Measuring diversity: the importance of species similarity

Tom Leinster[1,2,*] Christina A. Cobbold[1,2]

[1]School of Mathematics and Statistics, University of Glasgow, UK
[2]Boyd Orr Centre for Population and Ecosystem Health, University of Glasgow, UK

Abstract

Realistic measures of biodiversity should reflect not only the relative abundances of species, but also the differences between them. We present a natural family of diversity measures taking both factors into account. This is not just another addition to the already long list of diversity indices: instead, a single formula subsumes many of the most popular indices, including Shannon's, Simpson's, species richness, and Rao's quadratic entropy. These popular indices can then be used and understood in a unified way, and the relationships between them are made plain. The new measures are, moreover, effective numbers, so that percentage changes and ratio comparisons of diversity value are meaningful.

We advocate the use of diversity profiles, which provide a faithful graphical representation of the shape of a community; they show how the perceived diversity changes as the emphasis shifts from rare to common species. Communities can usefully be compared by comparing their diversity profiles. We show by example that this is a far more subtle method than any relying on a single statistic.

Some ecologists view diversity indices with suspicion, questioning whether they are biologically meaningful. By dropping the naive assumption that distinct species have nothing in common, working with effective numbers, and using diversity profiles, we arrive at a system of diversity measurement that should lay much of this suspicion to rest.

Key words: diversity, biodiversity, entropy, quadratic entropy, species similarity, model, effective number, diversity profile, microbial diversity.

1 Introduction

'A mathematical approach does not oblige a biologist to be modest about his ability to make biological distinctions', wrote Hurlbert in 1971. Yet modesty seems to prevail when it comes to measuring diversity: all the most commonly-used indices are based on a crude model in which distinct species are assumed to have nothing in common, contrary to what every biologist knows. Non-specialists are amazed to learn that a community of six dramatically different species is said to be no more diverse than a community

*Corresponding author. Email: Tom.Leinster@glasgow.ac.uk

of six species of barnacle. There is a mismatch between the general understanding of biodiversity as the variety of life, and the diversity indices used by biologists every day.

With the preservation of biodiversity a pressing global concern, this mismatch matters. 'Diversity' is one of those words that is used freely in both scientific and non-scientific contexts, often with different meanings (Adams et al. 1997). Politicians may understand diversity to mean one thing; the scientists advising them may use it to mean another. Misguided policies may be the result. The Organisation for Economic Co-operation and Development's guide to biodiversity for policy makers states that 'associated with the idea of diversity is the concept of "*distance*", i.e. some measure of the dissimilarity of the resources in question' (OECD 2002). But the conventional measures of diversity ignore this aspect altogether.

This unhappy situation may result from a lack of good diversity measures that reflect the varying dissimilarities between species, or a lack of understanding of how to use them. Let us call such measures *similarity-sensitive*. The best-known similarity-sensitive diversity measure is the quadratic entropy of Rao (1982a,b). This is receiving increasing attention, but is still a minor player. Perhaps theoretical ecologists have been hesitant to introduce new diversity indices when the profusion of similarity-*in*sensitive indices is already perceived to form an impenetrable jungle (Ricotta 2005)—although work of Jost (2006, 2007, 2009) dispels that myth.

We present a new family of similarity-sensitive diversity measures, and show how to use them. This family includes—either directly or upon applying a simple transformation—Rao's quadratic entropy, species richness, Shannon entropy, the Gini–Simpson index, the Berger–Parker index, the Hill numbers, the Patil–Taillie–Tsallis entropies, and the entropies of Ricotta and Szeidl (2006) (of which we give a new interpretation). We can also extract the indices of Hurlbert (1971) and Smith and Grassle (1977), and there are close connections with the phylogenetic indices of Faith (1992), Allen et al. (2009) and Chao et al. (2010). Once these many indices are all seen in the same context (Table 1), the relationships between them are clarified.

Our diversity measures take two inputs:

- Relative abundance data. We assume the members of the community to be divided into species; the relative abundance data describes the proportions in which they are present. The word 'species' can stand for any unit thought biologically meaningful.

- Similarity data: for each pair of species, a number specifying how similar they are. Again, 'similar' can be used in any biologically meaningful way: a genetic notion of similarity will lead to a measure of genetic diversity, a functional notion of similarity will lead to a measure of functional diversity, and so on. The traditional, naive model, in which commonalities between species are ignored, implicitly takes all similarities between distinct species to be zero. This leads to a naive measure of diversity.

The user also chooses a parameter q between 0 and ∞, indicating how much significance is attached to species abundance. For example, at one extreme ($q = 0$), species richness attaches as much significance to rare species as common ones. At the other ($q = \infty$), the index of Berger and Parker (1970) depends only on the most abundant species; rare species are ignored altogether.

Given the abundance and similarity data, and a choice of parameter q, our formula produces a number: the *diversity of order q* of the community. It is an effective number. This means that the diversity of order q of a community of S totally dissimilar species in equal proportions is simply S. Thus, if a community is assigned a diversity of 18.2, that

means that it is slightly more diverse than a community of 18 totally dissimilar equally abundant species. There are 'effectively' 18.2 species. Effective numbers 'enable us to speak naturally' (Hill 1973).

Jost (2006, 2007, 2009) has argued eloquently for the primacy of effective numbers. Among the many diversity indices, those that are effective numbers play a special role and deserve a special name. Jost calls them 'true diversities'; we call our measures simply 'diversities'. Any diversity index can be converted into an effective number by a few simple steps of algebra (Jost 2006). Adopting this as a common standard clears up much confusion.

Similarity-sensitive measures that are effective numbers have a crucial advantage over earlier similarity-sensitive indices such as Rao's (1982a, 1982b) and Ricotta and Szeidl's (2006). Chao et al. (2010) defined an important family of effective number similarity-sensitive measures, tailored specifically to phylogenetic diversity. As shown in the Appendix, they are closely related to our measures.

Given relative abundance and similarity data, one should calculate the diversity of order q for *every* q, and plot it against q. This graph is the community's *diversity profile*. Meaningful ecological information can be read off at a glance. We illustrate this with examples, arguing that the diversity profile of a community can be regarded as its fingerprint.

Microbial ecologists have long recognized the need for similarity or distance measures in the quantification of diversity (Mills and Wassel 1980), because of the complexities of microbial taxonomy. We show how to apply our measures to communities of microbes.

Taking species similarity into account gives a more accurate reflection of reality. It also sheds light on the hidden assumptions inherent in the naive model. Our more subtle approach is adaptable to the needs of the user, in that it allows for measurement of different types of diversity: genetic, morphological, functional, and so on. A plethora of diversity indices, both sensitive and insensitive to species similarity, is replaced by a single formula. Thus, our approach is not only more realistic and versatile: it also simplifies.

2 The diversity measures

We consider throughout a fully-censused community of S species, with relative abundances denoted by $p_1, \ldots, p_S$; thus, $p_i \geq 0$ and $\sum_{i=1}^{S} p_i = 1$. We write $\mathbf{p} = (p_1, \ldots, p_S)$. The similarities between species are encoded in an $S \times S$ matrix $\mathbf{Z} = (Z_{ij})$, with Z_{ij} measuring the similarity between the i^{th} and j^{th} species. We assume that $0 \leq Z_{ij} \leq 1$, with 0 indicating total dissimilarity and 1 indicating identical species; hence we also assume that $Z_{ii} = 1$.

Genetic measures of similarity or homology are often expressed as percentages, directly providing similarity coefficients Z_{ij} on a scale of 0 to 1. Other measures of interspecies distance d_{ij} lie on a scale of 0 to infinity, but can easily be transformed to lie on a scale of 0 to 1 by the formula $Z_{ij} = e^{-ud_{ij}}$ (Nei 1972), where u is a constant. Different transformations are possible, but this is probably the simplest. We return to this in the Discussion.

Although the most obvious examples of similarity matrices are symmetric ($Z_{ij} = Z_{ji}$), symmetry is *not* part of the definition of similarity matrix. Partly this is because none of our results require the assumption of symmetry. Partly it is because there are useful non-symmetric similarity matrices; for example, such matrices enable us to connect our diversity measures to certain existing measures of phylogenetic diversity

(Appendix, Proposition A7). If the prospect of a non-symmetric similarity matrix seems counterintuitive, it may be useful to consider the related concept of distance. In a general scientific context, the most obvious measures of distance are, again, symmetric. But there are many physical situations in which non-symmetric distances play an important role. For example, the work required to push a load up a slope is greater than that required to push it down again.

We now define our family of diversity measures. There is one measure for each value of the parameter q in the range $0 \leq q \leq \infty$. This is called the *sensitivity parameter*, and controls the relative emphasis that the user wishes to place on common and rare species; it is explained in Section 5.

For $q \neq 1, \infty$, the *diversity of order* q of the community is

$$ {}^qD^{\mathbf{Z}}(\mathbf{p}) = \left(\sum p_i (\mathbf{Zp})_i^{q-1} \right)^{\frac{1}{1-q}}, \tag{1} $$

where

$$ (\mathbf{Zp})_i = \sum_{j=1}^{S} Z_{ij} p_j. $$

The sum in (1) is over all values of $i = 1, \ldots, S$ such that $p_i \neq 0$. In other words, it is over all species that are actually present.

We justify our definition in three ways: by explaining the formula directly (this section), by exhibiting many well-known diversity measures as special cases (Section 3), and by listing its many desirable properties (Section 4), chief among which is that ${}^qD^{\mathbf{Z}}(\mathbf{p})$ is an effective number.

First we explain the significance of the quantity $(\mathbf{Zp})_i$. It is the expected similarity between an individual of the i^{th} species and an individual chosen at random. It therefore measures the ordinariness of the i^{th} species within the community. We call $(\mathbf{Zp})_i$ the *relative abundance of species similar to the* i^{th}. We always have $(\mathbf{Zp})_i \geq p_i$: the relative abundance of species similar to the i^{th} is at least as great as the relative abundance of the i^{th} species itself. (It follows that $\sum (\mathbf{Zp})_i$ usually exceeds 1.)

Since $(\mathbf{Zp})_i$ measures the ordinariness of the i^{th} species within the community, the average ordinariness of an individual from the community is

$$ \sum_{i=1}^{S} p_i (\mathbf{Zp})_i. \tag{2} $$

This quantity is large if most of the population is concentrated into a few very similar species. So, average ordinariness could be called *concentration*, and is inversely related to diversity. A measure of diversity is therefore provided by the reciprocal, $1 \big/ \sum p_i (\mathbf{Zp})_i$. This is precisely ${}^2D^{\mathbf{Z}}(\mathbf{p})$, the diversity of order 2.

The diversities of other orders $q \neq 2$ arise from other notions of average. The mean of $x_1, \ldots, x_S$ is $\sum \frac{1}{S} x_i$. More generally, for any weights $p_1, \ldots, p_S$ adding up to 1, the weighted mean is $\sum p_i x_i$. But there is also, for each real number $t \neq 0$, another kind of average: first transform each x_i into x_i^t, then take the weighted mean, then apply the inverse transformation. This is the *generalized mean* or *power mean* $(\sum p_i x_i^t)^{1/t}$ (Hardy et al. 1952). Taking $t = q - 1$ and $x_i = (\mathbf{Zp})_i$ gives

$$ \left(\sum p_i (\mathbf{Zp})_i^{q-1} \right)^{\frac{1}{q-1}} $$

as a measure of average ordinariness, or concentration, of the community. Its reciprocal, ${}^qD^{\mathbf{Z}}(\mathbf{p})$, is therefore a measure of diversity. Varying the parameter q varies the influence

	sensitive to rare species ←			→ insensitive to rare species	
	$q = 0$	$q = 1$	$q = 2$	$q = \infty$	Remarks
Diversity ${}^qD^{\mathbf{Z}}$	(compare Faith)		$\frac{1}{1-\text{Rao}}$		compare CCJ
Naive diversity ${}^qD = {}^qD^{\mathbf{I}}$ (Hill numbers)	species richness	exp(Shannon)	inverse Simpson concentration	$\frac{1}{\text{Berger–Parker}}$	${}^2D, \ldots, {}^mD$ give HSG
Entropy ${}^qH^{\mathbf{Z}}$ (Ricotta–Szeidl)		(compare AKB)	Rao's quadratic entropy	—	
Naive entropy ${}^qH = {}^qH^{\mathbf{I}}$ (Patil–Taillie–Tsallis)	species richness minus 1	Shannon entropy	Gini–Simpson	—	

Table 1: How some familiar diversity indices can be derived from our diversities ${}^qD^{\mathbf{Z}}$. For the the three entries marked 'compare', see the Appendix (pages 24–28). Abbreviations: CCJ = Chao–Chiu–Jost; HSG = Hurlbert–Smith–Grassle; AKB = Allen–Kon–Bar-Yam. References: Faith (1992), Rao (1982a,b), Chao et al. (2010), Hill (1973), Simpson (1949), Berger and Parker (1970), Hurlbert (1971), Smith and Grassle (1977), Ricotta and Szeidl (2006), Allen et al. (2009), Patil and Taillie (1982), Tsallis (1988), Gini (1912).

on diversity of ordinary species (those i for which $(\mathbf{Zp})_i$ is large) relative to unusual species (those for which it is small).

The cases $q = 1$ and $q = \infty$ have been excluded because the formula (1) for ${}^qD^{\mathbf{Z}}(\mathbf{p})$ does not make sense there. It does, however, converge to a limit as $q \to 1$ or $q \to \infty$. We define ${}^1D^{\mathbf{Z}}(\mathbf{p})$ and ${}^\infty D^{\mathbf{Z}}(\mathbf{p})$ as those limits, namely

$$
\begin{aligned}
{}^1D^{\mathbf{Z}}(\mathbf{p}) &= 1/(\mathbf{Zp})_1^{p_1}(\mathbf{Zp})_2^{p_2}\cdots(\mathbf{Zp})_S^{p_S}, \\
{}^\infty D^{\mathbf{Z}}(\mathbf{p}) &= 1/\max(\mathbf{Zp})_i
\end{aligned}
$$

(Appendix, Proposition A2), where any term 0^0 in the first formula is evaluated as 1, and the maximum in the second is over all $i = 1, \ldots, S$ such that $p_i \neq 0$.

We have taken care to cover the eventuality that $p_i = 0$ for some values of i. This may occur if, for instance, one conducts an annual survey of a site using a checklist of species: some years, some species may be absent. Propositions A1 and A2 of the Appendix show why this eventuality must be handled in the way that it is.

3 Relationships between diversity indices

Here we show that many familiar diversity indices arise from ours, or are closely related (Table 1). In some cases, the familiar index is equal to ${}^qD^{\mathbf{Z}}(\mathbf{p})$ for a particular value of q and/or $\mathbf{Z}$. In others, it becomes equal upon applying a simple transformation.

We also explain some of the new measures arising from the general definition, and show how they can be used to measure the diversity of a community of microbes—a problem area for many diversity indices.

The oldest and most common measure of diversity is species richness, the number $s \leq S$ of values of i such that $p_i \neq 0$. This measure takes no notice of the varying similarities between species; it uses the *naive model* of a community, in which the similarity coefficient Z_{ij} is taken to be 0 (total dissimilarity) if $i \neq j$, and 1 (total similarity) if $i = j$. Hence $\mathbf{Z}$

is the identity matrix $\mathbf{I}$, and $(\mathbf{Zp})_i = p_i$. Writing ${}^qD(\mathbf{p}) = {}^qD^{\mathbf{I}}(\mathbf{p})$, we have ${}^0D(\mathbf{p}) = s$: species richness is the naive diversity of order 0.

In general, the naive diversity ${}^qD(\mathbf{p})$ is the *Hill number* of order q (Hill 1973). The naive diversity ${}^1D(\mathbf{p})$ of order 1 is the exponential of Shannon entropy, a form advocated by MacArthur (1965) and Whittaker (1972).

Now take an arbitrary similarity matrix $\mathbf{Z}$. The diversity of order 0 is a similarity-sensitive version of species richness, given by

$$ {}^0D^{\mathbf{Z}}(\mathbf{p}) = \sum_{i:\ p_i \neq 0} \frac{p_i}{(\mathbf{Zp})_i}. $$

The contribution $p_i/(\mathbf{Zp})_i$ made by the i^{th} species is always between 0 and 1. It is close to 1 when there are few individuals of other similar species: a species makes the greatest contribution to diversity when it is unusual. Diversity of order 0 includes species richness (the case $\mathbf{Z} = \mathbf{I}$), as well as Faith's phylogenetic diversity measure when the phylogenetic tree is ultrametric (Appendix, pages 24–28). Neither of these depends on $\mathbf{p}$, except concerning whether each p_i is zero or not; but in general, diversity of order 0 does depend on $\mathbf{p}$.

The diversity of order 2 is

$$ {}^2D^{\mathbf{Z}}(\mathbf{p}) = \frac{1}{\sum_{i,j} p_i Z_{ij} p_j} = \frac{1}{\mu_2}, $$

where μ_2 is the expected similarity between a randomly-chosen pair of individuals. This is closely related to a common measure of genetic diversity, as we shall see. In the naive model, it is the inverse Simpson concentration $1/\sum p_i^2$.

More generally, take any whole number $q \geq 2$. Given q individuals of respective species $i_1, i_2, \ldots, i_q$, the product

$$ Z_{i_1,i_2} Z_{i_1,i_3} \cdots Z_{i_1,i_q} \tag{3} $$

is a measure of their similarity as a group. Call (3) their *group similarity*, and let μ_q be the expected similarity of a randomly-chosen group of q individuals (sampled with replacement). Then

$$ {}^qD^{\mathbf{Z}}(\mathbf{p}) = \mu_q^{1/(1-q)} \tag{4} $$

(Appendix, Proposition A3). Thus, diversity increases as the mean group similarity decreases.

Formula (4) can be applied in situations where many diversity indices are unusable. For example, we can use it to estimate the diversity of a community of microbes, where there are good notions of similarity but the question of what constitutes a species is highly problematic (Johnson 1973, Watve and Gangal 1996). To apply the formula for ${}^qD^{\mathbf{Z}}(\mathbf{p})$ we do not *need* to know what a species is: it is enough to have a measure of similarity between two isolates. An estimate for μ_q (hence ${}^qD^{\mathbf{Z}}$) is given by repeatedly taking q isolates from the community, calculating the group similarity for each, and taking the mean.

The naive diversity ${}^\infty D(\mathbf{p})$ of order ∞ is $1/\max p_i$, the reciprocal of the Berger–Parker index. This is a measure of dominance. The same can be said of ${}^\infty D^{\mathbf{Z}}(\mathbf{p})$ for general $\mathbf{Z}$, but now dominance is measured not merely in terms of how abundant each species is—it also takes into account how abundant *similar* species are. The species i for which $(\mathbf{Zp})_i$ is greatest need not itself be very abundant, as long as there are highly abundant species similar to it. For morphological diversity (Pavoine et al. 2005), diversity

of order ∞ will be highest when there are no clusters of populous species in any small region of morphometric space.

Much of the literature on diversity indices concerns 'entropies' of various kinds. These are not effective numbers, so we do not advocate using them as primary measures. However, in order to demonstrate the simplifying power of our definition, we now show that many such entropies are also just transformations of the diversities ${}^qD^{\mathbf{Z}}(\mathbf{p})$.

The explanation is in terms of 'surprise', a concept from information theory. (We need to invoke information theory only to make connections with historically established indices; it is not needed in order to justify our diversities themselves.) This extends the narrative of Patil and Taillie (1982) and Ricotta and Szeidl (2006).

When sampling from the community, our surprise at finding an individual of the i^{th} species decreases with its ordinariness—that is, with the abundance of organisms of the same or similar species. We are most surprised when we find a rare, distinctive species. Mathematically, we can quantify the surprise as $\sigma((\mathbf{Zp})_i)$, where $\sigma(x)$ is some decreasing function of x ($0 \leq x \leq 1$). For a randomly-chosen individual, the expected surprise is $\sum p_i \sigma((\mathbf{Zp})_i)$. This is an index of the diversity of the whole community.

Patil and Taillie (1982) defined a surprise function σ_q for each $q \geq 0$:

$$\sigma_q(x) = \begin{cases} \frac{1}{q-1}\left(1 - x^{q-1}\right) & \text{if } q \neq 1 \\ -\ln x & \text{if } q = 1. \end{cases}$$

(The second expression is the limit of the first as $q \to 1$.) This gives, for each $q \geq 0$, a diversity index ${}^qH^{\mathbf{Z}}$, the expected surprise according to σ_q:

$${}^qH^{\mathbf{Z}}(\mathbf{p}) = \sum p_i \sigma_q((\mathbf{Zp})_i) = \begin{cases} \frac{1}{q-1}\left(1 - \sum p_i (\mathbf{Zp})_i^{q-1}\right) & \text{if } q \neq 1 \\ -\sum p_i \ln(\mathbf{Zp})_i & \text{if } q = 1. \end{cases}$$

As usual, the sums are over all $i = 1, \ldots, S$ such that $p_i \neq 0$.

These indices ${}^qH^{\mathbf{Z}}$ are the entropies of Ricotta and Szeidl (2006), who explained them in terms of inter-species conflict; the interpretation as expected surprise is new. Ricotta and Szeidl used the *dis*similarity matrix $\mathbf{\Delta}$, with entries $\Delta_{ij} = 1 - Z_{ij}$, rather than the similarity matrix $\mathbf{Z}$; Proposition A4 of the Appendix proves the equivalence of their formula and ours. (We adopt the policy that the word *dissimilarity* and the symbol Δ_{ij} refer to a measure of difference on a scale of 0 to 1, while *distance* and d_{ij} refer to a measure of difference on a scale of 0 to ∞. So although Ricotta and Szeidl called their coefficients 'distances' and denoted them by d_{ij}, we call them dissimilarities Δ_{ij}, because they are measured on a scale of 0 to 1.)

The diversities ${}^qD^{\mathbf{Z}}$ and entropies ${}^qH^{\mathbf{Z}}$ are related by the transformation

$${}^qH^{\mathbf{Z}}(\mathbf{p}) = \begin{cases} \frac{1}{q-1}(1 - {}^qD^{\mathbf{Z}}(\mathbf{p})^{1-q}) & \text{if } q \neq 1 \\ \ln({}^1D^{\mathbf{Z}}(\mathbf{p})) & \text{if } q = 1. \end{cases}$$

The indices ${}^qH^{\mathbf{Z}}$ and ${}^qD^{\mathbf{Z}}$ carry the same information, and they always make the same judgement on which of two communities is the more diverse. (This is because the transformation is invertible and increasing.) However, only ${}^qD^{\mathbf{Z}}$ has the cardinal virtue of being an effective number.

The entropy of order 2 is

$${}^2H^{\mathbf{Z}}(\mathbf{p}) = 1 - \sum_{i,j=1}^{S} p_i Z_{ij} p_j.$$

This is the *quadratic entropy* of Rao (1982a,b), notably used to measure nucleotide diversity (Nei and Tajima 1981). Usually it is expressed in terms of the dissimilarity matrix $\mathbf{\Delta}$; then

$$ {}^2H^{\mathbf{Z}}(\mathbf{p}) = \sum_{i,j=1}^{S} p_i \Delta_{ij} p_j, $$

the expected dissimilarity between a random pair of individuals. The effective number form ${}^2D^{\mathbf{Z}}(\mathbf{p})$ was also derived by Ricotta and Szeidl (2009).

Quadratic entropy in the naive model is the Gini–Simpson index, ${}^2H(\mathbf{p}) = 1 - \sum p_i^2$. Many authors have used quadratic entropy with matrices in which some of the dissimilarities Δ_{ij} are greater than 1, or equivalently $Z_{ij} < 0$ (e.g. Izsák and Papp 1995). This still gives a meaningful index of diversity; but since $Z_{ij} < 0$ indicates 'more-than-total dissimilarity', it destroys the possibility of a meaningful relationship between quadratic entropy and the Gini–Simpson index.

In the naive model $\mathbf{Z} = \mathbf{I}$, the Ricotta–Szeidl entropies ${}^qH^{\mathbf{Z}}$ become a well-known family of entropies ${}^qH = {}^qH^{\mathbf{I}}$. These first appeared, in a slightly different form, in information theory (Havrda and Charvát 1967, Aczél and Daróczy 1975). In ecology they were introduced by Patil and Taillie (1982), and in physics, finally, by Tsallis (1988). See Table 1.

The apparent profusion of diversity indices is, then, partly an illusion. Many familiar indices are special cases of our measures, or simple transformations thereof.

Chao et al. (2010) proposed a family of diversity measures taking into account phylogenetic similarities, derived from a phylogenetic tree. Their measures, called the mean phylogenetic diversity of order q, have excellent properties and are closely related to ours. (See Proposition A7 of the Appendix. There is a subtlety concerning non-ultrametric trees, detailed there.) Chao et al. showed that, after applying a simple transformation, the phylogenetic diversity measure of Faith (1992) and the phylogenetic entropy of Allen et al. (2009) are special cases of mean phylogenetic diversity. Hence they, too, are closely related to ours. As we show in the next section, new insights into mean phylogenetic diversity are gained by connecting it with our measures.

Further diversity indices can be obtained by combining ${}^qD^{\mathbf{Z}}$ for several values of q. For example, Hurlbert (1971) and Smith and Grassle (1977) studied the expected number of species occurring in a random sample of m individuals. This turns out to be a combination of the diversities ${}^2D, {}^3D, \ldots, {}^mD$ (Appendix, Proposition A8). By incorporating as many indices as possible into the family $({}^qD^{\mathbf{Z}})$, we move towards a systematic understanding of diversity measures.

4 Properties

Here we state the principal properties of our diversity measures. These properties encode basic scientific intuition, and any diversity measure taking species similarity into account should satisfy them all. For each value of q $(0 \leq q \leq \infty)$, the diversity measure ${}^qD^{\mathbf{Z}}$ passes this test.

Some of these properties might seem so obvious as not to merit a mention. But the literature on diversity measurement is strewn with indices failing to satisfy properties that might seem 'obvious'. Some indices have been used for decades, and even become the textbook standard, before it is pointed out that the supposedly obvious properties are, in fact, false; see Jost (2008) for an example.

Comparable lists of properties can be found in Rényi (1961), Routledge (1979), Chakravarty and Eichhorn (1991), Suyari (2002) and Jost (2009)—but none of the indices discussed there take account of the varying differences between species.

The properties are arranged in three groups and proved in the Appendix (Propositions A9–A19).

1. Partitioning properties

Effective number: the diversity of a community of S equally abundant, totally dissimilar species is S.

Modularity: suppose that the community is partitioned into m subcommunities, with no species shared between subcommunities, and with species in different subcommunities being totally dissimilar. Then the diversity of the community is entirely determined by the sizes and diversities of the subcommunities.

Replication: if, moreover, these m subcommunities are of equal size and equal diversity, d, then the diversity of the whole community is md.

Modularity enables us to calculate the diversity of a partitioned community from the diversities and sizes of the subcommunities alone, *without* having to know the abundance and similarity data within the subcommunities. The formula is as follows. Write $w_1, \ldots, w_m$ for the relative sizes of the subcommunities, so that $\sum w_i = 1$. Write d_i for the diversity of order q of the i^{th} subcommunity. Then the diversity of order q of the whole community is

$$\begin{cases} \left(\sum w_i^q d_i^{1-q}\right)^{\frac{1}{1-q}} & \text{if } q \neq 1, \infty \\ {}^1D(\mathbf{w}) d_1^{w_1} d_2^{w_2} \cdots d_m^{w_m} & \text{if } q = 1 \\ \min(d_i/w_i) & \text{if } q = \infty \end{cases}$$

where the sum and the minimum are over all i such that $w_i \neq 0$. This is proved in the Appendix (Proposition A10).

For a simple example of replication, suppose that m islands are each populated by d species, with all the species totally dissimilar and equally abundant. Then the whole community consists of md totally dissimilar equally abundant species. Diversity is an effective number, so each island has diversity d and the whole community has diversity md, as claimed. When $m = 2$, replication is called 'doubling'. Its importance is explained in Hill (1973) and Jost (2006); and as shown by Jost (2009), replication is essential for reasoning logically about conservation.

The replication and modularity principles for our measures give some new results on existing measures. For example, Chao et al. (2010) proved that their mean phylogenetic diversity satisfies replication, but only under the assumption that the subcommunities all have the same mean evolutionary change. Our general results show that their measures satisfy replication even without this assumption (Appendix, Corollary A12). Moreover, our results provide modularity formulas for the mean phylogenetic diversity of a community partitioned into completely distinct subcommunities, even when the subcommunities have different sizes and different diversities.

2. Elementary properties

Symmetry: diversity is unchanged by the order in which the species happen to be listed.

Absent species: diversity is unchanged by adding a new species of abundance 0.

Identical species: if two species are identical, then merging them into one leaves the diversity unchanged.

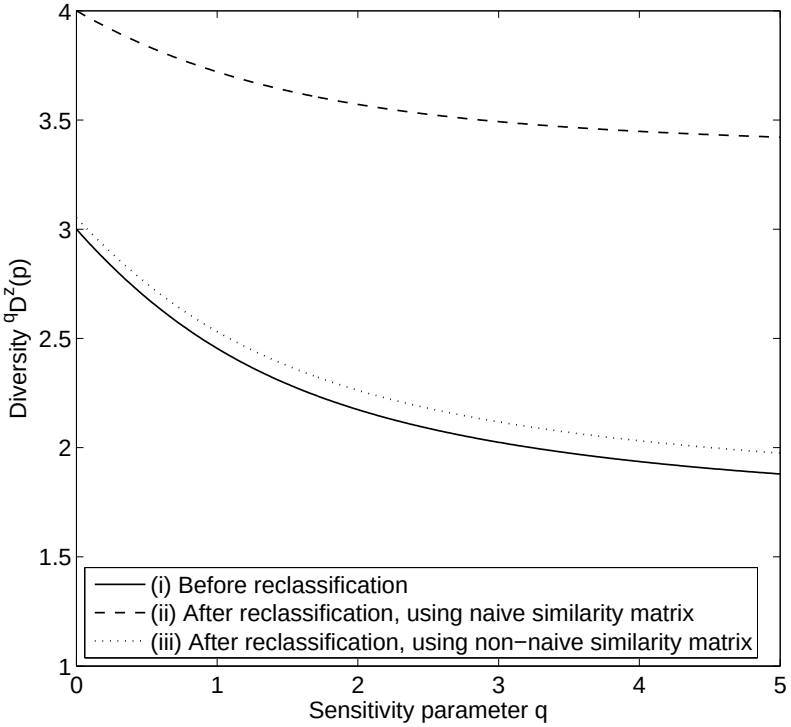

Figure 1: Diversity profiles of a hypothetical community, before and after taxonomic reclassification.

The identical species property is formulated mathematically in the Appendix (Proposition A16). It means that 'a community of 100 species that are identical in every way is no different from a community of only one species' (Ives 2007).

Uncontroversial as this may be, it has the important consequence that our measures are not oversensitive to decisions about taxonomy. Consider, for example, a system of 3 species with relative abundances $\mathbf{p} = (0.1, 0.3, 0.6)$, and with the species regarded as totally dissimilar. Suppose that on the basis of new genetic evidence, the last species is reclassified into two separate species of equal abundance, so that the relative abundances become $0.1, 0.3, 0.3, 0.3$.

Under the wholly unrealistic assumption that the two new species are totally dissimilar, the diversity profile jumps dramatically (Fig. 1). For example, the diversity of order ∞ jumps by 100%, from 1.67 to 3.33. But if, based on the genetic evidence, the two new species are given a high similarity, the diversity profile changes only slightly. Fig. 1 shows the profile with a similarity of $Z_{34} = Z_{43} = 0.9$ between the two new species (and $Z_{ij} = 0$ for $i \neq j$ otherwise).

This sensible behaviour is guaranteed by the identical species property. For if the two new species were deemed to be identical then, by that property, the profile would be unchanged. Since our measures are continuous, if the new species are deemed to be nearly identical then the profile is nearly unchanged.

3. Effect of species similarity on diversity

Monotonicity: when the similarities between species are increased, diversity decreases.

Naive model: when similarities between species are ignored, diversity is greater than when they are taken into account.

Range: the diversity of a community of S species is between 1 and S.

Monotonicity is formulated mathematically in Proposition A17 of the Appendix. It means that a community is more diverse when its species are more dissimilar. The naive model property is an extreme case: if a measure knows nothing of the commonalities between species, it will evaluate the community as more diverse than it really is. The naive model typically overestimates diversity.

This completes the list of fundamental properties satisfied by the diversity measures ${}^qD^{\mathbf{Z}}$. It is a logical consequence that they are also satisfied by the mean phylogenetic diversity of Chao et al. (2010) when the phylogenetic tree is ultrametric. This is proved in the Appendix (Proposition A7). For non-ultrametric trees, mean phylogenetic diversity can be greater than the number of species, contravening the naive model and range properties: see the supplement to Chao et al. (2010) and Example A20 in the Appendix.

The Ricotta–Szeidl entropies ${}^qH^{\mathbf{Z}}$ satisfy some of the properties, but not effective number, replication or range. This is a major advantage of the diversities ${}^qD^{\mathbf{Z}}$ over the entropies ${}^qH^{\mathbf{Z}}$.

5 Diversity profiles

We now have not just a single measure of a community's diversity, but a family of measures: ${}^qD^{\mathbf{Z}}(\mathbf{p})$, for each value of the sensitivity parameter q. The *diversity profile* of a community is the graph of ${}^qD^{\mathbf{Z}}(\mathbf{p})$ against q.

Diversity profiles convey a great deal of meaningful information. Although various other types of diversity profile have been discussed for decades (Hill 1973, Patil and Taillie 1979, 1982, Dennis and Patil 1986, Tóthmérész 1995, Patil 2002, Mendes et al. 2008), the idea has not achieved its full potential. When coupled with a model that takes species similarity properly into account, they are a powerful graphical tool for comparing ecological communities. The examples below show that one *should* draw the whole profile, rather than just calculating one or two indices. A diversity profile tells us more about ecological reality.

The left-hand end of a diversity profile gives information about species richness and rare species: when q is small, ${}^qD^{\mathbf{Z}}(\mathbf{p})$ is affected almost as much by rare species as common ones. The right-hand tail gives information about dominance and common species: when q is large, ${}^qD^{\mathbf{Z}}(\mathbf{p})$ is barely affected by rare species. For discussion in the naive case, see Whittaker (1972) and Hill (1973).

The sensitivity parameter q is, therefore, the *in*sensitivity to rare species. As it grows, the perceived diversity ${}^qD^{\mathbf{Z}}(\mathbf{p})$ drops. More precisely, the diversity profile is always a decreasing continuous curve (Appendix, Proposition A21).

In the first few examples, for the sake of exposition, we use the naive similarity matrix $\mathbf{Z} = \mathbf{I}$.

Example 1 Riegl et al. (2009) monitored coral cover on the Roatán fringing reef (western Caribbean) from 1996 to 2005. The diversity profiles for the first and last years, using the naive similarity matrix, are shown in Fig. 2(i). The profiles cross, so we cannot unambiguously say which of the two communities is the more diverse. An ecologist most

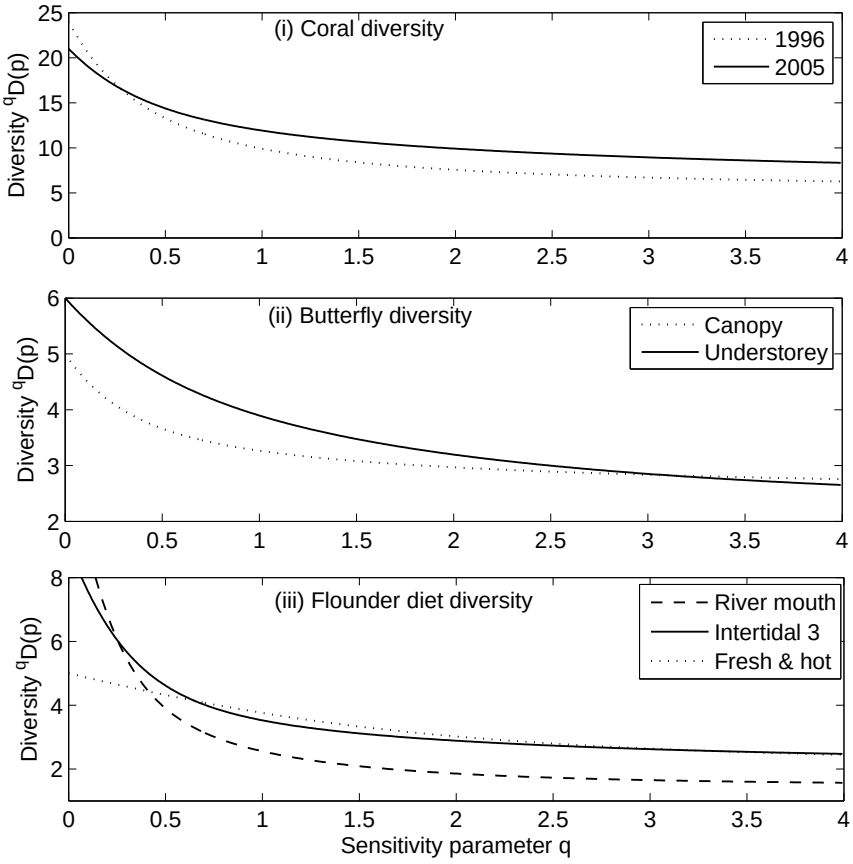

Figure 2: Diversity profiles using naive similarity matrix. (i) Coral: data from Table 4 of Riegl et al. (2009); (ii) butterflies of subfamily Nymphalinae: data from Table 5 of DeVries et al. (1997); (iii) flounder diet: data from Wirjoatmodjo (1980).

concerned with species richness would say that diversity had dropped; as the profiles show, 3 fewer species were observed in 2005 than 1996. But the profiles cross very far to the left ($q \approx 0.3$), so from almost any other point of view, the diversity increased. Indeed, for $q \geq 1$, the diversity of order q of the 2005 community is 2 to 3 species greater than in 1996. In short, the profiles indicate that the coral community became less rich in species but more even.

Individually, the two curves display properties typical of diversity profiles. Diversity tends to drop sharply between $q = 0$ and $q = 1$, levelling off soon after $q = 2$. (For this reason, the values of ${}^0D$, ${}^1D$, ${}^2D$ and ${}^\infty D$ usually give a good indication of the shape of the whole profile.) The abrupt drop of the 1996 curve in the region $0 \leq q \leq 1$ indicates that there were many rare species.

At the heart of Hurlbert's (1971) critique of the 'nonconcept' of diversity lay the observation that different diversity measures can make different judgements on which of two communities is the more diverse. All this means is that diversity profiles can cross.

In fact this happens frequently. For example, Table 1 of Ellingsen (2001) gives data on populations of soft-sediment macrobenthos at 16 sites on the Norwegian continental shelf. There are $1 + 2 + \cdots + 15 = 120$ pairs of sites, and the data show that for at least 53 of the 120 pairs, the profiles cross.

When the profile of one community is wholly above that of another, it can simply be called 'more diverse'. But when diversity profiles cross, the locations of the crossings give meaningful information about how the communities differ. Contrast Example 1 with the following.

Example 2 DeVries et al. (1997) surveyed butterfly populations in the canopy and understorey at a site in the Ecuadorian rainforest. The diversity profiles for the subfamily Nymphalinae, using the naive similarity matrix, are shown in Fig. 2(ii). In contrast to Example 1, the profiles cross at a high value of q (approximately 3.1). So if one is principally concerned with dominance, the population in the canopy appears to be fractionally more diverse, but from any other point of view, there is more diversity in the understorey.

In practice, diversity profiles do not usually cross more than once, although in principle there is no limit to the number of crossings. Figure 2(iii) shows diversity profiles for the diet of flounders (*Platichthys flesus*) at three different sites in an Irish estuary (Wirjoatmodjo 1980). One pair of profiles crosses twice, although with very small magnitude. The answer to the question 'where is flounder diet most diverse?' depends heavily on the sensitivity parameter: as q varies, the ranking of the three sites changes several times.

Patil and Taillie (1982) and Patil (2002) plotted diversity profiles using the entropies ${}^qH$ instead of the Hill numbers ${}^qD$. This conveys the same information, but, as Patil noted, often makes it hard to see where profiles cross. This is another advantage of effective numbers.

Diversity profiles give much more information than one or two diversity indices, but their biological relevance remains limited if they are used with the naive model. The following examples illustrate the effect of incorporating species similarity.

Example 3 Again we use the butterfly data of DeVries et al. (1997), this time taking the species of subfamily Charaxinae (Fig. 3(i)).

According to the naive model, the diversity profile of the canopy lies above that of the understorey until about $q = 5$, from which point they are almost identical. So for any sensitivity value, the canopy is more diverse than, or as diverse as, the understorey.

When no other species similarity data is available, one can fall back on taxonomy. Define a similarity matrix $\mathbf{Z}$ by

$$Z_{ij} = \begin{cases} 0 & \text{if the } i^{\text{th}} \text{ and } j^{\text{th}} \text{ species are of different genera} \\ 0.5 & \text{if the } i^{\text{th}} \text{ and } j^{\text{th}} \text{ species are different but congeneric} \\ 1 & \text{if } i = j. \end{cases}$$

The diversity profiles now tell a different story. For q greater than about 1, it is the understorey that is more diverse. It is easy to see why. Most of the population in the canopy is from the *Memphis* genus, whereas the understorey population is spread more evenly between genera. So when we build into the model the principle that species of the same genus tend to be somewhat similar, the canopy looks much less diverse than it did before.

All the diversity values drop when similarity is taken into account. This illustrates the 'naive model' property of the previous section.

Taxonomic models of this kind are certainly crude, and the similarity coefficient 0.5 was chosen arbitrarily. (Existing taxonomic models are just as arbitrary: e.g. Warwick and Clarke (1995), Shimatani (2001).) Some ecologists might prefer to stick to the naive model, ducking the question of how to choose the similarity coefficients. But to do so is to pretend that taxonomy says nothing about the commonalities and contrasts between species. It throws away relevant information.

Example 4 Turnbaugh et al. (2009) compared the microbial communities in the guts of lean and overweight humans. Here we compare the diversity profiles for two particular test subjects from that study, a lean child and an overweight mother. Since only a fraction of microbial species have been isolated and given taxonomic classifications, it is not possible to partition the microbes into species. Instead we work directly with DNA sequencing data, kindly supplied to us by Christopher Quince, treated with the noise removal algorithms described in the supplement of Turnbaugh et al. (2010).

Using the naive similarity matrix, the diversity profiles cross at $q \approx 1$ (Fig. 3(ii)). This suggests that the gut microbiome of the lean child has greater variety, but is less evenly distributed, than that of the overweight mother. However, using a genetic similarity matrix, the diversity in the lean child is seen to be greater for all values of q. This supports the results of Turnbaugh et al. (2009).

Example `R` code illustrating the calculation of diversity profiles is available at www.maths.gla.ac.uk/∼cc/supplements/diversity.html.

6 Discussion

We have described a general and biologically meaningful system for quantifying diversity. It is versatile enough to accommodate diversity of different types (functional, genetic, etc). It produces quantities that satisfy the ecologically intuitive properties of Section 4. The system acknowledges the spectrum of viewpoints on the relative importance of rare and common species. In this way, it meets a wide variety of ecological needs.

The field of diversity analysis has been criticized repeatedly over the years. We believe that our measures answer many of the criticisms. Let us consider some of them.

The varying differences between species are ignored. A basic fault of most diversity indices is that they behave as if different species had nothing in common. This has long been recognized: 'one would obviously regard [the diversity of a community] as greater if the species belonged to several genera than if they were all congeneric, and as greater still if these genera belonged to several families than if they were confamilial' (Pielou 1975). It is a glaringly obvious fault. Yet to this day, the most popular indices are wholly insensitive to the similarities between species.

There have been attempts to solve this problem. Some diversity indices depend only on species similarity, ignoring abundance (Faith 1992, Solow and Polasky 1994, Izsák and Papp 2000, Petchey and Gaston 2002). Rao's quadratic entropy takes both abundance and similarity into account, but represents a particular viewpoint on the relative importance of rare and common species. Ricotta and Szeidl's (2006) family of entropies ${}^{q}H^{\mathbf{Z}}$ allows for that viewpoint to be varied, by varying the parameter q; but their entropies suffer from not being effective numbers. The measures of Chao et al. (2010) are effective numbers, and do allow a varying q, but are particular to situations in which species similarity is derived from a tree (e.g. phylogenetic or taxonomic). We believe that ours is the first general system for quantifying diversity that takes species

Species	Canopy	Understorey
Prepona laertes	15	0
Archaeoprepona demophon	14	37
Zaretis itys	25	11
Memphis arachne	89	23
Memphis offa	21	3
Memphis xenocles	32	8

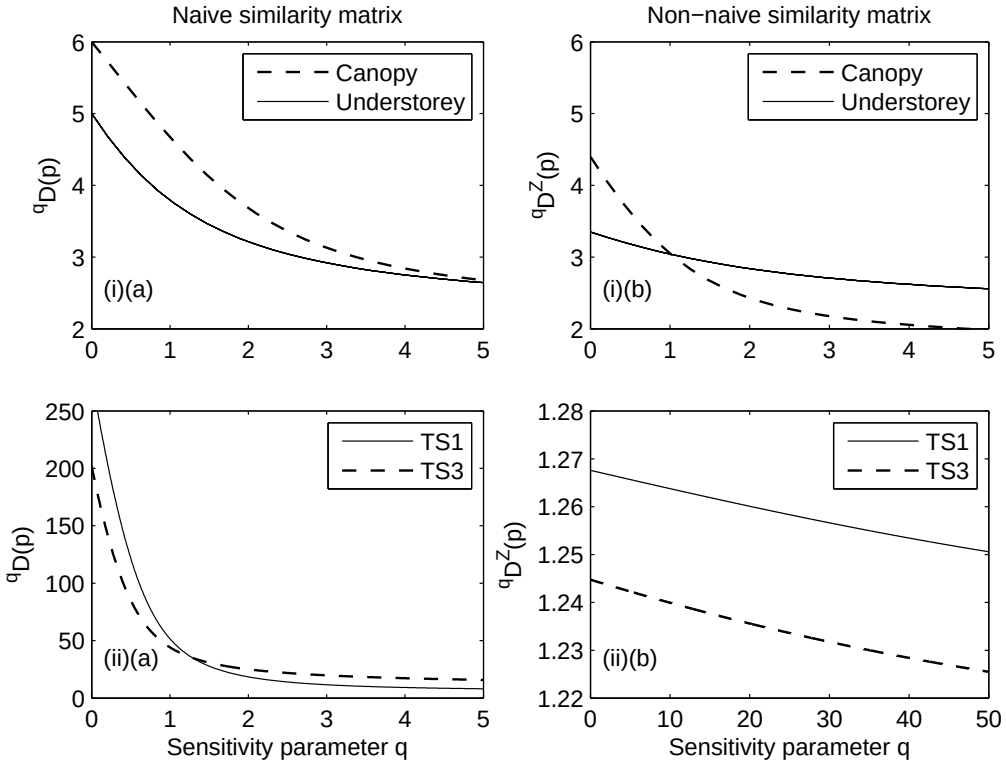

Figure 3: (i) Abundances of six butterfly species of subfamily Charaxinae, and their diversity profiles using (a) the naive similarity matrix, and (b) a taxonomic similarity matrix. Data from Table 5 of DeVries et al. (1997). (ii) Diversity profiles of the gut microbiomes in a lean child (TS1) and an overweight mother (TS3) (Turnbaugh et al. 2009), using (a) naive and (b) genetic similarity matrices. Note the different scales.

similarity into account, allows for any weighting of rare against common species, and produces an effective number.

The numbers produced by diversity indices are meaningless. This is often a fair criticism. Diversity indices that are *not* effective numbers can be very hard to interpret, and to speak of percentage changes in their value, or ratios between values, is perilous (Jost 2007). But our diversity measures ${}^qD^{\mathbf{Z}}$ are effective numbers, so their values have an intuitive interpretation and one can speak safely of percentage changes and ratios.

Diversity indices carry little information. A diversity index, being a single number, cannot carry much information; it must not be treated as a 'talisman' (Pielou 1975). This is why we advocate the use of diversity profiles. Diversity profiles allow for much more subtle inter-community comparison than a single number ever could. An ecologist who has enough information to calculate quadratic entropy (namely, a relative abundance vector $\mathbf{p}$ and a similarity matrix $\mathbf{Z}$) has enough information to graph the diversity profile, and should do so.

Earlier we called the diversity profile of a community its 'fingerprint'. There is a mathematical result justifying this, in the naive case at least: no two relative abundance vectors $\mathbf{p}$ have the same diversity profile, unless they consist of the same numbers $p_1, \ldots, p_S$ listed in different orders (Appendix, Proposition A22). This says that the diversity profile is simply the relative abundance data repackaged—and repackaged in a way that lets ecologists extract meaningful information at a glance.

Diversity indices depend too much on the notion of species. The division of living organisms into species is notoriously problematic. Conventional indices such as Shannon's, Simpson's and species richness depend wholly on this division, and behave badly in the face of taxonomic reclassification.

We have demonstrated two ways in which our measures answer this criticism. First, as shown in Section 4, our measures respond proportionately to changes in taxonomy. Second, we are able to measure diversity in situations where there is no clear division of organisms into species or other discrete units. We demonstrated this for microbes, and the method can be applied in other similar situations (e.g. soil types: McBratney and Minasny (2007)).

We also anticipate, and answer, a possible objection to our own diversity measures. In order to compute ${}^qD^{\mathbf{Z}}(\mathbf{p})$, one has to assign a similarity coefficient Z_{ij} to each pair of species. There is no canonical way to do this, so it might be objected that this makes the quantification of diversity too subjective.

Our answer is that diversity *is* subjective: it depends on which characteristics of organisms are taken to be important. This flexibility is, in fact, an advantage. If community A is genetically more diverse, but functionally less diverse, than community B, that is not a contradiction but a point of interest. Different ways of quantifying similarity lead to different measures of diversity. The word 'diversity' means little until one has specified the biological characteristics with which one is concerned.

Several methods for determining a similarity matrix $\mathbf{Z}$ have already been developed, principally in connection with Rao's quadratic entropy. Some are genetic (Hughes et al. 2008); others are functional (Botta-Dukát 2005, Petchey and Gaston 2006), taxonomic (Vane-Wright et al. 1991, Warwick and Clarke 1995, Shimatani 2001), morphological (Pavoine et al. 2005), or phylogenetic (Faith 1992, Hardy and Senterre 2007). Typically one begins by associating with each species some data embodying the characteristics deemed to be important: a list of functional traits, a DNA sequence, a location on a phylogenetic tree, etc. One then computes the similarity coefficients Z_{ij} in terms of some notion of difference between the associated data. There are as many possibilities

as there are quantifiable characteristics of living organisms.

Similarity and diversity vary according to perspective. Suppose, for example, that we are interested in the antigenic diversity of a collection of strains of the parasite *Plasmodium*. If similarity is measured using a nucleotide comparison of the entire genome then any two strains will look near-identical, giving the collection a very low diversity. But since we wish to measure *antigenic* diversity, we are really only concerned with the part of the genome that determines antigenicity. A nucleotide comparison localized to that region will reveal the sought-after differences, producing lower similarities and higher diversity.

The same question of perspective arises in other contexts. Chao et al. (2010) defined a measure of mean phylogenetic diversity since T years ago, which decreases as T increases. From the perspective of the history of all life on earth, all species of, say, eucalyptus look nearly identical, having diverged a relatively short time ago. Correspondingly, for large T the mean phylogenetic diversity of any eucalyptus community is very low. But if we wish to compare two eucalyptus communities, it would be sensible to take a smaller value of T, or, for a more complete picture, plot diversity against T (as in Fig. 3 of Chao et al. (2010)).

Similarly, inter-species distances $0 \leq d_{ij} \leq \infty$ can be transformed into similarities $0 \leq Z_{ij} \leq 1$ by putting $Z_{ij} = e^{-ud_{ij}}$, where the parameter u represents a choice of perspective. The complete picture emerges when we plot diversity against both q and u. (This transformation and the variation of the parameter u have deep mathematical roots (Leinster 2010).) Alternatively, we can choose a threshold $d_{\max}$ and define $Z_{ij} = 1 - d_{ij}/d_{\max}$, or $Z_{ij} = 0$ if $d_{ij} > d_{\max}$; this amounts to a piecewise linear approximation of the exponential transformation $e^{-(1/d_{\max})d_{ij}}$. Again, the choice of parameter $d_{\max}$ represents a choice of perspective on species similarity.

As a last resort, on the rare occasion that there is genuinely no information about species similarity—not even a taxonomic classification—one can use the naive model, $\mathbf{Z} = \mathbf{I}$. But the user should be aware that this represents an extreme assumption: distinct species have nothing whatsoever in common.

Every diversity index makes an assumption on the similarity of species. When no assumption is made explicit, there is invariably an implicit assumption of the naive model. For example, Shannon's and Simpson's indices use the naive model. To argue for the use of similarity-insensitive measures is to ignore the plain fact that some species are more similar than others. Deliberately ignoring biological reality is unlikely to lead to a helpful assessment of diversity.

At least two important questions remain. First, we have said very little about partitioning and α-, β- and γ-diversity. Jost (2007) (foreshadowed by Routledge (1979)) showed that in the naive context, there is no room for debate: if α- and β-diversity are to be independent, there is only one possible definition. (And in that context, $q = 1$ plays a special role.) It remains to extend this analysis to the similarity-sensitive context.

Second, we have deliberately avoided the evident statistical questions, preferring to separate the issue of principle (*what* are the meaningful quantities to measure?) from the issue of practice (*how* do we measure them?).

Our system of diversity measurement replaces a jumble of indices by a single formula. It behaves intuitively because it uses effective numbers. It allows for a nuanced comparison of communities because it produces diversity profiles, not just a single statistic. It provides a more faithful reflection of reality, because it takes into account the similarities between species. And it is highly versatile, since it allows similarity, hence diversity, to be measured in different ways according to ecologists' differing needs.

Acknowledgements We are very grateful to Anne Chao, Dan Haydon, Lou Jost, Louise Matthews and Richard Reeve for their helpful comments on earlier versions of this article. We also thank John Baez, David Corfield, André Joyal, Chris Quince, Urs Schreiber, Jiří Velebil and Simon Willerton. Leinster is supported by an EPSRC Advanced Research Fellowship, and thanks the School of Mathematics and Statistics at the University of Sheffield for its hospitality during part of this work.

References

Aczél, J. and Z. Daróczy, 1975. On Measures of Information and Their Characterizations. Academic, New York.

Adams, D. C., M. S. D. Bitetti, C. H. Janson, L. B. Slobodkin, and N. Valenzuela, 1997. An "audience effect" for ecological terminology: Use and misuse of jargon. *Oikos* **80**:632–636.

Allen, B., M. Kon, and Y. Bar-Yam, 2009. A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *The American Naturalist* **174**:236–243.

Berger, W. H. and F. L. Parker, 1970. Diversity of planktonic Foraminifera in deep-sea sediments. *Science* **168**:1345–1347.

Botta-Dukát, Z., 2005. Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *Journal of Vegetation Science* **16**:533–540.

Chakravarty, S. R. and W. Eichhorn, 1991. An axiomatic characterization of a generalized index of concentration. *The Journal of Productivity Analysis* **2**:103–112.

Chao, A., C.-H. Chiu, and L. Jost, 2010. Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B* **365**:3599–3609.

Dennis, B. and G. P. Patil, 1986. Profiles of diversity. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences, Vol. 7*, pages 292–296. John Wiley, New York.

DeVries, P. J., D. Murray, and R. Lande, 1997. Species diversity in vertical, horizontal and temporal dimensions of a fruit-feeding butterfly community in an Ecuadorian rainforest. *Biological Journal of the Linnean Society* **62**:343–364.

Ellingsen, K. E., 2001. Biodiversity of a continental shelf soft-sediment macrobenthos community. *Marine Ecology Progress Series* **218**:1–15.

Faith, D. P., 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61**:1–10.

Gini, C., 1912. Variabilità e mutabilità. Studi Economico-Giuridici della facoltà di Giurisprodenza dell' "Università" di Cagliari, III, parte II.

Hardy, G., J. E. Littlewood, and G. Pólya, 1952. Inequalities. Cambridge University Press, Cambridge, second edition.

Hardy, O. J. and B. Senterre, 2007. Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity. *Journal of Ecology* **95**:493–506.

Havrda, J. and F. Charvát, 1967. Quantification method of classification processes: Concept of structural α-entropy. *Kybernetika* **3**:30–35.

Hill, M. O., 1973. Diversity and evenness: A unifying notation and its consequences. *Ecology* **54**:427–432.

Hughes, A. R., B. D. Inouye, M. T. J. Johnson, N. Underwood, and M. Vellend, 2008. Ecological consequences of genetic diversity. *Ecology Letters* **11**:609–623.

Hurlbert, S. H., 1971. The nonconcept of species diversity: A critique and alternative parameters. *Ecology* **52**:577–586.

Ives, A. R., 2007. Diversity and stability in ecological communities. In R. M. May and A. R. McLean, editors, *Theoretical Ecology: Principles and Applications*. Oxford University Press, Oxford.

Izsák, J. and L. Papp, 1995. Application of the quadratic entropy indices for diversity studies of the drosophilid assemblages. *Environmental and Ecological Statistics* **2**:213–224.

Izsák, J. and L. Papp, 2000. A link between ecological diversity indices and measures of biodiversity. *Ecological Modelling* **130**:151–156.

Johnson, J. L., 1973. Use of nucleic-acid homologies in taxonomy of anaerobic bacteria. *International Journal of Systematic Bacteriology* **23**:308–315.

Jost, L., 2006. Entropy and diversity. *Oikos* **113**:363–375.

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

Jost, L., 2008. G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**:4015–4026.

Jost, L., 2009. Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). *Ecological Economics* **68**:925–928.

Leinster, T., 2010. The magnitude of metric spaces. Preprint arXiv:1012.5857, available from http://arXiv.org. Submitted.

MacArthur, R. H., 1965. Patterns of species diversity. *Biological Reviews* **40**:510–533.

McBratney, A. and B. Minasny, 2007. On measuring pedodiversity. *Geoderma* **141**:149–154.

Mendes, R. S., L. R. Evangelista, S. M. Thomaz, A. A. Agostinho, and L. C. Gomes, 2008. A unified index to measure ecological diversity and species rarity. *Ecography* **31**:450–456.

Mills, A. L. and R. A. Wassel, 1980. Aspects of diversity measurement for microbial communities. *Applied and Environmental Microbiology* **40**:578–586.

Nei, M., 1972. Genetic distance between populations. *The American Naturalist* **106**:283–292.

Nei, M. and F. Tajima, 1981. DNA polymorphism detectable by restriction endonucleases. *Genetics* **97**:145–163.

OECD, 2002. Handbook of Biodiversity Valuation: A Guide for Policy Makers. Organisation for Economic Co-operation and Development, Paris.

Patil, G. P., 2002. Diversity profiles. In A. H. El-Shaarawi and W. W. Piegorsch, editors, *Encyclopedia of Environmetrics*. John Wiley & Sons, Chichester, UK.

Patil, G. P. and C. Taillie, 1979. A study of diversity profiles and orderings for a bird community in the vicinity of Colstrip, Montana. In G. P. Patil and M. Rosenzweig, editors, *Contemporary Quantitative Ecology and Related Ecometrics*. International Co-operative Publishing House, Fairland, MD.

Patil, G. P. and C. Taillie, 1982. Diversity as a concept and its measurement. *Journal of the American Statistical Association* **77**:548–561.

Pavoine, S., S. Ollier, and D. Pontier, 2005. Measuring diversity from dissimilarities with Rao's quadratic entropy: Are any dissimilarities suitable? *Theoretical Population Biology* **67**:231–239.

Petchey, O. L. and K. J. Gaston, 2002. Functional diversity (FD), species richness and community composition. *Ecology Letters* **5**:402–411.

Petchey, O. L. and K. J. Gaston, 2006. Functional diversity: Back to basics and looking forward. *Ecology Letters* **9**:741–758.

Pielou, E. C., 1975. Ecological Diversity. John Wiley & Sons, New York.

Rao, C. R., 1982a. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology* **21**:24–43.

Rao, C. R., 1982b. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: The Indian Journal of Statistics* **44**:1–22.

Rényi, A., 1961. On measures of entropy and information. In J. Neymann, editor, *4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561. University of California Press.

Ricotta, C., 2005. Through the jungle of biological diversity. *Acta Biotheoretica* **53**:29–38.

Ricotta, C. and L. Szeidl, 2006. Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology* **70**:237–243.

Ricotta, C. and L. Szeidl, 2009. Diversity partitioning of Rao's quadratic entropy. *Theoretical Population Biology* **76**:299–302.

Riegl, B., S. J. Purkis, J. Keck, and G. P. Rowlands, 2009. Monitored and modeled coral

population dynamics and the refuge concept. *Marine Pollution Bulletin* **58**:24–38.
Routledge, R. D., 1979. Diversity indices: Which ones are admissible? *Journal of Theoretical Biology* **76**:503–515.
Shimatani, K., 2001. On the measurement of species diversity incorporating species differences. *Oikos* **93**:135–147.
Simpson, E. H., 1949. Measurement of diversity. *Nature* **163**:688.
Smith, W. and J. F. Grassle, 1977. Sampling properties of a family of diversity measures. *Biometrics* **33**:283–292.
Solow, A. R. and S. Polasky, 1994. Measuring biological diversity. *Environmental and Ecological Statistics* **1**:95–107.
Suyari, H., 2002. On the most concise set of axioms and the uniqueness theorem for Tsallis entropy. *Journal of Physics A: Mathematical and General* **35**:10731–10738.
Tóthmérész, B., 1995. Comparison of different methods for diversity ordering. *Journal of Vegetation Science* **6**:283–190.
Tsallis, C., 1988. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics* **52**:479–487.
Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon, 2009. A core gut microbiome in obese and lean twins. *Nature* **457**:480–484.
Turnbaugh, P. J., C. Quince, J. J. Faith, A. C. McHardy, T. Yatsunenko, F. Niazi, J. Affourtit, M. Egholm, B. Henrissat, R. Knight, and J. I. Gordon, 2010. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *PNAS* **107**:7503–7508.
Vane-Wright, R. I., C. J. Humphries, and P. H. Williams, 1991. What to protect?—Systematics and the agony of choice. *Biological Conservation* **55**:235–254.
Warwick, R. M. and K. R. Clarke, 1995. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series* **129**:301–305.
Watve, M. G. and R. M. Gangal, 1996. Problems in measuring bacterial diversity and a possible solution. *Applied and Environmental Microbiology* **62**:4299–4301.
Whittaker, R. H., 1972. Evolution and measurement of species diversity. *Taxon* **21**:213–251.
Wirjoatmodjo, S., 1980. Growth, Food and Movement of Flounder, *Platichthys flesus* (L.) in an Estuary. Ph.D. thesis, The New University of Ulster.

A Appendix: mathematical proofs

In this appendix, we write

$$\mathbb{P}_S = \{(p_1, \ldots, p_S) \in \mathbb{R}^S \mid p_i \geq 0, \sum p_i = 1\}$$

for the set of relative abundance vectors for S species. As usual, S denotes the *theoretical* number of species in the population. It may be that $p_i = 0$ for some values of i; we write $s \leq S$ for the number of values of i such that $p_i > 0$. A **similarity matrix** is an $S \times S$ matrix $\mathbf{Z}$ such that $0 \leq Z_{ij} \leq 1$ for all i and j, and $Z_{ii} = 1$ for all i.

In the definition

$$^qD^{\mathbf{Z}}(\mathbf{p}) = \left(\sum p_i (\mathbf{Zp})_i^{q-1}\right)^{\frac{1}{1-q}} \quad (q \neq 1, \infty),$$

the sum is over all $i = 1, \ldots, S$ such that $p_i \neq 0$. We now explain why the eventuality that some species are absent ($p_i = 0$) must be handled in this way.

We would like our measures of diversity to be continuous in $\mathbf{p}$, as far as possible. A small change in the abundance of a species should cause only a small change in the measured diversity. An exception is species richness (naive diversity of order 0): if the relative abundance of a species increases from 0 to 0.001, the species richness increases by 1. However, we will show that for all q with $0 < q < \infty$, diversity of order q is indeed continuous in $\mathbf{p}$.

When all S species are present, the sum in the definition of $^qD^{\mathbf{Z}}(\mathbf{p})$ is over all i from 1 to S. It follows easily that $^qD^{\mathbf{Z}}$ is continuous on the set

$$\mathbb{P}_S^\circ = \{\mathbf{p} \in \mathbb{P}_S \mid p_i > 0 \text{ for all } i\}.$$

Given a continuous function on $\mathbb{P}_S^\circ$, there is at most one way of extending it to a continuous function on the whole of $\mathbb{P}_S$. We show that $^qD^{\mathbf{Z}}$, defined as above, is indeed continuous on the whole of $\mathbb{P}_S$ (for $0 < q < \infty$). This implies that, once the definition has been decided for relative abundance vectors in which no p_i is zero, our formula is the *only* way of handling the case where one or more p_i is zero.

Proposition A1 *Let $0 < q < \infty$. Then the function $^qD^{\mathbf{Z}}$ on $\mathbb{P}_S$ is continuous.*

The delicacy of the proof arises from the possibility that $p_i = 0$ for some i: for then it may be that $(\mathbf{Zp})_i = 0$, and in that case $(\mathbf{Zp})_i^{q-1}$ is undefined for $q < 1$.

Proof When $1 < q < \infty$, the sum in the definition of $^qD^{\mathbf{Z}}(\mathbf{p})$ can equivalently be taken over *all* values of i from 1 to S. Then $^qD^{\mathbf{Z}}(\mathbf{p})$ is clearly continuous in $\mathbf{p}$.

Let $0 < q < 1$. Define functions $f_1, \ldots, f_S$ on $\mathbb{P}_S$ by

$$f_i(\mathbf{p}) = \begin{cases} p_i(\mathbf{Zp})_i^{q-1} & \text{if } p_i > 0 \\ 0 & \text{if } p_i = 0. \end{cases}$$

Since $^qD^{\mathbf{Z}}(\mathbf{p}) = (f_1(\mathbf{p}) + \cdots + f_S(\mathbf{p}))^{1/(1-q)}$, it suffices to prove that each f_i is continuous.

Certainly f_i is continuous on $\mathbb{P}_S^{(i)} = \{\mathbf{p} \in \mathbb{P}_S \mid p_i > 0\}$, so all we have to prove is that if $\mathbf{p} \in \mathbb{P}_S$ with $p_i = 0$, then $f_i(\mathbf{r}) \to 0$ as $\mathbf{r} \to \mathbf{p}$ in $\mathbb{P}_S$. Since $f_i(\mathbf{r}) = 0$ whenever $r_i = 0$, we might as well constrain $\mathbf{r}$ to lie in $\mathbb{P}_S^{(i)}$. We have $r_i \leq (\mathbf{Zr})_i \leq 1$, so

$$r_i^q = r_i \cdot r_i^{q-1} \geq r_i \cdot (\mathbf{Zr})_i^{q-1} \geq r_i \cdot 1 = r_i. \tag{A.1}$$

Now as $\mathbf{r} \to \mathbf{p}$ we have $r_i \to p_i = 0$, so $r_i^q \to 0$; hence $r_i(\mathbf{Zr})_i^{q-1} \to 0$ by (A.1), as required.

Finally, consider $q = 1$. It is enough to prove that $(\mathbf{Zp})_i^{p_i}$ is continuous in $\mathbf{p}$, for each i. (One evaluates 0^0 as 1.) Certainly it is continuous on $\mathbb{P}_S^{(i)}$, so all we have to prove is that if $\mathbf{p} \in \mathbb{P}_S$ with $p_i = 0$, then $(\mathbf{Zr})_i^{r_i} \to (\mathbf{Zp})_i^{p_i} = 1$ as $\mathbf{r} \to \mathbf{p}$ in $\mathbb{P}_S$. We have $r_i \leq (\mathbf{Zr})_i \leq 1$, so

$$r_i^{r_i} \leq (\mathbf{Zr})_i^{r_i} \leq 1^{r_i} = 1. \tag{A.2}$$

Observe also that $\lim_{x \to 0+} x^x = 1$. Now as $\mathbf{r} \to \mathbf{p}$ we have $r_i \to p_i = 0$, so we also have $r_i^{r_i} \to 1$; hence $(\mathbf{Zr})_i^{r_i} \to 1$ by (A.2), as required. □

Diversity of order 0 need not be continuous in $\mathbf{p}$ (depending on the similarity matrix). The same goes for diversity of order ∞. However, ${}^0D^{\mathbf{Z}}$ and ${}^\infty D^{\mathbf{Z}}$ do fit naturally into the family $({}^qD^{\mathbf{Z}})$ of diversity measures, in the sense made precise by the following proposition.

Proposition A2 *Let $\mathbf{p} \in \mathbb{P}_S$ and let $\mathbf{Z}$ be an $S \times S$ similarity matrix. Then:*

i. ${}^qD^{\mathbf{Z}}(\mathbf{p})$ *is continuous in q, for $0 < q < \infty$*

ii. $\lim_{q \to 0} {}^qD^{\mathbf{Z}}(\mathbf{p}) = {}^0D^{\mathbf{Z}}(\mathbf{p})$

iii. $\lim_{q \to \infty} {}^qD^{\mathbf{Z}}(\mathbf{p}) = {}^\infty D^{\mathbf{Z}}(\mathbf{p})$.

Proof All of this follows from standard results on generalized means (also called power means), which can be found in Hardy et al. (1952). Writing $x_i = (\mathbf{Zp})_i$, we have

$$1/{}^qD^{\mathbf{Z}}(\mathbf{p}) = \left(\sum_{i\colon p_i > 0} p_i x_i^{q-1} \right)^{1/(q-1)}$$

($q \neq 1$), which is the mean of order $q - 1$ of the family $(x_i)_{i\colon p_i > 0}$, weighted by the p_is. Similarly, $1/{}^1D^{\mathbf{Z}}(\mathbf{p})$ is the mean of order 0.

Parts (i) and (ii) are immediate except for continuity at $q = 1$, which follows from Theorem 3 of Hardy et al. Part (iii) follows from Theorem 4 of Hardy et al. □

Now consider the diversity ${}^qD^{\mathbf{Z}}(\mathbf{p})$ when q is an integer greater than 1. Let μ_q be the expected value of

$$Z_{i_1,i_2} Z_{i_1,i_3} \cdots Z_{i_1,i_q}$$

over all samples with replacement of q individuals from the community, whose respective species have been written as $i_1, i_2, \ldots, i_q$. Thus,

$$\mu_q = \sum_{i_1,i_2,\ldots,i_q} p_{i_1} p_{i_2} p_{i_3} \cdots p_{i_q} Z_{i_1,i_2} Z_{i_1,i_3} \cdots Z_{i_1,i_q}.$$

Proposition A3 *Let $q \geq 2$ be an integer. Then ${}^qD^{\mathbf{Z}}(\mathbf{p}) = \mu_q^{1/(1-q)}$.*

Proof Since $q > 1$, the sum in the definition of ${}^qD^{\mathbf{Z}}(\mathbf{p})$ might as well be over all $i = 1, \ldots, S$ (including those for which $p_i = 0$). We have

$$\begin{aligned} {}^qD^{\mathbf{Z}}(\mathbf{p})^{1-q} &= \sum_{i=1}^S p_i (\mathbf{Zp})_i^{q-1} = \sum_{i=1}^S p_i \left(\sum_{j=1}^S Z_{ij} p_j \right)^{q-1} \\ &= \sum_{i,j_1,\ldots,j_{q-1}} p_i Z_{i,j_1} p_{j_1} Z_{i,j_2} p_{j_2} \cdots Z_{i,j_{q-1}} p_{j_{q-1}} = \mu_q, \end{aligned}$$

as required. □

The next result states that our formula for ${}^qH^{\mathbf{Z}}(\mathbf{p})$ agrees with that of Ricotta and Szeidl (2006), except that they did not specify how their formula was to be interpreted in the case that some of the relative abundances p_i are 0. Recall from Section 3 of the main text that although Ricotta and Szeidl referred to their inter-species differences as 'distances' and denoted them by d_{ij}, we call them 'dissimilarities' and denote them by Δ_{ij}, since they are measured on a scale of 0 to 1.

Let $\mathbf{Z}$ be an $S \times S$ similarity matrix. Let $\mathbf{\Delta}$ be the corresponding dissimilarity matrix, defined by $\Delta_{ij} = 1 - Z_{ij}$.

Proposition A4 *For $0 \leq q < \infty$ and $\mathbf{p} \in \mathbb{P}_S$,*

$$ {}^qH^{\mathbf{Z}}(\mathbf{p}) = \begin{cases} \frac{1}{q-1}\left(1 - \sum_{i:\, p_i>0} p_i(1 - \sum_{j\neq i} \Delta_{ij}p_j)^{q-1}\right) & \text{if } q \neq 1 \\ -\sum_{i:\, p_i>0} p_i \ln(1 - \sum_{j\neq i} \Delta_{ij}p_j) & \text{if } q = 1. \end{cases} $$

Proof For $i = 1, \ldots, S$ we have

$$ (\mathbf{Zp})_i = \sum_{j=1}^{S} Z_{ij}p_j = \sum_{j=1}^{S}(1 - \Delta_{ij})p_j = 1 - \sum_{j=1}^{S} \Delta_{ij}p_j = 1 - \sum_{j\neq i} \Delta_{ij}p_j $$

since $\sum_{j=1}^{S} p_j = 1$ and $\Delta_{ii} = 1 - Z_{ii} = 0$. The result follows. □

Many authors have found it convenient to assume that the measure of dissimilarity or distance between species is a metric in the mathematical sense. For distances (d_{ij}) to define a **metric** means that (i) $d_{ij} = 0$ if and only if $i = j$; (ii) $d_{ij} = d_{ji}$; and (iii) the triangle inequality holds: $d_{ij} + d_{jk} \geq d_{ik}$.

Now, it may be that we started with inter-species distances d_{ij} measured on a scale of 0 to ∞ and converted them to similarities Z_{ij} by the formula $Z_{ij} = e^{-ud_{ij}}$, where u is a positive constant. These in turn correspond to dissimilarities $\Delta_{ij} = 1 - Z_{ij}$, measured on a scale of 0 to 1. Whether we are using (d_{ij}) or $\mathbf{\Delta} = (\Delta_{ij})$ matters: asking that (d_{ij}) defines a metric is not the same as asking that (Δ_{ij}) defines a metric, as the following proposition shows.

Proposition A5 *If (d_{ij}) is a metric then (Δ_{ij}) is a metric, but the converse implication fails.*

Proof The quantities d_{ij} and Δ_{ij} are related by $\Delta_{ij} = 1 - e^{-ud_{ij}}$, or equivalently, $d_{ij} = -(1/u)\ln(1 - \Delta_{ij})$. It is easy to see that axioms (i) and (ii) hold for d if and only if they hold for $\mathbf{\Delta}$. For the triangle inequality (iii),

$$ \begin{aligned} d_{ij} + d_{jk} \geq d_{ik} &\iff -\ln(1 - \Delta_{ij}) - \ln(1 - \Delta_{jk}) \geq -\ln(1 - \Delta_{ik}) \\ &\iff (1 - \Delta_{ij})(1 - \Delta_{jk}) \leq 1 - \Delta_{ik} \\ &\iff \Delta_{ij} + \Delta_{jk} - \Delta_{ij}\Delta_{jk} \geq \Delta_{ik}. \end{aligned} $$

This implies the triangle inequality for $\mathbf{\Delta}$, but not conversely: e.g. $S = 3$, $\Delta_{12} = \Delta_{23} = 1/2$, $\Delta_{13} = 4/5$. □

The next few results concern the relationships between our measures ${}^qD^{\mathbf{Z}}$ and other indices of diversity.

First we compare our measures with some measures of phylogenetic diversity. All of the latter are based on phylogenetic trees, some very simple examples of which are shown in Fig. A1. From a phylogenetic tree we extract the following information:

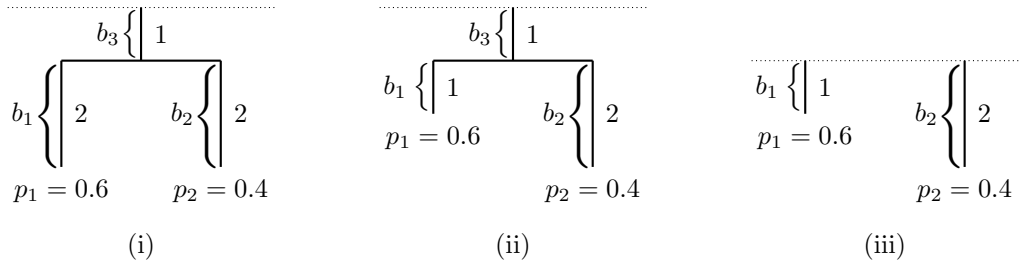

Figure A1: Three simple phylogenetic trees, each with two present-day species. The dotted horizontal lines show the beginning of the time period under consideration. In (i), the tree has three branches, b_1, b_2, b_3, with $L(b_1) = L(b_2) = 2$ and $L(b_3) = 1$. Tree (i) is ultrametric; trees (ii) and (iii) are not. In trees (i) and (ii), the present-day species have a common ancestor in the time period considered; in tree (iii), they do not.

- the number S of present-day species (which we put in order: $1, 2, \ldots, S$)
- the set of branches
- for each branch b, its length $L(b)$
- for each branch b, the set $I_b \subseteq \{1, 2, \ldots, S\}$ of present-day species descended from b.

We make the convention that the variable i always ranges over the present-day species $1, 2, \ldots, S$, and the variable b always ranges over the set of branches.

Before making the comparison, we review the phylogenetic measures concerned. The simplest is that of Faith (1992), which is just the total branch length,

$$\sum_b L(b).$$

Suppose now that we have a relative abundance vector $\mathbf{p} = (p_1, \ldots, p_S)$ for the present-day species. For each branch b, write

$$p(b) = \sum_{i:\ i \in I_b} p_i,$$

which is the total relative abundance of present-day species descended from b. For each present-day species i, write

$$L_i = \sum_{b:\ i \in I_b} L(b),$$

which is the total evolutionary change undergone by the ith species over the time-span considered. For the tree to be **ultrametric** means that $L_1 = L_2 = \cdots = L_S$.

Chao et al. (2010) write

$$\overline{T} = \sum_i p_i L_i = \sum_{i,b:\ i \in I_b} p_i L(b) = \sum_b p(b) L(b) \tag{A.3}$$

for the mean evolutionary change per present-day species. This reflects the time-span under consideration. When the tree is ultrametric, $\overline{T} = L_1 = \cdots = L_S$. For $0 \leq q < \infty$,

they define the **mean phylogenetic diversity** of order q as

$${}^q\overline{D}(\overline{T}) = \begin{cases} \Big(\sum_b \dfrac{L(b)}{\overline{T}} p(b)^q\Big)^{1/(1-q)} & \text{if } q \neq 1 \\ \prod_b p(b)^{-(L(b)/\overline{T})p(b)} & \text{if } q = 1. \end{cases}$$

The expression at $q = 1$ is the limit as $q \to 1$ of ${}^q\overline{D}(\overline{T})$. Although they do not mention it, there is also a limit as $q \to \infty$, namely

$${}^\infty\overline{D}(\overline{T}) = 1/\max_b p(b).$$

When the present-day species have a common ancestor in the time-span considered, the tree has a root b; then $p(b) = 1$ and so ${}^\infty\overline{D}(\overline{T}) = 1$.

The phylogenetic entropy of Allen, Kon, and Bar-Yam (2009) is

$$H^{\mathrm{AKB}}(\mathbf{p}) = -\sum_b L(b)p(b)\ln p(b).$$

They also implicitly propose a phylogenetic entropy of each order $q \geq 0$,

$${}^qH^{\mathrm{AKB}}(\mathbf{p}) = \sum_b L(b)p(b)\sigma_q(p(b))$$

where σ_q is the surprise function of Section 3. For example, ${}^1H^{\mathrm{AKB}} = H^{\mathrm{AKB}}$.

Chao et al. (2010) showed that Faith's measure and H^{AKB} can be recovered from their measures. We now connect our measures to Chao et al.'s, Faith's, and ${}^qH^{\mathrm{AKB}}$ (for any q). To do this, we begin by showing how, from the data given (a phylogenetic tree and relative abundances for the present-day species), we can extract a matrix $\mathbf{Z}$ and a relative abundance vector $\boldsymbol{\pi}$.

Since we are considering the evolution of species through time, our basic biological units (which would usually be called 'species') are not present-day species, but species *in a particular period of their evolutionary history*. That is, a unit is a pair (i, b) where b is a branch and $i \in I_b$. We call such a pair a **historical species**. Its relative abundance $\pi_{(i,b)}$ is weighted according to how great a portion of evolutionary time it occupies:

$$\pi_{(i,b)} = \frac{L(b)}{\overline{T}} p_i. \tag{A.4}$$

(Equation (A.3) implies that $\sum_{i,b:\ i\in I_b} \pi_{(i,b)} = 1$.) The matrix $\mathbf{Z}$ is defined by

$$Z_{(i,b),\,(j,c)} = \begin{cases} \overline{T}/L_j & \text{if } j \in I_b \\ 0 & \text{otherwise.} \end{cases} \tag{A.5}$$

We offer no intuitive interpretation of this last formula. But intuitive or not, this formula provides a strong logical connection (Proposition A7) between our diversity measures and the measures of phylogenetic diversity mentioned above.

If the tree is ultrametric then $\overline{T}/L_j$ is always 1, so the matrix $\mathbf{Z}$ consists entirely of 0s and 1s. It is not symmetric, but *is* a similarity matrix in the precise sense defined in the main text and at the beginning of this appendix. This demonstrates that non-symmetric similarity matrices can serve a useful purpose.

If the tree is not ultrametric then matters are more delicate. First, the matrix $\mathbf{Z}$ depends on $\overline{T} = \sum_i p_i L_i$, which in turn depends on the relative abundances $\mathbf{p}$ of the present-day species. (There is no such dependence when the tree is ultrametric.) When the entries of $\mathbf{Z}$ depend on species abundances, it can no longer be thought of as a 'similarity' matrix in the same way. Second, $\mathbf{Z}$ is *not* a similarity matrix in the precise sense, since some of its entries are strictly greater than 1. But this turns out to cause no mathematical difficulty. The situation is clarified by introducing a new piece of terminology, as follows.

Let us say that a **relatedness matrix** is a real square matrix $\mathbf{Y}$ such that $Y_{ij} \geq 0$ for all i, j and $Y_{ii} > 0$ for all i. Certainly every similarity matrix is a relatedness matrix, but not vice versa. Observe that for any relatedness matrix $\mathbf{Y}$ and relative abundance vector $\mathbf{r}$, if $r_1 > 0$ then $(\mathbf{Yr})_1 > 0$: for

$$(\mathbf{Yr})_1 = Y_{11} r_1 + \sum_{j=2}^{S} Y_{1j} r_j \geq Y_{11} r_1 > 0.$$

Similarly, for any $i \in \{1, \ldots, n\}$, if $r_i > 0$ then $(\mathbf{Yr})_i > 0$; so $(\mathbf{Yr})_i^{q-1}$ is a well-defined real number (even if $q < 1$). The definitions of the diversities ${}^qD^{\mathbf{Y}}(\mathbf{r})$ and entropies ${}^qH^{\mathbf{Y}}(\mathbf{r})$ therefore make mathematical sense for an arbitrary relatedness matrix $\mathbf{Y}$. The phylogenetic matrix $\mathbf{Z}$ in equation (A.5) is always a relatedness matrix. It only satisfies the stronger condition of being a similarity matrix if the tree is ultrametric. But it is a relatedness matrix in any case, so ${}^qD^{\mathbf{Z}}(\boldsymbol{\pi})$ and ${}^qH^{\mathbf{Z}}(\boldsymbol{\pi})$ are always mathematically well-defined quantities.

Section 4 of the main text contains a list of important properties satisfied by the diversity measures ${}^qD^{\mathbf{Z}}$. Most of them hold for an arbitrary relatedness matrix $\mathbf{Z}$. The stronger assumption that $\mathbf{Z}$ is a similarity matrix is only needed for the naive model and range properties, as shown in Propositions A9–A17. So the notion of relatedness matrix widens the scope of our results. It is also useful because it allows us to prove new results about the phylogenetic measures of Chao et al. (2010) (Corollary A12).

Example A6 In the ultrametric tree of Fig. A1(i), there are four historical species: $(1, b_1)$, $(1, b_3)$, $(2, b_2)$ and $(2, b_3)$. We have $\overline{T} = 2 + 1 = 3$, and

$$\mathbf{Z} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \qquad \boldsymbol{\pi} = \begin{pmatrix} \frac{2}{3} \times 0.6 \\ \frac{1}{3} \times 0.6 \\ \frac{2}{3} \times 0.4 \\ \frac{1}{3} \times 0.4 \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.2 \\ 0.267 \\ 0.133 \end{pmatrix}.$$

This $\mathbf{Z}$ is a similarity matrix, depending only on the structure of the phylogenetic tree (and independent of the species abundances). As observed above, these properties of $\mathbf{Z}$ are guaranteed by the fact that the tree is ultrametric.

In the non-ultrametric tree of Fig. A1(ii), there are the same four historical species (i, b). We have $L_1 = 1 + 1 = 2$, $L_2 = 2 + 1 = 3$, and $\overline{T} = 0.6 \times L_1 + 0.4 \times L_2 = 2.4$, giving

$$\mathbf{Z} = \begin{pmatrix} 1.2 & 1.2 & 0 & 0 \\ 1.2 & 1.2 & 0.8 & 0.8 \\ 0 & 0 & 0.8 & 0.8 \\ 1.2 & 1.2 & 0.8 & 0.8 \end{pmatrix}, \qquad \boldsymbol{\pi} = \begin{pmatrix} \frac{1}{2.4} \times 0.6 \\ \frac{1}{2.4} \times 0.6 \\ \frac{2}{2.4} \times 0.4 \\ \frac{1}{2.4} \times 0.4 \end{pmatrix} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.333 \\ 0.167 \end{pmatrix}.$$

This $\mathbf{Z}$ is a relatedness matrix but not a similarity matrix, and it does depend on the abundances of the present-day species.

The following result makes the connection between our measures and the phylogenetic measures of Chao et al. (2010), of Faith (1992), and of Allen et al. (2009).

Proposition A7 *Take a phylogenetic tree and a relative abundance vector for the present-day species. Then, defining* $\mathbf{Z}$ *and* $\boldsymbol{\pi}$ *as in equations* (A.4) *and* (A.5) *above,*

i. ${}^qD^{\mathbf{Z}}(\boldsymbol{\pi}) = {}^q\overline{D}(\overline{T})$*, the mean phylogenetic diversity of Chao et al. (2010), for all* $0 \leq q \leq \infty$*. In particular,* ${}^0D^{\mathbf{Z}}(\boldsymbol{\pi})$ *is* $\frac{1}{\overline{T}}$ *times Faith's phylogenetic diversity.*

ii. ${}^qH^{\mathbf{Z}}(\boldsymbol{\pi}) = \frac{1}{\overline{T}} \times {}^qH^{AKB}(\mathbf{p})$*, the phylogenetic entropy of Allen et al. (2009), for all* $0 \leq q < \infty$*. In particular,* ${}^1H^{\mathbf{Z}}(\boldsymbol{\pi}) = \frac{1}{\overline{T}} \times H^{AKB}(\mathbf{p})$.

So where Chao et al.'s measures are diversities, Allen et al.'s are (up to a multiplicative factor) the accompanying entropies.

Proof First we compute $\mathbf{Z}\boldsymbol{\pi}$. For each historical species (i, b),

$$\begin{aligned}(\mathbf{Z}\boldsymbol{\pi})_{(i,b)} &= \sum_{j,c\colon j\in I_c} Z_{(i,b),(j,c)}\pi_{(j,c)} = \sum_{j,c\colon j\in I_b\cap I_c} \frac{\overline{T}}{L_j}\frac{L(c)}{\overline{T}}p_j \\ &= \sum_{j\colon j\in I_b} \frac{p_j}{L_j} \sum_{c\colon j\in I_c} L(c) = \sum_{j\colon j\in I_b} p_j = p(b).\end{aligned}$$

It suffices to prove (i) and (ii) when $q \neq 1, \infty$, by Proposition A2 (which holds for arbitrary relatedness matrices, by exactly the same proof). For (i),

$$\begin{aligned}{}^qD^{\mathbf{Z}}(\boldsymbol{\pi}) &= \left(\sum_{\substack{i,b\colon i\in I_b,\\ \pi_{(i,b)}>0}} \pi_{(i,b)}(\mathbf{Z}\boldsymbol{\pi})_{(i,b)}^{q-1}\right)^{1/(1-q)} = \left(\sum_{\substack{i,b\colon i\in I_b,\\ p_i>0}} \frac{L(b)}{\overline{T}}p_i \cdot p(b)^{q-1}\right)^{1/(1-q)} \\ &= \left(\sum_b \frac{L(b)}{\overline{T}}p(b)^q\right)^{1/(1-q)} = {}^q\overline{D}(\overline{T}).\end{aligned}$$

In particular, ${}^0D^{\mathbf{Z}}(\boldsymbol{\pi}) = (1/\overline{T})\sum_b L(b)$.

For (ii),

$$\begin{aligned}{}^qH^{\mathbf{Z}}(\boldsymbol{\pi}) &= \sum_{\substack{i,b\colon i\in I_b,\\ \pi_{(i,b)}>0}} \pi_{(i,b)}\sigma_q\big((\mathbf{Z}\boldsymbol{\pi})_{(i,b)}\big) = \sum_{\substack{i,b\colon i\in I_b,\\ p_i>0}} \frac{L(b)}{\overline{T}}p_i \cdot \sigma_q(p(b)) \\ &= \sum_b \frac{L(b)}{\overline{T}}p(b)\cdot\sigma_q(p(b)) = \frac{1}{\overline{T}} \times {}^qH^{\mathrm{AKB}}(\mathbf{p}),\end{aligned}$$

as required. □

Next we turn to the indices studied by Hurlbert (1971) and Smith and Grassle (1977), showing that they can be derived from the naive diversities (Hill numbers) ${}^qD$. For each $m \geq 2$, let $H_m^{\mathrm{HSG}}(\mathbf{p})$ be the expected number of species represented in a random sample of m individuals. As Hurlbert observed,

$$H_m^{\mathrm{HSG}}(\mathbf{p}) = \sum_{i=1}^{S}(1 - (1-p_i)^m).$$

Proposition A8 *For each $m \geq 2$, the Hurlbert–Smith–Grassle index is given by*

$$H_m^{HSG}(\mathbf{p}) = m - \sum_{q=2}^{m} (-1)^q \binom{m}{q} {}^qD(\mathbf{p})^{1-q},$$

where $\binom{m}{q}$ is the binomial coefficient $m!/q!(m-q)!$.

Proof We have

$$\begin{aligned}
\sum_{i=1}^{S} (1 - (1-p_i)^m) &= S - \sum_{i=1}^{S} \sum_{q=0}^{m} \binom{m}{q} (-p_i)^q = S - \sum_{q=0}^{m} (-1)^q \binom{m}{q} \sum_{i=1}^{S} p_i^q \\
&= S - \left\{ \binom{m}{0} S - \binom{m}{1} 1 + \sum_{q=2}^{m} (-1)^q \binom{m}{q} {}^qD(\mathbf{p})^{1-q} \right\} \\
&= m - \sum_{q=2}^{m} (-1)^q \binom{m}{q} {}^qD(\mathbf{p})^{1-q},
\end{aligned}$$

as required. □

We now prove the properties of our diversity measures stated in Section 4.

> *Our standing assumption for the rest of this appendix is that matrices called* $\mathbf{Z}$*,* $\mathbf{Z}(i)$*, etc. are relatedness matrices. They are not required to be similarity matrices except where this is stated explicitly.*

We observe that although Proposition A2 is stated for similarity matrices, it holds for arbitrary relatedness matrices, by exactly the same proof.

Proposition A9 (Effective number) *Let $0 \leq q \leq \infty$. Then diversity of order q is an effective number; that is, if $p_1 = \cdots = p_S = 1/S$ then ${}^qD^{\mathbf{I}}(\mathbf{p}) = S$.*

Proof This follows immediately from the definition of ${}^qD^{\mathbf{Z}}(\mathbf{p})$, substituting $\mathbf{Z} = \mathbf{I}$ and $\mathbf{p} = (1/S, \ldots, 1/S)$. □

Suppose now that the community is divided into m subcommunities. No species appears in more than one subcommunity, and species in different subcommunities are totally dissimilar.

Write $w_1, \ldots, w_m$ for the relative sizes of the subcommunities (so that $\sum w_i = 1$). Within the i^{th} subcommunity, write S_i for the number of species, $\mathbf{r}_i = (r_{i1}, \ldots, r_{iS_i})$ for its relative abundance vector (so that $\sum_{j=1}^{S_i} r_{ij} = 1$), and $\mathbf{Z}(i)$ for its matrix (an $S_i \times S_i$ matrix).

Write S for the number of species in the whole community, $\mathbf{p}$ for the overall relative abundance vector, and $\mathbf{Z}$ for the overall, $S \times S$, matrix.

Proposition A10 (Modularity) *For $0 \leq q \leq \infty$,*

$${}^qD^{\mathbf{Z}}(\mathbf{p}) = \begin{cases} \left(\sum_{i:\, w_i > 0} w_i^q d_i^{1-q} \right)^{\frac{1}{1-q}} & \text{if } q \neq 1, \infty \\ {}^1D(\mathbf{w}) d_1^{w_1} d_2^{w_2} \cdots d_m^{w_m} & \text{if } q = 1 \\ \min_{i:\, w_i > 0} (d_i / w_i) & \text{if } q = \infty \end{cases}$$

where $d_i = {}^qD^{\mathbf{Z}(i)}(\mathbf{r}_i)$.

Proof We have

$$
\begin{aligned}
&S = S_1 + \cdots + S_m,\\
&\mathbf{p} = (w_1 r_{11}, \ldots, w_1 r_{1S_1}, \ \ldots \ , w_m r_{m1}, \ldots, w_m r_{mS_m}),\\
&\mathbf{Z} = \begin{pmatrix} \mathbf{Z}(1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}(2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}(m) \end{pmatrix},
\end{aligned}
$$

the last expression being a block sum of matrices. So for $1 \leq i \leq m$ and $1 \leq j \leq S_i$,

$$(\mathbf{Z}\mathbf{p})_{S_1 + \cdots + S_{i-1} + j} = w_i (\mathbf{Z}(i)\mathbf{r}_i)_j.$$

Hence for $q \neq 1, \infty$, letting i range over $1, \ldots, m$ and j range over $1, \ldots, S_i$,

$$
\begin{aligned}
{}^q D^{\mathbf{Z}}(\mathbf{p})^{1-q} &= \sum_{i,j \colon w_i r_{ij} > 0} w_i r_{ij} \big(w_i (\mathbf{Z}(i)\mathbf{r}_i)_j\big)^{q-1} \\
&= \sum_{i \colon w_i > 0} w_i^q \sum_{j \colon r_{ij} > 0} r_{ij} (\mathbf{Z}(i)\mathbf{r}_i)_j^{q-1} = \sum_{i \colon w_i > 0} w_i^q d_i^{1-q},
\end{aligned}
$$

as required. For $q = 1$,

$$
\begin{aligned}
{}^1 D^{\mathbf{Z}}(\mathbf{p}) &= \prod_{i,j} \big(w_i (\mathbf{Z}(i)\mathbf{r}_i)_j\big)^{-w_i r_{ij}} \\
&= \left(\prod_{i,j} w_i^{-w_i r_{ij}}\right) \left(\prod_{i,j} (\mathbf{Z}(i)\mathbf{r}_i)_j^{-w_i r_{ij}}\right) = {}^1 D(\mathbf{w}) \prod_i d_i^{w_i}
\end{aligned}
$$

since $\sum_j r_{ij} = 1$ for each i. Finally, for $q = \infty$,

$$
\begin{aligned}
{}^\infty D^{\mathbf{Z}}(\mathbf{p}) &= 1 \Big/ \max_{i,j \colon w_i r_{ij} > 0} w_i (\mathbf{Z}(i)\mathbf{r}_i)_j = 1 \Big/ \max_{i \colon w_i > 0} \Big(w_i \max_{j \colon r_{ij} > 0} (\mathbf{Z}(i)\mathbf{r}_i)_j \Big) \\
&= 1 \Big/ \max_{i \colon w_i > 0} (w_i / d_i) = \min_{i \colon w_i > 0} (d_i / w_i),
\end{aligned}
$$

as required. □

Proposition A11 (Replication) *If $w_1 = \cdots = w_m$ and $d_1 = \cdots = d_m = d$ then ${}^q D^{\mathbf{Z}}(\mathbf{p}) = md$ for all $0 \leq q \leq \infty$.*

Proof Substitute $w_i = 1/m$ and $d_i = d$ into Proposition A10. □

We can deduce some facts about the mean phylogenetic diversity of Chao et al. (2010), extending the replication principle that they proved.

Corollary A12 *Let $0 \leq q \leq \infty$. Take m completely distinct phylogenetic assemblages, with relative sizes $w_1, \ldots, w_m$. Write $\overline{T}_i$ for the mean evolutionary change per species in the ith assemblage, $d_i = {}^q\overline{D}(\overline{T}_i)$ for its mean phylogenetic diversity of order q, and $\overline{T}$ for the mean evolutionary change per species in the pooled assemblage.*

i. If $\overline{T}_1 = \cdots = \overline{T}_m$ then the mean phylogenetic diversity ${}^q\overline{D}(\overline{T})$ of the pooled assemblage is determined by $w_1, \ldots, w_m$ and $d_1, \ldots, d_m$ via the formula in Proposition A10.

ii. If $w_1 = \cdots = w_m = 1/m$ and $d_1 = \cdots = d_m = d$ then the mean phylogenetic diversity ${}^q\overline{D}(\overline{T})$ of the pooled assemblage is md.

Chao et al. (2010) proved (ii) under the further assumption that the m assemblages all have the same mean evolutionary change per species; but this assumption is unnecessary.

Proof We begin the proof without assuming either the hypothesis in (i) or that in (ii). We use the notation set up after Proposition A5.

Write $\mathbf{Z}(i)$ for the matrix (A.5) of the ith assemblage, and $\mathbf{Z}$ for the matrix (A.5) of the pooled assemblage. Then $\mathbf{Z}$ is the block sum

$$\mathbf{Z} = \begin{pmatrix} (\overline{T}/\overline{T}_1)\mathbf{Z}(1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & (\overline{T}/\overline{T}_2)\mathbf{Z}(2) & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & (\overline{T}/\overline{T}_m)\mathbf{Z}(m) \end{pmatrix}.$$

Write $\boldsymbol{\sigma}(i)$ for the relative abundance vector (A.4) of the ith assemblage, and $\boldsymbol{\pi}$ for the relative abundance vector (A.4) of the pooled assemblage. Then whenever a present-day species j belongs to the ith assemblage and is descended from a branch b,

$$\pi_{(j,b)} = \frac{\overline{T}_i}{\overline{T}} w_i \cdot \sigma(i)_{(j,b)}.$$

Note also that $\overline{T} = \sum_{i=1}^m w_i \overline{T}_i$.

We now apply Proposition A10. For $q \neq 1, \infty$, it gives

$$\begin{aligned} {}^qD^{\mathbf{Z}}(\boldsymbol{\pi}) &= \left(\sum_{i\colon w_i > 0} \left(\frac{\overline{T}_i}{\overline{T}} w_i \right)^q \left({}^qD^{(\overline{T}/\overline{T}_i)\mathbf{Z}(i)}(\boldsymbol{\sigma}(i)) \right)^{1-q} \right)^{1/(1-q)} \\ &= \left(\sum_{i\colon w_i > 0} \left(\frac{\overline{T}_i}{\overline{T}} w_i \right)^q \left(\frac{\overline{T}_i}{\overline{T}} \, {}^qD^{\mathbf{Z}(i)}(\boldsymbol{\sigma}(i)) \right)^{1-q} \right)^{1/(1-q)} \\ &= \left(\sum_{i\colon w_i > 0} \frac{\overline{T}_i}{\overline{T}} w_i^q \left({}^qD^{\mathbf{Z}(i)}(\boldsymbol{\sigma}(i)) \right)^{1-q} \right)^{1/(1-q)} \end{aligned}$$

where in the second step, we have used the easily verified fact that when a matrix $\mathbf{Z}$ is multiplied by a constant $\lambda > 0$, the resulting diversity (of any order) is divided by λ. So by Proposition A7,

$${}^q\overline{D}(\overline{T}) = \left(\sum_{i\colon w_i > 0} \frac{\overline{T}_i}{\overline{T}} w_i^q d_i^{1-q} \right)^{1/(1-q)}. \tag{A.6}$$

To prove (i), suppose that $\overline{T}_1 = \cdots = \overline{T}_m$: then $\overline{T} = \overline{T}_i$ for all i, so (A.6) gives

$${}^q\overline{D}(\overline{T}) = \left(\sum_{i\colon w_i > 0} w_i^q d_i^{1-q} \right)^{1/(1-q)},$$

and similarly for $q = 1, \infty$. To prove (ii), suppose instead that $w_1 = \cdots = w_m = 1/m$ and $d_1 = \cdots = d_m = d$: then $\overline{T} = (1/m)\sum \overline{T}_i$, so (A.6) gives

$${}^q\overline{D}(\overline{T}) = \left(\sum_{i=1}^m \frac{\overline{T}_i}{\overline{T}} \left(\frac{1}{m} \right)^q d^{1-q} \right)^{1/(1-q)} = md,$$

and similarly for $q = 1, \infty$, as required. □

The 'elementary properties' all follow from a single general lemma. For this we need some notation. Take $S, T \geq 1$ and a function $\theta\colon \{1, \ldots, S\} \to \{1, \ldots, T\}$. For each $\mathbf{r} \in \mathbb{P}_S$, define $\theta \cdot \mathbf{r} \in \mathbb{P}_T$ by

$$(\theta \cdot \mathbf{r})_j = \sum_{i\colon \theta(i)=j} r_i$$

($j \in \{1, \ldots, T\}$), where the sum is over all $i \in \{1, \ldots, S\}$ such that $\theta(i) = j$. For each $T \times T$ matrix $\mathbf{Z}$, define an $S \times S$ matrix $\mathbf{Z} \cdot \theta$ by

$$(\mathbf{Z} \cdot \theta)_{ii'} = Z_{\theta(i),\theta(i')}$$

($i, i' \in \{1, \ldots, S\}$).

Lemma A13 *${}^qD^{\mathbf{Z}\cdot\theta}(\mathbf{r}) = {}^qD^{\mathbf{Z}}(\theta \cdot \mathbf{r})$, for all $0 \leq q \leq \infty$.*

Proof By continuity of ${}^qD^{\mathbf{Z}}$ in q (Proposition A2), it is enough to prove this when $q \neq 1, \infty$. We will use the convention that indices i, i' range over $1, \ldots, S$ and indices j, j' range over $1, \ldots, T$.

We have

$$\begin{aligned}
((\mathbf{Z} \cdot \theta)\mathbf{r})_i &= \sum_{i'} (\mathbf{Z} \cdot \theta)_{ii'} r_{i'} = \sum_{i'} Z_{\theta(i),\theta(i')} r_{i'}, \\
(\mathbf{Z}(\theta \cdot \mathbf{r}))_j &= \sum_{j'} Z_{jj'} (\theta \cdot \mathbf{r})_{j'} = \sum_{j'} \sum_{i'\colon \theta(i')=j'} Z_{jj'} r_{i'} = \sum_{i'} Z_{j,\theta(i')} r_{i'},
\end{aligned}$$

so

$$((\mathbf{Z} \cdot \theta)\mathbf{r})_i = (\mathbf{Z}(\theta \cdot \mathbf{r}))_{\theta(i)}.$$

Hence

$$\begin{aligned}
{}^qD^{\mathbf{Z}\cdot\theta}(\mathbf{r})^{1-q} &= \sum_{i\colon r_i>0} r_i ((\mathbf{Z} \cdot \theta)\mathbf{r})_i^{q-1} = \sum_{i\colon r_i>0} r_i (\mathbf{Z}(\theta \cdot \mathbf{r}))_{\theta(i)}^{q-1} \\
&= \sum_{j\colon (\theta\cdot\mathbf{r})_j>0} \sum_{i\colon \theta(i)=j} r_i (\mathbf{Z}(\theta \cdot \mathbf{r}))_j^{q-1} \\
&= \sum_{j\colon (\theta\cdot\mathbf{r})_j>0} (\theta \cdot \mathbf{r})_j (\mathbf{Z}(\theta \cdot \mathbf{r}))_j^{q-1} = {}^qD^{\mathbf{Z}}(\theta \cdot \mathbf{r})^{1-q}.
\end{aligned}$$

The result follows. □

The three elementary properties of diversity can be deduced. In each, q may take any value in the range $0 \leq q \leq \infty$.

Proposition A14 (Symmetry) *Let θ be a permutation of $\{1, \ldots, S\}$, let $\mathbf{Z}$ be an $S \times S$ matrix, and let $\mathbf{p} \in \mathbb{P}_S$. Define $\mathbf{Z}'$ and $\mathbf{p}'$ by $Z'_{ij} = Z_{\theta(i),\theta(j)}$ and $p'_i = p_{\theta(i)}$. Then ${}^qD^{\mathbf{Z}'}(\mathbf{p}') = {}^qD^{\mathbf{Z}}(\mathbf{p})$.*

Proof In the notation above, we have $\mathbf{Z}' = \mathbf{Z} \cdot \theta$ and $\mathbf{p} = \theta \cdot \mathbf{p}'$, since

$$(\theta \cdot \mathbf{p}')_j = \sum_{i\colon \theta(i)=j} p'_i = \sum_{i\colon \theta(i)=j} p_{\theta(i)} = p_j.$$

Hence by Lemma A13,

$${}^qD^{\mathbf{Z}'}(\mathbf{p}') = {}^qD^{\mathbf{Z}\cdot\theta}(\mathbf{p}') = {}^qD^{\mathbf{Z}}(\theta \cdot \mathbf{p}') = {}^qD^{\mathbf{Z}}(\mathbf{p}),$$

as required. □

Proposition A15 (Absent species) *Let* $\mathbf{Z}$ *be an* $S \times S$ *matrix and let* $\mathbf{p} \in \mathbb{P}_S$ *with* $p_S = 0$. *Write* $\mathbf{Z}'$ *for the restriction of* $\mathbf{Z}$ *to the first* $(S-1)$ *species; that is,* $\mathbf{Z}$ *is the* $(S-1) \times (S-1)$ *matrix given by* $Z'_{ij} = Z_{ij}$. *Write* $\mathbf{p}' = (p_1, \ldots, p_{S-1}) \in \mathbb{P}_{S-1}$. *Then* ${}^qD^{\mathbf{Z}'}(\mathbf{p}') = {}^qD^{\mathbf{Z}}(\mathbf{p})$.

Proof Let $\theta\colon \{1, \ldots, S-1\} \to \{1, \ldots, S\}$ be the embedding $\theta(i) = i$. Then $\mathbf{Z}' = \mathbf{Z} \cdot \theta$ and $\mathbf{p} = \theta \cdot \mathbf{p}'$, so the result follows as in the previous proof. □

Proposition A16 (Identical species) *Let* $\mathbf{Z}$ *be an* $S \times S$ *matrix such that* $Z_{i,S} = Z_{i,S-1}$ *and* $Z_{S,i} = Z_{S-1,i}$ *for all* i. *Let* $\mathbf{p} \in \mathbb{P}_S$. *Write* $\mathbf{Z}'$ *for the restriction of* $\mathbf{Z}$ *to the first* $(S-1)$ *species, and define* $\mathbf{p}' \in \mathbb{P}_{S-1}$ *by*

$$p'_j = \begin{cases} p_j & \text{if } j < S-1 \\ p_{S-1} + p_S & \text{if } j = S-1. \end{cases}$$

Then ${}^qD^{\mathbf{Z}'}(\mathbf{p}') = {}^qD^{\mathbf{Z}}(\mathbf{p})$.

Proof Define a function $\theta\colon \{1, \ldots, S\} \to \{1, \ldots, S-1\}$ by

$$\theta(i) = \begin{cases} i & \text{if } i \leq S-1 \\ S-1 & \text{if } i = S. \end{cases}$$

Then $\mathbf{Z} = \mathbf{Z}' \cdot \theta$ and $\mathbf{p}' = \theta \cdot \mathbf{p}$. The result follows from Lemma A13. □

The final properties to be proved are those from group 3: 'effect of species similarity on diversity'.

Proposition A17 (Monotonicity) *Let* $\mathbf{Z}$ *and* $\mathbf{Z}'$ *be* $S \times S$ *matrices with* $Z_{ij} \leq Z'_{ij}$ *for all* i, j. *Then* ${}^qD^{\mathbf{Z}}(\mathbf{p}) \geq {}^qD^{\mathbf{Z}'}(\mathbf{p})$, *for all* $0 \leq q \leq \infty$ *and* $\mathbf{p} \in \mathbb{P}_S$.

Proof By continuity in q (Proposition A2), it is enough to prove this when $q \neq 1, \infty$. We have $(\mathbf{Zp})_i \leq (\mathbf{Z'p})_i$ for all i.

If $0 \leq q < 1$ then x^{q-1} is decreasing in $x > 0$ and $y^{1/(1-q)}$ is increasing in $y > 0$. Hence

$$\begin{aligned} (\mathbf{Zp})_i \leq (\mathbf{Z'p})_i \text{ for all } i &\implies (\mathbf{Zp})_i^{q-1} \geq (\mathbf{Z'p})_i^{q-1} \text{ for all } i \text{ such that } p_i > 0 \\ &\implies \sum_{i\colon p_i > 0} p_i (\mathbf{Zp})_i^{q-1} \geq \sum_{i\colon p_i > 0} p_i (\mathbf{Z'p})_i^{q-1} \\ &\implies {}^qD^{\mathbf{Z}}(\mathbf{p}) \geq {}^qD^{\mathbf{Z}'}(\mathbf{p}'). \end{aligned}$$

If $q > 1$ then x^{q-1} is increasing in x and $y^{1/(1-q)}$ is decreasing in y, and a similar argument applies. □

Proposition A18 (Naive model) *Let* $\mathbf{Z}$ *be a similarity matrix,* $\mathbf{p} \in \mathbb{P}_S$, *and* $0 \leq q \leq \infty$. *Then* ${}^qD(\mathbf{p}) \geq {}^qD^{\mathbf{Z}}(\mathbf{p})$.

Proof We have $I_{ij} = 0 \leq Z_{ij}$ for all $i \neq j$, and $I_{ii} = 1 = Z_{ii}$ for all i, so ${}^qD(\mathbf{p}) = {}^qD^{\mathbf{I}}(\mathbf{p}) \geq {}^qD^{\mathbf{Z}}(\mathbf{p})$ by Proposition A17. □

Proposition A19 (Range) *Let* $\mathbf{Z}$ *be a similarity matrix,* $\mathbf{p} \in \mathbb{P}_S$, *and* $0 \leq q \leq \infty$. *Then* $1 \leq {}^qD^{\mathbf{Z}}(\mathbf{p}) \leq S$.

Proof Let $\mathbf{Y}$ be the similarity matrix with $Y_{ij} = 1$ for all i, j. Then $Z_{ij} \leq Y_{ij}$ for all i, j, so ${}^qD^{\mathbf{Z}}(\mathbf{p}) \geq {}^qD^{\mathbf{Y}}(\mathbf{p})$ by Proposition A17. But ${}^qD^{\mathbf{Y}}(\mathbf{p}) = 1$ by repeated application of Proposition A16, or by direct calculation. Hence ${}^qD^{\mathbf{Z}}(\mathbf{p}) \geq 1$.

By Proposition A18, we have ${}^qD^{\mathbf{Z}}(\mathbf{p}) \leq {}^qD(\mathbf{p})$. But the Hill number ${}^qD(\mathbf{p})$ is maximized at the uniform distribution $\mathbf{p} = (1/S, \ldots, 1/S)$, where, being an effective number, it takes the value S. Hence ${}^qD^{\mathbf{Z}}(\mathbf{p}) \leq S$. □

In the last two properties (naive model and range), we explicitly assumed that $\mathbf{Z}$ was a similarity matrix, not merely a relatedness matrix. The following example shows that they can fail without that assumption.

Example A20 In Proposition A7, the mean phylogenetic diversity of Chao et al. (2010) was expressed as ${}^qD^{\mathbf{Z}}(\boldsymbol{\pi})$ for a suitable matrix $\mathbf{Z}$ and vector $\boldsymbol{\pi}$ (equations (A.4) and (A.5)). When the phylogenetic tree is ultrametric, $\mathbf{Z}$ is a similarity matrix; hence, mean phylogenetic diversity for ultrametric trees satisfies all the properties above. But when the tree is not ultrametric, the mean phylogenetic diversity can be greater than the number of species, contrary to the naive model and range properties. An example is given in the supplement to Chao et al. (2010). Another example is the tree of Fig. A1(iii). There,

$$\overline{T} = 0.6 \times 1 + 0.4 \times 2 = 1.4,$$

and the mean phylogenetic diversity of order 0 is

$$\frac{1}{\overline{T}} \times (\text{total branch length}) = \frac{1}{1.4} \times 3 = 2.142\ldots,$$

which is greater than 2, the number of species. Similarly, short calculations show that the mean phylogenetic diversities of orders 1 and 2 are also greater than the number of species.

We finish with some facts about diversity profiles.

Proposition A21 *Let $\mathbf{Z}$ be an $S \times S$ matrix and $\mathbf{p} \in \mathbb{P}_S$. Then ${}^qD^{\mathbf{Z}}(\mathbf{p})$ is decreasing in q. That is, whenever $0 \leq q \leq q' \leq \infty$, we have ${}^qD^{\mathbf{Z}}(\mathbf{p}) \geq {}^{q'}D^{\mathbf{Z}}(\mathbf{p})$.*

It may be that the diversity profile is constant: e.g. if $\mathbf{Z} = \mathbf{I}$ and $\mathbf{p} = (1/S, \ldots, 1/S)$ then ${}^qD^{\mathbf{Z}}(\mathbf{p}) = S$ for all q.

Proof This follows from standard results on generalized means: Theorems 5 and 16 of Hardy et al. (1952). □

The final result states that if two communities have the same naive diversity profiles then their relative abundance vectors are the same, except perhaps for the order in which the relative abundances p_i are listed.

Proposition A22 *Let $\mathbf{p}, \mathbf{p}' \in \mathbb{P}_S$ and suppose that ${}^qD(\mathbf{p}) = {}^qD(\mathbf{p}')$ for all $0 < q < \infty$. Then $(p'_1, \ldots, p'_S)$ is a permutation of $(p_1, \ldots, p_S)$.*

Proof We prove by induction on S that if $\mathbf{p}, \mathbf{p}' \in \mathbb{P}_S$ with $p_1 \leq \cdots \leq p_S$, $p'_1 \leq \cdots \leq p'_S$ and ${}^qD(\mathbf{p}) = {}^qD(\mathbf{p}')$ for all $q > 1$, then $\mathbf{p} = \mathbf{p}'$. Clearly this holds for $S = 1$. Now let $S \geq 2$, and write $f(q) = {}^qD(\mathbf{p}) = {}^qD(\mathbf{p}')$.

We have $p_S = 1/\lim_{q\to\infty} f(q) = p'_S$. If $p_S = p'_S = 1$ then $\mathbf{p} = \mathbf{p}' = (0, \dots, 0, 1)$. Otherwise, we may define $\mathbf{r}, \mathbf{r}' \in \mathbb{P}_{S-1}$ by

$$\mathbf{r} = \left(\frac{p_1}{1 - p_S}, \dots, \frac{p_{S-1}}{1 - p_S} \right)$$

and similarly $\mathbf{r}'$. Then for all $q > 1$,

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

But since $p_S = p'_S$, this expression is also equal to ${}^qD(\mathbf{r}')$. Hence by inductive hypothesis, $\mathbf{r} = \mathbf{r}'$; that is, $p_i = p'_i$ for all $i < S$. □