# microbiomeSeq: An R package for analysis of microbial communities in an environmental context.

Msc. Bioinformatics, Polyomics and Systems Biology

Alfred Ssekagiri (Student ID: 2286345)

Supervisor: Dr. Umer Zeeshan Ijaz

Co-supervisor: Prof. William T. Sloan

August 21, 2017

A report submitted in partial fulfillment of the requirements for MSc Bioinformatics, Polyomics and Systems Biology Degree at the University of Glasgow.

## 0.1 Acknowledgement

I would like to acknowledge my supervisor Dr Umer Zeeshan Ijaz for the guidance and support provided through the course of this project. In addition, I extend my heart felt gratitude to the Makerere University Center of Excellence for Infection and Immunity training and research for the financial support towards this degree.

# Contents

## 0.2 List of abbreviations/ acronyms

**OTU**: Operational Taxonomical Unit.

**PERMANOVA**: Permutational Analysis of Variance.

**ANOVA**: Analysis of Variance.

**PCoA**: Principle Coordinate Analysis.

**NMDS**: Non-metric Multidimensional Scaling.

**FSO**: Fuzzy Set Ordination.

**LCBD**: Local Contribution to Beta Diversity.

**KMDA**: Kernel-based Metabolite Differential Analysis.

**MDA**: Mean Decrease in Accuracy.

**lfc**: log fold Change.

## 0.3 Summary

microbiomeSeq is an R package developed to enhance the available statistical analysis procedures for microbial communities data obtained from 16S ribosomal RNA sequencing and to provide more informative visualisation of the results. It mainly focuses on community diversity within and across samples, differential expression analysis of taxa between conditions, co-occurrence patterns analysis at community level and the relationships between and environmental traits with community structure. Main features of the package include the following:

**Alpha diversity:** Taxa distribution within samples is measured using common diversity indices and then compared amongst specified groups or conditions by analysis of variance.

**Beta-diversity:** Taxa distribution across samples is explored using multivariate analysis techniques. samples are ordered such that similar samples are closer to each other than dissimilar ones, variability of diversity amongst multiple conditions is explored using permutation analysis of variance. In addition, homogeneity of variance between conditions or groups in the samples is explored.

**Differential expression analysis:** Abundance of each feature is compared between or among conditions to identify up and down regulated features using Kruskal-Wallis test and DESeq package. Features are assigned importance using random forest classifier. In addition, kernel based differential analysis is used to group taxa into sets basing on correlation that must exist between or among features that belong to the same set. Then, these sets are tested for differential expression using a distance based score test.

**Co-occurrence pattern analysis:** This is used to identify co-occurring taxa at community level under specified environmental conditions. Co-occurrence is measured as positive correlation whose threshold is specified. Amongst these features, pairwise co-occurrences which are outstanding within sub communities are detected. Taxa are assigned roles depending on their linkage within respective sub communities and the entire network.

**Relationships between microbial community and environmental traits:** Two procedures are implemented, the first one directly correlates taxa abundance and environmental variables and the other uses fuzzy set ordination to test effects of perturbation in environmental variables to community structure.

We use a dataset which was generated by 16S ribosomal RNA sequencing of various latrines from Tanzania and Vietnam at different depths.

## 0.4 Introduction

Microbial communty studies have increasingly gained popularity given the high importance of microbiodiversity to human and environmental health [Morgan and Huttenhower, 2012]. This calls for approaches to efficiently identify, characterize and evaluate microbial communities. Generally microbial community refers to a group of different populations of micro-organisms co existing under certain environmental conditions [Oulas et al., 2015]. Studying microbial communities reveals underlying characteristics of the micro organisms in a complex community for example, microbial composition: which microbes are present and how much they are in the community, function of different microbes within the community, relationships between or among different microbes, and how they respond to their environment.

Tremendous advancement of high-throughput sequencing also known as next generation sequencing has lead to great improvement in studying microbial communities [Morgan and Huttenhower, 2012] hence the field of metagenomics whose main goal is to understand composition and functional diversity of microbial communities. This technology enables direct analysis of genomes available in environmental samples as opposed to traditional techniques which require single-cell culture where microbes are grown on solid-media [Oulas et al., 2015]. The prominent sequencing technologies used for this include: 454 pyrosequencing and Illumina systems although the later in particular Miseq has been shown to have a better performance as compared to other technologies[D'Amore et al., 2016].

Here we focus on the analysis based on amplification of genes of interest commonly referred to as marker genes. These include the highly conserved 18S and 16S ribosomal RNA genes for eukaryotes and prokaryotes respectively [Oulas et al., 2015]. Basing on a defined similarity threshold usually $97\%$, sequences are clustered into operational taxonomic unit (OTU) [Morgan and Huttenhower, 2012]. Examples of algorithms designed for this include: UPARSE [Edgar, 2013] which has been shown to report less than $1\%$ incorrect biases USEARCH Edgar [2015] and QIIME [Kuczynski et al., 2012] which is more established than most of these.

Alternatively, the sequence reads are directly assigned species for example using DADA2 [Callahan et al., 2016] algorithm which provides a fine scale resolution to extent of characterizing sequences differing by a single nucleotide. This is a more informative procedure because the fine scale resolution has been shown to be critical to ecological niches and consequently crucial to microbiome associated phenotypes most especially in clinical studies [Eren et al.,

2014]. For that matter, grouping similar sequences into a single unit would imply neglecting biological variation amongst those sequences.

In both approaches, taxonomic assignment is done with reference to established databases such as Silva, Green Genes and Ribosomal Database Project (RDP). The community is described in terms of operational taxonomic units it contains and their phylogenetic relationships.

Main information that is output by above algorithms include: OTU or species abundance table which is a matrix type usually with samples in rows and OTUs as columns, corresponding taxonomic classification tables, OTUs as rows and taxonomic levels as columns and or phylogenetic tree mainly in NEWICK format. These can then be associated with corresponding environmental data mainly referred to as meta data which is also usually a matrix object with samples as rows and environmental variables in columns.

Statistical methods are used to evaluate and analyse the community data so that it can be turned into further biological insights. Various tools have been developed for this purpose but to the best of our knowledge, the analysis procedures and visualisation options can be improved. Therefore, rather than re-inventing the wheel, we choose to build upon existing packages such as such as vegan [Oksanen et al., 2007], phyloseq [McMurdie and Holmes, 2013], DESeq2 [Love et al., 2014] to enhance analysis procedure by extending some of the available functions and creating more informative plots with on-figure results but with less clutter. It is mainly aimed at measuring taxonomic distribution within and across samples that is alpha and beta diversity respectively, identifying up and down regulated taxa under a given pair or group of conditions, exploring co-occurrence patterns within communities and relationships between community composition and environmental traits.

## 0.5 Methods and dependencies

In this section, we describe statistical methods to evaluate and analyse the community data so that it can be turned into further biological insights.

**Alpha diversity:** This is the distribution of features within samples. Different diversity indices are used for this measure. Simpson index which accounts for the number of species present and their relative abundance in the community, Shannon which is a commonly used index to characterise species diversity, Pielou's evenness which measures the closeness of each species, Fisher alpha is a parametric index of diversity that models species as logseries distribution.

These measures provide useful information the community composition in terms of rarity and commonness of species in the community. Analysis of variance (ANOVA) is used to compare these diversity measures between pairs of groups. The generated p-values are used to assess the significance of differences between groups. The function *diversity* of vegan package *vegan* [Oksanen et al., 2007] is used for calculating diversity and *aov* function for analysis of variance.

**Beta diversity:** This is a measure of features distribution across samples. It is measured by using multi variate statistics procedures such as ordination. This is the ordering of samples such that those with similar diversity are closer to each than those with dissimilar diversity. The different methods of ordination aim to provide a low dimension representation of the samples though different procedure is followed for a particular method. The similarity or dissimilarity is based on distance between samples.

Distance measures used include: Bray Curtis, which takes into account abundance of species, Unifracs which use phylogenetic distance between the branch lengths of features. Weighted Unifrac is weighted by features abundance and UnWeighted Unifrac does not consider abundances of features. We use phyloseq [McMurdie and Holmes, 2013] package to calculate the distances. Below are the ordination methods that we have used.

Non multi dimensional scaling (NMDS) aims at representing pairwise dissimilarity between samples in low dimension space. The distance between samples is calculated using a selected distance measure or dissimilarity coefficient. Monotonic regression is then used to compare the ranked distances and the original dissimilarity matrix by optimizing a stress function which measures how similar the ranked distances are to the original distance on a scale of

[0,1]. Function *metaMDS* of the *vegan* package is used for NMDS ordination.

Another ordination method is principle coordinate analysis which seeks to represent the data in the lowest possible dimensions without much loss of information. Similar to NMDS, it also takes a dissimilarity object of samples and generates a low dimension representation with eigenvalues for each resultant dimension. The eigenvalues represent the amount of variation in the dataset explained by a particular dimension. This is implemented using a function *capscale* of the *vegan* package [Oksanen et al., 2007].

We use permutation analysis of variance to test whether the variance between samples for multiple conditions is the same or not. This is implemented using *adonis* function of vegan package. permutation analysis of variance is a non parametric method that performs distance based multivariate analysis of variance of objects between or among groups. It involves calculating within group and between group dissimilarities using a selected dissimilarity coefficient or distance measure. Then, within and between group variances are compared using F-test. Significance of the result is based on comparing F-test result and random permutations of samples between conditions.

To test whether spread of diversity between samples under multiple conditions or groups is the same or not, we use *betadisper* function of vegan package which tests for homogeneity of variances. Non-euclidean distance measure such as Bray Curtis and Unifracs are used to compute distance between samples and group centroids. These distances are reduced to principal coordinates and subjected to ANOVA to test whether they are different or not. The significance of the result is based on comparing generated p-values to a user defined threshold. A significant result implies that, the spread of diversity under a particular pair of conditions is different and vice versa. The implementation is as proposed by [Anderson et al., 2006] using an specified dissimilarity measures.

**Differential expression analysis:**

We explore features which are up or down regulated under given conditions. This is worth exploring because it reveals difference in expression among groups which could be directly linked to observed characteristics in the groups. For example, in case of case-control studies, the observed phenotypes can be attributed to features that are either down and up regulated in one group with respect to the other . In other words, differential expression analysis helps us to identify individual features which may be most linked or responsible for differences observed

among or between studied groups. The tools used for this purpose include DESeq2 [Love et al., 2014] package and Kruskal-Wallis test as explained below.

Deseq [Love et al., 2014] procedure which was originally developed for differential expression analysis of RNA-seq data is used. In this case, abundance of a feature in a given sample is modelled as a negative binomial distribution, whose mean depends on sample specific size factor and concentration of that feature in a sample. Wald test is used to test significance of coefficients based on sample estimates. A feature is described as differentially expressed between the conditions by a threshold p-value and log2 fold change.

Kruskal-Wallis is a non parametric method for testing whether samples originate from the same distribution. Unlike DESeq2, this method does not assume any particular distribution of taxa abundance. The abundance of each is tested between/among specified conditions. Since the the same test is performed multiple times on the same dataset, the p-values values generated are corrected for multiple testing using family wise error rate. Depending a p-value threshold specified by the user, a given feature is described as significantly differentially expressed or not.

The procedure above only identifies up and down regulated features but does not reveal which features are actually more important than others in the community. Therefore, we need a procedure which can attach a measure of relevance to these features. By that, we can point out with more confidence individual features which are most probably responsible for observed differences in phenotypes or any other difference being investigated between the groups/conditions.

Differentially expressed features are classified using random forest classifier implemented by the *importance* function of randomForest package [Liaw and Wiener, 2002] to find most important features. The measure used in this case is Mean Descrease in Accuracy. This is obtained by removing the relationship of a feature and measuring increase in error. Consequently, the feature with high mean decrease in accuracy is considered most important.

To explore differential analysis at group level, we use set level differential analysis. This provides information about groups of features that respond more or less the same way in a community under a given pair of conditions. A set of features is generated depending on a threshold correlation that must exist among features of a certain group. we use a kernel based metabolite differential analysis (KMDA) [Zhan and Ghosh, 2015] package which allows set-level differential analysis strictly under a specified pair of conditions. sets of features are generated using

*group.pearson* and to test whether these sets are differentially expressed under the specified conditions, we use functions *dscore* and *sscore* . Similar analysis is implemented for numerical variables of the sample data.

**Co-occurence pattern analysis:**

Co-occurrence patterns are useful since they reflect processes that maintain the co existence of different micro organisms within a given microbial community. Co-occurring species share similar ecological characteristics and as such, they may consequently be involved in common biological processes and may be depending on each other in one way or another. Applying this to microbial community analysis can identify traits of co-occurring taxa and interactions between taxa within the community.

This is used to identify co-occurring features at community level under specified environmental conditions.This implementation follows the procedure presented by [Williams et al., 2014]. Sub community of co-occurring features within the community and then identify pairwise co-occurrences within sub communities based on correlation between a given pair of features. p-values generated during pairwise correlations tests are adjusted for multiple comparisons by false discovery rate.

Co-occurrence is measured as positive correlation whose threshold is specified. Negative correlation is indicative of competition between a given pair of features or non overlapping niches. The correlation and associated p-values are calculated by functions *corAndPvalue* and *bicorAndPvalue* of WGCNA [Langfelder and Horvath, 2012] package. A network showing co-occurrence is generated with features as nodes and edges as correlation between the corresponding pair of features. The network statistics used to assign importance to features include betweenness, closeness and eigenvector. Packages used for to implement this include: igraph package [Csardi and Nepusz, 2006] which provides functions to calculate network statistics.

**Topological roles of taxa:** Taxa in identified sub communities are assigned roles in the network using a procedure provided by [Guimera and Amaral, 2005]. Two metrics used for this purpose. First is within-module degree z-score which measures how well a particular feature is connected to others in the same subcommunity. It is given by Equation (1)

$$z_i = \frac{k_{im} - \bar{k}_m}{\sigma_{k_m}}, \tag{1}$$

where $k_{im}$ is the number of links of taxon $i$ in sub community $m$, $\bar{k}_m$ and $\sigma_{k_m}$ are the respective

mean and standard deviation of number of links for sub community $m$.

The second metric is the among-module connectivity which measures how a feature is linked to other modules in the network also refered to as participation coeffiecient is given by Equation (2).

$$p_i = 1 - \sum_{h=1}^{N_m} \left( \frac{k_{ih}}{k_i} \right)^2, \tag{2}$$

where $k_i$ is the number of links of taxon $i$ in the entire network and $k_{ih}$ is the number of links of taxon $i$ to sub community $h$.

Features are assigned roles depending on where they lie in the z-p space as given by [Guimera and Amaral, 2005]. A taxon is a module hub if $z \geq 2.5$ and a non hub if $z < 2.5$. Non hubs classified into four groups that is ultra peripherals ($p \leq 0.05$), peripherals ($p \in (0.05, 0.62]$), connectors ($p \in (0.62, 0.80]$), kinless ($p > 0.8$). Module hubs are classified as provincial ($p \leq 0.30$), connector ($p \in (0.30, 0.75]$) and kinless ($p > 0.75$).

Other packages used include: ggplot2 package [Wickham, 2016] is designed for generating visualisations. The tools provided are used to generate plots for the analysis results with support from other packages including gtable and gridExtra [Auguie, 2016].

## 0.6 Product Design

This section entails the functional aspects available for this package using the methods described in Section 0.5 and visualisations which are produced by ggplot [Wickham, 2016]. The input data is required as phyloseq object containing taxa abundance, sample data, taxonomy assignment and associated tree. Components of these may not be necessary depending on the aspect being investigated, phyloseq has functions to manipulate the data and that makes it very useful in this package.

### 0.6.1 Alpha diversity

Given a phyloseq object, vector of indices' methods and a variable specifying groups in the dataset, diversity measures are computed and compared amongst groups by pairwise ANOVA. A visualisation showing the results is generated where groups that show significant variance in diversity measure are annotated with significance labels. The significance is based on comparing ANOVA p-values and a user set threshold.

### 0.6.2 Beta diversity

Given a phyloseq object, ordination method and grouping variable, an ordination is performed and samples are grouped for different groups. Mean ordination value is calculated and spread of points are drawn as ellipses.

Pairwise dispersion between groups in the samples are calculated and significance is assessed using p-values whose threshold is specified by the user. PERMANOVA of taxa abundance amongst conditions is performed and results are reported. The resulting p-value is compared to the user-set threshold to assume significance.

A plot of the first two dimensions of ordination is produced and the most significantly dispersed groups are annotated on the plot with corresponding significance labels. In case the PERMANOVA results are significant according to pre-set p-value threshold, then p-value and $r^2$ are also annotated on the ordination plot. For case of NMDS, the value of STRESS function is also annotated and for PCoA, the axes have percentage values which correspond to variance explained by respective axes in the dataset.

Local contribution to beta diversity (LCBD) is calculated and most abundant features are de-

tected by taking sums of observed abundance values for each feature in all samples. The number of most abundant taxa to be used in this calculation can be specified.
The results are visualised in a plot which has points at the bottom whose diameter corresponds to magnitude of LCBD value of a particular sample and bars which correspond to taxa that are most abundant with the top taxa sharing a bigger portion of the bar for each sample.

### 0.6.3 Differential Expression Analysis

Abundance of each feature is compared between or among conditions to identify up and down regulated features. In case of Kruskal-Wallis test, the p-values generated are corrected for multiple comparisons by family wise error rate. Similarly, DESeq is also used to identify differentially expressed features. These features are then assigned importance using random forest classifier.
A filename can be specified to which the significant features and corresponding log2 fold change, basemean, p-values and group where they are upregulated can be written. This is important for reproducibility and further inference. Plots produced include:

Significant features plot: It shows box plots of taxa abundance distribution in groups annotated with names and p-values of the taxa and corresponding ranks of importance.

MA plot: A plot of log2 fold change against mean abundance of most significant features with an option to label the taxa or not. This reduces clutter in case of very many significant features.

lfc plot: This plot shows down and up regulated features with base mean values annotated on top of bars. Size of the bars corresponds to magnitude of log fold and sign orientation refers to up or down regulation.

MDA plot: This is a standalone visualisation of mean decrease in accuracy measure for each of the significant features. Bigger values of mean decrease in accuracy (MDA) represent higher importance.

Plot of multi testing corrections only for Kruskal-Wallis test.

Kernel differential analysis: Taxa is grouped into sets basing on correlation threshold(s) that must exist between or among features that belong to the same set. Then, these sets are tested

for differential expression using a distance based score test. A file is generated containing taxa and corresponding score statistics. The results are visualised in a plot showing sets of taxa with adjusted p-values and significant labels annotated.

### 0.6.4 Correlation between abundance and environmental traits.

The relationship between most abundant taxa and numerical environmental variables are based on correlation. A user selected correlation coefficient is used to compute correlation between abundance of each taxon and selected environmental variables. A correlation test is performed and p-values adjusted for multiple comparisons using Benjamin-Hochberg method. Significance of correlation is specified by a user defined p-value which is compared adjusted p-values. A plot is generated to visualise the results in heat map where taxa are in rows, environmental variables are in columns and correlation with significance labels annotated in the cells.

### 0.6.5 ANOVA of environmental variables

Selected environmental traits are compared between or among specified groups using ANOVA. The results are visualised as plots the distribution of variables annotated with significance of variation in specified groups. A user defined threshold p-value is compared to ANOVA p-value in order to assess significance.

### 0.6.6 Co-occurrence analysis

**Generating the network and sub communities:** Co-occurrence pattern analysis is used to identify co-occurring taxa in community under specified environmental conditions. Co-occurrence is measured as positive correlation whose threshold(s) can be specified as by the user via arguments. Amongst these features, pairwise co-occurrence which are outstanding within sub communities are detected. p-values generated during pairwise correlations are adjusted for multiple comparisons by false discovery rate. The network statistics used to assign importance of taxa include betweenness, closeness, and eigenvector centrality.

Output includes a network with subcommittees and plots of betweenness versus eigenvector centrality for a selected correlation thresholds. These can be further used to detect roles of each taxa in the network and also to study response of sub community to environmental traits.

**Assigning roles in sub communities:** Features identified in sub communities are assigned roles using two metrics which include: within-module degree score (Z) which measures how well a particular feature is connected to others in the same sub community and among-module connectivity (P) which measures how a feature is linked to other modules in the network. Features are classified as ultra peripherals, peripherals, provincial, connectors, kinless, module hubs, or non hubs depending on where they lie in the Z-P space.

**Correlation of taxa with environmental traits:** In a given sub community, a feature with the highest betweenness centrality is the very influential and is therefore a good representation of the sub community. The correlation of such a feature with environmental variables shows the response of a corresponding sub community to the environmental traits. A file of the correlation results and a visualisation. The plot is annotated with correlation significance labels.

## 0.7 Evaluation

In this section, we test the functionality of implemented tools using a pitlatrine dataset which was generated by 16S ribosomal RNA sequencing of various latrines from Tanzania and Vietnam at different depths during a study aimed at assessing the influence of intrinsic environmental and geographical factors on the bacterial ecology of pit latrines [Torondel et al., 2016]. The two countries are chosen in order to have a contrasting set of different pit latrines systems. The selection of latrines was based construction materials, design characteristics and number of individuals using it. Faecal material was collected at 20 cm intervals from top to bottom in each pit latrine. Therefore the samples in the dataset are categorically classified by country, latrine number and depth at which the material was derived.

The dataset comprises of an abundance table of 8883 OTUs in 81 samples, taxonomy assignment of 8883 OTUs at 6 levels that is Kingdom, phylum, order, family, class and genus, corresponding phylogenetic tree in NEWICK format and Meta data which comprises of eleven numerical variables including Total Solids (TS), Volatile Solids (VS), Volatile Fatty acids (VFA), Protein (Prot), Ammonia $(NH_4)$, Temperature (Temp), Carbohydrates (Carbo), total Chemical Oxygen Demand (CODt), soluble chemical oxygen demand (CODs) and percentage of CODt converted to CODs (perCODsbyt) and three categorical variables specifying latrine, depth and country from which a particular sample was generated.
Details of how to reproduce these results and using the package as well for similar datasets are available in a tutorial document attached to this project on a portable disk and the tutorial is available at microbiomeSeq tutorial.

**ANOVA Alpha diversity**

Within sample diversity was investigated using three diversity indices that is Richness, Simpson and Shannon and compared between countries at a threshold p-value of $5\%$. Figure 1 shows the distribution of diversity between Tanzania and Vietnam for the three diversity indices. It is quite evident that the mean diversity is higher in Vietnam than Tanzania for all measures. Moreover, there is significant different in microbe distribution between the two countries as illustrated by significance labels annotated on the plots. This suggests that geographical location is critical to microbes distribution in the community.

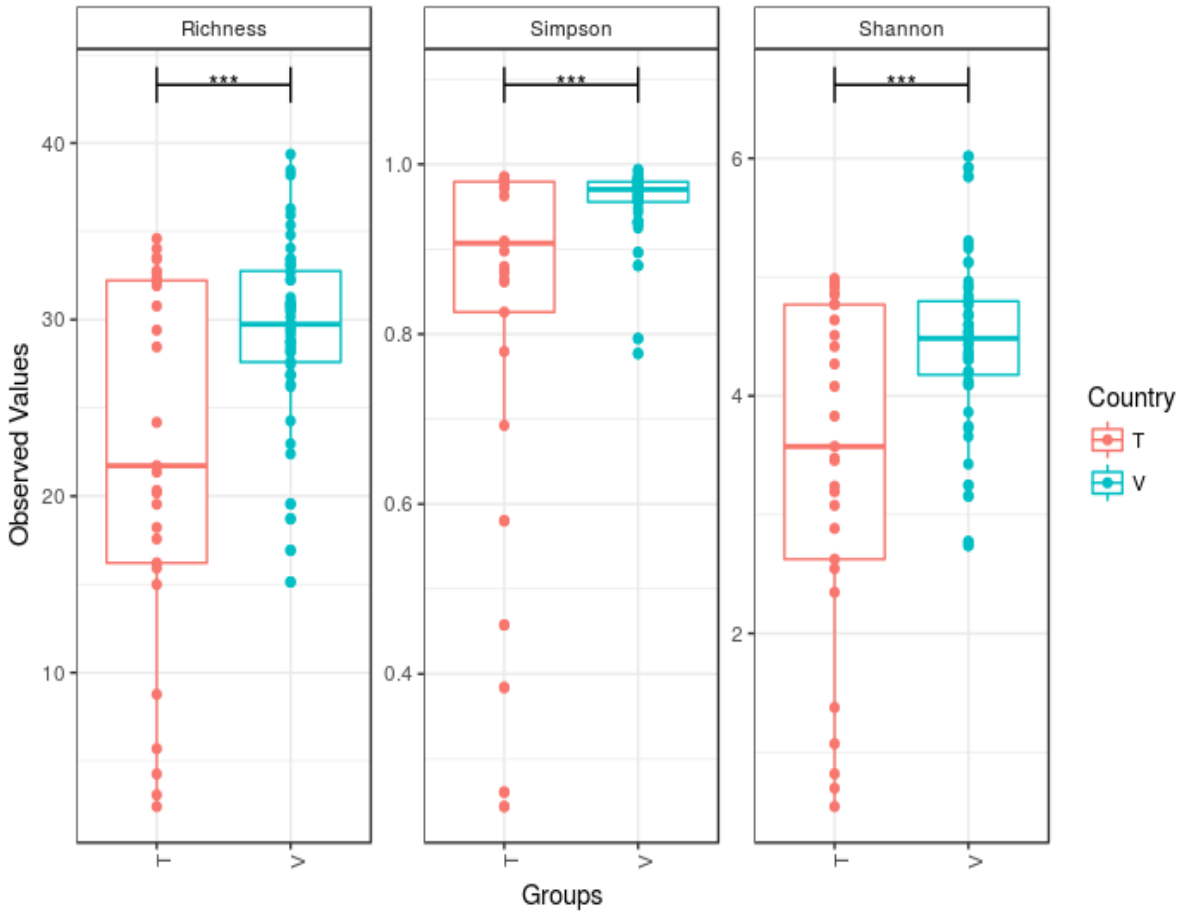


Figure 1: ANOVA diversity between Tanzania and Vietnam

**Ordination and beta dispersion**

We tested the combination of ordination and beta-dispersion using two techniques that is NMDS and PCoA using Depth as the grouping variable. Results were visualised as shown in Figure 3 and Figure 2 respectively. Threshold p-value for significance was set to $5\%$ for both beta-dispersion and PERMANOVA. The first and second dimension of PCoA explain $2.99\%$ and $3.51\%$ of the variance in sample diversity of microbes across samples.

PERMANOVA results show that diversity varies significantly amongst the different depths at which samples are obtained with a p-value of $0.01$ and that depth explains $15.4\%$ of the variance in diversity across samples. BETA-DISPERSION results show that depth 06 seems to have a different variance in diversity as compared to other depths. In case of NMDS the stress function value is $0.151$ which is more near to 0 than 1. This implies that the ordination is not a good representation of original dataset.

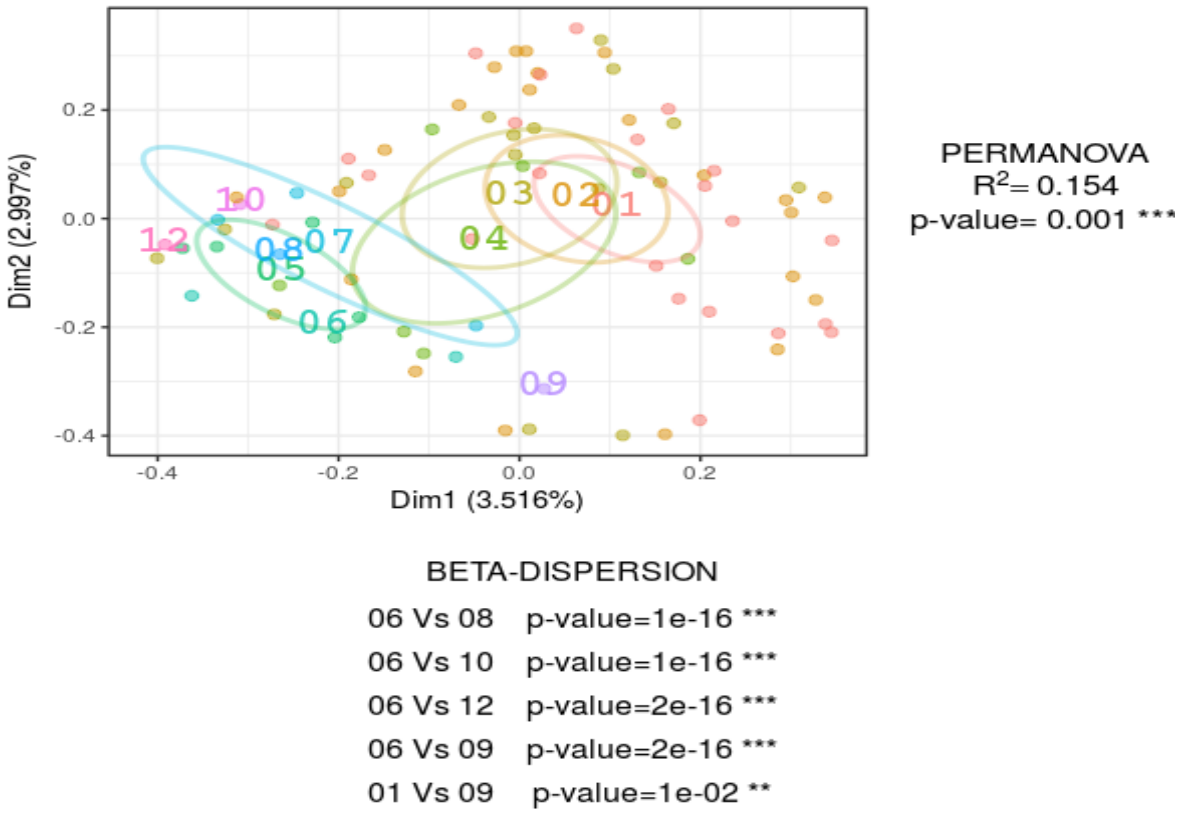


Figure 2: PCoA, PERMANOVA and beta-dispersion results.

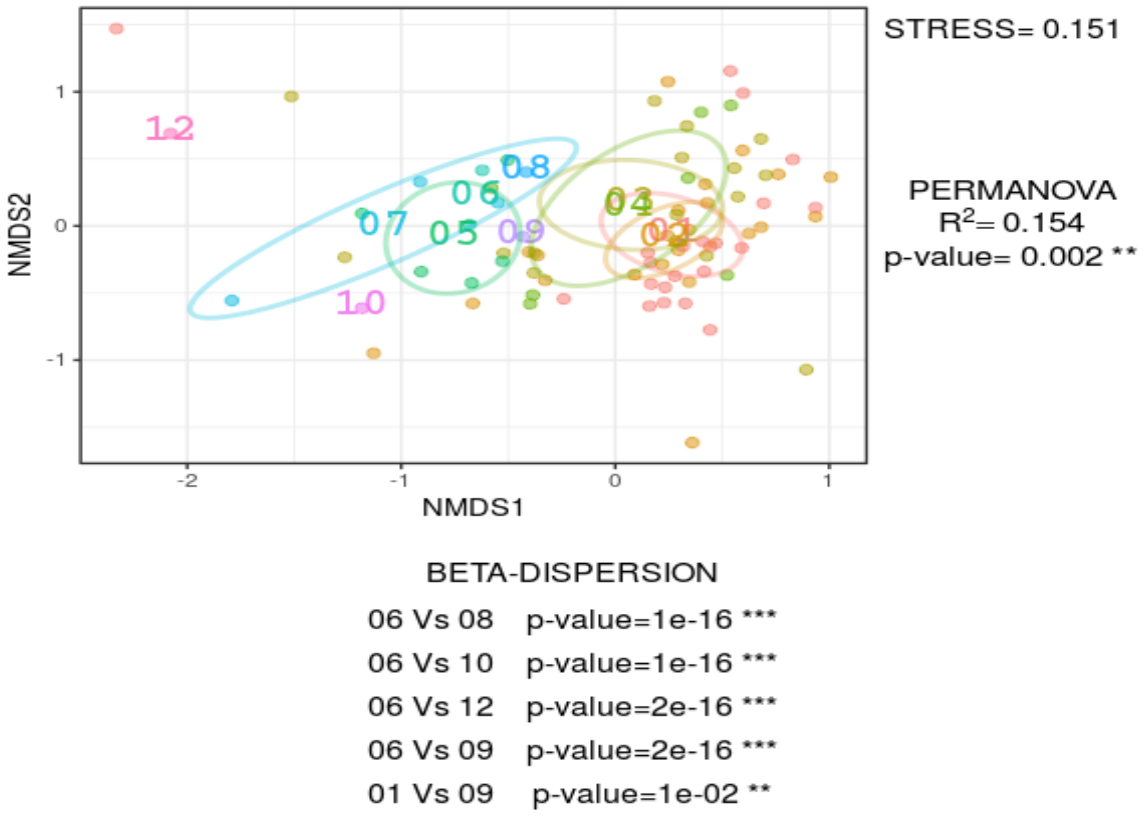


Figure 3: NMDS, PERMANOVA and beta-dispersion results.

Canonical correspondence analysis was used to identify environmental traits that best describe community structure of the two countries at a p-value threshold of $5\%$. As shown in Figure 4, proteins, carbohydrates, temperature and total solid are the best environmental variables. proteins, carbohydrates, temperature are mostly influential to the Tanzania community as Total solids to the Vietnam community.

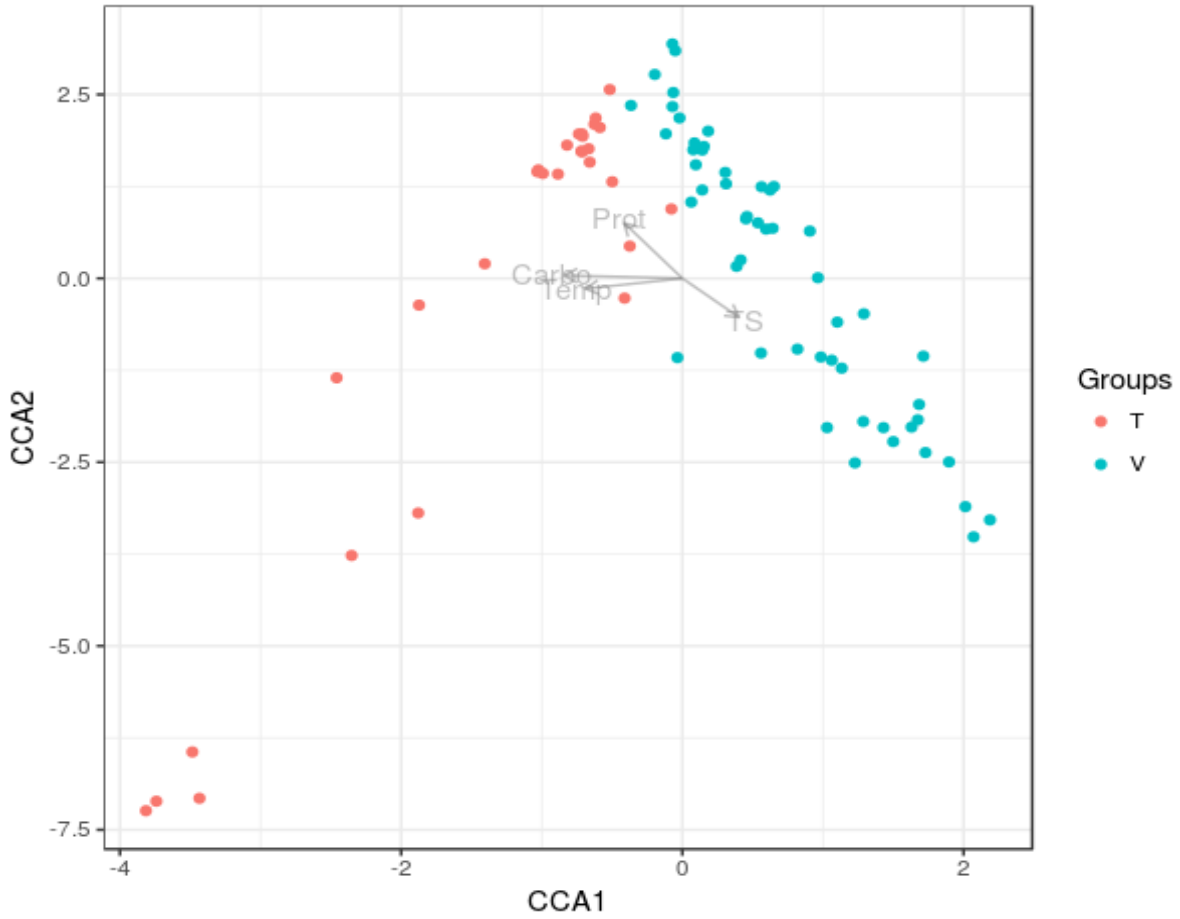


Figure 4: Canonical correspondence analysis results plot. Blue and red are for Vietnam and Tanzania respectively. The arrows show

We investigated the effects of small changes in environmental traits to community structure. Figure 5 shows a fuzzy set ordination result of community data between countries Tanzania and Vietnam for environmental variable Temp. As annotated, the correlation between the fuzzy set and original abundance is $0.62$ and significant as indicated by the significance label. This is a moderate positive correlation, therefore it suggests that small changes in temperature may lead to moderate changes the community structure.

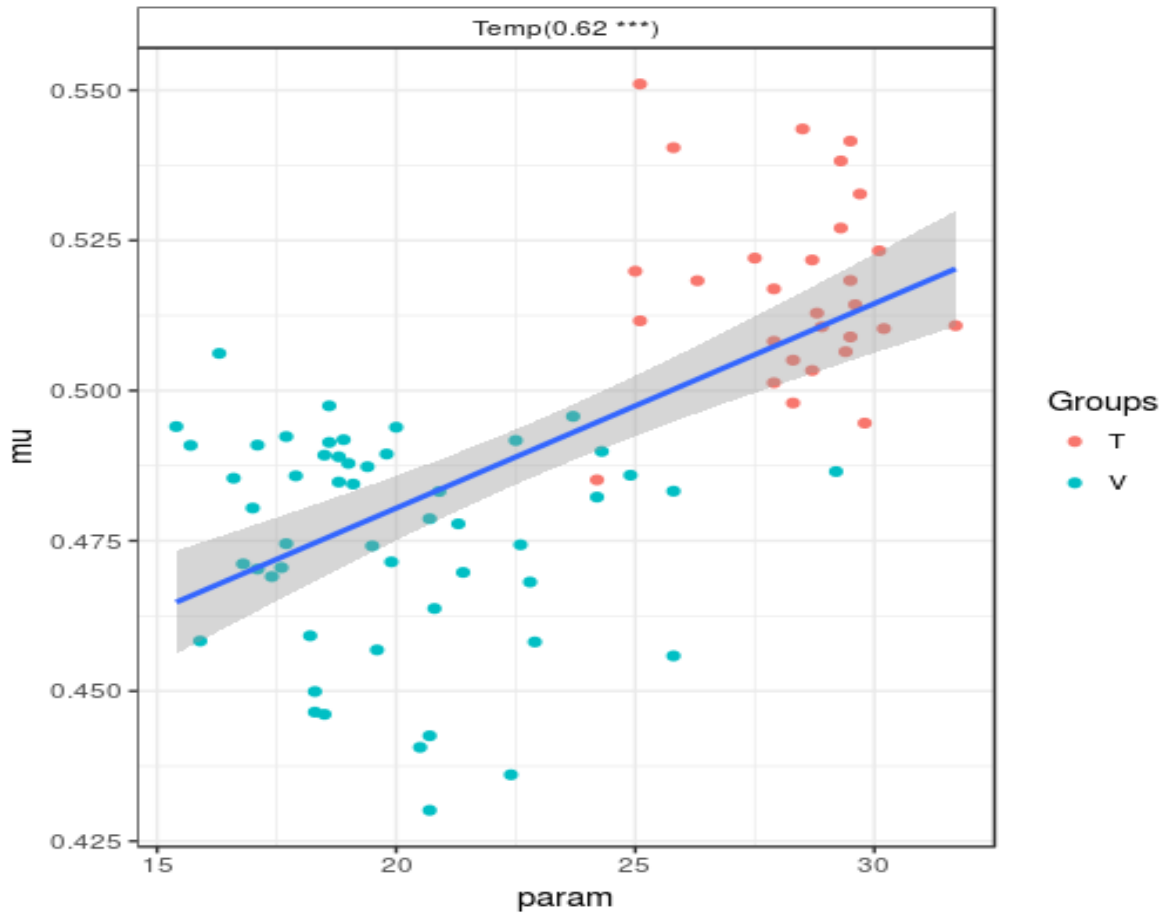


Figure 5: Fuzzy set ordination of samples between countries. Red and Blue are for Tanzania (T) and Vietnam (V).

Generally, a low correlation value is indicative of huge difference between fuzzy sets and original values and a high value shows a smaller difference between fuzzy sets and original values. This implies that the community is very sensitive to variables with a low correlation value and vice versa.

### Local Contribution to Beta diversity

We calculated LCBD by each of the samples at phylum taxonomic level and considered 20 top most taxa for visualisation as shown in Figure 6. Prior to that, abundance data was normalised by relative tansformation to obtain proportions accross samples.

Phyla *Firmicutes*, *Bacteriodes*, *Proteobacteria* and *Actinobateria* show high abundance and consequently contribute highly to LCBD in all the samples from Tanzania and Vietnam. We note that, the proportion of these phyla are clearly different for the two countries. In fact, phyla *Proteobacteria*, *Actinobateria* , *Deinococcus.Thermus* show high proportion of abundance in Vietnam and phyla *Firmicutes*, *Synergistes* whereas *Proteobacteria* show high relative abundance in Tanzania.

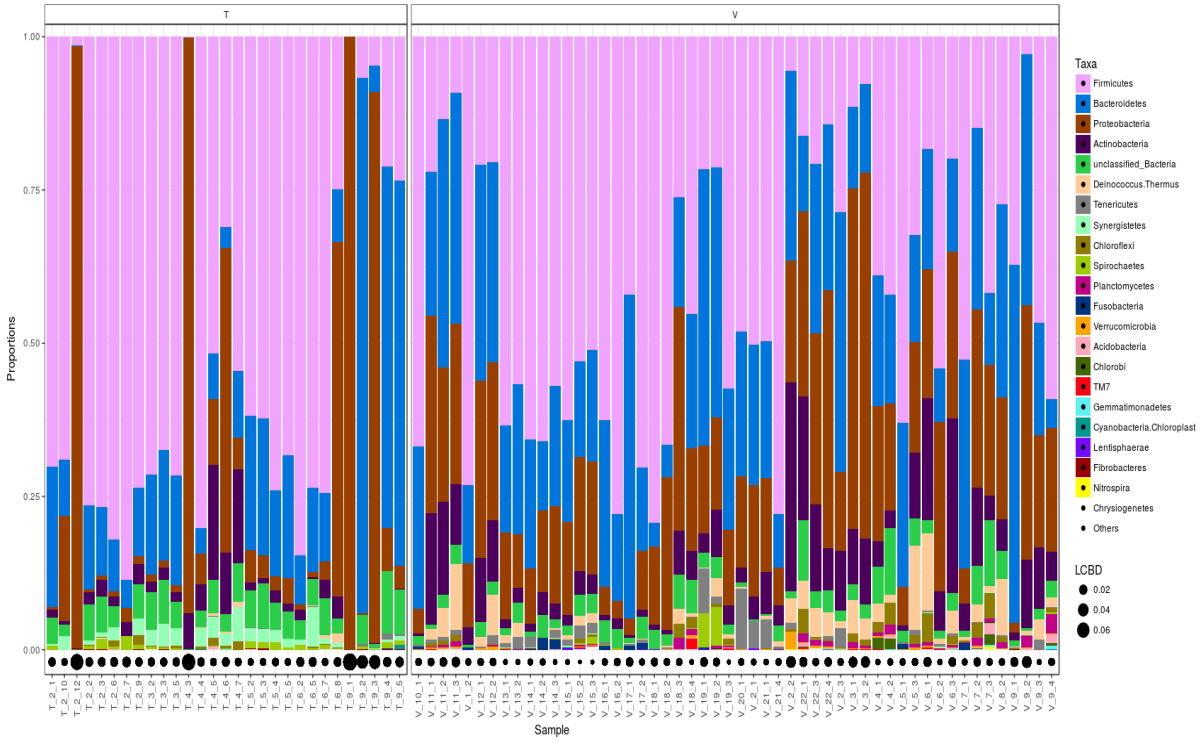


Figure 6: Most abundant taxa and local contribution to beta diversity for Tanzania (T) and Vietnam (V). Black points at the bottom whose diameter corresponds to magnitude of LCBD value coresponding to a particular sample, the bars correspond to taxa that are most abundant with the top taxa sharing a bigger portion of the bar for each sample.

**Differential expression analysis**

Differential expression of taxa at phylum taxonomic level was investigated between Tanzania (T) and Vietnam (V) using DESeq. Significance of differential expression was based on p-value threshold of $5\%$ and zero log2 fold change which are the defaults. The results are visualised in Figure 7. Amongst the significantly up or down regulated phyla,*Synergistes* are the most important, followed by *Proteobacteria*, *Fibrobacteres*, *Actinobacterial* among others. The p-values and rank of importance are annotated on the figure.

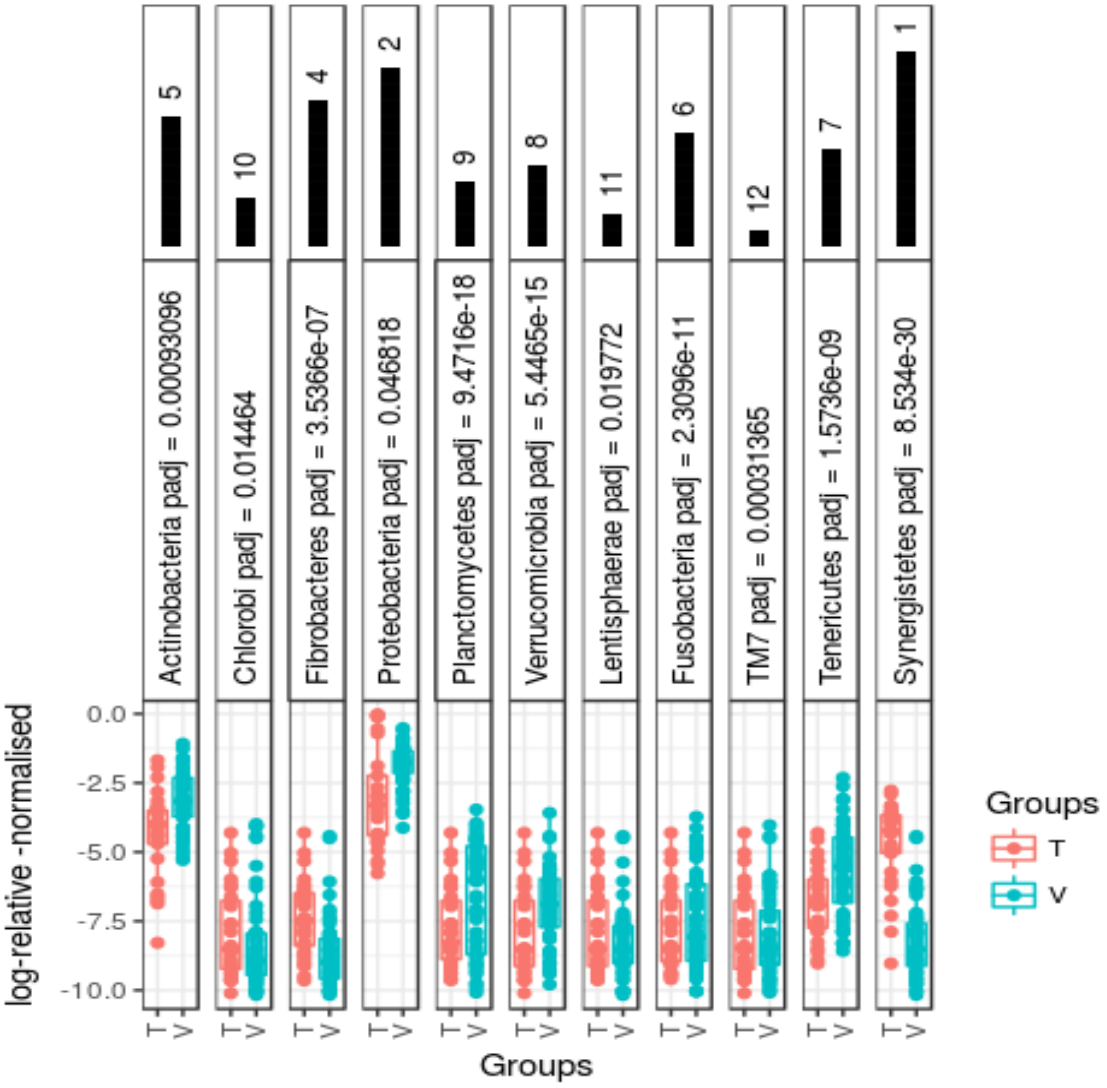


Figure 7: Significantly differentially expressed phyla between Tanzania and Vietnam: The black bars correspond to importance of a corresponding feature and on top of which the ranks based on mean decrese accuracy are indicated. The data was log relative normalised for purposes of output. The middle section of the plot indicates phyla description and corresponding adjusted p-value.

Figure 8 shows the relationship between log2 fold change and mean abundance of the phyla that are differentially expressed between the two countries. The significance depicted here is only as per log2 fold change and not based on p-value. *Synergistes* (mean abundance 145.9) and *Deinococcus-Thermus* (mean abundance 40.5) are respectively the most down and regulated phyla with 6 fold between the two countries.

An alternative visualisation is as shown in Figure 9. Abundance of each phyla is annotated on extremes of log2 fold bars.

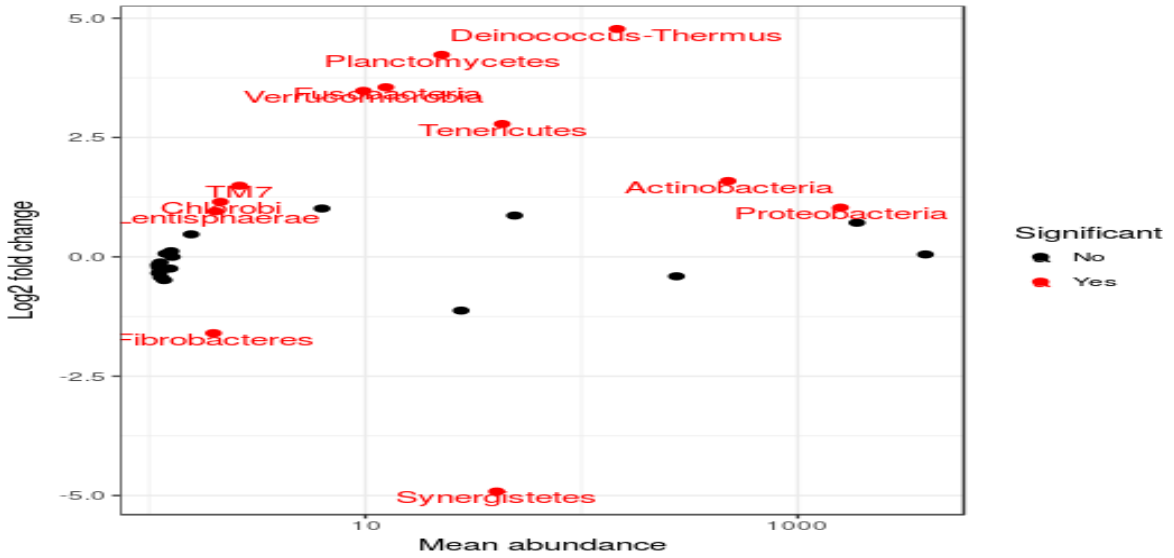


Figure 8: Mean abundance plot: Red and black points respectively correspond to phyla which shows significant differential expression between countries and those that do not.

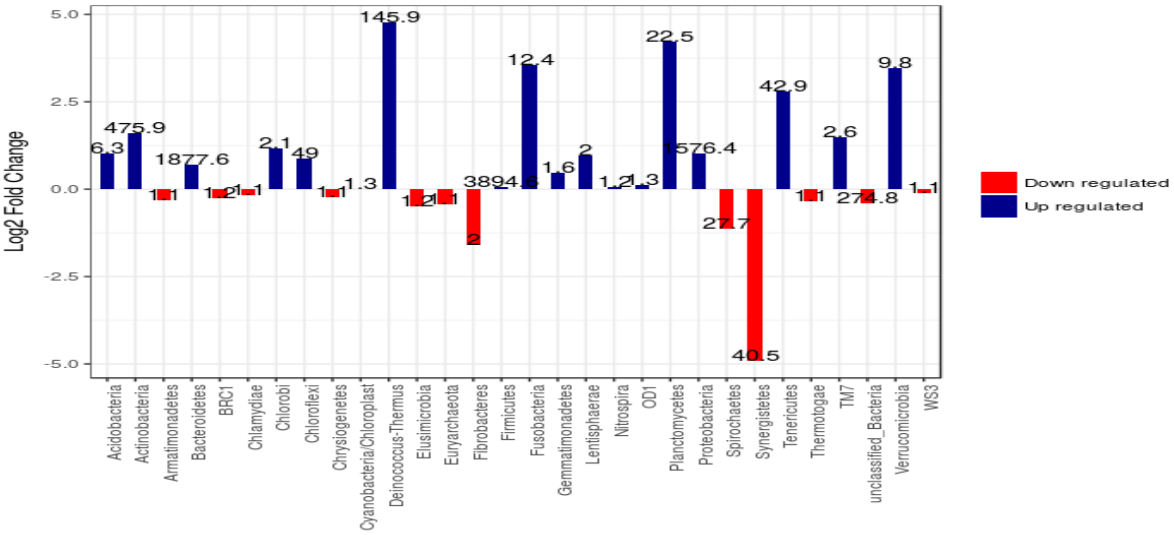


Figure 9: log2 fold change plot: Red and blue shows down and up regulated phyla respectively. Values annotated on bars are mean abundances for corresponding phyla.

Figure 10 shows mean decrease in accuracy for significantly expressed phyla between Tanzania and Vietnam. *Synergistes* are the most important with a mean decrease accuracy of 45 followed by *Proteobacteria* at 25 and *TM7* is least important with close to 5 mean decrease in accuracy. The ranking of phyla depicted in Figure 7 is based on this measure.

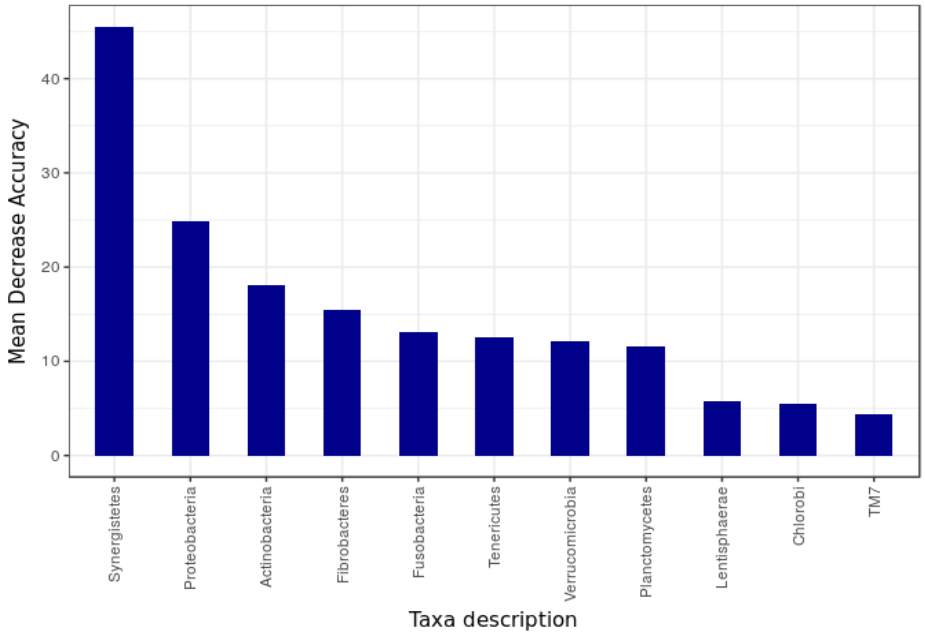


Figure 10: Mean decrease in accuracy of each of the significantly differentially expressed phyla.

Table 1 shows qualitatively the characteristics of differentially expressed features as visualised in the figures above showing the country in which a particular phyla is up regulated.

Table 1: DESeq differential expression results

| Phyla | Base Mean | Log2 Fold Change | P-value | Adjusted-pvalue | Country |
|---|---|---|---|---|---|
| Synergistetes | 40.480500 | -4.9153857 | 2.844673e-31 | 8.534020e-30 | T |
| Deinococcus-Thermus | 145.904957 | 4.7731870 | 1.112263e-23 | 1.668394e-22 | V |
| Planctomycetes | 22.530003 | 4.2313944 | 9.471601e-19 | 9.471601e-18 | V |
| Verrucomicrobia | 9.793022 | 3.4771679 | 7.261986e-16 | 5.446490e-15 | V |
| Fusobacteria | 12.441201 | 3.5526292 | 3.849406e-12 | 2.309644e-11 | V |
| Tenericutes | 42.907190 | 2.7828493 | 3.147201e-10 | 1.573600e-09 | V |
| Fibrobacteres | 1.983949 | -1.5982656 | 8.252045e-08 | 3.536591e-07 | T |
| TM7 | 2.617610 | 1.4901325 | 8.363945e-05 | 3.136480e-04 | V |

Similarly, we used kruskal Wallis to test differential expresion of phyla between Tanzania and Vietnam. Prior to testing, taxa abundance was normalised by log relative transformation. Corresponding plots similar to Figure 7 and Figure 10 were produced. Table 2 shows most significantly differentially expressed phyla with associated statistics.

Table 2: Kruskal-Wallis test results

| Phyla | p-value | E-value | FWER | q-value |
|---|---|---|---|---|
| Synergistetes | 7.436027e-11 | 2.230808e-09 | 2.230808e-09 | 2.230808e-09 |
| Spirochaetes | 3.524440e-07 | 1.057332e-05 | 1.057327e-05 | 5.286660e-06 |
| Deinococcus-Thermus | 3.524440e-07 | 1.057332e-05 | 1.057327e-05 | 3.524440e-06 |
| Fibrobacteres | 4.935991e-05 | 1.480797e-03 | 1.479738e-03 | 3.701993e-04 |
| Proteobacteria | 2.477304e-04 | 7.431913e-03 | 7.405278e-03 | 1.486383e-03 |
| Bacteroidetes | 3.762102e-04 | 1.128631e-02 | 1.122495e-02 | 1.881051e-03 |
| Actinobacteria | 3.762102e-04 | 1.128631e-02 | 1.122495e-02 | 1.612329e-03 |

*Synergistetes*, *Deinococcus-Thermus*, *Fibrobacteres* are among the phyla that are detected by both methods to be highly differentially expressed between countries. Therefore in case of a particular intervention towards decomposition of material would be directed to these in respective countries of up regulation.

A plot for multiple comparison corrections is produced from Kruskal-Wallis test is shown in Figure 11.

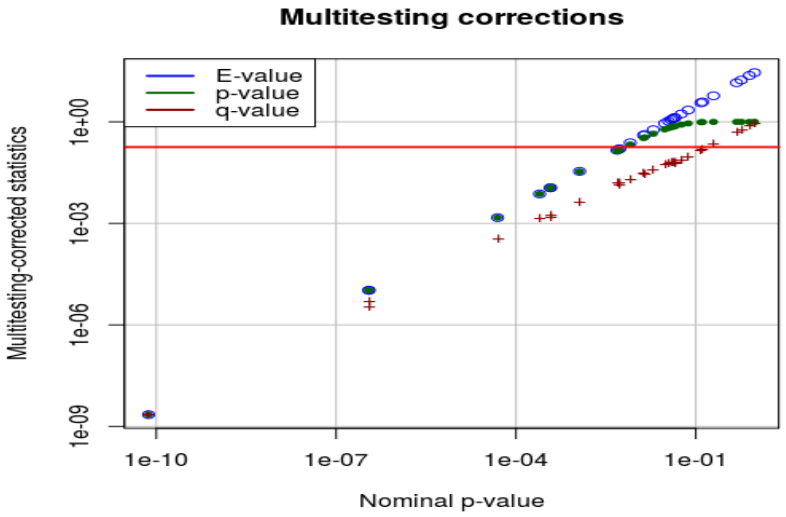


Figure 11: Multiple testing corrections: Red line shows the threshold q-value for significance.

Kernel based differential expression was explored between Tanzania and Vietnam at pyla taxonomic level. Taxa abundance was log-relative transformed prior to analysis, this is because the method is designed for fractional data. The results are visualised as shown in Figure 12. Threshold value for correlation and adjusted p-value were set to $0.99$ and $5\%$ respectively. P-values were adjusted by Benjamin- Hochberg method which is also the default.

At this correlation threshold, *Eukaryota*, *Thermotogue*, *Chlamydiae* and *Armatinonadetes* is a group of phyla found to be differentially expressed between Tanzania and Vietnam with adjusted p-value of $0.08$. This suggests that this set of phyla is also partly responsible for the difference between the microbial communities of the two countries. In addition, it also verifies the results obtained by using DESeq and Kruskal-Wallis, since the top most detected above are also found significantly differentially expressed by this method. Examples include: *Synergistetes* (p-value=2.19e-229), *Deinococcus-Thermus* (p-value=1.34e-47) and *Fibrobacteres* (p-value=4.88e-28).

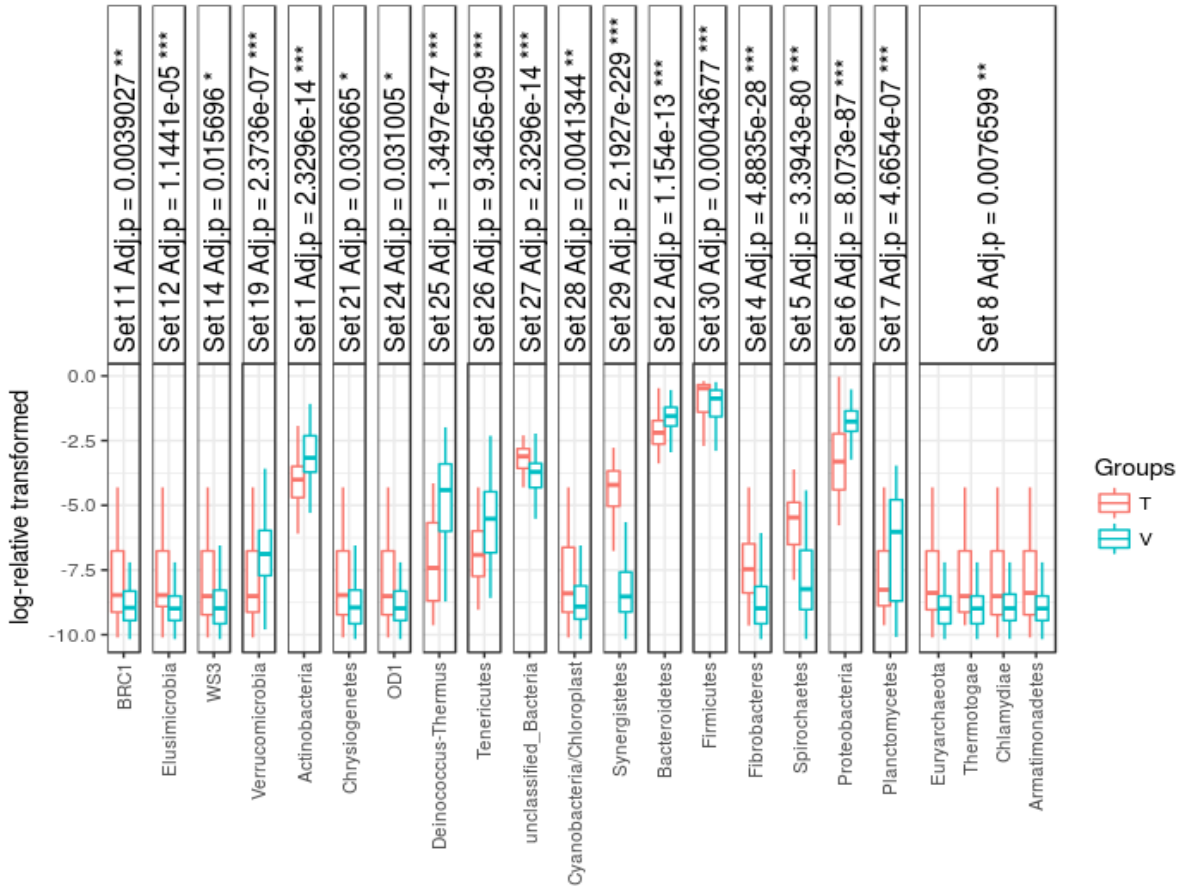


Figure 12: Sets of differentially expressed phyla between Tanzania (T) and Vietnam (V).

Similarly, kernel differential analysis was applied to the numerical environmental variables. The aim of this is to identify groups of environmental traits that are differentially correlated between the two countries at a correlation threshold of $0.09$. We scale normalised the meta data prior to analysis. The results show that individual variables are correlated differently between the countries. In other words, all traits have relationships amongst them and these relationships are different for both countries. However, there is no evidence for any groups of variables showing difference in correlation between the two countries as visualised in Figure 13.

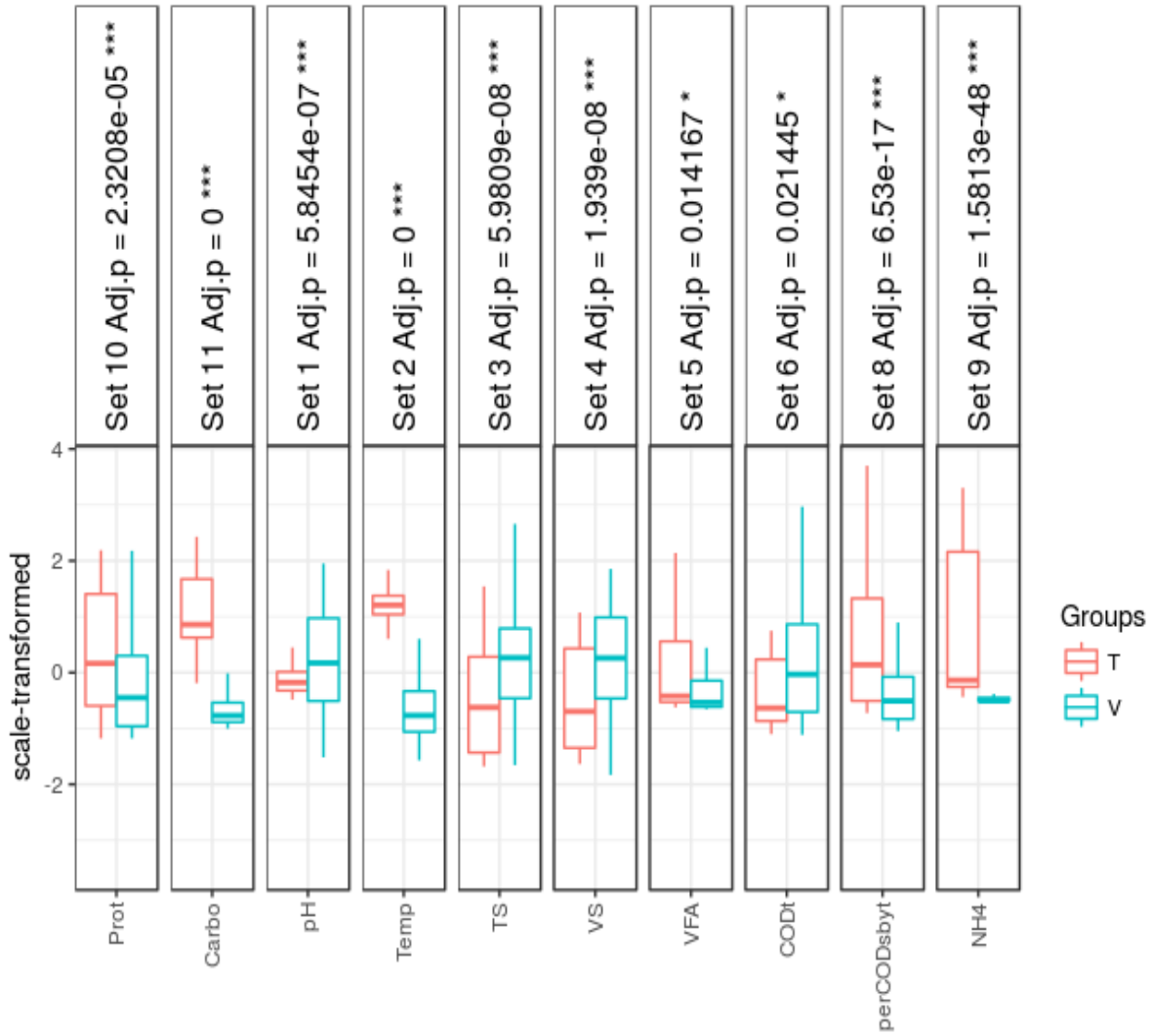


Figure 13: Sets of differentially correlated environmental variables between Tanzania (T) and Vietnam (V).

### Co-occurrence pattern analysis

Co-occurrence within ecosystems is explored between given conditions. In this case, we generated a co-occurrence network for Vietnam at correlation threshold of 0.35 and threshold q-value of 5% at genus taxonomic level. Nodes are coloured as per corresponding sub community. The entire network is shown in Figure 14.

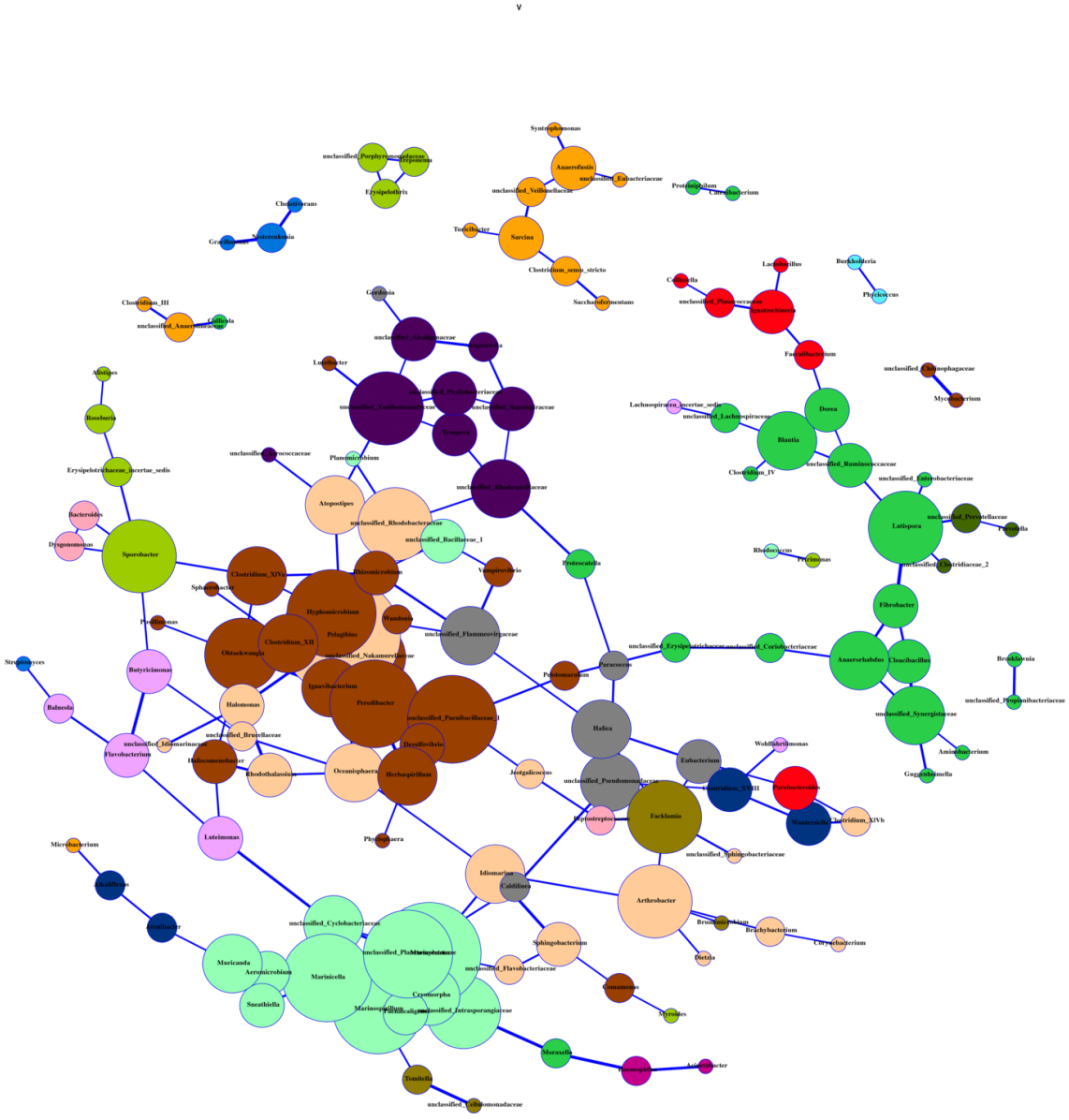


Figure 14: Network showing sub communities. The size of the nodes is proportional to its own total degree. The width of the edges is proportional to the correlation between the two nodes to which it corresponds. Positive and negative correlations between taxa(nodes) are indicated by blue and red colour of the edges respectively.

Figure 15 is a plot of betweenness versus eigenvector centrality at 0.35 correlation threshold for Vietnam. As noted earlier, these are measures of importance of taxa in the network. Betweenness of taxa in this case is a measure of taxa's control in the network. High betweenness centrality implies that a corresponding node has more influence in the network and vice versa. Eigenvector centrality measures taxa's linkage to others in the network taking into account how connected they are.Therefore, taxa with high eigenvector centrality is linked to highly linked taxa. The different colours correspond to a sub community (module) to which a particular taxon belongs.

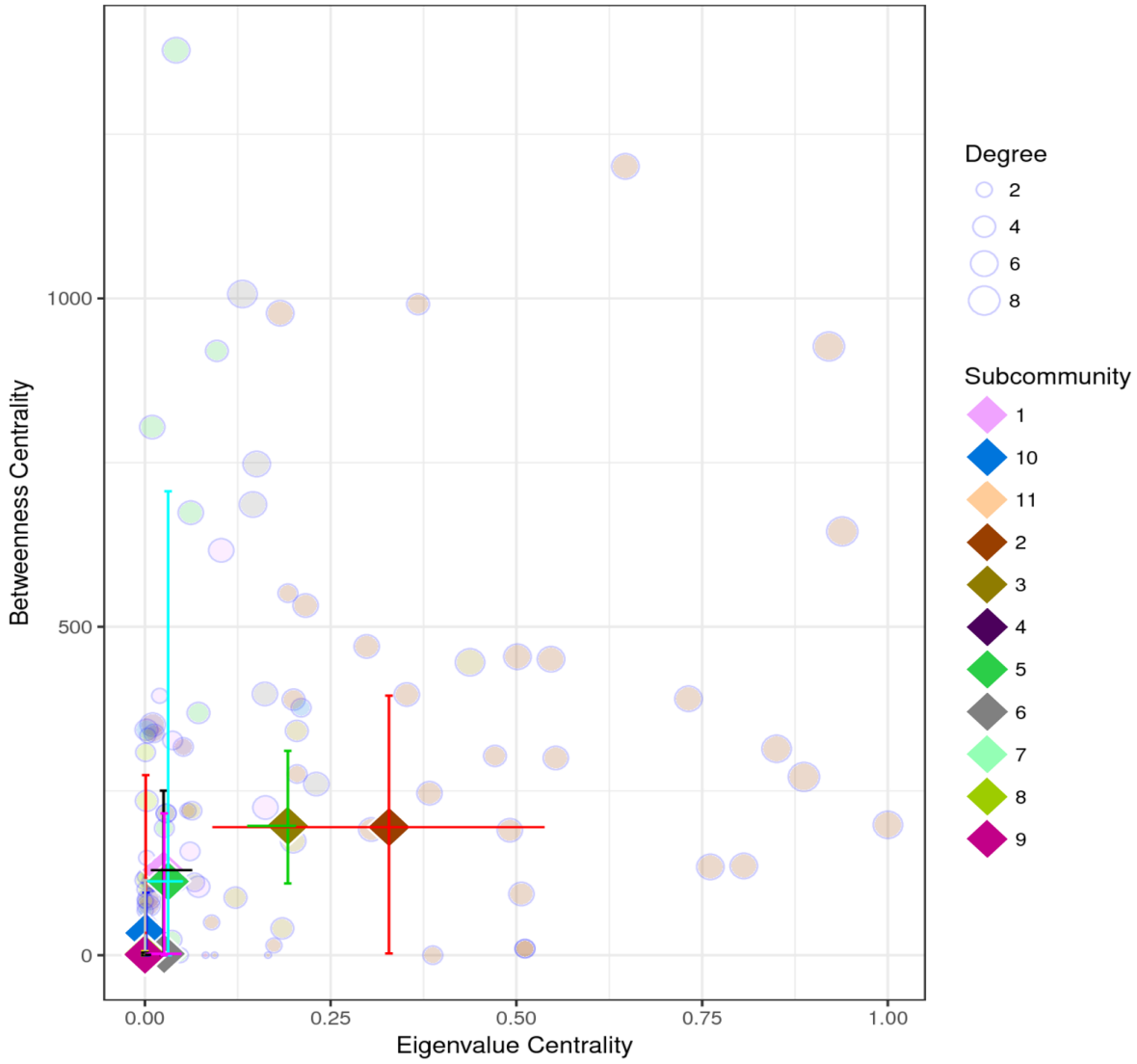


Figure 15: Relationship among network statistics: Betweenness, eigenvector centrality and degree.

Using the network obtained above, we assigned roles to each genus as shown in Figure 16. Most of the genera are identified as being non hub modules specifically being in ultra peripheral

and peripheral zone. This indicates that these genera are most probably involved in processes restricted to corresponding sub communities than across sub communities. There a few non hub connectors and these are ones that link own sub communities to others. This indicates that these genera are might have a wider range of habitat and or resource usage in the community. In addition, they might be involved in essential biochemical processes that cut across neighbouring sub communities. Non Hub connector genera include: Sub community #2 *Herbaspirillum* and *Geminicoccus*, sub community #8 *unclassified-Prevotellaceae*, *Lactobacillus* and *Lactococcus*, sub community #3 *unclassified-Propionibacteriaceae* and *Paludibacter*. Single genus *unclassified-Synergistaceae* of module #3 is the only connector hub.

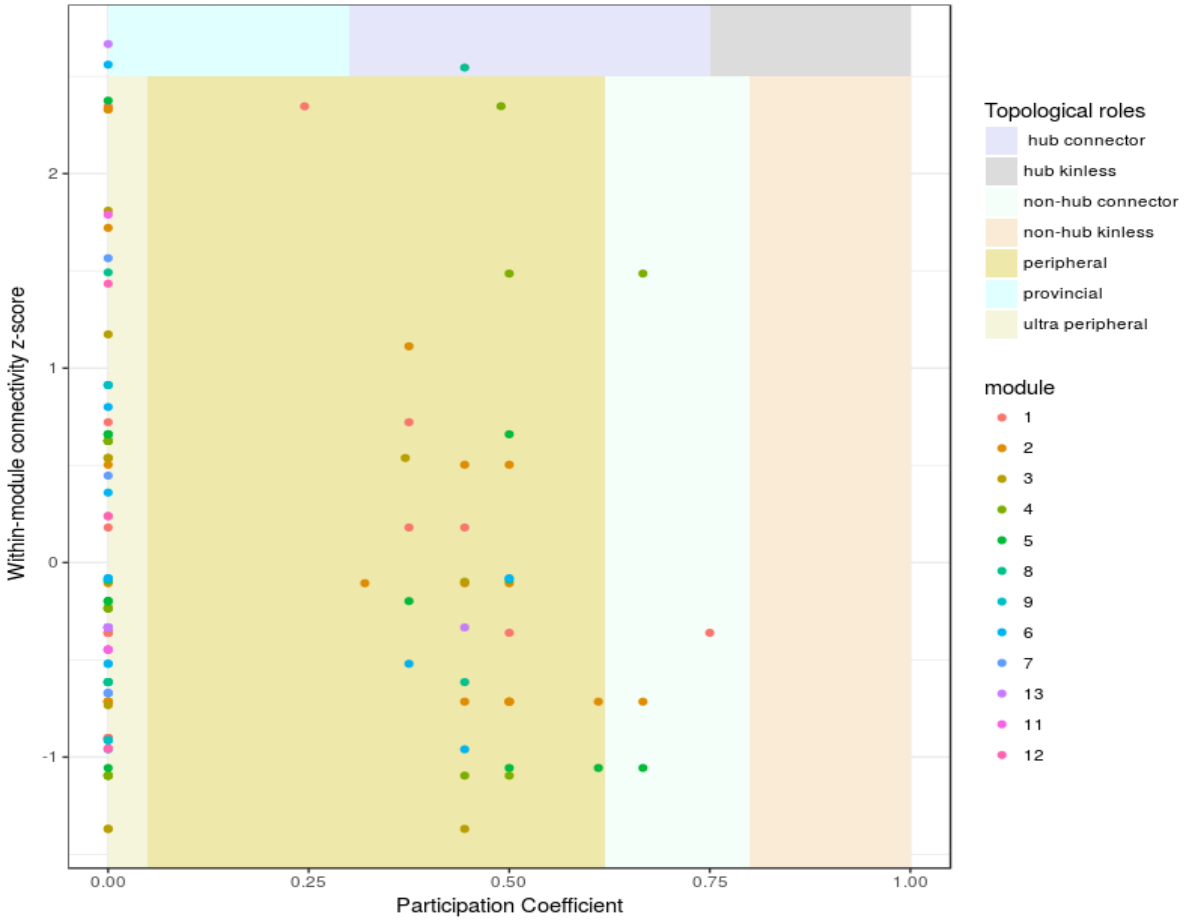


Figure 16: Topological roles of taxa: Based on within module degree (y-axis) and among participation coefficient (x-axis). Different regions correspond to different topological roles as shown in key. Each point is representative of a taxon and its colour signifies the sub community (labelled as module in key) to which it belongs.

In this case, most of the taxa non module hubs. Amongst these, most are ultra peripherals and peripherals with a few non-hub connectors. Therefore most of the taxa are are restricted to

their own sub communities with exception of a few that are in the provincial zone. This is also quite evident from the network as shown in Figure 14, there are not many connections between individual sub communities. We explored the sub communities' response to environmental traits by the correlation between keystone features in each sub community with environmental traits. The results are visualised in Figure 17. This was obtained using Pearson correlation method. Correlation test p-values were adjusted for multiple comparisons using Benjamin-Hochberg method and significance assessed at p-value threshold of $5\%$.

Sub community #2 is significantly positively correlated with Temp but negatively with pH. This indicates that members of this module may be activated and inhibited by increasing levels of temperature and pH respectively. In addition, #12 is significantly positively correlated with all CODs related traits, Prot and VFA but negatively with pH which suggests that, members of this sub community increase in abundance given resources but could be inhibited by increasing levels of pH. #8 is highly positively correlated with NH4 concentration and proteins. The rest of the sub communities are weakly correlated with the environmental traits.

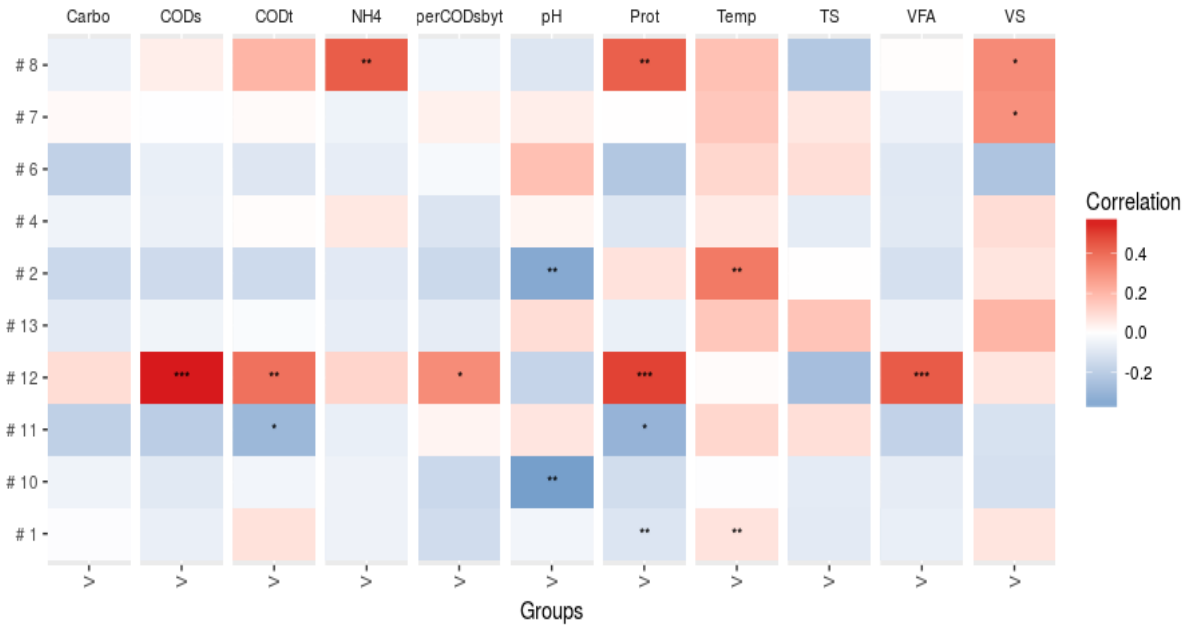


Figure 17: Relationship between sub communities' keystone features and environmental variables for Vietnam (V). Sub communities are represented by #number. Asterisks show level of significance of correlation (*p-value $< 0.05$, **p-value $< 0.01$, ***p-value $< 0.001$). Red, blue and white indicate positive, negative and no correlation respectively.

Table 3 shows genera that are identified as keystone genera for each of the sub communities with their correlation with CODs. As explained in the methods section, these genera are obtained on the criterion than they have the maximum betweenness relative to others in a corresponding sub community.

Table 3: Correlation between sub communities' keystone genera and CODs for Vietnam (V).

| Sub community | Genera | AdjPvalue | Correlation |
|---|---|---|---|
| #1 | Truepera | 0.26 | -0.16 |
| # 12 | Clostridium_IV | 0.004 | 0.40 |
| # 2 | Muricauda | 0.38 | -0.12 |
| # 3 | Gallicola | 0.66 | -0.06 |
| # 4 | Kangiella | 0.60 | -0.07 |
| # 5 | Alistipes | 0.25 | 0.16 |
| # 6 | Alkaliflexus | 0.97 | -0.01 |
| # 7 | Ignatzschineria | 0.55 | 0.08 |
| # 8 | Faecalibacterium | 0.01 | 0.36 |
| # 9 | Patulibacter | 0.67 | 0.06 |

We investigated the relationship between 50 most abundant phyla and numerical environmental variables based on Pearson rank coefficient in Tanzania (T) and Vietnam (V). P-values arising from the correlation test are adjusted for multiple testing using Benjamin-Hochberg and significance based on a threshold of $5\%$. The visualisation of this is shown in Figure 18.

Generally, phyla abundance is negatively correlated with the environmental variables in both countries with exception of a few instances. The observed relationship suggests that microbial biodiversity is reduced as more resources that is carbohydrates (carbo) and proteins (prot) are available. The deviations from the general pattern include: phyla such as *Firmicutes* and *Fusobacteria* which show a significant positive correlation between their abundance and traits with exception of Temp, pH and VS particularly in Vietnam.

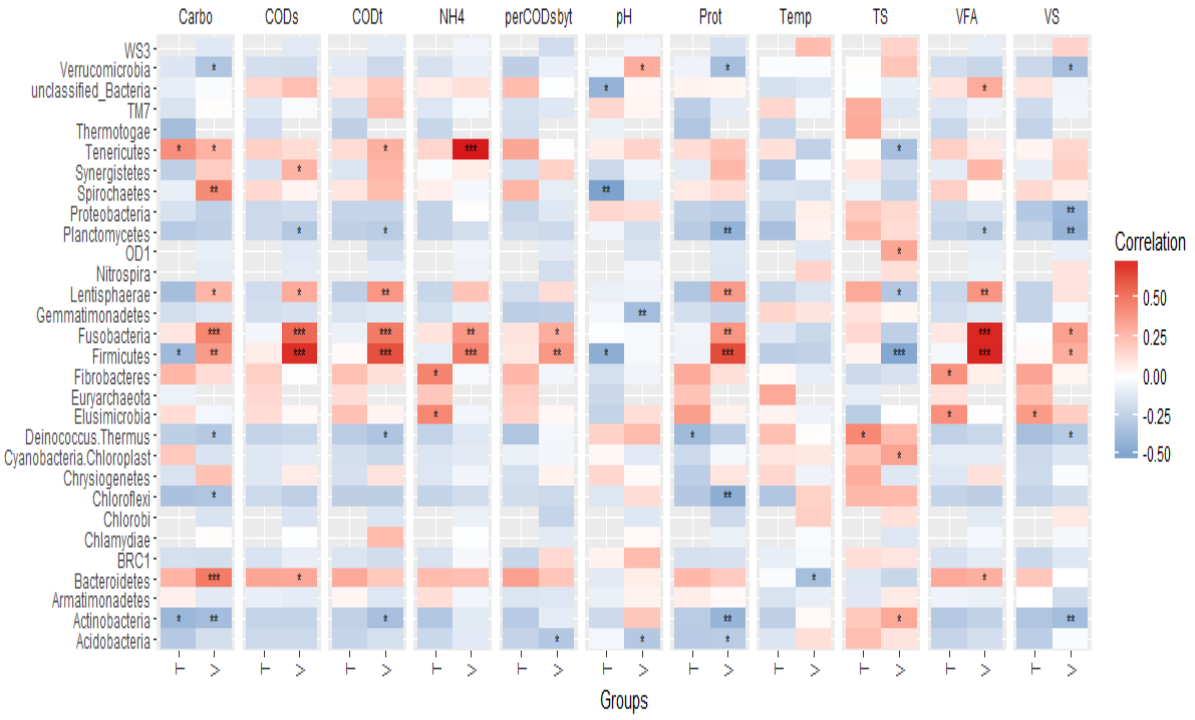


Figure 18: Relationship between most abundant phyla and environmental traits. T and V correspond to Tanzania and Vietnam respectively. Asterisks show level of significance of correlation (*p-value $< 0.05$, **p-value $< 0.01$, ***p-value $< 0.001$). Red, blue and white indicate positive, negative and no correlation respectively.

## 0.8 Discussion

The main goal of this project was to create functions that enhance analysis and visualisation of Amplicons sequenced microbial community data building upon existing tools and integrated into an R package. This was successfully accomplished and extensively tested on a real dataset.

The tools depend on a number of other packages which should be imported in order for the functions to work well. In addition, some methods require that data is transformed to some kind of type, this is crucial to avoiding false results.

Results are presented in form of files containing information which can be used for further analysis and visualisations which can be further manipulated by the user. As much as the visualisations present useful information, it can not all fit in a single figure, therefore it is worth exploring the files generated to develop more insights about the details of analysis results.

Some of the procedures for example kernel differential analysis take some time to run most especially when applied to abundance data, same applies to co-occurrence analysis procedures. Therefore, depending on the size of the dataset and/or taxonomic level at which the analysis is conducted, the execution time varies accordingly.

The evaluation of the tools on a real dataset does suggest that all parameters most especially those that relate with user set thresholds are highly critical to results. As such it is recommended that, these are set from an informed perspective.

It is important to note that for most of the methods implemented have a thing to do with comparing among or between conditions. Depending on the procedure being used, the specified groups should be having a reasonable number of observations, other wise, there may be higher chances of crashing. For this particular iteration of the package, a few warning points have been put in most especially for verifying input parameters. This will be improved in later releases.

The package has been entirely developed and tested on a ubuntu 16.04 system with a singe dataset. Therefore, it is important to test it on more datasets and accross operating systems including windows and mac to ensure that it is compatible to all the systems.

## 0.9 Conclusions and future work

Having developed microbiomeSeq package with goals limited to this project, it leaves too much room for further improvement. Currently the functions available allow exploring community diversity within and across samples, differential expression analysis of taxa between conditions, co-occurrence pattern analysis at community level and relationships between micro-organisms and their environment.

Since we are talking microbial community, it would be great to implement a procedure for analysing results from whole genome short gun sequencing such that we can explore functional patterns within the communities in relation to their composition.

Further options can be explored and incorporated into the respective main features of the package to allow more flexibility and also provide options for the users. In addition, since, some of the functions require a lot of time, therefore, in future work can be done to optimize the implementation more to allow faster computation.

# Bibliography

M. J. Anderson, K. E. Ellingsen, and B. H. McArdle. Multivariate dispersion as a measure of beta diversity. *Ecology letters*, 9(6):683–693, 2006.

B. Auguie. gridextra: Miscellaneous functions for grid graphics. 2016. URL `https://CRAN.R-project.org/package=gridExtra`. R package version 2.2.1.

B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, 13 (7):581–583, 2016.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

R. D'Amore, U. Z. Ijaz, M. Schirmer, J. G. Kenny, R. Gregory, A. C. Darby, M. Shakya, M. Podar, C. Quince, and N. Hall. A comprehensive benchmarking study of protocols and sequencing platforms for 16s rrna community profiling. *BMC genomics*, 17(1):55, 2016.

R. Edgar. Usearch: ultra-fast sequence analysis, 2015.

R. C. Edgar. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature methods*, 10(10):996–998, 2013.

A. M. Eren, G. G. Borisy, S. M. Huse, and J. L. M. Welch. Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences*, 111(28):E2875–E2884, 2014.

R. Guimera and L. A. N. Amaral. Functional cartography of complex metabolic networks. *nature*, 433(7028):895, 2005.

J. Kuczynski, J. Stombaugh, W. A. Walters, A. González, J. G. Caporaso, and R. Knight. Using qiime to analyze 16s rrna gene sequences from microbial communities. *Current protocols in microbiology*, pages 1E–5, 2012.

P. Langfelder and S. Horvath. Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11):1–17, 2012. URL `http://www.jstatsoft.org/v46/i11/`.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

P. J. McMurdie and S. Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4):e61217, 2013.

X. C. Morgan and C. Huttenhower. Human microbiome analysis. *PLoS computational biology*, 8(12):e1002808, 2012.

J. Oksanen, R. Kindt, P. Legendre, B. O'Hara, M. H. H. Stevens, M. J. Oksanen, and M. Suggests. The vegan package. *Community ecology package*, 10:631–637, 2007.

A. Oulas, C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis, and I. Iliopoulos. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology insights*, 9:75, 2015.

B. Torondel, J. H. Ensink, O. Gundogdu, U. Z. Ijaz, J. Parkhill, F. Abdelahi, V.-A. Nguyen, S. Sudgen, W. Gibson, A. W. Walker, et al. Assessment of the influence of intrinsic environmental and geographical factors on the bacterial ecology of pit latrines. *Microbial biotechnology*, 9(2):209–223, 2016.

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

R. J. Williams, A. Howe, and K. S. Hofmockel. Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. *Frontiers in microbiology*, 5, 2014.

X. Zhan and D. Ghosh. Kmda: Kernel-based metabolite differential analysis. 2015. URL https://CRAN.R-project.org/package=KMDA. R package version 1.0.