A new R library for discriminating groups based on abundance profile and biodiversity in microbiome metagenomic matrices

Clara I. Rodríguez Casado, Antonio Monleón-Getino

Abstract: The use of modern molecular techniques in the study of the human microbiome has revealed an extraordinary diversity of microorganisms. There has been great interest in associating specific groups of organisms with health and disease. However, little is known about changes in the structure of the microbial community due to changes in the health status of the host. To assess these changes in the composition and structure of the microbiome, we focused our interest on the modifications in relative species abundance using metagenomics. Species abundance patterns have been used extensively in ecology to describe the structure of living communities. We present **MetagenOutLDA**, a new R library that allows full statistical analysis of metagenomic matrices to be carried out using standard statistical techniques. This library performs three basic tasks divided into three functions: 1) estimating metagenomic abundance profiles (relative abundance of species) for each sample using robust regression and graphical representation; 2) estimating different metagenomic alpha biodiversity; and 3) performing discriminant analysis to distinguish between sample groups (healthy / sick, group1 / group2, etc.) and provide a percentage of correct classification using different methods (LDA, QDA, support vector machine, robust LDA, etc.). Case analyses are presented in this paper using **MetagenOutlineLDA** for a metagenomic study of the human microbioma with people affected by Crohn's disease. We present the mathematical base of the different functions involved in **MetagenOutlineLDA** and an explanation of both its use and the results obtained. The results seem to confirm the hypothesis that inflammatory diseases such as Crohn's disease alter not only the composition of the human microbiome, but also its structure. The innovative nature of this work lies in the development of a new library to support the metagenomic analysis of the microbiome and help confirm that the species abundance distribution of the microbial community discriminates more effectively than its composition, which can be helpful for diagnosing disease.

Index terms: metagenomic, microbiology, discriminant analysis, statistics, bioinformatics, medicine, R package

1 Introduction

Patterns of taxa abundance distributions are the result of the combined effects of historical and biological processes and, as such, are central to ecology. Anthropogenic disturbance is known to affect the structure of bacterial communities [1]. In this work we go a step further to determine whether these disturbances are present in the structure of bacterial communities (microbiota) of people affected by inflammatory diseases. Determination of microbiota structure may make it possible to identify the disease from a medical point of view.

Modern molecular techniques have revealed an extraordinary diversity of microorganisms, most of which are still uncharacterized. This is a major challenge for microbial ecologists and statisticians: how can we compare the microbial diversity of different environments when the vast majority of microbial taxa remain unknown? The solution lies in the use of nonparametric estimation of metagenomes using statistical techniques, associated parameter estimation techniques and phylogeny of the community applied to microbial ecology [2].The combination of these statistical techniques with those of molecular biology allows for rigorous estimation and comparison of microbial diversity in different environments, such as the presence of disease. This is a complex task due to the variability of possible situations and the scarcity of current knowledge, although it is interesting as an adjunct in the diagnosis of human diseases.

Researchers have often used values given by one or more diversity indices to quantify the diversity of species in a biological sample from an ecological point of view. Such indices include species richness, the Shannon index, the Simpson index and the complement of the Simpson index (also known as the Gini-Simpson index) [3]; [4]; [5].

Using replicated experimental treehole microcosms perturbed with different concentrations of the pollutant pentachlorophenol, Ager et al. (2010) [1] observed changes in the bacterial community structure using rank-abundance plots fitted with linear regression models. The slopes of the regression models were used as a descriptive statistic of changes in evenness over time [1].

The structure of bacteria has been studied before, and a reduction in bacteria richness and a change in the structure of the community are observed when there is a change such as a pollution event [1], but no observations have been made in the case of human diseases. None of the classical species abundance distribution models (log-series, geometric-series, log-normal) will fit a range [6] of communities in varying states of disturbance (unperturbed to perturbed) or impoverishment (species-rich to poor), which is why [1] propose a nonlinear regression model to describe richness and community structure and use the slope

of the regression line to describe changes in the structure and richness of the bacterial community.

Some research presents ecological distributions of health and adopts clinical metagenomic approaches, such as "Species abundance distributions and richness estimations in fungal metagenomics - lessons learned from community ecology" [7]. In that study, the authors observed that completely surveyed communities follow log-normal distributions, whereas power-law functions best describe incompletely surveyed communities. It is arguable whether the statistics behind those theories can be applied to voluminous next-generation sequencing data in microbiology by treating individual DNA sequences as counts of molecular taxonomic units (MOTUs). So here we define metagenomics as the study of the microbiome from the direct isolation of DNA in the environment.

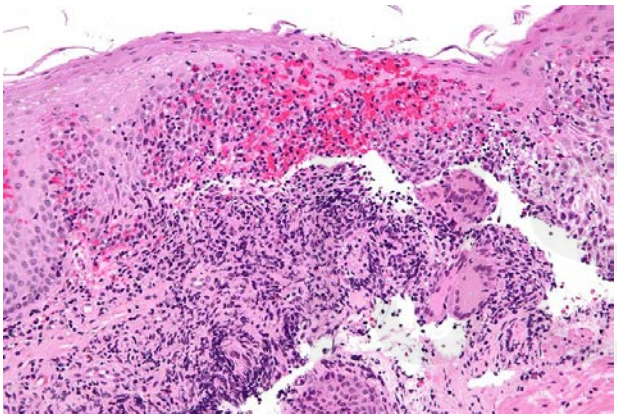

Fig. 1. Highly magnified micrograph of Crohn's disease. Biopsy of the oesophagus (from https://upload.wikimedia.org/wikipedia/commons/f/f9/Crohn%27s_disease_-_esophagus_-_high_mag.jpg., GNU li-cense)

In some human diseases, it has been shown that there is a relationship between the profiles of the metagenomic and metatranscriptomic biodiversity of microbial flora and the pathology of patients affected by periodontitis [8]; and more recently such a relationship has also been shown in other studies [9]. Those studies reveal the relationships between microbioma and human health. Study [9] reports the sequencing of 16 metagenomic samples collected from dental swabs and plaques representing four periodontal states. A strong correlation between microbiota community structure and disease status was observed and a core disease-associated community described. In the cases described [8], [10], [9], it was not possible to characterize the species abundance distribution or richness estimates using metagenomics, as occurred in [7]. However, periodontitis is not an isolated case, as a relationship seems to hold in other pathologies associated with the conduct of microbiological communities, such as Crohn's disease [10]. Figure 1 shows a highly magnified micrograph of an oesophageal biopsy from a patient with Crohn's disease.

The paper "Metagenomic Analysis of the Structure and Function of the Human Gut Microbiota in Crohn's Disease" also presents a case of differential behaviour between gut microbial communities in the case of healthy people and people affected by the disease [10]. Crohn's disease seems to be caused by a combination of environmental factors and genetic predisposition. Crohn's is the first genetically complex disease in which the relationship between genetic risk factors and the immune system has been understood in considerable detail in different studies [11].

The research carried out was initially based on the method used by Aget et al. [1] that involves observing changes in microbiome structure using rank-abundance plots fitted to linear regression models as a descriptive statistic for changes in evenness over time. We then also based it on the evidence that inflammatory diseases such as periodontitis and Crohn's disease alter not only the composition of the human microbiome, but also its structure, a complex trait that is difficult to measure and analyse. The aim of this paper is to present MetagenOutLDA, a new R library to help researchers in this field and confirm that the species abundance distribution of metagenomic microbial communities can discriminate groups of samples (e.g. disease/healthy) more effectively than the composition of the microbial community, and to show how this can help in disease diagnosis.

2 Materials and Methods

2.1 R package

MetagenOutLDA was created using the GNU project: R. R is a widely used free software environment and programming language for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS, and is supported by the R Foundation for Statistical Computing [12]. The R language is widely used among statisticians and data scientists for developing statistical software and data analysis. The popularity of R has increased substantially in recent years. [13].

The source code for the R software environment is written primarily in C, Fortran and R. R is freely available under the GNU General Public License on https://www.r-project.org/.

2.1 Metagenomic example to illustrate the use of MetagenOutLDA

The database Metagenomic Analysis of the Structure and Function of the Human Gut Microbiota in Crohn's Disease [10] illustrates the use of the R library MetagenOutLDA. In this study, 19 microbial metagenomic sequences were compared (12 patients affected by Crohn's disease and 7 healthy people). The central hypotheses proposed in the study are: (1) that specific members and/or functional activities of the gastrointestinal microbiota differ in patients with Crohn's disease and healthy individuals; and (2) that it will be possible to elucidate microbial signatures that correlate with the occurrence and progression of the disease by integrating data obtained from 16S rRNA-based molecular fingerprinting, metagenomic and metaproteomic approaches [10].

TABLE 1
FUNCTIONS CONTAINED ON METAGENOUTLDA

Function	Description
dmcmetagen()	A function to calculate and graphically represent a profile () of a metagenomic abundance using nonlinear regression and robust confidence intervals.
dmcbiodiv()	A function to calculate alpha-biodiversity indices (Shannon, Simpson and Inverse of Simpson).
dmcTable()	Generates a classification table, with the percentages of classification of different methods: LDA, QDA, RRLDA, MDA, SVM using the results obtained at dmcbiodiv() and dmcmetagen()

We have used this study to create a new R library that makes it possible to properly discriminate study groups and also allows other researchers with similar matrix types to do so.

Sample	1	2	...	k	Total
1	X_{11}	X_{12}	...	X_{1k}	$N_{1.}$
2	X_{21}	X_{22}	...	X2k	$N_{2.}$
...	⋮	⋮	⋱	⋮	⋮
p	X_{p1}	X_{p2}	...	X_{pk}	$N_{p.}$
Total	$N_{.1}$	$N_{.2}$	⋯	$N_{.k}$	N..

Fig. 2. Metagenomic matrix (M) structure (p rows: samples, k columns: taxa or OTU (operational taxonomic units)). This matrix format is used by the MetagenOutLDA library as a baseline for different statistical analysis.

The structure of a typical metagenomics matrix (M) is shown in Figure 2 [15]. The meaning of the rows and columns is also explained in the figure. This matrix shows the p samples in the rows (in our example 19 patients) and the taxa identified by the molecular method or the organism identified (OTU: operational taxonomic unit) in the columns. This matrix is usually named M and its dimension is defined by p files and k columns:

$$\mathrm{Dim}(M_{ij}) = p \cdot k \quad (1)$$

The $\log_2(M')$, where M' is the transpose of matrix M, is usually used in metagenomics and for MetagenOutLDA.

As a result of metagenomic analysis, M can be very large and usually has few samples and thousands of OTUs, most with small frequencies or 0 (Sparse matrix). Another characteristic is that each sample may be different:

$$M' = \begin{pmatrix} 45 & 12 & \dots & 2 \\ 22 & 8 & \dots & 6 \\ 12 & 13 & \dots & 0 \\ 6 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 2 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \end{pmatrix}$$

Finally, the number of files during the statistical analysis (i) can be different in each column j (samples or patients), since we discard the rows with an OTU frequency of 0 and usually transform the frequency in its $\log_2$(frequency), and the M and M' matrices are therefore algebraically difficult to manipulate and analyse, which constitutes a good argument for building this new library. Some authors [8] also recommend deleting the rows with an OTU of just 1, although this type of data filtering is highly variable and can lead to different results, but can easily

be implemented in the library.

2.2 Description of the MetagenOutLDA library

The **MetagenOutLDA** R library and the script used as an example, which contains all the results mentioned in the present article, can be downloaded from the Github repository:
www.github.com/amonleong/metagen

For more help on using R, and downloading and installing R libraries, go to:
https://www.r-project.org/

Within MetagenOutLDA, there is a matrix with the above mentioned data relating to the metagenomics of Crohn's disease described previously in [10].

All graphics, analysis and results discussed herein may be reproduced from the material available in the Github repository.

MetagenOutLDA has three easy-to-use functions described in Table 1 that analyse matrix M_{ij} (see Formula 1): 1) **dmcmetagen()**: to estimate the metagenomic profile of abundance and richness distribution (microbial community structure) of each patient using nonlinear regression, 2) **dmcbiodiv()**: to estimate different biodiversity indices, and 3) **dmcTable()**: to perform different discriminant analysis of the data and provide a percentage of correct classification.

For each example of analysis using the MetagenOutLDA function, we briefly describe the method used, the results of the use of the function, and the results obtained in the Crohn's disease example used in the present paper.

2.3 Mathematical base of the function dmcmetagen()

First of all, we estimate the parameters of a linear model or nonlinear polynomial multiple regression model (cubic model, or another model under consideration) of abundance within the microbiome from each matrix M' column (see 1 and Figure 2), taking into account that the number of files (i) may be different in each column j (samples or patients), since we discard the rows with an OTU frequency of 0 and usually transform the frequency in its log2(frequency) to normalize the distribution of potential Poisson probabilities f(x) (M_{ij} is the OTU frequency). Its species abundance profile is thus designated together with the goodness of fit (R^2). Subsequently, different biodiversity indicators are calculated. Finally, discriminant analysis is used to obtain a discriminant function that allows us to distinguish between health and disease. The classification matrix is used to assess the goodness of fit and quality of the classification by the function **dmcmetagen**.

Nonlinear regression model. The normal linear regression model can be represented as:

$$y_i = x_i^{'}\beta + \varepsilon_i \quad (2)$$

Where $x_i^{'}$ is a row vector of prediction for the i_{th} of n observations, usually with a 1 in the first position representing the regression constant β is the vector of regression parameters to be estimated and ε_i is a random error, assumed to be normally distributed, regardless of the errors in other observations, with expectation 0 and constant variance σ^2 (Gauss-Markov conditions), where the random error for each observation is:

$$\varepsilon_i \sim N(0, \sigma^2) \quad (3)$$

Where $N(0, \sigma^2)$ represents a Gaussian (normal) distribution of probability with mean 0 and variance σ^2.

In the more general normal nonlinear regression model [15]; [16] the function $f()$ that relates the response to the predictors (e.g. metagenomic abundance) is not necessarily linear:

$$y_i = f(\beta, x_i^{'}) + \varepsilon_i \quad (4)$$

As in the linear model, β is the vector of parameters and x_i' is a vector of predictors (however, in the nonlinear regression model, these vectors do not generally have the same dimension), where the error term for each sample is $\varepsilon_i \sim N(0, \sigma^2)$ [17].

The function dmcbiodiv() identifies which function $f()$ best fits the data observed; the function may be polynomial, exponential, logistic, etc. In our case, different functions were tested and the cubic function was found to be well suited to the abundance metagenomic data (R^2>0.9 and a good graphic fit). A possible model of abundance could be the cubic polynomial function, which has the form:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 + \varepsilon_{ij} \quad (5)$$

where y_{ij} is the predicted metagenomic abundance, i is the group (health, disease) and j is the sample (person).
Once the most representative metagenomic abundance model obtained by nonlinear regression has been established and the coefficient of determination of the model (R^2) determined, the means and confidence intervals of the model are estimated for each coefficient ($\beta_0, \beta_1, \beta_2, \ldots$) and study population.

2.4 Mathematical base of the function dmcbiodiv()

There is controversy regarding the nature of microbial communities [18]; [19]; despite this controversy, however, ecologists have made efforts to describe the properties of these communities. Species richness and abundance are the alpha diversity components that allow us to evaluate the community structure, which we conceive as the sum of its parts. Therefore, a relatively simple way to describe a community is through the study of the species richness and abundance within it. The term "richness" refers to the number of species that make up the community, while the term "abundance" refers to the number of individuals per species found in the community.

Species richness (S) is the number of species (hits) in the sample, which monotonically increases with the true number of species (hits) in the community [19]; [21]. Here, we define S as the taxonomic content of the metagenomic dataset of the number of 16S rRNA bands detected in the metagenomic analysis, using BLAST-P over the database MEGAN 3.7.5, as mentioned above. Metagenomic abundance and richness profiles for the samples analysed (Crohn's disease) are presented in Figure 3.

In addition, Whittaker (1972) [21] described three terms for measuring biodiversity over spatial scales: alpha, beta and gamma diversity. Alpha diversity refers to the diversity within a microbioma ecosystem, and is usually expressed as the number of species or hits found in the MEGAN 3.7.5 database (i.e. species richness) in that ecosystem. [22]. Some authors have suggested the use of the representative and widely recognized Shannon or Shannon-Weaver index (H'), or the Simpson index (D). Both are common in the ecology literature and are widely used to measure microbial community diversity [23]. However, these information-based indices are usually based on the quantification of species abundance [19] and, as mentioned above, defining bacterial species is no trivial task.

The Shannon diversity index (H') is one of the most enduring methods of measuring overall biodiversity used in diversity studies. H' is maximized when all species (hits) have the same number of samples (N), and is calculated from:

$$H' = -\sum_{i=1}^{N} p_i \log_2(p_i) \quad (6)$$

where p_i is the proportion of the community represented by species (hits) *i*, and the summation is over all species (hits). The most common practice is to use natural logarithms, although some argue for base = 2, which makes sense but no real difference. *H'* represents the uncertainty in predicting the species of an individual chosen at random [23].
The Simpson index, *D*, is a derivation of the original Simpson index, λ, which was the first diversity measurement proposed, where individuals chosen at random belong to the same species. 0 is infinite diversity and 1 is no diversity (all the same species). It is calculated from:

$$D = 1 - \lambda = 1 - \sum_{i=1}^{N} p_i^2 \quad (7)$$

The disadvantage of *S*, *H'* and *D* is that they are successively more sensitive to evenness [23], [20].

Moreover, parametric confidence intervals are estimated for the different estimators of the regression model coefficients and for the biodiversity indices.

2.5 Mathematical base of the function dmcTable()

This relates to analysis to discriminate between the populations considered**.** The problem of discrimination appears in many situations in which elements must be classified using incomplete information and was originally described by [24]. The goals of descriptive discriminant analysis include the following: 1) To identify the relative contribution of the variables (in this case the distribution associated with metagenomic abundance) to the separation of the groups (in the current example: disease and health), and also to find the optimal plane on which the points can be projected to best illustrate the discrimination between the groups, 2) To predict or allocate observations to groups, in which linear functions of the variables (classification functions) are used to assign an individual sampling unit to one of the groups. The values measured in the observation vector for an individual or object are evaluated by the classification functions to find the group to which the individual is most likely to belong.
To perform discriminant analysis between the species abun-

Authors
- *Clara I. Rodríguez Casado is currently teaching statistics at the University of Barcelona (Section of Statistics, Departament of genetics, Microbiology and Statistics, Faculty of Biology), Barcelona, Spain. Email: clara.rodriguez@ub.edu*
- *Correspondence author: Antonio Monleón-Getino is currently teaching statistics at the University of Barcelona (Section of Statistics, Departament of genetics, Microbiology and Statistics, Faculty of Biology), Barcelona, Spain. Statistics and Bioinformatics Research Group (**GRBIO**). Email: amonleong@ub.edu*

dance profiles of the health and disease groups, there are several possible approaches [25]. Here we present the classical discriminant analysis (LDA) developed by Fisher in 1936 [24], which is based on multivariate normality of the variables and is optimal under this assumption. If all the variables are continuous, it is often possible to transform the variables in the same ways used for normal datasets, even though the original data are not normal. It is important to remember that we are using estimated coefficients ($\beta_0, \beta_1, \beta_2, ...$) and biodiversity indices (*H'*, *D*, …) as input variables, and normality is not guaranteed. This sort of analysis is frequently used in statistical pattern recognition [25], but has recently appeared in studies related to metagenomics, such as the paper "Feature selection in omics prediction problems using cat scores and false nondiscovery rate control" [26].

Let P_1 and P_1 be the two populations studied (health and disease), where we have defined a random vector variable, x, with p variables (measures derived from metagenomic analysis: estimations of parameters associated with abundance profiles, and biodiversity indices). We assume that x is absolutely continuous and that the density functions of both populations, f_1 and f_2, are known. We are going to study the problem of classifying a new element, x_0, with known values of p variables, as one of these two populations. If we know the prior probabilities π_1 and π_2, with $\pi_1 + \pi_2 = 1$, and that the element comes from one or other of the two populations, its probability distribution will be a mixed distribution:

$$f_x = \pi_1 f_1(x) + \pi_2 f_2(x) \tag{7}$$

and once x_0 has been observed, we can compute the posterior probabilities that the element belongs to each of the two populations: $P(i/x_0)$, with i = 1, 2. These probabilities are calculated using Bayes' theorem:

$$P(1 \mid x_0) = \frac{P(x_0 \mid 1)\pi_1}{\pi_1 f_1(x) + \pi_2 f_2(x)} \tag{8}$$

We classify x_0 in the most probable posterior population. Since the denominators are equal, we classify x_0 in P_1 if:

$$\pi_2 f_2(x_0) > \pi_1 f_1(x_0) \tag{9}$$

We are going to apply this analysis to the case in which f_1 and f_2 are normal distributions with different mean vectors, but with identical covariance matrices. In order to establish a general rule, let us suppose that we wish to classify a generic element x, which, if it belongs to the population i = 1, 2, has a density function of:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\left|V\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'V^{-1}(x-\mu_i)\right\} \tag{10}$$

The optimal decision is to classify the element in population P_1 if:

$$\frac{f_2(x)\pi_2}{c(2\mid 1)} > \frac{f_1(x)\pi_1}{c(1\mid 2)} \tag{11}$$

Since both terms are positive, taking logarithms and replacing $f_i(x)$ with its corresponding expression, the above equation becomes:

$$-\frac{1}{2}(x-\mu_2)'V^{-1}(x-\mu_2) + \log\frac{\pi_2}{c(2\mid 1)} > -\frac{1}{2}(x-\mu_1)'V^{-1}(x-\mu_1) + \log\frac{\pi_1}{c(1\mid 2)} \tag{12}$$

Letting D_i^2 be the Mahalanobis distance between the observed point, x, and the mean of population i, defined by:

$$D_i^2 = (x-\mu_i)'V^{-1}(x-\mu_i) \tag{13}$$

we can then write:

$$D_1^2 - \log\left(\frac{\pi_1}{c(1\mid 2)}\right) > D_2^2 - \log\left(\frac{\pi_2}{c(2\mid 1)}\right) \tag{14}$$

And assuming that the costs and prior probabilities are equal, $c(1|2) = c(2|1)$; $\pi_1 = \pi_2$. The above rule can be simplified as: classify into 2 if $D_1^2 > D_2^2$.

Or, rather, classify the observation into the population with the smallest Mahalanobis distance [27] .

The function dmcmetagen provides a leader board featuring the results of linear discriminant analysis (LDA) with and without cross-validation and other more complex techniques, such as: quadratic discriminant analysis (QDA), robust regularized linear discriminant analysis (RRLDA), multiple discriminant analysis (MDA) and support vector machine (SVM). We used all these techniques to avoid the effects of non-normality and determine which is the most effective for separating the groups.

We used cross-validation, which is a statistical technique for estimating the performance of a predictive model. This model validation technique assesses how the results of a statistical analysis will generalize to an independent dataset. It is mainly used here to estimate how accurately a predictive model will perform in a new sample.

3 Results and Discussion

Here we present the use and parameters of the three functions that comprise the MetagenOutLDA library, which has the three easy-to-use functions described in Table 1 and presented in detail here.

3.1 Use of dmcmetagen()

To obtain the metagenomic distribution profile, a nonlinear polynomial multiple regression model (cubic model, see Formula 5) is fitted to each of the subjects in the sample. The parameters of the regression model and the goodness of fit in the form of the coefficient of determination (R^2) are reported, with robust confidence intervals in order to avoid the possible effects of non-normality of the M_{ij} data. The function calls the linear model lm() function in R to fit the regression model.

The function has the following arguments:

dmcmetgen(Mat,regres_type, group, label_group, conf_int=0.95, printGrups=TRUE, print=FALSE, adjust=TRUE, robust = TRUE, order = TRUE, graphTitle="")

The Mat parameter must be the matrix M ($\mathrm{Dim}(M_{ij}) = p \cdot k$, see Figure 2 and Formula 1) with data for the subjects in all the groups we want to study (group1/group2, healthy/disease, etc.). The data matrix must have a specific format, the subjects must be in the columns and the data (hits) must be in the rows; and the values of the data matrix must have the format shown in Figure 1. The order argument is used to indicate when it is necessary to order the data in the matrix. It is set to TRUE by default. The regres_type argument indicates to the function what type of regression model must be fitted to the data to estimate the parameters (linear or cubic model).

The purpose of this function is to obtain the parameter estimators (see Formula 5) of the regression model to subsequently use them for discriminant analysis. In order to use the results obtained with dmcmetagen(), it is necessary that the results differentiate the group to which each of the subjects belongs. In order to obtain this information, the arguments group reports the group to which the subject belongs and it must be a vector v that indicates this (1, 1, …, 2, 2); while and label_group reports the names of the different groups and should be a vector that indicates the different names of the groups (disease, disease, …, health, health).

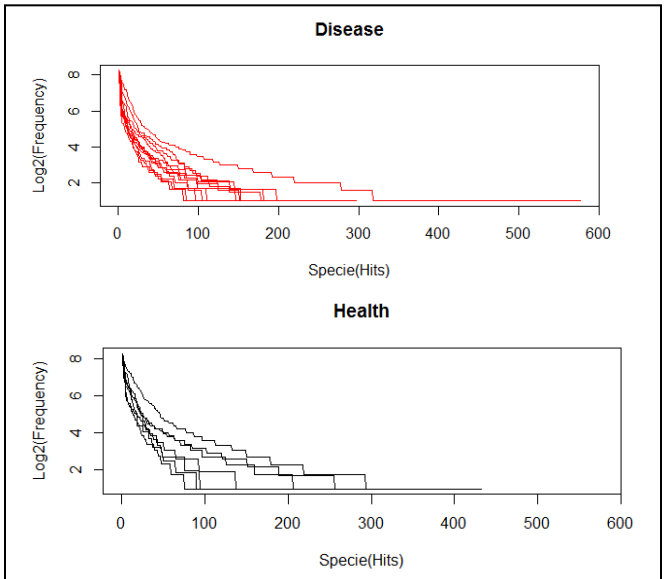

Figure 2: Metagenomic abundance profiles for the samples analysed, represented using the MetagenOutLDA library and the function dcmetagen(). The X axis represents species (hits or OTU: operational taxonomic unit) and the Y axis represents Log_2 of the OTU frequency (abundance) for every healthy (H) or disease (D) group for Crohn' s disease samples.

The function returns a matrix with the estimates of the parameters ($\beta_0, \beta_1, \beta_2, ...$) for the selected regression model, with a confidence interval of 95% and using the robust method by default, which can be modified by the arguments of the function conf_int and robust, respectively. In addition to these numerical results, graphic results are also obtained (see Figure 2); a graph for each of the groups is represented by default, printGrups=TRUE, without title, graphTitle="". It is also possible to represent all metagenomic profiles in the same graph by setting the print argument to TRUE.

TABLE 2

PERCENTAGE OF INDIVIDUALS CORRECTLY CLASSIFIED (HEALTH/DISEASE) USING DIFFERENT DISCRIMINANT ANALYSES. CV = CROSS VALIDATION

Num	β_0	β_1	β_2	β_3	Shannon Index (*H'*)	Simpson Index (*D*)	Gp
1	6.4518(6.3564,6.5473)	-0.0335(-0.0349,-0.0321)	1e-04(1e-04,1e-04)	0(0,0)	5.17	0.99	H
2	6.7989(6.6996,6.8982)	-0.0678(-0.0709,-0.0647)	3e-04(3e-04,3e-04)	0(0,0)	4.60	0.98	H
3	6.1528(6.0083,6.2974)	-0.0686(-0.074,-0.0632)	3e-04(3e-04,4e-04)	0(0,0)	4.37	0.96	H
4	6.2138(6.0115,6.4161)	-0.1473(-0.1636,-0.131)	0.0017(0.0014,0.0021)	0(0,0)	3.77	0.94	H
5	6.3078(6.1271,6.4885)	-0.1365(-0.1506,-0.1223)	0.0013(0.001,0.0016)	0(0,0)	3.84	0.95	H
6	5.47(5.3442,5.5958)	-0.0683(-0.0728,-0.0638)	4e-04(3e-04,4e-04)	0(0,0)	4.70	0.98	H
7	5.7814(5.6349,5.9279)	-0.0921(-0.1004,-0.0839)	7e-04(6e-04,8e-04)	0(0,0)	4.34	0.97	H
8	7.4564(7.2458,7.6671)	-0.1019(-0.1096,-0.0942)	6e-04(5e-04,7e-04)	0(0,0)	3.54	0.91	H
9	6.5917(6.4308,6.7525)	-0.1045(-0.1143,-0.0946)	7e-04(6e-04,9e-04)	0(0,0)	4.02	0.96	H
10	6.0485(5.8512,6.2458)	-0.132(-0.1429,-0.121)	0.0012(0.0011,0.0014)	0(0,0)	4.07	0.96	H
11	6.0846(5.9697,6.1995)	-0.0618(-0.0664,-0.0573)	3e-04(2e-04,3e-04)	0(0,0)	4.61	0.98	H
12	5.8261(5.7283,5.9239)	-0.0575(-0.0603,-0.0546)	3e-04(2e-04,3e-04)	0(0,0)	4.90	0.98	H
13	5.6022(5.5118,5.6927)	-0.0944(-0.0997,-0.089)	6e-04(6e-04,7e-04)	0(0,0)	4.40	0.98	D
14	4.3846(4.2993,4.4699)	-0.0458(-0.0488,-0.0428)	2e-04(2e-04,2e-04)	0(0,0)	5.15	0.99	D
15	4.2704(4.1503,4.3904)	-0.0568(-0.0627,-0.0508)	3e-04(2e-04,4e-04)	0(0,0)	4.81	0.99	D
16	6.1655(6.0332,6.2979)	-0.1387(-0.1488,-0.1286)	0.0013(0.0011,0.0015)	0(0,0)	4.01	0.97	D
17	5.7139(5.6412,5.7867)	-0.0339(-0.0354,-0.0325)	1e-04(1e-04,1e-04)	0(0,0)	5.42	0.99	D
18	5.309(5.2109,5.4071)	-0.0425(-0.0452,-0.0398)	2e-04(1e-04,2e-04)	0(0,0)	5.16	0.99	D
19	5.3231(5.5878,6.2196)	-0.0358(-0.1003,-0.0577)	1e-04(3e-04,8e-04)	0(0,0)	5.35	0.99	D

Gp= group

3.2 Use of dmcbiodiv()

This function has the following arguments:

dmcbiodiv(Mat,order= TRUE,index=c("shannon","simpson"))

As in the case of the function dmcmetagen() , this function also needs two arguments Mat, the matrix M_{ij} with the data, and order, the logical argument indicating whether it is necessary to order the data.

In addition to these two arguments, the argument index indicates which of the alpha diversity indices must be calculated (H', D, inv D). The different values that this argument can take are: "shannon", "simpson" or "invsimpson".

The purpose of this function is to obtain alpha biodiversity indices in order to use them when subsequently performing a discriminant analysis to classify the data correctly. The function returns an array with the biodiversity index values indicated for each of the subjects (all groups) that are present in the dataset. The biodiversity indices for each sample are shown in Table 2, with the estimation of coefficients done by dmcbiodiv().

As complementary results, Table 2 does not present statistically significant differences ($p>0.05$) between the Shannon (*H'*) and Simpson (*D*) indices between the groups (disease and health). Otherwise, significant statistically differences ($p<0.05$) between coefficients ($\beta_0, \beta_1, \beta_2, ...$) of the groups studied (disease and health) are observed. These statistical differences can confirm that the species abundance distribution of the metagenomic microbial community (microbial community structure) discriminates better than its composition (alpha biodiversity).

3.3 Use of dmcTable()

This function has the following arguments:

dmcTable(Matmet,Matbd,label_group,lambda=0.2,hp=0.75,NameCoeff=c("B0","B1","B2","B3","Shannon","Simpson"), tol=0.0001)

The arguments Matmet and Matbd are two matrices with the regression model coefficients and the biodiversity index, respectively. Both matrices must have the same row dimension (each subject is in a row), but the column dimension is variable. With the label_group argument, we indicate to the function the names of the different groups in the study.

The purpose of this function is to generate a summary table

(see Table 2) with the classification percentages of the different classification methods mentioned above, in order to determine which is better and whether it is possible to differentiate groups and thus create a diagnosis tool. This requires an indication of the data to be taken into account when the discriminant analysis is carried out. For this, we use the NameCoeff argument, which must be a vector with the parameter estimator ($\beta_0, \beta_1, \beta_2, ...$) and the biodiversity index (H', D) we want to use for the discriminant analysis.

TABLE 3
PERCENTAGE OF INDIVIDUALS CORRECTLY CLASSIFIED (HEALTH/DISEASE) USING DIFFERENT DISCRIMINANT ANALYSIS. CV = CROSS VALIDATION

Discriminant Method	healthy	Crohn	Total
LDA without CV	0.86	0.75	0.79
LDA with CV	0.71	0.75	0.74
QDA	1.00	1.00	1.00
RRLDA	0.71	0.83	0.79
MDA	1.00	1.00	1.00
SVM	1.00	0.86	0.89

The different methods are summarized in Table 3. The results of this function are: linear discriminant analysis (LDA) (with and without cross-validation), quadratic discriminant analysis (QDA), robust regularized linear discriminant analysis (RRLDA) (this method needs two specific arguments, lambda and hp), multiple discriminant analysis (MDA) and support vector machine (SVM).

To avoid misclassification due to the introduction of parameters that are below a certain value, the argument tol is used to indicate tolerance when entering the parameters in the classification model.

As Table 3 shows, it is possible to discriminate between groups using the library with a correct classification of between 75% and 100%. This result is encouraging as it suggests the library tool presented here can perform correct diagnosis between groups through metagenomic analysis. Unfortunately, such genetic analysis types are very expensive to carry out and could not be validated with larger samples that verify and validate their use.

4 CONCLUSION

Here we present the MetagenOutLDA R library: a set of three R robust and easy-to-use functions to calculate and discriminate OTU abundance profiles within microbial communities (health, disease). We explain how to use it by means of examples of people affected by Crohn's disease.
The function **dmcmetagen()** first fits the metagenomic profile of the abundance and richness distribution of each patient using nonlinear regression and robust confidence intervals of the parameters estimated. The function **dmcTable()** then analyses the biodiversity of the profile (the Shannon and Simpson indices); and finally the function **dmcmetagen()** performs discriminant analysis of the data obtained.
Early results indicate that it is possible to discriminate between healthy and affected groups in Crohn's disease (75%-100% correctly classified) based on metagenomic analysis of the microbioma. Furthermore, this work is novel in that it confirms that the species abundance distribution of the metagenomic microbial community (microbial community structure) discriminates better than its composition (alpha biodiversity) in these cases, and this can be helpful in the diagnosis of disease.

ACKNOWLEDGMENTS

The authors thank Dr Jorge Frías-Lopez from the Departament of Microbiology at the Forsyth Institute (Cambridge, MA, USA) for help and support, and Susan Keddie for revising the language and providing advice on the English.

REFERENCES

[1.] Ager D, Evans S, Li H, Lilley AK, van der Gast CJ. 2010. Anthropogenic disturbance affects the structure of bacterial communities. Environmental 2920.2009.02107.x. Epub 2009 Nov 25

[2.] Bohannan BJM, Hughesy J. 2003. New approaches to analyzing microbial biodiversity data. Current Opinion in Microbiology 6:282–287.

[3.] Ríos M, Monleón-Getino T. A graphical study of tuberculosis incidence and trends in the WHO's European region (1980-2006). Eur J Epidemiol. 2009;24(7):381-7. doi: 10.1007/s10654-009-9347-6. Epub 2009 May 19. PubMed PMID: 19452128.

[4.] Maceda-Veiga, A.; Monleón-Getino, A.; Caiola, N.; Casals, F.; De Sostoa, A. (2010). Fish assemblages in the iberian mediterranean watersheds over time: biodiversity, conservation status and introduced species.

Freshwater Biology, 15, 1-13. http://www3.interscience.wiley.com/journal/123353970/abstract ISSN: 0046-5070.

[5.] Figuerola, B.; Monleón-Getino, T.; Ballesteros, M.; Avila, C. (2012). Spatial patterns and diversity of briozoan communities from the Southern Ocean: South ShedlandIslands, Bouvet Island and Eastern Weddell Sea. Systematics And Biodiversity, 10(1), 109-123.

[6.] Tokeshi M. 1993. Species abundance patterns and community structure. Advanced Ecolology Res 24: 112-186

[7.] Unterseher M, Jumpponen A, Pik M, Tedersoo L, Moora M, Dormann CF, Martin Schninttler. Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology. Molecular Ecology (2011) 20, 275–285.

[8.] Frias-Lopez J, Duran-Pinedo A. Effect of periodontal pathogens on the metatranscriptome of a healthy multispecies biofilm model. J Bacteriol. 2012 Apr;194(8):2082-95. doi: 10.1128/JB.06328-11. Epub 2012 Feb 10. PubMed PMID: 22328675; PubMed Central PMCID: PMC3318478.

[9.] Jinfeng Wang, Ji Qi, Hui Zhao, Shu He, Yifei Zhang, Shicheng Wei, Fangqing Zhao 2013. Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. Nature Scientific Reports 3, Article number: 1843

[10.] Fraser-Liggett CM. 2010. Baltimore Metagenomic Analysis of the Structure and Function of the Human Gut Microbiota in Crohn's Disease. Natureprecedings 10.1038/npre.2010.4958.1.

[11.] Braat H, Peppelenbosch MP, Hommes DW (August 2006). "Immunology of Crohn's disease". Annals of the New York Academy of Sciences 1072: 135–54.

[12.] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[13.] Wikipedia. R programming language 22/7/2016 (https://en.wikipedia.org/wiki/R_(programming_language))

[14.] La Rosa PS, Elena Deych, Berkley Shands and William D.Shannon (2016). HMP: Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from HMP. R package version 1.4.3. http://CRAN.R-project.org/package=HMP . Mayo 2016.

[15.] Bates, D. M. & D. G. Watts. 1988. Nonlinear Regression Analysis and Its Applications. New York: Wiley.

[16.] Gallant, A. R. 1975. "Nonlinear Regression." The American Statistician 29:73—81.

[17.] Fox J. 2002. Nonlinear Regression and Nonlinear Least Squares. Appendix to An R and S-PLUS Companion to Applied Regression. The R Project for Statistical Computing [From http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonlinear-regression.pdf]

[18.] Curtis, T. P. & Sloan, W. T. Exploring microbial diversity - A vast below. Science 309, 1331-1333 (2005).

[19.] Magurran, A. E. Measuring biological diversity (Blackwell Publishing., 2004).

[20.] Magurran, A.E., 1988. In: Ecological Diversity and its Measure ment. Princeton University Press, Princeton, NJ.

[21.] Whittaker, R.H. (1972). Evolution and measurement of species diversity. Taxon, 21, 213-251.

[22.] Meffe, G.K., L.A. Nielsen, R.L. Knight, and D.A. Schenborn. (2002). Ecosystem management: adaptive, community-based conservation. Washington, D.C., U.S.A: Island Press.

[23.] Hedrick DB , Peacock A, Stephen JR, Macnaughton SJ, Bruggemannb J, Whiteb D Measuring soil microbial community diversity using polar lipid fatty acid and denaturing gradient gel electrophoresis data. Journal of Microbiological Methods 41 (2000) 235 –248.

[24.] Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics 7 (2): 179–188.

[25.] McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience.

[26.] Ahdesmäki, M.; Strimmer K. (2010) "Feature selection in omics prediction problems using cat scores and false nondiscovery rate control", Annals of Applied Statistics, 4 (1), 503–519.

[27.] Pena D. 2014. Discriminant analysis. Textos de docencia cap 13. Universidad Carlos III. [From http://halweb.uc3m.es/esp/Personal/personas/dpena/docencia/Sub-ingcf13.pdf]

OTHER REFERENCES

[1.] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.

[2.] Dalevi D et al. 2008. Annotation of metagenome short reads using Proxygenes. Bioinformatics (2008) 24: i7-13

[3.] Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. J. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge UK.

[4.] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res. 2014 Jan;42.

[5.] Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003 Jan 1;31(1):371-3.

[6.] Huson, DH and Mitra, S (2012). Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN. Evolutionary Genomics: Statistical and Computational Methods, ed. by Maria Anisimova. Springer, chap. 17, pp. 415-429.

[7.] Kanehisa M, Goto S (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes". Nucleic Acids Res 28 (1): 27–30.

[8.] Marrugan and Philip. 2001. Implications of species loss in freshwater fish assemblages. Ecography.Volume 24, Issue 6, pages 645–650,

December 200

[9.] T. Monleon-Getino, J. Frias-Lopez . New statistical function to discriminate metagenomic microbial community relative species abundance profiles. Comptat 2014. Book of Abstracts (Geniva, Swithzeland). C1584

[10.] Noguchi, H. et al. (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes.DNARes.15, 387-396.

[11.] Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. Jan 1;28(1):33-6.

[12] Elements of Statisical Learning - Data Mining, Inference and Prediction" (2nd edition, Chapter 12) by Hastie, Tibshirani and Friedman, 2009, Springer