Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, **2021**, *22*, 267-290

REVIEW ARTICLE

Modeling Microbial Community Networks: Methods and Tools

Marco Cappellato[1], Giacomo Baruzzo[1], Ilaria Patuzzi[2] and Barbara Di Camillo[1,*]

[1]*Department of Information Engineering, University of Padova, Padova, Italy;* [2]*Research & Development, Eubiome Srl, Padova, Italy*

ARTICLE HISTORY

Received: June 29, 2020
Revised: July 22, 2020
Accepted: July 29, 2020

DOI:
10.2174/1389202921999200905133146

Abstract: In the current research landscape, microbiota composition studies are of extreme interest, since it has been widely shown that resident microorganisms affect and shape the ecological niche they inhabit. This complex micro-world is characterized by different types of interactions. Understanding these relationships provides a useful tool for decoding the causes and effects of communities' organizations. Next-Generation Sequencing technologies allow to reconstruct the internal composition of the whole microbial community present in a sample. Sequencing data can then be investigated through statistical and computational method coming from network theory to infer the network of interactions among microbial species.

Since there are several network inference approaches in the literature, in this paper we tried to shed light on their main characteristics and challenges, providing a useful tool not only to those interested in using the methods, but also to those who want to develop new ones. In addition, we focused on the frameworks used to produce synthetic data, starting from the simulation of network structures up to their integration with abundance models, with the aim of clarifying the key points of the entire generative process.

Keywords: Microbiota, microbiota analysis, microbial interactions, network inference, relationship models, synthetic count data.

1. INTRODUCTION

The microbiota is the set of population microorganisms such as bacteria, archaea, viruses and unicellular fungi that characterize a specific environment and ecosystem, such as human gut or saliva, environmental microecological niches (such as animals and vegetables), water or soil. This complex micro-world is characterized by several interactions that determine its nature. Mainly, two types of relationships are observed in a bacterial community: microbial and ecological.

Microbial interactions refer to a number of different kinds of relationships that occur between different taxa, that can bring a positive (+), negative (-) or neutral (0) effect for each of the two taxa involved:

- (+, +) Mutualism is a common benefit relationship established between biological species. In some cases, microorganisms cooperate in carrying out the same physiological function, in others, they exchange the metabolic products for the mutual sustenance, as in syntrophy or cross-feeding case.
- (+, -) Parasitism and Predation are two phenomena related to the survival of an organism subject to host or prey's life, respectively. Parasitic associations can be observed between host and bacteriophage virus that infects bacteria and archaea for the replication purpose. *Bdellovibrio* is an example of a predator that attacks other bacteria and feeds on the biomolecules produced.
- (+, 0) Commensalism occurs when a member of the population benefits from the presence of others, without any advantage or harm to them. In the biodegradation process, commensal bacteria feed on others' products.
- (-, 0) Amensalism describes a relationship in which an organism harms another component of the community without positive or negative implications for itself. This type of interaction can happen when the metabolic products of one species alter the environment, making it adverse towards another species.
- (-, -) Competition takes place when two species inhabiting the same environment vie for a common resource. If the resource is in limited supply and the species niches totally overlap, the weaker competitor will be pushed toward extinction or will undergo a gradual shift toward a different ecological niche. This latter phenomenon is summarized in Gause's competitive exclusion principle.
- (0, 0) Neutralism indicates the absence or irrelevance of relationships.

Ecological interactions, on the other hand, occur between taxa and the environment. As an example, in the human gut, host cells live in symbiosis with the microbiota. Molecular signals from bacteria promote many physiologi-

*Address correspondence to this author at the Department of Information Engineering, University of Padova, Padova, Italy;
E-mail: barbara.dicamillo@unipd.it

1389-2029/21 $65.00+.00

cal functions and, in the other direction, host cells secrete metabolites that influence the microbial ecosystem [1, 2]. Furthermore, age, lifestyle and diet, influence the local environment through hormone and metabolite secretion, thus, in turn, changing the environment in which bacteria live [3]. Considering the entire time span from pregnancy to birth up to the first months of life, there are different factors that influence microorganisms transfer between the mother and the children: prenatal factors, such as mother's diet and lifestyle; the delivery mode, *i.e*., Caesarean or natural birth; the first contacts that occur between the mother and the skin or the mucous membrane surfaces of the new-born child [4]; the type of supply, *i.e*., breastfeeding or formula-fed [5]. Recently, airways or lung microbiota has also aroused interest in the study of the causes related to environmental exposures [6] or smoking habits [7]. Hanski *et al.* [8] have shown results regarding the impact of biodiversity in the natural landscape on the skin microbiota in relation to the allergic predisposition.

Deciphering the complex networks of associations among microbial communities, and between them and the environment, tries to shed light on questions like: "how do microbes interact?", "how does the environment change the microbial population?", "what is the effect of external perturbations on microbial dynamics?". These are the main reasons that guide the study of microbial ecosystems, where the answers are sought by exploiting the information contained in sequencing data.

The study of microbe-microbe, environment-microbes and host-microbes interactions is extremely important to understand community organization in relation to the factors that determine biodiversity. In addition, microbial networks could provide a powerful predictive and therapeutic tool in the field of human health. Information on how the community is modified due to an introduced stimulus could allow, for example, to act on the network by means of probiotics to restore the correct composition of the community [9].

To investigate the complex bacterial communities' landscape, network theory provides useful tools [10]. Graphs are frequently used in molecular biology to represent the relationships between entities, the nodes of the network, where edges correspond to some interactions between them. Edges may be directed when they link two nodes asymmetrically, from one to the other, or undirected, when they link two nodes symmetrically. Edges can be weighted if there is a strength score associated with the link between the nodes. Indeed, biological networks describe relationships that are established between different actors involved in physiological processes, such as proteins, genes or biomolecules (Table **1** for some examples).

In a microbial community landscape, the nodes of the network represent different members, while edges correspond to some of the previously described interactions that occur between them. The presence of a relationship between taxa is inferred from taxa abundance values, using different reverse engineering approaches [11-14] stemming from network theory. Microbial networks can also contain nodes related to ecological or physiological variables that present significant association patterns with the abundance values of microorganisms.

In this review, we will focus on the microbes-microbes interaction networks that shape the microbial community. The aim is to give an overview of the literature of microbial networks reconstruction, providing useful information not only for analysts looking for available methods, but also for researchers interested in developing new ones. In section 2, we introduce one of the most popular sequencing techniques used to produce abundance data, 16S rRNA gene sequencing, and into well-known standard analysis tools; in section 3, we summarize the main literature methods for the reconstruction of microbial interaction networks; in section 4, 5 and 6, we consider the need for benchmarking studies that evaluate the performance of the developed methods, the simulation frameworks used to generate the gold standard and the assessment scores. In section 7, we discuss the limits and challenges still open in the field.

2. 16S SEQUENCING AND STANDARD ANALYSIS METHODS

In the current research landscape, there are two main sequencing techniques used to carry out studies on microbial communities in terms of species abundance: the Whole Genome Sequencing (WGS) and the targeted amplicon sequencing of 16S ribosomal RNA (16S rDNA-seq). In this section, we will focus on the latter, which is still the most commonly used, since it is much cheaper than WGS.

16S rRNA gene encodes for a small subunit of the prokaryotic ribosome. The 16S rRNA gene contains highly conserved regions, mainly shared by all the species, and hypervariable regions, which are characteristic for each phylogenetic lineage [15] and that are used to discriminate and identify the different community members in the sample. Since the 16S rRNA gene is characteristic of prokaryotes, only microorganisms belonging to bacteria and archaea kingdoms can be detected by this sequencing technique.

Conserved regions, which flank the hypervariable ones, are used as binding sites for primers during the amplification phase preceding 16S rDNA-seq. The choice of primers is a crucial point for the correct characterization of the bacterial community [16] since one should maximize the balance between efficiency and specificity in the targeted amplicon amplification and maximize the coverage, in terms of the fraction of all bacterial 16S sequences matched by at least one primer pair. In the literature, several papers address the identification of primers that show better resolution in terms of taxonomic profiles [17-19]. An alternative is to use software for 16S primer design and optimization like SPYDER [20] or mopo16S [21], that might also account for possible variations in the conserved regions [22].

Targeted amplification produces a large number of 16S rRNA fragments, called amplicons, which are then sequenced using Next-Generation Sequencing (NGS) platforms, such as Illumina or Ion Torrent [23]. These widely used technologies allow deep, high throughput, in-parallel DNA sequencing, and produce a large amount of data in a timely and cost-effective fashion. Indeed, NGS can produce millions of short sequences, called reads, that can be used to detect the presence and abundance of different taxa in the original population. The obtained reads are preprocessed using different software tools like QIIME2 [24], Mothur

Table 1. Some examples of biological networks.

Biological Networks	Nodes	Edges' Meaning
Gene Co-Expression Networks	Genes	Co-expression level
Gene Regulatory Networks	Transcription factors and binding sites or genes and their regulators	Regulatory relationships
Metabolic Networks	Metabolites	Biochemical reactions
Microbial Interaction network	Taxa (and ecological or physiological variables)	Microbe-microbe, (environment-microbes and host-microbes) interactions
Protein-Protein interaction network (PPI)	Proteins	Interactions involving the activation of a molecular and cellular mechanism
Sequence Similarity Networks (SSNs)	Proteins or genes sequences	The similarity in the amino acid or nucleotide chain

[25] or USEARCH [26]. Generally, these tools incorporate several methods for denoising and quality filtering, to discard short and low-quality base pairs sequences. Furthermore, different algorithms perform reads clustering in Operational Taxonomic Units (OTUs), namely, clusters of organisms usually grouped by DNA sequence similarity. Clustering methods of more recent publication allow the reconstruction of the so-called Amplicon Sequence Variants (ASVs), a higher resolution version of the classic OTUs obtained by clustering sequences that differ by at least one single nucleotide. ASV methods infer the biological sequences in each sample using amplification and sequencing error models and can distinguish sequence variants differing by as little as one nucleotide, without setting an arbitrary dissimilarity threshold to cluster sequences as previously done for OTUs. The last step of all preprocessing pipeline is the taxonomy annotation assignment to each OTU/ASV by a classifier trained on a reference database (such as RDP classifier [27]), or read alignment to potential target sequences (as VSEARCH [28]).

The final output of the whole process is the so-called OTU/ASV table, where each element contains the number of times a read coming from a given sample was found to belong to a particular OTU/ASV, and the related taxonomy, which describes and characterizes each OTU/ASV at the deepest possible taxonomic level. All the processing steps require technological or methodological choices that have an impact on the final tables and, consequently, on the following analysis. In the literature, there are studies that try to find the best tool or tool configuration for each preprocessing step [29, 30], providing useful information to guide the choice of analysts. Unfortunately, there is currently no standard global preprocessing pipeline defined and the development of new methods and technologies is still in progress.

Sequencing count matrices have peculiar characteristics that are linked to both biological and technical features. Firstly, the limited obtainable sequencing depth together with the sample harvesting makes 16S rDNA-seq count data highly sparse (70-95% of null values). Secondly, count data do not reflect absolute abundance, but rather portions of a whole (the sequencing depth), that reflect the proportion of individuals belonging to a specific taxonomic group [31, 32]. Therefore, an increase in the absolute abundance of a community member causes a decrease in the other entries, an artifact called compositional bias. Finally, the total of counts experimentally obtained during the sequencing run usually differs from sample to sample.

Several methods have been proposed in the literature to normalize, correct for sampling and compositional biases and account for technical variability, in order to make the samples comparable. Considering the compositional nature of the data, the log ratios transformations are used to transform abundance value from the Simplex space (proportion space that keeps the relative information of counts) to the Real space. The most known transformations are the additive log-ratio (alr) and the centered log-ratio (clr), proposed by Aitchison [33], although another transformation, the isometric log-ratio (ilr) [34], is available. In short, alr transforms each sample with the logarithm of the ratio between each relative component (OTU/ASV) and a reference feature. Conversely, clr uses the geometric mean of the relative abundances of the sample as a denominator. The orthonormal bases of the clr-plane allow to retain the metric properties in mapping data in real Cartesian coordinates. Since the transformations involve the logarithm function, the presence of many null values represents a problem. To overcome this obstacle, some studies add a small constant amount to the data matrix or to just zero values, the pseudo-count. However, this approach alters the internal proportions by changing the relationships between the compositional parts. Martin-Fernandez *et al.* [35] developed a Bayesian framework to impute zero values in compositional data. The method assumes that count values follow a multinomial distribution with Dirichlet distribution as conditional prior. All null values are replaced with the *a posteriori* estimate obtained from the model, while the positive counts are multiplied by a quantity dependent on the replaced values, the so-called multiplicative replacement strategy. Therefore, this method reconstructs lost values by maintaining the relationships between the non-null components. In the R package zCompositions [36], several approaches that use different prior distribution or replacement strategies are implemented.

Hron *et al.* [37], instead, proposed two different alternatives that are implemented in the robComposition R package [38]. The first, *k*-nearest neighbor (knn) imputation, considers a sample with zero values and finds a set of other more similar to it, based on compositional distance measure. Then, the replacement step involves the information from the neighbors to replace the missing values. The second, called iterative model-based imputation, models the missing values as a regression of all the other non-null. Since regression cannot be applied to compositional data, the authors transform data into the simplex space using the ilr on the dataset where zeros are replaced using knn. Finally, the regression is performed on the transformed data by distinguishing the dependent variables based on their value in the original data.

As an alternative, there are some methods that deal with both sparsity and compositional bias. GMPR [39] is an inter-sample normalization method designed for microbiota sequencing data. In brief, the algorithm calculates for each pair of samples in the dataset, the median of the abundance ratios for taxa with non-zero values. Finally, the size factor for each sample is obtained from the geometric mean of the median ratios for all the other samples. Moreover, Wrench [40] exploits a Bayes normalization approach. Basically, after choosing the reference vector as the average proportions present in the dataset, the ratios of the taxa proportions pairs are modeled using a hurdle log-normal model. The identification of the model allows to estimate the true taxa-wise proportion ratios that are linked to the compositional scaling factors used to normalize the taxa between samples and intra-sample.

After the count data are normalized, statistical analysis is carried out. In most cases, alpha and beta diversity are computed to investigate and quantify the compositional complexity of a community within a sample and the variability between samples, respectively. There are several formulations of these metrics with different properties [41]. In many studies, contrary to the classic evaluation of the difference between abundance profiles, beta diversity is also evaluated in terms of the distance between the samples. The most used metrics belong to the UniFrac family [42] that considers phylogenetic distances between observed organisms in the computation. Principal Coordinate Analysis (PCoA) or Non-Metric Multidimensional Scaling (NMDS) is often used to visualize samples in a lower dimensional space and to identify subgroups of samples visually. These dimensionality reduction techniques are applied to dissimilarity or distance matrices. In addition, Lê Cao *et al.* [43] proposed the use of Principal Component Analysis directly on the clr or ilr transformed data as an alternative to the aforementioned techniques.

Differential Abundance (DA) testing is the first step of the downstream analysis. DA purpose is to quantify differences observed in the microbiota composition of groups of subjects identified by the covariates of interest. These methods test taxa abundance between two sets of subjects identified by a binary variable (*e.g.*, healthy and sick), thus finding the main culprits of differences. The most used statistical methods are mainly divided into nonparametric tests, such as the Wilcoxon rank-sum or Kruskal-Wallis test, and model-based approach taking into account the characteristic distribution of OTU/ASV samples, *e.g.* ALDEx2 [44] ANCOM [45], mixMC [43], BhGLM [46]. The application of DA methods has contributed to significant advances in identifying the role of taxa in disease status [47]. Despite this, it is necessary to further improve current methods performance, since some have shown poor control of false positive rate and low power to detect differences, especially for taxa with low abundance [48]. In the literature, most studies focus on alpha and beta diversity or DA. To get an idea of the number of articles in the literature dealing with downstream analyses, we queried the PubMed (https://www.ncbi.nlm.nih. gov/pubmed/) database searching for keywords related to the microbiota field and the name of each investigation. Fig. (**1**) shows the counts obtained from this research, starting from the rise of NGS technology. The trend relating to the analysis of the diversity in the sample composition (in square and triangle dots) has been continuously growing for several years. Although these analyses are part of the standard procedure in the microbiota field, the search for new, more performing methodologies is still active, as mentioned above. On the contrary, network analysis (circle dots) is a recent research topic and, given the remarkable insights it has brought, it represents the future of studies in this landscape. Probably, as usually happens with the introduction of new methods, it takes time for the user community to recognize their value and potential. Certainly, the complexity of the developed methods, the lack of numerous experimentally validated results, or the challenges still opened represent an obstacle to the spread of network inference methods. Aware of innovation and the potential of knowledge of interaction mechanisms, we propose this review in order to provide a useful tool to those interested in applying or developing these methods.

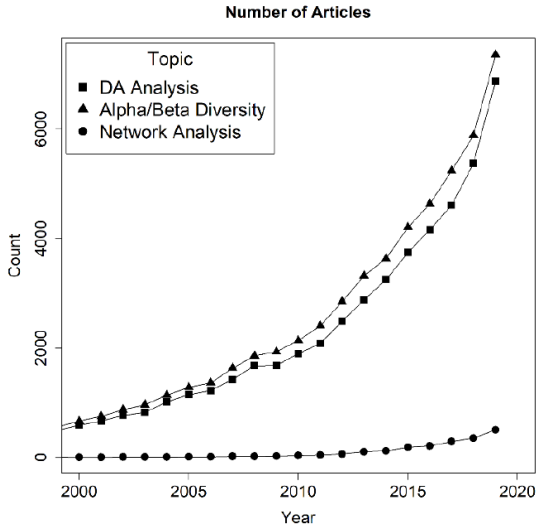

Fig. (1). Number of articles in the literature relating to the main downstream analyses. Counts are obtained by querying the PubMed database searching in the title/abstract: (network OR network analysis OR microbial interactions) AND (16S OR microbiota OR microbial communities) for Network Analysis; (Differential OR abundance OR analysis OR statistical) AND (16S OR microbiota OR microbial community) for DA Analysis; (Alpha OR beta OR diversity OR analysis) AND (16S OR microbiota OR microbial community) for Alpha/Beta Diversity.

3. INFERRING MICROBIAL INTERACTION NETWORK

The microbiota is a complex system in which a large number of variables act in concert, contributing to the community's biodiversity. Generally, two main categories of experiments can be found in the literature: cross-sectional and longitudinal studies.

Cross-sectional studies involve multiple samples at a specific time point. In this case, a snapshot of the bacterial community of several subjects is available, so any consideration of dynamics is impossible. The networks inferred on this type of dataset are usually indirect graphs that represent the pairwise relationships between taxa abundances. The basic idea is that two taxa co-occur when they have similar abundance values according to the defined metric, and therefore the arc between the two is present. Thus, from cross-sectional studies, static co-occurrence network can be explored.

Conversely, longitudinal studies consist of multiple measurements at different time points. Therefore, it is possible to study the evolution of bacterial interactions and consider pre-post relationships as putative cause-effect relationships, also correlated with external factors that perturb the system. Consequently, the inferred networks are generally directed, and represent a cause-effect relationship in the time window considered.

In this section of the review, we will present a number of statistical methods presented in the literature for the characterization of static and dynamic networks. Table **2** summarizes all the methods commented in this review, also providing useful indications regarding the availability of the software.

In order to facilitate the reading, we first introduce terms and concepts that will be widely used in the following sections. With i and j, we will indicate two generic taxa related to the row indices i and j of the OTU/ASV table. The final goal of each method presented in this review is the inference of a network structure that can be represented by an adjacency matrix θ in which element $\theta_{ij} = 0$ if there is no edge connecting the two nodes i and j, $\theta_{ij} = 1$ otherwise. To obtain θ and draw the interaction graph, some methods exploit the Covariance matrix Γ, where diagonal elements Γ_{ii} correspond to the variance of the i-th taxa abundance, while off-diagonal values Γ_{ij} are covariances between i and j, *i.e.*, the expected value of the products of their distances from the average. Alternatively, others use the Precision matrix Ω, that is the inverse of the covariance matrix previously defined.

3.1. Methods based on Pairwise Microbial Relationships

The first network inference methods proposed in the literature are non-parametric, as they do not make any assumptions about data distribution. Pairwise scores are defined with correlation metrics, such as Pearson or Spearman, or similarity and dissimilarity measures, such as Bray–Curtis or Kullback–Leibler [49]. The null model, *i.e.*, the probability distribution of the metric in a random situation, of the chosen metric is calculated on the OTU/ASV matrix with rows independently permuted a high number of times. Then, the p-value is defined as the probability that the correlation/similarity value observed for each taxa pair is greater than the one obtained by chance. In general, the correction for multiple tests, such as Bonferroni or False Discovery Rate, is used to correct the p-values, ensuring global False Positive Rate control. In a correlation-based approach, the sign of the interaction (positive or negative correlation) is also available.

Another metric used is Mutual Information (MI), a dimensionless number that quantifies mutual dependence between two random variables. MI can be interpreted as the expected reduction in uncertainty regarding one variable, given the observation of the other. In particular, in a study by Reshef *et al.* [50], the Maximal Information Coefficient (MIC) was used. MIC has the following properties: generality, *i.e.*, it can infer different types of non-linear relationships with sufficient sample size, and equitability, *i.e.*, same noisy associations give a similar score. Basically, for each node pair (i, j), MIC(i,j) is the maximum value of the mutual information overall i-by-j grid (up to a maximal grid resolution) normalized between grids of different dimensions. The network is not directed, but the edges identify nonlinear relationships that cannot be determined by linear correlation.

3.1.1. Ensemble Approach for Pairwise Metrics

In the literature, there are several ways to define the pairwise association from two variables, and the resulting networks may present some structural variations. Thus, some authors proposed to use ensemble approaches based on merging several inferred networks obtained from different metrics.

CCREPE [51] combines a linear model, 2 correlation metrics (Spearman and Pearson) and 2 similarity metrics (Bray–Curtis distance and Kullback–Leibler divergence). To determine the significance of the scores, the authors developed the ReBoot method. The aim of this procedure is to build a null distribution that reflects correlation caused only by compositional bias and a bootstrap distribution as a confidence interval for the observed value of correlation from data. More in detail, for each pair of taxa, the null distribution is obtained iterating the following steps: permute the relative abundances, normalize the data and finally calculate the correlation between them. On the other hand, bootstrap distribution is constructed by calculating several times the correlation between each pair of taxa on a resampled dataset. A pooled variance Z-test is used to assess differences between the two distributions, defining the p-value for the associations. For Bray–Curtis and Kullback–Leibler metrics, the authors calculate p-values comparing the bootstrap distribution with a null point value calculated with the permutation approach. Then, all the identified networks are merged with the Simes p-value combination method [52], and the edges are then corrected using the Benjamini-Hochberg-Yekutieli procedure. The final network represents all significant co-occurrence and co-exclusion links between taxa pairs. Co-exclusion refers to nonlinear patterns that underlie an adversity relationship between micro-organisms.

Another ensemble approach is CoNet [53]. The underlying rationale of this tool is that the identification of a network coming from the intersection of different methods reduces false-positive edge calls. In CoNet, there are various

Table 2. Table of methods covered in this review with indications for code availability, base approach and type of data.

Method	Software	Source	Approach	Developed for
Meta-network	Python code	http://www.microbioinformatics.org/software/Meta-Network.htm	Rule Mining Association	Cross-sectional
MDiNE	R package	https://github.com/kevinmcgregor/mdine	Bayesian Graphical Model	Cross-sectional
SPRING	-	-	Graphical Model	Cross-sectional
TIME	web app	https://web.rniapps.net/time/	Granger Causality	Time Series
MPLasso	R package	https://github.com/ChiehLo/MPLasso_Rpackage	Graphical Model	Cross-sectional
gCoda	R code	https://github.com/huayingfang/gCoda	Graphical Model	Cross-sectional
BAnOCC	R package	https://bitbucket.org/biobakery/banocc/src/master/	Bayesian Graphical Model	Cross-sectional
MTPLasso	-	-	gLV	Time Series
Ridenhour *et al.*	R code	https://www.nature.com/articles/ismej2017107#Sec10	ARIMA	Time Series
cooccur	R package	https://cran.r-project.org/web/packages/cooccur/index.html	Probability Theory	Cross-sectional
CoNet	Cytoscape plugin	http://apps.cytoscape.org/apps/conet	Ensemble Pairwise Metrics	Cross-sectional
metaMIS	Matlab (stand-alone GUI)	https://sourceforge.net/projects/metamis/	gLV	Time Series
SPIEC-EASI	R package	https://github.com/zdk123/SpiecEasi#analysis-of-american-gut-data	Graphical Model	Cross-sectional
REBACCA	R code	http://faculty.wcas.northwestern.edu/~hji403/REBACCA.htm	Covariance Estimation	Cross-sectional
CCLasso	R code	https://github.com/huayingfang/CCLasso	Covariance Estimation	Cross-sectional
RMN	-	-	rule-based algorithm	Time Series
LIMITS	Mathematica	http://physics.bu.edu/~pankajm/Code/code.html	gLV	Time Series
eLSA	Python package	https://bio.tools/elsa	LSA	Time Series
SparCC	Python package[1]	https://bitbucket.org/yonatanf/sparcc	Compositional Correlation	Cross-sectional
MENAP	Web App	http://ieg2.ou.edu/MENA	Random Matrix Theory	Cross-sectional
CCREPE	R package[2]	http://www.bioconductor.org/packages/release/bioc/html/ccrepe.html	Ensemble Pairwise Metrics	Cross-sectional
MIC	MINEv2.jar (Java), minepy (Python-Matlab), minerva (R)	http://www.exploredata.net/Downloads/MINE-Application	Pairwise Relationship	Cross-sectional

1. There is also an R implementation in SpiecEasi, gCoda and CCLasso package, 2. This information comes from the SpiecEasi paper.

correlations (*i.e.*, Pearson, Spearman, Kendall), similarity (*i.e.*, MI, Steinhaus, distance correlation) or dissimilarity (*i.e.*, Kullback-Leibler, Euclidean, Bray-Curtis, Jensen-Shannon) metrics implemented, that the user can choose to combine. The significance level of each pairwise comparison can be assigned with a permutation test or with the above-mentioned ReBoot procedure. The strategy recommended by the authors to evaluate the final score of each network edge is the Brown p-value merging method [54], because it weighs the results considering the dependence between the different metrics used.

3.1.2. Association Metrics Based on Network Topology

Recently, Yang *et al.* [55] proposed Meta-Network, a workflow based on association rule mining. The method transforms abundance data into a binary presence-absence matrix on which it calculates the probability of each (i,j) co-occurrence as the percentage of co-existence with respect to the total number of samples. Subsequently, the Pearson correlation between the abundance profile of (i,j) is calculated on pairs that exceed a predefined probability threshold. In order to identify the possible indirect interactions between nodes, for example, when taxa share a functional scheme, Meta-Network uses the functional similarity (FS) weight measure on the correlation-based network [56]. In short, FS-weight assigns a functional similarity score to each pair (i,j) based on the network topology considering the first and second level neighbors. Again, a threshold on the calculated weights filters network's arcs and draws new ones relating to indirect associations. In addition, the Meta-network framework contains another method based on Part Mutual Information (PMI), that takes into account nonlinear relationships. The algorithm starts by building a zero-order network using MI as an association metric. Then, partial information of each (i,j) pair conditioned by the abundance of N neighbors is calculated. Path Consistency Algorithm (PCA) with a defined correlation threshold is applied to adjust edges distribution in the network. PCA-PMI algorithm is applied iteratively considering an increasing value of N until the final network reaches convergence, *i.e.*, the increase in the number of considered neighbors does not change the topology.

3.1.3. Compositional Correlation Approach

Different methods are based on correlation to account for pairwise associations between taxa. However, since the abundances of the count matrix are compositional data, the correlation calculation can lead to reconstruct erroneous edges unless data are not previously transformed, so to lie on a Euclidean space. SparCC [57] is a proposed method to overcome the compositional effect in calculating the correlation. The authors refer to Aitchison's theory, which defines the variance of log-ratios as a metric to quantify the dependence between two compositional variables. In particular, the developed algorithm is based on a relationship between the variation matrix, *i.e.*, variance of the component log fractions in all the samples, and the unknown Γ of the true log-transformed variables. In order to infer Linear Pearson correlations between the log-transformed variables, some approximations are necessary. SparCC works under the hypothesis that the number of different taxa in the dataset must be large, and the number of strongly correlated variables is low. However, the authors demonstrated in a simulated context that the method is robust even when the hypothesis of the sparsity of Γ is not fully verified. In addition, they found that there is a relationship between alpha diversity, calculated with Shannon effective number (n_{eff}) [58], and compositional bias. For this reason, the method is recommended in datasets with low diversity (at least $n_{eff} = 50$).

3.2. Methods Based on Multivariate Approach

In this section, we will deal with different methods that have the aim of estimating the Covariance matrix Γ or its inverse, the Precision matrix Ω. The estimate of these entities takes place through the formulation of an optimization problem. Essentially, the desired model parameters are those that minimize an objective function (also known as loss function). In the case of *Least Absolute Shrinkage and Selection Operator* (Lasso) estimation, the loss function is defined as the difference between the expected values obtained by the model and the real ones. Alternatively, if a maximum likelihood estimator is used, the problem becomes the maximization of the probability distribution of the observed values given the model parameters, the so-called likelihood function. Usually, the optimization problem is formulated considering an element of penalization on the number of model parameters controlled by a regularization value.

3.2.1. Lasso-based Covariance Estimation

Regression-based methods build a linear system that relates the Γ matrix of the log-transformed relative abundances with the covariance matrix of the unknown real abundances, Γ'. The final estimate of the true Γ' is obtained by solving a Lasso problem with a regularization element that controls the sparsity of the associations.

CCLasso [59], for example, combines the loss function with an l_1-penalty on the off-diagonal components of the log-basis covariance matrix to take into account the sparsity. The authors have shown that the estimated matrix is positive-definite with the elements included in the range [-1,1], as opposed to SparCC, which does not guarantee this property.

REBACCA [60] uses the same regularization method but sets a different objective function. Furthermore, the two methods use distinct approaches to handle parameter tuning. CCLasso uses a K-fold cross-validation on the loss function choosing the parameter value that minimizes the mean K-fold error. Instead, REBACCA utilizes a stability resampling method [61], which regulates the number of selected variables taking into account their selection probability, based on the independent application of Lasso on two datasets deriving from the random split of the original data.

3.2.2. Graphical Lasso Models

Some network inference tools are based on the estimate of the undirected graphical model from the data. The structure of the interaction graph is reconstructed, starting from the concept of conditional independence between the variables involved in the model. Taxa i and j are conditionally independent when the abundance value of i, compared to all the other taxa in the dataset (X_{-i-j}), does not add information to the probability of occurrence of j, and vice versa. In the special case of the multivariate normal distribution of the variables, a null partial correlation between i and j corre-

sponds to their conditional independence $i \perp j | X_{-i-j}$. Remembering that $\Omega = \Gamma^{-1}$, the estimate of variables that satisfy the previous definition can be carried out based on precision matrix elements. If $\Omega_{ij} = 0$, it means that the partial correlation between i and j is zero, and the variables are conditionally independent. Consequently, the non-null elements of Ω identify the network structure θ, and therefore the conditionally dependent pairs of nodes. In the literature, there are methods that directly estimate the entire precision matrix based on graphical lasso (Glasso) [62], and those that try to estimate the individual null components Ω_{ij} by neighborhood selection of Meinshausen and Bühlmann (MB) [63].

SPIEC-EASI [64] is one of the methods based on graphical models to infer the structure of the underlying network based on conditional independence. The Γ matrix of the clr transformed data is related to the covariance matrix of the true log-transformed abundances Γ'. When the number of taxa is greater than the number of subjects, the estimator of Γ becomes an approximation of the estimator of Γ'. Moreover, in this case, the sparsity of the data represents a problem for the network identification and the hypothesis of graph sparsity is taken into consideration. The authors estimate the true abundances Ω using a penalized maximum likelihood method (Glasso problem). The regularization parameter, which is linked to the edges' sparsity, is identified by means of the StARS selection algorithm [65].

Similar to SPEAC-EASI, another method based on estimating the sparse inverse covariance from penalized maximum likelihood is gCoda [66]. There are two major assumptions characterizing this approach: the logarithm of the real absolute abundances is derived from the multivariate normal distribution and the edges density is sparse. The distribution hypothesis leads to the formulation of an optimization problem different from SPIEC-EASI, which is solved through a Majorization-Minimization algorithm developed by the authors.

Another graph-based method is MPLasso [67]. The Glasso optimization problem to estimate Ω requires the assumption of edge sparsity regulated by tuning parameters. On the other hand, in MPLasso, regularization takes place through the penalty on the l_1 component (the sum of the inverse correlation matrix elements) of the objective function in which the co-occurrence matrix prior (P) is also considered. The *a priori* knowledge on associations between taxa is extracted from the literature through a text mining algorithm. The authors, in reference to the @MInter method [68], access the PubMed database to perform queries. In particular, for each (i,j) pair, the number of papers which contain in the abstract only taxa i, only taxa j, both and neither, are obtained. Then, on the contingency matrix identified by the previous values, the Fisher's exact test with Bonferroni correction is carried out to find the prior probability of associations. Finally, Bayesian information criterion (BIC) in Gaussian Graphical models context is used to choose the best penalty parameter, since it considers a balance between the maximized value of the likelihood function and the corresponding edge number in relation to samples size.

Compared to the estimate of the global conditional independence performed by Glasso formulation, the MB method exploits the relationship between the partial correlations and the coefficients of the linear regression in which each node is the dependent variable of the other predictor nodes.

A variant of the previously described Glasso method exploiting the MB approach is also implemented in SPIEC-EASI. The algorithm defines each clr-transformed taxon as the linear regression of the others, and requires solving a penalized regression problem (Lasso). Then, for each node, the non-zero coefficients identify the set of its local neighbors, and the final edges are selected through the union or intersection of all these sets. The regularization parameter, which controls the sparsity of local associations, is still detected by the StARS algorithm.

SPRING [69] is a Semi-Parametric Rank-based approach based on the truncated Gaussian copula model [70] to estimate the true underlying correlation matrix for zero-inflated count data. In short, the authors rewrite the MB optimization problem by replacing the sample correlation matrix with Semi-Parametric Rank-based (SPR) correlation estimator. They use the same stability-based method (StARS) used in SPIEC-EASI to identify the tuning parameter. In addition, they propose a modified version of the clr transformation, called mclr, to overcome the limits of adding the pseudo-count in the calculation of the log-ratios. Briefly, mclr consists of calculating the clr on the non-zero elements of the relative abundance vector, and then adding a constant to shift the transformed values in the positive direction. This transformation is consistent if zero-counts are added, does not change the original zero measurements, and keeps the order of the measurements unchanged. The authors suggest transforming the data with the mclr and subsequently applying the SPRING framework for the conditional independence network inference.

3.2.3. Bayesian Formulation of Graphical Model

Usually, those methods that consider abundance data as compositional estimate Γ' of the unknown log true value, or its inverse Ω. Instead, BAnOCC [71] sets a Bayesian solution scheme that allows to estimate both matrices and a confidence value for the estimate. The method assumes that the unknown true count values follow a log-normal distribution. The authors identify the posterior likelihood function for Ω, given the observed compositional values. Then, a Glasso prior is introduced to estimate Ω under sparsity constraint.

MDiNE [72], on the other hand, uses a Bayesian framework to infer the two precision matrices of two groups of subjects (Ω^A and Ω^B) identified by a binary variable. In the hypothesis of multivariate normality of the alr transformed counts, the authors rely on a logistic normal multinomial model of counts to draw a Glasso problem that considers separately the two matrices Ω^A and Ω^B. To induce sparsity in the estimation of the elements of Ω, a Laplace prior with mean zero and scale parameter λ is introduced. This step is the Bayesian analogue of considering an l_1-penalty in the objective function of a usual Lasso problem. The penalty parameter λ has its exponential prior distribution. Finally, the model is fitted using the Hamiltonian Monte Carlo procedure. Therefore, MDiNE is useful in the presence of two groups in the dataset, because from the posterior probability function, final estimates of the two matrices are available, from which different interaction networks can be built. The

main advantage of MDiNE is modeling the difference between Ω^A and Ω^B. Indeed, the dataset is not split into two groups of subjects to independently infer the network, as would be done with the other methods. As a result, MDiNE addresses the problem of the reduced number of samples compared to the high number of taxa, an issue that is exacerbated when the dataset is split into groups for the investigation of possible differences.

3.2.4. Methods based on Probability Theory

In MENAP [73], the authors resort to the Random matrix theory by developing an algorithm that automatically finds the correlation threshold for network reconstruction. Briefly, when the eigenvalue spacing distribution of an adjacency matrix follows Poisson distribution, it means that the system has non-random properties, instead when it presents Gaussian Orthogonal Ensemble (GOE) statistics, the generative process is random. Under the hypothesis that the eigenvalues of θ obtained from a complex biological system also have the previous properties, the correlation threshold is iteratively defined at the transition point between GOE and Poisson distribution identified by a statistical test on eigenvalues probability density.

A probabilistic approach to define the type of association between pairs (*i*,*j*) has been used by Veech *et al.* [74] and implemented in the R package cooccur [75]. Although this method is not originally developed for metagenomic data, there are studies that use it to analyze association patterns also in 16 rDNA-seq field [76, 77]. Observed co-occurrence between *i* and *j* is defined as the number of co-presences among all samples (Q_{obs}). The exact probability that the two taxa co-occur in *s* samples (P_s) is given by the probability mass function of the hypergeometric distribution. Varying *s* from 0 to Q_{obs} and summing all the P_s give the probability that *i* and *j* co-occur in at least Q_{obs} samples if their co-presence patterns were random, (*i.e.*, the p-value of positive association). The p-values for negative associations are calculated similarly, and a threshold can be applied directly to identify network edges. Alternatively, the effect size (*ES*) can be calculated as the difference between Q_{obs} and its expected value $E(Q_{obs})=P_{obs}\cdot N$. Then, *ES* values obtained are normalized with respect to the total number of samples. The standardized *ES* has a range in [-1,1] which defines the sign of the association. The probabilistic model is based on the assumption that the probability of (*i*,*j*) co-occurrence in each subject is equal to its frequency calculated among all subjects. Note that the model is distribution-free, since no randomization is required to create the null hypothesis, and metric-free because it is based on presence-absence values.

3.3. Methods for Time Series Data

Longitudinal studies involve several microbiota sequencing measures in time. Therefore, for each involved subject, the data can be rearranged into a matrix containing the measurements of *p* taxa on the rows and the *n* instants on columns. In the rest of the manuscript, we will refer with $X_i = [x_i(1), x_i(2),\ldots, x_i(n)]$ to the vector of the i^{th} taxon abundances in the *n* temporal instants, $x_i(t)$ the measure of the i^{th} taxon in the generic time point t=1,...,*n*, $\underline{X}^t = [x_1(t), x_2(t),\ldots, x_p(t)]$ the vector of the values of all *p* taxa in *t*. In addition, we will use the indices *i*, *j*, *o*, for taxa, whereas *z*, *w* will be used for points in the time axis.

The methods developed for time series analysis try to infer the relationships between the *p* taxa in mainly two ways. In section 3.3.1, we will present algorithms based on pairwise association metrics between time profiles; in the remaining part of the section, we will deal with methods that involve different mathematical models to describe the entire observed dynamics of each taxon.

3.3.1. Local Association in Time-windows

In studies where measures of many different time points are available, the search for relationships based on the entire time span could obscure the associations that occur in short sub-intervals, perhaps crucial for the future evolution of the bacterial community. Local Similarity Analysis (LSA) is a useful tool for identifying dependencies between taxa that occur in short periods over time, also accounting for some delay in the putative cause-effect relationship. In the literature, there are several methods and applications of the LSA, starting from the problem of aligning sequences, the search for relationships between gene expression levels up to local association in the microbial population. The main idea is the calculation of a similarity score *LS* between two variables X_i and X_j (in our case, taxa) carried out on each possible time window of length $l<n$. To be more precise, *LS* is calculated finding the temporal interval [*w*,*w*+*l*-1] and [*z*,*z*+*l*-1] of length *l* that maximizes an association score *S* between the variables considered:

$$LS = \max_{l,w,z} S = \max_{l,w,z;|w-z|\leq D} |\sum_{k=0}^{l-1} X_i(w+k)\cdot X_j(z+k)| \qquad (1)$$

where the values of the variables X_i and X_j have previously been normally standardized, *l* denotes the value of the time shift sought, *w* and *z* are the possible beginning instants of the relationship to be identified and *k* is the index that defines the time windows in the interval considered. In general, the search for the maximum score for short delays can be limited by forcing the combinations of the starting point *w* and *z* to follow a constraint such as $|w-z|<D$, where *D* is the number of time units of the delay. After calculating the maximum association score in its relative time segment, statistical significance is identified by means of a permutation-based approach. Considering the high number of all taxa pairs and the time required for the dynamic programming algorithm used for calculating the *LS*, the main problem of the LSA is the runtime.

Xia *et al.* [78] proposed eLSA, which speeds up the approach described above, exploiting a theoretical approximation of the statistical significance distribution of the *LS* scores. The authors also provided indications on the number of time points necessary to maintain the validity of the approximation used, and extended the method in the presence of replicated measures.

In general, LSA is used for the study of associations over time and allows to reconstruct a direct network in which the direction is identified by the time-delay association, *i.e.*, an edge from *i* to *j* indicates that the relational pattern is observed later in *j* with respect to *i*. However, LSA can also be applied to static data, if the delay parameter is set to null. In this case, the inferred network will be without temporal causality indications of the significant associations found among the taxa.

3.3.2. Rule-based Interaction Inference

LSA-based methods make no assumptions about the type of relationship between taxa. As mentioned above, complex mechanisms of interaction are established within a bacterial community that usually involves more than one taxon. In a study by Tsai *et al.* [79], a rule-based algorithm, RMN, is used to infer cooperative and competitive associations that occur simultaneously. The authors borrow the idea from smooth response surface (SRS) algorithm developed for gene regulatory network inference [80], in which the relationship between the target, repressor and activator nodes is modeled by a 3-dimensional surface. The main assumption of RMN is that under the community regulation network, there is a cooperation-competition system. Considering a triplet (X_i, X_j, X_o) composed of the relative abundances of taxa i, j and o, in which cooperation between i and o and competition between j and o are hypothesized, the method models the two relations based on the values of the triplet. In practice, RMN describes each possible taxa triplet according to a piecewise nonlinear quadratic polynomial that involves the use of hyperbolic tangent functions (tanh):

$$\hat{X}_o = \begin{cases} 2\tanh(1.1X_i)\,(1-\tanh(1.1X_j)) & if\ X_i \in [0,0.5[\ and\ ]0.5,1] \\ 1-2(1-\tanh(1.1X_i))\tanh 1.1X_j & if\ X_i \in]0.5,1]\ and\ [0,0.5[\\ \tanh(1.1X_i)-\tanh 1.1X_j+0.5 & otherwise \end{cases} \quad (2)$$

where $\hat{X}_O$ the estimated standard relative abundance of X_o expressed as a function of the other taxa involved in the triplet. Briefly, RMN applies the model assumed for each triplet across all time points, and assigns a goodness score to the model using the lack of fit function L. Subsequently, L is adjusted by means of a function which assigns a reliability score to the measure of L. Finally, thresholds on the calculated scores allow for filtering the triplets that will build the nodes of the final network. Given the assumptions on the relationships of (X_i, X_j, X_o), if the triplet exceeds the filtering step, a directed edge from taxa i to taxa o will identify cooperation, and on the other hand, from i to j competition. The authors state that RMN may not be able to identify linear correlations that could instead be found with similarity-based methods. Consequently, the information obtained on the complex relationships between taxa *via* RMN could be complementary to that obtained from methods that exploit similarity to have a greater resolution on microbial interaction landscape.

3.3.3. Modelling Community Dynamic through the Lotka-Volterra Model

Other methods, on the other hand, seek to model the dynamics of changes in taxa abundances by considering the entire bacterial community. A known model for population dynamics is the generalized Lotka-Volterra (gLV) model, also called the Ricker model in its discrete form. gLV describes the temporal evolution of the abundance of taxa i in relation to its growth rate and to the bond strength that undertakes with all the other subjects of the population. In detail, the dynamic model is defined as follows:

$$\frac{d}{dt}X_i(t) = X_i(t)\left(r_i + \sum_{k=1}^{p} m_{ik}X_k(t) + \varepsilon_i(t)\right) \quad (3)$$

where r_i is the growth rate of taxa i, m_{ik} is a vector which takes into account the influence of taxa k on the growth of taxa i, and ε_i is an additive stochastic noise which considers the measurement error and possible environmental factors influencing abundance changes. The previous differential equation can be rewritten considering the definition of derivative of the logarithm as follows:

$$\frac{d}{dt}log\left(X_i(t)\right) = r_i + \sum_{k=1}^{p} m_{ik}X_k(t) + \varepsilon_i(t) \quad (4)$$

In this way, it is possible to refer to a standard linear regression problem in which the objective is the estimation of the model interaction coefficients:

$$Y = \Phi\Xi + E \quad (5)$$

where the element Y_{ti} of the response matrix Y^{nxp} is defined as the difference ln $X_i(t+1)$ - ln $X_i(t)$, E^{nxp} is the matrix containing errors, $\Phi^{nx(p+1)}$ is the design matrix of the abundance measures, and the parameters matrix $\Xi^{(p+1)xp}$ is formed by the vector of the growth rates g^p flanked by the coefficients matrix M^{pxp} (Ξ = [g; M]). In practice, $|m_{ij}|$ determines edges weight in the network, while the sign is linked to the direction of the association between taxa. As a consequence, $m_{ij} > 0$ involves a beneficial contribution of taxa j to i, vice versa $m_{ij} < 0$ a competitive relationship, and a null value means no association.

In the literature, there are several methods that parameterize the Lotka-Volterra model. LIMITS [81] is based on sparse linear regression. The design matrix Φ is composed of relative abundances measure, therefore it is singular. The non-invertibility of Φ causes impossibility in using the Least Square estimate procedure. As seen above, the sparsity hypothesis of the intersections is necessary in order to identify the model. The basic idea of LIMITS is to iteratively consider an increasing number of non-zero coefficients m_{ij} in the regression until the prediction error of the model reaches a predefined threshold. For each step of the greedy algorithm, a bootstrap aggregation or "bagging" method is used to control the instability that occurs in the forward stepwise regression procedure. Briefly, the entire available dataset is randomly half divided into training and test sets. Then, for each taxon i, the model coefficients on the training set are estimated considering at each step the j-th taxon, which leads to a lower prediction error on the test set. The error threshold determines the sparsity of the model, since the procedure ends at the j-th taxon if the prediction error reaches it. Finally, a network is obtained from the matrix of the coefficients M. The entire algorithm is repeated several times resulting in B coefficient matrices, where B indicates the bootstrap number. Finally, M^B matrices are aggregated by calculating the median of the coefficients of each model with respect to all B instances. The authors choose the median as the aggregation metric because it preserves the sparsity of the final interaction matrix and improves its stability.

MetaMIS [82], on the other hand, estimates the model coefficients by relying on the Partial Least Square Regression technique. After that, the estimated M interactions are replaced in the gLV model to assess the reliability of the reconstructed time profiles. The method sorts the taxa based on the average abundance value with respect to all the samples in the dataset. Several networks are reconstructed considering a set of taxa that grows with each new added element that follows a decreasing order of the ranking (from high to low abundance taxa). The application of the gLV

model for each set of taxa determines the success or failure in the reconstruction of the abundance profile in the established time interval. Successful interaction networks are maintained, and a final consensus network can be built using one sample Z-test calculated on the proportions. In practice, for each coefficient m_{ij}, there will be n^{+}_{ij} positive and n^{-}_{ij} negative interactions, so if the ratio $n^{+}_{ij} / (n^{+}_{ij} + n^{-}_{ij})$ is statistically greater than a predefined threshold, then the positive direction is consistent across networks, similarly for negative relationships.

In analogy with previous work in cross-sectional studies [67], MTPLasso [83] tries to solve the regression problem exploiting information on known interactions between taxa from the literature. Starting from the gLV model parameterized as in Eq. (**5**), the authors formulate the Lasso regression with the mean square error in the objective function flanked by l_1-penalty on the interaction matrix. To treat the high dimensional problem, the authors introduce a prior matrix P that multiplies (element-wise) the interaction matrix in the l_1 norm of the objective function. P is obtained by assigning a weight to the interactions identified by the precision matrix estimated with the MPLasso framework. The penalty parameter λ is selected by a 5-fold cross-validation procedure. To stabilize the accuracy of the model used, several interaction matrices are estimated following the bootstrap aggregating approach. Finally, for each m_{ij}, on the median interaction matrix, a confidence score is calculated, taking into account model performance with and without the presence of the corresponding coefficient. The final network is obtained by selecting the relationships based on the previously calculated score.

3.3.4. Modelling Community Dynamic through Autoregressive Model

Another widely-used tool to estimate the dynamics of the time series is the autoregressive integrated moving average (ARIMA) model. In short, the model is composed of: an autoregressive component "AR", which considers the time-variance of the variable of interest as a regression of its previous values; an integrated part "I" which treats the variable as a difference with its former values; finally a linear propagation of the error in the preceding instants is the "MA" element. To clarify, the ARIMA model for a generic taxon of interest is defined by:

$$ARIMA(u,d,q) = \sum_{k=1}^{u+d} \underline{X}^{t-k} \Psi_k + \sum_{k=0}^{q} \Lambda_k \varepsilon_{t-k} \quad (6)$$

where Ψ is the matrix of the interactions, Λ is the coefficients matrix of the residual error ε, while parameter u is the number of lag in AR part, d is the degree in the differencing process k, and q is the error propagation order.

In a study by Ridenhour *et al.* [84], the ARIMA model is used to describe the temporal evolution of the bacterial community as above. The authors choose to consider the error on the counts measurement as a Poisson process. As a consequence, they rewrite the model as a log-linked Poisson regression. The parameter estimation requires a regularization algorithm to achieve the best performance with the minimum number of interactions, due to the usual high-dimensionality problem. Therefore, the authors choose the elastic-net approach that introduces l_1 and l_2 norm penalties to the parameters vector in the objective functions. The optimal value of the penalty coefficients is obtained using the cross-validation procedure. The authors use a linear model corresponding to a first-order ARIMA model, *i.e.*, ARIMA(1,0,0), to analyse a real dataset because the limited number of samples suggests limiting the model complexity. However, a high-order model or a different choice of parameters u, d, q can be used to describe a more complex dynamic with the awareness that the number of parameters to be estimated in the model and the difficulty in interpreting the interaction coefficients will increase.

3.3.5. Modelling Community Dynamic through the Granger Model

Another approach used to determine a causal relationship is based on the concept of Granger Causality. Following the definition, if a time series X_i can be better predicted considering the past history of both X_i and X_j rather than just X_i's past, then taxon X_j G-causes X_i ($X_j \rightarrow X_i$) [85, 86]. Therefore, using an autoregressive representation, the comparison is made between the models:

$$X_i(t) = \sum_{k=1}^{\infty} \alpha_k X_i(t-k) + \varepsilon(t) \quad (7)$$

$$X_i(t) = \sum_{k=1}^{\infty} \alpha_k X_i(t-k) + \sum_{k=1}^{\infty} \beta_k X_i(t-k) + \xi(t) \quad (8)$$

where α and β are the time regression coefficients, while ε and ξ the autoregressive prediction error of the two models. X_j G-causes X_i , $X_j \rightarrow X_i$, when the variance of $\varepsilon(t)$ decreases with the introduction of the X_j series in the model. The concept can be extended to the multivariate case in which p taxa are involved in the model. In practice, the conditional Granger model can be rewritten as follows:

$$\underline{X}^t = A^1 \underline{X}^{t-1} + A^2 \underline{X}^{t-2} + \cdots + A^{t-1} \underline{X}^1 + \varepsilon^t \quad (9)$$

where A is the matrix of the regression coefficients. When $A_{ij}^{t^*}$ is statistically significant for a generic instant t^* then $X_j^{t^*} \rightarrow X_i^t$. Consequently, an edge can be drawn in the adjacency matrix θ^t, which describes the network of relationships for the corresponding time point. In general, the model leads to reconstruct a graph of $n \times p$ nodes.

Mainali *et al.* [87] developed TIME, where a Granger framework is used for finding causal relationships among all taxa. The authors implement three different approaches that can be chosen by the user. The first is based on the inference of the pairwise Granger causality for each pair X_i and X_j starting from the comparison of the two regressions described in Eqs. (**7** and **8**). The second one estimates the coefficients in the multivariate case using a Granger Lasso with a regularization procedure to deal with a high number of parameters in the model. The third method consists of an ensemble approach which selects the pairs of nodes $X_j \rightarrow X_i$ identified by both the previous options.

4. BENCHMARKING STUDIES

All the aforementioned methods deal with the interactions inference problem through different approaches or hypotheses, or formulations of the solution strategy. Since it is impossible to know the complex reality of the interactions that occur within a real microbial community, the authors usually demonstrate the best performance of their method on simulated data (section 5), or validate the associations found by comparing them with the literature. In Fig. (**2**), a graph is

shown to describe the comparisons made between methods presented in this review. Each node represents a network inference tool, with edges going from the method presenting a paper in which the comparison is proposed to the literature methods compared with it. Looking at the number of edges entering different nodes, it can be observed that CCREPE, SparCC and SPIEC-EASI are the most used methods in performance assessments. A possible reason is that they were developed less recently, so they are the most used tools in many analyses. The Spearman and Pearson correlation-based methods are used as the baseline for comparisons also by recently developed methods such as SPRING and BAnOCC. The figure also highlights the limited number of comparisons between time series methods (in the box square), although some of them use simulated data to test hypotheses or the effect of the parameters of the model used. The results of multiple comparisons between several methods and the same target tool are influenced by design used in each individual collation. Indeed, the use of different simulation frameworks or performance evaluation metrics complicates the possibility of drawing general conclusions regarding the reliability of the target.

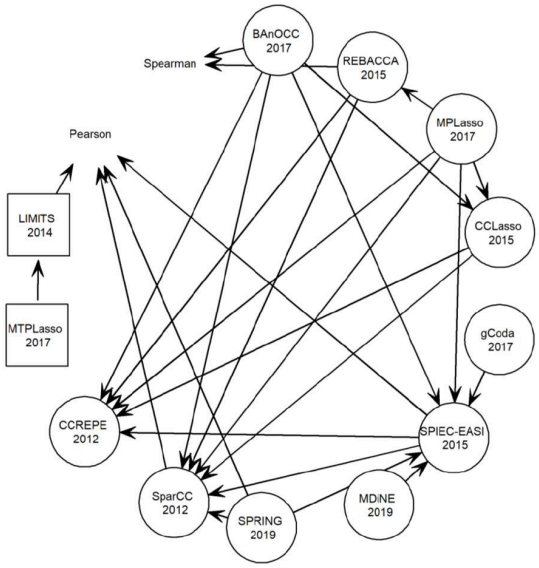

Fig. (2). Comparison graph. Each node corresponds to a method mentioned in the review. An edge from node A to node B means that in the article of method A there is a comparison between method A and method B in simulated context. The methods inside a circle deal with cross-sectional data, while in square with time series. Pearson and Spearman nodes do not have a specific form since they are correlation measures that can be applied to any profile vector. The graph was built with the R package *network* using circle layout option.

Benchmark articles represent an independent, simultaneous assessment between many network reconstruction methods. Consequently, they are a valid tool to understand the performance of the developed algorithms. Recently, Hirano *et al.* [88] tested several correlation-based methods, such as Pearson, Spearman, MIC, SparCC, REBACCA, and CCLasso, and graphical model-based, *i.e*., Pearson's partial correlation, Spearman's partial correlation, and SPIEC-EASI. Within the two categories, some methods are based on a compositional approach, while others on a measure of count association. The results seem to demonstrate that compositional approaches have comparable or sometimes lower performance compared to correlation-based. However, authors highlight the dependency of the results on the type of associations between taxa, which causes a difficult interpretation of co-occurrence networks. To the best of our knowledge, there is no benchmark in the literature that considers all the static methods mentioned in this review simultaneously.

The lack of independent comparative studies also characterizes the time series landscape. In general, there are comparisons between different configurations of the same approach. For example, Chen *et al.* [89] showed in a simulated study that the use of the Granger causality calculated on each pair of time series individually (pair-wise formulation) can lead to the reconstruction of spurious causal links. On the contrary, the multivariate approach (conditional Granger causality) is more robust to these errors.

The importance of this research area and the continuous development of new methods requires a careful and deeper evaluation in order to identify the best performing approaches.

An important step to assess methods performance in the absence of biological truth is the simulation of microbial community networks and, consistently, *in silico* sequencing count data originating from microbial interactions.

5. SIMULATING SYNTHETIC COUNT DATA

The tools infer the complex web of microbial interactions starting from abundance data, but in the absence of a ground truth network, it is impossible to verify the reliability of the estimates. In order to demonstrate the performance of inference methods, synthetic data generation is necessary. The challenge is to simulate count data that reflect a realistic output of a sequencing process by imposing a known dependency structure between variables. The simulation frameworks are, therefore, divided into two steps. First, there is a need to define an underlying structure of the interaction between each taxon, *i.e*., which variables are involved in the network and the strength of the relationship. The second goal is to find a count data model that simulates abundance profiles conditioned by the imposed relational structure. At the end of the entire procedure, a comparison between the network inferred by the methods and that imposed in the generative process needs to be carried out to evaluate the reconstruction performance.

Fig. (**3**) shows a scheme of the steps that constitute the general simulation strategy. In paragraphs 5.1 and 5.2, we describe in detail two different approaches used to define a ground truth of associations between variables (the two different ways in the figure). Subsequently, in section 5.3, we summarize several models used to produce the final matrices of the abundances (*i.e*. the last step in the framework). Besides, Table **3** provides an overview of the simulation procedures that the single tools use to perform a comparison with other methods.

5.1. Network Structure Generation

A graph of microbial interactions can have different topological characteristics that describe the natural mechanisms

established between the members of the community in a given ecological niche. Simulating a network similar to a possible ecological scenario corresponds to generate an adjacency matrix θ starting from different hypotheses on the properties of the graph. Remembering that the adjacency matrix only describes the presence or absence of the edges between the nodes, it is necessary to subsequently define weight of the interaction strength in order to obtain the ground truth of the associations. Many of the methods in the review use this approach to simulate synthetic data with different underlying topology. In subsections 5.1.1 and 5.1.2, we describe the different topologies considered and the common methodologies for assigning the strength scores.

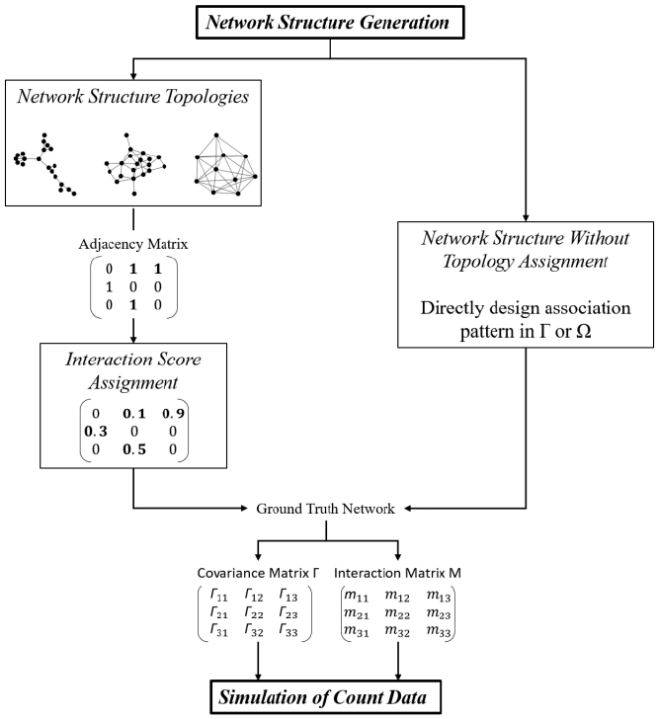

Fig. (3). Scheme of the overall synthetic count data procedure. The two approaches for generating association structures are shown separately. Both alternatives produce the ground truth network, which represents the input to the count data simulation step.

5.1.1. Network Structure Topologies

The simulation of a taxa interaction network of a bacterial community is challenging. Firstly, it is necessary to have an idea of the possible structure that could be observed within a real ecological scenario. Generally, the network topology is influenced by specific environmental conditions. For example, in the hypothesis that a particular niche favors only one species that dominates all the others, the resulting network will have a hub node (parent) highly connected to other sparsely connected nodes. In a simulated context, a network inference method must be robust to the variability of the possible relationship schemes. Therefore, some of the previously seen methods use a simulation framework, for performance comparison purposes, which considers different graph structures in the process of synthetic data generation. In the literature, there are several topological models used to create a graph with a defined structure and properties. In the following, we will summarize only the approaches that the tools use to create synthetic data with a known underlying network.

- **Random Graph:** It is a network built on a probability distribution of nodes or by a random process that models the graph, such as the Erdös-Renyi model. Considering the total number of possible edges *e*, to obtain a random configuration, it is necessary to set the probability of connection P_{RG} between two taxa. The sparsity level of the network depends on the chosen probability that determines the value 1 in the adjacency matrix.
- **Neighbor Graph:** It is a graph where each node is connected with a fixed number of neighbors. In a $[0,1]^2$ plane, *p* points are randomly selected. Subsequently, for each point, an edge is assigned linking the *K* nearest neighbors.
- **Band Graph:** It is a chain graph with each node linked to its *K* nearest. For each node pair (X_i,X_j), an edge is set in the corresponding adjacency matrix when $1<|i-j|<K$. Another approach (B2 in Table **3**) iteratively sets edges in the adjacency matrix for the next available off-diagonal vector, if the number of available edges is greater than the off-diagonal components.
- **Hub Graph:** It is a type of network that includes independent subnets characterized by hub nodes with a high degree. The first option is to choose *g* points at random as hub nodes. Then, the other *p-g* nodes are connected to hubs with a defined probability P_h and to non-hubs with P_{nh}. Another way (H2 in Table **3**) is to divide nodes into *g* random groups. Then, for each group, a center node is connected to the others.
- **Cluster Graph:** It is a network made up of independent subnets. Basically, *p* nodes are clustered in *g* independent groups with equal size. Then, the number of edges for each subnet is set to e_{sub}. Then, the Erdös-Renyi model is used to construct a random graph for each group with a defined edge probability P_c. Another approach used to create a cluster graph (denoted as C2 in Table **3**) consists of assigning an edge with probability P_c for each pair in the uniformly distinct g groups.
- **Block Graph:** In general, the network has several sparsely connected blocks. To obtain this type of network, methods divide the nodes set in blocks of the same dimension. An edge between node couples in the same block is assigned with high probability P_{sb}, while the connection between nodes belonging to different blocks has a lower probability P_{db}.
- **Scale-free Graph:** It is a network with the property that the node degree follows a power law, so the probability that a node is connected with other *h* nodes is a decreasing exponential with base *h*. Usually, methods use the Barabàsi-Albert (B-A) scheme [90] to generate this type of graph. In brief, the algorithm starts with 1 or 2 nodes, then each new node is iteratively connected to *c* nodes in the network with a probability dependent on the node degree in the building network. Alternatively, the Chung–Lu model [91] can be used to build scale free networks. Edges connect nodes based on node weight $(i+i_0-1)\zeta$, where ζ in [0,1] is node index, and i_0 is a constant. The output network follows a power law in which the exponential parameter is related to ζ.

- **Small-world Network:** It is a network where many nodes are not directly connected to each other, but each node can reach any node on the network through a short path. Watts–Strogatz model [92] is usually used to generate this type of network. Briefly, the algorithm starts building a ring lattice with p nodes and E(deg)/2 edges that connect each node neighbors on both sides, where E(deg) is the desired mean degree. Then, for each edge, the algorithm proceeds with rewiring the target node given a predefined probability P_{sw} avoiding duplicate and self-loop.

5.1.2. Interaction Score Assignment

When the network structure θ is simulated, the next step is to set the strength of the interaction, *i.e*., the weight of the edges. As seen previously, the inference methods are based on the estimation of the covariance matrix Γ, the precision matrix Ω or directly the network adjacency matrix θ. Therefore, the basic idea is to generate one of these matrices by choosing correlation values, conditional dependence, or coefficients for non-zero values of θ. There are mainly 4 different methodologies:

1. **Assign Values to Γ:** Some authors directly set the elements of Γ with a defined strength value, depending on the type of network considered. Then, the components in the diagonal of the matrix are chosen sufficiently large to ensure that the covariance matrix is positive-definite, and then the elements are normalized to 1.
2. **Assign Values to M:** In the dynamic model, M scores are usually obtained from different uniform distributions. If the authors decide to consider the type of interaction between taxa in design M, the M_{ij} and M_{ji} elements are set based on the sign score.
3. **Assign Values to Ω (first method):** The approach is to assign at the corresponding non-null entries in θ a correlation strength score sampled uniformly in a given interval. The certainty that the Ω is positive-definite with a predefined number of conditions κ is given by scaling the diagonal for a constant using binary search. The Γ obtained by reversing the precision matrix is then rescaled. A variant of the previous method consists of assigning the values to Ω from a uniform distribution to the corresponding non-null elements of the lower triangular matrix of θ. Then, the elements of the upper triangular part are set equal to their symmetrical element in the lower part. The diagonal is imposed on a constant in order to obtain an Ω with a predefined number of conditions κ.
4. **Assign Values to Ω (second method):** In the R huge package [93], often used to generate the types of networks cited in the previous section, a different strategy is adopted to ensure a positive-definite Ω. The smallest eigenvalue of $\theta \cdot v$ is calculated, where v is a parameter that controls the magnitude of partial correlations. The final precision matrix is $\theta = \theta \cdot v + (|\min(eig)| + 0.1 + u)\, I$, where $\min(eig)$ is the minimum eigenvalue calculated before, u is a shift parameter for the diagonal, and I the identity diagonal matrix. At the end of all the procedures described, Γ will be available inverting θ.

5.2. Network Structure without Topology Assignment

Testing the developed methods is important to provide the robustness of the method in relation to different bacterial communities' scenarios. However, some procedures do not take into account topologies in the synthetic count generative process, but directly design association pattern in Γ or Ω.

SparCC uses a simulation approach where communities are modeled by a multivariate log normal distribution. The elements of μ are the same except for one taxon, which is chosen in order to ensure a given n_{eff} on average for the community. In addition, variance values are constant and Γ is obtained by defining for each pair of taxa a certain probability of perfect positive or negative correlation. Finally, Γ is transformed into the nearest positive-definite matrix for invertibility assurance.

In BanOCC, the authors use the model adopted by the method to generate a small synthetic dataset. Basically, they start from a given correlation structure of the true abundances with no true correlations or at least one. Then, they set different parameters to obtain a final count matrix with several scenarios of negative spurious correlation induced. Besides, other simulations are conduced starting from a real-data template of correlation structure, where all the p/2 random selected off-diagonal values are set to the same strength.

In the MDiNE simulation approach, the Ω is obtained by Cholesky decomposition. In practice, the authors define a lower triangular matrix L_0 by sampling from two different uniform distributions for each entry and for the diagonal elements. Subsequently, Ω is obtained by decomposition $\Omega = L_0 L_0^T$. Then, the authors use the NorTA approach [94] with Γ derived from the previously precision matrix. We delve into this approach in the next section.

In LIMITS, the Ricker model is used to simulate the temporal abundance profile. In this discrete form, the regression of species i abundance involves the difference between all other species and their abundance at the equilibrium of the system. Therefore, the generative process starts sampling abundances X_i at equilibrium from a lognormal distribution. Then, the interaction matrix is built with the coefficients in the diagonal m_{ii} sampled from a uniform distribution dependent on the equilibrium point X_i. The off-diagonal parameters m_{ij} are added up one by one to the model stability maintenance and all the interaction coefficients are determined.

RMN builds, for assessment purposes, a predefined relational scheme using tanh functions to simulate a regulation network composed of 10 nodes. In detail, given a network of 10 nodes and 3 latent factors, the authors build a system of equations that describe the links between the i-th taxon and its connected nodes.

5.3. Simulation of True Abundances and Count Data

Each community sample is composed of a set of reads that characterize the different species present within it. The count values can be interpreted as the result of a random sampling of the reads, based on the fact that a DNA sample can be seen as a collection of fragments taken from the species present within it and then DNA sequencing can be compared to a random sampling of the species. A classical approach for modelling count data is by using a Poisson random variable, thinking that each DNA fragment has the same

Table 3. Summary of the generative processes used by the inference methods present in the review to evaluate performance against the ground truth structure.

Method	Network Structure	Score Assignment	Count Data Simulation	Assessment Metrics
MDiNE	Ω obtained by Cholesky decomposition	-	NorTA	-AUROC and AUROC of edge difference between two precision matrix Ω^A , Ω^B -Wnc
SPRING	-Band (B2) -Cluster -Scale-free (B-A)	method 3 (variant)	NorTA	-d_{ij} -scatter plot (estimate *vs* true correlation). -PR curve -AUPRC -d_H (in relation to the tuning parameter) -average number of overlapping edges between methods
MPLasso	-Cluster (C2) -Band -Scale-free -Random -Hub (H2)	method 4	-ln (y) ~N (μ, Γ) -y~NB	-AUPRC -ACC (also on edge sign recovery for log normal model) -L_1
gCoda	-Random -Neighbor -Band -Hub -Block	method 1	ln (y) ~N (μ, Γ)	-AUROC -ROC curve
BAnOCC	- 4 correlation scenarios - log-basis correlation set on different strength	-	- model of the method -sparseDOSSA	-heatmap of the estimates and significance of correlations -Type I error rate -Type II error rate -AUROC -ROC curve
MTPLasso	-Random -Scale-free	method 2	gLV	-AUPRC -IACC
Ridenhour *et al.*	-Small-world (Watts-Strogatz)	method 2	$X_i(t+1)=X_i(t)e^{CiX(t)}$	-ROC curve -MSE
SPIEC-EASI	-Band: (B2) -Cluster -Scale-free (B-A)	method 3	NorTA	-PR curve -AUPRC -node degree -betweenness centrality -geodesic distance
REBACCA	-3 fix structure: case 1: hierarchical structure case 2: 4 mostly negative correlated taxa case3: 3 correlated cluster groups -Scale-free (B-A) -Cluster	method 1	-y~log ratio normal (LRN) - y~Poisson log normal (LNP) - y~Dirichlet log normal (LND) + multinomial distribution with given sample size	-AUROC -ROC curve -FP rate

(Table 3) contd….

Method	Network Structure	Score Assignment	Count Data Simulation	Assessment Metrics
CCLasso	-Random -Neighbor -Band -Hub -Block	method 1	ln (y) ~N (μ, Γ)	-AUROC -ROC curve -RMSE -d_F
RMN	- Association structure imposed by the system	-	system of tanh equation with 3 latent factors considered	-TP rate -TN rate -accuracy -F measure -Pr values
LIMITS	- m_{ii} sampled from a uniform distribution, m_{ij} are iteratively added up to the model stability maintenance	method 2	gLV	-scatter plot (interactions *vs* correlation coefficients) -scatter plot (true interactions *vs* estimated interactions) for both abundance and relative abundance values. -correlation between true and inferred interactions -Frequency of error in interaction sign -sensitivity and specificity
SparCC	-random Γ where each OTU pair has a given probability of being perfectly correlated	method 1	ln (y) ~N (μ, Γ)	-Visual comparison of network -RMSE

Ω = precision matrix; m_{ij} = interaction coefficients of M ; Γ = covariance matrix; y = true compositional abundance; μ = the mean abundance vector; x(t)= abundance of taxon i in t; Wnc= Weighted natural connectivity; dij= pairwise absolute difference; d_H= Hamming distance; d_F= Frobenius norm distance; Pr=probability of prediction for pairs with less than 0.5 of non linear correlation coefficient.

chance of being selected for sequencing and the fragments are selected independently [95]. However, taking into account biological noise, *i.e*., the number of fragments for the same species among different samples, is affected by biological variability. The Negative Binomial (NB) distribution has been adopted in sequencing count data modelization [96, 97]. However, as previously mentioned, 16S sequencing data are also characterized by a high sparsity. To model this characteristic, a mixed-model zero-inflated approach, such as zero-inflated negative binomial distribution (ziNB) is often used, which is a mixture of Negative Binomial (ZINB) models with a point mass at zero [98, 99]. More recently, Patuzzi *et al.* [100] proposed a model based on a Gamma – Multivariate Hypergeometric distribution, able to describe the compositional nature of 16S sequencing count data, thus explicitly accounting for the constraint imposed by the fact that sequencing platforms can produce reads only up to their capacity (*i.e*., the sequencing depth).

In any case, the simulation of count data must proceed from the simulation of true abundances that should reflect the topology of a known network. For example, if we assume that the true compositional abundance y follows a log normal distribution with mean μ and covariance matrix Γ, $\ln(y) \sim N(\mu,\Gamma)$, then the simulated observed relative value can be calculated as $x_i = y_i / \Sigma y_i$ with $i = 1,2,\ldots, p$ where p is the number of taxa. In REBACCA, the generative process is essentially composed of two steps. First, the true basis proportion of each taxon is given by a log normal multivariate distribution with zero mean and a defined covariance matrix Γ. The true abundance values given the proportion obtained are modeled by a Poisson log normal distribution and proportions are recalculated. In alternative, the true basis proportions are obtained from the same normal distribution with given mean and Γ and the final values are sampled using a Dirichlet log normal. The second step draws counts values from a Multinomial distribution using the true proportions as probabilities and given sequencing depth (corresponding to the number of extractions). All models require the definition of the mean true abundances μ vector, which control the balance of the components. Generally, each element of μ vector is sampled by a multivariate uniform distribution.

SPIEC-EASI proposes a simulation strategy based on the Normal to Anything (NorTA) approach to simulate count matrices with a known underlying structure. In short, the method allows to generate a multivariate distribution with a defined correlation structure and with a defined univariate marginal distribution starting from a normal multivariate distribution. Firstly, a $p \times s$ matrix U with independent rows drawn from a multivariate normal distribution $N(\mu,\Gamma)$ is generated, where p is the number of taxa and s is the number of samples. Then, for each column of U, each marginal element is transformed using cumulative distribution function (CDF) of univariate normal distribution. Finally, data are generated from the inverse CDF of the desired marginal distribution for each column of the transformed U.

In SPIEC-EASI, the marginal distribution chosen by the authors is the ziNB. Parameters of the marginal distribution are estimated from real data using R package VGAM [101]. However, in SPRING, the authors show that taking the inverse of the empirical cumulative distribution function (ecdf) calculated independently for each OTU leads to synthetic data that are more reliable.

Another way to simulate abundance matrices that reflect a realistic data structure is proposed in BAnOCC. In particular, the authors use SparseDOSSA [102] simulator, which can consider known correlation structures in the generative process. Each taxon has a marginal distribution modeled by a zero-inflated, truncated log-normal distribution where parameters are obtained from a log normal distribution with a given correlation matrix.

The gLV model is also used to generate microbial population abundances given a defined interaction matrix M. In the previously cited comparison [88], taxa abundances derive from the model after 1000 time points in which the steady state has already been reached. The initial abundances values are sampled independently from a Poisson distribution of average 100 and a total population of $100n$, while the growth rates r are sampled from a uniform distribution. Finally, the relative abundances are obtained by dividing each X_i by the total sum of the population. The simulation procedure used in MTPLasso, sets the growth rates vector in a different interval, while the M elements on diagonal m_{ii} and outside the diagonal m_{ij} are sampled by two different uniform distributions excluding zero, because it would eliminate edges from the underlying structure.

In the simulation framework used by Ridenhour *et al.* [84], the abundance value of each taxon is chosen by a negative binomial distribution with defined parameters. The generative model assumed follows an exponential growth of the abundances of each i-th taxon over time $X_i(t+1)=X_i(t)e^{C_iX(t)}$, where C_i is the weights vector in the i-th row of the adjacency matrix.

5.4. Simulation Parameters

In a simulated context, researchers can manage covariates not possible in a real dataset. Usually, the above-mentioned approaches are used to simulate different synthetic data by varying the number p of taxa present in the dataset and the number s of samples or n of time point$s$. Then for each configuration and, where used, each network structure θ, several datasets are generated. In the time series scenario, an additional parameter is usually considered to introduce a noise component in the simulation. In MTPLasso, for example, different additive noise levels in the gLV model are tested, while in RMN, a random noise component is added to the equations in the system.

6. METRICS TO ASSESS METHOD PERFORMANCE

Synthetic networks represent the ground truth on which comparisons are made. Different metrics are then used to evaluate and summarize the differences between the simulated reality and the inferred networks.

The choice of metrics to be used in the comparison between the simulated ground truth and the inferred network is extremely important.

Indeed, the evaluation should be carried out in an unbiased way, without favoring some methods with respect to others. Generally, the performance evaluation is carried out using metrics that look at two different aspects: the correct edges identification, and the maintenance of the characteristic properties of the topology considered.

6.1. Edge Recovery Metrics

Correct edge recovery, *i.e*., the ability to reconstruct the true relationships between the nodes of the network, can be calculated in a static way, in terms of the number of true positives (TP) and true negatives (TN), also accounting for the statistical significance of the inferred edges. In this review, we have seen how many methods associate a p-value to each edge of the network. In the graphical-based methods, on the other hand, pseudo p-values are obtained from the stability score in correspondence with the chosen regularization parameter. The calculated p-values are sorted and a Precision-Recall curve is calculated as the threshold changes across the p-value range. In particular, the values of the contingency matrix (*i.e*., TP, TN, FP, FN) consider the presence-absence of the edge with respect to the underlying network and, where possible, the sign of the association. The performance can then be summarized with the area under the PR curve (AUPRC) to assess the performance in terms of precision and recall of network edge detection or, similarly, with the area under the Receiver Operating Characteristic curve (AUROC), to assess the performance in terms of specificity and recall. The AUPRC reaches its maximum when the estimated network is perfectly reconstructed (AUPRC = 1), whereas AUPRC = 0.5 corresponds to the performance of random relationship assignments. Different AUPRC, obtained across different simulations, can be averaged and their distribution can be considered to compare the performance of different methods under different scenarios. The same considerations regarding the interpretation of the values of the areas can be done for AUROC. The main difference is that AUPRC curves do not take into account true negative, so precision is affected by how rare the true values of the positive class are. Consequently, AUPRC is usually used when there is an unbalance between the two possible classes.

As a parallel approach, in MPLasso, the authors use an accuracy estimate (ACC), which takes into account the total number of pairwise correlations (n_c) defining ACC= (TP+TN)/n_c.

Other metrics are used to test the performance, not only in terms of identified edges, but also with respect to the weight of the association identified. Usually, the evaluation is done by comparing the values of the correlation, precision or interaction matrix with the related ground truth matrix, based on the type of inference method used.

In MTPLasso, the authors exploit the Interaction Type Classification Accuracy (IACC) calculated as the fraction of correctly estimated interacting interactions, since the method deals with time-series data. In addition, distance metrics between matrices are used to quantify the differences between estimates and simulated values, such as pairwise absolute difference d_{ij}, Hamming distance d_H, L_1 distance between matrix, mean square error (MSE) or its rooted version (RMSE), Frobenius norm distance d_F.

6.2. Network Topology Metrics

In the previous section, different types of networks have been described. Each graph has different properties characteristic of the topology considered. For example, scale-free or hub graphs have few nodes with a high degree, while in band graphs, the degree is constant on all nodes. As pointed out previously [49], the main properties concern node degree distribution, Hub nodes, modularity and average shortest path length. Usually, these aspects are investigated in networks reconstructed on real data in order to find known associations between taxa in the microbial population. In the literature, several topological measures have been proposed. In MENAP [73] paper, there is a useful comprehensive description of network topological indices.

In a simulated context, authors sometimes verify that network properties are efficiently reconstructed by their own methods. In SPIEC-EASI, the authors compare the node degree, betweenness centrality and geodesic distance distributions of the real and estimated network using the Kullback – Leibler divergence as a comparison measure. In MDiNE, the authors rely on Weighted natural connectivity (*Wnc*) [103] as a measure of the overall network structure. *Wnc* measures how much edge removal affects network connectivity. *Wnc* measures on inferred networks are compared with those on simulated data by calculating the log squared error.

7. OPEN CHALLENGES AND PERSPECTIVES

Despite the strong efforts of scientific research to reconstruct microbial network interactions, some challenges are still open. We have already seen how numerous are the tools that deal with the problem related to data compositionality and sparsity. However, the recent comparison by Hirano *et al.* [88] shows that the compositional approach does not always lead to a better inferred network. Accordingly, further insights through independent benchmarking studies are still needed to evaluate this aspect. Data imputation represents a frontier yet little explored in the microbiota research landscape. However, the identification of the real and technical zeros and the subsequent recovery of the information on the abundances lost due to the sequencing process could help reduce sparsity.

A consolidated phenomenon that also emerges from simulation studies concerns the positive effect of increasing the number of samples or time points in interactions recovery. In cross-sectional studies, a higher number of samples could allow to relax the sparsity hypothesis or to improve model identifiability, addressing the problem of a high number of taxa with respect to the number of samples. Although the costs of sequencing are continuously falling, increasing the number of samples can still be expensive. Furthermore, in many cases, it is difficult to gather or find numerous samples useful for the study of interest. The risk is an increase in the time required to carry out the entire research with a consequent greater expenditure of resources. Most importantly, design issues generally limit the possibility of overcoming the problem of high dimensionality in microbiota data. To reduce the number of rows in the OTU/ASV tables, one could think of carrying out studies at higher taxonomic levels such as family or order. In this way, however, some of the potential driver interactions of physiological phenomena may not be detectable because they are masked by the taxonomic resolution. A potential solution is to try to integrate the information inferred by the different networks obtained at increasing levels of the taxa classification tree. The increase in the samples present in the study is mainly linked to the resources available to carry out the research. Over the past few years, several consortia, international projects and multicenter studies have been created with the aim not only of collecting a large amount of data, but also of providing guidelines for the treatment and analysis of such data. In the future, we expect to see a consolidation of databases available to the scientific community. In addition, an important ever closer perspective is represented by multi-omics studies that seek to associate microbiome data with the genome, epigenome, transcriptome and metabolome of the human host.

Another aspect to consider is that many methods filter the data focusing only on the most abundant taxa present in the samples. The reason is to remove potential spurious interactions by focusing on core taxa. However, filtered taxa could play a crucial role in the development of some mechanisms that are established in the community. A possible technique could be considered not only the abundance percentage, but also the variance observed in the samples. Modelling the abundance profile noise could help filter data by reducing potential spurious interactions. Again, the correct imputation and recovery of the information on low abundance values could help to reduce this technical obstacle.

There are also limitations related to experimental design. In cross-sectional studies, the subjects are sampled at a single point in time, but in a bacterial community, interactions develop over time. Therefore, if the sampling of different subjects is not carried out by checking all the possible factors that convey the dynamics, the inferred patterns could be unrealistic. On the other hand, longitudinal studies enable to infer the causal direction and, therefore, to study the whole underlying generative mechanism behind the microbial ecosystem. The sampling frequency is a crucial point. Indeed, if samples are not measured at a temporal resolution suitable for dynamics variation of the microbial system, they will not convey the information to correctly infer the underlying interactions, the well-known aliasing phenomenon. System biology studies can help in defining the time ranges characteristic of the phenomena that guide microbial interactions. The knowledge of the time required for the development of a certain effect on the bacterial community due to a stimulus is of fundamental importance for outlining the sampling times. In the future, we expect a growing interest in studying the effect of time resolution on the network's inference.

Cross-sectional studies generally do not allow to estimate the sign of the interaction between species, complicating the interpretation of the relationship. In the absence of a direction, in fact, it is difficult to reconstruct the mechanisms that develop in the observed communities, but only the driver relationships that determine their composition can be identified. A potential solution to the absence of directed edges is integration with the results obtained through other studies in the literature, *e.g.*, from longitudinal studies. Furthermore, this prior information could be used to build predictive models related to the new co-occurrence interactions found.

CONCLUSION

In this review, we have summarized different approaches related with the microbial community network inference problem, specifically from 16S sequencing data.

We first introduced several methods based on pairwise association metrics that mainly differ in the strategy of assigning edge significance, in the use of information on the topological structure, and in considering the compositional nature of data. Then, we presented some multivariate models with the aim of estimating the entire interaction structure. The main distinction is between regression-based models, where the estimation of the covariance matrix of real abundances is obtained by solving a Lasso problem, and those that rely on the notion of conditional independence, where the estimation process based on Glasso involves the precision matrix. Other multivariate approaches, on the other hand, infer associations using probability theory. In addition, we presented some methods with the aim of reconstructing the evolution of microbial interactions exploiting the temporal information of the longitudinal data. We saw the concept of local association between time profiles as a metric of cause-effect relationship that can be established in different time-lagged windows. Finally, we have defined the different formulations of models that estimate the overall dynamic community.

Another aspect that we have covered is the importance of using simulated data as a necessary evaluation tool in the development of new reliable methods. For this purpose, we provided an overview of some simulation framework used by the methods to generate synthetic count data. We have not only seen how to generate several network structures, but also how to integrate them into the models used to produce synthetic abundance data.

Although the developed approaches have shown encouraging results in several applications, further efforts are still needed to ensure greater reliability of the inferred microbial interactions. To achieve this goal, not only new statistical or computational methods, but also a solid and reliable simulation framework must be improved. The lack of direct observation of biological ground truth makes it difficult to validate the many interactions that can arise in a complex community. However, if the method is robust, the results obtained are more trusted, even if in contrast with previously observed biological results. Unfortunately, modeling the biological ground truth is a difficult task to propose, since it is based on hypotheses that may not correspond to the reality of the phenomenon. For this reason, a simulation framework that considers different network structures, different count matrix generation approaches and finally, different parameters is desirable. The main idea is that if the inferred networks are robust not only with respect to the simulation parameters, but also with respect to the different possible biological scenarios. With these two objectives achieved, interaction networks will surely be one of the most useful tools for understanding how to control and manipulate the complex micro-world of the microbiota.

LIST OF ABBREVIATIONS

16S rRNA	=	Gene encodes for a small subunit of the prokaryotic ribosome
WGS	=	Whole Genome Sequencing
16S rDNA-seq	=	Targeted amplicon sequencing of 16S ribosomal RNA
NGS	=	Next-Generation Sequencing
OTU	=	Operational Taxonomic Unit
ASV	=	Amplicon Sequence Variant
alr	=	Additive Log-ratio
clr	=	Centered Log-ratio
ilr	=	Isometric Log-ratio
DA	=	Differential Abundance
MI	=	Mutual Information
MIC	=	Maximal Information Coefficient
Lasso	=	Least Absolute Shrinkage and Selection Operator
Glasso	=	Graphical Lasso
MB	=	Neighborhood Selection of Meinshausen and Bühlmann Method.
mclr	=	Modified version of clr proposed by SPRING's authors
LSA	=	Local Similarity Analysis
gLV	=	Generalized Lotka-Volterra Model
NorTA	=	Normal to Anything Method
NB	=	Negative Binomial
ziNB	=	Zero-inflated Negative Binomial Distribution
TP, TN, FP, FN	=	The values of the contingency matrix: true positive, true negative, false positive, false negative
AUPRC	=	Area Under the Precision-Recall Curve
AUROC	=	Area Under the Receiver Operating Characteristic Curve
ACC	=	Accuracy of the Prediction
IACC	=	Interaction Type Classification Accuracy
MSE	=	Mean Square Error
RMSE	=	Rooted Version of MSE
Wnc	=	Weighted Natural Connectivity

LIST OF SYMBOLS

i, j	=	Two generic taxa related to the row indices i and j of the OTU/ASV table
θ	=	Adjacency matrix
Γ	=	Covariance matrix
Ω	=	Precision matrix
N	=	Number of neighbors
Γ’	=	Covariance matrix of the true abundance
n_{eff}	=	Shannon effective number
X_{-i-j}	=	The abundance value of taxon i, compared to all the other j in the dataset
$i \perp j \mid X_{-i-j}$	=	Definition of conditional independence between two variables
P	=	Prior co-occurrence matrix
Ω^A, Ω^B	=	Precision matrix of two groups of subjects
λ	=	Penalty parameter

Q_{obs}	=	Number of co-presences among all samples
s	=	Number of samples
P_s	=	Exact probability that the two taxa co-occur in s samples
ES	=	Effect size
p	=	Number of taxa
n	=	Number of time points
X_i	=	$[x_i(1), x_i(2),\ldots, x_i(n)]$ = vector of the i^{th} taxon abundances in the n temporal instants
$x_i(t)$	=	The measure of the i^{th} taxon in the generic time point
t	=	Generic time point in [1,...,n]
$\underline{X}^t$	=	$[x_1(t), x_2(t),\ldots, x_p(t)]$ = vector of the values of all p taxa in t.
i, j, o	=	Generic indices for taxa
z, w	=	Generic indices for time points
LS	=	Measure of Local Similarity
l	=	Length of a time window
S	=	Association Score
k	=	General index for the summations involved in the formulas
D	=	The temporal unit of the interested interval in LS calculation
$\hat{X_o}$	=	Estimated standard relative abundance of X_o
r_i	=	The growth rate of taxa i in gLV model
m_{ik}	=	Interaction coefficient which takes into account the influence of taxa k on the growth of taxa i
ε_i	=	An additive stochastic noise in gLV model
Y^{nxp}	=	The response matrix of the regression in Eq. (**5**)
E^{nxp}	=	The matrix containing errors in Eq. (**5**)
$\Phi^{nx(p+1)}$	=	The design matrix of the abundance measures in Eq. 5
$\Xi^{(p+1)xp}$	=	The parameters matrix in Eq. (**5**)
g^p	=	Vector of the growth rates
M^{pxp}	=	Interaction coefficients matrix
B	=	Bootstrap number
n^+_{ij} , n^-_{ij}	=	Number of positive or negative interaction for a generic coefficient m_{ij}
Ψ	=	The matrix of the interactions in Eq. (**6**)
Λ	=	The coefficients matrix of the residual error ε in Eq. (**6**)
u	=	The number of lag in AR part in Eq. (**6**)
d	=	The degree in the differencing process k in Eq. (**6**)
q	=	The error propagation order in Eq. (**6**)
l_1 , l_2	=	Norm penalties to the parameters vector in the objective functions
$X_j \rightarrow X_i$	=	Definition of Granger causality (X_j G-causes X_i)
α , β	=	The time regression coefficients in Eq. (**7**-**8**)
ε, ξ	=	The autoregressive prediction error of the two models in Eq. (**7**-**8**)
A	=	The matrix of the regression coefficients in Eq. (**9**)
e	=	Total number of edges in a generic network
P_{RG}	=	Probability of connection between two taxa in random graph
K	=	Number of nearest neighbors
g	=	Number of hub points or cluster group chosen
P_h ,P_{nh}	=	Probability of connection between nodes and hubs
e_{sub}	=	The number of edges for each cluster graph subnet
P_c	=	Edge probability in the cluster subgraph
P_{sb}	=	Edge probability in the same block
P_{db}	=	Edge probability in the different block
$(i+i_0\text{-}1)\zeta$,	=	Node weight in Chung–Lu model
E(deg)	=	Mean degree
P_{sw}	=	Rewiring probability in Watts–Strogatz model
κ	=	Number of conditions in Ω
eig	=	The eigenvalues of $\theta \cdot v$ where v is a parameter that controls the magnitude of partial correlations.
$\theta=\theta \cdot v + (\mid \min(eig) \mid + 0.1 + u)I$	=	The precision matrix calculated in R huge package, where u is a shift parameter for the diagonal, and I the identity diagonal matrix
y	=	True compositional abundance
μ	=	Vector of mean abundances
L_0	=	Lower triangular matrix used in MDiNE simulation approach
U	=	A matrix with independent rows drawn from a multivariate normal distribution.
C_i	=	The weights vector in the *i*-th row of the adjacency matrix used in the exponential growth model
n_c	=	The total number of pairwise correlations
d_{ij}	=	Pairwise absolute difference
d_H	=	Hamming distance
L_1	=	Norm distance between matrix
d_F	=	Frobenius norm distance

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This work was supported in part by MIUR (Italian Ministry for Education) under the initiative "Departments of Excellence" (Law 232/2016)."

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Tremaroli, V.; Bäckhed, F. Functional interactions between the gut microbiota and host metabolism. *Nature,* **2012**, *489*(7415), 242-249. http://dx.doi.org/10.1038/nature11552 PMID: 22972297

[2] Chang, C.S.; Kao, C.Y. Current understanding of the gut microbiota shaping mechanisms. *J. Biomed. Sci.,* **2019**, *26*(1), 59. http://dx.doi.org/10.1186/s12929-019-0554-5 PMID: 31434568

[3] Tasnim, N.; Abulizi, N.; Pither, J.; Hart, M.M.; Gibson, D.L. Linking the gut microbial ecosystem with the environment: does gut health depend on where we live? *Front. Microbiol.,* **2017**, *8*, 1935. http://dx.doi.org/10.3389/fmicb.2017.01935 PMID: 29056933

[4] McDonald, B.; McCoy, K.D. Maternal microbiota in pregnancy and early life. *Science,* **2019**, *365*(6457), 984-985. http://dx.doi.org/10.1126/science.aay0618 PMID: 31488675

[5] Collado, M.C.; Cernada, M.; Baüerl, C.; Vento, M.; Pérez-Martínez, G. Microbial ecology and host-microbiota interactions during early life stages. *Gut Microbes,* **2012**, *3*(4), 352-365. http://dx.doi.org/10.4161/gmic.21215 PMID: 22743759

[6] Rylance, J.; Kankwatira, A.; Nelson, D.E.; Toh, E.; Day, R.B.; Lin, H.; Gao, X.; Dong, Q.; Sodergren, E.; Weinstock, G.M.; Heyderman, R.S.; Twigg, H.L., III; Gordon, S.B. Household air pollution and the lung microbiome of healthy adults in Malawi: a cross-sectional study. *BMC Microbiol.,* **2016**, *16*(1), 182. http://dx.doi.org/10.1186/s12866-016-0803-7 PMID: 27514621

[7] Huang, C.; Shi, G. Smoking and microbiome in oral, airway, gut and some systemic diseases. *J. Transl. Med.,* **2019**, *17*(1), 225. http://dx.doi.org/10.1186/s12967-019-1971-7 PMID: 31307469

[8] Hanski, I.; von Hertzen, L.; Fyhrquist, N.; Koskinen, K.; Torppa, K.; Laatikainen, T.; Karisola, P.; Auvinen, P.; Paulin, L.; Mäkelä, M.J.; Vartiainen, E.; Kosunen, T.U.; Alenius, H.; Haahtela, T. Environmental biodiversity, human microbiota, and allergy are interrelated. *Proc. Natl. Acad. Sci. USA,* **2012**, *109*(21), 8334-8339. http://dx.doi.org/10.1073/pnas.1205624109 PMID: 22566627

[9] Wang, Y.N.; Meng, X.C.; Dong, Y.F.; Zhao, X.H.; Qian, J.M.; Wang, H.Y.; Li, J.N. Effects of probiotics and prebiotics on intestinal microbiota in mice with acute colitis based on 16S rRNA gene sequencing. *Chin. Med. J. (Engl.),* **2019**, *132*(15), 1833-1842. http://dx.doi.org/10.1097/CM9.0000000000000308 PMID: 31268903

[10] Koutrouli, M.; Karatzas, E.; Paez-Espino, D.; Pavlopoulos, G.A. A guide to conquer the biological network era using graph theory. *Front. Bioeng. Biotechnol.,* **2020**, *8*, 34. http://dx.doi.org/10.3389/fbioe.2020.00034 PMID: 32083072

[11] Butte, A.J.; Kohane, I.S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.,* **2000**, *2000*, 418-429. PMID: 10902190

[12] Herrero, J.; Díaz-Uriarte, R.; Dopazo, J. An approach to inferring transcriptional regulation among genes from large-scale expression data. *Comp. Funct. Genomics,* **2003**, *4*(1), 148-154. http://dx.doi.org/10.1002/cfg.237 PMID: 18629097

[13] Schäfer, J.; Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics,* **2005**, *21*(6), 754-764. http://dx.doi.org/10.1093/bioinformatics/bti062 PMID: 15479708

[14] Basso, K.; Margolin, A.A.; Stolovitzky, G.; Klein, U.; Dalla-Favera, R.; Califano, A. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.,* **2005**, *37*(4), 382-390. http://dx.doi.org/10.1038/ng1532 PMID: 15778709

[15] Woese, C.R.; Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA,* **1977**, *74*(11), 5088-5090. http://dx.doi.org/10.1073/pnas.74.11.5088 PMID: 270744

[16] Hiergeist, A.; Reischl, U.; Gessner, A. Priority Program 1656 Intestinal Microbiota Consortium/ quality assessment participants. Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *Int. J. Med. Microbiol.,* **2016**, *306*(5), 334-342. http://dx.doi.org/10.1016/j.ijmm.2016.03.005 PMID: 27052158

[17] Bukin, Y.S.; Galachyants, Y.P.; Morozov, I.V.; Bukin, S.V.; Zakharenko, A.S.; Zemskaya, T.I. The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci. Data,* **2019**, *6*, 190007. http://dx.doi.org/10.1038/sdata.2019.7 PMID: 30720800

[18] Mancabelli, L.; Milani, C.; Lugli, G.A.; Fontana, F.; Turroni, F.; van Sinderen, D.; Ventura, M. The impact of primer design on amplicon-based metagenomic profiling accuracy: detailed insights into bifidobacterial community structure. *Microorganisms,* **2020**, *8*(1), 131. http://dx.doi.org/10.3390/microorganisms8010131 PMID: 31963501

[19] Klindworth, A.; Pruesse, E.; Schweer, T.; Peplies, J.; Quast, C.; Horn, M.; Glöckner, F.O. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.,* **2013**, *41*(1), e1. http://dx.doi.org/10.1093/nar/gks808 PMID: 22933715

[20] Thomas, M.C.; Thomas, D.K.; Selinger, L.B.; Inglis, G.D. spyder, a new method for *in silico* design and assessment of 16S rRNA gene primers for molecular microbial ecology. *FEMS Microbiol. Lett.,* **2011**, *320*(2), 152-159. http://dx.doi.org/10.1111/j.1574-6968.2011.02302.x PMID: 21554380

[21] Sambo, F.; Finotello, F.; Lavezzo, E.; Baruzzo, G.; Masi, G.; Peta, E.; Falda, M.; Toppo, S.; Barzon, L.; Di Camillo, B. Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. *BMC Bioinformatics,* **2018**, *19*(1), 343. http://dx.doi.org/10.1186/s12859-018-2360-6 PMID: 30268091

[22] Martinez-Porchas, M.; Villalpando-Canchola, E.; Ortiz Suarez, L.E.; Vargas-Albores, F. How conserved are the conserved 16S-rRNA regions? *PeerJ,* **2017**, *5*, e3036. http://dx.doi.org/10.7717/peerj.3036 PMID: 28265511

[23] Besser, J.; Carleton, H.A.; Gerner-Smidt, P.; Lindsey, R.L.; Trees, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin. Microbiol. Infect.,* **2018**, *24*(4), 335-341. http://dx.doi.org/10.1016/j.cmi.2017.10.013 PMID: 29074157

[24] Bolyen, E.; Rideout, J.R.; Dillon, M.R.; Bokulich, N.A.; Abnet, C.C.; Al-Ghalith, G.A.; Alexander, H.; Alm, E.J.; Arumugam, M.; Asnicar, F.; Bai, Y.; Bisanz, J.E.; Bittinger, K.; Brejnrod, A.; Brislawn, C.J.; Brown, C.T.; Callahan, B.J.; Caraballo-Rodríguez, A.M.; Chase, J.; Cope, E.K.; Da Silva, R.; Diener, C.; Dorrestein, P.C.; Douglas, G.M.; Durall, D.M.; Duvallet, C.; Edwardson, C.F.; Ernst, M.; Estaki, M.; Fouquier, J.; Gauglitz, J.M.; Gibbons, S.M.; Gibson, D.L.; Gonzalez, A.; Gorlick, K.; Guo, J.; Hillmann, B.; Holmes, S.; Holste, H.; Huttenhower, C.; Huttley, G.A.; Janssen, S.; Jarmusch, A.K.; Jiang, L.; Kaehler, B.D.; Kang, K.B.; Keefe, C.R.; Keim, P.; Kelley, S.T.; Knights, D.; Koester, I.; Kosciolek, T.; Kreps, J.; Langille, M.G.I.; Lee, J.; Ley, R.; Liu, Y.X.; Loftfield, E.; Lozupone, C.; Maher, M.; Marotz, C.; Martin, B.D.; McDonald, D.; McIver, L.J.; Melnik, A.V.; Metcalf, J.L.; Morgan, S.C.; Morton, J.T.; Naimey, A.T.; Navas-Molina, J.A.; Nothias, L.F.; Orchanian, S.B.; Pearson, T.; Peoples, S.L.; Petras, D.; Preuss, M.L.; Pruesse, E.; Rasmussen, L.B.; Rivers, A.; Robeson, M.S., II; Rosenthal, P.; Segata, N.; Shaffer, M.; Shiffer, A.; Sinha, R.; Song, S.J.; Spear, J.R.; Swafford, A.D.; Thompson, L.R.; Torres, P.J.; Trinh, P.; Tripathi, A.; Turnbaugh, P.J.; Ul-Hasan, S.; van der Hooft, J.J.J.; Vargas, F.; Vázquez-Baeza, Y.; Vogtmann, E.; von Hippel, M.; Walters, W.; Wan, Y.; Wang, M.; Warren, J.; Weber, K.C.; Williamson, C.H.D.; Willis, A.D.; Xu, Z.Z.; Zaneveld, J.R.; Zhang, Y.; Zhu, Q.; Knight, R.; Caporaso, J.G. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.,* **2019**, *37*(8), 852-857. http://dx.doi.org/10.1038/s41587-019-0209-9 PMID: 31341288

[25] Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J.; Sahl, J.W.; Stres, B.; Thallinger, G.G.; Van Horn, D.J.; Weber, C.F. Introducing mothur: open-source, platform-independent, community-supported software for describing and

comparing microbial communities. *Appl. Environ. Microbiol.,* **2009**, *75*(23), 7537-7541.
http://dx.doi.org/10.1128/AEM.01541-09 PMID: 19801464

[26] Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* **2010**, *26*(19), 2460-2461.
http://dx.doi.org/10.1093/bioinformatics/btq461 PMID: 20709691

[27] Wang, Q.; Garrity, G.M.; Tiedje, J.M.; Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.,* **2007**, *73*(16), 5261-5267.
http://dx.doi.org/10.1128/AEM.00062-07 PMID: 17586664

[28] Rognes, T.; Flouri, T.; Nichols, B.; Quince, C.; Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ,* **2016**, *4*, e2584.
http://dx.doi.org/10.7717/peerj.2584 PMID: 27781170

[29] Xue, Z.; Kable, M.E.; Marco, M.L. Impact of DNA sequencing and analysis methods on 16S rRNA gene bacterial community analysis of dairy products. *MSphere,* **2018**, *3*(5), e00410-e00418.
http://dx.doi.org/10.1128/mSphere.00410-18 PMID: 30333179

[30] Prodan, A.; Tremaroli, V.; Brolin, H.; Zwinderman, A.H.; Nieuwdorp, M.; Levin, E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One,* **2020**, *15*(1), e0227434.
http://dx.doi.org/10.1371/journal.pone.0227434 PMID: 31945086

[31] Gloor, G.B.; Macklaim, J.M.; Pawlowsky-Glahn, V.; Egozcue, J.J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.,* **2017**, *8*, 2224.
http://dx.doi.org/10.3389/fmicb.2017.02224 PMID: 29187837

[32] Quinn, T.P.; Erb, I.; Richardson, M.F.; Crowley, T.M. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics,* **2018**, *34*(16), 2870-2878.
http://dx.doi.org/10.1093/bioinformatics/bty175 PMID: 29608657

[33] Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. B,* **1982**, *44*(2), 139-160.
http://dx.doi.org/10.1111/j.2517-6161.1982.tb01195.x

[34] Egozcue, J.J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math. Geol.,* **2003**, *35*, 279-300.
http://dx.doi.org/10.1023/A:1023818214614

[35] Martín-Fernández, J.A.; Barceló-Vidal, C.; Pawlowsky-Glahn, V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.,* **2003**, *35*(3), 253-278.
http://dx.doi.org/10.1023/A:1023866030544

[36] Palarea-Albaladejo, J.; Martín-Fernández, J.A. zCompositions- R package for multivariate imputation of left-censored data under a compositional approach. *Chemom. Intell. Lab. Syst.,* **2015**, *143*, 85-96.
http://dx.doi.org/10.1016/j.chemolab.2015.02.019

[37] Hron, K.; Templ, M.; Filzmoser, P. Imputation of missing values for compositional data using classical and robust methods. *Comput. Stat. Data Anal.,* **2010**, *54*(12), 3095-3107.
http://dx.doi.org/10.1016/j.csda.2009.11.023

[38] Templ, M.; Hron, K.; Filzmoser, P. robCompositions: An R-package for robust statistical analysis of compositional data. *Compositional Data Analysis, Eds.; V. Pawlowsky-Glahn and A. Buccianti*; , **2011**.

[39] Chen, L.; Reeve, J.; Zhang, L.; Huang, S.; Wang, X.; Chen, J. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ,* **2018**, *6*, e4600.
http://dx.doi.org/10.7717/peerj.4600 PMID: 29629248

[40] Kumar, M.S.; Slud, E.V.; Okrah, K.; Hicks, S.C.; Hannenhalli, S.; Corrada Bravo, H. Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics,* **2018**, *19*(1), 799.
http://dx.doi.org/10.1186/s12864-018-5160-5 PMID: 30400812

[41] Finotello, F.; Mastrorilli, E.; Di Camillo, B. Measuring the diversity of the human microbiota with targeted next-generation sequencing. *Brief. Bioinform.,* **2018**, *19*(4), 679-692.
PMID: 28025179

[42] Wong, R.G.; Wu, J.R.; Gloor, G.B. Expanding the UniFrac toolbox. *PLoS One,* **2016**, *11*(9), e0161196.
http://dx.doi.org/10.1371/journal.pone.0161196 PMID: 27632205

[43] Lê Cao, K.A.; Costello, M.E.; Lakis, V.A.; Bartolo, F.; Chua, X.Y.; Brazeilles, R.; Rondeau, P.; Mix, M.C. A multivariate statistical framework to gain insight into microbial communities. *PLoS One,* **2016**, *11*(8), e0160169.
http://dx.doi.org/10.1371/journal.pone.0160169 PMID: 27513472

[44] Fernandes, A.D.; Reid, J.N.; Macklaim, J.M.; McMurrough, T.A.; Edgell, D.R.; Gloor, G.B. 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.
http://dx.doi.org/10.1186/2049-2618-2-15 PMID: 24910773

[45] Mandal, S.; Van Treuren, W.; White, R.A.; Eggesbø, M.; Knight, R.; Peddada, S.D. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.,* **2015**, *26*, 27663.
PMID: 26028277

[46] Pendegraft, A.H.; Guo, B.; Yi, N. Bayesian hierarchical negative binomial models for multivariable analyses with applications to human microbiome count data. *PLoS One,* **2019**, *14*(8), e0220961.
http://dx.doi.org/10.1371/journal.pone.0220961 PMID: 31437194

[47] Riquelme, E.; Zhang, Y.; Zhang, L.; Montiel, M.; Zoltan, M.; Dong, W.; Quesada, P.; Sahin, I.; Chandra, V.; San Lucas, A.; Scheet, P.; Xu, H.; Hanash, S.M.; Feng, L.; Burks, J.K.; Do, K.A.; Peterson, C.B.; Nejman, D.; Tzeng, C.D.; Kim, M.P.; Sears, C.L.; Ajami, N.; Petrosino, J.; Wood, L.D.; Maitra, A.; Straussman, R.; Katz, M.; White, J.R.; Jenq, R.; Wargo, J.; McAllister, F. Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell,* **2019**, *178*(4), 795-806.e12.
http://dx.doi.org/10.1016/j.cell.2019.07.008 PMID: 31398337

[48] Hawinkel, S.; Mattiello, F.; Bijnens, L.; Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.,* **2019**, *20*(1), 210-221.
http://dx.doi.org/10.1093/bib/bbx104 PMID: 28968702

[49] Faust, K.; Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.,* **2012**, *10*(8), 538-550.
http://dx.doi.org/10.1038/nrmicro2832 PMID: 22796884

[50] Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science,* **2011**, *334*(6062), 1518-1524.
http://dx.doi.org/10.1126/science.1205438 PMID: 22174245

[51] Faust, K.; Sathirapongsasuti, J.F.; Izard, J.; Segata, N.; Gevers, D.; Raes, J.; Huttenhower, C. Microbial co-occurrence relationships in the human microbiome. *PLOS Comput. Biol.,* **2012**, *8*(7), e1002606.
http://dx.doi.org/10.1371/journal.pcbi.1002606 PMID: 22807668

[52] Sarkar, S.K.; Chang, C. The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Am. Stat. Assoc.,* **1997**, *92*(440), 1601-1608.
http://dx.doi.org/10.1080/01621459.1997.10473682

[53] Faust, K.; Raes, J. CoNet app: inference of biological association networks using Cytoscape. *F1000 Res.,* **2016**, *5*, 1519.
http://dx.doi.org/10.12688/f1000research.9050.1 PMID: 27853510

[54] Brown, M.B. 400: A method for combining non-independent, one-sided tests of significance. *Biometrics,* **1975**, *31*(4), 987-992.
http://dx.doi.org/10.2307/2529826 PMID: 1203428

[55] Yang, P.; Yu, S.; Cheng, L.; Ning, K. Meta-network: optimized species-species network analysis for microbial communities. *BMC Genomics,* **2019**, *20*(S2)(Suppl. 2), 187.
http://dx.doi.org/10.1186/s12864-019-5471-1 PMID: 30967118

[56] Chua, H.N.; Sung, W.K.; Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics,* **2006**, *22*(13), 1623-1630.
http://dx.doi.org/10.1093/bioinformatics/btl145 PMID: 16632496

[57] Friedman, J.; Alm, E.J. Inferring correlation networks from genomic survey data. *PLOS Comput. Biol.,* **2012**, *8*(9), e1002687.
http://dx.doi.org/10.1371/journal.pcbi.1002687 PMID: 23028285

[58] Jost, L. Entropy and diversity. *Oikos,* **2006**, *113*(2), 363-375.
http://dx.doi.org/10.1111/j.2006.0030-1299.14714.x

[59] Fang, H.; Huang, C.; Zhao, H.; Deng, M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics,* **2015**, *31*(19), 3172-3180.
http://dx.doi.org/10.1093/bioinformatics/btv349 PMID: 26048598

[60] Ban, Y.; An, L.; Jiang, H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics,* **2015**, *31*(20), 3322-3329.
http://dx.doi.org/10.1093/bioinformatics/btv364 PMID: 26079350

[61] Shah, R.D.; Samworth, R.J. Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.,* **2013**, *75*(1), 55-80. http://dx.doi.org/10.1111/j.1467-9868.2011.01034.x

[62] Friedman, J.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics,* **2008**, *9*(3), 432-441. http://dx.doi.org/10.1093/biostatistics/kxm045 PMID: 18079126

[63] Meinshausen, N.; Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.,* **2006**, *34*(3), 1436-1462. http://dx.doi.org/10.1214/009053606000000281

[64] Kurtz, Z.D.; Müller, C.L.; Miraldi, E.R.; Littman, D.R.; Blaser, M.J.; Bonneau, R.A. Sparse and compositionally robust inference of microbial ecological networks. *PLOS Comput. Biol.,* **2015**, *11*(5), e1004226. http://dx.doi.org/10.1371/journal.pcbi.1004226 PMID: 25950956

[65] Liu, H.; Roeder, K.; Wasserman, L. In: Stability approach to regularization selection (stars) for high dimensional graphical models. *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS),* , pp. 1-14.**2010**

[66] Fang, H.; Huang, C.; Zhao, H.; Deng, M. gCoda: conditional dependence network inference for compositional data. *J. Comput. Biol.,* **2017**, *24*(7), 699-708. http://dx.doi.org/10.1089/cmb.2017.0054 PMID: 28489411

[67] Lo, C.; Marculescu, R. MPLasso: Inferring microbial association networks using prior microbial knowledge. *PLOS Comput. Biol.,* **2017**, *13*(12), e1005915. http://dx.doi.org/10.1371/journal.pcbi.1005915 PMID: 29281638

[68] Lim, K.M.K.; Li, C.; Chng, K.R.; Nagarajan, N. @MInter: automated text-mining of microbial interactions. *Bioinformatics,* **2016**, *32*(19), 2981-2987. http://dx.doi.org/10.1093/bioinformatics/btw357 PMID: 27312413

[69] Yoon, G.; Gaynanova, I.; Müller, C.L. Microbial networks in SPRING - semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. Genet.,* **2019**, *10*, 516. http://dx.doi.org/10.3389/fgene.2019.00516 PMID: 31244881

[70] Yoon, G.; Carroll, R.J.; Gaynanova, I Sparse semiparametric canonical correlation analysis for data of mixed types *ArXiv e-prints,* **2018**.https://arxiv.org/abs/1807.05274

[71] Schwager, E.; Mallick, H.; Ventz, S.; Huttenhower, C. A Bayesian method for detecting pairwise associations in compositional data. *PLOS Comput. Biol.,* **2017**, *13*(11), e1005852. http://dx.doi.org/10.1371/journal.pcbi.1005852 PMID: 29140991

[72] McGregor, K.; Labbe, A.; Greenwood, C.M.T. MDiNE: a model to estimate differential co-occurrence networks in microbiome studies. *Bioinformatics,* **2020**, *36*(6), 1840-1847. http://dx.doi.org/10.1093/bioinformatics/btz824 PMID: 31697315

[73] Deng, Y.; Jiang, Y.H.; Yang, Y.; He, Z.; Luo, F.; Zhou, J. Molecular ecological network analyses. *BMC Bioinformatics,* **2012**, *13*(1), 113. http://dx.doi.org/10.1186/1471-2105-13-113 PMID: 22646978

[74] Veech, J.A. A probabilistic model for analysing species co-occurrence. *Glob. Ecol. Biogeogr.,* **2013**, *22*(2), 252-260. http://dx.doi.org/10.1111/j.1466-8238.2012.00789.x

[75] Griffith, D.M.; Veech, J.A. Marsh, CJ cooccur: Probabilistic species co-occurrence analysis. *R.J. Stat. Soft.,* **2016**, *69*(2), 1-17.

[76] Adair, K.L.; Wilson, M.; Bost, A.; Douglas, A.E. Microbial community assembly in wild populations of the fruit fly Drosophila melanogaster. *ISME J.,* **2018**, *12*(4), 959-972. http://dx.doi.org/10.1038/s41396-017-0020-x PMID: 29358735

[77] Villette, P.; Afonso, E.; Couval, G.; Levret, A.; Galan, M.; Goydadin, A.C.; Cosson, J.F.; Giraudoux, P. Spatio-temporal trends in richness and persistence of bacterial communities in decline-phase water vole populations. *Sci. Rep.,* **2020**, *10*(1), 9506. http://dx.doi.org/10.1038/s41598-020-66107-5 PMID: 32528097

[78] Xia, L.C.; Ai, D.; Cram, J.; Fuhrman, J.A.; Sun, F. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics,* **2013**, *29*(2), 230-237. http://dx.doi.org/10.1093/bioinformatics/bts668 PMID: 23178636

[79] Tsai, K.N.; Lin, S.H.; Liu, W.C.; Wang, D. Inferring microbial interaction network from microbiome data using RMN algorithm. *BMC Syst. Biol.,* **2015**, *9*(1), 54. http://dx.doi.org/10.1186/s12918-015-0199-2 PMID: 26337930

[80] Xu, H.; Wu, P.; Wu, C.F.J.; Tidwell, C.; Wang, Y. A smooth response surface algorithm for constructing a gene regulatory network. *Physiol. Genomics,* **2002**, *11*(1), 11-20. http://dx.doi.org/10.1152/physiolgenomics.00060.2001 PMID: 12361986

[81] Fisher, C.K.; Mehta, P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One,* **2014**, *9*(7), e102451. http://dx.doi.org/10.1371/journal.pone.0102451 PMID: 25054627

[82] Shaw, G.T.; Pao, Y.Y.; Wang, D. MetaMIS: a metagenomic microbial interaction simulator based on microbial community profiles. *BMC Bioinformatics,* **2016**, *17*(1), 488. http://dx.doi.org/10.1186/s12859-016-1359-0 PMID: 27887570

[83] Lo, C.; Marculescu, R. Inferring microbial interactions from metagenomic time-series using prior biological knowledge. *ACM-BCB'17: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics,* Aug 20-23, **2017**.Boston, MA, USA. http://dx.doi.org/10.1145/3107411.3107435

[84] Ridenhour, B.J.; Brooker, S.L.; Williams, J.E.; Van Leuven, J.T.; Miller, A.W.; Dearing, M.D.; Remien, C.H. Modeling time-series data from microbial communities. *ISME J.,* **2017**, *11*(11), 2526-2537. http://dx.doi.org/10.1038/ismej.2017.107 PMID: 28786973

[85] Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica,* **1969**, *37*(3), 424-438. http://dx.doi.org/10.2307/1912791

[86] Wen, X.; Rangarajan, G.; Ding, M. Multivariate Granger causality: an estimation framework based on factorization of the spectral density matrix. *Philos. Trans.- Royal Soc., Math. Phys. Eng. Sci.,* **2013**, *371*(1997), 20110610. http://dx.doi.org/10.1098/rsta.2011.0610 PMID: 23858479

[87] Baksi, K.D.; Kuntal, B.K.; Mande, S.S. 'TIME': a web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Front. Microbiol.,* **2018**, *9*, 36. http://dx.doi.org/10.3389/fmicb.2018.00036 PMID: 29416530

[88] Hirano, H.; Takemoto, K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics,* **2019**, *20*(1), 329. http://dx.doi.org/10.1186/s12859-019-2915-1 PMID: 31195956

[89] Chen, Y.; Bressler, S.L.; Ding, M. Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. *J. Neurosci. Methods,* **2006**, *150*(2), 228-237. http://dx.doi.org/10.1016/j.jneumeth.2005.06.011 PMID: 16099512

[90] Barabasi, A.L.; Albert, R. Emergence of scaling in random networks. *Science,* **1999**, *286*(5439), 509-512. http://dx.doi.org/10.1126/science.286.5439.509 PMID: 10521342

[91] Chung, F.; Lu, L. Connected components in random graphs with given expected degree sequences. *Ann. Combin.,* **2002**, *6*(2), 125-145. http://dx.doi.org/10.1007/PL00012580

[92] Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature,* **1998**, *393*(6684), 440-442. http://dx.doi.org/10.1038/30918 PMID: 9623998

[93] Zhao, T.; Liu, H.; Roeder, K.; Lafferty, J.; Wasserman, L. The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.,* **2012**, *13*(37), 1059-1062. PMID: 26834510

[94] Nelsen, R.B. *An introduction to copulas. Springer Series in Statistics*; Springer, **1999**. http://dx.doi.org/10.1007/978-1-4757-3076-0

[95] Marioni, J.C.; Mason, C.E.; Mane, S.M.; Stephens, M.; Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.,* **2008**, *18*(9), 1509-1517. http://dx.doi.org/10.1101/gr.079558.108 PMID: 18550803

[96] Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.,* **2010**, *11*(10), R106. http://dx.doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621

[97] La Rosa, P.S.; Brooks, J.P.; Deych, E.; Boone, E.L.; Edwards, D.J.; Wang, Q.; Sodergren, E.; Weinstock, G.; Shannon, W.D. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One,* **2012**, *7*(12), e52078.

http://dx.doi.org/10.1371/journal.pone.0052078 PMID: 23284876

[98] Xu, L.; Paterson, A.D.; Turpin, W.; Xu, W. Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One,* **2015**, *10*(7), e0129606. http://dx.doi.org/10.1371/journal.pone.0129606 PMID: 26148172

[99] Lambert, D. Zero-Inflated poisson regression, with an application to defects in manufacturing. *Technometrics,* **1992**, *34*(1), 1-14. http://dx.doi.org/10.2307/1269547

[100] Patuzzi, I.; Baruzzo, G.; Losasso, C.; Ricci, A.; Di Camillo, B. metaSPARSim: a 16S rRNA gene sequencing count data simulator. *BMC Bioinformatics,* **2019**, *20*(Suppl. 9), 416. http://dx.doi.org/10.1186/s12859-019-2882-6 PMID: 31757204

[101] Yee, T.W. The VGAM package for categorical data analysis. *J. Stat. Softw.,* **2010**, *32*(10), jss.v032.i10. http://dx.doi.org/10.18637/jss.v032.i10

[102] Ren, B.; Schwager, E.; Tickle, T.L.; Huttenhower, C. sparseDOSSA: Sparse data observations for simulating synthetic abundance. **2016**.https://huttenhower.sph.harvard.edu/sparsedossa

[103] Zhang, X.; Wu, J.; Tan, Y.; Deng, H.; Li, Y. Structural robustness of weighted complex networks based on natural connectivity. *Chin. Phys. Lett.,* **2013**, *30*(10), 108901. http://dx.doi.org/10.1088/0256-307X/30/10/108901