Normalization methods for microbial abundance data strongly affect correlation estimates

Michelle Badri[1], Zachary D. Kurtz[2], Christian L. Müller[3], and Richard Bonneau[1,3*]

[1]Department of Biology, New York University, New York, 10012 NY, USA
[2]Lodo Therapeutics, New York, 10016 NY, USA
[3]Flatiron Institute, Simons Foundation, New York, 10010 NY, USA
[4]Computer Science Department, Courant Institute, New York, 10012 NY, USA
*rb133@nyu.edu

August 31, 2018

Abstract

Consistent normalization of microbial genomic survey count data is fundamental to modern microbiome research. Technical artifacts in these data often obstruct standard comparison of microbial composition across samples and experiments. To correct for sampling bias, library size, and technical variability, a number of different normalization methods have been proposed, including adaptations of RNA-seq analysis work flows and log-ratio transformations from compositional data analysis. Additional characterization of the effects of data normalization on higher-order summary statistics is required. We review and compare popular data normalization schemes and assess their effect on subsequent correlation estimation. Application of these normalization methods to the largest publicly available human gut microbiome dataset show substantial variation among patterns of correlation. We show that log-ratio and variance-stabilization transformations provide the most consistent estimates across experiments of different sample sizes. We also show that data analysis methods that rely on correlation, such as data clustering and network inference, differ depending on the normalization schemes. These findings have important implications for microbiome studies in multiple stages of analysis.

1 Author Summary

Scientific surveys of microbial communities have become ubiquitous due to decreasing sequencing costs and advances in our understanding of the roles of microbes in ecosystems. In the course of routine investigation, between the steps of sequencing marker genes and applying data analysis pipelines, we are given microbial count data. Data normalization and/or transformation is required to account for between-sample differences that arise from the way communities are measured. These methods can fail to take into account the underlying features of microbial community data. We use a large and publicly available dataset of microbes profiled from the human gut to test for consistency of normalization methods. We find that methods that explicitly account for the underlying compositional nature of the data are more consistent across independent subsamples and the inferred microbial associations better recapitulates phylogeny, as assessed by 16S gene sequences.

2 Introduction

Recent advances in large-scale microbial amplicon and metagenomics sequencing data collection efforts provide samples across different microbial habitats that are amenable to quantitative analysis. Following the organization of sequence data into OTUs (Operational Taxonomic Units), via pipelines such as qiime and mothur, taxa counts are then collected in OTU (or higher-level taxonomy) tables. Downstream analysis tasks include assessing community diversity, associating bacterial compositions to system-specific ecological and biomedical co-variates and learning microbe-microbe associations [4, 23].

However, technical artifacts inherent in microbial abundance data preclude the application of such analyses pipelines directly on raw counts. These tables of counts, where rows represent OTUs and columns represent samples, are sparse or contain a very high proportion of zeros. Crucially, due to the nature of microbial sampling, the raw counts should be considered only relative to the sum total of individual counts or contingent on the sampling depth. Total number of counts for any given observation is typically 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 taxa 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, in the majority of experimental designs the community size is unknown and thus our data is best thought of as compositional (each OTU fraction of total counts, total community size unknown) [3, 16, 27].

Robust analysis of microbiome data must therefore be performed using principles from the field of compositional data analysis [1, 11]. Additionally, technical variation due to sequencing such as differences in amplification biases, and batch effects due to multiple sequencing runs can mislead investigators [18]. To ameliorate these biases, various normalization methods have been proposed to correct for sampling bias, library size, and technical variability [13, 2, 12, 22]. We consider a variety of methods, either adapted from RNA-seq pre-processing workflows, or designed specifically for compositional data, e.g., log-ratio transformations. Individually, these methods have been used upstream of some analysis task [29], but lacking a gold standard complicates their proper evaluation for statistical inference.

In particular, the effect of data normalization methods on higher-order summary statistics has remained elusive. Previous work has assessed correlation from sequencing-generated count data. Weiss et al. (2016) examined various correlation detection strategies for microbial data and found that some commonly-used tools varied widely in their sensitivity and precision, using a synthetic data setup [28]. Careful consideration of data transformation has been given in gene expression sequencing studies [30]. However, isolating the effects of data transformations specifically for correlation detection is still an outstanding task for microbiome data analysis.

This is further complicated by the lack of verified microbial interactions in most microbial ecosystems. Although synthetic data is a useful for evaluating correlation detection algorithms, we risk over-fitting our methods and models to engineered assumptions. Therefore, we propose a consistency-based setup to evaluate correlation, reasoning that, if we have no *a priori* preference for a particular data transformation for a given dataset then we should evaluate each estimator under best possible conditions: an asymptotic sampling scheme. Therefore, we subsampled a large microbiome dataset, applied a data transformation and compared the estimated correlations to it's own large-sample 'gold standard'. This gives us an assumption-free evaluation of how a particular transform may disrupt correlations under a more restrictive (but still realistic) sample size.

We examined the effects of microbiome count data normalization methods on correlation estimates from amplicon-based genomic survey data. The estimation of pairwise OTU-OTU correlations represents one of the key statistical structure for many downstream analysis tasks, including discriminant analysis, clustering, and relevance networks. We observe that across a large range of sample sizes, normalization methods are empirically inconsistent in (i) resulting correlation estimates, (ii) hierarchical clustering of taxa, and (iii) community structures assessed by relevance networks.

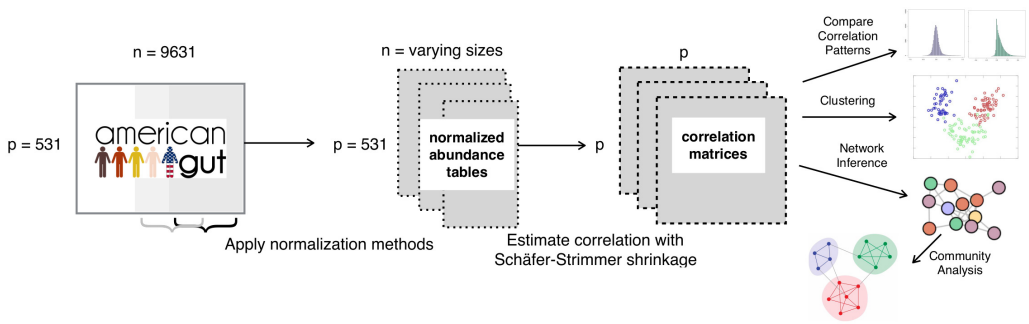

Figure 1: Our pipeline to examine the effects of normalization methods on correlation estimation using shrinkage: a common input to downstream analysis

3 Methods

3.1 Asymptotic sample sizes from real American Gut Project data

We examined the effects of the different normalization methods on data from the American Gut Project [17]. We filtered the dataset to contain only fecal samples whose sequencing depths fall above the 10th quantile and removed taxa that were present in fewer than 30% of all samples. This resulted in a data matrix that comprises $p = 531$ taxa and $n = 9631$ samples.

To examine the effects of the transformations on matrices of various sample sizes, we sub-sampled the OTU table to varying sizes of $25 <= n < 9631$. Therefore we are able to estimate correlation based on real data asymptotic n values. We use the correlation estimates at the largest available sample sizes as the 'gold-standard' reference, and applied various stability measures to profile consistency across a large range of sample sizes. Finally, to simulate data under null correlation, we shuffled OTU count data across samples prior to normalization.

3.2 Normalization methods

We are given OTU counts, collected over n samples and stored in a matrix $W \in \mathbb{N}_0^{n\times p}$, a collection of p-dimensional row vectors $w^{(j)} = [w_1^{(j)}, w_2^{(j)}, \ldots, w_p^{(j)}]$, where $j = 1, \ldots, n$ 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 $\{0, 1, 2, \ldots\}$. Let the total OTU count for sample j be $m^{(j)} = \sum_{i=1}^{p} w_i^{(j)}$. A pseudocount of 1 is included for calculation of geometric means and applied to non-negative input data to avoid the undefined log[0].

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. Otherwise known as total sum scaling (TSS) or total sum normalization, we have

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

where the sample space of the data is the $p-1$ Simplex.

Cumulative Sum Scaling In addition to generating compositional artifacts, TSS may also place undue influence on taxa that are highly sampled due to technological sequencing biases, i.e. by being over-represented in the scaling factor $m^{(j)}$ [18]. To reduce the influence of these highly abundant taxa in a sparse environment, Paulson et al (2013) develped the cumulative sum scaling (CSS), along with the metagenomeSeq package [22]. Rather than normalizing by the total sum, as in TSS, CSS selects a scaling factor that is a fixed quantile of OTUs counts. CSS is formally specified (changing only the notation for consistency) as

$$CSS(w^{(j)}) = \frac{w_i^{(j)}}{m_{\hat{l}}^{(j)}} N, \in \mathrm{R}_{\geq 0}^{n\times p}.$$

where the scaling factor for sample j is $m_{\hat{l}}^{(j)} = \sum_{i|w_{ij}\leq q_l^{(j)}} w_i^{(j)}$, the sum of read counts up to and including the lth quantile $q_l^{(j)}$. N is a pre-specified constant, chosen such that the resulting data vector resembles the units of the original counts. Herein N was selected as 1000 which is also the default value in the metagenomeSeq implementation of CSS. The sample space of CSS-transformed data is that of non-negative real numbers, denoted by $\mathrm{R}_{\geq 0}$.

For sample quantile index $\hat{l}$, and denoting the median lth quantile across all samples as $\bar{q}_l = med_j \left\{q_l^{(j)}\right\}$, $q_l^{(j)}$ is determined such that the quantity $\delta_l = med_j \mid q_l^{(j)} - \bar{q}_l \mid$, the median absolute deviation of sample quantiles, is empirically stable. For instance $\hat{l} := \inf\{\delta_{l+1} - \delta_l \geq \beta\delta_l : 1 \leq l < n\}$, for the threshold β, the relative difference between median differences, which is set to 0.1 in the original paper. Thus, the scaling factor is defined by summing all the counts up to the smallest value of l that is stable, on average, across all samples.

Common Sum Scaling Common sum scaling (COM), suggested by [18] as an alternative to rarefying, is simply dividing the counts scaled to the minimum depth of each sample:

$$COM(w^{(j)}) = \left[\left\lfloor w_1^{(j)} \frac{m^{(min)}}{m^{(j)}} \right\rfloor \ldots, \left\lfloor w_p^{(j)} \frac{m^{(min)}}{m^{(j)}} \right\rfloor\right] \in \mathbb{R}^{n\times p},$$

where $m^{(min)} = \inf\{m^{(1)}, m^{(2)}, \ldots, m^{(n)}\}$ and $\lfloor\cdot\rfloor$ is an operator that rounds the proportion to the nearest integer.

Relative Log Expression Although initially developed for gene expression data, the normalization methods introduced with the DESeq/edgeR package have been applied to with microbiome data [2]. The first proposed normalization method is relative log expression (RLE).

Let $g(x) = \left(\prod_{i=1}^{m} x_i\right)^{1/m}$ be the geometric mean of a length m vector x, let $w_i = W_i^T = \left[w_i^{(1)}, \ldots, w_i^{(n)}\right]$ be the vector of counts of OTU i over n samples (a transposed column vector of count matrix W), and define the numeric scaling factor for sample j:

$$\begin{aligned} s^{(j)} &= med^{(j)}\left[\frac{w_1^{(j)}}{g(w_1)}, \ldots, \frac{w_p^{(j)}}{g(w_p)}\right], \\ \bar{s}^{(j)} &= \frac{s^{(j)}}{g(s^{(\cdot)})}, \end{aligned}$$

where the function $med[x]$ selects the median value of vector x and $s^{(\cdot)} = \left[s^{(1)}, \ldots, s^{(n)}\right]$, a collection of the sample scaling factors. Finally, with a global scaling factor $\hat{s}_c = \frac{1}{n}\sum_{j=1}^{n} \bar{s}^{(j)}$, the arithmetic mean of all normalized scaling factors, we can define the RLE as

$$RLE(w^{(j)}) = \left[\hat{s}_c \frac{w_1^{(j)}}{\bar{s}^{(j)}}, \ldots, \hat{s}_c \frac{w_p^{(j)}}{\bar{s}^{(j)}} \right] \in \mathrm{R}_{\geq 0}^{n \times p}.$$

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.

Variance Stabilizing Transformation Also part of the DESeq package, the goal of variance stabilizing transform (VST) is to factor out the dependence of the variance on the mean (over-dispersion) [2]. Consider the mean-dispersion relation $v(\mu) = \frac{1}{n-1} \sum \left(\frac{w_i^{(j)}}{\hat{s}^{(j)}} - \hat{\mu}_i^{(j)} \right)^2$ where we compute "size factors" $\hat{s}^j = med_i \left(\frac{w_i^{(j)}}{g(w^{(j)})} \right)$. $\hat{\mu}_i^{(j)}$ is the average of counts to size factor ratios from sample j: $\mu_i^{(j)} = \frac{1}{n} \sum_{i=1}^{n} \frac{w_i^{(j)}}{\hat{s}^j}$. The reciprocal of the square root of the variance is then numerically integrated:

$$\mathrm{VST}(w^{(j)}) = \int^{w_i^{(j)}} \frac{\mathrm{d}\mu}{\sqrt{v(\mu)}} \in \mathbb{R}^{n \times p}$$

and the integral, where the fit is approximated by a spline function, is evaluated for each count value in the column. The fit in the DESeq package is accomplished 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: $\hat{v}(\mu) = v(\hat{\mu}_i^{(j)}) - z^{(j)}$ where $z^{(j)} = \frac{\mu_i^{(j)}}{n} \sum_{i=1}^{n} \frac{1}{\hat{s}^j}$ Within the local regression, parameters are chosen such that the for large counts the data is asymptotically scaled to be equal to the logarithm base 2 of normalized counts.

Centered Log-Ratio Proposed originally by Aitchison, log-ratio transformations are capable of removing the unit-sum constraint of compositional data, allowing ratios to be analyzed in the Euclidean spaceAitchison [1], Kurtz et al. [12]. The Centered log-ratio (CLR) is

$$CLR(w^{(j)}) = \left[\log \frac{w_1^{(j)}}{g(w^{(j)})}, \ldots, \log \frac{w_p^{(j)}}{g(w^{(j)})} \right] \in \mathbb{R}^{n \times p},$$

where the ratio is taken with respect to the geometric mean of the whole composition. The resulting data lies in a $p-1$ hyperplane of p-dimensional Euclidean space. Although other log-ratio transformations exist, i.e. that use a fixed OTU for the reference/denominator rather than geometric mean, we prefer the CLR here since the dimensionality after the transformation remains p Gloor et al. [11], Silverman et al. [26].

3.3 Covariance and Correlation Estimation

Given a transformation function $f : \mathbb{N}_0^{n \times p} \to \mathcal{X}^{n \times p}$, which represents the normalizations described above,a standard pipeline is to generate either the empirical (sample) covarinace or Pearson correlation matrix. It is, however, a well-known phenomenon that empirically estimated correlations/covariances are inadmissible in the setting where $p \gg n$. For example, type I errors may be grossly inflated and $\hat{S}$ cannot be inverted. This is of particular concern when the downstream analysis is an eigendecomposition of a covariance matrix (e.g. PCA). To account for this phenomenon, a common approach is to introduce structural assumptions into the covariance estimator. For example, if we assume that a true ecosystem is only sparsely correlated among species, we can estimate many fewer than $\frac{p(p-1)}{2}$ correlations. A principled, data-driven approach to introducing structural sparsity are SchÃďfer-Strimmer shrinkage estimates SchÃďfer and Strimmer [25]. Pearson correlation with shrinkage results in positive-definite correlation matrices (Fig 1), since correlations that do not have strong evidence (i.e. they are smaller than some threshold λ) are set to zero while all others are biased towards λ. We estimate covariance with shrinkage, s_{ij}^*

$$s_{ij}^* = \begin{cases} s_{ii} & i = j \\ r_{ij}^* \sqrt{s_{ii} s_{jj}} & i \neq j \end{cases}$$

where r_{ij}^* is defined as:

$$r_{ij}^* = \begin{cases} 1 & i = j \\ r_{ij} min(1, max(0, 1 - \widehat{\lambda}^*)) & i \neq j \end{cases},$$

where shrinkage intensity is

$$\widehat{\lambda}^* = \frac{\sum_{i \neq j} \widehat{Var}\left(r_{ij}\right)}{\sum_{i \neq j} \left(r_{ij}^2\right)}.$$

Variance of empirical correlations therein is defined as $\widehat{Var}\left(r_{ij}\right) = \frac{n}{(n-3)^2} \sum_{k=1}^{n} \left(w_{kij} - \bar{w_{ij}}\right)^2$, where $w_{kij} = \left(x_{ki} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_j\right)$ and $\bar{w_{ij}} = \frac{1}{n} \sum_{k=1}^{n} w_{kij}$. The covariance estimate is then converted to our final correlation estimate, $\bar{r}_{ij}$, using:

$$\bar{r}_{ij} = \begin{cases} 1 & i = j \\ \frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}} & i \neq j \end{cases}$$

3.4 Comparing Correlation Patterns

The matrices were then compared using Frobenius, Förstner and spectral distances [10]. Each of these distance metrics are useful in this application due to their natural relationship to both the eigenvalues and singular values of our correlation estimates. The Frobenius norm (widely used in non-negative matrix factorization and correlation matrix regularization) is defined as the differences in square root of the sum of the absolute squares of its elements. We define the Frobenius Distance as: $d\left(A, B\right) = \sqrt{\sum_{i,j} \left(A_{ij} - B_{ij}\right)^2}$. The spectral norm is a unitary invariant matrix norm under the property that eigenvalues and eigenvectors are conserved under unitary transformations. The spectral distance is the difference of the largest eigenvalues of the respective matrices. Spectral Distance: $d\left(A, B\right) = \sqrt{\lambda_{max}\left(A\right) - \lambda_{max}\left(B\right)}$

The resulting distance matrices were visualized using multidimensional scaling (MDS). The distance between correlation estimates, calculated from sub-samples of increasing size, were used as a measure of consistency.

3.4.1 Clustering

Clustering of identified taxa can uncover biological associations related to host covariates that affect the microbiome, (i.e. clusters of taxa associated with specific diets) as well as experimental and batch effects [17]. To perform spectral clustering, we calculated similarity as $1 - \sqrt{\frac{1-R}{2}}$ where R is the correlation estimate. Affinity matrices were computed using a k-nearest neighbor (k =3) to obtain a sparse and symmetric representation of the correlation estimates. The normalized graph Laplacian was computed and eigenvalues of the resulting matrices were analyzed. In spectral clustering the selection of the clusters, or the first k eigenvalues and eigenvectors of the graph Laplacian, is often based on the eigengap, that is, the the position of first non-zero eigenvalue in a ranked list of eigenvalues [21]. To perform hierarchical clustering, correlation matrices were converted to squared Euclidean dissimilarity using $\frac{1-R}{2}$. Clustering was then performed using Ward's method.

3.4.2 Relevance networks

To examine the effects of normalization on estimation of microbe-microbe interaction networks we analyze relevance networks derived from correlation estimates. Pairwise correlation values between OTUs were ranked by absolute value. We then select the correlation pairs with the highest absolute value of correlation to be interpreted as interactions or edges in our relevance network. Multiple studies have found a higher prevalence of positive correlations between phylogenetically related taxa in human gut datasets [6, 9, 8, 12, 14]. We use this observed potential for phylogenetic assortativity in our relevance networks to compare the results of different normalization methods in the absence of a gold-standard network.

3.4.3 Community Analysis

Microbe-microbe interaction networks and their underlying ecological properties can be investigated through community analysis of interaction patterns. We examined the underlying community structure in our relevance networks and compared these results to known taxonomic associations. In this context the sign of the correlation coefficients are interpreted as the sign of the association. Modularity was computed according to the fast-greedy modularity algorithm described in [5] and implemented in the igraph package in R [7]. This modularity grouping was compared to the phylogenetic annotation of the subgraphs in the network at the top 2000 edges. We examined assortativity by taxonomic grouping to compare how well the relevance networks captured phylogenetic relationships [20]. Assortativity within modules was computed by isolating the modules determined from the fast-greedy modularity into subgraphs and combining them into one network to create a graph composed of disconnected modules (only edges within modules, no edges between modules). Conversely, to analyze assortativity between modules the edges within each module were removed and assortativity was computed (only edges between modules, no edges within modules). We also examined modularity and assortativity across network sizes using a relevance network built using all (n=9631) samples.

4 Results

With a growing number of statistically robust analysis tools there is an urgent need for increased understanding of data pre-processing and its impact. In this contribution we reveal, through focus on a prevalent intermediate structure: correlation and covariance matrices, that normalization methods can greatly influence analysis outcome.

Normalization methods produce different correlation structures

Transformations that remove compositional artifacts (CLR) and stabilize the variance (VST) result in considerably different patterns of correlation than those that do not (Fig. 2). The Frobenius distance and spectral distance between correlation estimates of OTU tables with increasing numbers of samples show diverging structures in correlation between methods and convergence across sample size within the same method. Smaller numbers of samples lead to similar correlation patterns across normalization methods (Fig. 2A). In MDS plots of spectral and Frobenius correlation patterns between CLR and VST remain similar despite changing sample sizes (Fig. 2A, Supplementary Fig. 3). The distance between estimates between sample sizes was used to evaluate the rate at which normalization methods arrived at stable patterns of correlation. CLR and VST arrived at consistent estimates of correlation fastest, followed closely by CSS and DESEQ normalization methods (Fig. 2B, Supplementary Fig. 3). The distance between correlation estimates of larger random sub-samples is very small suggesting that the estimates are more consistent regardless of which samples were included.

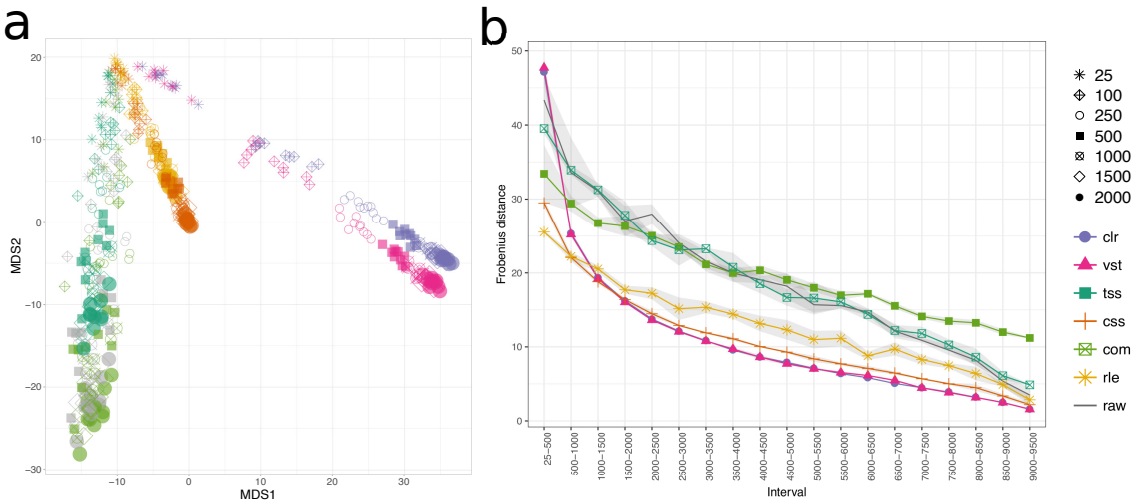

Figure 2: Transformations that remove compositional artifacts (CLR) and stabilize the variance (VST) result in substantially different patterns of correlation. A) Multidimensional scaling representation of Frobenius distance between correlation structures of varying size estimated from different normalization methods. These estimates are also compared to untransformed or raw count values (dark grey points). B) Frobenius distance between sub-samples of different sizes. Lines represent mean and grey ribbon represent standard deviation from the mean. (color scheme as in A)

Normalization methods that remove compositional artifacts (CLR) and stabilize the variance (VST) result in similar patterns of correlation that diverge from estimates without log or log-like transformations. By examining the distance between correlation matrices through changing sample sizes as a measure of consistency because we can expect that some spurious correlations vanish with increasing sample size[9]. At small values of n correlation structures are similar regardless of normalization prior to estimation -closer to random, more spurious associations. While we examine exceptionally large sample sizes in this work, we see that CLR and VST arrive at consistent estimates with far fewer samples than other methods. This assessment of consistency is relevant as many microbiome studies have small sample sizes due to clinical, environmental, and experimental limitations.

CLR and VST produce similar distributions of correlation values (that appear closer to normal distributions) than other methods (Fig 3A. and 3B.). The distribution resulting from TSS is very similar to that from raw counts.

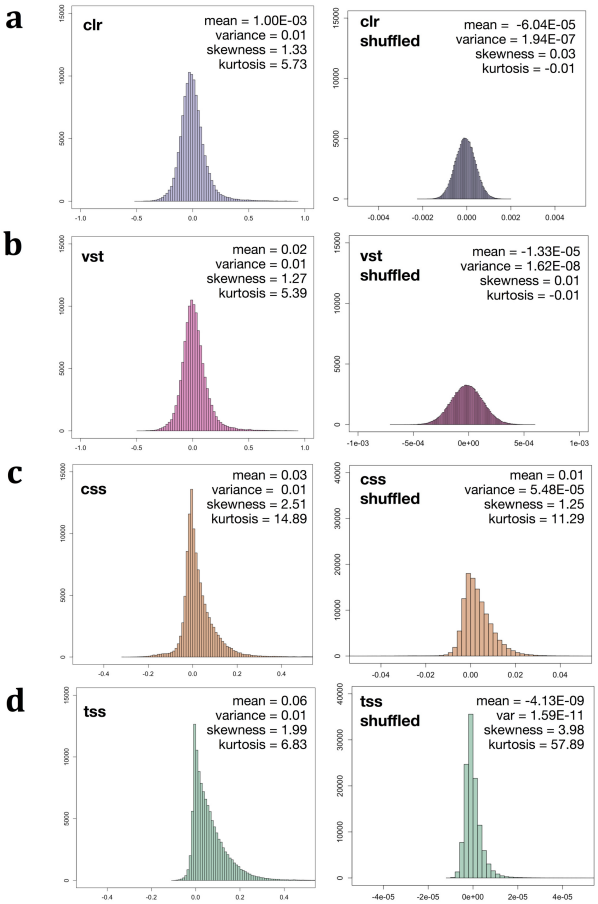

Figure 3: Frequencies of correlation values reveal different underlying distributions after transformation. Histograms of correlation frequency of matrices after application of different normalization methods. Frequencies were plotted with 200 bins and mean, variance, skewness and kurtosis are shown for each distribution.

Interestingly, histograms of correlation frequency reveal distributions of correlation values are positively skewed. CLR and VST, though capable of transforming counts to near-normal distribution, are unable to completely remove this phenomena from real OTU data. However these transformations are the only methods capable of removing positive skew from the shuffled OTU data. This suggests the underlying OTU data has a strong positive skew, and our results show the log and log-like transformations are best at ameliorating this effect. The measure of skewness in this regard has its limitations: it is not specific to the length of the right-handed tail and may reflect attributes of the left-handed tail that are being affected by normalization or permutation.

Differently normalized correlation structures lead to different clustering results

The eigengaps resulting from CLR and VST are relatively larger than all other normalization methods (Supplement 4). The clusters resulting from performing spectral clustering on the correlation matrix after each normalization methods also differ. The clusters predicted by CLR and VST are more homogeneous in terms of phylogenetic annotation. (Family level shown in Fig. 4, Supplement 5.).

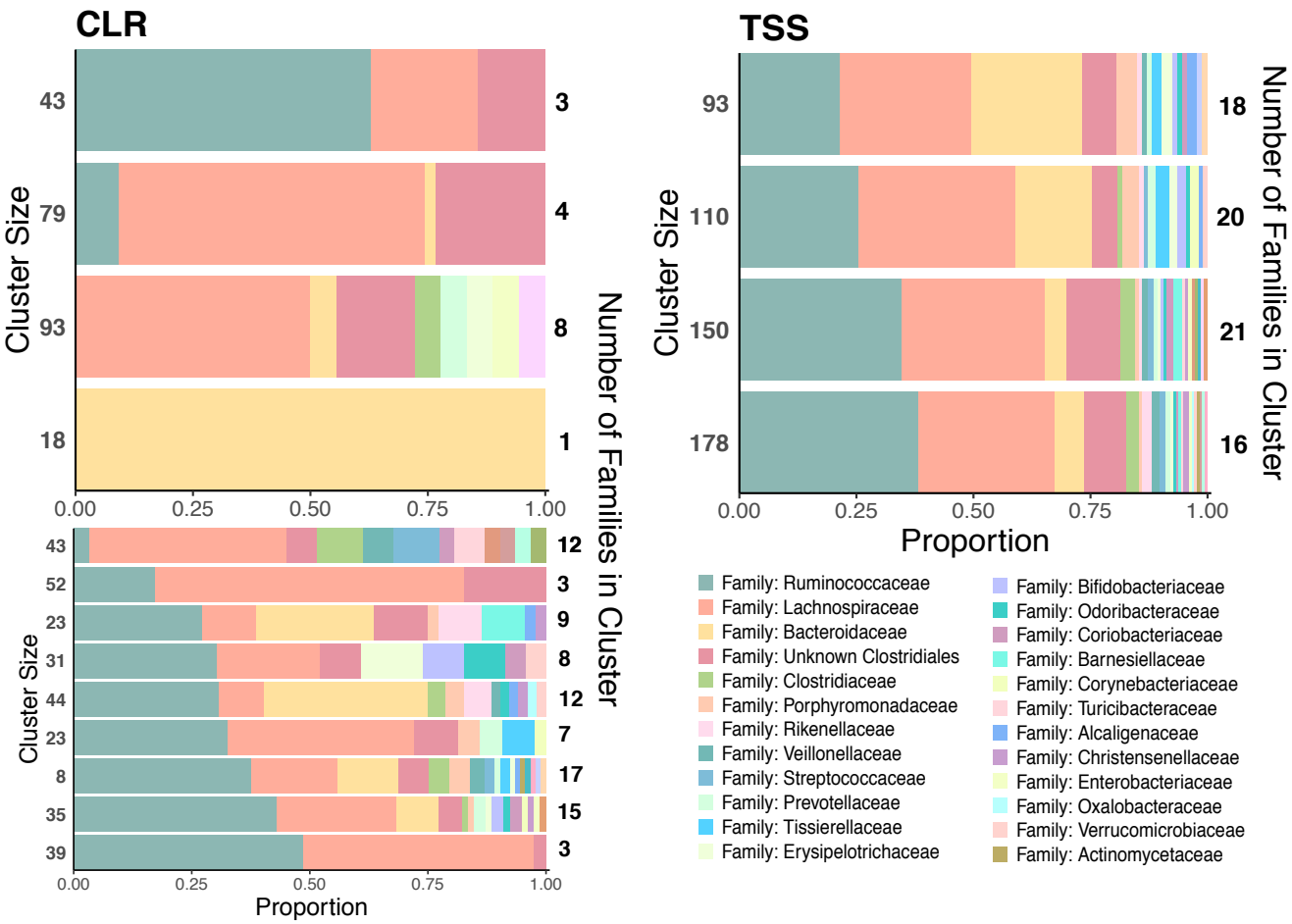

Figure 4: Clusters of OTUs after CLR transform produce more taxonomically homogeneous clusters than TSS The stacked bar plots represent the composition of OTUs in each cluster at the Family level. Clusters are vertically ordered from highest percentage of the most abundant Family: Rumincococcae. Horizontally the order represents the highest percentage in each cluster. Numbers on the left axis represent the number of OTUs in each cluster. Numbers on the right axis represent the number of taxonomic groups at the order level are in each cluster.

When analyzing eigenvalues associated with each normalization method CLR and VST have a larger eigengap (the first non-zero eigenvalue) (Supplementary Fig). The eigengap is an important and robust heuristic in choosing number of clusters for many methods of clustering [21, 15]. We show that these numbers of clusters represent an ability to recover groupings at lower phylogenetic levels for CLR and VST transformations. We also show that hierarchical clustering results in a similar homogeneous grouping of OTUs with similar taxonomic annotation (Fig. 5 and Supplement 6).

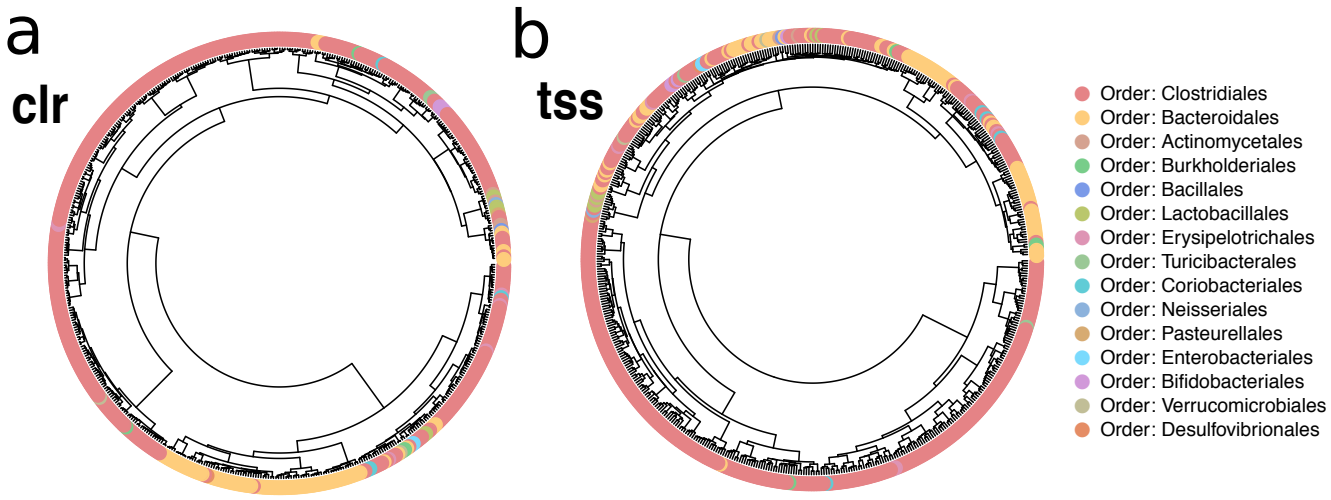

Figure 5: Circular dendrograms reveal that hierarchical clustering results in more homogenous clustering by taxonomic annotation. A larger number of taxa of orders Enterobacteriales and Bacteriodales cluster together in CLR than in TSS.

Normalization methods result in relevance networks with different community structures

To compare the structure of the relevance networks we examined their community structure in terms of modularity and assortativity. The proposed modular division from the fast-greedy modularity algorithm results in groups that largely correspond to subgraphs with high concentrations of inter-phyla edges. This division is more evident in the CLR and VST networks than in TSS and CSS. The Bacteriodetes form distinct modules with many intra-phyla edges and relatively few edges connecting to other phyla (Fig. 6). Proteobacteria comprise modules with very high numbers of intra-phyla edges in CLR, TSS and VST (Fig. 6 and Supplement 7). Firmicutes in CLR and VST networks are divided into a higher number of modules that distinguish the families Lachnospiraceae and Rumincococcae.

We assess the performance of normalization techniques by their recovery of taxon-taxon interactions among top edges in their relevance networks.

CLR and VST have the highest number of inter-family edges as determined by phylogenetic assortativity (Fig. 7A.). When examining edges only within modules CLR, VST and CSS have high assortativity coefficients signifying they have a relatively high number of inter-family edges(Fig. 7B.). The assortativity coefficients of edges only between modules has a much lower range of values and varies widely between networks. CLR and VST over an average of 20 networks produce a higher number of inter-genus edges between modules (Fig. 7C.).

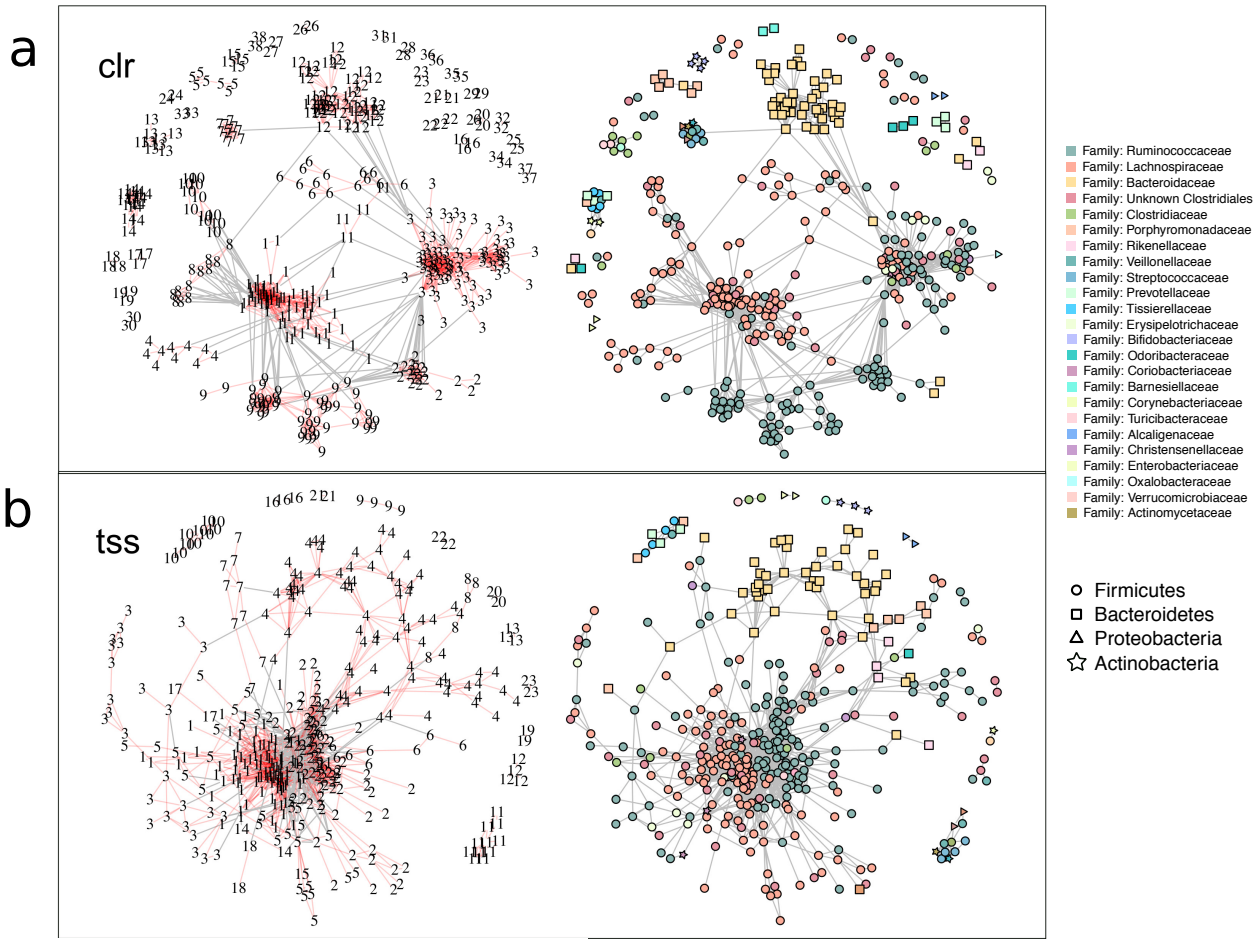

Figure 6: CLR produces more modular networks that separate communities at the Family level than TSS. For networks on the left of each panel every node represents an OTU labeled with module annotation as predicted by the Fast-Greedy modularity algorithm. The networks on the right represent the corresponding taxonomic annotation of the OTU at the family level. Layout using the force-directed Fruchterman-Reingold algorithm was conserved for both networks in each panel for comparison

We examine a simple approximation of co-occurrence based microbial network inference by performing community analysis on relevance networks. The resulting networks show considerable edge overlap between normalization methods. VST, CLR and to a lesser degree RLE form more modular networks (dense connections between the nodes within modules but sparse connections between nodes in different modules). For simplicity, we examined community structure at a fixed number of edges between the fixed number of nodes representing the 531 taxa in our relevance networks. Compared to the modularity score across edge sizes the observation that CLR and VST have more modular network structures is true for a large range of network sizes. CLR appears to produce slightly more modular networks with sample sizes above 1000. These highly interconnected modules representing key phyla have also been observed in previous publications examining similar subsets of AGP data [12, 6].

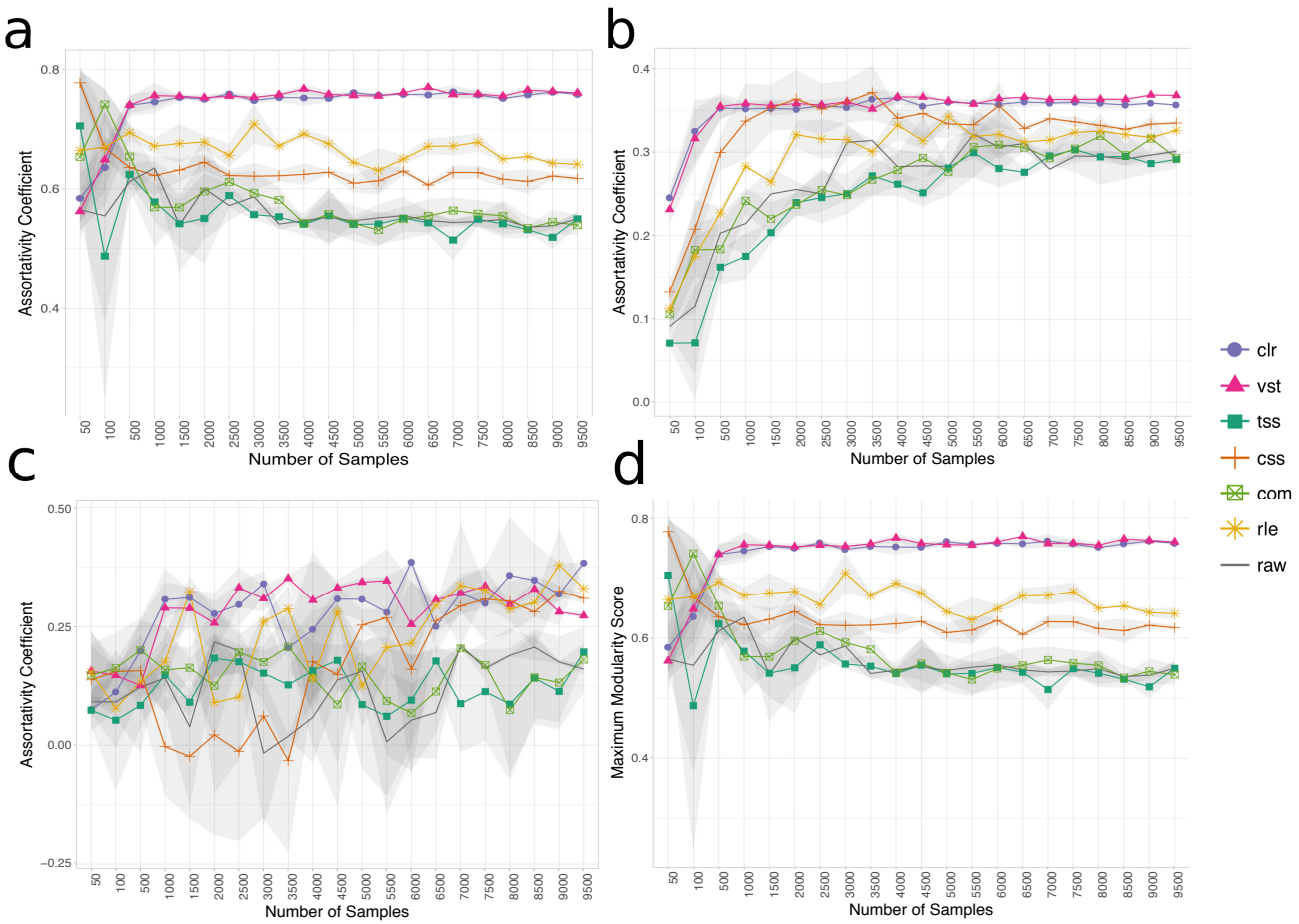

Figure 7: Relevance networks produced by CLR and VST transform contain more inter-family links independent of sample size at the top 2000 edges. A) Assortativity coefficient across sample size of family annotation. B) Assortativity coefficient across sample size of family annotation of only edges within modules. C) Assortativity coefficient across sample size of family annotation of only edges between modules. D) Maximum modularity score across sample size at 2000 edges. For all plots lines represent mean and grey ribbon represent standard deviation from the mean.

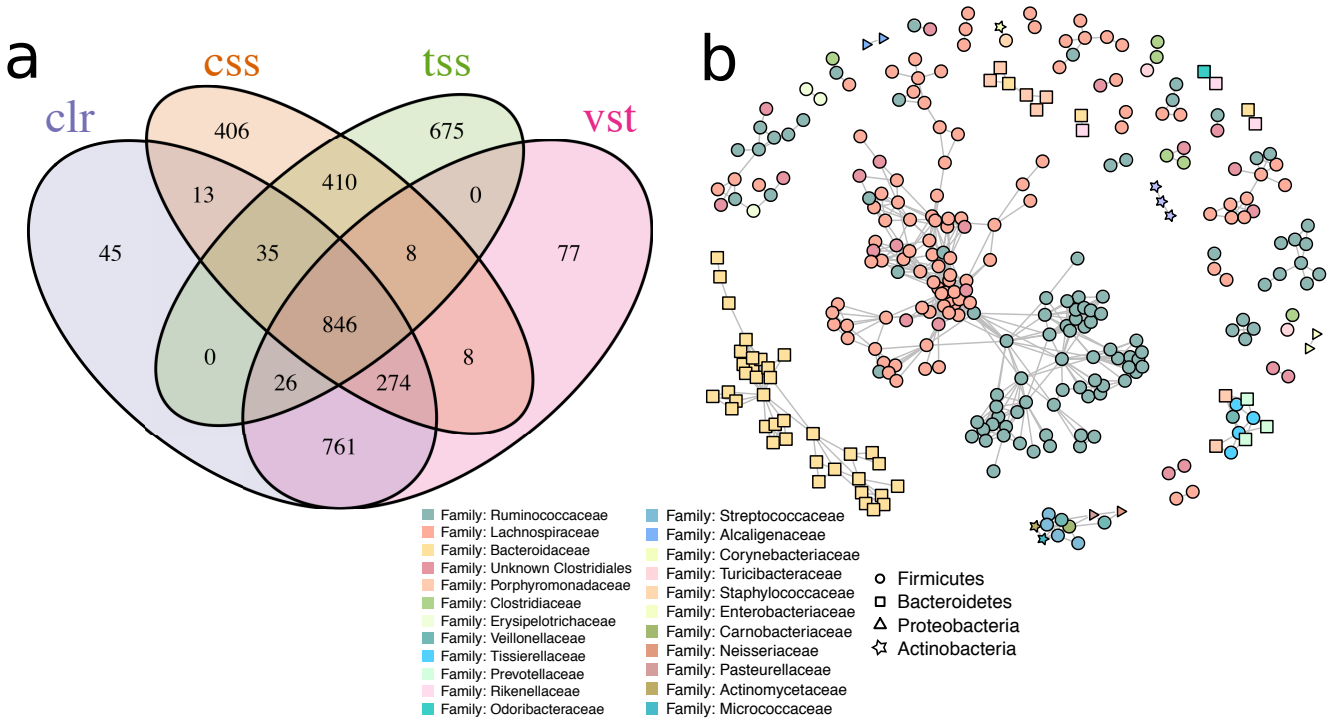

Figure 8: A) Venn Diagram of edges in common at top 2000 edges between four normalization methods. B) Network of edges in common between all methods. This consensus network contains 766 edges between 241 taxa. Node color represents Family annotation and node shape represents phylum. Network layout was generated using the force-directed Fruchterman-Reingold algorithm

When considering the top 2000 edges, normalization using CLR, CSS, VST and TSS agree upon 766 edges between 281 OTUs. The network contains several large sub-graphs representing mostly inter-family and inter-phyla interactions. The subgraphs themselves are also highly interconnected (Fig. 8B.). The methods with the most edges in common are CLR and VST. The normalization method with the most unique edges is TSS (Fig. 8A.). We have shown strikingly similar observations resulting from CLR and VST normalization. Despite being designed to behave like a log transformation at large counts, we show in this contribution the correlation derived from VST behaves similar to those of CLR across both sample and library size. When examining the per-OTU standard deviation (taken across all of our 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 Log2 transform (Supplement 5).

5 Discussion

In this work we have shown that methods of normalization that transform abundance data produce significantly different correlation structures, which in turn influence the outcome of nearly all downstream analysis. Evaluation of normalization methods based on stability of downstream analysis and tolerance to down-sampling suggests that VST and CLR are superior methods when dealing with compositional microbiome data. CLR and VST produce very similar, and more consistent estimates, of correlation than the other normalization methods examined in this work and the microbe-microbe interaction networks produced from CLR-normalized data were slightly more modular than VST. The current and widely used implementation of VST in the DESeq2 package contains statistical steps to handle repeated measures and data structures associated with microarray and RNA-seq data. The centered-log ratio, in terms of formulation as presented in this work with the addition of a pseudocount,is relatively easily applied to microbial abundance data. For this reason CLR has been applied to various pipelines for analysis of microbial abundance data and our work shows that its continued use is justified (when the added model complexity introduced by VST and by the much larger run times incurred by the VST normalization are considered). Correspondingly, we have shown that widely used methods like TSS and CSS (methods used in several prior studies) produce clusters and networks that perform relatively worse at grouping OTUs by taxonomy. This is not surprising as, mathematically, these methods do not address the compositional nature of microbial abundance data.

Future studies should consider additional normalization tools. For example, the isometric log ratio (ILR) is a family of compositionally-robust transforms, that has been widely adopted in a number of settings. Here we consider ILR beyond scope of this work, as it requires the selection of free parameters and should be guided by external information (uses additional data types not used by the other methods)[19, 26].

Here, we examined relevance networks by ranking interaction strengths. This allowed us to examine the effect of data normalization setups on a simple inference scheme. However, robust methods of inferring microbial associations from abundance data make use of correction for compositional effects and transitive correlations

in the face of limited sampling. For example, Bayesian models which incorporation prior information and penalized inverse covariance methods should be evaluated in future work [6, 9, 12, 24]. To reduce noise from sequencing and other technical artifacts when comparing normalization methods we only examined a filtered set of OTUs. Normalization in this context does not necessitate the selective removal of samples by library size, which is sometimes employed in the pre-processing of ecological data. While there is certainly merit to understanding how well proposed normalization methods remove technical bias, this endeavor is outside the scope of this contribution.

A common downstream analysis task is differential abundance analysis, i.e. to find the altered microbial components in response to some intervention. Recent contributions have already thoroughly examined the role of normalization on this task [18, 29], however integrating higher order correlations into this analysis is an outstanding objective. Our framework for examining the effect of normalization on correlation and downstream analyses can be adapted to examine the effect of simulated biologically-plausible correlated count data or experimentally validated community abundance data. In the future, the preference to measure absolute population of gene abundances may also alter the methods by which we normalize data, which could necessitate an entirely new scheme to normalize data.

References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., London, UK, UK, 1986. ISBN 0-412-28060-4.

[2] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-10-r106. URL `http://dx.doi.org/10.1186/gb-2010-11-10-r106`.

[3] V. Bucci, B. Tzen, N. Li, M. Simmons, T. Tanoue, E. Bogart, L. Deng, V. Yeliseyev, M. L. Delaney, Q. Liu, B. Olle, R. R. Stein, K. Honda, L. Bry, and G. K. Gerber. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biology*, 17(1):121, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0980-6.

[4] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. PeÃśa, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, May 2010. ISSN 1548-7105. doi: 10.1038/nmeth.f.303. URL `https://www.nature.com/articles/nmeth.f.303`.

[5] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), Dec. 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.70.066111. URL `http://arxiv.org/abs/cond-mat/0408187`. arXiv: cond-mat/0408187.

[6] J. C. Claussen, J. SkieceviÄĐienÄŮ, J. Wang, P. Rausch, T. H. Karlsen, W. Lieb, J. F. Baines, A. Franke, and M.-T. HÃijtt. Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLOS Computational Biology*, 13(6):e1005361, June 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005361. URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005361`.

[7] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

[8] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Computational Biology*, 8(7), July 2012. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1002606. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3395616/`.

[9] J. Friedman and E. J. Alm. Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology*, 8(9):e1002687, Sept. 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002687. URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002687`.

[10] W. FÃűrstner and B. Moonen. A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium*, pages 299–309. Springer, 2003.

[11] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8, 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.02224. URL `https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full`.

[12] Z. D. Kurtz, C. L. Muller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol*, 11, 2015. doi: 10.1371/journal.pcbi.1004226. URL `http://dx.doi.org/10.1371/journal.pcbi.1004226`.

[13] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL `http://dx.doi.org/10.1186/s13059-014-0550-8`.

[14] C. Lozupone, K. Faust, J. Raes, J. J. Faith, D. N. Frank, J. Zaneveld, J. I. Gordon, and R. Knight. Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome Research*, 22(10):1974–1984, Oct. 2012. ISSN 1549-5469. doi: 10.1101/gr.138198.112.

[15] U. v. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec. 2007. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-007-9033-z. URL `https://link.springer.com/article/10.1007/s11222-007-9033-z`.

[16] J. L. Mark Welch, Y. Hasegawa, N. P. McNulty, J. I. Gordon, and G. G. Borisy. Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *Proceedings of the National Academy of Sciences*, 114(43):E9105, Oct. 2017. doi: 10.1073/pnas.1711596114. URL `http://www.pnas.org/content/114/43/E9105.abstract`.

[17] D. McDonald, E. Hyde, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov, B. Behsaz, C. Brennan, Y. Chen, L. D. Goldasich, P. C. Dorrestein, R. R. Dunn, A. K. Fahimipour, J. Gaffney, J. A. Gilbert, G. Gogul, J. L. Green, P. Hugenholtz, G. Humphrey, C. Huttenhower, M. A. Jackson, S. Janssen, D. V. Jeste, L. Jiang, S. T. Kelley, D. Knights, T. Kosciolek, J. Ladau, J. Leach, C. Marotz, D. Meleshko, A. V. Melnik, J. L. Metcalf, H. Mohimani, E. Montassier, J. Navas-Molina, T. T. Nguyen, S. Peddada, P. Pevzner, K. S. Pollard, G. Rahnavard, A. Robbins-Pianka, N. Sangwan, J. Shorenstein, L. Smarr, S. J. Song, T. Spector, A. D. Swafford, V. G. Thackray, L. R. Thompson, A. Tripathi, Y. VÃązquez-Baeza, A. Vrbanac, P. Wischmeyer, E. Wolfe, Q. Zhu, T. A. G. Consortium, and R. Knight. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, 3(3):e00031–18, June 2018. ISSN 2379-5077. doi: 10.1128/mSystems.00031-18. URL `https://msystems.asm.org/content/3/3/e00031-18`.

[18] P. J. McMurdie and S. Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*, 10(4): e1003531, Apr. 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003531. URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531`.

[19] J. T. Morton, J. Sanders, R. A. Quinn, D. McDonald, A. Gonzalez, Y. VÃązquez-Baeza, J. A. Navas-Molina, S. J. Song, J. L. Metcalf, E. R. Hyde, M. Lladser, P. C. Dorrestein, and R. Knight. Balance Trees Reveal Microbial Niche Differentiation. *mSystems*, 2(1):e00162–16, Feb. 2017. ISSN 2379-5077. doi: 10.1128/mSystems.00162-16. URL `https://msystems.asm.org/content/2/1/e00162-16`.

[20] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2), Feb. 2003. ISSN 1063-651X, 1095-3787. doi: 10.1103/PhysRevE.67.026126. URL `http://arxiv.org/abs/cond-mat/0209450`. arXiv: cond-mat/0209450.

[21] A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 849–856, Cambridge, MA, USA, 2001. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2980539.2980649`.

[22] J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop. Robust methods for differential abundance analysis in marker gene surveys. *Nature methods*, 10(12):1200–1202, Dec. 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2658. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4010126/`.

[23] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23): 7537–7541, Dec. 2009. ISSN 1098-5336. doi: 10.1128/AEM.01541-09.

[24] E. Schwager, H. Mallick, S. Ventz, and C. Huttenhower. A Bayesian method for detecting pairwise associations in compositional data. *PLOS Computational Biology*, 13 (11):e1005852, Nov. 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005852. URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005852`.

[25] J. SchÃďfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:Article32, 2005. ISSN 1544-6115. doi: 10.2202/1544-6115.1175.

[26] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6, 2017. ISSN 2050-084X. doi: 10.7554/eLife.21887.

[27] F. StÃďmmler, J. GlÃďsner, A. Hiergeist, E. Holler, D. Weber, P. J. Oefner, A. Gessner, and R. Spang. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome*, 4, June 2016. ISSN 2049-2618. doi: 10.1186/s40168-016-0175-0. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4915049/.

[28] S. Weiss, W. V. Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, 10(7):1669–1681, July 2016. ISSN 1751-7370. doi: 10.1038/ismej.2015.235. URL https://www.nature.com/articles/ismej2015235.

[29] S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. VÃązquez-Baeza, A. Birmingham, E. R. Hyde, and R. Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5:27, Mar. 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0237-y. URL https://doi.org/10.1186/s40168-017-0237-y.

[30] I. Zwiener, B. Frisch, and H. Binder. Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. *PLOS ONE*, 9(1): e85150, Jan. 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0085150. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085150.

Supplementary Figures

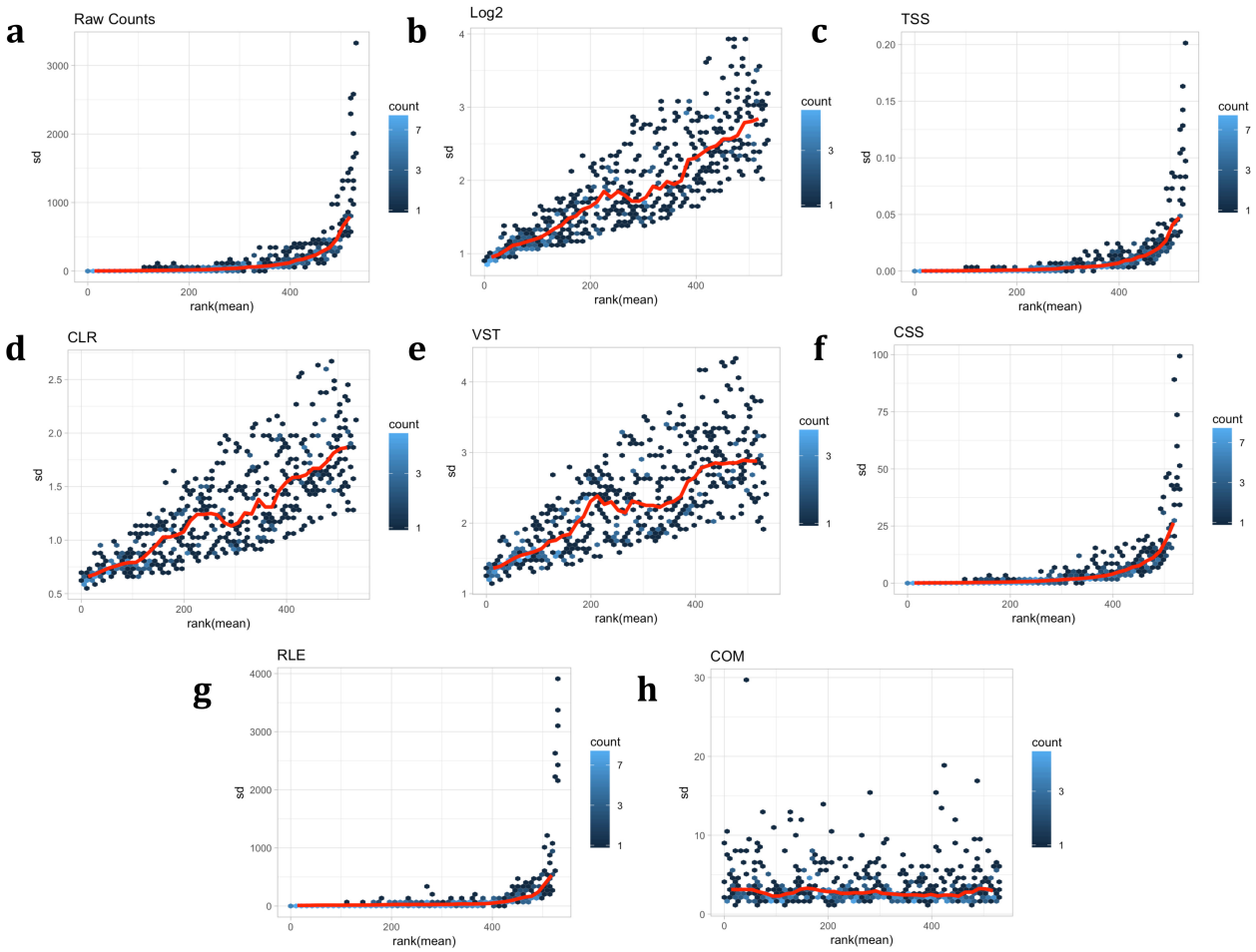

Figure 9: Supplement 1: Standard deviation per-OTU (taken across all 531 OTUs) plotted against the rank of the average count.

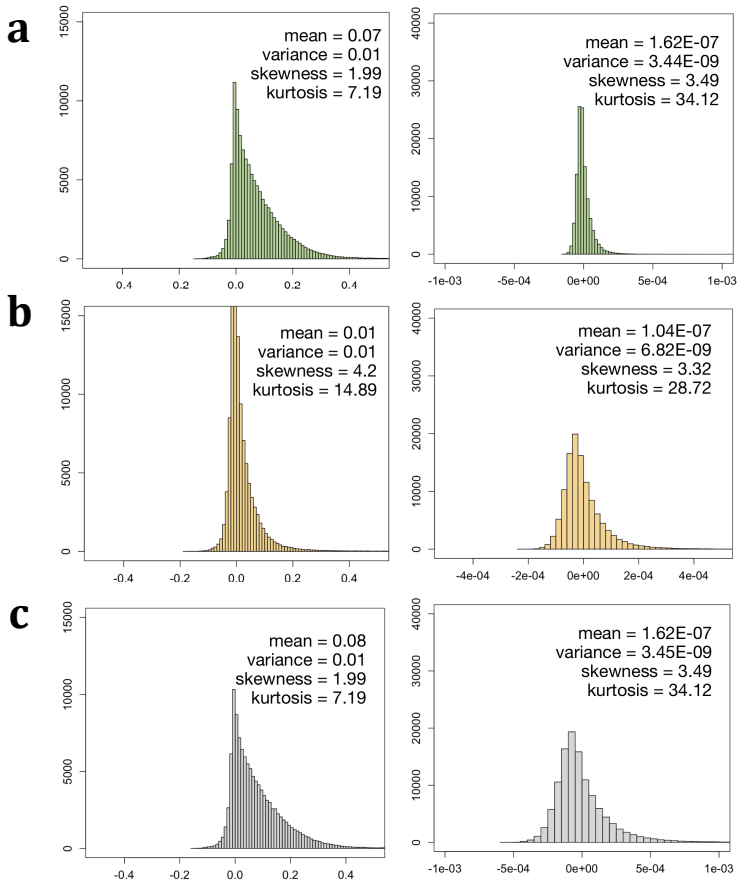

Figure 10: Supplement 2: Frequencies of correlation values reveal different underlying distributions after transformation. Histograms of correlation frequency of matrices after application of different normalization methods. Frequencies were plotted with 200 bins and mean, variance, skewness and kurtosis are shown for each distribution.

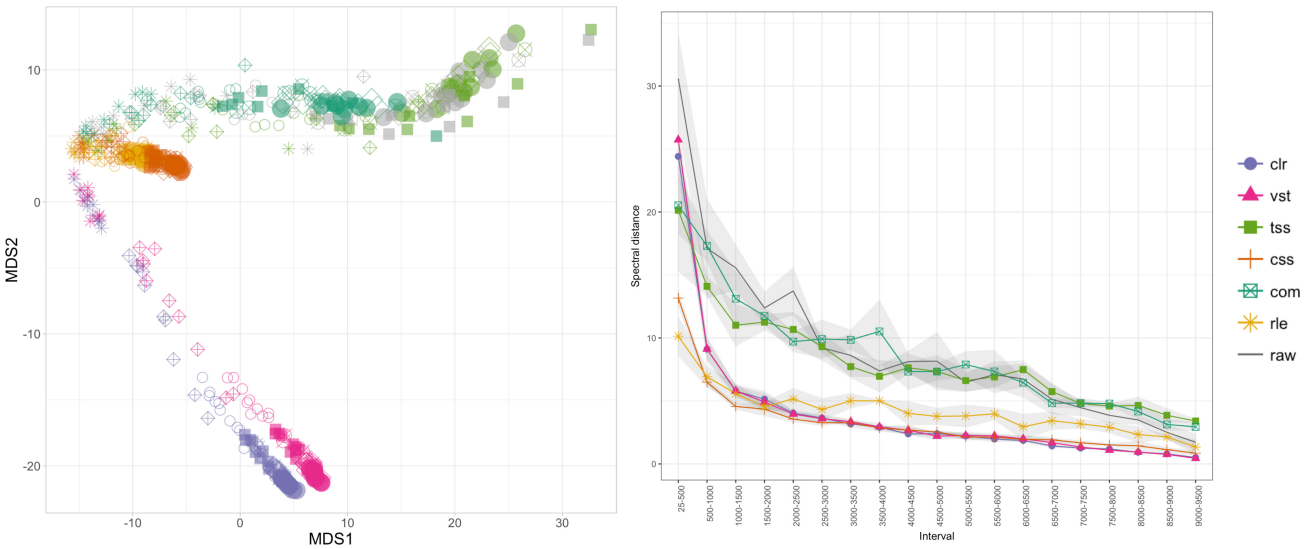

Figure 11: Supplement 3: Transformations that remove compositional artifacts (CLR) and stabilize the variance (VST) result in substantially different patterns of correlation. A,B) Multidimensional scaling representation of spectral and Förstner distances, respectively, between correlation structures of varying size estimated from different normalization methods. These estimates are also compared to untransformed or raw count values (dark grey points). C,D) Spectral distance and Förstner distance between sub-samples of different sizes. Lines represent mean and grey ribbon represent standard deviation from the mean of 20 random sub-samples. (color scheme as in A,B)

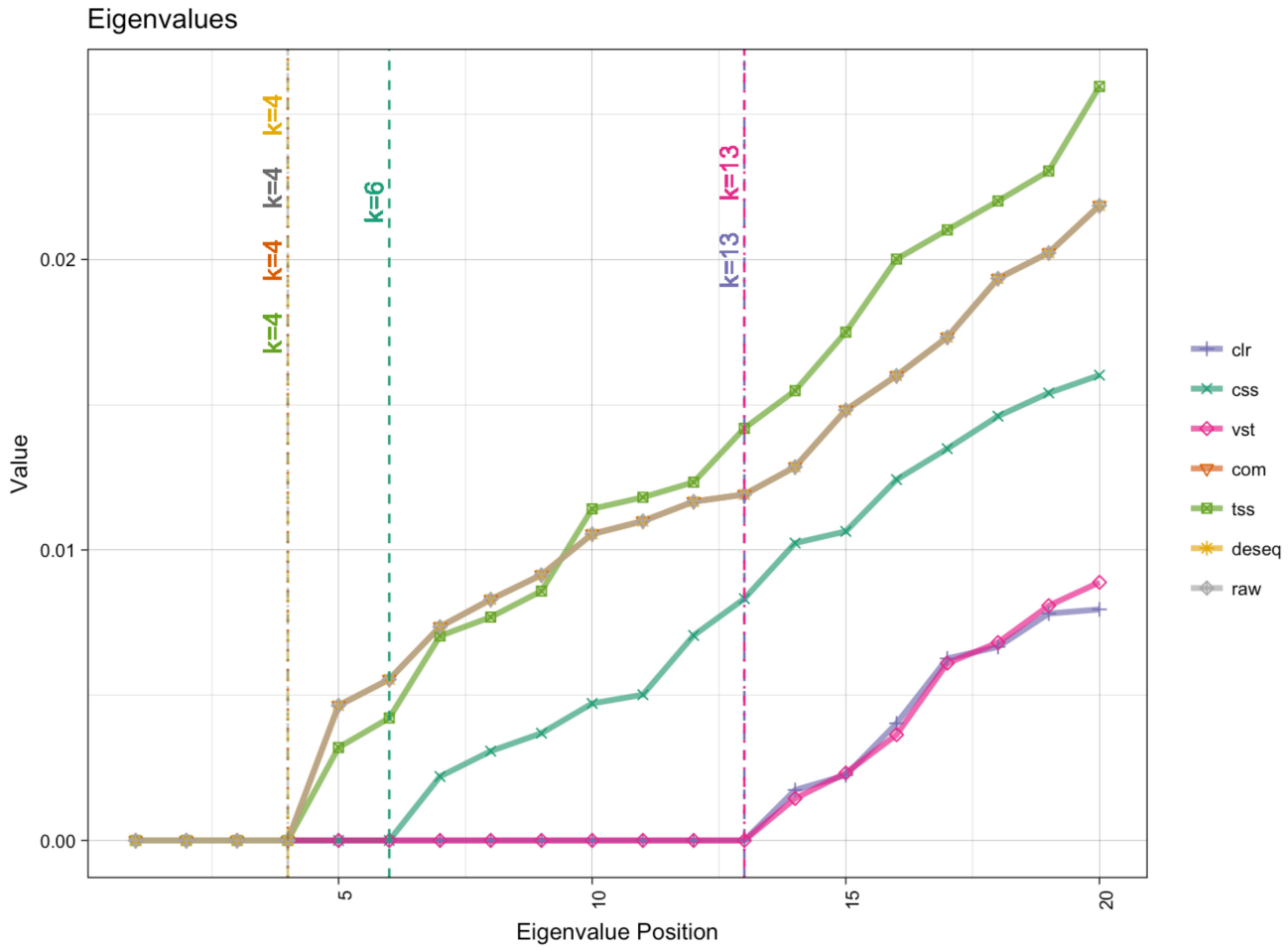

Figure 12: Supplement 4: CLR and VST produce higher numbers of clusters as selected by the eigengap metric. Vertical lines represent the first non-zero eigenvalue which is selected as the number of clusters.

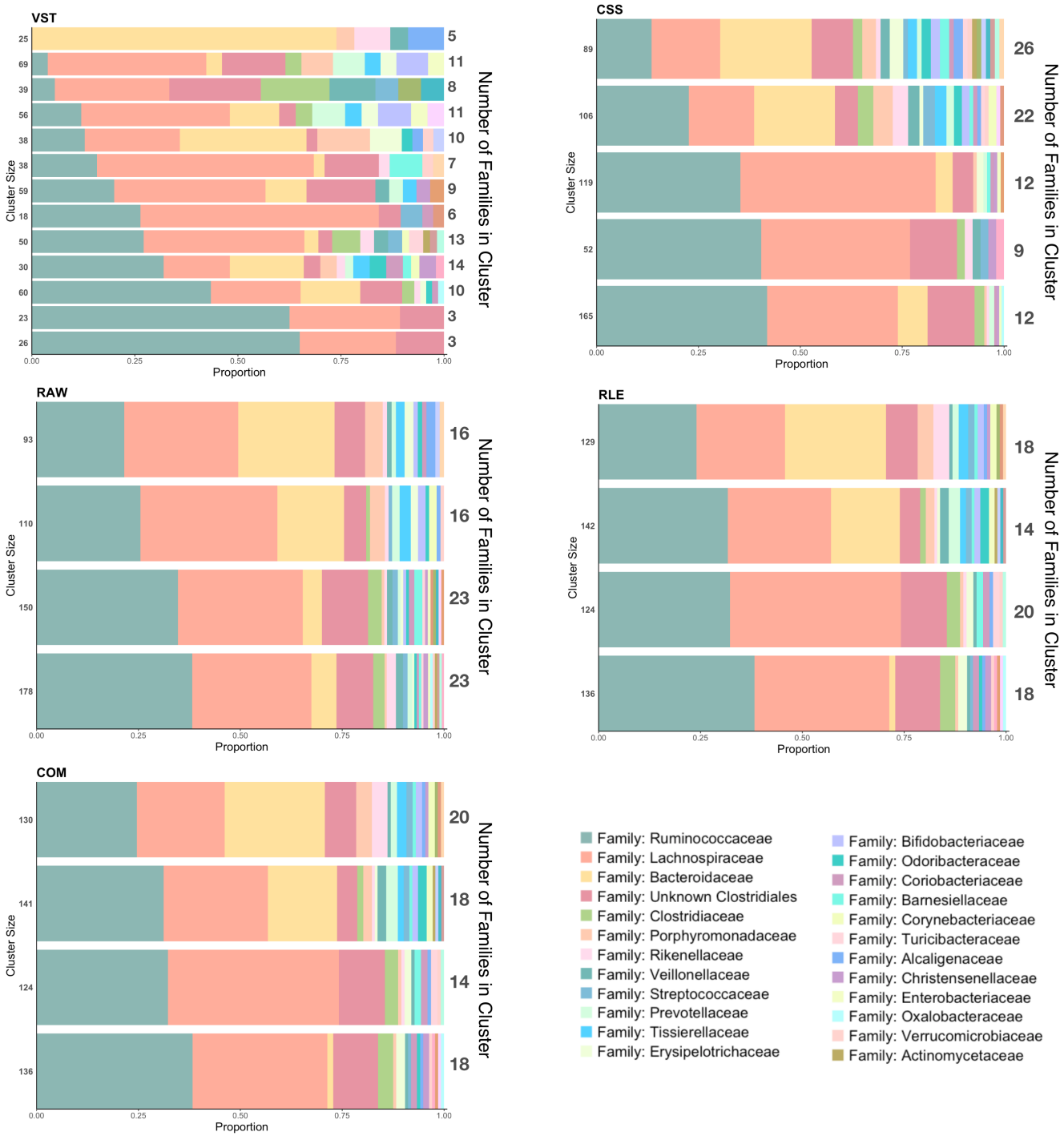

Figure 13: Supplement 5: Clusters of OTUs after CLR and VST transform produce more taxonomically homogeneous clusters than TSS and CSS (See Figure 4). The stacked bar plots represent the composition of OTUs in each cluster at the Family level. Clusters are vertically ordered from highest percentage of the most abundant Family: Rumincococcae. Horizontally the order represents the highest percentage in each cluster. Numbers on the left axis represent the number of OTUs in each cluster. Numbers on the right axis represent the number of taxonomic groups at the family level are in each cluster.

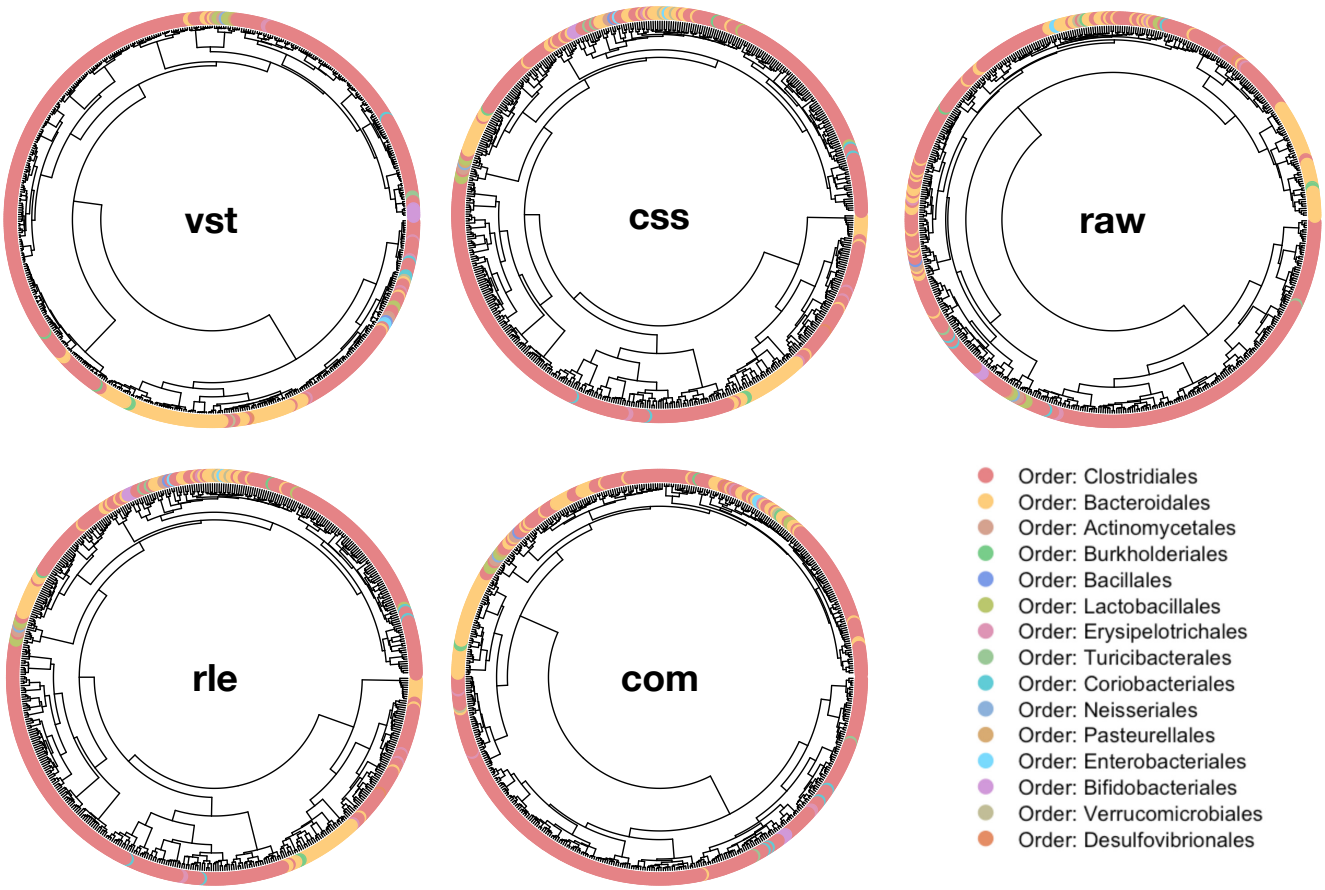

Figure 14: Supplement 6: Circular dendrograms reveal that hierarchical clustering results in more homogeneous clustering by taxonomic annotation. A larger number of taxa of orders Enterobacteriales and Bacteriodales cluster together in VST than in other normalization methods.

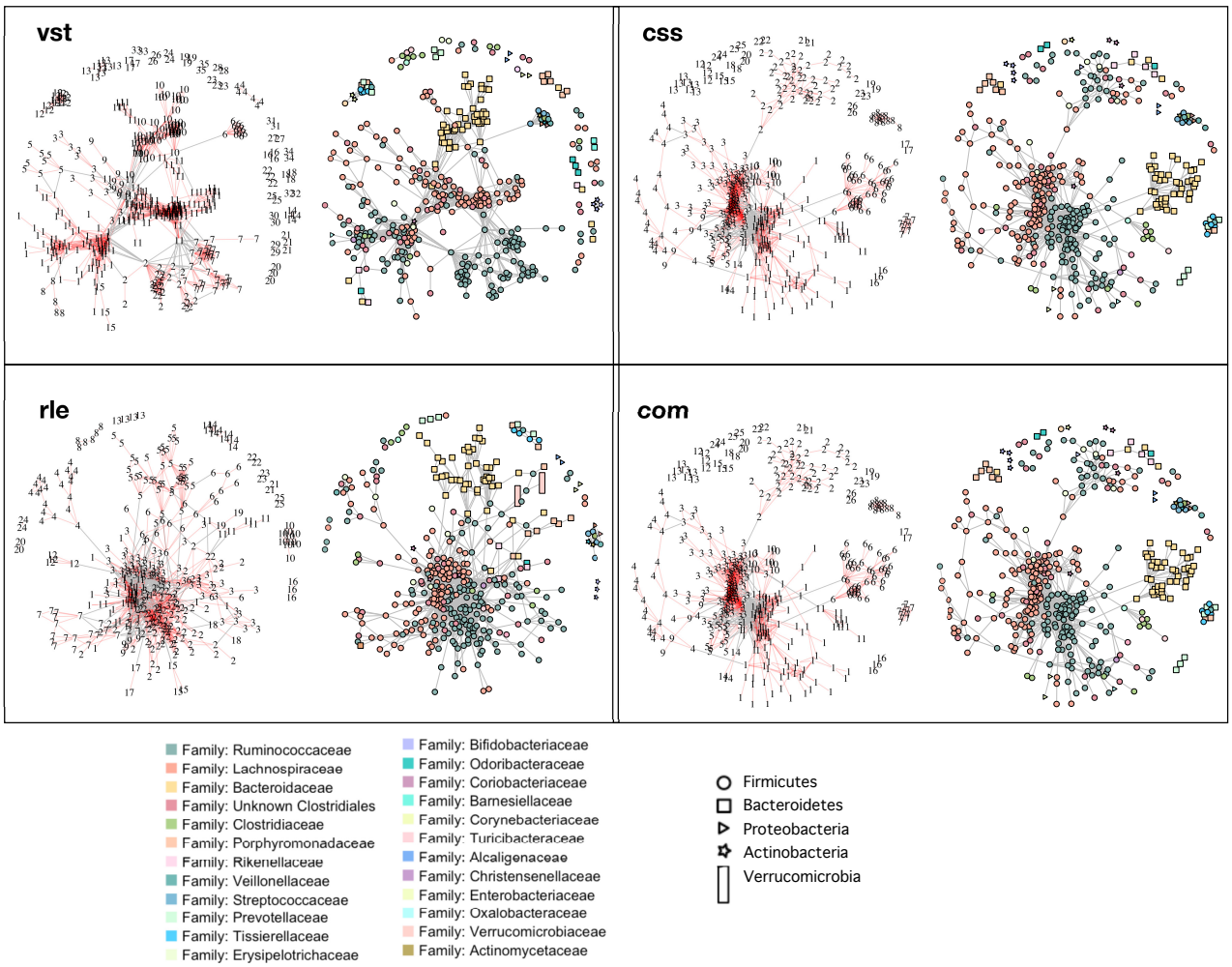

Figure 15: Supplement 7: CLR and VST produce more modular networks that separate communities at the Family level. For networks on the left of each panel every node represents an OTU labelled with module annotation as predicted by the Fast-Greedy modularity algorithm. The networks on the right represent the corresponding phylogenetic annotation of the OTU at the family level. Layout using the force-directed Fruchterman-Reingold algorithm was conserved for both networks in each panel for comparison