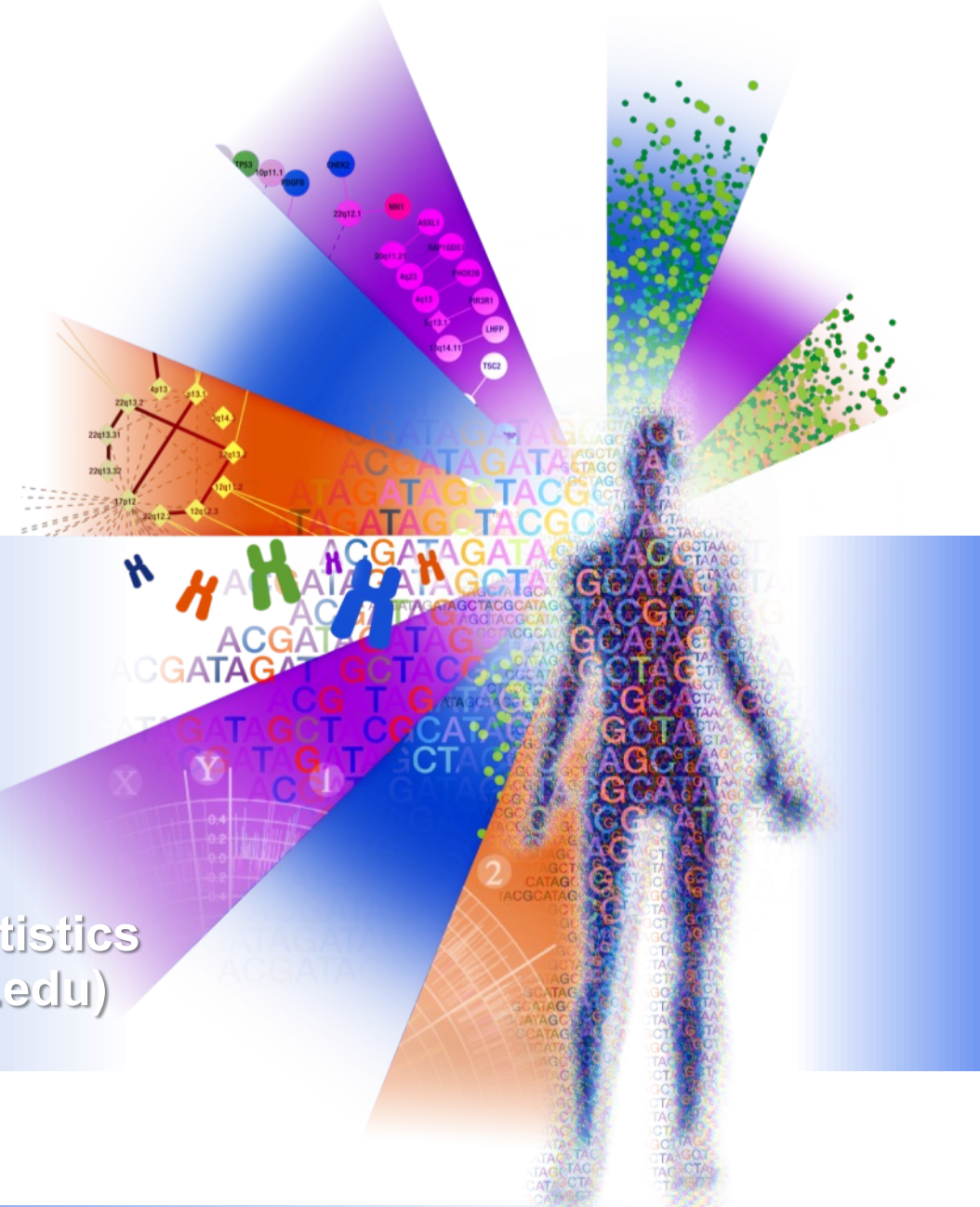
MAYO CLINIC



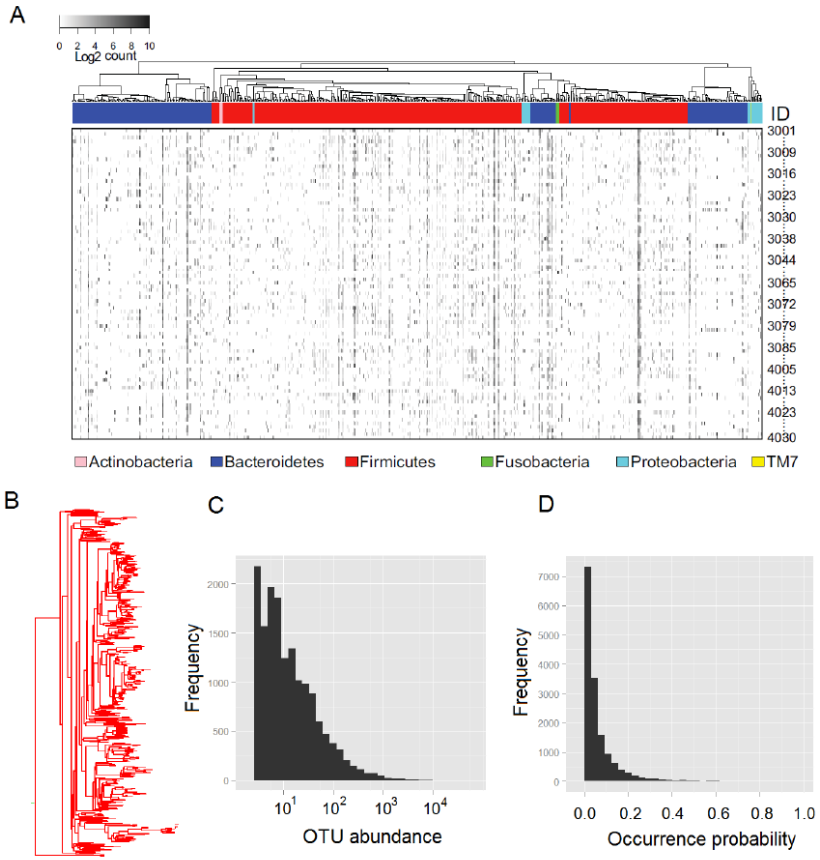
CENTER FOR INDIVIDUALIZED MEDICINE

Introductory Microbiome Statistics

Jun Chen (chen.jun2@mayo.edu)

©2012 MFMER | slide-1

Characteristics of microbiome data

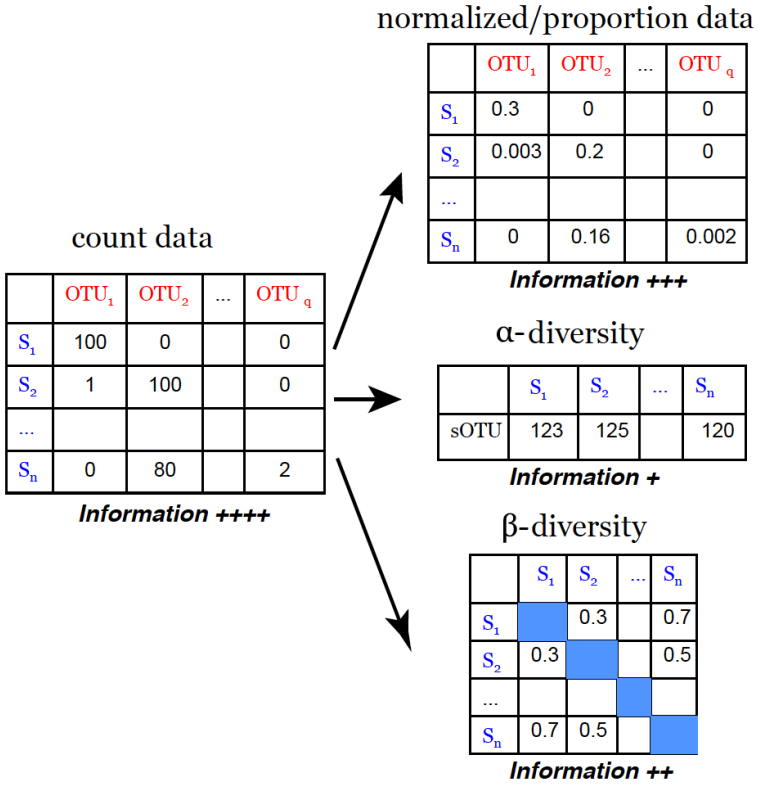


Data characteristics

- High-dimensional counts with variable library sizes
- Overdispersion, excessive zeros and outliers
- Phylogeny and hierarchical structure
- Compositional effects

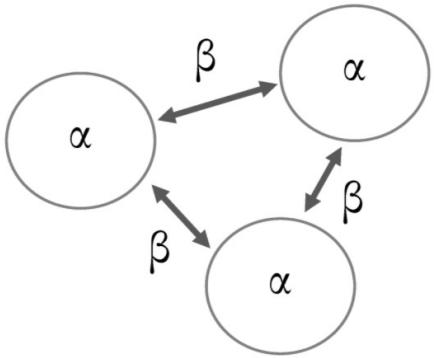
Multi-level summarization of microbiome data

- Raw count data
  - OTU/ASV level
  - Different taxonomic level
- Normalized/Proportion data
  - OTU/ASV level
  - Different taxonomic level
- Diversity (community-level) data
  - α-diversity
  - β-diversity



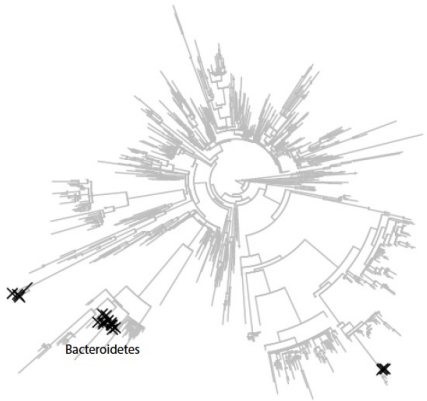
Diversity measures quantify community-level properties

- α-diversity: a measure of within-sample diversity (richness and evenness)
  - Not dependent on the composition of other samples
  - Sobs, Chao1, Shannon, Simpson, Phylogenetic diversity (PD)
- β-diversity: a measure of between-sample diversity in terms of pairwise ecological distances
  - Measures how different one community is to another
  - Bray-Curtis, Jaccard, UniFrac distance
- Can be split along two major axes:
  - Unweighted vs. weighted measure
  - Phylogenetic vs. Non-phylogenetic measure

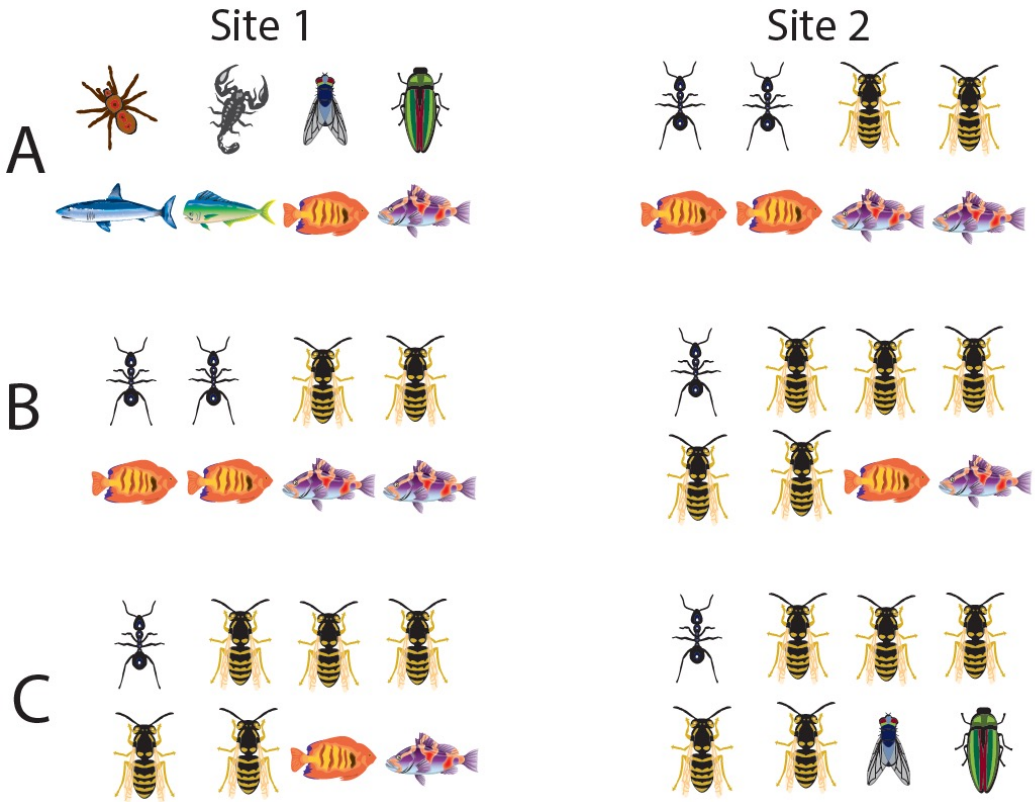


Each diversity measure captures a different aspect of the community properties

- Unweighted measure - community membership/structure;
  Weighted measure - community compositional profile;
- Unweighted measure also puts more emphasis on rare/low-abundance species than weighted measure
- Phylogenetic measure captures "clustered" signals
- It is important to study several representative diversity measures to have a full view of the community



Each diversity measure captures a different aspect of the community properties



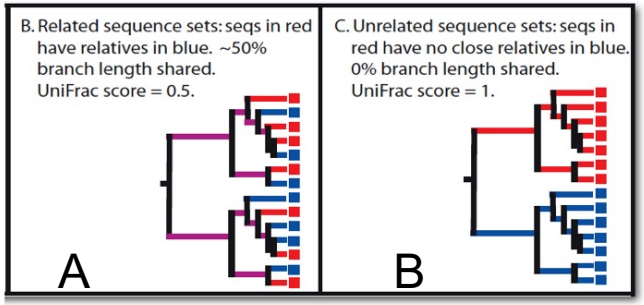
Different richness, same evenness (Sobs, Chao1)

Same richness, different evenness (Shannon, Simpson)

Same richness, same evenness, different phylogeny (Phylogenetic diversity)

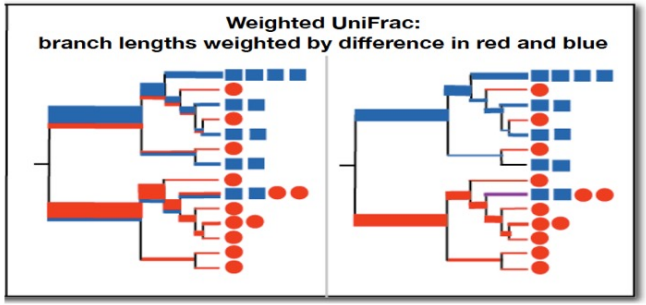
UniFrac distance (Knight group) captures phylogenetically related community difference

Unweighted UniFrac(d^U): the fraction of the branch length of the tree that is unique to each community/microbiome sample.



- Most efficient in detecting the difference in community membership
- Also difference in rare lineages, differs in probability of being picked up by the sequencing machine

Weighted UniFrac(d^W): branch length weighted by proportion difference



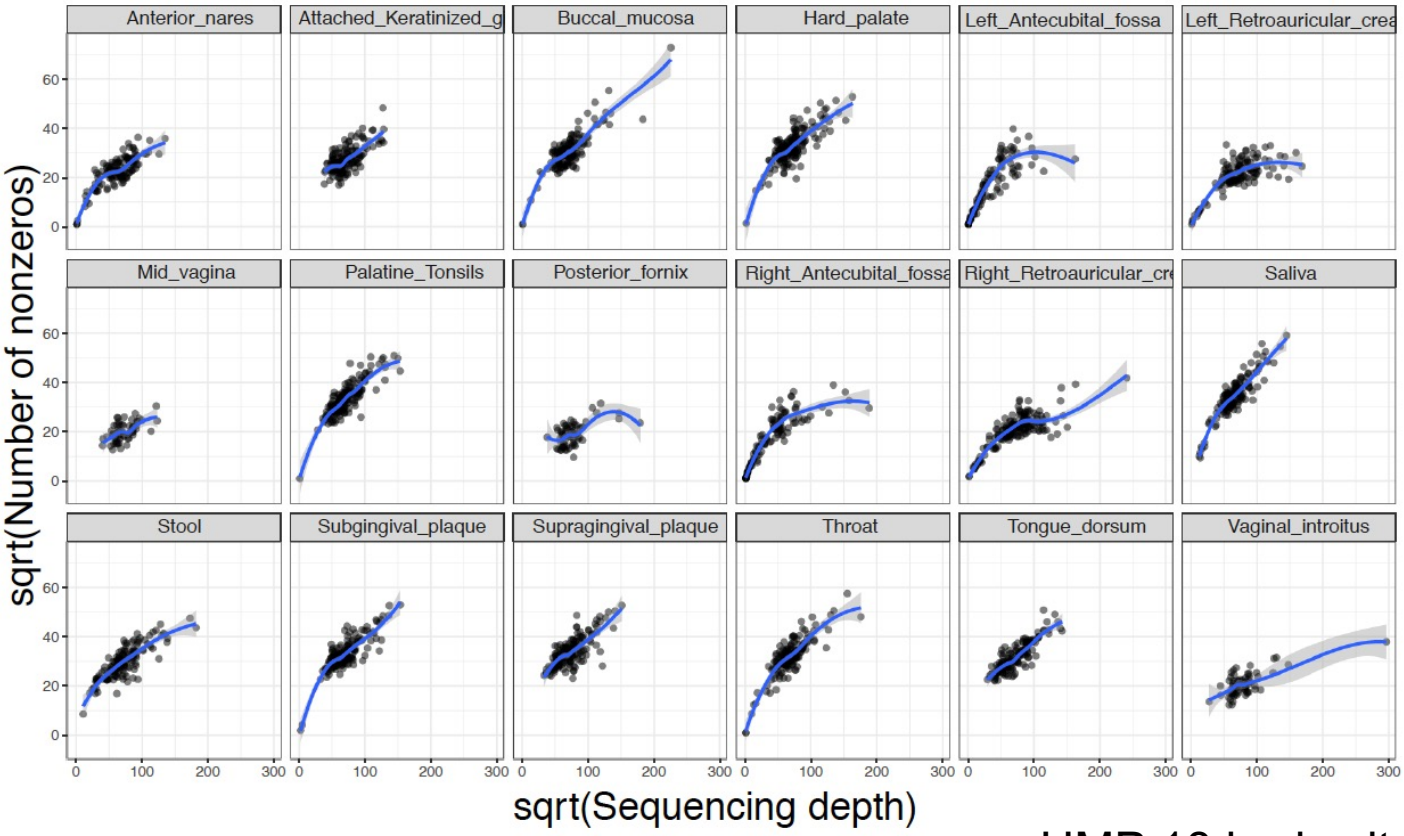
- Most efficient in detecting difference in abundant lineages
- Absolute difference puts too much weight on abundant lineages

Bray-Curtis/Jaccard distances are UniFrac equivalents without using phylogeny.

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, *71*(12), 8228-8235.

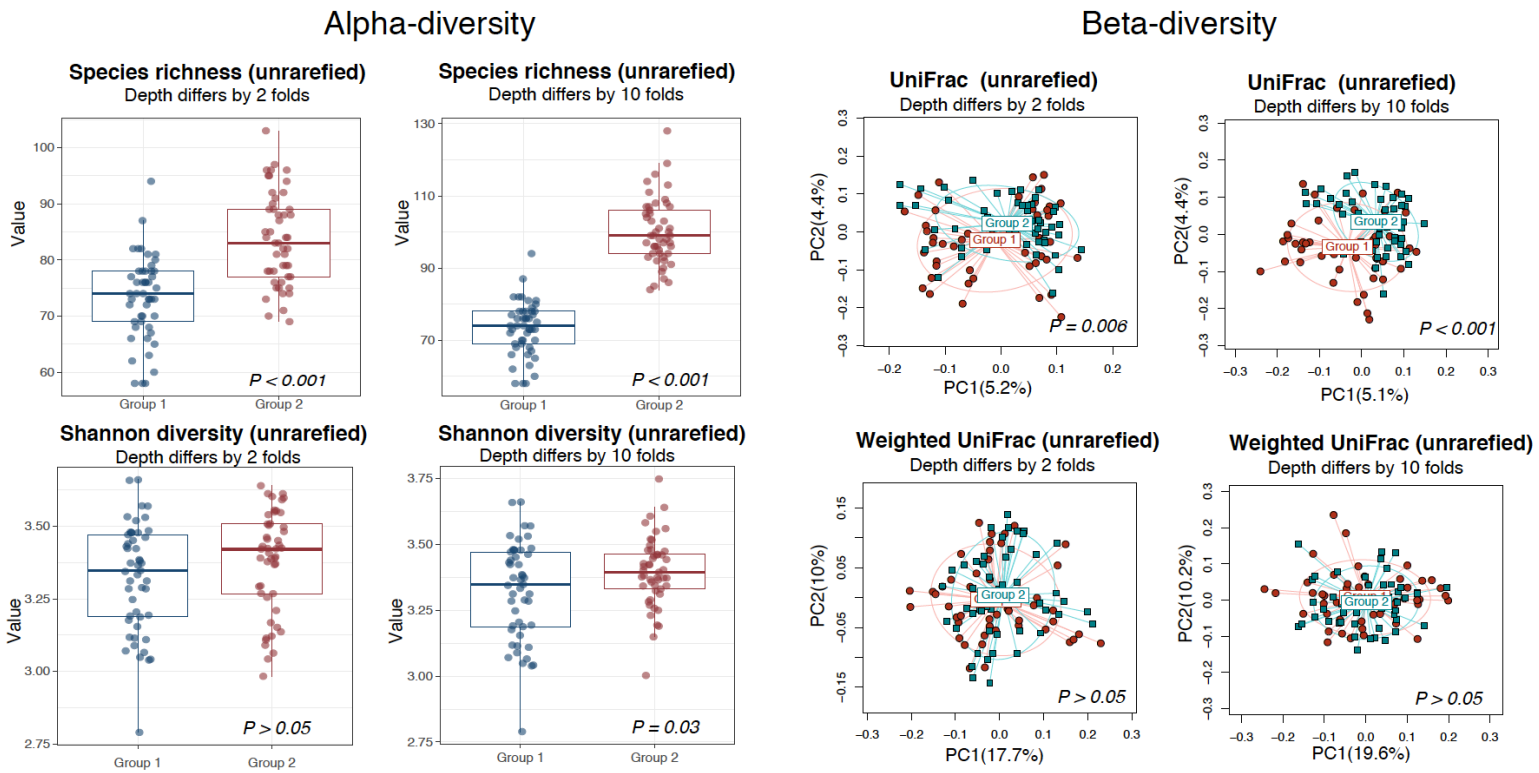
Diversity depends on the sampling effort (sequencing depth)

As we sequence more deeply, we will detect more rare species.



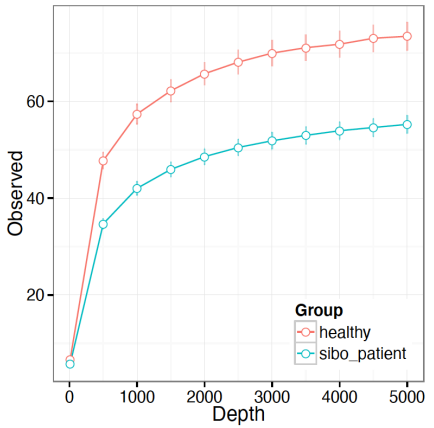
HMP 18 body sites

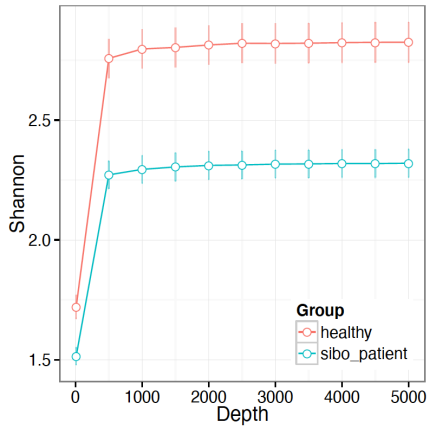
Differential sequencing depth could produce false patterns



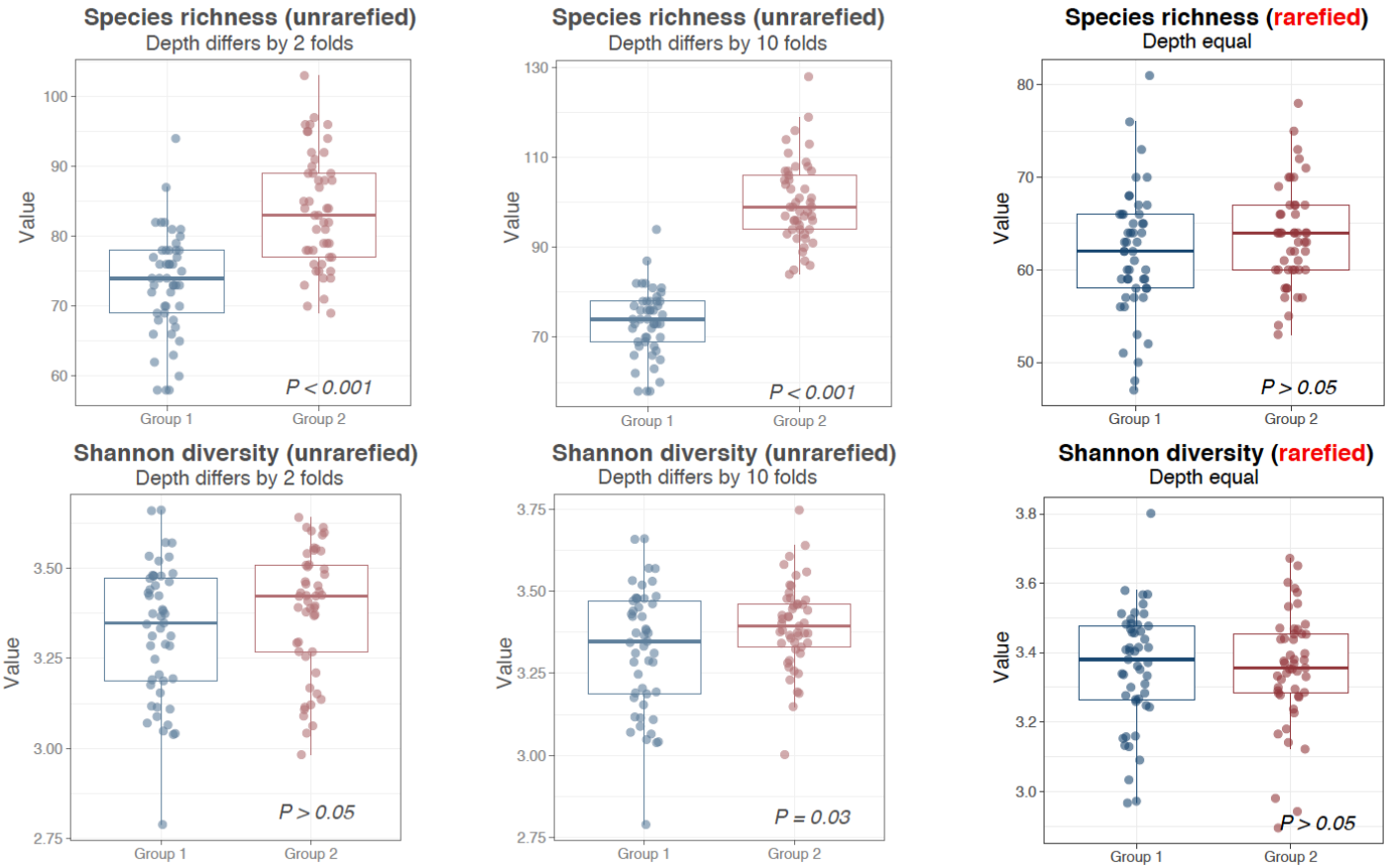
Diversity measure and rarefaction

- To conduct a fair comparison of diversity measures, esp. for unweighted measures, the sequencing depth should be equal. The process of down-sampling the sequences to the same depth is called rarefaction.
- We usually down-sample the sequences at different depths and create a rarefaction curve. The rarefaction curve can be used to assess the adequacy of the sampling depth.
- Unless rarefied to a very low depth, rarefaction does not lose much statistical power for diversity analysis. It could even increase the power for unweighted analysis.

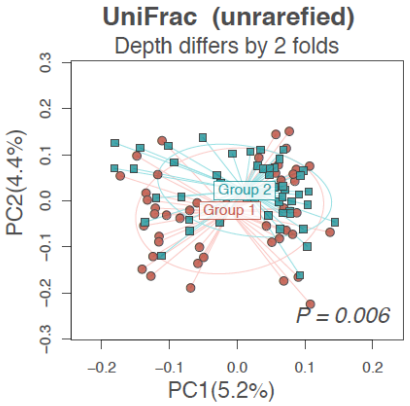


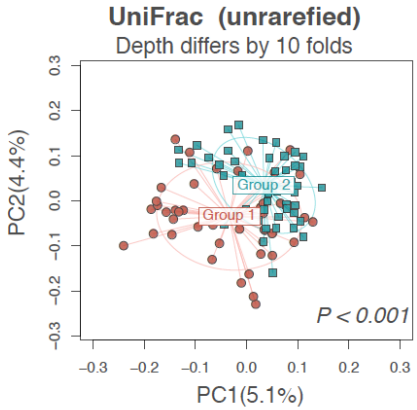


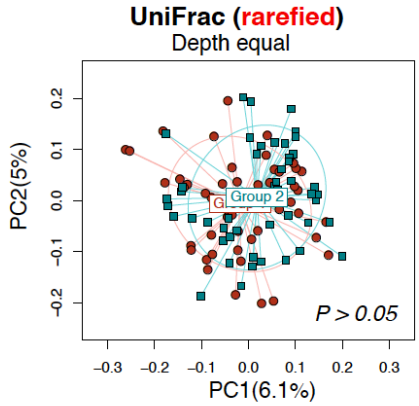
Rarefaction corrects the false pattern

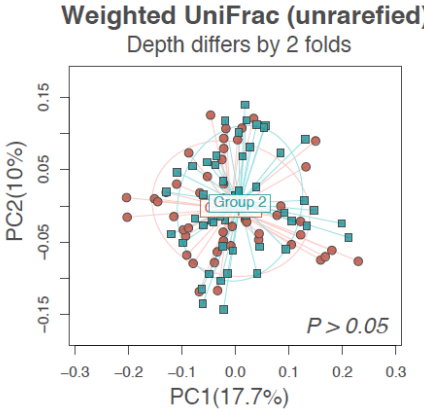


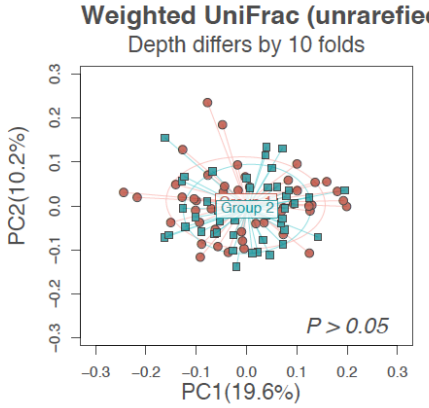
Rarefaction corrects the false pattern

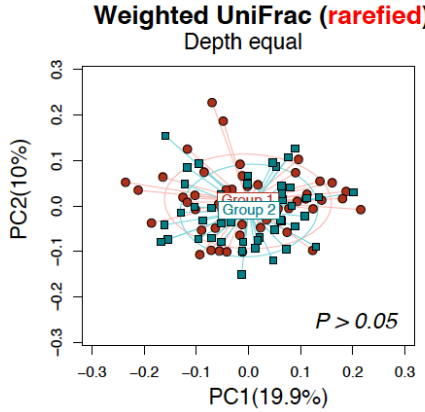




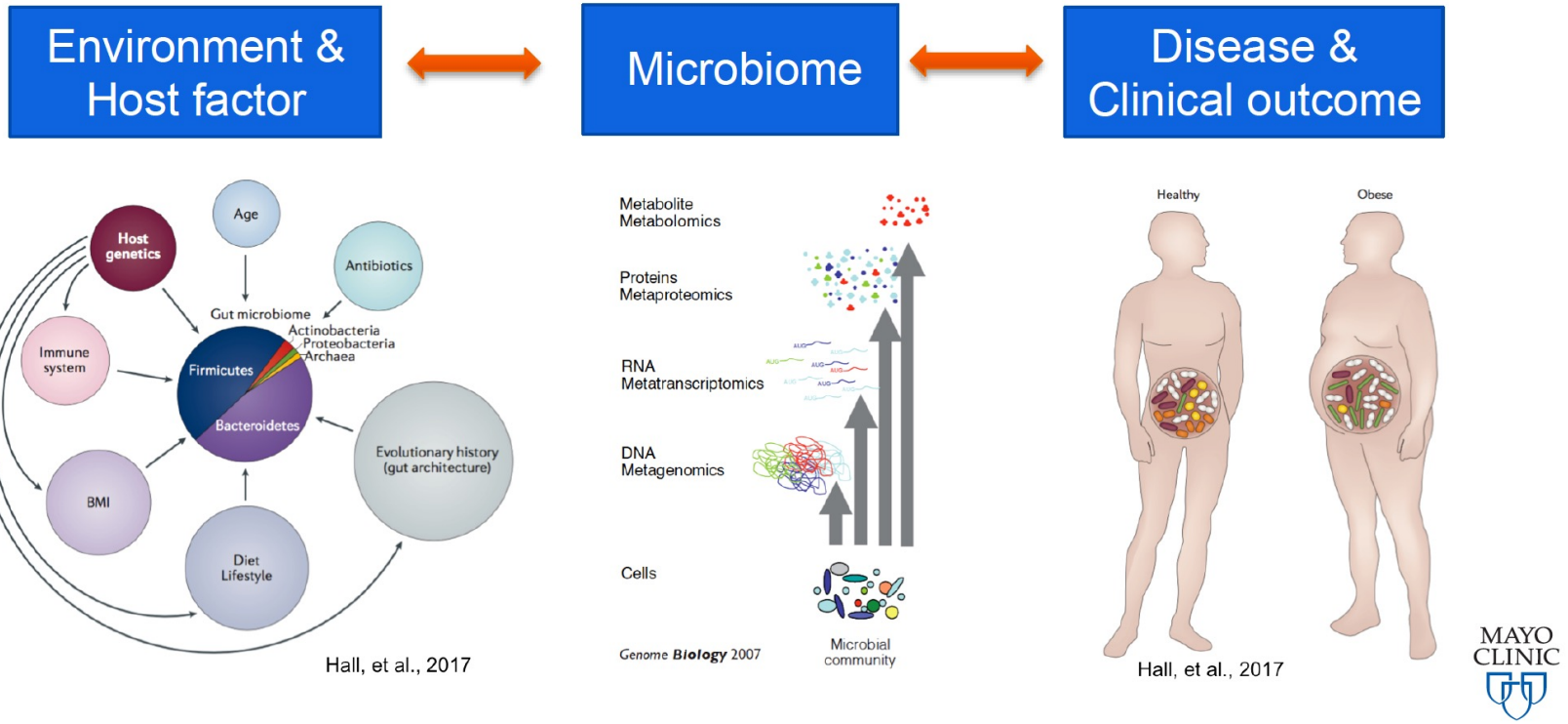






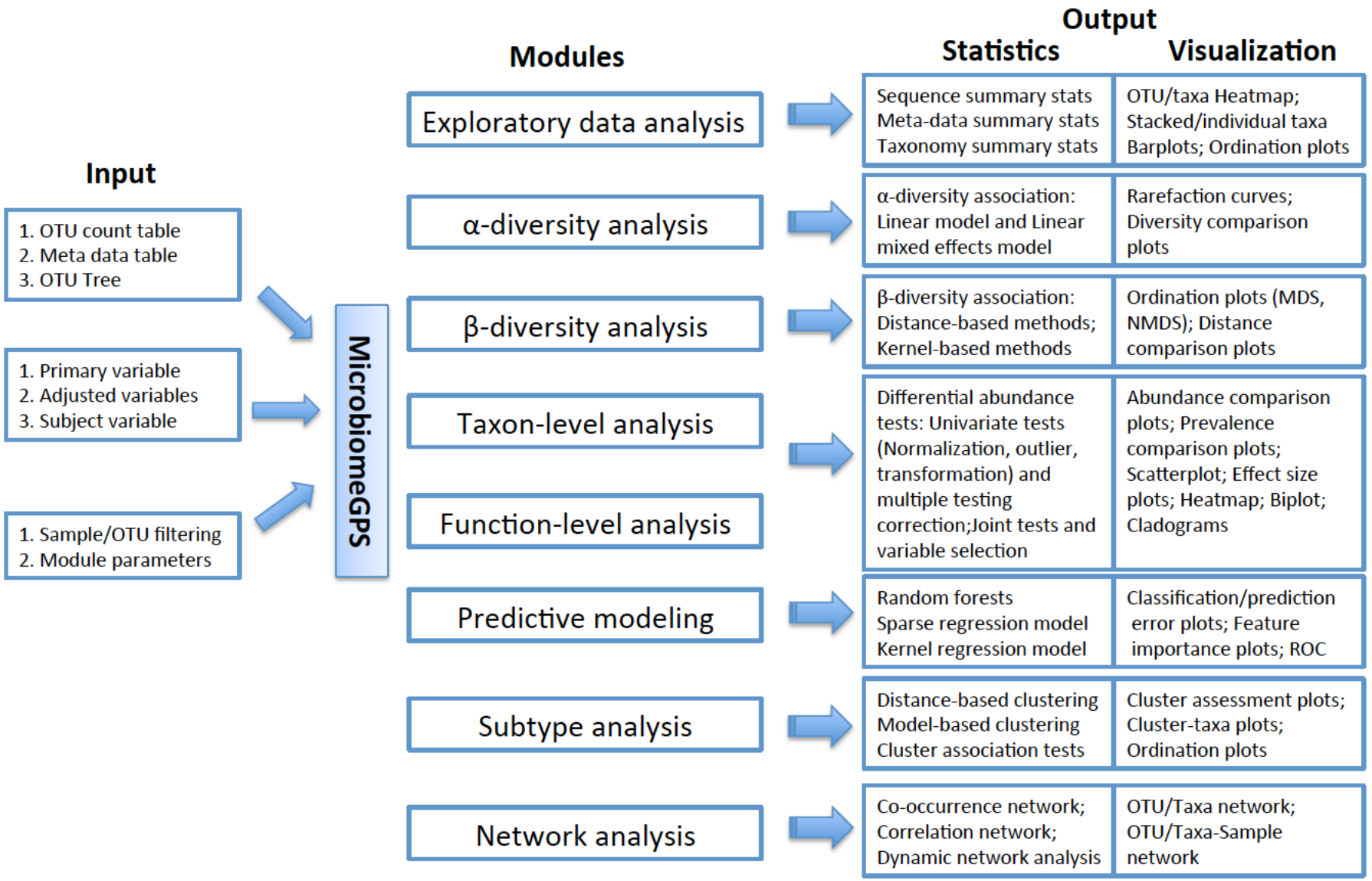


Central themes of microbiome studies



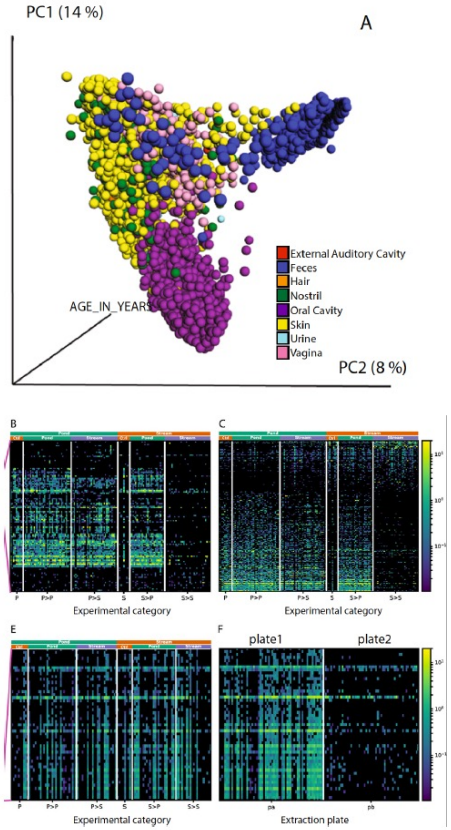
Hall, A. B., Tolonen, A. C., & Xavier, R. J. (2017). Human genetic variation and the gut microbiome in disease. *Nature Reviews Genetics*, *18*(11), 690-699.

Warnecke, F., & Hugenholtz, P. (2007). Building on basic metagenomics with complementary technologies. *Genome biology*, *8*, 1-5.


Input
1. OTU count table
2. Meta data table
3. OTU Tree
1. Primary variable
2. Adjusted variables
3. Subject variable
1. Sample/OTU filtering
2. Module parameters
MicrobiomeGPS
Modules
Exploratory data analysis
α-diversity analysis
β-diversity analysis
Taxon-level analysis
Function-level analysis
Predictive modeling
Subtype analysis
Network analysis
Output
Statistics
Visualization
Sequence summary stats
Meta-data summary stats
Taxonomy summary stats
OTU/taxa Heatmap; Stacked/individual taxa Barplots; Ordination plots
α-diversity association: Linear model and Linear mixed effects model
Rarefaction curves; Diversity comparison plots
β-diversity association: Distance-based methods; Kernel-based methods
Ordination plots (MDS, NMDS); Distance comparison plots
Differential abundance tests: Univariate tests (Normalization, outlier, transformation) and multiple testing correction;Joint tests and variable selection
Abundance comparison plots; Prevalence comparison plots; Scatterplot; Effect size plots; Heatmap; Biplot; Cladograms
Random forests
Sparse regression model
Kernel regression model
Classification/prediction error plots; Feature importance plots; ROC
Distance-based clustering
Model-based clustering
Cluster association tests
Cluster assessment plots; Cluster-taxa plots; Ordination plots
Co-occurrence network; Correlation network; Dynamic network analysis
OTU/Taxa network; OTU/Taxa-Sample network

Exploratory data analysis

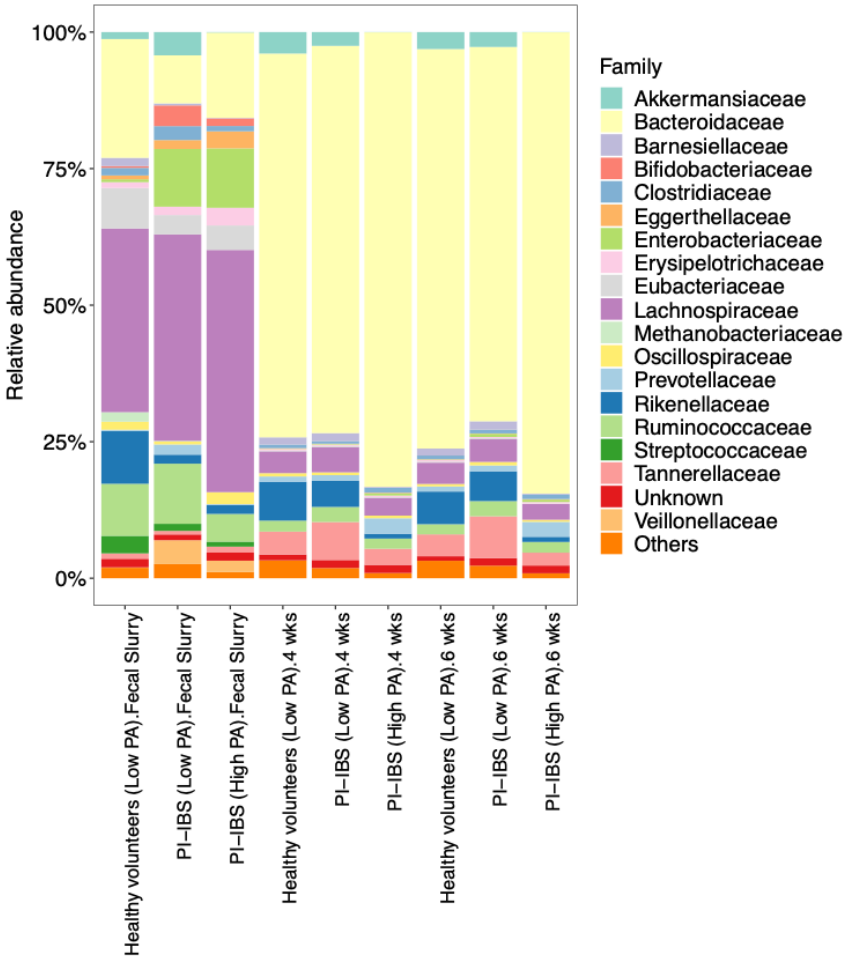
- Various visualizations through heat map, stacked bar plot, PCoA plot, cladogram
- Exploratory data analysis can
  - Detect outlier samples/features
  - Understand sources of variability, particularly batch effects
  - Detect contaminants by comparing to negative controls
  - Form hypotheses for statistical testing
- Two interactive tools
  - EMPeror (Vzquez-Baeza, et al., 2013, GigaScience): PCoA plot-based.
  - Calour (Xu et al., 2019, mSystems): heat map-based.

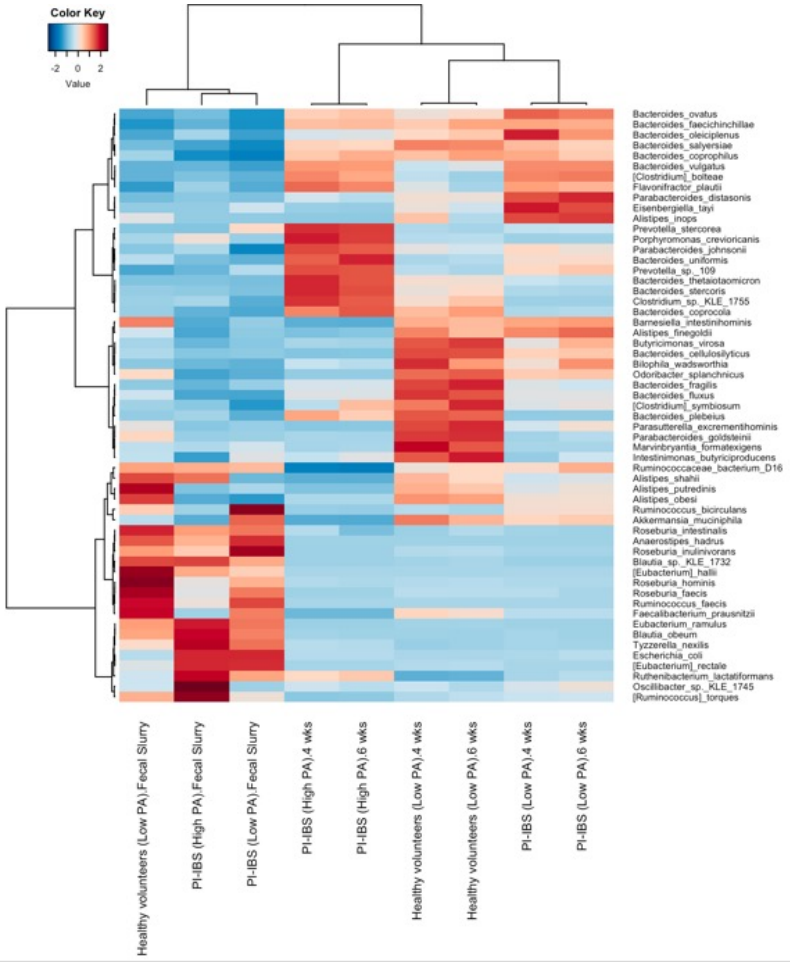


Xu, Z. Z., et al. (2019). Calour: an interactive, microbe-centric analysis tool. *Msystems*, *4*(1), 10-1128.

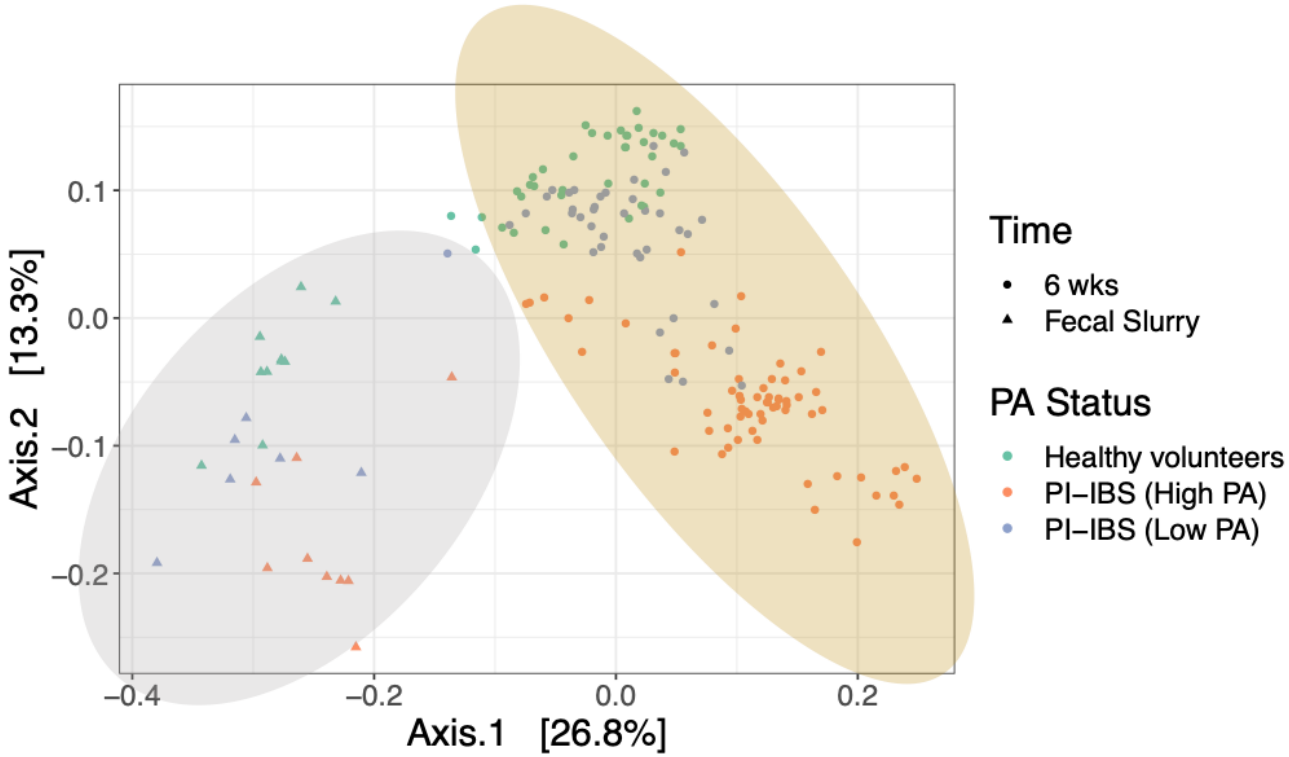
Vázquez-Baeza, Y., et al. (2013). EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience*, *2*(1), 2047-217X.

Exploratory data analysis (examples)





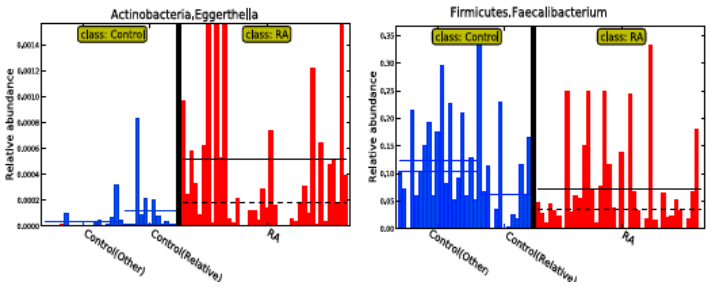
Exploratory data analysis (examples)



Microbiome association test

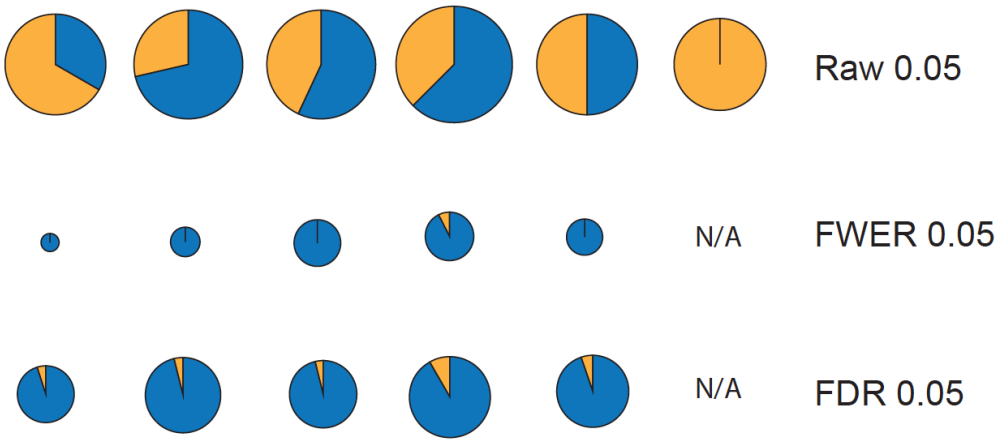
- Individual taxon vs community-level test
- Community-level test to establish an overall association
  - α-diversity: close to normal, linear (mixed) model can be applied
  - β-diversity: PERMANOVA and MiRKAT (Zhao et al., 2015, AJHG)
    - Idea: compare microbiome similarity to phenotype similarity
    - PERMANOVA: distance-based regression with the microbiome as the outcome
    - MiRKAT: kernel machine regression with the microbiome as the covariate
    - More powerful methods are under development.
- Individual taxon testing to identify specific taxa (differential abundance analysis)
  - Many methods are available but no consensus exists!
  - Multiple testing correction must be performed!!

	Case samples				Control samples			
Taxon 1	c_{11}	c_{12}	$\cdots$	c_{1m}	$c_{1,m+1}$	$c_{1,m+2}$	$\cdots$	c_{1n}
Taxon 2	c_{21}	c_{22}	$\cdots$	c_{2m}	$c_{2,m+1}$	$c_{2,m+2}$	$\cdots$	c_{2n}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
Taxon q	c_{q1}	c_{q2}	$\cdots$	c_{qm}	$c_{q,m+1}$	$c_{q,m+2}$	$\cdots$	c_{qn}
Library size	s_1	s_2	$\cdots$	s_m	s_{m+1}	s_{m+2}	$\cdots$	s_n



Multiple testing correction

- Assume we are testing 500 taxa, if we use the traditional type I error cutoff 0.05, by definition, we will detect an average of 25 differential taxa while none are truly differential!!
- P-values should be corrected for multiple testing.
  - Family-wise error rate (FWER) control (e.g. Bonferroni correction) controls the probability of making any false discovery in the results. May be too conservative.
  - False discovery rate (FDR) control (e.g. Benjamini-Hochberg procedure) controls the average percentage of false discoveries in the results. Balance of power and false positives.



Left shows the proportion of true positives (blue) and false positives (yellow) on six example datasets by using unadjusted, FWER-adjusted, and FDR-adjusted p-values at 0.05 level. The size of the circle represents the number of discoveries made.

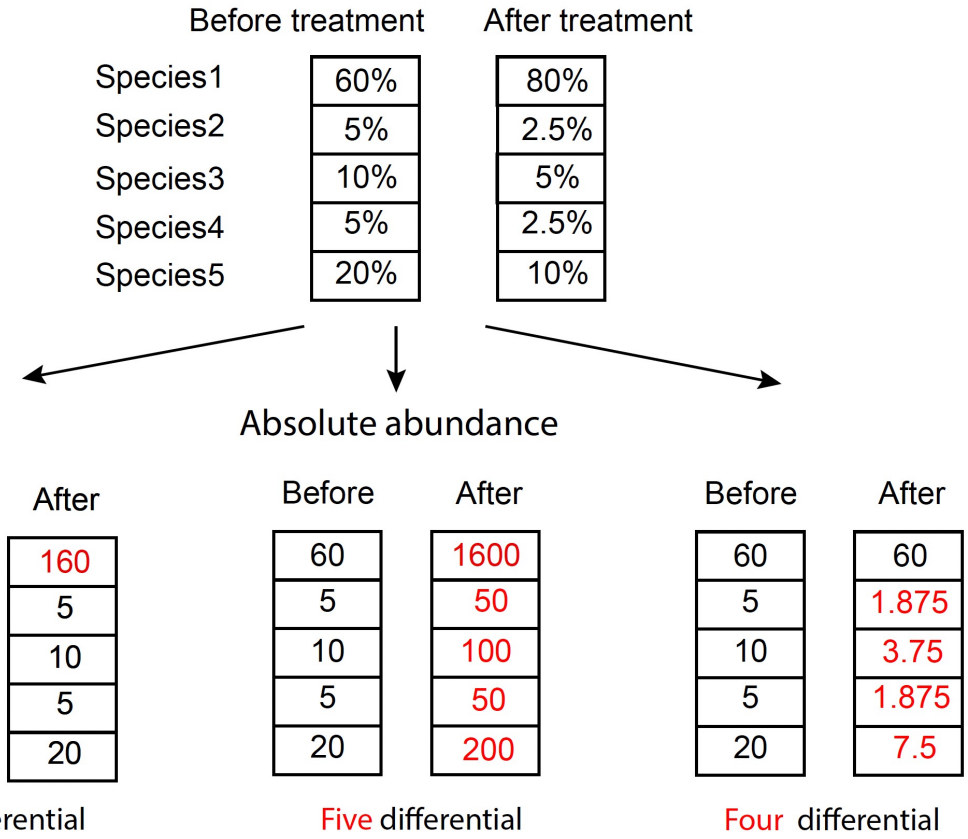
Challenges of differential abundance analysis

- Microbiome sequencing data are "compositional": the change of the abundance of one taxon will lead to the changes of the relative abundances of other taxa
- The severity of compositional effects depends on the number of differential taxa, their abundance level, and the direction of change

	Before treatment			After treatment	
Species1	60	60%	→	160	80%
Species2	5	5%		5	2.5%
Species3	10	10%		10	5%
Species4	5	5%		5	2.5%
Species5	20	20%		20	10%
Total	**100**	**100%**		**200**	**100%**

The "relative" abundances (relative to **the total sum**) are changed for all species!

Assumption for differential abundance analysis based on relative abundance data



Approaches to address compositional effects

- Select a suitable "divisor"
  - Use some robust normalizing factor (CSS, GMPR, CLR, etc.) to capture the sequencing effort of the non-differential part (sparse signal)
  - Find a single or a set of reference taxa, which are assumed to be relatively invariant with respect to some condition (DACOMP, RAIDA, ZicoSeq)
- Bias correction (ANCOM-BC, LinDA, fastANCOM)

Yang, L., & Chen, J. (2022). A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome*, *10*(1), 130.

Zhou, H., He, K., Chen, J., & Zhang, X. (2022). LinDA: linear models for differential abundance analysis of microbiome compositional data. *Genome biology*, *23*(1), 95.

Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature communications*, *11*(1), 3514.

A non-exhaustive list of existing differential abundance analysis methods

Method	Model	Address compositional effects	Handling zero	Covariate
DESeq2	Negative binomial model	Scale factor (RLE)	Model (Overdispersion)	Yes
edgeR	Negative binomial model	Scale factor (TMM)	Model (Overdispersion)	Yes
metagenomeSeq	Zero-inflated log-normal model	Scale factor (CSS/Wrench)	Model (Zero-inflation)	Yes/No
RAIDA	Zero-inflated log-normal model	Reference	Model (Zero-inflation)	No
ANCOM	Wilcoxon	Pairwise log ratio	Pseudo-count	No
ANCOM-BC	Log linear model	Bias correction	Pseudo-count	Yes
fastANCOM	Log linear model	Bias correction	Pseudo-count	Yes
LinDA	Log linear model	Bias correction	Pseudo-count	Yes
MaAsLin2	Log linear model	Scale factor	Pseudo-count	Yes
DACOMP	Wilcoxon	Reference	Not neccesary	No
ZINQ	Logistic+quantile regression	Scale factor	Model (Logistic)	Yes
LDM	Linear model	Scale factor (TSS)	Not necessary	Yes
ZicoSeq	Linear model	Reference	Empirical Bayes	Yes
MicrobiomeDDA (mbzinb)	Zero-inflated negative binomial model	Scale factor (GMPR)	Model (Zero-inflaton)	Yes
Aldex2	Wilcoxon/t test/glm	Scale factor (CLR)	Empirical Bayes	Yes
Beta-binomial (Corncob)	beta-binomial model	Scale factor (TSS)	Model (Overdispersion)	Yes
eBay	Wilcoxon/t test	Scale factor (CLR)	Empirical Bayes	No

LinDA: linear models for differential abundance analysis with bias correction (Zhou et al., 2022, Genome Biology)

https://cran.r-project.org/web/packages/MicrobiomeStat/

True model based on absolute abundance X

$$\log(X_{is}) = u_s \alpha_i + (1, \mathbf{c}_s^\top)\beta_i + \epsilon_{is},$$

where u_s is the variable of interest, $\mathbf{c}_s$ are covariates, and ϵ_{is} is the error term.

Model based on CLR-transformed relative abundance Y

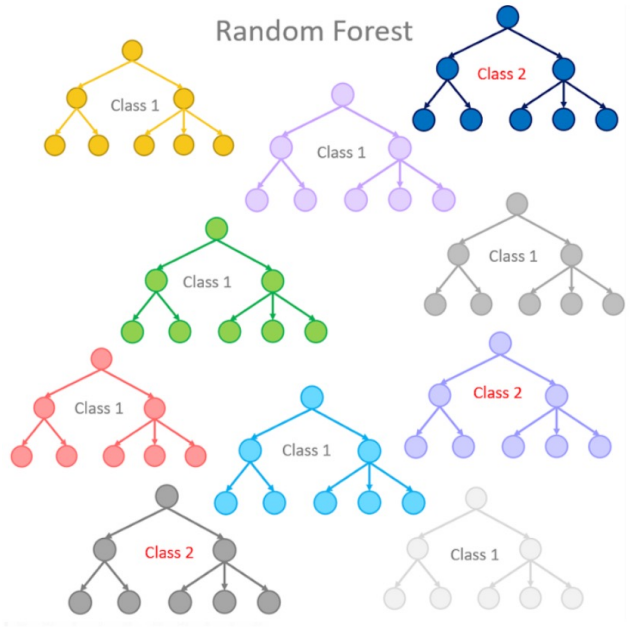
$$\begin{aligned} W_{is} :&= \log\left\{\frac{Y_{is}}{(\prod_{j=1}^m Y_{js})^{1/m}}\right\} \approx log\left\{\frac{X_{is}}{(\prod_{j=1}^m X_{js})^{1/m}}\right\} \\ &= \log(X_{is}) - \frac{1}{m}\sum_{j=1}^m \log(X_{js}) \\ &= u_s(\alpha_i - \bar{\alpha}) + (1, \mathbf{c}_s^\top)(\beta_i - \bar{\beta}) + \varepsilon_{is} - \bar{\varepsilon}_s \end{aligned}$$

where $\bar{\alpha} = m^{-1}\sum_{i=1}^m \alpha_i$, $\bar{\beta} = m^{-1}\sum_{i=1}^m \beta_i$, and $\bar{\varepsilon}_s = m^{-1}\sum_{i=1}^m \varepsilon_{is}$.

The OLS estimator for α based on the CLR-transformed data is **biased** with the bias term being $\bar{\alpha}$. LinDA estimates the bias and performs inference based on the de-biased estimate.

Microbiome-based prediction with Random Forest algorithm

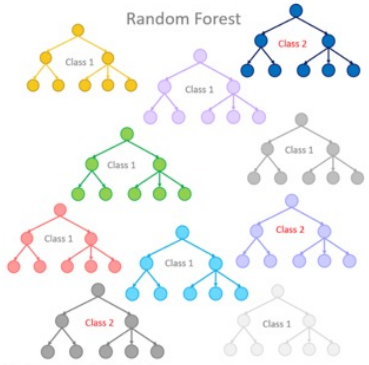
- Algorithm based on decision trees with bootstrap samples and splitting on a random set of features
- Advantages
  - Capture nonlinear effects
  - Capture interaction between taxa
  - Can accommodate a large number of taxa
  - Robust to outliers
  - Provide importance rank of the taxa
  - Can couple with Boruta feature selection to identify biomarker taxa
  - Out-of-bag (OOB) error gives a reasonable error estimate (may overestimate)
- Important parameters
  - Number of trees (default may be small)
  - Number of features to split (default is usually OK)



https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147

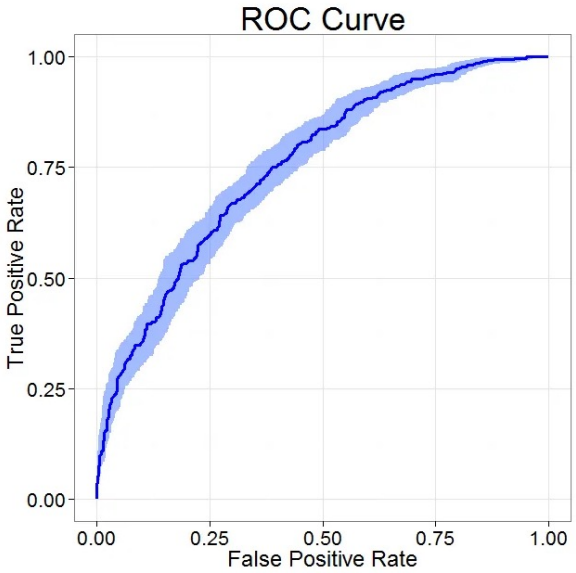
Microbiome-based prediction - considerations

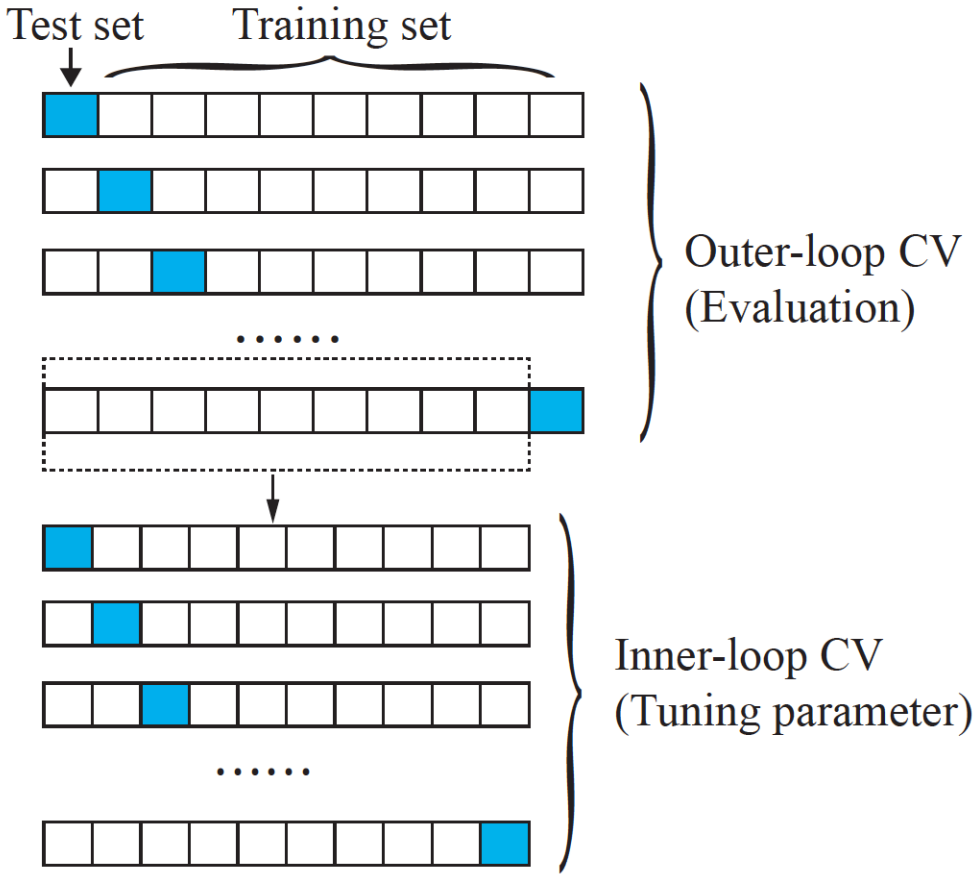
- What input to Random Forest?
  - Taxonomic or original OTU/ASV abundance
  - Whether to incorporate phylogenetic information
  - Taxonomic or functional abundance
  - How to normalize the data
  - Whether to perform feature filtering and selection



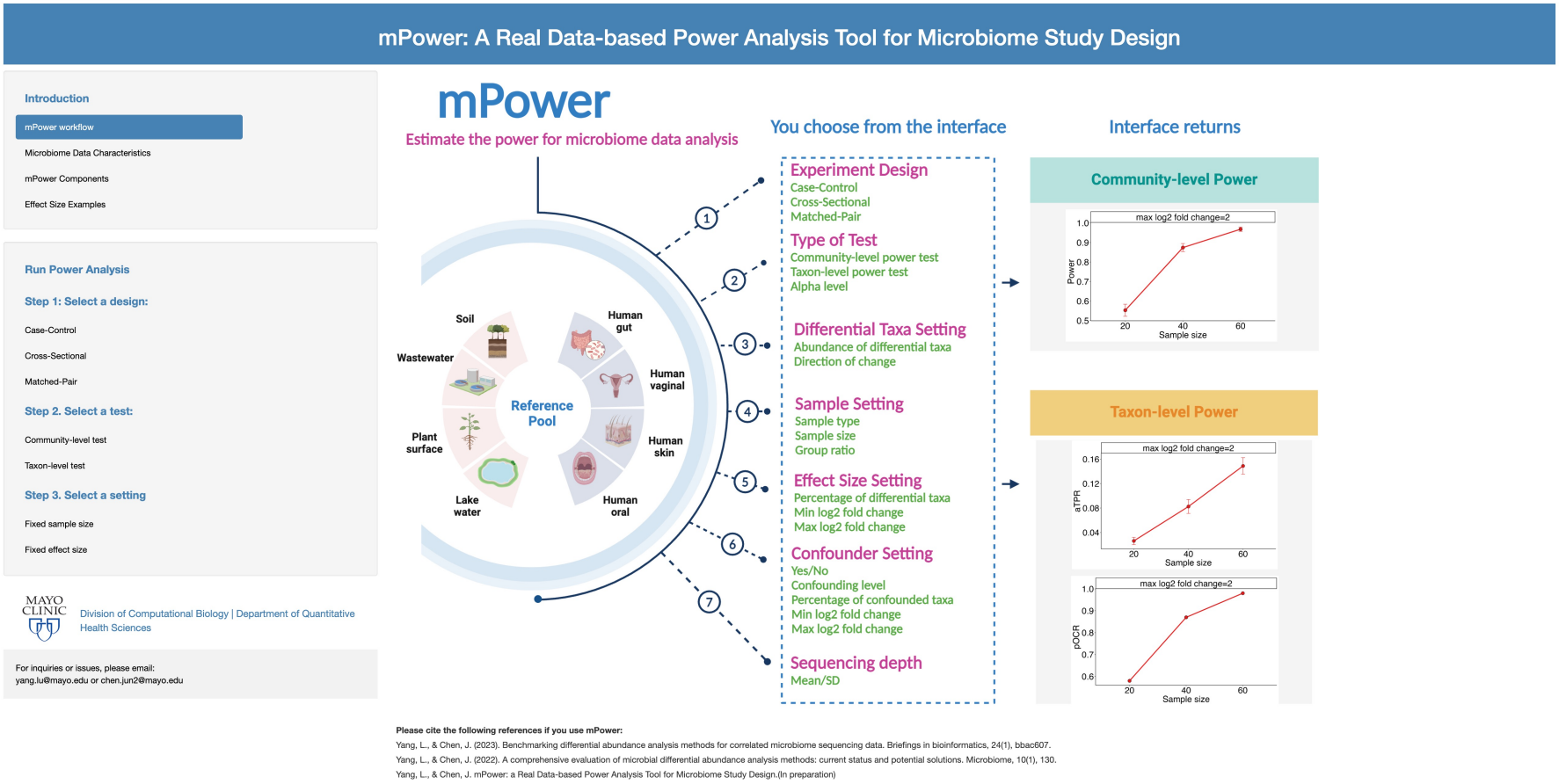
- Performance evaluation by cross-validation (CV)
  - If all the parameters are set as default, OOB error or traditional CV can be used to evaluate prediction performance.
  - If we tune the parameters and perform feature filtering/selection, two-stage nested CV should be used.
  - Feature filtering/selection should be performed in the training data only.

Nested cross-validation for objective assessment of prediction performance





mPower: a power calculation tool for microbiome study design



https://microbiomestat.shinyapps.io/mPower/ (ongoing project)

MicrobiomeStat: a R package supporting comprehensive microbiome data analysis

SINGLE-POINT ANALYSIS

Introduction

Alpha Diversity Analysis

Beta Diversity Analysis

Feature-level Analysis

One-Click Reports Generation

PAIRED SAMPLES ANALYSIS

Introduction

Alpha Diversity Analysis

Beta Diversity Analysis

Feature-level Analysis

One-Click Reports Generation

LONGITUDINAL ANALYSIS

Introduction

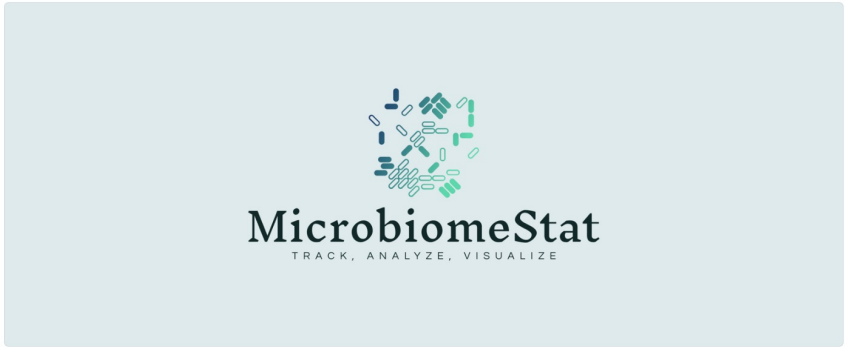
Alpha Diversity Analysis

Beta Diversity Analysis

Feature-level Analysis

One-Click Reports Generation

MicrobiomeStat: Supporting Longitudinal Microbiome Analysis in R



MicrobiomeStat is a specialized tool for analyzing longitudinal and paired microbiome data. It can also be used for longitudinal analysis of other omics datatypes as long as the data are properly normalized and transformed. As a special case, cross-sectional and case-control microbiome data analysis is also supported.

Gitbook LLM Integration

Maximize the potential of MicrobiomeStat by utilizing its Gitbook wiki documentation in conjunction with Gitbook LLM (Large Language Model). In the top right corner of the Gitbook, you'll find the "Ask or Search" feature. Feel free to pose any questions you might have, such as "How to filter microbiome data in MicrobiomeStat?"

Resources

- MicrobiomeStat Repository: GitHub
- MicrobiomeStat Shiny Repository: GitHub
- Shiny Web Application: Link
- MicrobiomeStat Github Pages: Link
- Function Documentation: Link
- Gitbook Wiki: Link

www.microbiomestat.wiki (ongoing development)