Briefings in Bioinformatics, 2016, 1–14

doi: 10.1093/bib/bbw119
Paper

Measuring the diversity of the human microbiota with targeted next-generation sequencing

Francesca Finotello, Eleonora Mastrorilli and Barbara Di Camillo

Corresponding authors: Francesca Finotello, Biocenter, Division of Bioinformatics, Medical University of Innsbruck, Innsbruck, Austria; Department of Information Engineering, University of Padova, Padova, Italy. E-mail: francesca.finotello@i-med.ac.at; Barbara Di Camillo, Department of Information Engineering, University of Padova, Padova, Italy. E-mail: dicamill@dei.unipd.it

Abstract

The human microbiota is a complex ecological community of commensal, symbiotic and pathogenic microorganisms harboured by the human body. Next-generation sequencing (NGS) technologies, in particular targeted amplicon sequencing of the 16S ribosomal RNA gene (16S-seq), are enabling the identification and quantification of human-resident microorganisms at unprecedented resolution, providing novel insights into the role of the microbiota in health and disease. Once microbial abundances are quantified through NGS data analysis, diversity indices provide valuable mathematical tools to describe the ecological complexity of a single sample or to detect species differences between samples. However, diversity is not a determined physical quantity for which a consensus definition and unit of measure have been established, and several diversity indices are currently available. Furthermore, they were originally developed for macroecology and their robustness to the possible bias introduced by sequencing has not been characterized so far. To assist the reader with the selection and interpretation of diversity measures, we review a panel of broadly used indices, describing their mathematical formulations, purposes and properties, and characterize their behaviour and criticalities in dependence of the data features using simulated data as ground truth. In addition, we make available an R package, DiversitySeq, which implements in a unified framework the full panel of diversity indices and a simulator of 16S-seq data, and thus represents a valuable resource for the analysis of diversity from NGS count data and for the benchmarking of computational methods for 16S-seq.

Key words: microbiome; 16S ribosomal RNA (rRNA) gene; evenness; richness; alpha diversity; beta diversity

Introduction

The human body hosts trillions of commensal, symbiotic and pathogenic microorganisms, mainly bacteria, which are collectively known as the human microbiota. In recent years, substantial efforts have been directed towards the investigation of the composition of the microbiota in the different body sites [1,2], its association with environmental factors, such as diet and lifestyle [3–5], and its role in human health and disease [2, 6–11]. The opening of this fascinating research frontier and the establishment of large-scale metagenomics projects like the Human Microbiome Project (HMP) [1, 12, 13] (http://hmpdacc.org) have become possible only a decade ago, with the advent of next-generation sequencing (NGS) technologies [14]. Amongst NGS methodologies, targeted amplicon sequencing of the 16S ribosomal RNA (16S rRNA) gene, referred to as '16S-seq' from here on, is currently the most used strategy for the identification and quantification of human-resident bacteria [15]. The 16S rRNA gene is used as target gene because it is present in all prokaryotes and contains regions that are highly conserved between species and regions that are, instead, species-specific. The conserved regions serve as amplification targets for polymerase chain reaction (PCR) primers to extract one or more variable

Francesca Finotello is a postdoctoral research fellow at the Division of Bioinformatics, Biocenter, Medical University of Innsbruck, Austria. She works on the development and application of computational methods for the analysis of high-throughput 'omics' data, primarily from NGS. Her research is currently focused on bioinformatics for precision immuno-oncology.
Eleonora Mastrorilli is a research fellow at Department of Food Safety, Istituto Zooprofilattico Sperimentale delle Venezie, Legnaro, Padova, Italy. Her work focuses on bioinformatic analysis of NGS data for microbial community ecology and food-borne pathogen characterization.
Barbara Di Camillo is an assistant professor of Bioengineering at the Department of Information Engineering, University of Padova, Italy. She works on bioinformatics, predictive modelling and statistical methods for high-throughput data analysis.
Submitted: 5 September 2016; **Received (in revised form):** 20 October 2016

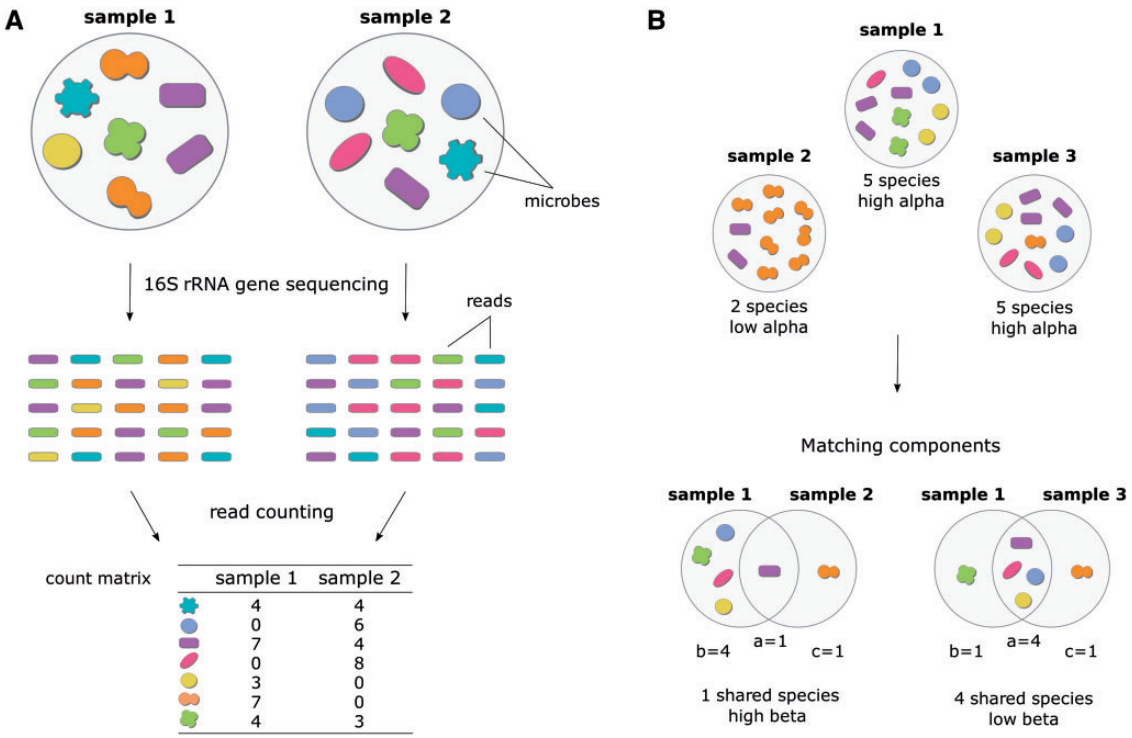

Figure 1. Schematization of 16S rRNA gene sequencing data generation and examples of microbial diversity. (**A**) 16S rRNA gene sequencing count data are generated through three main steps. First, the microbial communities of a set of samples are subjected to 16S rRNA gene sequencing. Then, the reads generated from all the sequenced 16S rRNA genes are computationally assigned to the different species of microbes. Finally, the reads assigned to each species are counted and summarized in a count matrix, representing the abundances of the different species in the assayed samples. (**B**) Alpha and beta diversity can be qualitatively described considering the species accounted by each sample or the species shared by pairs of samples, respectively. Sample 2 can be seen as an example of low alpha diversity, when compared with Samples 1 and 3, as it has a lower number of species and a lower evenness. The three samples can be also compared in terms of matching components, namely considering the number of common species *a* and number of unique species *b* and *c* (Koleff et al., 2003). Samples 1 and 2 have high beta diversity because they have few species in common. Differently, Samples 1 and 3 have low beta diversity because they have many shared species and only one unique species each. A colour version of this figure is available online at BIB online: http://bib.oxfordjournals.org.

regions of the 16S rRNA gene. The latter are then subjected to NGS, producing thousands of sequenced fragments, called 'reads'. Finally, the reads are mapped to known 16S rRNA gene sequences of different microbial taxa, such as genera or phyla, or clustered according to their sequence similarity into 'operational taxonomic units' (OTUs) [16]. The number of NGS reads assigned to each taxon/OTU, referred to as 'count', is used as a proxy of the abundance of that taxon/OTU in the original sample (Figure 1A). We refer the reader to a review [17] for a detailed description of 16S-seq data generation and analysis. Here and throughout the manuscript, we will use the term 'species' for its ease of interpretation, but all discussions and computations can be intended considering any taxonomical or OTU level.

Once microbial abundances are measured through 16S-seq, diversity indices provide valuable mathematical tools to investigate the dynamics of the human microbiota or to detect changes in its composition between different conditions or body sites. The terms alpha and beta diversity were originally introduced by the ecologist Whittaker [18, 19] in 1960 to describe the biodiversity of a landscape. According to Whittaker, total species diversity in a landscape (called 'gamma' diversity) is determined by two contributions: the average species diversity in sites belonging to a landscape (alpha diversity) and the differentiation among those sites (beta diversity). In other words, if the sites in a landscape are similar in terms of species composition, beta diversity is low. If instead some sites contain different sets of species, total gamma diversity exceeds the mean alpha diversity and beta diversity accounts for this additional variation. However, the definition of alpha and beta diversity has been revised several times, and a consensus on their mathematical formulation has not been reached yet. Despite some works have attempted to mathematically partition total diversity into alpha and beta diversity [20, 21], the standard practice is to use separately alpha and beta diversity indices to estimate a sort of intra- and inter-site diversity, respectively. In particular, in the study of the human microbiota, alpha diversity is used to describe the compositional complexity of a single sample, whereas beta diversity is used to describe taxonomical differences between samples. To simplify diversity indices down to an intuitive, qualitative definition, we can state that a sample has high alpha diversity when it contains a high number of equally abundant species, and low diversity otherwise. When comparing two samples, beta diversity is high if they share few species and low if most of their species are in common (Figure 1B).

Despite this seemingly simple summary explanation, diversity is not a determined physical quantity for which a consensus definition and unit of measure have been established. In fact, the multifaceted nature of the concept of diversity has motivated the development of several mathematical indices, which have been originally applied to ecology and, later, broadly used in the analysis of 16S-seq data. However, these indices have different purposes, units of measures and mathematical formulations, and thus might be more or less suited for responding to specific biological questions. Furthermore, they were originally developed for macroecology, and, to the best of our knowledge, their

Table 1. Indices of alpha diversity

Name	Alpha diversity index	Relation with ${}^{q}D$	Class	References
Observed richness	S^{obs}	${}^{0}D$	Richness	[19]
Chao1	$S^{Chao1} = S^{obs} + \frac{f_1 (f_1 - 1)}{2(f_2+1)}$		Richness	[23]
1st-order Jackknife	$S^{Jackk1} = S^{obs} + f_1$		Richness	[24]
2nd-order Jackknife	$S^{Jackk2} = S^{obs} + 2f_1 - f_2$		Richness	[24]
ACE	$S^{ACE} = S^{abund} + \frac{S^{rare}}{C^{ACE}} + \frac{f_1}{C^{ACE}} \cdot \gamma^{ACE2}$		Richness	[25]
Hill numbers	${}^{q}D = \left(\sum_{i=1}^{S^{obs}} p_i{}^{q} \right)^{\frac{1}{1-q}}$		Diversity	[27]
Berger–Parker	$BP = \frac{1}{max(p_i)}$	$\frac{1}{\infty D}$	Diversity	[28]
Rényi entropy	${}^{q}RE = \frac{1}{1-q} ln \left(\sum_{i=1}^{S^{obs}} p_i{}^{q} \right)$	$\ln({}^{q}D)$	Diversity	[29]
Inverse Simpson	$IS = \frac{1}{\sum_{i=1}^{S^{obs}} p_i{}^{2}}$	${}^{2}D$	Diversity	[30]
Gini–Simpson	$GS = 1 - \sum_{i=1}^{S^{obs}} p_i{}^{2}$	$1 - \frac{1}{{}^{2}D}$	Diversity	[30]
Shannon entropy	$H = -\sum_{i=1}^{S^{obs}} p_i \cdot ln(p_i)$	$\ln({}^{1}D)$	Diversity	[31]
Tail	$T = \sqrt{\sum_{i=1}^{S^{obs}} (i-1)^2 \cdot p_i}, \quad \text{with } p_1 \geq p_2 \geq \ldots \geq p_S$		Diversity	[32]
EF	${}^{q}EF = \frac{{}^{q}D}{{}^{0}D} = \frac{{}^{q}D}{S^{obs}}$		Evenness	[22]
RLE	${}^{q}RLE = \frac{\ln({}^{q}D)}{\ln({}^{0}D)} = \frac{\ln({}^{q}D)}{\ln(S^{obs})}$		Evenness	[22]
Pielou	$P = \frac{\ln({}^{1}D)}{\ln(S^{obs})}$		Evenness	[33]

robustness to the possible bias introduced by sequencing has not been characterized so far.

To assist the reader with the selection and interpretation of diversity indices, in the present work, we review their mathematical properties and explore their features trough numerical simulations. In particular, we: (i) describe the mathematical formalism, purposes and properties of alpha and beta diversity indices; (ii) use numerical simulations to characterize their behaviour in dependence on data characteristics; (iii) present an R package, DiversitySeq, which implements in a unified framework the full panel of indices, easing the computation and visualization of alpha and beta diversity, and a simulator of 16S-seq data. We believe DiversitySeq can be a valuable resource for the analysis of diversity from NGS count data and for the benchmarking of computational methods for 16S-seq.

Measures of alpha and beta diversity

Alpha diversity

Although the quantitative definition of alpha diversity has been disputed, we can qualitatively describe alpha diversity (also called 'local diversity') as a measure of the compositional complexity of a community within a site, which increases with the number of present species and with the evenness of their relative abundances [22]. Therefore, for a given number of different species S, alpha diversity is maximal when all species are equally abundant. Several indices of alpha diversity are available (Table 1): richness indices, which estimate the number of different species in a sample; evenness indices, which consider the species relative abundances, without focusing on their total number; and diversity indices, which account for both the species relative abundances and the total number of different species.

Richness indices

Richness provides an intuitive measure of alpha diversity by simply counting the number of different species present in a sample. However, during a site sampling and NGS, some rare species can be lost, resulting in an underestimation of species richness. To address this issue, richness estimators predict the true richness of a sample by correcting the observed richness S^{obs} for the number of lost species, estimated considering the distribution of the rarest species. In particular, Chao1 index [23] and the first- and second-order Jackknife indices [24] use the singletons f_1 and doubletons f_2, i.e. the number of different species with exactly one and two counts, respectively, to estimate the true number of different species present in a sample (Table 1). Here and in the following, we use the term 'count' to generally refer to the number of individuals or reads accounted by each species.

A slightly more complex approach, called Abundance Coverage Estimator (ACE) [25], considers f_x, namely the number of different species with exactly x counts, and computes species richness as:

$$S^{ACE} = S^{abund} + \frac{S^{rare}}{C^{ACE}} + \frac{f_1}{C^{ACE}} \cdot \gamma^{ACE2}.$$

Where $S^{rare} = \sum_{x=1}^{10} f_x$ is the number of rare species, with ≤ 10 counts.

$S^{abund} = \sum_{x=11}^{K} f_x$ is the number of abundant species, with >10 counts.

$C^{ACE} = 1 - \frac{f_1}{n^{rare}}$ is the proportion of counts accounted by singletons, with $n^{rare} = \sum_{x=1}^{10} x \cdot f_x$ representing the total number of counts accounted by all rare species.

$\gamma^{ACE} = \frac{S^{rare}}{C_{ACE}} \sum_{i=1}^{10} i(i-1) f_1 \frac{n^{rare}-1}{n^{rare}}$ is the coefficient of variation of the rare species [26, 34].

Diversity indices

More complex alpha diversity indices can be computed taking into account the relative abundance p_i of each species $i = 1, \ldots, S$. A commonly used set of alpha diversity indices is the so-called Hill numbers of order q [27], which are defined as:

$$ {}^qD = \left(\sum_{i=1}^{S^{obs}} {p_i}^q \right)^{\frac{1}{1-q}}. $$

For $q = 1$, the previous equation is undefined, and the following formula is used:

$$ {}^1D = \frac{1}{\prod_{i=1}^{S^{obs}} p_i^{p_i}} = \exp\left(-\sum_{i=1}^{S^{obs}} p_i \cdot \ln(p_i) \right). $$

For $q \to \infty$, Hill index corresponds the inverse of the Berger–Parker dominance index [28].

By re-expressing ${}^qD$ as:

$$ {}^qD = \frac{1}{\left(\sum_{i=1}^{S^{obs}} p_i {}^{q-1} p_i \right)^{q-1}}, $$

it can be noticed that ${}^qD$ gives a weighted mean of species abundances and that q defines the type of weighted mean computed at the denominator. For instance, for $q = 0$, the harmonic mean is computed; for $q = 1$, the geometric mean; and for $q = 2$, the arithmetic mean. Moreover, by increasing the value of q, the weight given to the most abundant species with respect to the rarest ones increases. For $q = 0$, species weights cancel out species abundances, and the index is equivalent to species richness ${}^0D = S^{obs}$, provided that the sum is computed only on present species with $p_i > 0$.

Another set of indices, called Rényi entropy of order q [29], can be computed by taking the natural logarithm of Hill numbers (Table 1).

Other commonly used diversity measures are the inverse Simpson index, the complementary Simpson or Gini–Simpson index and the Shannon entropy (Table 1). As it can be noted from Table 1, all these indices can be defined starting from Hill numbers ${}^qD$. We did not consider the original Simpson index [30] because it is not a measure of alpha diversity: it decreases when the number of species increases or when abundances are less evenly distributed across species.

Li and colleagues [32] recently proposed a diversity measure called Tail statistic to improve sensitivity in the presence of rare species. In Tail formula (Table 1), the relative abundances are supposed to be sorted in decreasing order, such that $p_1 \geq p_2 \geq \ldots \geq p_S$. Differently from Hill diversity, Tail weights relative abundances considering species ranks (with the most abundant species p_1 having weight zero).

Evenness indices

Evenness indices measure how evenly the relative abundances are distributed across the different species. Thus, besides being a valuable indicator of biodiversity, evenness also determines the stability and resilience of an ecosystem [35]. A commonly used evenness measure is the evenness factor ${}^qEF$, defined based on Hill diversity (Table 1). Its reciprocal is called inequality factor and defined as ${}^qIF = 1/{}^qEF$.

It must be noted that it is impossible for two communities with minimal evenness (i.e. ${}^qD_1 = {}^qD_2 = 1$), but different total number of species (e.g. $S_1 = 10$ and $S_2 = 1000$), to obtain the same EF. In particular, the community with more species will always have the lowest EF (${}^qEF_1 = 0.1$ and ${}^qEF_2 = 0.001$). For this reason, indices of relative evenness (RLE) and relative inequality (RLI), which scale to the range of values that are possible for a given richness S, have been developed. ${}^qRLE$ (see formula in Table 1) varies between 0 and 1, reaching its maximum when all species are equally abundant, while its corresponding RLI factor ${}^qRLI = 1 - {}^qRLE$ has the opposite behaviour. When $q = 1$, the RLE factor ${}^qRLE$ corresponds to Pielou index [33] (Table 1). Although other relative indices are available, the logarithmic transformation used in RLE and RLI formulation is preferable, as it preserves the complementarity between evenness and inequality [22]. In particular, Pielou index is broadly used to measure evenness in microbiota studies.

Beta diversity

As discussed in the 'Introduction' section, beta diversity is commonly used in microbiota studies to highlight taxonomical differences between pairs of samples. To this end, species abundances are usually not taken into account and presence–absence data are instead used to identify which species are shared by samples and which are not. In particular, the list of present species in two samples is expressed in terms of matching components [36] a, b and c (Figure 1B). a is the number of species shared by both samples, whereas b and c are the number species present only in the first or in the second sample, respectively. In the following, we adopt this notation and refer to the total number of species accounted by a pair of samples as $N = a + b + c$.

Beta diversity indices

From the beta diversity indices reviewed in Koleff *et al.* [36], widely used by the scientific community since their implementation in the R package Vegan [34], we considered a panel of 14 non-redundant indices expressed in terms of matching components a, b and c (Table 2, naming and formulation as in [36]). For sake of consistency and interpretability, we focused on dissimilarity indices, which return lower beta values when the number of shared species a increases. Accordingly, we transformed similarity indices into dissimilarity measures as $\beta_{dissimilarity} = 1 - \beta_{similarity}$. Moreover, we excluded β_{rlb} because it is not symmetric and β_{gl} because it is not an index of beta diversity, but rather a measure of local alpha diversity gradients (i.e. it measures relative differences in species richness between samples). Finally, we introduced an additional index β_{mn}, which is the normalized version of β_m, and scaled β_{-3} to $\beta_{-3n} = 2 \cdot \beta_{-3}$, so to have them ranging between 0 and 1 (Table 2).

Properties of diversity indices

To ease the interpretation of alpha and beta diversity indices (Tables 1 and 2) and to identify different classes of indices, we considered the following properties: number equivalents,

Table 2. Indices of beta diversity

Beta diversity index	Equivalent measures	References
$\beta_w = \frac{b+c}{2a+b+c}$	β_{-1}, β_t, β_{me}, $1-\beta_{sor}$, β_{hk}	[18, 37–43]
$\beta_c = \frac{b+c}{2}$	$\beta_{wb}/2$, β_l	[44–46]
$\beta_{cc} = \frac{b+c}{a+b+c}$	β_g, $1-\beta_j$	[37, 38, 48–50]
$\beta_{co} = 1 - \frac{a\cdot(2a+b+c)}{(a+b)(a+c)}$		[51]
$\beta_m = \frac{(2a+b+c)(b+c)}{(a+b+c)}$		[37]
$\beta_{mn} = \frac{(2a+b+c)(b+c)}{(a+b+c)^2}$		Derived from [37]
$\beta_{rs} = \frac{2\cdot(bc+1)}{(a+b+c)^2-(a+b+c)}$		[52, 53]
$\beta_r = \frac{2bc}{(a+b+c)^2-2bc}$		[37, 38, 54]
$\beta_{-3n} = \frac{2\cdot min(b,c)}{a+b+c}$		Derived from [52]
$\beta_{-2} = \frac{min(b,c)}{a+max(b,c)}$		[39]
$\beta_{sim} = \frac{min(b,c)}{a+min(b,c)}$		[55]
$\beta_I = \log(2a+b+c) - \frac{2a\cdot log(2)+(a+b)\cdot log(a+b)+(a+c)\cdot \log(a+c)}{2a+b+c}$		[40, 54]
$\beta_e = \exp(\beta_I) - 1$		[54]
$\beta_z = 1 - \frac{log(2a+b+c)-\log(a+b+c)}{\log(2)}$		[43, 55]

number of species, doubling and replication invariance for alpha diversity indices; homogeneity, measures of continuity or gain and loss, zero beta diversity in nested condition and beta diversity inverse to a/N for beta diversity indices, as detailed in the following.

Number equivalents. To clarify the concept of numbers equivalent (also called 'effective numbers') [22], suppose a diversity measure estimated from a data set has a value of $\bar{D}$. Amongst all the possible data sets having the same diversity $\bar{D}$, there is one composed by $\bar{S}$ equally abundant species. The number of species $\bar{S}$ making up this equivalent data set is the effective numbers of species. Hill numbers can be directly interpreted as effective number of species, and most of alpha diversity indices can be converted to effective number of species by re-expressing them as Hill numbers (Table 1). Although often neglected in microbiota studies, this property facilitates interpretability because diversity measures that can be rendered as number equivalents have the 'number of species' as unit of measure and possess the doubling property (discussed in what follows).

Number of species. Despite not mandatory, having the number of species as unit of measure of alpha diversity facilitates the quantitative interpretation of results. For example, if two communities A and B have diversity $D_A = 200$ and $D_B = 150$, we can easily state the community A has 50 species more than the community B.

Doubling. If we imagine to merge two equally large communities with the same diversity $\bar{D}$, but no species in common, when the doubling property holds, the diversity of the doubled community would be $\bar{D}' = 2\bar{D}$. To understand the usefulness of this property for data interpretation, we can imagine a site hosting a microbial community with 10 000 equally abundant species. A perturbation causes a loss of 5000 species. ${}^0D$ gives an intuitive measure of pre- and post-perturbation diversity, measuring a 50% loss:

$$\frac{{}^0D_{post} - {}^0D_{pre}}{{}^0D_{pre}} \cdot 100 = \frac{5000 - 10000}{10000} \cdot 100 = -50\%.$$

Differently, the interpretation of species loss measured by Shannon entropy H, here used as an example, is less straightforward:

$$\frac{H_{post} - H_{pre}}{H_{pre}} \cdot 100 = \frac{8.52 - 9.21}{9.21} \cdot 100 = -7.50\%.$$

None of the indices is wrong, as they correctly assay the direction of the change: both have a lower value after the perturbation, indicating a loss in diversity. However, the doubling property enables a straightforward and quantitative interpretation of diversity.

Replication invariance. Some alpha diversity measures that do not obey to the doubling property are instead replication invariant. For these measures, the merged community of the example described above would have the same evenness $\bar{D}' = \bar{D}$ as the two original communities. Evenness measures, which do not obey to the doubling property, are in some cases replication invariant. To understand the meaning of this property, let consider two communities: a community A with two species, one with 100 counts and one with only one count, $k_A = [100, 1]$; and a community B consisting of two replicates of A, $k_B = [100,\ 100,\ 1,\ 1]$. Replication invariant measures determine the same evenness for A and B. Differently, measures that are not replication invariant assign a lower evenness to A, as it shows greater differences between species relative abundances ($p_A = [0.010,\ 0.990]$) than B ($p_B = [0.005,\ 0.005,\ 0.495,\ 0.495]$).

Homogeneity. Beta diversity measures that satisfy the property of homogeneity are independent from the total number of species N, as long as the proportions of shared and unique species a/N, b/N and c/N are constant [36]. This property is important to guarantee comparability between data sets with different number of present species.

Measures of continuity or gain and loss. To highlight differences in how beta diversity indices combine a, b and c, a distinction between measures of continuity and measures of gain and loss has been proposed [36]. Measures of continuity only depend on the number of shared species a, while measures of gain and loss obtain high diversity when the number of shared species a is low, and the number of unique species b and c (i.e. the species gained by one sample or lost by the other) are similar.

Zero beta diversity in nested condition. Some beta diversity measures depend on the unique-species ratio, i.e. the ratio between the species that are unique to one or to the other sample. When the unique-species ratio equals 1, the number of unique species in both samples is the same, while lower values of the ratio indicate that one sample has more unique species than the other. When the unique-species ratio equals 0, one sample has no unique species and is entirely contained within the other one (nested condition). In this condition, some indices are constantly equal to 0, indicating minimum diversity, while others depend on the number of unique species in the largest sample.

Beta diversity inverse to a/N. The relationship between the fraction of shared species a/N and beta diversity is an important feature for the understanding of microbiota studies, where high beta diversity is interpreted in terms of large compositional differences between samples. Beta diversity should decrease as a/N increases and should obtain its maximal value when there are no species in common between the compared samples ($a = 0$).

Table 3. Summary of the properties of alpha diversity indices

Property	S^{obs}	Chao1	Jackknife1	Jackknife2	ACE	Hill, $q > 0$	Gini–Simpson	Shannon	EF	RLE
Number equivalents	x					x				
Number of species	x	x	x	x	x	x				
Doubling	x		x	x		x				
Replication invariance									x	

Table 4. Summary of the properties of beta diversity indices

Property	β_w	β_c	β_{cc}	β_{co}	β_m	β_{mn}	β_{rs}	β_r	β_{-3n}	β_{-2}	β_{sim}	β_I	β_e	β_z
Homogeneity	x		x	x		x		x	x	x	x	x	x	x
Continuity	x	x	x		x	x								x
Gain and loss				x			x	x	x	x	x	x	x	
Zero in nested condition							x	x	x	x	x			
Inverse to a/N	x	x	x	x		x	x	x	x	x	x			x

Characterization of diversity indices through numerical simulations

To facilitate the assessment and interpretation of the properties described above, we characterized the selected diversity indices using numerical simulations. The results are thoroughly presented in what follows and summarized in Tables 3 and 4.

Alpha diversity indices

The properties of the different alpha diversity indices are summarized in Table 3. Rényi entropy was considered only for $q = 1$, which corresponds to Shannon entropy. While inverse Simpson, Berger–Parker and Pielou indices correspond to ${}^2D$, $\frac{1}{{}^\infty D}$ and ${}^1RLE$, respectively. We adopted the ACE index implemented in Vegan [34] and replaced non-finite estimates with S^{obs}.

To facilitate the interpretation of the alpha diversity indices, we first simulated two simple scenarios so to introduce the basic properties of alpha diversity indices before proceeding with more complex assessments. In one scenario, all species are equally abundant, while in the other one, half of the species are rare, i. e. they account for a single count. Several communities were simulated considering an increasing number of different species, up to 100. For this simplistic simulation, we considered only a subset of indices to highlight the major hallmarks of the different classes of alpha measures (Figure 2), and excluded richness estimators and high-order Hill numbers, EF and RLE indices, as they require a finer grid of possible abundance values to be assessed (described in the following paragraphs).

From Figure 2, it can be noticed that, when ${}^0D$ is used, alpha diversity equals species richness S^{obs}: it exactly counts the number of present species, without accounting for their abundances. When q increases, the rarest species are down-weighted, and only the abundant species are considered for the computation, thus resulting in a lower diversity. In particular, when $q = 1$, each species is weighted by its proportional abundance. When $q < 0$, rare species are over-weighted, and the final diversity can exceed S, resulting in values of difficult interpretation. Therefore, from here on, we do not consider diversity and evenness indices of order $q < 0$.

Tail index also takes into account species abundances and obtains a lower diversity in the scenario with rare species, but its values cannot be directly interpreted as number of different species.

Shannon index encodes the same information of ${}^1D$, but computed on a logarithmic scale. Thus, also Shannon index values cannot be directly interpreted as number of different species.z

Gini–Simpson encodes the same information of ${}^2D$ because they are respectively the complement and the inverse of the Simpson dominance index, but only ^{2}D can be interpreted in terms of number of species because it is expressed as numbers equivalent.

Finally, RLE and EF do not measure the number of species, but the ratio between the number of abundant species and the number of total species, computed on a logarithmic or a natural scale, respectively. They reach the maximum of 1 when all species are equally abundant and decrease when more rare species are present. EF can be directly interpreted as the proportion of abundant species in the community (see the scenario with 50% or rare species in Figure 2) and is constant as long as the percentage of rare species is maintained, irrespectively of total species. Contrariwise, RLE accounts for differences in total species, reducing evenness in communities with smaller sample size.

The scenarios considered in Figure 2 are useful to understand the basic behaviour of the alpha indices, but are not realistic. Therefore, we simulated several data sets with different number of species (up to 1000 different species) and different relative abundances. The species-collection phase was simulated with a Poisson sampling, considering a scenario of deep sampling, were most of the species are likely to be represented with their correct abundances, and a scenario of shallow sampling, in which some rare species are lost and some abundances are biased (see Supplementary Figure S1 and 'Materials and methods' section for further details on data simulation).

As expected, species number could be robustly estimated when deep sampling was performed (Figure 3), but was underestimated in the shallow-sampling scenario (Figure 4), especially for low-evenness communities. The bias was corrected by Jackknife1 and 2, Chao1 and ACE estimators in most of the cases, but not for communities with low evenness.

Indices for diversity and evenness always decreased as evenness decreased. By using a finer grid of possible abundance values, it can be seen that as the order q increases (see for instance ${}^1D$ and ${}^\infty D$ in Figure 4), more and more species are considered rare and a lower diversity is obtained.

Gini–Simpson index varied little (median 0.96 $\pm$ 0.06 median absolute deviation, MAD, in both simulations), depending on species abundances and evenness, and quickly saturated to 1 when the number of species exceeded 10.

The same simulated data were also used to assess the properties of doubling and replication invariance. Hill numbers satisfied the doubling property, and the ratio between $\bar{D}'$ and $\bar{D}$ (called from here on 'replication diversity ratio') was exactly 1 (Figure 5). Amongst richness estimators, only Jackknife

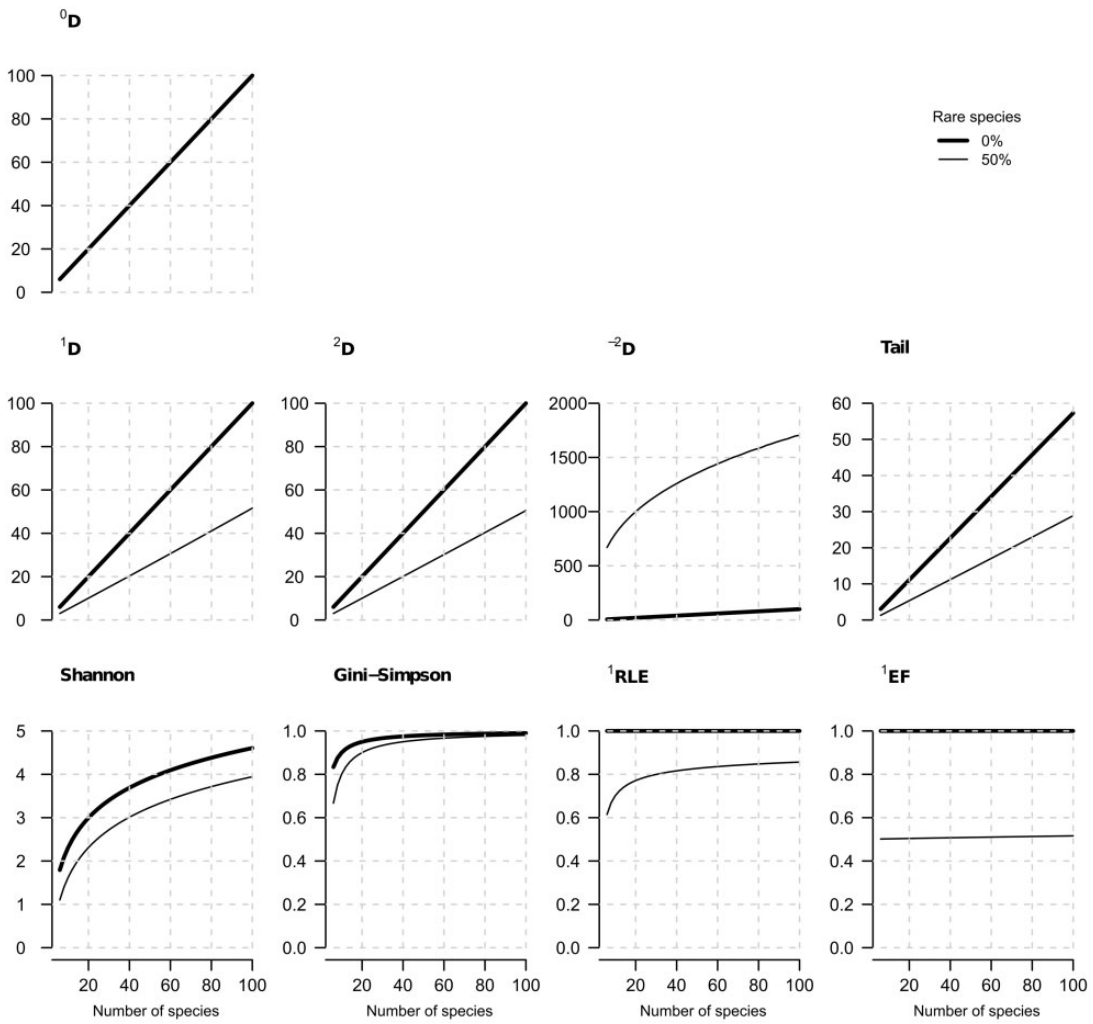

Figure 2. Alpha diversity indices applied to communities of 1:100 species, with all species equally abundant (thick line) or with 50% rare species and 50% abundant species (thin line).

estimators obeyed to the doubling property, while ACE and Chao1 presented some shifts from 1. Tail, Gini–Simpson and Shannon showed larger deviance from 1. As discussed in the 'Properties of diversity indices' section, only non-relative evenness measures EF are strictly replication invariant.

Beta diversity indices

We simulated different number of total (N), shared (a) and unique (b and c) species to characterize beta diversity indices in terms of range of values, dependence on N and a and impact of the unique-species ratio, as detailed in 'Materials and Methods' section. The properties of the different beta diversity indices are summarized in Table 4.

From Figure 6 and from the equations in Table 2, it can be noticed that most of beta indices range between 0, representing minimum diversity, and 1, representing maximum diversity. Only the maximum of β_m, β_c, β_I and β_{rs} indices is not equal to 1:

- $\max(\beta_m) = N$;
- $\max(\beta_c) = N/2$;
- $\max(\beta_I) = \log(2) = 0.69$;
- $\max(\beta_{rs}) = \frac{N^2+4}{2N^2-2N} = 0.5 + \varepsilon$, with $\varepsilon < 0.03$ for $N > 20$.

Conversely, the normalized versions, $\beta_{cc} = 2\beta_c/N$, $\beta_{mn} = \beta_m/N$ and $\beta_e = \exp(\beta_I) - 1$, range between 0 and 1.

Moreover, β_c and β_m do not obey to the homogeneity property and depend on the total number of species N, as it can be noticed from the formulae above and from Figure 6. β_{rs} is not strictly homogeneous, but for $N = 1,000$, the error ε can be considered negligible.

The measures of continuity (Table 4) do not depend on the unique-species ratio (see overlapping curves in Figure 6), while all the remaining beta indices depend on the unique-species ratio, presenting higher diversity when b and c are similar. To clarify the impact of b and c on beta diversity, we assessed the variation of beta diversity when a and c are fixed and b varies (i.e. when one sample gains more and more unique species), being lower, equal or higher than c (Supplementary Figure S2). The classification of indices proposed in [36] can be further particularized with subsets of indices with peculiar behaviours. For example, amongst the continuity measures, β_c and β_m can be though as special cases because of their lack of homogeneity (Table 3). In particular, β_m reached higher diversity when a was higher (see the orange curve over the purple one in Supplementary Figure S2). This behaviour is particularly counterintuitive, as diversity should decrease as the number of shared species increases. From Supplementary Figure S2, it can

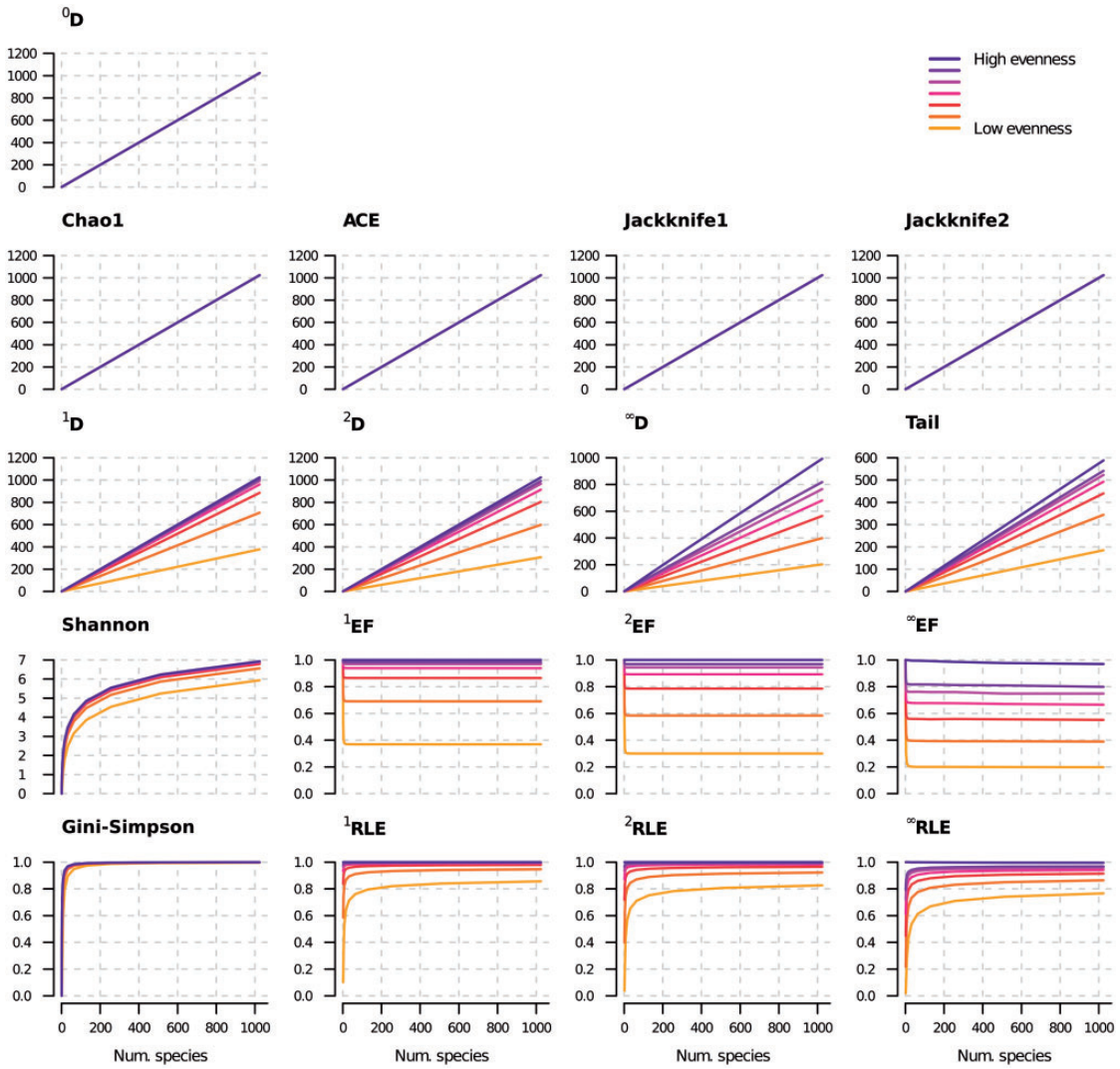

Figure 3. Alpha diversity indices applied to simulated communities with 1e7 individuals, different richness and different distributions of individuals across species (evenness). A colour version of this figure is available online at BIB online: http://bib.oxfordjournals.org.

be noticed also that, for $a \neq 0$, continuity measures except β_c and β_m tended to one as b increased, indicating maximal beta diversity. In case of no shared species ($a = 0$), they were constantly equal to one, returning maximal diversity independently from the number of unique species b and c. Conversely, most of the measures of gain and loss showed higher diversity when and b and c were similar (see the peaks in Supplementary Figure S2), and reduced diversity when the difference between b and c became larger.

We simulated the nested condition with different fractions of shared species a/N by varying the number of shared species a or the total number of species N. These two scenarios lead to equal beta diversity (or comparable in case of β_{rs}) for all indices, except β_m and β_c, for which the scale is affected by non-homogeneity (Supplementary Figure S3). In the nested condition, β_{rs}, β_r, β_{-3n}, β_{-2} and β_{sim} obtained their maximum diversity irrespectively from b and c (see flat curves in Figure 6 and Supplementary Figure S3), while the others depended on the number of unique species in the largest sample.

Overall, all beta measures decreased as the fraction of shared species a/N decreased as expected, except for β_m when $a = 0$ (Supplementary Figure S2), and for β_I and β_e, which decreased monotonically only when the number of unique species strongly differed (Figure 6 and Supplementary Figure S3).

Effect of low sequencing depth on alpha and beta diversity indices performance

As discussed above, the estimation of the true diversity from microbial data is not trivial because, during a site sampling, some rare species can be lost and cause underestimation of species richness. After site sampling, sequencing may introduce an additional bias in species abundances and worsen richness underestimation. Excluding technical artefacts, the main reason for this bias might be the limited *sequencing depth* dedicated to each sample: a maximum number of reads is generated from the sequencing of each sample — usually in the order of 10^5 reads per sample with the Illumina 16S-seq (http://www.illumina.com/). The expected counts for the i-th species in the j-th sample can be estimated as $E[k_{ij}] = p_{ij} \cdot \gamma_j$, where γ_j is the sequencing depth of the j-th sample, and p_{ij} is the relative abundance of the i-th species in the j-th sample after site sampling (see 'Materials and methods'). Therefore, samples that are sequenced at lower sequencing depths obtain lower counts per

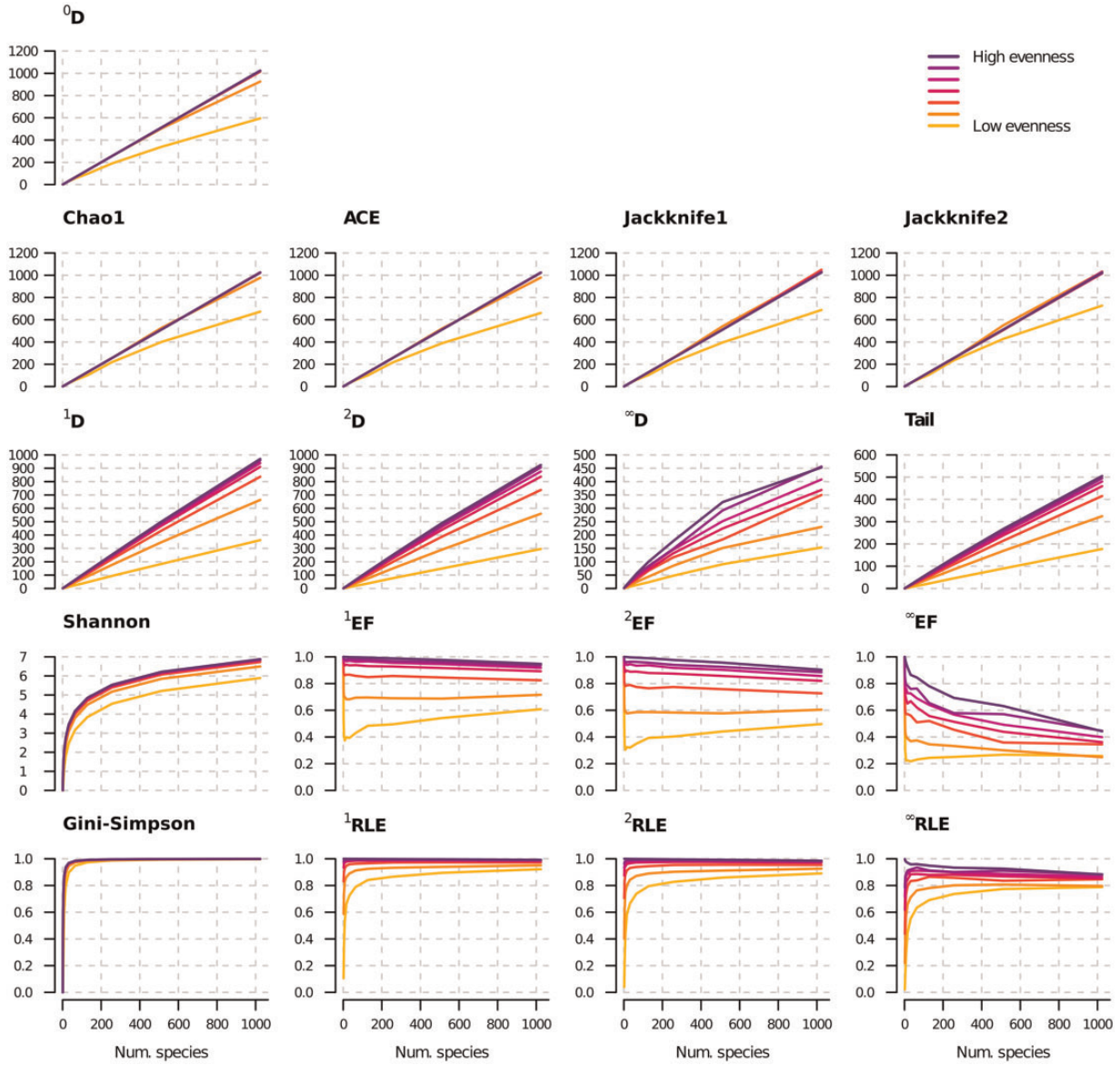

Figure 4. Alpha diversity indices applied to simulated communities with 1e4 individuals, different richness and different distributions of individuals across species (evenness). A colour version of this figure is available online at BIB online: http://bib.oxfordjournals.org.

species, and, above all, the rare species are less likely to be represented in the final count matrix $[k_{ij}]$ than the abundant ones. As a consequence, the richness observed in a 16S-seq experiment S_j^{seq}, computed by summing up the number of species with at least one count (Figure 1A), is generally lower than or equal to the richness measured from the collected sample S_j^{samp}, which in turn is lower than or equal to the true richness S_j^{true} of the site of interest:

$$S_j^{seq} \leq S_j^{samp} \leq S_j^{true}.$$

Diversity indices were originally developed for species-abundance data, but they can be applied also to count data $[k_{ij}]$ using the relative counts p'_{ij} as proxy of the relative abundances

$$p'_{ij} = \frac{k_{ij}}{\sum_i k_{ij}}.$$

However, the bias introduced by sequencing in the final count matrix $[k_{ij}]$ may challenge the estimation of true species diversity. To test the robustness of alpha and beta diversity indices to the bias introduced by sequencing, we generated a synthetic 16S-seq data set using a negative binomial (NB) model (see 'Materials and methods'). In particular, we simulated a matrix of species abundances over 8048 species and 20 samples, along with a corresponding matrix of counts resulting from its sequencing. The parameters for the simulation were extracted from a real data set of the HMP [12].

The real count data showed the typical overdispersion of NB distribution (Supplementary Figure S4), confirming that the NB distribution is suited for the description of NGS data [56, 57]. The simulated counts were comparable with the real counts both in terms of data distribution (Supplementary Figure S5) and overdispersion (Supplementary Figure S4).

The performance of the richness indices (Figure 7A) were tested in terms of *relative error*, computed by comparing the

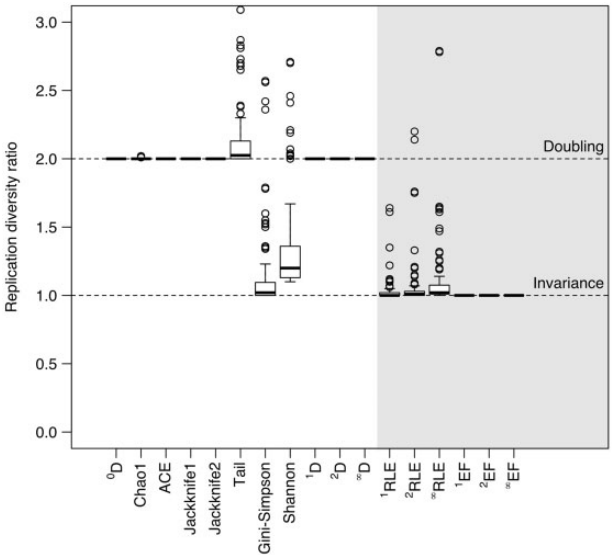

Figure 5. Ratio between the diversity of the doubled community and that of the original one (replication diversity ratio). Evenness indices are highlighted by the shaded grey area. The values of 1 and 2 represent the conditions of doubling and replication invariance, respectively.

estimated richness with the true number of present species in the matrix of abundances ('Materials and methods'). The sequencing bias resulted in richness underestimation for all the tested indices (relative error ranging between −71 and −27%). Jackknife2 index showed the smallest relative error.

We applied alpha diversity, evenness (Figure 7B) and beta diversity (Figure 7C) indices to both abundances and count matrices, and measured their relative variation (see 'Materials and methods' section). This variation does not represent a deviation from the true diversity because the latter would depend on the specific index adopted, but it is inversely related to the robustness of the index to the technical variability introduced by sequencing. Gini–Simpson, Shannon, ^2D, $^\infty D$, ^1RLE, ^2RLE, $^\infty RLE$ and $^\infty EF$ obtained a median variation equal or close to 0%, meaning that, in most of the cases, sequencing had a limited impact on their estimates. In particular, Gini–Simpson, ^2RLE and $^\infty RLE$ relative error ranged in ±2%. ^1D, ^1EF and Tail showed larger relative variations because of sequencing. Of note, evenness indices were computed slightly differently from their original formulation (Table 1): instead of the number of present species S^{obs}, the total number of assayed species (i.e. the number of rows in the count matrix) was used, so to make evenness comparable across samples. When used in their original formulation, larger relative variations were obtained (Supplementary Figure S6).

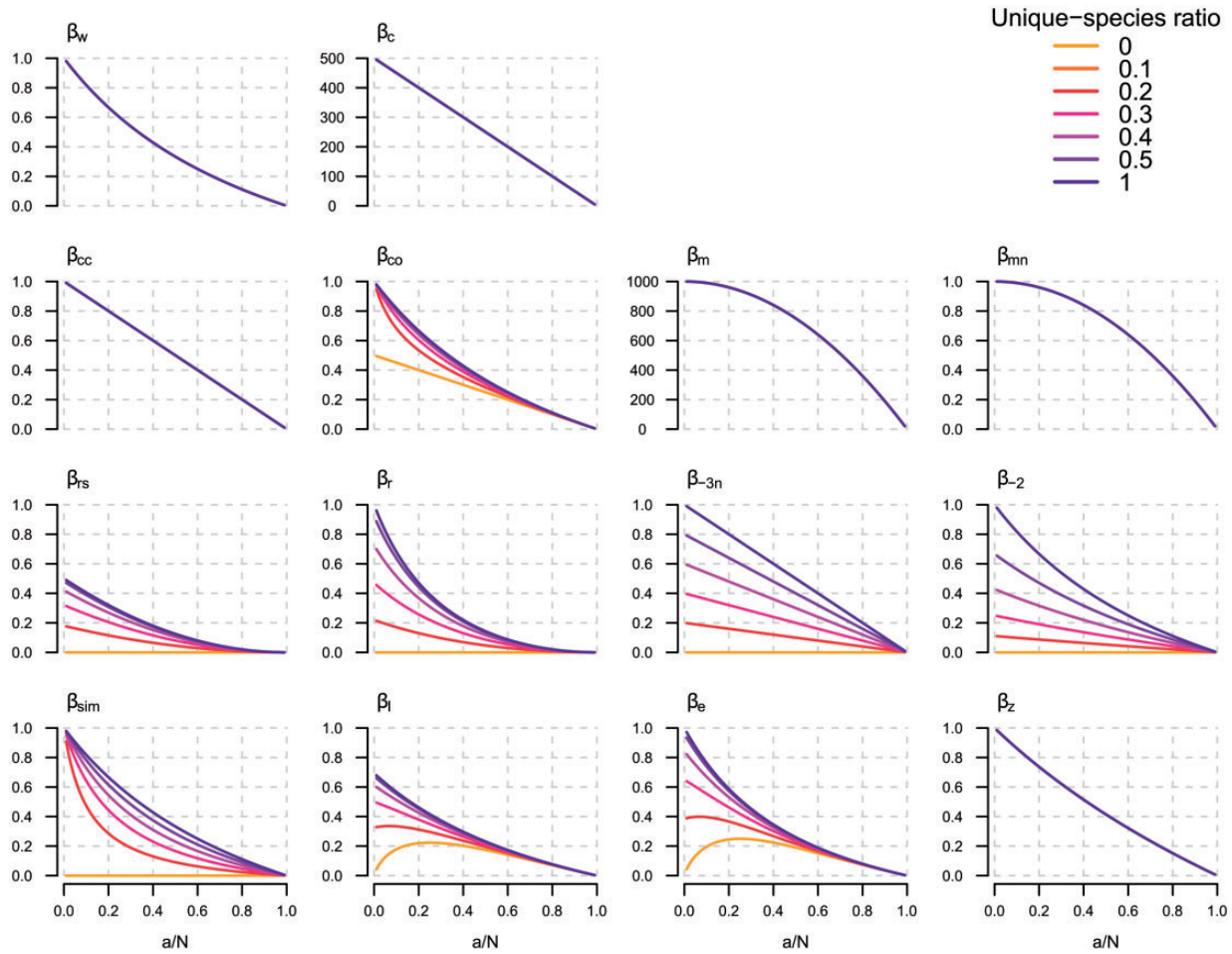

Figure 6. Values of beta diversity indices for different fractions of shared species a/N and different unique-species ratios b/c. A colour version of this figure is available online at BIB online: http://bib.oxfordjournals.org.

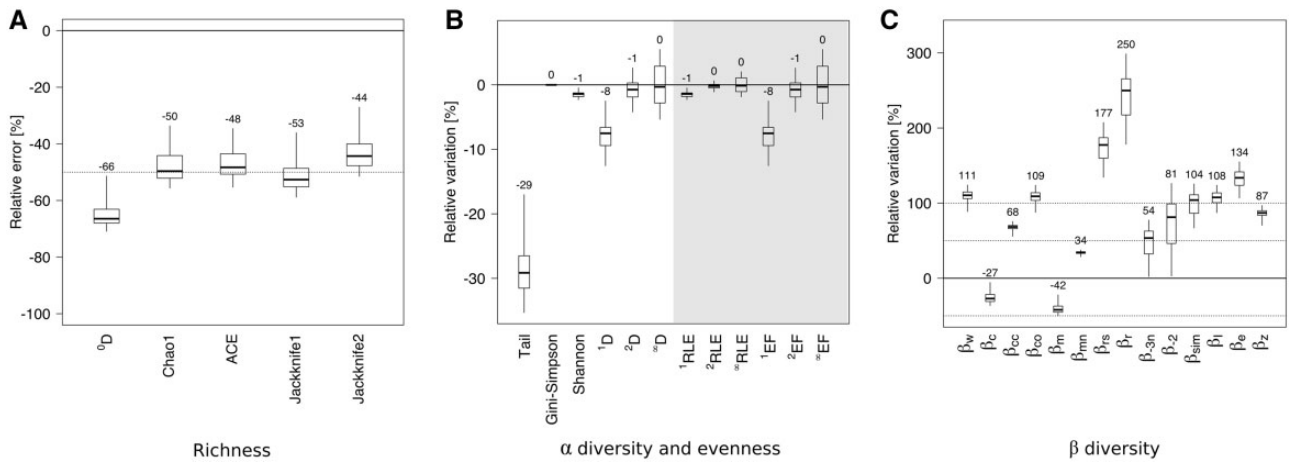

Figure 7. Boxplots of relative error of richness indices (**A**) and relative variation of alpha diversity (**B**) or evenness indices (**B**, grey shaded area) and beta diversity indices (**C**). The solid black line represents null relative error or variation. The medians of the distributions are reported over the boxplots.

Beta diversity indices were more heavily affected by the sequencing bias, and showed large relative variations, often exceeding 100%. Only β_c, β_m and β_{mn} ranged in $\pm 50\%$.

DiversitySeq: an R package for the simulation of 16S-seq count data and analysis of diversity

We developed an R package, named DiversitySeq, which implements a simulator of synthetic 16S-seq data sets and the functions to estimate and visualize alpha and beta diversity.

The simulator, implemented in the 'simcounts' function, takes as input a vector of species average abundances $\bar{\mu}_i$, a vector of sequencing depths γ_j for the m samples to be simulated and a coefficient of dispersion ϕ (see 'Materials and methods').

In addition, DiversitySeq enables the easy computation of alpha and beta diversity, with any of the indices described in the present work, using a single R function: 'aindex' or 'bindex', respectively. Evenness indices can be also computed on the total number of assayed species, as discussed in the 'Results' section, specifying the option 'keep0 = TRUE'. Diversity values can be visualized as boxplots using the 'divplot' function contained in the package. Finally, DiversitySeq can be easily downloaded and installed using the instructions in the package vignette available here: http://sysbiobig.dei.unipd.it/?q=Software#DiversitySeq.

Discussion

The application of a panel of diversity indices to simulated species abundance data highlighted that they are characterized by different properties and affected in different ways by the bias introduced by undersampling and sequencing.

The observed richness, computed by simply counting the number of present species, is strongly affected by this bias. More complex richness indices, such as Jackknife, ACE and Chao1, are able to estimate the true richness when rare species are lost during sample collection, but are strongly affected by low sequencing depth, which results in underestimated richness for all the tested indices. The bias is worsened under the condition of low species evenness, which is typical for a 16S-seq experiment. Apparently, when sequencing depth is higher and comparable across samples, e.g. 1×10^6 reads per sample, richness estimation improves (data not shown).

Evenness indices are affected as well by the sequencing depth, which hampers the distinction between absent and rare species. The choice of an order $q > 1$ limits the weight given to low-count species and consequently reduces the variation due to the sequencing bias. We suggest computing evenness considering the total number of assayed species (i.e. the number of rows of the count matrix) or, in case of different data sets, the union of all species assayed, so to guarantee comparability. The choice of the class of evenness indices to be used depends on the question under investigation: EF indices quantify the fraction of diversity accounted by the abundant species, whereas RLE indices highlight differences between species relative abundances.

For alpha diversity, a wide panel of indices is available. Differently from richness, diversity does not have an unambiguous definition, but describes a multifaceted phenomenon, whose components are accounted for differently by the available indices. In addition, some of the indices encode the same information, but with slightly different mathematical formulations, as in the case of Shannon entropy and ${}^1D$, or Gini–Simpson and ${}^2D$. Amongst these, the formulations that correspond to numbers equivalent have an easily interpretable unit of measurement — the number of species — and possess the doubling property. Similarly to evenness indices, Hill numbers computed for $q > 1$ are more robust to the sequencing bias. However, the choice of the order q should also depend on the weight one want to give to low-abundance species: the higher the order, the lower the influence of rare species on diversity computation. In addition, indices with high sensitivity to rare species, such as low-order Hill numbers and Tail index, should be used only when the sequencing depth is sufficient to distinguish the vanishingly rare species from the absent ones. Finally, we advise against using Gini–Simpson for data sets with >10 species: despite robust to the sequencing bias, it does not vary enough to capture differences between samples.

For beta diversity, several indices are also available. As discussed above, the sequencing bias hinders the distinction of rare and absent species and, consequently, renders the computation and comparison of lists of present species sensitive to noise. Therefore, beta diversity indices are strongly biased in case of low sequencing depth and species unevenness, which might be ameliorated using higher sequencing depth and *ad hoc* normalization procedures. If the focus of a study is the investigation of inter-sample microbial diversity in terms of fractions of shared and unique species, measures of continuity that also possess the homogeneity property return values that are easily interpretable and comparable across studies. For instance, β_{mn}

possesses these properties and, in our assessment, was among the beta diversity indices least affected by the sequencing bias.

To overcome their limitations, beta diversity measures could be complemented by multivariate analyses such as ordination [58, 59] to help to shed light on differences and similarities of the investigated samples. Ordination plots do not consider the lists of present species but the full distribution of species abundances and can be used to graphically represent ecological distances between samples.

Conclusions

In the present work, we reviewed a panel of widely used indices for alpha and beta diversity, described their mathematical formulation, purposes and properties, as well as characterized their behaviour and criticalities in dependence of data features. In particular, we assessed their robustness in the analysis of NGS data using a synthetic 16S-seq data set. The present work is not intended to provide a comprehensive ranking of the best performing approaches, that would have required a broader benchmarking, both in terms of tested indices and data sets, but is aimed at providing a simplified overview of the major mathematical tools for investigating the diversity of the human microbiota, leaving to the user the final choice of the most suited indices given the biological question under study.

The simulator and the code for computing the indices are available as R package, DiversitySeq, allowing possible extension of the benchmarking to other data sets. Moreover, the simulator can be a useful resource for the assessment of computational methods for metagenomic count data, such as differential abundance analysis [57, 60, 61].

We believe this work can assist the user with the selection and interpretation of diversity indices for the analysis of 16S-seq data, as well as of other count data sets with similar characteristics, such as from 5S rRNA gene sequencing or environmental metagenomics. More broadly, these indices are of interest for different applications where NGS counts are computed for a set of non-overlapping classes, like in the analysis of immune repertoires, where the relative counts represent the proportions of lymphocytes sharing a particular T-cell or B-cell receptor in a sequenced sample [62, 63].

Materials and methods

Simulation data for the assessment of alpha diversity indices

We generated two simulated data sets similarly to [22], simulating different communities of K individuals belonging to S species, with $K = 10^4$ and $R = 5:100$. In one simulation, all species are equally abundant, i.e. $p_i = 1/K$, whereas in the other one, half of the species are rare and account for only one individual, i.e. $p_i^{rare} = 1$, and half of the species are abundant and account for the remaining individual as $p_i^{abund} = \frac{2}{R} - \frac{1}{K}$.

We generated more realistic abundance data sets, considering communities with different number of species S and seven different evenness scenarios (Supplementary Figure S1). A Poisson distribution was used to simulate the sampling of K individuals per sample. We used $S = 1:2^{10}$ and $K = 10^4$ or 10^7.

Simulation data for the assessment of beta diversity indices

To test beta diversity indices, we performed three simulations, considering different values of total (N), shared (a) and unique (b and c) species. We tested the impact of the number of shared and unique species a by ranging a/N from 0 to 1, with unique-species ratio equal to 0, 0.1, 0.2, 0.3, 0.4, 0.5 or 1, keeping $N = 1000$ fixed. We tested the impact of b and c with a simulation where $c = 200$ was fixed, a was fixed and equal to 0 or 300, and b varied between 1 and 5000. Finally, we simulated the nested condition with $c = 0$, a/N ranging from 0 to 1, and varying N or a (with fixed $a = 20$ or $N = 1000$, alternatively).

Simulation of the 16S-seq data set

For the synthetic 16S-seq data set, we assumed a NB model to create a $n \times m$ matrix of counts. We modelled the biological variability with a gamma distribution and the technical variability with a Poisson distribution, as in [56]. In particular, from a vector of average relative abundances per species $\bar{\mu}_i$, we modelled the relative abundances across m samples as:

$$p_{ij} \sim Gamma\left(\bar{\mu}_i, \phi\bar{\mu}_i^2\right),$$

with $mean(p_{ij}) = \bar{\mu}_i$ and $var(p_{ij}) = \phi\bar{\mu}_i^2$. ϕ is the dispersion parameter that links the mean and the variance of the distribution. The final matrix of absolute abundances was computed as $\tau_{ij} = round(p_{ij} \cdot \sum_i \bar{\mu}_i)$.

The counts were modelled as:

$$k_{ij} \sim Poisson\left(p_{ij} \cdot \gamma_j\right),$$

where γ_j is the sequencing depth of the j-th sample.

The $\bar{\mu}_i$, ϕ, m, n and γ_j parameters for the simulation were estimated from real 16S-seq data of the HMP [12], as described in the following. The real matrix of counts over 43 140 OTUs and 187 stool samples was downloaded from http://hmpdacc.org/. OTUs with null counts on all samples were removed, preserving $n = 8048$ rows, and the $m = 20$ samples with the highest sequencing depths were selected.

Average counts per millions for each OTU of the matrix were estimated using the aveLogCpm function of the edgeR package [64], transformed into $\delta_i = round(CPM_i - min(CPM_i))$ and finally used as proxy of average relative abundances: $\bar{\mu}_i = \delta_i / \sum \delta_i$.

The dispersion parameter ϕ was assumed to be common to all OTUs and was estimated with the 'estimateGLMTagwiseDisp' function of edgeR, after count normalization via trimmed mean of M-values performed with the 'calcNormFactors' function. The overdispersion parameter ϕ of Supplementary Figure S4 was also estimated with the 'estimateGLMTagwiseDisp' function of edgeR [64]. The sequencing depths were computed as $\gamma_j = \sum_i k_{ij}$. The diversity indices were computed on the relative counts as follows:

$$p'_{ij} = \frac{k_{ij}}{\sum_i k_{ij}}.$$

The simulated 16S-seq data and the parameters and the code for the simulation are available in the R package DiversitySeq (http://sysbiobig.dei.unipd.it/?q=Software#DiversitySeq).

Assessment of relative error and variance

The relative error was computed as:

$$RE = (S_k - S_\tau)/S_\tau \cdot 100,$$

where S_τ is the richness estimated from abundance data $[\tau_{ij}]$ and S_k is the richness estimated from count data $[k_{ij}]$.

The relative variation was computed as:

$$RV = (D_k - D_\tau)/D_\tau \cdot 100,$$

where D_τ is the value of the diversity index applied to abundance data $[\tau_{ij}]$, and D_k is the value of the index applied to sequencing data $[k_{ij}]$.

Key Points

- Diversity indices provide valuable mathematical tools to describe the ecological complexity of the human microbiota profiled with targeted NGS.
- In the study of the human microbiota, alpha diversity describes the compositional complexity of a single sample, whereas beta diversity describes taxonomical differences between samples.
- Diversity indices have different purposes, properties, mathematical formulations, units of measures and robustness to the sequencing bias.
- The properties of the different diversity indices must be taken into account and evaluated depending on data characteristics and on the biological question under study.
- The R package DiversitySeq implements in a unified framework the panel of diversity indices reviewed in this work and a simulator of count data that can be used for the benchmarking of computational methods for microbiota count data.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

Publication of this article has been funded by PRAT 2010, grant number CPDA101217.

References

1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**:207–14.
2. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet* 2012;**13**:260–70.
3. Martínez I, Stegen JC, Maldonado-Gómez MX, *et al*. The gut microbiota of rural Papua new Guineans: composition, diversity patterns, and ecological processes. *Cell Rep* 2015;**11**: 527–38.
4. De Filippo C, Cavalieri D, Di Paola M, *et al*. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci USA* 2010;**107**:14691–6.
5. Kau AL, Ahern PP, Griffin NW, *et al*. Human nutrition, the gut microbiome and the immune system. *Nature* 2011;**474**: 327–36.
6. Qin J, Li Y, Cai Z, *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**:55–60.
7. Knip M, Siljander H. The role of the intestinal microbiota in type 1 diabetes mellitus. *Nat Rev Endocrinol* 2016;**12**:154–67.
8. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. *Genome Med* 2016;**8**:51.
9. Wu H, Tremaroli V, Bäckhed F. Linking microbiota to human diseases: a systems biology perspective. *Trends Endocrinol Metab TEM* 2015;**26**:758–70.
10. Garrett WS. Cancer and the microbiota. *Science* 2015;**348**:80–6.
11. Schwabe RF, Jobin C. The microbiome and cancer. *Nat Rev Cancer* 2013;**13**:800–12.
12. NIH HMP Working Group, Peterson J, Garges S, *et al*. The NIH Human Microbiome Project. *Genome Res* 2009;**19**:2317–23.
13. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* 2012;**486**:215–21.
14. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
15. Weinstock GM. Genomic approaches to studying the human microbiota. *Nature* 2012;**489**:250–6.
16. Schmidt TSB, Matias Rodrigues JF, von Mering C. Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput Biol* 2014;**10**:e1003594.
17. Kuczynski J, Lauber CL, Walters WA, *et al*. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet* 2012;**13**:47–58.
18. Whittaker RH. Vegetation of the Siskiyou mountains, Oregon and California. *Ecol Monogr* 1960;**30**:279–338.
19. Whittaker RH. Evolution and measurement of species diversity. *Taxon* 1972;213–51.
20. Tuomisto H. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography* 2010;**33**:2–22.
21. Jost L. Partitioning diversity into independent alpha and beta components. *Ecology* 2007;**88**:2427–39.
22. Jost L. The relation between evenness and diversity. *Diversity* 2010;**2**:207–32.
23. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987;783–91.
24. Smith EP, van Belle G. Nonparametric estimation of species richness. *Biometrics* 1984;119–29.
25. Chao A, Lee S-M. Estimating the number of classes via sample coverage. *J Am Stat Assoc* 1992;**87**:210–7.
26. O'Hara RB. Species richness estimators: how many species can dance on the head of a pin? *J Anim Ecol* 2005;**74**:375–86.
27. Hill MO. Diversity and evenness: a unifying notation and its consequences. *Ecology* 1973;**54**:427–32.
28. Berger WH, Parker FL. Diversity of planktonic foraminifera in deep-sea sediments. *Science* 1970;**168**:1345–7.
29. Rényi A. others. On measures of entropy and information. *Proc Fourth Berkeley Symp Math Stat Probab* 1961;**1**:547–61.
30. Simpson EH. Measurement of diversity. *Nature* **163**:688.
31. Shannon CEA. mathematical theory of communication. *ACM SIGMOBILE Mob Comput Commun Rev* 2001;**5**:3–55.
32. Li K, Bihan M, Yooseph S, *et al*. Analyses of the microbial diversity across the human microbiome. *PloS One* 2012;**7**: e32118.
33. Pielou EC. The measurement of diversity in different types of biological collections. *J Theor Biol* 1966;**13**:131–44.
34. Oksanen J, Blanchet FG, Kindt R, *et al*. *Vegan: Community Ecology Package*, 2015. R package version 2.4-1. https://CRAN.R-project.org/package=vegan
35. Wittebolle L, Marzorati M, Clement L, *et al*. Initial community evenness favours functionality under selective stress. *Nature* 2009;**458**:623–6.
36. Koleff P, Gaston KJ, Lennon JJ. Measuring beta diversity for presence–absence data. *J Anim Ecol* 2003;**72**:367–82.
37. Magurran A. Ecological diversity and its measurement Croom Helm London. 1988;179. p.
38. Southwood TRE, Henderson PA. Ecological methods. 2009.
39. Harrison S, Ross SJ, Lawton JH. Beta diversity on geographic gradients in Britain. *J Anim Ecol* 1992;151–8.
40. Wilson M, Shmida A. Measuring beta diversity with presence-absence data. *J Ecol* 1984;1055–64.

41. Mourelle C, Ezcurra E. Differentiation diversity of Argentine cacti and its relationship to environmental factors. *J Veg Sci* 1997;**8**:547–58.
42. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol Skr* 1948;**5**:1–34.
43. Harte J, Kinzig AP. On the implications of species-area relationships for endemism, spatial turnover, and food web patterns. *Oikos* 1997;417–27.
44. Lande R. Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* 1996;5–13.
45. Weiher E, Boylen CW. Patterns and prediction of α and β diversity of aquatic plants in Adirondack (New York) lakes. *Can J Bot* 1994;**72**:1797–804.
46. Cody ML. Towards a theory of continental species diversities: bird distributions over Mediterranean habitat gradients. *Ecol Evol Communities* 1975;**214**:257.
47. Colwell RK, Coddington JA. Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond B Biol Sci* 1994;**345**:101–18.
48. Pielou EC. The interpretation of ecological data: a primer on classification and ordination. 1984;
49. Gaston K, Rodrigues A, Van Rensburg B, *et al*. Complementary representation and zones of ecological transition. *Ecol Lett* 2001;**4**:4–9.
50. Jaccard P. The distribution of the flora in the alpine zone. *New Phytol* 1912;**11**:37–50.
51. Cody ML. Bird diversity components within and between habitats in Australia. *Species Divers Ecol Communities* 1993; 147–58.
52. Williams PH. Mapping variations in the strength and breadth of biogeographic transition zones using species turnover. *Proc R Soc Lond B Biol Sci* 1996;**263**:579–88.
53. Williams PH, Klerk HM, Crowe TM. Interpreting biogeographical boundaries among Afrotropical birds: spatial patterns in richness gradients and species replacement. *J Biogeogr* 1999; **26**:459–74.
54. Routledge R. On Whittaker's components of diversity. *Ecology* 1977;1120–7.
55. Lennon JJ, Koleff P, Greenwood J, *et al*. The geographical structure of British bird distributions: diversity, spatial turnover and scale. *J Anim Ecol* 2001;**70**:966–79.
56. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* 2015;**14**:130–42.
57. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014;**10**: e1003531.
58. Ramette A. Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol* 2007;**62**:142–60.
59. Kuczynski J, Liu Z, Lozupone C, *et al*. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods* 2010;**7**:813–9.
60. Paulson JN, Stine OC, Bravo HC, *et al*. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 2013;**10**:1200–2.
61. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 2009;**5**:e1000352.
62. Ruggiero E, Nicolay JP, Fronza R, *et al*. High-resolution analysis of the human T-cell receptor repertoire. *Nat Commun* 2015;**6**:8081.
63. Greiff V, Miho E, Menzel U, *et al*. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends Immunol* 2015;**36**:738–49.
64. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma Oxf Engl* 2010;**26**:139–40.