TYPE Methods
PUBLISHED 04 October 2022
DOI 10.3389/fmicb.2022.1016947

Effective data filtering is prerequisite for robust microbial association network construction

Mengqi Wang[1] and Qichao Tu[1,2,3,4]*

[1]Institute of Marine Science and Technology, Shandong University, Qingdao, China, [2]Joint Lab for Ocean Research and Education at Dalhousie University, Shandong University, Qingdao, China, [3]Joint Lab for Ocean Research and Education at Dalhousie University, Xiamen University, Qingdao, China, [4]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Guangzhou, China

Check for updates

OPEN ACCESS

EDITED BY
Decai Jin,
Research Center
for Eco-Environmental Sciences (CAS),
China

REVIEWED BY
Kai Feng,
Key Laboratory of Environmental
Biotechnology, Research Center
for Eco-Environmental Sciences (CAS),
China
Yujia Qin,
University of Hawai`i at Manoa,
United States

*CORRESPONDENCE
Qichao Tu
tuqichao@sdu.edu.cn

SPECIALTY SECTION
This article was submitted to
Microbial Symbioses,
a section of the journal
Frontiers in Microbiology

RECEIVED 11 August 2022
ACCEPTED 20 September 2022
PUBLISHED 04 October 2022

CITATION
Wang M and Tu Q (2022) Effective data filtering is prerequisite for robust microbial association network construction.
Front. Microbiol. 13:1016947.
doi: 10.3389/fmicb.2022.1016947

Microorganisms do not exist as individual population in the environment. Rather, they form complex assemblages that perform essential ecosystem functions and maintain ecosystem stability. Besides the diversity and composition of microbial communities, deciphering their potential interactions in the form of association networks has attracted many microbiologists and ecologists. Much effort has been made toward the methodological development for constructing microbial association networks. However, microbial profiles suffer dramatically from zero values, which hamper accurate association network construction. In this study, we investigated the effects of zero-value issues associated with microbial association network construction. Using the TARA Oceans microbial profile as an example, different zero-value-treatment approaches were comparatively investigated using different correlation methods. The results suggested dramatic variations of correlation coefficient values for differently treated microbial profiles. Most specifically, correlation coefficients among less frequent microbial taxa were more affected, whichever method was used. Negative correlation coefficients were more problematic and sensitive to network construction, as many of them were inferred from low-overlapped microbial taxa. Consequently, microbial association networks were greatly differed. Among various approaches, we recommend sequential calculation of correlation coefficients for microbial taxa pairs by excluding paired zero values. Filling missing values with pseudo-values is not recommended. As microbial association network analyses have become a widely used technique in the field of microbial ecology and environmental science, we urge cautions be made to critically consider the zero-value issues in microbial data.

KEYWORDS

microbial community, zero values, data filtering, correlation inference, association networks

Introduction

Rather than simple accumulation of individual populations, microorganisms in natural ecosystems may interact with each other and form complex assemblages, thereby execute essential ecosystem functions (e.g., biogeochemical cycling of various nutrients) and maintain ecosystem stability (Fuhrman et al., 2015). Therefore, in addition to the community structure and composition, their interactive relationships shall also be critically considered for better understanding the roles that microbial communities play in both natural and artificial ecosystems. However, due to the complexity and not yet cultured majority of microbial taxa, direct observation of such complex relationships is generally not feasible (Trosvik et al., 2010; Berry and Widder, 2014). Recent advances in high throughput metagenomic technologies make it possible to capture genomic information for thousands of microbial taxa in a single experiment (van Dijk et al., 2014; Slatko et al., 2018). Since then, multiple statistical approaches, such as SparCC (Friedman and Alm, 2012), MENA (Deng et al., 2012), LSA (Xia et al., 2013), CoNet (Faust and Raes, 2016), and SPIEC-EASI (Kurtz et al., 2015), have been developed to infer the complex relationship among microbial taxa based on the (relative) abundance profiles, in the form of association networks.

Notably, these different approaches mainly focus on different statistical methods for pairwise correlation coefficients calculation and network construction. For example, SparCC focuses on correlation inference of compositional data, and uses Aitchison's variance of log-ratios to solve ingredient problems (Friedman and Alm, 2012). MENA uses Random Matrix Theory (RMT) to identify a cutoff for constructing microbial association networks, by determining the transition point of nearest-neighbor spacing distribution of eigenvalues from Gaussian (random) to Poisson (non-random) distribution (Deng et al., 2012). LSA breaks down the global molecular similarity as local similarity at each grid point surrounding the molecules and is efficient to calculate statistical significance for pairwise local similarity analysis, making possible all-to-all local association studies otherwise prohibitive. The LSA approach is commonly used to construct association network from time series data (Xia et al., 2013). CoNet offers ensemble-based network construction, by combining a number of different correlations (Pearson, Spearman, and Kendall), similarities (Mutual information) or dissimilarities (Bray-Curtis and Kullback-Leibler) to improve the accuracy of the strength of the associations between objects (Faust and Raes, 2016). SPIEC-EASI is a computational framework that includes statistical methods for the inference of microbial ecological interactions from 16S rRNA gene sequencing datasets, and is equipped with a sophisticated synthetic microbiome data generator with controllable underlying species interaction topology (Kurtz et al., 2015).

Microbial communities in natural ecosystems are usually consisted by a few dozens of abundant and occasional taxa, and a long "tail" of rare taxa (Logares et al., 2014). As a result, zero values are especially common in microbial community profiles, especially for occasional and rare taxa, suggesting that these microbial taxa could be either biologically absent or not detected by the current sequencing effort. Consequently, inadequate considering the issues associated with zero values may lead to problematic microbial association network construction. However, current studies constructing microbial association networks generally input the (log-transformed) (relative) abundance profiles to the abovementioned approaches and do not specifically consider the situation of zero values in microbial profiles.

In this study, we aimed to investigate the issues associated with correlation coefficient inference and association network construction caused by zero values in microbial profiles. Multiple situations for dealing with zero values in microbial profiles were considered and comparatively evaluated for association network construction. The results demonstrated that different data filtering methods for zero values dramatically interfere microbial association network construction. Different data filtering methods may exert as strong variation in correlation coefficient values as network correlation inference methods. Such issues were especially critical for negative association inference. Therefore, we urge that the common zero values in microbial profiles be thoroughly considered for robust microbial association network construction.

Materials and methods

Methodological framework

A framework was presented to illustrate the overall methodology flow of this study (**Figure 1**). A typical microbial profile containing large number of zero values was selected as the example dataset. Different data filtering methods were then applied to process the zero values associated with the microbial profile. The commonly used correlation coefficient calculation methods including the Spearman, Pearson and Kendall were applied to calculate the pairwise correlation among microbial taxa. The SparCC method was also employed to construct microbial association networks based on the microbial profile containing large number of zero values. Association networks were then constructed with identical cutoffs and comparatively analyzed, in order to evaluate the effects of different data filtering methods.

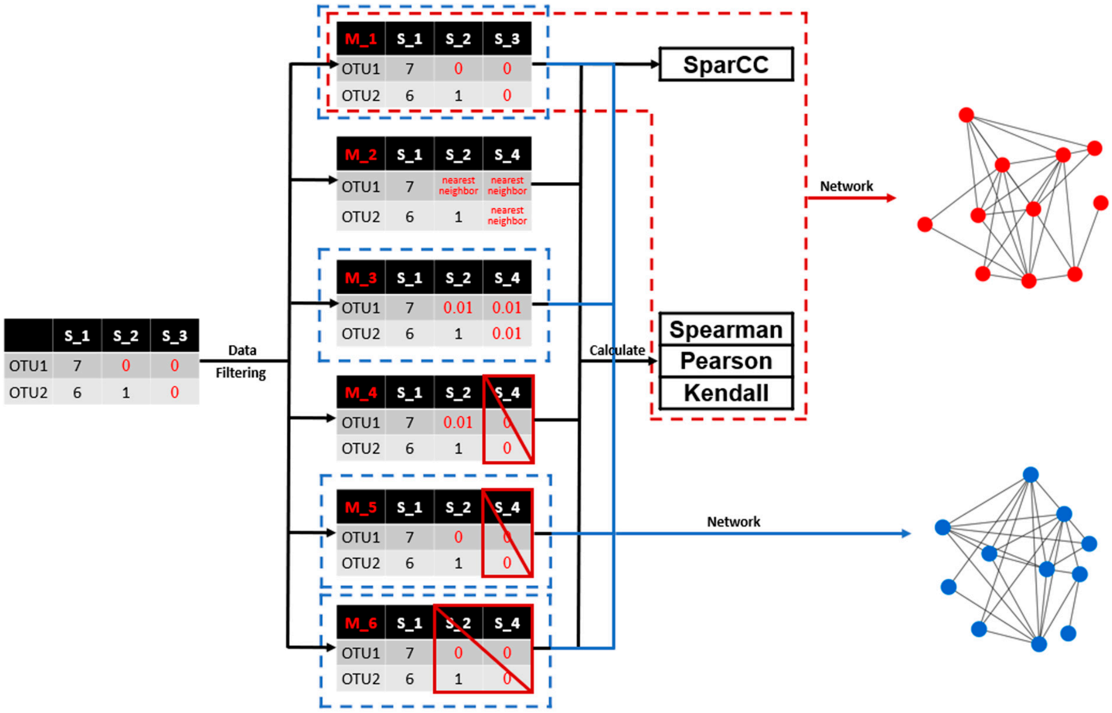


FIGURE 1
A flowchart illustrating the overall methodology workflow of this study. A typical microbial profile containing large number of missing values was selected as the example dataset. Different data filtering methods were then applied to process the missing values associated with the microbial profile. Commonly used correlation coefficient calculation methods including the SparCC, Spearman, Pearson, and Kendall were applied to calculate the pairwise correlation among microbial taxa. Association networks were then constructed and comparatively analyzed.

The TARA oceans microbial profile

We first tried to generate synthetic microbial profiles using models like generalized Lotka–Volterra (gLV) model (Hernández-Bermejo and Fairén, 1997). However, effective dataset containing comparable zero-values to experimentally generated datasets cannot be generated. Due to the lack of standard microbial profiles with zero values and known relationships, we used a real dataset, the TARA Oceans metagenome, to illustrate the effects of different data filtering methods on microbial correlation coefficient calculation. The TARA Oceans metagenomic dataset was selected because it is a representative dataset covering 139 microbial communities in the largest fluid ecosystem in the Earth's biosphere. Zero-values are common in this dataset, satisfying the requirement of this study. The miTAG based microbial taxonomic profile was downloaded from http://ocean-microbiome.embl.de/companion.html. The row annotated as unclassified taxa was first removed from the microbial profile for its representing an aggregation of many unknown microbial OTUs. The profile was then rarefied to a same sequencing depth (39,410 reads per sample). In the profile, a total of 35,650 OTUs were found. The absolute value of OTU abundance after rarefaction was retained for correlation coefficient calculation. In addition to direct using abundance data for correlation calculation, several studies suggest that microbial profiles are compositional (Jackson, 1997; Friedman and Alm, 2012). Therefore, in addition to rarefied abundance data, the microbial profile was also centered log transformed and subject to correlation calculation. The R package "compositions" was used for centered log transformation (van den Boogaart and Tolosana-Delgado, 2008).

Data filtering and correlation coefficient calculation

Six different data filtering methods were employed to deal with zero values (**Figure 1**). The first method is to treat zero values (i.e., NA values) as 0. The second method is to use a nearest neighbor algorithm instead of zero values (Cover and Hart, 1967), by which zero values were replaced by inferred expectations based on observed data. The R package "DMwR2" was used for nearest neighbor employment (Torgo, 2016). The third method is to use 0.01 instead of zero values, which is the default value recommended by the MENA pipeline (Deng et al., 2012). The fourth method is to exclude the samples from correlation calculation in which pairwise zero values are detected and the unpaired zero values are replaced with 0.01. The fifth method is similar to the fourth method but unpaired zero values are kept as is. The sixth method is to exclude all samples from correlation calculation as long as paired or unpaired zero values are observed.

Afterward, correlation methods including Spearman, Pearson, and Kendall, which reflect the direction and degree of the change trend between two variables, were employed for correlation coefficient calculation for the treated microbial profiles. Correlation coefficient values were sequentially calculated for each pair of OTUs meeting the minimum requirement (i.e., larger than 30 paired valid observations). Since the data filtering methods are not applicable to SparCC, only the raw microbial profile containing zero values was subject to correlation calculation and network construction by SparCC. Here, the python package SparCC3[1] was used, with the iteration value set to 20. Since SPIEC-EASI is based on penalized estimators, the edge weights are not directly comparable to SparCC and Pearson/Spearman correlation coefficients, this method was not investigated here. In some cases (e.g., methods 4, 5, and 6), OTU pairs that did not meet the minimum requirement (i.e., larger than 30 paired valid observations) for correlation calculation were excluded. The calculated correlation coefficients by different data filtering methods and correlation methods were comparatively analyzed. For better illustration, a total of 10,000 OTU pairs were randomly selected. The associated R code and documents for correlation calculation are available at https://github.com/qichao1984/DataFiltering.

Microbial association network construction

In many approaches, an empirical cutoff is selected for association network construction, except the MENA pipeline. Here, for evaluation purpose, a correlation cutoff with absolute value larger than 0.6 and *P*-value smaller than 0.001 were chosen for network construction. As in many other studies, we focus on the patterns of association networks and do not expect strong effects of this cutoff on the results and conclusions. In addition, Random Matrix Theory (RMT) was also employed via the iNAP pipeline (Feng et al., 2022) to identify thresholds for constructing highly confident microbial ecological networks, together with the *P*-value cutoff of 0.001 (Zhou et al., 2010; Feng et al., 2022). Two types of association networks, including co-occurrence and co-exclusive networks, were constructed and analyzed. In the co-occurrence network, only positive associations were included, whereas, in the co-exclusive network, only negative associations were included. Microbial association networks were constructed for microbial profiles subject to different data filtering methods and correlation methods. The structural and topological differences among the constructed networks were comparatively inspected.

To better illustrate how microbial association networks differ from each other, subnetworks were also constructed targeting the first neighbors of top ten most connected nodes identified in the fifth data filtering methods. In addition, consensus networks were constructed based on differently treated microbial profiles and different correlation methods, in order to see the variations caused by data filtering methods and correlation methods. The constructed networks were visualized and analyzed by the Cytoscape software (Shannon et al., 2003).

Results

Zero values were rarely critically considered in microbial association network constructions

We first investigated the current situation of implementing microbial association network analyses in microbial ecological and environmental studies by analyzing published literatures. Using the keywords "microbial ecological network" and "microbial co-occurrence network," it was found that the number of published studies in the NCBI PubMed database increased dramatically over the past 13 years. Specifically, the number of indexed literatures per year achieved to ~581 in December 2021, whereas the number was only ~15 in 2009 (**Figure 2A**). This means that, as a recently developed technology in the metagenomic era, microbial association network analyses have become a routine approach in microbial ecological and environmental studies. Among these studies, SparCC is the method with most citations, followed by SPIEC-EASI and MENA (**Figure 2B**).

Although commonly applied, it should be pointed out that issues associated with zero values have rarely been critically considered in many studies. For computational purposes, only minimal treatments have been employed by different association network construction approaches (**Table 1**). For instance, methods like SparCC and SPIEC-EASI assign a pseudo-count of 1 to zero values, to avoid log transformation of zero value (Friedman and Alm, 2012; Kurtz et al., 2015). The MENA pipeline (Deng et al., 2012), which mainly employs Pearson correlation coefficient and Spearman's rank order correlation for association inference, fills a pseudo-value of 0.01 to zero values by default, with the option of user defined values available. A common agreement reached by computational microbial ecologists is that OTUs showing up in only a few samples should be excluded from correlation analyses, as they may lead to false association inferences (Huse et al., 2010; Haas et al., 2011; Reshef et al., 2011; Degnan and Ochman, 2012; Faust, 2021). More critically, literature analysis suggested that no specific attention

[1] https://github.com/JCSzamosi/SparCC3

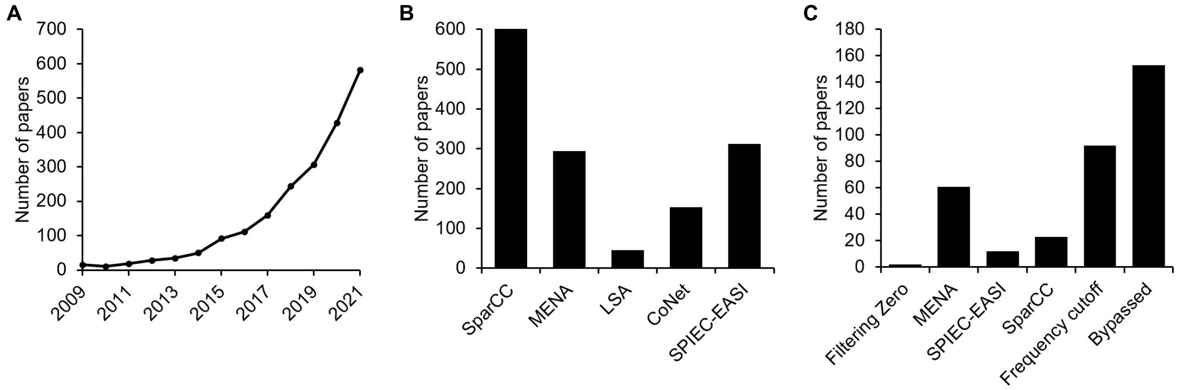


FIGURE 2
The current situation of implementing microbial association network analyses in microbial ecological and environmental studies by analyzing published literatures. **(A)** Using the keywords "microbial ecological network" and "microbial co-occurrence network," the number of papers published in the NCBI PubMed database in the past 13 years was analyzed. **(B)** Citations of different microbial association network construction methods. **(C)** The number of papers whether missing values were considered when constructing microbial association networks.

was paid to zero value issues by many studies (Figure 2C), as judged by whether zero values are specifically mentioned in the methods section, though it is possible that some studies do have considered zero values but did not mention it in the paper. Ignoring such issues may impose severe consequences to the conclusions drawn in the studies, such as inaccurate (false positive and false negative) inference of microbial association relationships. It is therefore of necessity to investigate how zero values may affect microbial association network inference.

Issues associated with microbial community profiles

Using the TARA Oceans microbiome data as an example, we first illustrated the typical data structure of microbial profiles and how different data filtering methods would affect them. Although the following issues may have already been recognized by the community, we believe they still deserve to be further emphasized here.

First, zero values are common in microbial profiles. As previously reported, microbial communities are usually composed by a small number of abundant taxa and a long "tail" of rare taxa (Logares et al., 2014), meaning the existence of many zero values in microbial profiles. Although thousands to tens of thousands microbial taxa could be identified in a routine study, only a small portion of them (6.34%) were present in multiple samples (e.g., 60 samples) (Figure 3A). In each sample of the profile, more than 88.69% observations are zero (Figure 3B). The meaning of these zero values is uncertain that they could either be biologically absent from the sample or technically not detected. In addition, random subsampling of microbial profiles to an equal sequencing depth may also result in zero values for rare taxa.

Second, not all OTUs shall be subject to correlation analyses for association network construction. Correlation coefficient between two OTUs can be calculated as long as they are present in three or more samples. And confident estimation of correlation coefficients between two OTUs requires more valid observations. However, the number of OTUs decreases with the number of samples they show up. For instance, 2 260 (6.34%) OTUs were present in ≥60 samples in the TARA Oceans expedition, though a total of 35,650 annotated OTUs were identified (Figure 3A). To our best knowledge, no standard or consensus criteria is available to define a sample number cutoff for selecting OTUs. Empirically, the MENA pipeline recommended a cutoff of showing up in 50% samples to select OTUs for correlation analyses. However, large scale studies (both spatial and temporal, e.g., studies encompassing hundreds of samples at the continent or global scale) may invalidate such empirical recommendations as very few microbial taxa may show up in ≥50% samples.

Third, treating zero values differently may lead to dramatically differed input microbial profiles for correlation analyses, further affecting association network inference. Multiple data filtering methods are available (Figure 1). Taking the TARA Oceans microbial profile containing OTUs showing up in ≥60 samples for example, the number of OTU pairs suitable for correlation analyses varied dramatically for different data filtering methods (Figure 3C).

Data filtering methods interfere microbial correlation inference

We then investigated how different data filtering methods may affect microbial correlation inference. As multiple microbial correlation inference methods are available, here only those more commonly used approaches were investigated, including the Pearson correlation coefficient, Spearman's rank order correlation, and the non-parametric Kendall rank correlation. The correlation coefficient values returned by the SparCC approach was also assessed. These are also the correlation methods used by most microbial association network construction methods. Correlation methods such as Bray-Curtis dissimilarity, Mutual Information, and Maximal Information Coefficient were not considered here due to their returned values always ≥0, failing to encompass negative correlations in microbial association relationships in the same way and criteria as other methods did.

TABLE 1 Characteristic and zero-value-treatment methods of the commonly used microbial association network construction methods.

Method	Characteristic	Zero value treatment
SparCC (6)	SparCC focuses on correlation inference of compositional data, and uses Aitchison's variance of log-ratios to solve ingredient problems. It could be used on any genomic survey data that has low diversity.	Add a pseudo-value of 1 to each element.
MENA (7)	MENA uses Random Matrix Theory (RMT) to identify a cutoff for constructing microbial association networks, by determining the transition point of nearest-neighbor spacing distribution of eigenvalues from Gaussian (random) to Poisson (non-random) distribution.	By default, unpaired zero values were filled by a pseudo-value of 0.01. More user defined options are also available.
LSA (8)	LSA breaks down the global molecular similarity as local similarity at each grid point surrounding the molecules and is efficient to calculate the statistical significance for pairwise local similarity analysis, making possible all-to-all local association studies otherwise prohibitive. It is commonly used to construct association network from time series data.	No mention of the treatment of zero values.
CoNet (9)	CoNet offers ensemble-based network construction, by combining a number of different correlations (Pearson, Spearman, and Kendall), similarities (Mutual information) or dissimilarities (Bray-Curtis and Kullback-Leibler) to improve the accuracy of the strength of the associations between objects.	Omitting sample pairs with zero values from the association strength calculation.
SPIEC-EASI (10)	SPIEC-EASI is a computational framework that includes statistical methods for the inference of microbial ecological interactions from 16S rRNA gene sequencing datasets. A sophisticated synthetic microbiome data generator with controllable underlying species interaction topology is also equipped.	Add a pseudo-value of 1 to zero values.

First, treating zero values differently resulted in dramatically different correlation coefficients. Here, microbial OTUs showing up in ≥60 samples were extracted and subjected to data filtering by all six methods. Correlation coefficients among different OTUs were calculated for the filtered microbial profiles. Using correlation coefficients among OTUs with paired abundance values (i.e., method 5) as standard, it was found that different data filtering methods resulted in different correlation coefficient values, no matter which correlation method was used (Figures 4A–C). Specifically, the correlation coefficient values for microbial profile whose zero values were filled by nearest neighbor algorithm showed the strongest variations from those by other methods. Correlation coefficient values for microbial profile whose zero values were filled with 0.01, which is the default approach used by the MENA pipeline, were much higher than the standard. Notably, correlation coefficient values were also sensitive to correlation methods. Pearson correlation coefficient seemed to be more robust than Spearman's rank correlation and Kendall's rank correlation that correlation coefficient values were more centered to the standard for abundance-based microbial profiles with different treatment of zero data, except that filled by the nearest-neighbor algorithm.

Second, correlation inference for less frequent microbial taxa was more affected. Microbial profiles consisted by OTUs showing up in ≥30 samples and ≤60 samples were also extracted and evaluated. In the nearest neighbor algorithm, missing values are inferred from its neighboring samples around. Therefore, missing values could not be inferred using the nearest neighbor algorithm for less frequent microbial taxa. As a result, comparing to the microbial profiles consisted of OTUs showing up in ≥60 samples, the correlation coefficient values for less frequent OTUs were much more sparsely distributed, no matter which correlation calculation method was used (Figures 4D–F). This suggested that cautions should be taken for microbial association network construction based on less frequent OTUs, either for choosing correlation calculation or data filtering methods.

Third, negative correlations could be more problematic. We also investigated how different data filtering methods affected positive and negative inference of microbial associations (Supplementary Figure 1). The number of strong positive and negative microbial associations ($|$ corr$| > 0.6$) greatly varied for different data filtering methods. This was especially critical for microbial profiles whose zero values were replaced by the nearest neighbor algorithm and subject to correlation calculation by Spearman and Kendall correlation methods. For positive inference, the majority of strong positive correlation were identified among microbial taxa whose overlapped observations were larger than 60 (Supplementary Figures 1A–C). For negative inference, a large portion of strong associations were identified among microbial

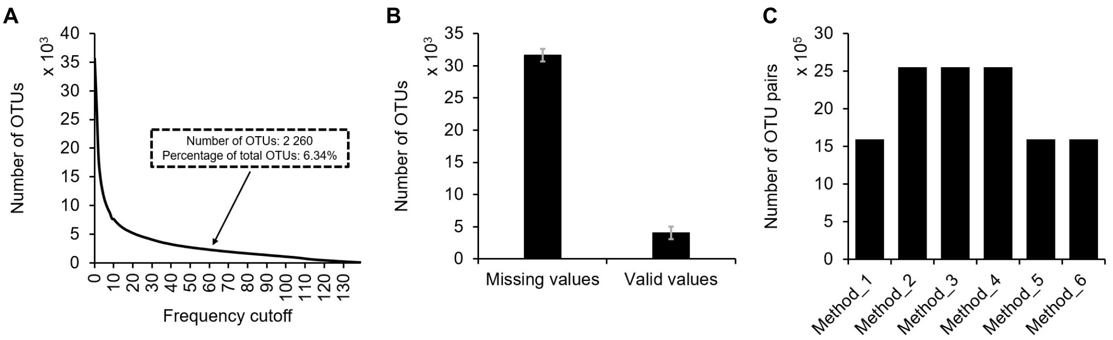


FIGURE 3

The situation of missing values in typical microbial profiles. Using the TARA Oceans miTAG microbial profile as an example, the following information was shown: **(A)** The number of OTUs under different frequency cutoff. With 60 samples as the frequency cutoff, the number of detected OTUs was 2,260, accounting for 6.34% of all OTUs. **(B)** The average number of missing values and valid values present in each sample with standard error. **(C)** The number of OTU pairs suitable for correlation analyses using different data filtering methods.

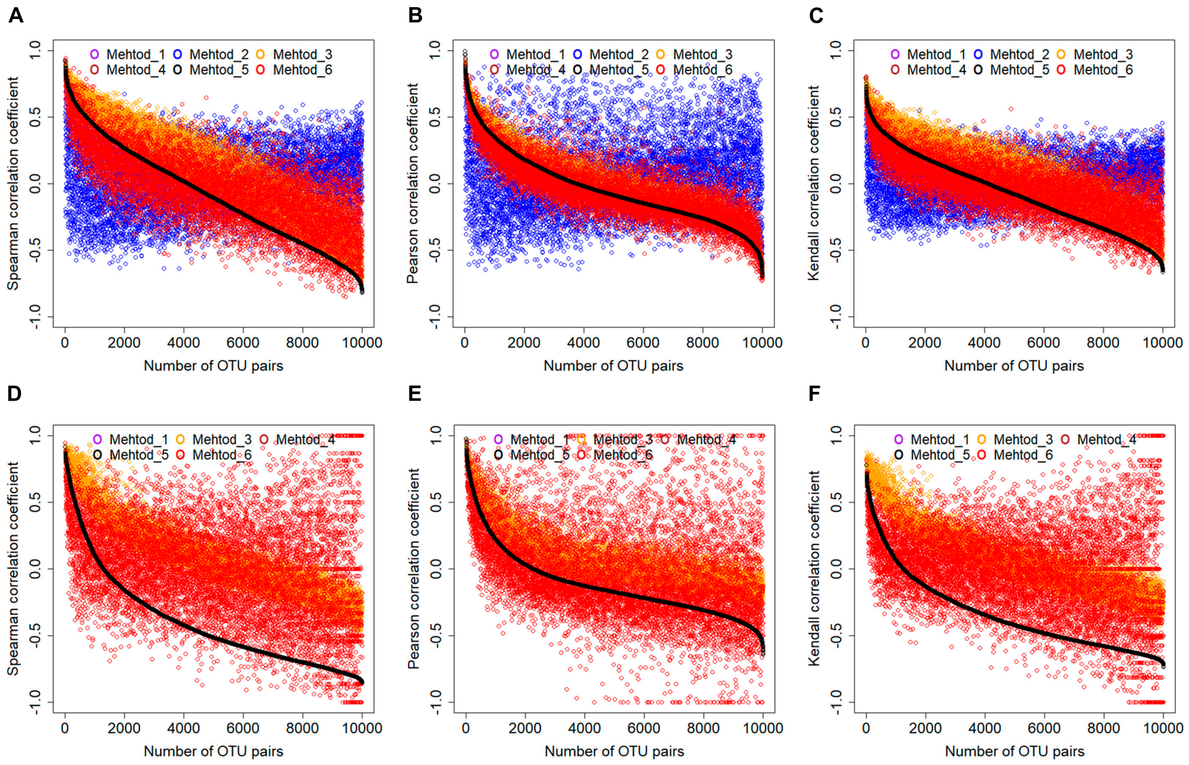


FIGURE 4

The effects of data filtering methods on correlation coefficient calculation. Two types of datasets, including OTUs showing up in ≥60 samples **(A–C)** and OTUs showing up in ≥30 samples and ≤60 samples **(D–F)** were analyzed. Three different correlation calculation methods were evaluated, including the Spearman, Pearson, and Kendall. For better visualization, correlation coefficient values of 10,000 randomly selected OTU pairs were plotted. For the less frequent dataset **(D–F)**, nearest neighbor algorithm was not applicable.

taxa with fewer overlapped observations, especially for those calculated by Spearman and Kendal correlation methods. Negative correlation coefficients were lowly detected using Pearson and Kendall correlation methods (Supplementary Figures 1D–F). Such results suggested that negative association inference for microbial association networks were more strongly affected by data filtering methods. The results also provided potential clues explaining why negative associations were rarely found in many studies.

Fourth, log-transformation of the abundance data improved the centrality of the distribution of correlation coefficient values to the standard. Several studies suggested that the abundance data of microbial profiles is compositional (Jackson, 1997; Friedman and Alm, 2012). A common approach to treat such data is log transformation. Here, the abundance profiles of microbial communities were centered log transformed and subjected to correlation coefficient calculation. As a result, comparing to untransformed microbial profiles, correlation coefficient values were overall more centralized to the standard. Such result was consistent across different correlation methods and different frequency cutoffs of microbial profiles (**Supplementary Figure 2**). Again, the correlation coefficient values for microbial profiles whose zero values were filled by nearest neighbor were always sparsely distributed from the standard. This suggested that the nearest neighbor algorithm seemed not a good way for microbial community data preprocessing, though this method has been widely applied in genomic studies (Parry et al., 2010; Belka et al., 2018; Tjarnberg et al., 2021; Xie et al., 2021).

Finally, correlation coefficient calculation method like SparCC may also be affected. Since data filtering methods are not applicable to SparCC, which is one of the representative methods that treat microbial community data as sparse, we compared the correlation coefficient values calculated by Spearman, Pearson, and Kendall correlation methods with that by SparCC (**Supplementary Figure 3**). As a result, the correlation coefficient values between SparCC and other methods were all strongly correlated ($R^2 > 0.95$, $P < 2.2e\text{-}16$). This suggested that the same issue encountered by other correlation methods also tended to exist in methods like SparCC.

Microbial association networks dramatically differed

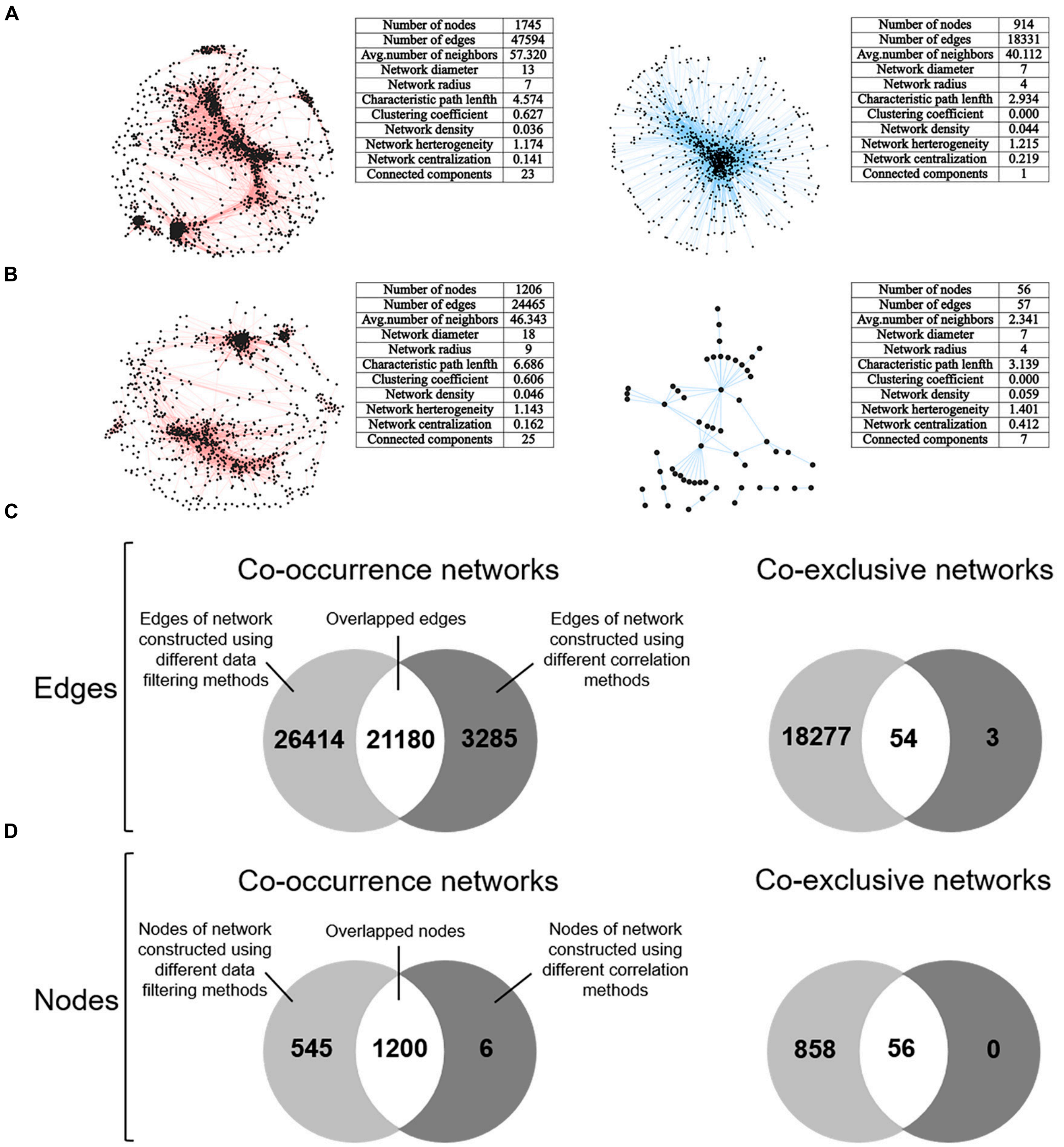
Finally, microbial association networks were constructed and comparatively analyzed. Correlation coefficient cutoffs of 0.6 and –0.6 ($P < 0.001$) were respectively used to identify positive and negative associations among different microbial taxa. Using the microbial association networks constructed based on Spearman correlation coefficient as examples, dramatically differed networks were obtained, both topologically and in content (**Supplementary Figure 4**). For both co-occurrence and co-exclusive networks, distinct network structures were observed for microbial profiles subject to different data filtering methods. Such distinct network topology resulted in distinct network parameters such as the number of connected nodes/edges, average number of neighbors, network diameter, network radius, characteristic path length, clustering coefficient, network density, network heterogeneity, network centralization, and connected components. We then extracted the subnetworks of top ten most connected nodes (**Supplementary Figure 5**), which is the commonly used approach to investigate the network topology of the most connected microbial taxa. More strikingly, the extracted subnetworks suggested that different data filtering methods resulted in different clues in finding out the most "important" microbial taxa in the ecosystem. Further, we constructed and compared the consensus networks by extracting the overlapped nodes and edges in the networks constructed using different data filtering methods and correlation methods (**Figure 5**). As a result, great difference among different data filtering methods could be found. Highly dissimilar consensus network was observed for different data filtering methods and different correlation methods.

In addition, the iNAP pipeline was also employed to determine the cutoffs for network construction based on RMT method. Similar analyses were then carried out (**Supplementary Figures 6–8**). In general, consistent result was observed that microbial association networks highly differed for different data filtering methods. However, as the cutoffs determined by RMT were much larger than 0.6, the constructed networks were strongly altered with much fewer nodes and links, especially for co-exclusive networks. Such scenario was expected considering the large cutoffs used for network construction. Based on the above results, we demonstrated that data filtering was also a critical issue as the selection of correlation methods.

Discussion

As a sophisticated computational approach developed about 10 years ago (Weiss et al., 2016; Röttjers and Faust, 2018), microbial association network analyses have been widely performed in microbial ecology and associated fields to reveal the potential relationships among microbial taxa (Barberán et al., 2012; de Vries et al., 2018; Guo et al., 2022), as well as those between microbes and environmental factors (Chaffron et al., 2010; Lima-Mendez et al., 2015). With proper experimental design and data interpretation, interesting studies that cannot be achieved conventionally have been carried out, such as accurate predicting the biotic interactions among microbes (Barzel and Barabási, 2013; Feizi et al., 2013; Lv et al., 2019), identifying microbial taxa belonging to the same ecological niche (Ma et al., 2020), and revealing the community diversity from the angle of microbial interconnected relationships (Ma et al., 2016; Tu et al., 2020). Foreseeably, the approach will be much more intensively employed by researchers from various fields to address important biological, ecological and environmental questions.

Similar to other approaches, microbial association network analyses also confront a series of caveats and challenges, from methodology to interpretation of the results. A series of issues have been recently discussed by Faust (2021) and Goberna and Verdú (2022). Among these, the issue of zero values is also emphasized in context with rare taxa, for which the majority values are filled with zero. However, the issue of zero values is



Number of nodes	1745
Number of edges	47594
Avg.number of neighbors	57.320
Network diameter	13
Network radius	7
Characteristic path lenfth	4.574
Clustering coefficient	0.627
Network density	0.036
Network herterogeneity	1.174
Network centralization	0.141
Connected components	23

Number of nodes	914
Number of edges	18331
Avg.number of neighbors	40.112
Network diameter	7
Network radius	4
Characteristic path lenfth	2.934
Clustering coefficient	0.000
Network density	0.044
Network herterogeneity	1.215
Network centralization	0.219
Connected components	1

Number of nodes	1206
Number of edges	24465
Avg.number of neighbors	46.343
Network diameter	18
Network radius	9
Characteristic path lenfth	6.686
Clustering coefficient	0.606
Network density	0.046
Network herterogeneity	1.143
Network centralization	0.162
Connected components	25

Number of nodes	56
Number of edges	57
Avg.number of neighbors	2.341
Network diameter	7
Network radius	4
Characteristic path lenfth	3.139
Clustering coefficient	0.000
Network density	0.059
Network herterogeneity	1.401
Network centralization	0.412
Connected components	7

FIGURE 5

Consensus networks representing microbial co-occurrence and co-exclusive associations were constructed and comparatively analyzed. **(A)** Consensus co-occurrence and co-exclusive networks constructed by extracting the overlapped nodes and edges based on networks constructed using the first, third, fifth, and sixth data filtering method (refer to **Figure 1**). The Spearman's rank order correlation coefficient was used for network construction. **(B)** Consensus co-occurrence and co-exclusive networks constructed by extracting the overlapped nodes and edges based on networks constructed using different correlation methods, including Spearman, Pearson, Kendall, and SparCC. The first data filtering method (filling missing values with zero) was used here. **(C)** Overlapped edges between the consensus networks of different data filtering methods and correlation methods. **(D)** Overlapped nodes between the consensus networks of different data filtering methods and correlation methods. In panels **(A,B)**, networks with red edges represent co-occurrence networks, and networks with blue edges represent co-exclusive networks.

not only associated with rare, but also abundant and occasional taxa, though with lesser extent. Most correlation coefficient measurements are not designed to consider zero values (i.e., zeros), except few such as Bray-Curtis dissimilarity, which automatically ignores matching zeros between microbial taxa pairs (Faust, 2021). However, dissimilarity values of Bray-Curtis range from 0 to 1, making it hard to define negative correlations in the same manner as other methods.

Current effort mainly focuses on methodology development and novel applications of the approach, leaving the zero-value issue less attended. In previous attempts, researchers generate simulated datasets to validate the accuracy and robustness of the network construction methods they developed (Friedman and Alm, 2012; Kurtz et al., 2015; Yang et al., 2021). However, to our best knowledge, the complex issue of zero values in microbial profiles cannot be well-simulated. Recently, Cougoul et al. (2019) proposed a framework to compute the upper bounds beyond which associations become less meaningful. This study demonstrated that the zero-value issue could be one of the most critical issues in microbial association network construction. As a computational approach, microbial association network analyses also comply the classic "garbage in, garbage out" bioinformatics principle, in which problematic outputs are given by problematic input data (Mellin, 1957; Babbage, 1994). From this point of view, proper data input could be even more important than the choice of correlation methods and network construction methods. This was also confirmed in this study that different data filtering methods resulted in dramatically varied correlation coefficient values using the same correlation method. Different data filtering methods may exert as strong variation on correlation coefficient values as different correlation methods.

Integrating all the results obtained in this study, we propose to critically consider the following caveats for more confident and robust microbial association network construction. First, the frequent zero values in microbial profiles may cause highly differed correlation inference in microbial association network construction. In general, paired valid values are expected to generate robust inference of both positive and negative associations, whereas unpaired valid values could be of great importance for negative associations. Not only shall attention be paid to rare taxa (Faust, 2021), but also abundant and occasional taxa are affected. Second, filling zero values with pseudo values is not recommended. Biased correlation values are observed for microbial profiles filled with pseudo values. The commonly used approaches in the genomic era, e.g., filling zero values using nearest neighbor algorithm (Torgo, 2016; Tjarnberg et al., 2021; Dann et al., 2022), do not seem to be effective in metagenomics. Third, rare and abundant taxa shall be differently treated prior to network construction. For rare taxa with less statistical power, either prevalence removal or forbidding computing correlation coefficients among rare taxa with large number of matching zeros shall be considered, though it may ignore valuable information carried by them (McMurdie and Holmes, 2014; Faust, 2021). Notably, cautions should be made to avoid altered relative abundance due to the removal of rare taxa. For correlation inference among abundant taxa, paired zeros shall be excluded for their biologically insignificance and reducing the statistical power of confident observations. As such, sequential correlation inference between taxa pairs is recommended to avoid miscellaneous issues associated with correlation inference based on whole microbial profile. Many R functions for correlation calculation usually treat and include zero values as is. Fourth, different data filtering methods result in dramatically different microbial association networks. It is therefore necessary for researchers to thoroughly consider the zero values associated with microbial profiles and choose the appropriate one for correlation calculation in network construction. This is an especially critical issue for negative association inference. Whether and how data filtering is carried out for association network construction shall be clearly stated in the manuscripts.

In conclusion, this study investigated the zero-value issue associated with microbial association network construction, which is an issue rarely critically considered in many of the current studies. Different data filtering methods resulted in highly differed association inference for network constructions, especially for negative association inference. As microbial association network approaches are being widely applied in various fields, even to infer relationships among microbial taxa (Barberán et al., 2012; de Vries et al., 2018; Guo et al., 2022), environmental factors (Chaffron et al., 2010; Lima-Mendez et al., 2015), and organic molecules (Li et al., 2018, 2019; Zhao et al., 2019), we urge this issue be critically considered in future studies. Whether and how zero values are treated shall be clearly stated in each study.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Author contributions

MW: conceptualization, software, investigation, formal analysis, and writing – original draft. QT: conceptualization, funding acquisition, resources, supervision, and writing – review and editing. Both authors contributed to the article and approved the submitted version.

Funding

This study was supported by National Key Research and Development Program of China (2020YFA0607600 and 2019YFA0606700), by the National Natural Science Foundation of China (92051110 and 31971446), by the Natural Science Foundations of Shandong Province (ZR2020YQ21), and by the Qilu Young Scholarship of Shandong University. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1016947/full#supplementary-material

References

Babbage, C. (1994). *Passages from the life of a philosopher*. Salt Lake City, UT: Project Gutenberg.

Barberán, A., Bates, S. T., Casamayor, E. O., and Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* 6, 343–351. doi: 10.1038/ismej.2011.119

Barzel, B., and Barabási, A.-L. (2013). Network link prediction by global silencing of indirect correlations. *Nat. Biotechnol.* 31, 720–725. doi: 10.1038/nbt.2601

Belka, A., Fischer, M., Pohlmann, A., Beer, M., and Höper, D. L. V. Q.-K. N. N. (2018). Composition-based DNA/RNA binning of short nucleotide sequences utilizing a prototype-based k-nearest neighbor approach. *Virus Res.* 258, 55–63. doi: 10.1016/j.virusres.2018.10.002

Berry, D., and Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* 5:219. doi: 10.3389/fmicb.2014.00219

Chaffron, S., Rehrauer, H., Pernthaler, J., and Von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* 20, 947–959. doi: 10.1101/gr.104521.109

Cougoul, A., Bailly, X., Vourch, G., and Gasqui, P. (2019). Rarity of microbial species: In search of reliable associations. *PloS One* 14:e0200458. doi: 10.1371/journal.pone.0200458

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D., and Marioni, J. C. (2022). Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* 40, 245–253. doi: 10.1038/s41587-021-01033-z

de Vries, F. T., Griffiths, R. I., Bailey, M., Craig, H., Girlanda, M., Gweon, H. S., et al. (2018). Soil bacterial networks are less stable under drought than fungal networks. *Nat. Commun.* 9, 3033–3012. doi: 10.1038/s41467-018-05516-7

Degnan, P. H., and Ochman, H. (2012). Illumina-based analysis of microbial community diversity. *ISME J.* 6, 183–194. doi: 10.1038/ismej.2011.74

Deng, Y., Jiang, Y.-H., Yang, Y., He, Z., Luo, F., Zhou, J., et al. (2012). Molecular ecological network analyses. *BMC Bioinformatics* 13:113. doi: 10.1186/1471-2105-13-113

Faust, K. (2021). Open challenges for microbial network construction and analysis. *ISME J.* 15, 3111–3118. doi: 10.1038/s41396-021-01027-4

Faust, K., and Raes, J. (2016). CoNet app: Inference of biological association networks using Cytoscape. *F1000 Res.* 5:1519. doi: 10.12688/f1000research.9050.2

Feizi, S., Marbach, D., Médard, M., and Kellis, M. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* 31, 726–733. doi: 10.1038/nbt.2635

Feng, K., Peng, X., Zhang, Z., Gu, S., He, Q., Shen, W., et al. (2022). iNAP: An integrated network analysis pipeline for microbiome studies. *iMeta* 1:e13. doi: 10.1002/imt2.13

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Fuhrman, J. A., Cram, J. A., and Needham, D. M. (2015). Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* 13, 133–146. doi: 10.1038/nrmicro3417

Goberna, M., and Verdú, M. (2022). Cautionary notes on the use of co-occurrence networks in soil ecology. *Soil Biol. Biochem.* 166:108534. doi: 10.1016/j.soilbio.2021.108534

Guo, B., Zhang, L., Sun, H., Gao, M., Yu, N., Zhang, Q., et al. (2022). Microbial co-occurrence network topological properties link with reactor parameters and reveal importance of low-abundance genera. *NPJ Biofilms Microbiomes* 8:3. doi: 10.1038/s41522-021-00263-y

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi: 10.1101/gr.112730.110

Hernández-Bermejo, B., and Fairén, V. (1997). Lotka-Volterra representation of general nonlinear systems. *Math. Biosci.* 140, 1–32. doi: 10.1016/S0025-5564(96)00131-9

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x

Jackson, D. A. (1997). Compositional data in community ecology: The paradigm or peril of proportions? *Ecology* 78, 929–940. doi: 10.1890/0012-96581997078[0929:CDICET]2.0.CO;2

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11:e1004226. doi: 10.1371/journal.pcbi.1004226

Li, H.-Y., Wang, H., Wang, H.-T., Xin, P.-Y., Xu, X.-H., Ma, Y., et al. (2018). The chemodiversity of paddy soil dissolved organic matter correlates with microbial community at continental scales. *Microbiome* 6:187. doi: 10.1186/s40168-018-0561-x

Li, X.-M., Chen, Q.-L., He, C., Shi, Q., Chen, S.-C., Reid, B. J., et al. (2019). Organic carbon amendments affect the chemodiversity of soil dissolved organic matter and its associations with soil microbial communities. *Environ. Sci. Technol.* 53, 50–59. doi: 10.1021/acs.est.8b04673

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., et al. (2015). Determinants of community structure in the global plankton interactome. *Science* 348:1262073. doi: 10.1126/science.1262073

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., et al. (2014). Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* 24, 813–821. doi: 10.1016/j.cub.2014.02.050

Lv, X., Zhao, K., Xue, R., Liu, Y., Xu, J., and Ma, B. (2019). Strengthening insights in microbial ecological networks from theory to applications. *mSystems* 4, e124–e119. doi: 10.1128/mSystems.00124-19

Ma, B., Wang, H., Dsouza, M., Lou, J., He, Y., Dai, Z., et al. (2016). Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *ISME J.* 10, 1891–1901. doi: 10.1038/ismej.2015.261

Ma, B., Wang, Y., Ye, S., Liu, S., Stirling, E., Gilbert, J. A., et al. (2020). Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. *Microbiome* 8:82. doi: 10.1186/s40168-020-00857-2

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Mellin, W. (1957). Work with new electronic 'brains' opens field for army math experts. *Hammond Times* 10:66.

Parry, R. M., Jones, W., Stokes, T. H., Phan, J. H., Moffitt, R. A., Fang, H., et al. (2010). K-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenomics J.* 10, 292–309. doi: 10.1038/tpj.2010.56

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., et al. (2011). Detecting novel associations in large data sets. *Science* 334, 1518–1524. doi: 10.1126/science.1205438

Röttjers, L., and Faust, K. (2018). From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiol. Rev.* 42, 761–780. doi: 10.1093/femsre/fuy030

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Curr. Protoc. Mol. Biol.* 122:e59. doi: 10.1002/cpmb.59

Tjarnberg, A., Mahmood, O., Jackson, C. A., Saldi, G.-A., Cho, K., Christiaen, L. A., et al. (2021). Optimal tuning of weighted kNN- And diffusion-based methods for denoising single cell genomics data. *PLoS Comput. Biol.* 17:e1008569. doi: 10.1371/JOURNAL.PCBI.1008569

Torgo, L. (2016). *Data mining with R: Learning with case studies, second edition*. London: Chapman and Hall/CRC, 137.

Trosvik, P., Rudi, K., Straetkvern, K. O., Jakobsen, K. S., Naes, T., and Stenseth, N. C. (2010). Web of ecological interactions in an experimental gut microbiota: Ecological web in microbiota. *Environ. Microbiol.* 12, 2677–2687. doi: 10.1111/j.1462-2920.2010.02236.x

Tu, Q., Yan, Q., Deng, Y., Michaletz, S. T., Buzzard, V., Weiser, M. D., et al. (2020). Biogeographic patterns of microbial co-occurrence ecological networks in six American forests. *Soil Biol. Biochem.* 148:107897. doi: 10.1016/j.soilbio.2020.107897

van den Boogaart, K. G., and Tolosana-Delgado, R. (2008). " compositions": A unified R package to analyze compositional data. *Comput. Geosci.* 34, 320–338. doi: 10.1016/j.cageo.2006.11.017

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi: 10.1016/j.tig.2014.07.001

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681. doi: 10.1038/ismej.2015.235

Xia, L. C., Ai, D., Cram, J., Fuhrman, J. A., and Sun, F. (2013). Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* 29, 230–237. doi: 10.1093/bioinformatics/bts668

Xie, K., Liu, K., Alvi, H. A. K., Chen, Y., Wang, S., and Yuan, X. K. N. N. C. N. V. (2021). A K-nearest neighbor based method for detection of copy number variations using NGS data. *Front. Cell Dev. Biol.* 9:796249. doi: 10.3389/fcell.2021.796249

Yang, Y., Wang, X., Xie, K., Zhu, C., Chen, N., and Chen, T. (2021). kLDM: Inferring multiple metagenomic association networks based on the variation of environmental factors. *Genomics Proteomics Bioinform.* 19, 834–847. doi: 10.1016/j.gpb.2020.06.015

Zhao, Z., Gonsior, M., Schmitt-Kopplin, P., Zhan, Y., Zhang, R., Jiao, N., et al. (2019). Microbial transformation of virus-induced dissolved organic matter from picocyanobacteria: Coupling of bacterial diversity and DOM chemodiversity. *ISME J.* 13, 2551–2565. doi: 10.1038/s41396-019-0449-1

Zhou, J., Deng, Y., Luo, F., He, Z., Tu, Q., Zhi, X., et al. (2010). Functional molecular ecological networks. *mBio* 1, e00169–e00169. doi: 10.1128/mBio.00169-10