NC STATE UNIVERSITY

BIOINFORMATICS RESEARCH CENTER

Phylogeny-Guided Microbiome OTU-Specific Association Test (POST)

Caizhi Huang[1], Benjamin J. Callahan[1,2], Michael C. Wu[3], Shannon T. Holloway[4], Hayden Brochu[5],, Wenbin Lu[4], Xinxia Peng[1,5], Jung-Ying Tzeng[1,4]

[1]Bioinformatics Research Center, North Carolina State University, Raleigh, 27606, USA. [2]Department of Population Health and Pathobiology, North Carolina State University, Raleigh, 27607, USA. [3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, 98109, USA. [4]Department of Statistics, North Carolina State University, Raleigh, 27606, USA. [5]Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, 27607, USA.

Abstract

Background: The relationship between host conditions and microbiome profiles, typically characterized by operational taxonomic units (OTUs), contains important information about the microbial role in human health. Traditional association testing frameworks are challenged by the **high-dimensionality**, **sparsity** and **compositionality** of typical microbiome profiles.

Methods: We propose a local collapsing test called POST. It is constructed under the **kernel machine framework** to accommodate complex OTU effects and extends kernel machine microbiome tests from community-level to OTU-level. In POST, whether or not to borrow information and how much information to borrow from the neighboring OTUs in the phylogenetic tree are supervised by **phylogenetic distance** and the **outcome-OTU association**.

Conclusions: Using POST, we show that adaptively leveraging the **phylogenetic information** can enhance the selection performance of associated microbiome features by improving the overall true positive and false positive detection. We developed a user-friendly R package ***POSTm*** which is freely available on CRAN.

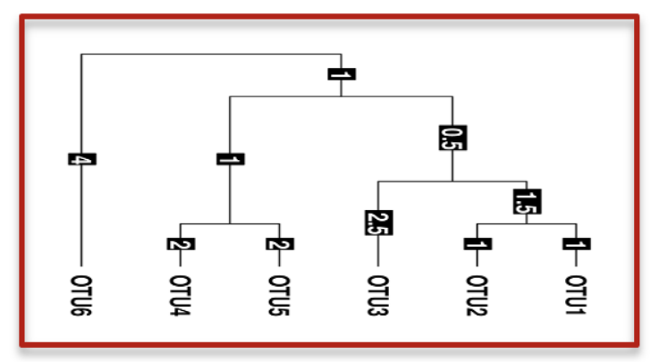
Methods

Step1: Quantify OTU Phylogenetic Correlation $r_{\ell m}$

$$r_{\ell m} = \exp\left(-\frac{d_{lm}^2}{c * s}\right)$$

- $d_{\ell m}$ is pairwise distance between OUT l and OTU m
- s is the standard deviation of $d_{\ell m}$'s
- c is a positive data-adaptive parameter that controls how fast the OTU correlation decreases when the between-OTU distance increases.

Phylogenetic Tree

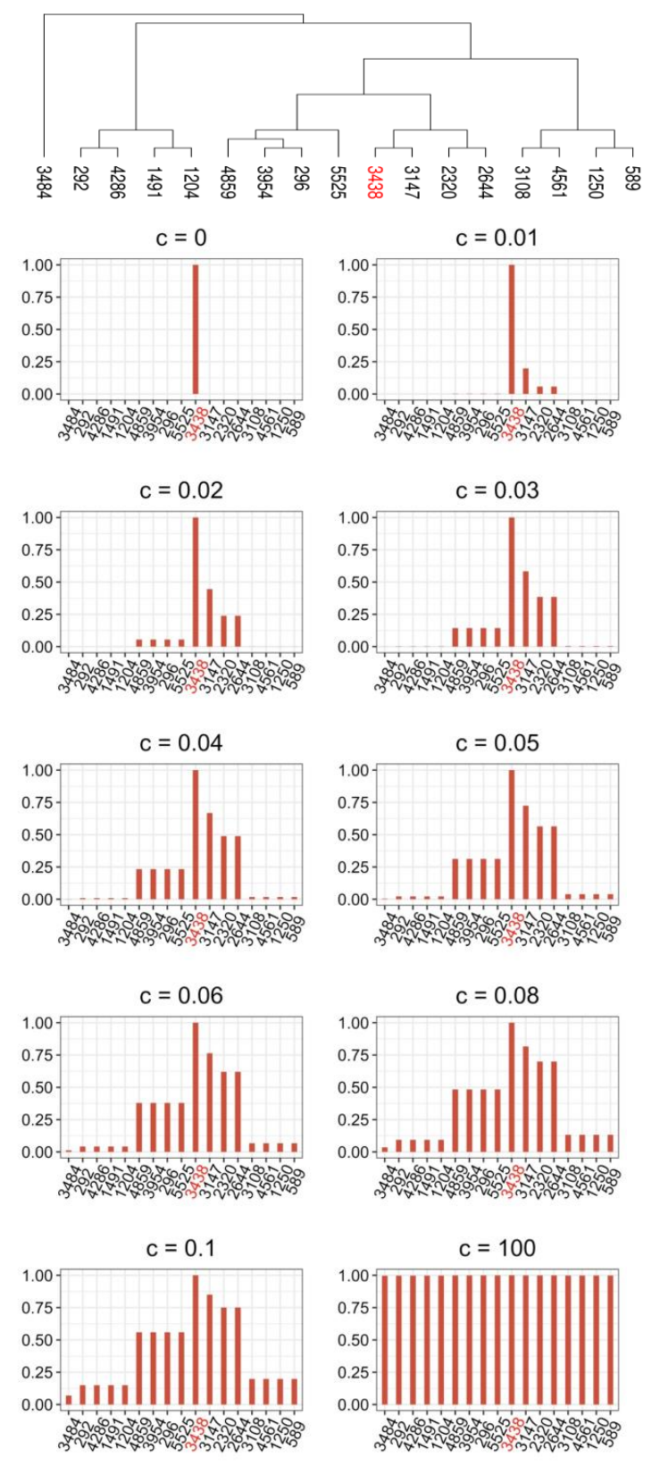


OTU Distance ($d_{\ell m}$)

	OTU1	OTU2	OTU3	OTU4	OTU5	OTU6
OTU1	0	2	5	6	6	8
OTU2	2	0	5	6	6	8
OTU3	5	5	0	6	6	8
OTU4	6	6	6	0	4	8
OTU5	6	6	6	4	0	8
OTU6	8	8	8	8	8	0

OTU Correlation ($r_{\ell m}$)

	OTU1	OTU2	OTU3	OTU4	OTU5	OTU6
OTU1	1	0.9	0.4	0.3	0.3	0.1
OTU2	0.9	1	0.4	0.3	0.3	0.1
OTU3	0.4	0.4	1	0.3	0.3	0.1
OTU4	0.3	0.3	0.3	1	0.6	0.1
OTU5	0.3	0.3	0.3	0.6	1	0.1
OTU6	0.1	0.1	0.1	0.1	0.1	1



Step2: Compute Subject Kernel Matrix K_m

OTU Table ($z_{\ell i}$)

Sub.	OTU1	OTU2	OTU3	OTU4	OTU5	OTU6
1	0	3	5	4	6	8
2	3	4	0	6	6	6
3	6	56	5	45	6	8
4	2	7	0	0	0	5
5	4	0	6	4	9	8
6	7	8	0	8	8	10

Aitchison distance of subjects i and j based on OTU m, applied on $z^*_{\ell i}$ (which is the centered log-ratio transformed $z_{\ell i}$), i.e.,

$$A^m = \begin{bmatrix} & \cdots & A^m_{ij} & \\ \vdots & \ddots & & \vdots \\ & \cdots & & \end{bmatrix} \qquad A^m_{ij} = \sqrt{\sum_{\ell=1}^{M} r_{\ell m} \times \left(z^*_{\ell i} - z^*_{\ell j}\right)^2}$$

$$K^m = -\frac{1}{2}(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}'/n)(\boldsymbol{A}^m)^2(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}'/n)$$

= Kernel matrix describing between-subject similarity based on OTU m.

Step3: Perform Association Test for OTU m

$$g(\boldsymbol{\mu}) = \boldsymbol{X\gamma} + h^m(\boldsymbol{Z})$$

- $\boldsymbol{\mu} = E(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{y}$ is $n \times 1$vector of outcome
- $\boldsymbol{X}$ is $n \times p$ covariate matrix
- $\boldsymbol{\gamma}$ is $p \times 1$ vector of covariate regression coefficients
- $\boldsymbol{Z}$ is $n \times M$ OTU table with M OTUs in total
- $h^m(\cdot)$ characterizes the effect of OTU m and can be specified using kernel method, i.e., $h^m(z_i) = \sum_{j=1}^{n} \alpha_j^m k^m(z_i, z_j)$

The association between OTU m and the outcome can be evaluated by testing

$$H_0: h^m(\boldsymbol{Z}) = 0$$

The above test is equivalent to the variance component test $H_0: \tau = 0$ giving $h^m(\boldsymbol{Z}) \sim N(0, \tau \boldsymbol{K}^m)$ with the following test statistic

$$T^{m.c} = \frac{1}{2\hat{\phi}}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)^T \boldsymbol{K}^m (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)$$

- $\hat{\boldsymbol{\mu}}_0 = g^{-1}(\boldsymbol{X}\hat{\gamma})$ with $\hat{\gamma}$ the estimated covariate coefficient under H_0
- $\hat{\phi}$ is the estimator of the dispersion parameter under H_0
- $T^{m.c}$ asymptotically follows a weighted mixtures of $\chi^2_{(1)}$ distribution under H_0

We consider A grid of $c \in \{c_1, \ldots, c_j\}$ based on simulation and use Cauchy combination method to aggregate p-values of different c's

$$p_m = \frac{1}{2} - \{\arctan(T^m/J)\}/\pi, \text{ where } T^m = \sum_{J=1}^{J} \tan\{(\tfrac{1}{2} - p_{m,c_j})\pi\}$$

Simulation Analysis

1. Modeling OTU counts data using Dirichlet-multinomial distribution

- Get the multinomial parameter.

$(p_{i1}, p_{i2}, \ldots, p_{iM}) \sim Dirichlet(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_M)$

- Get the counts data

$(Z_{i1}, Z_{i2}, \ldots, Z_{iM}) \sim Multinomial\ (p_{i1}, p_{i2}, \ldots, p_{iM}, n_i\ )$

2. Generating outcome with two approaches

- Case-control (Simulation A; Only Binary)
- Resampling procedure (Simulation B)

Continuous: $y_i \sim N(\eta_i, 1)$

Binary: $y_i \sim Bernoulli(\Pi_i)$, $\Pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$

- $\eta_i = 0.5\omega \times (\text{scale}(x_{1i}) + \text{scale}(x_{2i})) + \sum_{m=1}^{M} \beta_m \times \text{scale}(z_{im})$
- $\omega = 1\ or\ 0$ is a parameter controlling if there are covariate
- x_{1i} and x_{2i} are covariates that related or not related to causal OTUs, respectively.

3. Select causal OTU with five scenarios (**Figure 1**)

- Scenarios 1 to 3: consider larger "causal hubs", each containing about 7-10 causal OTUs
- Scenario 4 considers smaller causal hubs of 2-3 causal OTUs
- Scenario 5 considers causal OTUs with random positions in the phylogenetic tree.

Table 1: Type I error rates at the significance levels of 0.05, 0.01, and 0.001 for POST.

Simulation	Outcome	a=0.05	a=0.01	a=0.001	a=0.0001
A	Binary	0.047	0.007	0.0006	0.00005
B	Continuous	0.051	0.010	0.0010	0.00008
	Binary	0.047	0.008	0.0006	0.00007

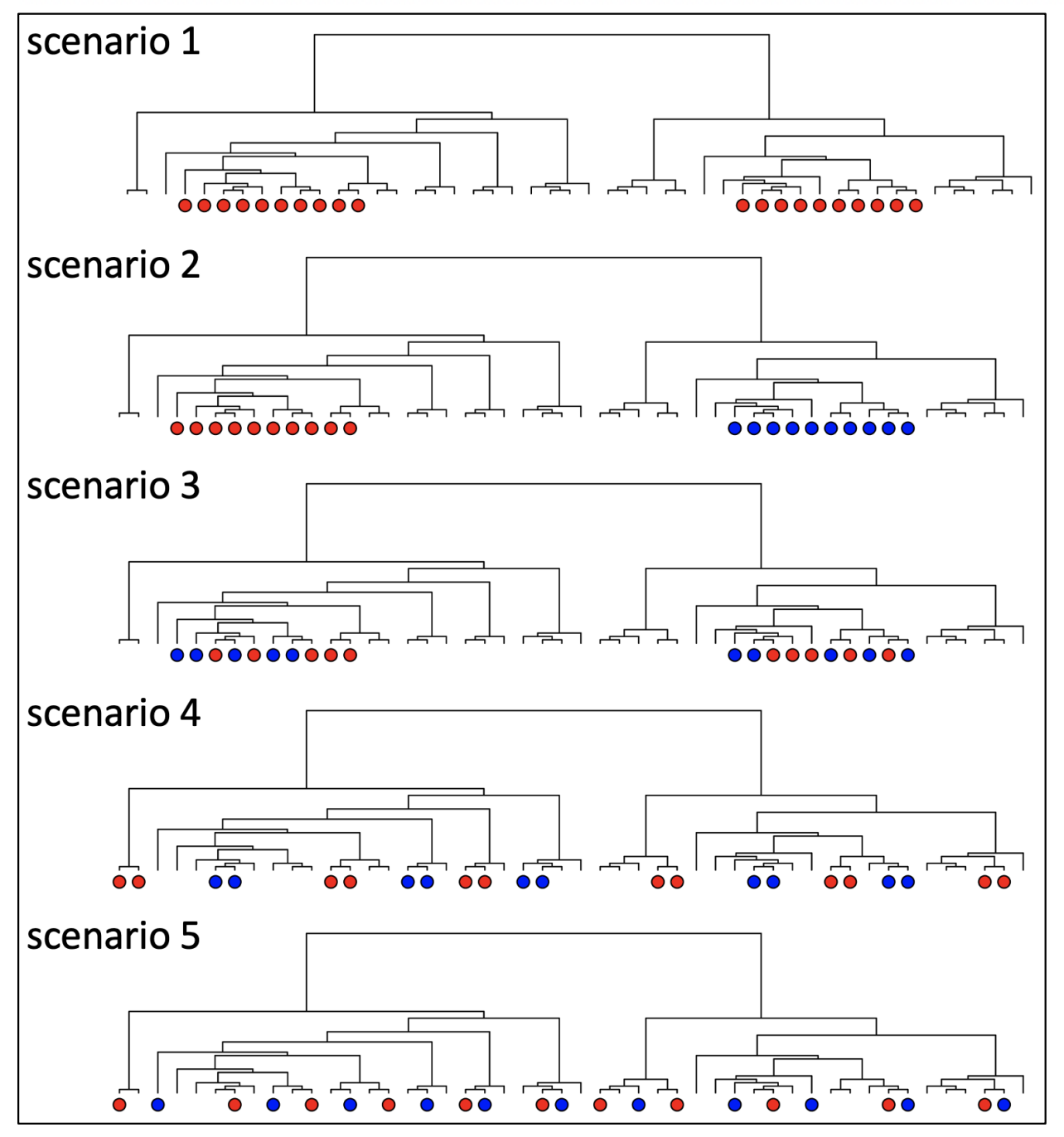


Figure 1: Illustration of the five causal-OUT scenarios. Red (blue) circles indicate + (-) causal effect.

Table 2: AUC of different methods in simulations A and B. Methods considered include POST, TreeFDR (TF), Single-OTU test (SO), DESeq2 (DE), ANCOM-BC (AB) and LinDA (LD), Wilcoxon rank-sum test for binary outcomes, and Spearman correlation test for continuous outcomes. Methods with the highest AUC are in red.

Simulation	Simulation A								Simulation B															
Outcome	Binary outcome								Continuous outcome								Binary outcome							
Method	POST	TF	SO	DE	AB	LD	WR-P	WR-R	POST	TF	SO	DE	AB	LD	SC-P	SC-R	POST	TF	SO	DE	AB	LD	WR-P	WR-R
Scenario1	**0.80**	0.78	0.62	0.69	0.54	0.56	0.51	0.51	**0.64**	0.52	0.58	0.61	0.57	0.61	0.57	0.61	0.58	0.53	0.55	0.52	0.57	**0.59**	0.54	0.55
Scenario2	**0.88**	0.77	0.71	0.71	0.70	0.67	0.62	0.58	**0.66**	0.50	0.62	0.65	0.62	0.62	0.57	0.51	0.60	0.48	0.56	0.57	**0.61**	0.56	0.52	0.51
Scenario3	**0.88**	0.78	0.72	0.70	0.64	0.64	0.60	0.60	**0.70**	0.49	0.62	0.63	0.61	0.62	0.58	0.51	**0.63**	0.50	0.59	0.55	0.58	0.58	0.57	0.51
Scenario4	**0.78**	0.75	0.70	0.69	0.70	0.67	0.62	0.58	**0.63**	0.48	0.61	**0.63**	0.58	0.61	0.58	0.52	**0.56**	0.49	**0.56**	0.54	0.55	0.55	0.55	0.51
Scenario5	0.66	**0.72**	0.67	0.64	0.65	0.63	0.57	0.55	0.59	0.48	0.60	**0.61**	0.59	0.60	0.54	0.51	0.54	0.50	0.54	0.52	**0.57**	0.55	0.53	0.50

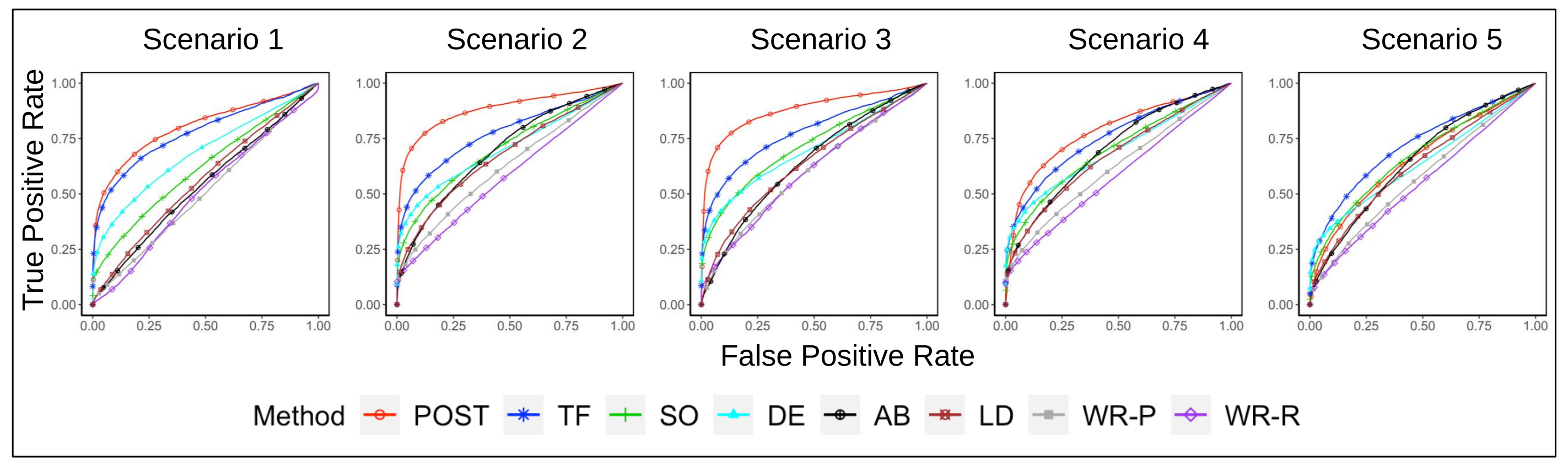


Figure 3: ROC curves of Simulation A with large effect size for POST, Single-OTU test (SO), TreeFDR (TF), DESeq2 (DE), ANCOM-BC (AB), LinDA (DA) and Wilcoxon rank-sum test (WR) under the 5 causal scenarios.

Real Data Analysis

Table 2: OTUs significantly associated with bacterial vaginosis (BV) at FDR level of 0.05.

OTU	FDR-adjusted p-value								Detected method	Genus/Species	Direction**
	POST	TF	SO	DE	AB	LD	WR-P*	WR-R*			
OTU3	0.039	0.089	0.045	0.002	0.006	0.041	0.178	0.025	POST/SO/DE/AB/LD/WR-R	*Lactobacillus crispatus*	-
OTU90	0.039	0.245	0.953	0.950	0.983	0.961	0.375	0.989	POST	*Lactobacillus sp.*	+
OTU7	0.039	0.653	0.881	0.147	0.766	0.729	0.760	0.932	POST	*Lactobacillus jensenii*	-
OTU82	0.039	0.517	0.985	0.970	0.963	0.976	0.258	0.989	POST	*Lactobacillus gasseri*	-
OTU66	0.039	0.544	0.701	0.027	0.379	0.615	0.378	0.932	POST/DE	*Lactobacillus sp.*	-
OTU2	0.039	0.180	0.917	0.876	0.766	0.923	0.235	0.989	POST	*Lactobacillus iners*	+
OTU58	0.039	0.839	0.917	0.876	0.871	0.923	0.791	0.989	POST	*Lactobacillus sp.*	+
OTU112	0.411	0.046	0.701	0.422	0.379	0.678	0.040	0.932	TF/WR-P	*Peptoniphilus sp.*	+
OTU85	0.470	0.046	0.906	0.372	0.766	0.870	0.097	0.989	TF	*Gemella sp.*	+
OTU11	0.918	0.286	0.943	0.000	0.869	0.923	0.220	0.989	DE	*Prevotella sp.*	+
OTU12	0.391	0.089	0.701	0.013	0.338	0.615	0.040	0.942	DE/WR-P	*Prevotella sp.*	+
OTU16	0.680	0.155	0.881	0.001	0.766	0.918	0.097	0.989	DE	*Prevotella timonensis*	+
OTU91	0.680	0.092	0.881	0.505	0.766	0.852	0.040	0.989	WR-P	*Prevotella sp.*	+

*: WR-P and WR-R did not adjust for race.

**: + (and −) indicates that the OTU is positively (and negatively) associated with BV risk from a logistic regression.

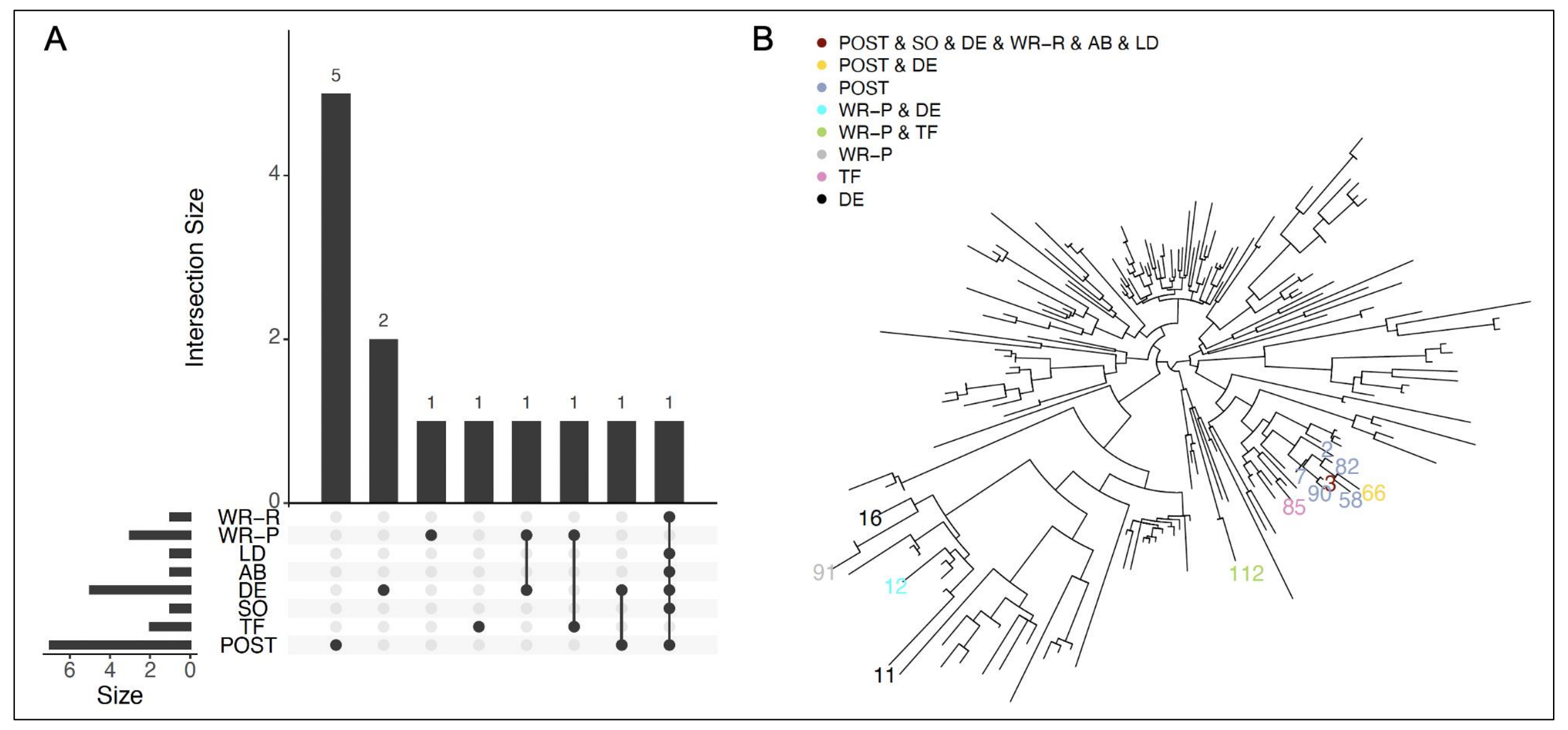


Figure 4: Upset plot of detected OTUs at FDR level of 0.05 and phylogenetic trees of the analyzed OTUs with detected OTUs for bacterial vaginosis study.

POSTm R Package

post *Phylogeney-Guided OTU-Specific Association Test for Microbiome Data*

Usage

```
post(
  y,
  OTU,
  tree = NULL,
  X = NULL,
  cValues = seq(from = 0, to = 0.05, by = 0.01)
)
```

Arguments

y	A numerical vector. The outcome of interest. Data can be binary or continuous.
OTU	A matrix object. The operational taxonomic units (OTU). Data can be provided as counts or as proportions. Each row indicates a single sample; each column a single OTU. NA/0 values are allowed, but their presence will trigger a shift of all data by a small internally defined value. The matrix must include column headers providing unique identification for each OTU; these identifiers are expected to be included in the tip labels of the input tree object. Any identifiers that are not included as tip labels are removed from the analysis.
tree	An object of class "phylo", "hclust", "phylog", a matrix object or NULL. If NULL, only the single OTU test will be estimated. Objects of class "phylo", "hclust", and "phylog" are phylogenetic trees, the tip labels of which must include all of the identifiers used as column headers of OTU. If a matrix, a square symmetric matrix containing the pairwise distances between OTUs as defined by the branch lengths. Note that the full tree should be provided/used and should not be subset or truncated, even if OTU does not contain all tips. See details for further information.
X	A data.frame object, matrix object or NULL. The covariates data. If NULL, an intercept only model is assumed. Factor covariates are allowed.
cValues	A numeric vector. The c values at which p-values are to be estimated. The default is a vector of evenly spaced values between zero and the recommended maximum value for OTUs defined at 97% sequence similiary, c_max = 0.05. If no tree is provided, cValues will be set to 0.

Reference

This work is published at **Huang, Caizhi, et al. "Phylogeny-guided microbiome OTU-specific association test (POST)." *Microbiome* 10.1 (2022): 1-15.**

JSM 2022