Compositionally aware estimation of cross-correlations for microbiome data

Ib Thorsgaard Jensen[1*], Luc Janss[2], Simona Radutoiu[1], Rasmus Waagepetersen[3**],

1 Department of Molecular Biology and Genetics, Aarhus University, Universitetsbyen 81, 8000 Aarhus C, Denmark
2 Center for Quantitative Genetics and Genomics, Aarhus University, C. F. Møllers Allé 3, bygn. 1130, 8000 Aarhus, Denmark
3 Department of Mathematical Sciences, Aalborg University, Skjernvej 4A, 9220 Aalborg Øst, Denmark

* ibstat@mbg.au.dk
** rw@math.aau.dk

Abstract

In the field of microbiome studies, it is of interest to infer correlations between abundances of different microbes (here referred to as operational taxonomic units, OTUs). Several methods taking the compositional nature of the sequencing data into account exist. However, these methods cannot infer correlations between OTU abundances and other variables. In this paper we introduce the methods SparCEV (Sparse Correlations with External Variables) and SparXCC (Sparse Cross-Correlations between Compositional data) for quantifying correlations between OTU abundances and either continuous phenotypic variables or components of other compositional datasets, such as transcriptomic data. We compare these new methods to empirical Pearson cross-correlations after applying naive transformations of the data (log and log-TSS). Additionally, we test the centered log ratio transformation (CLR) and the variance stabilising transformation (VST). We find that CLR and VST outperform naive transformations, except when the correlation matrix is dense. For large numbers of OTUs, SparCEV and SparXCC perform similarly to CLR and VST. SparCEV outperforms all other tested methods when the number of OTUs is small (less than 100). SparXCC outperforms all tested methods when at least one of the compositional datasets has few variables (less than 50), and more so when both datasets have few variables.

Author summary

Sequencing data of the microbiome posses a unique and challenging structure that renders many standard statistical tools invalid. Features such as compositionality and sparsity complicates statistical analysis, and as a result, specialized tools are needed. Practitioners have long been interested in the construction of correlation networks within the microbiome, and several methods for accomplishing this exist. However, less attention has been paid to the estimation of cross-correlations between microbial abundances and other variables (such as gene expression data or environmental and phenotypic variables). Here, we introduce novel approaches, SparCEV and SparXCC, for inferring such cross-correlations, and compare these to transformation-based approaches, namely log, log-TSS, CLR and VST. In some cases, SparCEV and SparXCC yield superior results, while in other cases, a simpler transformation-based approach suffices. The methods are used to study cross-correlations between bacterial abundances in the skin microbiome and the severity of atopic dermatitis, as well as cross-correlations between fungal and bacterial OTUs in the root microbiome of the legume *Lotus japonicus*.

Introduction

Sequencing data are ubiquitous in modern biology [1]. For example, RNA-seq data have been used to identify genes associated with clinical outcomes of cancer patients [2], for human disease profiling [3], and to identify genes with possible links to Rett Syndrome [4]. Microbiome data have drawn much attention in recent years, particularly regarding the human gut microbiome. Composition of the human gut microbiome has been shown to be associated with several aspects of human health, such as obesity [5] and metabolic disorders [6]. More recently, the integration of microbiome data with other omics data has received increasing interest [7–11].

Data from sequencing technologies pose many specific challenges. They produce count data with technical noise, which makes results from rare features difficult to interpret. Additionally, they are compositional, meaning that the observed variables are components of an arbitrary total.

Within the field of microbiome studies, several methods have been proposed to infer interactions between microbial abundances. These include Local Similarity Analysis (LSA) [12], which finds non-linear relationships through time using time series data; Sparse Compositional Correlations (SparCC), which infers correlations based on compositional data [13]; and Sparse Inverse Covariance Estimation for Ecological Association Inference (SPIEC-EASI), which infers relations through graphical models [14]. In this paper, we focus on the estimation of correlations.

When considering correlations in the context of compositional data, there are essentially three cases of interest: A) correlations between features of the same compositional dataset, B) cross-correlations between features of a compositional dataset and non-compositional variables, and C) cross-correlations between two compositional datasets. Correlations between bacterial abundances in a microbiome is an example of case A. An example of case B is cross-correlations between gut microbes and clinical features of patients [15], and an example of case C is cross-correlations between microbial abundances and gene expression levels from RNA-seq data [7]. For an overview of these cases, see Table 1.

Case	Correlations	Methods
A	Within a composition	SparCC SPIEC-EASI LSA Pearson
B	Between a composition and an external variable	SparCEV Pearson
C	Between two compositions	SparXCC SPIEC-EASI mmVec Pearson

Table 1. Cases A, B, and C along with explanations and and overview of applicable methods.

The methods mentioned above all operate in case A. More recently, case C has gained more interest, with new methods being developed. For example, SPIEC-EASI has been extended to infer interactions between variables from two compositional datasets [16]. Like the original SPIEC-EASI, pairs of variables that are conditionally independent are identified by estimating the precision matrix using a penalized estimation scheme to enforce sparsity. The method mmVec [17] is designed for identifying interactions between OTU abundances and metabolite concentrations. This method employs a neural network to estimate the probability of observing a metabolite, given that a specific OTU is observed. It was shown to perform with similar accuracy as the extended SPIEC-EASI and to outperform correlation-based procedures. However, the correlations were estimated using a flawed methodology, where the centered log-ratio (CLR) transformation was applied to both datasets simultaneously, rather than separately, thus biasing the results. This was noted in a response from Quinn and Erb [18], who showed that when the CLR-transformation was applied appropriately, correlation-based methods outperform both mmVec and SPIEC-EASI in the setup examined by Morton et al. [17]. However, in a counter-response Morton et al. showed that it was possible to construct scenarios, where mmVec outcompeted all alternatives. Futhermore they show that mmVec is better suited to handle sparse data than correlation-based methods [19].

In this paper, we focus on inferring cross-correlations in cases B and C. Inspired by SparCC, we introduce two novel compositionally aware methods, SparCEV (Sparse Correlations with External Variables) and SparXCC (Sparse Cross-Correlations between Compositional data). Using simulation studies, we compare these methods to Pearson cross-correlations applied to various transformations of the data. Theoretical comparisons of transformation-based methods and derivations of new methods are given in the supplementary material.

Materials and methods

Modelling Sequencing Data

Let a_i denote the absolute abundance of OTU i, $i = 1, \dots, p$, $A = \sum_{j=1}^{p} a_j$, and $r_i = a_i/A$ the relative abundance. The aim of this paper is to estimate the correlation between $\log a_i$ and other log transformed variables. However, we only have access to observed read counts, denoted x_i for OTU i. To theoretically compare the different strategies and to develop new methods, we adopt a simplified modelling framework, where

$$x_i = r_i N, \tag{1}$$

where $N = \sum_{j=1}^{p} x_j$ denotes the library size. We compare the methods considered using simulations from models that are more complex and realistic than (1), see Section Simulation models. In these models, x_i given (r_i, N) is not fixed. We use the term *technical variance* for the variance of x_i given (r_i, N) and the term *biological variance* for the variance of r_i. The more realistic models are, however, intractable for theoretical analysis.

All tested methods but one require log-transformation of the x_is, which is problematic if $x_i = 0$ is observed. As a remedy, we add 1 to all read counts prior to log-transformation of the data (the pseudo-count method).

Existing Strategies for Cross-correlation Estimation

Naive transformation

We use the term *naive transformation* to refer to any transformation that does not take the compositional nature of the data into account. Naive transformations considered in this paper are log and log total sum scaling (TSS). Theoretically, naive transformations do not adequately account for the compositional structure of the data (see S1 Text). Nonetheless, it remains a common practice to apply these transformations [20] or no transformations at all [15, 21–25]. As a result, any method that outperform naive transformations would constitute an improvement relative to common practice.

Adapted Transformations

We use the term *adapted transformations* to refer to transformations that are adapted to the particular structure of the data beyond differing library sizes. In this paper, we consider the centered log-ratio (CLR) and the variance-stabilising transformation (VST). See Table 2 for definitions of all transformations (naive and adapted) considered in this paper. Some common transformations, such as trimmed M-means (TMM) [26], DESeq's median-based transformation [27], and upper-quartile transformation (UQ) [28], are not included, since they do not correct for within-replicate biases and are thus not applicable for correlation estimation.

Transformation	Expression	Interpretation
log	$\log x_i$	Log-transformed observed read counts
log-TSS	$\log \mathrm{TSS}(x_i) = \log\left(\frac{x_i}{\sum_{j=1}^{p} x_j}\right)$	The log of estimated relative abundances
CLR	$\mathrm{CLR}(x_i) = \log x_i - \frac{1}{p}\sum_{j=1}^{p} \log x_j$	Abundances relative to the average abundance
VST	$\mathrm{VST}(x_{ik}) = \int_0^{x_{ik}/\hat{s}_k} f(\mu)\mathrm{d}\mu$, where $f(\mu) = \mu^2 \hat{a}_0 + \mu(1 + \hat{a}_1)$	Removal of mean-variance relationship

Table 2. An overview of the transformations used to assess cross-correlations. In VST, k is a replicate index.

VST makes use of the DESeq modelling framework [27]. Specifically, it assumes that $x_{ik} \sim \mathrm{NB}(\mu_{ik}, \phi_i)$, where $\mu_{ik} = s_k \lambda_i$ and $\phi_i = a_0 + a_1/\lambda_i$. The estimates $\hat{a}_1$, $\hat{a}_0$, $\hat{\lambda}_i$ and $\hat{s}_k$ are obtained using the DESeq2 estimation procedures [27, 29]. Here, k denotes the index of the biological replicate. VST requires at least one feature without any zero counts. This may be violated for data with small p which is frequently the case for our simulated microbiome data. We therefore only consider the VST transformation in case C for simulated gene expression data while applying CLR to the microbiome data.

For convenience, we use the term CLR for the method where empirical Pearson cross-correlations are applied to CLR-transformed data, and likewise for log, log-TSS, and VST.

Theoretical assessment of transformations

We examine theoretically whether empirical Pearson cross-correlations combined with the transformations presented in Table 2 are likely to yield good approximations of cross-correlations. We focus on case B, since it is simpler and the results in case C are analogous. We seek an approximation of $\mathbb{C}\mathrm{orr}\,[\log a_i, b]$, where b is a non-compositional variable, here referred to as a *phenotypic variable*. By (1) and Table 2 we have,

$$x_i = \frac{a_i}{A}N, \quad \mathrm{TSS}(x_i) = \frac{a_i}{A}, \quad \mathrm{CLR}(x_i) = \log a_i - \frac{1}{p}\sum_{j=1}^{p} \log a_j.$$

By definition,

$$\mathbb{C}\mathrm{orr}\left[\log a_i, b\right] = \frac{\mathbb{C}\mathrm{ov}\left[\log a_i, b\right]}{\sqrt{\mathbb{V}\mathrm{ar}\left[\log a_i\right]}\sqrt{\mathbb{V}\mathrm{ar}\left[b\right]}}, \tag{2}$$

and we require good approximations of the numerator and the denominator. Since b is not compositional, $\mathbb{V}\mathrm{ar}\left[b\right]$ can easily be estimated without the need for approximation. According to our derivations in S1 Text, log, log-TSS, and CLR all lead to reasonable approximations of the covariance $\mathbb{C}\mathrm{ov}\left[\log a_i, b\right]$ under the model in (1) given appropriate assumptions. Furthermore, we show that

$$\mathbb{V}\mathrm{ar}\left[\mathrm{CLR}(x_i)\right] \approx \mathbb{V}\mathrm{ar}\left[\log a_i\right]$$

when p is large. However, analogous results do not hold for log and log-TSS, demonstrating that naive transformations are not sufficient. Summing up, CLR can yield good approximations under the model in (1) with the following assumptions:

(Bi) $\frac{1}{p}\sum_{j\neq i} \mathbb{C}\mathrm{ov}\left[\log a_i, \log a_j\right] \approx 0$ for all i

(Bii) $\frac{1}{p}\sum_{i=1}^{p} \mathbb{C}\mathrm{ov}\left[\log a_i, b\right] \approx 0$ for all i

(Biii) The number of OTUs, p, is large.

In case C, we we need (Bi), (Biii), and the additional assumptions

(Ci) $\frac{1}{q}\sum_{l\neq k} \mathbb{C}\mathrm{ov}\left[\log b_k, \log b_l\right] \approx 0$ for all k

(Cii) $\frac{1}{q}\sum_{l=1}^{q} \mathbb{C}\mathrm{ov}\left[\log a_i, b_l\right] \approx 0$ for all i

(Ciii) $\frac{1}{p}\sum_{j=1}^{p} \mathbb{C}\mathrm{ov}\left[\log a_j, b_k\right] \approx 0$ for all k

(Civ) The number of genes, q, is large

Conditions (Bi), (Bii) and (Ci)-(Ciii) hold if the correlation matrix is sparse; thus, following the language of Friedman and Alm [13], we refer to these as *sparsity assumptions*. This is a slight abuse of terminology, since these conditions may also hold if all rows of the correlation matrix contain entries whose distributions are symmetric around zero, even though such a matrix is not sparse.

Compositionally Aware Methods

Inspired by SparCC, we introduce compositionally aware methods for cases B and C. In case B, we assume the same sparsity condition as for CLR in the previous section. We then show in S1 Text that

$$\mathbb{C}\mathrm{orr}\left[\log a_i, b\right] \approx \frac{1}{\sigma_b \alpha_i}\frac{1}{p-1}\sum_{j\neq i} \mathbb{C}\mathrm{ov}\left[\log\frac{x_i}{x_j}, b\right],$$

where $\alpha_i^2 = \mathbb{V}\mathrm{ar}\left[\log a_i\right]$ can be estimated by SparCC and $\sigma_b^2 = \mathbb{V}\mathrm{ar}\left[b\right]$ can be estimated in a standard fashion. In contrast to CLR this method only requires conditions (Bi) and (Bii), but not (Biii). Therefore, it is likely preferable when p is small. We name this method ***Spar****se* ***C****orrelations of* ***E****xternal* ***V****ariables* (SparCEV).

In Case C, we let b_k, $k = 1, \ldots, q$, denote the gene expression level of the kth gene, $B = \sum_{k=1}^{q} b_k$, and M the library size. Similar to (1), we assume the model

$$y_k = \frac{b_k}{B} M$$

for the observed gene expression level. In S1 Text, we obtain the relation

$$\mathbb{C}\text{orr}\left[\log a_i, \log b_k\right] \approx \frac{(p-1)q\alpha_i^2 + p(q-1)\beta_k^2 + q\sum_{j\neq i}\alpha_j^2 + p\sum_{l\neq k}\beta_l^2 - t_{ik}}{(p-1)(q-1)\alpha_i\beta_k}, \qquad (3)$$

where

$$t_{ik} = \sum_{j=1}^{p}\sum_{l=1}^{q} \mathbb{V}\text{ar}\left[\log\left(\frac{x_i/x_j}{y_k/y_l}\right)\right].$$

The parameters $\alpha_i^2 = \mathbb{V}\text{ar}\left[\log a_i\right]$ and $\beta_k^2 = \mathbb{V}\text{ar}\left[\log b_k\right]$ can be approximated by applying SparCC to the microbiome and gene expression datasets individually, and the variances in t_{ik} can be estimated in a standard fashion. Details regarding efficient computation of the t_{ik}s are given in the S1 Text. As with CLR, we need the assumptions (Bi) and (Ci)-(Ciii), but unlike CLR, we do not need (Biii) and (Civ). We refer to this method as ***Spar**se Cross-**C**orrelations of **C**ompositional data* (SparXCC, where ”X” represents ”cross”).

Simulation models

We adopt the parametric model employed by *SparseDOSSA2* [30], adapting the methodology slightly to handle cases B and C. A simulated dataset contains $n \geq 1$ replicates, where n is the number of microbiome samples sequenced. The individual simulated variables (e.g. abundances or gene expression levels) are characterized by the mean, μ_i, variance, σ_i^2, and zero-probability, π_i. The parameter π_i reflects the probability that OTU i is absent from a given replicate. We refer to this as a *biological zero*. The correlation between variables is characterized by the correlation matrix R. We simulate p OTU abundances, $a_1, \ldots, a_p$ and q other variables, $b_1, \ldots, b_q$. In case B, the latter q variables are non-compositional, typically $q = 1$, and we take $\pi_{p+k} = 0$, $k = 1, \ldots, q$, so that biological zeros do not occur for the b_ks. In case C, $q > 1$ and the b_ks are compositional. The library size N^a is simulated from a log normal distribution with parameters μ_a and σ_a^2.

The simulation algorithm is given in the following steps. For ease of presentation, we present the case where $n = 1$, but when $n \geq 1$ the steps would simply be repeated n times.

1. Simulate the $p + q$-dimensional variable $g \sim \text{N}(0, R)$.
2. Define the variables Z_i for $i = 1, \ldots, p + q$ such that $Z_i = 0$ if $g_i < \Phi^{-1}(\pi_i)$ and $Z_i = F_i^{-1}(\Phi(g_i))$ otherwise, where Φ is the standard normal cumulative distribution function (cdf) and $F_i(t) = \pi_i + (1-\pi_i)\Phi((\log t - \mu_i)/\sigma_i^2)$ is the cdf of a zero-inflated log-Gaussian distribution with parameters $(\pi_i, \mu_i, \sigma_i^2)$. We now have
$$\log Z_i | Z_i \neq 0 \sim \text{N}(\mu_i, \sigma_i^2) \text{ and } P\left(Z_i = 0\right) = \pi_i.$$
3. Set $a_i = Z_i$ for $i = 1, \ldots, p$ as the absolute OTU abundances. In case C, $b_j = Z_{j+p}$ for the absolute gene expression levels. In case B, we let $b_j = \log Z_{j+p}$ for the non-compositional phenotypic variables.
4. Set $r_i^a = a_i / \sum_{k=1}^{p} a_k$ as the relative abundances of the OTUs, and in case C, set $r_j^b = b_j / \sum_{k=1}^{q} b_k$ as the relative expression levels.

5. Simulate $N^a \sim \log \mathrm{N}\left(\mu_a, \sigma_a^2\right)$ and let $\lceil N^a \rceil$ be the library size.
6. Simulate the vector, $x = (x_1, \ldots, x_p)^\top$, of observed read counts of OTU 1,...,p as $x \sim \text{Multinom}\left(\lceil N^a \rceil, r_1^a, \ldots, r_p^a\right)$.
7. In case C, simulate $N^b \sim \log \mathrm{N}\left(\mu_b, \sigma_b^2\right)$ and let $\lceil N^b \rceil$ be the library size.
8. In case C, simulate the vector, $y = (y_1, \ldots, y_q)^\top$, of observed read counts of gene $j = 1, \ldots, q$ as $y \sim \text{Multinom}\left(\lceil N^b \rceil, r_1^b, \ldots, r_p^b\right)$.

In case B, steps 7–8 are skipped. Excluding the simulation of the b_js, steps 1–6 in the above procedure are identical to the procedure employed by SparseDOSSA2. The above simulation scheme differs from the model (1) by the multinomial noise generated in steps 6 and 8 where the multinomial model is a simplistic representation of the randomness generated in the sequencing procedure. The correlation $\mathbb{C}\text{orr}\left[\log a_i, \log a_j\right]$ agrees with R_{ij} when $\pi_i = \pi_j = 0$. This does not hold in the presence of biological zeros, $\pi_i > 0$ or $\pi_j > 0$, in which case $\log a_i$ or $\log a_j$ may not even be well defined. We nevertheless use R as a ground truth for comparison with our estimates, including the case of biological zeros. In that way, the presence of biological zeros is considered a source of noise relative to our method in addition to the multinomial noise.

The correlation matrix R is constructed using two methods described in S1 Text. The first is called the cluster method, and it works by assigning a portion of the OTUs to a "cluster". All OTUs in the cluster are correlated to each other with the same correlation coefficient and uncorrelated to every other OTU. All OTUs outside the cluster are also uncorrelated with each other. In case C, a similar portion of the genes are also assigned to the cluster, and in case B, all non-compositional variables are also assigned to the cluster. This gives us a high degree of control over the degree of sparsity and the strength of the correlations. The second method is called the loadings method, and it results in a correlation matrix without exact zero entries but where most variables are only weakly correlated, with a relatively small proportion of highly correlated variable pairs. The loadings method most likely results in more realistic systems than the cluster method.

Throughout the simulations in this paper, we simulate $n = 50$ replicates. In many practical settings, n is considerably lower than that. However, for the purposes of the present simulation study, it is important that we can detect biases in the estimators. If the bias is small relative to the variance of the estimator, it may be difficult to detect in a simulation study. Since the variance of an estimator increases as n decreases, it is counter-productive to perform simulation studies with small n. In other words, we construct a situation where the main bottleneck to producing accurate results is the chosen method, not the size of the dataset.

Selecting Parameter Values

In the simulations results shown in Figs 1 and 4, the log-scale parameters, μ_i and σ_i^2, and zero-probabilities, π_i, are chosen using a real dataset as a template. We estimate the mean, μ_{ri}, and variance, σ_{ri}^2 of the observed read counts for $i = 1, \ldots, p$. We then choose μ_i and σ_i^2 such that the simulated variables have mean μ_{ri} and variance σ_{ri}^2, on the linear scale. By the properties of the log-normal distribution, the means and variances are related by

$$\sigma_i^2 = \log\left(1 + \frac{\sigma_{ri}^2}{\mu_{ri}^2}\right), \quad \mu_i = \log \mu_{ri} - \frac{\sigma_i^2}{2}. \tag{4}$$

The parameters π_i are set to half the proportion of zeros for the ith variable, with the assumption that half of the zeros are biological and the other half are technical. In case

B, the parameters of the phenotype variable are somewhat arbitrarily chosen so that it has a mean of 30 and a variance of 1.

In Fig 2, considered later, we examine the impact of diversity on the accuracy of the correlation estimation methods. We measure diversity using the *effective number of OTUs*, p_{eff}. We have $p_{\text{eff}} = \mathrm{e}^H$, where $H = \sum_{i=1}^{p} r_i \log r_i$ is the entropy or *Shannon index*. The quantity p_{eff} can be interpreted as the minimal number of OTUs such that a replicate has entropy H. This occurs when all OTUs are equally abundant. We choose $\sigma_i^2 = 1$ and $\pi_i = 0$ for $i = 1, \ldots, p$ and μ_i in such a way that we get a specific value of p_{eff} in expectation. This is accomplished by selecting the linear-scale mean relative abundances $\nu_i = \frac{1-\nu_1}{p-1}$ for $i \geq 2$ and obtaining ν_1 by solving the equation

$$\log p_{\text{eff}} = \nu_1 \log \nu_1 + \sum_{i=2}^{p} \frac{1-\nu_1}{p-1} \log \frac{1-\nu_1}{p-1} \tag{5}$$

for ν_1 given a choice of p_{eff}. We then choose an arbitrary value for the microbial load, say 1000, and set $\mu_{ri} = 1000\nu_i$. Finally, the μ_i are obtained from the right part of (4).

Methods Assessment

We might assess the accuracy of correlation estimates $\hat{\rho}_{ij}$ by comparing the true correlations to the estimated correlations by computing, for example, the mean absolute error (MAE). However, suppose we use the estimate $\hat{\rho}_{ij} = 0$ for all i, j and pick, for example, $c = 0.05$ and $\rho = 0.75$ for the cluster method. Then, even though non-zero correlations are not well estimated, the MAE, $c\rho = 0.0375$, is quite low. Thus, we separately consider the MAE of the pairs whose true correlation is zero and the MAE of the pairs whose true correlation is non-zero. In case of the loadings method, no correlations are exactly zero, so we then instead assess the MAE of pairs whose true correlation-coefficient exceeds the thresholds 0, 0.1, . . . , 0.8.

In summary, for the cluster method, we use the criteria

$$\frac{1}{|S|} \sum_{(i,j)\in S} |R_{ij} - \hat{\rho}_{ij}|, \quad \frac{1}{pq - |S|} \sum_{(i,j)\notin S} |R_{ij} - \hat{\rho}_{ij}|,$$

where $S = \{(i,j) \in A_{pq} | R_{ij} \neq 0\}$ and $A_{pq} = \{(i,j) \in \mathbb{N}^2 | 0 \leq i \leq p,\ p < j \leq p+q\}$. For the loadings method, we use the criteria

$$\frac{1}{|S_t|} \sum_{(i,j)\in S_t} |R_{ij} - \hat{\rho}_{ij}|, \text{ where } S_t = \{(i,j) \in A_{pq} || R_{ij}| \geq t\}, \text{ for } t = 0, 0.1, \ldots, 0.8, \tag{6}$$

where $t = 0$ corresponds to the overall MAE.

Discriminating between Correlated and Uncorrelated pairs

Since sparsity is only an approximate assumption, any test-statistic used to derive p-values is likely to be biased. This is exacerbated by the technical noise, which has particularly high impact for low-abundance OTUs. We shall not attempt to remedy these challenges here. Instead, we choose a dynamic threshold based on the data. Pairs whose estimated absolute correlation exceeds this threshold are considered the most likely candidates for genuinely correlated pairs. The threshold is derived in the following way. Let X and Y be the two datasets under study (in case C, Y is compositional). Permute each dataset separately. This breaks all cross-correlation, but not the correlations within each dataset. Let S_{perm} be the set of cross-correlation estimates obtained from the permuted data and let m be the 100(1-1/(pq))-percentile of S_{perm}.

We might use m as a threshold, since almost all correlations should be smaller than m. However, since all correlations are broken under the permutation, m may be too low for real data where not all correlations are zero. More specifically,

$$\mathbb{C}\text{orr}\left[\log a_i, b\right] = \rho_i^{\text{SparCEV}} + \varepsilon,$$

where ρ_i^{SparCEV} is the SparCEV estimate and where

$$\varepsilon = -\frac{1}{p-1}\sum_{j\neq i}\frac{\alpha_i}{\alpha_j}\mathbb{C}\text{orr}\left[\log\frac{x_i}{x_j}, b\right]$$

in practice may differ from zero due to violation of sparsity condition (Bii). In practice, ε may be small relative to the noise of ρ_i^{SparCEV} if n is sufficiently low. However, as n grows, the noise of ρ_i^{SparCEV} shrinks, and the relative impact of ε grows. This may lead to an unacceptable amount of false discoveries, as we see in S10 Fig. To remedy this, we add an additional user-specified parameter, t, where only cross-correlations above t are considered to be of interest. We then use $m^* = \max\{m, t\}$ as the final threshold. By S9 Fig and S10 Fig, this sufficiently controls the false discovery rate when n is large.

Data Availability

All data used in this paper can be found at https://github.com/IbTJensen/Microbiome-Cross-correlations/. The raw sequencing data from Byrd et al. [31] can be found in NCBI Bioproject 46333, and the OTU table was originally obtained from Morton et al. [32] at https://github.com/knightlab-analyses/reference-frames. The raw sequencing data from Thiergart et al. [20] can be found at the European Nucleotide Archive (ENA). The 16S dataset has project accession no. PRJEB34100, and the ITS dataset has project accession no. PRJEB34099. The OTU tables was originally obtained at https://github.com/ththi/Lotus-Symbiosis.

Implementation and Code Availability

Illustrations are produced using ggplot2 version 3.4.1 [33], ggpubr version 0.6.0 [34], and GGally version 2.1.2 [35]. The VST transformation was performed using DESeq2 version 1.34.0 [29], hypothesis testing on cross-correlations were carried out using psych version 2.2.9 [36], and the SPIEC-EASI networks were estimated using the package SpiecEasi version 1.1.2 [37]. The scripts used for the simulations and the code for SparCEV and SparXCC can be found at https://github.com/IbTJensen/Microbiome-Cross-correlations/.

Results

In this section, we compare the different estimation methods on simulated datasets with the correlation matrices constructed using the cluster and the loadings methods.

Case B

Fig 1 shows the performance of the different correlation estimation methods, with correlation matrices generated by both the cluster and the loadings method. All MAEs are computed as means over 1000 simulated datasets, with $n = 50$ replicates. For the cluster method $\rho = 0.75$ and for the loadings method $k = 5$. With both correlation generation methods, poor results are obtained when only the log-transformation is applied, and all other methods yield better results. For the cluster method, CLR and SparCEV outperform log-TSS when $c = 0.1$, and SparCEV outperforms CLR when p is small. When $c = 0.4$, log-TSS performs the same or better than CLR and SparCEV for $p \geq 100$. For $c = 0.7$, 70% of pairs are correlated, and thus the sparsity assumption is severely violated. As expected, this is a substantial obstacle to accurate estimation, especially for CLR and SparCEV. In fact, log-TSS performs similarly or better than both, except when $p = 10$, where SparCEV still has a slight edge. For the loadings method, SparCEV outperforms all alternatives when $p = 10$. For $p = 100$, the difference between SparCEV and CLR is negligible, but both outperform log-TSS. When $p = 1000$, SparCEV and CLR perform practically identically, and they only outperform log-TSS at higher thresholds and only by a small margin. For both the cluster and the loadings methods, the difference between CLR and SparCEV shrinks as p increases, which is in agreement with the theory presented in S1 Text.

Caption for Fig 1: MAE of different cross-correlation methods for correlation matrices generated by the cluster method (left column) and the loadings method (right column). For the cluster method, different p (number of OTUs) and c (the proportion of OTUs in a cluster) are used. For the loadings method, threshold values $t = 0, 0.1, \ldots, 0.8$ and different p are used. The lines show the mean accuracy, and the edges of the envelopes show ± 1.96 standard errors (SE). The results are based on 1000 simulated datasets where each simulated dataset has 50 replicates.

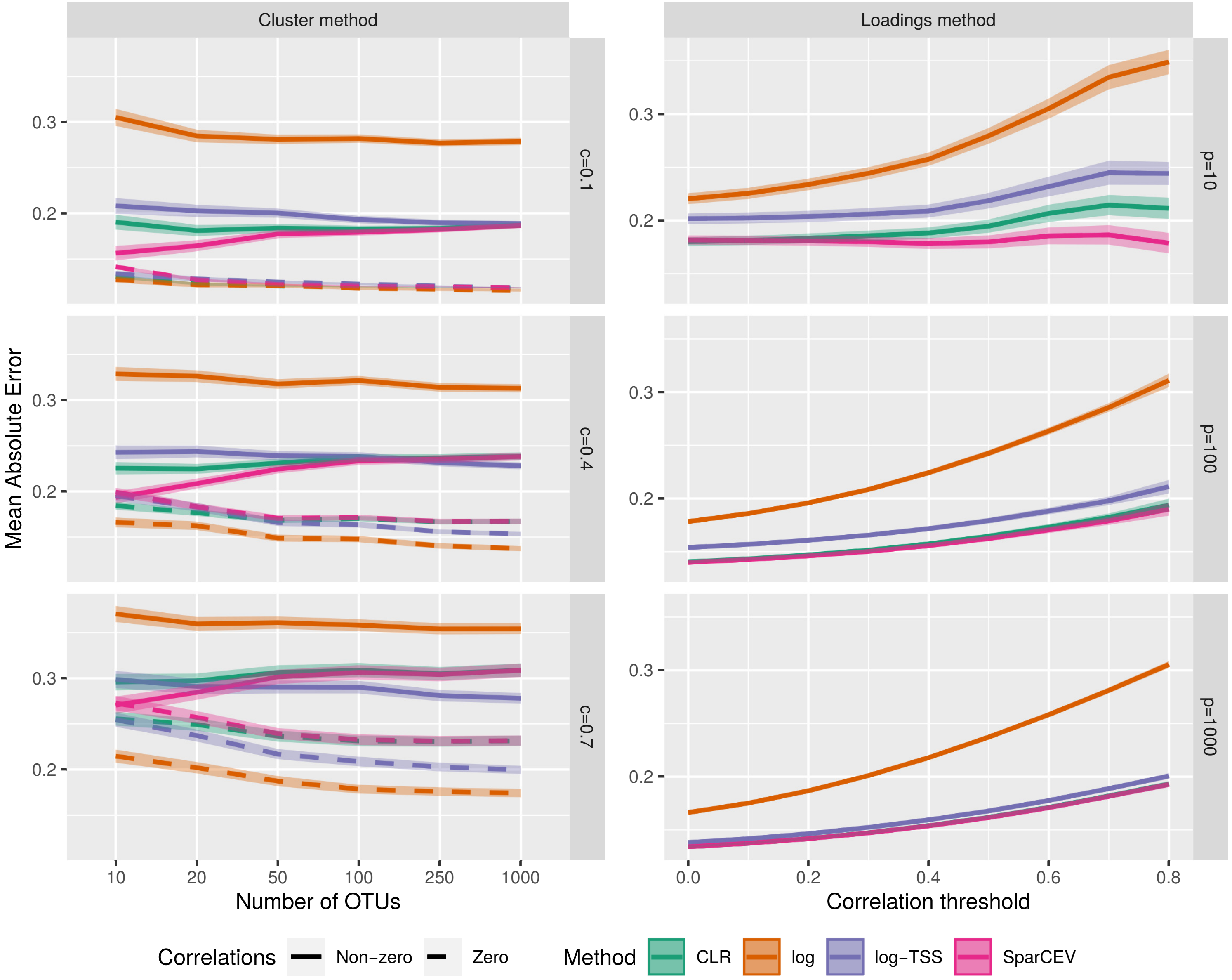

Fig 1. MAE of different cross-correlation methods for correlation matrices generated by the cluster method (left column) and the loadings method (right column). For the cluster method, different p (number of OTUs) and c (the proportion of OTUs in a cluster) are used. For the loadings method, threshold values $t = 0, 0.1, \ldots, 0.8$ and different p are used. The lines show the mean accuracy, and the edges of the envelopes show ±1.96 standard errors (SE). The results are based on 1000 simulated datasets where each simulated dataset has 50 replicates.

The general pattern observed in Fig 1 is that log yields the worst results, log-TSS is an improvement, CLR and SparCEV outperform log-TSS (except when sparsity is badly violated), and SparCEV outperforms CLR at low p. This behavior is consistent with the theory presented in S1 Text. Situations where p is small may be encountered in practice, for example, when abundances at high taxonomic levels are considered or when synthetic communities are employed, as is sometimes done in the plant field [25].

The Effect of Diversity

Friedman and Alm [13] showed that the accuracy of correlation estimates in case A depends on the diversity of the microbiome. They show that the accuracy of empirical Pearson correlation estimates decreases as p_{eff} increases, whereas SparCC is unaffected by p_{eff}. Fig 2 shows the impact of diversity in case B with $p = 100$ and different average p_{eff}. The simulation settings are identical to those in Fig 1 for $p = 100$, except that we choose $(\mu_i, \sigma_i^2, \pi_i)$ differently (see Parameter Selection under Material and Methods). Specifically, we set $\pi_i = 0$ for all i to avoid zero inflation. This is because a more zero-inflated dataset will tend to have lower entropy (and thus lower p_{eff}) than a less zero-inflated dataset. This introduces a chaotic element to the simulation process that may muddle the patterns we seek to investigate.

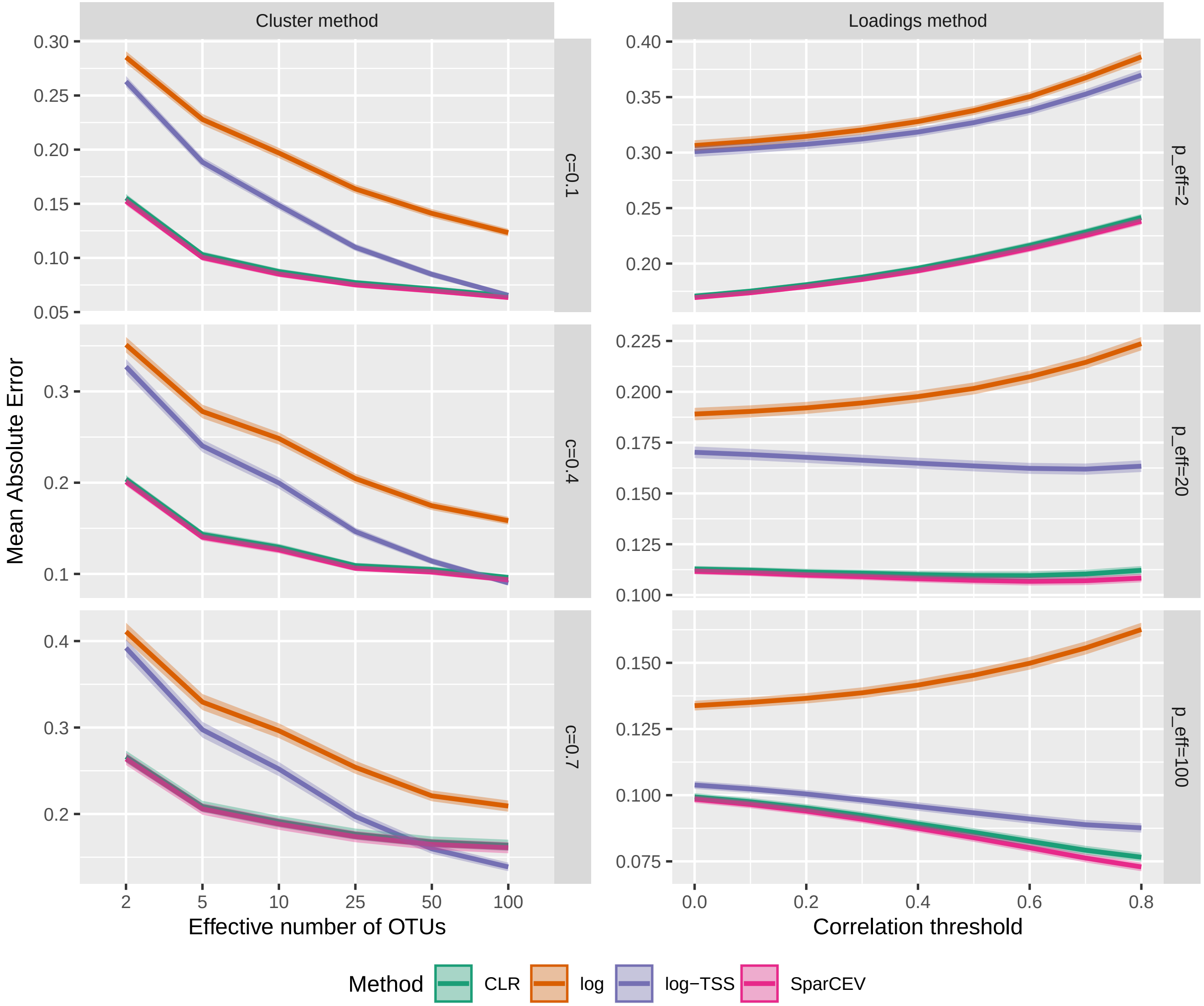

Fig 2. MAE of different cross-correlation methods for correlation matrices generated by the cluster method (left column) and the loadings method (right column). For the cluster method, different p_{eff} (effective number of OTUs) and c (the proportion of OTUs in a cluster) are used. For the loadings method, threshold values $t = 0, 0.1, \ldots, 0.8$ and different p_{eff} are used. The lines show the mean accuracy, and the edges of the envelopes show ±1.96 SE. The results are based on 1000 simulated datasets where each simulated dataset has 50 replicates.

In Fig 1, we show two sets of lines for the cluster method, one for correlated pairs and one for uncorrelated pairs. In Fig 2, these lines would have fallen on top of each other, so for ease of presentation, the lines for the uncorrelated pairs have been omitted. The results for uncorrelated pairs are instead shown in S2 Fig, where the overall pattern is similar to Fig 2. However, for uncorrelated pairs, both log and log-TSS perform better than CLR and SparCEV when p_{eff} is high and sparsity is violated.

In Fig 2, we see that SparCEV and CLR are both only mildly affected by the effective number of OTUs for correlation matrices generated by both the cluster method and the loadings method, regardless of threshold or density of the correlation matrix. SparCEV still consistently outperforms CLR, although the difference is negligible (for all p_{eff}, the difference is similar to the difference we saw in Fig 1 at $p = 100$). However, the accuracy of the results obtained from log and log-TSS depend heavily on p_{eff}. The accuracy of log-TSS is similar to that of CLR and SparCEV only for dense correlation matrices with uniformly distributed abundances, which are unlikely to occur in nature. In general, the benefit of using CLR or SparCEV is greater for less diverse microbiota. This is consistent with established knowledge in case A [13].

Application on Atopic Dermatitis Data

In this section, we analyze the correlations found in an atopic dermatitis dataset from Byrd et al. [31]. The severity of the symptoms was quantified by the widely used measure *objective SCORing of Atopic Dermatitis* (objective SCORAD) [38]. We estimate the correlations between objective SCORAD and bacterial abundances at the family level using SparCEV. We obtained a correlation threshold of 0.52 using the threshold selection approach described in Materials and Methods. Additionally, we calculated empirical bootstrap confidence intervals (CI) with the BC_a-method according to Efron [39]. The families with absolute correlation with SCORAD exceeding 0.52 are shown in Fig 3A with the exception of two families whose 95%-CIs included zero. These families were *Nocardioidaceae* (Estimate: -0.54, 95%-CI: [-0.76, 0.03]) and *Nakamurellaceae* (Estimate: -0.55, 95%-CI: [-0.77, 0.02]).

It is well known that colonization by *Staphylococcus aureus* can exacerbate the severity of atopic dermatitis [31]. Indeed, we find that *Staphylococcaceae* is the family whose abundance has the strongest positive correlation with the objective SCORAD (estimate: 0.65, 95%-CI: [0.34, 0.81]), see Fig 3A and Fig 3B. Some members of the fungal family *Malasseziaceae* are believed to play a pathogenic role in atopic dermatitis [40, 41], and this family also appears to be positively correlated with the objective SCORAD, although it is not above our permutation threshold (estimate: 0.42, 95%-CI: [0.08, 0.74]). Other studies found that the relative abundance of the genus *Propionibacterium* was depleted in patients with atopic dermatitis [42] and that the genus *Cutibacterium* may inhibit the growth of *Staphylococcus aureus* [43]. Both these genera are members of the family *Propionibacteriaceae*, but we did not find a correlation between the objective SCORAD and the abundance of this family (estimate: -0.20, 95%-CI: [-0.56, 0.20]). The strongest negative correlation detected was with the family *Hyphomicrobiaceae* (estimate: -0.73, 95%-CI: [-0.84, -0.51]), but to our knowledge, this family is not known to play a role in atopic dermatitis. According to Byrd et al., the two predominant viruses were polyomaviruses and papillomaviruses. The Byrd et al. data indicate a possible negative correlation with *Polyomaviridae* (estimate: -0.41, 95%-CI: [-0.71, -0.04]), albeit below our threshold, while *Papillomaviridae* is uncorrelated with SCORAD (estimate: -0.16, 95%-CI: [-0.62, 0.22]). All correlations and their confidence intervals can be found in S1 Table.

Fig 3C shows that the diversity (as measured by effective number of families) is negatively correlated with the objective SCORAD score. This is consistent with prior knowledge that the diversity of the skin microbiome is substantially reduced in atopic dermatitis patients [42, 43]. The effective number of families is only 18 (out of 407 observed families) even in the most diverse replicate (the effective number of families in the least diverse replicate is 1.2, with over 96% of the relative abundance occupied by *Staphylococcaceae*). Thus, the diversity in all replicates is low, and by Fig 2 we expect substantially more accurate correlation estimates from SparCEV or CLR compared with log-TSS. According to Fig 3D, the estimates using log-TSS are consistently smaller than those of SparCEV. Assuming the SparCEV estimates are more accurate, as is indicated by Fig 1 and Fig 2, the log-TSS estimates possibly mask potential positive correlations

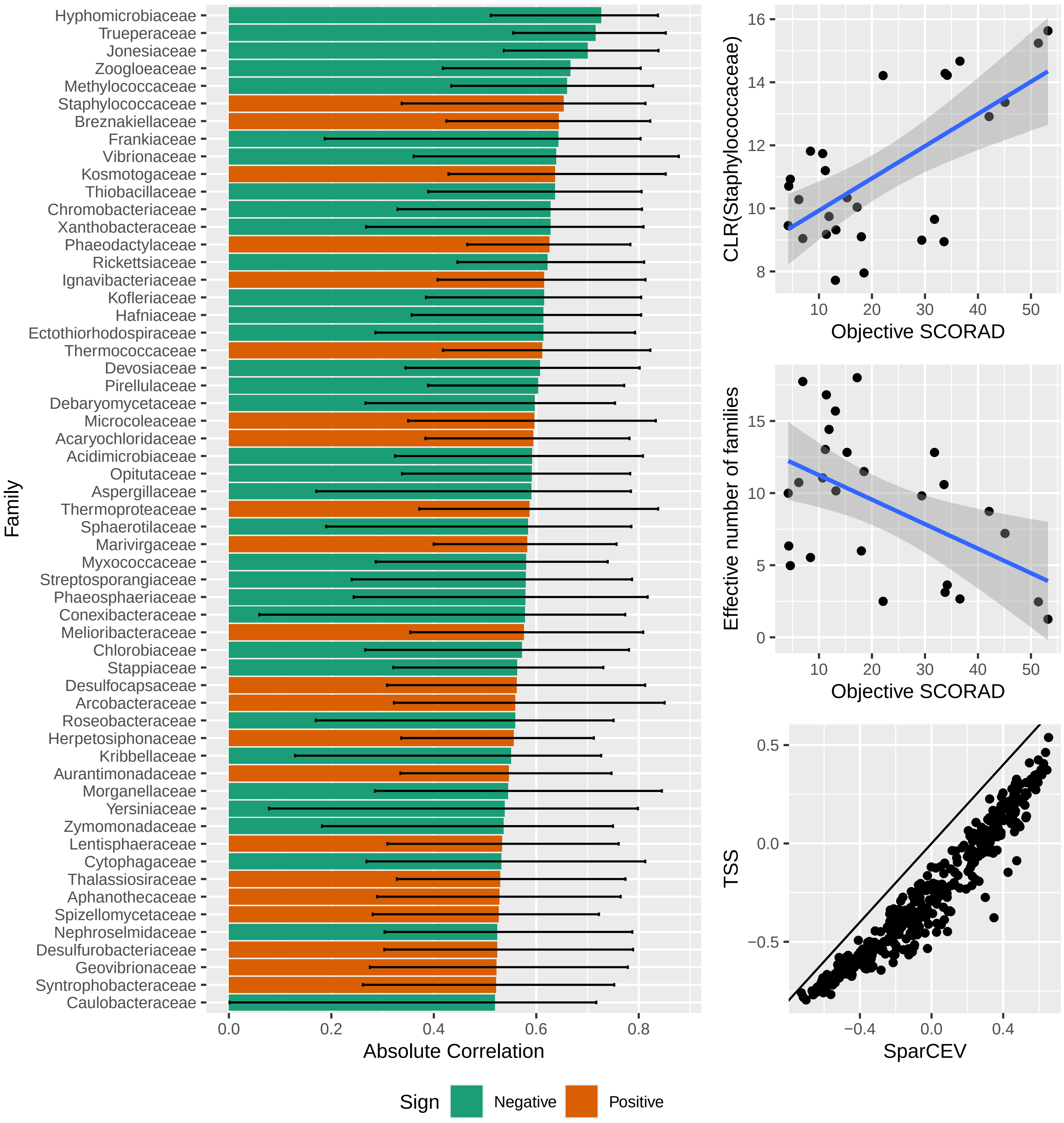

Fig 3. Results from a correlation analysis on atopic dermatitis data from Byrd et al. [31]. **A**: All correlations exceeding the permutation threshold of 0.52 with color according to the sign of the correlation and with error bars given by the empirical bootstrap 95%-confidence interval. **B**: Scatter plot between the CLR-transformed abundance of *Staphylococcaceae* and the objective SCORAD. The blue line is derived from a smooth line fitted to the data with error bars derived from the standard deviation. **C**: Scatter plot between the effective number of families and the objective SCORAD. **D**: Scatter plot between the estimated correlation using log-TSS and SparCEV. The straight line has slope 1 and intercept 0.

and exaggerate negative correlations.

Case C

We repeated the numerical studies from Fig 1 in case C. The left column in Fig 4 shows results with the correlation matrices obtained using the cluster method with $c = 0.1$. For results with other values of c, see S4 Fig. On Fig 4, all tested methods perform similarly on non-correlated pairs and on correlated pairs; CLR and CLR+VST yield almost identical results and outperform log-TSS, while SparXCC is superior when p or q is small, in agreement with the theory. On S4 Fig, we see that for $c = 0.4$, the performance lead for SparXCC is reduced and for $c = 0.7$ it is almost nonexistent. Just as in case B, a non-sparse correlation matrix is a considerable obstacle for SparXCC, CLR, and CLR+VST. They are all outcompeted by log-TSS in this setting. With the loadings method, we see a similar pattern as in case B. When $p = q = 10$, SparXCC outcompetes all alternatives, especially for higher thresholds. For $p = q = 100$ the difference between SparXCC and CLR or CLR+VST is reduced and for $p = q = 1000$, SparXCC, CLR, and CLR+VST all perform identically and all outperform log-TSS. In S5 Fig, we see results for all tested combinations of p and q. Here, we see that SparXCC outperforms all tested transformation-based methods when either p or q is sufficiently small. We also carried out the simulations without zero-inflation and found practically identical results to those seen on Fig 4. We note that situations where p, q or both are small may arise for example when examining correlations between 16S data (bacterial OTUs) and ITS data (fungal OTUs). Specifically, when synthetic communities are employed or when correlations at a high taxonomic level are of interest.

In case B, SparCEV did not have any meaningful drawbacks in relation to CLR. There are cases where they are practically identical, and then there are cases where SparCEV outcompetes CLR. It is not as simple in case C, since SparXCC requires considerably more computational resources than SparCEV. As a result, it is advisable to carefully consider one's specific dataset before choosing a method. When both p and q are large, there is little gain in using SparXCC over CLR, but it runs substantially slower.

Application on plant microbiome data

In this section, we analyze the correlations found in the root microbiome of *Lotus japonicus* in a dataset by Thiergart et al. [20]. We have two sequencing datasets (each compositional), one from 16S ribosomal RNA (rRNA) and one from internal transcribed spacers (ITS). The 16S data contains bacterial OTUs, and the ITS data contains fungal OTUs. The data are from plants of multiple genotypes, the wild type (Gifu) and the mutants *ccamk*, *symrk*, *ram1*, and *nfr5*. The data contains 15–22 replicates for each genotype. We estimate the correlations using SparXCC. The replicates within each genotype originate from three different experiments. This potentially has a confounding effect on the results. For the purposes of this example, we employ the function `RemoveBatcheffect` from the R-package `limma` [44] to correct for differing means between experiments. The results can be seen on Fig 5. More details on correlated OTUs can be found in S2 Table. A similar analysis was carried out on data collected from the rhizosphere of the plant. The results of this can be found in S7 Fig and S3 Table.

Thiergart et al. estimate cross-correlations between bacterial and fungal OTUs only on the Gifu data using Spearman correlations on TSS-transformed data (Spearman-TSS). They consider a pair correlated when the p-value is less than 0.001 and thus obtain 585 pairs with significant correlations. Using permutation thresholding on SparXCC, we find only 12 correlated pairs in Gifu when correcting for confounding

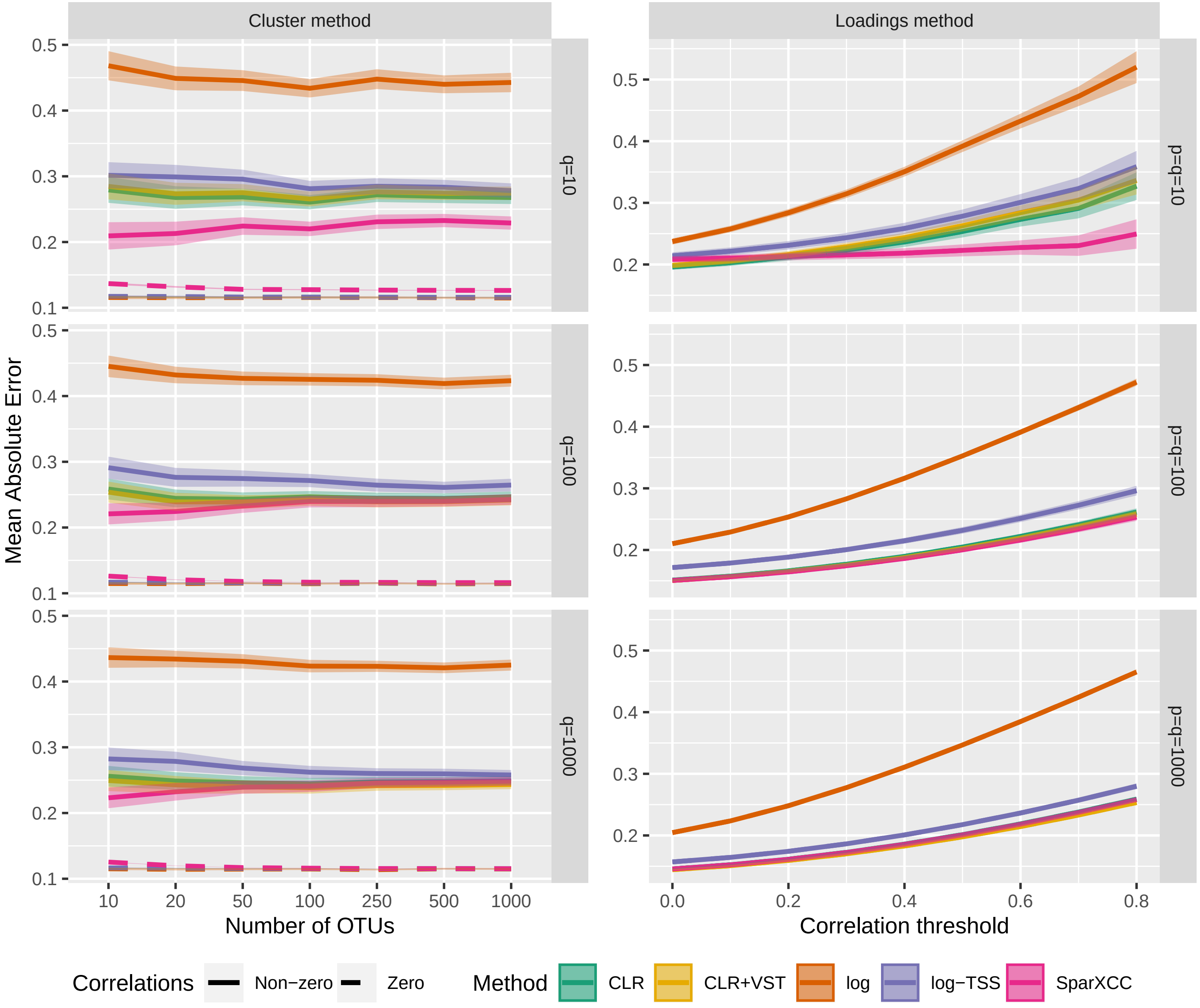

Fig 4. MAE of different cross-correlation methods for correlation matrices generated by the cluster method (left column) and the loadings method (right column) in case C. For the cluster method, different p (number of OTUs) and q (number of genes) are used. For the loadings method, threshold values $t = 0, 0.1, \ldots, 0.8$ and different p and q are used. The lines show the mean accuracy, and the edges of the envelopes show ±1.96 standard errors (SE). The results are based on 500 simulated datasets where each simulated dataset has 50 replicates.

effects. In order to make a direct comparison to the results in the original paper, we also carried out the correlation estimation without correcting for confounding effects. We then obtain 1229 correlations above the threshold. A substantial proportion of correlations identified by Thiergart et. al were also found by SparXCC (66 %). On S6 Fig and in S4 Table, we see that in most cases, the methods find similar estimates, but in some cases they may differ considerably. In fact, in some cases, the two methods disagree on the sign of the correlation. The reason for the differences between the methods may be that SparXCC approximates Pearson correlations, which measure

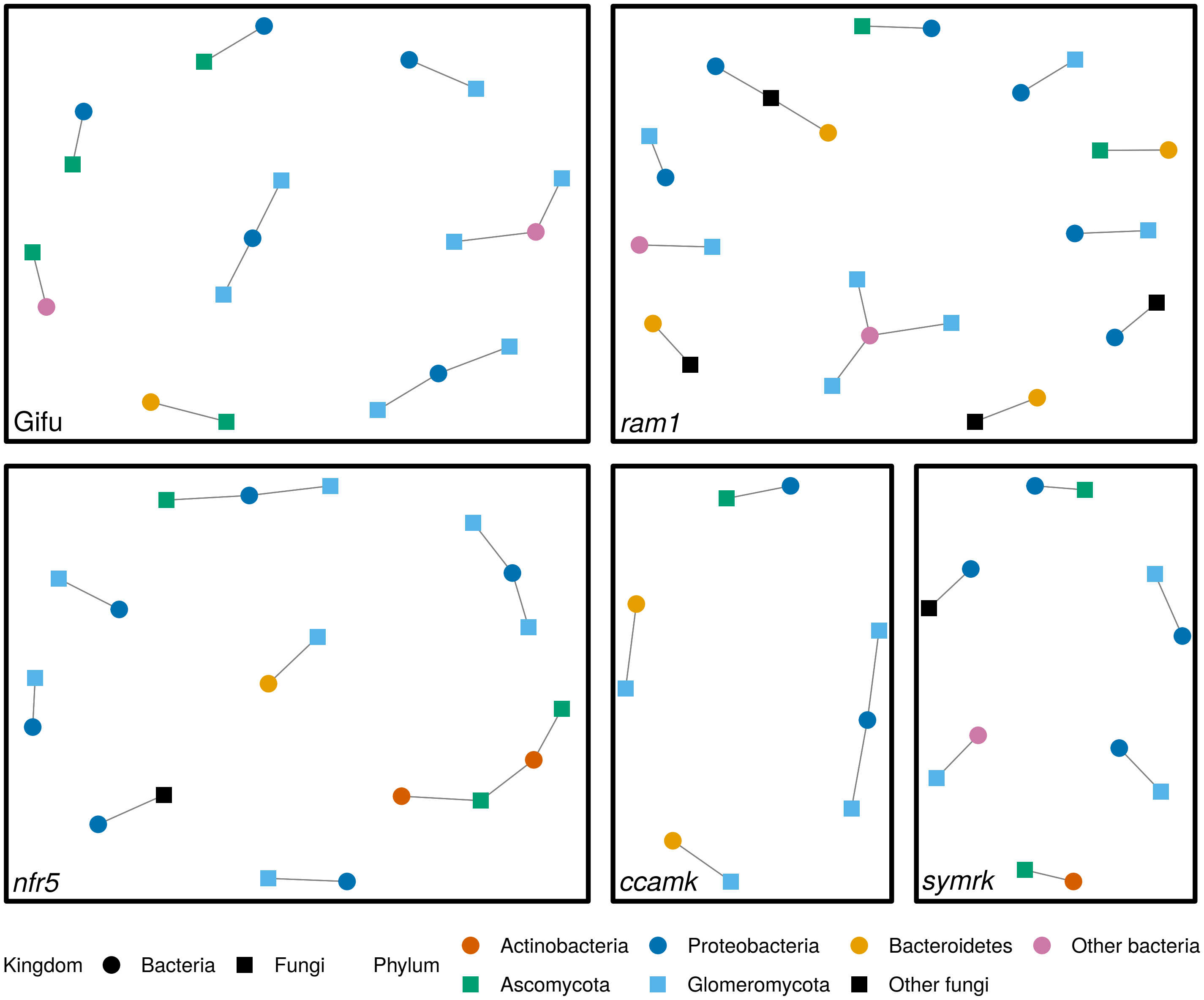
Gifu
ram1
nfr5
ccamk
symrk
Kingdom ● Bacteria ■ Fungi
Phylum ● Actinobacteria ● Proteobacteria ● Bacteroidetes ● Other bacteria
■ Ascomycota ■ Glomeromycota ■ Other fungi

Fig 5. Results from applying CLR to 16S and ITS sequencing data from the root microbiome of *Lotus japonicus*, from Thiergart et al. [20]. Each circular vertex represents a bacterial OTU from the 16S data and a square vertex represents a fungal OTU from the ITS data. Vertices are colored based on the phylum of the OTU it represents. Two vertices are connected by an edge if their estimated correlation is above the permutation threshold. The analysis is carried out separately for the genotypes Gifu, *ram1*, *nfr5*, *ccamk*, and *symrk*. Only cross-correlations are shown.

linear relationships, while Spearman correlations measure monotonic relationships. To examine this possibility, we also estimated the Pearson correlations of the log-TSS transformed data (Pearson-log-TSS) and found 1395 pairs with significant correlations. Interestingly, we found that Pearson-log-TSS showed a greater degree of overlap with Spearman-TSS than with SparXCC (87 % vs 72 %).

Of the pairs where SparXCC and Spearman-TSS disagreed on the sign, 21 were above the permutation threshold, but not detected as significant by the t-test. All of these pairs involved two specific fungal OTUs, both members of the phylum

Ascomycota. Both had many reads (ranging from 95 to 2576), so these results are not an artifact of low read counts. Additionally, all of these pairs showed the same pattern when comparing SparXCC to Pearson-log-TSS; the estimated correlations had different signs, but were not detected as significant by the t-test. This suggests that the model structure in SparXCC is able to capture some pattern of association that is lost with the transformation-based methods.

Discussion

For the theoretical considerations in this paper, we, like Friedman and Alm [13], assume that the data follow the model in (1). According to this model, the true relative abundances r_i are observed, which would only be the case with infinite sequencing depth. We nevertheless assess the different correlation estimation methods using data simulated under a more realistic setting where the x_is are noisy observations of the r_is. Specifically, SparseDOSSA2 assumes that the x_is are multinomial, given the library size N and the r_is. Friedman and Alm [13] suggests mitigating the impact of the technical variance of x_i given (r_i, N) by using a Monte Carlo sampling procedure based on a uniform Dirichlet prior. However, in our simulation setup, we find that this reduces accuracy compared to using a pseudo-count (See Fig S8 Fig). It is a topic of further research to investigate the nature of the technical variance (e.g. if it is truly generated by a multinomial model, as postulated by SparDossa2) and how to account for it in the cross-correlation estimations.

We did not consider testing null hypotheses of zero correlation. Due to various sources of bias, including the aforementioned technical variance, it is difficult to base hypothesis testing on theoretical results. Friedman and Alm [13] use a bootstrapping procedure when applying SparCC in case A. This is computationally demanding, however, not least since corrections for multiple testing are needed when carrying out hypothesis testing for a large number of correlations. Furthermore, in cases B and C, it is challenging to construct a bootstrap simulation scheme that respects the null hypothesis for a particular correlation while maintaining the remaining correlation structure. Due to these difficulties, we believe that it may be more appropriate to rely on the correlation estimates themselves, as we have done with the permutation threshold selection.

A fundamental assumption in this paper is that the interactions between microbes and other variables can be adequately described by a correlation matrix. To our knowledge, no alternatives have been unambiguously shown to universally better describe interactions between compositional datasets such as microbiome and RNA-seq data. Which metric is more sensible may depend on the underlying biology of the specific data under study. We compared the performance of SPIEC-EASI and correlation-based approaches in case C. SPIEC-EASI uses a penalized regression scheme to estimate the precision matrix, $R^{\text{-}1}$, and does not aim to estimate the correlation matrix, R, directly. Instead the primary aim is to discriminate between pairs that are conditionally independent and pairs that are not. With a correlation matrix constructed using the cluster method, a pair is uncorrelated if and only if it is conditionally independent (i.e $R_{ij} = 0 \Leftrightarrow R^{-1}_{ij} = 0$). Thus we can directly compare the power and false discovery rate (FDR) of SPIEC-EASI with those found using pair-wise correlations. To do this, we subjected the correlation estimates of CLR to a t-test and the estimates of SparXCC to the permutation thresholding scheme described in the material and methods section as well as using SPIEC-EASI to identify conditionally dependent pairs. Compared to the thresholding method, we found that SPIEC-EASI has higher power for $n \leq 50$ but with a much higher FDR. The t-test had better power than the other methods tested for $n \geq 50$, but FDR increased as n increased. However, it was able to adequately control the FDR at $n \leq 50$. Permutation thresholding saw relatively high FDR at $n = 20$, but was otherwise able to better control the FDR than the other methods tested, and it was nearly as powerful as the t-test. See S10 Fig for similar simulations in case B, comparing a t-test, permutation thresholding with m as a threshold, and permutation thresholding with m^* as a threshold.

The interactions present in a real biological system are likely to be more complicated than a correlation matrix generated by the cluster method can account for. In such cases, the methods may diverge (in general, $R^{-1}_{ij} = 0$ need not imply that $R_{ij} = 0$ or vice versa), and it may not be clear which is more appropriate.

Conclusion

When estimating correlations between compositional variables and non-compositional variables (case B), the results in Fig 1, Fig 2, and S1 Fig suggest that SparCEV should be the method of choice. When the number of compositional variables is sufficiently large (above 100 in our results), empirical Pearson correlations following the CLR-transformation perform essentially just as well. When estimating cross-correlations between two compositional datasets (case C), the results in Fig 4, S3 Fig, S4 Fig, and S5 Fig suggest that the method of choice should be SparXCC. However, when both p and q are large, CLR may be preferred, since it performs essentially just as well in these cases, but with considerably less computational complexity.

Supporting information

S1 Fig. Case B without biological zeros Accuracy of the different cross-correlation methods in case B, in the absence of biological zero by enforcing $\pi_j = 0$ for $j = 1, \ldots, p$. Otherwise, the same simulation settings as Fig 1 are used.

S2 Fig. Case B diversity and zero correlations Accuracy of the different cross-correlation methods in case B on uncorrelated pairs at different levels of diversity.

S3 Fig. Case C without biological zeros Accuracy of the different cross-correlation methods in case C, in the absence of biological zero by enforcing $\pi_j = 0$ for $j = 1, \ldots, p + q$. Otherwise, the same simulation settings as Fig 4 are used.

S4 Fig. Cluster method in case C for large *c* Accuracy of the different cross-correlation methods on correlation matrices generated by the cluster method in case C for $c = 04, 0.7$. Otherwise, the same simulation settings as Fig 4 are used.

S5 Fig. Loadings method in case C for all combinations of *p* and *q* Accuracy of the different cross-correlation methods on correlation matrices generated by the loadings method in case C for all combinations of $p = 10, 100, 1000$ and $q = 10, 100, 1000$. Otherwise, the same simulation settings as Fig 4 are used.

S6 Fig. Spearman correlations of relative abundances vs SparXCC The estimated correlation coefficients as estimated by Spearman correlations of relative abundances plotted against correlations approximated by SparXCC. For Spearman, a pair is considered correlated when a t-test returns a p-value less than 0.001. For SparXCC, a pair is considered correlated when it is above the permutation threshold.

S7 Fig. Cross-correlation network constructed on rhizosphere data Graph with edges between nodes when the cross-correlation is above a permutation threshold, estimated by SparXCC on rhizosphere data.

S8 Fig. Pseudo-count versus Dirichlet Monte Carlo sampling Accuracy of using a pseudo-count versus Dirichlet Monte Carlo for SparCEV.

S9 Fig. Separating correlated and uncorrelated pairs in Case C Power and FDR of CLR with a t-test, SparXCC with permutation thresholding, and SPIEC-EASI.

S10 Fig. Separating correlated and uncorrelated pairs in Case B Power and FDR of CLR with a t-test and SparCEV with permutation thresholding, using both m and m^* as defined in section Discriminating between Correlated and Uncorrelated pairs.

S1 Table Correlations between families and objective SCORAD score.

S2 Table Correlations between bacterial OTUs from 16S data and fungal OTUs from ITS data from the root of *Lotus japonicus*. Confounding effects from the experiment effect were removed and SparXCC was applied.

S3 Table Correlations between bacterial OTUs from 16S data and fungal OTUs from ITS data from the rhizosphere of *Lotus japonicus*. Confounding effects from the experiment effect were removed and SparXCC was applied.

S4 Table Correlations between bacterial OTUs from 16S data and fungal OTUs from ITS data from the root of *Lotus japonicus*. The data was not corrected for confounding effects prior to correlation estimation.

S1 Text Theoretical analysis of transformation-based correlations, derivation of compositionally aware methods, and construction of correlation matrices.

Acknowledgments

Funding: This work was supported by the Bill and Melinda Gates Foundation and from Foreign, Commonwealth & Development Office through Engineering the Nitrogen Symbiosis for Africa (ENSA; OPP11772165). We thank Adrián Gómez Repollés for assistance with the dermatitis data. We thank Thorsten Thiergart and Ruben Garrido-Oter for assistance with the plant microbiome data. We thank B Kirtley Amos and Max Gordon for critical reading. We thank Ke Tao and Sha Zhang for supplying the data used to construct the templates for the simulation studies. We thank Taylor Grace FitzGerald for copy-editing.

References

1. McCombie WR, McPherson JD, Mardis ER. Next-Generation Sequencing Technologies. Cold Spring Harbor Perspectives in Medicine. 2019;9(11):a036798. doi:10.1101/cshperspect.a036798.
2. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357(6352):eaan2507. doi:10.1126/science.aan2507.

3. Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. International Journal of Molecular Sciences. 2017;18(8):1652. doi:10.3390/ijms18081652.

4. Ehrhart F, Coort SL, Eijssen L, Cirillo E, Smeets EE, Bahram Sangani N, et al. Integrated analysis of human transcriptome data for Rett syndrome finds a network of involved genes. The World Journal of Biological Psychiatry. 2020;21(10):712–725. doi:10.1080/15622975.2019.1593501.

5. Cani PD. Human gut microbiome: hopes, threats and promises. Gut. 2018;67(9):1716–1725. doi:10.1136/gutjnl-2018-316723.

6. Wilson AS, Koller KR, Ramaboli MC, Nesengani LT, Ocvirk S, Chen C, et al. Diet and the Human Gut Microbiome: An International Review. Digestive Diseases and Sciences. 2020;65(3):723–740. doi:10.1007/s10620-020-06112-w.

7. Huang H, Ren Z, Gao X, Hu X, Zhou Y, Jiang J, et al. Integrated analysis of microbiome and host transcriptome reveals correlations between gut microbiota and clinical outcomes in HBV-related hepatocellular carcinoma. Genome Medicine. 2020;12(1):102. doi:10.1186/s13073-020-00796-5.

8. Zancarini A, Westerhuis JA, Smilde AK, Bouwmeester HJ. Integration of omics data to unravel root microbiome recruitment. Current Opinion in Biotechnology. 2021;70:255–261. doi:10.1016/j.copbio.2021.06.016.

9. Shaffer M, Armstrong AJS, Phelan VV, Reisdorph N, Lozupone CA. Microbiome and metabolome data integration provides insight into health and disease. Translational Research. 2017;189:51–64. doi:10.1016/j.trsl.2017.07.001.

10. Monteleone AM, Troisi J, Fasano A, Dalle Grave R, Marciello F, Serena G, et al. Multi-omics data integration in anorexia nervosa patients before and after weight regain: A microbiome-metabolomics investigation. Clinical Nutrition. 2021;40(3):1137–1146. doi:10.1016/j.clnu.2020.07.021.

11. Korenblum E, Dong Y, Szymanski J, Panda S, Jozwiak A, Massalha H, et al. Rhizosphere microbiome mediates systemic root metabolite exudation by root-to-root signaling. Proceedings of the National Academy of Sciences. 2020;117(7):3874–3883. doi:10.1073/pnas.1912130117.

12. Xia LC, Ai D, Cram J, Fuhrman JA, Sun F. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. Bioinformatics. 2013;29(2):230–237. doi:10.1093/bioinformatics/bts668.

13. Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. PLoS Computational Biology. 2012;8(9):e1002687. doi:10.1371/journal.pcbi.1002687.

14. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology. 2015;11(5):e1004226. doi:10.1371/journal.pcbi.1004226.

15. Xu X, Zhang W, Guo M, Xiao C, Fu Z, Yu S, et al. Integrated analysis of gut microbiome and host immune responses in COVID-19. Frontiers of Medicine. 2022;16(2):263–275. doi:10.1007/s11684-022-0921-6.

16. Tipton L, Müller CL, Kurtz ZD, Huang L, Kleerup E, Morris A, et al. Fungi stabilize connectivity in the lung and skin microbial ecosystems. Microbiome. 2018;6(1):12. doi:10.1186/s40168-017-0393-0.

17. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, et al. Learning representations of microbe–metabolite interactions. Nature Methods. 2019;16(12):1306–1314. doi:10.1038/s41592-019-0616-3.

18. Quinn TP, Erb I. Examining microbe–metabolite correlations by linear methods. Nature Methods. 2021;18(1):37–39. doi:10.1038/s41592-020-01006-1.

19. Morton JT, McDonald D, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, et al. Reply to: Examining microbe–metabolite correlations by linear methods. Nature Methods. 2021;18(1):40–41. doi:10.1038/s41592-020-01007-0.

20. Thiergart T, Zgadzaj R, Bozsóki Z, Garrido-Oter R, Radutoiu S, Schulze-Lefert P. *Lotus japonicus* Symbiosis Genes Impact Microbial Interactions between Symbionts and Multikingdom Commensal Communities. mBio. 2019;10(5):e01833–19. doi:10.1128/mBio.01833-19.

21. Yang Y, Misra BB, Liang L, Bi D, Weng W, Wu W, et al. Integrated microbiome and metabolome analysis reveals a novel interplay between commensal bacteria and metabolites in colorectal cancer. Theranostics. 2019;9(14):4101–4114. doi:10.7150/thno.35186.

22. Stagaman K, Cepon-Robins TJ, Liebert MA, Gildner TE, Urlacher SS, Madimenos FC, et al. Market Integration Predicts Human Gut Microbiome Attributes across a Gradient of Economic Development. mSystems. 2018;3(1):e00122–17. doi:10.1128/mSystems.00122-17.

23. Huang W, Sun D, Chen L, An Y. Integrative analysis of the microbiome and metabolome in understanding the causes of sugarcane bitterness. Scientific Reports. 2021;11(1):6024. doi:10.1038/s41598-021-85433-w.

24. Monteleone AM, Troisi J, Fasano A, Dalle Grave R, Marciello F, Serena G, et al. Multi-omics data integration in anorexia nervosa patients before and after weight regain: A microbiome-metabolomics investigation. Clinical Nutrition. 2021;40(3):1137–1146. doi:10.1016/j.clnu.2020.07.021.

25. Vorholt JA, Vogel C, Carlström CI, Müller DB. Establishing Causality: Opportunities of Synthetic Communities for Plant Microbiome Research. Cell Host & Microbe. 2017;22(2):142–155. doi:10.1016/j.chom.2017.07.004.

26. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010;11(3):R25. doi:10.1186/gb-2010-11-3-r25.

27. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106.

28. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11(1):94. doi:10.1186/1471-2105-11-94.

29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.

30. Ma S, Ren B, Mallick H, Moon YS, Schwager E, Maharjan S, et al. A statistical model for describing and simulating microbial community profiles. PLOS Computational Biology. 2021;17(9):e1008913. doi:10.1371/journal.pcbi.1008913.

31. Byrd AL, Deming C, Cassidy SKB, Harrison OJ, Ng WI, Conlan S, et al. *Staphylococcus aureus* and *Staphylococcus epidermidis* strain diversity underlying pediatric atopic dermatitis. Science Translational Medicine. 2017;9(397):eaal4651. doi:10.1126/scitranslmed.aal4651.

32. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. Establishing microbial composition measurement standards with reference frames. Nature Communications. 2019;10(1):2719. doi:10.1038/s41467-019-10656-5.

33. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org.

34. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots; 2022. Available from: https://CRAN.R-project.org/package=ggpubr.

35. Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, et al.. GGally: Extension to 'ggplot2'; 2021. Available from: https://CRAN.R-project.org/package=GGally.

36. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research; 2022. Available from: https://CRAN.R-project.org/package=psych.

37. Kurtz Z, Mueller C, Miraldi E, Bonneau R. SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference; 2022.

38. Wolkerstorfer A, De Waard Van Der Spek FB, Glazenburg EJ, Mulder AP, Oranje AP. Scoring the Severity of Atopic Dermatitis: Three Item Severity Score as a Rough System for Daily Practice and as a Pre-screening Tool for Studies. Acta Dermato-Venereologica. 1999;79(5):356–359. doi:10.1080/000155599750010256.

39. Efron B. Better Bootstrap Confidence Intervals. Journal of the American Statistical Association. 1987;82(397):171–185. doi:10.1080/01621459.1987.10478410.

40. Darabi K, Hostetler SG, Bechtel MA, Zirwas M. The role of Malassezia in atopic dermatitis affecting the head and neck of adults. Journal of the American Academy of Dermatology. 2009;60(1):125–136. doi:10.1016/j.jaad.2008.07.058.

41. Glatz M, Bosshard P, Hoetzenecker W, Schmid-Grendelmeier P. The Role of Malassezia spp. in Atopic Dermatitis. Journal of Clinical Medicine. 2015;4(6):1217–1228. doi:10.3390/jcm4061217.

42. Bjerre RD, Bandier J, Skov L, Engstrand L, Johansen JD. The role of the skin microbiome in atopic dermatitis: a systematic review. British Journal of Dermatology. 2017;177(5):1272–1278. doi:10.1111/bjd.15390.

43. Koh LF, Ong RY, Common JE. Skin microbiome of atopic dermatitis. Allergology International. 2022;71(1):31–39. doi:10.1016/j.alit.2021.11.001.

44. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47. doi:10.1093/nar/gkv007.