BMC Bioinformatics

RESEARCH

Open Access

Filtering ASVs/OTUs via mutual information-based microbiome network analysis

Elham Bayat Mokhtari and Benjamin Jerry Ridenhour*

*Correspondence: bridenhour@uidaho.edu

Department of Mathematics and Statistical Science, University of Idaho, Moscow, ID, USA

Abstract

Microbial communities are widely studied using high-throughput sequencing techniques, such as 16S rRNA gene sequencing. These techniques have attracted biologists as they offer powerful tools to explore microbial communities and investigate their patterns of diversity in biological and biomedical samples at remarkable resolution. However, the accuracy of these methods can negatively affected by the presence of contamination. Several studies have recognized that contamination is a common problem in microbial studies and have offered promising computational and laboratory-based approaches to assess and remove contaminants. Here we propose a novel strategy, MI-based (mutual information based) filtering method, which uses information theoretic functionals and graph theory to identify and remove contaminants. We applied MI-based filtering method to a mock community data set and evaluated the amount of information loss due to filtering taxa. We also compared our method to commonly practice traditional filtering methods. In a mock community data set, MI-based filtering approach maintained the true bacteria in the community without significant loss of information. Our results indicate that MI-based filtering method effectively identifies and removes contaminants in microbial communities and hence it can be beneficial as a filtering method to microbiome studies. We believe our filtering method has two advantages over traditional filtering methods. First, it does not required an arbitrary choice of threshold and second, it is able to detect true taxa with low abundance.

Keywords: Contamination, Microbiome, 16S rRNA, Mutual information, Graph theory

Introduction

High-throughput sequencing approaches are some of the most powerful tools for studying and characterizing microbial communities. Bacterial phylogeny and taxonomy can be characterized using marker genes, such as 16S rRNA gene sequences which are present in all bacteria, and it is sufficiently large for informatics and analysis purposes [20, 30]. However, the potential for contamination which is defined as non-intended introduction of bacteria during sample collection, DNA extraction, and PCR amplification into the sample of interest is high; thus a low signal-to-noise ratio poses a major problem in analyses of such data [7, 33, 42]. Contamination is particularly problematic when studying low yield samples because of significant impacts on results [33, 42]. Therefore, it is necessary to identify, minimize, and filter contaminants as a potential source of bias that leads to skew data analysis.

Attempts to experimentally control or eliminate sources of contamination can be challenging if not impossible. To minimize or identify contamination, strategies such as inclusion of negative controls or blanks for every batch of samples and use of them through the entire extraction, amplification, or library preparations have been suggested [5, 33]. One of the advantages of sequencing the blanks is the ability to detect and quantify the levels of contamination as well as their the sources. [5, 9, 27, 31, 33]. However, including an appropriate negative control is not always easy and in the majority of microbiome published studies controls have not been included [18, 42] and [33] recommended keeping records of kits and other reagents, performing technical replicates, and using sample randomization across kits and PCR runs into control measurement error. Some researchers have proposed using mock communities as a positive control during extraction, amplification, and sequencing alongside experimental samples [7]. Positive controls are commercially available in the form of defined communities, however their validity for a particular microbiome research is not guaranteed and standardized protocols for designing positive controls might not be available [18].

None of the above experimental methods are capable of eliminating existing contaminants completely, easily, and reliably in all cases. Therefore, strategies that use the power of bioinformatics and statistical methods to clean sequencing data must be introduced. For example, [21] identified and removed Operational Taxonomic Units (OTUs) as potential contaminants if they have strong negative correlation with amplicon counts after 16S library preparation. However, in many cases, contaminant OTUs might occur on the host as well as being present as contamination and therefore, this leads to a higher than desired false positive rate. Ad hoc methods such as removing genes or taxa with total read count or percentage smaller than or below an empirical threshold across all samples [2, 23, 32, 39, 43] are easy to implement and relatively common among microbiome studies. However, choosing an appropriate filtering threshold is a complex problem by itself and an arbitrary choice can bias the results. In addition, the impact of taxa or genes is not directly proportional to their numeric abundance and there might be biological signal among rare taxa—or genes—that is of interest; thus removing low abundance taxa could lead to loss of important information.

The `decontam` package in R introduced by [8] has been developed to identify contaminants using statistical models. [8] demonstrated the accuracy of their method to remove contaminants from a data set generated by [33]. However, a major limitation of `decontam` is that it assumes contaminants and true signals are distinct from one another, and this assumption is violated in the case of cross-contamination due to sequences from pooled samples. [26] developed the R package `microDecon` which is based on proportions of contaminant OTUs or Amplicon Sequence Variants (ASVs) in blank samples to identify and remove contaminant reads from meta-barcoding data. They demonstrated that their method is robust to both high and low contamination levels. They also showed that their approach can recover the real community from the contaminant community even with a large overlap between the two. However, similar to [8], in case of the existence of cross-contamination, this method is not effective as it assumes a common source of contamination. Recently, [36] introduced the R package `PERFect` for microbiome filtering using covariance matrices and compared them to traditional filtering procedures. They showed that for a very strong signal, `PERFect` provides a more effective contaminant reduction when the signal-to-noise ratio is high. A limitation of their methods is that it is skewed toward retaining dominant taxa, however, this is a common limitation among any filtering methods that does not take into account other types of information such as knowledge about blanks or negative controls.

Here, we propose and validate a method to identify and remove non-bacterial signals that are observed due to contamination or sequencing errors in microbiome data. We use the fact that bacteria live in communities where they rely on one another, and their interactions or coexistence are major drivers of microbial community and function. We utilize a graph model to represent and characterize these interactions and/or coexistence by assuming each taxon is a node and pairwise-bacterial associations are edges in this biological network. We use an information theoretic functional to estimate the strength of these interactions and remove isolated taxa that are not informative to the network as potential noise. We apply permutation and bootstrap based hypothesis testing to measure the probability of increase in information loss due to taxa removal is random. We validate our method using the [7] mock community data set. Finally, we compare the performance of commonly used ad hoc filtering methods with our proposed method.

The rest of this paper is organized as follows. In Sect. 2, we introduce our filtering method using graph models and information loss measurement. Statistical inference based on bootstrap and permutation hypothesis testing is presented in Sect. 2. Method validation and comparison with traditional filtering methods using [7] data set are provided in Sects. 3 and 4, respectively. Finally we conclude the paper in Sect. 5.

Materials and methods

We propose a method to identify and remove contaminated sequence reads from data sets, while accounting for the amount of information loss due to this removal. Note that the proposed method can be applied to both OTU and ASV count tables.

Mathematical definition

Here, we define notations which will be frequently used in the following sections. Consider a high-dimensional count matrix where each input represents the count of sequence reads of a taxon, which, for simplicity, we will assume to be a bacterial species or strain. Let $X_{n\times m}$ be a microbial abundance matrix. For each $i = 1, \cdots, n$ and $j = 1, \cdots, m$, let x_{ij} be the observed count of the j-th taxon in the i-th sample and X_j denotes the abundance of the j-th taxon across all n samples. Generally, the number of samples is considerably less than the number of taxa, that is $n << m$.

The proposed method: network-based contaminant identification

Graph theory is an important concept in statistics and can be used to describe the relationships between random variables [24, 40]. A network (or a graph) is defined as a set of nodes connected by edges [28]. Microbial interactions can be represented as a connectivity network, where nodes correspond to taxa and the edges represent the associations between taxa [41]. One potential association measure is mutual information (MI) which is a non-directional connectivity measure. MI was introduced by Shannon in 1948 [34] as a measure of statistical dependence between two random variables. Unlike Pearson or Spearman correlation coefficients, the most widely used association measures, that quantify linear and monotonic relationships, respectively, MI can be used to estimate non-linear relationships [10, 37].

MI measures the expected reduction in uncertainty about X that results from learning Y, or vice versa. This quantity can be formulated as

$$I(X;Y) = H(X) - H(X|Y), \tag{1}$$

where $H(X)$, known as "entropy," is the average amount of information, or surprise, a variable X has. It is defined to be

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x), \tag{2}$$

where p is the probability of observing the i-th value of the bin measurement data $x_i \in \mathcal{X}$ using partition-based methods such as histograms. The conditional entropy is the uncertainty of X given Y and it is formulated as

$$H(X|Y=y) = -\sum_{x\in\mathcal{X}} \frac{p(x,y)}{p(y)}\log\frac{p(x,y)}{p(y)}, \tag{3}$$

where $p(x, y)$ is the joint probability density of measurements X and Y.

From equation (3) we can derive the following identity

$$H(X|Y) = H(X,Y) - H(X), \tag{4}$$

where $H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y)$ is the joint entropy which measures the amount of uncertainty in the two random variables X and Y taken together.

MI possesses the following desirable properties.

1. It is symmetric: $I(X;Y) = I(Y;X)$,
2. $I(X;Y) \geq 0$, equality holds if and only if the two variables are independent,
3. $I(X;Y) \leq H(X,Y)$.

In situations where X is uniquely determined by Y, knowledge of Y dictates a single possible value of X. It then follows that the conditional entropy satisfies $H(X|Y) = 0$ and therefore MI has the maximum value of $I(X;Y) = H(X)$. Moreover, the stronger the relationship between two variables, the greater is the MI. Kinney and Atwal [22] proved that MI places the same importance on linear and nonlinear dependence.

Here, we use MI as an association measure and transform it into network adjacencies. A network adjacency $A = (A_{ij})$ satisfies the following conditions:

1. $0 \leq A_{ij} \leq 1$,
2. $A_{ij} = A_{ji}$,
3. $A_{ii} = 1$.

For m taxa $X_1, \cdots, X_m$ an adjacency matrix $\mathcal{I}$ is a m by m matrix where each entry is the amount of information shared between each pair of taxa. We construct our adjacency matrix based on MI by satisfying three above conditions: (1) transformation to [0, 1]; (2) symmetrization; and, (3) setting diagonal values to 1. It can be easily seen that MI is bounded below by 0 and it is symmetric. However, it is not bounded above by 1 and the diagonals are not equal to 1 but rather are the entropy of the variable, $H(X)$. To satisfy the above conditions, we divide each entry of the mutual information matrix $\mathcal{I}$ by one of its upper bound which is a joint entropy between each pair of taxa, resulting in adjusted adjacency matrix $\widetilde{\mathcal{I}}$.

Therefore, for each pair of taxa X_j and $X_{j'}$, the adjusted mutual information is calculated as

$$\widetilde{\mathcal{I}}_{jj'}(X_j; X_{j'}) = \frac{\mathcal{I}_{jj'}(X_j; X_{j'})}{H(X_j, X_{j'})}, \quad j, j' = 1, \cdots, m \tag{5}$$

The result of this transformation is a m by m matrix $\widetilde{\mathcal{I}}$ where each entry varies between 0 and 1. Also, if $j = j'$, then $\mathcal{I}_{jj'}(X_j; X_{j'}) = H(X_j)$ and $H(X_j, X_{j'}) = H(X_j)$ so $\widetilde{\mathcal{I}}_{jj'}(X_j; X_{j'}) = 1$. Thus our transformation (5) satisfies the conditions of a network adjacency.

In the following subsection, we describe an approach that results in an unweighted adjacency matrix based on the adjusted mutual information measure we defined above.

Filtering using unweighted network adjacency

A filtered unweighted network adjacency between taxa X_j and X_j' can be defined by hard thresholding the adjusted mutual information-based adjacency matrix $\widetilde{\mathcal{I}}$ using signum function.

$$\mathcal{I}^*_{jj'}(X_j; X_j') = \begin{cases} 1 & \text{if } \widetilde{\mathcal{I}}_{jj'}(X_j; X_j') \geq \tau \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where τ is the hard threshold parameter. Hard thresholding leads to the intuitive concept of taxa connectivity (i.e., a binary variable indicating whether two species do or do not interact), and it is commonly used to construct sparse covariance matrices [38, 44].

Choosing the threshold τ

In many biological networks, hard thresholding of the association adjacency matrix is based on the scale-free criteria (defined below) of a graph and often applied when $m << n$ [1, 3, 44]. In other words, it is assumed that the probability that a node is connected with k other nodes (the degree distribution of a network) is characterized by a power-law distribution

$$P(k) \sim k^{-\gamma}, \tag{7}$$

where k is the node degree, and γ is some exponent reported in some biological graphs to be $2 < \gamma < 3$ [4]. We choose the threshold τ by fitting a linear function $f(k) = -\hat{\gamma}k + \hat{b}$ to the empirical degree distribution in log space and estimating the coefficient of variation, (R^2), of the fit. We choose the threshold that results in the highest R^2 value . In addition to high R^2 values, it is recommended [38, 44] to have a high mean connectivity so that the network contains enough information. We compute the mean degree $\bar{k}$ for each threshold τ, by taking the average over the degree of all nodes. It is expressed as follows

$$\bar{k} = \frac{\sum_{j'=1}^{m} \sum_{j=1}^{m} \mathcal{I}^{*}_{jj'}}{m}$$

We use mean connectivity as a tie breaker for thresholds that could produce the same R^2 value. Choosing an appropriate threshold which provides us with the highest R^2 and a high $\bar{k}$, we build our network based on $\mathcal{I}^{*}_{jj'}$ and remove isolated nodes (taxa), i.e., nodes that have a connectivity degree of 0. Because isolated nodes do not share information with other taxa, we assume they are potential contaminants, and we may remove them without significant loss of information. Conversely, nodes (taxa) that create non-trivial subgraphs (i.e., subgraphs having more than one node) are assumed to be true taxa.

Subnetworks with minimal information loss

Hidaka et al. [17] proposed a method of searching graph partitions (separations of the vertex set) which leads to the minimal information loss. In another work, Smirnova et al. [36] proposed a filtering loss measure to remove taxa with insignificant contribution to the total covariance. Inspired by the these ideas, we propose a method to filter taxa in a network based on total mutual information.

To do this, first we define the connectivity degree d_j of the j-th node for $j = 1, \cdots, m$ in the weighted graph; this is the sum of the weights on all edges adjacent to node j. The formula for connectivity degree d_j is

$$d_j = \sum_{j'=1}^{m} \widetilde{\mathcal{I}}_{jj'}, \tag{8}$$

where we take $\widetilde{\mathcal{I}}_{jj'}$ to be the weight on the edge connecting nodes j and j'. Next, we sort the connectivity degree d_j in an increasing order. Following this, we remove nodes (taxa) based on sample quantile values of sorted connectivity degrees for all taxa $j = 1, \cdots, m$. Finally, we compute the information loss according to the following formula:

$$\Lambda_k = 1 - \frac{\|\widetilde{\mathcal{I}}'_k\|_F^2}{\|\widetilde{\mathcal{I}}\|_F^2}, \tag{9}$$

where $\| \cdot \|_F^2$ is the Frobenius norm, sometimes also called the Euclidean norm, $\widetilde{\mathcal{I}}'_k$ is the adjusted mutual information matrix after removing all taxa below the k^{th} quantile. Here, $\|\widetilde{\mathcal{I}}'_k\|_F^2$ represents the total information shared between taxa after removing certain number of taxa.

Statistical inference: hypothesis testing

Hypothesis testing using a permutation test

In this subsection we present an algorithm based on permutation testing, described in Algorithm 1, inspired by François et al. [14] to compare the difference in information loss due to various quantile removal. Let $q_1, \cdots, q_\ell$ be the quantile values. We are interested in testing if the information loss by removing the taxa with degree less than q_k is significantly different from removing taxa with degree less than q_{k+1}, i.e., $H_0 : \Lambda_k = \Lambda_{k+1}$. A permutation test is a non-parametric hypothesis test [16] and is commonly used to assess the statistical significance when the distribution of the test statistic is not known and needs to be empirically derived. Here, we introduce essential notations for Algorithm 1.

For all $1 \leq k \leq \ell$, define $\widetilde{\mathcal{I}}'_k$ to be the $\widetilde{\mathcal{I}}$ after removing taxa with degree less than q_k, and let r_k be the number of taxa removed. Let $\Delta_{k+1} = \Lambda_{k+1} - \Lambda_k$. If D is any subset of the columns of the full OTU table, define $\widetilde{\mathcal{I}}_D$ as the adjusted mutual information matrix of D.

```
Algorithm 1: ALGORITHM 1
  Input: significance level $\alpha$, number of permutations $M$, quantile vector
    $\boldsymbol{q} = (q_1, \cdots, q_\ell)$.
  Define : $\hat{\boldsymbol{\theta}} = (\Delta_{k+1})$   for   $1 \leq k \leq \ell - 1$; $\boldsymbol{\Theta} = [\hat{\boldsymbol{\theta}}^*_1, \hat{\boldsymbol{\theta}}^*_2, \cdots, \hat{\boldsymbol{\theta}}^*_M]$ .
  Calculate $||\widetilde{\mathfrak{I}}||^2_F$.
  Calculate $\hat{\boldsymbol{\theta}}$.
  For permutation $m = 1, \cdots, M$
    (1) Randomly shuffle columns of taxonomy count, call this matrix $D$.
    (2) For each k:
        (a) Remove the first $r_k$ columns from $D$.
        (b) Compute $\Lambda_k$ using $\widetilde{\mathfrak{I}}_D$.
    (3) Calculate the $m$-th column of $\boldsymbol{\Theta}$, such that $\hat{\boldsymbol{\theta}}^*_m = (\Delta_{k+1})$.
  For each $k$, compute $p_k = \frac{\#\boldsymbol{\Theta}_k \geq \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}}{M}$, where $\boldsymbol{\Theta}_k$ is the $k$-th row of $\boldsymbol{\Theta}$.
  Calculate p-values: $\boldsymbol{p} = (p_k)$.
  Find the index of the first entry of $\boldsymbol{p}$ that is less than or equal to $\alpha$. Call this index $ind$.
  Remove all taxa with degree less than $q_{ind-1}$.
```

Hypothesis testing using bootstrap

In the previous subsection we described a permutation test as a useful hypothesis testing tool. Here we use bootstrap methods [13], Algorithm 2, to test the same hypothesis. Again, we specifically wish to test $H_0 : \Lambda_k = \Lambda_{k+1}$. Similar to permutation tests, a bootstrap hypothesis test is based on a test statistic. Here, we introduce essential notations for Algorithm 2. Let $q_1, \cdots, q_\ell$ be quantile values. Let $X = (x_{ij})$ for $1 \leq i \leq n$ and $1 \leq j \leq m$ be the taxa count matrix. For all $1 \leq k \leq \ell$, define $\tilde{\mathcal{I}}'_k$ to be the columns of $\tilde{\mathcal{I}}$ after removing taxa with degree less than q_k. Define X_k to be the subset of the columns of X corresponding to the columns of $\tilde{\mathcal{I}}'_k$. Let Σ_k be the covariance matrix of X_k and m_k be the number of taxa in X_k. Consider the test statistic

$$t_k = \frac{\Lambda_{k+1} - \Lambda_k}{\sqrt{\|\Sigma_{k+1}\|^2/m_{k+1} + \|\Sigma_k\|^2/m_k}}. \tag{10}$$

We next describe our bootstrap process and bootstrap test statistic t^*. For each k and $b = 1, \ldots, B$, sample $m_k + m_{k+1}$ columns with replacement from (X_k, X_{k+1}) and name this matrix $\boldsymbol{N}^*$. In addition, we denote the first m_k columns of $\boldsymbol{N}^*$, $\boldsymbol{Z}^*$ and the remaining m_{k+1} columns $\boldsymbol{Y}^*$. Let $\Sigma^*(\boldsymbol{Z}^*)$ $(\tilde{\mathcal{I}}'(\boldsymbol{Z}^*))$ and $\Sigma^*(\boldsymbol{Y}^*)$ $(\tilde{\mathcal{I}}'(\boldsymbol{Y}^*))$ be the covariance (adjusted mutual information) matrices of $\boldsymbol{Z}^*$ and $\boldsymbol{Y}^*$, respectively. Let $\tilde{\mathcal{I}}'(\boldsymbol{N}^*)$ be the adjusted mutual information matrix of $\boldsymbol{N}^*$ and define,

$$\Lambda^*(\boldsymbol{Z}^*) = 1 - \frac{\|\tilde{\mathcal{I}}'(\boldsymbol{Z}^*)\|_F^2}{\|\tilde{\mathcal{I}}(\boldsymbol{N}^*)\|_F^2} \quad \text{and} \quad \Lambda^*(\boldsymbol{Y}^*) = 1 - \frac{\|\tilde{\mathcal{I}}'(\boldsymbol{Y}^*)\|_F^2}{\|\tilde{\mathcal{I}}(\boldsymbol{N}^*)\|_F^2}. \tag{11}$$

Lastly, define our bootstrap test statistic to be

$$t_{kb}^* = \frac{\Lambda^*(\boldsymbol{Y}^*) - \Lambda^*(\boldsymbol{Z}^*) - (\Lambda_{k+1} - \Lambda_k)}{\sqrt{\|\Sigma^*(\boldsymbol{Y}^*)\|^2/m_{k+1} + \|\Sigma^*(\boldsymbol{Z}^*)\|^2/m_k}}. \tag{12}$$

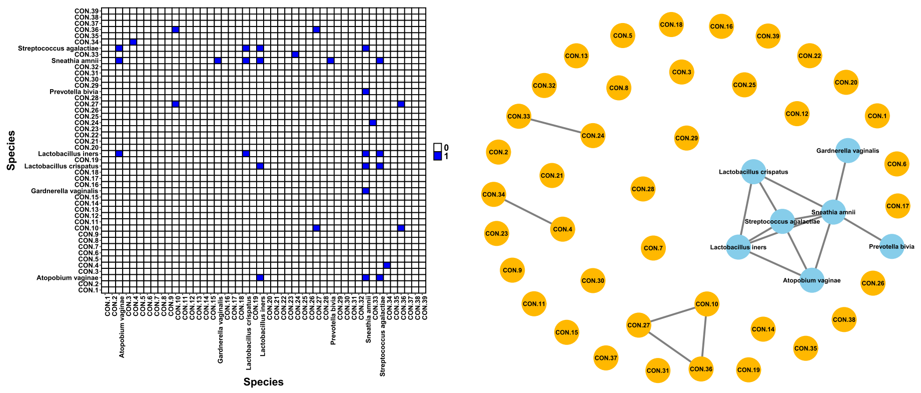
```
Algorithm 2: ALGORITHM 2
  Input: significance level $\alpha$, the number of bootstrap samples $B$, quantile vector
    $\boldsymbol{q} = (q_1, \cdots, q_\ell)$
  For each k:
    (1) Calculate $t_k$ (Eq. 10).
    (2) for each $b = 1, \cdots, B$:
        (a) Generate $\tilde{\mathfrak{I}}(\boldsymbol{N}^*)$, $\tilde{\mathfrak{I}}'(\boldsymbol{Z}^*)$, $\tilde{\mathfrak{I}}'(\boldsymbol{Y}^*)$, and Calculate $t_{kb}^*$ (Eq. 12).
    (3) Compute $Pval_k = \frac{\# t_{kb}^* \geq t_k}{B}$,
  $\boldsymbol{Pval}$ =$(Pval_1, \cdots, Pval_\ell)$.
  Find the index of the first entry of $\boldsymbol{Pval}$ that is less than or equal to $\alpha$. Call this index
    $ind$.
  Remove all taxa with total abundance less than $q_{ind-1}$.
```

Evaluating the filtering method

Mock microbial community

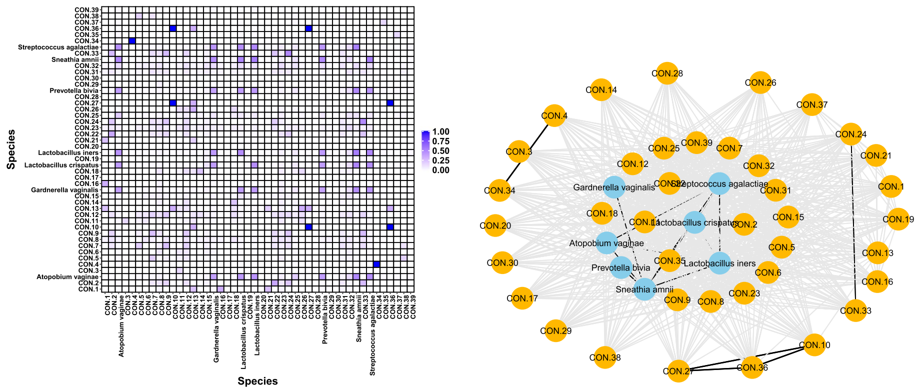
To test our method, we used a publicly available mock community data set given in Brooks et al. [7] where the ground-truth was known. These data consist of prescribed proportions of cells from seven vaginally-relevant bacterial strains: *Atopobium vaginae, Gardnerella vaginalis, Lactobacillus crispatus, Lactobacillus iners, Prevotella bivia, Sneathia amnii*, and *Streptococcus agalactiae* to quantify and characterize bias introduced in the sample processing pipeline such as DNA extraction, PCR amplification, and sequencing classification. The data consist of 240 sequenced samples; the resulting sequencing and ASV identification pipeline produced a table with 46 ASVs. Therefore, there were 39 false and 7 true ASVs produced in the upstream sequencing and analyses of the data. Of the approximately 3.67M total reads in the data set, 99.9% were attributed to the 7 true bacterial species. The most frequent of the contaminant species (*Pseudomonas gessardii*) was only present at a frequency of 6.81×10^{-5}.

We start by constructing an unweighted network of vaginal microbiome data. Table 1 reports the results for varying the threshold parameter τ for the mock community data. It can be seen that the coefficient of determination $R^2 = 0.97$ clearly favors $\tau = 0.45$. Based on these results, we use $\tau = 0.45$ to construct the unweighted network. Because of the large drop in R^2 after $\tau = 0.45$, we investigated the removal of individual edges with mutual information scores between 0.45 and 0.5. It seems the large drop was at least partially due to removing the edge between *Atopobium vaginae* and *Streptococcus agalactiae* (mutual information = 0.469, $R^2 = 0.38$); in



(a) (b)

Fig. 1 Schematic diagram of an unweighted microbiome network based on adjusted mutual information. **a** Adjacency matrix with $\tau = 0.45$ threshold; **b** the microbiome network diagram was formed according to the relationship among 46 taxa. We indicate contaminant taxa as CON.(arbitrary number) for convenience in illustrative purpose



(a) (b)

Fig. 2 Schematic diagram of a weighted microbiome network based on adjusted mutual information. **a** Adjacency matrix; **b** the microbiome network diagram was formed according to the relationship among 46 taxa. We indicate contaminant taxa as CON.(arbitrary number) for convenience in illustrative purpose

other words, removing this particular network edge significantly altered the topology such that it no longer fit a scale-free distribution nearly as well.

In Fig. 1 we have established the adjusted mutual information unweighted network of this dataset. It can be seen that $\mathcal{I}^*$ can reflect the true connection between the microbiome as a subnetwork and the majority of noise taxa are indicated as isolated nodes. In addition, we can define the weighted network where weights are adjusted mutual information ($\widetilde{\mathcal{I}}$), this is shown in Fig. 2. It can be seen that in these types of networks all the nodes are connected to all other nodes. Here, edges are colored based on the strength of the connectivity between adjacent nodes from very weak (light grey), moderate (grey), strong (black). Notice that the weight between majority of true taxa is strong, however we can see three subnetworks of noise that strongly share information.

Receiver operator characteristic (ROC)

Here we use an ROC curve to evaluate the classification accuracy of each taxon in this data set using a thresholding parameter τ in reference to the binary outcome D, which takes 0 (noise taxon) or 1 (true taxon). In order to do this, we measure the degree d_j of each node (taxon) in our unweighted network obtained by different hard thresholding parameter τ. For each taxon, convention dictates that a true taxon is defined as $d_j \geq \tau$. The classification accuracy of each taxon is then evaluated by considering a confusion matrix. It cross-classifies the predicted outcome for taxon with $d_j \geq \tau$ versus the true outcome D. For the fixed cutoff τ, the true positive fraction is the probability of identifying a taxon as a true signal, when it is truly a taxon.

In general, ROC analysis assesses the trade-offs between the test's fraction of true positives versus the false positives as τ varies over the range of 0 to 1.

$$\text{TPF}(\tau) = P(d_j \geq \tau | D = 1),$$

and the false positive fraction is the probability of identifying a taxon as a true signal, while it is a noise taxon.

$$\text{FPF}(\tau) = P(d_j \geq \tau | D = 0).$$

Because τ is not fixed in advance, one can plot TPF (sensitivity) y-axis against FPF (1-specificity) x-axis for all possible values of τ. If $\text{TPF}(\tau) = \text{FPF}(\tau)$, for all τ, it is a useless test with regards to the binary prediction. A perfect test that is completely informative about the signal status has $\text{TPF}(\tau) = 1$ and $\text{FPF}(\tau) = 0$ for at least one value τ. In other words, an excellent model has an area under the ROC curve (AUC) near 1 which indicates a good measure of separability. A model with area near 0 indicates a good measure of separability but a poor classification accuracy. An area under the ROC curve of 0.5 means the model has no class separation ability and is considered to be a random classifier. Figure 3 shows the ROC curve, assessing true versus false positive rate with $\text{AUC} = 0.86$ demonstrates the good performance of the method. We point out that using an ROC analysis to assess a diagnostic method is only possible when the truth is already known (as is the case with the mock data) and would not be part of analyzing normal microbiome data. The ROC also corroborates that $\tau = 0.45$ was the best threshold value, as determined by the R^2 analysis.

Information loss

We use the Brooks et al. [7] data set to estimate the amount of information loss by removing different percentages of taxa and investigate if the difference in information loss by removing percentages of taxa with degree less than q_k versus q_{k+1} is significant. Figure 4 illustrates the results for this mock community data set. The left panel displays the information loss and the right panel displays the difference in information loss. The data set was sorted according to the increasing connectivity degree of taxa using the adjusted mutual information adjacency matrix. For example, a cutoff assignment of 1% removes 1% of the taxa with the lowest connectivity degree. It is clear that applying percentile based filtering changes the amount of information loss, Fig. 4. For example, information loss has a drastic increase from 0.86 to 0.91 filtering threshold, while there is no or minimal change in information loss between removing 81% and 86% of taxa. This provides us with the intuition that 86% of taxa can be removed from the further analysis without loosing significant amount of information and these taxa could be the result of sequencing or PCR error, especially in high-throughput sequencing data sets.

Figure 5 shows information loss versus the number of taxa that are removed based on the lowest connectivity degree one at a time. We can see that after removing true signals (indicated by taxonomic name) the information loss values increase dramatically. From the Figs. 4 and 5, it is clear that information loss increases after removing a certain number or percentage of taxa. However, we need to investigate whether this rise of information loss is due to random errors or a real effect. In other word, we want to determine whether the information loss after removing less than q_k of taxa is significantly different from information loss after removing q_{k+1} of taxa, where q_k is the *k-th* quantile value of connectivity degree. To do this, we use permutation and bootstrapping approaches to test the null hypothesis which indicates that removing taxa with degree less than q_{k+1} versus q_k does not make any difference in information loss and hence we can remove them from further analysis. We follow the Algorithms 1 and 2 by setting $M = 500$, $B = 500$, $\alpha = 0.05$, and $\mathbf{q} = (0.01, \cdots, 0.96)$. The results of our permutation and bootstrapping tests are shown in Table 2; the second column shows that there is a significant loss of information after removing $\geq$ 91% of taxa (p-value$<$ 0.05) at 5% significance level. As an alternative, we follow Algorithm 2 to apply bootstrap method for hypothesis testing to approximate p-value. In Table 2 the third column shows that the bootstrap test gives similar results to permutation tests which also indicates that there is strong evidence that removing 91% of taxa will result in loosing significant amount of information.

Comparison study on mock community data from Brooks, et al.

Here, we use the data set in [7] to assess the performance of our method and compare results to alternative methods. More specifically, we consider four traditional methods which have been commonly employed for filtering of microbiome data: (1) Traditional 1: we retain taxa with more than 0.1%, 5%, and 1% relative abundance in at least one sample [11, 15, 25, 29]. (2) Traditional 2: we retain taxa with at least 5 reads in at least 3 samples [19]. (3) Traditional 3: we retain taxa presented in more than 5 samples [6]. (4) Traditional 4: we remove samples with fewer than 100 reads and taxa with fewer than 10 reads, as well as taxa which present in fewer than 1% of samples [12].

Results presented in Table 3 indicated that MI-based filtering method performs better than Traditional 2, 3, and 4 as well as Traditional 1 for the choice of 0.1%. In particular, we can see that our MI-based method removed 84.8% of taxa with minimum loss of information and preserved all taxa which were true signals. Traditional 1 filtering method with a retention threshold of 1% and 5% performed as well as the MI-based method, however, when the retention threshold was 0.1% this method retained 2.6% of contamination. For the Traditional 2 filtering method, following filtering 78.3% of taxa, this method retained 7.7% noise signals. Similarly for the Traditional 3 and 4 filtering methods, they filtered 58.7% and 60.9% of taxa, respectively; these two methods were the most permissive and retained 30.8% and 28.2% of noise, respectively. In comparison to the R `PERFect` package [36], the default settings using the "simultaneous" and "permutation" algorithms filtered 78.2% and 82.6% of the taxa and retained 7.7% and 2.6% of the noise, respectively (i.e., slightly worse than our proposed method). It should be noted that these `PERFect`

Table 1 Microbiome network characteristics for different hard thresholds τ

τ	R^2	$\bar{k}$	$-\gamma$
0.05	0.71	5.96	− 0.85
0.10	0.86	3.65	− 1.00
0.15	0.75	2.30	− 0.56
0.20	0.75	1.87	− 0.64
0.25	0.84	1.61	− 0.60
0.30	0.84	1.39	− 0.61
0.35	0.70	1.17	− 2.68
0.40	0.74	1.04	− 0.69
0.45	**0.97**	**0.70**	**− 0.97**
0.50	0.56	0.43	− 4.53
0.55	0.56	0.22	− 3.90
0.60	0.57	0.17	− 3.96
0.65	0.56	0.17	− 3.96
0.70	0.56	0.17	− 3.96
0.75	0.56	0.17	− 3.96
0.80	0.56	0.17	− 3.96
0.85	0.56	0.17	− 3.96
0.90	0.56	0.17	− 3.96
0.95	0.57	0.17	− 3.96

The second column contains coefficient of variation R^2 that varies between 0 and 1, where 0 indicates that the power-law model explains none of the variability of the empirical degrees, while 1 indicates the model perfectly fit the data. $\bar{k}$ and $-\gamma$ are the average node degree and the exponent of the power-law distribution, respectively. Bold values indicate the chosen level of thresholding (highest R^2)

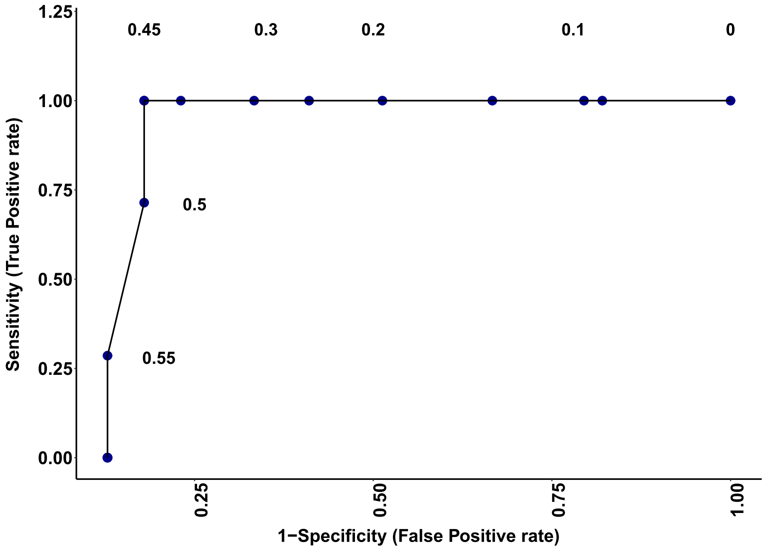
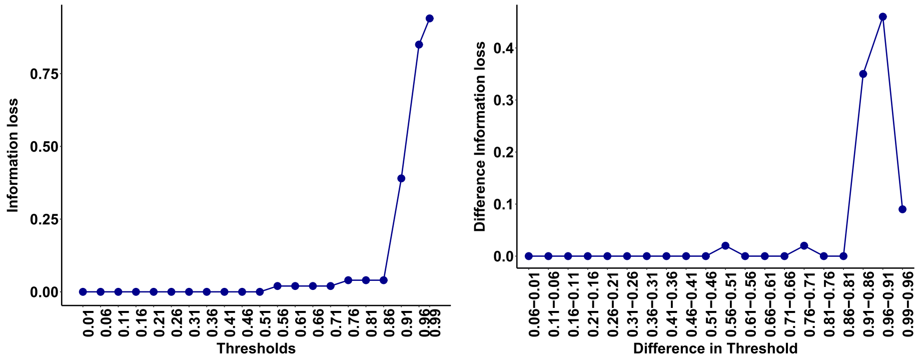


Fig. 3 ROC curve describing true positive versus false positive rate of the unweighted adjusted mutual information network model predicting true taxa for the Brooks et al. [7] data

results are depedent on parameter choice (see Table 1 of [36] for a full listing of their results under different parameter choices for the Brooks et al. [7] data).

Our proposed method was therefore successfully able to identify and remove contaminants from these data which results in dimensionality reduction of the data that



(a) (b)

Fig. 4 Information loss for the Brooks et al. [7] data. **a** Information loss as a function of threshold. Taxa are sorted according to the increasing connectivity degree of taxa and are removed based on different percentiles. **b** Difference in information loss that evaluates the slope at each taxon

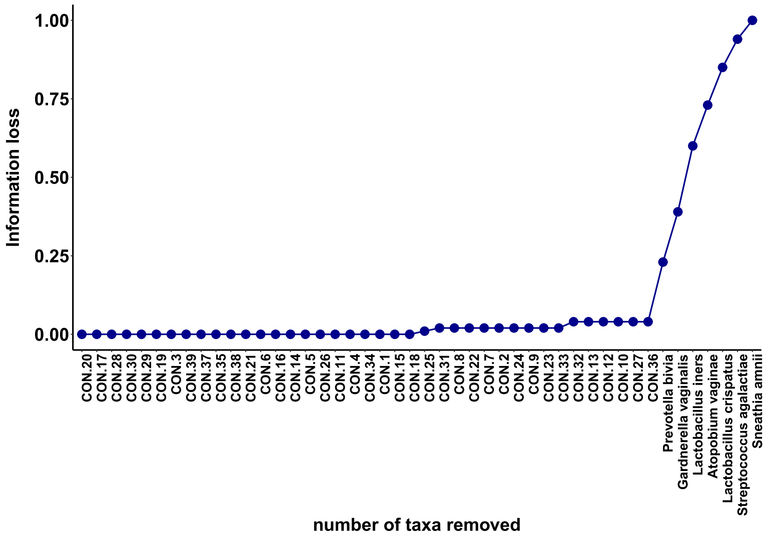


Fig. 5 Information loss for the Brooks et al. [7] data. Information loss as a function of number of taxa being removed. Taxa are sorted according to the increasing connectivity degree of taxa and are removed one at a time. We indicate contaminant taxa as CON.# for convenience in illustrative purpose

can reduce computation time and improve interpretation of downstream analyses. The proposed method has two advantages in addition to its superior performance in comparison with above mentioned traditional methods. First, it does not required a choice of threshold which is critical and not easy to obtain. Second, it is able to detect true taxa with low abundance. Most of the traditional methods have subjective predetermined thresholding value that might have adverse effects on the analysis due to loss of important information within filtered taxa. Our proposed method chooses a filtering threshold based on hypothesis testing and information loss. As mentioned earlier, these traditional methods remove taxa with low abundance and hence any important taxon with low abundance is removed leading to significant loss of information. However, MI-based filtering method, removes taxa based on their interactions with other taxa and therefore it can preserve low abundance taxa in case of their

Table 2 *P*-values by permutation and bootstrapping test based on 500 randomizations each

Percentage of taxa removal	*p*-value by permutation	*p*-value by bootstrapping
0.01–0.06	1.000	0.508
0.06–0.11	1.000	0.514
0.11–0.16	1.000	0.503
0.16–0.21	1.000	0.491
0.21–0.26	1.000	0.511
0.26–0.31	1.000	0.575
0.31–0.36	1.000	0.644
0.36–0.41	1.000	0.525
0.41–0.46	0.999	0.601
0.46–0.51	1.000	0.663
0.51–0.56	1.000	0.507
0.56–0.61	0.996	0.585
0.61–0.66	0.994	0.666
0.66–0.71	0.997	0.534
0.71–0.76	0.445	0.722
0.76–0.81	0.994	0.753
0.81–0.86	0.314	0.813
0.86–0.91	$<$ **0.002**	$<$ **0.002**
0.91–0.96	$<$ **0.002**	$<$ **0.002**

Bold values indicate levels of filtering that significantly altered the MI network

Table 3 Comparison of 6 commonly used traditional filtering method with MI-based filtering method for Mock community data set in [7]

Filtering method	# Taxa preserved	Filtered%	Preserved contamination%
Traditional 1-0.1%	8	82.6	2.6
Traditional 1-1%	7	84.8	0.0
Traditional 1-5%	7	84.8	0.0
Traditional 2	10	78.3	7.7
Traditional 3	19	58.7	30.8
Traditional 4	18	60.9	28.2
MI-based	7	84.8	0.00

strong association with other taxa. This allows us to study significant taxa that occur in low abundance. Our method also provides another advantage in comparison to methods such a `microDecon` [26] and `decontam` [8]: it does not require negative control samples to calibrate the algorithm. Therefore our methods reduce the cost of sequencing in comparison.

Obviously, the Brooks et al. [7] data are overly simplisitic. Because of the fact that this mock community has only 7 true species and 39 false species (contaminants/sequencing noise), all of the methods performed relatively well except for the Traditional 3 and 4 filtering methods. (Table 3). Unfortunately, data for large complex communities—like those that the method will typically be applied to—where the true composition is known are not available. However, our MI filtering based method has been applied to the gut microbiome of dairy calves [35]. In this analysis, the raw ASV table had 431 putative species. After filtering using the mutual information criterion, 76% of the ASVs were determined to be contamination and the resulting refined data only had 57 ASVs, thus making downstream analyses much easier and more interpretable (see Slanzon et al. [35] for further details). Of course, we do not know how accurate that filtering process was because the true composition of these data is unknown, but it does demonstrate the utility of our methods.

Conclusion

Removing contaminants prior to any downstream analysis is an essential step in metagenomic sequencing data research. Host associated contaminants significantly complicate analysis, particularly in low microbial biomass body sites. Contamination can cause analysis of sequencing reads to result in false positive or false negative and hence decreasing the reliability of the analysis. Here, we developed a simple method that uses a combination of graph theory and information theoretic functionals to identify and remove contaminants in metagenomic data sets. Our results suggest that mutual information based filtering method can improve the accuracy of detecting contaminants, especially in comparison with the commonly used traditional filtering methods.

To fully explore the strengths and weaknesses of our proposed filtering method, evaluation on different labeled mock community data sets is necessary. Unfortunately, labeled data sets are expensive and difficult to obtain and thus hindered our ability to test our method further. We believe that it is possible to improve the threshold selection in the unweighted graph given improved (ground-truthed) data with which to work. Looking solely for isolated nodes is not sufficient as it is unable to filter out random interaction between contaminants nodes. When the number of nodes increases we can expect more random interaction between contaminants, making the isolated node approach even less powerful. We believe more advanced methods from graph theory could remedy this short coming. Future work could include but are not limited to examining the efficiency of techniques such as dense community detection, dense subgraph selection, and vertex selection based on vertex centrality. These sophisticated node selection methods could provide a more powerful filtering method in the unweighted graph.

Acknowledgements
Not applicable.

Author Contributions
EBM and BJR contributed to all phases of this research and publication. Both authors read and approved the final manuscript.

Funding
Research reported in this publication was supported by the National Institutes of Health under Award Numbers R24OD020347 and 1R01AG065546 and Centers of Biomedical Research Excellence (COBRE) under Award Number P20GM104420 . The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials
Data sharing is not applicable to this article as no data sets were generated during the current study. The analyzed data set may be found in [7].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests

The authors report no competing interest.

Received: 24 May 2022 Accepted: 6 September 2022
Published online: 16 September 2022

References

1. Albert R. Scale-free networks in cell biology. J Cell Sci. 2005;118(21):4947–57.
2. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
3. Barabási A-L, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–12.
4. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13.
5. Barton HA, Taylor NM, Lubbers BR, Pemberton AC. DNA extraction from low-biomass carbonate rock: an improved method with reduced contamination and the low-biomass contaminant database. J Microbiol Methods. 2006;66(1):21–31.
6. Brigham A, Sadorf EGS (U.S.), Benthic invertebrate assemblages and their relation to physical and chemical characteristics of streams in the Eastern Iowa basins, 1996-98. Water-resources investigations report, U.S. Department of the Interior, U.S. Geological Survey. 2001.
7. Brooks JP, Edwards DJ, Harwich MD Jr, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Consortium VM, Strauss JF, Jefferson KK, Buck GA. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. 2015;15:66.
8. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6(1):226.
9. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, Peacock SJ, Smith GCS, Parkhill J. Recognizing the reagent microbiome. Nat Microbiol. 2018;3(8):851–3.
10. Dionisio A, Menezes R, Mendes DA. Mutual information: a measure of dependency for nonlinear time series. Phys A Stat Mech Appl. 2004;344(1):326–9 (**applications of Physics in Financial Analysis 4 (APFA4)**).
11. Dobbler P, Mai V, Procianoy RS, Silveira RC, Corso AL, Roesch LFW. The vaginal microbial communities of healthy expectant Brazilian mothers and its correlation with the newborn's gut colonization. World J Microbiol Biotechnol. 2019;35(10):159.
12. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8(1):1784.
13. Efron B, Tibshirani R. An introduction to the bootstrap. Boca Raton: CRC Press; 1994.
14. François D, Wertz V, Verleysen M. The permutation test for feature selection by mutual information. In: ESANN 2006, European Symposium on Artificial Neural Networks, pp. 239–244, 2006.
15. Gliniewicz K, Schneider GM, Ridenhour BJ, Williams CJ, Song Y, Farage MA, Miller K, Forney LJ. Comparison of the vaginal microbiomes of premenopausal and postmenopausal women. Front Microbiol. 2019;10:193.
16. Good P. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer series in statistics. New York, NY: Springer; 1994. https://doi.org/10.1007/978-1-4757-2346-5.
17. Hidaka S, Oizumi M. Fast and exact search for the partition with minimal information loss. PLoS One. 2018;13(9):1–14.
18. Hornung BVH, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. FEMS Microbiol Ecol. 2019;95(5).
19. Ingham AC, Kielsen K, Cilieborg MS, Lund O, Holmes S, Aarestrup FM, Müller KG, Pamp SJ. Specific gut microbiome members are associated with distinct immune markers in pediatric allogeneic hematopoietic stem cell transplantation. Microbiome. 2019;7(1):131.
20. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007;45(9):2761–4.
21. Jervis-Bardy J, Leong LEX, Marri S, Smith RJ, Choo JM, Smith-Vaughan HC, Nosworthy E, Morris PS, O'Leary S, Rogers GB, Marsh RL. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. Microbiome. 2015;3(1):19.
22. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proc Natl Acad Sci. 2014;111(9):3354–9.
23. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
24. Li L, Wang Z, He P, Ma S, Du J, Jiang R. Construction and analysis of functional networks in the gut microbiome of type 2 diabetes patients. Genomics Proteomics Bioinform. 2016;14(5):314–24.
25. ...Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R, Claverie J-M, Decelle J, Dolan J, Dunthorn M, Edvardsen B, Gobet A, Kooistra W, Mahé F, Not F, Ogata H, Pawlowski J, Pernice M, Romac S, Shalchian-Tabrizi K, Simon N, Stoeck T, Santini S, Siano R, Wincker P, Zingone A, Richards T, de Vargas C, Massana R. Patterns of rare and abundant marine microbial eukaryotes. Curr Biol. 2014;24(8):813–21.
26. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. microDecon: a highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. Environ DNA. 2019;1(1):14–25.
27. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, Hyde J, Doty MM, Stillwell K, Benardini J, Kim JH, Allen EE, Venkateswaran K, Knight R. KatharoSeq enables high—throughput microbiome analysis from low-biomass samples. mSystems. 2018;3(3):e00218–e0017.

28. Naqvi A, Rangwala H, Keshavarzian A, Gillevet P. Network-based modeling of the human gut microbiome. Chem Biodiv. 2010;7(5):1040–50.
29. Partula V, Mondot S, Torres MJ, Kesse-Guyot E, Deschasaux M, Assmann K, Latino-Martel P, Buscail C, Julia C, Galan P, Hercberg S, Rouilly V, Thomas S, Quintana-Murci L, Albert ML, Duffy D, Lantz O, Touvier M, Consortium tMI. Associations between usual diet and gut microbiota composition: results from the Milieu Intérieur cross-sectional study. Am J Clin Nutr. 2019;109(5):1472–83.
30. Patel JB. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. Mol Diagn. 2001;6(4):313–21.
31. Pollock J, Glendinning L, Wisedchanwet T, Watson M. The madness of microbiome: attempting to find consensus best practice for 16S microbiome studies. Appl Environ Microbiol. 2018;84(7):e02627.
32. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
33. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12(1):87.
34. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423.
35. Slanzon GS, Ridenhour BJ, Moore DA, Sischo WM, Parrish LM, Trombetta SC, McConnel CS. Fecal microbiome profiles of neonatal dairy calves with varying severities of gastrointestinal disease. PLoS One. 2022;17(1): e0262317.
36. Smirnova E, Huzurbazar S, Jafari F. PERFect: PERmutation filtering test for microbiome data. Biostatistics. 2018;20(4):615–31.
37. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinform. 2012;13(1):328.
38. Sulaimanov N, Koeppl H. 2016: graph reconstruction using covariance-based methods. EURASIP J Bioinform Syst Biol. 2016;1:19.
39. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008;321(5891):956–60.
40. Sun C, Yang F, Wang C, Wang Z, Zhang Y, Ming D, Du J. Mutual information-based brain network analysis in post-stroke patients with different levels of depression. Front Human Neurosci. 2018;12:285.
41. Tavakoli S, Yooseph S. Learning a mixture of microbial networks using minorization-maximization. Bioinformatics. 2019;35(14):i23–30.
42. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol. 2014;15(12):564.
43. Xia Y, Sun J, Chen D. Statistical analysis of microbiome data with R. ICSA book series in statistics. Singapore: Springer; 2018.
44. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005. https://doi.org/10.2202/1544-6115.1128.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.