

# Shrinkage improves estimation of microbial associations under different normalization methods

Michelle Badri<sup>1</sup>, Zachary D. Kurtz<sup>2</sup>, Richard Bonneau<sup>1,3,4,\*</sup> and Christian L. Müller<sup>5,6,7,\*</sup>

<sup>1</sup>Department of Biology, New York University, New York, NY 10012, USA, <sup>2</sup>Lodo Therapeutics, New York, NY 10016, USA, <sup>3</sup>Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY 10010, USA, <sup>4</sup>Computer Science Department, Courant Institute, New York, NY 10012, USA, <sup>5</sup>Center for Computational Mathematics, Flatiron Institute, Simons Foundation, New York, NY 10010, USA, <sup>6</sup>Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg 85764, Germany and <sup>7</sup>Department of Statistics, Ludwig-Maximilians-Universität München, Munich 80539, Germany

Received March 03, 2020; Revised October 27, 2020; Editorial Decision November 03, 2020; Accepted November 10, 2020

## ABSTRACT

**Estimation of statistical associations in microbial genomic survey count data is fundamental to microbiome research. Experimental limitations, including count compositionality, low sample sizes and technical variability, obstruct standard application of association measures and require data normalization prior to statistical estimation. Here, we investigate the interplay between data normalization, microbial association estimation and available sample size by leveraging the large-scale American Gut Project (AGP) survey data. We analyze the statistical properties of two prominent linear association estimators, correlation and proportionality, under different sample scenarios and data normalization schemes, including RNA-seq analysis workflows and log-ratio transformations. We show that shrinkage estimation, a standard statistical regularization technique, can universally improve the quality of taxon–taxon association estimates for microbiome data. We find that large-scale association patterns in the AGP data can be grouped into five normalization-dependent classes. Using microbial association network construction and clustering as downstream data analysis examples, we show that variance-stabilizing and log-ratio approaches enable the most taxonomically and structurally coherent estimates. Taken together, the findings from our reproducible analysis workflow have important implications for microbiome studies in multiple stages of analysis, particularly when only small sample sizes are available.**

## INTRODUCTION

Recent advances in microbial amplicon and metagenomic sequencing as well as large-scale data collection efforts provide samples across different microbial habitats that are amenable to quantitative analysis. Following the organization of sequence data into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs), via pipelines such as qime (1), mothur (2) or dada2 (3), the resulting count data are then available in tabular format for statistical analysis. Downstream analysis tasks include assessing community diversity (4), differential abundance analysis, associating bacterial compositions to system-specific ecological and biomedical covariates, and learning microbe–microbe associations.

However, technical artifacts inherent in microbial abundance data preclude the application of such analysis tasks directly on the measured counts. The data typically comprise a high proportion of zeros and carry only relative information about species abundance. The total number of read counts for any given observation is limited by the total amount of sequencing, quality of DNA preparations and other technical factors and does not represent the community abundance or total species abundance in the sample or ecosystem. For example, unequal amplicon library sizes can bias sequencing reads to OTUs from the larger sample, regardless of true abundance profiles. Although some recent studies have used controlled communities, spike-in controls and other innovations to obtain total community size (5–8), in the majority of experimental designs, the community size is unknown, and, thus, our data are best thought of as containing relative or compositional information (each OTU fraction of total counts, total community size unknown) (9,10). Additionally, technical variation due to sequencing such as differences in amplification biases and batch effects due to multiple sequencing runs can hamper proper quantification of microbial compositions (11).

\*To whom correspondence should be addressed. Tel: +49 89 2180 3466; Fax: +49 89 2180 5044; Email: cmueller@flatironinstitute.org  
Correspondence may also be addressed to Richard Bonneau. Tel: +1 646 603 3714; Fax: +1 646 603 3751; Email: rb133@nyu.edu

To ameliorate these biases, general data normalization methods have been proposed to correct for sampling bias, library size and technical variability, including workflows from RNA-seq pre-processing and compositional data analysis (12–15). Dedicated normalization and modeling strategies are also available for specific analysis tasks, most prominently, for differential abundance testing (16–19).

Here, we examine data normalization schemes in connection with a fundamental multivariate statistical estimation task: inferring pairwise linear associations from microbial count data. Two common strategies that have been adopted for microbial relative abundance data are correlation after data normalization (20,21) and proportionality (22,23) as association measure for compositional data. Under the assumption that ecological association can be captured by empirical Pearson’s correlation or proportionality, accurate correlation or proportionality estimation is of paramount importance for a host of downstream analysis tasks, including state-of-the-art diversity estimation that takes the connectivity of the community into account (4), direct microbial association network inference (14), discriminant analysis and microbial community clustering.

While previous work (20,24) has assessed the precision of correlation detection strategies on synthetic microbial sequencing count data, we took a different approach and investigated the behavior of linear association estimation on the largest-to-date citizen-science sample collection, the American Gut Project (AGP) (25). The large available sample size  $n > 9000$  allows us, for the first time, to critically measure the empirical consistency of combinations of data normalization and association estimation techniques. More specifically, given the lack of ‘gold standard’ microbial associations in gut microbial communities, we asked the question whether and how association patterns inferred from small but realistic sample sizes of tens to a few hundreds of samples resemble those inferred using the entire AGP dataset. This type of ‘sample size consistency’ evaluation was at the heart of the present study.

Using a comprehensive set of evaluation criteria and summary statistics, we first show that, independent of any specific data normalization scheme, standard linear association measures are unreliable in the small sample regime. We propose the application of shrinkage estimation (26) as an effective strategy for sample size consistent association estimation. We show that a popular correlation shrinkage approach from functional genomics (27) enjoys excellent performance in the microbiome context. For proportionality, we introduce a novel shrinkage estimator,  $\rho^*$  (rhoshrink), and assess its statistical behavior. In particular, we quantify the effects of sample size on data normalization and association estimates on several downstream analysis tasks, including microbial association (or relevance) network inference and clustering. Figure 1 shows the proposed analysis framework used in this study.

Our analysis revealed that all normalization-dependent association estimates in the AGP data can be broadly grouped into five categories and that variance-stabilizing and log-ratio approaches provide the most consistent estimation in terms of taxonomic and community structure coherence. Our findings, available in a fully reproducible statistical analysis R workflow at Synapse ID: syn21654780,

have important implications for microbiome studies in multiple stages of analysis, most prominently in the presence of small sample sizes. In particular, we believe that our developed shrinkage estimation framework will improve the reproducibility of future microbiome data analysis studies at almost no additional computational cost.

## MATERIALS AND METHODS

To examine the interplay of data normalization and association estimation methods, we first describe the four essential ingredients of our analysis: the processed AGP 16S rRNA dataset, the comprehensive list of data normalization methods, statistical estimation of linear associations and downstream statistical evaluation and analysis tools.

### AGP sample collection

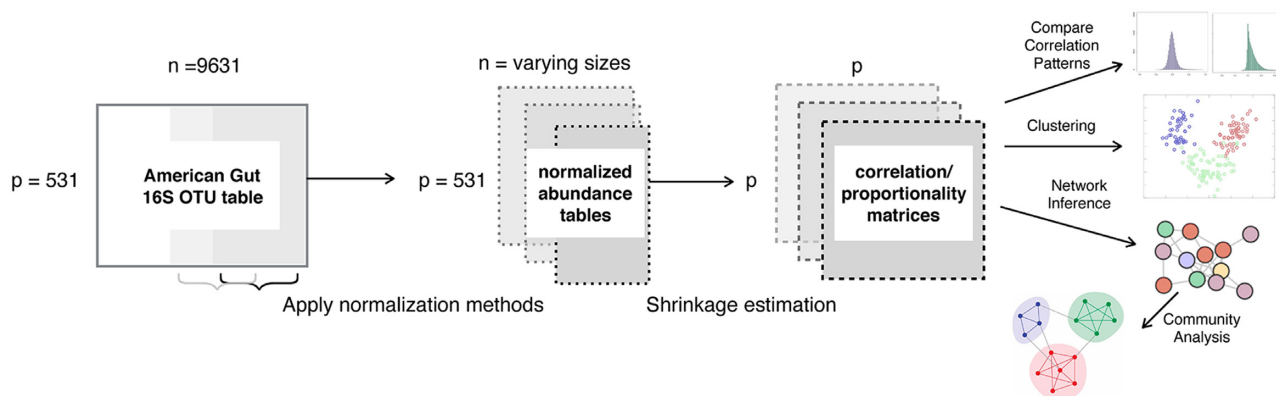
The AGP is a large public repository of human microbiome samples aiming to survey the diversity of microbiota.

Information including diet, disease status and lifestyle variables was measured for public health analysis, but the sheer size of the dataset makes it also a powerful profile of the microbial communities in the human gut. For each batch of samples, the V4 region of the 16S rRNA gene was amplified and sequenced using Illumina next-generation sequencing. The untrimmed data were then processed using sequence variant calling. We obtained OTU count tables and mapping files for unrarefied AGP samples (25) from the project website <ftp://ftp.microbio.me/AmericanGut/ag-2017-12-04/>. The original OTU table contained  $P = 35\,511$  OTU observations for  $n = 15\,148$  samples. We filtered the dataset to contain only fecal samples whose sequencing depths fall above the 10th percentile and removed OTUs that were present in  $<30\%$  of all samples. This resulted in a data matrix comprised of  $P = 531$  OTUs and  $n = 9631$  samples.

To investigate the sample size dependence of data normalization and association estimation on this dataset, we generated collections of random subsamples of varying sample sizes, ranging from  $25 \leq n \leq 9000$ , with 5 (random) replicates per sample size. While sample sizes of  $n \leq 25$  are not uncommon in clinical studies of the microbiome,  $n = 25$  was the lower bound for study inclusion in a recent meta-analysis (28). The large sample limit of  $n = 9000$  was chosen to ensure that the set of samples across the random subsamples is still relatively distinct, which also serves as our large sample reference for association estimates. To simulate reference data under null correlation or proportionality, we randomly shuffled OTU count data across samples prior to normalization.

### Normalization methods

All normalization methods require as input OTU counts, collected over  $n$  samples and stored in a matrix  $W \in \mathbb{N}_0^{n \times p}$ . Each row is a  $p$ -dimensional vector  $w^{(j)} = [w_1^{(j)}, w_2^{(j)}, \dots, w_p^{(j)}]$ , where  $j = 1, \dots, n$  is the sample index,  $w_i^{(j)}$  is the read count of OTU  $i$  in sample  $j$  and  $\mathbb{N}_0$  is the set of natural numbers  $\{0, 1, 2, \dots\}$ . Let the total OTU count



**Figure 1.** Framework for examining the effects of normalization methods on linear association estimation with increasing sample size. Comparative summary statistics of the resulting association matrices include distribution-based analysis, distance-based matrix comparison, hierarchical clustering and association network analysis.

for sample  $j$  be  $m^{(j)} = \sum_{i=1}^p w_i^{(j)}$ . Several methods require the application of a log transformation, thus requiring non-negative input data. We include a pseudocount of 1 to all OTU input data if zero counts are not explicitly handled by the respective normalization scheme. We consider the following data normalization or transformation schemes.

*Total sum scaling.* A standard approach for normalizing count data is to divide individual counts by the total OTU counts in a sample, thus scaling the count vector such that the total sum is fixed to 1. This normalization is known as total sum scaling (tss) or total sum normalization. It reads

$$\text{tss}(w^{(j)}) = \left[ \frac{w_1^{(j)}}{m^{(j)}}, \frac{w_2^{(j)}}{m^{(j)}}, \dots, \frac{w_p^{(j)}}{m^{(j)}} \right] \in \mathbb{S}^p.$$

The resulting sample space of the data is thus the  $(p - 1)$ -dimensional simplex.

*Cumulative sum scaling.* The tss approach may place unwanted influence on OTUs that are highly sampled due to sequencing biases by over-representing it in the scaling factor  $m^{(j)}$  (11). To reduce the influence of these highly abundant OTUs for sparse data, cumulative sum scaling (css) has been proposed in (15) and implemented in the metagenome-Seq R package. Rather than normalizing by the total sum, css selects a scaling factor that is a fixed quantile of OTU counts. Formally,

$$\text{css}(w^{(j)}) = \left[ \frac{w_1^{(j)}}{m_i^{(j)}} N, \frac{w_2^{(j)}}{m_i^{(j)}} N, \dots, \frac{w_p^{(j)}}{m_i^{(j)}} N \right] \in \mathbb{R}^{p \geq 0},$$

where the scaling factor for sample  $j$  is  $m_i^{(j)} = \sum_{i|w_{ij} \leq q_l^{(j)}} w_i^{(j)}$ . The quantity  $q_l^{(j)}$  is the sum of read counts up to and including the  $l$ th quantile.  $N$  is a pre-specified constant, e.g.  $N \triangleq 1000$ , chosen such that the resulting data vector resembles the units of the original counts. The sample space of css-transformed data is that of non-negative real numbers  $\mathbb{R}_{\geq 0}$ .

Let  $\hat{l}$  be the index of the  $q_l^{(j)}$ , the  $l$ th quantile for sample  $j$ ,  $\bar{q}_l = \text{med}_j \{q_l^{(j)}\}$  the median  $l$ th quantile across all samples and  $\mu_l^{(j)}$  be the mean  $l$ th quantile. css requires the median absolute deviation of sample quantiles to be empirically stable via the quantity  $\delta_l = \text{med}_j | \mu_l^{(j)} - \bar{q}_l |$ . A common choice is to set  $\hat{l} := \min\{\delta_{l+1} - \delta_l \geq 0.1\delta_l : 1 \leq l < n\}$  (15). The scaling factor is then defined by summing all the counts up to the smallest value of  $l$  that is stable, on average, across all samples that is greater than or equal to the median. The default choice for the median is the  $l$ th quantile.

*Common sum scaling.* Common sum scaling (com), as introduced in (11), is an alternative to rarefying OTU counts. Counts are scaled to the minimum depth of each sample via

$$\text{com}(w^{(j)}) = \left[ \left\lfloor w_1^{(j)} \frac{m^{(\min)}}{m^{(j)}} \right\rfloor, \dots, \left\lfloor w_p^{(j)} \frac{m^{(\min)}}{m^{(j)}} \right\rfloor \right] \in \mathbb{R}^p,$$

where  $m^{(\min)} = \inf\{m^{(1)}, m^{(2)}, \dots, m^{(n)}\}$ . The operator  $\lfloor \cdot \rfloor$  (floor) converts a real number to the greatest integer that is less than or equal to the input.

*Relative log expression.* The relative log expression (rle) is introduced for gene expression data and available in the DESeq/edgeR package (13). The rle method is defined as follows. Let  $g(x) = (\prod_{i=1}^m x_i)^{1/m}$  be the geometric mean of an  $m$ -dimensional vector  $x$ , and let  $w_i = W_i^T = [w_i^{(1)}, \dots, w_i^{(n)}]$  be the vector of counts of OTU  $i$  over  $n$  samples (a transposed column vector of count matrix  $W$ ). The numeric scaling factor for sample  $j$  is

$$s^{(j)} = \text{med}^{(j)} \left[ \frac{w_1^{(j)}}{g(w_1)}, \dots, \frac{w_p^{(j)}}{g(w_p)} \right],$$

$$\bar{s}^{(j)} = \frac{s^{(j)}}{g(s^{(\cdot)})},$$

where  $\text{med}[x]$  denotes the median of vector  $x$  and  $s^{(\cdot)} = [s^{(1)}, \dots, s^{(n)}]$  is a collection of the sample scaling factors. Let the



global scaling factor  $\hat{s}_c = \frac{1}{n} \sum_{j=1}^n \bar{s}^{(j)}$  be the arithmetic mean of all normalized scaling factors. The rle is then defined as

$$\text{rle}(w^{(j)}) = \left[ \hat{s}_c \frac{w_1^{(j)}}{\bar{s}^{(j)}}, \dots, \hat{s}_c \frac{w_p^{(j)}}{\bar{s}^{(j)}} \right] \in \mathbb{R}^{p \geq 0}.$$

In summary, the rle estimates a median library from the geometric mean over all samples. The median ratio of each sample to the median library is then taken as the scale factor.

*Inverse hyperbolic sine.* A standard variance-reducing transformation, often applied to flow and mass cytometry data (29,30), is the inverse hyperbolic sine function, defined as

$$\text{asinh}(w^{(j)}) = \log \left( w^{(j)} + \sqrt{(w^{(j)})^2 + 1} \right),$$

applied element-wise over the sample vector. The resulting data matrix is then mean centered prior to association estimation.

*Wrench.* The Wrench procedure, introduced in (17), estimates compositional correction factors in the presence of zero inflation. Wrench is defined as

$$\text{wren}(w^{(j)}) = \left[ \frac{w_1^{(j)}}{m^{(j)}\eta_j^{-1}}, \dots, \frac{w_p^{(j)}}{m^{(j)}\eta_j^{-1}} \right] \in \mathbb{R}^p.$$

The quantity  $\eta_j^{-1}$  represents a compositional scale factor where  $\eta_j^{-1} = \frac{1}{p} \sum_i e_i^{(j)} \frac{y_{ji}}{y_{+i}}$ . Here,  $y_{ji}$  is the proportion of each feature  $i$  in sample  $j$  and  $y_{+i}$  is the average proportion of each feature  $i$  across all samples. The weight  $e_i^{(j)}$  is estimated using the ‘ $W_2$ ’ scheme, the default choice in the Wrench R package [see also (17) for further details]. While Wrench is capable of incorporating information about sample grouping, e.g. for differential abundance testing, we consider all samples to be in a single group.

*Variance-stabilizing transform.* The goal of variance-stabilizing transformations (vst) is to factor out the dependence of the variance on the mean (overdispersion) (13). Consider the mean-dispersion relation  $v(\mu) = \frac{1}{n-1} \sum_{j=1}^n \left( \frac{w_i^{(j)}}{\bar{s}^{(j)}} - \hat{\mu}_i^{(j)} \right)^2$ . Here, the ‘size factors’ are  $\hat{s}^{(j)} = \text{med} \left( w_i^{(j)} / g(w^{(j)}) \right)^{1/n}$  and  $\hat{\mu}_i^{(j)} = \frac{1}{n} \sum_{j=1}^n w_i^{(j)} / \hat{s}^j$  is the average count to size factor ratio of sample  $j$ . The vst is then the integral quantity defined as

$$\text{vst}(w^{(j)}) = \int_0^{w^{(j)}} \frac{d\mu}{\sqrt{v(\mu)}} \in \mathbb{R}^p.$$

The function  $v(\mu)$  is approximated by a spline function and evaluated for each count value in the column. The vst normalization is available in the DESeq package where the numerical fitting is achieved using local regression on the graph  $(\hat{\mu}_i^{(j)}, v(\mu))$ . A smooth function  $v(\mu)$  is estimated using an estimate of raw variance:  $\hat{v}(\mu) = v(\hat{\mu}_i^{(j)}) - z_i$ , where

$z_i = \frac{\mu_i^{(j)}}{n} \sum_{j=1}^n \frac{1}{\hat{s}^j}$ . The local regression is parameterized such that large counts are scaled to be asymptotically equal to the logarithm base 2 of normalized counts. When we examined the per-OTU standard deviation [taken across all ( $P = 531$ ) OTUs] plotted against the rank of the average OTU count, it can be seen that vst produces similar counts to both clr and a logarithm base 2 transform (Supplementary Figure S1).

*Centered log-ratio transformation.* Log-ratio transformations, introduced in (9), transform positive compositional data from the simplex to Euclidean space (9,14). The centered log-ratio (clr) transform is defined as

$$\text{clr}(w^{(j)}) = \left[ \log \frac{w_1^{(j)}}{g(w^{(j)})}, \dots, \log \frac{w_p^{(j)}}{g(w^{(j)})} \right] \in \mathbb{R}^p,$$

where the ratio is taken with respect to the geometric mean of the composition. The resulting data lie in a  $p - 1$  hyperplane of  $p$ -dimensional Euclidean space.

### Estimation of linear associations

Following a transformation of count data under some function  $f: \mathbb{N}_0^p \rightarrow \mathcal{X}^p$ , we consider several estimation methods for linear associations among the  $p$  OTUs.

*Covariance and correlation estimation.* The standard way of estimating linear associations is the empirical (sample) covariances in the sample space  $\mathcal{X}^p$  that forms the basis for many downstream multivariate data analysis techniques, including principal component analysis (PCA), discriminant analysis, metric learning and network inference.

Formally, column centering the transformed data results in an  $n \times p$  data matrix  $X = f(W) \left( I_p - \frac{1}{n} \mathbf{1}_p \right)$ , where  $I_p$  is the  $p$ -dimensional identity matrix and  $\mathbf{1}_p$  is unit (all-ones) matrix. In matrix notation, the sample covariance matrix (cov) is  $\hat{S} = X^T X \circ \frac{1}{n-1} \mathbf{1}_p$ , where  $\circ$  indicates element-wise multiplication of two equal size matrices.

The estimate  $\hat{S}$  is a symmetric  $p \times p$  matrix with the sample variances along the diagonal and can be normalized to obtain a matrix containing Pearson correlation coefficients. Let  $\hat{D} = \text{diag}[\hat{S}]$  be a diagonal matrix with the  $p$  post-transformed OTU variances on the diagonal and zero elsewhere. The Pearson correlation matrix is then  $\hat{R} = \hat{D}^{-1/2} \hat{S} \hat{D}^{-1/2}$ . The matrix  $\hat{R}$  is a symmetric  $p \times p$  matrix where each entry  $\hat{r}_{ik} = \hat{R}[i, k]$  corresponds to the Pearson correlation between OTUs  $i$  and  $k$  under the data transformation.

The magnitude and sign of the values in  $\hat{R}$  are often interpreted as the association strength and direction, respectively. The sample correlation/covariance matrices are, however, inadmissible in the  $p \gg n$  setting, i.e. when fewer samples than OTUs are available. For example, type I errors may be grossly inflated, since the parameters under estimation are underdetermined. Standard operations for solving systems of linear equations such as PCA are then ill-posed.

*Proportionality estimation.* Covariance and correlation estimation on compositional data has long been criticized due

to the necessary presence of negative bias, scale dependence and subcompositional incoherence in the estimates (9,31). Association measures based on the concept of proportionality have thus been put forward as an alternative to correlation (22). Here, we consider the symmetric proportionality  $\rho_p$  (23,32) that, by default, operates on clr-transformed data  $X^{\text{clr}} \triangleq \text{clr}(W)$ . The measure is defined as

$$\rho_p(X_i^{\text{clr}}, X_k^{\text{clr}}) = 1 - \frac{\text{var}(X_i^{\text{clr}} - X_k^{\text{clr}})}{\text{var}(X_i^{\text{clr}}) + \text{var}(X_k^{\text{clr}})}, \quad (1)$$

where  $X_i^{\text{clr}}$  and  $X_k^{\text{clr}}$  are the columns of the matrix corresponding to OTUs  $i$  and  $k$ , respectively. The quantity  $\rho_p$  is a proportionality measure because differences of clr-transformed components are equivalent to log ratios of compositions. When  $p \gg n$ , the sample estimator for  $\rho_p$  faces similar challenges as sample correlation/covariance estimation.

### Shrinkage estimation of linear associations

One way to improve sample estimation of linear associations in the high-dimensional ( $p > n$ ) setting is via regularization. For instance, covariance/correlation estimators with stronger statistical properties can be derived in the  $p > n$  setting when imposing structural assumptions about the underlying population covariance. One ubiquitous structural assumption is sparsity where only a few strong pairwise correlations are assumed to be present in the data. An effective data-driven approach to realizing structural sparsity is shrinkage estimation. We next revisit a popular covariance/correlation shrinkage approach and introduce a novel shrinkage estimator for proportionality.

*Shrinkage covariance estimation.* While several regularized covariance estimators are available in the literature (33–35), we focus here on Schäfer–Strimmer shrinkage estimation (27). The principal idea of shrinkage estimation is to shrink small sample correlations toward entries of a prescribed target matrix. The standard target matrix is the  $p \times p$ -dimensional identity matrix. Shrinkage intensities are simultaneously estimated from data (33). In Schäfer–Strimmer shrinkage, as implemented in the R package `corpcor`, individual entries  $s_{ik}^*$  of the shrinkage covariance  $S^*$  and entries  $r_{ik}^*$  in the shrinkage correlation  $R^*$  are estimated as follows. For all off-diagonal elements in  $S^*$ , we compute

$$\hat{s}_{ik}^* = \hat{r}_{ik}^* \sqrt{\hat{s}_{ii} \hat{s}_{kk}} \quad \forall i \neq j,$$

where the shrunk correlation estimates are  $\hat{r}_{ik}^* = (1 - \hat{\lambda}_1^*) \hat{r}_{ik}$ . The variance (var) estimates  $\hat{s}_{ii}^*$  are shrunk in a separate procedure toward the median  $v = \text{med}[\hat{s}_{11}, \dots, \hat{s}_{pp}]$  via  $\hat{s}_{ii}^* = \hat{\lambda}_2^* v + (1 - \hat{\lambda}_2^*) \hat{s}_{ii}$ .

The shrinkage intensities  $\hat{\lambda}_1^*$  and  $\hat{\lambda}_2^*$  are determined empirically by estimating the variance within the sample covariance matrix (see Supplementary Methods).

*Shrinkage proportionality estimation.* To derive a shrinkage estimator for proportionality, we first consider an equiv-

alent formulation of (1) in terms of covariances and variances (23,32). The reformulation reads

$$\rho_p(X_i^{\text{clr}}, X_k^{\text{clr}}) = \frac{2 \times \text{cov}(X_i^{\text{clr}}, X_k^{\text{clr}})}{\text{var}(X_i^{\text{clr}}) + \text{var}(X_k^{\text{clr}})},$$

and the corresponding sample estimator is thus

$$\hat{\rho}_p(X_i^{\text{clr}}, X_k^{\text{clr}}) = \frac{2\hat{s}_{ik}}{\hat{s}_{ii} + \hat{s}_{kk}},$$

where  $\hat{s}_{ik}$  are elements of the covariance estimates  $\hat{S} = (X^{\text{clr}})^T X^{\text{clr}} \circ \frac{1}{n-1} \mathbf{1}_p$  on clr-transformed data. This formulation clarifies the link between the standard correlation matrix on clr-transformed data and  $\rho_p$ : the former uses the geometric mean of  $\hat{s}_{ii}$  and  $\hat{s}_{kk}$  in the denominator, whereas the latter uses the arithmetic mean. Since  $\hat{\rho}_p$  is completely determined by sample covariances and variances, we expect the measure to have the same drawbacks in the small sample setting as the sample covariance estimators. We thus propose the following shrinkage proportionality estimator  $\rho_*$  (rhoshrink) as

$$\rho_*(X_i^{\text{clr}}, X_k^{\text{clr}}) \triangleq \frac{2s_{ik}^*}{s_{ii}^* + s_{kk}^*}, \quad (2)$$

where  $s_{ik}^*$  are the elements of the Schäfer–Strimmer shrinkage covariance  $S^*$ . The estimator for  $\rho_*$  is thus completely determined by the Schäfer–Strimmer covariance shrinkage estimates, outlined in the previous paragraph.

### Comparing association patterns

The distribution of association patterns was visualized using density plots. For each method, we examine the distribution of values after the appropriate association metric is applied to a single normalized subsample. In each density distribution, we also calculate statistical moments: mean, variance, skewness and kurtosis. To quantify similarities of the estimated association patterns across different data normalization methods, association measures and sample sizes, we considered three different distance measures. These distances are then used for comparative low-dimensional embeddings of the different estimates as well as for measuring convergence of the estimators with sample size.

*Frobenius distance.* Given a pair of  $p \times p$ -dimensional association matrices  $\hat{R}$  and  $\hat{R}'$ , the Frobenius distance measures the sum of squared differences between the corresponding entries and is defined as

$$d_f(\hat{R}, \hat{R}') = \sqrt{\sum_k \sum_i (\hat{R}_{ik} - \hat{R}'_{ik})^2} \equiv \|\hat{R}_{ik} - \hat{R}'_{ik}\|_F.$$

*Spectral distance.* Given a square, symmetric matrix  $A$ , let  $A = U\Sigma U^T$  be its singular value decomposition, where  $\Sigma$  is a diagonal matrix with singular values along the diagonal entries, i.e.  $\Sigma_{ii} = \sigma_i$ . Let  $\sigma_{\max}(A)$  be the largest singular value of  $A$ . The spectral distance is

$$d_s(\hat{R}, \hat{R}') = \sqrt{\sigma_{\max}(\hat{R} - \hat{R}')} \equiv \|\hat{R}_{ik} - \hat{R}'_{ik}\|_2.$$

Due to its sole dependence on the ‘spectrum’ (the singular values) of the association matrix, the spectral distance is

invariant to unitary transformations (e.g. rotations) of the matrices.

**Correlation matrix distance.** The correlation matrix distance (CMD) (36) measures the orthogonality of two correlation matrices and is defined as

$$d_{\text{cmd}}(\hat{R}, \hat{R}') = 1 - \frac{(\hat{R}\hat{R}')}{\|\hat{R}\|_F \|\hat{R}'\|_F} \in [0, 1],$$

where the trace operator  $(A) = \sum_i A_{ii}$  is the sum of the diagonal entries of a square matrix.

### Downstream analysis

We considered two downstream exploratory data analysis tasks that require the estimation of microbial associations as input: (i) OTU clustering and (ii) microbial association network construction and community analysis.

**Clustering.** Unsupervised clustering of OTUs can help identify microbial subcommunities that may jointly affect host phenotype or reveal experimental and batch effects (25). We considered two popular clustering techniques: spectral and hierarchical clustering.

Spectral clustering requires the construction of an ‘affinity matrix’ from estimated associations, or a matrix expressing how similar pairwise entries are to each other. Here, to construct the affinity matrix we transformed associations into dissimilarity scores  $A_s = 1 - \sqrt{\frac{1-\hat{R}}{2}}$  and constructed a  $k$ -nearest neighbor graph ( $k = 2$ ) to obtain a sparse and symmetric affinity matrix  $A$ . Identification of OTU clusters is based on  $k$ -means clustering of the first  $m$  components of the eigendecomposition of the normalized Laplacian  $L = D^{1/2}(D - A)D^{1/2}$ , where  $D$  is the diagonal degree matrix with entries containing the row or column sums of  $A$  (37). We chose the target cluster size to be number of connected components of the associated affinity graph (37). To assess the taxonomic content of a particular clustering, we evaluated the homogeneity of each cluster with respect to the taxonomic families of the underlying OTUs. As a quantitative measure, we computed the ratio of the effective family number (exponential of the Shannon entropy of family counts) to the total number of families detected per cluster.

For hierarchical clustering, we converted association matrices to dissimilarity measures using  $A_h = \sqrt{\frac{1-\hat{R}}{2}}$ . Clustering was then performed using Ward’s method from the `hclust` package in R. Circular dendrograms were cut using the `cuttree` method, where  $k = 10$  was chosen to represent the number of class annotations.

**Relevance networks and community analysis.** Relevance networks (38) are a popular way of visualizing and analyzing the overall structure of the microbial ecosystem. Relevance network construction ranks all pairwise correlation or proportionality values between OTUs by absolute value, selects a certain percentage of highest ranked pairs and visualizes the resulting set of pairs as edges between OTUs in an association network. The ranking of pairwise interactions allows us to compare strong associations regardless of differences in scale.

Multiple studies have found a higher prevalence of positive associations between taxonomically related OTUs in human gut datasets (14,39–42). We thus use taxonomic coherence, measured by assortativity (43), as independent summary statistic for relevance networks. When categorical variables are available for each node, the assortativity coefficient of a network takes values in  $[-1, 1]$  and measures the tendency of adjacent nodes to belong to the same category. In the context of microbial networks, the nodes are OTUs and the associated categories are their inferred taxonomic rank at the genus level.

We also examined the presence of community structure in the relevance networks using the concept of modularity (44). Similar to clustering analysis, modularity analysis of a network enables the partitioning of nodes into tightly connected subcommunities. Modularity was computed using the fast-greedy algorithm, described in (45) and implemented in the `igraph` package in R. Network layout was generated using the force-directed Fruchterman–Reingold algorithm (46).

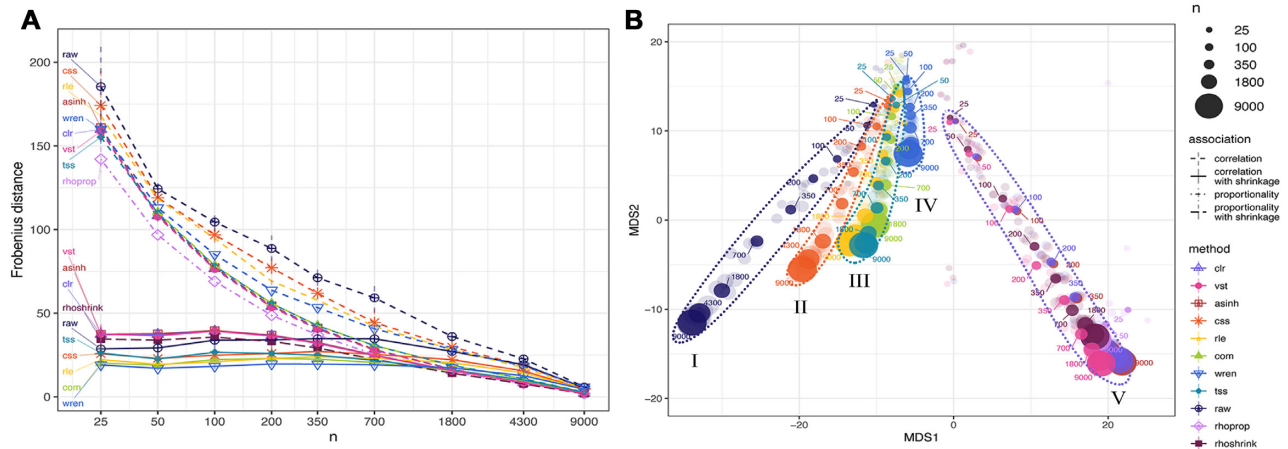
## RESULTS

Our comprehensive computation and analysis workflow produced several key results that are summarized below. We highlighted statistical properties of association estimation, followed by a comparison of downstream analysis results. For ease of presentation, we focused on tss and clr as representative data normalization/transformations as well as standard and shrinkage-based proportionality estimation in the main text. The complete analysis is available in the Supplementary Data.

### Shrinkage universally improves consistency of association estimation

We first analyzed the influence of shrinkage on the estimation of associations under different data normalization and sample sizes. We show the convergence properties of association estimation, as measured by Frobenius distance  $d_f$  with respect to the large sample limit, with increasing sample size in Figure 2A. Shrinkage universally improves estimates in the low sample regime compared to its sample estimation counterparts. Even when the sample size  $n$  exceeds the number of OTUs  $p$ , most shrinkage estimates remain more similar to their respective large sample. This behavior is also reflected in the distribution of association estimates at low ( $n = 50$ ) and large ( $n = 9000$ ) sample sizes, as highlighted in Figure 3 for the proportionality measures  $\rho_p$  (rhoprop) and  $\rho_*$  (rhoshrink). In the small sample limit, rhoprop produces extreme proportionality estimates compared to the large sample limit (third row in Figure 3). The shapes of the distributions of rhoshrink estimates, however, were more similar in the small and large sample limits, and the distribution covered a similar range of  $[-0.1, 0.1]$ . These phenomena were observed for all combinations of data normalization and association estimation (Supplementary Figures S4A and S5A). As expected, the influence of shrinkage vanished in the large sample limit, as reflected in decreasing shrinkage intensities with increasing sample size (Supplementary Figure S2).





**Figure 2.** Frobenius distance between estimates of association. (A) Average Frobenius distance between subsamples of the same sample size. Dashed lines represent the mean distance between normalized matrices after Pearson correlation. The solid lines represent the mean distance between normalized matrices where correlation/proportionality estimation with shrinkage was performed. The dot-dashed line represents rho, a proportionality metric. The long-dashed line represents rhoshrink, proportionality with shrinkage included. Vertical lines represent standard deviation from the mean for each corresponding method. (B) Multidimensional scaling (MDS) representation of Frobenius distance between correlation structures of varying sizes estimated from different normalization methods. The most opaque points represent the mean of five subsamples of the same size [color scheme as in (A)]. Points are labeled based on subsample size.

**Normalization methods induce distinct association patterns**

We next analyzed the similarity among the different association estimates with increasing sample size using MDS. Figure 2B shows a 2D MDS embedding of all shrinkage association estimates using the Frobenius distance. We identified five distinct classes. Association estimates following a variance-reducing/stabilizing transformations (clr, vst, asinh, rhoprop, rhoshrink) form a distinct linear trace in the embedding, ordered along sample size (V). Correlation estimates on raw count data form another distinct group (I). Correlation estimates after css (II) and wren (IV) normalization form two distinct traces in the embedding. Finally, correlation estimates following the com, rle and tss normalization form the fifth class of association patterns (III). For small sample sizes, association patterns are similar independent of the normalization methods. As each of the five classes forms a distinct linear trace in the embedding, we used the distances between estimates of different sample sizes to evaluate the rate at which normalization methods arrived at stable patterns of association. In agreement with Figure 2A, we observed that wren, vst and com arrived at consistent association estimates with the fewest samples, followed closely by rle and tss normalization methods (Figure 2B and Supplementary Figure S3). The observed grouping pattern and convergence behavior are largely invariant to the distance measure used (see Supplementary Figures S4B and S5B for spectral distance and CMD, respectively).

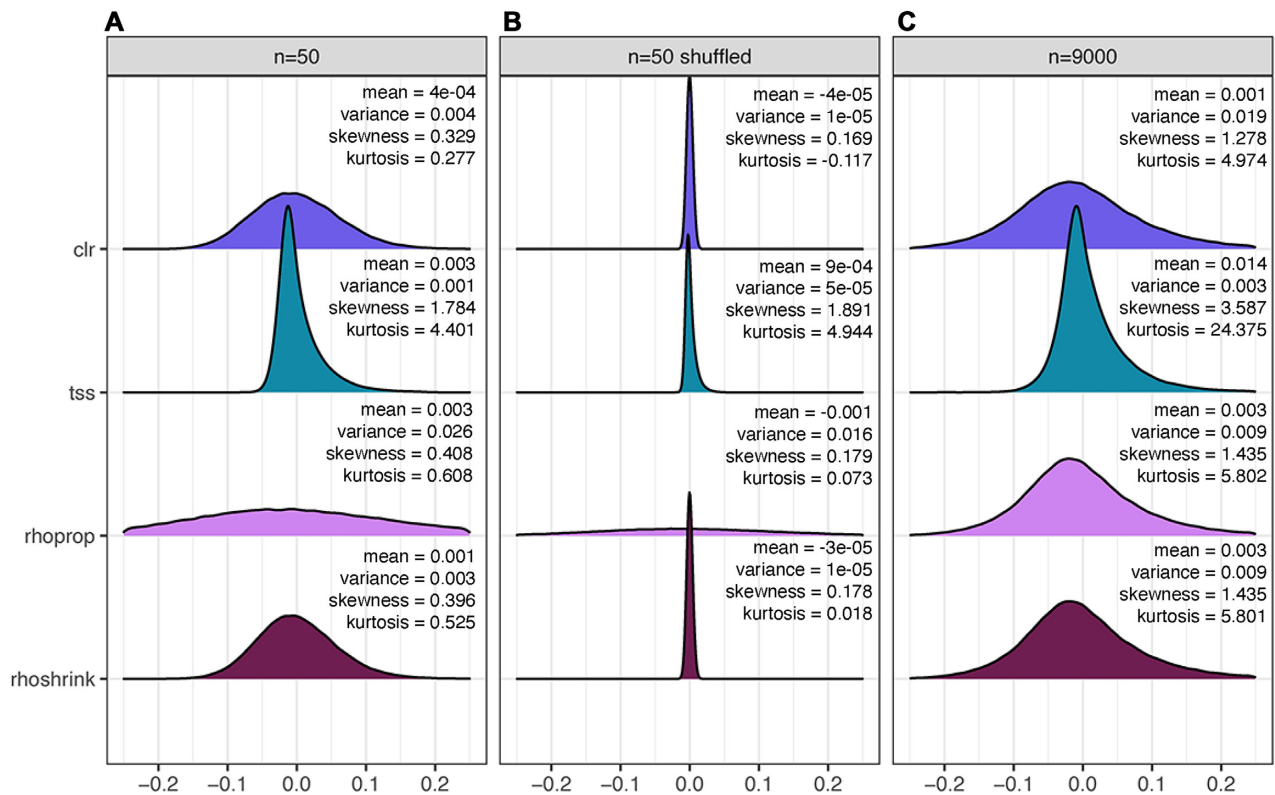
**Association estimates are positively skewed**

We next analyzed the shapes of empirical distributions of shrinkage association estimates for all normalization schemes in three different sample scenarios: small sample regime ( $n = 50$ ), randomly shuffled data in the small sample regime ( $n = 50$ ) and large sample regime ( $n = 9000$ ). Figure 3 shows clr, tss, rhoprop and rhoshrink distributions across

these scenarios (see Supplementary Figure S3 for all others). All correlation distributions are positively skewed. Estimates without shrinkage are considerably wider in the low sample regime (as exemplified for standard proportionality rhoprop versus rhoshrink in Figure 3). Overall, variance-reducing/stabilizing transformations (clr, vst, asinh, rhoprop, rhoshrink) induce wider, more symmetric association distributions. All other normalization schemes induce distributions with considerable positive skewness, resembling correlation distributions on raw count data (Supplementary Figure S3). Positive skewness also persists for association estimates on shuffled data. Although the shapes of shrinkage association distributions are visually similar in the small and large sample limits, we universally observed an increase in skewness and kurtosis with larger sample sizes independent of the normalization scheme.

**Clustering methods are sensitive to normalization and shrinkage estimation**

We next focused on analyzing the influence of normalization and association estimation on downstream data analysis tasks. We first considered clustering of OTUs using a large sample limit of  $n = 9000$  samples from the AGP dataset. For spectral clustering, we asked the question whether and how normalization and shrinkage influence (i) the standard selection of the number of cluster and (ii) the taxonomic composition of the resulting clusters. One common strategy for model selection in spectral clustering is the ‘spectral gap’ criterion. The number of selected clusters is considerably larger ( $k \geq 11$ ) for the variance-reducing/stabilizing transformations (clr, vst, asinh, rhoprop, rhoshrink) than for other normalization methods ( $k \leq 8$ ) (Supplementary Figure S6). Despite the large sample size, the spectral gap of rhoprop- and rhoshrink-based spectral clustering is still different, resulting in  $k = 11$  and  $k = 12$  clusters, respectively. The different number of clus-



**Figure 3.** Density of association values under different transformations and shrinkage. To represent clr and tss, data are normalized and correlation is calculated with shrinkage. Proportionality without shrinkage and proportionality with shrinkage are represented by rhoprop and rhoshrink, respectively. Each plot is a single random subsample of four representative methods at (A) 50 samples, (B) 50 samples with shuffled data and (C) 9000 samples. Mean, variance, skewness and kurtosis are shown for each distribution. Additional methods are provided in Supplementary Figure S3.

ters also contributed to marked differences in terms of the homogeneity of OTU compositions, as shown in Figure 4. Variance-reducing/stabilizing transformations produced taxonomically more homogeneous groups at the family level. rhoshrink-based clustering produced the highest mean cluster purity, indicating strong agreement between estimated OTU associations and taxonomic identity (as shown for the family level in Figure 4 and Supplementary Figure S5). rhoprop- and rhoshrink-based clustering formed very similar but not identical clusters in terms of composition and cluster purity. A larger number of OTUs of family Ruminococcaceae and class Bacteroidia cluster together in clr-based clustering compared to tss-based clustering. OTU clusters derived from css, rle and wren normalization resulted in no distinct taxonomic grouping (see Supplementary Figure S5).

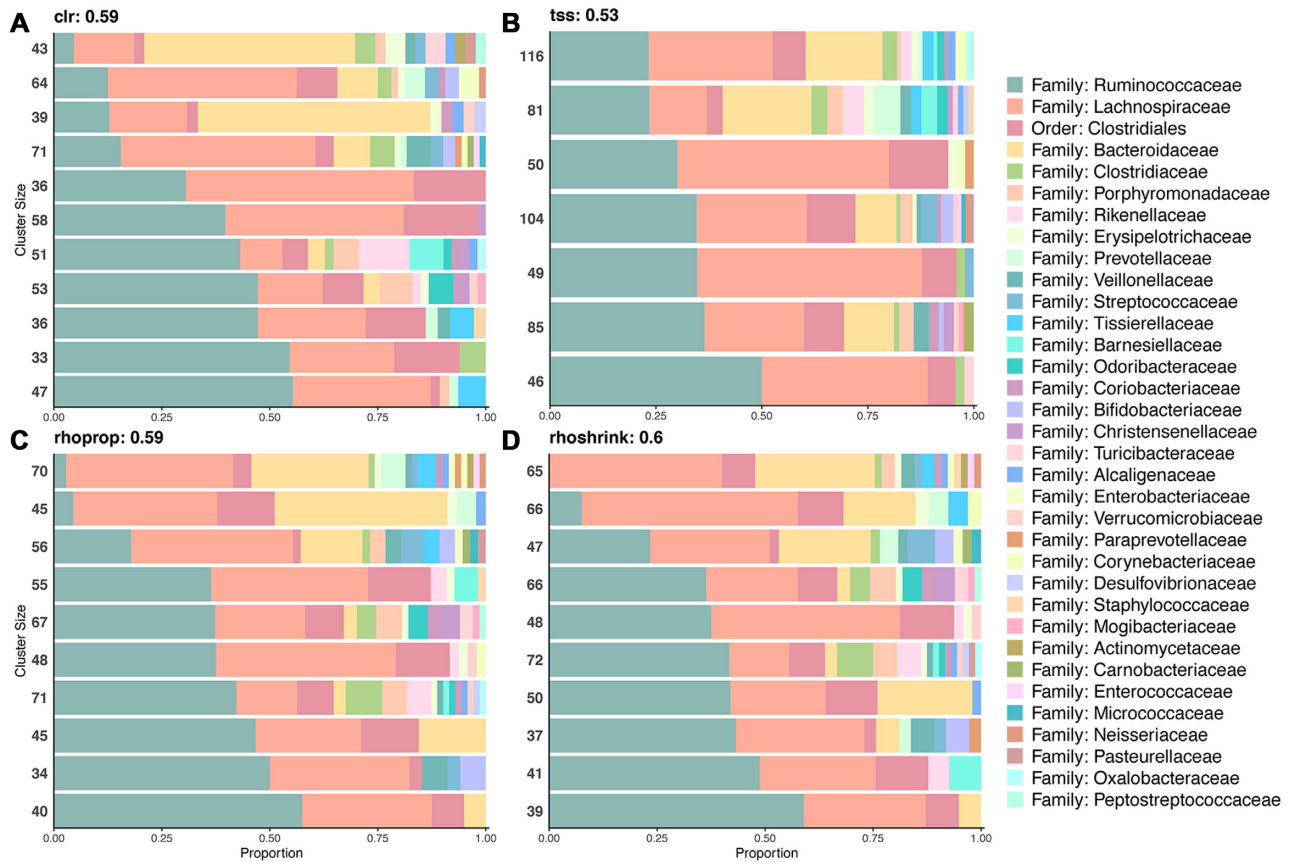
Hierarchical clustering largely confirmed the previous observations. For ease of comparison, we set the number of clusters to  $k = 10$  for inference workflows. Figure 5 shows the dendrograms for clr-, tss-, rhoprop- and rhoshrink-based clustering. While some distinct and homogeneous clusters can be found in the tss case, the majority of OTUs have been grouped into a single cluster comprising many families and classes of taxonomically unrelated bacteria. However, taxonomic grouping is well represented by hierarchical clustering of rho- and rhoshrink-based estimates (Figure 5). Similarly, vst and asinh have recovered large groups of the most prevalent family annotation: Ru-

minococcaceae, Lachnospiraceae and Bacteroidaceae (see Supplementary Figure S8).

### Normalization induces relevance networks with different community structures

We next considered the downstream statistical task of learning microbial relevance networks from AGP data. We estimated associations in the large sample limit  $n = 9000$  and selected the top 2000 associations for network construction in every data normalization/association estimation workflow. Figure 6 shows network visualizations for clr-, tss-, rhoprop- and rhoshrink-based relevance networks (see Supplementary Figure S9 for the other instances). We identified subcommunities of highly connected OTUs using modularity maximization. The number of identified modules ranged between 20 (using Wrench) and 38 (using vst normalization). Relevance networks derived from variance-reducing/stabilizing transformations (clr, vst, asinh, rhoprop and rhoshrink) were partitioned into 35–38 modules and achieved a maximum modularity score of  $\approx 0.8$  (compared to modularity scores of  $< 0.6$  for all other networks). Visual inspection of these networks revealed that members of the Bacteroidetes phylum (represented by square nodes in Figure 6) formed tightly connected modules with few edges connecting to other phyla. Firmicutes (represented by circular nodes) in networks were divided into a higher





**Figure 4.** OTU clusters from spectral clustering. (A–D) Each horizontal bar represents the composition of OTUs in a cluster at the family level. Clusters are in order of increasing percentage of the most abundant family: Ruminococcaceae. In each cluster, the colors represent the OTU families in each cluster. Numbers to the left of each bar represent the number of OTUs in each cluster. Values next to each method name represent cluster purity. Additional methods are provided in Supplementary Figure S7.

number of modules comprising distinct families, including Lachnospiraceae (represented by orange circles) and Ruminococcaceae (teal circles, Figure 6 and Supplementary Figure S9). This striking modularity is less pronounced in the tss-based relevance networks (Figure 6B).

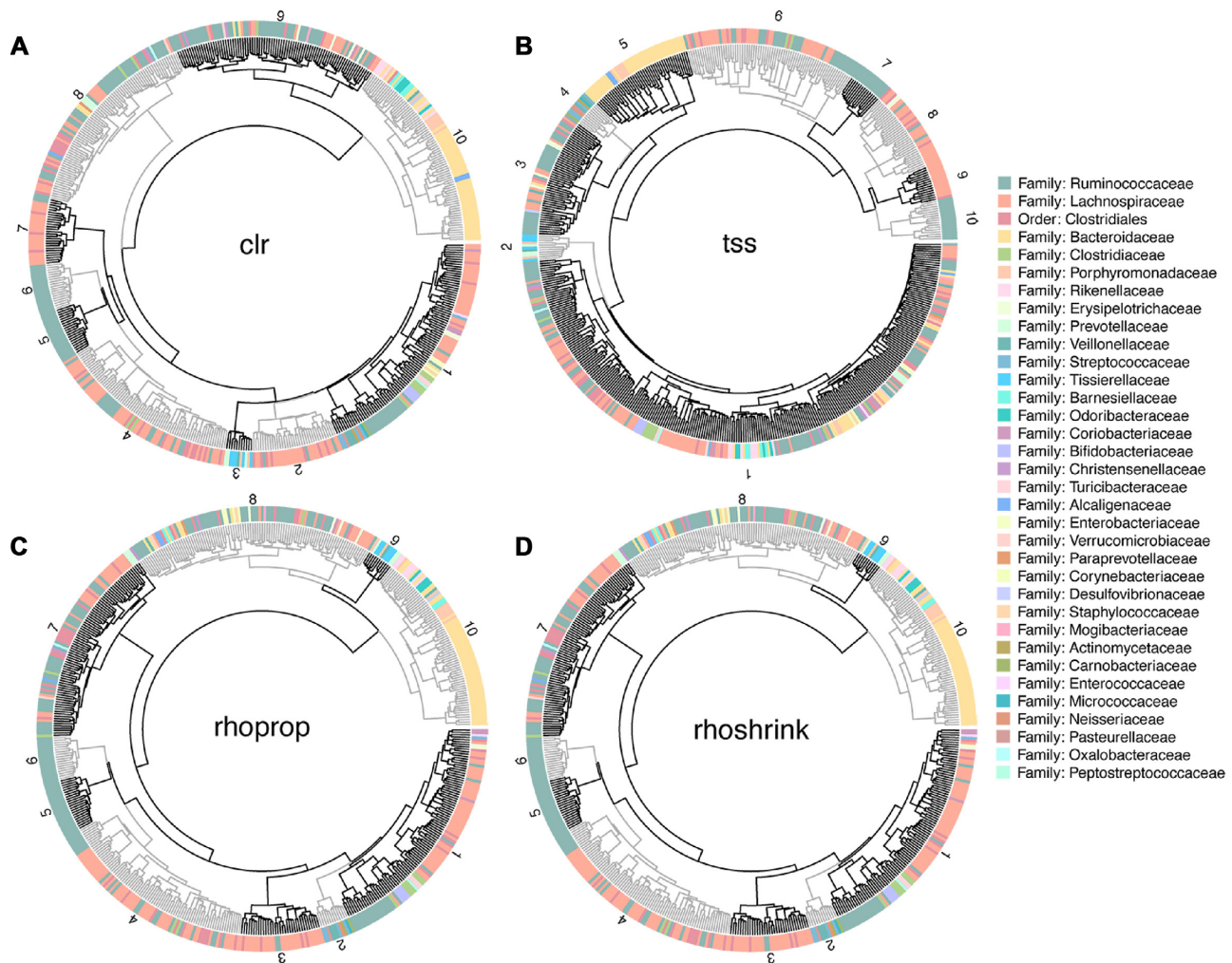
Similar to the clustering analysis, we next evaluated the taxonomic coherence of the different networks. Using assortativity on the genus level as a quantitative measure, we found that relevance networks derived from variance-reducing/stabilizing transformations showed the highest overall assortativity in the large sample limit ( $\approx 0.35$ ).

We next asked the question whether high-level network properties such as assortativity and modularity were consistent independent of the sample size used to estimate the association networks. We thus repeated the previous analysis for different sample sizes, ranging from  $n = 25$  to  $n = 9000$ . Figure 7 shows the estimated network assortativity and maximum modularity score estimates versus sample size. We found that for relevance networks derived from variance-reducing/stabilizing transformations, both assortativity and modularity monotonically increase with sample size. Both estimates stabilize around sample sizes  $n \approx p$ . For the remaining relevance networks, assortativity estimates monotonically increase with sample size, while mod-

ularity tends to decrease with sample size. In summary, this analysis implies that estimates of high-level network summary statistics such as assortativity and modularity are inconsistent compared to their large sample limit.

We next examined the edge overlap from correlation-based relevance networks (clr and tss transformations) and proportionality metrics (rhoprop and rhoshrink). We found a common core of 1086 edges between 349 OTUs that were present in all relevance networks. This consensus network also contained several tightly connected network modules with highly assortative inter-family associations (Figure 8A). Overall, we found that clr-, rhoprop- and rhoshrink-based networks shared the majority of common edges with rhoprop- and rhoshrink-based edge set differing only by a single edge. The tss-based relevance network comprised 779 unique edges not shared by any of the other networks (Figure 8B).

Additionally, we found that clr-, vst- and asinh-based correlation networks also shared a large common consensus core (Supplementary Figure S10A). Similarly, com-, rle- and tss-based networks showed a large edge set overlap (Supplementary Figure S10B). These observations confirmed the distinct groupings observed in the MDS embedding of Frobenius distances (Figure 2B).

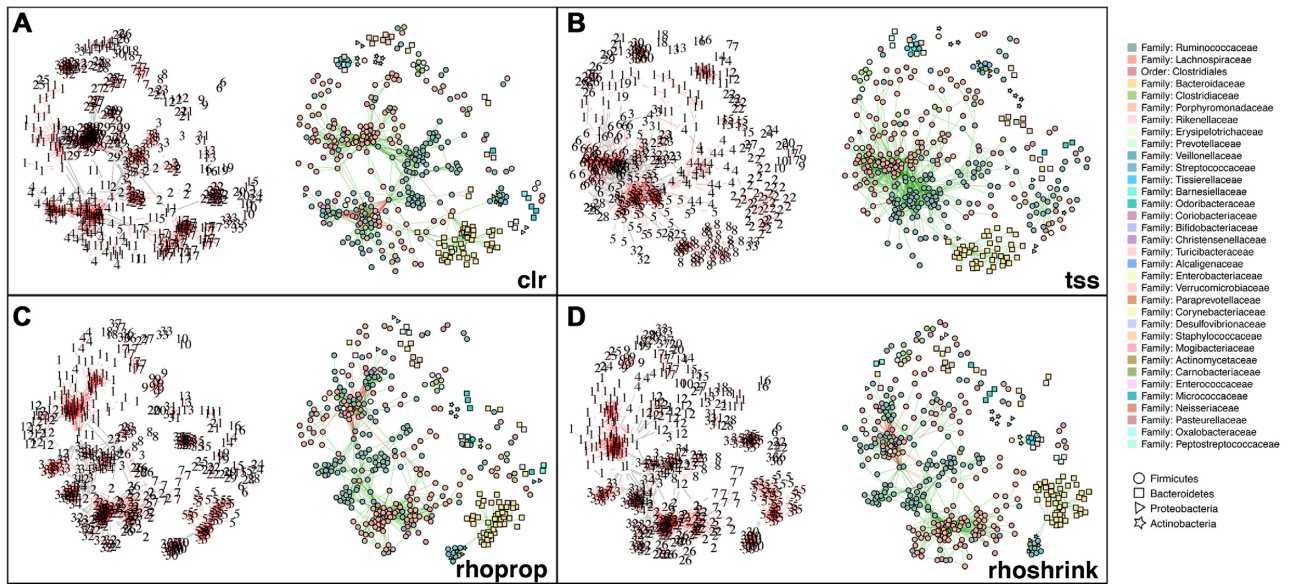


**Figure 5.** Circular dendrograms showing hierarchical clustering patterns among OTUs. Each point surrounding the circular dendrogram represents one of the 531 OTUs in our dataset. The color represents family annotation. Each dendrogram (A–D) has been cut hierarchically into 10 trees (representing the 10 orders to which these taxonomic families map). The gray and black shading is used to highlight different clusters that are numbered. Hierarchical clustering of clr-transformed OTUs is better at delineating taxonomic relationships than clustering of those using tss; rhoprop and rhoshrink produce similar clustering patterns. Additional methods are provided in Supplementary Figure S8.

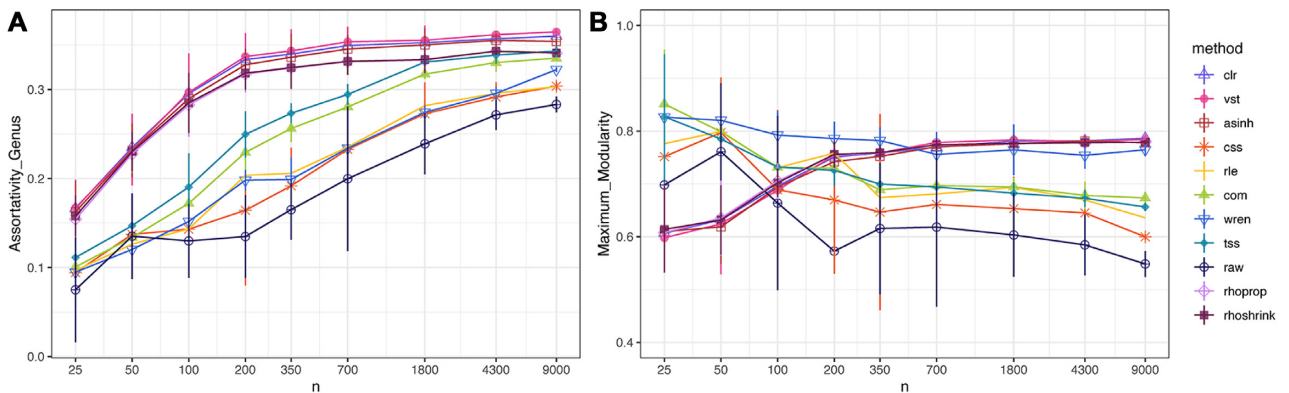
## DISCUSSION

Data normalization and inference of taxon–taxon associations from microbial genomic survey count data are two of the most basic statistical analysis tasks in modern microbiome research. To help the practitioner of microbial data analysis make informed choices about the different available normalization and association inference schemes, we have taken a closer look at the impact of data normalization on association estimation and several downstream exploratory data analysis tasks. Rather than asking what is the *best* method available for different analysis steps, we have leveraged the large available sample size in the AGP dataset and assessed the *consistency* of two ubiquitous linear association estimators for microbiome data, correlation and proportionality, under a wide range of realistic sample size scenarios, data normalization schemes and downstream data analysis tasks.

Our analysis revealed several important observations that have direct implications for best practice in microbiome data analysis workflows. First, we have confirmed that correlation and proportionality estimates are inconsistent in the low sample regime  $n < p$  when compared to large sample counterparts, in terms of both general large-scale association patterns (Figure 2) and downstream network summary statistics, including assortativity and modularity (Figure 7). While this phenomenon has been long appreciated in the statistical literature, we have established that shrinkage estimation, a popular statistical regularization scheme used in finance (33) and genomics research (27), can also improve association estimation for microbiome data, independent of the employed normalization method. Leveraging the close mathematical relationship between variance–covariance estimation and the concept of proportionality, we have also introduced a novel shrinkage proportionality estimator, rhoshrink, that is easy to compute and may prove



**Figure 6.** Community structure of relevance networks. (A–D) The left network of each panel shows module membership. Each numbered node represents the module annotation of an OTU in the graph. The networks on the right represent the corresponding taxonomic annotation of the OTU at the family (color) and phylum (shape) levels. Values stated next to method name represent the number of modules in the network. Node layout is conserved for both networks in each panel. Additional methods are provided in Supplementary Figure S9.



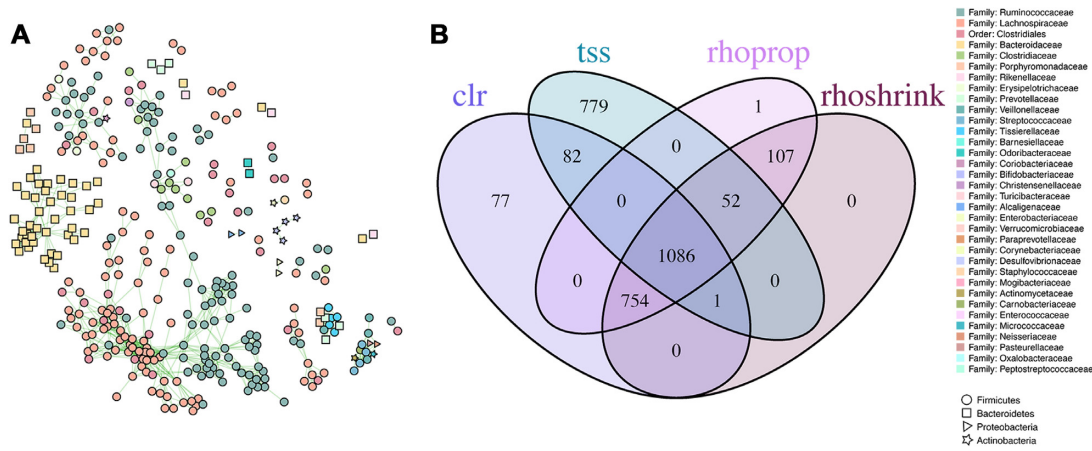
**Figure 7.** Community analysis of relevance network structure with increasing sample size. (A) Assortativity coefficient across sample size of genus annotation. (B) Maximum modularity score across sample size at 2000 edges. For all plots, lines represent mean and gray ribbons represent standard deviation from the mean.

useful in other scientific areas where compositional data are available.

On the AGP data, we have been able to categorize 10 data normalization/association estimation workflows into five coherent groups that show strong agreement across all sample size scenarios (Figure 2B). Most prominently, we have found that variance-reducing/stabilizing transformations lead to a high agreement of correlation or proportionality estimates. This was also confirmed in the downstream microbial relevance network comparison where clr-based correlation networks and proportionality association networks showed high agreement among the inferred edge sets (Figure 8). This implies that, in the presence of large sample sizes and large number of OTUs, differences between correlation and proportionality estimates are less pronounced than previously expected. An important observation on the

AGP dataset was that the empirical distributions of association estimates were universally right-skewed even in the randomly shuffled data scenario. This implies that irrespective of the data normalization/association inference workflow, one will observe a higher prevalence of positive associations. This phenomenon has been previously described in the context of microbial association inference across many different microbial habitats (14,47). While it is tempting to interpret these results as ecological features of the underlying microbial community in terms of higher prevalence of commensal rather than competitive microbial interactions, the positive skewness may also be due to technical limitations in the data generation process and shortcomings in current statistical estimation. For instance, truncation to zero effects for low sequencing read counts likely obstructs unbiased estimation of negative correlations and, in turn, proportional-





**Figure 8.** Shared interactions between relevance networks. (A) Consensus network of edges in common between four representative methods. Network contains 1086 edges between 346 OTUs. Node color represents family annotation and node shape represents phylum. (B) Venn diagram showing unique and shared interactions predicted from representative normalization methods.

ity. A possible remedy for this data-induced artifact is the application of more advanced semi-parametric correlation estimators that infer latent correlations under data truncation assumptions (21,48). A detailed investigation of semi-parametric and other estimators may provide a promising avenue for future research.

In many studies, microbial counting strategy has transitioned away from the use of OTUs and toward ASVs (3). We suspect that shrinkage may also improve association learning in these contexts where variance-stabilizing transformations are often used in analysis (49).

Despite the universal presence of positive skewness in association estimates for the AGP data, we have observed that variance-reducing/stabilizing transformations could reduce positive skewness in shrunk association estimates (Figure 3). Moreover, our results on microbial association network construction and clustering as typical downstream exploratory data analysis examples also revealed that variance-reducing/stabilizing approaches provided the most consistent estimation in terms of taxonomic and structural coherence, as measured by taxonomic cluster purity in spectral and hierarchical clustering (Figures 4 and 5) and network assortativity (Figures 6 and 7). Taken together, we can recommend any variance-reducing/stabilizing transformations followed by shrinkage estimation for association inference. However, transformations such as  $\text{asinh}$  and  $\text{clr}$  may be preferred since they are faster to compute than  $\text{vst}$ , while providing similar statistical properties. The resulting shrinkage correlation estimates can then also serve as input for more involved direct microbial network inference workflows that account for transitive correlations, adjust for additional covariates or model latent effects (14,39,50,51).

For relevance network estimation, consensus network construction, as put forward here for the AGP data (Figure 8), is a straightforward strategy to relax the influence of data normalization. For our AGP consensus network, we found that more than half of the top 2000 edges in the  $\text{tss}$ -,  $\text{clr}$ -,  $\text{rhoprop}$ - and  $\text{rhoshrink}$ -based relevance networks were in full agreement, connecting a subset of 346 OTUs. The inferred AGP consensus network comprised a majority of

positive edges and showed high assortativity at the genus level (0.39) and a maximum modularity of 0.8.

Assortativity increased in the consensus network compared to individual relevance networks. Notably, many taxa in the consensus network were frequently identified as key targets for microbiome therapeutics, such as prebiotic treatment and fecal microbiota transplants, including *Akkermansia muciniphila*, *Prevotella copri*, *Ruminococcus bromii* and *Faecalibacterium prausnitzii* (52,53).

Our computational data analysis workflow, available on GitHub and as Synapse project (see the ‘Data Availability’ section), is fully reproducible, provides all novel shrinkage estimators introduced here and allows easy extension and comparison to additional data normalization, estimation and downstream analysis tasks. For instance, future work could include the integration of more advanced zero-replacement strategies (54,55), application of popular data normalization schemes from single-cell data analysis (56) or the application of other correlation (21,48) or proportionality estimators, including those available in the `propr` package (23). Here, rather than using universal thresholding for sparsifying associations, more advanced selection strategies that control false discovery rates [as available in the `propr` package (23)] may improve the sample size consistency of the microbial association inference workflows.

Going forward, we believe that large-scale reproducible computational analysis workflows that focus on sample size-dependent consistency of statistical estimates are of paramount importance for deriving stable testable hypotheses about the complex interplay between host phenotype and the microbiome from large-scale microbial genomic survey data.

## DATA AVAILABILITY

The code and data used are available as a GitHub repository at <https://github.com/MichelleBadri/NormCorr-manuscript> and Synapse project syn21654780. Data used for this study were accessed from <ftp://ftp.microbio.me/AmericanGut/ag-2017-12-04/>.

The latest complete AGP dataset can be accessed on Qiita using study ID 10317 (25).

## SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

## ACKNOWLEDGEMENTS

We would like to thank the Simons Foundation.

## FUNDING

Flatiron Institute, a division of the Simons Foundation.  
Conflict of interest statement. None declared.

## REFERENCES

- Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.
- Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.
- Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**, 581–583.
- Willis, A.D. and Martin, B.D. (2020) Estimating diversity in networked ecological communities. *Biostatistics*, doi:10.1093/biostatistics/kxaa015.
- Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., Deng, L., Yeliseyev, V., Delaney, M.L., Liu, Q. *et al.* (2016) MDSINE: Microbial Dynamical Systems Inference Engine for microbiome time-series analyses. *Genome Biol.*, **17**, 121.
- Zwiener, I., Frisch, B. and Binder, H. (2014) Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PLoS One*, **9**, e85150.
- Mark Welch, J.L., Hasegawa, Y., McNulty, N.P., Gordon, J.I. and Borisov, G.G. (2017) Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *Proc. Natl Acad. Sci. U.S.A.*, **114**, E9105–E9114.
- Stömmeler, F., Gläser, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P.J., Gessner, A. and Spang, R. (2016) Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome*, **4**, 28.
- Aitchison, J. (1986) In: *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.
- Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V. and Egozcue, J.J. (2017) Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.
- McMurdie, P.J. and Holmes, S. (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.*, **10**, e1003531.
- Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
- Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
- Kurtz, Z.D., Muller, 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**, e1004226.
- Paulson, J.N., Stine, O.C., Bravo, H.C. and Pop, M. (2013) Robust methods for differential abundance analysis in marker gene surveys. *Nat. Methods*, **10**, 1200–1202.
- Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A. *et al.* (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**, 27.
- Kumar, M.S., Slud, E.V., Okrah, K., Hicks, S.C., Hannehalli, S. and Corrada Bravo, H. (2018) Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics*, **19**, 799.
- Martin, B.D., Witten, D. and Willis, A.D. (2020) Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann. Appl. Stat.*, **14**, 94–115.
- Brill, B., Amir, A. and Heller, R. (2019) Testing for differential abundance in compositional counts data, with application to microbiome studies. arXiv doi: <https://arxiv.org/abs/1904.08937>, 30 March 2020, preprint: not peer reviewed.
- Weiss, S., Treuren, W.V., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J. *et al.* (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.*, **10**, 1669–1681.
- Yoon, G., Gaynanova, I. and Müller, C.L. (2019) Microbial networks in SPRING: semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. Genet.*, **10**, 516.
- Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S. and Bähler, J. (2018) Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.*, **11**, e1004075.
- Quinn, T.P., Richardson, M.F., Lovell, D. and Crowley, T.M. (2017) propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.*, **7**, 16252.
- Jackson, D.A. (1997) Compositional data in community ecology: the paradigm or peril of proportions. *Ecology*, **78**, 929–940.
- McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y. *et al.* (2018) American Gut: an open platform for citizen science microbiome research. *mSystems*, **3**, e00031-18.
- Stein, C.M. (1981) Estimation of the mean of a multivariate normal distribution. *Ann. Stat.*, **9**, 1135–1151.
- Schäfer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol.*, **4**, 32.
- Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A. and Alm, E.J. (2017) Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.*, **8**, 1784.
- Holmes, S. and Huber, W. (2019) In: *Modern Statistics for Modern Biology*. Cambridge University Press, Cambridge.
- Burbidge, J.B., Magee, L. and Robb, A.L. (1988) Alternative transformations to handle extreme values of the dependent variable. *J. Am. Stat. Assoc.*, **83**, 123–127.
- 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. *Proc. R. Soc. Lond.*, **60**, 489–498.
- Erb, I. and Notredame, C. (2016) How should we measure proportionality on relative gene expression data? *Theory Biosci.*, **135**, 21–36.
- Ledoit, O. and Wolf, M. (2004) Honey, I shrunk the sample covariance matrix. *J. Portfolio Manage.*, **30**, 110–119.
- Rothman, A.J., Bickel, P.J., Levina, E. and Zhu, J. (2008) Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, **2**, 494–515.
- Bien, J. and Tibshirani, R.J. (2011) Sparse estimation of a covariance matrix. *Biometrika*, **98**, 807–820.
- Herdin, M., Czink, N., Özcelik, H. and Bonek, E. (2005) Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels. In: *Proceedings of the IEEE Vehicular Technology Conference*. Stockholm.
- Ng, A.Y., Jordan, M.I. and Weiss, Y. (2001) On spectral clustering: analysis and an algorithm. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01)*. MIT Press, Cambridge, MA, pp. 849–856.
- Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. and Kohane, I.S. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. U.S.A.*, **97**, 12182–12186.

39. Claussen, J.C., Skievcicicene, J., Wang, J., Rausch, P., Karlsen, T.H., Lieb, W., Baines, J.F., Franke, A. and Hütt, M.-T. (2017) Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLoS Comput. Biol.*, **13**, e1005361.
40. Friedman, J. and Alm, E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.*, **8**, e1002687.
41. Faust, K., Sathirapongsasuti, J.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**, e1002606.
42. Lozupone, C., Faust, K., Raes, J., Faith, J.J., Frank, D.N., Zaneveld, J., Gordon, J.I. and Knight, R. (2012) Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Res.*, **22**, 1974–1984.
43. Newman, M. E.J. (2003) Mixing patterns in networks. *Phys. Rev. E*, **67**, 026126.
44. Newman, M.E. (2006) Modularity and community structure in networks. *Proc. Natl Acad. Sci. U.S.A.*, **103**, 8577–8582.
45. Clauset, A., Newman, M. E.J. and Moore, C. (2004) Finding community structure in very large networks. *Phys. Rev. E*, **70**, 066111.
46. Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *Inter J.*, **1695**, 1–9.
47. Faust, K., Lima-Mendez, G., Lerat, J.-S., Sathirapongsasuti, J.F., Knight, R., Huttenhower, C., Lenaerts, T. and Raes, J. (2015) Cross-biome comparison of microbial association networks. *Front. Microbiol.*, **6**, 1200.
48. Yoon, G., Carroll, R.J. and Gaynanova, I. (2020) Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, **107**, 609–625.
49. Sprockett, D., Fischer, N., Boneh, R.S., Turner, D., Kierkus, J., Sladek, M., Escher, J.C., Wine, E., Yerushalmi, B., Dias, J.A. *et al.* (2019) Treatment-specific composition of the gut microbiota is associated with disease remission in a pediatric Crohn's disease cohort. *Inflamm. Bowel Dis.*, **25**, 1927–1938.
50. Biswas, S., McDonald, M., Lundberg, D.S., Dangl, J.L. and Jojic, V. (2016) Learning microbial interaction networks from metagenomic count data. *J. Comput. Biol.*, **23**, 526–535.
51. Kurtz, Z.D., Bonneau, R. and Müller, C.L. (2019) Disentangling microbial associations from hidden environmental and technical factors via latent graphical models. bioRxiv doi: <https://doi.org/10.1101/2019.12.21.885889>, 23 December 2019, preprint: not peer reviewed.
52. Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E.G., Abramson, S.B. *et al.* (2013) Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife*, **2**, e01202.
53. Scott, K.P., Jean-Michel, A., Midtvedt, T. and Hemert, S.V. (2015) Manipulating the gut microbiota to maintain health and treat disease. *Microb. Ecol. Health Dis.*, **26**, 25877.
54. Palarea-Albaladejo, J. and Martín-Fernández, J.A. (2015) zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometr. Intell. Lab.*, **143**, 85–96.
55. Cao, Y., Zhang, A. and Li, H. (2020) Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika*, **107**, 75–92.
56. Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S. and Marioni, J.C. (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods*, **14**, 565–571.