UNIVERSITY OF PADOVA

Master Degree in Bioengineering

Diversity indices and normalization approaches in microbiome studies

Supervisor: Prof. Barbara Di Camillo

Co-supervisor: Dott. Francesca Finotello

Student: Eleonora Mastrorilli

Registration Number 1056704

December 9th, 2014

INTRODUCTION

We are used to think of humans as individuals, but we actually are ecosystems. We all carry upon and inside our bodies an incredibly complex microbial population that helps us digest, synthetize useful components and fend off pathogens. Indeed, this huge community of bacteria, which are mainly symbiont, cooperates with us in maintaining a dynamic equilibrium with the world surrounding us. Although often neglected so far, recent developments revealed that the delicate balance between us and our bacterial community may play a great role in defining our health status. Moreover, several studies have already proven that a certain degree of correlation exists between alteration of this micro flora and severe human diseases, like cancer, BPCO and bowl inflammatory diseases. To understand how this is possible, it suffices to think that only one in ten cells that populates our body is really "human", while the remaining nine actually are microbial cells.

Studies evaluating the composition of this microbial community, trying to understand its interaction patterns, as well as many other issues, are all considered to be addressing the *human microbiota*. The term microbiota is used to refer to the bacterial population considered as a whole, while its genomic content is in turn called the human *microbiome*. Indeed, we perform these exploratory analysis by means of new DNA sequencing technologies, whose capability of sequencing up to hundreds of millions of DNA fragments in a single run allows us to study the microbiota in its own environment: the human body.

However, in order to control such a powerful tool, we first have to assess a standardized methodology to guide us towards trustworthy and truthful results. The aim of this thesis is therefore to understand a typical microbiome analysis pipeline, pointing out where a detailed examination of the available methods still has not been carried out and what the impact of such disambiguation might be. In details, we will give a brief introduction on what microbiome is, why and how it is investigated in Chapter 1, also providing some examples of published researches on the subject. In Chapter 2 we will give an overview of the available ecological measures used to assess biodiversity and we will evaluate their adequacy to microbiome studies' needs by means of a simulation. In the third chapter we will investigate why normalization approaches are particularly needed when dealing with microbiome dataset and what are the possible methods to be used; moreover, we will propose a new method that addresses data sparsity. Again we will examine the impacts these approaches have on the obtained outcomes, by testing them on a simulated datasets. Chapter 4, lastly, will describe how we structured our microbiome simulation, starting from real data analysis.

Microbiome exploration could really give us a new perspective on human health and disease, providing us with a novel instrument to investigate and monitor how genetic and environmental factors impact on our physical condition. Great potential lies in its analysis, with the possibility for new medical treatments approaches to develop, possibly beneficial both for us and our symbiotic microbes. Genomics and bioinformatics tools, therefore, play a major role in exploring it, trying to understand if particular alterations in the microbiota could be markers, targets, or even causes, of some particular diseases.

INDEX

CHAPTER 1:

WHAT IS MICROBIOME?

In recent years increasing attention has been devoted to explore the bacteria that inhabit our world, making use of technologies that allows us to observe them directly in their environment. Microbes are the most numerous and diverse kingdom in nature: indeed, the majority of the Earth's biomass is microbial. We usually relate microbes to infections and diseases, while their major role is quite the opposite: they play a part in maintaining the environmental equilibrium (e.g. in the carbon cycle) and, even when related to human, they often contribute to our health.

Although many studies have been developed to describe microbial community in ecology context, from salty lakes to deep sea, great potential lies in the analysis of microbial community that lives in contact with human body, the so-called *Human Microbiome*. Exploring and characterizing it could give us unprecedented information to understand its structure, function and role in human health and disease.

In this chapter, we will give a brief overview of what human microbiome is and why we study it; then we will describe how we explore it and, finally, we will report two recent studies to convey the possibilities these techniques give us for future developments.

1 OVERVIEW

As Joshua Lederberg first defined, microbiota is "*the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space*" [1], while we refer to microbiome as the collective genome that we can extract form it. Indeed, the human body really is an ecosystem, composed by an incredible number of different kinds of microbes, living inside and outside our bodies.

To convey an idea of the impact human microbiome might play in our lives, we can look it from two different perspectives. First of all, quantitatively, the human body contains over 10 times more microbial cells than human cells: so, if each of us consists of about 10 trillion human cells, we have as many as 100 trillion microbial cells on and inside of our bodies. Secondly, from a functional point of view, we know that our DNA-and particularly our genes-encode the functions that characterize us. Each of us has about 20,000 human genes, but again we have between 2 and 20 million microbial genes, helping us digest, fend off pathogens and react to external agents. Our microbiota does indeed carry out a number of metabolic reactions not encoded in our genome that are necessary for human health.

Many studies have targeted microbiota and its correlation with several diseases, from auto-immune pathologies to diabetes [2] [3] [4] and interest in understanding his part in human health is growing worldwide. Actually a collaborative worldwide project has been developed, the Common Fund's Human Microbiome Project (HMP) program, aimed at developing tools and dataset for the scientific community to explore the role of microbes in human health and disease, as we will briefly describe in the following section.

Traditionally, microorganisms have been studied by direct observation of cultures upon Petri dishes in the laboratory. Unfortunately, most of the microbial species composing human microbiota have never been successfully isolated in the laboratory, typically due to the inability to reproduce necessary growth conditions in the lab. Notably, microbes that grow well in culture may not be the most important nor the most abundant in a particular habitat (e.g. the well-known bacterium *Escherichia Coli*). This discrepancy between the number of bacteria directly observable from environmental samples and the number of cultured bacteria was defined “the great plate count anomaly”, by Staley and Konopka in 1984 [5] [6].

Advances in DNA sequencing technologies have allowed many new fields of research to flourish, one of which, called metagenomics, studies samples of genetic material recovered directly from their natural habitats. These culture independent techniques allowed to begin the exploration of microbial communities independently of bacterial cultivation, thus granting the possibility to detect even bacterial strains that went so far undercover because of culture limitations.

Most of the microbes our microbiota is composed by are difficult to grow on culture in laboratory. Therefore only the new techniques of DNA sequencing have allowed us to investigate microbiome by observing its composition in its own environment. Two different approaches are currently used to tackle this problem: marker gene targeted sequencing (using 16S rRNA) and whole genome shotgun sequencing. Although very (both from a technical point of view as well as for the actual throughput produced), these two techniques follow a pretty similar analysis pattern. Therefore, since target sequencing has a longer history of published studies and freely available databases and tools, in this thesis we will focus on this technique only, starting from describing what kind of gene 16S is and how it is commonly used to determine bacterial taxonomy [7].

In the next paragraphs we will describe briefly how the Human Microbiome Project developed, succeeding in using Next Generation Sequencing (NGS) on human microbiota's DNA samples to investigate microbial richness and structure. We will also describe a typical analysis pipeline used to explore microbiome samples, from sequencing steps to data analysis.

2 The Human Microbiome Project

The NIH Common Fund Human Microbiome Project (HMP) was founded with the primary goal of generating research resources that could aid comprehensive description of the human microbiota and analysis of its role in human health and disease [8] [9] [10].

> *"The NIH Human Microbiome Project is one of several international efforts designed to take advantage of large scale, high through multi 'omics analyses to study the microbiome in human health. As a community resource program, the HMP is a partner in an international collaboration to generate rich, comprehensive, and publicly available datasets of the microbiome. This information will be available worldwide for use by investigators and others in efforts to understand and improve human health."* [11]

The first phase of HMP (developed during 2007-2012) characterized the composition and diversity of microbial communities placed in the major mucosal surfaces of the human body, also evaluating the genetic and metabolic potential of these communities. The current phase of HMP (started in 2013 and ongoing until 2015) intend to create the first dataset integrating biological properties from both the microbiome and host from cohort studies of microbiome-associated diseases.

In the HMP new sequencing technologies for culture-independent microbiome analysis complements genetic analyses of existing reference strains, providing an incredibly high quantity of data about the complexity of human-associated microbiota.

Several goals have been set by this initiative:

- determining if there exist a core human microbiome;
- exploring the relation between changes in the human microbiome composition and disease conditions;
- developing a repository of sequenced high-quality reference genomes as well as new computational analysis tools;
- performing a complete characterization of the human microbiome.

The HMP cohort has targeted the microbial communities of 242 healthy volunteers, men and women, between the ages of 18 and 40, by sampling it from 15 or 18 body sites. Samples were collected non-invasively from 5 major sampling spots: oral cavity, nasal cavity, skin, gastrointestinal tract and urogenital tract [12]. Data were analyzed by both targeting 16S gene (elder analyses) and using a whole genome shotgun approach (more recent studies). The current reference database is planning to sequence up to 3000 genomes from both cultured and uncultured bacteria, to provide a comprehensive pool of high-quality sequences to be used in the analysis of human microbiome data.

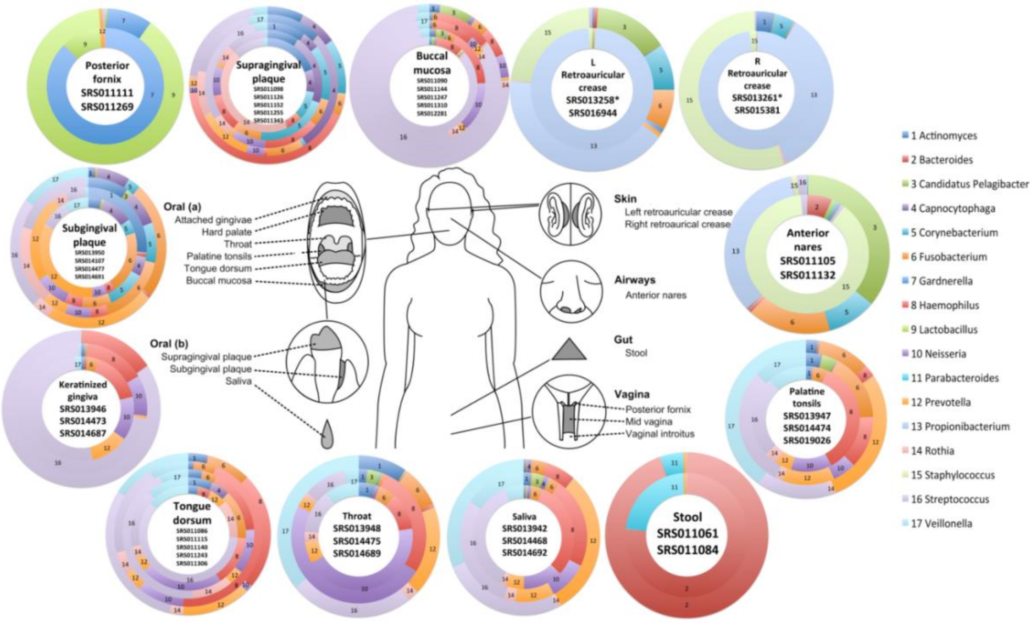


FIGURE 1 AN OVERVIEW OF THE HMP SAMPLING SPOTS AND THEIR MEAN POPULATION **[13]**

3 THE 16S RIBOSOMAL GENE

The 16S rRNA is a sequence, found in all organisms, that partly composes the 30S small subunit of prokaryotic ribosomes. This ribosomal gene has been long used to help identify taxonomic groups found in a sample, since the first studies from C. Woese and G. E. Fox in 1977 [6].

16S is composed by several regions, as depicted in Fig. 3 (here an example of 16S from *E. Coli*). Some of these regions are highly conserved, and are thus called *constant regions*: they are common to all organisms and allow to distinguish and detect 16S from the rest of the genome. Therefore, when 16S is used as a marker gene to be sequenced, common constant regions are targeted as primer binding sites. Other regions, defined *hypervariable regions*, are nine in number and are identified with symbols V1 through V9. These regions are used to infer taxonomic identity of organisms according to phylogeny, making 16S rRNA sequencing a keystone for metagenomic analysis. Actually, due to the rates of evolution of this regions, it is possible to infer taxonomy among bacterial strains by identification of species-specific signature sequences useful for bacterial classification. The amount of sequence difference between different organisms gives indeed a proxy of the amount of evolution that taxonomically separates the organisms.

Therefore, targeted 16S sequencing allows to discover what kinds of microbes live in different samples, simply by processing an environmental sample DNA. If all the ribosomal RNA genes in the sample are isolated, their sequences can be determined using multiplexed NGS. Sequences are then compared to collections of known sequences, stored in publicly available database to identify the microbes found in the original sample.

There are three main database used to determine RNA sequences:

- the *Ribosomal Database Project* (RDP) is a curated database that offers both ribosome data and analysis tools, including phylogenetical alignment of rRNA sequences and phylogenetic trees;
- the *Greengenes* is a web application providing access to the 16S rRNA gene sequence alignment, helping users to annotate sequences;
- the *Silva* database is an on-line resource for high quality alignment of ribosomal RNA sequence data to reference small and large subunits of rRNA [14].

4 Next Generation Sequencing (NGS)

The term *Next Generation* Sequencing refers to a plethora of technologies, developed in the last decades, characterized by an incredible decrease in sequencing cost and increase in sequencing speed compared to traditional methods like Sanger sequencing. These high-throughput techniques have allowed many new research field to develop, like whole genome sequencing and RNA-sequencing. However, the advantages of NGS are balanced by shorter read-lengths and lower accuracy (increased error profiles), therefore an accurate evaluation of both their potentiality and drawbacks is an important consideration to be done. In this section we will focus our attention on two technologies, the Roche/454 GS FLX Titanium sequencer and the Illumina Genome Analyzer [15] [16], since these are the two most widespread instrument used to develop microbiome analysis. We will describe in details how sequencing is performed using these two platforms, once sample collection and DNA extraction have been carried out.

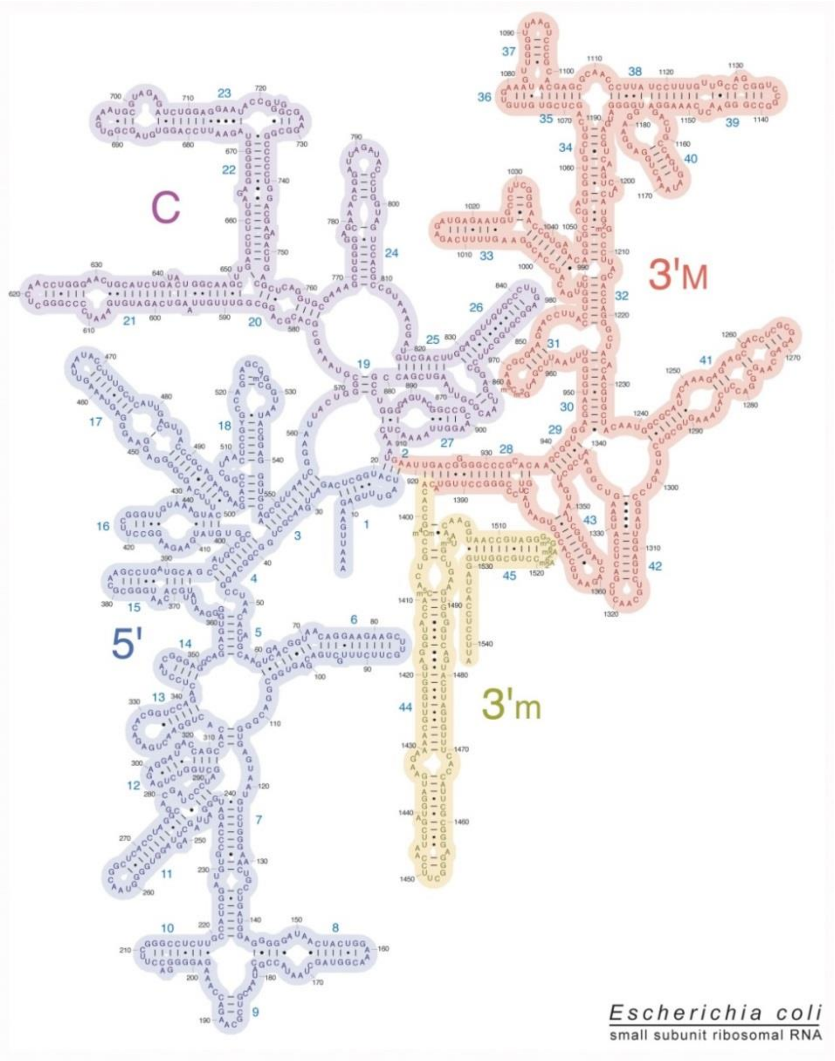


FIGURE 2 E.COLI 16S RRNA **[17]**

4.1 Roche/454 GS FLX Titanium sequencer

Released in 2005, this was the first high-throughput sequencing platform available. Its technology combines emulsion-PCR amplification of the fragments followed by *pyrosequencing*, a sequencing-by-synthesis approach that reads the nucleotide sequence simultaneously as the sequence extension proceeds.

First of all sequencing libraries are created, by adding two adaptors to each DNA fragment, one at the 3' end and the other at the 5' end of the molecule. These adaptors are really pre-synthetized oligos, one of which contains a specific sequence complementary to the oligonucleotides bound on 28-μm sequencing beads. Therefore, hybridization is possible between molecules and beads at a very low molecule-to-bead ratio, so that the probability that each bead captures more than one fragment is minimized.

Then, emulsion-PCR is performed by capturing the beads in an oily emulsion that keeps each bead separated. Within each droplet, PCR (i.e. Polymerase Chain Reaction) amplifies the DNA fragment, so that, at the end of the amplification process, each bead is covered by thousands of copies of the starting molecule. During PCR DNA is repeatedly denatured, primer annealed and copied using known primer sequences, DNA polymerase and nucleotides.

When emulsion is broken, each bead (carrying million copies of a single DNA template) is captured using the second adapter and deposited on a micro-fabricated array (picotiterplate) containing up to $2 \cdot 10^6$, 44μm well. Each bead fills a well and smaller beads, bearing useful enzymes like ATP sulfurylases and luciferases, are added to the plate before the reaction begins. The flow cell is exposed to a CCD camera for signal detection and to a stream carrying the nucleotides to be added in a fixed sequence.

During the pyrosequencing phase, one nucleotide at time is washed over the flow cell, so that DNA polymerase can incorporate it when complementary to the template sequence. When an incorporation event occurs, a phyrophosphate is released as a side product, which in turn can react with ATPsulfurylase to synthetize ATP. ATP then reacts with luciferase emitting light signal, which is measured by the CCD camera. Conversely, if no incorporation occurs, no signal is emitted and the exceeding nucleotides are removed by the aphyrase enzyme or are simply washed away.

This reaction allows the camera to sense light spot only in the position correspondent to a well in which an incorporation has occurred. Therefore, by combining the information from the sequence of nucleotides repeatedly washed on the array and from the emitted signal, a flow gram is produced, from which the sequencer is able to read the complementary sequence of the template investigated. The signal revealed is indeed proportional to the number of incorporation, at least for homopolimers sequences shorter than 8-mers.

The 454/Roche sequencer allows for up to 800nt long read sequences, with a mean throughput of 750Mb/day. However the possibility for long homopolimers to happen causes a high rate of insertion/deletion and possible interference with neighboring well's signal. Moreover, the error rate usually increases with the read length because of a reduction in both enzyme quantity and efficiency.

4.2 Illumina Genome Analyzer

This platform combines bridge-PCR amplification of the reference fragment with polymerase-based sequencing using reversible terminator technology. Indeed, similarly to Sanger sequencing, the incorporation reaction is stopped after each base ligation and the base calling is obtained thanks to fluorescent dyed labels incorporated in each base.

Again the first step is creating the sequencing library adding two different adapters to both ends of the targeted fragment. Both the adapters are also tethered to a solid substrate, the flow cell, therefore it is possible for single stranded fragments to bind to the surface through complementary hybridization. Since a very low concentration of single stranded reference DNA is pumped on the flow cell, they will most probably attach to the surface far from each other: this will allow in the next steps to obtain clusters of identical copies of the starting fragment, ensuring strong and clear sequencing signal.

Adding nucleotides and the required enzymes for PCR, bridge amplification can take place. The single stranded DNA molecule bends, so that its second adapter can hybridize with its complementary one, attached to the flow cell. In this way reverse strand can be synthetized, starting from the double-stranded adapters, thus creating the new strand covalently bound to the surface of the cell. If the double strand is denatured again, the single stranded fragments bend again, and another covalently bound reverse strand can be synthetized. By repeating this process of bending and reverse strand synthesis, the so called bridge amplification, clusters up to 1000 clonal amplicons are generated very closely on the array. Before proceeding with the sequencing step, one of the two strands population has to be cleaved from the cluster, in order to avoid base calling conflicts.

After cluster generation, amplicons are all single-stranded, identically oriented copies of the target molecule, that can be sequenced by annealing primers to adapter oligos. This polymerase-based sequencing uses reversible terminator chemistry in order to ensure single-base extension of the reference. Indeed, the nucleotides available for the extension are modified by a chemically cleavable blockage at the 3' position, preventing the DNA polymerase to incorporate more than one base at the time.

Moreover, each base is laballed with four different fluorescent markers that identify the nucleotide at each cycle. Indeed, an imaging phase follows the incorporation, in which laser excitation stimulates the marker to emit light signal, that can be read by a camera. In this way it is possible to add all four the nucleotides at each cycle, because automatic software will call the base using the specific fluorochrome. After base calling, it is possible to remove both the fluorophore and the blockage.

This platform is characterized by a huge throughput, up to 5000Mb/day at less than one dollar per Mb. On the other hand, its read length is quite poor, no longer than 100nt, although a paired end sequencing approach is also available to redress this issue. Its error rate is also quite high compared to Sanger sequencing, because of a combination of uncorrected PCR errors, reduction enzyme efficiency with read length and possible spurious sequences in the clusters.

5 A TYPICAL ANALYSIS PIPELINE

Mirobiome studies have to follow several phases to pass from marker gene data to diversity profiling. In this section we will revise the preliminary steps that we apply to raw data produced by a sequencer to generate a so-called OTU table, i.e. a matrix having species counts on its rows and data samples on its columns, as summarized in the shaded boxes of Fig.3. By doing so, we will implicitly refer to some of the most used and reliable pipeline already available, like QIIME and MG-RAST [18]. We will not describe the downstream analysis, because they depend on the study purpose and still there is no agreement on what processing are essential and what are not for a generic exploratory analysis. However, the reader should know that it is possible to develop diversity analysis, comparative metagenomics and network analysis starting from this very steps.

As described in the previous sections, suppose we decided to conduct our microbiome analysis by targeting 16S rRNA gene. This means we have already detected an appropriate constant region to be targeted by our primer and we have amplified and sequenced the hypervariable region of interest (for example, V1-V3). Now that we have our samples prepared, we sequence them (for example, using Illumina or Roche/454 as described above) and we obtain the output reads of our target sequences.

1) *Preanalysis step*: this phase requires the user to provide metadata containing information like sample ID, barcode sequences, primer sequences and information about the samples. It includes:
   a) Primer detection and read demultiplexing[1];

[1] A multiplex sequencing assay is a fast, high-throughput, cost-effective sequencing process in which a large number of different samples are pooled and sequenced together, attaching

b) Quality Filtering: usually a minimum quality Phred score of 25 is used to filter or trim the reads; minimum and maximum sequence length are fixed (usually around 200-1000 nucleotide range); maximum number of ambiguous bases and homopolymer length are set.

2) *16S rRNA Detection and Clustering*: during this phase an alignement tool is used to compare and identify 16S rRNAs against a well-known database (like RDP, Greengenes or Silva) using an identity threshold; this step is needed to eliminate every source of DNA/RNA content that might be sequenced that is not 16S rRNA derived from the microbial community under investigation. Then reads are clustered together at 97% identity, forming the so-called *OTUs* (Operational Taxonomic Units), which are operational group used in DNA sequencing as representative of species (or at least genera) present in the sample. The longest sequence or the most abundant one within the cluster after this OTU-picking procedure is used as the representative sequence for each OTU.

3) *Taxonomic Classification*: the representative sequences are used to be aligned against a known database to obtain the taxonomic description of the respective OTU: phylum, class, order, family, genus and species (when possible). There are several taxonomic assignment algorithms available, like the Rdp classifier, Blast, Mothur, Rtax, Pyrotagger. All the OTUs for which taxonomic classification isn't possible, are collected in the category "unknown" or "unclassified". Based upon taxonomic classification it is possible to build phylogenetic trees of the samples when the metadata available are rich enough.

As subsequent analysis, several possibility have to be taken into account:

- Diversity Analysis, as will be described in Chapter 2;
- Graphical inspection using tree plots, heat maps, bar plots and plot ordination methods (PCoA, NMDS, etc.), which are dimensional reduction methods useful to investigate trend in the data;
- Differential abundance estimation, assessed by statistical testing;
- Network analysis.

It still isn't univocally recognized by the scientific community whether or not it is necessary to apply a normalization step to the OTU table before moving on to downstream analysis: we will deal with this procedure ourselves in Chapter 3.

individual "barcode" sequences to each sample so that they can be distinguished in the data analysis phase. It is a useful technique when targeting specific genomic regions, as in the 16S case.

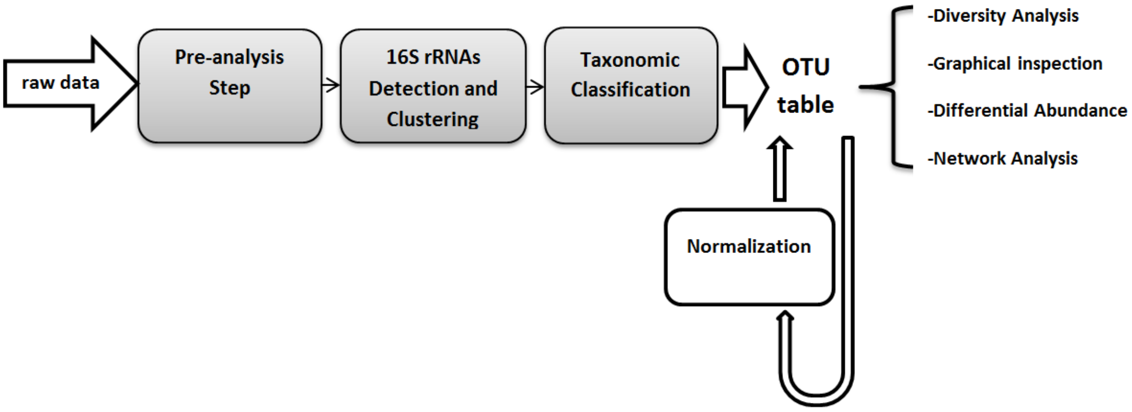


FIGURE 3 OVERVIEW OF A TYPICAL MICROBIOME ANALYSIS PIPELINE

6 Recent studies

6.1 An obesity-associated gut microbiome with increased capacity for energy harvest (Turnbaugh, P.J.; Ley, R.E.; Mahowald, M.A.; Magrini, V.; Mardis, E.R.; Gordon, J.I.; 2006)

In this article, the authors explored the already proven relationship existing between obesity and distal gut microbiota, seeking to establish a cause-effect association. Using both metagenomic and biochemical analyses they verified that change in the gut microbiome composition affect the metabolic potential of the host, by modifying its energy balance between intake and consumption.

The main goal was to verify whether or not the microbiota of obese individuals could be more efficient at energy extraction from food, compared to the microbiota of lean individuals. Two different phenomena supported this idea:

- obesity is characterized by variation in the relative abundance of the two dominant bacterial divisions, the *Bacteroidetes* and the *Firmicutes*, compared to lean individuals;
- germ-free mice tend to gain weight when colonized with distal gut microbial community derived from conventional mice, proving that the microbiota encodes metabolic capabilities to process some otherwise indigestible

components; this impacts energy balance by increasing energy extraction from food income.

By sequencing 16S rRNA samples for microbiome analysis the authors proved that an increase in the relative abundance of *Firmicutes* is typical of obese hosts, whose microbial profile tend to cluster together. Moreover, obese microbiome is enriched for gene tags encoding useful enzymes for polysaccharides digestion, that help the host degrading these complex molecules into simpler (and absorbable) ones. This entails that obese mice microbial composition do play a role in their ability to harvest energy from ingested food, thus favoring weight gain.

In order to assess a cause-effect relation between gut microbiome and obesity, a microbiota transplant experiment has been developed, in which both obese and lean mice were used as donors of harvested microbiota. Interestingly, mice colonized by obese-derived microbiota showed an increase in body fat (during the two weeks of observation) that was significally greater than their lean counterparts, showing a percentage increase of adipose tissue greater than 45%.

The results from this study therefore suggest that obesity-associated gut microbiome do play a role in increasing the host capacity to extract energy from dietary intake.

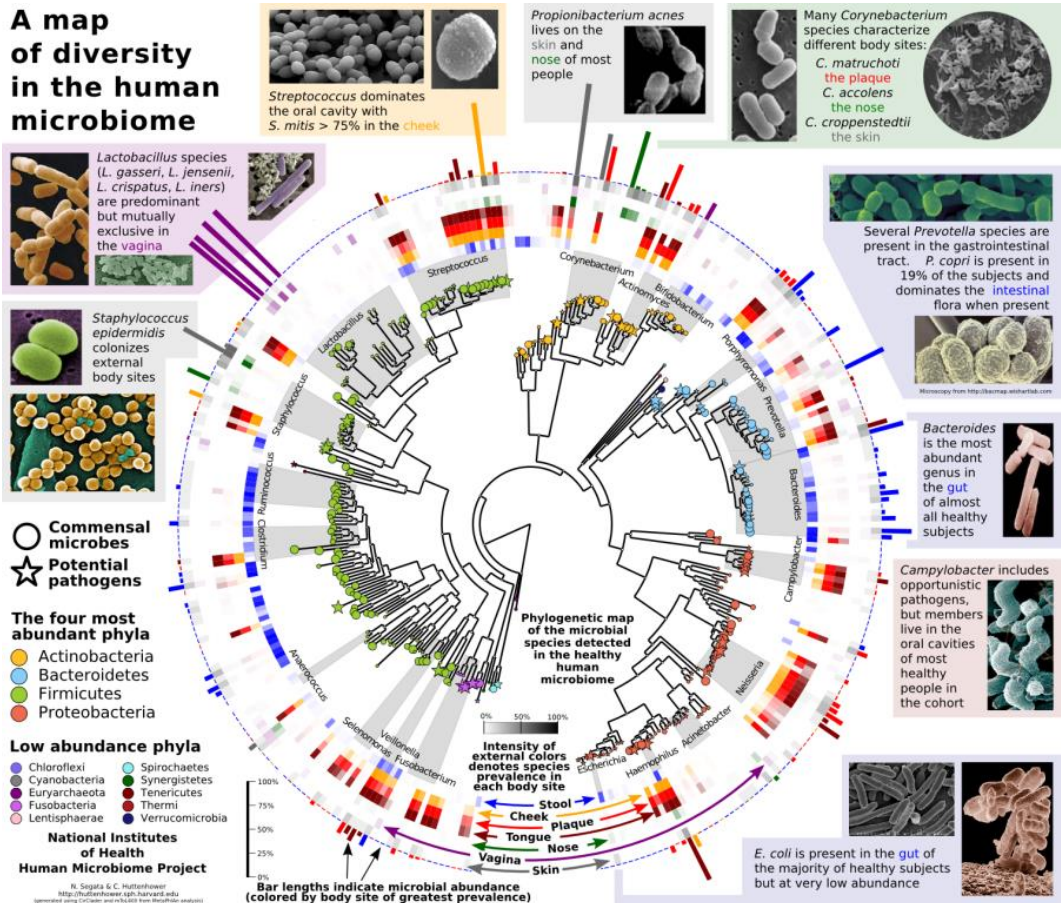
6.2 Development of the Human Gastrointestinal Microbiota and Insights From High-Throughput Sequencing (Dominguez-Bello, M.G.; Blaser, M.J.; Ley, R.E.; Knight, R.; 2013)

In this study, the authors investigated the development of the gastrointestinal tract microbiota, by exploiting high-throughput DNA sequencing and bioinformatics tools to compare bacterial population among individuals and time points. They collected data from many sampling spot of a small cohort of subjects to monitor developmental trajectories of microbiome composition at different stages of life, from newborn to older people.

By supposing that bacteria pioneering newborn microbiota will have a major impact on its development in future stages of life, they evaluated how this primordial ensemble could lead to a complex and stable adult ecosystem. First of all, they reversed the idea that infant are delivered in an almost sterile environment: on the contrary, the birth canal is highly colonized by communities dominated by *Lactobacillus* and *Prevotella* species. Indeed, the vaginal community undergoes several changes during pregnancy, in order to provide newborns with beneficial bacterial strains. Therefore, vaginally delivered babies have their founder species to be related mainly from their own mother's vaginal microbiota. Neonates' body sites are colonized by mainly this unique microbiota, showing largely undifferentiated bacterial communities throughout all their body sites. Before developing the highly differentiated adult microbial communities, infants have to be exposed to diverse human microbes during development that, together with genetic, physiochemical, and dietary factors will contribute to shape microbiome into a unique fingerprint of the individual. For example, breastfeeding has proven to reinforce vaginally acquired, lactic-acid producing bacteria in the infant's GI tracts.

However, an increasingly percentage of newborn babies are now delivered by C-section, that prevents them to get through the birth canal. The authors indeed proved that C-section babies are initially populated by bacterial communities resembling adult skin microbiota. These communities comprise *Staphylococcus*, *Corynebacterium* and *Propionibacterium*, and their intestinal microbiota has proven to remain highly different from natural newborns for several months after birth. The lack of mother-derived microbial colonies affects the development of their gastrointestinal tract microbiota; several hypothesis are being evaluated concerning the possible relationship between C-section babies microbiome and various pathologies, including asthma and allergies.

After birth, the gastrointestinal bacterial community increases rapidly in diversity, although with high instability as well. This trend is maintained over the first few years of life, contemporaneously with exposure to new environments, food and, therefore, bacterial strains. Infancy is indeed a timeframe of rapid colonization related to external events (like diet or health condition) in a cause-effect manner. Although it seems to be a rather individual evolution, major external factors like children's origin clearly cluster children's microbiota together, even if it is still unclear if diet, genetics or environmental factors impact on this patterns. Aging is itself a major feature in determining microbial population composition and diversity of gastrointestinal tract.



A map of diversity in the human microbiome

Streptococcus dominates the oral cavity with *S. mitis* > 75% in the cheek

Propionibacterium acnes lives on the skin and nose of most people

Many *Corynebacterium* species characterize different body sites:
C. matruchoti the plaque
C. accolens the nose
C. croppenstedtii the skin

Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina

Staphylococcus epidermidis colonizes external body sites

Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the intestinal flora when present

Microscopy from http://bacmap.wishartlab.com

Bacteroides is the most abundant genus in the gut of almost all healthy subjects

Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort

E. coli is present in the gut of the majority of healthy subjects but at very low abundance

Phylogenetic map of the microbial species detected in the healthy human microbiome

Intensity of external colors denotes species prevalence in each body site

Bar lengths indicate microbial abundance (colored by body site of greatest prevalence)

National Institutes of Health Human Microbiome Project

N. Segata & C. Huttenhower
http://huttenhower.sph.harvard.edu
(generated using CirCleader and mTol400 from MetaPhlAn analysis)

FIGURE 4 AN OVERVIEW OF TAXONOMIC ASSIGNEMENT AND COMPOSITION OF HUMAN MICROBIOME **[19]**

CHAPTER 2:

HOW TO MEASURE BIODIVERSITY

When comparing biological samples, it is interesting to evaluate how different they are, in order to relate them and assess their biodiversity. This concept, although intuitive, is very broad and sometimes difficult to define precisely: each of us has an inborn ability to detect differences among two specimens, however is far less obvious to agree on which features a diversity measure should focus on, or which properties it should have.

Ecology first dealt with the need of measuring and comparing quantitatively diversity among species and habitats, in order to investigate and preserve ecosystem's richness and variety. Therefore, ecologists felt the need to find suitable indices to measure both species richness and diversity among samples: some of them are derived from well-known theories, e.g. Shannon's information theory [20] or Fisher's log-normal distribution [21], while many others have been newly defined on purpose. Whittaker [22] [23], in 1972, first defined three different terms to measure ecological biodiversity: alpha, beta, and gamma diversity. Alpha diversity refers to the local diversity found within a particular site, area or ecosystem, and is often expressed as the total or mean number of species in that habitat. Beta diversity measures the observed differences among species belonging to different habitats or ecosystems. Gamma diversity is a measure of the overall diversity for different ecosystems or regions, and is determined both by local diversity (i.e. alpha diversity) and habitat diversity (i.e. beta diversity).

Microbiome analyses make use of the concepts expressed above, although with somehow different goals: to compare bacterial population among body sites, to evaluate differences between case and control patients or simply to explore microbiome richness in a new sampling spot. Therefore, they often employ *diversity indices* to quantitatively measure species richness within a single sample and diversity between different samples. When diversity indices are applied to ecology, they usually refer to the observed species in different habitats or regions. Microbiome studies can easily extend these concepts by referring to species or taxa presence, absence, abundance in samples belonging to different subjects or anatomical sites. In this chapter, as in most microbiome studies, attention is focused on alpha and beta diversity, applied within or between sample to assess biodiversity. Gamma diversity is indeed a disputed measure, since it can be obtained from alpha and beta diversity using different models. However, debating its definition goes beyond our purposes and we will therefore neglect it, keeping in mind that alpha and beta diversity already provide a complete description of what we need to calculate it.

The rationale behind this review has emerged by noticing that a vast and heterogeneous literature concerning diversity indices exists, spanning from ecology to economics and, last but not least, to microbiome studies. However, it is not clear in the literature what indices do really measure and when to use them, how they can be adapted to microbiome features and how they are influenced by them. Many authors before us already faced this problem, all of them having an ecological background [24] [25] [26] [27] [28] [29] [30], well aware that, as Henk Wolda said [30]:

> […] *the results depend largely on the index chosen, which suggests the dangerous possibility that one can choose an index to demonstrate whatever one wants the data to show, without necessarily being able to prove that this is indeed what they do show.*

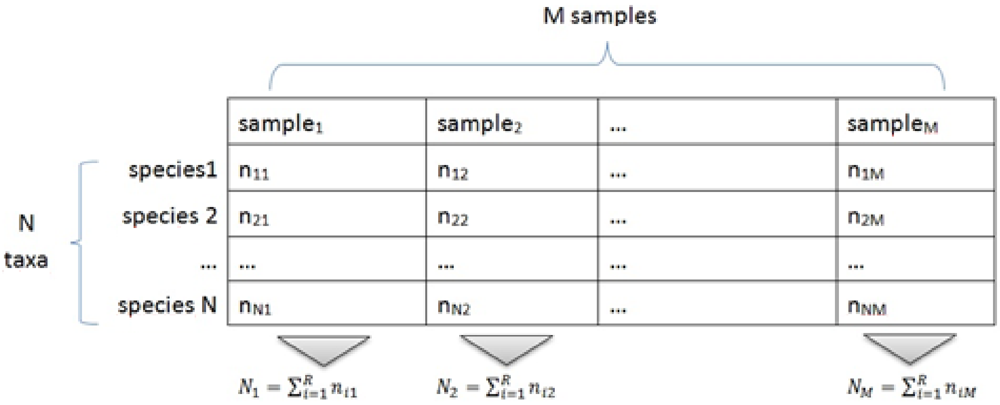
However, microbiome studies have seldom dealt with this problem, mainly borrowing measures that had already proven their effectiveness in other fields of study. Indeed, with the current computational capability of computers, there is no need to choose one particular index, since all of them can be calculated with ease. Nonetheless we believe that having many measures at hand without considering what properties they are inspecting or what can influence them, could worsen or mislead our understanding of the data, instead of shedding light on their inner characteristics. Thus we have structured this review, far from being exhaustive, to be focused on the pros and cons of applying some of the most used measures to microbiome data, aware of the biological question these indices are asked in this specific context and of the unique features this kind of data have. This is exactly the approach we are going to use, in the next sections, to explore alpha and beta diversity, separately, sure that it will be of some use to those approaching biodiversity investigation of microbiome data.

1 USEFUL NOTATION AND TERMINOLOGY

In the next sections we will review most of the proposed measures we found in microbiome literature to quantify biodiversity. In order to do so, we will here introduce a list of symbols and terms the reader may refer to in the next sections, for sake of clarity. Figures 5 and 6 provide further graphical explanation of the same concepts.

- *ecological niche*: we represent it as a matrix, containing M different samples as columns and N different species as rows.

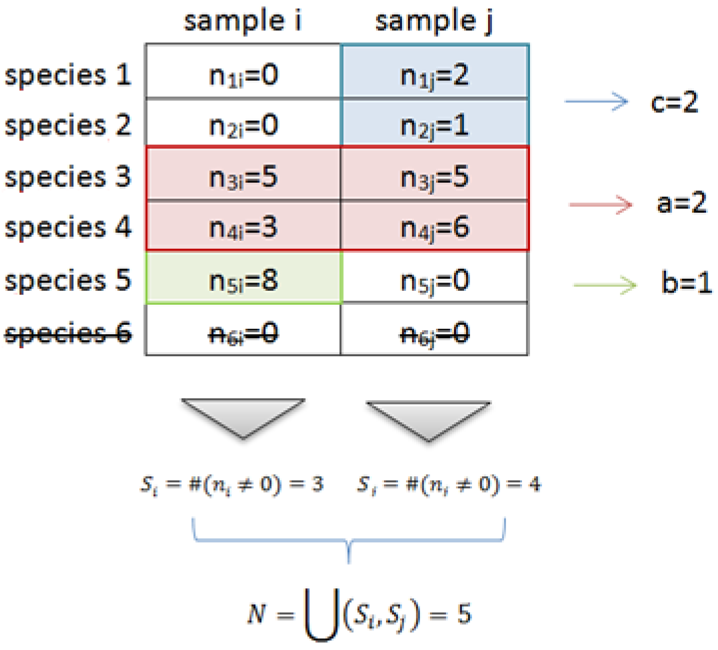
- *species*: in this context we will refer to species to indicate OTU, genera, or any taxonomic level we have decided to focus on.
- *true number of species R*: it describes the (unknown) real number of species present in the sample from which data are drawn.
- *observed number of species S*: it measures the total number of species within a sample having non zero abundance; it is always true that S≤R.
- *total number of species N*: it represents the number of species found in the union set composed by all the samples (columns) considered.
- *total abundance in a sample* N_k: total number of individuals/counts found in sample k; it equals $N_k = \sum_{i=1}^{N} n_{ik}$.
- *abundance* n_{ik}: the measured quantity in sample k belonging to species i; note that it may represent the number of individuals identified as well as the number of count mapped to the specific taxa, depending on the context.
- *relative abundance* p_i: it describes what proportion of the total individuals in a sample belongs to a particular species i; it is calculated as $p_i = \frac{n_{ik}}{\sum_j n_{jk}}$.
- *shared species (a):* when comparing two samples, the number of species having nonzero abundance in both of them.
- *unique species (b and c):* when comparing two samples, the number of species being nonzero in only one of them.



	sample$_1$	sample$_2$	...	sample$_M$
species1	n_{11}	n_{12}	...	n_{1M}
species 2	n_{21}	n_{22}	...	n_{2M}
...	...	...	...	...
species N	n_{N1}	n_{N2}	...	n_{NM}

$N_1 = \sum_{i=1}^{R} n_{i1}$ $N_2 = \sum_{i=1}^{R} n_{i2}$ $N_M = \sum_{i=1}^{R} n_{iM}$

FIGURE 5 DEFINITION OF M SAMPLES, N SPECIES, ABUNDANCES N$_{IJ}$



	sample i	sample j
species 1	$n_{1i}=0$	$n_{1j}=2$
species 2	$n_{2i}=0$	$n_{2j}=1$
species 3	$n_{3i}=5$	$n_{3j}=5$
species 4	$n_{4i}=3$	$n_{4j}=6$
species 5	$n_{5i}=8$	$n_{5j}=0$
~~species 6~~	~~$n_{6i}=0$~~	~~$n_{6j}=0$~~

$S_i = \#(n_i \neq 0) = 3$ $S_j = \#(n_j \neq 0) = 4$

$$N = \bigcup (S_i, S_j) = 5$$

FIGURE 6 DEFINITIONS OF TOTAL NUMBER OF SPECIES N, MEASURED NUMBER OF SPECIES S, SHARED SPECIES (A) AND UNIQUE SPECIES (B AND C)

2 Alpha Diversity

Alpha diversity is a measure, introduced in 1960 by Whittaker [22] [23], describing the ecological diversity found within a particular sample. In principle, the value of the diversity indices used to quantify it increases according to the number of species found in the sample (i.e. the so-called *richness*) and the evenness of their distribution. However, many different measures of alpha diversity are available. Here we consider measures of richness and evenness, both deriving from data distribution or species-abundance models, as summarized in Table1.

We can consider species richness and evenness as two independent characteristics of biological communities, that together contribute to its overall diversity. Indeed, species richness describes the contribution to the total alpha diversity brought by the number of different species found in the sample, while species evenness focuses on how different abundance distribution of individuals in the sample may affect its diversity. Nearly all richness and evenness indices are calculated starting from relative abundance of species, i.e. on $p_i = n_{ij}/N_j$, with n_{ij} the abundance of the i-th species in the sample j and N_j the total abundance of the sample.

The goal of this section is to clarify what experimental features may affect alpha diversity analysis, looking for robust, clearly understandable and comparable measures among the many proposed. Many factors indeed are involved in determining the alpha diversity estimated for a sample: how many species are in it, how many of them are effectively measurable, how individuals are distributed among species, what is the discriminating power of the measure being used, and many others. To all these sources of uncertainty has to be added the complexity brought by the experimental features of microbiome analysis. Therefore we developed an introductory analysis of some of the available indices commonly used to quantify alpha diversity, followed by a simulation analysis useful to investigate both measure's reliability and experimental factors' impacting on them.

2.1 Definition of alpha diversity indices

Here follows a description of the indices used in this survey, divided into three categories: measures focusing on sample's richness, measures dealing with evenness of their abundance, and measures summarizing them both in a unique index.

Species richness

- **Observed species, Margalef, Menhinick**

The simplest index available measures the total number of observed species S in the sample, i.e., the number of species showing a nonzero abundance in the dataset. Clearly, by its definition, this measure is correlated with sample size. Many authors have tried to correct this index, using different strategies: e.g. Margalef's diversity index ($D_{Mg} = \frac{(S-1)}{\ln N_j}$) and Menhinick's diversity index ($D_{Mn} = \frac{S}{\sqrt{N_j}}$) try somehow to disentangle this measure from the total abundance found in the sample. Both these indices show good discriminating ability, but their measures remain strongly influenced by the total abundance sampled.

- **Chao index**

This index, first proposed by Anne Chao in 1984 [31], aimed at correcting species richness, accounting for unknown missing species. It uses the number of rare species in the sample to evaluate how likely it is that there are more undiscovered species being ignored. Once we define n_i as the number of species with abundance i (or with i sequences) the S_{c1} index is composed by S, the observed number of species, n_1 the number of species with only one sequence (i.e. "singletons") and n_2 the number of species with only two sequences (i.e. "doubletons"), as explained by the formula:

$$S_{c1} = S + \frac{n_1(n_1 - 1)}{2(n_2 + 1)}$$

If a sample contains many singletons, it is likely that more undetected species exist, and the Chao index will then estimate greater species richness than it would for a sample without rare species.

<u>Species evenness</u>

- **Simpson index**

The original Simpson measure, named λ, was introduced in 1949 by Edward H. Simpson [32] as an estimate of concentration in a classification, and it equals the probability that two entities taken at random from a dataset belong to same species, or, in formulae:

$$\lambda = \sum p_i^2$$

λ has its minimum value to be 1/S, which is reached when all types are equally abundant, since proportional abundances are by definition values between 0 and 1, while its maximum value, 1, is obtained when no diversity is observed. It is considered a dominance index, indeed it is heavily weighted towards the most abundant species in the sample and is less sensitive to species richness. Notably, λ obtains smaller values for increasing number of species, showing an opposite behavior compared to all other diversity measures. Therefore transformations of λ were proposed, that increase accordingly with increasing diversity: the inverse Simpson index, also noted as E, that is an evenness measure, simply computes the reciprocal of λ; the complementary Simpson index, C, calculates its one complement, so that its range remains limited to [0,1].

- **Pielou index**

Pielou's index calculates species evenness from diversity measures, by dividing Shannon-Wiener index (H) by its highest value ($\log_2 S$), so that it converges for large samples. Even if this index has a large literature proving its poor performances [33], it is still one of the most widely used to assess species distribution evenness [34]. It is dependent from richness measure and particularly it depends on the correct estimation of R, the true number of species in the community. Using S, the number of observed species in the sample, as an estimate of the true number of species in the sample makes this index both highly dependent on sample size and very sensitive to inclusion or exclusion of rare species, thus it is difficult to achieve robust comparison between samples.

- **Camargo index**

Camargo index is an evenness measure independent of number of species that focuses only on the distribution of individuals among the species:

$$Cam = 1 - \sum_{i=1}^{S} \sum_{j=i+1}^{S} \frac{|p_{ik} - p_{jk}|}{N}$$

p_{ik} is the relative abundance of species i in the sample; p_{jk} is the relative abundance of species j in the sample while N is the total number of species. However Mouillot [35] proved it is less robust compared to Shannon and Pielou since it has a very high bias when the total species richness is high.

- **Log-series index**

This index is widely used because of its good discriminant ability and its independence from sample size. It derives from Fisher's log-series [36], where it is assumed that the abundance of species follows the log series distribution

$$\alpha x, \frac{\alpha x^2}{2}, \frac{\alpha x^3}{3}, \dots, \frac{\alpha x^n}{n}$$

(each term gives the number of species predicted to have 1,2,3,....n abundance in the sample) and the expected number of species with n_i observed individuals equals

$$E[n_i] = \frac{\alpha x^{n_i}}{n_i}$$

where α is the parameter used to describe diversity and $x = n_i/(n_i + \alpha)$. Usually both α and x are estimated with a maximum likelihood approach from the data, or with an iterative approach. Indeed, they can be found as the solution of the system

$$\begin{cases} S = \alpha \ln(1 - x) \\ N = \dfrac{\alpha x}{1 - x} \end{cases}$$

This index describes the way in which individuals are divided among different species, which is a measure of diversity. It shows a good discrimination power between sites, insensitivity to density fluctuations and has normal distribution, which allows for confidence intervals to be defined. Still, when data distribution deviates from the log-series, α becomes dependent on sample size.

<u>Both species richness and evenness</u>

- **Shannon-Wiener index**

This index resumes the measure originally proposed by Claude Shannon [37] to quantify the entropy (also called "uncertainty" or "information content") in strings of text. The Shannon index calculates the uncertainty in predicting the species an individual taken at random from the dataset belongs to.

When all the species in the dataset are equally common, all values equal 1/S, and the Shannon index hence takes its maximum value, equal to ln(S). Conversely, when there is only one type in the dataset, Shannon entropy exactly equals zero, since there is no uncertainty in predicting the type of the next randomly chosen entity. In its most common version, it chooses 2 to be the base of the logarithm used, although the meaningfulness of the consequent unity of measure, bits, is already been reported as doubtful in ecological applications [25].

- **Hill's number**

Hill proposed a unifying statistic [38] that encapsulates several diversity measures depending on the value assigned to an adjustable parameter *a*. The most interesting cases are:

- $a=0$, in this case the Hill's number coincides with the total number of species observed, S.
- $a=1$, $H_1=exp(H)$ is the equivalent of the Shannon-Wiener index, but expressed in terms of equivalent number of species.
- $a=2$, $H_2=1/\lambda$ coincides with the invSimpson index
- $a\rightarrow \infty$, $H_{inf}=1/p_x$, where p_x is the proportional abundance of the most common species and a dominance index (also known as Berger-Parker index [39]). Its reciprocal, H_{inf} increases accordingly to the diversity of the sample.

Aside from the convenience of having an only index containing several diversity measures of use, increasing the value assigned to *a* allows the user to give more weight to the most abundant species in the overall calculation of the diversity value.

All the measures revised so far are summarized for ease in Table 1. By simply inspecting the expressions of the indices proposed, we can already investigate their range and critical points. Clearly the total number of species S, the Chao index S_c, the Fisher log-series index α, the Margalef index and the Menhinick index (D_{Mg} and D_{Mn} respectively) do not have an upper bound (while they have a lower bound fixed to 0) because all of them depend upon the number of species detected, which may assume any value. Conversely, the Simpson index C, the Pielou index R and the Camargo index *Cam* have their range to be bounded within 0 and 1, because their value is obtained by calculation on probability values. Lastly, the remaining measures have a limited range but different from [0,1]: the inverse Simpson index E and the Hill number (here we consider mainly H_1 and H_{inf}) span from 1 to S, while the Shannon-Wiener index ranges between $[0, \ln S]$. Some additional characteristics of the indices proposed in the literature must be taken into account. The Pielou index cannot be evaluated if S=1 (only one species detected), because both its numerator and denominator equal zero; similarly the Margalef index cannot be calculated if $N_{tot}=1$, because in this case S=1 too and again both numerator and denominator equal zero. Obviously, if an empty sample is taken into account, the invSimpson, the H_{inf}, the Menhinck and the Camargo index will attain an undefined numerical result too. Therefore we suggest to trim away samples showing $N_{tot}=0$ before evaluating alpha diversity.

Equation	Referred to
S	Total number of species in the sample
$D_{Mg} = \frac{(S-1)}{\ln N_{tot}}$	Margalef diversity index
$D_{Mn} = \frac{S}{\sqrt{N_{tot}}}$	Menhinick diversity index
$H = -\sum_i p_i \log_2 p_i$	Shannon-Wiener index
$S_{c1} = S + \frac{n_1(n_1-1)}{2(n_2+1)}$	Chao index
$C = 1 - \lambda$ $E = 1/\lambda$	Complementary Simpson index (named Simpson index from now on) Inverse Simpson
$R = \frac{H}{\log_2 N}$	Pielou's regularity index
$Hill_a = \left(\sum_i p_i^a\right)^{\frac{1}{1-a}}$	Hill's diversity number
$Cam = 1 - \sum_{i=1}^{S}\sum_{j=i+1}^{S}\frac{p_i - p_j}{N}$	Camargo index
α estimated from $E[n_i] = \frac{\alpha x^{n_i}}{n_i}$	Log-series index

TABLE 1 ALPHA DIVERSITY MEASURES ANALYZED IN OUR REVIEW,

2.2 Desirable properties of alpha diversity indices

As we mentioned in the introduction, our aim is to test the ability of the revised measure to reliably detect alpha diversity, and to assess their dependence on experimental features like sequencing depth, number of species and evenness of species abundance distribution. Microbiome studies do indeed suffer from the difficulty to reliably estimate species composition, richness and diversity of a sample because of a combination of these factors. For example, if the sequencing depth available is smaller than the total number of species present, some taxa is going to remain inevitably undetected. However, even though we have a sequencing depth exceeding the total number of species, some of them could be so rare that the sampling process doesn't manage to detect them.

We would theoretically look for measures that are independent from sequencing depth, a common feature we are very concerned about in microbiome studies. It is indeed often variable among samples, thus preventing them to be reliably compared. Then we search for measures that combines all the characteristics we want to detect in our samples: species richness and evenness, true samples' dimensions and features. For instance, it would be desirable for an index aimed at assessing species evenness to be almost insensitive in variation of species richness, if the abundance distribution is held fixed. On the other hand, we would hope a richness index to be able to extract the same number of species even if species abundance distribution is varied. We investigated these properties and the effect of a combination of the three features highlighted by means of a simulation study.

2.3 SIMULATION STUDY

In order to evaluate the performances of the measures investigated, we built up a simulation study in which all these diversity indices are tested upon samples derived from a combination of six different sequencing depth (1e+03, 1e+04, 1e+05, 4e+05, 7e+05, 1e+06) and nine different total number of species, expressed as percentage of the sequencing depth (1%, 12.5%, 25.0%, 37.5%, 50.0%, 62.5%, 75.0%, 87.5%, 100%). Their abundance data are extracted from seven different cumulative counts shapes, to account for the difference among samples with evenly distributed species abundance and samples with a few species prevailing on the others, as showed for clarity in Fig.7.

Firstly we investigated the effect of total species variation on alpha diversity values, when sequencing depth is maintained constant and all seven abundance distribution are evaluated (Fig.8). We examined this scenario for all the indices discussed above when sequencing depth equals 1e+03, 1e+04, 1e+05, 4e+05, 7e+05, 1e+06 respectively. Then we evaluate how alpha diversity behave when the sequencing depth increases under all seven possible species abundance distribution (Fig.9). We studied this scenario for all the indices discussed above when total number of species equals 1%, 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, 87.5% and 100% of the total sequencing depth. Lastly we inspected alpha diversity when the total number of species varies and focusing our attention on different situation characterized by different sequencing depth (Fig. 10). We considered this scenario for all the indices discussed above when cumulative sum of count abundances follow shape 0, shape 1, shape 2, shape 3, shape 4, shape 5 and shape 6. Besides, we separately evaluate how using the measured number of species observed in the sample instead of the real total number of species present (here known because we are in simulation conditions) could affect some of the indices proposed.

2.4 RESULTS ON ALPHA INDICES

The next paragraphs and figures summarize the results of our simulation, showing how alpha diversity measures behave when one of these three aspects is investigated: variation of the sequencing depth, variation of the total number of species, variation of count abundance distribution shape (see Fig.7). Only the most interesting graphics are shown, in order to avoid redundancy. All the other figures, that may help the reader verifying the statement we developed, are available in Appendix A.

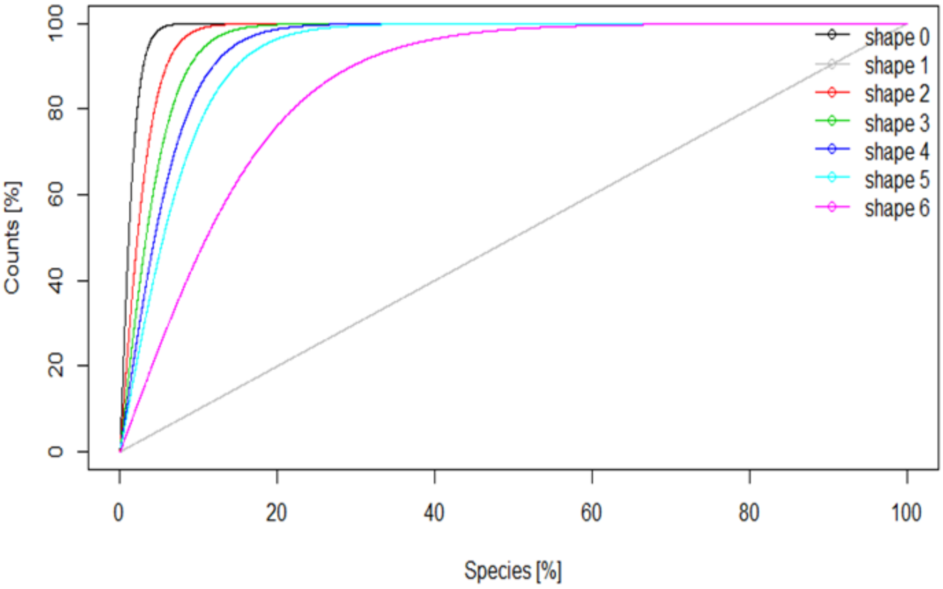


FIGURE 7 DIFFERENT SPECIES ABUNDANCE DISTRIBUTION SIMULATED

Dependency from sequencing depth

If we look at the results obtained in Fig. 8, 9 and 10, we notice that all the revised measures are influenced by sequencing depth. The measures having range in [0,1] rapidly attain their saturation value as sequencing depth increases. In details, both Simpson and Camargo index depend less evidently from the shape selected, and gain their maximum value of 1. However, if sequencing depth and total number of species are small the dependency from abundance distribution shape cannot be neglected. These measures therefore obtain the same value for samples having the same total number of species (here expressed as percentage of the sequencing depth), but they are unable to detect any difference in richness among samples having different number of species but the same S/N_{tot} ratio (where N_{tot} equals the sequencing depth). Notably, Camargo index has very poor discrimination ability, since in most of our simulation conditions obtains a value near to its upper bound: this evidence confirms that it suffers from high bias whenever the total number of species is high [35], which might be an important drawback for applicability to microbiome studies.

The other measures having a limited range, i.e. Shannon, inverse Simpson and Hill's indices, show an opposite trend: for high sequencing depth and high number of species the alpha diversity value obtained is determined mostly by the particular species abundance distribution shape, as desirable, regardless of the total species detected. The remaining indices, Fisher's alpha, Chao index, D_{mg}, D_{Mn} and the total species S show a peculiar behavior: likewise the other measures, the alpha diversity value they obtain is independent from total number of species if both sequencing depth and total species are sufficiently high (i.e. S≥10% of the sequencing depth, and $N_{tot}\geq 10^5$). However, for increasing sequencing depth, D_{Mg}, D_{Mn} and alpha show different trend depending on the different shape involved: for the most even ones their diversity value increases logarithmically for small sequencing depth and then decreases exponentially as soon as sequencing depth exceeds 10^5, while a saturating trend is shown for most uneven species abundance distributions. Interestingly, invSimpson confirms it can be used as an evenness index as stated in [33], so that Simpson and InvSimpson together might provide a complete description of both data richness and evenness. Simpson index shows under every condition a saturating trend towards one, reached exponentially when total species increase. Chao index and S show a similar pattern, as both increase logarithmically with sequencing depth and linearly with the total number of species, toward a saturating value that depends upon the abundance species distribution shape considered. On the other hand, Chao, invSimpson, alpha and Hill's indices increase linearly with total species in the sample, with slopes depending on cumulative counts shape.

In our simulation we decided to define the total number of species detected as a percentage of the available sequencing depth, therefore these two features have a coupled effect on the alpha diversity value obtained when one of these two characteristics is increased. We thus developed a quick example, in which the total species present are fixed, in order to disentangle these two aspects in our analysis.

We then set total species equal to 10, 100, 1000, 4000, 7000, 10000, while the sequencing depth considered are the same shown before except 1e+03. Fig. 11 shows the same condition as Fig. 8 when the number of species is held fixed. There are two main differences with the previous analysis: firstly only Simpson, Camargo and Pielou index attain a saturation value (which is, as already explained, equal to 1 for C and Cam and approximately equal to 0.7 for R), while all other indices increase accordingly with total species, following either a linear or a logarithmic pattern depending on the specific formula they use. Secondly, since the total number of species doesn't increase with sequencing depth anymore, only Chao, Margalef, Menhinick, Pielou index and S change their value with sequencing depth. This testifies the impact that sequencing depth has on the ability to detect the true total number of species present, since higher sequencing depth allows us a more powerful measure for S. All the indices listed indeed depend on the value obtained for S.

Dependency from species abundance distribution shape

In this section we evaluated the impact that species abundance distribution has on alpha diversity measures by simulating seven different cumulative counts layouts, as displayed in Fig.7. The most even one, named "shape 0" is not realistic in practice, since it assumes that all the species are perfectly equally abundant. However it served us as a reference to assess the different indices performances under the simplest condition, in which no abundant species may hide any rare one. This in some cases lead to a pattern that distinguishes this shape only from the others in determining the alpha diversity value measured. Most of the indices proposed follow the same pattern when sequencing depth or total number of species is increased, with the alpha diversity measure increasing monotonically from the most uneven distribution shape to shape 0: indeed, all the indices that look at data evenness gain higher alpha diversity value when the most even distribution is investigated. Each measure however shows a different sensibility to count abundance distribution and therefore has a different range of values. Simpson and Camargo index are almost insensitive to shape variation, while all the other measures are highly dependent from it, since it determines the total number of species detected, S, as discussed in the next paragraph.

Dependency from S

Most of the indices considered in this survey attain the same value when calculated using the observed total number of species S (that is, the number of taxa showing a nonzero count abundance) or using the real number of species R, fixed in our simulation to be a percentage of the sequencing depth. Indeed, most of them directly use the relative abundance of species, derived from the count distribution only.

However, as shown in Fig. 12, sequencing depth and abundance distribution shape influence greatly the discrepancy found between S and R, and therefore have a major impact on the alpha diversity measures that directly use this value. Three of the investigated indices, in detail Margalef index D_{Mg}, Menhinick index D_{Mn} and Pielou index show a different trend under the two conditions proposed, as displayed in Fig. 11, 12 and 13. D_{Mg} and D_{Mn} linearly increase with the total number of species and they are insensitive to the used shape if S=R is considered. Under the same condition, D_{Mg} increases linearly with the sequencing depth, while D_{Mn} shows a logarithmically increasing trend.

Conversely, if S equals the number of observed species, both D_{Mg} and D_{Mn} become sensible to the species abundance distribution shape and show a different trend depending on the sequencing depth available. Indeed, for low sequencing depth (10^3-10^4) these measures still increase linearly with the total number of species detected, with the slope depending on the abundance distribution shape used. For higher sequencing depth (10^5-10^6) both tend to a saturating value, that they maintain constant, irrespective of total species increase. This trend shows how, for high sequencing depth, the alpha diversity value obtained by those index is more influenced by the abundance distribution shape (that determines how easily all the present species can be detected) than from the real total number of species. Moreover, under the same condition, D_{Mg} and D_{Mn} show a peculiar behavior when sequencing depth increases, depending on the total number of species observed. If the sequencing depth greatly exceeds the total number of species present they still show the same trend obtained for S=R; in this case the sequencing depth is high enough to detect all the present species. However, as the total number of species increases, the alpha value obtained begin to increase less than linearly for D_{Mg}, while D_{Mn} shows a pattern where it increases logarithmically and then decrease exponentially as the sequencing depth increases. The Pielou index shows a different pattern from the measures described above. In both the theoretical and realistic condition it tends to a saturating value and shows to be sensitive to the abundance distribution shape used. However, the saturating value changes among the two situations, and the dependency from the shape used appear more markedly when S=R. Under this condition it shows to change trend, decreasing exponentially towards a lower saturating value when the sequencing value increases, if the total number of species is sufficiently high (for example if S>10% of sequencing depth, when it is$\geq 10^4$). We can explain this pattern by looking at the index formula: $R \geq S$ always, especially for the most unevenly distributed count abundance, therefore the index value obtained with the measured total number of species will always have a smaller denominator thus maintaining higher values.

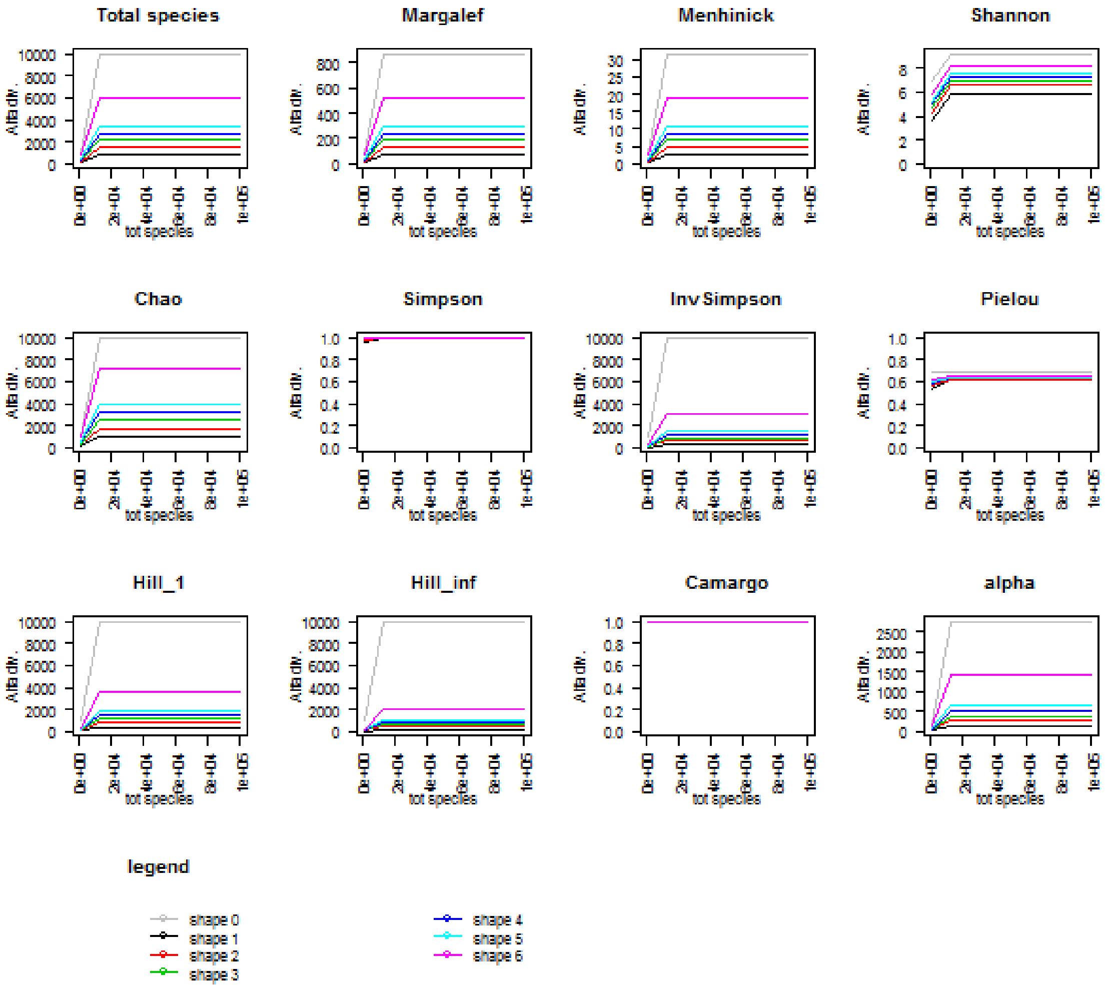


FIGURE 8 ALPHA DIVERSITY VALUES FOR SEQUENCING DEPTH=1E+05: MAXIMA AND MINIMA FOR EACH INDEX ARE

```
                      min          max
Total species 106.0000000 1.000000e+04
Margalef        9.1201841 8.685021e+02
Menhinick       0.3352014 3.162278e+01
Shannon         3.6093441 9.210340e+00
Chao          115.1666667 1.000000e+04
Simpson         0.9666726 9.999000e-01
InvSimpson     30.0053491 1.000000e+04
Pielou          0.5364725 6.931472e-01
Hill_1         36.9418156 1.000000e+04
Hill_inf       20.0160128 1.000000e+04
Camargo         0.9999988 1.000002e+00
alpha          11.7092627 2.766290e+03
```

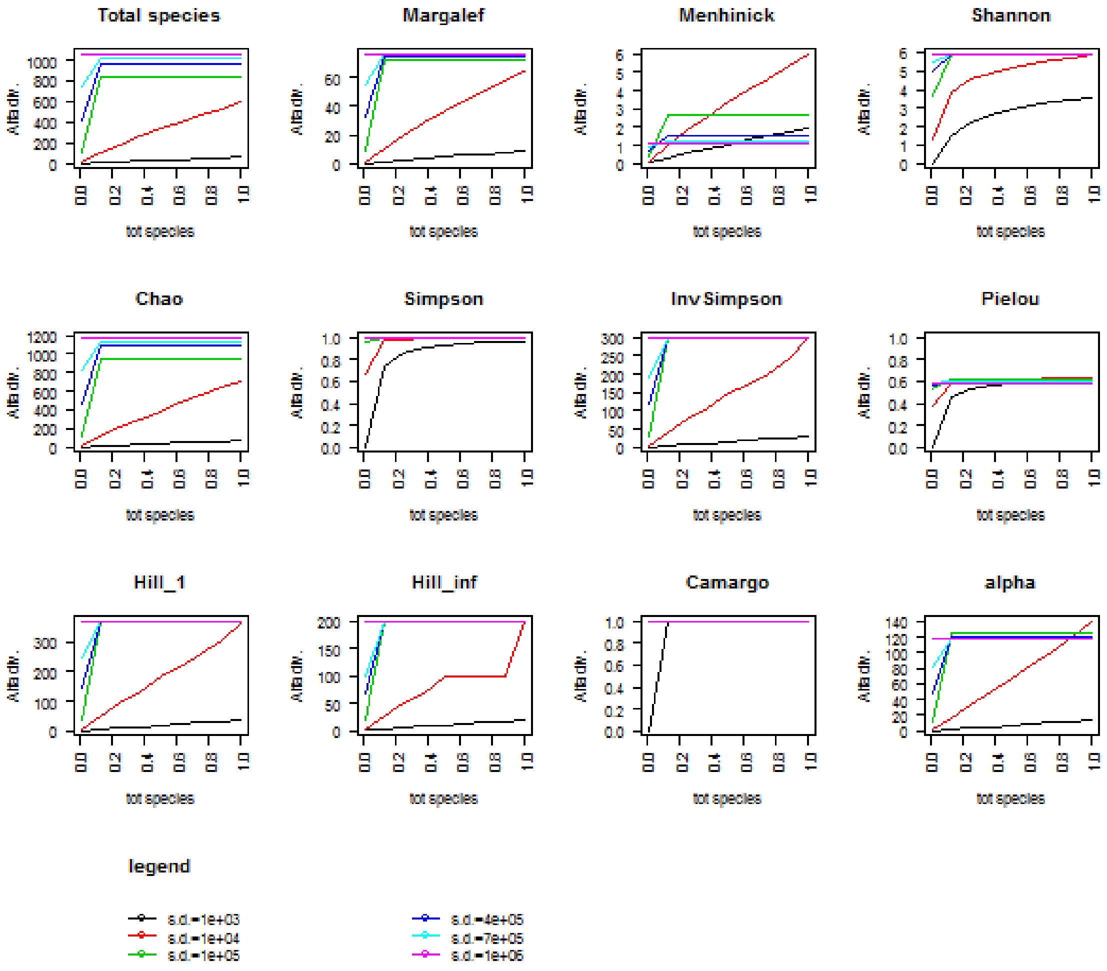


FIGURE 9 ALPHA DIVERSITY VALUES WHEN SHAPE=2: MAXIMA AND MINIMA FOR EACH INDEX ARE

```
                      min          max
Total species 1.00000000 1060.0000000
Margalef      0.00000000   76.6532313
Menhinick     0.03162278    6.0005101
Shannon       0.00000000    5.9116085
Chao          1.00000000 1175.2884615
Simpson       0.00000000    0.9966664
InvSimpson    1.00000000  299.9766592
Pielou        0.00000000    0.6380804
Hill_1        1.00000000  369.2996796
Hill_inf      1.00000000  199.9908000
Camargo       0.00000000    1.0000047
alpha         0.10967162  139.9756918
```

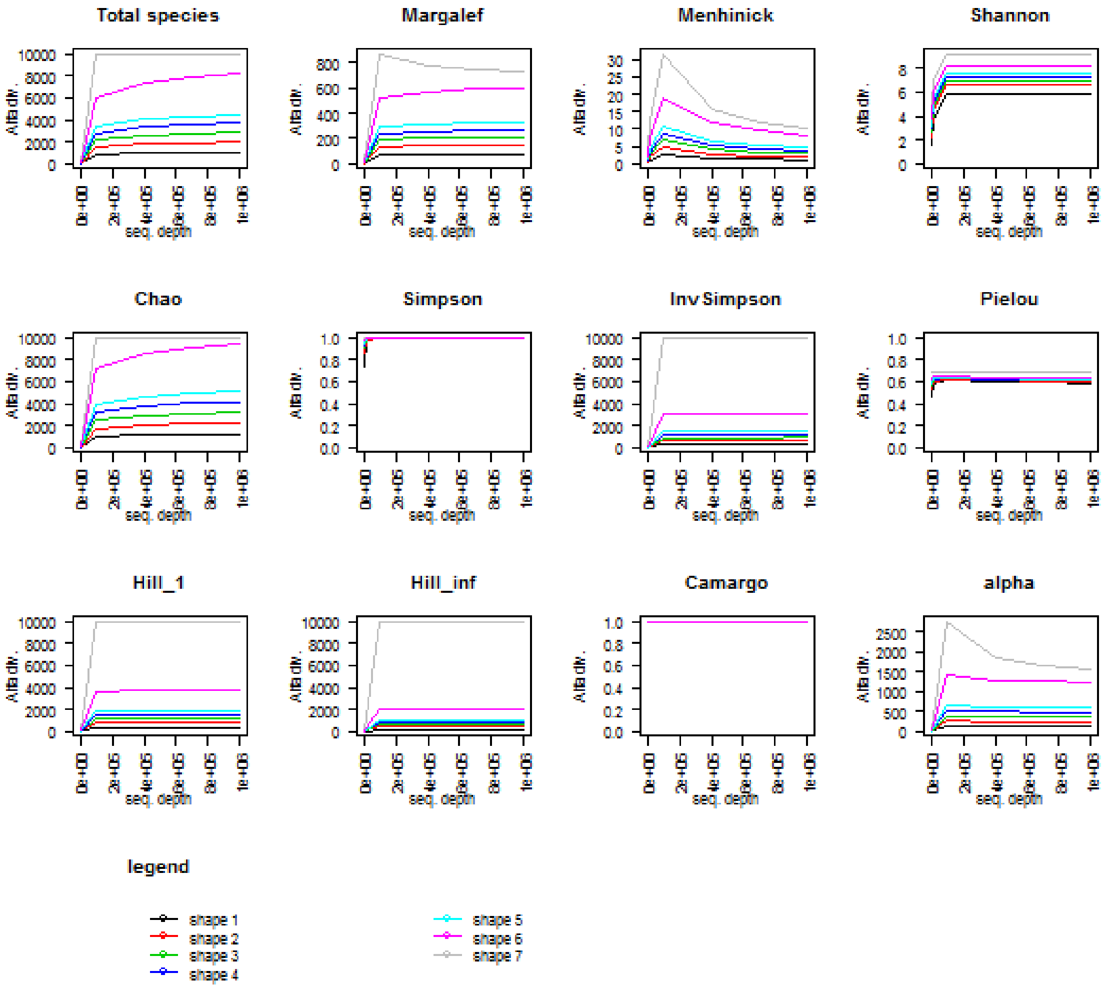


FIGURE 10 ALPHA DIVERSITY VALUE FOR TOTAL NUMBER OF SPECIES=12.5% OF SEQUENCING DEPTH: MAXIMA AND MINIMA FOR EACH INDEX ARE

```
                     min          max
Total species 10.0000000 1.000000e+04
Margalef       1.3028834 8.685021e+02
Menhinick      0.3162278 3.162278e+01
Shannon        1.5354869 9.210340e+00
Chao          10.0000000 1.000000e+04
Simpson        0.7360340 9.999000e-01
InvSimpson     3.7883667 1.000000e+04
Pielou         0.4622276 6.931472e-01
Hill_1         4.6435859 1.000000e+04
Hill_inf       2.6315789 1.000000e+04
Camargo        0.9999000 1.000004e+00
alpha          1.5445238 2.766290e+03
```

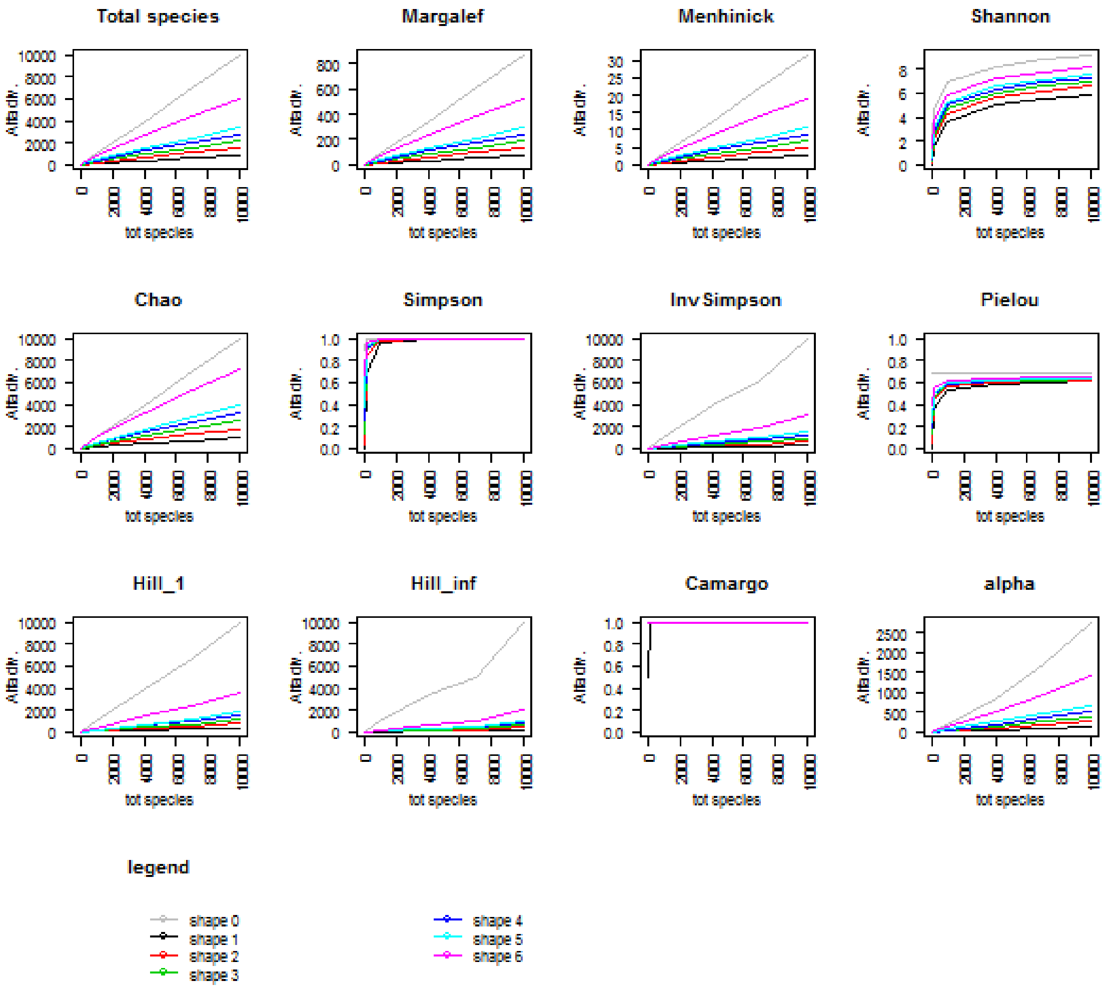


FIGURE 11 SAME AS FIG. 2, BUT WITH FIXED TOTAL NUMBER OF SPECIES: MAXIMA AND MINIMA FOR EACH INDEX ARE

```
                       min          max
Total species 2.0000000000 1.000000e+04
Margalef      0.0868588964 8.685021e+02
Menhinick     0.0063245553 3.162278e+01
Shannon       0.0009284090 9.210340e+00
Chao          2.0000000000 1.000000e+04
Simpson       0.0001799838 9.999000e-01
InvSimpson    1.0001800162 1.000000e+04
Pielou        0.0009284090 6.931472e-01
Hill_1        1.0009288401 1.000000e+04
Hill_inf      1.0000900081 1.000000e+04
Camargo       0.5000900000 1.000000e+00
alpha         0.1490730138 2.766290e+03
```

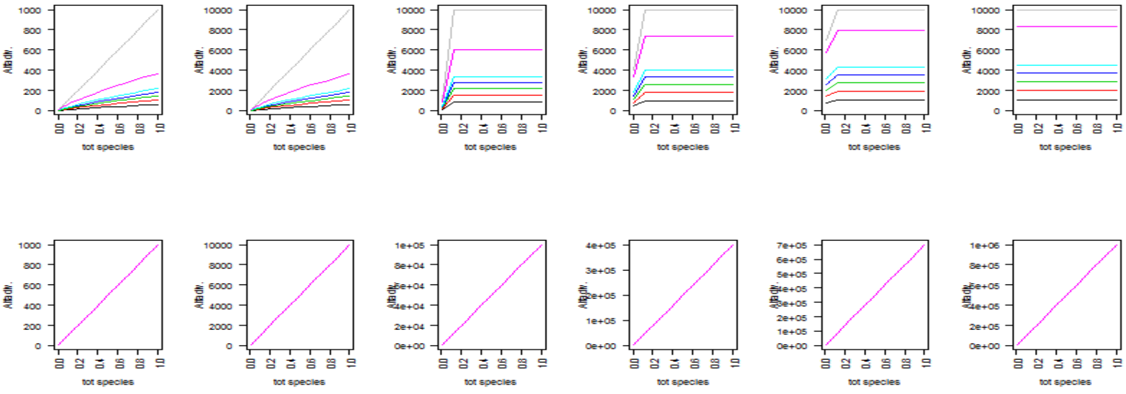


FIGURE 12 DIFFERENCE BETWEEN THE MEASURED TOTAL NUMBER OF SPECIES, S, AND THE TEORICAL (SIMULATED) ONE, UNDER ALL SIX SEQUENCING DEPTH CONDITIONS.

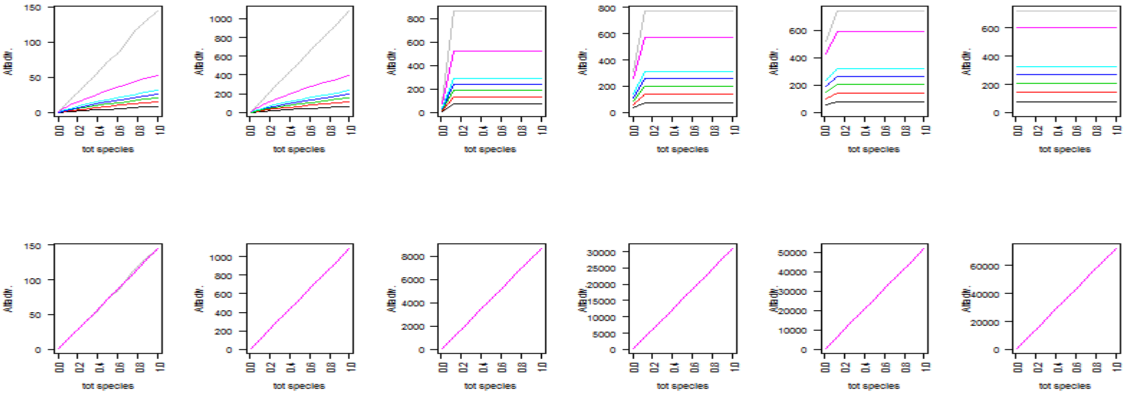


FIGURE 13 DIFFERENCE BETWEEN THE MARGALEF INDEX OBTAINED WITH THE MEASURED NUMBER OF SPECIES S AND WITH THE TEORICAL ONE, UNDER ALL SIX SEQUENCING DEPTH CONDITIONS.

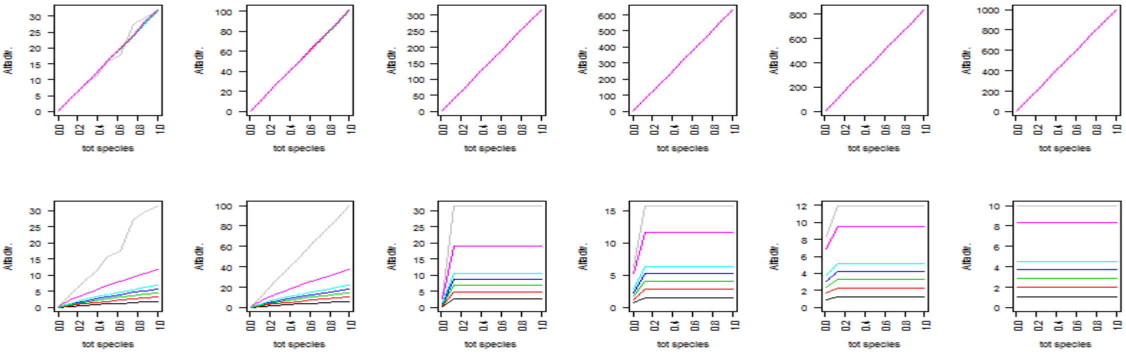


FIGURE 14 DIFFERENCE BETWEEN THE MENHINICK INDEX OBTAINED WITH THE MEASURED NUMBER OF SPECIES S AND WITH THE TEORICAL ONE, UNDER ALL SIX SEQUENCING DEPTH CONDITIONS.

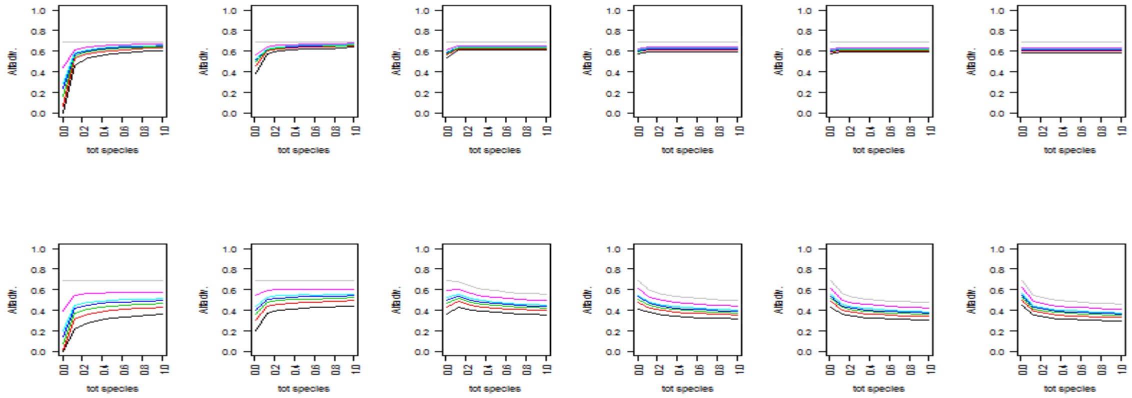


FIGURE 15 DIFFERENCE BETWEEN THE PIELOU INDEX OBTAINED WITH THE MEASURED NUMBER OF SPECIES S AND WITH THE TEORICAL ONE, UNDER ALL SIX SEQUENCING DEPTH CONDITIONS.

2.5 Conclusion

This paragraph reviewed some of the most common alpha diversity indices used in microbiome as well as in ecology studies in order to evaluate samples richness. For each measure we revised its range, its critical point (if possible) and its value when sequencing depth, total number of species or counts abundance distribution is varied. We underline that in most cases S, the total number of measured species, is different from R, the total number of species actually to be found in the sample, and that this undermines the reliability of some of the revised measures. Furthermore we suggest to make careful use of indices with range limited to $[0,1]$, since most microbiome studies show a sequencing depth and a total number of taxa sufficiently high to weaken their discriminating ability, pushing all the sample's alpha diversity values towards 1.

We suggest Hill's number of order 1 in association with S and Fisher's alpha could represent an informative basic subset of alpha diversity measures to be calculated when investigating microbiome samples. Indeed, as we stated above, S describes the total number of species observed, an important factor to be taken into account when exploring alpha diversity, since it affects all the other measures considered. H_1 calculates the equivalent number of species needed to obtain the same Shannon index as the one obtained from the data; since it takes values ranging from 1 to S, by looking at it we can gain an indication of our sample's evenness. In fact, the most its value approaches S the most all the species are evenly distributed, while the most it approaches 1 the most it is likely that one species only is prevailing in abundance over the others. This measure, depending upon the total number of species, might be difficult to use to compare samples with different S; however it uses the same unit of measure of total number of species, therefore is very easy to get some information about species abundance distribution. This last feature has a great impact on alpha diversity evaluation, therefore great caution must be used when dealing with uneven distributed samples. In fact two samples showing the same richness, i.e. the same number of species, might achieve a different alpha diversity value because one has more unevenly distributed abundances that hide rare species, and sequencing depth isn't high enough to counterbalance this effect. Lastly, Fisher's alpha shows a good discriminating ability and focuses on sample's evenness, although it assumes a less significant range of values.

3 Beta Diversity

Beta diversity is a measure, introduced in 1960 by Whittaker [23] of the variation of species composition among two or more sites: it compares diversity between ecosystems by quantifying the amount of species difference between them. Whittaker himself proposed several approaches to quantify species variation [22] [23] and different measures of beta diversity are reported in literature. As a result, beta diversity has become quite disputed.

Originally it was defined to measure diversity along ecologically relevant gradients, like time or space, however often beta diversity has been used as a generic term to describe any of the numerous available indices to measure compositional similarity or dissimilarity. This led to an overall confusion on what are the peculiar features and hypothesis of every single measure, thus preventing both aware choice of what measure to use and reliable comparison between different studies' results. The available indices of beta diversity are not completely equivalent and can, in fact, quantify distinct data characteristics and have different values for the same data set.

In this section our aim is therefore to assess indices properties and test their performance using a simulation which allows us to control for specific features. We tried to eliminate redundancy and favor clear understanding of the pros and cons of each of the reviewed measures, in order to provide the reader with an overview that can help him choose the best measure to use, according to his needs and data characteristics.

3.1 Definition of beta diversity indices

We started from the work of Koleff *et al.* [24], who reviewed 24 measures of beta diversity expressed in terms of "matching components", i.e. shared and unique species between two samples. All these measures have flourished after that the first one, introduced by Witthaker himself in 1960 [22], was proposed in two different versions. Each of these indices was originally justified to adapt to specific needs and to highlight different data characteristics. However it is clear that, starting from the 24 we revised, some of them are redundant having the exact same expression. Therefore we compared and tested all of them in order to narrow down their number to a basic core of indices with desirable properties (Table 2; for the meaning of the terms a, b and c, see Figure 16). In Appendix B we explain the reasoning and methods beyond this subsetting. Here, for sake of clarity, we will always refer to *b* and *c* as number of unique species (of the two sample considered, respectively), to *a* as number of shared species among the samples and to *N*=*a*+*b*+*c* as total number of species (Fig. 16).

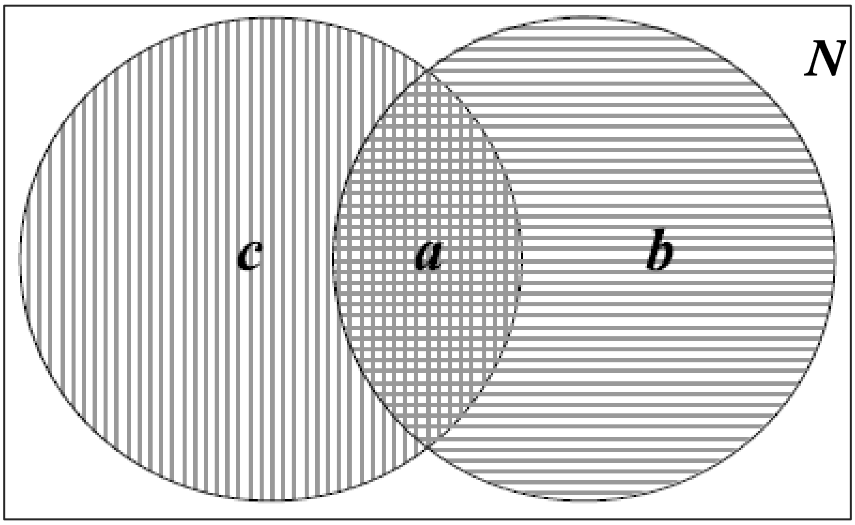


FIGURE 16 MEANING OF A, B, C AND N. **[24]**

- **Whittaker** [22] [23] [40]

This is the first dissimilarity index proposed, expressed as

$$\beta_w = \frac{b + c}{2a + b + c}$$

if the calculation is developed. Here we will retain only the dissimilarity version of this measure, as explained in Appendix B. It is the one complement of the so-called Czekanowsji similarity coefficient, weighting the shared species two times more than the unshared species. Its value is limited to be greater than or equal to 0 and less than or equal to 1, by its very formulation.

- **Cody** [41]

Cody's measure, simply expressed by

$$\beta_c = \frac{b + c}{2}$$

focuses on the number of unshared species found in the two samples. Clearly this measure shows no upper bound, while has its lower bound to be limited to 0 because of *a*, *b*, *c* are all strictly positive.

- **Routledge** [42] [43]

This measure, expressed as

$$\beta_r = \frac{2bc}{(a+b+c)^2 - 2bc}$$

looks at a complex combination of *a*, *b* and *c*. Because of its composition, it will always take values smaller than one and will achieve a value equal to 0 every time one of the two samples does not show any unshared species. Routledge defined two other indices to assess beta diversity, calculated by

$$\beta_I = log(2a+b+c) - \left(\frac{1}{(2a+b+c)} \cdot 2a \cdot log2\right) - \left[\frac{1}{2a+b+c} \cdot \left((a+b)log(a+b) + (a+c)\, log(a+c)\right)\right]$$

and

$$\beta_e = exp(\beta_I) - 1$$

both forced to have a log dependency on at least one of the components *a*, *b* and *c*, while maintaining their value always smaller than or equal to one. In particular, both β_I and β_e obtain a value equal to zero if and only if the two samples are composed by common species only ($a \neq 0$, $b = c = 0$).

- **Magurran** [44]

This measure, expressed as

$$\beta_m = \frac{(2a+b+c)(b+c)}{(a+b+c)}$$

weights the number of unshared species by the sum of species richness found in the two samples, balanced by the total number of species. Clearly, by its own definition, this index hasn't an upper bound, while reaching the minimum value of 0 if and only if the two samples share all the species.

- **Harrison** [45]

This index, proposed by Harrison et al. in 1992, is expressed as

$$\frac{\min(b,c)}{max(b,c) + a}$$

and therefore evaluates the proportion of the smallest amount of unshared species compared number of species found in the other sample. It has range in [0,1] because of its formulation.

- **Colwell & Coddington** [46]

The measure introduced by Colwell and Coddington, expressed as

$$\beta_{cc} = \frac{b+c}{a+b+c}$$

can be interpreted as a simple distance measure deriving from the so-called *simple matching coefficient*, the latter being the relative proportion of species shared among the two samples. This index clearly has value belonging to [0,1], reaching its maximum when there is no common species between the two samples and its minimum when the two samples are identical.

- **Williams** [47]

There are two measures this author proposed in 1996, expressed as

$$\beta_{-3} = \frac{min(b,c)}{a+b+c}$$

$$\beta_{19} = \frac{bc+1}{((a+b+c)^2-(a+b+c))/2}$$

Both of them introduce a nonlinear relation between *a*, *b* and *c*. Since all the three components are coerced to be positives, β_{-3} has range strictly positive but less than one; the same holds for β_{19} although it can never achieve a value equal to zero.

- **Lennon** [48]

The estimate of β_z in terms of matching component is developed by Koleff et al. in their study [24], and is expressed by

$$\beta_z = 1 - \left[\frac{log\left(\frac{2a+b+c}{a+b+c}\right)}{log2}\right]$$

It describes the estimated power law of the ratio between species richness of a larger quadrat with respect to a smaller one. However, since the physical area of the quadrats are involved in defining it, and since these concepts have no equivalent in the microbiome context, we will neglect such interpretation. Whenever the two samples considered share no species, the Lennon index will equal 1, independently of the value b or c show; on the other hand, it will equal 0 when the two samples are identical, i.e. they have all species in common.

In summary, the 13 beta diversity indices to be taken into account are reviewed for completeness in Table 2.

Symbol	Measure re-expressed	Reference
β_W	$\frac{a+b+c}{(2a+b+c)/2}-1=\frac{b+c}{2a+b+c}$	Whittaker (1960), Magurran (1988). Harrison et al. (1992), Wilson & Shmida (1984), Mourelle & Ezcurra (1997), Sørensen (1948) based on Dice (1945); Whittaker (1975), Magurran (1988), Southwood & Henderson (2000) (sim), Harte & Kinzig (1997).
β_c	$\frac{b+c}{2}$	Cody (1975). Weiher and Boylen (1994), Lande (1996).
β_r	$\frac{2bc}{(a+b+c)^2-2bc}$	Routledge (1977), Magurran (1988), Southwood & Henderson (2000)
β_l	$log(2a+b+c)-\left(\frac{1}{(2a+b+c)}\cdot 2a\cdot log2\right)-\left[\frac{1}{2a+b+c}\cdot\left((a+b)log(a+b)+(a+c)\,log(a+c)\right)\right]$	Routledge (1977), Wilson & Shmida (1984)

β_e	$exp(\beta_I) - 1$	Routledge (1977)
β_m	$\frac{(2a+b+c)(b+c)}{(a+b+c)}$	Magurran (1988).
β_{-2}	$\frac{min(b,c)}{max(b,c)+a}$	Harrison et al. (1992)
β_{co}	$1 - \frac{a(2a+b+c)}{2(a+b)(a+c)}$	Cody (1993)
β_{cc}	$\frac{b+c}{a+b+c}$	Colwell & Coddington (1994, “complementarity” measure), Pielou (1984). Gaston et al. (2001), Jaccard (1912), Magurran (1988), Southwood & Henderson (2000) (sim.)
β_{-3}	$\frac{min(b,c)}{a+b+c}$	Williams (1996)
β_{19}	$\frac{bc+1}{((a+b+c)^2-(a+b+c))/2}$	Williams (1996), Williams et al. (1999)
β_{sim}	$\frac{\min(b,c)}{\min(b,c)+a}$	Lennon et al. (2001), based on Simpson (1943)
β_z	$1 - \left[\frac{log\left(\frac{2a+b+c}{a+b+c}\right)}{log2}\right]$	Lennon et al. (2001), Harte & Kinzig (1997)

TABLE 2 SELECTED BETA DIVERSITY MEASURES: SYMBOL, ACCORDING TO KOLEFF ET AL., FORMULATION IN TERMS OF MATCHING COMPONENTS AND REFERENCES . “SIM.” INDICATES THAT THE MEASURE WAS ORIGINALLY DEFINED AS SIMILARITY INDEX.

Starting from a simple analysis of the beta diversity formulae, we can already detect some interesting information: for example we can understand that, by their very definition, all the measures considered but β_c and β_m will have their maximum value to be less than or equal to 1, while the two mentioned indices will have no upper bound. Moreover, we can evaluate whether or not these indices attain their maximum value when *a*=0, meaning that the samples do not share any species (maximum dissimilarity): under this condition only four out of thirteen measures (β_w, β_{cc}, β_{sim}, β_z) show to attain a maximum value equal to 1 (irrespective of the values of *b* or *c*). β_{co} can be calculated under this condition only if the limit for $a \rightarrow 0$ is considered, otherwise NaN is obtained. The other measures depend upon the value that *b* and *c* assume: as particular cases β_{-3}, β_{-2}, β_I and β_e always achieve a value lower than one, while β_c and β_m could attain every value belonging to the range [0, (b+c)/2] or [0, b+c] respectively. We moreover underline that great attention must be paid to limit condition that lead some indices to meaningless values: this is the case of β_I, β_e, β_{co} and β_{sim}. For these measures, if *a*=0 and *b* or *c*=0, a value equal to NaN is returned because their expressions are of the form $0 \cdot (-\infty)$ or $\frac{0}{0}$. For all the measures whose value depend on *b* and *c*, minimum diversity is detected if either of those component becomes zero, except for β_{co}, β_I, β_e.

3.2 Desirable properties of beta diversity indices

Here we define a list of properties we wish a beta diversity measure to fulfill:

- *Independence from the total number of species N*: we would try to measure diversity between two samples without being influenced by the species richness detected. This means, for example, that we would obtain the same beta diversity for two samples sharing half of their species, independently from the total number of species they contain.

- *Range limited to [0,1]*: since we might want to compare two samples using different indices, it might be useful to guarantee that each of them takes value in the same limited range. In particular we would require

  - a maximum value equal to 1 to be gained when two samples show maximum dissimilarity

  - a minimum value equal to 0 to be gained when two samples show maximum similarity

  These properties therefore will be satisfied not only if the measure has values in the desired range, but if it scales accordingly to its upper and lower bound as requested. Generally this properties would need independency from N to be satisfied too.

- *Linear scaling with a/N*: as a consequence of the properties described above, we would like our measure to range between [0,1] following a linear trend depending on the relative proportion of shared species with respect to the total number of species. However, logarithmic and exponential scaling will be considered as well.

- *No need for normalization*: we would require our measure to satisfy our range constraints without needing any form of normalization of *a*, *b*, *c* nor any scaling.

- *Independency form b/c*: this requirement is far less obvious, and depends on the user's need. If you want your measure to attain the same value, independently from the numeric values assigned to *b* and *c*, then measures satisfying this properties has to be preferred. On the other hand, if you want a measure able to distinguish among several conditions sharing the same amount of species, a measure for which this property doesn't hold should be chosen. For example, if two samples share half of the total number of species, you may be interested or not in the composition of the other half of total species, whether they are evenly divided between the two samples or belong to one of them preferentially.

- *Nested samples*: this can be seen as a subcase of the previous one, asking how the measure should behave when two nested sample differ because the number of unshared species varies.

We choose not to list here some basic properties, necessary in order to refer to our indices as measures, to focus the reader on some features we considered more interesting. However, all the listed measures satisfy non negativity, symmetry, triangular inequality and equality for identical samples; most of them, except those depending on N, verify the homogeneity[2] property as well.

We investigated these properties by means of a simulation study[3], aimed at assessing the effect of total number of species N, the effect of relative number of unique species a/N and the effect of variation in b/N under nested condition on our measures' performances.

3.3 SIMULATION STUDY

[2] $\beta(a,b,c)= \beta(2a,2b,2c)$: beta diversity should not be affected if all the matching components are multiplied by the same constant. All measures satisfy this criterion, except β_c, β_m [24]. We may refer to homogeneity as a particular case of independence from total species number N.

[3] See the section *Simulation Study* for details.

We investigated several properties of the retained measures by means of simulations on presence/absence data. We computed each index, using the measures implemented in the *vegan R* package, starting directly from values assigned to *a*, *b* and *c*, tailored to inspect three main aspects that could affect beta diversity analysis. First of all we analyzed the effect of total number of species N (N=a+b+c) when the two samples compared share an increasing percentage of species, ranging from 0 to 90% and have the same number of unique species (i.e. b=c). In Fig. 17 we have a visual description of the simulation conditions while in Fig. 20 we look at the beta diversity value obtained when b and c ratio is held fixed to 1:1, a/N varies from 0 to 0.9 (x-axis) and the total number of species is equal to N=1000/N=2000/N=10000, as highlighted with different colors.

Then we analyzed the effect of relative number of unique species (b and c, Fig. 16) when the two samples compared share an increasing percentage of species, ranging from 0 to 90%. In Fig. 21 we look at the beta diversity value obtained when the total number of species is fixed (N=1000), a/N varies from 0 to 0.9 (x-axis) and when the number of unique species (b and c respectively) are in proportion 1:1, 2:1 and 10:1, here highlighted with different colors. Fig. 18 graphically clarifies the simulation conditions.

Afterwards we analyzed the effect of variation in b and N under nested condition (c=0). Fig. 19 summarizes the simulation setup, while in Fig. 22 we graphically represent in black the beta diversity value obtained when the total species N is held fixed (N=1000) and the number of shared species (*a*) increases, ranging from 0 to 90% (x-axis), while in yellow the beta diversity value obtained when the number of shared species remain fixed (a=10) while N varies from 10000 to 10.

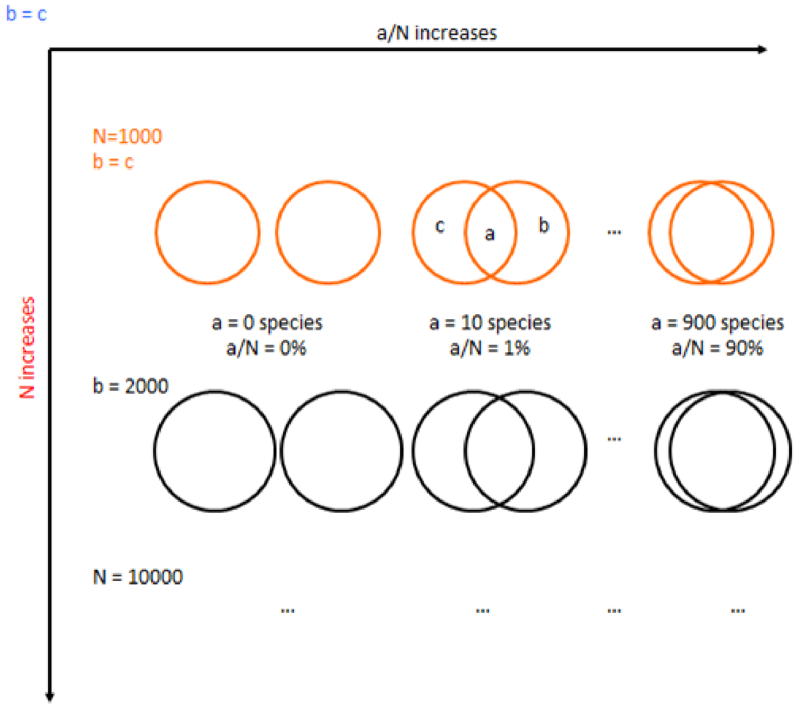


FIGURE 17 SIMULATION CONDITION 1 (IMAGE COURTESY OF DOTT. F. FINOTELLO, PRIVATE CONVERSATION)

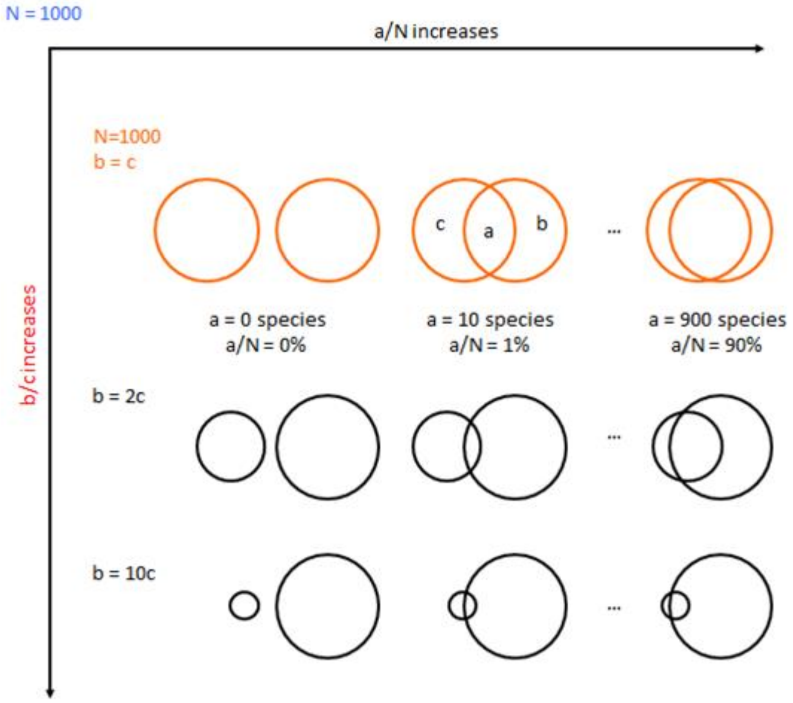


FIGURE 18 SIMULATION CONDITION 2(IMAGE COURTESY OF DOTT. F. FINOTELLO, PRIVATE CONVERSATION)

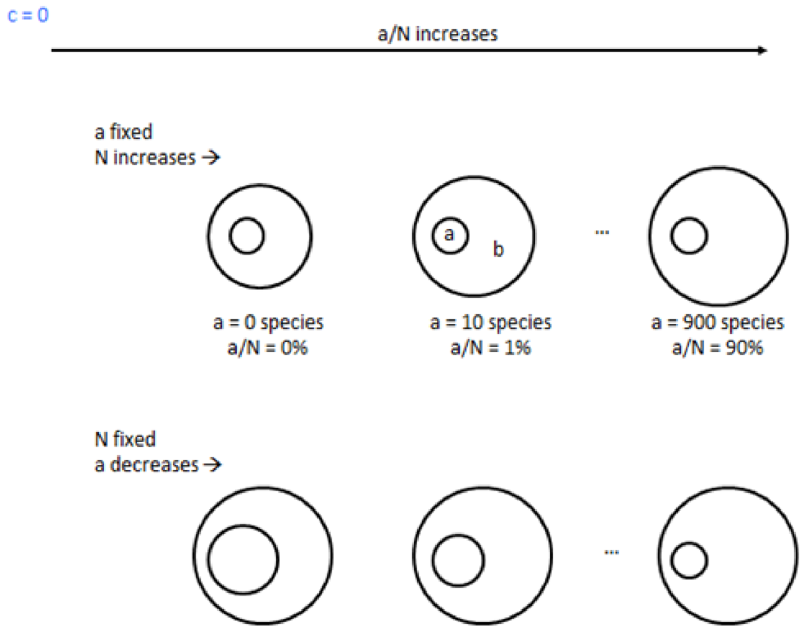


FIGURE 19 SIMULATION CONDITION 3(IMAGE COURTESY OF DOTT. F. FINOTELLO, PRIVATE CONVERSATION)

3.4 RESULTS ON BETA INDICES

This paragraph and the next figures summarize the results of our simulation, showing the beta diversity value obtained for each one of the investigated effects.

Several properties can be investigated by observing Fig. 20, 21 and 22. First of all, we inspect the influence of total species N on the beta diversity values obtained (Fig.20). Clearly, most of the beta diversity indices are insensible to variation in N, except β_m and β_c, that scale accordingly and for which normalization might be required, as we will suggest later on. Moreover, we can probe how beta diversity values change when a/N ratio increases while all the other condition are held fixed. β_c, β_{cc} and β_{-3} linearly decrease when a/N increases, β_m logarithmically decrease while all the other indices exponentially decrease under the same conditions. All of the measures here considered tend to 0 for a/N reaching 1, that is all of them achieve the smallest dissimilarity value possible when maximum similarity is obtained.

Then we examined how beta measures behave when b/c ratio is varied (Fig 21): β_r, β_I, β_e, β_{co}, β_{19}, β_{sim}, β_{-2} and β_{-3} show different trends depending on it, but remain unvaried with increasing sample size. This means that the total number of present species do not influence the β diversity value attained but, on the other hand, the relative proportion of unique species affects it somehow. If the former is a desirable property in β diversity measures, the latter is less desirable, since it suggests that differences in species richness in the samples (measured through α diversity) influences the beta diversity measured among the two samples. Moreover we claim it follows a counterintuitive pattern: the most the two sample have an uneven number of species present, the lower the values attained are, while higher diversity values are shown when the two samples are "even", meaning that they have the same number of unshared species. For example, we report the values obtained with β_r:

b=c	b=2c	b=10c
1.000	0.800	0.198

We argue that in no way two samples might be considered less diverse from each other only because one has a prevalent number of unique species present than the other. Other indices obtain the same value irrespective of variation in proportion among b and c: this is the case of β_w, β_c, β_{cc}, β_m, β_z. Those measures indeed focus mainly on the number of species shared (i.e. *a*) among the samples: therefore their dissimilarity value depends upon the relative proportion of shared species independently of the presence of unique –unshared– species.

Afterwards we examined the behavior of our beta diversity indices under nested condition (Fig. 22). This describes the particular situation in which one sample shows species that are all included in a second sample, but the vice versa doesn't hold true. First of all, we notice that β_r, β_{-2}, β_{-3}, β_{19} and β_{sim} maintain a constant null value under every condition, meaning that all of them focus only on the fact that two nested samples share all the possible species (for the less rich one) and are therefore similar. β_w, β_{co}, β_{cc} and β_z show linear decrease as a/N increases considering the species present in the two samples (although β_{co} does not attain its maximum value in nested condition); in particular they appear to be insensitive of how b decreases (whether *a* increases when *N* is fixed or *N* decreases when *a* is fixed). Moreover, all of them attain high dissimilarity values when the proportion of shared species for the richer sample is small compared to its total number of present species, while showing a value equal to zero if and only if the two samples are actually identical. β_c and β_m linearly decrease when N is held constant and only the proportion of shared species is varied, while exponentially decrease when only *a* is held fixed because of the already proven dependency upon total species number, N. Lastly, β_I and β_e behave peculiarly under nested condition: they show an increasing value for *c* (or *b*, because of symmetry property) less than or equal to *a*, and then a decreasing value if *c* (or *b*) continue increasing over *a*. We further developed a specific simulation to deepen this trend, as shown in Figures 23 and 24.

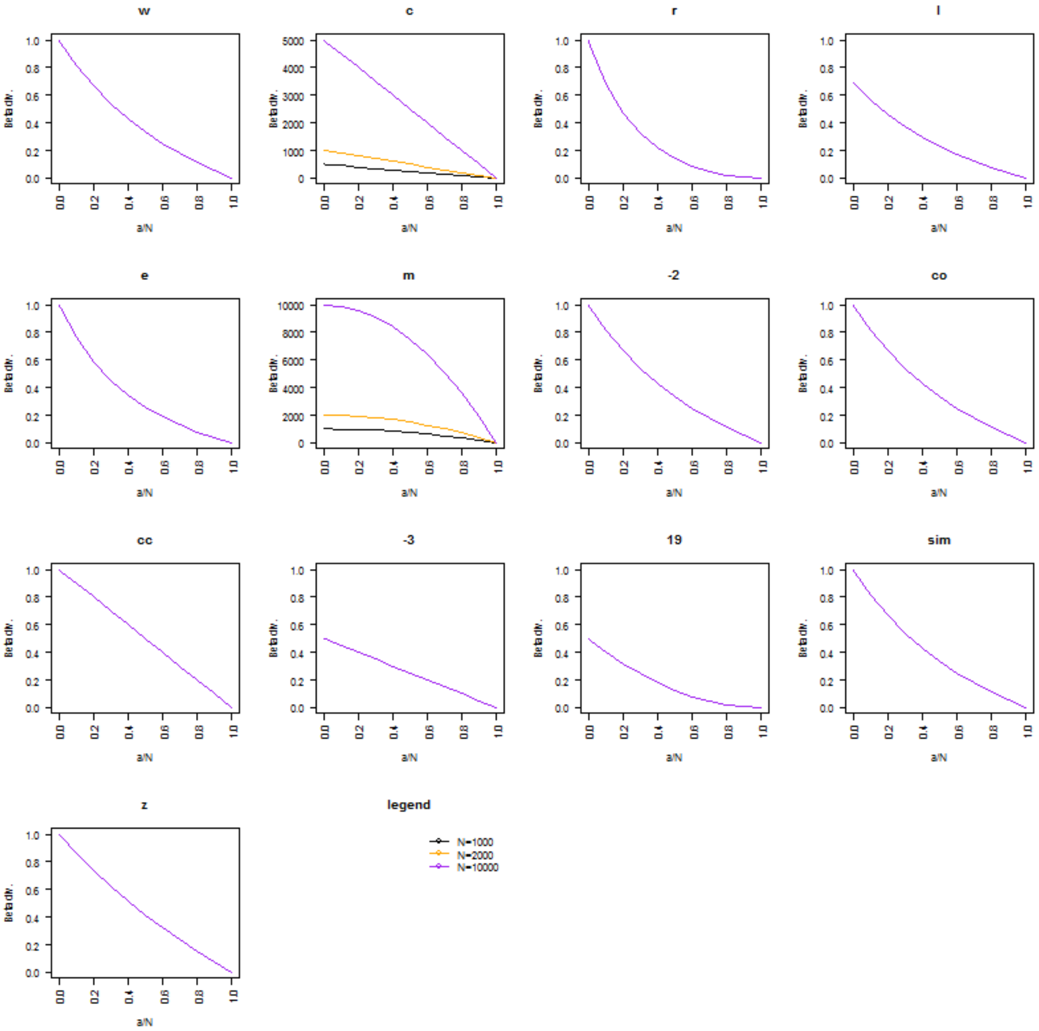


FIGURE 20 EFFECT OF TOTAL SPECIES NUMBER N VARIATION WHEN A/N RANGES FROM 0 TO 1. MAXIMUM AND MINIMUM ACHIEVED FOR EACH MEASURE:

```
                min          max
w          5.26e-02            1
c             5e+01        5e+03
r          5.03e-03            1
I          3.65e-02      6.9e-01
e          3.72e-02            1
m           1.9e+02        1e+04
-2         5.26e-02            1
co         5.26e-02            1
cc            1e-01            1
-3            5e-02        5e-01
19        4.997e-03   4.9995e-01
sim        5.26e-02            1
z           7.4e-02            1
```

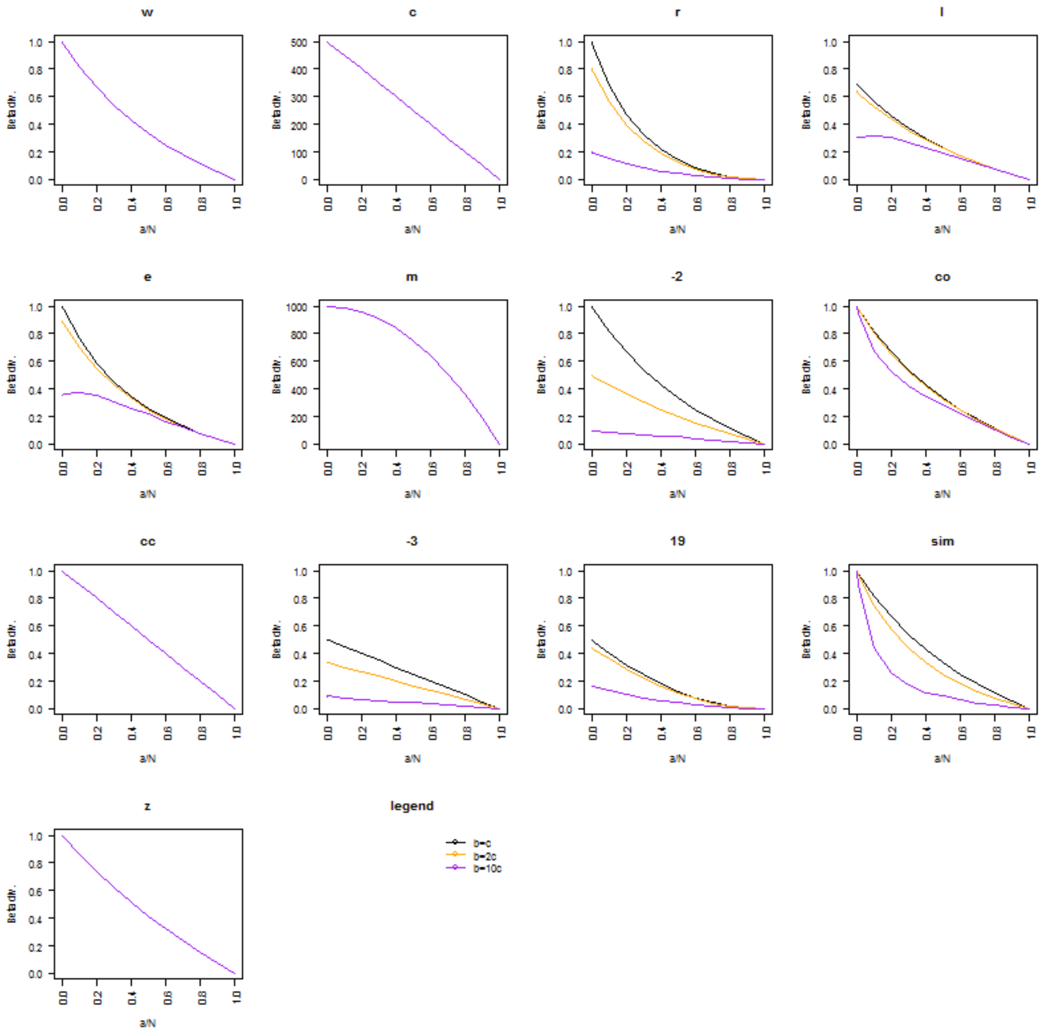


FIGURE 21 EFFECT OF VARIATION IN B AND C RATIO WHEN A/N RANGES FROM 0 TO 1. MAXIMUM AND MINIMUM ACHIEVED FOR EACH MEASURE:

```
              min          max
w        5.26e-02            1
c           5e+01          500
r         1.8e-03            1
I         3.6e-02    0.6931472
e         3.6e-02            1
m         1.9e+02         1000
-2       1.01e-02            1
co          5e-02            1
cc          1e-01            1
-3          1e-02          0.5
19        1.8e-03    0.4995025
sim        1.1-02            1
z         7.4e-02            1
```

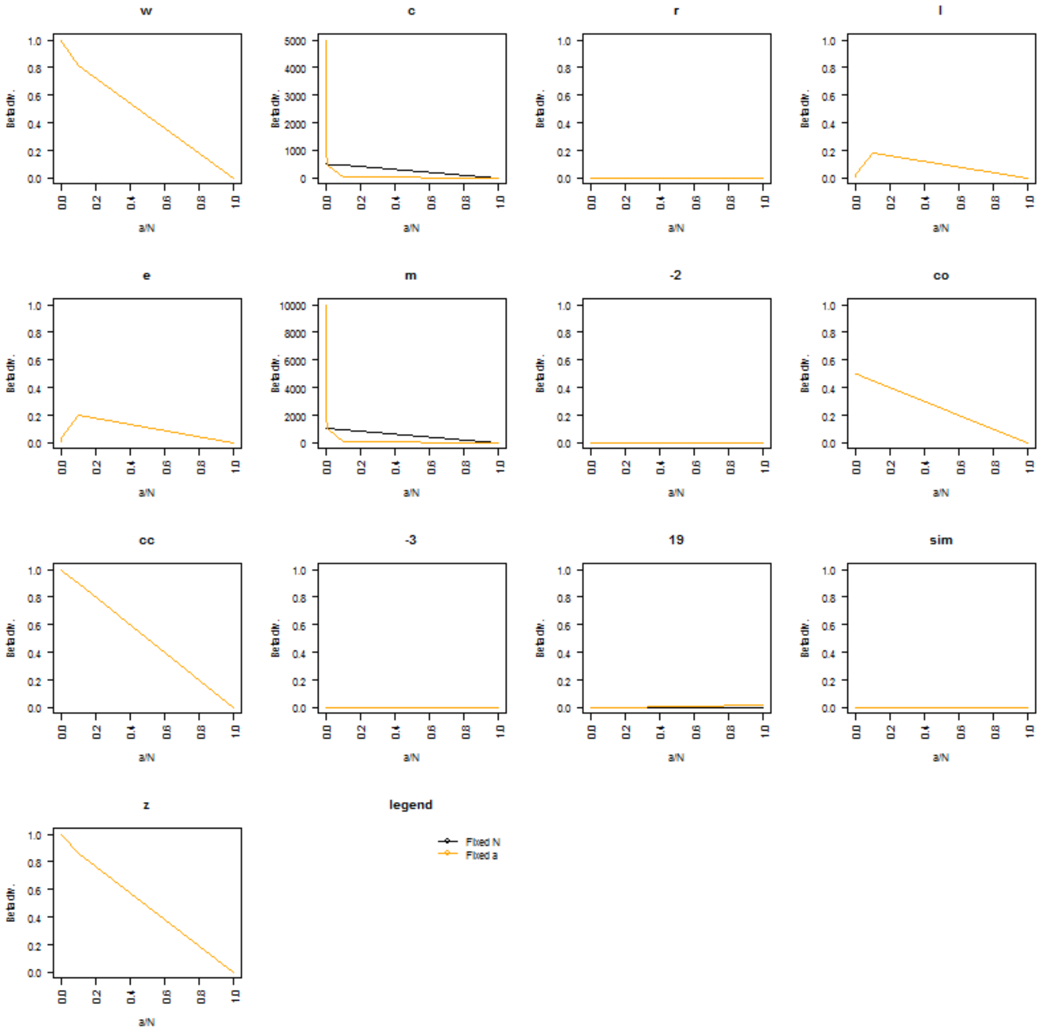


FIGURE 22 EFFECTS OF DECREASING B IN NESTED CONDITION. MAXIMUM AND MINIMUM ACHIEVED FOR EACH MEASURE:

```
              min          max
w               0     9.98e-01
c               0    4.995e+03
r               0            0
I       -4.44e-16    1.786e-01
e       -4.44e-16    1.955e-01
m               0  9.99999e+03
-2              0            0
co              0    4.995e-01
cc              0     9.99e-01
-3              0            0
19    1.9998e-08     1.81e-02
sim             0            0
z               0    9.986e-01
```

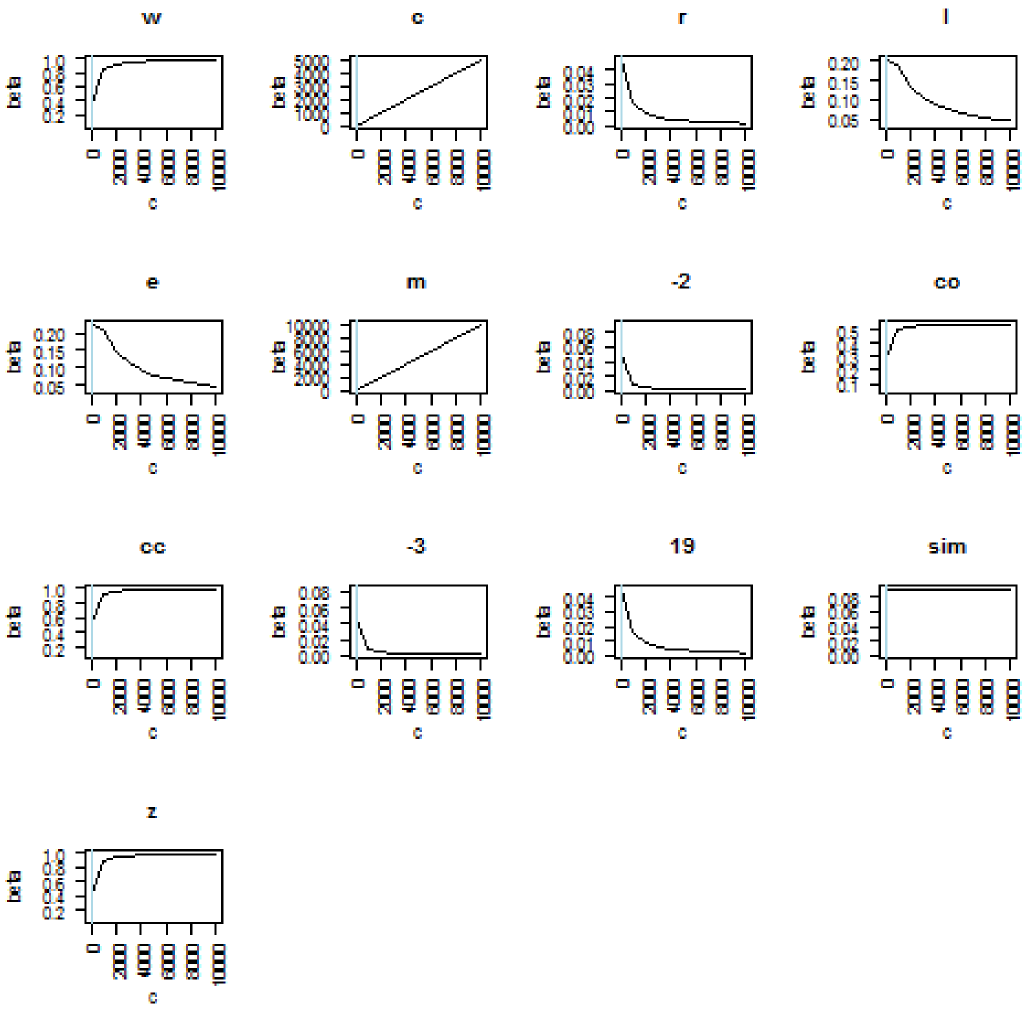


FIGURE 23 BETA DIVERSITY VALUE FOR A=100, B=10 AND C VARYING FOLLOWING X-AXIS. IN LIGHT BLUE HIGHLIGHTED C=A. SINGLE VALUES FOR EACH POINT ARE

```
       beta_1      beta_2      beta_3      beta_4      beta_5
w        0.05        0.09        0.35    8.35e-01     9.8e-01
c         5.5          10          55    5.05e+02   5.005e+03
r      0.0016       0.014      0.0475    1.65e-02    1.96e-03
I       0.035       0.063       0.203     1.9e-01    4.59e-02
e       0.036       0.065       0.225    2.09e-01     4.7e-02
m      20.909       36.67      162.38     1.1e+03    1.01e+04
-2    0.00909      0.0909        0.05    9.09e-03     9.9e-04
co       0.05      0.0909       0.295       5e-01     5.4e-01
cc      0.099       0.167       0.524   9.099e-01     9.9e-01
-3      0.009       0.083       0.048   9.009e-03    9.89e-04
19    0.00177      0.0139       0.045    1.62e-02   1.957e-03
sim    0.0099      0.0909      0.0909    9.09e-02    9.09e-02
z      0.0733       0.126       0.438   8.756e-01     9.86e-01
```

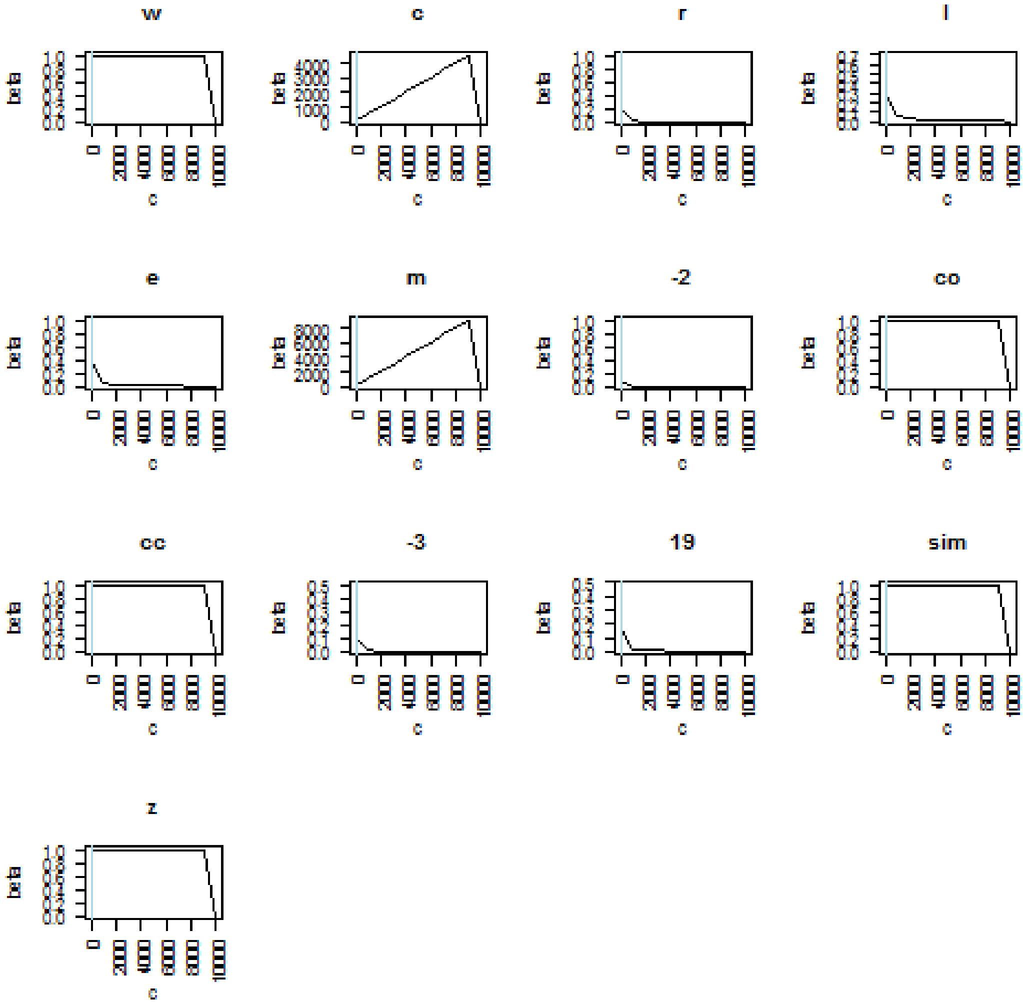


FIGURE 24 BETA DIVERSITY VALUE FOR A=0, B=10 AND C VARYING FOLLOWING X-AXIS. IN LIGHT BLUE HIGHLIGHTED C=A . SINGLE VALUES FOR EACH POINT ARE

```
           beta_1     beta_2       beta_3       beta_4        beta_5
w               1          1            1            1             1
c             5.5         10           55     5.05e+02     5.005e+03
r           0.198          1        0.198     1.99e-02      1.99e-03
I           0.305      0.693        0.305     5.55e-02      7.90e-03
e           0.356          1        0.356     5.71e-02      7.93e-03
m              11         20          110     1.01e+03     1.001e+04
-2            0.1          1          0.1        1e-02         1e-03
co              1          1            1            1             1
cc              1          1            1            1             1
-3         0.0909        0.5       0.0909     9.90e-03      9.99e-04
19          0.167      0.481        0.164     1.96e-02       1.99-03
sim             1          1            1            1             1
z               1          1            1            1             1
```

Here we find that β_r, $\beta_{I,}$ β_e, β_{19} increase their value when $c \leq a$, while begin decreasing if $c>a$; similarly, β_{-3}, β_{-2} increase if $c<a$, and start decreasing as soon as $c \geq a$. All the other indices simply confirmed their dependency/independency form b/c ratio and on $a=0$ or $a \neq 0$.

Lastly, we focused our attention on the range each measure shows under these different conditions. β_m and β_c show a range wider than [0,1] suggesting that normalization is needed if we want to obtain values comparable with other samples having different sample size. β_{-3}, β_I and β_{19} show a range smaller than [0,1], therefore they might need normalization too since no condition can be found in which two samples are more diverse than having no shared species. All the other indices have values always included in [0,1].

Even though different approaches are possible to tackle the normalization problem, here we choose to exploit only two of them:

1. In order to make the measures independent of variation in the total number of species, we re-expressed the terms *a*, *b*, *c* as percentages: $a'=a/N$, $b'=b/N$ and $c'=c/N$. The total number of species is $a+b+c$, therefore $a'+b'+c'=1$.

2. To gain independence from species richness, we divided a, b and c by $2a+b+c$ obtaining a'', b'' and c'': in fact $(2a+b+c)/2$ is the mean number of species present in one of the two samples.

The next figures show the same analysis we conducted before, expressed in terms of a', b' and c'.

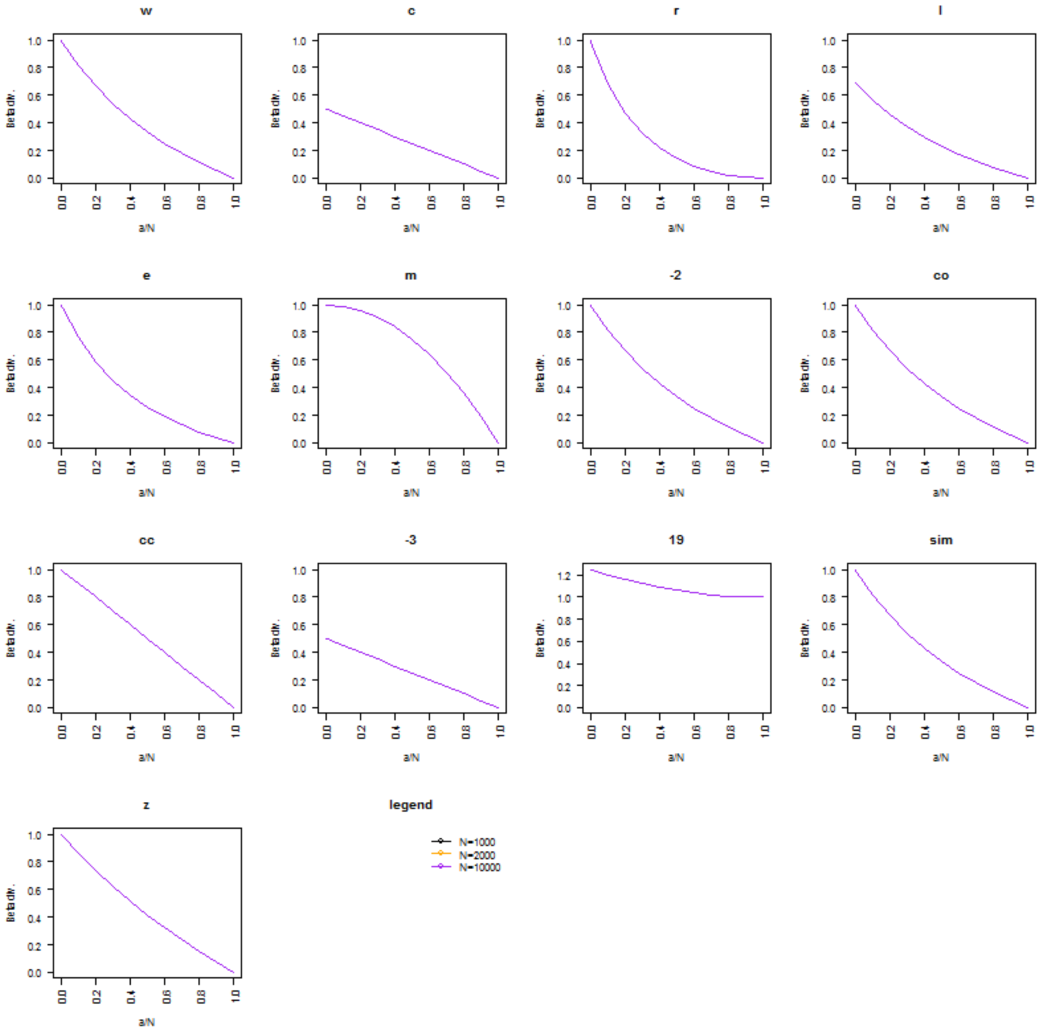


FIGURE 25 SAME AS FIGURE 20, BUT WITH A', B' AND C'.

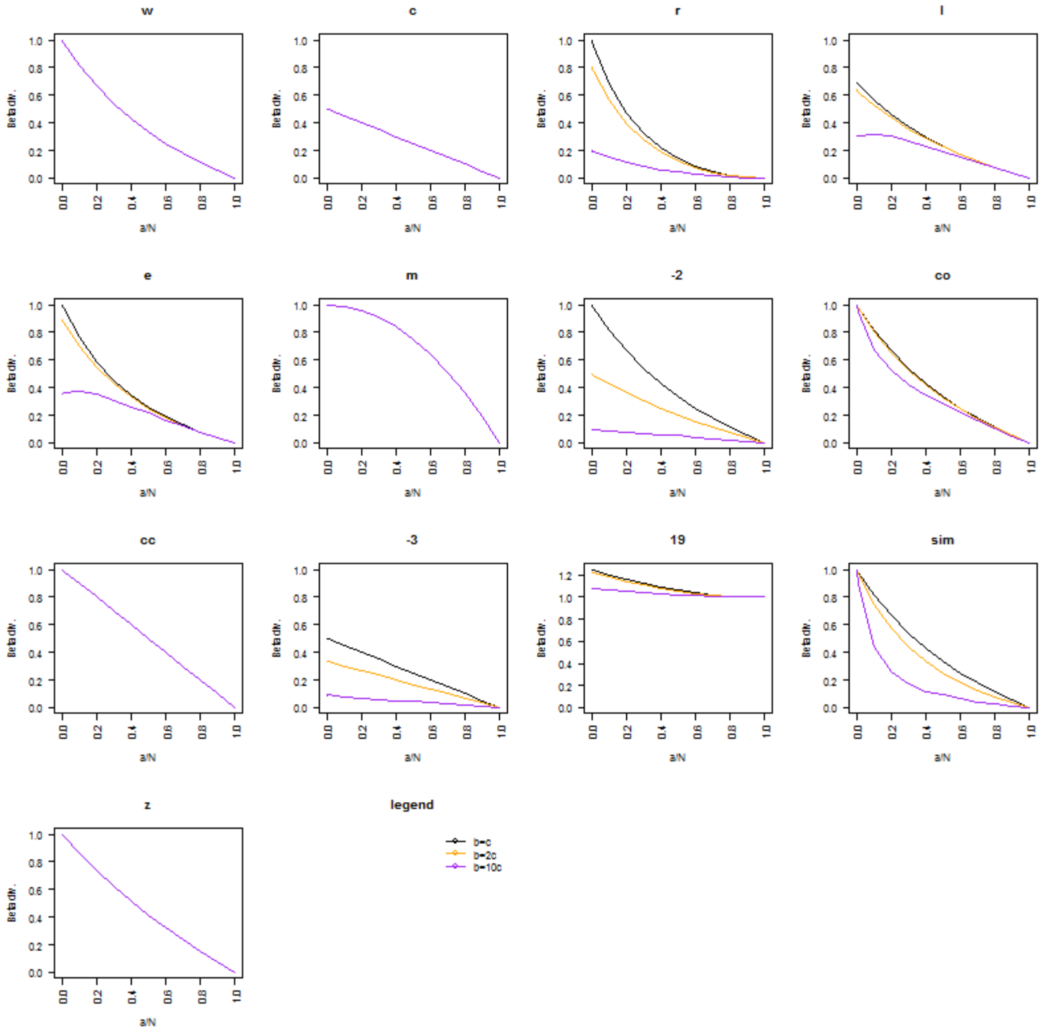


FIGURE 26 SAME AS FIGURE 21, BUT WITH A', B' AND C'.

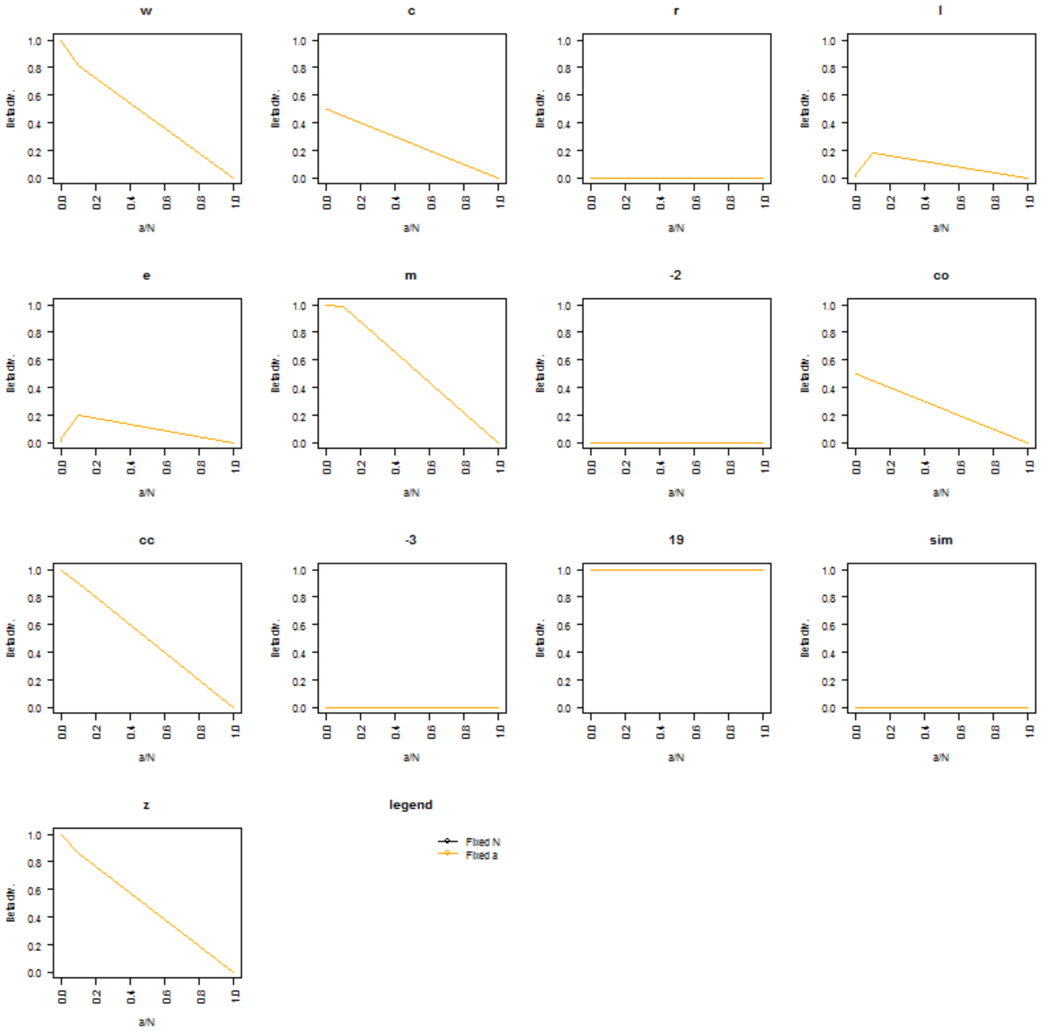


FIGURE 27 SAME AS FIGURE 22, BUT WITH A', B' AND C'.

We notice that measures showing desirable properties without need of these normalization factors remain unchanged by them or worsen their trend, therefore we claim not to apply these approaches by default, irrespective of the indices chosen. However, interestingly, some measure become redundant after normalization: β_c equals $\beta_{cc}/2$ if the first normalization approach is used and $\beta_w/2$ with the second one; β_m equals β_{cc} with the second method but gains no benefit by the first one. Therefore, since comparability among beta diversity measures is a highly desirable property, we could use these equalities to further reduce to eleven the total number of beta diversity indices to use, in order to avoid redundancy. However the other indices to be normalized do not benefit from this approach, partly because of the presence of the logarithm or the minimum/maximum value in their expressions, and are thus to be treated differently. For example β_{-3} can easily be normalized by doubling its value, since its range is limited to [0,0.5]. We tried to scale the other indices value by means of a multiplicative factor calculated under condition of complete dissimilarity of the two samples considered. Anyhow, we are still worried that this might bring to arguable results, hardly comparable on solid basis with the other measures available.

To summarize all the properties reviewed, we refer the reader to Table 3, that we suggest might be useful to revise before choosing to use some of the indices proposed.

	β_W	β_c	β_r	β_l	β_e	β_m	β_{-2}	β_{co}	β_{cc}	β_{-3}	β_{19}	β_{sim}	β_z
Range [0,1]	x		x				x	x	x			x	x
1 for max dissimilarity	x							x	x			x	x
0 for maximum similarity	x	x	x	x	x	x	x	x	x	x	x	x	x
Do not need normalization	x		x				x	x	x	x		x	x
Scales linearly with a/N		x				x	x		x	x			
Independent from N	x		x	x	x		x	x	x	x	x	x	x
Independent of b and c ratio	x	x				x		x	x				x
Nested condition: scales linearly with b	x							x	x				x
Nested condition: constant with b variation			x				x			x	x	x	

TABLE 3 REVISED PROPERTIES AND MEASURES SATISFYING THEM

3.5 Conclusion

This section aimed at obtaining a complete review of the dissimilarity indices proposed in Koleff *et al.* [24] and often used in microbiome as well as in ecology studies in order to evaluate beta diversity among different samples. We succeeded in reducing those measures from twenty-four to thirteen, thus avoiding redundant expressions, and we tested all of them to investigate their properties. We further suggested that some of these measures may benefit from a normalization approach, which could be useful in order to obtain comparable results. Moreover we studied their trend when the total number of species N varies, when the unique species ratio b/c varies, when the relative shared species quantity a/N varies, when the unique species b varies under nested condition.

We suggest the reader, when dealing with beta diversity, to always include at least one among β_{w}, β_{co}, β_{cc} and β_{sim} in their list of tested indices. Indeed, all these measures have a predictable behavior in most circumstances and could be useful to investigate the data characteristics we have focused on, to be used in comparison to other indices in exploring different features, or even to evaluate other measure's performances under specific conditions.

We reviewed several features, summarized in Table 3, that must be taken into account when those measures are used, in order to properly interpret their result. We suggest that all these properties and our remarks regarding some of them could be a useful guide to the user who needs to choose what indices to use among the many available, depending on his needs.

CHAPTER 3:
THE "NORMALIZATION ISSUE"

As we discussed in the first chapter, one of the main drawbacks of using NGS of tag sequences like 16S to explore the microbial composition of samples is that each one may show a different sequencing depth. This means that the total number of mapped reads for each sample may differ from one sample to another, sometimes by several orders of magnitude. As we saw in Chapter 2, sequencing depth has a huge impact in evaluating diversity too, since it undermines our ability to spot all the species present in a sample. Moreover, the presence of rare species or, as opposite, of predominant ones, may impair the ability of sequenced data to detect true species abundance distribution.

The main goal of normalization is to eliminate bias carried by all these effects and to make species abundance comparable between different samples or groups of samples. Indeed, DNA sequencing data consist of discrete numbers, the raw counts, of the equivalent 16S sequence reads, that we use as a proxy of the relative abundance of bacterial individuals in the sample. This means, we are making inferences on the presence/absence and relative abundance of bacterial strains starting from a digital measure, the count itself, which does not describe the exact number of individuals observed. Moreover, it does suffer from several error sources, both connected to experimental practice as well as NGS technology inner features. Therefore, many studies have underlined that careful normalization approaches are need to correct count datasets before applying any downstream analysis.

1 WHY TO NORMALIZE?

Since we know that the total reads per sample can vary broadly within a single sequencing run, what should worry us most is what impact this inner feature of the data might have on our ability to reliably explore and analyze the microbial community they represent. First of all, the major impact of different library sizes lies in the strong correlation that several studies have proven [49] [50] [51] between sequencing depth and number of nonzero species detected in high-throughput 16S studies. For example, Paulson [49] demonstrated that both in the Human Microbiome Project and in the Lung microbiome study the number of species do not stabilizes even for samples with high sequencing depth, indicating that in both cases library size may not be sufficient to obtain a complete profile of the microbial community underneath (see Fig.28). This strong dependency implies as well that, especially in samples with low coverage, zero counts aren't always describing absent taxonomic features; they might be the result of undersampling, while being misinterpreted as absent. Different coverage implies different levels of uncertainty in our data, too. This leads to two different observations: firstly, how can we compare – in terms of absolute values- samples having similar values but different reliability; secondly, if we choose to look at data in terms of relative proportions, how can we ignore that they result from the ratio of values having diverse variability. Fig.29 wants to give a hint of both these problems and to underline the importance of choosing consciously the approach to use.

Moreover most species, in marker-gene studies, are rare (that is, they are absent from a large number of samples). Sparsity of count data is a typical feature, caused by both biological and technical variability. We refer to *biological variability* (BV) to describe the natural variation of biological and physiology parameters due to differences among subject or within the same subject as time passes. Therefore, BV is often divided into two types, namely *interindividual* (differences between subjects) and *intra-individual* (differences in the same subject over time/condition). *Technical variability* (TV) is instead used to describe the intrinsic variability resulting from experimental processes, responsible of differences in parameters found within technical replication of the same biological sample.

The amount of variability affecting a dataset can lead to a strong bias, especially when data are scaled for comparison and tested for statistical differences. Indeed, even if variation in read counts between technical replicates has been often adequately modeled as a Poisson random variable [52] [53] [54] [55], what interest us most when making inferences on population distribution is the variation of features among biological replicates.

We consider counts n_{ik} as a raw signal describing the level of presence of the species i in sample *k;* it really represents the number of 16S sequenced reads assigned to the taxonomic level observed. We suppose that:

- there are N (unknown) number of species in sample *k*;
- the species *i* appears with T_{ik} individuals (unknown) in the sample;
- the DNA fragment sequenced have on average length *l*;
- the corresponding 16S fragment has fixed length L_i.

Since the sequencing process can be modeled as a random sampling process [56], then the probability that one read is assigned to the *i*-th species equals

$$p_{ik} = \frac{T_{ik} \cdot L_i/l}{\sum_{j=1}^{N} T_{jk} \cdot L_j/l}$$

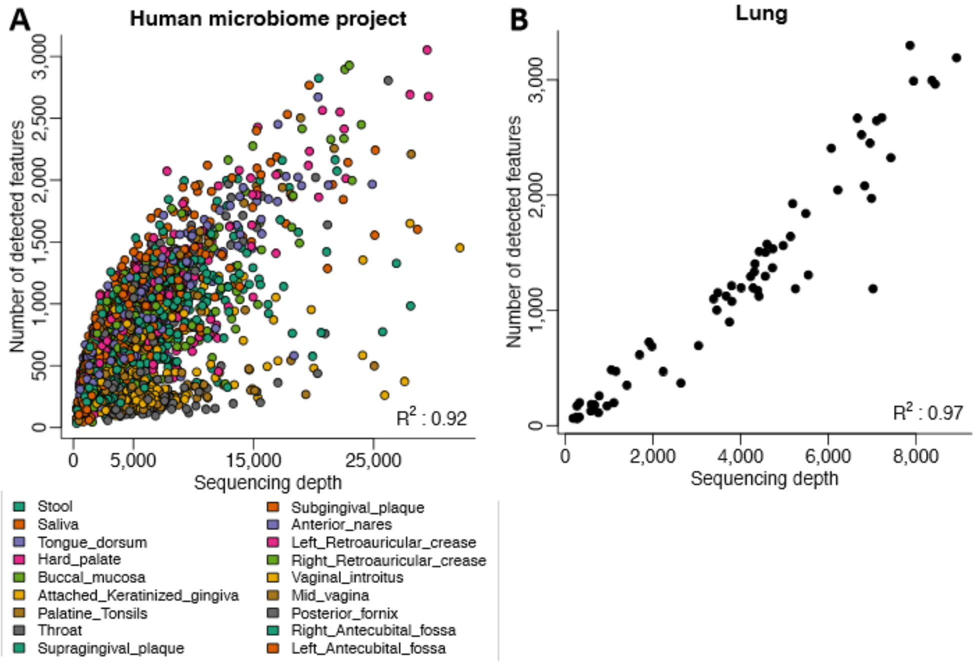
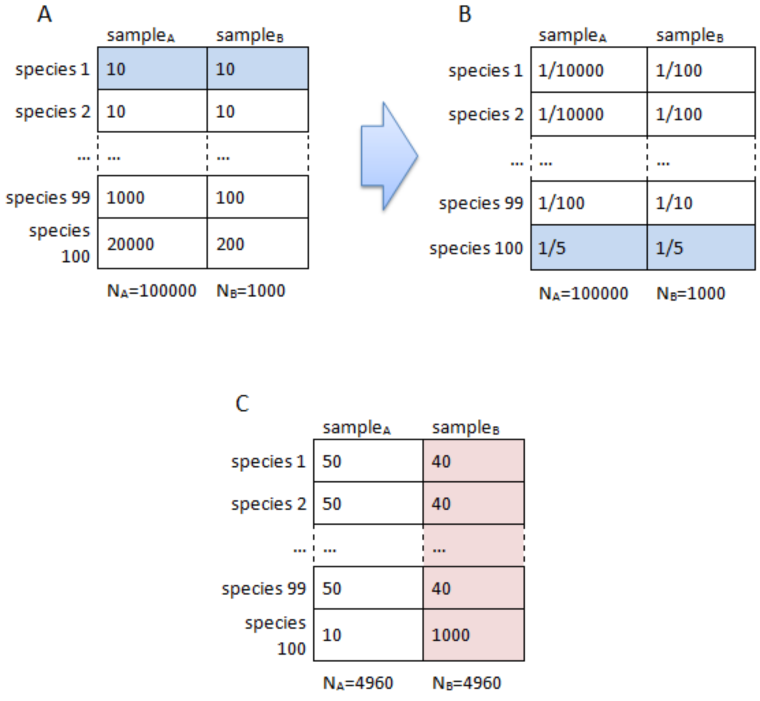


FIGURE 28 THE NUMBER OF OTUS DETECTED IN A SAMPLE DEPENDS ON SEQUENCING DEPTH. FROM **[49]**



A

	sample$_A$	sample$_B$
species 1	10	10
species 2	10	10
…	…	…
species 99	1000	100
species 100	20000	200
	N$_A$=100000	N$_B$=1000

B

	sample$_A$	sample$_B$
species 1	1/10000	1/100
species 2	1/10000	1/100
…	…	…
species 99	1/100	1/10
species 100	1/5	1/5
	N$_A$=100000	N$_B$=1000

C

	sample$_A$	sample$_B$
species 1	50	40
species 2	50	40
…	…	…
species 99	50	40
species 100	10	1000
	N$_A$=4960	N$_B$=4960

FIGURE 29 DIFFERENT EFFECTS OF SEQUENCING DEPTH ON COUNT DISTRIBUTION: A) SAME ABUNDANCE BUT DIFFERENT RELIABILITY, B) SAME PROPORTION, BUT DIFFERENT VARIABILITY, C) ONE ABUNDANT SPECIES "CONSUMES" MOST OF THE SEQUENCING DEPTH

where, for microbiome data, the L_i/l ratio can be neglected since there is no length bias. Then, the probability that n_{ik} read map to species *i* in sample *k* is obtained from a binomial distribution as

$$P(n_{ik}) = \binom{N_k}{n_{ik}} p_{ik}^{n_{ik}} (1 - p_{ik})^{N_k - n_{ik}}$$

where N_k is the total number of read in sample *k*, or correspondingly its sequencing depth. However, if N_k is high and p_{ik} is low, we know that this distribution can be approximated by a Poisson distribution having parameter $\lambda_{ik} = p_{ik} \cdot N_k$.

We allow our dataset to contain different biological samples belonging to the same experimental group; therefore p_{ik} is a random variable depending on the parameter λ_{ik} which is itself a random variable, having its mean and variance. This means that the abundance of species *i* is not the same across different biological replicates, causing the count data variability to be over-dispersed [57]: Poisson distribution cannot explain this additional dispersion, thus Negative Binomial distribution models are preferred [57] [58] [59].

The expected value for p_{ik} is then

$$E[p_{ik}] = p_{ik} \cdot N_k = \frac{T_{ik} \cdot \frac{L_i}{l}}{\sum_{j=1}^{N} T_{jk} \cdot \frac{L_j}{l}} \cdot N_k \qquad (*)$$

meaning that the average value of raw counts of species *i* in sample *k* is proportional (but not equal) to the true number of individuals T_{ik}. Indeed, the measured number of count n_{ik} depends upon several factors, included:

- the sequencing depth;
- the abundance of species *i* in sample *k*;
- the abundance of the other species found in sample *k*.

A latent problem underlying microbiome count data set is how counts are distributed, within a sample, among different species. Indeed, it is easily understood that samples having a prevalent species –and being analyzed with a finite sequencing depth- are more likely to detect many reads belonging to the overabundant strain and to neglect other, rarer, species. This may lead both to underestimate the abundance of some taxa and to leave some others completely undetected, being considered as absent. Fig. 29 (c) shows this case.

Often the denominator of eq. (*) is referred to as *size factor S_k*. It is clear that it plays a role in introducing a systematic bias when comparing two samples: as a result, normalization techniques are required to correct different size factors and make count data comparable. The aim of normalization is indeed to remove systematic technical effects in the data, thus ensuring that technical bias impact is minimized in favor of true biological variation. The hypothesis needed to estimate the systematic bias and normalize data are that we are monitoring a high number of different species and that most of the samples belonging to the same experimental group have species abundance being similarly expressed, both easily verified when dealing with microbiome datasets.

Many different normalization approaches are available, ranging from simple scaling to more complex methods; in the next sections we will describe the state-of-the-art techniques and discuss what approaches we have decided to investigate in our work.

2 Review of state-of-the-art normalization techniques applied to microbiome

Since no agreement is found on what normalization method have to be preferred when dealing with microbiome data, we decided to compare and evaluate the performance of six different approaches. Some of them are inherited from RNA-Seq pipelines, because of the intrinsic similarity of data features between these two fields, and some of them have already been evaluated in recent microbiome studies [60] [49].

- *Total sum scaling (TSS)* or *global scaling* [52]: it is the most common normalization technique. It simply divides raw counts by the total number of reads found in the sample (the sequencing depth), i.e. it transforms counts in the corresponding proportion within the sample by simply computing $p_{ij} = n_{ij}/N_j$. Because of its simplicity, this method is not implemented in any specific R function.
- *Trimmed mean of M-values (TMM)* [61]: in order to account for differences in sequencing depth between samples and to reduce the impact due to highly abundant species, TMM removes a percentage of the features showing the highest (in absolute value) log-fold-changes between the compared samples. Then it evaluates the normalization factors to align columns of the count matrix, and uses them to correct all data. Trimmed mean is implemented in the *edgeR* package [62] [63] [64] [65] [66] function *calcNormFactors.*
- *DESeq size factors* [58]: this scaling approach uses robust regression to estimate the size factors for each column. Each sample is divided by the

geometric means of the rows; then the median (or another location estimator specified by the user) of these ratios, skipping the genes with a geometric mean of zero, is used as the size factor for this column. This method uses the so-called "relative log expression", proposed by Anders and Huber in 2010, to compute the size factor to be used. It is implemented by calling the function *estimateSizeFactors* of the *DESeq* package.

- *cumulative sum scaling (CSS)* [49]: this method extends the precedent idea of quantile normalization proposed by Bullard *et al.* [53] [67] by looking for a data-specific quantile to use in order to normalize data in a coherent way. This is computed in a two-step procedure: first of all the percentile for which to scale by is calculated, then the cumulative sum scaling factor is applied to the dataset. If we denote the *l*th quantile of sample j as q_j^l(meaning that in sample j there are l species with counts smaller than q_j^l), and the sum of counts in sample j up to the *l*th quantile as s_j^l, then we look for a particular value $\hat{l}$ to produce our normalized counts as

$$\widetilde{n_{\iota J}} = {}^{n_{ij}}\!\big/_{s_j^{\hat{l}}}$$

(eventually scaled by a common constant so that normalized counts have interpretable units). $s_j^{\hat{l}}$ should be the median scaling factor across all samples. The choice of $\hat{l}$ is implemented in an adaptive, data-driven way: if we suppose that, up to this quantile, counts are derived from a common distribution, we can find $\hat{l}$ as the value for which sample count distribution deviates from the reference. In details, if $\overline{q^l} = med_j\left(q_j^l\right)$ is the median *l*th quantile across all samples, it can be used as the *l*th quantile of the reference distribution [68], and $d_l = med_j\left|q_j^l - \overline{q^l}\right|$ is the median absolute deviation of sample-specific quantile from the reference. Since d_l is stable for low quantiles but tends to instability as l grows, we choose $\hat{l}$ to be the smallest value for which instability is detected, in terms of relative first differences. Lastly, since CSS-normalized samples are well approximated by a log-normal distribution, a log transformation to the normalized count data is applied. It is worthwhile noting that, if the function is not able to detect an appropriate quantile for which to scale, or the identified quantile is too low, then the conventional quantile scaling is performed, with 75th quantile threshold. Cumulative sum scaling is implemented in the *metagenomeSeq* package [49], with the two step obtained using *cumNormStat* and *cumNormMat* functions respectively. As a side note, relative proportion of species remains unaffected by this normalization procedure.

- *rarefying*: rarefaction is not really a normalization method, although it is widely used if it is so. It actually is a subsampling procedure that aims at reducing all the samples at a common sequencing depth, usually the smaller one available. We used the rarefaction method wrapped in the *phyloseq* package's function *rarefy_even_depth* [69]. This method uses random subsampling to extract values from the actual sample data until the wanted library size is reached; for computational reasons sampling with replacement is the default approach, although the original idea proposed by Hurlbert [70] used a without replacement approach. We therefore implemented another version of it, more similar to the one wrapped in QIIME, which repeats the subsampling procedure without replacement several times (5 by default) and for different library sizes, specified by the user.

Two recent papers, by McMurdie and Paulson [49] [71] addressed this very same problem by testing several normalization approaches on real and simulated data, in order to assess their properties and their performance when applied to different microbiome-specific problems.

Joseph N. Paulson *et al.*, in his study [49] proposed as a new normalization technique the *cumulative-sum scaling* (CSS), implemented in the *metagenomeSeq* package, which is able to deal with under-sampled data, a common feature of microbiome datasets. To assess how well it performs, they compared it to the common-used normalization technique of *total-sum scaling* (TSS), DESeq size factors and trimmed mean normalization. TSS had already shown in RNA-seq experiments [53] [68] to add bias that prevent correct differential abundance estimates. CSS aim is to remove biases in count data matrix introduced by features being preferentially amplified in a sample-specific manner.

CSS proved to be the best method, among the tested ones, to separate samples belonging to a longitudinal study of gnotobiotic mice gut microbiome, according to diet on a multidimensional scaling analysis. Methods performance were assessed using the 1000 features with larger variance after normalization, on both a MDS and linear discriminant analysis, and then log ratio of class posterior probabilities were calculated using leave-one-out cross validation. CSS outperformed DESeq, TMM and TSS, also allowing CSS-normalized sample abundance to be better approximated by a log-normal distribution.

The second contribution of this study was a novel distribution mixture model, the so-called *Zero-Inflated Gaussian* (ZIG) distribution, whose main purpose was to explicitly account for undersampling when testing for differential abundance among sample groups. This model tries to mitigate bias in this kind of tests derived from the over-abundance of zero-counts species due to undersampling of microbial communities. Its performance was tested both on simulated and real data (the latter related to oral microbiome extracted by the Human Microbiome Project), and compared to the results obtained with others metagenomic tools like *Metastats*, *edgeR*, *DESeq*, *LEfSe* and *Xipe*. The insuccess of the other models investigated was attributed by the authors to the lack of robust modeling in *Metastats* and *LEfSe*, and the impossibility to meet the hypothesis upon which *DESeq* and *EdgeR* models are based. Therefore this new method, justified by the observed relationship in microbiome data between sequencing depth and number of non-zero species detected, was found to be able to estimate the probability that an observed zero is caused by undersampling or derives from the true absence of the species in the microbial community.

Paul J. McMurdie, in his work with Susan Holmes in 2013 [71] focuses on explaining why the common approach called *rarefaction* is statistically inadmissible. They indeed claim that it requires to throw away available valid information resulting in loss of statistical power when rarefied data are investigated to identify differential abundance[4]. Rarefying proofs to be inadequate in two different ways: first of all, it equalizes the variance between all the samples by imposing them to be equal to the worst value among them, thus increasing uncertainty; moreover, it adds additional uncertainty because of the random subsampling step, in which we lose data information.

Similarly, they criticize another common method, *total sum scaling*, because this approach does not take into account the *heteroscedasticity* problem. In other words, it is incorrect to compare the species simple proportions $p_{ij} = n_{ij}/N_j$ without taking into account the difference in the denominator (the sequencing depth of sample j) and the variance related to it.

In order to support their statements, they built up two different simulation, in which they evaluated the impact of different normalization approaches on the calculation of sample-wise distances and on differential abundance analysis. For both simulations, they used real microbiome count data from the *Global Patterns* dataset, repeatedly subsampled and ad hoc modified. In the first one, DESeq variance stabilization

[4] This term refers to, analogously to differential expression from RNA-seq, looking for statistically significant differences between the mean abundance of a species between two or more sample classes.

method[5] and edgeR upper-quantile log fold change normalization[6] were added to the two methods discussed earlier. For each of the normalization approaches used, sample-wise distance evaluation was then calculated using *Bray-Curtis*, *Euclidean*, *Poisson*, *top-MSD* and *UniFrac* distances. Unsupervised classification was then performed for each combination of experimental condition, looking at the proportion of simulated samples that were consistently assigned to their original sampling spot (*Ocean* and *Feces*, respectively), despite the modifications (mainly, mix between samples) applied.

Conversely, the second simulation aimed at detecting differentially abundant species between two classes (here, *target* and *non-target*). This was achieved by artificially perturbing one class by means of a fixed effect size, since both of them derived by random sampling from the same source environment of the *Global Pattern* dataset. In this way, it was possible to evaluate how well the species artificially inflated were identified, while accounting for false positives. For each simulated experiment the following statistical tests (all corrected for multiple comparisons) were applied: two-sided Welch version of t-test, exact binomial test, negative binomial test, negative binomial Wald test, an estimate of the posterior probability using a Zero Inflated Gaussian mixture model.

According to their analysis, McMurdie and Holmes demonstrated that rarefying count data or using proportions both undermines the performance of clustering methods and statistical tests, with the latter suffering from an unacceptably high false positive rate, with results worsening with effect-size[7]. In detail, count proportions tended to perform better than rarefaction, but showed a higher FP rate when the effect size was large. This undesired effect was independent of the other analysis features; it was common to TMM or DESeq normalization too, but not to RLE normalization approach. Interestingly, rarefied count suffered from an increase in both Type II (decreased sensitivity) and Type I (decreased specificity) error; those effects are linked to the added uncertainty brought along by random subsampling. However, the same analysis proved that modelling counts with a Negative Binomial

[5] This normalization approach transforms the count data (by dividing them by the size factor), yielding a matrix of values which are now approximately homoskedastic. It is obtained by calling the DESeq function *getVarianceStabilizedData.*

[6] This method adjusts counts so that the effective library sizes are made equal, but preserve fold-changes between groups and biological variability within each group. It is obtained by calling the edgeR function *equalizeLibSizes.*

[7] Here the authors refer to effect-size as the fold-change in species abundance artificially added among the true-positive OTUs.

with RLE normalization might be an effective approach: it was able to accurately and specifically identify differential abundance under every simulation condition. The newly proposed zero-inflated Gaussian mixture model, implemented in *metagenomeSeq* performed well too.

3 A NEW METHOD: ZERO IMPUTATION

As already noted above, differences in sequencing depth could affect our ability to sample rare species. This effect, however, cannot be corrected by scaling data using a normalization factor, because absent species would not benefit from it. Therefore, we propose a method that tries to correct specific absent species, after TMM normalization has been applied to scale data. This approach follows several steps:

- detect samples having zero minimum value, identify their sampleID and their group
- for each of them
    - spot the species showing zero abundance (if less than 5, skip the correction for this subject)
    - extract a subset of the data, composed by samples belonging to the same class of the sample analyzed and by all the species having nonzero abundance in the sample of interest
    - evaluate the distance matrix between the columns of this submatrix
    - order the samples according to increasing distance from the sample analyzed
    - extract the IDs of the four nearer samples
    - extract another subset of the data, composed by the four nearer samples and only the species showing zero abundance in the sample of interest
    - evaluate the weighted mean of the rows of this submatrix, with weights being the sequencing depths of each sample
    - replace only the zero abundant species of the sample of interest with the corresponding weighted mean.

The underling intuition is that similar samples, i.e. having similar patterns of count abundance distribution, may give us cross-information to distinguish among true zeros (absent species) and false zeros (rare species remained undetected). Therefore, it searches for a subset of similar and trustworthy samples from which to infer the possible abundance of the species. Their reliability is weighted using their sequencing depth, since higher library size have the highest probability to detect rare species.

4 METHODS ASSESSMENT

We assessed the impact of the different normalization approaches by applying them to a simulated dataset, whose structure will be discussed in the next chapter. Here we only report that it permitted us to disentangle technical and biological variability, and therefore allowed us to monitor to what extent the normalization approaches were able to suppress the former, leaving the latter unaltered. Indeed, we had the possibility to compare normalized count data with the individual abundances found in the corresponding sample. We performed such a comparison by evaluating:

1. Ecological distances, between the original samples and the normalized ones: in detail, we computed
   a. *Euclidean Distance*: even if it is not an efficient measure of similarity in ecological context, we used it as a reference measure. It is simply expressed by

$$d_j = \sqrt{\sum\nolimits_i (c_{ij} - s_{ij})^2}$$

where *j* is the sample analyzed, *i* indicates the species, *c* is the normalized count while *s* represents the number of individuals found in the sample.

   b. *Bray-Curtis distance*: this distance, computed as

$$d_j = \frac{\sum_i |c_{ij} - s_{ij}|}{\sum_i (c_{ij} + s_{ij})}$$

is a semimetric [8] distance measure commonly used to quantify differences between samples based on abundance data, whose value range in $[0,1]$.

   c. *Canberra distance*[9]: this distance, obtained as

[8] We refer to a measure as semimetric, if it is a function that satisfies the first three axioms, that are

- non-negativity: d(x,y)≥0
- coincidence axiom: d(x,y)=0 if and only if x=y
- symmetry: d(x,y)=d(y,x)

but not necessarily the triangle inequality (d(x,z)≤d(x,y)+d(y,z)).

[9] This is actually a normalized version of the original distance, here reported as it is computed by the *vegdist* function of the *vegan* package in R.

$$d_j = \frac{1}{NZ} \sum_i \frac{|c_{ij} - s_{ij}|}{(c_{ij} + s_{ij})}$$

(where *NZ* represents the number of nonzero entries) is a metric used to compare ranked lists of elements, having its range in $[0,1]$.

We computed each of these measures directly on raw data, on data proportions (that is, dividing each sample by the total number of counts/individuals found in it) and on ranked data. The first approach is of course pretty weak, since we do not expect to find the exact same values in samples having different unity of measures (individuals and counts). Both proportions and ranking look for a more general comparability, namely searching for relative abundances or at least scale ordination to remain unchanged between normalized counts and original individual abundances. In this context, the lower the value, the better the normalization approach performance, since it describes the ability of the method to reduce the added technical variability and to let count data describe directly real microbial abundances.

2. Alpha diversity: as we explained in the previous chapter, our ability to assess both data richness and evenness depends on several factors, among which the possibility to detect rare species and to resembles as much as possible real individuals abundance distribution. Based upon the review we conducted, we decided to evaluate the following measures: Fisher alpha, the measured number of species S, Simpson and Inverse Simpson indices, Hill's number of order 1. We will compute them for each normalized dataset, and we will compare their value with the ones obtained directly on real data, by means of two statistic tests: paired t-test and Wilcoxon rank sum paired test.

3. Beta diversity: we computed samples' diversity both within a single body site and among two different sites. Starting from the review carried out in the previous chapter, we evaluated β_w, β_{co}, β_{cc} and β_{sim}. Again, we will compare beta diversity values obtained from normalized dataset with the corresponding values computed on the biological replicates of our community. Statistically significant differences will be assessed using both the t-test and the Wilcoxon rank sum test in their paired version.

5 Results

First of all, we look at the ecological distances found between real and normalized data. When applied to compare sample's proportion with mean proportion found in the real biological replicates, Canberra distance always achieves his minimum value if the normalization approach using TMM is followed by the zero imputation procedure we have described. Bray-Curtis distance confirms this trend, even though sometimes DESeq Size Factor seems to outperform the zero imputation. Euclidean distance suggests a similar conclusion, even if a better performing measure is more difficult to be univocally determined in this case, because most distance values are really close to each other. In some (yet rare) cases rarefied sample's proportion manage to be particularly close to real data; however, this appears to be the exception rather than the rule, since most of the times rarefied data show very poor performances. DESeq size factor often shows to achieve proportions near to real ones, even though it outperforms zero imputation only in some specific samples. Most of the others normalization approaches do not improve in any way the distances found between real and normalized data proportions, since TMM, TSS and CSS simply leave relative abundance unaltered. However, when a paired comparison is performed in which real proportion are directly matched with normalized ones, less accord emerges from the three measures used. Indeed, both Canberra and Euclidean distance agree in defining trimmed mean with zero imputation the most effective way to draw proportions nearer to real ones, while Bray-Curtis seems to suggest that one of the equivalent methods between TSS, TMM and CSS achieve better performances.

Analogous conclusion might be drawn when evaluating distances on ranked proportions of data: in every case the trimmed mean and zero imputation combination outperforms all the other normalization techniques. In Fig.30 we reported an example, showing that on average the distance obtained with other normalization approaches is always greater than the one obtained with the best performing method. In general, therefore, zero imputation seems to have a positive influence in shifting normalized data proportion (or ranks) nearer to real data ones.

Then we compared alpha diversity (both richness and evenness) found in real and normalized data. Again, we firstly evaluate how normalization approaches perform when compared with the average value obtained from real data. Then we performed a pairwise comparison between all the alpha diversity values measured on real and normalized data. In addition, this time we assessed results comparability using two statistical tests: the t-test and the Wilcoxon test. Here the obtained results widely vary depending on the chosen index. Indeed, both Fisher's alpha index and the observed number of species S benefits from the zero imputation, while all the other indices (Simpson, Inverse Simpson index and Hill's number of order 1) seem to show results nearer to their true value when other normalization methods are employed. Statistical test results confirm that the measured species S detected on true data distribution and on normalized and zero imputed dataset aren't significantly different, therefore this approach can be a useful instrument to explore data richness.

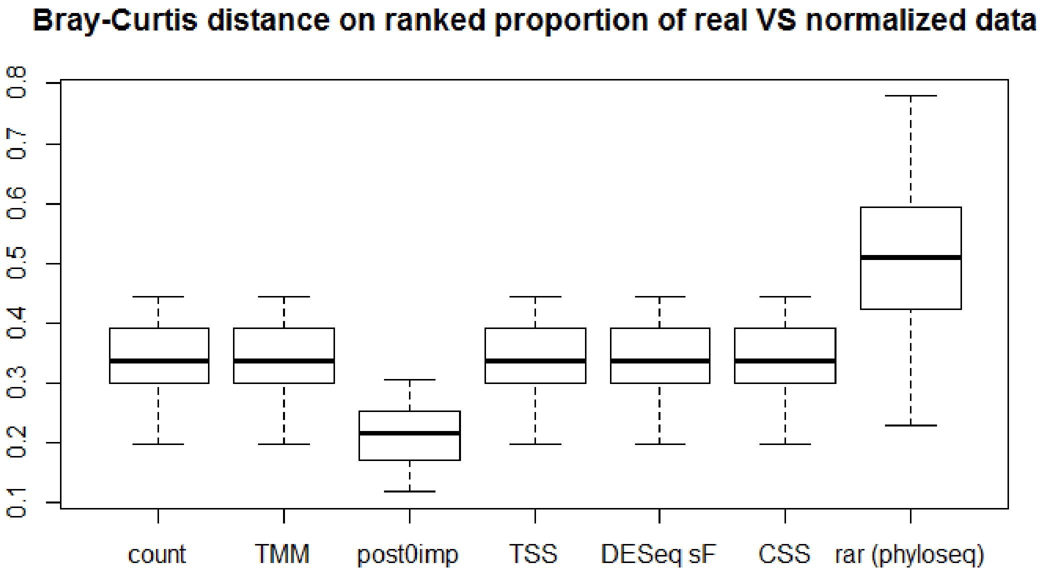


FIGURE 30 COMPARISON OF BRAY-CURTIS DISTANCE MEASURED BETWEEN REAL AND NORMALIZED RANKED PROPORTION WHEN ALL THE NORMALIZATION APPROACHES ARE CONSIDERED.

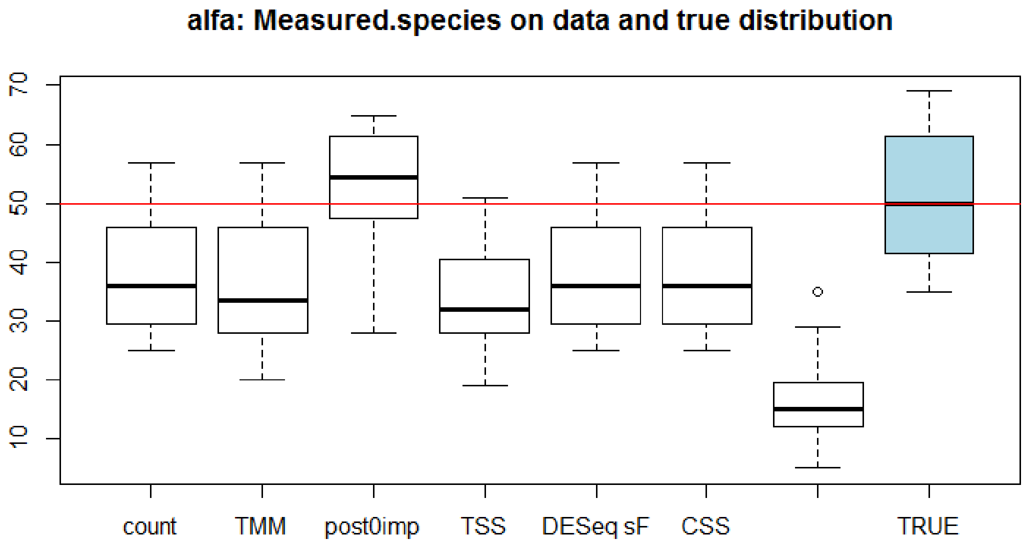


FIGURE 31 S VALUES OBTAINED WITH DIFFERENT NORMALIZATION APPROACHES AND REAL VALUE DISTRIBUTION.

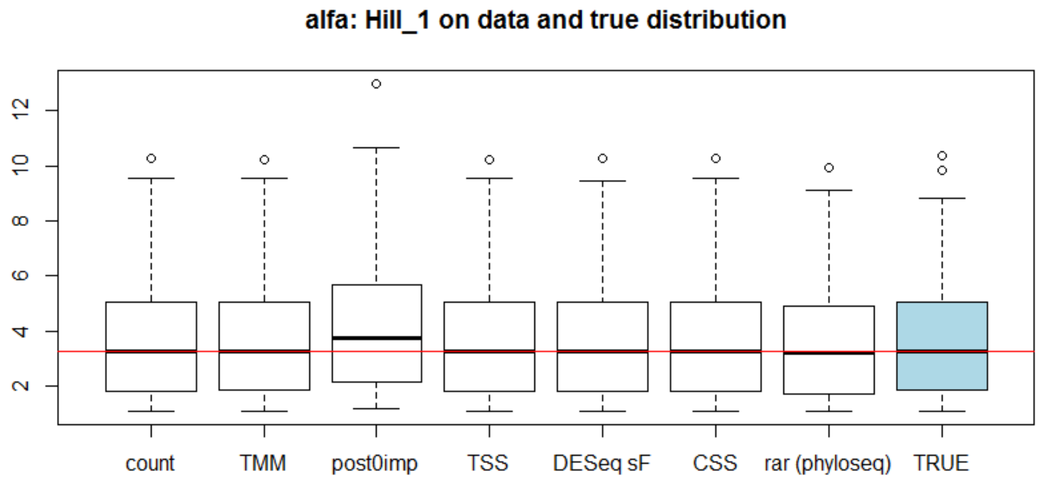


FIGURE 32 HILL'S NUMBER OBTAINED WITH DIFFERENT NORMALIZATION APPROACHES AND REAL VALUE DISTRIBUTION.

On the other hand, the S value obtained using all the remaining normalization methods are significantly different from reality. All the normalization approaches proposed show a Fisher alpha value which differs significantly from the real one; however, the p-value obtained when the zero imputation is used is several orders of magnitude greater than the one obtained for all the others normalization approaches (2e-04 compared to values smaller than 10e-9). These results suggest that the zero imputation truly helps count data to recover some rare species remained undetected: indeed S well approximates real data richness and Fisher's log series index may benefit from this recovery in better fitting the amount of rare species[10]. Conversely, when all the other alpha diversity indices are considered, both Wilcoxon and t-test agree in detecting zero imputation as the normalization techniques that achieves significantly different results. Moreover, in all cases the values obtained when this method is used is greater than the real one. All these indices focus on species evenness distribution, therefore zero imputation seems to inflate their values by making normalized counts more evenly distributed than real data. Fig. 31 and 32 visually support our discussion by presenting both a richness and an evenness index.

[10] Here we recall that the Fisher index α tries to fit species abundance using a log series distribution of the form $\alpha x, \frac{\alpha x^2}{2}, \frac{\alpha x^3}{3}, \ldots$ where each of these terms represents the number of species having abundance equal to 1, 2, 3, …

Afterwards we assessed normalization effect on beta diversity evaluation. We investigated both diversity within a single body site and between two different ones, and again statistical tests have been used to assess significant differences. The ability of normalized data to reproduce beta diversity found within a body site actually depends on the sampling spot itself. Indeed, in most cases all the normalization approaches show really poor performances, by greatly overestimating or dramatically underestimating diversity, independently from the measure chosen. It is even difficult to rank normalization approaches depending on their performances, since most of them have variable results. In all cases, although, scaling approaches like TSS, TMM and DESeq size factor tend to overestimate beta diversity: this can be imputed to the reduced ability to find common species if some of them are rare and therefore go undetected. On the other hand, zero imputation always underestimates beta diversity within a body site. This could be seen as a consequence of using similar data from the same sampling spot to correct zeros, since in this way all samples within it will look more similar to one another. Although an undesirable feature, it cannot be corrected, since there is no way to distinguish a zero to be corrected from a zero really deriving from biological variability. In all but one case, beta diversity assessed on normalized data resulted significantly different from the real value. The exception is obtained when CSS normalization is applied to the Buccal Mucosa dataset, as shown in Fig. 33. Indeed in this occasion only, when β_w and β_{cc} are used, the beta diversity estimated is comparable to the real one. However, no general conclusion can be drawn on the performance of CSS normalization approach, since it really depends on the underlying data distribution; to support our statement, Fig. 34 shows its poor performance when applied to the Mid Vagina dataset.

Finally, we investigated the performances of normalization approaches when beta diversity between two different body sites is computed. In this occasion only, we limited our analysis at comparing the Mid Vagina and the Vaginal Introitus dataset. Statistical testing assessed that no normalization approach is able to achieve beta diversity values comparable to real data; however, visual investigation of boxplot shows that some potential may lie in the TMM normalization and zero imputation pipeline. Indeed, it seems to obtain results resembling real beta diversity more than other normalization methods do. Again this method tends to underestimate true diversity, because of its inability to discern missing data from true absent species.

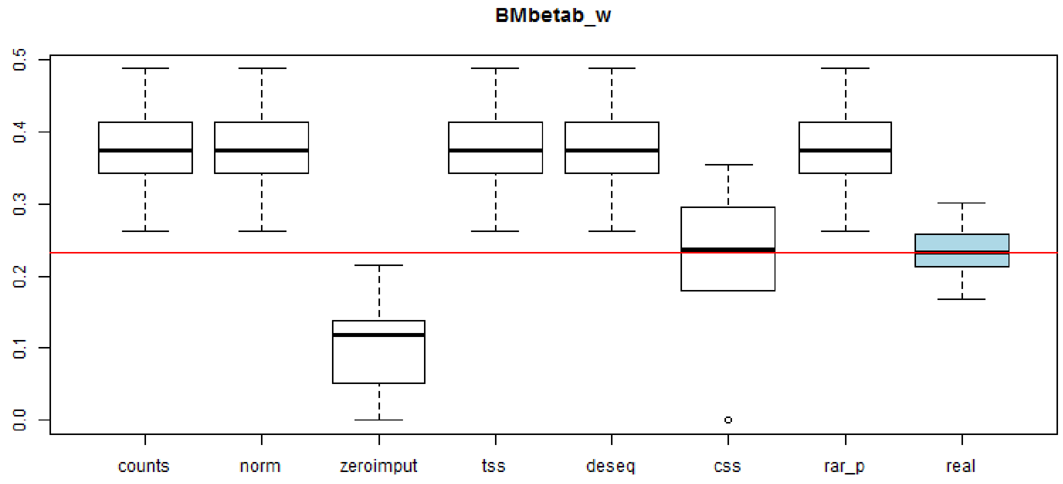


FIGURE 33 BETA DIVERSITY ASSESSED USING WHITTAKER'S INDEX ON BUCCAL MUCOSA REAL AND NORMALIZED DATA

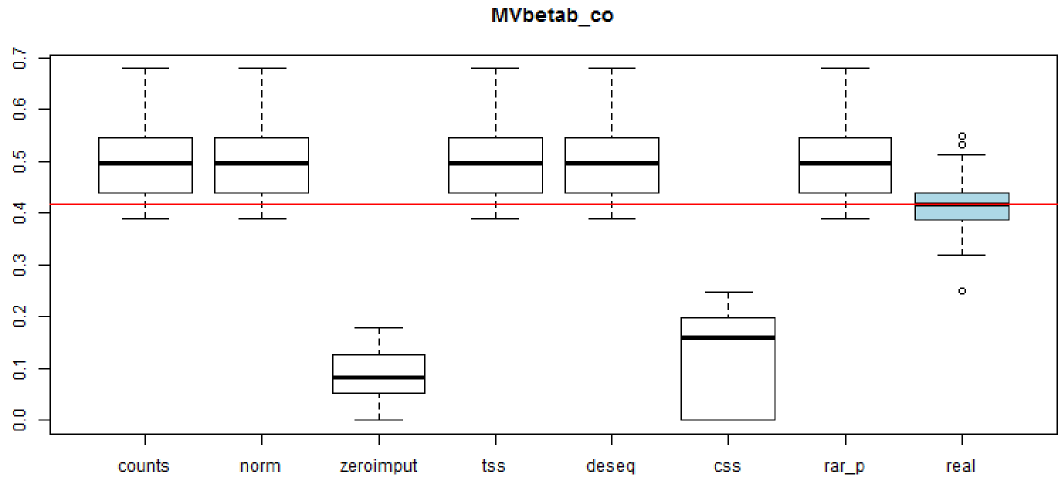


FIGURE 34 BETA DIVERSITY ASSESSED USING CODY'S INDEX ON MID VAGINA REAL AND NORMALIZED DATA

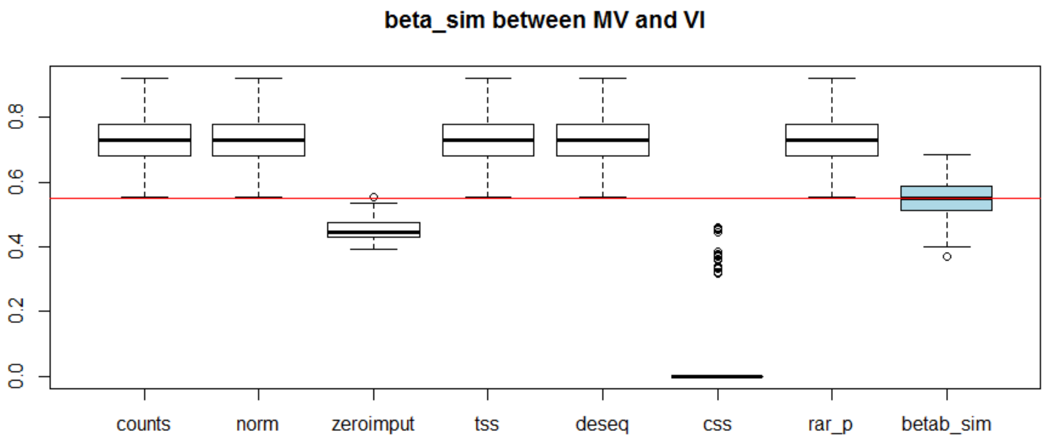


FIGURE 35 BETA DIVERSITY BETWEEN MID VAGINA AND VAGINA INTROITUS EVALUATED USING LENNON'S INDEX.

6 CONCLUSIONS

In this chapter we proved that microbiome data suffer from several variability sources. Therefore they may benefit from the application of a normalization technique in order to make data reliably comparable. However, no standard exists on what approach to prefer to normalize microbiome data. We performed a simulated experiment, in which a mock microbial community was generated *in-silico* to assess normalization method's performances. Unfortunately, there is poor accord in our result to set a golden standard approach; the technique to be preferred actually seem to depend on the aim of the investigation. In particular, none of the methods tested were able to correct data so that the estimated sample evenness or diversity is comparable with the real one. On the other hand, if sample's proportion or the observed number of species are to be recovered, the new zero imputation approach we proposed seems to work quite well. The user should keep in mind, when using it, that this approach tends to overestimate data richness and underestimate its diversity; however, richer data than the simulated one could lead to better results. Indeed, if technical replicates are available in the dataset, the zero imputation procedure is more likely to correct only the zeros which are really present species, by choosing as its similar nearer samples the replicates itself. In our simulation, in fact, each sample was a different biological community, therefore similar samples were chosen among data that could really differ because of biological variability, that we should really try to preserve.

As a side note, we would like to underline that the zero imputation procedure we proposed starts from a rather poor-performing normalization approach, the trimmed mean. It could be interesting to investigate if other –better performing- techniques, like DESeq normalization or CSS, could be a better starting point to choose before applying zero correction.

CHAPTER 4: MICROBIOTA DATA SIMULATION

In this chapter we will describe the steps and the statistical background we used to build a microbiome count data simulation, i.e. a pipeline that, based upon several real data observation, generates an *in-silico* microbial community and the count table derived from its sequencing. We have simulated a count-level dataset with N taxa (genus-level aggregated) and M samples, extracted from the following sampling spot: Buccal Mucosa, Tongue Dorsum, Vaginal Introitus and Mid Vagina, which data had been previously downloaded from the Human Microbiome Project database. According to precedent findings [60], microbiome data show evident over-dispersion, therefore we applied a Negative Binomial distribution to model them.

Since we believed the keystone effects to be monitored are both data sparsity and over-dispersion, we developed our simulation step-by-step, from bacterial individuals to sequenced counts, replicating *in-silico* the different stages of a marker gene sequencing experiment as follows.

1 THE COUNT TABLE GENERATION

We started by simulating a vector representing the mean abundance of each species in a bacterial community, specific for a body site. This vector, called m_{BS}[11], contains the mean number (included zero) of bacterial individuals for all N the simulated species:

$$m_{BS} = [m_1, m_2, \dots, m_N]$$

However, in order to be independent from the total number of microbial individuals found, we decided to model directly the proportions, i.e. the probabilities that an individual from a sample belongs to a particular species. Therefore we set

$$p_{BS}(i) = \frac{m_i}{\sum_j m_j}$$

to be the proportion of individuals, on average, belonging to the i-th species.

[11] *BS* here stands for body site.

When exploring the bacterial community of a specific body site, we would typically compare it among many subjects. As we introduced in the previous chapter, different individuals may show variation in their microbial population because of biological variability: we modelled such variability using a Gamma distribution, one for each species involved. For each body sites, we generated 10 biological replicates, their values being sampled from the Gamma distributions modelled.

Thus, if we indicate with s_{ij} the relative abundance of the species *i* in the subject *j*, it is distributed as a Gamma across all samples, with

$$s_{ij} \sim \Gamma(k, \theta)$$

where the two positive parameters, called *shape k* and *scale θ*, are related to mean and variance according to the following equations

$$E\left[s_{ij}\right] = k \cdot \theta$$

$$Var\left[s_{ij}\right] = k \cdot \theta^2$$

We fixed

$$\theta = \phi \cdot p(i)$$

$$k = 1/\,\phi$$

thus

$$E\left[s_{ij}\right] = p_{BS}(i)$$

$$Var\left[s_{ij}\right] = \phi \cdot p_{BS}(i)^2 \qquad (**)$$

meaning that the average proportion of species *i* across all subjects is equal to $p(i)$, while the variance for the same species depends on the square of the mean and from a constant value ϕ (**), which is equal for all the simulated species and whose meaning will be explained later.

To simulate the sequencing of the M samples, we use a Poisson distribution to generate the counts n_{ij} for each species *i* and sample *j*, which accounts for the technical variability of the sequencing step [72] [53] [52] [57]. Therefore, if s_{ij} is the proportion of individuals belonging to the *i*-th species in sample *j*, then the number of counts n_{ij} is distributed as a Poisson with parameter

$$\lambda_{ij} = s_{ij} \cdot N_j$$

where N_j represents the sequencing depth available for sample *j*. As known,

$$E[n_{ij}] = Var[n_{ij}] = \lambda_{ij} = s_{ij} \cdot N_j$$

Indeed, when a Poisson distribution's parameter λ is itself a random variable distributed as a Gamma, as in this case, Negative Binomial distribution arises. The resulting count matrix is therefore distributed following a Negative Binomial distribution for each species' counts.

Thus

$$E[n_{ij}] = \lambda_{ij} = s_{ij} \cdot N_j$$

$$Var[n_{ij}] = \lambda_{ij} \cdot (1 + \phi\lambda_{ij}) = s_{ij} \cdot N_j \cdot (1 + \phi \cdot s_{ij} \cdot N_j)$$

ϕ is therefore the overdispersion observed in sequencing data, which, as known, are distributed as a Negative Binomial [57].

This count generation approach has been repeated for four body sites, using different parameters for each of them, whose values are derived from real data as described in the following section.

2 Parameters value

We start by recalling that, in order to develop the simulation steps described above, there are a few parameters to be set:

- the data matrix dimension: the number of species N and the number of observed samples M;
- the original data proportions of the body site $p_{BS}(i)$, from which to extract the biological replicates;
- the overdispersion parameter ϕ;
- the sequencing depths.

To obtain realistic data distribution, therefore, we firstly had to investigate real datasets from which to derive the parameters we needed.

We started by downloading from the HMP database the count tables for four body sites: Buccal Mucosa, Tongue Dorsum, Vaginal Introitus and Mid Vagina. A few preprocessing steps were needed to rearrange all data in a unique count matrix, having rows describing speciess and columns describing different samples/subjects. Data were aggregated by genus level and replicate samples deriving from the same subject were eliminated, keeping only the sample showing higher sequencing depth. Lastly, we removed OTUs showing zero abundance (i.e. absent) in all the samples from all four sampling spots, while arbitrarily deciding to keep only 10 subject for each body site, selected among the ones showing the highest sequencing depth. These steps left us with a count table with 130 rows and 40 samples, which was then used in our simulation to be N and M respectively.

As we already stated, over-dispersed data are commonly distributed according to a Negative Binomial distribution [73] [74] [75]. The same holds true for the real data we investigated, as shown in the rightmost column of Fig.37. Therefore, we used a Negative Binomial model to fit them, separately for each body site. We calculated the common dispersion[12] for all the species (ϕ) to be used in our simulation, assuming that all species had the same mean-variance relationship. Moreover, we computed for each species the average log number of individuals per million[13], which is a useful descriptive measure of the "expression level" of the species itself, to be used in our simulation. However, before using such a variable as representative of microbial population, we had to correct it both for logarithm and for the added prior (used to avoid applying logarithm to zero values):

$$cpm_{BS} = 2^{AveLogCPM_{BS}}$$

$$cpm_{BS} = cpm_{BS} - \min(cpm_{BS})$$

We used cpm_{BS} as a proxy of the average abundance of each microbial strain in our mock community. In order to extract a coherent yet different microbial ensemble, we observed the histogram of the log distribution of cpm_{BS} and decided to fit it with a Gamma distribution (see Fig.36), from which we sample our simulated vector of individuals, m_{BS}. We decided to use a Gamma fit because of the positive support of this distribution ($x \in (0, \infty)$) and to exploit the flexibility its probability density function has, depending on the parameterization chosen. We fixed our parameters

[12] We performed this by calling the functions *calcNormFactors* and *estimateGLMCommonDisp*, found in the edgeR package

[13] Described by the AveLogCpm variable calculated by *estimateGLMCommonDisp.*

after several trials so that the simulated vectors contain a number of zeros (i.e., of absent species) that is similar to real data; that's the reason why we decided to adapt parameterization depending on sampling spot[14]. However, it is worth noting that our main goal is not the goodness of fit, but the possibility we have, by manipulating the Gamma parameters, to determine the number of zeros to be found in our sample, thus regulating data abundance evenness or unevenness.

The last parameter to be set in our simulation is the sequencing depth used when replicating the sampling step. In our simulation we used real sequencing depth, extracted from our real data subset, being the ten highest library size found for each body site; however, the user could easily use a fixed sequencing depth or generate random ones, by fitting realistic library sizes.

As a simple inner check, we show in Fig. 37 that the data obtained following these steps show a Negative Binomial trend, with over-dispersion comparable to real data. The main strength of this approach is related to our ability to control and manipulate the sparsity of the overall dataset by changing the initial parameters of the Gamma distribution.

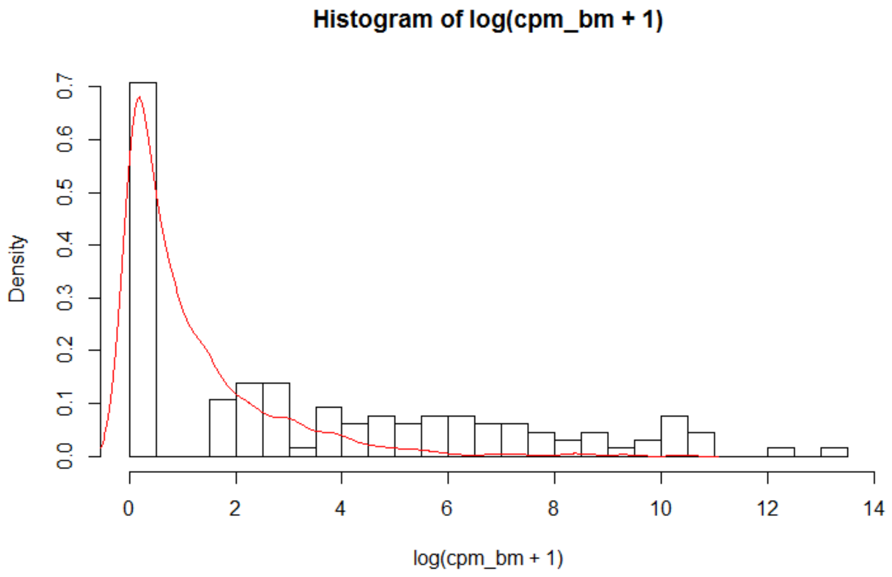
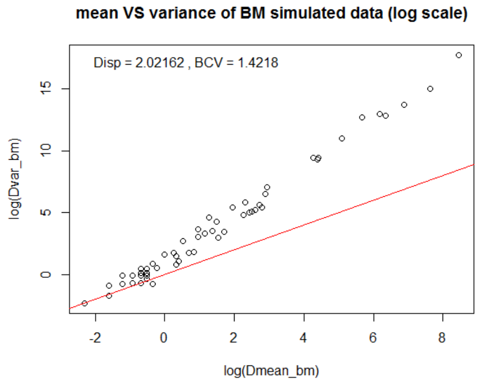
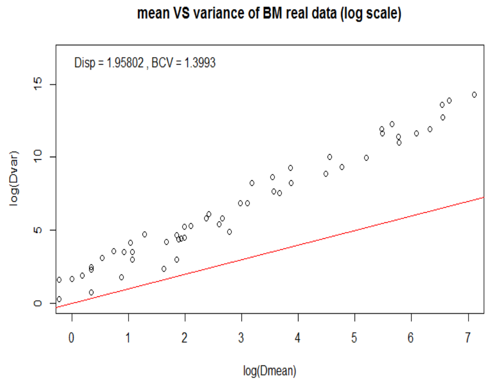


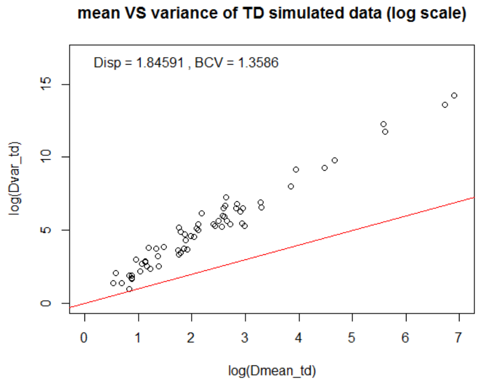
FIGURE 36 HISTOGRAM OF LOG COUNT PER MILLION ABUNDANCE FOR BUCCAL MUCOSA AND ITS GAMMA FIT

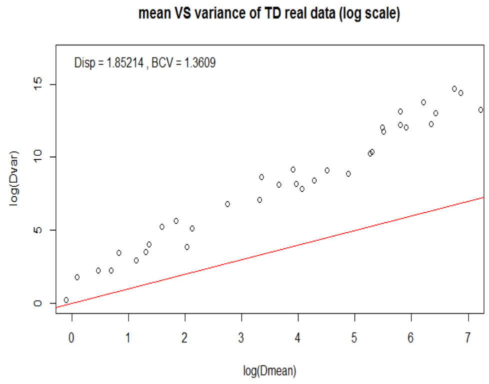
[14] For results reproducibility: Buccal Mucosa has k=0.6, θ=2; Tongue Dorsum has k=0.45, θ=2; Vaginal Introitus has k=0.65, θ=3; Mid-vagina has k=0.4, θ=3.

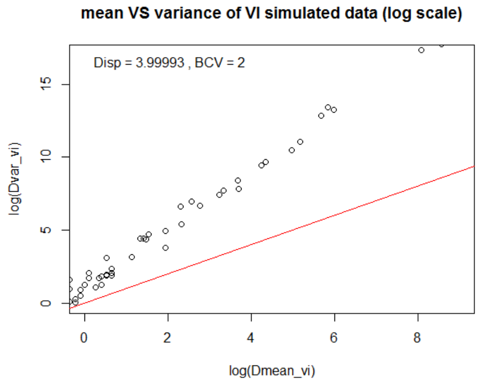
SIMULATED REAL

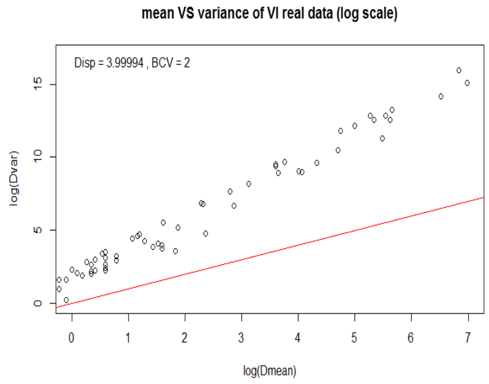












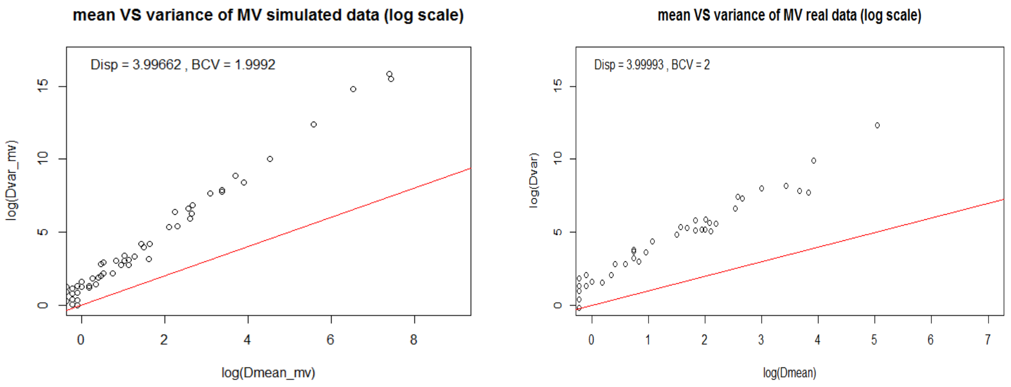


FIGURE 37 COMPARISON OF MVA PLOTS (LOG SCALE) FROM REAL AND SIMULATED DATA. BCV= BIOLOGICAL COEFFICIENT OF VARIATION; DISP=PHI.

CONCLUSION

Humans and their microbiota together build a complex symbiotic ecosystem, whose ensemble cooperates at the benefit of both parts. Therefore, a rising belief advocates that the microbiota composition and its equilibrium might have a big role in defining human health. As we have tried to highlight throughout this thesis, next generation sequencing techniques give us great opportunities to explore microbiota, by analyzing its genomic content. However, as we pointed out, microbiome data suffer from some inherent drawbacks which might prevent our investigation to target microbiota in a truthful way. We here recall that both limited sequencing depth and added technical variability alter our ability to recover true data distribution, and that these features further complicates the investigation of often rare and/or unevenly distributed data like microbiome ones.

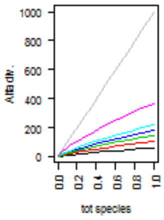
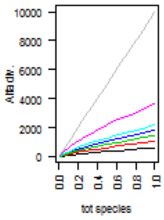
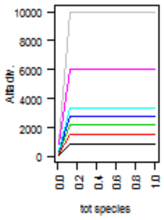
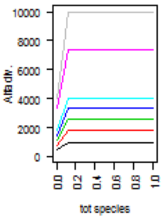
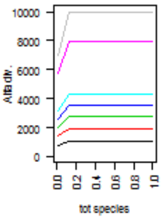
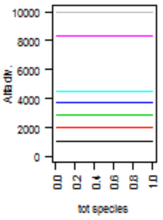
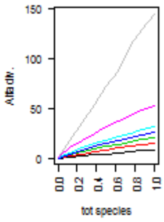
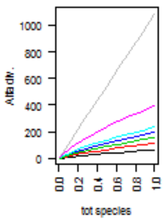
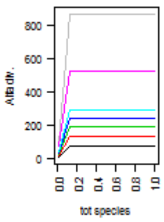
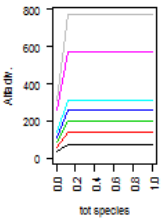
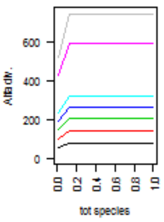
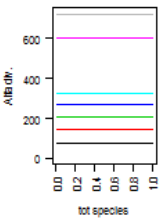
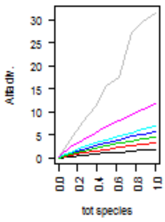
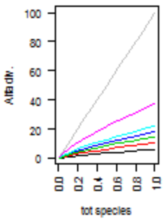
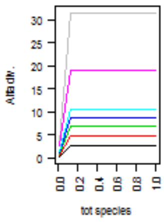
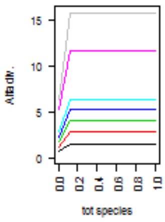
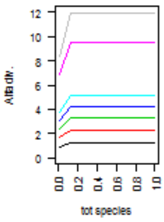
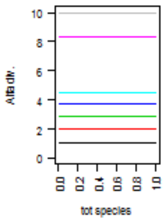
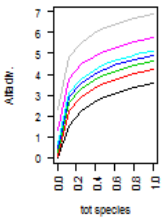
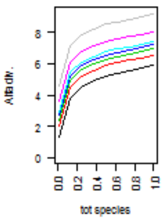
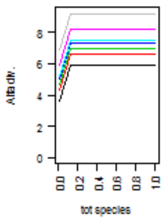
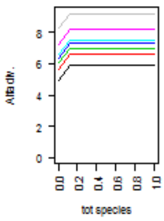
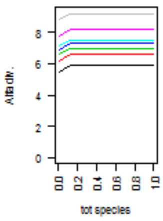
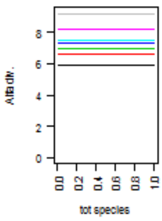
We therefore felt the need to study and, at least partly, try to account for this kind of issues. Thus, the main contributions developed in this thesis can be summarized in four main points. First of all, several biodiversity indices, both targeting alpha and beta diversity, have been reviewed, tested, and their properties have been assessed in a simulation context. Starting from previous work, we have tried to eliminate redundancies, enforce consistency and assess applicability in the microbiome context. Secondly, we have explored normalization approaches, often applied to microbiome data to reduce technical variability. We started from recent literature and analyzed several different methods, some of which are borrowed from other genomic disciplines, like RNA-Sequencing. Besides evaluating their performances, we proposed a new method, useful to complement normalization with a second step, the zero imputation. Our method is aimed at selectively correcting data for missed species, a problem none of the traditional normalization approaches can deal with. Lastly, we developed a microbiota simulation that we used to test all the reviewed methods.

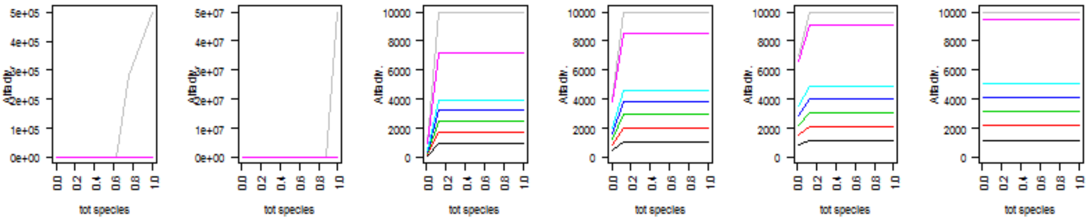
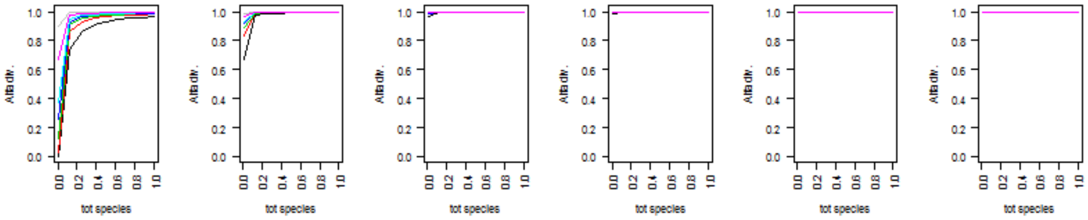
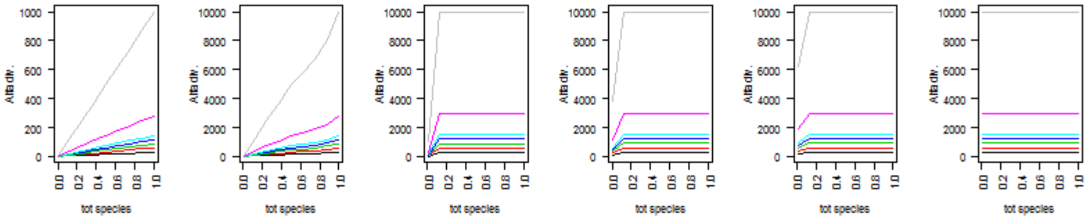
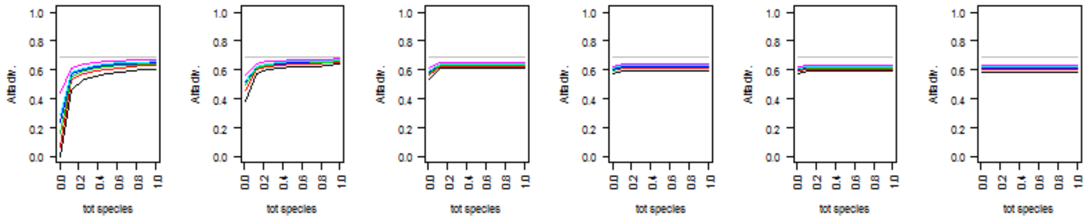
Several interesting results have been pointed out throughout this work. First of all, our revision of alpha and beta diversity underlined that clarity is needed when exploring microbiome biodiversity. An extensive yet not consistent literature exists, that we have explored in order to favor a conscious understanding of the specific properties being tested by each index. Moreover, a smaller, consistent subset of measures has been proposed to be used as a basic toolkit to explore data biodiversity. Then, our simulation confirmed that normalization of microbiome data is an essential step to be applied in order to mitigate added technical variability. Our suggestion is to supplement it with another step, the zero imputation, that is able to compensate for rare undetected data. Its first implementation and testing on microbiome simulated data has proven its potential in adjusting for species richness, although further work might be needed to improve data diversity detection too. However, this test has served as a proof of principle of its applicability and effectiveness in the microbiome data context. Lastly, we believe the microbiota data simulation we developed might be useful for further works were testing and evaluating of new methods are needed.

Microbiome data exploration could give us a picture of the bacterial population inhabiting our body space, therefore all the efforts are needed to extract the most complete and reliable information possible from them. Even if not exhaustive, this thesis aimed at detecting, understanding and trying to account for several different sources of error that could mislead data analysis and results. We are well aware that, before microbiome accesses clinical application, research still have to cover a long and uncertain path; however, if this works succeeds in arising curiosity for the subject and some consciousness on both its potential and its features, we will feel like we have fixed one brick of this difficult road.

APPENDIX A

The next figures summarize the obtained simulation results under different conditions:

Measure	Sequencing depth					
	1e+03	1e+04	1e+05	4e+05	7e+05	1e+06
S						
Margalef						
Menhinick						
Shannon						

Chao	
Simpson (C)	
InvSimpson (D)	
Pielou	

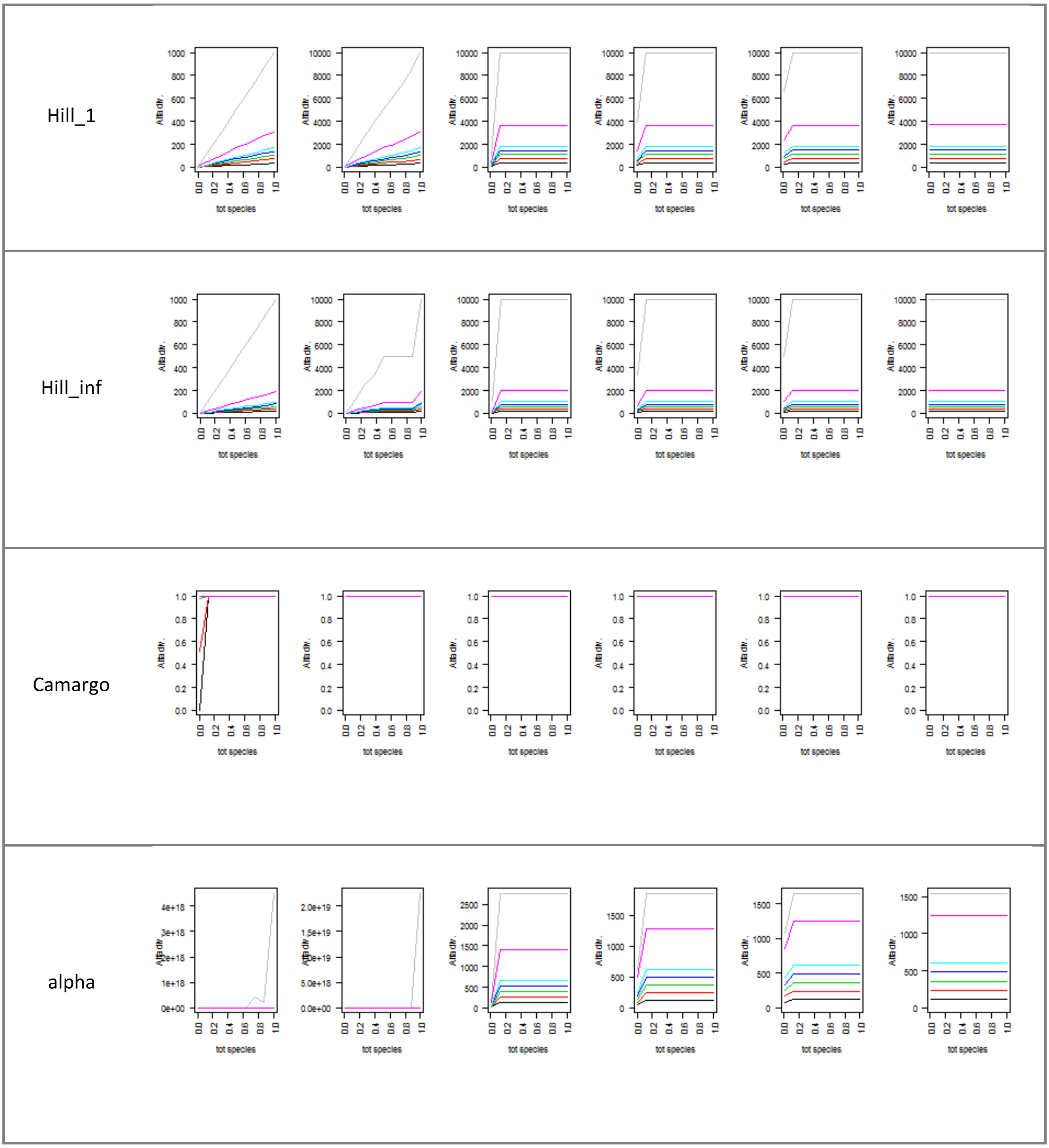
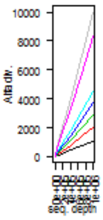
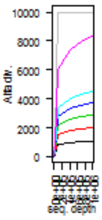
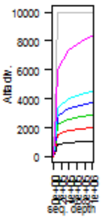
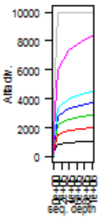
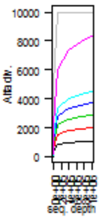
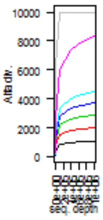
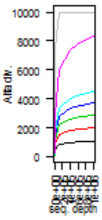
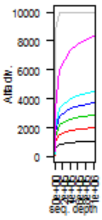
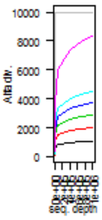
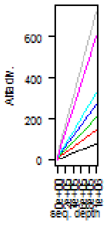
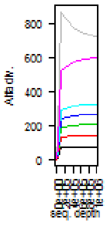
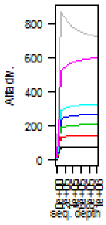
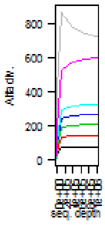
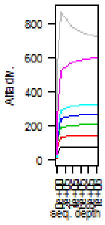
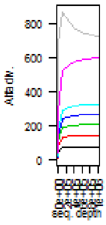
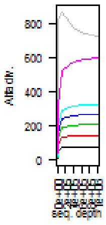
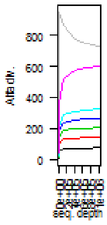
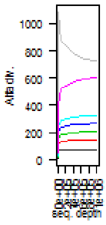
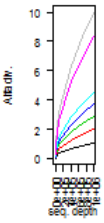
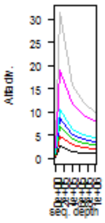
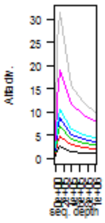
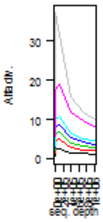
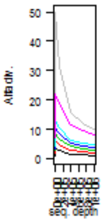
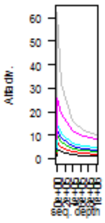
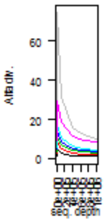
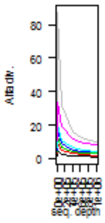
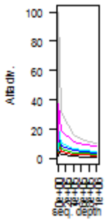
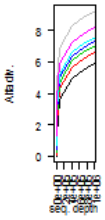
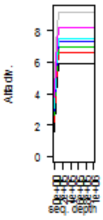
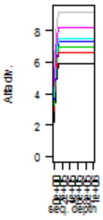
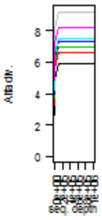
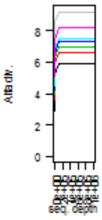
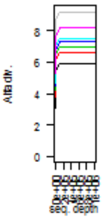
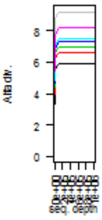
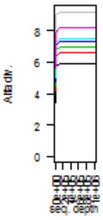
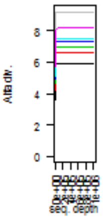
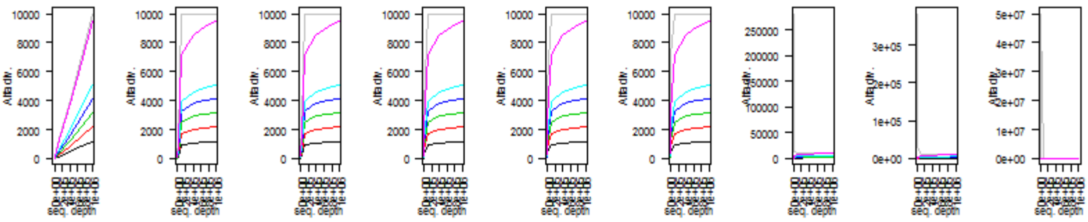
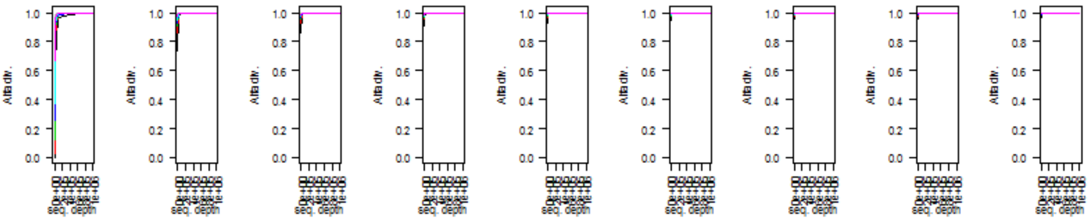
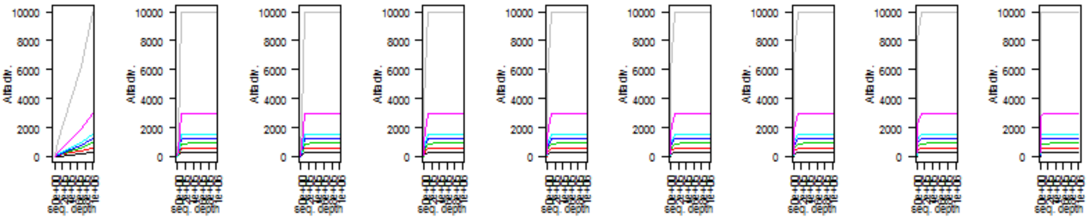
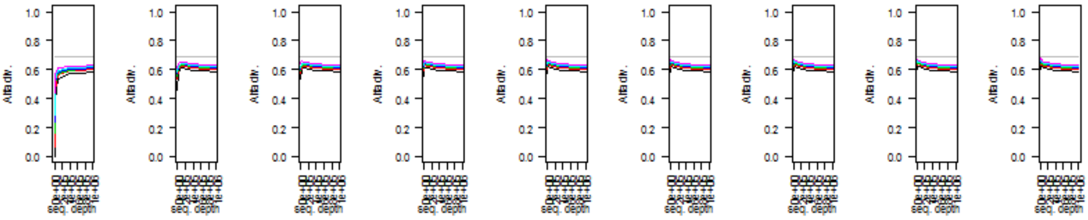
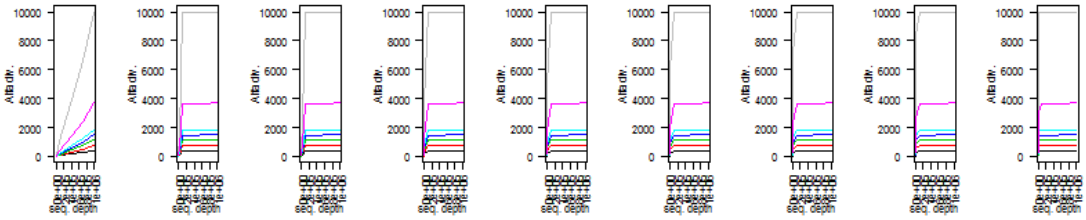


TABLE 3: ALPHA DIVERSITY VALUE (Y-AXIS) WHEN TOTAL NUMBER OF SPECIES INCREASES (X-AXIS), UNDER 6 DIFFERENT SPECIES ABUNDANCE DISTRIBUTION (COLOURS AS IN FIG 1), WHEN SEQUENCING DEPTH INCREASES (DIFFERENT COLUMNS).

Measure	Total species								
	0.01	**0.125**	**0.25**	**0.375**	**0.5**	**0.625**	**0.75**	**0.875**	**1.00**
S									
Margalef									
Menhinick									
Shannon									

Chao	
Simpson (C)	
InvSimpson (D)	
Pielou	
Hill_1	

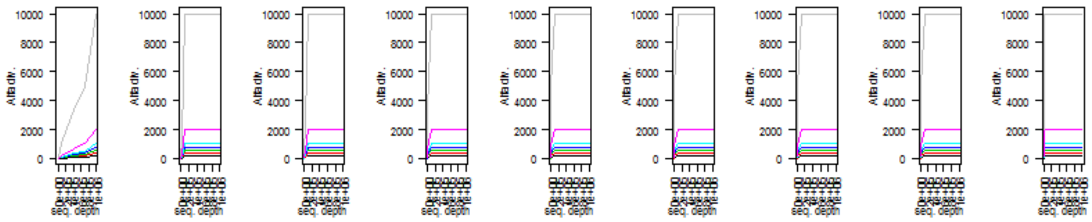
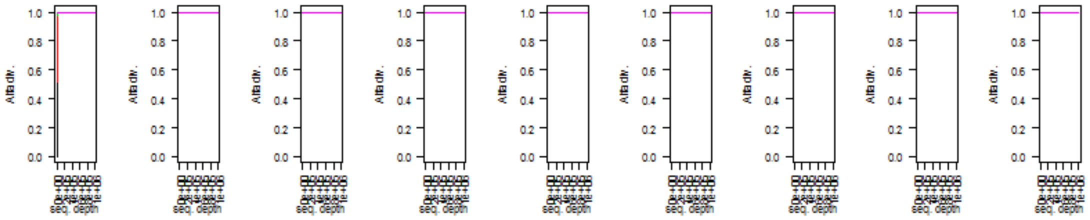
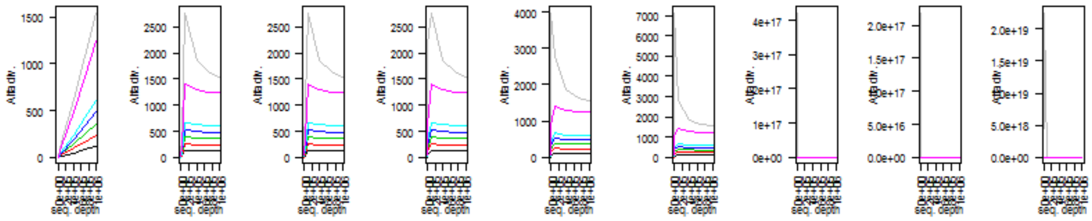
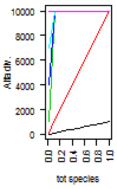
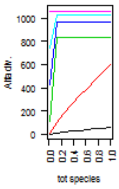
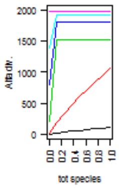
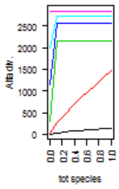
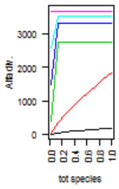
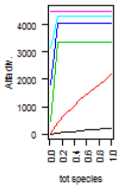
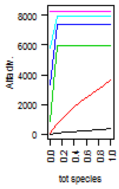
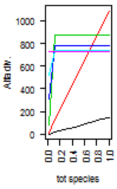
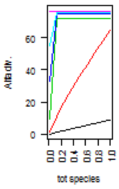
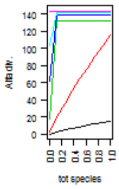
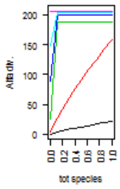
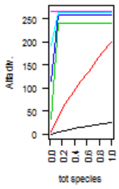
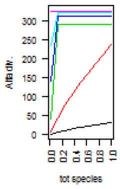
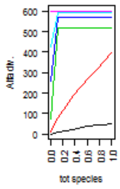
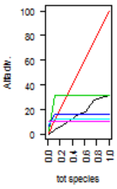
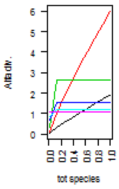
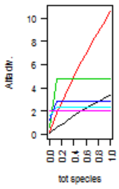
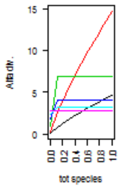
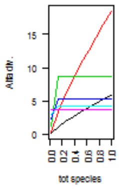
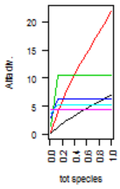
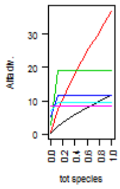
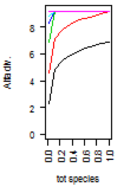
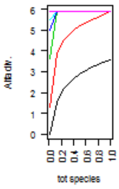
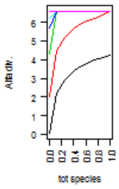
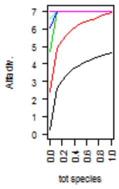
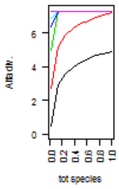
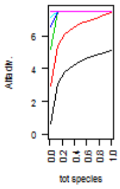
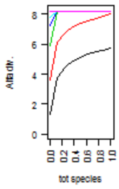
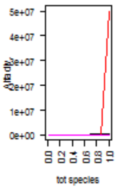
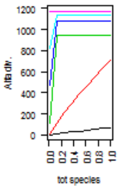
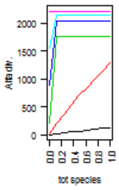
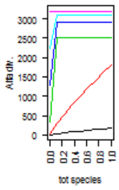
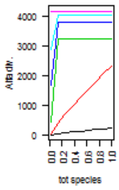
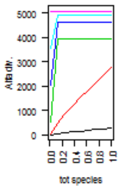
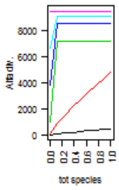
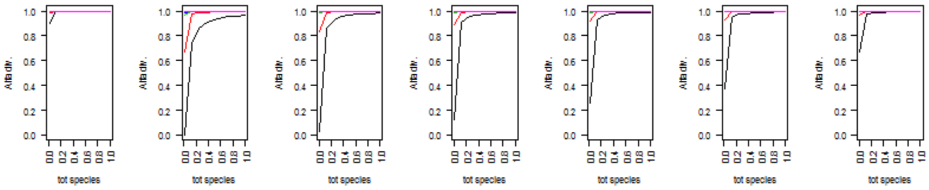
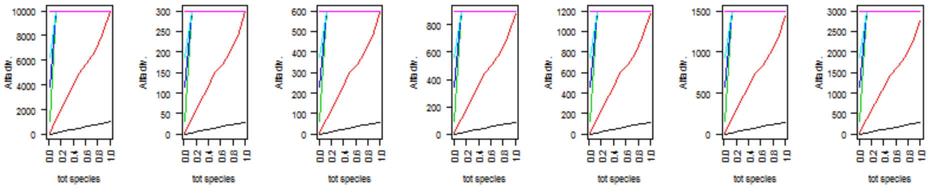
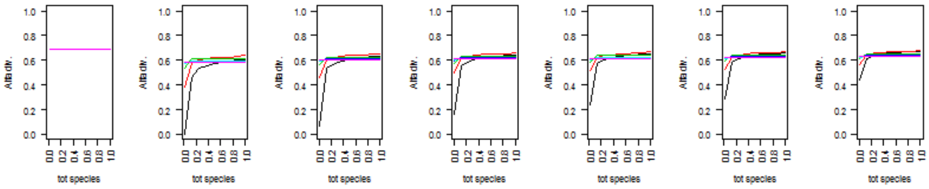
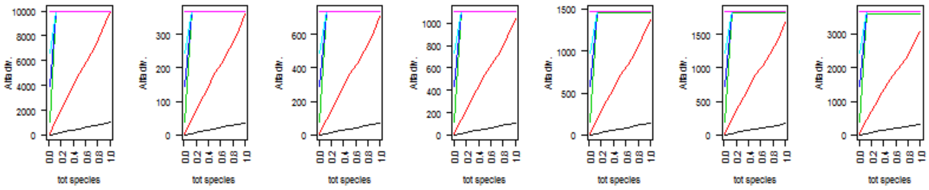
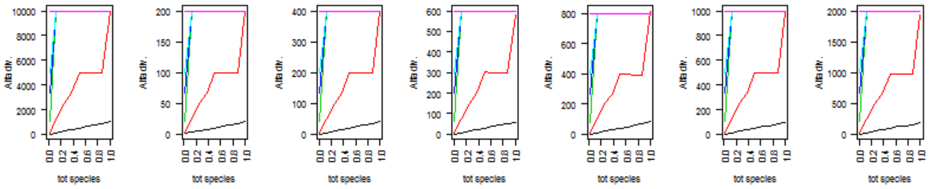
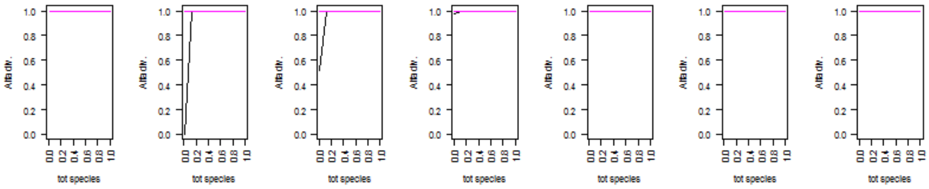
Hill_inf	
Camargo	
alpha	

TABLE 4: ALPHA DIVERSITY VALUE (Y-AXIS) WHEN SEQUENCING DEPTH INCREASES (X-AXIS), UNDER 6 DIFFERENT SPECIES ABUNDANCE DISTRIBUTION (COLOURS AS IN FIG 1), WHEN TOTAL NUMBER OF SPECIES INCREASES (DIFFERENT COLUMNS).

Measure	Shape 0	 1	 2	 3	 4	 5	 6
S							
Margalef							
Menhinick							
Shannon							
Chao							

Simpson (C)	
InvSimpson (D)	
Pielou	
Hill_1	
Hill_inf	
Camargo	

alpha

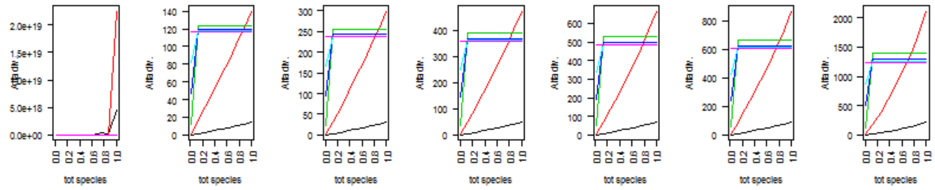


TABLE 5: ALPHA DIVERSITY VALUE (Y-AXIS) WHEN TOTAL NUMBER OF SPECIES INCREASES (X-AXIS), UNDER 4 DIFFERENT SEQUENCING DEPTH CONDITION, WHEN SPECIES ABUNDANCE DISTRIBUTION IS VARIED RANGING FROM THE MOST UNEVENLY DISTRIBUTED TO THE MOST EVENLY DISTRIBUTED ONE (DIFFERENT COLUMNS).

APPENDIX B

Our analysis started from the work of Koleff et al. [24], who already reviewed beta diversity measures and standardized their expressions in terms of shared or unique species. Table 6 contains all the 24 measures analyzed.

Since we are interested in measuring quantitatively sample's diversity, in our review we focus on dissimilarity indices only, meaning that when dealing with similarity indices we have always considered their complementary version (i.e. for each $\beta_{similarity}$, we considered $1\text{-}\beta_{similarity}$): this choice allowed us to compare values that scale accordingly with samples diversity. In addition, since some of the 24 indices reviewed by Koleff et al. are redundant, having the exact same expression, we searched for coincident measures, that we decided to aggregate. The next subparagraph are aimed at explaining our procedure.

- *Whittaker*

Whittaker proposed two expressions of the same index, as shown in Table 6, one scaling as a diversity index while the other scales as a similarity index. As we stated previously we choose to deal with diversity measures only, therefore we will discard the second variant. The retained one is computed as

$$\frac{a+b+c}{(2a+b+c)/2}-1$$

whose expression coincides with β_{-1} and, if we re-express it as

$$\frac{b+c}{2a+b+c}$$

it equals $\beta_{t,}$ β_{me} and β_{hk} too. Moreover, if we re-express β_{sor}, originally a similarity index, in terms of dissimilarity by evaluating its one complement, this measure too becomes coincident with β_W and can be therefore eliminated.

- *Cody*

His measure, expressed as

$$\frac{b+c}{2}$$

equals β_l and $\beta_{Wb}/2$.

- *Colwell & Coddington*

β_{cc}, calculated with the following equation

$$\frac{b+c}{a+b+c}$$

is the same as β_g. Once again, if we re-express β_j in terms of dissimilarity, we can exclude this index too since it coincides with β_{cc}.

We further decided to withdraw β_{gl} since it is not a true measure of beta diversity but rather a measure of local alpha diversity gradients (it represents the difference in species richness between samples) and β_{rlb} because it does not satisfy the basic property of symmetry [24]. Indeed, we choose to exclude any non-symmetric measure, since we want all the measures to remain unchanged if we symmetrically switch the two sets. By doing so, we obtained a coherent subset of non-redundant and comparable beta diversity measures to be used when testing samples for species dissimilarity.

	Symbol	Measure re-expressed	Reference
1	β_W	$\frac{a+b+c}{(2a+b+c)/2}-1 \ or \ \frac{a+b+c}{(2a+b+c)/2}$	Whittaker (1960), see also Magurran (1988)
2	β_{-1}	$\frac{a+b+c}{(2a+b+c)/2}-1$	Harrison et al. (1992)
3	β_c	$\frac{b+c}{2}$	Cody (1975)
4	β_{wb}	$b+c$	Weiher and Boylen (1994)
5	β_r	$\frac{2bc}{(a+b+c)^2-2bc} \ or \ \frac{2bc}{(a+b+c)^2-2bc}-1$	Routledge (1977), see also Magurran (1988), Southwood & Henderson (2000)
6	β_l	$log(2a+b+c)-\left(\frac{1}{(2a+b+c)}2alog2\right) - \left[\frac{1}{2a+b+c}\left((a+b)log(a+b) + (a+c)\,log(a+c)\right)\right]$	Routledge (1977), Wilson & Shmida (1984)
7	β_e	$exp(\beta_I)-1$	Routledge (1977)
8	β_t	$\frac{b+c}{2a+b+c}$	Wilson & Shmida (1984)
9	β_{me}	$\frac{b+c}{2a+b+c}$	Mourelle & Ezcurra (1997)
10	β_j	$\frac{a}{a+b+c}$	Jaccard (1912), see also Magurran (1988), Southwood & Henderson (2000)
11	β_{sor}	$\frac{2a}{2a+b+c}$	Sørensen (1948) based on Dice (1945); see also Whittaker (1975), Magurran (1988), Southwood & Henderson (2000)

12	β_m	$\frac{(2a+b+c)(b+c)}{(a+b+c)}$	Magurran (1988)
13	β_{-2}	$\frac{min(b,c)}{max(b,c)+a}$	Harrison et al. (1992)
14	β_{co}	$1-\frac{a(2a+b+c)}{(a+b)(a+c)}$	Cody (1993)
15	β_{cc}	$\frac{b+c}{a+b+c}$	Colwell & Coddington (1994, “complementarity” measure), see also Pielou (1984)
16	β_g	$\frac{b+c}{a+b+c}$	Gaston et al. (2001)
17	β_{-3}	$\frac{min(b,c)}{a+b+c}$	Williams (1996)
18	β_l	$\frac{b+c}{2}$	Lande (1996)
19	β_{19}	$\frac{bc+1}{((a+b+c)^2-(a+b+c))/2}$	Williams (1996), Williams et al. (1999)
20	β_{hk}	$1-\frac{2a}{2a+b+c}$	Harte & Kinzig (1997)
21	β_{rlb}	$\frac{a}{a+c}$	Ruggiero et al. (1998)
22	β_{sim}	$\frac{\min(b,c)}{\min(b,c)+a}$	Lennon et al. (2001), based on Simpson (1943)
23	β_{gl}	$\frac{2\lvert b-c\rvert}{2a+b+c}$	Lennon et al. (2001)
24	β_z	$1-\left[\frac{log\left(\frac{2a+b+c}{a+b+c}\right)}{log2}\right]$	Lennon et al. (2001), see also Harte & Kinzig (1997)

TABLE 6 ALL THE DIVERSITY INDICES REVIEWED IN KOLEFF ET AL. IN SHADED BOXES, THE MEASURES THAT WE HAVE RETAINED.

BIBLIOGRAPHY

[1] J. Lederberg e A. McCray, «'Ome Sweet 'Omics—a genealogical treasury of words,» *Scientist,* vol. 15, n. 8, 2001.

[2] A. Guss e e. al, «Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis.,» *The ISME Journal,* n. 5, p. 20–9, 2011.

[3] P. Turnbaugh e e. al, «A core gut microbiome in obese and lean twins,» *Nature,* vol. 457, n. 7228, p. 480–84, 2008.

[4] A. Giongo e e. al, «Toward defining the autoimmune microbiome for type 1 diabetes,» *The ISME Journal,* vol. 5, n. 1, p. 82–91, 2011.

[5] J. Staley e A. Konopka, «Measurements of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats,» *Annu. Rev. Microbiol.,* n. 39, pp. 321-46, 1985.

[6] C. Woese e F. G.E., «Phylogenetic structure of the prokaryotic domain: The primary kingdoms,» *Proceedings of the National Academy of Sciences,* vol. 74, n. 11, p. 5088–90, 1977.

[7] J. Barriuso, J. Valverde e R. Mellado, «Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows,» *BMC Bioinformatics,* vol. 12, n. 473, 2011.

[8] T. H. M. P. Consortium, «The NIH Human Microbiome Project,» *Genome Res.,* vol. 19, n. 12, p. 2317–23, 2009.

[9] T. H. M. P. Consortium, «A framework for human microbiome research,» *Nature,* n. 486, p. 215–21, 2012.

[10] T. H. M. P. Consortium, «Structure, function and diversity of the healthy human microbiome,» *Nature,* n. 486, p. 207–14, 2012.

[11] [Online]. Available: http://commonfund.nih.gov/hmp/index.

[12] [Online]. Available: http://www.hmpdacc.org/micro_analysis/microbiome_analyses.php.

[13] G. Xie, C. Lo, M. Scholz e P. Chain, «Recruiting Human Microbiome Shotgun Data to Site-Specific Reference Genomes,» *PLOS one,* 2014.

[14] C. Quast, E. Pruesse, P. Yilmaz e e. al., «The SILVA ribosomal RNA gene database project: improved data processing and web-based tools,» *Nucl. Acids Res.,* n. 41, pp. 590-96, 2013.

[15] «High-throughput DNA sequencing – concepts and limitations,» *Bioessays,* n. 32, pp. 524-36, 2010.

[16] F. Finotello e B. Di Camillo, «Sequencing Technologies,» 2014.

[17] [Online]. Available: http://rna.ucsc.edu/rnacenter/ribosome_images.html.

[18] V. D'Argenio, G. Casaburi, V. Vincenza Precone e F. Francesco Salvatore, «Comparative Metagenomic Analysis of Human Gut Microbiome,» *BioMed Research International,* n. 325340, pp. 1-11, 2013.

[19] [Online]. Available: http://huttenhower.sph.harvard.edu/metaphlan.

[20] C. Shannon, «A mathematical theory of communication,» *The Bell System Technical Journal,* n. 27, p. 379–423 and 623–656, 1948.

[21] R. Fisher, A. Corbet e C. Williams, «The relation between the number of species and the number of individuals in a random sample of an animal population,» *Journal of Animal Ecology,* n. 12, pp. 42-58, 1943.

[22] R. Whittaker, «Vegetation of the Siskiyou Mountains, Oregon and California.,» *Ecological Monographs,* n. 30, p. 279–338, 1960.

[23] R. Whittaker, «Evolution and measurement of species diversity,» *Taxon,* vol. 21, pp. 213-251, 1972.

[24] P. Koleff, K. Gaston e J. Lennon, «Measuring beta diversity for presence-absence data,» *Journal of Animal Ecology,* n. 72, pp. 367-82, 2003.

[25] C. Heip, P. Herman e K. Soetaert, «Indices of diversity and eveness,» *Océanis,* n. 24, pp. 61-87, 1998.

[26] M. Anderson e e. al, «Navigating the multiple meanings of beta diversity: a roadmap for

the practicing ecologist,» *Ecology Letters,* n. 14, pp. 19-28, 2011.

[27] A. Tuomisto, «A consistent terminology for quantifying species diversity? Yes, it does exist,» *Oecologia,* n. 164, p. 853–60, 2010.

[28] A. Tuomisto, «A diversity of beta diversities: straightening up a concept gone awry. Part 1.,» *Ecography,* n. 33, pp. 2-22, 2010.

[29] A. Tuomisto, «A diversity of beta diversities: straightening up a concept gone awry. Part 2.,» *Ecography,* vol. 33, pp. 2-22, 2010.

[30] H. Wolda, «Similarity indices, sample size and diversity,» *Oecologia,* n. 50, pp. 296-302, 1981.

[31] A. Chao e T. Shen, «Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample,» *Environmental and Ecological Statistics,* n. 10, pp. 429-443, 2003.

[32] E. Simpson, «Measurement of diversity,» *Nature,* vol. 163, n. 688, 1949.

[33] C. Heip, P. Herman e K. Soetaert, «Indices of diversity and eveness,» *Océanis,* n. 24, pp. 61-87, 1998.

[34] C. Mulder e et al, «Species evenness and productivity in experimental plant communities,» *Oikos,* vol. 107, p. 50–63, 2004.

[35] A. Lepetre e M. Mouillot, «A comparison of species diversity estimators,» *Res Popul Ecol,* vol. 41, pp. 203-15, 1999.

[36] R. Fisher, A. Corbet e C. Williams, «The relation between the number of species and the number of individuals in a random sample of an animal population,» *Journal of Animal Ecology,* n. 12, pp. 42-58, 1943.

[37] C. Shannon, «A mathematical theory of communication,» *The Bell System Technical Journal,* n. 27, p. 379–423 and 623–656, 1948.

[38] M. Hill, "Diversity and Evenness: A Unifying Notation and Its Consequences," *Ecological Society of America,* vol. 54, no. 2, pp. 427-32, 1973.

[39] W. Berger e F. Parker, «Diversity of Planktonic Foraminifera in Deep-Sea Sediments,»

Science, n. 168, p. 1345–47, 1970.

[40] R. Whittaker, «Evolution of species diversity in land communities,» *Evolutionary biology,* n. 10, pp. 1-67, 1977.

[41] M. Cody, Diversity, rarity and conservation in Mediterranean-climate regions, Sinauer Associates, Sunderland, Massachussetts, 1986, pp. 122-52.

[42] R. Routledge, «On Whittaker's components of diversity,» *Ecology,* vol. 58, pp. 1120-27, 1977.

[43] R. Routledge, «Estimating ecological components of biodiversity,» *Oikos,* vol. 42, pp. 23-9, 1984.

[44] A. Magurran, Ecological diversity and its measurement, London: Croom-Helm, 1988.

[45] S. Harrison, S. Ross e J. Lawton, «Beta diversity on geographical gradients in Britain,» *Journal of animal Ecology,* n. 61, pp. 151-58, 1992.

[46] R. Colwell e J. Coddington, «Estimating terrestrial biodiversity through extrapolation,» *Philosophical Transactions: Biological Sciences,* vol. 345, n. 1311, pp. 101-18, 1994.

[47] P. Williams, «Mapping variations in the strength and breadth of biogeographic transition zones using species turnover.,» *Proceedings of the Royal Society ,* vol. 263, n. 579-588.

[48] J. Lennon, P. Koleff, J. Greenwood e K. Gaston, «The geographical structure of British bird distribution: diversity, spatial turnover and scale,» *Journal of Animal Ecology,* n. 70, pp. 966-79, 2001.

[49] N. e. a. Paulson, "Differential abundance analysis for microbial marker-gene surveys.," *Nature Methods ,* p. 10: 1200–1202., 2013.

[50] J. e. a. White, «WhiteAlignment and clustering of phylogenetic markers - implications for microbial diversity studies.,» *BMC Bioinformatics,* vol. 11, p. 152 , 2010.

[51] K. e. a. Faust, «Microbial Co-occurrence Relationships in the Human Microbiome.,» *PLoS Computational Biology,* vol. 8, p. , 2012.

[52] J. M. C. M. S. e. a. Marioni, «Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays.,» *Genome research,* vol. 18, pp. 1509-17, 2008.

[53] J. P. E. H. K. &. D. S. Bullard, «Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.,» *BMC Bioinformatics,* n. 11, p. 94, 2010.

[54] D. R. Auer P, «Statistical Design and Analysis of RNA Sequencing Data Genetics.,» vol. 185, pp. 405-416, 2010.

[55] L. L. K. M. A. A. V. O. A. Y. L. N. V. McIntyre, «RNA-seq: technical variability and sampling.,» *BMC Genomics,* vol. 12, n. 293, 2011.

[56] W. W. Jiang H., «Statistical inferences for isoform expression in RNA-Seq.,» *Bioinformatics,* vol. 25, pp. 1026-32, 2009.

[57] A. B. B. G. D. e. a. Oberg, «Technical and biological variance structure in mRNA-Seq data: life in the real worls,» *BMC Biogenomics,* vol. 13, n. 304, 2012.

[58] A. S. a. H. W., «Differential expression analysis for sequence count data.,» *Genome Biology,* vol. 11, p. 106, 2010.

[59] S. D. C. J. e. a. Di Y., «The NBP negative binomial model for assessing differential gene expression from RNA-Seq.,» *Stat Appl Genet Mol Biol,* vol. 10, n. 24, 2011.

[60] P. McMurdie e S. Holmes, «Waste not, want not: why rarefying microbiome data is inadmissible,» 2013.

[61] M. O. A. Robinson, «A scalong normalization method for differential expression analysis of RNA-Seq data.,» *Genome Biol,* vol. 11, n. R25, 2010.

[62] M. D. a. S. G. Robinson MD, «edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.,» *Bioinformatics,* vol. 26, pp. -1, 2010.

[63] J. D. C. Y. S. a. K. G. McCarthy, «Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.,» *Nucleic Acids Research,* vol. 40, n. 10, pp. -9, 2012.

[64] R. M. a. S. GK, «Moderated statistical tests for assessing differences in tag abundance.,» *Bioinformatics,* vol. 23, pp. -6, 2007.

[65] R. M. a. S. GK, «Small-sample estimation of negative binomial dispersion, with applications to SAGE data.,» *Biostatistics,* vol. 9, pp. -11, 2008.

[66] L. H. a. R. M. Zhou X, «Robustly detecting differential expression in RNA sequencing data using observation weights.,» *Nucleic Acids Research,* vol. 42, p. 91, 2014.

[67] R. H. B. C. F. e. a. Irizarry, «Exploration, normalization and summaries of high density nucleotide array probe level data.,» *Biostatistics,* vol. 4, pp. 249-64, 2003.

[68] M. e. a. Dillies, «A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.,» *Briefings in Bioinformatics,* 2012.

[69] M. P. a. H. S, «phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data.,» *PLoS ONE,* vol. 8, n. 4, p. 61217, 2013.

[70] S. Hurlbert, «The nonconcept of species diversity: a critique and alternative parameters,» *Ecology,* vol. 52, n. 4, pp. 577-86, 1971.

[71] P. H. S. McMurdie, «Waste not, want not: why rarefying microbiome data is inadmissible.,» *PLoS Computational Biology,* vol. 10, n. 4, 2014.

[72] H. Jiang e W. Wong, «Statistical inferences for isoform expression in RNA-Seq,» *Bioinformatics,* n. 25, pp. 1026-32, 2009.

[73] S. D. C. J. e. a. Di Y, «The NBP negative binomial model for assessing differential gene expression from RNA-seq.,» *Stat Appl Genet Mol Biol,* vol. 10, n. 24, 2011.

[74] F. F. a. D. C. B, «Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis,» *Briefings in Functional Genomics,* n. 35, pp. 1-13, 2014.

[75] Y. Chen e D. McCarthy, «edgeR: differential expression analysis of digital gene expression data,» pp. 1-79, 2014.

[76] M. L. B. P. M. Ghodsi, «DNACLUST: accurate and efficient clustering of phylogenetic marker genes.,» *BMC Bioinformatics ,* vol. 12, n. 271, 2011.

[77] Q. G. G. T. J. C. J. Wang, «Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.,» *Appl. Environ. Microbiol. ,* n. 73, p. 5261–67, 2007.