Journal of Advanced Statistics, Vol. 4, No. 4, December 2019
https://dx.doi.org/10.22606/jas.2019.44001

Shannon Entropy Ratio, a Bayesian Biodiversity Index Used in the Uncertainty Mixtures of Metagenomic Populations

Toni Monleón-Getino[1,2,3*], Clara I Rodríguez-Casado[1,3] and Pablo Emilio Verde[4]

[1] Section of Statistics. Department of Genetics, Microbiology and Statistics. University of Barcelona, Barcelona, Spain

[2] GRBIO. Research Group in Biostatistics and Bioinformatics

[3] BIOST[3]. Research Group in Clinical Statistics, Bioinformatics and Computacional Biodiversity

[4] Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

Email: amonleong@ub.edu

Abstract. The microbial communities contain a unique complexity that makes difficult studying their diversity. Unfortunately the culture of microorganisms is very complex for their identification and is necessary study the microbial communities sampled directly from their natural environment using metagenomic approach. An important problem in metagenomics is measuring the diversity in a population (entropy) and the variation between subpopulations (beta-diversity) in uncertainty conditions. A good method that we propose can be use the Bayesian Shannon index and Shannon entropy ratio (SER) to estimate it, using a prior information based on a phylogenetic previously estimation. Bayesian methods improve the precision of parameter estimates, and uncertainty in parameter estimates can be easily propagated in calculations. The Bayesian diversity estimates were higher than their frequentist counterparts and had lower standard errors, so this approximation is present the diversity mixed index by means Markov Chain Monte Carlo (MCMC) simulation using JAGS with R.

Keywords: Entropy, Bayesian methods, biodiversity, probability, categorical data, metagenomics, microbiology.

1 Introduction

1.1 Metagenomics

The gut microbiota is home to more than 99% of the genetic information in humans [1] and there is important connection between the gut microbiome and metabolism, immune health, disease, autism, allergies, and obesity, it remains a largely unexplored area of science [2]. Microbial communities contain a unique complexity that makes difficult studying their diversity, for many questions the structure of the microbial community one only needs to know the relative order of diversity among samples rather than total diversity. Unfortunately, the culture of microorganisms is very complex for their identification and study, so they have had to develop new scientific methodologies for their study because of the large number of applications they have. One of these methodologies is metagenomics.

Metagenomics (also referred to as environmental and community genomics) is the study of genetic (genomic analysis of microorganisms) material recovered directly from environmental samples by direct extraction and cloning of DNA from an assemblage of microorganisms [3]. In any biological system information is ultimately linked to the DNA sequences present and microbial communities are no exception. In microbial communities we used 'word' frequency profiles of OTUs (operational taxonomic unit) as a proxy for the composition of the bacterial community at the genomic level, thus avoiding the need of defining bacterial species or taxonomical groups [4].

The broad field may also be referred to as environmental genomics, ecogenomics or community genomics. While traditional microbiology and microbial genome sequencing and genomics rely upon cultivated clonal cultures, early environmental gene sequencing cloned specific genes (often the 16S rRNA gene) to produce a profile of diversity in a natural sample [5].

The development of metagenomics stemmed from the ineluctable evidence that as-yet-uncultured microorganisms represent the vast majority of organisms in most environments on earth. This evidence was derived from analyses of 16S rRNA gene sequences amplified directly from the environment, an approach that avoided the bias imposed by culturing and led to the discovery of vast new lineages of microbial life [3].

An important problem in metagenomics is measuring the diversity in a population (entropy) and the variation between subpopulations (beta-diversity) in uncertainty conditions; therefore the aim of this paper is propose the use of new indices based on Bayesian methods, that can improve the precision of parameter estimates, and uncertainty in parameter estimates can be easily propagated in calculations. In order to focusing on the knowledge of these Bayesian diversity indexes, once proposed, several metagenomic populations and scenarios will be simulated and studied.

1.2 Metagenomic Diversity

Diversity measurement is important for understanding community structure and dynamics of the organisms, as is relatively solved by the ecologists, but has been particularly challenging for microbes. Microbiologists have recently rediscovered that ecologists and evolutionary biologists studying the diversity of microorganisms have developed a range of approaches to analyse the environmental diversity patterns and many of them can be applied to microorganisms. Basically the analysis of biodiversity that is carried out in metagenomics is fundamentally based on the one used in classical ecology like richness, abundance, alpha diversity and beta diversity often incorporating the concepts exposed as the phylogenetic relation and taking into account how the samples were obtained and the technical noise.

Most of the above-mentioned methods concentrate on comparing species richness, unfortunately not all of these statistics are applicable to taxonomic levels other than species, like OTUs (Operational taxonomic units) used in metagenomics. Although some OTU definitions try to capture species-like unit, one can ask valid questions about biodiversity at any level, as long as the OTU definition is clear and consistent [6].

In a classical ecology, we can divide the measure of the diversity in three dimensions (see Figure 1):

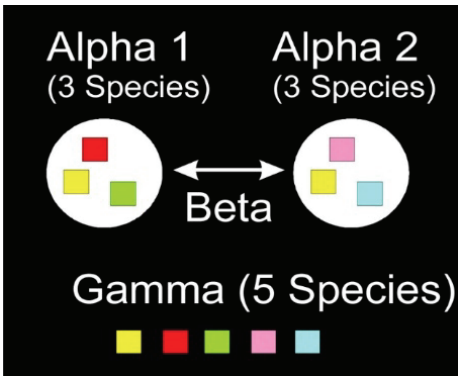

Figure 1. Schematic representation of the alpha, beta and gamma diversity (from http://www.geo.fu-berlin.de/en/v/fg736/subprojects/project-6/index.html)

The alpha diversity is denoted to the diversity within a community at a particular site. When we casually speak of diversity in an area, more often than not it refers to alpha diversity. This diversity is on a small scale, generally of the size of one ecosystem (e.g. for microbiome gut area, tooth, etc.):

Alpha diversity captures both the microbial richness of a sample and the evenness of the microbial abundance distribution. In metagenomics, alpha diversity is defined by the specific richness (S) and the structure of the community using the Shannon, Simpson or other indices.

We define the relative abundance, p_i, for the *ith* OTU as,

$$p_i = \lim_{n\to\infty}\left(\frac{n_i}{n}\right) \tag{1}$$

The Shannon index (H') is the most popular diversity index for the microbiologist when used metagenomics. Is also known as Shannon's diversity index, Shannon entropy, Shannon–Wiener index or

Shannon–Weaver index. The measure was originally proposed by Claude Shannon to quantify the entropy (uncertainty or information content) in strings of text [7]. *H'* quantifies the uncertainty (entropy or degree of surprise) associated with this prediction. It is most frequently calculated as follows:

$$H' = -\sum_{i=1}^{R} p_i \, log_2(p_i) \quad (2)$$

where p_i is the proportion of individuals belonging to the *ith* OTUs in the dataset of interest (sample or data set). Then the Shannon entropy quantifies the uncertainty in predicting the OTUs assigned to an individual that is taken at random from the dataset.

The beta diversity is referred to the diversity of species between two or more communities (or ecosystems), think that on the microbiome is very difficult talk about communities, in a sample is possible find several communities living together and changing genes between them. It is at a larger scale than alpha diversity, and looks to compare the diversity between two separate entities that are often divided by a clear biological/physiological barrier like (e.g.) the gut, skin or the vagina. Another important aspect of microbiome data is the heterogeneity in community structure across samples, as this can be an indication of the existence of "enterotypes". This can be addressed within the modelling framework by employing Bayesian nonparametric models that would allow to cluster selected associations across partitions of the samples [8].

Beta diversity refers to the measurement of the degree of difference in community memberships or structure between two samples of interest (in metagenomics the number of groups can be unknown). Beta diversity represents the similarity (or difference) in OTU composition between samples or also can be referred to the degree of difference between its genetical compositions. Again, many metrics such as Bray-Curtis, Sörensen quantitative index or Morisita-Horn measure in the case on non-phylogenetic analysis species-based or UniFrac, FST or DPCoA in the case of phylogenetic analysis are commonly employed to measure it.

The Gamma diversity is referred to a very large scale, where diversity is compared between many ecosystems (digestive system), at the level of a biome. It could range over areas like the entire gut, the entire mouth zone, etc. or compare human with others.

An old ecological problem when exploring a given environment is how many species are not observed. If we look the biological context of metagenomics, we realize that m*any* bacterial species cannot be grown artificially out of their natural environment and sets of species (OTUs) can only be studied all together, within their environment (e.g. human gut) using metagenomics via NGS by sampling and sequencing DNA (or RNA) from all species. In this context we can talk about species abundant distribution (SAD) where the observed counts X_i are truncated, meaning that OTUs with 0's are not observed.

Some classical distributions for SAD are: Poisson, Log-normal [9], Poisson-Gamma [10, 11], Poisson counts with Gamma intensities, Log-series, Geometric series and Mixture of discrete distributions [12, 13].

Unfortunately, none of the classical species abundance distribution models (log-series, geometric-series, log-normal) are useful with the complex microbial communities, especially in varying states of disturbance (unperturbed to perturbed) or impoverishment (species-rich to poor) [14].

Another question, still more complicated mathematically, is study the probability distribution that fits to the metagenomic matrix (species, abundance, samples and communities, see Table 1), not only to a simple sample (species vs abundance), where it might be interesting to study it's SAD. One of the possible causes that explains this difficulty is because microbial communities of this type are not homogeneous and are made up of several subpopulations or communities that establish particular genetic relationships for each individual in space and in the time.

From the statistical point of view, it is a problem and must be kept in mind in the development of statistical methods that encompass this metagenomic problem.

1.3 Probability Distribution of Abundance

In this section we briefly review statistical models that have been used to model the distribution of abundance for a metagenomics matrix.

1.3.1 The multinomial model with Dirichlet prior distribution

In the table 1 it is possible to see the general Metagenomic matrix ($\boldsymbol{M}$) structure (n rows: samples, p

columns: taxa or OTU (operational taxonomic units[1])). Usually, for convenience, we change the notation of p by k; also during the statistical analysis we use the transpose $\boldsymbol{M'}$, which shows the samples (e.g. patients) in the columns and the organism identified (OTU: operational taxonomic unit) in the rows (Table 1). $\boldsymbol{M'}$ has the dimension:

$$Dim(\boldsymbol{M'}) = kn \tag{3}$$

where k is the number of OTUs and n the number of samples.

As a result of metagenomic analysis, $\boldsymbol{M'}$ can be very large and usually has thousands of OTUs, most of them with small frequencies or 0, i.e. $\boldsymbol{M'}$ is typically a sparse matrix. This matrix is truncated, in the sense that there are species that have not been observed in the sampling.

Table 1. matrix structure of $\boldsymbol{M'}$ (metagenomics matrix input).

num	Taxon (OTU)	Sample 1	Sample 2	Sample *jth*	Sample n	
1	*otu*.1	m_{11}	m_{12}	…	m_{1n}	$N_{1.}$
2	*otu*.2	m_{21}	m_{22}	…	m_{2n}	$N_{2.}$
⋮	⋮	…	…	m_{ij}	…	…
k	*otu*.k	m_{k1}	m_{k2}	…	m_{kn}	$N_{k.}$
		$N_{.1}$	$N_{.2}$	…	$N_{.n}$	N

From the statistical point of view It is very convenient to formalize the probability distribution underlying this matrix structure, so each sample from $\boldsymbol{M'}$ can be represented by one k-dimensional random vector X_j; X_j=(m_{1j}, m_{2j}, …, m_{kj}), where m_{kj} represents the number of times that taxa k is observed in sample j.

The probability distribution of each random vector $X_{i.}$ (vector row) and $X_{.j}$ (vector column) can be associate individually to a multinomial distribution,

$$X_{.j} \sim MN\left(N_{.j}, \theta_{1j}, \ldots, \theta_{kj}\right);\ \forall j = 1, .., n \tag{4}$$

$$X_{i.} \sim MN(i., \theta_{i1}, \ldots, \theta_{in});\ \forall i = 1, .., k \tag{5}$$

The multinomial distribution is a multivariate generalization of the binomial distribution, where the marginal distribution of each X_{ij} is:

$$X_{ij} \sim Bin\left(m_{ij}, \theta_{ij}\right); 1 \le \theta_{ij} \le 1;\ \forall j = 1, .., n;\ \forall i = 1, .., k \tag{6}$$

e.g. if we consider the partition of all sample space Ω^j the j-sample space in k parts:

$$A_{1j} \quad A_{2j} \quad \ldots \quad A_{kj}$$

One individual selected randomly has the probability θ_{kj} of belongs to the taxon A_{kj} in the partition:

$$\left.\begin{array}{l} P\left(A_{1j}\right) = \theta_{1j} \\ P\left(A_{2j}\right) = \theta_{2j} \\ \vdots \\ P\left(A_{kj}\right) = \theta_{kj} \end{array}\right\} \sum_{i=1}^{k} \theta_{ij} = 1\ ;\ \forall j = 1, .., n \tag{7}$$

If we wish calculate for a sample j the probability of have $N_{.j}$ individuals, m_{1j} belonging to class A_{1j}, m_{2j} belongs to class A_{2j}, …, m_{kj} belongs to class A_{kj}, with the restriction

$$\sum_{i=1}^{k} m_{ij} = N_{.j};\ \forall j = 1, .., n \tag{8}$$

And using the multinomial function of density (mass function) we can calculate this probability, $MN(N_{.j}$; $\theta_j = (\theta_{1j}, \theta_{2j}, \ldots, \theta_{kj}))$:

$$P\left[\left(A_{1j} = m_{1j}\right) \cap \ldots \cap \left(A_{hj} = n_{kj}\right)\right] = \frac{N_{.j}!}{m_{1j}! m_{2j}! \ldots m_{kj}!} \theta_{1j}^{m_{1j}} \cdot \theta_{2j}^{m_{2j}} \cdot \ldots \cdot \theta_{kj}^{m_{kj}};\ \forall j \tag{9}$$

where $0 \le \theta_{kj} \le 1$ for all i in 1 to k, and $\theta_{1j} + \ldots + \theta_{kj} = 1(\forall j)$, and if k=1 the mass function reduces to the binomial, $\forall j = 1, \ldots, n$.

The conjugate prior of the Multinomial distribution is the Dirichlet distribution, the multivariate generalization of beta distribution. Hence the parameter vector $\theta_k = (\theta_{1j}, \theta_{2j}, \ldots, \theta_{kj})$; $\forall j$ has a prior distribution given by:

[1] Operational definition used to classify groups of closely related individuals. OUT is used as a pragmatic definition to group individuals by similarity, equivalent to but not necessarily in line with classical Linnaean taxonomy or modern evolutionary taxonomy. Extracted from Wikipedia "Operational taxonomic unit".

$$\theta_k \sim Dirichlet(\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{kj}); \forall j = 1, .., n \quad (10)$$

In (10) the density function is given by:

$$g(\theta|\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{kj}) = \frac{\Gamma(\sum_i^k \alpha_{ij})}{\prod_i^k (\Gamma\alpha_{ij})} \theta_{1j}^{(\alpha_{1j}-1)} \theta_{2j}^{(\alpha_{2j}-1)} \ldots \theta_{kj}^{(\alpha_{kj}-1)};$$

$$\alpha_{ij} > 0; 0 \leq \theta_{ij} \leq 1; \ \sum_i^k \theta_{ij} = 1; \ \forall j = 1, .., n \quad (11)$$

In bayesian inference, remember that $p(\theta|x)$ is known as posterior distribution and is proportional to likelihood ($p(x|\theta)$) x prior distribution ($p(x)$), so $p(\theta|x) \propto p(x|\theta) \cdot p(x)$.

The posterior distribution of θ_j given X is:

$$\theta_j|x \sim Dirichlet(x_{1j} + \alpha_{1j}, x_{2j} + \alpha_{2j}, \ldots, x_{kj} + \alpha_{kj}); \ \forall j = 1, .., n \quad (12)$$

1.3.2 The multinomial-Dirichlet distribution

Another model of probability that is considered as potentially explain ***M'*** is the Dirichlet-Multinomial distribution (DM) where reflects an over dispersed multinomial distribution. The Dirichlet-multinomial distribution is a family of discrete multivariate probability distributions on a finite support of non-negative integers, it is also called the Dirichlet compound multinomial distribution (DCM) or multivariate Pólya distribution (MPD).

DM is a compound probability distribution, where a probability vector $\boldsymbol{p}$ is drawn from a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$, and an observation drawn from a $MN(N, \theta_1, \ldots, \theta_k)$ with probability vector $\boldsymbol{p}$ and number of trials N. It is frequently encountered in Bayesian statistics, empirical Bayes methods and classical statistics as an over dispersed multinomial distribution.

DM approximates the multinomial distribution arbitrarily for a well for large $\boldsymbol{\alpha}$. DM is a multivariate extension of the Beta-binomial distribution, as the multinomial and Dirichlet distributions are multivariate versions of the binomial distribution and beta distributions, respectively.

If we consider the distribution DM of the abundance of a sample j,

$$X_{.j} \sim DM(N_{.j}, \alpha_{1j}, \ldots, \alpha_{kj}); \ \forall j = 1, .., n \quad (13)$$

The parameters are $N_{.j} > 0$; $\alpha_{1j}, \ldots, \alpha_{kj} > 0$; $\forall j = 1, \ldots, n$. DM is the marginal distribution of $X_{.j}$ after integrating θ.

In (13) the density function (pmf) is given by:

$$g(X_j|\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{kj}) = \frac{(N_{.j}!)\Gamma(\sum_i^k \alpha_{ij})}{\Gamma(n+\sum_i^k \alpha_{ij})} \prod_i^k \left(\frac{\Gamma(\sum_i^k x_{ij}+\alpha_{ij})}{(x_{ij}!)\Gamma(\alpha_{ij})}\right); \forall j = 1, .., n \quad (14)$$

DM distribution can also be motivated via an urn model for positive integer values of the vector $\boldsymbol{\alpha}$, known as the "Polya urn" model.

Very recently Wadsworth et al. [8] proposed an integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data, so this fact justifies also the use of DM as a good model to simulate metagenomic matrices.

1.3.3 The multinomial Dirichlet mixture model

As with SAD, several authors (Holmes et al, 2012) nowadays believe that because of the uncertainty of the data and their special characteristics, the most correct way to model ***M'*** is through mixtures of multivariate probability distributions.

Given a set of probability functions $p_1(X)$, ..., $p_n(X)$, or the corresponding cumulative distribution functions $P_1(X)$, ..., $P_m(X)$ and weights w_1, ..., w_m, where $w_i \geq 0$; $\sum_{i=1}^m w_i = 1$, a mixture distribution can be represented as a convex combination of m multinomial (MN) and/or Diritchlet-Multinomial mixtures (DMM):

$$F(x) = \sum_{i=1}^m w_i P_i(x) \quad (15)$$

$$f(x) = \sum_{i=1}^m w_i p_i(x) \quad (16)$$

where $p_i(x) = \frac{N_{.j}!}{m_{1j}!m_{2j}!\ldots m_{kj}!} \theta_{1j}^{m_{1j}} \cdot \theta_{2j}^{m_{2j}} \cdot \ldots \cdot \theta_{kj}^{m_{kj}}$; or $p_i(x) = \frac{(N_{.j}!)\Gamma(\sum_i^k \alpha_{ij})}{\Gamma(n+\sum_i^k \alpha_{ij})} \prod_i^k \left(\frac{\Gamma(\sum_i^k x_{ij}+\alpha_{ij})}{(x_{ij}!)\Gamma(\alpha_{ij})}\right)$ for the MN or DM pmf.

Holmes et al. [16] propose at the paper "Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics" the Dirichlet multinomial mixtures (DMM) as a generative models for microbial metagenomics. DMM suppose that samples have different size, and the matrix $\boldsymbol{M}$ is sparse, as communities are diverse and skewed to rare taxa. Most methods used previously to classify or cluster samples have ignored these features. Holmes et al. [16] describe each community by a vector of taxa probabilities $\boldsymbol{\alpha}$. These vectors are generated from one of a finite number of Dirichlet mixture

components each with different hyperparameters, using a model named N-mixture model to accommodate these multinomial sampling to a Bayesian perspective for each sample of ***M'***.

Finally Royle proposes [13], at his work "N-mixture models for estimating population size from spatially replicated counts", a more complex model to try to explain the abundance and richness in the case of replications and where the uncertainty is high where the number of species is unknown, this model was used to estimate the sample size in the classical ecology models. In this sense, this model of N-mixtures must continue to be explored, since it can be very promising both to find the calculation of the sample size in the case of metagenomics, as well as to study the new parameters of diversity that are proposed, especially in the case of much uncertainty, number of unknown species, etc. as the case of metagenomics that concerns us.

The key idea for N-mixtures is to view site-specific population sizes, N, as independent random variables distributed according to some mixing distribution (e.g., Poisson). Prior parameters are estimated from the marginal likelihood of the data, having integrated over the prior distribution for N.

Some explanations on this work are based on the Master Thesis (Master in Biostatistics and Bioinformatics UOC-UB) from Paloma Pizarro during the course 2015-2016 "Bacterial Metagenomics: Associated Probability Distributions and Profile Analysis" [16]. During this thesis different test of goodness of fit was applied to a different ***M'*** real matrices in order to test the fit to a multinomial distribution, Dirichlet and DM.

A function to compute the $C(\alpha)$-optimal test statistics of Kim and Margolin [17] for evaluating the Goodness-of-Fit of a Multinomial distribution (null hypothesis) versus a Dirichlet-Multinomial distribution (alternative hypothesis) was used to do the goodness of fit tests, this function can be found in the library HMP for R. The $C(\propto)$-optimal test- statistics is given by,

$$T = \sum_{j=1}^{K} \sum_{i=1}^{P} \frac{1}{\sum_{i=1}^{P} x_{ij}} \left(x_{ij} - \frac{N_i \sum_{i=1}^{P} x_{ij}}{N_g} \right)^2 \tag{17}$$

where K is the number of taxa or OTU, P is the number of samples, x_{ij} is the taxon or OTU j, $j = 1, \ldots, K$ from sample i, $i = 1, \ldots, P$, N_i is the number of reads (abundance) in sample i, and N_g is the total number of reads across samples. As the number of reads (abundance) increases, the distribution of the T statistic converges to a Chi-square with degrees of freedom equal to $(P–1)(K–1)$, when the number of sequence reads is the same in all samples. When the number of reads is not the same in all samples, the distribution becomes a weighted Chi-square with a modified degree of freedom (see Kim and Margolin, 1992, for more details and documentation about the use of the library HMP for R). Each taxa in ***M'*** should be present in at least 1 sample, a column with all 0's may result in errors and/or invalid results.

For the metagenomic data tested at Pizarro (2016) [16], T= 37.75 and p.value = 1, it was no possible demonstrate that ***M'*** was different to a multinomial distribution. An R script was developed in order to do this test to a future simulated matrices.

All of this models to explain ***M'***, above-mentioned (MN, DM-MN, N-mixtures) can be used to incorporate prior information (phylogenetic information, abundance model) and the uncertainty that arises in metagenomics (unknown number of species, number of subpopulations, etc) in order to obtain a posterior distribution predictive distribution and use it to estimate diversity and make inference, so we think that is necessary use the bayesian approach by MCMC method using BUGS and R.

2 Methods

2.1 MCMC Simulation

By using a prior information, Bayesian methods improve the precision of parameter estimates like abundance, and uncertainty in parameter estimates can be easily propagated in calculations.

Niane et al. [18] showed that the Bayesian diversity estimates were higher than their frequentist counterparts and had lower standard errors. The Bayesian estimates resulted in lower p-values than the frequentist approach for two cases reflecting in Bayesian method's higher power. Therefore, a Bayesian approach is recommended as it has a wider framework for inference on diversity studies.

Many situations arise in Bayesian Statistics in which the posterior distribution is difficult to either calculate or simulate from. This owes much to the necessity to calculate integrals of complex and commonly multi-dimensional expressions. Even if a posterior distribution can be found up to a constant

of proportionality, the presence of a large number of parameters in a Bayesian framework means that most methods of simulating random variables from this distribution break down. For many years such problems were either too difficult or too computationally intensive to be tackled. The discovery of Markov Chain Monte Carlo (MCMC) methods marked a new approach to these problems. The ability of MCMC to provide insight into large, complex Bayesian problems has been one of the most important developments in modern statistics and without which Bayesian Statistics would not be so popular today.

In the present work we choose the BUGS language (Bayesian inference Using Gibbs Sampling) [19] to simulate the posterior distribution of DM-MN used in "Bayesian Methods for Ecology" [20] in order to calculate *H'* and the other ecological parameters proposed later. Computations were performed using JAGS [21], an open source implementation of BUGS in C++, within R [22] by using the package rjags.

2.2 A Proposal of Diversity Indices Based on Simulation for Alpha and Beta Diversity

One of the problems with the ecological interpretation of metagenomic analyses is that the number of species is unknowns and other is the coexistence of diverse communities (k) with different composition of OTUs [23, 24]. It is in this situation that previous biodiversity measures cannot be applied.

A first proposal of the model and the simulation used is calculate *H'* using the Bayesian model and the implementation of (9). The advantage of this method is that a 95% credibility interval is obtained, which can be used to compare the diversities with other samples, as well as it is easy to calculate the alpha diversity of the different subpopulations to perform inference with them.

A simple form to measure the beta diversity between subpopulations (k) is use a Shannon entropy coefficient of variation between subpopulations ratio (SCV) between the different subpopulations that compose the sample. SCV is defined as:

$$SCV = 100\frac{sd(H')}{\overline{H'}} \tag{18}$$

where H'_i is the Shannon entropy for each *ith* subpopulation, so $H' = \{H'_1, \ldots, H'_k\}$, $\overline{H'} = \left(\left(\sum_{i=1}^{k} H'_i\right)/k\right)$ is the average Shannon entropy ($\overline{H'}$) for the k subsets or subsamples and $sd(H')$ is the standard deviation (sd) of *H'*. $SCV = 0$ when $sd(H')=0$. SCV is the coefficient of variation of Shannon entropy between subpopulations (For example phylogenetically determined). SCV can be a good form to estimate the variability of compositions between subpopulations, so the Bayesian credibility interval of *SER* includes the 0, all the *H'* are equal, homogeneous diversity. If SCV is larger and the credibility interval non-include the 0, *H'* are non-equal, heterogeneous diversity.

At this point is easy to establish the following a hypothesis test to compare alpha diversities (using the *H'* estimated and the credibility interval) between subpopulations and known the number of subpopulations that are diverse (ξ), if we have only two subpopulations:

$$\begin{cases} H_o : \theta_1 / \theta_2 = 1 \\ H_1 : \theta_1 / \theta_2 \neq 1 \end{cases} \tag{19}$$

where θ_1 is the Shannon entropy *H'* at all the first subpopulation, and θ_2 is the Shannon entropy *H'* for the second subpopulation. The hypothesis test of (17) can be reformulated as:

$$\begin{cases} H_o : \theta_1 = \theta_2 \\ H_1 : \theta_1 \neq \theta_2 \end{cases} \tag{20}$$

The choice of a null hypothesis in this case is important. Ideally, the null hypothesis should be such that is the rejection will have important logical consequences that lead to better important consequences (the alpha biodiversity is the same between this subpopulation an d all sample) , however, the genetists use nil nulls (predicting no difference) that are very unlikely to be correct [20]. Unfortunately is very difficult construct null hypothesis with a non-zero effect, however a priori information is known before start the test.

Classical ecological methods suggested use a non-parametric methods based on confidence intervals to solve this question and the use of re-sampling methods (e.g. jack-knife) to estimate the confidence interval between $\theta_1=\theta_2$, which is similar to the 95% credible interval used at "Bayesian Methods for

Ecology" [20]. So we propose calculate 95% credible Bayesian interval to compare θ_1 and θ_2 and decide between H_0 or H_1, to do it we propose the Shannon entropy ratio (SER), so we can define:

$$SER = \frac{H_1}{H_2} \tag{21}$$

where H'_1 is the Shannon entropy of the first subpopulation and H'_2 Shannon entropy of the second subpopulation. To decide between $\begin{cases} H_o: \theta_1/\theta_2 = 1 \\ H_1: \theta_1/\theta_2 \neq 1 \end{cases}$ we would use the Bayesian confidence (credible) interval of SER:

$$P((\theta_1/\theta_2)_L \leq (\theta_1/\theta_2) \leq (\theta_1/\theta_2)_U) = 1 - \alpha \tag{22}$$

where $SER = \theta_1/\theta_2$

Finally, we can use SER to know the number of subpopulations that are really different on the alpha diversity point of view (ξ) using (19), as the number of pairs where $\theta_1/\theta_2 \neq 1$, when H_1 was accepted.

3 Results

A BUGS script very similar as the above-mentioned proposed by McCarthy [20] was used to do a simulation of a DM distribution, as a multinomial distribution with a distribution prior Diritchlet, adding the calculation of the bayesian H', SER and SCV to simulate different scenarios. After run the BUGS scrip in R (using the library JAGS) and verify its function, 16 scenarios were generated and used to the calculations.

To simulate the DM probability distributions on the table 2 as a multinomial distribution with a priori Diritchlet parameters was used the library Learnbayes for R:

```
library(LearnBayes)
nsites<-100 #num OTUs
ab.total<-1000 #abundance
ppp <- rdirichlet(1, par = rep(1, nsites))
ppp
nsimulac<- 1
matriu <- array(0, dim=c(nsites, nsimulac))
for(i in 1:nsimulac){
    X <- as.vector(rmultinom(1, size =ab.total , prob = ppp))# DM-MN
    matriu[,i]<-X
    N <- sum(X)
}
matriu #vector of abundance-richness
N #total abundance
```

To simulate the probability distributions on the table 3, more complex than in the Table 2, we used a N-mixture model of Poisson-Binomial distributions proposed by Royle [13]. A BUGS scrip proposed in https://groups.nceas.ucsb.edu/non-linear-modeling/projects/nmix/WRITEUP/nmix.pdf was used to generate the probability distribution of abundance-richness used on the different scenarios of the Table 3.

The different scenarios were:

-presence of a subpopulation (replications): yes/not
-number of replications: 1-10
-number of OTUs: 10, 100, 1000.
-total abundance: 10-10000000

For each the 16 scenarios the Shannon entropy (H') was estimated using the classical formula and the Bayesian H', SCV and SER.

John Kruschke in "Doing Bayesian Data Analysis" [25] recommends that for parameters of interest, MCMC chains should be run until their effective sample size is at least 10,000.

To estimate classical H' Shannon we have used the package vegetarian in R, which uses a bootstrap method for standard error estimation. Herewith is the instruction used in R:

```
>H(t(as.data.frame(matriu)),lev="alpha",q=1,boot=T, boot.arg=list(num.iter=1000))
```


3.1 Bayesian Estimation of H'

Table 2. Results of MCMC simulations of different scenarios in order to calculate H' (classical/Bayesian) and to compare it (se = standart error, BCI=95% Bayesian credible interval)

Number of OTUs	Total abundance	Number of subpopulations	Classical H' (estimation and standart error)	Bayesian H' (mean and BCI)
100	10	1	2.302585 (se=0.181736)	4.10360901 BCI=(3.99509184- 4.1387734)
100	100	1	3.664233 (se=0.07023475)	4.08932381 BCI=(3.99451192-4.12187228)
100	1,000	1	4.109216 (se=0.02510306)	4.0835262 BCI=(4.0394458-4.1268858)
100	100,000	1	4.11355 (se=0.002699161)	4.027884193 BCI=(4.022506713-4.033219173)
500	10	1	2.302585 (se=0.1817862)	5.44736823 BCI=(5.40422477-5.46183720)
500	100	1	4.303583 (se=0.06447291)	5.4444885 BCI=(5.4016684-5.4846489)
500	1,000	1	5.54568 (se=0.02484851)	5.451367 BCI=(5.419685-5.462166)
500	100,000	1	5.776844 (se=0.002546979)	5.43471455 BCI=(5.43033198-5.43908440)
1000	10	1	2.302585 (se=0.1752076)	5.90820106 BCI=(5.88124265-5.90852462)
1000	100	1	4.433582 (se=0.0664647)	5.9065353 BCI=(5.8797588-5.9317666)
1000	1,000	1	5.982431 (se=0.02604336)	5.89943765 BCI=(5.87467233-5.92261028)
1000	10,000	1	6.413494 (se=0.008347227)	5.896549905 BCI=(5.885235051-5.900356613)
1000	100,000	1	6.488955 (se=0.002437707)	5.916427429 BCI=(5.912781727-5.920037143)
1000	10^6	1	6.488525 (se=0.0002512142)	5.9105585711 BCI=(5.9101794419-5.9109271541)

The Table 2 shows the Bayesian calculations using MCMC method of H'. This was obtained of different simulations of abundance-richness population matrices using the DM distribution of an abundance-richness sample on different scenarios (abundance, richness=number of OTUs), simulating only one multinomial sample. Here is possible observe that Bayesian H' always better estimates the "supposedly true value of H'", for all scenarios used.

As a clarification, to obtain the true value of H' (classical point of view, without supposing a priori distribution), the abundance should be infinite (population perspective) and it has been assumed that for the simulations, the entropy H' is based on the same proportion of abundance for each site, thus the supposed true entropy in the case of equiprobability of the OTU would be:

$$\text{supposedly true value of } H' = -\sum_{i=1}^{R}\left(\frac{nsites}{total\ abundance}\right) log_2(nsites/total\ abundance) \quad (23)$$

In practice and for the amount of OTU used, H' it can be considered true for an abundance of $n >$ 10000.

Bayesian H' is based on the probabilities of observed and unobserved OTUs, since the a priori probability is not 0 for these OTUs. On the other hand, H' (estimation of Entropy using classical Shannon) always gives different values depending on the presence or not of the samples (mean and BCI).

3.2 Bayesian Estimation of SER and SCV

Table 3. Results of MCMC simulations of different scenarios in order to calculate *H'*, H_1, H_2, *SER* and *SCV* (BCI=95% Bayesian credible interval)

Number of OTUs	Total abundance	Number of subpopulations	Replications	Bayesian *H'* (Mean and BCI=95% Bayesian credible interval)			
					mean	**2.5%**	**97.5%**
20 (n_1=10, n_2=10)	17	2 (λ_1=1, λ_2=1)	1	*H'*	2.682	2.508	2.82
				H_1	1.26	0.955	1.555
				H_2	1.422	1.127	1.692
				SER	0.906	0.575	1.357
				SCV	14.157	0.575	38.442
					mean	**2.5%**	**97.5%**
200 (n_1=100, n_2=100)	437	2 (λ_1=3, λ_2=3)	1	*H'*	4.872	4.832	4.909
				H_1	2.413	2.253	2.571
				H_2	2.459	2.3	2.617
				SER	0.984	0.861	1.116
				SCV	3.839	0.139	10.834
					mean	**2.5%**	**97.5%**
200 (n_1=100, n_2=100)	4232	2 (λ_1=3, λ_2=3)	10	*H'*	4.922	4.907	4.936
				H_1	2.463	2.402	2.524
				H_2	2.459	2.397	2.521
				SER	1.002	0.953	1.053
				SCV	1.418	0.052	4.024
					mean	**2.5%**	**97.5%**
200 (n_1=100, n_2=100)	5239	2 (λ_1=5, λ_2=3)	10	*H'*	4.945	4.932	4.958
				H_1	2.889	2.837	2.94
				H_2	2.0567	2	2.114
				SER	1.4049	1.343	1.470
				SCV	23.788	20.693	26.914
					mean	**2.5%**	**97.5%**
200 (n_1=100, n_2=100)	550	2 (λ_1=5, λ_2=3)	1	*H'*	4.916	4.882	4.947
				H_1	2.658	2.512	2.801
				H_2	2.258	2.107	2.408
				SER	1.18	1.044	1.328
				SCV	11.534	3.06	19.944
					mean	**2.5%**	**97.5%**
1000 (n_1=500, n_2=500)	698	2 (λ_1=1, λ_2=1)	1	*H'*	5.946	5.925	5.966
				H_1	2.981	2.852	3.111
				H_2	2.965	2.835	3.094
				SER	1.006	0.922	1.097
				SCV	2.51	0.093	7.072

The Table 3 shows the bayesian calculations using MCMC method of *H'*, H_1 and H_2 (H_1, H_2 are the *H'* of subpopulation 1 and 2, respectively) obtained of different simulations of abundance-richness population matrices using the N-mixture (N-Poisson-binomial) distribution proposed by Royle [13] of an abundance-richness sample on different scenarios (abundance-richness=number of OTUs), simulating only one multinomial sample (merging two independent OTU table in one final table). Here is possible observe that $SER \neq 1$ (the Bayesian credible interval doesn't contain the 1 in his interval. ex: 1.404875[1.342809;1.470086]) in all the cases where $\lambda_1 \neq \lambda_2$, without regard the number of OTUs. abundance or parameter value (λ_1, λ_2). *SCV* varies according to abundance and difference between lambda parameter values. It is interesting to see that *SCV* can provide information of the beta diversity between populations. So we think it can be a good indicator of this type of diversity since it is

standardized to a range of variation between 0 (same entropy) and 100% maximum difference of entropy between subpopulations.

The functions and calculations discussed in this work can be found with examples in the function *Shannon.Bayesian.Entropy (X=matrix, nsites, N=total abundance, N1=nsites 1st population, N2=nsites 2nd population)* of the library BDSbiost3 for R; link https://github.com/ amonleong/BDSbiost3 [23, 24].

4 Conclusion

One of the problems with the ecological interpretation of metagenomic analyses is that the number of species is unknown and other is the coexistence of diverse communities (k) with different composition of OTUs. To solve this we propose adapt the classical ecological framework frequentist to use Bayesian models, like the Dirichlet compound multinomial distribution (DM). The Bayesian diversity measures proposed can estimate better the simulated populations better than their frequentist counterparts and had lower standard errors. By using a prior information. Bayesian methods improve the precision of parameter estimates like abundance and uncertainty in parameter estimates can be easily propagated in calculations.

We have adapted the bayesian models of probability from McCarthy [20] and the N-mixtures of Royle [13] to its use in the metagenomic field where the uncertainty of the species (OTU) is high and with an unknown number of subpopulations. Also this models were used to simulate the different scenarios proposed. We have simulated 16 different scenarios in order to test the utility and validate the diversity proposed measures and it has been possible to observe in all of them the absence of errors. so it is proposed to use them with real metagenomic abundance-rich matrices.

This context convert the Bayesian framework in a really useful tool to calculate the diversity of the samples and determine if the diversity of these subpopulations are really different and its variability, using a new proposed measures in this work based on the classical Shannon coefficient of entropy *H'* but now calculated using MCMC. Another Bayesian measures proposed and studied in this work are Shannon entropy ratio (*SER*), a measure to establish differences between the entropies of the subpopulations studied on a metagenomic matrix that can help to known the number of subpopulations that are really different on the alpha diversity point of view (ξ).

Another advantage of using Bayesian methods is that hierarchical models like N-mixtures model of abundance (some parameters of the model are treated as a random variable) are easy to calculate to equivalent non- hierarchical models.

The only drawback of this Bayesian computing environment is the consumption of computational time, so for each simulation many minutes of computation are necessary.

Acknowledgments. My thanks to Paloma Pizarro-Tobias for her magnificent dissertation work and her calculations with metagenomic probability distributions.

References

1. C. M. Guinane and P. D. Cotter. “Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ”, Therapeutic advances in gastroenterology, vol. 6, no. 4, pp. 295–308, 2013.
2. M. Pollan. “Some of My Best Friends Are Germs”. 2013. Available: http://www.nytimes.com/2013/05/19/magazine/say-hello-to-the-100-trillion-bacteria-that-make-up-your-microbiome.html?_r=1
3. J. Handelsman. “Metagenomics: Application of Genomics to Uncultured Microorganisms”. Microbiology and Molecular Biology Review, vol. 68, no. 4, pp: 669–685, 2004.
4. CI. Rodríguez and T. Monleón-Getino. “A new R library for discriminating groups based on abundance profile and biodiversity in microbiome metagenomic matrices”. International Journal of Scientific and Engineering Research, vol. 7, no. 10, pp: 243-253, 2016.
5. D. Marco (editor). “Metagenomics: Current Innovations and Future Trends”. Caister Academic Press, 2011.
6. B. J. M. Bohannan and J. Hughesy. “New approaches to analyzing microbial biodiversity data” Current Opinion in Microbiology, vol. 6, pp: 282–287. 2003. Available: https://pages.uoregon.edu/bohannanlab/pubs/Bohannan%20and%20Hughes03%20copy.pdf

7. C. E. Shannon. "A Mathematical Theory of Communication". Bell System Technical Journal (PDF), vol. 27, no. 3, pp: 379–423, 1948.
8. WD. Wadsworth, R. Argiento, M. Guindani, J. Galloway-Pena, SA. Shelbourne and M. Vannucci. "An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data". BMC Bioinformatics, vol, 18, no, 94. 2017.
9. J. R. Doroghazi and D. H. Buckley. "Evidence from GC-TRFLP that Bacterial Communities in Soil Are Lognormally Distributed". PLoS One, vol. 3 no. 8, pp. e2910, 2008.
10. R. A. Fisher, Corbet A. S., and C. B. 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, vol. 12, pp. 42–58, 1943.
11. D. J. Golichier, R. B. O'Hara, L. Ruíz-M, and L. Cayuela. "Lifting a veil on diversity: a Bayesian approach to fitting relative-abundance models". Ecological Applications, vol. 16, no. 1, pp. 202–212, 2006
12. S. D. Hooper, D. Dalevi, A. Pati, K. Mavromatis, N. N. Ivanova and N. C. Kyrpides. "Estimating DNA coverage and abundance in metagenomes using a Gamma approximation". Bioinformatics, vol. 26, pp. 295–301, 2010.
13. JA. Royle "N-mixture models for estimating population size from spatially replicated counts". Biometrics, vol. 60, no. 1: pp. 108-115, 2004.
14. M.S. Lindner and B. Y. Renard. "Metagenomic abundance estimation and diagnostic testing on species level". Nucleic Acids research, vol. 41 n. 1, pp. e10, 2012.
15. I. Holmes, K. Harris and C. Quince. "Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics". PLoS ONE, vol. 7 no. 2, 2012.
16. P. Pizarro. "Bacterial Metagenomics: Associated Probability Distributions and Profile Analysis". Master thesis of the master in Biostatistics and Bioinformatics (UOC-OPC, Barcelona, Spain). Adviced by Toni Monleón Getino. 2016
17. B. S. Kim and B. H. Margolin. "Testing Goodness of Fit of a Multinomial Model Against Overdispersed Alternatives". Biometrics, vol, 48, pp. 711-719, 1992.
18. A.A. Niane, M. Singh and P. C. Strulk. "Bayesian estimation of shrubs diversity in rangelands under two management systems in northern Syria". Open Journal of Ecology, vol. 4, pp. 163-173, 2004
19. D. Lunn, D. Spiegelhalter, A. Thomas and N. Best. "The BUGS project: Evolution, critique and future directions". Statistics in Medicine, vol. 28, pp. 3049-67, 2009.
20. M. A. McCarthy. Bayesian Methods for Ecology". Cambridge University Press, 2007.
21. M. Plummer "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling". Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March 20–22. Vienna, Austria, 2003.
22. R Core Team. "R: A language and environment for statistical computing". R Foundation for Statistical Computing. Vienna. Austria. 2016. Available: http://www.R-project.org/.
23. CI. Rodríguez, T. Monleón-Getino, M. Cubedo, M. Ríos-Alcolea. "A priori groups based on Bhattacharyya distance and partitioning around medoids (PAM) with applications to metagenomics". IOSR Journal of Mathematics, vol. 13, no. 3, pp. 24-32, 2017.
24. A. Monleon-Getino, CI. Rodríguez-Casado and J. Méndez-Viera. "Sample size in metagenomics. a bayesian approach using BDSbiost3 for R". XVI Spanish Biometric Conference. CEB, Sevilla, Spain, 2017. Library for R BDSbiost3, avalaible at: at: https://github.com/amonleong/BDSbiost3
25. J. K. Kruschke. "Doing Bayesian Data Analysis A Tutorial with R. JAGS. and Stan". Academic Press / Elsevier, 2015.

