Distance-Based Biogeographical Analyses Differ in Their Ability to Provide Insight Into Microbial Ecology

Jai Ram Rideout[1], Antonio Gonzalez[2], Andrew Cochran[2], Damien Coy[1], Michael Dwan[1], Andrew King[3], Logan Knecht[1], Dan Knights[4,5], Justin Kuczynski[6], Levi McCracken[1], Jessica Metcalf[2], Laura Parfrey[2], Bharath Prithiviraj[2], Michael S. Robeson[7], Will Van Treuren[2], Jose Carlos Clemente[8], Rob Knight[2,9], J. Gregory Caporaso[1,10]

[1] Northern Arizona University, [2] University of Colorado, [3] Ecosystem Sciences, CSIRO, [4] University of Minnesota, [5] Harvard Medical School, [6] Second Genome, [7] Oak Ridge National Laboratory, [8] Mount Sinai School of Medicine, [9] Howard Hughes Medical Institute, [10] Argonne National Laboratory

Introduction

Microbial ecology studies are increasing in scope and complexity. As more samples are being taken across temporal, spatial, and environmental gradients, there is an increasing need for tested, reliable biogeographical statistical methods in the field of microbial ecology.

Many biogeographical methods have been used in traditional macro-scale ecology for years to address these types of study designs. However, these methods have not been properly evaluated on microscale ecological data, where sample sizes are **bigger**, there are often more **observations** (e.g., sequences), and **scales** are different, to verify that biologically meaningful results are obtained.

We present an evaluation of methods for detecting biogeographical patterns on microbial ecology-based positive and negative controls, using both empirical and simulated datasets to verify their efficacy, and provide recommendations on methods that are most applicable to studies of microbial ecology.

Methods under evaluation

Classes of biogeographical methods

Gradient analysis	Grouping analysis
Mantel	Adonis
Mantel Correlogram	ANOSIM
Moran's I	db-RDA
BEST	MRPP
PC-Correlation	PERMANOVA

Evaluation datasets (i.e., positive/negative controls)

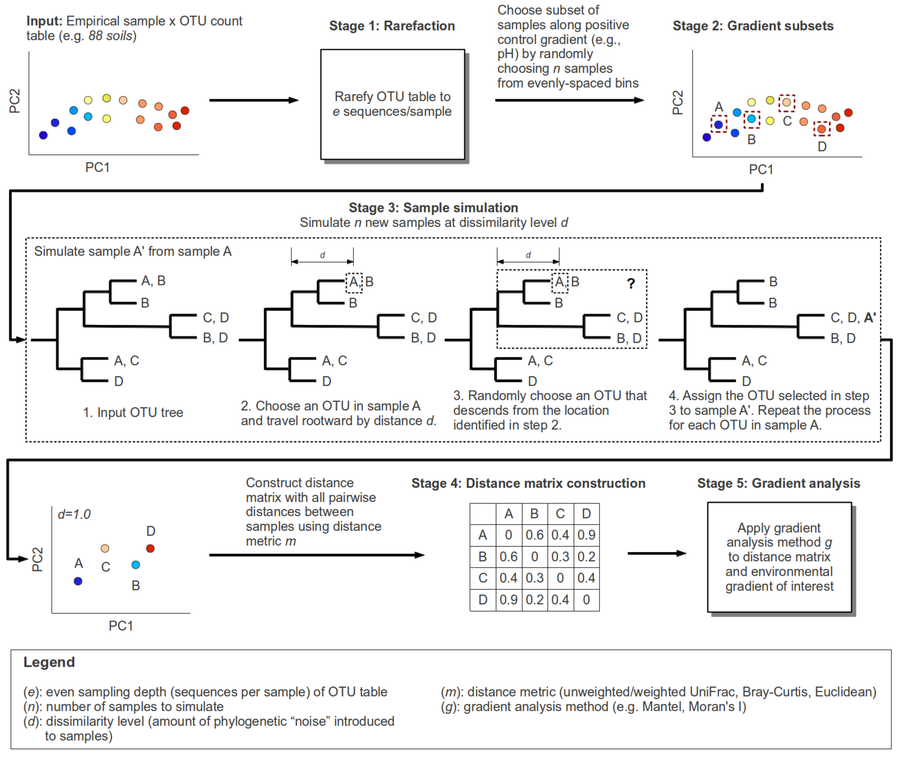
Empirical dataset	Parameter(s)	Citation
88 Soils	pH, Latitude	Lauber et al., 2009
Keyboards	Subject	Fierer et al., 2010
Whole Body	Body site, Sex	Costello et al., 2009
Guerrero Negro	Microbial mat layer	Harris et al., 2013

Availability

All methods are available in **QIIME** (versions 1.5.0 and higher) under an open source license (GPL), and are accompanied with **documentation**, **tutorials**, and **extensive unit tests**. More info can be found at www.qiime.org.

Evaluation strategy

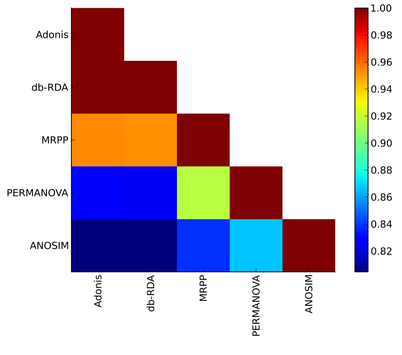
Simulated data workflow



	A	B	C	D
A	0	0.6	0.4	0.9
B	0.6	0	0.3	0.2
C	0.4	0.3	0	0.4
D	0.9	0.2	0.4	0

Methods are correlated, but differ in interpretability.

Each method's effect size / test-statistic varies in magnitude, making it difficult to differentiate positive and negative controls for some methods such as MRPP and Moran's I (data not shown).

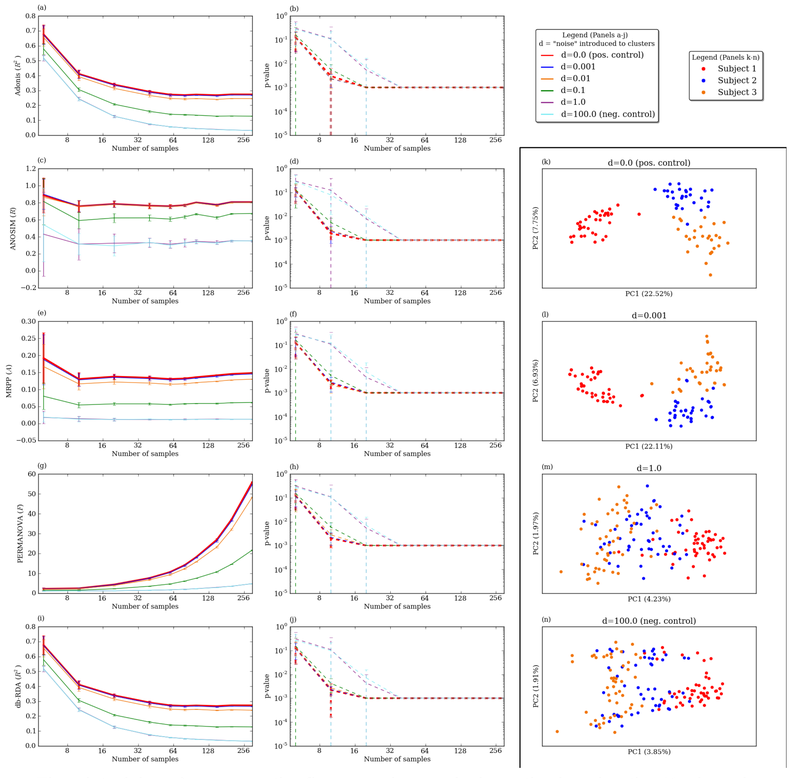


The heatmap to the left indicates Spearman's rank correlation coefficients between the five grouping analysis methods. The data used to compute the correlation coefficients were obtained from evaluating the empirical datasets using both positive and negative controls. Additionally, sampling depth, distance metric, and number of samples were varied.

The heatmap for gradient analysis methods is omitted here because analyses are still in progress.

Effect size matters: don't rely solely on the p-value!

As the number of samples increases, the p-value converges to zero, regardless of whether a positive or negative control is used. Thus, it is imperative to interpret **both** p-value and effect size / test-statistic.



The plots (above) compare the five grouping analysis methods using the *Keyboards* dataset at an even sampling depth of 390 sequences per sample with the unweighted UniFrac distance metric. The plots in the left column compare number of samples (x-axis, log scale) to effect size (y-axis) at varying dissimilarity levels (*d*). The plots in the central column compare number of samples (x-axis, log scale) to p-value (y-axis, log scale). Means are plotted with standard deviation based on 10 independent simulations. A dissimilarity level of 0.0 is represented by a bold red line and indicates the results of the original dataset without any simulation of the data. The rightmost column (panels *k*-*n*) are principal coordinates analysis (PCoA) plots that illustrate the effect of four dissimilarity levels on the clustering of samples.

Conclusions

Due to the ever-increasing number of samples used in modern microbial ecology study designs, it is necessary for researchers to consider both effect size magnitude *and* p-value when drawing conclusions. We plan to provide recommendations for specific methods that easily distinguish positive from negative controls based on effect size magnitude.