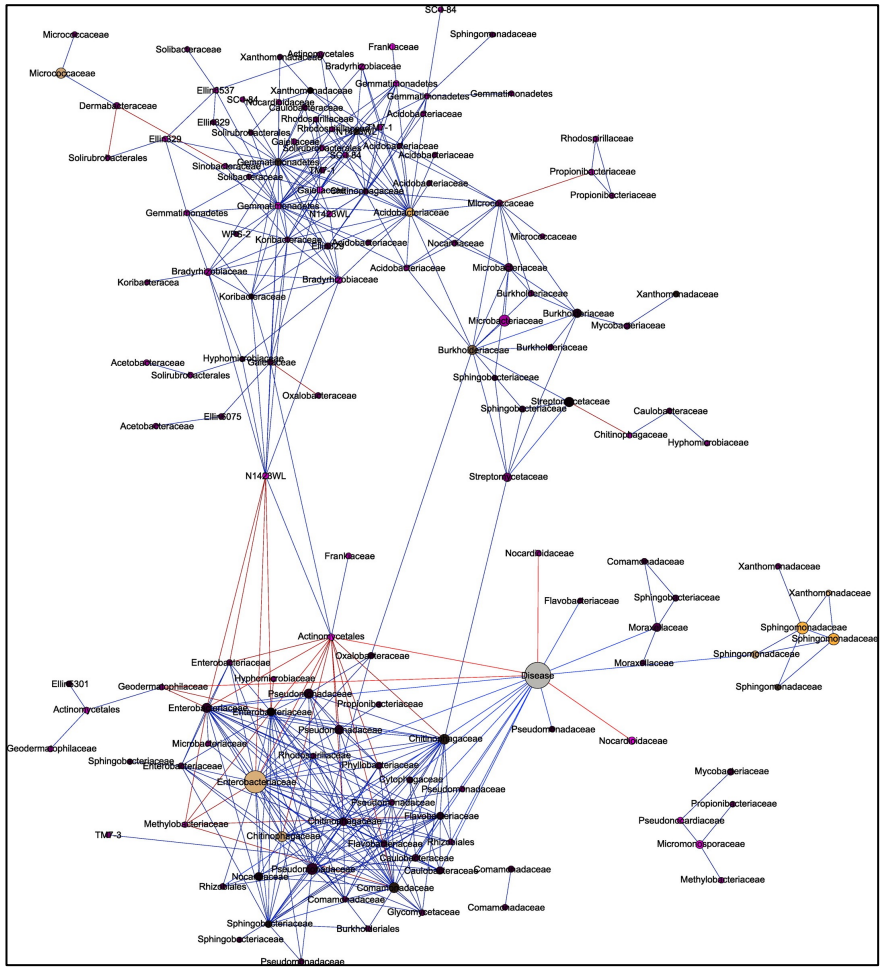
Microbiome network in R

Ashish Adhikari, PhD
adhikariashish@ufl.edu

@Garrettlab

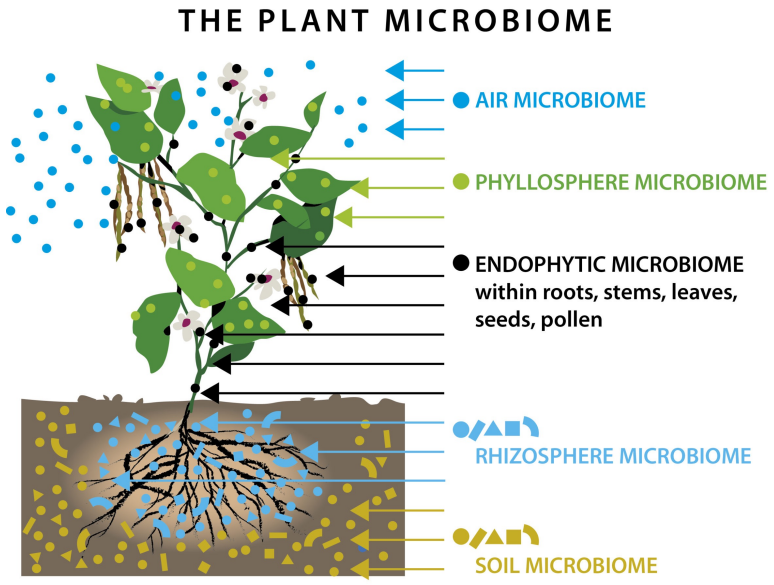
LEARNING GOALS FOR TODAY

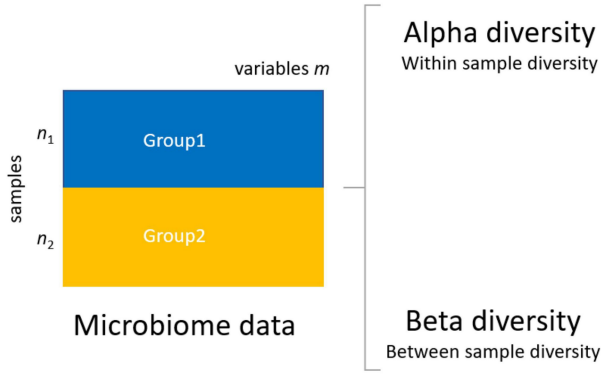
1. Understand and interpret microbiome networks
2. Understand features and algorithm used
3. Usefulness in context of plant pathology
4. Build microbiome network in R



Microbiome

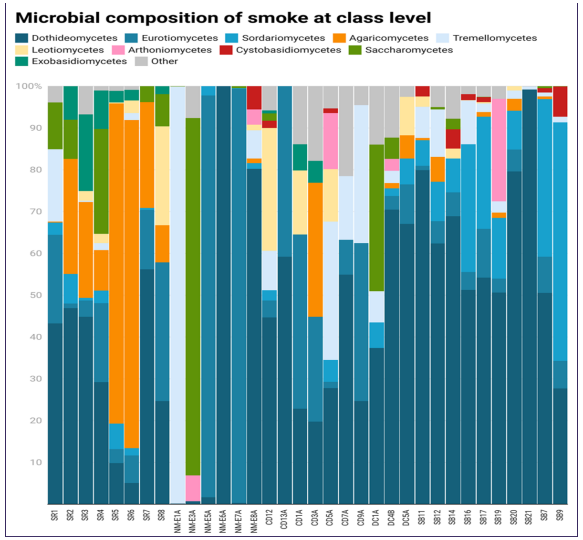
- Microbiomes are complex microbial communities in a system
- Structure and function are heavily influenced by microbe–microbe and microbe–host interactions





Gut microbiome
Phyllosphere microbiome
Soil microbiome
Wildfire smoke microbiome

"Who's who"



Zhao et al in Preparation

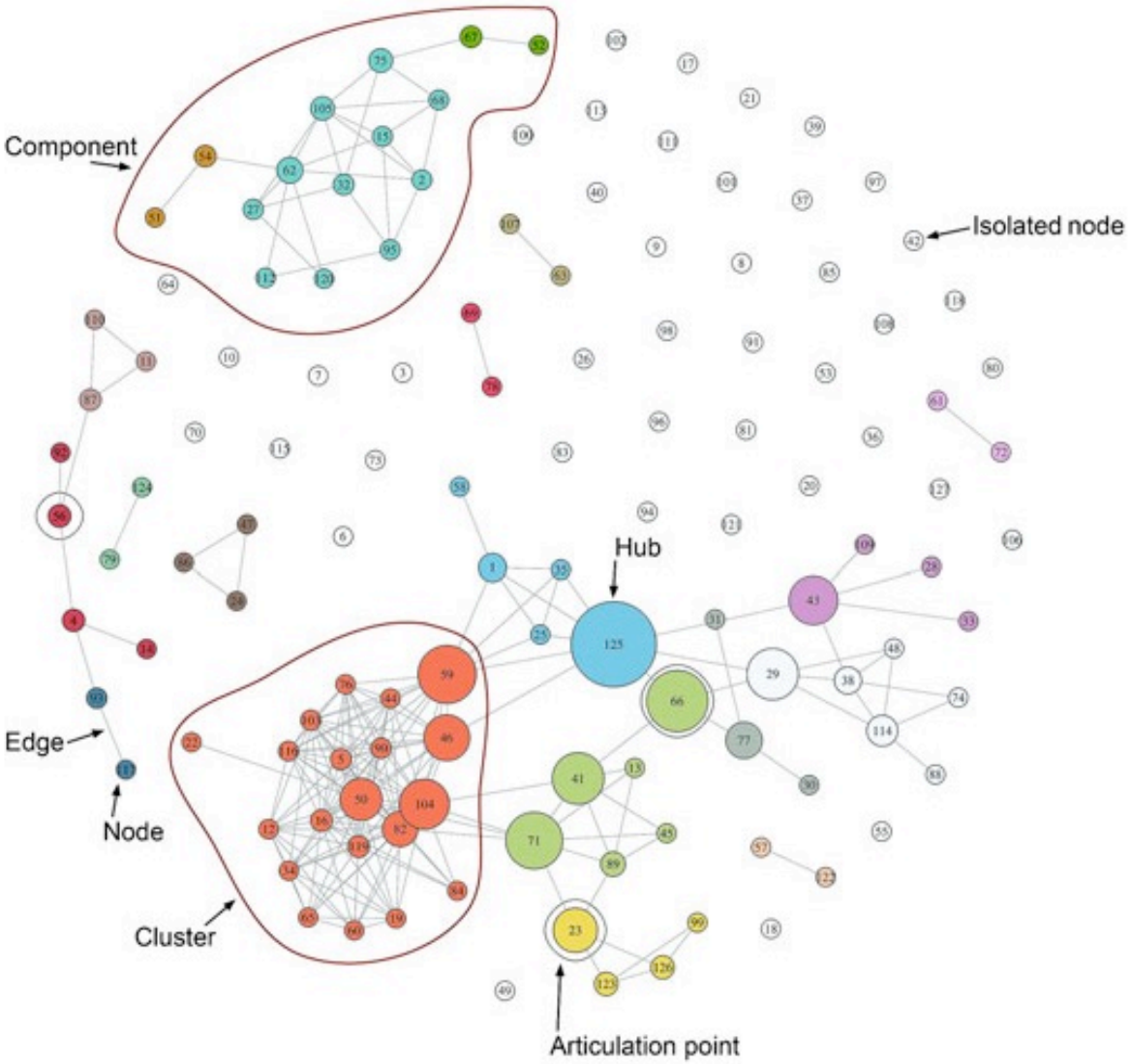


Figure adopted from Layeghifard et al. 2018

Nodes/Vertex: the building blocks/ OTUs/microbes

Edges/Links: the connections between two nodes within the network. (correlation of relative abundances/ or a direct interaction, such as nutrient competition, antagonism)

Module: a group of highly interconnected nodes with limited connections to nodes outside of the group

Complexity: the average links per node in the network

Modularity: a metric to summarize the number and isolation of modules within a network

Degrees: the number of edges connected to a particular node

Closeness centrality: the mean distance of a node to other nodes in the network

Betweenness centrality: indexes the degree to which an individual node connects to other nodes; the number of times a node lies on the shortest path between two other nodes in the network

Modular hubs: nodes that have high connectivity to other nodes within the same module

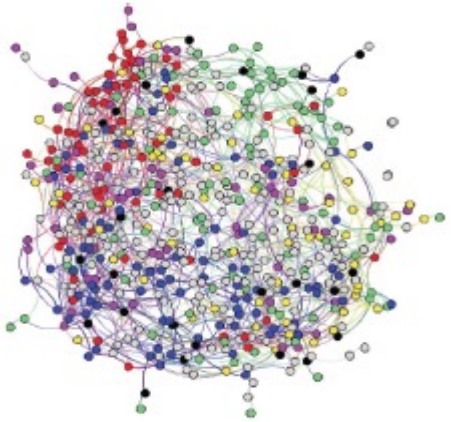
Biological and analytical questions??

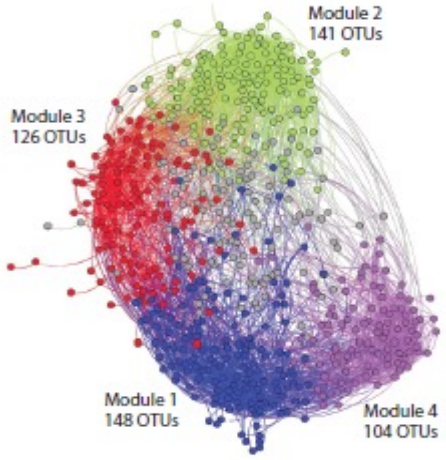
How are microbial taxa organized within plant-associated microbiomes?

How do network roles of microbes vary within and across microbiomes?

Why are specific microbes or collection of microbes coassociated?

How do coassociations among microbes influence microbiome functionality, ecology, or evolution?





Network structure	Nodes and connections	Modules and ecological clusters
1. Do networks provide evidence for ecological or evolutionary processes, or do they suggest random associations among microbes?	1. How does the total connectivity, or the distribution of node connectivities, vary among microbiomes?	1. What are the taxonomic or functional relationships between microbes within a module?
2. Are two (or more) networks substantially different from one another, and in what ways?	2. How connected is an individual taxon to other nodes, and does this vary among microbiomes?	2. How do the numbers, diversity, and/or composition of modules vary among microbiomes?
	3. Are connections between specific microbial taxa conserved across microbiomes?	3. Are diversity or composition of modules, or the abundances of taxa within modules, related to habitat, plant health, or plant/ecosystem productivity?

Bakker et al. 2014

A. Dissimilarity-Based Methods

- Simplest, based on distance to construct co-occurance network
- pairwise dissimilarity index such as Bray–Curtis or Kullback–Leibler

B. Correlation-Based Methods

- most popular
- significant pairwise associations between nodes using a correlation coefficient such as Pearson's correlation coefficient or Spearman's nonparametric rank correlation coefficient

Limitations: suffers from limitations such as detecting spurious correlations among low-abundance OTUs

C. Regression-Based Methods

- Use multiple regression analysis to infer the abundance of one species from the combined abundances of other taxa
- interpretation of regression results can be more difficult

Limitation: methods suffer from overfitting that increases with the number of predictor variables, and is associated with increase in the number of false positives.

D. Probabilistic Graphical Models

- Probability Graph Models (PGMs) use graphs to help us understand how things are connected and affect each other. These models are like maps showing relationships between different factors.
- Uses probability theory and graph theory to deal with uncertainly and complexity
- High accuracy and minimal bias.

Types of PGMs:

•**Directed vs. Undirected:** Directed graphs show one-way relationships (like parent to child), while undirected graphs show mutual relationships (like friendships where both people are friends with each other).

•**Static vs. Dynamic:** Static graphs show connections at one point in time, while dynamic graphs show how relationships change over time.

•**Probabilistic vs. Decisional:** Probabilistic models deal with uncertainty (like weather predictions), while decisional models consider decisions people make (like choosing the best route to drive).

D. Probabilistic Graphical Models

- Probability Graph Models (PGMs) use graphs to help us understand how things are connected and affect each other. These models are like maps showing relationships between different factors.
- Uses probability theory and graph theory to deal with uncertainly and complexity
- High accuracy and minimal bias.

Microbiome Networks Example: Imagine studying bacteria in your gut. Each type of bacteria is a variable. If we draw lines (edges) between bacteria, showing their relationships, we create a graph. In this graph:

- **Undirected Links:** If two bacteria are connected, it means they influence each other equally.

- **Time Series Data:** If we study changes over time, some bacteria changes may lead or follow others.

D. Probabilistic Graphical Models

- Probability Graph Models (PGMs) use graphs to help us understand how things are connected and affect each other. These models are like maps showing relationships between different factors.
- Uses probability theory and graph theory to deal with uncertainly and complexity
- High accuracy and minimal bias.

•**Network construction using PGMs**

•**Correlation Graphs:** Graphs show how things are related. Classical tests like Pearson's correlation can be used, but they might miss some connections

•**Partial Correlation Graphs:** Reveal specific relationship between nodes when other influences are removed.

•**Conditional Independence Graphs:** Conditional independence means two things are unrelated when considering a third factor.

These graphs provide detailed and accurate information but are tough to make with limited data.

Network Inference Methods Robust to Compositionality Bias

Microbiome data usually suffer from two problematic features that confound their analysis. Firstly, OTU data are compositional; meaning that microbial counts are interdependent due to the normalization of counts to the total number of counts in the sample. This interdependence can lead to spurious results when using traditional statistical methods such as Pearson's correlation. Secondly, the ratio of observations (samples) to the number of variables (OTUs) is small. Recently, there have been many efforts to develop network construction methods robust to these two issues. These methods are described in Box 3.

The concerns over correlation-based analyses have led to the development of methods that are robust to compositionality.

SparCC (Sparse Correlations for Compositional data), for example, is a technique that uses linear Pearson's correlations between the log-transformed components to infer associations in compositional data.

SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference) is another statistical method for the inference of microbial ecological networks that combines data transformations developed for compositional data analysis with a graphical model inference framework with the assumption that the underlying ecological association network is sparse.

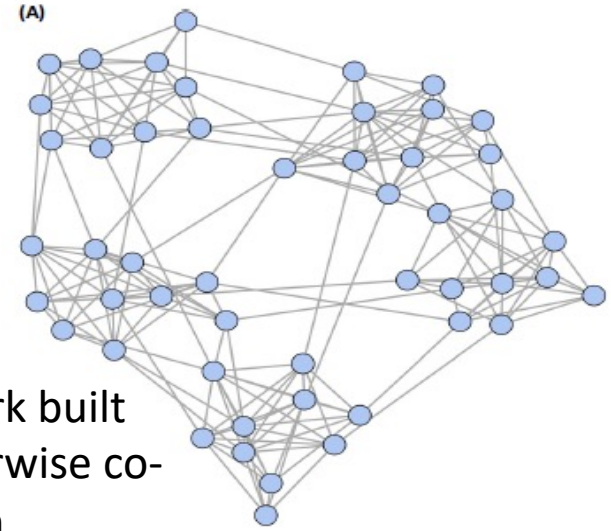
One method, EBICglasso,
estimates sparse undirected graphical models for continuous data with multivariate Gaussian distribution through the use of L1 (lasso) regularization before using an extended Bayesian information criteria (EBIC) to select the most fitting model.

Lasso regularization

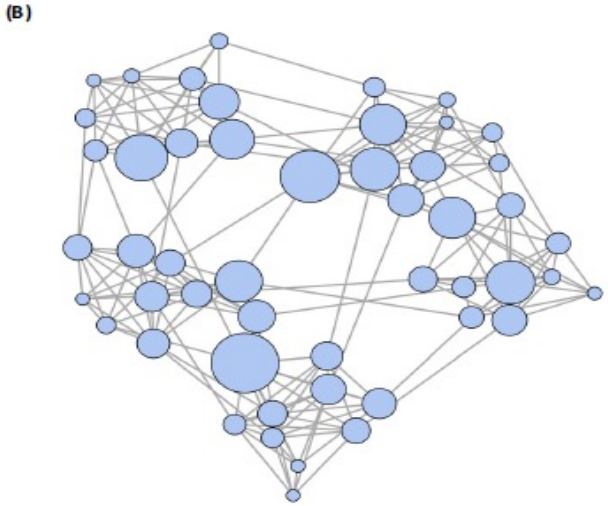
is a regression-analysis method that enforces a sparsity constraint on the data that can lead to simpler and more interpretable models less prone to overfitting. lasso essentially makes the data
smaller (i.e., sparse) and performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model by selecting only a subset of the provided covariates for use in the final model. Given a collection of graphical models for the data, information criteria enable us to estimate the relative quality of each model or tune parameters of penalization methods such as the graphical lasso.

Thus, EBIC provides a means for model selection and optimization.

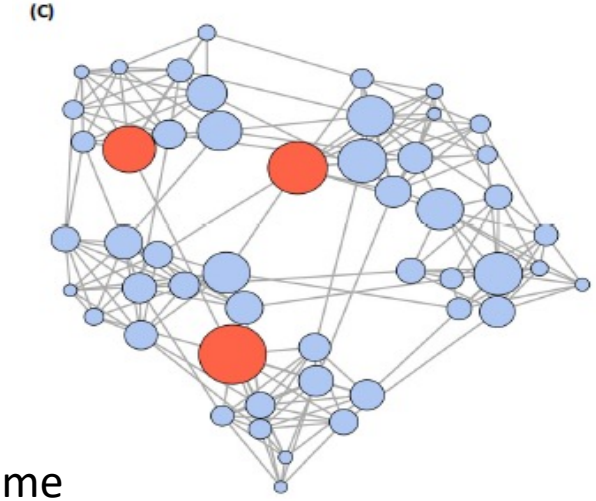
Microbiome Network Analysis



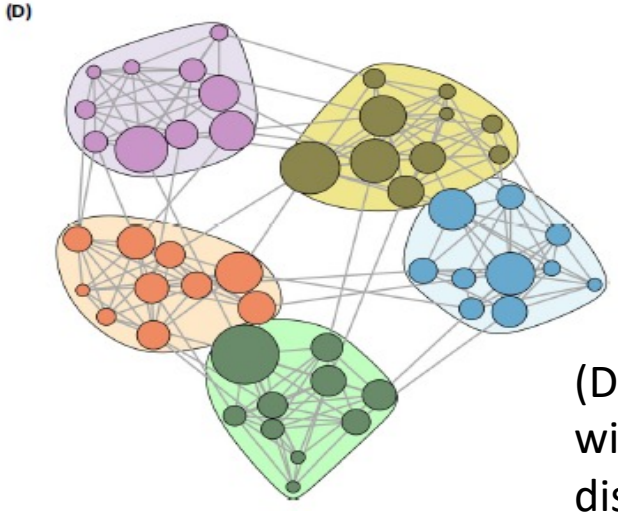
(A) A microbiome network built from an OTU table; a pairwise co-occurrence or interaction.



B) The same microbiome network with nodes' sizes proportionate to HITS scores computed for all the microbes.



(C) The same microbiome network with hub (keystone) species highlighted in red.



(D) The same microbiome network with microbes clustered into five distinct groups.

NETWORK ANALYSIS INSIGHTS INTO PLANT MICROBIOMES

1. Network Structure: Connectivity

- Simply put, greater network connectivity indicates consistent patterns of coexistence among taxa/ or mutual exclusive habitat preference.
- More stability

2. Network Structure: Modules

- A useful statistical means to focus on specific assemblages that may represent distinct, coselected, and plausibly functional ecological units.
- Habitat preference, similar function, ….

- Disease specific module/ Pathobiome

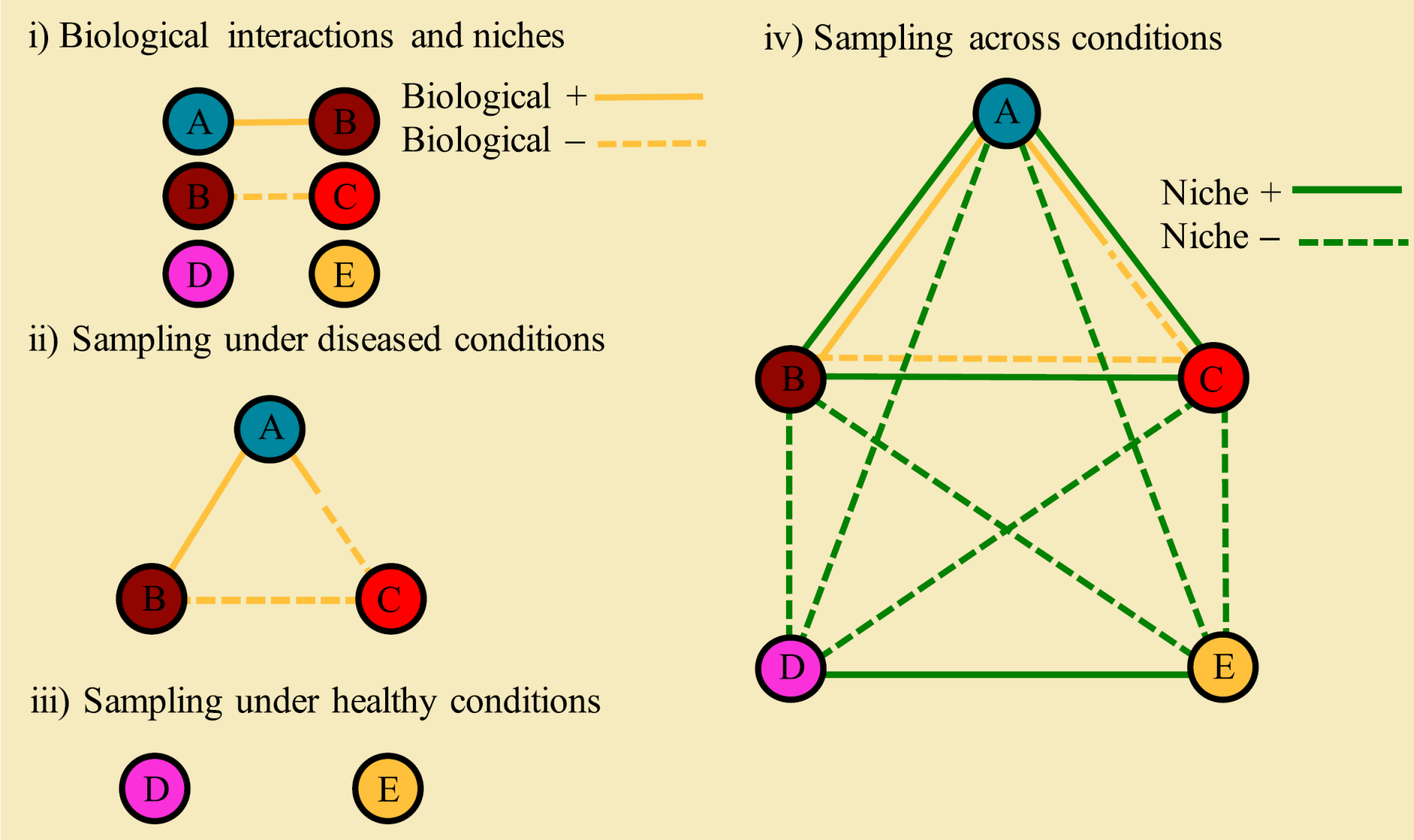
The pathobiome concept emphasizes the potential role of microbial assemblages and species interactions in mediating plant disease and disease symptoms, e.g., in both suppressing and enhancing disease development.

NETWORK ANALYSIS INSIGHTS INTO PLANT MICROBIOMES

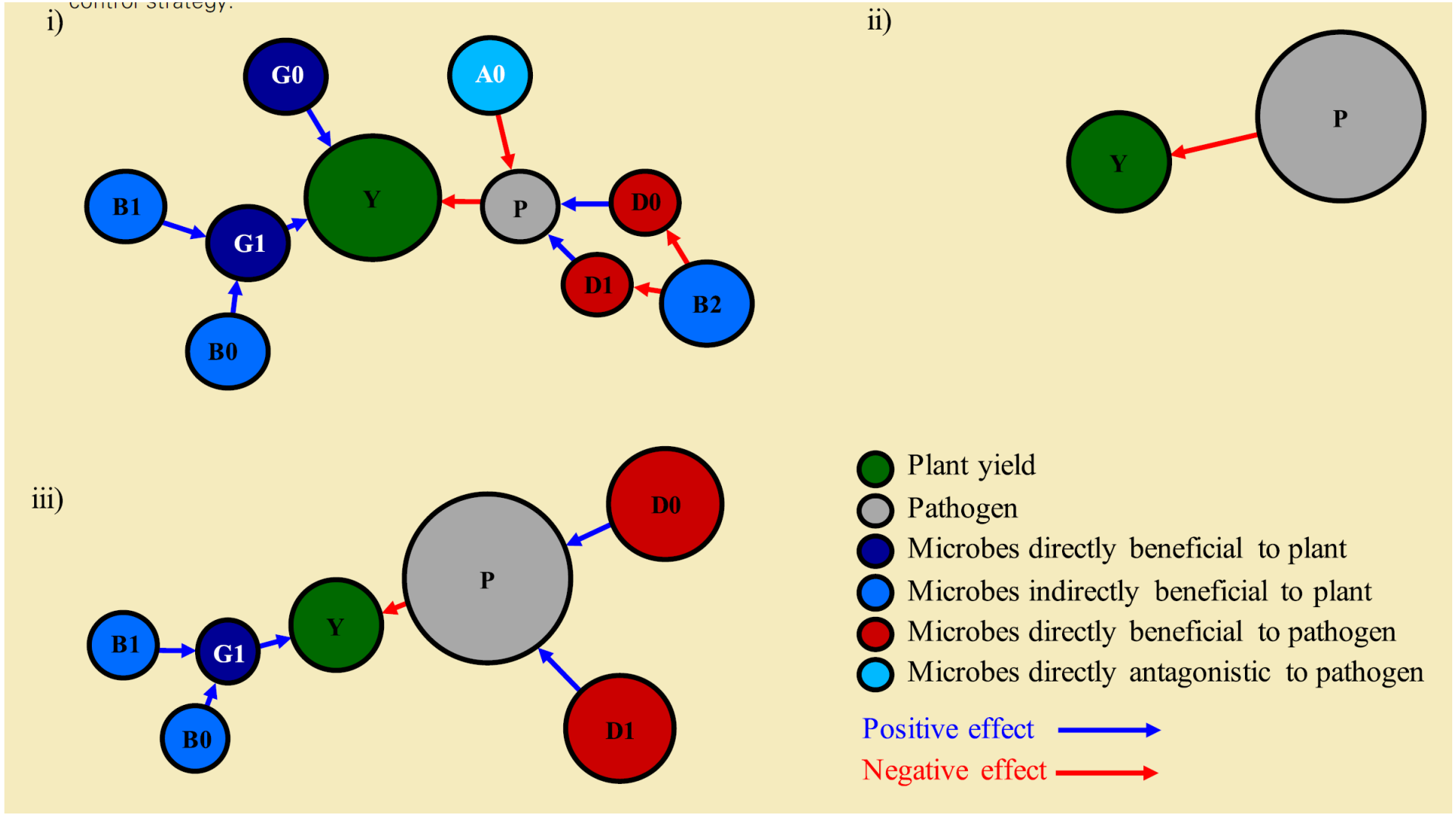
Network Hubs

- To identify an individual microbial taxa to have main role microbiome organization and functionality.
- Synthetic community: Top-Down/ Bottom-Up approaches

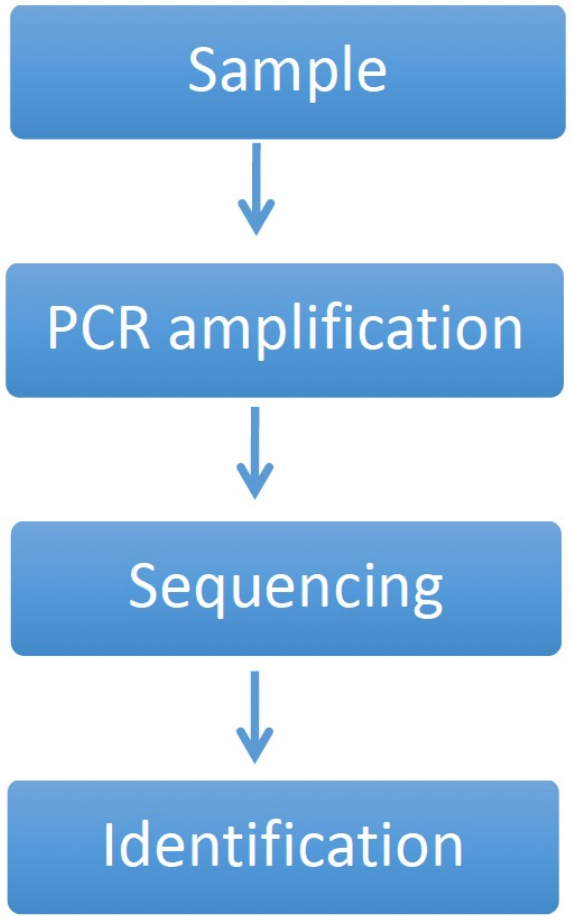
The effects of biological interactions and niche preference are combined in associations:

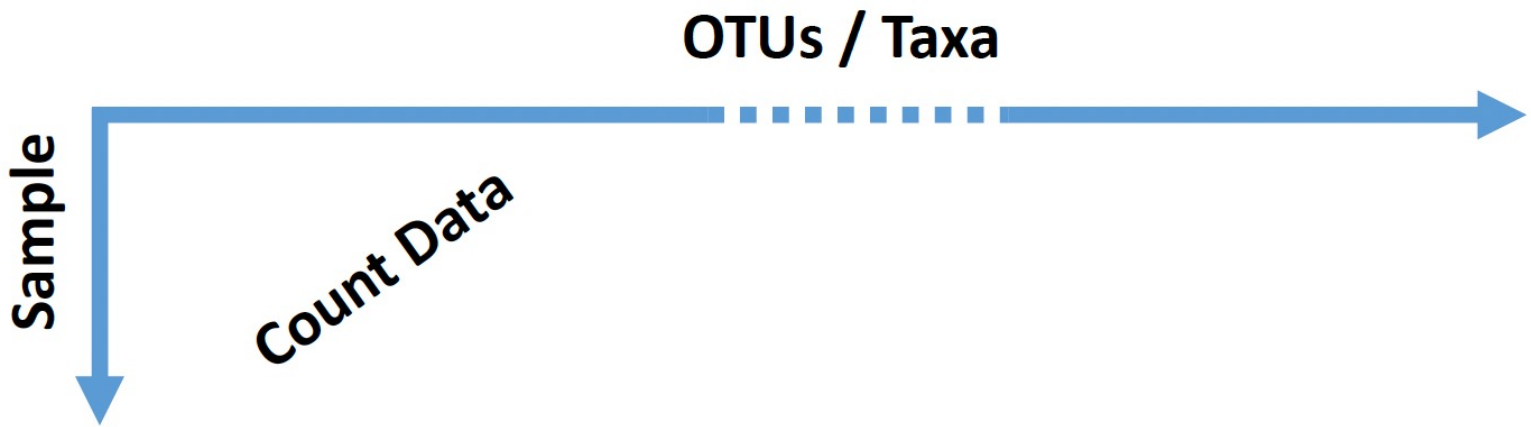


Examples of potential interactions



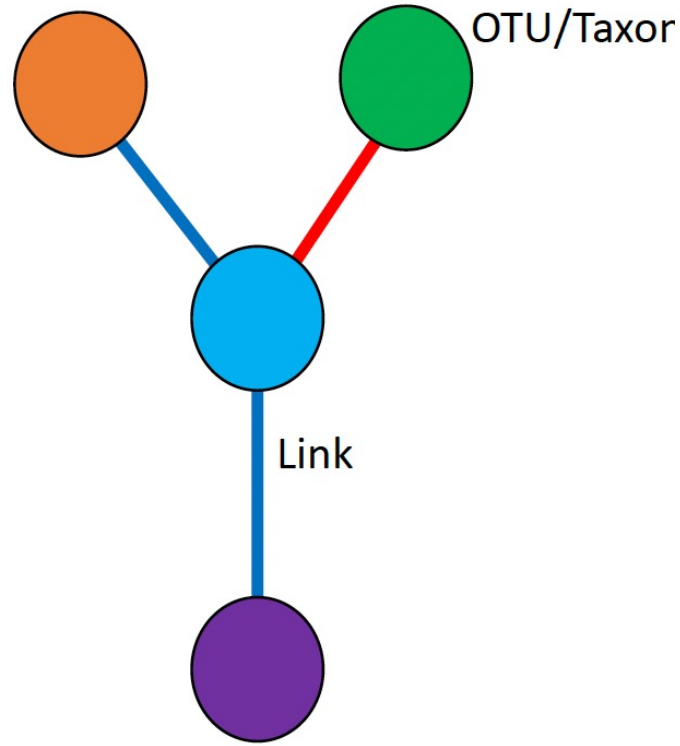
Example: 16S ribosomal RNA (rRNA) sequencing



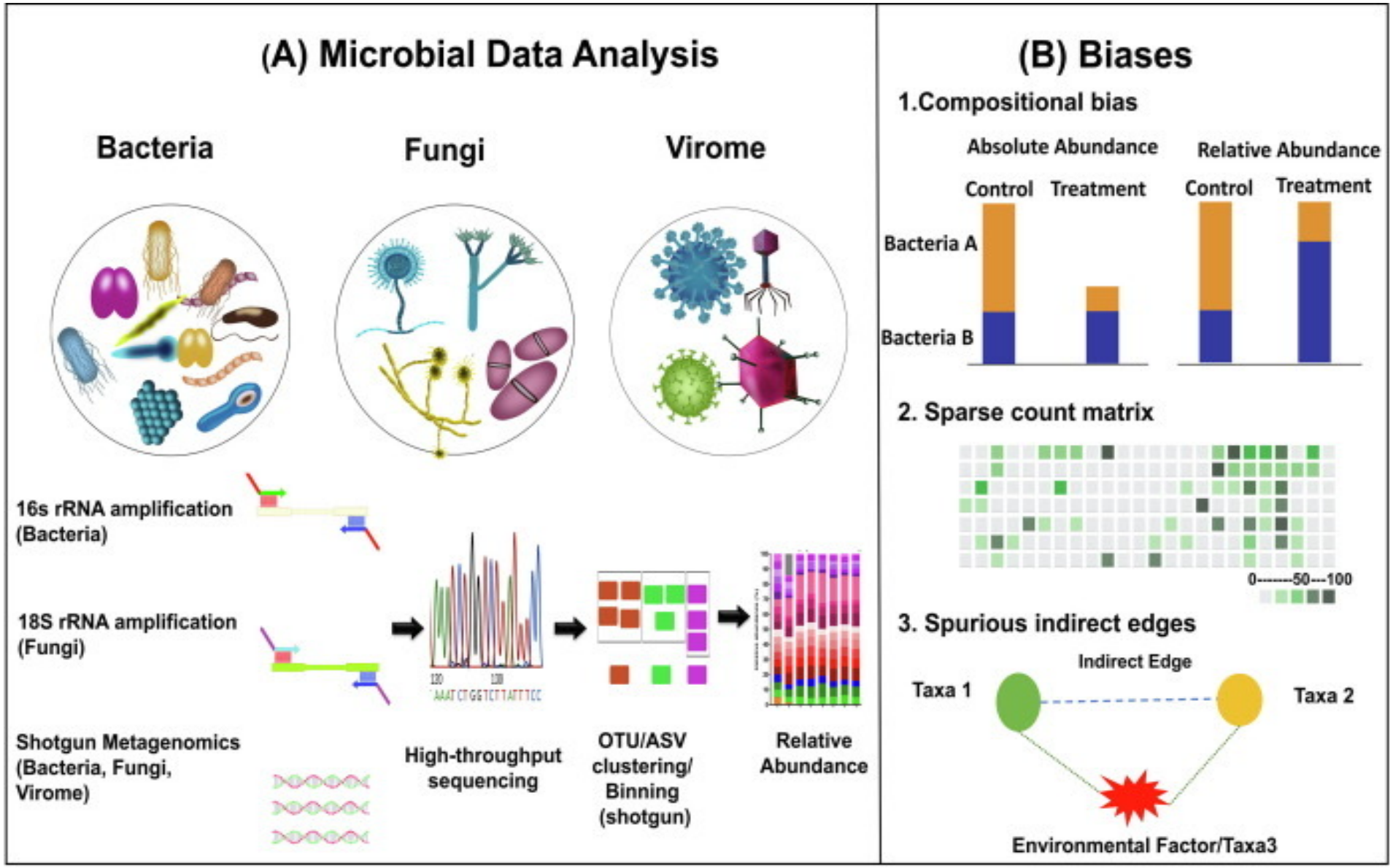


Case study- metagenomics and network models

- OTU/ taxon is represented as a node
- Links defines the relationship between two OTUs
- Various methods can be used to define the links
  - Presence or absence of association (co-occurrence)
  - Statistical approach to define the association
    - Such as correlation, proportionality
    - Methods robust for metagenomics data ~ deal with compositional bias and spurious correlation.
      - SparCC, SpiecEasi….. and so on…
      - Mostly, to reduce compositional bias associated with data type.



Three main challenges:



Adopted from Matchado et al. 2021

Matchado, M.S., Lauber, M., Reitmeier, S., Kacprowski, T., Baumbach, J., Haller, D. and List, M., 2021. Network analysis methods for studying microbial communities: A mini review. *Computational and structural biotechnology journal*, *19*, pp.2687-2698.

Poudel, R., Jumpponen, A., Schlatter, D.C., Paulitz, T.C., Gardener, B.M., Kinkel, L.L. and Garrett, K.A., 2016. Microbiome networks: a systems framework for identifying candidate microbial assemblages for disease management. *Phytopathology*, *106*(10), pp.1083-1096.

Layeghifard, M., Hwang, D.M. and Guttman, D.S., 2017. Disentangling interactions in the microbiome: a network perspective. *Trends in microbiology*, *25*(3), pp.217-228.

Bakker, M.G., Schlatter, D.C., Otto-Hanson, L. and Kinkel, L.L., 2014. Diffuse symbioses: roles of plant–plant, plant–microbe and microbe–microbe interactions in structuring the soil microbiome. *Molecular ecology*, *23*(6), pp.1571-1583.

Layeghifard, M., Hwang, D.M. and Guttman, D.S., 2018. Constructing and analyzing microbiome networks in R. *Microbiome analysis: Methods and protocols*, pp.243-266.

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Host-Microbe Biology

A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems

Begüm D. Topçuoğlu,[a] Nicholas A. Lesniak,[a] Mack T. Ruffin IV,[c] Jenna Wiens,[b] Patrick D. Schloss[a]

[a]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA
[b]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, USA
[c]Department of Family Medicine and Community Medicine, Penn State Hershey Medical Center, Hershey, Pennsylvania, USA

- This study trained seven models that used fecal 16S rRNA sequence data to predict the presence of colonic screen relevant neoplasias (SRNs) (*n* = 490 patients, 261 controls and 229 cases

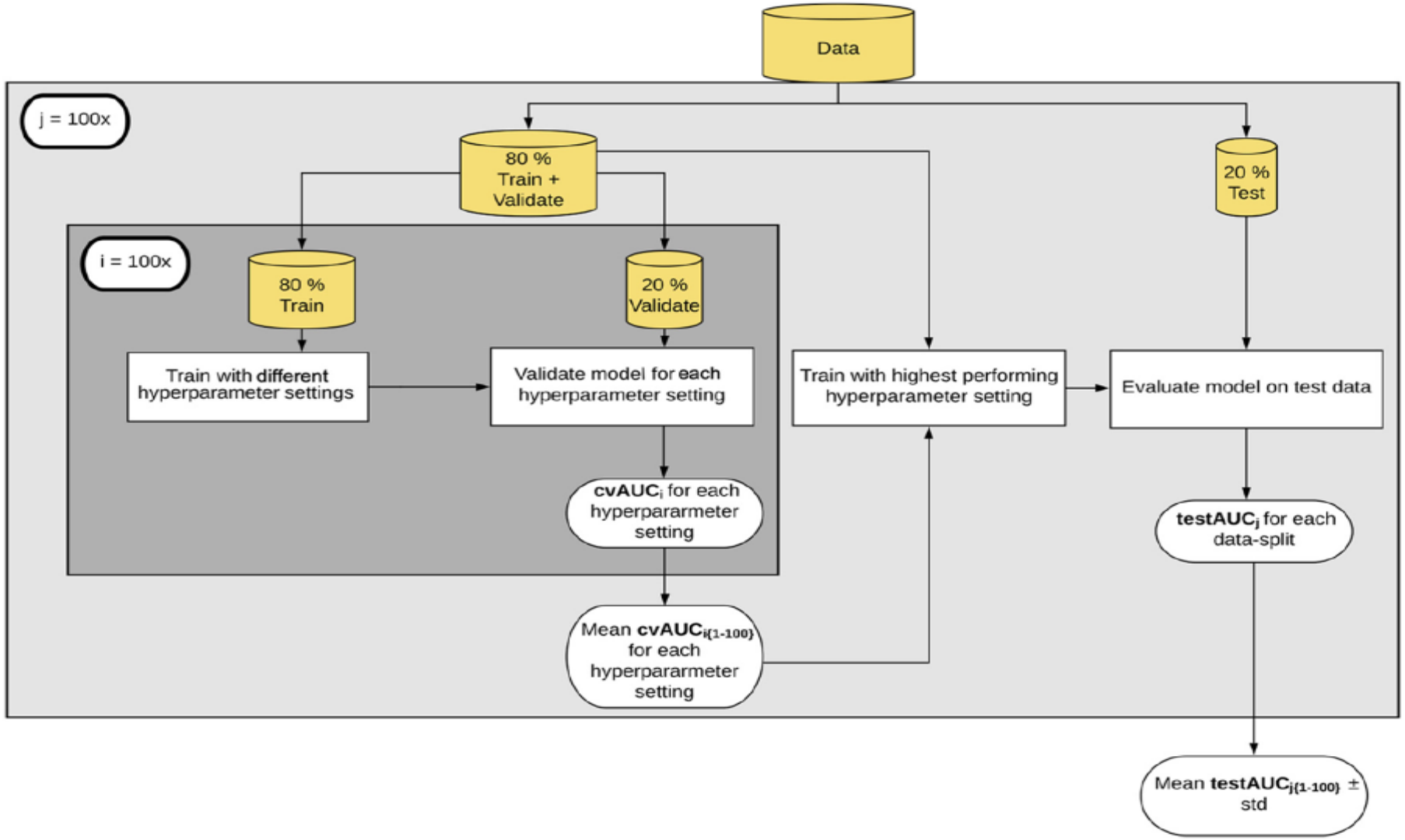


FIG 1 Machine learning pipeline. We split the data to create a training (80%) and held-out test set (20%). The splits were stratified to maintain the overall class distribution. We performed five-fold cross-validation on the training data to select the best hyperparameter setting and then used these hyperparameters to train the models. The model was evaluated on the held-out data set. Abbreviations: cvAUC, cross-validation area under the receiver operating characteristic curve.

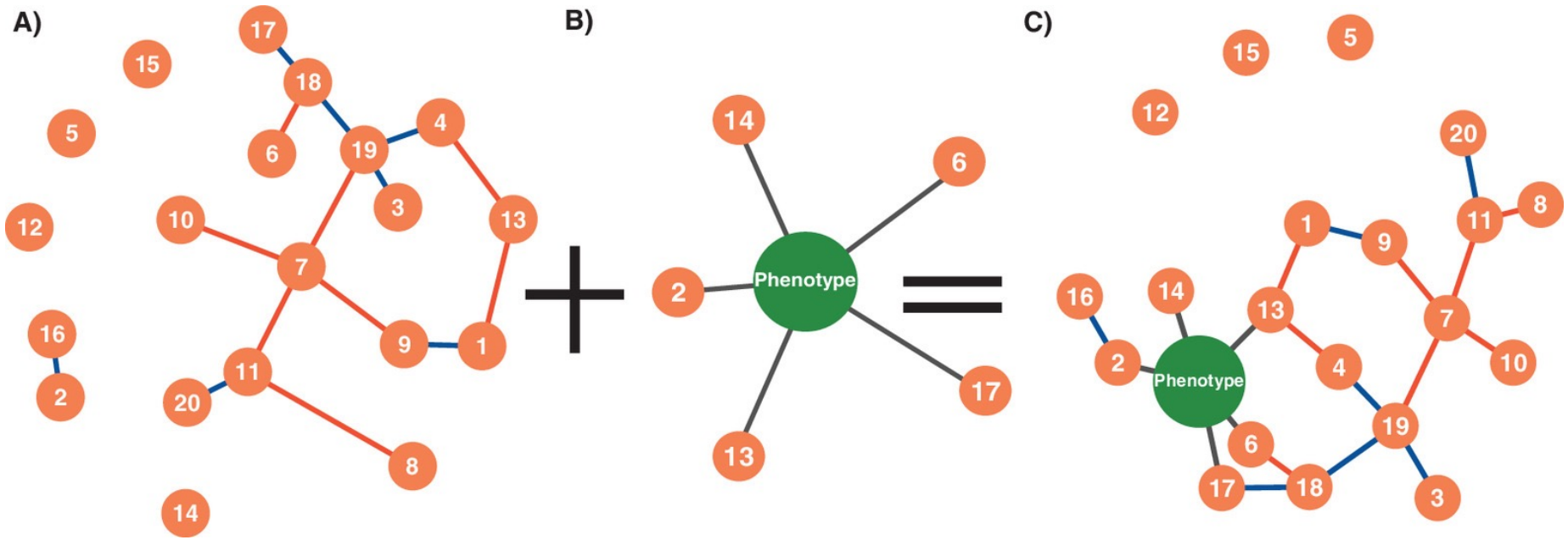
| Microbial Ecology | Research Article | 24 May 2023

Integration of Phenotypes in Microbiome Networks for Designing Synthetic Communities: a Study of Mycobiomes in the Grafted Tomato System

Authors: Ravin Poudel, Ari Jumpponen, Megan M. Kennelly, Cary Rivard, Lorena Gomez-Montano, Karen A. Garrett

AUTHORS INFO & AFFILIATIONS

DOI: https://doi.org/10.1128/aem.01843-22 • Check for updates



Hands-on experience on building network models with microbiome date

Let’s switch to Rstudio

```
# Loading required packages
library(igraph)
library(Hmisc)
library(Matrix)
```

```
# Load the data with the OTU table: otudata.csv
otu.table<-read.csv(file.choose(), header=T, row.names = 1)

# Read taxonomy file associated with OTU table into new object:
otu_taxonomy.csv
tax<-read.csv(file.choose(),header=T, row.names = 1)

# Check how many OTUs we have
dim(otu.table)

# Keep the OTUs with more than 10 counts
otu.table.filter<-otu.table[,colSums(otu.table)>10]

# Check for the decrease in the number of OTUs
dim(otu.table.filter)
```

```
# Calculate pairwise correlation between OTUs
otu.cor<-rcorr(as.matrix(otu.table.filter), type="spearman")


# Get p-value matrix
otu.pval <- forceSymmetric(otu.cor$P) # Self-correlation as NA

# Select only the taxa for the filtered OTUs by using rownames
of otu.pval
sel.tax <-tax[rownames(otu.pval),,drop=FALSE]


# Sanity check
all.equal(rownames(sel.tax), rownames(otu.pval))
```

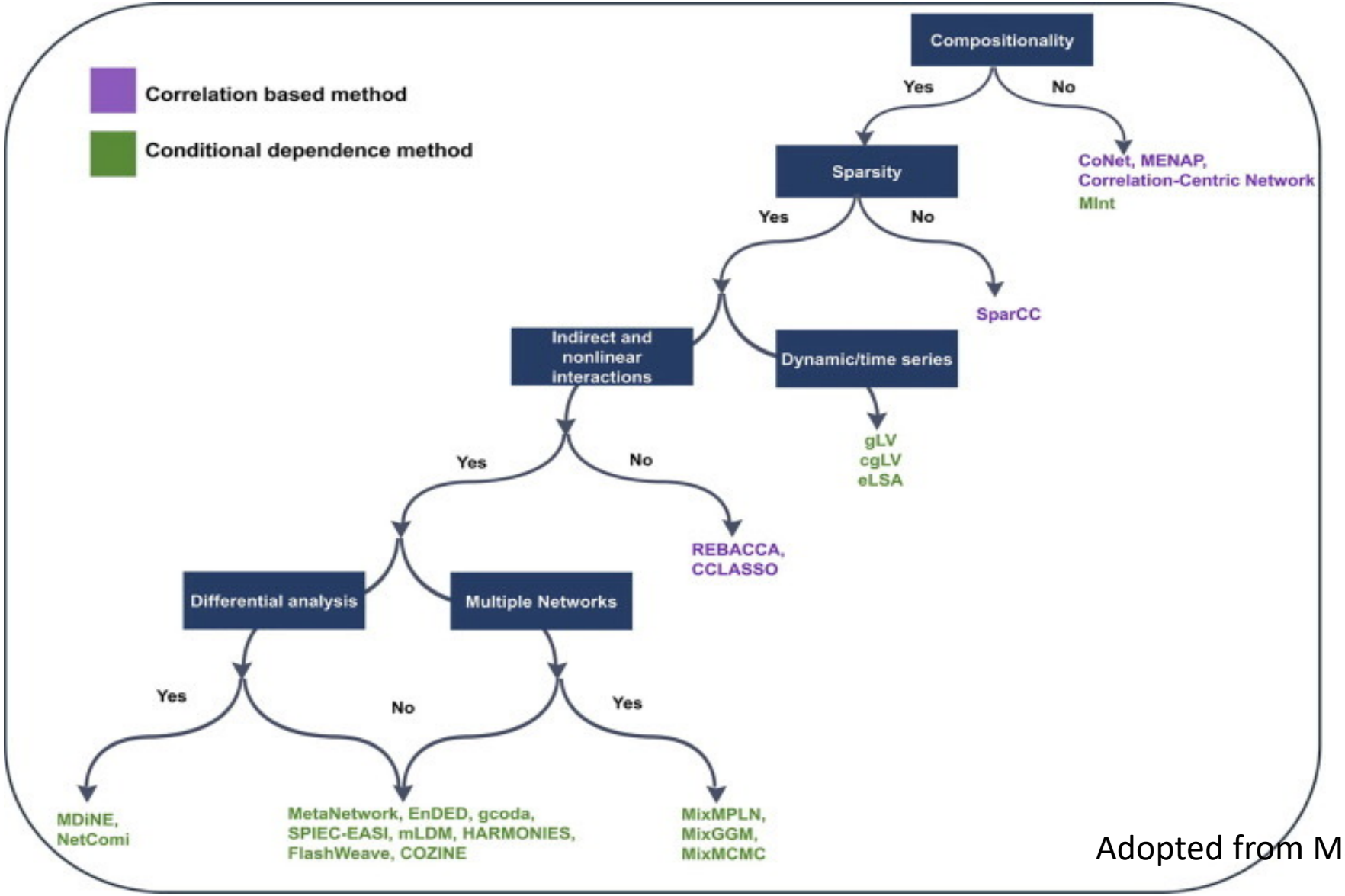
```
# Filter the association based on p-values and level of correlations
p.yes<-otu.pval<0.001


# Select the r values for p.yes
r.val=otu.cor$r # select all the correlation values
p.yes.r<-r.val*p.yes # only select correlation values based on p-value criterion


# Select OTUs by level of correlation
p.yes.r<-abs(p.yes.r)>0.75 # output is logical vector

p.yes.rr<-p.yes.r*r.val # use logical vector for subscripting.
```

```
# Create an adjacency matrix
adjm<-as.matrix(p.yes.rr)

# Add taxonomic information associated with adjacency matrix
colnames(adjm)<-as.vector(sel.tax$Family)
rownames(adjm)<-as.vector(sel.tax$Family)


# Create an adjacency matrix in igraph format
net.grph=graph.adjacency(adjm,mode="undirected",weighted=TRUE,
diag=FALSE)
```

```
# Calculate edge weight == level of correlation
edgew<-E(net.grph)$weight

# Identify isolated nodes
bad.vs<-V(net.grph)[degree(net.grph) == 0]

# Remove isolated nodes
net.grph <-delete.vertices(net.grph, bad.vs)
```

```
# Plot the graph object
plot(net.grph,
     vertex.size=4,
     vertex.frame.color="black",
     edge.curved=F,
     edge.width=1.5,
     layout=layout.auto,
     edge.color=ifelse(edgew<0,"red","blue"),
     vertex.label=NA,
     vertex.label.color="black",
     vertex.label.family="Times New Roman",
     vertex.label.font=2)

# Plot the graph object
plot(net.grph,
     vertex.size=5,
     vertex.frame.color="black",
     edge.curved=F,
     edge.width=1.5,
     edge.color=ifelse(edgew<0,"red","blue"),
     vertex.label.color="blue",
     vertex.label.family="Times New Roman",
     vertex.label.font=0.5)
```

Workflow indicating the suitable network approaches depending upon different challenges.



Adopted from Matchado et al. 2021