Correlation measures in metagenomic data: the blessing of dimensionality

Alessandro Fuschi[1☯], Alessandra Merlotti[1☯], Thi Dong Binh Tran[2], Hoan Nguyen[2], George M. Weinstock[2,3,†], Daniel Remondini[1*]

1 Department of Physics and Astronomy, University of Bologna, Bologna 40127, IT.
2 The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030 USA
3 Dept. Genetics and Genome Science, University of Connecticut Health Center, Farmington, CT 06032 USA

☯These authors contributed equally to this work.
†Deceased
* daniel.remondini@unibo.it

Abstract

Microbiome analysis has revolutionized our understanding of various biological processes, spanning human health, epidemiology (including antimicrobial resistance and horizontal gene transfer), as well as environmental and agricultural studies. At the heart of microbiome analysis lies the characterization of microbial communities through the quantification of microbial taxa and their dynamics. In the study of bacterial abundances, it is becoming more relevant to consider their relationship, to embed these data in the framework of network theory, allowing characterization of features like node relevance, pathway and community structure. In this study, we address the primary biases encountered in reconstructing networks through correlation measures, particularly in light of the compositional nature of the data, within-sample diversity, and the presence of a high number of unobserved species. These factors can lead to inaccurate correlation estimates. To tackle these challenges, we employ simulated data to demonstrate how many of these issues can be mitigated by applying typical transformations designed for compositional data. These transformations enable the use of straightforward measures like Pearson's correlation to correctly identify positive and negative relationships among relative abundances, especially in high-dimensional data, without having any need for further corrections. However, some challenges persist, such as addressing data sparsity, as neglecting this aspect can result in an underestimation of negative correlations.

Introduction

Techniques based on next-generation sequencing (NGS) can elucidate the complex functioning of natural microbial communities directly in their natural environment. New branches of research have been created such as the study of the human microbiota which showed heterogeneity between different anatomical sites and individual variability [1, 2], or the ability to characterize and monitor the presence of antimicrobial resistance worldwide [3]. Complementing the analyses conducted directly on the abundance of microbiota samples, it can be greatly beneficial to explore a second layer of information represented by the relationships among the observed species. Network

theory provides many essential tools to characterize collective properties of the ecology of a natural environment by defining central elements or communities in the system and allowing visualization of these results by exploiting network structural properties [4]. Consequently, the initial step in reconstructing any network involves the identification and quantification of relationships between species, often achieved by assessing correlations or conditional dependencies among each pairwise combination of variables. Independent from the NGS technique used like RNA-seq, 16s or whole genome shotgun, the underlying data are similar, composed of counts of sequencing reads mapped to a large number of references (taxa) and the unifying theoretical framework is their compositional nature [5, 6]. Taxa abundance is determined by the number of read counts, which is affected by sequencing depth and varies from sample to sample. Typically a sum constraint is imposed over all the samples (1 for probability, 100 for percentage or 10^6 for part per million) called $L1$ normalization, to remove the effect of sample depth. In this way, data are described as proportions and referred to as compositional data [7, 8]. However, as noted by Pearson at the end of 19th century [9], compositional data can generate spurious correlations between measurements. From a mathematical point of view the data lie on a simplex [8], thus it can be extremely dangerous to use Euclidean metrics for proximity and correlation estimations. These biases on correlation between relative abundances can be significant in some datasets but mild in others [fig:1], and the *diversity* within each sample, called α-diversity, (referred to as $\bar{P}$, see Materials and Methods) concurs to enforce this bias [10]. Correlation biases become more pronounced when counts are concentrated in a few taxa. Conversely, when counts are distributed more evenly across samples, these biases tend to decrease. Hence, it is imperative to take into account these compositional effects when reconstructing networks from metagenomic data. Failing to do so may lead to entirely incorrect conclusions [11], endangering the accuracy and reliability of inferred ecological interactions.

To improve correlation estimates on relative abundances, methods such as Sparse Correlations for Compositional data (SparCC) [10], Sparse and Compositionally Robust Inference of Microbial Ecological Networks (SPIEC-EASI) [12], Proportionality for Compositional data (Rho) [13] and many others [14–29] have been developed, almost all making extensive use of the compositional theory introduced by Aitchison [8]. Aitchison provided a family of transformations to handle this type of data, known as log-ratio transformations. The counts of each sample are expressed relative to a reference to enable comparisons, followed by the application of logarithm. One common choice is the centered log-ratio transformation (CLR), where each element is divided by the geometric mean of the sample in a logarithmic scale. This operation is both isomorphic and isometric, preserving distances. However, like L1 normalization, CLR also introduces a sum constraint where the sample sum is fixed to 0. This constraint is equivalent to mapping the counts on a Cartesian hyperplane instead of a simplex, and it also introduces spurious dependencies between variables.

Our work shows that, unlike L1 normalization, the bias introduced by the sum constraint in CLR strongly depends on the dataset dimensionality D, or more explicitly it is related to the number of taxa or references [fig:1]. In our study, we not only demonstrate but also quantify these biases, which diminish as the dimensionality increases. In metagenomic contexts, where dimensionality can extend to hundreds or more, the impact of spurious correlations introduced by CLR becomes negligible, making any subsequent step for correlation estimation less critical.

Furthermore, there are additional typical sources of error in the estimation of correlations in metagenomic datasets. Often a large part of taxa in the NGS experiments are under the detection limits of the sequencing techniques, producing very sparse abundance matrices. It's really common to find datasets where more than

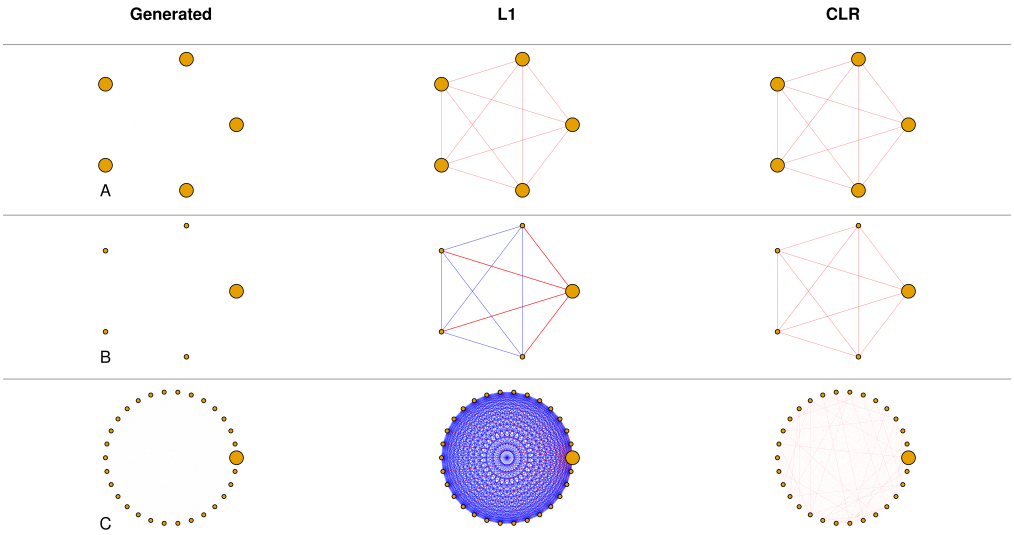

Fig 1. Impact of L1 and CLR Normalizations on Correlation Estimates. Three different cases (A, B, C) are shown, with data generated from uncorrelated multivariate standardized normal distributions sampled 10,000 times, in which data were shifted in order to be positive. Left figures describe the generated data with fixed number of species (dimensionality D) and node size proportional to the mean species abundance (α-diversity $\bar{P}$); central figures represent Pearson's correlation as links (red, negative; blue, positive) with width proportional to its value, after L1 data normalization; right figures represent the same situation after CLR data transform. The parameters for the presented cases were: A) $D = 5$ and $\bar{P} \approx 1$, B) $D = 5$ and $\bar{P} \approx 0.5$, C) $D = 30$ and $\bar{P} \approx 0.5$. In L1 normalization, biases are strongly associated with dataset diversity and do not decrease with dimensionality, while for CLR normalization these biases decrease with increasing dimensionality and are independent of diversity (see Results Section).

$70 - 80\%$ of species are undetected and typically it is assigned the value of 0. The unobserved species are not to be interpreted as the absence of that species but rather as a missing value in which we have no further information. Moreover, non-zero counts exhibit strongly non-normal distributions in non transformed data, with heavy tails that invalidate the assumptions of Pearson's correlation. The distribution that better describes the real NGS data is still a debated discussion, but in different context the zero-inflated negative binomial distribution (ZINB) is employed [12, 30]. The ZINB distribution can effectively capture the excess of zeros and the dispersion in the data, making it a suitable choice for representing counts in metagenomic datasets, particularly given its discrete nature similar to the counts.

The aim of this manuscript is to explore biases affecting correlation estimates, particularly in the context of compositionality and zero-excess issues commonly encountered in metagenomic datasets. In the absence of a ground truth, we create synthetic datasets across a wide range of conditions, varying dimensionality, diversity, data distribution and sparsity to characterize the biases in correlation estimation. To achieve this, we have developed a model focused on the 'Normal to Anything' approach that allows the generation of random variables with arbitrary marginal distributions starting from multivariate normal variables with desired correlation structure.

This work is structured to address three main considerations. The first is the examination of the biases introduced by L1 and CLR transformations in relation to dimensionality and within diversity. This involves a thorough analysis of how these transformations impact data interpretation across various compositional contexts. Importantly, we acknowledge that while CLR is extensively used in metagenomics as a crucial analytical tool, its application is often not accompanied by a deep understanding of its limitations and advantages.

The second consideration corroborates our findings regarding compositional biases arising from L1 and CLR transformations. For this, we compare various recently developed methods on real metagenomic data with the simplest approach of using Pearson correlation on CLR transformed abundances (Pearson+CLR). Our analysis reveals an almost complete overlap in the final results, emphasizing the significance of the CLR transformation.

The third aspect of our research evaluates the role of zero measurements in estimating correlation after minimizing compositional biases through optimal transformation. This involves assessing how zero counts affect the accuracy of correlation measures, thereby providing insights into the appropriate handling of sparse data in metagenomic studies.

Results

Compositional biases become negligible with high dimensionality

To comprehend and quantify the compositional biases inherent in Pearson correlation, we conducted a comprehensive comparative analysis. We compared the known correlation structure initially provided as input to the model with the correlation structures obtained after applying L1 and CLR normalizations, while systematically varying the dimensionality D and the within dataset diversity $\bar{P}$ (see Materials and Methods section). In total, we generated 1560 distinct datasets adjusting the dimensionality, ranging from 5 to 200 in steps of 5, and manipulating the within dataset diversity from $0,025$ to $0,975$ in increments of $0,025$, with a tolerance of $\pm 0,005$. To isolate the effects of the L1 and CLR transformations, we made deliberate efforts to minimize any known sources of error and chose the simplest experimental conditions to ensure the robustness of our findings. In line with these principles, we consistently

conducted the analysis with an uncorrelated covariance structure, and we chose to work with normally distributed variables to avoid potential errors in the Pearson correlations that may result from non-normally distributed data. Furthermore we choose for each experiment $N = 10000$ samples, in order to minimize possible random correlation between variables. Finally, we quantified the biases by calculating the mean absolute error (MAE) on all values of the matrix obtained by subtracting L1- and CLR-normalized correlation matrices, denoted as R^{L1} and R^{CLR}, to the original correlation matrix R, as follow:

$$\mathrm{MAE}_{D,\bar{P}}(K) = \frac{\sum_{i=1}^{D} |R_i^K - R_i|}{D^2} \quad K = \text{L1 or CLR} \tag{1}$$

MAE values range within the interval $[0, 2]$, where 0 implies a perfect accordance with the ideal correlation and 2 represents maximum distortion.

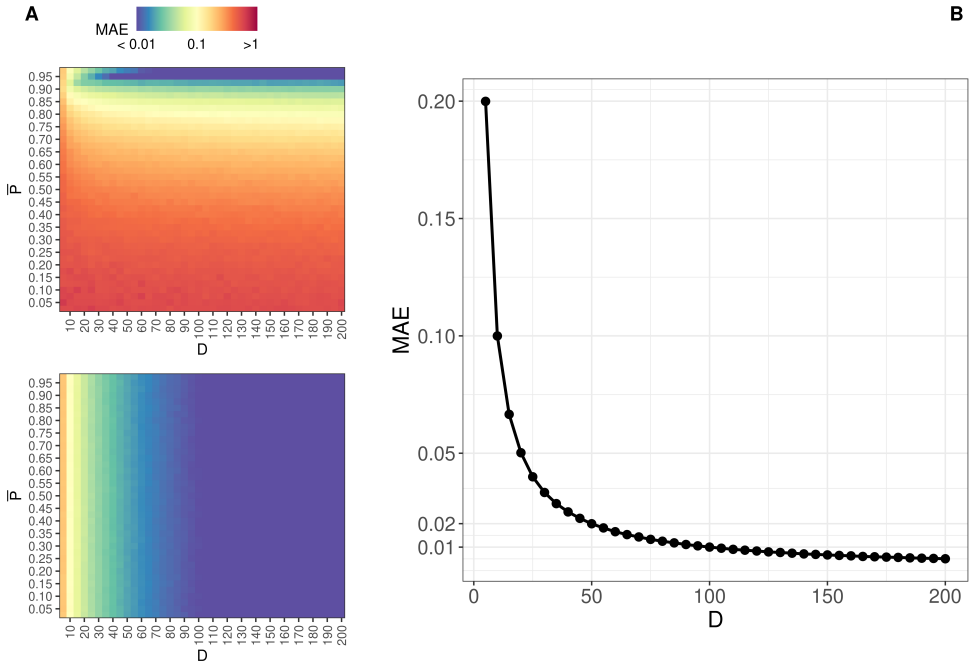

Fig 2. Different behavior of spurious correlations for L1 and CLR transforms : **A)** heatmap of MAE as a function of the dataset diversity $\bar{P}$ and dimensionality D for L1 normalization (top) and CLR normalization (bottom) in $\log_{10}$ scale. **B)** Scatter plot illustrating the MAE for CLR normalization on correlation as a function of dimensionality.

The distinct behaviors of the two normalizations are evident, as they introduce different biases on correlation (see Figure 2). Specifically, L1 correlations are primarily influenced by within dataset diversity, with the biases becoming more pronounced as the values within a sample become more heterogeneously distributed. On the other hand, CLR data exhibit biases that are independent of dataset diversity, and these distortions diminish rapidly with increasing dimensionality. Building upon the premise of complete independence of the CLR biases on correlation from dataset diversity, we can estimate this effect by calculating an average over all diversity values $\bar{P}$. We observe that the error decreases to less than 0.01 for dimensionality values greater than or equal to 100. Thus, we posit that in typical metagenomic scenarios, where the dimensionality often extends into the hundreds, the effects of compositionality are negligible.

Comparison between state of art methods

We opted to compare various computational approaches designed for inferring relationships within compositional datasets, contrasting these with the more straightforward Pearson correlation method applied to CLR-transformed data (Pearson+CLR). This comparison included evaluations of SparCC, the proportionality metric (ρ), and the SPIEC-EASI framework utilizing one of its two primary algorithms. All these tools employ similar and comparable concepts, even if developed with different methodologies. For instance, SparCC aims to approximate the Pearson correlation by assuming that the true underlying correlation network is sparse, meaning that highly correlated variables are relatively few compared to the total number. In contrast, ρ is based on the similar concept of proportionality as an alternative to traditional correlations, with the goal of mitigating compositional biases.
SPIEC-EASI employs graphical model inference to discern the conditional independence among variables, enhancing its efficacy through iterative evaluations across multiple dataset subsampling. Within SPIEC-EASI exist two inference schemes, we selected the graphical lasso (GLASSO) algorithm for its conceptual alignment with correlation analysis, as it similarly hinges on the covariance structure among the variables. The alternative, the Meinshausen-Bühlmann (MB) algorithm, departs from the correlation-based framework, instead drawing on principles of linear regression for inferring relationships.
All aforementioned methods make extensive use of the compositional theory starting their routine by normalizing data via a log-ratio transformation consistent with Aitchison's philosophy.
The comparison, conducted on the 51 samples of subject 69-001 in healthy condition from HMP2 (see Materials and Methods) shows an almost complete overlap of the final results, as in Fig. 3. The comparison between the Pearson+CLR with SparCC and Rho is direct, since these three methods produce values between -1 and 1: the scatter plot of the respective correlation values is ≈ 0.99 for both, in very good accordance to a $y = x$ linear relationship (Fig.3: 1A-1B).

Since SPIEC-EASI produces a binary output in terms of conditional independence between each pair of variables, we consider the histogram of Pearson+CLR values, and overlap bins corresponding to couples of variables significantly associated through SPIEC-EASI by imposing a threshold on overall stability equal to 5%. Most of the significant links for SPIEC-EASI are associated to high absolute values of Pearson+CLR (Fig.3: 1C). This analysis shows that significant SPIEC-EASI associations predominantly correspond to high absolute Pearson+CLR values.

Further comparison were performed between the networks inferred by the SPIEC-EASI GLASSO and Pearson+CLR through thresholding, considering as links the correlations with $p < 0.05$ after Bonferroni correction for multiple testing. Approximately 78% of the edges were common between the inferred networks (Fig 3: 2C), and a visual inspection of the network representations indicates that their collective properties are nearly identical (Fig 3:2A-2B).
In practical applications, despite the heterogeneity of their underlying methodologies, the considered methods converge towards equivalent outcomes. This observation underscores the central role of the Centered Log-Ratio in all considered algorithms, that is sufficient to minimize spurious correlations within high-dimensional contexts. Particularly in metagenomic studies, where the dimensionality often extends into the hundreds, the necessity for additional corrective measures appears redundant.

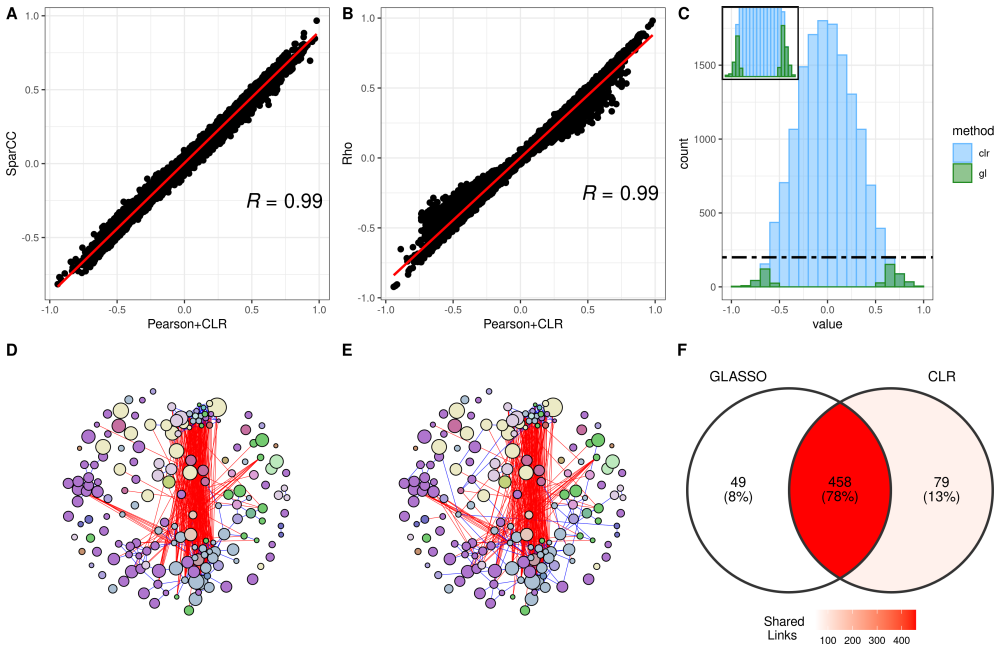

Fig 3. Comparison between state-of-the-art methods and Pearson+CLR on HMP2 data: **A-B)** Scatterplot of the weights associated with pairwise relationships between between the OTUs of Sparcc and ρ compared to Pearson+CLR. **C)** Significant links obtained through the SPIEC-GLASSO mapped to the Pearson+CLR correlation histogram. **D)** Reconstructed network from Pearson+CLR method with Bonferroni-corrected p-value threshold set at 0.05; **E)** Reconstructed network using the SPIEC-EASI GLASSO method with a stability threshold set to 0.05; we used the same vertices layout as for Pearson+CLR network. **F)** Venn diagram of the shared links between SPIEC-EASI GLASSO and Pearson+CLR networks.

Data sparsity remains a limitation

In this section we focus on the error on estimating correlation as a function of the ratio of zero values in the samples, similar to real-world scenarios. To achieve this, we have implemented a zero-inflated negative binomial distribution as the target distribution within our modeling framework based on NorTA approach (see Materials and Methods section). This distribution was selected to accurately capture the frequent occurrence of zero counts and the asymmetrical distributions seen in real data.

In the preceding section, we discussed measures taken to minimize the impact of spurious correlations introduced by the CLR transformation. To achieve this, we standardized the dimensionality (D) of all generated datasets to 200, a choice informed by its effectiveness in ensuring that correlation errors remain consistently below the threshold of 0.01. Even in this analysis we fixed the number of observations (N) to 10^4 to reduce errors within the estimated correlation matrix. Furthermore, we only took the CLR into consideration for the analysis given that the L1 in real situations, with more heterogeneously distributed data, is impractical as seen in the previous section.
We generate data that closely resemble real-world observations deriving the parameters *munb*, *size* and ϕ of the zero-inflated negative binomial distribution from the actual distributions of the OTUs of subject 69-001 in the HMP2 dataset, using the *fitdist* function from the R package *SpiecEasi* [31]. Each taxon was then generated using random parameters falling within the range of the first and ninth deciles of the

previously fitted ZINB parameters, distributed according to their empirical distribution using the *quantile* function of base R.

To quantify the error, we consider the absolute difference between the initial data correlation matrix R and the correlation on the same data transformed through NorTA approach and CLR, R_{CLR} with nonzero correlation only between two taxa labeled I and J. We build the correlation matrix specifically by varying only the value between I and J, labelled as r, from -0.9 to 0.9 in steps of 0.05, leaving all the others 198 taxa uncorrelated. In practice, all the other taxa other than I,J only contribute to reduce the biases introduced by the CLR transformation. Moreover, we varied the ratio of zero counts (ϕ_I and ϕ_J) of their respective marginal distributions from 0 to 0.95 in increments of 0.025. This process enables us to track the correlation error between taxa I and J across different levels of sparsity and correlation ($\text{err}_{\phi,r}$).

This process was repeated 100 times for every combination of ϕ and r, and MAE was calculated as follows (see Fig. 4.A):

$$\text{MAE}_{r,\phi} = \frac{\sum_{i=1}^{100} |R^i(r,\phi) - R^i_{CLR}(r,\phi)|}{100} \tag{2}$$

An important aspect to emphasize in our methodology is the deliberate decision to randomly generate parameters for each ZINB distribution. This approach was intended to observe the correlation phenomenon in a manner that is as independent as possible from any specific data distribution, ensuring that our findings are not biased by particular distributional characteristics of the data. The pseudo-code below summarizes our methodology:

```
// Fit ZINB Model parameters using OTUs from HMP2
params_ZINB_HMP2=fitZINBParameters(OTUs_HMP2);

// Perform 100 iterations of simulation
for (iteration in 1:100) {

    // generate ZINB random parameters using the ecdf of the real
    // distributions of the ZINB parameters
    random_params_ZINB=randomZINBParameters(params_ZINB_HMP2);

    // Generate synthetic dataset with D=200
    syntheticData = generateSyntheticDataset(D=200, par=random_params_ZINB);

    // Loop over varying levels of sparsity (phi) and correlation (r)
    for (phi in seq(0, 0.95, by=0.025)) {
        for (r in seq(-0.9, 0.9, by=0.05)) {

            // Modify variables I and J in the dataset
            modifyVariables(syntheticData, I, J, phi, r);

            // Record error for current sparsity and correlation
            err_phi_r = recordError(syntheticData, I, J);
        }
    }
}

// Calculate the Mean Absolute Error (MAE) for each phi and r
// over the 100 iterations
```

```
calculateMAE(err_phi_r);
```

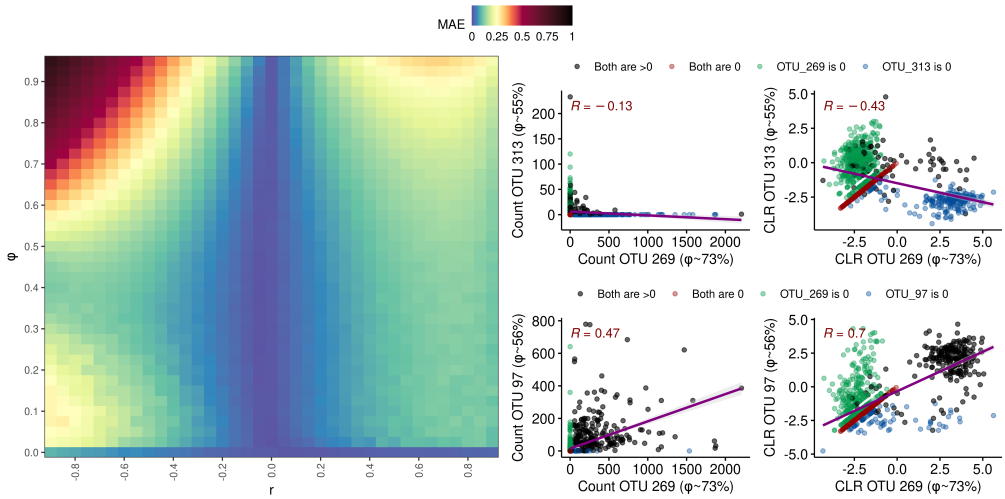

Fig 4. Impact of Sparsity on Correlation Coefficients in CLR-Transformed Data: **A)** heatmap depicting MAE of correlation coefficients across different values of sparsity Φ and correlation values r. **B)** effect of CLR transform on negative (top subplots) and positive (bottom subplots) correlation between selected pair of OTUs from HMP2 dataset, with a significant presence of zero counts, before (left subplots) and after (right subplots) CLR transform.

MAE significantly differs between positive and negative correlations, as clearly illustrated in (Fig. 4-A). While the error generally grows with an increasing number of zeros, this effect is particularly marked for taxa with negative correlations, as observed in the upper left section of the figure. Additionally, it is noteworthy that when variables are uncorrelated, the presence of zeros does not significantly impact the results.
An important aspect in the application of the CLR transformation is the number of zero counts, that requires the introduction of pseudo-counts to avoid logarithm divergence. This is illustrated in Fig. 4-B, using data from the HMP2 dataset, where we consider two pairs of OTUs with a high percentage of zeros and opposite sign of the correlation values. When examining negatively correlated variables in metagenomic studies, most of the nonzero values of one variable are matched with the pseudo-counts of the other. Such a pattern leads to a flattening on the x, y axes of the two OTU scatterplot, producing a hyperbolic-like pattern (Fig. 4-B top left) that tends to underestimate the value of negative correlation. We show that CLR significantly increases the negative correlation value mitigating this phenomenon, also in case of positively correlated OTUs (Fig. 4.B bottom).

Discussion

The network analysis framework is a robust tool for enhancing our comprehension of metagenomic studies, enabling us to unravel the intricate dynamics of microbial ecosystems. Although network reconstruction from second-order statistics such as correlation offers a straightforward methodology, the compositional nature of metagenomic data presents unique analytical challenges that require specialized techniques. Our study conducts a detailed investigation into the potential biases that affect the accuracy of correlation measures, considering factors such as dimensionality,

diversity, and sparsity of datasets,characteristics commonly associated with metagenomics data of any type.

Our analysis is focused on the effect of the Centered Log-Ratio (CLR) transformation when applied to compositional data. We discovered that the spurious correlations introduced by the CLR transformation decrease as a function of sample dimensionality. This contrasts with the L1 transformation, where spurious correlations are mainly influenced by the within-diversity of the dataset and do not decrease with sample dimensionality. Given the high dimensionality that characterizes metagenomic datasets— in the order of hundreds or more OTUs or taxa—the spurious correlations associated with CLR become thus negligible. The CLR transformation is also adequate to rectify the effect of diversity for sufficiently high-dimensionality data (in the order of hundreds) without additional adjustments, at difference with L1 transform for which high diversity remains an issue. We underscore the pivotal importance of the CLR transformation as a foundational step for metagenomic studies, streamlining the processing steps while ensuring data integrity.

To validate the role of the CLR transformation in compositional data analysis, we conducted a comparative study using various state-of-the-art algorithms specifically designed to estimate associations in metagenomic datasets. Our findings indicate a striking convergence of SparCC, ρ, and SPIEC-EASI GLASSO methods for correlation estimation towards Pearson's correlation on CLR-transformed data. This convergence suggests that the log-ratio transformation is the critical normalizing step across all methods, effectively neutralizing the compositional bias inherent to the data.

However, we must also acknowledge the substantial impact of dataset sparsity on correlation measures: the large number of zero counts associated with low-abundance taxa can significantly distort correlations, more severely affecting negative correlations. While CLR mitigates this distortions, the proportion of zero counts is the crucial parameter: the larger the zero count ratio, the larger the distortion. It is thus impossible to entirely eliminate the bias introduced by zero counts, unless eliminating any information about very rare species. A compromise must thus be found between minimizing correlation distortions and retaining low-abundance species in the analysis. This trade-off is fundamental for ensuring the accuracy and comprehensiveness of metagenomic data interpretation as a function of the study design.

Materials and Methods

Within Dataset Diversity $\bar{P}$

The within diversity of a dataset $\bar{P}$ is defined as the mean value over all the samples of the Pielou index [32], which is the Shannon entropy normalized to 1 with respect to the dimension. Given a dataset $\mathbf{X} \in \mathcal{N}^{N,D}$ composed of N distinct samples $\vec{x}$ of dimension D:

$$P(\vec{x}) = \frac{H(\vec{x})}{\log(D)} = \frac{-\sum_{i=1}^{D} p_i \log(p_i)}{\log(D)} \tag{3}$$

with $H(\vec{x})$ the Shannon entropy, p_i corresponding to the i-th taxa relative abundance in the sample. Finally, the diversity of a dataset $\bar{P}$ is calculated as:

$$\bar{P}(\mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{4}$$

Generation of Gaussian Data for Characterization of L1 and CLR Correlation Biases

Our examination of the biases introduced by L1 and CLR transformations began with the creation of synthetic datasets modeled on Gaussian distributions. This methodology was specifically crafted to underscore the compositional biases inherent in metagenomic datasets, with a concentrated focus on dimensionality (D) and within-sample diversity ($\bar{P}$)—elements that are fundamentally tied to the compositional nature of the data. Our objective was to isolate and examine biases arising specifically from these compositional attributes, recognizing their direct impact on correlation analysis. While we acknowledge that sparsity and non-Gaussian distribution patterns also affect correlation metrics, these elements are secondary in the context of compositional data analysis. They were thus delineated outside of this study's primary scope and are addressed in a subsequent section.

Utilizing the *mvtnorm* R package [33], we constructed a matrix of variables following a multivariate Gaussian distribution. In this matrix, the dimension D corresponds to the variables (or taxa), and N signifies the number of observations or samples, all governed by a predefined correlation matrix. To enable the calculation of the Pielou index without modifying the correlation structure, all generated values were shifted to be positive.

To tune the within dataset diversity of the generated Gaussian data, a simply but functional strategy was employed: applying a multiplicative factor to one selected variable from the Gaussian-generated dataset. This deliberate manipulation skewed the distribution towards this variable, thus altering the dataset's diversity ($\bar{P}$) without distorting the established correlation structure.

Following this adjustment for within dataset diversity, both L1 and CLR transformations were applied to the synthetic datasets. We then extracted the correlation matrices from these transformed datasets to analyze the biases each normalization method introduced.

Realistic Synthetic Data Generation for Sparsity Biases Characterization on Correlation Measurement

To generate realistic artificial data with specified characteristics such as dimensionality (D), correlation structure (R), and sparsity (Φ), we have extensively used the 'Normal to Anything' (NorTA) paradigm. This framework is capable of producing an arbitrary multivariate distribution that conforms to a pre-established correlation structure R, drawing upon the principles of copula functions theory [34]. Essentially, the NorTA method allows for the transformation of normally distributed data into any desired distribution while preserving the original correlation structure. The core principle of the NorTA approach involves two main steps: Firstly, generating a multivariate normal dataset with the desired correlation structure, and secondly, transforming this dataset to have the targeted distribution while maintaining the predetermined correlations. The transformation is mathematically represented as follows:

$$U_{Gen} = F^{-1}(\mathrm{CDF}(U)) \tag{5}$$

In this equation, U represents the multivariate normal data, CDF is the cumulative distribution function of the normal distribution, F^{-1} is the inverse CDF (quantile function) of the target distribution, and U_{Gen} is the transformed data with the desired distribution and correlation structure.

We have already defined key parameters of the generated dataset, indeed the dimensionality (D) and the correlation structure (R) are trivially integrated within the

NorTA framework. However, the delineation of dataset sparsity (Φ) is a less obvious aspect, it is determined by the selection of the marginal distribution ρ. To introduce sparsity we have to appeal to the zero-inflated or the hurdle versions of conventional distributions. These modified distributions include an additional parameter, commonly denoted as ϕ, which regulates the proportion of zero-valued data. Thus, the level of sparsity within the final dataset Φ depends from the ϕ_i parameters designated for each marginal distribution.

Finally, we perform L1 and CLR transformations on the tuned dataset U_{Gen}, yielding U_{L1} and U_{CLR}, respectively, each with their corresponding correlation matrices R_{L1} and R_{CLR}. The central goal of our model is to assess how these transformations impact the correlation matrices in comparison to the original matrix R, and not respect the empirical matrix from U_{Gen}. Specifically, we aim also to evaluate the CLR transformation's efficacy in addressing the skewness and normalizing data with heavy-tailed distributions through logarithmic scaling.

Since the CLR transformation is not defined for zero values, we replaced them with a value corresponding to the 65% of the sample detection limit, in order to minimize the distortion in the covariance structure, as in [35, 36].

HMP2 16S Human Gut Data

We utilized the Human Microbiome Project's second iteration (HMP2) dataset, which encompasses operational taxonomic unit (OTU) counts and taxonomic classifications from a longitudinal study on the microbiomes of healthy and prediabetic individuals over a period of up to four years [40]. The complete dataset includes 1122 samples encompassing 1953 OTUs derived from 96 subjects. Each sample is accompanied by metadata indicating the health status of the corresponding subject. To enhance the homogeneity of the dataset for our analysis, we narrowed the focus to a single subject coded as 69-001, who is classified as healthy and has contributed 51 samples. To refine the dataset further, we applied a filtering process based on OTU prevalence and median values of the abundances. Specifically, we retained OTUs with non-zero values in $> 33\%$ of the samples and a median value of non-zero counts ≥ 5. This stringent selection criterion was designed to eliminate the rarest OTUs and focus on those with a consistent presence across the samples, thereby facilitating a more robust subsequent analysis.

Data and Code Availability

For free access to all the code and data utilized, please visit the following URL: `https://github.com/Fuschi/Correlation-Biases-on-Metagenomics-Data` - GitHub Repository. This repository contains comprehensive resources for replicating the analyses based on R base [37], *VGAM* [38], *mvtnorm* [33], and *igraph* [39].

Acknowledgments

D. R. and A. F. acknowledge EU H2020 "VEO - Versatile Emerging infectious disease Observatory" Project n. 874735 and EU H2020 ERA-HDHL "SYSTEMIC - An integrated approach to the challenge of sustainable food systems" n. 696295.

The authors would like to thank G.W. for inspiring this work through fruitful discussions and joint work. His loss is a big miss for all of us.

References

1. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–214. doi:10.1038/nature11234.

2. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550(7674):61–66. doi:10.1038/nature23889.

3. Hendriksen RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nature Communications. 2019;10(1):1124. doi:10.1038/s41467-019-08853-3.

4. Newman M. Networks: An Introduction. 1st ed. Oxford ; New York: Oxford University Press; 2010.

5. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2(1):15. doi:10.1186/2049-2618-2-15.

6. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM. A field guide for the compositional analysis of any-omics data. GigaScience. 2019;8(9):giz107. doi:10.1093/gigascience/giz107.

7. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology. 2017;8:2224. doi:10.3389/fmicb.2017.02224.

8. Aitchison J. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society Series B (Methodological). 1982;44(2):139–177.

9. Pearson K. Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London. 1997;60(359-367):489–498. doi:10.1098/rspl.1896.0076.

10. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS computational biology. 2012;8(9):e1002687. doi:10.1371/journal.pcbi.1002687.

11. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: A Valid Alternative to Correlation for Relative Data. PLOS Computational Biology. 2015;11(3):e1004075. doi:10.1371/journal.pcbi.1004075.

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

13. Quinn T, Richardson M, Lovell D, Crowley T. Propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. Scientific Reports. 2017;7. doi:10.1038/s41598-017-16520-0.

14. Peschel S, Müller CL, von Mutius E, Boulesteix AL, Depner M. NetCoMi: network construction and comparison for microbiome data in R. Briefings in Bioinformatics. 2021;22(4):bbaa290. doi:10.1093/bib/bbaa290.

15. Deutschmann IM, Lima-Mendez G, Krabberød AK, Raes J, Vallina SM, Faust K, et al. Disentangling environmental effects in microbial association networks. Microbiome. 2021;9(1):232. doi:10.1186/s40168-021-01141-7.

16. Yang P, Tan C, Han M, Cheng L, Cui X, Ning K. Correlation-Centric Network (CCN) representation for microbial co-occurrence patterns: new insights for microbial ecology. NAR Genomics and Bioinformatics. 2020;2(2):lqaa042. doi:10.1093/nargab/lqaa042.

17. McGregor K, Labbe A, Greenwood CMT. MDiNE: a model to estimate differential co-occurrence networks in microbiome studies. Bioinformatics. 2020;36(6):1840–1847. doi:10.1093/bioinformatics/btz824.

18. Jiang S, Xiao G, Koh AY, Chen Y, Yao B, Li Q, et al. HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity. Frontiers in Genetics. 2020;11.

19. Ha MJ, Kim J, Galloway-Peña J, Do KA, Peterson CB. Compositional zero-inflated network estimation for microbiome data. BMC Bioinformatics. 2020;21(Suppl 21):581. doi:10.1186/s12859-020-03911-w.

20. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology. 2020;21(1):111. doi:10.1186/s13059-020-02015-1.

21. Yang P, Yu S, Cheng L, Ning K. Meta-network: optimized species-species network analysis for microbial communities. BMC Genomics. 2019;20(2):187. doi:10.1186/s12864-019-5471-1.

22. Tavakoli S, Yooseph S. Learning a mixture of microbial networks using minorization–maximization. Bioinformatics. 2019;35(14):i23–i30. doi:10.1093/bioinformatics/btz370.

23. Tackmann J, Matias Rodrigues JF, von Mering C. Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data. Cell Systems. 2019;9(3):286–296.e8. doi:10.1016/j.cels.2019.08.002.

24. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35(17):3055–3062. doi:10.1093/bioinformatics/bty1054.

25. Yang Y, Chen N, Chen T. Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model. Cell Systems. 2017;4(1):129–137.e5. doi:10.1016/j.cels.2016.12.012.

26. Fang H, Huang C, Zhao H, Deng M. gCoda: Conditional Dependence Network Inference for Compositional Data. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology. 2017;24(7):699–708. doi:10.1089/cmb.2017.0054.

27. Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape. F1000Research; 2016. 5:1519. Available from: https://f1000research.com/articles/5-1519.

28. Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for compositional data through Lasso. Bioinformatics. 2015;31(19):3172–3180. doi:10.1093/bioinformatics/btv349.

29. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial Co-occurrence Relationships in the Human Microbiome. PLOS Computational Biology. 2012;8(7):e1002606. doi:10.1371/journal.pcbi.1002606.

30. Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. Genome Biology. 2020;21:191. doi:10.1186/s13059-020-02104-1.

31. Kurtz Z, Mueller C, Miraldi E, Bonneau R. SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference; 2023. Available from: https://github.com/zdk123/SpiecEasi.

32. Pielou EC. The measurement of diversity in different types of biological collections. Journal of Theoretical Biology. 1966;13:131–144. doi:10.1016/0022-5193(66)90013-0.

33. Genz A, Bretz F. Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics. Heidelberg: Springer-Verlag; 2009.

34. Nelsen RB. An Introduction to Copulas. Springer Series in Statistics. New York, NY: Springer; 2006. Available from: http://link.springer.com/10.1007/0-387-28678-0.

35. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Mathematical Geology. 2003;35(3):253–278. doi:10.1023/A:1023866030544.

36. Lubbe S, Filzmoser P, Templ M. Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chemometrics and Intelligent Laboratory Systems. 2021;210:104248. doi:10.1016/j.chemolab.2021.104248.

37. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/.

38. R: The R Project for Statistical Computing;. Available from: https://www.r-project.org/.

39. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695.

40. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, et al. The Integrative Human Microbiome Project. Nature. 2019;569(7758):641–648. doi:10.1038/s41586-019-1238-8.