Title

Making sense of microbial populations from representative samples

Author

Morton, James

Publication Date
2018

UNIVERSITY OF CALIFORNIA SAN DIEGO

Making sense of microbial populations from representative samples

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

James T. Morton

Committee in charge:

Professor Rob Knight, Chair
Professor Pieter Dorrestein
Professor Rachel Dutton
Professor Yoav Freund
Professor Siavash Mirarab

2018

The dissertation of James T. Morton is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2018

DEDICATION

To my friends and family who paved the road and lit the journey.

EPIGRAPH

The ‘paradox’ is only a conflict between reality and your feeling of what reality ‘ought to be’

—Richard Feynman

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ANCOM	Analysis of Composition of Microbiomes
BIOM	Biological Observation Matrix
CF	Cystic Fibrosis
clr	Center Log-ratio
EMBAD	Earth Mover Band Aware Distance)
FN	False Negative
FP	False Positive
GLS	Generalized Least Squares
HGT	Horizontal Gene Transfer
ILR	Isometric Log-ratio
OTU	Operational Taxonomic Unit
PCA	Correspondence Analysis
PCA	Principal Component Analysis
PCM	Phylogenetic comparative methods
PCoA	Principal Coordinates Analysis
PERMANOVA	Permutational Multivariate Analysis of Variance
PGLS	Phylogenetic Generalized Least Squares
rRNA	Ribosomal RNA
SCORAD	Scoring Atopic Dermatitis

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to acknowledge Rob Knight, for his guidance that enabled me to grow as a person as well as a scientist. This work was largely enabled by him and I am greatly thankful for his support.

I would like to thank my committee members Pieter Dorrestein, Yoav Freund, Siavash Mirarab and Rachel Dutton not only for providing feedback for this thesis, but also for fostering years of collaborative science.

This work couldn't have been done with out the strong support of the members of the Knight lab. Many thanks to the individuals responsible for the finances, logistics, and computational hardware maintainence that ultimately facilitated the sharing of ideas across public venues, namely Ulla Westerman, Jerry Kennedy, Gail Ackermann, Jeff DeReus, Michiko Souza, and Sarah Adams. I would also like to acknowledge the multiple individuals that served as mentors, namely Jon Sanders, Daniel McDonald, Antonio Gonzalez, Yoshiki Vazquez-baeza, Zhenjiang Xu, Amnon Amir, Albert Barberán, Luke Thompson, Justine Debelius, Sejin Song, Jessica Metcalf, Greg Humphrey and Rob Quinn. My fellow peers brought the excitment into the lab, making my PhD experience thoroughly enjoyable, in particular Lisa Marotz, Qiyun Zhu, Stefan Janssen, Georg Loss, Anupriya Tripathi, Alison Vrbanac, Serene Lingjing Jiang, Antony Pearson, Peter Edge, Benjamin Pullman, Mingxun Wang, Richardo Silva, Alexander Aksenov, Louis-Felix Nothias-Scaglia and Brooke Anderson. Among the many collaborators, colleagues and mentors outside of UCSD responsible for contributing to this body of work, I would especially like to acknowledge Anna Edlund, Alex Washburne, Justin Silverman, John Karro, Iddo Friedberg, Manuel Lladser, Christopher Lowry, Noah Fierer, Benjamin Langmead, Jeff Leek and Susan Holmes for their thought-provoking insights.

Finally, I would like to thank my friends and family for their love and support. In particular, my parents Jade and John Morton for their drive and care early on distilling my love in engineering, mathematics and science. Rachael Morton for being a loveable sister and an angel

on my shoulder. And Juer Song, who is my pillar, my pillow and my greatest companion.

Chapter 1, in full, is a reprint of the material as it appears in "Methods for phylogenetic analysis of microbiome data" Alex D. Washburne, James T. Morton, Jon Sanders, Daniel McDonald, Qiyun Zhu, Angela M. Oliverio, Rob Knight *Nature Microbiology* 3, 2018. The dissertation author was the primary investigator and co-first author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in "Uncovering the Horseshoe Effect in Microbial Analyses" James T. Morton, Liam Toran, Anna Edlund, Jessica L. Metcalf, Christian Lauber, Rob Knight *mSystems*, 2, 2017. The dissertation author was the primary investigator and first author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in "Balance Trees Reveal Microbial Niche Differentiation" James T. Morton, Jon Sanders, Robert A. Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A. Navas-Molina, Se Jin Song, Jessica L. Metcalf, Embriette R. Hyde, Manuel Lladser, Pieter C. Dorrestein, Rob Knight *mSystems*, 2, 2017. The dissertation author was the primary investigator and first author of this paper.

Chapter 4 has been submitted for publication of the material as it may appear in Nature Biotechnology, 2019 "Establishing microbial measurement standards with reference frames" James T. Morton, Clarisse Marotz, Justin Silverman, Alex Washburne, Livia S. Zaramela, Anna Edlund, Karsten Zengler, Rob Knight. The dissertation author was the primary investigator and first author of this paper.

VITA

2014	B. S. in Computer Science *cum laude*, Miami University, OH
2014	B. S. in Mathematics and Statistics *cum laude*, Miami University, OH
2014	B. S. in Electrical Engineering *cum laude*, Miami University, OH
2014	B. S. in Engineering Physics *cum laude*, Miami University, OH
2018	Ph. D. in Computer Science, University of California San Diego

PUBLICATIONS

Author names marked with † indicate shared first co-authorship.

†Alex D. Washburne, †**James T. Morton**, Jon Sanders, Daniel McDonald, Qiyun Zhu, Angela M. Oliverio, Rob Knight "Methods for phylogenetic analysis of microbiome data" *Nature Microbiology* 3, 2018

James T. Morton, Liam Toran, Anna Edlund, Jessica L. Metcalf, Christian Lauber, Rob Knight "Uncovering the Horseshoe Effect in Microbial Analyses" *mSystems*, 2, 2017

James T. Morton, Jon Sanders, Robert A. Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A. Navas-Molina, Se Jin Song, Jessica L. Metcalf, Embriette R. Hyde, Manuel Lladser, Pieter C. Dorrestein, Rob Knight "Balance Trees Reveal Microbial Niche Differentiation" *mSystems*, 2, 2017

The following publications were not included as part of this dissertation, but were also significant byproducts of my doctoral training.

Stefan O Reber, Philip H Siebler, Nina C Donner, **James T Morton**, David G Smith, Jared M Kopelman, Kenneth R Lowe, Kristen J Wheeler, James H Fox, James E Hassell, "Immunization with a heat-killed preparation of the environmental bacterium Mycobacterium vaccae promotes stress resilience in mice" *Proceedings of the National Academy of Sciences*, 113, 2016

Jack A Gilbert, Robert A Quinn, Justine Debelius, Zhenjiang Z Xu, James Morton, Neha Garg, Janet K Jansson, Pieter C Dorrestein, Rob Knight, "Microbiome-wide association studies link dynamic microbial consortia to disease" *Nature*, 535, 2016

Yoshiki Vázquez-Baeza, Antonio Gonzalez, Larry Smarr, Daniel McDonald, **James T Morton**, Jose A Navas-Molina, Rob Knight, "Bringing the dynamic microbiome to life with animations" *Cell host & microbe*, 21, 2017

Albert Barberán, Robert R Dunn, Brian J Reich, Krishna Pacifici, Eric B Laber, Holly L Menninger, **James T Morton**, Jessica B Henley, Jonathan W Leff, Shelly L Miller, "The ecology of microscopic life in household dust" *Proc. R. Soc. B*, 282, 2015

Erin M Hill-Burns, Justine W Debelius, **James T Morton**, William T Wissemann, Matthew R Lewis, Zachary D Wallen, Shyamal D Peddada, Stewart A Factor, Eric Molho, Cyrus P Zabetian, “Parkinson’s disease and Parkinson’s disease medications have distinct signatures of the gut microbiome” *Movement disorders*, 32, 2017

Amnon Amir, Daniel McDonald, Jose A Navas-Molina, Justine Debelius, **James T Morton**, Embriette Hyde, Adam Robbins-Pianka, Rob Knight, “Correcting for microbial blooms in fecal samples during room-temperature shipping” *MSystems*, 2, 2017

Amnon Amir, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, **James T Morton**, Zhenjiang Zech Xu, Eric P Kightley, Luke R Thompson, Embriette R Hyde, Antonio Gonzalez, “Deblur rapidly resolves single-nucleotide community sequence patterns” *MSystems*, 2, 2017

Alison Vrbanac, Justine W Debelius, Lingjing Jiang, **James T Morton**, Pieter Dorrestein, Rob Knight, “An Elegan (t) Screen for Drug-Microbe Interactions” *Cell host & microbe*, 21, 2017

Sian MJ Hemmings, Stefanie Malan-Müller, Leigh L van den Heuvel, Brittany A Demmitt, Maggie A Stanislawski, David G Smith, Adam D Bohr, Christopher E Stamper, Embriette R Hyde, **James T Morton**, “The microbiome in posttraumatic stress disorder and trauma-exposed controls: an exploratory study” *Psychosomatic medicine*, 79, 2017

Laura-Isobel McCall, **James T Morton**, Jean A Bernatchez, Jair Lage de Siqueira-Neto, Rob Knight, Pieter C Dorrestein, James H McKerrow, “Mass spectrometry-based chemical cartography of a cardiac parasitic infection” *Analytical chemistry*, 89, 2017

Yoshiki Vázquez-Baeza, Chris Callewaert, Justine Debelius, Embriette Hyde, Clarisse Marotz, **James T Morton**, Austin Swafford, Alison Vrbanac, Pieter C Dorrestein, Rob Knight, “Impacts of the human gut microbiome on therapeutics” *Annual review of pharmacology and toxicology*, 58, 2018

Jessica L Metcalf, Se Jin Song, **James T Morton**, Sophie Weiss, Andaine Seguin-Orlando, Frédéric Joly, Claudia Feh, Pierre Taberlet, Eric Coissac, Amnon Amir, “Evaluating the impact of domestication and captivity on the horse gut microbiome” *Scientific reports*, 7, 2017

Lingjing Jiang, Amnon Amir, **James T Morton**, Ruth Heller, Ery Arias-Castro, Rob Knight, “Discrete false-discovery rate improves identification of differentially abundant microbes” *MSystems*, 2, 2017

Clifford A Kapono, **James T Morton**, Amina Bouslimani, Alexey V Melnik, Kayla Orlinsky, Tal Luzzatto Knaan, Neha Garg, Yoshiki Vázquez-Baeza, Ivan Protsyuk, Stefan Janssen, “Creating a 3D microbial and chemical snapshot of a human habitat” *Scientific reports*, 8, 2018

Daniel McDonald, Embriette Hyde, Justine W Debelius, **James T Morton**, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, “American Gut: an Open Platform for Citizen Science Microbiome Research” *mSystems*, 3, 2018

Robert A Quinn, William Comstock, Tianyu Zhang, **James T Morton**, Ricardo da Silva, Alda Tran, Alexander Aksenov, Louis-Felix Nothias, Daniel Wangpraseurt, Alexey V Melnik, "Niche partitioning of a pathogenic microbiome driven by chemical gradients" *Science advances*, 4, 2018

ABSTRACT OF THE DISSERTATION

Making sense of microbial populations from representative samples

by

James T. Morton

Doctor of Philosophy in Computer Science

University of California San Diego, 2018

Professor Rob Knight, Chair

Microbiomes make up the vast majority of life on Earth, and we are just beginning to understand how to study them using high-throughput omics. However, analysis of microbial populations is complicated by numerous statistical challenges. We first outline these challenges in the context of phylogenetically-aware methods, then focus on two concepts: the horseshoe effect and compositionality.

The horseshoe effect is a phenomenon that can lead to horseshoe patterns appearing in low dimensional representations of high dimensional data. For multiple decades, this pattern confounded ecologists when studying populations across multiple environmental conditions. Here, we show that the horseshoe effect arises from distance saturation, and can be indicative of microbial population displacement. This phenomenon is illustrated across a soil study and a decomposition study.

In the second part of the thesis, we will discuss identifiability due to representative sampling, also known as compositionality. Statistical laws have shown that it's possible to obtain unbiased estimators for population proportions from representative samples. However, based on representative samples alone, it is not possible to determine which species abundances have grown or declined, since there is an infinite number of outcomes that can explain the same change in proportions. In the biological sciences, this problem is also known as the differential abundance problem, which is critical for determining which microbes have been altered across experimental outcomes. Here, we show that in order to estimate which species have been altered, the total population size needs to be estimated.

We present two workarounds to this problem that ultimately negate the need to estimate total population size. The first solution is using ratios, analogous to concentrations in chemistry. We will showcase the usefulness of this technique on a soils study and a cystic fibrosis study. The second solution is using ranks as a proxy to feature importances. Rather than attempting to compute absolute change, we can compute relative change, ultimately ranking which microbes have increased or decreased the most across different experimental conditions. We show how these ranks can be computed using multinomial regression and can facilitate reproducible findings in the context of oral microbial communities and atopic dermatitis.

Chapter 1

Methods for phylogenetic analysis of microbiome data

How does knowing the evolutionary history of microbes affect our analysis of microbiological datasets? Depending on the research question, the common ancestry of microbes can be a source of confounding variation or can be a scaffolding used for inference. For example, when performing regression on traits the common ancestry is a source of dependence among observations, whereas when searching for clades with correlated abundances the common ancestry is the scaffolding for inference. The common ancestry of microbes and their genes is organized in trees – phylogenies – which can and should be incorporated into the analysis of microbial datasets.

While there has been a recent expansion of phylogenetically informed analytical tools, little guidance exists for which method best answers which biological questions. Here, we review methods for phylogeny-aware analyses of microbiome datasets, considerations for choosing the appropriate method, and challenges inherent in these methods. We introduce a conceptual organization of these tools, breaking them down into phylogenetic comparative methods, ancestral state reconstruction, analysis of phylogenetic variables, and analysis of phylogenetic distances. Careful consideration of the research question and ecological and evolutionary assumptions will help researchers choose a phylogeny and appropriate methods to produce novel, accurate, and biologically informative insights.

1.1 Introduction

High-throughput sequencing yields information about microbial communities in quantities that outstrip our ability to make sense of it. Most microbial taxa have never been cultivated or experimentally characterized. For many, we have only sequence fragments, whole genome sequence data for a few distant relatives, and a tree capturing the microbes' evolutionary histories. How can we organize and analyze the deluge of information about uncharacterized microbes and their sequence fragments?

Two main tools for organizing the diversity of life are the taxonomy and the phylogeny. The taxonomy classifies a microbe based on a hierarchy of taxonomic names ranging from one of three domains (Bacteria, Archaea and Eukarya) to one of several million species. The phylogeny is an estimation of the microbes' evolutionary history which classifies every organism by a series of splits corresponding to estimated events in which a most recent common ancestor speciated to form two daughter species.

Microbial taxonomy and phylogeny may eventually be equivalent, with every clade in the phylogeny having a taxonomic name. However, contemporary taxonomic classification is coarse relative to the phylogeny; modern taxonomic labels categorize a small fraction of the branches in the phylogeny. For the time being, the phylogeny is a more detailed source of knowledge about the common ancestry of microbes.

Phylogenies are a tool to organize and understand the microbial world[1, 2]. Because related organisms tend to have similar characteristics, phylogenies can incorporate those characteristics into our analyses even if we can't measure them directly. Phylogenies are a scaffold to classify lineages and infer functional ecological traits, even for lineages that have not been classified taxonomically or physiologically. Microbial ecology can be accelerated by high-throughput classification and inferences made possible with phylogenies. Resource consumption [3], habitat associations[4] and species interactions[5, 6] are causes and consequences of traits, and using phylogenies to infer or implicitly work with traits may enhance our ability to manipulate microbial communities to impact human health [7], biogeochemistry [8], and climate change [9].

How can a phylogeny assist analyses of microbiome data? Different research questions require different considerations about how to amend statistical analyses when considering a phylogeny. For example, studies testing associations between traits should consider the phylogeny as a source of dependence among observations, whereas studies looking for simpler ways of binning species should consider the phylogeny as a scaffolding for possible bins. There is a vast and growing literature on methods for analyzing phylogenetically-structured data, methods with subtle yet consequential differences in the questions they seek to answer. There is a need to simplify the diverse field into a set of conceptually distinct classes of methods and thereby provide a framework for instruction, comparison, and development of methods for analyzing phylogenetically-structured data.

In this review, we organize the field of phylogenetically-structured data analysis by discussing the major classes of methods. We first emphasize a fundamental issue in the field: the imperfection of estimated phylogenies. We then define four categories: (1) comparative methods, (2) ancestral state reconstruction and descendant trait imputation (3) variable analysis, and (4) phylogeny-aware distances (Table 1). Most statistical tools can be revisited for phylogeny-aware analyses, but the categories we cover capture the most commonly used and actively developed classes of methods. We discuss challenges of phylogenetically-aware analysis of microbiome data, including (Horizontal Gene Transfer (HGT)) and the choice of which genes to use when building phylogenies. By partitioning the literature into distinct conceptual classes of methods, we provide a common framework for the development and implementation of these important methods in microbiome data analysis. See Box 1 for a glossary of relevant terms in Appendix A.

1.2 Phylogenetic Inference

The tree of life is not known; it is estimated, and accurate phylogenies improve accuracy of phylogenetically-structured data analysis. Microbial phylogenies are commonly estimated by collecting gene sequences, aligning sequences based on homologies, and using models of mutation to infer most-likely evolutionary histories. The estimated phylogeny can vary depending on which genes are sequenced, how sequence positions are aligned, which model of evolution is used, and the method for inferring histories. Errors in phylogenetic inference can propagate to errors in phylogenetically-structured data analysis. Here, we discuss the interplay between phylogenetic inference and phylogenetically-aware analyses; for a review of methods for phylogenetic inference, readers can consult focused reviews of that topic [22, 23].

One can construct a phylogeny for any gene, and different genes will vary in the number of species containing the gene, the accuracy of the phylogeny, and phylogenetic signal of a set of traits. The 16S rRNA gene is commonly used for phylogenetic inference in Bacteria and Archaea, but one could also construct a phylogeny for other genes such as beta lactamases and their relatives, yielding a phylogeny with edges along which antibiotic resistance traits arose [22]. Microbial Eukaryotes likewise have many genes which can be used for phylogenetic inference, the 18S rRNA gene being most commonly used[24].

The genes chosen for phylogenetic inference ultimately determine the set of traits correlated with the phylogeny. Bacterial genome trees generally correlate with the 16S rRNA gene (16S)-derived phylogenies[25], but the correlation between a 16S tree and gene content varies over lineages and phylogenetic depths[26]. HGT disrupts the correlation between 16S trees and gene content by allowing bacteria with distant 16S genes to share common and consequential traits, such as pathogenicity islands and antibiotic resistance genes[27, 28]. Moreover, the 16S sequence has multiple variable regions, and can vary among multiple copies within the same genome, complicating phylogenetic inference[29]. More complicated scenarios, such as when epistasis underlies a functional ecological trait and one of the epistatic genes can be horizontally transmitted, prohibit a clear prescription for which gene's tree should be used for phylogenetic inference.

Different methods for analyzing phylogenetically-structured data use different features of the phylogeny. Distances and phylogenetic comparative methods which aggregate information over many branches in the phylogeny are more robust to errors in phylogenetic inference[30, 31] . Methods which rely on a few branches are more sensitive to errors in phylogenetic inference [32]. For methods relying on a few internal nodes or branches, the uncertainty in phylogenetic inference – particularly the bootstrap support for the monophyly of critical branches[33] – may be an important measure of uncertainty to incorporate into downstream data analysis.

Where monophyly is crucial, researchers can collapse resolved nodes into polytomies to improve the bootstrap support across the whole tree. A more certain yet coarse-grained phylogeny may be preferable to a less certain yet fully resolved phylogeny. Incorporating phylogenetic information has the capacity of drawing hypotheses on organisms never observed before. While the vast majority of microbial life on the planet is not cultureable,phylogenetic analyses allow us to extrapolate and infer characteristics about unknown organisms based on closely related, cultureable organisms.

1.3 Phylogenetic Comparative Methods

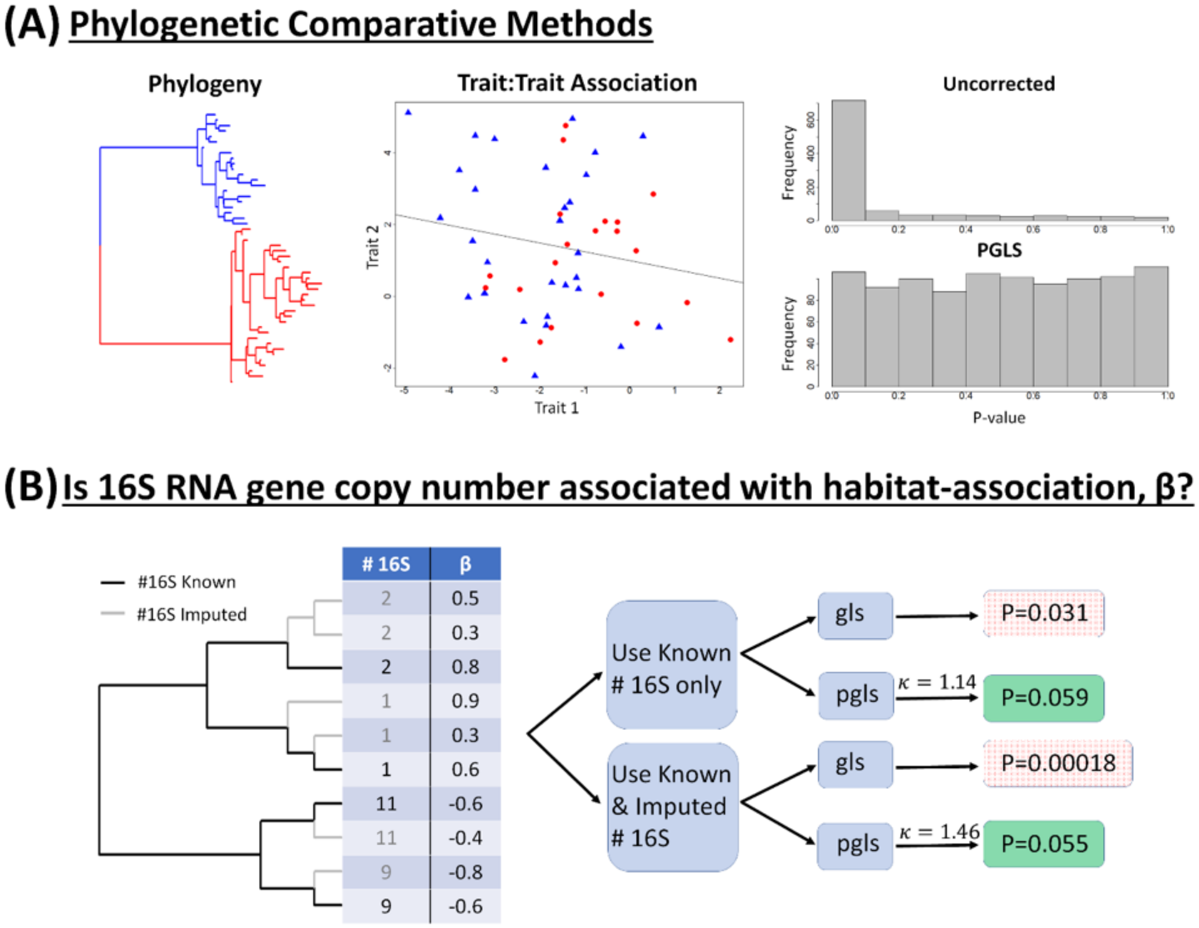
Phylogenetic comparative methods (PCM)s are used when comparing multiple traits across organisms. Closely related organisms often have similar traits due to inheritance from a common ancestor; such dependence of traits across organisms can affect tests of trait:trait and trait:habitat associations.

For example, we may find an association between 16S copy number (trait) and pH preference (habitat) through a correlation between 16S copy number and a measure of pH preference across 1,000 species of microbes (Figure 1.1a). Such an association could yield a false positive result if the taxa consist of a set of closely related of Acidobacteria with low 16S copy number and low pH preference, and a set of closely related Fusobacteria with high 16S copy number and a high pH preference[29]. Intuitively, the phylogenetic signal of these traits reduces our sample size because the observed traits represent samples from two lineages, not 1,000 independent species. More rigorously, the phylogeny affects the covariance structure of residuals under null models of trait evolution. Robust tests of trait-associations are done using PCMs[34, 11] (Figure 1.1b).

Generalized Least Squares (GLS) can control for dependence among observations when performing regression. In GLS, residuals—the difference between predictions and observations—are expected to covary and the covariance matrix is used to modify least-squares calculations. Random evolution produces close relatives whose observed traits will covary due to the shared variation acquired during their shared ancestry[34]. Phylogenetic generalized least squares[35] (PGLS; Figure 1.1), a tool for trait:trait and trait:habitat associations, implements GLS with residual covariances defined by a model of evolution.

A common first step in PCMs is to estimate and test the phylogenetic signal against a null model of no phylogenetic signal. The λ of Pagel [36] or the κ of Blomberg [37] are commonly used test statistics for phylogenetic signal. For PGLS, one must assume an evolutionary model; a Brownian motion, that is, a branching, random walk of trait values from an ancestral value at the root to the tips of the tree, is the default. The evolutionary model defines a covariance matrix for the residuals 1.1b. Under a Brownian motion model of evolution, the covariance between the residuals of two species' trait values is proportional to the amount of shared evolutionary history; more closely related species have more closely related traits even under a null model of random evolution. For more complicated models of evolution, one can jointly estimate the parameters for the evolutionary model and the regression coefficients [38].

PCMs extend to many statistical tests. Testing whether the volume of bacterial spores is smaller than the volume of daughter cells would involve a paired t-test, absent phylogenetic signal. A phylogenetic paired t-test[39] was developed to account for phylogenetic signal in such tests. There are many models of trait evolution, metrics of phylogenetic signal, and methods to control for phylogenetic signal when comparing traits. A recent scholarly edition of modern PCMs provides a review of the field and directions of current research[10].



# 16S	β
2	0.5
2	0.3
2	0.8
1	0.9
1	0.3
1	0.6
11	-0.6
11	-0.4
9	-0.8
9	-0.6

Figure 1.1: Advantages of using phylogenetic comparative methods. Phylogenetic comparative methods control for the statistical dependence among traits resulting from evolution of traits along the phylogenetic tree. (A) An exaggerated phylogeny with two distantly related clades. If trait evolution is simulated as a random walk on the phylogeny, the two distantly related clades will drive covariances between traits. Failing to correct for the effects of random trait evolution can lead to a high false-positive rate. Methods such as phylogenetic generalized least squares (PGLS) correct for the residual covariance expected under random trait evolution and produce more accurate statistical tests of association. (B) PGLS should be used when testing associations between traits, even trait quantities such as regression coefficients from abundance meta-data associations. To implement PGLS, a model of trait evolution needs to be assumed or estimated. Here, we estimate Blomberg's κ. PGLS should be used regardless whether the traits used are known or imputed through ancestral state reconstruction.

PCMs are not commonly used in microbiome studies, although a recent study[40] has employed PCMs to identify genes associated with colonization of the human gut (trait:habitat). Failure to correct for phylogenetic dependence in tests of trait:trait and trait:habitat association can yield a high false-positive rate (Figure 1.1). To amend this, we recommend researchers familiarize themselves with and utilize PCMs. Many methods can be implemented through the R packages [41] , phangorn [42], phytools[43], picante[44], caper [45], Geiger [46], and phylglm[47]. In the supplemental online tutorial (https://knightlab-analyses.github.io/phylogenetic-tutorials/), we illustrate how these packages can be used to simulate trait evolution and test associations between traits. We also illustrate the sensitivity of these methods to HGT.

1.4 Ancestral State Reconstruction

Estimating, or reconstructing, ancestral trait values assists imputation of traits in uncharacterized species and identification of historical lineages along which major trait differences arose. In studies of microorganisms, ancestral state reconstruction is commonly used to estimate genetic and metabolic profiles of extant communities using a set of reference genomes. In microbiome studies, this is commonly performed using PICRUSt[12], which uses pre-calculated ancestral state reconstructions to impute trait values, such as genes encoding glycoside hydrolase activity, for taxa whose traits are unknown.

PICRUSt operates on a phylogenetic tree, constructed from 16S sequences, connecting various sequenced genomes and environmental sequences. First, trait information observed in the sequenced genomes is used to infer ancestral trait profiles. Ancestral profiles are then used to predict the profiles of each organism in an environmental sample. An input sample's predicted metagenomic profile is then estimated by adding the product of OTU abundances in the sample and their corresponding profiles. Because this method relies strongly on the reference database, and the available sequenced genomes, it underperforms in environments where few or no genomes are known. Conversely, PICRUSt was able to predict the profiles of whole genome shotgun human fecal samples with a Spearman R^2 of > 0.9 [12], suggesting that the microbial phylogeny is highly predictive of microbial genome content.

The methodology underlying ancestral state reconstruction are very similar to phylogenetic comparative methods, as both require a consideration of models of evolution [48]. Three main types of algorithms are used to connect the tree, traits, a model of evolution, and estimates of ancestral states given the model of evolution: maximum parsimony, maximum likelihood and Bayesian inference[48]. Maximum parsimony reconstructs ancestral states by minimizing the number of trait changes between the ancestor and the present descendants. This approach assumes that trait changes are slow, and does not account for scenarios involving rapid evolution. In addition, maximum parsimony treats all branches the same and minimizes the number of changes on each branch; this can be problematic, particularly if not all of the species have been observed [49]. Maximum likelihood and Bayesian inference improve on maximum parsimony by incorporating explicit models of evolution – such as a Brownian motion model of trait evolution along the tree - into the estimation of ancestral states. Rather than simply assuming that changes are rare, these methods can account for some changes occurring more frequently than others—for example, assuming synonymous substitutions are more frequent than non-synonymous substitutions—and fit parameters to these models given an estimated phylogeny. However, maximum likelihood will often underestimate the number of changes within a single branch and can generate suboptimal results, particularly if the rate of evolution changes across the phylogeny [50]. Bayesian approaches can compute evolutionary parameters across a deep sampling of possible evolutionary trees and evaluate more complex models of evolution that account for non-uniform rates of evolution. While Bayesian methods can generate more accurate results than maximum parsimony or maximum likelihood, they can be computationally expensive with large numbers of species. Consequently, PICRUSt estimates microbial ancestral states using maximum parsimony or maximum likelihood.

As for phylogenetic comparative methods, estimates of ancestral states can be quickly confounded by HGT (see supplemental online tutorial), and thus applications of these methods to microbial datasets should be performed with consideration of the observed rates of transfer for the gene families of interest.

1.5 Analysis of phylogenetic variables

Locations on the Earth's surface can be described with three Cartesian (xyz) coordinates, but they are more naturally described using two spherical coordinates (latitude and longitude). A phylogeny, similar to a sphere, suggests natural coordinates. Phylogenetic variables are used to reduce the dimension of community ecological data, simplify calculations of distances, and describe meaningful features and directions of change in communities (Figure 1.2). We coin the term "phylogenetic variables" to describe variables constructed using features in the phylogeny to aggregate, contrast, and summarize data of species in the phylogenetic tree (Figure 1.2). Variables and distances are related, but contain distinct information: saying the city is east doesn't indicate how far it is, and saying a city is 80 kilometers away doesn't indicate which direction it is. Directions are described through phylogenetic variables (Figure 1.2A), and the magnitude of changes is measured through distances (Figure 1.2B). Phylogenetic variables include diversity metrics, taxonomic abundances, differences of abundance along all edges [44], differences of abundances between clades (Figures 1.2A, 1.2C, 1.2D) [13, 15], and more.

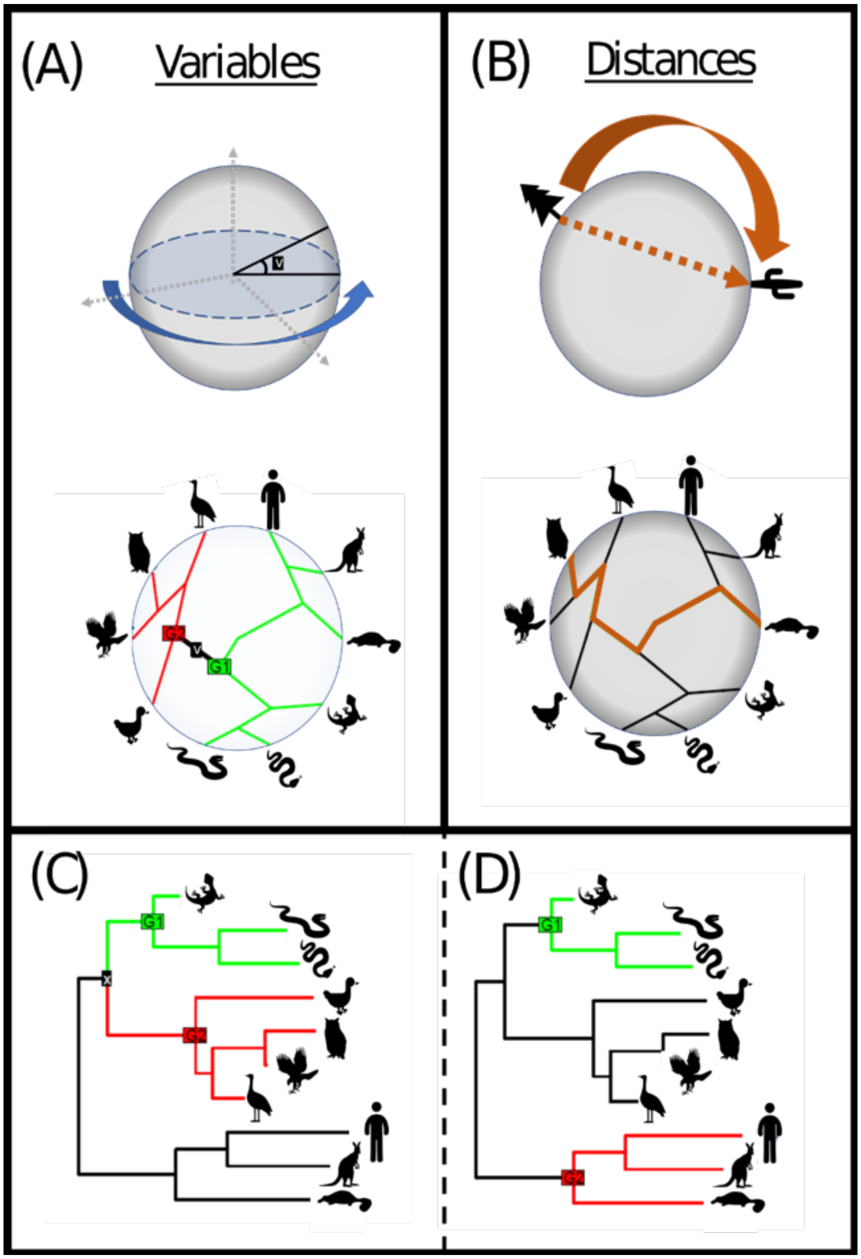


Figure 1.2: A comparison between phylogenetic variable analysis and phylogenetic distances. (a) Changing variables can allow more natural descriptions of complex topologies. A spherical Earth indicates spherical coordinates. Phylogenetic variables use the tree as a scaffolding for constructing coordinates corresponding to phylogenetic features. Phylofactorization constructs coordinates for contrasting groups, G1 and G2, separated by edges where traits, such as flight, arose. (b), A default path between two points is a straight line, but a more meaningful path on a sphere is a geodesic—that is, the shortest path along the surface of the sphere. Likewise, phylogeny-aware distances such as UniFrac define evolutionary paths and their distances between one community and another. (c), PhILR constructs coordinates between contrasting sister clades. (d), The space of possible phylogenetic variables and distances is infinitely large. Ratios between distant clades, as illustrated here, are viable but currently unused phylogenetic variables. Researchers should consider the biological interpretability of novel variables and distances, and their ability to inform future studies.

Phylogenetic variables simplify microbiome datasets by reducing the dimension of the data to a few variables carrying biological information. If a few monophyletic clades explain the majority of a microbiome dataset's variance along an environmental gradient, then there may be traits, shared among members of each clade, which are important determinants of abundance along the environmental gradient and underlie the observed community compositional changes.

The set of possible phylogenetic variables is infinitely large. Consequently, researchers must be deliberate in their choice of novel phylogenetic variables – what are important directions of change that carry implications for further research? Community changes along the direction of a phylogenetic variable, such as alpha diversity, does not necessarily convey useful biological information or immediate implications for future study design. Two common challenges in the analysis of phylogenetic variables can help guide the choice and development of phylogenetic variables: statistical dependence and biological interpretability.

Statistical independence, or well-characterized dependence, facilitates robust multivariate statistics and multiple comparisons corrections. For instance, when testing associations between species' abundances and environmental meta-data, and repeating the process for genera, families, orders, classes, and phyla, the variables analyzed have a nested dependence: if one taxon increases in abundance, all else being equal it will increase the abundance of all higher taxonomic groups in which it is found. For another example, if every sequence discovered is novel, the Shannon diversity of n sequences and n species will be $H = \log(n)$ and the species richness and evenness across samples will be correlated. Failing to account for the dependence among phylogenetic variables can increase error rates when performing multiple hypothesis tests.

Phylogenetic variables with a clear biological interpretation can carry implications for future study design and biological theory development. Changes in the abundance of a monophyletic clade may suggest a heritable trait driving changes in abundance; future experiments can focus on the clade to search for possible functional ecological traits. In macroscopic ecology, theoretical arguments justify the utility of various diversity metrics as proxies for extinction rates, island-biogeographic processes, ecosystem stability, and conservation goals[17, 19, 18]. Theoretical justification and interpretation of phylogenetic variables connects the analysis of phylogenetic variables (e.g. associations between diversity and meta-data) with experimental design and biological theory.

Two recently developed methods — PhILR [15] and phylofactorization [13] illustrate the challenges of phylogenetic variables analysis. Motivated by the compositional nature of sequence-count data [20, 21], both methods construct variables through average log-ratios of abundances between two clades in the phylogeny. PhILR variables measure the difference between sister clades (Figure 1.2C), and phylofactorization iteratively constructs variables measuring the difference between clades separated by edges in the tree (such as those in Figure 1.2A,D).

Changes in a PhILR coordinate may indicate a trait differentiating sister clades, whereas changes in coordinates from phylofactorization may indicate a trait arose along the identified edge. In both methods, significant associations between phylogenetic variables and meta-data motivate future work comparing genomes of two clades to search for functional traits. PhILR motivates comparison of sister clades (e.g. placental mammals to marsupials, or birds to crocodiles), whereas phylofactorization implicates comparison of clades separated by edges (e.g. birds to non-birds). In the supplementary tutorial, we illustrate these two methods for phylogenetic variables analysis, show how to construct these variables, compare them to EdgePCA, analyze a simulated dataset where rRNA gene copy number drives associations with disturbance frequency in soils[51], and interpret the results.

The goal of analyzing phylogenetic variables is to identify meaningful directions of change in microbiome data. Much like how principal components analysis can identify major directions/axes of variation in a dataset, phylogenetic variables can identify directions of change in microbiome data which explain variance in community composition and have implications for extinction risk, which organisms to cultivate, which genomes to search, and more.

1.6 Using Phylogeny-Aware Distances

Quantifying the dissimilarity between different species and between different communities comprising these species can facilitate accurate classification of meta-data (such as whether a patient has a disease), clustering of samples, and inferences of community function. Trees in forests sequester carbon in wood, whereas grasses do not. Consequently, measures of distance between communities containing trees from communities containing grasses may be indicative of differences in the ecosystem physiology of forests and grasslands. For the microbial world, traits driving ecosystem function are often unknown, yet accurate classification of disease states can have major consequences for human health and, where traits analogous to woody biomass underlie habitat associations, incorporating the phylogeny into distance measures can aid classification (Figure 1.3). Phylogeny-aware distances translate a dataset (figure 1.3a) into a distance matrix between samples (Figure 1.3b), which can be used to classify samples (Figure 1.3c).

One of the most widely used methods for phylogeny-aware analysis of microbiome data is the analysis of UniFrac distances between samples[16]. The UniFrac distance was motivated as a more biologically meaningful distance between communities than standard Euclidean and Bray-Curtis distances. The intuition behind UniFrac, and most phylogeny-aware distances, is that communities containing more phylogenetically distinct species are more different than communities with more closely related species. Incorporating phylogenetic distances along which functional changes occur may better quantify functional differences between communities.

Many extensions of Unifrac have been explored with the aim of controlling statistical artifacts in count data and tuning the importance of abundance in UniFrac distances. If counts are randomly distributed among species, clades with more species will have higher variances in total counts and thus have greater impact on UniFrac distances than clades with fewer species. To remedy this effect, VAW-UniFrac[52] stabilizes the variance of UniFrac distances. VAW-UniFrac was extended by the Generalized Unifrac Distance [53], which contains a tunable parameter to increase/decrease the importance of abundance in the distances between communities.

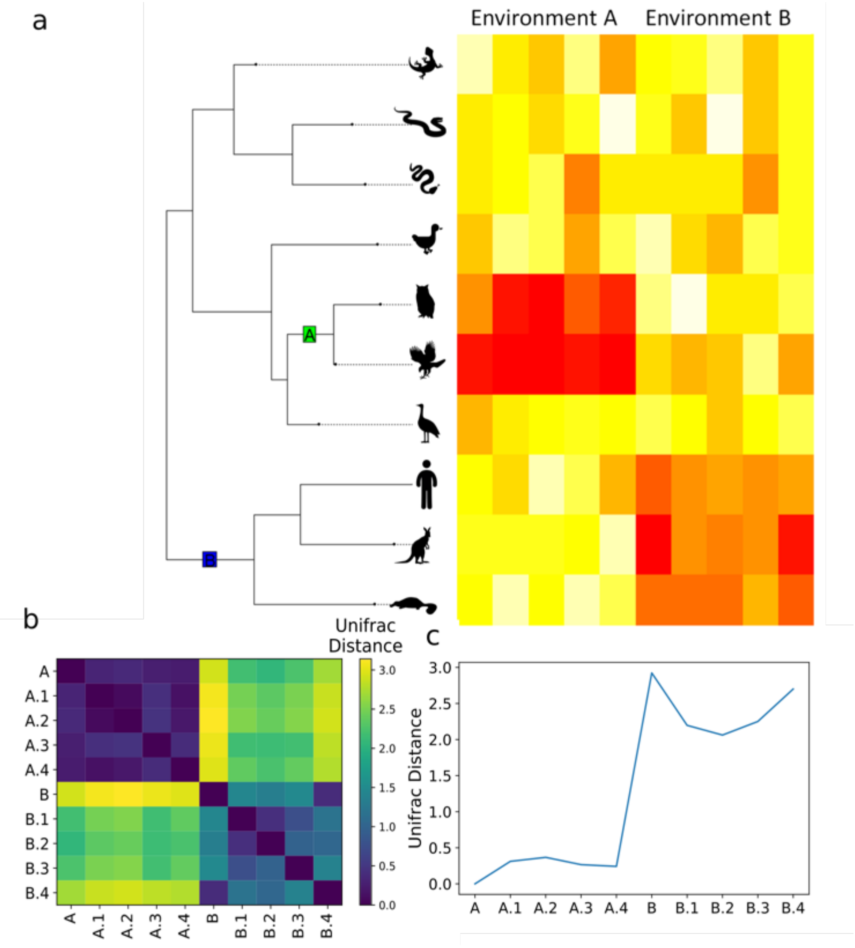


Figure 1.3: A demonstration of how to interpret Unifrac distances.(a) A heatmap of species abundances with red indicating high abundance and yellow indicating low abundance across different environments. The evolutionary history is represented by the phylogenetic tree, and the main differences between Environment A and Environment B are being driven by the abundances in clade A and clade B. (b) While variables contain information for each sample, distances relate two samples. Plotted are the pairwise Unifrac distances between the samples; distances between samples from Environments A and samples from Environment B are larger compared to distances between samples from Environment A or distances between samples from Environment B. (c) The Unifrac distance between a sample from the Environment A to all other samples illustrates how distances can be useful for sample-site classification. Phylogeny-aware distances can relate to functional distances by capturing flow of abundances through edges along which traits arose.

There have been a number of other phylogenetically informed distance metrics such as Sorensens' index, Rao's D and Rao's H that have been proposed alternative methods to incorporate evolutionary information [54]. Furthermore, standard statistical techniques such linear regression can be augmented to penalize differences between close relatives[55, 56, 57].The phylogeny is a scaffold for many variables, and can serve as the basis for many useful distance metrics. Which distance(s), of the possible distances, are of interest to a given microbiologist?

We suggest two main goals in the construction of phylogeny-aware distances: improving sample-site classification/visualization and providing meaningful interpretations of community differences.

If sample-site visualization is the goal of an experiment, a researcher may be inclined to search through a space of possible distances until finding one that looks the best, irrespective of the biological interpretability of the distance. Otherwise, searching too many distances risks dredging the data and presenting statistically significant patterns which were obtained by testing multiple candidates without proper corrections for multiple hypothesis tests performed. Correcting for such multiple tests will face the same challenges of unclear dependence among tests that arise in the analysis of multiple phylogenetic variables. While many existing distances can successfully classify samples across a range of site categories and clinical variables, the biological implications of discovered differences are often unclear. Does a larger distance indicate greater difficulty in bioremediation of one community into another? Does a larger distance imply a larger difference in ecosystem function or patient morbidity? What follow-up experiments should one conduct to better understand the biochemical and microbiological causes of community differences, given a large UniFrac distance?

Construction of new phylogeny-aware distances and their use in modified statistical methods should consider the performance gains relative to existing methods and whether they provide a new interpretation of discovered differences. Careful justification of new distances can improve the biological interpretation of results. For instance, macroscopic ecologists debate how beta diversity can be used for conservation[19]. Such discussions can improve the interpretation of existing and newly developed phylogeny-aware distances and help researchers understand any implications of high or low distances between communities. In addition, a high quality tree is critical for revealing ecologically relevant patterns [58]. As with phylogenetic comparative methods and phylogenetic variables, phylogeny-aware distances benefit from explicit consideration of ecological and evolutionary models to aid the biological interpretation of their results.

1.7 Challenges of phylogenetic analysis

There are challenges to phylogenetically structured data analysis, including HGT, the choice of which gene tree to use, the sensitivity to errors in phylogenetic inference, and the explicit consideration of ecological and evolutionary models. Here, we discuss broader challenges of phylogenetic analysis; for challenges especially relevant to microbial and microbiome datasets, (see Box 2 in Appendex A). HGT between microbial genomes complicates the evolutionary story of vertical transmission captured in a phylogenetic tree [59]. HGT raises the question of which phylogeny to use and how informative the phylogeny is for the research question. For PCM, HGT can lead to improper corrections and poorly calibrated statistical tests (illustrated in supplement). HGT of a major trait driving variation in the data can reduce the appropriateness of the phylogenetic variables or distances being used.

It is favorable to choose gene families that are insensitive to HGT for inferring phylogenies. Studies have evaluated the chance of HGT based on functional and ecological features[59, 60], providing guidelines for this task. Perhaps there is no gene absolutely HGT-free throughout the tree of life, including 16S [61]. Using multiple genes in phylogenetic inference can minimize the negative impact of HGT [62], and reveal genes influenced by HGT within the selected range of taxa[56]. Computational tools are available for assessing the probability of putative HGT events based on species/gene tree reconciliation[63]. Exploration of genomic context, sequence signature and atypical homology search results also help tracking HGTs[64].

HGT does not invalidate phylogeny-aware analyses of microbiome data. HGT of functional traits could be hypothesized through phylogeny-aware analyses by strong effects with little phylogenetic signal [65]. If phylofactorization identifies an unusually large number of tips of the tree associated with antibiotic exposure, HGT may be driving variation in the data and can be further tested by comparison of genomes among the phylogenetic factors identified. Nonetheless, HGT requires consideration when analyzing phylogenetically-structured data. The sensitivity of many methods to the horizontal transfer of functional traits is currently understudied. The combination of HGT and the existence of different phylogenies for each gene motivates careful justification of which genes to use to make phylogenies. Finally, all methods face the challenge of being interpretable and advancing our knowledge of microbiological systems. To that end, new methods should explicitly consider ecological and evolutionary models for how traits evolve and drive patterns in the data. One study simulated trait evolution on a tree and compared PGLS with phylogenetic eigenvector regression methods[66], which use eigenvectors from phylogenetic distance matrices as explanatory variables and do not correspond to a clear evolutionary model. The study found that PGLS produced more reliable and better-calibrated statistical results[66]. Considering evolutionary and population genetic models in method development promotes accurate understanding of the assumptions under which a phylogeny-aware analysis performs well and interpretation of findings in terms of the biological processes at play[67].

As more methods are developed, researchers should be aware of the tradeoffs between machine learning and human understanding: the former may produce more accurate predictions in the short term, whereas the latter produces theory that can generate more accurate and generalizable predictions in the long term.

1.8 Discussion

The common ancestry of microorganisms can be a source of confounding variation in our data, or a scaffolding on which we make inferences. There are many existing and emerging methods for analyzing microbiome datasets in light of evolution, and choosing the right method requires precise statements of the research question (Table 1).

First, decide which tree to use. Commonly, microbiome studies use the 16S tree for Bacteria and Archaea and the 18S tree for microbial eukaryotes, but there is a phylogeny for every gene and some questions are better analyzed with trees from other genes. The phylogeny obtained will be an estimate, and uncertainty in phylogenetic inference can translate to uncertainty in downstream phylogenetically-structured data analysis.

If the research question uses a trait as a response variable, the phylogeny may be a source of confounding variation. Phylogenetic comparative methods, such as PGLS, correct for dependence among traits one expects under null models of evolution along the tree.

If the research question is seeking historical trait values, or edges along which major trait differences arose, ancestral state construction is needed. If testing associations between imputed traits, researchers need to combine ancestral state reconstruction for imputation of missing traits with phylogenetic comparative methods which correct for confounding variation.

If the research question aims to simplify patterns of community composition, the phylogeny is a scaffolding that can be used to produce biologically informative variables and directions of change. The choice of variables should be made according to their ability to capture features in data, their statistical dependence, and their biological interpretation.

If the research question is to differentiate microbial samples, the phylogeny can define distances between samples. By re-defining distances, the phylogeny can be used to modify virtually any statistical method, but the choice of which distance to use should be based on the research goals of sample-site classification or biological interpretation of differences.

Phylogenetic analysis of microbiome data can allow researchers to categorize unclassified microorganisms, test evolutionary hypotheses about trait associations or traits driving habitat associations, and better understand how microbial communities differ and how they change over time, space, and treatments. There are several classes of methods for analyzing microbiome data in light of evolution. Careful consideration of the research question and the allowable ecological and evolutionary assumptions enables researchers to identify existing methods or produce novel methods that address their research question and produce novel, accurate, and biologically informative insights. The deluge of information about microbial sequences is producing phylogenetically-structured data which, given the right tools, can accelerate our understanding of microbial community structure and function.

1.9 Acknowledgements

J.T.M. was funded by NSF GRFP DGE-1144086. A.D.W. received support from Duke University Biology Department's provision of start-up funds for D. Nemergut (deceased) and the Defense Advanced Research Projects Agency (DARPA) grant D16AP0013. This paper is published in the spirit of D. Nemergut's contagious love of science.

Chapter 1, in full, is a reprint of the material as it appears in "Methods for phylogenetic analysis of microbiome data" Alex D. Washburne, James T. Morton, Jon Sanders, Daniel McDonald, Qiyun Zhu, Angela M. Oliverio, Rob Knight *Nature Microbiology* 3, 2018. The dissertation author was the primary investigator and co-first author of this paper.

Table 1.1: Comparison of different phylogentically aware methods. Different classes of methods for using the phylogeny in data analysis address different classes of questions. These methods can be summarized based on their use of given a dataset of abundance vectors, x, observed or imputed trait values, y, and the phylogeny, P.

Class of Methods	Brief Description	Specific Example	General Formula	Highlighted Method
Comparative Methods	Find associations between traits, controlling for evolution on phylogeny	Is 16S Ribosomal RNA (rRNA) gene copy number associated with growth rates in vivo?	$y_i - g(Y) + \varepsilon$ where Conv $[\varepsilon] = f(P)$	Phylogenetic Generalized Least Squares (PGLS)[10] Paired t-test[11]
Ancestral State Reconstruction	Impute trait values for historical lineages in the phylogeny and use ancestral traits to impute trait values for contemporary species	What is the best estimate of 16S rRNA gene copy number of an Operational Taxonomic Unit (OTU) based on the 16S rRNA copy numbers of its relatives?	Infer features of $P\|y$ Impute $y_{i,j}\|y_j$	PICRUSt[12]
Phylogenetic Variables	Use the phylogeny to construct variables that are biologically interpretable (for example, a clade's abundance) and simplify/summarize features in the community	Which interior edges in P separate taxa with different habitat associations? How does Faith's phylogenetic diversity change with pH?	Define variables: $v_i = f_i(x,P)$ Analyse, interpret and combine v_i	Diversity analyses Taxonomic analyses Phylofactorization [13] EdgePCA[14] PhIsometric Log-ratio (ILR)[15]
Phylogeny-Aware Distances	Use the phylogeny to construct distances between samples, which can then be used to modify statistical tools for classification, regularized regression and more	How different are two microbial communities?	Define distance $d[x_i,x_j] = K(x_i,x_j,P)$ Analyse and use to modify various statistical methods	UniFrac[16, 17, 18, 19] Inner product methods[20], [21]

Chapter 2

Uncovering the horseshoe effect in microbial analyses

The horseshoe effect is a phenomenon that has long intrigued ecologists. Commonly thought to be an artifact of dimensionality reduction, multiple techniques were developed to unravel this phenomenon and simplify interpretation. Here, we provide evidence that horseshoes arise as a consequence of distance metrics that saturate - a familiar concept in other fields but new to microbial ecology. This saturation property loses information about community dissimilarity, simply because it cannot discriminate between samples that do not share any common features. The phenomenon illuminates niche differentiation in microbial communities and indicates species turnover along environmental gradients. Here we propose a rationale to the observed horseshoe effect from multiple dimensionality reduction techniques applied to simulations, soil samples, and samples from postmortem mice. An intuitive-depth understanding of this phenomenon allows for the targeting of niche differentiation patterns from high-level ordination plots.

2.1 Introduction

Ecological datasets, particularly those observed in microbiome studies, are typically sparse and high-dimensional, frustrating most conventional statistical techniques. Many numerical ecology software packages make use of distance-based statistics by calculating the distance between ecological communities, to compare various ecosystems to each other over space and time. One of the most common exploratory analysis techniques is ordination, where the distances between the communities are embedded into a Euclidean space, and then visualized via Principal Components Analysis (Principal Component Analysis (PCA)) [68]. A widely used extension of this technique, where the distance metric can be varied, is called Principal Coordinates Analysis (Principal Coordinates Analysis (PCoA)) [68].

One phenomenon that commonly occurs in datasets containing ecological gradients is the horseshoe effect or Guttman effect [69]. This phenomenon is typified by a linear gradient that appears as a curve in ordination space. The horseshoe effect, or its relative the arch effect

[70] (where the ends of the gradient do not attract each other along the first principal coordinate as they do in the horseshoe effect), is observed using multiple types of ordinations, including Principal Components Analysis, Principal Coordinates Analysis, Non-Metric Multidimensional Scaling, Correspondence Analysis, and many others [68]. In 1982, the prevailing view of the horseshoe effect arose, when it was described by Gauch as a mathematical artifact that obscures the underlying ecological gradient. Soon thereafter, Detrending Correspondence Analysis [70] was invented to unbend the horseshoe using reciprocal averaging. Since then, detrending has become a commonly applied practice to ordinations in ecological datasets. Although these detrending techniques appear to provide a more intuitive visualization, they have been criticized as providing a distorted perspective of the underlying data, relying on many parameter settings that cannot be chosen in a principled way, and obscuring true underlying patterns in the data [71].

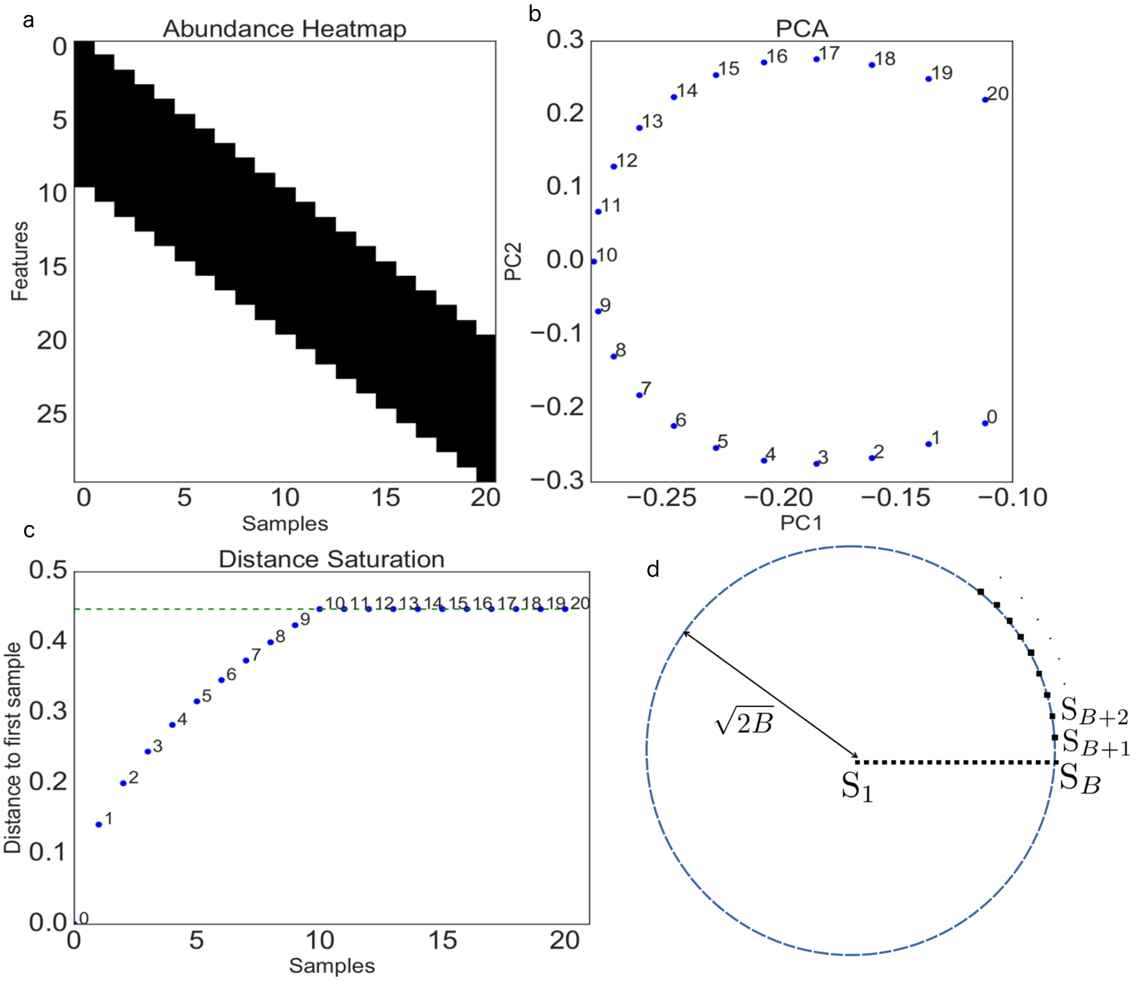


Figure 2.1: An explanation of the horseshoe effect arising from distance saturation. (a) A band table where the y axis encodes for individual Operational Taxonomic Unit (OTU)s and the x axis encodes for samples. Blocks that are colored black have a value of 1/10 while blocks that are colored white have a value of 0. (b) The first 2 components from a PCA of the band table, yielding the typical horseshoe shape. (c) The Euclidean distance from the point 0 to all of the other points. (d) An illustration of distance saturation property.

In previous studies, it was shown that horseshoes can arise from band tables [72, 73]. These tables consist of highly dense, non-zero values along the diagonal of the table, and sparse values everywhere else. This pattern can be apparent when the rows and columns are sorted in the proper order. Although the idea that band tables lead to horseshoes is not a new idea, it is commonly misunderstood how this concept applies to microbial analyses. Here we provide some intuition behind the mathematical structure of horseshoes.

In Figure 2.1a, we show a simulated band table, where each vertical band is represented by a sample, and contains 10 non-zero values. In typical microbiome datasets, these values could reflect OTU or species counts; for simplicity, here we will to refer to them as species counts, although this concept can also be generalized to multiple data types, such as gene counts, metabolite abundances. Each sample in the table is shifted by 1 row, creating the band effect. When PCA is applied directly to this table, the first 2 eigenvectors yield a horseshoe pattern (Figure 2.1b). Here, the band table is parameterized with a band size of 10, since each sample has exactly 10 non-zero values.

For close local points, the Euclidean distance grows linearly along the gradient (Figure 2.1c). However, after a certain point, the distance completely saturates. This property has been previously noted with Euclidean distance [70]. The overlap between the first sample in the band table, and sample 10 and beyond disappears, and the distance between these samples is maximized. This can yield unintuitive properties, sample 10 could be less dissimilar than sample 1 compared to sample 20. For instance, sample 10 could represent a medium pH environment, sample 1 could represent an acidic low pH environment and sample 20 could represent a high pHbasic environment. Sample 1 is expected to be a substantially more different microbial community to Sample 20 than Sample 10. The acidophiles found in Sample 1 are typically not found in basic environments. Sample 20 is expected to be more different to sample 1 than sample 10, since it contains very different microbes that thrive in high pH environments. But as far as Euclidean distance is concerned, sample 10 is just as dissimilar to sample 1 as sample 20, just because there are no common bacteria shared between these samples. It is apparent that the saturation property of Euclidean distance does not capture all of the information about community dissimilarity along a gradient, simply because it cannot discriminate between samples that do not share any common features. Once the distance is saturated, all samples that do not overlap lie within a ball of radius B where B is the band size lying within a ball of radius where B is the band size and the first point is the center of the ball as shown in Figure 2.1d.

This saturation property has been suggested to give rise to horseshoes in previous studies in other fields [72], and is an unintuitive property that can confound ecological interpretations if not understood properly. This property also restricts the possible trajectories of samples in the feature space, and gradients cannot be represented by linear trajectories in the real space (Supplemental proof 1 in Appendix B). This means that communities in the original high dimensional space do not arrange into linear trajectories in the first place, and when projected to lower dimensions do not fall into linear trajectories. These trajectories are what we refer to as horseshoes. The horseshoe phenomenon is analogous to the familiar concept of saturation in molecular evolution, where two randomly evolving sequences saturate at 75% DNA sequence identity (assuming equal nucleotide frequencies), even if infinite time has elapsed [74]. Consequently, distances that reflect a higher degree of molecular change need to be corrected for multiple substitutions in order to recover the molecular clock-like behavior obtained when comparing more similar sequences. This is why corrections according to models such as Jukes-Cantor or the Kimura 2-parameter model are required to obtain distances for reconstructing better phylogenetic trees. Analogous distance corrections are needed in microbial ecology for reconstructing better relationships among microbial communities [75].

It is important to note that horseshoes do not only arise from PCA, but also arise in PCoA with a variety of distance metrics. Arch effects have plagued every multidimensional reduction technique we have applied to a wide range of microbial ecology datasets [76]. In the following case studies, we'll show that these distance metrics also have the saturation property. In addition, if a distance doesn't have this saturation property, there won't be an observed horseshoe artifact (Figure S1).

Case Study 1 - 88 Soils

In this study, 88 soil samples were obtained from multiple locations across the United States having varying levels of pH [77]. The V4 region of the 16S rRNA gene (16S) within each organism was amplified and sequenced using 454 pyrosequencing to obtain relative abundances of microbial taxa. A matrix representing abundance values for each taxonomic unit per soil sample was used as input in correspondence analysis (Correspondence Analysis (PCA)) [78]. The resulting ordination showed clear separation of the communities based on pH (Figure 2.2a), which led to the same conclusion that pH is a major driving factor in soil biogeography, i.e. pH has major impacts on the distribution of bacterial taxonomic units in soil [77]. The PCA analysis in Figure 2.2a also shows the classic horseshoe shape. Here we revisited this study, to better understand the horseshoe shape behind this dataset.

To test the effect of another commonly used distance metric on the sample distribution, we analyzed the same soil dataset applying Chi Squared distance (Figure 2.2b). Similar to what was observed with Euclidean distance, which was applied in the simulation, the Chi Squared distance increased sharply at pH 3 and 4, but began to saturate at pH of 5. Also the band table similar to what we have observed in Figure 2.2a can be obtained when sorting. Also, when the the OTU table was sorted by sample pH and the mean pH of the samples that the OTUs were observed in mean pH of the OTUs (Equation 12), the same band table pattern appeared as we show in Figure 2.2a. While the diagonal isn't completely dense, there are more non-zero values compared to the corners of the heatmap. In line with the findings from the original study, this pattern is likely representative of niche differentiation of OTUs with respect to pH. The organisms that thrive in low pH environments tend not to exist in high pH environments and vice versa. Low pH and high pH samples are shown in Figure 2.2c to have few overlapping species, a pattern not observed in the original study as membership was evaluated at coarser levels of taxonomic resolution[77].

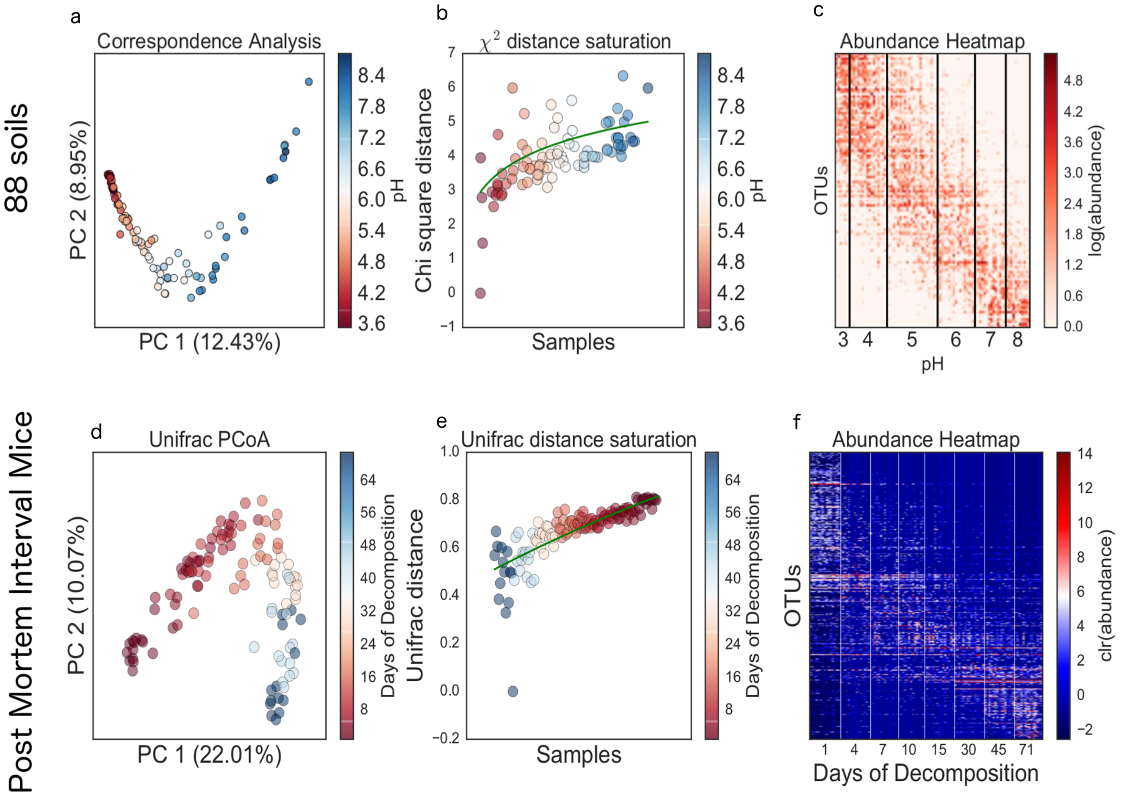


Figure 2.2: Two case studies show casing how horseshoes can appear in the context of soil microbial communities and post-mortem microbial communities. (a) Correspondence analysis of 88 soils. (b) Distance saturation of chi-squared metric, plotting the chi squared distance of the first sample versus all of the other samples. (c) Heatmap of log transformed OTU counts from the 88 soils with the samples sorted by pH and the OTUs sorted by mean pH. (d) Principal Coordinates Analysis of unweighted UniFrac distance. (e) UniFrac distance of a samples from the last time point versus all of the samples. (f) Heatmap of centred log ratio transformed (Equation 2) OTU counts sorted by harvest days.

Case Study 2 - Post Mortem Mice Study

In this study, 120 mice were sacrificed and allowed to decompose on soil. Mice were destructively sampled over approximately 8 weeks[79]. 16S sequencing libraries were generated from total DNA extracted from swabs of the skin on the head, and relative abundance values were calculated for each bacterial OTU. A relative abundance matrix was generated for each library and used as input in PCA. This analysis generated a clear horseshoe (Figure 2.2d) using unweighted UniFrac distance [77], with a gradient with respect to the time since death, possibly reflecting a changing skin microbiome during decomposition of the mouse carcass. When the samples were sorted by time since death using a similar strategy as noted above, a band table emerges (Figure 2.2f). Also, the unweighted UniFrac distance analysis appears to have the same saturation property as observed previously with Euclidean distance and Chi-squared distance. It is important to note that highest possible UniFrac distance is 1, suggesting that this distance metric can also be saturated. In Figure 2.2e, while the distance hasn't completely saturated, these distances are quickly approaching the theoretical maximal UniFrac distance.

The striking changes in microbial communities during decomposition are associated with dramatic environmental biochemical changes, including increased pH, ammonia, and total nitrogen, all measured in soil beneath the mouse carcasses. Correspondingly, microbial communities are predicted to increase in gene abundance of important nitrogen cycling pathways such as amino acid degradation (e.g. glutamate dehydrogenase, lysine decarboxylase, ornithine decarboxylase) and nitrate reduction (e.g. nitrate and nitrite reductase). Bacterial taxa in the families Chromatiaceae (OTU 46026, 4482362) and Rhizobiaceae (OTU 4301099) are involved in nitrogen metabolism and become abundant as mouse bodies progress through the stages of decomposition (e.g. Fresh, Active Decay, Advanced Decay). As shown in Figure S1, all of these OTUs peak at specific timepoints. The two Chromatiaceae OTUs peak during Active Decay (bloating and purge of fluids) at 15 days of decomposition. The Rhizobiaceae OTU peaks during Advanced Decay (sinking and sagging flesh) at 30 days of decomposition and when pH, ammonia, and total nitrogen were measured at their highest levels [79].

To further validate if saturation leads to horseshoes, a new distance metric Earth Mover Band Aware Distance) (EMBAD) (Earth Mover Band Aware Distance) was engineered to be non-saturating as a proof of concept (Supplemental Methods Appendix B). This distance metric uses prior knowledge about the ordering of the band table, and is determined by calculating the flow between two samples. As shown in Figure S1a, sample 1 and sample 2 each have 4 species proportions. To calculate the distance between sample 1 and sample 2, the probability mass of species 1 and species 2 needs to be shuffled over to species 3 and species 4. This concept is analogous to computing maximum flow along a pipe, and can be calculated using Earth Mover's distance [80, 81, 16].

For the 88 soils (Figure S1b), the EMBAD was applied to the pH sorted table. Therefore, even if two samples are not overlapping, samples closer together will have a smaller distance than samples farther apart in the gradient. This is because the distance is defined to be not saturating and explicitly accounts for the pH gradient. The same strategy was employed for the postmortem interval mice (Figure S1c), sorting the table by decomposition days. The PCoA plots resulting from these applications of EMBAD suggest that a non-saturating distance metric could remove the horseshoe effect from lower dimensional projections of these abundances. This provides further evidence that this saturation property could explain the the horseshoe phenomenon.

For the 88 soils study a Permutational Multivariate Analysis of Variance (PERMANOVA) test investigating the difference between soils with a pH less than 3, and soils with a pH greater than 8. With the EMBAD distance metric the PERMANOVA gave a pseudo F-statistic of 650.5 and a p-value of 0.0003, which has a much larger effect size compared to the original Chi-squared distance metric with a pseudo F-statistic of 3.8 and a p-value of 0.0004 with 9999 permutations. A similar trend was observed in the post mortem interval mice study when testing the first decomposition day to the last decomposition day using PERMANOVA. The EMBAD distance metric had a pseudo F-statistic of 439.8 and a p-value of 0.0001 with 9999 permutations, which has a larger effect size than the Unifrac distance metric, which had a pseudo F-statistic of 25.5 and a p-value of 0.0001. This method is relieved from misinterpretations of data due to horseshoes and arches and facilitates the interpretation of taxonomic units along biologically significant gradients that reflect the selective pressure of these factors on the distribution of microbes.

In light of the benefits of engineering a non-saturating distance metric, the EMBAD distance metric requires the gradient to be known a priori. Generalizing this approach in the absence of known gradients is a difficult problem would require an exhaustive using known algorithms. Specifically, this problem falls under the category of NP-hard problems (Supplemental Proof 2 in Appendix B). In the 88 soils study and the post mortem mice study, we were fortunate to be able to infer the underlying band table with known metadata.

The band patterns we observe here are probably very common in ecology studies investigating species distribution patterns across spatial or temporal gradients. The pattern confirms microbial ecological fundamentals, i.e. bacteria have acquired unique adaptations to the environment and occupy either a broad range or very specific niches. In our case studies of the 88 soils - and the postmortem mice we confirmed that by using a band table pattern analysis approach, bacterial species show different adaptations to pH and bacterial diversity changes over time during decomposition of mice carcasses. The band pattern approach we apply here represents an additional method to visualize differences between microbial communities.

On the basis of our observations described here, the horseshoe effect appears in dimensionality reduction techniques due to the saturation property of distance metrics. While we have tested only a few distance metrics, it is suspected that a vast majority of these distance metrics exhibit the same property, which would also explain why horseshoes are encountered so frequently across many different fields. The saturation property has also been observed in multiple other fields, and other studies from different disciplines have led to similar conclusions [72]. In spite of the saturation property of distance metrics, identifying horseshoes is still highly useful for identifying patterns concerning niche differentiation. These insights can ultimately guide additional statistical analyses, such as network analyses and indicator taxon analyses, to facilitate the targeted characterization of microbial niches.

2.2 Materials and Methods

All analyses can be found below on github `https://github.com/knightlab-analyses/horseshoe-analyses`. The mean gradient used for the 2 case studies was calculated as follows.

$$\overline{g}_x = \sum_{i=1}^{N} g_i \frac{x_i}{\sum_{j=1}^{D} x_j} \tag{2.1}$$

Where x_i is the proportion of OTU x in sample i , $\overline{g}_x$ is the mean gradient of OTU x, and g_i is the sample gradient at sample i. This calculation can be found in the gneiss package under the function **mean_niche_estimator**. The function used to sort the tables in Figure B.1c used **niche_sort**. In the 88 soils study, the table was sorted by sample pH and the mean pH of the samples that the organisms were observed in. In the post mortem mice study, the table was sorted by the days of decomposition and the mean day of the samples of that the organisms were observed in.

The heatmap in Figure B.2f and the abundances in Figure S2 were normalized using the center log ratio Center Log-ratio (clr) transformation given by the following equation.

$$clr(x) = \left[\log \frac{x_1}{g(x)}, ..., \log \frac{x_D}{g(x)}\right] = \log x - \overline{\log x} \tag{2.2}$$

Where $g(x) = \sqrt[n]{\prod_{i=0}^{n} x_i}$ is the geometric mean and $\overline{\log x} = \log g(x) = \frac{1}{n}\sum_{i=0}^{n} \log x$ is the average of the log transformed values. A pseudocount of 1 is added to all of the counts to prevent logarithms of zero occurring.

Analyses were performed using Scipy, Numpy, Matplotlib, Seaborn, Scikit-bio and Gneiss.

2.3 Acknowledgements

We thank Noah Fierer for his input on the analysis of the 88 soil samples and Dan Knights for discussion on the EMBAD metric. We also acknowledge Amnon Amir and Tomasz Kosciolek for their insights into the horseshoe effect. Finally, we thank Susan Holmes for her insights on previous work understanding the horseshoe effect.

J.T.M. was funded by NSF GRFP DGE-1144086. J.L.M. and R.K. were funded by the Office of Justice Programs National Institute of Justice (award NIJ-2011-DN-BX-K533).

Chapter 2, in full, is a reprint of the material as it appears in “Uncovering the Horseshoe Effect in Microbial Analyses” James T. Morton, Liam Toran, Anna Edlund, Jessica L. Metcalf, Christian Lauber, Rob Knight *mSystems*, 2, 2017. The dissertation author was the primary investigator and first author of this paper.

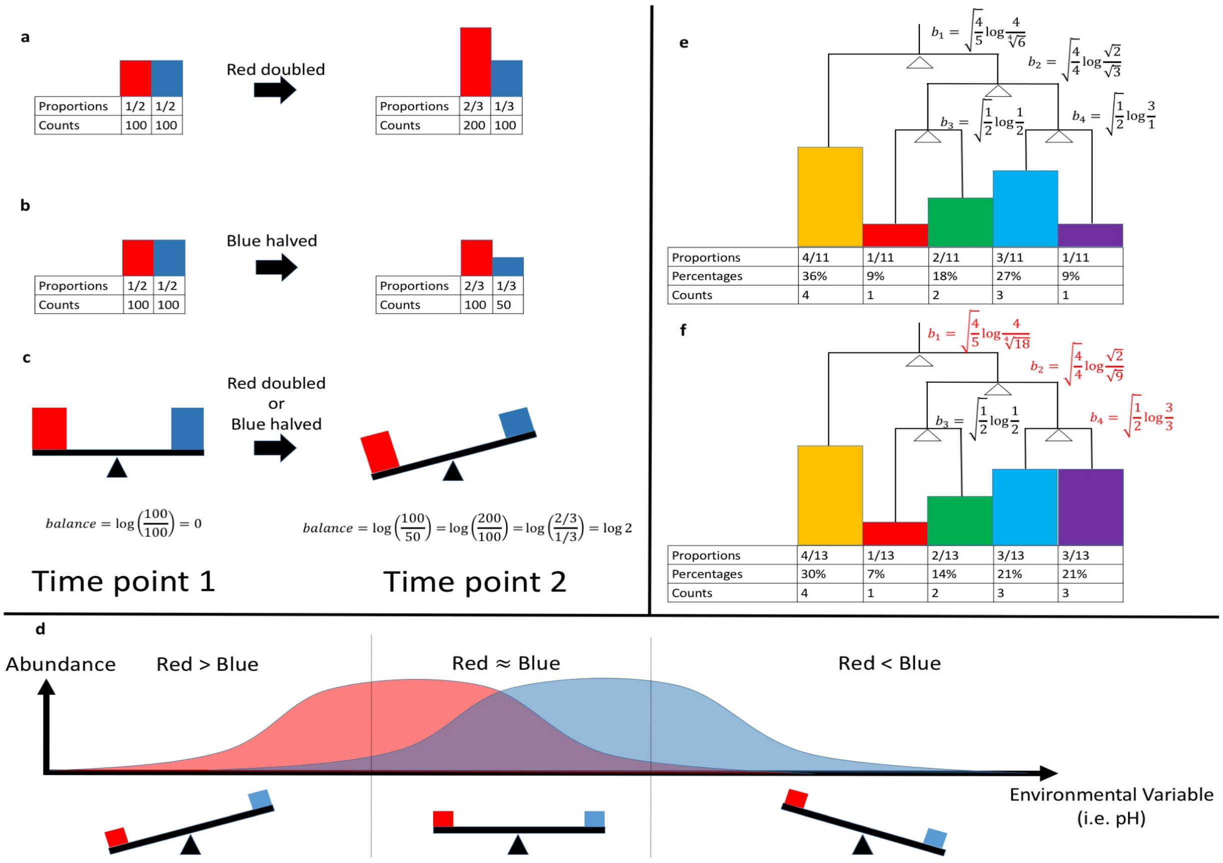
Chapter 3

Balance trees reveal microbial niche differentiation

Advances in sequencing technologies have enabled novel insights into microbial niche differentiation, from analyzing environmental samples, to understanding human diseases and informing dietary studies. However, identifying the microbial taxa that differentiate these samples can be challenging. These issues stem from the compositional nature of 16S Ribosomal RNA (rRNA) gene data (or, more generally, taxon or functional gene data), which changes in the relative abundance of one taxon influence the apparent abundance of the others. Here we acknowledge that inferring properties of individual bacteria is a difficult problem, and instead introduce the concept of balances to infer meaningful properties of sub-communities, rather than properties of individual species. We show that balances can yield insights about niche differentiation across multiple microbial environments including soil environments and lung sputum. These techniques have the potential to reshape how we carry out future ecological analyses aimed at revealing differences in relative taxonomic abundance across different samples.

3.1 Introduction

The ultimate goal for many microbial ecologists is to fully characterize niches of microbial organisms and understand interactions among taxa. An understanding of how microbial communities are affected by environmental conditions could yield insights into microbial interactions and their role in macro-ecological processes, such as nitrogen fixation [82] and acidification [83]. But despite the extraordinary increase in available data brought about by advances in DNA sequencing, characterizing niche differentiation in microbes remains an outstanding problem, partly due to the difficulty of correctly interpreting compositional data. Broadly speaking, a compositional dataset is represented by relative abundances, or proportions that individually carry no meaning on the absolute abundance of a specific feature (i.e. 20% of 100 and 20% of 10,000 are very different absolute abundances). The constraints associated with compositional data are well known, but unfortunately often neglected in microbial ecology, leading to conflicting interpretations and irreproducible analyses [84, 85] .


a
Red doubled
b
Blue halved
c
Red doubled
or
Blue halved
$balance = \log\left(\frac{100}{100}\right) = 0$
$balance = \log\left(\frac{100}{50}\right) = \log\left(\frac{200}{100}\right) = \log\left(\frac{2/3}{1/3}\right) = \log 2$
Time point 1
Time point 2
d
Abundance
Red > Blue
Red ≈ Blue
Red < Blue
Environmental Variable
(i.e. pH)
e
$b_1 = \sqrt{\frac{4}{5}}\log\frac{4}{\sqrt[4]{6}}$
$b_2 = \sqrt{\frac{4}{4}}\log\frac{\sqrt{2}}{\sqrt{3}}$
$b_3 = \sqrt{\frac{1}{2}}\log\frac{1}{2}$
$b_4 = \sqrt{\frac{1}{2}}\log\frac{3}{1}$
f
$b_1 = \sqrt{\frac{4}{5}}\log\frac{4}{\sqrt[4]{18}}$
$b_2 = \sqrt{\frac{4}{4}}\log\frac{\sqrt{2}}{\sqrt{9}}$
$b_3 = \sqrt{\frac{1}{2}}\log\frac{1}{2}$
$b_4 = \sqrt{\frac{1}{2}}\log\frac{3}{3}$

Proportions	1/2	1/2
Counts	100	100

Proportions	2/3	1/3
Counts	200	100

Proportions	1/2	1/2
Counts	100	100

Proportions	2/3	1/3
Counts	100	50

Proportions	4/11	1/11	2/11	3/11	1/11
Percentages	36%	9%	18%	27%	9%
Counts	4	1	2	3	1

Proportions	4/13	1/13	2/13	3/13	3/13
Percentages	30%	7%	14%	21%	21%
Counts	4	1	2	3	3

Figure 3.1: An explanation of balances and how to interpret them. (a, b) A hypothetical scenario where 2 samples of 2 proportions could explain two different scenarios in the environment. The balance between these 2 proportions is consistent for both scenarios. (c) The balance of Red and Blue species abundances. (d) balances of Red and Blue individuals across an environmental variable. (e, f) The comparison of proportions and balances of two environments in the scenario where the Purple population (i.e. the most right bin) triples. The balances were calculated using the groupings specified by the tree.

We illustrate an example of this problem in Figure 3.1. In this scenario, there are two species, “Red” and “Blue”. At the first time point, there are 100 Red individuals and 100 Blue individuals (Figure 3.1a). At the next time point, the number of Red individuals doubles, yielding 200 Red individuals, and the proportion of Red and Blue individuals becomes 2/3 and 1/3, respectively (Figure 3.1b). Suppose that we do not know the true total number of individuals in the given environment, and can only make inferences about the observed proportions – a common scenario in microbial ecology, where absolute quantification is rarely performed. In Figure 3.1b, the community has the exact same proportions at Time 1 and Time 2 as Figure 3.1a; however, instead of the Red individuals doubling at the second time point, the number of Blue individuals is halved (Figure 3.1c).

This is the problem with compositionality – based on proportions alone, it is impossible to determine whether the growth or decline of any individual species has truly occurred [86], and the inherent feature of one change in abundance driving abundance changes in another species violates assumptions of independence. Analyses that rely on such assumptions, as many statistical approaches do, are thus prone to misinterpretation. For example, traditional correlation metrics such as Pearson and Spearman can be misleading when estimating microbe-microbe correlations [87, 88, 89, 90]. As a result, it becomes a major challenge to specify types of interactions between microbes, such as parasitism, competition, predation or mutualism, as shown in correlations studies in oral, fecal and vaginal samples from the Human Microbiome Project [91, 87]. Even more advanced correlation-detection techniques such as SparCC [87] and SPIEC-EASI [89], struggle with this, and typically require additional assumptions such as sparse Operational Taxonomic Unit (OTU) correlations (i.e. few OTUs are actually correlated with each other). Furthermore, interpreting the resulting network is a major challenge, making it difficult to differentiate between true ecological relationships and random processes [91].

The compositionality problem is also problematic for statistically detecting differentially abundant microbes across environments or between groups — consequently, it is a major barrier to reliably drawing conclusions about realized microbial niches using community sequencing data. Conventional statistical tools such as t-test and Mann-Whitney can incorrectly identify nearly 100% of the taxa present in samples to be significantly different across environments (Figure S1), and univariate tests such as t-tests and ZIG [92] have been shown to mislabel microbes as significantly different across sample groups up to 60% of the time [93]. More advanced tools for differential abundance detection such as (Analysis of Composition of Microbiomes (ANCOM))

[93], are typically designed to control for false-positives and reliably detect differentially abundant species, but require multiple assumptions (i.e. the number of changing microbes across environments is small) and may require complex parameter tuning. To help overcome these issues of compositionality, we explore using the concept of balances, by moving away from inferring changes of individual species to instead inferring changes of microbial sub-communities to study niche differentiation of microbial communities.

3.2 Concept

Balances were first introduced as an exploratory technique in geology [94, 95]. Fundamentally, they overcome the problem of inferring changes in abundance from compositional data by sidestepping it, and instead inferring changes in the balance between particular subsets of the community. To understand the concept, let us revisit the scenario in Figure 3.1a and 3.1b. Instead of examining proportion changes, we can investigate the balance between Red and Blue individuals by taking the log ratio of Red and Blue counts (Figure 3.1c). By looking at the balance of these two species, we avoid incorrectly attempting to infer absolute increases or decreases in their abundances. Instead, we can focus on the balance of the Red and Blue individuals, and directly infer the transition of dominance between these species.

These balances can also be useful for understanding species distributions across different covariates — a key proximate goal of microbial ecology, and one that is both crucial to the larger goal of niche characterization and heavily impacted by problems inherent in compositionality. In Figure 3.1d, the Red individuals tend to exist in the low pH end of the spectrum, while the Blue individuals tend to exist in the high pH end of the spectrum. A single balance can capture information about the transition from a high relative abundance of Red individuals in low pH environments to a high relative abundance of Blue individuals in high pH environments. In low pH environments, the balance is positive, since there are proportionally more Red individuals than

Blue individuals. When the Red and Blue individuals are present in roughly equal proportions, the balance is roughly zero, representing a turning point, transitioning from a Red dominated community to a Blue dominated community. As the pH increases, the balances become increasingly negative, since there are more Blue individuals than Red individuals. This balance effectively encodes for the niche separation of Red and Blue individuals across the pH gradient.

This idea of balances can be extended to multiple dimensions — and more than two taxa — using bifurcating trees. A bifurcating tree can be built relating microbial taxa to each other using any criterion, and balances can be calculated on the internal nodes of the tree from the geometric means of the corresponding sub-trees. The appropriate criterion to build a tree depends on the question at hand. A phylogenetic tree could be used to investigate evolutionary relationships of microbes [96, 13], or hierarchical clustering of environmental variables could be used to explore environmental niches of microbes. To gain more intuition about this, consider Figure 3.1e, in which there are five species and 11 individuals. The four balances (internal nodes in the tree) are calculated by taking the log ratio of geometric means of sub-trees, also known as the isometric log ratio (Isometric Log-ratio (ILR)) transform. The full equation to calculate balances for a single sample is as follows,

$$b_i = \sqrt{\frac{|i_L|\,|i_R|}{|i_L| + |i_R|}} log \frac{g(i_L)}{g(i_R)} \tag{3.1}$$

where b_i is the balance of the at internal node i , i_L, is the set of all species proportions contained in the left sub-tree at internal node i, i_R, is the set of all species proportions contained in the right subtree at the internal node i, g(x), is the geometric mean of all of the proportions contained in vector x, $|i_R|$, is the number of species contained in i_R , and $|i_L|$ is the number of species contained in i_L (see Materials and Materials for more details). Following this equation, in Figure 3.1f b_1 is calculated by taking the log ratio of the Yellow species and the geometric mean of the Red, Green, Blue, and Purple species.

It's also important to note that the some of the balances don't impact each other. For instance, the changes in b_4 do not impact the changes in b_3, just because these balances don't share any common tips. This is crucial, because this property allows us to ignore some of the variance of the balances towards the tips of the tree, and focus on the balances closer to the root of the tree. These balances toward the root of tree capture the most information, since they contain a significant proportion of tree tips. As a result, these high level balances have the potential to explain large shifts in these microbial communities. The choice of the tree can allow for analysts to embed prior knowledge into the structure of the tree to test for these large community shifts.

Here, we will discuss two studies from which novel insights were gained from this application. While there are many compositionally aware tools available that are designed to identify microbial interactions and abundance fluctuations, we will refrain from benchmarking balances against these tools, as balances answer a conceptually different question. These analyses are not restricted to analyzing ratios of individual OTUs and can be easily extended to analyze ratios of subcommunities.

3.3 Results

3.3.1 Case Study #1 – Balances of pH-driven subcommunities in soils

In this study [77], 88 soil samples were collected from North and South America, along with many edaphic measurements. The study reported that there was a strong correlation between pH and species richness, suggesting that pH was a strong driver behind fluctuations in soil microbial communities. *Acidobacteria* were found to be negatively correlated with pH and *Actinobacteria* and *Bacteroidetes* to be positively correlated with pH, while alpha-, beta- and gammaproteobacteria were not correlated with pH at all. These correlation analyses are a little misleading, since the pH was correlated with each of the phyla independently. The problem with this approach is that it does not account for all of the other phyla: similar to the argument made in Figure 3.1b, the change in a single phylum could also be explained by correlated changes in all of the other phyla. Here, the negative correlation between *Acidobacteria* and pH could also be caused by the positive correlation between *Bacteroidetes* and pH. Additionally, we cannot determine whether the alpha-, beta- and gammaproteobacteria are correlated with pH or not. Another possibility is that these three phyla could be positively correlated with pH, while *Acidobacteria* is not correlated with pH. However, *Bacteroidetes* may be so strongly correlated with pH that *Acidobacteria* appears to be negatively correlated with pH, and the other three phyla not correlated with pH at all. This scenario is one of the infinite possible underlying relationships that can explain these observed correlations.

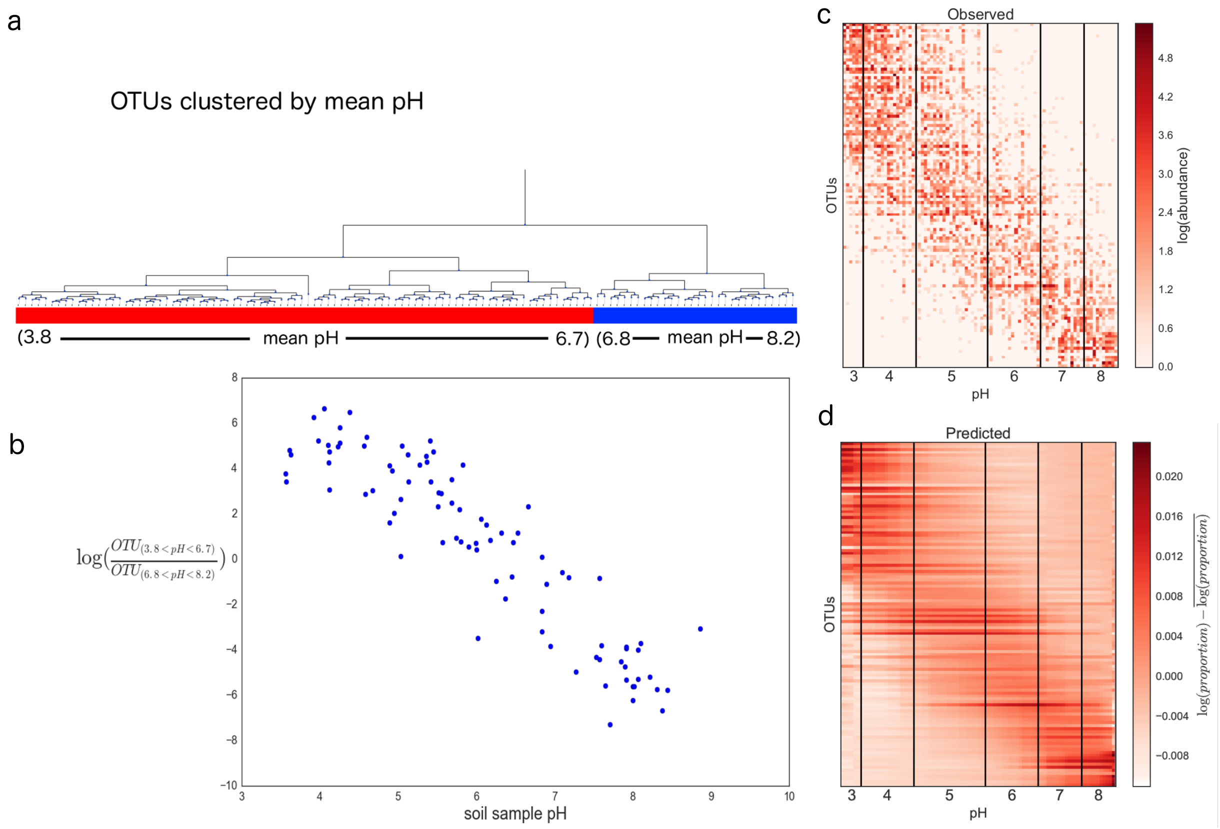


Figure 3.2: The application of balances on a soil microbial dataset to identify microbial partitioning with respect to pH.(a) Hierarchical clustering of closed ref OTUs based on mean pH. (b) The balance of low pH associated organisms ($3.8 <$ mean pH < 6.7) and high pH associated organisms ($6.8 <$ mean pH < 8.2). (c) Observed OTU counts sorted by pH. (d) Predicted OTU proportions from ordinary least squares linear regression on balances sorted by pH. The coefficient of determination was 35%, showing that 35% of the variation in the microbial community abundance data can be predicted by pH alone.

At a first glance, uncovering the true correlations correctly appears to be a hopeless cause. This is where balances become useful. Rather than attempting to correlate individual phyla against pH, we will group OTUs together according to their difference in mean pH (Figure 3.2a), and investigate how these balances of groups changes with respect to pH (See Materials and Methods on hierarchical clustering). This circumvents the dependence issue noted previously. We do not need to worry about subgroups within the left and right subtrees of a balance to be influencing each other, due to the independence property shown in Figure 3.1ef.

The balance concept proves to be a very powerful technique for investigating how these groups of organisms change relative to each other as pH increases. Recall the cartoon example in Figure 3.1d. If there are two distinct unimodal species distributions, the balance pivots from being weighted by Red in low pH, to being weighted by Blue in high pH. The exact same phenomenon is occurring here, except there are multiple species on the left end of the balance, and multiple species on the right end of the balance.

As shown in Figure 3.2b, there is a well defined trend of low pH OTUs ($3.8 <$ mean pH < 6.6) gradually being overtaken by high pH OTUs ($6.7 <$ mean pH < 8.2) as the pH increases, forming a nice linear trend defined by the top balance in the tree shown in Figure 3.2a. If we were to sort the samples by their mean pH, and the OTUs by their mean pH (Equation 3), a well defined band pattern appears. Here, it is clear that OTUs with a mean pH less than 3 rarely have nonzero counts above 8. Likewise, OTUs that have a mean pH more than 8 rarely have nonzero counts below 3. If we were to tie in this band pattern in Figure 3.2c together with the balance vs pH trends shown in Figure 3.2b, we would obtain a very different interpretation from the original study. OTUs tend to be observed in very specific pH ranges, but not commonly observed outside of these ranges. This ties together with some concepts in niche theory - OTUs are more suited to live within a designated range of pHs. And if they are placed outside of this pH range, they are outcompeted by other organisms who are more suited to live within the given pH range.

These patterns were completely missed when only looking at the phylum level in the original study. In fact, based on the calculated mean pH values for each OTUs, it is observed that OTUs from all of the phyla mentioned in the study are widely distributed across the pH gradient. As an extreme example, OTUs from the family *Bradyrhizobiaceae* were observed to be present in both ends of the spectrum, some present at pH values as low as 5.36, while others present at a pH as high as 6.75. These are astronomical differences, considering that 95% of the OTUs have a mean pH that falls between this range. This provides additional justification for building a tree based on mean pH, rather than bacterial phylogeny.

Finally, these balances can be used to build predictive models. Using ordinary least squares on the calculated balances, the entire microbial community profile can be predicted using pH alone with an R^2 of 0.35. This means that pH alone explains over 35% of the total variation in entire soil microbial communities across North and South America. The resulting fit can be transformed back to proportions to yield the predicted proportions (Figure 3.2d). From this heatmap, the key patterns are still retained, such as the band pattern apparent in Figure 3.2c. There are many regression techniques published that attempt to use microbial abundances to predict covariates, such as the post-mortem interval [79] or body mass index [97]. This approach is the first of its kind to attempt to address the reverse problem to predict entire microbial community distributions based on environmental variables. These predictions were enabled by the powerful fundamental properties of balances.

3.3.2 Case Study #2 – Balances of pH-driven subcommunities in a lung sputum culture microcosm

In this study, lung sputum samples were collected from 16 cystic fibrosis (Cystic Fibrosis (CF)) patients. These sputum samples were then grown in a capillary tube culture system (Winogradsky Cystic Fibrosis system) that mimics the conditions of a lung bronchiole [98]. These samples were placed into separate tubes and the pH of the media was adjusted from 5 to 8.5 at intervals of 0.5 to determine how the microbial community changed with respect to pH. After growth in the capillary tubes, the communities were assessed using 16S rRNA gene amplicon sequencing.

One of the difficulties in this study was characterizing pathogenic bacteria. Early on in this case study, the only significant finding discovered was that patients had different lung sputum microbiomes (Figure 3.3a). It was hypothesized that there was a subcommunity of low pH organisms and a subcommunity of high pH organisms that periodically appeared and disappeared in CF lung sputum. However, these changes could not be detected using available statistics, likely due to the compositionality problem. Since the different CF patients had idiosyncratic lung communities, they ended up having different OTUs responding across the laboratory pH gradient, yielding insufficient statistical power to detect changes in any given OTU. As a result, when these lung sputum communities were placed into different media and studied, it was not clear exactly what organisms were a part of this low pH or high pH subcommunity.

Balances are a natural solution to this problem. In addition to probing for similar patterns to those observed in the previous study, balances are well adapted as a transformation for standard statistical analyses. Since Euclidean operations directly translate into perturbation and powering operations on proportions [99, 100], many publicly available statistical tools can be applied to directly to balances. For this study, we opted to use Linear Mixed Effects models to test for pH differences while simultaneously accounting for all of the differences between lung microbiomes across CF patients. Based on prior analyses with pH in soils, the tree was built using the exact same strategy (See Methods and Materials). Significant balances testing for pH were determined with a p-value cutoff at 0.05 after Bonferroni correction.

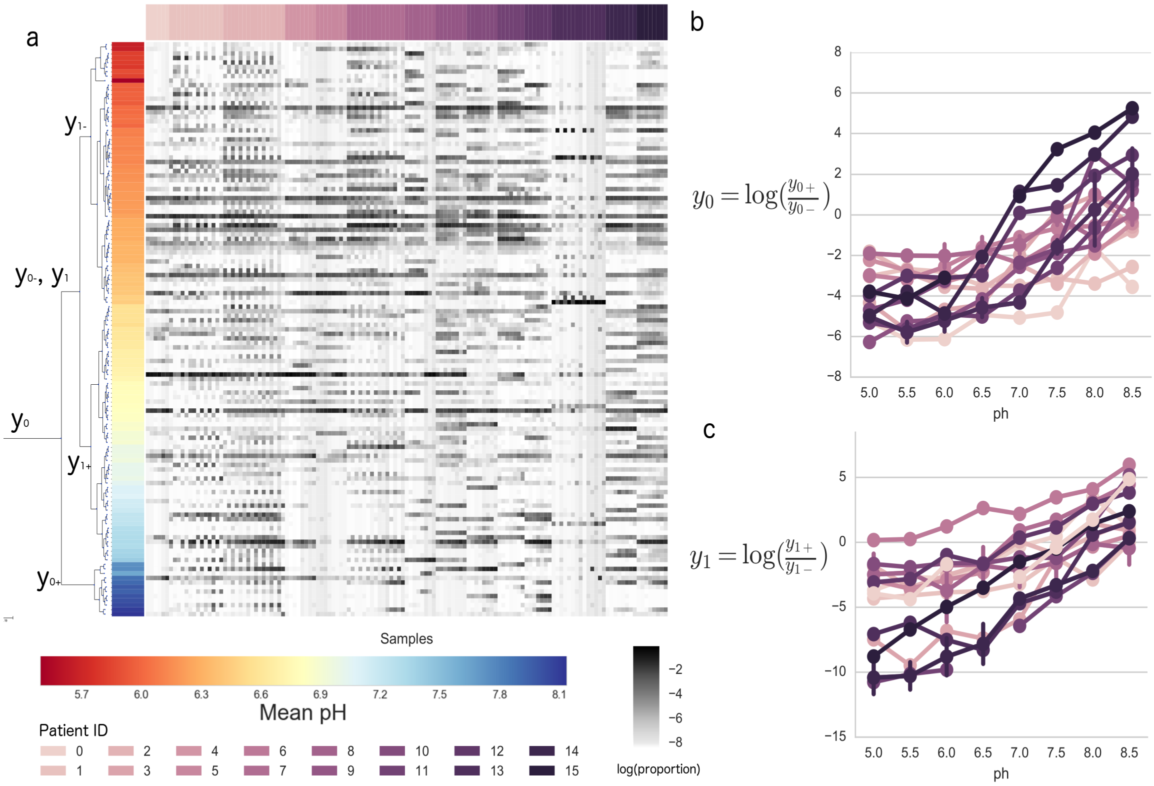


Figure 3.3: The application of balances on a cystic fibrosis dataset to identify microbial partitioning with respect to pH.(a) A bifurcating tree generated from hierarchical clustering of OTUs based on mean pH. The size of the internal nodes is inversely proportional to the p-value of the linear mixed effects model test on pH for that given balance. A heatmap of all of the OTU abundances sorted by patient. OTUs were log transformed and centered across rows and columns. These abundances are aligned with the tips of the tree. (c) The progression of the top balance over the pH for all of the patients. (d) The progression of the second top balance over pH for all of the patients.

A heatmap relating pH to OTU abundances across these samples does not yield clear trends (Fig 3a). But even though we don't see a clear pattern in the heatmap, with the balance approach, we can still observe niche differentiation across the pH gradient. In Figure 3.3b, y0 represents the log ratio of all of the high pH OTUs ($7.6 <$ mean pH < 8.12) over all of the low pH OTUs ($5.4 <$ mean pH < 7.4). As the pH of the samples increases, the balance increases, likely because the low pH OTUs are becoming increasingly less abundant compared to the high pH OTUs (p-value=7.5×10^{-46}). The same pattern is even more apparent in y1 (Figure 3.3c). The low pH OTUs ($5.4 <$ mean pH < 6.4) become increasingly less abundant than high pH OTUs ($6.5 <$ mean pH < 7.4) as the sample pH increases (p-value=2.25×10^{-67}). When Bonferroni multiple hypothesis correction was applied to these tests, the p-values were rounded down to zero. While these patterns were not obvious when looking at the raw proportions, the balance tree approach shows very well defined trends among groups of OTUs. This can be done because even though individual OTUs may be sporadically distributed across the original samples, OTUs that thrive in similar pH niches grouped together on the environmental balance tree. It is clear from Figure 3.3b and c that there is a transition from low pH organisms to high pH organisms along the pH gradient. Even though the CF patients don't have the same lung microbiomes, they contain OTUs that behave the same with respect to pH. This pattern would not have been nearly as apparent without clustering the OTUs by mean pH and accounting for the patient effects in the linear mixed models.

3.4 Discussion

In this study, we have demonstrated the benefits of applying balances to infer niche differentiation in microbes. In the first case study, we have outlined the challenge of performing correlations of OTUs versus environmental variables, and showed how balances can capture information about species turnover across the pH gradient, which allowed us to build a model to predict microbial proportions based on pH alone. In the second case study, we identified the challenges of studying individual OTUs due to similar niches being occupied by drastically different OTUs across different patients. Balances coupled with linear mixed models allowed us to obtain more statistically robust results, which were also more informative with respect to the differences in distribution of microbes across environmental niches.

There are numerous additional benefits of analyzing species balances instead of individual species counts. First, balances are known to be scale-invariant, so balance trees naturally correct for differences in sequencing depth without requiring rarefaction (Equation S1 in Appendix C) and avoid many of the limitations associated with this procedure [101]. Second, balances are sub-compositionally coherent, which means that changes in non-overlapping sub-communities do not impact each other. For instance, in Figures 1e and 1f, the Purple population triples and balances change because they explicitly contain the Purple species. In contrast, the balance red and green log ratio does not change between these two scenarios because it does not relate to the Purple species (in fact, it only accounts for the Red and Green species). This is not the case when observing the raw proportions, from which it appears as though everything is changing, even though the Purple species is the only changing species. This phenomenon has previously been noted [93] and can lead to extremely high false positive rates with some standard statistical techniques such as Pearson correlations or t-tests on proportions. More discussion about this issue can be found in Figure S1 (in Appendix C). Third, arithmetic operations on balances directly translate into perturbation and powering operations on proportions [99, 100], which can capture information about relative growth and decay of species. This ultimately opens the door for applying standard statistical techniques, such as multiple linear regression [102] and linear mixed effects models nested design statistics directly to balances, providing additional justification for the analyses performed in the case studies. We have shown this in the two case studies. Finally, balances are permutation invariant. Species can be sorted in any order deemed appropriate. Along the same lines, these species can be rearranged into any arbitrary grouping represented as a bifurcating tree. These trees can be built to address the questions at hand, whether it be studying species turnover across pH gradients, or even uncovering the relationships between phylogenetic clades. In fact, balances can be thought as being utilized as an ordination technique, since every bifurcating tree forms an orthonormal basis in the Aitchison Simplex [94].

Although the concept of balances does not address questions about properties of individual bacteria, it does answer higher-level questions concerning interactions among groups of organisms, which are arguably much more interesting from an ecological point of view. These questions can be based either on the phylogenetic tree of the bacterial community, or on environmental clustering. There is still room for improvement on utilizing balances. For example, the issue of zeroes still remains, because the logarithm of zero is undefined. Currently, the common approach is to add a pseudo-count [103]. However, an appropriate tree choice can mitigate this issue, because the zeroes can be explicitly aggregated in some scenarios (Figure S2 and Figure S3 in Appendix C). Along the same lines, issues can arise from low-coverage samples. If sampling is not saturated, many OTUs have low read counts, and the balances towards the tips of the trees can be highly volatile. This is because the absolute change between one or two reads may be small for low abundance OTUs, but this will lead to large changes in log ratios, which lead to spurious signals at the tips of the tree. As a rule of thumb, balances towards the root of the tree are more trustworthy than those at the tips of the tree.

The balances approach will be key for analyzing functional roles of OTUs. It is known that in environments like the human gut, people share very few OTUs with each other, but have roughly the same proportions of functional genes [77]. This suggests that there is substantial functional redundancy across OTUs, which has been observed previously in time series studies in the context of infection [104] — in other words, in these microbial communities many players might be sporadically distributed across similar niches. This phenomenon could explain the sparse nature of 16S relative abundance data, and why similar environments such as human guts share few common OTUs. Such distributions pose tremendous challenge to analyses based around identifying the niche occupancy of individual OTUs. By instead permitting the statistical comparisons to be performed across nested groups of OTUs with similar distributions, it becomes possible to robustly identify patterns of niche differentiation without requiring sufficient information be present in the abundances of each individual taxon. Identifying common functional roles of potentially diverse organisms, and analyzing the balances between these groups could significantly simplify analyses in future amplicon studies. The ability to construct such trees would enable rapid characterizations of environmental niches, and the corresponding functional roles of the microbes occupying in these niches.

All in all, balance trees are an extremely powerful tool for analyzing relative abundances and uncovering patterns associated with niche differentiation, while avoiding the issues associated with compositionality and enabling the application of conventional statistical tools. This will ultimately open the doors for extensive mining of ecologically relevant patterns.

3.5 Methods and Materials

All analyses can be found in the attached IPython notebooks. The core functions required to perform the balance basis calculations, tree visualization tools, and statistical analyses can be found in `https://github.com/biocore/gneiss`. The IPython notebooks used to carry out all of the analyses can be found in the gneiss repository. All code has been extensively unit-tested and documented.

The core compositional statistics and tree data structures were are part of scikit-bio 0.4.1 and beyond. The hierarchical clustering was performed using Scipy. Pandas and Biological Observation Matrix (BIOM) [105] were used to store and manipulate the OTU tables and the metadata files. Seaborn, matplotlib and ETE [106] were used for the visualizations.

The isometric log ratio transform is an isomorphism (i.e. a function) that can map proportions to balances one to one [99]. These balances can be calculated as shown in Equation 1. Alternatively, they can be calculated using a linear transformation with an orthonormal basis e. This orthonormal basis can be calculated as follows

$$e_l = C\left[\exp(\underbrace{0,...0}_{k},\underbrace{a,...a}_{r},\underbrace{b,...b}_{s},\underbrace{0,...0}_{t})\right] \tag{3.2}$$

$$a = \frac{\sqrt{s}}{\sqrt{r(r+s)}} \quad and \quad b = \frac{-\sqrt{r}}{\sqrt{s(r+s)}}$$

where e_l refers to the balance axis aligned with the internal node l. $C[x]$ denotes the normalization operation to normalize all of the OTU abundances to proportions that add up to 1. r refers to the number of tips in the left subtree, s refers to the number of tips in the right subtree, k refers to number of tips to the left of the left subtree and t refers to the number of tips to the right of the right subtree. Since e forms an orthonormal basis, it must have unit norm and every pair of axes in e must be orthogonal. The square root term in Equation 1 is a normalization factor which was required for unit norm in Equation 2 (12). Since it is not possible to take a logarithm of zero, a pseudocount of 1 was added to all of the abundance. While this is a problem being addressed by the field, this technique is one of the more commonly used techniques [103].

The mean pH used for the 2 case studies was calculated as follows.

$$\overline{g}_x = \sum_{i=1}^{N} g_i \frac{x_i}{\sum_{j=1}^{D} x_j} \tag{3.3}$$

Where x_i is the proportion of OTU x in sample i , g_x is the mean pH of OTU x, and g_i is the sample pH at sample i. This calculation can be found in the gneiss package under the function **mean_niche_estimator**. The function used to sort the tables in Figure 3.2c used **niche_sort**. The resulting tree was built using UPGMA [107] is shown in Figure 3.2a and Figure 3.3a, and can be generated using the scipy linkage function.

This regression model is implemented in gneiss under the **ols** function. The analysis can be found in the IPython notebooks on the gneiss repository under the ipynb folder in **88soils.ipynb**. To focus on the highest abundant organisms, only OTUs that had more than 100 reads in the entire study were considered.

The linear mixed effects model is implemented in gneiss under the **mixed** functions, and the analyses can also be found in the IPython notebooks in the ipynb folder in **cfstudy.ipynb** In case study 2, only OTUs that had more than 500 reads were considered.

The WinCF system was used according to the methods in [98], except only the pH dye media variable was used. The media was buffered at 0.5 units of pH from 5 to 8.5 using calculated proportions of phosphate buffer and NaOH or HCl. Sputum samples were collected from CF patients after expectoration or induced expectoration of sputum according to the UCSD IRB approved project #081500, and were inoculated in triplicate into capillary tubes containing the eight different pH buffered media. These eight sets of tubes in triplicate from 18 patients was then incubated at 37^{o}C for 48 hours. The media was then removed, bacterial DNA extracted, and variable region 4 of the 16S rRNA gene was amplified and sequenced on the Illumina MiSeq platform using Earth Microbiome Project benchmarked protocols [108, 109]. Data were processed using QIITA and OTUs were calculated using closed reference clustering at the 97% identity cutoff for both the 88 soils and the CF study.

3.6 Data availability

Data for case study 1 was retrieved from Qiita (study ID 103). Data from case study 2 was retrieved from Qiita (study ID 10511).

3.7 Acknowledgements

We first acknowledge Jonathan Friedman for the original idea of applying balances to analyze microbial communities. We also acknowledge Justin Silverman and Lawrence David, in addition to Liam Toran, Tomasz Kosciolek, and Amnon Amir, for their insights and discussion on balances. In addition, we are grateful for the input from Christian Lauber and Noah Fierer concerning case study 1. Finally, we thank all of the scikit-bio developers, especially Jorge Cañardo Alastuey, Evan Bolyen, Jai Rideout, and Greg Caporaso, for reviewing the compositional statistics submodule in scikit-bio.

J.T.M. was funded by NSF grant GRFP DGE-1144086 and NSF grant IGERT 1144807

under the IQ Biology program at the University of Colorado Boulder. R.A.Q. was funded under the Cystic Fibrosis Research Innovation Award from Vertex Pharmaceuticals. This work was funded under Alfred P. Sloan Foundation grants G-2015-13933 and G-2015-13979 and National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grant P01DK078669.

J.T.M. led the software development, benchmarking, and manuscript writing and developed the idea of applying regression to balances. J.S. contributed the idea of applying linear mixed-effects models to balances and named the software package. R.A.Q. collected the CF lung sputum samples. D.M., A.G., J.A.N.-M., and Y.V.-B. reviewed the code in Gneiss. M.L. reviewed the mathematical notation. All authors wrote and proofread the manuscript.

Chapter 3, in full, is a reprint of the material as it appears in “Balance Trees Reveal Microbial Niche Differentiation” James T. Morton, Jon Sanders, Robert A. Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A. Navas-Molina, Se Jin Song, Jessica L. Metcalf, Embriette R. Hyde, Manuel Lladser, Pieter C. Dorrestein, Rob Knight *mSystems*, 2, 2017. The dissertation author was the primary investigator and first author of this paper.

Chapter 4

Establishing microbial measurement standards with reference frames

Differential abundance analysis is controversial throughout microbiome research. Current gold standard approaches require laborious measurements of total biomass to accurately determine taxonomic shifts among samples. Therefore, most studies rely on making conclusions based off changes in relative abundance. We highlight commonly made pitfalls in comparing relative abundance across samples and identify two solutions that reveal microbial changes without the need to estimate total biomass. We define the notion of “reference frames”, which provide deep intuition about the compositional nature of microbiome data. In an oral time series experiment, reference frames alleviate false positives and produce consistent results on both raw and cell count normalized data. Furthermore, reference frames identify consistent, differentially abundant microbes previously undetected in two independent published datasets from subjects with atopic dermatitis. These methods allow re-assessment of published relative abundance data to reveal reproducible microbial changes from standard sequencing output without the need for new molecular assays.

Introduction

Next-generation sequencing data used to study the microbiome is inherently compositional and routinely provides information in the form of relative abundances, independent of the total biomass of the original sample. Numerous analytical approaches including rarefaction[110], median[111], and quantile normalization[111, 112] have been proposed for comparing compositional samples. However, these analytical solutions cannot control false discovery rates[113, 114], and their application contributes to lack of reproducibility among microbiome studies[115]. Here we illustrate mathematical challenges in analyzing compositional microbiome data from DNA sequence reads, and define the concept of “reference frames” for inferring changes in abundance.

4.0.1 Why using relative abundance data to evaluate changes in abundance can be misleading

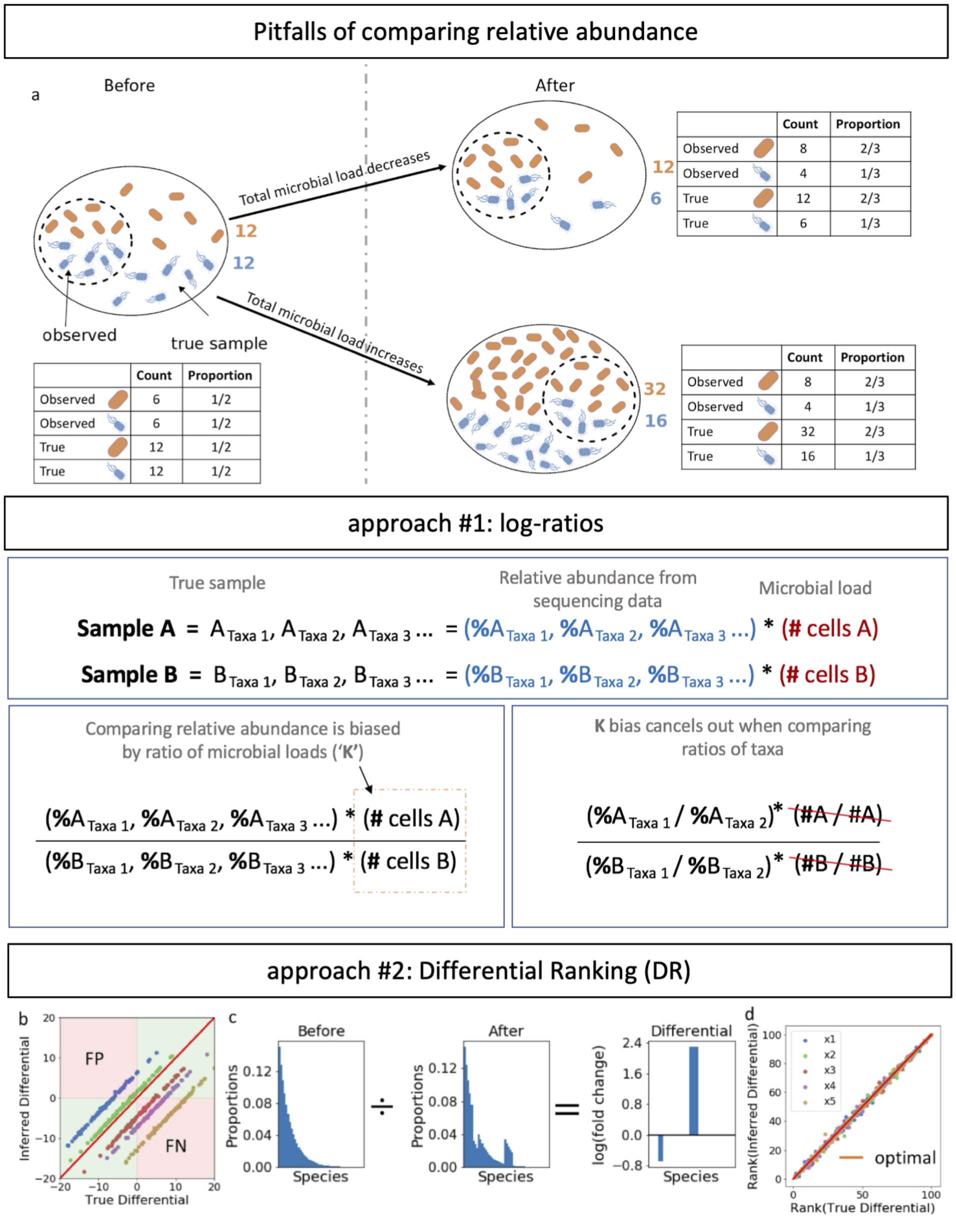
To illustrate the pitfalls of inferring changes in abundance among samples using relative abundance data, consider the following example (Fig 1). Samples from a population containing only two taxa (orange and blue) are collected pre- and post-treatment. Before treatment, the two taxa occur in equal proportions. After treatment, the orange taxon is twice as abundant as the blue taxon. Did orange increase and blue decrease?

Many different scenarios could actually lead to the same observation. For example, the orange taxon could quadruple and the blue taxon only double. The orange taxon could remain constant, and the blue taxon halved. Or the orange taxon could of halved, but the blue taxon could decrease four-fold. Because we only observe relative abundance data, we cannot differentiate among these outcomes, which have markedly different biological significance. An infinite number of different outcomes produce the same 2:1 ratio of orange to blue, greatly complicating the generation of a meaningful null hypothesis and therefore yielding misleading p-values when the incorrect null hypothesis is chosen.

4.0.2 Microbiome measurement data is inherently compositional

Multiple processing steps are required to generate microbiome sequencing data. Samples are collected from a much larger population (e.g. fecal material from the gut, or water sample from the ocean). From these samples, a subsample is used for DNA extraction (e.g. a swab from a fecal sample, or an aliquot of a water sample). Another subsample of the extracted DNA is then used as input for PCR, a subset of the resulting amplicon is pooled into a library, and a subset of the library is sequenced.

By the time quality-filtered sequencing data is obtained, the sequences reflect only a small subset of the population and are not an accurate representation of the microbial load in


Pitfalls of comparing relative abundance
a
Before
After
Total microbial load decreases
Total microbial load increases
12
12
12
6
32
16
observed
true sample
approach #1: log-ratios
True sample
Relative abundance from sequencing data
Microbial load
Sample A = $A_{Taxa\ 1}, A_{Taxa\ 2}, A_{Taxa\ 3} \ldots = (\%A_{Taxa\ 1}, \%A_{Taxa\ 2}, \%A_{Taxa\ 3} \ldots) * (\#\ cells\ A)$
Sample B = $B_{Taxa\ 1}, B_{Taxa\ 2}, B_{Taxa\ 3} \ldots = (\%B_{Taxa\ 1}, \%B_{Taxa\ 2}, \%B_{Taxa\ 3} \ldots) * (\#\ cells\ B)$
Comparing relative abundance is biased by ratio of microbial loads ('K')
$\frac{(\%A_{Taxa\ 1}, \%A_{Taxa\ 2}, \%A_{Taxa\ 3} \ldots) * (\#\ cells\ A)}{(\%B_{Taxa\ 1}, \%B_{Taxa\ 2}, \%B_{Taxa\ 3} \ldots) * (\#\ cells\ B)}$
K bias cancels out when comparing ratios of taxa
$\frac{(\%A_{Taxa\ 1} / \%A_{Taxa\ 2})^{*} \cancel{(\#A / \#A)}}{(\%B_{Taxa\ 1} / \%B_{Taxa\ 2})^{*} \cancel{(\#B / \#B)}}$
approach #2: Differential Ranking (DR)
b
FP
FN
Inferred Differential
True Differential
20
10
0
−10
−20
−20
−10
0
10
20
c
Before
After
Differential
Proportions
0.12
0.08
0.04
0.00
Species
÷
=
log(fold change)
2.4
1.6
0.8
0.0
−0.8
d
Rank(Inferred Differential)
Rank(True Differential)
100
80
60
40
20
0
0
50
100
x1
x2
x3
x4
x5
optimal

		Count	Proportion
Observed		8	2/3
Observed		4	1/3
True		12	2/3
True		6	1/3

		Count	Proportion
Observed		6	1/2
Observed		6	1/2
True		12	1/2
True		12	1/2

		Count	Proportion
Observed		8	2/3
Observed		4	1/3
True		32	2/3
True		16	1/3

Figure 4.1: Illustration demonstrating statistical limitations inherent in compositional datasets. (a) Two different biological scenarios can yield the exact same proportions of taxa in samples from a population pre- and post-treatment. (b) Simulated datasets plotting the true differential obtained using absolute abundance data on the x-axis, versus the inferred differential obtained using relative abundance data on the y-axis. Each dot represents a taxon in the dataset, and the colors represent datasets with various ratios of total microbial load ('K') between before and after samples. The red line represents the optimal scenario where the samples have equal microbial load. This illustrates the prevalence of either false positives (False Positive (FP)) or false negatives (False Negative (FN)) when performing differential abundance analysis on samples with unequal total microbial load. The presence of either (FP)s or (FN)s is dictated by a nonlinear function of the true differential (see online methods). (c) An illustration of differential proportions of bacterial species before and after treatment. (d) Same data as (b) but plotting the rank of the differentials, demonstrating that ranks are equivalent regardless of differences in microbial load.

the original sample[116]. Analyzing purely compositional data (e.g. DNA sequencing data) with conventional statistical tools has led to false discovery rates approaching 100%[117, 118]. Therefore, in addition to compositional data from sequencing, quantitative information about total microbial load is necessary to determine which microbes are changing.

4.0.3 Challenges to microbial load quantification

Multiple approaches at each level of sample processing have been proposed to quantify the total microbial load from environmental samples. Adding a known amount of reference DNA as an internal standard has been used to extrapolate the amount of starting nucleic material[119, 120]. Normalization by this method is complicated due to the calibration challenges of choosing the proper amount of internal standard[120, 119]. At the extraction level, quantitative PCR (qPCR) of genomic DNA with ‘universal’ primers against the 16S rRNA gene has be deployed to estimate total microbial load[121]. However, it is impossible to prevent primer bias, resulting in uneven amplification of rRNA genes across species. Further, quantification by both spike-in and qPCR is performed on multiple subsets of the original sample. Quantifying microbial load by flow cytometry is performed on the original sample, and is agnostic to nucleotide sequences. One recent study reported that adding quantitative information obtained by flow cytometry dramatically improved interpretation of 16S rRNA gene amplicon sequencing data[116]. However, flow cytometry requires expensive equipment, experienced users, and limits throughput.

The total microbial load of an environmental sample is only one dimension of measurement among the hundreds to thousands of dimensions measured by microbial relative abundances. If the abundance of a single taxon and the relative abundance of all taxa is known, it is feasible to compute the absolute abundance of all taxa. As such, considerable information rests in relative abundances, and important insights can be gleaned without costly microbial quantification methods. Below we describe two methods to evaluate relative differential abundance independent of microbial load information.

4.0.4 Using ratios circumvents bias without microbial load quantification

Computing changes in abundance from compositional data introduces a bias due to the lack of total microbial load (Fi 1. approach#1). Simulated data in Fig 1b shows how different biases (i.e. ratios between total microbial loads) can cause either false positives or false negatives. By simply comparing the ratio of taxa between samples, the bias constant introduced by unknown microbial load cancels out. For instance, if an observed taxon X changes from 10% to 20% in relative abundance, that observation can be irreproducible across studies or samples because of fluctuations in abundance of other taxa. Changes in the abundance of taxon X relative to taxon Y should be consistent. Taking the logarithm of this ratio (log-ratio) enforces symmetry around zero, giving equal weight to relative increases and relative decreases.

4.0.5 A novel approach to rank differential abundance

Comparing ratios of taxa can circumvent the bias introduced by unknown microbial loads. However, choosing taxa for comparison from the thousands in a given sample set can be challenging. By ranking the log-ratio abundance changes of each taxon (what we refer to as "differentials"), an accurate depiction of compositional change in a dataset can be obtained and taxa can be prioritized (Fig 1c). As shown in Fig 1d, the rank of each taxon's differential is independent of the changes in the absolute microbial load, yielding an identical ranking of microbial differences between the relative and absolute abundances. However, because of the unknown bias described above, we cannot infer based on rank alone if a microbe has changed, and therefore a coefficient of zero does not imply that the microbe has not changed abundance.

Differentials can be estimated directly by using explicit count-based regression models (see online methods). For example, multiple studies have shown that multinomial linear models can infer differentials without adding pseudocounts to handle sampling zeros[96, 122, 123, 124]. The coefficients from multinomial regression analysis can be interpreted as feature importances or rankings commonly employed by machine learning methods, and can be ranked to determine which taxa are changing the most between samples. We refer to this ranking procedure as differential ranking (DR).

4.0.6 Reference frames enable reproducible compositional data comparisons

We argue that analyzing compositional data requires a choice of "reference frames" for inferring changes in abundance. By "reference frame", we draw on the concept from physics where velocity is measured "relative to" another moving object. As microbial populations change, we can constrain our inferences to how microbial populations change relative to reference frames given by other microbial populations. The choice of numerator and denominator in a log-ratio determines the reference frame for inferring changes. In DR, the differential abundances of each taxon serve as a reference to each other when they are ranked numerically. To demonstrate these principles, we confirm the robustness of these two methods of employing reference frames in real life datasets.

4.0.7 Value of reference frames in the analysis of microbes in unstimulated saliva

We demonstrate the utility of DR in a sample set with dramatic differences in total microbial load. Unstimulated saliva samples were collected from 8 individuals before and after brushing their teeth (morning and night, n=32), and processed in parallel for microbial load quantification with flow cytometry and 16S rRNA gene amplicon sequencing. Importantly, participants were asked to provide unstimulated saliva for exactly 5 minutes, so in addition to estimating microbial concentration, we could obtain a proxy for the total microbial load taking into account salivary flow rate. As expected, the total microbial load significantly decreased after brushing teeth (Fig. 2a).

For both relative and cell count-normalized data, we performed paired t-tests to evaluate the change in abundance of each taxon before and after brushing teeth (Fig. 2b). Applying t-tests to the relative data had a high false-positive rate, as seen by the disagreements between the cell count-normalized and relative t-statistics (Spearman r=0.53). Further, there was absolutely no correlation in p-value distribution between the relative and cell count-normalized data (Spearman r=0.09), highlighting the fallacy in calculating a p-value without a valid null hypothesis.

Alternatively, evaluating the ratio between *Actinomyces* and the remaining taxa produced identical t-statistics and p-values between the relative and cell count-normalized data (Spearman r=1.0). Ratio-based analyses are unaffected by microbial load (see Methods, equation 3) and result in identical interpretations as one obtains from costly and rate-limiting flow-cytometry measurements.

From the DR analysis (Fig. 2c), we can identify which taxa are changing the most (highest and lowest log-fold change). Here, we highlight *Actinomyces* and *Haemophilus* species, which have very different ranks: *Actinomyces* have low ranks and *Haemophilus* have high ranks. The difference in ranks between these taxa correctly suggests that *Haemophilus* taxa are more prevalent relative to other taxa before brushing, and *Actinomyces* taxa are more prevalent relative to other taxa after brushing. When inspecting t-test results on individual taxon in the relative data, it appears that *Actinomyces* significantly increased (t-statistic=2.89, p-value=0.013) after brushing teeth and that *Haemophilus* significantly decreased (t-statistic=-2.593, p-value=0.023). However, cell count data revealed that only *Haemophilus* significantly decreased (t-statistic=-2.477, p-value=0.029) (Fig. 2d).

The log-ratio of *Actinomyces* and *Haemophilus* between the relative and the cell count-normalized data is identical. While we cannot observe the decrease of *Haemophilus* or the consistency of *Actinomyces* abundance, with the log-ratio of their relative abundance, we can observe the interaction between these two taxa and the increase of *Actinomyces* relative to

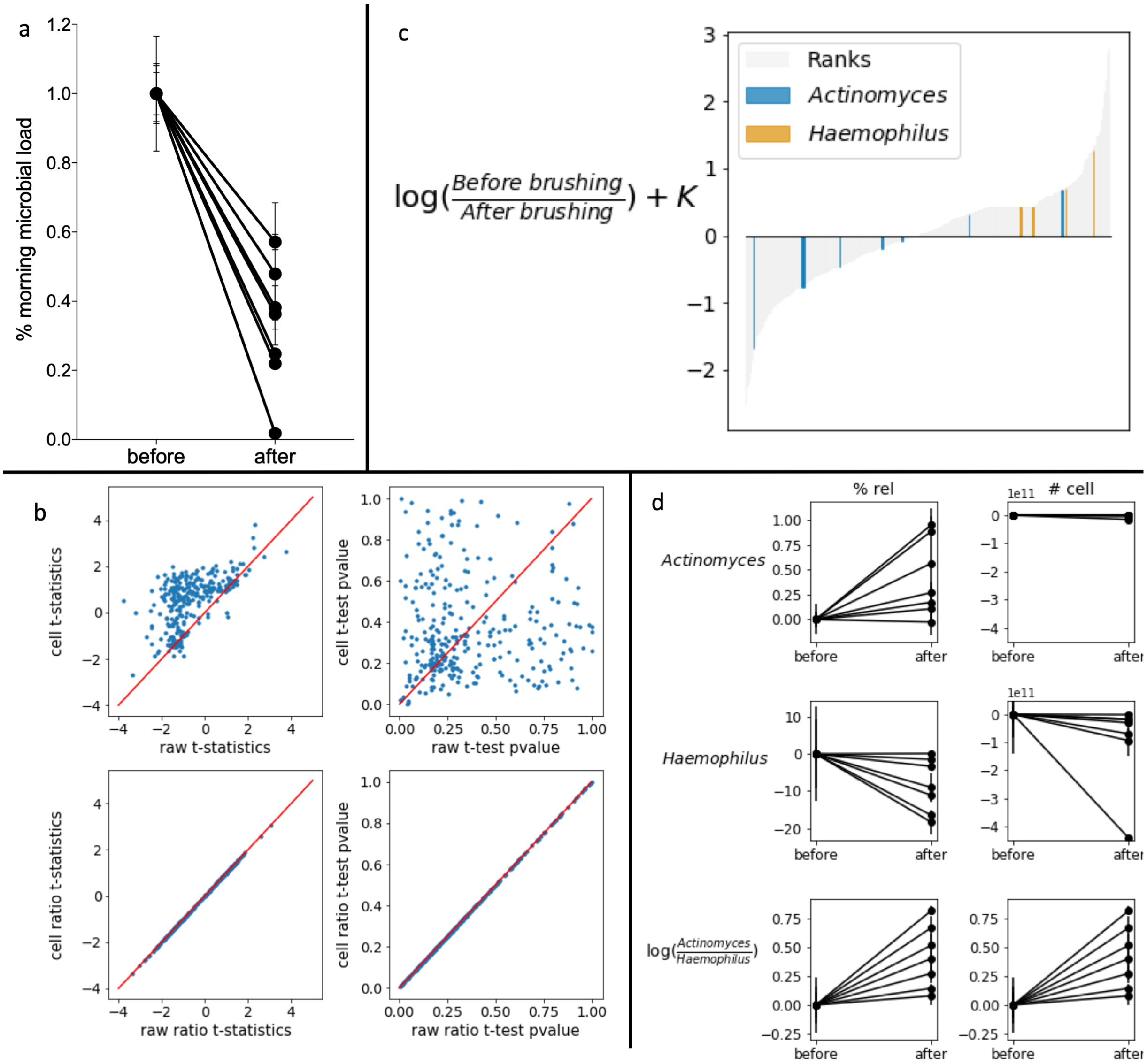


Figure 4.2: Analysis of salivary microbiota before and after brushing teeth. (a) Flow cytometry-quantified microbial load in unstimulated saliva collected for 5 minutes normalized to before brushing teeth. Each line corresponds to a different volunteer. (b) A comparison of t-statistics (left) and p-values (right) on individual taxa (top) and ratio between each taxa to *Actinomyces* (bottom) between relative abundance data (x-axis) and cell count-normalized data (y-axis). (c) Microbial ranks estimated from multinomial regression applied to oral time series dataset with *Actinomyces* and *Haemophilus* highlighted. The y-axis represents the log-fold change that is known up to some bias constant K, and the x-axis numerically orders the ranks of each taxa in the analysis (d) A comparison of relative abundance vs cell counts of *Actinomyces*, *Haemophilus* and log(*Actinomyces*:*Haemophilus*) before and after brushing teeth. Only the differences since the before time point are visualized.

Haemophilus after brushing teeth (t-statistic=2.833, p-value=0.015) These results are consistent with our knowledge about oral biogeography. *Haemophilus* is typically found on the periphery of oral biofilms and was likely removed from the biofilm during the brushing process, whereas *Actinomyces* is generally found on the surface of the tooth and acts as an anchor for biofilm attachment[125]. Importantly, this experiment demonstrates the potential fallibility of relying on relative abundance; It is illogical to conclude that *Actinomyces* increases after tooth brushing despite the increase in relative abundance. As demonstrated by flow cytometry, total microbial load decreases, and while both *Haemophilus* and *Actinomyces* decrease, *Haemophilus* decreases more.

Next, we demonstrate the utility of log-ratios and DR to identify consistent microbial changes across previously published datasets where quantification of microbial load is unavailable.

4.0.8 Discovery of interkingdom relationships in atopic dermatitis using reference frames

The tooth brushing example provides ground truth for using log-ratios and DR, but many clinically relevant microbiome questions involve less obvious differences. Using data from patients with atopic dermatitis (AD), an important skin disease, we demonstrate how viewing relative abundances alone can produce false negatives.

AD has a complex etiology. Many microbiome studies performed using next-generation sequencing have focused on bacterial changes associated with AD, especially the pathogen *Staphylococcus aureus*. The yeast genus *Malassezia* has also been implicated in AD, although conflicting results have been published as to which *Malassezia* species are involved and whether they are more or less prevalent in AD[126]. A recent shotgun metagenomic study examined the skin microbiome over time during an AD flare and recovery. The authors observed a decrease in *Staphylococcus aureus* relative abundance in the healthy, recovered skin (non-lesioned) compared to AD flare (lesion), but no significant changes in the relative abundance of *Malassezia* species over time in these AD patients[127].

Applying compositionally coherent methods to this dataset revealed new insights. Observing the DR results (Fig. 3a), it is apparent that, compared to lesioned skin, *S. aureus* is one of the taxa to decrease the most relative to all other microbes in the non-lesioned sites, followed by *S. epidermidis*, and *M. globosa*. Consistent with the analysis of relative abundance in Fig. 3b, the ratio of *S. aureus : P. acnes* was significantly increased in flare (t-statistic=2.973, p-value=7.811×10^{-3}) and correlated with SCORAD score, a clinical assessment of AD severity (Pearson=0.747, p-value=3.516×10^{-6}). Contrary to previous findings, both *S. epidermidis : P. acnes* and *M. globosa : P. acnes* were also significantly increased in lesioned skin (t-statistic=3.197, p-value=4.748×10^{-3}, and t-statistic=4.030, p-value=7.16×10^{-4}, respectively) and correlated with SCORAD score (Pearson r=0.464, p-value=6.975×10^{-4}, and Pearson r=0.668, p-value=1.125×10^{-7}, respectively) (Fig. 3c).

To validate this observation, we analyzed shotgun data from an independent AD dataset[128]. In this dataset, the relative abundance of *M. globosa* significantly increased between lesioned and non-lesioned skin (Fig. 3e, t-statistic=4.135, p-value=0.0001). But the ratio of *M. globosa : P. acnes* increased even more dramatically in lesioned skin (Fig. 3d, t-statistic=5.79, p-value=8.6×10^{-7}) (Fig. 3d). These results are congruent with a previous report that *M. globosa* was cultivated more successfully from lesioned versus non-lesioned sites in AD[129]. Thus, DR analysis can identify novel, clinically significant microbial changes which can be validated across cohorts by choosing insightful reference frames.

Discussion

Adding information about absolute microbial load between samples can highlight issues inherent in compositional data analysis. However, there are multiple practical and technical challenges in quantifying microbial load. For example, skin swabs are often difficult to use in flow

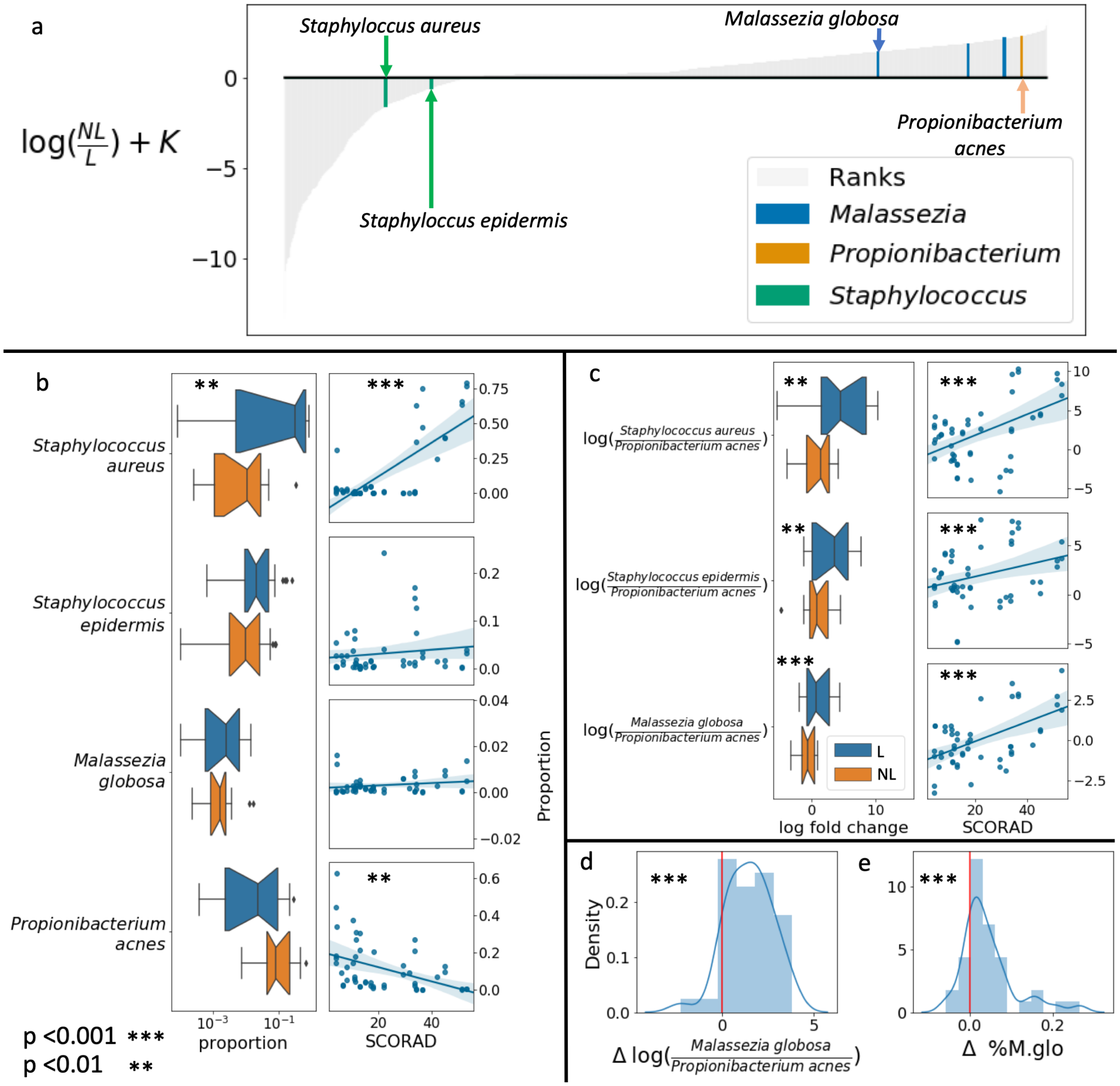


Figure 4.3: Comparison of lesioned (L) versus non-lesioned (NL) skin in two atopic dermatitis studies; Byrd et al.[127], (a-c) and Leung et al.[128], (d-e). (a) Microbial ranks estimated from multinomial regression applied to shotgun metagenomics from Byrd et al[127] with key genera highlighted. The y-axis represents the log-fold change that is known up to some bias constant K. (b) Proportions of *S. aureus*, *S. epidermidis*, *M. globosa*, and *P. acnes* in lesioned (blue) and non-lesioned (orange) skin (left) and correlation of relative abundance with Scoring Atopic Dermatitis (SCORAD) score (right) (c) Log-ratios of *S. aureus* : *P. acnes* , *S. epidermidis* : *P. acnes* , and *M. globosa* : *P. acnes* (left) and correlation of ratio with SCORAD score. (d) Change in log ratio of *M. globosa* : *P. acnes*. (e) Change in relative abundance of M. globosa between lesioned and non-lesioned skin from (Leung et al.[128]).

cytometry due to very low biomass and difficulty in transferring intact cells from swabs into liquid solution. Furthermore, skin samples are notoriously sensitive to 16S rRNA gene primer choice making qPCR quantification challenging[130]. Similarly, for historically collected samples that exist only as DNA in a freezer or as sequences in a database, flow cytometry approaches to determine absolute microbial load are not feasible.

However, absolute abundances of a community are only one dimension of measurement, and robust, alternative techniques eliminate the need to estimate total biomass. We have demonstrated the validity of using microbial ratios and differential rankings to determine significant changes in microbiome studies by comparing compositional inferences with absolute abundance inferences.

By using flow cytometry to quantify total microbial load, we validated these analytical tools in 16S rRNA gene amplicon sequencing data from unstimulated saliva. We found evidence of false positives when looking exclusively at changes in relative abundance before and after brushing teeth. By evaluating the ratio of *Actinomyces : Haemophilus*, we reached an identical conclusion to our cell-count normalized data without the need for microbial load quantification. The consistency of our results rests in the use of ratios defining reference frames for inferring compositional changes.

Furthermore, we highlighted an example of a false negative in previously generated shotgun metagenomic data from the skin of individuals with AD. We were able to reproduce the findings that *S. aureus*, and to a lesser extent *S. epidermidis*, are differentially abundant in AD lesions. Additionally, using log-ratios and differential ranking, we were also able to show a more subtle but statistically significant change in *M. globosa* abundance in AD lesions. This same result was obtained in two independent metagenomic studies of AD patients and agrees with previous cultivation-based work quantifying increased colony forming units of *M. globosa* in AD lesions.

Consistency between inferences made based on relative and cell count-normalized data is crucial, because in many circumstances it is not possible or practical to estimate total microbial load. The seeming contradiction between microbial load-corrected abundances and relative abundances does not invalidate data from the existing 100,000+ experiments utilizing 16S rRNA gene amplicon or metagenomic sequencing[131, 132]. Importantly, these techniques are not limited to next generation microbiome sequencing, but can be applied to any experiments involving compositional data (e.g. metabolomics, proteomics, etc.).

Multiple tools have already been developed that can facilitate analysis using log-ratios. For instance, tools such as PhILR, Phylofactoriation and Gneiss provide different means to compute reference frames for log-ratio analysis. However, these methods rely heavily on pseudocounts, because the logarithm of zero is undefined. This can add a substantial amount of bias, especially in sparse datasets. The Differential ranking (DR) procedure circumvents this problem through the use of the multinomial regression.

While various methods of multinomial-based models have been developed[96, 122, 123, 124], the interpretation of the resulting model coefficients is usually incorrect. A zero valued coefficient does not imply that the corresponding species abundance hasn't changed, due to the total biomass bias as discussed in Fig 1. DR provides a novel means to correctly interpret the coefficients of these models. By ranking the coefficients we can determine which taxa have changed the most relative to each other. This subtle distinction acknowledges the limits of analysis of compositional data, and as demonstrated above can have dramatic impacts on data interpretation.

While there are widespread misconceptions concerning how to interpret microbial abundances, there is still much hope for resolving these outstanding controversies. Ongoing efforts at the NIH and EMBL-EBI have already stored petabytes of multi-omics datasets ready to be re-analyzed, and databases, such as Qiita and gcMeta, contain curated data and metadata from hundreds of thousands of samples[132, 131]. There is much promise for resolving outstanding controversies by re-analyzing these datasets using reference frames to make stable inferences of compositional change.

4.1 Methods

If we wanted to compute the change between two samples containing compositions (e.g. relative abundance of microbes) $\boldsymbol{A} = (a_1, ..., a_D)$ and $\boldsymbol{B} = (b_1, ..., b_D)$, it would look like

$$\frac{\boldsymbol{A}}{\boldsymbol{B}} = (\frac{a_1}{b_1}, \dots \frac{a_D}{b_D}) \tag{4.1}$$

If we are only able to measure relative abundances, as is the case with next generation amplicon sequencing, we can only estimate the proportion p_{a_i} for species i in the sample A i.e. $p_{a_i} = \frac{a_i}{N_a}$. Estimating the true abundance can be done via $a_1 = N_a p_{a_1}$, where N_a is the total abundance of sample A. To estimate the true change, the following can be done

$$\frac{\boldsymbol{A}}{\boldsymbol{B}} = \frac{N_a \times p_{a_1}}{N_b \times p_{b_1}}, \dots \frac{N_A \times p_{a_D}}{N_B \times p_{b_D}} = \frac{\boldsymbol{p_A}}{\boldsymbol{p_B}} \times \frac{N_A}{N_B} \tag{4.2}$$

To determine if species i abundance has changed between samples A and B, we test to see if $\frac{a_i}{b_i} = 1$. However as shown above, we cannot perform this test, since the results of this test would be confounded by the total biomass bias $\frac{N_A}{N_B}$.

In many cases, the total biomass cannot be estimated, so any techniques to identify important species will need to alleviate this bias. One alternative is to use ratios. If we choose species D to be the reference species, it is clear that the total biomass cancels as follows

$$\frac{a_1/a_D}{b_1/b_D} = \frac{p_{a_1}/p_{a_D}}{p_{b_1}/p_{b_D}} \tag{4.3}$$

Another alternative is to use ranks. Since the bias is applied uniformly across the differential, it will not affect the ordering of the species. Hence, ranks are agnostic to the total biomass bias.

$$rank(\frac{\boldsymbol{A}}{\boldsymbol{B}}) = rank(\frac{\boldsymbol{p_A}}{\boldsymbol{p_B}} \times \frac{N_A}{N_B}) = rank(\frac{\boldsymbol{p_A}}{\boldsymbol{p_B}}) \tag{4.4}$$

This differential is also commonly referred to as a perturbation in the context of the compositional literature[100]. It is important to note that this does not justify using Spearman correlation or other non-parameteric tests such as Kruskal-Wallis applied to relative abundance data since these tests do not satisfy scale invariance[133, 86].

Both of the log-ratios and the differential ranking techniques satisfy scale invariance, meaning that both of these techniques are agnostic to the total biomass. This concept is critical when analyzing relative abundance data, since this is one step closer to maintaining consistent conclusions between the original environment and the observed sequences.

Estimating log-fold differential expression from relative abundances can result in either false positives (FP) or false negatives (FN) depending on the distribution of true differential expression. Whether FNs or FPs are observed depends on a nonlinear relationship involving the true (unobserved) differential expression. For a p-vector x let us define $LSE(x) = \log(e^{x_1} + \ldots + e^{x_p})$. If $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_D)$ denotes the true differential expression of the D species between two conditions, let $\boldsymbol{\alpha} = LSE(\boldsymbol{\delta})$. Further, let $\hat{\boldsymbol{\delta}} = (\hat{\delta_1}, \ldots, \hat{\delta_D})$ represent the inferred differential expression from proportional data and $\hat{\boldsymbol{\alpha}} = LSE(\hat{\boldsymbol{\delta}})$. If $LSE(\delta) > LSE(\hat{\delta})$ then FPs will be observed. In contrast if $LSE(\delta) < LSE(\hat{\delta})$ then FNs will be observed.

4.1.1 Multinomial regression

To perform the differential ranking (DR) analysis, we used multinomial regression. Multinomial regression and related count regression models are commonly used in the context of microbiome analysis. Here, we use the multinomial regression model since these models can reliably estimate the means and can be easily reinterpreted in the context of compositional data analysis.

Counts from the multinomial regression can be formulated in the following generative model

$$\begin{aligned}
&\beta_{jk} \sim \mathcal{N}(0, \mu_\beta) \\
&\boldsymbol{\eta_i} = alr^{-1}(\boldsymbol{X_i}\boldsymbol{\beta}) \\
&\boldsymbol{Y_i} \sim Multinomial(\boldsymbol{\eta_i}) \ ,
\end{aligned}$$

where $\boldsymbol{\beta}$ represents the coefficients of the model across all measured covariates k. $\boldsymbol{X_i}$ represents the metadata covariates for sample i. $\boldsymbol{Y_i}$ represents the measure microbial counts for sample i. A normal prior centered around zero was placed on the covariates $\boldsymbol{\beta}$ to serve as regularization to combat issues associated with high dimensionality.

The inverse alr function is a function commonly used in the context of compositional data, give as follows

$$alr^{-1}(x) = C[\exp(0, x_1, \ldots, x_{D-1})] \qquad C[x] = \left[\frac{x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i}\right] .$$

This is also referred to as a degenerate softmax function, which is commonly used in the context of neural networks. This function is also isomorphic between $\mathbb{R}^{D-1}$ and $\mathcal{S}^D$ (the space of proportions), so this will ward against identifiability issues when estimating these model parameters. The models were estimated using a maximum a posteriori priori (MAP) estimation using stochastic gradient descent.

If we examine the model parameters $\boldsymbol{\beta_k} \in \mathbb{R}^{D-1}$, we reinterpret the quantities given by $alr^{-1}(\boldsymbol{\beta}_k)$ as differentials as discussed in Fig 1.

It is also worthwhile to note the connection between $\boldsymbol{\beta_k}$ and balances. Since $\boldsymbol{\beta_k}$ is expressed in alr coordinates, there is also a direct connection to ilr coordinates, meaning that $\boldsymbol{\beta_k}$

can also be transformed into balances. More explicitly, the ilr coordinates of these coefficients can be computed as follows

$$\boldsymbol{\beta}_{\boldsymbol{k}}^{(ilr)} = ilr_{\boldsymbol{\Psi}}(alr^{-1}(\boldsymbol{\beta}_{\boldsymbol{k}})) \ .$$

The resulting coefficients are represented as coordinates given by the orthonormal basis $\boldsymbol{\Psi}$. An example of such a basis can be dervied from bifurcating trees discussed in Morton et al[118], Silverman et al[15] and Washburne et al[13]. This can allow for relative changes in abundances as given by $alr^{-1}(\boldsymbol{\beta}_{\boldsymbol{k}})$ to inform which balances are changing in ancestral states given by the tree . The multinomial regression serves as an alternative means to compute regression coefficients discussed in PhILR, Phylofactor and Gneiss, while avoiding issues with imputation and zeros.

The multinomial regression was implemented using Tensorflow[134] and can be found in https://github.com/mortonjt/songbird.

4.1.2 Saliva sample collection

Eight volunteers provided unstimulated saliva so that salivary flow rate could be measured according to a standardized protocol [135]. Briefly, individuals were asked to allow saliva to flow for exactly five minutes through a disposable funnel (Simport, SIM F490-2)into a sterile, 15 mL conical tube preloaded with 2 mL sterile glycerol for bacterial preservation. Participants were asked to provide samples before brushing and after brushing teeth in the morning and in the evening. Samples inverted several times to mix with the glycerol and stored at -20°C immediately after collection. This study was approved by an Institutional Review Board (IRB# 150275) and written informed consent was acquired before sample collection

4.1.3 Flow cytometry

Unstimulated saliva samples were thawed on ice, and aliquots were diluted tenfold with sterile, 1x PBS. To remove human cells and salivary debris, samples were filtered using a sterile 5 μm syringe filter (Sartorius Stedim Biotech GmbH). 5 μl 20x SYBR green (SYBR^{TM} Green I Nucleic Acid Gel Stain, Invitrogen) was added to 1 mL of the microbial suspension (0.1x final concentration) and incubated in the dark for 15 minutes at 37°C. Finally, 50 μl AccuCount Fluorescent Particles (Spherotech, ACFP-70-10) were added for assessment of microbial load. Samples were processed on a SH800 Cell Sorter (Sony Biotechnology) using a 100 μm chip with the threshold set on FL1 at 0.06%, and gain settings as follows; FSC=4, BSC=25%, FL1=43%, FL4=50%. The gating strategy was adapted from Vandeputte et al.,[116]. Briefly, fluorescent microbial cells were gated from background on a FL1-Fl4 density plot, and remaining background was removed by eliminating large events detected on a FSC-BSC density plot. Negative controls (sterile PBS stained identically to samples) were run between each sample set to exclude cross-contamination. Settings were identical among all samples.

4.1.4 Amplicon sequencing

DNA extraction and 16S rRNA amplicon sequencing were done using Earth Microbiome Project (EMP) standard protocols (http://www.earthmicrobiome.org/protocols-and-standards/16s). 500 μl of unstimulated saliva was used for gDNA extraction with MagAttract PowerSoil DNA Kit (QIAGEN) as previously described [136]. Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair 515f to 806r with Golay error-correcting barcodes on the reverse primer. 240 ng of each amplicon was pooled and purified with the MO BIO UltraClean PCR cleanup kit and sequenced on the Illumina MiSeq sequencing platform.

The sequences and biom tables [137] can be found on Qiita (http://qiita.microbio.me) under study ID 11896. Demultiplexed fastq files were processed using QIIME2 (https://qiime2.org)[138].

Deblur was used to denoise the sequences [139]. 16S taxonomy was assigned using RDP classifier [100, 140]. Songbird was used to perform multinomial regression - repository can be found here: https://github.com/mortonjt/songbird Paired t-tests were performed to evaluate the differences before and after brushing teeth. All log-ratios that were evaluated to either positive or negative infinity are dropped prior to statistical analysis. These numerical issues occur due to particular microbes not observed, and we treat them as missing data respectfully.

4.1.5 Shotgun metagenome studies

We used supplementary data from Byrd et al [127] and Donald Leung[128]. The provided relative abundances were compared to the log ratio of the raw count data. Paired t-tests were performed to evaluate the differences between lesion and non-lesion skin samples. All log-ratios that were evaluated to either positive or negative infinity are dropped prior to statistical analysis. These numerical issues occur due to particular microbes not observed, and we treat them as missing data respectfully.

4.2 Acknowledgements

We would like to Doris Vandeputte for her insights behind the total microbial load quantification. J.T.M. was funded by NSF GRFP DGE-1144086.

Chapter 4 has been submitted for publication of the material as it may appear in Nature Biotechnology, 2019 “Establishing microbial measurement standards with reference frames” James T. Morton, Clarisse Marotz, Justin Silverman, Alex Washburne, Livia S. Zaramela, Anna Edlund, Karsten Zengler, Rob Knight. The dissertation author was the primary investigator and first author of this paper.

Appendix A

Supplemental material for Chapter 1

A.0.1 Box 1: Glossary of terms

Ancestral state: The traits of the ancestral species, typically an estimate of the phenotype and the genotype of the ancestral organism.

Ancestral state reconstruction: Imputing the ancestral states at various points in the phylogeny.

Bayesian inference: Given a prior set of beliefs about the ancestral states, and observed phenotypes/genotypes of existing species, Bayesian methods will attempt to obtain a more informed estimate of the ancestral states, along with confidences of the prediction.

Blomberg's κ: A more common, modern measure of phylogenetic signal compared to Pagel's λ (see below), ranging from 0 to infinity, which indicates the extent of acceleration or deceleration of evolution over time.

Bootstrapping (phylogenetics): Repeated, stochastic reconstruction of a phylogeny proposed by Felsenstein [33], often used to assess the percentage of reconstructions in which each clade is found.

Brownian motion: A continuous random walk where jumps are normally distributed random variables. Commonly used in PCMs as a null model of continuous trait evolution from the ancestral node towards the tips of the tree, where the random walk branches with those in the phylogeny. Under a Brownian motion model of evolution, the covariances between species' observed traits is proportional to the branch length of their shared ancestry.

Classification: Regression or other efforts to predict categorical dependent variables.

Clustering: Creation of classifiers (categorical variables) identifying groups of variables, such as groups of species with high within-group similarity and low between-group similarity.

DNA amplicons: DNA products of artificial amplification events, such as the resultant products of polymerase chain reaction amplification of 16S rRNA genes that are later sequenced and counted to assemble microbiome datasets. Amplicons may sometimes be used to construct accurate phylogenies for microorganisms.

Edge: A structure in a phylogeny representing a hypothesized distinct, unbroken lineage during a point in time.

Edge lengths: Edge lengths may represent either the time over which an historical lineage persisted or the number of mutation events separating its ancestral from daughter nodes.

EdgePCA: A method which performs principal component analysis on a set of variables, v i , corresponding to differences of abundances along each edge, i.

Epistasis: When two or more genetic loci interact to determine a phenotypic trait.

Evenness: A general term for a variety of metrics indicating how close a community is to having equal abundances across all species.

ILR: Isometric log-ratio. A standardized difference of arithmetic means of log-transformed data. Often used in microbiological datasets that are more appropriately analysed on a log scale, but an analogous difference for non-log-transformed data is the t-statistic for a two-sample t-test.

Maximum likelihood: Maximum likelihood methods treat ancestral states as unknown parameters. Given a probabilistic model of evolution, maximum likelihood methods will attempt to optimize these parameters to try to find the most likely ancestral states that yield the traits that we observe in known species in the present.

Maximum parsimony: Maximum parsimony attempts to reconstruct ancestral states by minimizing the number of trait changes between the ancestor and the present descendants.

Monophyletic: A set of species is called 'monophyletic' relative to a larger set of species if their most recent common ancestor has no other descendants besides those within the set of species.

Node: A structure in a phylogeny representing a hypothesized timing of speciation, when one lineage splits into two or more distinct lineages.

Pagel's λ: A measure of phylogenetic signal, ranging between 0 and 1, which indicates the relative extent to which a traits' correlations among close relatives match a Brownian motion model of trait evolution.

PhILR: Phylogenetic isometric log-ratio. A transform of the data requiring a fully resolved phylogeny (that is, no polytomies). Instead of representing data with one variable for each species, the PhILR transform represents the data with one variable for each node in the phylogeny. Variables are constructed using the ILR transform to contrast sister clades descending from each node.

Phylofactorization: A method of choosing variables by a generalized graph-partitioning algorithm. Variables are constructed by first considering contrasts along edges, such as differences or ILRs contrasting birds and non-birds, and then finding out which variable maximizes a researcher's objective function. The phylogeny is partitioned along that edge and the process is repeated, limiting contrasts only to sub-phylogenies in which the edges are found (for example, after partitioning birds/non-birds, the edge separating doves/non-doves instead separates doves/non-dove-birds).

Phylogenetic comparative methods: Statistical methods which correct for correlations of trait observations among close relatives, to be used whenever traits, broadly defined as heritable features at the tips of a phylogeny, are a dependent variable or when testing differences between two traits. PCMs often use models of evolution to calculate correlations between observations of close relatives expected under random evolution.

Phylogenetic distance: The sum of edge lengths along the path connecting two species in a phylogeny.

Phylogenetic inference: The estimation of the evolutionary history of a set of genes.

Phylogenetic variables: Variables constructed with the aid of a phylogeny (including the star phylogeny in which all species originate from the same polytomy). In contrast to phylogenetic distances, variables indicate directions and curves along which variation has biological meaning.

Phylogeny: A diagrammatic hypothesis of the evolutionary history of a set of genes. The phylogeny can be rooted, implying knowledge of the most basal common ancestor of the set of genes, or unrooted.

Polytomy: A node with more than two daughter lineages. Often, polytomies represent uncertainty about the precise timing of historical speciation events.

Regression: A predictive mathematical model that will attempt to estimate relationships between variables.

Shannon diversity: A particular measure of evenness, H, defined by a set of relative abundances, p_i, summing to 1: $H = -\sum_i (p_i \log(p_i))$.

A.0.2 Box 2: Challenges to phylogenetic analysis of microbiome data

Horizontal gene transfer: Horizontal Gene Transfer (HGT) disrupts the correlation between evolutionary histories of genes and raises important questions about which gene trees to use for phylogenetic analysis. While the 16S gene tree correlates to the bulk of genomic content in microorganisms, important horizontally transmitted genes such as β-lactamases have phylogenies that are different from the 16S. Analysing a β-lactamase gene tree will allow analysis of β-lactamase traits—such trees may be appropriate for studying the composition of antibiotic resistant genes in the environment. We discuss this further in the 'Challenges of phylogenetic analysis' section of the text.

Phylogenetic inference: The 16S Ribosomal RNA (rRNA) gene tree is most commonly used, but other genes, such as β-lactamase genes, can be used to make phylogenies. Regardless of the gene, phylogenetic inference is an estimate of evolutionary history and the estimate is most accurate with large and even taxon sampling [59]. Uneven taxon sampling can produce erroneous phylogenies resulting in similar traits being misinterpreted as homologies. Phylogenetic reconstruction using skin and skeletal structure of many species of lizards, one species of bird and one species of bat may incorrectly estimate that birds and bats are sister taxa, whereas a more complete sampling of taxa to include mammals may correctly group bats with mammals. Building trees de novo within each study site, with the limited taxon sampling of each locale, risks producing many erroneous trees that are difficult to compare across studies. Global consensus trees built from commonly used genes, even taxon sampling and standardized methods for adding new sequences to the existing tree can ensure that researchers make comparable inferences on the same, reasonably accurate scaffold of microorganisms' evolutionary history.

Ancestral state reconstruction: As with phylogenetic inference, sparse taxon sampling can increase the error rate of ancestral state reconstruction. Methods such as PICRUSt, which draw on genomes and traits of organisms from relatively well-sampled environments such as the human microbiome, will probably have high error rates for organisms in less well-sampled environments.

Vast number of species: A recent study estimates there to be upwards of a trillion microbial species [141]. While large datasets for macroscopic organisms exist, the regularity of species-rich datasets in microbial ecology and the ease of collecting many samples warrant special consideration of the computational costs, visualization and interpretation of methods often developed for smaller datasets. Parallelization, emphasis on lineages of common knowledge or importance, common knowledge of phylogenetics among microbiologists, and simplified representations of phylogenies by collapsing clades may allow researchers to perform thorough phylogenetic analysis of microbial big data.

Evolutionary model for microorganisms: There is no microbial fossil record. In macroecology, fossil records are used to calibrate evolutionary rates necessary for phylogenetic inference and ancestral state reconstruction [142, 143]. While we know that different genes within different species have different mutation rates [144, 145], the calibration and validation of evolutionary models for microorganisms is still an open area of research. The correct evolutionary model can produce accurate effect sizes and measurements of uncertainty (significance, confidence intervals, and so on), ensuring accuracy and reproducibility of inferences in phylogenetically structured data analysis.

Appendix B

Supplemental material for Chapter 2

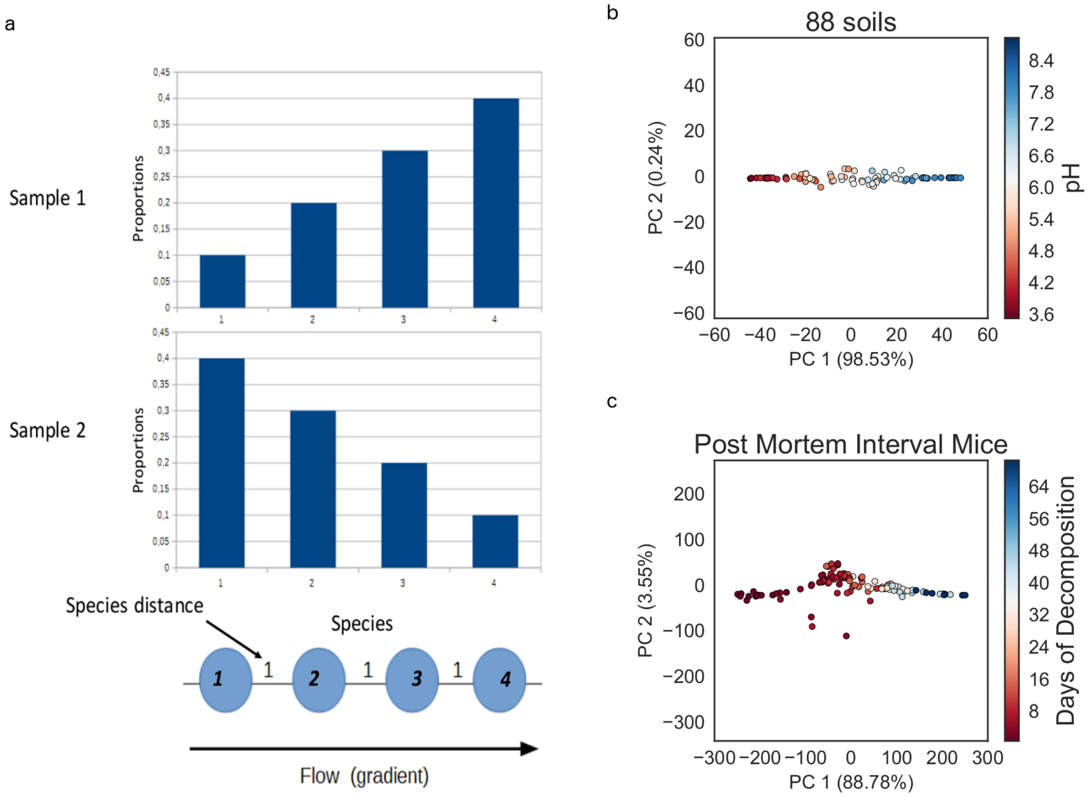


Figure B.1: An illustration of a distance metric that is engineered not to saturate. (a) An illustration of the Earth Mover Band Distance. (b) Demonstration of the EMBAD metric on the 88 soils dataset. (c) Demonstration of the EMBAD metric on the Post Mortem Interval Mice dataset

The idea behind the EMBAD distance metric is as follows. Suppose that we have obtained a scrambled matrix of OTU abundances but there exists an underlying band pattern when the table is sorted. Specifically, this table can be reordered and sorted by a value, such as sample pH. In addition, the species can also be sorted by the sample value ranges that they are observed in. In the pH example from the 88 soils study, the microbes (OTUs) were ordered based on the pH ranges they were found in. This ordering of species can be used to construct a pipe where the lowest ordered species is placed on one end of the pipe, and the highest ordered species is placed on the opposite end of the pipe. Once this pipe is constructed the species abundances from different samples can be imposed on the pipe, and the flow between samples can be computed using the Earth Mover's distance. Consider the example in Figure S1a. There are two samples where sample 1 is dominated by species 4 and sample 2 is dominated by species 1. If the ordering of species is already known, we can compute the proportions of individuals in sample 1 that need to be shuttled along the pipe in order to transform sample 1 into sample 2. In this scenario, about 0.3 of species 1, 0.1 of Species 3 need to be distributed across species 1 and 2. If we can determine the ordering of species, we can effectively compute how dissimilar Sample 1 and Sample 2 are from each other.

Furthermore, by imposing an ordering across all species, this distance metric is designed to be non-saturating. If there are samples that do that overlap, the dissimilarity between these samples is weighed by how far away they are in the pipe. Two samples that appear close together in the pipe will have a smaller distance since there the proportions will travel a smaller distance along the pipe.

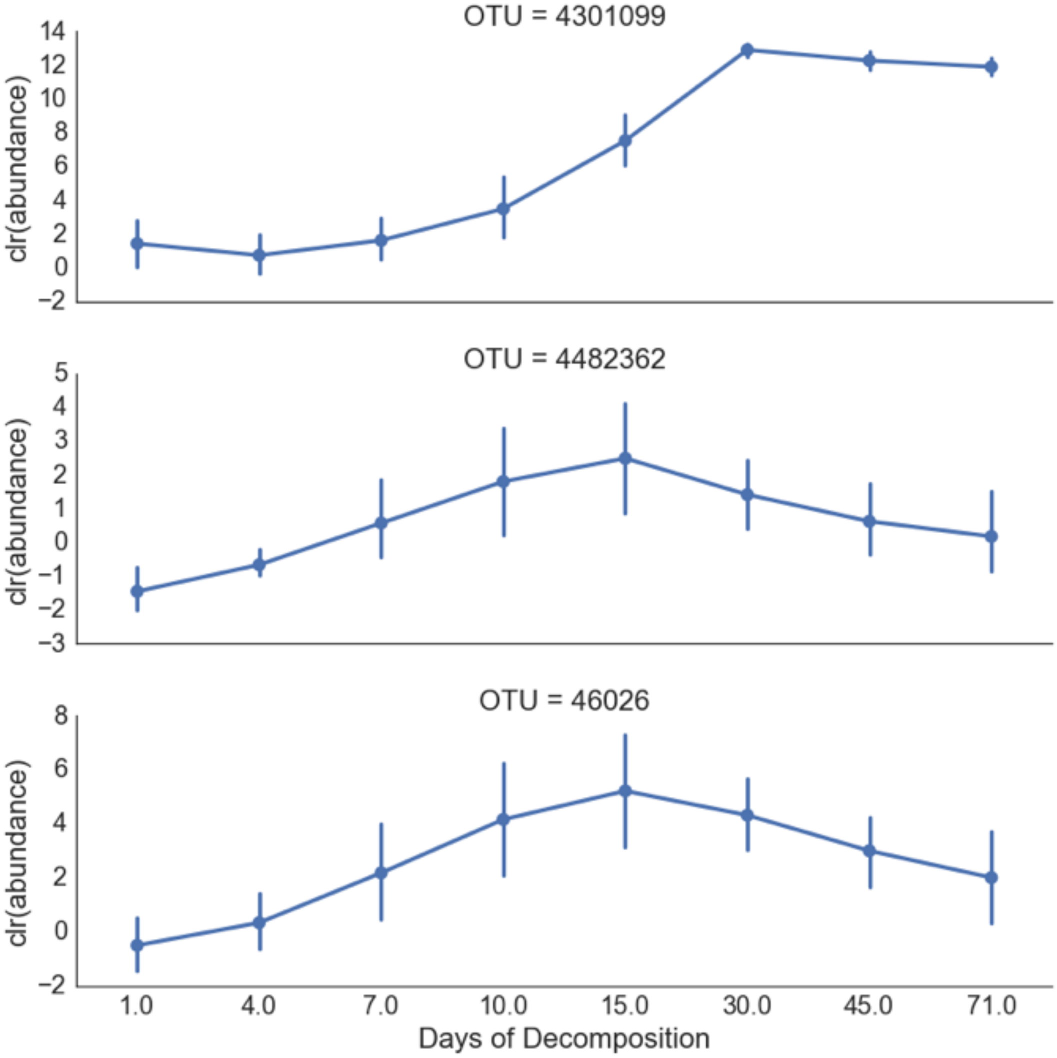


Figure B.2: Abundances of taxa across time in the post-mortem experiment. The center log ratio (Equation 2) transformed abundances of Rhizobiaceae (OTU 4301099) and Chromatiaceae (OTU 46026, 4482362) versus time. These demonstrate distinct relationships of different taxa as a function of decomposition.

B.1 Distance saturation proof

Theorem:

Let $(S_i)_{1 \leq i \leq N}$ be a set of N different samples along a linear trajectory

Let $d = \min_{i,j,i\neq j} \|S_i - S_j\|_2$ be the minimum Euclidean distance between every pair of samples in $(S_i)_{1\leq i\leq N}$ and $C = \max_{i,j} \|S_i - S_j\|_2$ be the maximum Euclidean distance between every pair of samples in $(S_i)_{1\leq i\leq N}$.

We have

$$N \leq \left\lfloor \frac{C}{d} \right\rfloor + 1$$

Proof:

Since all our samples samples are on a linear trajectory, without loss of information we can project the samples on this line. Now consider our samples as points in $\mathbb{R}$ the real number line.

Without loss of generality, we can suppose that our samples are ordered long the real line:

$$S_1 < S_2 < \cdots < S_N$$

We have $d > 0$ because all samples are different.

Thanks to the structure of the real line, we have

$$C = \max_{i,j} \|S_i - S_j\|_2 = S_n - S_1$$

and all samples are part of the interval of length $C : I = [S_1, S_n]$

We can include the interval I into the reunion of $\lfloor \frac{C}{d} \rfloor + 1$ intervals of length d.

Since $d = \min_{i,j,i\neq j} \|S_i - S_j\|_2$ two samples cannot be in the same sub-interval I_k.

Therefore by the pigeon-hole principle

$$N \leq \left\lfloor \frac{C}{d} \right\rfloor + 1$$

Corollary:

When there is a distance-saturation, we have $N \leq \left\lfloor \frac{C}{d} \right\rfloor + 1$, therefore N samples cannot be on a linear trajectory.

B.2 NP hardness of finding an optimal linear embedding

Suppose that we have a scrambled table and has an underlying band pattern.

In order to define an EMBAD distance to infer an underlying band pattern in the absence of a known gradient, we need to (1) be able to determine the define the trajectory of points that define the horseshoe and (2) determine the optimal ordering of OTUs based on (1). In order to resolve (1), we need to obtain the shortest path through the horseshoe. Specifically we would need to find the optimal ordering of points $x_1, \ldots, x_N \in \mathbb{R}_+^D$ such that the following objective function is minimized.

$$\min \sum_{i=1}^{n} \|x_i - x_{i-1}\|_2$$

If there exists an algorithm to find the shortest path through the horseshoe, then this solution can be used to solve the Traveling Salesman problem. Therefore, defining an EMBAD distance metric in the absence of a known gradient is NP-hard.

Appendix C

Supplemental material for Chapter 3

C.1 Scale invariance of balances

$$\log \frac{\prod_{x_j \in i_L} (x_j)^{1/|i_L|}}{\prod_{x_k \in i_R} (x_j)^{1/|i_R|}} = \log \frac{\prod_{p_j \in i_L} (np_j)^{1/|i_L|}}{\prod_{p_k \in i_R} (np_j)^{1/|i_R|}} = \log \frac{\prod_{p_j \in i_L} (p_j)^{1/|i_L|}}{\prod_{p_k \in i_R} (p_j)^{1/|i_R|}} \tag{C.1}$$

n is the true sequencing count, x_j is the true abundances of species j and p_j is the proportion of species j. i_L is the set of all species proportions contained in the left sub-tree at internal node i, i_R is the set of all species proportions contained in the right sub-tree at the internal node i, and, $g(x)$ is the geometric mean of all of the proportions contained in x, $|i_R|$ is the number of species contained i_R and $|i_L|$ is the number of species contained i_L. As shown above, the sequencing depth constant gets effectively canceled out. Thus, log ratios are a natural normalization for sequencing depth, especially if there are no zero abundances present and the samples have sufficient coverage.

C.2 Benchmark of compositional coherence

Supplemental Figure 1

The simulation consisted of a uniform population of 1000 individuals. A blooming was simulated across 9 time points, where a single organism eventually grew 100,000x fold. At each time point, 30 compositions were simulated using multinomial sampling with replacement. At each time point, a statistical test was performed comparing the sample at that time point to the original time point. Since we know beforehand that only 1 species is changing, any other tests that don't involve the first set of proportions that is determined to be significant with p-value $<$ 0.05 is a false positive Figure S1a-b). In fact, if the bloom of a given species is high enough, all of the other individuals can be detected to change. In Figure S1e, if 1 species has changed by 100,000x, then all of the pvalues will be less than 10-10, giving a false positive rate close to 100%. The same procedure is performed using balances, any balance that does not contain x1 that is determined to be significant from a t-test is considered a false positive (Figure S1c). Note that this is highly dependent on the choice of the tree. In this case, we used a tree where the blooming species x1 was to the far right of the tree (Figure S1f). In this way, only the balance between x1, and x2 through x1000 should be changing. But if we were to flip the tree, and place x1 to the far right of the tree, every balance in the tree will contain x1 (Figure S1g). So as x1 blooms, the number of significant balances will increase (Figure S1d).

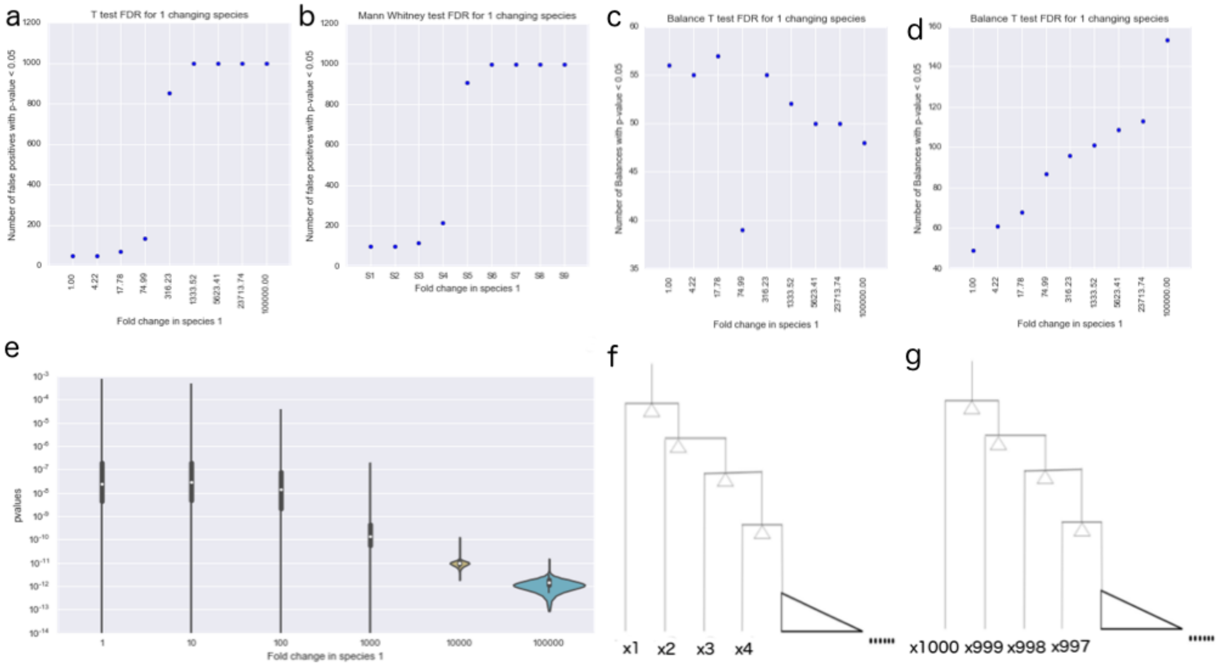


Figure C.1: A benchmark of statistical tests on compositional data. Simulations illustration the occurence of false positives in traditional statistical tests

While it may be deemed biologically irrelevant, blooms do happen frequently in microbial studies, with individual species sometimes blooming 5 orders of magnitude within a short period of time. And the 10,000x fold growth of a single species will have the exact same effect as the 1,000x fold growth change of 10 species. This suggests that there could be many subtle scenarios where we could be misinterpreting biologically relevant signals by testing individual proportions of microbes.

Supplemental Figure 2

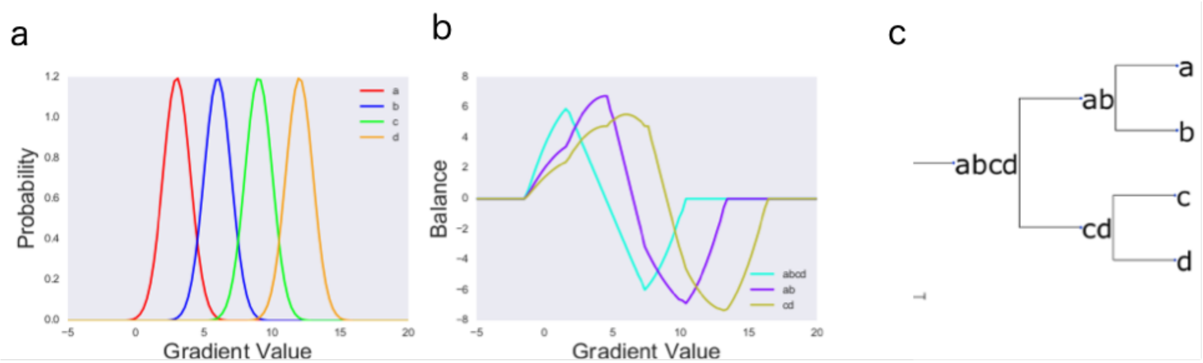


Figure C.2: An ecological intrepretation of balances. A simulation of 4 species, where each species is normally distributed along some environmental gradient. Each species has a normal distribution with a variance of 3 and a mean of 3, 6, 9 and 12 respectively as shown in Figure S2a. The resulting balances can be calculated as follows.

$$abcd = \log\frac{\sqrt{ab}}{\sqrt{cd}} \qquad ab = \log\frac{\sqrt{a}}{\sqrt{b}} cd = \log\frac{\sqrt{c}}{\sqrt{d}}$$

Note, it is not possible to take a logarithm of zero. A commonly used approach around this problem is to add a pseudocount. Here we add a pseudocount of 1 after multiplying all of the species probabilities by 10000. These species abundances are transformed into balances as shown in Figure S2b. Because of the zero phenomenon, the balances yield something reminiscent to a triangular wave when applied to a pair of unimodal distributions. Take balance ab for example. At the far left around -5, neither a or b are present, so both of their abundances are zero. But since we are adding pseudocounts, the resulting balance is given by log(1/1)=0 . When the gradient value increases to 0, the abundance of a approaches the peak of the distribution, while the abundance of b is still zero, causing the ab balance to increase. By the time the gradient value is around 4, the abundances of b starts to appear, causing the ab to peak. When the gradient value is around 8, the abundance of a starts disappearing while the abundance of b starts approaching the maximum peak in Figure S2a. At a gradient value of 10, the abundance of b also begins to dwindle, and the ab balance spirals towards zero. This same triangular wave pattern appears in all 3 of these balances, and portions of this also appear in the 88 soils study as shown in Figure S3.

It is also important to note that a balance of zero also indicates that the abundances between the ratios are equal. So if a balance is zero, and a pseudocount scheme was used, either the proportions between the numerator and denominator are truly equal, or both the numerator and the denominator are zero.

C.3 Analysis of balances in soils

Supplemental Figure 3

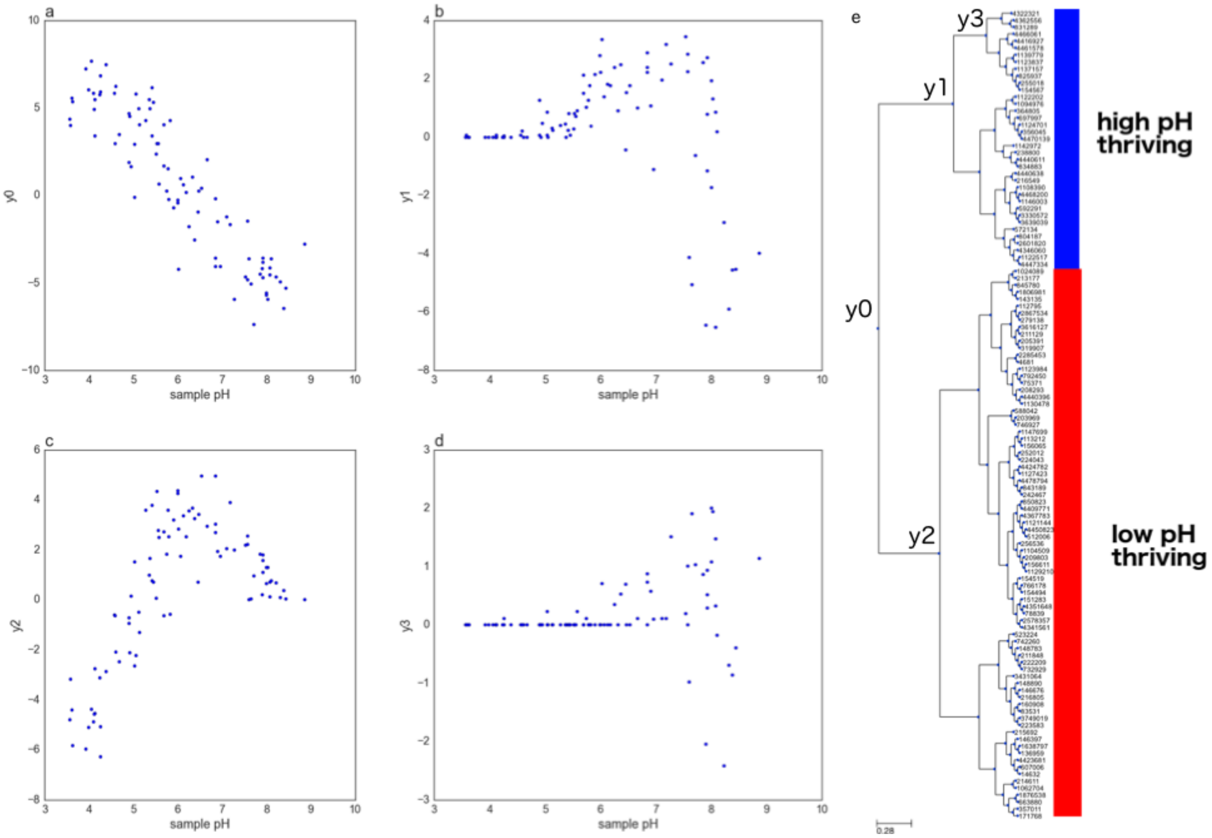


Figure C.3: Other balances from the 88 soils study. A perspective of different balances in the 88 soils study.

If there is truly a unimodal species distribution along pH, we'd expect to see the same sort of triangular wave pattern as shown in Figure 2S. If this is the case, then the top balance y0 is likely to be resulting from the midsection of the triangular wave between the minimum and the maximum (Figure S3a). The peaks of the triangular wave are a bit more apparent in the Figure S3b-d. In Figure S3b, the lower subtree in y1 is probably reaching a maximum in the true abundance around a pH of 7. In Figure S3c, the upper subtree in y2 is also likely approaching a maximum in the true abundance around a pH of 6. The same sort pattern could be happening in Figure S3d with the lower subtree in y3. These glimpses of triangular waves in these graphs suggest that there could be unimodel distributions of OTUs across the pH gradient.

As daunting as the zeros problem is, the zeros present in data sets such as the 88 soils follow predictable patterns. Even with a simple pseudo count strategy, we can still extract sensible information about balances of microbes across different pH values.

Appendix D

A brief overview of Aitchison Geometry

Given that many of the techniques presented in the dissertation were based on the concepts based on Aitchison geometry, we will provide a brief overview behind Aitchison geometry.

Aitchison geometry is a framework focused on the analysis of quantities including proportions, percentages, probabilities and concentrations. These quantities are also referred to as compositions. At the heart of the framework is the characterization of the Aitchison simplex, where each element of the space is a composition. A composition can be thought of as a set of proportions, or percentages.

Linear operations can be defined on compositions, known as "perturbation" and "powering operations". These operations are linear in the Aitchison simplex and can be transformed into traditional addition and multiplication operations in Euclidean space through the use of log-ratio transformations. Inner products can be defined in the Aitchison simplex, giving rise to the distance metrics such as the Aitchison distance. It can also be shown that the Aitchison simplex forms a finite Hilbert space [100].

D.1 Definition

The Aitchison simplex is formally defined for D species as follows

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

Where $\lambda > 0$ can be any positive real-valued constant. By definition, the compositions are the quantities denoted by $\mathbf{x}$.

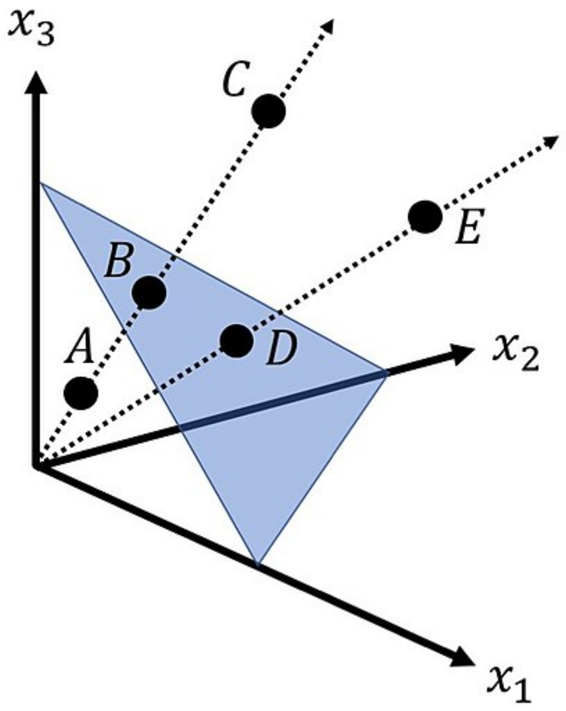


Figure D.1: An illustration of the Aitchison simplex. Here, there are 3 parts, x_1, x_2, x_3 represent values of different proportions. A, B, C, D and E are 5 different compositions within the simplex. A, B and C are all equivalent and D and E are equivalent.

There are three core axioms that the Aitchison simplex, namely

D.1.1 Scale invariance

Whether data is represented as proportions, percentages or probabilities, in the context of the Aitchison simplex all of these measurements are equivalent since they only differ by a constant scaling factor.

D.1.2 Subcompositional coherence

Observations on shared species should be consistent. For example, supposed that there are 2 biologists that visited the exact same rainforest to count insects. One biologist observed 3 species of spiders and 10 species of ants whereas the other biologist only observed 2 species of spiders and 7 species of ants. If these biologists observed the same 2 spider species and the same 7 species of ants, their conclusions about those species should be the same. For instance, they should notice that the ratio of those two spider species are consistent between their observations. While the universal formal definition is still not clearly established, this concept can be formalized to distance metrics as follows

$$d(x_k, y_k) \leq d(x, y) \qquad \forall x, y \in S^D,\ x_k, y_k \in S^k,\ S^k \subset S^D$$

D.1.3 Permutation invariance

The ordering of how the proportions or were measured or counted doesn't matter. This is analogous to how combinations are invariant to the order of selection.

D.2 Vector Space Structure

D.2.1 Properties

The Aitchison simplex has the following operators defined using the **closure** operation as follows

Perturbation

$$x \oplus y = [\frac{x_1y_1}{\sum_{i=1}^D x_iy_i}, \frac{x_2}{\sum_{i=1}^D x_iy_i}, \ldots, \frac{x_Dy_D}{\sum_{i=1}^D x_iy_i}] = C[x_1y_1, ..., x_Dy_D] \qquad \forall x, y \in S^D$$

Powering

$$\alpha \odot x = [\frac{x_1^\alpha}{\sum_{i=1}^D x_i^\alpha}, \frac{x_2^\alpha}{\sum_{i=1}^D x_i^\alpha}, \dots, \frac{x_D^\alpha}{\sum_{i=1}^D x_i^\alpha}] = C[x_1y_1, ..., x_Dy_D] \qquad \forall x \in S^D, \quad \alpha \in \mathbb{R}$$

Inner product

$$\langle x, y \rangle = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j} \qquad \forall x, y \in S^D$$

Under these these operations alone, it is sufficient to show that the Aitchison simplex forms a Euclidean vector space.

D.2.2 Orthonormal bases

Since the Aitchison simplex forms a finite Hilbert space, it is possible to construct orthonormal bases in the simplex. Every composition can be decomposed as follows

$$x = \bigoplus_{i=1}^{D} x_i \odot e_i$$

Where $e_1, \dots e_{D-1}$ forms an orthonormal basis in the simplex [99].

D.3 Linear transformations

There are 3 well-characterized isomorphisms that transform from the Aitchison simplex to real space. All of these transforms satisfy linearity and as given below

D.3.1 Additive Log-ratio Transform

The additive log ratio (alr) transform is an where $alr : S^D \to \mathbb{R}^{D-1}$. This is given by

$$alr(x) = \left[\log \frac{x_1}{x_D} \dots \log \frac{x_{D-1}}{x_D}\right]$$

The choice of denominator is arbituary, and could be any specified component. This transform is commonly used in chemistry with measurements such as pH. In addition, this is the transform most commonly used for Multinomial logistic regression. The alr transform is not an isometry, meaning that distances on transformed values will not be equivalent to distances on the original compositions in the simplex.

D.3.2 Center Log-ratio Transform

The center log ratio (clr) tranform is both an isomorphism and an isometry where $clr : S^D \to \mathbb{U}, \quad U \subset \mathbb{R}^D$

$$clr(x) = \left[\log \frac{x_1}{g(x)} \dots \log \frac{x_{D-1}}{g(x)}\right]$$

The inverse of this function is also known as the softmax function commonly used in neural networks.

D.3.3 Isometric Log-ratio Transform

The isometric log ratio (ilr) tranform is both an isomorphism and an isometry where $ilr : S^D \to \mathbb{R}^{D-1}$

$$ilr(x) = \left[\langle x, e_1 \rangle, \dots \langle x, e_{D-1} \rangle\right]$$

There are multiple ways to construct orthonormal bases, including using the Gram–Schmidt process Singular-value decomposition of clr transformed data. Another alternative is to construct log contrasts from a bifurcating tree. If one is given a bifurcating tree, we can construct a basis from the internal nodes in the tree.

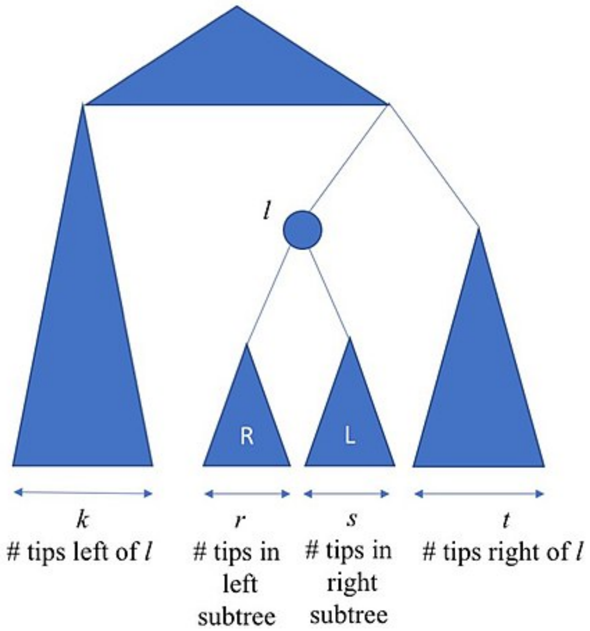


Figure D.2: An illustration of the bifurcating trees as an orthonormal basis. A representation of a tree in terms of its orthogonal components. l represents an internal node, an element of the orthonormal basis. This is a precursor to using the tree as a scaffold for the ilr transform.

Each vector in the basis would be determined as follows

$$e_l = C[exp(\underbrace{0,...0}_{k},\underbrace{a,...,a}_{r},\underbrace{b,...,b}_{s},\underbrace{0,...0)}_{t}]$$

The elements within each vector are given as follows

$$a = \frac{\sqrt{s}}{\sqrt{r(r+s)}} \quad \text{and} \quad b = \frac{-\sqrt{r}}{\sqrt{s(r+s)}}$$

where k,r,s,t are the respective number of tips in the corresponding subtrees shown in the figure. It can be shown that the resulting basis is orthonormal [94].

Once the basis Ψ is built, the ilr transform can be calculated as follows

$$ilr(x) = C[\exp(clr(x)\Psi)]$$

where each element in the ilr transformed data is of the following form

$$b_i = \sqrt{\frac{rs}{r+s}} \log \frac{g(x_R)}{g(x_S)}$$

where x_R and x_S are the set of values corresponding to the tips in the subtrees R and S.

Bibliography

[1] Jennifer BH Martiny, Stuart E Jones, Jay T Lennon, and Adam C Martiny. Microbiomes in light of traits: a phylogenetic perspective. *Science*, 350(6261):aac9323, 2015.

[2] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, et al. A new view of the tree of life. *Nature Microbiology*, 1:16048, 2016.

[3] David Tilman. *Resource competition and community structure*. Princeton university press, 1982.

[4] Robert H MacArthur. Environmental factors affecting bird species diversity. *The American Naturalist*, 98(903):387–397, 1964.

[5] Robert McCredie May. *Stability and complexity in model ecosystems*, volume 6. Princeton university press, 2001.

[6] Roger Arditi and Lev R Ginzburg. *How species interact: altering the standard view on trophic ecology*. Oxford University Press, 2012.

[7] Human Microbiome Project Consortium et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[8] Paul G Falkowski, Tom Fenchel, and Edward F Delong. The microbial engines that drive earth's biogeochemical cycles. *science*, 320(5879):1034–1039, 2008.

[9] Richard D Bardgett, Chris Freeman, and Nicholas J Ostle. Microbial contributions to climate change through carbon cycle feedbacks. *The ISME journal*, 2(8):805–814, 2008.

[10] László Zsolt Garamszegi. Modern phylogenetic comparative methods and their application in evolutionary biology. *Concepts and Practice. London, UK: Springer*, 2014.

[11] Emilia P Martins and Thomas F Hansen. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667, 1997.

[12] Morgan GI Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron E Burkepile, Rebecca L Vega Thurber, Rob Knight, et al. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature biotechnology*, 31(9):814–821, 2013.

[13] Alex D Washburne, Justin D Silverman, Jonathan W Leff, Dominic J Bennett, John L Darcy, Sayan Mukherjee, Noah Fierer, and Lawrence A David. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5:e2969, February 2017.

[14] Frederick A Matsen IV and Steven N Evans. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS One*, 8(3):e56859, 2013.

[15] Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife*, 6:e21887, 2017.

[16] Catherine Lozupone and Rob Knight. UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and environmental microbiology*, 71(12):8228–8235, 2005.

[17] Jacob Socolar and Alex Washburne. Prey carrying capacity modulates the effect of predation on prey diversity. *The American Naturalist*, 186(3):333–347, 2015.

[18] K. S. McCann. The diversity-stability debate. *Nature*, 405:228, 2000.

[19] Jacob B Socolar, James J Gilroy, William E Kunin, and David P Edwards. How should beta-diversity inform biodiversity conservation? *Trends in ecology and evolution*, 31(1):67–80, 2016.

[20] John Aitchison. *The statistical analysis of compositional data*. Chapman and Hall London, 1986.

[21] Gregory B Gloor and Gregor Reid. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62(8):692–703, 2016.

[22] Masatoshi Nei and Sudhir Kumar. *Molecular evolution and phylogenetics*. Oxford university press, 2000.

[23] Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature reviews. Genetics*, 13(5):303, 2012.

[24] D. M. Hillis and M. T. Dixon. Ribosomal dna: molecular evolution and phylogenetic inference. *Q. Rev. Biol.*, page 411–453, 1991.

[25] Berend Snel, Peer Bork, and Martijn A Huynen. Genome phylogeny based on gene content. *Nature genetics*, 21(1):108–110, 1999.

[26] Jesse R Zaneveld, Catherine Lozupone, Jeffrey I Gordon, and Rob Knight. Ribosomal rna diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic acids research*, 38(12):3869–3879, 2010.

[27] Barry G Hall and Miriam Barlow. Evolution of the serine β-lactamases: past, present and future. *Drug Resistance Updates*, 7(2):111–123, 2004.

[28] J Peter Gogarten, W Ford Doolittle, and Jeffrey G Lawrence. Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution*, 19(12):2226–2238, 2002.

[29] Tomáš Větrovský and Petr Baldrian. The variability of the 16s rrna gene in bacterial genomes and its consequences for bacterial community analyses. *PloS one*, 8(2):e57923, 2013.

[30] C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, 73:1576–1585, 2007.

[31] E. A. Stone. Why the phylogenetic regression appears robust to tree misspecification. *Syst. Biol.*, 60:245–260, 2011.

[32] S. J. Riesenfeld and K. S. Pollard. Beyond classification: gene-family phylogenies from shotgun metagenomic reads enable accurate community analysis. *BMC Genomics*, 14, 2013.

[33] Joseph Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, pages 783–791, 1985.

[34] Alan Grafen. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 326(1233):119–157, 1989.

[35] Simon P Blomberg, James G Lefevre, Jessie A Wells, and Mary Waterhouse. Independent contrasts and pgls regression estimators are equivalent. *Systematic Biology*, 61(3):382–391, 2012.

[36] Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, 1999.

[37] Simon P Blomberg, Theodore Garland Jr, Anthony R Ives, and B Crespi. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4):717–745, 2003.

[38] Karasov Lavin, S. R., Ives W. H., A. R., K. M. Middleton, and T.Jr. Garland. Morphometrics of the avian small intestine compared with that of nonflying mammals: a phylogenetic approach. *Physiol. Biochem. Zool.*, 81:526–550, 2008.

[39] Patrik Lindenfors, Liam J Revell, and Charles L Nunn. Sexual dimorphism in primate aerobic capacity: a phylogenetic test. *Journal of evolutionary biology*, 23(6):1183–1194, 2010.

[40] Patrick H Bradley, Stephen Nayfach, and Katherine S Pollard. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLoS computational biology*, 14(8):e1006242, 2018.

[41] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.

[42] Klaus Peter Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011.

[43] Liam J Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012.

[44] Steven W Kembel, Peter D Cowan, Matthew R Helmus, William K Cornwell, Helene Morlon, David D Ackerly, Simon P Blomberg, and Campbell O Webb. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26(11):1463–1464, 2010.

[45] David Orme. The caper package: comparative analysis of phylogenetics and evolution in r. *R package version*, 5(2), 2013.

[46] Weir Harmon, L. J., J. T., C. D. Brock, R. E. Glor, and W. Challenger. Geiger: investigating evolutionary radiations. *Bioinformatics*, page 129–131, 2007.

[47] Ls Tung Ho and C. Ané. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst. Biol.*, 63:397–408, 2014.

[48] Clifford W Cunningham, Kevin E Omland, and Todd H Oakley. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution*, 13(9):361–366, 1998.

[49] Jeffrey B Joy, Richard H Liang, Rosemary M McCloskey, T Nguyen, and Art FY Poon. Ancestral reconstruction. *PLoS computational biology*, 12(7):e1004763, 2016.

[50] M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11:459–468, 1994.

[51] Joel A Klappenbach, John M Dunbar, and Thomas M Schmidt. rrna operon copy number reflects ecological strategies of bacteria. *Applied and environmental microbiology*, 66(4):1328–1333, 2000.

[52] Qin Chang, Yihui Luan, and Fengzhu Sun. Variance adjusted weighted unifrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC bioinformatics*, 12(1):118, 2011.

[53] Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

[54] Nathan G Swenson. Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PloS one*, 6(6):e21264, 2011.

[55] Jun Chen, Frederic D Bushman, James D Lewis, Gary D Wu, and Hongzhe Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2013.

[56] Elizabeth Purdom. Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. *The Annals of Applied Statistics*, pages 2326–2358, 2011.

[57] J. et al. Fukuyama. Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput. Biol.*, 13, 2017.

[58] Micah Hamady, Catherine Lozupone, and Rob Knight. Fast unifrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data. *The ISME journal*, 4(1):17, 2010.

[59] J Peter Gogarten and Jeffrey P Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature reviews. Microbiology*, 3(9):679, 2005.

[60] Ofir Cohen, Uri Gophna, and Tal Pupko. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Molecular biology and evolution*, 28(4):1481–1489, 2010.

[61] Kei Kitahara and Kentaro Miyazaki. Revisiting bacterial phylogeny: natural and experimental evidence for horizontal gene transfer of 16s rrna. *Mobile genetic elements*, 3(1):e24210, 2013.

[62] Nicola Segata, Daniela Börnigen, Xochitl C Morgan, and Curtis Huttenhower. Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, 4:2304, 2013.

[63] Cuong Than, Derek Ruths, and Luay Nakhleh. Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC bioinformatics*, 9(1):322, 2008.

[64] Matt Ravenhall, Nives Škunca, Florent Lassalle, and Christophe Dessimoz. Inferring horizontal gene transfer. *PLoS computational biology*, 11(5):e1004095, 2015.

[65] C. A. Lozupone and R. Knight. Species divergence and the measurement of microbial diversity. *FEMS Microbiol. Rev.*, 32:557–578, 2008.

[66] José Alexandre Felizola Diniz-Filho, Carlos Eduardo Ramos Sant'Ana, and Luis Mauricio Bini. An eigenvector method for estimating phylogenetic inertia. *Evolution*, 52(5):1247–1262, 1998.

[67] Rob P Freckleton, Natalie Cooper, and Walter Jetz. Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. *The American Naturalist*, 178(1):E10–E17, 2011.

[68] Pierre Legendre and Louis Legendre. *Numerical Ecology*. 2003.

[69] J Podani and I Miklos. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*, 83(12):3331–3343, 2002.

[70] M.O. Hill and H.G. Gauch. Detrended Correspondence Analysis: An Improved Ordination Technique. *Vegetatio*, 42(Kendall 1971), 1980.

[71] R Ejrnaes. Can we trust gradients extracted by Detrended Correspondence Analysis? *Journal of Vegetation Science*, 11(Minchin 1987):565–572, 2000.

[72] Persi Diaconis, Sharad Goel, Susan Holmes, et al. Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, 2(3):777–807, 2008.

[73] Sergio Camiz. The guttman effect: its interpretation and a new redressing method. 2005.

[74] J. M. Smith and N. H. Smith. Synonymous nucleotide divergence: What is 'saturation'? *Genetics*, 142(3):1033–1036, 1996.

[75] N J Tourasse and M Gouy. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. *Molecular biology and evolution*, 14:287–298, 1997.

[76] Justin Kuczynski, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Noah Fierer, and Rob Knight. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature methods*, 7(10):813–819, 2010.

[77] Christian L Lauber, Micah Hamady, Rob Knight, and Noah Fierer. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Applied and Environmental Microbiology*, 75(15):5111–5120, 2009.

[78] Michael J Greenacre. Theory and applications of correspondence analysis. 1984.

[79] Jessica L. Metcalf, Zhenjiang Zech Xu, Sophie Weiss, Simon Lax, Will Van Treuren, Embriette R Hyde, Se Jin Song, Amnon Amir, Peter Larsen, Naseer Sangwan, Daniel Haarmann, Greg C. Humphrey, Gail Ackermann, Luke R Thompson, Christian Lauber, Alexander Bibat, Catherine Nicholas, Matthew J Gebert, Joseph F Petrosino, Sasha C Reed, Jack A Gilbert, Aaron M. Lynne, Sibyl R. Bucheli, David O. Carter, and Rob Knight. Mammalian Corpse Decomposition. *Science*, 351(6269):158–162, 2016.

[80] Ofir Pele and Michael Werman. A linear time histogram metric for improved SIFT matching. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5304 LNCS(PART 3):495–508, 2008.

[81] Steven N Evans and Frederick A Matsen. The phylogenetic kantorovich–rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

[82] Sven Hoefman, David van der Ha, Nico Boon, Peter Vandamme, Paul De Vos, and Kim Heylen. Niche differentiation in nitrogen metabolism among methanotrophs within an operational taxonomic unit. *BMC Microbiology*, 14(1):83, 2014.

[83] Shabnam Qureshi, Brian K. Richards, Tammo S. Steenhuis, Murray B. McBride, Philippe Baveye, and Sylvie Dousset. Microbial acidification and pH effects on trace element release from sewage sludge. *Environmental Pollution*, 132(1):61–71, 2004.

[84] Gregory B. Gloor, Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. It's all relative: Analyzing microbiome data as compositions. *Annals of Epidemiology*, 26(5):322–329, 2015.

[85] Matthew C.B. Tsilimigras and Anthony A. Fodor. Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges. *Annals of Epidemiology*, 26(5):330–335, 2016.

[86] David Lovell, Warren Muller, Jennifer Taylor, Alec Zwart, and Chris Helliwell. Caution ! compositions ! can constraints on omics data lead analyses astray. (August 2016), 2010.

[87] Jonathan Friedman and Eric J Alm. Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology*, 8(9):1–11, 2012.

[88] David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: a valid alternative to correlation for relative data. *PLoS computational biology*, 11(3):e1004075, 2015.

[89] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, and Dan R Littman. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, pages 1–25, 2015.

[90] Sophie J Weiss, Zhenjiang Xu, Amnon Amir, Shyamal Peddada, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vazquez-Baeza, Amanda Birmingham, and Rob Knight. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ PrePrints*, 3:e1408, 2015.

[91] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.*, 10(8):538–550, 2012.

[92] Joseph N Paulson, O Colin Stine, Hector Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200, 2013.

[93] Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1):27663, 2015.

[94] J J Egozcue and Vera Pawlowsky-glahn. Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology*, 37(7), 2005.

[95] J J Egozcue. CoDa-dendrogram: A new exploratory tool. 2005.

[96] Justin D Silverman, Heather Durand, Rachael J Bloom, Sayan Mukherjee, and Lawrence A David. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts, 2018.

[97] Pixu Shi, Anru Zhang, and Hongzhe Li. Regression analysis for microbiome compositional data. *Annals of Applied Statistics*, 10(2):1019–1040, 2016.

[98] Robert A Quinn, Katrine Whiteson, Yan-wei Lim, Peter Salamon, Barbara Bailey, Simone Mienardi, Savannah E Sanchez, Don Blake, Doug Conrad, and Forest Rohwer. ORIGINAL ARTICLE A Winogradsky-based culture system shows an association between microbial fermentation and cystic fibrosis exacerbation. 9(4):1024–1038, 2014.

[99] J J Egozcue and C Barcel. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[100] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modelling and Analysis of Compositional Data*. 2015.

[101] Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4):e1003531, 2014.

[102] Juan Jos, Vera Pawlowsky-glahn, Karel Hron, and Peter Filzmoser. Simplicial regression . The normal model. *Journal of Applied Probability and Statistics*, pages 1–22.

[103] Josep A Martín-Fernández, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003.

[104] Lawrence A David, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm. Host lifestyle affects human microbiota on daily timescales. *Genome biology*, 15(7):R89, 2014.

[105] Daniel McDonald, Jose C Clemente, Justin Kuczynski, Jai Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, Rob Knight, and J Caporaso. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(1):7, 2012.

[106] Jaime Huerta-Cepas, Joaquín Dopazo, and Toni Gabaldón. ETE: a python Environment for Tree Exploration. *BMC bioinformatics*, 11(1):24, 2010.

[107] William H E Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24, 1984.

[108] J Gregory Caporaso, Christian L Lauber, William A Walters, Donna Berg-lyons, James Huntley, Noah Fierer, Sarah M Owens, Jason Betley, Louise Fraser, Markus Bauer, Niall Gormley, Jack A Gilbert, Geoff Smith, and Rob Knight. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, 6(8):1621–1624, 2012.

[109] J Gregory Caporaso, Christian L Lauber, William A Walters, Donna Berg-lyons, Catherine A Lozupone, Peter J Turnbaugh, Noah Fierer, and Rob Knight. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings in the National Academy of Sciences*, 2010.

[110] Sophie J Weiss, Zhenjiang Xu, Amnon Amir, Shyamal Peddada, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vazquez-Baeza, Amanda Birmingham, and Rob Knight. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ PrePrints*, 3:e1408, 2015.

[111] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.

[112] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods*, 10(12):1200–1202, December 2013.

[113] Jakob Russel, Jonathan Thorsen, Asker D Brejnrod, Hans Bisgaard, Soren J Sorensen, and Mette Burmolle. DAtest: a framework for choosing differential abundance or expression method, 2018.

[114] Stijn Hawinkel, Federico Mattiello, Luc Bijnens, and Olivier Thas. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.*, August 2017.

[115] Gregory B Gloor, Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. It's all relative: Analyzing microbiome data as compositions. *Ann. Epidemiol.*, 26(5):322–329, 2015.

[116] Doris Vandeputte, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, Raul Y Tito, Lindsey De Commer, Youssef Darzi, Séverine Vermeire, Gwen Falony, and Jeroen Raes. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681):507–511, November 2017.

[117] Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. 1:1–7, 2015.

[118] James T Morton, Jon Sanders, Robert A Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A Navas-Molina, Se Jin Song, Jessica L Metcalf, Embriette R Hyde, Manuel Lladser, Pieter C Dorrestein, and Rob Knight. Balance trees reveal microbial niche differentiation. *mSystems*, 2(1):e00162–16, February 2017.

[119] Wenke Smets, Jonathan W Leff, Mark A Bradford, Rebecca L McCulley, Sarah Lebeer, and Noah Fierer. A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing, 2015.

[120] Andrzej Tkacz, Marion Hortala, and Philip S Poole. Absolute quantitation of microbiota abundance in environmental samples. *Microbiome*, 6(1):110, June 2018.

[121] Mangala A Nadkarni, F Elizabeth Martin, Nicholas A Jacques, and Neil Hunter. Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology*, 148(Pt 1):257–266, January 2002.

[122] Tarmo Äijö, Christian L Müller, and Richard Bonneau. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics*, 34(3):372–380, February 2018.

[123] Neal S Grantham, Brian J Reich, Elizabeth T Borer, and Kevin Gross. MIMIX: a bayesian Mixed-Effects model for microbiome data from designed experiments. March 2017.

[124] Fan Xia, Jun Chen, Wing Kam Fung, and Hongzhe Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, December 2013.

[125] Jessica L Mark Welch, Blair J Rossetti, Christopher W Rieken, Floyd E Dewhirst, and Gary G Borisy. Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences*, 113(6):E791–E800, 2016.

[126] Martin Glatz, Philipp P Bosshard, Wolfram Hoetzenecker, and Peter Schmid-Grendelmeier. The role of malassezia spp. in atopic dermatitis. *J. Clin. Med. Res.*, 4(6):1217–1228, May 2015.

[127] Allyson L Byrd, Clay Deming, Sara K B Cassidy, Oliver J Harrison, Weng-Ian Ng, Sean Conlan, NISC Comparative Sequencing Program, Yasmine Belkaid, Julia A Segre, and Heidi H Kong. Staphylococcus aureus and staphylococcus epidermidis strain diversity underlying pediatric atopic dermatitis. *Sci. Transl. Med.*, 9(397), July 2017.

[128] Donald Y.M. Leung, Agustin Calatroni, Livia S. Zaramela, Nathan T. Dyjack1, Kanwaljit Brar1, Petra LeBeau, Gloria David, Keli Johnson, Susan Leung, Marco Ramirez-Gama1, Bo Liang, Cydney Rios, Michael T. Montgomery, Brittany N. Richers, Cliff F. Hall1, Kathryn A. Norquest1, John Jung, Irina Bronova, Simion Kreimer, Jr C. Conover Talbot, Debra Crumrine, Robert Cole, Peter Elias, Karsten Zengler, Max A. Seibold, Evgeny Berdyshev, Elena Goleva1, and on behalf of the NIH/NIAID funded Atopic Dermatitis Research Network. Biomarkers of the non-lesional skin surface identify atopic dermatitis with food allergy as a unique endotype. *Science Translational Medicine*, Submitted 2018 (Under Review).

[129] M H S Falk, M T Linder, C Johansson, and others. The prevalence of malassezia yeasts in patients with atopic dermatitis, seborrhoeic dermatitis and healthy controls. *Acta Derm. Venereol.*, 2005.

[130] Jacquelyn S Meisel, Geoffrey D Hannigan, Amanda S Tyldsley, Adam J SanMiguel, Brendan P Hodkinson, Qi Zheng, and Elizabeth A Grice. Skin microbiome surveys are strongly influenced by experimental design. *J. Invest. Dermatol.*, 136(5):947–956, May 2016.

[131] Wenyu Shi, Heyuan Qi, Qinglan Sun, Guomei Fan, Shuangjiang Liu, Jun Wang, Baoli Zhu, Hongwei Liu, Fangqing Zhao, Xiaochen Wang, Xiaoxuan Hu, Wei Li, Jia Liu, Ye Tian, Linhuan Wu, and Juncai Ma. gcmeta: a global catalogue of metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Res.*, October 2018.

[132] Antonio Gonzalez, Jose A Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D Swafford, Stephanie B Orchanian, Jon G Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J Gregory Caporaso, Pieter C Dorrestein, and Rob Knight. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods*, 15(10):796–798, October 2018.

[133] Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. 8(9):1–11, 2012.

[134] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[135] Mahvash Navazesh, Satish K S Kumar, and University of Southern California School of Dentistry. Measuring salivary flow: challenges and opportunities. *J. Am. Dent. Assoc.*, 139 Suppl:35S–40S, May 2008.

[136] Clarisse Marotz, Amnon Amir, Greg Humphrey, James Gaffney, Grant Gogul, and Rob Knight. DNA extraction for streamlined metagenomics of diverse environmental samples. *Biotechniques*, 62(6):290–293, June 2017.

[137] Daniel McDonald, Jose C Clemente, Justin Kuczynski, Jai Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, Rob Knight, and J Caporaso. The biological observation matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience*, 1(1):7, 2012.

[138] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7(5):335–336, May 2010.

[139] Amnon Amir, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, Luke R Thompson, Embriette R Hyde, Antonio Gonzalez, and Rob Knight. Deblur rapidly resolves Single-Nucleotide community sequence patterns. *mSystems*, 2(2), March 2017.

[140] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73(16):5261–5267, August 2007.

[141] Kenneth J Locey and Jay T Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, 2016.

[142] Christy A Hipsley and Johannes Müller. Beyond fossil calibrations: realities of molecular clock practices in evolutionary biology. *Frontiers in genetics*, 5:138, 2014.

[143] Félix Forest. Calibrating the tree of life: fossils, molecules and evolutionary timescales. *Annals of Botany*, 104(5):789–794, 2009.

[144] Ziheng Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996.

[145] Alan Hodgkinson and Adam Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature reviews genetics*, 12(11):756, 2011.