A PHYLOGENETIC TRANSFORM ENHANCES ANALYSIS OF COMPOSITIONAL MICROBIOTA DATA

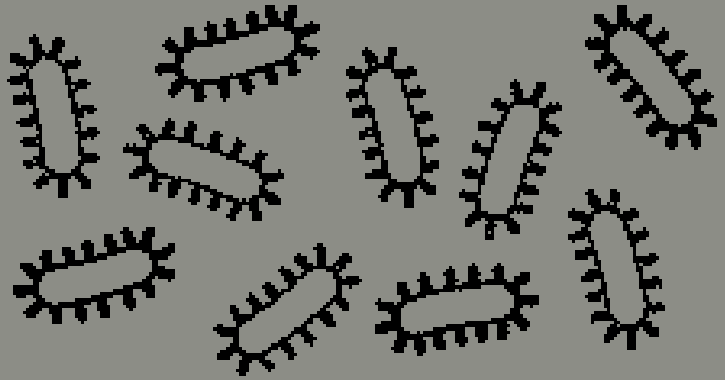
Justin D Silverman, Alex D Washburne,
Sayan Mukherjee, Lawrence A David

Duke University

Presented by Roee Wodislawski

eLife 2017;6:e21887

OUTLINE



A PHYLOGENETIC TRANSFORM ENHANCES ANALYSIS OF COMPOSITIONAL MICROBIOTA DATA

- ➢ Introduction
- Challenges
- Method
- Results
- Benchmarks
- Implementation
- Summary & Conclusions

Introduction - Microbiota Research

- Identifying relationships between bacterial taxa or microbes and their environment.

- Mostly analyzed by relative abundance of bacterial taxa, whose compositional nature can lead to spurious statistical analyses.

- Instead of rederiving statistical tools, the compositional data can be transformed to another space where existing statistical models can be applied as-is.

- Wait a second… doesn't all that sound familiar?

Introduction - CoDA Theory

- CoDA stands for Compositional Data Analysis

- Sample space

$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, x_2, \dots, x_D] \in \mathbb{R}^D \,\middle|\, x_i > 0, i = 1, 2, \dots, D; \sum_{i=1}^{D} x_i = \kappa \right\}$$

- Closure

$$\mathcal{C}[x_1, x_2, \dots, x_D] = \left[\frac{x_1}{\sum_{i=1}^{D} x_i}, \frac{x_2}{\sum_{i=1}^{D} x_i}, \dots, \frac{x_D}{\sum_{i=1}^{D} x_i} \right]$$

- Centered log-ratio transform

$$\mathrm{clr}(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_D}{g(x)} \right]$$

isomorphic and isometric, inverse is given by the softmax function.

Introduction - CoDA Theory

- Centered log-ratio transform cont.

$$\mathrm{clr}(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_D}{g(x)}\right]$$

$$log(g(x)) = log\left(exp\left[\frac{1}{n}\sum log(x_i)\right]\right) = \mathbb{E}(log(x))$$

$$\sum log\frac{x_i}{g(x)} = \sum \left[log(x_i) - \mathbb{E}(log(x))\right] = 0$$

subcompositionally dominant, but not subcompositionally coherent as the covariance matrix is singular.

- Isometric log-ratio transform

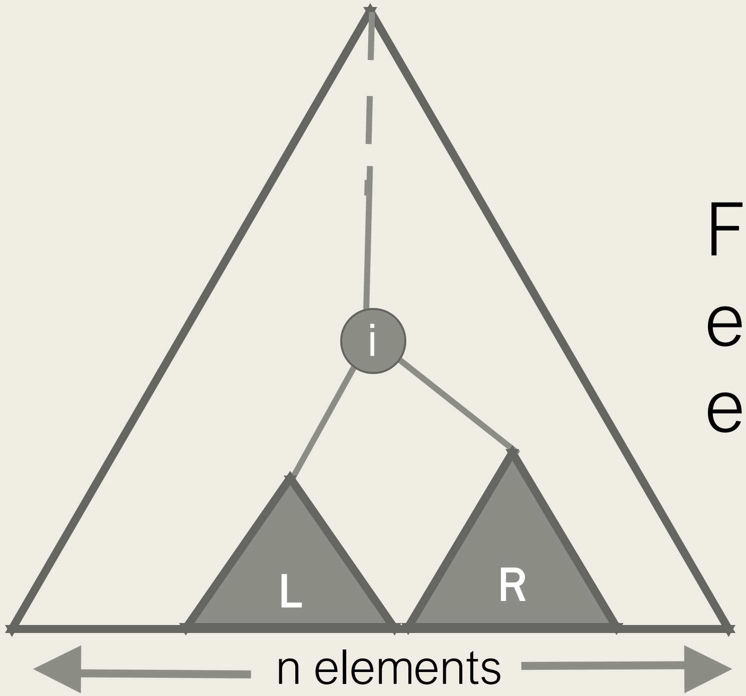
$$ilr(x)_i = \sqrt{\frac{l \cdot r}{l + r}}\, log\, \frac{g(x_L)}{g(x_R)}$$

defines the i-th element in the ILR transformed vector for a specific sequential binary partition where R and L are the two subdivisions created by the i-th division.

Introduction - CoDA Theory

- Isometric log-ratio transform cont.

$$ilr(x)_i = \sqrt{\frac{l \cdot r}{l + r}}\ log\ \frac{g(x_L)}{g(x_R)}$$

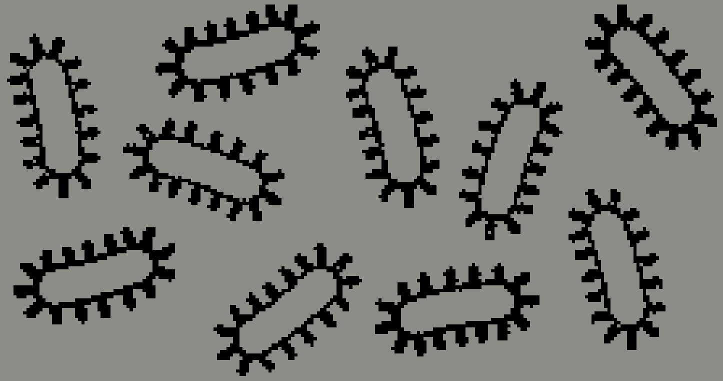


For a relative abundances vector with n elements, ILR calculates a vector of n – 1 elements, each called a **balance.**

isomorphic, isometric and subcompositionally coherent (there are orthonormal bases).

It can be written in terms of the CLR transform and thus inverted in a similar manner.

OUTLINE

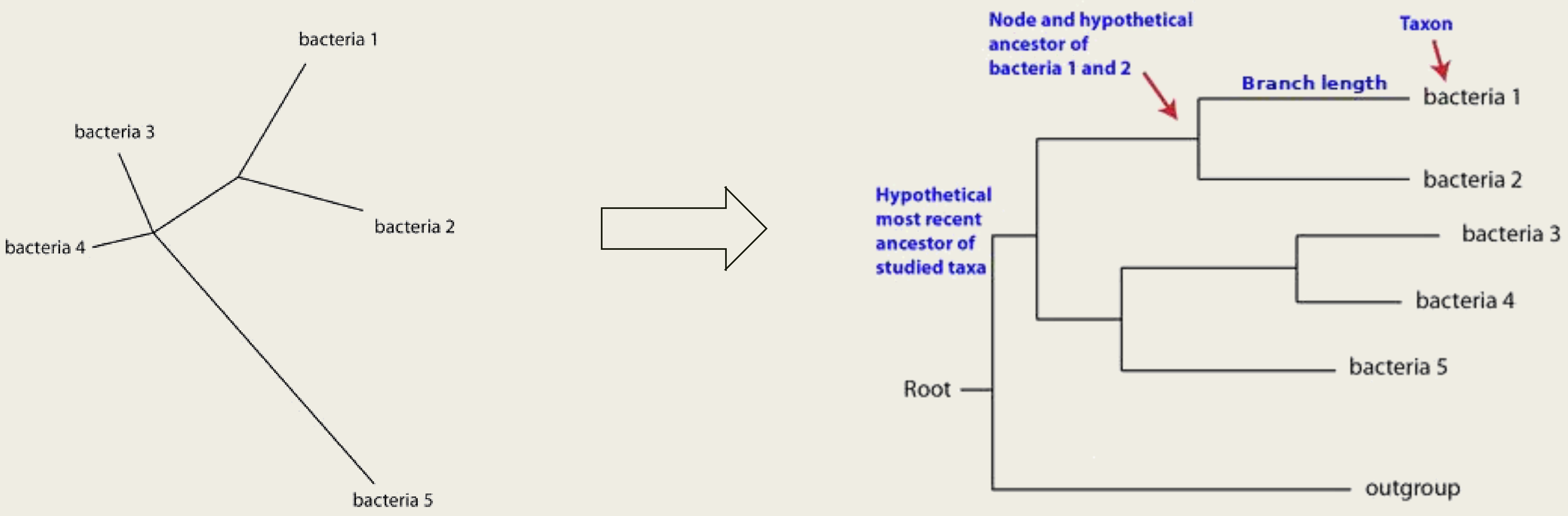


A PHYLOGENETIC TRANSFORM ENHANCES ANALYSIS OF COMPOSITIONAL MICROBIOTA DATA

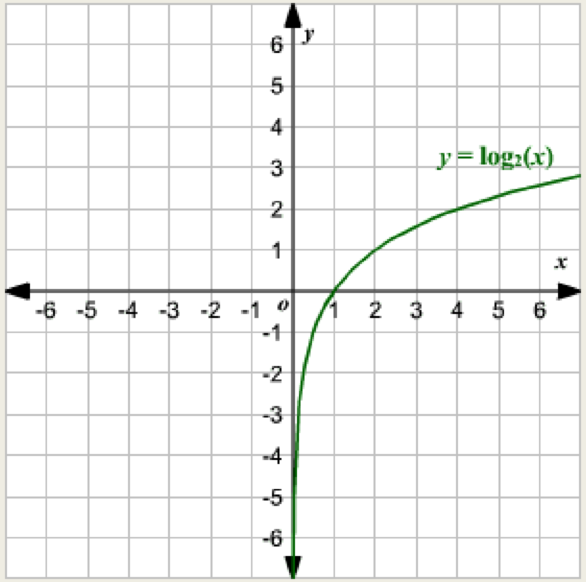
- ✓ Introduction
- ➢ Challenges
- Method
- Results
- Benchmarks
- Implementation
- Summary & Conclusions

Challenges - Partition

- ILR transform requires choice of a sequential binary partition.
- Resulting coordinates should be meaningful.
- Bacterial phylogenetic tree is a natural and informative partition.
- Branch lengths indicate genetic change (avg. nucleotide substitutions per site).



Challenges - Zero Values



- Zero values cause issues with computing log ratios.

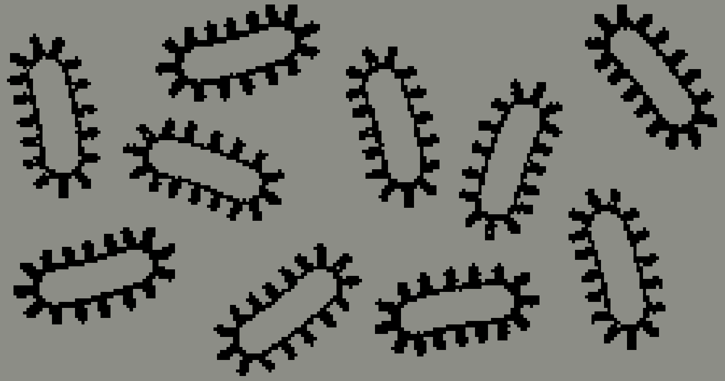
- Zero-replacement can fix that, but it may introduce bias.

- Taxa with many zero and near-zero counts are less reliable.

- Hard filtering thresholds may remove a substantial fraction of observed taxa.

- Weights will be attached to individual taxa in order to obtain soft-thresholding.

OUTLINE



A PHYLOGENETIC TRANSFORM ENHANCES ANALYSIS OF COMPOSITIONAL MICROBIOTA DATA

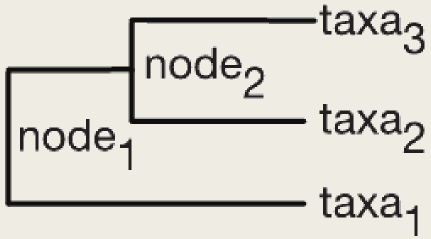
- ✓ Introduction
- ✓ Challenges
- ➢ Method
- Results
- Benchmarks
- Implementation
- Summary & Conclusions

Method - Weighted ILR Transform

Based on Egozcue and Pawlowsky-Glahn, 2016

counts c_j for taxa $j \in \{1,\ \ldots,\ D\}$ $\Rightarrow$ $\boldsymbol{x} = C[(c_1,\cdots,c_D)] = \left(\frac{c_1}{\sum_j c_j},\ldots,\frac{c_D}{\sum_j c_j}\right)$

positive weights, $\mathbf{p} = (p_1, p_2, \ldots, p_D)$ $\Rightarrow$ shifted composition $\boldsymbol{y} = \boldsymbol{x}/\boldsymbol{p} = (x_1/p_1,\ldots,x_D/p_D)$



$\Rightarrow$

$$\Theta \quad \begin{array}{c|ccc} & \text{taxa}_1 & \text{taxa}_2 & \text{taxa}_3 \\ \hline \text{node}_1 & +1 & -1 & -1 \\ \text{node}_2 & 0 & +1 & -1 \end{array}$$

$$y_i^* = \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} \log \frac{g_p(\boldsymbol{y}_i^+)}{g_p(\boldsymbol{y}_i^-)} \qquad g_p(\boldsymbol{y}_i^\pm) = \exp\left(\frac{\sum_{(\theta_{ij}=\pm 1)} p_j \log y_j}{\sum_{(\theta_{ij}=\pm 1)} p_j}\right) \qquad n_i^\pm = \sum_{\theta_{ij}=\pm 1} p_j$$

All discussed properties of original transform are met!

Method - Taxa Weighting

- Central tendency of counts for a single taxon can be measured in multiple ways.

- Geometric mean with a pseudocount of one outperformed both arithmetic mean and median as a measure of central tendency.

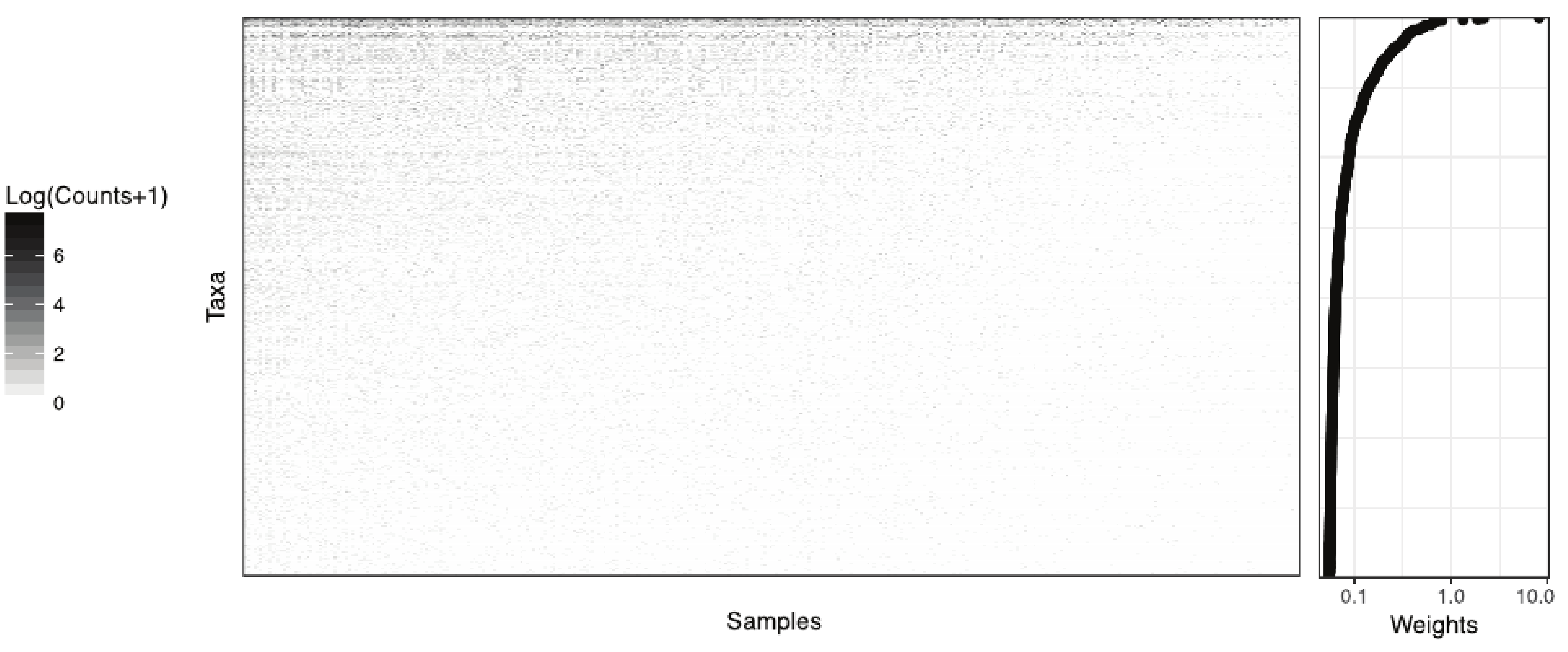
- Given a single taxon, Euclidean norm of the relative abundances vector (across all samples) captures its site-specifity.

- Inclusion of the Euclidean norm together with the geometric mean for taxa weights calculation has shown benchmark performance improvements.

$$p_j = \sqrt[N]{(c_{j1}+1)\cdot\ldots\cdot(c_{jN}+1)}\ \cdot \left\|x_j\right\|$$

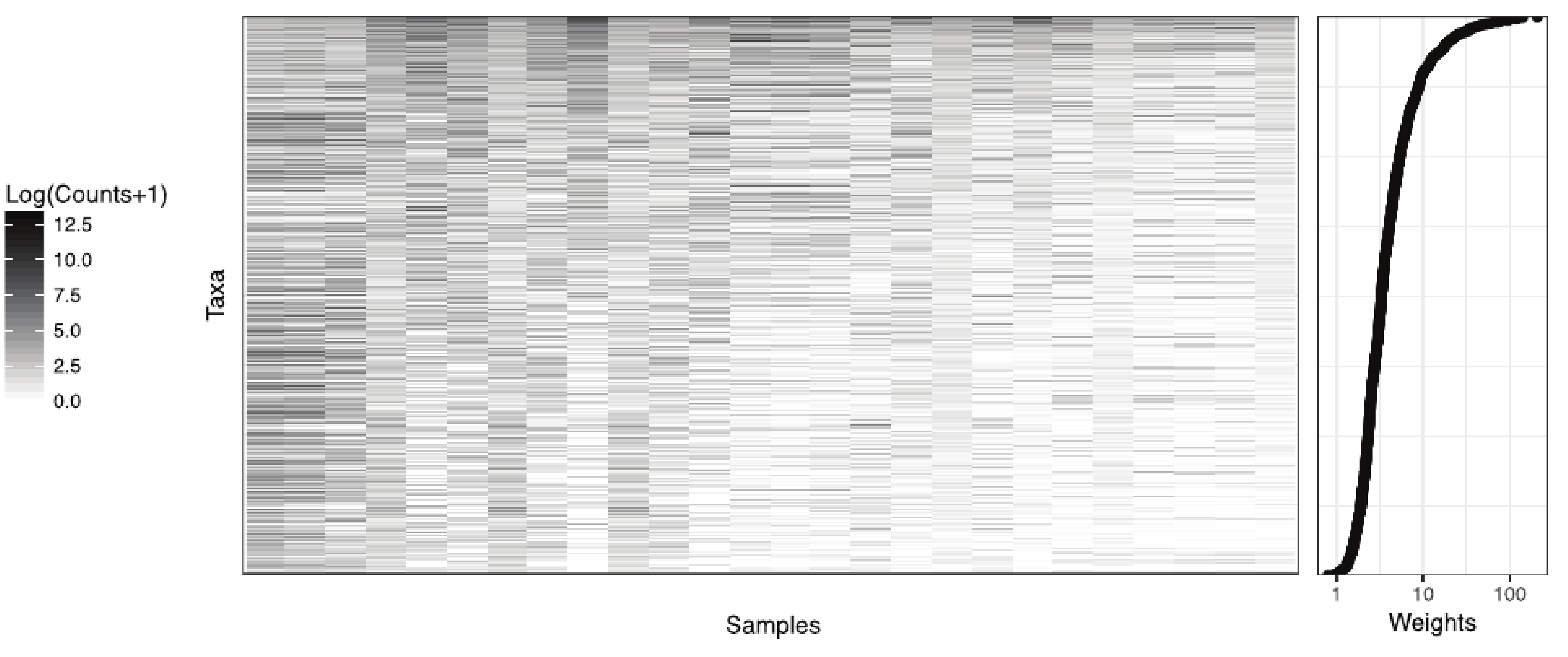
Method - Taxa Weighting

Visualization



Method - Taxa Weighting

Visualization Cont.



Method - Branch Length Weighting

- Incorporating information on evolutionary distances between taxa.

- ILR balances will be linearly scaled using the phylogenetic distance between the relevant neighboring clades:

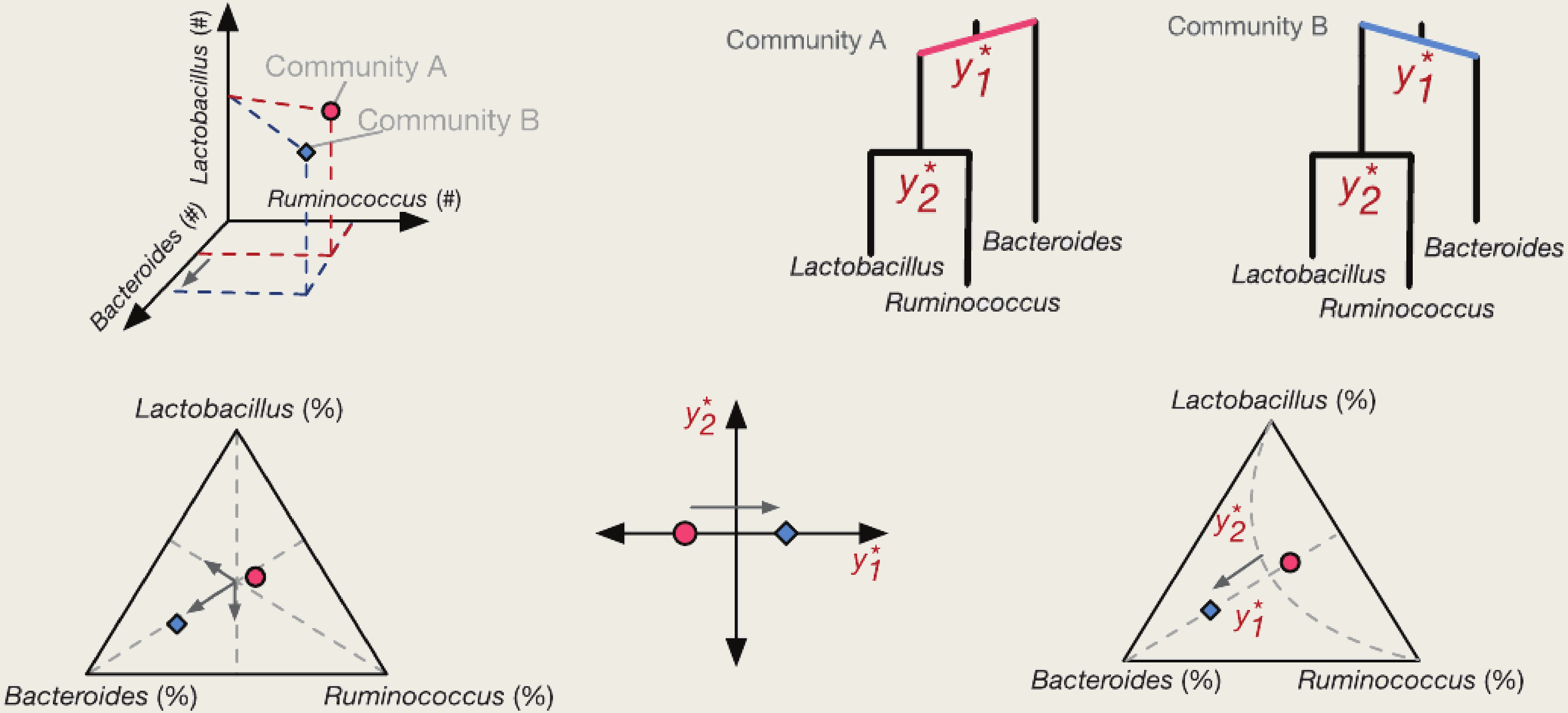
$$y_i^{*,blw} = y_i^* \cdot f\left(d_i^+, d_i^-\right)$$

- What function should be used?

$$f\left(d_i^+, d_i^-\right) = d_i^+ + d_i^- \qquad \boxed{f\left(d_i^+, d_i^-\right) = \sqrt{d_i^+ + d_i^-}} \qquad f\left(d_i^+, d_i^-\right) = 1$$

Method - Summary

y_1^* Balance of *Bacteroides* to *Ruminococcus* and *Lactobacillus*
y_2^* Balance of *Ruminococcus* to *Lactobacillus*



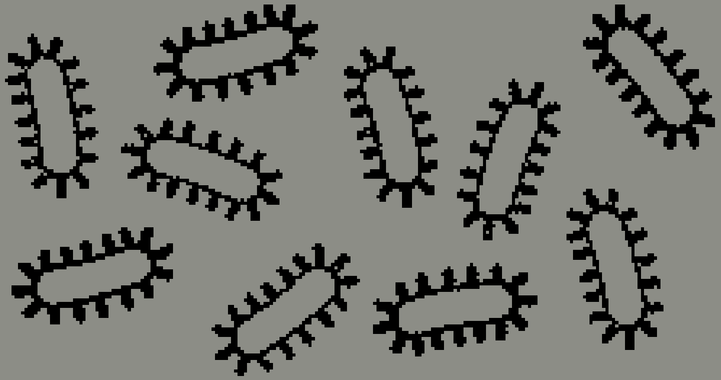
Method - Summary

ILR transformation with the phylogenetic tree as the choice of sequential binary partition, soft thresholding through taxa weights and phylogenetic distances embedded to incorporate evolutionary information:

$$y_i^* = \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} \log \frac{g_p(\mathbf{y}_i^+)}{g_p(\mathbf{y}_i^-)} \sqrt{d_i^+ + d_i^-}$$

Phylogenetic ILR

OUTLINE



A PHYLOGENETIC TRANSFORM ENHANCES ANALYSIS OF COMPOSITIONAL MICROBIOTA DATA

- ✓ Introduction
- ✓ Challenges
- ✓ Method
- ➢ Results
- • Benchmarks
- • Implementation
- • Summary & Conclusions

Results - Datasets

- Usage of previously published OTU tables, taxonomic classification and phylogenies in environmental and human-associated 16S rRNA datasets.

- Datasets:
  - *Human Microbiome Project (HMP - 4743 samples)*
  - *Global Patterns (GP – 26 samples)*
  - *Costello Skin Sites (CSS – 357 samples)*

- Preprocessing:
  - OTU tables filtering
  - Phylogeny pruning and outgroup rooting
  - Minimal pseudocount

Results - Site-Distinguishing Balances

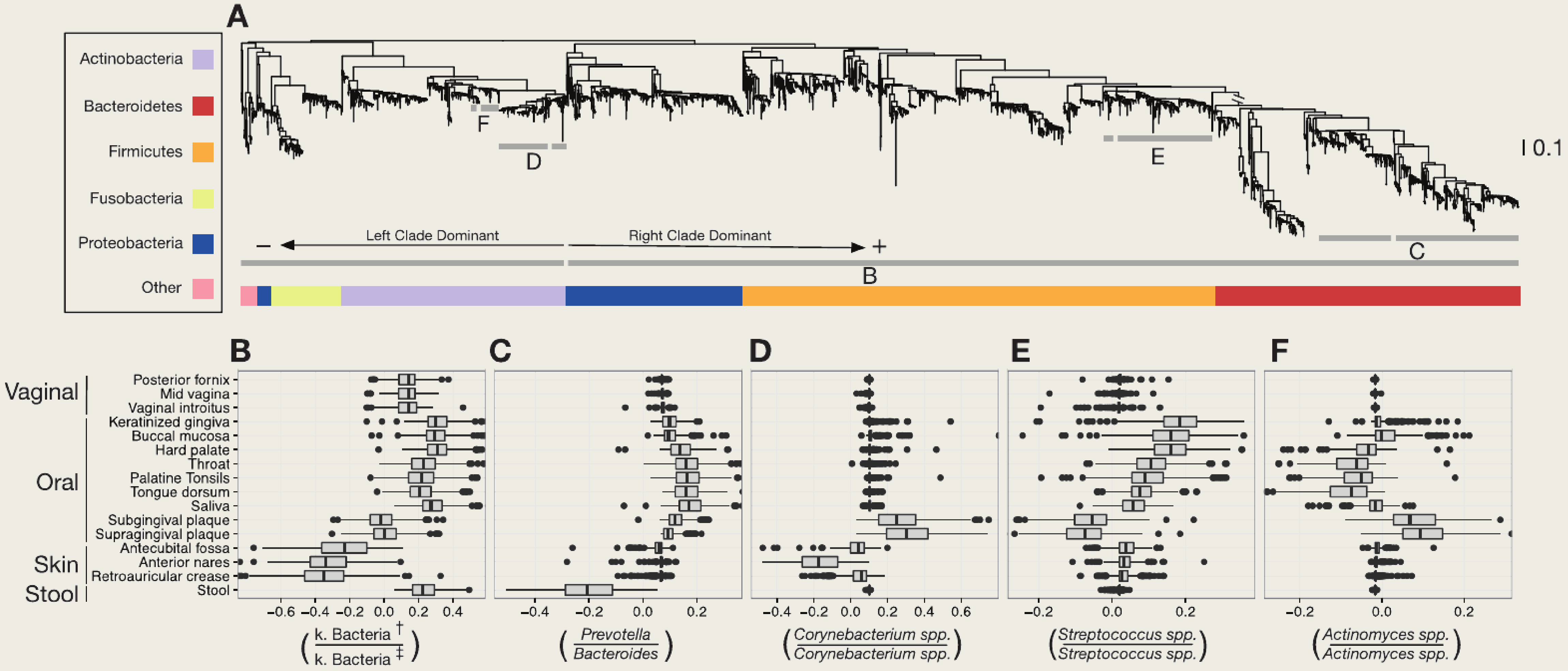
- Neighboring bacterial clades that differ by body site preference are interesting, as they may reflect functional specialization or adaptation to new environments.

- Investigating how distances between communities de-compose along PhILR balances can highlight which balances distinguish microbiota by site.

- Sparse logistic regression was used to identify such balances, for which the regression coefficients should be non-zero.

Results - Site-Distinguishing Balances

HMP Dataset Analysis



Results - Site-Distinguishing Balances

- Several balances were found to distinguish different sites.
- These balances indicate which neighboring clades may have adapted to human body site environments.
- While there were discriminatory ancestral balances, more recent balances (closer to the tips) also separated sites in nearby habitats.

Results - Balance Variance and Depth

- Balance variance is a measure of association (covariation) between neighboring bacterial clades.

- The relationship between balance variance and phylogenetic depth will be investigated using (log-log) linear regression:

$$\log var(y^*) = \beta \log d + \alpha$$

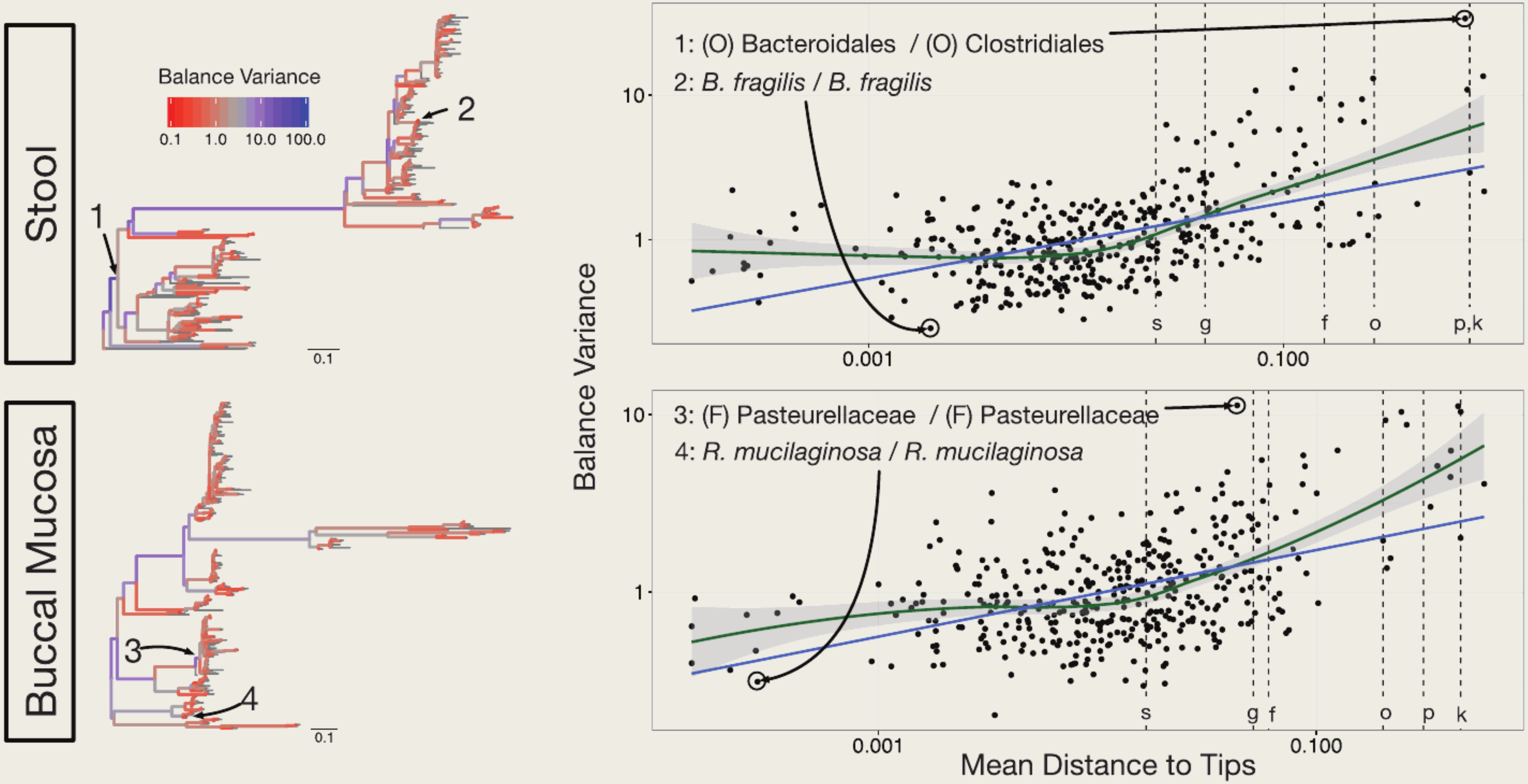
- The regression is combined with a permutation scheme of the tree tips to test the null hypothesis that β = 0 (covariation is independent of depth).

- Both weighting schemes were omitted, together with more stringent filtering thresholds, to validate the results.

Results - Balance Variance and Depth

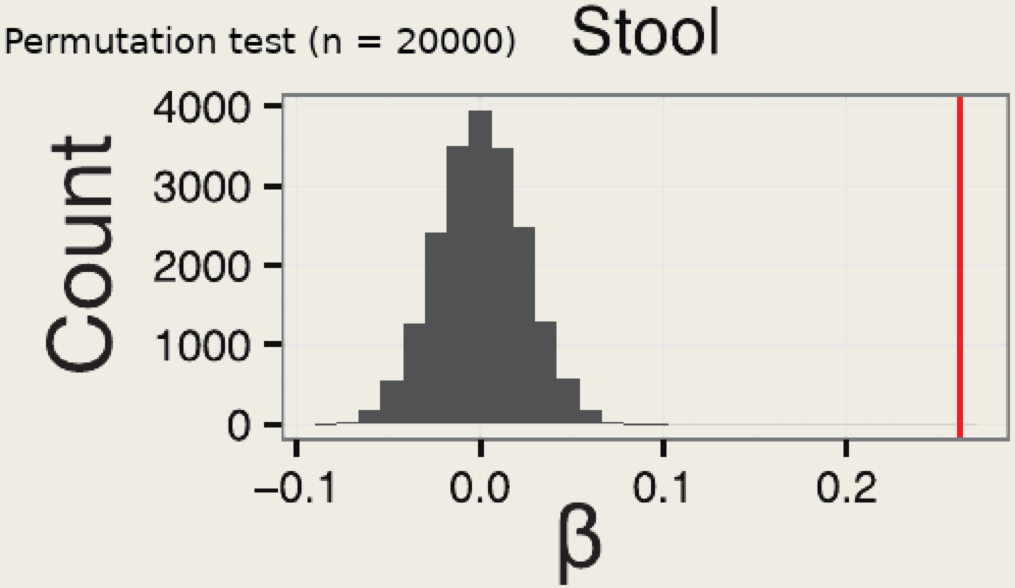
HMP Dataset Analysis

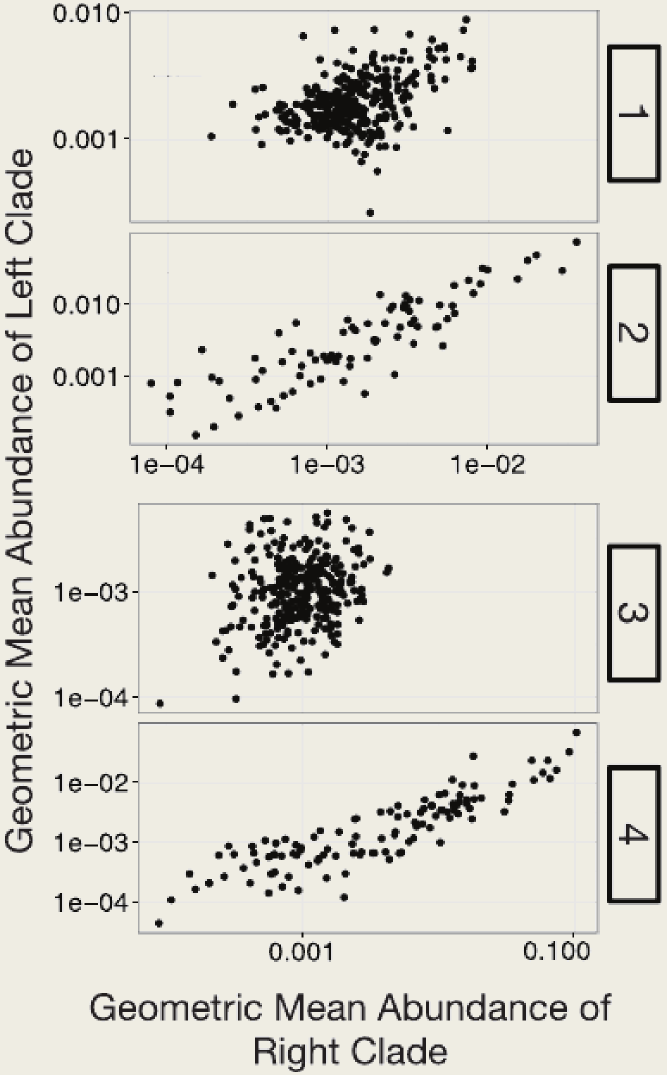
p-value < 0.01 log-scale LOESS regression



Results - Balance Variance and Depth

HMP Dataset Analysis Cont.



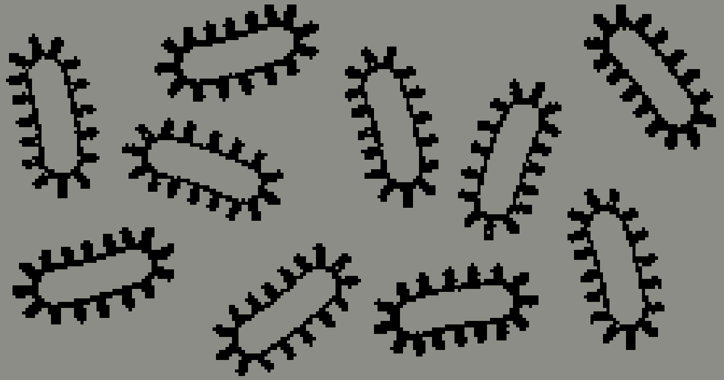


Results - Balance Variance and Depth

- Observed pattern of decreasing balance variance near the tips of the phylogenetic tree suggests that closely related bacteria tend to covary in human body sites.

- Trends between variance and phylogenetic depth were stronger above the species level than below it.

OUTLINE



A PHYLOGENETIC TRANSFORM ENHANCES ANALYSIS OF COMPOSITIONAL MICROBIOTA DATA

✓ Introduction

✓ Challenges

✓ Method

✓ Results

➢ Benchmarks

- Implementation
- Summary & Conclusions

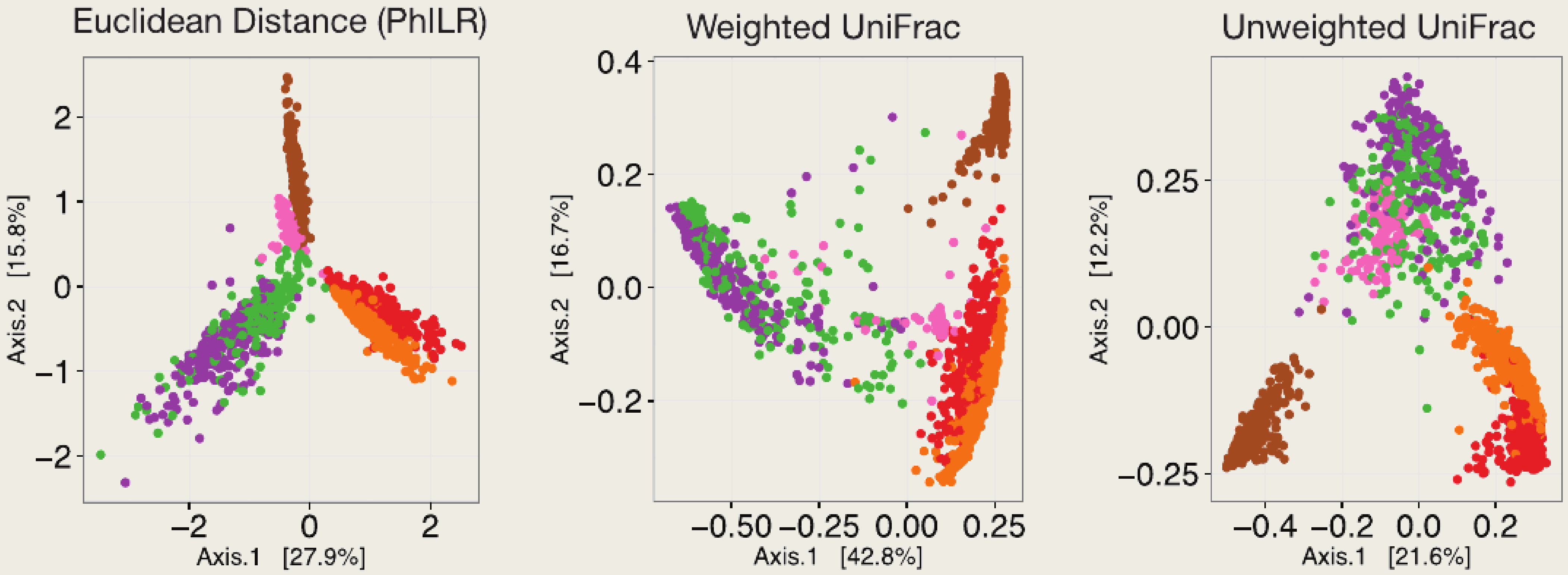
Benchmarks - PCoA

- No, it shouldn't be PCA.
- Principal Coordinates Analysis (classical multidimensional scaling) is a method to explore and to visualize data in a low dimensional Euclidean space, given a distance matrix.
- Intuitive example: cities map.
- If the distance metric is Euclidean, PCoA and PCA are identical.

Benchmarks - PCoA

Visualization of Body Sites Separation

🟩 Antecubital Fossa
🟪 Anterior Nares
🟧 Keratiznized Gingiva
🟥 Saliva
🟫 Stool
🟪 Mid Vagina

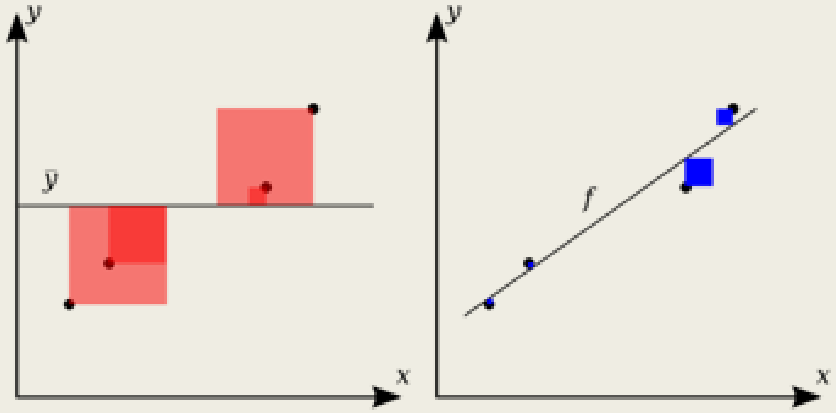


Benchmarks - PERMANOVA R^2

- R^2 is a statistical measure for the proportion of variance in the dependent variable that can be explained by the independent variable.

- Also called coefficient of determination, usually calculated by:

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

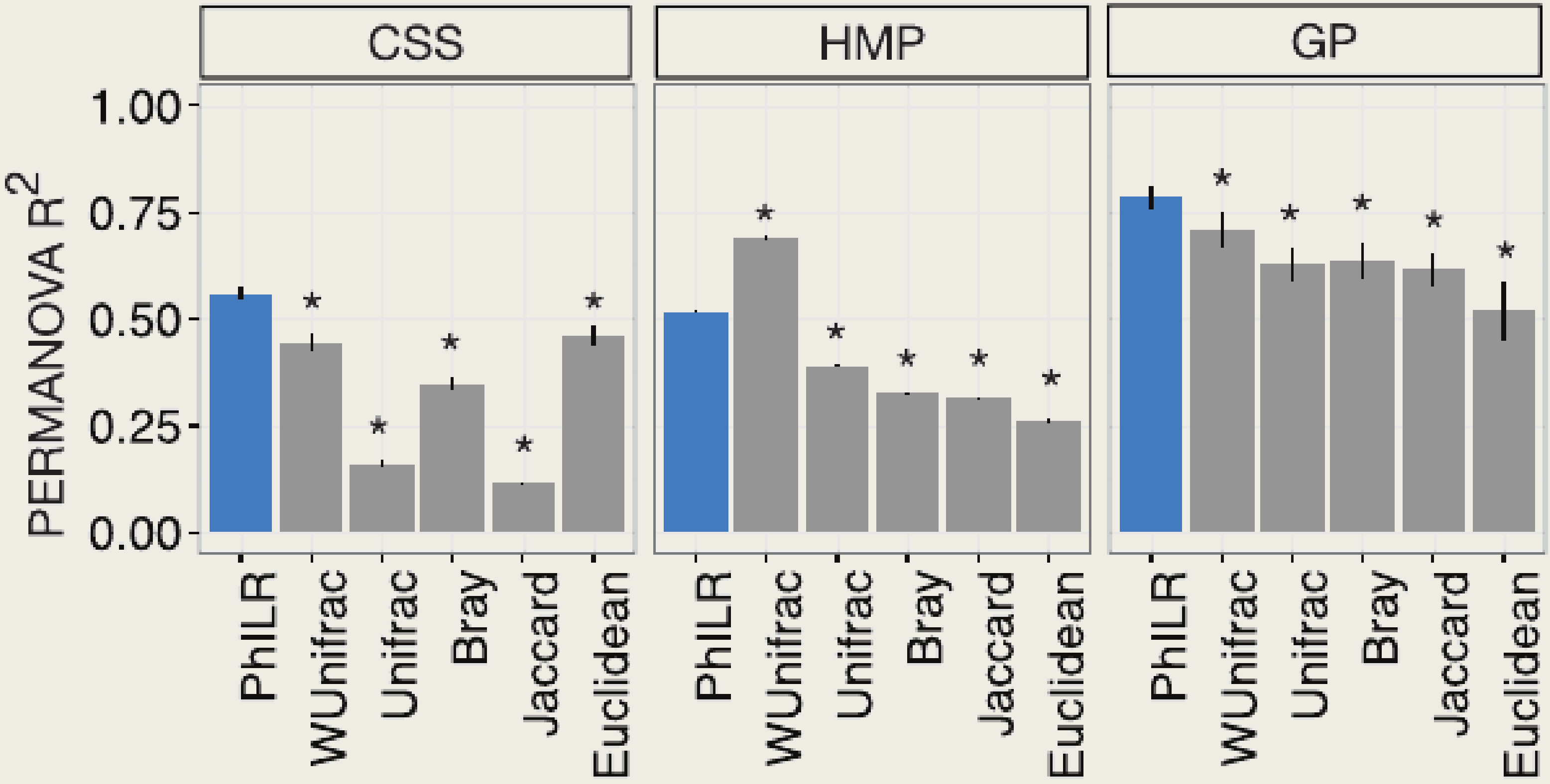


- It is a proportion - always a number between 0 and 1.

- In our case: PERMANOVA R^2 values should represent how well sample identity explained the variability in sample pairwise distances.

Benchmarks - PERMANOVA R^2

* indicates p-value < 0.01 of pairwise test against PhILR

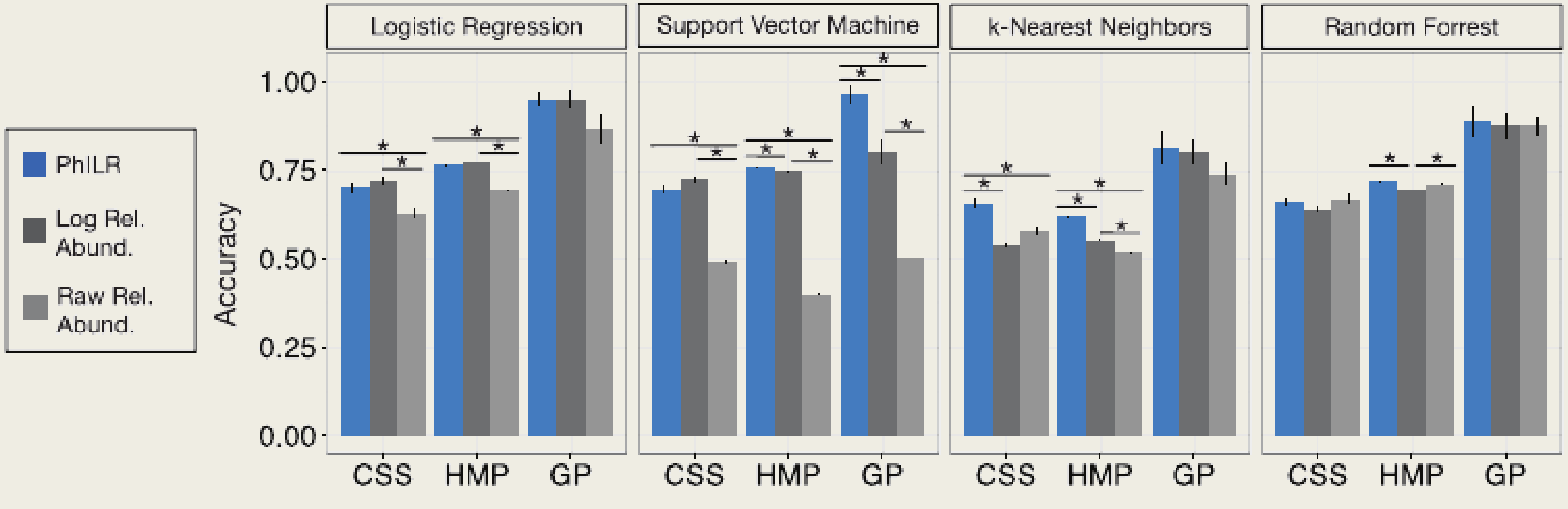


Benchmarks - Supervised Classification

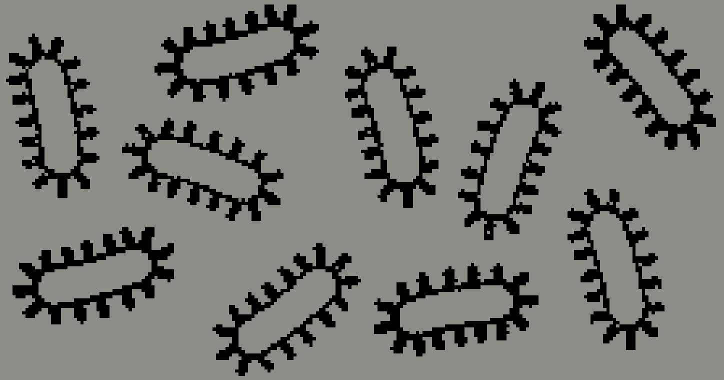
- Performance of predictive statistical models in the PhILR coordinate system.

- Supervised classification techniques were applied to the same datasets with different distance metrics.

* indicates p-value < 0.01 of pairwise tests



OUTLINE



A PHYLOGENETIC TRANSFORM ENHANCES ANALYSIS OF COMPOSITIONAL MICROBIOTA DATA

- ✓ Introduction
- ✓ Challenges
- ✓ Method
- ✓ Results
- ✓ Benchmarks
- ➢ Implementation
- Summary & Conclusions

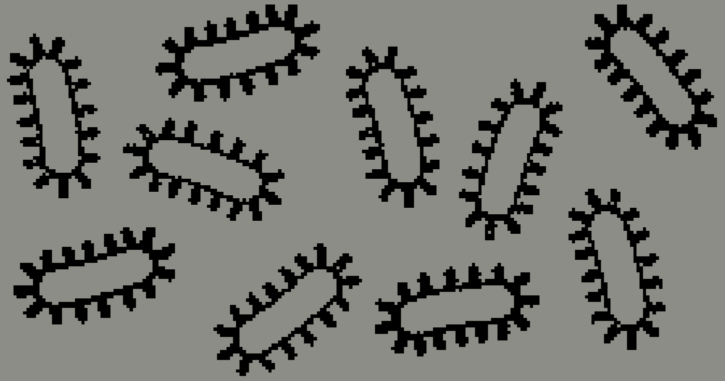
Implementation

- Implemented for the R programming language as a package named *philr,* available at: https://bioconductor.org/packages/release/bioc/html/philr.html

- The package contains the PhILR transform, with both weighting schemes (taxa & branch length) integrated.

- Open source 😊 https://github.com/jsilve24/philr

OUTLINE



A PHYLOGENETIC TRANSFORM ENHANCES ANALYSIS OF COMPOSITIONAL MICROBIOTA DATA

- ✓ Introduction
- ✓ Challenges
- ✓ Method
- ✓ Results
- ✓ Benchmarks
- ✓ Implementation
- ➢ Summary & Conclusions

Summary

- The PhILR transform incorporates microbial evolutionary model with the isometric log-ratio transform.

- Possible biological insights such as adaptation of bacterial clades to sites and relations between the covariation of neighboring bacterial clades to the phylogenetic depth.

- Benchmarks of community-level analyses depicted Euclidean distances on PhILR transformed data as a compositionally robust measure.

Conclusions

- Statistical methods can now be applied to metagenomic datasets as-is.
- Substitution of the transform into existing bioinformatics piplines should be seamless.
- Other arenas of bioligical research may also benefit from the PhILR transform.

Conclusions

Also Based on Editorial Decision Letter and Response

- The proposed weighthing schemes may be viewed as preliminary heuristics, additional effort is needed to determine when weightings are optimal.

- In some cases other sequential binary partitions may lead to similar benchmark performance but the phylogenetic nature allows for bioligical insights from coordinates interpretation.

- Specially designed tools, even if not grounded in CoDA theory, are still expected to perform equally or better at the respective benchmark.