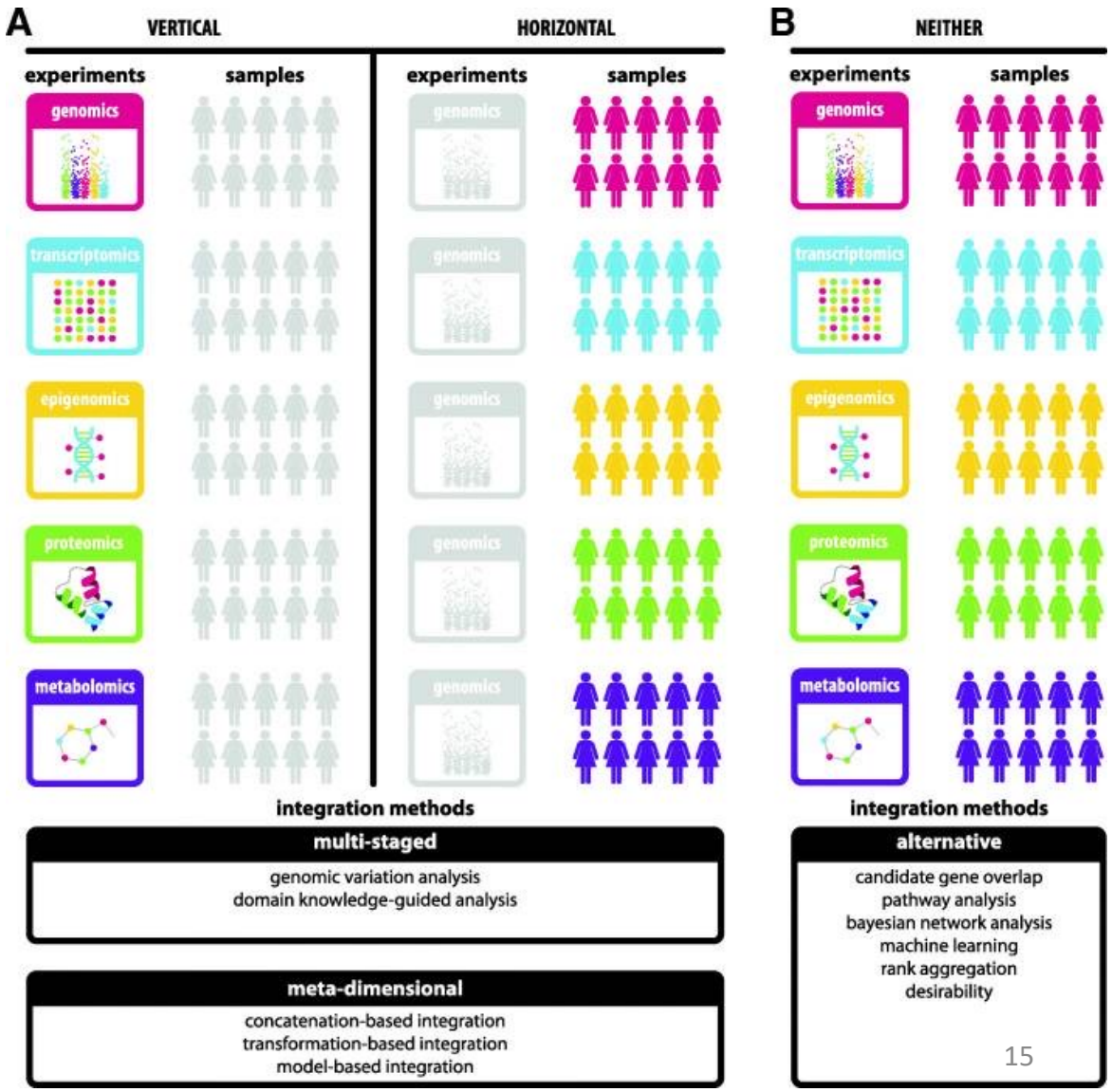
- Multiple omics data types from the same N subjects.
- Network of association among variables.

Horizontal or meta-analysis design

- Single omics data type from multiple studies/cohorts
- Cross-laboratory or cross-platform comparisons

Heterogeneous or unpaired design

- Different omics data collected study by study
- Prioritization of candidates for individual data.



Eidem et al. 2018. BMC Med Genomics. PMID:30453955

Main Approaches for Data Integration

Pathway-based integration

- Datasets are analyzed individually using selected candidates, e.g., differentially expressed genes, metabolites, proteins.
- Integration is performed at the pathway level – what pathways are overlapped?
- Tools: MetaboAnalyst, iPEAP, MetScape, MetaCore.

Network-based integration

- Multiple datasets are combined simultaneously and globally
- Pathway analysis tools are used to interpret the integrated data.
- Examples: 3Omics, mixOmics, xMWAS.

Main Approaches for Data Integration

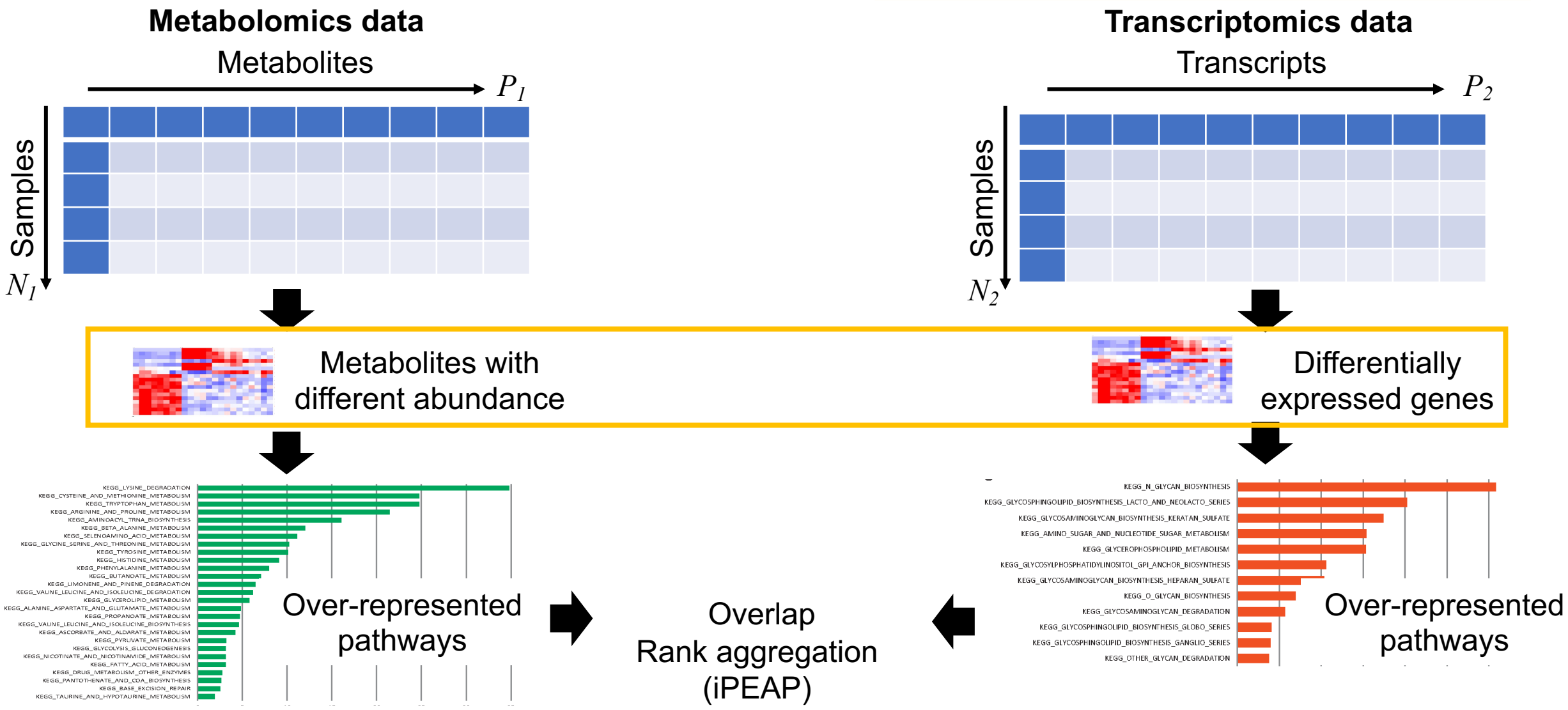
Pathway-based integration

- Datasets are analyzed individually using selected candidates, e.g., differentially expressed genes, metabolites, proteins.
- Integration is performed at the pathway level – what pathways are overlapped?
- Tools: MetaboAnalyst, iPEAP, MetScape, MetaCore.

Network-based integration

- Multiple datasets are combined simultaneously and globally
- Pathway analysis tools are used to interpret the integrated data.
- Examples: 3Omics, mixOmics, xMWAS.

Pathway-based Integration: General Workflow



Pathway-based Integration: Common Tools

Name	Platform	Function	Link	Reference
IMPaLa	Web	Joint pathway analysis	http://impala.molgen.mpg.de	Kamburov et al. (2011)
Ingenuity Pathway Analysis	License, Web, Local	Joint pathway analysis and mapping	Ingenuity Pathway Analysis | QIAGEN Digital Insights	Krämer et al. (2014)
MetaCore	License, Web, Local	Functional, joint and network pathway analysis	https://portal.genego.com/	
InCroMAP	JAVA	Joint pathway (enrichments, visualization) visualization	http://www.ra.cs.uni-tuebingen.de/software/InCroMAP/index.htm	Wrzodek et al. (2012)
MetaboAnalyst	Web	Joint pathway analysis and mapping, Network analysis	https://www.metaboanalyst.ca/	Xia et al. (2009)

MetaboAnalyst 5.0 Joint-Pathway Analysis

Please upload a gene list and a metabolite list below

Specify organism: Homo sapiens (human)

Try our example data

Gene list with optional fold changes

```
#Official  logFC
AASS  -0.139042168
ACAA2  1.401267672
ACADL  -2.608712824
ACADM  -0.876538515
ACADS  0.150535255
ACADSB  -1.637743607
ACHE  2.567118372
ACSM1  -2.348501729
ACTA2  -0.282176735
ACTB  1.559623747
ACTC1  -1.690352151
ADCY1  2.916857724
ADH1A  -0.87610472
AGL  -0.399133917
AGTR1  -1.078340189
AKR1A1  2.178398898
AKR1B1  -1.077265882
```

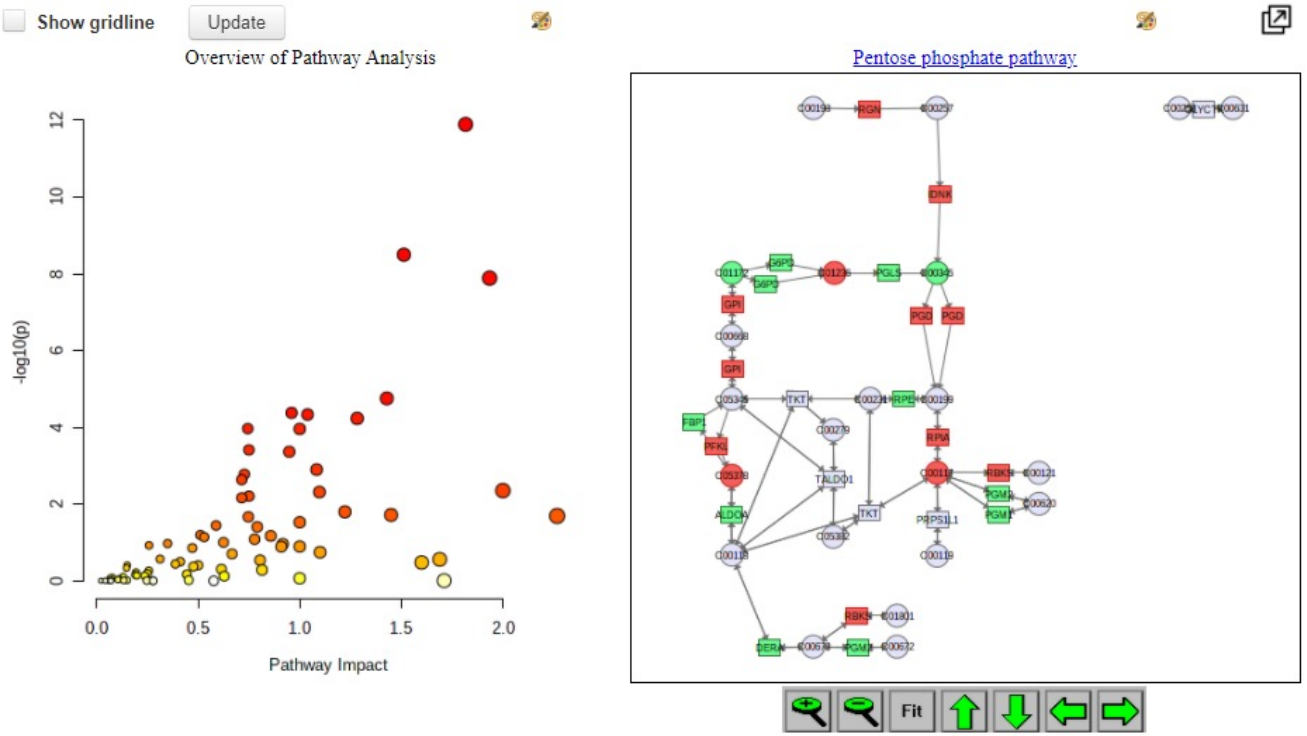
Compound list with optional fold changes

```
#KEGG  logFC
C00006  0.512160717
C00024  0.351757155
C00026  -2.669056963
C00029  0.379186578
C00031  1.669222153
C00047  -2.492289379
C00049  2.963835134
C00062  -2.558919927
C00064  1.77810046
C00072  0.632536475
C00077  -2.09045808
C00084  0.347392968
C00089  -1.460843412
C00097  3.046798674
C00101  -1.495004303
C00109  0.476718643
C00111  -2.672997377
```

ID Type: Official Gene Symbol

ID Type: KEGG ID

Submit



Tight integration by combining queries in which genes and metabolites are pooled into a single query and used to perform enrichment analysis within their "pooled universe".

Loose integration by combining p values in which enrichment analysis is performed separately for genes and metabolites in their "individual universe", and then individual p-values are combined via **weighted Z-tests.**

Pathway-based Integration: Pros & Cons

Advantages

- User-friendly, many web-based.
- Easy to reduce and prioritize each dataset by user-defined cutoffs.
- Easy to visualize and interpret in biological context.
- More confidence in overlapped pathways.
- Can be applied in unpaired study design – datasets do not have to be collected from the same cohort.

Limitations

- Rely on prior knowledge defined pathways and database.
- Limited application in relatively new field, such as microbiome.
- Bias toward certain pathways, gene sets or diseases.
- Lack of direct information on interactions.
- Cannot discover new interactions.

Main Approaches for Data Integration

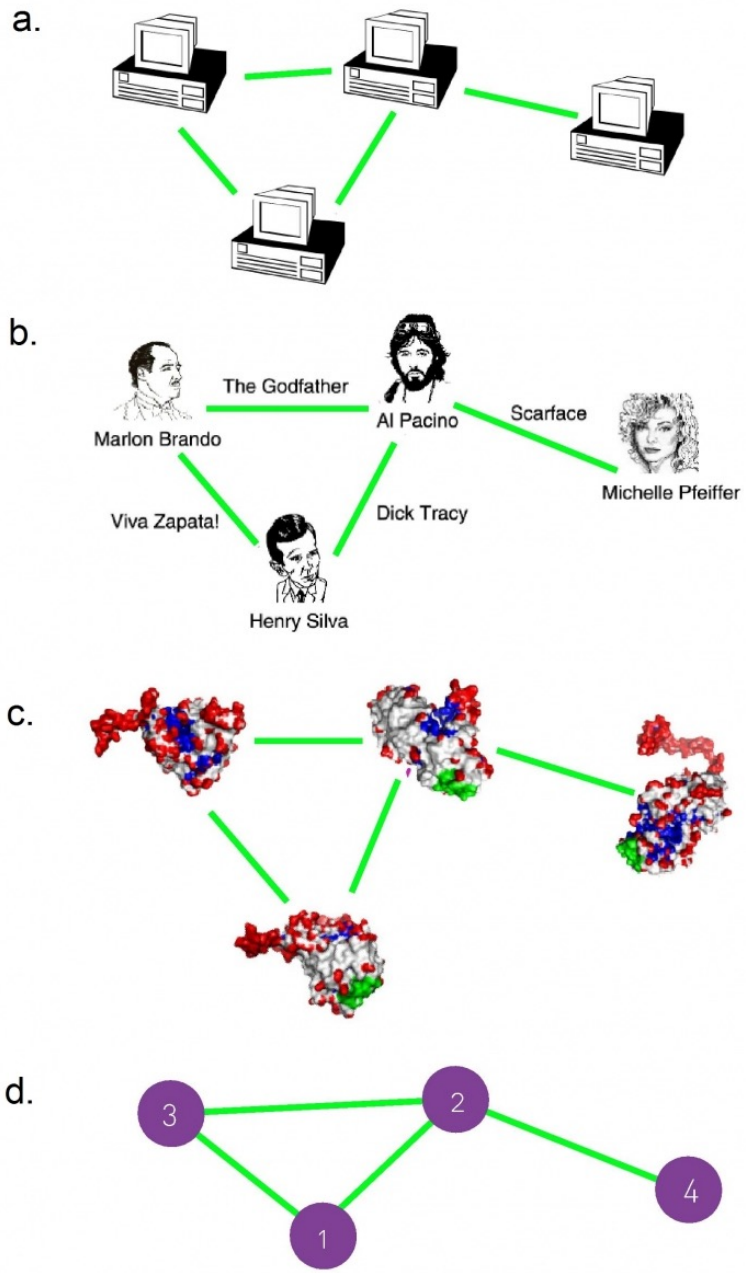
Pathway-based integration.

- Datasets are analyzed individually using selected candidates, e.g., differentially expressed genes, metabolites, proteins.
- Integration is performed at the pathway level – what pathways are overlapped?
- Tools: MetaboAnalyst, iPEAP, MetScape, MetaCore.

Network-based integration

- Multiple datasets are combined simultaneously and globally.
- Pathway analysis tools are used to interpret the integrated data.
- Examples: 3Omics, mixOmics, xMWAS.

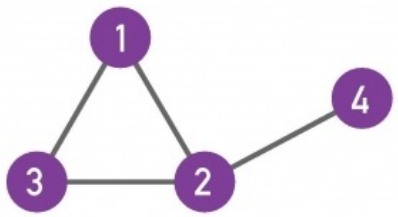
Let's start a network



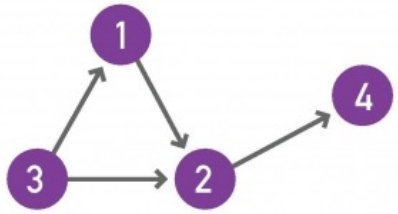
a. **Adjacency matrix**

$$A_{ij} = \begin{matrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{matrix}$$

b. **Undirected network**



c. **Directed network**



$$A_{ij} = \begin{matrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

$$A_{ij} = \begin{matrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

Barabási. Network Science

Creation of biological network maps

Protein networks: proteins that are linked to each other by physical (binding) interaction; based on immunoprecipitation and high-throughput mass spectrometry.

Metabolic networks: metabolites that are linked if they participate in the same biochemical reactions.

RNA networks: which capture the role of interactions between regulatory RNAs, such as small non-coding microRNAs (miRNAs) and small interfering RNAs (siRNAs), and DNA in regulating gene expression.

Regulatory networks: directed links represent either regulatory relationships between a transcription factor and a gene, or post-translational modifications

Barabási et al. 2011, Nat Rev Genet: doi: 10.1038/nrg2918

KEGG: Kyoto Encyclopedia of Genes and Genomes

https://www.genome.jp/kegg/

KEGG - Table of Contents

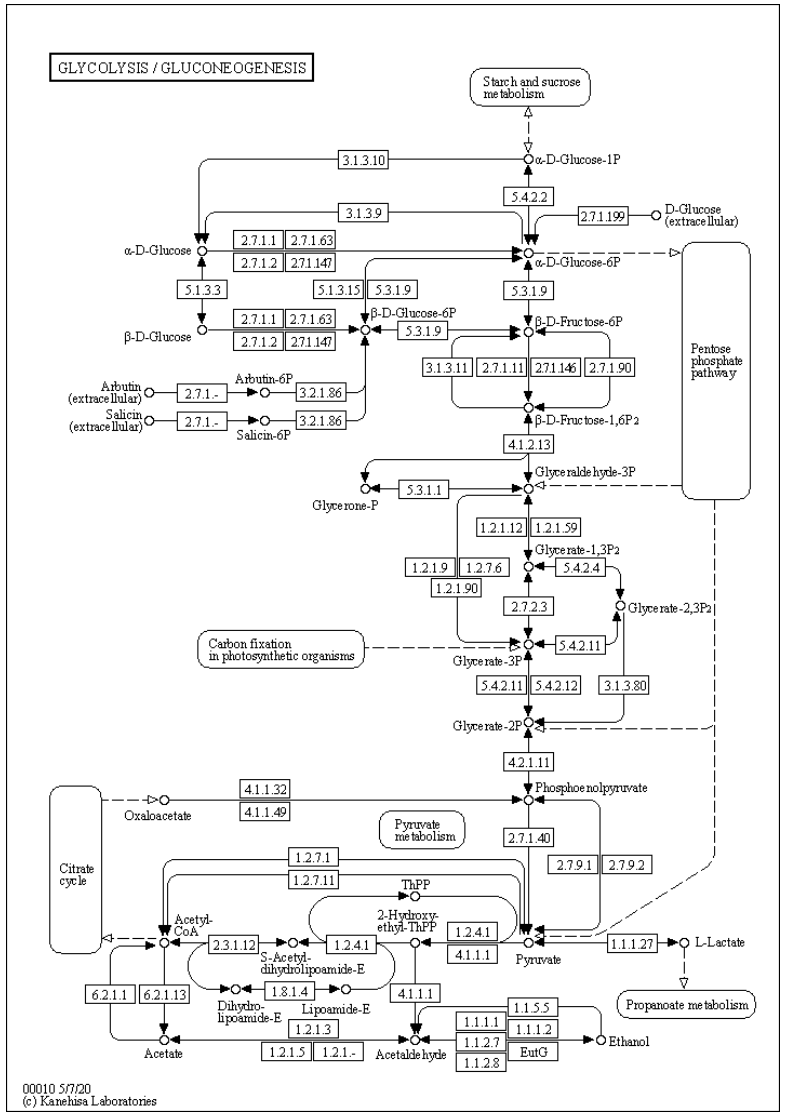
KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND DISEASE DRUG

Search KEGG for Go

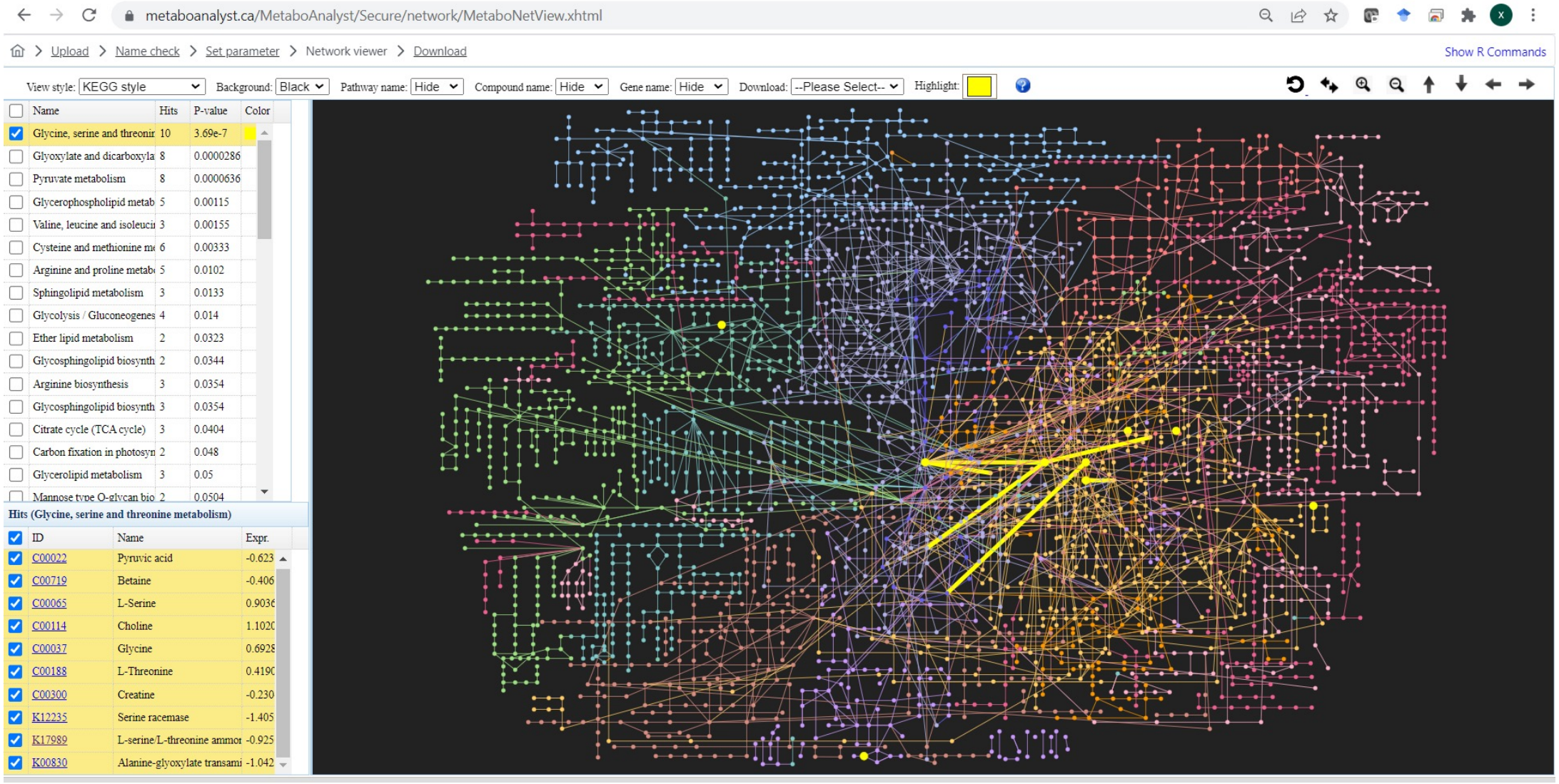
Data-oriented entry points

KEGG databases

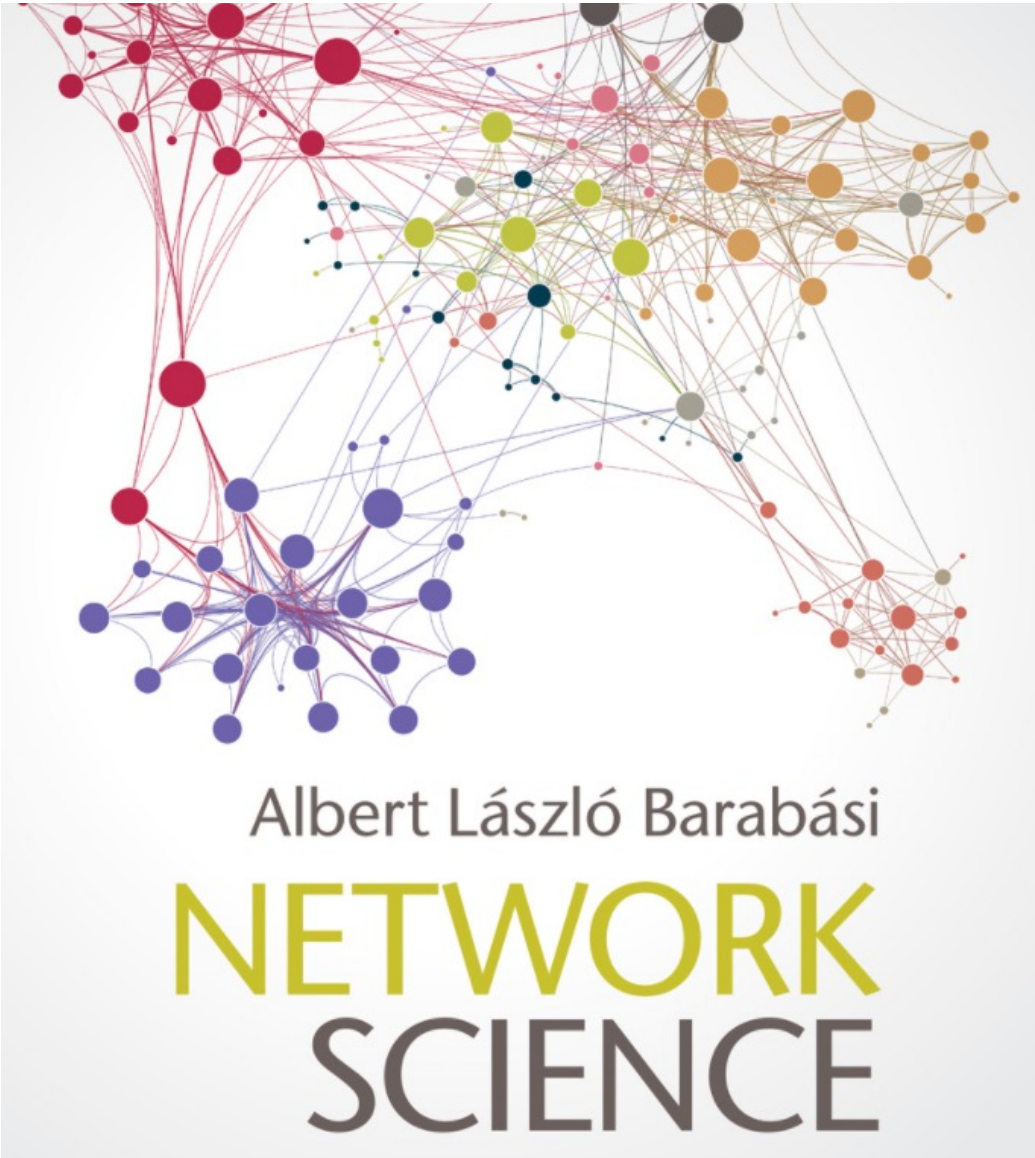
Category	Entry point	Database	Content	Classification
Systems information	KEGG PATHWAY	PATHWAY	KEGG pathway maps	Pathway maps
	KEGG BRITE	BRITE	BRITE hierarchies and tables	Brite hierarchies Brite tables
	KEGG MODULE KEGG RModule	MODULE	KEGG modules and reaction modules	Modules Reaction modules
Genomic information	KEGG ORTHOLOGY KEGG Annotation	KO	Functional orthologs	KO
	KEGG GENES KEGG SeqData	GENES	Genes and proteins	
	KEGG GENOME KEGG Virus KEGG Taxonomy	GENOME	Genomes of cellular organisms and viruses	Organisms Viruses
Chemical information	KEGG COMPOUND	COMPOUND	Metabolites and other small molecules	Compounds
	KEGG GLYCAN	GLYCAN	Glycans	
	KEGG REACTION	REACTION RCLASS	Biochemical reactions Reaction class	
	KEGG Enzyme	ENZYME	Enzyme nomenclature with sequence data	EC sequence data
Health information	KEGG NETWORK	NETWORK VARIANT	Disease-related network variations Human gene variants	Network variation maps
	KEGG DISEASE	DISEASE	Human diseases	Human diseases Infectious diseases
	KEGG DRUG	DRUG DGROUP	Drugs Drug groups	Drugs (ATC) Drugs (target) Antiinfectives



MetaboAnalyst 5.0: Network Analysis Tool – driven by existing biological knowledge



Development of Network Science: A data-driven approach



http://networksciencebook.com/

REVIEWS

Network medicine: a network-based approach to human disease

Albert-László Barabási[*‡§], *Natali Gulbahce*[*‡||] *and Joseph Loscalzo*[§]

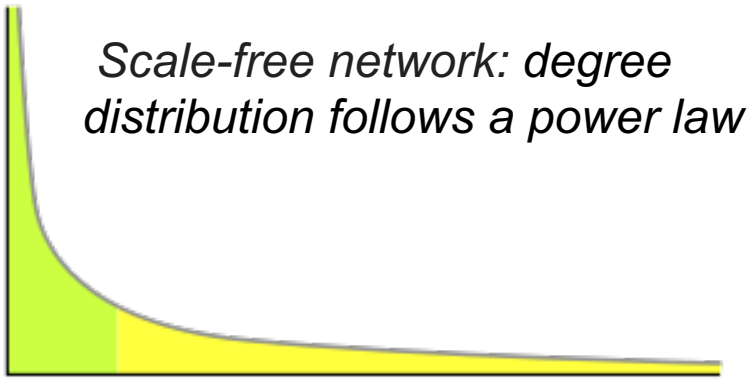
Abstract | Given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems. The emerging tools of network medicine offer a platform to explore systematically not only the molecular complexity of a particular disease, leading to the identification of disease modules and pathways, but also the molecular relationships among apparently distinct (patho)phenotypes. Advances in this direction are essential for identifying new disease genes, for uncovering the biological significance of disease-associated mutations identified by genome-wide association studies and full-genome sequencing, and for identifying drug targets and biomarkers for complex diseases.

Elements of network theory

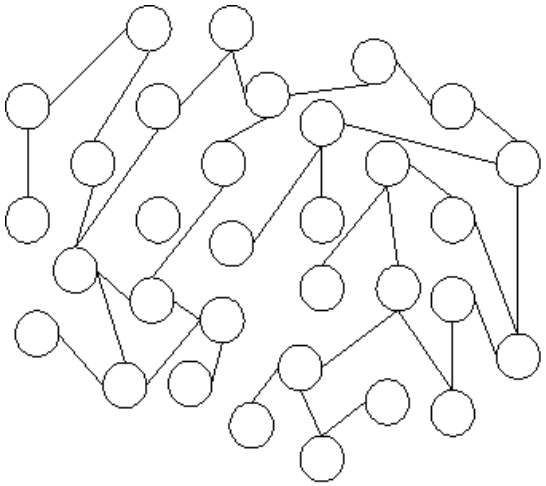
Basic principle: Real networks (*e.g.* found in natural, technological and social systems) are not random, but follow a series of basic organizing principles in their structure and evolution that distinguish them from randomly linked networks.

Scale-free property: many real networks, including human protein–protein interaction and metabolic networks, are scale free. The probability of observing a high-degree node, or hub, is several orders of magnitude higher in a scale-free than in a random network.

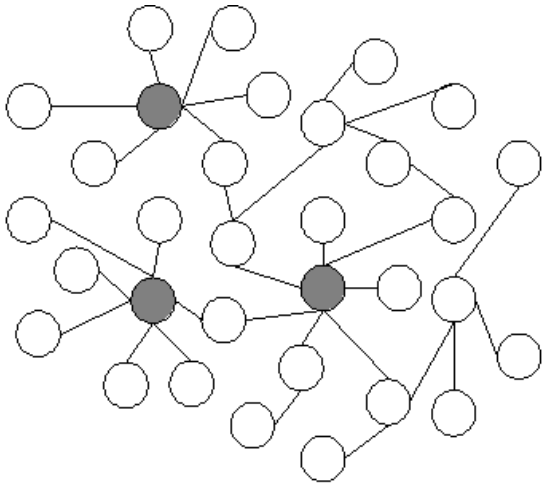
Hubs: a few highly connected nodes.



An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate.

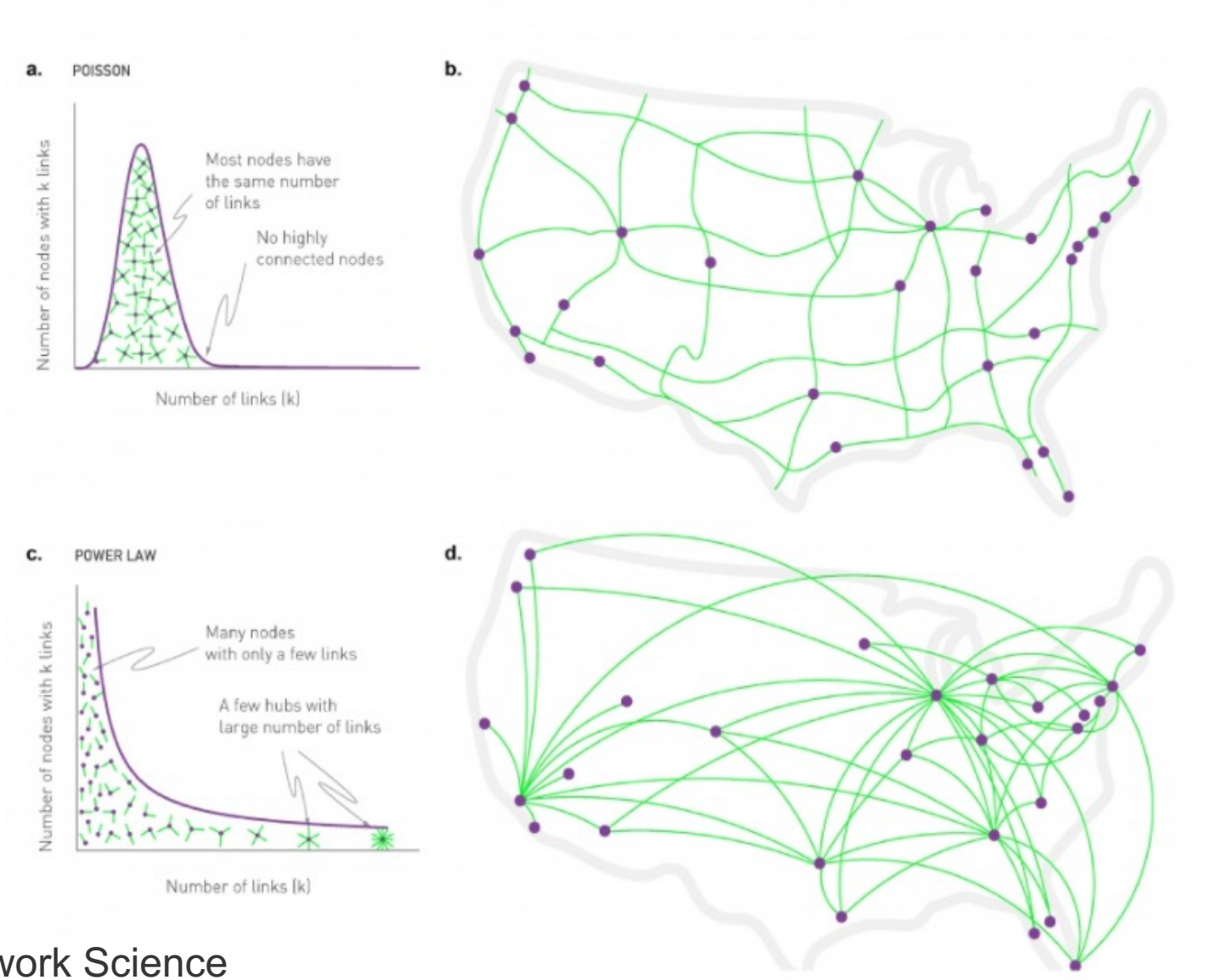


(a) Random network



(b) Scale-free network

Random vs. Scale-free Networks



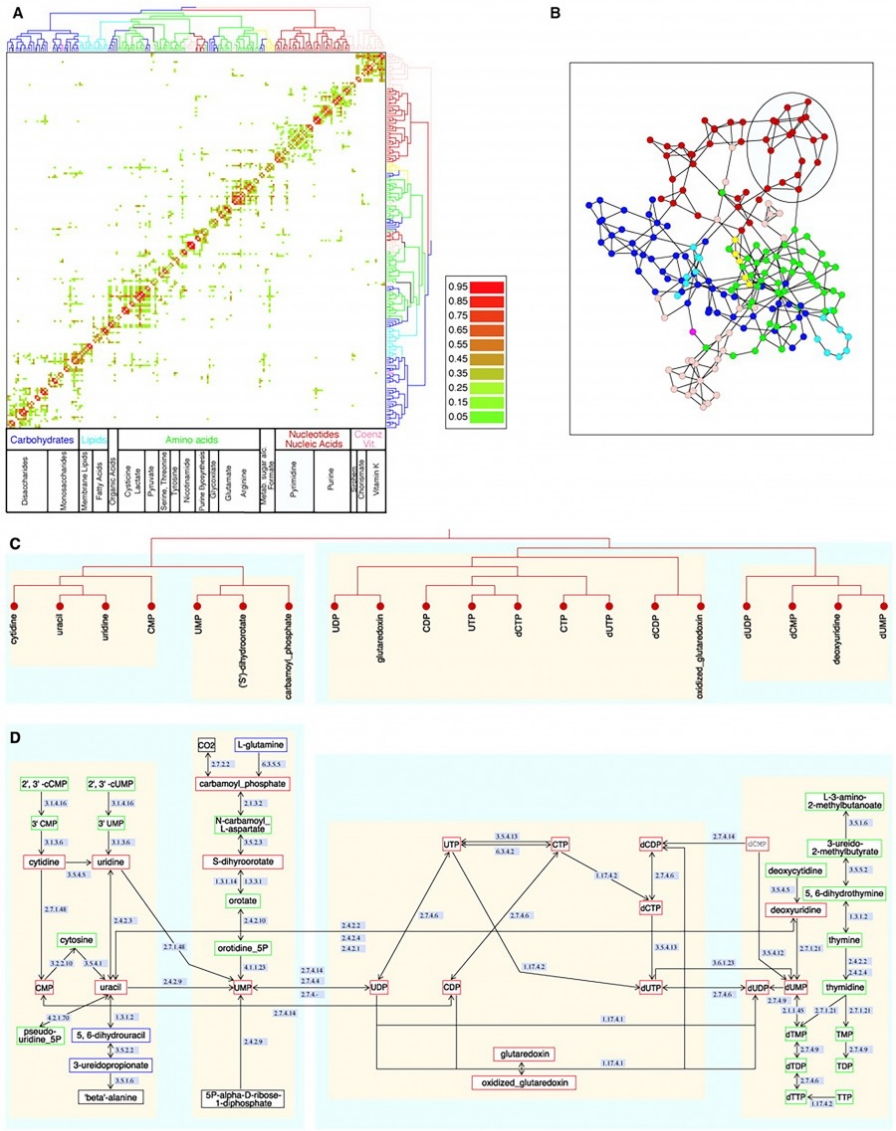
National Highway Network

vs

Air-traffic network

Barabási. Network Science

Towards modular biology



Communities in E.coli Metabolic Networks

Biology must move beyond its focus on single genes. It must explore instead how groups of molecules form functional modules to carry out a specific cellular functions. – Lee Harwell

A network's community structure is uniquely encoded in its wiring diagram.

A community is a locally dense connected subgraph in a network.

Randomly wired networks lack an inherent community structure.

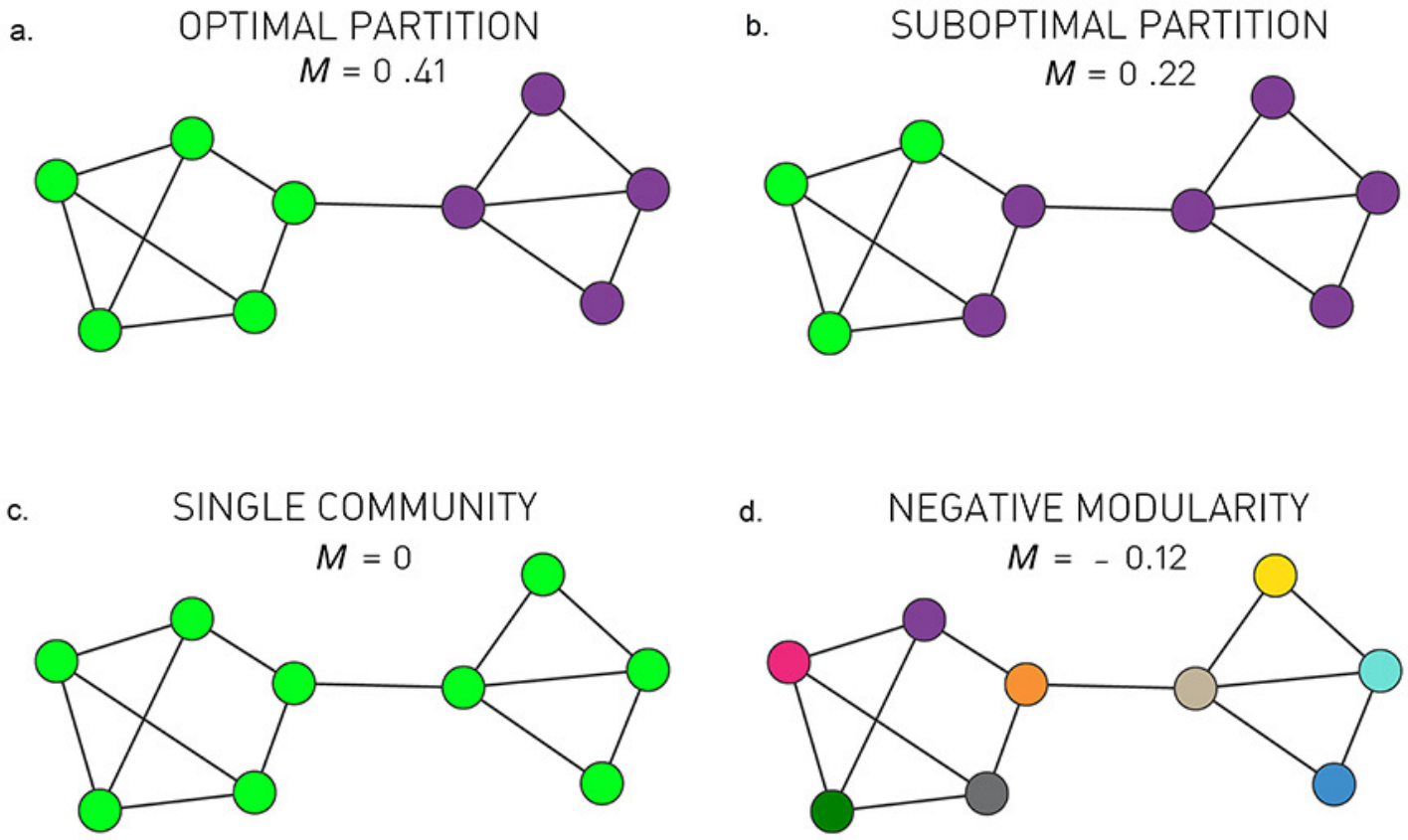
For a given network the partition with maximum modularity corresponds to the optimal community structure.

Barabási. Network Science

Topology-based Community Detection

Multilevel community detection:

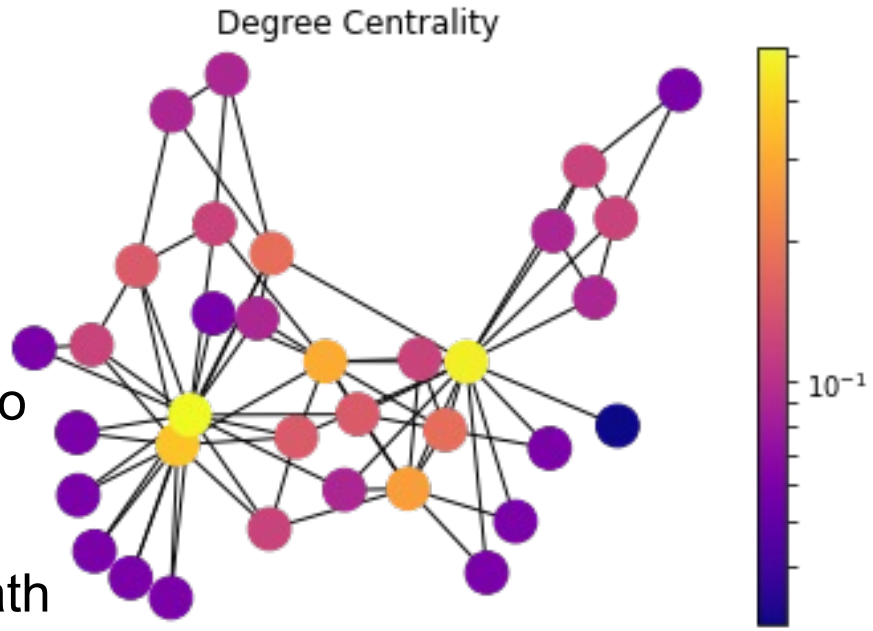
i. Each node is assigned to a different community;
ii. Each node is moved to a community with which it achieves the highest positive contribution to modularity
iii. Step 2 is repeated for all nodes until no improvement can be achieved
iv. Each community after step 3 is now considered a node and step 2 is repeated until there is a single node left or the modularity can no longer be improved



Newman 2006, PNAS

Centrality Analysis

Centrality: Measure of importance of a node in the network.



Common centrality measures:

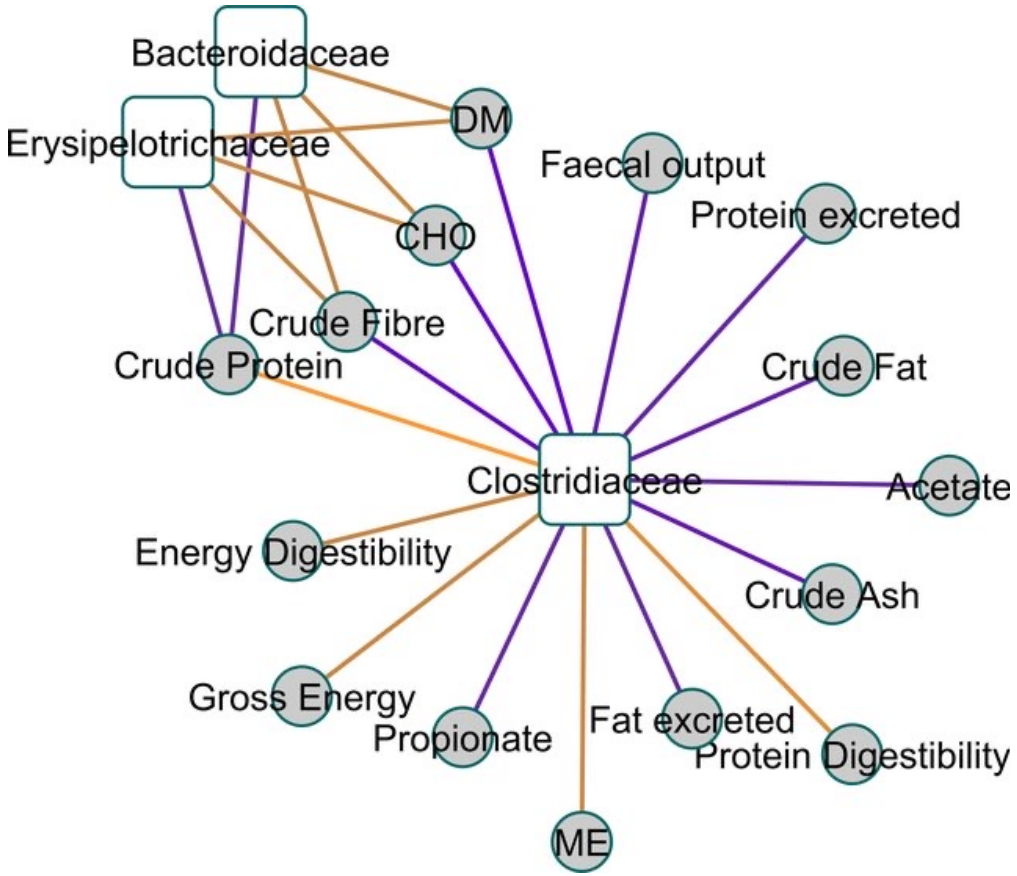
- Degree: based on the number of connections.
- Eigenvector: based on the number and quality of connections. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.
- Betweenness: based on the extent to which a node lies on the path between other nodes. Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.
- Closeness: average length of the shortest path between the node and all other nodes in the graph. The more central a node is, the closer it is to all other nodes.

Differential centrality analysis: difference between centrality under two conditions (e.g. $|\text{centrality}_{\text{exposed}} - \text{centrality}_{\text{control}}|$.

Relevance Network

Network of highly-correlated biomedical/clinical entities (Butte et al. 2000)

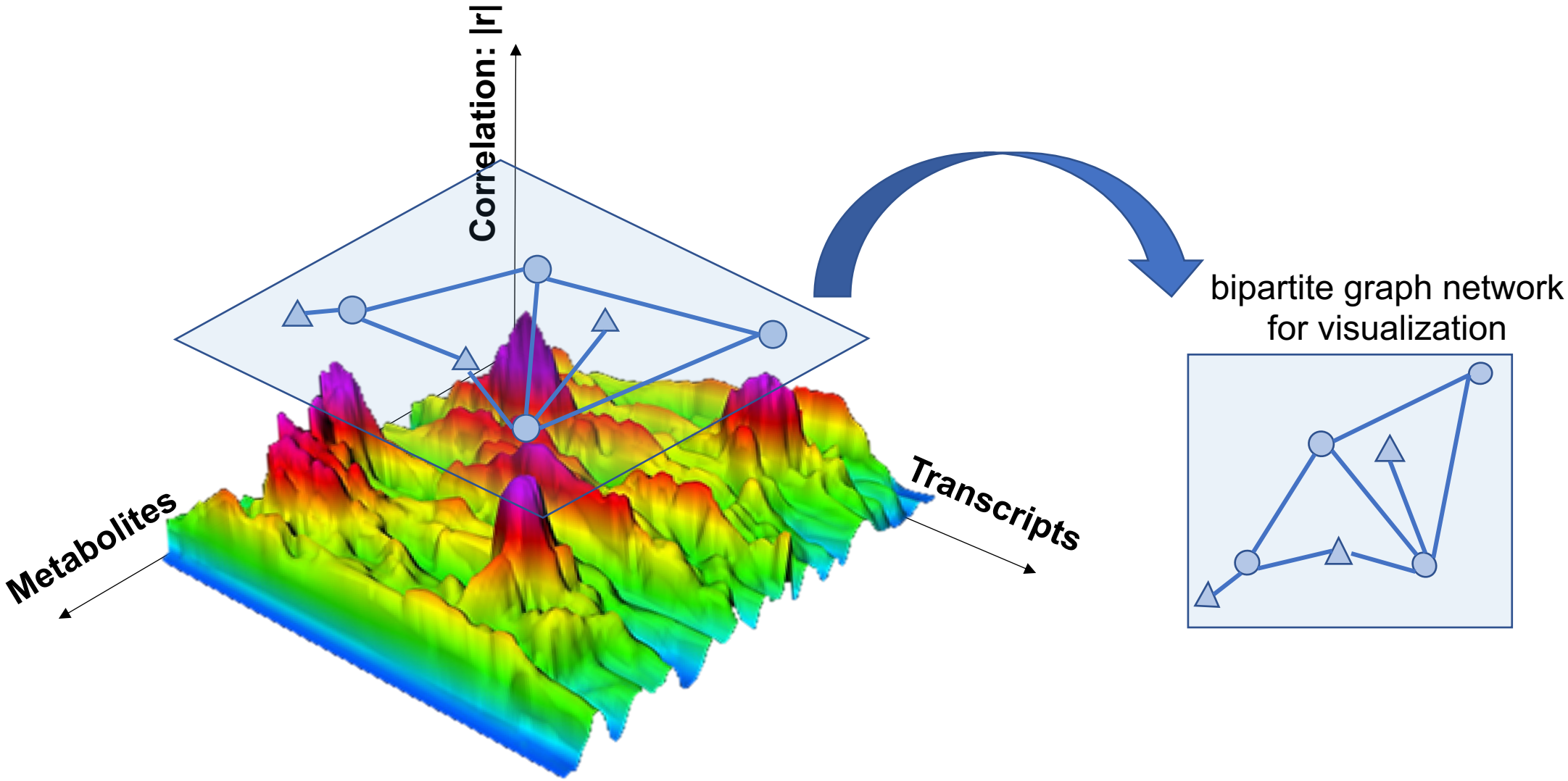
- Metabolomics x Proteomics, Transcriptomics x Proteomics, Metabolomics x Microbiome, Metabolomics x Clinical outcome, etc.
- Can generate a bipartite graph network for visualization.



Integration of faecal 16S genomic DNA amplicon data with physical measurement and metabolomics data.

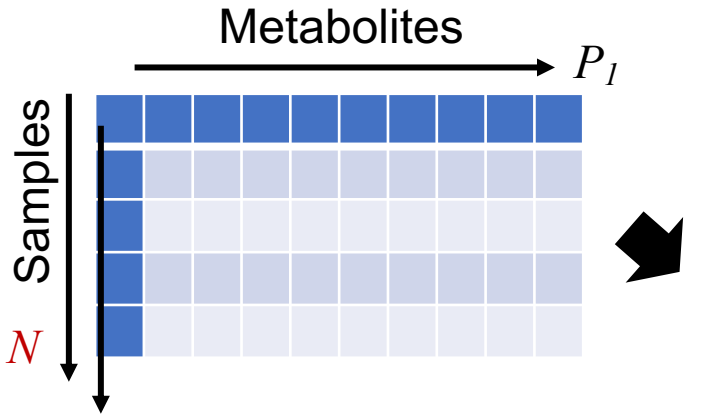
Bermingham et al. PeerJ. 2019

Transcriptome-metabolome wide association study (TMWAS)

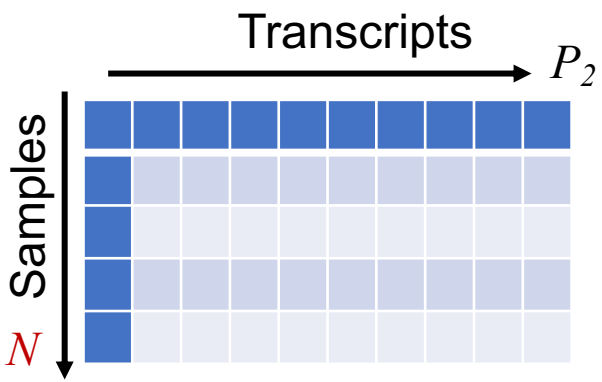


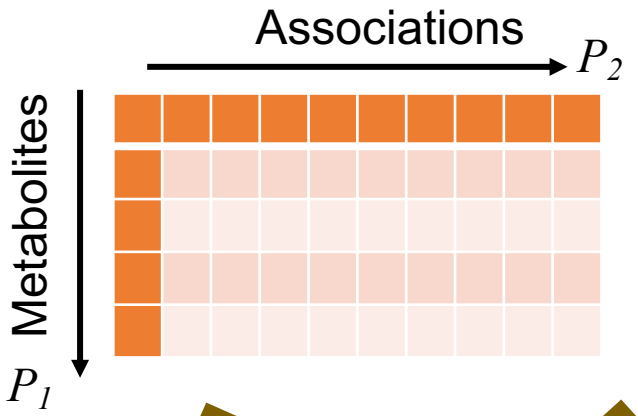
Data-driven Integration: General Workflow

Metabolomics data

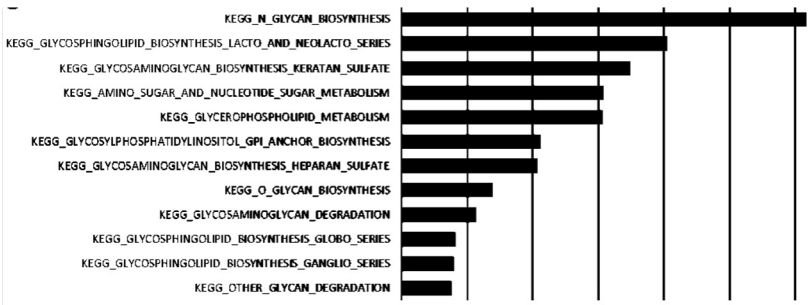


Transcriptomics data

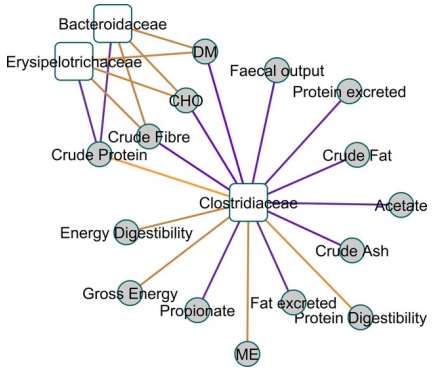




Over-represented pathways

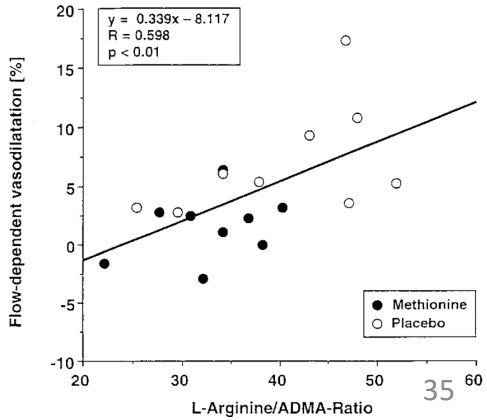


Relevance Network



Clustering analysis

Targeted analysis



Methods for Generating Relevance Networks: Univariate

Univariate methods: Pairwise Pearson or Spearman correlation between different omics data.

Name	Platform	Function	Link	Reference
3Omics	Web	Correlation analysis and network visualization of up to three data set	http://3omics.cmdm.tw/	Kuo et al. (2013)
MetabNet	R	Correlation analysis of metabolites	https://rdrr.io/github/kuppal2/xmsPANDA/man/metabnet.html	Uppal et al. (2015)

Methods for Generating Relevance Networks: Multivariate

Multivariate methods: Multivariate regression techniques such as partial least squares (PLS), sparse partial least squares regression (sPLS), multilevel sparse partial least squares (msPLS) regression, etc.

Name	Platform	Function	Link	Reference
mixOmics	R	Provides a wide range of linear multivariate methods for data exploration, integration, dimension reduction and visualization of biological data sets	http://mixomics.org/	Rohart et al. (2017)
xMWAS	R, web	Data-driven integration with multivariate regression and differential network analysis	https://kuppal.shinyapps.io/xmwas (Online) and http://github.com/kuppal2/xMWAS/ (R)	Uppal et al. (2018)

PLS and sPLS Regression in Data-driven Integration

Partial Least Squares (PLS) regression is not limited to uncorrelated variables. It can handle many noisy, collinear (correlated) and missing variables, and can also simultaneously model several response variables Y.
Regression mode: the goal is to predict Y from X (Y and X play an asymmetric role).
Canonical mode: X and Y play a symmetric role.

Sparse Partial Least Squares (sPLS) regression: In addition to PLS, sPLS performs simultaneous variable selection in the two data sets, by introducing LASSO penalization on the pair of loading vectors. The user has to specify the number of variables to select, *keepX*, *keepY*.

Multilevel sPLS: accounts for repeated measurements.

http://mixomics.org/methods/spls/

xMWAS Workflow with Differential Network Analysis

Data matrices from multiple assays

↓

Integrative and association analysis
Pairwise integrative and network analysis (e.g. X<->Y; X<->Z, Y<->Z) using (sparse) Partial Least Squares regression

↓

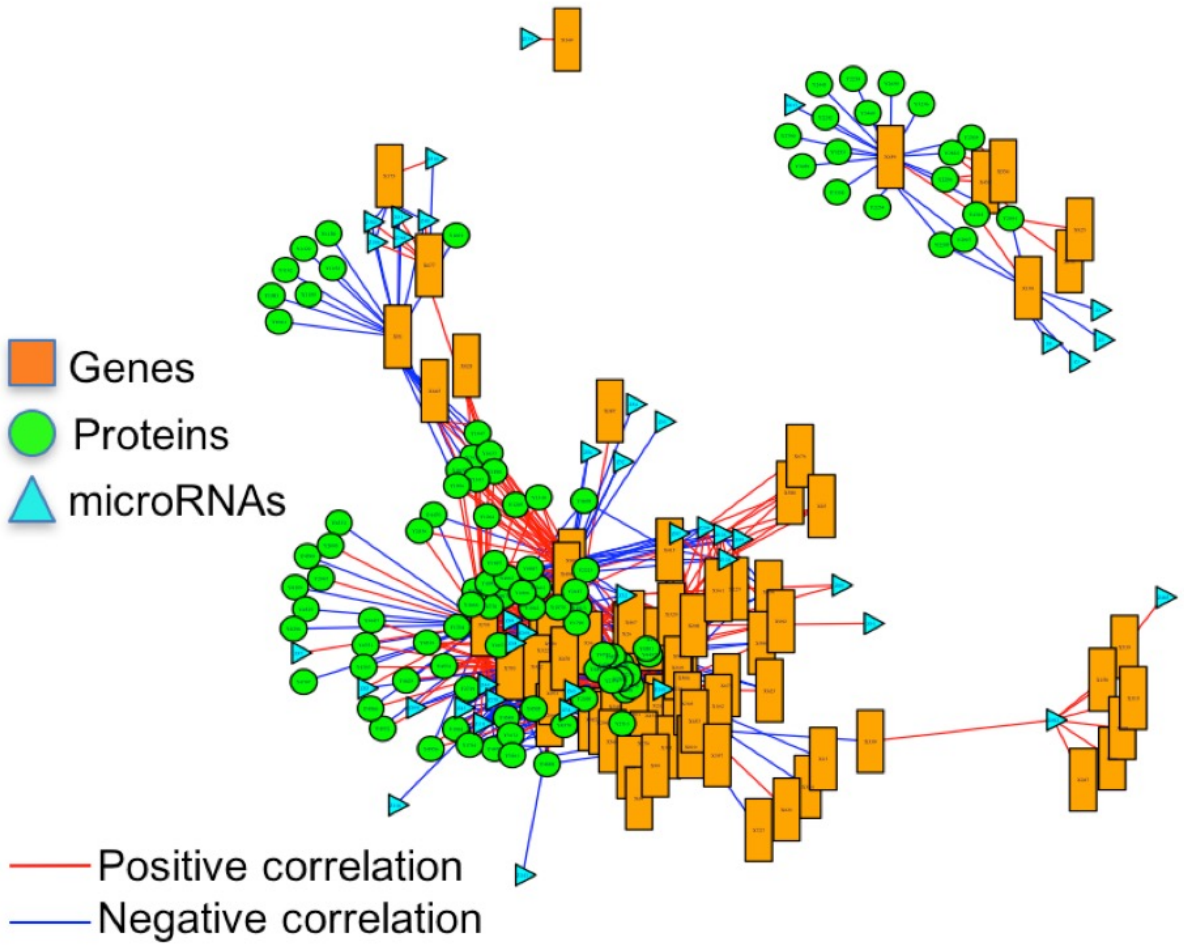
Generate an edge list matrix, L_e, with the list of edges that meet the significance criteria and association score threshold

↓

Generate a graph, G, using the union of edge list matrices

↓

Community detection, differential centrality and rewiring analysis, and network visualization



xMWAS (Web): Step 1 – Uploading Data

http://kuppal.shinyapps.io/xmwas/

xMWAS - a data-driven integration and network analysis tool (v0.552)

Introduction | Analysis | Help and Support

Input Files

Choose Files (see help and support)

Parameter Settings

1. Data preparation and filtering
2. Integration and association analysis
3. Centrality analysis
4. Graphical options

Input file for dataset A ('.csv' or '.txt', 100MB limit)

Browse... No file selected

Name for dataset A:

datasetA

Input file for dataset B ('.csv' or '.txt', 100MB limit)

Browse... No file selected

Name for dataset B:

datasetB

Up to 4 data sets

Add more datasets: + −

Choose a class labels file ('.csv' or '.txt'):

Browse... No file selected

More Options

Start processing | Download results

xMWAS (Web): Step 2 – Data Preprocessing

http://kuppal.shinyapps.io/xmwas/

Input Files

Choose Files (see help and support)

Parameter Settings

1. Data preparation and filtering

2. Integration and association analysis

3. Centrality analysis

4. Graphical options

Start processing | Download results

Relative Standard Deviation (RSD) Threshold (rows):

1

Maximum #datasetA variables to select based on RSD (change according to your dataset):

10000

Maximum #datasetB variables to select based on RSD (change according to your dataset):

10000

Minimum non-missing sample ratio (rows):

0

How are the missing values represented in the data?:

0

xMWAS (Web): Step 3 – Integration parameters

http://kuppal.shinyapps.io/xmwas/

Input Files

Choose Files (see help and support)

Parameter Settings

1. Data preparation and filtering

2. Integration and association analysis

3. Centrality analysis

4. Graphical options

Start processing | Download results

Pairwise integrative analysis

Choose a data integration method:

PLS: Partial least squares

Choose PLS mode (not applicable to RCC option):

regression

Number of components to use in PLS model:

5

Find optimal number of PLS components? (Note: turning this option ON may increase run time)

True False

Association analysis

Correlation Threshold:

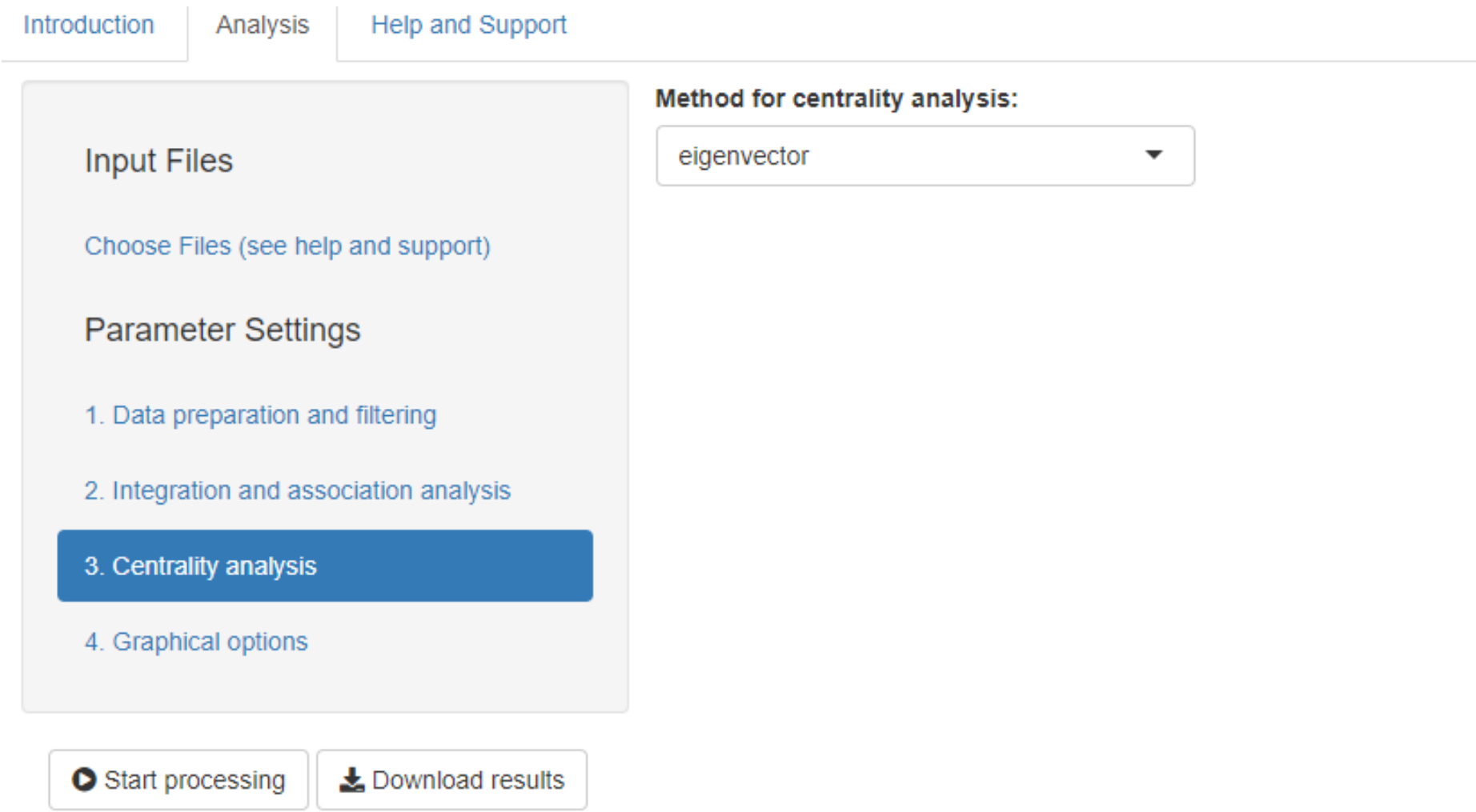
0.4

P-value Threshold For Student's T-test:

0.05

xMWAS (Web): Step 4 – Methods for Centrality

http://kuppal.shinyapps.io/xmwas/



xMWAS (Web): Step 5 – Graphic Options

http://kuppal.shinyapps.io/xmwas/

Input Files

Choose Files (see help and support)

Parameter Settings

1. Data preparation and filtering

2. Integration and association analysis

3. Centrality analysis

4. Graphical options

Start processing | Download results

Size of the Labels:

0.25

Size of the Nodes:

7

Seed for Random Number Generator:

100

Maximum number of associations to include in the network (any numeric value >0 or -1 to use all):

-1

Use dataset A as reference?

○ True ◉ False

Node shape for dataset A:

square

Node shape for dataset B:

circle

Node shape for dataset C:

triangle

Node shape for dataset D:

star

xMWAS (Web): Step 6 – Download and Enjoy!

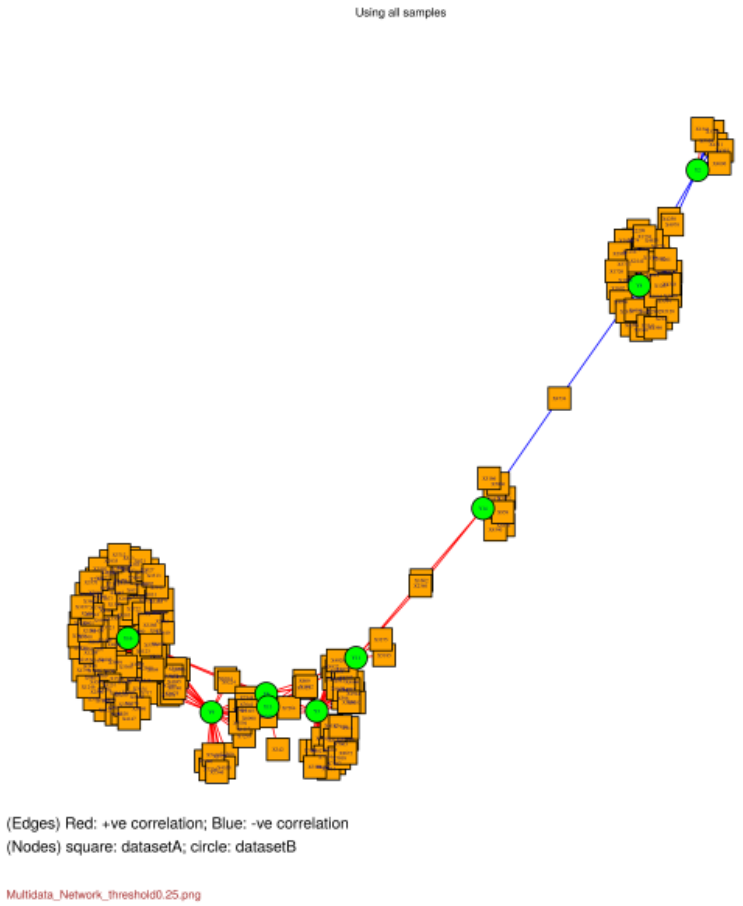
Start processing | Download results

Output

Slide to go to next figure:

0 1 2 3

Name	Type	C
pairwise_results	File folder	
cluster_membership_centrality_tab...	TXT File	
InputParameters.txt	TXT File	
interactive.network.communities.h...	Microsoft Edge HTML Do...	
LogApr01_21_11.txt	TXT File	
Multidata_Network_threshold0.2.p...	PNG File	
Multidata_Network_threshold0.2_c...	PNG File	
Multidata_Network_threshold0.2_li...	TXT File	
Multidata_Network_threshold0.2cy...	GML File	
Multidata_Network_threshold0.25....	PNG File	
Multidata_Network_threshold0.25_...	PNG File	
Multidata_Network_threshold0.25_l...	TXT File	
Multidata_Network_threshold0.25c...	GML File	
Network_stats.csv	Microsoft Excel Comma S...	
NodeID_Name_mapping.txt	TXT File	
README.txt	TXT File	



Data-driven Association-based Integration: Pros & Cons

Advantages

- Avoid bias generated by prior knowledge.
- Feasible in any field with any level of curated database.
- Keeps information on interactions.
- Provides systems-level overview and visualization.
- Reveals novel hypothesis – useful for experimental studies.

Limitations

- Must use paired data collected from the same cohort of samples
- Less user-friendly.
- Higher computational requirement.
- Biological interpretation challenges.

Case Study: *Haemophilus ducreyi* Infection Induces Oxidative Stress, Central Metabolic Changes, and a Mixed Pro- and Anti-inflammatory Environment in the Human Host

Collaboration between UAB, Emory and Indiana Univ.

Julie A. Brothwell, Kate R. Fortney, Hongyu Gao, Landon S. Wilson, Caroline F. Andrews, Tuan M. Tran, Xin Hu, Teresa A. Batteiger, Stephen Barnes, Yunlong Liu, Stanley M. Spinola

Healthy adult volunteers are infected with *H. ducreyi* on the upper arm until they develop pustules. Here, we characterized host-pathogen interactions in pustules using transcriptomics and metabolomics and examined interactions between the host transcriptome and metabolome using integrated omics.

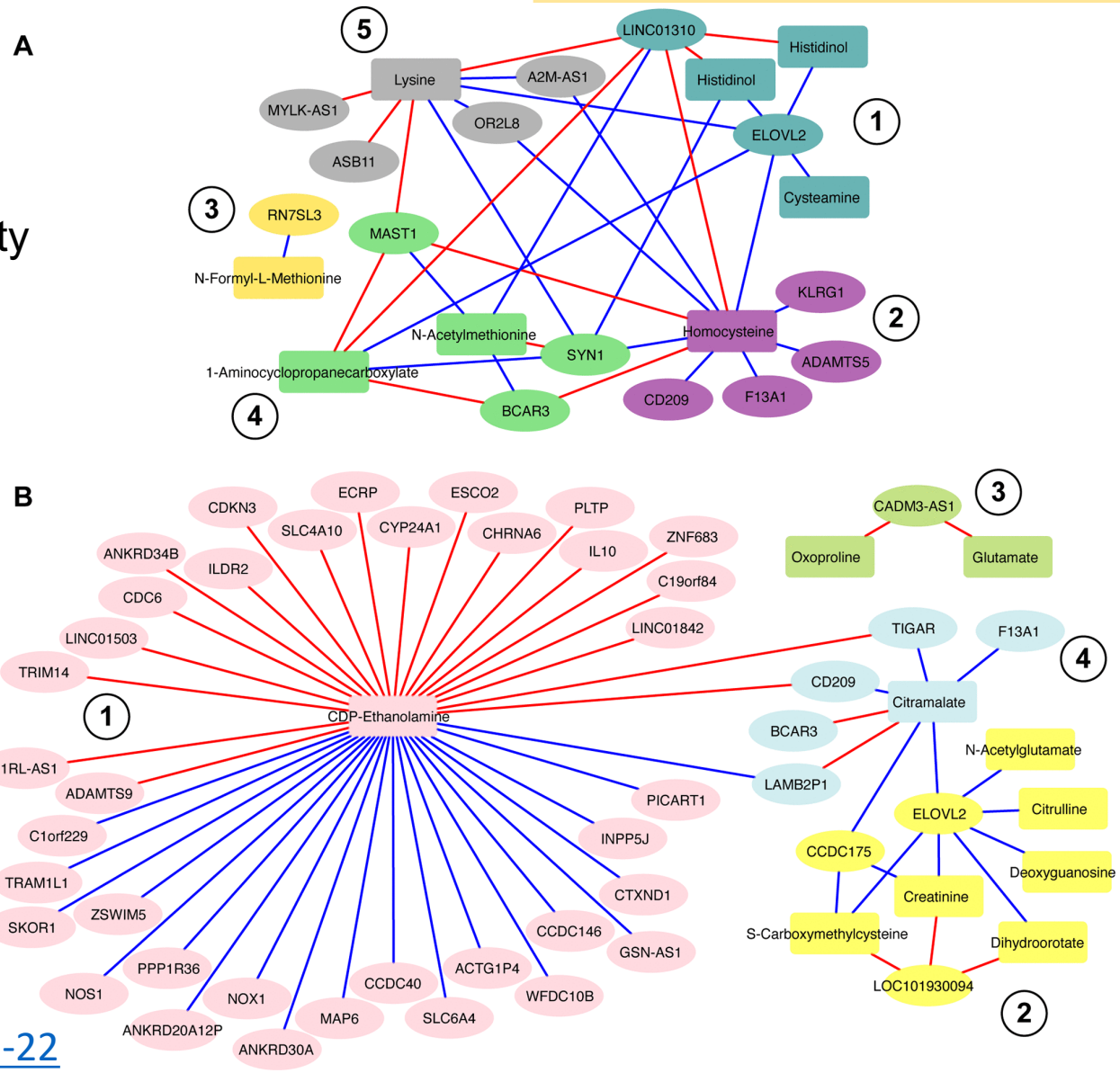
Metabolome and human transcriptome interaction networks

PLS regression; correlation
$P < 0.05$
$|\rho|>0.75$

Each color represents a community

☐ Metabolites (A: positive ions; B: negative ions)

○ Host transcripts



DOI: https://doi.org/10.1128/mbio.03125-22

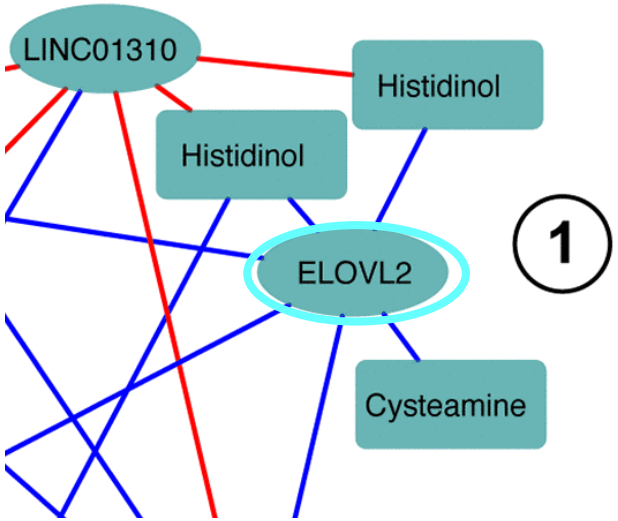
ELOVL2: a key regulator in inflammation

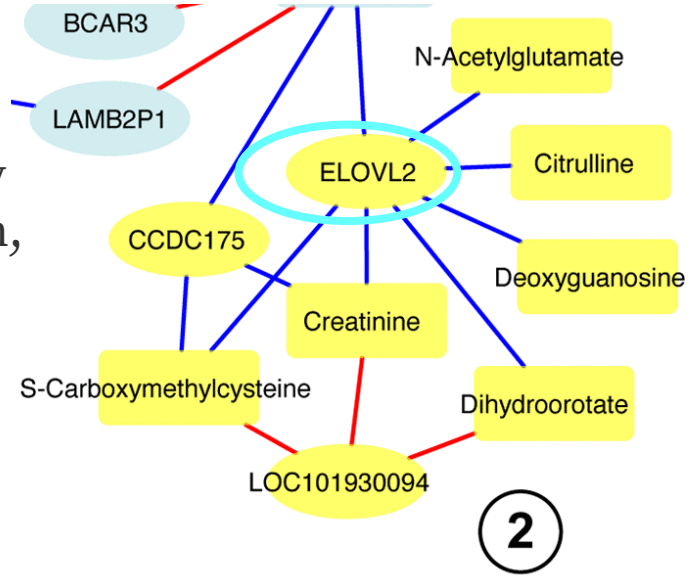
PLS regression; correlation
$P < 0.05$
$|\rho|>0.75$

Each color represents a community

□ Metabolites (A: positive ions; B: negative ions)

○ Host transcripts





ELOVL2, which elongates very long polyunsaturated fatty acids, is correlated with changes in fatty acid metabolism, and anti-inflammatory metabolites.

DOI: https://doi.org/10.1128/mbio.03125-22

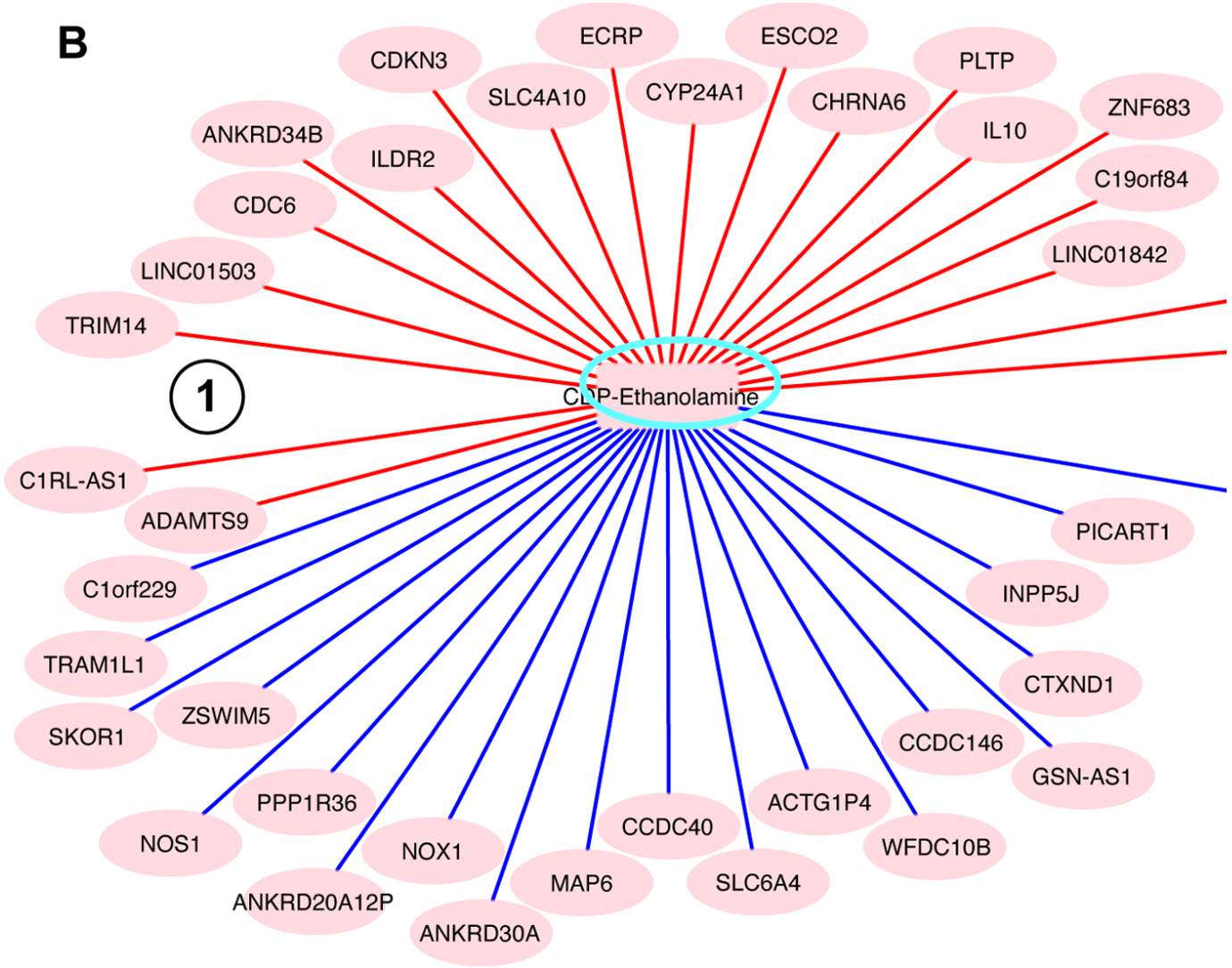
The role of lipid synthesis in infection

PLS regression; correlation
$P < 0.05$
$|\rho|>0.75$

Each color represents a community

□ Metabolites (A: positive ions; B: negative ions)

○ Host transcripts



CDP-ethanolamine, used to synthesize lipid phosphoethanolamine. The abundance of CDP-ethanolamine is associated with cell growth and skin repair, whereas it is negatively associated with the immune response and neuronal function

DOI: https://doi.org/10.1128/mbio.03125-22

xMWAS Integration: Pros & Cons

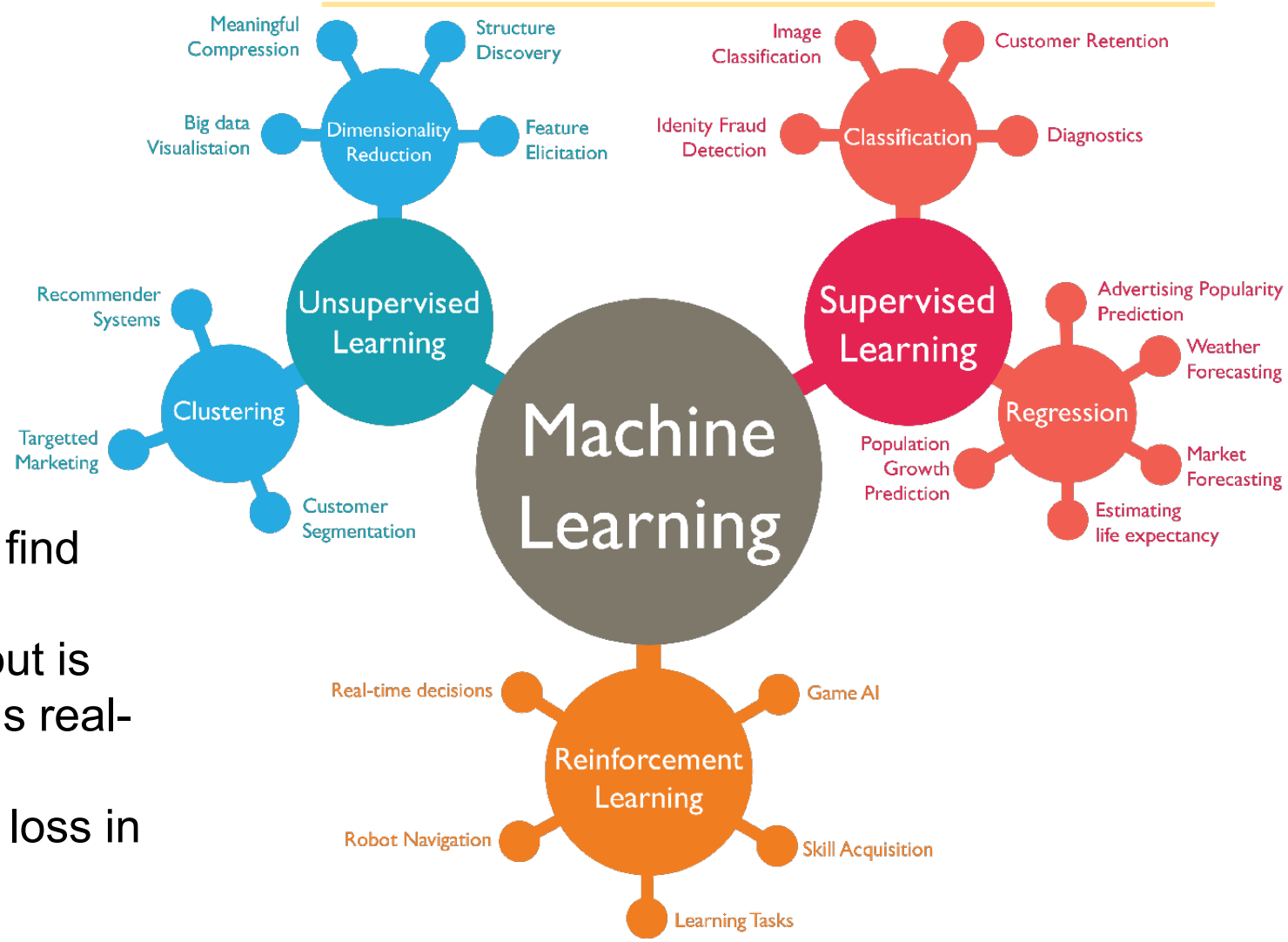
Advantages

- Do not rely on prior knowledge defined pathways and database.
- Can be used in relatively new field.
- No bias toward certain pathways, gene sets or diseases.
- Can provide direct information on interactions and discover new biological mechanisms.

Limitations

- Require collection of multi-omics data on the same or very similar subjects.
- Bipartite - Does not provide intra-correlations within the same layer.
- Does not consider directionality of interactions, weight of interactions or feed-back loops.

Machine Learning Methods for Network Reconstruction

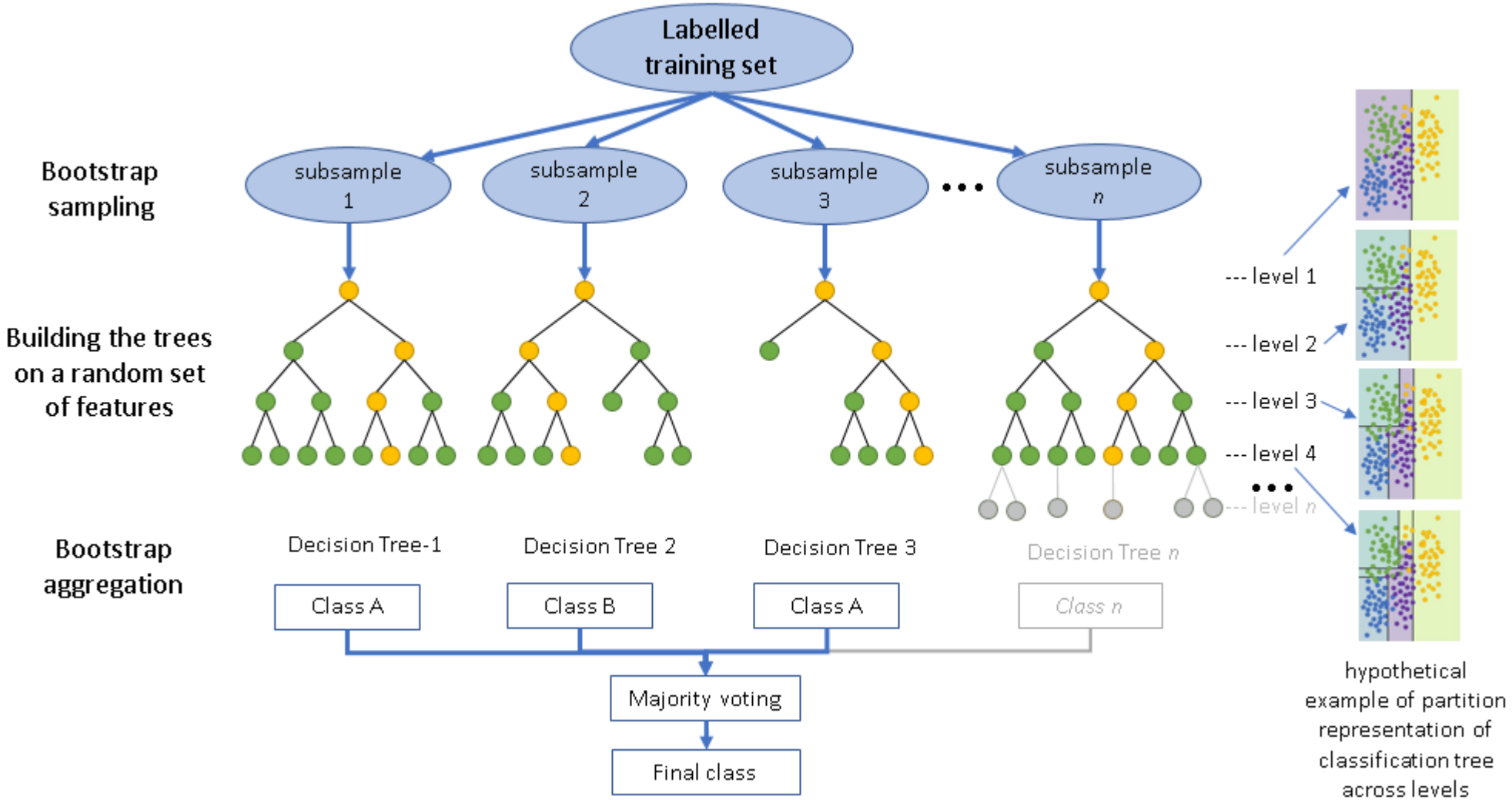


Unsupervised Machine Learning (ML) to find clusters.
Supervised ML for classification if the output is categorical, or regression when the output is real-valued.
Reinforcement ML learns from the gain or loss in the output result through trial-and-error interactions with a dynamic environment.

Common Examples of ML Algorithms in Data Integration

Name	Function	Reference
Support vector machine (SVM)	Creates a linear hyperplane, maintaining the largest possible distance between different classes of example data points	Boser et al., 1992, Guyon et al., 1993, Vapnik et al., 1997
Random forest (RF)	Composed of many decision trees. Each tree is grown using a training set and a random vector and works as a classifier. Each tree votes for the most popular class, and the most voted class is chosen	Breiman (2001)
Autoencoders	Consists of an encoder and a decoder. The encoder extracts features from large input data, and the decoder tries to construct an output very similar to the input using only the extracted features. In this way, it excludes the redundant data.	Murphy (2012)

Random Forest Classifier

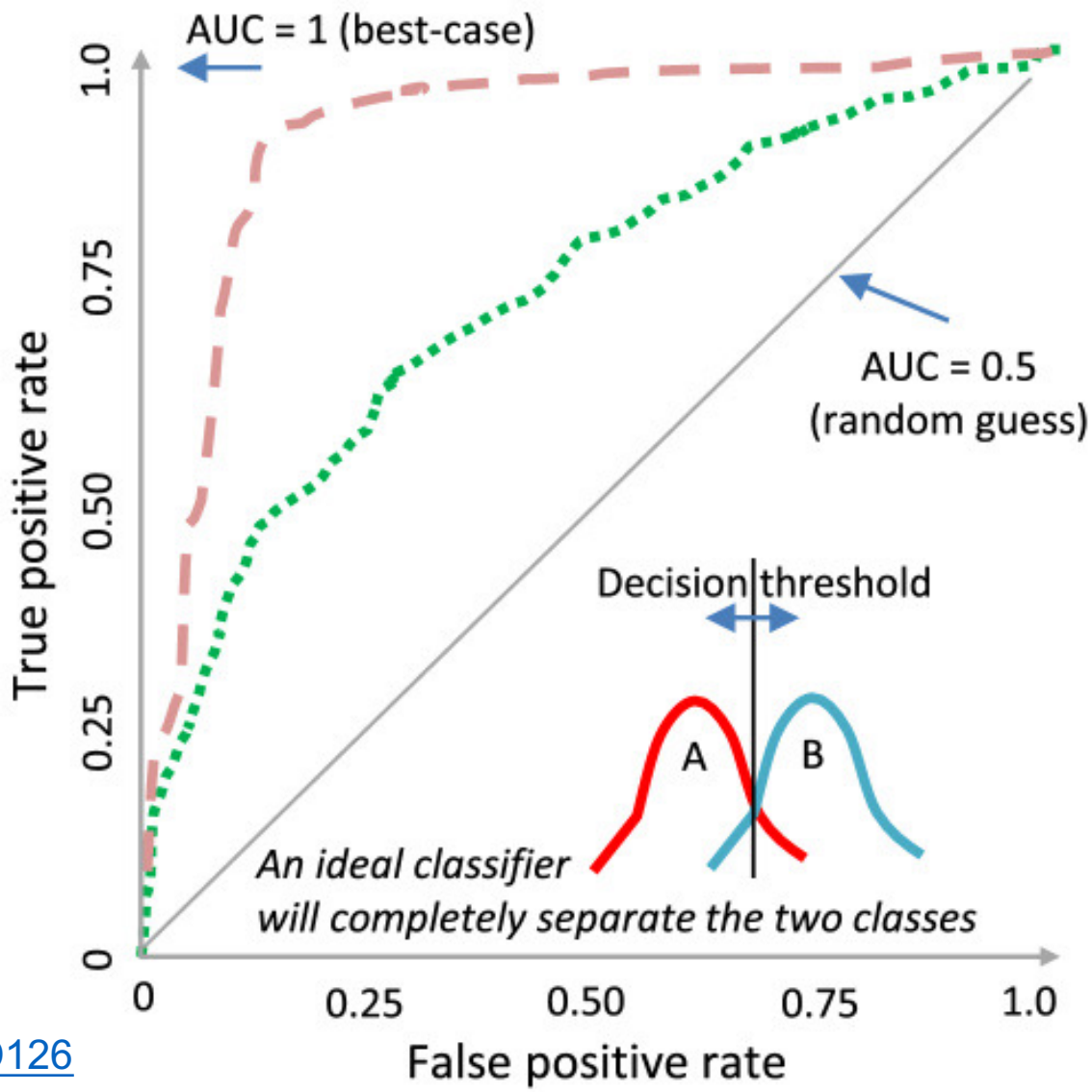


https://medium.com/nerd-for-tech/random-forest-sturdy-algorithm-d60b9f9140d4
https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Performance assessment

Confusion matrix

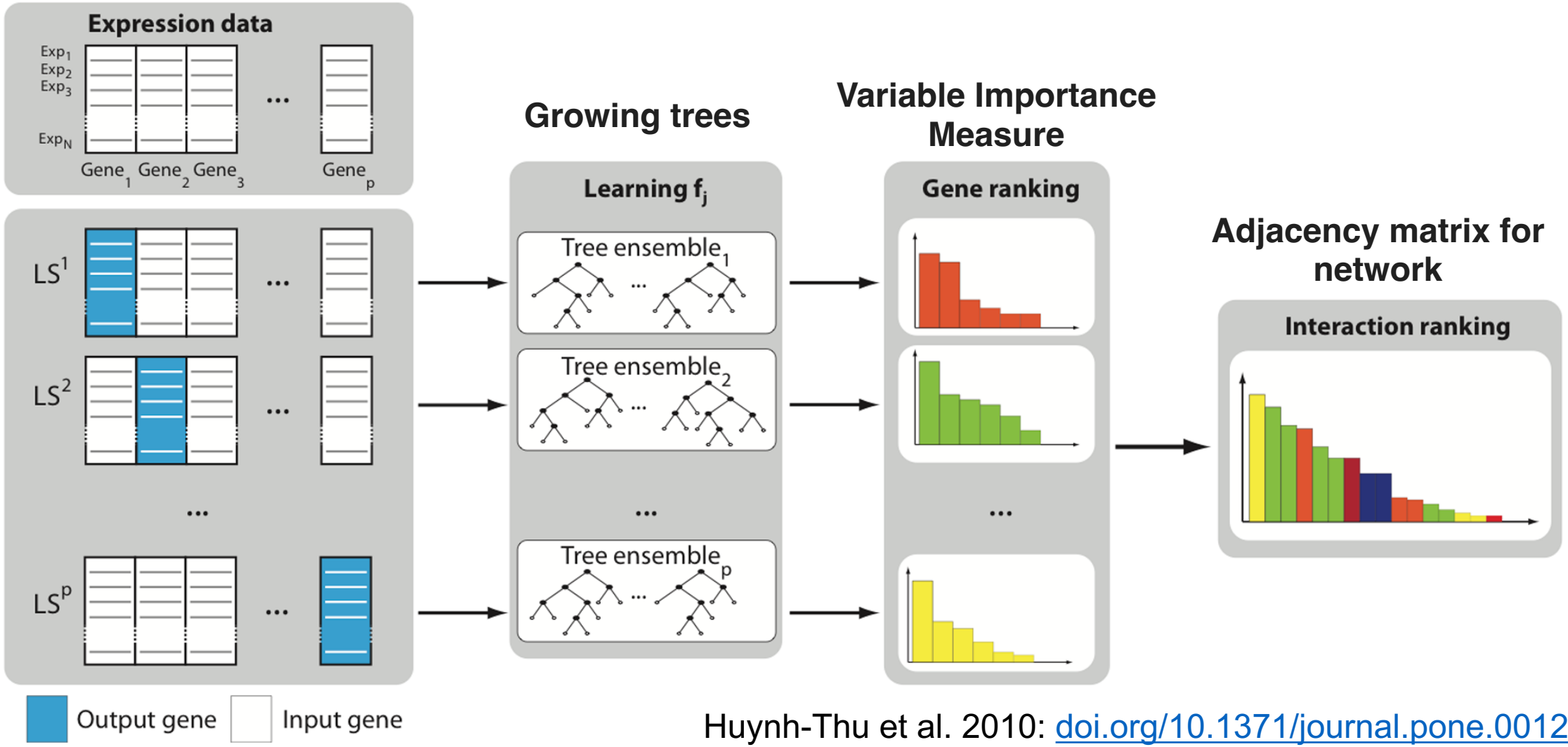
Total population	Positive by label	Negative by label
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative



Danish Rafique and Luis Velasco, 2018: doi.org/10.1364/JOCN.10.00D126

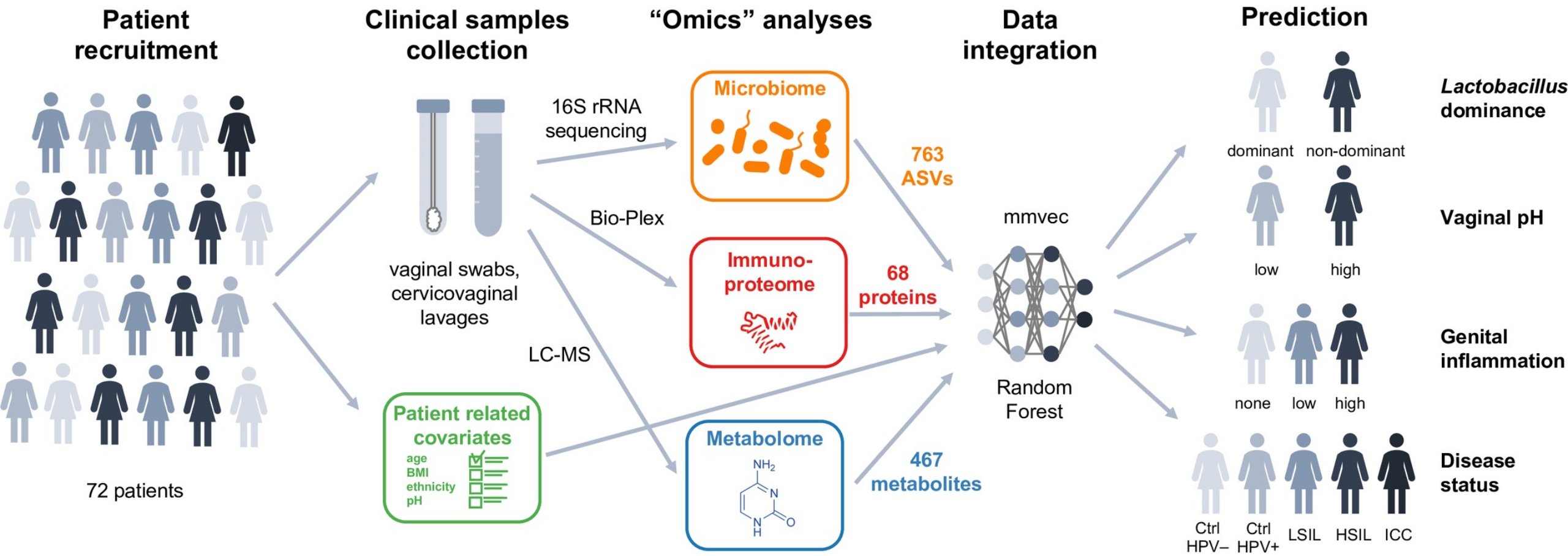
Tree-based Network Reconstruction

GENIE3: Assuming that the expression of each gene in a given condition is a function of the expression of the other genes in the network (plus some random noise) for reconstruction of Gene Regulatory Networks



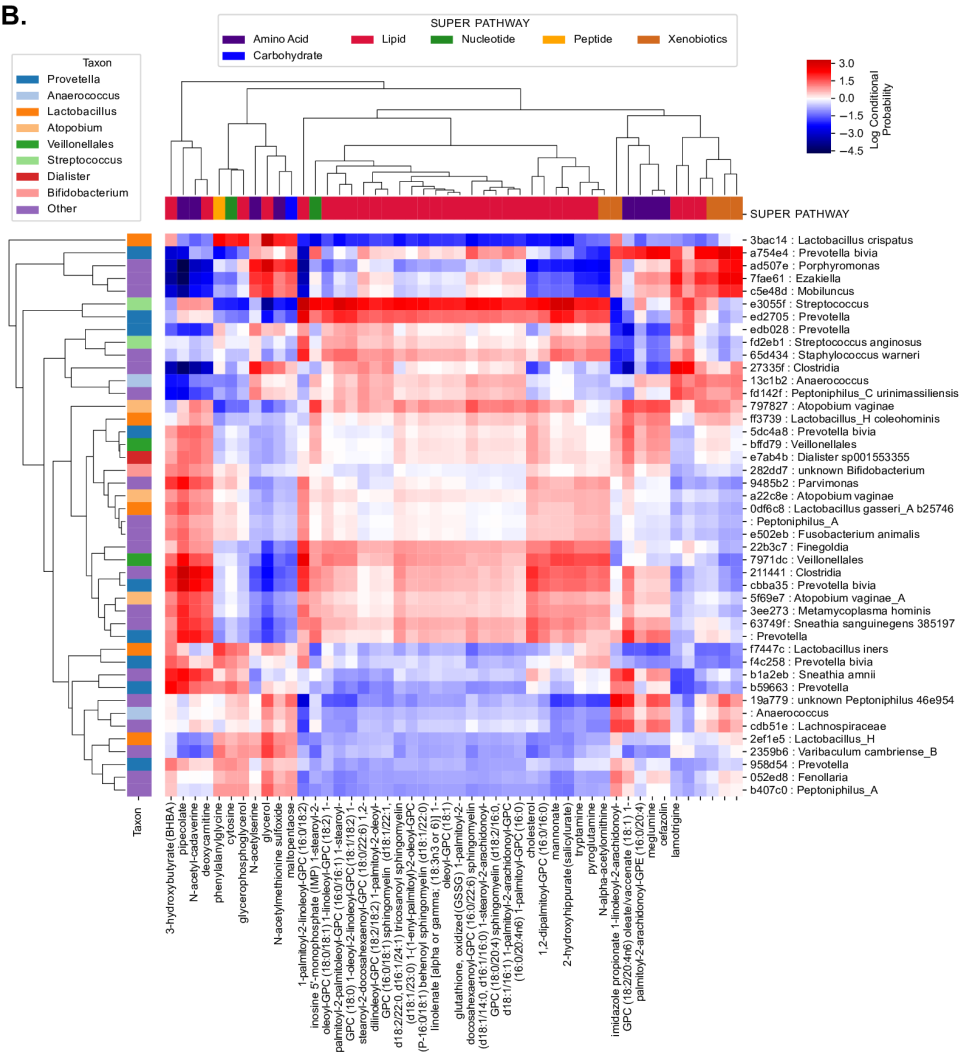
Huynh-Thu et al. 2010: doi.org/10.1371/journal.pone.0012776

Multi-omics data integration reveals metabolome as the top predictor of the cervicovaginal microenvironment



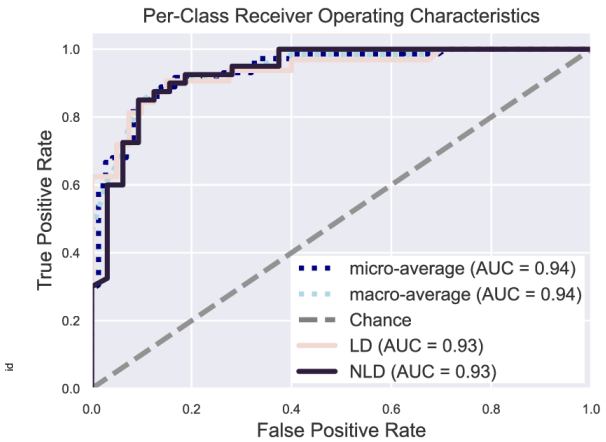
Bokulich et al. 2022 https://doi.org/10.1371/journal.pcbi.1009876

Multi-omics data integration reveals metabolome as the top predictor of the cervicovaginal microenvironment

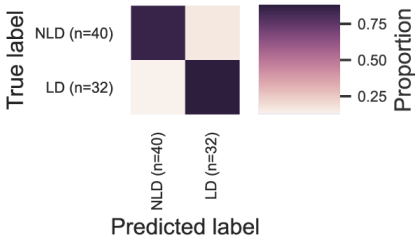


Lactobacillus dominance

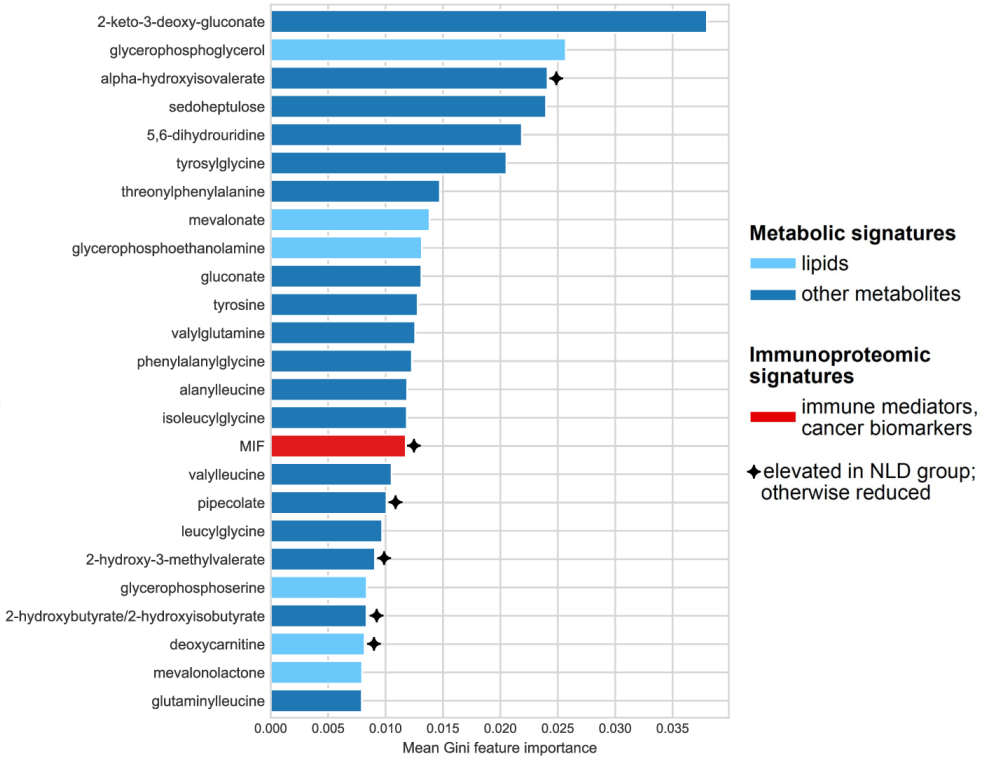
A. Receiver operating curve



B. Confusion matrix



C. Most predictive features



Bokulich et al. 2022 https://doi.org/10.1371/journal.pcbi.1009876

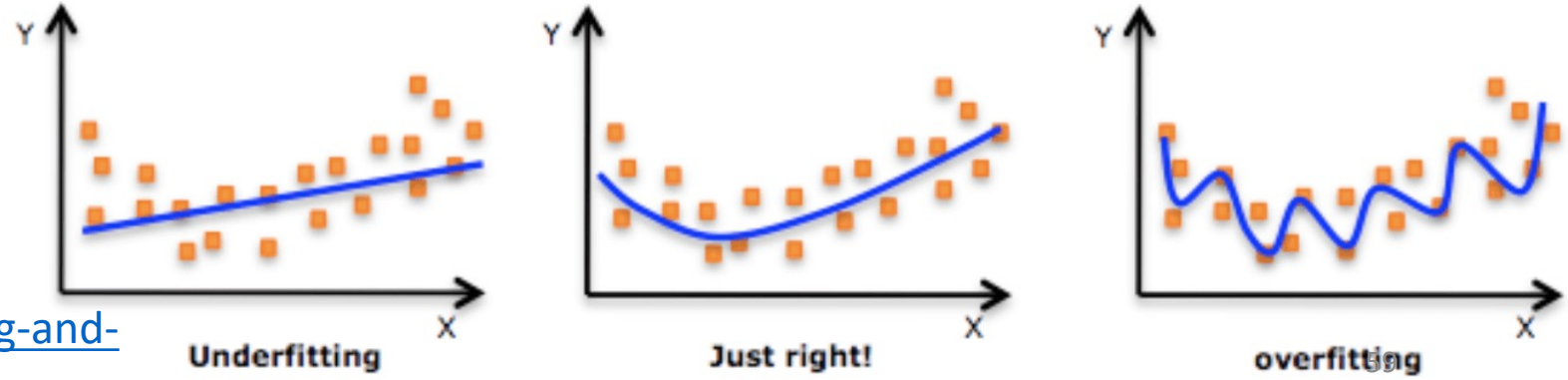
Overfitting and Underfitting

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively **impacts** the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.
More likely with nonparametric and nonlinear models that **have more flexibility** when learning a target function.

Solution:

1. Prune a decision tree after it has learned in order to remove some of the detail it has picked up or set parameters or techniques to limit and constrain how much detail the model learns.
2. Use matrices of co-expression network modules rather than all the data.
3. Use a resampling technique (e.g. k-fold cross validation) to estimate model accuracy
4. Use a validation dataset.

Underfitting: a model that can neither model the training data nor generalize to new data.



https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

ML Data Integration: Pros & Cons

Advantages

- Can deal with very large, high-dimensional and heterogenous datasets.
- Useful in non-linear complex predictive analysis.
- No human intervention needed (automation).
- Do not require existing knowledge.
- Reveals novel hypothesis.
- Can be improved over time.

Limitations

- High computational demand.
- Prone to overfitting.

Summary and Conclusions

An overview of general omics data types and study design is discussed. Various tools and techniques are available for integrating and visualizing multi –omics data.

High dimensional data with collinearity and missing values are typical challenges for omics integration. Each statistical strategy has advantages and limitations; choose based on your research questions.

The data-driven integration approach based on network theory is useful in studying the global behavior of the systems and can lead to discovery of new biological connections.

THANK YOU!

QUESTIONS:
XIN.HU2@EMORY.EDU

AI Based Multi OMICS Integration Design

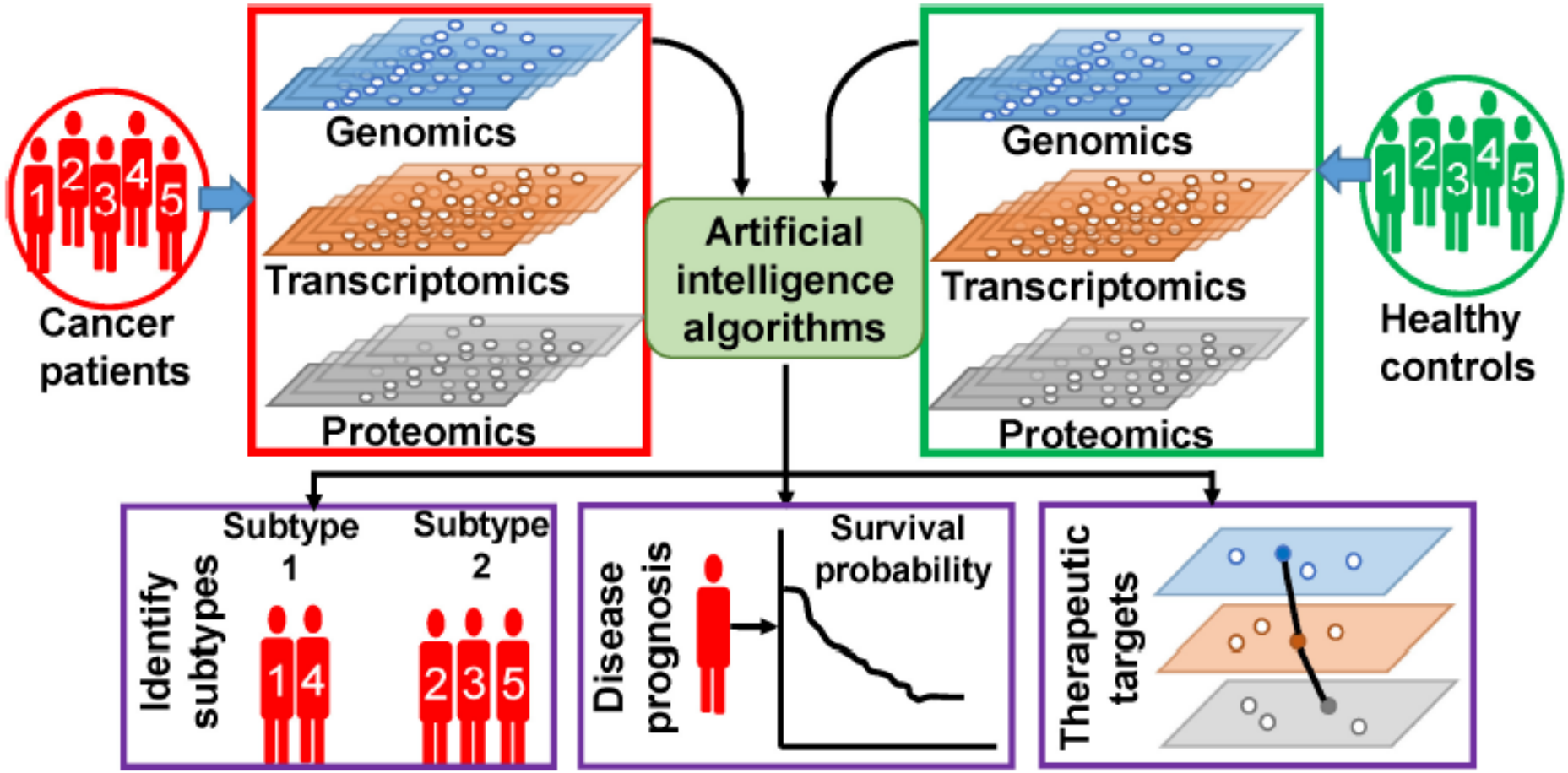


FIGURE 1 | Artificial intelligence (AI)-based analysis of multi-omics data. Different types of omics data are integrated and analyzed by AI algorithms to extract patient-specific information.

Biswas and Saikat Chakrabarti 2020 Frontiers in Oncology

AI Based Data Integration: Pros & Cons

Advantages

- Can deal with very large, high-dimensional and heterogenous datasets.
- Useful in non-linear complex predictive analysis.
- No human intervention needed (automation).
- Do not require existing knowledge.
- Reveals novel hypothesis.
- Can be improved over time.

Limitations

- High computational demand.
- Prone to overfitting.

"...is well established and truly powerful, but it cannot be used as a black-box."

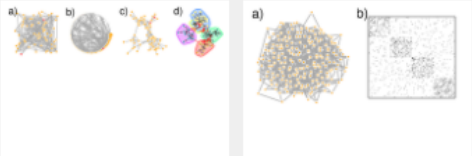
Nine quick tips for analyzing network data

Vincent Miele, Catherine Matias, Stéphane Robin, Stéphane Dray

Published: December 19, 2019 • https://doi.org/10.1371/journal.pcbi.1007434

Article | Authors | Metrics | Comments | Media Coverage

Figures



Citation: Miele V, Matias C, Robin S, Dray S (2019) Nine quick tips for analyzing network data. PLoS Comput Biol 15(12): e1007434. https://doi.org/10.1371/journal.pcbi.1007434

Editor: Francis Ouellette, University of Toronto, CANADA

Published: December 19, 2019

Funding: Funding was provided by the French National Center for Scientific Research (CNRS) to SD, CM, and VM; the French National Institute for Agricultural Research (INRA) to SR and the French National Research Agency (ANR) grant ANR-18-CE02-0010-01 EcoNet to SD, CM, VM, and SR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Modules: topological modules that represent highly interlinked local regions in the network.

Intra-modular hubs (blue nodes) mostly connect nodes within the same module and have relatively short connection distances; characterized by the PLS1. Intra-modular hubs (red nodes) have a more diverse connectivity profile with connections extending long distances and connecting nodes from different modules; characterized by the PLS2. Size and color saturation of the nodes in the connectome corresponds to the regional scores on PLS1 (Intra-modular hub) and PLS2 (Inter-modular hub) to represent the spatial pattern of transcriptional profiles [adapted and modified from (Vértes et al., 2016)]