Tree-Aggregated Predictive Modeling of Microbiome Data

Jacob Bien [1,*], Xiaohan Yan [2], Léo Simpson [3,4], and Christian L. Müller[4,5,6,*]

[1]Department of Data Sciences and Operations, University of Southern California, CA, USA

[2]Microsoft Azure, Redmond, WA, USA

[3]Technische Universität München, Germany

[4]Institute of Computational Biology, Helmholtz Zentrum München, Germany

[5]Department of Statistics, Ludwig-Maximilians-Universität München, Germany

[6]Center for Computational Mathematics, Flatiron Institute, Simons Foundation, NY, USA
[*]correspondence to: jbien@usc.edu, cmueller@flatironinstitute.org

September 1, 2020

Abstract

Modern high-throughput sequencing technologies provide low-cost microbiome survey data across all habitats of life at unprecedented scale. At the most granular level, the primary data consist of sparse counts of amplicon sequence variants or operational taxonomic units that are associated with taxonomic and phylogenetic group information. In this contribution, we leverage the hierarchical structure of amplicon data and propose a data-driven, parameter-free, and scalable tree-guided aggregation framework to associate microbial subcompositions with response variables of interest. The excess number of zero or low count measurements at the read level forces traditional microbiome data analysis workflows to remove rare sequencing variants or group them by a fixed taxonomic rank, such as genus or phylum, or by phylogenetic similarity. By contrast, our framework, which we call `trac` (`t`ree-aggregation of `c`ompositional data), learns data-adaptive taxon aggregation levels for predictive modeling making user-defined aggregation obsolete while simultaneously integrating seamlessly into the compositional data analysis framework. We illustrate the versatility of our framework in the context of large-scale regression problems in human-gut, soil, and marine microbial ecosystems. We posit that the inferred aggregation levels provide highly interpretable taxon groupings that can help microbial ecologists gain insights into the structure and functioning of the underlying ecosystem of interest.

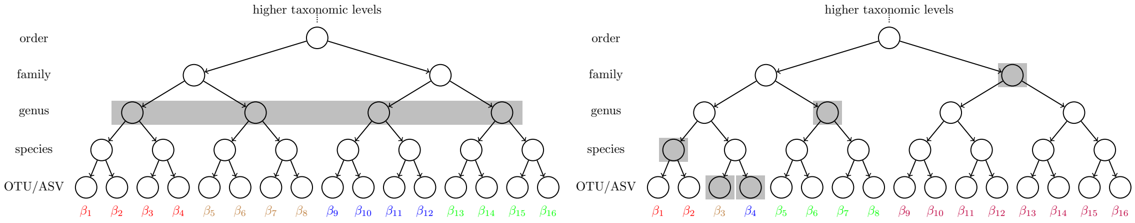
Introduction

Microbial communities populate all major environments on earth and significantly contribute to the total planetary biomass. Current estimates suggest that a typical human-associated microbiome consists of $\sim 10^{13}$ bacteria (Sender et al., 2016) and that marine bacteria and protists contribute to as much as 70% of the total marine biomass (Bar-On et al., 2018). Recent advances in modern targeted amplicon and metagenomic sequencing technologies provide a cost effective means to get a glimpse into the complexity of natural microbial communities, ranging from marine and soil to host-associated ecosystems (Sunagawa et al., 2015; Bahram et al., 2018; McDonald, 2018). However, relating these large-scale observational microbial sequencing surveys to the structure and functioning of microbial ecosystems and the environments they inhabit has remained a formidable scientific challenge.

Microbiome amplicon surveys typically comprise sparse read counts of marker gene sequences, such as 16S rRNA, 18S rRNA, or internal transcribed spacer (ITS) regions. At the most granular level, the data are summarized in count or relative abundance tables of operational taxonomic units (OTUs) at a prescribed sequence similarity level or denoised amplicon sequence variants (ASVs) (Callahan et al., 2017). The special nature of the marker genes enables taxonomic classification (Wang et al., 2007; Chaudhary et al., 2015) and phylogenetic tree estimation (Schliep, 2011), thus allowing a natural hierarchical grouping of taxa. This grouping information plays an essential role in standard microbiome analysis workflows. For instance, due to the excess number of zero or low count measurements at the OTU or ASV level, amplicon data pre-processing uses the grouping information for count aggregation where sequencing variants are pooled together at a higher taxonomic rank, for example, at the genus level, or according to phylogenetic similarity (Zhang et al., 2012; Chen et al., 2013; Xia et al., 2013; Lin et al., 2014; Randolph et al., 2015). In addition, rare sequence variants with incomplete taxonomic annotation are often simply removed from the sample.

This common practice of aggregating to a fixed taxonomic or phylogenetic level and then removing rare variants comes with several statistical and epistemological drawbacks. A major limitation of the fixed-level approach to aggregation is that it forces an awkward tradeoff between, on the one hand, using low-level taxa that are too rare to be informative (requiring throwing out many of them) and, on the other hand, aggregating to taxa that are at such a high level in the tree that one has lost much of the rich granularity in the original data. Aggregation to a fixed level attempts to impose an unrealistic “one-size-fits-all” mentality onto a complex, highly diverse system with dynamics that likely vary appreciably across the range of species represented. A fundamental premise of this work is that the decision of how to aggregate should not be made globally across an entire microbiome in preprocessing but should rather be integrated into the particular statistical analysis being performed. Many factors contribute to the question of how one should aggregate: the dynamics of the ecosystem under study, the abundance of different taxa, the goal of the statistical analysis, and the available quality of the sequencing data, including sequencing technology, sample sequencing depth, and sample size.

Another important factor when considering the practice of aggregating counts is that standard amplicon counts only carry relative (or “compositional”) information about the microbial abundances and thus require dedicated statistical treatment. When working with relative abundance data, Aitchison (1982); Egozcue and Pawlowsky-Glahn (2005); Gloor et al. (2017) show that counts should be combined with geometric averages rather than arithmetic averages. The practice of performing arithmetic aggregation of read counts before switching over to the geometric-average-based compositional data analysis workflow is incongruous.


higher taxonomic levels
order
family
genus
species
OTU/ASV
β_1 β_2 β_3 β_4 β_5 β_6 β_7 β_8 β_9 β_{10} β_{11} β_{12} β_{13} β_{14} β_{15} β_{16}
higher taxonomic levels
order
family
genus
species
OTU/ASV
β_1 β_2 β_3 β_4 β_5 β_6 β_7 β_8 β_9 β_{10} β_{11} β_{12} β_{13} β_{14} β_{15} β_{16}

Data	n	p	**Kingdom**	**Phylum**	**Class**	**Order**	**Family**	**Genus**	**Species**	**OTU**
Gut (HIV): sCD14	152	539	1	1	0	0	2	1	0	0
Gut (AGP): BMI	6266	1387	1	1	1	6	16	17	6	84
Central Park Soil: pH	580	3379	1	2	2	1	0	0	0	0
Central Park Soil: Moisture	580	3379	1	5	2	3	1	0	0	0
Fram Strait (PA): Leucine	26	3320	0	1	0	0	1	0	0	0
Fram Strait (FL): Leucine	25	4510	0	0	2	0	1	0	0	0
Ocean (TARA): Salinity	136	8916	1	1	2	0	0	0	0	0

Figure 1: A common preprocessing step is to aggregate microbiome amplicon data to a fixed level, such as genus, before performing subsequent analysis (upper left); `trac` instead performs a flexible tree-based aggregation in which the choice of what level to aggregate can vary across the tree (upper right) and is determined in a data-adaptive fashion with the goal of optimizing to the particular prediction task. Applying `trac` to seven different prediction tasks, which are characterized by differing environments, covariates of interest, number of samples n, and number of OTUs/ASVs p, reveals that aggregating at a wide range of levels can be useful for prediction (lower). In one case, we perform pH and moisture prediction on the same dataset (Central Park Soil), demonstrating that for different prediction tasks, different aggregations may be optimal.

To address these concerns, we propose a flexible, data-adaptive approach to tree-based aggregation that fully integrates aggregation into a statistical predictive model rather than relegating aggregation to preprocessing. Our method `trac` (`tr`ee-`a`ggregation of `c`ompositional data) learns dataset-specific taxon aggregation levels that are optimized for predictive regression modeling, thus making user-defined aggregation obsolete (see top panel of Figure 1 for illustration). Our framework is designed to mesh seamlessly with the compositional data analysis framework by combining log-contrast regression (Bacon-Shone and Aitchison, 1984) with tree-guided regularization, recently put forward in Yan and Bien (2020). Thanks to the convexity of the underlying penalized estimation problem, `trac` can deliver interpretable aggregated solutions to large-scale microbiome regression problems in a fast and reproducible manner.

We demonstrate the versatility of our framework by analyzing seven regression problems on five datasets covering human-gut, soil, and marine microbial ecosystems. Figure 1 illustrates our key idea and summarizes the properties of the analyzed microbial datasets, highlighting the heterogeneity of `trac`-inferred taxonomic aggregation levels for the respective regression tasks. For instance, we found that in Central Park soil, the phylum to family aggregation levels that were optimized for predicting soil pH were different from those for predicting soil moisture. In contrast, primary productivity prediction in the Fram Strait of the North Atlantic revealed almost identical microbial aggregations, independent of size class, namely log-contrasts of the Flavobacteriaceae family to the phylum Proteobacteria for particle-associated (PA) microbiota and Flavobacteriaceae to the Alpha- and Gammaproteobacteria classes, for free-living (FL) microbes.

Our `trac` framework complements other statistical approaches that make use of the available taxonomic or phylogenetic structure in microbial data analysis. For example, Lozupone and Knight (2005) use phylogenetic information in the popular `unifrac` metric to measure distances between microbial compositions. Washburne et al. (2017); Silverman et al. (2017); Morton et al. (2017); Washburne et al. (2019) combine tree information with the idea of "balances" from compositional data analysis (Egozcue and Pawlowsky-Glahn, 2005) to perform phylogenetically-guided factorization of microbiome data. Zhai et al. (2018); Xiao et al. (2018) include the tree structure in linear mixed models, Khabbazian et al. (2016) uses phylogenetic-tree-based regression for detecting evolutionary shifts in trait evolution, and Wang and Zhao (2017); Bradley et al. (2018) integrate tree-information into regression models for microbiome data.

In addition to our novel statistical formulation, we also offer an easy-to-use and highly scalable software framework for simultaneous taxon aggregation and regression. We believe that `trac`, available as an R package at `https://github.com/jacobbien/trac`, can be a valuable tool to microbial ecologists, significantly speeding up exploratory data analysis by delivering parsimonious and interpretable associations between microbiome data and variables of interest. This, in turn, will help formulate ecological questions and testable hypotheses about microbial niche differentiation, host-microbiome interactions, and ecosystem functioning.

Materials and methods

Modeling strategy

Let $y \in \mathbb{R}^n$ be n observations of a variable we wish to predict and let $X \in \mathbb{R}_+^{n\times p}$ be a matrix with X_{ij} giving the number of reads assigned to microbe j in sample i. The total number of reads $\sum_j X_{ij}$ in sample i is a reflection of the sequencing process and therefore should not be interpreted as providing meaningful information about the biological sample itself. This observation has motivated the adoption of compositional data methods, which ensure that analyses depend only on *relative* abundances. Following the foundational work of Bacon-Shone and Aitchison (1984), one appropriate model for regression with relative abundance data is the log-contrast model where the outcome of interest is modeled as linear combinations of log-ratios (i.e., log-contrasts) of relative abundance features. For high-dimensional microbiome data, Lin et al. (2014) propose solving an ℓ_1-penalized regression estimator that includes a zero-sum constraint on the coefficients. Writing $\log(X)$ for the matrix with ijth entry $\log(X_{ij})$, their estimator is of the form

$$\text{minimize}_{\beta\in\mathbb{R}^p} \quad L\left(y - \log(X)\beta\right) + \lambda\mathcal{P}(\beta) \;\text{ s.t. }\; 1_p^T\beta = 0. \tag{1}$$

Here, $L(r) = (2n)^{-1}\|r\|^2$ is squared error loss and $\mathcal{P}(\beta) = \|\beta\|_1$ is a lasso penalty (Tibshirani, 1996). The zero-sum constraint ensures that this model is equivalent to a log-contrast model

(Combettes and Müller, 2020) and invariant to sample-specific scaling. To see this, observe that replacing X by DX, where D is an arbitrary diagonal matrix, leaves Eq. (1) unchanged:

$$[\log(DX)\beta]_i = \sum_{j=1}^{p} \log(D_{ii}X_{ij})\beta_j = \sum_{j=1}^{p} [\log(D_{ii})\beta_j + \log(X_{ij})\beta_j] = 0 + [\log(X)\beta]_i.$$

Lin et al. (2014)'s choice of the ℓ_1 penalty was motivated by the high dimensionality of microbiome data and the desire for parsimonious predictive models. However, for the reasons discussed in Yan and Bien (2020), such a penalty is not well-suited to situations in which large numbers of features are highly rare, a well-known feature of amplicon data. Thus, Lin et al. (2014) adopt the common approach of preprocessing the data by aggregating read counts to the genus level and then screening out all but the most abundant genera. The left panel of Figure 1 depicts this standard practice: taxonomic (or phylogenetic) information in the form of a tree $\mathcal{T}$ is used to aggregate data, usually by taking the arithmetic mean, to a *fixed level* of the tree.

Our goal is to make aggregation more flexible (as shown in the right panel of Figure 1), to allow the prediction task to inform the decision of how to aggregate, and to do so in a manner that is consistent with the log-contrast framework introduced above. A key insight is that aggregating features can be equivalently expressed as setting elements of β equal to each other. For example, suppose we partition the p microbes into K groups $G_1, \ldots, G_K$ and demand that β be constant within each group. Doing so yields K aggregated features. If all of the β_j in group G_k are equal to some common value γ_k, then

$$\sum_j \beta_j \log(X_{ij}) = \sum_{k=1}^{K} \gamma_k \left(\sum_{j \in G_k} \log(X_{ij}) \right) = \sum_{k=1}^{K} \gamma_k |G_k| \cdot \log \left[(\prod_{j \in G_k} X_{ij})^{1/G_k} \right].$$

Thus, we are left with a linear model with K aggregated features, each being proportional to the log of the geometric mean of the microbe-level counts.

Associating the elements of β with the leaves of $\mathcal{T}$, the above insight tells us that if our estimate of β is constant within subtrees of $\mathcal{T}$, then that corresponds to a regression model with tree-aggregated features. In particular, each subtree with constant β-values will correspond to a feature, which is the log of the geometric mean of the counts within that subtree. The `trac` estimator uses a convex, tree-based penalty $\mathcal{P}_{\mathcal{T}}(\beta)$ for the penalty in Eq. (1) that is specially designed to promote β to have this structure of being constant along subtrees of $\mathcal{T}$. The mathematical form of $\mathcal{P}_{\mathcal{T}}(\beta)$ is given in Appendix A. In that appendix, we show that the `trac` estimator reduces to solving the following optimization problem:

$$\text{minimize}_{\alpha \in \mathbb{R}^{|\mathcal{T}|-1}} \quad L\left(y - \log(\text{geom}(X; \mathcal{T}))\alpha\right) + \lambda \sum_{u \in \mathcal{T}-\{r\}} w_u |\alpha_u| \text{ s.t. } 1_{|\mathcal{T}|-1}{}^T \alpha = 0, \qquad (2)$$

where $\text{geom}(X; \mathcal{T}) \in \mathbb{R}^{n \times (|\mathcal{T}|-1)}$ is a matrix where each column corresponds to a non-root node of $\mathcal{T}$ and consists of the geometric mean of all the microbe counts within the subtree rooted at u. Comparing this form of the `trac` optimization problem to Eq. (1) reveals an alternate perspective: `trac` can be interpreted as being like a sparse log-contrast model but instead of the features corresponding to microbes, they correspond instead to the geometric means of all non-root taxa in $\mathcal{T}$ (i.e., X is replaced by geom$(X;\mathcal{T})$). This also facilitates model interpretability since we can directly combine positive and negative predictors into pairs of log-ratio predictors. The particular choice of penalty is a weighted ℓ_1-norm. The trac package allows the user to specify general choices of weights $w_u > 0$; all results in this paper take w_u to be the inverse of the number of leaves in the subtree rooted at u.

The regularization parameter λ is a positive number determining the tradeoff between prediction error on the training data and how much aggregation should occur. By varying λ, we can trace out an entire solution path $\hat{\alpha}(\lambda)$, from highly sparse solutions (large λ) to more dense solutions involving many taxa (small λ). This "aggregation path" can itself be a useful exploratory tool in that it provides an ordering of the taxa as they enter the model.

Computation, model selection, and model quality assessment

Using trac in practice requires the efficient and accurate numerical solution of the convex optimization problem, specified in Eq. (2), across the full aggregation path. We experimented with several numerical schemes and found the path algorithm of Gaines et al. (2018) particularly well-suited for this task. The trac R package internally uses the path algorithm implementation from the classo Python module (https://github.com/Leo-Simpson/c-lasso), enabling solving high-dimensional trac problems on a standard laptop. Other R packages used in this paper include reticulate (Ushey et al., 2020), ggplot2 (Wickham, 2016), ape (Paradis and Schliep, 2019), igraph (Csardi and Nepusz, 2006), and ggtree (Yu et al., 2017).

To find a suitable aggregation level along the solution path, we use cross validation (CV) with mean squared error to select the regularization parameter $\lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}]$ for all the results presented in this paper. In particular, we perform 5-fold CV with the "one-standard-error rule" (1SE) (Hastie et al., 2009), which identifies the largest λ whose CV error is within one standard error of the minimum CV error. This heuristic purposely favors models that involve fewer taxa and are therefore easier to interpret. To assess how well a trac model generalizes to "unseen" data, we randomly select 2/3 of the samples in each of the considered datasets for model training and selection. On the remaining 1/3 of the samples, we compute out-of-sample test mean squared error as well as the correlation between model predictions and actual measurements on the test set.

Data collection

We consider a collection of five publicly available and previously analyzed datasets, spanning human gut, soil, and marine ecosystems (see Figure 1 lower panel). All datasets consists of 16S rRNA amplicon data of Bacteria and Archaea in form of OTU count tables, taxonomic classifications, and measured covariates, as provided in the original publications. Prior to trac analysis, zero counts in the datasets were replaced by a pseudo-count of one. No aggregation was performed on the data. For ease of interpretability, we leverage the taxonomic tree information rather than phylogeny in our aggregation framework. To investigate potential human host-microbiome interactions, we re-analyze two human gut datasets, one cohort of HIV patients (Gut (HIV)), available in (Rivera-Pinto et al., 2018), comprising $p = 539$ OTUs and $n = 152$ samples, and the other a subset of the American Gut

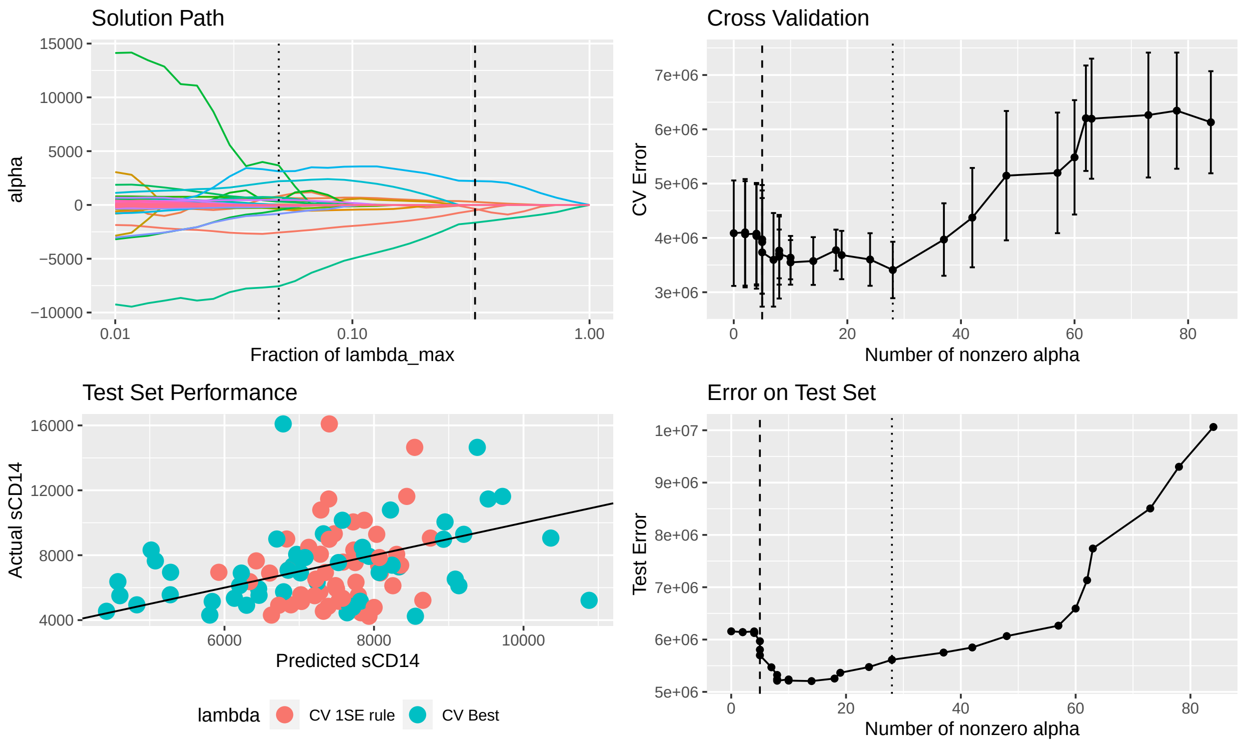
Project data (Gut (AGP)) (McDonald, 2018), provided in (Badri et al., 2020), comprising $p = 1387$ OTUs present in at least 10% of the $n = 6266$ samples. To study niche partitioning in terrestrial ecosystems, we use the Central Park soil dataset (Ramirez et al., 2014), as provided by Washburne et al. (2017), which consists of $p = 3379$ OTUs and $n = 580$ samples with a wide range of soil property measurements. For marine microbial ecosystems, we consider a sample collection from the Fram Strait in the North Atlantic (Fadeev et al., 2018), available at `https://github.com/edfadeev/Bact-comm-PS85`. The data set consists of $n = 26$ samples for $p = 3320$ OTUs in the particle-associated size class, and $n = 25$ samples for $p = 4510$ OTUs in the free-living size class. The second marine dataset is the Tara global surface ocean water sample collection (Sunagawa et al., 2015), available at `http://ocean-microbiome.embl.de/companion.html`. In Tara, each of the $p = 8916$ OTUs is present in at least 10% of the $n = 136$ samples. All data and analysis scripts used here are available as fully reproducible R workflows in the Supplementary Material.

Results and Discussion

Immune marker prediction in HIV patients

Infection with HIV is often paired with additional acute or chronic inflammation events in the epithelial barrier, leading to disruption of intestinal function and the microbiome. The interplay between HIV infection and the gut microbiome has been posited to be a "two-way street" (Dillon et al., 2016) since HIV-associated mucosal pathogenesis potentially leads to perturbation of the gut microbiome and, in turn, altered microbial compositions could result in ongoing disruption in intestinal homeostasis and secondary HIV-associated immune activation and inflammation.

Here, we investigate one aspect of this complex relationship by learning predictive models of immune markers from gut amplicon sequences. While Nowak et al. (2015) were among the first to provide evidence that gut microbial *diversity* is a predictor of HIV immune status (as measured by CD4+ cell counts), we consider soluble CD14 (sCD14) measurements in HIV patients as the variable to predict and learn an interpretable regression model from gut microbial amplicon data. sCD14 is a marker of microbial translocation and has been shown to be an independent predictor of mortality in HIV infection (Sandler et al., 2011). Following Rivera-Pinto et al. (2018), we analyze a HIV cohort of $n = 151$ patients where sCD14 levels (in pg/ml units) and fecal 16S rRNA amplicon data were measured. Using all available $p = 539$ bacterial and archaeal OTUs, we illustrate the typical `trac` prediction and model selection outputs in Figure 2. In the top left panel of Figure 2, we visualize the solution of the α coefficients associated with each aggregation along the regularization path. The vertical lines indicate the aggregations that were selected via cross-validation (CV) with the Minimum Mean Squared Error (MSE, dotted line) and one-standard-error rule (1SE, dashed line) (see top right panel in Figure 2). On the test data, we highlight the relationship between test prediction performance of the `trac` models versus the number of inferred aggregations (Figure 2 middle right panel). Models between five and 28 aggregations show excellent performance on the test set. `trac` with the 1SE rule identified a parsimonious model with aggregation to five main taxa (Figure 2 bottom panel): the kingdom Bacteria, phylum Actinobacteria and the family Lachnospiraceae are negatively associated, and the family Ruminococcaceae and the genus Bacteroides are positively associated with sCD14 counts, thus resulting in a `trac` model with three log-contrasts. To assess the stability of the present results, we repeated the same analysis on one hundred randomized training and test sample splits and recorded the selection frequency of taxon aggregations under the 1SE and MSE rules. The aggregations reported here are indeed the most stable ones. In addition, we found a genus with uncertain placement (*Incertae_Sedis*) classified into the Lachnospiraceae family to be frequently selected. This genus was also found to be a stable predictor of sCD14 in the genus-level analysis of Rivera-Pinto et al. (2018).



Kingdom	Phylum	Class	Order	Family	Genus	Species	OTU	α
Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae				2221.75
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae				-1644.86
Bacteria	Actinobacteria							-501.43
Bacteria								-362.27
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides			286.80

Figure 2: Varying the `trac` regularization parameter λ produces an solution path (top left). To select an interpretable aggregation, cross-validation (CV)is performed (top right) using either the solution with minimum CV error (dotted vertical line) or the 1SE rule (dashed vertical line). We hope to select a model along the solution path that is sparse and has small test error (middle right). The actual vs. predicted values of sCD14 on a test set (middle left) gives a closer look at the performance of the two models. In this case, the CV Best solution attains better test set performance (test correlation: 0.37) than CV 1SE rule (test correlation: 0.23) at the expense of a denser model. The `trac` model selected with the 1SE rule comprises five taxa across four levels, listed in the bottom panel.

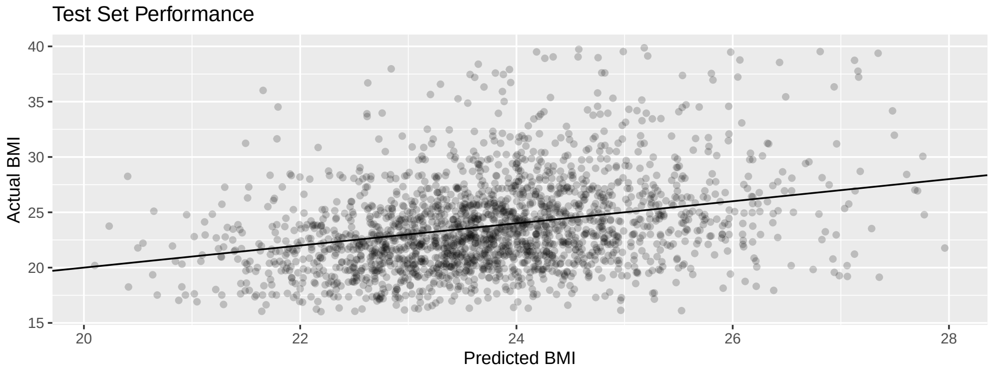
Overall, our analysis suggests a strong role of the Ruminococcaceae/Lachnospiraceae family ratio and, to a lesser extent, the Ruminococcaceae/Actinobacteria ratio in predicting mucosal disruption (as measured by sCD14). The protective or disruptive roles of Ruminococci or Lachnospiraceae in HIV patients is typically considered to be highly species-specific. Moreover, few consistent microbial patterns are known that generalize across studies (Dubourg, 2016). For instance, Monaco et al. (2016) report high variability and diverging patterns of the differential abundances of individual OTUs belonging to the Ruminococcaceae and Lachnospiraceae family in HIV-negative and HIV-positive participants. Our model posits that, on the family level, consistent effects of these two families are detectable in amplicon data. This also suggests that, with the right aggregation level, a re-analysis of recent HIV-related microbiome data may, indeed, reveal reproducible patterns of different taxon groups in HIV infection.

BMI prediction from American Gut microbiome profiles

It has proven to be difficult to find consistent gut microbial signatures that are predictive of a person's body mass index (BMI). Several early studies argued that obesity is associated with phylum-level changes in the microbiome (Turnbaugh et al., 2009), including increased Firmicutes to Bacteroidetes phyla ratios (Ley et al., 2006), often referred to as a hallmark predictor of obesity. On the COMBO dataset (Wu et al., 2011), Lin et al. (2014) and Shi et al. (2016) were among the first to identify a small set of microbial genera that were predictive of host BMI under a sparse log-contrast model.

Using `trac`, we revisit BMI prediction from microbial abundance data using a subset of the American Gut Project (AGP) data comprising $p = 1387$ OTUs across $n = 6266$ participants in the lean to obese BMI range. The `trac` model with the 1SE rule identified a model with 132 predictors, consisting of aggregations across *all* taxonomic levels. The lower panel in Figure 3 summarizes the 15 strongest predictors which include the kingdom Bacteria (vs. Archaea) as negative baseline, the phylum Bacteroidetes and several families and genera in the class Clostridia (which belongs to the Firmicutes phylum) with positive associations. The strongest positive OTU level predictor is an unknown species belonging to the Ruminococcaceae family. The top panel of Figure 3 shows the corresponding `trac` model BMI predictions (with 1SE rule) vs. measured BMI on the test set. The out-of-sample test correlation is 0.33 and the BMI test error is 15.31. For reference, Lin et al. (2014) reported on the COMBO data a BMI prediction test error (albeit with a different train/test split) of about 30 using a genus-level log-contrast model, consisting of the four genera Alistipes, Clostridium, Acidaminococcus, and Allisonella.

By contrast, our model contains considerably more predictive aggregations across all taxonomic levels. For instance, on the genus level, `trac` selects Blautia, Dorea, and Ruminococcus as positive predictors. The strongest overall positive predictors are the Bacteroidetes phylum, and the Ruminococcaceae, Lachnospiraceae, and Clostridiales families. The Lachnospiraceae/Bacteria ratio is also the first log-contrast to enter the `trac` aggregation path on the AGP data. The Erysipelotrichaceae and the Mogibacteriaceae families are the strongest negative predictors. Consistent with our model, Mogibacteriaceae were shown to be more abundant in lean individuals (Oki et al., 2016), and Erysipelotrichaceae were recently reported to be more abundant in normal compared to obese people or subjects



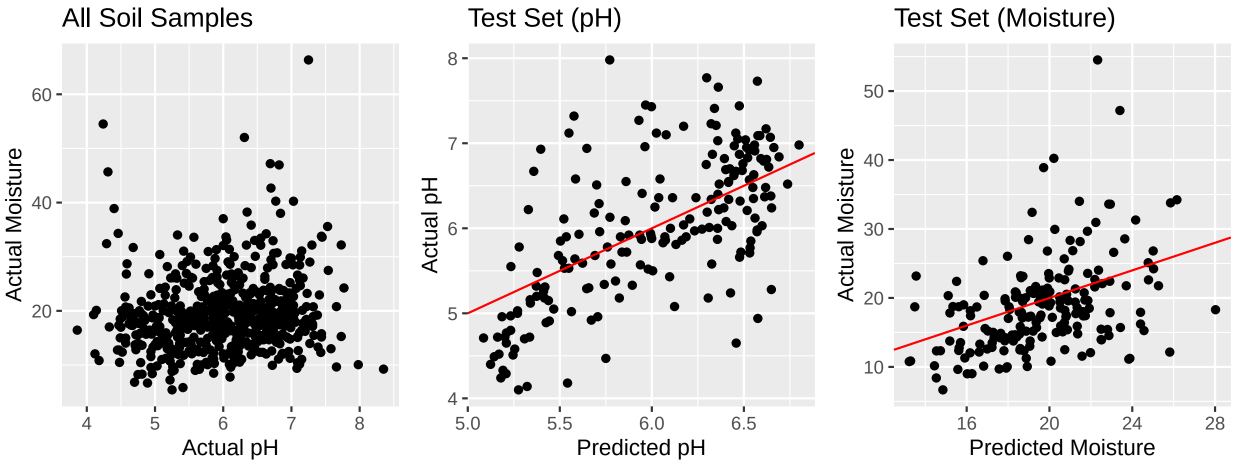
Kingdom	Phylum	Class	Order	Family	Genus	Species	OTU	α
Bacteria								-11.95
Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae				2.86
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae				2.23
Bacteria	Bacteroidetes							1.45
Bacteria	Firmicutes	Clostridia	Clostridiales					1.18
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae				0.90
Bacteria	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae				-0.80
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia			0.73
Bacteria	Firmicutes	Bacilli	Lactobacillales					0.72
Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae				0.71
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Dorea			0.51
Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus			0.49
Bacteria	Firmicutes	Clostridia	Clostridiales	[Mogibacteriaceae]				-0.36
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	[Barnesiellaceae]				0.32
Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	-	-	4356062	0.30

Figure 3: A scatter plot of measured BMI vs. `trac` model BMI predictions on a test set of $n = 2088$ AGP participants (upper) shows that predicted BMIs largely cover the "normal" BMI range between 20 and 28 with an overall test set correlation of 0.33. This model has 132 selected taxa, ranging from Kingdom to OTU levels (lower panel shows the top 15 aggregations with largest α-coefficients).

with metabolic disorder (Chávez-Carbajal et al., 2019). However, the fact that even our `trac` model could not identify a simple sparse predictive aggregation model for BMI suggests that more complex statistical models are required for predictive modeling, including adjustment for available covariates such as diet, sex, and overall life style.

Predicting Central Park soil properties from microbial communities

Soil microbial compositions vary considerably across spatial scales and are shaped by myriads of biotic and abiotic factors. Using univariate regression models, Fierer and Jackson (2006) found that soil habitat properties, in particular pH and soil moisture deficit (SMD), can predict overall microbial "phylotype" diversity. Using $n = 88$ soil samples from North and South America, Lauber et al. (2009) showed that soil pH concentrations are strongly associated with amplicon sequence compositions, as measured by pairwise `unifrac` distances. Moreover, they found that soil pH correlated positively with the relative abundances of Actinobacteria and Bacteroidetes phyla, negatively with Acidobacteria, and not at all with Beta/Gammaproteobacteria ratios.



Kingdom	Phylum	Class	Order	Family	Genus	Species	OTU	α
Bacteria								-0.74
Bacteria	Acidobacteria	Acidobacteria-6						0.58
Bacteria	Bacteroidetes							0.45
Bacteria	Proteobacteria	Gammaproteobacteria						-0.19
Bacteria	Acidobacteria	Acidobacteriia	Acidobacteriales					-0.13
Bacteria	Verrucomicrobia							0.03

Kingdom	Phylum	Class	Order	Family	Genus	Species	OTU	α
Bacteria								-20.36
Bacteria	Proteobacteria							19.53
Bacteria	Proteobacteria	Deltaproteobacteria						5.12
Bacteria	Verrucomicrobia							-2.78
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae				-2.48
Bacteria	Acidobacteria							1.55
Bacteria	Bacteroidetes							1.20
Bacteria	Actinobacteria							-1.05
Bacteria	Acidobacteria	Acidobacteriia	Acidobacteriales					-0.38
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales					-0.33
Bacteria	Proteobacteria	Betaproteobacteria						-0.26
Bacteria	Acidobacteria	DA052	Ellin6513					0.24

Figure 4: Top panel: Scatter plot between moisture and pH (left panel; correlation 0.13), `trac` models show that the microbiome is predictive of both pH (middle panel; test set correlation 0.65) and moisture (right panel; test set correlation 0.42). The 45° lines in the middle and right panels are included for reference. Lower panel: List of selected aggregations for pH (top) and moisture prediction (bottom).

Here, we used `trac` on the Central Park soil data collection comprising $n = 580$ samples and $p = 3379$ bacterial and archaeal OTUs (Ramirez et al., 2014; Washburne et al., 2017) to provide a refined analysis of the relationship between soil microbiome and habitat properties. Rather than looking at the univariate correlative pattern between soil properties and phyla, we build multivariate models that take soil pH and moisture as response variables of interest and optimize taxa aggregations using our predictive framework.

For pH, `trac` found an interpretable model with six aggregated taxonomic groups: the two phyla Bacteroidetes and Verrucomicrobia and the class Acidobacteria-6 were positively associated, whereas the order Acidobacteriales, the class Gammaproteobacteria, and the overall kingdom of Bacteria (compared to Archaea) were negatively associated with pH (see the table in the middle panel of Figure 4). We can thus associate a log-contrast model with three log-ratios of aggregated taxonomic groups with soil pH in Central Park. The overall correlation between the `trac` predictive model and the training data was 0.68. On the test data, the model still maintained a high correlation of 0.65 (see also Figure 4 top panel).

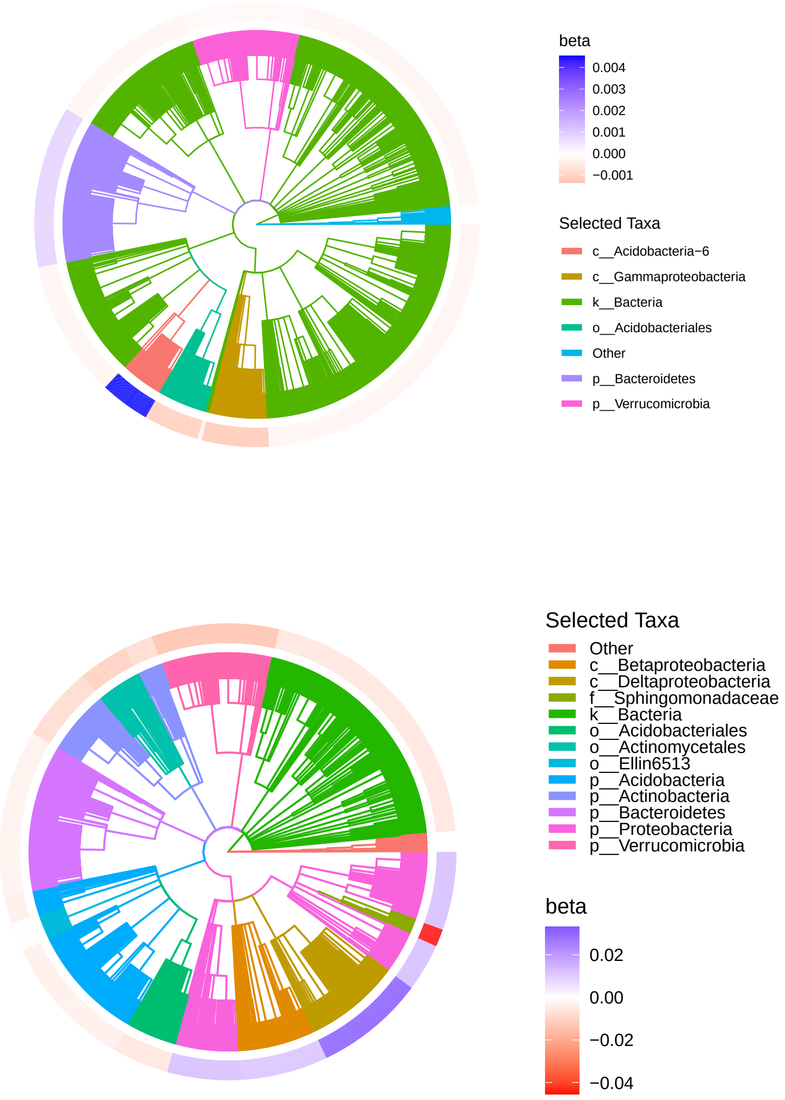


Figure 5: Taxonomic aggregations (as highlighted by branch colors) inferred by `trac`, that are predictive of Central Park soil pH and moisture, respectively. The color coding on the outermost ring corresponds to the estimated leaf coefficients β (see also Appendix A).

With the standard caveat that regression coefficients do not have the same interpretation (or even necessarily have the same sign) as their univariate counterparts, our model also supports a positive relationship between the Bacteroidetes phylum and pH and gives refined insights into the role of the Acidobacteria phylum. The model posits that the class Acidobacteria-6 is positively related and the order Acidobacteriales (in the Acidobacteriia class) is negatively related with pH. Washburne et al. (2017) observed similar groupings in their phylofactorization of the Central Park data.

There, the classes Acidobacteria-6 and Acidobacteriia belonged to different "binned phylogenetic units" whose relative abundances increased and decreased along the pH gradient, respectively. Finally, the phylum Verrucomicrobia and the class Gammaproteobacteria, included in our model, have been reported to be highly affected by pH with several species of Gammaproteobacteria particlarly abundant in low pH soil (Bartram et al., 2014).

We next investigate the relationship between soil microbiome and gravimetric moisture (% water) measurements in Central Park. As shown in the left plot in the upper panel of Figure 4, there is no apparent pattern between measured pH and moisture in the dataset. Using `trac`, we inferred a predictive model of moisture consisting of twelve taxonomic aggregations, including the phyla Verrucomicrobia and Actinobacteria, and the family Sphingomonadaceae as strong negative predictors, and the phylum Proteobacteria and the class Deltaproteobacteria as strong positive predictors (see lower table of Figure 4). Overall, there are five positive and seven negative coefficients, leading to a log-contrast model with at least seven log-ratios. On the test data, the correlation between model predictions and measurements was 0.42 (Figure 4 right plot of upper panel). Our reduced predictive power is in agreement with Fierer and Jackson (2006)'s observation about the smaller influence of SMD compared to pH on microbial composition. Nonetheless, `trac`'s taxonomic groupings provide meaningful information about the taxonomic structure of soil microbiota along moisture gradients. For example, the model supports the positive association between Proteobacteria and moisture, as previously observed in a study along a vegetation gradient on the Loess Plateau in China (Zeng et al., 2016), and the negative effect of moisture on the phylum Verrucomicrobia and the positive effect on Deltaproteobacteria in the Giessen free-air CO2 enrichment (Gi-FACE) experiment (de Menezes et al., 2016). The Gi-FACE study, however, also reported several relationships between the microbiome and the soil moisture that are incongruent with our model, including the role of Acidobacteria.

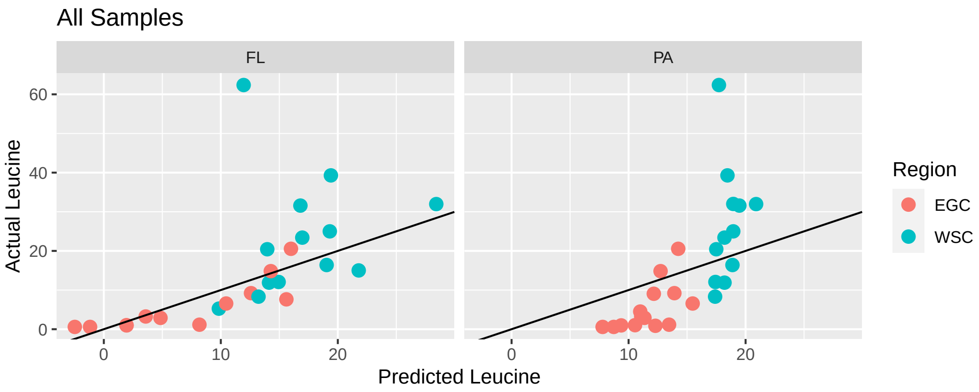
Finally, Figure 5 compares the aggregations across the taxonomic tree that were found by `trac` for soil pH and moisture prediction, respectively. We observe that only the phyla Bacteroidetes and Verrucomicrobia, and the order Acidobacteriales are common in both models, confirming that the relevant taxonomic aggregations depend on the response variable being predicted.

Primary bacterial production in the Fram Strait

Current estimates suggest that the ocean microbiome could be responsible for about half of all primary production occurring on Earth (Longhurst et al., 1995; Moran, 2015). While net primary production is known to be highly influenced by a multitude of environmental drivers, including light, nutrients, and temperature (Boyd et al., 2014), it is not yet established whether amplicon sequencing data alone contain enough information to serve as a stable predictor of (regional) marine primary production.

To investigate this relationship we use a recent dataset from Fadeev et al. (2018) covering the Fram Strait, the main gateway between the North Atlantic and Arctic Oceans, to tackle this question with `trac`. The Fram Strait comprises two distinct oceanic regions, the northward flowing West Spitsbergen Current (WSC), and the East Greenland Current (EGC) flowing southward along the Greenland shelf. Recent ocean simulations, however, suggest substantial horizontal mixing and exchange by eddies between the two regions. We thus learn regression models from amplicon data across both regions and considered the available leucine incorporation (as proxy to bacterial production) as the outcome (Fadeev et al., 2018). We train separate models for the two different size fractions ($p = 4530$ free-living (FL) taxa in the $0.22\mu m$ fraction and $p = 3320$ particle-associated (PA) taxa in $3\mu m$ fraction).

On the FL dataset, `trac` identifies a parsimonious model, comprising three aggregated taxonomic groups, strongly associated with bacterial production. The two classes Gammaproteobacteria and Alphaproteobacteria are negatively associated, and the family Flavobacteriaceae is positively associated with bacterial production, leading to a two-factor log-contrast model. The overall correlation between the `trac` prediction and the training data is 0.85. On the test data, the model maintains a correlation of 0.57. On the PA dataset, `trac` infers a single predictive log-contrast with the Flavobacteriaceae family being positively associated and the entire phylum Proteobacteria negatively associated with primary production. On the test data, the PA model predictions show a correlation of 0.90 with the measurements. Figure 6 summarizes the scatter plots of leucine measurements vs. `trac` predictions for the two size fractions, colored by region WSC and EGC, respectively.



Kingdom	Phylum	Class	Order	Family	Genus	Species	OTU	α
Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae				27.90
Bacteria	Proteobacteria	Alphaproteobacteria						-23.40
Bacteria	Proteobacteria	Gammaproteobacteria						-4.49

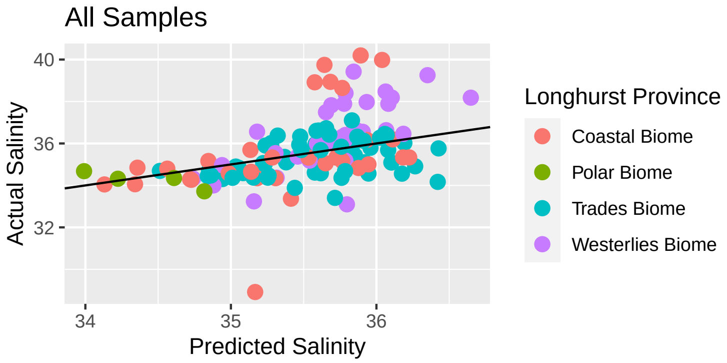
Kingdom	Phylum	Class	Order	Family	Genus	Species	OTU	α
Bacteria	Proteobacteria							-13.93
Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae				13.93

Figure 6: Top panel: `trac` models of primary production (leucine) from free living (FL) and particle associated (PA) taxa. The correlation between predicted and measured leucine (restricted to the test sets) is 0.57 for FL taxa and 0.90 and PA taxa, respectively. Lower panel: List of selected FL (top) and PA taxa (bottom) and aggregation coefficients α.

We observe that the PA model appears to serve as an implicit region classifier since predicted leucine values of < 17 belong uniquely to samples in the low-productivity EGC region (see top right panel in Figure 6). Our model suggests an important positive association of the heterotrophic Flavobacteriaceae with primary production, independent of size class. Flavobacteriaceae are known to strongly contribute to mineralization of primary-produced organic matter (see Bowman and Nichols 2005 and references therein), thus suggesting an indirect relationship between Flavobacteriaceae and primary production. However, previous studies in South polar front and antarctic zone postulated a strong role of Flavobacteriaceae for polar primary production (Abell and Bowman, 2005).

Global predictive model of ocean salinity from Tara data

Integrative marine data collection efforts such as Tara Oceans (Sunagawa et al., 2020) or the Simons CMAP (`https://simonscmap.com`) provide the means to investigate ocean ecosystems on a global scale. Using Tara's environmental and microbial survey of ocean surface water (Sunagawa et al., 2015), we next illustrate how `trac` can be used to globally connect environmental covariates and the ocean microbiome. As an example, we learn a global predictive model of ocean salinity from $n = 136$ samples and $p = 8916$ miTAG OTUs (Logares et al., 2014). `trac` identifies four taxonomic aggregations, the kingdom Bacteria and the phylum Bacteroidetes being negatively associated and the classes Alpha- and Gammaproteobacteria being positively associated with marine salinity (see Figure 7 lower panel). Figure 7 shows the scatter plot of salinity measurements vs. `trac` model predictions (with correlation on the out-of-sample test set of 0.55).



Kingdom	Phylum	Class	Order	Family	Genus	Species	OTU	α
Bacteria	Proteobacteria	Alphaproteobacteria						4.00
Bacteria								-2.92
Bacteria	Bacteroidetes							-1.38
Bacteria	Proteobacteria	Gammaproteobacteria						0.30

Figure 7: Top panel: Measured salinity (y-axis) vs. `trac` model prediction (x-axis) on the Tara data. The correlation between prediction and actual salinity on the test set is 0.55. Each sample is colored by one of the four Longhurst Biome definitions. Outliers to the model are located in Coastal and Westerlies Biomes. Lower panel: List of selected taxa and aggregation coefficients α.

Our model shows good global predictive capabilities with a few high salinity outliers located in the Red Sea (Coastal Biome) and the Mediterranean Sea (Westerlies Biome). Despite the fact that salinity is known to be an important environmental factor in marine microbial ecosystems, most studies thus far have investigated the connection between the microbiome and salinity gradients on a local marine scale, in particular estuaries. For instance, Bouvier and Del Giorgio (2002); Cottrell and Kirchman (2003) observed a marked increase of Alphaproteobacteria with increasing salinity, consistent with the estimated positive relationship in the `trac` model. In a global marine microbiome meta-analysis, Yilmaz et al. (2016) reported Spearman rank correlations between phyla relative abundances and physicochemical water properties, including salinity. There, four out of five orders in the Bacteroidetes phylum, and three out of four orders belonging to Gammaproteobacteria were reported to be negatively correlated with salinity. This suggests that our aggregation model does not universally agree with standard univariate assessments of the influence of environmental factors. Nevertheless, we believe that our `trac`-derived two-factor log-ratio model of ocean salinity can further contribute to the understanding of the large-scale biogeography in global ocean surface water.

Conclusions

Finding predictive and interpretable relationships between microbial amplicon sequencing data and ecological, environmental, or host-associated covariates of interest is a cornerstone of exploratory data analysis in microbial biogeography and ecology. To this end, we have introduced `trac`, a novel scalable tree-aggregation regression framework for compositional amplicon data. The framework leverages the hierarchical nature of microbial sequencing data to learn parsimonious log-ratios of aggregated microbial compositions that best predict continuous environmental or host-associated response variables. `trac` seamlessly generalizes prior approaches to sparse log-contrast modeling (Lin et al., 2014; Shi et al., 2016; Rivera-Pinto et al., 2018) and shares similarities with ideas from tree-guided, *balance* modeling of compositional data (Egozcue and Pawlowsky-Glahn, 2005; Silverman et al., 2017; Washburne et al., 2017), albeit with a stronger focus on finding *predictive* relationships.

In the human gut microbiome context, the estimated `trac` model of immune marker sCD14 concentrations in HIV patients asserted a particularly strong predictive role of the Ruminococcaceae/Lachnospiraceae family ratio, thus delivering a testable hypothesis for future HIV-microbiome studies. In contrast, `trac` prediction of Body Mass Indices (BMIs) of participants in the American Gut Project revealed a dense model with more than one hundred aggregations across all available taxonomic levels. Our analysis is consistent with the complexity found in other recent large-scale approaches aiming at discovering taxonomic signatures of obesity (Peters et al., 2018).

The `trac` analysis of environmental microbiomes in soil and marine habitats consistently provided parsimonious taxonomic aggregations for predicting covariates of interest. Rather than describing univariate relationships between single bacterial compositions or microbial diversity and soil properties, we asked for microbial aggregations that best “predict” soil pH and moisture measurements. This revealed distinct microbial taxa ratios that aligned with the underlying environmental gradients. Similarly, we found predictive aggregated taxa signatures in marine ecosystems. For example, Flavobacteriaceae/Proteobacteria ratios accurately predicted regional leucine incorporation (as proxy for primary production), and Alpha- and Gammaproteobacteria/Bacteroidetes ratios well-aligned with sea surface water salinity on a global scale, reminiscent of the ubiquitous Firmicutes/Bacteroidetes ratio in the context of the gut microbiome and obesity.

The `trac` framework naturally lends itself to several methodological extensions that are easy to implement and may prove valuable in microbial ecology. Firstly, as highlighted in the gut microbiome context, inclusion of additional factors such as diet and life style would likely improve prediction performance. This can be addressed by combining `trac` with standard (sparse) linear regression to allow the incorporation of (non-compositional) covariates into the statistical model. Secondly, while we focused on predictive regression modeling of continuous outcomes, it is straightforward to adopt our framework to classification tasks when binary outcomes, such as, e.g., case vs. control group, or healthy vs. sick participants, are to be predicted. Thirdly, due to the compositional nature of current amplicon data, we presented `trac` in the common framework of log-contrast modeling. However, alternative forms of tree aggregations over compositions are possible, for instance, by directly modeling the relative abundances as features rather than log-transformed quantities. Tree aggregations would then amount to grouped relative abundance *differences* and not log-ratios, thus resulting in a different interpretation of the estimated model features.

In summary, we believe that our methodology and its implementation in the R package `trac`, together with the presented reproducible application workflows, provide a valuable blueprint for future data-adaptive aggregation and regression modeling in microbial biogeography and ecology research. This, in turn, should contribute to the generation of new interpretable and testable hypotheses about the factors that shape microbial ecosystems in their natural habitats.

Acknowledgments

We thank Dr. Javier Rivera-Pinto for providing the processed OTU tables for the Gut(HIV)-sCD14 data set. This work was supported by the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems/CBIOMES (Grant ID: 549939, JB). Jacob Bien was also supported in part by NIH Grant R01GM123993 and NSF CAREER Award DMS-1653017. Christian L. Müller acknowledges support from the Center for Computational Mathematics, Flatiron Institute, and the Institute of Computational Biology, Helmholtz Zentrum München. An earlier version of this work appeared as Chapter 4 of Xiaohan Yan's PhD dissertation (Yan, 2018).

Appendix

A Derivation of Optimization Problem

We design a convex tree-based penalty $\mathcal{P}_{\mathcal{T}}(\beta)$ that promotes β to be constant along branches of $\mathcal{T}$. We encode $\mathcal{T}$ through a binary matrix $A \in \{0,1\}^{p\times(|\mathcal{T}|-1)}$ indicating whether feature j is a leaf of each non-root node $u \in \mathcal{T}-\{r\}$, that is $A_{ju} = 1\{j \in \mathcal{L}(u)\}$ where $\mathcal{L}(u)$ is the set of leaves that descend from u. In particular, we take

$$\mathcal{P}_{\mathcal{T}}(\beta) = \min_{\gamma \in \mathbb{R}^{|\mathcal{T}|-1}} \{\|\gamma\|_1 \quad \text{s.t.} \quad \beta = A\gamma\}.$$

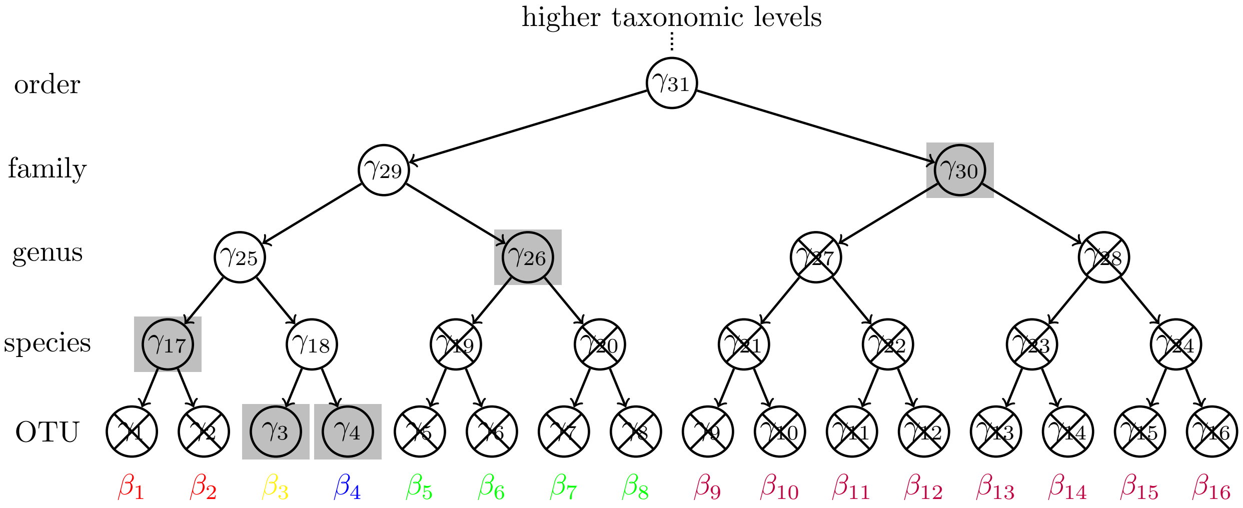

higher taxonomic levels
order
family
genus
species
OTU
γ_{31}
γ_{29}
γ_{30}
γ_{25}
γ_{26}
γ_{27}
γ_{28}
γ_{17}
γ_{18}
γ_{19}
γ_{20}
γ_{21}
γ_{22}
γ_{23}
γ_{24}
γ_1 γ_2 γ_3 γ_4 γ_5 γ_6 γ_7 γ_8 γ_9 γ_{10} γ_{11} γ_{12} γ_{13} γ_{14} γ_{15} γ_{16}
β_1 β_2 β_3 β_4 β_5 β_6 β_7 β_8 β_9 β_{10} β_{11} β_{12} β_{13} β_{14} β_{15} β_{16}

Figure 8: Schematic of the tree aggregation process.

Figure 8 shows a schematic of the tree aggregation idea. The vector $\gamma \in \mathbb{R}^{|\mathcal{T}|-1}$ can be thought of as a latent parameter vector with an entry associated with each node of the tree (see Figure 8). We associate a β_j to each leaf of $\mathcal{T}$, and the constraint $\beta = A\gamma$ expresses a particular relationship between these, namely that each coefficient β_j is the sum of the γ_u for which $j \in \mathcal{L}(u)$ (i.e., each β_j is the sum of its ancestor γ-values in the tree). This relationship implies that when all the γ-values in a subtree are zero (denoted by crossed out nodes in the figure), then all the β coefficients within the subtree are equal. Thus, the sparsity inducing ℓ_1-norm on γ in $\mathcal{P}_{\mathcal{T}}(\beta)$ induces β to tend to be constant within subtrees of $\mathcal{T}$. Using this penalty in Eq. (1) leads to the `trac` method, which is computed by solving,

$$\text{minimize}_{\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^{|\mathcal{T}|-1}} \quad L\left(y - \log(X)\beta\right) + \lambda\|\gamma\|_1 \ \text{s.t.}\ 1_p^T\beta = 0, \quad \beta = A\gamma. \tag{3}$$

This estimator is built on the tree-based aggregation penalty of Yan and Bien (2020), developed for general situations in which features are rare and a tree relating the features is available. In their setting, features are not compositional, so they do not introduce a sum-to-zero constraint or take the log of the features. The `trac` problem can be written more simply, entirely in terms of γ, as

$$\text{minimize}_{\gamma \in \mathbb{R}^{|\mathcal{T}|-1}} \quad L\left(y - \log(X)A\gamma\right) + \lambda\|\gamma\|_1 \ \text{s.t.}\ 1_p^T A\gamma = 0.$$

The $n \times (|\mathcal{T}| - 1)$ matrix $\log(X)A$ has the sum of the log counts of each of the $|\mathcal{T}| - 1$ subtrees of $\mathcal{T}$ (excluding $\mathcal{T}$ itself). Changing variables to $\alpha_u = \gamma_u \cdot |\mathcal{L}(u)|$ and using properties of logarithms establishes the equivalence with problem Eq. (2) in the main paper.

References

Abell, G. C. and Bowman, J. P. (2005). Ecological and biogeographic relationships of class Flavobacteria in the Southern Ocean. *FEMS Microbiology Ecology*, 51(2):265–277.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177.

Bacon-Shone, J. and Aitchison, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*.

Badri, M., Kurtz, Z. D., Bonneau, R., and Müller, C. L. (2020). Shrinkage improves estimation of microbial associations under different normalization methods. *bioRxiv*.

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Põlme, S., Sunagawa, S., Ryberg, M., Tedersoo, L., and Bork, P. (2018). Structure and function of the global topsoil microbiome. *Nature*, 560(7717):233–237.

Bar-On, Y. M., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25):6506–6511.

Bartram, A. K., Jiang, X., Lynch, M. D., Masella, A. P., Nicol, G. W., Dushoff, J., and Neufeld, J. D. (2014). Exploring links between pH and bacterial community composition in soils from the Craibstone Experimental Farm. *FEMS Microbiology Ecology*, 87(2):403–415.

Bouvier, T. C. and Del Giorgio, P. A. (2002). Compositional changes in free-living bacterial communities along a salinity gradient in two temperate estuaries. *Limnology and Oceanography*, 47(2):453–470.

Bowman, J. P. and Nichols, D. S. (2005). Novel members of the family Flavobacteriaceae from Antarctic maritime habitats including Subsaximicrobium wynnwilliamsii gen. nov., sp. nov., Subsaximicrobium saxinquilinus sp. nov., Subsaxibacter broadyi gen. nov., sp. nov., Lacinutrix copepodicola gen. nov., sp. nov., and novel species of the genera Bizionia, Gelidibacter and Gillisia. *International Journal of Systematic and Evolutionary Microbiology*, 55(4):1471–1486.

Boyd, P. W., Sundby, S., and Pörtner, H.-O. (2014). Net primary production in the ocean. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 133–136.

Bradley, P. H., Nayfach, S., and Pollard, K. S. (2018). *Phylogeny-corrected identification of microbial gene families relevant to human gut colonization*, volume 14.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, 11(12):2639–2643.

Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A., and Sharma, V. K. (2015). 16S classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS ONE*, 10(2):e0116106.

Chávez-Carbajal, A., Nirmalkar, K., Pérez-Lizaur, A., Hernández-Quiroz, F., Ramírez-Del-Alto, S., García-Mena, J., and Hernández-Guerrero, C. (2019). Gut microbiota and predicted metabolic pathways in a sample of Mexican women affected by obesity and obesity plus metabolic syndrome. *International Journal of Molecular Sciences*, 20(2):1–18.

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., and Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258.

Combettes, P. L. and Müller, C. L. (2020). Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *Statistics in Biosciences*, pages 1–26.

Cottrell, M. T. and Kirchman, D. L. (2003). Contribution of major bacterial groups to bacterial biomass production (thymidine and leucine incorporation) in the Delaware estuary. *Limnology and Oceanography*, 48(1 I):168–178.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

de Menezes, A. B., Müller, C., Clipson, N., and Doyle, E. (2016). The soil microbiome at the Gi-FACE experiment responds to a moisture gradient but not to CO2 enrichment. *Microbiology (United Kingdom)*, 162(9):1572–1582.

Dillon, S. M., Frank, D. N., and Wilson, C. C. (2016). The gut microbiome and HIV-1 pathogenesis: A two-way street. *Aids*, 30(18):2737–2751.

Dubourg, G. (2016). Impact of HIV on the human gut microbiota : Challenges and perspectives. *Human Microbiome Journal*, 2:3–9.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828.

Fadeev, E., Salter, I., Schourup-Kristensen, V., Nöthig, E. M., Metfies, K., Engel, A., Piontek, J., Boetius, A., and Bienhold, C. (2018). Microbial communities in the east and west fram strait during sea ice melting season. *Frontiers in Marine Science*, 5(NOV):1–21.

Fierer, N. and Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *PNAS*, 103(3).

Gaines, B. R., Kim, J., and Zhou, H. (2018). Algorithms for Fitting the Constrained Lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8(November):2224.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Khabbazian, M., Kriebel, R., Rohe, K., and Ané, C. (2016). Fast and accurate detection of evolutionary shifts in ornstein–uhlenbeck models. *Methods in Ecology and Evolution*, 7(7):811–824.

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). 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.

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. *Nature*.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, 101:785–797.

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., Hingamp, P., Ogata, H., de Vargas, C., Lima-Mendez, G., Raes, J., Poulain, J., Jaillon, O., Wincker, P., Kandels-Lewis, S., Karsenti, E., Bork, P., and Acinas, S. G. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*.

Longhurst, A., Sathyendranath, S., Platt, T., and Caverhill, C. (1995). An estimate of global primary production in the ocean from satellite radiometer data. *Journal of Plankton Research*, 17(6):1245–1271.

Lozupone, C. and Knight, R. (2005). UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and environmental microbiology*, 71(12):8228–8235.

McDonald, D. e. a. (2018). American gut: an open platform for citizen science microbiome research. *mSystems*, 3(3).

Monaco, C. L., Gootenberg, D. B., Zhao, G., Handley, S. A., Ghebremichael, M. S., Lim, E. S., Lankowski, A., Baldridge, M. T., Wilen, C. B., Flagg, M., Norman, J. M., Keller, B. C., Luévano, J. M., Wang, D., Boum, Y., Martin, J. N., Hunt, P. W., Bangsberg, D. R., Siedner, M. J., Kwon, D. S., and Virgin, H. W. (2016). Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host and Microbe*, 19(3):311–322.

Moran, M. A. (2015). The global ocean microbiome. *Science*, 350(6266).

Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., Navas-Molina, J. A., Song, S. J., Metcalf, J. L., Hyde, E. R., Lladser, M., Dorrestein, P. C., and Knight, R. (2017). Balance Trees Reveal Microbial Niche Differentiation. *mSystems*, 2(1):e00162–16.

Nowak, P., Troseid, M., Avershina, E., Barqasho, B., Neogi, U., Holm, K., Hov, J. R., Noyan, K., Vesterbacka, J., Svärd, J., Rudi, K., and Sönnerborg, A. (2015). Gut microbiota diversity predicts immune status in HIV-1 infection. *Aids*, 29(18):2409–2418.

Oki, K., Toyama, M., Banno, T., Chonan, O., Benno, Y., and Watanabe, K. (2016). Comprehensive analysis of the fecal microbiota of healthy Japanese adults reveals a new bacterial lineage associated with a phenotype characterized by a high frequency of bowel movements and a lean body type. *BMC Microbiology*, pages 5–11.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Peters, B. A., Shapiro, J. A., Church, T. R., Miller, G., Trinh-Shevrin, C., Yuen, E., Friedlander, C., Hayes, R. B., and Ahn, J. (2018). A taxonomic signature of obesity in a large study of American adults. *Scientific Reports*, 8(1):1–13.

Ramirez, K. S., Leff, J. W., Barberán, A., Bates, S. T., Betley, J., Crowther, T. W., Kelly, E. F., Oldfield, E. E., Ashley Shaw, E., Steenbock, C., Bradford, M. A., Wall, D. H., and Fierer, N. (2014). Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B: Biological Sciences*, 281(1795).

Randolph, T. W., Zhao, S., Copeland, W., Hullar, M., and Shojaie, A. (2015). Kernel-Penalized Regression for Analysis of Microbiome Data. *ArXiv e-prints*.

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. (2018). Balances: a New Perspective for Microbiome Analysis. *mSystems*, 3(4):1–12.

Sandler, N. G., Wand, H., Roque, A., Law, M., Nason, M. C., Nixon, D. E., Pedersen, C., Ruxrungtham, K., Lewin, S. R., Emery, S., Neaton, J. D., Brenchley, J. M., Deeks, S. G., Sereti, I., and Douek, D. C. (2011). Plasma levels of soluble CD14 independently predict mortality in HIV infection. *Journal of Infectious Diseases*, 203(6):780–790.

Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14(8):1–14.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Stat.*, 10(2):1019–1040.

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:1–20.

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Acinas, S. G., Babin, M., Bork, P., Boss, E., Bowler, C., Cochrane, G., de Vargas, C., Follows, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Iudicone, D., Jaillon, O., Kandels, S., Karp-Boss, L., Karsenti, E., Lescot, M., Not, F., Ogata, H., Pesant, S., Poulton, N., Raes, J., Sardet, C., Sieracki, M., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Wincker, P., Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., Lombard, F., Ogata, H., Pesant, S., Sullivan, M. B., Wincker, P., and de Vargas, C. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology*, 18(8):428–445.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., D'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sieracki, M., Velayoudon, D., Bowler, C., De Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., and Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237):1–10.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484.

Ushey, K., Allaire, J., and Tang, Y. (2020). *reticulate: Interface to 'Python'*. R package version 1.16.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73(16):5261–5267.

Wang, T. and Zhao, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *The Annals of Applied Statistics*, 11(2):771–791.

Washburne, A. D., Silverman, J. D., Leff, J. W., Bennett, D. J., Darcy, J. L., Mukherjee, S., Fierer, N., and David, L. A. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5:e2969.

Washburne, A. D., Silverman, J. D., Morton, J. T., Becker, D. J., Crowley, D., Mukherjee, S., David, L. A., and Plowright, R. K. (2019). Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. *Ecological Monographs*, 89(2):1–27.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., and Lewis, J. D. (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science*, 334(6052):105–108.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063.

Xiao, J., Chen, L., Johnson, S., Yu, Y., Zhang, X., and Chen, J. (2018). Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Frontiers in Microbiology*, 9(JUN):1–14.

Yan, X. (2018). *Statistical Learning for Structural Patterns with Trees*. PhD thesis, Cornell University.

Yan, X. and Bien, J. (2020). Rare feature selection in high dimensions. *Journal of the American Statistical Association*, 0(just-accepted):1–30.

Yilmaz, P., Yarza, P., Rapp, J. Z., and Glöckner, F. O. (2016). Expanding the world of marine bacterial and archaeal clades. *Frontiers in Microbiology*, 6(JAN):1–29.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36.

Zeng, Q., Dong, Y., and An, S. (2016). Bacterial community responses to soils along a latitudinal and vegetation gradient on the Loess Plateau, China. *PLoS ONE*, 11(4):1–17.

Zhai, J., Kim, J., Knox, K. S., Twigg, H. L., Zhou, H., and Zhou, J. J. (2018). Variance Component Selection With Applications to Microbiome Taxonomic Data. *Front Microbiol*, 9:509.

Zhang, T., Shao, M.-F., and Ye, L. (2012). 454 pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *The ISME Journal*, 6(6):1137–1147.