

The effects of pre-processing techniques on network analysis of microbiome sequencing data

– Bachelor Thesis –

to be awarded

Bachelor of Science in Artificial Intelligence

submitted by

Colino Sprockel

February 5, 2021

Colino Sprockel
c.j.sprockel@students.uu.nl
Student ID: 3979822

Supervisor

1st: dr. H. W. Uh

2nd: dr. M. van Ommen

Abstract

The human microbiome is a growing area of research. Enabled by advances in sequencing techniques, vast amounts of microbiome data are being generated. Using this data to answer research questions is challenging due to the compositionality, sparsity and high-dimensionality of the data. Network techniques have been successfully applied to make sense of microbiome data, yet difficulties still remain. Here, we compare different methods of preparing data for the use in weighted gene co-expression network analysis (WGCNA), a popular framework within bioinformatics utilizing network theory. Three different methods were applied: one based on simple compositionality, one based on the centered log-ratio transform, and one using SparCC: an advanced method specifically designed to infer correlations in microbiome sequencing data. We found that network results vary widely depending on the method used.

Contents

List of Figures	iii
1 Introduction	1
2 Microbiome and its statistical challenges	2
2.1 The human microbiome	2
2.2 Statistical challenges	3
2.2.1 Compositionality	3
2.2.2 Sparsity	4
2.2.3 High-dimensionality	5
2.2.4 Diversity	5
3 Methods for statistical analysis of microbiome	6
3.1 pre-processing	6
3.1.1 compositional	6
3.1.2 Centered log ratio transform + 1	6
3.2 SparCC	7
3.3 network notations	9
3.4 Weighted Gene Co-Expression Network Analysis	10
3.4.1 Scale-free networks	11
3.4.2 Topological Overlap Matrix	11
3.4.3 Hierarchical clustering	12
4 Application to Immuno deficiency microbiome dataset	12
4.1 introduction	12
4.2 Network construction	14
4.3 Consensus Analysis	15
5 Conclusion and discussion	17
References	20
A Appendix	22

List of Figures

1	Correlation heatmaps that result from the three methods	14
2	Dendrograms and clustering of all three methods	15
3	Table showing the modules generated by the different approaches and the names of the bacteria in those modules. Their respective networks are shown at the bottom, the circles indicating bacteria and their size is the degree of the bacteria in the network. Edges over 0.05 are shown as lines between nodes.	16
4	Consensus tables comparing all three methods	17
5	Relative abundances of the bacteria, colors show the cluster the bacteria belongs to. Clusters from Compositional approach	22
6	Relative abundances of the bacteria, colors show the cluster the bacteria belongs to. Clusters from CLR1 approach	22
7	Relative abundances of the bacteria, colors show the cluster the bacteria belongs to. Clusters from SparCC approach	23
8	Network of brown module from the SparCC approach	23
9	Network of turquoise module from the SparCC approach	24
10	Network of yellow module from the SparCC approach	24
11	Network of blue module from the CLR1 approach	25
12	Network of brown module from the CLR1 approach	25
13	Network of yellow module from the CLR1 approach	26
14	Network of yellow module from the Compositional approach	26
15	Network of brown module from the Compositional approach	27
16	Network of pink module from the Compositional approach	27

1 Introduction

In and on the human body countless bacteria reside, this collection of bacteria is commonly known as the microbiome. Microbiomes are complex microbial communities that are heavily influenced by microbe-microbe, and host-microbe interactions. The decreasing costs of sequencing techniques expand our possibilities to sequence microbiome samples and discover more about the relation between personal health and our microbiome. Vast volumes of microbiome data are being generated yet disentangling these datasets and generating valuable insight remains challenging.

Many methods have been developed for microbiome data, often based on univariate testing of single bacteria [1]. Given the complex nature of microbial communities, univariate methods might not be able to fully grasp the underlying relations. Several novel approaches based on AI have been recently proposed, such as network techniques [2]. Network theory enables us to examine complex systems using a holistic approach: one can model and analyse a microbiome and all its complex interactions in a single network [3]. A widely used framework using network techniques on gene sequencing data is Weighted Gene Co-expression Network Analysis (WGCNA) [4] [5].

The multivariate nature of WGCNA enables us to look at more complex relationships between bacteria, like finding groups (clusters) of bacteria that strongly impact each other. The starting point of many network techniques like WGCNA is a correlation matrix containing pairwise correlations between all possible pairs of bacteria.

The compositionality, sparsity, and high-dimensionality of microbiome data pose a series of unique challenges to construct the correlation matrices [2]. Due to the novelty of applying network techniques to analyse microbiome data, best practices regarding how to handle these complex issues have yet to be solidified, resulting in a multitude of different approaches. Due to these challenging properties, results using network techniques vary greatly depending on the pre-processing approaches used to generate the correlation matrix. Which is a problem given researchers' aims to achieve reliable and reproducible results.

Unfortunately there currently is not one set approach to prepare data for network analysis, researchers use a wide array of different methods, resulting in a fragmented research field.

Given the data's unique properties, advanced pre-processing methods may be warranted¹. One of these methods is SparCC, which has been specifically designed to deal with microbiome sequencing data [6]. SparCC generates correlation matrices using complex techniques more commonly found in machine learning, a subfield of AI.

The goal of this paper is to show that the varying methods of constructing correlation matrices that are actively used by researchers, lead to different network results. I will argue for the wider adoption of SparCC and other complex approaches that appropriately deal with the unique challenges of microbiome sequencing data.

2 Microbiome and its statistical challenges

2.1 The human microbiome

An increasing number of studies show correlations between microbiome composition and health outcomes, ranging from inflammatory bowel disease [7] [8] and cancer [9], to autism [10] and major depressive disorder [11]. The microbiome is emerging as a new frontier of human healthcare. Altering the microbiome of patients as an effective form of treatment or managing people's microbiome as a measure to prevent future illness might become feasible in the future. To enable this we must first have a solid understanding of the microbiome and the factors that influence this complex system.

It is currently estimated that the average human microbiome consists of between 500-1000 different species of bacteria. These bacterial species combined contain around 100 times the estimated number of human genes (2,000,000 vs 20,000 genes) [12]. This quantitative difference helps shed some light on the difficulties in trying to extract meaningful information from microbiome sequencing data: the microbiome contains magnitudes more variables, requiring vastly larger datasets than the human genome project, or vastly superior statistical methods.

¹When mentioning pre-processing I generally mean all data handling starting from the raw count data up until and including construction of the correlation matrix.

2.2 Statistical challenges

Microbiome data science is facing serious challenges caused by various statistical properties of microbiome data, such as compositionality, sparsity, and high-dimensionality. These properties can introduce various biases when not appropriately taken into account. In this section, I will discuss four important properties and how they can introduce biases when they are not properly dealt with.

2.2.1 Compositionality

One of the special properties of microbiome data is its compositionality. This property, not often seen in other fields of data science is the result of a necessary standard manipulation to deal with microbiome sequencing data (MSD). MSD is commonly generated using high-throughput sequencing techniques. These techniques match genetic material in a sample to known gene sequences, thus identifying the species present in the sample.

Results are often presented using counts: a relatively abundant bacteria will have high counts and rare bacteria will have low counts. The dataset will be presented as a matrix $M = m_{ij}$, with each row i being a sample from a volunteer, and each column j a species of bacteria. Elements m_{ij} contain the sequence count of bacteria j found in sample i . The read count of sample j is the sum of all elements of that sample:

$$read_count(j) = \sum_k m_{kj} \quad (1)$$

Due to inadequacies of the sequencing process, the read count varies widely between samples. Because of this variation, the generated data only contains meaningful information concerning the relative abundance of bacteria (how the bacteria relate to one another within a given sample). Changing the data into compositional data ensures that the meaningless absolute counts can't impact statistical analysis downstream. Making data compositional is done by dividing the vector containing the bacteria counts of a sample by the sum of the vector. This way the resulting vector sums to 1, with each element representing the relative abundance of a bacteria in this sample.

More precisely, define a count matrix S , with elements s_{ij} being the count of the i^{th} sample and the j^{th} bacteria. Then c_{ij} is the relative abundance of bacteria j in the i^{th} sample.

$$c_{ij} = \frac{s_{ij}}{\sum_{k=1}^n s_{ik}} \quad (2)$$

Though necessary, the now compositional structure introduces a new statistical challenge. Because of its compositionality, all bacteria are now negatively correlated with each other, since the abundances of the bacteria sum up to one. If one bacteria happens to have a high relative abundance it will be negatively correlated with every other bacteria, since it monopolizes part of the available sum to 1, suppressing the values of the other bacteria.

Log-ratio transformations have been proposed as a solution to normalise compositional data: [13] [14]. First developed to deal with rock and soil samples, log-ratio transforms create pairwise abundance ratios that contain true knowledge about the relationship between these two variable [15]. This will be discussed in detail at section 3.1.2.

Unfortunately, these transforms are unable to fully deal with the spurious correlations caused by the compositionality of the data mentioned above. Furthermore, log-ratio transforms typically cannot be directly applied due to another property of microbiome data: sparsity.

2.2.2 Sparsity

Microbiome data is Sparse, which means that many elements in the dataset are zeros. This is problematic since part of our solution to the problem of compositionality is the log-ratio transform, yet we cannot take logarithms of zero. The cause of these zeros is also unknown: a bacteria might truly not be present in the sample or it was present and the sequencing device was not sensitive enough to sense it. Silverman *et al.* [16] show that the underlying source of zeros can impact results in a meaningful way.

The two obvious solutions are to remove zero elements from the data, or adding a pseudocount to all elements. Since the sparsity is so large, removing all zero elements would not leave much remaining. The more widely used strategy is adding

a pseudocount to all elements, this way no elements are deleted and logarithms can be used.

Pseudocounts can cause spurious correlations because of the compositionality. Since original zeros in a sample have the same count after the pseudocount, low abundant bacteria will become correlated simply because their counts have the same number across samples. The correlation between two bacteria effectively becomes an indication of the number of zero counts on these bacteria.

2.2.3 High-dimensionality

The high-dimensionality of microbiome data is a more common statistical problem that occurs when one has a large number of variables relative to the number of samples, also called the small N large P problem. Because of the high number of variables correlations between variables may be found by chance, without there being any real correlation between them. This problem is often solved by being very cautious when interpreting results from small N large P datasets, and requiring incredibly strong p-values. The number of variables are high in microbiome datasets since there are many different kinds of bacteria co-existing on and inside our bodies. Test subjects for microbiome studies rarely number more than a couple hundred people.

2.2.4 Diversity

The three properties mentioned above can be even more challenging depending on the amount of diversity within the ecosystem. With low diversity exacerbating the spurious correlations caused by the compositionality of the dataset. Unfortunately this is often the case in microbiome communities.

Friedman & Alm [6] show that when diversity is low, shuffling the dataset and computing Pearson correlation yield many correlations (where none should exist because the data is random). This effect lessens as species diversity increases, though some spurious correlations remain at high diversity. This indicates that species diversity influences the magnitude of the other statistical challenges of microbiome data.

3 Methods for statistical analysis of microbiome

Microbiome sequencing data will be used to create networks in three steps: the first is pre-processing, the second is creation of a correlation matrix, and the third is the construction of networks based on these matrices.

3.1 pre-processing

Pre-processing is necessary to transform the count data into relative abundances and deal with zeros

3.1.1 compositional

With a count matrix S , and element s_{ij} being the count of the j^{th} bacteria in the i^{th} sample. And c_{ij} being the relative abundance of the j^{th} bacteria in the i^{th} sample.

$$c_{ij} = \frac{s_{ij}}{\sum_{k=1}^n s_{ik}} \quad (3)$$

which guarantees that the relative abundances of sample i sum to 1.

$$\sum_{k=1}^n c_{ik} = 1 \quad (4)$$

3.1.2 Centered log ratio transform + 1

We start by adding 1 to all the counts to deal with the zeros in our count matrix S , creating a count + 1 matrix S^* .

$$s_{ij}^* = s_{ij} + 1 \quad (5)$$

S^* will be turned into the compositional matrix C^* using equation 3.

taking the log of each element

$$l_{ij} = \log(c_{ij}^*) \quad (6)$$

The mean of row i

$$\mu_i = \frac{\sum_{k=1}^n l_{ik}}{n} \quad (7)$$

Each element minus the mean of its row.

$$clr_{ij} = l_{ij} - \mu_i \quad (8)$$

3.2 SparCC

SparCC is a technique for inferring correlations from compositional data developed by Friedman & Alm [6]. SparCC requires two conditions to be met for it to be a valid method of inference: "(1) the number of different components is large, and (2) the true correlation network is 'sparse' (i.e., most components are not strongly correlated with each other)" [6].

The biggest benefit from using this approach lies in the lower rate of spurious correlations compared to conventional methods. Especially when the number of effective species is low (i.e., a few species dominate the ecosystem) the risk of finding correlations without any biological cause is very high. In microbial networks of low diversity, inferred connections are often dominated by negative correlations to the dominant species, which leads to positive correlations among the remaining species [6].

SparCC starts with a log-ratio transform, the advantages are that the ratios between bacteria are independent of other bacteria, solving part of the problem of compositionality.

$$y_{ij} = \log \frac{x_i}{x_j} = \log x_i - \log x_j \quad (9)$$

Where x_i is the compositional fraction of bacterium i .

Note that the element y_{ij} contains is a fraction, containing information regarding the relation of the two bacteria i and j . This means that applying equation 9 to a single sample/row results in a matrix of size $P \times P$ where P is the number of

bacteria. Applying this equation to a complete count matrix results in a 3D matrix of dimensions $P \times P \times N$ where N is the number of samples.

$$t_{ij} \equiv Var[y_{ij}] \quad (10)$$

is used to describe the variance across all samples. If two bacteria are perfectly correlated, their ratio remains constant and $t_{ij} = 0$. When two bacteria are not correlated, the ratio will vary widely and the corresponding t_{ij} will be large. t_{ij} is still hard to interpret because it lacks a scale. The following equation helps:

$$t_{ij} = \omega_i^2 + \omega_j^2 - 2p_{ij}\omega_i\omega_j \quad (11)$$

where ω_i^2 and ω_j^2 are the variances of the log-transformed basis abundances of bacterium i and j .

$$\omega_i^2 = \frac{(\sum x_i - \bar{x})^2}{n - 1} \quad (12)$$

and p_{ij} is the correlation between them [6]

$t_{ij} < \omega_i^2 + \omega_j^2$ indicates a positive correlation and $t_{ij} > \omega_i^2 + \omega_j^2$ a negative correlation. Solving equation 11 for all variables gives us the correlation matrix. Unfortunately equation 11 cannot be solved as there are more unknown variables than equations. Rewriting equation 11 gives

$$p_{ij} = \frac{\omega_i^2 + \omega_j^2 - t_{ij}}{2\omega_i\omega_j} \quad (13)$$

To solve equation 13, SparCC approximates the base variances ω_i^2 for each bacterium by assuming average correlations p_{ij} are small. Approximating the base variances is done through estimating the component fractions using a bayesian framework and a Dirichlet distribution. Since the estimates of the component fractions are taken from the Dirichlet distribution, the SparCC method is non-deterministic. Any further mathematical details can be found in the paper by Friedman & Alm [6] and are beyond the scope of this paper

Iterative SparCC

Iterative SparCC improves upon SparCC by excluding components (pairs of bacteria) under certain conditions to stop them from impacting other correlations. After removing a component, the algorithm is run again but with all remaining components. A pair of Components can be removed under two conditions:

1. The most strongly correlated pair is removed if the magnitude of this correlation exceeds a certain threshold.
2. If two components form an exclusive pair, meaning that all other combinations have been excluded through (1), the two components are removed from analysis.

In case (2), the removed components are completely removed from analysis since the components violate the sparsity assumption of the system. Meeting requirement (2) indicates that all other possible pairs of the two components i and j have been excluded through (1), meaning all of these possible pairs have a correlation that exceeds the given threshold.

3.3 network notations

A network is a system comprised of Nodes and Edges. Nodes are the hubs in the network and the Edges, also called Vertices are the connections between Nodes. Take a social network for example: the Nodes will be people and the Edges will be connections between people, representing friendships for instance.

Adjacency matrix:

Networks are often represented in an adjacency matrix format. Consider a network with n nodes. To construct an $n \times n$ matrix $A = [a_{ij}]$ an adjacency function is used, with the most commonly used being:

$$a_{ij} = \begin{cases} 1 & \text{if } s_{ij} \geq \tau \\ 0 & \text{if } s_{ij} < \tau \end{cases} \quad (14)$$

where s_{ij} is the pairwise correlation between elements i and j . Eq 14 is a "hard" adjacency function because connections between nodes at or above the threshold τ result in an edge and connections below τ are ignored. All adjacency functions need to map onto $[0, 1]$.

Networks can be discrete or continuous. discrete networks consist only of simple binary connections between nodes; there either is a connection or there isn't one. However, The networks we will use are weighted networks. This means that edges have a continuous value instead of a 0 or 1, often called a weight. The value represents the strength of the connection between two nodes.

The degree of a node i is the sum of its weights:

$$p(i) = \sum_{k=1}^n a_{ik} \quad (15)$$

meaning highly connected nodes have the highest degree. This definition is needed for when we discuss scale-free networks.

3.4 Weighted Gene Co-Expression Network Analysis

Weighted gene co-expression network analysis was originally created to deal with gene sequencing data and has been incredibly popular over the years. Though microbiome data is very different (bacteria instead of genes), it has been promoted by some researchers as a way of analysing microbiome data [2]

The WGCNA framework can be used to construct biologically relevant networks, it includes methods to reduce noise, cluster groups, and further methods of analysis. Clustering bacteria helps mitigate the high-dimensionality discussed at 2.2.3 The WGCNA framework proposes the power adjacency function:

$$a_{ij} = \text{power}(s_{ij}, \beta) \equiv |s_{ij}|^\beta \quad (16)$$

where $|s_{ij}|$ is the absolute value of s_{ij} , necessary because adjacency functions must map onto $[0, 1]$, and soft power parameter $\beta \leq 1$ The effect of this function is the shrinkage of the correlation matrix ie., correlations get smaller, with the most pronounced effect on the smallest correlations. The choice of β has far reaching consequences for the pairwise connection strengths. For example, increasing β will shrink more of the smaller correlations, decreasing the amount of noise in the network but also leaving only the largest pairs with a noticeable connection.

3.4.1 Scale-free networks

Many biological networks have been shown to approach a scale free topology. Therefore B. Zhang & Horvath [4] proposed that a good way of picking a β can be achieved by measuring how closely networks generated with various β fit a scale-free topology.

A scale-free network is defined as a network in which the degree distribution follows a power law. Which roughly means that the number of nodes with a certain degree monotonically decreases when the degree gets higher: $p(k) \sim k^{-\gamma}$

A network's scale-free fit can be calculated and plotted using the WGCNA framework. The protocol describes that the best method for picking a β is by generating networks using varying β , and plotting the resulting model fit. From the plot one can find the lowest β which achieves a R^2 which is sufficient.

An alternative approach, described by Bartzis *et al.* [17] calculates the minimally required β from the number of samples and variables:

$$\frac{P(P-1)}{2(\sqrt{N})^\beta} < 1 \quad (17)$$

where P is the number of variables and N the number of samples. Solving equation 17 gives the minimally required β .

3.4.2 Topological Overlap Matrix

B. Zhang & Horvath [4] propose using the topological overlap dissimilarity of nodes as a basis for identifying nodes that are tightly connected to each other. A Topological Overlap Matrix (TOM) $\Omega = [\omega_{ij}]$ shows how similar nodes are to each other, with higher values for ω_{ij} indicating higher similarity.

$$\omega_{ij} = \frac{g_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \quad (18)$$

Where k_i and k_j are the degrees of node i and node j . a_{ij} is the weight of the connection between node i and node j . Note that

$$g_{ij} = \sum_u a_{iu}a_{uj} \quad (19)$$

which measures the connections with a third node that nodes i and j have in common. $w_{ij} = 1$ if all the neighbors of the node with the lower degree are connected to the other node, and nodes i and j are connected to each other. $w_{ij} = 0$ if the pair of nodes are not connected to each other, and the nodes are not connected with each other through a direct neighbor.

The dissimilarity is easily calculated by

$$d_{ij}^w = 1 - \omega_{ij} \quad (20)$$

B. Zhang & Horvath [4] show that this method leads to more distinct clusters than the alternative measure: $d_{ij} = 1 - |p_{ij}|$ where $|p_{ij}|$ is the absolute pairwise correlation.

3.4.3 Hierarchical clustering

Once the dissimilarity matrix has been constructed, module detection is the next step. These groups of highly interconnected bacteria are found by using the hierarchical clustering algorithm (HCL). The built-in R function `hclust()` is used to perform this operation.

Hierarchical clustering is a form of unsupervised learning, meaning that no group labels are available: the algorithm only utilizes the data itself (the dissimilar TOM in this case) to find clusters. One drawback of this clustering method is that the number of clusters in a system is not objectively known, thus making it difficult to separate the bacteria in the correct number of clusters. Hierarchical clustering is presented as the default approach used to find modules in the WGCNA framework and was therefore used, any further discussion of this algorithm is beyond the scope of this paper.

4 Application to Immuno deficiency microbiome dataset

4.1 introduction

To compare the various methods of creating networks for microbiome analysis, a 16S rRNA sequencing dataset was used. This dataset contains oral swabs from 41 healthy controls, and 103 patients diagnosed with various forms of common variable

immunodeficiency disease (CVID)[18]. Of the 8 remaining samples we did not have patient data but I decided to include them since their unknown medical status has no impact on the results. Given the significant interactions between the immune system and microbiome[19], it is expected that the bacterial compositions vary widely between samples, enabling us to easily find networks of related bacteria.

We will show varying results depending on different pre-processing methods. And then compare these results, all code can be found on [Github](#).

data handling

The dataset consists of 152 samples and 170 bacteria classified at genus level, also called L6 (this is the level between family (L5) and species (L7)). This dataset was cleaned and as is standard practice in MB data analysis, any genus with a zero count in over 90% of the samples was removed. All samples were checked to have a minimum read count of over 8000, this was found to be true so no samples had to be removed from the dataset.

this cleaned dataset of 152 samples and 59 bacteria is used for all further analysis. Three different methods were used to create the correlation matrices necessary for network analysis, these will be referred to as compositional transform, clr1 transform and SparCC. The resulting correlation heatmaps are shown in figure 1. Note that the SparCC correlation matrix shown in the figure is less colorful than the other matrices, indicating weaker correlations. This is in line with our expectations: the other two methods remain prone to finding spurious correlations due to the properties of microbiome data, driving up correlations which result in brighter graphs. Spearman correlation was chosen since it is non-parametric.

compositional transform

Compositional transformation as described 3.1.1 is applied to the dataset. The result is used to construct a correlation matrix using Spearman's correlation. The correlation matrix has a median of 0.06 and an interquartile range (IQR) of 0.32

clr1 transform

Centered log ration transform as described 3.1.2 is applied to the dataset. The result is used to construct a correlation matrix using Spearman's correlation. The resulting correlation matrix has a median of 0.01 and an IQR of 0.45

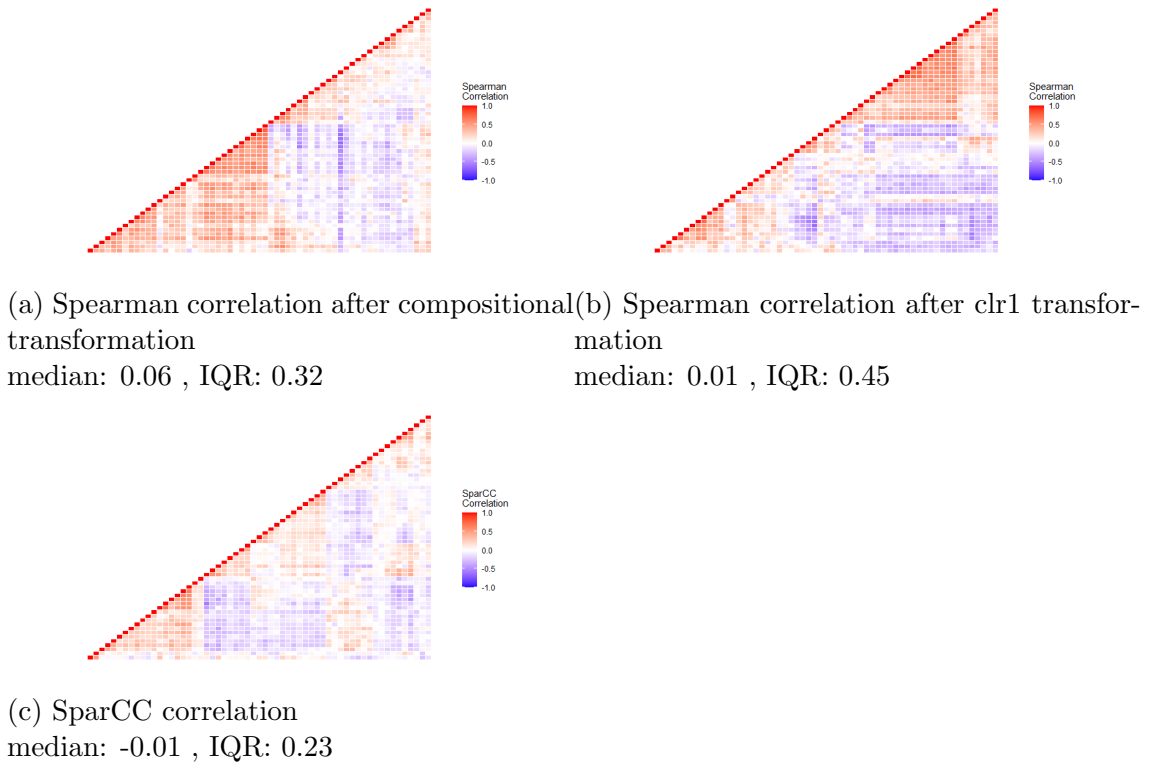


Figure 1: Correlation heatmaps that result from the three methods

SparCC

The SparCC algorithm is applied using the SpiecEasi R package [20]. Parameters are: 20 outer iterations, 10 inner iterations, and a correlation threshold of 0.1. Since SparCC is non-deterministic, this procedure is run 10 times and the correlation matrix is averaged. The resulting correlation matrix has a median of -0.01 and an IQR of 0.23. This IQR that is significantly lower than those of the CLR1 and Compositional transform, indicating that SparCC is more conservative in attributing correlation.

4.2 Network construction

TO find a good soft power β , scale-free topology fit is plotted for all three methods using WGCNA's inbuild functions. The resulting plots show extremely high variance making it impossible to reliably pick a good β . The plots can be found in the appendix/adjacent material. The alternative approach described by Bartzis *et al.* [17] was used. We take the smallest whole number for β that satisfies equation 17, which is 3.

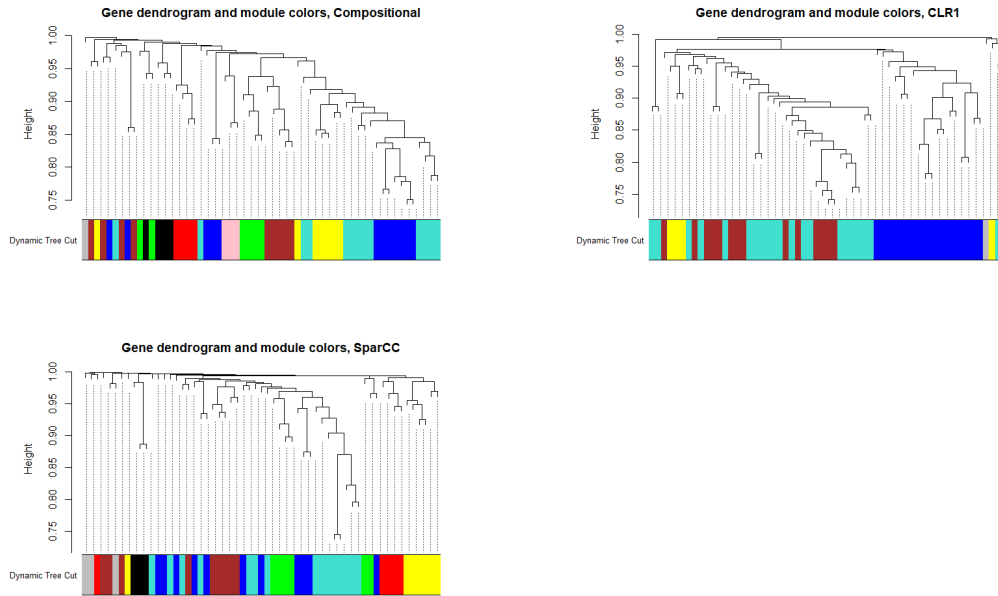


Figure 2: Dendrograms and clustering of all three methods

Network adjacency matrices are constructed from the correlation matrices using the function *adjacency.fromSimilarity()* from the WGCNA R package. Topological Overlap matrix and dissimilar TOM are calculated. Using hierarchical clustering, dendrograms are generated using AVERAGA and the dendrogram is separate into clusters using WGCNA’s function *cutreeDynamic()*. The parameter for minimum module size is set to three bacteria. All parameters are kept the same for all three methods.

The resulting modules and networks can be seen in figure 3.

One can easily see that the three methods result in networks that look very dissimilar visually. As expected, the number of edges that satisfy the 0.05 threshold are much fewer in SparCC network than the other two.

4.3 Consensus Analysis

To examine the effects of the different pre-processing approaches we will now take a more detailed look at the clusters we found. If the pre-processing methods had little impact on the results, we would find that the clusters found following pre-processing would have significant overlap. However, if we find large difference in the generated clusters between methods, we can conclude that the three approaches cause different

4 APPLICATION TO IMMUNO DEFFICIENCY MICROBIOME DATASET 16

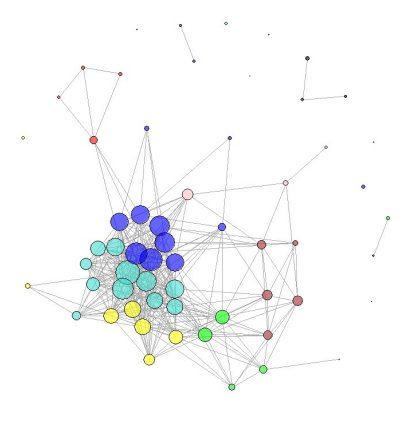
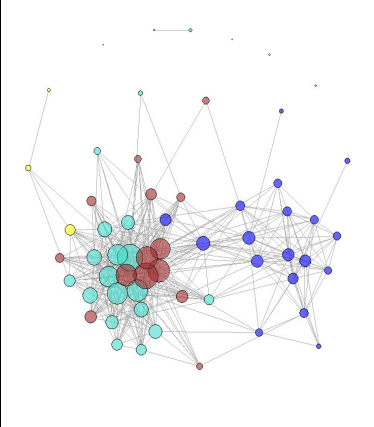
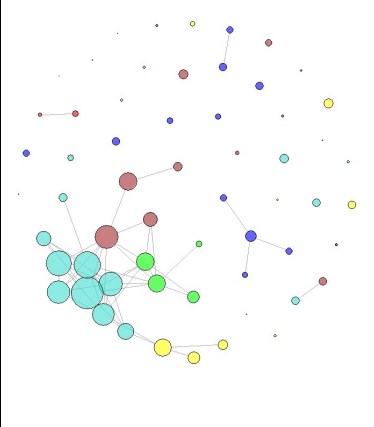
Compositional		CLR1		SparCC	
Module (#nodes)	Bacteria	Module (#nodes)	Bacteria	Module (#nodes)	Bacteria
Turquoise (13)	Alloprevotella, Prevotellaceae, Mogibacterium, Eubacterium_nodatum_group, Butyrivibrio_2, Lachnoanaerobaculum, Stomatobaculum, Lachnospiraceae, Ruminococcaceae_UCG-014, Solobacterium, uncultured.1, Leptotrichia, Kingella	Turquoise (21)	F0332, Bifidobacterium, Bifidobacteriaceae, Rothia, Bergeyella, Chloroplast, Granulicatella, Lactobacillus, Catonella, Oribacterium, Lachnospiraceae, Dialister, uncultured.1, Fusobacterium, Streptobacillus, Eikenella, Kingella, Neisseriaceae, Aggregatibacter, Treponema_2, Corynebacterium_all	Turquoise (15)	Actinomyces, Atopobium, Prevotellaceae, Bergeyella, Abiotrophia, Mogibacterium, Butyrivibrio_2, Stomatobaculum, Ruminococcaceae_UCG-014, Solobacterium, Megasphaera, Veillonella, Neisseriaceae, Prevotella_all, Selenomonas_all
Blue (12)	Actinomyces, Rothia, Atopobium, Abiotrophia, Granulicatella, Streptococcus, Megasphaera, Veillonella, Actinobacillus, Aggregatibacter, Prevotella_all, Selenomonas_all	Blue (18)	Actinomyces, Atopobium, Porphyromonas, Capnocytophaga, Gemella, Lactobacillales, Mogibacterium, Butyrivibrio_2, Lachnoanaerobaculum, Stomatobaculum, Ruminococcaceae_UCG-014, Megasphaera, Veillonella, Leptotrichia, Neisseria, Haemophilus, Prevotella_all, Selenomonas_all	Blue (13)	Bifidobacterium, Bifidobacteriaceae, Alloprevotella, Chloroplast, Lactobacillus, Peptostreptococcus, Dialister, Fusobacterium, Phyllobacterium, Bradyrhizobium, Sphingomonas, Achromobacter, Corynebacterium_all
Brown (9)	F0332, Porphyromonas, Capnocytophaga, Bergeyella, Lautropia, Eikenella, Neisseria, Neisseriaceae, Haemophilus	Brown (14)	Alloprevotella, Campylobacter, Abiotrophia, Streptococcus, Eubacterium_nodatum_group, Peptostreptococcus, Solobacterium, uncultured_bacterium.5, Phyllobacterium, Bradyrhizobium, Sphingomonas, Achromobacter, Lautropia, Actinobacillus	Brown (10)	F0332, Gemella, Streptococcus, Lactobacillales, Eubacterium_nodatum_group, Streptobacillus, Lautropia, Kingella, Actinobacillus, Aggregatibacter
Yellow (7)	Campylobacter, Catonella, Oribacterium, Candidatus_Saccharimonas, uncultured_bacterium.5, Saccharimonadales, Corynebacterium_all	Yellow (4)	Prevotellaceae, Tannerella, Parvimonas, Peptococcus	Yellow (9)	Campylobacter, Catonella, Lachnoanaerobaculum, Oribacterium, Lachnospiraceae, Leptotrichia, Candidatus_Saccharimonas, uncultured_bacterium.5, Eikenella
Green (6)	Tannerella, Parvimonas, Peptococcus, Peptostreptococcus, Fusobacterium, Treponema_2			Green (5)	Porphyromonas, Capnocytophaga, uncultured.1, Neisseria, Haemophilus
Red (4)	Phyllobacterium, Bradyrhizobium, Sphingomonas, Achromobacter			Red (4)	Rothia, Tannerella, Granulicatella, Treponema_2
Black(4)	Bifidobacterium, Bifidobacteriaceae, Lactobacillus, Dialister				
Pink (4)	Gemella, Lactobacillales, Streptobacillus				
					

Figure 3: Table showing the modules generated by the different approaches and the names of the bacteria in those modules. Their respective networks are shown at the bottom, the circles indicating bacteria and their size is the degree of the bacteria in the network. Edges over 0.05 are shown as lines between nodes.

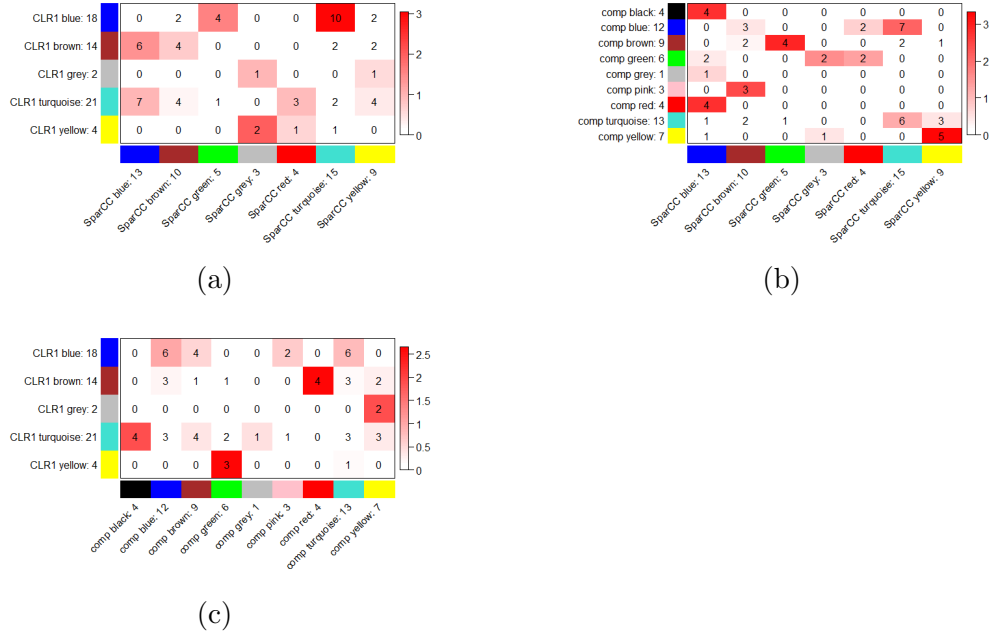


Figure 4: Consensus tables comparing all three methods

outcomes, since it's the only part of the process that has been modified. All other parameters have been kept equal across the three methods.

figure 4c shows the module overlap between the CLR1 clusters and the Compositional clusters. One can see that everything is spread out, there are hardly any modules where both methods agree on. Some of the best agreements are comp: red and CLR1: brown, all bacteria that belong to the comp: red module, also belong to the CLR1: brown module. CLR1: blue shows a more illustrative story: the bacteria that are clustered together in the blue module when using CLR1 are spread out across 4 modules when using a Compositional only approach. Similar lack of agreement can be seen when comparing SparCC with CLR1 or SparCC with Compositional. Bacteria that are unable to become a member of a cluster end up in the grey module. Some examples of modules shown as networks can be seen in the A

5 Conclusion and discussion

Considering the medical relevance of the human microbiome, standard statistical methods are necessary to extract useful information from microbiome sequencing data. Network techniques have been shown to capture important relations often missed by univariate approaches and have become one of the most popular methods

of analysing microbiome data. To create networks, one needs the pairwise correlations between all bacteria.

Unfortunately, the challenging nature of the data, caused by its compositionality, sparsity and high-dimensionality introduce various statistical artifacts that are difficult to deal with. There is currently not one set approach to deal with these challenges, with researchers using a wide array of different methods, resulting in a fragmented research field.

To investigate the effects of various methods they were applied to the oral section of the CVID dataset. This dataset contains oral samples of 152 volunteers, part healthy controls, part patients with a compromised immune system.

Three different methods (compositional, CLR1, and SparCC) have been implemented to produce correlation matrices necessary for network techniques. The networks generated from these matrices, and the clusters found with them, have been shown to be heavily impacted by the pre-processing approaches used. Especially the clusters found using hierarchical clustering were shown to be extremely dissimilar. Code can be found on [Github](#).

SparCC is most likely the superior method since it was conservative in attributing correlations, while we know that the other two approaches produce spurious correlations caused by the statistical properties of the data.

This paper shows that complex techniques often originating in fields like computer science and AI can help solve challenging problems in a different domain.

One approach similar to SparCC is Spiec Easi (SE) [21]. SE is another complex technique that promises to deal with the challenges of microbiome data. It would be interesting to see how these two methods compare. No strong advice can be given based on this thesis since the underlying biological relationships remain unknown. We can say with certainty that different methods lead to different results, not which result is correct.

Further research utilizing simulated datasets with full knowledge of the underlying system might provide valuable information and enable us to separate approaches. Best case would be that this leads to a clear result that accelerates adoption of the singular best approach throughout the field.

The dataset used in this paper is considered very small (59×152), network techniques are often applied to much larger datasets with many more variables and

samples. Since larger networks are considered more robust, its small size might have exacerbated the differences found between the different methods.

References

1. Waldron, L. Data and statistical methods to analyze the human microbiome. *Msystems* **3** (2018).
2. Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J. & Jiang, Y. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Frontiers in Genetics* **10**, 995. ISSN: 1664-8021 (2019).
3. Layeghifard, M., Hwang, D. M. & Guttman, D. S. in *Microbiome analysis* 243–266 (Springer, 2018).
4. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4** (2005).
5. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559 (2008).
6. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**, e1002687 (2012).
7. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature microbiology* **2**, 1–7 (2017).
8. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature microbiology* **4**, 293–305 (2019).
9. Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V. & Wargo, J. A. The microbiome, cancer, and cancer therapy. *Nature medicine* **25**, 377–388 (2019).
10. Vuong, H. E. & Hsiao, E. Y. Emerging roles for the gut microbiome in autism spectrum disorder. *Biological psychiatry* **81**, 411–423 (2017).
11. Jiang, H. *et al.* Altered fecal microbiota composition in patients with major depressive disorder. *Brain, behavior, and immunity* **48**, 186–194 (2015).
12. Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V. & Knight, R. Current understanding of the human microbiome. *Nature medicine* **24**, 392–400 (2018).
13. Aitchison, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44**, 139–160 (1982).

14. Aitchison, J. & Egozcue, J. J. Compositional data analysis: where are we and where should we be heading? *Mathematical Geology* **37**, 829–850 (2005).
15. Erb, I., Gloor, G. B. & Quinn, T. P. Compositional data analysis and related methods applied to genomics—a first special issue from NAR Genomics and Bioinformatics. *NAR Genomics and Bioinformatics* **2**, lqaa103 (2020).
16. Silverman, J. D., Roche, K., Mukherjee, S. & David, L. A. Naught all zeros in sequence count data are the same. *Computational and structural biotechnology journal* **18**, 2789 (2020).
17. Bartzis, G., Deelen, J., Maia, J., Ligterink, W., Hilhorst, H. W., Houwing-Duistermaat, J.-J., van Eeuwijk, F. & Uh, H.-W. Estimation of metabolite networks with regard to a specific covariable: applications to plant and human data. *Metabolomics* **13**, 1–17 (2017).
18. Berbers, R.-M. *et al.* Low IgA Associated With Oropharyngeal Microbiota Changes and Lung Disease in Primary Antibody Deficiency. *Frontiers in immunology* **11**, 1245 (2020).
19. Levy, M., Kolodziejczyk, A. A., Thaïss, C. A. & Elinav, E. Dysbiosis and the immune system. *Nature Reviews Immunology* **17**, 219–232 (2017).
20. Kurtz, Z., Mueller, C., Miraldi, E. & Bonneau, R. *SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference* R package version 1.1.0 (2020).
21. 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* **11**, e1004226 (2015).

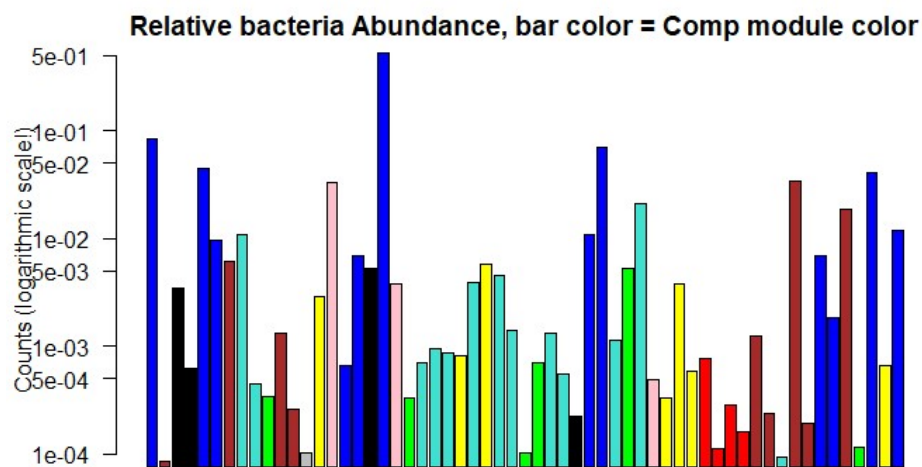


Figure 5: Relative abundances of the bacteria, colors show the cluster the bacteria belongs to. Clusters from Compositional approach

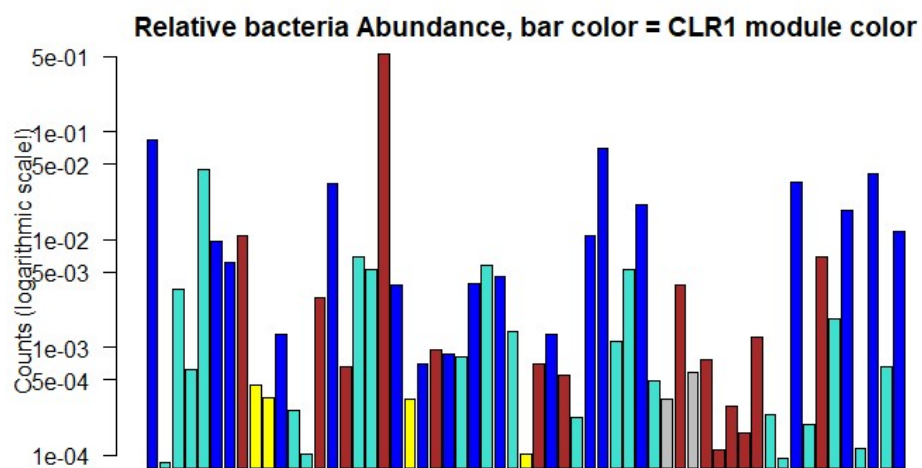


Figure 6: Relative abundances of the bacteria, colors show the cluster the bacteria belongs to. Clusters from CLR1 approach

A Appendix

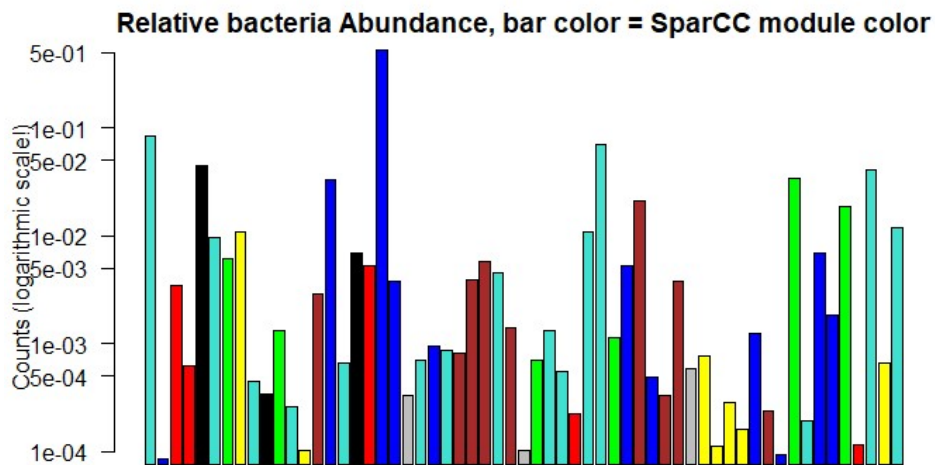


Figure 7: Relative abundances of the bacteria, colors show the cluster the bacteria belongs to. Clusters from SparCC approach

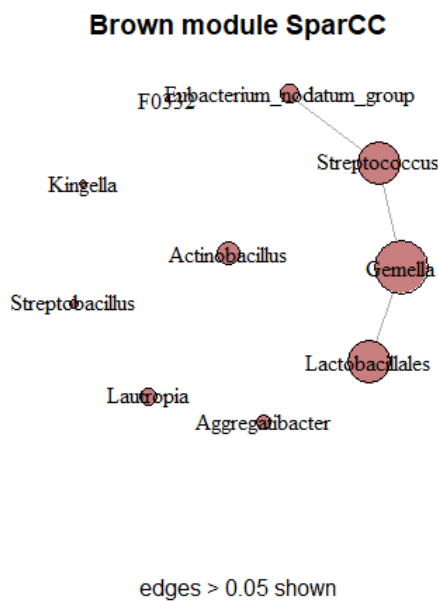


Figure 8: Network of brown module from the SparCC approach

Turquoise module SparCC

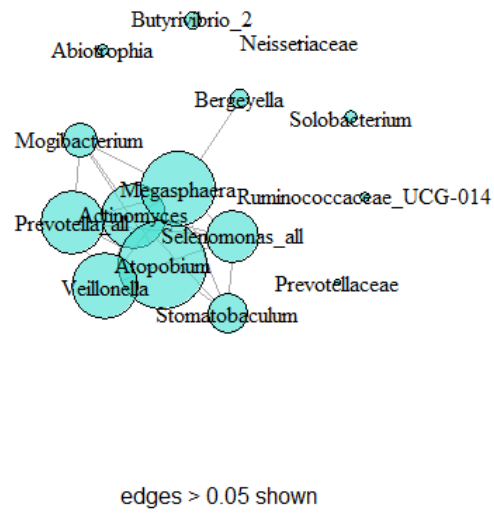


Figure 9: Network of turquoise module from the SparCC approach

yellow module SparCC

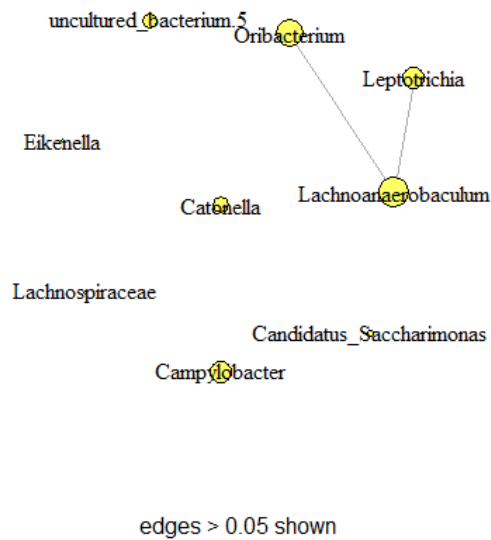


Figure 10: Network of yellow module from the SparCC approach

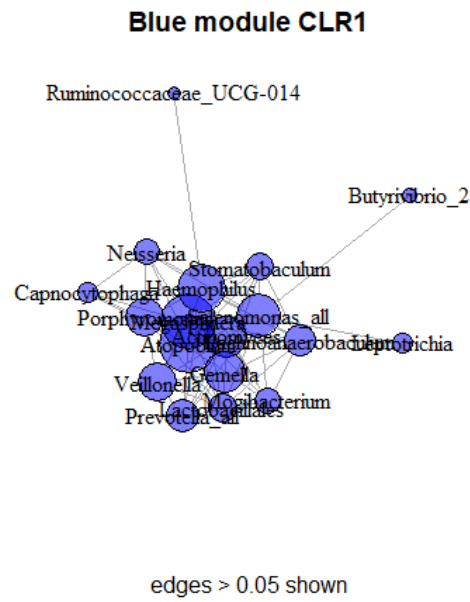


Figure 11: Network of blue module from the CLR1 approach

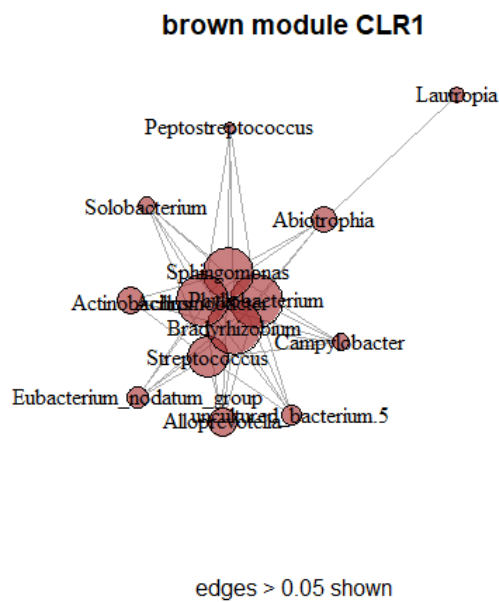


Figure 12: Network of brown module from the CLR1 approach

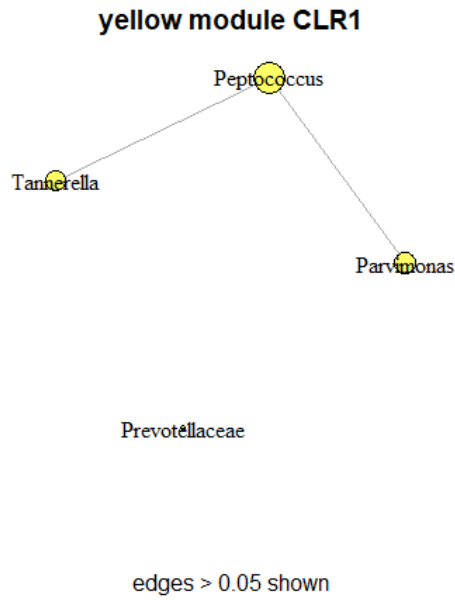


Figure 13: Network of yellow module from the CLR1 approach

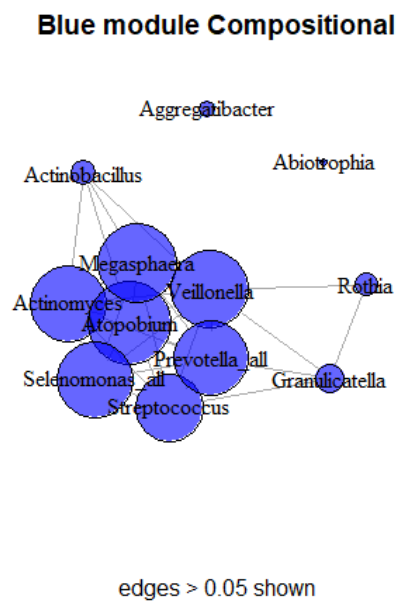


Figure 14: Network of yellow module from the Compositional approach

Brown module Compositional

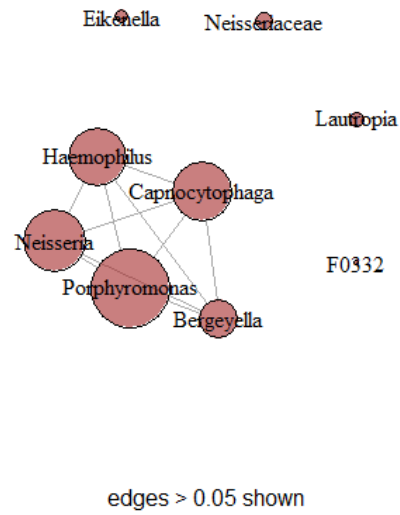


Figure 15: Network of brown module from the Compositional approach

Pink module Compositional

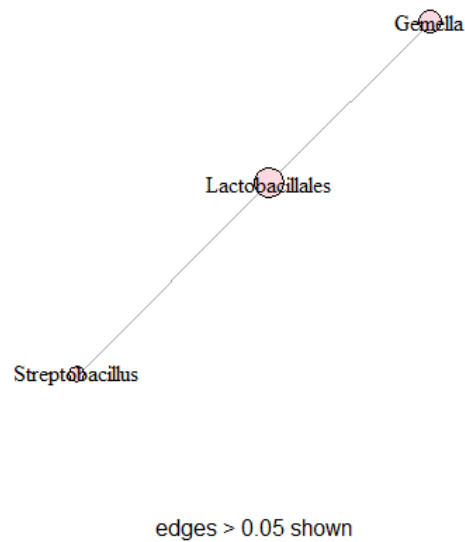


Figure 16: Network of pink module from the Compositional approach