Methods in Ecology and Evolution

*Methods in Ecology and Evolution* 2014, **5**, 452–461 doi: 10.1111/2041-210X.12173

# A new frontier in biodiversity inventory: a proposal for estimators of phylogenetic and functional diversity

**Pedro Cardoso[1,2]*, François Rigal[2], Paulo A. V. Borges[2] and José C. Carvalho[2,3]**

[1]*Finnish Museum of Natural History, University of Helsinki, POBox 17 (Pohjoinen Rautatiekatu 13), 00014 Helsinki, Finland;* [2]*Azorean Biodiversity Group (CITA-A) and Portuguese Platform for Enhancing Ecological Research & Sustainability (PEERS), University of the Azores, 9700-042 Angra do Heroísmo, Portugal; and* [3]*CBMA – Molecular and Environmental Centre, Department of Biology, University of Minho, 4710-057 Braga, Portugal*

## Summary

**1.** Complete sampling of all dimensions of biodiversity is a formidable task, even for small areas. Undersampling is the norm, and the underquantification of diversity is a common outcome. Estimators of taxon diversity (TD) are widely used to correct for undersampling. Yet, no similar strategy has been developed for phylogenetic (PD) or functional (FD) diversity.

**2.** We propose three ways of estimating PD and FD, building on estimators originally developed for TD: (i) correcting PD and FD values based on the completeness of TD; (ii) fitting asymptotic functions to accumulation curves of PD and FD; and (iii) adapting nonparametric estimators to PD and FD data.

**3.** Using trees as a common framework for the estimation of PD and FD, we tested the approach with European mammal and Azores Islands arthropod data. We demonstrated that different methods were able to considerably reduce the undersampling bias and often correctly estimated true diversity using a fraction of the samples necessary to reach complete sampling.

**4.** Besides the utility of knowing the true diversity of an assemblage from incomplete samples, the use of estimators may present further advantages. For instance, comparisons between sites or time periods are possible only if either sampling is complete or sampling effort is equivalent and sufficient to allow sensible comparisons. Also, as PD and FD asymptote faster than TD, comparisons between these different dimensions may require unbiased values. The framework now proposed combines taxon, phylogenetic and functional diversity into a single framework, offering a tool for future developments involving these different facets of biological diversity.

**Key-words:** accumulation curves, alpha diversity, arthropods, Azores, European mammals, extrapolation, functional diversity, nonparametric estimators, phylogenetic diversity, sampling bias

## Introduction

From scales as small as a single tree, which can house thousands of species (Erwin 1982), to the entire planet, which is home to millions (Mora *et al.* 2011), we are always far from being able to count every single species from most groups. Even studies restricted in space, time and taxonomic scope face intractable problems when trying to reach complete lists of species (Cardoso 2009; Coddington *et al.* 2009). As species richness increases with the number of individuals or samples, observed richness almost invariably underestimates true richness (Coddington *et al.* 2009).

The problem of undersampling and consequent bias in diversity descriptors has long been recognised as both ubiquitous and fundamental to correct. Without complete inventories of communities, comparisons of species richness cannot be reliably made if one does not consider the sampling effort or completeness attained (Gotelli & Colwell 2001). Many researchers address this concern by estimating alpha diversity using techniques that adjust for sampling effort or completeness (Longino, Colwell & Coddington 2002). The same problems have been identified for beta diversity, since underestimation of similarity also occurs because of the failure to account for unseen shared species (Chao *et al.* 2000, 2005; Cardoso, Borges & Veech 2009). However, the same attention given to taxon diversity (TD) bias due to undersampling and how to correct it has never been given to other facets of biodiversity, namely phylogenetic diversity (PD) and functional diversity (FD).

Taxon diversity is the most common measure of biodiversity, usually expressed in terms of species richness where alpha or gamma diversity is concerned. However, underlying the use of TD is the simplistic assumption that the taxa are equally distinct from one another, disregarding the fact that communities are composed of species with different evolutionary histories and a diverse array of ecological functions. Thus, the last decade has seen a growing interest in complementary representations of biodiversity, including PD and FD (Devictor *et al.* 2010; Cardoso *et al.* 2014).

*Correspondence author. E-mail: pedro.cardoso@helsinki.fi

Phylogenetic diversity takes evolutionary relationships between taxa into account (Faith 1992) and reflects how much evolutionary history is behind the species constituting the communities. Assemblages with identical TD may be considerably different with respect to their evolutionary past, depending on how far the species diverge from their nearest common ancestor (Webb *et al.* 2002; Graham & Fine 2008). PD has been measured based on phylogenetic trees or cladograms, reflecting the amount of phylogenetic information conveyed by the assemblage (Faith 1992). Related measures reflecting the degree of (un)relatedness of taxa have also been proposed (Webb *et al.* 2002; but see Helmus *et al.* 2007 for alternative measures).

Functional diversity quantifies components of biodiversity that influence how an ecosystem operates or functions (Tilman *et al.* 2001). We refer here to FD as the amount of biological functions or traits displayed by the species occurring in given assemblages. This takes into account that species often overlap in their traits, and as such, their relations may be depicted much in the same way as species are related in a phylogenetic framework. Communities with completely different species composition may be characterised by low variation in functional traits, with unrelated species replacing others with similar roles in the network. FD has also been quantified in many different ways, such as quadratic entropy (Rao 1982), dendrogram-based measures (Petchey & Gaston 2002, 2006) or the functional hypervolume occupied by taxa (Cornwell, Schwilk & Ackerly 2006; Villéger, Mason & Mouillot 2008).

Despite the wide recognition that TD is often underestimated and that estimators are frequently needed to correct for bias, few studies have been made to verify whether the same problems are present in PD or FD measures (Walker, Poos & Jackson 2008; Ricotta *et al.* 2012) and, more importantly, no estimators have been proposed to date. In this study, we focus on the estimation of the alpha component of TD, PD and FD, although beta diversity will probably require a similar approach. Three main strategies have been followed in the past to estimate species richness (TD) from incomplete samples (Soberón & Llorente 1993; Colwell & Coddington 1994; Longino, Colwell & Coddington 2002; Gotelli & Colwell 2011): (i) fitting the log-normal distribution to species abundance data and estimating the part of the distribution to the left of the veil line; (ii) fitting asymptotic curves to randomised accumulation curves; and (iii) using nonparametric estimators based on the abundance or incidence of rare species. After hundreds of published tests with many different data sets, some of these alternatives were found to perform generally better (although still far from perfect), namely the latter two (Longino, Colwell & Coddington 2002; Walther & Moore 2005). In this study, we adapt some of the mostly used and best-performing species richness (i.e. TD) estimators to PD and FD measures and test their accuracy and ability to correct for undersampling bias with a variety of theoretical and empirical data sets. We show that different methods can be successfully used for estimating PD and FD, with a wide application to all kinds of organisms and sampling schemes.

## Materials and methods

Two kinds of data may be used for estimating diversity: (i) abundance data, which may or may not be divided by sample, and (ii) incidence data, that is, the presence or absence of each species in each sample. From such data, a number of approaches for estimating true diversity are possible.

### ESTIMATION OF TAXON DIVERSITY

Here, we will focus on two of the mostly used and promising methods. If indeed PD and FD accumulate with effort (often a measure of space or time) in the same fashion as TD, then the approaches used to estimate TD will be useful, albeit with possible adaptations, for PD and FD.

### *Fitting asymptotic curves*

Usually sampling is not instantaneous, that is, features (species – TD, clades – PD and traits – FD) accumulate with the increasing number of individuals or samples (Fig. 1). If at first the accumulation is fast, as most features are yet to be sampled, the accumulation rate constantly decreases, as an increasing number of features are already sampled and it becomes harder to find novelty. Because the accumulation is not random, but made in a stepwise way, after randomisation of the accumulation process, it is possible to fit asymptotic curves to the data points and calculate the asymptote, which should approximate the true total diversity (observed plus non-observed).

One of the mostly used equations is the Clench or Michaelis–Menten curve:

$$S_{\mathrm{obs}} = \frac{aQ}{1 + bQ}$$

where $S_{\mathrm{obs}}$ = observed richness, $Q$ = number of samples ($x$-axis) and $a$ and $b$ are fitting parameters. After finding these parameters, the asymptote, that is, estimated richness ($S^*$), is:

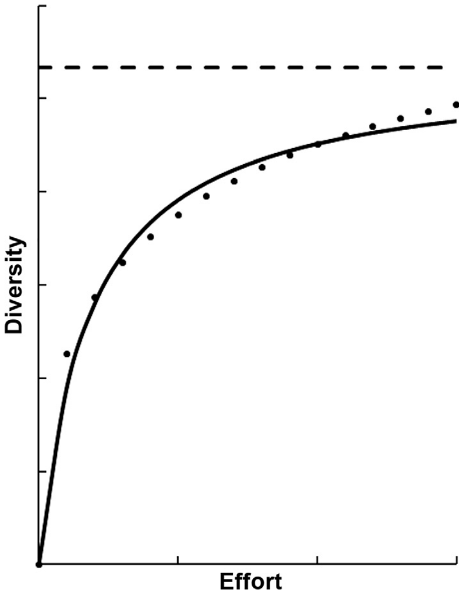

**Fig. 1.** Randomised accumulation of diversity with increasing effort (dots), a fitted asymptotic equation (continuous line) and its asymptote representing true diversity (dashed line).

$$S^* = \frac{a}{b}$$

Other equations have been tested with variable success (e.g. negative exponential, Weibull; Soberón & Llorente 1993; Flather 1996).

### Nonparametric estimators

Nonparametric estimators, based on the abundance or incidence of rare species, are often considered the best option for estimating species richness with most data sets (Gotelli & Colwell 2011). Among the first to be suggested for this purpose and still the mostly used are the jackknife (Heltshe & Forrester 1983) and Anne Chao's formulas (Chao 1984, 1987). The jackknife 1 based on abundance data is:

$$S^* = S_{\text{obs}} + S_1$$

where $S_1$ = singletons, that is, the number of species known from a single individual in the samples.

Its incidence data equivalent further includes a correction factor and is:

$$S^* = S_{\text{obs}} + Q_1\left(\frac{Q-1}{Q}\right)$$

where $Q_1$ = uniques, that is, the number of species known from a single sample.

The Chao 1 (Chao 1984) is also based on the abundance of rare species:

$$S^* = S_{\text{obs}} + \frac{S_1(S_1-1)}{2(S_2+1)}$$

where $S_2$ = doubletons, that is, the number of species known from exactly two individuals in the samples. A corresponding incidence estimator is known as Chao 2 (Chao 1987) and is based on the presence of rare species in samples:

$$S^* = S_{\text{obs}} + \frac{Q_1(Q_1-1)}{2(Q_2+1)}$$

where $Q_2$ = duplicates, that is, the number of species known from exactly two samples.

Recently, Lopez *et al.* (2012) have proposed a correction factor based on the percentage of singletons or uniques in the sampled universe, with the reasoning that this proportion ($P$) is directly correlated with undersampling. Irrespective of the function, its correction for abundance estimators is:

$$S^*P = S^*\left(1 + \left(\frac{S_1}{S_{\text{obs}}}\right)^2\right)$$

and for incidence estimators:

$$S^*P = S^*\left(1 + \left(\frac{Q_1}{S_{\text{obs}}}\right)^2\right)$$

Many other nonparametric estimators have been proposed to date, but here we adapt these eight (four formulas, with and without $P$ correction) with the same reasoning being easily translated to other methods.

## MEASURES OF PHYLOGENETIC AND FUNCTIONAL DIVERSITY

Phylogenetic diversity and FD sensu lato can be expressed in a large variety of ways (Mouchet *et al.* 2010; Schleuter *et al.* 2010). All measures may be divided into two main groups. First, those that reflect overall diversity or the amount of evolutionary history for PD and the amount of functions for FD contained in a given community, which may be called phylogenetic and functional richness, respectively. These tend to increase with increasing number of taxa, as new taxa tend to provide new branches in the phylogenetic or functional tree or occupy new space in the functional hypervolume. Second, there are measures that reflect average distance between taxa, which may be called phylogenetic or functional dispersion or differentiation (Webb *et al.* 2002; Laliberté & Legendre 2010). These are often not influenced by the number of taxa, being largely insensitive to sampling effort (Weiher, Clarke & Keddy 1998).

Given that estimates of PD and FD are especially useful for the first group of measures and that our study aims to estimate PD and FD in comparison with TD as measured by species richness, we will focus on measures that capture the notion of richness. Therefore, in the following sections, we will measure PD and FD as the sum of the edge length of a phylogenetic or functional tree (Faith 1992; Petchey & Gaston 2002), although other representations could be used, such as functional hyperspace (see Schleuter *et al.* 2010; for other measures of functional richness). Hereafter, PD refers to the standard definition sensu Faith (1992) covering both a conventional root in the sense of out-group and a common ancestor. Note that since TD can also be visualised using a tree with each taxon linked directly to the root by an edge of unit length (star tree), tree diagrams provide a common basis for unequivocal estimation and comparison of TD, PD and FD (Cardoso *et al.* 2014).

## ESTIMATION OF PHYLOGENETIC AND FUNCTIONAL DIVERSITY

Here, we propose three ways of estimating PD and FD, building on the estimators originally developed for TD: first, correcting PD and FD values based on the completeness of the taxon inventory; second, fitting asymptotic functions to accumulation curves of PD and FD; and third, adapting nonparametric estimators to PD and FD data.

### Correcting with taxon inventory completeness

Inventory completeness (sensu Sørensen, Coddington & Scharff 2002) can be defined as:

$$c = \frac{S_{\text{obs}}}{S^*}$$

where $S^*$ is computed with any of the available methods (estimators).

The most straightforward way to estimate PD and FD is probably to: (i) compute both observed TD and PD or FD; (ii) compute taxon inventory completeness as above; and (iii) correct observed PD or FD with the completeness values:

$$D^* = \frac{D_{\text{obs}}}{c}$$

where $D^*$ and $D_{\text{obs}}$ are PD or FD estimated and observed, respectively.

This approach should be particularly efficient if the sampling is made randomly along the phylogenetic or functional tree. If unique taxa in either phylogenetic or functional terms are much rarer or more abundant, or alternatively much harder or easier to sample than the rest of the assemblage, the bias may be large. Very unique parts of the tree will be disproportionally sampled compared with that remaining, and this makes the correction of PD or FD values with TD completeness difficult.

### *Fitting asymptotic curves*

This approach follows the same strategy as for TD. The idea is to fit asymptotic functions to randomised or analytical accumulation curves of PD or FD values (see Nipperess & Matsen 2013 for an analytical solution) and to use the asymptote value as the estimate. A randomised curve is generated by taking 1, 2,…, *n* random samples, *n* being the total number of samples. For each sampling level, the observed PD or FD is calculated as the branch length already sampled from the global tree (with *n* samples). The randomisation procedure is repeated a large number of times (we recommend at least 1000 to guarantee smoothness of the curves) allowing obtaining a mean PD or FD for each sampling level. In fact, the R package picante (Kembel *et al.* 2010) includes a function, specaccum.psr, that builds randomised accumulation curves for phylogenetic species richness, and software is also available to compute analytical curves (pplacer suite – matsen.fhcrc.org/pplacer/ or David Nipperess's R functions – davidnipperess.blogspot.com.au). We propose to use such curves to fit different asymptotic functions, such as the Clench, negative exponential or Weibull.

### *Adapting nonparametric estimators*

Here, we propose an adaptation based on phylogenetic or functional trees. Given a tree (Fig. 2) that is progressively completed with the accumulation of samples, singleton or doubleton edge length may be calculated as the sum of length of edges represented by a single or two individuals, respectively. Conversely, unique or duplicate edges may be calculated as the sum of length of edges represented in a single or two samples, respectively. Other abundance or incidence classes may be calculated similarly. This way, many nonparametric estimators, including the eight tested here, may be used to estimate PD or FD.

The adaptation of the jackknife estimators from TD to PD and FD is straightforward and requires no further changes. As for the Chao functions, because PD and FD are measured in real numbers, the use of the original functions is not possible. If $S_1$ is smaller than 1, the numerator $S_1(S_1–1)$ will be negative and the estimated diversity will be smaller than the observed. The same is true for $Q_1$. We propose to substitute these unit correction constants by the minimum branch length of terminal branches found in the tree. This is in fact the minimum value that *S*1 or *Q*1 may take. This minimum value should exclude null distances, as these do not bring any new information. Because in TD the minimum distance is 1 (if TD is measured in a star tree of unit length), it is possible to apply these generalisations of the original Chao formulas to TD, PD or FD. In fact, if this minimum value changes from 1 to close to 0, the formula seamlessly changes from the bias-corrected to the classic form of the Chao estimators (Gotelli & Colwell 2011). The Chao 1 adaptation is:

$$S^* = S_{\text{obs}} + \frac{S_1(S_1 - \min)}{2(S_2 + \min)},$$

where min = minimum taxon, phylogenetic or functional distance between any two species, excluding null distances. The corresponding Chao 2 adaptation is:

$$S^* = S_{\text{obs}} + \frac{Q_1(Q_1 - \min)}{2(Q_2 + \min)}$$

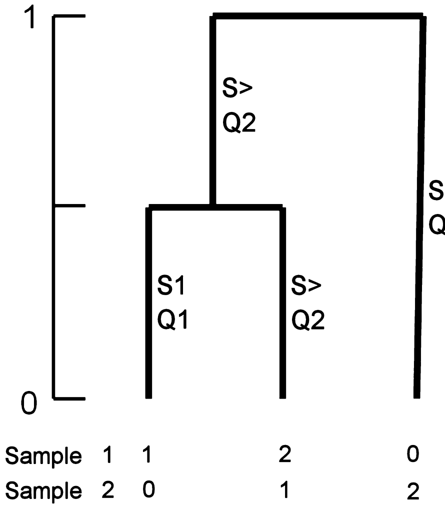

**Fig. 2.** Hypothetical phylogenetic or functional tree (with total phylogenetic or functional diversity = 2·5) with three species and two samples with abundance of species per sample. $S_1$ = singleton edges, $S_2$ = doubleton edges, $S \geq$ = edges represented by more than two individuals, $Q_1$ = unique edges, $Q_2$ = duplicate edges.

Finally, it should be noted that usually nonparametric estimators underestimate diversity when very few samples are used. For that reason, it is common to calculate and plot accumulation curves for the estimated values in addition to the usual observed diversity curves (Sørensen, Coddington & Scharff 2002; Lopez *et al.* 2012). If the estimator curve asymptotes, it may be a good indicator of estimator reliability. In the case of the estimation of PD or FD, using real numbers instead of integers (as in TD) means that depending on the particular samples randomly chosen for building the accumulation curve, particularly the first few samples, some of the estimates may reach unrealistic values and these will disproportionately influence average estimates. This is critical for the Chao estimators, as both use the number of doubletons or duplicates in the denominator. Instead of using averages as usual for TD accumulation curves, we propose to use medians when plotting estimated PD or FD accumulation curves calculated with the Chao estimators. This option should provide very similar values to averaging in most cases, but avoids spurious inflation of estimates in particular cases.

## PERFORMANCE TESTING OF THE NEW ESTIMATORS

Two methods may be used to test the performance of the new estimators. The first is to create artificial communities with a known number of species, species abundance distribution and spatial distribution of each species in a computer-simulated landscape. Artificial trees depicting theoretical phylogenetic or functional relationships between species must also be created for PD or FD analyses. Then individuals or plots are randomly sampled, and these data are used to estimate diversity. Estimators are compared based on their ability to recover true, known, diversity (Brose, Martinez & Williams 2003).

The second method is to use empirical data from areas exhaustively sampled, so that we may assume that the observed diversity is near the real one. Random fractions of the data are extracted from the data set, and these are used to estimate total diversity. Estimates are then compared with known diversity (Colwell & Coddington 1994). Here, we used both methods in performance testing.

All analyses were performed in the R statistical environment (R Development Core Team 2013) using the vegan (Oksanen *et al.* 2011) and spatstat (Baddeley & Turner 2005) packages. An R script to estimate TD, PD and FD using the three proposed methods is given in the Supporting Information (Estimate.r, Data S1).

### Theoretical data sets

We created one artificial community with 10 000 individuals distributed in 100 species following a log-normal species abundance distribution. To simulate a phylogenetic or functional tree, we randomly attributed 10 binary alleles/trait values to each species and from this matrix built a dendrogram using UPGMA (although any other method could have been used, as this tree was purely random). The individuals of each virtual species were then distributed in three simulated square surfaces with varying levels of intraspecific interaction. This way species followed aggregated, random or uniform (overdispersed) distributions in space depending on whether interactions were positive, neutral or negative, respectively. Next, we simulated the sampling of each of the three communities by superimposing a 10 × 10 grid over each square, each of the 100 cells being a sample. Finally, by randomising the order of the samples 1000 times, we built randomised taxon and phylogenetic/functional diversity accumulation curves based on the sum of the branch length and the respective estimated values for all levels of undersampling (from 1 to 99 samples).

### Empirical data sets

We used two empirical data sets previously used by our team for beta diversity partition (see Cardoso *et al.* 2014 for details). These are both exhaustive, with known total diversity values and adequate phylogenetic and functional data, and represent very different organisms and spatial scales.

The Atlas of European Mammals (Mitchell-Jones *et al.* 1999) provided the distribution of the 160 native mammals in Europe, east of roughly 30°E, in a 50 × 50 km resolution. We built the phylogenetic tree for these species by extracting the phylogenetic relationship from the world-wide mammal supertree provided by Bininda-Emonds *et al.* (2007). We tested the ability of the estimators to calculate the PD of all native European mammals with increasing number of cells, how many samples/cells would be needed to correctly estimate total European mammal PD and how much of the bias of observed PD is eliminated. Because no abundance data are available (and in fact would probably be uninformative at this scale), we did not calculate abundance-based estimators.

The North-Atlantic Azorean archipelago, with nine islands, presents a mosaic of land uses, which replace the once almost homogeneous cover of laurel forest (Cardoso *et al.* 2013). A total of 36 sites in the natural forests of Terceira Island were sampled for epigean arthropods using 30 pitfall traps per site (Gaspar, Borges & Gaston 2008). Functional characteristics related to resource use were collated for all the arthropod species, and a functional tree was built (see Cardoso *et al.* 2014 for details). Based on the distribution of the 28 endemic species sampled and the respective functional relationships in the global functional dendrogram, we then calculated the overall FD that can be found in natural forests. We tested the ability of the estimators to calculate the true FD if instead of using 30 pitfall traps we had used anything from 1 to 29 traps per site. We also calculated how many traps per site would be needed to correctly estimate total FD and how much of the bias inherent to observed diversity was eliminated.

### Accuracy measurement

The behaviour of the estimators and their ability to correct undersampling bias can be tested in a number of different ways (Walther & Moore 2005). To test the new estimators with the empirical data sets, we used the scaled mean square error:

$$\mathrm{SMSE} = \frac{1}{A^2 Q}\sum_{j=1}^{Q}(S_j - A)^2$$

where $A$ = real total diversity and $S_j$ = estimated diversity for the *j*th sample. We chose this measure among the many possible options as it is: (i) scaled to true diversity, so that similar absolute differences are weighted according to how much they represent of the real value; (ii) scaled to the number of samples, so that values are independent of sample size; (iii) squared, so that small, mostly meaningless fluctuations around the true value are down-weighted; and (iv) independent of positive or negative deviation from the real value, as such differentiation is usually not necessary. This value was calculated for both data sets, using the sample numbers $n/2^9$ to $n/2^1$ (nine values) for the European data set, so that the first, highly biased, part of the curve was heavily weighted compared with the last, mostly unbiased, part of the curve, and all 29 values for the Azorean data set.

## Results

THEORETICAL DATA SETS

### Fitting asymptotic curves

For illustrative purposes, we show the results of curve fitting with only five samples (Fig. 3). Even with only 5% of the theoretical assemblages sampled, the percentage of observed species is higher than 80% in all cases. However, most of the real-world assemblages are aggregated, and this is, probably, the worst-case scenario for sampling and estimation of diversity as such assemblages require much more effort to be thoroughly sampled than assemblages where species are randomly or uniformly distributed in space and hence much more accessible when sampling effort concentrates in few sites or microhabitats. But even in aggregated assemblages, the asymptote of the Clench function reaches a value very close to the real.

### Nonparametric estimators

In general, nonparametric estimators are capable of correcting much of the bias of observed diversity, with a similar behaviour for taxon, phylogenetic or functional diversity (Fig. 4). Their behaviour does, however, differ, with the jackknife formulas for incidence data often overshooting with very few samples and both the Jack and Chao formulas for abundance data underestimating with aggregated communities. The adaptations here proposed for PD and FD are, however, at least as efficient as the original formulations for TD.

EMPIRICAL DATA SETS

### Phylogenetic diversity of European mammals

For the nonparametric estimators, between 100 and 500 of the 2000 cells are needed to reach the true species richness (TD) value (Fig. 5), with both jackknifes being particularly efficient. For PD, the correction based on taxon inventory completeness is remarkably efficient, with between 5 and 100 cells being needed to reach the true values, with the P-corrected versions

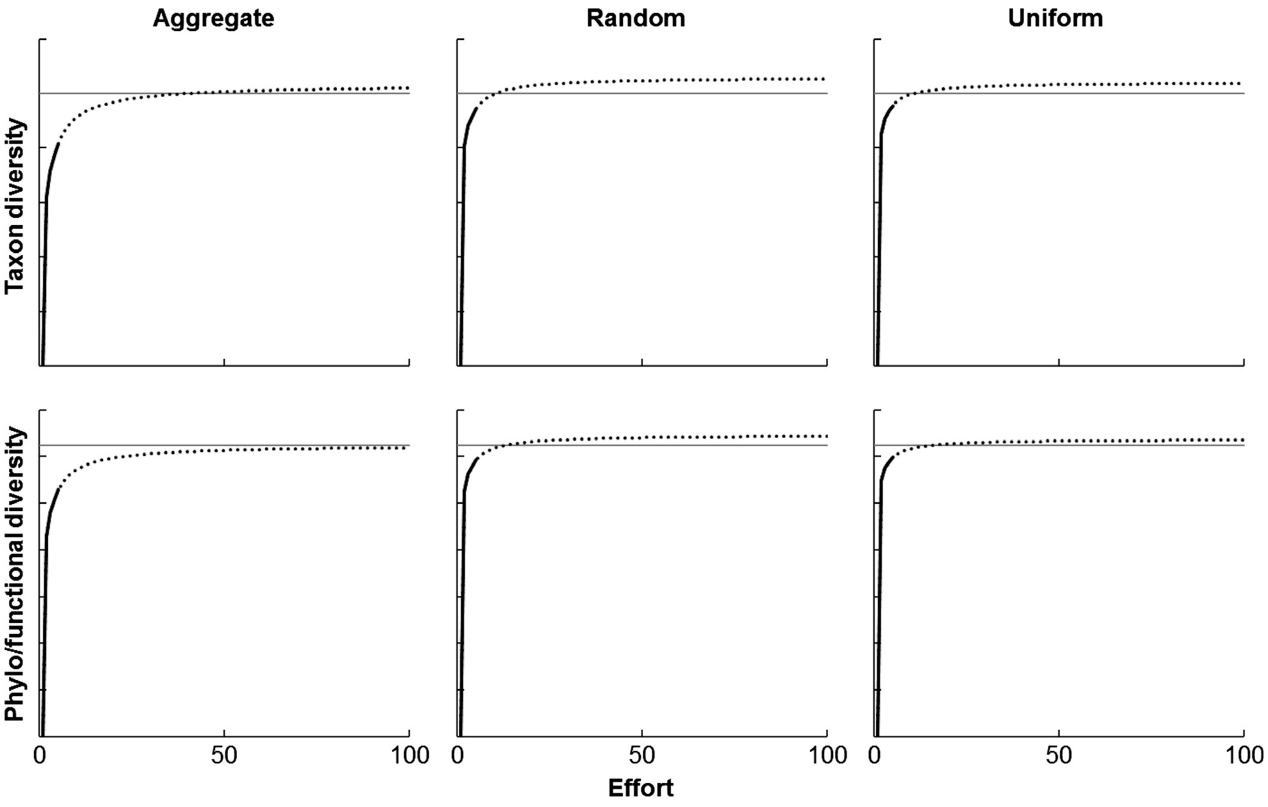

**Fig. 3.** Randomised accumulation curves for observed diversity up to five samples (full black line) and the respective fitted Clench functions extrapolated up to 100 samples (dotted black line). Theoretical assemblages following aggregated, random and uniform spatial distributions were built, as was a random phylogenetic/functional tree connecting all species (see text for details). The true diversity of the assemblages is also shown (full grey line).

of estimators performing particularly well. The direct nonparametric estimation approach requires about 100 cells for all functions. The Clench curve clearly underperforms in all cases compared with the nonparametric estimators. Regarding the accuracy of the different methods, all estimators largely reduced the bias of the observed values (Appendix S1, Table S1, Supporting information).

### *Functional diversity of Azorean arthropods*

In all cases the estimators predict higher diversity than effectively observed in the 36 sampled sites, which may indicate that in fact our sampled universe is still incomplete (Fig. 6). The FD estimates based on TD completeness reach the asymptote very quickly, at about five traps, while in all other cases about 15–20 samples (traps) are needed (Fig. 6). Regarding the accuracy, all estimators had rather similar abilities to correct bias, although the Clench method seems to perform slightly better (Appendix S1, Table S2, Supporting information). As the SMSE values are based on the total observed diversity, and this may be underestimated, these values should, however, be interpreted with caution.

## Discussion

This study sought to explore the impact of undersampling on observed TD, PD and FD and to adapt some of the mostly used and best-performing species richness estimators to PD and FD measures.

In general, the simple correction of observed PD or FD with taxon inventory completeness values seems to perform surprisingly well. This approach even outperforms TD estimates, for which the estimators were specifically built. This may be due to two factors in combination. First, TD estimators are known to consistently undercorrect with a low number of samples (O'Hara 2005; Cardoso 2009; Lopez *et al.* 2012). Second, observed PD and FD are intrinsically less biased than TD with a low number of samples, as the first taxa sampled add on average more diversity than the latter taxa, which should share parts of the tree with the ones previously sampled. The two effects combined mean that with a low number of samples, it is possible to very effectively correct PD or FD values. This would not be the case if species richness estimators were more efficient with low sampling, in which case estimated PD or FD would initially overshoot the true value.

Both asymptotic functions and nonparametric estimators used for PD or FD seem to perform as well as for TD, for which they were first created. Although results are variable, this means that in many empirical data sets at least 70–80% completeness is needed to have reliable estimates (Sørensen, Coddington & Scharff 2002; Cardoso 2009). Reaching such levels is, however, not guaranteed in many studies, particularly when dealing with hyperdiverse taxa, such as most arthropods (Coddington *et al.* 2009). Overall, better estimators are needed

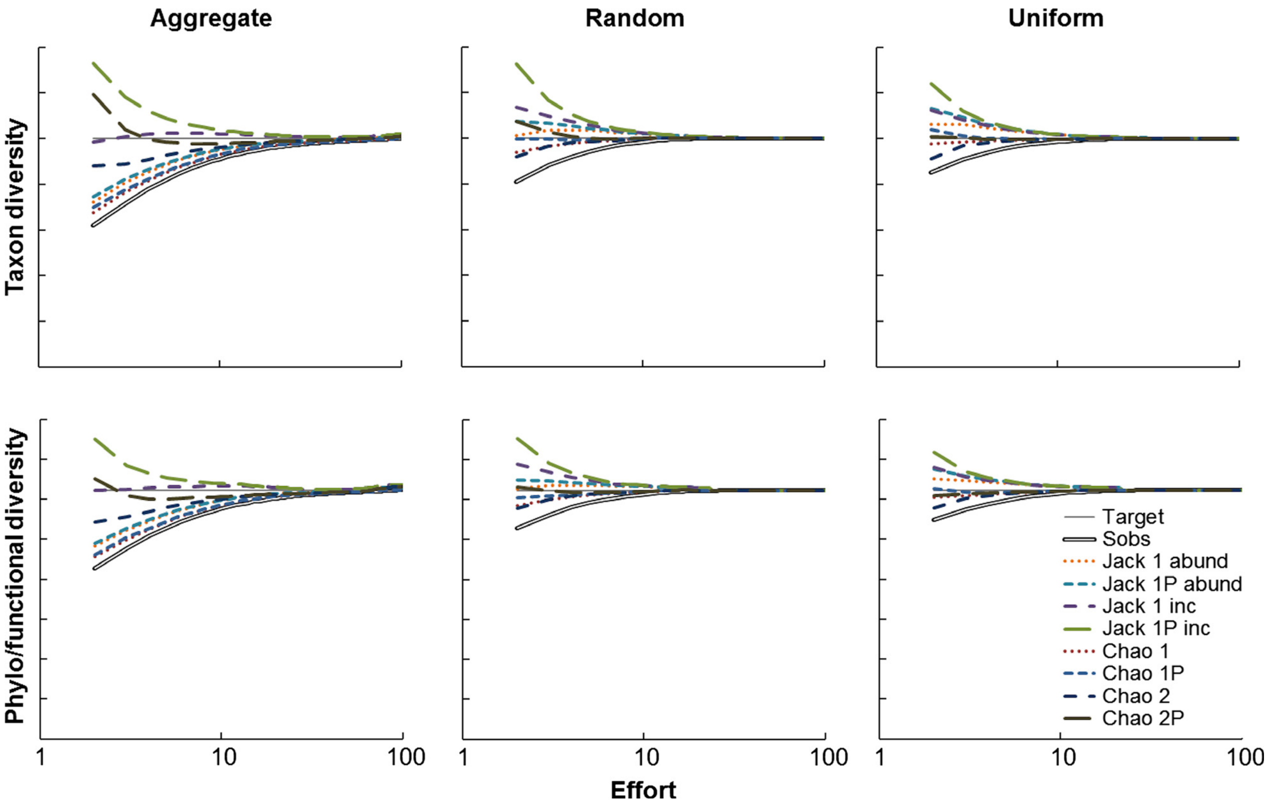

**Fig. 4.** Randomised accumulation curves for observed and estimated diversity. Theoretical assemblages following aggregated, random and uniform spatial distributions were built, as was a random phylogenetic/functional tree connecting all species (see text for details). The true diversity of the assemblages is also shown (target). Note that the *x*-axis is in $\log_{10}$ scale to facilitate comparisons with very low sampling levels. Sobs = observed diversity, Jack 1 (P) abund = jackknife 1 estimator with abundance data, Jack 1(P) inc = Jackknife 1 estimator with incidence data, Chao 1(P) = Chao estimator with abundance data, Chao 2(P) = Chao estimator with incidence data (P = P-corrected version for all Jack and Chao formulas).

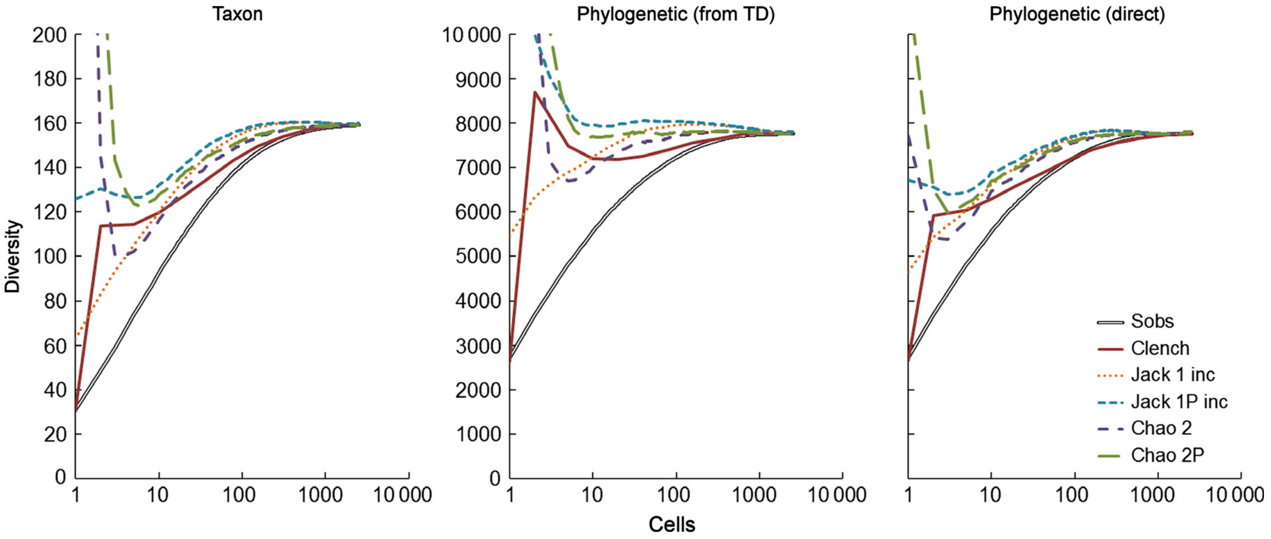

**Fig. 5.** Observed and estimated species richness (TD) and phylogenetic diversity (PD) of European mammals with the randomised accumulation of cells in Europe. The curves in the middle panel were calculated correcting the observed PD values using the completeness of TD (see text for details). The curves in the right panel were calculated either fitting the Clench asymptotic function to each point in the observed PD curve or using nonparametric estimators. Note that the *x*-axis is in $\log_{10}$ scale to facilitate comparisons with very low sampling levels. Sobs = observed diversity, Clench = fitted asymptotic curve, Jack 1(P) inc = Jackknife 1 estimator with incidence data, Chao 2(P) = Chao estimator with incidence data (P = P-corrected version for both Jack and Chao formulas).

for TD, PD and FD, but in all cases, the bias inherent to observed diversity is in fact reduced, so their utility is certainly confirmed.

We should also mention that we did not try adapting the analytical estimators of unconditional variance existing for accumulation curves of nonparametric formulas (Gotelli &

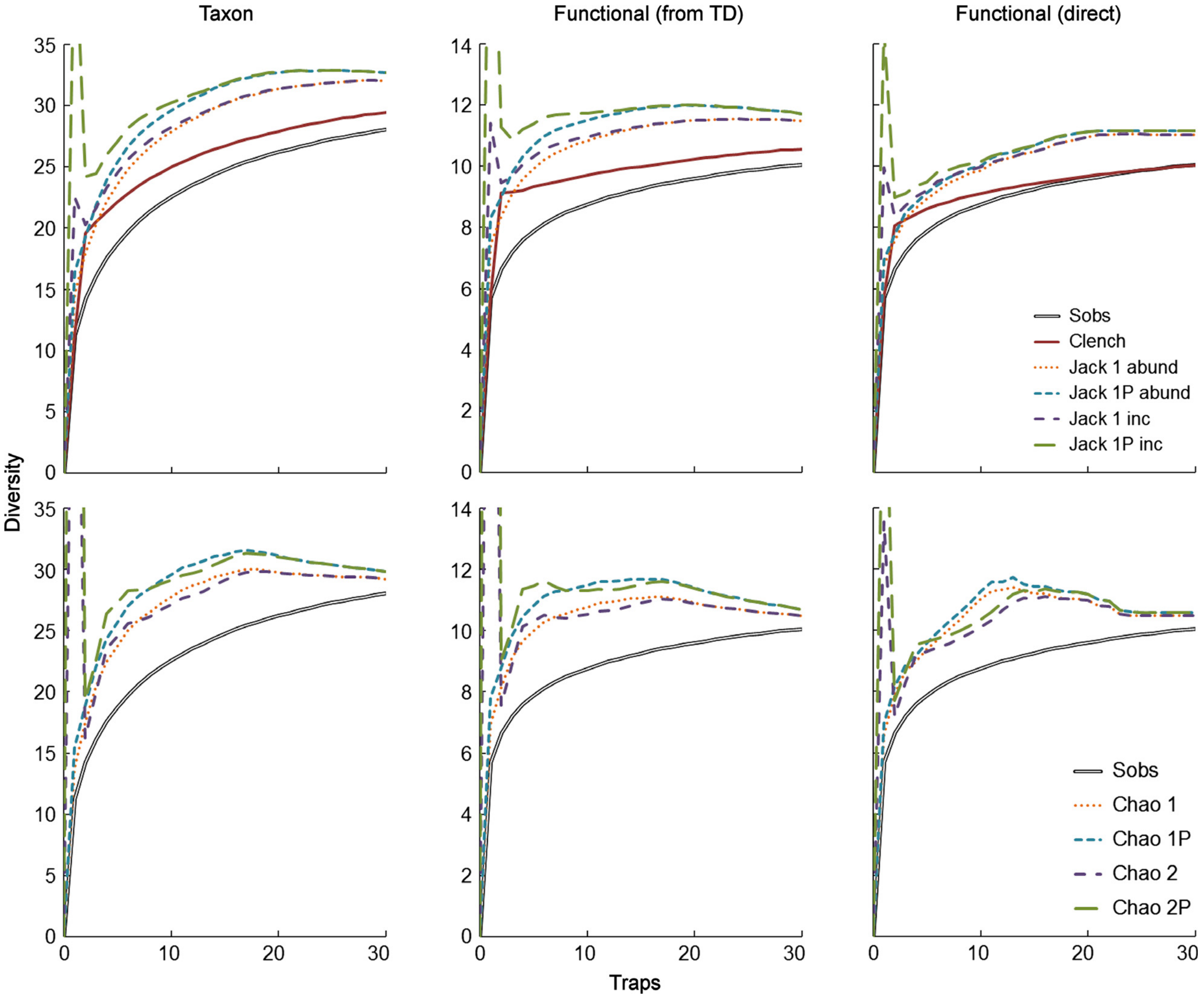

**Fig. 6.** Observed and estimated taxon (TD) and functional (FD) diversity of endemic Azorean arthropods with the randomised accumulation of traps. The curves in the middle panels were calculated correcting the observed FD values using the completeness of TD (see text for details). The curves in the right panels were calculated either fitting the Clench asymptotic function to each point in the observed FD curve or using nonparametric estimators. Sobs = observed diversity, Clench = fitted asymptotic curve, Jack 1(P) abund = jackknife 1 estimator with abundance data, Jack 1(P) inc = jackknife 1 estimator with incidence data, Chao 1(P) = Chao estimator with abundance data, Chao 2(P) = Chao estimator with incidence data (P = P-corrected version for all Jack and Chao formulas).

Colwell 2011). Only conditional variances are possible to calculate, and these approach 0 when all samples are included.

### ALTERNATIVE METHODS

Although we have focused on asymptotic functions and nonparametric estimators, taxa richness (TD) has been estimated through other means. Asymptotic functions have been criticised for not relying on a strong theoretical basis, being a strictly phenomenological method. Different functions may fit a curve equally well, yet results may vary dramatically (Soberón & Llorente 1993), and residuals often reveal that many functions do not correctly fit curve shapes (O'Hara 2005). For these reasons, Gotelli & Colwell (2011) do not recommend the method. Possible better alternatives exist (Colwell *et al.* 2012), but these are probably much harder to adapt for PD and FD.

Another often used approach for species richness estimation is to fit a species abundance distribution to a truncated parametric distribution and estimate the portion to the left of the 'veil line'. This unseen part of the distribution corresponds to the undetected species. However, PD and FD do not follow such theoretical abundance distributions. In fact, it would be unclear to define what abundance is in this context.

### GOING BEYOND DIVERSITY ESTIMATES

Besides the obvious appeal and utility of knowing the true diversity of an assemblage from incomplete samples, the use of estimators may present further advantages. Comparisons between sites, regions and time periods are only possible if either the sampling is complete, which often is not the case, or the sampling effort is equivalent and sufficient enough to allow sensible comparisons. The way this problem has been typically corrected is through rarefaction. Rarefaction allows the comparison of diversity among assemblages on an equal-effort basis (Gotelli & Colwell 2001; Chao & Jost 2012). Rarefaction of PD or FD, either through randomisation procedures or analytically (Walker, Poos & Jackson 2008; Ricotta *et al.* 2012; Nipperess & Matsen 2013), is an alternative to estimation when total diversity values are not needed, but the objective is to compare different assemblages with similar effort. Rarefaction does, however, prevent one from knowing true diversity numbers; estimating these, if without significant bias, should be a preferable option.

Another advantage of using estimated in lieu of observed or rarefied diversity is clear in studies relating TD with either PD or FD. Many studies try to measure the level of redundancy in assemblages by comparing differences in TD between assemblages with respective differences in PD or FD (Mayfield *et al.* 2010; Gerisch *et al.* 2012; Whittaker *et al.* in press). However, as seen (Appendix S1, Supporting information), PD and FD are less biased than TD in the presence of undersampling. This means that, due to pure sampling effects, phylogenetic or functional redundancy may be underestimated in many cases, as PD and FD may reach an asymptote much before TD. Obviously, this use of the estimators does prevent the application of our first approach, the reliance on TD completeness values, as all dimensions are corrected by the same proportion.

### LIMITS AND FUTURE PERSPECTIVES

In this study, we choose to focus on trees or dendrograms, not only because trees allow a straightforward adaptation of existing estimators but also because they allow the comparison of TD, PD and FD under the same representation (Cardoso *et al.* 2014). Although there is no other obvious alternative for PD, the use of dendrograms to compute FD has been the subject of strong debate (Schleuter *et al.* 2010). We emphasise that the user needs to consider that the choice of the distance and the clustering method may greatly affect the FD values obtained; some publications and associated scripts are available to guide the researcher in this process (Mouchet *et al.* 2008; Mérigot, Durbec & Gaertner 2010). The framework proposed here can also be adapted to other representations such as multidimensional hypervolumes, a common representation for FD. It should be noted, however, that the adaptation is not as straightforward as with dendrograms. Due to the nature of the calculation of convex-hull volumes, each sample may occupy a relatively small portion of the functional space, yet when considering several samples simultaneously, the space in between samples may be occupied, even when none of the samples individually occupies it. To what extent this shortcoming causes biases in practice remains to be studied. Additionally, the calculation of the convex-hull volume for each sample requires that the number of species always exceeds the number of traits (Laliberté & Legendre 2010), which, in turn, constrains the user either to reduce the dimensionality of the functional space or to remove species-poor samples. In any case, FD as measured with hypervolumes can always be estimated using the correction with TD completeness or fitting asymptotic functions.

## Conclusions

Using a range of both theoretical and empirical examples, we show that current approaches to estimating taxon diversity (TD, usually species richness) are as efficient or even more efficient when used to estimate phylogenetic (PD) or functional (FD) diversity. The framework used here combines TD, PD and FD into a single representation, offering a tool for future developments involving these different facets of biological diversity. A number of topics are in need of future development, namely (i) the comparison of different approaches to estimating PD and FD, not only the ones presented here but also other asymptotic functions or nonparametric estimators; (ii) verification of the circumstances in which each approach is preferable, if this is predictable at all; and (iii) the development of estimators specifically for PD and FD data, potentially superior to the current adaptations of TD estimators.

## Acknowledgements

We thank Nick Gotelli, Jack Longino, Jon Coddington, David Nipperess and Dan Faith for comments on previous versions of the manuscript. Clara Gaspar compiled all the functional data on Azorean arthropods. The Societas Europaea Mammalogica and particularly Tony Mitchell-Jones provided the European Mammals Atlas data.

## Data accessibility

R scripts: uploaded as online supporting information (Estimate.r).

## Funding

Azorean data derived from the DRCT projects M.2.1.2/I/003/2008 and M.1.1.2/FRCT. F.R. and P.B. were partly supported by FCT project PTDC/BIA-BIC/119255/2010.

## References

Baddeley, A. & Turner, R. (2005) Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**, 1–42.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R. *et al.* (2007) The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Brose, U., Martinez, N.D. & Williams, R.J. (2003) Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, **84**, 2364–2377.

Cardoso, P. (2009) Standardization and optimization of arthropod inventories – the case of Iberian spiders. *Biodiversity and Conservation*, **18**, 3949–3962.

Cardoso, P., Borges, P.A.V. & Veech, J.A. (2009) Testing the performance of beta diversity measures based on incidence data: the robustness to undersampling. *Diversity and Distributions*, **15**, 1081–1090.

Cardoso, P., Rigal, F., Fattorini, S., Terzopoulou, S. & Borges, P.A.V. (2013) Integrating landscape disturbance and indicator species in conservation studies. *PLoS One*, **8**, e63294.

Cardoso, P., Rigal, F., Carvalho, J.C., Fortelius, M., Borges, P.A.V., Podani, J. & Schmera, D. (2014) Partitioning taxon, phylogenetic and functional beta diversity into replacement and richness difference components. *Journal of Biogeography*, **41**, 749–761.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, **11**, 265–270.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.

Chao, A. & Jost, L. (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, **93**, 2533–2547.

Chao, A., Hwang, W.-H., Chen, Y.-C. & Kuo, C.-Y. (2000) Estimating the number of shared species in two communities. *Statistica Sinica*, **10**, 227–246.

Chao, A., Chazdon, R.L., Colwell, R.K. & Tsung-Jen, S. (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**, 148–159.

Coddington, J.A., Agnarsson, I., Miller, J.A., Kuntner, M. & Hormiga, G. (2009) Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. *Journal of Animal Ecology*, **78**, 573–584.

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London – Biological Sciences*, **345**, 101–118.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L. & Longino, J.T. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, **5**, 3–21.

Cornwell, W.K., Schwilk, D.W. & Ackerly, D.D. (2006) A trait-based test for habitat filtering: convex hull volume. *Ecology*, **87**, 1465–1471.

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

Erwin, T.L. (1982) Tropical forests: their richness in Coleoptera and other arthropod species. *Coleopterists Bulletin*, **36**, 74–75.

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

Flather, C. (1996) Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, **23**, 155–168.

Gaspar, C., Borges, P.A.V. & Gaston, K.J. (2008) Diversity and distribution of arthropods in native forests of the Azores archipelago. *Arquipelago Life and Marine Sciences*, **25**, 1–30.

Gerisch, M., Agostinelli, V., Henle, K. & Dziock, F. (2012) More species, but all do the same: contrasting effects of flood disturbance on ground beetle functional and species diversity. *Oikos*, **121**, 508–515.

Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Gotelli, N.J. & Colwell, R.K. (2011) Estimating species richness. *Frontiers in Measuring Biodiversity* (eds A.E. Magurran & B.J. McGill), pp. 39–54. Oxford University Press, New York.

Graham, C. & Fine, P. (2008) Phylogenetic beta diversity: linking ecological and evolutionary processes across space and time. *Ecology Letters*, **11**, 1265–1277.

Helmus, M.R., Bland, T.J., Williams, C.K. & Ives, A.R. (2007) Phylogenetic measures of biodiversity. *American Naturalist*, **169**, 68–83.

Heltshe, J. & Forrester, N.E. (1983) Estimating species richness using the jackknife procedure. *Biometrics*, **39**, 1–11.

Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., Blomberg, S.P. & Webb, C.O. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**, 1463–1464.

Laliberté, E. & Legendre, P. (2010) A distance-based framework for measuring functional diversity from multiple traits. *Ecology*, **91**, 299–305.

Longino, J., Colwell, R.K. & Coddington, J.A. (2002) The ant fauna of a tropical rainforest: estimating species richness three different ways. *Ecology*, **83**, 689–702.

Lopez, L.C.S., Fracasso, M.P.A., Mesquita, D.O., Palma, A.R.T. & Riul, P. (2012) The relationship between percentage of singletons and sampling effort: a new approach to reduce the bias of richness estimates. *Ecological Indicators*, **14**, 164–169.

Mayfield, M.M., Bonser, S.P., Morgan, J.W., Aubin, I., McNamara, S. & Vesk, P.A. (2010) What does species richness tell us about functional trait diversity? Predictions and evidence for responses of species and functional trait diversity to land-use change. *Global Ecology and Biogeography*, **19**, 423–431.

Mérigot, B., Durbec, J.P. & Gaertner, J.C. (2010) On goodness-of-fit measure for dendrogram-based analyses. *Ecology*, **91**, 1850–1859.

Mitchell-Jones, A.J., Amori, G., Bogdanowicz, W., Krystufek, B., Reijnders, P.J.H., Spitzenberger, F. *et al.* (1999) *The Atlas of European Mammals.* T. & A.D. Poyser Ltd and Academic Press, London.

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B. & Worm, B. (2011) How many species are there on Earth and in the Ocean? *PLoS Biology*, **9**, e1001127.

Mouchet, M., Guilhaumon, F., Villéger, S., Mason, N.W., Tomasini, J.A. & Mouillot, D. (2008) Towards a consensus for calculating dendrogram-based functional diversity indices. *Oikos*, **117**, 794–800.

Mouchet, M.A., Villéger, S., Mason, N.W. & Mouillot, D. (2010) Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology*, **24**, 867–876.

Nipperess, D.A. & Matsen, F.A. (2013) The mean and variance of phylogenetic diversity under rarefaction. *Methods in Ecology and Evolution*, **4**, 566–572.

O'Hara, R.B. (2005) Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology*, **74**, 375–386.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B. *et al.* (2011) Vegan: community ecology package. R package version 2.0-2. http://CRAN.R-project.org/package = vegan.

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

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

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. http://www.R-project.org.

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

Ricotta, C., Pavoine, S., Bacaro, G. & Acosta, A.T.R. (2012) Functional rarefaction for species abundance data. *Methods in Ecology and Evolution*, **3**, 519–525.

Schleuter, D., Daufresne, M., Massol, F. & Argillier, C. (2010) A user's guide to functional diversity indices. *Ecological Monographs*, **80**, 469–484.

Soberón, M.J. & Llorente, J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**, 480–488.

Sørensen, L.L., Coddington, J.A. & Scharff, N. (2002) Inventorying and estimating subcanopy spider diversity using semiquantitative sampling methods in an afromontane forest. *Environmental Entomology*, **31**, 319–330.

Tilman, D., Reich, P.B., Knops, J., Wedin, D., Mielke, T. & Lehman, C. (2001) Diversity and productivity in a long-term grassland experiment. *Science*, **294**, 843–845.

Villéger, S., Mason, N.W. & Mouillot, D. (2008) New multidimensional functional diversity indices for a multifaceted framework in functional ecology. *Ecology*, **89**, 2290–2301.

Walker, S.C., Poos, M.S. & Jackson, D.A. (2008) Functional rarefaction: estimating functional diversity from field data. *Oikos*, **117**, 286–296.

Walther, B.A. & Moore, J.L. (2005) The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, **28**, 815–829.

Webb, C.O., Ackerly, D.D., McPeek, M.A. & Donoghue, M.J. (2002) Phylogenies and community ecology. *Annual Review of Ecology, Evolution and Systematics*, **33**, 475–505.

Weiher, E., Clarke, G.P. & Keddy, P.A. (1998) Community assembly rules, morphological dispersion, and the coexistence of plant species. *Oikos*, **81**, 309–322.

Whittaker, R.J., Rigal, F., Borges, P.A.V., Cardoso, P., Terzopoulou, S., Casanoves, F. *et al.* (in press) Functional biogeography of oceanic islands and the scaling of functional diversity in the Azores. *Proceedings of the National Academy of Sciences USA*.

*Received 3 December 2013; accepted 26 February 2014*
*Handling Editor: Daniel Faith*

## Supporting Information

Additional Supporting Information may be found in the online version of this article.

**Appendix S1.** Scaled mean square error (SMSE) of each descriptor of species richness (TD) or phylogenetic diversity (PD) of accumulation curves for European mammals (only incidence estimators are shown) and Azorean arthropods.

**Table S1.** Scaled mean square error (SMSE) of each descriptor of species richness (TD) or phylogenetic diversity (PD) of accumulation curves for European mammals (only incidence estimators are shown).

**Table S2.** Scaled mean square error (SMSE) of each descriptor of species richness (TD) or functional diversity (FD) of accumulation curves for Azorean arthropods.

**Data S1**. An R script to estimate TD, PD and FD.

**Appendix S1.** Scaled mean square error (SMSE) of each descriptor of species richness (TD) or phylogenetic diversity (PD) of accumulation curves for European mammals (only incidence estimators are shown) and Azorean arthropods.

**Table S1.** Scaled mean square error (SMSE) of each descriptor of species richness (TD) or phylogenetic diversity (PD) of accumulation curves for European mammals (only incidence estimators are shown). Smaller values are best, 0 would represent an ideal descriptor. Percentages represent remaining bias in relation to observed values.

| SMSE | TD | PD (completeness) | PD (curve fit and non-parametric) |
|---|---|---|---|
| $S_{obs}$ | 0.112 | 0.032 | 0.032 |
| Clench | 0.030 (27%) | 0.002 (7%) | 0.014 (45%) |
| Jack 1 incidence | 0.043 (39%) | 0.002 (7%) | 0.009 (27%) |
| Jack 1P incidence | 0.012 (11%) | 0.001 (4%) | 0.005 (17%) |
| Chao 2 | 0.026 (23%) | 0.003 (11%) | 0.012 (38%) |
| Chao 2P | 0.044 (39%) | 0.0002 (1%) | 0.008 (25%) |

**Table S2.** Scaled mean square error (SMSE) of each descriptor of species richness (TD) or functional diversity (FD) of accumulation curves for Azorean arthropods. Smaller values are best, 0 would represent an ideal descriptor. Percentages represent remaining bias in relation to observed values.

| SMSE | TD | FD (completeness) | FD (curve fit and non-parametric) |
|---|---|---|---|
| $S_{obs}$ | 0.038 | 0.017 | 0.017 |
| Clench | 0.014 (37%) | 0.002 (12%) | 0.007 (44%) |
| Jack 1 abundance | 0.019 (49%) | 0.015 (86%) | 0.009 (55%) |
| Jack 1P abundance | 0.023 (59%) | 0.026 (153%) | 0.010 (58%) |
| Chao 1 | 0.012 (32%) | 0.007 (39%) | 0.009 (51%) |
| Chao 1P | 0.013 (33%) | 0.015 (89%) | 0.011 (65%) |
| Jack 1 incidence | 0.015 (40%) | 0.014 (84%) | 0.007 (41%) |
| Jack 1P incidence | 0.019 (49%) | 0.030 (176%) | 0.007 (42%) |
| Chao 2 | 0.013 (33%) | 0.006 (38%) | 0.008 (48%) |
| Chao 2P | 0.010 (27%) | 0.015 (86%) | 0.009 (50%) |

⍰

```
##### Code to calculate observed and estimated taxon, phylogenetic and functional diversity
#####
#
# This code contains several functions to calculate estimated taxon, phylogenetic and
functional diversity, according to the methods given in
# Cardoso et al (2014) A new frontier in biodiversity inventory: a proposal for estimators
of phylogenetic and functional diversity.
# Methods in Ecology and Evolution, 5: 452-461.
#
# xTree : a function necessary to calculate branch lengths
# Auxiliary functions : functions necessary to preprocess the data
# estimate : the main function that calculates the estimators of TD, FD and PD
#
# Both xTree and auxiliary functions need to be uploaded into R before running the function
estimate.
#
# The function estimate has two main arguments:
# comm - a data frame with species as columns, plots/sites as rows, and presence/absence or
abundance as entries
# tree - which could be either a phylogenetic or functional tree.
# The tree can be either a phylo or an hclust object. The function estimate extracts branch
lengths from an
# hclust object but we include the function as.hclust to automatically convert phylo objects
to
# class "hclust".
#
# The authors ensure that care has been taken in writing the code and it is
# believed to be accurate. However, users of this code are cautioned that it has not been
# extensively tested and its use and results are solely the responsibility of the user.
#
##### Code by Jose Carlos Carvalho, Francois Rigal & Pedro Cardoso
#
# xTree function adapted from
http://owenpetchey.staff.shef.ac.uk/Code/Code/calculatingfd_assets/Xtree.r
# by Jens Schumacher (described in Petchey & Gaston 2002, 2006)
#
xTree <- function(h) {
#
# h : a tree or dendrogram. It could be an hclust or phylo object.
# If it is a phylo object, the tree needs to be ultrametric (check with function
is.ultrametric from package ape)
#
    h <- as.hclust(h)
    nSpecies = nrow(as.data.frame(h['order']))
    H1 <- matrix(0, nSpecies, 2 * nSpecies - 2)
    l <- vector("numeric", 2 * nSpecies - 2)
    for(i in 1:(nSpecies - 1)) {
        if(h$merge[i, 1] < 0) {
            l[2 * i - 1] <- h$height[order(h$height)[i]]
            H1[ - h$merge[i, 1], 2 * i - 1] <- 1
        } else {
            l[2 * i - 1] <- h$height[order(h$height)[i]]-h$height[order(h$height)[h$merge[i,
            1]]]
            H1[, 2 * i - 1] <- H1[, 2 * h$merge[i, 1] - 1] + H1[ , 2 * h$merge[i, 1]]
        }
        if(h$merge[i, 2] < 0) {
            l[2 * i] <- h$height[order(h$height)[i]]
            H1[ - h$merge[i, 2], 2 * i] <- 1
        } else {
            l[2 * i] <- h$height[order(h$height)[i]] - h$height[order(h$height)[h$merge[i,
            2]]]
            H1[, 2 * i] <- H1[, 2 * h$merge[i, 2] - 1] + H1[, 2 *h$merge[i, 2]]
        }
    }
    rownames(H1)= h$labels
    list(l, H1)
}

##### Auxiliary functions #####
#
# preprocess data
prep <- function(comm, tree, abund = TRUE){
    xTree.object = xTree(tree)
    len = xTree.object[[1]]             ## length of each branch
    A = xTree.object[[2]]           ## matrix species X branches
    minBranch = min(len[colSums(A)==1])     ## minimum branch length of terminal branches
    BA = comm%*%A           ## matrix samples X branches
    if (!abund) BA = ifelse(BA >= 1, 1, 0)
return (list(lenBranch = len, sampleBranch = BA, minBranch = minBranch))
}

# observed TD
sobs <- function(comm){
    value = colSums(comm)
    return (sum(value > 0))
}

# TD of rare species for abundance - singletons, doubletons, tripletons, etc
srare <- function(comm, abund){
    value = colSums(comm)
    return (sum(value == abund))
}

# TD of rare species for incidence - uniques, duplicates, triplicates, etc
qrare <- function(comm, incid){
    value = colSums(ifelse(comm > 0, 1, 0))
    return (sum(value == incid))
}

# observed P/FD
dSobs <- function(comm, tree){
    data = prep(comm, tree)
    value = ifelse (colSums(data$sampleBranch)>0, 1, 0)     ## vector of observed branches
    return (sum(value*data$lenBranch))
}

# P/FD of rare species for abundance - singletons, doubletons, tripletons, etc
dSrare <- function(comm, tree, abund){
    data = prep(comm, tree)
    value = ifelse (colSums(data$sampleBranch)==abund, 1, 0)    ## branches with given
    abundance
    return (sum(value*data$lenBranch))
}

# P/FD of rare species for incidence - uniques, duplicates, triplicates, etc
dQrare <- function(comm, tree, incid){
    data = prep(comm, tree, FALSE)
    value = ifelse (colSums(data$sampleBranch)==incid, 1, 0)    ## branches with given
    incidence
    return (sum(value*data$lenBranch))
}

# minimum terminal branch length
minBranch <- function(comm, tree){
    data = prep(comm, tree)
    return(data$minBranch)
}

##### Main function
#
estimate <- function(comm, tree, func = "species", runs = 1000){
#
# Arguments--
#
# comm : matrix of species x samples (with species ordered as in the tree)
# tree : an hclust object with all the species of comm
# func : computes the estimates for TD, PD and FD. This argument has four options:
## "species" - computes TD with non-parametric estimators
## "completeness" - computes P/FD with TD completeness correction
## "dendrogram" - computes P/FD with non-parametric estimators
## "curve" computes P/FD with curve fitting
# nruns : number of runs used to build the accumulation curves

    func <- match.arg(func, c("species", "completeness", "dendrogram", "curve"))

    ##### species (TD with non-parametric estimators)
    switch(func, species = {
        results = matrix(0,nrow(comm),15)
        for (r in 1:runs){
            comm = comm[sample(nrow(comm)),, drop=FALSE]    ## shuffle rows (samples)
            data = matrix(0,0,ncol(comm))           ## reset data
            runData = matrix(0,0,15)
            for (q in 1:nrow(comm)){
                data = rbind(data, comm[q,])
                n = sum(rowSums(data))
                obs = sobs(data)
                s1 = srare(data, 1)
                s2 = srare(data, 2)
                q1 = qrare(data, 1)
                q2 = qrare(data, 2)
                jack1ab = obs + s1
                jack1abP = jack1ab * (1+(s1/obs)^2)
                jack1in = obs + q1 * ((q-1)/q)
                jack1inP = jack1in * (1+(q1/obs)^2)
                chao1 = obs + (s1*(s1-1))/(2*(s2+1))
                chao1P = chao1 * (1+(s1/obs)^2)
                chao2 = obs + (q1*(q1-1))/(2*(q2+1))
                chao2P = chao2 * (1+(q1/obs)^2)
                runData = rbind(runData, c(q, n, obs, s1, s2, q1, q2, jack1ab, jack1abP,
                jack1in,
jack1inP, chao1, chao1P, chao2, chao2P))
            }
            results = results + runData
        }
        results = results / runs

    ##### completeness (P/FD with TD completeness correction)
    }, completeness = {
        results = estimate(comm, , "sp", runs)
        results = results[-1,]
        obs = matrix(0,nrow(comm),1)
        for (r in 1:runs){
            comm = comm[sample(nrow(comm)),]            ## shuffle rows (samples)
            for (s in 1:nrow(comm))     obs[s,1] = obs[s,1] + dSobs(comm[1:s,], tree)
        }
        obs = obs / runs
        for (i in 4:7) results[,i] = -1
        for (i in 8:15) results[,i] = obs * (results[,i] / results[,3])
        results[,3] = obs

    ##### dendrogram (P/FD with non-parametric estimators)
    }, dendrogram = {
        resArray = array(0, dim = c(nrow(comm), 15, runs))
        for (r in 1:runs){
            comm = comm[sample(nrow(comm)),, drop=FALSE]    ## shuffle rows (samples)
            data = matrix(0,0,ncol(comm))           ## reset data
            runData = matrix(0,0,15)
            for (q in 1:nrow(comm)){
                data = rbind(data, comm[q,])
                n = sum(rowSums(data))
                obs = dSobs(data, tree)
                s1 = dSrare(data, tree, 1)
                s2 = dSrare(data, tree, 2)
                q1 = dQrare(data, tree, 1)
                q2 = dQrare(data, tree, 2)
                mb = minBranch(data, tree)
                jack1ab = obs + s1
                jack1abP = jack1ab * (1+(s1/obs)^2)
                jack1in = obs + q1 * ((q-1)/q)
                jack1inP = jack1in * (1+(q1/obs)^2)
                chao1 = obs + (s1*(s1-mb))/(2*(s2+mb))
                chao1P = chao1 * (1+(s1/obs)^2)
                chao2 = obs + (q1*(q1-mb))/(2*(q2+mb))
                chao2P = chao2 * (1+(q1/obs)^2)
                runData = rbind(runData, c(q, n, obs, s1, s2, q1, q2, jack1ab, jack1abP,
                jack1in,
jack1inP, chao1, chao1P, chao2, chao2P))
            }
            resArray[,,r] = runData
        }

        ##### calculate median of all runs
        results = matrix(0,nrow(comm),15)
        v = array(0, dim = c(runs))
        for (i in 1:nrow(comm)){
            for (j in 1:15){
                for (k in 1:runs){
                    v[k] = resArray[i,j,k]
                }
                if (j < 12 ) results[i,j] = mean(v)
                else results[i,j] = median(v)
            }
        }

    ##### curve (P/FD with curve fitting)
    }, curve = {
        results = matrix(0,nrow(comm),4)
        for (r in 1:runs){
            comm = comm[sample(nrow(comm)),, drop=FALSE]    ## shuffle rows (samples)
            runData = matrix(0,0,4)
            for (s in 1:nrow(comm)){
                n = sum(rowSums(comm[1:s,,drop=FALSE]))
                obs = dSobs(comm[1:s,], tree)
                runData = rbind(runData, c(s,n,obs,obs))
            }
            results = results + runData
        }
        results = results / runs
        for (s in 3:nrow(comm)){                ## fit curves only with 3 or more samples
            x = results[1:s,1]
            y = results[1:s,3]
            clench = try(nls(y ~ (a*x)/(1+b*x), start = list(a = 1, b = 1), control =
            nls.control(maxiter = 10000)),
silent = TRUE); ## does not stop in the case of error
            if(class(clench) != "try-error"){
                a = coef(clench)[1]
                b = coef(clench)[2]
                results[s,4] = a/b
            }
        }
        colnames(results) = c("Samples", "Ind", "Obs", "Clench")
        return (results)
    })
    results = rbind(rep(0,15), results)
    colnames(results) = c("Samples", "Ind", "Obs", "S1", "S2", "Q1", "Q2", "Jack1ab",
    "Jack1abP", "Jack1in",
"Jack1inP", "Chao1", "Chao1P", "Chao2", "Chao2P")
    return(results)
}

# Example of a tipical R session running this code
# tr # species X trait matrix
# comm # sites X species matrix
# tr.hc= hclust (tr)
# estimate (comm, tr.hc)
```