Association Analysis of Microbiome Presence-Absence Data using Logic Regression

Gen Li

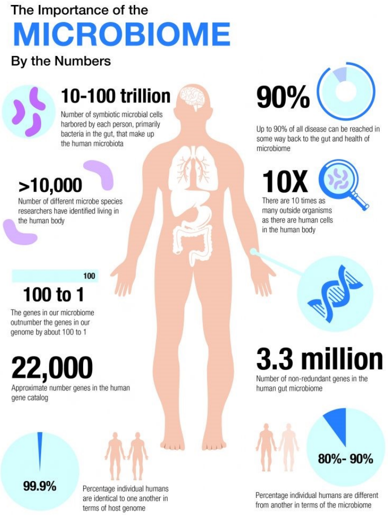
Associate Professor
Department of Biostatistics, University of Michigan

November 16, 2022

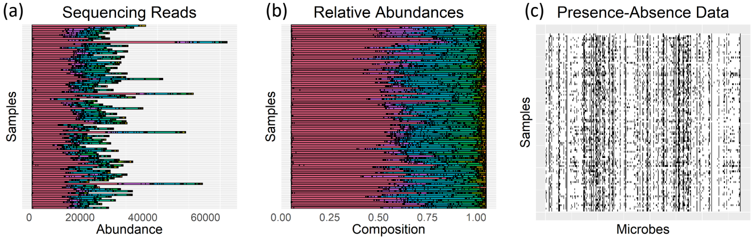
Joint work with Yiwen Chen

Prominence of Microbiome Research

- Numerous associations with health conditions
- Promising therapeutic innovations
- Potential causal effects on disease
- A burgeoning field with lots of unknowns

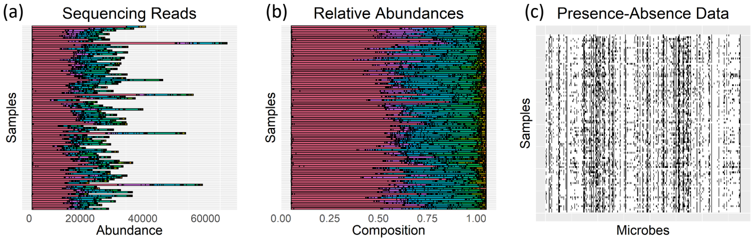

The Importance of the
MICROBIOME
By the Numbers
10-100 trillion
Number of symbiotic microbial cells harbored by each person, primarily bacteria in the gut, that make up the human microbiota
90%
Up to 90% of all disease can be reached in some way back to the gut and health of microbiome
>10,000
Number of different microbe species researchers have identified living in the human body
10X
There are 10 times as many outside organisms as there are human cells in the human body
100
100 to 1
The genes in our microbiome outnumber the genes in our genome by about 100 to 1
22,000
Approximate number genes in the human gene catalog
3.3 million
Number of non-redundant genes in the human gut microbiome
99.9%
Percentage individual humans are identical to one another in terms of host genome
80%- 90%
Percentage individual humans are different from another in terms of the microbiome

Microbiome Data



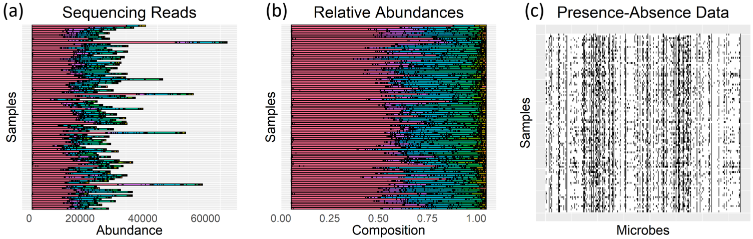
- Read counts from amplicon or metagenomic sequencing data
    - Heterogeneous sequencing depths

Microbiome Data



- Read counts from amplicon or metagenomic sequencing data
- Relative abundances after normalization
  - Compositionality; skewness; zero-inflation

Microbiome Data



- Read counts from amplicon or metagenomic sequencing data
- Relative abundances after normalization
- Dichotomized presence/absence (P/A) states

Statistical Challenges

- Compositional data analysis is tricky
  - ⋆ Non-Euclidean
- Highly skewed
  - ⋆ Dominant *vs.* rare taxa
- Excessive zeros
  - ⋆ Typically over 50%
- Tree structure
  - ⋆ Taxonomy or phylogeny
- Measurement errors
  - ⋆ Esp. at finer levels

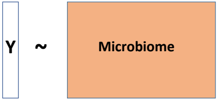
Chapman & Hall/CRC
Interdisciplinary Statistics Series
COMPOSITIONAL DATA ANALYSIS IN PRACTICE
Vera Pawlowsky-Glahn
Antonella Buccianti
Compositional Data Analysis
Theory and Applications
Modeling and Analysis of Compositional Da
VERA PAWLOWSKY-GLAHN
JUAN JOSÉ EGOZCUE
RAIMON TOLOSANA-DELGADO
STATISTICS IN PRACTICE
WIL
Peter Filzmoser
Karel Hron
Josep Antoni Martín-Fernández
Javier Palarea-Albaladejo Editors
Advances in Compositional Data Analysis
Festschrift in Honour of Vera Pawlowsky-Glahn
Springer

Benefits of P/A Analysis

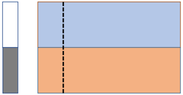
- P/A states of taxa have intrinsic health implications
  - E.g., presence of *E. coli* causes UTI
- Binary data are much easier to analyze
- More robust against measurement errors
- Better suited for rare taxa analysis
- *When conducted tactically*, P/A dichotomization may preserve almost all the abundance information*

Fundamental Tasks in Microbiome Research

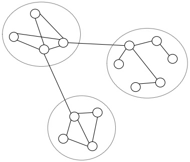
- **Regression/Association Analysis**



- Differential Abundance Analysis



- Network Inference



Regression Analysis

Novel Tree-Guided Logic Regression for P/A Data

State-of-the-Art

Log-contrast models

$$Y = \beta_0 + \beta_1 \log(X_1/X_p) + \cdots + \beta_{p-1} \log(X_{p-1}/X_p) + \varepsilon$$

or

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j \log X_j + \varepsilon, \quad \text{where } \sum_{j=1}^{p} \beta_j = 0$$

- $(X_1, \ldots, X_p)$ is the vector of relative abundances
- Inadequate to handle zeros
- Lacking straightforward biological interpretations
- Subject to measurement errors

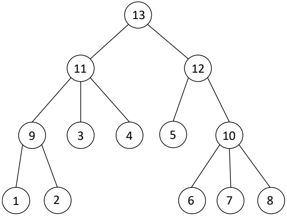
BIOSTATISTICS

P/A Dichotomized Predictors

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{k=1}^{q} \theta_k L_k + \varepsilon,$$

- $X_j \in \{0, 1\}$ is the P/A status of Taxon j
- L_k is a logic expression (Boolean operation of $X_1, \ldots, X_p$)
  - Regular interaction (e.g., $L_k = X_j X_{j'}$)
    - ✓ Pros: existing methods available (e.g., quadratic regression)
    - × Cons: undesirable interpretation; heredity constraint
  - Arbitrary logic expression (e.g., $L_k = X_1 \vee X_2 \vee X_3 \wedge X_4^c$)
    - ✓ Pros: flexible; logic regression (Ruczinski et al., 2003, JCGS)
    - × Cons: obscure biological meaning; slow to fit

Tree-Guided Logic Expression



- Phylogenetic tree or taxonomic tree
- Leaf nodes: $X_1, \dots, X_p$
- Internal nodes: $L_1, \dots, L_q$
- Only use *OR* to combine descendant leaf nodes; for example,
  - $L_9 = X_1 \vee X_2$
  - $L_{12} = X_5 \vee X_6 \vee X_7 \vee X_8$

Model Fitting

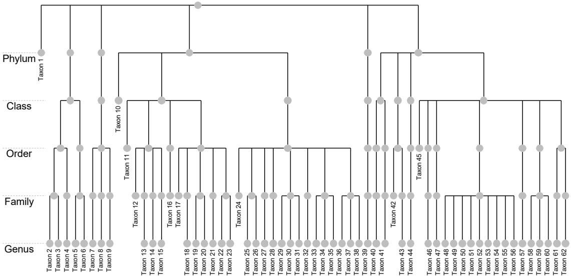
$$\min \; \|Y - \boldsymbol{X}\beta - \boldsymbol{L}\boldsymbol{\theta}\|^2 + \mathcal{P}_\lambda\left((\beta^T, \boldsymbol{\theta}^T)^T\right)$$

- $\mathcal{P}_\lambda(\cdot)$ is a sparsity-inducing penalty (e.g., LASSO)
  - ✓ Pros: ready to implement
  - × Cons: collinearity; overlapping features
  - ★ **Proposal: a bottom-up combination/selection procedure**

BIOSTATISTICS

Bottom-Up Procedure

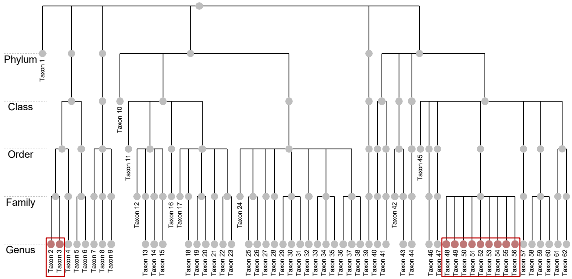
- A greedy heuristic algorithm
- First, **combine** eligible nodes that lead to the steepest decrease in BIC (tradeoff btw *goodness-of-fit* and *parsimony*)
- Once BIC stops decreasing, further **select** features



BIOSTATISTICS

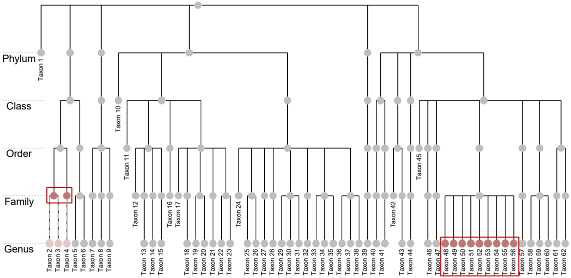
Bottom-Up Procedure

- A greedy heuristic algorithm
- First, **combine** eligible nodes that lead to the steepest decrease in BIC (tradeoff btw *goodness-of-fit* and *parsimony*)
- Once BIC stops decreasing, further **select** features



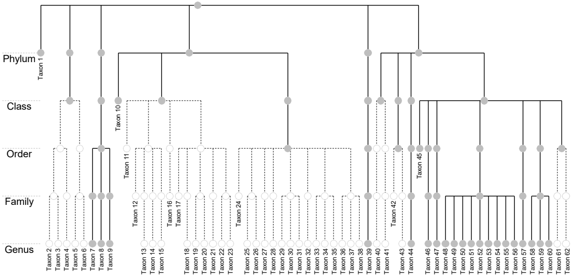
Bottom-Up Procedure

- A greedy heuristic algorithm
- First, **combine** eligible nodes that lead to the steepest decrease in BIC (tradeoff btw *goodness-of-fit* and *parsimony*)
- Once BIC stops decreasing, further **select** features



Bottom-Up Procedure

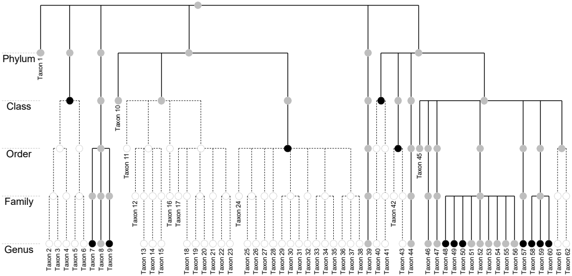
- A greedy heuristic algorithm
- First, **combine** eligible nodes that lead to the steepest decrease in BIC (tradeoff btw *goodness-of-fit* and *parsimony*)
- Once BIC stops decreasing, further **select** features



Bottom-Up Procedure

Introduction
Microbiome Data
Statistical Challenges
Regression Analysis
Logic Regression
Bottom-Up Procedure
ORIGINS Analysis
Differential Abundance Analysis
Censoring Transformation
Log-Rank Test
Summary

- A greedy heuristic algorithm
- First, **combine** eligible nodes that lead to the steepest decrease in BIC (tradeoff btw *goodness-of-fit* and *parsimony*)
- Once BIC stops decreasing, further **select** features

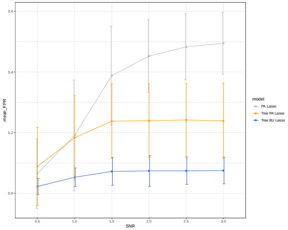


Properties

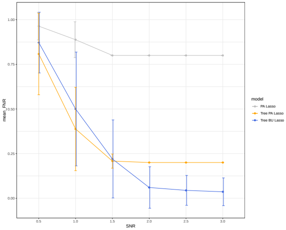
Compared to naive variable selection methods (either for leaf nodes only or for all nodes), the BU method

- Accommodates the tree structure
- Better alleviates the collinearity
- Selects fewer variables
- Has better interpretability

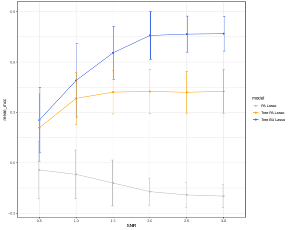
Simulation



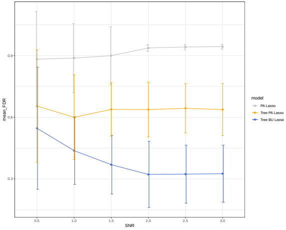
(a) FPR



(b) FNR



(c) MCC



(d) FDR

BIOSTATISTICS

ORIGINS Data Analysis

Oral Infections, Glucose Intolerance and Insulin Resistance Study (ORIGINS)

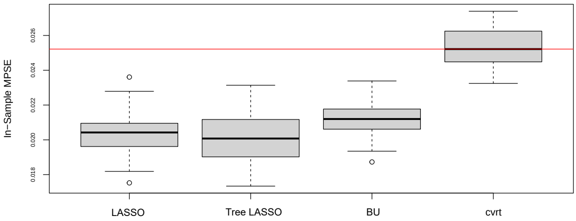
- Goal: associate oral microbiota with periodontal health
- 757 diabetes-free individuals (WAVE II)
- 16S rRNA sequencing on subgingival plaque samples
- 530 taxa at the OTU level with known taxonomic structure
- Periodontal status (percent bleeding on probing) as outcome
- Sex, age, BMI as covariates

BIOSTATISTICS

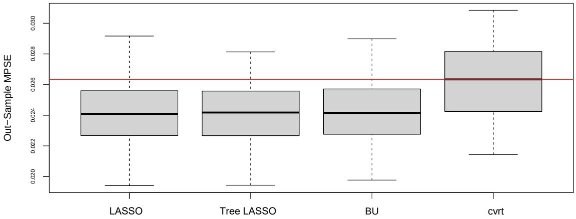
Evaluation Metric

- Randomly selected 500 samples for training and 257 for testing
- Repeated 50 times
- Compared BU (proposed), LASSO (leaf nodes only), Tree-LASSO (all nodes)
- Each time, evaluated the following
  - In-sample MSE (goodness of fit)
  - Out-sample MSE (prediction performance)
  - Selected features (interpretability)

Results: MSE

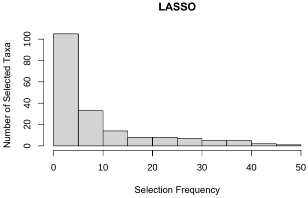


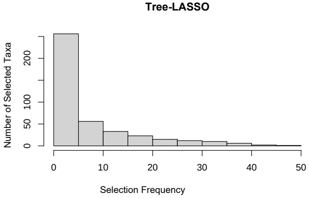
Model Fitting Comparison

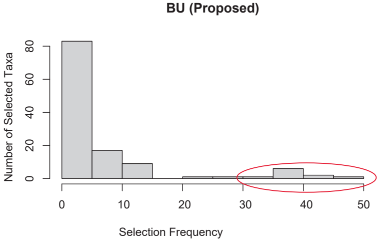


Prediction Comparison

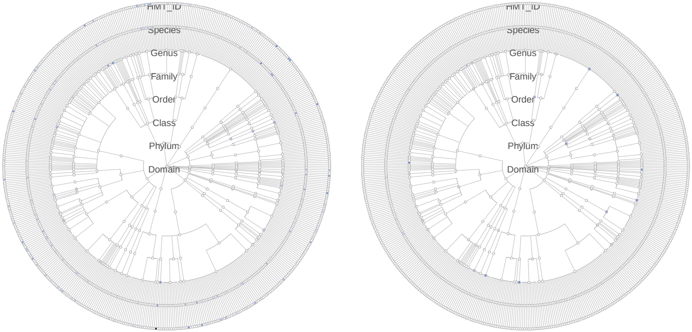
Results: Feature Selection


LASSO


Tree-LASSO


BU (Proposed)

Results: Feature Selection

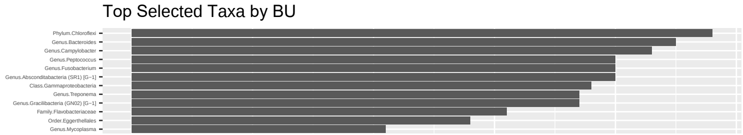


(a) Tree-Lasso

(b) BU

- BU selection is much more sparse and stable

Results: Feature Selection



Top Selected Taxa by BU

- Top selected taxa (>40%) at different taxonomic levels by BU
- Their P/A is associated with periodontal status
- The evidence in literature corroborates with our findings

Differential Abundance Analysis

Novel Transformation and Censored Data Analysis

Overview

Goal: identify differentially abundant taxa in different groups

- Parametric tests rely on zero-replacement transformations
- Nonparametric tests are not good at handling ties
- Inadequate covariate adjustment

Is there a better way to handle “0”?

Where is “0” from

- Structural zeros
  - “True” zeros
  - Absence of a taxon in a sample
- Sampling zeros
  - “Pseudo” zeros
  - Fail to detect the existence due to low abundance or insufficient sequencing depth

Where is "0" from

- Structural zeros
  - "True" zeros
  - Absence of a taxon in a sample
- Sampling zeros
  - "Pseudo" zeros
  - Fail to detect the existence due to low abundance or insufficient sequencing depth
- Both are due to **actual abundance below detection limit**

Where is "0" from

- Structural zeros
  - "True" zeros
  - Absence of a taxon in a sample
- Sampling zeros
  - "Pseudo" zeros
  - Fail to detect the existence due to low abundance or insufficient sequencing depth
- Both are due to **actual abundance below detection limit**

That's **censoring**!

Treat “0” as Censored

Assume detection limit is 1

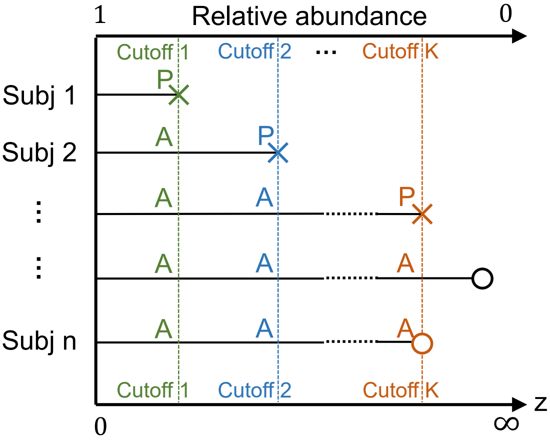
Samples			OTU			Library Size
x_1	19	1	78	0	0	m_1=98
x_2	5	0	41	2	0	m_2=48
x_1^*	19	1	78	1^-	1^-	m_1^*=100
x_2^*	5	1^-	41	2	1^-	m_2^*=50
$C(x_1^*)$	0.19	0.01	0.78	0.01^-	0.01^-	
$C(x_2^*)$	0.10	0.02^-	0.82	0.04	0.02^-	
z_1	-log(0.19)	-log(0.01)	-log(0.78)	$\{-\log(0.01)\}^+$	$\{-\log(0.01)\}^+$	
z_2	-log(0.10)	$\{-\log(0.02)\}^+$	-log(0.82)	-log(0.04)	$\{-\log(0.02)\}^+$	

Interpretation

	Microbiome Data	**Survival Data**
Type	Right-censored	Right-censored
Range	$[0, \infty)$	$[0, \infty)$
Time	Abundance cutoff (high to low)	Time duration (short to long)
Event	Presence	Death
Censoring	Zero count	Dropout
At-risk	Abundance at or below (P/A!)	Survival time at or above

BIOSTATISTICS

Visualization (for one taxon)



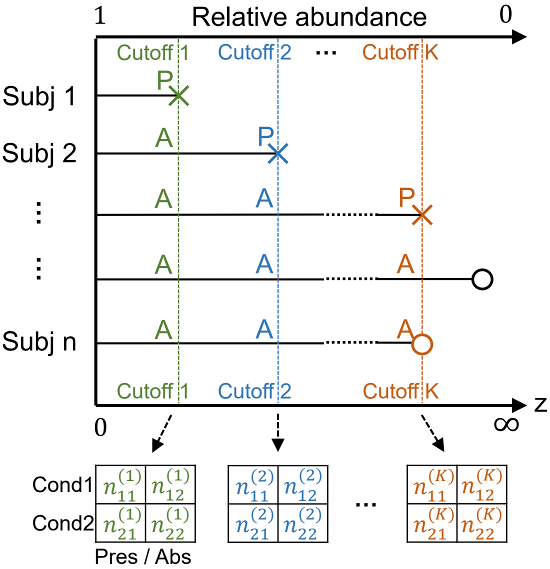
Differential Abundance Analysis

For each taxon, test for equality of distribution

$$(z_1^{(1)}, \cdots, z_{n_1}^{(1)}) \quad \text{vs.} \quad (z_1^{(2)}, \cdots, z_{n_2}^{(2)})$$

- Classic two-sample test in survival analysis
  - Without covariate: log-rank test
  - With covariates: Cox model
- No distributional assumption on z

Log-Rank Test (for one taxon)



Cond1	$n_{11}^{(1)}$	$n_{12}^{(1)}$
Cond2	$n_{21}^{(1)}$	$n_{22}^{(1)}$

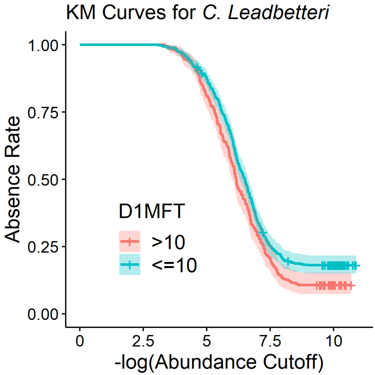
$n_{11}^{(2)}$	$n_{12}^{(2)}$
$n_{21}^{(2)}$	$n_{22}^{(2)}$

⋯

$n_{11}^{(K)}$	$n_{12}^{(K)}$
$n_{21}^{(K)}$	$n_{22}^{(K)}$

Additional Remarks

- **Better zero handling** (by aggregating P/A information across different cutoffs)
- **More powerful** in detecting differences at lower abundance levels (suitable for rare taxa comparison)
- **Highly flexible** (different variants available; Log-rank test is equivalent to the score test in Cox)



Summary

BIOSTATISTICS

Summary

P/A-based methods have untapped potential for microbiome studies

- Easier to analyze
- Less sensitive to measurement error
- Better suited for rare taxa
- New method developments
  - Interpretable regression analysis
  - Differential abundance analysis
  - Co-occurrence network inference

Thank you!

Also working on methods for **longitudinal microbiome data***.*
Interested to know more?

Contact: `ligen@umich.edu`

*Support: R01HG010731; R03DE027773; R03DE031296