environmental microbiology

Environmental Microbiology (2013) **15**(10), 2642–2657 doi:10.1111/1462-2920.12156

Minireview

A unifying quantitative framework for exploring the multiple facets of microbial biodiversity across diverse scales

Arthur Escalas,[1]* Thierry Bouvier,[1] Maud A. Mouchet,[2] Fabien Leprieur,[1] Corinne Bouvier,[1] Marc Troussellier[1] and David Mouillot[1]

[1]*UMR 5119 CNRS-UM2-UM1-IRD-Ifremer Ecologie des systèmes marins côtiers, Université Montpellier 2 cc 093, 34 095 Montpellier Cedex 5, France.*

[2]*UMR 5553 LECA, BP 53, 2233 Rue de la Piscine, 38041 Grenoble Cedex 9, France.*

Summary

Recent developments of molecular tools have revolutionized our knowledge of microbial biodiversity by allowing detailed exploration of its different facets and generating unprecedented amount of data. One key issue with such large datasets is the development of diversity measures that cope with different data outputs and allow comparison of biodiversity across different scales. Diversity has indeed three components: local (α), regional (γ) and the overall difference between local communities (β). Current measures of microbial diversity, derived from several approaches, provide complementary but different views. They only capture the β component of diversity, compare communities in a pairwise way, consider all species as equivalent or lack a mathematically explicit relationship among the α, β and γ components. We propose a unified quantitative framework based on the Rao quadratic entropy, to obtain an additive decomposition of diversity ($\gamma = \alpha + \beta$), so the three components can be compared, and that integrate the relationship (phylogenetic or functional) among Microbial Diversity Units that compose a microbial community. We show how this framework is adapted to all types of molecular data, and we highlight crucial issues in microbial ecology that would benefit from this framework and propose ready-to-use R-functions to easily set up our approach.

Received 22 October, 2012; revised 5 May, 2013; accepted 7 May, 2013. *For correspondence. E-mail arthur.escalas@univ-montp2.fr; Tel. (+33) 4 67 14 92 27; Fax (+33) 4 67 14 37 19.

Introduction

Through their important biomass and their diversified metabolic abilities, micro-organisms play a key role in the regulation of ecological processes, such as organic matter degradation, in all ecosystems (Sleator *et al.*, 2008). However, the diversity of micro-organisms, representing a large portion of all biological diversity, has been only recently and partially described thanks to major advances in molecular methods, which is particularly true for bacteria and microbial eukaryotes (Fierer and Lennon, 2011). Indeed, the use of nucleic acid-based analyses has revealed that microbial diversity levels have been underestimated by several orders of magnitude over the past few decades (Ward *et al.*, 1990; Rappe and Giovannoni, 2003; DeSantis *et al.*, 2005). By improving our capacity to assess microbial community composition (i.e. the number and identity of taxonomical units) and structure (i.e. the abundance distribution among these units as well as their phylogenetic relatedness), the era of molecular microbiology already does and will even more provide a better understanding the ecological processes that shape these communities and underpin ecosystem functioning (Bell *et al.*, 2005). To gain from this unprecedented mass of information, we need to unify the calculations of different diversity indices while considering the wide range of available and forthcoming data from genes to functions. Moreover, diversity indices should be able to quantify the biodiversity of microbial communities and its variations across spatial and temporal scales, and along environmental gradients (Christen, 2008).

Quantifying the diversity of ecological communities has become a multifaceted issue for all groups of organisms including microbes (Devictor *et al.*, 2010). Species belonging to a community can differ in their abundance, their taxonomic affiliation, their phylogenetic relatedness along with their ecological functions (Bryant *et al.*, 2008;

Peter *et al*., 2011). These last three facets of biodiversity were coined as taxonomic, phylogenetic and functional diversity respectively. Beyond the simple assessment of community composition, in terms of microbial taxa, species or OTU (operational taxonomic unit) richness, the phylogenetic structure sheds light on evolutionary constraints and historical contingencies shaping microbial communities, while their functional structure indubitably relates to ecosystem functioning (Salles *et al*., 2009; Bissett *et al*., 2011; Peter *et al*., 2011; Pommier *et al*., 2012). In other words, while taxonomic diversity is the facet of diversity that provides information about how many and which microbes are present in the community, phylogenetic diversity informs about their evolutionary history, and functional diversity quantifies the breadth of roles that they play. However, the current indices used to estimate the level of microbial diversity are mainly restricted to community composition and rarely embrace the different facets of community structure by including phylogenetic and functional information along with the distribution of abundances among lineages or functional groups (Lozupone *et al*., 2007; Bryant *et al*., 2008). This is even more critical as environmental filtering may operate at the level of lineages instead of isolated species, and specific microbial processes may rely on a single phylotype or functional group (Bryant *et al*., 2008; Peter *et al*., 2011). This could be the case, for example, if a niche-based process selects lineages or species according to biological traits associated with specific phylogenetic groups (Pommier *et al*., 2012).

The importance of measuring the different facets of biodiversity (taxonomic, phylogenetic and functional) while taking into account scale effects is fundamental when quantifying microbial diversity with a biogeographical perspective or tackling macroecological issues (Odonnell *et al*., 1994; Christen, 2008; Fierer and Lennon, 2011). This could benefit the development of theories and hypotheses in regards to the factors that structure microbial communities, their response to environmental pressures and the connections between diversity and function (Kembel, 2009; Stegen *et al*., 2012). Surprisingly, only few microbiological studies have used the historical Whittaker's biogeographical framework (Whittaker, 1960; 1972), along with its biodiversity decomposition into local (α), inter-sites (β) and regional (γ) components (Griffiths *et al*., 2011). Here, the term 'decomposition of biodiversity' refers to the idea that there are different components (or levels) that, together, constitute the biodiversity at a certain scale. This approach is for instance particularly useful in biogeography where one wants to estimate the biodiversity across a landscape composed of different localities inhabited by different biological communities. Hence, the biodiversity estimation in each locality corresponds to the α level, the difference between these localities is the β level and the overall biodiversity across all localities is the γ level.

The estimation of these biodiversity components can be achieved either with indices related to community composition (species richness) or with more sophisticated indices embracing the different facets of community structure. Recently, Lozupone and Knight (2008) plainly reviewed the indices designed to estimate separately the α and β components of the Whittaker's decomposition for different types of data. While the use of these indices led to valuable insights in microbial ecology and biogeography such as the existence of universal patterns (e.g. taxa–area relationship and community assembly rules; Fuhrman, 2009), in our view, they exhibit several limitations. Indeed, former classical indices dedicated to the estimation of community differences (e.g. Jaccard, Sorensen and Bray–Curtis) only capture the β component of diversity, leaving apart the α and γ components. They also often compare communities in a pairwise way and consider all species as equivalent. Similarly, recent indices including species abundance distributions, phylogenetic or functional differences between species (e.g. Unifrac, VAW-UniFrac and βMNTD) only account for the β component of diversity (Lozupone *et al*., 2007; Kembel, 2009; Chang *et al*., 2011; Stegen *et al*., 2012). This can be a limiting view of biodiversity across scales as estimating differences between communities (β-diversity) does not provide any information about the biodiversity of local communities (α) and of the whole system (γ). For instance, for a given level of β-diversity estimated between two communities, using the Bray–Curtis or the Unifrac dissimilarity index, these two communities can have either a low or a high level of α diversity, and these two scenarios cannot be discriminated. In this case, the estimation of biodiversity is only partial and would require the comparison of the β-diversity value to α and γ-values to fully assess the biodiversity across the system. Moreover, the α, β and γ components are often estimated using different and independent indices, which prevent any mathematically explicit relationship among them and thus any comparison, contrary to the Whittaker's framework. Finally, no consensus has emerged about how and in which cases these indices should be used according to the earlier-mentioned limitations. We, thus, urgently need to improve our way of analysing the different facets of microbial diversity across scales through a unified and flexible framework that allows us (i) to estimate β-diversity for a set of two or more communities, (ii) to estimate simultaneously the three components using a similar unit, (iii) to compare the three components using a decomposition of γ diversity and (iv) to integrate data of different nature (e.g. abundance, taxonomy, phylogeny and function). New indices recently developed for biogeography and community analyses of macro-organisms may fulfil

this gap and may successfully be applied to the microbial world (Allen *et al*., 2009; Ricotta and Szeidl, 2009; Devictor *et al*., 2010).

Here we aim to (i) briefly review the different kinds of methods used to assess the different facets of biodiversity in microbial communities and classify the generated data; (ii) propose a unified, flexible and multifaceted framework to estimate microbial diversity based on taxonomic, phylogenetic or functional data and across temporal and spatial scales; (iii) present some possible applications of this framework in microbial ecology and finally (iv) provide appropriate ready-to-use resources for organizing and analysing multiple microbiological data.

Measuring the diversity of microbial communities: different methods for different data

We first present the available methods to study the different facets of microbial biodiversity along with their advantages, disadvantages, some key considerations for their application and the associated data (Table 1).

Methods for studying the different facets of biodiversity in microbial communities

The presented methods were selected based on their potential to be used for extensive biodiversity studies. They also need to be well established in the literature, independent of cultivation, allow the characterization of microbes' biodiversity while focusing on the community level, and allow the analysis and the comparison of a large number of samples in a standardized and reproducible way. Moreover, in order to be used with the proposed framework, these methods should provide a table output depicting the microbial composition (in rows) of the studied communities (in columns). Based on these criteria, we set aside several methods used to assess microbial diversity such as the family of fluorescent *in situ* hybridization methods because their sample size limitation and their low resolution or the community-level physiological profile methods such as Biolog Ecoplates because they are not culture-independent methods and are biased by inoculums variability (Kirk *et al*., 2004). Hence, we focus on high-throughput nucleic acids-based methods even if they have well-known limitations such as nucleic acid extraction step (Petric *et al*., 2011). The following section briefly reviews some key characteristics of the chosen methods, along with their advantages and disadvantages.

One of the most common approaches to study the diversity of microbial communities is the use of ribosomal RNA (rRNA) gene analysis using fingerprinting methods. These fingerprint techniques are based on electrophoretic separation of PCR products (or amplicons) amplified from nucleic acid extracted from a sample (Nocker *et al*., 2007). The separation step can be performed by gel electrophoresis, chromatographic or capillary electrophoresis, and is based on amplicon length (T-RFLP, ARISA and LH-PCR) or nucleotide composition (SSCP, DGGE and DHPLC). For each community (i.e. sample), the resulting output is a profile specific of the studied community with respect to migration distance and relative intensity of band or peak, which refer theoretically to a unique sequence (Loisel *et al*., 2006). All fingerprints are affected by the same nucleic acid extraction and PCR biases (unspecific amplification, generation of chimeric sequences, formation of heteroduplexes and nucleotide misincorporation), and are inherently limited in the maximum number of microbial units detected (Von Wintzingerode *et al*., 1997; Loisel *et al*., 2006). Moreover, the diversity estimation is biased by migration issues leading to comigration of multiple amplicons (up to 15) under one band or peak, or the formation of several bands or peaks for one amplicon (Kisand and Wikner, 2003). For detailed comparisons of fingerprint methods, see elsewhere (Kirk *et al*., 2004; Nocker *et al*., 2007).

During the last decade, the development of metagenomic tools with even higher resolution has revolutionized the description of biodiversity in microbial communities. The metagenomic term refers to the culture-independent analysis of complete genomes of microbial communities, directly isolated from an environmental sample (water, soil and air) or living on plants or animal hosts (Sleator *et al*., 2008; Petrosino *et al*., 2009; Metzker, 2010). Metagenomic methods considered here (sequencing and microarrays) offer the highest throughput and resolution for microbial diversity assessment and could avoid biases introduced during PCR amplification of marker genes (von Mering *et al*., 2007).

Historically, the most standard metagenomic approach is gene-based cloning and sequencing, which involves the cloning and sequencing of amplicons using the Sanger method (Suenaga, 2012). This chain-termination method (Sanger *et al*., 1977), improved and still in use, produces high-quality reads (sequences) up to 1000 bp and has a wider coverage of the targeted sequences and a better resolution than other sequencing approaches (Xiong *et al*., 2010). In counterpart, the cloning step can be time consuming and laborious, reducing the sample throughput or leading to arbitrary loss of genomic DNA (Metzker, 2010; Zinger *et al*., 2012). The use of direct pyrosequencing of amplified fragments can avoid the cloning step, but the length of the reads is shorter with this method, which makes the process of genome assembly more difficult (Fierer and Lennon, 2011). Moreover, gene-based sequencing (with cloning or not) is restricted to PCR-amplified sequences and so is affected by PCR biases. Overall, the limitation here lies in the reliability of

Table 1. Summary of the most widely used molecular methods to assess the diversity of microbial communities and their associated data.

	Method characteristics							Nature of Microbial Diversity Units (MDUs)				
Method type	Diversity source	Amplification	Output	Sample throughput	Cost	Advantages	Disadvantages	Nucleic acid fragments	Taxa	Phylogenetic groups	Functions	Relative abundance
Nucleic acid-based fingerprints	Group-specific gene Functional encoding gene	PCR	Band/peak pattern	Medium-high	+	Quick and easy to set up	Limited resolution Restricted sampling effort	✓	(✓)	(✓)	(✓)	(✓)
Gene-based cloning and sequencing	Group-specific gene	PCR	Sequences	Low	+++	Adaptable sampling effort The highest resolution and coverage (reads >1000 bp)	Cloning step and associated costs and quantitative biases Focus on a specific sequence Reliance on existing databases		✓	✓	(✓)	(✓)
Shotgun cloning and sequencing	Metagenome	New amplification methods No amplification	Sequences	Low	+++	Adaptable sampling effort Long reads Whole metagenome screening	Cloning step and associated costs and quantitative biases Reliance on existing databases		✓	✓	✓	✓
Next generation sequencing	Metagenome	New amplification methods No amplification	Sequences	High	+++	Adaptable sampling effort No cloning step Whole metagenome screening	Potential diversity overestimation Reliance on existing database Short reads Sample throughput limited by the cost		✓	✓	✓	✓
Microarrays	Group-specific gene Functional encoding gene	New amplification methods PCR No amplification	Array image	High	++++	Adaptable and standardized sampling effort Whole metagenome screening	Only predefined genes can be detected Reliance on existing databases Sample throughput limited by the cost		✓	✓	✓	✓

Ticks in brackets correspond to method–data associations that are not unanimously accepted in the literature. See text section 1 for further details on the methods and data type, definition of 'group-specific gene', 'functional encoding gene' and the corresponding references.

relative abundance distributions for detected sequences because the extraction, amplification and cloning steps of nucleic acid may introduce quantitative biases. These gene-based sequencing approaches are progressively being replaced by whole metagenomic sequencing, which produces reads from the whole metagenome sequences and could benefit from the development of new amplification methods such as whole genome amplification or emulsion PCR (Shendure and Ji, 2008; Petrosino *et al*., 2009). Whole metagenomic DNA can be randomly sheared before cloning and sequencing, this is the case in the shotgun approach, or can be directly sequenced using next generation sequencing (NGS) methods (Petrosino *et al*., 2009; Roh *et al*., 2010); for a detailed comparison of NGS methods see elsewhere (Shendure and Ji, 2008; Metzker, 2010). These whole metagenomic sequencing approaches generate thousands of short sequences (reads) that can be assembled in longer sequences using bioinformatic automated pipelines (Hirsch *et al*., 2010; Santamaria *et al*., 2012). Recent developments in high-throughput sequencing are limited by the computing step, but rapid improvements are underway (von Mering *et al*., 2007; Metzker, 2010).

Another recently improved metagenomic approach is the microarray, which are glass slides on which DNA fragments are spotted and serve as probes for the hybridization of the labelled metagenomic DNA. Next, fluorescent label intensity is measured, which reveals the presence of hybridized DNA on the slides (Zhou, 2003). Microarrays overcome some of the restrictions of other methods such as the low resolution of fingerprint methods, the laborious cloning step of some sequencing approaches and avoids the amplification step by direct hybridization of metagenomic DNA on the slides (Roh *et al*., 2010). However, microarrays only detect already sequenced genes as the probes spotted on the slides are designed using reference databases (Zhou, 2003; Hirsch *et al*., 2010).

For detailed comparisons of metagenomic approaches see Metzker (2010), Roh and colleagues (2010) and Su and colleagues (2012).

Markers to assess the biodiversity of microbial communities

Basically there are two categories of marker genes, group-specific genes and functional encoding genes (Stahl, 2007). Group-specific genes (taxonomic or phylogenetic) correspond to conserved biopolymers that can be used to infer taxonomic or phylogenetic relationships among the organisms. Functional encoding genes code for specific proteins and could be used to evaluate specific chemical transformations or potential activity of microbial populations (Stahl, 2007).

The taxonomic and phylogenetic relationships between prokaryotes can be deduced from sequence comparisons of conserved group-specific genes. To do so, these genes should be widely distributed, should not be frequently transmitted horizontally and should be present as a single copy. In addition, they should not be too long for being easily amplified and sequenced, but not too short in order to contain enough information. Moreover, they should have a 'good' level of resolution, that is they should not be too conserved nor too variable (Gevers and Coenye, 2007). The group-specific markers that are the most commonly used in microbial ecology are the rRNA genes because they are universally present, functionally constant and are composed of conserved and more variable domains (Vos *et al*., 2012). Modern microbial taxonomy and phylogeny benefit from the PCR sequencing of the genes coding for the small (16S) or large (23S) rRNA molecules (Gevers and Coenye, 2007; Stahl, 2007). However, other genes had been used as group-specific marker genes to delineate relationships between microorganisms such as *recA*, *gyrB*, *rpoB*, *rpoD*, *hsp60*, *soda*, *atpD* and *infB* (Gevers and Coenye, 2007; Bonilla-Rosso *et al*., 2012; Vos *et al*., 2012).

Functional encoding genes have been used as markers for studying the diversity of microbial functions in different nucleic acid-based approaches such as fingerprints (Hirsch *et al*., 2010), microarrays (He *et al*., 2008) or various sequencing approaches (Dinsdale *et al*., 2008; Carvalhais *et al*., 2012).

In any of the earlier presented nucleic acid-based methods, the target molecule should be either DNA or RNA, the latter being the transcribed version of the former, and the use of reverse transcription is required to generate cDNA from RNA (Stahl, 2007). In an ecological perspective, DNA and RNA targets provide different views of the microbial communities, which correspond, respectively, to metagenomic and metatranscriptomic approaches. When using a group-specific marker and its corresponding DNA and RNA versions, 16S rDNA and 16S rRNA for instance, we can differentiate present versus active bacteria respectively (Gremion *et al*., 2003). In the case of functional encoding genes, one can differentiate present (potential) versus expressed (realized) functions depending on the use of DNA or mRNA respectively (Sleator *et al*., 2008; Roh *et al*., 2010). Note that total environmental RNA extracted from microbial communities is mainly composed of rRNA and transfer RNA, with approximately 1–5% mRNA (Carvalhais *et al*., 2012).

Data generated and associated microbial biodiversity facets

The data outputs of selected methods are represented in Fig. 1. In the rest of the article, we will use the generic

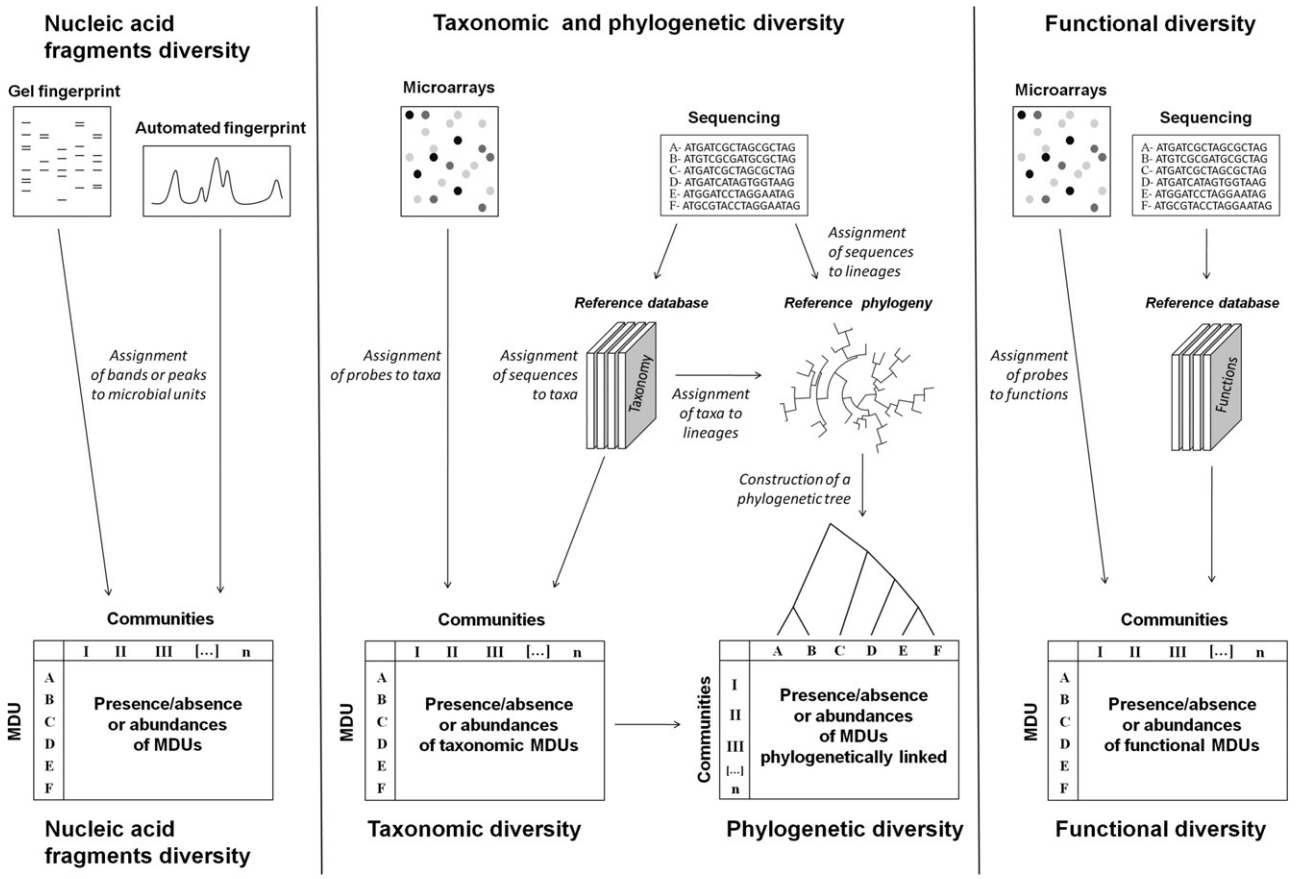

Fig. 1. The different ways to obtain data for the different facets of microbial biodiversity using the methods presented in Table 1. See text and references therein for a detailed description of the different steps presented here.

term Microbial Diversity Unit (MDU) to refer to the different diversity units that compose a microbial community. These MDUs can correspond to different facets of biodiversity: the diversity of nucleic acid fragments, taxa, phylogenetic lineages or functions depending on the method and the genetic marker.

The data obtained using fingerprint methods can be represented as a presence–absence matrix for each MDU in each sample (Fig. 1). Although there is still an ongoing debate about how to name the units detected by fingerprints (e.g. OTU, biotype, ribotype, genotype, phylotype and ribosomal genotypes), all these terms refer to the same entities. They are in fact nucleic acid fragments (amplicons) that are discriminated in various ways, which depend on the method, to give a snapshot of the complexity of the microbial community. Consequently, in the case of fingerprints, the generic term MDU will refer to these entities, and the associated diversity will be called 'fingerprint's nucleic acid fragments diversity'. Considering the earlier-mentioned PCR biases, there is still a debate on the possibility to use band intensities, peak heights or areas as relative abundances of MDUs (Bent *et al.*, 2007). Taking into account the limited resolution of the method and the earlier definition of MDU, it is important to note that these MDUs may not correspond to any taxonomic or phylogenetic group. Hence, the resolution of fingerprints methods is limited as they do not provide any clues about which microbes compose the community. However, it is possible to collect and sequence the MDUs (e.g. bands in DGGE) to know their taxonomic affiliation. This allowed, in the context of many experiment, the identification of the organism of interest, but this approach remains complicated in practice and is still considered limited (Zinger *et al.*, 2012). The amplified fragment may also correspond to a functional encoding gene (*mer*, *amoA*, *nifH*, *nozZ*, *mcrA*, etc.). In such a case, what is estimated is the diversity of the gene encoding the function within the community but not the whole diversity of the community functions because fingerprints can analyse only one gene at a time (Stahl, 2007; Hirsch *et al.*, 2010).

Measuring the taxonomic diversity of micro-organisms requires their identification and classification into the different levels of taxonomic hierarchy such as genus, family, order and phylum (Odonnell *et al.*, 1994; Santamaria *et al.*, 2012). This approach is still widely used to study microbial diversity, but affiliation of microbes into

taxonomic levels is no longer done on the basis of their phenotypic characteristics but rather using their genetic similarity with known taxa (Huse *et al*., 2008). This similarity is estimated using nucleotide sequences of group-specific marker genes (usually the 16S rRNA gene) present in the community metagenome (Odonnell *et al*., 1994). The sequences can be obtained using earlier presented sequencing approaches and can then be compared with reference databases to obtain their taxonomic affiliation (Christen, 2008; Santamaria *et al*., 2012; Fig. 1). The description of this step (called binning) is beyond the scope of this work (but see Kunin *et al*., 2008; Santamaria *et al*., 2012 for further details). Although there is some quantitative biases associated with the nucleic acids extraction, PCR and cloning steps, metagenomics approaches provide the relative abundances of detected MDUs (i.e. taxa or OTUs).

The taxonomic diversity of microbial communities can be also estimated using microarrays designed with probes corresponding to group-specific marker genes of different microbial taxa. For instance, the PhyloChip G3 is a microarray able to detect more than 60 000 different MDUs (i.e. taxa) (Kellogg *et al*., 2012). The probes hybridized on the array provide the MDU composition of the community, and their hybridization intensities can be linked to the relative abundances of MDUs to fully assess the quantitative structure of microbial communities, i.e. the abundance distribution among MDUs (Fig. 1; DeSantis *et al*., 2005; Handley *et al*., 2012). The resulting data are a table with detected MDUs in rows and the studied samples in column (Fig. 1).

The investigation of phylogenetic diversity of microbial communities has increased since the development of metagenomic methods, providing deeper insight into the processes that influence their composition and structure (Martin, 2002; Christen, 2008). To assess the phylogenetic diversity of microbial communities, one needs to know the phylogenetic relatedness between the microbes present in the community. A first way is to determine the taxonomic diversity of the community using the sequencing approaches described previously and then to assign the identified MDUs to lineages of a reference phylogenetic tree (Petrosino *et al*., 2009; Liggenstoffer *et al*., 2010). Phylogenetic relatedness between MDUs can also be obtained by comparing directly their respective sequences with referenced sequences (Kembel *et al*., 2011). As in the case of taxonomic MDUs, these approaches provide the relative abundances of the phylogenetic MDUs (i.e. the leaves of the phylogenetic tree).

The phylogenetic diversity of microbial communities can also be estimated using microarrays. Indeed, the list of positive hybridized probes (positive MDUs) can be used to prune the phylogenetic tree relating the entire spotted probes on the array. This gives the tree relating only the MDUs composing the studied community (Holmes *et al*., 2010).

In both cases (sequencing and microarrays), the resulting dataset is a table containing the detected MDUs and the studied samples, associated with a phylogenetic tree or a phylogenetic distance matrix depicting the relatedness among MDUs (Fig. 1).

The functional diversity of micro-organisms is now widely considered as the biodiversity component underpinning ecosystem functioning (Christen, 2008). Its estimation drastically differs depending on the size of the studied organisms. Indeed, for macro-organisms (e.g. plants, fishes and birds), functions performed by each species or populations are usually approximated by measuring functional traits on individuals and finally calculating the functional diversity of communities (Mendez *et al*., 2012; Villéger *et al*., 2012). Even if this approach is possible for some groups of micro-organisms such as zooplankton or flagellates, it becomes more difficult to set up as the size of studied organisms decreases (Barnett *et al*., 2007; Kruk *et al*., 2010). When communities are composed of bacterial, archaeal or microbial eukaryotes populations, this species-centred approach is currently impossible because we cannot separate all the populations that compose a community to measure their specific functional traits. In the near future, the development of technologies such as MicroFISH, NanoSIMS and flow cytometry may allow assigning simultaneously both identity and functions to specific microbes forming natural communities (Amann and Fuchs, 2008). However, today, the functional diversity of certain groups of micro-organisms such as bacteria, archaea and microbial eukaryotes is only assessed at the community level using nucleic acid-based methods (Dinsdale *et al*., 2008; He *et al*., 2008; Yavitt *et al*., 2012). Metagenomic data provides extensive information about gene content and their potential functions, and metatranscriptomic assesses what genes may be expressed. Whatever the approach, the functional data are obtained as lists of functional encoding genes associated with the whole community (Fig. 1). These genes can, however, be treated as discrete functional units (MDUs) composing the functional diversity of microbial communities, exactly the same way as taxa compose their taxonomic diversity. This diversity of functions can be obtained using whole metagenomic sequencing approaches by comparing the obtained sequences to databases containing sequences of functional genes (Dinsdale *et al*., 2008; Prakash and Taylor, 2012). As it is the case for taxa or phylogenetic groups, relative abundances of these functional MDUs can be extracted, providing a full quantitative assessment of functional community structure. Similarly, using probes corresponding to sequences of functional genes,

microarrays can provide the functional diversity of microbial communities along with the relative abundances of detected functions (He *et al*., 2008).

The resulting dataset is a table containing the detected MDUs in rows and the studied samples in column (Fig. 1). In the near future, we may use functional data in the same way as we use phylogenetic data, as it is done for macro-organisms. Then, combining a taxonomic MDU composition table and a functional matrix depicting the functions performed by each MDU, we will be able to assess microbial functional diversity using the tools primarily developed for macro-organisms (Mouchet *et al*., 2010).

Decomposing the biodiversity into α, β and γ components

Biodiversity across scales

Biodiversity is classically decomposed across temporal and spatial scales into three levels considered as components: (i) local diversity (α), (ii) regional diversity (γ) and (iii) the difference among local communities (β). β-diversity is also referred as differentiation diversity and turnover (Whittaker, 1960; Vellend, 2001; Jurasinski *et al*., 2009; Anderson *et al*., 2011). Although these last two terms are often synonymous in the literature, they actually apply to different concepts (Tuomisto, 2010a,b; Anderson *et al*., 2011). 'Differentiation diversity' refers to the variation in community structure regardless of any external gradient and often estimated using (dis)similarity or distance estimators (Jurasinski *et al*., 2009; Anderson *et al*., 2011). 'Turnover' can be defined as a directional (along a gradient) pairwise estimation of change in community structure (according to Jurasinski *et al*., 2009 and Anderson *et al*., 2011, but criticized in Tuomisto, 2010a,b). To prevent ambiguity, we use the generic term β-diversity throughout the paper to refer to between community diversity based on an additive partitioning of biodiversity components.

Since Whittaker's seminal works on β-diversity (Whittaker, 1960; 1972), the number of proposed indices to quantify the three components of diversity has drastically increased (Koleff *et al*., 2003; Tuomisto, 2010a,b; Anderson *et al*., 2011). β-diversity indices can be divided into two classes whether they are based on a dissimilarity metric or deduced from the decomposition of diversity into α, β and γ components. In the first case, β-diversity is simply estimated as a pairwise intercommunity distance using a chosen dissimilarity metric (e.g. Sorensen, Jaccard or Bray–Curtis dissimilarity). This is achieved regardless of α and γ components (Koleff *et al*., 2003; Zinger *et al*., 2012). An important limitation of this dissimilarity-based approach of β-diversity is that the same intercommunity dissimilarity value (e.g. calculated with the Bray–Curtis index) may be obtained between two pairs of communities, which have different local diversity values (estimated with another index). In the second case, diversity is decomposed into α, β and γ components, all being related within an additive, $\beta = \gamma - \bar{\alpha}$ or a multiplicative framework, $\beta = \gamma/\bar{\alpha}$, in which $\bar{\alpha}$ corresponds to the mean local diversity across samples (Whittaker, 1960).

Additive decomposition of the Rao quadratic entropy

The Rao quadratic entropy (Q) is a measure of diversity that combines species-relative abundances and pairwise interspecies differences. By combining these two features of diversity, this index measures the community structure rather than its composition. This approach therefore complements the classic estimation of diversity using indices based on species richness, evenness or community composition (i.e. basically who is present in the community). In the context of microbial ecology, species can be replaced by any MDUs, such as phylotypes, OTU, taxa, species or functional genes, according to the method used (Table 1).

Here, we propose to use additive partitioning of the Rao quadratic entropy, which has several valuable properties in comparison to independent α, β and γ diversity estimations (Rao, 1982; Ricotta and Szeidl, 2009). Additive partitioning has the advantage, over its multiplicative counterpart, to express the three components (α, β and γ) in the same unit (phylotype, taxa, OTU, functional genes, microbial unit, etc.) so they can be compared directly (Lande, 1996). Another advantage of this framework is that α-diversity values do not influence the calculation of β-diversity values (Jost, 2007; 2010). The additive property also enables the calculation of the relative contributions of α- and β-diversity to the γ-diversity and in doing so, to compare their values among multiple scales and studies (Lande, 1996). Finally, using this framework, β-diversity can be estimated globally for a set of communities or between pairs of communities.

At the local scale, Rao quadratic entropy Q_α represents the expected dissimilarity between two randomly chosen MDUs from a sampled community and hence can be defined as the extent of dissimilarity between MDUs in a community (e.g. the phylogenetic distance between taxa in a community):

$$Q_\alpha = \sum_{i=1}^{s} \sum_{j=1}^{s} d_{ij} p_i p_j \quad (1)$$

Where d_{ij} is the distance (taxonomic, phylogenetic or functional) between the i-th and the j-th MDUs in the local community; distances need to be ultrametric to ensure the monotonicity of the Q with the richness (Pavoine *et al*., 2005). In an ultrametric tree, the branch lengths are scaled in a way that all distances from the root to the tips (or leaves) of the tree (MDUs in our case) are the same

(Vellend *et al.*, 2010). When these distances are unknown, d_{ij} can be set to unity. p_i and p_j are the relative local abundances of the *i*-th and the *j*-th MDUs respectively; p_i and p_j can be set equal for presence–absence data. *s* is the number of MDUs in the local community.

At the regional scale γ, sampled communities are pooled together into a single regional community. The Rao quadratic entropy at this regional scale Q_γ can be defined as the extent of dissimilarity between two randomly chosen MDUs in the regional community.

$$Q_\gamma = \sum_{i=1}^{S}\sum_{j=1}^{S} d_{ij}P_iP_j \quad (2)$$

Where d_{ij} is the distance (taxonomic, phylogenetic or functional) between the *i*-th and the *j*-th MDUs in the regional community. P_i and P_j are the relative regional abundances of the *i*-th and the *j*-th MDUs respectively. *S* is the number of MDUs in the regional community. The relative regional abundances are commonly quantified as the mean relative abundances over local communities for MDUs. Likewise, the quantification of local diversity, d_{ij}, P_i and P_j, can be set to a unique value when information is missing, e.g. $d_{ij} = 1$ or $P_i = P_j = 1/S$.

The mean intracommunity ($\bar{\alpha}$) quadratic entropy $Q_{\bar{\alpha}}$ is simply the mean of the local quadratic entropy values across the *k*th studied communities. The local quadratic entropy (Q_α) can be weighted by a parameter w_k or not (see de Bello *et al.*, 2010 for more details). For instance, this parameter could correspond to local community abundances:

$$Q_{\bar{\alpha}} = \sum_{n=1}^{N} w_k Q_{\alpha_k} \quad (3)$$

Subtracting the regional quadratic entropy (Q_γ) and the mean local quadratic entropy ($Q_{\bar{\alpha}}$) allows quantifying the β component of the quadratic entropy (Q_β) using an additive framework and hence the intercommunity diversity (Ricotta and Szeidl, 2009; de Bello *et al.*, 2010):

$$Q_\beta = Q_\gamma - Q_{\bar{\alpha}} \quad (4)$$

Standardized indices

Recent studies show that many diversity indices, including the Rao quadratic entropy, might have counterintuitive ecological properties (Jost, 2007; Ricotta and Szeidl, 2009; de Bello *et al.*, 2010). Indeed, when α-diversity increases, the β-diversity decreases and approaches zero, even in cases where there are no shared species between sampling units. Consequently, estimated β-diversity would be low regardless of the actual species overlap and the change in diversity across sampling units (Jost, 2007). We therefore applied the correction proposed by Jost (2007) derived from equivalent numbers (see de Bello *et al.*, 2010 for further details). Following its definition, the equivalent number of species is the number of maximally dissimilar species having equal abundance, which produces maximal entropy. Thus, by replacing $Q_{\bar{\alpha}}$ and Q_γ by their equivalent numbers in Eq. 4, we obtain the unbiased measures of intracommunity, regional and intercommunity diversity as follows (Ricotta and Szeidl, 2009):

$$Q_{\bar{\alpha}_{(corrected)}} = 1/(1 - Q_{\bar{\alpha}}) \quad (5)$$

$$Q_{\gamma_{(corrected)}} = 1/(1 - Q_\gamma) \quad (6)$$

$$Q_{\beta_{(corrected)}} = Q_{\gamma_{(corrected)}} - Q_{\bar{\alpha}_{(corrected)}} \quad (7)$$

It is worth noting that the correction is applied on $Q_{\bar{\alpha}}$ and not on the local Q_α so the relationship that makes $Q_{\bar{\alpha}}$ the mean of local Q_α is lost (see de Bello *et al.*, 2010 for more details).

To quantify the relative proportion of α and β components of diversity within the γ diversity, the corrected $Q_{\beta(corrected)}$ component of quadratic entropy can be expressed as the percentage of the corrected $Q_{\gamma(corrected)}$ component (total regional diversity):

$$Q_{\beta st} = \frac{Q_{\beta_{(corrected)}}}{Q_{\gamma_{(corrected)}}} \quad (8)$$

We compiled a function under the free R software (R Development Core Team, 2011) to estimate all Rao indices described earlier. This function, along with an R-script is available in Appendix S1.

Theoretical examples

Our described decomposition is illustrated with four simplified but realistic cases in Fig. 2, according to the nature of the MDU data (presence/absence versus relative abundance, and known versus unknown relationships among MDUs). We built an artificial regional pool of five MDUs (A–E) scattered across three local communities (I–III) of similar size, i.e. three individuals [cases (A) and (B)) and 100 individuals [cases (C) and (D)] distributed into three MDUs taken from the regional pool. For each community, we estimated α-diversity (Q_α), the mean α-diversity ($Q_{\bar{\alpha}}$), the regional γ-diversity (Q_γ), the β-diversity (Q_β) and the standardized β-diversity ($Q_{\beta st}$). The Jost correction was applied in order to quantify diversity while accounting for the equivalent number of species, i.e. the number of maximally dissimilar ($d_{ij} = 1$) and evenly distributed MDUs required to obtain the same index value (Q) as estimated with our dataset. In case (A), data are presence–absence of MDUs, while pairwise distances are unknown ($d_{ij} = 1$). In case (C), the relative abundances of MDUs are known.

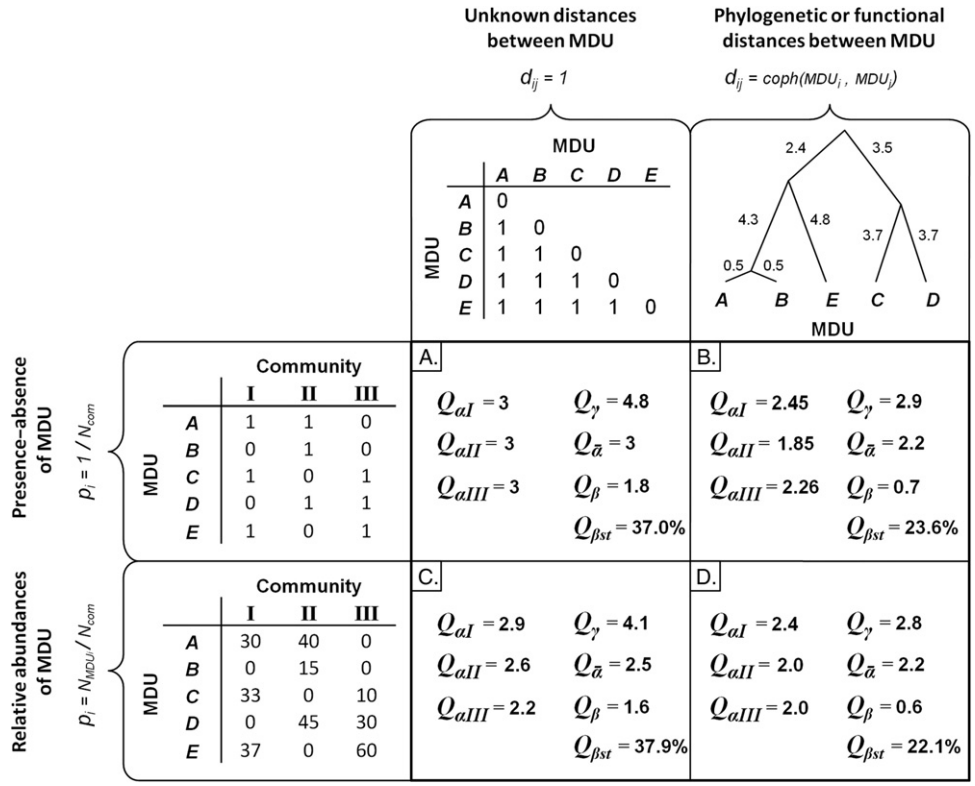

Fig. 2. Schematic example illustrating calculation of additive partitioning of Rao's quadratic entropy ($Q_{\bar{\alpha}}$, Q_β and Q_γ) for different types of data. We estimated the different components of Rao entropy using the Rao function: α-diversity (Q_α), the mean α-diversity ($Q_{\bar{\alpha}}$), the regional γ-diversity (Q_γ), the β-diversity (Q_β) and the standardized β-diversity ($Q_{\beta st}$) (see Appendix S1 for data, R-script and R-function). We applied the Jost correction, but the weighting of local communities was not applied because all communities contain the same number of individuals (100).
A. Presence–absence data of equally distant Microbial Diversity Units (MDUs, A–E) within local communities (I, II and III).
B. Presence–absence of phylogenetically related MDUs within local communities.
C. Abundance data of equally distant MDUs within local communities.
D. Abundance data of phylogenetically related MDUs within local communities. d_{ij} corresponds to the distance between MDU_i and MDU_j; *coph* (MDU_i and MDU_j) is the cophenetic distance between MDU_i and MDU_j, that is the length of the branches relating these two MDUs on a phylogenetic or functional tree; N_{com} is the total number of individuals in the studied communitym, and N_{MDU} is the number of individuals of the MDU_i. Note that $Q_{\bar{\alpha}}$ values do not correspond to the mean local Q_α values because the Jost correction was applied after the calculation of local Q_α, $Q_{\bar{\alpha}}$, Q_β and Q_γ.

Cases (B) and (D) are based on the same community matrices as cases (A) and (C), respectively, but the phylogenetic relatedness among the five MDUs are known in the formers. This relatedness corresponds to the pairwise cophenetic distances between MDUs, which is the amount of branch length relating all MDU pairs on an ultrametric phylogenetic tree.

In case (A), all communities have the same Q_α diversity as they contain the same number of MDUs (i.e. three). The regional Q_γ value does not equal the number of MDUs (i.e. five) because Q_γ decreases with the proportion of shared units among communities. Here, MDUs A, C, D and E are shared by two communities. In case (C), the relative abundances of MDUs are known. Local diversity (Q_α) is maximized when individuals are evenly distributed (community I) and minimized with unbalanced distributions of individuals amongst units (community III).

Cases (A) and (C) exhibit the highest $Q_{\beta st}$ values among the four cases; β-diversity represents more than 37% of the estimated regional diversity Q_γ. In cases (B) and (D), the highest Q_α value is estimated for community I, which has the highest phylogenetic diversity (distance = 19.2) as each MDU (A, C and E) belongs to scattered lineages. Moreover, MDUs have similar abundances in this community, thus increasing the estimated diversity. Community II has the lowest Q_α value in both cases. This is due to an uneven abundance distribution among MDUs but more importantly to the low phylogenetic diversity (distance = 14.9), explained by the presence of close relative MDUs such as A and B. Finally, community III has intermediate Q_α values because the phylogenetic diversity is intermediate (distance = 18.1) and abundances are unevenly distributed (case D).

The estimated β-diversity ($Q_{\beta st}$) represents 23.6% and 22.1% of the regional Q_γ diversity in cases (B) and (D)

respectively. These values are lower than in cases (A) and (B) because taking into account the phylogenetic relationships between MDUs reduces the dissimilarity between communities that share common branches on the phylogenetic tree in addition to sharing some species.

All the data, R-scripts and R-functions required to run these theoretical cases are available in a user friendly format in Appendix S1.

How does this framework enrich the microbial ecologist's toolbox?

What do we need in microbial ecology?

Microbial ecology faces, at least, two major challenges. The first one relies on the need to elucidate the role of microbes not only on ecosystem functioning, but also on ecosystem resilience and stability in the context of environmental changes (Bell *et al.*, 2005). The second challenge is to use the enormous amount of available and forthcoming microbial data generated through molecular approaches in a quantitative way that is more ecologically relevant (Jones *et al.*, 2012). These two challenges are interrelated because the former needs to generate a large number of samples (which is realistic in terms of sampling strategy and collection), and the second will see an increasing amount of information per individual and species that differ in their nature (abundance, identity, phylogeny, activity, physiology and function). To address these challenges, microbial ecology calls for a much better description of biodiversity, microbial processes and interactions in space and time. Decomposing diversity, as described earlier, in a way that fulfil Whitaker's framework but with more flexibility, is of high priority because it will allow the comparison of communities in a standardized way (time point and sites) and the integration of data of various sorts.

Measuring community structure to complement existing tools

Given the myriad of indices already available for microbiologists, the proposal of a new framework is only valuable if it brings additional and complementary information to existing tools (Lozupone and Knight, 2008). Using four theoretical cases (Fig. 3), we compared β-diversity values estimated using the additive Rao quadratic entropy framework, based on MDUs phylogenetic relatedness and their relative abundance, with those obtained using a classical additive composition, based on MDUs composition only (i.e. MDUs presence/absence).

In case (A), the two communities have no MDU in common, hence explaining the compositional β-diversity of 50%, which represents the highest possible value for this number of communities (see Appendix S2). The two MDUs within each community are closely related phylogenetically (low $Q_{\bar{\alpha}}$ value), but these two pairs of MDUs are phylogenetically distant between communities ($Q_{\beta st} = \beta_{st} = 50\%$, highest possible value). In this case, the taxonomic composition and the phylogenetic structure of the two communities differ in a similar way, i.e. the maximum level.

In case (B), the communities contain four MDUs while they share three of them. The estimated β-diversity, $Q_{\beta st}$ and β_{st} are all low because only one out of four species differs between the communities and because the two unshared species (A and B) are closely related phylogenetically (this reduces the $Q_{\beta st}$). In this case, the taxonomic composition and the phylogenetic structure are close between the two communities.

In case (C), the communities have no MDU in common, and the two MDUs that compose each community have marked unequal abundances. The β-diversity estimated using only community composition is 50% (β_{st}), which is the highest possible value, as in case (A). However, phylogenetic β-diversity ($Q_{\beta st}$) is much lower than is case (A) as closely phylogenetically related MDUs have the same abundances in their respective communities (A–B in community I, and C–D in community II). In this case, the taxonomic composition maximally differs between the two communities, but their phylogenetic structures are very close.

In case (D), the MDU composition is the same between the two communities, explaining the lowest compositional β-diversity value ($\beta_{st} = 0$) suggesting no turnover. However, the estimated phylogenetic β-diversity ($Q_{\beta st}$) is high (44.7% over a maximum of 50%) as the most abundant MDUs are phylogenetically distant between the two communities. In this case, while the community composition is perfectly identical between the two communities, their phylogenetic structure markedly differs.

Here, using four theoretical cases, we show that compositional and phylogenetic β-diversity are not trivially related and that the Rao framework deserves to be applied in addition to classical taxonomic-based analyses in order to reveal complementary biodiversity patterns.

Potential applications in microbial ecology

To illustrate possible uses of the presented framework, we identify three common issues that may necessitate a scaling of biodiversity (Fig. 4).

The first example corresponds to the monitoring of one or several bacterial communities over time (Fig. 4A). This can refer, for instance, to the dynamic of an *in situ* microbial community in different seasons, in experiments after input of contaminants, nutrients or after a modification in land use (Jones *et al.*, 2012; Perez-Leblic *et al.*, 2012;

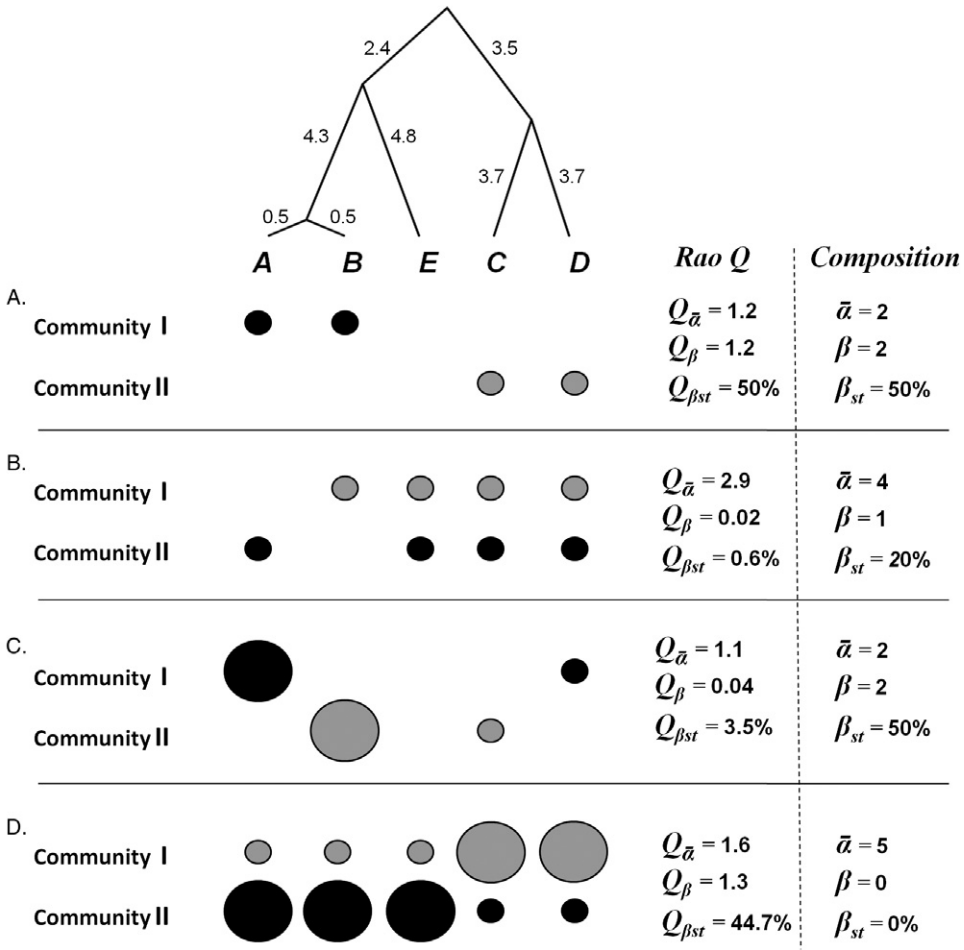

Fig. 3. Comparison of the additive framework for the Rao quadratic entropy with a composition-based additive partition of Microbial Diversity Unit (MDU) diversity. In each case (A–D), we calculated an additive partition of the regional diversity. We estimated $Q_{\bar{\alpha}}$, Q_{β} and $Q_{\beta st}$ components of diversity using the Rao quadratic entropy framework (Q). We estimated $\bar{\alpha}$, β and βst using an additive composition-based framework, i.e. a framework that considers all MDUs as equivalent and uses only presence/absence of MDUs within communities, where: γ = the number of MDUs across communities; $\bar{\alpha}$ = the mean number of MDUs per community; $\beta = \gamma - \bar{\alpha}$; $\beta st = \beta/\gamma$. The big circles correspond to 20 individuals and the small circles to 1.

Zhou *et al*., 2012). The objective would be then to determine whether the structure of these communities varies through time by estimating the relative contribution of $Q_{\bar{\alpha}}$ and Q_{β} values to Q_{γ} values. If $Q_{\bar{\alpha}}$ explains most of Q_{γ}, then microbial communities remain stable through time, while a higher contribution of Q_{β} to Q_{γ} would mean a major change in community structure.

Another possibility for use of our approach (Fig. 4B) is when investigating spatial processes across systems, metacommunities, and the dynamics of biodiversity and ecosystem processes from the nano to the regional scale (Jones *et al*., 2012; Yavitt *et al*., 2012). The relative contribution of $Q_{\bar{\alpha}}$ and Q_{β} values to Q_{γ} values, and their interactions with biotic and abiotic factors, may shed light on the processes underpinning empirical community patterns, i.e. the patch dynamics, species sorting, source–sink effects and neutral model frameworks (Logue *et al*., 2011).

The last example refers to micro-organism–macro-organism associations (Fig. 4C) and corresponds to the study of host-associated microbial communities. In the last years, there has been increased interest in understanding the structure of the indigenous microbiota that inhabit the surface or the inside of terrestrial and aquatic animals and plants (Fierer *et al*., 2012; Mouchet *et al*., 2012). The application of ecological theory to the host-associated microbiota, through description of α, β and γ components of diversity and their interactions, may push us beyond simple descriptions of community structure towards the understanding of mechanisms that structure their diversity and functions. Consequently, this could lead us to a better understanding of their role in animal and plant health (Fierer *et al*., 2012).

The three potential applications described earlier are not exhaustive. Within natural communities, the microbial taxa coexist with a wide range of physiologic states,

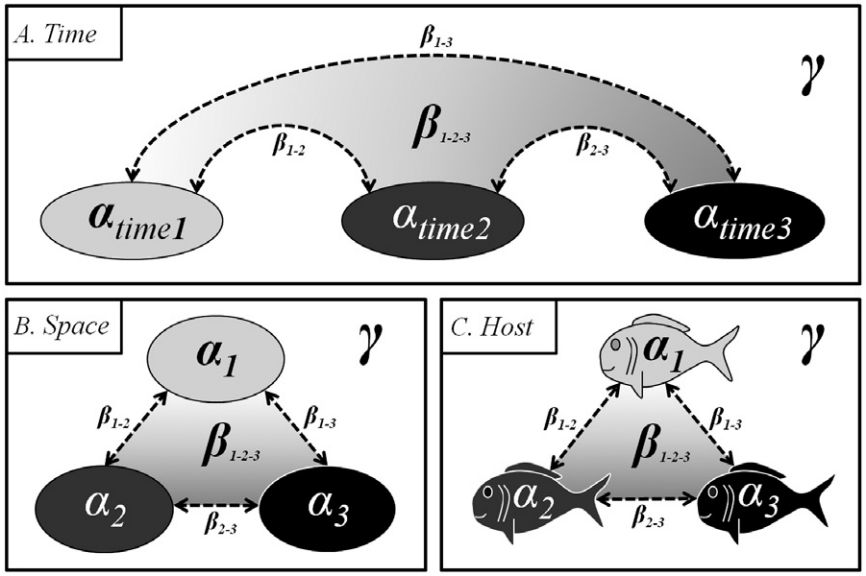

Fig. 4. Potential designs for studying the different components (α, β and γ) of microbial biodiversity across scales. In each case, the sampling unit is the local community (α_1, α_2 and α_3), γ represents the regional diversity across all communities and β the intercommunity diversity, in a global ($\beta_{1\text{-}2\text{-}3}$) or in a pairwise way ($\beta_{1\text{–}2}$, $\beta_{1\text{–}3}$ and $\beta_{2\text{–}3}$). In case (A), the three local communities correspond to the same community sampled at different times; in case (B) to spatially dispersed communities; and in case (C) to fish gut microbial communities.

expressed as different activity levels from very active to latent or even dead states (Del Giorgio and Gasol, 2008). Moreover, communities are dominated by species represented by few individuals with wide functional potential (Szabo *et al*., 2007). The loss and persistence of these categories of cells in the assemblage, as well as their relative importance in assembly processes are still unknown. However, there may be key players to explain the persistence of species and their global diversity patterns. These issues may benefit from the assessment of α, β and γ diversity of these cell categories at the different scales described in Fig. 4.

Concluding remarks

The modern molecular tools presented here allow estimating different facets of microbial diversity (taxonomic, phylogenetic and functional) with different but high levels of resolution and standardization. Many recent papers rely on this multifaceted approach to address questions such as the understanding of biogeographical patterns (Griffiths *et al*., 2011; Nemergut *et al*., 2011), the multiscale assessment of diversity and the comprehension of microbial community assembly rules (Lozupone and Knight, 2007; Fierer *et al*., 2012; Zinger *et al*., 2012). While these studies have facilitated the development of concepts and test major theories, they are not consistent in the way they measure the different components of diversity (α, β and γ) across scales, phylogenies and functions, so comparison between studies is not possible. The framework proposed by de Bello *et al*. (2010) and that we have adapted to the microbial world is unique by combining the dissimilarity and the relative abundances among the community members (here MDUs), and being flexible to cope with different kinds of data that are, or will be, generated by molecular tools. In addition, it provides a standardized methodology for the comparison of α, β and γ components across different facets of microbial diversity. Thus, large datasets covering microbial cell identity and function that are currently methodologically accessible, as well as the unified framework of diversity calculations described here, are key ingredients for successful findings in spatio-temporal distributions of microbial life, along with comparisons between case studies.

Acknowledgements

This work was partially funded by an EC2CO project FDFish (2008PRJ1) and the ANR project BIODIVNEK. Authors would like to thank Alison Duncan for improving the English language of this manuscript, and two anonymous reviewers for insightful comments on this manuscript. The authors declare that they have no conflict of interest.

References

Allen, B., Kon, M., and Bar-Yam, Y. (2009) A new phylogenetic diversity measure generalizing the shannon index and its application to phyllostomid bats. *Am Nat* **174:** 236–243.

Amann, R., and Fuchs, B.M. (2008) Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nat Rev Microbiol* **6:** 339–348.

Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L., *et al*. (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol Lett* **14:** 19–28.

Barnett, A.J., Finlay, K., and Beisner, B.E. (2007) Functional diversity of crustacean zooplankton communities: towards a trait-based classification. *Freshwater Biol* **52:** 796–813.

Bell, T., Newman, J.A., Silverman, B.W., Turner, S.L., and Lilley, A.K. (2005) The contribution of species richness and composition to bacterial services. *Nature* **436:** 1157–1160.

de Bello, F., Lavergne, S., Meynard, C.N., Lepš, J., and Thuiller, W. (2010) The partitioning of diversity: showing Theseus a way out of the labyrinth. *J Veg Sci* **21:** 992–1000.

Bent, S.J., Pierson, J.D., and Forney, L.J. (2007) Measuring species richness based on microbial community fingerprints: the emperor has no clothes. *Appl Environ Microbiol* **73:** 2399–2401.

Bissett, A., Richardson, A.E., Baker, G., Wakelin, S., and Thrall, P.H. (2011) Life history determines biogeographical patterns of soil bacterial communities over multiple spatial scales. *Mol Ecol* **19:** 4315–4327.

Bonilla-Rosso, G., Eguiarte, L.E., Romero, D., Travisano, M., and Souza, V. (2012) Understanding microbial community diversity metrics derived from metagenomes: performance evaluation using simulated data sets. *FEMS Microbiol Ecol* **82:** 37–49.

Bryant, J.A., Lamanna, C., Morlon, H., Kerkhoff, A.J., Enquist, B.J., and Green, J.L. (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proc Natl Acad Sci USA* **105:** 11505–11511.

Carvalhais, L.C., Dennis, P.G., Tyson, G.W., and Schenk, P.M. (2012) Application of metatranscriptomics to soil environments. *J Microbiol* **91:** 246–251.

Chang, Q., Luan, Y., and Sun, F. (2011) Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics* **12:** 118.

Christen, R. (2008) Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. *Microbes Environ* **23:** 253–268.

Del Giorgio, P.A., and Gasol, J.M. (2008) Physiological structure and single-cell activity in marine bacterioplankton. In *Microbial Ecology of the Oceans*. Kirchman, D.L. (ed.). New York, USA: John Wiley & Sons, pp. 243–298.

DeSantis, T.Z., Brodie, E.L., Moberg, J.P., Zubieta, I.X., Piceno, Y.M., and Andersen, G.L. (2005) Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiol Lett* **245:** 271–278.

Devictor, V., Mouillot, D., Meynard, C., Jiguet, F., Thuiller, W., and Mouquet, N. (2010) Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: the need for integrative conservation strategies in a changing world. *Ecol Lett* **13:** 1030–1040.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., *et al*. (2008) Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632.

Fierer, N., and Lennon, J.T. (2011) The generation and maintenance of diversity in microbial communities. *Am J Bot* **98:** 439–448.

Fierer, N., Ferrenberg, S., Flores, G.E., González, A., Kueneman, J., Legg, T., *et al*. (2012) From animalcules to an ecosystem: application of ecological concepts to the human microbiome. *Annu Rev Ecol Evol Syst* **43:** 137–155.

Fuhrman, J.A. (2009) Microbial community structure and its functional implications. *Nature* **459:** 193–199.

Gevers, D., and Coenye, T. (2007) Phylogenetic and genomic analysis. In *Manual of Environmental Microbiology*. Hurst, C.J., Crawford, R.L., Garland, J.L., Lipson, D.A., Mills, A.L., and Stetzenbach, L.D. (eds). Washington, DC, USA: ASM Press, pp. 157–168.

Gremion, F., Chatzinotas, A., and Harms, H. (2003) Comparative 16S rDNA and 16S rRNA sequence analysis indicates that Actinobacteria might be a dominant part of the metabolically active bacteria in heavy metal-contaminated bulk and rhizosphere soil. *Environ Microbiol* **5:** 896–907.

Griffiths, R.I., Thomson, B.C., James, P., Bell, T., Bailey, M., and Whiteley, A.S. (2011) The bacterial biogeography of British soils. *Environ Microbiol* **13:** 1642–1654.

Handley, K.M., Wrighton, K.C., Piceno, Y.M., Andersen, G.L., DeSantis, T.Z., Williams, K.H., *et al*. (2012) High-density PhyloChip profiling of stimulated aquifer microbial communities reveals a complex response to acetate amendment. *FEMS Microbiol Ecol* **81:** 188–204.

He, Z.L., Van Nostrand, J.D., Wu, L.Y., and Zhou, J.Z. (2008) Development and application of functional gene arrays for microbial community analysis. *T Nonferr Metal Soc* **18:** 1319–1327.

Hirsch, P.R., Mauchline, T.H., and Clark, I.M. (2010) Culture-independent molecular techniques for soil microbial ecology. *Soil Biol Biochem* **42:** 878–887.

Holmes, S., Alekseyenko, A., Timme, A., Nelson, T., Pasricha, P.J., and Spormann, A. (2010) Visualization and statistical comparisons of microbial communities using R packages on phylochip data. *Pac Symp Biocomput***:** 142–153.

Huse, S.M., Dethlefsen, L., Huber, J.A., Welch, D.M., Relman, D.A., and Sogin, M.L. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* **4:** e1000255.

Jones, S.E., Cadkin, T.A., Newton, R.J., and McMahon, K.D. (2012) Spatial and temporal scales of aquatic bacterial beta diversity. *Front Microbiol* **3:** 318.

Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology* **88:** 2427–2439.

Jost, L. (2010) Independence of alpha and beta diversities. *Ecology* **91:** 1969–1974.

Jurasinski, G., Retzer, V., and Beierkuhnlein, C. (2009) Inventory, differentiation, and proportional diversity: a consistent terminology for quantifying species diversity. *Oecologia* **159:** 15–26.

Kellogg, C.A., Piceno, Y.M., Tom, L.M., DeSantis, T.D., Zawada, D.G., and Andersen, G.L. (2012) PhyloChip™ microarray comparison of sampling methods used for coral microbial ecology. *J Microbiol Methods* **88:** 103–109.

Kembel, S.W. (2009) Disentangling niche and neutral influences on community assembly: assessing the performance of community phylogenetic structure tests. *Ecol Lett* **12:** 949–960.

Kembel, S.W., Eisen, J.A., Pollard, K.S., and Green, J.L. (2011) The phylogenetic diversity of metagenomes. *PLoS ONE* **6:** e23214.

Kirk, J., Beaudette, L.A., Hart, M., Moutoglis, P., Khironomos, J.N., Lee, H., and Trevors, J.T. (2004) Methods of studying soil microbial diversity. *J Microbiol Methods* **58:** 169–188.

Kisand, V., and Wikner, J. (2003) Limited resolution of 16S

rDNA DGGE caused by melting properties and closely related DNA sequences. *J Microbiol Methods* **54:** 183–191.
Koleff, P., Gaston, K.J., and Lennon, J.J. (2003) Measuring beta diversity for presence-absence data. *J Anim Ecol* **72:** 367–382.
Kruk, C., Huszar, V.L.M., Peeters, E.T.H.M., Bonilla, S., Costa, L., *et al*. (2010) A morphological classification capturing functional variation in phytoplankton. *Freshwater Biol* **55:** 614–627.
Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008) A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* **72:** 557–578.
Lande, R. (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* **76:** 5–13.
Liggenstoffer, A.S., Youssef, N.H., Couger, M.B., and Elshahed, M.S. (2010) Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores. *ISME J* **4:** 1225–1235.
Logue, J.B., Mouquet, N., Peter, H., Hillebrand, H., and Working, M. (2011) Empirical approaches to metacommunities: a review and comparison with theory. *Trends Ecol Evol* **26:** 482–491.
Loisel, P., Harmand, J., Zemb, O., Latrille, E., Lobry, C., Delgenes, J.P., and Godon, J.J. (2006) Denaturing gradient electrophoresis (DGE) and single-strand conformation polymorphism (SSCP) molecular fingerprintings revisited by simulation and used as a tool to measure microbial diversity. *Environ Microbiol* **8:** 720–731.
Lozupone, C.A., and Knight, R. (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104:** 11436–11440.
Lozupone, C.A., and Knight, R. (2008) Species divergence and the measurement of microbial diversity. *FEMS Microbiol Rev* **32:** 557–578.
Lozupone, C.A., Hamady, M., Kelley, S.T., and Knight, R. (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* **73:** 1576–1585.
Martin, A.P. (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* **68:** 3673–3682.
Mendez, V., Gill, J.A., Burton, N.H.K., Austin, G.E., Petchey, O.L., and Davies, R.G. (2012) Functional diversity across space and time: trends in wader communities on British estuaries. *Divers Distrib* **18:** 356–365.
Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nat Rev Genet* **11:** 31–46.
Mouchet, M.A., Villéger, S., Mason, N.W.H., and Mouillot, D. (2010) Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Funct Ecol* **24:** 867–876.
Mouchet, M.A., Bouvier, C., Bouvier, T., Troussellier, M., Escalas, A., and Mouillot, D. (2012) Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Funct Ecol* **24:** 867–876.
Nemergut, D.R., Costello, E.K., Hamady, M., Lozupone, C.A., Jiang, L., *et al*. (2011) Global patterns in the biogeography of bacterial taxa. *Environ Microbiol* **13:** 135–144.
Nocker, A., Burr, M., and Camper, A.K. (2007) Genotypic microbial community profiling: a critical technical review. *Microb Ecol* **54:** 276–289.
Odonnell, A.G., Goodfellow, M., and Hawksworth, D.L. (1994) Theoretical and practical aspects of the quantification of biodiversity among microorganisms. *Philos Trans R Soc Lond B Biol Sci* **345:** 65–73.
Pavoine, S.A., Ollier, S., and Pontier, D. (2005) Measuring diversity from dissimilarities with Rao's quadratic entropy: are any dissimilarities suitable? *Theor Popul Biol* **67:** 231–239.
Perez-Leblic, M.I., Turmero, A., Hernandez, M., Hernandez, A.J., Pastor, J., Ball, A.S., *et al*. (2012) Influence of xenobiotic contaminants on landfill soil microbial activity and diversity. *J Environ Manage* **95:** 285–290.
Peter, H., Ylla, I., Gudasz, C., Romani, A.M., Sabater, S., and Tranvik, L.J. (2011) Multifunctionality and diversity in bacterial biofilms. *PLoS ONE* **6:** e23225.
Petric, I., Philippot, L., Abbate, C., Bispo, A., Chesnot, T., Hallin, S., *et al*. (2011) Inter-laboratory evaluation of the ISO standard 11063 'Soil quality – method to directly extract DNA from soil samples'. *J Microbiol Methods* **84:** 454–460.
Petrosino, J.F., Highlander, S., Luna, R.A., Gibbs, R.A., and Versalovic, J. (2009) Metagenomic pyrosequencing and microbial identification. *Clin Chem* **55:** 856–866.
Pommier, T., Douzery, E.J.P., and Mouillot, D. (2012) Environment drives high phylogenetic turnover among oceanic bacterial communities. *Biol Lett* **8:** 562–566.
Prakash, T., and Taylor, T.D. (2012) Functional assignment of metagenomic data: challenges and applications. *Brief Bioinform* **13:** 711–727.
R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: Foundation for Statistical Computing [WWW document]. URL http://www.r-project.org/.
Rao, C.R. (1982) Diversity and dissimilarity coefficients: a unified approach. *Theor Popul Biol* **21:** 24–43.
Rappe, M.S., and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu Rev Microbiol* **57:** 369–394.
Ricotta, C., and Szeidl, L. (2009) Diversity partitioning of Rao's quadratic entropy. *Theor Popul Biol* **76:** 299–302.
Roh, S.W., Abell, G.C.J., Kim, K.H., Nam, Y.D., and Bae, J.W. (2010) Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends Biotechnol* **28:** 291–299.
Salles, J.F., Poly, F., Schmid, B., and Le Roux, X. (2009) Community niche predicts the functioning of denitrifying bacterial assemblages. *Ecology* **90:** 3324–3332.
Sanger, F., Nicklen, S., and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74:** 5463–5467.
Santamaria, M., Fosso, B., Consiglio, A., De Caro, G., Grillo, G., Licciulli, F., *et al*. (2012) Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform* **13:** 682–695.
Shendure, J., and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26:** 1135–1145.
Sleator, R.D., Shortall, C., and Hill, C. (2008) Metagenomics. *Lett Appl Microbiol* **47:** 361–366.

Stahl, D.A. (2007) Molecular approaches for the measurement of density, diversity, and phylogeny. In *Manual of Environmental Microbiology*. Garland, J.L. (ed.). Washington, DC, USA: ASM Press, pp. 139–156.

Stegen, J.C., Lin, X., Konopka, A.E., and Fredrickson, J.K. (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J* **6:** 1653–1664.

Su, C., Lei, L., Duan, Y., Zhang, K.Q., and Yang, J. (2012) Culture-independent methods for studying environmental microorganisms: methods, application, and perspective. *Appl Microbiol Biotechnol* **93:** 993–1003.

Suenaga, H. (2012) Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environ Microbiol* **14:** 13–22.

Szabo, K.E., Itor, P.O.B., Bertilsson, S., Tranvik, L., and Eiler, A. (2007) Importance of rare and abundant populations for the structure and functional potential of freshwater bacterial communities. *Aquat Microb Ecol* **47:** 1–10.

Tuomisto, H.A. (2010a) 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* **33:** 2–22.

Tuomisto, H.A. (2010b) A diversity of beta diversities: straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. *Ecography* **33:** 23–45.

Vellend, M. (2001) Do commonly used indices of β-diversity measure species turnover? *J Veg Sci* **12:** 545–552.

Vellend, M., Cornwell, W.K., Magnuson-Ford, K., and Mooers, A.O. (2010) Measuring phylogenetic biodiversity. In *Biological Diversity: Frontiers in Measurement and Assessment*. Magurran, A., and McGill, B. (eds). Oxford, UK: Oxford University Press, pp. 194–207.

Villéger, S., Ramos Miranda, J., Flores Hernandez, D., and Mouillot, D. (2012) Low functional β-diversity despite high taxonomic β-diversity among tropical estuarine fish communities. *PLoS ONE* **7:** e40679.

Von Mering, C., Hugenholtz, P., Raes, J., Tringe, S.G., Doerks, T., Jensen, L.J., *et al*. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315:** 1126–1130.

Von Wintzingerode, F., Göbel, U.B., and Stackebrandt, E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* **21:** 213–229.

Vos, M., Quince, C., Pijl, A.S., de Hollander, M., and Kowalchuk, G.A. (2012) A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. *PLoS ONE* **7:** e30600.

Ward, D.M., Weller, R., and Bateson, M.M. (1990) 16S ribosomal-RNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345:** 63–65.

Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol Monogr* **30:** 279–338.

Whittaker, R.H. (1972) Evolution and measurement of species diversity. *Taxon* **21:** 213–251.

Xiong, J., Wu, L., Tu, S., Van Nostrand, J.D., He, Z., Zhou, J., and Wang, G. (2010) Microbial communities and functional genes associated with soil arsenic contamination and the rhizosphere of the arsenic-hyperaccumulating plant Pteris vittata L. *Appl Environ Microbiol* **76:** 7277–7284.

Yavitt, J.B., Yashiro, E., Cadillo-Quiroz, H., and Zinder, S.H. (2012) Methanogen diversity and community composition in peatlands of the central to northern Appalachian Mountain region, North America. *Biogeochemistry* **109:** 117–131.

Zhou, J. (2003) Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* **6:** 288–294.

Zhou, X., Zhang, Y., and Downing, A. (2012) Non-linear response of microbial activity across a gradient of nitrogen addition to a soil from the Gurbantunggut Desert, northwestern China. *Soil Biol Biochem* **47:** 67–77.

Zinger, L., Gobet, A., and Pommier, T. (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* **21:** 1878–1896.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. This is the R-script that explains the content of the supplementary files and how to use the Rao function.
Appendix S2. This file contains complementary information about the uses and behaviour of the Rao function.
F(Rao).R. This is the R-code of the Rao function.
PhyloTree.tre. This file is a NEWICK phylogenetic tree and served as example in the Appendix S1 R-script.
Fig. S1. Fig2_community.abundance.txt: This is a matrix of the data used for Fig. 2. These data served as example in the Appendix S1 R-script.
Fig. S2. Fig2_community.presence.absence.txt: This is a matrix of the data used for Fig. 2. These data served as example in the Appendix S1 R-script.
Fig. S3. Fig3_caseA, Fig3_caseB, Fig3_caseC, Fig3_caseD: These files contain the data used in Fig. 3. These data served as example in the Appendix S1 R-script.