DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY

Statistical Methods for Sequence-Based Microbial Community Assays

Citation

Schwager, Emma Holdrich. 2017. Statistical Methods for Sequence-Based Microbial Community Assays. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:42061500

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Statistical Methods for Sequence-based Microbial Community Assays

A dissertation presented

by

Emma Holdrich Schwager

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

August, 2017

©2017 Emma Holdrich Schwager
All rights reserved.

Dissertation Advisor: Curtis Huttenhower Emma Holdrich Schwager

Statistical Methods for Sequence-based Microbial Community Assays

Abstract

The human microbiome comprises the totality of micro-organisms residing in and on the human body. Of late, it has been the subject of much intensive research into how this microbial community is involved in diseases, either directly (as in the case of periodontitis or bacterial vaginosis) or indirectly (as in the case of obesity or type II diabetes). The implication of microbial involvement in these and other diseases suggests that the microbiome can be used as a therapeutic agent, because unlike the human genome, it is both measurable and plastic. Studies on the microbiome typically collect data using sequencing methodologies, such as 16S rRNA gene sequencing (which sequences a single gene universal among bacteria), whole metagenome shotgun sequencing (which sequences all DNA in a given sample), or metatranscriptomic sequencing (which sequences all RNA in a given sample). The abundance data generated by these technologies have unique characteristics which must be accounted for in any statistical analysis. Particularly, microbiome data tend to be highly zero-inflated, often having 80% or more zeros; high-dimensional, often having orders of magnitude more features than samples; and compositional because the abundances are constrained by the total number of sequencing reads in a sample. In this dissertation, I address these three challenges in two key areas of microbiome analysis: detecting microbial interactions and pre-computing study power. I develop a Bayesian correlation-detection method appropriate for relative abundance data to

explore ecological interactions between taxa. I use this method to elucidate the community ecological structure in the human microbiome at the species level, laying the foundation for further understanding of the behaviors of communities in the host and how they respond to perturbations. I also use simulation to provide a set of guidelines for practitioners performing pre-study power analysis in microbial epidemiology.

Contents

List of Figures

List of Tables

Acknowledgments

This thesis would have been impossible for me to finish without the unceasing support of many people.

To my husband Randall: thank you for listening to me, encouraging me, and sacrificing your time and energy to make this possible. I assuredly could not have done this without you.

To my son Bill: thank you for being your charming, loving self. Thank you for your joy in life that encourages me to look past the details of the moment and to see the wonders of the world God has made.

To my parents Martin and Sallie: thank you for your advice, your practical help, and your prayers. I have learned so much from you about how to love people in all circumstances.

To my parents-in-love Tim and Leslie: thank you for your prayers and your listening ears.

To my siblings Will, Gus, and Charlotte: thank you for your encouragement and for reminding me of how blessed my life really is.

Introduction

The human microbiome consists of the collection of micro-organisms that reside in and on the human body. While the microbiome does include bacteria, archaea, yeast, and viruses, most studies focus on bacteria. These organisms are found in almost every part of the human body: from the highly populous communities in the gastrointestinal tract [Finegold et al., 1983] and female urogenital tract [Hyman et al., 2005], to the sparsely populated skin [Noble, 1984] and upper respiratory tract [Charlson et al., 2011], and even sites considered until recently to be sterile, such as the placenta [Aagaard et al., 2014; Stout et al., 2013] and lower respiratory tract [Erb-Downward et al., 2011; Hilty et al., 2010]. The gut alone contains an estimated 0.2 kg of bacteria, which comprise approximately 3.8e13 cells [Sender et al., 2016]. These symbionts collectively encode a large pool of genetic diversity and flexibility: it has been estimated that our microbiomes contain over 100 times more genes than our own genome [Qin et al., 2010]. The large number and great diversity of microbes in and on the human body have lately been the focus of much intensive research regarding their effects on human health [Lloyd-Price et al., 2016].

Individual microbes have long been known to cause certain specific diseases such as whooping cough (*Bordetella pertussis*), tuberculosis (*Mycobaterium tuberculosis*), strep throat (*Streptococcus pyogenes*), and cholera (*Vibrio cholerae*). In the last 15 years, we have begun to understand more about the health effects of our symbiotic microbial communities and how they contribute to either health or disease. Microbial communities are necessary for proper immune system development [Hooper and

Gordon, 2001]. In the gut, they aid in extracting nutrients from food by breaking down fiber into easily-absorbed short-chain fatty acids [Hooper et al., 2002] and synthesizing vitamins [Hill, 1997; Hooper et al., 2002]. The resident microbial flora also provides a barrier to colonization by pathogens such as *Candida* [Boris et al., 1998; Sobel et al., 1981] and *Clostridium difficile* [Wilson, 1993]. Conversely, the microbiome has been implicated in a variety of diseases, both directly and indirectly. Some infections are characterized as poly-microbial, whereby a community of organisms (as opposed to a single pathogen) result in the disease: these include bacterial vaginosis [Hill, 1993], periodontitis [Socransky and Haffajee, 2005], and dental caries (cavities) [Jenkinson and Lamont, 2005]. Other diseases have a microbial component whose role in pathogenesis is unclear, such as eczema [Leyden et al., 1974], inflammatory bowel disease [Khor et al., 2011], and obesity [Turnbaugh et al., 2006]. The involvement of the microbiome in these various diseases suggests that we can use the microbiome to predict disease onset and to manipulate the disease course by guiding the microbiome to a healthy state. Unlike genetic risk factors, which can be mitigated but not changed, the microbiome is plastic, allowing interventions to be designed and targeted. Further, the microbiome is measurable at both the individual microbe and community levels.

For most of its history, microbial ecology focused on studying the subset of microbes that could be cultured and studied in isolation. The relatively recent introduction of sequencing technology allowed communities to be studied as a whole and in context. The first platform developed was 16S rRNA sequencing. Virtually all bacteria have the 16S ribosomal subunit RNA gene in their genomes, and its sequence is very stable across evolutionary history [Woese, 1987]. This allows nearly-universal primers to be used, which amplify this gene from all bacteria in a sample [Caporaso et al., 2011; Weisburg et al., 1991]. The slow evolution of this gene across the bacterial phylogenetic tree allow the identification of taxa from this gene alone on

the basis of sequence similarity [Pace, 2009; Woese, 1987]. Most often, the resulting sequences are clustered at 97% similarity to give taxonomic clumps known as operational taxonomic units (OTUs) [Konstantinidis and Tiedje, 2005]. 16S rRNA sequencing is inexpensive widely applicable because it relies solely on the sequencing of the one gene. For the same reason, it is also limited in its taxonomic resolution (it has a hard time distinguishing between recently diverged species or genera [Fox et al., 1992; Pace, 2009]), and is unable to identify non-bacterial organisms such as archaea, eukaryotes and viruses. The desire for higher resolution and more complete profiling combined with decreasing sequencing costs led to the sequencing of all the DNA in a sample, termed whole metagenome shotgun (WMS) sequencing [Tringe and Rubin, 2005]. Such sequencing allows species and even strains to be identified from the genes present in a sample, and can capture viruses and eukaryotes. Even more recently, metatranscriptomic sequencing techniques allow the RNA of a sample to be sequenced, resulting in a picture of the transcriptional activity of the microbial community [Frias-Lopez et al., 2008]. Regardless of which technique is employed, the resulting data take the form of a table with features (e.g., OTU, species, genus, transcript) in one dimension and samples in the other: each element in the table is the abundance of that feature in that sample.

Data resulting from microbial community studies pose unique statistical problems that must be accounted for when performing any analyses. Among these are zero-inflation, high-dimensionality, and compositionality. There is a high level of zero-inflation of microbial data–often more than half the features have at least 50% zeros, and it is not uncommon for features to have 80% zeros. For such sparse features, standard continuous distributions such as normal or gamma, or discrete models such as negative binomial require the addition of a zero-inflation parameter [Paulson et al., 2013; Xu et al., 2015]. Microbial data are also high-dimensional, especially relative to the number of samples. Quite commonly, the number of features measured is orders

of magnitude larger than the number of samples collected [Kurtz et al., 2015]. Often zero-inflation and high-dimensionality are addressed simultaneously by removing features with too many zeros before analysis. Lastly, microbial data are compositional because any sequencing analysis is constrained by the total number of reads sequenced (the read depth), yielding a multinomial sampling scheme [La Rosa et al., 2012]. The resulting sum-constraint leads to a number of analysis complications, including apparent correlations between features and the necessity of modelling the entire data generation mechanism in any multivariate scenario.

In this dissertation, I address these statistical challenges in two key areas of microbial community analysis: microbial interactions and power analysis. Microbial interactions are key to understanding the behavior of communities in the host and their response to challenges such as antibiotics or dietary changes. From sequencing data, these can be inferred from correlations among taxa. I develop a Bayesian model to identify such correlations and apply it to a species-level dataset to study the ecology of microbial interactions at the species level. In microbial epidemiology, pre-study power analysis is crucial to accurately estimating sample size needs or detection abilities given a particular study design. Guidelines for such analysis have not been fully developed for microbial community studies. I use simulation to evaluate existing power analysis methods along with simplistic approximations to develop guidelines for such analysis.

1

A Bayesian Method for Detecting Pairwise Associations in Compositional Data

Abstract

Scientific abstract

Compositional data consist of vectors of proportions normalized to a constant sum from an unconstrained basis. The sum constraint makes inference on correlations between basis features challenging due to the information loss from normalization. However, such correlations are of long-standing interest in fields including ecology. We propose a novel Bayesian framework (BAnOCC: Bayesian Analysis of Compositional Covariance) to estimate a sparse precision matrix through a LASSO prior. The resulting posterior, generated by MCMC sampling, allows uncertainty quantification of any function of the precision matrix, including the correlation matrix. We also use a first-order Taylor expansion to approximate the transformation from the basis to the composition in order to investigate what characteristics of the basis can make the correlations more or less difficult to infer. On simulated datasets, we show that BAnOCC infers the true network as well as previous methods while offering the advantage of posterior inference. Larger and more realistic simulated datasets further showed that BAnOCC performs well as measured by type I and type II error rates. Finally, we apply BAnOCC to a microbial ecology dataset from the Human Micro-

biome Project, which in addition to reproducing established ecological results revealed unique, competition-based roles for Proteobacteria in multiple distinct habitats.

Lay summary

Data from many fields are available primarily in the form of proportions, also referred to as compositions, which impose mathematical constraints on identifying interactions among components in the underlying systems. In particular, correlations cannot be calculated directly from proportions or from count data that give rise to them. Methods that work around this difficulty generally do so by imposing strong assumptions about the distribution of underlying data or associated correlations, and these in turn often prevent quantifying uncertainty in the resulting estimates of correlation. We developed a statistical model (BAnOCC: Bayesian Analysis of Compositional Covariance) that both estimates correlations between counts or proportions and provides a posterior distribution for each correlation that quantifies how uncertain the estimate is. BAnOCC does well at controlling the number of false positives in simulated data and can be practically applied to a wide range of proportional data types.

1.1 Introduction

A long-standing goal of applied statistics in many fields has been identifying features associated significantly by a measure such as correlation [Pearson, 1896; Spearman, 1904]. When the features to be associated form a composition, inference of the correlation matrix is subject to the well-known problem of spurious correlation [Aitchison, 1981; Chayes, 1960; Chayes and Kruskal, 1966; Pearson, 1897]. Compositional data in particular are vectors of proportions that sum to a fixed constant (typically one); they are usually thought of as the result of sum-normalizing an unobserved (or unrecorded) and unconstrained basis, following the terminology of [Aitchison, 1981]. The resulting sum-constraint of the compositional data means that any pairwise correlation mea-

sured using such data can be non-zero even if all the pairwise correlations in the basis are zero, a phenomenon called spurious correlation [Pearson, 1897]. The fact that all the features sum to one also makes the basis correlation matrix (that is, the correlation matrix of the unnormalized counts) non-identifiable without untestable, though perhaps not unreasonable, assumptions [Ban et al., 2015; Fang et al., 2015; Faust, K. and Sathirapongsasuti, F. et al., 2012; Friedman and Alm, 2012]. Any method thus offers at best a partial reconstruction of the basis correlation matrix, and the interest in characterizing such correlations in fields from geology to ecology has led to a variety of approaches.

In the context of microbial ecology, several methods have been proposed to identify significant ecological relationships from compositions; virtually all rely on some form of sparsity assumption and infer quantities relating to the log-transformed basis (hereafter referred to as the log-basis). The only technique that does not rely on a sparsity assumption is ReBoot [Faust, K. and Sathirapongsasuti, F. et al., 2012], which estimates a "compositionally-corrected" correlation matrix using a permutation-based method. Friedman and Alm [Friedman and Alm, 2012] proposed SparCC, which estimates the log-basis correlation matrix under the assumption that the correlations are on average small in magnitude. Fang et al. [Fang et al., 2015] noted that the resulting estimate is not guaranteed to be positive definite or that the elements will lie inside [-1, 1] and proposed CCLasso to estimate the log-transformed basis correlation matrix using a LASSO penalty on the off-diagonal elements of the variance-covariance matrix. Ban et al. [Ban et al., 2015] similarly proposed REBACCA to estimate the log-basis correlation matrix; they use the same LASSO penalty function but a different likelihood function. Kurtz et al. [Kurtz et al., 2015] proposed SPIEC-EASI to estimate the log-basis precision matrix when the number of features is large by using sparse graph estimation techniques.

These approaches have difficulty quantifying uncertainty in the estimates, cannot

incorporate uncertainty from the choice of tuning parameter, and are not flexible in the quantities they estimate. Friedman and Alm [Friedman and Alm, 2012] proposed an inferential procedure based on the bootstrap, but offered no theoretical justification. Fang et al. [Fang et al., 2015] and Kurtz et al. [Kurtz et al., 2015] focused solely on estimation, while Ban et al. [Ban et al., 2015] used a subsampling method from Shah and Samworth [Shah and Samworth, 2013] to stabilize the selection error rate. The LASSO-based methods [Ban et al., 2015; Fang et al., 2015; Kurtz et al., 2015] typically choose a shrinkage parameter and subsequently infer the log-basis covariance or precision matrix. Friedman and Alm [Friedman and Alm, 2012], Fang et al. [Fang et al., 2015], and Ban et al. [Ban et al., 2015] all use the log-basis covariance matrix for network construction, while Kurtz et al. [Kurtz et al., 2015] use the log-basis precision matrix. This means that investigators typically must choose whether a precision or correlation matrix is best, and often use the resulting estimate with little guidance as to its uncertainty.

We address these issues by providing a flexible, fully Bayesian approach to identify correlations in compositional data. It is able to quantify uncertainty through the associated posterior and estimates both the log-basis correlation and precision matrix by modeling the composition directly. The graphical LASSO prior of [Wang, 2012] is used to estimate a sparse log-basis precision matrix (and hence a sparse log-basis correlation matrix) through a LASSO penalty, mitigating the non-identifiable nature of the basis correlation matrix. We have implemented the resulting method as BAnOCC (Bayesian Analysis of Compositional Covariance). In this study, we also use a first-order Taylor expansion to approximate the compositional covariance as a function of the mean and variance of the basis. While not necessary to the development of our method, this expansion helps us explore the situations in which a naïve approach (ignoring the sum-constraint) might work. This approximation shows not only that the spurious correlation between two features can take any value in $[-1, 1]$ even if none

of the features are correlated in the basis, but also that both the basis variance and basis means control the magnitude and direction of the spurious correlation. Thus, we provide a novel characterization of the surprisingly broad circumstances under which compositionality can impede straightforward identification of the correlation matrix, and we provide the BAnOCC model to overcome this in datasets where it is possible.

1.2 Methods

Per-subject basis and composition notation

The model assumes that a single subject's composition, $\boldsymbol{C}_i = (C_{i,1}, \ldots, C_{i,p})^T$, is generated by the normalization of that subject's unobserved basis, $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})^T$. That is, $\boldsymbol{C}_i = \frac{\boldsymbol{X}_i}{\sum_{j=1}^{p} X_{i,j}}$. We also assume that the bases for all subjects are independent and identically distributed (iid); this implies that the compositions are iid as well because the transformation is per-subject.

Feature correlations and covariances in composition and basis

We also introduce notation for the covariance and correlation among the features. The basis covariance matrix is denoted by $\boldsymbol{\Sigma}_X = [\sigma_{X,jk}]$, to be inferred from $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n$. Similarly, the covariance matrix of the composition is denoted by $\boldsymbol{\Sigma}_C = [\sigma_{C,jk}]$. To construct the network of feature interactions, the relevant null hypotheses (one for each feature pair j and k) are that features j and k have a covariance of zero ($\sigma_{X,jk} = 0$); this is equivalent to testing if they are uncorrelated ($\rho_{X,jk} = 0$). We then define the basis and compositional correlation matrices as $\boldsymbol{R}_X = [\rho_{X,jk}]$ and $\boldsymbol{R}_C = [\rho_{C,jk}]$, respectively.

1.2.1 BAnOCC: Bayesian Analysis of Compositional Covariance

BAnOCC assumes that the basis follows a log-normal distribution and that the basis correlation matrix is sparse; it is parametrized with the log-basis precision matrix and the log-basis mean (**Figure 1.1**). Posteriors for the parameters of the model (and thus functions of them which are of interest) are inferred using MCMC sampling. This fully Bayesian treatment of the problem gives several advantages: a full posterior distribution to quantify the uncertainty in the estimates, the ability to place a prior on the sparsity parameter, and estimates of any function of the log-basis precision matrix, including the log-transformed basis covariance and correlation matrices.

BAnOCC models the basis counts using a log-normal distribution with parameters based on the moments of the log-basis: $\boldsymbol{X}_i \overset{iid}{\sim} \mathcal{LN}(\boldsymbol{m}, \boldsymbol{S})$, such that $\boldsymbol{m} = \mathbb{E}(\log\{\boldsymbol{X}\})$ and $\boldsymbol{S} = \mathrm{Var}(\log\{\boldsymbol{X}\})$. This continuous approximation of the count data that form the basis is expected to perform well when the underlying counts have a large dynamic range. In ecology, for example, the log-normal distribution is used to model the (discrete) abundance across species [Magurran and Henderson, 2003; Preston, 1948]. In microbial ecology specifically, the logistic normal is sometimes assumed to be the generating distribution of the composition [Ban et al., 2015; Kurtz et al., 2015]; further, the (discrete) read counts are often simulated using a log-normal distribution [Koren et al., 2013; Paulson et al., 2013]. The log-normal distribution also allows the totals to be easily integrated out of the likelihood.

Parametrization of the likelihood

The likelihood is parametrized by the log-basis precision matrix $\boldsymbol{O} = \boldsymbol{S}^{-1}$ and the log-basis mean $\boldsymbol{m}$, and other parameters of interest like the log-basis covariance matrix $\boldsymbol{S}$ are sampled as transformations of these. By parametrizing using $\boldsymbol{O}$, we are able to leverage a graphical LASSO prior to enforce sparsity on $\boldsymbol{O}$ and by extension $\boldsymbol{S}$. Conveniently, the assumption of the log-normal distribution obviates the need to

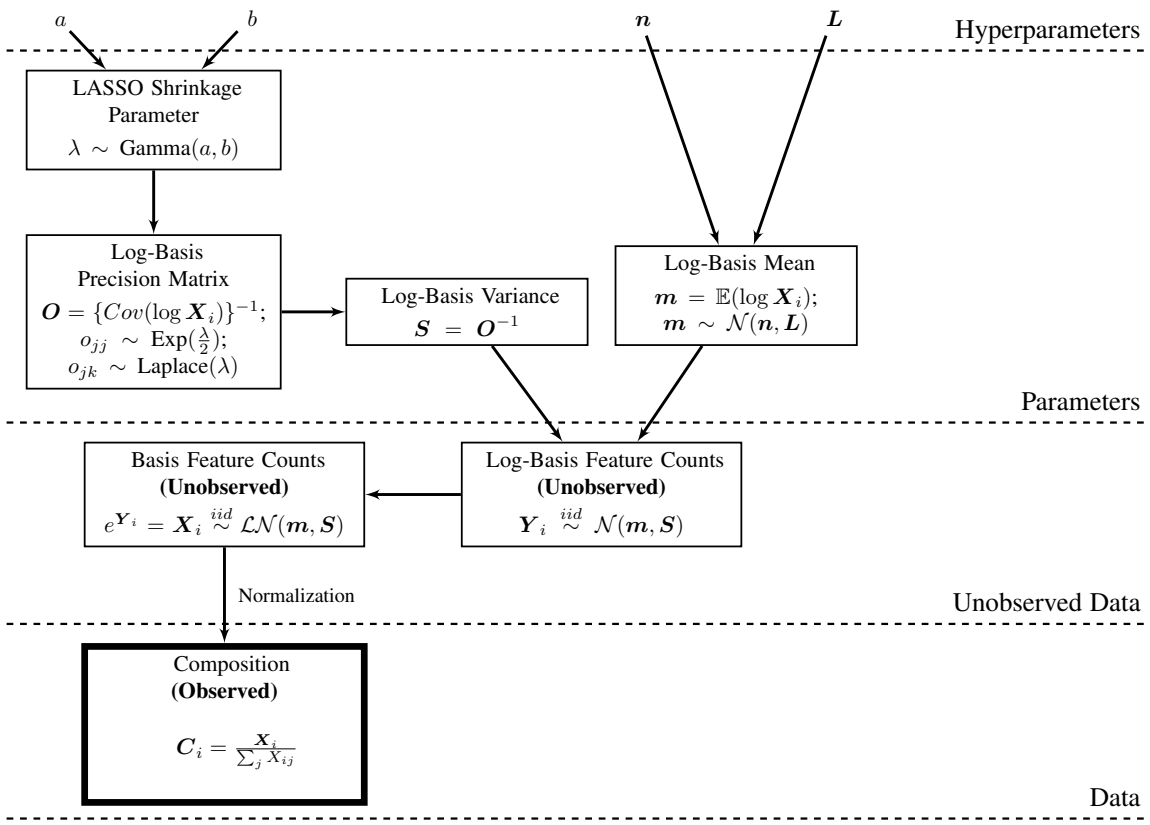

Figure 1.1: **BAnOCC infers log-basis correlation and precision matrices from compositions by modeling unobserved basis counts.** In the BAnOCC model, the observed compositions, $\boldsymbol{C}_i$, are derived by normalizing the basis counts $\boldsymbol{X}_i$. The BAnOCC model assumes that the $\boldsymbol{X}_i$ follow a log-normal distribution, parametrized by the log-basis mean $\boldsymbol{m}$ and covariance $\boldsymbol{S}$. It places a normal prior on $\boldsymbol{m}$, the GLASSO prior of [Wang, 2012] on the log-basis precision matrix $\boldsymbol{O}$, and a hyperprior on the GLASSO shrinkage parameter λ (**Section 1.2**).

sample the basis covariance to determine the existence and direction of an association between two features in the basis. This results because when some element of $\boldsymbol{S}$, s_{jk}, is zero, then the corresponding element of $\boldsymbol{\Sigma}_X$, $\sigma_{X,jk} \propto e^{s_{jk}} - 1$ will also be zero; further, the non-zero elements of $\boldsymbol{S}$ and $\boldsymbol{\Sigma}_X$ will have the same sign (though not the same magnitude).

Under the log-normal assumption, the complete likelihood of the observed composition $\boldsymbol{c}_i$ and the latent total $t_i = \sum_{j=1}^{p} x_{i,j}$ is given by

$$\mathcal{L}\left(\boldsymbol{m}, \boldsymbol{O} \,|\boldsymbol{c}_i, t_i\right) = \frac{\exp\left[-\frac{1}{2}\left\{\log\left(\boldsymbol{c}_i t_i\right) - \boldsymbol{m}\right\}^T \boldsymbol{O}\left\{\log\left(\boldsymbol{c}_i t_i\right) - \boldsymbol{m}\right\}\right]}{\left(t_i\right)(2\pi)^{p/2}\left|\boldsymbol{O}\right|^{-1/2}\prod_{j=1}^{p} c_{i,j}}, \tag{1.1}$$

where $\boldsymbol{c}_i = \left(c_{i,1}, \ldots, c_{i,p-1}, 1 - \sum_{j=1}^{p-1} c_{i,j}\right)$. A detailed derivation can be found in **Section S1.3**. Fitting this likelihood directly is computationally expensive, as the presence of the latent totals necessitates exploring a space whose dimension depends on both n and p. However, (1.1) factors into two portions: a part dependent on the compositions $\boldsymbol{c}_i$, and the kernel of a log-normal distribution for the totals $t_i = \sum_{j=1}^{p} x_{i,j}$ with parameters $m_i^* = \mathbf{1}^T\boldsymbol{O}(\boldsymbol{m} - \log\{\boldsymbol{c}_i\})s^{2^*}$ and $s^{2^*} = \frac{1}{\mathbf{1}^T\boldsymbol{O}\mathbf{1}}$ (where $\mathbf{1}$ is a vector of 1's). Integrating over the totals in (1.1) gives the more computationally tractable marginal likelihood

$$\mathcal{L}\left(\boldsymbol{m}, \boldsymbol{O} \,|\boldsymbol{c}_i\right) = \left|\boldsymbol{O}\right|^{1/2} \frac{\exp\left\{-\frac{1}{2}\left(\boldsymbol{m} - \log \boldsymbol{c}_i\right)^T \boldsymbol{O}\left(\boldsymbol{m} - \log \boldsymbol{c}_i\right) - \frac{(m_i^*)^2}{s^{2^*}}\right\}(2\pi)^{1/2}(s^{2^*})^{1/2}}{(2\pi)^{p/2}\prod_{j=1}^{p} c_{ij}}.$$

Prior distributions

In order to mitigate the non-identifiability of the precision matrix $\boldsymbol{O}$, BAnOCC uses a shrinkage prior to conservatively estimate the sparsest $\boldsymbol{O}$ consistent with the observed relative abundance data. This is the graphical LASSO prior of [Wang, 2012]:

$$p(\boldsymbol{O}|\lambda) = C^{-1}\prod_{j=1}^{p}\mathrm{Exp}\left(o_{jj}\left|\frac{\lambda}{2}\right.\right)\left\{\prod_{k=i+1}^{p}\mathrm{Laplace}\left(o_{jk}|\lambda\right)\right\}1_{\boldsymbol{O}\in M^+},$$

where $1_{\boldsymbol{O}\in M^+}$ is an indicator function that $\boldsymbol{O}$ is positive definite, $\mathrm{Exp}(x|\lambda)$ has the exponential density of the form $p(x) = \lambda e^{-\lambda x}1_{x>0}$, and $\mathrm{Laplace}(x|\lambda)$ has the Laplace

density of the form $p(x) = \frac{\lambda}{2}e^{-\lambda|x|}$. In comparison to variable selection priors such as spike-and-slab [Mitchell and Beauchamp, 1988], the graphical LASSO prior is more scalable to high dimensions at the cost of being unable to generate estimates that are exactly zero [Mallick and Yi, 2013]. We deal with this by using the resulting posterior samples to conclude whether a correlation is likely to be zero or not. The choice of λ is key to the degree of shrinkage imposed by this prior. We placed a gamma prior on λ in lieu of specifying it *a priori*; this is possible because [Wang, 2012] showed that the normalizing constant C does not depend on λ. The prior for $\boldsymbol{m}$ is the conditionally-conjugate normal prior $\mathcal{N}(\boldsymbol{n}, \boldsymbol{L})$ with mean $\boldsymbol{n}$ and covariance matrix $\boldsymbol{L}$. Hyperparameter choice for the two priors (on $\boldsymbol{m}$ and λ) is discussed in more detail below.

Implementation and inference

BAnOCC samples the posterior using Stan's C++ implementation and R interface [Stan Development Team, 2014]. Multiple quantities can be estimated from BAnOCC, including the log-basis precision, covariance, and correlation matrices. In our simulations and application, we estimated the log-basis correlation $\boldsymbol{R}_{\log \boldsymbol{X}}$ because it is interpretable and nicely scaled; we used the posterior median as the point estimate and the 95% credible intervals for w_{jk} to determine whether the correlation between features j and k was non-zero.

1.2.2 Choosing hyperparameters

The interpretation of the prior parameters on $\boldsymbol{m}$ is relatively straightforward, while that of the shrinkage parameter λ is less clear. Because log-basis means $\boldsymbol{m}$ have a normal distribution, $e^{\boldsymbol{m}}$ represents the median basis counts, which conveniently have a log-normal distribution with parameters $\boldsymbol{n}$ and $\boldsymbol{L}$. Therefore, we could parametrize the prior on $\boldsymbol{m}$ by the expected median basis count $\boldsymbol{n}_{LN} = \exp\{\boldsymbol{n} + 0.5diag(\boldsymbol{L})\}$ and uncertainty of the median basis count $\boldsymbol{L}_{LN} = \boldsymbol{n}_{LN}\boldsymbol{n}_{LN}^T(e^{\boldsymbol{L}} - 1)$. The prior

on the shrinkage parameter λ has a shape parameter a that determines how much prior probability mass is placed on λ values close to zero, and a rate parameter b that determines how the probability mass is spread across the entire domain. In particular, $a \leq 1$ forces an asymptote at zero, while $a > 1$ does not.

When little or no prior data is available, weakly informative priors can be used. Any prior on λ should have high probability mass close to zero and so should have $a \leq 1$. Larger values of a will "soften" the asymptotic behavior at zero (**Figure S1.5**). The value of the rate parameter b should be chosen to so that most prior probability mass is on sensible values for λ. The degree of shrinkage implied by λ does not appreciably change for $\lambda > 1$ (**Figure S1.6**), and so a b of around 5 will give a reasonable uninformative prior distribution for λ. For the log-basis means, $\boldsymbol{m} \sim \mathcal{N}(\mathbf{0}, l\boldsymbol{I})$ can be used, with l a large value such as 100. An overlarge value for l can make computation less efficient and put prior mass on grossly implausible values of $e^{\boldsymbol{m}}$, so an l of 500 or less is reasonable.

Prior subject-matter information can be incorporated into the priors for both λ and $\boldsymbol{m}$, but most easily into the prior on $\boldsymbol{m}$. If the data have few features, a smaller shape hyperparameter a should be employed to upweight values of λ that yield high shrinkage. The implied prior on the median basis counts $e^{\boldsymbol{m}}$ could be sampled to provide an empirical distribution of the total counts $\sum_{j=1}^{p} e^{m_j}$; this could be assessed for gross deviations from what might be considered reasonable, or agreement with known ranges if such data are available.

1.2.3 Software

The implementation of BAnOCC is publicly available with source code, documentation, tutorial data, and as an R/Bioconductor package at `http://huttenhower.sph.harvard.edu/banocc`.

1.3 Results

1.3.1 Basis mean and covariance determine spurious correlation sign and magnitude

We first aimed to identify what characteristics of compositional data impede or facilitate the accurate estimation of basis correlation matrices in general. Such characteristics should delineate when BAnOCC or any other technique for estimating the basis correlation would perform well. A first-order Taylor expansion approximates the compositional covariance as a function of the basis mean and covariance. Because the compositional correlation is a function of the compositional covariance, the resulting approximation also explains how the correlation behaves. Letting $\boldsymbol{X}$ represent the basis and $\boldsymbol{C}$ the composition, with the basis mean denoted by $\boldsymbol{\mu}_X = (\mu_{X,j})^T$ and the approximate average proportions by $\boldsymbol{\omega} = \left(\frac{\mu_{X,1}}{\sum_{j=1}^p \mu_{X,j}}, \ldots, \frac{\mu_{X,p}}{\sum_{j=1}^p \mu_{X,j}}\right)^T$, the Taylor expansion yields

$$\boldsymbol{\Sigma}_C \approx \left(\frac{1}{\sum_{j=1}^p \mu_{X,j}}\right)^2 (\boldsymbol{I} - \boldsymbol{\omega}\mathbf{1}^T)\boldsymbol{\Sigma}_X(\boldsymbol{I} - \boldsymbol{\omega}\mathbf{1}^T)^T. \tag{1.2}$$

Here $\boldsymbol{I}$ is the $p \times p$ identity matrix, and $\mathbf{1}$ is a p-dimensional vector of 1's. Eq (1.2) allows us to approximate the behavior of the compositional covariance from the basis parameters that generate it.

Spurious correlation can take any value between -1 and 1

Surprisingly, when no features in the basis are correlated, the spurious correlation can take any value in $[-1, 1]$ depending on the properties of the basis (**Figure 1.2**). This is suggested by considering Eq (1.2) when $\sigma_{X,jk} = 0$ for all $j \neq k$, then

$$\sigma_{C,jk} \approx \left(\frac{1}{\sum_{l=1}^p \mu_{X,l}}\right)^2 \left[\omega_j\omega_k \sum_l \sigma_{X,ll} - \omega_j\sigma_{X,kk} - \omega_k\sigma_{X,jj}\right] \text{ for } j \neq k. \tag{1.3}$$

The weights ω_j and the variances $\sigma_{X,ll}$ can be configured arbitrarily to force $\sigma_{C,jk}$ either to the extreme positive or extreme negative end of the spectrum. In particular,

we see three types of strong spurious correlations (**Figure 1.2B-D**): "negative dominant", "positive dominant", and "negative mixed". These three types of correlations are thus representative of a range of expected real-world behaviors, and we included them in subsequent simulation studies of BAnOCC and previous models.

"Negative dominant" spurious correlation (**Figure 1.2B**) occurs when features j and k in the basis have (1) high mean and (2) high variability compared to the remaining ($l \neq j, k$) features. Intuitively, the remaining features must contribute minimally to the total mean or total variance in the basis. When normalized, the sum-constraint thus forces a negative correlation between features j and k because they behave as if they were the only two features in the composition.

In the "positive dominant" spurious correlation type (**Figure 1.2C**), features j and k in the basis have (1) small variability and (2) high mean relative to the remaining ($l \neq j, k$) features. The positive correlation in the composition results because the variability in the sum of the remaining feature abundances causes the compositions for features j and k to be shrunk or stretched in the same direction when the data are normalized.

Finally, "negative mixed" spurious correlations are the result of "positive dominant" type bases where feature k and the remaining features have switched roles (**Figure 1.2D**). After normalization, the variability in feature k forces feature j to move in the opposite direction to accommodate the remaining features.

Extending and improving current assumptions about compositional correlation

Eq (1.3) also offers an alternative explanation for the negative covariance between features in a Dirichlet distribution. A Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_p$ results when each feature in the basis is independent and has a $Gamma(\alpha_j, \beta)$. The mean and variance of a Gamma distribution are $\frac{\alpha}{\beta}$ and $\frac{\alpha}{\beta^2}$, respectively, implying that in the basis, a feature with high mean will also have high

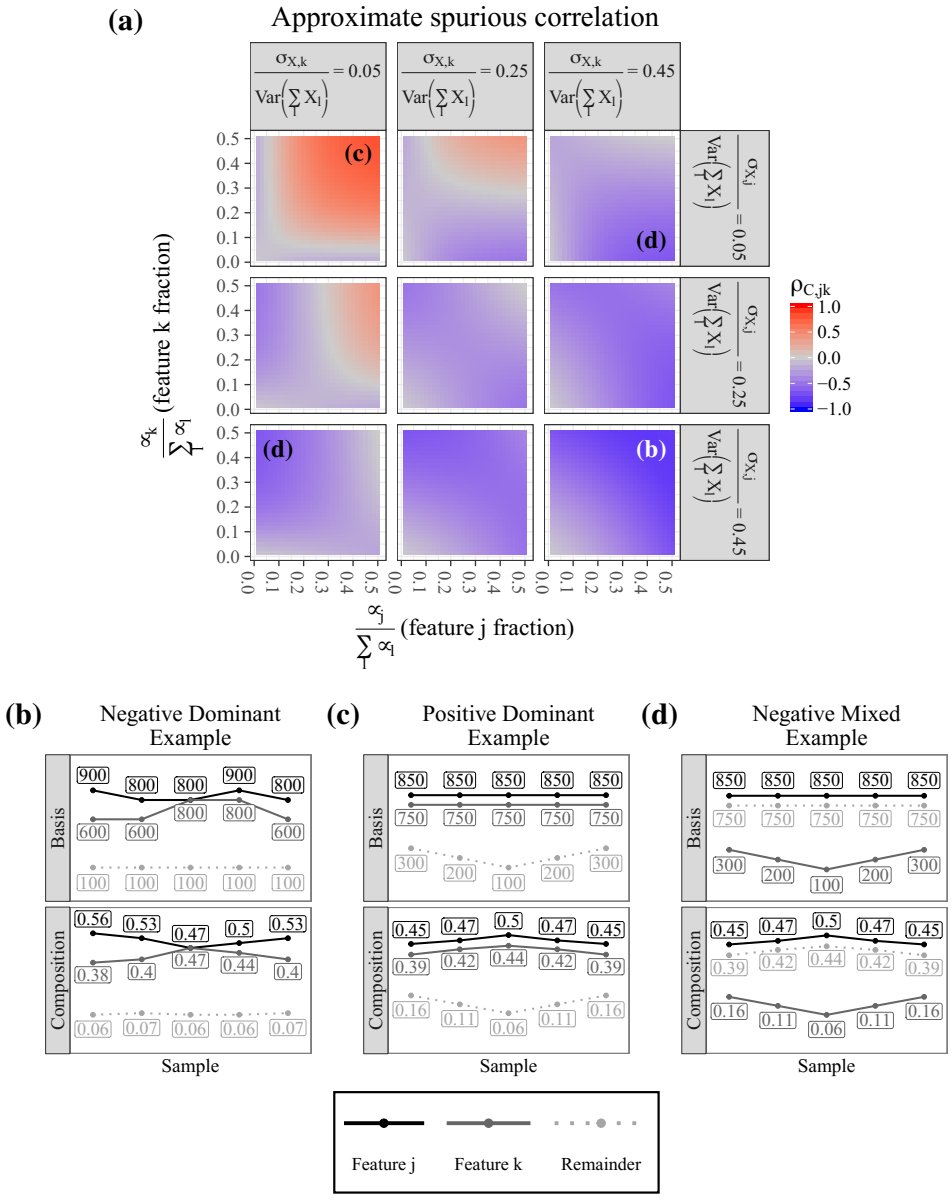

Figure 1.2: **Spurious correlation is not constrained as a function of feature count, mean, and variance.** (**A**) The approximate compositional correlation (based on Eq (1.3)) between features j and k when $\sigma_{X,jk} = 0$, as a function of the proportion of the total mean and proportion of total variability they contribute. (**B**)-(**D**) Examples of compositions that display positive (**B**) and negative (**C**)-(**D**) compositional correlations; in each, the top panel shows the correlation of the basis abundances across samples, while the bottom panel shows the correlation of the relative abundances across samples. The spurious correlation can be positive or negative, and of arbitrary magnitude, depending on the characteristics of the basis.

variance, and vice versa. This captures "negative dominant" correlations well, but fails to capture "positive dominant" or "negative mixed" correlations, which result when at least one feature has high mean but *low* variance in the basis.

Eq (1.2) and (1.3) further suggest that the overall effect of normalization on the correlation estimate as the number of features p increases depends on the characteristics of $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$. In ecological applications, it is often assumed that if p is large and the compositional means are similar across the p features, then the correlation estimates based on the composition and basis are not likely to be very different [Ban et al., 2015; Friedman and Alm, 2012]. Part of the appeal of this reasoning is that it does not rely on information about the unobserved basis. Expanding Eq (1.2), we can see that $\boldsymbol{\Sigma}_C \propto \boldsymbol{\Sigma}_X - \boldsymbol{\omega}\mathbf{1}^T\boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_X\mathbf{1}\boldsymbol{\omega}^T + \boldsymbol{\omega}\mathbf{1}^T\boldsymbol{\Sigma}_X\mathbf{1}\boldsymbol{\omega}^T$. If the means are very similar to each other, this affects only the weights $\boldsymbol{\omega}$ given to the offset $\boldsymbol{\omega}\mathbf{1}^T\boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_X\mathbf{1}\boldsymbol{\omega}^T + \boldsymbol{\omega}\mathbf{1}^T\boldsymbol{\Sigma}_X\mathbf{1}\boldsymbol{\omega}^T$. Small weights render the offset negligible only in the case where the unobserved basis variance $\boldsymbol{\Sigma}_X$ is not too large: the behavior of the offset as the number of features increases depends on the similarity of the means (through $\boldsymbol{\omega}$) *and* on the basis variances of the additional features (through $\boldsymbol{\Sigma}_X$).

Thus when analyzing compositional data, one cannot know with certainty in which data the correlations are strongly affected by the normalization, much less the magnitude and direction of the change in correlation structure induced by normalizing. The information loss due to normalization implies that $\boldsymbol{\Sigma}_X$ is non-identifiable without assumptions about its structure. However, knowing how the unobserved basis affects the spurious correlation allows simulation of datasets that have specific types of spurious correlation for testing the performance of estimation methods in these cases.

1.3.2 Simulation studies

Data Generation Methods

Using the information from this theoretical analysis, we tested BAnOCC on two types of datasets. The first comprised small datasets generated using the model itself but designed to be challenging by incorporating negative dominant correlations. Second, we also simulated larger, more realistic datasets using an independent model specific to microbial community structure, sparseDOSSA [Ren et al., 2016].

For the former, four small datasets with 1,000 samples and nine features each were generated according to four scenarios. The "simple" scenario had no basis correlations and no negative dominant correlation; the "high spurious" scenario had no basis correlations but the presence of a negative dominant correlation; the "retained spike" scenario had several basis correlations and no negative dominant correlation; and the "reversed spike" scenario had several basis correlations and a negative dominant correlation between two features that are positively correlated in the basis (see details in **Section S1.4**). On these data, we used hyperparameters $n_j = 0$, $\boldsymbol{L} = 1000\boldsymbol{I}$, $a = 0.5$ and $b = 5$ (**Figure S1.7**).

Realistic data were generated using the SparseDOSSA model [Ren et al., 2016], which generates each feature from a zero-inflated, truncated log-normal distribution with subsequent rounding and estimates the feature-specific parameters by fitting to a given real-world template dataset. We induced correlations between features by using a multivariate distribution with a log-basis correlation that had off-diagonal elements set to one of four different correlation strengths ($\{-0.7, -0.3, 0.3, 0.7\}$). To ensure that strong compositional effects were present, we used a template with low-diversity community structure [The Human Microbiome Consortium, 2012] with 14 pseudomicrobial features. The correlations were set so that the non-zero elements of the log-basis precision matrix and the log-basis covariance matrix would be the same; we used seven correlations (see details in **Section S1.4**). We used hyperparameters

$a = 0.5$, $b = 5$, $n_j = 3$, and $\boldsymbol{L} = 30\boldsymbol{I}$ (**Figure S1.8**).

BAnOCC and CCLasso Perform Comparably in Difficult Scenarios

Using our first set of simulated data for evaluation, we compared the estimation and inference from BAnOCC with that from CCLasso [Fang et al., 2015], a frequentist LASSO-based method that chooses the shrinkage parameter using K-fold cross validation (**Figure 1.3**). BAnOCC had much lower false positive rates than CCLasso, resulting from the model's ability to use the posterior distribution to account for estimate uncertainty while CCLasso, being LASSO-based, used a non-zero point estimate to determine significance of an effect.

BAnOCC and CCLasso both estimate the log-basis correlation matrix accurately, and both are a substantial improvement on a naïve approach (row 2 of **Figure 1.3**). In particular, both BAnOCC and CCLasso have much lower false positive rates than Pearson correlation. Over all the null associations, Pearson correlation had a staggering false positive rate of 82%; CCLasso had almost 14% false positives as a result of many small but non-zero estimates; BAnOCC, because it uses the posterior credible intervals to evaluate uncertainty, had a false positive rate of about 3%. BAnOCC cannot estimate the log-basis correlations w_{jk} to be exactly zero because of the continuous prior, but the null associations whose 95% credible intervals cover zero have very small estimates (all are less than 0.15, 75% are less than 0.05).

The association between features 1 and 5 in the "reversed spike" dataset was difficult for both BAnOCC and CCLasso. Both gave a small, negative estimate (-0.001 for BAnOCC and -0.113 for CCLasso). BAnOCC displays a slight bias toward positive correlations instead of the moderate negative correlation that was present in the underlying basis, as shown by several false positive associations in this dataset. This behavior is common among many methods, including SparCC and SPIEC-EASI (**Figure S1.9**). It results from the fact that when a negative-dominant structure is

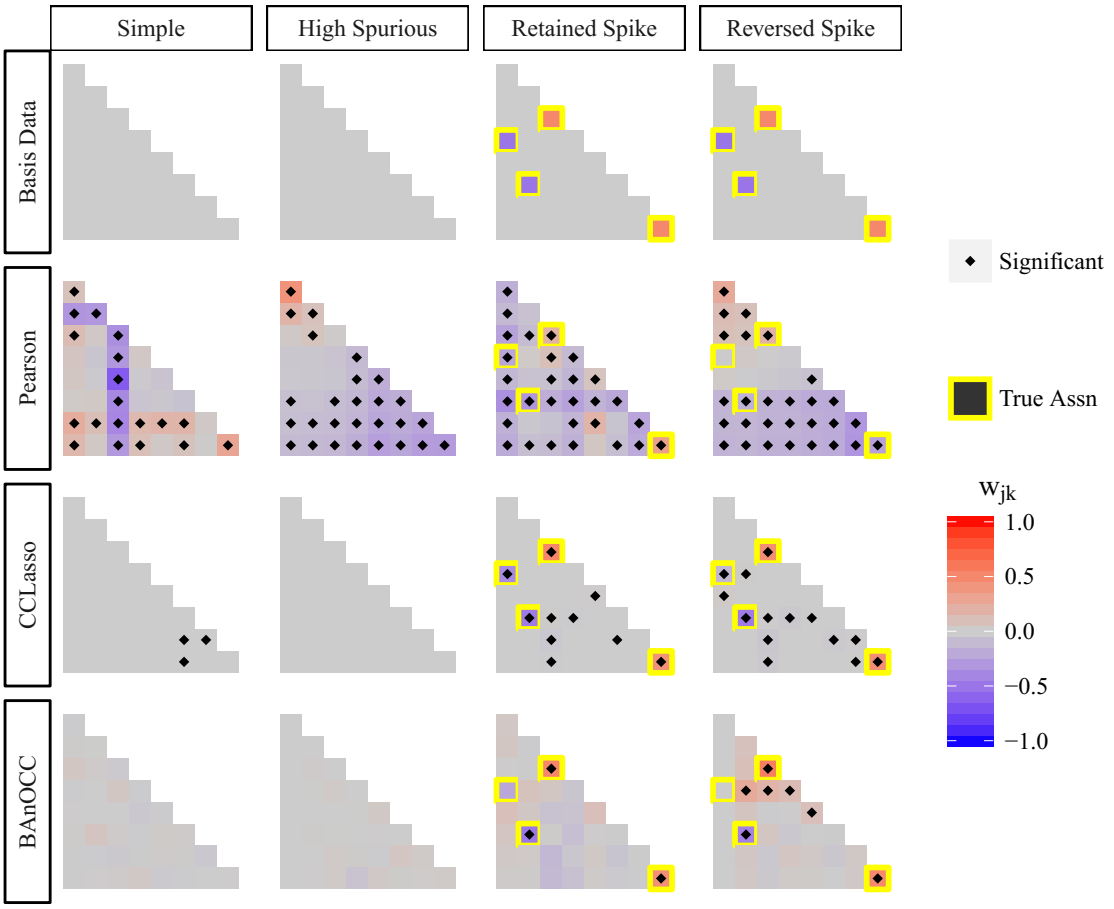

Figure 1.3: **BAnOCC infers the correct basis correlation matrix in four scenarios simulated to be challenging.** Each column represents one of the four datasets simulated to evaluate methods for identification of correlations from compositional data: "simple", with no basis correlations and no negative dominant correlation; "high spurious", with no basis correlations and the presence of a negative dominant correlation; "retained spike" with several basis correlations and no negative dominant correlation; and "reversed spike" with several basis correlations and a negative dominant correlation between two positively correlated features. The top row shows the true correlation matrix. The second row shows the uncorrected compositional correlations $\widehat{\boldsymbol{R}_C}$ as estimated using the 1,000 samples in the simulated data, along with associated inference. Each of the subsequent rows shows the log-basis correlation estimate $\hat{\boldsymbol{R}}_{\log \boldsymbol{X}}$ and the associated inference using the compositional data for CCLasso and BAnOCC, respectively.

present, positive correlations become much more likely to be real than negative ones, an interesting observation to consider when interpreting real-world results from any of these methods.

BAnOCC and CCLasso agree well with the true magnitude and direction of the non-zero associations that both methods conclude are significant. For these associations, the relative difference with the true value is less than 15% for both methods. When the associations were rejected, the 95% credible interval from BAnOCC covered the true value, indicating its utility for evaluating the uncertainty of the estimate. The false negative rates were 25% for BAnOCC and 0% for CCLasso, a direct result of the higher tolerance for false positives CCLasso exhibits. In practice, this has the expected effect of dramatically lowering BAnOCC's false positive rate in recovering true correlations from compositional data.

Comparison of Type I and Type II Error Rates

We compared BAnOCC's performance as measured by type I and type II error rates to a range of previous methods (**Figure 1.4**): simplicial variation [Aitchison, 2003], SparCC [Friedman and Alm, 2012], CCLasso [Fang et al., 2015], SPIEC-EASI [Kurtz et al., 2015], and Spearman correlation (directly on the composition as a negative control). For a positive control, we also applied Spearman correlation to the basis (**Table 1.1** and **Section S1.5**).

Only BAnOCC and SparCC controlled type I error while maintaining high power for all correlation strengths (see also AUC boxplots in **Figure S1.10**). Both behaved similarly to Spearman correlation applied to the basis abundances, which represents the best possible performance as it uses the unconstrained data rather than the composition (impossible in practice when only the composition is available). As other authors have noted, SparCC does not guarantee that its log-basis correlation estimate has bounded elements nor that it is positive definite [Fang et al., 2015]. By

Table 1.1: **Methods included in an evaluation on simulated data.** Type I and type II error rates were determined for these methods by the correct or incorrect rejection of H_0; for CCLasso and SPIEC-EASI, no inferential methodology was provided and so the correct or incorrect estimation of w_{jk} as zero was used. Note that although SPIEC-EASI infers the precision matrix, construction of the true correlation matrix in the simulated data guarantees that the same elements will be non-zero in the precision and covariance matrix.

Method	H_0	Error calls	Inference Method
Simplicial Variation	$\frac{1}{t_{jk}} = 0$	inference	one-sided permutation test
SparCC	$w_{jk} = 0$	inference	authors' bootstrap-based method
CCLasso	$w_{jk} = 0$	estimation	-
SPIEC-EASI	$w_{jk} = 0$	estimation	-
BAnOCC	$w_{jk} = 0$	inference	95% credible interval
Spearman (composition)	$\rho_{X,jk} = 0$	inference	two-sided permutation test
Spearman (basis)	$\rho_{X,jk} = 0$	inference	two-sided permutation test

contrast, BAnOCC not only estimates a positive definite correlation matrix with bounded elements, but also can infer network edges based on the precision matrix as well.

Several methods proved to control the type I error rate poorly: Spearman correlation exemplifies this as a negative control, but simplicial variation, SPIEC-EASI using GLASSO and to a lesser extent CCLasso were comparable. This is somewhat expected in simplicial variation, but SPIEC-EASI using GLASSO may not be performing as expected, especially since in contrast the Meinshausen-Bühlmann neighborhood selection method did control type I error. This may also possibly be because the neighborhood selection infers each element of the matrix one at a time, while GLASSO infers the matrix all at once; this makes the GLASSO optimization a more difficult problem.

Feature 5 in the template dataset has a large mean and variance, while feature 3 has a small mean and variance. This results in a strong negative spurious correlation in the composition, which gives rise to interesting behavior of essentially all methods when detecting this association. When the true association is negative, many

compositionally-appropriate methods such as BAnOCC, SparCC, and SPIEC-EASI (MB) do poorly at detecting the true correlation (**Figure 1.4B**) because the negative correlation is difficult to attribute to the basis rather than spurious correlation. Conversely, more naïve methods such as simplicial variation and Spearman correlation do very well at detecting a weak negative correlation between these two features because this becomes a strong negative correlation in the composition. This simulated example thus provides some insight into the form of sensitivity / specificity tradeoff that applies in the constrained, information-loss setting of identifying true correlations from compositions.

1.3.3 A Microbial Interaction Network from the Human Microbiome Project

As an example application, we inferred a correlation network among microbial taxa profiled using ecological data from the Human Microbiome Project [The Human Microbiome Consortium, 2012] (**Figure 1.5**). Microbial community sequencing generates compositions by assigning sequencing reads to microorganisms; since nucleotide sequencing depth is arbitrary, the resulting counts are not informative regarding the basis and are often normalized to relative abundances. Co-variation patterns in such data are of interest because they suggest ecological interactions, such as mutualism (positive correlation) or predation (negative correlation) [Faust, K. and Sathirapongsasuti, F. et al., 2012].

The microbial taxonomic relative abundance data used here consisted of 523 microbial features measured across 700 total samples using MetaPhlAn2 v2.0_beta1 [Truong et al., 2015] in July of 2014, further excluding from all networks markers removed in the subsequent version's database (v2.0_beta2). These samples were in turn drawn from 127 individuals at six distinct body sites. Microbial ecology differs at each body site [The Human Microbiome Consortium, 2012], providing examples

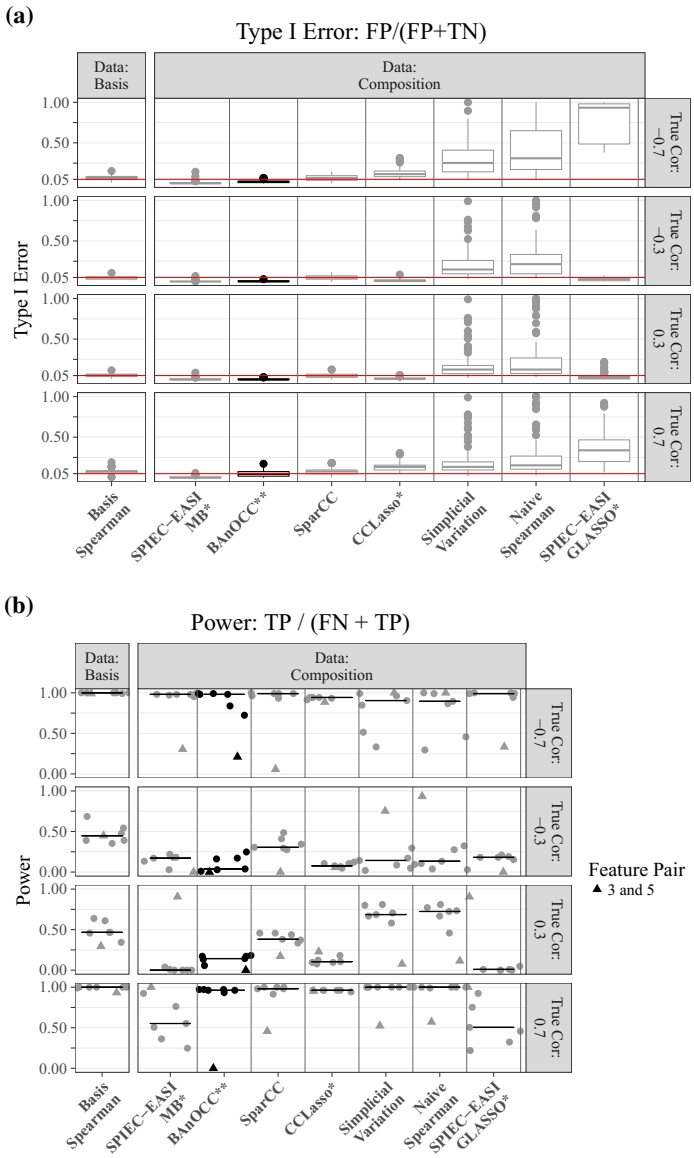

Figure 1.4: **The BAnOCC model controls type I error while maintaining power.** Results on simulated data comprising SparseDOSSA-derived compositions modeled on a low-diversity dataset with 14 features. The type I error rate is controlled at the 0.05 level for BAnOCC and approximately so for SparCC, CCLasso, and SPIEC-EASI (MB), but not for simplicial variation or Spearman correlation (on the composition, a negative control). BAnOCC maintains good power across all true correlation values, but as expected has better power for stronger true correlation values. Type I and type II error rates are determined by correct or incorrect rejection of H_0 based on inference (simplicial variation, SparCC, Spearman correlation, and BAnOCC) or estimation (CCLasso and SPIEC-EASI). (**Figure S1.10** and **Figure S1.11**).

for BAnOCC analysis that ranged from diverse, relatively even communities (such as stool) to less diverse, highly skewed ecologies (such as the vaginal posterior fornix). For each of three representative body sites (stool, posterior fornix, and buccal mucosa), we selected the first time point from each subject, collapsed taxonomic information to the genus level, and then removed features with relative abundance less than 0.0001 in at least 50% of samples. With too few features, little to nothing can be concluded about the true correlations; so if fewer than 10 features remained we lowered the prevalence cutoff until 10 features were retained.

The hyperparameters for the gamma prior on λ were $a = 0.5$ and $b = 5$ for all body sites, ensuring that we gave substantial weight to sparser precision matrices. For all body sites, we used the prior variability of the log-basis means $\boldsymbol{L} = 30\boldsymbol{I}$; each body site, however, had a different n_j so that the distribution of the sums of medians were similar across different body sites (see **Figure S1.12 - S1.15**). We further compared BAnOCC's inferred network using the log-basis correlation matrix with that from CCLasso, and BAnOCC's inferred network using the log-basis precision matrix with SPIEC-EASI. There is broad agreement between the methods as to which edges are significant, with very few edges discrepant between the methods (**Figure S1.16**).

In stool, BAnOCC inferred several positive associations between genera belonging to the same family in the stool (**Figure 1.5A**), in particular *Bacteroides*, *Odoribacter*, *Parabacteroides* and *Alistipes* in the family Bacteroidales. Until recently, these genera were classified as part of the same genus [Rajilić-Stojanović and de Vos, 2014]. This supports the common observation that closely (but not too closely) related taxa tend to have positive ecological associations [Barberán et al., 2012]. Additionally, positive associations in the buccal mucosa (**Figure 1.5B**) connect taxa that are known to physically co-aggregate; in particular, *Fusobacterium* interactions with species from the *Porphyromonas* and *Capnocytophaga* genera (among others) are crucial in biofilm formation [Kolenbrander et al., 2010] and have been previously recovered from 16S-

based ecological analyses [Faust, K. and Sathirapongsasuti, F. et al., 2012]. Lastly, we can see the well-documented negative association between the *Lactobacillus* genus in the posterior fornix with several genera associated with dysbiosis such as *Gardnerella* and *Prevotella* [Gajer et al., 2012] (**Figure 1.5C**).

Two interactions newly suggested by this analysis involved the Proteobacteria across multiple body sites, and specifically in stool and the oral cavity (buccal mucosa). The genera *Escherichia* and *Haemophilus* represent the two major proteobacerial residents in these habitats, respectively, and both were involved in predominantly negative interactions with more typical, abundant members of these communities (e.g. *Faecalibacterium* and *Eubacterium* in the gut, *Leptotrichia* or *Corynebacterium* in the mouth). These clades are highly phylogenetically diverged and tend to carry larger, more generalize genomes and pan-genomes [Hogg et al., 2007; Rasko et al., 2008]; this suggests that they will overgrow in these habitats only in unusual situations, exemplified by *E. coli*'s abundance in the gut primarily during inflammation [Arthur et al., 2012]. Further details may be provided by future analyses using BAnOCC or related methods on species or strain-level ecological profiling.

1.4 Discussion and Conclusions

Here, we describe BAnOCC, a Bayesian method for inferring the log-basis correlation structure from compositional data. Assuming a log-normal distribution on the unknown basis counts, the model estimates the log-basis correlations using a sparsity-inducing shrinkage prior on the log-basis precision matrix. Unlike existing correlation-inference methods that summarize pairwise associations using a single point estimate, BAnOCC yields uncertainty estimates of the precision, covariance, and correlation parameters. Simulation results show that BAnOCC performs as well as or better than existing methods in controlling type I error while maintaining power for network edge detection from compositional data. Finally, we applied the method

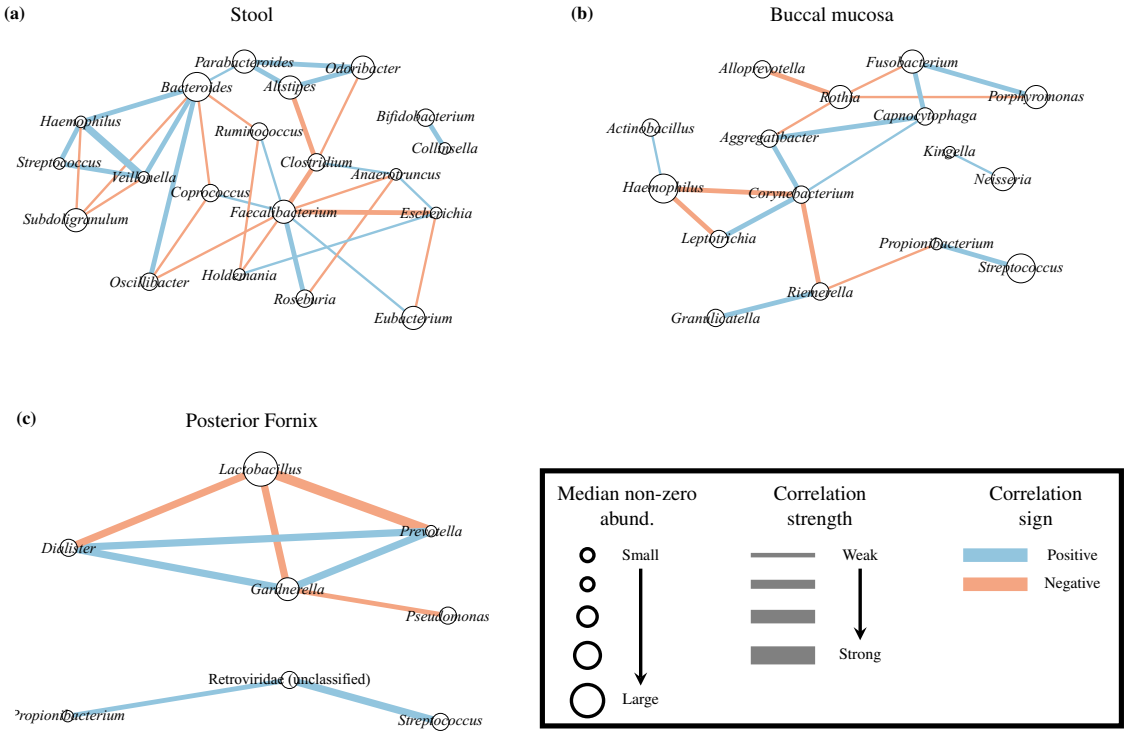

Figure 1.5: **BAnOCC association networks from the Human Microbiome Project.** The association networks inferred from three HMP body sites: stool (**A**), buccal mucosa (**B**), and posterior fornix (**C**). Using four chains with a minimum of 5000 iterations, we ran BAnOCC until convergence (see details in **Section S1.5**). Only significant correlations stronger than 0.15 are shown (see **Table S1.3**, **Table S1.1**, **Table S1.2**). The GLASSO prior results in sparse networks for these datasets, highlighting individual associations between taxa.

to assess microbial relationships in the human microbiome, confirming established interactions and suggesting novel ones for future validation.

Analysis using a Taylor series approximation provided one of the first characterizations of properties that make true correlations "difficult" to recover from compositions, or conversely "easy" to miss as false negatives. In particular, this depends not only on the more intuitive number and evenness of feature means, but also on the distribution of their variance. This allowed us to simulate designedly difficult test cases for BAnOCC and a variety of published methods, in contrast to previous simulation studies that relied primarily on relatively simple synthetic data [Ban et al., 2015; Fang et al., 2015; Faust, K. and Sathirapongsasuti, F. et al., 2012; Friedman and Alm, 2012]. In most studies, spurious correlation is noted to be commonly present and of varying magnitudes and directions [Kurtz et al., 2015]. However, the possible sensitivity of methods to the type of spurious correlation encountered has not been explored and is an important contribution to the characterization of existing and future methods.

We anticipate several computational and statistical refinements that may further improve BAnOCC's performance. While BAnOCC uses 95% credible intervals for inference, these can be overly conservative [Li and Lin, 2010]. Alternative thresholding methods may improve on this, such as the scaled neighborhood criterion [Li and Lin, 2010] or the partial-correlation based approach of [Carvalho et al., 2010] and [Wang, 2012]. A discrete-continuous mixture prior such as the G-Wishart prior [Dawid and Lauritzen, 1993] or the covariance selection prior [Wong et al., 2003] on the log-basis correlation matrix would further allow the posterior probability that $w_{jk} = 0$ to be nonzero, and this quantity could be used as a threshold.

For applications specifically on count data, such as microbial compositions, the data could be modeled more accurately by adding a hierarchical layer. This would generate measurement counts conditional on the unobserved basis counts, making the

observed compositions a function of normalized measurement counts. The degree of zero-inflation observed in ecological data could also be modeled directly using a hurdle or mixture model, or a multinomial distribution for the measurement counts. This would provide a particularly targeted approach for microbial ecology, in which more detailed data (at the species or strain level [Truong et al., 2015]) could be further incorporated. We thus hope to refine both the accuracy of compositional correlation inference and the applications to microbial community data in future studies.

2

Metagenomic discovery of interactions among species in the human microbiome

Abstract

Most biochemical activities in microbial communities are carried out by individual clades, often in groups as specific as a species or individual strain. Microbial taxa can now be identified at this resolution using metagenomic approaches, providing opportunities to computationally explore such interactions at scale. This study provides the first investigation of potential ecological interactions among human-associated microbial species across six body site habitats. We identified interactions in 2,077 metagenomes from the Human Microbiome Project using a novel compositionally-aware Bayesian method, BAnOCC. The resulting network included 136 taxa and 1,961 significant edges, with associations concentrated in the oral cavity and gut. Interaction patterns were in part explained by ecological factors, including oxygen tolerance, phylogenetic relatedness, and niche differentiation. *Streptococcus cristatus* and *Corynebacterium matruchotii* from the oral cavity, for example, were identified as known co-aggregators, as were other pairs that interact metabolically (e.g. streptococci and veillonella). Co-occurrence and exclusion motifs were significantly enriched across the body and represent potential ecological functional groups for future characterization. This work thus represents a novel high-resolution profile of the large-scale

intermicrobial relationships in the human microbiome.

2.1 Introduction

Microbial interactions drive the formation and maintenance of ecological communities, but they are challenging to identify from sequencing data. Such interactions can include physical co-adherence and co-aggregation, as well as metabolic interactions and competitions through cross-feeding, or metabolic inhibition [Kolenbrander et al., 2010]. As in any ecological food web, the inter-microbial interactions upon which a community relies can be extremely specific, depending on the exact species or strains of organisms that are present [Kolenbrander et al., 2006]. However, methods have only recently been developed to profile taxa at this level of detail from culture-independent sequencing [Segata et al., 2012] Also due to the limited statistical methods capable of accurately inferring ecological interaction networks from metagenomes [Fang et al., 2015; Friedman and Alm, 2012; Kurtz et al., 2015], potential microbial species interactions in the human microbiome have not yet been well-studied.

Most previous studies of interaction networks in the human microbiome have relied on less-precise data to provide a first look at this complex ecology. 16S rRNA gene amplicon sequencing, under most circumstances, yields taxonomic profiles at the genus level, often with substantial technical noise due to the PCR amplification process [Fox et al., 1992; Schloss et al., 2011]. Faust and Sathirapongsasuti and colleagues [Faust, K. and Sathirapongsasuti, F. et al., 2012], for example, used amplicon data to explore genus-level associations, finding strong segregation by body area and a relationship between the direction of an association with the relatedness of the organisms [Faust, K. and Sathirapongsasuti, F. et al., 2012]. Other studies have focused on the development of novel methods for detecting associations from microbial taxonomic abundance data, with limited ecological interpretation of the resulting networks [Fang et al., 2015; Friedman and Alm, 2012; Kurtz et al., 2015]. Large-scale networks from

other habitats such as soil [Barberán et al., 2012] and activated sludge [Ju et al., 2014] have revealed general patterns such as prevalent co-occurrences between taxa within a phylum, but it is unclear to what extent these results generalize to human-associated communities.

Although much microbial physiology–including interaction partners–is species- or strain-specific, high-throughput sequencing techniques have only recently been developed for taxonomic profiling at this resolution [Luo et al., 2015; Truong et al., 2015, 2017]. The area has seen extensive prior work, however, typically using lower-throughput methods such as direct co-culture or fluorescent microscopy [Mark Welch et al., 2016]. Many examples of individual interactors are known from such studies, such as the mutual inhibition of *Streptococcus mutans* and *Streptococcus sanguinis* [Kreth et al., 2005], co-aggregation between streptococci and other oral colonizers [Handley et al., 2005; Lancy et al., 1980, 1983; Mark Welch et al., 2016], and oral biofilm formation structure and dynamics [Mark Welch et al., 2016]. These methods are difficult to scale past tens of organisms or to apply to microbes that are not amenable to *ex vivo* or isolate culture.

To overcome these limitations in the first species-level study of human-associated microbial interactions at scale, we used over 1,500 metagenomes from the Human Microbiome Project (HMP) [Lloyd-Price et al., in press]. These profiled nearly 1,000 species from over 250 individuals across six body sites (anterior nares, buccal mucosa, tongue dorsum, supragingival plaque, stool, and posterior fornix) at two time points (separated by an average of seven months). We employed BAnOCC [Schwager, 2017], a recently developed, compositionally-appropriate Bayesian correlation-based method, to infer inter-species interactions at each body site, and combined the information across the two time points to generate a robust network of replicated edges. Several associations were supported by known co-aggregations and metabolic interactions, particularly in the oral cavity. Ecological factors, including oxygen tolerance,

phylogenetic relatedness, and niche-differentiation, further explained the connectivity of the network. Sub-structure analysis revealed enrichment of inter-organism positive and negative feedback motifs and the presence of several distinct sub-communities in supragingival plaque and buccal mucosa. Taken together, these analyses represent the first high-throughput characterization of inter-species (rather than inter-generic) associations in the human microbiome.

2.2 Results

2.2.1 An ecological network of potential inter-species interactions

We inferred a network of microbial species interactions across the human body using the the Human Microbiome Project (HMP) population [The Human Microbiome Consortium, 2012], which targeted a North American cohort of adult men and women lacking overt disease. The samples analyzed here were drawn from 265 individuals who were sampled at up to 18 body sites over up to three time points [Human Microbiome Project, 2012; Lloyd-Price et al., in press]. The resulting 2,355 metagenomes were drawn primarily from six targeted body sites (anterior nares, supragingival plaque, buccal mucosa, tongue dorsum, stool, and posterior fornix) at two of the three possible time points. After standard sequence- and sample-level quality control [Lloyd-Price et al., in press], remaining samples were taxonomically profiled using MetaPhlAn2 [Truong et al., 2015], and subsequent quality control further consolidated technical replicates and removed species present in only few samples at either time point (Methods). This retained an average of 134 samples per body site at the first time point and 124 samples per body site at the second (**Table S2.1**, **Figure 2.1A**).

Using these data, we constructed a network for each body site and time point using BAnOCC [Schwager, 2017], a Bayesian method for inferring correlations from

compositional data (**Figure 2.1B**). BAnOCC explicitly models the data generation process that gives rise to compositional counts and uses sparsity assumptions to approximate the correlation between unconstrained counts using relative abundance data. This is particularly important for microbial ecology as many observed correlations on relative abundance data can be spurious (either reversed in direction, or altered in magnitude). BAnOCC inferred significant associations between microbes within the same body site; cross-body-site correlations for each time point were generated using a permutation test on Pearson correlation with FDR correction [Benjamini and Hochberg, 1995] (**Section 2.4.2**) at a significance level of 0.05. The resulting networks at each time point were combined by retaining only edges significant at both time points: this final network was the subject of our investigation.

Beginning with a high-level view of the network (**Figure 2.2A**), oral body sites (buccal mucosa, supragingival plaque, and tongue dorsum) tended to have more edges than other body sites, resulting in larger and denser sub-networks. This could reflect the complex, organized microbial structures in well-studied oral biofilms [Kolenbrander et al., 2010] or the interactions necessary to protect anaerobes from oxygen toxicity in an aerobic environment. We observed that there were many more negative edges in supragingival plaque and buccal mucosa than in tongue dorsum, suggesting that these body sites are comprised of distinct alternative sub-communities that tend to co-exclude. Despite the density of the network at the oral sites, most species pairs at each body site were not significantly related, which is expected as BAnOCC is intended to produce a sparse network (**Figure S2.1**). Cross-body-site correlations were also very sparse, with all significant edges occurring within the oral cavity, and most of these connected the same species across multiple body sites (**Figure S2.2**).

Several of the most confidently predicted interactions in each body site (**Figure 2.2B**) were confirmed by previous lower-throughput studies. This includes, for example, the co-exclusion of *Lactobacillus iners* and *Lactobacillus crispatus* in the vaginal

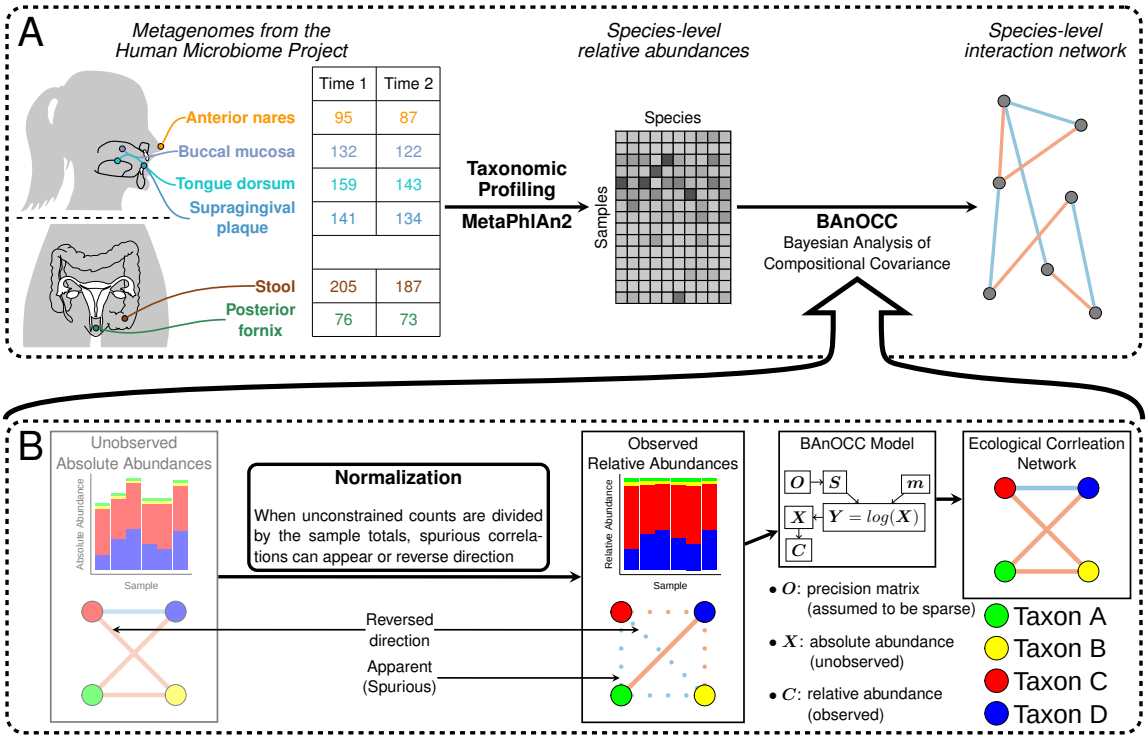

Figure 2.1: **A network of species-level microbial interactions from across the human body.** (**A**) We taxonomically profiled [Truong et al., 2015] species from 1,554 samples from six body sites in the Human Microbiome Project (HMP) population (263 individuals). Ecological interaction networks were inferred from these profiles using BAnOCC [Schwager, 2017], a Bayesian and compositionally-appropriate method. The resulting network included 215 species-body-site combinations (nodes) and 1,961 edges retained across both time points analyzed. (**B**) Compositional sequence counts or relative abundances cannot be used to directly compute correlations for ecological network inference. Instead, BAnOCC includes a Bayesian model of the unobserved true microbial organism abundances, for which the observed relative abundances are a proxy. By incorporating a set of sparsity assumptions, the underlying ecological network can be approximately inferred. Accounting for the process of "normalizing" from (unobserved) microbial abundances to (observed) sequence counts is important because, if ignored, it can cause incorrect, spurious correlations to appear.

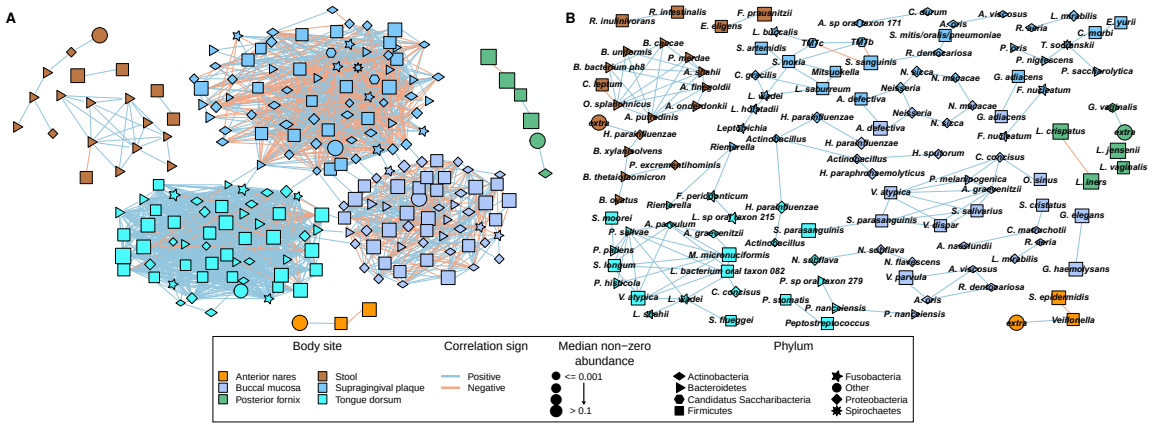

Figure 2.2: **Ecological interactions in the human microbiome.** (**A**) Ecological interaction network of all 1,961 correlations significant among 136 species at both time points, and (**B**) the strongest 30 edges in each body site that were significant at both time points. For complete network data see **S2.4 Data**

community [DiGiulio et al., 2015; Gajer et al., 2012]. In the oral cavity, species of *Veillonella* and *Streptococcus* interact both metabolically and physically [Chalmers et al., 2008; Hughes et al., 1988], and positive associations between species of these genera were predicted in the buccal mucosa. The cluster of *Bacteroides*, *Alistipes*, *Odoribacter*, and *Parabacteroides* species in the gut reflects their phylogenetic similarity [Rajilić-Stojanović and de Vos, 2014]; they are similar enough to flourish in the same environment, but sufficiently distinct to niche-adapt in order to coexist non-competitively. Many of the strong positive associations body-wide were between closely related species, reflecting a similar ecological pattern. This observation motivated us to look more closely at the ecological drivers of the network connectivity.

2.2.2 Ecological drivers of network topology

We next tested whether a series of microbial physiological properties were able to explain aspects of the inferred network structure (**Figure 2.3**). Oxygen tolerance was one of the most apparent drivers of the observed topology. In aerobic environments, strict anaerobes require the presence of other microbes to shelter them from the toxic

effects of oxygen. In any ecological network, this should manifest as a bias towards strong positive edges between strictly anaerobic species. To test this, we used a "weighted edge sum" measure for each taxon within a body site (**Figure 2.3A**), since this measure is less sensitive to the particular significance cutoff used (**Section 2.4.3**). A positive value indicates that a taxon has more positive edges than negative ones. Consistent with the already-anaerobic environment of the gut, the weighted edge sum did not strongly correlate with oxygen tolerance in stool, though this is also partly due to the sparse interaction network at this body site. On the other hand, in supragingival plaque and tongue dorsum–both sites with exposure to oxygen–weighted edge sums were significantly higher for anaerobes and obligate anaerobes than for microaerophiles and facultative anaerobes ($p < 0.001$, Wilcoxon rank sum test). Unlike tongue dorsum and supragingival plaque, which allow microbes to persist and create anoxic environments, the buccal mucosa consists of an unprotected mucin layer. This manifests as a dearth of strong positive interactions involving anaerobes and obligate anaerobes; these tend to be less prevalent and abundant in the buccal mucosa, possibly because they are transient rather than permanent members of this community [The Human Microbiome Consortium, 2012].

A substantial component of network structure was also explained by phylogenetic relatedness among species (**Figure 2.3B**). Specifically, we found that microbes with lower phylogenetic distances (as defined by the nucleotide divergence between the last common ancestors of the species' reference genomes co-occurred more often than they were co-excluded, and that this pattern decreased with phylogenetic distance. Edges involving two species from the same genus were much more likely to be positive than negative (**Figure S2.3**). As the phylogenetic distance increased past approximately the level of class, however, the numbers of negative and positive interactions become approximately equal at most body sites (**Figure S2.3**). The corresponds to the expectation that sufficiently similar organisms will co-occur in similar environments

[MacArthur, 1965].

Finally, differences among ecological contexts also explained a subset of network structure (**Figure 2.3C**). For example, two species might physically interact to form a biofilm on a hard surface but not when suspended in saliva. We tested which species had the greatest differences in weighted edge sum between sites, while controlling for density between sites (**Section 2.4.3**). The analysis was restricted to the oral sites, as these had the largest numbers of species in common. Several of the largest differences in weighted edge sum occurred when a taxon had many positive correlations in the buccal mucosa but negative correlations in plaque. This could be an artifact of plaque development, whereby species that are early colonizers (such as Streptococcus species) are displaced by others in the process; this would appear as a negative correlation when measured across people with differing stages of plaque development. It could also suggest that certain species become more competitive when confined in the nutrient-limited structure of a biofilm. Overall, among analyzable organisms (species occurring in all three oral sites), we observed an inverse relationship between the differences in weighted edge sum and a niche-association score of Lloyd-Price et al. [in press] (**Figure S2.4**), which measures the degree to which different strains of a species have specialized to different body sites. This suggests that some taxa are able to persist in multiple body sites by altering their interactions with other community members rather than by genomic alteration.

Several typical microbial physiological properties did not appear to be associated with network structure. Gram type, for example, was not significantly associated with species' connectivity, with the exception of enriched positive interactions among gram-negative organisms in the stool (corresponding to Bacteroidetes) than for gram-positive organisms (corresponding to Firmicutes) (**Figure S2.5A**). Genomic measures of generality such as genome size, number of genes, or the ratio of core-to-pan genome size, also did not associate significantly with network structure,

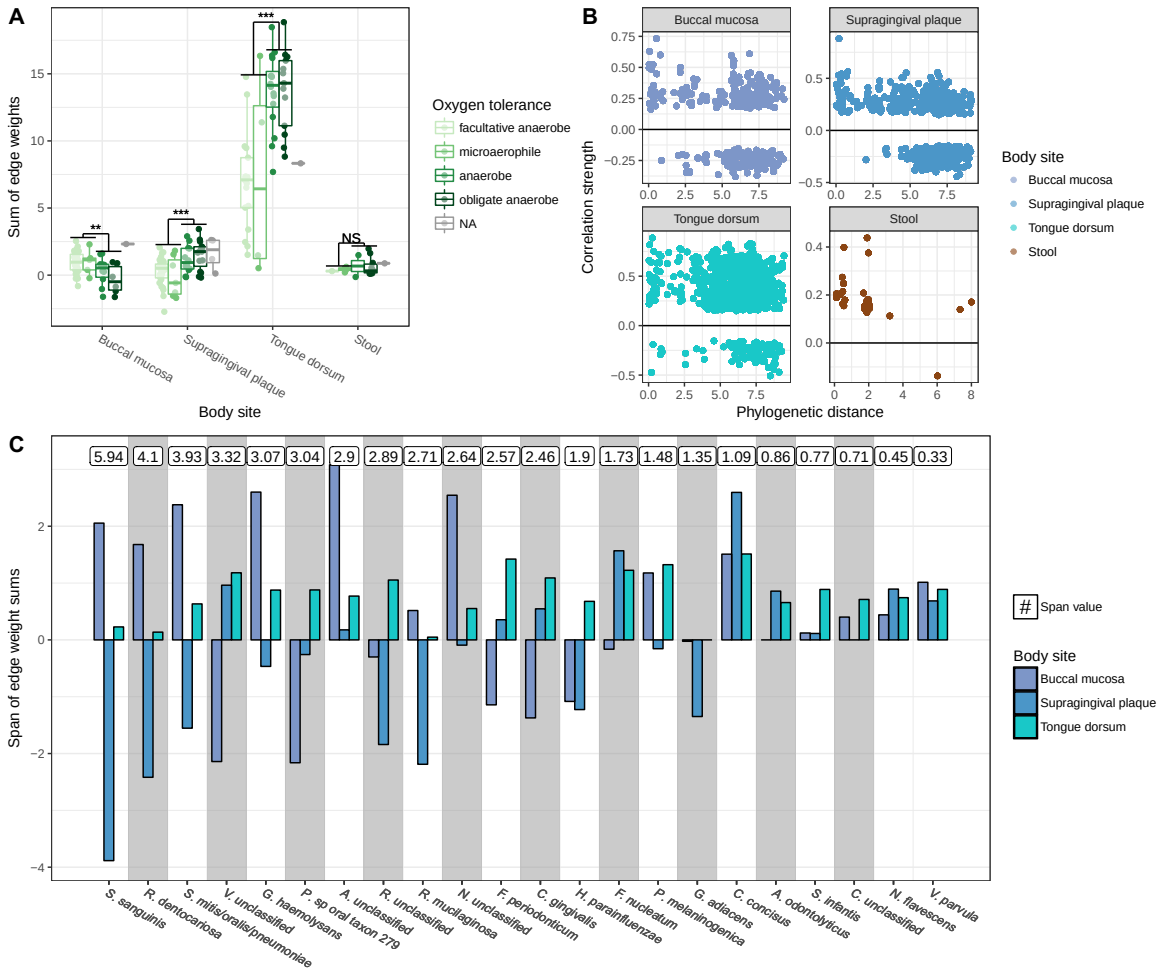

Figure 2.3: **Relation of taxon properties to network properties.** (**A**) Differences in species' aerotolerance were significantly associated with network structure, as summarized by the weighted sum of interactions. Clades containing more than one oxygen tolerance phenotype are represented as NA. Wilcoxon rank sum test: **, $p < 0.01$; ***, $p < 0.001$; NS, not significant. (**B**) Phylogenetic distance was also associated with network structure, here summarized as nucleotide divergence (calculated by using the last common ancestor of all sequenced strains for a species) versus their correlation measure as estimated by BAnOCC. (**C**) A subset of species occurred in multiple body sites with different predicted interactors. This shows taxa with the greatest differences in scaled sum of incident correlations when detected in all three oral body sites (**Section 2.4.3**). To correct for differences in body-site-specific networks, the sum of the correlations is normalized by the average weighted edge sum in each body site.

although our ability to analyze these within the network was limited by the low number of relevant reference genomes available (**Figure S2.5B-D**). Particularly given the strain-specificity of interaction and metabolic phenotypes, this highlights the utility of increased reference genome coverage even among relatively well-sequenced human-associated microbes.

2.2.3 Evaluation of network substructure

Next, we assessed the presence of specific substructures within the inferred interaction network, specifically inter-species motifs and modules. "Motifs" in this context represent combinations of positive and/or negative interactions that occurred more frequently than expected by chance, suggesting functional selection for inter-organismal synergy or competition. "Modules" are groups of densely connected taxa suggesting groups of organisms that commonly interact. For both types of substructure discovery, we restricted our analyses to the three oral sites because these were by far the most dense.

First, we specifically explored "triangular" motifs consisting of three taxa that interact with each other via: (i) three positive interactions, in which all microbes flourish together; (ii) two positive and one negative interactions, in which one pair of microbes compete but each flourishes in the presence of the a third; (iii) two negative and one positive, in which one pair of microbes co-occur together but each competes with the third; and (iv) three negatives, in which all of the microbes compete. The counts of each type in the three oral networks were compared to 1,000 randomized networks (**Section 2.4.4**). Positive triangles, in addition to two negative/one positive triangles, were both enriched, the former suggesting groups of synergistic organisms (beyond pairwise interactions) and the latter indicating single organisms that might compete with mutualistic groups. Two positive/one negative and fully negative interactions were both significantly depleted in the true network, which cor-

responds with neither being biologically stable (as they would require competition with a co-occurring organism; **Figure 2.4A**).

Next, we identified dense modules of mutually interacting taxa, representing potential larger multi-organism functional groups (**Figure 2.4B**). We identified six modules (two in supragingival plaque and four in buccal mucosa) using information mapping [Rosvall and Bergstrom, 2008], although these were generally consistent across several methods (**Section 2.4.4**, **Figure S2.6**). There was little evidence of larger clusters, likewise based on a variety of clustering methods and evaluations (**Section S2.1**; **Figure S2.7 - S2.8**).

Among smaller modules, supragingival plaque, as an example, formed two clusters: one predominantly consisting of species associated with periodontal disease progression and one smaller cluster associated with early stage plaque. From the former, *Treponema socranskii*, *Prevotella nigrescens*, *Selenomonas noxia*, and *Campylobacter gracilis* were all members of the "orange" complex of supragnigival plaque, and Leptotrichia buccalis was associated with the "red" complex, both indicators of disease [Haffajee et al., 2008]. Further, other taxa in this cluster (particularly TM7 members) have also been associated with periodontitis in previous studies [Brinig et al., 2003; Kumar et al., 2003]. Conversely, the second cluster contained known early plaque colonizers such as *Streptococcus mitis* and *Streptococcus oralis* as well as *Actinomyces*, *Neisseria*, *Gemella*, *Rothia*, and *Haemophilus* species [Kolenbrander et al., 2005; Mark Welch et al., 2016]. This early-stage cluster also had a few members, such as *Streptococcus sanguinis*, which have been associated with periodontal health and the absence of caries [Becker et al., 2002].

The four modules in the buccal mucosa broke into two co-exclusive pairs, each with one cluster enriched for facultative anaerobes and one cluster with more evenly distributed oxygen tolerances (**Figure 2.4B**). In one pair, the facultatively enriched cluster includes species of *Gemella*, *Neisseria*, and *Haemophilus*, which co-exclude

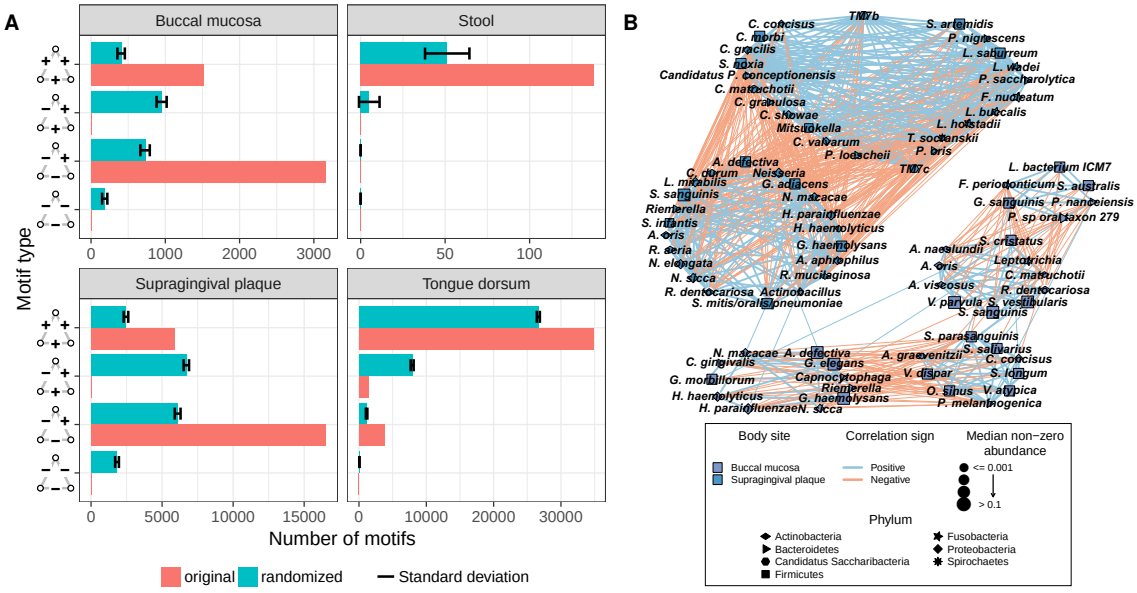

Figure 2.4: **Network sub-structure reveals higher-order community organization.** (**A**) We tested for enrichment of "triangular" network motifs (three-edge recurring patterns of positive and negative interactions) enriched beyond expectation by chance (based on 1,000 randomized networks; **Section 2.4.4**). Groups of positive (three +) or single positive (one +, two -) interactions were enriched, while single negative (one -, two +) and negative (three -) triangles were depleted in the true network. (**B**) Two modules in supragingival plaque and four modules in buccal mucosa were consistently found among the densely connected subparts of the network (**Section 2.4.4**). The former may correspond, for example, to early vs. late plaque colonizers [Haffajee et al., 2008; Kolenbrander et al., 2005; Mark Welch et al., 2016].

with the *Veillonella* and *Streptococcus* species included in the evenly distributed cluster. In the other pair, facultative anaerobes *Streptococcus* and *Actinomyces* co-exclude with *Prevotella nanceiensis*, *Streptococcus australis*, *Fusobacterium periodonticum*, *Lachnospiraceae*, and *Prophyromonas*. This structure suggests that in the buccal mucosa, two types of communities are present: one with more facultative anaerobes and another with more mixed oxygen tolerance levels. Finally, because of its high density, the tongue dorsum had variable numbers of clusters across methods and was excluded from more detailed analysis (**Figure S2.9**).

2.2.4 *Streptococcus cristatus* and *Corynebacterium matruchotii* interact in oral biofilm

Streptococcus cristatus (formerly known as *S. crista* [Trüper and De'clari, 1997] and *S. sanguis* serotype I [Handley et al., 1991]) has long been known to co-aggregate with *Corynebacterium matruchotii* (formerly *Bacterionema matruchotii* [Collins, 1982]) in co-aggregation assays [Lancy et al., 1980]. "Corncob" structures involving these two bacteria, whereby long rods of *C. matruchotii* are covered with coccoid *S. cristatus* (forming the "kernels"), are readily detectable by microscopy [Lancy et al., 1980]. This is facilitated in part by tufts of fimbrils along the surface of *S. cristatus*, composed of long and short types [Handley et al., 1985], the former of which interact with *C. matruchotii* via the SrpA protein [Handley et al., 2005]. Further, species from these two genera have been observed to interact and co-localize in supragingival plaque [Mark Welch et al., 2016], indicating that this interaction is likely to occur during *in vivo* plaque formation.

This interaction was well-supported by our data (**Figure 2.5**). Both bacteria were highly prevelant in two body sites (buccal mucosa and supragingival plaque), and in both of these locations we inferred a positive association between *S. cristatus* and *C. matruchotii* (**Figure 2.5A**). The relationship was strong when compared to

other associations involving these two species (**Figure 2.5B**) and when compared to all other edges (**Figure S2.10A**), giving strong evidence in favor of truly non-zero correlation even after compositional effects were accounted for. BAnOCC measures interactions using a correlation matrix, so that they are interpretable; this can result in non-"causal" correlations that result from relationships among other taxa [Kurtz et al., 2015]. To confirm that these correlations were unlikely to be such an artifact, we looked at the interaction in the precision matrix, which approximately measures the conditional correlation [Kurtz et al., 2015]. The interaction was significant at the precision scale as well as the correlation scale (**Figure S2.10B**). The association can further be observed visually in the raw relative abundance data (**Figure 2.5C**). Together, this indicated that our model strongly replicated a well-verified co-aggregation between two species in the human oral cavity.

2.3 Discussion and Conclusions

We explored associations between human-associated microbial species from six body sites across two time points using metagenomes from the expanded Human Microbiome Project (HMP1-II) [Lloyd-Price et al., in press]. The interaction network was inferred using BAnOCC, a compositionality-aware Bayesian model for identifying significant correlations among microbial community members. Our analysis confirmed several previously-documented associations and revealed potential ecological drivers of human microbiome structure, including oxygen tolerance, phylogenetic relatedness, and niche differentiation. The analysis further revealed several sub-structures in the resulting interaction network, which included enrichment for inter-species community regulatory motifs, as well as co-excluding modules in the buccal mucosa and supragingival plaque that are well-explained by established ecological trade-offs.

As with many recent studies [Franzosa et al., 2015], this investigation supports the importance of using the highest possible taxonomic resolution when characteriz-

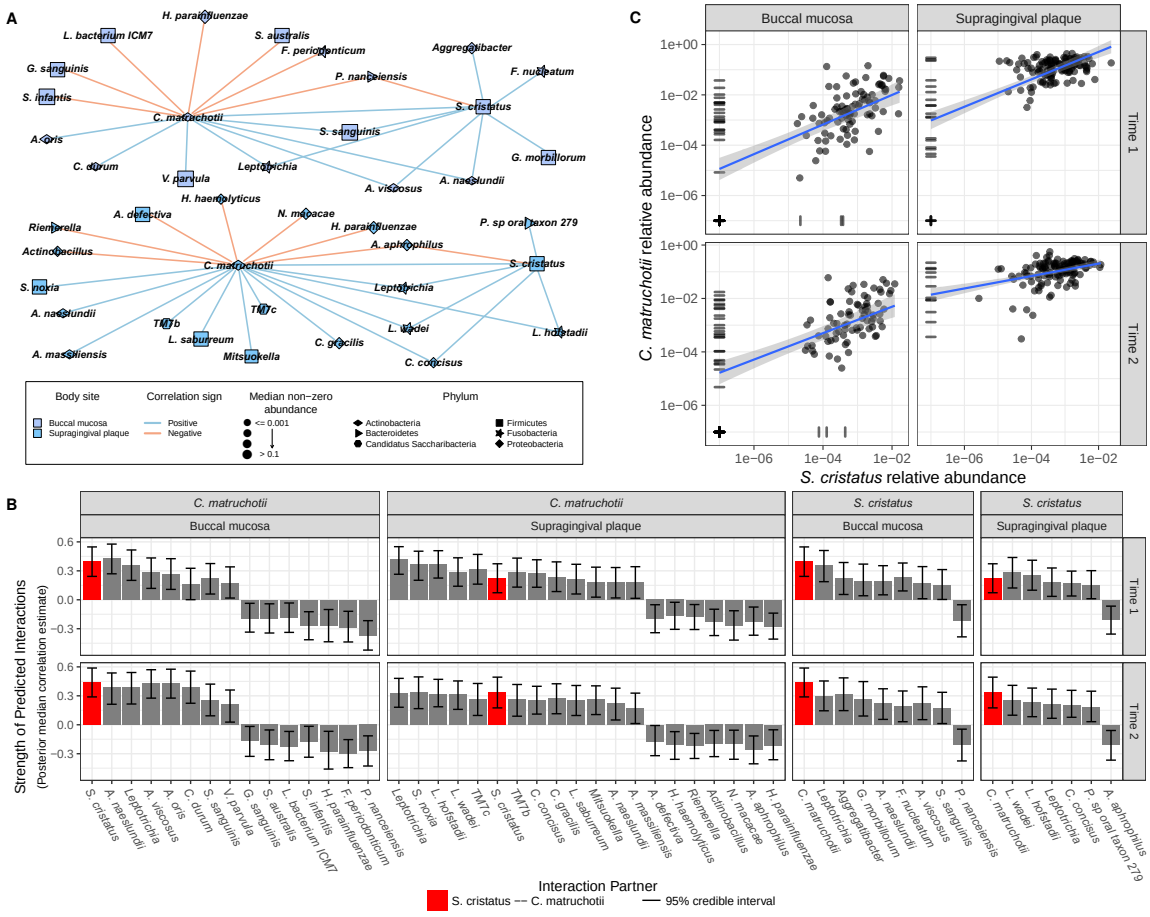

Figure 2.5: ***C. matruchotii* and *S. cristatus* have strong predicted and validated interactions.** (**A**) *Corynebacterium matruchotii* and *Streptococcus cristatus* nodes and their neighbors in buccal mucosa and supragingival plaque. (**B**) Inferred association strengths for the relationship between *C. matruchotii* and *S. cristatus* when compared with their neighbors. This interaction is strong relative to almost all other predicted association partners. (**C**) The raw (relative abundance) data make an apparent strong positive relationship between these taxa visible, in both body sites and at both time points.

ing members of the human microbiome. Microbial phenotypes are often species- or strain-specific, which can include interaction partners [Cisar et al., 1979; Palmer et al., 2017]. The specificity of co-aggregation partners among oral bacteria has been well-established. *Streptococcus* and *Actinomyces* species co-adhere during initial plaque formation, for example, by several well-characterized mechanisms that can rely on differing cell surface molecules among taxa [Kolenbrander et al., 2006]. Even for interactions among streptococci, the gene products responsible for mediating interactions can differ depending on strain [Palmer et al., 2017]. Species-resolved profiles are more easily obtained from metagenomic sequencing than from other high-throughput data types, and their importance is further visible here, in examples such as the co-exclusion of vaginal lactobacillus species (**Figure 2.1B**), which would not be visible at the genus level [DiGiulio et al., 2015; Gajer et al., 2012]. While it is possible to obtain strain-level resolution from metagenomic sequencing data [Scholz et al., 2016], strains are generally unique to particular individuals rather than broadly carried by many people [Franzosa et al., 2015]. This suggests that for future work identifying strain-specific interactions from culture-independent data, the gene sets responsible for these interactions may provide more generalizable biomarkers than the taxonomy itself.

There are several limitations to the current study, the most important one being the cross-sectional design. Observational cross-sectional data can only inform undirected interactions, and therefore, cannot be used to mechanistically infer ecology or causation. For example, a negative association in such data cannot distinguish whether microbe A actively inhibits microbe B or vice versa, or whether another mechanism is at play such as competition for common resources. Experimental data is therefore necessary to confirm the presence and mechanism of interactions detected as undirected associations. Densely sampled time courses would allow the direction of edges to be determined, and indeed have been used in previous work to detect

other types of potential microbial interactions in specific environments such as the murine gut [Bucci et al., 2016; Marino et al., 2014; Stein et al., 2013]. Additionally, inferring significant associations from relative abundance or count data is methodologically challenging [Fang et al., 2015; Friedman and Alm, 2012; Kurtz et al., 2015], and calibration information that provided or approximated absolute counts of microbial abundances remove this constraint. This network of potential interactions thus provides a starting point for more efficiently establishing how specific pairs or groups of taxa interact, which would otherwise require more resource-intensive approaches for co-aggregation [Kolenbrander et al., 2006] or metabolic cross-feeding [Belenguer et al., 2006; Falony et al., 2006] screens.

Our study thus provides a first look at potential species-level ecological networks across the human microbiome, detailing potential interactions between specific microbes in a baseline population. This high-level overview showed several ecological drivers of microbial interaction patterns, including oxygen tolerance, phylogenetic relatedness, and niche adaptation. Future work investigating the mechanism by which these drivers influence the assembly and stability of the microbiome would reveal the degree to which these interactions via co-adherence, cross-feeding, metabolic competition, or other mechanisms drive the formation of microbial communities. Further, we delineated several functional groups in the oral microbiome that are partially explained by known biology. Additional studies addressing the distribution of these groups across people and through time could provide insight into how microbial communities shift in response to perturbations. Finally, this work provides a way to prioritize specific pairwise interactions for in-depth chracterization, yielding new insights about how microbial communities persist and regenerate after perturbations such as oral hygiene or dietary changes. Future studies that more precisely delineate the causality and mechanisms of interactions in the human microbiome will thus continue to contribute to our understanding of microbial ecology and its role in health

and disease.

2.4 Methods

2.4.1 Data and quality control

The data were obtained from the Human Microbiome Project; metagenomic relative abundance profiles for each body site were downloaded from http://hmpdacc.org/hmsmcp2 during December, 2016, yielding 2,355 samples and totaling 2,437 taxa across 17 body sites. These were profiled using MetaPhlAn2 v2.2.0 [Truong et al., 2015] and quality-controlled as in [Lloyd-Price et al., in press], removing sixty outlier samples. We removed any taxonomic features that were unclassified at the genus level or higher. The resulting species-level relative abundance table consisted of 951 species for 265 subjects sampled at up to three time points each, separated by, on average, 7 months (minimum 1 month, maximum 12 months).

We only used the six body sites that had at least 20 subjects at each time point (**Table S2.2**), and randomly selected a representative when multiple technical replicates were available to ensure that each sample corresponded uniquely to one individual, body site, and time point. We filtered features by retaining those with abundance of at least 0.0001 in at least 50% of samples. Due to the lower complexity of the posterior fornix, we used a lower cutoff of 42% to ensure that at least 10 features passed the filtering (see also **Table S2.3**). The third time point was not considered as it constituted very few samples per body site (at most 80 in any one site). This resulted in two time points for each of the six body sites with feature counts ranging from 10 (posterior fornix) to 109 (supragingival plaque) and sample counts ranging from 73 to 205 (**Table S2.1**). To ensure that sample relative abundances summed to one after filtering the features, we added an "extra" feature consisting of the remaining unaccounted-for abundance.

2.4.2 Generating a network of microbial interactions

We used BAnOCC version 1.0.0 to infer ecological associations between species within a body site [Schwager, 2017]. Briefly, BAnOCC uses MCMC sampling to infer the pairwise correlations between log-transformed species abundances by incorporating sparsity-inducing shrinkage prior on the covariances. and their average log-transformed abundances. The sparsity-inducing prior allows substantial mass around zero to ensure strong shrinkage of spurious correlations, while simultaneously allowing recovery of true signals through its heavy tails. Because of its Bayesian approach, it requires priors for both of these parameters as well as starting values. We used an agnostic normal prior for the average log-transformed abundances ((**Figure S2.11A**). The prior on the correlations in the model is specified as a distribution on the degree of shrinkage imposed by the model through a shrinkage parameter ; our specified prior on was such that high weight was given to higher degrees of shrinkage (specified by smaller values; see (**Figure S2.11B**). The Markov chain was run for at least 5,000 iterations, of which at least 1,500 were discarded as burn-in and the rest were used as draws from the posterior distribution for making inference (**Table S2.4**). This choice of running parameters appeared to work satisfactorily based on convergence diagnostics (Rhat statistic [Gelman and Rubin, 1992] ¡1.1). Further, we thinned the MCMC samples by keeping every other or third observation from each chain and discarding the rest in order to eliminate any unwanted autocorrelation. Edges were considered significant if the 95% posterior credible interval across all posterior samples for that edge excluded zero.

To infer associations among species from different body sites, we employed a two-sided permutation test that permuted the sample labels in order to preserve the compositional structure of each body site. The correlations across each of the 15 () body site pairs at each time point were measured across 1,000 permutations and compared with the observed value. The resulting p-values were corrected for multiple-hypothesis

testing using Benjamini-Hochberg FDR correction [Benjamini and Hochberg, 1995] across all the tests (from all body sites and time points). We considered an edge significant if it has a q-value of less than 0.05. A larger cutoff did not yield appreciable differences in the edges included (**Figure S2.12**).

2.4.3 Ecological measures

To characterize the ecological drivers of the network, we employed several measures of ecological properties. These included a weighted edge sum to measure the degree to which a taxon engages in positive interactions, and phylogenetic distance to measure the relatedness of two taxa.

Weighted edge sum

To measure how positively-biased the interactions of a particular taxon are, we used the weighted edge sum, which is the sum of all edges incident on a particular node. This measure is less sensitive to the significance threshold than degree because it accounts for the magnitude as well as the existence of the edges. In detail, for each taxon in each body site of the final network, we calculated the sum of the correlations as $\sum_{i=1}^{n_j} \rho_{ij}$, where node j is the node in question, ρ_{ij} is the correlation between node i and node j and n_j is the number of nodes adjacent to node j in the final network.

phylogenetic distance

A PhyloPhlAn [Segata et al., 2013] tree for all isolate genomes downloaded in December 2013 was used to determine phylogenetic distances between species. Phylogenetic distance was calculated by measuring branch length between the pair of last common ancestors of all reference genomes for any given pair of taxa in the PhyloPhlAn phylogenetic tree; branch length represents the average number of mutations between the two taxa.

2.4.4 Network substructure

We examined two types of network structure: small, repeated motif patterns and large inter-connected modules. Triangular motifs were counted and compared to the counts across randomized networks. Modules were detected using the densest parts of the network in each of the three oral sites. Subsequently, several detection methods were employed; they gave general agreement.

Randomizing networks for motif discovery

We randomized the networks at each time point and body site such that the degree for each node remained the same using a Markov chain algorithm [Gkantsidis et al., 2003]. Briefly, this involves selecting two edges in the network, A–B and C–D and replacing them with A–C and B–D so long as the degree of the four nodes does not change.

Community subtype modules

For the module analysis, we were interested solely in modules connected via positive interactions, and we therefore excluded negative edges from the detection process.

To extract the densest subpart of the network for module detection, we included members of all "fuzzy maximal cliques". A fuzzy maximal clique consists of a maximal clique of n nodes combined with all nodes that have at least $n*(1-fuzziness)$ edges to the maximal clique. A fuzziness of 0 includes only the maximal cliques, while a fuzziness of 1 includes all nodes connected to the clique. We used a fuzziness of 0.6 in our analyses, which provided the best balance between density of the resulting fuzzy maximal cliques and inclusion of as many nodes as possible.

To detect modules we used seven methods as implemented by the igraph R package version 1.0.1 [Csardi and Nepusz, 2006]. All of these methods required edge weights for edges in the network (which excluded the negative edges), and for which

we used $1 - \rho_{ij}$. Seven community-detection methods were used from the igraph R package (version 1.0.1) [Csardi and Nepusz, 2006] were used: the walktrap algorithm [Pons and Latapy, 2005], the edge-betweenness algorithm [Newman and Girvan, 2004], the fast greedy modularity optimization algorithm [Clauset et al., 2004], information mapping [Rosvall and Bergstrom, 2008], the label propagation algorithm [Raghavan et al., 2007], the leading eigenvector method [Newman, 2006], and the multi-level modularity optimization algorithm [Blondel et al., 2008]. All of the methods aim to find densely connected modules within the networks, but they use slightly different approaches. For visualization and analysis, we used the information mapping algorithm, but all seven gave very similar results for both buccal mucosa and supragingival plaque (**Section S2.1**).

3

Methods for Power Calculation in Human Microbiome Population Studies

Abstract

When conducting large-scale population studies it is crucial to accurately estimate the number of samples needed or the magnitudes of effects which can be detected. Inaccurate estimation can result in an under-powered study, in which many effects of interest are overwhelmed by noise in the data and few biological conclusions can be drawn. However, in microbial epidemiology, guidelines and recommended models for pre-study power analysis have not been fully developed even in the simple two-group case. In order to develop guidelines for conducting power analyses in microbial studies, we compare the analytical power approximations of several commonly used models with the corresponding power of the tests under model mis-specification. This mis-specification captures the zero-inflated and compositional characteristics of real data from microbial sequencing studies. We conclude that a Dirichlet-multinomial model can be accurate for an omnibus test, even under model mis-specification, while a normal approximation to the read counts has the highest power while controlling type I error in even communities. These results provide the first set of guidelines for power analysis in microbial epidemiology studies.

3.1 Introduction

When conducting large-scale population studies it is crucial to properly estimate either the number of samples needed or the magnitudes of effects that can be detected. An accurate power analysis can determine either the number of samples sufficient to detect a certain effect or the range of effect sizes likely to be detected given a certain number of samples. Inaccurate estimation can result in an under-powered study, in which many effects of interest are overwhelmed by noise in the data and few biological conclusions can be drawn. A power analysis typically involves assuming a statistical distribution (such as normal or multinomial) for the data. Under the assumed distribution, the statistical power of the study is typically an analytical function of three factors: the effect size, the number of samples, and the parameters of the distribution.

Guidelines for such analysis have not been developed for microbial community studies, even for relatively simple two-group study designs. This is partly because microbiome data have unique characteristics that make it unclear to what extent prior methodologies are applicable. Microbiome data are highly zero-inflated and compositional, which can make them poor fits for standard distributions such as the normal distribution [McMurdie and Holmes, 2014; Paulson et al., 2013]. Previous work on microbial data power analysis has sometimes adopted the Dirichlet-multinomial distribution [La Rosa et al., 2012]. Alternative methods such as modeling transformed or normalized data using the normal distribution are also used [Morgan, XC and Tickle, TL et al., 2012; Paulson et al., 2013]. None of these methods have been rigorously compared for use in pre-study power analysis in microbiome studies. To evaluate their utility for pre-study power analysis in microbiome data, we use simulated data with zero-inflation and compositionality to compare both the testing power and the analytical power functions of these methods.

Two broad categories of tests can be used to assess differences in microbial communities between two groups, which we label "omnibus" and "per-feature" tests. Omnibus tests ask, "Are the two groups different overall?" and typically compare the overall relative abundance distributions (as the Dirichlet-multinomial [La Rosa et al., 2012]) or use a single measure of the community (such as the first principal component). Conversely, per-feature tests ask, "Is a particular feature different between the two groups?" which compare the abundance distributions for a single feature across the two groups. Per-feature tests involve many more comparisons, which can attenuate power (because most or all of the features are tested); omnibus tests have fewer comparisons at the expense of specific conclusions about the drivers of the difference. Both types of tests have been used to compare groups of microbiome samples, sometimes in conjunction with each other [La Rosa et al., 2012; Segata et al., 2011]. For this reason, we include both types in our evaluations.

We evaluate omnibus and per-feature tests using comprehensive simulations of realistic synthetic data. The simulated data capture the zero-inflation, high-dimensionality, and compositionality of microbial data while maintaining a gold-standard of which features are truly different between the two groups. To understand how well the analytical power function works for microbiome data, we compare the function with the testing power of the different methods. To evaluate which methods are most likely to be of use for a pre-study power analysis, we compare the power and type I error rates of the different methods. We use several models: the Dirichlet-multinomial model for omnibus tests and normal model for per-feature tests. We further assess several normalization and transformation methods in conjunction with a normal distribution. We find that the Dirichlet-multinomial model accurately predicts power in our simulations, and that the normal distribution on the raw read counts has most accurate analytical power function and the highest testing power while controlling the type I error rate.

3.2 Results

3.2.1 Study Overview

To evaluate the accuracy of the analytical power functions of these methods, we compared these functions with against the testing power across a variety of effect sizes. The analytical power function relies on a measure of effect size, so each method had a method-specific effect size measure (e.g., difference in means or difference in scaled means or ratio of means). The simulated data were generated using sparseDOSSA [Ren et al., 2016], which employs a t-statistic-like measure of effect size, and we used this measure as a standard effect size in our simulated datasets. The method-specific effect sizes were functions of the parameters of the sparseDOSSA model (**Section 3.4**). The simulated data were used to evaluate the testing power for a given sample size, while the effect sizes were used for the analytical power function for a given sample size (**Figure 3.1**).

Each method comprised a distribution, normalization, and transformation. A Dirichlet-multinomial distribution was used for omnibus testing, while a normal distribution was used for per-feature testing. For the normal distribution only, we employed several normalization methods and transformations as well as neither normalization nor transformation (**Table 3.1**). The normalization methods included the cumulative-sum-scaling normalization [Paulson et al., 2013] and total-sum-scaling normalization; both are intended to correct for the variability in read depth across samples (**Section 3.4.2**). The transformations comprised the arcsine-square-root transformation, the log transformation, and the logit transformation (**Section 3.4.2**); all three were employed on relative abundance data as a means of ensuring that the data might be a better fit for a normal distribution.

Simulated data were generated using the sparseDOSSA model [Ren et al., 2016], which samples each feature independently from a feature-specific zero-inflated trun-

Table 3.1: **Methods utilized.** A summary of the distributions, along with normalizations and transformations, employed.

Distribution	Normalization	Transformation	Effect size	Baseline parameters
Normal	-	-	$\delta = \mu_1 - \mu_2$	$\sigma = \sigma_1 = \sigma_2$
Normal	CSS	-	$\delta = \mu_1 - \mu_2$	$\sigma = \sigma_1 = \sigma_2$
Normal	TSS	-	$\delta = \mu_1 - \mu_2$	$\sigma = \sigma_1 = \sigma_2$
Normal	TSS	arcsine-square-root	$\delta = \mu_1 - \mu_2$	$\sigma = \sigma_1 = \sigma_2$
Normal	TSS	log	$\delta = \mu_1 - \mu_2$	$\sigma = \sigma_1 = \sigma_2$
Normal	TSS	logit	$\delta = \mu_1 - \mu_2$	$\sigma = \sigma_1 = \sigma_2$
Dirichlet-multinomial	-	-	ϕ_{ssd}	$\theta = \theta_1 = \theta_2$

cated log-normal distribution. The distributional parameters for each feature were determined by the feature-specific log-mean (**Section 3.4.1**), and the means generated using the sparseDOSSA R package defaults [Ren et al., 2016]. A certain percentage of features were randomly sampled and given a common t-statistic-like effect size for each run. We generated datasets combinatorially with numbers of features P either 100 or 200; percentage of spiked features $\in \{\frac{1}{P}, 0.02, 0.05, 1\}$; spike strengths $\in \{1, 5, 10, 20\}$; and numbers of samples $\in \{50, 100, 200\}$. For each combination of parameters, we simulated 500 i.i.d. datasets in order to obtain accurate estimates of error rates and testing powerfor each method. This model is distinct from the distributions we assume for the power analysis, allowing each method to be assessed for robustness to deviation from the assumed model.

3.2.2 Overview of Performance

The Dirichlet-multinomial model displays the expected shape for a power curve (**Figure 3.2**). The Dirichlet-multinomial power is also accurately predicted by certain parameter values, indicating that using realistic estimates of the overdispersion parameter θ can accurately inform predictive power before the commencement of data collection.

All of the per-feature tests have accurately predicted power for some parameter values (**Figure 3.2**). The choice of parameter value should be informed by the

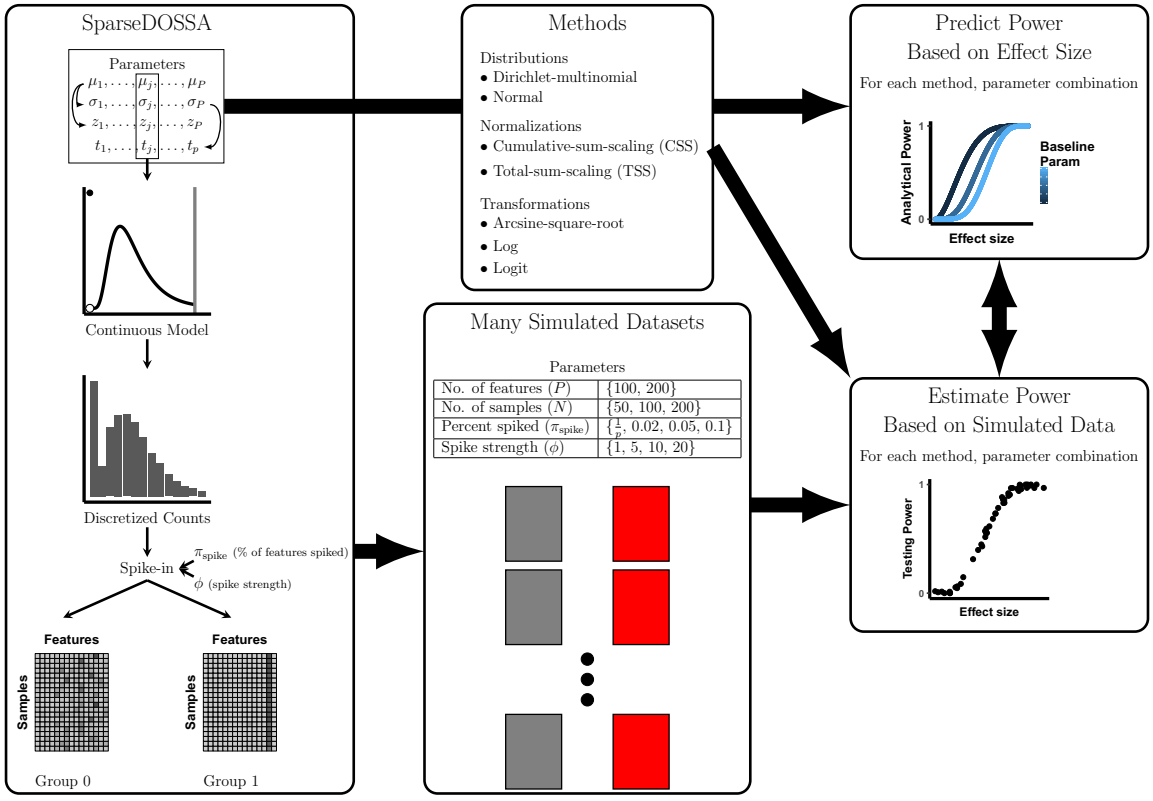

Figure 3.1: **Overview of study.** SparseDOSSA was used to generate realistic datasets with varying spike-in strengths. The model parameters were used to determine model-specific effect sizes for use in the analytical power functions for each method. The simulated datasets were used to evaluate testing power and calculate type I error rates.

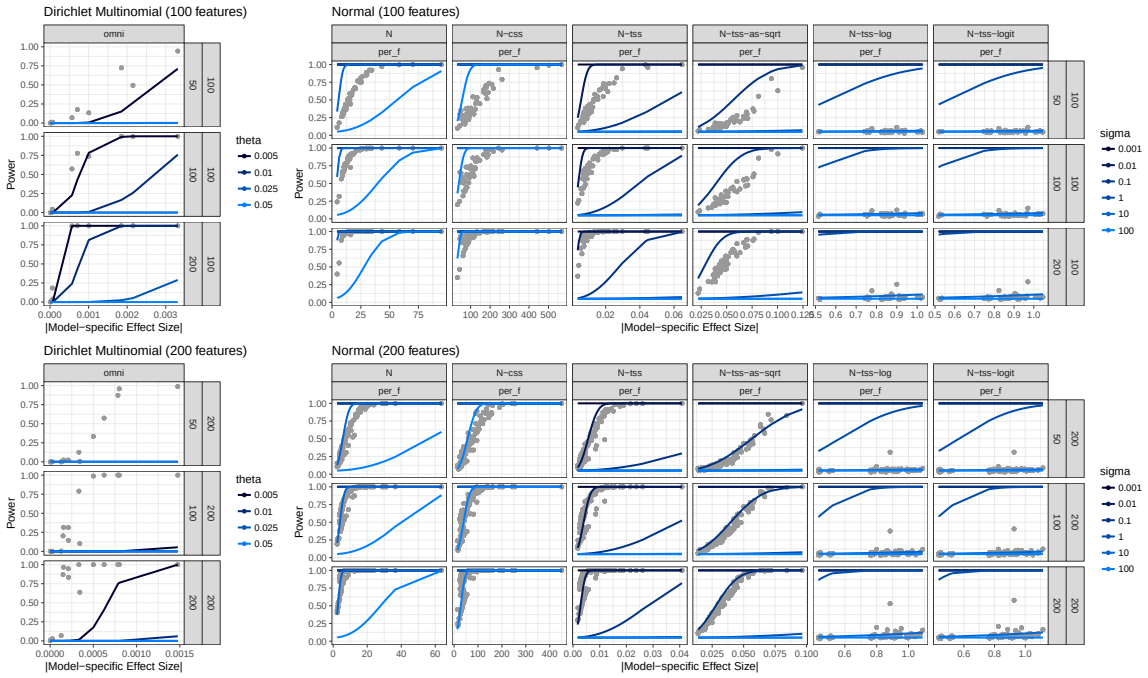

Figure 3.2: **Testing power compared with the analytical power function for each method.** Power of the tests (gray points) and the analytical power function curves (colored lines) across the method-specific effect sizes calculated from the sparseDOSSA parameters. The analytical power functions are shown for a range of different baseline parameters for each model. The fact that the testing power points fall between the analytical function curves in most scenarios indicates that the analytical power function would agree well with the testing power for some value of the baseline parameter.

scale of the data at hand. For example, a parameter value between 0.01 and 0.1 makes most sense for total-sum-scaled data, as the data is constrained between 0 and 1. By contrast, a parameter value of around 100 is a better choice for log- or logit-transformed total-sum-scaled data because the log-transformation amplifies the variability in the data. The low power of the log- and logit-transformed data is a result of this high variability relative to the effect size (measured as the difference in means). Count or cumulative-sum-scaled count data fall somewhere in the middle in terms of the parameter values that yield accurate power predictions.

3.2.3 Type I error rates

While most methods have good agreement between the analytical power function (for some baseline parameter values) and the testing power, each method must also control the type I error rate to the nominal level (in this case, 0.05). We evaluated this for the per-feature methods by looking at the false positive rates for all null features across the 500 datasets (**Figure 3.3**). The type I error rate is controlled for the normal distribution when read counts and log- or logit-transformed total-sum-scaled data is used. By contrast, cumulative-sum-scaled data, untransformed total-sum-scaled data, and arcsine-square-root-transformed total-sum-scaled data have much higher average false positive rates than the nominal level; this is especially evident when looking at the distribution of type I error rates across all features (**Figure S3.1**), showing that some features have particularly high type I error rates for these methods.

3.2.4 Effect Sizes

SparseDOSSA specifies a t-statistic-like effect size on the count level as a mean-difference on zero-mean, unit-variance data subsequently scaled. This effect size is most efficient to generate simulated data with pre-specified metadata interactions but does not directly correspond to any of the effect sizes of the various methods that we test (**Section 3.4**). To ensure that the effect sizes for each of the methods are comparable, we elected to pre-specify the sparseDOSSA effect size and then use a method-specific effect size derived from the parameters of the sparseDOSSA model. Briefly, for the omnibus Dirichlet-multinomial test we used the scaled sum-of-squared-differences between the normalized feature means, while for the per-feature tests we used the per-feature difference in the normalized and transformed feature means (**Section 3.4**). To understand how these relate to one another, and also the interpretation of the sparseDOSSA effect size in real data, we compared these model-specific effect sizes with the sparseDOSSA effect sizes.

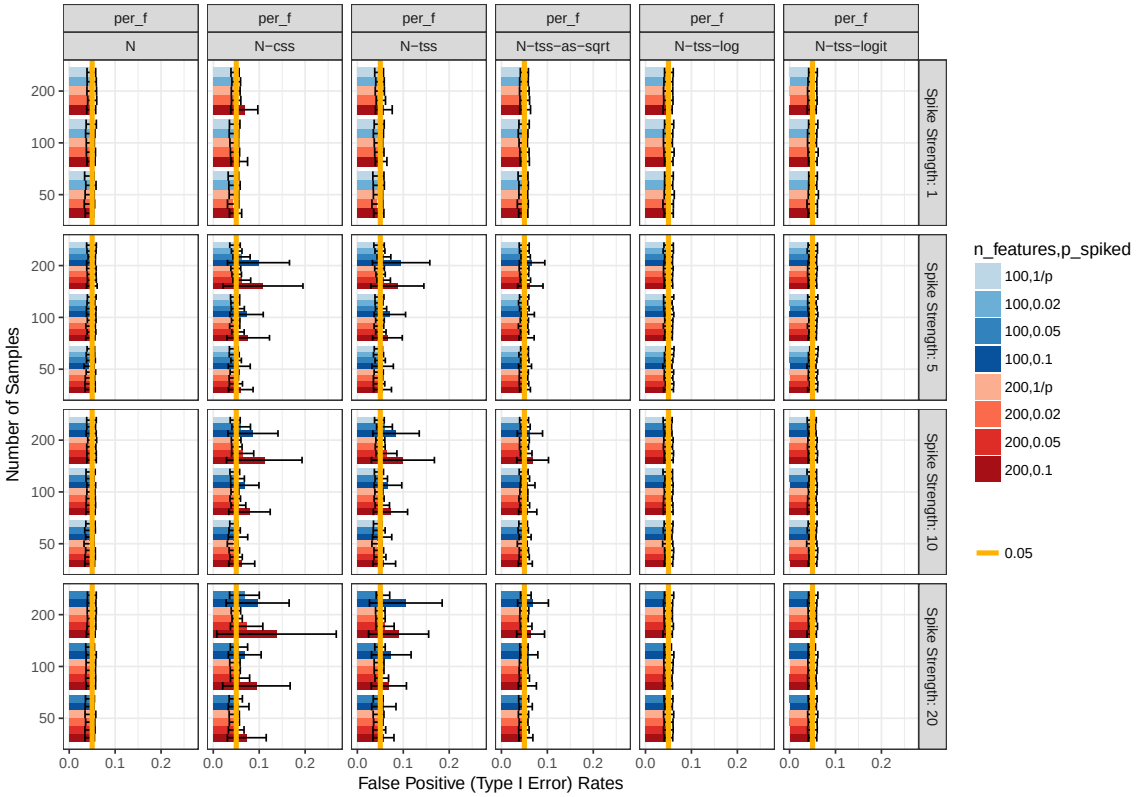

Figure 3.3: **Average type I error rates.** The type I error rates of the different methods across all simulation parameters. Bar heights are the mean type I error rates (measured using 500 datasets) across all null features. Error bars are standard deviations.

All of the tests display an approximately monotonic pattern, with larger sparseDOSSA effect sizes corresponding to larger model-specific effect sizes (**Figure 3.4**). The slight non-monotonicity is in all cases a result of sparseDOSSA's random choice of features to spike in, which results in some effect sizes being stronger depending on the characteristics of the spiked-in features. Averaging model-specific effect sizes over multiple different spike-ins should reveal a strictly monotone pattern in expectation. Interestingly, the log- and logit-transformed relative abundance methods have the most monotonic relationship.

3.3 Conclusions

In this paper, we present the groundwork for the first guidelines regarding power analysis techniques in two-group microbiome analyses. We use simulations to evaluate several tools in data with similar characteristics to real microbiome data: zero-inflated, compositional, and high-dimensional. For general omnibus tests, we recommend using a Dirichlet-multinomial model, which had high agreement between the analytical power function and the power of the test. We find that for per-feature tests, in most cases a simple normal approximation after appropriate normalization and transformation of the data also has high agreement between the analytical power function and the testing power.

There are a few limitations of this study. Most glaringly, we have only used a single distribution to simulate the data used to evaluate these methods. All models are to some degree mis-specified because the data generating model is not the same as any of the methods' distributions. The degree to which this mis-specification matters for each model is unclear, and the log-normal assumption might agree more closely with the generating model of some distributions than of others. Alternative distributions (including generative distributions from some of the models) could be explored to assess the robustness of our results. Further, the extent to which the

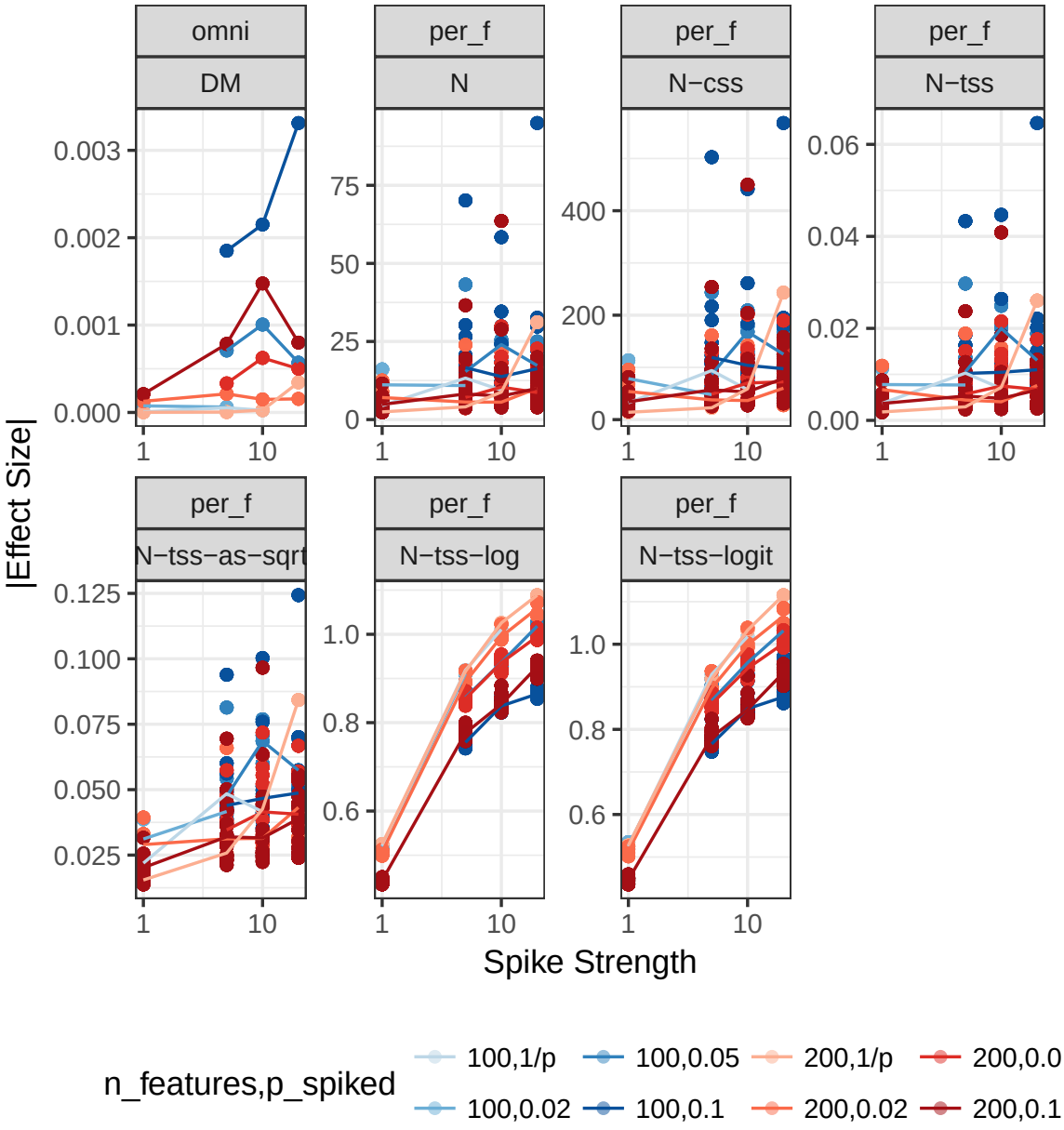

Figure 3.4: **SparseDOSSA effect sizes versus method-specific effect sizes.** Method-specific effect sizes compared with the effect size specified through sparseDOSSA. The method-specific effect sizes are based on the true means for each feature in the two groups. For the per-feature tests, the line indicates the median effect size across all spiked-in features.

effect sizes from this simulation study are an accurate reflection of real data is not yet clear.

Immediate future work will include comparisons with real datasets, more simulated datasets, and methods from other fields. Including real microbial data will allow us to evaluate how our assumptions and recommendations translate into practice. Simulating a larger number of simulated datasets of varying ecologies assesses the robustness of our conclusions across multiple microbial habitats. RNA-seq data also displays high zero-inflation rates and high-dimensionality; in this context, negative binomial and Poisson models have been proposed for analysis and power calculations [Aban et al., 2008; Fang and Cui, 2011; Ng and Tang, 2005]. Including these models in our analysis will allow for a more comprehensive recommendations.

This study lays the groundwork for microbiome power analysis by developing accurate methods for power analysis in a two-group microbial analysis setting. Many microbial studies utilize more complicated designs, such as multi-group or longitudinal data [Morgan, XC and Tickle, TL et al., 2012; Vatanen et al., 2016]. Future work will extend our methods to these settings in order to provide comprehensive recommendations for the majority of microbial analyses. In addition, these methods do not take into account false discovery rate corrections, which are commonly employed in microbial studies because of the large number of tests (particularly when doing per-feature analysis). In practice, power analyses often use the family-wise error rate for the level of the test, but some methods have been proposed to allow the less stringent false discovery rate to be accounted for [Jung, 2005; Liu and Hwang, 2007; Müller et al., 2004]. Extending some of these methods to power analysis in microbiome data will allow a more typical study design to be accounted for in advance.

3.4 Methods

3.4.1 SparseDOSSA

Realistic data were generated using the SparseDOSSA model [Ren et al., in review], which generates each feature from a zero-inflated, truncated log-normal distribution with subsequent rounding and estimates the feature-specific parameters by fitting to a given real-world template dataset. Concretely, each feature X_j is sampled from a zero-inflated, truncated log-normal distribution with zero-inflation parameter z_j, truncation point t_j, $E(log(X_j)) = \mu_j$, and $Var(log(X_j)) = \sigma_j^2$. Given a feature-specific log-mean μ_j, the zero-inflation parameter, truncation point, and log-variance σ_j^2 are computed as functions of μ_j based on relationships across all P taxa. Associations are generated replacing the observed counts with a scaled function of the covariates [Ren et al., in review]. After sampling and spiking, the samples are rounded to the nearest integer to produce count data.

The sparseDOSSA parameters in conjunction with the spike-in procedure can be used to derive the expectations for the features in each of the two groups using well-known properties of expectations. Briefly, if $F(x, \mu, \sigma)$ is the cumulative density function of a log-normal distribution with parmeters μ and σ, the null features have the expectation of $E(X_j|K = 0) = E(X_j|K = 1) = \frac{1-z_j}{F(t_j,\mu_j,\sigma_j)}\Phi\left(\frac{log(t_j)-\mu_j-\sigma_j^2}{\sigma_j}\right)e^{\mu_j+0.5\sigma_j^2}$ (where Φ is the cdf of a standard normal distribution), while the spiked features have the expectation of $E(X_j|K) = E(X_j) + \frac{(1-z_j)\psi}{1+\psi}(\sqrt{Var(X_j|X_j > 0)})\left(\frac{(K-E(K)}{\sqrt{Var(K)}}\right)$. The complete derivations are in **Section S3.1**.

3.4.2 Methods implemented

Dirichlet-multinomial distribution

The Dirichlet-multinomial test was implemented using the HMP R package, version 1.4.3 [La Rosa et al., 2012]. This assumes that the counts for each group k arise

from a Dirichlet-multinomial distribution parameterized with the relative abundance vector π_k and over-dispersion parameter θ:

$$f(X_{ik}|\theta_k, \pi_k, R_{ik}) = \binom{R_{ik}}{x_{ik}} \frac{\prod_{j=1}^{P} \prod_{r=1}^{x_{ikj}} \pi_{jk}(1-\theta_k) + (r-1)\theta_k}{\prod_{r=1}^{R_{ik}}(1-\theta_k) + \theta_k(r-1)}$$

They test $\mathrm{H}_0 : \pi_0 = \pi_1$ against $\mathrm{H}_\mathrm{a} : \pi_0 \neq \pi_1$ using a Wald-type chi-squared test with test statistic:

$$X^2(\pi_0, \pi_1, \theta_0, \theta_1, R_0, R_1) = (\hat{\pi}_0 - \hat{\pi}_1)^T S^{-1} (\hat{\pi}_0 - \hat{\pi}_1).$$

The authors suggest an effect size $\phi = \frac{X^2(\pi_0,\pi_1,\theta_0,\theta_1,R_0,R_1)}{X^2([0,1],[1,0],\theta_0,\theta_1,R_0,R_1)}$, where $X^2([0,1],[1,0],\theta_0, \theta_1, R_0, R_1)$ is the maximal possible test statistic for comparing two groups. However, this does not have an analogue in non-Dirichlet-multinomial data because it requires true (or estimated) values for θ_k. We therefore used instead the scaled sum of squared differences between the normalized sparseDOSSA expectations:

$$\phi_{ssd} = \frac{1}{2} \sum_{j=1}^{P} \left(\frac{E(X_j|K=0)}{\sum_i E(X_i|K=0)} - \frac{E(X_j|K=1)}{\sum_i E(X_i|K=1)} \right)^2.$$

The power is calculated using a simulation-based approach implemented in the R package. In order to force the power to be a function of the effect size ϕ_{ssd}, we assumed that the read depths would be the same for all samples in the two groups. Based on P, the read depths, the sample sizes, and ϕ_{ssd}, we generated π_0 and π_1 with effect size ϕ_{ssd} based on an iterative procedure as follows. We first assumed that $\pi_0 = [\sqrt{2\phi_{ssd}}, \frac{1-\sqrt{2\phi_{ssd}}}{P-1}, \ldots, \frac{1-\sqrt{2\phi_{ssd}}}{P-1}]$ and that $\pi_1 = [2\sqrt{2\phi_{ssd}}, \frac{1-2\sqrt{2\phi_{ssd}}}{P-1}, \ldots, \frac{1-2\sqrt{2\phi_{ssd}}}{P-1}]$. At each iteration, if $\phi_{ssd}(\pi_0, \pi_1, R_0, R_1) > \phi$, we subtracted $|\phi(\pi_0, \pi_1, \theta, R_0, R_1) - \phi|^{7/8}$ from the first element of π_1 and added $\frac{|\phi(\pi_0,\pi_1,R_0,R_1)-\phi|^{7/8}}{P-1}$ to each of the remaining elements. If $\phi(\pi_0, \pi_1, R_0, R_1) < \phi$, then the process was reversed. We iterated until $\phi(\pi_0, \pi_1, R_0, R_1)$ was within 1e-10 of ϕ. We then used these calculated π_0 and π_1 values, along with the given read depths, in the R package approach.

Normal distribution

The data for each group is assumed to be normally distributed:

$$X_{ik} \sim N(\mu_k, \sigma_k)$$

We tested $H_0 : \mu_1 = \mu_0$ against $H_a : \mu_1 \neq \mu_0$ using a t-test, as implemented in R. The effect size is given by $\delta = \mu_1 - \mu_0$. The power is calculated using the standard two-sided formula, and assuming that $\sigma_0 = \sigma_1 = \sigma$.

Normalizations

The two normalizations employed were cumulative-sum scaling and total sum scaling. Cumulative-sum scaling was used as implemented in the metagenomeSeq package, version 1.12.1 [Paulson et al., 2013]. Total sum scaling consisted in dividing every count by the total counts for that sample to produce relative abundances.

The effect sizes for each normalization were calculated by applying the normalizations on the sparseDOSSA expectations. That is, for no normalization, the effect size for each feature was $E(X_j|K = 1) - E(X_j|K = 0)$; for CSS normalization, the effect size for each feature was $CSS\left([E(X_1|K = 1), \ldots, E(X_P|K = 1)]\right)_j - CSS\left([E(X_1|K = 0), \ldots, E(X_P|K = 0)]\right)_j$; for untransformed TSS normalization, the effect size for each feature was $\frac{E(X_j|K=1)}{\sum_i E(X_i|K=1)} - \frac{E(X_j|K=0)}{\sum_i E(X_i|K=0)}$.

Transformations

Three transformations were used: arcsine-square-root, log, and logit. These were only employed with total-sum-scaled data. For the log and logit transformations, any zero values were replaced with 1e-7; for the logit transformation, any values of 1 were replaced with 1-1e-7.

As for the normalizations, the method-specific effect sizes for the transformed data used the transformation on the normalized feature expectations. If we let $\pi_{jk} = \frac{E(X_j|K=k)}{\sum_i E(X_i|K=k)}$, then the effect size for each transformation τ was $\tau(\pi_{j1}) - \tau(\pi_{j0})$.

Discussion

In this dissertation, I have addressed the statistical challenges of inferring correlations from compositional data and the choice of methods for performing pre-study power analysis. I developed a Bayesian model with a GLASSO penalty to appropriately account for the compositional constraint while inferring the correlations between the unconstrained counts that generate the compositions. This framework provides both accurate estimates and uncertainty quantification for the resulting correlations. Using simulation studies, I find that the normal model on the raw read counts is most effective for performing pre-study power analysis when comparing each feature separately, while a dirichlet-multinomial model is effective for such analysis when comparing both groups overall.

This work moves the field of microbiome studies forward in several important ways. Firstly, the ability to use compositional sequencing data to infer interactions is key to understanding the large-scale ecology of the microbiome in its community context. Secondly, using this analysis method to infer correlations between species in the healthy human microbiome provides a foundation for understanding the response of the microbial community as a whole to perturbations such as antibiotics or dietary changes as well as offering potential targets for more in-depth mechanistic studies of interactions. Lastly, microbial epidemiologists can now use tools for pre-study power analysis whose accuracy has been evaluated in data with microbial characteristics, including zero-inflation and compositionality. Accurate power analysis is key to conducting efficient studies and indispensable for the FDA approval process for

any microbial-based therapy. These concrete contributions therefore further the goal of developing ways to cure disease by manipulating the microbiome.

Appendix for Chapter 1

S1.3 Detailed likelihood derivation

For a single observation, let $\boldsymbol{x} = (x_1, \ldots, x_p)$ be the basis and $t = \sum_{j=1}^{p} x_j$ be the total. Under our model, $\boldsymbol{x} \sim \mathcal{LN}(\boldsymbol{m}, \boldsymbol{O})$. The normalized composition is $\boldsymbol{c} = (c_1, \ldots, c_p))$, where $c_j = \frac{x_j}{t}$.

The density of $\boldsymbol{x}$ is given by:

$$f(\boldsymbol{x}) = \frac{exp\left\{\frac{-1}{2}\left[(\log(\boldsymbol{x}) - \boldsymbol{m})^T \boldsymbol{O} (\log(\boldsymbol{x}) - \boldsymbol{m})\right]\right\}}{(2\pi)^{p/2} |\boldsymbol{O}|^{-1/2} \prod_{j=1}^{p} x_j}$$

Note that $\boldsymbol{x}$ can be alternatively represented by $\boldsymbol{y} = g(\boldsymbol{x}) = (t, c_1, \ldots, c_{p-1})$, with inverse transformation $g^{-1}(\boldsymbol{y}) = \left(c_1 t, \ldots c_{p-1} t, \left(1 - \sum_{j=1}^{p-1} c_j\right) t\right)$ and determinant of the Jacobian t^{p-1}. The density of $\boldsymbol{y}$ is then:

$$\begin{aligned} f(\boldsymbol{y}) &= f(g^{-1}(\boldsymbol{y}))|J| \\ &= \frac{exp\left\{\frac{-1}{2}\left[(\log(t\boldsymbol{c}) - \boldsymbol{m})^T \boldsymbol{O} (\log(t\boldsymbol{c}) - \boldsymbol{m})\right]\right\} t^{p-1}}{(2\pi)^{p/2} |\boldsymbol{O}|^{-1/2} \prod_{j=1}^{p} t c_j} \\ &= \frac{exp\left\{\frac{-1}{2}\left[(\log(t\boldsymbol{c}) - \boldsymbol{m})^T \boldsymbol{O} (\log(t\boldsymbol{c}) - \boldsymbol{m})\right]\right\}}{(2\pi)^{p/2} |\boldsymbol{O}|^{-1/2} t \prod_{j=1}^{p} c_j} \end{aligned}$$

S1.4 Detailed description of simulated datasets

Small Datasets for Challenging Scenarios

Four challenging scenarios tested BAnOCC's performance on "edge cases" with strong spurious correlations. Given a fixed correlation matrix ($\boldsymbol{R}_{\log \boldsymbol{X}}$), the log-basis means ($m_j$) and standard deviations (s_j) were sampled from scenario-specific parameter generating distributions. Nine features and 1,000 samples formed each dataset; the small number of features made the correlation difficult to infer in all cases. Two scenarios were "null" in having no basis correlations, while two were "spiked" in having at least one true basis correlation.

No basis correlations were present for two of the scenarios, which differed by the presence of a negative-dominant type spurious correlation. The "simple" scenario had parameter generating distributions with small variance (**Figure S1.1**) which meant that the spurious correlations present were due primarily to the small number of features. The "high spurious" scenario had a strong negative-dominant type correlation induced by combining a parameter generating distribution for m_j with small mean and variance with a parameter generating distribution for s_j with large mean and variance (**Figure S1.2**). This meant that the basis means ($\mu_{X,j} = e^{m_j+0.5s_j^2}$) and variances ($\sigma^2_{X,j} = \mu^2_{X,j}(e^{s_j^2-1})$) were determined by the value of s_j so that features with high $\mu_{X,j}$ would also have large $\sigma^2_{X,j}$, inducing a negative-dominant type spurious correlation.

Several true correlations were present in the remaining scenarios, which were again distinguished by the presence or absence of a negative-dominant type spurious correlation. The "retained spike" scenario had the same parameter generating distributions as the "simple" scenario (**Figure S1.3**), while the "reversed spike" scenario had the same parameter generating distributions as the "high spurious" scenario (**Figure S1.4**). Because the spurious correlations in the "retained spike" scenario were due primarily to the small numbers of features, the magnitude and direction of the true associations were approximately retained between the log-basis and the composition. The features in the "reversed spike" scenario were ordered such that the negative dominant spurious correlation was between two features with a positive true correlation; this caused the correlation to reverse direction when comparing the log-basis and the composition.

Realistic Data for Performance Comparison

To understand how BAnOCC performs in practice and would compare with other methods, we used more realistic simulated data generated by the sparseDOSSA soft-

ware, version 1.1.0 [Ren et al., 2016], which generates synthetic data that closely resemble datasets from microbial ecology. Briefly, it models each feature as a zero-inflated, truncated log-normal distribution with subsequent rounding and estimates the feature-specific parameters using a template dataset. This generative model differs from our assumed model by the incorporation of truncation, zero inflation, and rounding. These differences make the simulated data more realistic and also allow us to assess the robustness of BAnOCC to model violations.

We induced correlations between the features by setting off-diagonal elements of $\boldsymbol{R}_{\log \boldsymbol{X}}$, the log-basis correlation, to non-zero values. By default, sparseDOSSA assumes that the features are uncorrelated. We added at most $p/2$ correlations by randomly selecting several elements of $\boldsymbol{R}_{\log \boldsymbol{X}}$ to be non-zero. The apparent value of the correlation in the simulated data is somewhat attenuated from the specified value due to the truncation, zero inflation, and rounding. All correlations were given the same value, and we used four different correlation strengths, $\{-0.7, -0.3, 0.3, 0.7\}$, to compare performance on positive or negative, weak or strong correlations.

For sparseDOSSA calibration, we chose a template dataset that was likely to have strong compositional effects. We used a vaginal dataset from the human microbiome project [The Human Microbiome Consortium, 2012], which has few features with one tending to dominate the others. Ecologically, the vagina is typically dominated by *Lactobacillus* species with few others present at low levels. For each template, we simulated 100 datasets with 100 samples and 14 features.

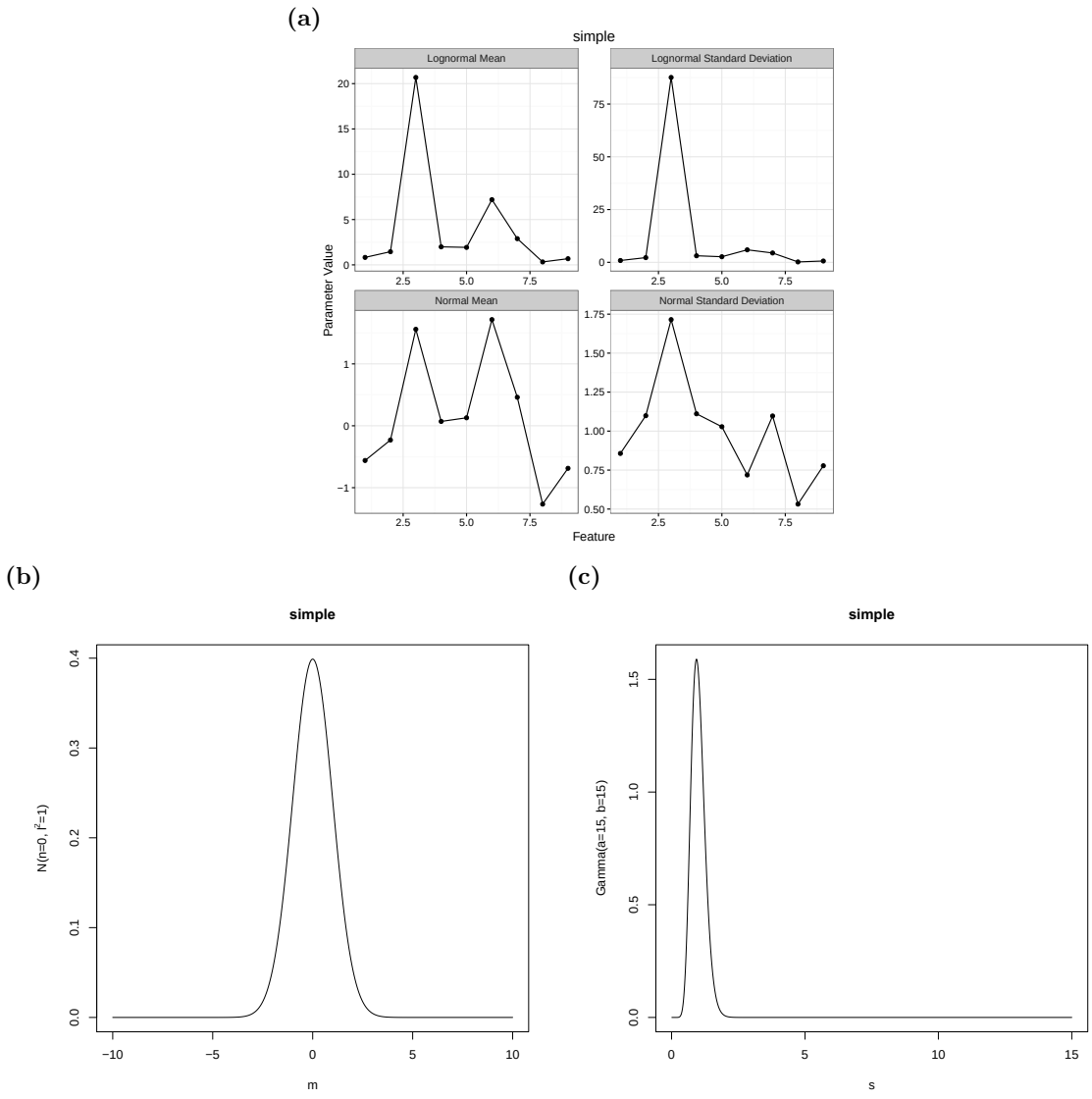

Figure S1.1: **Parameters and parameter-generating distributions for the "simple" simulation scenario.** The parameters for the "simple" scenario were pretty similar to each other on the normal scale (**A**) because the parameter-generating distributions had small variance (**C**)-(**D**). This meant that if the number of features was large, the spurious correlation should not be very strong.

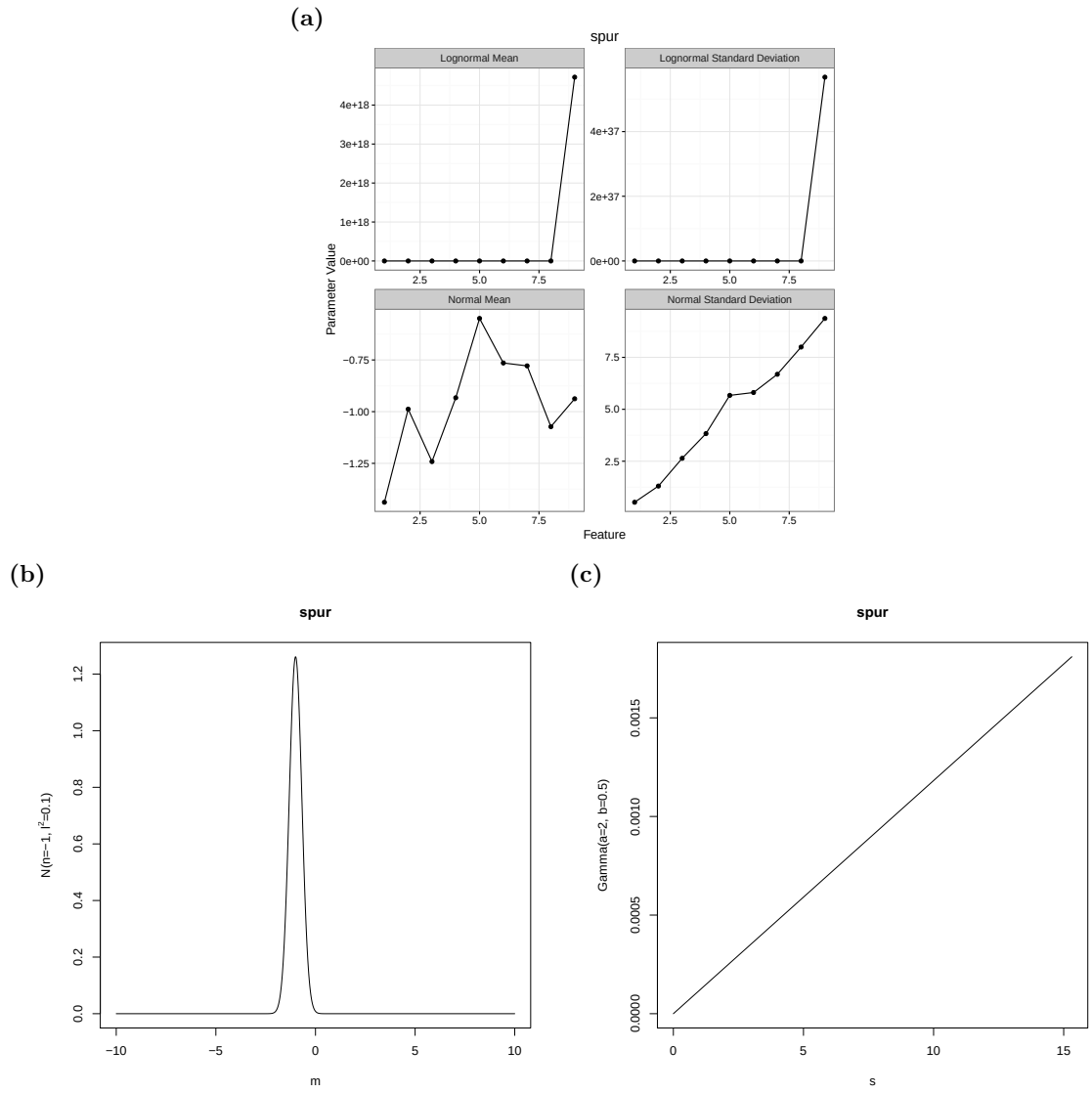

Figure S1.2: **Parameters and parameter-generating distributions for the "high spurious" simulation scenario.** The "high spurious" scenario had a negative dominant correlation resulting from a positive relationship between the basis (lognormal) means and variances (**A**). This resulted from a parameter-generating distribution for m_j with small mean and variance combined with a parameter-generating distribution for s_j with large mean and variance (**B**)-(**C**). This implied that the basis mean ($\mu_{X,j} = e^{m_j + \frac{1}{2}s_j^2}$) and variance ($\sigma_{X,j} = \mu_{X,j}^2(e^{s_j^2} - 1)$) were determined by the value of s_j.

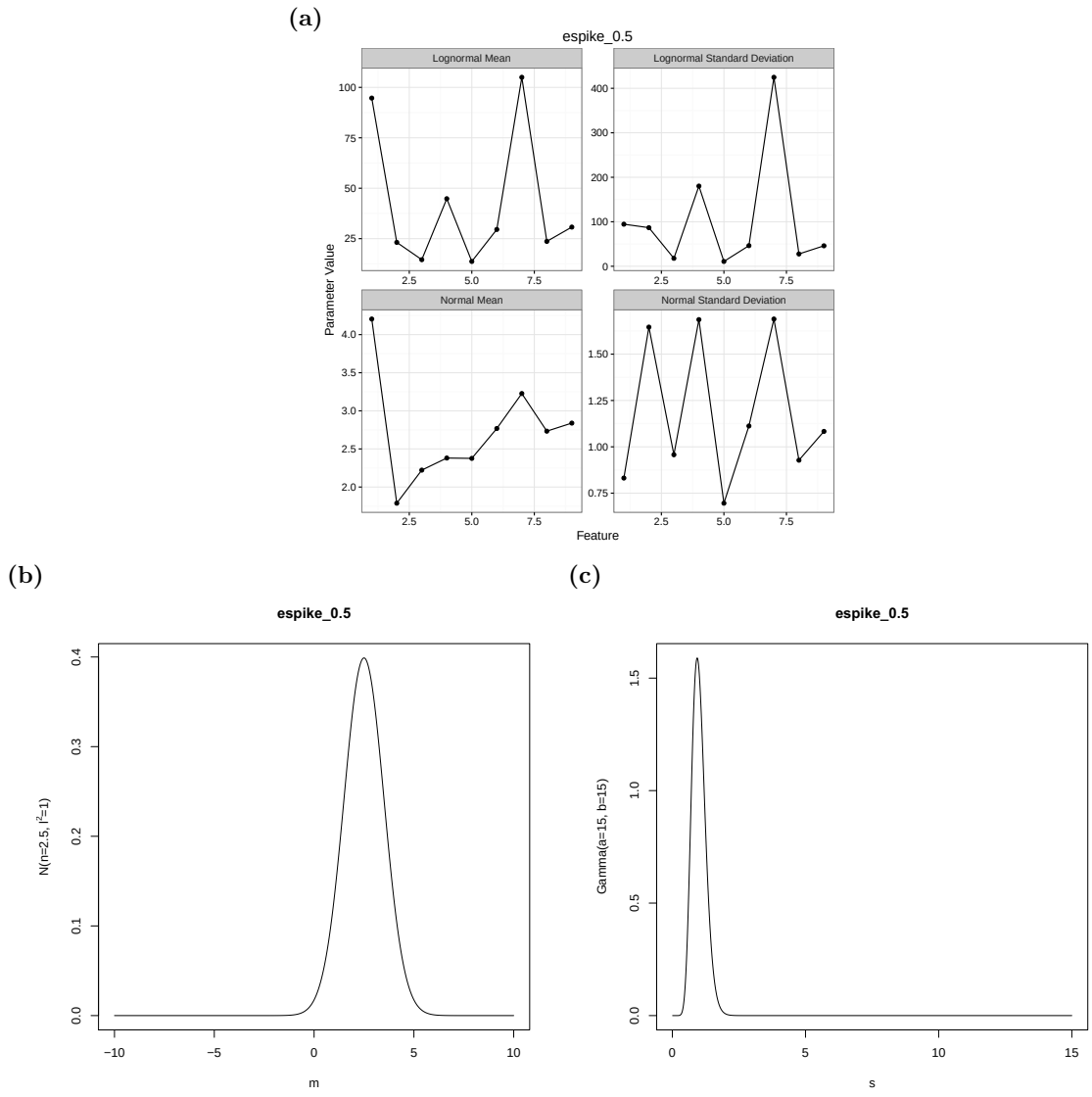

Figure S1.3: **Parameters and parameter-generating distributions for the "retained spike" simulation scenario.** The features in the "retained spike" scenario had parameter values very similar to each other (**A**) as a result of the small variance of the parameter-generating distributions (**C**)-(**D**).

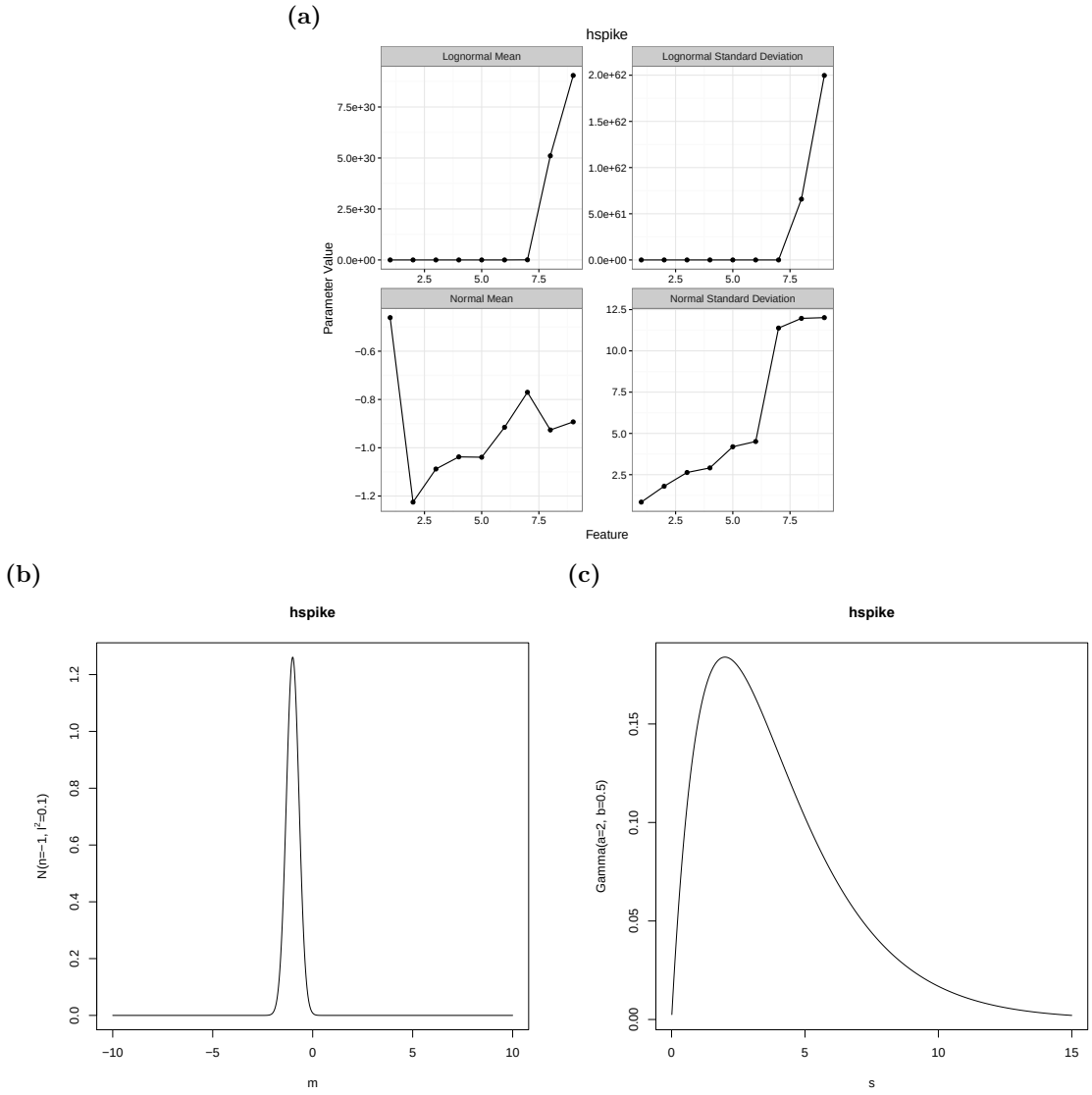

Figure S1.4: **Parameters and parameter-generating distributions for the "reversed spike" simulation scenario.** The "reversed spike" scenario had a negative dominant spurious correlation introduced due to the fact that the two features with the highest mean also had the highest variance (**A**). This resulted from very low variance in the log-basis mean (m_j) generating distribution but high variance in the log-basis standard deviation (s_j) generating distribution (**B**)-(**C**). This implied that the log-basis mean ($\mu_{X,j} = e^{m_j + \frac{1}{2}s_j^2}$) and log-basis variance ($\sigma_{X,j} = \mu_{X,j}^2(e^{s_j^2} - 1)$) were determined primarily by the value of s_j and therefore positively correlated.

S1.5 Implementation of methods compared

It is important to remember that H_0: $\rho_{X,jk} = 0$ is equivalent to H_0: $\rho_{\log X,jk} = 0$ when $\boldsymbol{X}$ follow a log-normal distribution.

Simplicial Variation

Simplicial variation [Aitchison, 2003] is based on the variance of logrations. Specifically, the statistic comparing features j and k is the variance of log-ratios

$$t_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\log\frac{c_{ij}}{c_{ik}} - \overline{\log\frac{c_{ij}}{c_{ik}}}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\log\frac{x_{ij}}{x_{ik}} - \overline{\log\frac{x_{ij}}{x_{ik}}}\right)^2$$

It can be seen that t_{jk} will be zero if $X_{ij} \propto X_{ik}$ for $i = 1, \ldots, n$, and in this case features j and k would be perfectly correlated.

To test whether the basis correlation between features j and k ($\rho_{X,jk}$) was zero, we used a one-sided permutation test on $\frac{1}{t_{jk}}$, which should be large if X_j and X_k are correlated. This test is approximate, as after sample permutation, $\frac{c_{i_1 j}}{c_{i_2 k}}$ is not necessarily equal to $\frac{x_{i_1 j}}{x_{i_2 k}}$.

The permutation p-value was used to determine type I and type II error rates. A positive result was obtained if the p-value was less than the test level (0.05), and these were compared with the true zero or non-zero values of the correlation matrix.

sparCC

sparCC [Friedman and Alm, 2012] relies on the simplicial variation statistic, but uses an assumption about $\sum_{i=1}^{p}\sum_{j=i+1}^{p} w_{jk}$ to impose sparsity and estimate w_{jk} through a system of iteration.

We used the authors' software from `https://bitbucket/yonatanf/sparcc`, downloaded on December 14, 2015. The authors propose a bootstrap based method for inference, and we used this method with the software's default parameters of 100 bootstrap datasets and two-sided p-values. The type I and type II errors were deter-

mined based on the correct or incorrect rejection of H_0: $w_{jk} = 0$ where rejection was determined by a two-sided p-value of less than 0.05.

CCLasso

CCLasso [Fang et al., 2015] uses a LASSO penalty on the off-diagonal elements of $\boldsymbol{R}_{\log \boldsymbol{X}}^{-1}$ with a loss function of

$$\text{LOSS}(\boldsymbol{R}_{\log \boldsymbol{X}}) = \frac{1}{2}\text{tr}\left[\left\{\boldsymbol{G}\left(\boldsymbol{R}_{\log \boldsymbol{X}} - \widehat{\boldsymbol{R}_C}\right)\boldsymbol{G}^T\right\} diag\left(\boldsymbol{G}\widehat{\boldsymbol{R}_C}\boldsymbol{G}^T\right)^{-1} \left\{\boldsymbol{G}\left(\boldsymbol{R}_{\log \boldsymbol{X}} - \widehat{\boldsymbol{R}_C}\right)\boldsymbol{G}^T\right\}\right],$$

where $\boldsymbol{G} = \boldsymbol{I} - \frac{1}{p}\boldsymbol{1}\boldsymbol{1}^T$, which corresponds to a centered log-ratio transformation of $\boldsymbol{C}_i$. The tuning parameter λ is chosen using k-fold cross validation. We downloaded the authors' software from `https://github.com/huayingfang/CCLasso` on December 10, 2015. We used the default parameters of 3-fold cross-validation, a tuning parameter interval of $[0.0001, 1]$, and a maximum number of selection iterations of 20.

Because CCLasso is LASSO-based, there is no accompanying inference method. We determined type I and type II errors as the authors do in their paper: by correct or incorrect estimation of w_{jk} to be exactly zero. That is, a positive result was w_{jk} estimated to be non-zero.

SPIEC-EASI

SPIEC-EASI [Kurtz et al., 2015] estimates $\boldsymbol{R}_{\log \boldsymbol{X}}^{-1}$ using either the neighborhood selection method of [Meinshausen and Bühlmann, 2006], or the graphical LASSO method of [Friedman et al., 2008]. The tuning parameter selection is done using stability selection via the StARS algorithm of [Liu et al., 2010]. We downloaded the authors' software from `https://github.com/zdk123/SpiecEasi` on June 7, 2016. We used a minimum lambda ratio of 0.01, 500 subsamplings for stability selection, and 100 lambda values.

SPIEC-EASI is also LASSO-based, and thus provided no inference method. As for CCLasso (above), we determined type I and type II errors by correct or incorrect estimation of w_{jk} to be exactly zero. Our simulated datasets were structured such that if w_{jk} was zero, w_{jk}^{-1} would also be zero, so this approach was appropriate.

Spearman Correlation

Spearman correlation served as a "naive" estimate of $\boldsymbol{R}_X$ that does not take into account the compositional structure of the data. We used a two-sided permutation test with 1000 permutations to evaluate significance, and determined type I and type II errors by correct or incorrect rejection of H_0: $\rho_{X,jk} = 0$.

We applied Spearman correlation twice: once on the simulated basis counts (which, of course, are unobserved in practice) and once on the simulated compositions. The former should have controlled type I error rates and good power because Spearman correlation would be appropriate if the data are unconstrained; the latter should have very high type I error rates because the compositions are constrained.

BAnOCC

BAnOCC samples the posterior using No-U-Turn Sampling (NUTS) [Hoffman and Gelman, 2014], a Hamiltonian Monte Carlo (HMC)-based algorithm, as provided in the rstan R package [Stan Development Team, 2014]. For all datasets, both simulated and real we started with 1000 iterations of warmup (necessary to choose the appropriate step size for HMC), and 4000 iterations of sampling for each of at least three chains. We evaluated convergence of the chains using the R-hat statistic [Gelman and Rubin, 1992] and increased the number of iterations until the R-hat statistics for all sampled parameters were less than 1.1.

S1.6 Supplemental data

Data S1.1: **Simulated data for difficult scenarios.** The simulated data for each of four difficult simulation scenarios described (**Section 1.3**; details in **Section S1.4**).

Data S1.2: **Realistic simulated data.** All simulated datasets from sparseDOSSA, as well as the template dataset used. For details on how these were generated, see **Section S1.4**.

Data S1.3: **HMP taxonomic profiles.** The taxonomic profiles from the Human Microbiome Project data as processed with MetaPhlAn version 2.0_beta1 [Truong et al., 2015].

S1.7 Supplemental figures

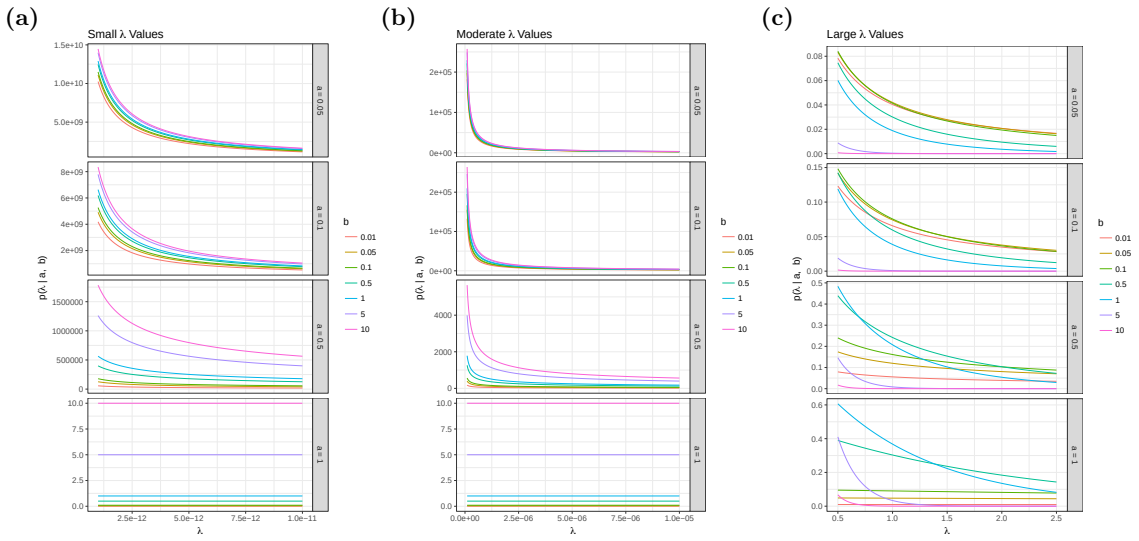

Figure S1.5: **A relatively informative prior on λ is effective.** The densities of different priors on λ for different ranges of λ values. The shape parameter a determines how quickly the prior density decreasys, while the rate parameter b determines how much prior weight is placed on small λ values rather than large λ values.

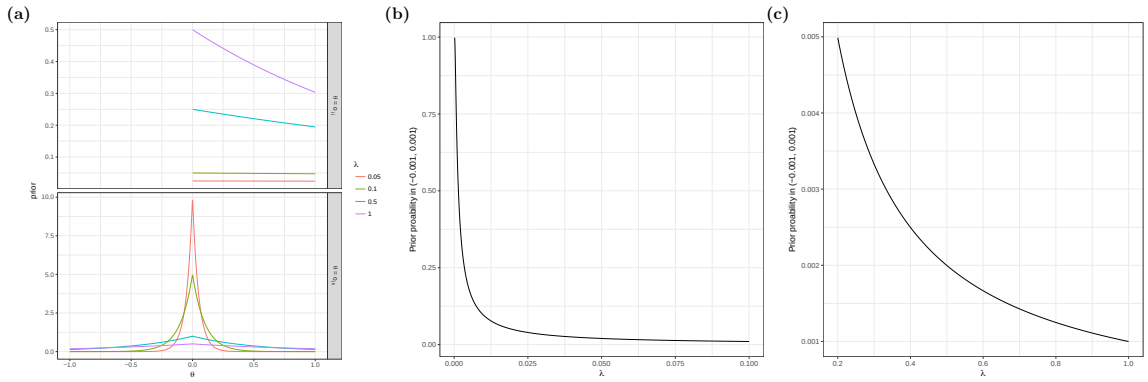

Figure S1.6: **Shrinkage increases for smaller λ.** (**A**) The shape of the prior on o_{jk} and o_{jj} for several values of λ. Smaller λ results in greater shrinkage towards zero. (**B**)-(**C**) The prior probability in the interval $(-0.001, 0.001)$ for each off-diagonal element $o_{jk}|\lambda \sim Laplace(\lambda)$ across small (**B**) or large (**C**) values of λ. Small values (< 0.1) of λ show the greatest shrinkage, while beyond $\lambda = 1$ the shrinkage becomes negligible.

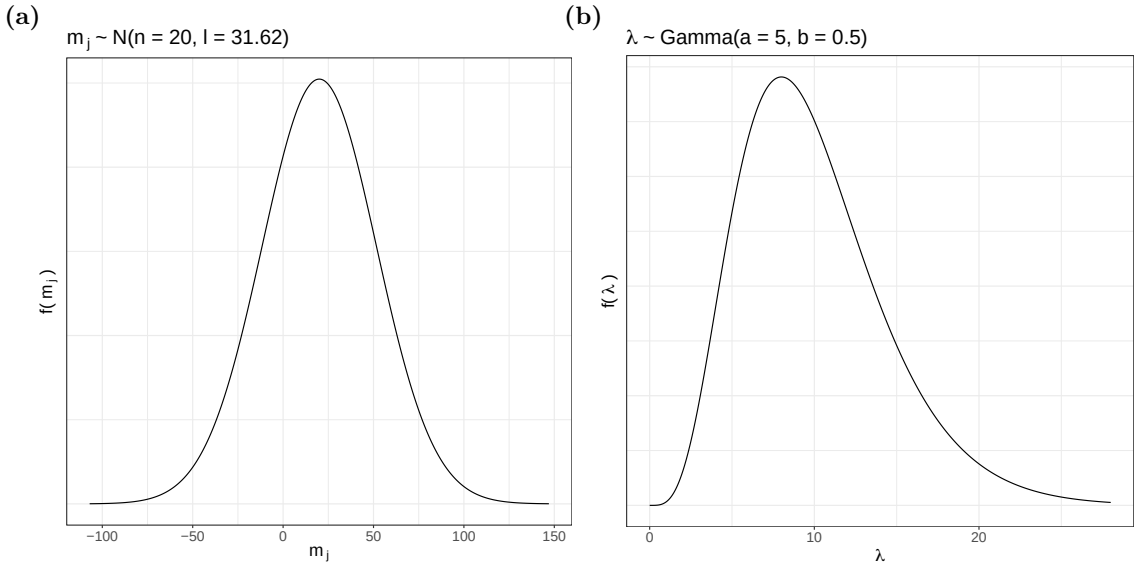

Figure S1.7: **Prior distributions for test cases.** The prior distributions for the test cases used a prior on $\boldsymbol{m}$ that was very uninformative, being centered at 0 and with a large variance. The prior on λ put most prior weight on λ values less than one and had narrow tails to encourage shrinkage of the correlation estimates (**B**).

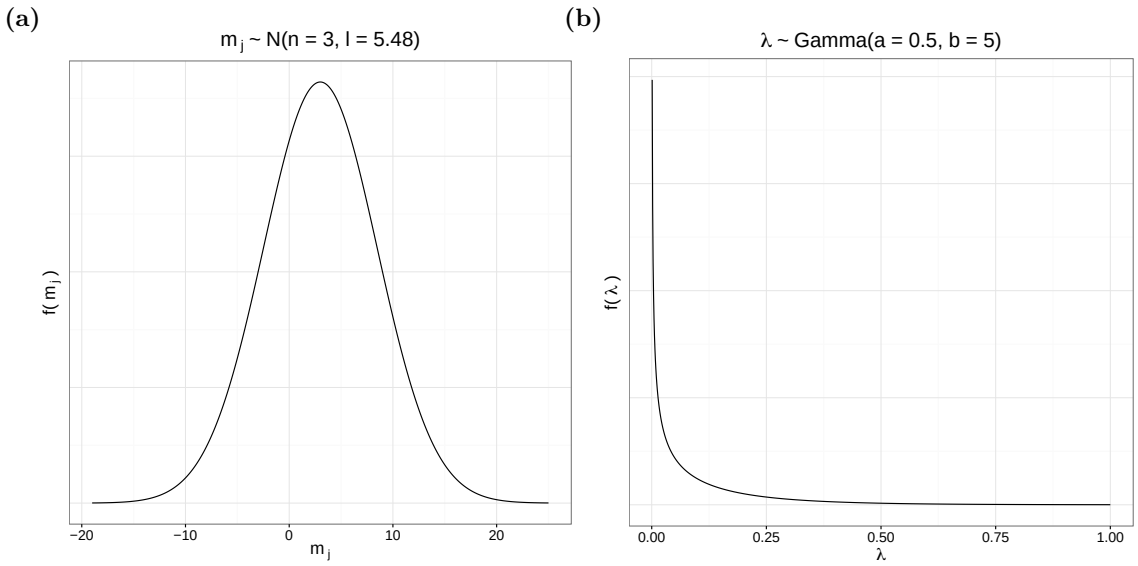

Figure S1.8: **Prior distributions for realistic simulated data.** We used a prior for $\boldsymbol{m}$ that gave reasonable behavior for the sum of the basis medians $\sum_{j=1}^{14} e^{m_j}$ (**A**). The prior on λ put most prior weight on λ values less than one and had narrow tails to encourage shrinkage of the correlation estimates (**B**). (See also **Figure S1.15**).

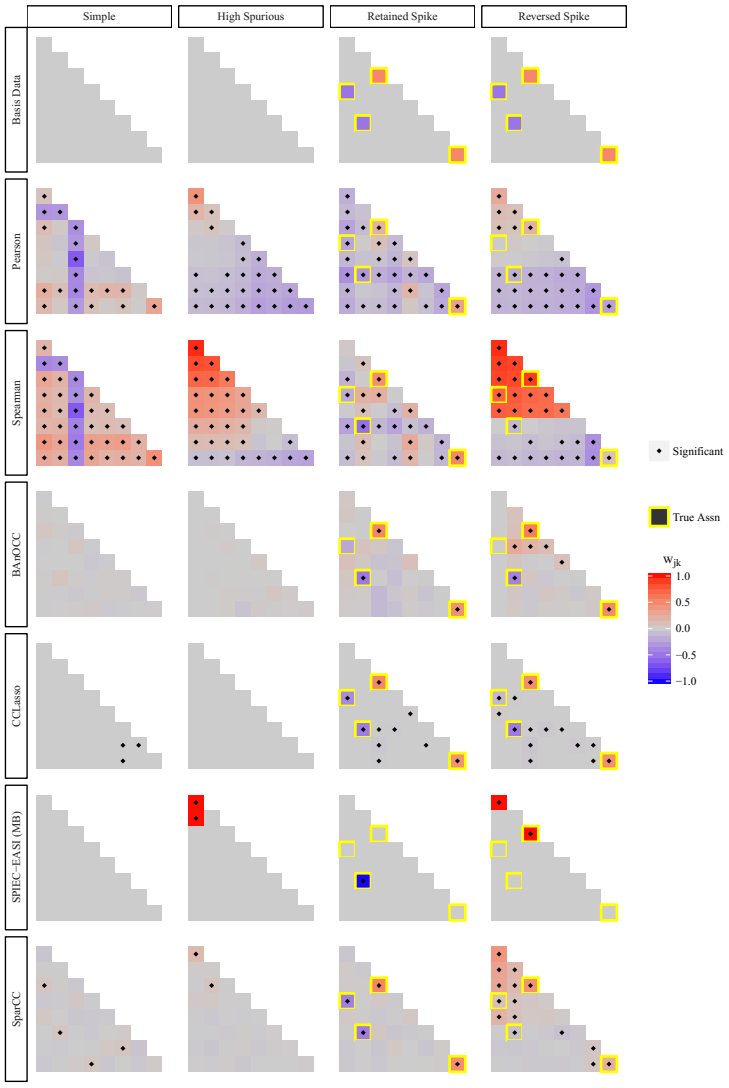

Figure S1.9: **Additional results for difficult scenarios.** The estimates and significance of several methods on the four scenarios (columns, see also **Section S1.4**): simple and retained spike, with no negative dominant spurious correlations; and high spurious and reversed spike, with strong negative dominant spurious correlations. The top row is the true log-basis correlation $\hat{\boldsymbol{R}}_{\log \boldsymbol{X}}$. The second row is the compositional correlation (and significance) using the 1,000 samples from the data. BAnOCC evaluates significance using 95% credible intervals. CCLAsso and SPIEC-EASI (MB) are significant if they are non-zero. SPIEC-EASI (MB) colors indicate the *sign* rather than the magnitude of the estimated correlations as the estimates are not possible to compute. SparCC evaluates significance using a bootstrap-based method. All the methods do poorly at detecting and correctly estimating the negative correlation between features 1 and 5 in the reversed spike scenario, and instead tend to falsely detect several positive correlations.

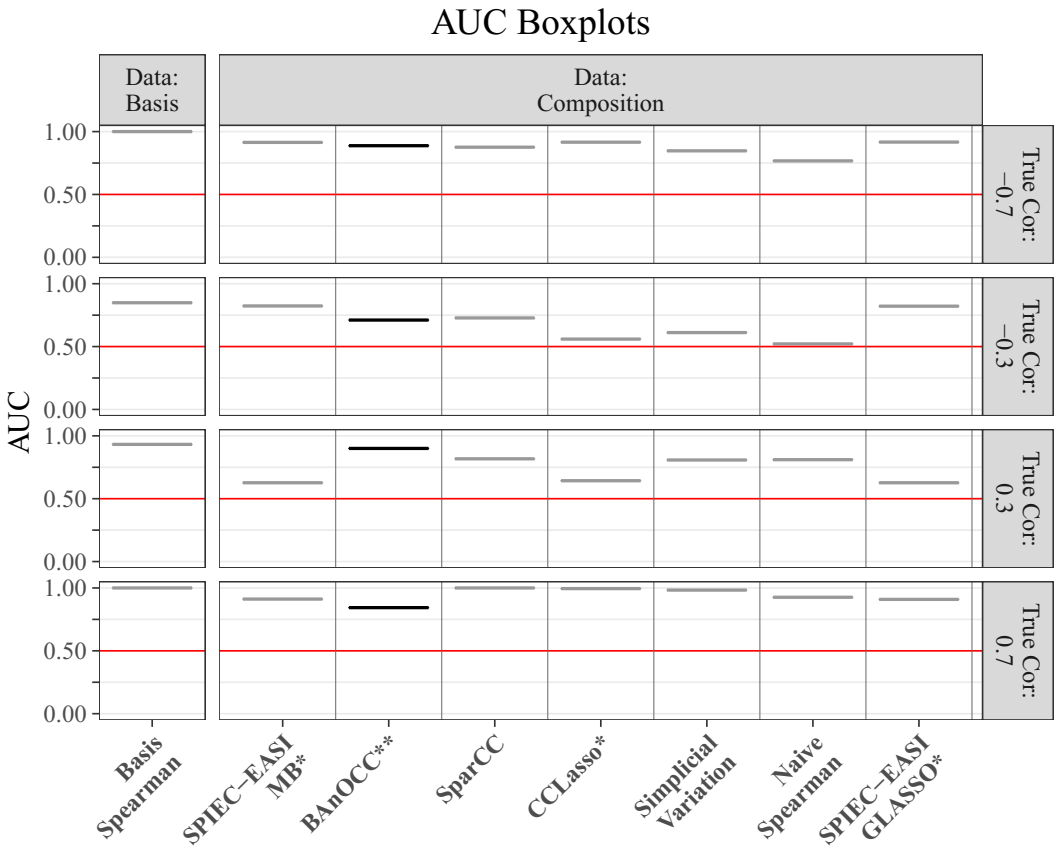

Figure S1.10: **AUC boxplots of method performance on realistic simulated datasets.** For a given correlation strength and template dataset, the AUC was calculated for each of the 100 simulated datasets based on p-values (Spearman correlation, simplicial variation, SparCC), credible intervals (BAnOCC), correlation estimate (CCLasso) or stability score (SPIEC-EASI). The AUCs are over seven true correlations, and all of the methods do better than random guessing (red line), but BAnOCC has overall the highest average AUC.

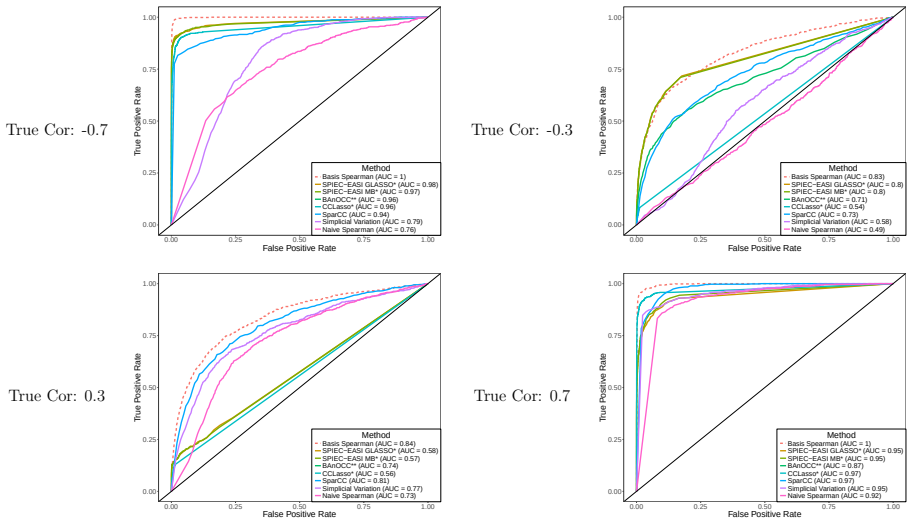

Figure S1.11: **Average ROC curves of method performance on realistic simulated datasets.** For a given correlation strength, calculate the ROC curve over all 700 true associations in the 100 simulated datasets. The cutoffs used are based on p-values (Spearman correlation, simplicial variation, SparCC), credible interval width (BAnOCC), correlation estimate (CCLasso) or stability score (SPIEC-EASI).

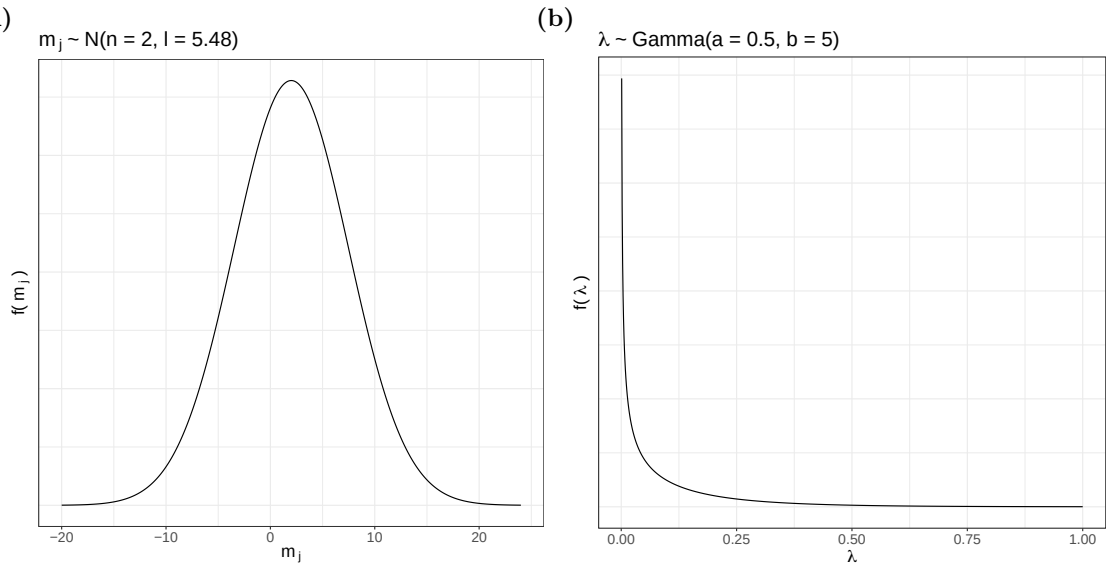

Figure S1.12: **Prior distributions for the stool body site.** We used a prior for $\boldsymbol{m}$ that gave reasonable behavior for the sum of the basis medians $\sum_{j=1}^{24} e^{m_j}$ (**A**). The prior on λ put most prior weight on λ values less than one and had narrow tails to encourage shrinkage of the correlation estimates (**B**). (See also **Figure S1.15**).

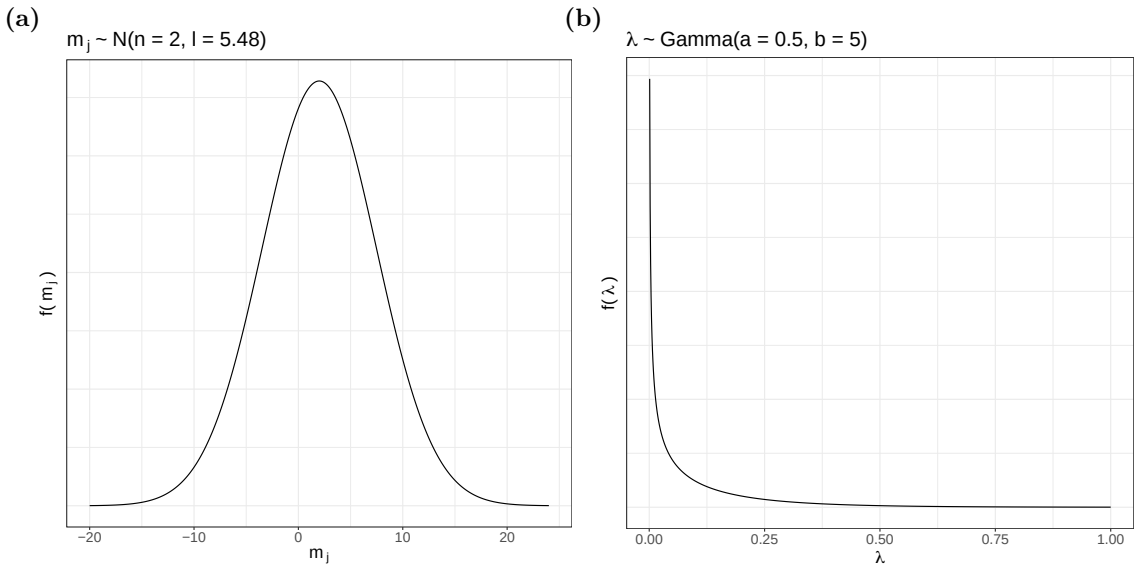

Figure S1.13: **Prior distributions for the buccal mucosa body site.** We used a prior for $\boldsymbol{m}$ that gave reasonable behavior for the sum of the basis medians $\sum_{j=1}^{21} e^{m_j}$ (**A**). The prior on λ put most prior weight on λ values less than one and had narrow tails to encourage shrinkage of the correlation estimates (**B**). (See also **Figure S1.15**).

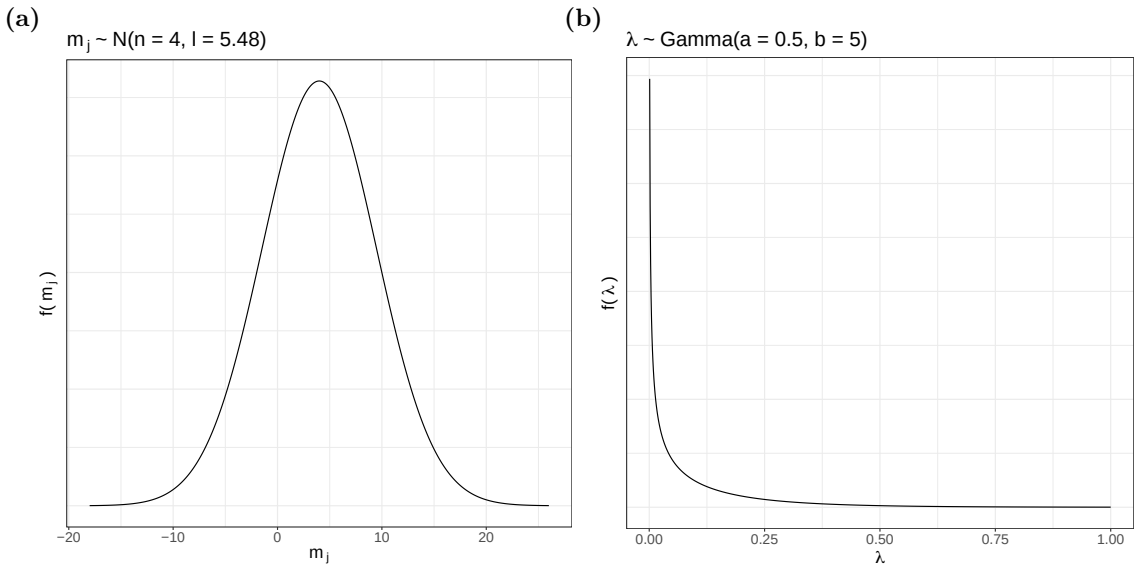

Figure S1.14: **Prior distributions for the posterior fornix body site.** We used a prior for $\boldsymbol{m}$ that gave reasonable behavior for the sum of the basis medians $\sum_{j=1}^{11} e^{m_j}$ (**A**). The prior on λ put most prior weight on λ values less than one and had narrow tails to encourage shrinkage of the correlation estimates (**B**). (See also **Figure S1.15**).

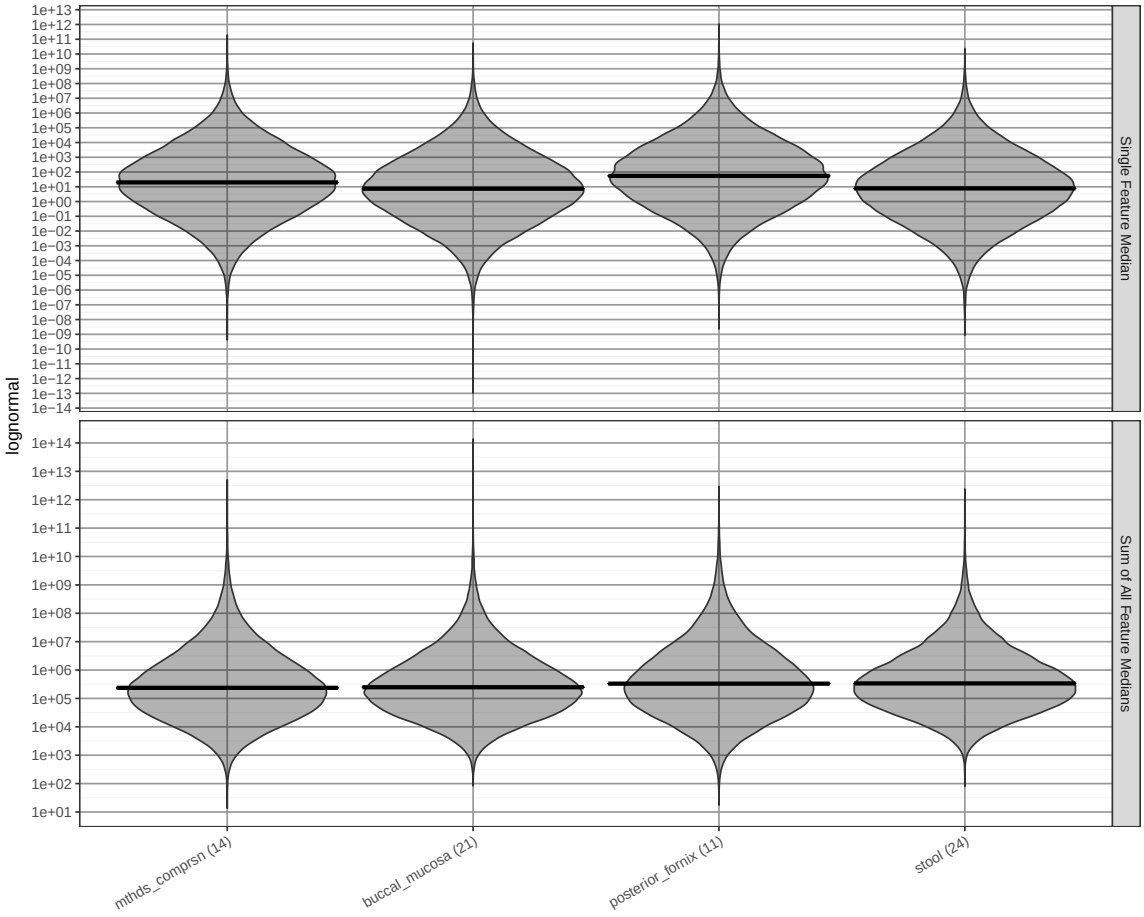

Figure S1.15: **Implied priors on median basis counts.** The implied priors on the median basis counts e^{m_j} (top panel) and the sum of the median basis counts $\sum_{j=1}^{p} e^{m_j}$ (bottom panel) for the SparseDOSSA simulated data and the body sites from the application. Each distribution is estimated using 100,000 random samples. The mean of m_j was selected such that the sum of the median basis counts approximately shared the same average.

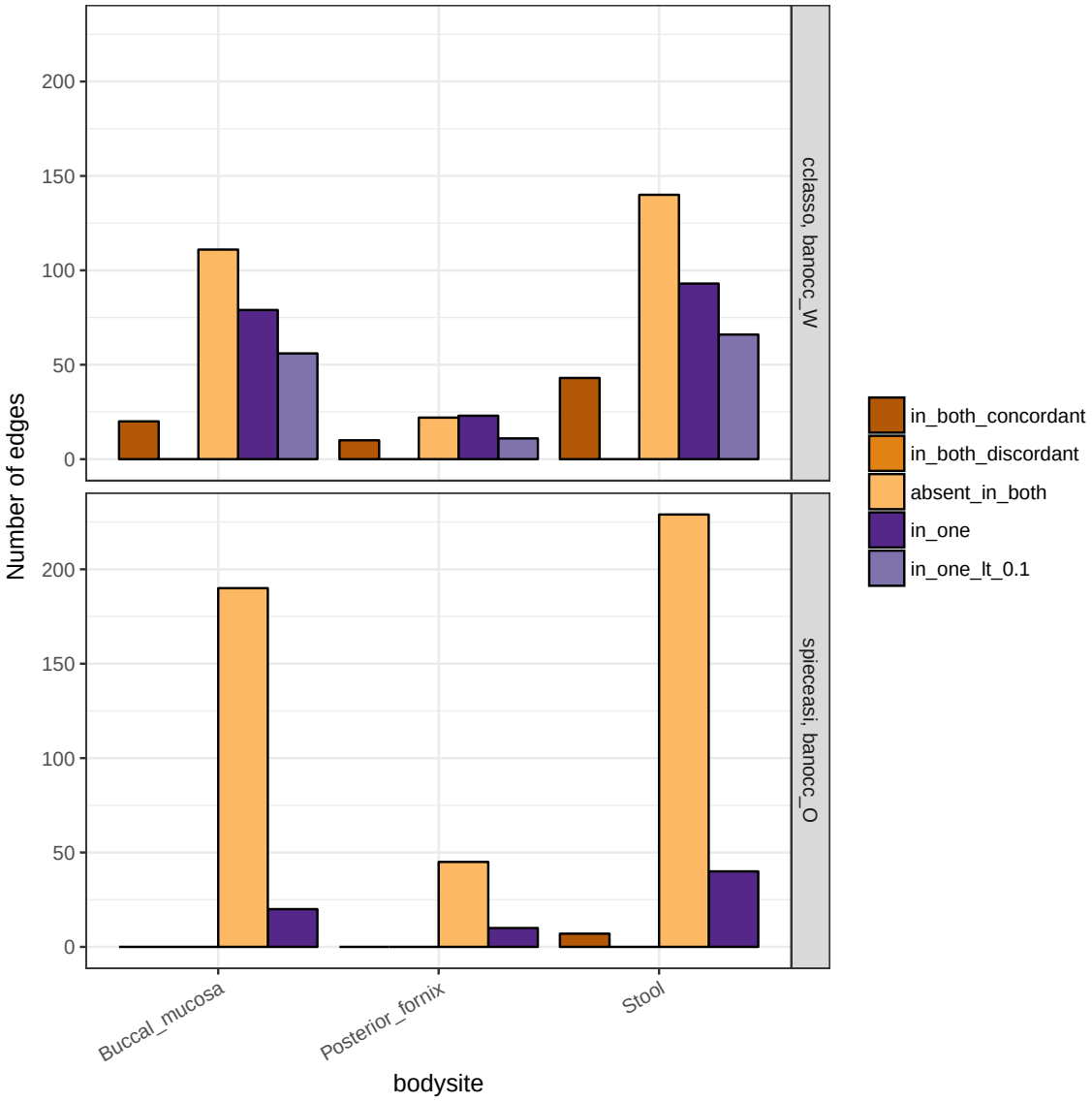

Figure S1.16: **Comparison of inferred networks on HMP data.** The number of edges significant in both methods, neither method, or only one method broken down by body site and whether the methods use the log-basis precision or correlation matrix. Most edges are either significant or not significant with both methods; few are found by only one method.

S1.8 Supplemental tables

Table S1.1: **BAnOCC buccal mucosa network.** The significant edges from running BAnOCC on the buccal mucosa body site with 5,500 warmup iterations and 12,000 total iterations. Edges are ordered by posterior median correlation magnitude. "hpd.95.ci" indicates the highest posterior density 95% credible intervals.

feature1	feature2	posterior.median	hpd.95.ci.lower	hpd.95.ci.upper
Propionibacterium	Viruses	0.42	0.22	0.60
Leptotrichia	*Haemophilus*	-0.35	-0.55	-0.12
Porphyromonas	*Fusobacterium*	0.34	0.08	0.57
Propionibacterium	*Streptococcus*	0.33	0.09	0.56
Corynebacterium	*Riemerella*	-0.33	-0.54	-0.10
Capnocytophaga	*Aggregatibacter*	0.32	0.08	0.54
Corynebacterium	*Leptotrichia*	0.32	0.10	0.53
Capnocytophaga	*Fusobacterium*	0.30	0.06	0.54
Corynebacterium	*Aggregatibacter*	0.28	0.05	0.49
Riemerella	*Granulicatella*	0.28	0.05	0.49
Corynebacterium	*Haemophilus*	-0.27	-0.49	-0.05
Rothia	*Alloprevotella*	-0.26	-0.47	-0.03
Propionibacterium	*Riemerella*	-0.25	-0.46	-0.02
Rothia	*Porphyromonas*	-0.24	-0.45	-0.03
Actinobacillus	*Haemophilus*	0.24	0.04	0.47
Rothia	*Fusobacterium*	-0.23	-0.45	-0.04
Streptococcus	Viruses	0.23	0.03	0.46
Rothia	*Aggregatibacter*	-0.23	-0.44	-0.02
Corynebacterium	*Capnocytophaga*	0.23	0.02	0.44
Kingella	*Neisseria*	0.22	0.00	0.43

Table S1.2: **BAnOCC posterior fornix network.** The significant edges from running BAnOCC on the posterior fornix body site with 1,500 warmup iterations and 5,000 total iterations. Edges are ordered by posterior median correlation magnitude. "hpd.95.ci" indicates the highest posterior density 95% credible intervals.

feature1	feature2	posterior.median	hpd.95.ci.lower	hpd.95.ci.upper
Prevotella	*Lactobacillus*	-0.77	-0.92	-0.56
Prevotella	*Dialister*	0.75	0.40	0.92
Gardnerella	*Dialister*	0.70	0.34	0.89
Gardnerella	*Prevotella*	0.68	0.40	0.85
Streptococcus	Retroviridae	0.62	0.30	0.83
Lactobacillus	*Dialister*	-0.61	-0.83	-0.28
Gardnerella	*Lactobacillus*	-0.55	-0.78	-0.28
Propionibacterium	Retroviridae	0.41	0.04	0.70
Gardnerella	*Pseudomonas*	-0.41	-0.71	-0.03
Retroviridae	Viruses	0.36	0.04	0.67

Table S1.3: **BAnOCC stool network.** The significant edges from running BAnOCC on the stool body site with 5,500 warmup iterations and 12,000 total iterations. Edges are ordered by posterior median correlation magnitude. "hpd.95.ci" indicates the highest posterior density 95% credible intervals.

feature1	feature2	posterior.median	hpd.95.ci.lower	hpd.95.ci.upper
Veillonella	*Haemophilus*	0.68	0.57	0.78
Streptococcus	*Veillonella*	0.50	0.34	0.65
Bacteroides	Viruses	0.49	0.31	0.63
Bacteroides	*Haemophilus*	0.45	0.27	0.61
Streptococcus	*Haemophilus*	0.41	0.25	0.56
Bacteroides	*Oscillibacter*	0.39	0.20	0.58
Haemophilus	Viruses	0.38	0.22	0.55
Parabacteroides	*Alistipes*	0.38	0.21	0.54
Odoribacter	*Alistipes*	0.37	0.20	0.54
Veillonella	Viruses	0.32	0.16	0.50
Bacteroides	*Veillonella*	0.31	0.13	0.47
Clostridium	*Faecalibacterium*	-0.29	-0.47	-0.09
Roseburia	*Faecalibacterium*	0.27	0.07	0.47
Faecalibacterium	*Escherichia*	-0.27	-0.44	-0.09
Bifidobacterium	*Collinsella*	0.27	0.08	0.44
Alistipes	*Clostridium*	-0.27	-0.44	-0.08
Odoribacter	*Parabacteroides*	0.26	0.09	0.44
Streptococcus	Viruses	0.25	0.06	0.43
Oscillibacter	Viruses	0.25	0.07	0.43
Anaerotruncus	*Escherichia*	0.25	0.05	0.45
Coprococcus	*Oscillibacter*	-0.25	-0.42	-0.07
Subdoligranulum	*Haemophilus*	-0.25	-0.42	-0.07
Subdoligranulum	Viruses	-0.23	-0.41	-0.05
Holdemania	*Escherichia*	0.23	0.05	0.44
Bacteroides	*Parabacteroides*	0.23	0.04	0.41
Bacteroides	*Subdoligranulum*	-0.23	-0.40	-0.05
Eubacterium	*Escherichia*	-0.21	-0.38	-0.04
Ruminococcus	*Holdemania*	-0.21	-0.39	-0.02
Bacteroides	*Coprococcus*	-0.21	-0.38	-0.04
Odoribacter	*Clostridium*	-0.20	-0.38	-0.03
Eubacterium	*Faecalibacterium*	0.20	0.02	0.40
Coprococcus	*Faecalibacterium*	0.20	0.04	0.40
Coprococcus	Viruses	-0.19	-0.37	-0.01
Roseburia	*Anaerotruncus*	-0.19	-0.37	-0.01
Subdoligranulum	*Veillonella*	-0.19	-0.35	-0.03
Faecalibacterium	*Ruminococcus*	0.19	0.02	0.38
Oscillibacter	*Faecalibacterium*	-0.19	-0.36	-0.02
Eubacterium	Viruses	-0.18	-0.36	-0.02
Bacteroides	*Ruminococcus*	-0.18	-0.35	-0.01
Escherichia	Viruses	0.17	0.00	0.36
Faecalibacterium	*Holdemania*	-0.17	-0.35	-0.02
Clostridium	*Anaerotruncus*	0.17	0.00	0.37
Anaerotruncus	*Faecalibacterium*	-0.16	-0.31	-0.03

Appendix for Chaper 2

S2.1 Supplemental methods

Supplementary ecological measures

Several alternative ecological measures were explored (**Figure S2.5**), including core-to-pan-genome ratio, pan-genome size, and genome size.

For each species with at least two isolate genomes in the NCBI at the time of MetaPhlAn2 v2.2.0 release (2014), we counted the number of total genes across the isolate genomes to give the pan-genome size. Core genes were defined as either (1) present in all genomes of that species or (2) absent in k out of N genomes with $1 - F(k-1, N, 0.05) > 0.95$ where $F(\cdot)$ is the binomial cdf. The second criterion assumes that about 5% of genes are erroneously missing from any genome, and ensures that genes can still be core if the probability of that gene being erroneously excluded from the genome is high (at least 95%). This is more likely when N is small than when N is large.

The number of nucleotides for each isolate genome of a particular taxon were obtained from the NCBI genomes for isolates in the MetaPhlAn2 database. The size of the genome for each taxon was the average of the genome size of the associated isolate genomes.

Supplemental network substructure

Clusters were detected using a distance of $1 - \rho_{ij}$ if an edge was significant, and 1 if it was not. Three cluster detection methods from the fpc R package (version 2.1.10) [Hennig, 2015]: hierarchical clustering with average linkage, clara [Rousseeuw and Kaufman, 1990], and k-medoids.

Three cluster evaluation methods from the fpc R package (version 2.1.10) [Hennig, 2015] were employed: silhouette width [Rousseeuw, 1987], Calinski-Harabasz index

[Caliński and Harabasz, 1974], and prediction strength [Tibshirani and Walther, 2005].

S2.2 Supplemental data

Data S2.4: **BAnOCC Network.** Table of the complete network, including significant and non-significant edges. Each row indicates an edge. Fields are: feature1 and feature2 are the the two taxa; bs1 and bs2 are the two body sites; bodysite is the bodysite containing both nodes (this is "btw" if the nodes are in different body sites); time is the time point; estimate is the correlation estimate (for intra-body-site edges, this is the posterior median from BAnOCC, while for inter-body-site edges, this is pearson correlation); sig indicates whether the edge is significantly different from zero (for intra-body-site edges, this is whether the 95% credible interval excludes zero; for inter-body-site edges, this is whether the q-value is less than 0.05).

S2.3 Supplemental figures

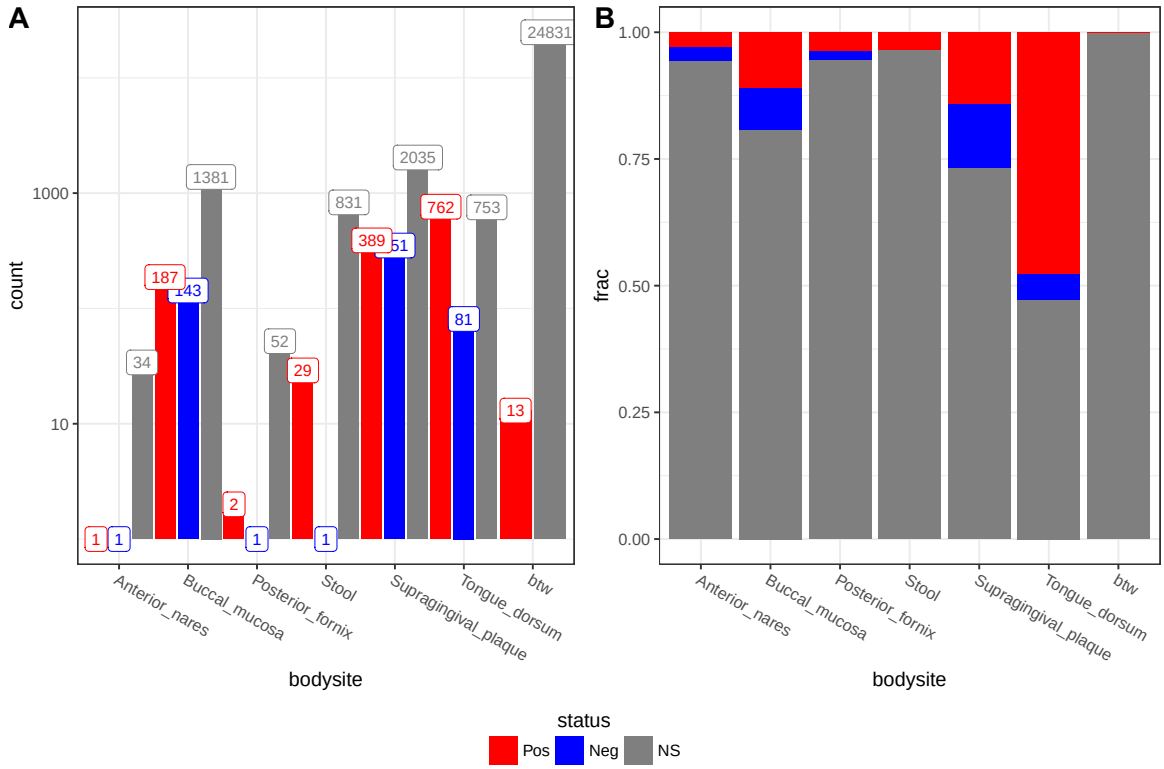

Figure S2.1: **Breakdown of edges by significance and direction.** Counts (**A**) or fractions (**B**) of significant or non-significant edges. Significant edges were significant at both timepoints, while non-significant edges were significant for at most one time point. The sign of the edge is indicated only for significant edges.

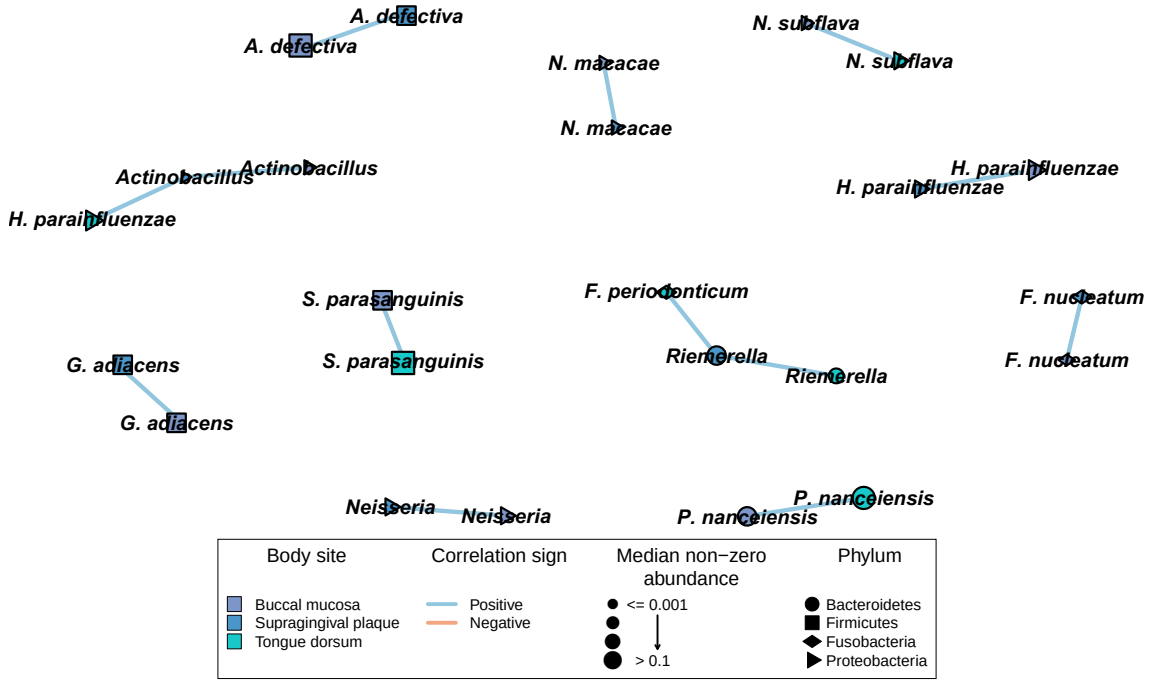

Figure S2.2: **Edges between body sites.** All edges that were significant at both time points and connect taxa across different body sites. Without exception, they connect species found in the three oral sites, and most connect the same species in different sites.

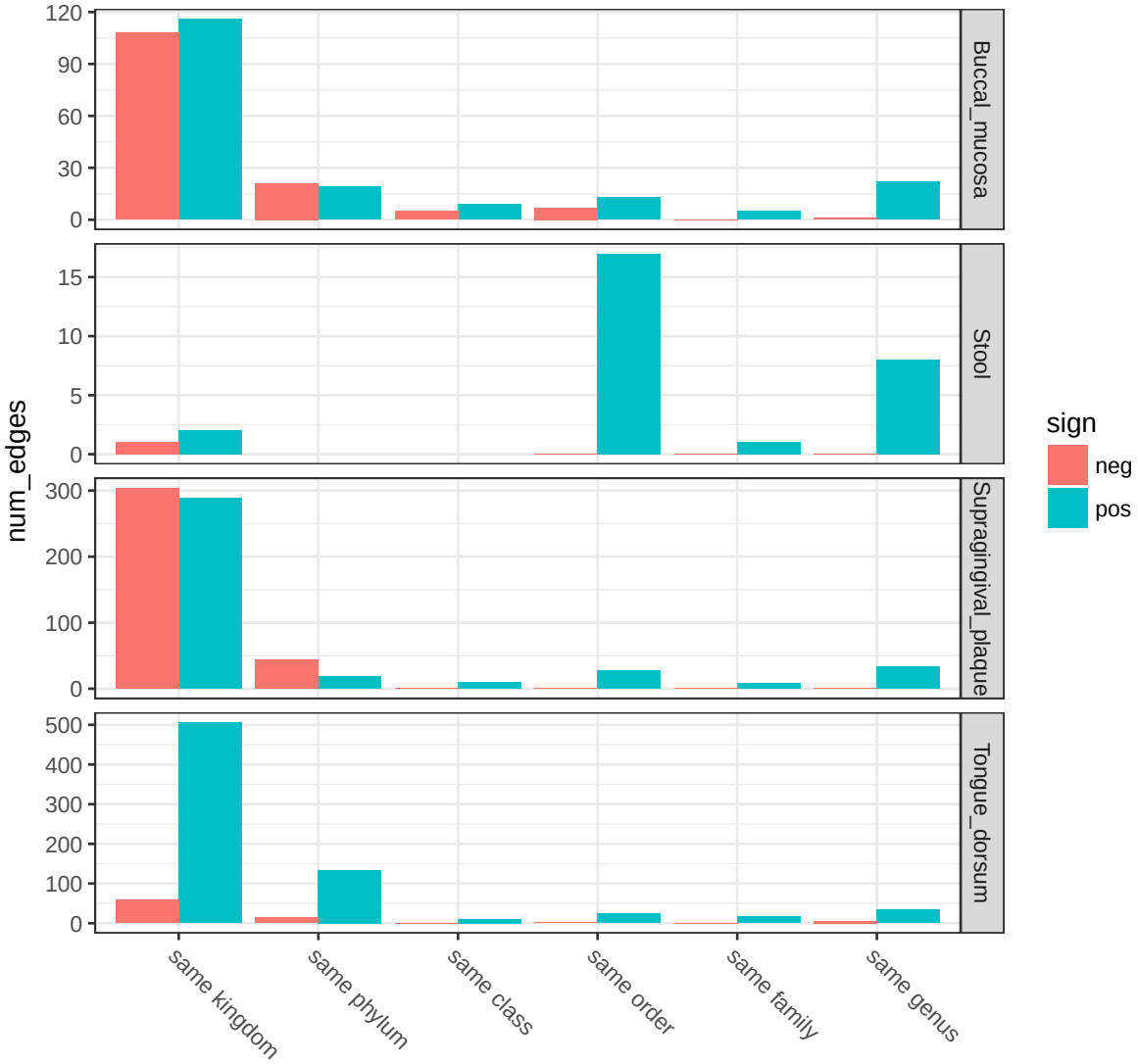

Figure S2.3: **Breakdown of correlation sign by relatedness of species involved.** A breakdown of the sign of the correlation between species by whether the two species involved share genus, family, etc. A greater proportion of positive edges involve the species of the same genus or family, while the proportions of positive and negative edges corresponding to different classes or phyla are approximately equal.

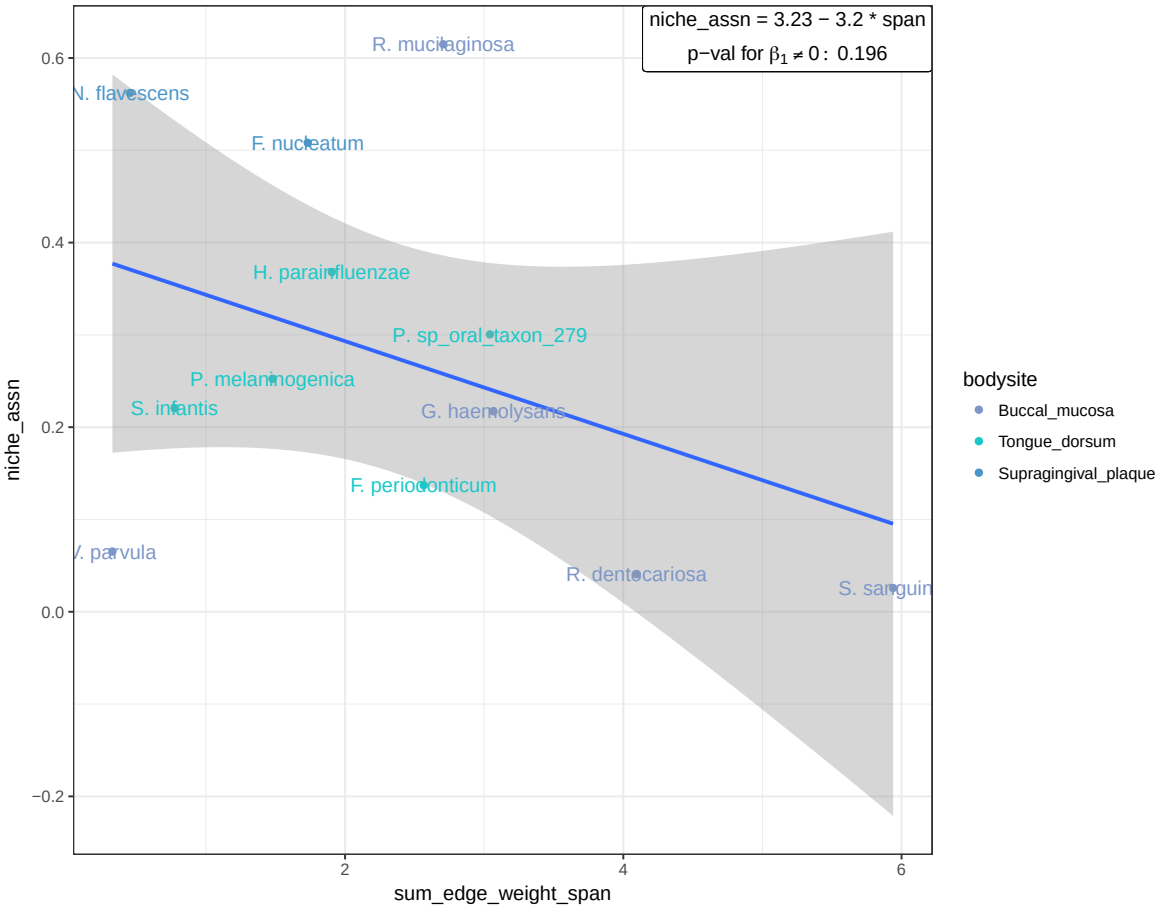

Figure S2.4: **Correlation between niche-association measure and degree span.** The niche-association measure of Lloyd-Price et al. [in press] measures how differentiated strains are within and between body sites. The span of the edge weight sums is the difference between the maximal and minimal edge weight sums for each node in **Figure 2.3**.

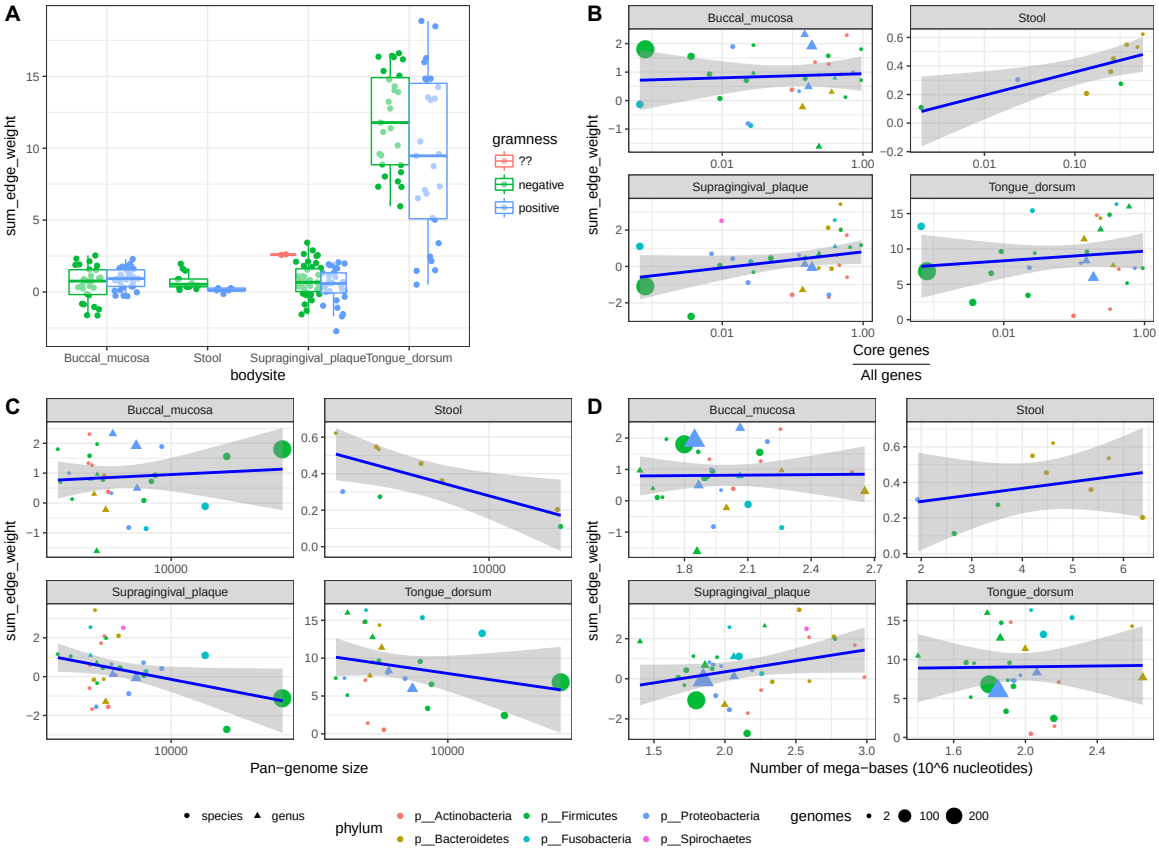

Figure S2.5: **Measures of generality versus weighted edge sum.** The relationship of the weighted edge sum to gramness (**A**), pan-to-core genome ratio (**B**), pan-genome size (**C**), and genome-size (**D**). For panels (**B**))-(**D**), the color of the nodes indicates the phylum for each species.

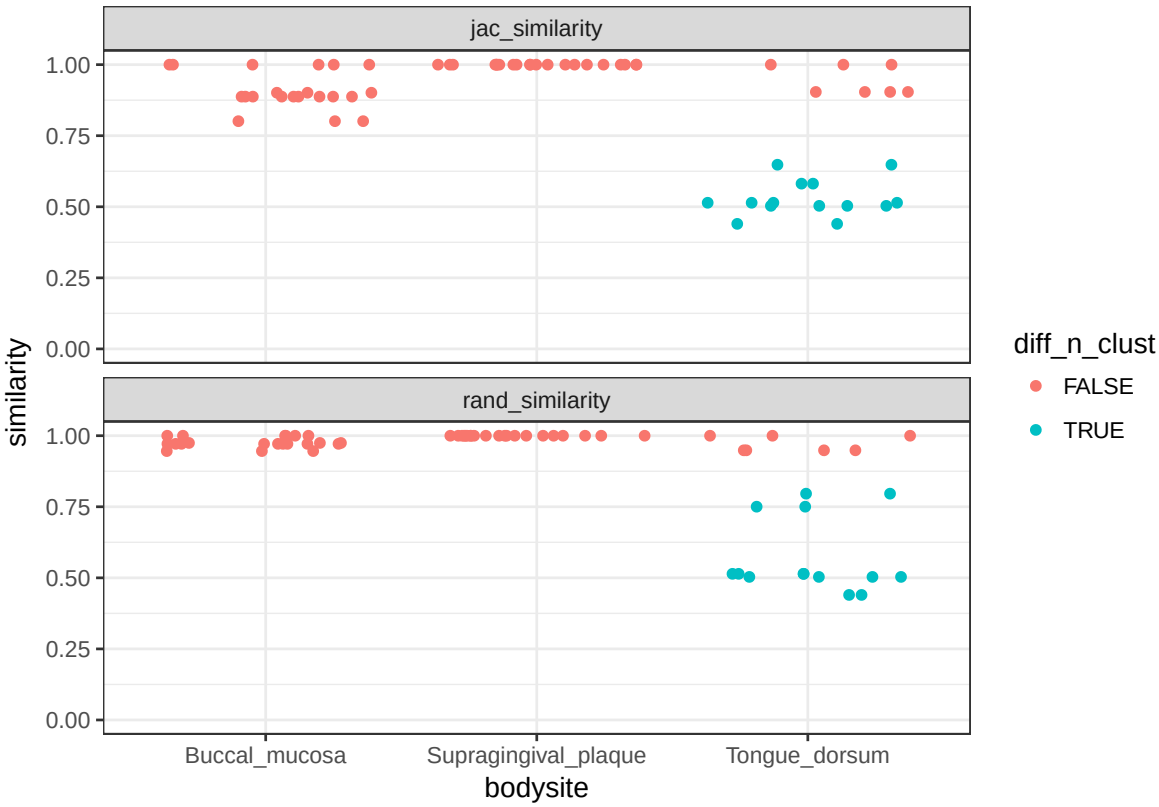

Figure S2.6: **Consistency of discovered modules across different methods.** The Jaccard and rand distance between the clusters from each of the different methods for each of the body sites. High values indicate high similarity between the clusters, with 1.0 indicating perfect agreement for both measures.

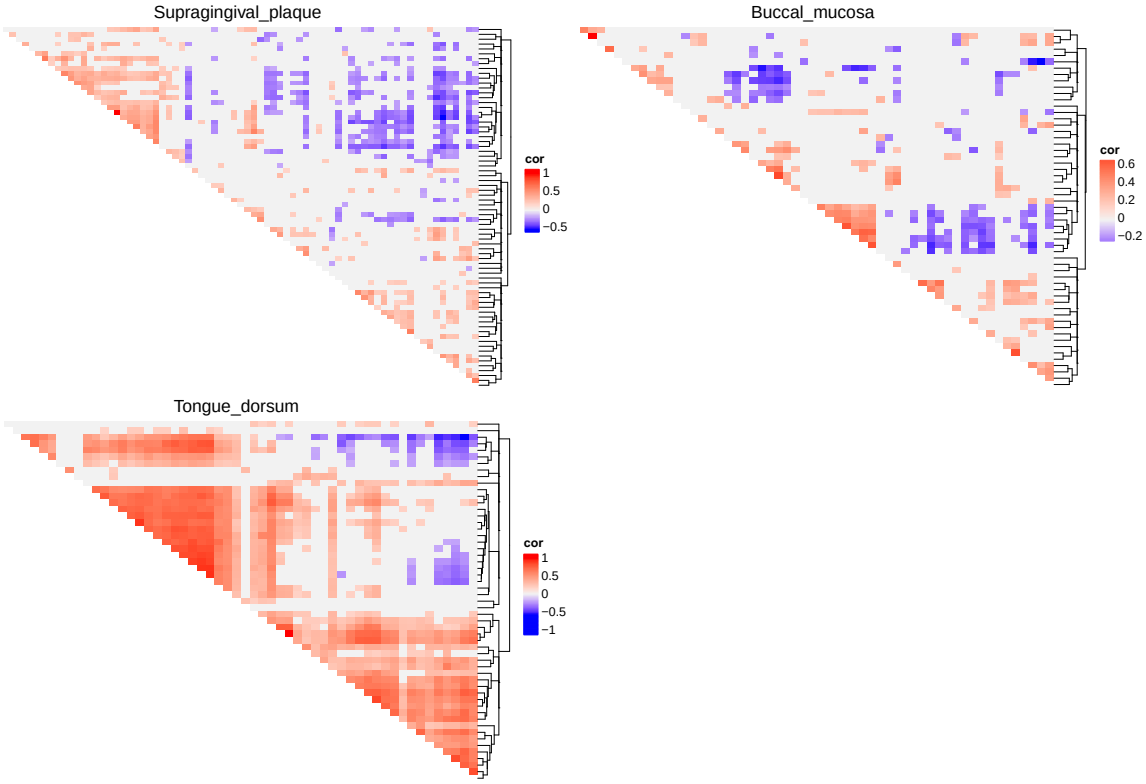

Figure S2.7: **Heatmap views of the three oral sites.** Heatmaps and dendrograms of the three oral sites show a high degree of sparsity that could contribute to the difficulty of finding substructure.

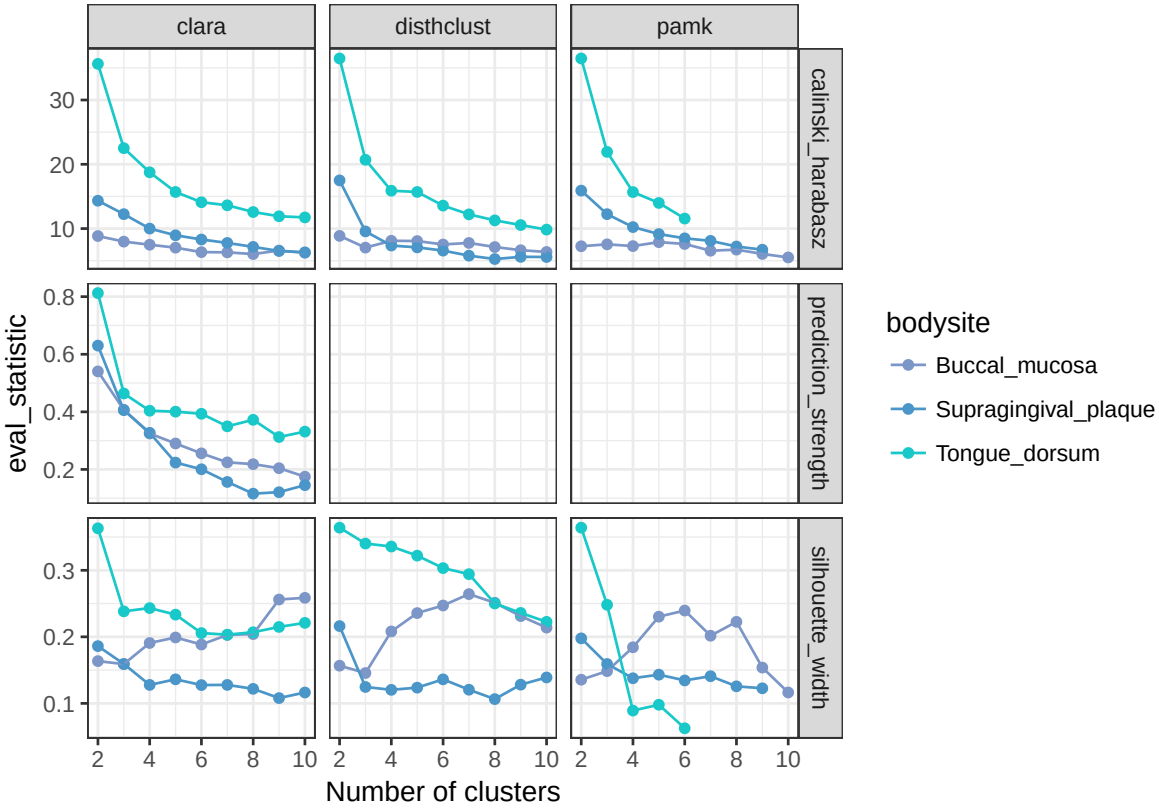

Figure S2.8: **Evaluation of clustering behavior in oral body sites.** Three measures of clustering strength: Calinski-Harabasz [Caliński and Harabasz, 1974], prediction strength [Tibshirani and Walther, 2005], and silhouette width [Rousseeuw, 1987] were used to assess the number of strong clusters present in each of the three oral sites. For all of these measures, high values indicate strong evidence, and a peak at the optimal number of clusters is expected. Three clustering methods were employed: CLARA [Rousseeuw and Kaufman, 1990], hierarchical clustering (disthclust), and K-medoids clustering (pamk). Within the fpc package, neither hierarchical clustering nor K-medoids clustering was able to run successfully with prediction strength.

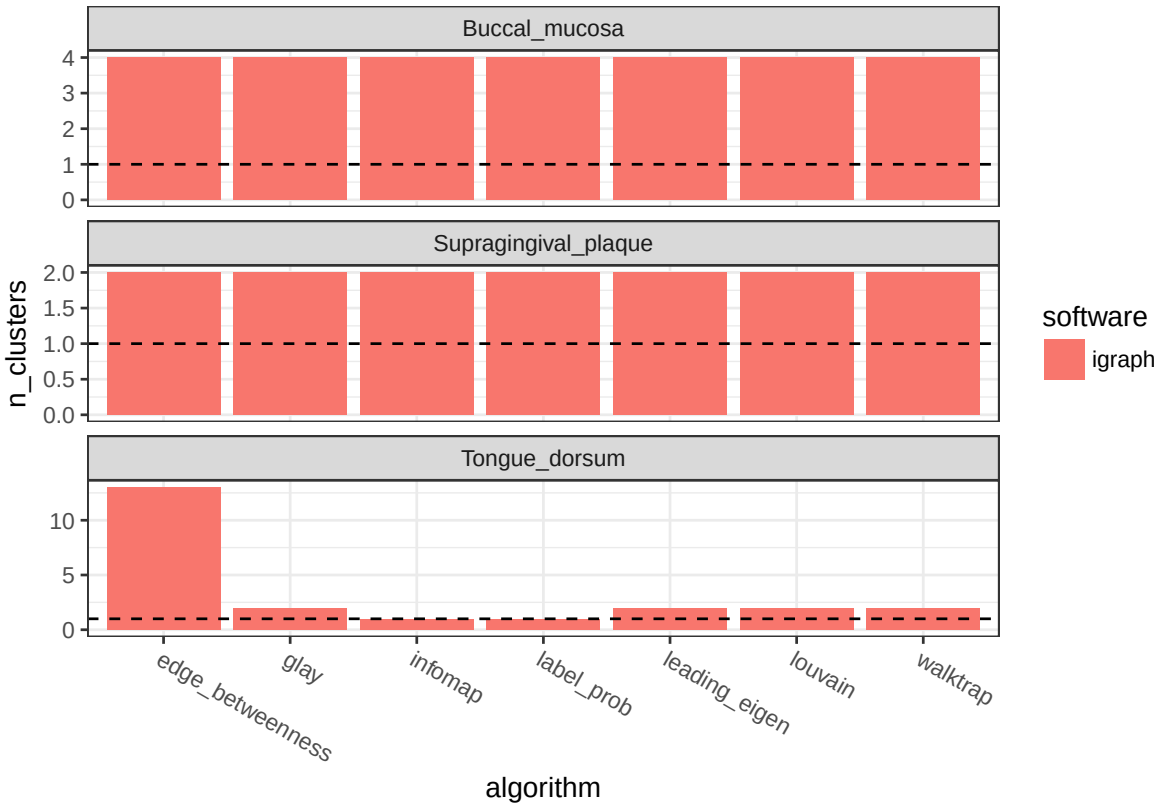

Figure S2.9: **Number of discovered modules across different methods.** The number of modules discovered for each body site and method. For supragingival plaque and buccal mucosa, the numbers are very consistent. Tongue dorsum had variable numbers of modules as a result of the methods' differing approaches to a strong and dense network with few absent edges.

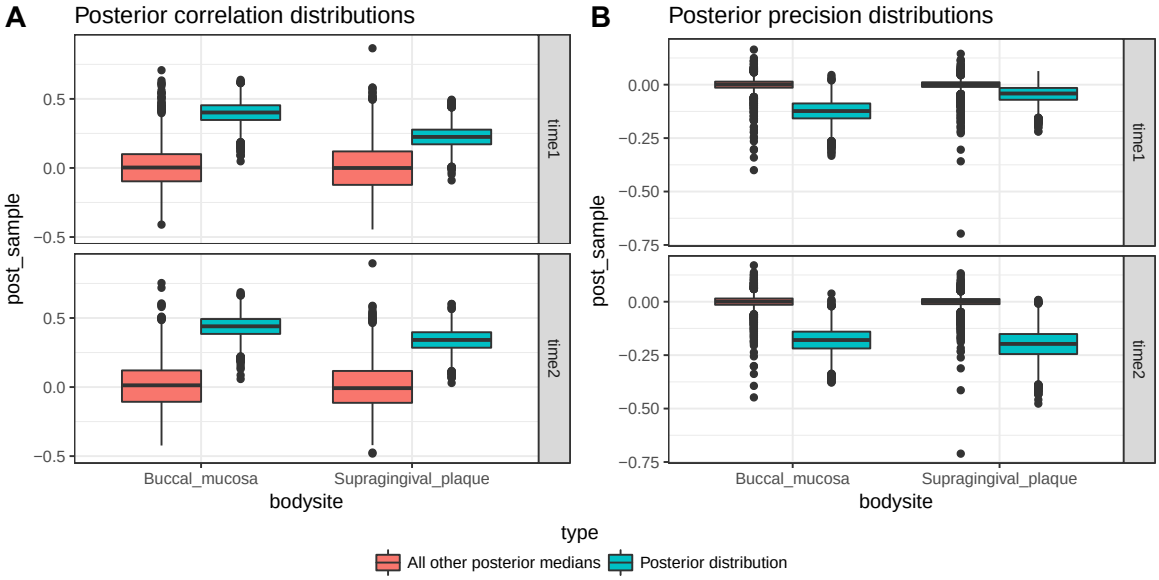

Figure S2.10: **Posterior distributions of edge weights.** Posterior distributions of edges weights for the correlations (**A**) or precision matrix entries (**B**). For the edge between *S. cristatus* and *C. matruchotii*, the posterior distribution of the edge weights, compared with the distribution of posterior medians from all other edges (significant and not significant) in the networks at each time point. A precision matrix element being negative indicates that the corresponding conditional correlation is positive.

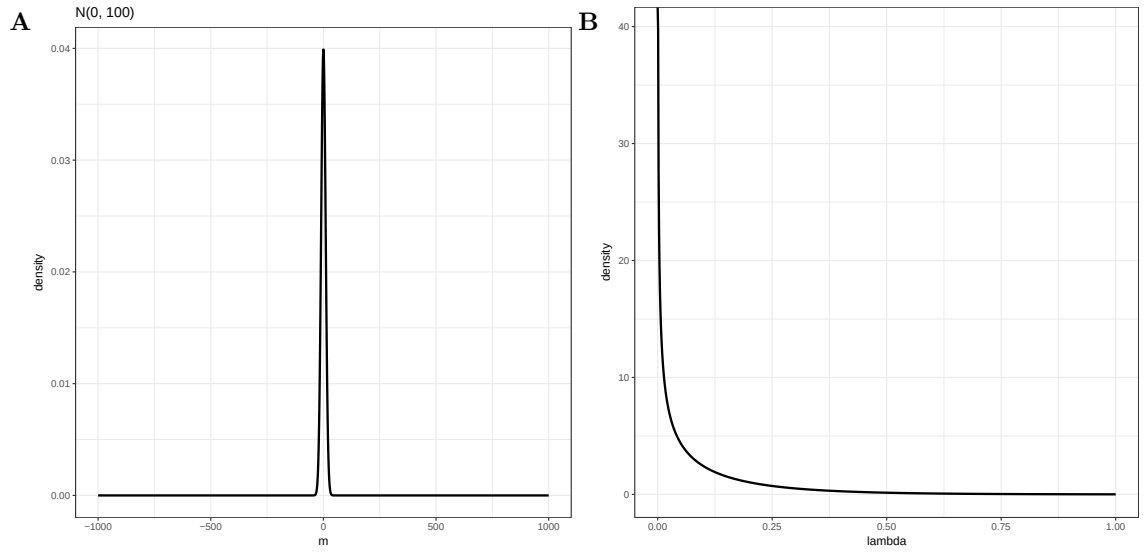

Figure S2.11: **Prior distributions for BAnOCC parameters.** Prior distributions on the log-basis means (**A**) and the LASSO shrinkage parameter (**B**) used for all body sites and time points.

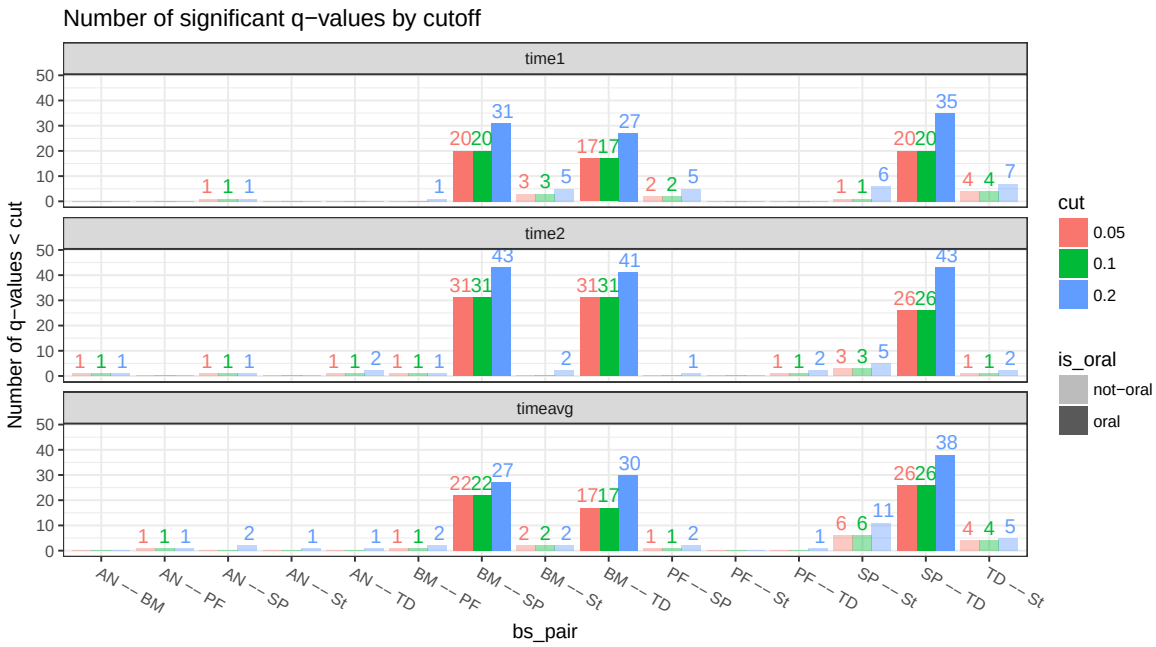

Figure S2.12: **Q-value cutoff does not appreciably change which between-body site edges are included.** The number of q-values less than a particular cutoff for three different cutoffs. Whether the cutoff is 0.05 or 0.1 makes no difference in the number of edges included.

S2.4 Supplemental tables

Table S2.1: **Dimensions of the datasets after filtering.** The number of unique taxa, subjects, and samples from each body site at the first two time points after removing (1) body sites that have fewer than 20 subjects at any time point; (2) technical replicates; and (3) taxa that have prevalences less than the cutoffs in Table S2.3 at either time point for a body site.

STSite	num_visno	n_taxa	n_subj	n_samp
Anterior_nares	1.00	8	95	95
Anterior_nares	2.00	8	87	87
Buccal_mucosa	1.00	58	132	132
Buccal_mucosa	2.00	58	122	122
Posterior_fornix	1.00	10	76	76
Posterior_fornix	2.00	10	73	73
Stool	1.00	41	205	205
Stool	2.00	41	187	187
Supragingival_plaque	1.00	74	141	141
Supragingival_plaque	2.00	74	134	134
Tongue_dorsum	1.00	56	159	159
Tongue_dorsum	2.00	56	143	143

Table S2.2: **Initial dimensions of the dataset split by bodysite and time point.** The number of unique taxa, subjects, and samples within each body site at each time point before any filtering

STSite	time_1	time_2	time_3	min
Anterior_nares	95	87	48	48
Buccal_mucosa	132	122	78	78
Hard_palate	–	1	–	1
Keratinized_gingiva	5	6	6	5
L_Retroauricular_crease	6	10	7	6
Mid_vagina	1	5	8	1
Palatine_Tonsils	4	12	9	4
Posterior_fornix	76	73	40	40
R_Antecubital_fossa	–	–	1	1
R_Retroauricular_crease	7	13	10	7
Saliva	3	2	2	2
Stool	205	187	79	79
Subgingival_plaque	6	10	8	6
Supragingival_plaque	141	134	70	70
Throat	5	9	4	4
Tongue_dorsum	159	143	77	77
Vaginal_introitus	2	4	5	2

Table S2.3: **Prevalence cutoffs for each body site and time point.** The final prevalence cutoffs used for each body site and time point, as well as the resulting number of taxa included. Prevalence for each taxon in each body site and time point was determined as the percentage of samples having abundance at least 1e-04 relative abundance. The initial prevalence filter was 50%, but where this resulted in less than 10 taxa for any time point from a body site, the prevalence filter was decreased so that all time points retained at least 10 taxa.

STSite	num_visno	prev_cut_0	n_taxa_0	min_cut	final_cut	n_taxa
Anterior_nares	1.00	0.50	8	0.45	0.45	10
Anterior_nares	2.00	0.50	9	0.48	0.45	11
Buccal_mucosa	1.00	0.50	64	0.50	0.50	64
Buccal_mucosa	2.00	0.50	59	0.50	0.50	59
Posterior_fornix	1.00	0.50	2	0.23	0.23	11
Posterior_fornix	2.00	0.50	3	0.27	0.23	19
Stool	1.00	0.50	43	0.50	0.50	43
Stool	2.00	0.50	41	0.50	0.50	41
Supragingival_plaque	1.00	0.50	74	0.50	0.50	74
Supragingival_plaque	2.00	0.50	82	0.50	0.50	82
Tongue_dorsum	1.00	0.50	56	0.50	0.50	56
Tongue_dorsum	2.00	0.50	61	0.50	0.50	61

Table S2.4: **MCMC parameters for running BAnOCC.**

Body site	Timepoint	Chains	Warmup iterations	Total iterations	Thinning
Posterior fornix	time 1	4	1500	5000	2
Posterior fornix	time 2	4	1500	5000	2
Anterior nares	time 1	4	5500	12000	3
Anterior nares	time 2	4	5500	12000	3
Tongue dorsum	time 1	4	1500	5000	2
Tongue dorsum	time 2	4	1500	5000	2
Supragingival plaque	time 1	4	5000	10000	3
Supragingival plaque	time 2	4	5000	10000	3
Buccal mucosa	time 1	4	5000	10000	3
Buccal mucosa	time 2	4	5000	10000	3
Stool	time 1	4	5000	10000	3
Stool	time 2	4	5000	10000	3

Appendix for Chapter 3

S3.1 Derivation of feature-specific expectations

If $Y \sim \mathcal{LN}(\mu, \sigma)$, it is easily shown that $E(Y) = e^{\mu+0.5\sigma^2}$. Let $Y_t = Y|Y \leq t$ be a right-truncated version of Y with truncation point t. Then

$$f(Y_t) = \frac{1}{F(t,\mu,\sigma)y\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right),$$

where $F(y, \mu, \sigma)$ is the cdf of a log-normal distribution with parameters μ and σ. The expectation of the truncated log-normal, therefore, is

$$\begin{aligned}
E(Y_t) &= \int_0^t \frac{y}{F(t,\mu,\sigma)y\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right)dy \\
&= \int_{-\infty}^{\frac{\ln t-(\mu+\sigma^2)}{\sigma}} \frac{\exp\left(\sigma z + \mu + \sigma^2\right)}{F(t,\mu,\sigma)\sqrt{2\pi}}\exp\left(-\frac{(\sigma z + \mu + \sigma^2 - \mu)^2}{2\sigma^2}\right)dz \\
&= \frac{e^{\mu+0.5\sigma^2}}{F(t,\mu,\sigma)}\int_{-\infty}^{\frac{\ln t-(\mu+\sigma^2)}{\sigma}} \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi}} \\
&= \frac{e^{\mu+0.5\sigma^2}}{F(t,\mu,\sigma)}\Phi\left(\frac{\ln t - (\mu+\sigma^2)}{\sigma}\right),
\end{aligned}$$

where Φ is the cdf of a standard normal distribution.

The variance of a truncated log-normal distribution is $Var(Y_t) = E(Y_t^2) - E(Y_t)^2$. Similarly to above,

$$\begin{aligned}
E(Y_t^2) &= \int_0^t \frac{y^2}{F(t,\mu,\sigma)y\sigma\sqrt{(2\pi)}}\exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right)dy \\
&= \int_{-\infty}^{\frac{\ln t-\mu-2\sigma^2}{\sigma}} \frac{\exp\left(\sigma z + \mu + 2\sigma^2\right)^2}{F(t,\mu,\sigma)\sqrt{2\pi}}\exp\left(-\frac{(\sigma z + \mu + 2\sigma^2 - \mu)^2}{2\sigma^2}\right)dz \\
&= \frac{\exp(2\mu+2\sigma^2)}{F(t,\mu,\sigma)}\Phi\left(\frac{\ln t - \mu - 2\sigma^2}{\sigma}\right).
\end{aligned}$$

Then

$$Var(Y_t) = \frac{\exp(2\mu+2\sigma^2)}{F(t,\mu,\sigma)}\Phi\left(\frac{\ln t - \mu - 2\sigma^2}{\sigma}\right) - \left[\frac{e^{\mu+0.5\sigma^2}}{F(t,\mu,\sigma)}\Phi\left(\frac{\ln t - (\mu+\sigma^2)}{\sigma}\right)\right]^2$$

S3.2 Null features

Now let X_{null} be a zero-inflated version of Y_t with zero-inflation parameter $z = P(X_{\text{null}} = 0)$. Specifically, let $X_{\text{null}} = Y_t Z$, where Z is an independent Bernoulli random variable with parameter $1 - z$. Then

$$E(X_{\text{null}}) = E(Y_t Z) = E(Y_t)E(Z) = (1 - z)E(Y|Y \leq t).$$

.

S3.3 Spiked features

Let X_{spike} be the spiked-in version of X_{null}. Further assume that the group identifiers are given by $K \sim Bernoulli(0.5)$. SparseDOSSA assumes that:

$$(X_{\text{spike}}|K = k) = \frac{1}{1+\psi}\left[Y_t Z + \psi\left(\frac{k - E(K)}{\sqrt{Var(K)}}\sqrt{Var(Y_t)} + E(Y_t)\right)Z\right],$$

where $I(\cdot)$ is the indicator function, $E(K)$ is the expectation of the group assignment (0.5 when the samples are evenly split), and $Var(K)$ is the variance of the group assignment.

This then implies that

$$\begin{aligned}
E(X_{\text{spike}}|K = k) &= \frac{1}{1+\psi}E\left[Y_t Z + \psi\left(\frac{k - E(K)}{\sqrt{Var(K)}}\sqrt{Var(Y_t)} + E(Y_t)\right)Z\right] \\
&= \frac{1}{1+\psi}E(X_{\text{null}}) + \left(\frac{\psi}{1+\psi}\right)\left(\frac{k - E(K)}{\sqrt{Var(K)}}\right)\sqrt{Var(Y_t)}E(Z) + \\
&\quad \frac{\psi}{1+\psi}E(Y_t)E(Z) \\
&= E(X_{\text{null}}) + (1 - z)\left(\frac{\psi}{1+\psi}\right)\left(\frac{k - E(K)}{\sqrt{Var(K)}}\right)\sqrt{Var(Y_t)}
\end{aligned}$$

S3.4 Supplemental figures

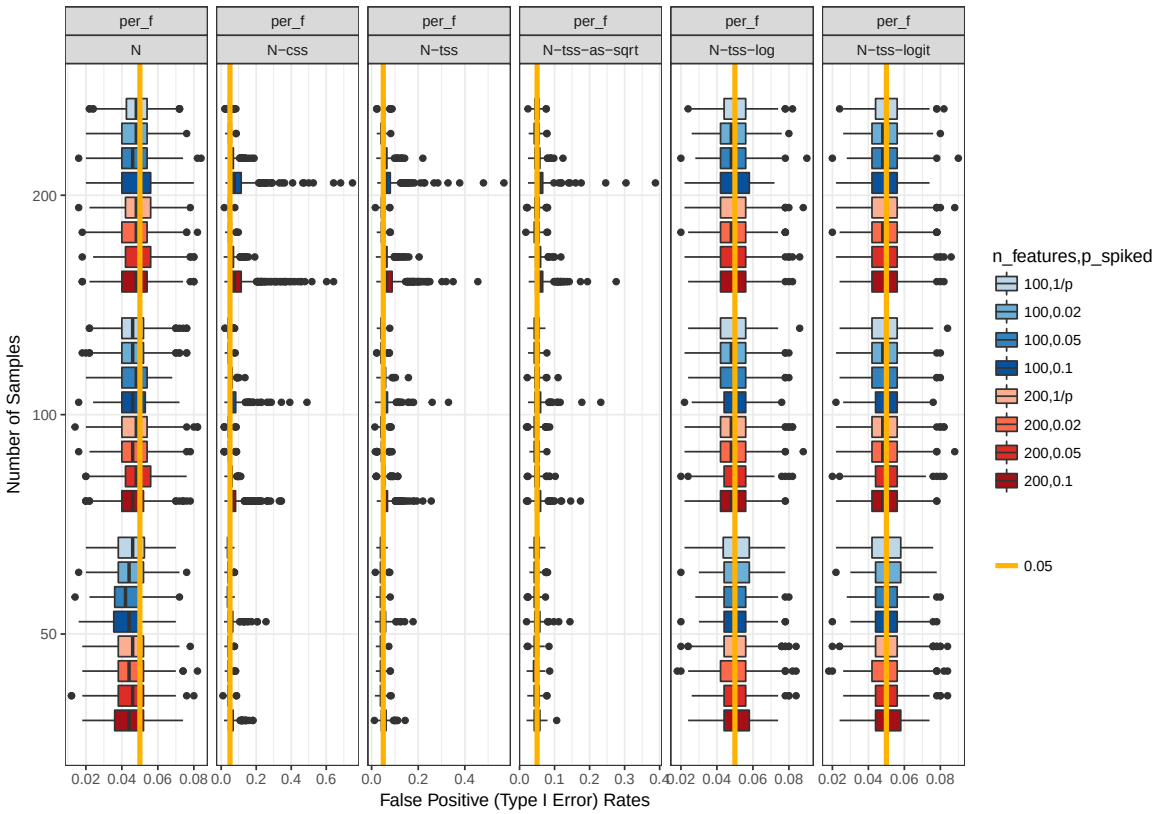

Figure S3.1: **Per-feature type I error rates.** The type I error rates of the different methods across all simulation parameters. Boxplots are the distribution of type I error rates for all null features.

References

Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., and Versalovic, J. (2014). The placenta harbors a unique microbiome. *Sci Transl Med*, 6(237):237ra65.

Aban, I. B., Cutter, G. R., and Mavinga, N. (2008). Inferences and power analysis concerning two negative binomial distributions with an application to mri lesion counts data. *Comput Stat Data Anal*, 53(3):820–833.

Aitchison, J. (1981). A new approach to null correlations of proportions. *Mathematical Geology*, 13(2):175–189.

Aitchison, J. (2003). A concise guide to compositional data analysis. In *2nd Compositional Data Analysis Workshop*.

Arthur, J. C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J. M., Fan, T.-J., Campbell, B. J., Abujamel, T., Dogan, B., Rogers, A. B., Rhodes, J. M., Stintzi, A., Simpson, K. W., Hansen, J. J., Keku, T. O., Fodor, A. A., and Jobin, C. (2012). Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science*, 338(6103):120–123.

Ban, Y., An, L., and Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*, 31(20):3322–3329.

Barberán, A., Bates, S. T., Casamayor, E. O., and Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J*, 6(2):343–351.

Becker, M. R., Paster, B. J., Leys, E. J., Moeschberger, M. L., Kenyon, S. G., Galvin, J. L., Boches, S. K., Dewhirst, F. E., and Griffen, A. L. (2002). Molecular analysis of bacterial species associated with childhood caries. *J Clin Microbiol*, 40(3):1001–1009.

Belenguer, A., Duncan, S. H., Calder, A. G., Holtrop, G., Louis, P., Lobley, G. E., and Flint, H. J. (2006). Two routes of metabolic cross-feeding between *Bifidobacterium adolescentis* and butyrate-producing anaerobes from the human gut. *Appl Environ Microbiol*, 72(5):3593–3599.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:10008.

Boris, S., Suarez, J. E., Vazquez, F., and Barbes, C. (1998). Adherence of human vaginal lactobacilli to vaginal epithelial cells and interaction with uropathogens. *Infect Immun*, 66(5):1985–1989.

Brinig, M. M., Lepp, P. W., Ouverney, C. C., Armitage, G. C., and Relman, D. A. (2003). Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease. *Appl Environ Microbiol*, 69(3):1687–1694.

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., Deng, L., Yeliseyev, V., Delaney, M. L., Liu, Q., Olle, B., Stein, R. R., Honda, K., Bry, L., and Gerber, G. K. (2016). MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol*, 17(1):121.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., and Knight, R. (2011). Global patterns of 16s rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A*, 108 Suppl 1:4516–4522.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Chalmers, N. I., Palmer, R. J., Cisar, J. O., and Kolenbrander, P. E. (2008). Characterization of a *Streptococcus* sp.-*Veillonella* sp. community micromanipulated from dental plaque. *Journal of Bacteriology*, 190(24):8145–8154.

Charlson, E. S., Bittinger, K., Haas, A. R., Fitzgerald, A. S., Frank, I., Yadav, A., Bushman, F. D., and Collman, R. G. (2011). Topographical continuity of bacterial populations in the healthy human respiratory tract. *Am J Respir Crit Care Med*, 184(8):957–963.

Chayes, F. A. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research*, 65(12):4185–4193.

Chayes, F. A. and Kruskal, W. (1966). Approximate statistical test for correlations between proportions. *The Journal of Geology*, 74(5):692–702.

Cisar, J. O., Kolenbrander, P. E., and McIntire, F. C. (1979). Specificity of coaggregation reactions between human oral streptococci and strains of *Actinomyces viscosus* or *Actinomyces naeslundii*. *Infect Immun*, 24(3):742–752.

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *ArXiv e-prints*, 70(6):066111.

Collins, M. D. (1982). Reclassification of *Bacterionema matruchotii* (mendel) in the genus *Corynebacterium*, as *Corynebacterium matruchotii* comb. nov. *Zentralblatt für Bakteriologie Mikrobiologie und Hygiene: I. Abt. Originale C: Allgemeine, angewandte und ökologische Mikrobiologie*, 3(3):364–367.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317.

DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., Sun, C. L., Goltsman, D. S. A., Wong, R. J., Shaw, G., Stevenson, D. K., Holmes, S. P., and Relman, D. A. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences of the United States of America*, 112(35):11060–11065.

Erb-Downward, J. R., Thompson, D. L., Han, M. K., Freeman, C. M., McCloskey, L., Schmidt, L. A., Young, V. B., Toews, G. B., Curtis, J. L., Sundaram, B., Martinez, F. J., and Huffnagle, G. B. (2011). Analysis of the lung microbiome in the "healthy" smoker and in copd. *PLoS One*, 6(2):e16384.

Falony, G., Vlachou, A., Verbrugghe, K., and De Vuyst, L. (2006). Cross-feeding between *Bifidobacterium longum* bb536 and acetate-converting, butyrate-producing colon bacteria during growth on oligofructose. *Appl Environ Microbiol*, 72(12):7835–7841.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). CCLasso: Correlation inference for compositional data through lasso. *Bioinformatics*, 31:3172–3180.

Fang, Z. and Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Brief Bioinform*, 12(3):280–287.

Faust, K. and Sathirapongsasuti, F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol*, 8(7):e1002606.

Finegold, S. M., Sutter, V. L., and Mathisen, G. E. (1983). *Human Intestinal Microflora in Health and Disease*, volume 1, chapter Normal indigenous intestinal flora, pages 3–31. Academic Press, New York.

Fox, G. E., Wisotzkey, J. D., and Jurtshuk, P. J. (1992). How close is close: 16s rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol*, 42(1):166–170.

Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J. M., and Huttenhower, C. (2015). Identifying personal microbiomes using metagenomic codes. *Proc Natl Acad Sci U S A*, 112(22):E2930–8.

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., and Delong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A*, 105(10):3805–3810.

Friedman, J. and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput Biol*, 8(9):e1002687.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9(3):432–441.

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., Koenig, S. S. K., Fu, L., Ma, Z. S., Zhou, X., Abdo, Z., Forney, L. J., and Ravel, J. (2012). Temporal dynamics of the human vaginal microbiota. *Science Translational Medicine*, 4(132):132ra52–132ra52.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gkantsidis, C., Mihail, M., and Zegura, E. (2003). The markov chain simulation method for generating connected power law random graphs. In *ALENEX*.

Haffajee, A. D., Socransky, S. S., Patel, M. R., and Song, X. (2008). Microbial complexes in supragingival plaque. *Oral Microbiology and Immunology*, 23(3):196–205.

Handley, P., Coykendall, A., Beighton, D., Hardie, J. M., and Whiley, R. A. (1991). *Streptococcus crista* sp. nov., a viridans streptococcus with tufted fibrils, isolated from the human oral cavity and throat. *Int J Syst Bacteriol*, 41(4):543–547.

Handley, P. S., Carter, P. L., Wyatt, J. E., and Hesketh, L. M. (1985). Surface structures (peritrichous fibrils and tufts of fibrils) found on *Streptococcus sanguis* strains may be related to their ability to coaggregate with other oral genera. *Infect Immun*, 47(1):217–227.

Handley, P. S., Correia, F. F., Russell, K., Rosan, B., and DiRienzo, J. M. (2005). Association of a novel high molecular weight, serine-rich protein (SrpA) with fibril-mediated adhesion of the oral biofilm bacterium *Streptococcus cristatus*. *Oral microbiology and immunology*, 20(3):131–140.

Hennig, C. (2015). *fpc:Flexible Procedures for Clustering*. R package version 2.1-10.

Hill, G. B. (1993). The microbiology of bacterial vaginosis. *Am J Obstet Gynecol*, 169(2 Pt 2):450–454.

Hill, M. J. (1997). Intestinal flora and endogenous vitamin synthesis. *Eur J Cancer Prev*, 6 Suppl 1:S43–5.

Hilty, M., Burke, C., Pedro, H., Cardenas, P., Bush, A., Bossley, C., Davies, J., Ervine, A., Poulter, L., Pachter, L., Moffatt, M. F., and Cookson, W. O. C. (2010). Disordered microbial communities in asthmatic airways. *PLoS One*, 5(1):e8578.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn Sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Hogg, J. S., Hu, F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C., and Ehrlich, G. D. (2007). Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biology*, 8(6):R103.

Hooper, L. V. and Gordon, J. I. (2001). Commensal host-bacterial relationships in the gut. *Science*, 292(5519):1115–1118.

Hooper, L. V., Midtvedt, T., and Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu Rev Nutr*, 22:283–307.

Hughes, C. V., Kolenbrander, P. E., Andersen, R. N., and Moore, L. V. (1988). Coaggregation properties of human oral *Veillonella* spp.: relationship to colonization site and oral ecology. *Applied and Environmental Microbiology*, 54(8):1957–1963.

Human Microbiome Project, C. (2012). A framework for human microbiome research. *Nature*, 486(7402):215–221.

Hyman, R. W., Fukushima, M., Diamond, L., Kumm, J., Giudice, L. C., and Davis, R. W. (2005). Microbes on the human vaginal epithelium. *Proc Natl Acad Sci U S A*, 102(22):7952–7957.

Jenkinson, H. F. and Lamont, R. J. (2005). Oral microbial communities in sickness and in health. *Trends Microbiol*, 13(12):589–595.

Ju, F., Xia, Y., Guo, F., Wang, Z., and Zhang, T. (2014). Taxonomic relatedness shapes bacterial assembly in activated sludge of globally distributed wastewater treatment plants. *Environmental Microbiology*, 16(8):2421–2432.

Jung, S.-H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics*, 21(14):3097–3104.

Khor, B., Gardet, A., and Xavier, R. J. (2011). Genetics and pathogenesis of inflammatory bowel disease. *Nature*, 474(7351):307–317.

Kolenbrander, P. E., Egland, P. G., Diaz, P. I., and Palmer, R. J. J. (2005). Genome-genome interactions: bacterial communities in initial dental plaque. *Trends Microbiol*, 13(1):11–15.

Kolenbrander, P. E., Palmer, R. J. J., Rickard, A. H., Jakubovics, N. S., Chalmers, N. I., and Diaz, P. I. (2006). Bacterial interactions and successions during plaque development. *Periodontol 2000*, 42:47–79.

Kolenbrander, P. E., Palmer Jr, R. J., Periasamy, S., and Jakubovics, N. S. (2010). Oral multispecies biofilm development and the key role of cell–cell distance. *Nat Rev Micro*, 8(7):471–480.

Konstantinidis, K. T. and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A*, 102(7):2567–2572.

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C., and Ley, R. E. (2013). A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol*, 9(1):1–16.

Kreth, J., Merritt, J., Shi, W., and Qi, F. (2005). Competition and coexistence between *Streptococcus mutans* and *Streptococcus sanguinis* in the dental biofilm. *Journal of Bacteriology*, 187(21):7193–7203.

Kumar, P. S., Griffen, A. L., Barton, J. A., Paster, B. J., Moeschberger, M. L., and Leys, E. J. (2003). New bacterial species associated with chronic periodontitis. *J Dent Res*, 82(5):338–344.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol*, 11(5):1–25.

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE*, 7(12):1–13.

Lancy, P., Appelbaum, B., Holt, S. C., and Rosan, B. (1980). Quantitative in vitro assay for "corncob" formation. *Infection and Immunity*, 29(2):663–670.

Lancy, P., Dirienzo, J. M., Appelbaum, B., Rosan, B., and Holt, S. C. (1983). Corncob formation between *Fusobacterium nucleatum* and *Streptococcus sanguis*. *Infection and Immunity*, 40(1):303–309.

Leyden, J. J., Marples, R. R., and Kligman, A. M. (1974). *Staphylococcus aureus* in the lesions of atopic dermatitis. *Br J Dermatol*, 90(5):525–530.

Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian Anal.*, 5(1):151–170.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1432–1440. Curran Associates, Inc.

Liu, P. and Hwang, J. T. G. (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23(6):739–746.

Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, 8:51.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., Brady, A., Creasy, H. H., McCracken, C., Giglio, M. G., McDonald, D., Franzosa, E. A., Knight, R., White, O., and Huttenhower, C. (in press). Strains, functions, and dynamics in the expanded human microbiome project. *Nature*.

Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J., and Gevers, D. (2015). Constrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol*, 33(10):1045–1052.

MacArthur, R. H. (1965). Patterns of species diversity. *Biological reviews*, 40(4):510–533.

Magurran, A. E. and Henderson, P. A. (2003). Explaining the excess of rare species in natural species abundance distributions. *Nature*, 422(6933):714–716.

Mallick, H. and Yi, N. (2013). Bayesian methods for high dimensional linear models. *Journal of biometrics & biostatistics*, 1:005–.

Marino, S., Baxter, N. T., Huffnagle, G. B., Petrosino, J. F., and Schloss, P. D. (2014). Mathematical modeling of primary succession of murine intestinal microbiota. *Proc Natl Acad Sci U S A*, 111(1):439–444.

Mark Welch, J. L., Rossetti, B. J., Rieken, C. W., Dewhirst, F. E., and Borisy, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences of the United States of America*, 113(6):E791–E800.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):1–12.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Morgan, XC and Tickle, TL, Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., Bousvaros, A., Korzenik, J., Sands, B. E., Xavier, R. J., and Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*, 13(9):R79.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing. *Journal of the American Statistical Association*, 99(468):990–1001.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E Stat Nonlin Soft Matter Phys*, 74(3 Pt 2):036104.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2):026113.

Ng, H. K. T. and Tang, M.-L. (2005). Testing the equality of two Poisson means using the rate ratio. *Stat Med*, 24(6):955–965.

Noble, W. C. (1984). Skin microbiology: coming of age. *J Med Microbiol*, 17(1):1–12.

Pace, N. R. (2009). Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev*, 73(4):565–576.

Palmer, R. J. J., Shah, N., Valm, A., Paster, B., Dewhirst, F., Inui, T., and Cisar, J. O. (2017). Interbacterial adhesion networks within early oral biofilms of single human hosts. *Appl Environ Microbiol*, 83(11).

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat Meth*, 10(12):1200–1202.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318.

Pearson, K. (1897). Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60:489–498.

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. *ArXiv Physics e-prints*.

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology*, 29(3):254–283.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76(3 Pt 2):036106.

Rajilić-Stojanović, M. and de Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *Fems Microbiology Reviews*, 38(5):996–1047.

Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson,

I. R., Sperandio, V., and Ravel, J. (2008). The pangenome structure of *Escherichia coli*: Comparative genomic analysis of e. coli commensal and pathogenic isolates. *Journal of Bacteriology*, 190(20):6881–6893.

Ren, B., Moon, Y. S., Schwager, E., Tickle, T. L., Lu, Y., Franzosa, E. A., and Huttenhower, C. (in review). A hierarchical probabilistic model of microbial community structure.

Ren, B., Schwager, E., Tickle, T. L., and Huttenhower, C. (2016). sparseDOSSA: Sparse data observations for simulating synthetic abundance.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A*, 105(4):1118–1123.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

Rousseeuw, P. J. and Kaufman, L. (1990). *Finding Groups in Data*. Wiley Online Library.

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One*, 6(12):e27310.

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., Tett, A., Morrow, A. L., and Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods*, 13(5):435–438.

Schwager, E. (2017). Banocc: Bayesian analysis of compositional covariance.

Segata, N., Bornigen, D., Morgan, X. C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun*, 4:2304.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol*, 12(6):R60.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*, 9(8):811–814.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*, 14(8):1–14.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80.

Sobel, J. D., Myers, P. G., Kaye, D., and Levison, M. E. (1981). Adherence of *Candida albicans* to human vaginal and buccal epithelial cells. *J Infect Dis*, 143(1):76–82.

Socransky, S. S. and Haffajee, A. D. (2005). Periodontal microbial ecology. *Periodontology 2000*, 38(1):135–187.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Stan Development Team (2014). RStan: the R interface to Stan, version 2.6.0.

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., Sander, C., and Xavier, J. B. (2013). Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol*, 9(12):1–11.

Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., Roehl, K. A., Nelson, D. M., Macones, G. A., and Mysorekar, I. U. (2013). Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. *Am J Obstet Gynecol*, 208(3):226.e1–7.

The Human Microbiome Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528.

Tringe, S. G. and Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet*, 6(11):805–814.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Meth*, 12(10):902–903.

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res*, 27(4):626–638.

Trüper, H. G. and De'clari, L. (1997). Taxonomic note: necessary correction of specific epithets formed as substantives (nouns)"in apposition". *International Journal of Systematic and Evolutionary Microbiology*, 47(3):908–909.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–1031.

Vatanen, T., Kostic, A. D., d'Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T. D., Hamalainen, A.-M., Peet, A., Tillmann, V., Uibo, R., Mokurov, S., Dorshakova, N., Ilonen, J., Virtanen, S. M., Szabo, S. J., Porter, J. A., Lahdesmaki, H., Huttenhower, C., Gevers, D., Cullen, T. W., Knip, M., and Xavier, R. J. (2016). Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell*, 165(4):842–853.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.*, 7(4):867–886.

Weisburg, W. G., Barns, S. M., Pelletier, D. A., and Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol*, 173(2):697–703.

Wilson, K. H. (1993). The microecology of *Clostridium difficile*. *Clin Infect Dis*, 16 Suppl 4:S214–8.

Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev*, 51(2):221–271.

Wong, F., Carter, C. K., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, 90(4):809–830.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One*, 10(7):e0129606.