G&I
Genomics & Informatics

Review article

eISSN 2234-0742
Genomics Inform 2019;17(1):e6
https://doi.org/10.5808/GI.2019.17.1.e6

Received: January 14, 2019
Revised: February 16, 2019
Accepted: February 20, 2018

*Corresponding author:
E-mail: malu.calle@uvic.cat

Statistical analysis of metagenomics data

M. Luz Calle*

Biosciences Department, Faculty of Science and Technology, University of Vic – Central University of Catalonia, Vic 08500, Spain

Understanding the role of the microbiome in human health and how it can be modulated is becoming increasingly relevant for preventive medicine and for the medical management of chronic diseases. The development of high-throughput sequencing technologies has boosted microbiome research through the study of microbial genomes and allowing a more precise quantification of microbiome abundances and function. Microbiome data analysis is challenging because it involves high-dimensional structured multivariate sparse data and because of its compositional nature. In this review we outline some of the procedures that are most commonly used for microbiome analysis and that are implemented in R packages. We place particular emphasis on the compositional structure of microbiome data. We describe the principles of compositional data analysis and distinguish between standard methods and those that fit into compositional data analysis.

Keywords: biomarkers, DNA sequence analysis, metagenome, microbiota, statistical models

Introduction

The study of the human microbiome and its role in human health is an active area of research. The human microbiome is involved in a large number of essential functions, like food digestion and modulation of the immune system, and alterations in microbiome composition may have important effects on human health. Many diseases have already been found to be associated with changes in the human microbiome. Different studies have shown that obesity is indeed partly determined by the composition of our gut microbiome. Chronic inflammatory skin conditions such as psoriasis, atopic dermatitis, acne and chronic skin ulcers have been associated to cutaneous microbiome changes. The colonic microbiota is suspected to be involved in the development of colorectal cancers. Inflammatory bowel diseases have long been associated to interactions between microbes and the host since the microbiome is essential for the activation of host immune responses. Microbial diversity is significantly diminished in Crohn disease. Early childhood antibiotic exposure has been associated with significantly increased risk for Crohn disease [1,2]. Understanding the role of the microbiome in human health and how it can be modulated is becoming increasingly relevant for preventive medicine and for the medical management of chronic diseases.

The terms microbiome and microbiota are used indistinctly to describe the community of microorganisms that live in a given environment. High-throughput DNA sequencing technologies have powered microbiome research by enabling the study of the genomes of all microorganisms of a given environment and a more precise quantification of microbiome abundances and function. Fig. 1 summarizes the main steps of a microbiome study: (1) microbial DNA extraction and sequencing according to two main approaches, amplicon sequencing and shotgun sequencing; (2) bioinformatics sequence processing; and (3) statistical analysis.

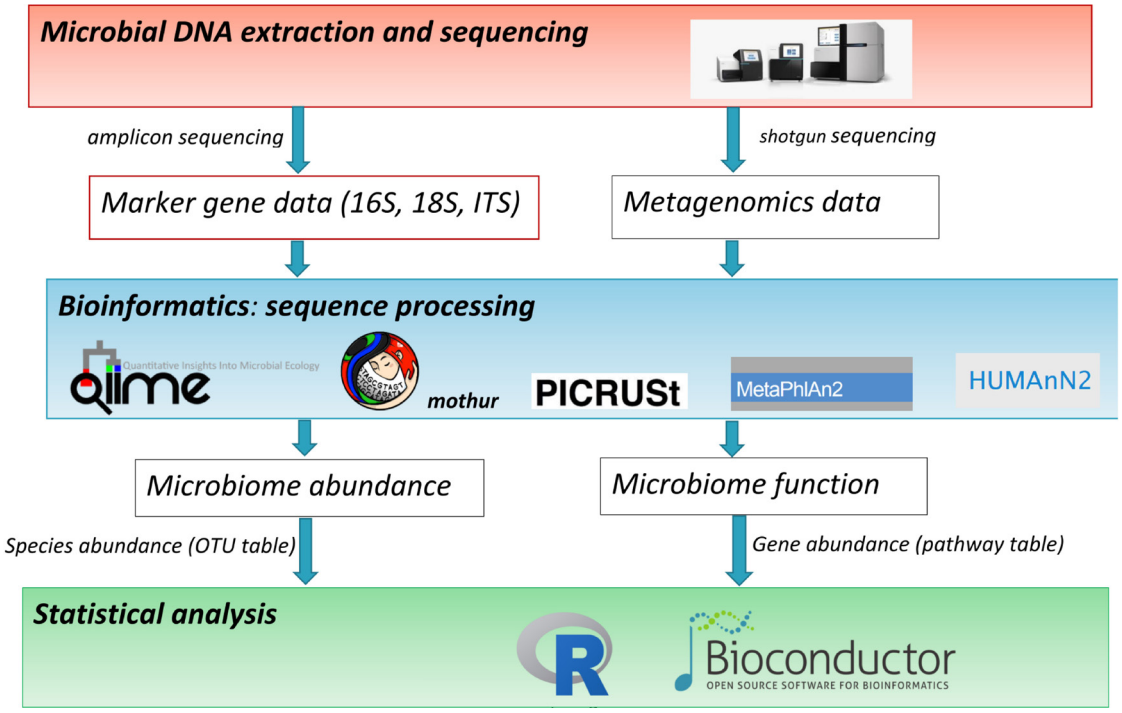

Fig. 1. Main steps of a microbiome study: (1) microbial DNA extraction and sequencing, (2) bioinformatics sequence processing, and (3) statistical analysis.

Amplicon sequencing relies on sequencing a phylogenetic marker gene after polymerase chain reaction (PCR) amplification. For bacteria and archaea, the marker gene is the 16S ribosomal RNA gene that encodes the RNA component of the small ribosomal subunit. The 16S rRNA gene contains both highly conserved areas and hypervariable sites, denoted as V1–V9. The conserved regions can be targeted with PCR primers while the hypervariable regions are specific to each microbial species and make possible to distinguish the different microbes. The V1–V3 and V4 regions are most commonly targeted. PCR amplification creates thousands to millions of copies (amplicons) of the DNA target region. PCR amplicons are then sequenced using high-throughput sequencing platforms and multiple nucleotide sequences, also known as reads, are obtained [3].

There are a number of bioinformatic pipelines available for processing microbiome 16S sequence data, the two most popular for amplicon sequencing are mothur [4] and QIIME [5]. Both pipelines are user-friendly and produce similar results. The bioinformatics pipeline consists of five main steps: Preprocessing and quality control filtering, operational taxonomic unit (OTU) binning, taxonomy assignment, construction of the abundance table and phylogenetic analysis.

Preprocessing and quality control filtering consists on first assign the sequences to samples (demultiplexing) and then sequences are quality filtered to remove too short sequences, too many ambiguous base pairs and chimeras. OTU binning is the process of clustering similar DNA sequences into OTUs, that is, groups of DNA sequences with at least 97% similarity. The different sequences assigned to an OTU are represented by a consensus sequence determined by the most common nucleotide at each position. Taxonomy assignment is then obtained by comparing OTU consensus sequences to microbial 16S rRNA reference databases such as GreenGenes (http:// greengenes.second.genome.com), SILVA (https://www.arb-silva.de), or RDP (http://rdp.cme.msu.edu). Taxonomy assignment provides the available annotation of each OTU to the different taxonomy levels (domain, kingdom, phylum, class, order, family, genus, and species). In practice, many OTUs are not completely annotated, especially for low taxonomy levels. Next, an OTU abundance table is built where each entry in the table corresponds to the number of sequences (reads) observed for each sample corresponding to each OTU. OTU tables may be extremely sparse with many OTUs only observed in a few samples. In this case it is convenient to agglomerate OTUs at broader taxonomic groups or taxa. The last step of the bioinformatics pipeline is phylogenetic analysis. Phylogenetic trees can be used to obtain phylogenetic distances between samples.

Shotgun metagenomics sequencing involves sequencing the total microbial DNA of a sample, instead of just a particular marker gene. With this technique, we can infer the relative abundance of every

https://doi.org/10.5808/GI.2019.17.1.e6

microbial gene and quantify specific metabolic pathways to predict the potential functionality of the entire community. This is achieved by mapping the obtained sequences against a database such as Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/pathway.html). A gene pathway table resulting from this type of functional study provides the number of sequences associated to a particular function for each sample. HumanN2 [6] and MetaPhlAn 2 [7] are two bioinformatics pipelines for metagenomics analysis.

From a statistical point of view, the output of both microbiome approaches, amplicon and shotgun sequencing, is similar: an abundance table of counts representing the number of sequences per sample for a specific taxon or the number of sequences matching a specific gene function. In this paper we illustrate the methodologies with data from 16S rRNA amplicon sequencing but most approaches also apply for microbiome shotgun metagenomics.

There are many reasons why the analysis of microbiome data is so challenging. On one hand, we face the usual challenges of count data analysis, i.e., skewed distribution, zero inflation and over-dispersion. Because of the experimental process and quality control filtering, microbiome data is very noisy and the total number of counts per sample is highly variable, which requires some normalization prior to the analysis so that the microbiome abundances among the different samples are comparable. Abundance tables are usually sparse since many species are infrequent. There is much redundant information because of co-abundance of many species. Moreover, the total number of counts per sample is constrained by the maximum number of sequence reads that the DNA sequencer can provide. This total count constraint induces strong dependencies among the abundances of the different taxa characterizing the compositional structure of microbiome data. Ignoring the compositionality of microbiome data may yield spurious results. In section 2, we describe the main principles of compositional data analysis.

The statistical analysis of microbiome abundance data usually starts with the normalization of the data followed by an exploratory study of the microbiome composition for the identification of possible data structures. The exploratory part consists of the analysis of diversity measures and their visualization through ordination plots, a term used in ecology to refer to several multivariate techniques for visualization of species abundance in a low-dimensional space. Subsequently, an inference analysis is performed where microbiome composition is tested for association with a variable of interest; this is known as differential abundance testing when the outcome of interest is dichotomous (i.e., disease status). These association tests can be multivariate, when the interest is to assess for global differences in microbial composition between sample groups, or univariate, with the aim of identifying which taxa are differentially abundant between sample groups. However, as we discuss later, univariate approaches for microbiome analysis are questionable and their results should be regarded with caution.

In sections 3 and 4, we describe the procedures that are commonly performed in a microbiome statistical analysis: normalization, diversity analysis, ordination and differential abundance testing, both, multivariate and univariate. This is not intended to be an exhaustive or systematic review of all the available methods. We outline some of the most widely used techniques for microbiome analysis, especially those that are implemented in R packages. We distinguish between standard methods and those that fit into compositional data analysis.

Microbiome Compositional Data

Microbiome data is compositional because the information that abundance tables contain is relative. In a microbiome abundance table, the total number of counts per sample is highly variable and constrained by the maximum number of DNA reads that the sequencer can provide. This total count constraint induces strong dependencies among the abundances of the different taxa; an increase in the abundance of one taxon implies the decrease of the observed number of counts for some of the other taxa so that the total number of counts does not exceed the specified sequencing depth. Moreover, observed raw abundances and the total number of reads per sample are non-informative since they represent only a fraction or random sample of the original DNA content in the environment. These characteristics of microbiome abundance data clearly fall into the notion of compositional data.

Compositional data are defined as a vector of strictly positive real numbers

$$x = (x_1, \cdots, x_k);\ x_i > 0, i \in \{1, \cdots, k\}$$

with a constraint or non-informative total sum. The elements of a composition are called components or parts. In a composition the value of each component is not informative by itself and the relevant information is contained in the ratios between the components or parts [8]. Except for the fact that microbiome abundance tables contain many zeros, microbiome data fit the definition of compositional data and, as already acknowledged by many authors [9,10], their analysis requires the use of a proper mathematical theory [11]. Aitchison introduced the log-ratio approach and laid the foundations of Compositional Data Analysis (CoDA).

Mathematically, the assertion that the relevant information is contained in the ratios between the components implies that two proportional compositions are equally informative and this induces equivalence classes of vectors carrying the same information. Two

vectors are compositionally equivalent if they are proportional. Each equivalence class has a representative in the unit simplex defined as:

$$S^k = \{ x = (x_1, \cdots, x_k), x_i > 0, \sum_{i=1}^{k} x_1 = 1 \}.$$

The simplex is thus the sample space of compositional data. In microbiome analysis, for example, both the raw counts and their transformation into relative abundances or proportions belong to the same equivalence class and they carry the same relative information.

Three important conditions should be fulfilled for a proper analysis of compositions: permutation invariance, scale invariance and sub-compositional coherence [11]. Permutation invariance states that a change in the order of the parts in the composition should not affect the results. Scale invariance establishes that any function used for the analysis of compositional data must be invariant for any element of the same compositionally equivalent class. Sub-compositional coherence requires that the results obtained when a subset of components is analyzed is coherent with the results for the whole composition. In the context of microbiome analysis this principle is important because we usually work with sub-compositions obtained after filtering out the most low-abundant taxa. Ignoring the compositional nature of microbiome data can result in spurious correlations and sub-compositional incoherencies.

Aitchison [11] put the basis of CoDA by introducing what is now called the Aitchison's log-ratio approach. The log-ratio analysis was introduced in order to meet the principle of scale invariance; as stated by Aitchison [11], "any meaningful (scale-invariant) function of a composition can be expressed in terms of ratios of its components." Because the logarithmic transformation makes ratios mathematically more tractable, the simplest invariant function is given by the log-ratio between two components, that is:

$$f(x) = \log\left(\frac{x_i}{x_j}\right), \; i, j \in \{1, \cdots, k\}.$$

The generalization of a log-ratio is a log-contrast function defined as a linear combination of logarithms of the components with the restriction that the sum of the coefficients is equal to 0:

$$f(x) = \sum_{i=1}^{k} a_i \log(x_i); so \;\; that \;\; \sum_{i=1}^{k} a_i = 0.$$

Log-contrast functions are suitable for CoDA because they are scale invariant.

As an alternative to working in the simplex, several data transformations have been proposed that transform compositional data to the real space where classical statistical analysis can be applied. All of them are based on log-ratios between components.

The additive log-ratio transformation (alr) is the first proposal introduced by Aitchison [11]. Taking one part as the reference, for instance x_k, the alr transformation is defined as:

$$alr(x_1, \cdots, x_k) = \left(\log\left(\frac{x_1}{x_k}\right), \ldots, \log\left(\frac{x_{k-1}}{x_k}\right)\right).$$

Aitchison also defined the centered log-ratio transformation (clr) to treat the parts symmetrically. The clr transformation is given by:

$$clr(x_1, \cdots, x_k) = \left(\log\left(\frac{x_1}{g(x)}\right), \ldots, \log\left(\frac{xk}{g(x)}\right)\right),$$

where $g(x) = (\prod x_i)^{1/k}$ is the geometric mean of the composition. One characteristic of the clr transformation is that the transformed components are restricted to have a sum equal to zero and this implies that some common statistical analyses cannot be applied after the clr transformation because of a singular covariance matrix.

The third alternative is the isometric log-ratio transformation (ilr) and consists in the representation of a composition given a particular orthonormal basis in the simplex. It overcomes the problem of the singular covariance matrix present in the clr-transformation. For a detailed description see Egozcue et al. [12].

Exploratory Analysis of Microbiome Data

The main element of a microbiome study is the microbiome abundance table, a matrix of counts, X, with n rows (samples) and k columns (taxa) where each entry x_{ij} provides the number of sequences (reads) corresponding to taxon *j* in sample *i*. Sometimes abundance tables are transposed, rows are taxa and columns are samples. Apart from the abundance table, other elements that may be available for microbiome analysis are the sample data, the taxonomy table, and the phylogenetic tree. Several R and Bioconductor packages, such as phyloseq, are designed to facilitate the integration of all these elements in a microbiome analysis [13].

Normalization

The large variability of the total counts per sample prevents meaningful comparisons of raw abundances between individuals. This is usually addressed through normalization of raw counts before the analysis. The most simple and frequently used normalization is the computation of relative abundances by dividing the raw abundances by the total number of counts per sample. Another popular normalization approach is rarefaction, which consists on subsampling the same number of reads for each sample so that all samples have the same number of total counts. Rarefaction is not recommended because it entails the loss of important information [14]. More sophisticated normalization techniques are implemented in some R packages, such as, DESeq [15] or edgeR [16], initially developed

https://doi.org/10.5808/GI.2019.17.1.e6

for RNA-seq analysis, that are also used for microbiome differential abundance testing. See Weiss et al. [17] for a comparison and discussion on the performance of different normalization methods for microbiome analysis.

CoDA techniques do not require the normalization step because the log-ratio approach involves working with ratios between components and this cancels the effect of the total counts per sample. Instead, CoDA methods entail the imputation of zeros. Microbiome abundance tables are sparse, they contain many zeros, and this should be properly addressed before compositional data methods can be applied. The simplest approach is to replace zeros by a small pseudo-count or to add a small constant to all the elements of the abundance matrix. As an alternative, Martín-Fernández et al. [18] propose the Bayesian-Multiplicative treatment, a zero replacement involving Bayesian inference and a modification of the non-zero values so that the original ratios between the non-zero components are preserved.

Diversity analysis

The diversity of the microbiome is an important indicator of the good or bad conditions of the ecosystem, with larger microbiome diversity being usually associated to better health status. Microbiome diversity can be assessed through multiple ecological indices that can be divided into two kind of measures, alpha and beta diversity. Alpha diversity measures the variability of species within a sample while beta diversity accounts for the differences in composition between samples. The R package vegan provides a large set of diversity measures [19].

Alpha diversity: within sample diversity

The most important measure of alpha diversity is richness, defined as the number of different species present in an environment. Richness is estimated by the observed richness, R_{obs}, the number of different species observed in the sample. The observed richness tends to underestimate the real richness in the environment, where the less frequent species are likely to be undetected. There are different indices that adjust for this and try to estimate the hidden part that has not been detected. One of the most extended richness measure is Chao1 index defined as

$$R_{Chao1} = R_{obs} + \frac{f_1(f_1 - 1)}{2(f_2 + 1)},$$

where f_1 is the number of species observed only once and f_2 is the number of species observed twice.

Another important indicator of alpha diversity is evenness, which measures the homogeneity in abundance of the different species in a sample. A commonly used measure of evenness is the Shannon index defined as

$$R_{S\ hannon} = -\sum_{i=1}^{k} p_i \log(p_i),$$

where p_i represents the relative abundances of the i-th taxon.

Beta diversity: between samples diversity

Beta diversity measures the differences in microbiome composition between samples. There is a wide range of ecological distances or dissimilarities for measuring how close are two microbial compositions. The most commonly used are Bray-Curtis, UniFrac and weighted UniFrac distances. We also define the Aitchison distance which is a proper distance for compositional data.

Let $p_1 = (p_{11},\ldots,p_{1k})$ and $p_2 = (p_{21},\ldots,p_{2k})$ denote the microbiome relative abundance of two different samples.

Bray-Curtis is defined as follows:

$$d_{BC}(p_1,p_2) = \frac{\sum_{i=1}^{k} |p_{1i} - p_{2i}|}{\sum_{i=1}^{k} (p_{1i} + p_{2i})}.$$

UniFrac family of distances [20] consider the phylogenetic tree that represents the evolutionary relationships among the different taxa. The phylogenetic tree can be obtained from the bioinformatic pipelines, such as mothur and QIIME. For a tree with r branches, let $b = (b1,\ldots,br)$ represent the length of the different branches in the phylogenetic tree, and $q_1 = (q_{11}, \ldots, q_{1r})$, and $q2 = (q_{21},\ldots, q_{2r})$ the relative abundances associated to each branch for the first and the second sample, respectively.

The unweighted UniFrac distance measures the relative length of those branches that lead exclusively to species present in only one of the two samples with respect to the total length of all branches in the tree:

$$d_U(b,q1,q2) = \frac{\sum_{i=1}^{r} bi\,|I(q_{1i} > 0) - I(q_{2i} > 0)|}{\sum_{i=1}^{r} b_i I(q_{1i} + q_{2i} > 0)}.$$

The unweighted UniFrac distance only takes into account the presence or absence of the taxa but Lozupone et al. [20] also introduced the weighted UniFrac distance that includes information on the relative abundance of each taxa and is defined as follows:

$$d_W(b,q1,q2) = \frac{\sum_{i=1}^{r} |q_{1i} - q_{2i}|}{\sum_{i=1}^{r} (q_{1i} + q_{2i}) I(q_{1i} + q_{2i} > 0)}.$$

For a proper CoDA analysis, a distance must be subcompositionally dominant, which means that the distance between two points in a multi-dimensional space should always be larger than their distance when projected in a lower dimensional space (sub-composition). Most commonly used distances in microbiome analysis, such as, the Bray-Curtis and the weighted and unweighted UniFrac distances are not sub-compositionally dominant, and this may induce

sub-compositionally incoherencies that question the reliability of the results of any distance-based analysis [8,11,21].

The Aitchison distance is a sub-compositionally coherent distance defined as the Euclidean distance after the clr-transformation of the compositions. Given two compositions x_1 and x_2, the Aitchison distance is given by

$$d_A(x_1,x_2) = d_E(clr(x_1),clr(x_2)),$$

where d_E denotes Euclidean distance.

Ordination

The goal of ordination plots is the visualization of beta diversity for identification of possible data structures. The multidimensional data is represented into a reduced number of orthogonal axes while keeping the main trends of the data and preserving the distances among samples as much as possible. Most commonly used ordination methods for microbiome data are principal coordinates analysis (PCoA), also known as multidimensional scaling, and non-metric multidimensional scaling (NMDS) [22,23].

PCoA an extension of Principal Components Analysis (PCA). Given a distance or dissimilarity matrix, D, PCoA performs eigenvalue decomposition of $D_c'D_c$ where D_c is the centered distance matrix. When D is the Euclidean distance, PCoA results exactly the same as PCA. Care must be taken with PCoA if the selected distance is not metric, because some eigenvalues may be negative and then, the graphical representation will not perform properly.

In order to avoid this problem NMDS is more commonly used. Also based on a distance matrix D, NMDS maximizes the rank-based correlation between the original distances and the distances between samples in the new reduced ordination space. The procedure starts with a random configuration and the optimal representation is obtained following an iterative procedure that at each steps improves the rank correlation.

Ordination plots can be obtained with the R and Bioconductor packages vegan and phyloseq, among others [13,19]. Alternatively, a CoDA ordination approach can be followed by performing PCA after the clr or ilr transformation as implemented by Le Cao et al. [24] in the context of the multivariate statistical framework mixMC.

Microbiome Statistical Inference

Multivariate differential abundance testing

Multivariate differential abundance testing refers to a global test of differences in microbial composition between two or more groups of samples. We can distinguish between distance-based or model-based approaches.

Permutational Multivariate Analysis of Variance Using Distance Matrices, PERMANOVA [25], is perhaps the most widely used distance-based method for multivariate community analysis. The null hypothesis of no differences in composition among groups is formulated by the condition that the different groups of samples have the same center of masses. Implemented in the function "adonis" of the vegan R package, it consists of a multivariate ANOVA based on dissimilarities. The variability within groups is compared against the variability between groups with the usual ANOVA F statistic, but partition of sums-of-squares is applied directly to dissimilarities. Significance is evaluated through permutations to generate a distribution of the pseudo F statistic under the null.

A related and popular distance-based approach is the analysis of similarities [26], implemented in the function "anosim" of the vegan R package.

An interesting model-based approach for multivariate microbiome analysis is Kernel machine regression (KMR), that extends PERMANOVA to a regression framework [27]. KMR is a semi-parametric regression model that includes a nonparametric component. The model can be expressed as a semiparametric linear regression model when the response variable is continuous

$$y_i = \beta_0 + \beta' Z_i + h(X_i) + \epsilon_i$$

or as a semiparametric logistic regression model for a dichotomous response variable

$$logit(y_i) = \beta_0 + \beta' Z_i + h(X_i) + \epsilon_i .$$

In the context of microbiome analysis, X is the microbiome abundance matrix and the non-parametric component $h(X)$ measures the relationship between microbiome composition and the outcome. This association can be tested according to the following hypothesis:

$$H_0: h(X) = 0 \quad vs \quad H_1: h(X) \neq 0 .$$

The nonparametric component is related to a Kernel matrix that is a transformation of the distance matrix D of pairwise distances between individuals. KMR is implemented for microbiome analysis in the R package MiRKAT [28]. KMR can be adapted to CoDA by using a subcompositionally dominant function, such as, the Aitchison distance. Rivera-Pinto [29] has implemented this adaptation in the R package MiRKAT-CoDA. The algorithm also includes a weighted version that allows the identification of the taxa that are more relevant for the joint association.

Among the different model-based methods for microbiome differential abundance testing we highlight the work by La Rosa et al. [30] that consider the Dirichlet-Multinomial distribution for hypothesis testing, power and sample size calculations. The proposed methods are implemented in the R package HMP.

https://doi.org/10.5808/GI.2019.17.1.e6

Le Cao et al. [24] propose the multivariate statistical framework mixMC where they perform sparse partial least squares discriminant analysis (sPLS-DA), implemented in the R package mixOmics [31]. In order to acknowledge the compositional structure of microbiome data, they apply sPLS-DA after the clr transformation. PLS-DA maximizes the covariance between linear combinations of the taxa and the response variable. The sparse version of PLS-DA uses Lasso penalized regression [32] and thus, it performs variable selection that enables the identification of the taxa that are most associated with the outcome.

Univariate differential abundance testing

When significant global differences in microbiome composition are detected between groups of samples, a natural question arises: which particular taxa are responsible of that global difference? A common strategy to answer this question is to test every taxa separately for association with the response variable. When the response variable is dichotomous this is known as univariate differential abundance testing.

Below we describe both, classical and CoDA approaches for univariate differential abundance testing. However, we advise that classical univariate approaches are notably affected by the compositional structure of microbiome data and their results, with large false discovery rates, might be questioned [17,33].

Nonparametric tests, like the Wilcoxon rank-sum test or the Kruskal-Wallis test, can be applied. However, more powerful parametric approaches are available, such as the Bioconductor packages edgeR [16] and DESeq2 [34], initially proposed for transcriptomics analysis (RNA-Seq data). Both fit a generalized linear model and assume that read counts follow a Negative Binomial distribution. The NB distribution extends the Poisson distribution by allowing the variance to be different from the mean. edgeR and DESeq2 mainly differ in the way they normalize the data. DESeq2 uses size factors that account for differences in sequencing depth between samples and shrinkage for large variances correction. edgeR can be implemented with different normalization methods but the most recommended is TMM, the trimmed mean of M-values normalization method, that indirectly attempts to overcome the problem of compositional DNA sequencing data ("the proportion of reads attributed to a given gene in a library depends on the expression properties of the whole sample rather than just the expression level of that gene") [16].

Two CoDA methods that explicitly accounts for the compositional nature of microbiome data are ANCOM [35] and ALDEx2 [36]. In ANCOM, the log-ratio of all pairs of variables is tested for differences in means. The number of significant results involving each variable is used to determine its significance. The ALDEx2 algorithm uses a Dirichlet-multinomial model to infer the multivariate abundance distribution from counts. After clr transformation it performs the Wilcoxon rank test (two groups) or Kruskal-Wallis tests (more than two groups).

Microbial signatures

Recently, Rivera-Pinto et al. [37] have proposed a new CoDA approach for microbiome analysis that is aimed to the identification of microbial signatures, groups of microbial taxa that are predictive of a phenotype of interest. The identification of microbial signatures involves both modeling and variable selection: modeling the response variable and identifying the smallest number of taxa with the highest prediction or classification accuracy. In order to fulfill the principles of CoDA, instead of analyzing individual abundances, we analyze the relative abundances between two groups of taxa, also referred as the abundance balance between the two groups, a concept that is formally defined as follows:

Let $x = (x_1,x_2,...,x_k)$ be the microbial composition of k taxa and, among these k taxa, let's consider two disjoint subgroups of taxa, group A and group B, with composition abundances denoted by x_A and x_B, each group with k_A and k_B different taxa and indexed by I_A and I_B, respectively. The abundance balance between A and B, denoted by $B(A,B)$, is defined as the log-ratio between the geometric mean abundances of the two groups of taxa as follows:

$$B(A,B) = C \cdot log \frac{\left(\prod_{i \in I_A} x_i\right)^{\frac{1}{k_A}}}{\left(\prod_{j \in I_B} x_i\right)^{\frac{1}{k_B}}},$$

where C is a normalization constant. The larger the values of balance $B(A,B)$, the more abundant is group A with respect to group B. Positive values of $B(A,B)$ arise when group A is more abundant than group B while negative values of $B(A,B)$ correspond to larger abundance of group B relative to group A abundance. A value of $B(A,B) = 0$ correspond to a perfect balance between the abundances of both groups of taxa.

The goal of the proposed algorithm is the identification of the two groups of microbial taxa, group A and group B, whose abundance balance $B(A,B)$ is most associated with an outcome of interest Y. For instance, for a binary outcome Y corresponding to disease status ($Y = 1$ for diesased and $Y = 0$ for not diseased), if we are able to identify the two groups of taxa A and B whose balance is associated with Y we may use $B(A,B)$, the relative abundance between groups A and B, as a microbial signature of disease risk. If large values of $B(A,B)$ are associated with $Y = 1$, we will infer that a person with larger relative abundances of group A with respect to group B will have higher risk of disease than other people with lower relative abundances between A and B.

The algorithm for the selection of microbial balances is implemented in the R package selbal. It starts with a first thorough search of the two taxa whose balance, or log-ratio, is most associated with the response variable. Once the first two-taxon balance is selected, the algorithm performs a forward selection process where, at each step, a new taxon is added to the existing balance such that the specified optimization criterion is improved (area under the receiver operating characteristic or mean squared error). The algorithm stops when there is no additional variable that improves the current optimization parameter or when the maximum number of components to be included in the balance is achieved. This number is established with a cross-validation procedure, which is also used to explore the robustness of the identified balance.

Discussion

In this work we present some of the techniques that are most commonly used for microbiome analysis. We place a particular emphasis on those methods that preserve the principles of compositional data analysis.

Classical methods that ignore the compositional nature of microbiome data can result in spurious correlations and sub-compositional incoherencies. This is especially relevant for classical univariate test where the strong dependencies between microbial abundances results in an important increase of type I error. Simulation studies show that the false discovery rate increases as the true-positive fold change increases and that it can achieve unacceptable extremely large values [17,33]. Moreover, from a biological point of view univariate approaches are questionable because they ignore that the microbiome is an ecosystem with complex interactions between its members and with the environment.

There is an increasing awareness of the need of using proper CoDA methods for microbiome analysis [10,38]. In this work we make clear that proper CoDA methods are available for all steps of a microbiome statistical analysis: normalization, diversity analysis, ordination and differential abundance testing, both, multivariate and univariate.

Normalization is not required and only zero imputation is needed. Diversity analysis and ordination can be performed after clr or ilr transformations, for instance, Aitchison distance and PCA. CoDA adapted Kernel machine regression can be used for multivariate differential abundance testing. Univariate approaches are not recommended. Penalized multivariate regression, such as sPLS-DA, is an alternative for the identification of the taxa that are most associated with the outcome. The algorithm selbal for the selection of microbial signatures is also an alternative to univariate selection of taxa when the main interest is prediction.

Even so, more research is still needed to fully understand the performance and limitations of the current available CoDA methods for microbiome analysis that will probably lead to their improvement or to the proposal of new approaches.

ORCID

M. Luz Calle: https://orcid.org/0000-0001-9334-415X

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was partially supported by the Ministerio de Economía y Competitividad, Spain, reference MTM2015-64465-C2-1-R.

References

1. Young VB. The role of the microbiome in human health and disease: an introduction for clinicians. BMJ 2017;356:j831.
2. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet 2012;13:260–270.
3. Amato KR. An introduction to microbiome analysis for human biology applications. Am J Hum Biol 2017;29:e22931.
4. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75:7537–7541.
5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7:335–336.
6. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 2018;15:962–968.
7. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 2015;12:902–903.
8. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Modelling and Analysis of Compositional Data. New York: John Wiley & Sons, 2015.
9. Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. It's all relative: analyzing microbiome data as compositions. Ann Epidemi-

 https://doi.org/10.5808/GI.2019.17.1.e6

ol 2016;26:322–329.
10. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol 2016;62:692–703.
11. Aitchison J. The Statistical Analysis of Compositional Data. London: Chapman & Hall, 1986.
12. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric logratio transformations for compositional data analysis. Math Geol 2003;35:279–300.
13. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 2013;8:e61217.
14. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 2014;10:e1003531.
15. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.
16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–140.
17. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 2017;5:27.
18. Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat Model 2015;15:134–158.
19. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. R package version 2.5-2. The Comprehensive R Archive Network, 2018. Accessed 2018 Dec 20. Available from: https://CRAN.R-project.org/package=vegan.
20. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005;71:8228–8235.
21. Aitchison J. A concise guide to compositional data analysis. 2005. Accessed 2019 Feb 14. Available from: http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf.
22. Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol 2007;62:142–160.
23. Greenacre M, Primicerio R. Multivariate Analysis of Ecological Data. Bilbao: Fundación BBVA, 2014.
24. Le Cao KA, Costello ME, Lakis VA, Bartolo F, Chua XY, Brazeilles R, et al. MixMC: a multivariate statistical framework to gain insight into microbial communities. PLoS One 2016;11:e0160169.
25. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol 2001;26:32–46.
26. Clarke KR. Non-parametric multivariate analyses of changes in community structure. Aust J Ecol 1993;18:117–143.
27. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 2007;63:1079–1088.
28. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, et al. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. Am J Hum Genet 2015;96:797–807.
29. Rivera-Pinto J. Statistical methods for the analysis of microbiome compositional data in HIV studies. Ph.D. Dissertation. Barcelona: University of Vic - Central University of Catalonia, 2018.
30. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS One 2012;7:e52078.
31. Le Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 2011;12:253.
32. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Methodol 1996;58:267–288.
33. Thorsen J, Brejnrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome 2016;4:62.
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
35. Mandal S, Van Treuren W, White RA, Eggesbo M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis 2015;26:27663.
36. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One 2013;8:e67019.
37. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. Balances: a new perspective for microbiome analysis. mSystems 2018;3:e00053. –18.
38. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol 2017;8:2224.